Non-convex Optimization and Resource Allocation in Wireless Communication Networks

Ravi R. Mazumdar

School of Electrical and Computer Engineering
Purdue University
E-mail: mazum@ecn.purdue.edu

Joint Work with Prof. Ness B. Shroff and Jang-Won Lee
Outline

- Introduction and non-convexity
- Joint power and rate allocation for the downlink in (CDMA) wireless systems
- Opportunistic power scheduling for the downlink in multi-server wireless systems
- Conclusion and future work
Motivation

- Tremendous growth in the number of users in communication networks
- Increasing demand on various services that can provide QoS
- Scarce network resources

*Need to efficiently design and engineer resource allocation schemes for heterogeneous services*
Motivation

- Most services are *elastic*
  - can adjust the amount of resource consumption to some degree
- By appropriately exploiting the elasticity of services
  - can maintain high efficiency and fairness
  - can alleviate congestion within the network
- Need appropriate model for the elasticity
- **Utility**
  - degree of user’s (service’s) satisfaction or performance by acquiring a certain amount of resource
  - different elasticity with different utility functions
  - example: expected throughput as a function of power allocation in wireless system
Total system utility maximization

\[
\max \sum_{i=1}^{M} U_i(\bar{x}) \\
\text{s. t. } g_k(\bar{x}) \geq 0, \ k = 1, 2, \ldots, K \\
\bar{x} \in X
\]

- If all \( U_i \) and \( g_k \) are concave and \( X \) is a convex set,
  - convex optimization problem
  - can be solved by using standard techniques
- Otherwise,
  - non-convex optimization problem
  - difficult to solve requiring a complex algorithm
Non-convexity in resource allocation

In general, three types of utility functions:

- **concave**: traditional data services on the Internet
- **sigmoidal-like (“S”)**: many wireless services and real-time services on the Internet
- **convex**: some wireless services
Non-convexity (cont’d)
Non-convexity (cont’d)

- Increasing demand for wireless and real-time services
  - non-concave utility functions becoming important
  - non-convex optimization problem
    ⇒ complex algorithm for a global optimum

Can we develop a simple algorithm for the approximation to the global optimum?
Inefficiency of naive approach

- 11 users and 10 units of a resource
- Utility function for each user: $U(x)$
- Approximate $U(x)$ with concave function $V(x)$

With $V(x)$, for each user,

$$x^* = \frac{10}{11}$$

- however, $U(x^*) = 0$
- zero total system utility

By allocating one unit to 10 users and zero to one user:

- 10 units of total system utility

Need resource allocation algorithms taking into account the properties of non-concave functions
Dual approach

Primal problem

\[
\text{max} \quad \sum_{i=1}^{M} U_i(\bar{x}) \\
\text{s. t.} \quad g_k(\bar{x}) \geq 0, \quad k = 1, 2, \ldots, K \\
\bar{x} \in X
\]

Dual problem

\[
\text{min} \quad Q(\bar{\lambda}) \\
\text{s. t.} \quad \bar{\lambda} \geq \bar{0},
\]

\[
Q(\bar{\lambda}) = \max_{\bar{x} \in X} \left\{ \sum_{i=1}^{M} U_i(\bar{x}) + \sum_{k=1}^{K} \lambda_k g_k(\bar{x}) \right\}
\]

- convex optimization
- simpler constraints
- smaller dimension

In many cases, the dual is easier to solve than the primal
Dual approach (cont’d)

Primal problem

\[
\begin{align*}
\text{max} & \quad \sum_{i=1}^{M} U_i(\bar{x}) \\
\text{s. t.} & \quad g_k(\bar{x}) \geq 0, \\
& \quad k = 1, 2, \ldots, K \\
& \quad \bar{x} \in X
\end{align*}
\]

Dual problem

\[
\begin{align*}
\text{min} & \quad Q(\bar{\lambda}) \\
\text{s. t.} & \quad \bar{\lambda} \geq 0,
\end{align*}
\]

\[
Q(\bar{\lambda}) = \max_{\bar{x} \in X} \left\{ \sum_{i=1}^{M} U_i(\bar{x}) + \sum_{k=1}^{K} \lambda_k g_k(\bar{x}) \right\}
\]

- non-convex optimization
- convex optimization
- simpler constraints
- smaller dimension

May not guarantee the feasible and optimal primal solution
Part I

Joint power and rate allocation for the downlink in (CDMA) wireless systems
Why joint power and rate allocation?

