QoS and Scheduling in Wireless Networks

Vijay Subramanian

Network Business,

Motorola Inc.

Email: Vijay.Subramanian@motorola.com

Introduction

Three broad class of problems:

- Broadcast channel - DL of most systems, typically "centralized", MAIN FOCUS.
- Multiple-access channel - UL of most systems, usually distributed but sometimes "centralized" (eg. HSUPA).
- General wireless network - mesh, adhoc or sensor network, typically distributed algorithms used, interference management a key component of scheduling.

Setting

- Scheduling done at discrete times - some framing structure.
- Centralized scheduler - knowledge of queue-lengths, QoS and channel conditions.
- Generous resource allocation mechanisms - scheduler has access to a generous bag of modulation and coding schemes.
- Power is allocated to each user:
  - From a common pool on the DL.
  - Limiting the net received power based upon per user limits on the UL.
- Bandwidth-time slices allocated to each user
  - CDMA - spreading codes across time.
  - TDMA/FDMA - time and/or frequency.
  - OFDM - time-carrier slices.

Simple Case

Consider best-effort/rate-adaptive traffic - TCP flows

Two key notions emerge:

- Fairness versus efficiency trade-off
- Opportunistic scheduling/Multi-user diversity

Fairness versus efficiency

Assume a static channel and a TDMA-case with a total of I users in the system.

- User i gets rate R_i when he is allowed to transmit.
- User i is allowed to transmit ρ_i fraction of the time.
- Thus, throughput for user i is $T_i = \rho_i R_i$.
- Let utility obtained for user i be based only on throughput - $U_i(\rho_i R_i)$, concave, increasing function.

Let us choose ρ_i to solve the following problem:

$$\begin{aligned} \max \quad & \textstyle\sum_{i \in I} U_i(\rho_i R_i) \\ \text{subject to} \quad & \textstyle\sum_{i \in I} \rho_i \leq 1 \qquad \rho_i \geq 0, \ \forall i \in I \end{aligned}$$

Choose the following family of utility functions - $U_i(x) \in U^\alpha(x)$ for $\alpha \leq 1$:

$$U^\alpha(x) = \begin{cases} \frac{\mathrm{sgn}(\alpha) x^\alpha}{|\alpha|}, & \alpha \leq 1,\ \alpha \neq 0 \\ \log(x), & \alpha = 0 \end{cases}$$

Solution $\rho_i = \frac{R_i^{\frac{\alpha_i}{1-\alpha_i}}}{\sum_{j \in I} R_j^{\frac{\alpha_j}{1-\alpha_j}}}$. Couple of interesting cases:

1. Max Rate scheduler ($\alpha_i = 1\ \forall i$) - $i^* = \arg\max_{i \in I} R_i$, $\rho_{i^*} = 1$, $\rho_i = 0,\ \forall i \neq i^*$, User throughput - $T_{i^*} = R_{i^*}$, $T_i = 0,\ \forall i \neq i^*$, Sum throughput - $\max_i R_i$ (maximum).
2. Proportionally Fair scheduler ($\alpha_i = 0\ \forall i$) - $\rho_i = \frac{1}{|I|}$, User throughput - $T_i = \frac{R_i}{|I|}$, Sum throughput - $\frac{\sum_{i \in I} R_i}{|I|}$ (arithmetic mean).
3. Equal throughput scheduler ($\alpha_i = -\infty\ \forall i$) - $\rho_i = \frac{R_i^{-1}}{\sum_{j \in I} R_j^{-1}}$, User throughput - $T_i = \frac{1}{\sum_{j \in I} R_j^{-1}}$, Sum throughput - $\frac{|I|}{\sum_{j \in I} R_j^{-1}}$ (harmonic mean).

Multi-user diversity

Consider I users with i.i.d. channels under the following two cases, both with $\rho_i = \frac{1}{|I|}$:

1. Round-robin scheduling - User throughput $\frac{E[R]}{|I|}$, Sum throughput $E[R]$.
2. At time k schedule user $i^* = \arg\max_{i \in I} R_i(k)$ - User throughput $\frac{E[\max_{i \in I} R_i]}{|I|}$, Sum throughput $E[\max_{i \in I} R_i]$.

This gives substantial improvement in throughput - asymptotically $\log(|I|)$ under some conditions on the distribution.

Thus, by choosing good times to schedule, every user's throughput is improved - for time-varying channels. In fact, as the number of users increases, the gain increases (slowly) to infinity.

Abstract Model of a Wireless Network

Setting:

Wireless communication system with d users.

Time-varying channel conditions captured by stochastic state $\mathbf{e}_k \in S$ at time k, stationary and ergodic with stationary distribution γ.

$\forall \mathbf{e} \in S$ we have a rate-region $\mathcal{R}(\mathbf{e}) \subseteq K \subset \Re_+^d$ where K is compact with $\mathcal{R}(\mathbf{e})$ convex, coordinate-convex and closed for all $\mathbf{e}$.

Steady-state capacity region

$$\bar{\mathcal{R}} := \left\{ w \in \Re_+^k \ : \exists v(\mathbf{e}) \in \mathcal{R}(\mathbf{e}) \ \forall \ \mathbf{e} \in \mathcal{S} \text{ s. t. } w = \textstyle\int_{\mathcal{S}} v(\mathbf{e}) \gamma(d\mathbf{e}) \right\}. \qquad (1)$$

$\bar{\mathcal{R}}$ is convex, coordinate convex and compact.

Note: Technology details (including feedback capabilities) are abstracted into the rate-regions.

General Analysis for Best effort traffic

Problem Statement

A Gradient-based Algorithm

Convergence to ODE

Some Examples

Analysis of ODE

Simple Numerical Results

Problem Statement

Assume that the d users (in the model) have rate-adaptive streams that want to share the above channel fairly and efficiently.

Summarize this as follows

$$\sup_{w \in \bar{\mathcal{R}}} U(w) \stackrel{\triangle}{=} \sum_{i=1}^{d} U_i(w_i). \tag{2}$$

Assume $U_i(\cdot)$ increasing, strictly concave and continuously differentiable utility function on $\Re_+$.

Hence, there exists unique w^* maximiser to (2) characterized by

$$\nabla U(w^*)^T (w - w^*) \leq 0 \quad \forall w \in \bar{\mathcal{R}}. \tag{3}$$

Question: Can we achieve w^* using only online policies?

