LIDS-P-2243

OPTIMAL BUFFER CONTROL FOR VARIABLE-RATE LOSSY COMPRESSION

David Tse, Robert Gallager and John Tsitsiklis
Laboratory for Information and Decision Systems, M.I.T.
email: dntse@lids.mit.edu

Presented at the 31st Allerton Conference, Sept. 1993

Abstract

In many lossy data compression systems, such as the one for HDTV, the data stream generated by the source coder has a variable rate, while the channel on which the data is to be transmitted operates in a fixed-rate manner. This necessitates the use of a buffer and a buffer control scheme whereby parameters of the compression system are adjusted based on the state of the buffer, to avoid overflow and at the same time to maintain adequate performance in terms of average distortion. We study a simple model and derive asymptotically optimal control scheme in the regime of large buffer size.

1 Introduction

Problems in rate-distortion theory are usually posed in terms of minimizing the average distortion in quantizing the source given a constraint in the *average* rate of the encoded bit stream. Very often, however, there are more stringent constraints on the behavior of the output process from a source coder. For example, there is usually a fixed bandwidth allocated for the broadcasting of compressed digital television signals, implying that the encoded bit stream must be fed into a fixed-rate channel. In data networks, a variable-rate output bit stream can usually be accommodated, but there are often rate-control schemes which regulate the burstiness of the stream by enforcing a peak rate as well as an average rate.

While it is theoretically possible to approach the rate-distortion limit using vector quantization and a fixed-length encoding of the reproduction vectors, the block length required of the vector quantizer is much too large to be useful in practice. With a small block length, variable-length encoding of the reproduction vectors can attain a significantly smaller distortion for a given average bit rate, by taking advantage of the statistical variability of the source. However, this necessitates the use of a buffer between the source coder and the channel in order to satisfy the constraints described above.

A general model is as follows. The input real-valued data are represented by the stochastic process $\mathbf{Y_1}, \mathbf{Y_2}, \ldots$, where $\mathbf{Y_t} \in \Re^k$ and k is the block length of the data to be quantized. The statistics of $\{\mathbf{Y_t}\}$ are assumed to be known and the block length k is pre-specified. At each time t, a quantizer Q is chosen which maps the vector $\mathbf{Y_t}$ into a variable-length binary codeword $\mathbf{C_t} = Q(\mathbf{Y_t})$ representing the reproduction vector into which $\mathbf{Y_t}$ is quantized. (Here, we are regarding Q as a composition of both the vector quantization and the lossless encoding of the reproduction vector.) The codeword $\mathbf{C_t}$ is then put into a buffer of size L. At each time instant t, a fixed number, R_c, of bits is taken out of the buffer and put into the channel. A *buffer control* scheme chooses the quantizer Q for the next source vector, based on the number of bits S_t currently in the buffer and the previous codewords $\mathbf{C_1}, \mathbf{C_2}, \ldots, \mathbf{C_t}$. (Assuming no transmission errors, this information is also available to the decoder so it can always keep track of the quantizer used at each time.) The problem is to find the buffer control scheme that minimizes the average distortion while keeping the buffer from overflowing.

While the above formulation naturally encompasses the case when the channel has a strictly fixed rate, it is also appropriate to model some situations where burstiness is allowed. One such scheme is the leaky bucket [GB92], which, when given two parameters R_c and W, implements a rate-control scheme such that the average rate of bits entering into the channel is R_c and the maximum number of bits going into the channel per unit time slot is W bits. One view of this scheme is to imagine that permits are generated at a rate of R_c at each time slot and are put into a bucket of size W. Excess permits are discarded. In time slot t, A_t data bits enter a buffer (of size L) and the ones at the bottom of the buffer can leave it and enter the channel as long as each bit can obtain a permit from the bucket. Thus, no more than W bits can enter the channel in any one time slot and the average rate is R_c. It is not difficult to show that if P_t is the number of permits in the bucket and S_t is the number of data bits waiting in the buffer at the beginning of time slot t, then $\tilde{S}_t$, defined by:

$$\tilde{S}_t = \begin{cases} S_t + W & \text{if } S_t > 0 \\ W - P_t & \text{if } S_t = 0 \end{cases}$$

is the state of a fictitious buffer of size $L + W$, with arrival process $\{A_t\}$ and a constant departure rate R_c bits per unit time. Moreover, since there is a one-to-one correspondence between the fictitious buffer state $\tilde{S}_t$ and the buffer-bucket state pair (S_t, P_t), any optimal control scheme for the fictitious buffer will translate into an optimal control scheme for the original leaky bucket problem. Thus, our formulation is general enough

to include this case.

