Optimal Erasure Protection for Scalably Compressed Video Streams With Limited Retransmission

David Taubman, *Member, IEEE,* and Johnson Thie

***Abstract*—This paper shows how the priority encoding transmission (PET) framework may be leveraged to exploit both unequal error protection and limited retransmission for RD-optimized delivery of streaming media. Previous work on scalable media protection with PET has largely ignored the possibility of retransmission. Conversely, the PET framework has not been harnessed by the substantial body of previous work on RD optimized hybrid forward error correction/automatic repeat request schemes. We limit our attention to sources which can be modeled as independently compressed frames (e.g., video frames), where each element in the scalable representation of each frame can be transmitted in one or both of two transmission slots. An optimization algorithm determines the level of protection which should be assigned to each element in each slot, subject to transmission bandwidth constraints. To balance the protection assigned to elements which are being transmitted for the first time with those which are being retransmitted, the proposed algorithm formulates a collection of hypotheses concerning its own behavior in future transmission slots. We show how the PET framework allows for a decoupled optimization algorithm with only modest complexity. Experimental results obtained with Motion JPEG2000 compressed video demonstrate that substantial performance benefits can be obtained using the proposed framework.**

***Index Terms*—Error protection, hybrid-ARQ, limited retransmission priority encoding transmission (LR-PET), PET, retransmission, robust transmission, scalable video.**

I. Introduction

THIS paper is concerned with the robust transmission of streaming scalable data through lossy communication channels, for applications in which limited retransmission is possible. We limit our attention to scalably compressed video frames, although the framework which we develop here may also be applied to the protection of other real-time media, such as audio.

Traditionally, the possibility of transmission errors has been addressed either through the use of forward error correction (FEC) or by automatic repeat request (ARQ) retransmission schemes. FEC approaches are often advocated for the transmission of real-time compressed data, based on the assumption that retransmission might cause unacceptable delays. Much recent research into the protection of scalable compressed imagery over packet-based networks [1]–[8] has adopted this perspective, exploiting the priority encoding transmission (PET) scheme of Albanese *et al.* [9] as a framework for unequal error protection.

Manuscript received September 12, 2003; revised May 13, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Aria Nostratinia.

The authors are with the School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney 2052, Australia (e-mail: d.taubman@unsw.edu.au; j.thie@ee.unsw.edu.au).

Digital Object Identifier 10.1109/TIP.2005.846028

In some cases, retransmission of lost data may indeed not be possible, due to stringent constraints on the delivery time, or because the transmission medium provides no means for the transmitter to discover whether or not a packet has been correctly received.[1] In many applications, however, delivery time constraints merely serve to limit the number of round trips, and, hence, the number of retransmission opportunities which exist. This perspective is adopted by a considerable body of literature [10]–[15], which deals with the optimized delivery of streaming media with limited retransmission. These works include retransmission decisions within the optimization framework. The combination of both limited retransmission and FEC has also been considered in a variety of settings [16]–[22]. We shall mention some of these again shortly. For the moment, however, we note that the advantages of the PET framework have not previously been considered in the context of limited retransmission. The one exception to this, of which we have recently become aware, is the work of Gan and Ma [22], [23]. As we shall see, our proposed strategy differs in a number of important respects from that advanced by Gan and Ma.

A. Scalable Source Frames

To study the delivery of streaming scalable media in the context of limited retransmission, we model the scalable data source as a sequence of "frames," each of which is compressed independently. Highly scalable image compression schemes have been widely investigated over the past 15 years. Some of the most notable algorithms to emerge from these efforts are the EZW [24] and SPIHT [25] algorithms, and the JPEG2000 [26] image compression standard. The latter is based on a variety of important concepts, including context-adaptive bit-plane coding [27], context-adaptive bit-plane reordering [28], [29], and embedded block coding with optimized truncation [30].

These algorithms all produce compressed bit-streams, with the property that low-quality representations are embedded within higher quality representations. Equivalently, the compressed bit-stream may be viewed as a succession of elements $\mathcal{E}_q$, $1 \leq q \leq Q$, having coded lengths L_q, which progressively augment the received image quality. For the embedded representation to be efficient, the image quality obtained when the image is reconstructed from an initial $Q' \leq Q$ of these elements should be comparable to that which could be obtained if the image were compressed using any other efficient algorithm, be it scalable or otherwise, to the same total length $L = \sum_{q=1}^{Q'} L_q$.

[1]It is worth noting, though, that existing protection schemes universally assume the existence of some feedback mechanism, whereby the transmitter is able to discover the prevailing network conditions and, hence, estimate loss probabilities.

Modern scalable image compression algorithms such as SPIHT and JPEG2000 are efficient in this sense.

Scalable video compression algorithms have also been the subject of intense research [31]–[37], with recent results demonstrating performance approaching that of the most advanced nonscalable techniques, while offering a large number of embedded elements (see, for example, [38]). The key difference between scalable image compression and scalable video compression is that a motion-compensated temporal transform is required, to efficiently exploit interframe redundancies.

For the purpose of this paper, it is convenient to model any scalably compressed video stream as a sequence of independently compressed "source frames" $\mathcal{F}[n]$, each of which has its own collection of embedded elements. In the simplest case, each original frame of the video sequence is independently compressed using a scalable image coder. More generally, when the original video frames are compressed jointly, so as to exploit motion redundancy, the compressed representation produced by any of the highly scalable video compressors of which we are aware may still be modeled as a sequence of element groups, where the elements in each group are coded independently from those in other groups. These element groups are identified variously in the literature as "groups of pictures," "frame slots," and so forth. The theory and algorithms presented in this paper may be applied directly to such data streams by treating each independent element group as a separate "source frame," $\mathcal{F}[n]$. For the discussion which follows, however, it is simplest to think of the $\mathcal{F}[n]$ as actual compressed video frames. Indeed, the experimental results which we present in Section IV-F are obtained by compressing each frame of a video sequence separately using JPEG2000; this is known as Motion JPEG2000.[2]

B. PET Framework

The lossy communication channels considered in this work are packet-based, where the packet transport model is that of an "erasure channel." An erasure channel is one in which each packet either arrives intact or is entirely lost. A key property of erasure channels is that the receiver knows exactly which packets have been lost (the "erasures"). This is a good model for most packet-based communication applications, although we note that packet loss in IP networks is usually assessed on the basis of excessive arrival delay, rather than certain knowledge that it has been lost.

In the context of erasure channels, Albanese *et al.* [9] introduced PET. The PET scheme works with a family of (N, k) channel codes, all of which have the same codeword length N, but different source lengths $1 \leq k \leq N$. We consider only "maximum distance separable" (MDS) codes, which have the key property that receipt of any k out of the N symbols in a codeword is sufficient to recover the k source symbols. The amount of redundancy $R_{N,k} = N/k$ determines the strength of the code, where smaller values of k correspond to stronger codes. It is convenient to augment this set of channel codes with the special value $k = \infty$, for which $R_{N,\infty} = 0$, meaning that the element is not transmitted at all. Note that redundancy $R_{N,k}$ is the reciprocal of the more commonly used "code rate" k/N. We find it more convenient to refer to redundancy, rather than code rate, in large part due to the importance of the case $R_{N,\infty} = 0$.

[2]As with Motion JPEG, the term "motion" refers to the fact that motion video is being compressed, rather than suggesting that scene motion is exploited.

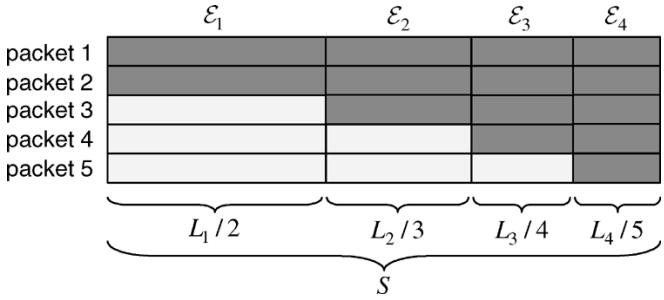

Fig. 1. Example of a PET frame consisting of $N = 5$ packets, into which $Q = 4$ elements are coded. The elements $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$, and $\mathcal{E}_4$, having equal lengths L_q, are assigned channel codes with $k = 2$, 3, 4, and 5, respectively. The dark and light shaded boxes correspond to source and parity symbols, respectively.

We measure the length L_q of each scalable source element $\mathcal{E}_q$ in "symbols." In our experiments, each source symbol corresponds to one byte, although other symbol sizes may be used. Given a collection of source elements $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_Q$, having uncoded lengths $L_1, L_2, \ldots, L_Q$, and channel code redundancies $R_{N,k_1} \geq R_{N,k_2} \geq \cdots \geq R_{N,k_Q}$, the PET scheme packages the encoded elements into N network packets, which we call a "PET frame." L_q/k_q source symbols are placed in each of the first k_q packets, while each of the remaining $N - k_q$ packets contains L_q/k_q parity symbols. This arrangement guarantees that receipt of any $k' \geq k_q$ packets is sufficient to recover element $\mathcal{E}_q$. The total encoded transmission length is $\sum_q L_q R_{N,k_q}$, which must be arranged into N packets, each having S symbols. Fig. 1 shows an example of arranging $Q = 4$ elements into a PET frame consisting of $N = 5$ packets. Consider element $\mathcal{E}_2$, which is assigned a (5,3) code. Since $k_2 = 3$, three out of the five packets together contain the source element's L_2 symbols, while the remaining two packets contain parity symbols. Hence, receiving any three packets guarantees recovery of element $\mathcal{E}_2$ and also $\mathcal{E}_1$, but not $\mathcal{E}_3$ or $\mathcal{E}_4$.

The PET frame construction described here ignores the possibility that L_q might not be exactly divisible by k_q. To deal with this eventuality in practice, some symbols from $\mathcal{E}_{q+1}$ may be protected with the (N, k_q) code used for $\mathcal{E}_q$, rather than the (potentially weaker) code which is otherwise used for the elements of $\mathcal{E}_{q+1}$. Such a policy tends to slightly increase the overall encoded length of the PET frame, forcing the selection of channel codes whose redundancies may be lower overall than suggested by the equations developed in this paper. These effects, while usually small, may slightly degrade the optimality of the channel code assignment algorithms presented in the sequel.

