# Learning to Broadcast with Layered Division Multiplexing

Roy Karasik*, Osvaldo Simeone†, and Shlomo Shamai (Shitz)*
*Department of Electrical and Computer Engineering, Technion, Haifa 32000, Israel
†King's Communications, Learning & Information Processing (KCLIP) lab, Centre for Telecommunications Research,
Department of Engineering, King's College London, London WC2R 2LS, U.K.
{royk@campus.technion.ac.il, osvaldo.simeone@kcl.ac.uk, sshlomo@ee.technion.ac.il}

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834785

***Abstract*—A broadcast/multicast communication system is studied in which layered division multiplexing (LDM) is applied to support differential quality-of-service (QoS) levels. Focusing on a practical scenario in which the transmitter does not know the fading distribution, layer allocation is optimized based on a dataset sampled during deployment. The optimality gap caused by the availability of limited data is bounded via a generalization analysis, and is shown to be monotonically decreasing as the dataset grows larger. Numerical experiments demonstrate that LDM improves spectral efficiency even for small datasets; and that, for sufficiently large datasets, the proposed mirror-descent-based layer optimization scheme achieves an expected rate close to that achieved when the transmitter knows the fading distribution.**

## I. Introduction

Layered division multiplexing (LDM) has been introduced in several standards as an effective means to support differential quality-of-service (QoS) in broadcast and multicast services. With LDM, multiple independent sub-messages, or layers, are superimposed, enabling the decoding of a different number of messages depending on the channel conditions, thus supporting communication at a variable rate [1]–[4]. The most common use of LDM is for multimedia broadcast, as adopted by the Advanced Television Systems Committee (ATSC 3.0) [4], [5], in which LDM supports a robust configuration for mobile receivers and a high-capacity connection for fixed receivers. Other applications include Machine-Type Communication (MTC) and Industry 4.0, in which LDM is considered as a tool to deliver critical control services and best-effort monitoring services [6]–[8]. Maximizing the expected achievable rate, or average rate across all receivers, requires adjusting the layers' rates and power levels as a function of the channel distribution [9]. However, in practice, this distribution is unknown. Accordingly, in this paper, we assume the transmitter has access to a dataset sampled during deployment, from which the rate and power allocation for each layer are optimized. We explore theoretic and algorithmic aspects of this design problem.

*Related Work:* LDM, also known as the *broadcast approach*, has been extensively studied as means to improve spectral efficiency in various scenarios. A comprehensive survey of the state-of-the-art is available in [1], and we mention here some representative examples. The broadcast approach for slowly fading single-user channels was investigated in [9], where it was shown that transmitting multiple layers can increase the expected achievable rate, and the optimal power allocation density was derived for an infinite number of layers. The gain of the broadcast approach was also demonstrated in [10] for finite number of layers. Specifically, for quasi-static Rayleigh fading channel, two layers were shown to achieve most of the throughput gain. Importantly, unlike our work, both references [9] and [10] assume that the transmitter knows the fading distribution. In [2], for broadcasting fixed and mobile services, LDM with two layers was shown to outperform time division multiplexing (TDM) and frequency division multiplexing (FDM) in terms of the mobile service's capacity-coverage trade-off. Multicast beamforming was studied in [11] with the goal of minimizing the outage probability for unknown fading distribution, and several gradient-based algorithms were proposed to optimize beamforming based on a dataset of channel samples. Similarly, an alternating gradient descent algorithm was recently proposed in [12] for the joint optimization of the precoding weights and the reconfigurable intelligent surface (RIS) reflection pattern in RIS-aided communication system.

*Main Contributions:* In this paper, we study the LDM-based broadcasting/multicasting system illustrated in Fig. 1, in which a single-antenna base station (BS) serves single-antenna clients. The channel coefficients and the the fading distribution are assumed to be unknown to the BS. In order to maximize the expected achievable rate, the BS optimizes layer allocation based on a dataset sampled during deployment. At a theoretical level, we bound the optimality gap caused by the availability of limited data via a generalization analysis [13], and characterize the number of samples required to maintain a desired optimality gap. At an algorithmic level, we introduce a mirror-descent based scheme [14] to maximize an empirical estimate of the expected rate. Numerical results demonstrate that broadcasting multiple layers improves spectral efficiency even for small datasets, and that, for sufficiently large datasets, the expected rate is close to that achieved when the BS knows the fading distribution, confirming the sample complexity analysis.

*Notation:* Random variables and vectors are denoted by lowercase and boldface lowercase Roman-font letters, respectively. Realizations of random variables and vectors are denoted by lowercase and boldface lowercase italic-font letters, respectively. For example, $x$ is a realization of random variable $\mathrm{x}$ and $\boldsymbol{x}$ is a realization of random vector $\mathbf{x}$. For any

This work has been supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement Nos. 694630 and 725731).

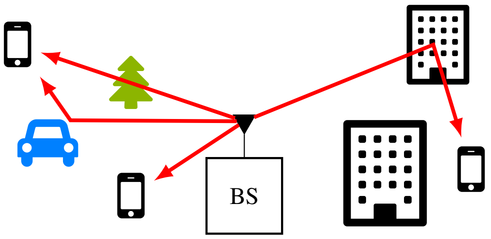

Fig. 1. Illustration of the broadcast setting under study. A single-antenna base station (BS) broadcasts a common message to single-antenna clients. The signal to each client undergoes a fading channel in which the fading coefficient is drawn from a common fading distribution $p_{\mathrm{h}}(h)$.

positive integer $K$, we define the set $[K] \triangleq \{1, 2, \ldots, K\}$. The cardinality and convex hull of a set $\mathcal{L}$ are denoted by $|\mathcal{L}|$ and $\mathrm{conv}(\mathcal{L})$, respectively. The $\ell^1$-norm and $\ell^2$-norm of a vector $\boldsymbol{s}$ are denoted by $\|\boldsymbol{s}\|_1$ and $\|\boldsymbol{s}\|_2$, respectively. For two scalars $a$ and $b$, the indicator of the event $a \geq b$ is denoted by $\mathbf{1}_{a \geq b}$. That is, $\mathbf{1}_{a \geq b}$ equals one if $a \geq b$ and zero otherwise. The set of non-negative real numbers is denoted by $\mathbb{R}_+$. $\mathrm{diag}(\boldsymbol{u})$ represents a diagonal matrix with diagonal given by the vector $\boldsymbol{u}$.

