Rate Splitting for MIMO Wireless Networks: A Promising PHY-Layer Strategy for 5G and Beyond

Bruno Clerckx

Dept. of Electrical and Electronic Engineering
Imperial College London

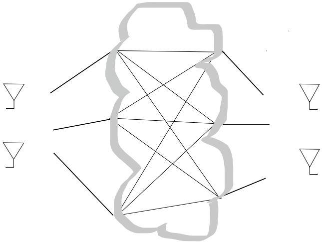
IEEE International Symposium on Wireless Communication Systems, August 2018

Contents

Introduction to MIMO Networks

Point-to-Point MIMO

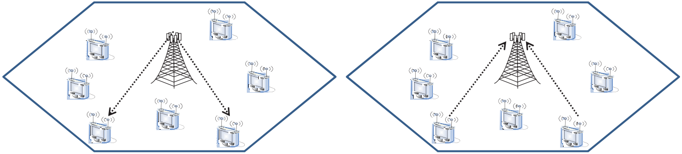
- MIMO channel with M transmit and N receive antennas



- MISO ($M > 1$,$N = 1$), SIMO ($M = 1$,$N > 1$), SISO ($M = 1$,$N = 1$)
- Multiplexing gain
  - Also called Degrees of Freedom - DoF
  - Slope of the achievable rate vs SNR (at high SNR)
  - Number of interference-free streams transmitted in parallel
  - $R \approx g_s \log_2(\rho)$ with $g_s \leq \min\{M, N\}$
  - Spatial Multiplexing/BLAST (no CSIT)
  - Multiple eigenmode transmission with waterfilling power allocation (CSIT)
  - DoF resilient to CSIT inaccuracy in point-to-point

Multi-User MIMO

- Most systems are multi-user!
- How to deal with K users? Benefit of MIMO in a multi-user setting?
- MIMO Broadcast Channel (BC) and Multiple Access Channel(MAC)



(a) Broadcast Channel - Downlink **(b) Multiple Access Channel - Uplink**

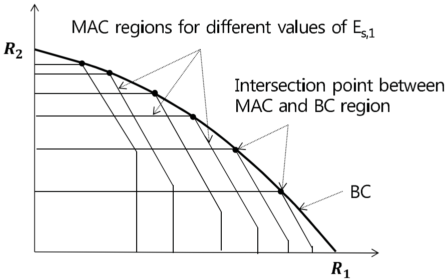
Differences between BC and MAC:

- Multiple independent additive noises in BC **vs** a single noise term in MAC.
- A single Tx power constraint in BC **vs** multiple Tx power constraints in MAC.
- The desired signal and the interference propagate through the same channel in the BC **vs** they propagate through different channels in the MAC.

We focus on Downlink settings.

Downlink Multi-User MIMO (BC)

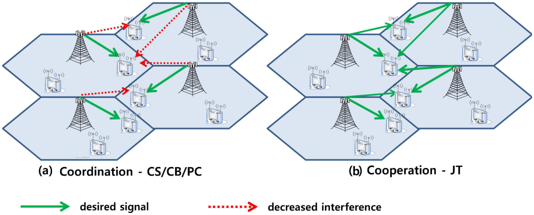
Transmitter sends independent streams to multiple receivers.



- SISO:
  - Users can be ordered according to strength.
  - Superposition coding and SIC achieve capacity region.
  - DPC can be used (transmitter side interference cancellation).
- MISO and MIMO:
  - Users cannot be ordered.
  - SC-SIC leads to performance loss.
  - DPC necessary to achieve capacity region (in general).

Multi-Cell MIMO: Coordination and Cooperation

- Jointly allocate resources across the whole network (and not for each cell independently) and use the antennas of multiple cells to improve the received signal quality at the mobile terminal and to reduce the co-channel interferences.
- Two categories: Coordination and Cooperation



- Coordination
  - No data sharing (user data is available at a single transmsitter)
  - CSI sharing
  - Modelled by an Interference Channel
- Cooperation
  - Data sharing (user data is available at multiple transmsitters)
  - CSI sharing
  - Modelled by BC (for Downlink)

Massive MIMO

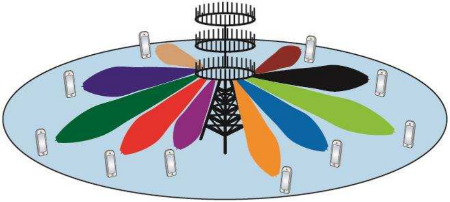


Figure: Downlink beamforming is centralized Massive MIMO deployment [1].

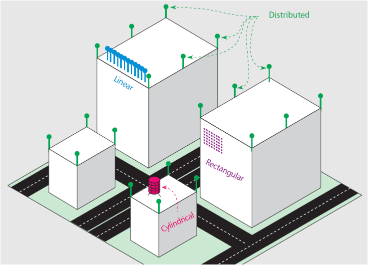


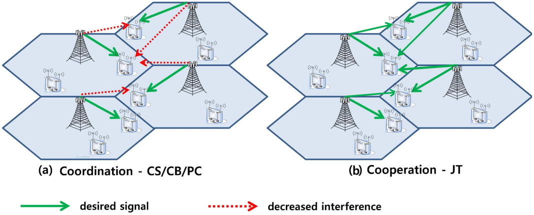
Figure: Various centralized and distributed massive MIMO deployments [2].

- Number of transmitting antennas at the transmitter is (massively) increased.
- Energy can be focused in very narrow beams (reduce multi-user interference).
- Simple precoder design based on matched beamforming (MRC).
- Simultaneously serve many users in the same resource block, simplified scheduling.
- With a massive number of antennas, comes a massive demand for CSIT.

Limitations of Current 4G and Emerging 5G Architecture

MIMO Networks: a central problem...the role of CSIT

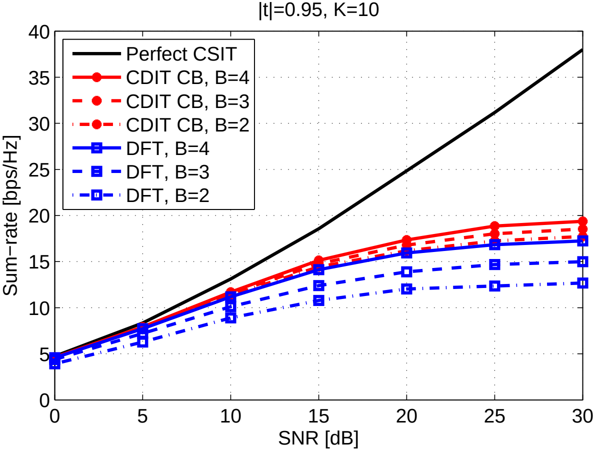
- MIMO Networks exploit more and more channel state information at the transmitter (CSIT)



- Performance crucially rely on accurate CSIT
- CSIT impairments - plenty of sources:
  - Quantization errors.
  - Estimation errors.
  - Delays.
  - Channel acquisition at RB/Subband level.
  - Hardware impairments (phase noise, additive/multiplicative RF impairments, calibration of RF chains).

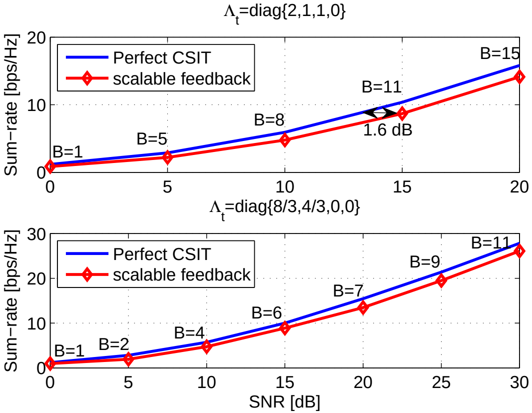
MU-MISO with Quantized Feedback - The Ceiling Effect

- MU-MISO with linear precoding and quantized feedback: the sum-rate saturates due to multi-user interference
- MU-MISO $=$ Full cooperation $=$ Upper-bound for any deployment



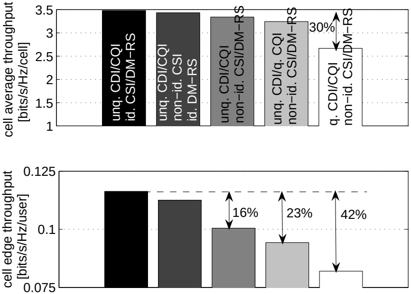
MU-MISO with Quantized Feedback - The Scaling Law

- Number of feedback bits necessary to maintain a rate loss of $\Delta\bar{R} \leq \log_2(\delta)$ bps/Hz per user
  - i.i.d. Rayleigh fading channels: $B \approx (n_t - 1)\log_2(P)$ [3].
  - spatially correlated Rayleigh fading channels $B \approx (r-1)\log_2(P)$ (r the rank of the transmit correlation matrix)[4].

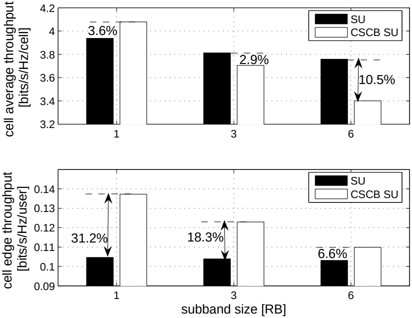


LTE-A SLS Observations

MU-MIMO



Coordinated Scheduling and Beamforming



Observations:

- Big loss due to imperfect CSIT.
- High CSIT accuracy is getting increasingly difficult to satisfy due to increasing number of antennas and access points in 5G (dense HetNet, Massive MIMO).

Motivation 1 for a New Physical Layer

- MU-MISO with linear precoding and quantized feedback: the sum-rate saturates due to multi-user interference.
- Big loss as the CSIT accuracy decreases.
- High CSIT accuracy has become increasingly difficult to satisfy due to increasing number of antennas and access points in 5G (dense HetNet, Massive MIMO).
- So far, techniques designed for perfect CSIT applied to imperfect CSIT scenarios.
- Imperfect CSIT hardly avoidable.
- Wiser to design wireless networks from scratch accounting for imperfect CSIT and its resulting multi-user interference?

Motivation 1 for a New Physical Layer

Information theoretic channel (e.g. MISO BC)

$\Downarrow$

Information theoretic limits (Capacity region)

$\Downarrow$

Communication scheme (e.g. DPC)

$\Downarrow$

Suboptimal scheme (Linear precoding)

$\Downarrow$

Signal processing (Precoder optimization)

$\Downarrow$

Imperfect CSIT (Robust optimization)

For example, robust optimization of $\mathbf{p}_1, \ldots, \mathbf{p}_K$ in

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{p}_k s_k.$$

BUT !!! The design is motivated by perfect CSIT to start with.

A Bottom-up Approach

Information theoretic channel (e.g. MISO BC with Imperfect CSIT)
$\Downarrow$
Information theoretic limits (Capacity region - unkown)
$\Downarrow$
Alternative information theoretic limits (DoF region)
$\Downarrow$
Communication scheme (Based on Rate-Splitting)
$\Downarrow$
Suboptimal scheme (Linear precoding)
$\Downarrow$
Signal processing (Precoder optimization)

For example, optimizing $\mathbf{p}_{\mathrm{c}}, \mathbf{p}_1, \ldots, \mathbf{p}_K$ in

$$\mathbf{x} = \mathbf{p}_{\mathrm{c}} s_{\mathrm{c}} + \sum_{k=1}^{K} \mathbf{p}_k s_k$$

where $\mathbf{p}_{\mathrm{c}} s_{\mathrm{c}}$ comes from Rate-Splitting.
Motivated by optimality in a DoF sense (multiplexing gain)

Motivation 2 for a New Physical Layer

- MIMO networks rely on two extreme interference management strategies: **fully decode interference** and **treat interference as noise**
  - NOMA based on superposition coding with successive interference cancellation relies on strong users to fully decode and cancel interference created by weaker users
  - SDMA (MU-MIMO, CoMP, Massive MIMO, millimetre wave MIMO based on linear precoding) rely on fully treating any multi-user interference as noise
- Rate-Splitting as a **more general** and **more powerful** transmission framework: **partially** decode interference and **partially** treat interference as noise
  - Softly bridge and therefore reconcile the two extreme strategies
  - RS encompasses NOMA and SDMA as special cases

$$\mathbf{x} = \mathbf{p}_{\mathrm{c}} s_{\mathrm{c}} + \sum_{k=1}^{2} \mathbf{p}_k s_k$$

where $\mathbf{p}_{\mathrm{c}} s_{\mathrm{c}}$ comes from Rate-Splitting.