Using the PET framework, several strategies [1]–[8] have been proposed for finding the optimal channel code assignment for each source element, under the condition that the total encoded transmission length should not exceed a specified limit, $L_{\max} = NS$. Generally, the optimization objective is an expected utility U, which must be an additive function of the source elements that are correctly recovered. That is

$$U = U_0 + \sum_{q=1}^{Q} U_q P_{N,k_q} \tag{1}$$

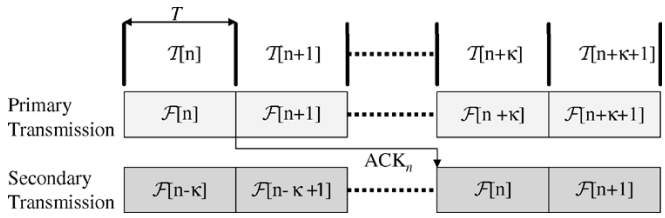

Fig. 2. Relationship between source frames $\mathcal{F}[n]$ and transmission slots $\mathcal{T}[n]$. ACK_n signifies the flow of acknowledgment information, whereby the receiver informs the transmitter of the number of packets which were lost during transmission slot $\mathcal{T}[n]$.

where U_0 is the amount of utility at the receiver when no source element is received, U_q is the utility associated with receiving $\mathcal{E}_q$, and P_{N,k_q} is the probability of recovering the source element $\mathcal{E}_q$, which is assigned an (N, k_q) channel code. Specifically, P_{N,k_q} is equal to the probability of receiving at least k_q out of N packets. As already mentioned, the special value $k_q = \infty$ is reserved for source elements which are not transmitted, having probability $P_{N,\infty} = 0$. Commonly, $-U$ represents the mean-squared error (MSE) of the reconstructed image or video frame, and U_q corresponds to a reduction in MSE associated with recovery of element $\mathcal{E}_q$. The term U_0 is included only for completeness; it plays no role in the intuitive or computational aspects of the optimization problems with which we are concerned.

C. Limited Retransmission

As mentioned already, we are interested in applications where limited retransmission of lost data is possible. To simplify matters, we limit our attention to the case in which the scalable source elements which constitute each source frame have two opportunities for transmission. Each source frame $\mathcal{F}[n]$ is assigned a primary "transmission slot" $\mathcal{T}[n]$, and a secondary transmission slot $\mathcal{T}[n+\kappa]$, during which information may be retransmitted. Each transmission slot has a fixed duration of T s, so that source frames are transmitted at a constant rate of $1/T$ frames/s. The primary and secondary transmission slots are separated by at least $(\kappa - 1)T$ s, and the value of κ is chosen so that this separation is long enough for the transmitter to discover whether or not packets sent during the primary transmission slot arrived successfully. Fig. 2 illustrates the relationship between source frames and transmission slots.

In the context of Fig. 2, it is possible to appreciate the key challenges of the protection assignment problem with which we are concerned. During any given transmission slot $\mathcal{T}[n]$, the transmitter must distribute the available bandwidth between the primary transmission of elements from source frame $\mathcal{F}[n]$ and the secondary transmission (retransmission) of elements from source frame $\mathcal{F}[n-\kappa]$. We write $L_{\max}$ for the maximum number of symbols which may be transmitted within slot $\mathcal{T}[n]$ and adopt the PET framework to encode elements from both source frames within a single PET frame, having N packets, each with at most S symbols, where $NS = L_{\max}$. Since the source data must be transmitted in real time, and the channel conditions may be subject to change over time, the transmitter must determine how best to distribute the total of $L_{\max}$ symbols between two different source frames, without knowing how many packets will be lost or what protection might be applied to the retransmission of any lost data from frame $\mathcal{F}[n]$ in its secondary transmission slot $\mathcal{T}[n+\kappa]$.

In this paper, we propose a general framework for solving the protection assignment problem mentioned above. To do this, we formulate a collection of hypotheses concerning the current frame's secondary transmission, which will take place in the future. We coin the term limited retransmission PET (LR-PET) to describe this novel framework, and we show that an optimal solution to the LR-PET assignment problem may be found with modest computational complexity.

To appreciate the potential advantage of LR-PET over frame-by-frame PET, consider again the example of Fig. 1, and suppose that only packets 1 and 3 arrive at the receiver. Without the possibility of retransmission, only element $\mathcal{E}_1$ may be successfully recovered. If retransmission is permitted, however, only the $L_2/3$ source symbols of $\mathcal{E}_2$ which belong to packet 2 need be retransmitted in order for the receiver to recover the whole of element $\mathcal{E}_2$. The relative value of these symbols is, thus, much higher during retransmission than it was during the original transmission, since the full utility of $\mathcal{E}_2$ can be realized at one third of its full transmission cost. A similar argument may be applied to the retransmission of symbols from the other elements, $\mathcal{E}_3$ and $\mathcal{E}_4$. Thus, LR-PET essentially augments the set of source symbols which may be transmitted (and protected) during a transmission slot with retransmitted source symbols, where the latter generally have elevated utility-length ratios. Higher utility-length ratios allow a larger expected utility to be achieved with LR-PET than with PET, subject to the transmission slot's length budget, $L_{\max}$.

In any given transmission slot, $\mathcal{T}[n]$, the temptation to favor retransmitted elements from frame $\mathcal{F}[n-\kappa]$, with their higher effective utility-length ratios, over elements from frame $\mathcal{F}[n]$, must be balanced by the fact that the utility-length advantage of future retransmissions (in slot $\mathcal{T}[n+\kappa]$) hinges on assigning the elements of $\mathcal{F}[n]$ some bandwidth during slot $\mathcal{T}[n]$. This emphasizes the importance of including hypotheses concerning the future retransmission of elements from $\mathcal{F}[n]$, when optimizing the distribution of bandwidth between elements from frames $\mathcal{F}[n]$ and $\mathcal{F}[n-\kappa]$.

D. Related Work

At this point, it is possible to compare and contrast our proposed approach with other relevant protection optimization frameworks. Perhaps the most relevant previous work is that of Chou and his coauthors, which is well represented by [14], [17], [18], amongst other papers. Chou *et al.* advance a powerful framework for rate-distortion optimal delivery of streaming scalable media, allowing for both FEC and retransmission with deadlines. At each transmission opportunity (or transmission slot), the server explores the impact of various policies for both the current transmission and hypothetical future retransmissions, searching for the policy which optimizes the expectation of a Lagrangian cost function. In this regard, their approach is actually identical to our own, as developed in Section IV. Related ideas may be found to varying degrees in [10]–[12]. In each case, the idea is to model the effect of different transmission policies on the properties of a Markov decision process,

finding the policy which maximizes the expectation of a utility objective over all paths through the state transition trellis.

A key challenge associated with such schemes is to manage the explosion in complexity (size of the state space) which arises from the interaction between transmission policies for each source element, especially in the context of a rich set of error correction options and/or acknowledgment failure possibilities. To the best of our knowledge, the present paper is the first to cast this optimization problem within the PET framework. We show that the PET framework allows the optimization problem to be decomposed into a set of independent optimization problems, associated with each successive source element in the primary frame. This allows us to avoid sub-optimal iterative policy optimization schemes, such as the SA algorithm which underpins the work of Chou *et al.* Besides simplifying the optimization algorithm, the PET framework is also fundamentally more efficient than one in which each source element is separately packetized and channel encoded, as envisaged by previous works which consider retransmission. This is because whenever a source element arrives at the decoder, all higher priority source elements on which it depends are also guaranteed to arrive. It is this property which yields the linear utility expression in (1), as opposed to the necessarily smaller polynomial utility expressions manipulated by Chou and others.

Very recently, Gan and Ma [22], [23] proposed a transmission optimization framework which does combine PET with the possibility of limited retransmission. Their approach, however, differs in many important respects from that proposed here. Gan and Ma do not exploit the fact that it is generally only necessary to retransmit a subset of the symbols associated with a lost source element. As explained in the example above, this has the important effect of increasing the effective utility-length ratio for retransmitted elements, accounting for much of the advantage of LR-PET over frame-by-frame PET. Quite to the contrary, results reported in [22] indicate a loss in overall expected utility, while those reported in [23] indicate only small gains and occasional overall losses, as retransmission is added to frame-by-frame PET. Further, the optimization objective in [22], [23] does not account for the potential retransmission of elements which are currently being transmitted/protected for the first time. The formulation of hypotheses to account for future retransmission is a central theme in our framework. It should be noted that Gan and Ma's work is motivated principally by the desire to minimize quality fluctuations from frame to frame, rather than maximizing an overall expected utility objective, as formulated in our present work. Finally, we note that our present work involves the simplifying assumption of at most two transmission opportunities for each source element, while Gan and Ma consider larger numbers of transmission opportunities.

E. Organization of the Paper

The remainder of this paper is organized as follows. Section II explores the erasure channel model more thoroughly, introducing the notation which we shall use throughout the paper. Section III then reviews the problem of optimal protection assignment for a single-scalable data source (e.g., a compressed image), with a linearly dependent sequence of elements $\mathcal{E}_q$, without the possibility of retransmission. Section IV describes our proposed LR-PET framework and the associated protection assignment strategy. Experimental results in Section IV-F then demonstrate that substantial performance benefits can be obtained by allowing limited retransmission within the LR-PET framework.

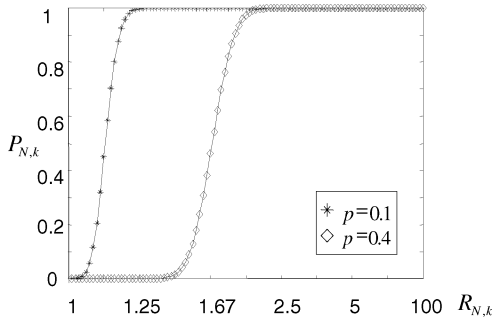

Fig. 3. Two $P_{N,k}$ versus $R_{N,k}$ characteristics, corresponding to $N = 100$, with $p = 0.1$ and $p = 0.4$.

II. Channel Model and Coding

The channel model we use is that of an erasure channel, where the receiver knows exactly which packets have been lost. Packet loss may result from either corruption or congestion. For the experimental results reported in this paper, we consider only an IID packet loss model, with packet loss probability p. However, the theoretical results and analytical methods described in this paper are not restricted to IID models. The Gilbert–Elliott [39] model, for example, is commonly used to model the bursty loss patterns observed in networks which are subject to congestion.

With the IID loss model, the probability of receiving at least k out of N packets with no error is given simply by

$$P_{N,k} = \sum_{i=k}^{N} \binom{N}{i} (1-p)^i \, p^{N-i}. \tag{2}$$

Fig. 3 shows an example of the relationship between $P_{N,k}$ and $R_{N,k}$, for the cases $p = 0.1$ and $p = 0.4$, with codeword length $N = 100$. Evidently, $P_{N,k}$ is monotonically increasing with $R_{N,k}$, but note that the curve is not convex.[3] As described in the next section, optimal solutions to the protection assignment problem for a single-scalable data source must inevitably be drawn from the upper convex hull of this curve. However, this need not necessarily be the case for the limited retransmission scenario with which we are ultimately concerned.

It is convenient to parametrize $P_{N,k}$ and $R_{N,k}$ by a single parameter r, defined by

$$r = \begin{cases} N + 1 - k, & k \neq \infty \\ 0, & k = \infty. \end{cases}$$

We think of r as a "redundancy index." Evidently, r lies in the range 0 to N, with $r = 0$ corresponding to the case where no information is transmitted, and $r = N$ corresponding to the case where only 1 of the N transmitted packets need be received for

[3]Throughout this paper, we use the term "convex" to mean "convex $\cap$." Convex $\cap$ functions are also sometimes called "concave," although we do not use that term here.