- Power is fundamental radio resource
  - trade off between performance of each user

- Variable data rate
  - trade off between data rate and the probability of packet transmission success for a given power allocation

- By jointly optimizing power and data rate allocation, the system performance can be further improved
Related work

- Oh and Wasserman [MOBICOM99]: Uplink power and rate control for a single class system without constraint on the maximum data rate
  - if applied to downlink, single server transmission is optimal

- Bedekar et al. [GLOBECOM99] and Berggren et al. [JSAC01]: Downlink power and rate control without constraint on the maximum data rate
  - single server transmission is optimal
Our work

- CDMA system that supports variable data rate by variable spreading gain
- Downlink in a single cell
- Snapshot of a time-slot
  - Constant Path gain and interference level during the time-slot
- Base-station has the total transmission power limit $P_T$
- Each user $i$ has
  - $R_{i}^{max}$: maximum data rate
  - $f_i$: function for packet transmission success probability
Signal to Interference and Noise Ratio (SINR)

SINR for user $i$

\[
\gamma_i(R_i, \bar{P}) = \frac{W}{R_i} \frac{P_i}{\theta(\sum_{m=1}^{M} P_m - P_i) + A_i}
\]

- $M$: number of users in the cell
- $W$: chip rate
- $\theta$: orthogonality factor
- $P_i$: power allocation for user $i$
- $R_i$: data rate of user $i$
- $A_i = I_i/G_i$: transmission environment of user $i$
  - $I_i$: background noise and intercell interference at user $i$
  - $G_i$: path gain from the base-station to user $i$

SINR is a function of power and rate allocation
Packet transmission success probability: $f_i$

- $f_i$ is an increasing function of $\gamma_i$

- For a given $R_i$, if $\sum_{m=1}^{M} P_m = P_T$, $f_i$ is
  - concave function,
  - “S” function, or
  - convex function

of its own power allocation $P_i$
Problem formulation

\[(A) \quad \max_{P_i, R_i} \sum_{i=1}^{M} R_i f_i(\gamma_i(R_i, \bar{P})) \]

s. t. \[\sum_{i=1}^{M} P_i \leq P_T \]
\[0 \leq P_i \leq P_T, \quad \forall i \]
\[0 \leq R_i \leq R_{i_{max}}, \quad i \in V \]
\[R_i = R_i^*, \quad i \notin V \]

- \(V\): a subset of users that have variable data rate
- \(R_i f_i(\gamma_i(R_i, \bar{P}))\): expected throughput of user \(i\)

**Goal**: Obtaining power and rate allocation that maximizes the expected total system throughput with constraints on the total transmission power limit of the base-station and the maximum data rate of each user
Optimal rate allocation

- To maximize the expected total system throughput, the base-station must transmit at the maximum power limit.

- Redefine SINR for user $i$ as

$$\gamma_i(R_i, P_i) \triangleq \frac{WP_i}{R_i \left( \frac{P_i}{P_T - P_i + A_i} \right)} = \frac{W}{R_i} \frac{P_i}{\sum_{j=1}^{M} P_j - P_i + A_i} = \gamma_i(R_i, \bar{P})$$

- For a given power allocation $P_i$, the optimal rate of user $i$,

$$R_i^*(P_i) = \begin{cases} \frac{WP_i}{\gamma_i^*(P_T - P_i + A_i)}, & \text{if } i \in V, \quad P_i \leq \frac{R_i^{max} \gamma_i^*(P_T + A_i)}{W + R_i^{max} \gamma_i^*} \\ R_i^{max}, & \text{if } i \in V, \quad P_i > \frac{R_i^{max} \gamma_i^*(P_T + A_i)}{W + R_i^{max} \gamma_i^*} \\ R_i^*, & \text{if } i \not\in V, \end{cases}$$

where $\gamma_i^* = \arg \max_{\gamma \geq 1} \left\{ \frac{1}{\gamma} f_i(\gamma) \right\}$. 
Equivalent power allocation problem

