Active Data Acquisition with Side Information

An Vuong Thinh Nguyen
School of Electrical and
Computer Engineering
Oregon State University
Corvallis, OR, 97331
Email: {vuonga2, thinhq}@oregonstate.edu

Anthony Q Nguyen
Email: anth.nguyen2357@gmail.com

Thuan Nguyen
Department of Engineering
East Tennessee State University
Johnson City, 37614
Email: nguyent11@etsu.edu

***Abstract*—This paper addresses the challenge of active data acquisition by proposing a method that balances the trade-off between acquisition cost and data fidelity. Lowering acquisition costs can compromise data quality, leading to the loss of valuable information and potentially resulting in misclassifications in subsequent supervised learning tasks. Conversely, high-fidelity data acquisition often incurs increased costs in terms of power consumption, storage, and the risk of collecting irrelevant data for the target application. To address this, we introduce an information theoretic framework that balances between specificity and generality, using mutual information as a measure of data relevance, as it provides a robust metric for a wide range of target applications. We demonstrate that the data acquisition problem is inherently challenging. However, it can be effectively approximated using a proposed heuristic approach. Furthermore, in scenarios where the probabilistic model of data acquisition is unknown, we illustrate through a synthetic example how a deep neural network (DNN) can be trained to learn effective data acquisition actions.**

Keywords: Entropy, mutual information, convexity.

I. Introduction

Modern data science increasingly relies on analyzing large volumes of data, often, the more data, the better. However, the quality of the data is heavily influenced by the methods used to acquire it. While accurate and comprehensive data acquisition techniques can yield high-fidelity data, they often come at the cost of increased power consumption and processing time, and may also raise the risk of collecting irrelevant data for specific task. On the other hand, reducing these costs may result in lower data fidelity, leading to loss of information that are crucial to the performance of downstream analysis and tasks.

Consider the scenario of object detection using radar. In this case, the radar emits signals and analyzes the reflected waves to identify and classify objects. Because certain objects may only respond to signals within specific frequency and power ranges, it is crucial to optimize the transmission parameters to target these effective ranges, ensuring accurate and efficient data acquisition while avoiding unnecessary use of power and frequency resources. This example underscores the importance of tailoring data acquisition to align with specific goals or tasks, especially under resource constraints. However, a highly goal-specific approach can lead to an overly specialized dataset that may exclude valuable information needed for related tasks. Ideally, we aim to collect a dataset that is rich and versatile enough to support not only the target task but also similar, yet unspecified, ones. For instance, a dataset containing a wide range of animal images is more useful than one restricted to only cats and dogs.

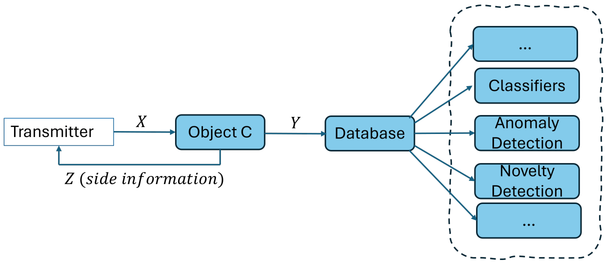

Figure 1. Active data acquisition with side information

We introduce an information theoretic perspective that offers a valuable approach to balancing specificity and generality in the active data acquisition problem. Instead of solely optimizing data collection under resource constraints to improve performance on a specific task, our objective is to acquire data that captures the most relevant information for a broad set of potential tasks without relying on any particular learning algorithm. To achieve this, mutual information is used as a measure of relevance. Unlike simpler metrics such as correlation, mutual information is more robust and not tied to any specific type of learning model [1]. In this paper, we describe an architecture for incorporating *side information* into the design of data acquisition strategies under resource constraints. As shown in Fig. 1, we aim to select the measurement action X based on the side information Z to yield a measured result Y that shared the most information with target C. We demonstrate that the data acquisition problem is inherently challenging. However, it can be effectively approximated using a proposed heuristic approach. Furthermore, in scenarios where the probabilistic model of data acquisition is unknown, we illustrate through a synthetic example how a deep neural network (DNN) can be trained to learn effective data acquisition actions.

II. Related Work

We begin with a brief review of the literature on target detection, a key example of active data acquisition. Target identification has been extensively studied across various scientific and engineering domains. Traditionally, these techniques have been model-driven, favored for their mathematical simplicity, efficiency, and the limited availability of training data. Model-based approaches leverage prior knowledge, derived from physical laws or well-established heuristics to estimate a small set of parameters for accurate object identification. As a result, they remain highly effective in domains like radar systems [2], [3], where object behavior and characteristics can be captured through a few well-defined parameters.

