Post-Clustering Soft Vector Quantization with Inverse Power-Function Distribution, and Application on Discrete HMM-Based Machine Learning

Mohamed Attia, Abdul-Aziz Al-Mazyad, Mohamed El-Mahallawy, Mohamed Al-Badrashiny, Walid Nazih

***Abstract*-In this paper, we introduce a soft vector quantization scheme with inverse power-function distribution, and analytically derive an upper bound of the resulting quantization noise energy in comparison to that of typical (hard-deciding) vector quantization.**

We also discuss the positive impact of this kind of soft vector quantization on the performance of machine-learning systems that include one or more vector quantization modules. Moreover, we provide experimental evidence on the advantage of avoiding over-fitting and boosting the robustness of such systems in the presence of considerable parasitic variance; e.g. noise, in the runtime inputs. The experiments have been conducted with two versions of one of the best reported discrete HMM-based Arabic OCR systems; one version deploying hard vector quantization and the other deploying our herein presented soft vector quantization. Test samples of real-life scanned pages are used to challenge both versions; hence the recognition error margins are compared.

***Index Terms*-Machine Learning, Over-fitting, Quantization Noise, Soft Vector Quantization.**

I. INTRODUCTION

Given a codebook of centroids[2]; i.e. set of centers of classes/clusters $\underline{c}_i \in \Re^n; 1 \le i \le L$, vector quantization (VQ) is a fundamental signal processing operation that seeks to attribute a given point $\underline{q} \in \Re^n$ to one of those centroids: $\underline{c}_{i_0}$ according to some optimization criterion. [4] Typical VQ deploys the minimum-distance criterion that:

$$\underline{q} \xrightarrow{VQ} i_0 : i_0 = \underset{\forall j; 1 \le j \le L}{\arg\min} \left\{ d(\underline{q}, \underline{c}_j) \right\} \qquad (1)$$

…where $d(\underline{q}_1, \underline{q}_2)$ is any legitimate distance measure between $\underline{q}_1, \underline{q}_2 \in \Re^n$. The quantization noise energy due to this operation is given by:

$$e_{VQ}^2 = \min_{\forall j; 1 \le j \le L} \left(d(\underline{q}, \underline{c}_j) \right)^2 \qquad (2)$$

The total quantization noise energy over a population of points[3] in this space of size s versus that codebook of centroids [4, 7, 10] is hence given by:

$$E_{VQ}^2 = \sum_{i=1}^{s} \min_{\forall j; 1 \le j \le L} \left(d(\underline{q}_i, \underline{c}_j) \right)^2 \qquad (3)$$

VQ in the form of eq. (1) is a hard-deciding operation following the *winner-takes-all policy* which may not be quite fair especially with rogue points which are almost equidistant from more than one centroid in the codebook.[10, 13, 17] With machine-learning systems that include a hard-deciding VQ module, the quantized observations (or observation sequences) corresponding to some inputs during the training phase may be significantly different from those corresponding to the same inputs in the runtime that may have only experienced just a slight variance in the observation space! Regardless to the deployed machine-learning methodology, that difference will inevitably cause some deterioration in the performance of such systems.

In order to boost the robustness of the run-time performance in the presence of all kinds of variances; e.g.

Manuscript received June 30th, 2010. The work presented in this paper is a part of a 12-month project funded by the Deanship of Scientific Research of AlKharj University, AlKharj-Kingdom of Saudi Arabia (KSA), during the interval of April 2010 to Mar. 2011.

Mohamed Attia is an HLT consultant for The Engineering Company for the Development of Computer System; RDI www.RDI-eg.com 12613 Giza-Egypt, a consultant for Luxor Technologies Inc.; Oakville-L6L6V2 Ontario-Canada, and a visiting assistant professor in the College of Computing and IT of the Arab Academy for Science & Technology; AAST www.AAST.edu Heliopolis Campus – Cairo - Egypt (e-mail: m_Atteya@RDI-eg.com).

Abdul Aziz Al-Mazyad is the dean of the College of Computer Engineering and Sciences in AlKharj University - AlKharj - KSA, and is also a professor in the dept. of Computer Engineering in the College of Computer and Information Sciences - King Saud University ksa.edu.sa Riyadh - KSA (e-mail: Mazyad@ccis.edu.sa).

Mohamed El-Mahallawy is an assistant professor in the College of Engineering and Technology of the Arab Academy for Science & Technology; AAST www.AAST.edu Heliopolis Campus – Cairo - Egypt (e-mail: Mahallawy@AAST.edu).

Mohamed-Al-Badrashiny is a PhD student in the dept. of Electronics & Electrical Communications – Faculty of Engineering – Cairo University – 12613 Giza – Egypt (Phone +20-23 812 5979, Fax: +20-23 338 21 66; e-mail: Mohammed.Badrashiny@RDI-eg.com).

Walid Nazih is a lecturer in the College of Computer Engineering and Sciences in AlKharj University - AlKharj – KSA (e-mail: wNazeeh@ksu.edu.sa).

[2] All the material presented in this paper is independent of the algorithm used for inferring that codebook; e.g. k-means, LBG ... etc.

[3] Typically, any adaptive methodology for inferring the codebook works in the offline phase on a sample population that is assumed to have the same statistical properties of the phenomenon being modeled.

noise, in the inputs to these systems, soft VQ is proposed so that there is a non-zero chance of the belonging of any given point to each centroid in the codebook. Intuitively, the closer is the point to some centorid than the other ones in the codebook; the higher is the probability of the attribution of this point to that centroid.

