Transform Coding with Backward Adaptive Updates

Vivek K Goyal, *Member, IEEE*, Jun Zhuang, and Martin Vetterli, *Fellow, IEEE*

This work was initiated while the first and second authors were with the University of California, Berkeley. Preliminary results were presented at IEEE Int. Conf. Image Proc. Sept. 16–19, 1996 and IEEE Data Compression Conf. March 25–27, 1997. Manuscript submitted April 1998.

V. K. Goyal is with Bell Labs, Lucent Technologies, Murray Hill, NJ.

J. Zhuang is with SBC Technology Resources, Inc., Pleasanton, CA.

M. Vetterli is with the Laboratoire de Communications Audiovisuelles, École Polytechnique Fédérale de Lausanne, Switzerland and the Dept. of Electrical Eng. & Comp. Sci. University of California, Berkeley, CA.

Abstract

The Karhunen–Loève transform (KLT) is optimal for transform coding of a Gaussian source. This is established for all scale invariant quantizers, generalizing previous results. A backward adaptive technique for combating the data-dependence of the KLT is proposed and analyzed. When the adapted transform converges to a KLT, the scheme is universal among transform coders. A variety of convergence results are proven.

Index terms: lossy data compression, universal source coding, transform coding, dithered quantization

I. Introduction

The essence of transform coding is to apply a linear transform to a source vector and then apply scalar quantization, as opposed to applying scalar quantization directly to the source vector. Heuristically, transform coding works because the transform can eliminate correlation between components of the source vector, producing a vector of transform coefficients more amenable to scalar quantization and entropy coding. Transform codes are popular because they provide an attractive compromise between computational complexity and performance. In the parlance of vector quantization, the point-density and oblongitis losses of scalar quantization are eliminated or reduced, leaving predominantly only a space-filling loss [1].

With a Gaussian source model, the optimal transform is a Karhunen–Loève transform (KLT), an orthonormal transform that produces uncorrelated transform coefficients. The optimality of the KLT is well-known for high rates [2] or when optimal fixed rate quantizers are employed [3], but holds more generally (see Appendix A). However, the KLT is rarely used in practice for a variety of reasons. One prominent reason is that the KLT is signal dependent; the transform used in the encoder and decoder must be adjusted to correspond to the covariance of the source in order to maintain optimality. A second reason is that since the KLT has no special structure, it requires more operations to compute than a harmonic transform such as a discrete cosine transform. For vectors of length of N, the complexity difference is roughly N^2 compared to $N \log N$, which is not overwhelming for small values of N.

This correspondence addresses only the first issue—the matching of transform to source. A *backward adaptive* method for transform adaptation is proposed and analyzed. In backward adaptation the encoder and decoder adapt in unison based on the coded data without the explicit transmission of coder parameters. Backward adaptation is also called *adaptation without side*

information or *on-line adaptation*.

The use of backward adaptation for transform adaptation in transform coding seems to be unprecedented, though backward adaptive techniques have a long history. For example, adaptation of prediction filters in speech coders is often backward adaptive [4], [5] and ADPCM includes not only backward adaptation of filter taps but also of quantizer scaling [6]. Similar to the quantizer scaling in ADPCM is the backward adaptive context modeling and quantizer scaling of the EQ image coder [7]. It is also possible to adapt a quantizer more generally without side information [8].

The incompletely realized aim of our work is to show that backward adaptation can result in a transform code that is *universal* for Gaussian sources. "Universal" is used here to mean that the performance approaches that of an ideal *transform code* designed with *a priori* knowledge of the source distribution. The results along these lines are asymptotic in the data length, but the transform or block size is fixed. Empirical evidence and partial analyses are provided. Such a code would be an "on-line" alternative to the "universal codebook" approach to universal transform coding by Effros and Chou [9].[1] Forward adaptive techniques that are not necessarily universal are discussed, *e.g.*, in [11].

The results of [9] were inspiring to this study because they indicated superior performance of weighted universal transform coding over weighted universal vector quantization for image compression with reasonable vector dimensions. It was also shown that there are sizable gains to be realized by varying the transform, a result that runs counter to the conventional wisdom in image compression.

In the remainder of the correspondence, the aforementioned ideas are made more precise. The sources and coding structures under consideration are described in Section II. Unable to satisfactorily analyze the original coding structure, we give several analyses based on simplifying assumptions. The main results are stated in Section III and proven in Appendix B. Section IV describes ways in which the encoding algorithms can be modified to reduce computational complexity or to track a varying source. Concluding comments appear in Section V.

[1]See the taxonomy of universal coding methods by Zhang and Wei [10] for explanations of the quoted terms.

II. Proposed Backward Adaptive Coding Structure

Let $\{x_n\}_{n\in\mathbb{Z}^+}$ be a sequence of independent, identically distributed (i.i.d.), zero-mean Gaussian random vectors of dimension N with covariance matrix $R_x = E[xx^T]$.[2] If R_x is not diagonal, *i.e.*, the components of x are correlated, one obtains better rate–distortion performance with transform coding than with direct scalar quantization and scalar entropy coding of the source vectors.

In transform coding, a square, invertible linear transform T is applied to each source vector to get a vector of *transform coefficients* $y_n = Tx_n$. The transform coefficients undergo scalar quantization and scalar entropy coding. Ideally, the transform should be selected such that the transform coefficients are uncorrelated and hence, since the source is Gaussian, independent. This was first shown by Huang and Schultheiss [3] under assumptions of optimal fixed-rate quantization and a mild, commonsense condition on the bit allocation. (Earlier work by Kramer and Mathews [12] did not involve quantization and was not in an operational rate–distortion framework.) Using high-resolution quantization theory, the same result can be obtained for optimal variable-rate (entropy coded) quantization or uniform quantization [2]. A new extension is given in Appendix A that relies only on the scalar quantizers having performance invariant to scaling (Theorem 6).

To mathematically describe an optimal transform T, simply note that by linearity of the expectation operator $R_y = E[(Tx)(Tx)^T] = TR_xT^T$. Thus T may be an orthonormal similarity transform composed of eigenvectors of R_x. This makes R_y a matrix with the eigenvalues of R_x on its main diagonal and zeros elsewhere. Such a transform is a Karhunen–Loève transform (KLT) of the source.

Since the optimal transform T depends on an ensemble average R_x, it is generally unknown at the encoder. (It may also be the case that $E[x_nx_n^T]$ varies slowly with n, though we will deal with this case only in passing.) We consider here systems that periodically adjust the transform at the encoder and decoder in a backward adaptive manner. A block diagram for such a system is shown in Figure 1. In this system, the quantizer Q is a scalar quantizer with uniform quantizer

[2]Throughout the paper, R_v will be used to denote the (exact) covariance matrix $E[vv^T]$ of a random vector v. $\widehat{R_v}$ denotes an estimate of R_v obtained from a finite length observation. Aside from this convention, subscripts indicate the time index of a variable, except where two subscripts are given to indicate the row and column indices of a matrix. A superscript T indicates a transpose.

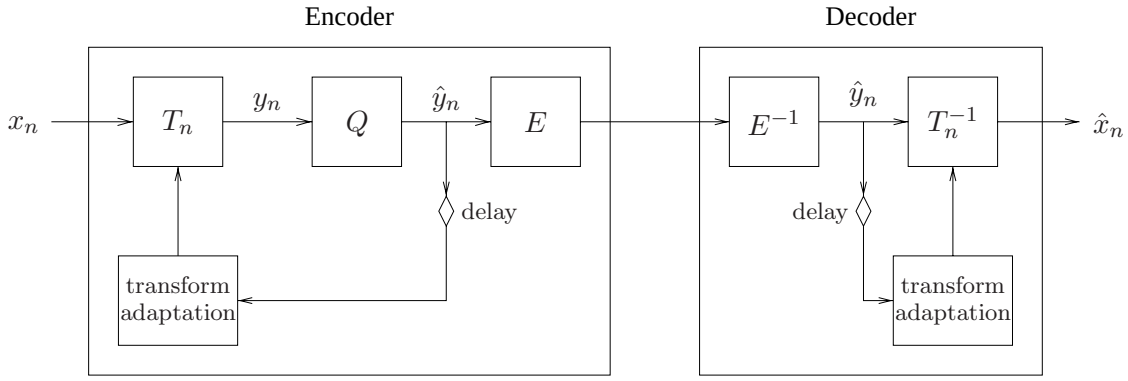


Fig. 1. Block diagram of transform coding system with backward adaptive transform updates. T_n is a time-varying orthogonal transform, Q is a scalar quantizer, and E is a universal scalar entropy coder.

q applied to each component:

$$q(x_i) \;=\; k_i\Delta, \qquad \text{for} \quad (k_i - \frac{1}{2})\Delta \le x_i < (k_i + \frac{1}{2})\Delta, \quad k_i \in \mathbb{Z}, \quad i = 1, 2, \ldots, N. \tag{1}$$

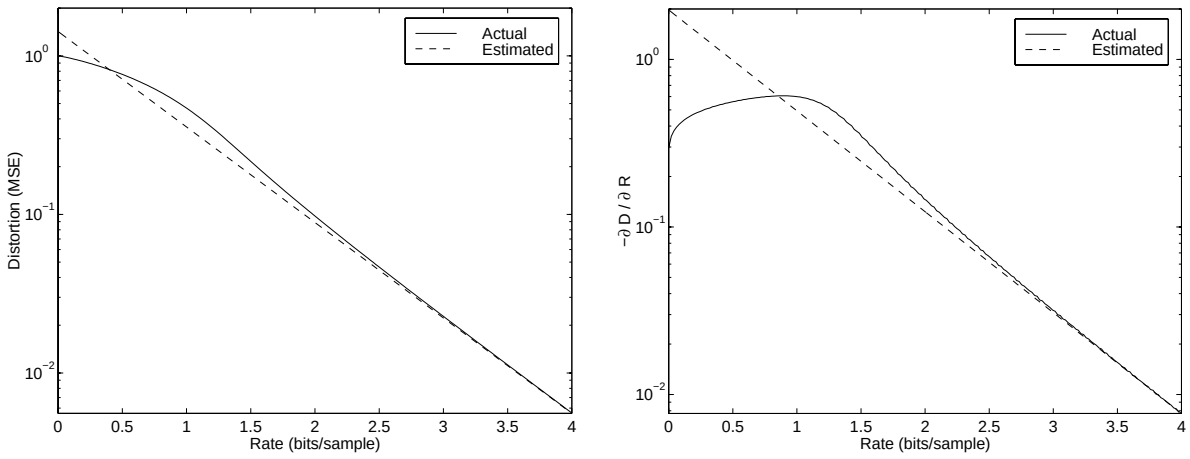
The entropy coder has N separate universal lossless codes for the N transform coefficient streams.