Flavour of Rate-Splitting: A Two-User Case

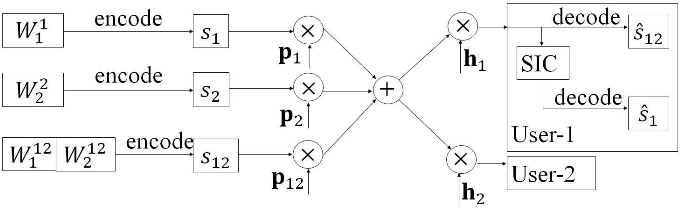
Transmitter

- W_1, W_2 split into $\{W_1^{12}, W_1^1\}$ for user-1 and $\{W_2^{12}, W_2^2\}$ for user-2
- W_1^{12}, W_2^{12} are encoded together into a common stream s_{12}
- W_1^1 and W_2^2 encoded into private stream s_1 for user-1 and s_2 for user-2
- Data streams are linear precoded $\mathbf{x} = \mathbf{p}_{12}s_{12} + \mathbf{p}_1 s_1 + \mathbf{p}_2 s_2$

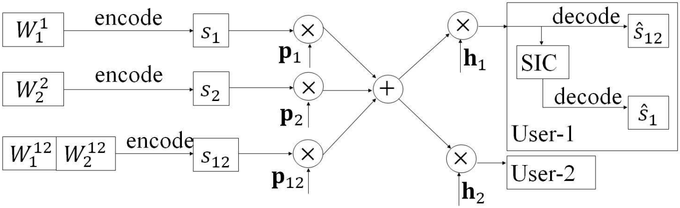
Receiver

- Both users first decode s_{12} by treating s_1 and s_2 as noise
- Both users perform SIC and retrieve s_1 and s_2, respectively

SIC (or joint decoding) **needed** to separate common and private streams



Flavour of Rate-Splitting: A Two-User Case



SDMA (classical multi-user linear precoding)

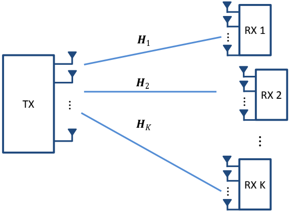
- Simply allocate no power to s_{12} and treat multi-user interference as noise
  $\mathbf{x} = \mathbf{p}_1 s_1 + \mathbf{p}_2 s_2$

SC-SIC (NOMA)

- Forcing user-1 to fully decode the message of user-2
- Allocate no power to s_2, encode W_1 into s_1 and encode W_2 into s_{12}
  $\mathbf{x} = \mathbf{p}_{12} s_{12} + \mathbf{p}_1 s_1$
- User-1 and user-2 decode s_{12} by treating s_1 as noise and user-1 decodes s_1 after canceling s_{12}

The MISO Broadcast Channel and Partial CSIT

System model



$$y_k(t) = \mathbf{h}_k^H(t)\mathbf{x}(t) + n_k(t)$$

- M transmit antennas and K single-antenna users ($M \geq K$).
- Channel state (matrix): $\mathbf{H}(t) = [\mathbf{h}_1(t), \ldots, \mathbf{h}_K(t)]$.
- In each t, transmitter obtains the estimate $\widehat{\mathbf{H}}(t)$ (i.e. CSIT).

System model: Transmission and Linear precoding

Linear precoding signal model:

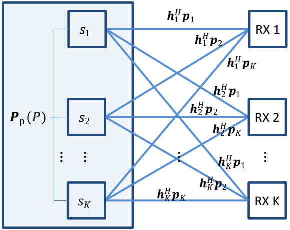
- Independent symbol streams: $W_1, \ldots, W_K \mapsto s_1, \ldots, s_K$.
- t is dropped for simplicity.
- Unity average power: $\mathbb{E}\{s_i s_k^*\} = 1$ if $i = k$, and 0 if $i \neq k$.
- Linear Precoding:

$$\mathbf{x} = \mathbf{p}_1 s_1 + \ldots + \mathbf{p}_K s_K.$$

- Average power constraint: $\sum_{k=1}^{K} \|\mathbf{p}_k\|^2 \leq P$.
- $\mathbf{P}_{\mathrm{p}} = [\mathbf{p}_1, \ldots, \mathbf{p}_K]$ can be adapted based on CSIT

$$\mathbf{P}_{\mathrm{p}}\big(\widehat{\mathbf{H}}(1)\big), \mathbf{P}_{\mathrm{p}}\big(\widehat{\mathbf{H}}(2)\big), \ldots, \mathbf{P}_{\mathrm{p}}\big(\widehat{\mathbf{H}}(T)\big).$$

System model: SINR and Rate

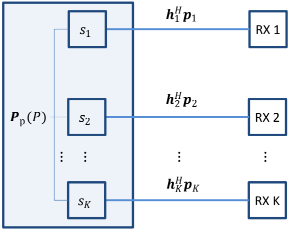


$$y_k = \overbrace{\mathbf{h}_k^H \mathbf{p}_k s_k}^{\text{desired signal}} + \overbrace{\mathbf{h}_k^H \sum_{i \neq k} \mathbf{p}_i s_i}^{\text{interference}} + \overbrace{n_k}^{\text{noise}}$$

- SINR (instantaneous): $\mathrm{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{p}_k|^2}{\sum_{i \neq k} |\mathbf{h}_k^H \mathbf{p}_i|^2 + \sigma_{\mathrm{n}}^2}$.
- Rate (instantaneous): $R_k = \log_2(1 + \mathrm{SINR}_k)$.
- Ergodic Rate (for $T \gg 1$): $\mathbb{E}\{R_k\}$.

Perfect CSIT

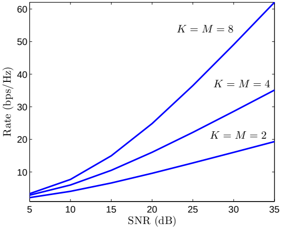
- Perfect CSIT: $\widehat{\mathbf{H}} = \mathbf{H}$.
- Zero-Forcing (ZF) precoding:
  - $\mathbf{P}_\mathrm{p} = \mathbf{H}(\mathbf{H}^H\mathbf{H})^{-1}\mathbf{B}$ where $\mathbf{B}$ is diagonal.
  - This yields: $\mathbf{p}_k \in \mathrm{null}\left([\mathbf{h}_1, \ldots, \mathbf{h}_{k-1}, \mathbf{h}_{k+1}, \ldots, \mathbf{h}_K]^H\right)$.



$$y_k = \mathbf{h}_k^H \mathbf{p}_k s_k + n_k$$

- Each user receives an interference-free stream.
- In other words, each user gets one full DoF.

Perfect CSIT:Degrees of Freedom (DoF)



- DoF: fraction of an interference-free stream's capacity as $P \to \infty$.
- Considering the Ergodic rate:

$$d_k = \lim_{P \to \infty} \frac{\mathbb{E}\{R_k\}}{\log_2(P)}.$$

- For MISO, we have $d_k \leq 1$ due to single-antenna receivers.
- Under perfect CSIT, ZF and equal power allocation achieves full DoF:

$$\sum_{k=1}^{K} d_k = K.$$

Imperfect CSIT

What happens when CSIT is imperfect?

Imperfect CSIT model:

$$\mathbf{H} = \widehat{\mathbf{H}} + \widetilde{\mathbf{H}}$$

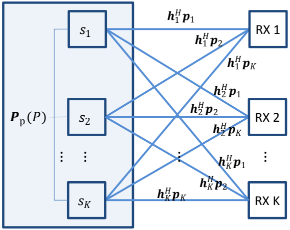
$$\mathbf{h}_k = \underbrace{\widehat{\mathbf{h}}_k}_{\text{estimate}} + \underbrace{\widetilde{\mathbf{h}}_k}_{\text{error}}$$

Estimate obtained through feedback or UL training [5].

- CSIT error power: $\mathbb{E}\left\{\|\widetilde{\mathbf{h}}_k\|^2\right\} = \sigma_{\mathrm{e},k}^2$.
- CSIT error scaling: $\alpha_k = \lim_{P\to\infty} -\frac{\log(\sigma_{\mathrm{e},k}^2)}{\log(P)}$
- It follows that: $\mathbb{E}\left\{\|\widetilde{\mathbf{h}}_k\|^2\right\} \sim P^{-\alpha_k}$.
- Assume: $\alpha_1, \ldots, \alpha_K = \alpha$.
  - $\alpha > 0$: CSIT improves with P (e.g. increasing number of feedback bit).
  - $\alpha = 0$: CSIT fixed with P (e.g. fixed number of feedback bit).
  - $\alpha = 1$: CSIT perfect in a DoF sense (as we see next).

Imperfect CSIT: Zero-Forcing

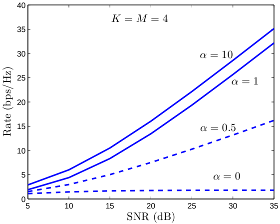
- ZF over the imperfect channel estimate:
  - $\mathbf{P}_{\mathrm{p}} = \widehat{\mathbf{H}}(\widehat{\mathbf{H}}^H\widehat{\mathbf{H}})^{-1}\mathbf{B}$.
  - This yields: $\mathbf{p}_k \in \mathrm{null}\left(\left[\widehat{\mathbf{h}}_1, \ldots, \widehat{\mathbf{h}}_{k-1}, \widehat{\mathbf{h}}_{k+1}, \ldots, \widehat{\mathbf{h}}_K\right]^H\right)$.



$$y_k = \overbrace{\mathbf{h}_k^H \mathbf{p}_k s_k}^{\text{desired signal}} + \overbrace{\widetilde{\mathbf{h}}_k^H \sum_{i \neq k} \mathbf{p}_i s_i}^{\text{residual interference}} + \overbrace{n_k}^{\text{noise}}$$

- Each user cannot enjoy an interference-free stream anymore.
- What happens to the DoF?

Imperfect CSIT: DoF loss



- ZF and equal power allocation: $\|\mathbf{p}_1\|^2 = \ldots = \|\mathbf{p}_K\|^2 = \frac{P}{K}$.

$$y_k = \overbrace{\mathbf{h}_k^H \mathbf{p}_k s_k}^{\text{desired signal } \sim P} + \overbrace{\widetilde{\mathbf{h}}_k^H \sum_{i \neq k} \mathbf{p}_i s_i}^{\text{residual interference } \sim P^{1-\alpha}} + \overbrace{n_k}^{\text{noise} \sim P^0}$$

- Assume $\alpha \in [0, 1]$.
- $\mathrm{SINR}_k \sim P^\alpha$ from which $\mathbb{E}\{R_k\} = \log_2(P^\alpha) + O(1)$.
- $d_k = \alpha$ from which the sum DoF [3, 5]:

$$\sum_{k=1}^{K} d_k = K\alpha.$$

Imperfect CSIT: Interference

Perfect CSIT:

- Inter-user interference can be fully eliminated.
- Full DoF is achieved.

Partial CSIT with $\alpha \geq 1$:

- Inter-user interference can be reduced to the level of noise.
- No DoF loss.

Partial CSIT with $\alpha < 1$:

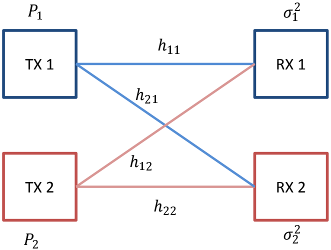
- Inter-user interference cannot be reduced to the level of noise.
- Treating interference as noise causes DoF loss.

If interference cannot be eliminated or reduced to noise level, why not decode it and remove it from the received signal (fully or in part)?

Let us first take a step back, and look at the 2-user Interference Channel (IC).

Fundamentals of Rate Splitting

Two-User Interference Channel (IC)



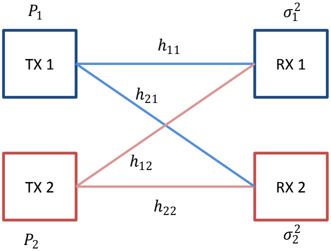
$$y_k = h_{k1}x_1 + h_{k2}x_2 + n_k$$

- Message W_k from TX-k to RX-k.
- Encoding: $W_k \mapsto x_k$.
- Decoding: $y_k \mapsto \widehat{W}_k$.

