Optimal Media Streaming in a Rate-Distortion Sense For Guaranteed Service Networks

Olivier Verscheure and Pascal Frossard
IBM Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598
USA

Jean-Yves Le Boudec
Swiss Federal Institute of Tech.
Lausanne 1015
Switzerland

Abstract

We present an optimal low-complexity scheduling strategy of continuous media in a rate-distortion sense for guaranteed service networks. In this context, the output of the smoother is constrained by the traffic envelope defined at the network entry point, the guaranteed service curve, the playback delay budget and the decoding buffer size. First we consider a stored and offline-compressed media stream. We tackle the problem of whether there exists one optimal strategy at the smoother which minimizes the playback delay and the receive buffer size, given the traffic envelope and the service curve. We show that there does exist such an optimal smoothing strategy, and give an explicit representation for it. We also obtain a simple expression for the smallest playback delay and playback buffer size which can be achieved over all possible smoothing and playback strategies. Then we provide the theoretical bounds on the media rate such that (i) the optimal smoothing solution meets some constraints on the admissible playback delay and maximum decoding buffer size, and (ii) the media size is maximum. This set of bounds leads to a useful separation principle, which allows us to consider scheduling and coding as two independent processes. Thus we cast the rate-distortion problem as a piece-wise linear convex optimization algorithm, which is solved efficiently using state-of-the-art linear programming techniques. Finally, experimental results exhibit significant improvements in terms of total average distortion compared to the smoothing of a fixed media encoder output, under equivalent traffic parameters and decoding constraints.

1 Introduction

We consider the transmission of *variable bit rate* (VBR) media streams over a network offering a guaranteed service such as ATM VBR or the guaranteed service of the IETF [1]. The

guaranteed service class provides firm end-to-end delay guarantees. This service guarantees both delay and bandwidth. The guaranteed service requires that the flow produced by the output device conforms with a traffic envelope σ, namely over any window of size t, the amount of data does not exceed $\sigma(t)$. With the Resource Reservation Protocol (RSVP), σ is derived from the T-SPEC field in messages used for setting up the reservation, and is given by $\sigma(t) = \min(M + pt, rt + b)$, where M is the maximum packet size, P the peak rate, r the sustainable rate and b the burst tolerance [2]. The function σ is also called an arrival curve.

The media streamer must thus produce an output conforming with the arrival curve constraint. One approach for achieving this is called *rate control* [3]. It consists in modifying the encoder output, by acting on the quantization parameters. Rate control is a delicate issue in video coding since it significantly affects the rendered quality. An alternative approach is to smooth the multimedia stream, using a smoother fed by the media multiplexer [4, 5]. This work combines both approaches.

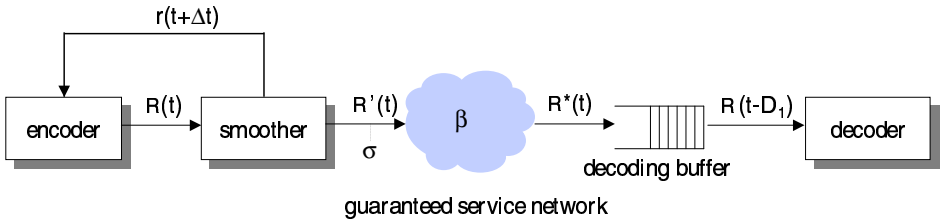

Figure 1: Scenario and notation used in this paper.

Our scenario is illustrated on Figure 1. A video signal is encoded, and then input into a smoother. The smoother writes the stream into a network for transmission. The smoother possibly feedbacks the optimal channel rate for the next time interval $(t + \Delta t)$. We call $R(t)$ the total number of bits observed on the encoded flow, starting from time $t = 0$, and $R'(t)$ the output of the smoother. The smoother output must satisfy the traffic envelope constraint given by some function σ negotiated with the network. At the destination, the receiver stores incoming bits into a decoding buffer before passing them to the decoder. The decoder starts reading from the decoding buffer after a delay D, and then reads the decoding buffer so as to reproduce the original signal, shifted in time. Thus the output of the decoding buffer is equal to $R(t-D_1)$, where D_1 is equal to D plus the transfer time for the first packet of the flow. The delay D is called *playback delay* at the receiver.

We assume that the network offers to the flow R' a guaranteed service, such as defined for example by the IETF. Call $R^*(t)$ the cumulative function at the output of the network. The transformation $R' \rightarrow R^*$ can be decomposed into a fixed delay, and a variable delay. Without loss of generality, we can reduce to the case where the fixed delay is zero, since it does not impact the smoothing method. The variable delay is due to queuing in, for example, guaranteed rate schedulers. The relationship between R' and R^* cannot be known exactly by the sending side, because it depends to some extend on traffic conditions; however, the

guarantee provided by the network can be formalized by a condition of the form [6, 7]

$$\forall t \geq 0, \exists s \leq t, \text{ such that } R^*(t) \geq R'(s) + \beta(t-s) \tag{1.1}$$

In the condition, β is a function, called the network service curve, which is negotiated during the reservation setup phase. For example, the Internet guaranteed service assumes the form $\beta(t) = \rho(t-L)^+$ where L is called the latency and ρ the rate.

