# SCALABLE PREDICTIVE CODING AS THE WYNER-ZIV PROBLEM

*Anshul Sehgal, Ashish Jagmohan, Narendra Ahuja*

University of Illinois
*asehgal, jagmohan, n-ahuja@uiuc.edu*

## ABSTRACT

An alternative to scalable predictive coding of first order Gauss-Markov processes is proposed in this paper. It is shown that conventional scalable predictive coding is inherently suboptimal. An alternative to scalable predictive coding, which achieves the rate-distortion performance of predictive coding for first-order Gauss-Markov processes is then proposed. The proposed approach is posed as a variant of the well-known Wyner-Ziv problem. By using coset codes with nested lattices, the present paper proves that the proposed approach achieves the predictive coding bound asymptotically at all scales while simultaneously providing the functionality of scalable coding.

## 1. INTRODUCTION

Predictive coding is a commonly used technique for efficient removal of temporal redundancy in video and audio compression systems. In the case of video, the alternative for removal of temporal redundancy is the use of 3D subband coding. However, 3D subband coding suffers from poor compression, introduces artifacts in the reconstructed video, requires high computation and has significant latency (due to the large number of video frames that need to be accumulated before compression). Owing to these drawbacks, predictive coding remains the most viable means of removing temporal redundancy from video streams.

In addition to removal of redundancy, layered coding (also called scalable coding) is an important requirement of any streaming system aimed at reliable coding and transmission of audio/video data over the Internet. A multimedia communication session over the Internet benefits from a compressed multimedia stream that can be decoded at multiple bit rates by serving users with vast variations in their available bandwidth simultaneously. Current day image compression standards such as JPEG and JPEG-2000 have options that allow the generation of an embedded representation of the media being compressed. These standards remove spatial redundancy in images efficiently, while allowing partial decoding of the compressed stream, thus achieving scalable coding/decoding. Video compression, on the other hand, not only requires removal of spatial redundancy, but also temporal redundancy. As mentioned above, this is often accomplished by predictive coding. The problem of video compression becomes especially difficult, if the stipulation of scalable encoding/decoding is also made along with efficient removal of temporal redundancy via predictive coding. The only approach that addresses this problem in an information theoretic setting is the work of Rose and Regunathan [1], however, their scheme does not achieve the rate-distortion bound. To the best of our knowledge, there are no algorithms in literature that achieve the rate-distortion performance of predictive coding while simultaneously offering the functionality of scalability.

The generation of a predictively encoded scalable stream is a difficult task due to the problem of predictive mismatch. Scalable encoding implies that each source symbol can be reconstructed to multiple fidelities. Further, if the source stream is predictively encoded, the reconstruction of each source symbol is used as a predictor for future source symbols. However, since there are multiple possible reconstructions of a scalably coded source symbol, there are multiple possible predictors for future symbols. The difficulty lies in choosing a good predictor from these multiple predictors. In order to achieve good compression, it is desirable to use the highest fidelity reconstruction of each source symbol as a predictor for future symbols, however, it is not necessary that the same predictor will be available while decoding. On the other hand, if a low fidelity reconstruction of a source sample is used as the predictor, this predictor will always be available while decoding (because of the scalability property, a low fidelity reconstruction can be obtained from a high fidelity reconstruction). Unfortunately though, in this case, compression efficiency is sacrificed by the use of an inferior predictor.

In this paper, we propose an alternative to scalable predictive coding that inherits the advantages of predictive coding, and at the same time, mitigates the multiple predictors problem mentioned above. The key observation made is that predictive coding can be posed as a variant of the well-known Wyner-Ziv side-information problem [2]. Section 2 introduces the problem of scalable predictive coding in a formal way. We elaborate on the link between predictive coding and the Wyner-Ziv problem in Section 3. Motivated by this, we use coset-codes (which achieve the Wyner-Ziv bound asymptotically) to design a predictive scalable coder for first-order Gauss-Markov processes. Section 4

0-7803-7510-6/02/$17.00 ©2002 IEEE

proves that the proposed encoding strategy asymptotically achieves the rate-distortion performance for predictive coding of these processes *at all scales*. Finally, conclusions are presented in Section 5.

## 2. SCALABLE PREDICTIVE CODING

Consider the *non-scalable* encoding of the output of a real-valued source with memory $\mathbb{S} = \{v_k\}_{k=1}^{\infty}$ using one-step predictive encoding. Given the reconstruction of source symbol $v_{k-1}$ (denoted $\widehat{v}_{k-1}$), symbol $v_k$ is encoded by generating the innovation $t_k = v_k - E[v_k|\widehat{v}_{k-1}]$, where $E[\cdot]$ denotes the expectation operator. Symbol $t_k$ is lossily compressed to yield $\widehat{t}_k$ which is communicated to the decoder losslessly. The decoder reconstruction of $v_k$ is given by $\widehat{v}_k = \widehat{t}_k + E[v_k|\widehat{v}_{k-1}]$, Thus, the decoder reconstruction $\widehat{v}_{k-1}$ serves as a *predictor* which is used for encoding symbol $v_k$.

