MMSE Scaling Enhances Performance in Practical Lattice Codes

Nuwan S. Ferdinand *, Matthew Nokleby †, Brian M. Kurkoski ‡ and Behnaam Aazhang §
nuferdin@ee.oulu.fi *, matthew.nokleby@duke.edu †, kurkoski@jaist.ac.jp‡, aaz@rice.edu§

***Abstract*—We investigate the value of MMSE scaling for practical lattice codes. For ideal lattices, MMSE scaling has been shown to be a key ingredient in achieving the capacity of the AWGN channel. We demonstrate that MMSE scaling enhances the performance, particularly at low SNR, for practical lattice codes. For example, a dimension $n = 10000$ LDLC lattice exhibits approximately 0.6 dB gain when MMSE scaling is used for a rate of 1 bit/dimension. Furthermore, we provide a novel derivation of the MMSE scaling rule, showing that it emerges naturally from principles of belief propagation decoders which account for the transmit power constraint.**

I. Introduction

In the search for structured codes for continuous channels, researchers have shown that several constructions of *lattice codes* achieve the capacity of the point-to-point AWGN channel. First, it was shown that an appropriate sequence of lattices, intersected with a "thin" shell corresponding to the transmit power constraint, form a capacity-achieving codebook [1]. Then, it was shown that, under a near maximum-likelihood decoding rule, lattices intersected with a hyperspherical shaping region achieve capacity [2]. In each case, the decoding rule is rather complex and eliminates most of the advantages promised by structured codes.

It was then shown that, under a (relatively) low-complexity *lattice decoding* rule, in which codewords are mapped to the nearest lattice point without regard for shaping, rates below $1/2\log_2(\mathrm{SNR})$ are achieved; however, this falls well short of capacity when the SNR is small [3]. Finally, it was shown that lattice decoding can achieve the full capacity when the receiver *scales* the incoming signal prior to decoding. In particular, the receiver scales the signal by the coefficient corresponding to optimal MMSE estimation.

Since the advent of capacity-achieving lattice codes, researchers have searched for lattice codes suitable for practical implementation [4]–[6]. At present, such codes employ simple lattice decoding without regard for the shaping region, which [3] predicts has poor performance at low SNR. Furthermore, it is unclear whether the MMSE scaling proposed in [7] improves performance in practical codes, or if it is only a proof technique.

In this paper, we show that MMSE scaling is beneficial for practical, finite-dimensional lattices. We focus on *low-density lattice codes* or LDLCs, for which efficient power-shaping algorithms, as well as low-complexity lattice decoders, exist. We show experimentally that the use of MMSE scaling improves performance for lattice codes of moderate block length, particularly for low SNR and low rates. Indeed, scaling offers improvements of up to 0.6dB over standard decoders. Finally, we provide an alternate derivation of the MMSE scaling rule from a belief propagation perspective. We show that if we modify the iterative lattice decoder to account for the transmit power constraint, the result is equivalent to applying MMSE scaling at the receiver. That is, MMSE scaling emerges naturally from the design of iterative decoders over power-constrained channels.

II. Preliminaries

A. System Model

We consider a standard additive white Gaussian noise (AWGN) channel in which the transmitter sends a signal $\mathbf{x} \in \mathbb{R}^n$, constrained to have average transmit power $\|\mathbf{x}\|^2 \leqslant nP_x$. the receiver obtains the transmitted signal, corrupted by noise:

$$\mathbf{y} = \mathbf{x} + \mathbf{z} \tag{1}$$

where $\mathbf{z}$ is AWGN with per-element variance P_z.

Each transmit signal $\mathbf{x}$ is a codeword from a finite codebook $\mathcal{C} \subset \mathbb{R}^n$. Define the rate of the codebook as

$$R = \frac{\log_2 |\mathcal{C}|}{n}, \tag{2}$$

measured in bits per channel use. The receiver uses a decoder $D : \mathbb{R}^n \to \mathcal{C}$ that maps received signals $\mathbf{y}$ to estimates $\hat{\mathbf{x}}$ of the transmitted codeword. We leave the decoder abstract for now; later on we consider several specific examples. Define the *average symbol error probability* as

$$P_e = \frac{1}{|\mathcal{C}|} \sum_{\mathbf{x} \in \mathcal{C}} \Pr(\hat{\mathbf{x}} \neq \mathbf{x} | \mathbf{x}), \tag{3}$$

where each term in (3) is the probability of a decoding error supposing that the codeword $\mathbf{x}$ is sent.