Problem P1: Given an arrival curve $\sigma(t)$ and a service curve $\beta(t)$, an admissible playback delay D and decoder buffer size X (*decoding constraints*), find the joint scheduling and source coding strategy *at the smoother* and *at the media encoder* respectively that minimizes the total distortion of the media stream such that the decoding constraints are verified.

Assumptions: We allow the smoother to perform some look-ahead (also called pre-fetching), namely, we do not require that $R'(t) \leq R(t)$. Look-ahead is commonly used with pre-recorded streams, for which the smoother is composed of both a disk server and a scheduler. Our study is restricted to the guaranteed service; we do not consider other frameworks, such as the best effort of the differentiated service of the IETF, where multiple media streams would share the same resources without individual guarantees.

The paper is organized as follows. Section 2 derives the *optimal* smoothing strategy for a stored, offline-compressed media stream. Optimal smoothing is defined as the scheduling strategy that minimizes simultaneously the playback delay and the required decoding buffer size under given traffic parameters $\sigma(t)$ and $\beta(t)$. Section 2 originally appeared in [5]. Section 3 demonstrates that the set of media streams for which optimal smoothing leads to the same minimum playback delay and required decoding buffer size under given traffic parameters is upper- and lower-bounded. This set of bounds does not depend on the media rate, allowing for a separation of the smoothing strategy from the media compression algorithm. Section 3 partly appeared in [8]. The *separation principle* is used in Section 4 where we propose the solution to *Problem P1*. In Section 5, we show experimentally that significant improvements in terms of total average distortion compared to the smoothing of a fixed media encoder output may be attained, under equivalent traffic parameters and decoding constraints (i.e., $\{(\sigma \otimes \beta), X, D\}$). Finally Section 6 summarizes the main results of this work.

2 Optimal Smoothing

A number of results exist on smoothing. In [4], smoothing is studied from the viewpoint of reducing the required network resources, with the assumption that connections are of the renegotiated CBR type. Optimality is sought in the sense of reducing the variability of the connection rate. In [9] the authors go one step further and address, among others, the issue of minimizing playback delay and buffer, for the case of a CBR connection. They also study the

cascaded scenario where playback and smoothing is performed at multiple points, typically as would occur with internetworking. Our results differ from these in two directions. Firstly, we are interested only in the end-system viewpoint, assuming that the sole information obtained by a source is what is available by signalling or by a protocol such as RSVP. Secondly, we focus on VBR rather than CBR or renegotiated CBR. Moving from CBR to VBR requires some sophistication in the method, which we try to use parsimonously. In [9], the authors find a representation of the latest optimal smoother output in the particular case of a CBR traffic envelope and a null network. As discussed in Section 2.3, we find a generalization of this result to the VBR case; we also give a simple, physical interpretation of this result in terms of time inversion.

One smoothing strategy is called *shaping* (it is called "optimal shaping" in [10]). It consists in putting the encoded flow $R(t)$ into a buffer, and outputting bits as soon as doing so does not violate the arrival curve constraint. It is shown in [10] that an optimal shaper minimizes the buffer requirement and the delay experienced in the smoother. However, a shaper is optimal only at the sender side. In this paper we consider another problem, namely, we would like to minimize the playback delay D and the buffer size at the receiver. Another difference with shaping is that we allow our smoothing strategy to look-ahead, which a shaper does not.

2.1 A formal definition of the admissible smoother output

Consider again the model illustrated in Figure 1. Assume first that we fix the value of the playback delay D. The job of the smoother is to produce an output whose cumulative function is R'. We take as time origin the beginning of the operation of the smoother, thus we must have

$$R'(t) = 0 \text{ if } t \leq 0 \tag{2.2}$$

We assume that R' is constrained by the traffic envelope σ, namely

$$R'(t) - R'(s) \leq \sigma(t-s) \text{ for all } s \leq t \tag{2.3}$$

We also assume that the network offers a service curve β to the flow, namely, Equation (1.1) is satisfied. It is more convenient to re-write Equation (1.1) as follows

$$R^*(t) \geq \inf_{0 \leq s \leq t} \{R'(s) + \beta(t-s)\} \tag{2.4}$$

As a convenient notation, the right-handside in the above equation is also traditionally written as $(R' \otimes \beta)(t)$, and is called the "min-plus" convolution of functions R' and β [10, 11]. This gives the equivalent writing for Equation (2.4):

$$R^*(t) \geq (R' \otimes \beta)(t) \tag{2.5}$$

The system must also satisfy the real-time constraint at the decoding buffer. This is expressed by

$$R^*(t) \geq R(t - D_0 - D) \tag{2.6}$$

where D is the playback delay and D_0 the transfer time for the first packet of the flow. Now we assume that the smoother cannot know the individual packet delays, but only the network service curve β. Thus, R' must be such that Equation (2.6) is true for *any* realization R^* satisfying Equation (2.4). Now remember that we have reduced our study to the case where the fixed part of the transfer delay is 0. Consider a particular realization R^* such that the first packet has a zero transfer delay, and for the rest (namely $t \geq t_1 =$ the arrival time of the second packet) satisfies the worst case $R^*(t) = (R' \otimes \beta)(t)$. We must thus have, for all $t > 0$:

$$(R' \otimes \beta)(t) \geq R(t - D) \tag{2.7}$$

Conversely, if this equation holds, then clearly $R^*(t) \geq R(t - D) \geq R(t - D_0 - D)$ and thus the real time condition is satisfied.