Next, consider an encoder connected to multiple decoders via proxy nodes (such as routers). Each decoder is directly lined to a proxy node through a low-delay lossless link. However, the available bandwidth on each link may differ and vary with time. The proxy nodes are connected to the encoder using a low-delay, lossless link with a large bandwidth. The encoder wishes to predictively encode the process $\{v_k\}$ and communicate it to the decoders such that each decoder can utilize its available bandwidth as best as possible. In such a scenario, the non-scalable predictive encoding approach mentioned above will not serve well, as each decoder will operate at the same (fixed) bandwidth (equal to the smallest instantaneous bandwidth among all link at the transmission instant). This would lead to under utilization of resources. In such a scenario, it is desirable to design a system such that each decoder can decode $\{v_k\}$ to the best possible fidelity, given its instantaneous bandwidth constraint. This goal can be accomplished by using scalable predictive coding; at time $k$, a decoder with large instantaneous bandwidth subscribes to a large number of layers and is able to reconstruct $v_k$ to a lower distortion as compared to a decoder with small instantaneous bandwidth at time $k$. This scenario models multicasting to users on heterogeneous links over the Internet. It also demonstrates the functional advantages of layered coding. We allude to this example as an application of layered coding throughout the paper.

Next, consider the scalable predictive coding of $\{v_k\}_{k=1}^{\infty}$ for transmission over the aforementioned setup. The reconstruction of source symbol $v_k$ (denoted as $\widehat{v}_k$) takes one of multiple possible values from a set $\mathbf{R}_k$. For clarity, we consider a two-layer reproduction of each source sample $v_k$; extensions to a larger number of layers is straightforward. Denote the base layer reconstructions of $v_k$ as $\widehat{v}_k^b$ and the enhancement layer reconstruction of $v_k$ as $\widehat{v}_k^e$. Therefore $\mathbf{R}_k = \{\widehat{v}_k^b, \widehat{v}_k^e\}$. If the encoder had precise knowledge of $\widehat{v}_k$ for each decoder, one-step predictive coding and decoding could be performed identically as above. However, this is not always the case for scenarios such as multicasting (as mentioned above) or server based streaming of stored media (where encoding and transmission processes are performed independently of each other). In fact, it is straightforward to see that if conventional predictive coding is used, the set of possible predictors for a source symbol grows exponentially with time index $k$. Thus, optimal predictive coding cannot be performed in these scenarios. The proposed algorithm, on the other hand, avoids the exponential growth of predictors.

In practice, scalable predictive coding is performed as follows: symbol $v_k$ is encoded by generating the innovation $t_k = v_k - E[v_k|\widehat{v}_{k-1}^b]$. Symbol $t_k$ is lossily coded to generate two layers $\widehat{t}_k^b$ and $\widehat{t}_k^e$, the base layer and the enhancement layer, respectively. A coarse reconstruction of $t_k$ is got from $\widehat{t}_k^b$, while a fine reconstruction of $t_k$ is got by adding $\widehat{t}_k^e$ to $\widehat{t}_k^b$. The decoder reconstructs the coarse and fine descriptions of $v_k$ as $\widehat{v}_k^b = \widehat{t}_k^b + E[b_k|\widehat{v}_{k-1}^b]$ and $\widehat{v}_k^e = \widehat{t}_k^e + \widehat{v}_k^b$. We note that this encoding mechanism is inherently suboptimal – even if $\widehat{v}_{k-1}^e$ is available at the decoder, $\widehat{v}_{k-1}^b$ is used as the predictor for $v_k$. In this paper, we present an alternative to this encoding scheme based on coset codes that asymptotically achieves the rate-distortion performance of predictive coding.

### 2.1. Rate-distortion calculations

Consider the problem of predictively encoding a zero-mean first-order Gauss-Markov process $x_k = \rho x_{k-1} + z_k$, where $E[x_k^2] = \sigma_x^2$, $E[z_k^2] = \sigma_z^2$, $\rho^2 = 1 - \sigma_z^2/\sigma_x^2$. The encoder compresses $x_k$ by computing the innovation $t_k = x_k - E[x_k|\widehat{x}_{k-1}] = x_k - \rho\widehat{x}_{k-1}$ and compressing it lossily. It can be shown that with a squared error distortion measure, the predictive coding rate-distortion function for this process is given by:

$$\begin{aligned} R_{PC}(D) &= \tfrac{1}{2}\log\tfrac{\sigma_x^2+\rho^2 D}{D} & 0 \le D < \sigma_x^2 \\ &= 0 & D \ge \sigma_x^2 \end{aligned} \qquad (1)$$

Next, consider a two-layer scalable predictive coding system, as described above. Let $D_1$ denote the base layer distortion and let $D_2$ denote the enhancement layer distortion ($D_2 < D_1$). It can be shown that in this scenario, if base layer decoding is performed for all $x_k$, then the corresponding base-layer R-D function $R^b(D_1) = R_{PC}(D_1)$. If enhancement layer decoding is performed for all $x_k$, the R-D function is given by:

$$\begin{aligned} R^e(D_1, D_2) &= \tfrac{1}{2}\log\tfrac{\sigma_z^2+\rho^2 D_1}{D_2} & 0 \le D_2 < D_1 < \sigma_x^2 \\ &= 0 & D_2 \ge \sigma_x^2 \end{aligned} \qquad (2)$$

Since $D_2 < D_1$, $R^e(D_1, D_2) > R_{PC}(D_2)$. Thus, conventional scalable predictive coding is inherently suboptimal.