Soft VQ in this sense will *shake up* the over-fitting of the training by introducing smoother and more expressive distributions of quantized observations in the statistical learning models, which will in turn be more capable to coping with run-time variances than those resulting from hard-deciding VQ.

Formally, soft VQ may in general be formulated as:

$$P(\underline{q} \xrightarrow{SoftVQ} i) = \frac{f\left(d(\underline{q},\underline{c}_i)\right)}{\sum_{j=1}^{L} f\left(d(\underline{q},\underline{c}_j)\right)} = \frac{f\left(d_i\right)}{\sum_{j=1}^{L} f\left(d_j\right)} \qquad (4)$$

The function $f(d_i)$ must obey the following conditions:

1. $f(d_i) \geq 0 \;\; \forall d_i \geq 0$
2. $f(d_i)$ is continuous $\forall d_i \geq 0$
3. $f(d_i)$ is a monotonically decreasing function $\forall d_i \geq 0$
4. $d_i = 0 \Rightarrow P(\underline{q} \xrightarrow{SoftVQ} i) = 1 \wedge P(\underline{q} \xrightarrow{SoftVQ} j \neq i) = 0$

It is crucial to note that the quantization noise energy due to the soft VQ of each given point $\underline{q}$ is given by:

$$e^2_{SoftVQ} = \sum_{j=1}^{L}\left(d_j^2 \cdot P(\underline{q} \xrightarrow{SoftVQ} i)\right) = \frac{\sum_{j=1}^{L}\left(d_j^2 \cdot f(d_j)\right)}{\sum_{j=1}^{L} f\left(d_j\right)} \qquad (5)$$

In eq. (5): each $d_j^2 \geq \left(d_{\min}^2 = d(\underline{q},\underline{c}_{i_0})^2\right) \forall j; 1 \leq j \leq L$ is weighted by probabilities ≥ 0, and together with eq. (2) and eq. (3), we conclude:

$$e^2_{SoftVQ} \geq e^2_{VQ} \Rightarrow r \equiv \frac{e^2_{SoftVQ}}{e^2_{VQ}} \geq 1 \Rightarrow \frac{E^2_{SoftVQ}}{E^2_{VQ}} \geq 1 \qquad (6)$$

This means that the price we pay for soft VQ is a higher harmful quantization noise energy, which in turn indicates the necessity to compromise that price with the gains of avoiding over-fitting for a more robust performance to inputs' variance.

The inverse power-function distribution for soft VQ is defined in the next section of this paper, and then section III is devoted to a detailed analytic investigation of the quantization noise energy resulting from this distribution relative to that of the typical hard VQ.

In section IV, the experimental setup with a discrete HMM-based Arabic OCR is described, and the experimental results are analyzed to see how good they match our claims on the benefits of our proposed soft VQ scheme for machine learning systems with one or more VQ modules.

II. INVERSE POWER-FUNCTION BASED SOFT VQ

In addition to satisfying the four conditions mentioned above, it is much desirable for the design of the function *f(x)* to have the following features:

1. Simplicity.
2. Having tuning-parameters that control the probability attenuation speed with increasing distance.
3. Minimum r_{avg} over all the possible instances of $\underline{q}$'s.

While its realization of the third desirable feature is subject to a detailed analysis over section III, the inverse power-function realizes all the necessary conditions and the first two desirable features above. It is defined as:

$$f\left(d_j\right) = d_j^{-m}; m > 0 \qquad (7)$$

III. NOISE ENERGY OF OUR SOFT VQ vs. HARD VQ

Substituting the formula of eq. (7) in eq. (5) gives:

$$e^2_{SoftVQ} = \frac{\sum_{j=1}^{L}\left(d_j^{2-m}\right)}{\sum_{j=1}^{L}(d_j^{-m})} \qquad (8)$$

…then, substituting (2) and (8) in (6), we get:

$$r = \frac{e^2_{SoftVQ}}{e^2_{VQ}} = \frac{\sum_{j=1}^{L}\left(d_j^{2-m}\right) \Big/ \sum_{j=1}^{L}\left(d_j^{-m}\right)}{d^2_{\min}} = \frac{\sum_{j=1}^{L}\left({d_{\min}}/{d_j}\right)^{m-2}}{\sum_{j=1}^{L}\left({d_{\min}}/{d_j}\right)^{m}} \qquad (9)$$

Let us define:

$$\alpha_j \equiv \frac{d_{\min}}{d_j} \leq 1; \alpha_j \in [0,1] \qquad (10)$$

…then eq. (9) can be re-written more conveniently as:

$$r = \frac{1 + \sum_{\substack{j=1 \\ j \neq i_0}}^{L} \alpha_j^{m-2}}{1 + \sum_{\substack{j=1 \\ j \neq i_0}}^{L} \alpha_j^{m}}; 1 \leq j \leq L \qquad (11)$$

