Bit Loading Using Imperfect CSIT for Prediction-Based Resource Allocation in Mobile OFDMA

Jorge F. Schmidt, *Student Member, IEEE*,
Juan E. Cousseau, *Senior Member, IEEE*, Risto Wichman, and
Stefan Werner, *Senior Member, IEEE*

***Abstract*—We present a prediction-based resource allocation algorithm (RA) for orthogonal frequency-division multiple-access (OFDMA) downlink, where inaccuracies in the wireless channel predictions are accounted for in the problem formulation. As the prediction quality significantly degrades with the prediction horizon, we propose a solution based on the histogram of the prediction error. This characterization also enables different mobile stations (MSs) to use different channel predictors as it does not rely on a specific prediction scheme. Using this characterization of the prediction error and based on classical resource allocation strategies, we derive an algorithm that incorporates imperfect channel prediction information of future time slots. We evaluate the proposed algorithm using a practical low-complexity channel predictor suitable for implementation at the MSs. Simulation results show that the proposed algorithm outperforms previous prediction-based RA strategies without the characterization of the prediction error, and the system throughput is comparable with the case with perfect channel state information in the transmitter (CSIT).**

***Index Terms*—Channel state information, histograms, orthogonal frequency-division multiple-access (OFDMA), prediction algorithm, scheduling algorithm, time-varying channels, wireless communications.**

I. Introduction

Orthogonal frequency-division multiple-access (OFDMA) is a multiple-access technique capable of exploiting multiuser diversity in a frequency-selective fading scenario. Because of its orthogonal structure, OFDMA allows multiple users to simultaneously transmit on the different subcarriers of one orthogonal frequency-division-multiplexing (OFDM) symbol. Considering time-varying environments, the problem of allocating specific subcarriers to users over successive time slots, taking into account overall data throughput as well as fairness constraints, has received a lot of attention over the past few years [1]–[8].

Predicted channel state information in the transmitter (CSIT) plays an important role on the efficient sharing of channel resources in time-varying wireless channels. For example, in long-term evolution (LTE) downlink [9], the physical layer scheduler allocates channel resources between mobile stations (MSs) in 1-ms resolution. The resource allocation algorithm (RA) is based on achievable rate values reported by the MSs through a feedback channel. In this context, channel prediction has already been considered for compensating feedback latency in rapidly varying scenarios; see, e.g., [1] and [2]. However, the use of channel predictors to improve RA algorithms has not been fully explored yet.

With the development of long-range channel predictors [10]–[14], it becomes feasible to include information about the channel state on future time slots in the RA algorithm. However, to exploit this new information, prediction accuracy has to be taken into account. Specifically, in high-mobility scenarios, channel prediction appreciably degrades as the prediction horizon increases [10], and the assumptions of perfect CSIT or constant prediction error are no longer valid. Several RA algorithms for OFDMA downlink considering imperfect CSIT have recently been proposed [1]–[4]. A modification of the proportional fair scheduler (PFS) is proposed in [4] to predict the allocation for several time slots. This method is shown to improve fairness among users when compared with the case without prediction. However, in high-mobility scenarios and with practical implementation of channel predictors, two main issues related to prediction-based RA in rapidly varying channels can be identified: 1) the tradeoff between prediction horizon and prediction accuracy and 2) characterization of the prediction error to improve system performance.

In this paper, we propose a prediction-based RA algorithm for high-mobility scenarios. The main concern is to characterize the prediction error associated with practical channel predictors. Imperfect CSIT due to prediction error has an important impact on the overall system performance as the achievable rates of MSs are computed from CSIT by evaluating the channel power on each subchannel.

The main contributions of this paper are the following: 1) the derivation of a prediction-based RA algorithm, which is aware of prediction error and capable of providing close to perfect CSIT system throughput at the system bit error rate (BER) constraint, and 2) a statistical description of the prediction error that does not assume a specific error model (usually associated with a specific prediction technique) and is thus feasible to be applied to general long-range channel predictors. To the best of our knowledge, general theoretical models for long-range predictors have yet to be developed.

Specifically, it is shown that based on the typical parameters of practical OFDMA systems, a set of histograms can be calculated and updated periodically to approximate the statistics of the prediction error. To corroborate the effectiveness of the proposed algorithm in practical OFDMA systems, we consider realistic system parameters as well as a realistic channel model for each MS. In addition, as OFDMA downlink channel prediction is performed at the MSs, we consider in this paper a computationally inexpensive channel predictor that is practical for implementation at the MSs [10].

The outline of this paper is as follows: In Section II, we define the system model, introduce the notation, review the prediction-based RA literature, and describe a long-range channel predictor suitable for application on prediction-based RA schemes. The error-aware prediction-based RA is derived in Section III, where the effect of imperfect predicted CSIT is analyzed, and an empirical characterization is proposed to compensate the prediction error. An analysis of the computational cost associated with the proposed scheme is presented in Section IV, and the performance of the proposed prediction correction scheme is evaluated in Section V. Finally, Section VI provides our conclusions.

II. Problem Statement

A. System Model

We consider an OFDMA downlink transmission where resources are allocated on subcarrier basis. On the other hand, in LTE and mobile WiMAX, resource blocks consist of several subcarriers, either

Manuscript received November 30, 2010; revised April 1, 2011 and July 2, 2011; accepted August 9, 2011. Date of publication August 18, 2011; date of current version October 20, 2011. This work was supported in part by the Academy of Finland, by Smart Radios, by the Wireless Research Center of Excellence, Agencia Nacional de Promocion Cientifica y Tecnologica under Grant PICT 2008-0182, and by the Universidad Nacional del Sur, Argentina, under Project 24/K043. The review of this paper was coordinated by Prof. H.-H. Chen.

J. F. Schmidt and J. E. Cousseau are with Instituto de Investigaciones en Ingeniería Eléctrica (IIIE)-Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Universidad Nacional del Sur, Bahía Blanca 8000, Argentina (e-mail: schmidt@uns.edu.ar; jcousseau@uns.edu.ar).

R. Wichman and S. Werner are with the School of Electrical Engineering, Aalto University, Aalto 00076, Finland (e-mail: risto.wichman@aalto.fi; stefan.werner@aalto.fi).

Digital Object Identifier 10.1109/TVT.2011.2165305

contiguous or distributed within the OFDMA band, and the signaling of rate values corresponding to the resource blocks is vendor specific. Nevertheless, we chose to minimize the system assumptions and study a generic case where a resource block consists of one subcarrier. However, the concept can be adapted to other types of resource blocks as well.

