Massive Random Access with Massive MIMO

Wei Yu

Joint Work with Zhilin Chen, Foad Sohrabi, Ya-Feng Liu
Liang Liu, Justin Kang

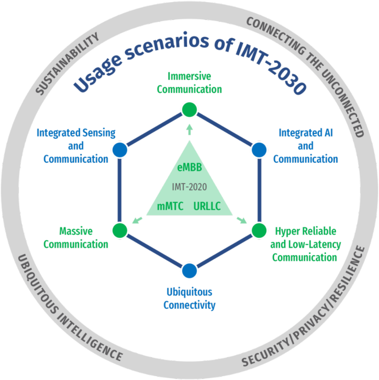
University of Toronto

2024

Introduction

Massive device connectivity is a key requirement for 5G/6G cellular networks

- Machine-type (M2M) communications, Internet of Things (IoT), Sensors...
- Sporadic traffic with low latency requirement
- Large number of devices but only a few are active at a time


Usage scenarios of IMT-2030
SUSTAINABILITY
CONNECTING THE UNCONNECTED
UBIQUITOUS INTELLIGENCE
SECURITY/PRIVACY/RESILIENCE
Immersive Communication
Integrated Sensing and Communication
Integrated AI and Communication
eMBB
IMT-2020
mMTC
URLLC
Massive Communication
Hyper Reliable and Low-Latency Communication
Ubiquitous Connectivity

Massive Connectivity for Internet-of-Things (IoT)

How to design wireless systems to support massive connectivity?

- Contention vs scheduling in massive random access
- Sparse device activity detection algorithms utilizing massive MIMO
- Minimum feedback for collision-free scheduling

Massive Connectivity for Internet-of-Things (IoT)

How to design wireless systems to support massive connectivity?

- Contention vs scheduling in massive random access
- Sparse device activity detection algorithms utilizing massive MIMO
- Minimum feedback for collision-free scheduling

The main points of this talk:

- To support massive connectivity, the use of non-orthogonal pilots is inevitable.

Massive Connectivity for Internet-of-Things (IoT)

How to design wireless systems to support massive connectivity?

- Contention vs scheduling in massive random access
- Sparse device activity detection algorithms utilizing massive MIMO
- Minimum feedback for collision-free scheduling

The main points of this talk:

- To support massive connectivity, the use of non-orthogonal pilots is inevitable.
- Compressed sensing techniques are indispensable for device detection.

Massive Connectivity for Internet-of-Things (IoT)

How to design wireless systems to support massive connectivity?

- Contention vs scheduling in massive random access
- Sparse device activity detection algorithms utilizing massive MIMO
- Minimum feedback for collision-free scheduling

The main points of this talk:

- To support massive connectivity, the use of non-orthogonal pilots is inevitable.
- Compressed sensing techniques are indispensable for device detection.
- Massive MIMO can significantly enhance device activity detection.

Massive Connectivity for Internet-of-Things (IoT)

How to design wireless systems to support massive connectivity?

- Contention vs scheduling in massive random access
- Sparse device activity detection algorithms utilizing massive MIMO
- Minimum feedback for collision-free scheduling

The main points of this talk:

- To support massive connectivity, the use of non-orthogonal pilots is inevitable.
- Compressed sensing techniques are indispensable for device detection.
- Massive MIMO can significantly enhance device activity detection.
- Channel estimation is the main bottleneck.

Massive Connectivity for Internet-of-Things (IoT)

How to design wireless systems to support massive connectivity?

- Contention vs scheduling in massive random access
- Sparse device activity detection algorithms utilizing massive MIMO
- Minimum feedback for collision-free scheduling

The main points of this talk:

- To support massive connectivity, the use of non-orthogonal pilots is inevitable.
- Compressed sensing techniques are indispensable for device detection.
- Massive MIMO can significantly enhance device activity detection.
- Channel estimation is the main bottleneck.
- Scheduling is a viable alternative to contention for massive random access.

Massive Connectivity for Internet-of-Things (IoT)

How to design wireless systems to support massive connectivity?

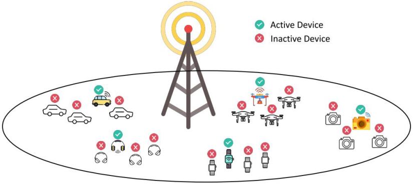
- Contention vs scheduling in massive random access
- Sparse device activity detection algorithms utilizing massive MIMO
- Minimum feedback for collision-free scheduling

The main points of this talk:

- To support massive connectivity, the use of non-orthogonal pilots is inevitable.
- Compressed sensing techniques are indispensable for device detection.
- Massive MIMO can significantly enhance device activity detection.
- Channel estimation is the main bottleneck.
- Scheduling is a viable alternative to contention for massive random access.
- Feedback for collision-free scheduling requires only very low rate.

