# Optimal Correlating Transform for Erasure Channels

Gianmarco Romano, Pierluigi Salvo Rossi, and Francesco Palmieri

***Abstract*—In this letter, we derive a gradient-based algorithm for computing the optimal transform when coefficients are transmitted over an erasure channel whose statistics are known. The discrete transform introduces correlation among the coefficients with consequent performance improvement against losses. Simulations show appreciable improvements over standard schemes and also good robustness when loss probabilities are only roughly estimated.**

***Index Terms*—Erasure channel, multiple description coding, transform coding.**

## I. Introduction

PACKET losses often have been associated to a multiple-channel model in which each packet, assumed to travel independently from the others, is associated to a different channel. Coding for such a channel is known as *multiple description coding* (MDC) as the different messages (packets) sent over the various channel are considered different source "descriptions": The idea is that when all the descriptions are received, high-quality reconstruction of the source is possible, while if only a small number of them is available, a smooth transition to lower quality reconstruction can be obtained.

The question of optimum correlating transform for multiple description coding has been addressed by Wang *et al.* [3] for $2 \times 2$ transforms and, more generally, by Goyal and Kovačević [4]. They pointed out the role of a *correlating transform* that, with the introduction of dependence among the coefficients (descriptions), "helps" the receiver to recover some information related to lost packets. Correlated coefficients are sent into different bit streams, each representing a description of the source. Goyal and Kovačević [4] find optimal $2 \times 2$ integer transforms for erasure channels and use them in practical encoders as building blocks for transforming vectors of any size. They also found some general analytical results for correlating transform in some special cases, i.e., for a specific number of descriptions (number of channels) and for a specific number of coefficients per description [4].

In this letter, we derive an algorithm for designing an optimal correlating transform of any dimension and with any number of descriptions. Our scheme is different with respect to those presented in [3] and [4] because we do not invert quantization and transform coding (integer transform). More comments about this are included in the following.

Manuscript received November 30, 2004; revised March 11, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dimitris A. Pados.

G. Romano and F. Palmieri are with the Dipartimento di Ingegneria dell'Informazione, Seconda Università di Napoli, Real Casa dell'Annunziata, 29-81031 Aversa, CE, Italy (e-mail: gianmarco.romano@unina2.it; francesco.palmieri@unina2.it).

P. Salvo Rossi is with the Dipartimento di Informatica e Sistemistica, Università di Napoli "Federico II," 21-80125 Napoli, Italy (e-mail: salvoros@unina.it).

Digital Object Identifier 10.1109/LSP.2005.855557

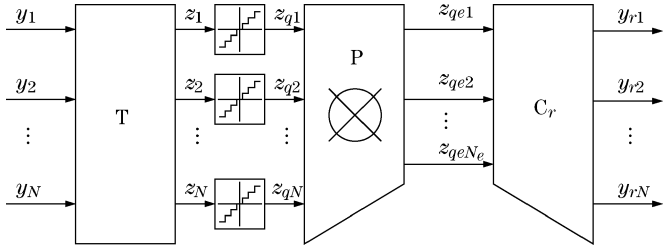


Fig. 1. Transform coding with erasures.

This letter is organized as follows. In Section II, we define the basic model for transform coding with erasures and define the optimization problem. Then, the gradient-based algorithm is derived by using techniques from matrix differential calculus. In Section III, the gradient algorithm is applied to a Markov-1 process on simulated channels that exhibit independent Bernoulli losses. Results are also reported about the robustness of our scheme against variations or inaccurate estimates of the erasure probabilities.

## II. System Description: Transform Coding With Erasures

The coding–decoding cascade model is shown in Fig. 1, where in addition to the traditional transform coding, we have included an erasure mechanism that randomly cancels some coefficients. The source is as an $N$-dimensional i.i.d. random vector $\mathbf{y} = (y_1, y_2, \ldots, y_N)^T$ with known correlation matrix. The $N \times N$ matrix $\mathbf{T}$ is the transform that is to be found. The $N$ transform coefficients $\mathbf{z} = (z_1, \ldots, z_N)^T$, where $\mathbf{z} = \mathbf{T}^T\mathbf{y}$, are quantized, with $N$ scalar quantizers $\mathbf{z}_q = Q(\mathbf{z})$, and sent over the channel. The matrix $\mathbf{P}$ models the erasure channel that can cancel a number of components of $\mathbf{z}_q$, i.e., $N_e(\leq N)$ random *erasures* may happen. At the receiver, only $N - N_e$ components are available for reconstructing $\mathbf{y}$.