Symmetric setup:

- $|h_{11}|^2 = |h_{22}|^2 = |h_\mathrm{d}|^2$ and $|h_{12}|^2 = |h_{21}|^2 = |h_\mathrm{c}|^2$
- $P_1 = P_2 = P$ and $\sigma_1^2 = \sigma_2^2 = \sigma^2$

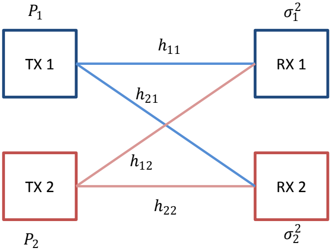
Two-User IC: Very weak interference



Very weak interference $|h_{\mathrm{c}}|^2 \ll |h_{\mathrm{d}}|^2$:

- Interference is so weak, it may be **treated as noise**.
- E.g. RX-1 decodes x_1 while treating x_2 as noise.
- $R_k \leq \log_2 \left(1 + \frac{P|h_{\mathrm{d}}|^2}{\sigma^2 + P|h_{\mathrm{c}}|^2}\right)$.

Two-User IC: Strong interference



Strong interference $|h_\mathrm{c}|^2 > |h_\mathrm{d}|^2$:

- Interfering signal is stronger than desired signal, may as well **decode** it.
- E.g. RX-1 decodes both x_2 and x_1 (MAC).
- $R_k \leq \log_2\left(1 + \frac{P|h_\mathrm{d}|^2}{\sigma^2}\right)$
- $R_1 + R_2 \leq \log_2\left(1 + \frac{P|h_\mathrm{d}|^2 + P|h_\mathrm{c}|^2}{\sigma^2}\right)$ (comes from cross-decoding)

Two-User IC: Rate-Splitting

Weak interference $|h_c|^2 < |h_d|^2$ **(or general case)**:

- Not strong enough to **decode**, or weak enough to **treat as noise**.
- **Rate-Splitting**: part **decoded** by other and part **treated as noise**.
  - Split messages: $W_k \mapsto W_{k0}, W_{k1} \mapsto x_{k0}, x_{k1}$.
  - Split power: $P_k \mapsto P_{k0}, P_{k1}$.
  - RX-1 decodes x_{20} and x_1 (composed of x_{10}, x_{11}).
  - RX-2 decodes x_{10} and x_2 (composed of x_{20}, x_{21}).
- Reduces to **treat as noise** when $P_{10} = P_{20} = 0$.
  - i.e. $|W_{10}| = |W_{20}| = 0$.
  - $W_k \mapsto x_{k1}$.
- Reduces to **decode** interference when $P_{11} = P_{21} = 0$.
  - i.e. $|W_{11}| = |W_{21}| = 0$.
  - $W_k \mapsto x_{k0}$.
- **Bridges** the two in general [6].

The MISO-BC with imperfect CSIT revisited

Rate-Splitting for MISO-BC[7]:

- The general idea is to split messages.
- One part decoded by all, while the other treated as noise.

But!

- In what proportion are messages split?
- How much power to allocate?
- How to transmit each part?

Strategy:

- Private messages:
  - Parts which are treated as noise.
  - Received at the level of noise
- Common message(s):
  - Parts which are decoded by all.
  - Transmitted in a public manner.

MISO-BC: Parts to treat as noise (private messages)

Interference reduction through power control:

- Reduce allocated power to P^{α}.
- Note that $P^{\alpha} \leq P$ for $\alpha \in [0, 1]$.
- Equal power allocation: $\|\mathbf{p}_1\|^2 = \ldots = \|\mathbf{p}_K\|^2 = \frac{P^{\alpha}}{K}$.

$$y_k = \overbrace{\mathbf{h}_k^H \mathbf{p}_k s_k}^{\text{desired signal } \sim P^{\alpha}} + \overbrace{\widetilde{\mathbf{h}}_k^H \sum_{i \neq k} \mathbf{p}_i s_i}^{\text{residual interference } \sim P^{\alpha-\alpha} = P^0} + \overbrace{n_k}^{\text{noise} \sim P^0}$$

- Interference is reduced to noise level $\sim P^0$.
- This also limits desired power $\sim P^{\alpha}$.
- DoF is maintained: $d_k = \alpha$ and $\sum_{k=1}^{K} d_k = K\alpha$.
- Only power levels (scalings) from 0 to α are occupied.
- The remaining power levels (α to 1) are freed for the other parts.

MISO-BC: Parts to decode (common message)

Superpose $W_c \mapsto s_c$ (with precoder $\mathbf{p}_c$) to be decoded by all users.

$$\mathbf{x} = \mathbf{p}_c s_c + \sum_{k=1}^{K} \mathbf{p}_k s_k$$

where $\|\mathbf{p}_c\|^2 = P - P^\alpha \sim P$ and $\|\mathbf{p}_1\|^2 = \ldots = \|\mathbf{p}_K\|^2 = \frac{P^\alpha}{K} \sim P^\alpha$

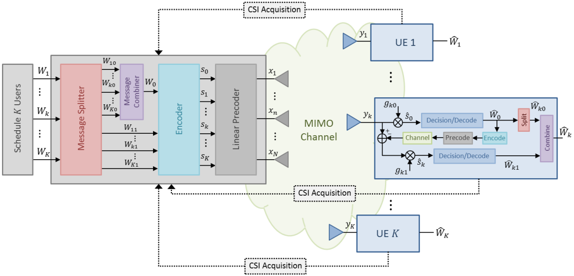
$$y_k = \overbrace{\mathbf{h}_k^H \mathbf{p}_c s_c}^{\sim P} + \overbrace{\mathbf{h}_k^H \mathbf{p}_k s_k}^{\sim P^\alpha} + \overbrace{\widetilde{\mathbf{h}}_k^H \sum_{i \neq k} \mathbf{p}_i s_i}^{\sim P^0} + \overbrace{n_k}^{\sim P^0}$$

- $\mathrm{SINR}_{c,k} \sim P^{1-\alpha}$ from which $\mathbb{E}\{R_{c,k}\} = \log_2(P^{1-\alpha}) + O(1)$.
- DoF of common message: $d_c = 1 - \alpha$.
- SIC is used to remove s_c, as it is decoded by all.
- DoF of private messages is maintained: $d_k = \alpha$.
- Sum DoF is boosted: $d_c + \sum_{k=1}^{K} d_k = (1-\alpha) + K\alpha$ [14].

What remains is to load both parts (private and common) with user data.

MISO-BC: Rate-Splitting

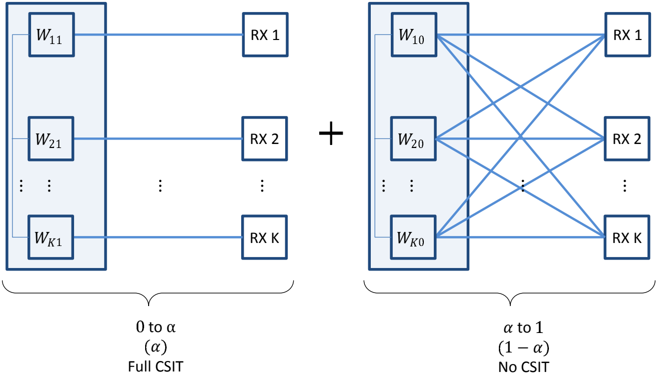
Instead of a new common message, s_c is loaded with part of user messages.



- Split message of user-1 : $W_1 \mapsto W_{10}, W_{11}$.
- Common part: $W_{10} \mapsto s_c$, decoded by all users but intended to users-1.
- Private part: $W_{11} \mapsto s_1$ decoded by user-1.
- $W_2, \ldots, W_K \mapsto s_2, \ldots, s_K$ decoded by corresponding users.

Splitting can be done for other (or all) users as in figure [28].

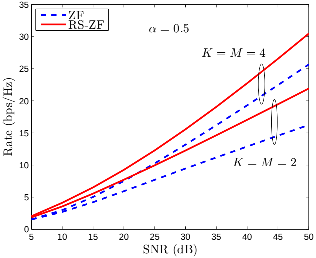
MISO-BC: Weighted sum interpretation



Decomposed into a weighted superposition of two networks [22]

- Perfect CSIT.
  - Achieves sum DoF of K.
  - Weighted by α.
- No CSIT
  - Achieves sum DoF of 1.
  - Weighted by $1-\alpha$.

MISO-BC: DoF with RS



Proposition

In the K user MISO-BC with partial CSIT, sum DoF achieved by ZF is given by

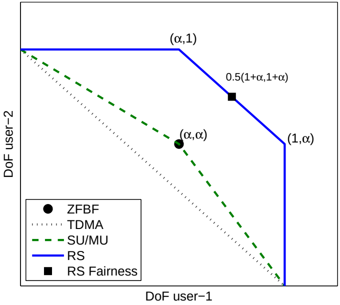
$$d_{\Sigma}^{\mathrm{ZF}} = K\alpha$$

while the sum DoF achieved by RS-ZF is given by

$$d_{\Sigma}^{\mathrm{RS}} = 1 + (K-1)\alpha.$$

Optimality of RS to achieve the entire DoF region of the K-user MISO BC shown in [31]. Converse based on [8].

MISO-BC: Two-User DoF region

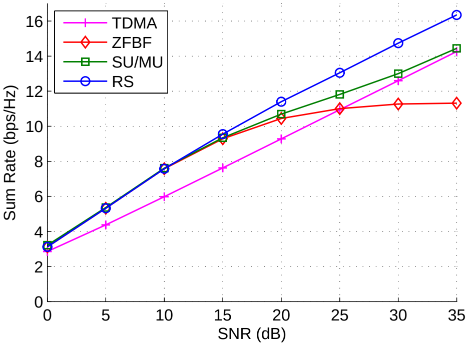


- Assume splitting for user-1
  - user-1 DoF: $d_c + d_1 = (1-\alpha) + \alpha = 1$.
  - user-2 DoF: $d_2 = \alpha$.
- Time-sharing between splitting for user-1 and user-2.
- Compared to time-sharing between ZF and TDMA.

Sum-Rate enhancement and Feedback reduction

From DoF to rate analysis [9]:

- So far we have looked at the DoF gains of RS $(P \to \infty)$.
- Sum-rate enhancement (slope gain and/or SNR gain) over ZF, TDMA, switching between TDMA/ZF (SU/MU)[9].
- $M = 4$ antennas, $K = 2$ users, and $B = 15$ bits.



Precoder Optimization

Precoder Optimization

Recall that the RS (linearly precoded) signal model is:

$$\mathbf{x} = \mathbf{p}_c s_c + \sum_{k=1}^{K} \mathbf{p}_k s_k$$

- Precoding matrix: $\mathbf{P} = [\mathbf{p}_c, \mathbf{p}_1, \ldots, \mathbf{p}_K]$.
- Power constraint: $\mathrm{tr}\left(\mathbf{P}\mathbf{P}^H\right) \leq P$.
- So far we considered simple barely optimized designs (ZF, random).
- The choice of $\mathbf{P}$ influences $R_c, R_1, \ldots, R_K$.

Challenges

- Transmitter only knows $\widehat{\mathbf{H}}$ and not $\mathbf{H}$.
- Instantaneous $R_c, R_1, \ldots, R_K$ not known by the transmitter.
- Transmission should be carried out at reliable (decodable) rates.