In this work we concentrate on the update mechanism for the transform and the effect of the transform updates. This is partly a matter of taste, but it is also motivated by the insensitivity of the optimal quantizer to the source and transform. The use of uniform scalar quantization with equal step sizes for each component is discussed in Section II-A and transform update procedures are considered in Section II-B.

A. Focusing on the Transform

Consider the quantization and entropy coding of a single transform coefficient branch in Figure 1. Since the quantizer indices are entropy-coded, the proper optimization criterion for the quantizer is to minimize the distortion for a given entropy coder output rate. Assuming that the transform and the universal lossless codes converge, this rate is well-approximated by the entropy rate of the quantizer output sequence. With this approximation one is left with an *entropy-constrained scalar quantizer* to design.

Even assuming that the variance of the transform coefficient is known, the best quantizer will generally be known only through a numerical optimization procedure. However, a uniform quantizer is optimal asymptotically for high rates [13] and, more importantly, is close to optimal at moderate rates [14]. This is an important distinction between fixed-rate and variable-rate scalar quantization that partially justifies our use of fixed uniform quantizers. (Alternatively, it



(a) (b)

Fig. 2. Comparisons between the actual performance of entropy coded uniform scalar quantization and high resolution approximations: (a) Actual distortion–rate performance compared to (2); and (b) Derivative of the actual distortion–rate performance compared to (4).

was shown in [8] that backward adaptation of fixed-rate quantizers can be successful, but this is not pursued here.)

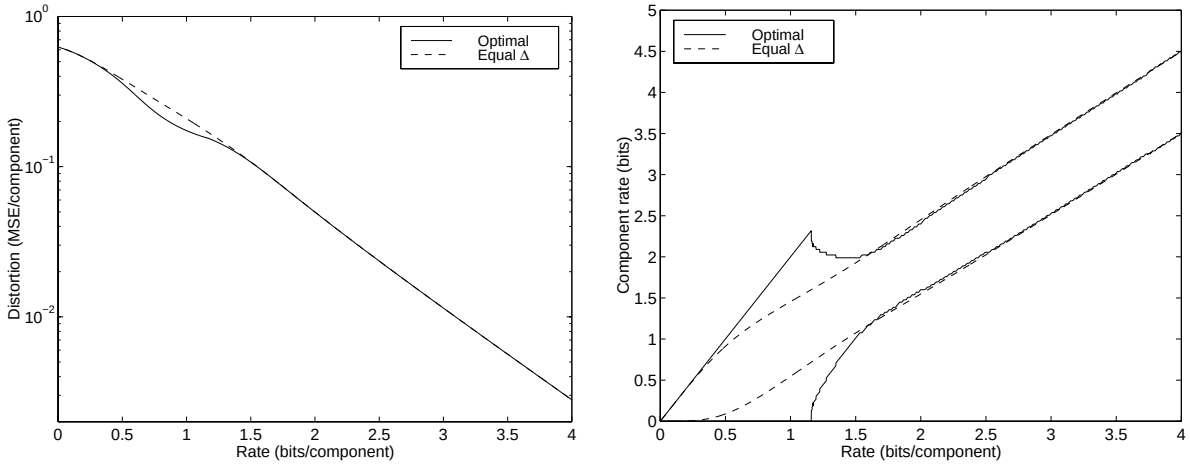
Now consider the joint optimization of the set of scalar quantizers. Using high-resolution analysis, it is easy to show that the optimal allocation of rates between the transform coefficients results in equal distortions and equal quantization step sizes for each transform coefficient [2]. Though this result is well-known, the minimum rate at which this is a good approximation is not; thus, we present some numerical calculations. At high rates, the operational distortion–rate performance of entropy-coded uniform quantization (ECUQ) of a Gaussian source with variance σ^2 is given approximately by

$$D \;=\; \frac{\pi e}{6}\,\sigma^2\,2^{-2R}. \tag{2}$$

This is easily obtained by combining the $D \approx \Delta^2/12$ distortion of fine, uniform quantization with Rényi's relation between the differential entropy of a continuous source and its uniformly quantized version [15]:

$$H(q(X)) \;\approx\; h(X) - \log_2 \Delta. \tag{3}$$

The inaccuracy of (2) at low rates is apparent from the fact that the maximum distortion should be σ^2; the distortion given by (2) exceeds σ^2 for rates below ≈ 0.255 bits. The actual distortion–rate behavior is compared to (2) in Figure 2(a).



(a) (b)

Fig. 3. Comparisons between the performance of optimal bit allocation and equal quantization step sizes for variables with variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 1/4$: (a) distortion–rate performances; and (b) bit allocations.

The simplicity of bit allocation using (2) is due to the form of $\partial D/\partial R$. Consider the allocation of R_1 and R_2 bits between components with variances σ_1^2 and σ_2^2, respectively. Since

$$\frac{\partial D_i}{\partial R_i} \;=\; -\frac{\pi e \log 2}{3}\,\sigma_i^2\, 2^{-2R_i}, \qquad i = 1, 2, \tag{4}$$

operating at equal slopes demands $\sigma_1^2 2^{-2R_1} = \sigma_2^2 2^{-2R_2}$. This in turn makes the component distortions equal and, again using high-resolution approximations, the quantization step sizes equal. This analysis demonstrates that using equal quantization step sizes is a good approximation to optimal bit allocation when (4) is accurate. This is true for rates above about 1 bit per sample (see Figure 2(b)).

To conclude the discussion of bit allocation, let us look at the effect of optimal bit allocation in one simple example. Variables with variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 1/4$ are quantized by ECUQ either with optimal bit allocation or with equal quantization step sizes. Figure 3(a) compares the distortion–rate performances and Figure 3(b) compares the bit allocation. It is apparent that optimal bit allocation provides little improvement. Note also that the optimal bit allocation is predicted well by the high-resolution analysis when the lower rate is at least 1 bit per sample.

For the remainder of the paper, ECUQ with equal quantization step sizes for all components is employed exclusively. With this restriction, we may fix the quantization step size Δ and focus on the entropies of the quantizer outputs; for small Δ the distortion is insensitive to the choice

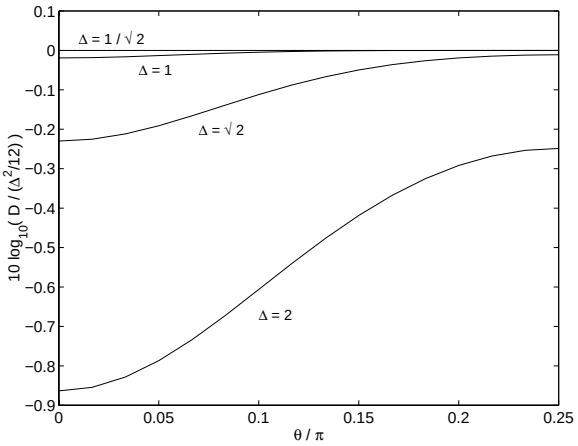


Fig. 4. Dependence of overall distortion on the choice of transform for a two-dimensional source. The dependence is mild and vanishes as the quantization step size Δ shrinks.

of transform. In the limit as Δ approaches zero, this insensitivity is clear because the distortion approaches $\Delta^2/12$ per component. It turns out that the deviation from this approximation is less than 5% for rates above 1 bit per sample. This is demonstrated in two dimensions by Figure 4. Sources with covariance matrices $R_x = J(\theta)^T \operatorname{diag}(1, 1/4) J(\theta)$, where $J(\theta)$ is a Jacobi rotation of θ radians,[3] were quantized with various quantization step sizes. The distortion, normalized by $\Delta^2/12$, is shown on a logarithmic scale as a function of θ. In this example, the distortion differs little from, and is bounded above, by $\Delta^2/12$.

