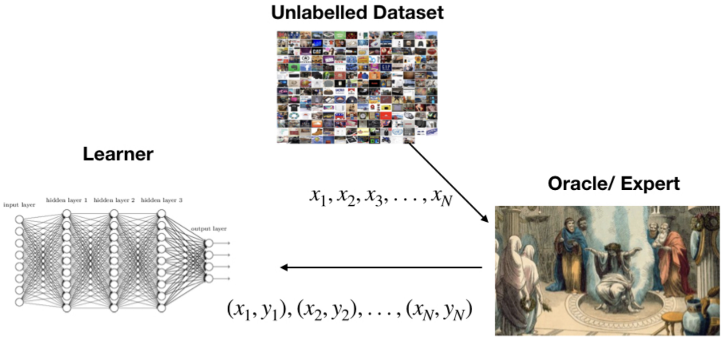
Active Learning for Individual Data via Minimal Stochastic Complexity

Shachar Shayovitz and Meir Feder

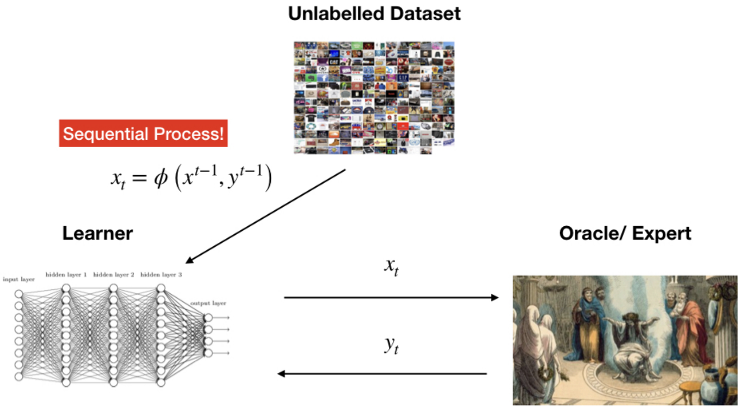
58th Annual Allerton Conference on Communication, Control, and Computing 2022