B. Lattice Codes

We focus on codebooks constructed from lattices. An n-dimensional lattice Λ is a discrete additive subgroup of $\mathbb{R}^n$. Any lattice Λ may be expressed in terms of a (non-unique) generator matrix $\mathbf{G} \in \mathbb{R}^{n \times n}$ by the following definition:

$$\Lambda = \{\lambda = \mathbf{G}\mathbf{b} \ : \mathbf{b} \in \mathbb{Z}^n\}. \tag{4}$$

We introduce a few basic definitions used throughout the paper; for further details about lattice codes we refer the reader to [8]. The *lattice quantizer* $\mathcal{Q}_\Lambda$ maps any point in $\mathbb{R}^n$ to the nearest lattice point in Λ in the Euclidean sense, with ties

broken arbitrarily:

$$\mathcal{Q}_\Lambda(x) = \arg\min_{\lambda\in\Lambda} \|\lambda - x\|^2. \tag{5}$$

The *fundamental Voronoi region* of Λ is the region of $\mathbb{R}^n$ closer to the origin than to any other lattice point:

$$\mathcal{V} = \{\mathbf{x} : \mathcal{Q}_\Lambda(x) = \mathbf{0}\} \tag{6}$$

The $\bmod$ operation with respect to Λ returns the quantization error:

$$[\mathbf{x}] \bmod \Lambda = \mathbf{x} - \mathcal{Q}_\Lambda(\mathbf{x}). \tag{7}$$

The $\bmod$ operation partitions $\mathbb{R}^n$ into Voronoi regions associated with each $\lambda \in \Lambda$, or the subsets of $\mathbb{R}^n$ closest to each λ.

Next, $V = \mathrm{Vol}(\mathcal{V})$ denotes the volume of the fundamental Voronoi region. The *second moment* of a lattice, defined as

$$\sigma_\Lambda^2 = \frac{1}{\mathrm{Vol}(\mathcal{V})}.\frac{1}{n}\int_{\mathcal{V}} \|\mathbf{x}\|^2 \, d\mathbf{x}, \tag{8}$$

specifies the average power of a random variable uniformly distributed across $\mathcal{V}$. The *normalized second moment* (NSM) of Λ is given by

$$G(\Lambda) = \frac{\sigma_\Lambda}{V^{\frac{2}{n}}}. \tag{9}$$

For any lattice, $\frac{1}{2\pi e} \leqslant G(\Lambda) \leqslant \frac{1}{12}$, where the minimum value corresponds to a hyperspherical Voronoi region, and the maximum corresponds to a hypercube. Poltyrev [9] showed that there exists a sequence of lattices such that $G(\Lambda) \to \frac{1}{2\pi e}$ as $n \to \infty$.

Suppose codewords are lattice points in Λ and the receiver uses $\mathcal{Q}_\Lambda(\mathbf{y})$ as the decoder. Then, an error occurs precisely when the noise escapes the fundamental Voronoi region, and the symbol error probability is

$$P_e = Pr(\mathbf{z} \notin \mathcal{V}) \tag{10}$$

Then, we define the *volume to noise ratio* (VNR) as

$$\mu(\Lambda, P_e) = \frac{V^{\frac{2}{n}}}{P_z}, \tag{11}$$

where P_z is the noise variance such that the probability of error is P_e. Poltyrev [9] showed also that there exists a sequence of lattices such that $\mu(\Lambda, P_e) \to 2\pi e$, which is the maximum theoretical value.

C. Low-Density Lattice Codes

A *low-density lattice* $\Lambda \subset \mathbb{R}^n$ is defined by a generator matrix $\mathbf{G} \in \mathbb{R}^{n\times n}$ such that the inverse $\mathbf{G}^{-1} = \mathbf{H}$ exists and is sparse [4]. In order to construct a finite-rate codebook, we choose a finite subset of the low-density lattice Λ, defined by a subset of $\mathbb{Z}^n$ to which we apply the generator matrix $\mathbf{G}$. Define the set

$$\mathcal{L} = \{\mathbf{b} \in \mathbb{Z}^n : 0 \leqslant b_i \leqslant L_i - 1\}, \tag{12}$$

where each L_i is a positive integer. Then we define the codebook

$$\mathcal{C} = \{\mathbf{Gb} : \mathbf{b} \in \mathcal{L}\} \subset \Lambda. \tag{13}$$