The system under consideration has N available subchannels and K active MSs, and the resources are allocated on a time slot basis, where a time slot consists of M consecutive OFDM symbols. For each time slot s, the base station (BS) scheduler decides which MS is assigned to a particular subchannel n. More than one subchannel might be assigned to an MS, depending on its rate requirement or channel state.

P_t symbols carrying pilot subcarriers are evenly distributed over a time slot, and P_f pilot subcarriers are evenly distributed in frequency over subcarriers on the pilot symbols to aid channel estimation and prediction at the MSs. Pilot symbols are quaternary phase-shift keying, whereas data symbols can be taken from a set of available quadratic-amplitude modulation (QAM) constellations $[\beta_1, \ldots, \beta_{\mathcal{K}}]$. Furthermore, convolutional coding of the data symbols is used to allow a finer grid of possible bit rates. The BS transmits on each subchannel using one of the modulation and coding schemes subject to the system BER constraint. The MSs are assumed to experience independent channel fading with the same statistics. Regarding the time selectivity of the channel, we assume that the channel varies significantly from one time slot to the next. As the RA is performed on a time slot basis, we assume that the channel variation over a time slot can be neglected. Thus, the channel frequency response for time slot s is given by

$$H_k(s,n) \approx \sum_{l=1}^{L} h_k(s,l) e^{-j\frac{2\pi l n}{N}} \tag{1}$$

which is a complex Gaussian random variable with zero mean and $\sigma^2_{H_k}$ variance. For $l = 1, \ldots, L$, $h_k(s,l)$ denotes the lth channel tap for user k over time slot s, with L being the length of the channel impulse response. CSIT is estimated for the current time slot and predicted for the following W time slots. There are two alternatives to acquire CSIT at the BS. The prediction can be performed in either BS or MSs. The latter case requires a larger overhead in the feedback channel as W estimates are fed back for each slot. In the former case, the overhead is less at the expense of a high computational load at the BS. The first alternative is implemented in this paper, whereas the derivations apply in both cases. In the following, the transmitted power is assumed to be the same in all subchannels to emphasize the bit loading. We denote the rate achieved by user k on time slot s by $R_k(s)$, which is given as the sum rate over all subchannels assigned to user k as $R_k(s) = M \sum_{n \in \mathcal{I}_k} r_k(m,n)$, where $\mathcal{I}_k$ is the index set for the subchannels assigned to user k, and $r_k(m,n)$ is the rate for user k at symbol time m on subchannel n. This way, the overall system throughput for time slot s is given by

$$R(s) = \sum_{k=1}^{K} R_k(s). \tag{2}$$

The design goal is to maximize the system throughput defined in (2) with constraints on total transmit power, system BER, and fairness among users.

B. Prediction-Based RA

A fair and efficient sharing of radio resources among users is an important design factor in wireless networks. PFS is a popular solution to provide a fair distribution of resources among users when CSIT is available. Denoting the kth MS average data rate by $\bar{R}_k$, based on a fairness reasoning, it aims to maximize MSs $\bar{R}_k$ over time based on allocations on previous time slots. That is, $U(s+1) = \arg\max_{\mathcal{U}} \sum_{k=1}^{K} \log(\bar{R}_k(s))$, where $U(s+1)$ is the utility function to be maximized for allocating time slot $s+1$ using information up to time slot s, and $\mathcal{U}$ is the set of all possible allocations for the considered time slot. The algorithm keeps track of the average throughput $\bar{R}_k(s)$ of each MS in an exponentially weighted window of length τ.

The performance of this scheduling algorithm can be improved if channel predictions from MSs are available for future time slots. However, long-range channel predictors have not been fully exploited within RA algorithms. An extension of the PFS to include CSIT prediction has been derived in [4]. This prediction-based PFS (P-PFS) calculates the achievable rates in a prediction window from slot $s+1$ to $s+W$, and the average rates at the end of this window are maximized to allocate time slot $s+1$. The prediction-based allocation scheme can be expressed as

$$U(s+W) = \arg\max_{\mathcal{U}} \sum_{k=1}^{K} \log\left(\bar{R}_k^W\right) \tag{3}$$

where $\bar{R}_k^W$ denotes the average rate of user k at the end of the next W time slots and is computed as

$$\bar{R}_k^W = \left(1 - \frac{1}{\tau}\right)^W \bar{R}_k(s) + \frac{1}{\tau} \sum_{w=1}^{W} \left(1 - \frac{1}{\tau}\right)^{W-w} \check{R}_k(s+w) \tag{4}$$

where $\check{R}_k$ indicates a "virtual" allocated rate, as time slots $s+2$ to $s+W$ have yet to be allocated. A practical algorithm to implement (3) is also given in [4], which will be used in this paper.

C. Channel Prediction

To be able to use information on future time slots, the achievable rates must be estimated within the prediction window. In what follows, we briefly describe a low-complexity channel predictor [10] that will be used both as a motivating example and for testing the proposed prediction-based RA scheme.

To simplify the notation, we focus in this section on only one MS and drop MS index k. It is shown in [10] that the time variation over $\bar{M}$ (several times larger than M) OFDM symbols of the channel coefficient corresponding to one subcarrier can be well described in terms of a size $\bar{M}$ discrete cosine transform (DCT) truncated to its first $G \ll \bar{M}$ basis functions. The basis dimension G is determined from the maximum expected channel Doppler shift and the DCT energy compression characteristics. By interpreting the DCT basis functions as the impulse responses of ideal bandpass filters centered at the cosine frequencies, it is also shown that an approximation $\hat{H}_{(FB)}(m,n)$ of $H(m,n)$ for $m = 1, \ldots, \bar{M}$ can be obtained by recursive filtering of $H(m,n)$ with a second-order infinite impulse response filter bank based on a normalized all-pass lattice realization, whose frequency response is given by

$$H_{FB}(e^{j\omega}) = \frac{0.5(1 - s_{20})(1 + e^{-j2\omega})}{1 - s_{20} e^{-j2\omega}} + \sum_{i=1}^{G-1} \frac{0.5(1 - s_{2i})(1 - e^{-j2\omega})}{1 + (s_{2i} + 1) s_{1i} e^{-j\omega} + s_{2i} e^{-j2\omega}} \tag{5}$$

where $s_{1i} = -\cos(\pi i / \bar{M})$ is the lattice parameter defining the filter central frequencies, and $s_{2i} (0 < s_{2i} < 1)$ is related to the 3-dB bandwidth of each narrowband filter. By defining $\mathbf{x}_i(m,n)$, $0 \le i \le$