## 3. PREDICTIVE CODING AS THE WYNER-ZIV PROBLEM

In [2], Wyner and Ziv introduced the problem of encoding of a continuous random variable $X$ with correlated side-information $Y$ available only at the decoder. Consider two continuous valued correlated Gaussian random variables $X$ and $Y$. The decoder has knowledge of $Y$, but the encoder only knows the joint statistics of $X$ and $Y$. The encoder wishes to compress $X$ as best as it can, even though it does not know $Y$. The encoder accomplishes compression of $X$ by leveraging the correlation between $X$ and $Y$. For $X, Y$ jointly Gaussian, it was shown in [2] that the rate-distortion function for the Wyner-Ziv problem, denoted by $R_{WZ}(D)$, is

$$R_{WZ}(D) = R_{X|Y}(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2_{X|Y}}{D} & 0 \le D \le \sigma^2_{X|Y} \\ 0 & D \ge \sigma^2_{X|Y} \end{cases} \quad (3)$$

where $\sigma^2_{X|Y}$ is the variance of the random variable $X|Y$ and $D$ is the squared error distortion measure.

Next, we show that Equations (1) and (3) are equivalent. In Equation (3), if we substitute $Y$ with $\rho\widehat{x}_{k-1}$ ($\widehat{x}_{k-1}$ is the reconstruction of $x_{k-1}$) and $X$ with $x_k$, $\sigma^2_{X|Y} = \sigma^2_{x_k|\widehat{x}_{k-1}} = \sigma_z^2 + \rho^2 D$. Also, $0 \le D < \sigma^2_{X|Y}$ in Equation (3) implies $D < \sigma_z^2 + \rho^2 D \Rightarrow D < \sigma_z^2/(1-\rho^2) \Rightarrow D < \sigma_x^2$. Making these substitutions in Equation (3), we see that Equations (1) and (3) are the same. Thus, any code that achieves the WZ bound for Gaussian side-information problems, should also achieve the rate-distortion function of a predictively coded first-order Gauss-Markov process. Fortunately, there are codes that achieve the WZ bound for Gaussian sources. Of course, it is not apparent what benefit framing predictive coding as the WZ problem will have. However, as we shall see, if the process $\{x_k\}_{k=1}^{\infty}$ is scalably predictively coded using the WZ approach, it is possible to achieve the predictive coding rate-distortion bound for both, the base and the enhancement layers. As discussed earlier, this is not the case if $\{x_k\}_{k=1}^{\infty}$ is scalably predictively coded using the conventional approach (as described in Section 2).

## 4. SCALABLE PREDICTIVE CODING USING COSET CODES

This section describes the proposed algorithm for scalable predictive coding. Zamir and Shamai [3] have shown that nested lattice codes achieve the WZ bound for correlated Gaussian sources asymptotically. Further details of their approach can be found in [3, 4]. We modify their solution for the problem of scalable predictive coding.

### 4.1. Preliminaries

The proposed coding algorithm uses nested lattices to encode the source process. An $n$-dimensional lattice $\Lambda$ is defined by a set of $n$ basis vectors $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_n$ in $\mathbb{R}^n$. Each lattice point is generated as a linear combination of these basis vectors with integer coefficients, i.e. $\Lambda = \{\mathbf{p} = \mathbf{G}\mathbf{a} : \mathbf{a} \in \mathbb{Z}^n\}$, $\mathbf{G} = [\mathbf{g}_1|\mathbf{g}_2|\ldots|\mathbf{g}_n]$.

The nearest neighbor quantizer $Q(\cdot) : \mathbb{R}^n \to \Lambda$ is defined as

$$Q(\mathbf{x}) = \arg\min_{\mathbf{p}\in\Lambda} ||\mathbf{x} - \mathbf{p}|| \quad (4)$$

where $\mathbf{x} \in \mathbb{R}^n$. The basic Voronoi region $\mathcal{V} = \{\mathbf{x} : Q(\mathbf{x}) = 0, \mathbf{x} \in \mathbb{R}^n\}$. The volume of the basic Voronoi cell and its second moment are denoted by $V = \int_{\mathcal{V}} dx$ and $\sigma^2 = \frac{1}{nV}\int_{\mathcal{V}} ||x||^2 dx$, respectively. The normalized second moment of lattice $\Lambda$ is denoted by $G = \sigma^2/V^{n/2}$ [5]. The "mod $\Lambda$" operation on $\mathbf{x} \in \mathbb{R}^n$ is defined as

$$\mathbf{x} \bmod \Lambda = \mathbf{x} - Q(\mathbf{x}) \quad (5)$$

Thus, $\mathbf{x} \bmod \Lambda$ is the quantization error when $\mathbf{x}$ is quantized on lattice $\Lambda$.

