Optimal Training Design for Channel Estimation in Decode-and-Forward Relay Networks With Individual and Total Power Constraints

Feifei Gao, *Student Member, IEEE*, Tao Cui, *Student Member, IEEE*, and
Arumugam Nallanathan, *Senior Member, IEEE*

Abstract—In this paper, we study the channel estimation and the optimal training design for relay networks that operate under the *decode-and-forward* (DF) strategy with the knowledge of the interference covariance. In addition to the total power constraint on all the relays, we introduce individual power constraint for each relay, which reflects the practical scenario where all relays are separated from one another. Considering the individual power constraint for the relay networks is the major difference from that in the traditional point-to-point communication systems where only a total power constraint exists for all colocated antennas. Two types of channel estimation are involved: maximum likelihood (ML) and minimum mean square error (MMSE). For ML channel estimation, the channels are assumed as *deterministic* and the optimal training results from an efficient multilevel waterfilling type solution that is derived from the majorization theory. For MMSE channel estimation, however, the second-order statistics of the channels are assumed known and the general optimization problem turns out to be nonconvex. We instead consider three special yet reasonable scenarios. The problem in the first scenario is convex and could be efficiently solved by state-of-the-art optimization tools. Closed-form waterfilling type solutions are found in the remaining two scenarios, of which the first one has an interesting physical interpretation as pouring water into caves.

Index Terms—Cave-filling, channel estimation, decode-and-forward, majorization theory, maximum likelihood, minimum mean square error, optimal training, relay networks, waterfilling.

I. INTRODUCTION

E MPLOYING multiple antennas can boost the system capacity by simultaneously transmitting multiple data streams [1], [2] and enhance the transmission reliability by using space-time coding (STC) techniques [3], [4]. Unfortunately, packing more than one antenna onto a small mobile terminal faces many difficulties, such as the size limitation and the hardware complexity. In order to overcome these difficulties, one would resort to the relay network, where the spatial diversity is achieved when relays are deemed as "virtual antennas" for the desired user [5]–[13]. These relay nodes either

can be provided by the telecommunication agency or could be obtained from other cooperative users [9]–[11], where the latter scenario is also referred to as *cooperative communication* since each user, although acts as a relay for a certain period, still has its own information to transmit.

The relay-based transmission is usually divided into two phases. During Phase I, the source broadcasts its information symbol to all relays. During Phase II, the relays would either choose to amplify and retransmit the received signal, or to decode the information bits first and then transmit newly encoded symbols. The former process is referred to as *amplify-and-forward* (AF) relaying and the latter is referred to as *decode-and-forward* (DF) relaying. Various cooperative diversity schemes and STC techniques have been developed in [7]–[13].

Channel estimation and optimal training design for AF relay networks have been recently introduced in [14], where it is shown that the estimation scheme in AF relay networks is quite different from the traditional point-to-point communication systems. For DF relay networks, however, the transmissions during Phase I and Phase II are actually separated by the decoding and re-encoding strategy. Hence, the channel estimation is similar to that in the multiple-input multiple-output (MIMO) system and can be separately performed for the two phases. However, since relays are geographically distributed, the individual power constraint for each relay has to be considered. These individual power constraints form the major challenge and, most of the time, bring difficulties to the optimization approach. Although there exist many training-based channel estimation methods for traditional point-to-point systems [15]–[19], channel estimation with individual power constraint for each antenna has not yet been considered either in relay networks or in the traditional multiple access systems, to the best of the authors' knowledge.

In DF relay networks, nevertheless, a total power constraint is also included when there exists a central control unit (CCU). Although the CCU cannot allocate the power to each relay from a common power pool, it can still determine how much power each relay will spend within its own power constraint.[1]

In this paper, we derive the optimal training for both the maximum likelihood (ML) channel estimation and the MMSE

[1]Consider the scenario where one source needs several relays to help forwarding the message to the destination due to certain reasons. It is then natural for the source to afford the power consumed at all relays. It is also reasonable that the source has it own budget on how much power it can afford. All these facts justify our introduction of a total power constraint onto all the distributed relays. Note that the total power constraint is a unique property resulting from the relaying nature, which does not exist in multiple-access systems.
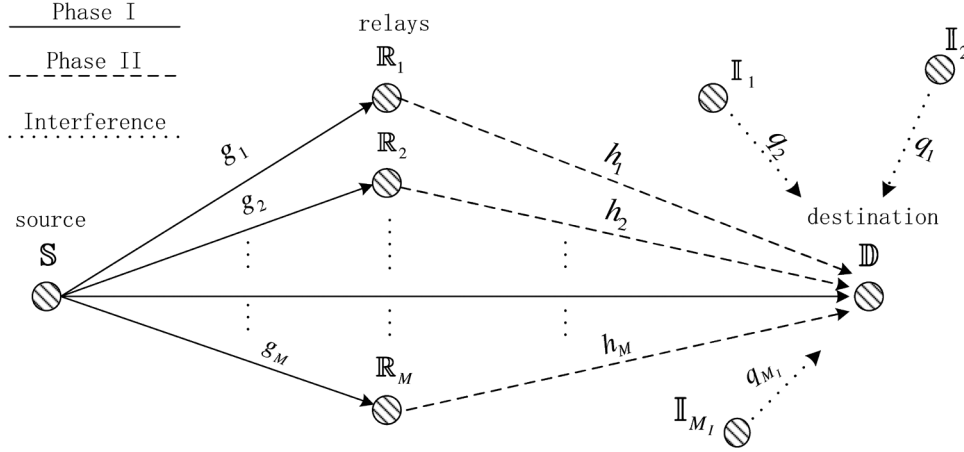
Fig. 1. Wireless relay networks with one source, one destination, M relays, and M_I interferences.

channel estimation based on the knowledge of the interference covariance matrix and with individual and overall power constraints. For ML channel estimation, a multilevel waterfilling type solution is obtained by using majorization theory [20]. However, the problem of MMSE channel estimation turns out to be nonconvex and is hard to solve. We instead consider three special yet reasonable scenarios: 1) white noise and correlated channels; 2) white noise and uncorrelated channels; and 3) equal power constraint and independent and identically distributed (i.i.d.) channels. The optimization in the first scenario can be converted into a semidefinite programming (SDP) problem and can be efficiently solved by SDP tools. The solution to the second scenario has waterfilling structure but has both the ground and the ceiling restrictions. Due to this specific physical meaning, we name this new structure *cave-filling*. The solution to the third scenario is shown to be similar to the one in [18] and [19], after some tricky reformulation.

The rest of this paper is organized as follows. Section II provides the system model of DF-based relay networks. Sections III and IV present the ML channel estimation and the MMSE channel estimation as well as their respective optimal training design. Section V displays simulation results to corroborate the proposed studies. Conclusions are drawn in Section VI.

Notations: Vectors and matrices are boldface small and capital letters, respectively; the transpose, Hermitian, and inverse of the matrix $\mathbf{A}$ are denoted by $\mathbf{A}^T$, $\mathbf{A}^H$, and $\mathbf{A}^{-1}$, respectively; $\mathrm{tr}(\mathbf{A})$ is the trace of $\mathbf{A}$ and $[\mathbf{A}]_{ij}$ is the (i,j)th entry of $\mathbf{A}$; $\mathrm{diag}\{\mathbf{a}\}$ denotes a diagonal matrix with the diagonal element constructed from $\mathbf{a}$; $\mathbf{d}(\mathbf{A})$ and $\boldsymbol{\lambda}(\mathbf{A})$ denote the vectors formed by the diagonal elements and the eigenvalues of $\mathbf{A}$, both arranged in nondecreasing order; $\mathbf{x} \preceq \mathbf{y}$ implies the element-wise inequality for vectors $\mathbf{x}$ and $\mathbf{y}$; $\mathbf{A} \preceq \mathbf{B}$ means that matrix $(\mathbf{B} - \mathbf{A})$ is positive semidefinite; $\mathbf{I}$ is the identity matrix; $\mathrm{E}\{\cdot\}$ denotes the statistical expectation; and the imaginary unit is $\mathrm{j} = \sqrt{-1}$.

II. SYSTEM MODEL OF DF RELAY NETWORKS

Consider a wireless network with M randomly placed relay nodes $\mathbb{R}_i$, $i = 1, \ldots, M$; one source node $\mathbb{S}$; one destination

node $\mathbb{D}$; and M_I interfering nodes $\mathbb{I}_j$, $j = 1, \ldots, M_I$ operating in the same frequency band, as shown in Fig. 1. Every node has only a single antenna that cannot transmit and receive simultaneously. The channel between any two nodes is assumed quasi-stationary Rayleigh flat fading in that it is constant within one frame but may vary from frame to frame. Denote the channel from $\mathbb{S}$ to $\mathbb{R}_i$ as g_i, from $\mathbb{R}_i$ to $\mathbb{D}$ as h_i, from $\mathbb{I}_j$ to $\mathbb{R}_i$ as f_{ji}, and from $\mathbb{I}_j$ to $\mathbb{D}$ as q_j, respectively,[2] namely, $g_i \in \mathcal{CN}(0, \sigma_{g_i})$, $h_i \in \mathcal{CN}(0, \sigma_{h_i})$, $f_{ji} \in \mathcal{CN}(0, \sigma_{f_{ji}})$, and $q_j \in \mathcal{CN}(0, \sigma_{q_j})$. We assume perfect synchronization among $\mathbb{S}$, $\mathbb{R}_i$, and $\mathbb{D}$. However, no synchronization assumption is made for interfering nodes, and only the statistics of the interference are known at $\mathbb{R}_i$s and $\mathbb{D}$.

The training is accomplished by the following two phases, each containing N consecutive time slots. For Phase I, the transmitter broadcasts the training signal $\mathbf{s}$ to $\mathbb{R}_i$s and $\mathbb{D}$. The received signals at $\mathbb{R}_i$ are expressed as

$$\mathbf{r}_i = g_i \mathbf{s} + \sum_{j=1}^{M_I} f_{ji} \mathbf{s}_{j1} + \mathbf{n}_{ri} \qquad (1)$$

where $\mathbf{s}_{j1}$ is the equivalent baseband signal from $\mathbb{I}_j$ during Phase I and $\mathbf{n}_{ri}$ is the white complex Gaussian noise at the ith relay. During Phase II, $\mathbb{R}_i$ sends out the training signal[3] $\mathbf{s}_i$ of length N and $\mathbb{D}$ receives

$$\mathbf{y} = \underbrace{[\mathbf{s}_1 \, \mathbf{s}_2 \, \ldots \, \mathbf{s}_M]}_{\mathbf{C}} \underbrace{\begin{bmatrix} h_1 \\ \vdots \\ h_M \end{bmatrix}}_{\mathbf{h}} + \underbrace{\sum_{j=1}^{M_I} q_j \mathbf{s}_{j2} + \mathbf{n}_{d2}}_{\mathbf{n}_d} = \mathbf{C}\mathbf{h} + \mathbf{n}_d$$

$$(2)$$

where $\mathbf{s}_{j2}$ is the signal from $\mathbb{I}_j$ during Phase II and $\mathbf{n}_{d2} \in \mathcal{CN}(0, N_0\mathbf{I})$ represents the complex white Gaussian

---

[2]Note that the interference, if any, affects both the relays and the destination, which is a highly undesired scenario.

[3]The relays do not need to decode the training signal $\mathbf{s}$ of Phase I but rather send new training signals to the destination.

noise vector at $\mathbb{D}$. The equivalent colored noise $\mathbf{n}_d$ has the covariance

$$\mathbf{R}_n = \mathrm{E}\left\{\mathbf{n}_d\mathbf{n}_d^H\right\} = N_0\mathbf{I} + \mathrm{E}\left\{\left(\sum_{j=1}^{M_I} q_j\mathbf{s}_{j2}\right)\left(\sum_{j=1}^{M_I} q_j\mathbf{s}_{j2}\right)^H\right\} \tag{3}$$

which is assumed known to the destination.

