USING FLEXIBLE TURBO-LIKE CODES FOR SIMPLIFIED WATER-FILLING

Keith M. Chugg
Communication Sciences Institute
Electrical Engineering Dept.
Viterbi School of Engineering
University of Southern California

Paul Gray
TrellisWare Technologies
16516 Via Esperillo, Suite 300
San Diego, CA 92127

ABSTRACT

A method for approaching the water-filling (WF) capacity of parallel Gaussian channels is presented. This method relies on the existence of modern turbo-like codes (TLCs) with a high degree of rate flexibility. The algorithm is a constant power WF algorithm with the active sub-channels and associated modulations selected based on the performance limits of the available TLC and modulation formats. A single TLC block is used over all sub-channels with rate imposed by the modulation assignments and the total capacity limits of the channel. Simulation results demonstrate that this approach can yield performance within approximately 3-5 dB of the ideal WF capacity for Rayleigh channel gains.

1. INTRODUCTION

A number of modern communication links can be modeled as a set of parallel additive white Gaussian noise (AWGN) channels with the signal-to-noise ratio (SNR) of each being known at the transmitter and the receiver. These include orthogonal frequency division multiplexing (OFDM) systems, closed-loop antenna array systems (multiple-input, multiple output (MIMO)) systems, and MIMO-OFDM systems (e.g., [1, 2]).

Water-filling (WF) is the well-known solution for achieving the Shannon capacity for such channels [3]. Given a fixed amount of transmit power, the optimal strategy allocates more power and rate to those sub-channels with lower inverse SNR (ISNR) – i.e., to the high SNR sub-channels. Given the power allocation, capacity is obtained by using a capacity achieving code and Gaussian signals on each sub-channel.

This ideal WF is a conceptual solution, however, since practical systems are constrained to use finite block length encoding schemes and a limited number of non-Gaussian signal constellations. Many approaches to this problem have been considered with practical constraints in mind. These approaches are based on “classical” error correcting codes which operate far from the capacity limits. Nonetheless, power allocation according to WF, or a suitable approximation, is often effective. Given a power allocation, bit loading is typically performed by including a margin associated with the reduced coding gain. Various other strategies have been suggested in the literature and utilized in practice; some based on incremental power allocation and bit loading [4], and others based on minimizing the resulting bit error probability [5, 6]. Constant power WF (CP-WF) algorithms exploit the fact that most of the WF benefits can be obtained by simply determining which sub-channels not to use and equally allocating power across the active sub-channels. A CP-WF algorithm is described in [1] with a compatible bit-allocation algorithm in [7]. The effectiveness of CP-WF is further justified in [8].

These practical WF approaches do not result in near-capacity communications because the lack of capacity approaching coding and modulation on each sub-channel. In fact, when little or no coding gain is available it is unclear that power-allocation based on the assumption of capacity achieving codes is the best approach [9].

Turbo-like codes (TLCs), however, do provide a practical method for achieving near-optimal performance. Turbo-like codes (or modern codes) are codes that are modeled using cyclic graphs and decoded using the effective heuristic of iterative message-passing (e.g., [10, 11]); this includes parallel and serial concatenated convolutional codes, as well as low density parity check (LDPC) codes.

Several recent results regarding the “universality” of TLCs suggest the approach in this paper. First, it has been demonstrated that a single TLC over all sub-channels can provide reliable data transfer at rates near the total mutual information, regardless of the particular SNR distribution over the sub-channels [12]. Second, simple methods for accurately predicting the performance of a good TLC for finite block sizes are available [13, 14]. Third, a family of simple TLCs have been developed that have significant rate and block-size flexibility and operate near the predicted limits over a wide range of operational scenarios [15, 16, 17].

This paper suggests a simple CP-WF algorithm based on
the properties of such a universally good TLC. This involves using a single TLC over all sub-channels and selecting its rate so as to approach the total capacity of the channel. The problem and the WF solution are described in further detail in Section 2. The TLC details are contained in Section 3 and the proposed approach is described in Section 4. Simulation results are presented in Section 5.