## II. System Model and Problem Definition

We consider the system depicted in Fig. 1 in which a single-antenna BS broadcasts a common message to single-antenna clients over a fading broadcast channel. The fading coefficient for each client is drawn from a common fading distribution $p_{\mathrm{h}}(h)$, and is assumed to remain constant for the duration of a coding block consisting of $n$ symbols. The common fading distribution $p_{\mathrm{h}}(h)$ may take the form of a mixture model, as in [15], in order to account for heterogeneous long-term effects such as path loss and shadowing.

The signal received by a client at time $t \in [n]$, denoted by $\mathrm{y}(t)$, can be expressed as

$$\mathrm{y}(t) = \sqrt{P}\mathrm{h}\mathrm{x}(t) + \mathrm{z}(t), \tag{1}$$

where $P > 0$ denotes the BS transmission power; $\mathrm{x}(t) \in \mathbb{C}$ denotes the signal transmitted at time $t$, which is subject to the average power constraint

$$\mathbb{E}\left[|\mathrm{x}(t)|^2\right] \leq 1; \tag{2}$$

channel coefficient $\mathrm{h} \sim p_{\mathrm{h}}(h)$ denotes the quasi-static fading coefficient; and $\mathrm{z}(t) \sim \mathcal{CN}(0,1)$ denotes the additive white Gaussian noise (AWGN).

We assume that the BS does not know the fading realizations nor the common fading distribution $p_{\mathrm{h}}(h)$, while each client knows its own channel h. Due to the lack of channel state information (CSI), the BS applies layered division multiplexing (LDM) [9] with $M$ layers, or sub-messages, in order to enable differential quality of service at the clients. The transmitted signal $\mathrm{x}(t)$ in (1) is accordingly given as

$$\mathrm{x}(t) = \sum_{m=1}^{M} \mathrm{x}_m(t), \tag{3}$$

where $\mathrm{x}_m(t) \sim \mathcal{CN}(0, \lambda_m)$, with $m \in [M]$, denotes a symbol from a Gaussian random codebook with average power $\lambda_m$ that is used to encode sub-message $\mathrm{w}_m \in [2^{n\rho_m}]$ of rate $\rho_m \geq 0$. To satisfy the normalized power constraint in (2), the power-allocation vector $\boldsymbol{\lambda} \triangleq (\lambda_1, \ldots, \lambda_M)$ must thus lie in the simplex

$$\Delta_c^M \triangleq \left\{\boldsymbol{\lambda} \in \mathbb{R}_+^M : \sum_{m=1}^{M} \lambda_m \leq 1\right\}. \tag{4}$$

We refer to message $\mathrm{w}_m$ and corresponding encoded signal $\mathrm{x}_m(t)$ as the $m$th layer.

Each client decodes sub-messages by applying *successive cancellation decoding* (SCD) with the order $\mathrm{w}_1, \ldots, \mathrm{w}_M$. When decoding layer $m \in [M]$, all subsequent layers are treated as AWGN. Each client can hence decode only a subset of layers depending on its channel gain $\mathrm{g} \triangleq |\mathrm{h}|^2$. We denote by $I_m \triangleq \sum_{i=m+1}^{M} \lambda_i$ the normalized power level of the inter-layer interference affecting the decoding of layer $m$, and as $p_{\mathrm{g}}(g)$ the distribution of the channel gain g.

We parametrize the rate $\rho_m$ of layer $m$ as [9]

$$\rho_m(\boldsymbol{s}^m, \boldsymbol{\lambda}) \triangleq \log_2\left(1 + \frac{\|\boldsymbol{s}^m\|_1 \lambda_m P}{1 + \|\boldsymbol{s}^m\|_1 I_m P}\right), \tag{5}$$

where $\boldsymbol{s} \triangleq (s_1, \ldots, s_M) \in \mathbb{R}_+^M$ is a non-negative vector set by the BS, and vector $\boldsymbol{s}^m \triangleq (s_1, \ldots, s_m) \in \mathbb{R}_+^m$ consists of the first $m$ elements of $\boldsymbol{s}$. Assuming that all previous layers are correctly decoded, the rate achievable for layer $m$ by a client with channel gain g is $\log_2(1 + \mathrm{g}\lambda_m P/(1 + \mathrm{g}I_m P))$. Therefore, the client can decode all layers up to layer $m$ if and only if its channel gain satisfies the inequality $\mathrm{g} \geq \|\boldsymbol{s}^m\|_1$. Accordingly, given the power and rate allocation vectors $\boldsymbol{\lambda}$ and $\boldsymbol{s}$, the total rate that can be decoded by a client with channel gain $g$ is given as

$$R(\boldsymbol{s}, \boldsymbol{\lambda}, g) \triangleq \sum_{m=1}^{M} \rho_m(\boldsymbol{s}^m, \boldsymbol{\lambda}) \mathbf{1}_{g \geq \|\boldsymbol{s}^m\|_1}. \tag{6}$$

