Convergence of Vector Quantization–Based Classifiers to the Bayes Optimal Classifier with Applications to Hybrid System Identification

Aneesh Raghavan ANEESH@KTH.SE
Christos N. Mavridis MAVRIDIS@KTH.SE
Karl H. Johansson KALLEJ@KTH.SE
DCS Division, KTH, Royal Institute of Technology, Stockholm

John S. Baras BARAS@UMD.EDU
Department of Electrical and Computer Engineering, University of Maryland, College Park

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

Vector quantization techniques have been extensively explored as interpretable, data-driven approaches within machine learning, demonstrating significant utility in hybrid system identification. In this study, we establish convergence guarantees for a general framework of quantization-based classifiers, encompassing histogram-based methods, variants of the generalized Lloyd's algorithm, learning vector quantization, and online deterministic annealing techniques. Utilizing principles from histogram estimation, we analyze the conditions under which these algorithms converge to the Bayes optimal error. These findings provide a rigorous theoretical foundation for the application of quantization-based algorithms in machine learning tasks associated with cyber-physical systems. An illustrative application in hybrid system identification is also presented.

Keywords: Statistical learning for dynamical and control systems, Optimization for machine learning, System identification.

1. Introduction

Vector Quantization (VQ) methods, initially introduced over three decades ago for data compression (Gersho and Gray, 2012; Gray, 1990), have since been extensively studied and employed as both supervised and unsupervised learning algorithms, as they offer explainability, robustness, and topology-preserving properties (Uriarte and Martín, 2005). Owing to their well-developed mathematical foundations, VQ methods often serve as compelling alternatives to state-of-the-art neural network architectures. Consequently, they continue to be studied alongside modern neural network models (Saralajew et al., 2018; Villmann et al., 2017a) and applied to a variety of applications including classification (Villmann et al., 2017b), clustering (Shah and Koltun, 2018), and time series and speech analysis (Melchert et al., 2016; Wang et al., 2019).

In particular, the role of VQ methods in hybrid system identification has been well demonstrated and continues to be studied (Mavridis and Johansson, 2025; Mavridis et al., 2024; Mavridis and Baras, 2023b; Wang et al., 2020; Ferrari-Trecate et al., 2003; Gegundez et al., 2008; Bianchi et al., 2020; Yu et al., 2023). Moreover, recent research has demonstrated that VQ methods exhibit remarkable robustness against adversarial attacks, indicating a potential advantage over neural network architectures in security-critical applications (Saralajew et al., 2019).

VQ methods can take many different forms, from standard histogram approaches, to more advanced methods, including Expectation-Maximization (EM) (Banerjee et al., 2005; Devroye et al., 2013),

stochastic approximation (Mavridis and Baras, 2020; Bottou, 1998; Baras and LaVigna, 1991), and annealing optimization methods (Rose, 1998; Mavridis and Baras, 2023c,a). In this work, we provide convergence guarantees for a general framework of VQ-based classification methods, commonly found in hybrid system identification applications.

We consider the binary classification problem. Every point in a given domain is classified either as 0 or 1. The joint distribution between samples from the domain and their corresponding label is unknown. The true classifier is also unknown. Given two independent sequences of data, one labeled and another unlabeled, the following algorithm is considered. Using the unlabeled data, VQ based partitions of the domain is obtained. Using the labeled data, the classifier for each partition, and thus for the entire domain, is obtained using a "majority" vote. The objective is to prove that under certain sufficient conditions, the sequence of classifiers obtained through the partitioning converges to the Bayes classifier, available only if the true joint distribution was known.

Under the assumptions that the true conditional densities are continuous, the VQ generated partitions are "well behaved", i.e., the size of the partitions converges to zero, and at every iteration each partition is rich enough with data points which carry both labels, we prove that the sequence of classifiers converges to the true Bayesian classifier. The progression of the proof follows similar principles as in Raghavan and Baras (2021) and is achieved in three steps: proving the convergence of the distribution estimators to the true conditional distributions, providing a measure-theoretic definition of the Bayes classifier, and showing the convergence of empirical classifiers and loss to the Bayes classifier and loss. These findings provide a rigorous theoretical foundation for the application of quantization-based algorithms in machine learning tasks associated with cyber-physical systems. An illustrative application in hybrid system identification is also presented.

The outline of the paper is as follows: In Section 2, we formally define the problem to be solved in this paper. In Section 3, we present the solution to the problem in three steps as described previously. In Section 4, we present examples demonstrating the result for a general classification problem and in the context of hybrid system identification. We conclude with future directions in Section 5.

2. Problem Formulation

We begin this section with the measure theoretic definition of conditional density which is essential for the rest of the paper. The optimization problem to be solved for obtaining VQ based partitions is then defined. For every iteration n, given the partitions and the labeled data, the empirical joint distribution is defined which is then utilized in defining the empirical classifier. The sufficient conditions for the convergence and the problem to be solved is stated in Problem 1.