Ergodic Sum-Rate Maximization

RS problem [14]: design precoder for given $\widehat{\mathrm{H}}$ to maximize ASR

$$\mathcal{R}_{\mathrm{RS}}(P): \begin{cases} \max\limits_{\bar{R}_c, \mathbf{P}} & \bar{R}_{\mathrm{c}} + \sum_{k=1}^{K} \bar{R}_k \\ \text{s.t.} & \bar{R}_{\mathrm{c},k} \geq \bar{R}_{\mathrm{c}}, \ \forall k \in \mathcal{K} \\ & \mathrm{tr}\left(\mathbf{P}\mathbf{P}^H\right) \leq P \end{cases}$$

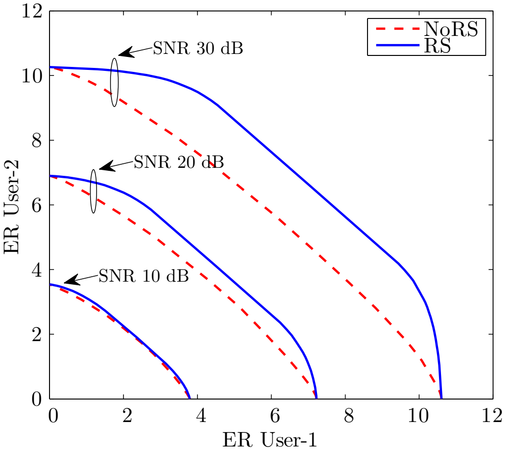
with $\bar{R}_{\mathrm{c},k} = \mathbb{E}_{\mathrm{H}|\widehat{\mathrm{H}}}\left\{R_{\mathrm{c},k} \mid \widehat{\mathbf{H}}\right\}$ and $\bar{R}_k = \mathbb{E}_{\mathrm{H}|\widehat{\mathrm{H}}}\left\{R_k \mid \widehat{\mathbf{H}}\right\}$,
as opposed to the conventional (NoRS) formulation

$$\mathcal{R}(P): \begin{cases} \max\limits_{\mathbf{P}_{\mathrm{p}}} & \sum_{k=1}^{K} \bar{R}_k \\ \text{s.t.} & \mathrm{tr}\left(\mathbf{P}_{\mathrm{p}}\mathbf{P}_{\mathrm{p}}^H\right) \leq P. \end{cases}$$

- Stochastic optimization problem (due to expectations inside the ARs).
- Even a deterministic version is non-convex and very difficult.
- WMMSE approach can efficiently handle sum rate problems.

Ergodic Sum-Rate Maximization: Two-user ER region

- More generally, we can solve the Weighted ESR problem [14].



- Shows the ER trade-offs between the two users.

Robust Max-Min Fairness

Non-Ergodic transmission over $T = 1$ random state $\{\mathbf{H}, \widehat{\mathbf{H}}\}$.

- For kth user, CSIT errors bounded by sphere with radius δ_k:

$$\mathbb{H}_k = \left\{ \mathbf{h}_k \mid \mathbf{h}_k = \widehat{\mathbf{h}}_k + \widetilde{\mathbf{h}}_k, \|\widetilde{\mathbf{h}}_k\| \leq \delta_k \right\}$$

- For any $\mathbf{P}$, worst-case rates defined as:

$$\bar{R}_{c,k} = \min_{\mathbf{h}_k \in \mathbb{H}_k} R_{c,k}(\mathbf{h}_k) \quad \text{and} \quad \bar{R}_k = \min_{\mathbf{h}_k \in \mathbb{H}_k} R_k(\mathbf{h}_k).$$

- For given $\widehat{\mathbf{H}}$, transmission at worst-case rates is reliable (robust).

Rate-Splitting revisited [15]: Sharing the common message

- $W_k \mapsto W_{k0}, W_{k1}$ for all $k \in \{1, \ldots, K\}$.
- $W_{10}, \ldots, W_{K0} \mapsto s_c$.
- $W_{11}, \ldots, W_{K1} \mapsto s_1, \ldots, s_K$.

Robust Max-Min Fairness

$$\mathcal{R}_{\mathrm{RS}}(P): \begin{cases} \max\limits_{\bar{\mathbf{c}},\mathbf{P}} & \min\limits_{k\in\mathcal{K}} \left(\bar{R}_k + \bar{C}_k\right) \\ \text{s.t.} & \bar{R}_{\mathrm{c},k} \geq \sum_{i=1}^{K} \bar{C}_i,\ \forall k \in \mathcal{K} \\ & \bar{C}_k \geq 0,\ \forall k \in \mathcal{K} \\ & \mathrm{tr}\left(\mathbf{P}\mathbf{P}^H\right) \leq P. \end{cases}$$

where $\bar{\mathbf{c}} = [\bar{C}_1, \ldots, \bar{C}_M]$.

- Portion of the common message rate given to user k: $\bar{C}_k$.
- Sum of all portions: $\sum_{k=1}^{K} \bar{C}_k = \bar{R}_{\mathrm{c}} = \min_i \bar{R}_{\mathrm{c},i}$.
- Rate of user k: $\bar{R}_k + \bar{C}_k$ (private and common portions).

Classical (NoRS) problem formulated as:

$$\mathcal{R}(P): \begin{cases} \max\limits_{\mathbf{P}_{\mathrm{p}}} & \min\limits_{k\in\mathcal{K}} \bar{R}_k \\ \text{s.t.} & \mathrm{tr}\left(\mathbf{P}_{\mathrm{p}}\mathbf{P}_{\mathrm{p}}^H\right) \leq P. \end{cases}$$

Robust Max-Min Fairness: Simulation results

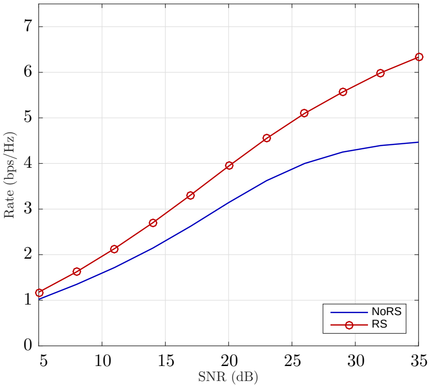


Figure: $K = M = 3$ and $\delta_1, \delta_2, \delta_3 = 0.1$.

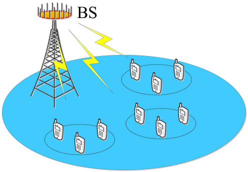
- NoRS saturates due to non-scaling CSIT errors.
- RS avoids saturation and performs better across all SNRs [15].

Massive MISO

Massive MIMO challenge: the huge demand for accurate CSIT.

The use of Rate-Splitting[13]:

- The constraint: $R_{\mathrm{c}} = \min_k \{R_{\mathrm{c},k}\}$.
- This highly reduces the gain when K is large.
- Channel statistics $\mathbf{R}_k$ can be further exploited.
- Large training and feedback overhead.



User grouping based on spatial correlation:

- Two-tier precoding [18, 19, 20]

$$\mathbf{x} = \sqrt{\frac{P}{K}} \sum_{g=1}^{G} \mathbf{B}_g \mathbf{W}_g \mathbf{s}_g,$$

- Users in g-th group share the same channel statistics: $\mathbf{R}_g$.
- $\mathbf{B}_g$: outer-precoding matrix based on channel statistics.
- $\mathbf{W}_g$: inner-precoding matrix designed based on short-term effective channel estimates: $\widehat{\widehat{\mathbf{H}}}_g = \mathbf{B}_g^H \widehat{\mathbf{H}}_g$.

Massive MISO: Hierarchical Rate-Splitting (HRS)

- Overlap between the eigen-subspaces $\Rightarrow$ inter-group interference.
- Imperfect CSIT $\Rightarrow$ intra-group interference.

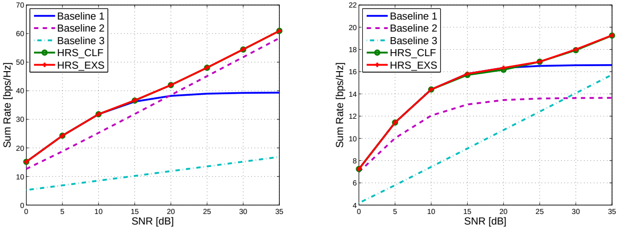
- **Hierarchical Rate-Splitting**[13]: a hierarchy of common messages to combat the inter-group and intra-group interference in Massive MIMO

$$\mathbf{x} = \overbrace{\sqrt{P_{\mathrm{sc}}}\mathbf{w}_{\mathrm{sc}}\, s_{\mathrm{sc}}}^{\text{system common msg.}} + \sum_{g=1}^{G} \mathbf{B}_g \left(\overbrace{\sqrt{P_{cg}}\mathbf{w}_{cg}\, s_{cg}}^{\text{group common msg.}} + \overbrace{\sqrt{P_{gk}}\mathbf{W}_g\, \mathbf{s}_g}^{\text{private msgs.}} \right)$$

- System common msg. decoded by all users: for inter-group interference.
- Group common msg. decoded by group: for intra-group interference

Massive MISO: Simulation results

- HRS under imperfect CSIT, $M = 100$, $K = 12, \tau^2 = 0.4$

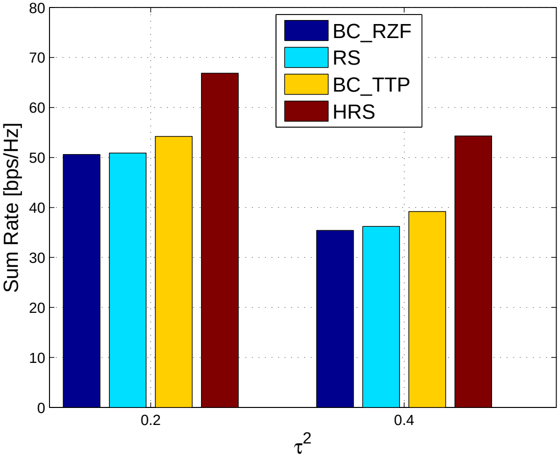


(a) disjoint eigen-subspace (b) overlapping eigen-subspace

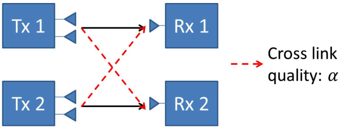
- HRS behaves as two-tier BC at low to medium SNR.
- HRS achieves a non-saturating sum rate.
- HRS decreases the complexity of precoder design and scheduling.
- HRS increases the complexity of the encoders and decoders.

Massive MISO: Simulation results

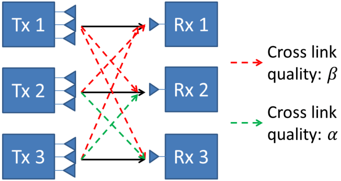
- $M = 100,\ K = 12, \tau^2 = 0.4,\ SNR = 30dB$, disjoint eigen-subspaces



Multi-Cell Coordination: Topological Rate-Splitting (TRS)



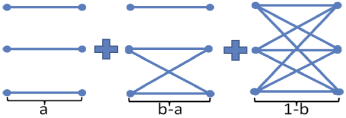
(c) two-cell scenario [21]



(d) three-cell scenario [22]

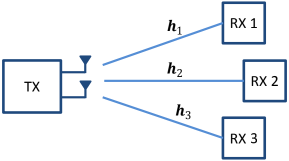
	Tx1	Tx2	Tx3
Rx1		b	b
Rx2	b		a
Rx3	b	a	

(e) CSIT pattern



(f) Weighted-sum interpretation [22]

Overloaded systems



- Overloaded scenarios: $K > M$.
- Scheduling over orthogonal resource blocks (time/frequency).
- Serve at most M users at a time.
- Reduces to conventional MISO BC in each block.
- With perfect CSIT, achieves DoF M in each block.

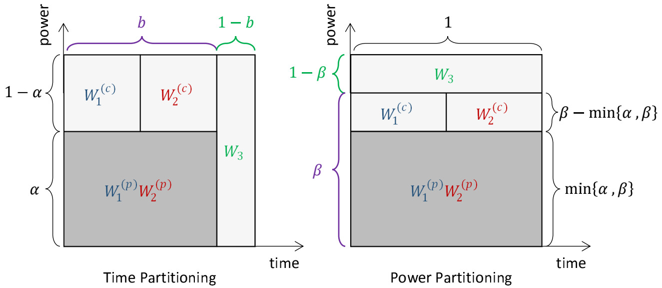
Consider a scenario where some user have little or no CSIT:

- IoT with many devices.
- Low-power sensor-like receivers.
- Can be served using the common message in the RS scheme [23].

Overloaded systems: Three-User example

- System: $M = 2$ antennas and $K = 3$ users.
- CSIT: $\alpha_1 = \alpha_2 = \alpha$ and $\alpha_3 = 0$.

