Optimal Operating Point in MIMO Channel for Delay-Sensitive and Bursty Traffic

Tara Javidi

University of California, San Diego

Joint work with my graduate student: Somsak Kittipiyakul

Problem Motivation

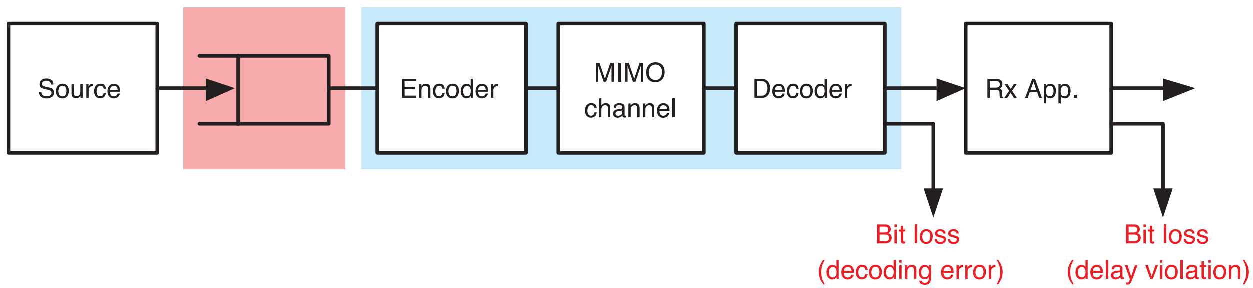
- Consider a MIMO system providing various types of gains:
  - Spatial Diversity gain: Multiple antennas increase diversity to combat channel fading.
    - The spatial path diversity allows for an lower error probability
  - Multiplexing gain: Multiple antennas increase data rate by multiplexing independent data over spatial channels.
    - Independent data can be multiplexed over spatially separated channels
  - Multi-user Diversity gain: Existence of multiple users allow for opportunistic use of channel
  - These gains are to be traded off against each other
    - Depends on the particulars of the problem, e.g. availability of channel state information

Problem Motivation

- Consider a MIMO system providing various types of gains:
  - Spatial Diversity gain: Multiple antennas increase diversity to combat channel fading.
    - The spatial path diversity allows for an lower error probability
  - Multiplexing gain: Multiple antennas increase data rate by multiplexing independent data over spatial channels.
    - Independent data can be multiplexed over spatially separated channels
  - Multi-user Diversity gain: Existence of multiple users allow for opportunistic use of channel
  - These gains are to be traded off against each other
    - Depends on the particulars of the problem, e.g. availability of channel state information
- What does this mean from a cross-layer perspective?
- How to go about trading these gains, from the perspective of a delay sensitive user with stochastic traffic demand?

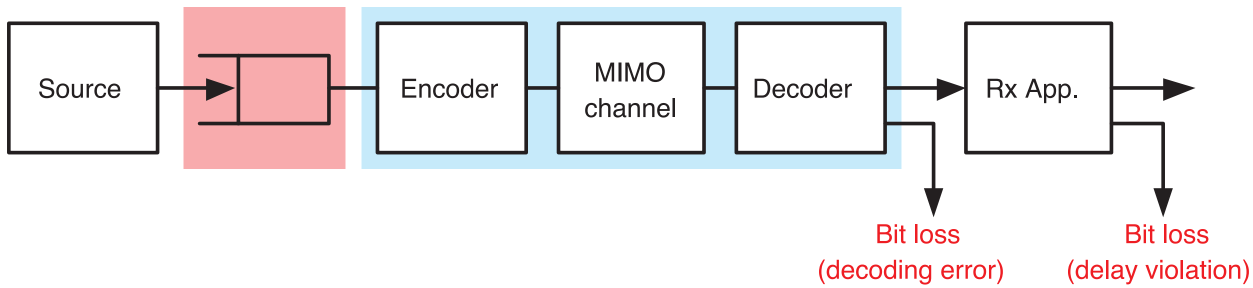
MIMO Network Design for Delay Sensitive Applications

- Consider an information source concatenated by a MIMO channel



MIMO Network Design for Delay Sensitive Applications

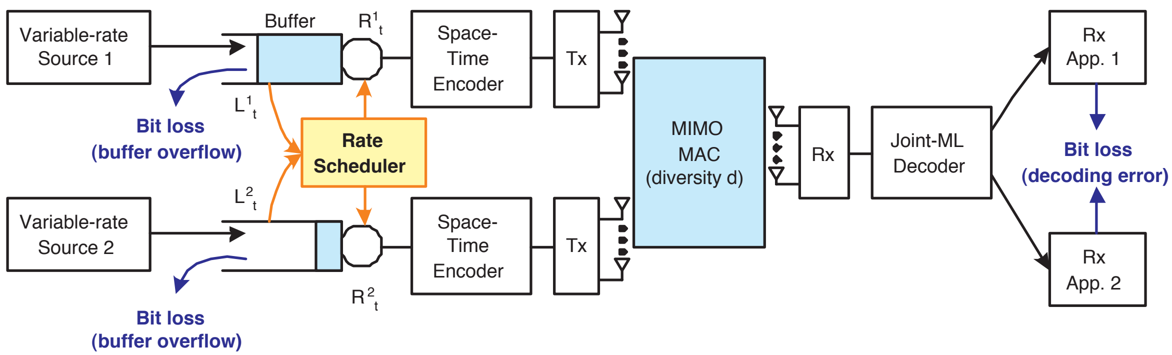
- Consider an information source concatenated by a MIMO channel



- Source of information bits
  - Bits arrive according to a Stochastic Process (modeling bursty traffic)
  - Delay sensitivity of application $\Rightarrow$ Delay QoS measure
- MIMO channel
  - M transmit and N receive antennas
  - Perfect CSI-R and no CSI-Tx
  - Space-time coding block by block (Raleigh fading, T_c large)
  - No retransmissions are considered

MIMO Network Design for Delay Sensitive Applications

- Consider an information source concatenated by a MIMO channel



- Source of information bits
  - Bits arrive according to a Stochastic Process (modeling bursty traffic)
  - Delay sensitivity of application ⇒ Delay QoS measure
- MIMO channel
  - M transmit and N receive antennas
  - Perfect CSI-R and no CSI-Tx
  - Space-time coding block by block (Raleigh fading, T_c large)
  - No retransmissions are considered

QoS and Performance Measure

- Overall performance degradation consists of two factors
  - Decoding error over the wireless medium
    - erroneous bits are lost
  - Large delay in the queue
    - bits facing large delays are obsolete, i.e. are considered lost

QoS and Performance Measure

- Overall performance degradation consists of two factors
  - Decoding error over the wireless medium
    - erroneous bits are lost
  - Large delay in the queue
    - bits facing large delays are obsolete, i.e. are considered lost

Objective: Joint design of optimal PHY and MAC layers to account for both forms of performance degradation

QoS and Performance Measure

- Overall performance degradation consists of two factors
  - Decoding error over the wireless medium
    - erroneous bits are lost
  - Large delay in the queue
    - bits facing large delays are obsolete, i.e. are considered lost

Objective: Joint design of optimal PHY and MAC layers to account for both forms of performance degradation

Intuition: Given a fixed scheduler design

- The decoding error over MIMO channel is decreasing as diversity increases
- The queueing delay is decreasing as multiplexing rate increases

Prior Related Work

- The question of optimal operating point in MIMO was posed by Holliday and Goldsmith (Allerton '04):
  - Context: Cross-layer optimization of source and MIMO encoders
  - Objective: Minimize end-to-end distortion due to source quantization error and channel decoding error
- In this context, our paper compliments the above by accounting for bursty arrivals and queuing delay
  - We need to find the right combination of performance measures

A sense of dèjá vu?

A sense of dèjá vu?