Formally, at each channel use, the erasures can be described by a random binary vector $\mathbf{e} = (e_1, e_2, \ldots, e_N)^T$, with $e_i = 0$ if the $i$th component is erased and $e_i = 1$ otherwise. A compact description of the erasure process can be done by defining the residual vector $\mathbf{z}_e$ containing the $N - N_e$ survivor components, kept in the same order, and by defining an $(N - N_e) \times M$ matrix $\mathbf{P}(\mathbf{e})$ such that $\mathbf{z}_{qe} = \mathbf{P}(\mathbf{e})\mathbf{z}_q$. The random matrix $\mathbf{P}(\mathbf{e})$ is a permutation matrix with "0" and "1" as its elements, with a number of rows equal to the survivors and with one "1" per row in the position of the survivor. The structure of $\mathbf{P}(\mathbf{e})$ models the type of erasure that can happen on the channel: For example, $\mathbf{P}(\mathbf{e})$ may have a block structure to model packet-wise (group) losses. Note that matrix $\mathbf{P}$ has all zeros in the columns corresponding to the erasures and no rows with all zeros.

The receiver, which knows which coefficients have been erased,[1] provides a linear reconstruction of $\mathbf{y}$, $\mathbf{y}_r = \mathbf{C}_r^T(\mathbf{e})\mathbf{z}_{qe}$, via an $(N - N_e) \times N$ Wiener filter [6] $\mathbf{C}_{ro}(\mathbf{e}) = E[\mathbf{z}_{qe}\mathbf{z}_{qe}^T]^{-1}E[\mathbf{z}_{qe}\mathbf{y}^T]$, which, for each erasure configuration $\mathbf{e}$, minimizes the mean square error $D(\mathbf{e}) = (1/N)E[\|\mathbf{y} - \mathbf{y}_r\|^2]$.

Following a standard approach, if quantization is sufficiently fine, quantization noise $\eta$ is additive, has zero mean, has uncorrelated components, and is uncorrelated with the input signal $\mathbf{z}$ [7]. Therefore, the quantization noise correlation matrix is $\mathbf{R}_\eta = \mathrm{diag}(\beta_1\sigma_{z_1}^2, \ldots, \beta_N\sigma_{z_N}^2)$, where $\sigma_{zi}^2$ is the variance of the $i$th coefficient $z_i$. In general, $\beta_i$ has a form $a2^{-bn_b}$, where $n_b$ is the number of bits allocated for the $i$th component, and $a$ and $b$ depend on the type of quantizer and on the distribution of $z_i$ [8, Table 4.4]. The correlation matrix of the quantized coefficient is simply $\mathbf{R}_{z_q} = \mathbf{R}_z + \mathbf{R}_\eta = \mathbf{T}^T\mathbf{R}_y\mathbf{T} + \mathbf{R}_\eta$. By defining $\mathbf{B} = \mathrm{diag}(\beta_1, \ldots, \beta_N)$, $\mathbf{R}_\eta$ can be written in compact form as $\mathbf{R}_\eta = \mathbf{T}^T\mathbf{R}_y\mathbf{T} \odot \mathbf{B}$, where $\odot$ is the Hadamard product (element-by-element matrix product). With such a model for quantization, the optimal reconstruction filter is $\mathbf{C}_{ro}(\mathbf{e}) = (\mathbf{P}\mathbf{T}^T\mathbf{R}_y\mathbf{T}\mathbf{P}^T + \mathbf{P}\mathbf{R}_\eta\mathbf{P}^T)^{-1}\mathbf{P}\mathbf{T}^T\mathbf{R}_y$, and the corresponding distortion for each erasure is

$$D_o(\mathbf{e}) = \frac{1}{N}\mathrm{tr}(\mathbf{R}_y) - \frac{1}{N}\mathrm{tr}(\mathbf{W}\mathbf{T}^T\mathbf{R}_y) \tag{1}$$