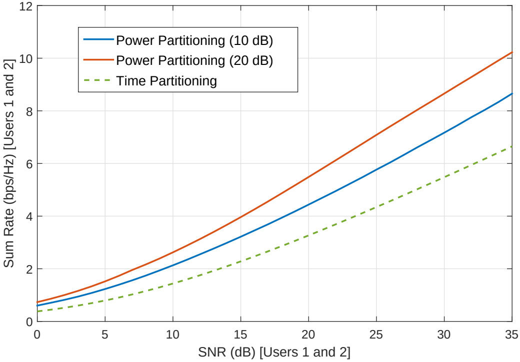
Scheduling approach



Power partitioning

- A **superposition of non-orthogonal layers and an orthogonal layer**
- Power partitioning achieves the optimum DoF region [23]

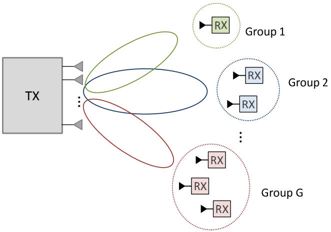
Overloaded systems: Numerical Results



- sum rate of RX-1 and RX-2 while maintaining the same rate for RX-3.
- Long-term SNR for RX-3 is 10 dB and 20 dB lower.
- Parameters: quality $\alpha = 0.5$, resource allocation $b = 0.5$.

Multigroup multicast beamforming

Users clustered into groups depending on content demand.



- K users grouped into $\mathcal{G}_1, \ldots, \mathcal{G}_G$.
- One message for each group: $W_1, \ldots, W_G$.
- Classical beamforming:

$$\mathbf{x} = \sum_{g=1}^{G} \mathbf{p}_g s_g.$$

Achieving max-min fairness (perfect CSIT):

$$\mathcal{R}(P): \begin{cases} \max\limits_{\mathbf{P}_\mathrm{p}} & \min\limits_{g \in \{1,\ldots,G\}} \min\limits_{i \in \mathcal{G}_g} R_i \\ \text{s.t.} & \sum\limits_{g=1}^{G} \left\|\mathbf{p}_g\right\|^2 \leq P. \end{cases}$$

- **Overloaded scenarios**: M is not enough for interference nulling [17].
- Rate saturation (even with perfect CSIT) due to inter-group interference.

Multigroup multicast beamforming: Simulation results

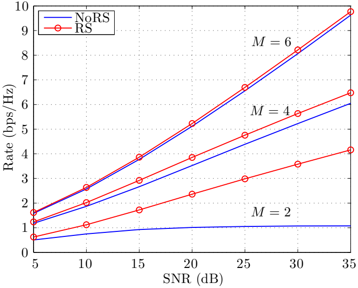


Figure: $K = 6$ users, $G = 3$ groups, $|\mathcal{G}_1| = 1$, $|\mathcal{G}_2| = 2$ and $|\mathcal{G}_3| = 3$.

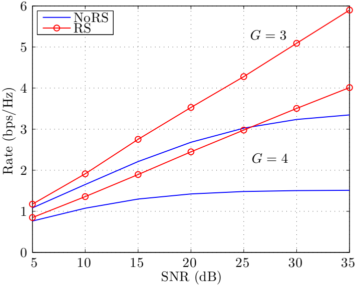
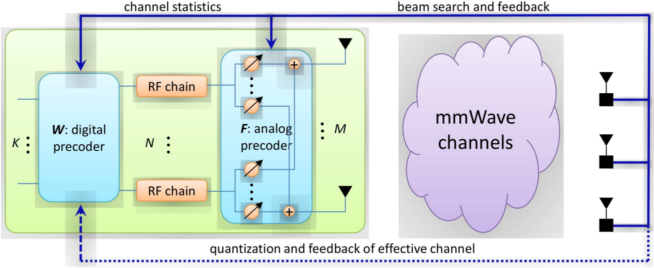


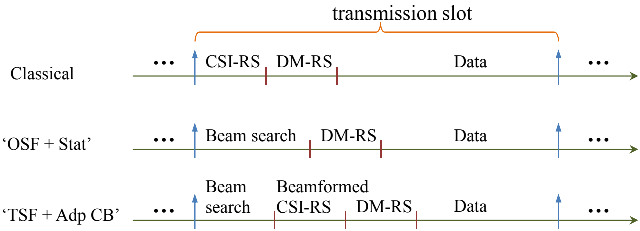
Figure: $M = 4$ antennas, $|\mathcal{G}_g| = 2$ users per group, $G = 3$ and 4.

RS to mitigate inter-group interference in overloaded scenarios [16, 17].

Multiuser Millimeter Wave Beamforming



Signalling and feedback procedure [29]



Multiuser Millimeter Wave Beamforming

RS to save second-stage channel training and feedback [29].

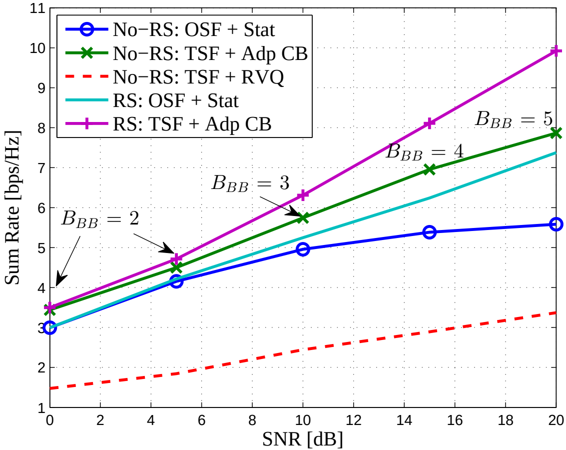
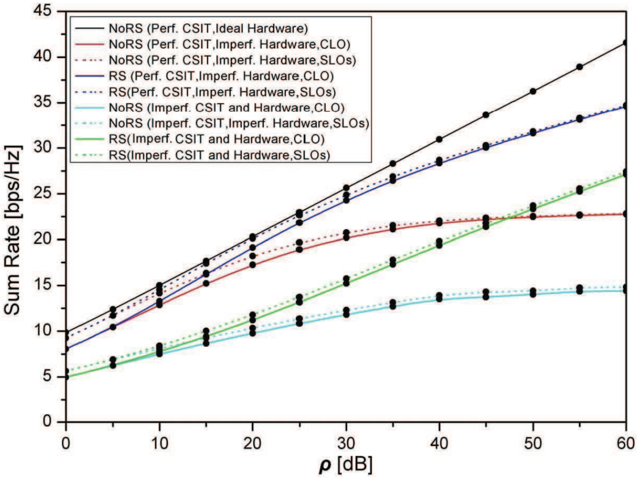


Figure: No-RS with extra feedback versus RS. $M = 32$, $K = 4$, $B_{RF} = 4$.

RF Impairments

RS to mitigate phase noise impairments [30].



RSMA: Generalizing SDMA and NOMA [34]

Space-Division Multiple Access (SDMA): multiplex users in **spatial domain** using **Multi-User Linear Precoding** (MULP)

- MU-MIMO, CoMP, network MIMO, mmw MIMO and Massive MIMO

Pros:

1. Reap all spatial multiplexing (DoF) benefits of a MISO BC with perfect CSIT
2. Low precoder and receiver complexity

Cons:

1. Suited to underloaded regime, not overloaded regime
2. Suited to semi-orthogonal users with similar channel strengths, not general settings
3. DoF optimal with perfect CSIT, not with imperfect CSIT

RSMA: Generalizing SDMA and NOMA

Non-Orthogonal Multiple Access (NOMA): multiplex users in **power (and spatial) domain** using (linearly precoded) **Superposition Coding** with **SIC** (SC-SIC)

Pros:

1. Cope with an overloaded regime with diversity of user channel strengths

Cons:

1. SC-SIC motivated by a SISO BC. DoF loss in MISO BC.
2. Suited to aligned users with diverse channel strengths, not general settings
3. Complexity at both the transmitter (opt. of precoders, groups and decoding orders) and the receivers (multi-layer SIC)
4. DoF loss in imperfect CSIT

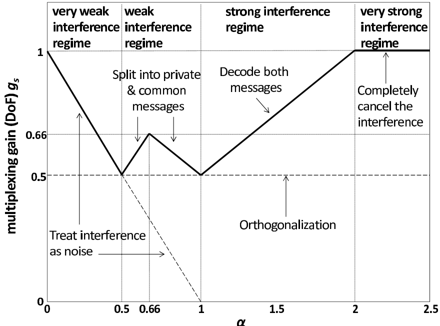
RSMA: Generalizing SDMA and NOMA

Two extreme interference management strategies: fully treat interference as noise and fully decode interference

SDMA: **fully treat** any residual multi-user **interference as noise**

NOMA: some users **fully decode and cancel interference** created by other users

Analogy with the two-user Gaussian SISO IC



Rate-Splitting Multiple Access (RSMA)

Multiplex users in **spatial and power domains** using **linearly precoded Rate-Splitting** (RS) with **SIC**

Decode **part** of the interference and treat the remaining **part** as noise

- Bridge the extremes
- General and powerful multiple access framework

Pros:

1. Encompass SDMA and NOMA as special cases
2. RSMA rate $\geq$ SDMA and NOMA rates
3. Optimal from a DoF perspective in both perfect and imperfect CSIT
4. Cope with any user deployments (diversity of channel strengths and directions), CSIT inaccuracy and network load
5. Lower computational complexity than NOMA for both the transmit scheduler and the receivers

Cons:

1. Higher encoding complexity than SDMA and NOMA

RSMA: Two-User Example

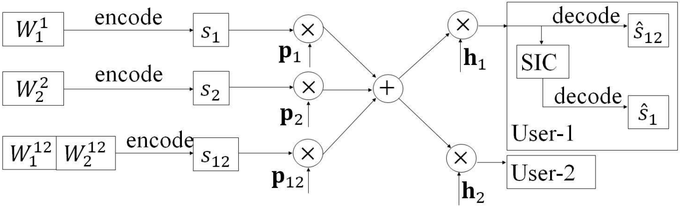
Transmitter

- W_1, W_2 split into $\{W_1^{12}, W_1^1\}$ for user-1 and $\{W_2^{12}, W_2^2\}$ for user-2
- W_1^{12}, W_2^{12} are encoded together into a common stream s_{12}
- W_1^1 and W_2^2 encoded into private stream s_1 for user-1 and s_2 for user-2
- Data streams are linear precoded $\mathbf{x} = \mathbf{p}_{12}s_{12} + \mathbf{p}_1 s_1 + \mathbf{p}_2 s_2$

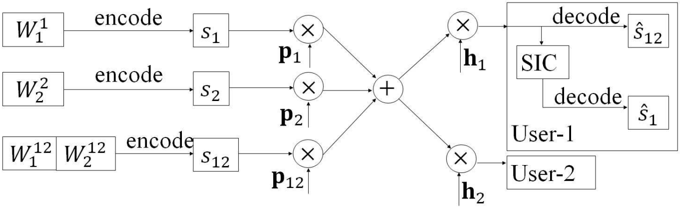
Receiver

- Both users first decode s_{12} by treating s_1 and s_2 as noise
- Both users perform SIC and retrieve s_1 and s_2, respectively

SIC (or joint decoding) **needed** to separate common and private streams



SDMA and NOMA: subsets of RSMA



SDMA based on MU-LP

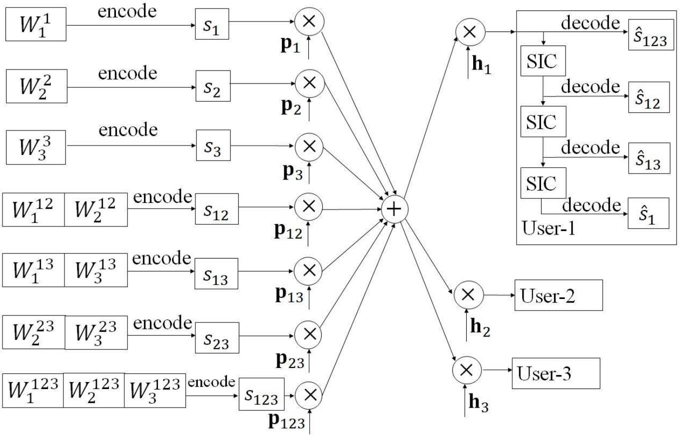
- Simply allocate no power to s_{12} and treat multi-user interference as noise
  $\mathbf{x} = \mathbf{p}_1 s_1 + \mathbf{p}_2 s_2$

NOMA based on SC-SIC

- Forcing user-1 to fully decode the message of user-2
- Allocate no power to s_2, encode W_1 into s_1 and encode W_2 into s_{12}
  $\mathbf{x} = \mathbf{p}_{12} s_{12} + \mathbf{p}_1 s_1$
- User-1 and user-2 decode s_{12} by treating s_1 as noise and user-1 decodes s_1 after canceling s_{12}
- NOMA more restrictive with $\mathbf{p}_{12} = \mathbf{p}_1$