However, with the growing availability of large datasets, the field has increasingly shifted toward supervised learning. In areas such as computer vision [4], [5] and radar-based target recognition [6], data-driven methods have replaced traditional models, allowing algorithms to learn directly from data.

In contrast to these approaches, our work does not propose a specific target detection algorithm. Instead, we focus on optimizing the acquisition of information that can be used by downstream classification algorithms, aiming to maximize the relevance of the collected data under resource constraints.

Our work also shares connections with active learning [7], [8], which aims to select informative samples for labeling in order to improve the performance of a learning algorithm, particularly when data collection is expensive. However, unlike most active learning methods that are tailored to specific models, our approach does not optimize for any particular learning algorithm. Instead, we focus on maximizing mutual information, aligning more closely with the information-theoretic perspective presented in [9].

In addition, recent work by Shayovitz et al. introduced a universal active learning framework that minimizes conditional information [10]. While this method is conceptually related to our own, both being grounded in information theory, our approach differs in its specific objectives and emphasis on data acquisition strategy rather than active label querying.

III. Problem Description

A. Motivated Scenario

Object Identification: In this setup, a detector (e.g., radar) sends a signal $x \in \mathbb{R}^m$ toward an object $c \in \mathcal{C} = \{c_0, c_1, \ldots, c_{k-1}\}$, and receives a reflected signal $y \in \mathbb{R}^n$ that depends on the object's identity. The goal is to identify c_i based on the received signal y. Traditional methods rely on physical models to identify the object from its signature. In contrast, supervised learning methods like deep neural networks learn a classifier $f(y) : \mathbb{R}^n \to \mathcal{C}$ from labeled data pairs (y, c), requiring a large dataset before training. Typically, the signal x is selected based on prior knowledge, which then determines the observed signature y. However, choosing x comes with costs, stronger signals, e.g., larger $x^T x$ require more power, and high-dimensional y increases storage and training complexity. Moreover, the performance of the classifier is highly influenced by the choice of x. Ideally, the selection of x should also incorporate additional side information about c, such as real-time data from other sensors when available, to improve accuracy and efficiency. Hence, we propose an information-theoretic framework that incorporates the costs associated with both x and y. We demonstrate how to optimally select x, both with and without side information, i.e., additional knowledge about the object c, to ensure that the resulting signal y carries the most relevant information about the object c, while satisfying all cost constraints. Crucially, the resulting dataset is not tailored to any specific classifier. Instead, our approach generates a general-purpose dataset suitable for different downstream tasks, effectively decoupling data acquisition from the choice of classifier.

B. Information Theoretic Formulation

To aid our discussion, we use the following notations:

Symbol	Description
$X, p(x)$	Sensing signal and its distribution
$C, p(c)$	Target and its distribution
$Y, p(y)$	Measurements and its distribution
$Z, p(z)$	Side information and its distribution

Table I
Notation summary for sensing and measurement model

Definition 1. *An active data acquisition model with side information is specified by the prior $p(c)$, the conditionals $p(z|c)$ and $p(y|x, c)$.*

$p(c)$ represents the prior distribution over the target variable c, $p(z|c)$ models the side information conditioned on the target, and $p(y|x, c)$ describes the measurement distribution given the sensing action x and the target c. Additionally, instead of searching for a single optimal action or sensing signal x^*, we aim to find the optimal distribution $p^*(x)$ over possible actions.

Figure 2 illustrates a graphical model for a special case of data acquisition where x and C are independent. This corresponds to a setting where no information about C is available beyond its prior $p(c)$ and the measurement model $p(y|x, c)$. In this case, given $p(c)$ and $p(y|x, c)$, the goal is to find the optimal sensing distribution $p^*(x)$ that maximizes either $I(Y; C)$ or $I(X, Y; C)$.

In contrast, Figure 3 depicts the more general scenario where the optimal action x^* can depend on side information Z related to the target C. Here, the objective is to determine the optimal conditional distribution $p^*(x|z)$ that maximizes either $I(Y; C)$ or $I(X, Y; C)$, given the known distributions $p(c)$, $p(z|c)$ and $p(y|x, c)$.