We study the maximization of the expected achievable rate

$$\bar{R}(\boldsymbol{s}, \boldsymbol{\lambda}) \triangleq \mathbb{E}_{\mathrm{g}}\left[R(\boldsymbol{s}, \boldsymbol{\lambda}, \mathrm{g})\right], \tag{7}$$

where the expectation is over the fading distribution $p_{\mathrm{g}}(g)$, with respect to the power and rate allocation vectors $\boldsymbol{\lambda}$ and $\boldsymbol{s}$. That is, we consider the optimization problem

$$(\boldsymbol{s}^*, \boldsymbol{\lambda}^*) \in \underset{(\boldsymbol{s}, \boldsymbol{\lambda}) \in \mathbb{R}_+^M \times \Delta_c^M}{\arg\max} \ \bar{R}(\boldsymbol{s}, \boldsymbol{\lambda}). \tag{8}$$

## III. Empirical Average Rate Maximization

In this paper, we assume that the BS does not know the fading distribution $p_{\mathrm{g}}(g)$, and hence it cannot directly optimize the expected achievable rate $\bar{R}(\boldsymbol{s}, \boldsymbol{\lambda})$. Instead, we assume that the BS has access to a dataset

$$\mathcal{G} = \{g_1, \ldots, g_N\} \tag{9}$$

consisting of $N$ fading realizations sampled in an independent and identically distributed (i.i.d.) manner from distribution $p_{\mathrm{g}}(g)$. Based on dataset $\mathcal{G}$, which is collected offline, e.g., during deployment, the BS approximates the expected achievable rate with the empirical average

$$\bar{R}^{\mathcal{G}}(\boldsymbol{s}, \boldsymbol{\lambda}) = \frac{1}{N} \sum_{i=1}^{N} R(\boldsymbol{s}, \boldsymbol{\lambda}, g_i). \tag{10}$$

The maximization of the average rate (10) over power and rate allocation vectors $\boldsymbol{\lambda}$ and $\boldsymbol{s}$ can be expressed as the optimization problem

$$(\boldsymbol{s}^{\mathcal{G}},\boldsymbol{\lambda}^{\mathcal{G}}) \in \underset{(\boldsymbol{s},\boldsymbol{\lambda})\in\mathbb{R}_+^M\times\Delta_c^M}{\arg\max}\ \bar{R}^{\mathcal{G}}(\boldsymbol{s},\boldsymbol{\lambda}). \tag{11}$$

A solution to problem (11) can be practically obtained via an iterative optimization scheme as detailed in Section IV.

We emphasize that optimizing the average rate $\bar{R}^{\mathcal{G}}(\boldsymbol{s},\boldsymbol{\lambda})$ via problem (11) is useful not only when the fading distribution $p_{\mathrm{g}}(g)$ is unknown, but also when the direct optimizations in (8) based on knowledge of the distribution $p_{\mathrm{g}}(g)$ is not tractable. In this latter case, one can potentially generate the dataset $\mathcal{G}$ with an arbitrary number of fading realizations $N$.

### A. Optimality Gap and Sample Complexity

An important theoretical question is whether the expected achievable rate obtained under the power and rate allocation vectors (11) approaches the ground-truth maximum expected achievable rate obtained with vectors (8) as the size of the dataset increases. If so, it would also be interesting to quantify how many samples $N$ are required to achieve a desired level of approximation. This is the subject of this subsection.

To proceed, we define the optimality gap

$$e^{\mathcal{G}} \triangleq \bar{R}\big(\boldsymbol{s}^*,\boldsymbol{\lambda}^*\big) - \bar{R}\big(\boldsymbol{s}^{\mathcal{G}},\boldsymbol{\lambda}^{\mathcal{G}}\big) \tag{12}$$

as the difference between the expected rate achieved with optimal power and rate allocation vectors (8) and the expected rate achieved by the empirical rate maximization (11). The optimality gap is random due to the stochastic nature of the dataset $\mathcal{G}$.

To bound the optimality gap, we assume that the norms of the optimal vectors $\boldsymbol{s}^*$ and $\boldsymbol{s}^{\mathcal{G}}$ in (8) and (11), respectively, can be bounded as $\max\{\|\boldsymbol{s}^*\|_1, \|\boldsymbol{s}^{\mathcal{G}}\|_1\} \leq S$ for some known constant $S > 0$. Note that this assumption is not restrictive since, in practice, $S$ represents the largest fading gain $g$ that a client is expected to experience. The following proposition bounds the optimality gap under this assumption.

*Proposition 1:* Let $\mathcal{G} = \{g_1,\ldots,g_N\}$ be a dataset of $N$ fading realizations drawn independently from the fading distribution $p_{\mathrm{g}}(g)$, and let $\delta \in (0,1]$. With probability at least $1-\delta$, the optimality gap (12) is bounded, for rate allocation vectors with bounded norms $\max\{\|\boldsymbol{s}^*\|_1, \|\boldsymbol{s}^{\mathcal{G}}\|_1\} \leq S$, as

$$e^{\mathcal{G}} \leq \left(4\sqrt{\frac{(2N+1)\ln(N+1)}{3N(N+1)}} + \sqrt{\frac{2\ln(2/\delta)}{N}}\right) 2\log_2\left(1+SP\right). \tag{13}$$

*Proof:* See Appendix A. ■

This result shows that the optimality gap scales with number of data points, $N$, as $\mathcal{O}(\sqrt{\ln(N)/N})$, implying that any level of accuracy can be attained as the dataset grows larger, i.e., as $N \to \infty$. Furthermore, for a given desired optimality gap $e^{\mathcal{G}} \leq \epsilon$, the required number of data points $N$, i.e., the sample complexity, satisfies the approximate inequality

$$\frac{N}{\ln(N)} \gtrapprox \left(\frac{\log_2(SP)}{\epsilon}\right)^2 \tag{14}$$

for large $N$. Intuitively, the sample complexity increases with the signal-to-noise ratio (SNR) metric $SP$ since, as the achievable rate increases, a better approximation is required to achieve the same subtractive optimality gap.