- Delay optimal rate adaptive channel codes for bursty traffic
  - What is the delay optimal coding strategy?
    - Hagenauer et. al. '90, Negi et. al. '04, Javidi et. al. '06
  - Intuitively, the larger the coding block the higher the rate, while there is a delay associated with waiting for a long block

A sense of dèjá vu?

- Delay optimal rate adaptive channel codes for bursty traffic
  - What is the delay optimal coding strategy?
    - Hagenauer et. al. '90, Negi et. al. '04, Javidi et. al. '06
  - Intuitively, the larger the coding block the higher the rate, while there is a delay associated with waiting for a long block
- Delay optimal power control over wireless channels for bursty traffic
  - What is the energy-delay trade-off?
    - Negi et. al '00, Berry et. al. '02, Rajan et. al., '04
  - Intuitively, lower transmission rates are more energy efficient while resulting in longer delay

A sense of dèjá vu?

- Delay optimal rate adaptive channel codes for bursty traffic
  - What is the delay optimal coding strategy?
    - Hagenauer et. al. '90, Negi et. al. '04, Javidi et. al. '06
  - Intuitively, the larger the coding block the higher the rate, while there is a delay associated with waiting for a long block
- Delay optimal power control over wireless channels for bursty traffic
  - What is the energy-delay trade-off?
    - Negi et. al '00, Berry et. al. '02, Rajan et. al., '04
  - Intuitively, lower transmission rates are more energy efficient while resulting in longer delay
- In both cases, the greatest challenge arises at empty buffer boundaries
  - queue dynamics are non-linear at the boundary and the impact grows away from the boundary

A sense of dèjá vu?

- Delay optimal rate adaptive channel codes for bursty traffic
  - What is the delay optimal coding strategy?
    - Hagenauer et. al. '90, Negi et. al. '04, Javidi et. al. '06
  - Intuitively, the larger the coding block the higher the rate, while there is a delay associated with waiting for a long block
- Delay optimal power control over wireless channels for bursty traffic
  - What is the energy-delay trade-off?
    - Negi et. al '00, Berry et. al. '02, Rajan et. al., '04
  - Intuitively, lower transmission rates are more energy efficient while resulting in longer delay
- In both cases, the greatest challenge arises at empty buffer boundaries
  - queue dynamics are non-linear at the boundary and the impact grows away from the boundary
- In both cases, large deviation techniques provide significant insight by focusing on the atypical sequence of arrivals (causing buffer overflow) and the atypical sequence of channel errors

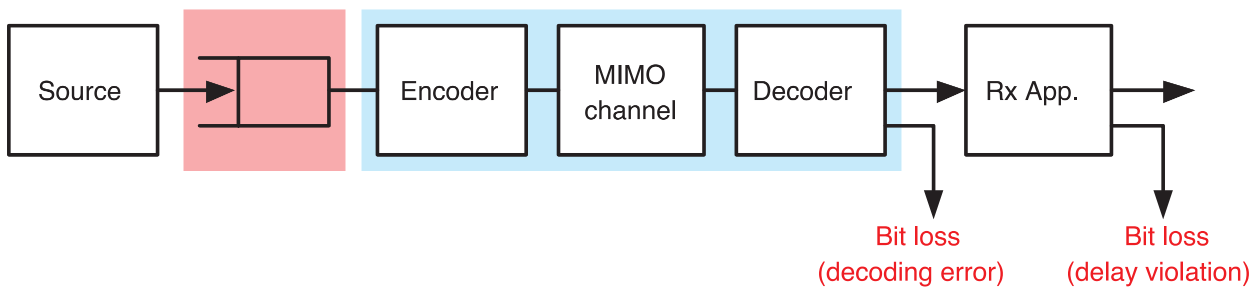
Outline

→ Problem formulation and a summary of results

- Delay optimal diversity gain in point-point setting
  - Characterize components of the system: MIMO channel, and delay sensitive bursty traffic source
  - Compute the total loss probability in terms of choice of MIMO parameter diversity d, in high SNR regime
  - Optimization of diversity gain d
- Delay optimal diversity in MIMO MAC with rate scheduling
  - Revisiting the above issues in a multi-user context
- Conclusion and Summary

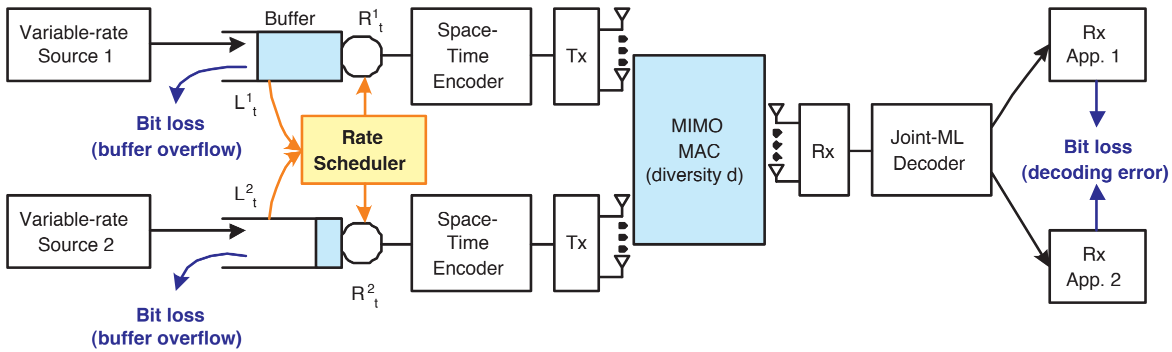
Problem Formulation: Model

- Consider an information source concatenated by a MIMO channel



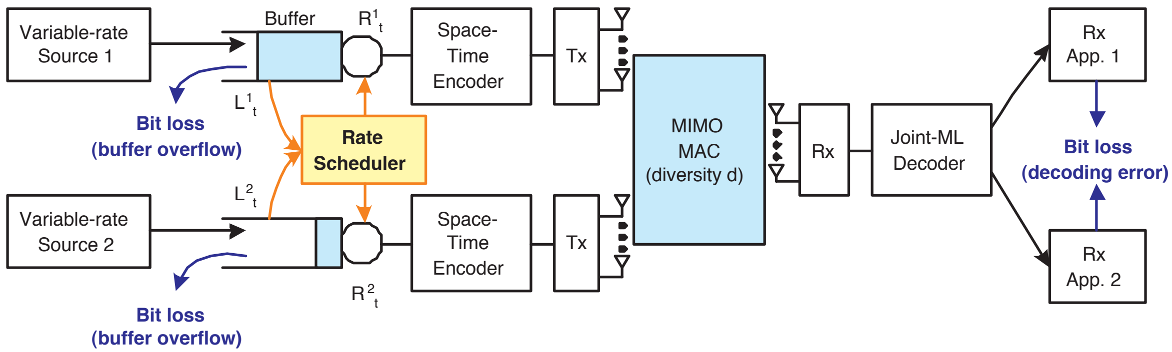
Problem Formulation: Model

- Consider an information source concatenated by a MIMO channel



Problem Formulation: Model

- Consider an information source concatenated by a MIMO channel



- Source of information bits
  - Bits arrive according to a Stochastic Process (modeling bursty traffic)
  - Quality of Service is of a (large) delay violation form
- MIMO channel
  - M transmit and N receive antennas
  - Perfect CSI-R and no CSI-Tx
  - Raleigh block fading; Space-time coding block by block
- No retransmissions are considered

Problem Formulation: Objective

- Overall performance degradation is measured via
  - Probability of decoding error over the wireless medium
    - $P_e =$ Probability of bit decoding error
  - Problem of buffer overflow resulting in long delay
    - $P_q = P(Q \geq B)$
  - Describe the total bit loss probability
    - $P_t \approx P_q + P_e$