The task of the channel estimation includes estimating each g_i at $\mathbb{R}_i$ and estimating all h_i at $\mathbb{D}$. The former can be carried out using the same algorithm as in the traditional single-input single-output (SISO) system. We omit details for brevity. In the remainder of this paper, we will only focus on estimating h_i. Meanwhile, $N \geq M$ is required since there are M unknown channels to be estimated. Assume, during the training process, $\mathbb{R}_i$ can maximally provide the power of p_i. Then the individual power constraint of $\mathbb{R}_i$ could be expressed as

$$[\mathbf{C}^H\mathbf{C}]_{ii} \leq p_i. \tag{4}$$

To offer a more general discussion at this point, we assume that there exists CCU, and the overall training power consumed from relays is limited by P, namely

$$\mathrm{tr}(\mathbf{C}^H\mathbf{C}) \leq P. \tag{5}$$

Note that CCU in a distributed relay network cannot allocate power to each relay from a common power pool but rather control the power level of each relay within its individual power constraint to meet certain purpose. There are two degenerated cases. First, if $P \geq \sum_{i=1}^{M} p_i$, the total power constraint is redundant. Secondly, if $P \leq \min_i p_i$, all the individual constraints are redundant. In the following, we assume that $\min_i p_i < P < \sum_{i=1}^{M} p_i$.

Remark: The model here also applies to the multiple-access system if only the individual power constraint is imposed. However, whether there can be a total power constraint should be based on some reasonable assumption. The related discussion is out of the scope of this paper and will be omitted here.

III. MAXIMUM LIKELIHOOD BASED CHANNEL ESTIMATION

A. Problem Formulation

The ML estimation considers the *deterministic* channel, and the channel $\mathbf{h}$ should be estimated from

$$\hat{\mathbf{h}}_{ML} = \left(\mathbf{R}_n^{-\frac{1}{2}}\mathbf{C}\right)^\dagger \mathbf{R}_n^{-\frac{1}{2}}\mathbf{y}$$
$$= \left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{y} \tag{6}$$

with the error covariance matrix

$$\mathrm{E}\left\{(\hat{\mathbf{h}}_{ML} - \mathbf{h})^H(\hat{\mathbf{h}}_{ML} - \mathbf{h})\right\} = \left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}. \tag{7}$$

The mean square error (MSE) is then $\mathrm{tr}((\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C})^{-1})$ and the optimal $\mathbf{C}$ can be found by solving the following constrained optimization problem P1:

$$\begin{aligned} \text{P1}: \quad &\min_{\mathbf{C}} \quad \mathrm{tr}\left(\left(\mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\right) \\ &\text{s.t.} \quad [\mathbf{C}^H\mathbf{C}]_{ii} \leq p_i, \quad i = 1, \ldots, M \\ &\quad\quad\quad \mathrm{tr}(\mathbf{C}^H\mathbf{C}) \leq P. \end{aligned} \tag{8}$$

Without loss of generality, we assume p_is are arranged in non-decreasing order and define $\mathbf{p} = [p_1, p_2, \ldots, p_M]^T$.

Before we proceed, we give several definitions in majorization theory. More results on majorization theory can be found in Appendix I and [20].

Definition 1: For any $\mathbf{u} = [u_1, u_2, \ldots, u_n]^T \in \mathbb{R}^n$, let $u_{(i)}$ denote the reordering of the components of $\mathbf{u}$ such that

$$u_{(1)} \leq \ldots \leq u_{(n)}. \tag{9}$$

Definition 2: For any two $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, we say $\mathbf{u}$ is majorized by $\mathbf{v}$ (or $\mathbf{v}$ majorizes $\mathbf{u}$) and write $\mathbf{u} \prec \mathbf{v}$ if

$$\sum_{i=1}^{k} u_{(i)} \geq \sum_{i=1}^{k} v_{(i)}, \quad 1 \leq k \leq n \tag{10}$$

$$\sum_{i=1}^{n} u_{(i)} = \sum_{i=1}^{n} v_{(i)}. \tag{11}$$

If only (10) holds, we say $\mathbf{u}$ is weakly majorized by $\mathbf{v}$ and write $\mathbf{u} \prec^w \mathbf{v}$ [20, A.2]. Note that $\mathbf{u} \prec \mathbf{v}$ implies $\mathbf{u} \prec^w \mathbf{v}$.

Theorem 1: Define a new problem P2

$$\begin{aligned} \text{P2}: \quad &\min_{\mathbf{D}} \quad \mathrm{tr}\left(\left(\mathbf{D}^H\mathbf{R}_n^{-1}\mathbf{D}\right)^{-1}\right) \\ &\text{s.t.} \quad \mathbf{D}^H\mathbf{D} \text{ is diagonal} \\ &\quad\quad\quad \mathrm{d}(\mathbf{D}^H\mathbf{D}) \succ^w \mathbf{p} \\ &\quad\quad\quad \mathrm{tr}(\mathbf{D}^H\mathbf{D}) \leq P \end{aligned} \tag{12}$$

where $\succ^w$ is defined in Definition 2. Suppose the optimal solution to P2 is $\mathbf{D}_o$. There exists a unitary matrix $\mathbf{U}$ such that the optimal $\mathbf{C}_o$ for P1 can be obtained as $\mathbf{D}_o\mathbf{U}^H$.

Proof: See Appendix II. The way to find $\mathbf{U}$ will be exhibited in Section III-C. ∎

Since $\mathbf{D}^H\mathbf{D}$ is diagonal, we can represent $\mathbf{D}$ as $\mathbf{Q}\mathbf{\Sigma}_D^{1/2}$, where $\mathbf{Q}$ is an $N \times M$ orthonormal matrix and $\mathbf{\Sigma}_D^{1/2}$ is a real diagonal matrix with diagonal element $\sigma_{D,i}^{1/2} \geq 0$. Since the column order of $\mathbf{Q}$ can be changed arbitrarily with the

corresponding interchange of $\sigma_{D,i}^{12}$, we can always assume that $\sigma_{D,i}^{12}$ are arranged in nondecreasing order. Then P2 becomes

$$
\begin{aligned}
\min_{\mathbf{Q},\sigma_{D,i}} \quad & \mathrm{tr}\left(\left(\mathbf{\Sigma}_D^{\frac{1}{2}}\mathbf{Q}^H\mathbf{R}_n^{-1}\mathbf{Q}\mathbf{\Sigma}_D^{\frac{1}{2}}\right)^{-1}\right) \\
\text{s.t.} \quad & \mathbf{Q}^H\mathbf{Q} = \mathbf{I} \\
& \sum_{i=1}^{k}\sigma_{D,i} \le \sum_{i=1}^{k}p_i, \quad k = 1,\dots,M \\
& \sum_{i=1}^{M}\sigma_{D,i} \le P \\
& \sigma_{D,i} \le \sigma_{D,i+1}, \quad \sigma_{D,i} \ge 0.
\end{aligned} \tag{13}
$$

Suppose the eigenvalue decomposition of $\mathbf{R}_n$ is $\mathbf{R}_n = \mathbf{U}_n\mathbf{\Sigma}_n\mathbf{U}_n^H$, where $\mathbf{U}_n$ is an $N \times N$ unitary matrix and $\mathbf{\Sigma}_n = \mathrm{diag}\{\sigma_{n,1},\dots,\sigma_{n,N}\}$ is a diagonal matrix. Since the column order of $\mathbf{U}_n$ can be changed arbitrarily if the diagonal elements in $\mathbf{\Sigma}_n$ are interchanged accordingly, we can always assume that $\sigma_{n,i}$ are arranged in nondecreasing order. Then, we get into the following theorem.

Theorem 2: The optimal $\mathbf{Q}$ to (13) is $\mathbf{U}_n[\mathbf{I}_M, \mathbf{0}_{M,N-M}^T]^T$ and the optimal $\sigma_{D,i}$ can be found from

$$
\begin{aligned}
\min_{\sigma_{D,i}} \quad & \sum_{i=1}^{M}\frac{\sigma_{n,i}}{\sigma_{D,i}} \\
\text{s.t.} \quad & \sum_{i=1}^{k}\sigma_{D,i} \le \sum_{i=1}^{k}p_i, \quad k = 1,\dots,M \\
& \sum_{i=1}^{M}\sigma_{D,i} \le P, \ \sigma_{D,i} \le \sigma_{D,i+1}, \ \sigma_{D,i} \ge 0.
\end{aligned} \tag{14}
$$

Proof: See Appendix III. ∎

We can remove the constraints $\sigma_{D,i} \ge 0$ and $\sigma_{D,i} \le \sigma_{D,i+1}$ since an optimal solution always satisfies them. This point will be made clear later.

B. Convex Optimization via Karush–Kuhn—Tucker (KKT) Conditions

Clearly, (14) is a convex optimization problem with respect to unknown $\sigma_{D,i}$s. Since $p_1 < P < \sum_{i=1}^{M}p_i$, there must exist an integer $k^* \in \{1,2,\dots,M-1\}$, such that $\sum_{i=1}^{k^*}p_i < P$ while $\sum_{i=1}^{k^*+1}p_i \ge P$. Therefore, the constraints $\sum_{i=1}^{k}\sigma_{D,i} \le \sum_{i=1}^{k}p_i$, for $k = k^*+1,\dots,M$, are redundant and can be removed for the time being.

The Lagrangian of the optimization problem is written as

$$
\mathcal{L} = \sum_{i=1}^{M}\frac{\sigma_{n,i}}{\sigma_{D,i}} + \sum_{k=1}^{k^*}\mu_k\left(\sum_{i=1}^{k}\sigma_{D,i} - \sum_{i=1}^{k}p_i\right) \\
+ \nu\left(\sum_{i=1}^{M}\sigma_{D,i} - P\right) \tag{15}
$$

where μ_k and ν are Lagrange multipliers and the KKT conditions are listed as

$$
\begin{aligned}
-\frac{\sigma_{n,k}}{\sigma_{D,k}^2} + \sum_{i=k}^{k^*}\mu_i + \nu = 0, \quad & 1 \le k \le k^* \\
-\frac{\sigma_{n,k}}{\sigma_{D,k}^2} + \nu = 0, \quad & k^*+1 \le k \le M \\
\mu_k\left(\sum_{i=1}^{k}\sigma_{D,i} - \sum_{i=1}^{k}p_i\right) = 0, \quad & 1 \le k \le k^* \\
\nu\left(\sum_{i=1}^{M}\sigma_{D,i} - P\right) = 0, \quad & \mu_k \ge 0, \quad \nu \ge 0.
\end{aligned}
$$

The solutions to the KKT conditions can be found from the following algorithm.

Algorithm 1: Multilevel Waterfilling

---

1) Set $j = 0$, $k_0 = 0$.
2) For each $k_j + 1 \le t \le k^*$, calculate the water level $\sum_{i=k_j+1}^{t}p_i / \sum_{i=k_j+1}^{t}\sqrt{\sigma_{n,i}}$ (hypothesizing patches $k_j + 1$ to t are saturated) and the water level $P - \sum_{i=1}^{k_j}p_i / \sum_{i=k_j+1}^{M}\sqrt{\sigma_{n,i}}$ (hypothesizing patch $k_j + 1$ to M are saturated). If the water level $P - \sum_{i=1}^{k_j}p_i / \sum_{i=k_j+1}^{M}\sqrt{\sigma_{n,i}}$ is the lowest, go to 3). Otherwise, if index t_0 gives the lowest water level, set $k_{j+1} = t_0$ and calculate $\sigma_{D,k} = (\sqrt{\sigma_{n,k}}\sum_{i=k_j+1}^{k_{j+1}}p_i)/(\sum_{i=k_j+1}^{k_{j+1}}\sqrt{\sigma_{n,i}})$ for $k_j + 1 \le k \le k_{j+1}$. If $k_{j+1} = k^*$, then $j := j + 1$ and go to 3); otherwise, $j := j + 1$ and go back to 2).
3) Calculate $\sigma_{D,k} = (\sqrt{\sigma_{n,k}}(P - \sum_{i=1}^{k_j}p_i))/(\sum_{i=k_j+1}^{M}\sqrt{\sigma_{n,i}})$, for $k_j + 1 \le k \le M$.