For $0 < m < 2$; it is obvious that the numerator grows indefinitely faster than the denominator for arbitrarily infinitesimal values of some $\alpha_j; j \in \Omega \subset \{1,2,..., L\}$ so that:

$$\lim_{\alpha_k \to 0 \forall k \in \Omega} r\Big|_{0<m<2} = \frac{1 + \sum_{\forall k \in \Omega}\left(\lim_{\alpha_k \to \delta \to 0} \alpha_k^{m-2}\Big|_{0<m<2}\right)}{1 + \sum_{\forall k \in \Omega}\left(\lim_{\alpha_k \to \delta \to 0} \alpha_k^{m}\Big|_{0<m<2}\right)} \qquad (12)$$

$$= \frac{1 + \omega \cdot \lim_{\delta \to 0}(\frac{1}{\delta})^{2-m}}{1 + \omega \cdot 0} = \frac{1+\infty}{1+0} = \infty; \omega = SizeOf\,(\Omega)$$

This result necessitates the avoidance of the interval of $0 < m < 2$ as the unlimited growth of the soft quantization noise will be enormously devastating to whatever machine learning process!

For $m = 2$; eq. (11) reduces into:

$$r\big|_{m=2} = \frac{L}{1+\sum_{\substack{j=1\\ j\neq i_0}}^{L}\alpha_j^2};1\le j\le L \qquad (13)$$

…and one can easily guess that:

$$\max\left(r\big|_{m=2}\right) = \lim_{\alpha_j\to 0\forall j\neq i;1\le j\le L}\left(r\big|_{m=2}\right) = L \qquad (14)$$

As the size of the codebook used with non-trivial problems is typically a large number, the worst case of eq. (14) still indicates a huge soft quantization noise that can ruin machine learning esp. as that worst case happens at the dominant situation of point so close to one centroid only and far from the others!

For $m > 2$; considering eq. (10), the following three special cases of eq. (11) can easily be noticed:

$$\lim_{m\to\infty} r = \frac{1+\sum_{\substack{j=1\\ j\neq i_0}}^{L}(\lim_{m\to\infty}\alpha_j^{m-2})}{1+\sum_{\substack{j=1\\ j\neq i_0}}^{L}(\lim_{m\to\infty}\alpha_j^{m})} = \frac{1+0}{1+0} = 1 \qquad (15)$$

…which shows that the quantization noise energy of our proposed soft VQ with the power *m* growing larger is approaching the one of the hard-deciding VQ, however its distributions are turning less smooth and more similar to those of the hard-deciding VQ.

$$\lim_{\forall\alpha_j\to 0;j\neq i_0} r = \frac{1+\sum_{\substack{j=1\\ j\neq i_0}}^{L}(\lim_{\alpha_j\to 0}\alpha_j^{m-2})}{1+\sum_{\substack{j=1\\ j\neq i_0}}^{L}(\lim_{\alpha_j\to 0}\alpha_j^{m})} = \frac{1+(L-1)\cdot 0}{1+(L-1)\cdot 0} = 1 \qquad (16)$$

…which occurs only when $\underline{q} = \underline{c}_{i_0}$.

$$\lim_{\forall\alpha_j\to 1} r = \frac{1+\sum_{\substack{j=1\\ j\neq i_0}}^{L}(\lim_{\alpha_j\to 1}\alpha_j^{m-2})}{1+\sum_{\substack{j=1\\ j\neq i_0}}^{L}(\lim_{\alpha_j\to 1}\alpha_j^{m})} = \frac{1+(L-1)\cdot 1}{1+(L-1)\cdot 1} = 1 \qquad (17)$$

…which occurs when $d(\underline{q},\underline{c}_i)$ is exactly the same $\forall i;1\le i\le L$.

Only for these special cases $r = 1$ otherwise $r > 1$. It is crucial to calculate the maximum value of *r*; i.e. the worst case, which is an upper bound of the ratio between the total quantization noise energy of the proposed soft VQ to that of the conventional hard-deciding VQ.

To obtain $r_{\max}$, the $(L-1)$-dimensional sub-space within $\alpha_{j\neq i_0}\in[0,1]\ \forall j;1\le j\le L$ has to be searched for those $\hat{\alpha}_{j\neq i_0}$ where that maximum is realized. This can be done analytically by solving the following set of $(L-1)$ equations:

$$\partial r\big/\partial\alpha_k\Big|_{\forall k\neq i_0} = 0;1\le k\le L \qquad (18)$$

For the sake of convenience, let us re-write eq. (11) as:

$$r = \frac{A_k+\alpha_k^{m-2}}{B_k+\alpha_k^{m}};A_k\equiv 1+\sum_{\forall j\neq i_0,j\neq k}\alpha_j^{m-2},B_k\equiv 1+\sum_{\forall j\neq i_0,j\neq k}\alpha_j^{m} \qquad (19)$$

Then:

$$\partial r\big/\partial\alpha_k\Big|_{\forall k\neq i_0} = 0 \Rightarrow \frac{(m-2)\cdot\hat{\alpha}_k^{m-3}}{A_k+\hat{\alpha}_k^{m-2}}\Bigg|_{\forall k\neq i_0} = \frac{m\cdot\hat{\alpha}_k^{m-1}}{B_k+\hat{\alpha}_k^{m}}\Bigg|_{\forall k\neq i_0} \qquad (20)$$

…that reduces into:

$$\frac{A_k+\hat{\alpha}_k^{m-2}}{B_k+\hat{\alpha}_k^{m}}\Bigg|_{\forall k\neq i_0} = r_{\max} = (\frac{m-2}{m})\cdot\hat{\alpha}_k^{-2}\Bigg|_{\forall k\neq i_0} \qquad (21)$$

