A Dual Stochastic DFP algorithm for Optimal Resource Allocation in Wireless Systems

Aryan Mokhtari and Alejandro Ribeiro

***Abstract*—A stochastic implementation of the Davidon-Fletcher-Powell (DFP) quasi-Newton method to minimize dual functions of optimal resource allocation problems in wireless systems is introduced. While the use of dual stochastic gradient descent algorithms is widespread, they suffer from slow convergence rate. Application of second order methods, on the other hand, is impracticable because computation of dual Hessian inverses incurs excessive cost. The proposed method utilizes stochastic gradients in lieu of deterministic gradients for both, the determination of descent directions and the approximation of the dual function's curvature. Since stochastic gradients can be computed at manageable computational cost stochastic DFP is realizable and retains the convergence rate advantages of its deterministic counterparts. Convergence results show that lower and upper bounds on the instantaneous form of the dual Hessian are sufficient to guarantee convergence to a small neighborhood of optimality. Numerical experiments illustrate that for ill conditioned dual functions stochastic DFP outperforms stochastic gradient descent by an order of magnitude.**

I. Introduction

This paper develops a stochastic version of the Davidon-Fletcher-Powell (DFP) quasi-Newton method to solve optimal resource allocation problems in wireless systems. In particular, we consider wireless communication systems characterized by a vector block fading channel $\mathbf{h} \in \mathcal{H}$ and corresponding resource allocation vectors $\mathbf{p}(\mathbf{h}) \in \mathcal{P}(\mathbf{h})$. Allocation of $\mathbf{p}(\mathbf{h})$ units of resource when the channel coefficient is $\mathbf{h}$ results in instantaneous performance $\mathbf{f}(\mathbf{h}, \mathbf{p}(\mathbf{h}))$. We want to find an optimal resource allocation function $\mathbf{p}^* := \{\mathbf{p}^*(\mathbf{h})\}_{\mathbf{h}\in\mathcal{H}}$ that optimizes a given utility of the ergodic performance $\mathbf{x} = \mathbb{E}[\mathbf{f}(\mathbf{h}, \mathbf{p}(\mathbf{h}))]$. Formally, consider a compact convex set $\mathcal{X}$; a set of compact, not necessarily convex, sets $\mathcal{P}(\mathbf{h})$; a concave function $f_0(\mathbf{x})$; and bounded functions $\mathbf{f}(\mathbf{h}, \mathbf{p}(\mathbf{h}))$. The pair of optimal ergodic performance $\mathbf{x}^*$ and associated optimal resource allocation $\mathbf{p}^*$ is defined as

$$(\mathbf{x}^*, \mathbf{p}^*) := \begin{array}{ll} \operatorname{argmax} & f_0(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} = \mathbb{E}[\mathbf{f}(\mathbf{p}(\mathbf{h}), \mathbf{h})], \quad \mathbf{x} \in \mathcal{X}, \quad \mathbf{p} \in \mathcal{P}, \end{array} \tag{1}$$

where $\mathcal{P} := \{\mathbf{p} : \mathbf{p}(\mathbf{h}) \in \mathcal{P}(\mathbf{h})\}_{\mathbf{h}\in\mathcal{H}}$ is the set of feasible resource allocation functions $\mathbf{p}$. Problems with the structure in (1) have null duality gap even if the functions $\mathbf{f}(\mathbf{p}(\mathbf{h}), \mathbf{h})$ are not concave or the sets $\mathcal{P}(\mathbf{h})$ are not convex [1], [2]. Lack of duality gap allows solution in the dual domain where the separable structure of the Lagrangian results in reduced complexity.

The simplest particular instance of (1) is power allocation in a point to point wireless channel. For each fading state $\mathbf{h}$ we allocate power $\mathbf{p}(\mathbf{h})$. The function $\mathbf{f}(\mathbf{p}(\mathbf{h}), \mathbf{h})$ represents the stacking of the instantaneous rate and instantaneous power consumption. The variable $\mathbf{x}$ represents the ergodic capacity and average power and our goal is to maximize a utility $f_0(\mathbf{x})$ of the ergodic capacity and average rate. The set $\mathcal{X}$ represents a restriction on allowable ergodic capacities and powers and the sets $\mathcal{P}(\mathbf{h})$ represent constraints on instantaneous powers. The fundamental interplay of instantaneous resource allocation conducive to desirable ergodic performance is widespread in wireless systems. Examples more interesting than point-to-point communication include optimization of orthogonal frequency division multiplexing [3], beamforming [4], cognitive radio [5], random access [6], communication with imperfect channel state information [7], and various flavors of wireless network optimization [8]–[10].

Work supported by ARO W911NF-10-1-0388, NSF CAREER CCF-0952867, and ONR N00014-12-1-0997. The authors are with the Dept. of Electrical and Systems Engineering, University of Pennsylvania, 200 South 33rd Street, Philadelphia, PA 19104. Email: {aryanm, aribeiro}@seas.upenn.edu.

In most of the problems considered in [3]–[10] it is possible to compute stochastic gradients of the dual function with manageable complexity. It is therefore possible, even desirable, to implement a dual stochastic gradient descent algorithm to find optimal dual variables – with optimal primal variables obtained as a byproduct of this implementation [11]. Practical appeal of dual stochastic gradient descent remains limited, however, because the convergence rate of these first order algorithms is slow. Resort to second order methods, on the other hand, is of little use because the computational cost of computing the dual Hessians' inverses necessary to find Newton steps is prohibitive. Quasi-Newton methods arise as the natural alternative because they rely on gradients to compute curvature estimates thereby achieving superlinear convergece rates in deterministic settings [12]–[15]. Since gradients can be estimated by stochastic gradients at manageable computational cost stochastic generalizations of quasi-Newton methods are realizable and expected to retain the convergence rate advantages of their deterministic counterparts.

In this paper we propose a stochastic version of the DFP quasi-Newton method that operates in the dual domain to find solutions to problems with the generic structure in (1). We begin with a discussion of stochastic gradient descent algorithms (Section II) and move on to introduce the DFP method as well as a regularized version that results in Hessian approximations with more amenable spectral properties (Section II-A). This regularized version is leveraged to introduce the dual stochastic DFP algortihm (Section II-B). Stochastic DFP differs from regular DFP in the use of a regularization and on the use of stochastic gradients in lieu of deterministic gradients for both, the determination of descent directions and the approximation of the dual function's curvature. Convergence results are then presented to show that lower and upper bounds on the instantaneous form of the dual Hessian are sufficient to guarantee convergence to a small neighborhood of optimality (Section III). Simulation results are presented for a simple frequency division multiple access channel with two users (Section IV). For well conditioned dual functions stochastic DFP and gradient descent exhibit comparable performance. For ill conditioned dual functions, however, stochastic DFP outperforms stochastic gradient descent by an order of magnitude.