RSMA: Three-User Example

General RS framework for 3-user



SDMA and NOMA again subsets of RSMA Framework extendable to K-user

Low-Complexity RSMA

Adjust the number of **SIC layers** and **common messages**

1-layer RS ($\mathcal{K} = \{1, \ldots, K\}$)

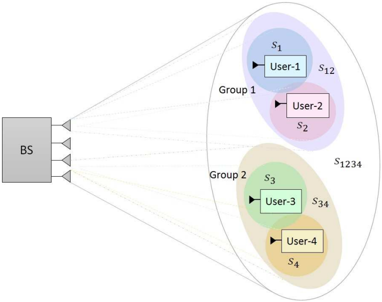
$$\mathbf{x} = \mathbf{P}\mathbf{s} = \mathbf{p}_{\mathcal{K}} s_{\mathcal{K}} + \sum_{k \in \mathcal{K}} \mathbf{p}_k s_k$$

- Only one SIC required at each receiver
- No user ordering/grouping at the transmitter
- MU-LP subset of 1-layer RS
- SC-SIC not a subset of 1-layer RS (for $K > 2$)

2-layer Hierarchical RS (HRS)

$$\mathbf{x} = \mathbf{P}\mathbf{s} = \mathbf{p}_{\mathcal{K}} s_{\mathcal{K}} + \sum_{g \in \mathcal{G}} \mathbf{p}_{\mathcal{K}_g} s_{\mathcal{K}_g} + \sum_{k \in \mathcal{K}} \mathbf{p}_k s_k$$

- Two layers of SIC required at each receiver
- User grouping but no user ordering at the transmitter
- MU-LP subset of 1-layer HRS
- SC-SIC not necessarily subset of 1-layer RS



Optimization

$$
\begin{aligned}
R_{\mathrm{RS}_3}(\mathbf{u},\pi) = \arg\max_{\mathbf{P},\mathbf{c}} \sum_{k=1}^{3} u_k R_{k,tot} \\
\text{s.t.} \quad & C_1^{123} + C_2^{123} + C_3^{123} \leq R_{123} \\
& C_1^{12} + C_2^{12} \leq R_{12} \\
& C_1^{13} + C_3^{13} \leq R_{13} \\
& C_2^{23} + C_3^{23} \leq R_{23} \\
& \mathrm{tr}(\mathbf{P}\mathbf{P}^H) \leq P_t \\
& R_{k,tot} \geq R_k^{th}, k \in \{1,2,3\} \\
& \mathbf{c} \geq \mathbf{0}
\end{aligned}
$$

SDMA vs NOMA vs RSMA

Multiple Access	NOMA		SDMA	RSMA
Strategy	**SC–SIC**	**SC–SIC per group**	**MU–LP**	**All forms of RS**
Design Principle	Fully decode interference	Fully decode interference in each group and treat interference between groups as noise	Fully treat interference as noise	Partially decode interference and partially treat interference as noise
Decoder architecture	SIC at receivers	SIC at receivers	Treat interference as noise	SIC at receivers
User Deployment Scenario	Users experience aligned channel directions and a large disparity in channel strengths.	Users in each group experience aligned channel directions and a large disparity in channel strengths. Users in different groups experience orthogonal channels.	Users channels are (semi-) orthogonal with similar channel strengths.	Any angle between channels and any disparity in channel strengths
Network load	More suited to overloaded network	More suited to overloaded network	More suited to underloaded network	Suited to any network load

Complexity SDMA vs NOMA vs RSMA

Multiple Access	NOMA		SDMA	RSMA	
Strategy	**SC–SIC**	**SC–SIC per group**	**MU–LP**	**RS**	**1-layer RS**
Encoder complexity	encode K streams	encode K streams	encode K streams	encode K private streams plus additional common streams	encode $K+1$ streams
Scheduler complexity	Very complex as it requires to find aligned users and decide upon suitable user ordering.	Very complex as it requires to divide users into orthogonal groups, with aligned users in each group and decide upon suitable user ordering in each group.	Complex as MU–LP requires to pair together semi-orthogonal users with similar channel gains.	Complex as it requires to decide upon suitable decoding order of the streams with the same stream order	Simpler user scheduling as RS copes with any user deployment scenario, does not rely on user grouping and user ordering.
Receiver complexity	Requires multiple layers of SIC. Subject to error propagation.	Requires multiple layers of SIC in each group and a single layer of SIC if groups are made of 2 users. Subject to error propagation.	Does not require any SIC.	Requires multiple layers of SIC. Subject to error propagation.	Requires a single layer of SIC for all users. Less subject to error propagation.

Numerical Results

Compare **SDMA, NOMA and RSMA**

Effect of **user channel alignment/orthogonality** θ and **channel strength disparity** γ

$$\mathbf{h}_1 = [1, 1, 1, 1]^H,$$

$$\mathbf{h}_2 = \gamma \times \left[1, e^{j\theta}, e^{j2\theta}, e^{j3\theta}\right]^H$$

Effect of **load**: underloaded $(K \leq N_t)$ and overloaded $(K > N_t)$ regime

Effect of **CSIT inaccuracy**

Numerical Results: $K = 2$, $N_t = 4$, Perfect CSIT

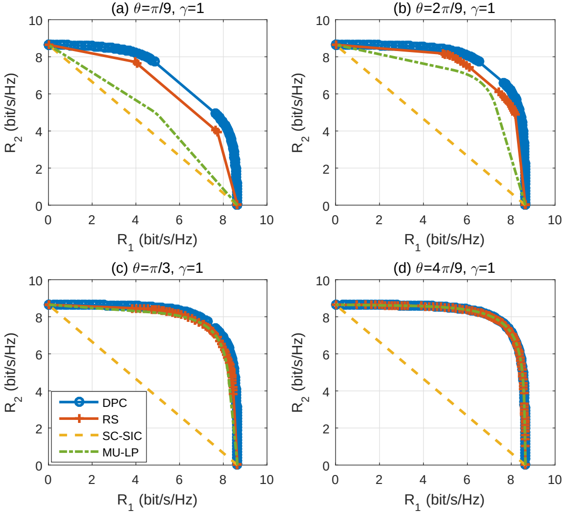


Figure: Achievable rate region of different strategies when $\gamma = 1$, SNR=20 dB.

Numerical Results: $K = 2$, $N_t = 4$, Perfect CSIT

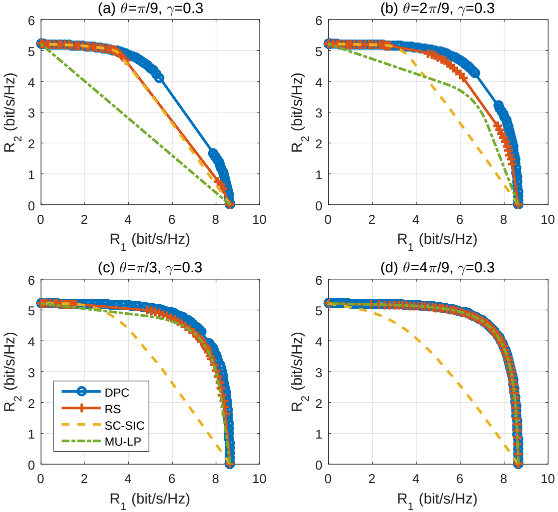


Figure: Achievable rate region with different strategies when $\gamma = 0.3$, $N_t = 4$, SNR=20 dB.

Numerical Results: $K = 2$, $N_t = 4$, Imperfect CSIT

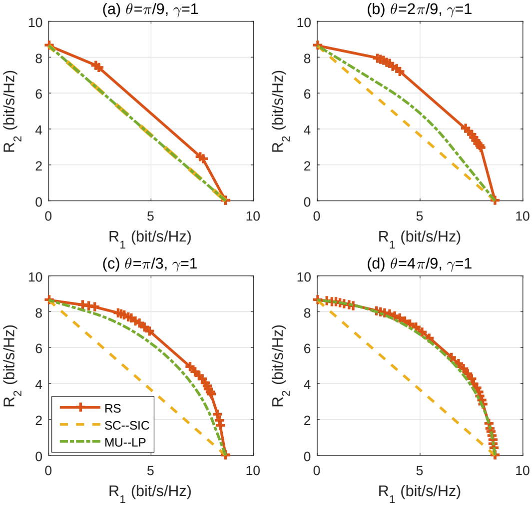


Figure: Achievable rate region of different strategies when $\gamma = 1$, $N_t = 4$, SNR=20 dB.

Numerical Results: $K = 2$, $N_t = 4$, Imperfect CSIT

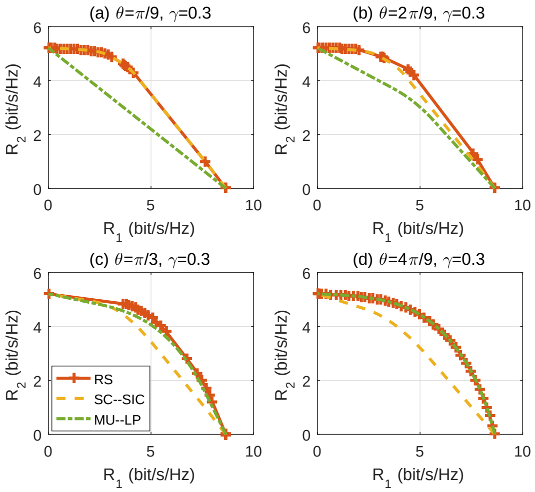


Figure: Achievable rate region with different strategies when $\gamma = 0.3$, $N_t = 4$, SNR=20 dB.

Numerical Results: $K = 3$, $N_t = 2$, Perfect CSIT

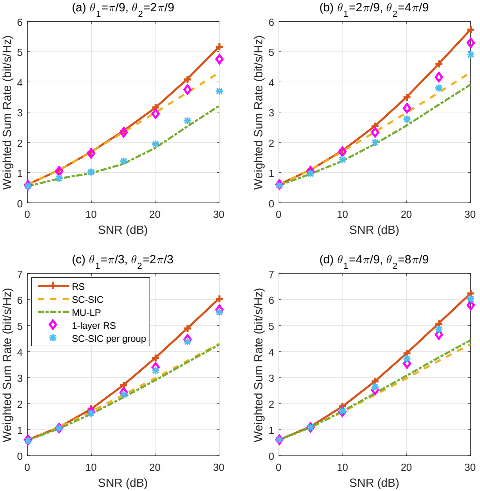


Figure: Weighted sum rate versus SNR comparison of different strategies for overloaded three-user deployment with perfect CSIT, $\gamma_1 = 1, \gamma_2 = 0.3$, $u_1 = 0.4, u_2 = 0.3, u_3 = 0.3,\ N_t = 2,\ \mathbf{r}_{th} = [0.02, 0.08, 0.19, 0.3, 0.4, 0.4, 0.4]$

Numerical Results: $K = 3$, $N_t = 1$, Perfect CSIT

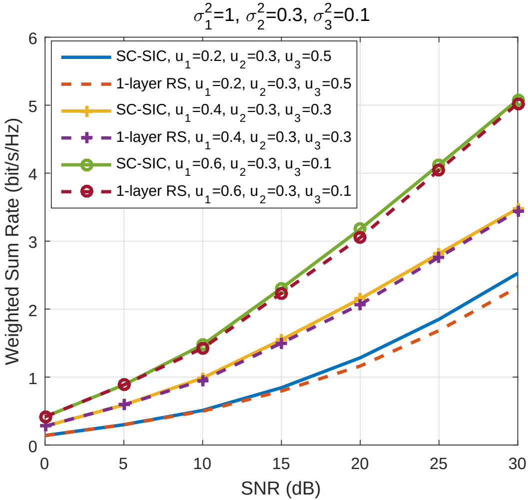


Figure: Weighted sum rate versus SNR comparison of different strategies for overloaded three-user deployment with perfect CSIT, $\sigma_1^2 = 1, \sigma_2^2 = 0.3, \sigma_3^2 = 0.1$, $N_t = 1$, $\mathbf{r}_{th} = [0, 0, 0.01, 0.03, 0.1, 0.2, 0.3]$