A pair of $n$-dimensional lattices $(\Lambda_1, \Lambda_2)$ is nested if $\mathbf{p} \in \Lambda_2 \Rightarrow \mathbf{p} \in \Lambda_1$. We denote a pair of nested lattices as $\Lambda_2 \subseteq \Lambda_1$. This implies that the corresponding generator matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ are related as $\mathbf{G}_2 = \mathbf{G}_1 \cdot \mathbf{J}$, where $\mathbf{J}$ is an $n \times n$ matrix with integer entries and $\det(\mathbf{J}) \ge 1$. For n-dimensional lattices $\Lambda_1$ and $\Lambda_2$, with $\Lambda_2 \subseteq \Lambda_1$, the volume of the basic Voronoi cells [5] are related as $V_2/V_1 = \det(\mathbf{J}) = (G_2/G_1)^{2/n}$, where $G_1$, $G_2$ are the normalized second moments of lattices $\Lambda_1$,$\Lambda_2$, respectively.

### 4.2. Proposed approach

We describe the proposed approach for a two-layer encoding system, i.e. each source sample can be decoded to two fidelity levels. The proposed lattice coding algorithm uses two pairs of unbounded lattices $\Lambda_i^c$, $\Lambda_i^s$, $i = 1, 2$, such that $\Lambda_i^c \in \mathbb{R}^n$ and $\Lambda_i^s \in \mathbb{R}^n$. Lattices $\Lambda_1^c$, $\Lambda_2^c$ denote the channel-coding lattices and lattices $\Lambda_1^s$, $\Lambda_2^s$ denote the source coding lattices. Any source sample $\mathbf{x}$ is quantized onto $\Lambda_i^s$ using the nearest neighbor quantizing functions $Q_i^s(\cdot) : \mathbb{R}^n \to \Lambda_i^s$. Similarly, the quantization function for the channel coding lattices $\Lambda_i^c$ is denoted by $Q_i^c(\cdot) : \mathbb{R}^n \to \Lambda_i^c$. The basic Voronoi cell, volume of the basic Voronoi cell, second moment and normalized second moment of $\Lambda_i^s$ lattices are defined as $\mathcal{V}_i^s$, $V_i^s$, $\sigma_i^{s\,2}$ and $G_i^s$ respectively, for $i = 1, 2$. The

corresponding quantities for the channel coding lattices are defined as $\mathcal{V}_i^c$, $V_i^c$, ${\sigma_i^c}^2$ and $G_i^c$ respectively, for $i = 1, 2$.

Consider the two-layer scalable predictive coding of a vector zero-mean first-order Gauss-Markov process $\mathbf{x}_k = \rho\mathbf{x}_{k-1} + \mathbf{z}_k$ with $\mathbf{x}_k \in \mathbb{R}^n$ and $\mathbf{z}_k \in \mathbb{R}^n$; $E[\mathbf{x}_k\mathbf{x}_k^t] = \sigma_x^2\mathbf{I}_n$; $E[\mathbf{z}_k\mathbf{z}_k^t] = \sigma_z^2\mathbf{I}_n$, where $\mathbf{I}_n$ denotes an $n \times n$ identity matrix. Let $d_1$ and $d_2$ (such that $d_1 > d_2$) denote the desired per symbol expected squared-error distortion for layer 1 and layer 2, respectively.

The lattice pairs $(\Lambda_1^c, \Lambda_1^s)$ and $(\Lambda_2^c, \Lambda_2^s)$ are chosen such that they satisfy the following properties:

I $\Lambda_i^c \subseteq \Lambda_i^s$, $i = 1, 2$.

II $\sigma_i^s = \sqrt{d_i}$, $i = 1, 2$.

III $\sigma_i^c = \sqrt{(1+\rho^2)d_i + \sigma_z^2 + \epsilon}$, $i = 1, 2$, where $\epsilon > 0$ is an arbitrarily small number.

IV $\Pr\{Q_i^c(\mathbf{z}^*) \neq 0\} < \epsilon$, $i = 1, 2$, where $\mathbf{z}^* \in \mathcal{R}^n$ is zero-mean Gaussian with variance $((1+\rho^2)d_i + \sigma_z^2)$.

V As the dimensionality of the lattices $n \to \infty$, $G_i^c \to \frac{1}{2\pi e}$ and $G_i^s \to \frac{1}{2\pi e}$, i.e. the shape of $\mathcal{V}_i^c$ and $\mathcal{V}_i^s$ tends to a sphere.

The reader is referred to [3, 4] for a discussion on the existence of such lattices.

#### 4.2.1. *Encoding algorithm*

It is desired that at time $k$, each decoder have the choice of decoding vector $\mathbf{x}_k$ to one of two fidelity levels, $d_1$ or $d_2$. In order to accomplish this task, the encoder could simply quantize $\mathbf{x}_k$ onto the two lattices $\Lambda_1^s$ and $\Lambda_2^s$ and transmit the index of the quantizer output to the decoder. If at time $k$, the decoder had adequate bandwidth to receive $Q_1^s(\mathbf{x}_k)$, it could decode $\mathbf{x}_k$ to distortion $d_1$ and if it had adequate bandwidth to receive $Q_2^s(\mathbf{x}_k)$, it could decode $\mathbf{x}_k$ to distortion $d_2$. However, by encoding/decoding in this manner, the correlation between $\mathbf{x}_k$ and $\widehat{\mathbf{x}}_{k-1}$, the decoder reconstruction of $\mathbf{x}_{k-1}$, is not exploited to improve the compression efficiency of the system. Lattices $\Lambda_1^c$ and $\Lambda_2^c$ accomplish this task of improving the compression efficiency.