where $\mathbf{W} = \mathbf{C}_{ro}(\mathbf{e})\mathbf{P}$. Given that the receiver acts optimally on each "erased" vector, the objective of our transform is to minimize the overall distortion averaged over all possible erasure events $D = (1/N)\sum_{\mathbf{e}} D_o(\mathbf{e})\Pr\{\mathbf{e}\}$. Even if the sample space of vector $\mathbf{e}$ can be quite large as $\mathbf{e}$ can assume $2^N$ configurations, in most applications, it is likely that the coefficients are grouped into a small number of packets, and therefore, the number of loss configurations is limited to manageable values. More generally, the overall cost function could be defined via an appropriately chosen weight function $w(\mathbf{e})$ instead of $\Pr\{\mathbf{e}\}$, i.e., $D = (1/N)\sum_{\mathbf{e}} D_o(\mathbf{e})w\{\mathbf{e}\}$. More specifically for various applications, the designer could choose to assign different importance to various loss configurations, not only as the result of their occurrence probability but also on the basis of different fidelity criteria determined by the application. For example, the designer may choose to emphasize *central distortion*, which is $D_o(\mathbf{e})$ when $\mathbf{e} = (1, 1, \ldots, 1)^T$(all coefficients are received) or vice-versa *side distortion* (not all coefficients are received) or something else. This generality is what make this framework quite flexible and different from previously proposed solutions. In a different scheme proposed in the context of multiple description coding [4],[9], transform and quantization were inverted, leading to an order with a quantizer followed by an integer-to-integer (I2I) correlating transform [4]. The use of the I2I transform was also motivated by the need to reduce the quantization error when a nonorthogonal transform is used. The side effect is that the I2I transform cannot change the central distortion that is due to the quantization error, and to get better side distortions, some redundancy in the rate is needed. Our scheme is quite different because the use of a nonorthogonal transform before quantization acts also on the quantization error and allows for a control mechanism also for the central distortion improving overall system performance.

The cost function to minimize is then

$$D = \frac{1}{N}\mathrm{tr}(\mathbf{R}_y) - \frac{1}{N}\sum_{\mathbf{e}} w(\mathbf{e})\mathrm{tr}(\mathbf{W}\mathbf{T}^T\mathbf{R}_y). \tag{2}$$

Therefore, the problem of nontrivial optimal choice for $\mathbf{T}$ is to find

$$\begin{cases} \mathbf{T}_o = \arg\max_{\mathbf{T}} \phi(\mathbf{T}) \\ \phi(\mathbf{T}) = \sum_{\mathbf{e}} w(\mathbf{e})\mathrm{tr}(\mathbf{W}\mathbf{T}^T\mathbf{R}_y) \end{cases}$$

subject to a constraint on the number $N_b$ of bits allocated to each vector, which corresponds to a rate $r = n_b/N$ bit/dim.

No simple structure for matrix $\mathbf{T}$ could be inferred from the cost function, and a closed-form solution for $\mathbf{T}$ is not likely to exist, also because $\mathbf{T}$ appears inside an inversion expression, and the performance is weighted over all loss configurations.

We search for the optimal matrix $\mathbf{T}$ with a gradient-ascent algorithm with the recursion

$$\mathbf{T}(n) = \mathbf{T}(n-1) + \mu\nabla_{\mathbf{T}}\phi(\mathbf{T}(n-1)) \tag{3}$$

where $\nabla_{\mathbf{T}}\phi$ is the gradient matrix (gradient flow) of $\phi$ with respect to $\mathbf{T}$, and $\mu$ is a scalar parameter that controls the speed of convergence. Such a gradient is computed (see the Appendix for the derivation) using techniques from matrix differential calculus [10] to be

$$\nabla_{\mathbf{T}}\phi = 2\sum_{\mathbf{e}} w(\mathbf{e})\left(\mathbf{R}_y\mathbf{W} - \mathbf{R}_y\mathbf{T}^T\mathbf{W}^T\mathbf{W} - \mathbf{R}_y\mathbf{T}\left((\mathbf{W}^T\mathbf{W}) \odot \mathbf{B}\right)\right). \tag{4}$$

Note that the gradient evaluation requires the correlation matrix of the quantization noise that depends on $\mathbf{T}$ and on bit allocation. We constrain the overall rate, i.e., the number of bits to be allocated to the quantizers, and the algorithm is organized as follows.