Numerical Results: $K = 4$, $N_t = 2$, Perfect CSIT

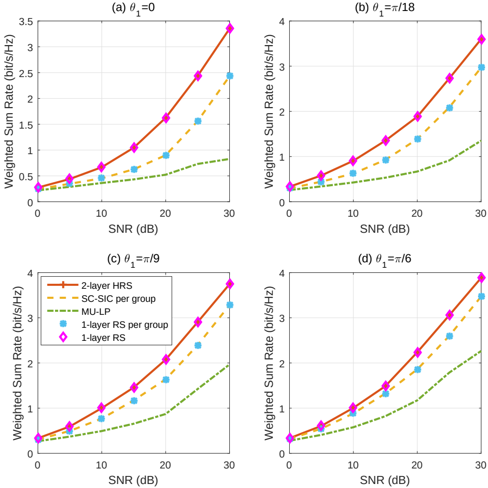


Figure: Sum rate versus SNR comparison of different strategies for overloaded four-user deployment with perfect CSIT, $\gamma_1 = 0.3$, $\theta_2 = \theta_1 + \frac{\pi}{9}$

Numerical Results: $K = 10$, $N_t = 2$, Perfect CSIT

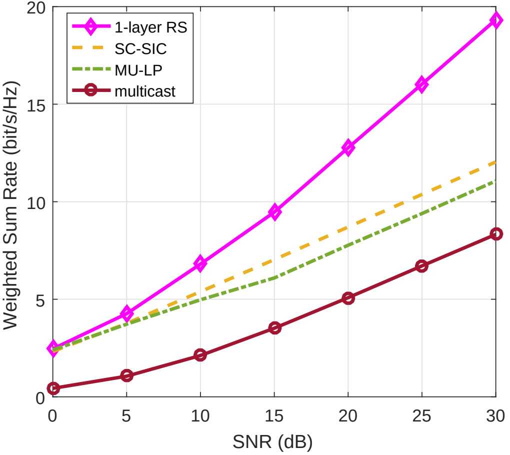


Figure: Weighted sum rate versus SNR comparison of different strategies for overloaded ten-user deployment with perfect CSIT, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_{10}^2 = 1$, $N_t = 2$, SNR=30 dB,$\mathbf{r}_{th} = [0.01, 0.03, 0.05, 0.1, 0.1, 0.1, 0.1]$

Numerical Results: $K = 10$, $N_t = 2$, Perfect CSIT

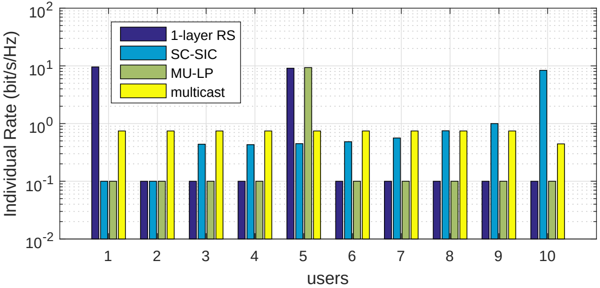


Figure: Individual rate comparison of different strategies for overloaded ten-user deployment with perfect CSIT for 1 randomly generated channel estimate, SNR=30 dB, $N_t = 2$, $\mathbf{r}_{th} = [0.01, 0.03, 0.05, 0.1, 0.1, 0.1, 0.1]$

DoF and rate gains with only **1 SIC layer** vs NOMA that requires **9 SIC layers**!

Partially decode interference and **partially treat interference as noise**: enhanced throughput and QoS, increased robustness and lower complexity

RSMA

New multiple access called **Rate-Splitting Multiple Access** (RSMA)

SDMA and NOMA subject to many limitations: high system complexity and a lack of robustness to user deployments, network load and CSIT inaccuracy

General multiple access framework based on rate-splitting (RS)

Partially decode interference and **partially** treat interference as noise

1-layer RS: low scheduler and receiver complexity and good performance in any user deployments, CSIT inaccuracy and network load

RSMA has the **potential to change** the design of the PHY and MAC layers of **next generation communication systems** by unifying existing approaches and relying on a **superposed transmission of common and private messages**

Unicast and Multicast Transmission

Unicast: one-to-one

Multicast: one-to-many

Non-orthogonal unicast and multicast: superimposed in the power domain

Efficient bandwidth and transmit power utilization/allocation

Interference between multicast and unicast and among the unicast messages

Applications:

- B5G: scarcity of radio resources and heterogeneity of applications
- Layered Division Multiplexing (LDM) in digital television systems
- ...

Unicast and Multicast Transmission

N_t antennas serving K single-antenna users

Transmitter

- multicast message W_0 intended for all users
- K unicast messages $W_1, \ldots, W_K$ intended for different users
- $W_0, W_1, \ldots, W_K$ independently encoded into data streams $s_0, s_1, \ldots, s_K$
- Streams are precoded by $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_K$

$$\mathbf{x} = \mathbf{P}\mathbf{s} = \mathbf{p}_0 s_0 + \sum_{k \in \mathcal{K}} \mathbf{p}_k s_k$$

Receiver

$$y_k = \mathbf{h}_k^H \mathbf{x} + n_k$$

- Decode multicast stream by treating unicast streams as interference
- Perform SIC and decode its intended unicast stream

SIC needed to separate multicast and unicast

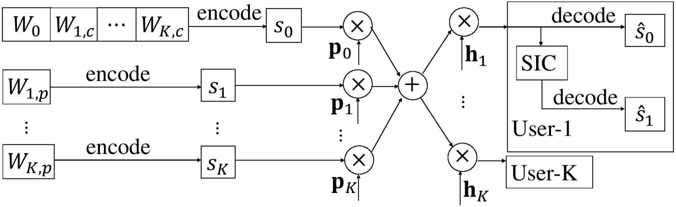
Rate-Splitting for Unicast and Multicast Transmission [35]

SIC used for a dual purpose

- Separate the unicast and multicast streams (as before)
- Better manage interference among unicast streams

Rate-Splitting (RS) needed at the transmitter

- W_k split into common part $W_{k,c}$ and private part $W_{k,p}$, $\forall k$
- $W_{1,c}, \dots, W_{K,c}$ encoded along with W_0 into super-common stream s_0
- s_0 includes multicast message and *parts* of the unicast messages
- $W_{1,p}, \dots, W_{K,p}$ independently encoded into private streams $s_1, \dots, s_K$



Optimization

Conventional (MU-LP)

$$R_{\mathrm{MU-LP}}(\mathbf{u}) = \max_{\mathbf{P}} \sum_{k\in\mathcal{K}} u_k R_k$$

$$\text{s.t.} \quad R_{k,0} \geq R_0^{th}, \forall k \in \mathcal{K}$$

$$\mathrm{tr}(\mathbf{P}\mathbf{P}^H) \leq P_t$$

Rate-Splitting (RS)

$$R_{\mathrm{RS}}(\mathbf{u}) = \max_{\mathbf{P},\mathbf{c}} \sum_{k\in\mathcal{K}} u_k R_{k,tot}$$

$$\text{s.t.} \quad C_0 + \sum_{k\in\mathcal{K}} C_{k,0} \leq R_{k,0}, \forall k \in \mathcal{K}$$

$$C_0 \geq R_0^{th}$$

$$C_{k,0} \geq 0, \forall k \in \mathcal{K}$$

$$\mathrm{tr}(\mathbf{P}\mathbf{P}^H) \leq P_t$$

Numerical Results for Two Users

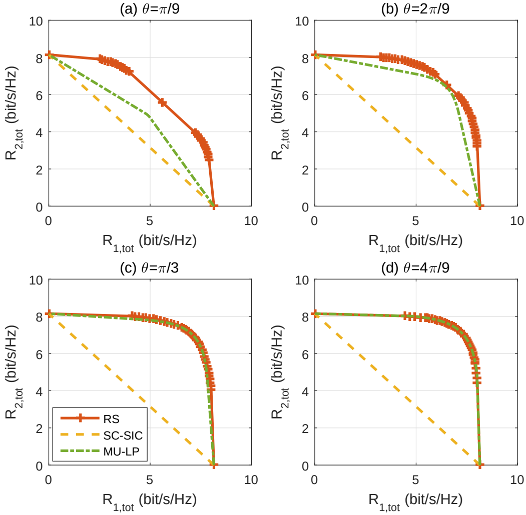


Figure: Achievable rate region comparison of different strategies in perfect CSIT, $\gamma = 1$, $R_0^{th} = 0.5$ bit/s/Hz

Numerical Results for Two Users

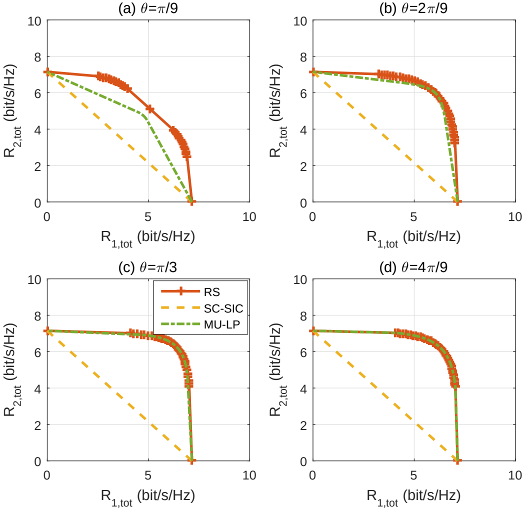


Figure: Achievable rate region comparison of different strategies in perfect CSIT, $\gamma = 1,\ R_0^{th} = 1.5$ bit/s/Hz

Numerical Results for Two Users

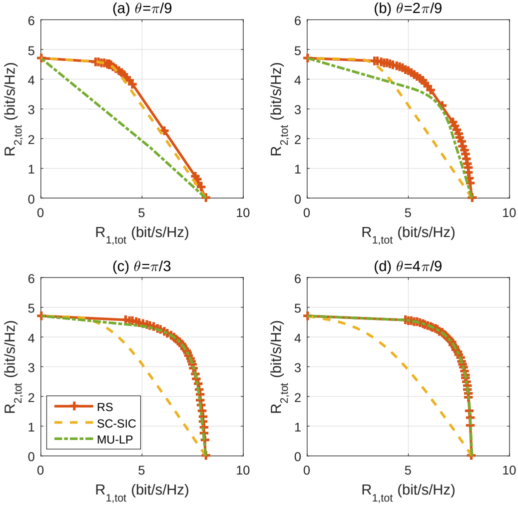


Figure: Achievable rate region comparison of different strategies in perfect CSIT, $\gamma = 0.3$, $R_0^{th} = 0.5$ bit/s/Hz

Rate-Splitting for Unicast and Multicast Transmission

RS efficiently exploits **existing SIC** receiver architecture.

SIC used for **dual purpose** of separating unicast and multicast streams and better manage the multi-user interference between the unicast streams.

RS **outperforms** existing Multi-User Linear-Precoding (MULP) and power-domain Non-Orthogonal Multiple Access (NOMA) in a wide range of user deployments (with a diversity of channel directions and channel strengths).

RS provides rate and QoS enhancements at **no extra cost for the receivers**.

Rate-Splitting in 5G

Standardization Issues

- (H/T)RS is a generalized strategy
  - Conventional SU/MU-MIMO and CoMP as special cases.
  - new SU/MU/RS mode switching in 5G depending on the SNR and the CSIT quality.
- New transmission mode indicator (DCI format)
  - Inform the Tx mode and the relevant information required for demodulation
- New signaling from BS to UEs
  - Number and type of messages (common/private)
  - Modulation and coding scheme of all common/private message
  - Information about whether common message is intended for the user or not
  - Transmit power of each message.
- CSI feedback mechanisms and signaling
  - Knowledge about the CSIT accuracy to allocate power to the common and private messages, e.g. computed by a UE and reported back to the BS.
  - Scheduling and Tx strategy decided based on CSIT accuracies from all users in all subbands.
  - CSI reporting on PUCCH and PUSCH.
  - Some CSIT patterns lead to a higher DoF than others [25].

From LTE Rel. 13/14 to RS

- NAICS in Rel-12
  - Network-Assisted Interference Cancellation and Suppression.
  - Providing knowledge about interfering transmissions at the receivers.
  - Allows the use of more advanced receivers (joint decoding, SIC).
- Downlink Multiuser Superposition Transmission (MUST)
  - Study item approved (3GPP).
  - Non-Orthogonal Multiple Access (NOMA).
  - Uses superposition coding at the transmitter.
  - Relies on SIC at the receivers.
  - Metric of interest: sum-rate, fairness, delays etc.

