Soft Vector Quantization with Inverse Power-Function Distributions for Machine Learning Applications

Mohamed Attia, Abdulaziz Almazyad, Mohamed El-Mahallawy, Mohamed Al-Badrashiny, Waleed Nazih

***Abstract*-This paper discusses the positive impact of soft vector quantization on the performance of machine-learning systems that include one or more vector quantization modules. The most impactful gains here are 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 paper then introduces a soft vector quantization scheme with inverse power-function distributions, and analytically derives an upper bound of its relative quantization noise energy to that of typical (hard-deciding) vector quantization. This relative noise is expressed as a closed-form function of the power in order to allow the selection of its optimal values of that compromise both a soft enough vector quantization with a stable performance via small enough relative quantization noise.

Finally, we present empirical evidence obtained via experimenting with two versions of the best reported OCR system for cursive scripts - that happened to deploy discrete HMMs - one version with hard vector quantization and the other with our herein presented soft quantization. Test samples of real-life scanned Arabic text pages are used to challenge both versions; hence the recognition error margins are compared.

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

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. [5] Typical VQ deploys the minimum-distance criterion that:

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

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

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

The total quantization noise energy over a population of points[3] in this space of size *s* versus that codebook of centroids [5, 8, 11] is hence given by:

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

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. [1, 11, 14, 18] With machine-learning/classification 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 subsequent 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. 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

Manuscript received Dec. 20th, 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).

Abdulaziz Almazyad is the dean of the College of Computer Engineering and Sciences in AlKharj University - AlKharj - KSA, and is also an assistant professor in the dept. of Computer Engineering in the College of Computer and Information Sciences - King Saud University ksu.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).

Waleed 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.

point to each centroid in the codebook. Intuitively, the closer is the point to some centroid 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 cope with run-time variances than those resulting from hard-deciding VQ.

Formally, soft VQ may in general be formulated as:

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

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:

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

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

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

This means that the price we pay for a soft VQ is a higher harmful quantization noise energy that may hinder any machine-learning method, 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 the best performing - according to the published literature [2, 6] - OCR system for type-written cursive scripts - that happened to deploy discrete HMMs - 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 DISTRIBUTION

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 emergences 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:

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

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

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

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

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

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

Let us define:

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

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

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

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

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

$$= \frac{1 + \xi_N + \omega \cdot \lim_{\delta \to 0}\left(\frac{1}{\delta}\right)^{2-m}}{1 + \xi_D + \omega \cdot 0} = \frac{1 + \xi_N + \infty}{1 + \xi_D + 0} = \infty$$

…where $\omega = SizeOf(\Omega)$ and $0 < \xi_N, \xi_D < L - \omega - 1$.

This result necessitates the avoidance of the interval of $0 < m < 2$ as the possible unlimited growth of the soft quantization noise energy with respect to that of hard quantization would be devastating to the stability of whatever machine learning procedure!

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:

$$\max\left(r\big|_{m=2}\right) = \lim_{\alpha_j \to 0 \forall j \neq i_0; 1 \le j \le L} \left(r\big|_{m=2}\right) = L \qquad (14)$$

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 ratio of soft VQ noise energy vs. that of hard VQ which can still ruin machine learning esp. as this worst case occurs at the dominant situation of points being 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:

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

…where ℓ is the number of α_j's that are exactly equal to one. This 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 also 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 - according to eq. (6) - is an upper bound of the ratio between the total quantization noise energy of the proposed soft VQ to that of the conventional hard 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:

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

…that reduces into:

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

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:

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

$$r_{\max} = \frac{A + \hat{\alpha}^{m-2}}{B + \hat{\alpha}^{m}} = \frac{1 + (L-1)\cdot \hat{\alpha}^{m-2}}{1 + (L-1)\cdot \hat{\alpha}^{m}} = \left(\frac{m-2}{m}\right)\cdot \hat{\alpha}^{-2} \qquad (23)$$

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

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

$$m > 2, L \ge 2, \hat{\alpha} \in [0,1]$$

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

$\therefore g(\hat{\beta})$ 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 algebraically extractable only for $(m/2)\in\{2,3,4,5\}$. [10]

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:

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

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 (see [10]):

$$\hat{\beta}=-\frac{1}{3}\cdot\sqrt[3]{\left(\frac{1}{2}\right)\cdot\left(27\cdot\eta_0-\sqrt{27\cdot\eta_0^2+4\times27\cdot\eta_1^3}\right)}$$
$$-\frac{1}{3}\cdot\sqrt[3]{\left(\frac{1}{2}\right)\cdot\left(27\cdot\eta_0+\sqrt{27\cdot\eta_0^2+4\times27\cdot\eta_1^3}\right)}$$