In summary, the constraints for the smoother is to produce an output R' which satisfies simultaneously Equations (2.2), (2.3) and (2.7).

2.2 Minimal Playback Delay

The first result in this paper is the following theorem.

Theorem 2.1. *There exists one minimum value of the playback delay D for which the smoother equations (2.2), (2.3) and (2.7) have a solution. It is given by*

$$\bar{D} = \inf\{t \geq 0 \mid \forall u \geq 0, v \geq 0 : R(u + v - t) \leq \sigma(u) + \beta(v)\}$$

The proof of the theorem is given in [12]. We now discuss the content and the implications of the theorem.

The theorem gives the smallest value of the playback delay that can be obtained by any smoothing strategy satisfying the arrival curve constraint σ, given that the network service curve guaranteed to the flow is β. The minimum delay $\bar{D}$ can be better interpreted using the concept of horizontal deviation [7], which we now recall. Figure 2 gives an intuitive definition.

Definition 2.1. *For two functions α and β, define the horizontal deviation $h(\alpha, \beta)$ by*

$$h(\alpha, \beta) = \sup_{s \geq 0} \left(\inf \{T : T \geq 0 \text{ and } \alpha(s) \leq \beta(s + T)\}\right) \tag{2.8}$$

It is shown in [12] that the value of the minimum playback delay $\bar{D}$ in the theorem is given by

$$\bar{D} = h(R, \sigma \otimes \beta) \tag{2.9}$$

In the formula, $\sigma \otimes \beta$ is the min-plus convolution defined as in the discussion following Equation (2.4), and which can be interpreted as follows [10, 7]. Consider for a second a

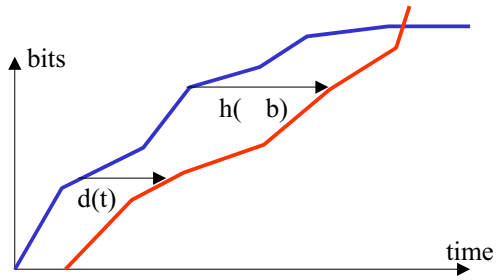

Figure 2: Definition of horizontal deviation for two functions α and β. Determine $d(t)$ for all t by drawing the horizontal distance from α to β. The horizontal deviation $h(\alpha, \beta)$ is the maximum of all $d(t)$.

hypothetical shaper, as defined in the Introduction, with traffic envelope σ. Assume that σ is a "good" function, namely sub-additive, as explained for example in [10]. The arrival curves used with RSVP or for ATM VBR connections are good functions. We know from [10, 7] that, if the input flow to the shaper is $S(t)$, and if the shaper is large enough to avoid losing data, then the output is equal to $(\sigma \otimes S)(t)$. Thus we can interpret $\sigma \otimes \beta$ as follows. Imagine a flow with cumulative function $S(t) = \beta(t)$; put this imaginary flow into a shaper in order to make it conform to the traffic envelope σ. The resulting, shaped flow is $\sigma \otimes \beta$. Then the minimum playback delay achievable with a look-ahead smoother is the horizontal deviation between the original signal $R(t)$ and the curve $(\sigma \otimes \beta)(t)$.

2.3 Optimal Smoother Output

So far we have given a result for the minimum playback delay. We now show a more global result, namely, there exists one smoother output which is better than any other output, at any time instant, in a sense which we define now.

Definition 2.2. *For a given signal $R(t)$, define $R^-(t)$ for all $t \in \mathbb{R}$ by*

$$R^-(t) = \sup_{u \geq 0, v \geq 0} \{R(t+u+v) - \sigma(u) - \beta(v)\}$$

Note that, unlike R, the function R^- is non-zero even for some negative times. After appropriate time-shifting, R^- is the optimal smoother output, as the following theorem shows.

Theorem 2.2. *This theorem is divided in two parts:*

1. *The minimal delay defined in Theorem 2.1 is the smallest t such that $R^-(-t) \leq 0$*
2. *For any admissible smoother output R', with playback delay D, we have, for all $t \geq 0$, $R'(t) \geq R^-(t - D)$*

The proof is given in [12]. We can interpret the theorem as follows. The first item relates the minimal delay $\bar{D}$ to the optimal output. It says that $\bar{D}$ is the smallest time shift which is necessary to make the flow described by R^- start at time 0. Second, note that, since $\bar{D}$ is the minimum playback delay, we must have $D \geq \bar{D}$. Now call $\overline{R'}(t) = R^-(t - \bar{D})$ the optimal output, namely the shifted version of R^- that starts at time 0. Then the theorem means that if we time-shift $\overline{R'}$ so that the first packet for this solution is played back at the same time as the first packet for some other solution R', then $\overline{R'}$ is, at every time instant, no earlier than R'. The shifted optimal output $\overline{R'}(t - (D - \bar{D})) = R^-(t - D)$ thus gives the latest time at which *every* packet of the flow should be scheduled. Figure 3 illustrates this.

