DATA-DRIVEN OPTIMIZATION FOR ZERO-DELAY LOSSY SOURCE CODING WITH SIDE INFORMATION

Elad Domanovitz, Daniel Severo, Ashish Khisti and Wei Yu

Department of Electrical and Computer Engineering
University of Toronto, Toronto, ON M5S 3G4, Canada
E-mails: elad.domanovitz@utoronto.ca, d.severo@mail.utoronto.ca,
akhisti@ece.utoronto.ca, weiyu@ece.utoronto.ca

ABSTRACT

This paper proposes a data-driven architecture for zero-delay lossy source coding with side information (i.e., Wyner-Ziv coding) for sources with memory. The overall architecture involves designing suitable filters at the encoder and the decoder and performing fixed-rate scalar quantization followed by one-dimensional binning of quantization indices. Unlike previous work, which uses an exhaustive search to optimize the system parameters, this paper proposes a lower-complexity data-driven method that does not require a priori knowledge of source and side information statistics. The main ingredients of the proposed approach include modeling the quantization process by an additive quantization noise process, modeling the modulo operation by a continuous approximation, and approximating the decoding process by a softmin function, which makes the system amenable to training using stochastic gradient descent. Experimental results on Gauss-Markov sources with different memory orders demonstrate that our proposed system can match the performance of systems optimized using an exhaustive search.

1. INTRODUCTION

This paper considers a source coding problem in which a source sequence is to be compressed in a lossy fashion at the encoder and to be reconstructed by a decoder which has access to a side information sequence unknown to the encoder. In this setting, known as Wyner-Ziv coding [1], significant improvement in compression rate is theoretically possible even though the side-information sequence is only available at the decoder. In fact, for memoryless Gaussian source and side information sequences, the rate-distortion function for Wyner-Ziv coding is same as the case when the side information is also available at the encoder. To implement Wyner-Ziv coding, Zamir et al [2] proposed a structured algebraic binning scheme based on a pair of nested linear/lattice codes for binary symmetric and quadratic Gaussian sources and demonstrated that the Wyner-Ziv rate-distortion function is asymptotically achievable as the dimensions of the lattices go to infinity. High-dimensional lattice codes are, however, difficult to implement in practice. An alternative approach is to use low-dimensional lattices to perform quantization, then to apply lossless source coding with side information over long blocks of quantization indices [3]. References [4, 5] showed that if ideal lossless coding is assumed, such methods can also come close to Wyner-Ziv rate-distortion function for Gaussian sources.

The aforementioned works involve coding over long blocks of source samples and are not suitable for real-time applications such as traffic monitoring, hazard detection and intrusion surveillance, where decisions need to be made under strict delay constraints. This present work considers the case of strict zero delay. In such a setting, the theoretical Wyner-Ziv rate-distortion bound cannot be achieved. For memoryless sources, a symbol-by-symbol coding scheme that consists of scalar quantization followed by a one-dimensional binning is proposed in [4,6] and shown to achieve within approximately 10dB in signal-to-quantization-noise ratio (SQNR) from the ideal Wyner-Ziv bound. Zero-delay setting when the source has memory is considered in [6, 7] and extended in [8]. The authors of [9] propose an architecture involving predictive coding of the input, followed by one-dimensional scalar quantization and scalar binning. These works show significant improvement over ignoring side information at the receiver, and are able to approach within approximately 10dB from the Wyner-Ziv bound for source and side information with memory just as in the memoryless case. However in these prior works, exhaustive numerical search is required for optimizing over the suggested architecture, and the numerical experiments are limited to first-order Gauss-Markov sources.

This paper suggests a data-driven iterative optimization for zero-delay Wyner-Ziv coding for sources with arbitrary memory. While it builds upon the architectures in [7, 8], the proposed approach has lower complexity as it does not require an exhaustive numerical search. It also does not rely on the knowledge of the statistics of the source or side-information sequences. The proposed method is based on stochastic gradient descent (SGD) and involves simultaneously updating the filter coefficients and quantization parameters to minimize the reconstruction error. Towards this end, we replace the feedback prediction filters in [8] with feedforward filters and replace the quantization and decoding operators with approximations that are differentiable and amenable to training. The proposed method demonstrates performance close to that of exhaustive search in [8].