Consider the communication of $Q_1^s(\mathbf{x}_k)$ using the lattice pair $(\Lambda_1^c, \Lambda_1^s)$. Instead of transmitting $Q_1^s(\mathbf{x}_k)$ to the decoder, $\mathbf{d}_k^1 = Q_1^s(\mathbf{x}_k) \text{mod } \Lambda_1^c$ is transmitted to the decoder. Because of the "mod " operation, when the decoder receives $\mathbf{d}_k^1$, the only knowledge it has of $Q_1^s(\mathbf{x}_k)$ is that $Q_1^s(\mathbf{x}_k)$ is from the set $\mathcal{C} = \{\mathbf{p} + \mathbf{d}_k^1 : \mathbf{p} \in \Lambda_1^c\}$. The decoder leverages its knowledge of $\widehat{\mathbf{x}}_{k-1}$ to ascertain which member of $\mathcal{C}$ is $Q_1^s(\mathbf{x}_k)$. Let $d_{min}(\Lambda_1^c)$ denote the smallest distance between any two lattice points of $\Lambda_1^c$. A little thought reveals that if it can be ensured that $||Q_1^s(\mathbf{x}_k) - \rho\widehat{\mathbf{x}}_{k-1}|| < d_{min}(\Lambda_1^c)/2$, the decoder can unambiguously decode $Q_1^s(\mathbf{x}_k)$ as the point nearest to $\rho\widehat{\mathbf{x}}_{k-1}$ in the set $\mathcal{C}$ [4]. In general though, the condition $||Q_1^s(\mathbf{x}_k) - \rho\widehat{\mathbf{x}}_{k-1}|| < d_{min}(\Lambda_1^c)/2$ cannot be ensured. However, based on typicality arguments, as the dimensionality of the lattices $\Lambda_1^c$ and $\Lambda_1^s$ tends to $\infty$, it can be shown that for Gaussian sources, the above mentioned condition can be ensured with probability arbitrarily close to one. Similarly, $Q_2^s(\mathbf{x}_k)$ can be communicated to the decoder by transmitting $\mathbf{d}_k^2 = Q_2^s(\mathbf{x}_k) \text{ mod } \Lambda_2^c$. To facilitate switching between the two layers, a third description $\mathbf{d}_k^3 = Q_2(\mathbf{x}_k - Q_1(\mathbf{x}_k))$ also needs to be generated.

Hence, the proposed encoder works as follows: the encoder generates three descriptions at each time step $k$. A subset of these descriptions is required for decoding $\mathbf{x}_k$ to a particular fidelity level, the elements of the subset depend on the desired distortion to which $\mathbf{x}_k$ is to be decoded and the distortion to which $\mathbf{x}_{k-1}$, the predictor vector for $\mathbf{x}_k$ was decoded.

The three descriptions for $\mathbf{x}_k$ are:

1. $\mathbf{d}_k^1 = Q_1(\mathbf{x}_k) \text{mod } \Lambda_1^c$
2. $\mathbf{d}_k^2 = Q_2(\mathbf{x}_k) \text{mod } \Lambda_2^c$
3. $\mathbf{d}_k^3 = Q_2(\mathbf{x}_k - Q_1(\mathbf{x}_k))$

At time $k$, each decoder requests its proxy node for a subset of the above three descriptions generated by the encoder at time $k$. The subset requested depends on the bandwidth available on the proxy-decoder link at time $k$ and the fidelity to which $\mathbf{x}_{k-1}$ was decoded by the decoder.

Before describing the decoding procedures, we calculate the per-symbol encoding rates $R_i$ of $\mathbf{d}_k^i$, $i = 1, 2, 3$. The encoding rate $R_1$ is given by

$$\begin{aligned} R_1 &= \frac{1}{n}\log\left(\frac{V_1^c}{V_1^s}\right) + O\left(\frac{1}{n}\right) \\ &= \frac{1}{2}\log\left(\frac{{\sigma_1^c}^2 G_1^c}{{\sigma_1^s}^2 G_1^s}\right) + O\left(\frac{1}{n}\right) \quad (6) \\ &= \frac{1}{2}\log\left(\frac{\sigma_z^2 + \rho^2 d_1}{d_1} + 1\right) + O\left(\frac{1}{n}\right) + O(\epsilon) \quad (7) \end{aligned}$$

Equation (7) is got from Equation (6) by substituting conditions II, III and V for $i = 1$. Thus, as $n \to \infty$, the rate of transmission of $\mathbf{d}_k^1$ can be made arbitrarily close to $\frac{1}{2}\log\left(\frac{\sigma_z^2+\rho^2 d_1}{d_1} + 1\right)$.