Definition 1 *Let (X, c) be random variables on the abstract probability space $(\Omega, \mathbb{F}, \mathbb{P})$ such that $X \in S \subseteq \mathbb{R}^{d_X}$, S compact, and $c \in \{0, 1\}$. The canonical probability space for this pair of random variables is $\left(\mathbb{R}^{d_X} \times \{0,1\}, \sigma\Big(\mathcal{B}(\mathbb{R}^{d_X}) \times \{\varnothing, \{0\}, \{1\}, \{0,1\}\}\Big), \mathcal{P}\right)$. The conditional measure $\mathcal{P}^i(\cdot)$ is defined on the σ-algebra $\mathcal{B}(\mathbb{R}^{d_X})$ as*

$$\mathcal{P}^i(E) = \frac{\mathcal{P}(X \in E \ \cap c = i)}{\mathcal{P}(c = i)}$$

Let λ denote the Lebesgue measure on $\mathbb{R}^{d_X}$. Suppose $\mathcal{P}^i$ is absolutely continuous with respect to λ, $\mathcal{P}^i \ll \lambda$, i.e., $\lambda(E) = 0 \implies \mathcal{P}^i(E) = 0$. By the Radon-Nikodym theorem, there exists a

measurable function $f^i : \mathbb{R}^{d_X} \to [0, \infty)$ *such that*

$$\mathcal{P}^i(E) = \int_E f^i(x)\, d\lambda(x), \quad \forall E \in \mathcal{B}(\mathbb{R}^{d_X}).$$

The function f^i *is the density of* $\mathcal{P}^i$ *(or the Radon-Nikodym derivative) with respect to the Lebesgue measure, is denoted by* $f^i = \frac{d\mathcal{P}^i}{d\lambda}$*, and is referred to as conditional density. The measure* $\mathcal{P}^i$ *is the distribution associated with the density* f^i*.*

For $n \geq 1$, we are given two sets of samples:

$$\mathcal{D}_n^1 \ = \ \{X_j\}_{j=1}^n \quad \text{(unlabeled sample)}, \qquad \mathcal{D}_n^2 \ = \ \{(X_j, c_j)\}_{j=1}^n \quad \text{(labeled sample)},$$

where each $X_j \in S \subseteq \mathbb{R}^{d_X}$ and $c_j \in \{0, 1\}$. Within each set, the samples are assumed i.i.d. The two sets, $\mathcal{D}_n^1$ and $\mathcal{D}_n^2$, are assumed to be independent. The joint distribution of (X, c), $\mathcal{P}$, defined on the the product σ-algebra is unknown. $\mathcal{P}$ is the true distribution from which $\mathcal{D}_n^1$ and $\mathcal{D}_n^2$ are sampled. Using the data set, $\mathcal{D}_n^1$, the set S is partitioned as follows. Let $d : S \times ri(S) \to [0, \infty)$ be a divergence measure, where $ri(S)$ represents the relative interior of S. Let $V_n := \{S_h\}_{h=1}^{K_n}$ be a partition of S with respect to d and $M_n := \{\mu_h\}_{h=1}^{K_n}$ a set of codevectors, such that $\mu_h \in ri(S_h)$, for all $h = 1, \ldots, K_n$. A quantizer, $Q_n : S \to M_n$, is defined as the mapping $Q(X) = \sum_{h=1}^{K} \mu_h \mathbf{1}_{\{X \in S_h\}}$. The vector quantization problem at iteration n is formulated as an empirical loss minimization problem:

$$\min_{M_n, V_n} L(Q_n) := \frac{1}{n} \sum_{j=1}^{n} d(X_j, Q_n(X_j)),$$

Since vector quantization is a hard-clustering algorithm, the problem is equivalent to:

$$\min_{\{\mu_h\}_{h=1}^{K_n}} \frac{1}{n} \sum_{j=1}^{n} \sum_{h=1}^{K_n} d(X, \mu_h) \mathbf{1}_{\{X \in S_h\}}$$

where S_h is defined as

$$S_{n,h} = \{x \in S : h = \underset{k=1,\ldots,K_n}{\arg\min}\ d(x, \mu_k)\},\ h = 1, \ldots, K_n.$$

The above problem is solved using the Lloyd's algorithm which is the k-means algorithm invoked in the context of vector quantization. Given (M_n, V_n), we define empirical class-conditional cell frequencies using $\mathcal{D}_n^2$ as follows. For each cell $S_{n,h}$ let,

$$N_{n,h} := \sum_{j=1}^{n} \mathbf{1}_{\{X_j \in S_{n,h}\}}, \qquad N_{n,h}^{(i)} := \sum_{j=1}^{n} \mathbf{1}_{\{X_j \in S_{n,h},\, c_j = i\}},\ i = 0, 1,\ \{X_j,\, c_j\} \in \mathcal{D}_n^2.$$

The empirical distributions are defined as follows:

$$\mathcal{P}_n(X \in S_{n,h},\, c = 1) = \frac{N_{n,h}^1}{n}, \mathcal{P}_n(X \in S_{n,h},\, c = 0) = \frac{N_{n,h} - N_{n,h}^1}{n}, h = 1, \ldots, K_n.$$
$$\mathcal{P}_n(c = 1 | X \in S_{n,h}) \ = \widehat{\eta}_{n,h} := \ N_{n,h}^1 / N_{n,h},\ \text{if } N_{n,h} > 0,\ 0,\ \text{otherwise.}$$

and the empirical class-priors are

$$\mathcal{P}_n(c=i) \;=\; \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{\{c_j=i\}},\; i\in\{0,1\},\; \{X_j,\,c_j\}\in\mathcal{D}_n^2.$$