$G-1$, the state vector for each passband filter on subcarrier n, the filter bank structure of (5) can be expressed in state space form as

$$\begin{aligned}\mathbf{x}_i(m+1,n) &= \mathbf{A}_i\mathbf{x}_i(m,n) + \mathbf{b}_i H(m,n)\\ \hat{H}_{(FB)i}(m,n) &= \mathbf{c}_i\mathbf{x}_i(m,n) + d_i H(m,n)\\ \hat{H}_{(FB)}(m,n) &= \sum_{i=0}^{G-1}\hat{H}_{(FB)i}(m,n)\end{aligned} \tag{6}$$

where $\mathbf{A}_i$, $\mathbf{b}_i$, $\mathbf{c}_i$, and d_i are the state space descriptions of the filters in (5). Using (6), the optimum algorithm for estimating $H(m,n)$ in additive white Gaussian noise is a set of G scalar Kalman filters for the filter bank of (6). Further, as the DCT basis functions are fixed for the whole transmission, the Kalman gains have a steady-state solution resulting in a low-complexity limiting Kalman filter. Having the Kalman estimate $\hat{H}(m,n)$ as input, a long-range channel predictor can be obtained by decimating on T (determined according to the maximum channel Doppler shift such that the decimated channel is sampled above the Nyquist frequency) the input estimates, scaling up in frequency the passband filters by the same factor T, and using $\mathcal{L}$ step extrapolation on the limiting Kalman filter. In this manner, the following limiting Kalman predictor can be obtained as

$$\begin{aligned}e^p(m+\ell,n) &= \hat{H}(m+\ell-T,n)\\ &\quad - \hat{H}^p(m+\ell-T,n)\\ \mathbf{x}^p_{i\ell}(m+\ell,n) &= \left(\mathbf{A}^p_i\right)^{\mathcal{L}}\mathbf{x}^p_{i\ell}(m+\ell-T\mathcal{L},n)\\ &\quad + \mathbf{k}^p_i e^p(m+\ell,n)\\ \hat{H}^p_i(m+\ell,n) &= \mathbf{c}^p_i\mathbf{x}^p_{i\ell}(m+\ell,n)\\ \hat{H}^p\left(m+\ell+T(\mathcal{L}-1),n\right) &= \sum_{i=0}^{G-1}\hat{H}^p_i(m+\ell,n)\end{aligned} \tag{7}$$

where $\ell = 0,\ldots,(T\times\mathcal{L})-1$ are the state vector samples used at each iteration, $\mathbf{x}^p_{i\ell}$ is the predictor state vector, $\mathbf{A}^p_i$ and $\mathbf{c}^p_i$ are the transition matrix and output vector of the passband filters scaled up in frequency by T, and $\mathbf{k}^p_i$'s are the corresponding steady-state Kalman gains. Finally, to obtain the predicted channel over all subcarriers, one limiting Kalman filter predictor is run over each of the P_f available pilot subcarriers and interpolated for all the other subcarriers using DCT interpolation in frequency.

III. Error-Aware Prediction-Based Resource Allocation Algorithm

In this section, we evaluate achievable rates $\tilde{r}_k(s+w,n)$ for future time slots (input to the P-PFS of Section II-B) when imperfect CSIT due to prediction errors is available. When perfect CSIT is available, extending [17] for coded modulation, the instantaneous BER for subchannel n in symbol time m can be approximated by

$$P_e(s,n) \approx c_1 \exp\left\{\frac{-c_2\gamma\delta(s,n)\left|H(s,n)\right|^2}{2^{\beta(s,n)}-1}\right\} \tag{8}$$

where $c_1 = 0.2$, $c_2 = 1.6$, $\beta(s,n)$ is the number of bits per symbol of the QAM constellation used for a subchannel and time slot, and $\delta(s,n)$ is the coding gain, with respect to the uncoded case, for the corresponding QAM constellation. The SNR value γ accounts for the path loss, whereas the small-scale fading effects are represented by $H(s,n)$. By evaluating (8) for the possible pairs of $\beta(s,n)$ and $\delta(s,n)$, the one that satisfies the system BER constraint is selected and determines the number of bits per symbol $\tilde{r}(s,n)$ for transmission.

In a practical scenario, it is impossible to have perfect CSIT, and the determination of $\tilde{r}(s,n)$ is based on an estimate $\hat{H}(s,n)$ of $H(s,n)$. In this case, the use of (8) no longer guarantees that the BER constraint is satisfied. In the case of imperfect CSIT [17], an average BER $\bar{P}_e(s,n) = E_{|H(s,n)|_{|\hat{H}(s,n)}}\{P_e(s,n)\}$ is considered instead of the instantaneous BER of (8), where the expectation is evaluated over $|H(s,n)|_{|\hat{H}(s,n)}$. By defining the random variable $\alpha = |H(s,n)|_{|\hat{H}(s,n)}$, it results in

$$\begin{aligned}\bar{P}_e(s,n) &= \int_0^\infty P_e(s,n)f(\alpha)d\alpha\\ &= \int_0^\infty c_1\exp\left\{\frac{-c_2\gamma\delta(s,n)\alpha^2}{2^{\beta(s,n)}-1}\right\}f(\alpha)\,d\alpha\end{aligned} \tag{9}$$

where $f(\alpha)$ is the probability density function (pdf) of $|H(s,n)|_{|\hat{H}(s,n)}$. When $\hat{H}(s,n)$ is an estimate or a slightly delayed estimate of the true channel coefficient $H(s,n)$, several assumptions hold regarding $f(\alpha)$ [2], [17], [18], and closed-form solutions can be derived to evaluate (9).

Unfortunately, for the prediction-based RA application considered here, $\hat{H}(s,n)$ is a prediction of $H(s,n)$ based on significantly older CSIT, and a derivation of $f(\alpha)$, besides being complex to obtain, depends on the characteristics of the channel predictor used to obtain $\hat{H}(s,n)$. In the following paragraphs, we reformulate (9) in terms of the prediction error instead of α, as the former is the available observation in the case of the prediction-based RA.