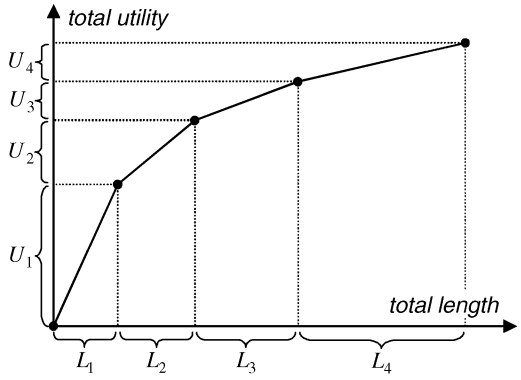

Fig. 4. Example of the convex ∩ utility-length characteristic for a source consisting of four elements.

correct decoding. The redundancy index is more intuitive than k, as a measure of the strength of the channel code, particularly in view of the importance of the case $r = 0$.

Since each of our PET frames will always involve N packets, we find it convenient to drop the explicit dependence on N. This leads to the simplified notation

$$P(r) = P_{N,k_{\min}(r)}, \quad R(r) = R_{N,k_{\min}(r)} \tag{3}$$

where

$$k_{\min}(r) = \begin{cases} N+1-r, & r > 0 \\ \infty, & r = 0 \end{cases} \tag{4}$$

is the minimum number of packets which must be received if an element which has been assigned redundancy index r is to be recovered.

III. PET Protection Assignment for a Single Source

This section reviews the problem of assigning an optimal set of channel codes to the elements $\mathcal{E}_1$ through $\mathcal{E}_Q$ of a single-source frame $\mathcal{F}$, subject to a constraint $L_{\max}$ on the length of the PET frame. For applications where retransmission is not possible, we use the methods described here to protect each source frame $\mathcal{F}[n]$ within its own transmission slot $\mathcal{T}[n]$. The material presented here also serves to facilitate our later discussion of the more complex LR-PET assignment problem, where primary and secondary transmissions of two different frames must be jointly protected.

Several schemes have been reported for finding optimal channel codes within the PET framework [1]–[8]. For the purpose of this review, we follow the development in [8]. In that work, a general set of results is provided for sources whose utility-length characteristic need not be convex and channel codes whose probability-redundancy characteristic need also not be convex. For our present purposes, however, we shall assume that the source utility-length characteristic is convex ∩, as illustrated in Fig. 4. That is

$$\frac{U_1}{L_1} \geq \frac{U_2}{L_2} \geq \cdots \geq \frac{U_Q}{L_Q}. \tag{5}$$

We assume that the source elements exhibit a linear dependency structure, $\mathcal{E}_1 \prec \mathcal{E}_2 \prec \cdots \prec \mathcal{E}_Q$, meaning that elements $\mathcal{E}_1$ through $\mathcal{E}_q$ must all be available before $\mathcal{E}_{q+1}$ can be correctly decoded. To each element $\mathcal{E}_q$, we assign a channel code with codeword length N and redundancy index r_q, placing all encoded elements within a PET frame with N packets. Our goal is to find the set of redundancy indices r_q, which maximize the expected utility, subject to the length constraint

$$L = \sum_{q=1}^{Q} L_q R(r_q) \leq L_{\max}. \tag{6}$$

Without any loss of generality, we may restrict our attention to solutions which satisfy $r_1 \geq r_2 \geq \cdots \geq r_Q$. To see this, suppose the expected utility is maximized by a set of redundancy indices $\{r'_q\}$, where $r'_q < r'_{q+1}$ for some q. An alternate set of redundancy indices $\{r''_q\}$ may be formed from $\{r'_q\}$, by setting $r''_{q+1} = r'_q$ and leaving all other indices unchanged. This alternate set has a smaller encoded length, since $R(r''_{q+1}) = R(r'_q) < R(r'_{q+1})$. However, it has exactly the same expected utility, since $\mathcal{E}_{q+1}$ cannot contribute to the utility unless $\mathcal{E}_q$ is successfully recovered, so there is no point in protecting $\mathcal{E}_{q+1}$ more strongly than $\mathcal{E}_q$. Iterating the construction, if necessary, we see that any optimal set of redundancy indices $\{r'_q\}$ may be progressively converted into a (possibly different) set $\{r''_q\}$, having exactly the same expected utility, at most the same encoded length, and satisfying $r''_1 \geq r''_2 \geq \cdots \geq r''_Q$. It follows that the indices $\{r''_q\}$ also maximize the expected utility, subject to the same length constraint.

Since $r_1 \geq r_2 \geq \cdots \geq r_Q$, the PET framework guarantees that elements $\mathcal{E}_1$ through $\mathcal{E}_q$ will all be correctly received, so long as $\mathcal{E}_q$ is correctly received. The expected utility U, available at the receiver, can then be expressed as

$$U = U_0 + \sum_{q=1}^{Q} U_q P(r_q)$$

and our task is to maximize U, subject to $r_1 \geq r_2 \geq \cdots \geq r_Q$ and the length constraint of (6).

The length-constrained optimization problem may be converted to a family of unconstrained optimization problems, parametrized by a quantity $\lambda > 0$. Specifically, let $U^{(\lambda)}$ and $L^{(\lambda)}$ denote the expected utility and transmission length associated with the set of redundancy indices $\left\{r_q^{(\lambda)}\right\}$, which maximize the Lagrangian-style functional

$$J^{(\lambda)} = U^{(\lambda)} - \lambda L^{(\lambda)} = \sum_{q=1}^{Q} \left(U_q P\left(r_q^{(\lambda)}\right) - \lambda L_q R\left(r_q^{(\lambda)}\right) \right) \tag{7}$$

subject to $r_1^{(\lambda)} \geq r_2^{(\lambda)} \geq \cdots \geq r_Q^{(\lambda)}$. Evidently, it is impossible to increase U beyond $U^{(\lambda)}$, without also increasing L beyond $L^{(\lambda)}$. Thus, if we can find λ such that $L^{(\lambda)} = L_{\max}$, the set $\left\{r_q^{(\lambda)}\right\}$ will form an optimal solution to our problem. In practice, the discrete nature of the problem may prevent us from finding a value λ such that $L^{(\lambda)}$ is exactly equal to $L_{\max}$, but if the elements are small enough, we should be justified in ignoring this small source of sub-optimality and selecting the smallest value of λ such that $L^{(\lambda)} \leq L_{\max}$.

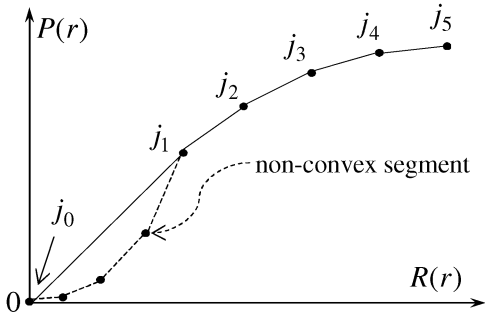

Fig. 5. Points $j_0, j_1, \ldots, j_5$, which belong to the upper convex hull $\mathcal{H}_C$, of a nonconvex $P(r)$ versus $R(r)$ characteristic.

Now, suppose we temporarily ignore the constraint that the redundancy indices $r_q^{(\lambda)}$ must be nonincreasing. We shall find that the resulting solutions will then be automatically guaranteed to satisfy this constraint. The unconstrained optimization problem decomposes into a collection of Q separate maximization problems. In particular, we seek $r_q^{(\lambda)}$ which maximizes

$$J_q^{(\lambda)} = U_q P\left(r_q^{(\lambda)}\right) - \lambda L_q R\left(r_q^{(\lambda)}\right)$$

for each $q = 1, 2, \ldots, Q$. Equivalently

$$r_q^{(\lambda)} = \underset{r}{\operatorname{argmax}} \left[P(r) - \lambda_q R(r)\right]$$

where $\lambda_q = \lambda L_q / U_q$.

This optimization problem arises in other contexts, such as the optimal truncation of embedded compressed bit-streams [26, chapter 8]. It is known that the solution $r_q^{(\lambda)}$ must belong to the set $\mathcal{H}_C$ which describes the upper convex hull of the $P(r)$ versus $R(r)$ characteristic, as shown in Fig. 5. If $0 = j_0 < j_1 < \cdots$ is an enumeration of the elements in $\mathcal{H}_C$, and

$$S_C(i) = \begin{cases} \frac{P(j_i) - P(j_{i-1})}{R(j_i) - R(j_{i-1})}, & i > 0 \\ \infty, & i = 0 \end{cases} \tag{8}$$

are the "slope" values on the convex hull, then $S_C(0) \geq S_C(1) \geq \cdots$. The solution to our optimization problem can then be found from

$$\begin{aligned} r_q^{(\lambda)} &= \max\left\{j_i \in \mathcal{H}_C \mid S_C(i) \geq \lambda_q\right\} \\ &= \max\left\{j_i \in \mathcal{H}_C \mid S_C(i) \frac{U_q}{L_q} \geq \lambda\right\}. \end{aligned} \tag{9}$$

Moreover, since the source is convex (5), we must have $\lambda_q \leq \lambda_{q+1}$, from which we deduce that $r_q^{(\lambda)} \geq r_{q+1}^{(\lambda)}$. That is, the redundancy indices which maximize the unconstrained objective of (7) are guaranteed to satisfy the constraint $r_1^{(\lambda)} \geq r_2^{(\lambda)} \geq \cdots \geq r_Q^{(\lambda)}$, as promised.

In summary, the optimal set of channel code redundancy indices is found by searching for the smallest value of λ such that the solutions $r_q^{(\lambda)}$ yielded by (9) produce a PET frame length $L^{(\lambda)}$ which is no larger than $L_{\max}$. The monotonic dependence of $r_q^{(\lambda)}$ and, hence, $U^{(\lambda)}$ and $L^{(\lambda)}$, on λ, renders this search problem particularly simple. A trivial bisection search, for example, limits the number of distinct values for λ which must be tried to at most the numerical precision (number of bits) used to represent slope values.

IV. LR-PET Framework

In the previous section, we assumed that each frame $\mathcal{F}[n]$ has its own transmission slot $\mathcal{T}[n]$, without any possibility for retransmission of lost data in future transmission slots. This led to a relatively simple strategy for assigning optimal redundancy indices $r_q[n]$ to its elements $\mathcal{E}_q[n]$. In this section, we consider the more general context, in which each source frame $\mathcal{F}[n]$ has both a primary transmission slot $\mathcal{T}[n]$, and a secondary transmission slot $\mathcal{T}[n+\kappa]$. Fig. 2 illustrates the relationship between source frames and transmission slots.

During transmission slot $\mathcal{T}[n]$, the transmitter selects redundancy indices $r_q[n]$ for each element of the primary frame $\mathcal{F}[n]$, as well as retransmission redundancy indices $s_q[n-\kappa]$ for each element of frame $\mathcal{F}[n-\kappa]$, subject to a limit $L_{\max}$ on the total number of symbols which can be transmitted within any slot. We shall consistently use the notation r_q and s_q to refer to the redundancy indices associated with primary and secondary transmission, respectively. We also remind the reader that the case in which an element is not transmitted at all is captured by the assignment of a redundancy index of 0.