Let $\mathcal{C} = \{C : S \to \{0,1\}, C \text{ measurable}\}$ be the set of all classifiers. The loss function, $l : S\times\{0,1\}\times\{0,1\}$, for the classification problem is defined as:

$$l(X,c,C(X)) = \begin{cases} \widehat{l}_{10}, & \text{if } c=1, C(X)=0 \\ \widehat{l}_{01}, & \text{if } c=0, C(X)=1 \\ 0, & \text{otherwise} \end{cases}$$

The classification problem with respect to the empirical distribution, $\mathcal{P}_n$, is

$$l_n^\star(C_n) = \min_{C_n\in\mathcal{C}} \mathbb{E}_{\mathcal{P}_n}\Big[l(X,c,C_n(X))\Big].$$

The above problem is equivalent to:

$$\min_{C_n\in\mathcal{C}} \mathbb{E}_{\mathcal{P}_n}\Big[\mathbb{E}_{\mathcal{P}_n}\Big[l(X,c,C_n)\Big|\{X\in S_{n,h}\}\Big]\Big].$$

By monotonicity of expectation, it suffices to solve the following problem for each cell $S_{n,h}$,

$$c_{n,h}^* \;=\; \arg\min_{b\in\{0,1\}}\Big\{\widehat{l}_{10}\,\mathcal{P}_n(c=1|X\in S_{n,h})\,\mathbf{1}_{\{b=0\}} + \widehat{l}_{01}\,\mathcal{P}_n(c=0|X\in S_{n,h})\,\mathbf{1}_{\{b=1\}}\Big\},$$

From the above it follows that,

$$c_{n,h}^\star \;=\; \mathbf{1}_{\{\widehat{\eta}_{n,h}\geq\tau\}}, \quad \text{with } \tau := \frac{\widehat{l}_{10}}{\widehat{l}_{10}+\widehat{l}_{01}}.$$

or for 0–1 loss simply $c_{n,h}^* = \mathbf{1}_{\{\widehat{\eta}_{n,h}\geq 1/2\}}$. Define the classifier $C_n^\star : S\to\{0,1\}$ by

$$C_n^\star(x) \;:=\; C_{M_n,V_n}^{c_{n,h}^\star}(x) \;=\; \sum_{h=1}^{K_n} c_{n,h}^*\,\mathbf{1}_{\{x\in S_{n,h}\}}.$$

Thus the procedure yields a sequence of classifiers $\{C_n^\star\}_{n\geq 1}$. Let $l(C)$ denote the true misclassification (Bayes) loss, $l^\star$ the optimal loss, and $C^\star$ the optimal classifier, i.e.,

$$l(C) \;=\; \mathbb{E}_{\mathcal{P}}\Big[l(X,c,C(X))\Big],\; l^\star = \min_{C\in\mathcal{C}} l(C),\; C^\star = \arg\min_{C\in\mathcal{C}} l(C),$$

Problem 1 *Under the following regularity conditions:*

- *A1. The conditional densities f^0, f^1 are continuous or satisfy Lebesgue-differentiation condition.*
- *A2. The partition sequence (V_n) satisfies $\max_h \operatorname{diam}(S_{n,h}) \to 0$ and $\min_h(n\,|S_{n,h}|)\to\infty$ (cells shrink and receive enough labeled points).*
- *A3. The codevector design via $\mathcal{D}_n^1$ yields "regular" partitions (no pathological cell shapes, volumes comparable to $1/\kappa_n$), and $\kappa_n\to\infty$ such that $\kappa_n = o(n/\log n)$.*

Prove that,

$$C_n^\star \to C^\star \;\; \text{a.e on } S \text{ and } l_n^\star(C_n^\star) \;\to\; l^\star,$$

that is, the proposed sequence of classifiers converges to the Bayes decision surface $C^\star$.

3. Theoretical Results

The proof of the desired result is obtained in three steps. In Lemma 2, the convergence of the empirical distributions to the true distribution is proven using all the three conditions mentioned in Problem 1. In Lemma 3, the measure theoretic derivation of the Bayes classifier is obtained. Invoking Lemmas 2 and 3, in Theorem 4, we solve Problem 1.

Lemma 2 *Let* $\widehat{n}^i = \sum_{j=1}^n \mathbf{1}_{\{c_j=i\}}$. *For any* $x \in S$, *let* $S_{n,x}$ *denote the Voronoi cell containing* x *and* $N_{n,x}^i := \sum_{j=1}^n \mathbf{1}_{\{X_j \in S_{n,x},\, c_j=i\}}$. *Define the estimator of* $f^i(\cdot)$ *as:*

$$\widehat{f}_n^i(x) = \frac{N_{n,x}^i}{\widehat{n}^i \lambda(S_{n,x})}$$

Under regularity conditions A1, A2, A3, $\widehat{f}_n^i \to f^i$, *a.s.*

Proof From Definition 1, $\mathcal{P}^i(S_{n,x}) = \int_{S_{n,x}} f^i(x)\, d\lambda(x)$ and its Lebesgue measure is $\lambda(S_{n,x})$. From the Strong law of large numbers, $\lim_{n\to\infty} \mathcal{P}_n(c=i) = \mathcal{P}(c=i)$. Thus $\widehat{n}^i \to \infty$ when $n \to \infty$ except for the pathological case when $\mathcal{P}(c=i) = 0$. Utilizing the same,

$$\widehat{f}_n^i(x) = \frac{N_{n,x}^i}{\widehat{n}^i\, \lambda(S_{n,x})} = \frac{\mathcal{P}^i(S_{n,x})}{\lambda(S_{n,x})} + \frac{\dfrac{N_{n,x}^i}{\widehat{n}^i} - \mathcal{P}^i(S_{n,x})}{\lambda(S_{n,x})}. \tag{1}$$