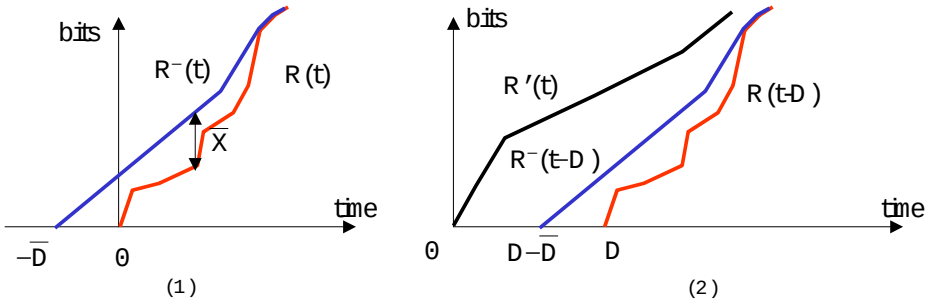

Figure 3: Optimal smoothing: (1) computation of $R^-(t)$ from the encoded signal $R(t)$. The minimum playback delay $\bar{D}$ is the point where $R^-(-t)$ hits 0. (2) For any admissible smoother output $R'(t)$ with playback delay D, the shifted version $R^-(t - D)$ is no earlier that R'.

Representation of Optimal Smoother Output with Time Inversion: The shifted optimal output R^- can be computed using its definition; however, we can reduce its complexity with a time inversion transformation. At this point we need to introduce a classical min-plus construct, called min-plus deconvolution, noted $\oslash$, and defined [13] by:

$$(f \oslash g)(t) = \sup_{u \in \mathbb{R}} \{f(t + u) - g(u)\} \tag{2.10}$$

Note that $f \oslash g$ may be non-zero for negative times even if this is not the case for f and g. With this notation, the function $R^-(t)$ can be written in a more compact way as $R^- = R \oslash (\sigma \otimes \beta)$.

It is shown in [12] that min-plus deconvolution can be computed easily by means of time inversion. Thus, R^- can be computed as follows. First invert time; then compute, in the inverted time domain, the min-plus convolution of the resulting function on one hand, of $\sigma \otimes \beta$ on the other hand; lastly, invert time again and obtain R^-.

In [9], the authors find a representation of the optimal smoother output in the particular case of a CBR traffic envelope and a null network. Their representation can be easily interpreted as the time inverted signal, shaped to a constant bit rate. Thus, their representation is a particular case of our result.

Required Buffer at the Decoder: Consider now the buffer size that must be provisioned at the decoder. Remember that we can remove any fixed delay. Thus, for a given scheduler output $R'(t)$, all we can know about the decoder input decoder R^* is that $R(t-D) \leq R^*(t) \leq R'(t)$. The decoder buffer content at some time t is $R^*(t) - R(t-D)$. Thus the buffer size that must be provisioned is $\sup_{t\geq 0}\{R'(t) - R(t-D)\}$. A simple examination of Figure 3 shows the following corollary.

Corollary 2.1. *The buffer size that need to be provisioned at the decoder is minimum for solution* $\overline{R'}(t) = R^-(t - \bar{D})$. *It is equal to*

$$\begin{aligned} \bar{X} &= \sup_{t\geq 0}\{R^-(t) - R(t)\} \\ &= \sup_{(t,u,v)\geq 0}\{R(t+u+v) - R(t) - \sigma(u) - \beta(v)\} \end{aligned}$$

We show in [12] that the formula for $\bar{X}$ can be interpreted in terms of network calculus abstractions, which leads to the following simplification.

The complexity of computing $\bar{X}$ with this method is $O(n^2)$, where n is the number of samples in the trace $R(t)$. In [12] we give an alternative method using the time inversion representation, which has a complexity of $O(n)$. It is the same representation as in [9], Section IV.A., for the particular case of a null network and a CBR traffic envelope.

2.4 Null network case

Consider the case where the network service provides a constant transfer delay. This occurs for example with a circuit switched service, or, as an approximation, with ATM constant bit rate (CBR) services if the delay variation if very small. In our framework, a constant delay network is equivalent to a null network.

The null network case is a straightforward application of the general case, by letting $\beta(t) = +\infty$ for all $t \geq 0$. Equivalently, simply remove β from all formulas: for example, the minimum playback delay becomes

$$\bar{D} = h(R, \sigma) = \inf\{t \geq 0 \mid \forall u \geq 0 : R(u-t) \leq \sigma(u)\}$$

3 Set of Bounds for Media Rate $R(t)$

So far we showed that there exists one optimal strategy at the smoother that minimizes the playback delay and the decoding buffer size, given some traffic parameters and a media flow $R(t)$. Now we fix the values of the playback delay and decoding buffer size, which we call decoding constraints. We study the set of flows $R(t)$ such that the optimal smoothing solution given some traffic parameters does not violate the decoding constraints.

We consider the null network case only. That is, $\beta(t) = +\infty$ for all $t \geq 0$ and $(\sigma \otimes \beta)(t)$ reduces to $\sigma(t)$. We further assume that $\sigma(t)$ is of the form $\min(M + pt, rt + b)$ and $\sigma(u) = 0$ for all $u \leq 0$.