Our objective is to jointly optimize the parameters $r_q[n]$ and $s_q[n-\kappa]$, subject to the length constraint $L_{\max}$ on slot $\mathcal{T}[n]$. To appreciate the nature of this problem, note that the utility of frame $\mathcal{F}[n]$ at the receiver depends on both $r_q[n]$ and $s_q[n]$, yet we must optimize the $r_q[n]$ and $s_q[n-\kappa]$ parameters together, without precise knowledge of which packets from frame $\mathcal{F}[n]$ will be lost, or what protection might be assigned to the retransmission of lost data. We adopt the PET framework, jointly protecting all source elements transmitted in slot $\mathcal{T}[n]$ with a single PET frame of N packets.

To address the dependence of $r_q[n]$ on the $s_q[n]$ values, which will not be assigned until the future transmission slot $\mathcal{T}[n+\kappa]$, we formulate a collection of hypotheses on what might happen in that slot. In particular, let ρ_k denote the probability that exactly k of the N packets transmitted in $\mathcal{T}[n]$ is correctly received. Each value of k generates a separate hypothesis. The expected utility of frame $\mathcal{F}[n]$ may then be expressed as

$$E[U[n]] = U_0[n] + \sum_{k=0}^{N} \rho_k \sum_q U_q[n] P\left(k, r_q[n], s_q^k[n]\right) \tag{10}$$

where $P\left(k, r_q[n], s_q^k[n]\right)$ is the probability that element $\mathcal{E}_q[n]$ will be correctly received if it is assigned a redundancy index of $r_q[n]$ in slot $\mathcal{T}[n]$, if only k of the N packets transmitted in that slot are received correctly, and if the remaining data is retransmitted with a redundancy index of $s_q^k[n]$ in slot $\mathcal{T}[n+\kappa]$. It is important to note that the additive formulation in (10) is valid only so long as the real and hypothetical redundancy indices satisfy

$$r_q[n] \geq r_{q+1}[n], \ \forall q \quad \text{and} \quad s_q^k[n] \geq s_{q+1}^k[n], \ \forall q, k. \tag{11}$$

A similar argument to that presented for the case of simple source protection in Section III shows that no loss of generality is incurred by imposing these constraints.[4]

The expected total transmission length associated with frame $\mathcal{F}[n]$ may be expressed as

$$E[L[n]] = \sum_{k=0}^{N} \rho_k \sum_q L_q[n] R\left(k, r_q[n], s_q^k[n]\right) \tag{12}$$

where $R\left(k, r_q[n], s_q^k[n]\right)$ is the ratio between the total coded length (in its primary and secondary slots) and the uncoded length of $\mathcal{E}_q[n]$, if redundancy index $r_q[n]$ is assigned in the primary slot $\mathcal{T}[n]$, if only k of the N packets transmitted in that slot are received correctly, and if the remaining data is retransmitted with a redundancy index of $s_q^k[n]$ in slot $\mathcal{T}[n+\kappa]$.

In order to account for the complex interaction between current and future redundancy indices, we adopt a global perspective. Consider for a moment the global objective of maximizing the total expected utility $E[\sum_n U[n]]$, subject to some constraint on the total expected length $E[\sum_n L[n]]$. Following the same reasoning presented in Section III, solutions which maximize such a global objective form a family, parametrized by λ. The redundancy indices associated with each member of this family (i.e., with each λ) are those which maximize

$$E\left[\sum_n U[n]\right] - \lambda E\left[\sum_n L[n]\right]. \tag{13}$$

The basic idea behind our formulation is as follows. In each transmission slot $\mathcal{T}[n]$, we select redundancy indices which maximize the Lagrange-style objective of (13), where the expectations are conditioned on whatever information is already known at the start of $\mathcal{T}[n]$, adjusting λ until the redundancy indices selected in this way satisfy the encoded length constraint for slot $\mathcal{T}[n]$ as tightly as possible, i.e., (14), shown at the bottom of the page. Note that $R(k_{n-\kappa}, r_q[n-\kappa], s_q[n-\kappa]) - R(r_q[n-\kappa])$ represents the ratio between the encoded length contribution to $\mathcal{T}[n]$ and the uncoded length of element $\mathcal{E}_q[n-\kappa]$.

Before moving to the optimization algorithm, it is worth mentioning that the actual value of λ which satisfies the length constraint in (14) may vary from slot to slot. This may happen for any number of reasons which cannot be foreseen ahead of time – e.g., variations in the source statistics, variations in the channel error rate, and so forth. As a result, the redundancy indices found in each successive transmission slot may be solutions to different global optimization problems, jeopardizing the global optimality of the solution found in any given transmission slot.

[4]Actually, the value of $s_q^k[n]$ is immaterial if $k \geq k_{\min}(r_q[n])$, since then no retransmission of $\mathcal{E}_q$ is required in slot $\mathcal{T}[n+\kappa]$. As a result, the requirement that $s_q^k[n] \geq s_{q+1}^k[n]$ strictly applies only for those $k < k_{\min}(r_q[n])$. Nevertheless, it remains true that the constraints of (11) can be imposed as is, without adversely affecting the expected utility.

One might attempt to account for this by generating statistical models for the evolution of λ in future transmission slots. Such modeling, however, is beyond the scope of the present paper. Instead, when formulating the expectations found in (10) and (12), we assume that the hypothetical redundancy indices $s_q^k[n]$ will be selected in accordance with the same global optimization objective, having the same value of λ as that selected for the present transmission slot. In [14], Chou and Miao also advocate the selection of transmission policies which maximize the global objective in (13). They suggest that λ may be interpreted as a type of quality parameter, which should be kept as constant as possible, subject to prevailing constraints on the encoded transmission rate.

A. Rewriting the Optimization Objective

Eliminating all terms from (13) which do not depend upon $r_q[n]$ and $s_q[n-\kappa]$, and writing $k_{n-\kappa}$ for the actual number of packets from slot $\mathcal{T}[n-\kappa]$ which were correctly received, we see that our objective in transmission slot $\mathcal{T}[n]$ is to maximize

$$\begin{aligned} &J_\lambda\left(\{r_q[n]\}_q, \{s_q^0[n]\}_q, \ldots, \{s_q^N[n]\}_q, \{s_q[n-\kappa]\}_q\right) \\ &= \sum_{k=0}^{N} \rho_k \sum_q U_q[n] P\left(k, r_q[n], s_q^k[n]\right) \\ &\quad + \sum_q U_q[n-\kappa] P\left(k_{n-\kappa}, r_q[n-\kappa], s_q[n-\kappa]\right) \\ &\quad - \lambda \sum_{k=0}^{N} \rho_k \sum_q L_q[n] R\left(k, r_q[n], s_q^k[n]\right) \\ &\quad - \lambda \sum_q L_q[n-\kappa] R\left(k_{n-\kappa}, r_q[n-\kappa], s_q[n-\kappa]\right) \end{aligned} \tag{15}$$

subject to the constraints

$$\begin{aligned} &r_q[n] \geq r_{q+1}[n], \ \forall q, \quad s_q^k[n] \geq s_{q+1}^k[n], \ \forall k, q \quad \text{and} \\ &s_q[n-\kappa] \geq s_{q+1}[n-\kappa], \ \forall q \end{aligned} \tag{16}$$

finding the smallest value of λ for which the actual transmission cost in slot $\mathcal{T}[n]$ satisfies (14).

For the sake of clarification, we note that the values of $r_q[n-\kappa]$ are known when we come to solve this optimization problem. For this reason, we omit the superscript from the redundancy indices $s_q[n-\kappa]$ which are not hypothetical. It is particularly worth noting that the optimization must be performed jointly over both the actual redundancy indices and the hypothetical redundancy indices, $s_q^k[n]$, even though the latter do not contribute to the length constraint in (14).

At this point, it is convenient to simplify the notation in (15), using primes to distinguish quantities which belong to frame $\mathcal{F}[n-\kappa]$ from those belonging to $\mathcal{F}[n]$. Thus, for example,

$$\sum_q L_q[n] R(r_q[n]) + \sum_q L_q[n-\kappa] \cdot [R(k_{n-\kappa}, r_q[n-\kappa], s_q[n-\kappa]) - R(r_q[n-\kappa])] \leq L_{\max} \tag{14}$$

$U'_q \equiv U_q[n-\kappa]$, $s'_q \equiv s_q[n-\kappa]$, $s_q^k \equiv s_q^k[n]$ and $k' \equiv k_{n-\kappa}$. With this notation, our optimization objective becomes

$$\begin{aligned} & J_\lambda\left(\{r_q\}_q, \{s_q^0\}_q, \ldots, \{s_q^N\}_q, \{s'_q\}_q\right) \\ &= \underbrace{\sum_{k=0}^{N} \rho_k \sum_q \left[U_q P\left(k, r_q, s_q^k\right) - \lambda L_q R\left(k, r_q, s_q^k\right)\right]}_{\Psi_\lambda\left(\{r_q\}_q, \{s_q^0\}_q, \ldots, \{s_q^N\}_q\right)} \\ &\quad + \underbrace{\sum_q \left[U'_q P\left(k', r'_q, s'_q\right) - \lambda L'_q R\left(k', r'_q, s'_q\right)\right]}_{J'_\lambda\left(\{s'_q\}_q\right)} \end{aligned} \tag{17}$$

and the parameter constraints become

$$r_q \geq r_{q+1}, \quad s_q^k \geq s_{q+1}^k, \quad \text{and} \quad s'_q \geq s'_{q+1}, \quad \forall k, q. \tag{18}$$

Note, also, that the task of maximizing J_λ subject to (18) is equivalent to the two independent tasks of maximizing J'_λ subject to $s'_q \geq s'_{q+1}$ and Ψ_λ subject to $r_q \geq r_{q+1}$ and $s_q^k \geq s_{q+1}^k$, where J'_λ and Ψ_λ are identified parenthetically in (17).

The remainder of this section is organized as follows. We first study the mechanics of retransmission, providing expressions for the terms $P(k, r_q, s_q)$ and $R(k, r_q, s_q)$. Next, we consider the solution to our optimization objective, in the simple case where each source frame has only one element. Finally, we show that the solution to the general problem is essentially no more complex than that developed for the single-element case since, under reasonable conditions, J_λ may be optimized separately for each q, and the resulting solutions are guaranteed to satisfy the constraints of (18).

B. Mechanics of Retransmission

Following the definition in (4), the number of packets which must be received if element $\mathcal{E}_q$ is to be recovered from its primary transmission slot is $k_{\min}(r_q)$. If $k \geq k_{\min}(r_q)$, there is no need for any retransmission. Suppose that $k < k_{\min}(r_q)$ primary transmission packets are received. Since the transmitter knows exactly which packets were received, it has only to retransmit the information from $\mathcal{E}_q$ which is contained within any $k_{\min}(r_q) - k$ of the $N - k$ packets which were lost. Thus, for the purpose of retransmission, the length of element $\mathcal{E}_q$ is effectively reduced to $\theta_{k,r_q} L_q$, where

$$\theta_{k,r_q} = \frac{k_{\min}(r_q) - k}{k_{\min}(r_q)} = 1 - \frac{k}{k_{\min}(r_q)}, \; 0 \leq k < k_{\min}(r_q).$$

Note that this expression is also valid for the special case $r_q = 0$, for which $k_{\min}(0) = \infty$ and $\theta_{k,0} = 1$.