## IV. Mirror Gradient Descent

In this section, we introduce a gradient-based iterative optimization procedure to tackle the empirical average rate maximization problem (11). The approach is based on the introduction of a surrogate smooth objective and on mirror descent, as described in the rest of this section and summarized in Algorithm 1.

**Algorithm 1:** Empirical average rate maximization

```
Input: Dataset $\mathcal{G}$
Initialization: Initialize $\boldsymbol{u} \in \mathbb{R}^M$ and $\boldsymbol{\lambda} \in \Delta_c^M$
set $i = 0$
set $\boldsymbol{u}^{(i)} = \boldsymbol{u}$ and $\boldsymbol{\lambda}^{(i)} = \boldsymbol{\lambda}$
while not converged do
    set $i \leftarrow i+1$
    set $\boldsymbol{u}^{(i)} \leftarrow \mathrm{GD}(\boldsymbol{u}^{(i-1)}; \mathcal{G}, \boldsymbol{\lambda}^{(i-1)})$ (defined in (19))
    set $\boldsymbol{\lambda}^{(i)} \leftarrow \mathrm{EG}(\boldsymbol{\lambda}^{(i-1)}; \mathcal{G}, \boldsymbol{u}^{(i-1)})$ (defined in (20))
return $\big(\boldsymbol{s}^{(i)} = \exp(\boldsymbol{u}^{(i)}), \boldsymbol{\lambda}^{(i)}\big)$
```

### A. Smooth Surrogate Objective

A first challenge in developing iterative solutions to problem (11), is that the partial derivative of the indicator in the achievable rate expression (6) with respect to vector $\boldsymbol{s}$ equals zero almost everywhere. Therefore, in order to facilitate the application of a gradient-based optimization procedure, we replace the rate $R(\boldsymbol{s},\boldsymbol{\lambda},g)$ in (6) with the smooth surrogate objective

$$R_{\sigma}(\boldsymbol{s},\boldsymbol{\lambda},g) \triangleq \sum_{m=1}^{M} \rho_m(\boldsymbol{s}^m,\boldsymbol{\lambda})\sigma(c(g-\|\boldsymbol{s}^m\|_1)), \tag{15}$$

where $\sigma(x) \triangleq 1/(1+\exp(-x))$ is the sigmoid function, and the parameter $c > 0$ determines the trade-off between smoothness and accuracy of the surrogate approximation. As $c \to \infty$, the surrogate (15) tends uniformly to the original rate (6), while smaller values of $c$ yield non-zero partial derivatives with respect to $\boldsymbol{s}$.

Using the approximation (15), we define the *surrogate empirical average rate maximization problem* as

$$\big(\tilde{\boldsymbol{s}}^{\mathcal{G}},\tilde{\boldsymbol{\lambda}}^{\mathcal{G}}\big) = \underset{(\boldsymbol{s},\boldsymbol{\lambda})\in\mathbb{R}_+^M\times\Delta_c^M}{\arg\max}\ \tilde{R}^{\mathcal{G}}(\boldsymbol{s},\boldsymbol{\lambda}), \tag{16}$$

where $\tilde{R}^{\mathcal{G}}(\boldsymbol{s},\boldsymbol{\lambda})$ denotes the *surrogate average rate*

$$\tilde{R}^{\mathcal{G}}(\boldsymbol{s},\boldsymbol{\lambda}) \triangleq \frac{1}{N}\sum_{i=1}^{N} R_{\sigma}(\boldsymbol{s},\boldsymbol{\lambda},g_i). \tag{17}$$

### B. Mirror Descent

Although the objective in (16) is smooth, plain-vanilla gradient descent cannot be applied to address the optimization (16) due to the domain constraints on the optimization variables $(\boldsymbol{s},\boldsymbol{\lambda}) \in \mathbb{R}_+^M \times \Delta_c^M$. To tackle the constraint $\boldsymbol{s} \in \mathbb{R}_+^M$, we parametrize the rate-allocation vector $\boldsymbol{s}$ with a vector $\boldsymbol{u} \in \mathbb{R}^M$ as

$$\boldsymbol{s} = \exp(\boldsymbol{u}) \triangleq (\exp(u_1),\ldots,\exp(u_M)). \tag{18}$$

Furthermore, to satisfy the constraint $\boldsymbol{\lambda} \in \Delta_c^M$, we consider a mirror-descent based scheme which adapts the updates to the geometry of the simplex $\Delta_c^M$ via the exponentiated gradient [16]. Overall, this leads to the updates

$$\boldsymbol{u} \leftarrow \boldsymbol{u} + \eta\, \mathrm{diag}(\exp(\boldsymbol{u}))\, \nabla_{\boldsymbol{s}} \tilde{R}^{\mathcal{G}}(\boldsymbol{s},\boldsymbol{\lambda})\Big|_{\boldsymbol{s}=\exp(\boldsymbol{u})} \triangleq \mathrm{GD}(\boldsymbol{u};\mathcal{G},\boldsymbol{\lambda}) \tag{19}$$

and

$$\lambda_m \leftarrow \frac{\lambda_m \exp\left(\gamma[\nabla_{\boldsymbol{\lambda}}\tilde{R}^{\mathcal{G}}(\exp(\boldsymbol{u}),\boldsymbol{\lambda})]_m\right)}{\sum_{m'=1}^{M}\lambda_{m'}\exp\left(\gamma[\nabla_{\boldsymbol{\lambda}}\tilde{R}^{\mathcal{G}}(\exp(\boldsymbol{u}),\boldsymbol{\lambda})]_{m'}\right)} \triangleq \mathrm{EG}(\boldsymbol{\lambda};\mathcal{G},\boldsymbol{u}),\ \forall m \in [M]. \tag{20}$$

The resulting procedure to optimize the empirical average rate is summarized in Algorithm 1.