II. Problem Formulation

To solve (1) we work in the dual domain. For that purpose introduce the vector Lagrange multiplier $\boldsymbol{\lambda}$ and define the Lagrangian associated with the optimization problem in (1) as

$$\mathcal{L}(\mathbf{x}, \mathbf{p}, \boldsymbol{\lambda}) := f_0(\mathbf{x}) + \boldsymbol{\lambda}^T \Big[\mathbb{E}\left[\mathbf{f}(\mathbf{p}(\mathbf{h}), \mathbf{h})\right] - \mathbf{x}\Big] \tag{2}$$

For the methods derived here the concept of Lagrangian maximizer is important

$$(\mathbf{x}(\boldsymbol{\lambda}), \mathbf{p}(\boldsymbol{\lambda})) := \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}, \mathbf{p}\in\mathcal{P}} \mathcal{L}(\mathbf{x}, \mathbf{p}(\mathbf{h}), \boldsymbol{\lambda}). \tag{3}$$

Because the structure of the Lagrangian is separable, the Lagrangian maximizers in 3 can be computed as

$$\mathbf{x}(\boldsymbol{\lambda}) = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}} \; f_0(\mathbf{x}) - \boldsymbol{\lambda}^T \mathbf{x} \tag{4}$$

$$\mathbf{p}(\mathbf{h}, \boldsymbol{\lambda}) = \operatorname*{argmax}_{\mathbf{p}(\mathbf{h})\in\mathcal{P}(\mathbf{h})} \boldsymbol{\lambda}^T \mathbf{f}(\mathbf{p}(\mathbf{h}), \mathbf{h}). \tag{5}$$

978-1-4673-5577-3/13/$31.00 ©2013 IEEE

Algorithm 1 Computation of stochastic gradients

1: **function** STOCHASTIC GRADIENT($\boldsymbol{\lambda}$, $\tilde{\mathbf{h}} = [\mathbf{h}_1; ...; \mathbf{h}_L]$)
2: Determine ergodic Lagrangian maximizers [cf. (4)]

$$\mathbf{x} = \mathbf{x}(\boldsymbol{\lambda}) = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}} f_0(\mathbf{x}) - \boldsymbol{\lambda}^T\mathbf{x}$$

3: Determine resource allocations for all $\mathbf{h}_l$ [cf. (5)]

$$\mathbf{p}(\mathbf{h}_l, \boldsymbol{\lambda}) = \operatorname*{argmax}_{\mathbf{p}(\mathbf{h}_l)\in\mathcal{P}(\mathbf{h}_l)} \boldsymbol{\lambda}^T\mathbf{f}(\mathbf{p}(\mathbf{h}_l), \mathbf{h}_l)$$

4: **return** stochastic gradient [cf. (9)]

$$\hat{\mathbf{s}}(\boldsymbol{\lambda}, \tilde{\mathbf{h}}) = \frac{1}{L}\sum_{l=1}^{L} \mathbf{f}(\mathbf{p}(\mathbf{h}_l, \boldsymbol{\lambda}), \mathbf{h}_l) - \mathbf{x}(\boldsymbol{\lambda})$$

5: **end function**

Observe that (5) is not convex, which in practice restricts applicability to problems with a moderate number of terminals. The dual function is then defined as

$$g(\boldsymbol{\lambda}) = \mathcal{L}(\mathbf{x}(\boldsymbol{\lambda}), \mathbf{p}(\boldsymbol{\lambda}), \boldsymbol{\lambda}). \tag{6}$$

The dual problem for (1) is defined as a minimization of the dual function over all positive dual variables, i.e.,

$$D = \min_{\boldsymbol{\lambda}\geq 0} g(\boldsymbol{\lambda}) \tag{7}$$

It has been proven that the duality gap for problem (1) is zero and the optimal value of the primal function and dual function are equal. Therefore, we have the ability to solve problem (7) instead of solving problem (1). Since the dual function in (6) is finite dimensional, its minimization is simpler than the solution of the infinite dimensional primal problem in (1). Given that the duality gap is null, optimal primal arguments can be computed as $\mathbf{x}^* = \mathbf{x}(\boldsymbol{\lambda}^*)$, $\mathbf{p}^* = \mathbf{p}(\boldsymbol{\lambda}^*)$ once an optimal dual variable $\boldsymbol{\lambda}^* = \operatorname{argmin} g(\boldsymbol{\lambda})$ is available.

Since the dual function is convex, descent algorithms can be used for its minimization. This is not difficult in principle because gradients $\nabla g(\boldsymbol{\lambda})$ of the dual function at $\boldsymbol{\lambda}$ are given by the constraint slack corresponding to the Lagrangian maximizers associated with $\boldsymbol{\lambda}$,

$$\mathbf{s}(\boldsymbol{\lambda}) := \nabla g(\boldsymbol{\lambda}) = \mathbb{E}[\mathbf{f}(\mathbf{p}(\mathbf{h}, \boldsymbol{\lambda}), \mathbf{h})] - \mathbf{x}(\boldsymbol{\lambda}). \tag{8}$$

However, exact evaluation of the gradient in (8) is not possible in general because it requires evaluation of the expectation $\mathbb{E}[\mathbf{f}(\mathbf{h}, \mathbf{p}(\mathbf{h}, \boldsymbol{\lambda}))]$ which depends on the Lagrangian maximizing function $\mathbf{p}(\boldsymbol{\lambda})$ obtained by solving the program in (5) and in most cases not available in closed form.

This difficulty is overcome by using stochastic gradients in lieu of the actual gradients in (8). To do so consider a given set of L channel realizations $\tilde{\mathbf{h}} = [\mathbf{h}_1; ...; \mathbf{h}_L]$ and define the stochastic gradient

$$\hat{\mathbf{s}}(\boldsymbol{\lambda}, \tilde{\mathbf{h}}) = \frac{1}{L}\sum_{l=1}^{L} \mathbf{f}(\mathbf{p}(\mathbf{h}_l, \boldsymbol{\lambda}), \mathbf{h}_l) - \mathbf{x}(\boldsymbol{\lambda}). \tag{9}$$

To compute the stochastic gradient $\hat{\mathbf{s}}(\boldsymbol{\lambda}, \tilde{\mathbf{h}})$ we solve the maximization in (4) to determine $\mathbf{x}(\boldsymbol{\lambda})$, L problems of the form in (5) for channels $\mathbf{h}_1, \ldots, \mathbf{h}_L$ to determine the resource allocations $\mathbf{p}(\mathbf{h}_l, \boldsymbol{\lambda})$, and proceed to evaluate the sum in (9); see Algorithm 1.

The dual stochastic gradient descent algorithm consists of recursive computation of primal Lagrangian maximizers $\mathbf{x}(\boldsymbol{\lambda}_t)$ and $\mathbf{p}(\mathbf{h}_{t,l}, \boldsymbol{\lambda}_t)$ for given dual iterate λ_t and channel samples $\tilde{\mathbf{h}}_t$, followed by evaluation of the stochastic gradient $\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$, and the projected descent step

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \epsilon\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t). \tag{10}$$

As long the step size ϵ is judiciously selected and the channel samples $\tilde{\mathbf{h}}_{t,l}$ are drawn independently from the channel probability distribution the sequence of dual variables generated by (10) approaches the optimal multiplier $\boldsymbol{\lambda}^*$ from where convergence of primal iterates follows, see e.g., [2], [11]. However, the convergence rate is slow, motivating the introduction of the stochastic DFP algorithm that we describe in the following section.