Observing that $|\mathcal{L}| = \prod_{i=1}^n L_i$, the rate of this code is

$$R = \frac{\sum_{i=1}^n \log_2(L_i)}{n}. \tag{14}$$

In general, the codewords in $\mathcal{C}$ do not obey the power constraint, so we must employ shaping methods in order to create a suitable codebook. In particular, we employ shaping methods developed in [6] for LDLCs. They require LDLCs with a lower-triangular parity-check matrix $\mathbf{H}$, which enables fast computation and manipulation of lattice points. As in [7], the shaping algorithms are based on nested lattices. In *hypercube shaping*, the shaping lattice Λ_1 is a scaled integer lattice; in *nested lattice shaping* Λ_1 is an integer multiple of the coding lattice Λ. In each case, the codewords are $\mathbf{x} = [\mathbf{c}] \bmod \Lambda_1$ for lattice points $\mathbf{c} = \mathcal{C}$. See [6] for details.

Low-density lattice codes can be decoded with a linear-complexity, iterative algorithm similar to those for low-density parity check codes [4], [5], [10]. Here we briefly sketch the algorithm from [5], which is based on a Gaussian-mixture approximation.

The message-passing algorithm has two types of nodes. Let V_k be the kth *variable node*, which is ensures that the estimated codeword matches the kth element of the received signal, and let C_l be the lth *check node*, which ensures that the lth element of $\mathbf{Hx}$ is an integer. Variable and check nodes iteratively exchange "messages", which are density functions of the codeword from each node's vantage point. These messages are computed as follows.

- Let $y_k = x_k + z_k$ be the kth element of $\mathbf{y}$. With knowledge only of y_k, the density of x_k is $\mathcal{N}(y_k, P_z)$. The variable node V_k sends this density to every connected check node, or every node C_l such that $H_{k,m} \neq 0$.
- Each check node C_l enforces the integer constraint. Given the incoming densities from each connected variable node and the integer constraint, the density of x_k is the convolution of the incoming densities, scaled down by $H_{k,l}$, and periodically extended by the integers. For tractability, the scaled convolution is approximated by a Gaussian density, and only a few integer replicas are computed, resulting in a Gaussian mixture. For each connected V_k, the approximate density is computed and sent.
- Returning to the variable node V_k, the density of x_k given y_k and the incoming densities is the product of the incoming densities and the Gaussian $\mathcal{N}(y_k, P_z)$. The variable node approximates the product by a Gaussian distribution, and sends the result to each connected check node.
- After sufficiently many iterations, the variable nodes V_k take the product of the incoming densities, and find the peak values x_k^* of the result. Finally, the decoder

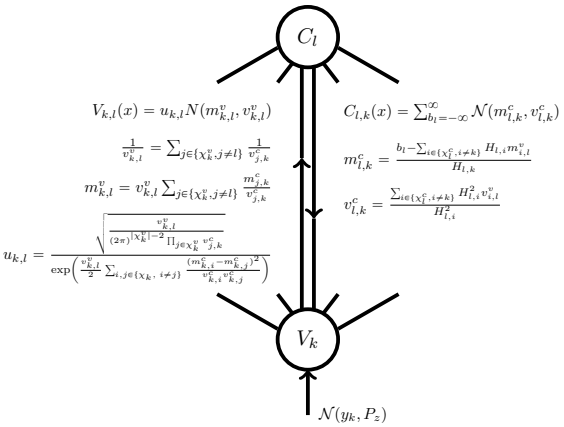

Fig. 1. LDLC Gaussian approximation algorithm.

estimates the information vector $\mathbf{b}$ by choosing

$$\mathbf{b}^* = \lfloor \mathbf{H}\mathbf{x}^* \rceil, \tag{15}$$

where $\lfloor \cdot \rceil$ denotes element-wise rounding to the nearest integer.

Fig. 1 illustrates the Gaussian mixture LDLC algorithm where it shows the operation of m^{th} check node and k^{th} variable node. We refer the reader to [4], [5], [10] for details.