## V. Numerical Results

In this section, we evaluate the expected rate $\bar{R}(\tilde{\boldsymbol{s}}^{\mathcal{G}},\tilde{\boldsymbol{\lambda}}^{\mathcal{G}})$ for parameters $(\tilde{\boldsymbol{s}}^{\mathcal{G}},\tilde{\boldsymbol{\lambda}}^{\mathcal{G}})$ obtained via Algorithm 1 with learning rates $\eta = \gamma = 0.01$ and sigmoid smoothness parameter $c = 10$. The expected rate is averaged over 1000 datasets $\mathcal{G}$, which we denote as $\mathbb{E}_{\mathcal{G}}\left[\bar{R}(\tilde{\boldsymbol{s}}^{\mathcal{G}},\tilde{\boldsymbol{\lambda}}^{\mathcal{G}})\right]$.

In Fig. 2, we plot the expected achievable rate as a function of the number of layers $M$ with power $P = 20$dB, Rayleigh fading distribution, and dataset of size $N = 10$, 100, and 1000. For this special case, the ideal optimal solution obtained by using infinite layers and assuming that the fading distribution is known was derived in [9], and is used as an upper bound. Furthermore, we plot for reference the expected rate achieved with finite number of layers when the BS knows the fading distribution, which is obtained by replacing the surrogate empirical average rate $\tilde{R}^{\mathcal{G}}(\boldsymbol{s},\boldsymbol{\lambda})$ with the expected rate $\bar{R}(\boldsymbol{s},\boldsymbol{\lambda})$ in the gradient-based updates (19)–(20). First, confirming the sample complexity analysis in Section III-A, for sufficiently large datasets, the expected rate is close to that achieved when the BS knows the fading distribution. Furthermore, using multiple layers provides notable gain over a single layer, even for small datasets. Finally, the expected rate achieved with $M = 6$ layers and sufficiently large dataset is seen to be close to the upper bound.

In Fig. 3, we plot the ratio of the expected rate achieved via LDM with $M$ layers to the expected rate achieved with a single layer as a function of the power $P$ with Rayleigh fading distribution and dataset of size 1000. It is observed that the gain of LDM increases with power $P$. Intuitively, this is because, for sufficiently high power, splitting the last layer, while keeping the same norm $\|\boldsymbol{s}\|_1$, has a negligible impact on the rate $\rho_M(\boldsymbol{s},\boldsymbol{\lambda})$ but adds another layer that is much more likely to be decoded (see eqs. (5)–(7)).

## VI. Conclusion

In this work, we have studied LDM as an enabler of differential QoS for broadcast/multicast communication systems. We have focused on a practical model in which the fading distribution is unknown, and the transmitter optimizes rate and power allocation for each layer based on a dataset sampled during deployment. The optimality gap caused by the availability of limited data was bounded via a generalization analysis ans was shown to monotonically decrease as the dataset grows larger. To optimize the rate and power allocation parameters, a mirror-descent based scheme was introduced, which, for sufficiently large datasets, was demonstrated via numerical experiments to achieve an expected rate close to that achieved when the BS knows the fading distribution. Among related problems left open by this study, we mention the extension to multiple transmit antennas [11] and to channels with multiple uncoordinated transmitters [17], [18]. An extended version of this work, which introduces the *conditional value-at-risk* (CVaR) rate performance measure for ultra-reliable communication, and considers meta-learning as a means to reduce sample complexity by leveraging data from previous deployments, is available in [19].

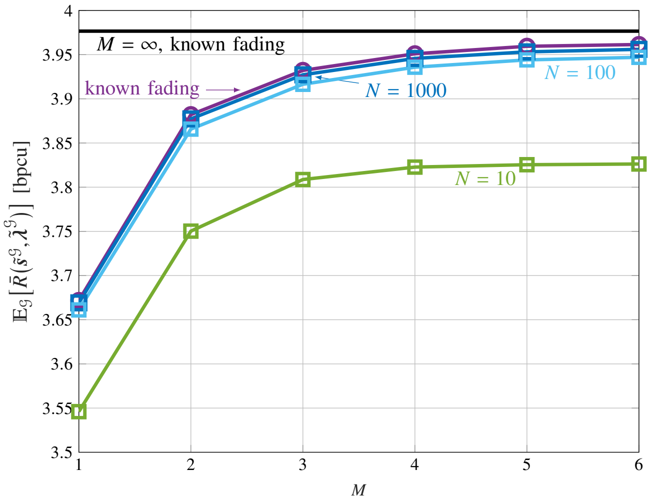

Fig. 2. The expected achievable rate as a function of $M$ with $P = 20$dB and $N = 10$, 100, and 1000.

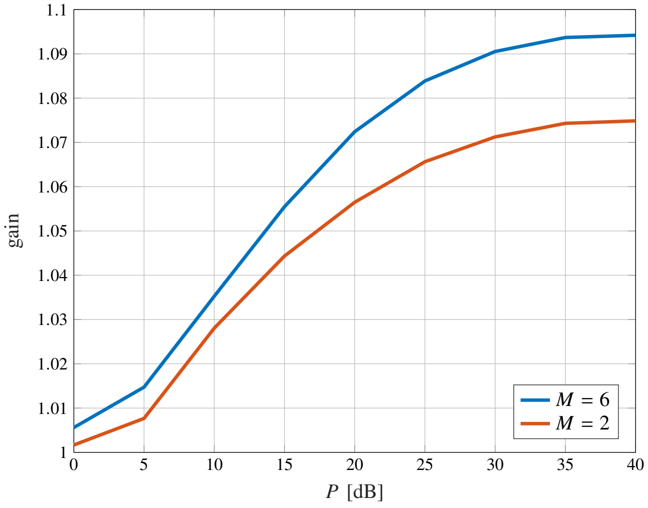

Fig. 3. The expected achievable rate gain as a function of $P$ with $N = 1000$, and $M = 2$ and 6.