A. Regularized DFP

To speed up convergence of (10) resort to second order methods is of little use because evaluating Hessians of the dual function is computationally intensive. A better suited methodology is the use of quasi-Newton methods whereby gradient descent directions are premultiplied by a matrix $\mathbf{B}_t^{-1}$,

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \epsilon\mathbf{B}_t^{-1}\mathbf{s}(\boldsymbol{\lambda}_t). \tag{11}$$

The idea is to select matrices $\mathbf{B}_t$ close to the dual Hessian $\mathbf{H}(\boldsymbol{\lambda}_t) := \nabla^2 g(\boldsymbol{\lambda}_t)$. Various methods are known to select matrices $\mathbf{B}_t$, including those by Broyden e.g., [12]; Broyden, Fletcher, Goldfarb, and Shanno (BFGS) e.g., [15]; and Davidon, Fletcher, and Powell (DFP) e.g., [12]. We work here with the matrices $\mathbf{B}_t$ used in the DFP method.

In DFP – and all other quasi Newton methods for that matter – the function's curvature is approximated by a finite difference. Specifically, define the variable and gradient variations at time t as

$$\boldsymbol{\mu}_t = \boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t, \quad \mathbf{r}_t = \mathbf{s}(\boldsymbol{\lambda}_{t+1}) - \mathbf{s}(\boldsymbol{\lambda}_t), \tag{12}$$

respectively. We select the matrix $\mathbf{B}_{t+1}$ to be used in the next time step so that it satisfies the secant condition $\mathbf{B}_{t+1}\boldsymbol{\mu}_t = \mathbf{r}_t$. The rationale for this selection is that the Hessian $\mathbf{H}(\boldsymbol{\lambda}_t)$ satisfies this condition for vanishing $\boldsymbol{\mu}_t$, i.e., for $\boldsymbol{\lambda}_{t+1}$ tending to $\boldsymbol{\lambda}_t$.

Notice however that the secant condition $\mathbf{B}_{t+1}\boldsymbol{\mu}_t = \mathbf{r}_t$ is not enough to completely specify $\mathbf{B}_{t+1}$. Part of this indeterminacy can be resolved by requiring matrices $\mathbf{B}_t \succeq \mathbf{0}$ to be (symmetric) positive semidefinite. To resolve the remaining indeterminacy we observe that it is reasonable to expect the Hessians $\mathbf{H}(\boldsymbol{\lambda}_t)$ and $\mathbf{H}(\boldsymbol{\lambda}_{t-1})$ to be close to each other and therefore select $\mathbf{B}_{t+1}$ as the closest matrix to the previous Hessian approximation $\mathbf{B}_t$ among all those that satisfy the secant condition $\mathbf{B}_{t+1}\boldsymbol{\mu}_t = \mathbf{r}_t$. Closeness between $\mathbf{B}_t$ and $\mathbf{B}_{t+1}$ is usually specified in terms of a weighted Frobenius norm. For the purposes of this paper it is better to specify closeness in terms of the relative Gaussian entropy and define the matrix $\mathbf{B}_{t+1}$ as the solution of the semidefinite program

$$\begin{aligned} \mathbf{B}_{t+1} \;=\; & \operatorname{argmin} \;\; \operatorname{tr}(\mathbf{B}_t\mathbf{X}^{-1}) - \log\det(\mathbf{B}_t\mathbf{X}^{-1}) - n, \\ & \text{s. t.} \qquad \boldsymbol{\mu}_t = \mathbf{X}^{-1}\mathbf{r}_t, \quad \mathbf{X} \succeq \mathbf{0}. \end{aligned} \tag{13}$$

The constraint $\mathbf{X} \succeq \mathbf{0}$ restricts the feasible space to positive semidefinite matrices whereas the constraint $\boldsymbol{\mu}_t = \mathbf{X}^{-1}\mathbf{r}_t$ requires $\mathbf{X}$ to satisfy the secant condition. The objective $\operatorname{tr}(\mathbf{B}_t\mathbf{X}^{-1}) - \log\det(\mathbf{B}_t\mathbf{X}^{-1}) - n$ represents the differential entropy between random variables with zero-mean Gaussian distributions $\mathcal{N}(0, \mathbf{B}_t)$ and $\mathcal{N}(0, \mathbf{X})$ having covariance matrices $\mathbf{B}_t$ and $\mathbf{X}$. The differential entropy is nonnegative and equal to zero if and only if $\mathbf{X} = \mathbf{B}_t$. The solution $\mathbf{B}_{t+1}$ of the semidefinite program in (13) is therefore closest to $\mathbf{B}_t$, in the sense of minimizing the Gaussian differential entropy, among all positive semidefinite matrices that satisfy the secant condition $\boldsymbol{\mu}_t = \mathbf{X}^{-1}\mathbf{r}_t$.

If the variations $\boldsymbol{\mu}_t$ and $\mathbf{r}_t$ are such that $\boldsymbol{\mu}_t^T\mathbf{r}_t > 0$ the semidefinite program in (13) has a unique solution that is explicitly given by

$$\mathbf{B}_{t+1} = \left(\mathbf{I} - \frac{\mathbf{r}_t\boldsymbol{\mu}_t^T}{\boldsymbol{\mu}_t^T\mathbf{r}_t}\right)\mathbf{B}_t\left(\mathbf{I} - \frac{\boldsymbol{\mu}_t\mathbf{r}_t^T}{\boldsymbol{\mu}_t^T\mathbf{r}_t}\right) + \frac{\mathbf{r}_t\mathbf{r}_t^T}{\boldsymbol{\mu}_t^T\mathbf{r}_t}. \tag{14}$$

The Hessian approximation update in (14) is the one most often associated with the definition of DFP [15].

It follows from (14) that $\mathbf{B}_t^{-1}$ stays positive definite for all iterations t as long as the initial matrix is $\mathbf{B}_1^{-1} \succ \mathbf{0}$ positive definite and the variations satisfy $\boldsymbol{\mu}_t^T\mathbf{r}_t > 0$. However, it is possible for the smallest eigenvalue of $\mathbf{B}_t^{-1}$ to become arbitrarily close to zero. This has proven not to be an issue in DFP implementations but is a more significant challenge in the stochastic version proposed here. To avoid this problem we introduce a

regularization of (13) that requires the minimum eigenvalue of $\mathbf{B}_{t+1}^{-1}$ to be at least δ,

$$\begin{aligned}\mathbf{B}_{t+1} = \operatorname{argmin}\ & \operatorname{tr}[\mathbf{B}_t(\mathbf{X}^{-1}-\delta\mathbf{I})]-\log\det[\mathbf{B}_t(\mathbf{X}^{-1}-\delta\mathbf{I})]-n, \\ \text{s.t.}\quad & \boldsymbol{\mu}_t = \mathbf{X}^{-1}\mathbf{r}_t, \quad \mathbf{X}\succeq\mathbf{0}. \end{aligned} \tag{15}$$

Since the negative logarithm determinant $-\log\det[\mathbf{B}_t(\mathbf{X}^{-1}-\delta\mathbf{I})]$ diverges as the smallest eigenvalue of $\mathbf{X}^{-1}$ approaches δ, the smallest eigenvalue of the Hessian inverse approximation matrices $\mathbf{B}_{t+1}^{-1}$ computed as solutions of (15) exceed the lower bound δ. In the following lemma we show that the regularized approximations in (15) can be computed by a formula similar to the update in (14)[1].