(B) \[
\max \sum_{i=1}^{M} U_i(P_i)
\]

s.t. \( \sum_{i=1}^{M} P_i \leq P_T \)
\( 0 \leq P_i \leq P_T, \quad \forall i, \)

\[
U_i(P_i) = \begin{cases} 
\frac{W}{\gamma^*_i} \frac{P_i}{P_T - P_i + A_i} f_i(\gamma^*_i), & \text{if } i \in V, P_i \leq \frac{R^\max_i \gamma^*_i (P_T + A_i)}{W + R^\max_i \gamma^*_i} \\
R^\max_i f_i(\gamma_i(R^\max_i, P_i)), & \text{if } i \in V, P_i > \frac{R^\max_i \gamma^*_i (P_T + A_i)}{W + R^\max_i \gamma^*_i} \\
R^*_i f(\gamma_i(R^*_i, P_i)), & \text{if } i \notin V
\end{cases}
\]

\( U_i(P_i) \) is a convex, concave, or “S” function of \( P_i \).
Power allocation

- Amount of power maximizing net utility

\[ P_i(\lambda) = \arg\max_{0 \leq P_i \leq P_T} \{U_i(P_i) - \lambda P_i\} \]

- Maximum willingness to pay per unit power

\[ \lambda_i^{max} = \min\{\lambda \geq 0 \mid \max_{0 \leq P \leq P_T} \{U_i(P) - \lambda P\} = 0\}, \forall i \]

- unique for each user \( i \)
- if \( \lambda > \lambda_i^{max} \), then \( P_i(\lambda) = 0 \)
- if \( \lambda < \lambda_i^{max} \), then \( P_i(\lambda) > 0 \)
Power allocation (cont’d)

Assume that \( \lambda_{1}^{\text{max}} \geq \lambda_{2}^{\text{max}} \geq \cdots \geq \lambda_{M}^{\text{max}} \)

- **User selection**
  - Select users from 1 to \( K \) that satisfies

\[
K = \max_{1 \leq j \leq M} \left\{ \sum_{i=1}^{j} P_{i}(\lambda_{j}^{\text{max}}) \leq P_{T} \right\}
\]

- **Users are selected in a decreasing order of** \( \lambda_{i}^{\text{max}} \)

- **Power allocation**
  - Find \( \lambda^{*} \) such that \( \sum_{i=1}^{K} P_{i}(\lambda^{*}) = P_{T} \)
  - Allocate power to each selected user \( i \) as \( P_{i}(\lambda^{*}) \)
  - Optimal power allocation for the selected users
Optimality

- $\bar{P}^*$: our power allocation
- $\bar{P}^o$: optimal power allocation
- If $\sum_{i=1}^{M} U_i(\gamma_i(P^o_i)) \rightarrow \infty$ as $M \rightarrow \infty$,

$$\frac{\sum_{i=1}^{M} U_i(\gamma_i(P^*_i))}{\sum_{i=1}^{M} U_i(\gamma_i(P^o_i))} \rightarrow 1, \text{ as } M \rightarrow \infty$$

- Our power allocation is
  - asymptotically optimal
  - a good approximation of the optimal power allocation with a large number of small users
Multiple access strategy

If

\[ R_{i}^{\text{max}} A_i \geq \frac{P_T W}{\gamma_i^*}, \quad \forall i, \]

single server transmission is optimal

- when users have high maximum data rate or are experiencing poor transmission environment
- when there is no constraint on the maximum data rate

- \( W \): chip rate
- \( \gamma_i^* \): constant that depends on \( f_i \)
- \( A_i = I_i/G_i \): transmission environment of user \( i \)
- \( I_i \): intercell interference and background noise at user \( i \)
- \( G_i \): path gain from the base-station to user \( i \)
Multiple access strategy (cont’d)