Proof: See Appendix IV. ∎

Algorithm 1 actually describes a way of multilevel waterfilling, as shown in Fig. 2. Each of the M patches corresponds to one unknown variable $\sigma_{D,k}$ and has a patch-width of $\sqrt{\sigma_{n,k}}$. A total amount of water P is going to be poured into all the patches. As the water is being poured, the water level for all patches will increase simultaneously. However, for each patch, there is a maximum possible water level that is computed from the step 2). Once the water level of a patch reaches its maximum, the water level of this patch will not be enhanced, and the remaining amount of the water can only be poured into the other patches. After all the water is poured, the final water level on the kth patch is the optimal value $\sigma_{D,k}$. Step 2) in fact guarantees that the final water level of the kth patch is always lower than or equal to that of the $k+1$th patch for $k = 1,\dots,M-1$. From Algorithm 1, the water level needs to be calculated $(k^*(k^*+3))/2$ times in the worst case.

C. Algorithm to Find C From D

After obtaining the optimal $\sigma_{D,i}$, we need to construct the original $\mathbf{C}$ for problem P1 such that all the constraints are satisfied. From proof of Theorem 1, we know that the $\sigma_{D,i}$s are the eigenvalues of $\mathbf{C}^H\mathbf{C}$ and $\mathbf{d}(\mathbf{C}^H\mathbf{C}) \le \mathbf{p}$ is required. Therefore,
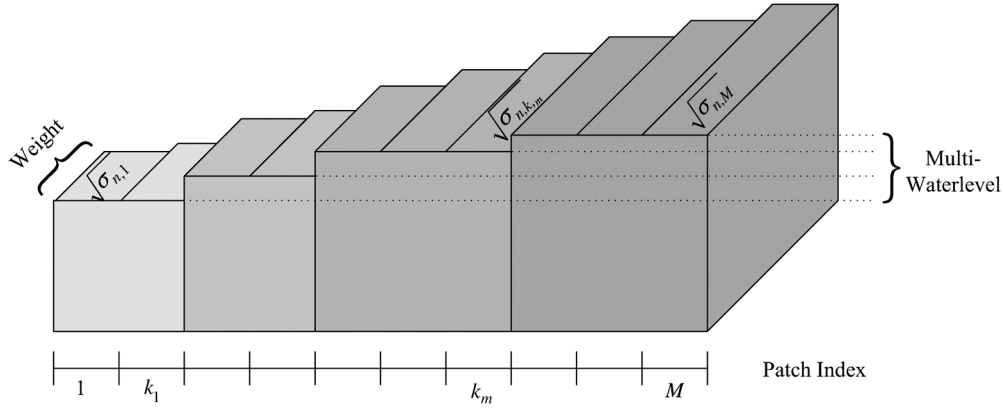
Fig. 2. Illustration on weighted multilevel waterfilling.

as shown in Corollary 1, we first need to find the diagonal values of $\mathbf{C}^H\mathbf{C}$ such that $\mathbf{d}(\mathbf{C}^H\mathbf{C}) \leq \mathbf{p}$ and $\mathbf{d}(\mathbf{C}^H\mathbf{C}) \prec \mathbf{d}(\boldsymbol{\Sigma}_D)$. The solution is obviously not unique. Nonetheless, we here provide a simple way to find one $\mathbf{d}(\mathbf{C}^H\mathbf{C})$. Denote $\mathbf{d}(\mathbf{C}^H\mathbf{C}) = [c_1, c_2, \ldots, c_M]$.

Algorithm 2: Finding Diagonal Elements of $\mathbf{d}(\mathbf{C}^H\mathbf{C})$

---

1) Set $c_i = \sigma_{D,i}$ for all i.
2) For $i = M : -1 : 2$, if $c_i > p_i$, then set $c_i := p_i$ and set $c_{i-1} := c_{i-1} + (c_i - p_i)$.

Proof: See Appendix V. ∎

After obtaining $\mathbf{d}(\mathbf{C}^H\mathbf{C})$, we can find a unitary matrix $\mathbf{U}_C$ via the algorithm provided in [21, Sec. IV-A] (as explained in Lemma 2), such that $\mathbf{U}_C\mathbf{D}^H\mathbf{D}\mathbf{U}_C^H$ has diagonal elements $\mathbf{d}(\mathbf{C}^H\mathbf{C})$. The final optimal $\mathbf{C}$ is $\mathbf{D}\mathbf{U}_C^H$, as has been indicated in Appendix II.

IV. MINIMUM MEAN SQUARE ERROR BASED CHANNEL ESTIMATION

Denote the covariance of $\mathbf{h}$ as $\mathbf{R}_h$, which is assumed known at $\mathbb{D}$. The linear MMSE estimator of $\mathbf{h}$ is expressed as

$$\hat{\mathbf{h}}_{\text{MMSE}} = \mathrm{E}\{\mathbf{h}\mathbf{y}^H\}\left(\mathrm{E}\{\mathbf{y}\mathbf{y}^H\}\right)^{-1}\mathbf{y}$$
$$= \mathbf{R}_h\mathbf{C}^H(\mathbf{C}\mathbf{R}_h\mathbf{C}^H + \mathbf{R}_n)^{-1}\mathbf{y}. \qquad (16)$$

The error covariance of the MMSE estimator is

$$\mathrm{Cov}(\Delta\mathbf{w}) = \left(\mathbf{R}_h^{-1} + \mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1} \qquad (17)$$

and the optimal training should be obtained from

$$\min_{\mathbf{C}} \quad \mathrm{tr}\left(\left(\mathbf{R}_h^{-1} + \mathbf{C}^H\mathbf{R}_n^{-1}\mathbf{C}\right)^{-1}\right)$$
$$\text{s.t.} \quad [\mathbf{C}^H\mathbf{C}]_{ii} \leq p_i, \quad i = 1, \ldots, M$$
$$\mathrm{tr}(\mathbf{C}^H\mathbf{C}) \leq P. \qquad (18)$$

The general solution to (18) is currently unknown. However, (18) can be converted to a convex problem under several special scenarios. For example, when each relay can offer sufficiently large power, the individual power constraints can be removed. Then, the problem becomes the same as the one in the traditional

colocated transmission [19]. In this section, however, we consider three special yet reasonable transmission scenarios under which the convexity of (18) can also be obtained.

A. White Interference and Correlated Channels

The case is valid when the interfering users also transmit the white data sequence[4] or when there is no interference at all. Let $\mathbf{R}_n = \sigma_n\mathbf{I}$ in this case. Then the cost function (18) becomes $\mathrm{tr}((\mathbf{R}_h^{-1} + \sigma_n^{-1}\mathbf{C}^H\mathbf{C})^{-1})$. Denoting $\mathbf{W} = \mathbf{C}^H\mathbf{C}$ and using an auxiliary matrix $\mathbf{T}$, the optimization can be rewritten as [14]

$$\min_{\mathbf{T},\mathbf{W}} \quad \mathrm{tr}(\mathbf{T})$$
$$\text{s.t.} \quad \begin{bmatrix} \mathbf{T} & \mathbf{I} \\ \mathbf{I} & \mathbf{R}_h^{-1} + \sigma_n^{-1}\mathbf{W} \end{bmatrix} \succeq \mathbf{0}$$
$$[\mathbf{W}]_{i,i} \leq p_i$$
$$\mathrm{tr}(\mathbf{W}) \leq P$$
$$\mathbf{W} \succeq \mathbf{0}. \qquad (19)$$

Therefore, (19) is the so called semidefinite programming (SDP) for the variables $\mathbf{T}$ and $\mathbf{W}$. Since both the cost function and the constraints are convex, the SDP formulation could be solved efficiently by interior point methods [22]. The convexity of (19) ensures that its global minimum can be found in polynomial time. The arithmetic complexity of the interior point methods for solving the SDP (19) is $O(M^{6.5}\log(1/\epsilon))$, where $\epsilon > 0$ is a constant to control the algorithm accuracy [22]. After getting $\mathbf{W}$, the training matrix $\mathbf{C}$ can be obtained by the corresponding decomposition.

B. White Interference and Independent Channels

Assuming independent channels is reasonable for relay networks since relays are geographically distributed over a certain region. The optimization problem (18) remains unchanged, except that $\mathbf{R}_h = \mathrm{diag}\{\sigma_{h,1}, \ldots, \sigma_{h,M}\}$ is a diagonal matrix. The following theorem characterizes the optimal solution in this scenario.

Theorem 3: The optimal $\mathbf{C}^H\mathbf{C}$ must be diagonal under the white interference and independent channels.

---

[4]The data from any user, no matter the primary user or the interfering users, are normally white after the interleaving.

Proof: For any matrix $\mathbf{C}$, we can always find a $\tilde{\mathbf{C}}$ such that $\tilde{\mathbf{C}}^H\tilde{\mathbf{C}} = \text{diag}(\mathbf{C}^H\mathbf{C})$, which means that $\tilde{\mathbf{C}}^H\tilde{\mathbf{C}}$ satisfies all the constraints. From the following inequality [23]:

$$\text{tr}(\mathbf{A}^{-1}) \geq \sum_{i=1}^{M} ([\mathbf{A}]_{ii})^{-1} \qquad (20)$$

where $\mathbf{A}$ is an arbitrary $M \times M$ positive definite matrix and the equality holds if and only if $\mathbf{A}$ is diagonal, $\tilde{\mathbf{C}}^H\tilde{\mathbf{C}}$ provides a lower objective value $\text{tr}((\mathbf{R}_h^{-1} + \sigma_n^{-1}\tilde{\mathbf{C}}^H\tilde{\mathbf{C}})^{-1})$ than $\text{tr}((\mathbf{R}_h^{-1} + \sigma_n^{-1}\mathbf{C}^H\mathbf{C})^{-1})$. Therefore, the optimal $\mathbf{C}^H\mathbf{C}$ must be diagonal. ∎

Let $\mathbf{C}^H\mathbf{C} = \text{diag}\{\sigma_{C,1}, \sigma_{C,2}, \ldots, \sigma_{C,M}\}$. The optimization problem (18) can be written as

$$\min_{\sigma_{C,i}} \quad \sum_{i=1}^{M} \frac{1}{\frac{1}{\sigma_{h,i}} + \frac{\sigma_{C,i}}{\sigma_n}}$$
$$\text{s.t.} \quad 0 \leq \sigma_{C,i} \leq p_i, i = 1, \ldots, M$$
$$\sum_{i=1}^{M} \sigma_{C,i} \leq P. \qquad (21)$$

The corresponding Lagrangian is

$$\mathcal{L} = \sum_{i=1}^{M} \frac{1}{\frac{1}{\sigma_{h,i}} + \frac{\sigma_{C,i}}{\sigma_n}} + \lambda\left(\sum_{i=1}^{M} \sigma_{C,i} - P\right)$$
$$+ \sum_{i=1}^{M} \mu_i(\sigma_{C,i} - p_i) - \sum_{i=1}^{M} \nu_i\sigma_{C,i} \quad (22)$$

where λ, μ_i, and ν_i are the corresponding Lagrange multipliers. The KKT conditions are

$$\sum_{i=1}^{M} \sigma_{C,i} \leq P, \qquad \sigma_{C,i} \leq p_i, \quad \sigma_{C,i}, \mu_i, \nu_i, \lambda \geq 0$$
$$\lambda\left(\sum_{i=1}^{M} \sigma_{C,i} - P\right) = 0$$
$$\mu_i(\sigma_{C,i} - p_i) = 0, \quad \nu_i\sigma_{C,i} = 0$$
$$-\frac{1}{\sigma_n\left(\frac{1}{\sigma_{h,i}} + \frac{\sigma_{C,i}}{\sigma_n}\right)^2} + \lambda + \mu_i - \nu_i = 0. \qquad (23)$$

The optimal $\sigma_{C,i}$ is derived as

$$\sigma_{C,i} = \begin{cases} 0, & \lambda \geq \frac{\sigma_{h,i}^2}{\sigma_n} \\ \sqrt{\frac{\sigma_n}{\lambda}} - \frac{\sigma_n}{\sigma_{h,i}}, & \frac{1}{\sigma_n\left(\frac{1}{\sigma_{h,i}} + \frac{p_i}{\sigma_n}\right)^2} < \lambda < \frac{\sigma_{h,i}^2}{\sigma_n} \\ p_i, & \lambda \leq \frac{1}{\sigma_n\left(\frac{1}{\sigma_{h,i}} + \frac{p_i}{\sigma_n}\right)^2} \end{cases} \quad (24)$$

or, more concisely, $\sigma_{C,i} = \min\{\max\{0, \sqrt{(\sigma_n/\lambda)} - (\sigma_n/\sigma_{h,i})\}, p_i\}$.