As related works, data-driven optimization for distributed source coding has been considered in [10,11]. In these works, the optimization of each component of the system (prediction at the encoder, binning, prediction at the decoder) is performed separately at each iteration while keeping the other components fixed. In contrast, the proposed method is based on SGD and involves simultaneously updating all the parameters in each iteration. Further, these papers consider the problem of multi-terminal source coding [12] rather than source coding with side information as in [8] and the present work.

2. PRACTICAL WYNER-ZIV CODING

We first introduce symbol-by-symbol Wyner-Ziv coding for memoryless sources as in [2] then show how the scheme can be extended to sources in memory following [6, 8].

This work was supported by Huawei Technologies Canada.

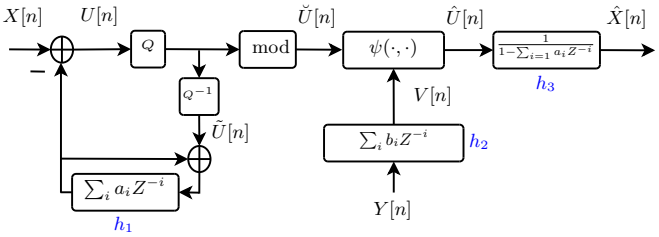

Fig. 1. Wyner-Ziv Coding with Predictive Filters

2.1. Symbol-by-symbol Wyner-Ziv coding

Assume that the source, side-information pair $(X, Y) \in \mathbb{R} \times \mathbb{R}$ has a probability density function (pdf) that satisfies $p_{XY}(x, y) > 0$ everywhere. A fixed-length symbol-by-symbol Wyner-Ziv code of rate R consists of a function $\phi : \mathbb{R} \to \{0, 1\}^R$ for encoding X, and a decoding function $\psi : \{0, 1\}^R \times \mathbb{R} \to \mathbb{R}$ for reconstructing X based on the encoded symbol and the side information Y, and a target distortion function, which is assumed to be the mean-squared distortion $D = \mathbb{E}[(X - \psi(\phi(X), Y))^2]$.

This paper considers a *nested scalar quantization* approach [5] to Wyner-Ziv coding on a symbol-by-symbol basis. Let $\mathcal{I}_\Delta = \{k\Delta : k \in \mathbb{Z}\}$ denote the set of representation points of a uniform scalar quantizer where the spacing between the points is Δ. Given any $X \in \mathbb{R}$ the quantization function $\tilde{x} = Q(x)$ is given by

$$Q(x) = \arg \min_{\tilde{x} \in \mathcal{I}_\Delta} |x - \tilde{x}|. \tag{1}$$

Next, apply a one-dimensional modulo operation to the quantized output to reduce the output to one of J values. In particular, if $Q(x) = k\Delta$ for some $k \in \mathbb{Z}$ then the output is given by $\phi(x) = j\Delta$ where $j = k \mod J$, which is in the set $\{0, \ldots, J - 1\}$. The quantization rate is given by $R = \log_2 J$.

Given the side information $Y = y$, and $\breve{x} = \phi(x) \in \{0, \Delta, \ldots, (J - 1)\Delta\}$, the decoding function is given by [8],

$$\hat{x} = \psi(\breve{x}, y) = \arg \min_{\tilde{x} \in \mathcal{I}_\Delta : \phi(\tilde{x}) = \breve{x}} |\tilde{x} - \mathbb{E}[X|Y = y]|. \tag{2}$$

Intuitively, the modulo operation is used to reduce the encoding rate, while relying on the decoder to search among the reconstruction points in $\mathcal{I}_\Delta$ whose bin index is consistent with the received $\breve{x}$ to find the one closest to the estimate provided by side information y.

Since we use a uniform scalar quantizer and a uniform modulo operation, the decoding rule can be expressed as follows:

$$l^\star = \arg \min_{l \in \mathbb{Z}} |\breve{x} + lJ\Delta - \mathbb{E}[X|Y = y]| \tag{3}$$

and $\hat{x} = \psi(\breve{x}, y) = \breve{x} + l^\star J\Delta$. Intuitively $l^\star$ denotes index of the *coarse* quantizer [13] with quantization intervals of size $J\Delta$, and $\breve{x}$ provides the offset associated with x that should be used in the reconstruction.