For a prediction-based RA algorithm considering CSIT up to W time slots in the future, W different pdfs should be evaluated. Let us define for each considered prediction horizon the random variables $e_w = (\alpha_w - |\hat{H}(s+w,n)|_{|H(s,n)})/\sigma^2_H$ such that the pdfs of α_w and e_w are related by

$$f_{\alpha_w}(\alpha_w) = \frac{1}{\sigma^2_H}f_{e_w}(e_w). \tag{10}$$

Replacing (10) in (9), the estimated BER for each prediction horizon can be evaluated in terms of the defined normalized prediction error e_w using

$$P_{e_w}(s+w,n) = \hat{P}_{e_w}(s+w,n)\times\rho_w(n) \tag{11}$$

where $\hat{P}_{e_w}(s+w,n)$ denotes the evaluation of (8) for the available predicted value of the channel coefficient

$$\begin{aligned}&\hat{P}_{e_w}(s+w,n)\\ &\quad = c_1\exp\left\{\frac{-c_2\gamma\delta(s+w,n)\left|\hat{H}(s+w,n)\right|^2_{|H(s,n)}}{2^{\beta(s+w,n)}-1}\right\}\end{aligned} \tag{12}$$

and $\rho_w(n)$ is the correction factor compensating the imperfect CSIT through the normalized prediction error

$$\rho_w(n) = \frac{1}{\sigma^2_H}\int_{-\infty}^{\infty}\exp\left\{\frac{-c_2\gamma\delta(s+w,n)\theta(s+w,n)}{2^{\beta(s+w,n)}-1}\right\}f_{e_w}(e_w)\,de_w \tag{13}$$

where $\theta(s+w,n) = 2\sigma^2_H|\hat{H}(s+w,n)|_{|H(s,n)}e_w + \sigma^4_H e^2_w$. It is clear that for perfect CSIT, $\rho_w(n) = 1$. We derive next an estimator of $f_{e_w}(e_w)$ in (13) to evaluate this correction factor.

Assuming that the prediction error e_w is identically distributed over the system subchannels and is independent of $H(s,n)$, the number Q of prediction error samples for the different subchannels can be used to construct a histogram to estimate $f_{e_w}(e_w)$. Following [19, Ch. 5.5], to find $\hat{f}_{e_w}(e_w)$, we use the empirical pdf. We expect that as $Q \to \infty$, $\hat{f}_{\epsilon}(e_w)$ approaches the true density $f_{e_w}(e_w)$.

To derive the estimator, let the ith histogram interval be centered on ϵ_i, and let ω be its width. Hence, the ith interval will be denoted by $\Delta_i = \{e_w : \epsilon_i - (\omega/2) < e_w \leq \epsilon_i + (\omega/2)\}$ so that the histogram $\hat{f}_{\epsilon}(e_w)$ is defined by

$$\hat{f}_{\epsilon_i}(e_w) = \frac{Q_i}{Q\omega}, \qquad \text{for } e_w \in \Delta_i \tag{14}$$

with Q_i being the number of observations for which $e_w \in \Delta_i$. The interval width ω, together with the setting of a maximum error threshold ϵ_Q, determine the number of probability mass points of the estimator. The bias of $\hat{f}_{\epsilon}(e_w)$ as an estimator of $f_{e_w}(e_w)$ can be shown for each interval to be $b(\hat{f}_{\epsilon_i}(e_w)) \approx f''_{e_w}(\epsilon_i)(\omega^2/24)$, where $f''_{e_w}(\epsilon_i)$ is the second derivative of $f_{e_w}(\epsilon_i)$, and the normalized standard error σ_ϵ to estimate $f_{e_w}(\epsilon_i)$ is $\sigma_\epsilon = \sqrt{1 - \omega f_{e_w}(\epsilon_i)/Q\omega f_{e_w}(\epsilon_i)}$. It can be noted that the effect of the cell width ω on σ_ϵ is opposite to that on the bias such that an appropriate choice of ω is to set $\omega \propto 1/\sqrt{Q}$ [19].

Using the estimated pdf for the prediction error, (13) can be rewritten in terms of (14) as

$$\hat{\rho}_w(n) = \frac{1}{\sigma_H^2} \sum_{-\epsilon_Q}^{\epsilon_Q} \exp\left\{ \frac{-c_2 \gamma \delta(s+w,n)\tilde{\theta}(s+w,n)}{2^{\beta(s+w,n)} - 1} \right\} \hat{f}_{\epsilon_i}(\epsilon_i) \tag{15}$$

where $\tilde{\theta}(s+w,n) = 2\sigma_H^2 |\hat{H}(s+w,n)|_{|H(s,n)\epsilon_i} + \sigma_H^4 \epsilon_i^2$. This last expression is then inserted into (11), replacing $\rho_w(n)$, and used to evaluate $P_{e_w}(s+w,n)$ on each subchannel such that

$$P_{e_w}(s+w,n) = \hat{P}_{e_w}(s+w,n) \times \hat{\rho}_w(n). \tag{16}$$

The pair $\beta^*(s+w,n)$, $\delta^*(s+w,n)$ satisfying (16) determines the achievable rate $\tilde{r}(s+w,n)$ on subchannel n. It is worth noting that as the number of subcarriers in typical OFDMA systems is usually large (above 512 subcarriers), the bias and standard error will be small, and the W histograms built for each MS will describe the statistics of the prediction error accurately. Further, the histograms for estimating $f_{e_w}(\epsilon_i)$ are constructed assuming perfect CSIT at the MS for current time slot. In a practical implementation, $H_k(s,n)$ is estimated at the MS for symbol detection. Assuming that the estimation error is much lower than the prediction error in the prediction range, which is a reasonable assumption for practical estimators/predictors (as will be shown in Section V), the former can be used as a reference for the characterization of the latter without affecting much the resulting $\hat{f}_{\epsilon_i}(\epsilon_i)$. In addition, if we further assume that the statistics of e_w are slowly time variant, then the histograms can periodically be improved with the incorporation of new data points. That is

$$\hat{f}^{(s)}_{\epsilon_i}(\epsilon_i) = \xi \hat{f}^{(s)}_{\epsilon_i}(\epsilon_i) + (1-\xi)\hat{f}^{(s-W)}_{\epsilon_i}(\epsilon_i) \tag{17}$$

where $0.5 \leq \xi \leq 1$ is a tuning parameter selected to compensate for the possibly time-varying statistics of e_w. Finally, the resulting values of $\tilde{r}_k(s+w,n)$ corresponding to each MS are fed to the prediction-based RA algorithm.