Problem Formulation: Objective

- Overall performance degradation is measured via
  - Probability of decoding error over the wireless medium
    - $P_e =$ Probability of bit decoding error
  - Problem of buffer overflow resulting in long delay
    - $P_q = P(Q \geq B)$
  - Describe the total bit loss probability
    - $P_t \approx P_q + P_e$
- Channel and queue resources to drive $P_t \rightarrow 0$

Problem Formulation: Objective

- Overall performance degradation is measured via
  - Probability of decoding error over the wireless medium
    - $P_e =$ Probability of bit decoding error
  - Problem of buffer overflow resulting in long delay
    - $P_q = P(Q \geq B)$
  - Describe the total bit loss probability
    - $P_t \approx P_q + P_e$
- Channel and queue resources to drive $P_t \to 0$
  - When $SNR \to \infty$, $P_e \to 0$
    - The rate of decay is determined by the well-known Diversity-Multiplexing trade-off for MIMO channel
    - The arrival rate scales with $\log SNR$

Problem Formulation: Objective

- Overall performance degradation is measured via
  - Probability of decoding error over the wireless medium
    - $P_e =$ Probability of bit decoding error
  - Problem of buffer overflow resulting in long delay
    - $P_q = P(Q \geq B)$
  - Describe the total bit loss probability
    - $P_t \approx P_q + P_e$
- Channel and queue resources to drive $P_t \to 0$
  - When $SNR \to \infty$, $P_e \to 0$
    - The rate of decay is determined by the well-known Diversity-Multiplexing trade-off for MIMO channel
    - The arrival rate scales with $\log SNR$
  - As buffer size $B \to \infty$, $P_q \to 0$
    - The rate of decay is determined by the source's *effective bandwidth*

Problem Formulation: Objective

- Overall performance degradation is measured via
  - Probability of decoding error over the wireless medium
    - $P_e =$ Probability of bit decoding error
  - Problem of buffer overflow resulting in long delay
    - $P_q = P(Q \geq B)$
  - Describe the total bit loss probability
    - $P_t \approx P_q + P_e$
- Channel and queue resources to drive $P_t \rightarrow 0$
  - When $SNR \rightarrow \infty$, $P_e \rightarrow 0$
    - The rate of decay is determined by the well-known Diversity-Multiplexing trade-off for MIMO channel
    - The arrival rate scales with $\log SNR$
  - As buffer size $B \rightarrow \infty$, $P_q \rightarrow 0$
    - The rate of decay is determined by the source's *effective bandwidth*
- Put $B \propto \log SNR$ ($\Rightarrow P(delay > D) \approx P_q$)

Summary of Approach

- For a given large SNR, the rates of error events are determined by MIMO encoder's spatial diversity gain d (the PHY parameter)

Summary of Approach

- For a given large SNR, the rates of error events are determined by MIMO encoder's spatial diversity gain d (the PHY parameter)
- In presence of multi-user, the rates of above events are also determined by scheduler design (the MAC parameter)

Summary of Approach

- For a given large SNR, the rates of error events are determined by MIMO encoder's spatial diversity gain d (the PHY parameter)
- In presence of multi-user, the rates of above events are also determined by scheduler design (the MAC parameter)

Objective: Choose an order of spatial diversity, d, such that it minimizes the total loss probability

$$P_t(d) \approx P_q(d) + P_e(d) \doteq SNR^{-f(d)} + SNR^{-d}$$

Summary of Approach

- For a given large SNR, the rates of error events are determined by MIMO encoder's spatial diversity gain d (the PHY parameter)
- In presence of multi-user, the rates of above events are also determined by scheduler design (the MAC parameter)

Objective: Choose an order of spatial diversity, d, such that it minimizes the total loss probability

$$P_t(d) \approx P_q(d) + P_e(d) \doteq SNR^{-f(d)} + SNR^{-d}$$

- $P_e(d)$ is the decoding error over MIMO channel; decreasing in d
- $P_q(d)$ is the queueing delay violation; increasing in d
- In the case of MIMO MAC, we seek an optimal scheduler

Summary of Approach

- For a given large SNR, the rates of error events are determined by MIMO encoder's spatial diversity gain d (the PHY parameter)
- In presence of multi-user, the rates of above events are also determined by scheduler design (the MAC parameter)

Objective: Choose an order of spatial diversity, d, such that it minimizes the total loss probability

$$P_t(d) \approx P_q(d) + P_e(d) \doteq SNR^{-f(d)} + SNR^{-d}$$

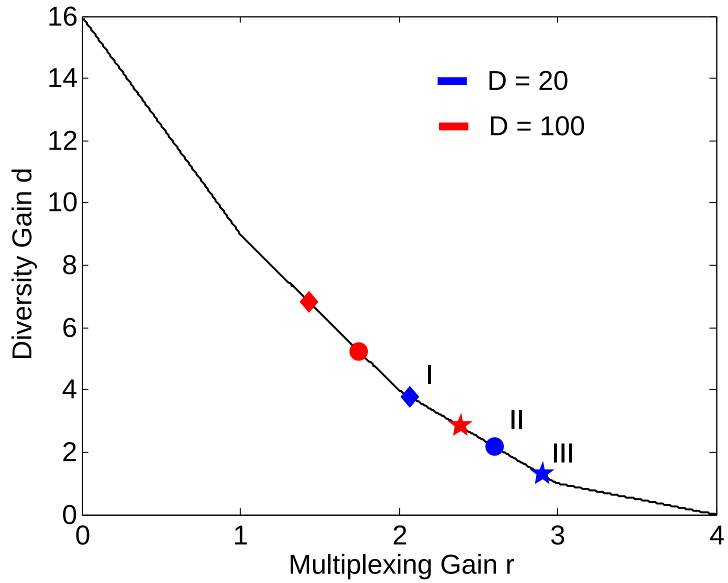
  - $P_e(d)$ is the decoding error over MIMO channel; decreasing in d
  - $P_q(d)$ is the queueing delay violation; increasing in d
  - In the case of MIMO MAC, we seek an optimal scheduler

- Find the optimal spatial diversity gain d^*
  - Balance the exponents of both terms of $P_t(d)$ at the same order.

Overview of the Result
Point-Point Case

- At high SNR, source is able to capitalize on diversity offered at MIMO as
  - The delay tolerance is increased
  - The source becomes less bursty
- We consider special cases of three source models:

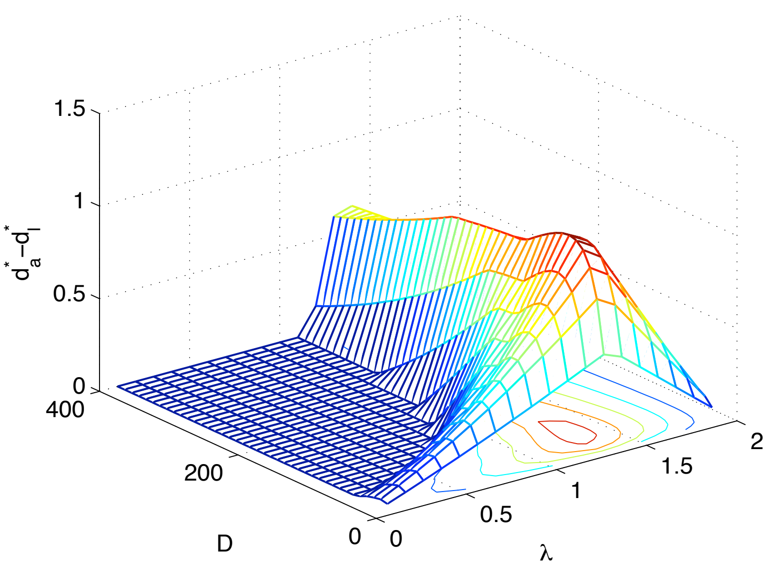
Order of source burtiness: Source III $>$ II $>$ I