## Appendix

### A. Proof of Proposition 1

The optimality gap $e^{\mathcal{G}}$ (12) can be upper bounded as

$$\begin{aligned} e^{\mathcal{G}} &= \bar{R}(\boldsymbol{s}^*,\boldsymbol{\lambda}^*) - \bar{R}^{\mathcal{G}}(\boldsymbol{s}^*,\boldsymbol{\lambda}^*) + \bar{R}^{\mathcal{G}}(\boldsymbol{s}^*,\boldsymbol{\lambda}^*) - \bar{R}^{\mathcal{G}}(\boldsymbol{s}^{\mathcal{G}},\boldsymbol{\lambda}^{\mathcal{G}}) \\ &\quad + \bar{R}^{\mathcal{G}}(\boldsymbol{s}^{\mathcal{G}},\boldsymbol{\lambda}^{\mathcal{G}}) - \bar{R}(\boldsymbol{s}^{\mathcal{G}},\boldsymbol{\lambda}^{\mathcal{G}}) \\ &\leq \left(\bar{R}(\boldsymbol{s}^*,\boldsymbol{\lambda}^*) - \bar{R}^{\mathcal{G}}(\boldsymbol{s}^*,\boldsymbol{\lambda}^*)\right) + \left(\bar{R}^{\mathcal{G}}(\boldsymbol{s}^{\mathcal{G}},\boldsymbol{\lambda}^{\mathcal{G}}) - \bar{R}(\boldsymbol{s}^{\mathcal{G}},\boldsymbol{\lambda}^{\mathcal{G}})\right), \end{aligned}$$

(21)

where the inequality holds since $(\boldsymbol{s}^{\mathcal{G}},\boldsymbol{\lambda}^{\mathcal{G}})$ maximize the average rate $\bar{R}^{\mathcal{G}}(\boldsymbol{s},\boldsymbol{\lambda})$. Next, to further bound the optimality gap, we bound, uniformly, the difference $|\bar{R}(\boldsymbol{s},\boldsymbol{\lambda}) - \bar{R}^{\mathcal{G}}(\boldsymbol{s},\boldsymbol{\lambda})|$ for all $\boldsymbol{\lambda} \in \Delta_c^M$ and $\boldsymbol{s} \in \mathbb{R}_+^M$ with $\|\boldsymbol{s}\|_1 \le S$. Note that the expected achievable rate (7) can be expressed as

$$\bar{R}(\boldsymbol{s},\boldsymbol{\lambda}) = \mathbb{E}_{\mathrm{g}}\left[R(\boldsymbol{s},\boldsymbol{\lambda},\mathrm{g})\right] = \sum_{m=1}^{M} \rho_m(\boldsymbol{s}^m,\boldsymbol{\lambda})\bar{F}_{\mathrm{g}}(\|\boldsymbol{s}^m\|_1), \tag{22}$$

where $\bar{F}_{\mathrm{g}}(\|\boldsymbol{s}^m\|_1)$ denotes the complementary cumulative distribution function (CCDF)

$$\bar{F}_{\mathrm{g}}(\|\boldsymbol{s}^m\|_1) \triangleq \Pr\left[\mathrm{g} \ge \|\boldsymbol{s}^m\|_1\right]. \tag{23}$$

Similarly, the average rate (10) can be expressed as

$$\bar{R}^{\mathcal{G}}(\boldsymbol{s},\boldsymbol{\lambda}) = \sum_{m=1}^{M} \rho_m(\boldsymbol{s}^m,\boldsymbol{\lambda})\bar{F}_{\mathrm{g}}^{\mathcal{G}}(\|\boldsymbol{s}^m\|_1), \tag{24}$$

where $\bar{F}_{\mathrm{g}}^{\mathcal{G}}(\|\boldsymbol{s}^m\|_1)$ denotes the empirical CCDF

$$\bar{F}_{\mathrm{g}}^{\mathcal{G}}(\|\boldsymbol{s}^m\|_1) \triangleq \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{g_i \ge \|\boldsymbol{s}^m\|_1}. \tag{25}$$

Therefore, to uniformly bound the difference $|\bar{R}(\boldsymbol{s},\boldsymbol{\lambda}) - \bar{R}^{\mathcal{G}}(\boldsymbol{s},\boldsymbol{\lambda})|$, we first uniformly bound $|\bar{F}_{\mathrm{g}}(s) - \bar{F}_{\mathrm{g}}^{\mathcal{G}}(s)|$ using the following proposition.