The above scheme performs symbol-by-symbol coding, which incurs zero delay. For the case of memoryless Gaussian source and side information, the best performing symbol-by-symbol codes achieve a signal-to-quantization-noise ratio (SQNR) of approximately 10dB from the information theoretic infinite blocklength Wyner-Ziv limit [4, 8]. The difference in performance can be attributed to the decoding error in (3), which can output incorrect quantization index $l^\star$ leading to a large distortion. The advantage of nested scalar quantization is that it is much easier to implement than nested lattice quantization, and it already significantly improves upon the baseline of ignoring the side information.

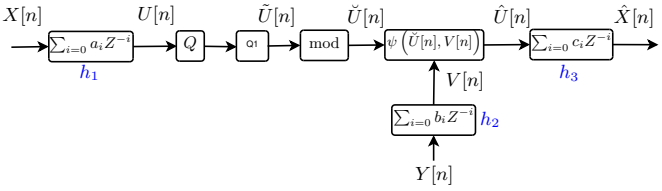

Fig. 2. Scalar Wyner-Ziv for Sources with Memory

2.2. Extension to Sources with Memory

Fig. 1 shows a natural architecture that can be used to extend the scheme in the previous section to handle case when both the source and side-information sequences can have memory [6, 8]. Following the idea of differential pulse coded modulation [9], the input source $X[n]$ is passed through a causal prediction filter in the feedback loop with coefficients $\mathbf{a} = [a_0 = 1, a_1, \ldots, a_{M_1-1}]$ at the encoder to generate the sequence $U[n]$, and an identical filter is used to map the reconstruction $\hat{U}[n]$ to the output $\hat{X}[n]$. The side information sequence is processed using a causal feedforward prediction filter with coefficients $\mathbf{b} = [b_0 = 1, \ldots, b_{M_2-1}]$ to produce a sequence $V[n]$. Uniform scalar quantization and uniform one-dimensional binning is applied to $U[n]$ to generate $\tilde{U}[n]$ and $\breve{U}[n]$ respectively, then the reconstruction function (2) is applied to generate $\hat{U}[n] = \psi(\breve{U}[n], V[n])$ at the decoder.

In general the optimal choice of filter coefficients $\mathbf{a}$ and $\mathbf{b}$, as well as the choice of the quantization interval Δ can only be found through an exhaustive search even when the statistics of the source and side information are known [8]. We note though that the complexity of this exhaustive search scales exponentially with the filter length, thus limit its applicability only to low-order prediction filters. For first-order Gauss-Markov source and side-information sequences, it is observed in [8] that the best performing scheme also exhibits approximately a gap of 10dB in SQNR from the ideal infinite blocklength Wyner-Ziv limit, suggesting that 10dB is the cost of symbol-by-symbol coding. The goal of this paper is to achieve the same performance as exhaustive search through a data-driven approach with complexity which does not necessarily scales exponentially with the filter length.

3. DATA-DRIVEN DESIGN FOR WYNER-ZIV CODING

We now introduce the proposed data-driven approach to designing a zero-delay Wyner-Ziv coding system for source and side information with memory. Fig. 2 illustrates the system used during the testing phase, while Fig. 3 illustrates the system used in the training phase for optimizing the parameters. The followings are the main differences as compared to the system in Fig. 1. These modifications are made in order to make the system amenable to training using SGD.

- *Prediction and Reconstruction Filters*: We use a feedforward, open-loop prediction filter with coefficients $\mathbf{a} = [a_0, \ldots, a_{L_1-1}]$ for the input source sequence instead of the feedback system in Fig. 1, as such a system is easier to train. In experiments we observe that when the length of the filter L_1 is much greater than the length of the feedback filter M_1, the performance loss was negligible. We also do not force the reconstruction filter to be identical to the source filter, but use a different set of coefficients $\mathbf{c} = [c_0, \ldots, c_{L_3-1}]$ so that $\hat{X}[n] = \sum_{i=0}^{i=L_3-1} c_i \hat{U}[n - i]$. Finally we use a feed-forward filter $\mathbf{b} = [b_0, \ldots, b_{L_2-1}]$ to process the side-information sequence.