Overview of the Result
Multi-access Case

- At high SNR, source is able to capitalize on diversity offered at MIMO as
  - The delay tolerance is increased
- Furthermore, quantified the improvements due to MAC layer scheduling
- The relationship to burstiness, though, is non-monotonic
  - At medium arrival rates, burstiness helps but not at high rates

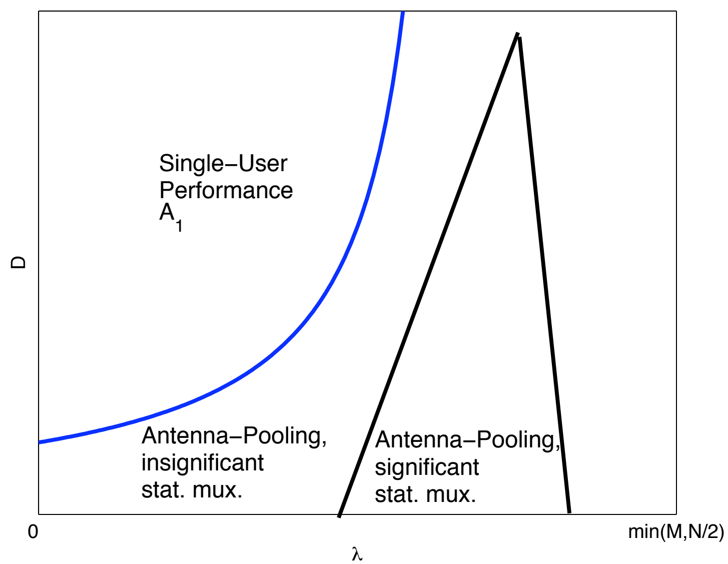
Statistical Multiplexing vs. delay bound and arrival rates



Overview of the Result
Multi-access Case

- At high SNR, source is able to capitalize on diversity offered at MIMO as
  - The delay tolerance is increased
- Furthermore, quantified the improvements due to MAC layer scheduling
- The relationship to burstiness, though, is non-monotonic
  - At medium arrival rates, burstiness helps but not at high rates

Performance Regimes and Scheduling Gain



Outline

- Problem formulation and a summary of results

→ Delay optimal diversity gain in point-point setting

  - Characterize components of the system: MIMO channel, and delay sensitive bursty traffic source
  - Compute the total loss probability in terms of choice of MIMO parameter diversity d, in high SNR regime
  - Optimization of diversity gain d

- Delay optimal diversity in MIMO MAC with rate scheduling

  - Revisiting the above issues in a multi-user context

- Conclusion and Summary

Point-Point MIMO Channel Model

- MIMO Channel: M transmit and N receive antennas, Rayleigh block fading (coherence time $\approx T_c \geq T \geq KM + N - 1$), perfect CSI-R, no CSI-Tx
- For each SNR value, interested in code with data rate $R(SNR)$ scaled with SNR and average error probability $P_e(SNR)$, achieving multiplexing and diversity gains, r and d

$$r := \lim_{SNR \to \infty} \frac{R(SNR)}{\log SNR} \qquad d := \lim_{SNR \to \infty} -\frac{\log P_e(SNR)}{\log SNR}$$

Diversity-Multiplexing tradeoff: The optimal diversity gain $d^*(r)$ achieved by any code scheme of block length T and multiplexing gain r is given by

$$d^*(r) = (M - r)(N - r)$$

piecewise linear, for $0 \leq r \leq \min(M, N)$.

- The channel error probability is

$$P_e(r) \doteq SNR^{-d^*(r)}$$

  – $P_e(r)$ is increasing in r

Source Model

- The source is assumed to be stochastic, buffered, and delay sensitive
  - If any given bit arrives to a queue whose length is greater than a threshold B, it will be obsolete and considered lost
  - Performance measure of interest

$$P_q = \mathbf{P}\left[Q > B\right]$$

Source Model

- The source is assumed to be stochastic, buffered, and delay sensitive
  - If any given bit arrives to a queue whose length is greater than a threshold B, it will be obsolete and considered lost
  - Performance measure of interest

$$P_q = \mathbf{P}\left[Q > B\right]$$

- Stationary and ergodic arrival stream $\{X_t\}$, with known distribution and *effective bandwidth*

$$\alpha(\delta) = \lim_{t \to \infty} \frac{1}{\delta t} \log \mathbf{E}\left[e^{\delta(X_1 + \ldots + X_t)}\right]$$

Source Model

- The source is assumed to be stochastic, buffered, and delay sensitive
  - If any given bit arrives to a queue whose length is greater than a threshold B, it will be obsolete and considered lost
  - Performance measure of interest

$$P_q = \mathbf{P}\left[Q > B\right]$$

- Stationary and ergodic arrival stream $\{X_t\}$, with known distribution and *effective bandwidth*

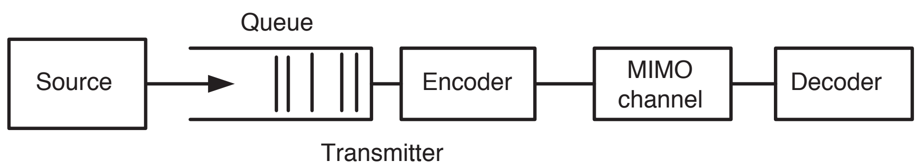
$$\alpha(\delta) = \lim_{t\to\infty} \frac{1}{\delta t} \log \mathbf{E}\left[e^{\delta(X_1+\ldots+X_t)}\right]$$

- Intuitively, a server with capacity $\alpha(\delta)$ is needed to guarantee low buffer overflow probability

$$\mathbf{P}\left[Q > B\right] \approx e^{-\delta B}$$

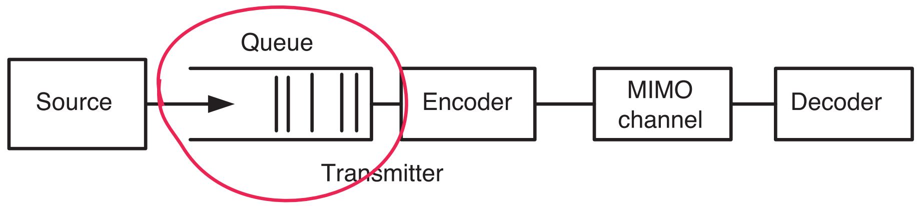
  - $\alpha(\cdot)$ is an increasing function
  - $\alpha(0)$ and $\alpha(\infty)$ are average arrival rate and the peak rate

Buffer Scaling and Delay Violation Probability

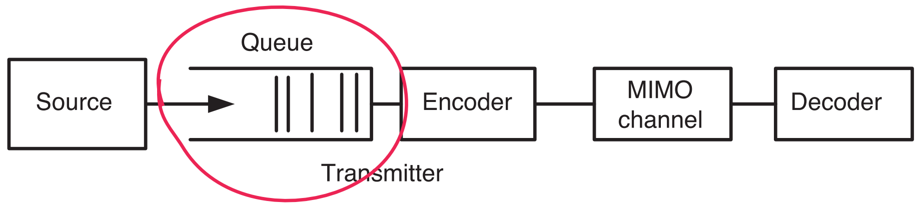


(1)

Buffer Scaling and Delay Violation Probability



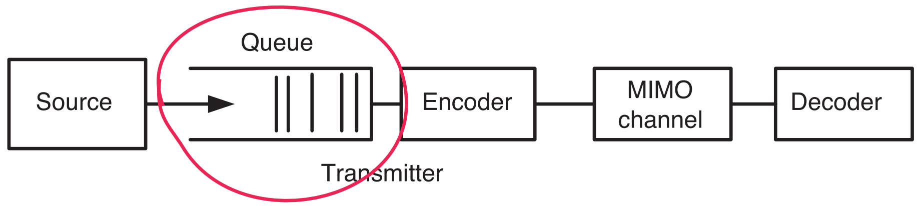
Buffer Scaling and Delay Violation Probability