In order for eq. (21) to hold true, all $\hat{\alpha}_{k\neq i_0}$ must be equal so that:

$$\hat{\alpha}_k\Big|_{\forall k\neq i_0} = \hat{\alpha} \qquad (22)$$

…that reduces eq. (19) into:

$$A = 1+(L-2)\cdot\hat{\alpha}^{m-2}, B = 1+(L-2)\cdot\hat{\alpha}^{m} \Rightarrow$$

$$r_{\max} = \frac{A+\hat{\alpha}^{m-2}}{B+\hat{\alpha}^{m}} = \frac{1+(L-1)\cdot\hat{\alpha}^{m-2}}{1+(L-1)\cdot\hat{\alpha}^{m}} = (\frac{m-2}{m})\cdot\hat{\alpha}^{-2}$$

$$\underline{\qquad\quad} \ (23)$$

Re-arranging the terms of (23), we get the polynomial equation:

$$\hat{\alpha}^{m} + \left(\frac{m}{2}\cdot\frac{1}{L-1}\right)\cdot\hat{\alpha}^{2} - \frac{m-2}{2}\cdot\frac{1}{L-1} = 0;$$
$$m > 2, L\ge 2, \hat{\alpha}\in[0,1] \qquad (24)$$

For any $m > 2$ that is an even number, eq. (24) can be shown to have one and only one real solution in the interval $\hat{\alpha}\in[0,1]$ through the following three-step proof:

1. Put $\hat{\beta} = \hat{\alpha}^2, c = \dfrac{1}{L-1}$, and re-write eq. (24) as:

$$g(\hat{\beta}) = \hat{\beta}^{m/2} + \frac{m}{2} \cdot c \cdot \hat{\beta} - \frac{m-2}{2} \cdot c = 0$$

2.

$$g(\hat{\beta} = 0) = -\frac{m-2}{2} \cdot \frac{1}{L-1} < 0$$

$$g(\hat{\beta} = 1) = \frac{2 \cdot L - 2 + m - m + 2}{2 \cdot (L-1)} = \frac{L}{L-1} > 0$$

$\therefore g(\hat{\beta})$ has roots $\in [0,1]$

3. $\because \dfrac{dg(\hat{\beta})}{d(\hat{\beta})} = \dfrac{m}{2} \cdot \hat{\beta}^{m/2-1} + \dfrac{m}{2 \cdot (L-1)} > 0$

$\therefore g(\hat{\beta})$ is a monotonically increasing function.

4. From steps 2 & 3, $g(\hat{\beta})$ has only one root $\in [0,1]$.

A closed-form solution of eq. (24) is extractable only for $(m/2) \in \{2,3,4,5\}$. [9]

When m=4, for example, eq. (24) turns into essentially a quadratic equation of the form:

$$\hat{\beta}^2 + 2 \cdot c \cdot \hat{\beta} - c = 0$$

$$\therefore \hat{\beta} = \hat{\alpha}^2 = \frac{-2 \cdot c \pm \sqrt{4 \cdot c^2 + 4 \cdot c}}{2} = \sqrt{c^2 + c} - c$$

And the solution in this case is:

$$\left.\hat{\alpha}^2\right|_{m=4} = \frac{\sqrt{L} - 1}{L-1}, \left.r_{\max}\right|_{m=4} = \frac{1}{2} \cdot \frac{L-1}{\sqrt{L} - 1}$$

$$\left.\lim_{L\to\infty} \hat{\alpha}^2\right|_{m=4} = \frac{1}{\sqrt{L}}, \lim_{L\to\infty} \left.r_{\max}\right|_{m=4} = \frac{1}{2} \cdot \sqrt{L} \qquad (25)$$

As another example, when m=6, eq. (24) turns into:

$$\hat{\alpha}^6 + 3 \cdot c \cdot \hat{\alpha}^2 - 2 \cdot c = 0$$

…that is a 3rd order equation of the form:

$$\hat{\beta}^3 + \eta_1 \cdot \hat{\beta} + \eta_0 = 0$$

…whose closed-form solution is [9]:

$$\hat{\beta} = -\frac{1}{3} \cdot \sqrt[3]{\left(\frac{1}{2}\right) \cdot \left(27 \cdot \eta_0 - \sqrt{27 \cdot \eta_0^2 + 4 \times 27 \cdot \eta_1^3}\right)}$$
$$-\frac{1}{3} \cdot \sqrt[3]{\left(\frac{1}{2}\right) \cdot \left(27 \cdot \eta_0 + \sqrt{27 \cdot \eta_0^2 + 4 \times 27 \cdot \eta_1^3}\right)}$$

$\hat{\alpha}^2$ and $r_{\max}$ are hence given by:

$$\hat{\alpha}^2 = \frac{\sqrt[3]{\left(1 + \sqrt{1 + \frac{1}{L-1}}\right)} - \sqrt[3]{\left(1 - \sqrt{1 + \frac{1}{L-1}}\right)}}{\sqrt[3]{L-1}}$$

$$r_{\max} = \frac{(2/3) \cdot \sqrt[3]{L-1}}{\sqrt[3]{\left(1 + \sqrt{1 + \frac{1}{L-1}}\right)} - \sqrt[3]{\left(1 - \sqrt{1 + \frac{1}{L-1}}\right)}}$$

$$\left.\lim_{L\to\infty} \hat{\alpha}^2\right|_{m=6} = \sqrt[3]{\frac{2}{L}}, \lim_{L\to\infty} \left.r_{\max}\right|_{m=6} = \frac{2}{3} \cdot \sqrt[3]{\frac{L}{2}}$$