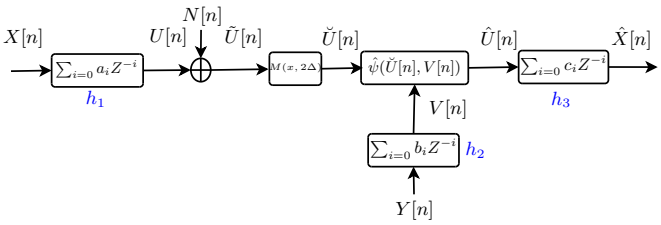

Fig. 3. Training Architecture

- *Decoding Function*: The decoding function $\psi(\cdot,\cdot)$ as defined in (2) requires a conditional expectation which is difficult to compute. We simplify (3) by replacing the conditional expectation with just the value of the associated side information:

$$l^\star = \arg\min_{l\in\mathbb{Z}} |\breve{U}[n] + lJ\Delta - V[n]| \tag{4}$$

then compute $\hat{U}[n] = \psi(\breve{U}[n], V[n]) = \breve{U}[n] + l^\star J\Delta$. This is justified because when optimizing over the filter coefficients, the proposed simplification would automatically force $V[n] \approx \mathbb{E}[U[n]|V[n]]$ in order to minimize the reconstruction error.
- *Soft Minimization* The minimization (4) in the decoding function is not differentiable thus not amenable to data-driven training. We replace it with the following approximation:

$$\hat{l}^\star = \sum_{k=-\lceil K/2\rceil}^{\lceil K/2\rceil} k\sigma_k \tag{5}$$

where $\sigma_k = \frac{e^{-\beta v_k}}{\sum_j e^{-\beta v_j}}$, and $v_k = (\breve{U}[n] + kJ\Delta - V[n])^2$. Note that $\hat{U}[n] = \hat{\psi}(\breve{U}[n], V[n]) = \breve{U}[n] + \hat{l}^\star J\Delta$. Here $\frac{1}{\beta}$ is the temperature parameter [14] and the parameter K is selected to limit the summation range during numerical computation. The intuition behind the *softmin* operation in (5) is that $\hat{l}^\star$ is a weighted sum of the indices between $-\lceil K/2\rceil$ and $\lceil K/2\rceil$, and with sufficiently large β, we have that $\sigma_k \approx 1$ for the value of k that minimizes v_k and $\sigma_k \approx 0$ for all other values of k.
- *Quantization* The quantization step $\tilde{U}[n] = Q(U[n])$ is also not differentiable. To make it amendable to data-driving training, we model it as an additive test channel: $\tilde{U}[n] = U[n] + N[n]$, where following [6], $N[n]$ is Gaussian[1] with variance $\sigma_N^2 = \frac{\Delta^2}{3}2^{-2R}$. The choice of noise variance matches the quantization error for uniform quantization [15] as in [16, 17].
- *Approximation of Modulo Function*: We use the following approximation for the modulo operation [18], for which differentiation with respect to the modulo size is implemented in the software library and found to perform reasonably well in the range of numerical values in the experiments:

$$M(x,\alpha) \approx \frac{\alpha}{\pi}\tan^{-1}\left(\tan\left(\pi\left(\frac{x}{\alpha}-\frac{1}{2}\right)\right)\right). \tag{6}$$

As shown in Fig. 4, the function $M(x,\alpha)$ takes in any $x \in \mathbb{R}$ as input and folds it into the interval $[0,\alpha]$ by performing the continuous version of the modulo operation i.e.,

$$M(x,\alpha) \approx \{x + n\alpha, n \in \mathbb{Z}\} \cap [0,\alpha). \tag{7}$$

[1]The uniform noise model is also a suitable choice here.