Proof: See Appendix VI. ∎

Substituting the expression for $\sigma_{C,i}$ into $\sum_{i=1}^{M} \sigma_{C,i} = P$, we obtain

$$\sum_{i=1}^{M} \min\left\{\max\left\{0, \sqrt{\frac{\sigma_n}{\lambda}} - \frac{\sigma_n}{\sigma_{h,i}}\right\}, p_i\right\} = P \qquad (25)$$

from which we can calculate the optimal value of $\sqrt{1/\lambda}$. The left-hand side of (25) is a piecewise-linear nondecreasing function of $1/\lambda$, with breakpoints at $\sigma_n/\sigma_{h,i}^2$ and $\sigma_n((1/\sigma_{h,i}) + (p_i/\sigma_n))^2$, so the equation has a unique solution. This solution also has a waterfilling type structure for the following reasons. We may think of $\gamma = \sqrt{\sigma_n/\lambda}$ as the water level associated with the ith patch, while thinking of $\sigma_n/\sigma_{h,i}$ and $(\sigma_n/\sigma_{h,i}) + p_i$ as the ground level and the ceiling level of patch i, respectively. The patch structure with both ground level and ceiling level is illustrated in Fig. 3. Then we flood the region with water to a depth γ. Note that those patches that have ceiling levels lower than γ will be saturated and no more water will exceed the corresponding ceiling levels. The total amount of water used is then $\sum_{i=1}^{M} \min\{\max\{0, \gamma - (\sigma_n/\sigma_{h,i})\}, p_i\}$. We keep on flooding the patches until we have used a total amount of water equaling to P. The depth of water above patch i is then the optimal value $\sigma_{C,i}$. There exists a difference between this new type of waterfilling and that of the multilevel waterfilling, where in the former we only consider one water level γ during the optimization. Due to its specific physical meaning, we will name the new structure as cave-filling.

Algorithm 3: Cave-filling

1) Sort the ground level according to another index set $[i]$ such that $\sigma_n/\sigma_{h,[i]} \leq, \ldots, \leq \sigma_n/\sigma_{h,[M]}$.
2) Set $k = 1$.
3) Find index M_k such that $(\sigma_n/\sigma_{h,[M_k]}) \leq (\sigma_n/\sigma_{h,k}) + p_k$. Calculate

$$P_k = \sum_{i=1}^{M_k}\left(\frac{\sigma_n}{\sigma_{h,k}} + p_k - \frac{\sigma_n}{\sigma_{h,[i]}}\right)$$
$$- \sum_{i=1}^{k-1}\left(\left(\frac{\sigma_n}{\sigma_{h,k}} + p_k\right) - \left(\frac{\sigma_n}{\sigma_{h,i}} + p_i\right)\right). \quad (26)$$

If $P_k = P$, set $\gamma = (\sigma_n/\sigma_{h,k}) + p_k$. If $P_k > P$, go to 4); If $P_k < P$, then $k := k + 1$ and go back to 3).
4) Apply the traditional waterfilling algorithm over patches k to M with a total power $P - \sum_{i=1}^{k-1} p_i$; namely, calculate γ from

$$\sum_{i=k}^{M} \max\left\{0, \gamma - \frac{\sigma_n}{\sigma_{h,i}}\right\} = P - \sum_{i=1}^{k-1} p_i. \qquad (27)$$

Proof: See Appendix VII. ∎

Corresponding to outer iteration in step 3), the traditional waterfilling in step 4) is referred to as the inner iteration.

C. Equal Power Constraints Under i.i.d. Channels

The assumption of i.i.d. channel is reasonable when distances between different relays and the destination are relatively the same. The assumption of the same maximum power consumption $p = p_i$ is also valid when relays are the same type of mobile terminals.
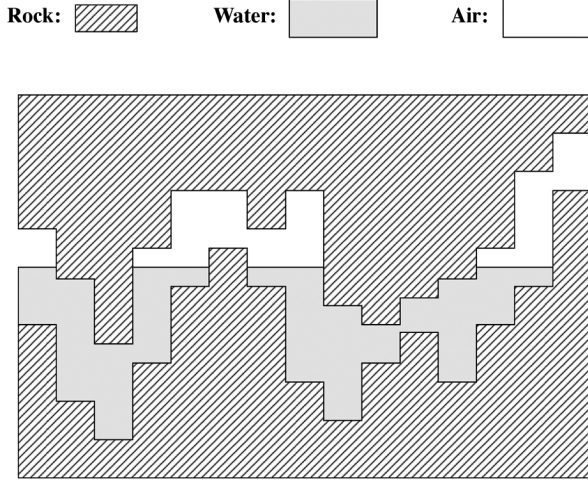
Rock: ▨ **Water:** ▢ **Air:** ▢

Fig. 3.   Illustration on cave-filling with both ground level and ceiling.

Under this circumstance, we may denote $\mathbf{R}_h = \sigma_h \mathbf{I}$, and the optimization is rewritten as

$$\min_{\mathbf{C}} \quad \text{tr}\left(\left(\sigma_h^{-1}\mathbf{I} + \mathbf{C}^H \mathbf{R}_n^{-1} \mathbf{C}\right)^{-1}\right)$$
$$\text{s.t.} \quad [\mathbf{C}^H \mathbf{C}]_{ii} \leq p$$
$$\text{tr}(\mathbf{C}^H \mathbf{C}) \leq P. \tag{28}$$

Theorem 4: The optimal $\mathbf{C}^H \mathbf{C}$ would have equal diagonal values under equal power constraints and i.i.d. channels.

Proof: Denote the singular value decomposition of the optimal $\mathbf{C}$ as $\mathbf{C} = \mathbf{U}_C \mathbf{\Sigma}_C^{1/2} \mathbf{V}_C^H$, where $\mathbf{U}_C$ is an $N \times M$ orthonormal matrix, $\mathbf{V}_C$ is an $M \times M$ unitary matrix, and $\mathbf{\Sigma}_C = \text{diag}\{\sigma_{C,1}, \ldots, \sigma_{C,M}\}$ is a diagonal matrix with nonnegative diagonal elements. Define $\mathbf{F}$ as the $M \times M$ normalized discrete Fourier transform matrix with $[\mathbf{F}]_{ij} = (1/\sqrt{M})e^{-j2\pi(i-1)(j-1)/M}$ and construct a new matrix $\tilde{\mathbf{C}} = \mathbf{C}^H \mathbf{V}_C \mathbf{F}$. Note that $\tilde{\mathbf{C}}^H \tilde{\mathbf{C}} = \mathbf{F}^H \mathbf{\Sigma}_C \mathbf{F}$ is a circulant matrix[5] and, therefore, has equal diagonal elements $\text{tr}(\mathbf{C}^H\mathbf{C})/M$. Meanwhile, the objective function $\text{tr}((\sigma_h^{-1}\mathbf{I} + \tilde{\mathbf{C}}^H \mathbf{R}_n^{-1} \tilde{\mathbf{C}})^{-1})$ remains the same as $\text{tr}((\sigma_h^{-1}\mathbf{I} + \mathbf{C}^H \mathbf{R}_n^{-1} \mathbf{C})^{-1})$. Since $[\tilde{\mathbf{C}}^H \tilde{\mathbf{C}}]_{ii} = \text{tr}(\mathbf{C}^H\mathbf{C})/M \leq \max_i [\mathbf{C}^H\mathbf{C}]_{ii} \leq p$, using $\tilde{\mathbf{C}}$ would satisfy all the individual power constraints. The total power constraint is also satisfied since $\text{tr}(\tilde{\mathbf{C}}^H \tilde{\mathbf{C}}) = \text{tr}(\mathbf{C}^H \mathbf{C})$. So, we can always consider the optimal $\mathbf{C}$ that has equal diagonal elements in $\mathbf{C}^H \mathbf{C}$.   ∎

The optimization is rewritten as

$$\min_{\mathbf{C}} \quad \text{tr}\left(\left(\sigma_h^{-1}\mathbf{I} + \mathbf{C}^H \mathbf{R}_n^{-1} \mathbf{C}\right)^{-1}\right)$$
$$\text{s.t.} \quad \mathbf{C}^H \mathbf{C} \text{ has equal diagonal elements}$$
$$\text{tr}(\mathbf{C}^H \mathbf{C})/M \leq p$$
$$\text{tr}(\mathbf{C}^H \mathbf{C}) \leq P. \tag{29}$$

As we only consider the nondegenerated case with $P < Mp$, we can remove the constraint $\text{tr}(\mathbf{C}^H\mathbf{C})/M < p$. Meanwhile, since we can always find a unitary matrix $\tilde{\mathbf{C}} = \mathbf{C}\mathbf{V}_C\mathbf{F}$ such that $\tilde{\mathbf{C}}^H \tilde{\mathbf{C}}$ has equal diagonal elements and gives the same value

[5]From [24], we know that $\mathbf{X} = \mathbf{F}^H \mathbf{Y} \mathbf{F}$ is a circulant matrix for any diagonal matrix $\mathbf{Y}$.

of objective function, we can first look into the following optimization:

$$\min_{\mathbf{D}} \quad \text{tr}\left(\left(\sigma_h^{-1}\mathbf{I} + \mathbf{D}^H \mathbf{R}_n^{-1} \mathbf{D}\right)^{-1}\right)$$
$$\text{tr}(\mathbf{D}^H \mathbf{D}) \leq P. \tag{30}$$

Now the problem becomes the classical one that has been discussed in [12] and [18], and the solution is $\mathbf{D} = \mathbf{U}_n \mathbf{\Sigma}_D^{1/2} \mathbf{V}_D^H$, where $\mathbf{U}_n$ is the eigenmatrix of $\mathbf{R}_n^{-1}$, $\mathbf{V}_D$ is any $M \times M$ unitary matrix, and $\mathbf{\Sigma}_D^{1/2} = \text{diag}\{\sqrt{\sigma_{D,1}}, \sqrt{\sigma_{D,2}}, \ldots, \sqrt{\sigma_{D,M}}\}$ is a diagonal matrix. If the eigenvalues of $\mathbf{R}_n$ are arranged in nondecreasing order, then the optimal $\sigma_{D,i}$ follows the weighted waterfilling structure

$$\sigma_{D,i} = \max\left\{0, \sqrt{\frac{\sigma_{n,i}}{\nu}} - \frac{\sigma_{n,i}}{\sigma_h}\right\}. \tag{31}$$

The water level $\sqrt{1/\nu}$ should be found from

$$\sum_{i=1}^{M} \max\left\{0, \sqrt{\frac{\sigma_{n,i}}{\nu}} - \frac{\sigma_{n,i}}{\sigma_h}\right\} = P. \tag{32}$$

Finally, the optimal solution $\mathbf{C}$ to the original problem (28) is $\mathbf{C} = \mathbf{U}_n \mathbf{\Sigma}_D^{1/2} \mathbf{F}$.

Corollary 2: If $P \leq \min_i M p_i$, the optimization can be solved similarly as (30).