Lemma 1 *Consider the semidefinite program in* (15) *where the matrix* $\mathbf{B}_t \succ \mathbf{0}$ *is positive definite and the inner product* $(\boldsymbol{\mu}_t - \delta\mathbf{r}_t)^T\mathbf{r}_t > 0$. *Then, the solution* $\mathbf{B}_{t+1}$ *of* (15) *satisfies*

$$(\mathbf{B}_{t+1}^{-1}-\delta\mathbf{I})^{-1} = \left(\mathbf{I}-\frac{\mathbf{r}_t\tilde{\boldsymbol{\mu}}_t^T}{\tilde{\boldsymbol{\mu}}_t^T\mathbf{r}_t}\right)\mathbf{B}_t\left(\mathbf{I}-\frac{\tilde{\boldsymbol{\mu}}_t\mathbf{r}_t^T}{\tilde{\boldsymbol{\mu}}_t^T\mathbf{r}_t}\right)+\frac{\mathbf{r}_t\mathbf{r}_t^T}{\tilde{\boldsymbol{\mu}}_t^T\mathbf{r}_t}, \tag{16}$$

where $\tilde{\boldsymbol{\mu}}_t := \boldsymbol{\mu}_t - \delta\mathbf{r}_t$ *is the corrected variable variation.*

The expression in (16) permits efficient computation of the regularized Hessian approximations $\mathbf{B}_{t+1}$. To implement the descent step (11) the approximation $\mathbf{B}_{t+1}$ needs to be inverted. This inversion can be avoided by using the Sherman-Morrison formula in (16) to write

$$\mathbf{B}_{t+1}^{-1} = \mathbf{B}_t^{-1} + \frac{\tilde{\boldsymbol{\mu}}_t\tilde{\boldsymbol{\mu}}_t^T}{\mathbf{r}_t^T\tilde{\boldsymbol{\mu}}_t} - \frac{\mathbf{B}_t^{-1}\mathbf{r}_t\mathbf{r}_t^T\mathbf{B}_t{}^{-1}}{\mathbf{r}_t^T\mathbf{B}_t^{-1}\mathbf{r}_t} + \delta\mathbf{I}. \tag{17}$$

When $\delta = 0$ the update in (17) coincides with standard non-regularied DFP. Therefore, the differences between DFP and regularized DFP are the replacement of the variable variation $\boldsymbol{\mu}_t$ in (12) by the corrected variation $\tilde{\boldsymbol{\mu}}_t := \boldsymbol{\mu}_t - \delta\mathbf{r}_t$ and the addition of the regularization term $\delta\mathbf{I}$. We use (17) in the construction of the stochastic DFP algorithm in the following section.

B. Stochastic DFP

As can be seen from (16) the regularized DFP curvature estimate $\mathbf{B}_{t+1}$ is obtained as a function of previous estimates $\mathbf{B}_t$, dual iterates $\boldsymbol{\lambda}_t$ and $\boldsymbol{\lambda}_{t+1}$, and corresponding gradients $\mathbf{s}(\boldsymbol{\lambda}_t)$ and $\mathbf{s}(\boldsymbol{\lambda}_{t+1})$. We can then think of a method in which gradients $\mathbf{s}(\boldsymbol{\lambda}_t)$ are replaced by stochastic gradients $\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$ in both, the curvature approximation update in (16) and the descent iteration in (11). These substitutions lead to the dual stochastic DFP algorithm that we introduce in the following.

Start at time t with current dual iterate $\boldsymbol{\lambda}_t$ and let $\hat{\mathbf{B}}_t$ stand for the Hessian approximation computed by stochastic DFP in the previous iteration. Obtain a batch of channel samples $\tilde{\mathbf{h}}_t = [\mathbf{h}_{t,1}; ...; \mathbf{h}_{t,L}]$ and for each of the $\mathbf{h}_{t,l}$ samples determine the values $\mathbf{p}(\mathbf{h}_{t,l}, \boldsymbol{\lambda}_t)$ of the Lagrangian maximizer resource allocation function associated with $\mathbf{h}_{t,l}$ as per (5). Further determine the ergodic Lagrangian maximizers $\mathbf{x}(\boldsymbol{\lambda}_t)$ as per (4) and evaluate the stochastic gradient $\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$ as per (9) with $\mathbf{x}(\boldsymbol{\lambda}) = \mathbf{x}(\boldsymbol{\lambda}_t)$ and $\mathbf{p}(\mathbf{h}_l, \boldsymbol{\lambda}) = \mathbf{p}(\mathbf{h}_{t,l}, \boldsymbol{\lambda}_t)$. Descend then on the dual function along the direction $\hat{\mathbf{B}}_t^{-1}\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$ moderated by the stepsize ϵ

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \epsilon\hat{\mathbf{B}}_t^{-1}\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t). \tag{18}$$

For the multiplier iterate $\boldsymbol{\lambda}_{t+1}$ compute the stochastic gradient $\hat{\mathbf{s}}(\boldsymbol{\lambda}_{t+1}, \tilde{\mathbf{h}}_t)$ associated with the *same* set of channel samples $\tilde{\mathbf{h}}_t$ used to compute the stochastic gradient $\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$. This requires determination of ergodic variables $\mathbf{x}(\boldsymbol{\lambda}_{t+1})$ as per (4), resource allocations $\mathbf{p}(\mathbf{h}_{t,l}, \boldsymbol{\lambda}_{t+1})$ as per (5), and evaluation of the stochastic gradient $\hat{\mathbf{s}}(\boldsymbol{\lambda}_{t+1}, \tilde{\mathbf{h}}_t)$ as per (9) with $\mathbf{x}(\boldsymbol{\lambda}) = \mathbf{x}(\boldsymbol{\lambda}_{t+1})$ and $\mathbf{p}(\mathbf{h}_l, \boldsymbol{\lambda}) = \mathbf{p}(\mathbf{h}_{t,l}, \boldsymbol{\lambda}_{t+1})$. Define then stochastic gradient variation at time t as

$$\hat{\mathbf{r}}_t = \hat{\mathbf{s}}(\boldsymbol{\lambda}_{t+1}, \tilde{\mathbf{h}}_t) - \hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t), \tag{19}$$

[1]Proofs are available in [16]

Algorithm 2 Stochastic DFP

Require: Dual variable $\boldsymbol{\lambda}_1$. Hessian approximation $\hat{\mathbf{B}}_1^{-1} \succ \delta\mathbf{I}$.