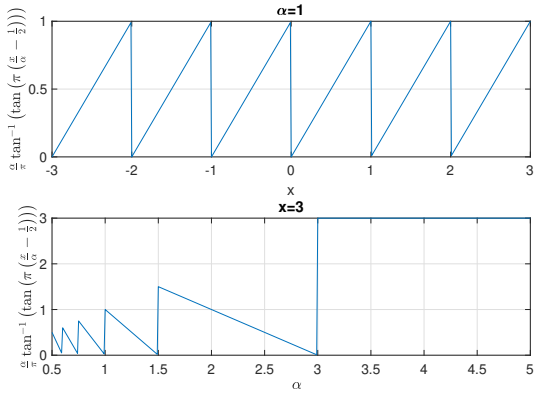

Fig. 4. The function $M(x,\alpha)$ in (6) plotted for fixed range $\alpha = 1$ and varying input, and for fixed input $x = 3$ and varying range α.

Thus to realize the modulo operation in Section 2.1 that maps $\tilde{x} = k\Delta$ to $\breve{x} = (k \mod J)\Delta$, it suffices to select $\alpha = J\Delta$. In practice, when optimizing Δ or equivalently α under the constraint $\alpha \geq 0$, we define $\alpha = e^{\tilde{\alpha}}$ and optimize over $\tilde{\alpha}$.

The proposed system performs data-driven optimization of the prediction filters coefficients and the modulo range for the following objective function:

$$\mathcal{L}(\mathbf{a},\mathbf{b},\mathbf{c},\Delta) = \mathbb{E}^{\text{emp}}\left[\|X[n] - \hat{X}[n]\|^2\right], \tag{8}$$

where $\mathbb{E}^{\text{emp}}[\cdot]$ denotes the empirical distribution with respect to the training samples. We optimize over the variables of interest using SGD [19]. In particular, at each iteration t we select a vector of sample inputs at random, compute the associated loss $\mathcal{L}^t(\mathbf{a}_t,\mathbf{b}_t,\mathbf{c}_t,\Delta_t)$ and update the parameters by computing the gradient with respect to the loss function. For example, the parameters of the filter with coefficients $\mathbf{a}$ are updated as $\mathbf{a}_{t+1} \leftarrow \mathbf{a}_t - \epsilon\nabla_{\mathbf{a}_t}\mathcal{L}^t(\mathbf{a}_t,\mathbf{b}_t,\mathbf{c}_t,\Delta_t)$, where ϵ denotes the learning rate and $\nabla_{\mathbf{a}}$ denotes the partial gradient with respect to the vector $\mathbf{a}$. We note though, that SGD might converge to a local minimum.

4. EXPERIMENTAL RESULTS

The proposed approach is implemented using the automatic differentiation of PyTorch [20]. We used Adam optimizer [21] with learning rate of $\epsilon = 0.01$. Throughout these experiments we performed optimization with 5-tap filters ($L_1 = L_2 = L_3 = 5$), with $K = 7$ (see (5)) and temperature value of $\beta = 20$. In general, $\{L_i\}_{i=1}^3$, K and β should be viewed as tunable hyperparameters. The filters are initialized as a discrete-time delta function and the initial value of Δ is randomly chosen. To prevent overfitting, we run over epochs built from 10 different realizations of source and side information, each serving as input to the optimization sequentially. Finally, the testing phase is performed on a different pair of source and side information (which are not part of the training). Rather than plotting the distortion we plot the SQNR (in dB) which is defined as $10\log_{10}(\frac{\sigma^2}{D})$, where σ^2 denotes the variance of the input samples $X[n]$.

4.1. First Order Gauss-Markov Source and Side Information

Let $T[n]$ be a first-order Gauss-Markov process, $T[n] = \rho T(n-1) + W[n]$ where $W[n]$ is i.i.d. zero-mean Gaussian and $\sigma_W^2 =$

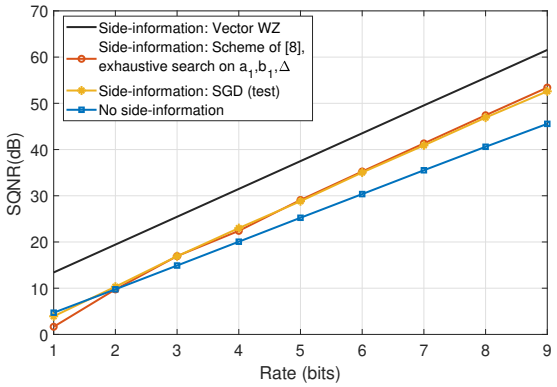

Fig. 5. Comparison of adaptive optimization and exhaustive search for first-order Gauss-Markov source and side information with $\rho = 0.7$, $\sigma_U = 0.1$, $\sigma_V = 0.1$.