______ (26)

For $(m/2) > 5$: one can only derive an expression for $\hat{\alpha}^2$ and $r_{\max}$ with any even degree m at $L \to \infty$; i.e. with a large code book. From eq. (25) and eq. (26) one can guess the generalization that:

$$\lim_{L\to\infty} \hat{\alpha} = \left(\frac{2 \cdot L}{m-2}\right)^{-1/m}, \lim_{L\to\infty} r_{\max} = \frac{m-2}{m} \cdot \left(\frac{2 \cdot L}{m-2}\right)^{\frac{2}{m}} \qquad (27)$$

Substituting that guess in the terms of eq. (24) gives:

$$\lim_{L\to\infty} \frac{T_3}{T_1} = \lim_{L\to\infty} \left(-\left(\frac{2 \cdot L}{m-2}\right)^{-1} \Big/ \left(\frac{2 \cdot (L-1)}{m-2}\right)^{-1} \right) = -1$$

$$\lim_{L\to\infty} \frac{T_2}{T_1} = \left(\frac{m}{m-2}\right) \Big/ \lim_{L\to\infty} \left(\frac{2 \cdot L}{m-2}\right)^{2/m} = 0$$

$$\lim_{L\to\infty} \frac{T_2}{T_3} = \left(\frac{m}{m-2}\right) \Big/ \lim_{L\to\infty} \left(\frac{2 \cdot L}{m-2}\right)^{2/m} = 0$$

…which confirms the validity of eq. (27) as an approximation at *L>>1*. Table 1 below summarizes the expressions of $\hat{\alpha}$ and $r_{\max}$ with large codebooks.

Table 1: Noise energy of soft VQ with a large codebook

m	$\lim_{L\to\infty} \hat{\alpha}$	$\lim_{L\to\infty} r_{\max}$	L=1,024		L=2,048	
			$\hat{\alpha} \approx$	$r_{max} \approx$	$\hat{\alpha} \approx$	$r_{max} \approx$
2	0	L	0	1,024	0	2,048
4	$L^{-1/4}$	$\frac{1}{2} \cdot \sqrt{L}$	0.177	16	0.149	22.63
6	$\left(\frac{L}{2}\right)^{-1/6}$	$\frac{2}{3} \cdot \sqrt[3]{\frac{L}{2}}$	0.354	5.333	0.315	6.720
m	$\left(\frac{2 \cdot L}{m-2}\right)^{-1/m}$	$\frac{m-2}{m} \cdot \left(\frac{2 \cdot L}{m-2}\right)^{\frac{2}{m}}$				
∞	1	1	1	1	1	1

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Fig. 1 below illustrates the simplest case of application of our proposed soft VQ scheme with a codebook of 2 centroids only: $\underline{c}_1, \underline{c}_2 \in \Re^1$ with different values of the power $m \in \{2,4,6\}$. The figure shows in each case the probability distributions of the belonging of any point in this 1D to both centroids. Fig. 2 is a zoom-in on the narrower interval around the two centroids in fig. 1. It is clear that with higher values of the power, the probability distribution gets sharper and more similar to those of hard VQ.