During retransmission, the $\theta_{k,r_q} L_q$ source symbols from element $\mathcal{E}_q$ are assigned the secondary redundancy index s_q, leading to a transmission cost of $\theta_{k,r_q} L_q R(s_q)$, and an expected utility of $U_q P(s_q)$. The receiver is able to recover $\mathcal{E}_q$ so long as $k_{\min}(s_q)$ of the N packets sent during the element's secondary transmission slot are received. To do so, the receiver must first decode the secondary channel code, recovering the $\theta_{k,r_q} L_q$ missing source symbols which are required to decode the primary channel code. These procedures are illustrated in Fig. 6. The following expressions are readily obtained for $P(k, r_q, s_q)$ and $R(k, r_q, s_q)$:

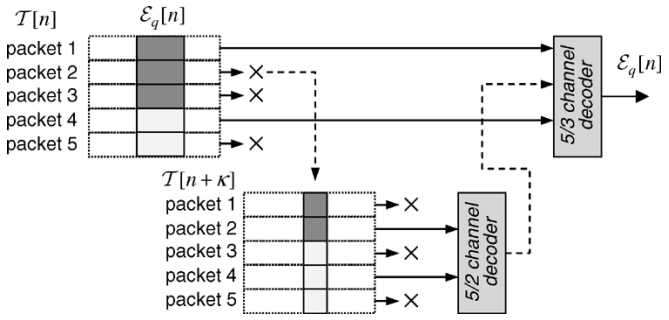

Fig. 6. Example showing how an element $\mathcal{E}_q[n]$ in source frame $\mathcal{F}[n]$ may be decoded from its primary transmission in slot $\mathcal{T}[n]$ and its secondary transmission in slot $\mathcal{T}[n+\kappa]$. The primary and secondary channel code redundancies are $R(r_q[n]) = 5/2$ and $R(s_q[n]) = 5/3$; only one third of the element's symbols must be retransmitted in the secondary slot.

$$P(k, r_q, s_q) = \begin{cases} P(s_q), & k < k_{\min}(r_q) \\ 1, & k \geq k_{\min}(r_q) \end{cases} \tag{19}$$

$$R(k, r_q, s_q) = R(r_q) + \begin{cases} \theta_{k,r_q} R(s_q), & k < k_{\min}(r_q) \\ 0, & k \geq k_{\min}(r_q) \end{cases}. \tag{20}$$

C. Single-Element Frames

To provide a more accessible introduction to the channel code assignment problem, we first consider the case in which each frame $\mathcal{F}[n]$ has only one source element. Let $\mathcal{E}_q[n]$ denote this source element; the subscript q is superfluous in this case, but we preserve it for the sake of later extension to source frames with multiple elements. Substituting the expressions from (19) and (20) into (17), we obtain

$$\begin{aligned} & \Psi_{\lambda,q}\left(r_q, s_q^0, \ldots, s_q^N\right) \\ &= \left(U_q \sum_{k=k_{\min}(r_q)}^{N} \rho_k\right) - \lambda R(r_q) L_q \\ &\quad + \sum_{k=0}^{k_{\min}(r_q)-1} \rho_k \left(U_q P\left(s_q^k\right) - \lambda \theta_{k,r_q} R\left(s_q^k\right) L_q\right) \\ &= U_q P(r_q) - \lambda R(r_q) L_q \\ &\quad + \sum_{k=0}^{k_{\min}(r_q)-1} \rho_k \left(U_q P\left(s_q^k\right) - \lambda \theta_{k,r_q} R\left(s_q^k\right) L_q\right) \end{aligned} \tag{21}$$

and

$$J'_{\lambda,q}\left(s'_q\right) = \begin{cases} U'_q P\left(s'_q\right) - \lambda R\left(r'_q\right) L'_q - \lambda \theta_{k',r'_q} R(s'_q) L'_q, & k' < k_{\min}(r'_q) \\ U'_q - \lambda R(r'_q) L'_q, & k' \geq k_{\min}(r'_q) \end{cases}.$$

Recall that k' and r'_q are known indices. Noting that s'_q is relevant only when $k' < k_{\min}\left(r'_q\right)$, the task of maximizing $J'_{\lambda,q}\left(s'_q\right)$ is equivalent to that of maximizing

$$U'_q P\left(s'_q\right) - \lambda' R\left(s'_q\right) L'_q, \quad \text{where} \quad \lambda' = \lambda \theta_{k',r'_q}.$$

As explained in Section III, the maximizing s'_q must be drawn from $\mathcal{H}_C$, the upper convex hull of the channel coding proba-

bility-redundancy characteristic. In fact, the maximizing s'_q is found using (9) to be

$$s'_q = \max\left\{ j_i \in \mathcal{H}_C \mid S_C(i)\frac{U'_q}{L'_q} \geq \lambda\theta_{k',r'_q} \right\}.$$

Similarly, for any given value of $r_q = r$, the hypothetical retransmission indices $s_q^{k,r}$ which maximize $\Psi_{\lambda,q}\left(r, s_q^{0,r}, \ldots s_q^{N,r}\right)$ are simply those which independently maximize the terms $U_q P\left(s_q^k\right) - \lambda\theta_{k,r} R\left(s_q^k\right) L_q$, the solution to which is

$$s_q^{k,r} = \max\left\{ j_i \in \mathcal{H}_C \mid S_C(i)\frac{U_q}{L_q} \geq \lambda\theta_{k,r} \right\}. \tag{22}$$

Here, we use the notation $s_q^{k,r}$ to express the dependence of the optimal hypothetical redundancy indices on the primary redundancy index, $r_q = r$. We are then left with the problem of determining r_q. Due to the complex interaction between r_q and the s_q^k, our current approach is simply an exhaustive search through the set $\mathcal{R}$ of all available redundancy indices. It is not generally sufficient to consider only those indices which belong to the convex hull, $\mathcal{H}_C \subset \mathcal{R}$. In the case of Reed–Solomon codes, the set $\mathcal{R}$ consists of all r in the range 0 through N. For the purpose of assessing complexity, we shall henceforth assume this maximally sized set of possible redundancy indices.

At first glance, an exhaustive search may appear to be computationally unattractive. For each of the $N+1$ possible indices $r \in \mathcal{R}$, we must solve the $k_{\min}(r) - 1 < N$ hypothetical optimization problems represented by (22), each of which involves up to $\|\mathcal{H}_C\| \leq N+1$ comparisons. This would suggest a complexity of at most $N(N+1)^2 \approx N^3$ comparisons. We can readily simplify this by exploiting the monotonicity of $S_C(i)$ in a binary search for the solution to (22), leaving us with a complexity of $N(N+1)\log_2\|\mathcal{H}_C\| \approx N^2 \log_2 N$ comparisons.

Fortunately, the exhaustive search may be even further simplified by observing that the $s_q^{k,r}$ values recovered using (22) are monotonically nondecreasing with r. To see this, observe that $k_{\min}(r)$ is a strictly decreasing function of r, hence $\theta_{k,r} = 1 - k/k_{\min}(r)$ is also strictly decreasing with r, but $s_q^{k,r}$ is a nonincreasing function of $\lambda\theta_{k,r}$. This property may be exploited by the following strategy for finding the $s_q^{k,r}$ for each k and each $r \in \mathcal{R}$.

For each k, $s_q^{k,r}$ is evaluated at $r = 0, 1, \ldots, N$ in sequence. Thus, when we come to evaluate $s_q^{k,r}$, we already know the value of $s_q^{k,r-1}$; we also know that $s_q^{k,r} \geq s_q^{k,r-1}$. The inequality in (22) is evaluated starting with the first $j_i \in \mathcal{H}_C$ which is greater than $s_q^{k,r-1}$, and continuing until some $j_i \in \mathcal{H}_C$ yields $S_C(i)\,U_q/L_q < \lambda\theta_{k,r}$. In this way, the inequality must be evaluated exactly $1 + \left\|\left(s_q^{k,r-1}, s_q^{k,r}\right] \cap \mathcal{H}_C\right\|$ times. It follows that the total number of times the inequality must be evaluated, in order to compute $s_q^{k,r}$ for all $r \in \mathcal{R}$, is bounded by

$$(N+1) + \left\|\left(0, s_q^{k,N}\right] \cap \mathcal{H}_C\right\| \leq (2N+1). \tag{23}$$

Actually, the bound on the right hand side of (23) can always be reduced by 1. To see this, note that the complexity expression on the left assumes that we always test the inequality of (22) until it fails. However, $\left\|\left(0, s_q^{k,N}\right] \cap \mathcal{H}_C\right\|$ can only equal N if $s_q^{k,N} = N$, in which case we will never perform a test for $s_q^{k,N}$ in which the inequality fails. The bound can, thus, be reduced to $2N$. This observation will prove more significant in Section IV-E. To conclude, we note that the number of values for k which must be considered is bounded by N. Thus, the total number of comparisons required to maximize $\Psi_{\lambda,q}$ is at most $2N^2$. A tighter bound, of the same order, could be developed by exploiting the fact that $s_q^{k,r}$ need not be evaluated for $k \geq k_{\min}(r)$.

D. Extension to Multiple Source Elements

We now tackle the general problem of maximizing the objective function J_λ in (17), subject to the constraints in (18). As in the single-element case, maximization of the second term $J'_\lambda(s'_1, s'_2, \ldots)$, subject to $s'_q \geq s'_{q+1}$, is an independent problem whose solution has already been described in Section III. For the first term, $\Psi_\lambda\left(\{r_q\}, \left\{s_q^k\right\}\right)$, coupling between the primary and hypothetical redundancy indices, r_q and s_q^k, would appear to present some difficulties. While exhaustive search through all possible r_q values is tractable for a single-element, the number of possible combinations for the indices r_1 through r_Q grows exponentially with Q. Fortunately, however, this is not necessary. Evidently

$$\Psi_\lambda\left(\{r_q\}_q, \left\{s_q^0\right\}_q, \ldots, \left\{s_q^N\right\}_q\right) = \sum_q \Psi_{\lambda,q}\left(r_q, s_q^0, \ldots s_q^N\right)$$

where the terms $\Psi_{\lambda,q}$ are given by (21). Thus, it is sufficient to maximize each $\Psi_{\lambda,q}$ independently, following the method outlined in Section IV-C, so long as we can be guaranteed that the independent solutions will satisfy the constraints in (18). We now show that this is the case, subject to some reasonable assumptions.

As in Section III, we assume that the source utility-length characteristic is convex,[5] satisfying (5). The following lemma shows that the hypothetical retransmission indices, s_q^k, will satisfy the necessary constraints so long as the primary redundancy indices r_q do. This simplifies our task to that of showing that the solutions which maximize each $\Psi_{\lambda,q}$ independently satisfy $r_q \geq r_{q+1}$, $\forall q$.

Lemma 1: Suppose the redundancy indices $\{r_q\}_q$ and $\left\{s_q^k\right\}_{q,k}$ which independently maximize each $\Psi_{\lambda,q}$ in (21) satisfy $r_q \geq r_{q+1}$, $\forall q$. Suppose also that the source utility-length characteristic is convex, following (5). Then $s_q^k \geq s_{q+1}^k$, $\forall q, k$.