The problem of buffer overflow for *lossless* variable-length encoding has a relatively long history. It was first looked at by Jelinek [Jel66], who gave conditions on codeword lengths that minimize the probability of overflow. This work was followed by Humblet [Hum92] who presented an algorithm to actually compute the optimal code. Wyner [Wyn74] gave an approximate formula for the average fraction of data that will be lost due to overflow for any given variable-length lossless coding scheme. These results are all on lossless coding, where there is no tradeoff between average distortion and overflow probability and hence no opportunity for control. More recent and closer to the problem considered here is the work of Farvardin and Modestino [FM86] and Harrison and Modestino [HM90]. They presented specific buffer control schemes for variable-rate lossy coding of memoryless sources and computed numerically and via simulation their average distortion performance. No attempt was made to derive optimal schemes.

The objective of this paper is to gain a better understanding of the structure of *optimal* buffer control schemes. For a fixed source and buffer size L, the optimal control scheme can be obtained numerically as a solution to an average-cost dynamic programming problem. Such an approach, however, yields little insights to qualitative features of optimal schemes. Moreover, sources like video are very complex and a complete statistical characterization is often impossible. To get more theoretical insights, we shall rather consider the asymptotic problem when the buffer size L becomes large, and analytically derive optimal buffer control schemes for that regime. As a first step in this program, we will treat the case of memoryless sources. More complex sources will be considered in future papers.

In the next section, we will give a more precise formulation of the problem. Section 3 contains our results. We will present a lower bound on the achievable performance as well as a very simple control scheme that almost attains that lower bound. The performance is also compared against what is achievable without any control. Section 4 contains the conclusions. In this extended abstract, the proofs will only be sketched. Also, the following short-hand notations will be used to compare rates of convergence: $a_n = O(b_n)$ if a_n goes to zero at least as fast as b_n; $a_n = o(b_n)$ if a_n goes to zero strictly faster than b_n; $a_n = \Omega(b_n)$ if a_n goes to zero no faster than b_n; $a_n = \Theta(b_n)$ if they go to zero at the same rate.

2 Formal Problem Statement

Let $\mathbf{Y_1}, \mathbf{Y_2}, \ldots$ be an i.i.d. sequence of source samples, where $\mathbf{Y_t} \in \Re^k$ has a density p_y, which is assumed to be sufficiently smooth. . A quantizer $Q : \Re^k \rightarrow \{0,1\}^*$ is a composition of two maps, $Q = V \circ U$ where $U : \Re^k \rightarrow \Re^k$ maps the data sample to a reproduction vector, and $V : \Re^k \rightarrow \{0,1\}^*$ assigns a binary codeword to the reproduction vector. The lossless code V is assumed to be uniquely decodable and satisfy the prefix-condition. Two important attributes of a quantizer are its mean-square distortion and its rate:

$$D(Q) = E(\|\mathbf{Y} - U(\mathbf{Y})\|^2), \qquad R(Q) = E(|Q(\mathbf{Y})|)$$

(Here, $|\cdot|$ denotes the length of the codeword.)

To characterize the optimal distortion that can be achieved for a given *average* rate, define the operational distortion-rate curve as:

$$D_{op}(R) = \inf_{\{Q:E(|Q(\mathbf{Y})|)\leq R\}} D(Q)$$

This is the minimum distortion one can achieve using the same quantizer on each data vector, with only an average rate constraint. It is not true in general that this curve is convex. If it is not, then in fact one can do better by time-sharing between points on this curve, i.e. we can use two quantizers, each for a fraction of the time, and achieve a better average distortion than one single quantizer at the same average rate. Let us now define the time-shared distortion-rate curve, $D_T(R)$, as the lower convex hull of $D_{op}(R)$. This curve will be a union of strictly convex sections, on which the optimum can be attained by a single quantizer, and straight-line segments, on which time-sharing takes place. Points on $D_T(R)$ represent the minimum distortion that can be achieved for a given source block length and average rate, and without any buffer constraints. The problem is how well we can do relative to this optimum when there is a finite buffer.