2. MODEL AND WF SOLUTIONS

Sub-channel \( i \) provides an AWGN channel with a gain \( \sqrt{\gamma[i]} \) so that the complex-valued (i.e., in-phase, quadrature (I/Q)) observation model is

\[
z_j[i] = \sqrt{\gamma[i]} e[i] x_j[i] + w_j[i], \quad i = 0, 1, \ldots, I - 1
\]  

(1)

where \( j \) indexes time along a given sub-channel and \( i \) indexes the sub-channels. The variable \( e[i] \) is the fraction of total energy (power) assigned to sub-channel \( i \) at the transmitter and \( x_j[i] \) is the information-bearing signal sent on the \( j^{th} \) use of sub-channel \( i \). The AWGN is assumed to be independent whenever the sub-channel index or the time index differs. The sub-channel gains are assumed to be known at both the transmitter and the receiver although this is not assumed to be independent whenever the sub-channel index or the time index differs. The sub-channel gains are assumed to be invariant with respect to the time index \( j \).

The model is normalized so that \( \mathbb{E} \{ |w_j[i]|^2 \} = N_0 \), \( \mathbb{E} \{ |x_j[i]|^2 \} = 1 \), and

\[
\sum_{i=0}^{I-1} e[i] = I \mathbb{E}, \quad \epsilon[i] \geq 0 \quad \forall i
\]  

(2)

With this normalization, the received energy per symbol to noise spectral level for sub-channel \( i \) is \( \frac{E_s[i]}{N_0} = \frac{x_j[i]}{N_0} \). The value of \( \mathbb{E}/N_0 \) will be used to characterize performance, where \( \mathbb{E} \) can be viewed as the average transmitted energy over the \( I \) sub-channels and \( \mathbb{E}/N_0 \) is constrained. If many sets of channel gains \( \{ \gamma[i] \} \) are considered and \( \gamma[i] \) tends to have unit average over these, then \( \mathbb{E}/N_0 \) can be viewed as the average received \( E_s/N_0 \) value.

The problem considered is the assignment of a set of \( e[i] \) (i.e., power allocation) and a set of signaling schemes, \( x_j[i] \), (i.e., bit loading) for the sub-channels. Given a set of sub-channel gains and power allocation, the capacity for sub-carrier \( i \) is \( C_i = \log_2(1 + E_s[i]/N_0) \), which is achieved when the signals are Gaussian. Thus, the standard constrained optimization problem is

\[
\max_{\{ e[i] \}} \sum_{i=0}^{I-1} \log_2(1 + \gamma[i] e[i]/N_0); \quad \text{subject to: (2)}
\]  

(3)

The solution to (3) is the well-known WF solution, where more power is allocated to channels with larger \( \gamma[i] \). Specifically, the optimizing values of \( e[i] \) are given by

\[
\epsilon_{\text{opt}}[i] = \max(0, W - N_0/\gamma[i])
\]  

(4)

where \( W \) is the "water-level" selected to ensure that the constraint in (2) is satisfied. Given this optimal power allocation, the capacity of the channel of (1) is given by the expression in (3) with \( \epsilon[i] = \epsilon_{\text{opt}}[i] \) and where sub-channel \( i \) carries \( C_i \) bits of information.

The CP-WF algorithm in [8] is also considered. This algorithm requires that the sub-channels be sorted according to sub-channel ISNR as does the CP-WF algorithm in [1].

3. GOOD, FLEXIBLE TURBO-LIKE CODES

The proposed WF algorithm is based upon the assumption of a TLC that can operate at virtually any code rate, block size, desired error rate, and given I/Q modulation format and do so at an SNR near that of the theoretical limits for a single AWGN channel. A family of very low complexity TLLCs with these characteristics are described in [15, 18, 16] and are commercially available in high-speed hardware [17]. In this paper, the code in Fig. 1 is considered. This code has been referred to as the Flexible LDPC (F-LDPC) because it can be viewed as an LDPC code, but it exhibits fine rate and block-size flexibility typically not associated with LDPC codes [16].

The overall code is systematic with parity generated by a serial concatenation of a 2-state convolutional code and an inner convolutional parity mapper\(^1\) that may be viewed as a punctured accumulator. The outer code is a 2-state, \( r = 1/2 \) convolutional code with generator polynomials \( G_1(D) = G_2(D) = 1 + D \). The inner parity mapper is a recursive single parity check (RSPC) generator with input/output relation (i.e., \( d \) in, \( p \) out)

\[
p_k = p_{k-1} \oplus (d_{k,J} \oplus d_{k,J+1} \ldots \oplus d_{k,J+k})
\]  

(5)

where \( J \) is an integer parameter. Note that for every \( J \geq 1 \) input bits, the RSPC outputs one bit. The overall rate of the code is therefore \( r_c = J/(J + 2) \). Considering the values of \( J = 2, 4, 8, 16, 32 \), the rates \( r_c = 1/2, 2/3, 4/5, 8/9, 16/17 \) are achieved, respectively.