- If $\sum_{i=1}^{M} P_i(\lambda_{M}^{max}) \leq P_T$, selecting all users is optimal
- If $P_1(\lambda_{2}^{max}) \geq P_T$, selecting only user 1 is optimal
- Otherwise, selecting a subset of users can be optimal
- Condition for optimal multiple access strategy depends on time-varying parameters such as
  - number of users
  - type of users (utility functions)
  - channel condition of users
- Static multiple access strategy could be inefficient

Need dynamic multiple access strategy (dynamic multi-server transmission)
User selection strategy

- If all users are *homogeneous*, selecting users according to *transmission environment* is optimal
  - higher priority to a user in a better transmission environment

- However, if users are *heterogeneous*, *no* simple optimal user selection strategy

- Our user selection strategy provides a simple and unified selection strategy for *heterogeneous* users
User selection strategy

- User $i$ is called *more efficient* than user $j$ if

\[ U_i(\gamma_i(P)) \geq U_j(\gamma_j(P)), \forall P \]

- More efficient user has a higher priority to be selected
- When other conditions are the same, user $i$ has a higher priority to be selected than user $j$ if
  - $R_{i}^{max} > R_{j}^{max}$ (maximum data rate),
  - $f_i(\gamma) > f_j(\gamma), \forall \gamma$ (transmission scheme), or
  - $A_i < A_j$ (transmission environment)

*Our user selection strategy provides a simple and efficient selection strategy for heterogeneous users*
Numerical results

- Model path gain considering distance loss and log-normally distributed slow shadowing
- Two classes of users, for a user in class $i$,
  \[ f_i(\gamma) = c_i \left\{ \frac{1}{1 + e^{-a_i(\gamma - b_i) - d_i}} \right\} \]
- Compare with the single-server system
Numerical results (cont’d)

| $R_1^{max}$ | 1562.5 | 6250 | 25000 |
| Selection ratio of class 1 | 0.501 | 0.388 | 0.198 |
| Selection ratio of class 2 | 0.568 | 0.392 | 0.020 |
| Utility (Our)/Utility (Single) | 3.415 | 3.854 | 1.016 |

- $f_1 = f_2$
- $R_1^{max} \neq R_2^{max} \ (R_2^{max} = 6250)$
- Selection ratio of class $i$: the ratio of the number of selected users to the number of users in class $i$
## Numerical results (cont’d)

<table>
<thead>
<tr>
<th>$b_1$</th>
<th>2.5</th>
<th>3.5</th>
<th>4.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection ratio of class 1</td>
<td>0.566</td>
<td>0.391</td>
<td>0.230</td>
</tr>
<tr>
<td>Selection ratio of class 2</td>
<td>0.288</td>
<td>0.389</td>
<td>0.484</td>
</tr>
<tr>
<td>Utility (Our)/Utility (Single)</td>
<td>4.196</td>
<td>3.852</td>
<td>3.525</td>
</tr>
</tbody>
</table>

- $R_1^{max} = R_2^{max}$
- $f_1 \neq f_2 \ (a_1 = a_2, \ b_2 = 3.5)$
- If $b_i < b_j$, then $f_i(\gamma) \geq f_j(\gamma), \ \forall \gamma$
  - class $i$ has a more efficient transmission scheme than class $j$
Numerical results (cont’d)

<table>
<thead>
<tr>
<th></th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio of class 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selection ratio of class 1</td>
<td>0.849</td>
<td>0.653</td>
<td>0.499</td>
</tr>
<tr>
<td>Selection ratio of class 2</td>
<td>0.004</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Utility (Our)/Utility (Single)</td>
<td>3.409</td>
<td>3.912</td>
<td>3.980</td>
</tr>
</tbody>
</table>

- $R_{1}^{max} = R_{2}^{max}$
- $f_1 = f_2$
- Class 1: inner region
- Class 2: outer region
Part II

Opportunistic power scheduling for the downlink in multi-server wireless systems
Why opportunistic scheduling?