We emphasize that this decoder does not consider the shaping region in choosing the optimum codeword, which increases the probability of error. As we show in the next section, the use of scaling can greatly improve error performance, particularly at low SNR.

III. Scaling for LDLCs

A. Scaling Algorithm

The difficulty with lattice decoding is that it is easy for the receiver to erroneously decode points near the boundary of the shaping region to lattice points that are not in the codebook. As the dimension increases, a rather large fraction of the codewords are near the boundary, and the error performance degrades. Scaling obviates this problem by "inflating" the equivalent lattice so that codewords are further from the boundary. However this advantage comes at the cost of self-noise, as we describe in the nest subsection.

Scaling itself is a simple process.The received signal $\mathbf{y}$, as given in (1), is multiplied by α at the receiver:

$$\begin{aligned} \mathbf{y}' = \alpha\mathbf{y} &= \alpha(\mathbf{x} + \mathbf{z}) \\ &= \mathbf{x} + \mathbf{z}' \end{aligned} \tag{16}$$

where the effective noise $\mathbf{z}' = (\alpha - 1)\mathbf{x} + \alpha\mathbf{z}$. Then it uses the LDLC decoder with input probability distribution $\mathcal{N}(\alpha\mathbf{y}, (\alpha - 1)^2 P_x + \alpha^2 P_z)$.

B. Analytical Evaluation

The achievable rate for optimal nested lattice codes with large block length is known and equal to AWGN capacity [7]. In this section we analyze the achievable rates of practical lattice codes, both with and without MMSE scaling. Let generator matrices $\mathbf{G}$ and $\mathcal{L}\mathbf{G}$ correspond to the fine lattice Λ_1 and coarse lattice Λ_2. Let $\mathcal{V}_1$ and $\mathcal{V}_2$ be the Voronoi regions of Λ_1 and Λ_2 respectively. We first obtain the second moment of effective noise $\mathbf{z}'$ as given in (16)

$$\begin{aligned} \sigma^2_{z'} &= \frac{1}{n} E[\|(1-\alpha)\mathbf{x} + \alpha\mathbf{z}\|^2] \\ &= (1-\alpha)^2 P_x + \alpha^2 P_z \end{aligned} \tag{17}$$

Simple optimization shows that the average power of the effective noise is minimized when α equals the MMSE coefficient

$$\alpha = \alpha_m = \frac{P_x}{P_x + P_z} \tag{18}$$

For this α, the corresponding effective noise power is $\sigma^2_{z'} = P_x P_z/(P_x + P_z)$. If the receiver uses an unconstrained lattice decoder on $\mathbf{y}'$, the error probability is

$$P_e = Pr[\mathbf{z}' \notin \mathcal{V}_1] \tag{19}$$

According to the definition of VNR, the volume of the fine lattice should be $V_1 = \left[\mu(\Lambda_1, P_e)\sigma^2_{\mathbf{z}'}\right]^{\frac{n}{2}}$ for a target probability of P_e. Now we find the achievable rate

$$R = \frac{1}{n} \log_2 \left(\frac{V_2}{V_1}\right) \tag{20}$$

According the the definition of normalized second moment

$$\begin{aligned} R_s &= \frac{1}{2} \log_2 \left[\frac{P_x/G(\Lambda_2)}{\mu(\Lambda_1, P_e)\sigma^2_{\mathbf{z}'}}\right] \\ &= \frac{1}{2} \log_2 \left(1 + \frac{P_x}{P_z}\right) - \frac{1}{2} \log_2 \left[G(\Lambda_2)\mu(\Lambda_1, P_e)\right] \end{aligned} \tag{21}$$

When block length $n \to \infty$, there exists a *good nested lattice* pair for *quantization* and AWGN *coding* such that $G(\Lambda_2) \to 1/2\pi e$ and $\mu(\Lambda_1, P_e) \to 2\pi e$. In that scenario $\frac{1}{2} \log_2 \left[G(\Lambda_2)\mu(\Lambda_1, P_e)\right] \to 0$; hence, lattice codes achieve the AWGN capacity [7].