A Gradient-based Algorithm

Let $V_k \in \mathcal{R}(\mathbf{e}_k)$ be the rate vector selected at time k.
Consider the IIR filtered average throughput

$$\begin{aligned} W_{k+1} &= W_k + \mu(V_k - W_k) \\ W_{k+1} &= \mu \sum_{l=0}^{k} (1-\mu)^l V_{k-l} \end{aligned}$$

where $\mu \in (0,1)$ controls the time constant in the averaging.

$$U(W_{k+1}) \approx U(W_k) + \mu \nabla U(W_k)^T (V_k - W_k) \quad \text{when } \mu \ll 1.$$

Would like to optimise $U(W_{k+1})$ given $U(W_k)$. Best choice given past decisions (*myopic and greedy view*) choose

$$V_k = \arg \max_{w \in \mathcal{R}(\mathbf{e}_k)} \nabla U(w_k)^T w. \tag{4}$$

More generally we will consider choosing $V_k = F(W_k, \mathbf{e}_k)$.

Convergence to ODE

We are interested in the case where $\mu \ll 1$ and will be looking at asymptotics as $\mu \to 0$.

Define the continuous time process

$$W_\mu(t) := W_{[t/\mu]}, \quad t \geq 0, \quad \text{where } [x] := \sup\{i \in Z : i \leq x\}.$$

Let

$$\bar{F}(w) := \int_S F(w, \mathbf{e})\gamma(d\mathbf{e}).$$

Theorem 1 *Under the assumption that $\bar{F}$ is continuous and that $W_\mu(0) \to w_0$ in probability, it follows that any limit point W of $\{W_\mu\}$ satisfies the ODE*

$$\dot{W} = \bar{F}(W) - W. \tag{5}$$

Define

$$\begin{aligned} \Xi_\mu(t) &= \frac{1}{\sqrt{\mu}}(W_\mu(t) - W(t)) \text{ and} \\ L_\mu(t) &= \sqrt{\mu}\sum_{k=1}^{[t/\mu]} \left(F(W(k\mu), \mathbf{e}_k) - \bar{F}(W(k\mu))\right). \end{aligned}$$

Assume (**C1**) that $L_\mu \Longrightarrow L$, where L is a zero-mean Brownian motion. Mixing conditions on $\mathbf{e}_k$ will imply this.

Additionally assuming (**C2**) that $F(w, \mathbf{e})$ is continuously differentiable in w with bounded derivative $\partial_w F(w, \mathbf{e})$.

We then have

Theorem 2 *Assume* **C1-C2** *and that the solution to (5) exists for all* $t \geq 0$*, and that* $\Xi_\mu(0) \to \xi_0$ *in probability. Then* $\Xi_\mu \Longrightarrow \Xi$ *satisfying*

$$\Xi(t) = \xi_0 + L(t) + \int_0^t \partial\bar{F}(W(s))\Xi(s)ds. \qquad (6)$$

Some Examples

Complete knowledge of current channel state:

This is the original algorithm that we designed in (4) and

$$F(w, \mathbf{e}) = \arg \max_{u \in \mathcal{R}(\mathbf{e})} \nabla U(w)^T u.$$

Define a compact and convex set $\mathcal{Q} \subset \Re_+^d$ to be strictly-convex if for all $a \geq 0,\ \sum_{i=1}^d a_i = 1$, there is a unique maximiser of $a^T u$ in $\mathcal{Q}$.

Under strict convexity of $\bar{\mathcal{R}}$

$$\bar{F}(w) = \arg \max_{u \in \bar{\mathcal{R}}} \nabla U(w)^T u.$$

Note:

1. Gradient-based scheduling algorithm - related to the conditional gradient/Frank-Wolfe algorithm.
2. With convex rate regions assumption (4) is an “easy” problem to solve.
3. If every rate-region is a simplex, then we can restrict attention to a TDM-type algorithm where only one user is chosen at any given time.

Analysis of ODE

We consider the ODE

$$\dot{W} = \bar{F}(W) - W, \tag{7}$$

where based upon the various examples given before $\bar{F}(W)$ takes the form

$$\bar{F}(w) = \arg\max_{u \in \mathcal{Q}} \nabla U(w)^T u, \tag{8}$$

for some convex, coordinate convex, and compact subset $\mathcal{Q} \subseteq \bar{\mathcal{R}}$.

Denote by $w^*(\mathcal{Q})$ the (unique) maximiser of the following problem

$$\sup_{w \in \mathcal{Q}} U(w).$$

Under these conditions we have the following result

Proposition 1 *Under strict-convexity of $\mathcal{Q}$, it follows that $w^*(\mathcal{Q})$ is the unique equilibrium point of the differential equation (8) and $W(t) \to w^*(\mathcal{Q})$ as $t \to +\infty$ starting with any state $W(0) = w_0 \in \mathcal{Q}$.*

Numerical Results

Given state $\mathbf{e} \in [0, 1]$, $(R_1(\mathbf{e}), R_2(\mathbf{e})) = (1 - \mathbf{e}, 0.5\mathbf{e})$. The continuity assumption is satisfied for the TDM-type scheduling algorithm with current rates.