We show that the terms on the right-hand side converge a.s. to $f^i(x)$ and 0 respectively. From condition A1, f^i is continuous at x. For every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\|y - x\| < \delta \Rightarrow\ |f(y) - f(x)| < \varepsilon.$$

From A2, since $\max_h \operatorname{diam}(S_{n,h}) \to 0$, for sufficiently large n we have $\operatorname{diam}(S_{n,x}) < \delta$, and hence

$$\left| \frac{1}{\lambda(S_{n,x})} \int_{S_{n,x}} f^i(y)\, d\lambda(y) - f^i(x) \frac{1}{\lambda(S_{n,x})} \int_{S_{n,x}} d\lambda(y) \right| \le \frac{1}{\lambda(S_{n,x})} \int_{S_{n,x}} |f(y) - f(x)|\, d\lambda(y) < \varepsilon.$$

Thus, $\dfrac{\mathcal{P}^i(S_{n,x})}{\lambda(S_{n,x})} \to f^i(x)$ as $n \to \infty$. When f^i satisfies Lebesgue-differentiation condition the above argument continues to hold. Conditional on $\mathcal{D}_n^1$, the test sample is i.i.d. and independent of the partition. Therefore

$$N_{n,x}^i \,|\, \mathcal{D}_n^1 \sim \operatorname{Bin}(\widehat{n}^i, \mathcal{P}^i(S_{n,x})).$$

By Bernstein's inequality, for all $\epsilon > 0$,

$$\mathbb{P}\left(|N_{n,x}^i - \widehat{n}^i \mathcal{P}^i(S_{n,x})| \ge \epsilon \,\middle|\, \mathcal{D}_n^1 \right) \le 2 \exp\left(- \frac{\epsilon^2/2}{\widehat{n}^i \mathcal{P}^i(S_{n,x})(1 - \mathcal{P}^i(S_{n,x})) + \epsilon/3} \right).$$

Setting $\epsilon = \widehat{n}^i \lambda(S_{n,x}) \varepsilon$ yields

$$\mathbb{P}\left(\left| \frac{N_{n,x}^i}{\widehat{n}^i} - \mathcal{P}^i(S_{n,x}) \right| \ge \lambda(S_{n,x})\varepsilon \,\middle|\, \mathcal{D}_n^1 \right) \le 2 \exp\left(-\alpha \widehat{n}^i \lambda(S_{n,x}) \varepsilon^2 \right),$$

for some constant $\alpha > 0$ independent of $\mathcal{D}_n^1$ and n (since $\mathcal{P}^i(S_{n,x}) \leq 1$). Taking expectations removes the conditioning:

$$\mathbb{P}\left(\left|\frac{N_{n,x}^i}{\widehat{n}^i} - \mathcal{P}^i(S_{n,x})\right| \geq \lambda(S_{n,x})\varepsilon\right) \leq 2\exp\left(-\alpha\widehat{n}^i\lambda(S_{n,x})\varepsilon^2\right).$$

By assumption A3, $\dfrac{\widehat{n}^i\lambda(S_{n,x})}{\log n} \to \infty$; hence

$$\sum_{n=1}^{\infty}\mathbb{P}\left(\left|\frac{N_{n,x}^i}{\widehat{n}^i} - \mathcal{P}^i(S_{n,x})\right| \geq \lambda(S_{n,x})\varepsilon\right) < \infty.$$

By the Borel-Cantelli lemma, with probability one there exists $N(\omega) < \infty$ such that, for all $n \geq N(\omega)$ the inequality below holds. Dividing by $\lambda(S_{n,x})$ and since $\varepsilon > 0$ is arbitrary, it follows that,

$$\left|\frac{N_{n,x}^i}{\widehat{n}^i} - \mathcal{P}^i(S_{n,x})\right| < \lambda(S_{n,x})\varepsilon \implies \left|\frac{\frac{N_{n,x}^i}{\widehat{n}^i} - \mathcal{P}^i(S_{n,x})}{\lambda(S_{n,x})}\right| \xrightarrow{\text{a.s.}} 0.$$

From (1),

$$\left|\widehat{f}_n^i(x) - f^i(x)\right| \leq \left|\frac{\mathcal{P}^i(S_{n,x})}{\lambda(S_{n,x})} - f^i(x)\right| + \left|\frac{\frac{N_{n,x}^i}{\widehat{n}^i} - \mathcal{P}^i(S_{n,x})}{\lambda(S_{n,x})}\right|.$$