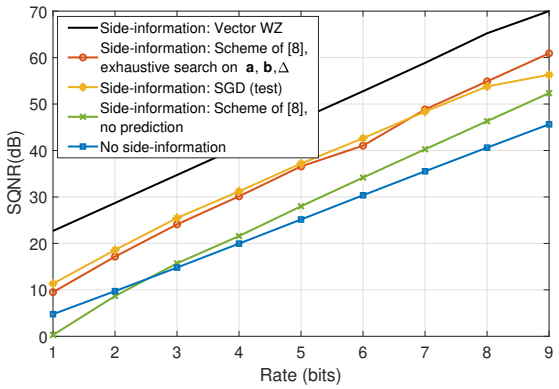

Fig. 6. Adaptive optimization vs. exhaustive search for first-order Gauss-Markov source and second-order side information sequence.

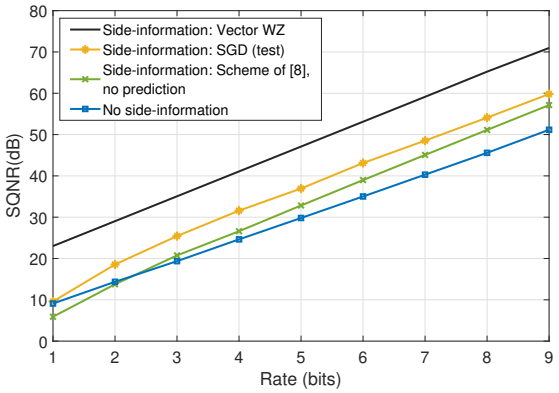

Fig. 7. Adaptive optimization vs. exhaustive search for third-order Gauss-Markov source and second-order side information sequence.

$1-\rho^2$, hence $\sigma_T^2 = 1$. We assume $T(1)$ is zero mean and has unit variance, so that the process is stationary. The source $X[n]$ and side information $Y[n]$ are defined as $X[n] = T[n] + N_x[n]$ and $Y[n] = T[n] + N_y[n]$, where $N_x[n]$ and $N_y[n]$ are i.i.d. zero-mean Gaussians independent of each other.

Prediction filters of order one are assumed in the benchmark scheme. The filter coefficients are found via exhaustive search while assuming $-1 < a_1, b_1 < 1$, along with an exhaustive search over the modulo range as in [8]. Note that exhaustive search for higher-order filters becomes increasingly difficult, because its complexity is exponential in the order of the filters. We have verified consistency with the values reported in [8]. In Fig. 5, we show the results of exhaustive search and also illustrate the infinite-blocklength information theoretic limit by a solid black line.

We note that the output of the data-driven approach is very close to exhaustive search. Both have about 10dB gap with respect to the ideal Wyner-Ziv limit which is described in Theorem 1 of [8] for this particular source and side-information pair. We also plot the performance of ignoring the side information (thus not performing binning but rather optimize the tradeoff between granular and overload distortion). It can be seen that except at low rates, taking side information into account provides better performance.

4.2. Higher Order Gauss-Markov Source and Side Information

Next, we consider an M-th order Gauss-Markov source: $T_M[n] = \sum_{l=1}^{M} \rho_l T(n-l) + W[n]$, where $W[n]$ is i.i.d. zero-mean Gaussian, and $T(1)$ is zero mean and has unit variance. By varying M at the source and side information we can have different order Gauss-Markov processes, i.e., $X[n] = T_{M_1}[n] + N_x[n]$ and $Y[n] = T_{M_2}[n] + N_y[n]$, where $N_x[n]$ and $N_y[n]$ are i.i.d. zero-mean Gaussians independent of each other and $T_{M_1}[n] = \sum_{l=1}^{M_1} \rho_{x,l} T(n-l) + W[n]$ and $T_{M_2}[n] = \sum_{l=1}^{M_2} \rho_{y,l} T(n-l) + W[n]$.