The calculation of $R_2$ is similar to that of $R_1$. Thus, $R_2$ is given by

$$R_2 = \frac{1}{2}\log\left(\frac{\sigma_z^2 + \rho^2 d_2}{d_2} + 1\right) + O\left(\frac{1}{n}\right) + O(\epsilon). \quad (8)$$

Next, we calculate $R_3$. We note that $R_3$ is the rate of encoding the quantization error $\mathbf{x}_k - Q_1^s(\mathbf{x}_k)$ quantized on

lattice $\Lambda_2^s$. From high resolution quantization theory, it is known that in the limit as $d_1 \to 0$, the quantization error $\mathbf{x}_k - Q_1^s(\mathbf{x}_k)$ tends to a uniform distribution over the basic Voronoi region $\mathcal{V}_1^s$ of lattice $\Lambda_1^s$. As the dimensionality $n$ of lattice $\Lambda_1^s \to \infty$, from condition V, we note that the shape of $\mathcal{V}_1^s$ tends to a sphere. From the asymptotic Gaussianity of a uniform distribution over a sphere, we note that $\mathbf{x}_k - Q_1^s(\mathbf{x}_k)$ tends to a Gaussian distribution with variance $(\sigma_1^s)^2$. Thus, $R_3$ is given by

$$R_3 = \log\left(\frac{\sigma_1^s}{\sigma_2^s}\right) = \frac{1}{2}\log\left(\frac{d_1}{d_2}\right). \tag{9}$$

The decoding procedures are described in the following Section.

#### 4.2.2. *Decoding algorithm*

The decoder reconstruction of symbol $\mathbf{x}_k$ (denoted $\widehat{\mathbf{x}}_k$) takes values from the set $\mathbf{R}_k = \{\widehat{\mathbf{x}}_k^b, \widehat{\mathbf{x}}_k^e\}$, where $\widehat{\mathbf{x}}_k^b = Q_1^s(\mathbf{x}_k)$ and $\widehat{\mathbf{x}}_k^e = Q_2^s(\mathbf{x}_k)$. As mentioned earlier, the subset of descriptions that the decoder retrieves from the proxy at time $k$ depends on the bandwidth available on the decoder-proxy link at time $k$ and the decoder reconstruction of $\mathbf{x}_{k-1}$. The decoder could reconstruct $\mathbf{x}_k$ to distortion $d_1$ or $d_2$. The bandwidth required for either case depends on whether $\mathbf{x}_{k-1}$ was reconstructed to $d_1$ or $d_2$. Hence, there are four scenarios, corresponding to the distortion to which $\mathbf{x}_k$ and $\mathbf{x}_{k-1}$ are decoded. Each of these is considered individually *below*.

*Case I:* $\widehat{\mathbf{x}}_{k-1} = \widehat{\mathbf{x}}_{k-1}^b$, $\widehat{\mathbf{x}}_k = \widehat{\mathbf{x}}_k^b$ – Consider the case when symbol $\mathbf{x}_{k-1}$ was decoded to distortion $d_1$ and the decoder wishes to reconstruct $\mathbf{x}_k$ to $d_1$ as well. Therefore, $\widehat{\mathbf{x}}_{k-1} = \widehat{\mathbf{x}}_{k-1}^b$. In this case, the decoder requests the proxy for symbol $\mathbf{d}_k^1$ and reconstructs $\widehat{\mathbf{x}}_k$ using $\mathbf{d}_k^1$ and $\widehat{\mathbf{x}}_{k-1}^b$ (the base layer reconstruction of $\mathbf{x}_{k-1}$). Symbol $\widehat{\mathbf{x}}_k$ is given by

$$\widehat{\mathbf{x}}_k = \arg\min_{\mathbf{p}\in\Lambda_1^c+\mathbf{d}_k^1} \|\rho\widehat{\mathbf{x}}_{k-1}^b - \mathbf{p}\|. \tag{10}$$

As shown in [4], Equation (10) can be rewritten as

$$\widehat{\mathbf{x}}_k = \rho\widehat{\mathbf{x}}_{k-1}^b + (\widehat{\mathbf{x}}_k^b \bmod \Lambda_1^c - \rho\widehat{\mathbf{x}}_{k-1}^b) \bmod \Lambda_1^c \tag{11}$$

$$= \rho\widehat{\mathbf{x}}_{k-1}^b + (\widehat{\mathbf{x}}_b^k - \rho\widehat{\mathbf{x}}_{k-1}^b) \bmod \Lambda_1^c. \tag{12}$$

Conditioned on correct decoding (i.e. $\widehat{\mathbf{x}}_k = \widehat{\mathbf{x}}_k^b$), the expected per symbol squared distortion of $\widehat{\mathbf{x}}_k$ is the second moment of the lattice $\Lambda_1^s$, $d_1$. Next, we show that the probability of incorrect decoding tends to zero asymptotically, as the dimensionality of the lattices $\Lambda_1^s$ and $\Lambda_1^c \to \infty$ and $d_1 \to 0$. The probability of decoding failure is given by

$$\begin{aligned} Pr(\widehat{\mathbf{x}}_k \neq \widehat{\mathbf{x}}_k^b) &= Pr(\rho\widehat{\mathbf{x}}_{k-1}^b + (\widehat{\mathbf{x}}_b^k - \rho\widehat{\mathbf{x}}_{k-1}^b) \bmod \Lambda_1^c \neq 0) \\ &= Pr((\widehat{\mathbf{x}}_b^k - \rho\widehat{\mathbf{x}}_{k-1}^b) \bmod \Lambda_1^c \\ &\quad \neq (\widehat{\mathbf{x}}_b^k - \rho\widehat{\mathbf{x}}_{k-1}^b)) \\ &= Pr\{Q_1^c(\widehat{\mathbf{x}}_b^k - \rho\widehat{\mathbf{x}}_{k-1}^b) \neq 0\} \\ &= Pr((\widehat{\mathbf{x}}_k^b - \rho\widehat{\mathbf{x}}_{k-1}^b) \notin \mathcal{V}_1^c). \end{aligned} \tag{13}$$