Further refinement in code rate can be achieved by puncturing the parity bits. In this paper, a regular puncture pattern of length 16 is considered and the fraction of parity

\(^1\)Note that this inner parity mapping is not a code since the operation is not invertible.
bits maintained from each block of 16 parity bits is \( q = 16/16, 15/16, \ldots 8/16 \). This yields an overall rate of \( r = r_c + q(1 - r_c) \) which, for the values stated, provides 45 code rates from 1/2 to 32/33.

The code is decoded using the standard rules of iterative decoding (e.g., [10, 11]). In this paper, the forward-backward algorithm (FBA) is used for soft-in, soft-out (SISO) decoding of the outer code and the inner RSPC, starting with the outer code. One iteration is an activation of the outer SISO and inner SISO.

The set of \( M \)-ary modulations considered is BPSK \((M = 2)\), QPSK \((M = 4)\), 16QAM, 64QAM, and 256QAM. For a given \( M \), the spectral efficiency \( \eta = r \log_2(M) \) bits/2D-channel use is achieved. In order to use these modulations with the binary F-LDPC code, a simple bit interleaver is used along with Gray labeling. The incoming I/Q matched-filter samples are marginalized to provide bit-level log likelihoods to the iterative F-LDPC decoder. While some gain can be had by including the SISO processor in this modulation mapping on subsequent iterations, this is not considered in this paper.

With these 45 code rates and these five modulations, 225 different values of spectral efficiency, ranging from \( \eta = 1/2 \) to \( \eta = (32/33)8 \), can be targeted. Given an input block size and a target error rate, this covers a wide range of operating \( E_s/N_0 \) values. As an example, Fig. 2 shows the performance of these configurations for an input block size of \( K = 8000 \) bits. Note that this gives a 0.25 dB resolution in SNR \((E_s/N_0)\) over the range from approximately \(-3 \) dB to \(27 \) dB.

3.1. TLC Performance Guideline

Once a particular non-Gaussian modulation has been selected for an AWGN channel the capacity is reduced. The mutual information rate is given in terms of integrals involving the probability density of the AWGN channel output \( z \), conditioned on the input signal and the marginal a-priori probabilities for the signal constellation. If these a-priori probabilities are taken to be \( 1/M \), the resulting information rate is the symmetric information rate (SIR), which can be computed very efficiently (e.g., [19]). Given a particular value of the SIR, there is a minimum value of \( E_s/N_0 \) that can achieve that rate (i.e., denoted by \((E_s/N_0)_{\min,SIR}\)).

To account for finite block size and non-zero error rate, it has been demonstrated that the value of \((E_s/N_0)_{\min,SIR}\) should be adjusted by [13, 14]

\[
\left( \frac{E_s}{N_0} \right)_{\min,dB} = \left( \frac{E_s}{N_0} \right)_{\min,SIR,dB} + \Delta dB
\]

\[
\Delta dB = \sqrt{\frac{20 \log_2(2^n + 1)[10 \log_{10}(1/P_{CW})]}{K \ln(10) (2^n - 1)}}
\]

where \( K \) is the input block size, \( \eta \) is the operating spectral efficiency in bits per I/Q channel use, and \( P_{CW} \) is the operational codeword error rate (block error rate (BLER)). A survey of the literature reveals that the best TLCs designed for a single operational scenario operate within 0.5 dB of \( E_s/N_0 \) of this guideline. The F-LDPC operates approximately 1 dB or closer to this guideline over a large range of operational scenarios. This is illustrated in Fig. 3.

4. CODE-DRIVEN CP WF

The proposed approach first performs CP-WF and then does bit allocation by exploiting the flexibility of the TLC. This is based on identifying SNR "switch points" that define the proper modulation to be used as a function of SNR. For example, considering Fig. 3, a range of SNR should be defined for each of the 5 modulations. The boundaries of these regions could reasonably be selected as: \( \Gamma_1 = 1.5 \) dB, \( \Gamma_2 = 6.0 \) dB, \( \Gamma_3 = 13 \) dB, and \( \Gamma_4 = 22 \) dB so that,
for example, QPSK would be used if $\frac{E_s}{N_0} \in (1.5, 6.6)$ dB. Also, if $\frac{E_s}{N_0} < \Gamma_0 = -2$ dB, the SNR is too low for lowest rate coding and modulation scheme (rate 1/2, BPSK) to operate as desired.