Maximizing $I(X, Y; C)$ assumes that both the sensing signal X and the corresponding measurement Y are available and used to extract information about the target C, such as in training scenarios where both are stored. In contrast, when X is not available or not used during training, the goal shifts

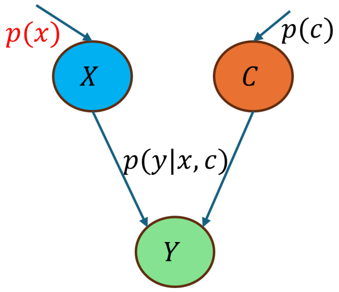

Figure 2. Active data acquisition model: X and C are independent

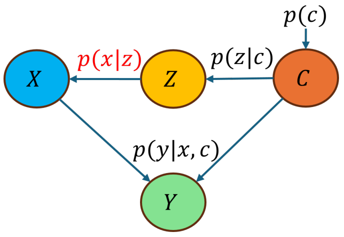

Figure 3. Active data acquisition model: X and C are dependent through Z

to maximizing the mutual information between Y and C alone. Based on these settings, we formulate the following optimization problems for active data acquisition without side information:

Problem P1:

$$\begin{aligned} \max_{p(x)} \quad & I(X,Y;C) \\ \text{s.t.} \quad & g_i(p(x)) \leq 0, i = 1, \ldots, N \\ & p(x) \geq 0,\ p(x) \leq 1,\ \sum_x p(x) = 1. \end{aligned} \tag{1}$$

Problem P2:

$$\begin{aligned} \max_{p(x)} \quad & I(Y;C) \\ \text{s.t.} \quad & g_i(p(x)) \leq 0, i = 1, \ldots, N \\ & p(x) \geq 0,\ p(x) \leq 1,\ \sum_x p(x) = 1. \end{aligned} \tag{2}$$

In both **P1** and **P2**, $g_i(p(x))$ are the given constraints modeling the costs associated the acquisition while the constraints on $p(x)$ enforce the validity of a distribution. Assume that $g_i(p(x))$ are convex functions with respect to $p(x)$, which is typical in real-world scenarios. For example, consider a scenario where the average transmit power must not exceed a certain threshold, expressed as $g(p(x)) = \mathbb{E}[X^2] < T$.

Similarly, for active data acquisition with side information, with Z being a latent representation modeled by $p(z|c)$, we have the following optimization problems:

Problem P3:

$$\begin{aligned} \max_{p(x|z)} \quad & I(X,Y;C) \\ \text{s.t.} \quad & g_i(p(x|z)) \leq 0, i = 1, \ldots, N \\ & p(x|z) \geq 0,\ p(x|z) \leq 1,\ \sum_x p(x|z) = 1. \end{aligned} \tag{3}$$

Problem P4:

$$\begin{aligned} \max_{p(x|z)} \quad & I(Y;C) \\ \text{s.t.} \quad & g_i(p(x|z)) \leq 0, i = 1, \ldots, N \\ & p(x|z) \geq 0,\ p(x|z) \leq 1,\ \sum_x p(x|z) = 1. \end{aligned} \tag{4}$$

IV. Solution Approach

A. Solution Characterization

Theorem 1. *If $g_i(p(x))$ are linear, then* **P1** *is a linear programming problem. Furthermore, $p^*(x)$ must be one of the extreme points (vertices) formed by intersections among the constraints. When there is no constraint $g_i(p(x))$, $p^*(x)$ must lie on one of the vertices of the probability simplex, i.e., the optimal solution is to send $X = x^*$ for some constant x^*.*

Proof. See Appendix □

Theorem 2. *If $g_i(p(x))$ are convex functions, then* **P2** *is a convex maximization problem. Furthermore, $p^*(x)$ must be one of the extreme points (vertices) formed by intersections of the given convex constraints. When there is no constraint $g_i(p(x))$, $p^*(x)$ must lie on one of the vertices of the probability simplex, i.e., the optimal solution is to send $X = x^*$ for some constant x^*.*

Proof. See Appendix □

Theorem 3. *If $g_i(p(x|z))$ are linear, then* **P3** *is a convex maximization problem. Furthermore, $p^*(x|z)$ must be one of the extreme points (vertices) from by intersections of the given convex constraints. When there is no constraint $g_i(p(x|z))$, $p^*(x|z)$ must lie on one of the vertices of the probability simplex.*

Proof. See Appendix □

Theorem 4. *If $g_i(p(x|z))$ are convex functions then* **P4** *is a convex maximization problem. Furthermore, $p^*(x|z)$ must be one of the extreme points (vertices) from by intersections of the given convex constraints. When there is no constraint $g_i(p(x|z))$, $p^*(x|z)$ must lie on one of the vertices of the probability simplex.*

Proof. Similar to the proof of Theorem 2 and 3. □

The optimal solutions to problems **P1**-**P4** lie at extreme points of the feasible region. To find these solutions, one must enumerate and search over these extreme points. While this process is generally NP-hard in high dimensions, there exist algorithms that can reliably yield locally optimal solutions [11], [12]. Notably, when the constraints involve entropy functions, the approach proposed in [13] can be directly applied.