Proof: We showed at the end of Section IV-C that $\theta_{k,r}$ is a strictly decreasing function of r. Using (22) then, we must have

$$\begin{aligned} s_q^k &= \max\left\{ j_i \in \mathcal{H}_C \mid \frac{U_q}{L_q} \geq \frac{\lambda\theta_{k,r_q}}{S_C(i)} \right\} \\ &\geq \max\left\{ j_i \in \mathcal{H}_C \mid \frac{U_q}{L_q} \geq \frac{\lambda\theta_{k,r_{q+1}}}{S_C(i)} \right\} \\ &\geq \max\left\{ j_i \in \mathcal{H}_C \mid \frac{U_{q+1}}{L_{q+1}} \geq \frac{\lambda\theta_{k,r_{q+1}}}{S_C(i)} \right\} = s_{q+1}^k. \end{aligned}$$

The second inequality above follows from the fact that the set of j_i for which $U_q/L_q \geq \lambda\theta_{k,r_{q+1}}/S_C(i)$ necessarily contains the set of j_i for which $U_{q+1}/L_{q+1} \geq \lambda\theta_{k,r_{q+1}}/S_C(i)$, since $U_{q+1}/L_{q+1} \leq U_q/L_q$. ■

[5]Alternatively, we first find the convex hull and then enforce the constraint that any element not on the convex hull must receive the same protection assignment as the next element which does lie on the convex hull.

To show that $r_q \geq r_{q+1}$, we begin by writing $\Psi_{\lambda,q}(r) = \Psi_{\lambda,q}\left(r, s_q^{0,r}, \ldots, s_q^{N,r}\right)$ where the $s_q^{k,r}$ are given by (22). That is, $\Psi_{\lambda,q}(r)$ is the maximum value which the objective function can attain, given that r_q is set to r. The following lemma embodies the key result which we must show.

Lemma 2: Let r_A and r_B be any two redundancy indices satisfying $r_A > r_B$ and suppose that $\Psi_{\lambda,q}(r_A) > \Psi_{\lambda,q}(r_B)$ for some $q > 1$. Then we must also have $\Psi_{\lambda,q-1}(r_A) > \Psi_{\lambda,q-1}(r_B)$, so long as the source utility-length characteristic is convex, satisfying (5), and the set of channel code redundancies which belong to $\mathcal{H}_C$ is sufficiently dense.

We defer the proof of this key result to the Appendix. The last condition, however, deserves some comment here. Our proof currently relies upon passing to a continuous model for the convex hull of the channel coding probability-redundancy characteristic. This indicates that the result might hold only when a very dense collection of channel codes is available. In practice, extensive numerical studies suggest that the result always holds, even in the discrete case, but we do not have a formal proof that this must be true. Our idealized proof in the appendix at least shows that enforcing the constraint $r_q \geq r_{q+1}$ while passing from the optimization of $\Psi_{\lambda,q}$ to that of $\Psi_{\lambda,q+1}$ is unlikely to produce sub-optimal results, and that any such likelihood diminishes as the available set of channel code redundancies increases.

Corollary 3: If $r = r_A$ maximizes $\Psi_{\lambda,q}(r)$ then it is always possible to find $r = r_B$ which maximizes $\Psi_{\lambda,q+1}(r)$ such that $r_B \leq r_A$.

Proof: Suppose, to the contrary, that any $r = r_B$ which maximizes $\Psi_{\lambda,q+1}(r)$ has $r_B > r_A$. This means, in particular, that $\Psi_{\lambda,q+1}(r_B) > \Psi_{\lambda,q+1}(r_A)$. According to Lemma 2 then, we must also have $\Psi_{\lambda,q}(r_B) > \Psi_{\lambda,q}(r_A)$ so that r_A cannot be the value of r which maximizes $\Psi_{\lambda,q}(r)$. ■

E. Complexity With Multiple Source Elements

We saw in Section IV-C that the inequality in (22) must be evaluated at most $2N^2$ times in order to maximize $\Psi_{\lambda,q}$. Since each source frame has Q elements, it is clear that all Q optimization problems can be solved using at most $2N^2Q$ comparisons. For large Q, however, we can exploit the monotonicity relationships of Lemma 1 and Corollary 3 to substantially reduce this complexity. This can be achieved using the following "branch and bound" strategy.

We start by solving the single-element optimization problem associated with element $\mathcal{E}_{q_1}$, where $q_1 \approx Q/2$, at a cost of $2N^2$ comparisons (we are ignoring the much lower cost of maximizing the $J'_{\lambda,q}$ terms here). The solution to this problem divides the set of redundancy indices which may be selected for r_q and s_q^k into two halves, depending on whether $q < q_1$ or $q > q_1$. We next solve the two single-element optimization problems associated with $q_2 \approx Q/4$ and $q_3 \approx 3Q/4$. In solving these problems, we need only consider $r_{q_2} \in [r_{q_1}, N]$, $r_{q_3} \in [0, r_{q_1}]$, $s_{q_2}^k \in \left[s_{q_1}^k, N\right]$ and $s_{q_3}^k \in \left[0, s_{q_1}^k\right]$. A trivial modification of the argument provided at the end of Section IV-C shows that $s_{q_2}^{k,r}$ may be evaluated for all $r \in [r_{q_1}, N]$ using at most $(N - r_{q_1} + 1) + \left\|\left(s_{q_1}^k, s_{q_2}^{k,N}\right] \cap \mathcal{H}_C\right\|$ comparisons, while $s_{q_3}^{k,r}$ may be evaluated for all $r \in [0, r_{q_1}]$ using at most $(r_{q_1} + 1) + \left\|\left(0, s_{q_1}^k\right] \cap \mathcal{H}_C\right\|$ comparisons. The total number of comparisons required to evaluate both $s_{q_2}^{k,r}$ and $s_{q_3}^{k,r}$ for all relevant r is, thus, bounded by $(N+2) + (\|\mathcal{H}_C\| - 1) \leq 2N + 2$.

Following the same argument as in Section IV-C, we note that the bound of $2N+2$ can always be reduced by 2. This is because the above complexity expressions are based on the assumption that we always test the inequality in (22) until it fails, but there is no need to test $s_{q_2}^{k,r}$ and $s_{q_3}^{k,r}$ values which exceed the known bounds of N and $s_{q_1}^k$, respectively. As before, we note that the number of values for k which must be considered is bounded by N (although tighter bounds can be found), so that the total number of comparisons required to maximize both Ψ_{λ,q_2} and Ψ_{λ,q_3} is at most $2N^2$.

We proceed by solving the four single-element optimization problems associated with $q_4 \approx Q/8$, $q_5 \approx 3Q/8$, $q_6 \approx 5Q/8$ and $q_7 \approx 7Q/8$. The reader may verify that the total number of comparisons required to evaluate $s_{q_4}^{k,r}$, $s_{q_5}^{k,r}$, $s_{q_6}^{k,r}$ and $s_{q_7}^{k,r}$, for all relevant r, is again given by $2N^2$. Continuing in this way, after a total of V subdivision steps, we solve a total of $2^V - 1$ single-element optimization problems using at most $2N^2V$ comparisons. Associating $2^V - 1$ with Q, we see that the total complexity is of order $2N^2 \log_2(Q+1)$, which is substantially less than $2N^2Q$, for large Q.

The above analysis is concerned only with maximization of the $\Psi_{\lambda,q}$ terms. As already noted, maximization of $J'_{\lambda,q}$ is significantly simpler, having complexity no larger than that of regular frame-by-frame PET. Of course, the optimization of $\Psi_{\lambda,q}$ and $J'_{\lambda,q}$ must be conducted within an outer loop, which adjusts λ until (14) is satisfied sufficiently tightly. While more sophisticated λ search strategies exist, a simple bisection search requires only one iteration for each bit in the numerical representation with which we choose for λ. For most practical applications, a logarithmic fixed point representation of λ suffices, with between 8 and 16 bits of precision. Chou and Miao [14] suggest an alternate rate control strategy, in which λ is adapted slowly, subject to a conventional "leaky bucket" model which constrains the average transmission rate. In such a scenario, $\Psi_{\lambda,q}$ and $J'_{\lambda,q}$ must be optimized for only one value of λ within any given transmission slot.

F. Experimental Results

We present experimental results in two parts. The first part compares the performance of frame-by-frame PET with that of LR-PET, using a retransmission delay of $\kappa = 2$. The second part investigates the effect of the transmission delay κ on the performance of the proposed LR-PET scheme.

For our scalable data source, we use a video sequence whose frames have been independently compressed using JPEG2000; i.e., a Motion JPEG2000 video stream. The video sequence consists of 30 monochrome frames, with 720P resolution. That is, each frame has a progressive scan with 720 rows and 1280 columns. Each JPEG2000-compressed frame has 15 precincts and 12 quality layers, for a total of $Q = 180$ elements, $\mathcal{E}_q[n]$. We choose a code length of $N = 30$, which is also the number of packets in each PET frame. The maximum transmission length is set to $L_{\max} = 230{,}000$ bytes, and we consider various packet loss probabilities in the range $0.1 \leq p \leq 0.5$.

For the PET results, we use the algorithm described in Section III to independently assign an optimal set of channel codes

to each source frame. In this case, there is no opportunity for retransmission so each source frame $\mathcal{F}[n]$ occupies its own transmission slot $\mathcal{T}[n]$, with its own PET frame of N packets. For the LR-PET results, each frame is assigned both a primary transmission slot and a secondary retransmission slot. The algorithm described in Section IV is used to optimize the distribution of transmission bandwidth between the primary transmission for frame $\mathcal{F}[n]$ and secondary retransmission for frame $\mathcal{F}[n-k]$, assigning appropriate redundancy indices $r_q[n]$ and $s_q[n-\kappa]$.

We use negative mean-squared error (MSE) as our measure of utility, taking averages over a large number of experiments. The actual number of experiments is adjusted according to the packet loss probability, so as to minimize the impact of nonrepresentative outcomes. For LR-PET, the utility of each frame is evaluated after its secondary transmission. Although our source material consists of only 30 frames, we cycle through these frames many times, effectively creating a much larger sequence, so that all but a few frames at the beginning of the sequence have to share their primary transmission with another frame's secondary transmission. When computing average utilities, we ignore these initial frames, as well as those at the end of the sequence which have no retransmission event. Finally, we find it convenient to express the utility values, after averaging, in terms of peak signal-to-noise rate (PSNR).[6]

Fig. 7 compares LR-PET with PET at a variety of different packet loss probabilities, p. The results in this figure are obtained by averaging the utilities at all frames in the video sequence. This figure clearly reveals the benefits of limited retransmission. At a packet loss probability of $p = 0.3$, the availability of just one opportunity for retransmission improves the PSNR by 4.2 dB. The benefits of retransmission appear to decline at lower and higher packet loss rates. At lower packet loss rates, most packets can be received without error, so that little protection is required and most elements are recovered from their primary transmission. We do not currently have any explanation for the apparent decline in retransmission benefits at high packet loss rates, although at $p = 0.5$ the advantage is still 2.5 dB in this experiment.