In order to compute this probability, we compute the distribution of $\widehat{\mathbf{x}}_k^b - \rho\widehat{\mathbf{x}}_{k-1}^b$. It is noted that as $d_1 \to 0$, for any $\mathbf{v} \in \mathbb{R}^n$, the distribution of $\mathbf{v} \bmod \Lambda_1^s$ tends to an independent uniform distribution over $\mathcal{V}_1^s$. Further, as $n \to \infty$, from condition V, $\mathcal{V}_1^s$ tends to a sphere. From the asymptotic Gaussianity of a uniform distribution over a sphere, we note that the distribution of $\mathbf{v} \bmod \Lambda_1^q$ tends to a Gaussian distribution with mean zero and variance $d_1$. Another assumption that is often made is that $\mathbf{v} \bmod \Lambda_1^s$ is independent of $\mathbf{v}$ [6]. Hence, asymptotically,

$$\begin{aligned} \widehat{\mathbf{x}}_k^b - \rho\widehat{\mathbf{x}}_{k-1}^b &= Q_1^s(\mathbf{x}_k) - \rho Q_1^s(\mathbf{x}_{k-1}) \\ &= \mathbf{x}_k - \mathbf{x}_k \bmod \Lambda_1^s - \rho\mathbf{x}_{k-1} \\ &\quad + \rho\{\mathbf{x}_{k-1} \bmod \Lambda_1^s\} \\ &= \mathbf{z}_k - \mathbf{x}_k \bmod \Lambda_1^s + \rho\{\mathbf{x}_{k-1} \bmod \Lambda_1^s\} \\ &\sim \mathcal{N}(0, \sigma_z^2 + (1+\rho^2)d_1), \end{aligned} \tag{14}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean $\mu$ and variance $\sigma^2$. Let $\mathbf{z}'' \sim \mathcal{N}(0, \sigma_z^2 + (1+d_1)\rho^2)$ Thus, $\Pr((\widehat{\mathbf{x}}_k^b - \rho\widehat{\mathbf{x}}_{k-1}^b) \notin \mathcal{V}_1^c) = \Pr(\mathbf{z}'' \notin \mathcal{V}_1^c)$. From property IV, $\Pr(\mathbf{z}'' \notin \mathcal{V}_1^c) < \epsilon$. Hence, asymptotically, the probability of error can be made arbitrarily close to 0.

Finally, consider a decoder which decodes $\{\mathbf{x}_k\}_{k=1}^{\infty}$ to distortion $d_1$ *for all* $k$. From Equation (7), asymptotically, as $d_1 \to 0$, and the dimensionality of the lattices $n \to \infty$, it is possible to achieve a per-symbol rate of transmission, $R$ arbitrarily close to

$$R = \frac{1}{2}\log\left(\frac{\sigma_z^2 + \rho^2 d_1}{d_1} + 1\right) \tag{15}$$

Equation (15) and Equation (1) are almost the same. The small loss in rate at small values of $\frac{\sigma_z^2+\rho^2 d_1}{d_1}$ is due to the effect of self-noise [4]. As shown by De Buda [7], a similar loss in capacity is observed when lattice decoding is performed on bounded lattices. Thus, in the limit as $d_1 \to 0$ and $n \to \infty$, *the proposed algorithm approaches the rate-distortion performance of predictive coding* for decoding at layer 1.

*Case II:* $\widehat{\mathbf{x}}_{k-1} = \widehat{\mathbf{x}}_{k-1}^e$, $\widehat{\mathbf{x}}_k = \widehat{\mathbf{x}}_k^e$ – Next, consider the case when the decoder reconstruction of the predictor vector $\widehat{\mathbf{x}}_{k-1} = \widehat{\mathbf{x}}_{k-1}^e$, (i.e. the decoder reconstructs $\mathbf{x}_{k-1}$ to distortion $d_2$), and wishes to decode $\mathbf{x}_k$ to distortion $d_2$ as well. In this case, the decoder requests the proxy for $\mathbf{d}_k^2$. The decoding rule, the distortion of the reconstructed vector $\widehat{\mathbf{x}}_k$ and the proof that $\Pr(\widehat{\mathbf{x}}_k \neq \mathbf{x}_k^e) \to 0$ is similar to

the case when the predictor and the current symbol are decoded to distortion $d_1$ with quantities defined for the lattice pair $(\Lambda_1^c, \Lambda_1^s)$ replaced by those for the lattice pair $(\Lambda_2^c, \Lambda_2^s)$. The per-symbol rate of transmission in this case is given by Equation (8).

We note that the asymptotic performance of *a decoder decoding at layer 2 also approaches the rate-distortion performance of predictive coding*. Moreover, this is simultaneously achieved with asymptotically optimal performance for a decoder decoding at layer 1. It is also noted that this is not possible with conventional scalable predictive coding. As mentioned in the Introduction, achieving the same performance with conventional layered predictive coding results in an exponential growth in the number of predictors with time. An alternative strategy to achieve the predictive coding distortion-rate bound would be to use two independent encoding loops operating at distortion levels $d_1$ and $d_2$. However, if this were done, it would not be possible for a decoder to switch between the two encoders due to the problem of predictive mismatch. As shown in Cases III and IV below, the proposed approach offers the decoders the option of switching between the two distortion levels, thus achieving scalability.