B. Transform Update Mechanisms

Referring again to Figure 1, for decoder tracking without side information it is necessary that the transform T_{n+1} depend only on $\{T_k\}_{k=1}^n$ and $\{\hat{y}_k\}_{k=1}^n$. We assume that the covariance estimate

$$\widehat{R_{\hat{x}}}^{(n)} \;=\; \frac{1}{n} \sum_{k=1}^{n} \hat{x}_k \hat{x}_k^T \tag{5}$$

is computed and that T_{n+1} is chosen such that $T_{n+1} \widehat{R_{\hat{x}}}^{(n)} T_{n+1}^T$ is diagonal with nonincreasing diagonal elements. This amounts to using $\widehat{R_{\hat{x}}}^{(n)}$ as an estimate for R_x. The calculation of T_{n+1} will have sign ambiguities[4] and if the eigenvalues of $\widehat{R_{\hat{x}}}^{(n)}$ are not distinct, there will be additional ambiguities; these can be resolved arbitrarily. The initial transform T_1 can also be arbitrary.

[3] Jacobi rotations are defined in equation (12) of Appendix A.

[4] If $T_{n+1} \widehat{R_{\hat{x}}}^{(n)} T_{n+1}^T$ is diagonal, then negating any row of T_{n+1} will not change the product $T_{n+1} \widehat{R_{\hat{x}}}^{(n)} T_{n+1}^T$.

More complicated update mechanisms are possible, but using an eigendecomposition of (5) has the attractive property of requiring only constant storage: As the data vectors are coded, only the $N(N+1)/2$ independent components of (5) must be stored. Adjustments to (5) to compensate for quantization effects are possible, but are not used so as to not rely too heavily on the Gaussian model for the source data.

At first glance it may seem that we expect $\widehat{R_{\hat{x}}}^{(n)}$ to converge to R_x, which would result in the transform converging to the desired KLT. In fact, we do not need $\widehat{R_{\hat{x}}}^{(n)} \to R_x$ to have the desired transform convergence. Suppose for the moment that the effect of quantization is to add a zero-mean signal z independent of x with $E[zz^T] = (\Delta^2/12)I_N$. Then $R_{\hat{x}} = R_x + (\Delta^2/12)I_N$ and since $R_{\hat{x}}$ and R_x have the same eigenvectors, the transform converges to the correct transform. Of course, this is an overly simplistic model of quantization. As detailed below, the difference between $E[xx^T]$ and $E[Q(x)Q(x)^T]$ is generally not a scaled identity. Nevertheless, we assert that the system works: The transform converges to the optimal transform, resulting in a universal system. We cannot prove this convergence precisely, but results suggesting the observed convergence are given in the following section.

III. Main Results

The main results of the paper are summarized in this section. Proofs are given in Appendix B.

A. Transform convergence implies universality

Theorem 1: Fix a quantization step size Δ and suppose $\{T_n\}$ converges elementwise to T, a KLT of the source. Let L_n denote the per component code length for coding the first n vectors using the adaptive scheme and let $L_n^\star$ denote the per component code length for coding the first n vectors with the fixed, optimal transform T. Then the average excess rate $n^{-1}(L_n - L_n^\star)$ converges in mean square to zero.

As discussed in Section II-A, given a quantization step size, the distortion of a transform coder depends only slightly on the transform. Thus Theorem 1 indicates that the backward adaptive scheme will have performance asymptotically almost equal to an optimal transform coder whenever the transform converges to a KLT. Transform convergence can be established when using an independence assumption similar to that used in heuristic analyses of the LMS algorithm. In such an analysis the sequence of transforms is assumed to be independent, though this assumption is clearly false [16, App. 3.B].

The following two sections give different types of convergence results that are suggestive of the convergence seen in simulations. In Section III-B the stochastic variation of (5) is ignored. The transform updates are then described by a deterministic iteration. As an alternative, the quantizer can be replaced by a subtractive dithered quantizer in order to insure nice behavior of the transform sequence. This is considered in Section III-C.

B. Deterministic analysis

In the original system, the distribution of $\hat{x}_n$ depends on T_n, which in turn depends on T_1 and $\{x_k\}_{k=1}^{n-1}$. Because of this complicated interdependence between quantization and stochastic effects, it is very difficult to analyze the convergence of the transform.

One way to reduce the complexity of the analysis is to neglect the stochastic aspect, meaning to assume there is no variance in moment estimates despite the fact that moments are estimated from finite length observations. The effect is to replace (5) with

$$R_{\hat{x}}^{(n)} \;=\; E[\hat{x}_n \hat{x}_n^T] \tag{6}$$

and update the transform such that $T_{n+1} R_{\hat{x}}^{(n)} T_{n+1}^T$ is diagonal with nonincreasing diagonal elements. We are left with a deterministic iteration summarized by

$$R_{\hat{x}}^{(n)} \;=\; T_n^T R_{\hat{y}}^{(n)} T_n \;=\; T_n^T \tilde{Q}(R_y^{(n)}) T_n \;=\; T_n^T \tilde{Q}(T_n R_x T_n^T) T_n$$

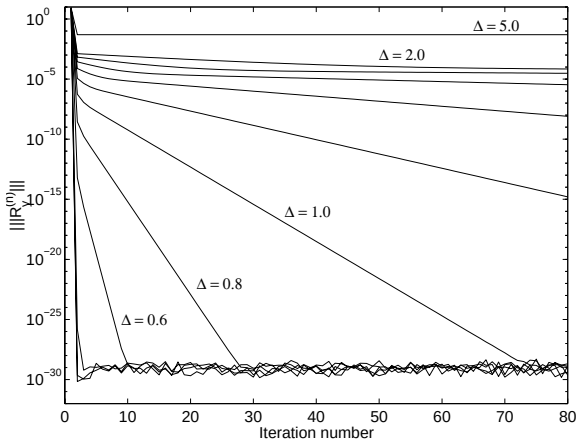
$$T_{n+1} R_{\hat{x}}^{(n)} T_{n+1}^T \;=\; \Lambda_n \text{ (diagonal with nonincreasing diagonal elements)},$$

where $\tilde{Q} : \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}$ gives the effect of quantization on the covariance matrix. $\tilde{Q}$ depends on the source distribution and Δ and can be described by evaluating expressions from [17].

Since R_x and $R_{\hat{x}}^{(n)}$ generally have different eigenvectors, it is not obvious that this iteration will converge. The following theorem gives a limited convergence result.

Theorem 2: Let R_x and T_1 be given. Then there exists a sequence of quantization step sizes $\{\Delta_n\} \subset \mathbb{R}^+$ such that the deterministic iteration described above converges to a KLT of the source. Since the KLT is ambiguous if the eigenvalues of R_x are not distinct, convergence is indicated by $R_{\hat{y}}^{(n)}$ approaching a diagonal matrix in Frobenius norm.

Theorem 2 does not preclude the possibility that the iteration will converge only with $\inf \Delta_n = 0$. However, numerical calculations suggest that the iteration actually converges for constant sequences of sufficiently small step sizes. Figure 5 shows numerical results for a four-dimensional Gaussian source with $(R_x)_{ij} = 0.9^{|i-j|}$, $T_1 = I$ and various values of Δ. To show the degree



Δ	Rate (bits/component)
0.1	4.475
0.2	3.488
0.4	2.536
0.6	2.006
0.8	1.620
1.0	1.322
1.2	1.097
1.4	0.929
1.6	0.802
1.8	0.702
2.0	0.623
5.0	0.218

Fig. 5. Simulations for various fixed quantization step sizes suggest that the deterministic iteration converges more generally than predicted by Theorem 2. The source vector length is $N = 4$ and the initial transform is the identity transform. The accompanying table provides an approximate correspondence between quantization step sizes and rates.

to which T_n diagonalizes R_x, $|||R_y^{(n)}|||$ is plotted as a function of the iteration number n, where $|||A||| = \sum_{i \neq j} a_{ij}^2$. An approximate correspondence between quantization step size and rate is also given.

Starting from an arbitrary initial transform, $|||R_y^{(n)}|||$ becomes small after a single iteration (note the logarithmic vertical axis). Then, to the limits of machine precision, it converges exponentially to zero with a rate of convergence that depends on Δ. (For $\Delta > 3$, loss of significance problems in the computation combined with very slow convergence make it difficult to ascertain convergence numerically.)

The results shown in Figure 5 are representative of the performance with an arbitrary R_x. The convergence, as measured by $|||R_y^{(n)}|||$, is unaffected by the multiplicities of the eigenvalues of R_x. The eigenspace associated with a multiple eigenvalue can be rotated arbitrarily without affecting $|||R_y^{(n)}|||$ or the decorrelation and energy compaction properties of the transform.

Theorem 3: Let $N = 2$ and let R_x be given. There exists $\Delta_{\max} > 0$ such that for any $\Delta < \Delta_{\max}$ the deterministic iteration converges, in the same sense as before, for any initial transform T_1.

C. Using dithered quantization

For the sake of analysis, let us alter the system to use subtractive dithered quantization [18], [19]. Replace the quantizer $q(\,\cdot\,)$ defined in (1) by

$$q_{\text{dither}}((y_n)_i) \;=\; q((y_n)_i + z_{ni}) - z_{ni}, \tag{7}$$

where the z_{ni}'s are independent and each is uniformly distributed on $[-\Delta/2, \Delta/2]$. We assume that the dither signal $\{z_{ni}\}_{n\in\mathbb{Z}^+,1\le i\le N}$ is somehow available at the decoder so that each component of the quantizer input can be reconstructed up to an error of magnitude $\Delta/2$. The dither signal is not used in the entropy coder.

The effect of the dither is to make the quantization error independent of the data and transform sequences. The following result is then straightforward.