In Fig. 6 we consider the case of $M_1 = 1$, $M_2 = 2$, $\rho_{x,1} = 0.51$, $\rho_{y,1} = 0.05$, $\rho_{y,2} = 0.5$, $\sigma_W^2 = 0.74$ and $\sigma_U^2 = \sigma_V^2 = 0.01$. To perform exhaustive search, we use a coarse grid for the prediction filters as in [8] over $-1 < a_1, b_1, b_2 < 1$,$2.5 < \Delta < 6$. As can be seen in Fig. 6, the SGD performs close to exhaustive search despite having much lower complexity. As reference, we show the distortion when ignoring side information and with zero prediction filter.

Finally, in Fig. 7 we consider the case where $M_1 = 3$, $M_2 = 2$, $\rho_{x,1} = 0.05$, $\rho_{x,2} = 0.5$, $\rho_{x,3} = 0.25$, $\rho_{y,1} = 0.3$, $\rho_{y,2} = 0.6$, and $\sigma_W^2 = 0.8485$. Further, $\sigma_U^2 = \sigma_V^2 = 0.01$. We do not present outcome of the grid search since its complexity is too high. As reference, we show the distortion when ignoring side information and with zero prediction filter. The SGD based approach improves over all other baseline schemes. Finally, we compare the performance in Fig. 6 and Fig. 7 with the ideal Wyner-Ziv limit, which we conjecture is achievable at infinite blocklength by re-deriving Theorem 1 of [8] for the higher-order Gauss-Markov source and side information, and confirm the 10dB gap due to the zero-delay scalar quantization.

5. CONCLUSION

This paper advocates a data-driven approach to optimizing the encoding and decoding schemes for zero-delay Wyner-Ziv coding for source and side information with memory. Whereas exhaustive numerical search is required in prior work, this paper shows that a lower-complexity data-driven approach which does not require source and side information statistics is feasible. Towards this end, we propose modifications in filter design, the quantization process, and the reconstruction function to make the overall model amenable to training using SGD. Numerical results demonstrate negligible loss as compared to the exhaustive search based approach.

6. REFERENCES

[1] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, 1976.

[2] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1250–1276, 2002.

[3] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[4] Z. Liu, S. Cheng, A. Liveris, and Z. Xiong, "Slepian-wolf coded nested lattice quantization for wyner-ziv coding: High-rate performance analysis and code design," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4358–4379, 2006.

[5] R. Zamir, *Lattice Coding for Signals and Networks: A Structured Coding Approach to Quantization, Modulation, and Multiuser Information Theory*. Cambridge University Press, 2014.

[6] R. Zamir, Y. Kochman, and U. Erez, "Achieving the gaussian rate–distortion function by prediction," *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 3354–3364, 2008.

[7] E. Tuncel, "Predictive coding of correlated sources," in *Information Theory Workshop*, 2004, pp. 111–116.

[8] X. Chen and E. Tuncel, "Low-delay prediction-and transform-based wyner–ziv coding," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 653–666, 2010.

[9] N. S. Jayant and P. Noll, "Digital coding of waveforms: principles and applications to speech and video," *Englewood Cliffs, NJ*, pp. 115–251, 1984.

[10] M. Fleming, Q. Zhao, and M. Effros, "Network vector quantization," *IEEE Transactions on Information Theory*, vol. 50, no. 8, pp. 1584–1604, 2004.

[11] A. Saxena and K. Rose, "Distributed predictive coding for spatio-temporally correlated sources," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 4066–4075, 2009.

[12] S.-Y. Tung, "Multiterminal source coding," Ph.D. dissertation, Cornell University, 1978.

[13] R. Zamir, "Lattices are everywhere," in *Information Theory and Applications Workshop*, 2009, pp. 392–421.

[14] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[15] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Springer Science & Business Media, 2012, vol. 159.

[16] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *International Conference on Learning Representations (ICLR)*, 2017. [Online]. Available: https://arxiv.org/abs/1611.01704

[17] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://openreview.net/forum?id=rkcQFMZRb

[18] "Does there exist a smooth approximation of x mod y?" Mathematics Stack Exchange, November 25, 2019. [Online]. Available: https://math.stackexchange.com/q/2491494

[19] L. Bottou, *Stochastic learning*. Springer, Berlin, Heidelberg, 2003.

[20] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980