1: **for** $t = 1, 2, \ldots$ **do**
2: Acquire L independent channel samples $\tilde{\mathbf{h}}_t = [\mathbf{h}_{t,1}, \ldots, \mathbf{h}_{t,L}]$
3: Compute $\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$ = STOCHASTIC GRADIENT($\boldsymbol{\lambda}_t$, $\tilde{\mathbf{h}}_t$)
4: Descend along direction $\hat{\mathbf{B}}_t^{-1}\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$ [cf. (18)]

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \epsilon\hat{\mathbf{B}}_t^{-1}\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t).$$

5: Compute $\hat{\mathbf{s}}(\boldsymbol{\lambda}_{t+1}, \tilde{\mathbf{h}}_t)$ = STOCHASTIC GRADIENT($\boldsymbol{\lambda}_{t+1}$, $\tilde{\mathbf{h}}_t$)
6: Compute stochastic gradient variation [cf. (19)]

$$\hat{\mathbf{r}}_t = \hat{\mathbf{s}}(\boldsymbol{\lambda}_{t+1}, \tilde{\mathbf{h}}_t) - \hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$$

7: Compute modified variable variation [cf. (20)]

$$\tilde{\boldsymbol{\mu}}_t = (\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t) - \delta\hat{\mathbf{r}}_t.$$

8: Update approximation of Hessian inverse [cf. (22)]

$$\hat{\mathbf{B}}_{t+1}^{-1} = \hat{\mathbf{B}}_t^{-1} + \frac{\tilde{\boldsymbol{\mu}}_t\tilde{\boldsymbol{\mu}}_t^T}{\hat{\mathbf{r}}_t^T\tilde{\boldsymbol{\mu}}_t} - \frac{\hat{\mathbf{B}}_t^{-1}\hat{\mathbf{r}}_t\hat{\mathbf{r}}_t^T\hat{\mathbf{B}}_t^{-1}}{\hat{\mathbf{r}}_t^T\hat{\mathbf{B}}_t^{-1}\hat{\mathbf{r}}_t} + \delta\mathbf{I}.$$

9: **end for**

as well as the modified dual variable variation

$$\tilde{\boldsymbol{\mu}}_t = \boldsymbol{\mu}_t - \delta\hat{\mathbf{r}}_t = (\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t) - \delta\hat{\mathbf{r}}_t. \tag{20}$$

The Hessian approximation $\hat{\mathbf{B}}_{t+1}$ for the next iteration is defined as the matrix that satisfies the stochastic secant condition $\boldsymbol{\mu}_t = \hat{\mathbf{B}}_{t+1}^{-1}\hat{\mathbf{r}}_t$ and is closest to $\hat{\mathbf{B}}_t$ in the sense of (15). As per Lemma 1 we can then compute $\hat{\mathbf{B}}_{t+1}$ explicitly as the matrix that satisfies

$$(\hat{\mathbf{B}}_{t+1}^{-1}-\delta\mathbf{I})^{-1} = \left(\mathbf{I}-\frac{\hat{\mathbf{r}}_t\tilde{\boldsymbol{\mu}}_t^T}{\tilde{\boldsymbol{\mu}}_t^T\hat{\mathbf{r}}_t}\right)\hat{\mathbf{B}}_t\left(\mathbf{I}-\frac{\tilde{\boldsymbol{\mu}}_t\hat{\mathbf{r}}_t^T}{\tilde{\boldsymbol{\mu}}_t^T\hat{\mathbf{r}}_t}\right)+\frac{\hat{\mathbf{r}}_t\hat{\mathbf{r}}_t^T}{\tilde{\boldsymbol{\mu}}_t^T\hat{\mathbf{r}}_t}, \tag{21}$$

as long as $(\boldsymbol{\mu}_t - \delta\hat{\mathbf{r}}_t)^T\hat{\mathbf{r}}_t = \tilde{\boldsymbol{\mu}}^T\hat{\mathbf{r}}_t > 0$. Conditions to guarantee that $\tilde{\boldsymbol{\mu}}^T\mathbf{r}_t > 0$ are introduced in Section III. The expression in (21) is used in the convergence analysis in Section III. For practical implementation we use the Sherman-Morrison formula in (16) to write the analogous of (17),

$$\hat{\mathbf{B}}_{t+1}^{-1} = \hat{\mathbf{B}}_t^{-1} + \frac{\tilde{\boldsymbol{\mu}}_t\tilde{\boldsymbol{\mu}}_t^T}{\hat{\mathbf{r}}_t^T\tilde{\boldsymbol{\mu}}_t} - \frac{\hat{\mathbf{B}}_t^{-1}\hat{\mathbf{r}}_t\hat{\mathbf{r}}_t^T\hat{\mathbf{B}}_t^{-1}}{\hat{\mathbf{r}}_t^T\hat{\mathbf{B}}_t^{-1}\hat{\mathbf{r}}_t} + \delta\mathbf{I}. \tag{22}$$

The dual stochastic DFP algorithm is summarized in Algorithm II-B. The two core steps in each iteration are the dual descent in Step 4 and the update of the Hessian approximation inverse $\hat{\mathbf{B}}_t^{-1}$ in Step 8. Step 2 comprises the observation of L channel samples that are required to compute the stochastic gradients in steps 3 and 5. The stochastic gradient $\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$ in Step 3 is used in the descent iteration in Step 4 and is computed using the function in Algorithm II. The stochastic gradient of Step 3 along with the stochastic gradient $\hat{\mathbf{s}}(\boldsymbol{\lambda}_{t+1}, \tilde{\mathbf{h}}_t)$ of Step 5 are used to compute the variations in steps 6 and 7 that permit carrying out the update of the Hessian approximation inverse $\hat{\mathbf{B}}_t^{-1}$ in Step 8. Iterations are initialized at arbitrary nonnegative multiplier $\boldsymbol{\lambda}_1$ and positive definite matrix $\hat{\mathbf{B}}_1^{-1}$ with the smallest eigenvalue larger than δ as indicated in Step 1.