TEL AVIV UNIVERSITY אוניברסיטת תל אביב

Passive Learning

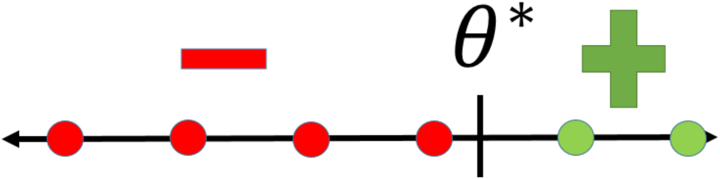


Active Learning

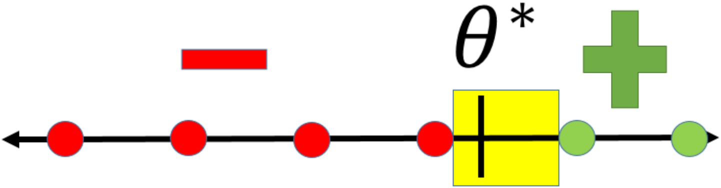


Motivating Example

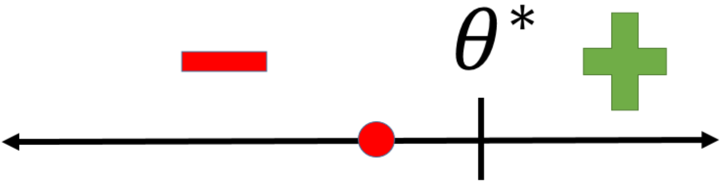
Motivating Example



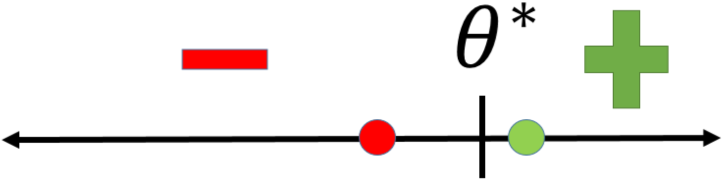
Motivating Example



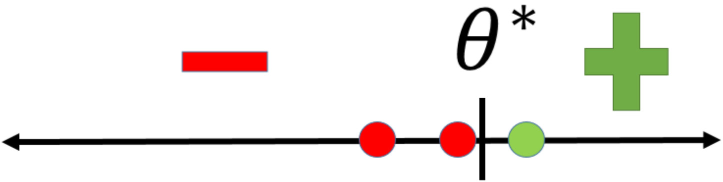
Motivating Example



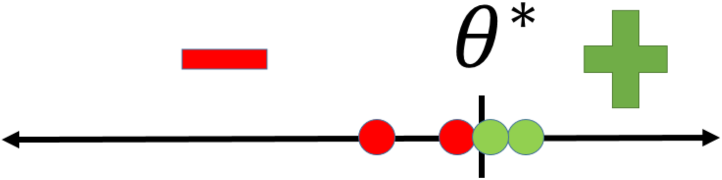
Motivating Example



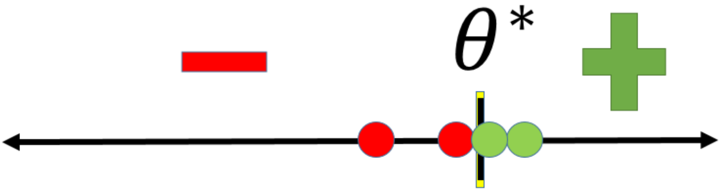
Motivating Example



Motivating Example



Motivating Example



Main Objective

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

Active Learning Criteria

- Maximum Uncertainty (MU)
  - $\hat{x}_n = \arg\max_{x_n} H\left(y_n | x^n, y^{n-1}\right)$.
  - Sensitive to noise.
- Bayesian Active Learning by Disagreement (BALD) [Houlsby, et al 2011]
  - $\hat{x}_n = \arg\max_{x_n} I\left(\theta; y_n | x^n, y^{n-1}\right)$.
  - Heuristic criteria.
- Universal Active Learning (UAL) [Shayovitz & Feder 2021]
  - $\hat{x}_n = \arg\min_{x_n} I\left(\theta; y | x, x^n, y^n\right)$.
  - Derived using the Capacity - Redundancy Theorem.
  - Takes into account the un-labelled test set.

Active Learning Criteria

- Maximum Uncertainty (MU)
  - $\hat{x}_n = \arg\max_{x_n} H\left(y_n | x^n, y^{n-1}\right)$.
  - Sensitive to noise.
- Bayesian Active Learning by Disagreement (BALD) [Houlsby, et al 2011]
  - $\hat{x}_n = \arg\max_{x_n} I\left(\theta; y_n | x^n, y^{n-1}\right)$.
  - Heuristic criteria.
- Universal Active Learning (UAL) [Shayovitz & Feder 2021]
  - $\hat{x}_n = \arg\min_{x_n} I\left(\theta; y | x, x^n, y^n\right)$.
  - Derived using the Capacity - Redundancy Theorem.
  - Takes into account the un-labelled test set.

Data assumed to follow some parametric distribution

Cannot validate for real world data!

Learning in Individual Setting

Assumptions

- No underlying parametric distribution.
- Training pool: $z^N = (x^N, y^N)$
- Test pair: (x, y)
  - x can be accessed.
  - y is not available.
- Probabilistic learners: $q(y|x)$.
- Log-loss cost function: $-\log\left(q\left(\cdot|x, z^N\right)\right)$.

Learning in Individual Setting

Assumptions

- No underlying parametric distribution.
- Training pool: $z^N = (x^N, y^N)$
- Test pair: (x, y)
  - x can be accessed.
  - y is not available.
- Probabilistic learners: $q\,(y|x)$.
- Log-loss cost function: $-\log\left(q\left(\cdot|x, z^N\right)\right)$.

Fundamental Problem

Minimizing the log-loss in the individual setting is ill-posed.