Theorem 4: With the dithered quantizer (7) and any initial transform T_1,

$$\widehat{R_{\hat{x}}}^{(n)} \quad \textit{converges in mean square to} \quad R_x + \frac{\Delta^2}{12} I \quad \textit{as } n \to \infty.$$

Also, the sequence of transforms $\{T_n\}$ converges in mean square to a KLT for the source.

Although we are assuming Gaussian signals throughout, the proof of the theorem does not depend on the distribution of the source. The transform converges to a transform that maximizes coding gain for any i.i.d. source; however, for non-Gaussian sources maximizing coding gain may not be ideal.

When the source is Gaussian, the KLT is the optimal transform and the entropies of the quantized variables can be easily estimated. This leads to the following theorem:

Theorem 5: Denote the eigenvalues of R_x by λ_1, λ_2, ..., λ_N. Define L_n and $L_n^\star$ as in Theorem 1. Then the average excess rate $n^{-1}(L_n - L_n^\star)$ converges in mean square to a constant ρ. Estimating discrete entropies with (3),

$$\rho < \frac{1}{2N} \sum_{i=1}^{N} \log_2 \left(1 + \frac{\Delta^2}{12\lambda_i} \right). \tag{8}$$

The constant ρ can be interpreted as the asymptotic redundancy of the system. It is the excess rate, in bits per source component, of the adaptive system, as compared to a fixed, optimal transform code designed with knowledge of R_x. The bound (8) comes simply from the variance added by the dither signal.[5] As Δ approaches zero, the power of the dither signal vanishes and

[5]When the dither signal is known at the entropy coder, performance better than the worst case given by (8) can be expected [20].

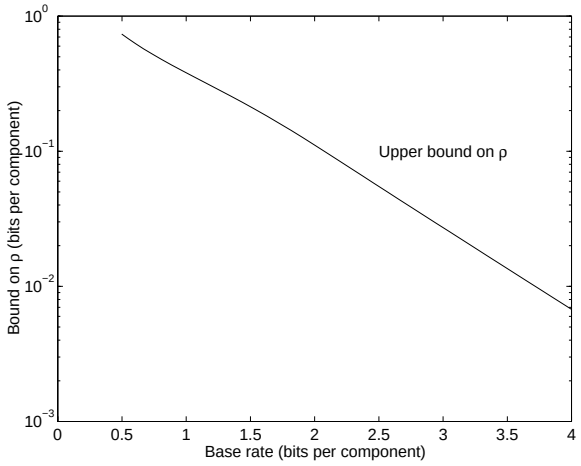


Fig. 6. Bound (8) on the excess rate ρ as a function of the coding rate for a Gaussian source with $(R_x)_{ij} = 0.8^{|i-j|}$.

accordingly ρ approaches 0. Thus the dithered system is universal for high-rate coding.

At moderate rates, ρ is quite small. For example, consider the coding of an eight-dimensional Gaussian source with $(R_x)_{ij} = 0.8^{|i-j|}$. By computing the bound (8) and the correspondence between Δ and the rate of a KLT coder for this particular source we get the curve shown in Figure 6. This roughly indicates that ρ must decay exponentially with the overall coding rate. In fact, using the high-rate approximation

$$\frac{\Delta^2}{12} \;\approx\; \frac{\pi e}{6}\left(\prod_{i=1}^{N}\lambda_i\right)^{1/N} 2^{-2R},$$

where R is the rate of the optimal transform coder, (8) can be written as

$$\rho \;<\; \frac{1}{2N}\sum_{i=1}^{N}\log_2\left(1+\frac{\Delta^2}{12\lambda_i}\right) < \frac{1}{2N\ln 2}\sum_{i=1}^{N}\frac{\Delta^2}{12\lambda_i} \approx \frac{\pi e}{12\ln 2}\left(\prod_{i=1}^{N}\lambda_i\right)^{1/N}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda_i}\right)2^{-2R}.$$

At rates of 2 or 3 bits per component, the excess rate is less than 6% or 1%, respectively.

D. Remark

The deterministic analyses and the analysis of the system with dither can be combined to form a heuristic argument for convergence. Soon after the system is initialized, the variance of (5) is high and thus the variation of the transform is also high; this has the effect of a dither. Later the changes to the transform are much smaller, but the transform cannot settle at an incorrect value because incorrect transforms are not fixed points of the deterministic iteration.

IV. Variations on the Basic Algorithms

Certain modifications to the basic algorithms can be made to reduce the computational complexity or to facilitate the coding of non-i.i.d. sources. All of the modifications mentioned in this section apply equally well to the dithered and undithered systems.

The most complicated step in these algorithms is the computation of the updated transform; thus, the complexity can be reduced by suppressing this computation. Instead of computing an eigendecomposition of $\widehat{R_{\hat{x}}}^{(n)}$ at each step, one can compute the eigendecomposition every L steps, holding the transform constant in between. L need not be constant, but if it is to vary it must be computable from coded data. Having constant $L > 1$ does not affect the conclusions in Theorem 4.

The coding of a non-i.i.d. source poses many problems. First of all, we must assume that $R_{\hat{x}}^{(n)}$ varies slowly, or that the source is "locally stationary." If this is not the case, an on-line algorithm will fail because the coding of x_n is based on an estimate of $R_{\hat{x}}^{(n)}$ from (recent) past samples.[6] Secondly, the covariance matrix estimate $\widehat{R_{\hat{x}}}^{(n)}$ should be local, *e.g.*,

$$\widehat{R_{\hat{x}}}^{(n)} \;=\; \frac{1}{K} \sum_{k=n-K+1}^{n} \hat{x}_k \hat{x}_k^T \tag{9}$$

or

$$\widehat{R_{\hat{x}}}^{(n)} \;=\; \omega \widehat{R_{\hat{x}}}^{(n-1)} + (1-\omega)\hat{x}_n \hat{x}_n^T, \tag{10}$$

with appropriate initialization. If the update interval L divides K in (9), it is not necessary to store a full window of K past samples [22].

A technique which simultaneously reduces the computational complexity and introduces a covariance estimate equivalent to (10) is to replace the eigendecomposition computation with an *incremental* change in the transform based on $\hat{x}_n$. This is explored in [16, Ch. 4],[23].

V. Conclusions

This correspondence has proposed a backward adaptive structure for transform adaptation in transform coding. Since there is no side information, the system is universal for Gaussian sources when the transform converges to a Karhunen–Loève transform. Simulations indicate convergence, and convergence can be shown under certain simplifying assumptions such as when the estimation noise is ignored or when the quantization is dithered. The problem of optimally combining forward and backward adaptive methods remains open.

[6] Without local stationarity, a forward adaptive method would presumably be superior; see [9], [21].

Gaussian sources were assumed throughout. This assumption was used in two ways: to justify maximizing coding gain and to concretely describe the effect of quantization on moment estimation. The availability of universal lossless coders is assumed, but, in contrast to [24], they are applied only to sequences of scalars. This potentially decreases the memory requirement and speeds convergence.

Appendix

I. Optimality of the Karhunen–Loève transform

This appendix provides a new result, with two proofs, on the optimality of the Karhunen–Loève transform for transform coding of Gaussian sources. It is more general than earlier results relying on optimal fixed-rate quantization [3] or high-resolution quantization theory [2], [3] because it relies only on a scale-invariance property of quantizer distortion–rate performance; in particular, it encompasses the earlier results and applies to entropy-coded uniform scalar quantization with equal step sizes for each component, as utilized in this correspondence.

Theorem 6 (Optimality of Karhunen–Loève transform) Consider the transform coding of a Gaussian source subject to MSE distortion. Assume that the distortion–rate performance of a scalar quantizer applied to a component with variance σ^2 is $D = \sigma^2 f(R)$. Then a KLT is an optimal transform, *i.e.*, for any given maximum per component rate, it minimizes the distortion.

First note that if $f(\cdot)$ is not nonincreasing, there will be rates that are useless: if $R_1 > R_2$ but $f(R_1) > f(R_2)$, rate R_1 can be replaced in any purportedly optimal solution by rate R_2 without increasing the distortion and without violating the rate constraint. Thus we henceforth assume that $f(\cdot)$ is nonincreasing.

Two proofs are given: The first is simple to describe from first principles but relies on an iterative construction. The second, more elegant proof relying on the theory of majorization (see [25]) is due to Telatar [26].

Proof 1: Let $(R_1, R_2, \ldots, R_N)$ be any bit allocation vector, *i.e.*, suppose that R_i bits are allocated to transform coefficient y_i. Given any orthogonal transform T, we will show that there exists a KLT $\widetilde{T}$ that yields distortion at most as high as yielded by T.