Initialize $\mathbf{T}$ to a random matrix and by optimal bit allocation, and assign to each quantizers a number of bits; now, matrix $\mathbf{B}$ can be computed, and the calculation of a new $\mathbf{T}$ with the gradient algorithm is possible. The new variances on the outputs of $\mathbf{T}$ will determine the new bit allocations and a new $\mathbf{B}$, etc. The iteration usually converges in a few steps to a minimum. This is conceivably a local minimum, but performance evolution shows almost invariably excellent improvement with respect to a unitary matrix $\mathbf{T}$ (no transform) or a randomly chosen $\mathbf{T}$.

The gradient expression requires the weighted sum over all the erasure configurations. When coefficients are grouped into descriptions, the number of all possible configurations is $2^{N_d}$, where $N_d$ is the number of descriptions. In typical applications, however, the number of descriptions is likely to be small, and we have a manageable number of loss configurations.

Parts of this algorithm could be replaced with Monte Carlo runs, for example, in the evaluations of the quantization variances, where we typically use theoretical formulas. In our experience, however, if at least three bits are assigned to each coefficient, the overall performance is relatively insensitive to the approximations. Extensions of the procedure may be implemented on line, as cooperative terminals may exchange channel loss statistics.

[1] In a packet communication link, the receiver knows which packets have been lost via a progressive numbering system (such as in RTP [5]).

## III. Simulations

We have performed many simulations on various types of signals such as images and one-dimensional sequences obtaining consistently excellent results [11], [12]. We report here typical results with reference to transform coding of a Markov sequence $\mathbf{x}$ with autocorrelation matrix $\mathbf{R}_x = \text{toeplitz}(r(0), \ldots, r(M-1))$, with $r(i) = \rho^{|i|}$. This is the correlation structure that corresponds to a sliding window on a Markov sequence. This model approximates rather well the autocorrelation structure of signals of practical interest, such as image blocks and some speech segments [8]. We apply the KL transform to such a sequence of length $M = 63$ and obtain a diagonal correlation matrix for the resulting coefficients. We consider only $N = 36$ transform coefficients, those with highest energy, and values of $\rho$ that range from 0.7 to 0.95. Coefficients are partitioned into three groups of dimension 12, by picking them up in a round-robin fashion so that descriptions have approximately the same importance. These groups are transmitted on independent erasure channels with erasure probability $p$. In order to evaluate the performance of the correlating transform, we have also computed the average mean-square error after losses when no correlating transform is applied. In such a case, packets are formed with interleaved quantized coefficients with no modifications. Such a distortion is compared to the one obtained after the inclusion of the correlating matrix $\mathbf{T}$. The gain $G = 10\log_{10}(D/D(\mathbf{T} = \mathbf{I}))$ is the measure of the improvement (in decibels) when the optimum correlating transform is applied.

Matrix $\mathbf{T}$ is computed with the gradient ascent algorithm from knowledge of the correlation matrix of the received data and knowledge of the loss probability. At each iteration, the gradient $\nabla\phi$ is evaluated according to (4), and a new matrix $\mathbf{T}$ is evaluated from (3), where $\mu$ is a scalar constant parameter chosen to be 0.25, which is a value that has shown to assure a local minimum within a few thousand iterations. Iterations stop when $\|\nabla\phi\|_2 < \epsilon$, with $\epsilon$ in the order of $10^{-3}$. We use uniform quantizers and a standard integer optimal bit allocation algorithm [8] with a total number of bits $n_b = 144$ to evaluate in turn $\mathbf{R}_\eta$. The rate is $r = 144/36 = 4$ bit/dim if considered after KL compression. Otherwise, $r = 144/63 = 2.29$ bit/dim for the original vector. We have used as a weight function $w(\mathbf{e})$ the occurrence probabilities.

The theoretical gain, from (2), is shown in Fig. 2. Results show gains ranging from about 1.5 to over 7 dB, with the highest gain for each sequence for a loss probability of about 0.2. Controlled correlation determines an excellent improvement compared to the standard transform coding based on KL, especially for low-to-moderate erasure probabilities. Within this range of erasure probabilities, the algorithm assigns automatically good side distortions and sacrifices central distortion, since it is more likely to have only a subset of descriptions rather than all of them. As the erasure probability increases, the probability of not receiving any description increases to a level that makes the advantage of using a correlating transform less drastic since the receiver begins to have not much information from which it can reconstruct the transmitted signal. In Fig. 2, we show also the gain obtained when a transmission over the erasure channel is simulated with the corresponding optimal correlating transform.