*Case III:* $\widehat{\mathbf{x}}_{k-1} = \widehat{\mathbf{x}}_{k-1}^b$, $\widehat{\mathbf{x}}_k = \widehat{\mathbf{x}}_k^e$ – Next, consider the case when a decoder had adequate bandwidth, so as to be able to decode $\mathbf{x}_{k-1}$ to distortion $d_1$, but wishes to decode $\mathbf{x}_k$ to distortion $d_2$. In this case, the decoder requests the proxy for coefficients $\mathbf{d}_k^1$ and $\mathbf{d}_k^3$. It decodes $\mathbf{x}_k$ to distortion $d_1$ as shown above. It then refines $\widehat{\mathbf{x}}_k = \widehat{\mathbf{x}}_k^b$ by adding $\mathbf{d}_k^3$ to it. Thus, $\mathbf{x}_k$ is reconstructed to distortion $d_2$. From Equations (7) and (9), the transmission rate required to accomplish this is

$$R = R_1 + R_2 = \frac{1}{2}\log\left(\frac{\sigma_z^2 + (1+\rho^2)d_1}{d_2}\right) \qquad (16)$$

*Case IV:* $\widehat{\mathbf{x}}_{k-1} = \widehat{\mathbf{x}}_{k-1}^e$, $\widehat{\mathbf{x}}_k = \widehat{\mathbf{x}}_k^b$– Lastly, consider the case when $\widehat{\mathbf{x}}_{k-1} = \mathbf{x}_{k-1}^e$ (i.e. the predictor is decoded to distortion $d_2$) and the decoder wishes to decode $\mathbf{x}_k$ to distortion $d_1$. In this case, the decoder requests the proxy for $\mathbf{d}_k^1$ and decodes $\mathbf{x}_k$ in a manner similar to Equation (10),

$$\widehat{\mathbf{x}}_k = \arg\min_{\mathbf{p}\in\Lambda_2^s+\mathbf{d}_k^2} ||\rho\widehat{\mathbf{x}}_{k-1}^e - \mathbf{p}||. \qquad (17)$$

Asymptotically, $\Pr(\widehat{\mathbf{x}}_k = \widehat{\mathbf{x}}_k^b) \to 1$, since the "distance" of $\widehat{\mathbf{x}}_{k-1}^e$ from $\widehat{\mathbf{x}}_k^b$ is smaller than that of $\widehat{\mathbf{x}}_{k-1}^b$ from $\widehat{\mathbf{x}}_k^b$ and we have already shown that the probability of decoding error tends to zero when $\widehat{\mathbf{x}}_{k-1}^b$ is used for decoding.

To summarize, we have shown that using the proposed approach, in the limit as $d_1 \to 0$, $d_2 \to 0$ (such that the ratio $d_1/d_2$ is fixed) and $n \to \infty$, the performance of the proposed algorithm approaches the rate-distortion function of predictively coded first-order Gauss-Markov processes. Also, the performance of the proposed approach is superior to that of conventional scalable predictive coding, which is inherently suboptimal.

## 5. CONCLUSIONS

In summary, the proposed approach achieves the distortion-rate lower bound for predictive coding, while achieving the functionality of layered coding. We note that even though the optimality of the proposed approach was proven using lattices, it might be easier to build practical systems based on the proposed idea using other approaches such as trellis based approaches [8], since these have lower decoding complexity as opposed to lattices, or training based codebook designs, since they offer more design flexibility as compared to lattices.

It should be noted that the sum total of the transmission rate from the encoder to the proxy is $R_1 + R_2 + R_3$ per source symbol. This is greater than the rate of transmission for conventional scalable predictive coding. However, the rate of tranmssion on the decoder-proxy link is reduced to that of the rate-distorion function for predictive coding.

# Acknowledgements

The authors wish to thank Dr. P. A. Chou, Microsoft Research for his help through out the course of this work.

## 6. REFERENCES

[1] K. Rose and S.L. Regunathan, "Toward optimality in scalable predictive coding," *IEEE Transactions on Image Processing*, vol. 10, pp. 965–976, July 2001.

[2] A.D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Info. theory*, vol. 22, pp. 1–10, Jan. 1976.

[3] R. Zamir and S. Shamai, "Nested linear / lattice codes for wyner-ziv encoding," in *Information Theory Workshop, Killarney, Ireland*, June 1998, pp. 92–93.

[4] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Info. theory*, vol. 48, pp. 1250–1276, June 2002.

[5] J. H. Conway and N. J. A. Sloane, *Sphere Packing, Lattices and Groups*, Springer-Verlag, 44 Hartz Way, Secaucus, NJ 07096, 3 edition, 1998.

[6] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Info. theory*, vol. 44, pp. 2325–2383, Oct. 1998.

[7] R. de Buda, "Some optimal codes have structure," *IEEE Journal on Selected Areas in Communications*, vol. 7, pp. 893–899, Aug. 1989.

[8] S.S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (discus): design and construction," in *IEEE Data Compression Conference*, 1999, pp. 158–167.