Before proceeding with more complicated constructions, note that the variances of the transform coefficients should have the same ordering as the rates. If $\sigma_{y_i}^2 > \sigma_{y_j}^2$ but $R_i < R_j$, then the distortion is reduced or unchanged by swapping the ith and jth rows of T. The resulting change

in distortion is

$$\Big(\sigma_{y_i}^2 f(R_j) + \sigma_{y_j}^2 f(R_i)\Big) - \Big(\sigma_{y_i}^2 f(R_i) + \sigma_{y_j}^2 f(R_j)\Big) = \underbrace{(\sigma_{y_i}^2 - \sigma_{y_j}^2)}_{>0}\underbrace{(f(R_j) - f(R_i))}_{\leq 0} \leq 0.$$

In the remainder of the proof we assume that T has the property:

$$\text{for any } i \text{ and } j, \qquad \sigma_{y_i}^2 > \sigma_{y_j}^2 \quad \text{implies} \quad R_i \geq R_j. \tag{11}$$

There is nothing to prove if T is a KLT, so we may assume that the (i,j) component of $R_y = TR_xT^T$ is nonzero for some (i,j) pair. Construct a new transform $T_1 = J(i,j,\theta)^T T$, where $J(i,j,\theta)$ is a Jacobi rotation defined by

$$J(i,j,\theta) \;=\; \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \cos\theta & \cdots & \sin\theta & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -\sin\theta & \cdots & \cos\theta & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \begin{matrix} \\ \\ i \\ \\ j \\ \\ \\ \end{matrix} \quad , \qquad \theta \in [-\pi/4,\, \pi/4), \tag{12}$$

$$\begin{matrix} & & i & & j & & \end{matrix}$$

and θ is chosen such that the (i,j) element of $T_1 R_x T_1^T$ is zero.

This choice of transform has a few important properties. The first is that $T_1 R_x T_1^T$ is closer to a diagonal matrix than TR_xT^T, where closeness is measured by the Euclidean norm of the off-diagonal elements. Thus repeatedly cycling through all (i,j) pairs, defining $T_{k+1} = J^T T_k$, eventually yields a diagonal matrix $\widetilde{T} R_x \widetilde{T}^T$, where $\widetilde{T} = \lim_{k\to\infty} T_k$.[7]

The second property is that among the diagonal elements, only the ith and jth are altered. These are altered such that the larger of the two is increased by a positive increment δ and the smaller is decreased by the same amount. This is easily verified by expanding

$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}^T \begin{bmatrix} a_1 & a_3 \\ a_3 & a_2 \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} b_1 & 0 \\ 0 & b_2 \end{bmatrix}$$

[7]This is the well-known classical Jacobi algorithm for computing eigendecompositions of symmetric matrices; for details, including convergence results, see [27, §8.5].

and solving for $\theta \in [-\pi/4, \pi/4)$. If $a_1 \geq a_2$, one finds

$$\begin{aligned} b_1 &= a_1 + \delta \\ b_2 &= a_2 - \delta \end{aligned} \qquad \text{where} \quad \delta = \frac{-(a_1 - a_2) + \sqrt{(a_1 - a_2)^2 + 4a_3^2}}{2} \geq 0,$$

with equality if and only if $a_3 = 0$.

Suppose $\sigma_{y_i}^2 \geq \sigma_{y_j}^2$. Then using the second property, the change in distortion by using $J(i, j, \theta)^T T$ in place of T is

$$\Big((\sigma_{y_i}^2 + \delta) f(R_i) + (\sigma_{y_j}^2 - \delta) f(R_j)\Big) - \Big(\sigma_{y_i}^2 f(R_i) + \sigma_{y_j}^2 f(R_j)\Big) = \underbrace{\delta}_{>0} \underbrace{(f(R_i) - f(R_j))}_{\leq 0} \leq 0.$$

Thus as we iterate to find $\widetilde{T}$, the distortion is decreased or unchanged at each step. The $\widetilde{T}$ thusly constructed is both a KLT and at least as good as T in distortion–rate performance. ■

Proof 2: The second proof is based on elementary properties of majorization, which are detailed in [25]. A vector $(\alpha_1, \alpha_2, \ldots, \alpha_N)$ is said to be *majorized* by another vector $(\beta_1, \beta_2, \ldots, \beta_N)$ if

$$\sum_{i=1}^{k} \alpha_{[i]} \leq \sum_{i=1}^{k} \beta_{[i]}, \qquad k = 1, 2, \ldots, N-1$$

and

$$\sum_{i=1}^{N} \alpha_{[i]} = \sum_{i=1}^{N} \beta_{[i]},$$

where the $[\cdot]$ notation indicates a decreasing ordering $\alpha_{[1]} \geq \alpha_{[2]} \geq \cdots \geq \alpha_{[N]}$.

Again let $(R_1, R_2, \ldots, R_N)$ be any bit allocation vector. The problem is to minimize the function $D = \sum_{i=1}^{N} \sigma_{y_i}^2 f(R_i)$ by manipulating the $\sigma_{y_i}^2$'s through the choice of T. Let $\sigma = (\sigma_{y_1}^2, \sigma_{y_2}^2, \ldots, \sigma_{y_N}^2) = \mathrm{diag}(T R_x T^T)$. For a Hermetian matrix, the diagonal elements are majorized by the eigenvalues, so σ is majorized by a vector λ of eigenvalues of R_x. Now the majorization of σ by λ is equivalent to σ being in the convex hull of the $N!$ permutations of λ. We are thus left with minimizing D over the convex polytope defined by the permutations of λ.[8] In minimizing a linear function over a convex polytope, the optimum is always attained at a corner point. This establishes that the optimal transform is a KLT. Furthermore, the arguments given in Proof 1 indicate that the optimal KLT (the optimal permutation) is one that satisfies (11). ■

[8]We have not carefully argued that all points in the polytope are feasible, but the achievability of the optimizing value will be clear.

II. Proofs

A. Proof of Theorem 1

Let $N\ell_n$ denote the number of bits used to code x_n. Because of the convergence of the sequence of transforms and the universality of the entropy coder, $E[\ell_n]$ converges to some limit, say $\bar{\ell}$. For the static coder, the number of bits used by the optimal coder $N\ell_n^\star$ will satisfy $E[\ell_n^\star] = \bar{\ell}^\star$. Since $\{T_n\}$ converges to T and the entropy rate of the quantizer output depends only on the transform, $\bar{\ell} = \bar{\ell}^\star$. Now since $\{E[\ell_k - \ell_k^\star]\}$ converges to zero, the mean of the sequence $n^{-1}\sum_{k=1}^n(\ell_k - \ell_k^\star) = n^{-1}(L_n - L_n^\star)$ converges to zero in mean square.

B. Proof of Theorem 2

The proofs of Theorems 2 and 3 rely on properties of $\tilde{Q}$, the function that describes the effects of quantization on the covariance matrix.

Let η_1 and η_2 be jointly Gaussian with $E[\eta_1] = E[\eta_2] = 0$, $E[\eta_1^2] = \nu_1^2$, $E[\eta_2^2] = \nu_2^2$, and $E[\eta_1\eta_2] = \nu_{12}$. Define $\hat{\nu}_1^2 = E[q(\eta_1)^2]$, $\hat{\nu}_2^2 = E[q(\eta_2)^2]$, and $\hat{\nu}_{12} = E[q(\eta_1)q(\eta_2)]$, where $q(\,\cdot\,)$ was defined by (1). Then using expressions from [17], one can show

$$\hat{\nu}_i^2 = \nu_i^2 + \frac{\Delta^2}{12} + \sum_{m=1}^{\infty}(-1)^m e^{-2m^2\pi^2\nu_i^2/\Delta^2}\left(\frac{\Delta^2}{m^2\pi^2} + 4\nu_i^2\right), \qquad i = 1,\, 2, \tag{13}$$

and

$$\hat{\nu}_{12} = (1+\delta)\nu_{12} + \mu, \tag{14}$$

where

$$\delta = 2\left(\sum_{m_1=1}^{\infty}(-1)^{m_1}e^{-2m_1^2\pi^2\nu_1^2/\Delta^2} + \sum_{m_2=1}^{\infty}(-1)^{m_2}e^{-2m_2^2\pi^2\nu_2^2/\Delta^2}\right)$$

and

$$\mu = \sum_{m_1,m_2=1}^{\infty}(-1)^{m_1+m_2}\frac{\Delta^2}{m_1m_2\pi^2}\exp\left(\frac{-2\pi^2(m_1^2\nu_1^2 + m_2^2\nu_2^2)}{\Delta^2}\right)\sinh\left(\frac{4\pi^2\nu_{12}m_1m_2}{\Delta^2}\right). \tag{15}$$

For any covariance matrix R, the diagonal elements of $\tilde{Q}(R)$ are described by (13) and the off-diagonal elements are described by (14). For the purpose of this theorem, we need only the following simple property of $\tilde{Q}$:

$$\tilde{Q}(R) = R + \frac{\Delta^2}{12}I + C, \;\text{ where }\; C \to 0 \;\text{ elementwise as }\; \Delta \to 0. \tag{16}$$

To measure the degree to which T_n diagonalizes R_x, define a distance measure $|||\cdot|||$ between a matrix A and the set of diagonal matrices by $|||A||| = \sum_{i\neq j} a_{ij}^2$. The strategy of the proof is to show that for sufficiently small Δ, the inequality $|||R_y^{(n+1)}||| \leq \frac{1}{2}|||R_y^{(n)}|||$ holds for all $n \geq 1$.

Combining $R_{\hat{x}}^{(n)} = T_n^T R_{\hat{y}}^{(n)} T_n$ with $R_{\hat{x}}^{(n)} = T_{n+1}^T \Lambda_n T_{n+1}$ gives $T_{n+1}^T \Lambda_n T_{n+1} = T_n^T R_{\hat{y}}^{(n)} T_n$. Define $H_n = T_n T_{n+1}^T$ so that

$$R_{\hat{y}}^{(n)} = H_n \Lambda_n H_n^T. \tag{17}$$

Also notice that

$$R_y^{(n+1)} = T_{n+1} R_x T_{n+1}^T = T_{n+1} T_n^T T_n R_x T_n^T T_n T_{n+1}^T = H_n^T R_y^{(n)} H_n.$$

As a final preparation, define $Z_n = R_y^{(n)} - R_{\hat{y}}^{(n)}$.