Proof: Consider a new problem by changing individual power constraints to $[\mathbf{C}^H\mathbf{C}]_{ii} \leq \max_i p_i$ while keeping the total power constraints the same. The new problem should have an optimal objective value less than or equal to that of the original problem. From (30), we know the final optimization solution $\tilde{\mathbf{C}}^H \tilde{\mathbf{C}}$ to this new problem has equal diagonal value P/M. Since $P/M \leq \min_i p_i$, all the individual power constraints are also included in the original individual power constraints. So the optimization to the new and original problem are the same.   ∎

V. SIMULATION RESULTS

In this section, we numerically examine the performance of our proposed channel estimation algorithms as well as the optimal training designs under various scenarios. The signal-to-noise-ratio is defined as $\text{SNR} = P/MN/N_0$ with $N_0 = 1$ (average power over time and spatial index).

The channels h_is are assumed as circularly symmetric complex Gaussian random variables with variances $\sigma_{h,i}$ normalized such that $\sum_{i=1}^{M} \sigma_{h,i} = M$. The channel covariance matrices $\mathbf{R}_h$ have the following structures:

$$[\mathbf{R}_h]_{i,j} = \sqrt{\sigma_{h,i}\sigma_{h,j}} \varepsilon_1^{|i-j|}$$

where $\varepsilon_1 < 1$ is a real scalar that affects the correlation between channels. Interference covariance matrices $\mathbf{R}_n$ in our example have a similar structure as $\mathbf{R}_h$, where a real scalar $\varepsilon_2 < 1$ is used to control the correlation between noise. The average interference power is assumed to be ten times the noise so that $\text{tr}(\mathbf{R}_n)/M = 11N_0$.

The training sequence $\hat{\mathbf{s}}_i$ that is the scalar multiple of the optimal $\mathbf{s}_i$ will be named as the proposed training sequence (Proposed T). Correspondingly, the L2 norm of the optimal $\mathbf{s}_i$ will be referred to as the proposed power allocation (Proposed P).
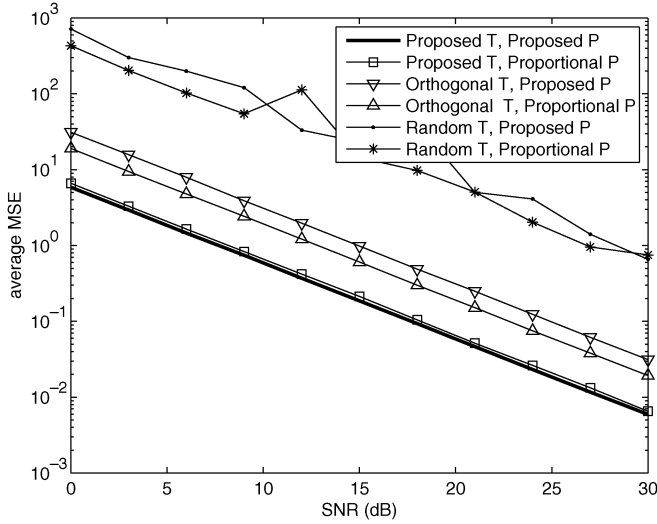
Fig. 4. Comparison between different training and power allocation for ML-based channel estimation, with $\varepsilon_1 = 0.9, \varepsilon_2 = 0.9, M = N = 4$.
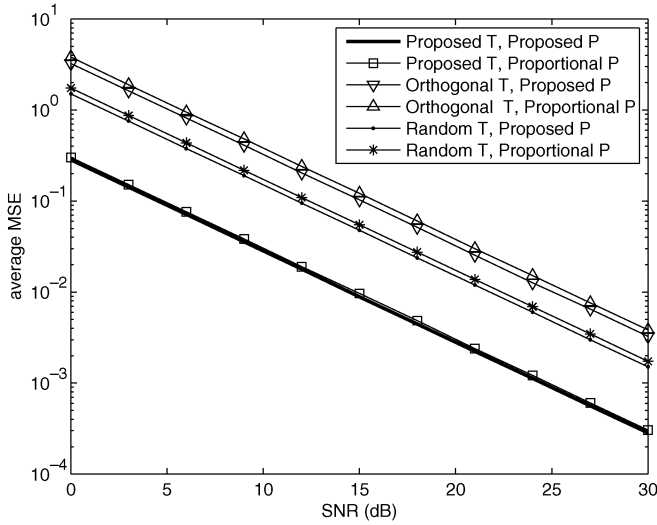

Fig. 5. Comparison between different training and power allocation for ML based channel estimation, with $\varepsilon_1 = 0.9, \varepsilon_2 = 0.9, M = N/2 = 4$.

The proportional power allocation (Proportional P) is defined as $\hat{p}_i = (p_i / \sum_{i=1}^{M} p_i)P$. We mainly compare the proposed training sequence with both orthogonal training (Orthogonal T) and random training (Random T). Therefore, the following six different types of the training scenarios will be examined: "Proposed T, Proposed P," "Proposed T, Proportional P," "Orthogonal T, Proposed P," "Orthogonal T, Proportional P," "Random T, Proposed P," "Random T, Proportional P." For all numerical examples, we use 10 000 Monte Carlo runs.

A. ML Channel Estimation

To exhibit the effect of the correlated channel and the colored interference, we adopt a relatively large ε_1 and ε_2 as $\varepsilon_1 = \varepsilon_2 = 0.9$. In Fig. 4, we display the MSEs of the ML channel estimation versus SNR for different training scenarios where $M = N = 4$. We can see that the proposed training with

the proposed power (the optimal solution) is slightly better than the proposed training with the proportional power. The orthogonal training under both power allocations has more than 6 dB SNR loss compared to the optimal one. The performance of the random training has around 20 dB SNR loss compared to the optimal one and is not stable.[6] since we assume the smallest possible N We then increase N to eight while keeping all other parameters fixed and show different MSEs in Fig. 5. Most observations are the same as those in Fig. 4 except that the performance of the random training becomes more stable and is better than that of the orthogonal training. For orthogonal and random training, although the proportional power allocation gives better performance than the proposed power allocation in Fig. 4, it gives worse performance in Fig. 5.

B. MMSE Channel Estimation

1) White Interference and Correlated Channel: To exhibit the effect of the correlated channel, we here adopt a relatively large ε_1 as $\varepsilon_1 = 0.9$. The convex optimization is conducted by the SDP tool SeDuMi v1.1 [25]. The MSE of different algorithms as a function SNR is shown in Fig. 6 with $M = N = 4$. We find that the proposed training sequence under the proposed power allocation gives the best performance. Interestingly, the proposed training sequence with the proportional power allocation is always parallel to the optimal one but has 1 dB SNR loss. Meanwhile, the orthogonal training with the proposed power allocation performs worse at lower SNR region but performs close to the optimal one at high SNR region. This is reasonable and agrees with the intuition that under the white interference and with high SNR, MMSE estimation will become similar to ML estimation whose optimal training sequence should be orthogonal training. Nonetheless, with proportional power allocation, the orthogonal training still has a 1-dB loss at high SNR. For random training, the one with the proposed power allocation is 2 dB better than the one with proportional training. However, both of them perform much worse than the proposed training under the proposed power allocation.

2) White Interference and Uncorrelated Channel: In this case, the proposed training is orthogonal. Therefore we only compare it with random training. The MSE performance of the four different training scenarios are shown in Fig. 7 with $M = N = 4$. Similarly, the proposed training with proportional power performs 1 dB worse than the proposed training with the proposed power allocation, and the random training suffers from a larger SNR loss.

3) Equal Power Constraints and i.i.d Channel: To exhibit the effect of the colored interference, we choose a relatively large ε_2 as $\varepsilon_2 = 0.9$. In this case, we find that the proportional power allocation is the same as the proposed power allocation. We thus only compare different schemes under the proposed power allocation. The MSEs are shown in Fig. 8. It is seen that the orthogonal training incurs a 2-dB loss over the optimal training, while the random training suffers from a significant loss.

From simulations in the MMSE case, we find that: 1) the proposed training sequence always performs better than other training sequences with the same power allocation and 2) the

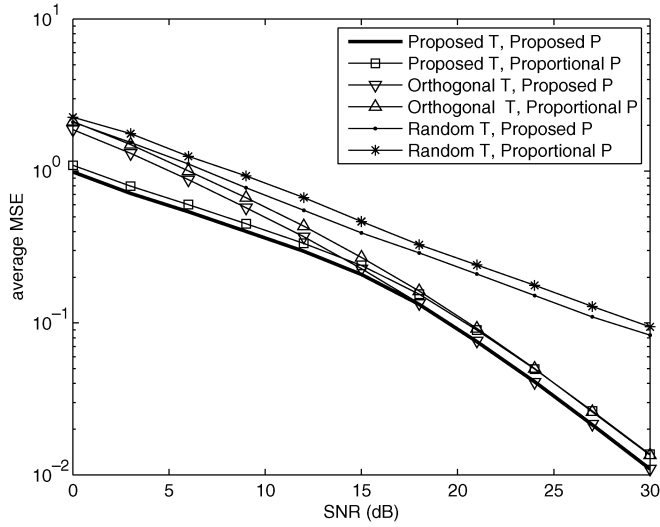[6]The same phenomenon has been observed in [14].

Fig. 6. Comparison between different training and power allocation for MMSE-based channel estimation, with $\varepsilon_1 = 0.9, \varepsilon_2 = 0$, under $M = N = 4$.

Fig. 8. Comparison between different training and power allocation for MMSE-based channel estimation, with $\varepsilon_1 = 0, \varepsilon_2 = 0.9$, under $M = N = 4$.
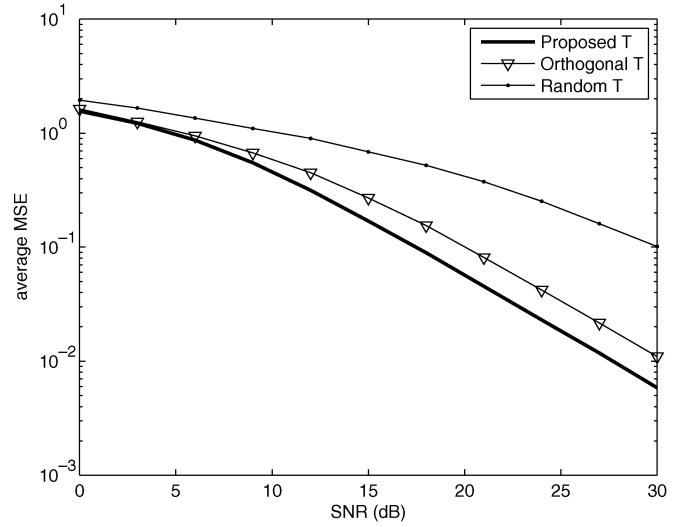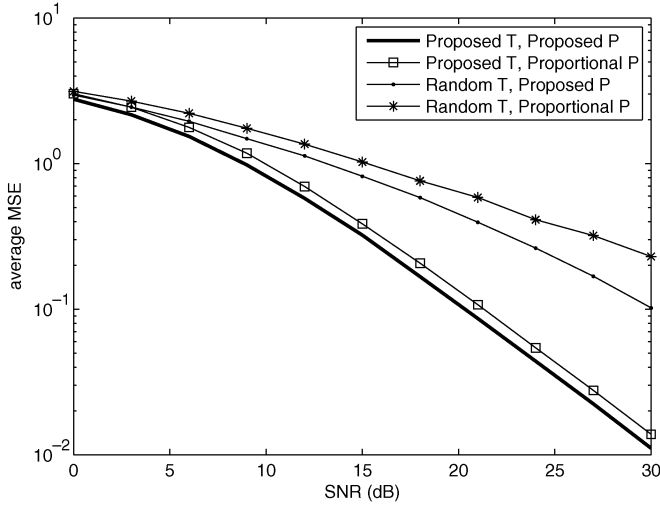

Fig. 7. Comparison between different training and power allocation for MMSE-based channel estimation, with $\varepsilon_1 = 0, \varepsilon_2 = 0$, under $M = N = 4$.
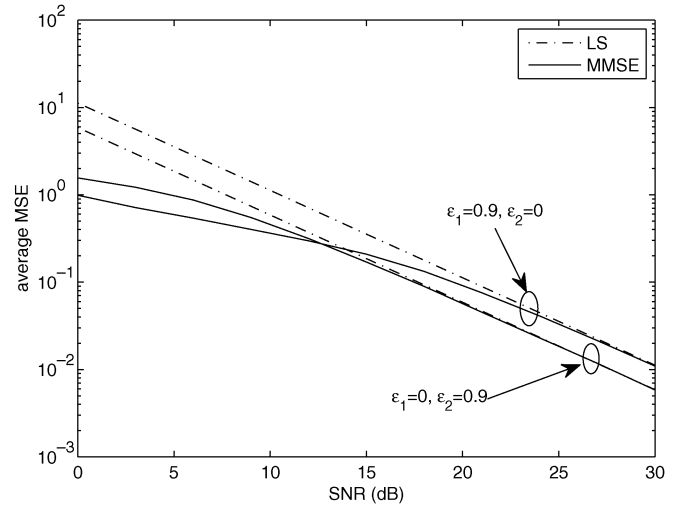

Fig. 9. Comparison between ML channel estimation and MMSE channel estimation for case 1 and case 3, respectively, under $M = N = 4$.

proposed power allocation always performs better than the proportional power allocation under the same training sequence.