However, in practice $G(\Lambda_2) > 1/2\pi e$ due suboptimal shaping, and $\mu(\Lambda_1, P_e) > 2\pi e$ due to suboptimal coding. Let β_s and β_c be the shaping and coding loss for any practical lattice scheme in general. Then we obtain second moment of Λ_2 and VNR for LDLC

$$G(\Lambda_2) = \frac{\beta_s}{2\pi e}, \tag{22}$$

$$\mu(\Lambda_1, P_e) = 2\pi e \beta_c \tag{23}$$

Then, we find the achievable rate for practical lattices, such as LDLCs, in terms of their shaping and coding loss. With MMSE scaling,

$$\begin{aligned} R_s &= \frac{1}{2} \log_2 \left[\frac{P_x/G(\Lambda_2)}{\mu(\Lambda_1, P_e)\sigma^2_{\mathbf{z}'}}\right] \\ &= \frac{1}{2} \log_2 \left(1 + \frac{P_x}{P_z}\right) - \frac{1}{2} \log_2 \left(\beta_s \beta_c\right) \end{aligned} \tag{24}$$

If we use $\alpha = 0$, that is, if we do not employ scaling, the rate

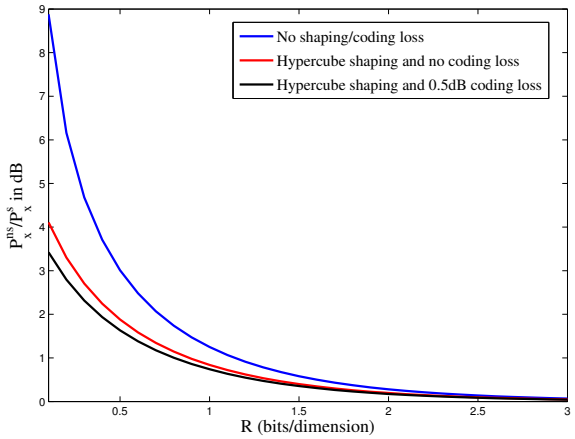

Fig. 2. Required power comparison to achieve a fix rate R with and without scaling.

is

$$R_{ns} = \frac{1}{2}\log_2\left(\frac{P_x}{P_z}\right) - \frac{1}{2}\log_2\left(\beta_s\beta_c\right) \tag{25}$$

Example: For LDLC for block length equal to 10000 and rate = 2.935bits/dimension the shaping and coding combine loss is roughly 2.8dB for hypercube shaping and 2.4dB for nested lattice shaping [6].

We obtain the transmit power requirement to achieve fixed rate of R for scaling and non-scaling cases assuming noise variance $P_z = 1$ without of loss of generality.

$$P_x^s = 2^{2R}\beta_s\beta_c - 1 \tag{26}$$

$$P_x^{ns} = 2^{2R}\beta_s\beta_c \tag{27}$$

In order to observe how scaling is effective for different rates, we have plotted P_x^{ns}/P_x^s in dB scale for different rates in Figure 2. Three cases are plotted; 1) The perfect nested lattice scenario where $\beta_s = \beta_c = 1$, 2) The hypercube shaping scenario where $\beta_s = 2\pi e/12$ (1.53dB) and no coding loss and 3) The hypercube scenario with β_c =0.5dB [1].

It is observed in Figure 2 that the scaling plays a significant impact for low rates even for the practical lattice codes with shaping and coding losses due to the addition of 1 in the achievable rate expression. One can observe that with the increase of coding and shaping losses, the power gain with the use of scaling decreases. This is due to the fact that increase of losses, increases the power requirement to achieve the fix rate R irrespective of scaling. Hence, with the increase of transmit power, the effect of scaling is reduced.

C. Lattice Scaling Using Belief Propagation

Belief propagation principles can be applied to obtain the MMSE scaling coefficient given by Erez and Zamir. While the MMSE scaling coefficient was obtained using information theoretic approaches [7], the significance here is that belief propagation principles alone can be used to arrive at the same result [11]. Recall that when the MMSE scaling coefficient $\alpha_m = P_x/(P_x+P_z)$ is used, the effective noise is $P_xP_z/(P_x+P_z)$ [7, eqn. (29)].

[1]For LDLC with block length 100000, the coding loss is 0.5dB

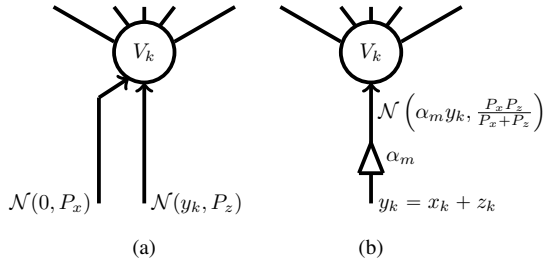

Fig. 3. (a) Modification of the "belief" at the variable node to include the transmit power constraint P_x. (b) Equivalent representation, after multiplying by Gaussians.