Consider an arrival curve $\sigma(t)$, a limited client buffer X and a maximum playback delay D. We define the set $\Omega_{\sigma,(X,D)}$ of flows $R(t)$ such that the optimal smoothing strategy applied to any $R(t) \in \Omega_{\sigma,(X,D)}$ respects the following set of constraints:

- Continuous Media: $R^-(t - \bar{D}) \geq R(t - D)$,
- Decoding buffer X: $R^-(t - \bar{D}) - R(t - D) \leq X$,
- Playback delay D: $R^-(t) \leq 0, \forall\ t \leq -D$.

Clearly the cardinality of the set $\Omega_{\sigma,(X,D)}$ may be greater than one. Figure 4 illustrates this fact with two input flows $R_1(t)$ and $R_2(t)$, both belonging to $\Omega_{\sigma,(X,D)}$. This leads us to the following Theorem:

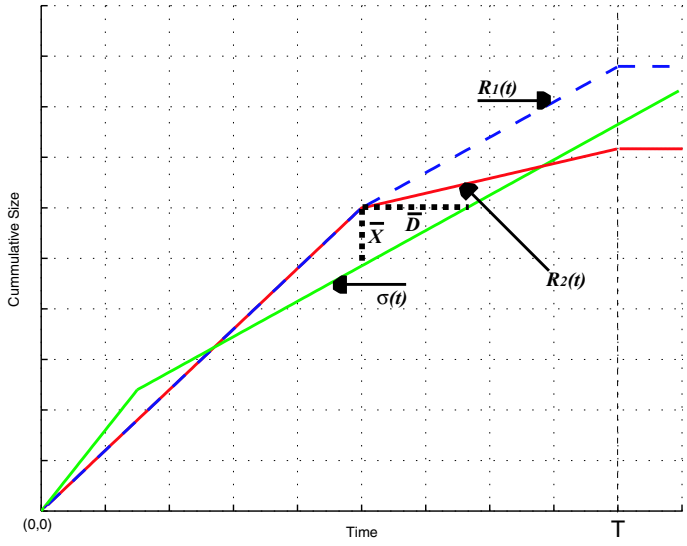

Figure 4: An arrival curve $\sigma(t)$ of the form $\min(M + pt, rt + b)$ and two streams $R_1(t)$ and $R_2(t)$. The respective optimal smoothing solutions require the same decoding buffer size $\bar{X}$ and playback delay $\bar{D}$.

Theorem 3.1. *(i) The flows $R(t) \in \Omega_{\sigma,(X,D)}$ are upperbounded by the function $R^{max}(t)$ defined as:*

$$R^{max}(t) = \delta_0(t) \wedge (\sigma(t) + \bar{X}) \wedge \sigma(t + \bar{D}),$$

where δ_0 is the 'impulse' function defined by $\delta_0(t) = +\infty$ for $t > 0$ and $\delta_0(t) = 0$ for $t \leq 0$, and $a(t) \wedge b(t)$ is the point-wise minimum between functions $a(t)$ and $b(t)$.

(ii) The flows $R(t)$ of equal duration T and such that $R(+\infty) = R^{max}(T)$ are lowerbounded by the function $R^{min}(t)$ written as:

$$R^{min}(t) = R^{max}(T) - R^{max}(T - t),$$

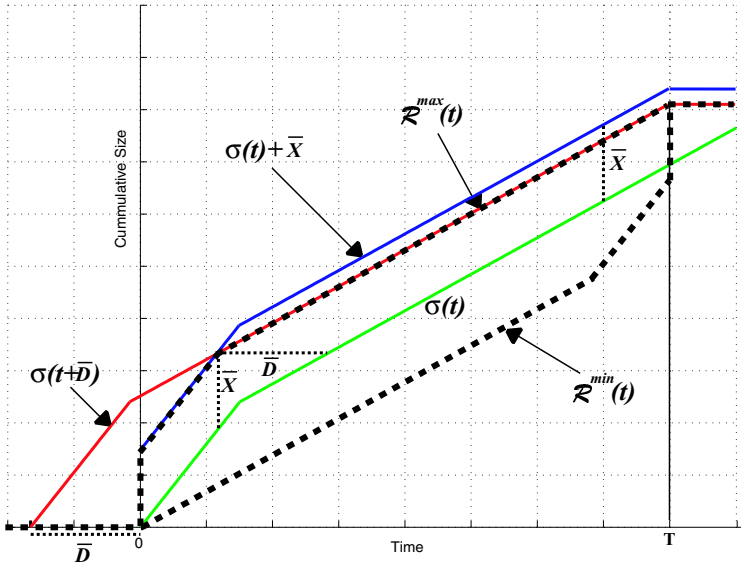

Figure 5: *An arrival curve $\sigma(t)$. According to Theorem 3.1, the maximum input flow $\mathcal{R}^{max}(t) \in \Omega_{(\bar{X},\bar{D})}$ for which the optimal smoothing solution requires a decoding buffer size of $\bar{X} \leq X$ and playback delay of $\bar{D} \leq D$ is given by the point-wise minimum between the three functions $\delta_0(t)$, $\left(\sigma(t) + \bar{X}\right)$ and $\sigma\left(t + \bar{D}\right)$. And the minimal input $\mathcal{R}^{min}(t)$ of flows $\mathcal{R}(t)$ with equal duration T and such that $\mathcal{R}(+\infty) = \mathcal{R}^{max}(T)$ is obtained by time reverting $\mathcal{R}^{max}(t)$. Any trajectory within these bounds respects the channel and client constraints when optimally smoothed.*

The interested reader may refer to [8] for the proof of the upperbound. The proof for the lowerbound is similar but in the reverse time domain. Figure 5 illustrates the Theorem.