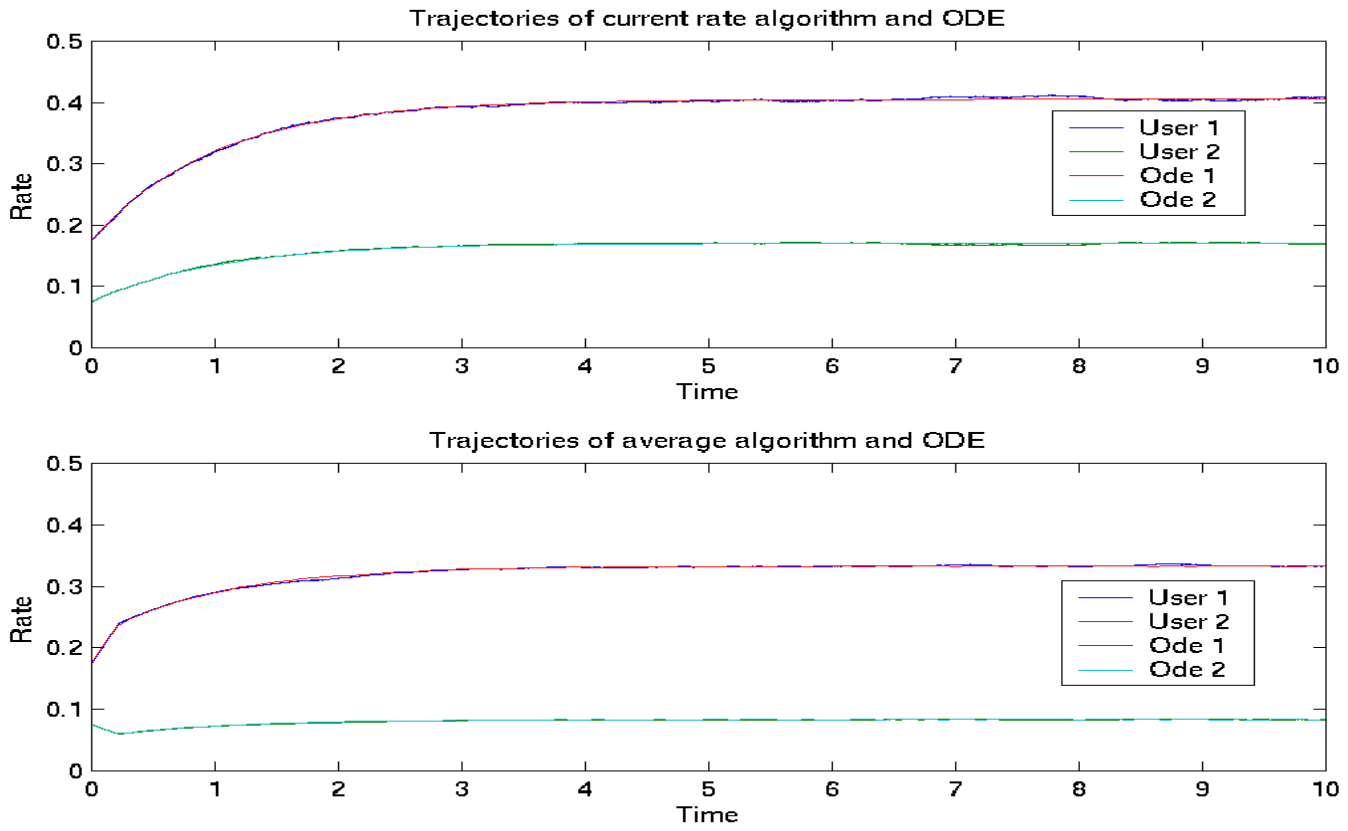

Figure 1: Trajectories of the different algorithms and comparison with the respective ODEs for Case 2.

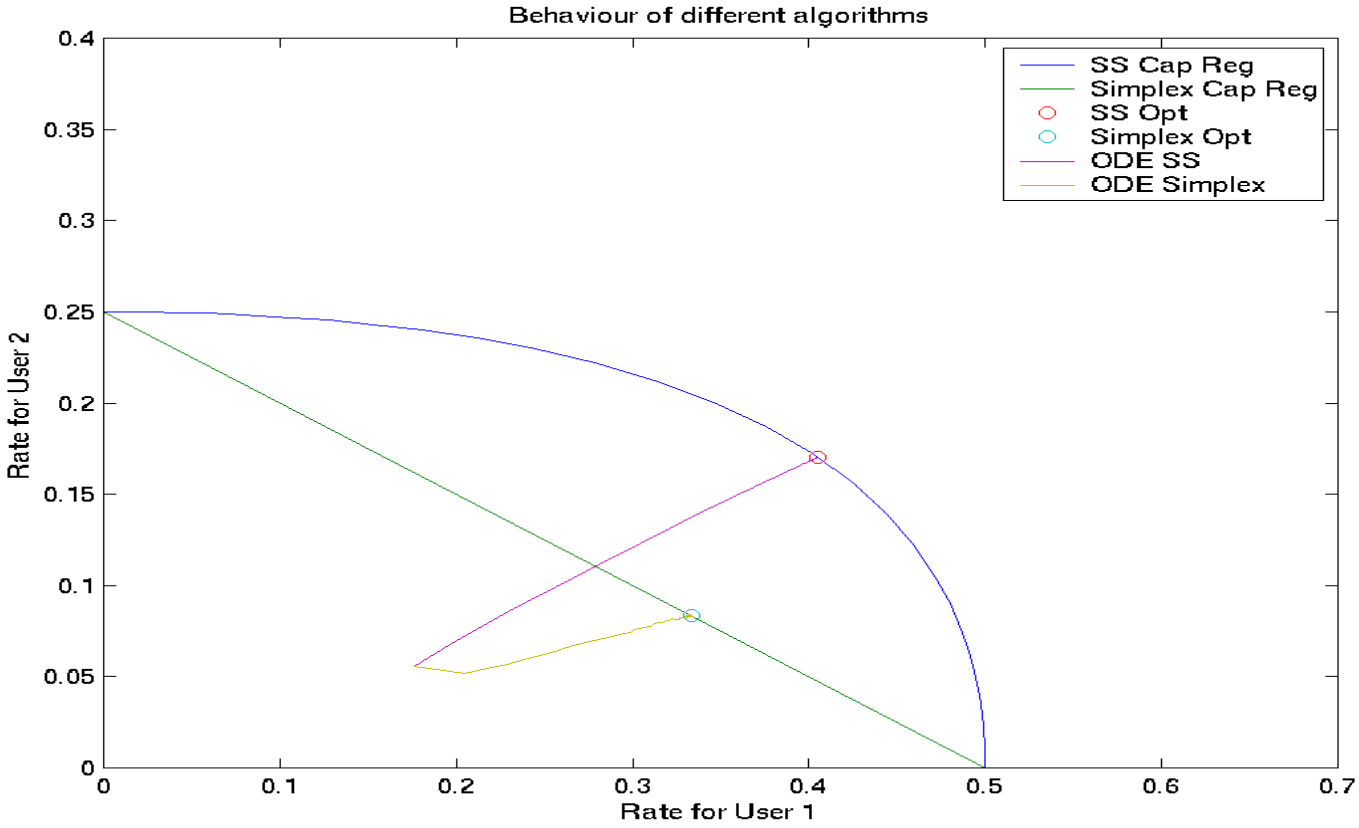

Figure 2: Trajectories of ODEs for Case 2.

General Analysis for real-time traffic

Review of stability

General framework for stabilizing policies

Largest weighted delay policy

Minimum draining time policy

Exponential Rule

Review of Stability

Rate Region: Assuming the earlier model this is the set of rate vectors that can be served by the channel/system.

Stabilizing policies: Set of policies that will keep queues stable for all rate vectors within the rate region. Classic Example of Rybko Stolyar:

- Buffers b_2 and b_4 given higher priority at stations.
- System unstable for $\lambda = 1$ if $\rho_1 = \frac{\lambda}{\mu_1} > 0.5$ and $\rho_2 = \frac{\lambda}{\mu_2} > 0$.
- However, stabilizing policies exist for $\rho_1 + \rho_2 < 1$.