*Proposition 2:* Let $\mathcal{G} = \{g_1,\ldots,g_N\}$ be a dataset of $N$ fading realizations drawn independently from the fading distribution $p_{\mathrm{g}}(g)$, and let $\delta \in (0,1]$. With probability at least $1-\delta$, uniformly over all $s \in \mathbb{R}_+$, we have

$$\left|\bar{F}_{\mathrm{g}}(s) - \bar{F}_{\mathrm{g}}^{\mathcal{G}}(s)\right| \le 4\sqrt{\frac{(2N+1)\ln(N+1)}{3N(N+1)}} + \sqrt{\frac{2\ln(2/\delta)}{N}}. \tag{26}$$

*Proof:* See Appendix B. ∎

Proposition 2 implies that, with probability at least $1-\delta$, we can bound the difference $|\bar{R}(\boldsymbol{s},\boldsymbol{\lambda}) - \bar{R}^{\mathcal{G}}(\boldsymbol{s},\boldsymbol{\lambda})|$, uniformly over all $\boldsymbol{\lambda} \in \Delta_c^M$ and $\boldsymbol{s} \in \mathbb{R}_+^M$ with $\|\boldsymbol{s}\|_1 \le S$, as

$$\begin{aligned}
&\left|\bar{R}(\boldsymbol{s},\boldsymbol{\lambda}) - \bar{R}^{\mathcal{G}}(\boldsymbol{s},\boldsymbol{\lambda})\right| \qquad (27)\\
&\overset{\text{(a)}}{=} \left|\sum_{m=1}^{M} \rho_m(\boldsymbol{s}^m,\boldsymbol{\lambda})\left[\bar{F}_{\mathrm{g}}(\|\boldsymbol{s}^m\|_1) - \bar{F}_{\mathrm{g}}^{\mathcal{G}}(\|\boldsymbol{s}^m\|_1)\right]\right|\\
&\overset{\text{(b)}}{\le} \sum_{m=1}^{M} \rho_m(\boldsymbol{s}^m,\boldsymbol{\lambda})\left|\bar{F}_{\mathrm{g}}(\|\boldsymbol{s}^m\|_1) - \bar{F}_{\mathrm{g}}^{\mathcal{G}}(\|\boldsymbol{s}^m\|_1)\right|\\
&\overset{\text{(c)}}{\le} \left(4\sqrt{\frac{(2N+1)\ln(N+1)}{3N(N+1)}} + \sqrt{\frac{2\ln(2/\delta)}{N}}\right)\sum_{m=1}^{M} \rho_m(\boldsymbol{s}^m,\boldsymbol{\lambda})\\
&\overset{\text{(d)}}{\le} \left(4\sqrt{\frac{(2N+1)\ln(N+1)}{3N(N+1)}} + \sqrt{\frac{2\ln(2/\delta)}{N}}\right)\log_2(1+SP),
\end{aligned}$$

where (a) follows from (22) and (24); (b) follows from from the triangle inequality and since the rate of each layer is non-negative; (c) follows from Proposition 2; and (d) holds since $S \ge \|\boldsymbol{s}\|_1$. Finally, based on inequalities (21) and (27), we can upper bound the optimality gap as (13).

### B. Proof of Proposition 2

Let function $\ell : \mathbb{R}\times\mathbb{C} \mapsto \{0,1\}$ be defined as

$$\ell(s,g) \triangleq \mathbf{1}_{g\ge s}. \tag{28}$$

The true and empirical CCDF can hence be expressed as

$$\bar{F}_{\mathrm{g}}(s) = \mathbb{E}_{\mathrm{g}}\left[\ell(s,\mathrm{g})\right] \tag{29}$$

and

$$\bar{F}_{\mathrm{g}}^{\mathcal{G}}(s) = \frac{1}{N}\sum_{i=1}^{N}\ell(s,g_i), \tag{30}$$

respectively, where $g_1,\ldots,g_N \in \mathcal{G}$ are $N$ fading realizations. In addition, let $\mathcal{L}(g_1,\ldots,g_N) \subset \{0,1\}^N$ be the set

$$\mathcal{L}(g_1,\ldots,g_N) \triangleq \{(\ell(s,g_1),\ldots,\ell(s,g_N)) : s \in \mathbb{R}\}. \tag{31}$$

Furthermore, denote by $\mathrm{Rad}(\mathcal{L}(g_1,\ldots,g_N))$ the *Rademacher complexity* of set $\mathcal{L}(g_1,\ldots,g_N)$, i.e.,

$$\mathrm{Rad}(\mathcal{L}(g_1,\ldots,g_N)) \triangleq \frac{1}{N}\mathbb{E}_{\mathbf{b}}\left[\sup_{\boldsymbol{\ell}\in\mathcal{L}(g_1,\ldots,g_N)}\sum_{i=1}^{N}\mathrm{b}_i\ell_i\right], \tag{32}$$

where the elements of random vector $\mathbf{b} = (\mathrm{b}_1,\ldots,\mathrm{b}_N) \in \{\pm1\}^N$ are i.i.d. with $\Pr[b_i = 1] = \Pr[b_i = -1] = 1/2$. Since $|\ell(s,g)| \le 1$ for all $g \in \mathbb{R}_+$ and $s \in \mathbb{R}$, by [13, Thm 26.5] and [20, Prop. 8], for random variables $\mathrm{g}_1,\ldots,\mathrm{g}_N$ that are i.i.d. according to $p_{\mathrm{g}}(g)$, we have, with probability of at least $1-\delta$, for all $s \in \mathbb{R}$,

$$\left|\bar{F}_{\mathrm{g}}(s) - \bar{F}_{\mathrm{g}}^{\mathcal{G}}(s)\right| \le 4\mathbb{E}\left[\mathrm{Rad}(\mathcal{L}(\mathrm{g}_1,\ldots,\mathrm{g}_N))\right] + \sqrt{\frac{2\ln(2/\delta)}{N}}. \tag{33}$$