Fig. 8 provides a more detailed view of the behavior of LR-PET at different frame instants and with various retransmission delays, κ. For these results, the packet loss probability is fixed at $p = 0.3$. MSE values for each frame of the original 30 frame sequence are averaged over all packet loss experiments and all occurrences of the frame within the extended video. As before, averaged MSE values are expressed in terms of PSNR. In the absence of channel errors, our original source content experiences only relatively small fluctuations in PSNR as a function of frame number, and the same is observed in the PET and LR-PET results when channel errors are present. Of greater interest is the fact that the benefits gained from retransmission are substantially insensitive to the retransmission delay κ.

One might at first suspect that the performance of LR-PET should be independent of retransmission delay, since κ does not enter directly into the expected utility expression of (10) and is not involved in the optimization objective of (17). As discussed toward the beginning of Section IV, however, our hypotheses regarding retransmission are formulated based on the assumption that λ will not change between a frame's primary and secondary transmission slot. The larger the retransmission delay κ, the more likely it is that the λ values selected in a frame's primary and secondary transmission slot will differ significantly. In particular, we expect λ to be sensitive to changes in the source statistics from slot to slot. In the example of Fig. 8, λ evolves only slowly, with typical frame-to-frame variations below 10% and an overall range from 0.8 to 1.4. This observation explains the low sensitivity to retransmission delay observed in this particular experiment.

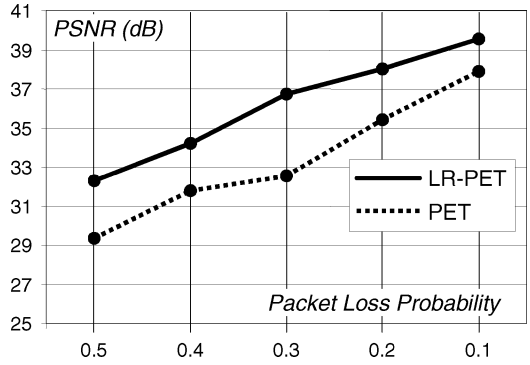

Fig. 7. Comparison between LR-PET with retransmission delay $\kappa = 2$, and PET without retransmission, for various packet loss probabilities p.

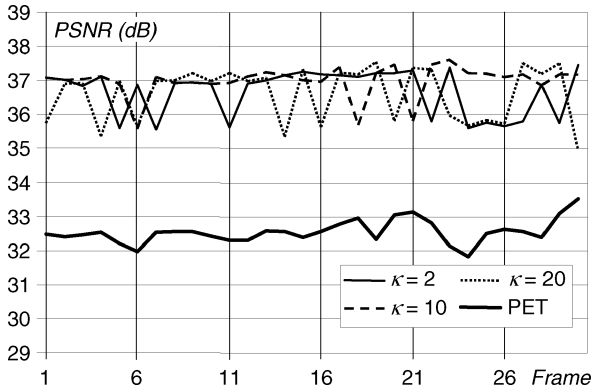

Fig. 8. Effect of retransmission delay on average LR-PET utility (expressed as PSNR) at each frame instant.

V. Conclusion

Previous work on optimal protection assignment strategies using the PET framework has largely assumed that retransmission of lost data is not possible. Relaxing this assumption, we propose the LR-PET framework for jointly optimizing the protection assigned to source elements which are being transmitted for the first time and elements which are being retransmitted. In the context of a real-time service with a single opportunity for retransmission, we provide an LR-PET optimization algorithm. The algorithm exhibits a complexity of order $N^2 \log_2 Q$, for N packets and Q source elements; this is remarkably low compared to that of other schemes proposed for the RD optimization of hybrid FEC/ARQ protocols. Our experimental results

[6]PSNR is defined as $10 \log_{10}(A^2/\mathrm{MSE})$, where A is the peak-to-peak signal amplitude. In this case, $A = 255$, since we are working with 8-bit video samples.

reveal large performance advantages from the inclusion of limited retransmission within LR-PET. We envisage that this framework may form the basis for much future work. Possible directions for such future work include the protection of streaming sources with interframe dependencies, extension to more general retransmission environments with larger or unpredictable numbers of possible retransmission opportunities, and experimental work with non-IID packet loss models.

APPENDIX
PROOF OF LEMMA 2

A. Proof

We are given $r_A > r_B$ and $\Psi_{\lambda,q}(r_A) > \Psi_{\lambda,q}(r_B)$, from which we wish to show that $\Psi_{\lambda,q-1}(r_A) > \Psi_{\lambda,q-1}(r_B)$. Writing $S_q = U_q/L_q$ and noting that $S_{q-1} \geq S_q$, it is sufficient to show that

$$\Psi_{A-B}(S_q) = \frac{\Psi_{\lambda,q}(r_A) - \Psi_{\lambda,q}(r_B)}{L_q}$$

is an increasing function of S_q. Using (21), we expand $\Psi_{A-B}(S_q)$ as

$$\begin{aligned}\Psi_{A-B}(S_q) = &\left(S_q \sum_{k=k_{\min}(r_A)}^{N} \rho_k\right) - \lambda R(r_A) \\ &+ \sum_{k=0}^{k_{\min}(r_A)-1} \rho_k\left(S_q P\left(s_q^{k,r_A}\right) - \lambda\theta_{k,r_A} R\left(s_q^{k,r_A}\right)\right) \\ &- \left(S_q \sum_{k=k_{\min}(r_B)}^{N} \rho_k\right) - \lambda R(r_B) \\ &+ \sum_{k=0}^{k_{\min}(r_B)-1} \rho_k\left(S_q P\left(s_q^{k,r_B}\right) - \lambda\theta_{k,r_B} R\left(s_q^{k,r_B}\right)\right).\end{aligned}$$

Now recall that $P(r) = \sum_{k=k_{\min}(r)}^{N} \rho_k$, so we can write $R(r)$ as $\sum_{k=k_{\min}(r)}^{N} \rho_k R(r)/P(r)$. This, together with the fact that $k_{\min}(r_A) < k_{\min}(r_B)$ allows us to express $\Psi_{A-B}(S_q)$ as a sum of terms, weighted by ρ_k; specifically, see the equation shown at the bottom of the page. Note that the terms $s_q^{k,r}$ depend only on k, r, λ, and S_q. Specifically, from (22), we have

$$s_q^{k,r} = \max\{j_i \in \mathcal{H}_C \mid S_q S_C(i) \geq \lambda\theta_{k,r}\}. \qquad (24)$$

To complete the proof, we show that each term in the summations appearing at the bottom of the page is an increasing function of S_q. The terms in the third summation do not depend on S_q, either directly or through s_q^{k,r_A} and s_q^{k,r_B}, so we need only concern ourselves with the first two summations. Specifically, we show that

$$\begin{aligned}W_{k,r_A,r_B}(S_q) &= S_q\left[P\left(s_q^{k,r_A}\right) - P\left(s_q^{k,r_B}\right)\right] \\ &\quad - \lambda\left[\theta_{k,r_A} R\left(s_q^{k,r_A}\right) - \theta_{k,r_B} R\left(s_q^{k,r_B}\right)\right] \\ \text{and } Z_{k,r_B}(S_q) &= S_q\left[1 - P\left(s_q^{k,r_B}\right)\right] + \lambda\theta_{k,r_B} R\left(s_q^{k,r_B}\right)\end{aligned}$$

are both increasing functions of S_q.

As noted in connection with the statement of Lemma 2, we find it necessary to pass to a continuous model of the channel coding probability-redundancy characteristic on its convex hull. This allows us to express (24) as

$$S_q G_{k,r}(S_q) = \lambda\theta_{k,r}, \quad \text{where} \quad G_{k,r}(S_q) = \frac{\partial P}{\partial R}\left(s_q^{k,r}\right). \qquad (25)$$

Here, we are using the fact that $S_C(i)$ is the slope of the channel coding probability-redundancy characteristic on its convex hull, as expressed by (8). In the limit, as the density of available channel codes grows without bound, we may replace $S_C(i)|_{j_i = s_q^{k,r}}$ with the derivative, $\partial P/\partial R\left(s_q^{k,r}\right)$. The notation $G_{k,r}(S_q)$ makes explicit the fact that $s_q^{k,r}$, and, hence, this derivative, depends only on k, r and S_q. In the same way, it is convenient to write $P_{k,r}(S_q)$ for $P\left(s_q^{k,r}\right)$ and $R_{k,r}(S_q)$ for $R_{k,r}\left(s_q^{k,r}\right)$, from which we obtain

$$\begin{aligned}W_{k,r_A,r_B}(S) &= S\cdot\Big[\left(P_{k,r_A}(S) - P_{k,r_B}(S)\right) \\ &\quad - \Big(G_{k,r_A}(S) R_{k,r_A}(S) - G_{k,r_B}(S) R_{k,r_B}(S)\Big)\Big] \\ &= S\cdot\Big[\left(P_{k,r_A}(S) - G_{k,r_A}(S) R_{k,r_A}(S)\right) \\ &\quad - \Big(P_{k,r_B}(S) - G_{k,r_B}(S) R_{k,r_B}(S)\Big)\Big] \text{ and} \\ Z_{k,r_B}(S) &= S\cdot\left[1 - P_{k,r_B}(S) + G_{k,r_B}(S) R_{k,r_B}(S)\right].\end{aligned}$$

Taking the derivative of $W_{k,r_A,r_B}(S)$ with respect to S, and using the fact that

$$\frac{\partial P_{k,r}}{\partial S} = \frac{\partial P_{k,r}}{\partial R_{k,r}} \frac{\partial R_{k,r}}{\partial S} = G_{k,r} \frac{\partial R_{k,r}}{\partial S}$$

$$\begin{aligned}\Psi_{A-B}(S_q) = &\sum_{k=0}^{k_{\min}(r_A)-1} \rho_k\left(S_q\left[P\left(s_q^{k,r_A}\right) - P\left(s_q^{k,r_B}\right)\right] - \lambda\left[\theta_{k,r_A} R\left(s_q^{k,r_A}\right) - \theta_{k,r_B} R\left(s_q^{k,r_B}\right)\right]\right) \\ &+ \sum_{k=k_{\min}(r_A)}^{k_{\min}(r_B)-1} \rho_k\left(S_q\left[1 - P\left(s_q^{k,r_B}\right)\right] - \lambda\left[\frac{R(r_A)}{P(r_A)} - \theta_{k,r_B} R\left(s_q^{k,r_B}\right)\right]\right) \\ &+ \sum_{k=k_{\min}(r_B)}^{N} \rho_k\left(-\lambda\left[\frac{R(r_A)}{P(r_A)} - \frac{R(r_B)}{P(r_B)}\right]\right)\end{aligned}$$

we get

$$\begin{aligned}\frac{\partial W_{k,r_A,r_B}}{\partial S} &= (P_{k,r_A}-G_{k,r_A}R_{k,r_A})-(P_{k,r_B}-G_{k,r_B}R_{k,r_B}) \\ &\quad + S\left(\frac{\partial P_{k,r_A}}{\partial S}-R_{k,r_A}\frac{\partial G_{k,r_A}}{\partial S}-G_{k,r_A}\frac{\partial R_{k,r_A}}{\partial S}\right) \\ &\quad - S\left(\frac{\partial P_{k,r_B}}{\partial S}-R_{k,r_B}\frac{\partial G_{k,r_B}}{\partial S}-G_{k,r_B}\frac{\partial R_{k,r_B}}{\partial S}\right) \\ &= (P_{k,r_A}-G_{k,r_A}R_{k,r_A})-(P_{k,r_B}-G_{k,r_B}R_{k,r_B}) \\ &\quad - SR_{k,r_A}\frac{\partial G_{k,r_A}}{\partial S}+SR_{k,r_B}\frac{\partial G_{k,r_B}}{\partial S}.\end{aligned}$$