Given any $\epsilon > 0$, pick $\varepsilon = \epsilon/2$ and use the almost-sure bounds from to find $N_1, N_2(\omega)$ such that for all $n \geq \max\{N_1, N_2(\omega)\}$,

$$\left|\widehat{f}_n^i(x) - f^i(x)\right| \leq \epsilon.$$

Hence $\widehat{f}_n^i(x) \to f^i(x)$ almost surely as $n \to \infty$. ■

Lemma 3 *Let $\lambda \times \mu$ denote the product of the Lebesgue measure on $\mathcal{B}(\mathbb{R}^{d_X})$ and counting measure on $2^{\{0,1\}}$. Then,*

$$f_{X,c}(x,i) := \frac{d\mathcal{P}}{d(\lambda\times\mu)}(x,i) = f^1(x)\mathcal{P}(c=1)\mathbf{1}_{i=1} + f^0(x)\mathcal{P}(c=0)\mathbf{1}_{i=0}.$$

$$f_X(x) = f^1(x)\mathcal{P}(c=1) + f^0(x)\mathcal{P}(c=0).\ \mathcal{P}_X(E) := \mathcal{P}(X\in E) = \int_E f_X(x)d\lambda(x).$$

$$\nu_{\bar{E}}(E) := P(X\in E,\, c\in\bar{E}).\ \mathcal{P}(c=i \mid X=x) := \frac{d\nu_{\bar{E}}}{d\mathcal{P}_X} = \frac{f_{X,c}(x,i)}{f_X(x)} \textit{ for } f_X(x) > 0.$$

Proof For any $E \in \mathcal{B}(\mathbb{R}^{d_X})$, from Bayes rule it follows that,

$$\mathcal{P}(c=i|X\in E) = \frac{\mathcal{P}(X\in E\ \cap c=i)}{\mathcal{P}(X\in E)} = \frac{\int_E f^i(x)\,d\lambda(x)\mathcal{P}(c=i)}{\sum_{i=0,1}\int_E f^i(x)\,d\lambda(x)\mathcal{P}(c=i)}$$

However, the above does not lead to the conditional probability of the class being i given $X = x$. From Definition 1, it follows that the joint law of (X, c) is absolutely continuous with respect to the product of Lebesgue and counting measure and admits athe density $f_{X,c}(x, i)$. For any $\bar{E} \subset \{0, 1\}$ and any Borel set $E \subset \mathcal{B}(\mathbb{R}^{d_X})$,

$$\nu_{\bar{E}}(E) = \mathcal{P}(X \in E,\, c \in \bar{E}) = \sum_{i \in \bar{E}} \int_E f_{X,c}(x, i)\, d\lambda(x).$$

Let the measure $\nu_{\bar{E}}(\cdot)$ be defined on $\mathcal{B}(\mathbb{R}^{d_X})$ as above. If $\mathcal{P}_X(E) = 0$, then $\int_E f_X(x) d\lambda(x) = 0$, hence $f_{X,c}(x, c) = 0$ a.e. on E for each c, and thus $\nu_{\bar{E}}(E) = 0$. Hence $\nu_{\bar{E}} \ll \mathcal{P}_X$ and by the Radon-Nikodym theorem there exists a $\mathcal{P}_X$-a.e. unique measurable function $g_{\bar{E}}(x)$ such that

$$\nu_{\bar{E}}(E) = \int_E g_{\bar{E}}(x) d\mathcal{P}_X(x) = \int_E g_{\bar{E}}(x) f_X(x)\, d\lambda(x).$$

Comparing with the explicit formula for $\nu_{\bar{E}}(E)$ we obtain for Lebesgue-a.e. x

$$g_{\bar{E}}(x) f_X(x) = \sum_{i \in \bar{E}} f_{X,c}(x, i).$$

Thus, whenever $f_X(x) > 0$,

$$g_{\bar{E}}(x) = \frac{\sum_{x \in \bar{E}} f_{X,c}(x, i)}{f_X(x)}.$$

For singletons $\bar{E} = \{i\}$ this yields

$$\mathcal{P}(c = i \mid X = x) = \frac{f_{X,c}(x, i)}{f_X(x)}, \quad \text{for } f_X(x) > 0,$$

which defines a regular conditional probability kernel. Uniqueness holds $\mathcal{P}_X$-almost everywhere. ■

Theorem 4 *Let* $\widehat{\eta}_n(x) = \dfrac{N^1_{n,x}}{N^1_{n,x} + N^0_{n,x}}$, *where* $N^1_{n,x}$ *is defined in Lemma 2. Under conditions A1, A2, and A3,*

$$\widehat{\eta}_n(x) \to \mathcal{P}(c = 1 \mid X = x) \text{ a.e on } S,\ C^\star_n \to C^\star \text{ a.e on } S,\ \text{and } l^\star_n(C^\star_n) \ \to\ l^\star.$$

Proof We express $\widehat{\eta}_n(x)$ as below:

$$\widehat{\eta}_n(x) = \frac{N^1_{n,x}}{N^1_{n,x} + N^0_{n,x}} = \frac{\underbrace{\dfrac{N^1_{n,x}}{\widehat{n}^1 \lambda(S_{n,x})}}_{(1)} \underbrace{\dfrac{\widehat{n}^1}{n}}_{(2)}}{\frac{N^1_{n,x}}{\widehat{n}^1 \lambda(S_{n,x})} \frac{\widehat{n}^1}{n} + \underbrace{\dfrac{N^0_{n,x}}{\widehat{n}^0 \lambda(S_{n,x})}}_{(3)} \underbrace{\dfrac{\widehat{n}^0}{n}}_{(4)}}$$

From Lemma 2 and strong law of large numbers, it follows that (1) converges to f^1, (2) converges to $\mathcal{P}(c=1)$, (3) converges to f^0, and (4) converges to $\mathcal{P}(c=0)$. From Lemma 3, it follows that $\widehat{\eta}_n(x) \to \mathcal{P}(c=1 \mid X=x)$ a.e on S. The loss for the classification problem with respect to the true distribution can be expressed as,

$$\int_{S\times\{0,1\}} l(x,i,C(x))d\mathcal{P}(x,i) = \int_{S\times\{0,1\}} l(x,i,C(x))f_{X,c}(x,i)d(\lambda\times\mu)$$
$$= \int_S \Big[\widehat{l}_{10}\mathbf{1}_{c=1,C(x)=0}\mathcal{P}(c=1 \mid X=x) + \widehat{l}_{01}\mathbf{1}_{c=0,C(x)=1}\mathcal{P}(c=0 \mid X=x)\Big] f_X(x)d\lambda(x)$$