Let $R(t)$ denote the output of a lossy media compression algorithm (e.g., MPEG-x, H.26x). The quantization step has been adjusted to produce the expected amount of traffic at time t, $\forall\ 0 \leq t < T$, with T being the duration of the input media sequence. A higher quantization step usually results in a higher compression factor, and conversely. Also, the higher the quantization step, the higher the degradation (see Sec. 5). We are interested in the trajectory $R^{opt}(t)$ that minimizes the total distortion given some traffic parameters and under some decoding constraints; namely $[\sigma, (X, D)]$.

The rate-distortion curve at time t is highly dependent on the spatio-temporal complexity of the underlying signal. That is, two different media segments compressed at the same rate usually result in different degradation levels. An efficient rate control algorithm increases the source rate whenever the spatio-temporal complexity of the underlying media signal increases, and conversely. Clearly, among the set of functions $R(t) \in \Omega_{\sigma,(X,D)}$, the solution $R(t) = R^{max}(t)$ does not necessarily lead to the minimal total distortion. Indeed it is unlikely that, given the parameters $\{\sigma, (X, D)\}$, the cumulative spatio-temporal complexity of the media signal follows the concave function $R^{max}(t)$.

However we observe that the flow $R^{opt}(t)$ that minimizes the total distortion is part of

the subset of functions in $\Omega_{\sigma,(X,D)}$ such that $R(+\infty) = R^{max}(T)$. This is straightforward from the property of strict convexity of the rate-distortion curves. Thus, from Theorem 3.1, the optimal trajectory $R^{opt}(t)$ is such that $R^{min}(t) \leq R^{opt}(t) \leq R^{max}(t)$. Finally, given an uncompressed continous media of total duration T and known time-varying R-D characteristics, and the traffic parameters and decoding constraints $[\sigma, (X, D)]$, the optimal scheduling strategy in a rate-distortion sense $\bar{R}'(t)$ is simply given by the following:

$$\bar{R}'(t) = (R^{opt} \ominus \sigma)(t - \bar{D}). \tag{3.11}$$

The next section proposes efficient techniques to obtain the rate-distortion optimal trajectory $R^{opt}(t)$.

4 Optimal Rate-Distortion trajectory

In Section 2 we relied on a stored static media stream $R(t)$. Now we dynamically build the stream such that we minimize a given cost function (e.g., average distortion of the video stream) while insuring that the output of the optimal smoother leads to a playback delay not greater than D and a required buffer size not larger than X.

Many rate-distortion (R-D) optimization methods have been proposed in the literature [3]. These methods typically perform a pre-analysis of the media sequence to measure the time-varying R-D characteristics before applying a rate allocation strategy. A popular approach has been to rely on a R-D model [14]. That is, a function $f_t(x)$ that models the relation between distortion and rate at time t; namely $d(t) = f_t(r(t))$, where $r(t)$ and $d(t)$ respectively denote the instantaneous bit rate and distortion at time t. The function $f_t(x)$ is strictly convex and positive. Thus, given this function, the problem of finding $R^{opt}(t)$ (that is, the trajectory $R(t) \in \Omega_{\sigma,(X,D)}$ that minimizes the total distortion) can be cast as a separable convex optimization problem.

Let us divide the time axis in intervals of fixed duration Δ (display duration of a frame or a group of frames). Let r_i denote the instanteanous rate in the time interval $I_i = [(i-1)\Delta, i\Delta)$, for all $1 \leq i \leq N$ and $N\Delta = T$ is the duration of the continuous media. We can write $R_i = \sum_{j=0}^{i} r_j$. Similarly we define d_i as the media distortion measured in the interval I_i. Let the function $f_i(x)$ denote the relation between distortion and rate in I_i; namely $d_i = f_i(r_i)$. We can solve the following problem using state-of-the-art convex programming techniques:

$$\begin{aligned} \text{Find} \quad & \{r_i\}, \ \forall\, 1 \leq i \leq N \\ \text{that minimizes} \quad & \sum_{i=1}^{N} f_i(r_i) \\ \text{under} \quad & R_i^{min} \leq \sum_{j=1}^{i} r_j \leq R_i^{max}, \end{aligned} \tag{4.12}$$

from which R_i^{opt} is simply given by: $R_i^{opt} = \sum_{j=1}^{i} r_j$ for all $1 \leq i \leq N$.

Unfortunately, model-based R-D optimizations have shown their limitations. Usually the model error is large and not strictly positive (i.e., the model is not an upper-bound of the exact R-D function). Therefore, we propose an approximation method based on computing a few R-D points and interpolating the remaining points using linear functions. The resulting piece-wise linear model is always greater or equal to the exact R-D function and has shown great potentials for rate-distortion optimization [15]. Moreover our problem becomes the minimization of a separable piece-wise linear convex function subject to linear constraints, which can be solved via extremely efficient linear programming (LP) techniques.