We can now make the calculation

$$|||R_y^{(n+1)}||| = |||H_n^T R_y^{(n)} H_n||| = |||H_n^T (Z_n + R_{\hat{y}}^{(n)}) H_n||| = |||H_n^T Z_n H_n|||,$$

where the last equality follows from $H_n^T R_{\hat{y}}^{(n)} H_n$ being diagonal (see (17)). From (16), it is clear that if Δ is small enough, $|||Z_n||| \leq \frac{1}{4}|||R_y^{(n)}|||$. It remains now to relate $|||Z_n|||$ and $|||H_n^T Z_n H_n|||$.

Substitute $R_{\hat{y}}^{(n)} = R_y^{(n)} + \Delta^2 I/12 + C_1$, where $\|C_1\| \to 0$ as $\Delta \to 0$, in (17) to get

$$R_y^{(n)} + \frac{\Delta^2}{12} I + C_1 = H_n \Lambda_n H_n^T. \tag{18}$$

Decrementing the index and rearranging gives

$$H_{n-1}^T R_y^{(n-1)} H_{n-1} + \frac{\Delta^2}{12} I + H_{n-1}^T C_1 H_{n-1} = \Lambda_{n-1}. \tag{19}$$

Since $H_{n-1}^T R_y^{(n-1)} H_{n-1} = R_y^{(n)}$, comparing (18) and (19) gives

$$H_n \Lambda_n H_n^T = \Lambda_{n-1} + C_1 - H_{n-1}^T C_1 H_{n-1}. \tag{20}$$

Now let $C_2 = H_n - I$. Substituting in (20) and expanding, we conclude that $\|C_2\| \to 0$ as $\Delta \to 0$. Thus by expanding $H_n^T Z_n H_n$ we see that $|||H_n^T Z_n H_n||| - |||Z_n||| \to 0$ faster than $|||Z_n||| \to 0$ as $\Delta \to 0$, so by choosing Δ small enough we have the bound $|||H_n^T Z_n H_n||| \leq 2|||Z_n|||$.

Combining all these calculations gives

$$|||R_y^{(n+1)}||| = |||H_n^T Z_n H_n||| \leq 2|||Z_n||| \leq 2 \cdot \frac{1}{4}|||R_y^{(n)}||| = \frac{1}{2}|||R_y^{(n)}|||.$$

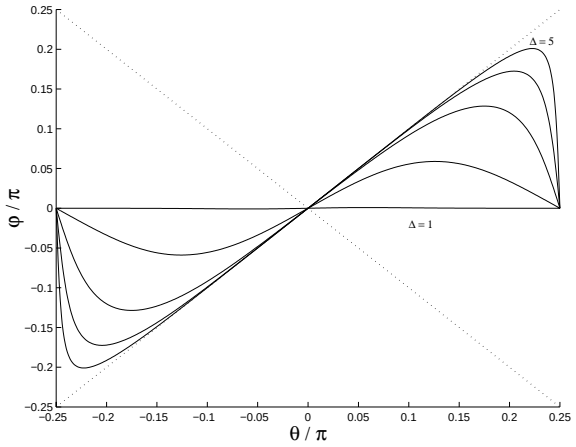


Fig. 7. Simulations of the deterministic iteration for $N = 2$ suggest convergence for any quantization step size Δ. The eigenvalues of R_x are 1 and 1/4. The step sizes shown are $\Delta = 1, 2, \ldots, 5$. φ is the iterate that follows after θ. Global convergence is indicated by the curves lying inside the cone $|\varphi| \leq |\theta|$, which is marked by dotted lines.

C. Proof of Theorem 3

Without loss of generality (rotating the coordinate system and initial transform, if necessary), assume $R_x = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$, $\sigma_1 \geq \sigma_2$. The transform iterates are all in $SO_2(\mathbb{R})$ and can be parameterized as

$$T_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix},$$

where $\theta \in [-\pi/4, \pi/4]$. We assume $\sigma_1 > \sigma_2$; if $\sigma_1 = \sigma_2$ the situation is uninteresting because R_y is diagonal for any T_θ.

Denote the transform iterate that follows after θ by φ. The proof will be completed by showing that there is a constant $\Delta_{\max}$, independent of θ, such that $\Delta \leq \Delta_{\max}$ implies $\sin^2 2\varphi \leq \sin^2 2\theta$ with equality only when $\theta = 0$. This will show global convergence to the fixed point zero, which is an optimal transform. As a preview of this result—and to motivate the rest of the proof—we compute and plot the "next iterate" map $\theta \mapsto \varphi$. Figure 7 shows the map $\theta \mapsto \varphi$ when $\sigma_1 = 1$ and $\sigma_2 = 1/2$ for $\Delta = 1, 2, \ldots, 5$. The iteration globally converges as long as the graph of $\varphi(\theta)$ lies inside the cone $|\varphi| \leq |\theta|$. From the plot, it seems this may be true for any Δ; we endeavor to show this for Δ less than some $\Delta_{\max}$.

The first step is to relate φ to θ. By looking at the general form of $T_\theta R_{\hat{x}} T_\theta^T$, one can show

that

$$\varphi = \frac{1}{2}\arctan\left(\frac{-2(R_{\hat{x}})_{1,2}}{(R_{\hat{x}})_{1,1} - (R_{\hat{x}})_{2,2}}\right). \tag{21}$$

$R_{\hat{x}}$ is related to θ through R_y and $R_{\hat{y}}$:

$$R_y = T_\theta R_x T_\theta^T = \begin{bmatrix} \sigma_1^2\cos^2\theta + \sigma_2^2\sin^2\theta & \frac{1}{2}(\sigma_1^2-\sigma_2^2)\sin 2\theta \\ \frac{1}{2}(\sigma_1^2-\sigma_2^2)\sin 2\theta & \sigma_1^2\sin^2\theta + \sigma_2^2\cos^2\theta \end{bmatrix} = \begin{bmatrix} \nu_1^2 & \nu_{12} \\ \nu_{12} & \nu_2^2 \end{bmatrix}, \tag{22}$$

$$R_{\hat{y}} = \tilde{Q}(R_y) = R_y + \frac{\Delta^2}{12}I + \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix},$$

where α, β, and γ depend on θ, σ_1, and σ_2 as given by (13), (14), and (22). Now, after computing $R_{\hat{x}} = T_\theta^T R_{\hat{y}} T_\theta$, one finds

$$(R_{\hat{x}})_{1,1} - (R_{\hat{x}})_{2,2} = \sigma_1^2 - \sigma_2^2 + (\alpha-\gamma)\cos 2\theta + 2\beta\sin 2\theta$$

and

$$(R_{\hat{x}})_{1,2} = \frac{1}{2}(\gamma-\alpha)\sin 2\theta + \beta\cos 2\theta. \tag{23}$$

Since $\sin^2(\arctan\phi) = \phi^2/(1+\phi^2)$,

$$\begin{aligned} \sin^2 2\varphi &= \frac{\left(\frac{-2(R_{\hat{x}})_{1,2}}{(R_{\hat{x}})_{1,1}-(R_{\hat{x}})_{2,2}}\right)^2}{1+\left(\frac{-2(R_{\hat{x}})_{1,2}}{(R_{\hat{x}})_{1,1}-(R_{\hat{x}})_{2,2}}\right)^2} = \frac{(-2(R_{\hat{x}})_{1,2})^2}{((R_{\hat{x}})_{1,1}-(R_{\hat{x}})_{2,2})^2 + (-2(R_{\hat{x}})_{1,2})^2} \\ &= \frac{[(\alpha-\gamma)^2\sin 2\theta - 2\beta\cos 2\theta]^2}{(\sigma_1^2-\sigma_2^2)^2 + 2(\sigma_1^2-\sigma_2^2)[(\alpha-\gamma)^2\cos 2\theta - 2\beta\sin 2\theta] + (\alpha-\gamma)^2 + 4\beta^2} \\ &\le \frac{[(\alpha-\gamma)^2\sin 2\theta - 2\beta\cos 2\theta]^2}{[\sigma_1^2-\sigma_2^2-\sqrt{(\alpha-\gamma)^2+4\beta^2}]^2}, \end{aligned} \tag{24}$$

where (24) follows from minimizing the denominator over θ. The following three lemmas allow us to complete the bounding of $\sin^2 2\varphi$.

Lemma 1: $|\alpha-\gamma| < c_1(\Delta,\sigma_1,\sigma_2)$ *uniformly in* θ, *with* $c_1(\Delta,\sigma_1,\sigma_2)\to 0$ *as* $\Delta\to 0$.

Proof: The series in (13) is an alternating series with terms that monotonically decrease in absolute value. Thus it can be bounded (with appropriate sign) by any partial sum [28]. Using simply the first term,

$$-e^{-2\pi^2(R_{\hat{y}})_{1,1}/\Delta^2}\left(\frac{\Delta^2}{\pi^2} + 4(R_{\hat{y}})_{1,1}\right) < \alpha < 0,$$

and similarly for γ. Finally,

$$|\alpha-\gamma| \le \max\{|\alpha|,|\gamma|\} < e^{-2\pi^2\sigma_2^2/\Delta^2}\left(\frac{\Delta^2}{\pi^2} + 4\sigma_1^2\right) = c_1(\Delta,\sigma_1,\sigma_2),$$

since α and γ have the same sign and $\sigma_1 < \sigma_2$. ■

Lemma 2: $|\beta| \le c_2(\Delta, \sigma_1, \sigma_2)|\sin 2\theta|$ *uniformly in* θ, *with* $c_2(\Delta, \sigma_1, \sigma_2) \to 0$ *as* $\Delta \to 0$.