Let us now turn to the setting with a finite buffer and an adaptive choice of quantizers. Since the data samples are i.i.d., it suffices to consider only memoryless buffer control schemes where the choice of the quantizer for the next sample is only a function of the buffer state and not of the previous quantized samples. Given a current buffer state $S_t^L = s$, let the quantizer selected for the next sample be Q_s^L. (The explicit dependence on the buffer size L is shown to emphasize that a different control scheme can be chosen for each L.) Let the net number of bits entering the buffer from the

next sample be $X_{t+1}^{L,s} \equiv |Q_s^L(Y_{t+1})| - R_c$. If this results in an underflow of the buffer, i.e. $X_{t+1}^{L,s} + S_t^L < 0$, the deficit is simply padded with zeros and sent onto the channel. (Assuming the decoder also keeps track of the buffer state, there is no problem in decoding here.) On the other hand, if the current sample results in an overflow of the buffer, i.e. $X_{t+1}^{L,s} + S_t^L > L$, then the part of the codeword that fits into the buffer is sent, thus filling up the buffer. We assume that in decoding this incomplete codeword, a fixed and large distortion of D_0 is committed, where $D_0 = E(\|\mathbf{Y} - E(\mathbf{Y})\|^2)$. In other words, the decoder treats this as a garbled codeword and just decodes it to the mean vector. In practice, one can do something more sophisticated, such as sending the nearest codeword which will fit into the buffer [BCTK80]. But the specific overflow handling scheme will have little bearing on our asymptotic analysis, so in this paper we will concentrate on the simpler scheme.

Given the buffer size L and a specific buffer control scheme, the buffer state process $\{S_t^L\}$ forms a finite-state Markov chain. Assuming that the chain is ergodic, let its steady state distribution be π_L. Since a large distortion is committed whenever the buffer becomes full, the tradeoff between the steady-state probability the buffer is full, $p_f(L) \equiv \pi_L(L)$, and the steady state average distortion in normal operation when the buffer is not full, $D_q(L) \equiv \sum_s \pi_L(s) D(Q_s^L)$, is of interest. For reasonable control schemes, one would expect that $p_f(L) \to 0$ and $D_q(L) \to D(R_c)$ as the buffer size L becomes large, where $D_T(R_c)$ is the minimum distortion achievable with an average rate constraint of R_c. The key question that will be addressed is the optimal tradeoff between these two rates of convergence, and which control scheme can achieve this optimal tradeoff. Another way of viewing the problem is to consider the total distortion as the sum of the average distortion in normal operation and the large distortion when the buffer is full, ie. $D_q(L) + p_f(L)D_0$, and the control objective is to make both terms small.

3 Results

The key tool used to estimate the steady-state probability of a full buffer is Wald's identity [Wal44], which yields information on hitting probabilities and hitting times of random walks.

Lemma 3.1 *(***Wald's Identity** *) Let $\{W_n\}$ be a negative-drift random walk starting at the origin, and let $\Lambda(r) \equiv \log E(e^{rW_1})$ be its log moment generating function, which*

is assumed to be finite for all r. Let $a \geq 0$ and $b \leq 0$ be two given barriers, and let N be the first $n \geq 1$ such $W_n \geq a$ or $W_n \leq b$. Then

$$E(e^{r^* W_N}) = 1$$

where r^* is the unique positive root of $\Lambda(r) = 0$. Furthermore, the expected hitting time is given by:

$$E(N)E(W_1) = E(W_N)$$

The tradeoff between the average distortion in normal operation and the fullness probability depends on whether the channel rate R_c lies in the time-shared regions of the distortion-rate curve or in the regions achievable by single quantizer. In the time-shared regions of the distortion-rate curve, one can achieve essentially optimal distortion with negligibly small probability of filling the buffer.