- Single server queue with server capacity $C = Tr \log SNR$

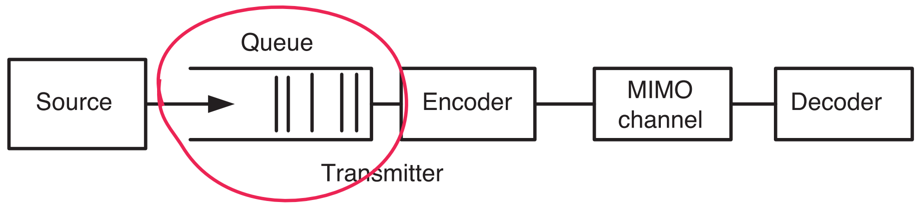
(3)

Buffer Scaling and Delay Violation Probability



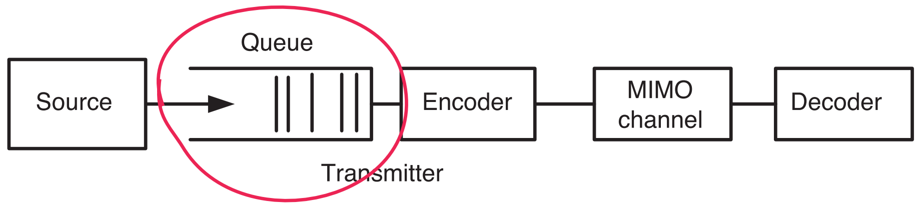
- Single server queue with server capacity $C = Tr \log SNR$
- Since server capacity is logarithmic in SNR, we consider a scaling of
  - Arrival rate $\lambda T \log SNR$
  - Buffer size $B = DC = DTr \log SNR$

Buffer Scaling and Delay Violation Probability



- Single server queue with server capacity $C = Tr \log SNR$
- Since server capacity is logarithmic in SNR, we consider a scaling of
    - Arrival rate $\lambda T \log SNR$
    - Buffer size $B = DC = DTr \log SNR$
- Intuitively, queue length bound $B = DC$ results in a delay bound D

Buffer Scaling and Delay Violation Probability



- Single server queue with server capacity $C = Tr \log SNR$
- Since server capacity is logarithmic in SNR, we consider a scaling of
  - Arrival rate $\lambda T \log SNR$
  - Buffer size $B = DC = DTr \log SNR$
- Intuitively, queue length bound $B = DC$ results in a delay bound D
- For B sufficiently large, the buffer overflow probability is

$$P_q(r) = \mathbf{P}\left[Q > DC\right] \doteq SNR^{-\delta(r)DTr} \tag{6}$$

where

$$\alpha(\delta(r)) = Tr \log SNR.$$

  - $P_q(r)$ is decreasing in r

Optimal Operating Point in MIMO Point-Point

- With $P_q(r)$ and $P_e(r)$ at hand we can compute the asymptotic optimal operating point

- Total loss probability is given as

$$P_t(r) \approx P_q(r) + P_e(r) \doteq SNR^{-\delta(r)DTr} + SNR^{-d^*(r)} \tag{7}$$

- Optimal r^* (minimizing P_t) in asymptotic high SNR regime is given by

$$\delta(r^*)DTr^* = d^*(r^*)$$

Example: Source Model I

- The arrivals are iid Compound Poisson with fixed packet size

- The number of bits arrived in timeslot t is

$$X_t = \sum_{i=1}^{N} Y_i$$

  - N is Poisson random variable with rate ν
  - $Y_1, Y_2, \ldots$ are fixed length packets of size $1/\mu$
  - The average bit arrival rate ν/μ scaled with $\log SNR$, i.e. $\nu/\mu = \lambda T \log SNR$ for a given constant λ.

- Compute

$$\text{Effective bandwidth:} \qquad \alpha(\delta) = \frac{\nu(e^{\delta/\mu} - 1)}{\delta}$$

Example: Source Model II

- The arrivals are iid Compound Poisson
- The number of bits arrived in timeslot t is

$$X_t = \sum_{i=1}^{N} Y_i$$

  - N is Poisson random variable with rate ν
  - $Y_1, Y_2, \ldots$ are iid exponential random variables with mean $1/\mu$
  - The average bit arrival rate ν/μ scaled with $\log SNR$, i.e. $\nu/\mu = \lambda T \log SNR$ for a given constant λ.

- Compute

Effective bandwidth: $$\alpha(\delta) = \begin{cases} \frac{\nu}{\mu-\delta} & \text{if } 0 \leq \delta < \mu, \\ \infty & \text{if } \delta \geq \mu. \end{cases}$$

Overflow exponent: $$\delta(r) = \mu(1 - \frac{\lambda}{r})$$

Example: Source Model III

- The arrivals are Markov-Modulated (ON-OFF)
- When OFF, there are no arrivals
- When ON, the number of bits arrived in timeslot t is

$$X_t = \sum_{i=1}^{N} Y_i$$

  - N is Poisson random variable with rate ν
  - $Y_1, Y_2, \ldots$ are iid exponential random variables with mean $1/\mu$.
  - The average bit arrival rate ν/μ scaled with $\log SNR$, i.e. $\nu/\mu = \lambda T \log SNR$ for a given constant λ.

- Compute the effective bandwidth:

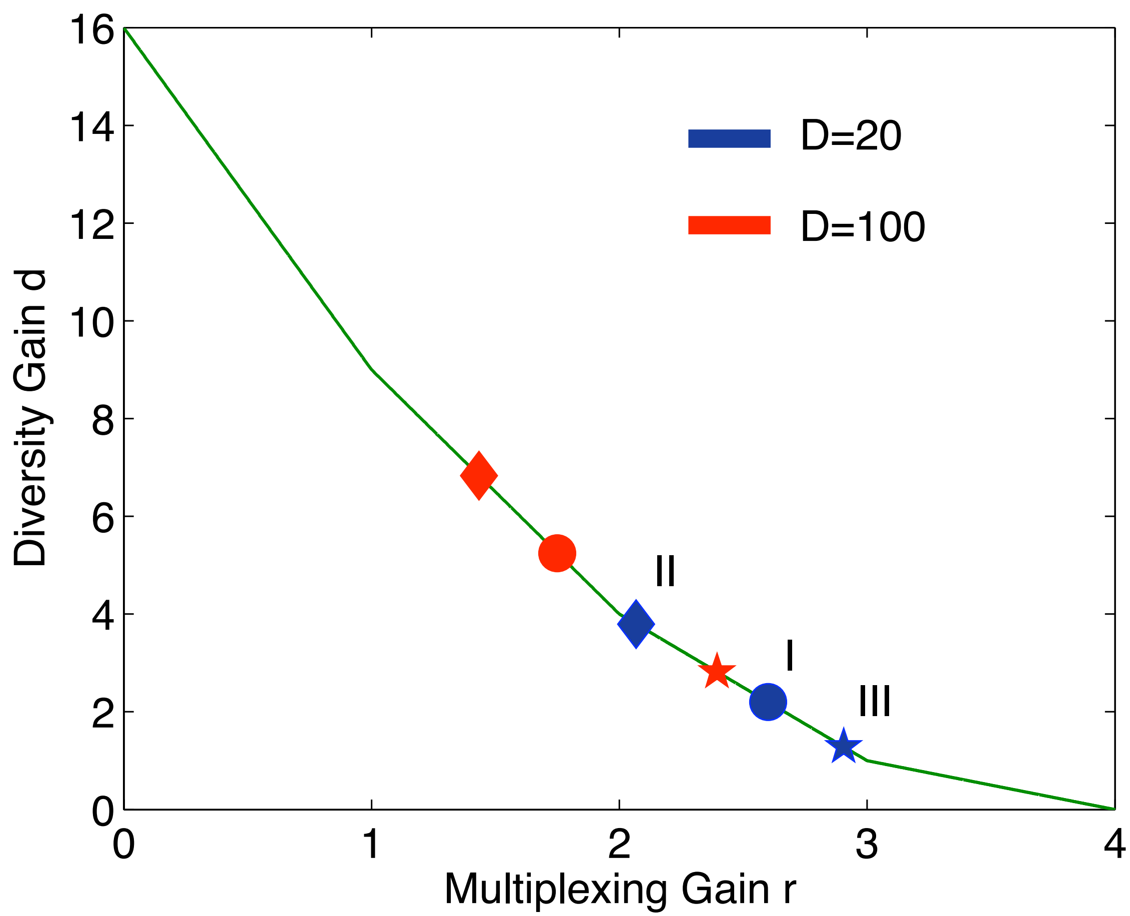
$$\alpha(\delta) = \frac{1}{\delta} \log \left[\frac{1}{2} (a(\delta) + \sqrt{a^2(\delta) + 4b(\delta)}) \right]$$