$\left.\hat{\alpha}^2\right|_{m=6}$ and $r_{\max}|_{m=6}$ are hence given by:

$$\left.\hat{\alpha}^2\right|_{m=6}=\frac{\sqrt[3]{\left(1+\sqrt{1+\frac{1}{L-1}}\right)}-\sqrt[3]{\left(1-\sqrt{1+\frac{1}{L-1}}\right)}}{\sqrt[3]{L-1}}$$

$$r_{\max}|_{m=6}=\frac{(2/3)\cdot\sqrt[3]{L-1}}{\sqrt[3]{\left(1+\sqrt{1+\frac{1}{L-1}}\right)}-\sqrt[3]{\left(1-\sqrt{1+\frac{1}{L-1}}\right)}}$$

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

_____ (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 speculate the generalization that:

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

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: Relative noise energy of soft VQ with large codebooks

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
8	$\left(\frac{L}{3}\right)^{-1/8}$	$\frac{3}{4}\cdot\sqrt[4]{\frac{L}{3}}$	0.482	3.224	0.442	3.834
m	$\left(\frac{2\cdot L}{m-2}\right)^{-1/m}$	$\frac{m-2}{m}\cdot\left(\frac{2\cdot L}{m-2}\right)^{2/m}$				
∞	1	1	1	1	1	1

Fig. 1 below illustrates the simplest case of application of our proposed soft VQ scheme with a codebook of 2 centroids only: $\underline{c}_1, \underline{c}_2 \in \Re^1$ at different values of the power $m \in \{2,4,6\}$. The curves show in each case the probability distribution of the chance of belonging of any point in this 1D space to $\underline{c}_1$ (with continuous lines) and to $\underline{c}_2$ (with dotted lines). 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 *m*, the probability distribution gets sharper and more similar to those of hard VQ.

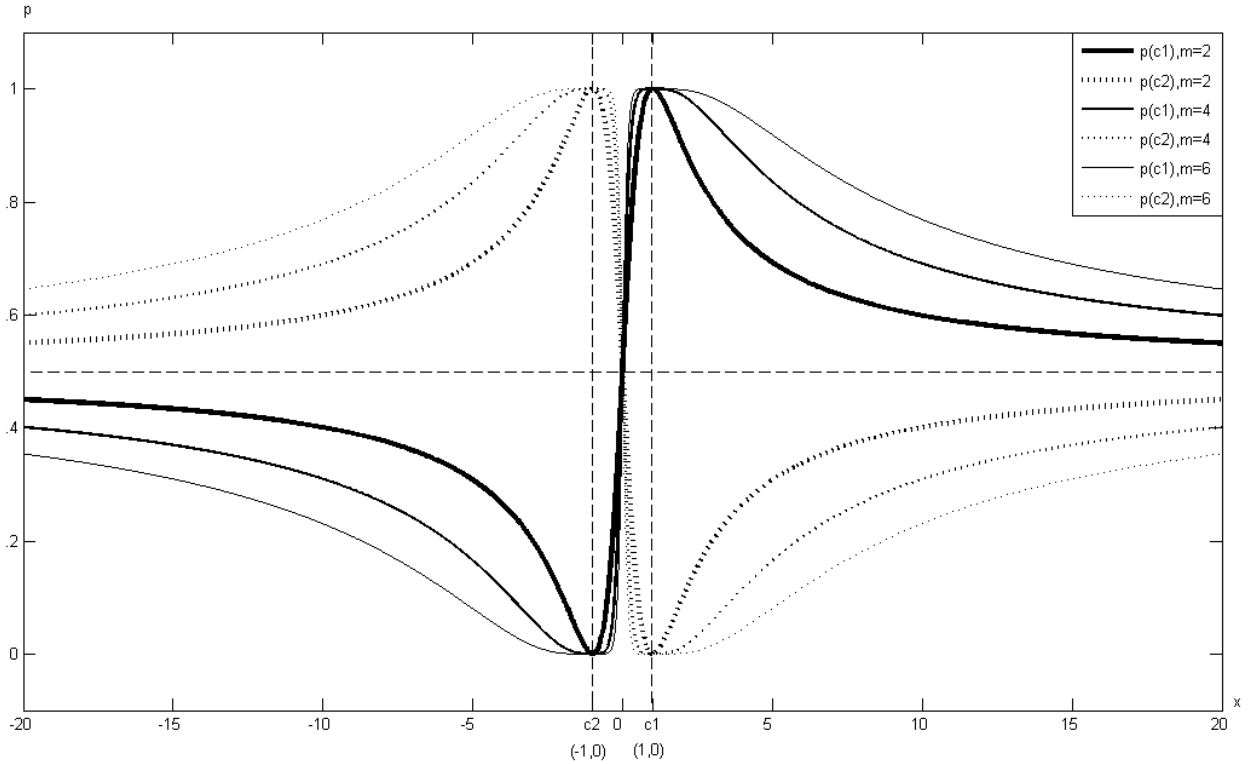

Fig. 1: Probability distributions of the proposed soft VQ with two-centroid codebook at different values of the power *m*.