- Trade-off between efficiency and fairness due to:
  - multi-class users
  - *time-varying and location-dependent channel condition*

- Our previous problem:
  - high system efficiency
  - however, unfair to some (inefficient) users

- Fairness:
  - achieved by an appropriate scheduling scheme

- **Opportunistic scheduling** considering each user’s:
  - delay tolerance
  - fairness or performance constraint
  - *time-varying channel condition*
Single-server vs. Multi-server

- **Single-server scheduling**
  - Only one user can be scheduled in a time-slot
  - In every time-slot, must decide
    - *which user* must be selected

- **Multi-server scheduling**
  - Multiple users can be scheduled in a time-slot
  - In every time-slot, must decide
    - *how many and which users* must be selected
    - *how much power* is allocated to each selected user

- Most work studied single-server scheduling

- *However, single-server scheduling can be inefficient*
  - *Need dynamic multi-server scheduling*
Related work

- **Single-server scheduling**
  - Qualcomm’s HDR: proportional fairness
  - Borst and Whiting [INFOCOM01]: constraint on utility based fairness
  - Liu, Chong, and Shroff [JSAC01, COMNET03]: constraints on minimum performance, and utility and resource based fairness

- **Multi-server scheduling**
  - Kulkarni and Rosenberg [MSWiM03]: static multi-server scheduling with independent interfaces
  - Liu and Knightly [INFOCOM03]: dynamic multi-server scheduling with constraint on utility based fairness assuming orthogonality among users and linear relationship between data rate and power allocation
Our work

- Dynamic multi-server scheduling for downlink in a single cell
- Allow users to interfere with each other
- $P_T$ is total transmission power at the base-station
- Utility function $U_i$ for user $i$: convex, concave, or "S" function
- In each time-slot, system is in one of the states $\{1, 2, \ldots, S\}$
  - corresponds to channel conditions of all users
  - stationary stochastic process with $\text{Prob}\{\text{state } s\} = \pi_s$
  - time-varying channel condition of each user is modeled as a discrete state stationary stochastic process
- Requirement for each user
  - resource based fairness
  - utility based fairness
  - minimum performance
SINR and utility function

- SINR for user $i$ when system is in state $s$

$$\gamma_{s,i}(P_{s,i}) = \frac{N_i P_{s,i}}{\theta(P_T - P_{s,i}) + A_{s,i}}$$

- Define

$$U_{s,i}(P_{s,i}) \triangleq U_i(\gamma_{s,i}(P_{s,i}))$$

- The utility function varies randomly according to the channel condition
Problem formulation with minimum performance

\[
(C) \quad \max_{P_{s,i}} \sum_{i=1}^{M} E\{U_i\}(= \sum_{i=1}^{M} \sum_{s=1}^{S} \pi_s U_{s,i}(P_{s,i}))
\]

s. t. \quad E\{U_i\}(= \sum_{s=1}^{S} \pi_s U_{s,i}(P_{s,i})) \geq C_i, \quad i = 1, 2, \cdots, M

\[
\sum_{i=1}^{M} P_{s,i} \leq P_T, \quad s = 1, 2, \cdots, S
\]

\[
0 \leq P_{s,i} \leq P_T, \quad \forall s, i
\]

Goal: Obtaining power scheduling that maximizes the expected total system utility with constraints on the minimum expected utility for each user and the total transmission power limit for the base-station
Problem with minimum performance (cont’d)

Main difficulties

- **Feasibility**
  - assume that the system has call admission control ensuring a feasible solution

- **Non-convexity**
  - dual approach

- **No knowledge for the underlying probability distribution a priori**
  - stochastic subgradient algorithm
Power scheduling

In each time-slot $n$, power is allocated to users by solving the dual of

\[
\text{(E)} \quad \max \sum_{i=1}^{M} U_{s(n),i}^{mp} (\bar{\mu}^{(n)}, P_{s(n),i})
\]

s. t. \quad \sum_{i=1}^{M} P_{s(n),i} \leq P_T

\quad 0 \leq P_{s(n),i} \leq P_T, \quad i = 1, 2, \cdots, M

- $U_{s(n),i}^{mp} (\bar{\mu}^{(n)}, P_{s(n),i}) \triangleq (1 + \mu_{i}^{(n)})U_{s(n),i}(P_{s(n),i})$

- Similar to our previous problem
Power scheduling (Cont’d)

- The utility function \( \mu_i \) is adjusted to guarantee the minimum performance constraint by using a stochastic subgradient algorithm

\[
\mu_i^{(n+1)} = [\mu_i^{(n)} - \alpha(n) v_i^{(n)}]^+, \forall i
\]