where

$$a(\delta) = p + q \exp(\frac{\delta \nu}{\mu - \delta}) \quad \text{and} \quad b(\delta) = (1 - p - q) \exp(\frac{\delta \nu}{\mu - \delta}).$$

Comparative Results for Sources I-III

Source burtiness ordering: Source III $>$ II $>$ I.



Sources with strict delay requirements and/or bursty traffic patterns must sacrifice diversity

Outline

- Problem formulation and a summary of results
- Delay optimal diversity gain in point-point setting
  - Characterize components of the system: MIMO channel, and delay sensitive bursty traffic source
  - Compute the total loss probability in terms of choice of MIMO parameter diversity d, in high SNR regime
  - Optimization of diversity gain d

$\rightarrow$ Delay optimal diversity in MIMO MAC with rate scheduling

  - Revisiting the above issues in a multi-user context

- Conclusion and Summary

Multiple Access MIMO
Source Model Revisited

- At times users need to pool (share) the antenna resources so the notion of effective bandwidth is not as useful
- Instead, we use an alternative (and equivalent) description of source

Definition: An ergodic source S is said to satisfy an LDP with decay function $\Lambda^* : \mathbb{R} \to [0, \infty]$ if, for large enough t and for small $\epsilon > 0$,

$$\mathbf{P}\left[\frac{S_t}{t} \in (a - \epsilon, a + \epsilon)\right] \approx e^{-n\Lambda^*(a)} \tag{8}$$

where Λ^* is a lower semicontinuous function and is related to limiting log moment generating function $\Lambda(\theta) := \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}[e^{\theta S_t}]$

$$\Lambda^*(a) = \sup_{\theta}[\theta a - \Lambda(\theta)]$$

Multi Access MIMO Channel Model

Definition: Let r_t^i be the multiplexing gain of user i, $i = 1, \ldots, K$, at time t. Given a common diversity d for all users, the spatial multiplexing gains $(r_t^1, \ldots, r_t^K)$ at any timeslot t must be within a multiplexing gain region

$$\mathcal{R}(d) = \left\{ (r^1, \ldots, r^K) : \sum_{s \in S} r^s \leq r^*_{|S|M,N}(d), \quad \forall S \subseteq \{1, \ldots, K\} \right\}. \tag{9}$$

where $r^*_{m,n}(d)$ is the largest multiplexing gain achieved for a given diversity d.

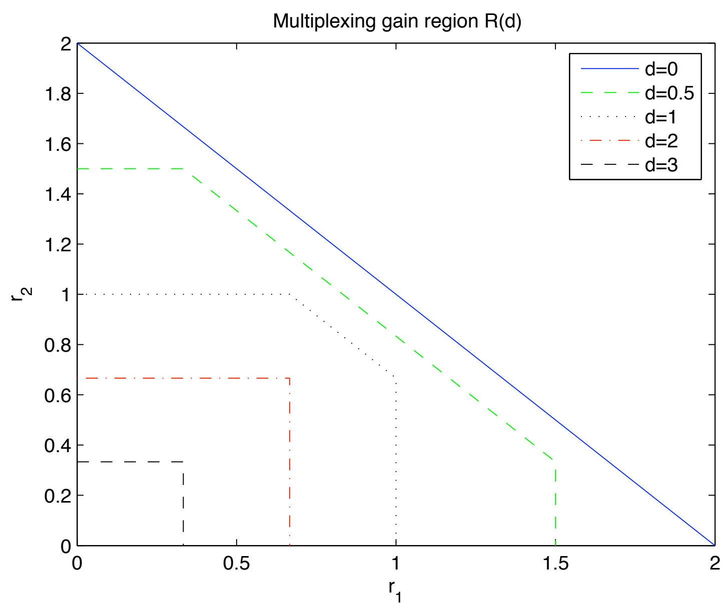


Figure 1: Example of the multiplexing gain region $\mathcal{R}(d)$ for $M = N = 2$ case. In this case, $d_0 = 2$.

Multi Access MIMO Channel Model

- Given any fixed diversity gain d, $\mathcal{R}(d)$ determines feasible (simultaneous) multiplexing rates

Multi Access MIMO Channel Model

- Given any fixed diversity gain d, $\mathcal{R}(d)$ determines feasible (simultaneous) multiplexing rates

- Not only the total rate, but also the shape of the multiplexing region $\mathcal{R}(d)$ depends on the value of d

Multi Access MIMO Channel Model

- Given any fixed diversity gain d, $\mathcal{R}(d)$ determines feasible (simultaneous) multiplexing rates
- Not only the total rate, but also the shape of the multiplexing region $\mathcal{R}(d)$ depends on the value of d
- The shape of the region determines how users share the common resources
  - The shape of the region determines the possibility of statistical multiplexing
  - The lower the diversity gain, the more conflict the users are in and vice versa

Multi Access MIMO Channel Model

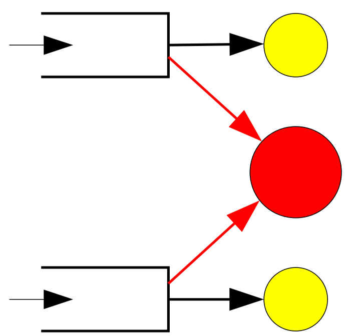
- Given any fixed diversity gain d, $\mathcal{R}(d)$ determines feasible (simultaneous) multiplexing rates
- Not only the total rate, but also the shape of the multiplexing region $\mathcal{R}(d)$ depends on the value of d
- The shape of the region determines how users share the common resources
  - The shape of the region determines the possibility of statistical multiplexing
  - The lower the diversity gain, the more conflict the users are in and vice versa
  - There exists a unique d_0 such that for $d \geq d_0$ the shape is square and for $d < d_0$ the region is a polymatriod
  - When $d \geq d_0$, users operate in isolation

Scheduler Model
(Two User Case)

- A scheduler allocates multiplexing (hence transmission) rate vector, i.e.

$$\mathcal{S} : (Q_1^t, Q_2^t) \Longrightarrow (r_1^t, r_2^t) \in \mathcal{R}(d)$$

- So give a fixed SNR and a fixed diversity, we have the following queueing model

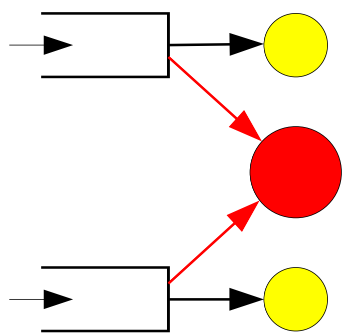


Scheduler Model
(Two User Case)