Remark 1 One may think that the natural substitution of the gradient variation $\mathbf{r}_t = \mathbf{s}(\boldsymbol{\lambda}_{t+1}) - \mathbf{s}(\boldsymbol{\lambda}_t)$ is the stochastic gradient variation $\hat{\mathbf{s}}(\boldsymbol{\lambda}_{t+1}, \tilde{\mathbf{h}}_{t+1}) - \hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$ instead of the one that we actually use $\hat{\mathbf{r}}_t = \hat{\mathbf{s}}(\boldsymbol{\lambda}_{t+1}, \tilde{\mathbf{h}}_t) - \hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$. This would have the advantage that $\hat{\mathbf{s}}(\boldsymbol{\lambda}_{t+1}, \tilde{\mathbf{h}}_{t+1})$ is the stochastic gradient used to descend in iteration $t+1$ whereas $\hat{\mathbf{s}}(\boldsymbol{\lambda}_{t+1}, \tilde{\mathbf{h}}_t)$ is not and is just computed for the purposes of updating $\mathbf{B}_t$. Therefore, using the variation $\hat{\mathbf{r}}_t = \hat{\mathbf{s}}(\boldsymbol{\lambda}_{t+1}, \tilde{\mathbf{h}}_t) - \hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$ requires twice as many primal maximizations as using the variation $\hat{\mathbf{s}}(\boldsymbol{\lambda}_{t+1}, \tilde{\mathbf{h}}_{t+1}) - \hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$. However, the use of the variation $\hat{\mathbf{r}}_t = \hat{\mathbf{s}}(\boldsymbol{\lambda}_{t+1}, \tilde{\mathbf{h}}_t) - \hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$ is necessary to ensure that $(\boldsymbol{\mu}_t - \delta\mathbf{r}_t)^T\mathbf{r}_t = \tilde{\boldsymbol{\mu}}^T\mathbf{r}_t > 0$. This is necessary for (21) to be true and cannot be guaranteed if we use the variation $\hat{\mathbf{s}}(\boldsymbol{\lambda}_{t+1}, \tilde{\mathbf{h}}_{t+1}) - \hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$. See Lemma (2) for details.

978-1-4673-5577-3/13/$31.00 ©2013 IEEE

III. Convergence

For the subsequent analysis it is convenient to define the instantaneous dual function associated with samples $\tilde{\mathbf{h}} = [\mathbf{h}_1, \ldots, \mathbf{h}_L]$

$$\hat{g}(\boldsymbol{\lambda}, \tilde{\mathbf{h}}) = \frac{1}{L}\sum_{l=1}^{L} f_0(\mathbf{x}(\boldsymbol{\lambda})) + \boldsymbol{\lambda}^T\Big[\mathbf{f}(\mathbf{p}(\mathbf{h}_l, \boldsymbol{\lambda}), \mathbf{h}_l) - \mathbf{x}(\boldsymbol{\lambda})\Big], \quad (23)$$

where $\mathbf{x}(\boldsymbol{\lambda})$ and $\mathbf{p}(\mathbf{h}_l, \boldsymbol{\lambda})$ are primal Lagrangian maximizers defined in (4). The definition in (23) is so that the dual function $g(\boldsymbol{\lambda})$ in (6) can be written as the expectation of the instantaneous functions $\hat{g}(\boldsymbol{\lambda}, \tilde{\mathbf{h}})$, i.e.,

$$g(\boldsymbol{\lambda}) = \mathbb{E}[\hat{g}(\boldsymbol{\lambda}, \mathbf{h})]. \quad (24)$$

Our goal here is to show that as time progresses the dual iterates $\boldsymbol{\lambda}_t$ approach the optimal multiplier $\boldsymbol{\lambda}^*$. In proving this result we make the following assumptions.

Assumption 1 The instantaneous functions $\hat{g}(\boldsymbol{\lambda}, \tilde{\mathbf{h}})$ are twice differentiable and the eigenvalues of the instantaneous dual Hessian $\hat{\mathbf{H}}(\boldsymbol{\lambda}, \tilde{\mathbf{h}}_t) = \nabla^2 \hat{g}(\boldsymbol{\lambda}, \mathbf{h})$ are bounded between constants $\tilde{m} > 0$ and $\tilde{M} < \infty$ for all channel realizations $\tilde{\mathbf{h}}$,

$$\tilde{m}\mathbf{I} \preceq \hat{\mathbf{H}}(\boldsymbol{\lambda}, \tilde{\mathbf{h}}_t) \preceq \tilde{M}\mathbf{I}. \quad (25)$$

Assumption 2 The second moment of the norm of the stochastic gradient is bounded for all $\boldsymbol{\lambda}$. I.e., there exists a constant $\hat{S}^2$ such that for all dual variables $\boldsymbol{\lambda}$ it holds

$$\mathbb{E}\big[\|\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)\|^2\big] \leq S^2, \quad (26)$$

Assumption 3 There exists a constant Γ that upper bounds all the eigenvalues of the inverse Hessian approximation matrix $\hat{\mathbf{B}}_t^{-1}$,

$$\hat{\mathbf{B}}_t^{-1} \preceq \Gamma\mathbf{I}. \quad (27)$$

As a consequence of Assumption 1 similar eigenvalue bounds hold for the (average) dual function $g(\boldsymbol{\lambda})$. Indeed, it follows from the linearity of the expectation operator and the expression in (24) that the dual Hessian is $\mathbf{H}(\boldsymbol{\lambda}) = \mathbb{E}[\hat{\mathbf{H}}(\boldsymbol{\lambda}, \tilde{\mathbf{h}})]$. Combining this observation with the bounds in (25) it follows that there are constants $m \geq \tilde{m}$ and $M \leq \tilde{M}$ such that

$$\tilde{m}\mathbf{I} \preceq m\mathbf{I} \preceq \mathbf{H}(\boldsymbol{\lambda}) \preceq M\mathbf{I} \preceq \tilde{M}\mathbf{I}. \quad (28)$$

The bounds in (28) are customary in convergence proofs of descent methods. For the results here the stronger condition spelled in Assumption 1 is needed. The restriction imposed by Assumption 2 is typical of stochastic descent algorithms, its intent being to limit the random variation of stochastic gradients. Assumption 3 is valid as long as the components of the matrices $\hat{\mathbf{B}}_t^{-1}$ stay bounded.

According to Lemma 1 the update in (21) is a solution to (15) – with the substitutions $\hat{\mathbf{B}}_t$ for $\mathbf{B}_t$ and $\boldsymbol{\mu}_t = \mathbf{X}^{-1}\hat{\mathbf{r}}_t$ for the secant condition $\boldsymbol{\mu}_t = \mathbf{X}^{-1}\mathbf{r}_t$ – as long as the inner product $(\boldsymbol{\mu}_t - \delta\hat{\mathbf{r}}_t)^T\hat{\mathbf{r}}_t = \tilde{\boldsymbol{\mu}}^T\mathbf{r}_t > 0$ is positive. Our first result is to show that selecting $\delta < 1/\tilde{M}$ guarantees that this inequality is satisfied for all times t.