IV. Complexity Analysis

The proposed prediction error characterization/compensation scheme involves an additional computational cost when compared with the baseline scenario of uncompensated imperfect CSIT. For the uncompensated case, the computational cost for the bit loading is given by the evaluation of (8) for the possible pairs of β and δ. Using the binary search algorithm to this end, the average number of required evaluations of (8) for this search is $\log_2(\tilde{\mathcal{K}}) - 1$ and at most $\log_2(\tilde{\mathcal{K}})$ probes in the worst case, where $\tilde{\mathcal{K}}$ is the number of available β–δ pairs, which is a small number for practical systems. On the other hand, for the proposed scheme, the computational cost involved in the histogram construction of (14) and in the evaluation of the correction factor $\hat{\rho}_w(n)$ of (15) should be considered besides the evaluation of $\hat{P}_{e_w}$, similar to (8).

The constructed histograms have $2\epsilon_Q + 1$ bins (probability mass points). The value of Q_i in (14) for each probability mass point can also be evaluated using the binary search algorithm requiring on average $\log_2(2\epsilon_Q + 1)$ trials for each sample. Thus, the computational cost associated with W histogram construction is evaluated as $\mathcal{O}(WQ\log_2(2\epsilon_Q + 1))$, and assuming that the histograms are updated every W time slots, it reduces to $\mathcal{O}(Q\log_2(2\epsilon_Q + 1))$. Regarding the computation of the correction factor $\hat{\rho}_w(n)$ of (15), it must be evaluated for all possible pairs of β and δ to be then inserted into (16). The evaluation of (15) involves $\mathcal{O}(2\epsilon_Q + 1)$ complex operations, which leads to a computational cost of $\mathcal{O}(\tilde{\mathcal{K}}(2\epsilon_Q + 1))$ complex operations.

Summarizing the discussion from the previous paragraphs, the additional computational cost of the proposed scheme, when compared with the baseline of the uncompensated case, is $\mathcal{O}(Q\log_2(2\epsilon_Q + 1) + \tilde{\mathcal{K}}(2\epsilon_Q + 1))$. Thus, in practice, the complexity is dominated by the number of error samples Q, which is much larger than the number of available modulation/coding pairs $\tilde{\mathcal{K}}$.

V. Numerical Evaluation

In this section, we evaluate the impact of the prediction error aware bit loading presented in Section III to the performance of the prediction-based RA algorithm in [4]. We seek to quantify the performance improvement achieved with the proposed empirical characterization of the prediction error.

We consider the downlink of an OFDMA system based on the 3rd Generation Partnership Project specification [9] operating at a carrier frequency $f_C = 2$ GHz with 10-MHz bandwidth, 15-kHz subcarrier spacing, and 5-μs cyclic prefix. The fast Fourier transform size is 1024, and $N = 600$ subcarriers are in use, occupying 90% of the bandwidth. The length of the time slot is $M = 15$ OFDMA symbols, which is equivalent to 1 ms, and the system uses 5% pilot ratio. 4-, 16-, and 64-QAM constellations are available for data transmission. In addition, the system employs convolutional coding for data transmission using the (133,171) rate 1/2 code with puncturing to obtain 2/3 and 3/4 code rates. A bit loading function similar to [8] is used corresponding to a system target BER of 10^{-3}. All the simulation results are obtained by averaging over time to average the channel fading statistics.

Based on the derivations in Section III, the parameters for the construction of the histograms are chosen as $Q = 600$ (all system subchannels used), $\omega = 0.078$, and the limits for the histogram set to ± 0.8. Finally, the parameter ξ for updating of the histograms is set to $\xi = 0.7$. To evaluate a realistic propagation environment, we consider independent wireless fading channel for each MS following the ITU-Vehicular A model, which results in a 27-tap frequency-selective channel for the given system parameters. Each channel tap varies in time according to Jakes Doppler spectrum, and we employ the practical channel estimator/predictor structure of [10]. Following [10], the parameters for this channel predictor are $\bar{M} = 12$ M such that $s_{1i} = -\cos(\pi i/\bar{M})$. It is shown in [10] that the performance is robust to the typical values of the selectivity parameter s_{2i} such that

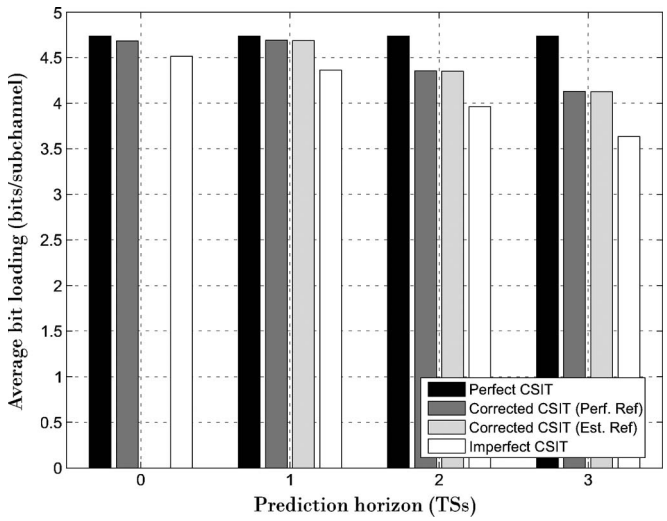

Fig. 1. Average bit loading per subcarrier without prediction and for $w = 1, 2,$ and 3 time slot predictions for a single MS. Results are averaged over 2000 time slots, as well as over MSs and system subchannels. Mobile speed is set to 50 km/h, SNR is 25 dB, and the system target BER is $10e^{-3}$.

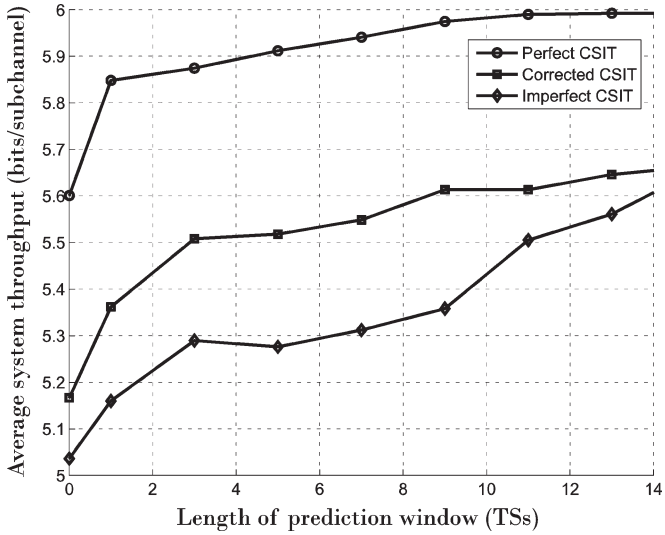

Fig. 2. System throughput versus prediction window length for P-PFS. Results are averaged over 1000 time slots, as well as over MSs and system subchannels to obtain system level results. Mobile speed is set to 50 km/h, SNR is 25 dB, and the system target BER is $10e^{-3}$.

its value is set to 0.998. In addition, for the MS speeds of interest, a value of $G = 8$ results for the basis expansion dimension. Finally, the extrapolation factor for the Kalman predictor is fixed to $\mathcal{L} = 3$ and the decimation factor T adjusted to meet the different prediction horizons.