- A scheduler allocates multiplexing (hence transmission) rate vector, i.e.

$$\mathcal{S} : (Q_1^t, Q_2^t) \Longrightarrow (r_1^t, r_2^t) \in \mathcal{R}(d)$$

- So give a fixed SNR and a fixed diversity, we have the following queueing model



- **Fair static** scheduler operates at the maximum allowable equal rate $r_1^t = r_2^t = r^*_{2M,N}(d)/2$
  - Reduces to decoupled point-point case
- **Longest Queue Highest Rate** scheduler allocates the shared server dynamically to the longer queue

Buffer Overflow Decay Exponent

- Given diversity d, the total service capacity is given by $C_{tot}(d) = Tr_{tot}(d) \log SNR$ where

$$r_{tot}(d) := \begin{cases} Kr^*_{M,N}(d) & \text{if } d \geq d_0, \\ r^*_{KM,N}(d) & \text{if } d < d_0. \end{cases}$$

- Since total capacity is logarithmic in SNR, we consider a scaling of $B = \frac{1}{k} DC_{tot} = \frac{1}{k} DTr_{tot} \log SNR$
  - Intuitively, the end-to-end delay bound D converts to queue length bound B as $B \approx DC$
- For B sufficiently large, the buffer overflow probability is

$$P_q(d) := \mathbf{P}[\text{delay violation at any user}] \approx \mathbf{P}[\max_{i=1,\dots,K} L^i_t > DC_{tot}(d)]$$

Buffer Overflow Decay Exponent

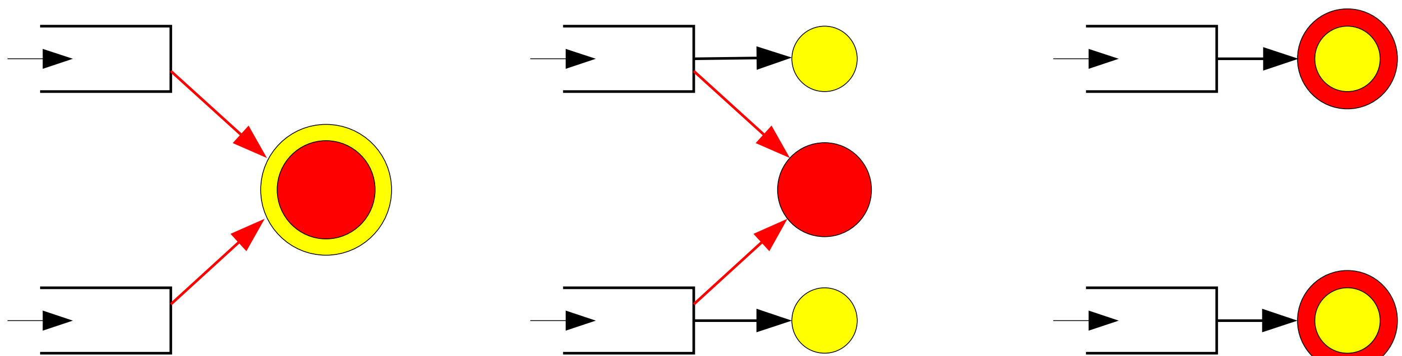
- for $d > d_0$, we derive the buffer overflow form the point-point analysis

Buffer Overflow Decay Exponent

- for $d > d_0$, we derive the buffer overflow form the point-point analysis
- for $d < d_0$, a large deviation analysis of queueing model (M) is required
  - Is a very difficult problem; working with R. Williams
  - Instead we work with bounds for case of $K = 2$

Buffer Overflow Decay Exponent

- for $d > d_0$, we derive the buffer overflow form the point-point analysis
- for $d < d_0$, a large deviation analysis of queueing model (M) is required
  - Is a very difficult problem; working with R. Williams
  - Instead we work with bounds for case of $K = 2$



$$e^{-\frac{1}{2}\hat{\sigma}_L(d)DC_{tot}(d)} \quad \leq \quad P_q(d) \quad \leq \quad e^{-\frac{1}{2}\sigma_s(d)DC_{tot}(d)}$$

where

$\sigma_s(d) := \sup\{\theta : \theta > 0, \Lambda(\theta) - \theta C_{tot}(d) = 0\}$ and $\hat{\sigma}_L(d) := \min\{2f(C_{av}(d)), f(2C_{tot}(d))\}$.

Minimizing Asymptotic Total Loss Probability

- The asymptotic approximation of the total loss probability $P_t(d) = P_q(d) + P_e(d)$ is
  - For $d_0 \leq d \leq MN$,
  $$P_t(d) \doteq e^{-\frac{1}{2}\sigma_s(d)DC_{tot}(d)} + e^{-d \log SNR}$$
  - For $0 < d < d_0$,
  $$e^{-\frac{1}{2}\hat{\sigma}_L(d)DC_{tot}(d)} + e^{-d \log SNR} \;\dot{\leq}\; P_t(d) \;\dot{\leq}\; e^{-\sigma_s(d)DC_{tot}(d)} + e^{-d \log SNR}$$
- The minimum of $P_t(d)$ or its bounds happens when the exponents of the two terms are within $o(1)$ of each other
  - Doing so, we arrive at an order optimal d or an interval where it lies.

Optimal Operating Interval in MIMO MAC

Algorithm 1:

1. Solve for d which is a solution of

$$\sigma_s(d) DTr_{tot}(d) = 2d$$

   If $d \geq d_0$, then $d^* = d_u^* = d_l^* = d$ and stop. Otherwise, set the lower bound $d_l^* = d$. Go to Step 2.

2. Solve for d which is a solution of

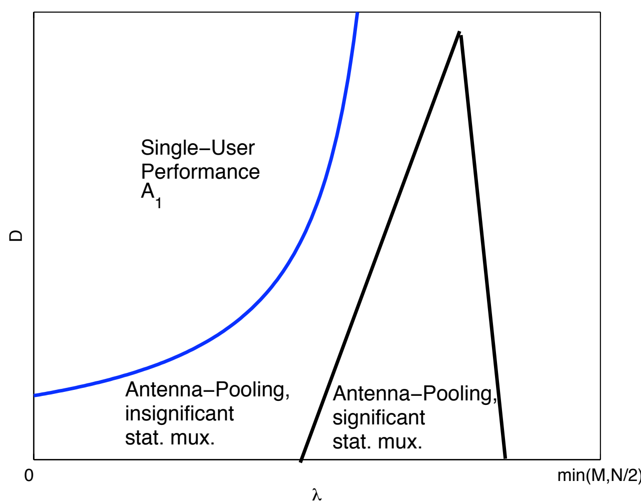
$$\hat{\sigma}_L(d) DTr_{tot}(d) = 2d$$

   and set $d_u^* = \min(d, d_0)$.

Theorem: Algorithm 1 results in a closed interval $[d_l^*, d_u^*]$ in which the optimal common diversity gain d^* lies.