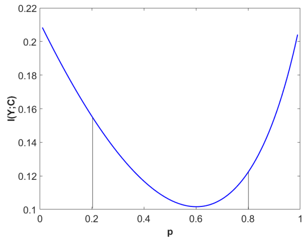

Figure 4. $I(Y;C)$ of additive channel; $a = 0.2$ and $b = 0.8$.

B. Illustrative Examples

In general, identifying extreme points in high-dimensional spaces is an NP-hard problem. However, in this section, we present simple examples defined over low dimensional probability simplexes, where optimal solutions can be derived either analytically in closed form or through straightforward numerical methods.

Example IV.1. (Additive Channel - P1 and P2) *Let $X \sim Bern(p)$ represent the active signal X. Let $C \sim Bern(q)$ represent the two possible objects: $C = 0$ and $C = 1$. Let*

$$Z_0 \sim Bern(q_0)$$
$$Z_1 \sim Bern(q_1).$$

If a signal X is transmitted, then the received signal is

$$Y = \begin{cases} X + Z_0 & \text{if object 0 is present} \\ X + Z_1 & \text{if object 1 is present} \end{cases}$$

The goal is to determine $p^*(x)$ that maximizes $I(X, Y; C)$ subject to $a \leq \mathbb{E}[X^2] \leq b$,, for some constants $a < b$.

Assuming C is independent with X, then

$$I(C; X, Y) = H(q_0 + qq_0 - qq_1) - H(q_0)(1 - q) - H(q_1)q$$

where $H(r) \overset{\triangle}{=} -r \log r - (1 - r) \log (1 - r)$ denotes the entropy of a binary random variable. This is a special case where the mutual information $I(C; X, Y)$ is constant regardless of $p(x)$. Hence, any distribution $p(x)$ would yield the same mutual information. On the other hand, to satisfy the constraint $a \leq \mathbb{E}[X^2] \leq b$, we can choose any $p^*(x) = (1 - p, p)$ where $a \leq p \leq b$. Also, Fig. 4 shows $I(Y; C)$ (derivation included in the Appendix) with varying p, where $q_0 = 0.4, q_1 = 0.9, q = 0.6, a = 0.2, b = 0.8$. The optimal solution occurs at $p^* = 0.2$ which corresponds to maximum $I(Y; C) = 0.1556$. Note that $p^*(x) = (0.8, 0.2)$ is an extreme point which confirms Theorem 2. Derivation details can be found in the appendix.

Example IV.2. (Noisy OR Channel - P3 and P4) *In this example, we follow the model introduced in **P3** and **P4**. We consider $p(c) = Bern(q)$ $p(z|C = 1) = Bern(\alpha_1)$, $p(z|C = 0) = Bern(\alpha_0)$, $p(x|Z = 1) = Bern(\beta_1)$, $p(x|Z = 0) = Bern(\beta_0)$. The channel is then modelled as the Noisy OR Channel $p(Y = 1|X = x, C = c) = (1-\gamma_X)^x(1-\gamma_C)^c(1-\epsilon)$, where $\gamma_X, \gamma_C, \epsilon$ are parameters. We put an energy constraint on $p(x|z)$, i.e. $a \leq \mathbb{E}._{p(x|z)}[X^2] \leq b$*

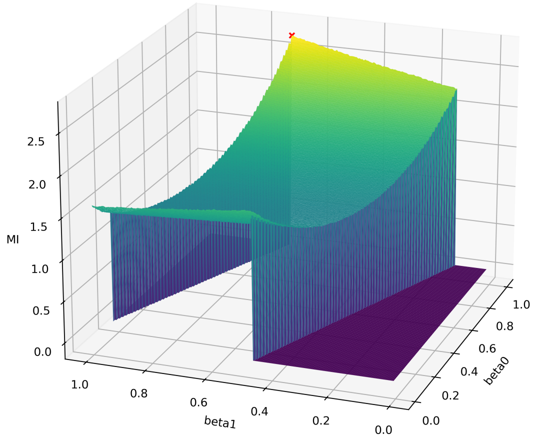

Figure 5. $I(Y;C)$ of Noisy OR Channel

For $q = 0.2$, $\alpha_0 = 0.7$, $\alpha_1 = 0.9$, $\gamma_X = 0.8$, $\gamma_C = 0.3$, $\epsilon = 0.2$, $a = 0.35$, $b = 0.78$, the maximum $I(Y; C) = 2.672$ and is attained at $(\beta_0, \beta_1) = (1, 0.701)$. The simulation is shown in Fig. 5, note that the optimal point (red cross) is one of the extreme points. Simulation details and more examples are provided in the appendix.