Stability Definitions

Many related definitions exist:

1. Queueing system modelled as a Markov Chain - then positive recurrence of Markov Chain or Harris recurrence.
2. Policy π stable if there exists $B^\pi < +\infty$ such that $\sup_{t \geq 0} \sup_i E[W_i^\pi(t)] < B^\pi$, i.e., expected workload is bounded.
3. More general definitions are to show that time to empty (hitting time to 0) is finite, or has finite mean.

Proof methodologies

1. Using Lyapunov functions: Generalized Foster-Lyapunov Criterion
   There exists $V(X) \geq 0$, norm-like with

$$\begin{aligned} E[V(X(t+1))|X(t)] &< +\infty \\ E[V(X(t+1)) - V(X(t))|X(t)] &< -c \text{ for } \|X(t)\| \geq b \end{aligned}$$

2. Using Fluid Limits: Let $S^{(n)}$ denote process S with initial condition such that $\|S^{(n)}(0)\| = n$.
   Suppose there exists $\epsilon > 0$ and an integer $T > 0$ such that for any sequence of processes $\{S^{(n)}, n = 1, 2, \cdots\}$, we have

$$\limsup_{n \to +\infty} E[\frac{1}{n}\|S^{(n)}(nT)\|] \leq 1 - \epsilon.$$

   Then S is stable.
   Fluid limit $s(t)$ is obtained as limit of sequence of these scaled processes. The key idea is to show that $s(t)$ starting from $\|s(t)\| = 1$ hits 0 in finite time T and stays there.

Largest Weight Delay First

First, we concentrate on TDM channels (Simplex rate regions) to illustrate good policies.

Tassiulas and Ephremides 1993 - Dynamic Server allocation to Parallel Queues with randomly varying Connectivity.

System - N users each with mean number of packets served m_n, i.i.d. service rates and i.i.d. arrivals, i.i.d. connectivity of queue to server.

Policy - serve the longest connected queue first (LCQ). Proof of stability using Lyapunov function $V(x) = \sum_{n=1}^{N} \frac{x_n^2}{m_n}$.

With statistically identical arrival, service and connectivity processes and a binary arrival process, LCQ minimizes delays - stochastically smallest. Uses a sample-path coupling argument.

Modified Largest Weighted Delay/Workload First

M. Andrews, *et al.*: At time t serve user $i^* = \arg\max_i \gamma_i V_i^{\beta}(t) R_i(t)$ where $\beta > 0$, $R_i(t)$ is the rate for user i at time t and $V_i(t) = Q_i(t) + W_i(t)$ with $Q_i(t)$ the queue-length of user i and $W_i(t)$ the delay of the head-of-the-line packet.

Proof uses fluid limit coupled with Lyapunov function $L(y) = \frac{1}{1+\beta} \sum_i \gamma_i y_i^{1+\beta}$. Additionally, show that in the fluid limit, after a finite time a Little's law relationship holds, namely, $\lambda_i w_i = q_i$ to prove stability for the delay-based policy using the queue-length based policy.

(Myopic and greedy) Policy tries to minimize drift of Lyapunov function at every instant.

Properties:

1. Under heavy traffic (arrival rates close to boundary) with Resource Pooling M-LWDF minimizes $L(V)$ over all policies - nice property for online policy that does not know arrival rates.
2. Using Large Deviations one can show an optimality property for such a policy too.

Minimum Draining Time Policy

Agrawal and Leelahakriengkrai: Arrival process assumed to be such that

$$\forall T \ (A_i(t+T) - A_i(t) | A(s),\ s \le t) \quad \le Z_i \ \forall t, \quad E[e^{\theta_i Z_i}] < +\infty$$

$$E[(\frac{A_i(t+T) - A_i(t)}{T} - \lambda_i)^+ | A(t),\ s \le t] \qquad < \qquad \epsilon \ \forall t$$

Let $\mathcal{L}$ be a control set - N-dimensional, bounded, closed and convex, not necessarily $\bar{R}$. Then

$$\mathcal{L} \quad = \quad \cap_{k \in K} S^k$$

$$S^k \quad = \quad \{x \in R_+^N : \sum_{n=1}^{N} \frac{x_n}{r_n^k} \le 1\} \ - \ \text{a simplex}$$

K can be an infinite set.

Given $w \in R_+^N$ there exists $k \in K$ - corresponding to simplex S^k such that

$$< \frac{1}{r^k}, w >= \max_{l \in K} < \frac{1}{r^l}, w >= \sum_{n=1}^{N} \frac{w_n}{r_n^l}.$$

Minimum Draining Time Policy

At time $t \geq 0$, channel $H(t) = h$ and workload $W(t) = w$, policy π selects bit rate $r^{\pi}(h, \frac{w}{\|w\|})$ from $R_c(h)$ such that

1. $\frac{w}{\|w\|} \in R^k$ implies

$$< \frac{1}{r^k}, r^{\pi} > \quad = \quad \max_{r \in R_c(h)} < \frac{1}{r^k}, r >$$

$$R^k \quad = \quad \{w \in R_+^N :< \frac{1}{r_k}, w >= \max_{k \in K} < \frac{1}{r_k}, w >\}$$

2. $r_i^{\pi} = 0$ if $w_i = 0$.

3. "Continuity" of policy in $\frac{w}{\|w\|}$.

If $\mathcal{L} = \bar{R}$, then minimum Draining Time (MDT) Policy.

Minimum Draining Time Policy

Stability by using Lyapunov function - $L(w) = \max_{l \in K} < \frac{1}{r_l}, w >$.

$$\begin{aligned} E[\frac{L(W(t+T)) - L(W(t))}{T} | W(s),\ s \leq t] &\leq \epsilon_0 \text{ if } L(W(t)) > b\ \forall t \geq 0 \\ (L(W(t+T)) - L(W(t)) | W(s),\ s \leq t) &\leq Z\ \forall t \leq 0 \end{aligned}$$

M-LWDF special case of this:

$$\begin{aligned} \mathcal{L} &= \{x : \sum_{i=1}^{N} \frac{x_i^2}{\bar{R}_i} \leq 1\} \\ r^{\pi} &= \arg \max_{r \in R_c(h)} \sum_{i=1}^{N} \frac{w_i r_i}{\bar{R}_i} \end{aligned}$$

M-LWDF for more general rate regions than simplexes. By good choice of $\mathcal{L}$ we can tune the delays properly.

Exponential Rule

Choose

$$\begin{aligned} r &= \arg \max_{r \in R_c(h)} < w, r > \\ w_i &= \gamma_i e^{\frac{a_i W_i(t)}{\beta + (\bar{W}(t))^\eta}}, \ \eta \in (0,1) \\ \bar{W}(t) &= \frac{1}{N} \sum_{n=1}^{N} a_i w_i(t) \end{aligned}$$

Proof: Uses local fluid limit - separation of time-scales.