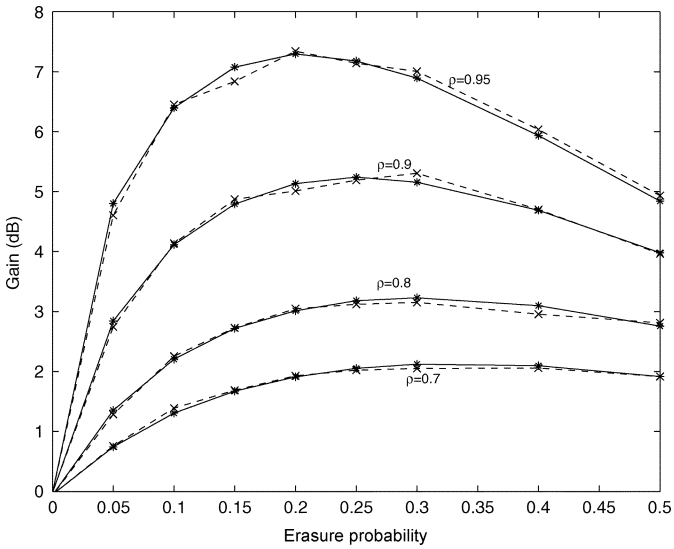


Fig. 2. Theoretical (asterisk $*$) and simulated (cross $\times$) gain over KLT for various $\rho$.

The results confirm the validity of the quantization noise model at different values of $\rho$ for the number of bits assigned to each coefficient. The reported results are to be considered typical as also varying $n_b$, $N$, and the number of packets, we obtain similar behaviors. We noted that the quantization assumption is satisfied as far as at least a mean rate of 3 b/coefficient is utilized.

The above simulations are based on the perfect knowledge of the channel behavior in the sense that the optimum correlating transform for the erasure probability $p$ is used for a simulated channel with erasure probability $p$. To address the natural question of what happens if $p$ is not exactly known, or it is varying, we have also performed some simulations to test the robustness of our scheme when a correlating transform determined for a specific erasure probability is used on a different channel. Fig. 3 shows the results. A plot of SNR versus the erasure probability is obtained for various optimal correlating transforms. Each one is optimized for a different erasure probability. Also shown are a plot of the SNR when no correlating transform is applied and one for a randomly chosen matrix $\mathbf{T}$. The distance between these curves is the gain due to the introduction of the correlating transform and the gain we get from optimization. These plots show that if the channel changes its probability, the presence of the correlating transform is still convenient with respect to the case with no correlating transform. From low to moderate erasure probabilities, the difference in the SNR among the different curves is small, confirming good robustness. The random correlating transform improves the performances, but the optimization algorithm gives always appreciable improvements. The curves also show that when no losses occur, performance degrades.

A designer may consider the performance in a no-loss scenario as the minimum acceptable one and optimize for his best estimate of channel loss.

In Fig. 4, a comparison with the optimal transform, as derived in [4], is shown. We have considered optimal $2 \times 2$ transforms and as performance parameter the overall distortion averaged over all erasure configurations, $D$, and a graph of $D$ versus the

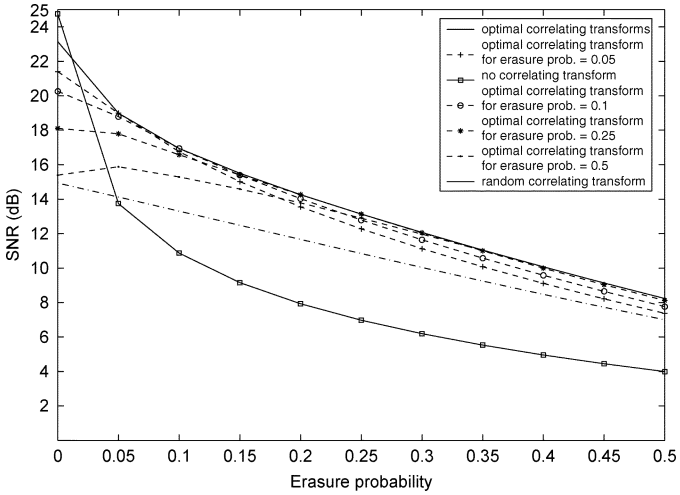


Fig. 3. Robustness against different probability of erasures.