C. Iterative Algorithm using Concave-Convex Procedure (CCCP)

In this section, we propose the use of CCCP [11] to find a solution for **P3-4** iteratively. For demonstration purpose, we focus on **P4**, for simplicity, we limit $g_i()$ to the power constraint $E[X^2]$. Then we can solve the equivalent problem, with some parameter β

$$\begin{aligned} \min_{p(x|z)} \quad & E[X^2] - \beta I(Y; C) \\ \text{s.t.} \quad & p(x|z) \geq 0, \; p(x|z) \leq 1, \; \sum_x p(x|z) = 1. \end{aligned} \tag{5}$$

Since $I(Y; C)$ is a function of $p(x|z)$, at each iteration of CCCP, we can use Taylor's expansion to linearize $I(Y; C)$ about $p_{t-1}(x|z)$, which is the solution to (5) at iteration $t-1$. By doing so, (5) becomes a linear programming problem over the probability simplex. In order to perform the procedure, access to the gradient $\nabla_{p(x|z)} I(Y; C)$ is required. We provide the detailed computation in the appendix. The procedure is presented in Algorithm 1, for each β, we run this algorithm multiple times to obtain the optimal results. Applying this algorithm to the same Noisy OR Channel presented in Example IV.2 yields $(\beta_0, \beta_1) = (1, 0.72)$.

We note that in the case where the constraint is the entropy function, the CCCP can be simplified to a clustering approach [12][13].

Algorithm 1 Concave-Convex Procedure

```
Input: $p(c), p(z|c), p(y|x,c) g(\cdot)$, and $\beta$.
Output: $p(x|z)$
Initialization: Randomly initialize $p(x|z)$.
Step 1: Evaluate the gradient:
    $\frac{\partial I(Y;C)}{\partial p(x_k|z_l)}\Big|_{p_{t-1}(x_k,z_l)}$
Step 2: Solve the linear programming problem for each $Z_l$ over the probability simplex
    $p_t(x|Z_l) = \arg\min_{p(x|Z_l)} E[X^2] - \beta\big\langle \nabla_{p(x|z))} I(Y;C)\big|_{p_{t-1}(x|Z_l))} , p(x|Z_l)\big\rangle$
    $\langle . \rangle$ denotes the inner product.
Step 3: Go to Step 1 until the output $p(x|z)$ stop changing or the maximum number of iterations has been reached.
```

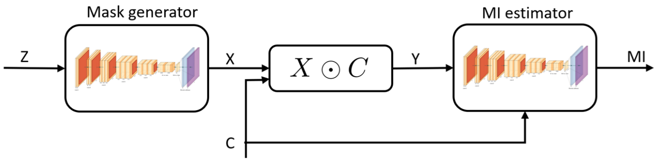

Figure 6. Neural networks architecture

V. Blind Estimation using Neural Networks

A. Motivation

In practice, we rarely have access to the underlying probabilistic models needed, i.e. $p(y|x,c)$ in **P1-2**, $p(z|c)$ and $p(y|x,c)$ in **P3-4**. To tackle this challenge, we now employ deep neural networks to jointly estimate the probabilistic models and maximizing the mutual information. In particular, focusing on the setup outlined in **P4**, our method involves the use of two networks, the first one is to represent the conditional $p(x|z)$, and the second one is to estimate the mutual information $I(Y;C)$ by leveraging the MINE framework [14]. To keep things simple, we use a deterministic model $Y = X \odot C$, where $\odot$ denotes the element-wise multiplication, and $Z = G * C$, where G is a Gaussian blur kernel. Our setup for 2D images is outlined in Fig. 6. Since we want X to be sparse, i.e. bottlenecking the information flowing from $C \rightarrow Y$, the following loss function is used during training:

$$\mathcal{L} = \mathcal{L}_{\text{MINE}} + \mathcal{L}_{\text{mask}} \tag{6}$$

$$\mathcal{L}_{\text{mask}} = \mathbb{E}_k\Big[\frac{1}{\epsilon + \sum_i (1 - \mathrm{X}^k)_i}\Big] \tag{7}$$

where k denotes the batch, i denotes the i-th element, and $\mathcal{L}_{\text{MINE}}$ is introduced in [14]. This $\mathcal{L}_{\text{mask}}$ encourages the mask to be sparse, as it grows logarithmically as number of 1's in X grows, ϵ is a small number for numerical stability.