Learning in Individual Setting

Define a hypothesis class:

$$P_\Theta = \{p(y|x,\theta)\,|\theta \in \Theta\}$$

Define the learning problem:

$$R(x; z^n) = \min_q \max_{y \in \mathbb{Y}} \log \left(\frac{p\left(y|x,\hat{\theta}\right)}{q(y|x,z^n)} \right)$$

where:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \left[\log p(y, y^n|x, x^n, \theta) + \log(w(\theta))\right]$$

and

$$p\left(y|x,\hat{\theta}\right) \in P_\Theta$$

Predictive Normalized Maximum Likelihood (pNML) / Stochastic Complexity

Theorem (Fogel and Feder 2018)

The universal learner, q_{pNML}, minimizes $R(x; z^n)$:

$$q_{pNML}(y|x, z^N) = \frac{p\left(y|x, \hat{\theta}\right)}{\sum_y p\left(y|x, \hat{\theta}\right)}$$

$$R(x; z^n) = \log \sum_{y \in \mathbb{Y}} p\left(y|x, \hat{\theta}\right)$$

The pNML regret is exactly the stochastic complexity of P_Θ.

Active Learning in Individual Setting

What is a "good" training set, z^n?

Small $R(x; z^n)$ on many test features x

- Optimizing over z^n is not possible!
- Find training features x^n which minimize the worst case labels y^n:
  - Average mini-max regret:

$$C_n^A = \min_{x^n \in \mathbb{X}^n} \max_{y^n \in \mathbb{Y}^n} \sum_x R(x; z^n)$$

  - Worst mini-max regret:

$$C_n^W = \min_{x^n \in \mathbb{X}^n} \max_{y^n \in \mathbb{Y}^n} \max_x R(x; z^n)$$

Active Learning in Individual Setting

Using Fogel and Feder 2018:

Individual Active Learning (IAL)

$$C_n^A = \min_{x^n \in \mathbb{X}^n} \max_{y^n \in \mathbb{Y}^n} \sum_x \sum_{y \in \mathbb{Y}} p\left(y|x, \hat{\theta}\left(x, y, z^n\right)\right)$$

$$C_n^W = \min_{x^n \in \mathbb{X}^n} \max_{y^n \in \mathbb{Y}^n} \max_x \sum_{y \in \mathbb{Y}} p\left(y|x, \hat{\theta}\left(x, y, z^n\right)\right)$$

Active Learning in Individual Setting

- In the next slides we examine IAL for different hypothesis classes:
  - One dimensional Barrier
  - Linear Regression
  - Gaussian Process Classification
- It will be shown that IAL coincides with known class specific criteria and thus is a unified framework for active learning!

One Dimensional Barrier - Separable Data

The 1-dimensional barrier hypotheses class, P_Θ, is defined as:

$$p(y = 1|x, \theta) = \mathbb{1}(\theta < x)$$

where the input $x \in [0, 1]$, output $y \in \{0, 1\}$ and the unknown barrier $\theta \in [0, 1]$.

One Dimensional Barrier - Separable Data

The 1-dimensional barrier hypotheses class, P_Θ, is defined as:

$$p(y = 1|x, \theta) = \mathbb{1}(\theta < x)$$

where the input $x \in [0, 1]$, output $y \in \{0, 1\}$ and the unknown barrier $\theta \in [0, 1]$.

Theorem (Shayovitz & Feder 2022)

For one dimensional separable data, greedy IAL induces a selection policy which coincides with binary search and thus optimal.