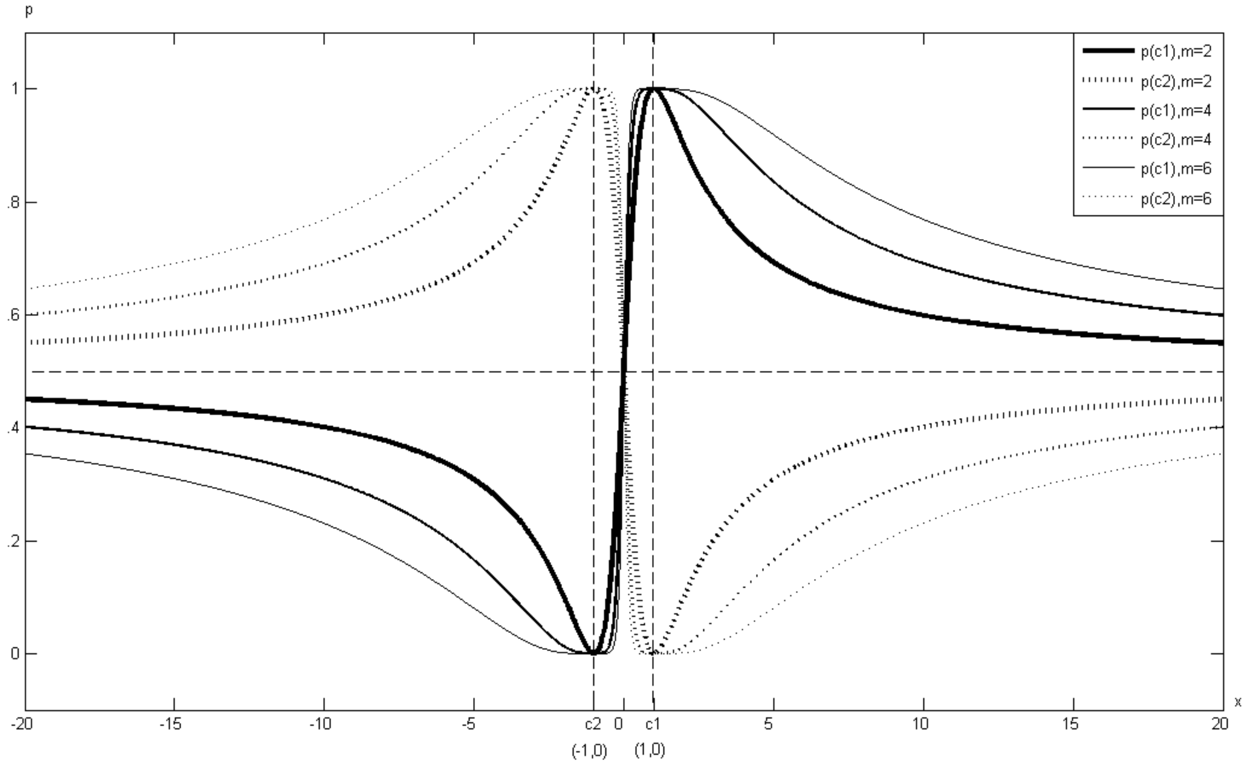

Fig. 1: Probability distributions of the proposed soft VQ with two-centroid codebook with different value of the power *m*.

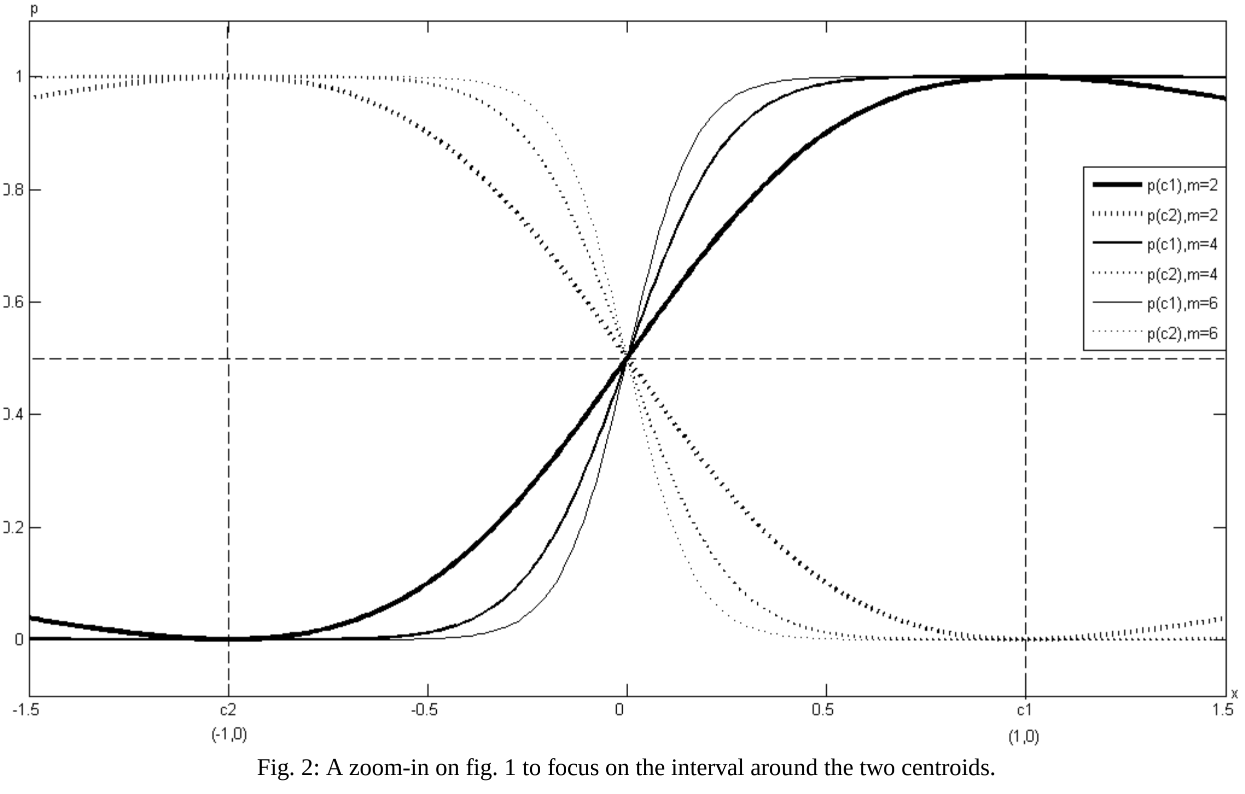

Fig. 2: A zoom-in on fig. 1 to focus on the interval around the two centroids.

Fig. 3 below illustrates the curve of eq. (11) with $L = 1024 >> 1$ at different even values of the power m. It is clear that $r_{\max}$ gets lower with increasing values of $m > 2$ in accordance with eq. (27).