Differentiating (25) yields $G_{k,r} = -S(\partial G_{k,r}/\partial S)$, which we substitute into the above to obtain

$$\begin{aligned}\frac{\partial W_{k,r_A,r_B}}{\partial S} &= (P_{k,r_A}-G_{k,r_A}R_{k,r_A})-(P_{k,r_B}-G_{k,r_B}R_{k,r_B}) \\ &\quad + R_{k,r_A}G_{k,r_A}-R_{k,r_B}G_{k,r_B} \\ &= P_{k,r_A}-P_{k,r_B} \\ &= P\left(s_q^{k,r_A}\right)-P\left(s_q^{k,r_B}\right)\geq 0.\end{aligned}$$

Here, we have used the fact that $s_q^{k,r}$ is a nondecreasing function of r, as shown in Section IV-C and that $P(s)$ is an increasing function of s. Similarly, taking the derivative of $Z_{k,r_B}(S)$ with respect to S, we also find that

$$\begin{aligned}\frac{\partial Z_{k,r_B}}{\partial S} &= [1-P_{k,r_B}+G_{k,r_B}R_{k,r_B}] \\ &\quad + S\left[-\frac{\partial P_{k,r_B}}{\partial S}+G_{k,r_B}\frac{\partial R_{k,r_B}}{\partial S}+R_{k,r_B}\frac{\partial G_{k,r_B}}{\partial S}\right] \\ &= 1-P_{k,r_B}+G_{k,r_B}R_{k,r_B}+SR_{k,r_B}\frac{\partial G_{k,r_B}}{\partial S} \\ &= 1-P_{k,r_B}+G_{k,r_B}R_{k,r_B}-G_{k,r_B}R_{k,r_B} \\ &= 1-P_{k,r_B}>0.\end{aligned}$$

■

REFERENCES

[1] R. Puri and K. Ramchandran, "Multiple description source coding using forward error correction codes," in *Proc. 33rd Asilomar Conf. Signals, Systems, and Computers*, vol. 1, Oct. 1999, pp. 342–346.

[2] A. Mohr, E. Riskin, and R. Ladner, "Unequal loss protection: Graceful degradation of image quality over packet erasure channels through forward error correction," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 6, pp. 819–828, Jun. 2000.

[3] A. Mohr, R. Ladner, and E. Riskin, "Approximately optimal assignment for unequal loss protection," in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, Sep. 2000, pp. 367–370.

[4] T. Stockhammer and C. Buchner, "Progressive texture video streaming for lossy packet networks," presented at the 11th Int. Packet Video Workshop, May 2001.

[5] S. Dumitrescu, X. Wu, and Z. Wang, "Globally optimal uneven error-protected packetization of scalable code streams," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Apr. 2002, pp. 73–82.

[6] V. Stankovic, R. Hamzaoui, and Z. Xiong, "Packet loss protection of embedded data with fast local search," in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, Sep. 2002, pp. 165–168.

[7] J. Thie and D. Taubman, "Optimal protection assignment for scalable compressed images," in *Proc. IEEE Int. Conf. Image Processing*, vol. 3, Sep. 2002, pp. 713–716.

[8] ——, "Optimal erasure protection assignment for scalable compressed data with small channel packets and short channel codewords," presented at the EURASIP JASP: Multimedia over IP and Wireless Networks, 2004.

[9] A. Albanese, J. Blomer, J. Edmonds, M. Luby, and M. Sudan, "Priority encoding transmission," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 1737–1744, Nov. 1996.

[10] M. Podolsky, M. Vetterli, and S. McCanne, "Limited retransmission of real-time layered multimedia," in *Proc. IEEE 2nd Workshop Multimedia Signal Processing*, Dec. 1998, pp. 591–596.

[11] M. Podolsky, S. McCanne, and M. Vetterli, "Soft ARQ for layered streaming media," *J. VLSI Signal Process. Signal, Image, Video Technol.*, vol. 27, pp. 81–97, Feb. 2001.

[12] V. Chande, N. Farvardin, and H. Jafarkhani, "Image communication over noisy channels with feedback," in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, Oct. 1999, pp. 540–544.

[13] C.-Y. Hsu and A. Ortega, "Rate control for robust video transmission over burst-error wireless channels," *IEEE J. Sel. Areas Comm.*, vol. 17, no. 5, pp. 756–773, May 1999.

[14] P. Chou and Z. Miao, "Rate-distortion optimized streaming of packetized media," Microsoft Tech. Rep. MSR-TR-2001–35, 2001.

[15] G. Cheung, W.-t. Tan, and T. Yoshimura, "Rate-distortion optimized application-level retransmission using streaming agent for video streaming over 3g wireless network," in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, Sep. 2002, pp. 529–532.

[16] R. Puri, K. Ramchandran, and A. Ortega, "Joint source channel coding with hybrid FEC/ARQ for buffer constrained video transmission," in *IEEE 2nd Workshop Multimedia Signal Processing*, Dec. 1998, pp. 567–572.

[17] P. Chou, A. Mohr, A. Wang, and S. Mehrotra, "FEC and pseudo-ARQ for receiver-driven layered multicast of audio and video," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 2000, pp. 440–449.

[18] J. Chakareski, P. Chou, and B. Aazhang, "Computing rate-distortion optimized policies for streaming media to wireless clients," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Apr. 2002, pp. 53–62.

[19] G. Wang, Q. Zhang, W. Zhu, and Y.-Q. Zhang, "Channel-adaptive error control for scalable video over wireless channel," presented at the 7th Int. Workshop on Mobile Multimedia Communications (MoMuC), Oct. 2000.

[20] T. Stockhammer, H. Jenkac, and C. Weiss, "Feedback and error protection strategies for wireless progressive video transmission," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 465–482, Jun. 2002.

[21] X. Zheng, S.-H. Chan, Q. Zhang, W.-W. Zhu, and Y.-Q. Zhang, "Feedback-free packet loss recovery for video multicast," in *IEEE Int. Conf. Communications*, vol. 2, Apr. 2003, pp. 870–874.

[22] T. Gan and K.-K. Ma, "Sliding-window packetization for forward error correction based multiple description transcoding," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 5, Apr. 2003, pp. 756–759.

[23] ——, "Sliding-window packetization for unequal loss protection based multiple description coding," in *Proc. IEEE Int. Conf. Image Processing*, vol. 3, Sep. 2003, pp. 641–644.

[24] J. Shapiro, "An embedded hierarchical image coder using zerotrees of wavelet coefficients," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, 1993, pp. 214–223.

[25] A. Said and W. Pearlman, "A new, fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 6, pp. 243–250, Jun. 1996.

[26] MAD. Taubman and M. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Boston: Kluwer, 2002.

[27] D. Taubman and A. Zakhor, "Multi-rate 3-D subband coding of video," *IEEE Trans. Image Processing*, vol. 3, no. 5, pp. 572–588, May 1994.

[28] J. Li and S. Lei, "Rate-distortion optimized embedding," in *Proc. Picture Coding Symp.*, Berlin, Germany, Sep. 1997, pp. 201–206.

[29] E. Ordentlich, M. Weinberger, and G. Seroussi, "A low-complexity modeling approach for embedded coding of wavelet coefficients," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 1998, pp. 408–417.

[30] D. Taubman, "High performance scalable image compression with EBCOT," *IEEE Trans. Image Processing*, vol. 9, no. 7, pp. 1158–1170, Jul. 2000.

[31] J. Ohm, "Three dimensional subband coding with motion compensation," *IEEE Trans. Image Processing*, vol. 3, no. 4, pp. 559–571, Apr. 1994.

[32] D. Taubman and A. Zakhor, "Highly scalable low-delay video compression," in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, Nov. 1994, pp. 740–744.

[33] S. Choi and J. Woods, "Motion compensated 3-D subband coding of video," *IEEE Trans. Image Processing*, vol. 8, no. 2, pp. 155–167, Feb. 1999.

[34] A. Secker and D. Taubman, "Motion-compensated highly scalable video compression using an adaptive 3D wavelet transform based on lifting," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 2001, pp. 1029–1032.

[35] B. Pesquet-Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 2001, pp. 1793–1796.

[36] A. Secker and D. Taubman, "Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression," *IEEE Trans. Image Processing*, vol. 12, no. 12, pp. 1530–1542, Dec. 2003.
[37] G. Van der Auwera, P. Munteanu, P. Schelkens, and J. Cornelis, "Bottom-up motion compensated prediction in wavelet domain for spatially scalable video coding," *Electron. Lett.*, vol. 38, no. 21, pp. 1251–1253, Oct. 2002.
[38] D. Taubman, "Successive refinement of video: Fundamental issues, past efforts and new directions," in *Proc. Int. Symp. Visual Communication and Image Processing*, vol. 5150, Jul. 2003, pp. 791–805.
[39] E. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell Syst. Tech. J.*, vol. 42, pp. 1977–1997, Sep. 1963.

David Taubman (M'92) received the B.S. and B.Eng. degrees from the University of Sydney, Sydney, Australia, in 1986 and 1988, respectively, and the M.S. and Ph.D. degrees from the University of California, Berkeley, in 1992 and 1994, respectively.

From 1994 to 1998, he was with Hewlett-Packard's Research Laboratories, Palo Alto, CA, joining the University of New South Wales, Sydney, in 1998, where he is an Associate Professor with the School of Electrical Engineering and Telecommunications. He is the coauthor, with M. Marcellin, of the book *JPEG2000: Image Compression Fundamentals, Standards and Practice* (Boston, MA: Kluwer, 2001). His research interests include highly scalable image and video compression, inverse problems in imaging, perceptual modeling, joint source/channel coding, and multimedia distribution systems.

Dr. Taubman was awarded the University Medal from the University of Sydney; the Institute of Engineers, Australia, Prize; and the Texas Instruments Prize for Digital Signal Processing, all in 1998. He has received two Best Paper awards, one from the IEEE Circuits and Systems Society for the 1996 paper, "A Common Framework for Rate and Distortion Based Scaling of Highly Scalable Compressed Video," and from the IEEE Signal Processing Society for the 2000 paper, "High Performance Scalable Image Compression with EBCOT."

Johnson Thie received the B.E. (electrical) and M.Biomed.E. (biomedical) degrees in 2000 and the Ph.D. (electrical) degree in 2004 from the University of New South Wales, Sydney, Australia.

For the summers of 1998 and 1999, he was with the Centre of Telecommunications and Industrial Physics (CSIRO), Sydney. Currently, he is a Research Engineer at Emotiv Systems, Sydney. His research interests include the distribution of compressed media over packet-based networks and applications of signal/image processing in biomedical engineering.