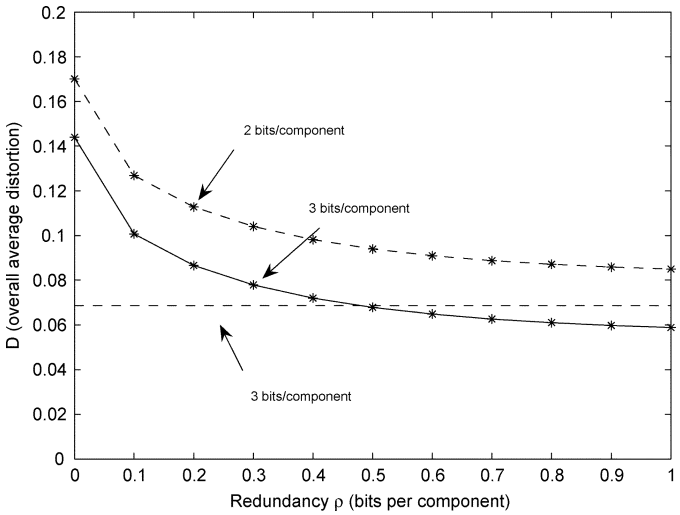


Fig. 4. Comparison with Goyal's transform. The overall average distortion for the optimal transform derived with the gradient method is provided for reference only and does not vary with redundancy. The average base rate is 3 b/component. The overall average distortion for the Goyal's MDCT is obtained for equal channel failure probabilities and is derived for a base rate of 2 and 3 b/component. Increase of redundancy improves side distortions, but the overall performances are worse than that obtained with the transform derived with the gradient method.$\mathbf{R}_y = \mathrm{diag}(1, 0.0625)$.

redundancy $\rho$ is reported for the Goyal transform for different base rates (2 and 3 b/component). The probability of erasure for each description is $p = 0.25$, and the correlation matrix is $\mathbf{R}_y = \mathrm{diag}(1, 0.0625)$. Note that, as expected, adding redundancy improves the performances of the Goyal transform because side distortions are better, but it has no effect on central distortion that contributes to $D$ as well. Also in the same figure, a reference value of distortion $D$ obtained with the optimal transform derived with the gradient method is shown for comparison purpose. In fact, no redundancy is added to the base rate used to determine the optimal transform. The comparison shows that for an average rate per component of 3 bits/component, the optimal transform determined with the gradient method performs better, since the overall rate needed by Goyal transform to obtain the same $D$ is always greater. Our scheme does not need redundancy because the algorithm sacrifices the central distortion to get better side distortions, which is something that the I2I transform cannot achieve since the central distortion depends only on the base rate.

## APPENDIX
## DERIVATION OF THE GRADIENT

We start with the computation of the differential of the trace

$$\mathrm{d}(\cdot) = \mathrm{d}\left(\mathrm{tr}(\mathbf{W}\mathbf{T}^T\mathbf{R}_y)\right).$$

Using the properties $\mathrm{d}(\mathrm{tr}(\mathbf{C}\mathbf{D})) = \mathrm{tr}(\mathrm{d}\mathbf{C}\mathbf{D} + \mathbf{C}\mathrm{d}\mathbf{D})$, $\mathrm{d}(\mathbf{F}\mathbf{G}\mathbf{F}^T) = \mathbf{F}\mathrm{d}\mathbf{G}\mathbf{F}^T$ ($\mathbf{F}$ is a constant matrix) [10, p. 174, eq. (5)], and $\mathrm{d}(\mathbf{G}^{-1}) = -\mathbf{G}^{-1}\mathrm{d}\mathbf{G}\mathbf{G}^{-1}$, [10, p. 183, eq. (17)], we have

$$\begin{aligned}\mathrm{d}(\cdot) &= \mathrm{tr}(2\mathbf{R}_y\mathrm{d}\mathbf{T}\mathbf{W}^T - 2\mathbf{W}\mathbf{T}^T\mathbf{R}_y\mathrm{d}\mathbf{T}\mathbf{W}^T - \mathbf{W}\mathrm{d}\mathbf{R}_\eta\mathbf{W}^T)\\ \mathrm{d}\mathbf{R}_\eta &= \left((\mathrm{d}\mathbf{T})^T\mathbf{R}_y\mathbf{T}\right)\odot\mathbf{B} + (\mathbf{T}^T\mathbf{R}_y\mathrm{d}\mathbf{T})\odot\mathbf{B}. \qquad (5)\end{aligned}$$