Proof Outline

- IAL can be written as:

$$C^A_{n|n-1} = \min_{x_n \in \mathbb{X}} \max_{y_n \in \mathbb{Y}} \sum_{v \in \mathbb{V}} \int_{u \in \mathbb{U}} p\left(v|u, \hat{\theta}^n\right) du$$

where $\hat{\theta}^n$ is the maximum likelihood estimation based on training and test data:

$$\hat{\theta}^n = \arg\max_{\theta \in \Theta} p\left(y^n, v|x^n, u, \theta\right)$$

- Likelihood for z^{n-1}:

$$p\left(y^{n-1}|x^{n-1}, \theta\right) \propto \mathbb{1}\left(\theta \geq \theta^{n-1}_{min}\right) \mathbb{1}\left(\theta < \theta^{n-1}_{max}\right)$$

where θ^{n-1}_{min} and θ^{n-1}_{max} represent the support of the posterior on θ given x^{n-1}, y^{n-1}.

Proof Outline

- Select feature point, x_n, which can be any point in the support of $p\left(y^{n-1}|x^{n-1},\theta\right)$.
- For $y_n = 0$:

$$p\left(y^n|x^n,\theta\right) \propto \mathbb{1}\left(\theta \geq x_n\right)\mathbb{1}\left(\theta < \theta_{max}^{n-1}\right)$$

- For $y_n = 1$:

$$p\left(y^n|x^n,\theta\right) \propto \mathbb{1}\left(\theta \geq \theta_{min}^{n-1}\right)\mathbb{1}\left(\theta < x_n\right)$$

Proof Outline

- IAL can be simplified to:

$$\max_{y_n \in \mathbb{Y}} \int_{u \in \mathbb{U}} \left(\mathbb{1} \left(\hat{\theta}^n_{v=1} < u \right) + 1 - \mathbb{1} \left(\hat{\theta}^n_{v=0} < u \right) \right) du = \max\{l_0, l_1\}$$

where

$$l_0 = |1 - \theta^{n-1}_{max}| + 2|x_n - \theta^{n-1}_{max}| + |x_n|$$

and

$$l_1 = |\theta^{n-1}_{min}| + 2|\theta^{n-1}_{min} - x_n| + |1 - x_n|$$

Proof Outline

We note that:

$$l_0 = 1 + |x_n - \theta_{max}^{n-1}|$$

and

$$l_1 = 1 + |\theta_{min}^{n-1} - x_n|$$

Therefore,

$$C_{n|n-1}^{A} = \min_{x_n \in \mathbb{X}} \max\{|x_n - \theta_{max}^{n-1}|, |\theta_{min}^{n-1} - x_n|\}$$

The point x_n which minimizes the maximal length is the mid point of the interval $\left[\theta_{min}^{n-1}, \theta_{max}^{n-1}\right]$

Linear Regression

The linear regression hypothesis class:

$$\underline{y} = X\underline{\theta} + \underline{z}$$

where:

- $X \in \mathbb{R}^{nxp}$ is a design matrix of n feature vectors.
- $\underline{y} \in \mathbb{R}^n$ is the vector of observable responses.
- $\underline{z} \sim N\left(0, \sigma^2 \mathbb{I}_n\right)$.

The error covariance of the OLS solution is:

$$\Sigma^{-1} = \sigma^2 \left(X^T X\right)^{-1}$$

Experimental Design

- The design problem reduces to find a design matrix X which "minimizes" the covariance matrix $\Sigma^{-1} = \left(X^T X\right)^{-1}$.
- Extensive research over the last decade under the mathematical field of "Optimal Experimental Design": [Pukelsheim 2006]
  - **A** Optimal Design: $f_A(\Sigma) = \frac{1}{p} Tr\left(\Sigma^{-1}\right)$
  - **D** Optimal Design: $f_D(\Sigma) = det\left(|\Sigma|\right)^{-\frac{1}{p}}$
  - **G** Optimal Design: $f_G(\Sigma) = \max_x \text{diag}\left(X\Sigma^{-1}X^T\right)$
  - **V** Optimal Design: $f_V(\Sigma) = Tr\left(X\Sigma^{-1}X^T\right)$

IAL for Linear Regression

Theorem (Shayovitz & Feder 2022)

For linear regression, IAL becomes:

$$C_n^A = \min_{x^n} \mathrm{Tr}\left(X\left(X_n^T X_n + \lambda I\right)^{-1} X^T\right)$$

$$C_n^W = \min_{x^n} \max_{x} \mathrm{diag}\left(X\left(X_n^T X_n + \lambda I\right)^{-1} X^T\right)$$

where X is a matrix which is a concatenation of the test vectors x and λ is a regularization factor.