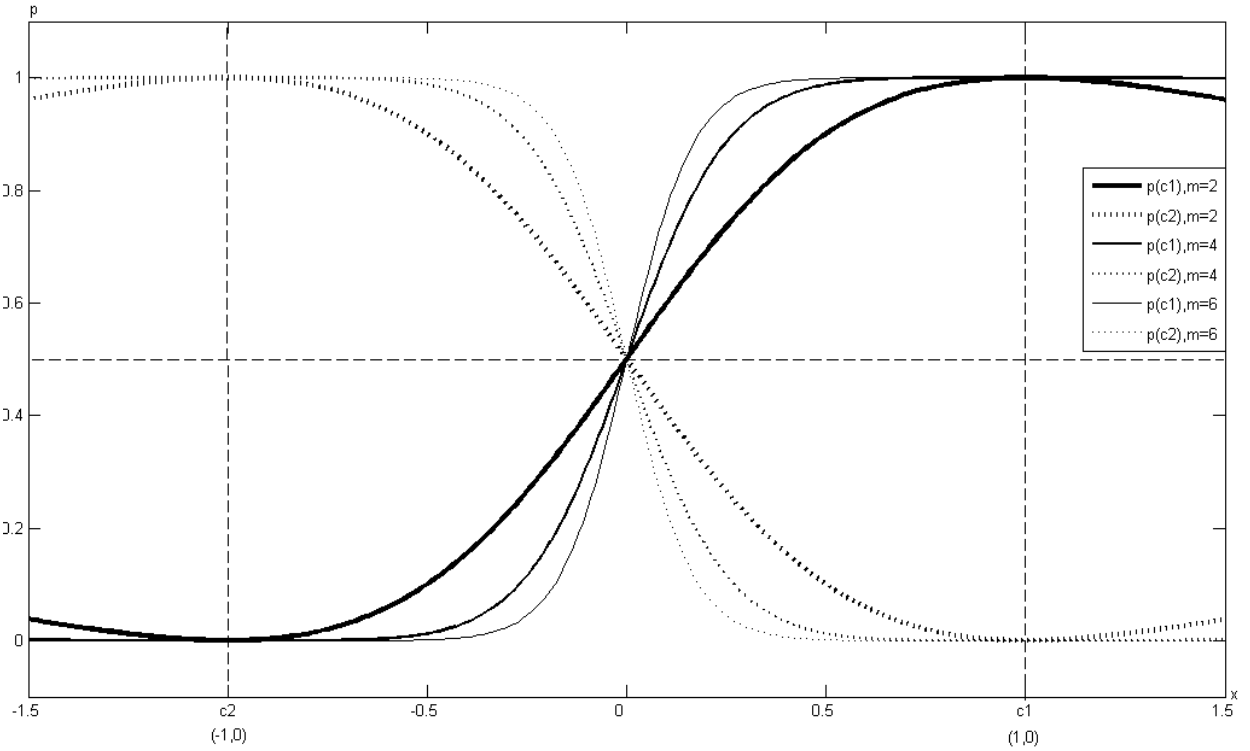

Fig. 2: A zoom-in on fig. 1 with a focus on the interval around the two centroids.

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).