Using the property [10, p. 46, Th. 7] $\mathrm{tr}(\mathbf{A}^T(\mathbf{B}\odot\mathbf{C})) = \mathrm{tr}((\mathbf{A}^T\odot\mathbf{B}^T)\mathbf{C}) = \mathrm{tr}((\mathbf{A}^T\odot\mathbf{C}^T)\mathbf{B})$, we have

$$\begin{aligned}\mathrm{d}(\cdot) = 2\mathrm{tr}\Big(&\mathbf{W}^T\mathbf{R}_y\mathrm{d}\mathbf{T} - \mathbf{W}^T\mathbf{W}\mathbf{T}^T\mathbf{R}_y\mathrm{d}\mathbf{T}\\ &- \left((\mathbf{W}^T\mathbf{W})\odot\mathbf{B}\right)\mathbf{T}^T\mathbf{R}_y\mathrm{d}\mathbf{T}\Big).\end{aligned}$$

Let $\mathbf{C} = 2(\mathbf{W}^T\mathbf{R}_y - \mathbf{W}^T\mathbf{W}\mathbf{T}^T\mathbf{R}_y - ((\mathbf{W}^T\mathbf{W})\odot\mathbf{B})\mathbf{T}^T\mathbf{R}_y)$. Since $\mathrm{d}(\cdot) = \mathrm{tr}(\mathbf{C}\mathrm{d}\mathbf{T}) = (\mathrm{vect}\mathbf{C}^T)^T\mathrm{d}(\mathrm{vect}\mathbf{T})$, we can see that [10, p. 176, Tab. 2] $(\partial\phi/\partial\mathbf{T}) = \sum_{\mathbf{e}} w(\mathbf{e})\mathbf{C}^T$ and the expression (4) results.

## REFERENCES

[1] V. Goyal, "Multiple description coding: compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74–93, Sep. 2001.
[2] G. Romano and F. Palmieri, "An algorithm for transform coding on lossy packet networks," in *Proc. 37th Conf. Inf. Sci. Syst.*, Johns Hopkins Univ., Baltimore, MD, 2003.
[3] Y. Wang, M. Orchard, V. Vaishampayan, and A. Reibman, "Multiple description coding using pairwise correlating transforms," *IEEE Trans. Image Process.*, vol. 10, no. 3, pp. 351–366, Mar. 2001.
[4] V. Goyal and J. Kovačević, "Generalized multiple description coding with correlating transforms," *IEEE Trans. Inf. Theory*, vol. 47, no. 9, pp. 2199–2224, Sep. 2001.
[5] H. Shulzrinne, S. Casner, R. Frederick, and V. Jacobson, RTP: A transport protocol for real-time applications, in Request for Comments 1889, Jan. 1996.
[6] S. Haykin, *Adaptive Filter Theory*, second ed. Englewood Cliffs, NJ: Prentice-Hall, 1991.
[7] B. Widrow, I. Kollar, and M. C. Liu, "Statistical theory of quantization," *IEEE Trans. Instrum. Meas.*, vol. 45, no. 2, pp. 353–361, Apr. 1996.
[8] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[9] Y. Wang, M. Orchard, and A. Reibman, "Multiple description image coding for noisy channels by pairing transform coefficients," in *Proc. IEEE Workshop Multimedia Signal Process.*, Princeton, NJ, Jun. 1997, pp. 419–424.
[10] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus With Applications in Statistics and Econometrics*. New York: Wiley, 1988.
[11] F. Palmieri, P. Guerra, G. Micera, G. Romano, and S. Zampognaro, "Optimal transform coding for audio and images on lossy packet channel," in *Proc. Int. Conf. Packet Video*, Forte Village, Italy, Mar. 2000.
[12] G. Romano, P. Salvo Rossi, and F. Palmieri, "Multiple description image coder using correlating transforms," in *Proc. 12th Eur. Signal Process. Conf.*, Vienna, Austria, Sep. 6–10, 2004.