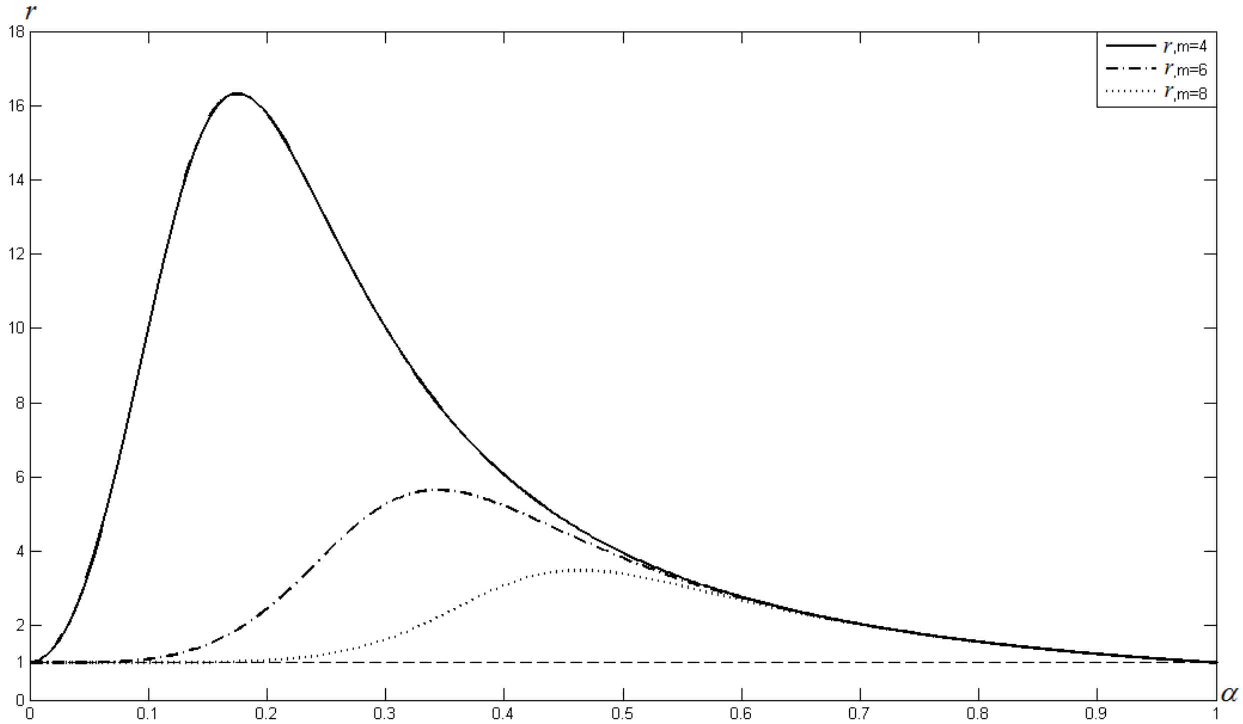

Fig. 3: Quantization noise energy of the proposed soft VQ compared with that of hard VQ with different values of the power *m*.

IV. EXPERIMENTAL SETUP AND RESULTS

In this section, we present the experimentation we conducted to attest the benefits we claimed - in the introduction above - of our proposed soft VQ for machine-learning systems that include a VQ module. The tough problem of font-written Arabic OCR has been selected for this challenge [2] that many groups have been trying to tackle since 3 decades! It is evident in the literature that the most effective methodology for dealing with this problem is the HMM-based one as Arabic is a cursive-scripted language. [1, 3, 5, 6, 11, 12, 15, 16] Among the many attempts made in this regard, we have selected the one that is reported with best performance to experiment with, esp. as its trainer and recognizer are discrete HMM-based and obviously deploy a VQ module. [1, 5, 16]

We setup two versions of this discrete HMM-based Arabic OCR system with *L=2,048*: one with hard-deciding VQ, and the other with our proposed soft VQ with *m=8*. Then we challenged both versions with two sets of test data: assimilation test data and generalization test data.

Assimilation test data consists of 5 test pages for each font/size that the system has seen upon its training phase. Of course the pages themselves are different than the ones used for training, however, they are written in fonts/sizes that the system experienced upon its training.

Moreover, these assimilation test pages are produced in the same conditions as the training ones; i.e. LASER-printed and scanned at 600 dpi. This sets up the favorable runtime conditions of least runtime variance from the training conditions.

Generalization test data, on the other hand, are sample pages selected randomly from some paper-books that are scanned also at 600 dpi. Obviously, there is no control on the fonts/sizes used in these pages. The tilting distortion and the printing noise are also quite apparent.[4] Of course, this sets up the less favorable runtime conditions of more considerable variance from the training conditions.

Table 2 below gives the measured word error rate (WER) of each version with each test data set.

Table 2: Results with hard and soft VQ versions of the OCR

WER of Assimilation Test		WER of Generalization Test	
Hard VQ	Soft VQ	Hard VQ	Soft VQ
3.08%	3.71%	16.32%	13.98%

While the OCR version with hard VQ produced a smaller WER than the version with soft VQ with the assimilation test, the soft VQ version outperformed the hard VQ version with the generalization test.

[4] The generalization test data set and the corresponding output of both versions are downloadable at http://www.rdi-eg.com/Soft_Hard_VQ_OCR_Generalization_Test_Data.RAR.

As the models built upon training of the hard VQ version were more over-fitted to the training data than those built with soft VQ, it was easier for the former one to recognize the “similar” inputs from assimilation test data with a narrower WER. On the other hand, the more “flexible” models built with soft VQ were more capable to absorb the much more variance in the inputs from the generalization test data than those built with hard VQ.