Allows for good tuning of delays - the $\bar{W}(t)$ term allows for equalizing the weighted delays.

Detailed Example - HSDPA Scheduling

Motivation for physical setting - Optimization Set

Additional Motivation for cost function

Restatement with additional constraints

Optimal Power Allocation

Dual Formulation

Optimal Algorithm

Key Steps

Suboptimal Algorithms

Results

Problem Statement

Given $\mathbf{e}$ - channel gain by interference plus noise

$$\max_{\mathbf{r}\in\mathcal{R}} < \mathbf{w}, \mathbf{r} > \tag{9}$$

where $\mathbf{w} \geq \mathbf{0}$ and

$$\mathcal{R}(\mathbf{e}) = \left\{ \mathbf{r} \geq 0 : r_i = n_i B \log\left(1 + \frac{p_i e_i}{n_i}\right), \right. \\ \left. n_i \leq N_i \ \forall i, \sum_i n_i \leq N, \sum_i p_i \leq P \right\}, \tag{10}$$

where p_i is the amount of power assigned to user i and n_i the number of codes assigned to user i.

Gaussian broadcast channel - CDMA with orthogonal codes over flat-faded channels.

Li and Goldsmith[*IT, March'01*] - characterization of rate region.

Motivation - Physical Setting

WCDMA evolution - High-Speed Downlink Packet Access (HSDPA), Release 5.

1. Co-exists with UMTS - partitioned resources.
2. Fine grain scheduling (2msec) at the base-sites.
3. Fixed spreading factor (16) Walsh codes. Upto 15 codes assigned. Can code-division multiplex (CDM) multiple users in one time-frame.
4. Dedicated reverse channels give (regular) channel quality feedback (quantised SINR of pilot is typically used) and transmission feedback.
5. Transmitter has big bag of modulation and coding schemes to choose from - **Use turbo codes and under Gaussianity assumption (knowing channel) come close enough to achieving Shannon capacity.**
6. Shared control channels to indicate users scheduled to and modulation format: Modulation schemes - QPSK, 8-PSK, 16-QAM, MCS signalled using shared control channel.
7. User equipment has different capability based upon class - number of codes that they can simultaneously decode (at least 5!).

Additional Motivation - Cost Function

Let $W_{i,t}$ be measure of throughput acheived by user i up to time t.

Let $Q_{i,t}$ be queue-length of user i at time t.

1. Since rate region is convex different choices of $\mathbf{w} \geq 0$ help trace out capacity region.
2. Rate Adaptive Sources[AgrawalSubramanian02, Stolyar03, WhitingKushner02] - Choosing

$$\mathbf{r}_t^* = \arg \max_{\mathbf{r}_t \in \mathcal{R}(\mathbf{e}_t)} \sum_i c_i (W_{i,t})^{\alpha - 1} r_{i,t} \qquad \alpha \leq 1$$

   yields asymptotic convergence of $\mathbf{W}_t$ to $\mathbf{W}^* = \arg\max_{\mathbf{r} \in \bar{\mathcal{R}}} U(\mathbf{r})$.

3. Stabilizing Policies[Tasioulas *et al.*, Agrawal *et al.*, Stolyar *et al.*, Shakottai *et al.*, YehCohen, ...]: Choosing

$$\mathbf{r}_t^* = \arg \max_{\mathbf{r}_t \in \mathcal{R}(\mathbf{e}_t)} \sum_i d_i (Q_{i,t})^{p-1} r_{i,t} \qquad p > 1$$

   yields a stabilizing policy - Lyapunov function $\sum_i d_i (Q_{i,t})^p$.

Problem restatement - Additional constraints

Solve for $(\mathbf{n}^*, \mathbf{p}^*)$ that yield

$$V^* = \max_{(\mathbf{n},\mathbf{p})\in\mathcal{X}} V(\mathbf{n},\mathbf{p})$$

subject to

$$\sum_i n_i \quad \leq \quad N \qquad (N \leq 15) \tag{11}$$

$$\sum_i p_i \quad \leq \quad P \tag{12}$$

where

$$V(\mathbf{n},\mathbf{p}) \quad = \quad \sum_i w_i n_i \log(1 + \frac{p_i e_i}{n_i}) \tag{13}$$

$$\mathcal{X} \quad = \quad \{(\mathbf{n},\mathbf{p}) \geq (\mathbf{0},\mathbf{0}) : n_i \leq N_i \ \forall i\} \quad (N_i \geq 5) \tag{14}$$

Relaxing integral constraints on $\mathbf{n}$.

Additional constraints:

1. Max SINR (per code)

$$\frac{p_i e_i}{n_i} \leq s_i \Leftrightarrow p_i \leq \frac{n_i s_i}{e_i}$$

2. Max rate per code constraints

$$\frac{r_i}{n_i} = \log\left(1 + \frac{p_i e_i}{n_i}\right) \leq (R/N)_i \Leftrightarrow p_i \leq \frac{n_i}{e_i}\left(e^{(R/N)_i} - 1\right)$$

Equivalent and they arise from a spectral efficiency / modulation order constraint.

Can be written as $p_i \leq s_i(n_i)\frac{n_i}{e_i}$. Thus, we have

$$\mathcal{X} = \left\{(\mathbf{n}, \mathbf{p}) \geq (\mathbf{0}, \mathbf{0}) : n_i \leq N_i,\ p_i \leq s_i(n_i)\frac{n_i}{e_i}\ \forall i\right\}$$

For convexity of $\mathcal{X}$ need $s_i(n_i)n_i$ to be concave in n_i.

Optimal Power Allocation

Assume $\sum_i N_i > N$, else set codes N_i, optimize over power.

Power optimization for given code allocation

Need to find $\mathbf{p}^*$ that solves

$$V^*(n) = \max_{\mathbf{p} \geq 0 : p_i \leq s_i(n_i) \frac{n_i}{e_i} \ \forall i} V(\mathbf{n}, \mathbf{p})$$

1. Check if problem is non-trivial, i.e., $\sum_i \frac{s_i(n_i) n_i}{e_i} > P$, else solution is obvious. Note that we may not fill the power budget in this case.
2. Using a dual formulation get a water-filling solution.

$$\exists \tilde{\lambda} \geq 0 \text{ s.t. } p_i^* = \frac{n_i}{e_i} \left[\left(\frac{w_i e_i}{\tilde{\lambda}} - 1 \right) \wedge s_i(n_i) \right]^+ \ \forall i, \ \sum_i p_i^* = P.$$

3. In conjunction with value of $\mathbf{p}^*$. leads to a simple finite-time algorithm (complexity $O(M \log M)$) to determine $\tilde{\lambda}$.