Proposition 3.2 *If $(R_c, D_T(R_c))$ lies on a straight-line time-shared segment of the distortion-rate curve, then there is a control scheme such that the average distortion $D_q(L)$ approaches $D_T(R_c)$ exponentially fast in L, and a fullness probability $p_f(L)$ decaying exponentially to zero with L.*

Sketch of Proof. Let (R_1, D_1) and (R_2, D_2) $(R_1 < R < R_2)$ be the two ends of the straight line segment containing $(R_c, D_T(R_c))$, and let Q_1 and Q_2 be the quantizers achieving these points respectively. Our scheme is as follows. For each L, use quantizer Q_2 when the buffer is less than half full and use Q_1 otherwise. Note that the selected quantizers are independent of the buffer size.

To compute $p_f(L)$, we view the times at which the buffer becomes full as the epochs of a renewal process and apply renewal theory to obtain $p_f(L) = \frac{1}{E(T)}$, where T is the duration between successive times when the buffer becomes full. Denote the upper and lower halves of the buffer by A_1 and A_2 respectively. To compute $E(T)$, we condition and decompose the evolution of the buffer state process with respect to events of barrier hitting, where the barriers are the top of the buffer, the center of the buffer, and the bottom of the buffer. Since in between these events, the process is simply a random walk, we can express $E(T)$ in terms of the expected barrier hitting times and hitting probabilities, which can in turn be estimated via Wald's identity. One finds that

$$
\begin{aligned}
E(T) &= \Theta\left(\frac{1}{P(\text{hitting the top before returning to center, starting from center})}\right) \\
&= \Theta(\exp(\frac{1}{2}r^*(Q_1)L)) \qquad \text{(by Wald's Identity)}
\end{aligned}
$$

where $r^*(Q_1)$ is the unique positive zero of the log moment generating function of the random variable $|Q_1(\mathbf{Y})| - R_c$.

We next compute the average distortion $D_q(L)$ for L large. This is given by:

$$D_q(L) = \pi_L(A_1)D_T(Q_1) + \pi_L(A_2)D_T(Q_2)$$

Since the buffer state process is bounded, the expected drift over its steady-state distribution is zero. Hence we can write the balance equation:

$$\pi_L(A_1)(R_1 - R_c) + \pi_L(A_2)(R_2 - R_c) + \text{ effects due to top and bottom of buffer } = 0$$

Since the fraction of time spent at the top and bottom of the buffer is exponentially small as L becomes large, it can be argued that the third term is also exponentially small, and

$$
\begin{aligned}
D_q(L) &= \frac{R_2 - R_c}{R_2 - R_1}D_T(Q_1) + \frac{R_c - R_1}{R_2 - R_1}D_T(Q_2) + \text{ exponentially small term} \\
&= D_T(R_c) + \text{ exponentially small term}
\end{aligned}
$$

□

Hence, if R_c is in the time-shared region, then both the average distortion in normal operation and the fullness probability approach their respective optimal values exponentially fast. Thus, there is really not much need to worry about the tradeoff between these two quantities. The reason for this nice situation is that one can use two fixed quantizers (for all buffer sizes) which, while achieving near-optimal average distortion during normal operation, also leads to very small probability of filling the buffer because of the large negative net drift of the quantizer used in the upper half of the buffer.

In the single-quantizer achievable region, on the other hand, there is a more stringent tradeoff involved. Since the distortion-rate curve is strictly convex at $R = R_c$, it is clear that the rates of the quantizers have to be close to R_c in order that their average distortion in normal operation can get close to the optimal distortion $D_T(R_c)$. The

closer the average distortion is to the optimal, therefore, the smaller the negative drifts have to be, and the larger the chance of having a full buffer. And there lies the tradeoff.

We make two assumptions on the class of quantizers $\mathcal{Q}$ that are used in the buffer control schemes we shall be dealing with:

1) There exist $\epsilon > 0$ such that $P(|Q(Y)| > R_c) > \epsilon$ for all quantizers $Q \in \mathcal{Q}$.

2) There exists a uniform bound M on the codeword lengths for all $Q \in \mathcal{Q}$.[1]

The following result gives a lower bound on the tradeoff between the average distortion $D_q(L)$ and the fullness probability $p_f(L)$.