C. ML Channel Estimation Versus MMSE Channel Estimation

Finally, we compare the ML channel estimation and the MMSE channel estimation with $M = N = 2$. We consider two cases: $\varepsilon_1 = 0.9$, $\varepsilon_2 = 0$ (case 1 in Section IV) and $\varepsilon_1 = 0, \varepsilon_2 = 0.9$ (case 3 in Section IV). The MSEs of different algorithms as a function of SNR are shown in Fig. 9. In both cases, the MMSE estimator outperforms the ML estimator in lower SNR, while the two estimators have nearly the same performance at higher SNR. This agrees with the phenomenon in the traditional SISO or MISO channel estimation [17].

VI. CONCLUSION

In this paper, we studied the training based channel estimation in relay networks using DF strategy. The major challenge is that there exists an individual power constraint for each relay node as well as a total power constraint over the whole network.

Both ML and MMSE estimators have been investigated. The ML-based channel estimation was solved thoroughly by using a multilevel waterfilling algorithm. For MMSE estimation, however, the general problem turns out to be nonconvex and is difficult to solve. We instead consider three special yet reasonable scenarios, all of which can be converted to convex optimization problems, and the last two scenarios have the waterfilling type solutions. Meanwhile, we name a new type of waterfilling structure as *cave-filling* where there are both grounds and ceilings for water patches. Numerical examples have been provided from which we find that the proposed training and the proposed power allocation are both important to achieve the best channel estimation.

APPENDIX I
MAJORIZATION THEORY

Majorization theory has been used to convert some matrix-valued nonconvex problems into scalar-valued convex ones in

[26]. Here, we briefly introduce some basic results on majorization theory [20].

Lemma 1 [20, 9.B.1]: For any $n \times n$ Hermitian matrix $\mathbf{A}$, there is $\mathbf{d}(\mathbf{A}) \prec \boldsymbol{\lambda}(\mathbf{A})$.

Lemma 2 [20, 9.B.1]: For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ satisfying $\mathbf{u} \prec \mathbf{v}$, there exists a real symmetric matrix $\mathbf{A}$ whose eigenvalues are $\mathbf{v}$ and diagonal elements are $\mathbf{u}$.

The matrix $\mathbf{A}$ can be eigendecomposed as $\mathbf{U}_A \text{diag}\{\mathbf{v}\}\mathbf{U}_A^H$. A practical algorithm to find $\mathbf{U}_A$ was proposed in [21, Sec. IV-A].

Lemma 3 [20, 5.A.9.a]: For any $\mathbf{u} \prec^w \mathbf{v}$, there must exist an $\tilde{\mathbf{u}} \le \mathbf{u}$ such that $\tilde{\mathbf{u}} \prec \mathbf{v}$.

Corollary 1: For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ satisfying $\mathbf{u} \prec^w \mathbf{v}$, there exists a real symmetric matrix $\mathbf{A}$ whose eigenvalues are $\mathbf{v}$ and diagonal elements $\mathbf{d}(\mathbf{A}) \le \mathbf{u}$.

Proof: Straightforwardly from Lemma 2 and Lemma 3. ∎

APPENDIX II
PROOF OF THEOREM 1

We need to first proof the equivalence between P1 and P2. It suffices to show that for any feasible point $\mathbf{C}$ in P1, there is a corresponding feasible point $\mathbf{D}$ in P2 which gives the same objective value, and vice versa. The proof needs some basic knowledge of majorization theory, which has been briefly introduced in Appendix I.

1) $P1 \longrightarrow P2$: Let $\mathbf{C}$ be any matrix in the space restricted by the constraints in P1, and the corresponding objective value is $\text{tr}((\mathbf{C}^H \mathbf{R}_n^{-1} \mathbf{C})^{-1})$. From the constraints of P1, we get $\mathbf{d}(\mathbf{C}^H \mathbf{C}) \le \mathbf{p}$ and therefore $\mathbf{d}(\mathbf{C}^H \mathbf{C}) \succ^w \mathbf{p}$ by definition. From Lemma 1 in the Appendix, we know $\boldsymbol{\lambda}(\mathbf{C}^H \mathbf{C}) \succ \mathbf{d}(\mathbf{C}^H \mathbf{C})$ so $\boldsymbol{\lambda}(\mathbf{C}^H \mathbf{C}) \succ^w \mathbf{p}$. Suppose the eigendecomposition of $\mathbf{C}^H \mathbf{C}$ is $\mathbf{C}^H \mathbf{C} = \mathbf{U}_C \boldsymbol{\Sigma}_C \mathbf{U}_C^H$ and define $\mathbf{D} = \mathbf{C} \mathbf{U}_C$. Then, we have $\text{tr}((\mathbf{D}^H \mathbf{R}_n^{-1} \mathbf{D})^{-1}) = \text{tr}((\mathbf{C}^H \mathbf{R}_n^{-1} \mathbf{C})^{-1})$. Since $\mathbf{D}^H \mathbf{D} = \boldsymbol{\Sigma}_C$ is a diagonal matrix, there is $\mathbf{d}(\mathbf{D}^H \mathbf{D}) = \boldsymbol{\lambda}(\mathbf{D}^H \mathbf{D}) = \boldsymbol{\lambda}(\mathbf{C}^H \mathbf{C}) \succ^w \mathbf{p}$. Moreover, $\text{tr}(\mathbf{D}^H \mathbf{D}) = \text{tr}(\mathbf{C}^H \mathbf{C}) \le P$. Therefore, for any feasible solution $\mathbf{C}$ to P1, $\mathbf{D} = \mathbf{C} \mathbf{U}_C$ is always a feasible point in P2 with the same objective value.

2) $P2 \longrightarrow P1$: Let $\mathbf{D}$ be any feasible solution to P2. Since $\mathbf{D}^H \mathbf{D}$ is diagonal, then $\mathbf{d}(\mathbf{D}^H \mathbf{D}) = \boldsymbol{\lambda}(\mathbf{D}^H \mathbf{D}) \succ^w \mathbf{p}$. From Corollary 1 in Appendix I, we know there exists a real symmetric matrix $\mathbf{A}$ such that $\mathbf{d}(\mathbf{A}) \le \mathbf{p}$ and $\boldsymbol{\lambda}(\mathbf{A}) = \boldsymbol{\lambda}(\mathbf{D}^H \mathbf{D})$. Therefore, $\mathbf{A}$ is positive semidefinite and can be expressed as $\mathbf{U} \mathbf{D}^H \mathbf{D} \mathbf{U}^H$ for some unitary matrix $\mathbf{U}$. Define $\mathbf{C} = \mathbf{D} \mathbf{U}^H$. Note that $\mathbf{d}(\mathbf{C}^H \mathbf{C}) = \mathbf{d}(\mathbf{A}) \le \mathbf{p}$, $\text{tr}(\mathbf{C}^H \mathbf{C}) = \text{tr}(\mathbf{D}^H \mathbf{D}) \le P$, and $\text{tr}((\mathbf{C}^H \mathbf{R}_n^{-1} \mathbf{C})^{-1}) = \text{tr}((\mathbf{D}^H \mathbf{R}_n^{-1} \mathbf{D})^{-1})$. Therefore, for any $\mathbf{D}$ in P2, there is also a corresponding feasible point in P1 with the same objective value.

Theorem 1 is implicitly proved from the above proof.

APPENDIX III
PROOF OF THEOREM 2

We first prove that the optimal $\boldsymbol{\Sigma}_D^{1/2} \mathbf{Q}^H \mathbf{R}_n^{-1} \mathbf{Q} \boldsymbol{\Sigma}_D^{1/2}$ must be a diagonal matrix. Note that the optimization can be separately

conducted for $\mathbf{Q}$ and $\sigma_{D,i}$. The objective function can be equivalently written as

$$\text{tr}\left(\left(\boldsymbol{\Sigma}_D^{\frac{1}{2}} \mathbf{Q}^H \mathbf{R}_n^{-1} \mathbf{Q} \boldsymbol{\Sigma}_D^{\frac{1}{2}}\right)^{-1}\right) = \text{tr}\left(\left(\mathbf{Q}^H \mathbf{R}_n^{-1} \mathbf{Q}\right)^{-1} \boldsymbol{\Sigma}_D^{-1}\right). \tag{33}$$

Suppose the eigenvalues of $\mathbf{Q}^H \mathbf{R}_n^{-1} \mathbf{Q}$ are λ_i, $i = 1, \dots, M$, which are arranged in nondecreasing order. From [27, Eq. (4)], we know

$$\text{tr}\left(\left(\mathbf{Q}^H \mathbf{R}_n^{-1} \mathbf{Q}\right)^{-1} \boldsymbol{\Sigma}_D^{-1}\right) \ge \sum_{i=1}^{M} \frac{1}{\lambda_{M-i+1} \sigma_{D,i}} \ge \sum_{i=1}^{M} \frac{\sigma_{n,i}}{\sigma_{D,i}} \tag{34}$$

where the second inequality comes from [28, Th. 10, pp. 209] and the property that $\sigma_{n,i}$s are arranged in nondecreasing order is also utilized here. The first equality holds when the eigenmatrix of $\mathbf{Q}^H \mathbf{R}_n^{-1} \mathbf{Q}$ is an appropriate permutation matrix. Clearly the lower bound of the objective function is $\sum_{i=1}^{M}(\sigma_{n,i}/\sigma_{D,i})$ and is achieved when $\mathbf{Q} = \mathbf{U}_n[\mathbf{I}_M, \mathbf{0}_{M,N-M}^T]^T$. Note that this $\mathbf{Q}$ is derived when we assume that $\sigma_{n,i}$ and $\sigma_{D,i}$ are arranged in nondecreasing orders. Otherwise, $\mathbf{Q}$ should be $\mathbf{U}_n$ left-multiplied by some appropriate permutation matrix.

Remark: This structure of $\mathbf{Q}$ tells that the optimal training should apply all energy on the eigenmodes that correspond to the smallest interference levels, i.e., the smallest $\sigma_{n,i}$.