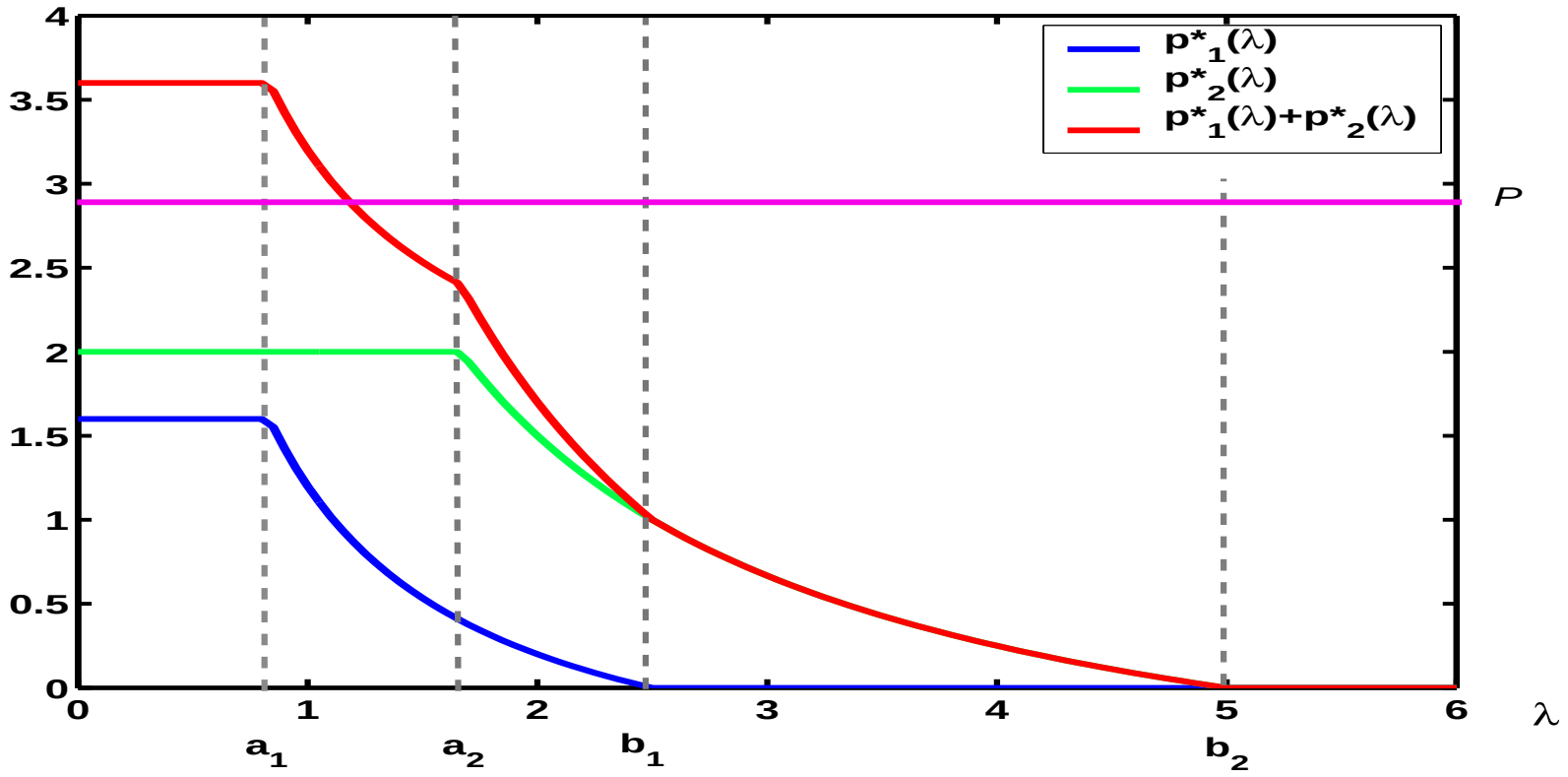

Figure 3: Algorithm Illustration.

Dual Formulation

Define

$$L(\mathbf{p}, \mathbf{n}, \lambda, \mu) = \sum w_i n_i \log\left(1 + \frac{p_i e_i}{n_i}\right) + \lambda(P - \sum p_i) + \mu(N - \sum n_i)$$

Dual function

$$L(\lambda, \mu) = \max_{(\mathbf{n},\mathbf{p}) \in \mathcal{X}} L(\mathbf{p}, \mathbf{n}, \lambda, \mu)$$

Dual problem: Solve for

$$L^* = \min_{(\lambda,\mu) \geq 0} L(\lambda, \mu)$$

Define $L(\lambda) = \min_{\mu \geq 0} L(\lambda, \mu)$, this is convex.

(Using Slater's Condition)
No duality gap and existence of Lagrange multipliers.

Optimal Algorithm

The solution to the optimization problem takes the following steps:

1. Given λ and μ for every $\mathbf{n}$ construct $\mathbf{p}^*(\mathbf{n})$ with $(\mathbf{n}, \mathbf{p}^*(\mathbf{n})) \in \mathcal{X}$ with $\mathbf{p}^*(\mathbf{n}) = \arg\max_{\mathbf{p}:(\mathbf{n},\mathbf{p})\in\mathcal{X}} L(\mathbf{p}, \mathbf{n}, \lambda, \mu)$. Note that this is not the same as the optimizing power vector from earlier. (Trivial)
2. Find $L(\lambda, \mu) = \max_{\mathbf{n}\geq\mathbf{0}:n_i\leq N_i} L(\mathbf{n}, \mathbf{p}^*(\mathbf{n}), \lambda, \mu)$. (Easy but important)
3. Find $L(\lambda)$ and $\mu^*(\lambda) = \arg\min_{\mu\geq 0} L(\lambda, \mu)$. (Simple but reveals key structure)
4. Find $\lambda^* = \arg\min_{\lambda\geq 0} L(\lambda)$ - Numerical search, unimodal function (convex) with $0 \leq \lambda \leq \max w_i e_i$, thus, Golden section method works well, there are means to speed up search.
5. Given λ^* find feasible $\mathbf{n}, \mathbf{p}$ to solve for primal. (Easy except for ...)