Next, we bound the expected Rademacher complexity $\mathbb{E}\left[\mathrm{Rad}(\mathcal{L}(\mathrm{g}_1,\ldots,\mathrm{g}_N))\right]$ in (33). We assume, without loss of generality (w.l.o.g.), that the channel realizations $g_1,\ldots,g_N \in \mathcal{G}$ are ordered such that $g_i \ge g_j$ for all $j \in [i]$. Note that, if $\ell(s,g_j) = 1$ for some $s \in \mathbb{R}$ then $\ell(s,g_i) = 1$ for all $j \le i \le N$. Therefore, we have

$$|\mathcal{L}(g_1,\ldots,g_N)| = N+1. \tag{34}$$

Denote by

$$\bar{\boldsymbol{\ell}} \triangleq \frac{1}{N+1}\sum_{\boldsymbol{\ell}\in\mathcal{L}(g_1,\ldots,g_N)}\boldsymbol{\ell} = \frac{1}{N+1}(1,2,\ldots,N) \tag{35}$$

the average vector in $\mathcal{L}(g_1,\ldots,g_N)$. Note that

$$\max_{\boldsymbol{\ell}\in\mathcal{L}(g_1,\ldots,g_N)}\left\|\boldsymbol{\ell}-\bar{\boldsymbol{\ell}}\right\|_2 = \left\|\bar{\boldsymbol{\ell}}\right\|_2 = \sqrt{\frac{N(2N+1)}{6(N+1)}}. \tag{36}$$

Hence, by *Massart Lemma* [13, Lemma 26.8], we have

$$\begin{aligned}
\mathrm{Rad}(\mathcal{L}(g_1,\ldots,g_N)) &\le \sqrt{\frac{N(2N+1)}{6(N+1)}}\cdot\frac{\sqrt{2\ln(N+1)}}{N}\\
&= \sqrt{\frac{(2N+1)\ln(N+1)}{3N(N+1)}}
\end{aligned} \tag{37}$$

for any channel realizations $g_1,\ldots,g_N \in \mathbb{R}_+$. This implies that the upper bound in (37) bounds the expected Rademacher complexity $\mathbb{E}\left[\mathrm{Rad}(\mathcal{L}(\mathrm{g}_1,\ldots,\mathrm{g}_N))\right]$ as well. By substituting (37) in (33) we get (26).

## References

[1] A. Tajer, A. Steiner, and S. Shamai (Shitz), "The broadcast approach in communication networks," *Entropy*, vol. 23, no. 1, p. 120, 2021.

[2] D. Gómez-Barquero and O. Simeone, "LDM versus FDM/TDM for unequal error protection in terrestrial broadcasting systems: An information-theoretic view," *IEEE Trans. Broadcast.*, vol. 61, no. 4, pp. 571–579, 2015.

[3] S. Verdu and S. Shamai (Shitz), "Variable-rate channel capacity," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2651–2667, 2010.

[4] L. Zhang, W. Li, Y. Wu, X. Wang, S.-I. Park, H. M. Kim, J.-Y. Lee, P. Angueira, and J. Montalban, "Layered-division-multiplexing: Theory and practice," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 216–232, 2016.

[5] S. I. Park, J.-Y. Lee, S. Myoung, L. Zhang, Y. Wu, J. Montalbán, S. Kwon, B.-M. Lim, P. Angueira, H. M. Kim, N. Hur, and J. Kim, "Low complexity layered division multiplexing for ATSC 3.0," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 233–243, 2016.

[6] E. Arruti, M. Mendicute, and M. Barrenechea, "QoS in industrial wireless networks using LDM," in *Proc. IEEE Int. Workshop Elect., Cont., Meas., Sig. App. Mecha. (ECMSM)*, 2017, pp. 1–6.

[7] ——, "Unequal error protection with LDM in inside carriage wireless communications," in *Proc. Int. Conf. ITS Telecommunications (ITST)*, 2017, pp. 1–5.

[8] J. Montalban, E. Iradier, P. Angueira, O. Seijo, and I. Val, "NOMA-based 802.11n for industrial automation," *IEEE Access*, vol. 8, pp. 168 546–168 557, 2020.

[9] S. Shamai and A. Steiner, "A broadcast approach for a single-user slowly fading MIMO channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2617–2635, 2003.

[10] Y. Liu, K. Lau, O. Takeshita, and M. Fitz, "Optimal rate allocation for superposition coding in quasi-static fading channels," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, 2002, pp. 111–111.

[11] Y. Shi, A. Konar, N. D. Sidiropoulos, X.-P. Mao, and Y.-T. Liu, "Learning to beamform for minimum outage," *IEEE Trans. Signal Process.*, vol. 66, no. 19, pp. 5180–5193, 2018.

[12] W. Fang, M. Fu, Y. Shi, and Y. Zhou, "Outage minimization for intelligent reflecting surface aided MISO communication systems via stochastic beamforming," in *Proc. IEEE Sens. Arr. Multichann. Sig. Process. Workshop (SAM)*, 2020.

[13] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[14] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[15] V. Ntranos, N. D. Sidiropoulos, and L. Tassiulas, "On multicast beamforming for minimum outage," *IEEE Trans. Wireless Commun.*, vol. 8, no. 6, pp. 3172–3181, 2009.

[16] J. Kivinen and M. K. Warmuth, "Exponentiated gradient versus gradient descent for linear predictors," *information and computation*, vol. 132, no. 1, pp. 1–63, 1997.

[17] M. Zohdy, A. Tajer, and S. Shamai (Shitz), "Broadcast approach to multiple access with local CSIT," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7483–7498, 2019.

[18] M. Zohdy, A. Tajer, and S. Shamai, "Distributed interference management: A broadcast approach," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 149–163, 2021.

[19] R. Karasik, O. Simeone, H. Jang, and S. Shamai (Shitz), "Learning to broadcast for ultra-reliable communication with differential quality of service via the conditional value at risk," *arXiv preprint arXiv:2112.02007*, 2021.

[20] N. Weinberger, "Generalization bounds and algorithms for learning to communicate over additive noise channels," *IEEE Trans. Inf. Theory*, 2021.