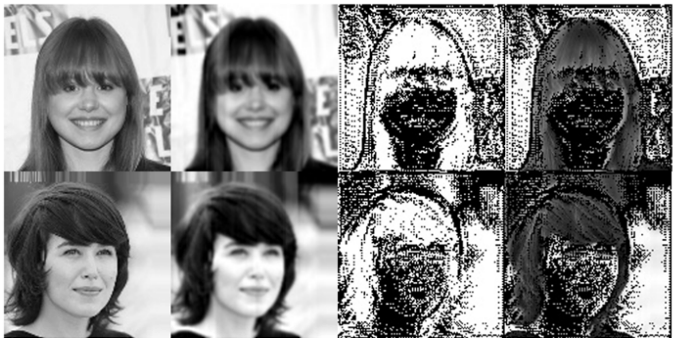

Figure 7. Samples generated from the Neural network. From left to right: Original images (C), Gaussian blurred images (Z), Masks generated (X), Reconstructed images ($Y = X \odot C$)

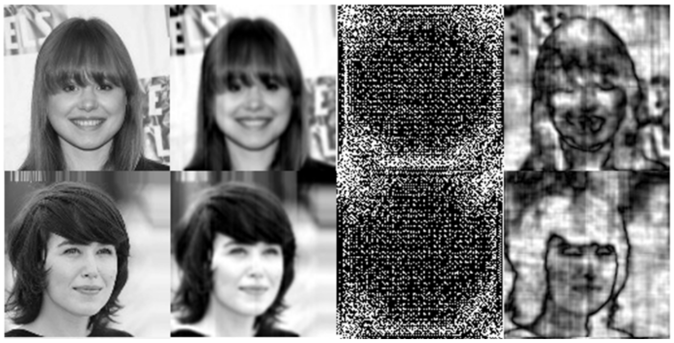

Figure 8. (Fourier domain) Samples generated from the Neural network. From left to right: Original images (C), Gaussian blurred images (Z), Masks generated in the frequency domain (X), Reconstructed images ($Y = \mathcal{F}^{-1}(X \odot C)$)

B. Experiments

We employ UNet [15] as the mask generator, and a simple 4-layer CNN with a Linear layer at the end for the MI estimator. Training was done using CelebA dataset [16]. We hypothesize that the generated masks will prioritize the high-frequency components of the image, i.e. edges, since these are where most information is stored. Some illustrative samples are presented in Fig. 7, we can observe that the masks mainly focus on the hair of the subjects, which contains high amount of information in this particular dataset. To actually verify our hypothesis, we re-ran the experiment in the Fourier domain, by applying the FFT to X and C, the results are shown in Fig. 8. We can see that in the Fourier domain, the masks X focus most of its weights on the four corners, whose regions contain high frequency components. This confirms our hypothesis. Thus, the network proposed in Fig. 6 can be used as a tool to locate interesting regions (across the dataset) by running it in the spatial domain, or detecting high-frequency component in the Fourier domain.

VI. Conclusion

We present an information-theoretic framework to tackle the challenges of data acquisition and classification under resource constraints. This approach provides a balanced solution to the active data acquisition problem by addressing both specificity and generality. Rather than optimizing data collection solely to enhance the accuracy of a specific learning algorithm, our objective is to gather data with the most relevant information for a wide range of potential tasks, all within the constraints of available resources, and without relying on any particular algorithm. We leverage mutual information to measure data relevance, as it is more robust than other metrics and not tied to specific learning algorithms. We also provided simulation results to demonstrate the effectiveness of our approach.

References

[1] A. Vuong, Anthony Nguyen, and Thinh Nguyen, “Active data acquisition - information theoretic approach,” in *IEEE Workshop on Computing, Networking and Communications (CNC)*. IEEE, 2025.

[2] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*, Prentice Hall, 1998.

[3] Sudan Han, Linjie Yan, Yuxuan Zhang, Pia Addabbo, Chengpeng Hao, and Danilo Orlando, “Adaptive radar detection and classification algorithms for multiple coherent signals,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 560–572, 2021.

[4] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu, “Object detection with deep learning: A review,” *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[5] Shivang Agarwal, Jean Ogier Du Terrail, and Frédéric Jurie, “Recent advances in object detection in the age of deep convolutional neural networks,” *arXiv preprint arXiv:1809.03193*, 2018.

[6] Uttam Majumder, Erik Blasch, and David Garren, *Deep Learning for Radar and Communications Automatic Target Recognition*, Artech, 2020.

[7] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang, “A survey of deep active learning,” *CoRR*, vol. abs/2009.00236, 2020.

[8] Ozan Sener and Silvio Savarese, “Active learning for convolutional neural networks: A core-set approach,” 2018.

[9] Yuhong Guo and Russell Greiner, “Optimistic active-learning using mutual information.,” in *IJCAI*, 2007, vol. 7, pp. 823–829.