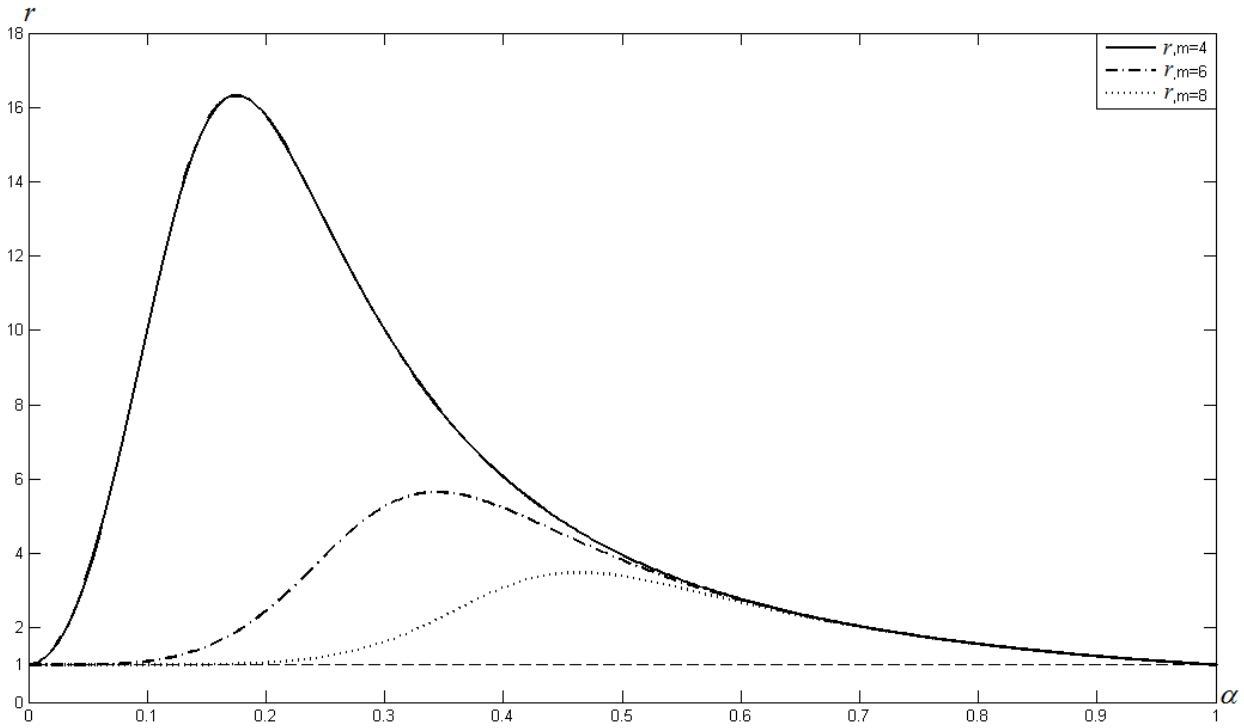

Fig. 3: Quantization noise energy of the proposed soft VQ relative to that of hard VQ at different values of the power *m*.

IV. EXPERIMENTAL SETUP AND RESULTS

This section presents 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.

For this challenge, we have chosen the task of Optical Character Recognition (OCR) for Arabic script [3] which is an instance of a broader family of cursive scripts including Persian, Urdu, Kurdish, etc. This is a 4-decade old tough digital pattern recognition problem as the connected characters in the script need to be both identified and segmented simultaneously. It is evident in the literature that the most effective methodology for dealing with such a problem is the HMM-based one. [2, 4, 6, 7, 12, 13, 16, 17]

Among the many recent attempts made on this problem, we have selected *CleverPage*© to experiment with[4]. *CleverPage*© is an ASR[5]-like HMM[6]-based Arabic Omni Type-Written OCR system that is reported to achieve the highest accuracy in this regard. [2, 6] The most characterizing innovation in this OCR system that puts *CleverPage*© ahead of its rivals is its lossless recognition features which are autonomously normalized horizontal differentials encoded in 16-component vectors sequence.

Each features vector is computed to differentially encode the pixels pattern in each single-pixel width slice (i.e. frame) included within the right-to-left horizontally sliding window. Given a sequence of such features, one can retrieve the shape of the digitally scanned type-written script with only a scaling factor along with an inevitable limited digital distortion.

Each single-pixel frame of a given Arabic word is assumed to have a limited maximum number of vertical dark segments (4 segments are found to be enough). Each of these segments is coded into 4 components differentially representing the topological and agglomerative properties of the segment as follows:

1st component: The segment's center of gravity with respect to that of the previous frame.
2nd component: The segment's length with respect to that of the previous frame.
3rd and 4th components: The orders of the most bottom, and the top segments in the preceding frame which are in 8-connection with the current segment. Special negative codes are given in case of non-connected segments.

Empty segments are padded by zeros. The dimensionality of this features vector is 16 = 4 segments × 4 components per segment. While half of these components are sharply discrete in nature, the others are analog, which made it quite hard to find Gaussian mixtures that properly represent that kind of hybrid data. So, discrete HMMs rather than GMMs have been resorted to as the recognition vehicle of this system. [2, 6, 17]

[4] See http://www.rdi-eg.com/technologies/OCR.htm

[5] Automatic Speech Recognition

[6] Hidden Markov Models

This made *CleverPage*© an excellent candidate for our experimentation especially as its producer[7] was generous enough to allow us set up 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. This is an empirical selection of the value of m that compromises between both minimal quantization noise energy of soft VQ relative to that of hard VQ, along with soft enough distributions of our soft VQ scheme.