Observations

- IAL coincides with G and V optimal designs:
- Note that IAL is a function of the training features x^n only and have no dependence on their respective labels y^n.
- Therefore, no need for online feedback and the training set selection can be cast as a subset selection problem performed offline.
- This problem is NP hard and approximate solutions are needed.

Gaussian Process Classification

Gaussian Process Classification (GPC) is a powerful, non-parametric kernel-based model.

$$f \sim GP(\mu(\cdot), k(\cdot, \cdot))$$

$$y|x, f \sim Bernoulli\left(\Phi\left(f_x\right)\right)$$

- f is a function of a feature point x and is assigned a Gaussian process prior with mean $\mu(\cdot)$ and covariance function $k(\cdot, \cdot)$.
- The label y is Bernoulli distributed with probability $\Phi(f_x)$, where Φ is the Gaussian CDF.

Variational Inference

- Given a training set, the posterior over f becomes non-Gaussian and complicated.
- Approximate inference is used to model the posterior distribution.

The MAP estimators $\hat{f}_{x_n}$ and $\hat{f}_u$ are computed based on:

$$\hat{f}_{x_n}^{y_n}, \hat{f}_u^{v} = \arg\max_{f_{x_n}, f_u} p(v|f_u)\, p(y_n|f_{x_n})\, q\left(f_{x_n}, f_u | y^{n-1}, x^{n-1}\right)$$

where
$q\left(f_{x_n}, f_u | y^{n-1}, x^{n-1}\right)$ is a Gaussian distribution.

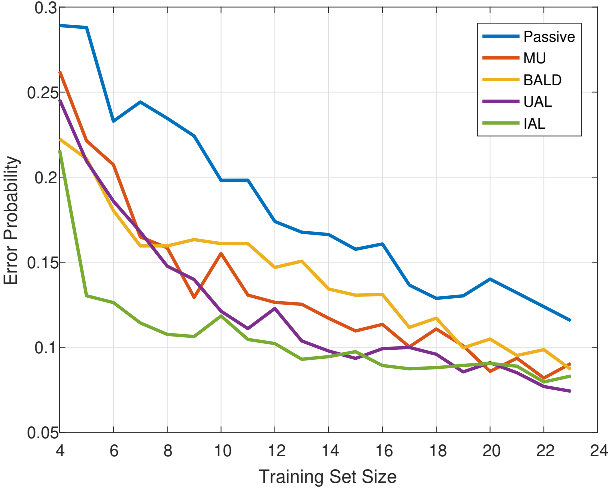
Data Set

- USPS hand-written digits data set.
- Total of 9298 handwritten single digits between 0 and 9.
- Each image consists of 16×16 pixels.
- Half of 9298 digits are designated as training and the other half are as test.
- Pixel values are normalized to be in the range of [-1, 1].

Empirical Comparison

- Binary classification task: the digit 0 vs $\{2, 4, 7, 8\}$.
- PCA is computed using the un-labeled training data.
- After centering and PCA, the 5 largest Eigenvalues of the PCA are used as the feature space for classification.
- A small random subset of the unlabeled test set is given to the learner (15 random samples) along with an initial labelled training set (3 random examples).
- IAL is compared to UAL, BALD, MU and passive learning.

Empirical Results



Summary

- Presented a novel AL criterion:
  - IAL takes into account the **un-labelled** test set.
  - IAL is not constrained by the assumption that the data is generated by some class of distributions.
- IAL can be viewed as a unified framework for active learning in a variety of hypothesis classes:
  - For binary classification, this criterion coincides with binary search
  - For linear regression, this criterion coincides with G and V optimal designs.
- In empirical comparison with state of the art AL criteria, IAL proved to be superior in terms of sample complexity.

Thank You!