The belief propagation perspective is as follows. The belief propagation decoder should be modified to take into account the transmitted codebook. In the ideal case, the transmitted signal is a mean zero, variance P_x Gaussian, due to the shaping region. Under belief-propagation principles, this contribution of the shaping region should be multiplied at each variable node k, since it is information about the transmitted signal x_k. Because this message is iteration-independent, multiply the shaping region message and the channel message $\mathcal{N}(y_k, P_z)$ to obtain a new variable node input message, as in Fig. 3-(a). Since both are Gaussian, the product will also be a Gaussian (exactly):

$$\mathcal{N}(y_k', P_z') = \mathcal{N}(0, P_x) \times \mathcal{N}(y_k, P_z), \tag{28}$$

with mean y_k' and variance P_z'. The mean and variance of the new Gaussian can be found as:

$$y_k' = \frac{P_x}{P_x + P_n} y_k, \tag{29}$$

and

$$P_z' = \frac{P_xP_z}{P_x + P_z}. \tag{30}$$

Observe that the new mean y_k' is scaling by the MMSE coefficient, that is, we found the MMSE coefficient by using belief-propagation arguments alone. In addition, the corresponding new variance is the same as the effective noise found using information theoretic arguments. This is illustrated in Fig. 3-(b).

IV. Numerical Section

Fig. 4 shows the symbol error rate versus average SNR for block length $n = 100$. We have fixed the rate to 0.65 bits/dimension by having zeros for last 35 integer locations and $L_i = 2, \forall i = \{1, \ldots 65\}$. One can notice a better symbol error rate (SER) performance when scaling is used. We observe $\sim$ 1dB gain by scaling before decoding. This is encouraging since this gain is obtained without adding any complexity to existing LDLC decoding algorithm. Moreover, nested lattice

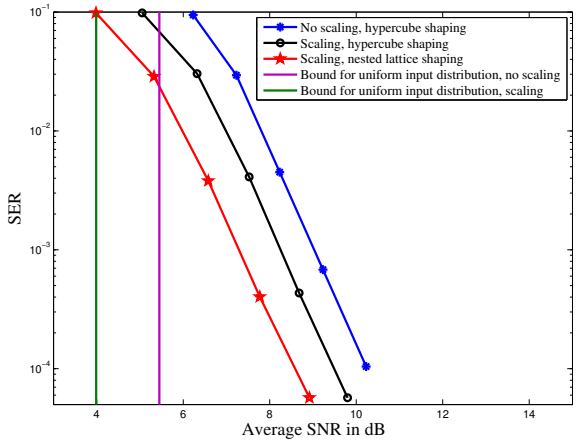

Fig. 4. Symbol error rate versus SNR for $n = 100$ with $R = 0.65$bits/dimension.

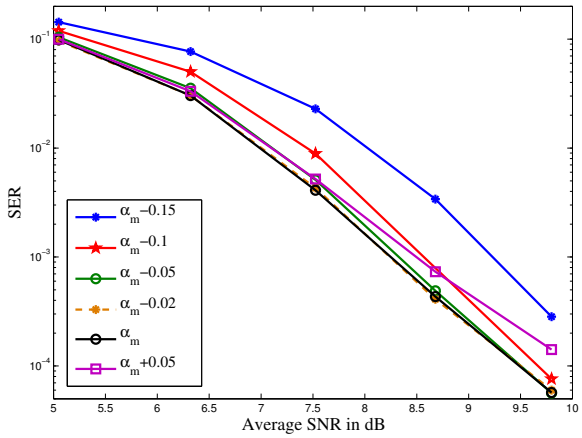

Fig. 5. Symbol error rate versus SNR for different α and $n = 100$ with $R = 0.65$bits/dimension.

shaping (with scaling) curve shows it has close to $\sim$ 2dB gain over hypercube shaping without scaling.