The training data of both setups covers 10 different popular Arabic fonts with 6 sizes per each font. It contains 25 LASER-printed pages per each size of each font with all the pages scanned at 600dpi B&W along with the correct digital text content of each page. [6]

Then we challenged both versions with two sets of test data: assimilation test data and generalization test data.

The assimilation test data consists of 5 test pages for each of the 60 (font, size) pairs represented in the training phase. Of course the pages themselves do not belong to the training data. These assimilation test pages are produced in the same conditions as the training ones. This provides the favorable runtime conditions of least runtime variance from the training conditions.

The generalization test data, on the other hand, are sample pages picked randomly from some Arabic paper-books that are scanned also at 600 dpi.[8] Obviously, there is no control on the fonts or sizes used in these pages. The tilting distortion and the printing noise are also quite apparent. Of course, this provides the less favorable runtime conditions of more considerable variance from the training conditions.

Table 2 below compares the measured word error rates of both versions with each of the two test data sets.

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

Error Rate of Assimilation Test		Error Rate of Generalization Test	
WER_A		WER_G	
Hard VQ	Soft VQ	Hard VQ	Soft VQ
3.08%	3.71%	16.32%	13.98%
CER_A		CER_G	
Hard VQ	Soft VQ	Hard VQ	Soft VQ
0.77%	0.93%	4.08%	3.50%
Degradation Due to Soft VQ		**Enhancement Due to Soft VQ**	
-20.45%		+14.34%	

CER, in the table above, stands for Character Error Rate while *WER* stands for Word Error Rate. Under the assumption of a single character error per word - which is valid in our case - the relation $WER \approx h \cdot CER; CER << 1$ holds true with $h \approx 4$ for Arabic. [2] While *WER* is the rate perceived by the OCR end user, most of the researchers and vendors prefer to use CER to optimistically express the performance of their systems.

While the OCR version with hard VQ produced smaller error rates than the version with soft VQ in the assimilation test, the soft VQ version outperformed the hard VQ version in the generalization test.

As the models built upon the 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 to recognize the "similar" inputs from assimilation test data with a narrower error margin. 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 claimed benefits of our proposed soft VQ for machine-learning systems as mitigating over-fitting and rendering their performance more robust with runtime variances like noise.

V. CONCLUSION

This paper discussed the virtues of soft vector quantization over the conventional hard-deciding one esp. for machine-learning applications. It then proceeded to propose 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. This formula reveals the proper values of the power where it is safe to use with such distributions without ruining the stability of machine-learning.

To attest the claimed benefits of our proposed soft VQ for machine-learning, we have experimented with a recent Omni Type-Written OCR for cursive scripts whose recognition error margin with Arabic printed text is reported to be the minimum. This OCR has an ASR-like HMM-based architecture with lossless recognition features vector combining both analog and sharply discrete components, which necessitated the usage of discrete HMMs hence the deployment of VQ. We setup two versions of this 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 smaller error rates than the version with soft VQ in the assimilation test, the latter outperformed the former in the generalization test. These results 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.

[7] RDI; www.RDI-eg.com

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

REFERENCES

[1] Attia, M., Almazyad, A., El-Mahallawy, M., Al-Badrashiny, M., Nazih, W., *Post-Clustering Soft Vector Quantization with Inverse Power-Function Distribution, and Application on Discrete HMM-Based Machine Learning*, Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2010, WCECS 2010, 20-22 October, 2010, San Francisco, USA, pp.574-580.

[2] 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.

[3] Attia, M., *Arabic Orthography vs. Arabic OCR; Rich Heritage Challenging A Much Needed Technology*, Multilingual Computing & Technology magazine www.Multilingual.com, USA, Dec. 2004. This paper is downloadable at the last section of the page: http://www.rdi-eg.com/technologies/papers.htm.

[4] 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.

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

[6] 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. The thesis is available at the last section of the page: http://www.rdi-eg.com/technologies/papers.htm.

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

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

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

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

[11] 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.

[12] 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.

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

[14] Kövesi, B., Boucher, J-M., Saoudi, S., *Stochastic K-means Algorithm for Vector Quantization*, Pattern Recognition Letters-Elsevier, Vol. 22, Issues 6-7, pp. 603-610, May 2001.

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

[16] 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.

[17] 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. This paper is downloadable at the last section of the page: http://www.rdi-eg.com/technologies/papers.htm.

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