Proof: Rearranging (14) gives $\beta = \delta(R_{\hat{y}})_{1,2} + \mu$, where the definitions of δ and μ must use $(R_{\hat{y}})_{1,1}$, $(R_{\hat{y}})_{2,2}$, and $(R_{\hat{y}})_{1,2}$ in place of ν_1^2, ν_2^2, and ν_{12}. As in the proof of Lemma 1, δ can be bounded by using the first term in each series:

$$|\delta| < 2\left(e^{-2\pi^2(R_{\hat{y}})_{1,1}/\Delta^2} + e^{-2\pi^2(R_{\hat{y}})_{2,2}/\Delta^2}\right) < 4e^{-2\pi^2\sigma_2^2/\Delta^2}. \tag{25}$$

Assume for the moment that the absolute value of the summand of (15) decreases monotonically with both m_1 and m_2. Then computing the double summation (15) in either order gives alternating series, so the same bounding technique can be used. We get

$$\begin{aligned}
|\mu| &\le \frac{\Delta^2}{\pi^2}\exp\left(\frac{-2\pi^2((R_{\hat{y}})_{1,1}+(R_{\hat{y}})_{2,2})}{\Delta^2}\right)\sinh\left(\frac{4\pi^2(R_{\hat{y}})_{1,2}}{\Delta^2}\right)\\
&= \frac{\Delta^2}{\pi^2}\exp\left(\frac{-2\pi^2(\sigma_1^2+\sigma_2^2)}{\Delta^2}\right)\sinh\left(\frac{2\pi^2(\sigma_1^2-\sigma_2^2)\sin 2\theta}{\Delta^2}\right)\\
&\le \frac{\Delta^2}{\pi^2}\exp\left(\frac{-2\pi^2(\sigma_1^2+\sigma_2^2)}{\Delta^2}\right)\sinh\left(\frac{2\pi^2(\sigma_1^2-\sigma_2^2)}{\Delta^2}\right)\sin 2\theta\\
&\le \frac{\Delta^2}{2\pi^2}e^{-2\pi^2\sigma_2^2/\Delta^2}\sin 2\theta.
\end{aligned} \tag{26}$$

Combining (25) and (26) gives

$$\begin{aligned}
|\beta| &= |\delta(R_{\hat{y}})_{1,2} + \mu| = \left|\frac{1}{2}\delta(\sigma_1^2-\sigma_2^2)\sin 2\theta\right|\\
&\le \frac{1}{2}(\sigma_1^2-\sigma_2^2)|\delta||\sin 2\theta| + |\mu|\\
&< \underbrace{\left(2(\sigma_1^2-\sigma_2^2)+\frac{\Delta^2}{2\pi^2}\right)e^{-2\pi^2\sigma_2^2/\Delta^2}}_{c_2(\Delta,\sigma_1,\sigma_2)}|\sin 2\theta|.
\end{aligned}$$

In general, the terms of (15) are not monotonically decreasing. However, the terms are monotonically decreasing (in absolute value) outside of $(m_1, m_2) \in \{1, 2, \ldots, M\}^2$ for some $M < \infty$. Since each individual term for $(m_1, m_2) \in \{1, 2, \ldots, M\}^2$ can be bounded as above, the bound can be extended to the general case. ■

Lemma 3: $|\beta| < c_2(\Delta, \sigma_1, \sigma_2)$ *uniformly in* θ, *with* $c_2(\Delta, \sigma_1, \sigma_2) \to 0$ *as* $\Delta \to 0$.

Proof: This follows immediately from Lemma 2. ■

By combining Lemmas 1 and 3, there exists $\Delta_1 > 0$ such that $\Delta < \Delta_1$ implies $(\alpha-\gamma)^2+4\beta^2 \le (\sigma_1^2-\sigma_2^2)^2/4$, uniformly in θ. Thus assuming $\Delta < \Delta_1$ we have

$$\sin^2 2\varphi \le \frac{[(\alpha-\gamma)^2\sin 2\theta - 2\beta\cos 2\theta]^2}{\frac{1}{4}(\sigma_1^2-\sigma_2^2)^2}.$$

Applying Lemmas 1 and 2,

$$\sin^2 2\varphi \leq (c_1 + 2c_2)^2 \sin^2 2\theta$$

and there exists $\Delta_2 > 0$ such that $\Delta < \Delta_2$ implies $(c_1 + 2c_2)^2 < 1$. The proof is complete with $\Delta_{\max} = \min\{\Delta_1, \Delta_2\}$.

The bounds in this theorem are rather complicated but we can check that the requirements on Δ are reasonable. Suppose $\sigma_1 = 1$ and $\sigma_2 = 1/2$. Then $\Delta_1 > 1.366$ and $\Delta_2 > 1.565$, so the theorem guarantees convergence for any $\Delta < 1.366$. (For this range of Δ, (15) can be bounded by the $m_1 = m_2 = 1$ term for any θ.) As we found for Theorem 2, numerical calculations suggest convergence for any Δ (see Figure 7).

D. Proof of Theorem 4

The mean square convergence of $\widehat{R_{\hat{x}}}^{(n)}$ follows from the Chebyshev law of large numbers [29] once we establish that each term of (5) has common expected value $R_x + \Delta^2 I/12$, has finite variance, and is elementwise uncorrelated with every other term. The second conclusion follows easily.

First note that

$$\hat{x}_k = T_k^T \hat{y}_k = T_k^T (y_k + (\hat{y}_k - y_k)) = x_k + T_k^T (\hat{y}_k - y_k).$$

Because of the use of subtractive dither, $\hat{y}_k - y_k$ is uniformly distributed on the hypercube $[-\Delta/2, \Delta/2]^N$ and independent of x_k and T_k [18], [19]. (The overall error $\hat{x}_k - x_k$ is uniformly distributed on a rotated hypercube, independent of x_k but not independent of T_k. Its components are uncorrelated but not independent.) Now any term of (5) can be expanded as

$$\hat{x}_k \hat{x}_k^T = x_k x_k^T + x_k(\hat{y}_k^T - y_k^T)T_k + T_k^T(\hat{y}_k - y_k)x_k^T + T_k^T(\hat{y}_k - y_k)(\hat{y}_k^T - y_k^T)T_k. \tag{27}$$

Since $\hat{y}_k - y_k$ depends on T_k but x_k and T_k are independent, computing the expectation of $\hat{x}_k \hat{x}_k^T$ is simplified by first conditioning on T_k:

$$\begin{aligned} E[\hat{x}_k \hat{x}_k^T \mid T_k] &= E[x_k x_k^T \mid T_k] + E[x_k(\hat{y}_k^T - y_k^T)T_k \mid T_k] + E[T_k^T(\hat{y}_k - y_k)x_k^T \mid T_k] \\ &\quad + E[T_k^T(\hat{y}_k - y_k)(\hat{y}_k^T - y_k^T)T_k \mid T_k] \\ &= R_x + 0 + 0 + T_k^T \frac{\Delta^2}{12} I T_k \\ &= R_x + \frac{\Delta^2}{12} I, \end{aligned}$$

where we have used the independence of x_k and $\hat{y}_k - y_k$ and the fact that each has mean zero.

The (i,j) element of $\widehat{R_{\hat{x}}}^{(n)}$ is the average of n random observations of $(\hat{x}_k \hat{x}_k^T)_{ij}$, which we denote $A_{ij}^{(k)}$. The calculation above shows that each $A_{ij}^{(k)}$ has mean $(R_x)_{ij} + \Delta^2 \delta_{ij}/12$. It can furthermore be shown that each $A_{ij}^{(k)}$ has variance bounded by a constant and that $A_{ij}^{(k)}$ is uncorrelated with $A_{ij}^{(\ell)}$ for $k \neq \ell$ [16, App. 3.C]. Thus by the Chebyshev law of large numbers we have that $\widehat{R_{\hat{x}}}^{(n)} \to R_x + \Delta^2 I/12$ elementwise in mean square. The second conclusion follows from the fact that R_x and $R_x + \Delta^2 I/12$ have the same eigenvectors.

Note that the dither is essential to the proof because it makes the quality of the estimate $\widehat{R_{\hat{x}}}^{(n)}$ independent of the sequence of transforms.

E. Proof of Theorem 5

The convergence of $n^{-1}(L_n - L_n^\star)$ to a constant follows by mimicking the proof of Theorem 1. In this case, the constant ρ is not zero because the entropy rate of the quantizer output depends not only on the transform but on the dither. It remains to estimate ρ.