Proposition 3.3 *Assume that as the buffer size becomes large, the sequence of buffer control scheme satisfies:* $\lim_{L\to\infty} \max_s |\mu_{L,s}| = 0$, *where* $\mu_{L,s} \equiv E(|Q_s^L(\mathbf{Y})|) - R_c$ *is the net drift of the quantizer. Then the rates of convergence of the steady-state normal distortion* $D_q(L)$ *and of the fullness probability* $p_f(L)$ *are constrained as follows: if* $p_f(L) = o(1/L^2)$, *then* $D_q(L) - D_T(R_c) = \Omega(1/L^2)$.

In essence, the result says that no buffer control scheme can make both $p_f(L)$ and $D_q(L)$ converging at a rate faster than $1/L^2$. If we take the total average distortion as $D_q(L) + D_0 p_f(L)$, then the results implies that a lower bound on the convergence rate of the total distortion to the optimal distortion $D_T(R_c)$ is $1/L^2$.

Sketch of Proof. Here, we consider only buffer control schemes which have drifts symmetrically about the center of the buffer, i.e. $\mu_{L,\frac{L}{2}+s} = -\mu_{L,\frac{L}{2}-s}$ for $s = 0, 1, \ldots, \frac{L}{2}$. The proof for non-symmetrical schemes use similar ideas but require a more elaborate argument, and will not be sketched here.

The key idea is to write an appropriate balance equation for the drifts. Because the drifts are always pointed towards the buffer center, one can argue that the center is the most probable state, so that $\pi_L(\frac{L}{2}) = \Omega(\frac{1}{L})$. If we define a new Markov chain by grouping the entire lower half of the buffer into one state s^* and look at the steady-state balance equation for the new chain, we get:

0 = amount of drift from buffer top + amount of drift from the interior of top half + amount of drift from s^*

[1]This assumption is stronger than necessary, and, with more technical work, can probably be replaced by a condition on the uniform boundedness on certain moments of $|Q(\mathbf{Y})|$.

The first term is due to the effect of reflection at the buffer top, and is proportional to the fullness probability, $p_f(L)$; the second term is lower bounded by $E_{\pi_L}(\mu_{L,s}|L > s > \frac{L}{2})$. Since s^* contains the center state, the third term is lower bounded by a term proportional to $\pi_L(\frac{L}{2})$. Hence, if $p_f(L) = o(\frac{1}{L})$, then $-E_{\pi_L}(\mu_{L,s}|s > \frac{L}{2}) = \Omega(\frac{1}{L})$. By the symmetry of the control scheme, we get

$$E_\pi(|\mu_{L,s}|) = \Omega(\frac{1}{L}) \tag{3.1}$$

On the other hand, by a second-order Taylor series approximation of the distortion-rate function around $R = R_c$, the average normal distortion $D_q(L)$ is roughly given by $D_T(R_c) + \frac{D_T''(R_c)}{2} E_{\pi_L}(\mu_{L,s}^2)$. (The first-order term disappears due to the symmetry in the control scheme.) Since $E_{\pi_L}(\mu_{L,s}^2) \geq (E_{\pi_L}(|\mu_{L,s}|))^2$, the desired lower bound follows from Eq. (3.1). □

We now give a very simple buffer control scheme which can achieve a near-optimal convergence rate for the normal distortion and at the same time having the fullness probability decay at a very fast rate. We first need the following lemma.

Lemma 3.4 *Let $(R_o, D_T(R_o))$ be a point in the interior of a single-quantizer achievable region of the distortion-rate curve. For every R in some neighborhood of R_o, assume that the optimal quantizer is unique and let Q_R be the optimal quantizer. Then*

$$\frac{dr^*(Q_R)}{dR}\,|_{R=R_o} = -\frac{2}{Var(|Q_{R_o}(Y)|)}$$

where, for $R \neq R_o$, $r^*(Q_R)$ is the unique nonzero root of the log moment generating function of the random variable $|Q_R(Y)| - R_o$.