[10] Shachar Shayovitz and Meir Feder, “Universal active learning via conditional mutual information minimization,” *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 2, pp. 720–734, 2021.

[11] Alan L Yuille and Anand Rangarajan, “The concave-convex procedure (cccp),” *Advances in neural information processing systems*, vol. 14, 2001.

[12] Thuan Nguyen and Thinh Nguyen, “Minimizing impurity partition under constraints,” 2019.

[13] Daniel J Strouse and David J Schwab, “The deterministic information bottleneck,” *Neural computation*, vol. 29, no. 6, pp. 1611–1630, 2017.

[14] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm, “Mine: Mutual information neural estimation,” 2021.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” 2015.

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, “Deep learning face attributes in the wild,” in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[17] Vašek Chvátal, *Linear programming*, Macmillan, 1983.

[18] Thomas Cover and Joy Thomas, *Elements of information theory*, Wiley-Interscience, 2 edition, 7 2006.

[19] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, USA, 2004.

[20] Philip B Zwart, “Global maximization of a convex function with linear inequality constraints,” *Operations Research*, vol. 22, no. 3, pp. 602–609, 1974.

Appendix

A. Proofs of Theorems

Theorem 1 *If $g_i(p(x))$ are linear, then* **P1** *is a linear programming problem. Furthermore, $p^*(x)$ must be one of the extreme points (vertices) formed by intersections among the constraints. When there is no constraint $g_i(p(x))$, $p^*(x)$ must lie on one of the vertices of the probability simplex, i.e., the optimal solution is to send $X = x^*$ for some constant x^*.*

Proof. For clarity, assume X, Y, and C are discrete random variables.

$$I(X,Y;C) = \sum_{x,y,c} p(x,y,c) \log \frac{p(x,y,c)}{p(x,y)p(c)} \quad (8)$$

$$= \sum_{x,y,c} p(y|x,c)p(x)p(c) \log \frac{p(y|x,c)p(x)p(c)}{p(c)p(x)p(y|x)}$$

$$= \sum_{x,y,c} p(y|x,c)p(x)p(c) \log \frac{p(y|x,c)}{p(y|x)}. \quad (9)$$

Since $p(y|x) = \sum_c p(y|x,c)p(c)$ which is a constant. Thus, Eq. (9) is a linear with $p(x)$. Thus, maximizing $I(X,Y;C)$ with respect to $p(x)$ is a linear programming problem subject to the constraints on a valid distribution and linear $g_i(p(x))$.

From the well-known linear programming result [17], if $I(X,Y;C)$ is not a constant function then $p^*(x)$ must be at an extreme point formed by the intersections between the probability simplex and the linear constraints.

Similar proof can be carried out for continuous random variables X. Furthermore, without any constraint $g_i(p(x))$, $p^*(x) = \delta(x - x^*)$ for some x^*. □

Theorem 2 *If $g_i(p(x))$ are convex functions, then* **P2** *is a convex maximization problem. Furthermore, $p^*(x)$ must be one of the extreme points (vertices) formed by intersections of the given convex constraints. When there is no constraint $g_i(p(x))$, $p^*(x)$ must lie on one of the vertices of the probability simplex, i.e., the optimal solution is to send $X = x^*$ for some constant x^*.*

Proof. As before, assume that X, Y, C are discrete random variables. From a well-known result in information theory[18], for a fixed $p(c)$, $I(C;Y)$ is a convex function with respect to $p(y|c)$ where $p(y|c)$ is considered as a channel matrix, C as input and Y as output. Since $p(y|c) = \sum_x p(y|c,x)p(x) = T(p(x))$, where T is a linear map specified by the given $p(y|x,c)$. For any convex function $f(z)$ in z, let $z = T(w)$ where T is any affine map, then $f(T(w))$ is also convex in w [19]. A quick proof is as follows: If $f(z)$ is a convex function, then if and only if for any z_0, there exists an affine map $g(z)$ such that $g(z_0) = f(z_0)$ and $g(z) \leq f(z)$ for all z. Applying an affine map T, $z = T(w)$ in the inequality, and note that $g(T(w))$ remains an affine map since the composite of two affine maps $g \circ T$ is a new affine map for which the inequality hold for all w. Therefore $f(w)$ is also convex in w. Consequently, since $g_i(p(x))$ are convex functions, then finding the optimal $p^*(x)$ that maximizes $I(Y;C)$ is a convex maximization problem. The proof for continuous random variables can be carried out in the same manner.