Fig. 4 illustrates the SER versus average SNR for different α. We have plotted for $\alpha = \alpha_m + [0 \ +0.05 \ -0.02 \ -0.05 \ -0.1 \ -0.15]$. Since we have not used dither, effective noise is correlated with signal [7]. Due to this correlation, the optimal α is not equal to α_m for small dimension as shown in [7], however, for large dimensions $\alpha \to \alpha_m$. It is observed from the figure that α_m outperforms other cases and as expected the SER increases as α deviates from α_m. Hence, we conclude that the optimal $\alpha = \alpha_m$ for dimension as small as $n = 100$.

The symbol error rate versus average SNR for $n = 10000$ is plotted in Fig. 6. The degree of rows of $\mathbf{H}$ and constellation sizes are selected as in [6] which correspond to $R = 1$ bits/dimension and used nested lattice shaping. We observe ~ 0.6 dB performance gain with the use of scaling. Further, it is only $\sim$ 0.7dB away from the bound of uniform input distribution without scaling. We have to note that due the use

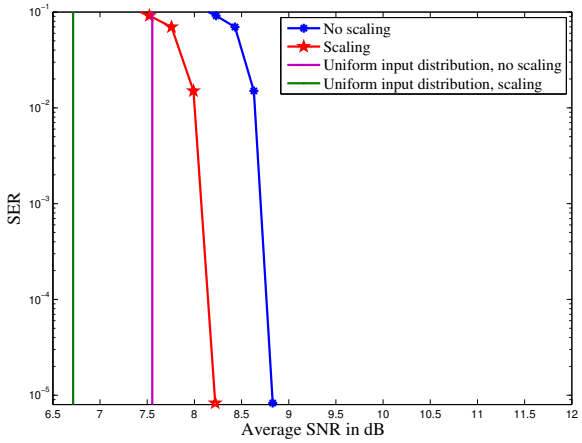

Fig. 6. Symbol error rate versus SNR for $n = 10000$ with $R = 1$bits/dimension.

of coarse constellation for the unprotected integers, there is a power loss of ~ 0.8dB and this loss can be further reduce by selecting better constellation sizes. We conclude that with the use of scaling, the LDLC even perform well at low rates.

V. Conclusion

We have studied the value of MMSE scaling for low density lattice codes, showing significant performance improvement at low SNR and low rates. Further, we have shown that, belief propagation principles alone can be applied to obtain the MMSE scaling coefficient by considering the shaping region. MMSE scaling has potential to substantially improve the performance of practical lattice schemes.

References

[1] R. de Buda, "Some optimal codes have structure," *IEEE J. Select Areas Commun.*, vol. 7, no. 6, pp. 893–899, Aug. 1989.

[2] R. Urbanke and B. Rimoldi, "Lattice codes can achieve capacity on the AWGN channel," *IEEE Trans. Inform. Theory*, vol. 44, no. 1, pp. 273–278, Jan. 1998.

[3] H.-A. Loeliger, "Averaging bounds for lattices and linear codes," vol. 43, no. 6, pp. 1767–1773, Nov. 1997.

[4] N. Sommer, M. Feder, and O. Shalvi, "Low-density lattice codes," *IEEE Trans. Info. Theory*, vol. 54, no. 4, pp. 1561–1585, July 2008.

[5] B. Kurkoski and J. Dauwels, "Message-passing decoding of lattices using gaussian mixtures," in *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, 2008, pp. 2489–2493.

[6] N. Sommer, M. Feder, and O. Shalvi, "Shaping methods for low-density lattice codes," in *Information Theory Workshop, 2009. ITW 2009. IEEE*, 2009, pp. 238–242.

[7] U. Erez and R. Zamir, "Achieving $\frac{1}{2}\log(1 + \mathrm{SNR})$ on the AWGN channel with lattice encoding and decoding," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2293–2314, Oct. 2004.

[8] R. Zamir, "Lattices are everywhere," in *Information Theory and Applications Workshop, 2009*, San Diego, CA, Feb. 2009, pp. 392–421.

[9] G. Poltyrev, "On coding without restrictions for the AWGN channel," *IEEE Trans. Inform. Theory*, vol. 40, no. 2, pp. 409–417, Mar. 1994.

[10] Y. Yona and M. Feder, "Efficient parametric decoder of low density lattice codes," in *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, 2009, pp. 744–748.

[11] H.-A. Loeliger, Personal communication, 2008.