5 Experimental Results

The main objective of this section is to show experimentally that significant improvements in terms of total average distortion compared to the smoothing of a fixed media encoder output may be attained, under equivalent traffic parameters and decoding constraints (i.e., $\{\sigma, X, D\}$).

5.1 Experimental Setup

Experiments have been conducted on a 168-frame long sequence conforming to the ITU-R 601 format (720*576, 25 frames per second). The sequence is composed of 2 video scenes that differ in terms of spatial and temporal complexities. The time axis is divided into fixed intervals of a group of pictures (GoP) duration (i.e., approximately 0.5 s.). The sequence was Open-Loop VBR (OL-VBR) compressed with the TM5 MPEG-2 video encoder using 5 different quantizer scale factors (MQUANT), ranging from 10 to 56. Figure 6 shows the cumulative trace resulting from OL-VBR encoding the sequence at MQUANT=56 (the mean squared error is 83.75; equivalently, the peak signal-to-noise ratio (PSNR) is 28.9 dB). The figure also shows the piece-wise linear approximation of the rate-distortion function at time $t = t_4$ and the experimental fitting of a common R-D model ($d = r^{\sigma}$). Note that the R-D model is not an upper-bound of the rate-distortion function.

5.2 Experimental R^{opt} and R^{-}

Figure 7 shows the bounds $R^{max}(t)$ and $R^{min}(t)$, and the solution $R^{opt}(t)$ for the following constraints: (i) $\beta(t) = \delta_0(t)$ and $\sigma(t) = \min\{pt, rt+b\}$ with peak rate $p = 6$ Mbps, sustainable rate $r = 4$ Mbps and bucket level $b = 4$ Mbits, and (ii) decoding buffer $X = 3$ Mbits and admissible playback delay $D = 1$ GoP (0.5 s.). The media rate $R^{opt}(t)$ is the optimal solution to the piece-wise linear R-D optimization problem. The minimal average distortion is MSE = 39.1 (i.e., PSNR = 32.2 dB). The media sequence $R_{28}(t)$ compressed with a constant quantizer MQUANT=28 also achieves the same distortion level but requires a decoding buffer $X_{28} = 1.5X$ and playback delay $D_{28} = 3D$ under the same traffic parameters.

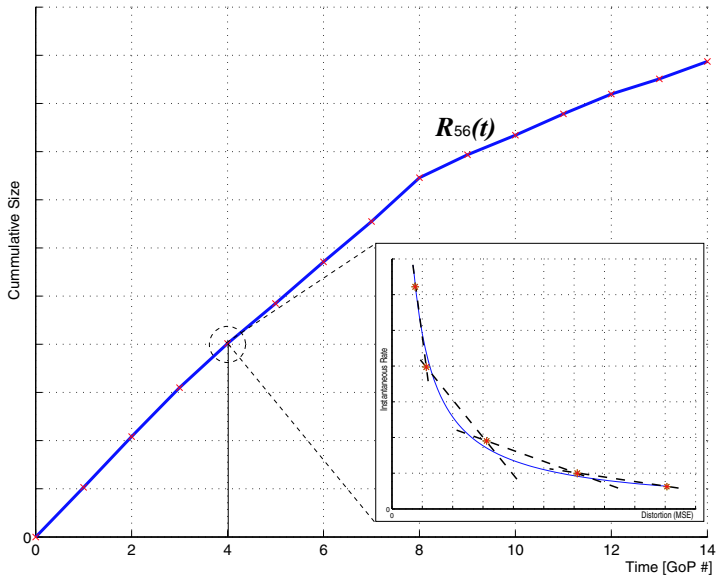

Figure 6: The MPEG-2 trace $R(t)$ compressed at a constant MQUANT=56 (cumulative representation). The average distortion (MSE) is 83.75 (i.e., PSNR=28.9 dB). The piece-wise linear approximation of the rate-distortion function at time $t = t_4$ and the experimental fitting with a R-D model: $d = r^{\sigma}$.

The minimal distortion achieved by the constant bit rate (CBR) media sequence is 55.6 (i.e., PSNR = 30.6 dB). Finally the optimal scheduling trajectory $\bar{R}'(t)$ is simply obtained from $\bar{R}'(t) = (R^{opt} \ominus \sigma)(t - \bar{D})$.

6 Summary of the Main Results

We have analyzed the scenario where a multimedia source uses the guaranteed service; the flow is assumed to receive a certain fixed network service curve, but has to comply with some traffic envelope. First we were interested in minimizing playback delay and required buffer at the decoder. In this context, we found that there exists one minimum playback delay, and obtained one scheduling strategy at the source which achieves this minimum. This strategy is also the one that sends data as late as possible. This result is explicit and easy to compute, but requires a complete knowledge of the entire signal. Nonetheless, the existence of and the expression for an explicit optimum is a fundamental result which can be used to analyze practical scheduling strategies.