Key Steps

We have

$$L(\mathbf{n}, \mathbf{p}^*(\mathbf{n}), \lambda, \mu) = \sum n_i \left(w_i h(w_i e_i, s_i(n_i), \lambda) - \mu\right) + \mu N + \lambda P$$

When $s_i(n_i) \equiv s_i$ (constraints (1) and (2)), easy to optimize over $\mathbf{n}$ to yield

$$L(\lambda, \mu) = \sum \left[\mu_i(\lambda) - \mu\right]^+ N_i + \mu N + \lambda P$$

where $\mu_i(\lambda) = w_i h(w_i e_i, s_i(n_i), \lambda)$.

Optimizing over μ is now very easy!

1. Order users in decreasing order of $\mu_i(\lambda)$ - Π_λ ordering.
2. Pack code budget as per Π_λ order.

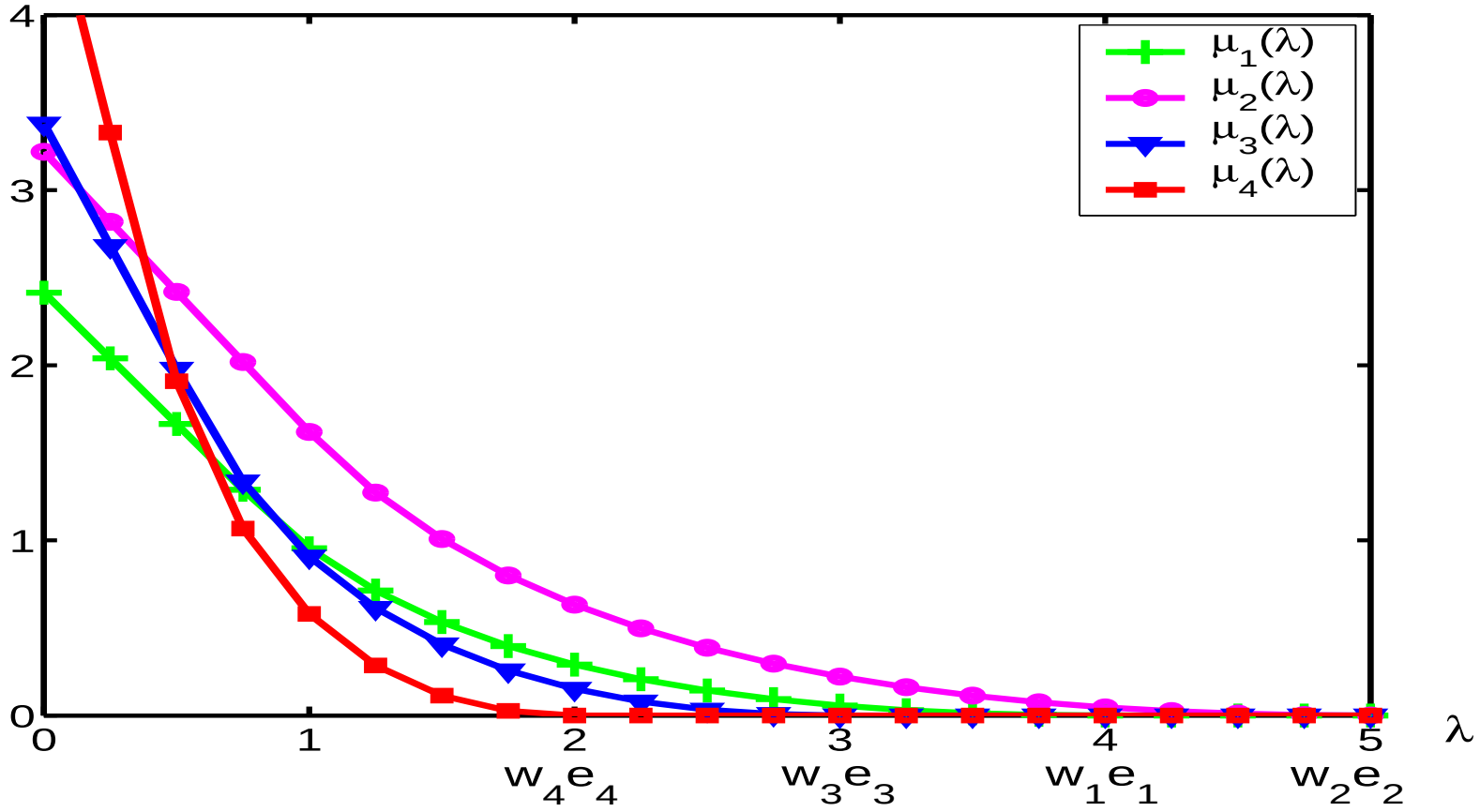

Figure 4: Example of $\mu^*(\lambda)$ computation.

We get the following structure for the optimal solution

Lemma 1 *For constraints of the type (1) and (2), an optimal code allocation can be found with the following properties:*

1. *For the case of $N_i = N$ at most two users will be scheduled.*
2. *If a certain condition holds or at most two users are involved in a tie, then at most $\lceil N/N_{\min} \rceil + 1$ users will be scheduled, where $N_{min} := \min_i N_i$. All but two users will have their full code allocation.*

Remark: For the HSDPA constraints this yields that a maximum of 4 users need to be scheduled - the simplest setting in the standards!!!

Suboptimal Algorithms

Owing to processing requirements - only 2 msec to schedule, code, etc. - it would be good to have lower complexity algorithms that perform well-enough.

1. **Truncated Optimal**: Guess at λ^*, compute code, get optimal power allocation.
2. **Greedy**: Extend single-user scheduling method (eg. HDR) by sequentially adding users.

Numerical Investigations

Simulation set-up:

40 users with average e chosen as per measure distribution with maximum code capability of 5.

Real e varied over time by i.i.d. processes with the Clarke spectrum with 10Hz Doppler.

Power budget 11.9W and code budget 15.

Considered utility maximization problem for different αs and infinite buffers.