**Power Allocation** is performed first considering uniform power allocation of $\frac{E_s[i]}{\sum E_s[i]}$. If $\frac{\gamma[i]E_s[i]}{N_0} < \Gamma_0$, sub-channel $i$ is not used, otherwise it is used. The power is then evenly allocated over all active sub-channels so that $\frac{E_s[i]}{\sum E_s[i]} = \frac{1}{\sum I_a[i]}$, where $I_a$ is the number of active channels.

**Rate Allocation** is performed in three stages. First a modulation is assigned to each active sub-channel according the switch point SNRs – i.e., set $M[i] = M_v$, where $v$ is the largest value such that $\Gamma_v \leq \frac{E_s[i]}{N_0}$. Second, compute the total number of binary digits that can be carried by the modulation assignment: $B = \sum \log_2(M[i])$. The third step is computing the operational limit for the number of information bits $R$ that can be transmitted over one set of sub-channel uses. Estimating $R$ could be accomplished using (6). However, if there is sufficient overlap between the SNR ranges accommodated by the different modulation formats, the degradation relative to AWGN capacity will be approximately constant over all modulations. Therefore $R$ can be computed as $R = \sum \log_2\left(1 + \alpha(\frac{E_s[i]}{N_0})\right)$ where $\alpha$ is a margin accounting for the effects of non-Gaussian signals, the block size, and the target BLER. In the example of Fig. 3 $\alpha = 0.66$ (i.e., approximately 1.8 dB from AWGN capacity).

Finally, the binary TWC code rate is selected to be the closest rate possible to rate $r = R/B$. Then, this single TWC block is loaded onto the active sub-channels as illustrated in Fig. 1. The targeted spectral efficiency is then $\frac{1}{2} R = \frac{1}{2} \sum \log_2(M[i])$ bits per total channel use (i.e., all $I$ sub-channels). When the input block size $K$ is large enough, multiple channel uses will be contained in one TWC block.

5. CDCP-WF SIMULATION RESULTS

In the following simulations $\left\{\sqrt{\frac{\gamma[i]}{N_0}}\right\}$ are generated using independent, Rayleigh model with $E\{\gamma[i]\} = 1$.

First, consider the effectiveness of the power allocation approach without enforcing a modulation constraint. Fig. 4 compares the code-based sub-channel selection (with $\Gamma_0 = -2$ dB) to the CP-WF approach in [8]. Note that this yields acceptable sub-channel selection and the degradation relative to optimal WF is negligible.

The performance of the complete CDCP-WF algorithm (i.e., including modulation and F-LDPC rate assignment) is shown in Figs. 5-6. For each given F-LDPC code block, independent realizations of $\{\gamma[i]\}$ were generated, the CDCP-WF algorithm was run to set the target average spectral efficiency for the target BLER of 0.01, and then decoding was performed for the code rate and modulation assignments. This process was continued until at least 100 F-LDPC block errors were observed. The observed average spectral efficiency was found to closely match the targeted value. Fig. 5 shows the observed spectral efficiency vs. the value of $\frac{E_s}{N_0}$ required reach the target BLER of 0.01 for various $K$ and $i = 64$. The performance is approximately 4 dB from the ideal WF capacity for $K = 8000$. Note that the degradation in performance for smaller block sizes is well predicted by (6). Fig. 6 shows similar results for a fixed input block size and various values of $I$. Over the range of $I$ and SNR considered, the performance is relatively insensitive to $I$. Results were also obtained using the recommended MIMO-OFDM channel models for the IEEE 802.11n standardization process with similar results [16].

6. CONCLUSIONS

The existence of good, flexible modern codes, such as the Flexible LDPC code used in this paper, enables one to approach the ideal WF capacity for parallel AWGN channels.
Fig. 6. Spectral efficiency of the CDCP-WF algorithm for various numbers of sub-channels and $K = 8000$ input bits.

Furthermore, this is achieved with a simple CP-WF driven by the universally good performance properties of the TLC and requires relatively small encoding (decoding) latency. Digital hardware for such codes is currently available with throughputs approaching gigabits/sec. This technique is only limited by the range of spectral efficiencies supported by the TLC and the modulation formats. In this paper, this was from 0.5 bps/Hz to approximately 8 bps/Hz.

7. REFERENCES