Lemma 2 *Consider the stochastic gradient variation $\hat{\mathbf{r}}_t$ defined in* (19) *and the modified variable variation $\tilde{\boldsymbol{\mu}}_t$ defined in* (20)*. Let Assumption 1 hold and recall the upper bound $M_{\tilde{\mathbf{h}}} \leq \tilde{M}$ on the largest eigenvalue of the instantaneous Hessians. Then, for all constants $\delta < 1/\tilde{M}$ it holds*

$$\tilde{\boldsymbol{\mu}}^T\hat{\mathbf{r}}_t = (\boldsymbol{\mu}_t - \delta\hat{\mathbf{r}}_t)^T\hat{\mathbf{r}}_t > 0 \quad (29)$$

Initializing the curvature approximation matrix $\hat{\mathbf{B}}_1^{-1} \succ \delta\mathbf{I}$, which implies $\hat{\mathbf{B}}_1 \succ \mathbf{0}$, and setting $\delta < 1/\tilde{M}$ it follows from Lemma (2) that the hypotheses of Lemma (1) are satisfied for $t = 1$. Hence, the matrix $\hat{\mathbf{B}}_2^{-1}$ computed from (22) is the solution of the semidefinite program in (15) and as such satisfies $\hat{\mathbf{B}}_2^{-1} \succ \delta\mathbf{I}$, which in turn implies $\hat{\mathbf{B}}_2 \succ \mathbf{0}$. Proceeding recursively we can conclude that $\hat{\mathbf{B}}_t \succ \mathbf{0}$ and $\hat{\mathbf{B}}_t^{-1} \succ \delta\mathbf{I}$; i.e., that the minimum eigenvalue of $\hat{\mathbf{B}}_t^{-1}$ is at least δ for all times t. Observe that since the constant $\tilde{M}$ is not known in general we interpret the hypothesis $\delta < 1/\tilde{M}$ as requiring δ to be sufficiently small. Adaptive selection of δ will be used in practice.

Having matrices $\hat{\mathbf{B}}_t^{-1}$ that are strictly positive definite and not approaching the border of the set of positive semidefinite matrices leads to the conclusion that if $\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$ is a descent direction, the same holds true of $\hat{\mathbf{B}}_t^{-1}\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$. The stochastic gradient $\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$ is not a descent direction in general, but we know that its conditional expectation $\mathbb{E}[\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t) \,\big|\, \boldsymbol{\lambda}_t] = \nabla g(\boldsymbol{\lambda}_t)$ is a descent direction. Therefore, we conclude that $\hat{\mathbf{B}}_t^{-1}\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t)$ is an average descent direction because $\mathbb{E}[\hat{\mathbf{B}}_t^{-1}\hat{\mathbf{s}}(\boldsymbol{\lambda}_t, \tilde{\mathbf{h}}_t) \,\big|\, \boldsymbol{\lambda}_t] = \hat{\mathbf{B}}_t^{-1}\nabla g(\boldsymbol{\lambda}_t)$. Stochastic optimization algorithms whose displacements $\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t$ are descend directions on average are expected to approach optimal arguments. This is indeed true as we claim in the following theorem.

Theorem 1 *Consider the stochastic DFP algorithm as defined by* (18)-(22)*. If assumptions 1-3 hold true the limit infimum of the squared distance to optimality $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}^*\|^2$ satisfies*

$$\liminf_{t\to\infty} \|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}^*\|^2 \leq \frac{\epsilon M\Gamma^2 S^2}{2m^2\delta}, \quad (30)$$

with probability 1 over realizations of the channel samples $\{\tilde{\mathbf{h}}_t\}_{t=1}^{\infty}$.

Theorem 1 shows that the dual iterates $\boldsymbol{\lambda}_t$ converge to an area near the optimal vector $\boldsymbol{\lambda}^*$ with probability 1. The important observation about Theorem 1 is that the volume of the area to which $\boldsymbol{\lambda}_t$ converges depends on the step size ϵ. There is a tradeoff between speed of convergence and accuracy of convergence. If we decrease ϵ, we converge to a point closer to the optimal value $\boldsymbol{\lambda}^*$, but the speed of convergence decreases. If we increase ϵ we have faster convergence to a point farther away from the optimal argument $\boldsymbol{\lambda}^*$.

IV. Simulations

The goal of this section is comparing the performance of Stochastic DFP with Stochastic graidents in good-condition and ill-condition problems. As it was mentioned in section III, the Stochastic gradient's convergence rate depends on the condition number of problem, but the Stochastic DFP's convergence rate does not depend on the problem condition number. To explore this fact, we consider a FDMA channel problem. Consider a central access point (AP) administers tones $\mathcal{F}$ and average power budget P_0 to serve J terminals $T_i[1:J]$. The goal is to develop an algorithm that allocates power and frequency to maximize the given utility function. At time t AP observes the fading channels vector $\mathbf{h}_f = [h_{1f}, ..., h_{Jf}]^T$ for all frequencies $f \in \mathcal{F}$. Based on the fading channels vector AP determines the terminal should use channel f and power it can use for transmission. To formalize the problem we need to define a vector to show that which node has access to frequency f at each time. The frequency allocation vector $\boldsymbol{\alpha}_f(t) = [\alpha_{1f}(t), ..., \alpha_{Jf}(t)]^T$ which $a_{if}(t) \in \{0, 1\}$ determines which terminal can use the frequency f at time t. Variable $\alpha_{if}(t) = 1$ if and only if frequency f is allocated to node i at time t. If $\alpha_{if}(t) = 1$, the power allocated for such communication is $p_{if}(t)$. If channel i in communicating with AP at time t has channel $h_{if}(t)$ and power allocation $p_{if}(t)$, then the delivered information is $\log(1 + h_{if}(t)p_{if}(t)/N_0)$. If the amount of information that AP send to node i on Frequency f, the total amount of information that node i receives at time t is $\sum_{f\in\mathcal{F}} \alpha_{if}(t)\log(1 + h_{if}(t)p_{if}(t)/N_0)$. If $c_i(t)$ is the units of information that node i accepts for delivery at time t, then to guaranty delivery of packets it suffices to ensure stability of information queues by requiring

$$\lim_{t\to\infty}\frac{1}{t}\sum_{u=1}^{t} c_i(u) = \lim_{t\to\infty}\frac{1}{t}\sum_{u=1}^{t}[\sum_{f\in\mathcal{F}} \alpha_{if}(u)\log(1 + \frac{h_{if}(u)p_{if}(u)}{N_0})] \quad (31)$$

Similarly, the amount of power consumed at time t is the sum of power used on all the frequencies for communication with all terminals, i.e., $\sum_{i=1}^{J}\sum_{f\in\mathcal{F}} \alpha_{if}(u)p_{if}(t)$. As there is a limitation for the amount of

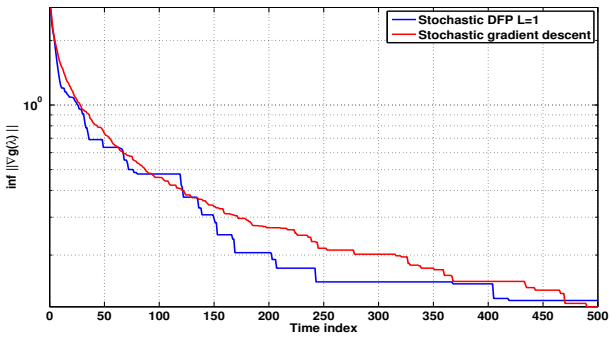

Fig. 1. Convergence of Stochastic gradient descent and Stochastic DFP for well conditioned problem.(L=1 in both algorithms)