Proposition 3.5 *Given any $K \geq 2$, there exists a buffer control scheme such that $p_f(L) = O(\frac{1}{L^K})$ and $D_q(L) - D_T(R_c) = \Theta(\frac{\ln^2 L}{L^2})$.*

Sketch of Proof. Consider the following buffer control scheme. For each buffer size L, use two quantizers Q_1^L and Q_2^L which achieve points on the distortion-rate curve at rates $R_c - \frac{C \ln L}{L}$ and $R_c + \frac{C \ln L}{L}$ respectively, with Q_1^L used in the top half of the buffer $(A_1)$and Q_2^L used in the bottom half (A_2). Note that in contrast to the scheme for the case when R_c is in the time-shared region, the quantizers are dependent on the buffer size L, with their rates approaching R_c as L becomes large. Using Wald's identity and

renewal theory as in the proof of Prop. (3.2), one can show that the fullness probability is given by:

$$p_f(L) = O(\exp(-\frac{1}{2}r^*(Q_1^L)L))$$

where $r^*(Q_1^L)$ is the unique positive zero of the log moment generating function of the random variable $|Q_1^L(Y)| - R_c$. By Lemma (3.4), we have the approximation

$$r^*(Q_1^L) \approx \frac{2}{Var(|Q_{R_c}(Y)|)}\frac{C \ln L}{L}$$

By choosing $C = K$ Var$(|Q_{R_c}(Y)|)$, we get $p_f(L) = O(L^{-K})$.

When we approximate the average distortion of Q_1^L and Q_2^L via the Taylor series expansion of the distortion-rate function around $R = R_c$, the first-order term disappears due to the symmetry of the scheme, and we have the approximation $D_q(L) = D_T(R_c) + O(\frac{\ln^2 L}{L^2})$, thus giving us the claimed rate of convergence. □

The above result shows that we can almost achieve the lower bound of $1/L^2$ on the rate of convergence of the total distortion while keeping the fullness probability decaying at arbitrarily fast (polynomial) rate. Moreover, the buffer control scheme is exceedingly simple, using two quantizers only. We now compare this performance with the lower bound for the case when there is no buffer control.

Proposition 3.6 *Suppose we are only allowed to use one quantizer for a given buffer size. Then if the fullness probability $p_f(L)$ decays faster than $1/L$, the average distortion $D_q(L)$ converges to the optimal distortion no faster that $1/L$. Hence, for any single-quantizer scheme, the rate of convergence of the total distortion is lower bounded by $1/L$.*

Hence, by using two quantizers with opposing drifts, the rate of convergence is squared. Intuitively, the gain is because with one quantizer of negative drift, the buffer is empty a significant fraction of time, thus wasting the allocated channel rate. In the two quantizer case, this is avoided by keeping the buffer half-full most of the time. The quantizer for the lower half of the buffer actually operates at a rate greater than the channel rate, thus decreasing the average distortion.

4 Conclusions

We have considered the problem of buffer control of variable-rate quantization system for a fixed-rate channel. Asymptotically optimal control schemes , in the regime of

large buffer sizes, have been derived for memoryless sources. Such optimal schemes have a simple form: they consist only of switching between two quantizers, one to be used when the buffer is more than half full and the other when the buffer is less than half full. An improvement over the best achievable without any buffer control is also demonstrated.

References

[BCTK80] Toby Berger, M.U. Chang, and S.Y. Tung-Kleinberg. Quantization-permutation codes and buffer-adapted huffman codes. In *Proceedings of the 18th Allerton Conference on Communications, Control and Computing, Monticello, IL*, pages 433–436, 1980.

[FM86] N. Farvardin and J.W. Modestino. Adaptive buffer-instrumented entropy-coded quantizer performance for memoryless sources. *IEEE Transactions on Information Theory*, 32(1):9–22, January 1986.

[GB92] Robert Gallager and Dimitri Bertsekas. *Data Networks.* Prentice Hall, second edition, 1992.

[HM90] D. Harrison and J.W. Modestino. Analysis and further results on adaptive entropy-coded quantization. *IEEE Transactions on Information Theory*, 36(5):1069–1088, September 1990.

[Hum92] Pierre Humblet. Generalization of huffman coding to minimize the probability of buffer overflow. *IEEE Transactions on Information Theory*, 27(2):230–232, March 1992.

[Jel66] Frederic Jelinek. Buffer overflow in variable length coding of fixed rate sources. *IEEE Transactions on Information Theory*, 14(4):490–501, May 1966.

[Wal44] A. Wald. On cumulative sums of random variables. *Annals on Mathematics and Statistics*, 15:283–296, 1944.

[Wyn74] Aaron Wyner. On the probability of buffer overflow under an arbitrary bounded input-output distribution. *SIAM Journal on Applied Mathematics*, 27:544–569, December 1974.