Resource Pooling, Scheduling, and Statistical Multiplexing

- Users are forced to sacrifice diversity for multiplexing gain when
  - Traffic intensity increases
  - The delay tolerance D decreases
  - Similar to P-P case
- Unlike P-P case, though, here burstiness of traffic at low to moderate traffic loads results in "higher" perceived diversity gain



- Algorithm 1 results in quantifying the statistical multiplexing gain
  - $d_u^* - d_l^*$ is an upper bound on benefit of dynamic scheduling

A Simple Example: Source Model I

- Recall the number of bits arrived in timeslot t is

$$X_t = \sum_{i=1}^{N} Y_i$$

  - N is Poisson random variable with rate ν
  - $Y_1, Y_2, \ldots$ are iid exponential random variables with mean $1/\mu$
  - The average bit arrival rate ν/μ scaled with $\log SNR$

Proposition: For the compound Poisson source with exponential packet length, the $\sigma_s(d)$ and $\hat{\sigma}_L(d)$ are given as

$$\sigma_s(d) = \mu(1 - \frac{\lambda}{r_{av}(d)}),$$

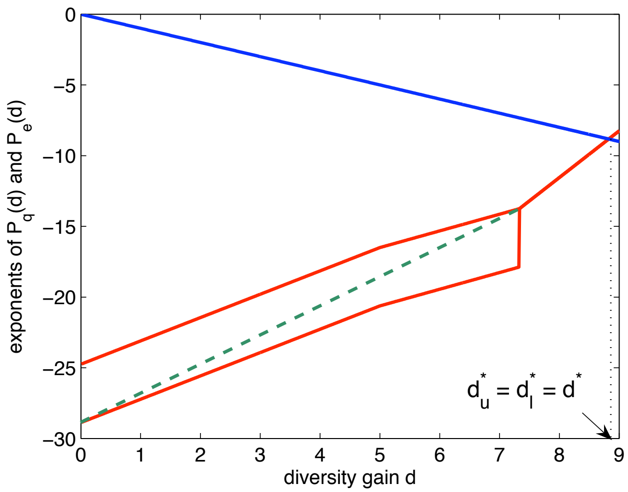
and

$$\begin{aligned} \hat{\sigma}_L(d) &= \min\left\{2\mu(1 - \frac{\lambda}{r_{av}(d)}), \mu(1 - \frac{\lambda}{2r_{av}(d)})\right\} \\ &= \begin{cases} \mu(1 - \frac{\lambda}{2r_{av}(d)}) & \text{if } 0 < \frac{\lambda}{r_{av}(d)} \leq 2/3, \\ 2\mu(1 - \frac{\lambda}{r_{av}(d)}) & \text{if } 2/3 \leq \frac{\lambda}{r_{av}(d)} < 1. \end{cases} \end{aligned}$$

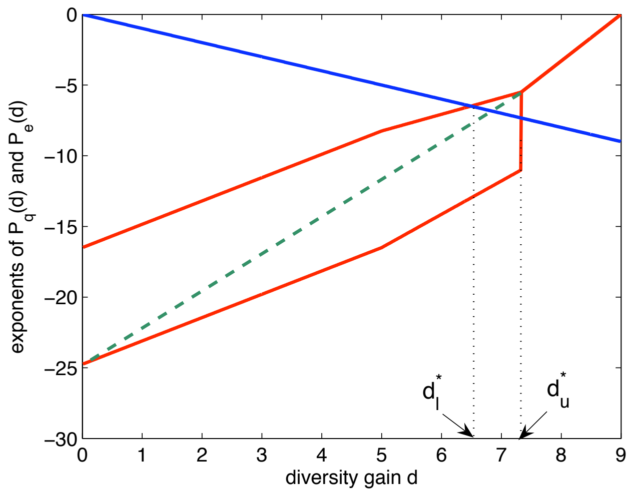
A Simple Example: Source Model I

- Assume $M = N = 4$
- The average packet size $1/\mu$ is 100 nats
- Block space-time coding over $T = 2M + N - 1 = 11$ symbols

Plots of the exponents of $P_q(d)$ and $P_e(d)$ for two different arrival rates ($\lambda = 0.5$ and 1) and delay bound $D = 150$



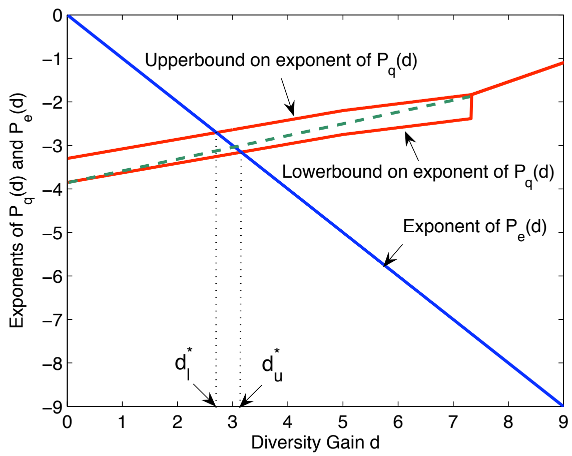
(a) $\lambda = 0.5, D = 150$



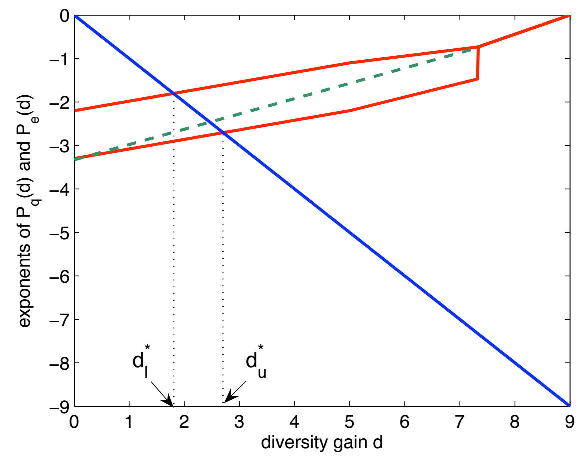
(b) $\lambda = 1, D = 150$

A Simple Example: Source Model I

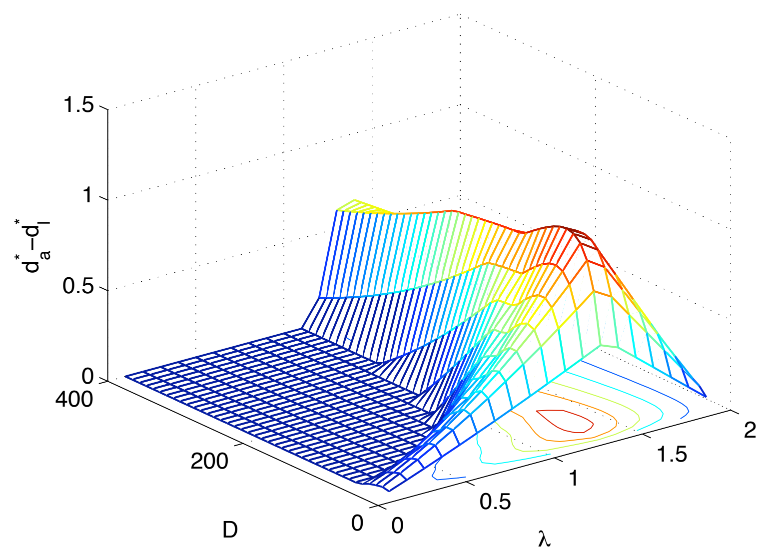
Plots of the exponents of $P_q(d)$ and $P_e(d)$ for two different arrival rates ($\lambda = 0.5$ and 1) and delay bound $D = 20$



(c) $\lambda = 0.5, D = 20$



(d) $\lambda = 1, D = 20$



Summary and Future Work

- In this paper, we
  - Considered bursty delay-sensitive source over MIMO channel.
  - Found optimal operating point $(r^*, d^*(r^*))$ on DMT curve, minimizing total bit loss probability P_t.
  - Compared r^* for three types of sources, illustrating relation between source burstiness, delay bound, and r^*.
- Future Work
  - Extend the study to account for time diversity w/ and w/o ARQ

Future Work: Time diversity

- Coding across multiple coherence blocks
  - Longer blocks allow for higher degree of diversity but cause larger transmission delay

- Time-diversity with ARQ:
  - Diversity-multiplexing-time diversity tradeoff [El Gamal, et. al. '04]
  - Dynamic operating points on MIMO-ARQ channel [Holliday, Goldsmith, and Poor '06]