Before addressing the performance of P-PFS, Fig. 1 illustrates the virtual rates $\check{R}_k(s+w)$ in (4) (input to P-PFS) when imperfect CSIT is available. The figure shows the average computed $\check{R}_k(s+w)$ for a single MS over three successive time slots based on the predicted CSIT with an SNR of 25 dB. The input to the P-PFS of (3) is the average rate $\bar{R}_k^W$ in (4) of user k at the end of the $W = 3$ time slots considered computed based on these virtual rates $\check{R}_k(s+w)$. In general, it noticed that imperfect CSIT degrades the virtual rates. However, even with imperfect CSIT, the prediction increases the system throughput, as will be seen later on. The proposed bit loading based on (16) is compared against the uncompensated case (8), and the case of perfect CSIT is also shown for reference. As expected, the average bit loading is independent on the prediction horizon when perfect CSIT is available. Two bars for imperfect CSIT are shown for the proposed bit loading when the true channel (Perf. Ref.) and channel estimates (Est. Ref.) are employed for the histogram construction. It can be noted that the proposed scheme effectively reduces the loading gap between the perfect and imperfect cases, obtaining a reduction from 0.35, 0.75, and 1.1 to 0.1, 0.35, and 0.6 bits/subchannel for $w = 1, 2,$ and 3 time slot prediction, respectively. Although not used by P-PFS, bit loading without prediction $(w = 0)$ is also plotted for reference.

Fig. 2 shows, for a ten MSs, 50 km/h, and 25-dB SNR scenario, the improvement in system throughput obtained by the P-PFS as the prediction window increases (0 stands for the conventional PFS). It can be noted that for all prediction window lengths, the performance is significantly improved using the proposed scheme (16), whereas the improvement by the corrected CSIT is larger for shorter prediction windows as the associated prediction error is smaller. This figure clearly shows the capabilities of the prediction-based approach. At the same time, the impact of the degraded virtual rates due to imperfect CSIT $\check{R}_k(s+w)$ (illustrated in Fig. 1) on the performance of P-PFS is evident.

In Fig. 3, system throughput is evaluated for a prediction window of length $W = 3$ and up to 20 MSs moving at 25 km/h in high and low SNR scenarios. The channel power for MSs is uniformly distributed between -3 and 0 dB. The 25- and 10-dB SNR settings result in

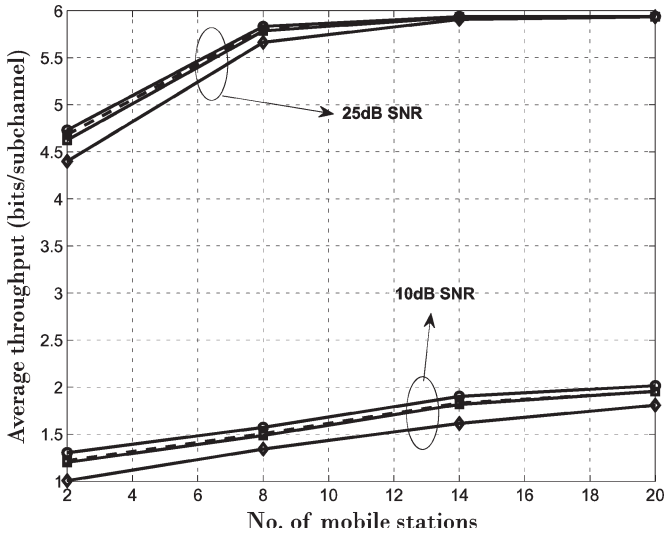

Fig. 3. System throughput versus the number of MSs for $W = 3$ P-PFS. Results are averaged over 500 time slots, as well as over MSs and system subchannels to obtain system level results and are shown for the case where the predicted CSIT is error free (circle marker), the predicted CSIT is corrected (square marker) as proposed in (16), and for uncompensated predicted CSIT using (8) (diamond marker). Mobile speed is set to 25 km/h, and the system target BER is $10e^{-3}$.

channel estimate references with MSE of -60.15 and -33.64 dB, respectively, for the construction of the histograms. In both scenarios, the results for perfect CSIT are plotted as a best performance bound. As the RA for the next time slot is based on the computation of the average rates achieved by MSs at the end of the prediction window, if the computed achievable rates for future time slots differ too much from the true achievable rates, the prediction-based RA performance will degrade in terms of system throughput or in terms of the system BER constraint. P-PFS with imperfect CSIT is not able to exploit the available resources to full extent, and the system throughput is reduced, in particular for the case of few MSs. On the other hand, it can be seen that the proposed correction scheme gives close to perfect CSIT performance, which demonstrates its effectiveness. This happens because, as the histograms constructed in (14) are not restricted to

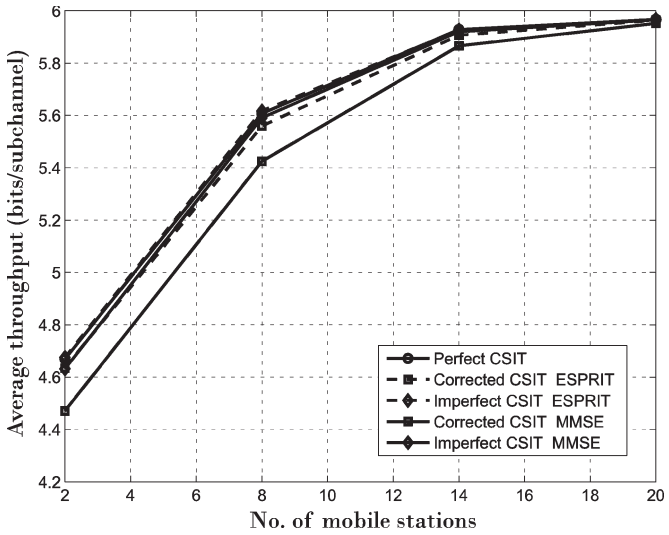

Fig. 4. System throughput versus the number of MSs for $W = 3$ P-PFS and different prediction schemes. Results are averaged over 500 time slots, as well as over MSs and system subchannels to obtain system-level results. Mobile speed is set to 25 km/h, SNR is 25 dB, and the system target BER is $10e^{-3}$.