This observed behavior nicely matches our theoretical claims on the positive impact of our proposed soft VQ on machine-learning systems as mitigating over-fitting and rendering their performance more robust with runtime variances such as noise.

V. CONCLUSION

In this paper, we have discussed the virtues of soft vector quantization over the conventional hard-deciding one, and then proposed a soft vector quantization scheme with inverse power-function distributions. The quantization noise energy of this soft VQ compared with that of hard VQ is then analytically investigated to derive a formula of an upper bound on the ratio between the quantization noise energy in both cases.

To attest our claims on the advantages of the proposed soft VQ for machine-learning, we have experimented with one of the best reported discrete HMM-based Arabic OCR systems. We setup two versions of the OCR system: one with the conventional hard VQ and the other with the proposed soft VQ. We then challenged each version with two sets of test data; assimilation test data and generalization test data.

While the OCR version with hard VQ realized a smaller word error rate than the version with soft VQ with the assimilation test, the latter outperformed the former one with the generalization test. These results nicely match our claimed positive impact of our proposed soft VQ on machine-learning systems as mitigating over-fitting and rendering their performance more robust with runtime variances; e.g. noise.

REFERENCES[5]

[1] Attia, M., Rashwan, M., El-Mahallawy, M., *Autonomously Normalized Horizontal Differentials as Features for HMM-Based Omni Font-Written OCR Systems for Cursively Scripted Languages*: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5478619, IEEE International Conference on Signal & Image Processing Applications (ICSIPA09); http://www.SP.ieeeMalaysia.org/ICSIPA09, Kuala Lumpur-Malaysia, Nov. 2009.

[2] Attia, M., *Arabic Orthography vs. Arabic OCR; Rich Heritage Challenging A Much Needed Technology*, Multilingual Computing & Technology magazine www.Multilingual.com, USA, Dec. 2004.

[3] Bazzi, I., Schwartz, R., Makhoul, J., *An Omnifont Open-Vocabulary OCR System for English and Arabic*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, No. 6, June 1999.

[4] Duda, R.O., Hart, P.E., *Pattern Classification and Scene Analysis*, (2nd ed.), John Wiley & Sons, New York, 2000.

[5] El-Mahallawy, M.S.M., *A Large Scale HMM-based Omni Front-Written OCR System for Cursive Scripts*, PhD thesis, Dept. of Computer Engineering, Faculty of Engineering, Cairo University, Apr. 2008.

[6] Gouda, A. M., *Arabic Handwritten Connected Character Recognition*, PhD thesis, Dept. of Computer Engineering, Faculty of Engineering, Cairo University, Nov. 2004.

[7] Gray, R.M., *Vector Quantization*, IEEE Signal Processing Magazine, pp. 4-29, Apr. 1984.

[8] Gray, R.M., Neuhoff, D.L., *Quantization*, IEEE Transactions on Information Theory, Vol. 44, No. 6, pp. 2325-2383, October 1998.

[9] Jacobson, N., *Basic algebra*, Vol. 1 (2nd ed.), Dover, ISBN 978-0-486-47189-1, 2009.

[10] Jain, A.K., *Data Clustering: 50 Years Beyond K-means* http://biometrics.cse.msu.edu/Presentations/FuLectureDec5.pdf, Plenary Talk at The IAPR’s 19th International Conference on Pattern Recognition http://www.icpr2008.org/, Tampa-Florida-USA, Dec. 2008.

[11] Kanungo, T., Marton, G., and Bulbul, O., *OmniPage vs. Sakhr: Paired Model Evaluation of Two Arabic OCR Products*, Proc. SPIE Conf. Document Recognition and Retrieval (VI), pp. 109-121, 1999.

[12] Khorsheed, M.S., *Offline Recognition of Omnifont Arabic Text Using the HMM ToolKit (HTK)*, Pattern Recognition Letters, Vol. 28 pp. 1563–1571, 2007.

[13] Kövesi, B., Boucher, J-M., Saoudi, S., *Stochastic K-means Algorithm for Vector Quantization*, Pattern Recognition Letters-Elsevier, Vol. 22, Issues 6-7, pp. 603-510, May 2001.

[14] Linde, Y., Buzo, A., Gray, R.M., *An Algorithm for Vector Quantizer Design*, IEEE Trans. Communications, Vol. COM-28, pp. 84–95, Jan. 1980.

[15] Mohamed, M., Gader, P., *Handwritten Word Recognition Using Segmentation-Free Hidden Markov Modeling and Segmentation-Based Dynamic Programming Techniques*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No. 5, pp.548–554, 1996.

[16] Rashwan, M., Fakhr, M., Attia, M., El-Mahallawy, M., *Arabic OCR System Analogous to HMM-Based ASR Systems; Implementation and Evaluation*, Journal of Engineering and Applied Science, Cairo University, www.Journal.eng.CU.edu.eg, Vol. 54 No. 6, pp. 653-672, Dec. 2007.

[17] Seo, S., Obermayer, K., *Soft Learning Vector Quantization*, ACM’s Neural Computation, Volume 15-Issue 7, pp. 1589-1604, MIT Press, July 2003.

[5] References number [1], [2], [5], and [16] are freely downloadable at http://www.rdi-eg.com/technologies/papers.htm (last section in the page).