Thus, $C(x)=1$ if

$$\widehat{l}_{10}\mathcal{P}(c=1 \mid X=x) \geq \widehat{l}_{01}(1-\mathcal{P}(c=1 \mid X=x)), \text{ i.e. } C^\star(x) = \mathbf{1}_{\mathcal{P}(c=1|X=x)\geq\tau}.$$

Since $\widehat{\eta}_n(x) \to \mathcal{P}(c=1 \mid X=x)$ a.e on S, and $C_n^\star(x) = \sum_{h=1}^{K_n} c_{n,h}^* \mathbf{1}_{\{x\in S_{n,h}\}} = \mathbf{1}_{\widehat{\eta}_n(x)\geq\tau}$, it follows that $C_n^\star \to C^\star$ a.e on S. The minimal cost at every n can be expressed as,

$$l_n^\star(C_n) = \sum_{h=1}^{K_n} \min(\widehat{l}_{10}\mathcal{P}_n(c=1|X\in S_{n,h}), \widehat{l}_{01}\mathcal{P}_n(c=1|X\in S_{n,h}))\mathcal{P}_n(X\in S_{n,h})$$
$$= \int_S \min(\widehat{l}_{10}\widehat{\eta}_n(x), \widehat{l}_{01}(1-\widehat{\eta}_n(x)) \sum_{i=0,1} \widehat{f}_n^i(x)\mathcal{P}_n(c=i)d\lambda(x).$$

From condition A1 and compactness of S, it follows that f^1 and f^0 are uniformly bounded. From Lemma 2, it follows that histogram estimators, $\widehat{f}_n^i(\cdot)$, are uniformly bounded. By the bounded convergence theorem it follows that,

$$\lim_{n\to\infty} l_n^\star(C_n) = \int_S \lim_{n\to\infty} \min(\widehat{l}_{10}\widehat{\eta}_n(x), \widehat{l}_{01}(1-\widehat{\eta}_n(x)) \sum_{i=0,1} \widehat{f}_n^i(x)\mathcal{P}_n(c=i)d\lambda(x)$$
$$= \int_S \min(\widehat{l}_{10}\mathcal{P}(c=1 \mid X=x), \widehat{l}_{01}\mathcal{P}(c=0 \mid X=x))f_X(x)d\lambda(x) = l^\star,$$

thus solving problem 1. ■

4. Simulation Examples

We demonstrate the convergence of empirical estimators to the Bayes optimal classifier using three different VQ-based partitioning algorithms: histogram approximation, Lloyd's (k-means) algorithm, and the Online Deterministic Annealing (ODA) method Mavridis and Baras (2023c,a). A hybrid system identification application is also illustrated, highlighting the relevance of these results in inference and control applications.

4.1. Binary Classification

We consider a binary classification problem in $S=[0,1]\times[0,1]\subset\mathbb{R}^2$ with Gaussian distributions specified by the mean vectors $(\mu_1,\mu_2) = ([0.25,0.25]^{\mathrm{T}}, [0.75,0.75]^{\mathrm{T}})$, the covariance matrices

$(\Sigma_1, \Sigma_2) = (0.04 \cdot I_2, 0.06 \cdot I_2)$, and prior probabilities $(p_1, p_2) = (0.4, 0.6)$. The closed form expression for the Bayes decision surface can be derived by the condition:

$$p_1 \mathcal{N}(x; \mu_1, \Sigma_1) = p_2 \mathcal{N}(x; \mu_2, \Sigma_2),$$

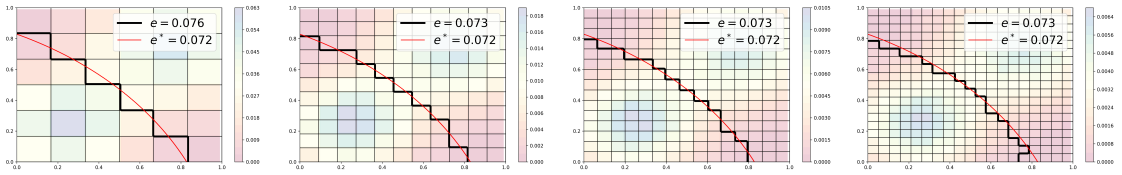
which yields a quadratic Bayes decision boundary of the form:

$$(x-\mu_2)^\top \Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1) + \log|\Sigma_2| - \log|\Sigma_1| + 2\log\frac{p_1}{p_2} = 0.$$

Substituting the parameters, the boundary equation becomes:

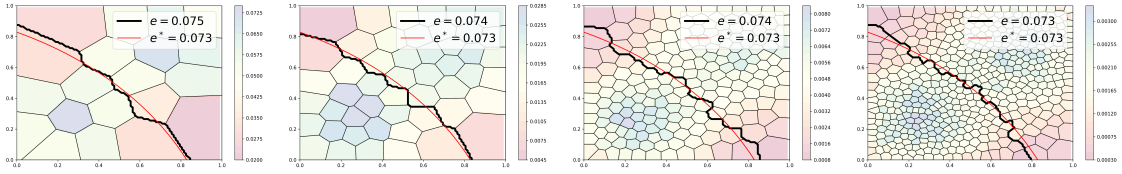
$$8(x_1^2 + x_2^2) + 12(x_1 + x_2) - 15 = 0.$$