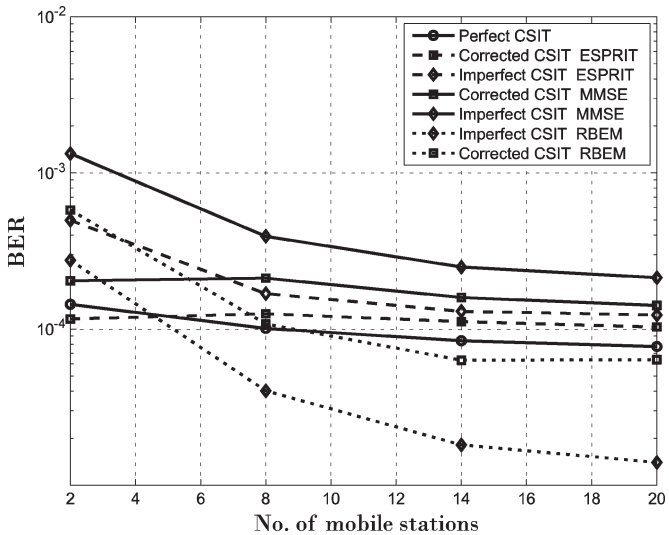

Fig. 5. System BER versus the number of MSs for $W = 3$ P-PFS using different prediction schemes. Mobile speed is set to 25 km/h, SNR is 25 dB, and the system target BER is $10e^{-3}$.

have zero mean, the correction factor of (15) has the capability to compensate a bias on the predicted CSIT. As expected, in this low SNR regime, system throughput is reduced as the channel quality for the MSs is degraded. However, the proposed error correction scheme still effectively reduces the throughput gap for both references used in the histogram construction. It is worth noting that the performance improvement obtained with the proposed scheme is larger in this last case as the uncompensated bit loading cannot approach the perfect CSIT case, even for a large number of users (where the probability of having a user with good channel state at each time slot increases). This latter conclusion justifies the added complexity of the proposed scheme.

Fig. 4 shows again the system throughput for the high SNR setting in Fig. 3, this time evaluated for a linear predictor [12] and a sum of sinusoid predictor [11], which we refer to as minimum mean square error (MMSE) predictor and estimation of signal parameters via rotational invariance (ESPRIT) based predictor, respectively. It is worth noting that the computational cost of these predictors is significantly larger than that of the predictor of [10] so that their implementation at the MSs may not be feasible. It is observed for both predictors that the proposed scheme also gives close to perfect CSIT performance, showing that it can be applied to different prediction schemes as the characterization of the prediction error is independent of the channel prediction scheme.

Regarding BER behavior using the proposed error correction scheme, Fig. 5 shows the BER results for the high SNR setting in Fig. 3 and all channel predictors tested. As the bit loading function used is discrete, all the curves are below the BER constraint, as expected, and thus, the perfect CSIT case is used as a reference for performance comparison. When the channel power is underestimated, the uncompensated bit loading algorithm assigns modulation/coding schemes that meet the BER constraint for this underestimated power. If the channel state is in reality better, then the BER attained with the assigned more robust modulation/coding pair will be lower, indicating that the system resources are not fully exploited. On the other hand, if the channel power is overestimated, then the assigned modulation/coding scheme will not be supported by the real channel state, thus leading to a higher BER, indicating a loss in transmission reliability. It is easy to see in Fig. 5 that while the recursive basis expansion model predictor in Section II-C underestimates the channel power, the MMSE and ESPRIT predictors overestimate the channel

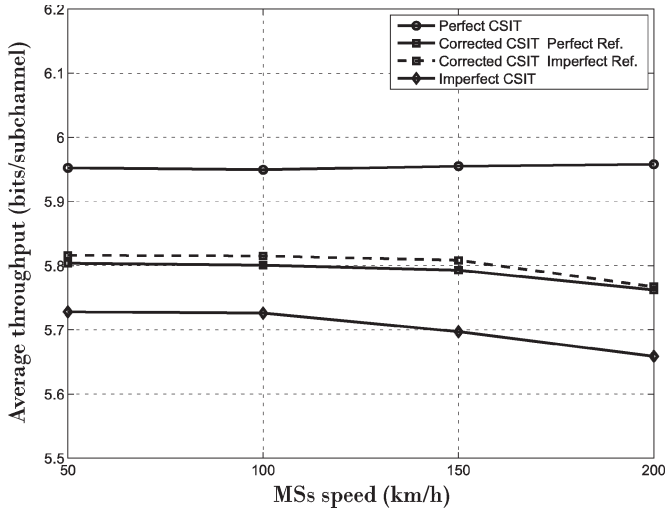

Fig. 6. System throughput for $W = 3$ prediction-based RA. Results are averaged over 2000 time slots, as well as over MSs and system subchannels to obtain system level results. There are ten MSs moving at speeds ranging from 50 to 200 km/h and an SNR of 25 dB.

power. However, in all cases, the proposed compensation scheme effectively reduces the BER gap relative to the perfect CSIT case. As observed in Fig. 3, the constructed histograms take into account a possible under/overestimation of the channel power and, thus, can compensate this effect in bit loading. In the high SNR scenarios of Figs. 3 and 4, for a large number of users, all the schemes converge to the same total system throughput. This happens because as the number of users increases, the probability of having a user with good channel conditions at all time slots increases, as does the system throughput. However, the achievable rates computed by the different schemes are not necessarily the same. Thus, the MSs scheduled by P-PFS might not be the same in all cases, leading to different BER performances for each scheme. This figure shows that the proposed technique approaches the perfect CSIT case for all channel predictors used, demonstrating that it does not rely on a specific prediction structure.

A last simulation setting is shown in Fig. 6, where we further investigate the performance dependence on the MS speed. We consider again a prediction window of $W = 3$ time slots and ten MSs moving at

speeds from 50 to 200 km/h with 25-dB SNR. This figure shows how the bit loading gap increases for higher speeds (the prediction error increases for fixed prediction range), resulting in a throughput loss for high MS speeds. In all cases, the proposed scheme reduces this loss significantly, compared with the uncompensated case.

VI. Conclusion

We have proposed a characterization of the prediction error for prediction-based resource allocation for OFDMA downlink over mobile wireless channel when imperfect channel state information is available. Based on the large amount of frequency data samples available in a typical OFDMA system, we derived an empirical approach based on histograms for the characterization of the prediction error for the different prediction horizons considered in the prediction window.