- \( v_i^{(n)} = U_{s(n),i}(P_{s(n),i}(\bar{\mu}(n))) - C_i \)

  - stochastic subgradient of the dual
  - \( P_{s(n),i}(\bar{\mu}(n)) \) is power allocation of user \( i \) in time-slot \( n \)

- \( \bar{\mu}(n) \) converges to \( \bar{\mu}^* \) that solves the dual problem
Feasibility

- Always satisfies the constraint on total transmission power limit

- If \( \frac{Q}{M} \rightarrow 0 \) as \( M \rightarrow \infty \), then

\[
\frac{\sum_{i \in H} U_i^* - C_i}{M} \rightarrow 0 \quad \text{as} \quad M \rightarrow \infty
\]

- \( Q \): expected number of users with the same channel conditions
- \( H \): set of users whose performance constraints are not satisfied
- \( U_i^* \): expected utility of user \( i \) in our power scheduling

- Asymptotically feasible on average

- \textit{Increase in the randomness of the system improves the degree of users’ satisfaction}
Optimality

If \( \sum_{i=1}^{M} U_i^o \to \infty \) and \( \frac{Q}{M} \to 0 \) as \( M \to \infty \), then

\[
\frac{\sum_{i=1}^{M} U_i^*}{\sum_{i=1}^{M} U_i^o} \geq 1 - \epsilon \quad \text{and} \quad \epsilon \to 0 \quad \text{as} \quad M \to \infty
\]

- \( U_i^o \): expected utility of user \( i \) in optimal power scheduling
- Asymptotically optimal
Numerical results

- The same cellular model as our previous problem
- Four users and each user $i$
  - same sigmoid utility function $U_i \ (U_i(0) = 0 \text{ and } U_i(\infty) = 1)$
  - same performance constraint $C_i = 0.59$
  - distance from the base-station to user $i$: $d_i$
    
    $d_1 < d_2 < d_3 < d_4$
  
- Performance comparison with
  - Non-opportunistic scheduling
  - Greedy scheduling
### Comparison of average utilities ($10^4$ time-slots)

<table>
<thead>
<tr>
<th>User</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-opportunistic</td>
<td>0.590</td>
<td>0.590</td>
<td>0.590</td>
<td>0.590</td>
<td>2.360</td>
</tr>
<tr>
<td>Greedy</td>
<td>0.973</td>
<td>0.964</td>
<td>0.796</td>
<td>0.168</td>
<td>2.901</td>
</tr>
<tr>
<td>Our opportunistic</td>
<td>0.951</td>
<td>0.736</td>
<td>0.591</td>
<td>0.591</td>
<td>2.869</td>
</tr>
</tbody>
</table>
Ratio of average total system utility of our opportunistic power scheduling to that of non-opportunistic power scheduling

$\sigma$: standard deviation of each user’s channel condition
Conclusion

- Utility framework
  - suitable for resource allocation with multi-media and data services
  - a useful tool for resource allocation in the next generations of communication networks
  - non-convex optimization problems in many cases

- Dual approach provides
  - efficient solution in many cases
  - simple algorithm that can be easily implemented with a (distributed) network protocol
Conclusion (cont’d)

In wireless systems

- Single server transmission is optimal only when all users have high data rate
- In general, need dynamic multiple access (dynamic multi-server system)
- Trade-off between efficiency and fairness
  - Opportunistic scheduling achieves both of them
- Randomness of the system could be beneficial to efficient and fair resource allocation, if appropriately exploited
Conclusion (cont’d)

Other problems

- **Pricing based base-station assignment**
  - considers both transmission environment of the user and congestion level of the base-station

- **Congestion control on the Internet**
  - algorithms for concave utility functions cause instability and congestion in the presence of real-time services with non-concave utility functions
  - self-regulating property stabilizes the system and alleviates congestion
Future work

- Scheduling considering
  - user dynamics
  - non-stationary environment
  - delay or short-term fairness constraints

- Resource allocation considering upper layer protocols (e.g., TCP)

- Resource allocation for uplink and multi-cellular system