Finally, we have shown that improvements in terms of total average distortion could be attained by adding a source rate selection mechanism to the optimal smoothing strategy. We presented the optimal low-complexity streaming strategy of continuous media in a rate-distortion sense for guaranteed service networks. First we computed the theoretical bounds on the cumulative media rate such that the optimal smoothing solution meets the decoding

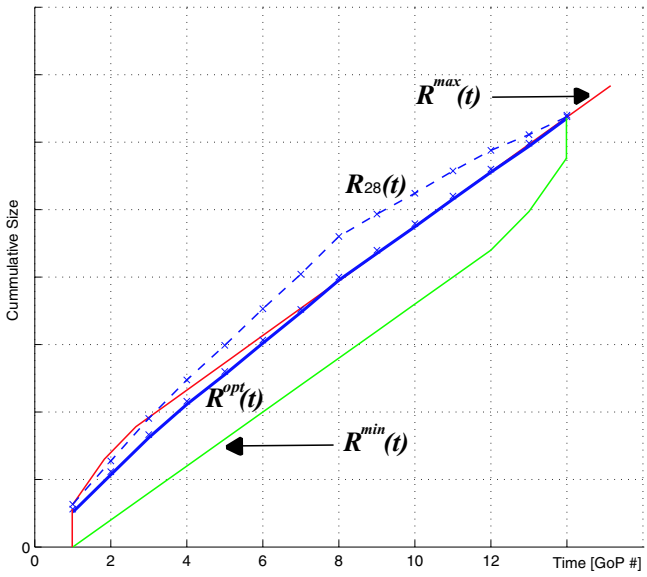

Figure 7: The theoretical bounds $R^{max}(t)$ and $R^{min}(t)$, the experimental optimal solution $R^{opt}(t)$ and the rate trajectory for constant MQUANT=28. The constraints are: (i) $\beta(t) = \delta_0(t)$ and $\sigma(t) = \min\{pt, rt + b\}$ with peak rate $p = 6$ Mbps, sustainable rate $r = 4$ Mbps and bucket level $b = 4$ Mbits, and (ii) decoding buffer $X = 3$ Mbits and admissible playback delay $D = 1$ GoP (0.5 s.).

constraints given some traffic parameters. Then we cast the rate-distortion problem as a piece-wise linear convex optimization algorithm where the bounds hereabove translate into linear constraints. The proposed joint technique may also be used as a benchmark tool for more practical frameworks (e.g., partial knowledge of the continuous media or partial knowledge of the behavior of the network).

References

[1] T. V. Lakhsman, A. Ortega and A. R. Reibman, "VBR video: Trade-offs and potentials," *Proceedings of the IEEE*, July 1997.

[2] R. Guérin and V. Peris, "Quality-of-service in packet networks - basic mechanisms and directions," *Computer Networks and ISDN, Special issue on multimedia communications over packet-based networks*, 1998.

[3] A. Ortega C.-Y. Hsu and A. R. Reibman, "Joint selection of source and channel rate for vbr video transmission under atm policing constraints," *IEEE Journal on Selected Areas in Communications*, 1997.

[4] J. Salehi, Z. Zhang, J. Kurose and D. Towsley, "Supporting stored video: Reducing rate variability and end-to-end resource requirements through optimal smoothing," *ACM SIGMETRICS*, May 1996.

[5] J.-Y. Le Boudec and O. Verscheure, "Optimal Smoothing for Guaranteed Service," *IEEE/ACM Transactions on Networking*, December 2000.

[6] R. L. Cruz, "Quality of service guarantees in virtual circuit switched networks," *IEEE Journal on Selected Areas in Communications*, pp. 1048–1056, August 1995.

[7] J.-Y. Le Boudec, "Application of network calculus to guaranteed service networks," *IEEE Transactions on Information Theory*, , no. 44, pp. 1087–1096, May 1998.

[8] O. Verscheure, P. Frossard and J.-Y. Le Boudec, "Joint Smoothing and Source Rate Selection for Guaranteed Service Networks," *IEEE INFOCOM*, April 2001.

[9] J. Rexford and D. Towsley, "Smoothing variable bit rate video in an internetwork," *IEEE/ACM transactions on networking*, vol. 23, no. 7, pp. 202–215, 1999.

[10] C.S. Chang, "On deterministic traffic regulation and service guarantee: A systematic approach by filtering," *IEEE Transactions on Information Theory*, , no. 44, pp. 1096–1107, August 1998.

[11] F. Baccelli, G. Cohen, G. J. Olsder and J.-P. Quadrat, *Synchronization and Linearity, An Algebra for Discrete Event Systems*, John Wiley and Sons, 1992.

[12] J.-Y. Le Boudec and O. Verscheure, "Optimal Smoothing for Guaranteed Service," Technical Report DSC/2000/014, EPFL-DSC, 2000, available at http://dscwww.epfl.ch/EN/publications/ps_files/tr00_014.pdf.

[13] R. Agrawal, R. L. Cruz, C. Okino, and R. Rajan, "Performance bounds for flow control protocols," *IEEE/ACM Transactions on Networking (7) 3*, pp. 310–323, June 1999.

[14] P. Frossard and O. Verscheure, "Joint Source/FEC Rate Selection for Quality-Optimal MPEG-2 Video Delivery," *IEEE Transactions on Image Processing*, vol. 10, no. 12, pp. 1815–1825, December 2001.

[15] L.-J. Lin and A. Ortega, "Bit-Rate Control Using Piecewise Approximated Rate-Distortion Characteristics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 4, pp. 446–459, August 1998.