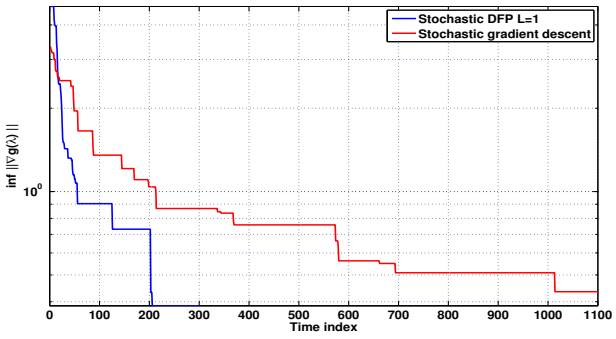

Fig. 2. Convergence of Stochastic gradient descent and Stochastic DFP for ill conditioned problem.(L=1 in both algorithms)

power that AP can consume, the average amount of power that AP uses must be less than or equal P_0.

$$P_0 = \lim_{t\to\infty} \frac{1}{t}\Big(\sum_{u=1}^{t}\Big[\sum_{i=1}^{J}\sum_{f\in\mathcal{F}} \alpha_{if}(u)p_{if}(u)\Big]\Big) \tag{32}$$

If we replace the time of average of variables with their ergodic limit, then the optimization problem can be represented as following

$$\begin{aligned} &\max \sum_i \log(c_i) \\ s.t. \quad &c_i = \mathbb{E}\Big[\sum_{f\in\mathcal{F}} \alpha_{if}(\mathbf{h}) \log\Big(1+\frac{h_{if}p_{if}(\mathbf{h})}{N_0}\Big)\Big], \\ &P_0 = \mathbb{E}\Big[\sum_{i=1}^{J}\sum_{f\in\mathcal{F}} \alpha_{if}(\mathbf{h})p_{if}(\mathbf{h})\Big]. \end{aligned} \tag{33}$$

This optimization problem is of the form of Problem 1. Defining the Lagrangian multipliers λ_i associated with capacity constraints and μ with the power constraint, the primal iteration which is the first step of Stochastic DFP will be

$$c_i(t) = \operatorname{argmax} \quad \log(c_i) - \lambda_i(t)c_i \tag{34}$$

$$p_{if}(t) = \frac{1}{L}\sum_{l=1}^{L} \operatorname*{argmax}_{i} \quad \lambda_i(t)\log\Big(1+\frac{h_{if}(t,l)p_{if}}{N_0}\Big) - \mu p_{if} \tag{35}$$

$$i_f(t) = \frac{1}{L}\sum_{l=1}^{L} \operatorname*{argmax}_{i} \quad \lambda_i(t)\log\Big(1+\frac{h_{if}(t,l)p_{if}(t)}{N_0}\Big) - \mu p_{if}(t) \tag{36}$$

and set $a_{i_f(t)}(t) = 1$ and $a_{if}(t) = 0$ for all other $i \neq i_f(t)$. The Stochastic DFP algorithm for optimal resource allocation in an FDMA broadcast channel is simulated for a system with $J = 10$ nodes using 2 frequency tones for communication. We consider two different cases. In the first scenario the Fading channels are i.i.d. Rayleigh with average powers 1 for all nodes. In the second case the fading channels are Rayleigh with the average 1 for the first five nodes and 10^3 for the second five nodes which corresponds to nodes that are about 10 times farther away from the access point. We call the first case well conditioned optimization problem and the second one is an ill conditioned optimization problem. Noise power is 1 and average power budget is $P_0 = 1$.

As you can see in the Figure 1, the speed of convergence for stochastic gradient descent and stochastic DFP are very similar when the condition number is small. The reason is for both algorithms the largest stepsize that we can use for small condition number is $\epsilon = 0.1$. However, it is obvious in Figure 2 the stochastic DFP has a faster convergence in comparison with stochastic gradient descent. The reason is by changing the condition number the largest step size that we can use for the stochastic gradient descent changes and we should set $\epsilon = 0.01$ to make sure that the sequence converges. While for the Stochastic DFP the same stepsize $\epsilon = 0.1$ also works for the problem with large condition number.

V. Conclusions

Optimal resource allocation problems in wireless systems were considered. A stochastic version of the DFP algorithm was introduced to find optimal dual variables. The proposed method inherits manageable computational complexity from stochastic gradient descent and reasonable convergence speed from deterministic DFP. Future research directions include further characterization of convergence properties, more exhaustive numerical experiments, and stochastic generalizations of other quasi-Newton methods.

References

[1] A. Ribeiro and G. Giannakis, "Separation principles in wireless networking," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4488–4505, September 2010.

[2] A. Ribeiro, "Optimal resource allocation in wireless communication and networking," *EURASIP J. Wireless commun.*, vol. 2012, no. 272, pp. 3727–3741, August 2012.

[3] X. Wang and G. B. Giannakis, "Resource allocation for wireless multiuser ofdm networks," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4359 – 4372, July 2011.

[4] N. D. Sidiropoulos, T. N. Davidson, and Z. Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, June 2006.

[5] Z. Quan, S. Cui, and A. H. Sayed, "Optimal linear cooperation for spectrum sensing in cognitive radio networks," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 28–40, February 2008.

[6] Y. Hu and A. Ribeiro, "Optimal wireless networks based on local channel state information," *IEEE Trans. Signal Process.*, vol. (submitted), no. 5, May 2011, available at http://www.seas.upenn.edu/ aribeiro/wiki.

[7] ——, "Optimal wireless multiuser channels with imperfect channel state information," in *Proc. Int. Conf. Acoustics Speech Signal Process.*, vol. (to appear). Kyoto Japan, March 25-30 2012.

[8] Y. Yi and S. Shakkottai, "Hop-by-hop congestion control over a wireless multi-hop network," *IEEE/ACM Trans. Netw.*, vol. 15, no. 133-144, pp. 1548–1559, February 2007.

[9] M. Chiang, S. H. Low, R. A. Calderbank, and J. C. Doyle, "Layering as optimization decomposition," *Proc IEEE*, vol. 95, no. 1, pp. 255–312, January 2007.

[10] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1–144, 2006.

[11] A. Ribeiro, "Ergodic stochastic optimization algorithms for wireless communication and networking," *IEEE Trans. Signal Process..*, vol. 58, no. 12, pp. 6369–6386, December 2010.

[12] C. G. Broyden, J. E. D. Jr., Wang, and J. J. More, "On the local and superlinear convergence of quasi-newton methods," *IMA J. Appl. Math*, vol. 12, no. 3, pp. 223–245, June 1973.

[13] R. H. Byrd, J. Nocedal, and Y. Yuan, "Global convergence of a class of quasi-newton methods on convex problems," *SIAM J. Numer. Anal.*, vol. 24, no. 5, pp. 1171–1190, October 1987.

[14] M. J. D. Powell, *Some properties of the variable metric method*, 2nd ed. London, UK: Academic Press, 1971.

[15] J. Nocedal and S. J. Wright, *Numerical optimization*, 2nd ed. New York, NY: Springer-Verlag, 1999.

[16] A. Mokhtari and A. Ribeiro, "A dual stochastic dfp algorithm for optimal resource allocation in wireless systems," *https://fling.seas.upenn.edu/ aryanm/wiki/index.php?n=Research.Publications.*