We evaluated the proposed scheme under realistic channel conditions, system parameters, and a practical channel predictor that is feasible for implementation at MSs. Simulation results indicate that the proposed scheme outperforms similar prediction-based resource RAs that disregard the prediction error.

References

[1] S. Stefanatos and N. Dimitriou, "Downlink OFDMA resource allocation under partial channel state information," in *Proc. IEEE ICC*, Jun. 2009, pp. 1–5.

[2] I. Wong and B. Evans, "Optimal resource allocation in the OFDMA downlink with imperfect channel knowledge," *IEEE Trans. Commun.*, vol. 57, no. 1, pp. 232–241, Jan. 2009.

[3] A. Falahati and M. Ardestani, "An improved low-complexity resource allocation algorithm for OFDMA systems with proportional data rate constraint," in *Proc. IEEE ICACT*, Feb. 2007, vol. 1, pp. 606–611.

[4] J. Hajipour and V. C. M. Leung, "Proportional fair scheduling in multi-carrier networks using channel predictions," in *Proc. IEEE ICC*, May 2010, pp. 1–5.

[5] S. Najeh, H. Besbes, and A. Bouallegue, "Predictive approach for OFDMA resource allocation over fixed wireless channels," in *Proc. IEEE ICSCS*, Nov. 2009, pp. 1–5.

[6] D. Ngo, C. Tellambura, and H. Nguyen, "Efficient resource allocation for OFDMA multicast systems with spectrum-sharing control," *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, pp. 4878–4889, Nov. 2009.

[7] Y. Zhang and C. Leung, "Subchannel power-loading schemes in multiuser OFDM systems," *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, pp. 5341–5347, Nov. 2009.

[8] D. Kivanc, G. Li, and H. Liu, "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Trans. Wireless Commun.*, vol. 2, no. 6, pp. 1150–1158, Nov. 2003.

[9] "Physical layer aspects for evolved UTRA," 3GPP Tech. Rep., TR 25.814, Ver. 1.0.3, Feb. 2006.

[10] J. Schmidt, J. Cousseau, R. Wichman, and S. Werner, "Low-complexity channel prediction using approximated recursive DCT," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, no. 10, 2011, to be published.

[11] I. Wong and B. Evans, "Low-complexity adaptive high-resolution channel prediction for OFDM systems," in *Proc. IEEE GLOBECOM*, Dec. 2006, pp. 1–5.

[12] D. Schafhuber and G. Matz, "MMSE and adaptive prediction of time-varying channels for OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 593–602, Mar. 2005.

[13] S. Semmelrodt and R. Kattenbach, "Investigation of different fading forecast schemes for flat fading radio channels," in *Proc. IEEE VTC*, Oct. 2003, vol. 1, pp. 149–153.

[14] A. Duel-Hallen, "Fading channel prediction for mobile radio adaptive transmission systems," *Proc. IEEE*, vol. 95, no. 12, pp. 2299–2313, Dec. 2007.

[15] N. Mokari, M. Javan, and K. Navaie, "Cross-layer resource allocation in OFDMA systems for heterogeneous traffic with imperfect CSI," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 1011–1017, Feb. 2010.

[16] M. Awad, V. Mahinthan, M. Mehrjoo, X. Shen, and J. Mark, "A dual-decomposition-based resource allocation for OFDMA networks with imperfect CSI," *IEEE Trans. Veh. Technol.*, vol. 59, no. 5, pp. 2394–2403, Jun. 2010.

[17] S. Ye, R. Blum, and L. Cimini, "Adaptive OFDM systems with imperfect channel state information," *IEEE Trans. Wireless Commun.*, vol. 5, no. 11, pp. 3255–3265, Nov. 2006.

[18] Y. Yao and G. Giannakis, "Rate-maximizing power allocation in OFDM based on partial channel knowledge," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1073–1083, May 2005.

[19] M. Jeruchim, P. Balaban, and K. Shanmugan, *Simulation of Communication Systems*. New York: Plenum, 1992.

Achieving the Outage Capacity of the Diamond Relay Network to Within One Bit and Even Less

Chi Wan Sung, *Member, IEEE*, Mingjun Dai, and Ping Hu

***Abstract*—A new forwarding strategy is proposed for the wireless diamond relay network under slow fading. The key feature is that it can adaptively switch between decode–forward and compress–forward according to the instantaneous received signal strength. It has four control parameters, which can be optimized by an alternating optimization procedure. Its outage performance is compared with two lower bounds. Analytically, it is proven to achieve the outage capacity to within 1 bit and within 50% for any signal-to-noise ratio (SNR). Empirically, it is shown to be nearly optimal for some fading scenarios.**

***Index Terms*—Alamouti code, diamond relay network, outage capacity.**

I. Introduction

The parallel relay network, in which a pair of source–destination nodes is connected by a number of parallel relay nodes, was proposed in [1]. Due to its potential application to mobile cellular systems, it has attracted much attention [2]–[6]. The particular case where there are two relays is called the diamond relay network, as the topology looks like a diamond, as shown in Fig. 1. This model was first studied in [7] and later investigated in [8]–[11]. Its channel capacity, however, remains unknown, even for the nonfading case with only additive white Gaussian noise (AWGN).

In a slow-fading environment with channel state information (CSI) available at receivers but not at transmitters, full diversity of the parallel relay network can be achieved by using distributed space-time code, together with the decode-and-forward (DF) method [2]. For the diamond relay network, one can adopt the Alamouti code, and this particular scheme is called ACDF [11]. The full-diversity result ensures that the outage probability curve of ACDF can be made the steepest possible at high signal-to-noise ratio (SNR). On the other hand, it does not rule out the possibility that the curve can be

Manuscript received February 10, 2011; revised July 7, 2011; accepted July 20, 2011. Date of publication July 29, 2011; date of current version October 20, 2011. This work was supported by the City University of Hong Kong under Grant SRG 7002703. The review of this paper was coordinated by Prof. H. H. Nguyen.

C. W. Sung and M. Dai (Corresponding author) are with the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong (e-mail: albert.sung@cityu.edu.hk; mingjudai2@student.cityu.edu.hk).

P. Hu was with the City University of Hong Kong, Kowloon, Hong Kong. She is now with the Alcatel-Lucent Shanghai Bell Company, Shanghai 201206, China (e-mail: hping2@hotmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2011.2163327