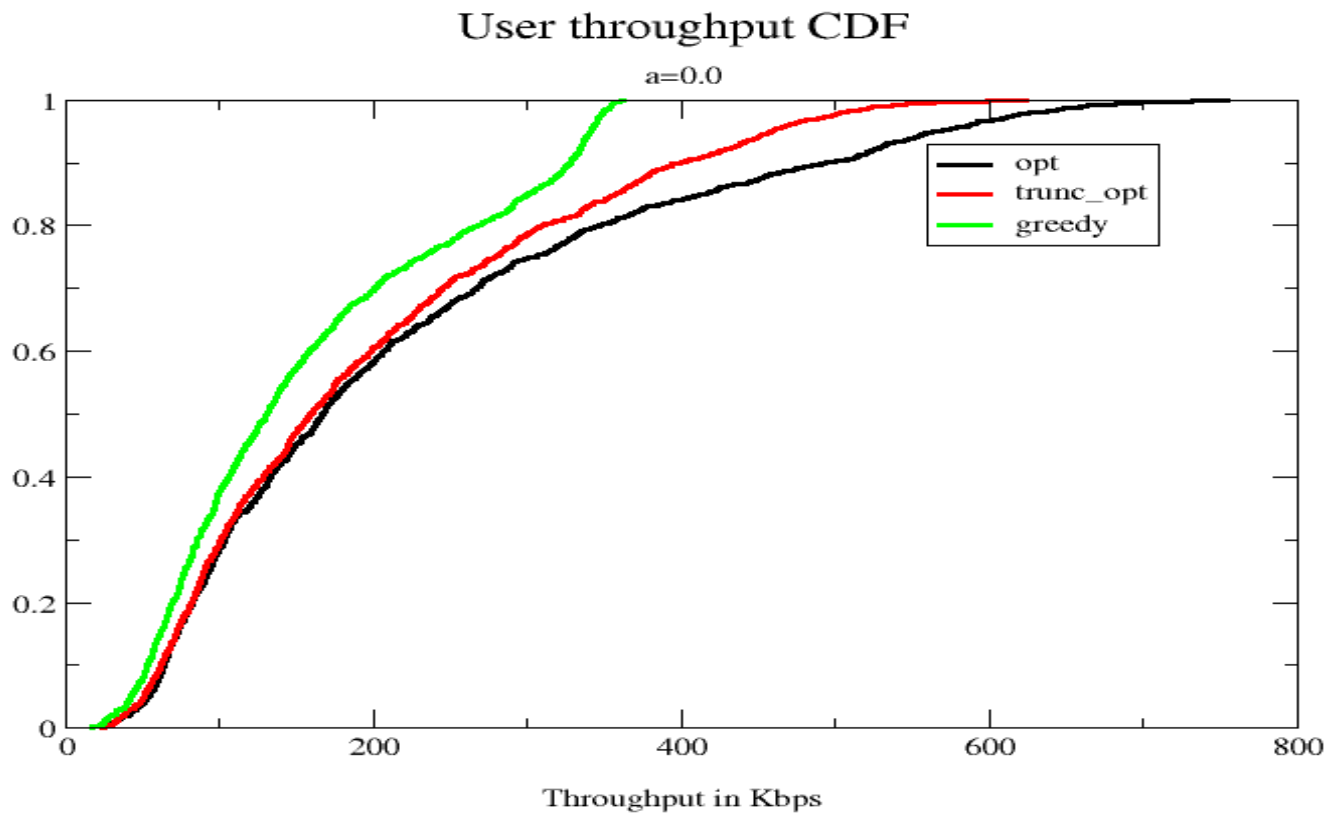

Figure 5: Empirical CDF of users throughputs for $\alpha = 0$.

Table 1: Simulation Results

α	Algorithm	Utility	Log Utility	M	N_s	P_s	Sector Throughput (Mbps)
0.0	Optimal	231.944	231.944	3.35461	15	11.8997	8.8145
0.0	Truncated optimal	229.282	229.282	3	15	11.2689	7.87875
0.0	Greedy baseline	222.222	222.222	3	15	10.9659	6.36075
0.25	Optimal	173.646	231.669	3.33331	15	11.8998	9.28545
0.25	Truncated optimal	170.275	228.886	3	15	10.7793	8.54505
0.25	Greedy baseline	163.798	222.663	3	15	10.6948	7.2903
0.5	Optimal	806.085	228.404	3.36408	15	11.899	11.1392
0.5	Truncated optimal	749.531	224.379	3	15	9.83421	9.127
0.5	Greedy baseline	725.4	220.801	3	15	9.72985	8.6008
0.75	Optimal	4129.16	213.411	3.36341	15	11.8903	12.6934
0.75	Truncated optimal	3579.71	207.866	3	15	7.82554	10.1799
0.75	Greedy baseline	3538.96	201.87	3	15	7.79743	10.2524

Figure 6: Optimal gives 38% improvement over greedy.

CDMA broadcast channel

1. Optimal code and power allocation via dual formulation
   - Power
   - Code
   - μ^*
   - λ^* one-dimensional numerical optimization.
2. Structural properties - need to consider at most $\lceil N/N_{\min} \rceil + 1$ users.
3. Optimal power allocation given a code allocation.
4. Lower complexity sub-optimal algorithms.
5. Naive extension of TDM policy significantly underperforms optimal CDM policy.

Other examples

- GPRS and EDGE: These are TDM systems: algorithms are a simple sorting type, followed by picking of the maximum.
- UMTS:

  Problem is similar to that of HSDPA but the functional relationship between power and code is different.

  Power budget is what is obtained after the common and dedicated channel powers have been removed.

  Code budget relationship is different since different spreading factor codes are used: need $\sum_i 2^{-SF_i} < 1 - \sum_{j \in \text{ resv}} 2^{-SF_j}$.

- HRPD/HRPD-A: With HRPD since it is TDM rate-regions, solution is easy, with HRPD-A can use the same formulation to decide the set of users to schedule.
- M. Andrews *et al.*: Possible to extend rate-adaptive streams policy to cases with minimum and maximum rate guarantees - use counters to detect overflow and underflow and use an exponential weight for this.

Conclusions

A broad class of scheduling policies have at the heart a gradient problem to be solved.

This problem is a weighted rate maximization over the current (possible) rate region.

In the case of rate-adaptive sources, these algorithms converges to utility maximizing solutions.

In the case of real-time sources, these algorithms are stabilizing in nature, and by tweaking the parameters appropriately, good delay performance can be obtained.

References

1. R. Agrawal and V. Subramanian, ”Optimality of Certain Channel Aware Scheduling Policies,” Proceedings Allerton Conference Oct 2002.
2. R. Agrawal, V. Subramanian and R. Berry, ”Joint Scheduling and Resource Allocation in CDMA Systems,” Proceedings WiOpt04, Cambridge, UK, March 2004.
3. S. Shakkottai and A. Stolyar, ”Scheduling for Multiple Flows sharing a Time-Varying Channel: The Exponential Rule,” Proceedings of the AMS, 2001.
4. M. Andrews *et al.*, ”Scheduling in a Queueing System with Asynchronously Varying Rates,” Bell Systems Technical Journal.
5. R. Agrawal and R. Leelahakriengkrai, ”Minimum Draining Time Policies,” ITC 2001.
6. M. Andrews *et al.*, ”Optimal Utility Based Multi-user Throughput Allocation Subject to Throughput Constraints,” IEEE Infocom2005.