The machinery required for RS is already being studied, discussed and developed.

- RS schemes can utilize such developments (or the other way around!).
- RS can fit in nicely.
- RS complements other schemes, and vice versa.

Conclusions and Future Challenges

Conclusions and Future Challenges

- 4G and current 5G candidates (MU-MIMO, CoMP, Massive MIMO, millimetre wave MIMO) rely on *private message transmissions*
  - *Treat interference as noise*
  - Such a strategy is only motivated in the presence of perfect CSIT
  - *Apply techniques designed for perfect CSIT to imperfect CSIT*

- NOMA forces strong users to *fully decode* and cancel interference created by weaker users:
  - Works only for degraded channels (SISO BC or MISO BC with aligned channels)

- RS **partially** decodes interference and **partially** treats interference as noise
  - Superposed transmission of common and private messages
  - Motivated by information theory for the realistic scenario of imperfect CSIT
  - A more general and powerful transmission framework
  - Benefits: unified framework, spectral/energy efficiencies, reliability, CSI feedback overhead reduction

- RS leads to fundamental changes in the design of PHY and Lower MAC
  - A gold mine of research problems for academia and industry

- The standardization of rate-splitting can leverage 3GPP current study/work items

Future Challenges: A gold mine of research problems

Introduction

- Overview, open problems, impact on standard specifications and operational challenges [28].

Fundamental Limits

- DoF region for K-user MISO BC with imperfect CSIT [8, 31].
- Capacity region of K-user MISO BC with imperfect CSIT: DPC + RS?
- DoF region for MIMO BC with imperfect CSIT [10, 11, 12].
- DoF region of overloaded MISO BC with imperfect CSIT [23].
- DoF region for MISO IC with imperfect CSIT [22]. TRS?
- DoF region for MIMO IC with imperfect CSIT [11]. RS + IA?
- Interplay between RS and coded caching [24, 33].

Optimization

- Ergodic sum-rate maximization for BC [14].
- Robust Max-Min Fairness for BC [15].
- RS beamforming optimization for other types of channels.

PHY challenges

- Finite SNR rate analysis [9].
- Energy efficiency of RS-based transmission [37].
- Space-time/frequency RS [26, 27, 9].
- RS with multi-carrier transmissions.

Future Challenges

PHY challenges (continued)

- RS with non-linear precoding [38].
- Diversity (and BER) performance of RS-based strategies.
- RS for Multigroup Multicast [16, 17].
- RS/HRS for Massive MIMO [13].
- RS as a way to combat pilot contamination.
- RS to mitigate hardware impairments [30].
- RS in higher frequency bands operation (e.g. millimeter-wave) [29].
- RS-based network MIMO [36].
- Coordination/cooperation among distributed antennas in homogeneous and heterogeneous network deployments.
- RS in half-duplex relay.
- RS in full duplex.
- RS in overloaded systems [23].
- RS and NOMA/MUST [34].
- RS and superposition of multicast and unicast messages [35].
- RS and physical layer security.
- RS in D2D and cognitive radio [32].

Future Challenges

PHY/MAC challenges

- User pairing and scheduling of common and private messages.
- RS design with Quality of Experience (QoE) and traffic constraints.

Performance Analysis

- Performance analysis of RS using stochastic geometry.

Standardization

- Link and system-level evaluations of RS.
- MIMO receiver implementation.
- Transmission schemes/mode.
- CSI feedback mechanisms.
- Downlink and uplink signaling.

References I

[1] E. Bjornson, E. G. Larsson and T. L. Marzetta, “Massive MIMO: ten myths and one critical question,” IEEE Communications Magazine, vol. 54, no. 2, pp. 114-123, February 2016.

[2] E. G. Larsson, O. Edfors, F. Tufvesson and T. L. Marzetta, “Massive MIMO for next generation wireless systems,” IEEE Communications Magazine, vol. 52, no. 2, pp. 186-195, February 2014.

[3] N. Jindal, “MIMO Broadcast Channels With Finite-Rate Feedback,” Trans. on Info Theory, vol. 52, no. 11, pp. 5045-5060, Nov. 2006.

[4] B. Clerckx, G. Kim, and S. Kim, “MU-MIMO with channel statistics-based codebooks in spatially correlated channels,” in IEEE Global Communications Conf. (Globecom), Nov 2008, pp. 15.

[5] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, Multiuser MIMO achievable rates with downlink training and channel state feedback, IEEE Trans. Inf. Theory, vol. 56, no. 6, pp. 28452866, Jun. 2010.

[6] A. El Gamal and Y.-H. Kim, *Network information theory*, Cambridge university press, 2011.

[7] S. Yang, M. Kobayashi, D. Gesbert, and X. Yi, Degrees of freedom of time correlated MISO broadcast channel with delayed CSIT, IEEE Trans. Inf. Theory, vol. 59, no. 1, pp. 315328, Jan. 2013.

[8] A.G. Davoodi, S.A. Jafar, “Aligned Image Sets under Channel Uncertainty: Settling a Conjecture by Lapidoth, Shamai and Wigger on the Collapse of Degrees of Freedom under Finite Precision CSIT,” arXiv:1403.1541.

[9] C. Hao, Y. Wu and B. Clerckx, “Rate analysis of two-receiver MISO broadcast channel with finite rate feedback: A rate-splitting approach,” IEEE Trans. on Comm., vol. 63, no. 9, pp. 3232-3246, Sept. 2015.

[10] J. Chen and P. Elia, Symmetric two-user MIMO BC with evolving feedback, in Inf. Theory and Applications Workshop (ITA), 2014, San Diego, CA, Feb 2014, pp. 15.

[11] C. Hao, B. Rassouli and B. Clerckx, “Achievable DoF Regions of MIMO Networks with Imperfect CSIT,” IEEE Trans. on Info Theory, vol. 63, no. 10, pp. 6587-6606, Oct 2017.

[12] B. Yuan, S.A. Jafar, “Elevated Multiplexing and Signal Space Partitioning in the 2 User MIMO IC with Partial CSIT,” IEEE SPAWC 2016.

[13] M. Dai, B. Clerckx, D. Gesbert and G. Caire, “A Rate Splitting Strategy for Massive MIMO with Imperfect CSIT,” IEEE IEEE Trans. on Wireless Comm., vol. 15, no. 7, pp. 4611-4624, July 2016.

[14] H. Joudeh and B. Clerckx, “Sum-Rate Maximization for Linearly Precoded Downlink Multiuser MISO Systems with Partial CSIT: A Rate-Splitting Approach,” IEEE Trans. on Comm., IEEE Trans. on Comm. vol. 64, no. 11, pp. 4847-4861, Nov 2016.

References II

[15] H. Joudeh and B. Clerckx, "Robust Transmission in Downlink Multiuser MISO Systems: A Rate-Splitting Approach," IEEE Trans. on Sig. Proc. Vol. 64, No. 23, pp. 6227-6242 , Dec 2016.

[16] H. Joudeh and B. Clerckx, "A Rate-Splitting Strategy for Max-Min Fair Multigroup Multicasting," IEEE SPAWC 2016.

[17] H. Joudeh and B. Clerckx, Rate-Splitting for Max-Min Fair Multigroup Multicast Beamforming in Overloaded Systems, IEEE Trans. on Wireless Commun., vol 16, no 11, pp 7276-7289, Nov 2017.

[18] C. Lim, T. Yoo, B. Clerckx, B. Lee, and B. Shim, "Recent trend of multiuser MIMO in LTE-advanced," IEEE Commun. Mag., vol. 51, no. 3, pp. 127135, March 2013.

[19] A.Adhikary, J.Nam, J.Ahn and G.Caire. "Joint spatial division and multiplexing - the large-scale array regime," IEEE TIT, Oct. 2013.

[20] J. Chen and V. Lau, "Two-tier precoding for FDD multi-cell massive MIMO time-varying interference networks," IEEE J. Sel. Areas Commun., vol. 32, no. 6, pp. 12301238, June 2014.

[21] X. Yi, D. Gesbert, S. Yang, and M. Kobayashi, On the dof of the multiple-antenna time correlated interference channel with delayed csit, in 2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Nov 2012, pp. 15661570.

[22] C. Hao and B. Clerckx, "MISO Networks with Imperfect CSIT: A Topological Rate Splitting Approach," IEEE Trans. on Comm., Vol 65, No 5, pp. 2164-2179, May 2017

[23] E. Piovano, H. Joudeh and B. Clerckx, "Overloaded MU-MISO Transmission with Imperfect CSIT," 2016 Asilomar Conference on Signals, Systems, and Computers.

[24] J. Zhang, F. Engelmann and P. Elia, "Coded caching for reducing CSIT-feedback in wireless communications," Allerton, Monticello, IL, 2015, pp. 1099-1105.

[25] B. Rassouli, C. Hao and B. Clerckx, "DoF Analysis of the MIMO Broadcast Channel with Alternating/Hybrid CSIT," IEEE Trans. on Info Theory, vol. 62, no. 3, pp. 1312-1325, Mar. 2016.

[26] R. Tandon, S. Jafar, S. S. Shitz, and H. Poor, On the synergistic benefits of alternating CSIT for the MISO broadcast channel, IEEE Trans. Inf. Theory., vol. 59, no. 7, pp. 41064128, Jul. 2013.

[27] J. Chen and P. Elia, Optimal DoF region of the two-user MISO-BC with general alternating CSIT, in Proc. Asilomar Conf. Signals, Syst. Comput., Pacific Grove, CA, USA, Nov. 2013, pp. 18601864.

References III

[28] B. Clerckx, H. Joudeh, C. Hao, M. Dai and B. Rassouli, “Rate Splitting for MIMO Wireless Networks: A Promising PHY-Layer Strategy for LTE Evolution,” IEEE Comm. Mag, pp 98-105, May 2016.

[29] M. Dai and B. Clerckx, “Multiuser Millimeter Wave Beamforming Strategies with Quantized and Statistical CSIT,” IEEE Trans. on Wireless Comm., vol 16, no 11, pp 7025-7038, Nov 2017.

[30] A. Papazafeiropoulos, B. Clerckx, and T. Ratnarajah, “Rate-Splitting to Mitigate Residual Transceiver Hardware Impairments in Massive MIMO Systems,” IEEE Trans. on Veh. Tech., vol 66, no 9, pp 8196-8211, Sept 2017.

[31] E. Piovano and B. Clerckx, “Optimal DoF region of the K-User MISO BC with Partial CSIT,” IEEE Commun. Letters, vol 21, no 11, pp 2368-2371, Nov 2017.

[32] A. Zappone, B. Matthiesen, E. Jorswieck, “Energy Efficiency in MIMO Underlay and Overlay Device-to-Device Communications and Cognitive Radio Systems,” IEEE Transactions on Signal Processing, 2016.

[33] E. Piovano, H. Joudeh and B. Clerckx, “On Coded Caching in the Overloaded MISO Broadcast Channel,” IEEE ISIT 2017.

[34] Y. Mao, B. Clerckx and V.O.K. Li, “Rate-Splitting Multiple Access for Downlink Communication Systems: Bridging, Generalizing and Outperforming SDMA and NOMA,” EURASIP Journal on Wireless Communications and Networking, May 2018.

[35] Y. Mao, B. Clerckx and V.O.K. Li, “Rate-Splitting for Multi-Antenna Non-Orthogonal Unicast and Multicast Transmission,” IEEE SPAWC 2018.

[36] Y. Mao, B. Clerckx and V.O.K. Li, “Rate-Splitting Multiple Access for Cooperative Multi-Cell Networks,” arXiv:1804.10516.

[37] Y. Mao, B. Clerckx and V.O.K. Li, “Energy Efficiency of Rate-Splitting Multiple Access, and Performance Benefits over SDMA and NOMA,” IEEE ISWCS 2018.

[38] Andre R. Flores, Bruno Clerckx and Rodrigo C. de Lamare, “Tomlinson-Harashima Precoded Rate-Splitting for Multiuser Multiple-Antenna Systems,” IEEE ISWCS 2018.

See also recent special sessions on RS at IEEE SPAWC 2018 and IEEE ISWCS 2018.