From the well-known optimization result [20], the solution to the convex maximization problem must be one of the extreme points formed by the intersections of the constraints. □

Theorem 3 *If $g_i(p(x|z))$ are linear, then* **P3** *is a convex maximization problem. Furthermore, $p^*(x|z)$ must be one of the extreme points (vertices) from by intersections of the given convex constraints. When there is no constraint $g_i(p(x|z))$,*

$p^*(x|z)$ *must lie on one of the vertices of the probability simplex.*

Proof. Similar to the proof of Theorem 2. Since $I(X,Y;C)$ is a convex function with respect to $p(x,y|c)$ for a fixed $p(c)$:

$$\begin{aligned} p(x,y|c) &= \frac{p(x,y,c)}{p(c)} = \frac{1}{p(c)} p(y|x,c)p(x,c) \\ &= \frac{1}{p(c)} p(y|x,c) \sum_z p(x|z)p(z|c)p(c) \\ &= \frac{1}{p(c)} \sum_z p(y|x,c)p(x|z)p(z,c) \end{aligned}$$

which is linear in $p(x|z)$. Thus, $I(Y;C)$ is convex in $p(x|z)$, using the same argument as in the proof for Theorem 2. The proof for the optimal solution follows as before. □

B. Gradient Computation

We provide the calculation of $\nabla_{p(x|z)} I(Y;C)$, which is needed for Algorithm 1.

$$\begin{aligned} I(Y;C) &= \sum_{y,c} p(y,c) \log \frac{p(y,c)}{p(y)p(c)} \\ &= \underbrace{\sum_{y,c} p(y,c) \log p(y,c)}_{I_1} - \underbrace{\sum_{y,c} p(y,c) \log p(c)}_{I_2} \\ &\quad - \underbrace{\sum_{y,c} p(y,c) \log p(y)}_{I_3} \end{aligned} \tag{10}$$

For simplification, we calculate the partial gradients for each pair (x_k, z_l) as follows:

$$p(y_i, c_j) = \sum_k p(y_i, c_j, x_k) \tag{11}$$

$$= \sum_k p(y_i|c_j, x_k) \sum_l p(x_k|z_l)p(z_l|c_j)p(c_j) \tag{12}$$

$$\frac{\partial p(y_i, c_j)}{\partial p(x_k|z_l)} = p(y_i|c_j, x_k)p(z_l|c_j)p(c_j) \tag{13}$$

$$\frac{\partial I(Y;C)}{\partial p(x_k|z_l)} = \frac{\partial I(Y;C)}{\partial p(y_i, c_j)} \frac{\partial p(y_i, c_j)}{\partial p(x_k|z_l)} \tag{14}$$

$$\frac{\partial I_1(Y;C)}{\partial p(x_k|z_l)} = \sum_{i,j} \big(\log p(y_i, c_j) + 1\big) p(y_i|c_j, x_k)p(z_l|c_j)p(c_j) \tag{15}$$

$$= \sum_{i,j} \Big(\sum_k p(y_i|c_j, x_k) \sum_l p(x_k|z_l)p(z_l|c_j)p(c_j) + 1 \Big) \times p(y_i|c_j, x_k)p(z_l|c_j)p(c_j) \tag{16}$$

$$\frac{\partial I_2(Y;C)}{\partial p(x_k|z_l)} = \sum_{i,j} p(y_i|c_j, x_k)p(z_l|c_j)p(c_j) \log p(c_j) \tag{17}$$

$$= \sum_j p(z_l|c_j)p(c_j) \log p(c_j) \sum_i p(y_i|c_j, x_k) \tag{18}$$

$$= \sum_j p(z_l|c_j)p(c_j) \log p(c_j) \tag{19}$$

Since $\frac{\partial I_3(Y;C)}{\partial p(x_k|z_l)}$ is difficult to calculate directly, we use a common approximation. At every iteration t, we assume $p_t(y)$ is constant, computed as:

$$p_t(y_i) \approx \sum_j p_{t-1}(y_i, c_j) \tag{20}$$

$$p_{t-1}(y_i, c_j) \approx \sum_k p(y_i|c_j, x_k) \sum_l p_{t-1}(x_k|z_l)p(z_l|c_j)p(c_j) \tag{21}$$

Thus,

$$\frac{\partial I_3(Y;C)}{\partial p(x_k|z_l)} = \log p_t(y_i)p(y_i|c_j, x_k)p(z_l|c_j)p(c_j) \tag{22}$$

$$= \sum_{i,j} p(y_i|c_j, x_k)p(z_l|c_j)p(c_j) \times \log \Bigg(\sum_j \sum_k p(y_i|c_j, x_k) \sum_l p_{t-1}(x_k|z_l)p(z_l|c_j)p(c_j) \Bigg) \tag{23}$$