APPENDIX IV
PROOF OF ALGORITHM 1

First, it is observed that $\sum_{i=1}^{M} \sigma_{D,i} = P$ must hold at the optimal point. Otherwise, $\nu = 0$ and $-(\sigma_{n,k}/\sigma_{D,k}^2) + \nu = 0$ cannot hold for $k^* + 1 \le k \le M$. Without loss of generality, suppose that at the optimal point, only m out of k^* μ_ks are nonzero, or the equality of the corresponding constraint holds. Denote these m μ_ks as μ_{k_i}, $i = 1, \dots, m$ with $k_1 < k_2 < \dots < k_m$. The assumption indicates $\mu_k = 0$ for $1 \le k < k_1$. Then,

$$-\frac{\sigma_{n,k}}{\sigma_{D,k}^2} + \sum_{i=1}^{m} \mu_{k_i} + \nu = 0, \quad 1 \le k \le k_1 \tag{35}$$

$$\sum_{i=1}^{k_1} \sigma_{D,i} = \sum_{i=1}^{k_1} p_i. \tag{36}$$

Define $\nu_1 = \sum_{i=1}^{m} \mu_{k_i} + \nu$; we have $\sigma_{D,k} = \sqrt{\sigma_{n,k}/\nu_1}$ for $1 \le k \le k_1$. This is exactly the weighted waterfilling by considering $\sqrt{1/\nu_1}$ as the water level and $\sqrt{\sigma_{n,k}}$ as weight for patch k, $1 \le k \le k_1$. Note that the waterfilling here is different from the traditional one [26] in that the water patches here have zero bottom level for all k. Therefore, the water level $\sqrt{1/\nu_1}$ can be explicitly calculated as $\sum_{i=1}^{k_1} p_i / \sum_{i=1}^{k_1} \sqrt{\sigma_{n,i}}$. In fact, when we pour the water into different patches, the water quantity for patch $1 \le k \le k_1$ increases while the ratio

$$\sigma_{D,1} : \dots : \sigma_{D,k_1} = \sqrt{\sigma_{n,1}} : \dots : \sqrt{\sigma_{n,k_1}} \tag{37}$$

is kept until the overall water quantity reaches $\sum_{i=1}^{k_1} p_i$. Obviously, this ratio indicates that $\sigma_{D,k_1} \ge \dots \ge \sigma_{D,1}$.

Next, we consider $\mu_{k_2}, \mu_{k_3}, \ldots, \mu_{k_m}$ in a similar way. When it comes to j, $2 \leq j \leq m$, we need to solve

$$-\frac{\sigma_{n,k}}{\sigma_{D,k}^2} + \sum_{i=j}^{m} \mu_{k_i} + \nu = 0, \quad k_{j-1} < k \leq k_j \quad (38)$$

$$\sum_{i=1}^{k_j} \sigma_{D,i} = \sum_{i=1}^{k_j} p_i. \quad (39)$$

As (39) holds for all $1 \leq j' < j$, (39) is equivalent to

$$\sum_{i=k_{j-1}+1}^{k_j} \sigma_{D,i} = \sum_{i=k_{j-1}+1}^{k_j} p_i. \quad (40)$$

Define $\nu_j = \sum_{i=j}^{m} \mu_{k_i} + \nu$; there is $\sigma_{D,k} = \sqrt{\sigma_{n,k}/\nu_j}$ for $k_{j-1} < k \leq k_j$ and the corresponding water level is $\sqrt{1/\nu_j}$. At this point, we see that multiple water levels coexist for the proposed algorithm. For the same reason, $\sigma_{D,k}$ is in nondecreasing order for $k_{j-1} < k \leq k_j$. Moreover, since $\nu_j = \nu_{j-1} - \mu_{k_{j-1}} \leq \nu_{j-1}$, the water level $\sqrt{1/\nu_j}$ is also arranged in nondecreasing order. Considering the fact that $\sqrt{\sigma_{n,i}}$ is arranged in nondecreasing order, we know that the optimal $\sigma_{D,k_{j-1}+1}$ is greater than or equal to $\sigma_{D,k_{j-1}}$. Therefore, $\sigma_{D,i}$, $i = 1, \ldots, k_m$, should be in nondecreasing order. Meanwhile, the water level $\sqrt{1/\nu_j}$ can be explicitly calculated as $\sum_{i=k_{j-1}+1}^{k_j} p_i / \sum_{i=k_{j-1}+1}^{k_j} \sqrt{\sigma_{n,i}}$.

Lastly, we have

$$-\frac{\sigma_{n,k}}{\sigma_{D,k}^2} + \nu = 0, \quad k_m < k \leq M, \quad (41)$$

$$\sum_{i=k_m+1}^{M} \sigma_{D,i} = P - \sum_{i=1}^{k_m} p_i. \quad (42)$$

The corresponding water level is $\sqrt{1/\nu} = (P - \sum_{i=1}^{k_m} p_i) / \sum_{i=k_m+1}^{M} \sqrt{\sigma_{n,i}}$ and $\sigma_{D,k} = \sqrt{\sigma_{n,k}/\nu}$ for $k_m < k \leq M$. Similarly, $\sigma_{D,k}$ should be in nondecreasing order for $k_m < k \leq M$ and $\sigma_{D,k_m+1} \geq \sigma_{D,k_m}$.

The above discussion not only provides the way to design the algorithm but also confirms the validity of omitting the constraints $\sigma_{D,i} \leq \sigma_{D,i+1}$ and $\sigma_{D,i} \geq 0$ in the first place. The solution structure follows a weighted multilevel waterfilling with multiple water levels at $\{\sqrt{1/\nu_j}, j = 1, \ldots, m, \sqrt{1/\nu}\}$ and the weight for the ith patch is $\sqrt{\sigma_{n,i}}$. The illustration of the proposed weighted multilevel waterfilling is given in Fig. 2, where it is seen that different patches may have different water level and different weight. The area of the cross-section, which is weight $\times$ water-level, will be the power that is poured into this specific patch. The cutting point k_j, $j = 1, \ldots, m$, can be obtained from the testing, which is given by step 2) in Algorithm 1.

APPENDIX V
PROOF OF ALGORITHM 2

1) Proof of $\mathbf{d}(\mathbf{C}^H \mathbf{C}) \prec \mathbf{d}(\mathbf{\Sigma}_D)$: From the initialization, we know $\sum_{i=1}^{k} c_i = \sum_{i=1}^{k} \sigma_{D,i}$ for all k. From the algorithm, the

excessive part $(c_k - p_k)$ will be included in c_{k-1}. This does not change the equality $\sum_{i=1}^{M} c_i = \sum_{i=1}^{N} \sigma_{D,i}$. Meanwhile, since more value are included into c_{k-1}, the inequality $\sum_{i=1}^{k} c_i \geq \sum_{i=1}^{k} \sigma_{D,i}$ for $k = 1, \ldots, M-1$ will hold.

2) Proof of $c_1 \leq c_2 \leq \ldots \leq c_M$ *and* $\mathbf{d}(\mathbf{C}^H \mathbf{C}) \leq \mathbf{p}$: From the algorithm, we know $c_k \leq p_k$ and c_k is already in nondecreasing order after the initialization. If at the current step c_k is smaller than p_k, then c_{k-1} will be kept unchanged and $c_{k-1} \leq c_k$ still holds (remember c_k will not be decreased in all previous steps). If on the other side c_k is greater than or equal to p_k, then c_k is updated to p_k and c_{k-1} is updated to $c_{k-1} + (c_k - p_k)$. However, at the next step, this c_{k-1} will be upper bounded by p_{k-1} and the excessive part $c_{k-1} - p_{k-1}$ will be added to c_{k-2}. Bearing in mind that p_ks are arranged in nondecreasing order, we know that $c_{k-1} \leq c_k$ still holds. This process continues until $k = 2$.

The speciality happens for c_1 since there is no behavior regarding whether c_1 is greater or less than p_1. Therefore, we only need to prove that the final c_1 satisfy $c_1 \leq p_1$ and $c_1 \leq c_2$. These two things can be proved together. If $c_2 \leq p_2$ still holds after getting the increment, then there will be no increment for c_1. In this case, the final c_1 is the same as the initial c_1, which is exactly $\sigma_{D,1}$, and the proof is completed. Otherwise, $c_2 > p_2$ and the excessive part will be added to c_1. Bearing in mind that c_2 may also receive the increment from the previous steps, we suppose a maximal integer $r_0 \in \{2, 3, \ldots, M\}$, such that c_k is equal to p_k for $2 \leq k \leq r_0$ when the algorithm goes to the last step. Then, the final c_1 is $\sigma_{D,1} + \sum_{i=2}^{r_0} (\sigma_{D,i} - p_i)$. From the optimization process, we know

$$\sum_{i=1}^{r_0} p_i \geq \sum_{i=1}^{r_0} \sigma_{D,i}. \quad (43)$$

Then

$$p_1 \geq \lambda_{D,1} + \sum_{i=2}^{r_0} (\sigma_{D,i} - p_i) = c_1 \quad (44)$$

can be derived. Since the final value of c_2 is p_2 in this case, we arrive at $c_1 \leq p_1 \leq c_2 = p_2$.

APPENDIX VI
PROOF OF (24)

Multiplying both side of (23) by $\sigma_{C,i}$ eliminates ν_i, and the following equation results:

$$\sigma_{C,i} \left(-\frac{1}{\sigma_n \left(\frac{1}{\sigma_{h,i}} + \frac{\sigma_{C,i}}{\sigma_n} \right)^2} + \lambda + \mu_i \right) = 0. \quad (45)$$

If $\lambda \geq \sigma_{h,i}^2/\sigma_n$, then $\lambda + \mu_i \geq \sigma_{h,i}^2/\sigma_n$. In this case, $\sigma_{C,i} > 0$ is not possible since if $\sigma_{C,i} > 0$, then $-(1/\sigma_n((1/\sigma_{h,i}) + (\sigma_{C,i}/\sigma_n))^2) + \lambda + \mu_i > 0$ and (45) cannot hold. Therefore, $\sigma_{C,i} = 0$ if $\lambda \geq \sigma_{h,i}^2/\sigma_n$. Similarly, multiplying both sides of
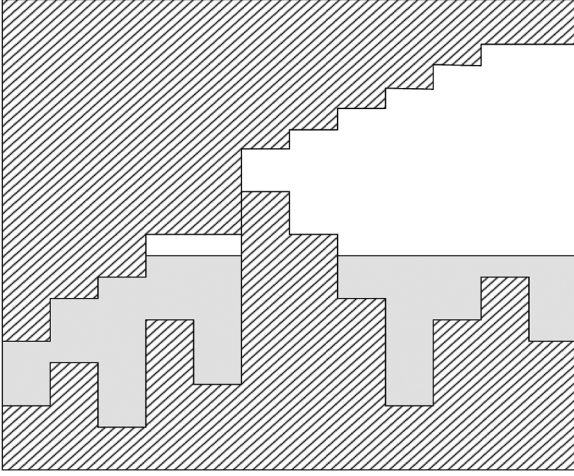
Fig. 10. Illustration on the practical cave-filling algorithm.

(23) by $(p_i - \sigma_{C,i})$ eliminates μ_i, and the following equation results:

$$(p_i - \sigma_{C,i}) \left(-\frac{1}{\sigma_n \left(\frac{1}{\sigma_{h,i}} + \frac{\sigma_{C,i}}{\sigma_n} \right)^2} + \lambda - \nu_i \right) = 0. \quad (46)$$

If $\lambda \leq 1/\sigma_n((1/\sigma_{h,i}) + (p_i/\sigma_n))^2$, then $\lambda - \nu_i \leq 1/\sigma_n((1/\sigma_{h,i}) + (p_i/\sigma_n))^2$. In this case, $\sigma_{C,i} < p_i$ is not possible since if $\sigma_{C,i} < p_i$, then $-(1/\sigma_n((1/\sigma_{h,i}) + (\sigma_{C,i}/\sigma_n))^2) + \lambda - \nu_i < 0$ and (46) does not hold. Therefore, $\sigma_{C,i} = p_i$ if $\lambda \leq 1/\sigma_n((1/\sigma_{h,i}) + (p_i/\sigma_n))^2$.