Simulations were performed with $N = 10^5$ samples. Figures 1, 2 and 3 depict the progression of the approximation of the Bayes decision surface as the number of codevectors increases and the volume of the cells decreases. It is observed that the empirical cost converges to the Bayesian cost. However, although the empirical classification boundary is shown to converge to the true quadratic surface, this is not observed in practice. This confirms our intuition that assumptions A2 and A3 of Problem 1 can rarely be met in practice, since they require infinite number of samples even in regions with near-zero measure. In that case, the classifiers converge to the Bayes error without perfectly reconstructing the quadratic decision surface.



(*a*) $k = 36$ (*b*) $k = 121$ (*c*) $k = 225$ (*d*) $k = 400$

Figure 1: Approximation of the Bayes decision surface (red) using histogram approximation.



(*a*) $k = 20$ (*b*) $k = 50$ (*c*) $k = 200$ (*d*) $k = 500$

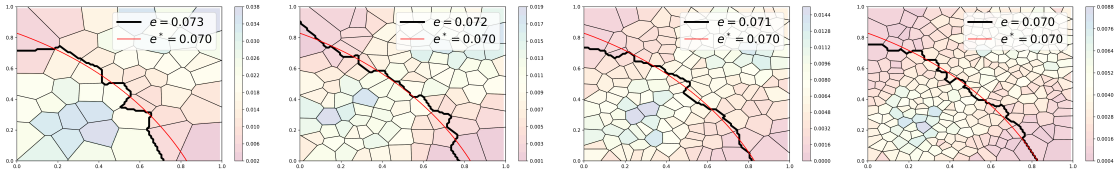
Figure 2: Approximation of the Bayes decision surface (red) using Lloyd's algorithm (k-means).

4.2. Hybrid System Identification

A benchmark PWARX system, adopted from (Mavridis and Johansson, 2025) is given by:

$$y_t = \begin{cases} \theta_2^{\mathrm{T}} \phi_t + e_t, & \text{if } \phi_t \in S_2 \\ \theta_1^{\mathrm{T}} \phi_t + e_t, & \text{otherwise} \end{cases}, \tag{2}$$

where $y_t \in \mathbb{R}^1$, $r_t \in P = [-4, 4]$, $\phi_t = [r_t \ 1]^{\mathrm{T}}$, $S_2 = \{\phi = [r \ 1]^{\mathrm{T}} : r \in (-1, 2)\}$, and $(\theta_1, \theta_2) = ([1, 2]^{\mathrm{T}}, [-1, 0]^{\mathrm{T}})$. The simplicity of this example allows for the graphical depiction of

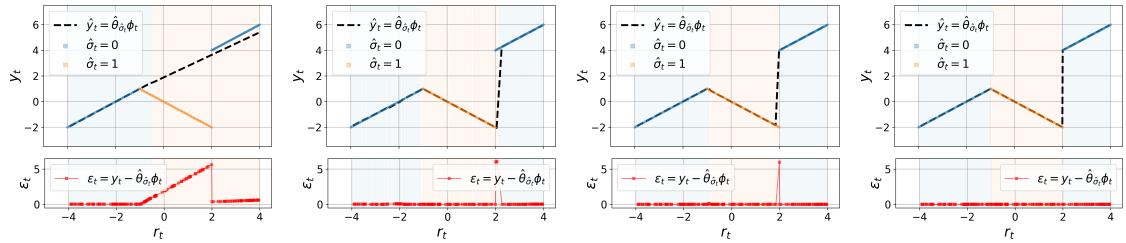


(a) $k = 51$ (b) $k = 110$ (c) $k = 165$ (d) $k = 302$

Figure 3: Approximation of the Bayes decision surface (red) using Online Deterministic Annealing.

both the mode-switching partition and the convergence of the model parameters. At the same time, it presents the same dynamics for non-convex regions of the input space. A total of $N = 1500$ observations under Gaussian noise ($e_t \sim \mathcal{N}(0, 0.2)$) are accessible. Following Corollary 4.1 in (Mavridis and Johansson, 2025), the modes of system (2) can be identified by a stochastic approximation VQ-based classification algorithm, while another stochastic approximation algorithm, running at a higher timescale, reconstructs the mode dynamics. Therefore, convergence of the mode switching classifier to the Bayes classifier is critical for the stable reconstruction of the hybrid system.

The progression of the mode classification and identification of system (2) are shown in Figure 4. In this case, the classification problem in isolation is one-dimensional and the Bayes decision boundary is the points $r = -1$ and $r = 2$, separating the region S_2 from the rest of the domain. Because of the simplicity of the boundary, the Bayes error is achieved with $k \geq 9$ codevectors. In a noise-free system the Bayes error is zero and is achieved with $k \geq 3$ codevectors.



(a) $k = 2$ (b) $k = 3$ (c) $k = 5$ (d) $k = 9$

Figure 4: Progression of hybrid system identification using Online Deterministic Annealing.