Using (3) and the differential entropy of a Gaussian random variable

$$h(\mathcal{N}(0,\sigma^2)) = \frac{1}{2}\log_2 2\pi e\sigma^2 \text{ bits},$$

the entropy of a Gaussian random variable with variance σ^2, uniformly quantized with bin width Δ, is approximately $2^{-1}\log_2 \Delta^{-2} 2\pi e$ bits. Thus the rate of the static optimal system is approximately

$$R_{\text{opt}} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\log_2 \frac{2\pi e \lambda_i}{\Delta^2} \quad \text{bits/component.} \tag{28}$$

The adaptive scheme converges to an optimal transform. However, because of the dithering, the signal at the input to the quantizer is not Gaussian and does not have component variances equal to the λ_i's. Since $\{z_n\}$ is independent of $\{y_n\}$, the variances simply add, giving $\lambda_i + \Delta^2/12$, $i = 1, 2, \ldots, N$. Since a Gaussian p.d.f. has the largest differential entropy for a given variance, the asymptotic rate of the universal coder can be bounded as

$$R_{\text{univ}} < \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\log_2 \frac{2\pi e(\lambda_i + \Delta^2/12)}{\Delta^2}. \tag{29}$$

Subtracting (28) from (29) and pairing terms gives (8).

References

[1] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Th.*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[2] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Acad. Pub., Boston, MA, 1992.

[3] J. J. Y. Huang and P. M. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Comm. Syst.*, vol. 11, pp. 289–296, Sept. 1963.

[4] J. D. Gibson, "Adaptive prediction in speech differential encoding systems," *Proc. IEEE*, vol. 68, no. 4, pp. 488–525, Apr. 1980.

[5] A. Gersho, "Advances in speech and audio compression," *Proc. IEEE*, vol. 82, no. 6, pp. 900–918, June 1994.

[6] R. V. Cox, "Speech coding," in *The Digital Signal Processing Handbook*, chapter 45, pp. 45.1–45.19. CRC and IEEE Press, 1998.

[7] S. LoPresto, K. Ramchandran, and M. T. Orchard, "Image coding based on mixture modeling of wavelet coefficients and a fast estimation–quantization framework," in *Proc. IEEE Data Compression Conf.*, J. A. Storer and M. Cohn, Eds., Snowbird, Utah, Mar. 1997, pp. 221–230, IEEE Comp. Soc. Press.

[8] A. Ortega and M. Vetterli, "Adaptive scalar quantization without side information," *IEEE Trans. Image Proc.*, vol. 6, no. 5, pp. 665–676, May 1997.

[9] M. Effros and P. A. Chou, "Weighted universal transform coding: Universal image compression with the Karhunen-Loève transform," in *Proc. IEEE Int. Conf. Image Proc.*, Washington, DC, Oct. 1995, vol. II, pp. 61–64.

[10] Z. Zhang and V. K. Wei, "An on-line universal lossy data compression algorithm via continuous codebook refinement—Part I: Basic results," *IEEE Trans. Inform. Th.*, vol. 42, no. 3, pp. 803–821, May 1996.

[11] R. D. Dony and S. Haykin, "Optimally adaptive transform coding," *IEEE Trans. Image Proc.*, vol. 4, no. 10, pp. 1358–1370, Oct. 1995.

[12] H. P. Kramer and M. V. Mathews, "A linear coding for transmitting a set of correlated signals," *IRE Trans. Inform. Th.*, vol. 23, no. 3, pp. 41–46, Sept. 1956.

[13] H. Gish and J. P. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inform. Th.*, vol. IT-14, no. 5, pp. 676–683, Sept. 1968.

[14] R. C. Wood, "On optimum quantization," *IEEE Trans. Inform. Th.*, vol. IT-15, no. 2, pp. 248–252, Mar. 1969.

[15] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Math. Acad. Sci. Hungar.*, vol. 10, pp. 193–215, 1959.

[16] V. K Goyal, *Beyond Traditional Transform Coding*, Ph.D. thesis, Univ. California, Berkeley, 1998, Published as Univ. California, Berkeley, Electron. Res. Lab. Memo. No. UCB/ERL M99/2, Jan. 1999. Available on-line at `http://cm.bell-labs.com/who/vivek/Thesis/`.

[17] L. Cheded and P. A. Payne, "The exact impact of amplitude quantization on multi-dimensional, high-order moments estimation," *Signal Proc.*, vol. 39, no. 3, pp. 293–315, Sept. 1994.

[18] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical survey," *J. Audio Eng. Soc.*, vol. 40, no. 5, pp. 355–375, May 1992.

[19] R. M. Gray and T. G. Stockham, Jr., "Dithered quantizers," *IEEE Trans. Inform. Th.*, vol. 39, no. 3, pp. 805–812, May 1993.

[20] D. W. E. Schobben, R. A. Beuker, and W. Oomen, "Dither and data compression," *IEEE Trans. Signal Proc.*, vol. 45, no. 8, pp. 2097–2101, Aug. 1997.

[21] V. K Goyal, M. Vetterli, and N. T. Thao, "Quantized overcomplete expansions in $\mathbb{R}^N$: Analysis, synthesis, and algorithms," *IEEE Trans. Inform. Th.*, vol. 44, no. 1, pp. 16–31, Jan. 1998.

[22] V. K Goyal, J. Zhuang, and M. Vetterli, "Universal transform coding based on backward adaptation," in *Proc. IEEE Data Compression Conf.*, J. A. Storer and M. Cohn, Eds., Snowbird, Utah, Mar. 1997, pp. 231–240, IEEE Comp. Soc. Press.

[23] V. K Goyal and M. Vetterli, "New gradient algorithms for Karhunen–Loève basis tracking," *IEEE Trans. Signal Proc.*, submitted Feb. 1999.

[24] J. Ziv, "On universal quantization," *IEEE Trans. Inform. Th.*, vol. IT-31, no. 3, pp. 344–347, May 1985.

[25] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorizations and Its Applications*, Academic Press, San Diego, 1979.

[26] I. E. Telatar, personal communication, July 1999.

[27] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore, MD, second edition, 1989.

[28] T. M. Apostol, *Mathematical Analysis*, Addison-Wesley, second edition, 1974.

[29] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 1965.

Vivek K Goyal (S'92–M'98) was born in Waterloo, Iowa, in 1971. He received the B.S. degree in mathematics and the B.S.E. in electrical engineering (both with highest distinction), in 1993, from the University of Iowa, Iowa City. He received the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1995 and 1998, respectively.

He was a Research Assistant in the Laboratoire de Communications Audiovisuelles at École Polytechnique Fédérale de Lausanne, Switzerland, in 1996. He worked in the Mathematics of Communications Department at Lucent Technologies' Bell Laboratories in 1997, where since 1998 he has been a Member of Technical Staff. His research interests include source coding theory, quantization theory, practical compression, and computational complexity.

Dr. Goyal is a member of Phi Beta Kappa, Tau Beta Pi, Sigma Xi, Eta Kappa Nu and SIAM. In 1998 he received the Eli Jury Award of the University of California, Berkeley, awarded to a graduate student or recent alumnus for outstanding achievement in systems, communications, control, or signal processing.

Jun Zhuang was born in Chang Sha, People's Republic of China, in 1971. He received the B.E. degree from TsingHua University, BeiJing, China in 1994, and the M.Sc. degree from University of California at Berkeley, both in Electrical Engineering.

From September 1996 to May 1998, he was with the Broadband Strategy & Engineering department at Pacific Bell, San Ramon, CA. Since May 1998, he has been a senior member of technical staff at SBC Technology Resources Inc., Pleasanton, CA. His current interests are in the areas of network design, simulation and new IP services.

Martin Vetterli received the Dipl. El.-Ing. degree from ETH Zürich (ETHZ), Switzerland, in 1981, the MS degree from Stanford University in 1982, and the Doctorat ès Science degree from EPF Lausanne (EPFL), Switzerland, in 1986.

He was a Research Assistant at Stanford and EPFL, and has worked for Siemens and AT&T Bell Laboratories. In 1986, he joined Columbia University in New York where he was last an Associate Professor of Electrical Engineering and co-director of the Image and Advanced Television Laboratory. In 1993, he joined the University of California at Berkeley where he was a Professor in the Dept. of Electrical Engineering and Computer Sciences until 1997, and holds now Adjunct Professor position. Since 1995, he is a Professor of Communication Systems at EPF Lausanne, Switzerland, where he chaired the Communications Systems Division (1996/97), and heads of the Audio-Visual Communications Laboratory. He held visiting positions at ETHZ (1990) and Stanford (1998).

He is a fellow of the IEEE, a member of SIAM, and was the Area Editor for Speech, Image, Video, and Signal Processing of the IEEE Transactions on Communications. He is also on the editorial boards of Annals of Telecommunications, Applied and Computational Harmonic Analysis and The Journal of Fourier Analysis and Applications.

He received the Best Paper Award of EURASIP in 1984 for his paper on multidimensional subband coding, the Research Prize of the Brown Bovery Corporation (Switzerland) in 1986 for his doctoral thesis, the IEEE Signal Processing Society's Senior Award in 1991 and in 1996 (for papers with D. LeGall and K. Ramchandran, respectively). and he is a IEEE Signal Processing Distinguished lecturer in 1999. He received the Swiss National Latsis Prize in 1996 and the SPIE Presidential award in 1999.

He was a plenary speaker at various conferences (e.g. 1992 IEEE ICASSP) and is the co-author, with J. Kovačević, of the book **Wavelets and Subband Coding** (Prentice-Hall, 1995). He has published about 75 journal papers on a variety of topics in signal and image processing and holds 5 patents.

His research interests include wavelets, multirate signal processing, computational complexity, signal processing for telecommunications, digital video processing and compression and wireless video communications.