Now let us prove that when $1/\sigma_n((1/\sigma_{h,i}) + (p_i/\sigma_n))^2 < \lambda < \sigma_{h,i}^2/\sigma_n$, both μ_i and ν_i must be zero. From (46), we know $(p_i - \sigma_{C,i})(-(1/\sigma_n((1/\sigma_{h,i}) + (\sigma_{C,i}/\sigma_n))^2) + \lambda) \geq 0$. If $\lambda < \sigma_{h,i}^2/\sigma_n$, then $\sigma_{C,i}$ cannot be zero, which gives $\nu_i = 0$. From (45), we know $\sigma_{C,i}(-(1/\sigma_n((1/\sigma_{h,i}) + (\sigma_{C,i}/\sigma_n))^2) + \lambda) \leq 0$. If $\lambda > 1/\sigma_n((1/\sigma_{h,i}) + (p_i/\sigma_n))^2$, then $\sigma_{C,i}$ cannot be p_i, which gives $\mu_i = 0$. So, for $1/\sigma_n((1/\sigma_{h,i}) + (p_i/\sigma_n))^2 < \lambda < \sigma_{h,i}^2/\sigma_n$, (23) becomes

$$-\frac{1}{\sigma_n \left(\frac{1}{\sigma_{h,i}} + \frac{\sigma_{C,i}}{\sigma_n} \right)^2} + \lambda = 0 \quad (47)$$

and $\sigma_{C,i}$ is calculated as $\sigma_{C,i} = \sqrt{\sigma_n/\lambda} - (\sigma_n/\sigma_{h,i})$.

Finally, let us prove that $\lambda > 0$. If $\lambda = 0$, then from the previous discussion, ν_i must be zero. From (23), we know $\mu_i = 1/\sigma_n((1/\sigma_{h,i}) + (\sigma_{C,i}/\sigma_n))^2 > 0$ for any i, which indicates that $\sigma_{C,i} = p_i$ for all i. This forms contradiction since we assume $P < \sum_{i=1}^{M} p_i$. Therefore, λ cannot be zero, and $\sum_{i=1}^{M} \sigma_{C,i} = P$ can be drawn from KKT.

APPENDIX VII
PROOF OF ALGORITHM 3

According to the physical waterfilling, the patch with the lowest ceiling level $(\sigma_n/\sigma_{h,i}) + p_i$ will saturate first. Without loss of generality, we assume the ceiling levels are

originally ordered as[7] $(\sigma_n/\sigma_{h,1}) + p_1 \leq (\sigma_n/\sigma_{h,2}) + p_2 \leq ,\ldots,\leq (\sigma_n/\sigma_{h,M}) + p_M$, as shown in Fig. 10. We also sort the ground level according to another index set $[i]$ such that $\sigma_n/\sigma_{h,[1]} \leq \sigma_n/\sigma_{h,[2]} \leq,\ldots,\leq \sigma_n\sigma_{h,[M]}$. We first need to find all the saturated patch k, and this process is called outer iteration. Obviously, when pouring the water, the saturation gradually happens from the smallest index to the largest.
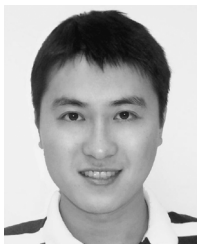
We first assume that patch 1 saturates exactly; then, there is a maximal integer M_1 such that $\sigma_n/\sigma_{h,[M_1]} \leq (\sigma_n/\sigma_{h,1}) + p_1$ and water will only be poured into patches with index set $\{[1],\ldots,[M_1]\}$. We then calculate the required total power $P_1 = \sum_{i=1}^{M_1}((\sigma_n/\sigma_{h,1}) + p_1 - (\sigma_n/\sigma_{h,[i]}))$. If this P_1 is greater than P, then we conclude that P is not large enough for any patch to saturate so that the traditional waterfilling could be applied directly on all patches. If P_1 is equal to P, then the water level γ is $(\sigma_n/\sigma_{h,1}) + p_1$. However, if P_1 is less than P, we need to go ahead and assume that patch 2 saturates exactly. Then, there is a number M_2 such that $\sigma_n/\sigma_{h,[M_2]} \leq (\sigma_n/\sigma_{h,2}) + p_2$. The required total power is $P_2 = \sum_{i=1}^{M_2}((\sigma_n/\sigma_{h,2}) + p_2 - (\sigma_n/\sigma_{h,[i]})) - (((\sigma_n/\sigma_{h,2}) + p_2) - ((\sigma_n/\sigma_{h,1}) + p_1))$, where $(((\sigma_n/\sigma_{h,2}) + p_2) - ((\sigma_n/\sigma_{h,1}) + p_1))$ is the power that should not be counted due to the saturation of the patch 1. If P_2 is greater than P, we can apply the traditional waterfilling over patches 2 to M with a total amount power $P - p_1$ (since patch 1 must be saturated from the previous step). If P_2 is equal to P, then $(\sigma_n/\sigma_{h,2}) + p_2$ is the water level. If P_2 is less than P, we need to go ahead and assume that patch 3 saturates exactly. This process should go on until we find the true water level.

REFERENCES

[1] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, pp. 585–595, Nov. 1999.

[2] G. J. Foschini, "Layered space time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs. Tech. J.*, vol. 1, pp. 41–59, 1996.

[3] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 744–765, Mar. 1998.

[4] S. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 16, pp. 1451–1458, Oct. 1998.

[5] T. M. Cover and A. A. El Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inf. Theory*, vol. IT-25, pp. 572–584, Sep. 1979.

[6] R. U. Nabar, H. Bolcskei, and F. W. Kneubuhler, "Fading relay channels: Performance limits and space time signal design," *IEEE J. Sel. Areas Commun.*, vol. 22, pp. 1099–1109, Aug. 2004.

[7] J. Boyer, D. D. Falconer, and H. Yanikomeroglu, "'Multihop diversity in wireless relaying channels," *IEEE Trans. Commun.*, vol. 52, pp. 1820–1830, Oct. 2004.

[8] J. N. Laneman and G. W. Wornell, "Distributed space time block coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2415–2425, Oct. 2003.

[9] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, pp. 3062–3080, Dec. 2004.

[10] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity—Part I: System description," *IEEE Trans. Commun.*, vol. 51, pp. 1927–1938, Nov. 2003.

[11] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity—Part II: System description," *IEEE Trans. Commun.*, vol. 51, pp. 1939–1948, Nov. 2003.

[12] S. Yiu, R. Schober, and L. Lampe, "Distributed space time block coding," *IEEE Trans. Commun.*, vol. 54, pp. 1195–1206, Jul. 2006.

[7]In MMSE estimation, we do not assume any specific ordering of either $\sigma_{h,i}$ or p_i at the very beginning.

[13] Y. Jing and B. Hassibi, "Distributed space time coding in wireless relay networks," *IEEE Trans. Wireless Commun.*, vol. 5, pp. 3524–3536, Dec. 2006.

[14] F. Gao, T. Cui, and A. Nallanathan, "On channel estimation and optimal training design for amplify and forward relay network," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1907–1916, May 2008.

[15] I. Barhumi, G. Leus, and M. Moonen, "Optimal training design for MIMO OFDM systems in mobile wireless channels," *IEEE Trans. Signal Process.*, vol. 51, no. 6, pp. 1615–1624, Jun. 2003.

[16] H. Minn and N. Al-Dhahir, "Optimal training signals for MIMO OFDM channel estimation," *IEEE Trans. Wireless Commun.*, vol. 5, pp. 1158–1168, May 2006.

[17] M. Biguesh and A. B. Gershman, "Training based MIMO channel estimation: A study of estimator tradeoffs and optimal training signals," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 884–893, Mar. 2006.

[18] T. F. Wong and B. Park, "Training sequence optimization in MIMO systems with colored interference," *IEEE Trans. Commun.*, vol. 52, pp. 1939–1947, Nov. 2004.

[19] Y. Liu, T. F. Wong, and W. W. Hager, "Training signal design for estimation of correlated MIMO channels with colored interference," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1486–1497, Apr. 2007.

[20] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*. New York: Academic, 1979.

[21] P. Viswanath and V. Anantharam, "Optimal sequences and sum capacity of synchronous CDMA systems," *IEEE Trans. Inf. Theory*, vol. 45, pp. 1984–1991, Sep. 1999.

[22] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Rev.*, vol. 39, pp. 49–95, Mar. 1996.

[23] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[24] Z. Wang and G. B. Giannakis, "Wireless multicarrier communications," *IEEE Signal Process. Mag.*, vol. 17, pp. 29–48, May 2000.

[25] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optim. Meth. Softw.*, vol. 11-12, pp. 625–653, Aug. 1999.

[26] D. P. Palomar, M. A. Lagunas, and J. M. Cioffi, "Optimum linear joint transmit-receive processing for MIMO channels with QoS constraints," *IEEE Trans. Signal Process.*, vol. 52, no. 5, pp. 1179–1197, May 2004.

[27] F. Zhang and Q. Zhang, "Eigenvalue inequalities for matrix product," *IEEE Trans. Autom. Control*, vol. 51, pp. 1506–1509, Sep. 2006.

[28] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus With Applications in Statistics and Econometrics*. New York: Wiley, 1999.
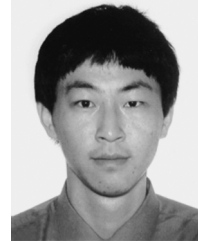
array signal processing. He has coauthored more than 40 refereed IEEE journal and conference papers.

Mr. Gao was a recipient of the President Scholarship from the National University of Singapore. He was a Technical Program Committee member for the IEEE ICC (2008, 2009), IEEE VTC (2008), and IEEE GLOBECOM (2008).

Tao Cui (S'04) received the M.Sc. degree from the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada, in 2005 and the M.S. degree from the Department of Electrical Engineering, California Institute of Technology, Pasadena, in 2006, where he is currently pursuing the Ph.D. degree.

His research interests are in the interactions between networking theory, communication theory, and information theory.

Mr. Cui received the Best Paper Award at the IEEE International Conference on Mobile Ad-hoc and Sensor Systems in 2007 and Second Place in the ACM Student Research Competition at the 2007 Richard Tapia Celebration of Diversity in Computing Conference. He received postgraduate scholarships from the Alberta Ingenuity Fund and the Alberta Informatics Circle of Research Excellence.

Arumugam Nallanathan (S'97–M'00–SM'05) received the B.Sc. degree (with honors) from the University of Peradeniya, Sri Lanka, in 1991, the CPGS degree from Cambridge University, U.K., in 1994, and the Ph.D. degree from the University of Hong Kong, Hong Kong SAR, China, in 2000, all in electrical engineering.

He was an Assistant Professor in the Department of Electrical and Computer Engineering, National University of Singapore, from August 2000 to December 2007. Currently, he is a Senior Lecturer in the Department of Electronic Engineering, King's College London, U.K. His research interests include cooperative communications, cognitive radio, MIMO-OFDM systems, ultrawide bandwidth communication, and localization. In these areas, he has published more than 130 journal and conference papers. He was a Technical Program Committee Member for more than 30 IEEE international conferences.

Dr. Nallanathan was a corecipient of the Best Paper Award at the 2007 IEEE International Conference on Ultra-Wideband (ICUWB'2007). He currently serves on the Editorial Board of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY as an Associate Editor. He was a Guest Editor for the *EURASIP Journal of Wireless Communications and Networking* "Special issue on UWB Communication Systems Technology and Applications." He also was General Track Chair for IEEE VTC'2008-Spring. He currently is Cochair of IEEE GLOBECOM'2008 Signal Processing for Communications Symposium and IEEE ICC'2009 Wireless Communications Symposium.

Feifei Gao (S'05) received the B.Eng. degree in information engineering from Xi'an Jiaotong University, Xi'an, Shaanxi, China, in 2002, the M.Sc. degree from McMaster University, Hamilton, ON, Canada, in 2004, and the Ph.D. degree from the National University of Singapore in 2007.

He is currently working as a Research Fellow at the Institute for Infocomm Research, A*STAR, Singapore. His research interests are in communication theory, broadband wireless communications, signal processing for communications, MIMO systems, and