5. Conclusion and Future Work

In this work we established convergence guarantees for a broad class of quantization-based classifiers, including histogram methods, variants of the generalized Lloyd's algorithm, and online deterministic annealing methods. An illustrative application in hybrid system identification underscores the practical relevance and adaptability of these methods.

Simulation results confirm that conditions A2 and A3 of Problem 1 are computationally expensive in practice. Future directions include finding probabilistic bounds of the form $\mathbb{P}(|C_n^\star - C^\star| > \epsilon) < \delta$, under weaker assumptions leading to computational tractability. Simultaneous partitioning and empirical classification will also be investigated leading to construction of partitions concentrated around the true Bayes decision surface.

Acknowledgments

This work was supported in part by Swedish Research Council Distinguished Professor Grant 2017-01078, Knut and Alice Wallenberg Foundation Wallenberg Scholar Grant, the Swedish Strategic Research Foundation FUSS SUCCESS Grant, and the Swedish Foundation for Strategic Research (SSF) grant IPD23-0019.

References

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

John S Baras and Anthony LaVigna. Convergence of a neural network classifier. In *Advances in Neural Information Processing Systems*, pages 839–845, 1991.

Federico Bianchi, Alessandro Falsone, Luigi Piroddi, and Maria Prandini. An alternating optimization method for switched linear systems identification. *IFAC-PapersOnLine*, 53(2):1071–1076, 2020.

Léon Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Giancarlo Ferrari-Trecate, Marco Muselli, Diego Liberati, and Manfred Morari. A clustering technique for the identification of piecewise affine systems. *Automatica*, 39(2):205–217, 2003.

ME Gegundez, Javier Aroba, and José Manuel Bravo. Identification of piecewise affine systems by means of fuzzy clustering and competitive learning. *Engineering Applications of Artificial Intelligence*, 21(8):1321–1329, 2008.

Allen Gersho and Robert M Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.

Robert M Gray. Vector quantization. *Readings in speech recognition*, 1(2):75–100, 1990.

Christos Mavridis and John S. Baras. Annealing optimization for progressive learning with stochastic approximation. *IEEE Transactions on Automatic Control*, 68(5):2862–2874, 2023a.

Christos Mavridis and Karl Henrik Johansson. Real-time switched system identification with online deterministic annealing. *IEEE Transactions on Automatic Control*, 2025.

Christos N Mavridis and John S Baras. Convergence of stochastic vector quantization and learning vector quantization with bregman divergences. *IFAC-PapersOnLine*, 53(2), 2020.

Christos N Mavridis and John S Baras. Identification of piecewise affine systems with online deterministic annealing. In *IEEE Conference on Decision and Control*, pages 4885–4890, 2023b.

Christos N. Mavridis and John S. Baras. Online deterministic annealing for classification and clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10):7125–7134, 2023c. doi: 10.1109/TNNLS.2021.3138676.

Christos N Mavridis, Aris Kanellopoulos, John S Baras, and Karl Henrik Johansson. State-space piece-wise affine system identification with online deterministic annealing. In *European Control Conference*, pages 3110–3115, 2024.

Friedrich Melchert, Udo Seiffert, Michael Biehl, B Hammer, T Martinetz, and T Villmann. Functional approximation for the classification of smooth time series. In *GCPR Workshop on New Challenges in Neural Computation 2016*, pages 24–31, 2016.

Aneesh Raghavan and John S Baras. Binary hypothesis testing with learning of empirical distributions. *IFAC-PapersOnLine*, 54(9):671–676, 2021.

Kenneth Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, 1998.

Sascha Saralajew, Lars Holdijk, Maike Rees, and Thomas Villmann. Prototype-based neural network layers: Incorporating vector quantization. *arXiv preprint arXiv:1812.01214*, 2018.

Sascha Saralajew, Lars Holdijk, Maike Rees, and Thomas Villmann. Robustness of generalized learning vector quantization models against adversarial attacks. *arXiv preprint arXiv:1902.00577*, 2019.

Sohil Atul Shah and Vladlen Koltun. Deep continuous clustering. *arXiv preprint arXiv:1803.01449*, 2018.

E Arsuaga Uriarte and F Díaz Martín. Topology preservation in som. *International journal of applied mathematics and computer sciences*, 1(1):19–22, 2005.

Thomas Villmann, Michael Biehl, Andrea Villmann, and Sascha Saralajew. Fusion of deep learning architectures, multilayer feedforward networks and learning vector quantizers for deep classification learning. In *12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM)*, pages 1–8. IEEE, 2017a.

Thomas Villmann, Andrea Bohnsack, and Marika Kaden. Can learning vector quantization be an alternative to svm and deep learning?-recent trends and advanced variants of learning vector quantization for classification learning. *Journal of Artificial Intelligence and Soft Computing Research*, 7(1):65–81, 2017b.

Jiaorao Wang, Chunyue Song, Jun Zhao, and Zuhua Xu. A PWA model identification method for nonlinear systems using hierarchical clustering based on the gap metric. *Computers & Chemical Engineering*, 138:106838, 2020.

Jixuan Wang, Kuan-Chieh Wang, Marc Law, Frank Rudzicz, and Michael Brudno. Centroid-based deep metric learning for speaker recognition. *arXiv preprint arXiv:1902.02375*, 2019.

Miao Yu, Federico Bianchi, and Luigi Piroddi. A randomized method for the identification of switched NARX systems. *Nonlinear Analysis: Hybrid Systems*, 49:101364, 2023.