Contention-Based vs Coordinated Scheduling

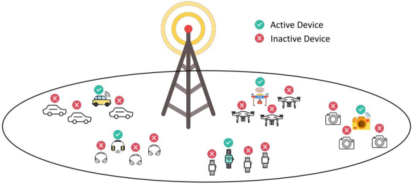
- Up to $10^5 \sim 10^6$ devices with sporadic traffic are connected to each BS.



- **Uncoordinated Random Access:**
  - Classic Slotted ALOHA: Contention-based uncoordinated scheduling.
  - Coded ALOHA can alleviate some of the inefficiencies of classic ALOHA.
- **Coordinated Random Access:**
  - Coordinated random access requires the BS to detect the active users.
  - Coordinated scheduling also requires feedback from the BS to the active users.

Contention-Based vs Coordinated Scheduling

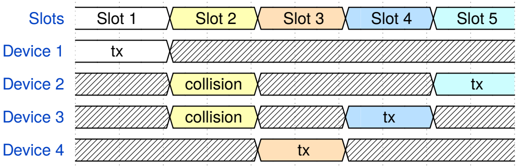
- Up to $10^5 \sim 10^6$ devices with sporadic traffic are connected to each BS.



- **Uncoordinated Random Access:**
  - Classic Slotted ALOHA: Contention-based uncoordinated scheduling.
  - Coded ALOHA can alleviate some of the inefficiencies of classic ALOHA.
- **Coordinated Random Access:**
  - Coordinated random access requires the BS to detect the active users.
  - Coordinated scheduling also requires feedback from the BS to the active users.
- **This talk is about the cost and benefit of coordination:**
  - How to design sparse user activity detection algorithms?
  - What is the minimum required feedback for collision-free scheduling?

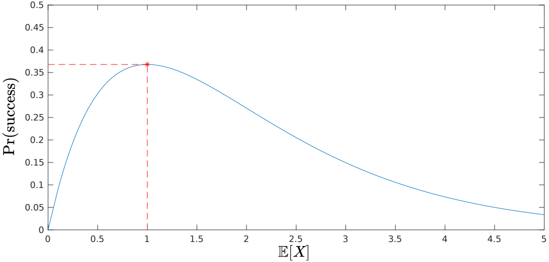
Classic Uncoordinated Solution: Slotted ALOHA

Slotted ALOHA involves contention and is uncoordinated
involving no communication between BS and users.



- Fix a finite set of *orthogonal* pilot sequences.
- An active user picks one of the pilot sequences at random.
- Transmission is successful only if no other users pick the same sequence.
- If there is a collision, users must re-transmit.

Slotted ALOHA: Analysis

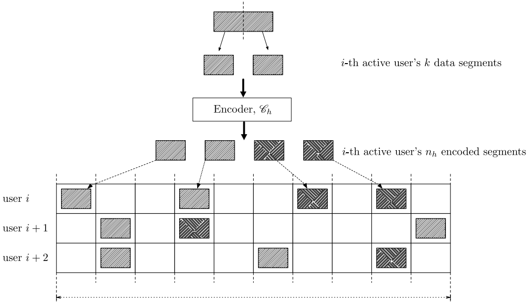


- Let X be the number of users that transmit in a particular slot.
- Since X is sum of independent Bernoulli trials, it follows Poisson distribution

$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda k}}{k!}, \quad \text{where } \mathbb{E}[X] = \lambda. \tag{1}$$

- Successful transmission only when $k = 1$, with probability $\lambda e^{-\lambda}$.
- Optimize over λ. Throughput is maximized when $\lambda = 1$ with P(success) $= \frac{1}{e}$.
- Slots with collision or slots with no transmission (i.e., 63% slots) are wasted.

Coded Slotted ALOHA


i-th active user's k data segments
Encoder, $\mathscr{C}_h$
i-th active user's n_h encoded segments
user i
user $i+1$
user $i+2$

- Coded Slotted ALOHA: Use packet-level erasure codes and successive interference cancellation (SIC) to extract information from collisions.
- Each user chooses an (n_h, k) erasure code $\mathscr{C}_h$ to encode their k segments.
- Code is chosen from a finite set $\{\mathscr{C}_h\}_{h=1}^{\theta}$ according to some p.m.f., and the n_h packets are transmitted randomly over a fixed frame.

E. Paolini, G. Liva, and M. Chiani, "Coded Slotted ALOHA: A Graph-Based Method for Uncoordinated Multiple Access," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6815–6832, 2015.

Coded Slotted ALOHA: Graph Representation

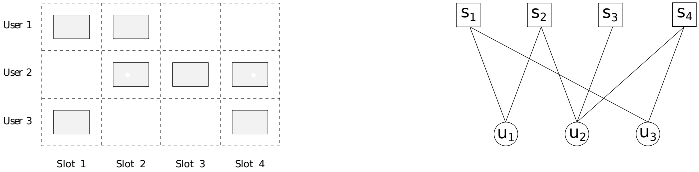


Figure: Bipartite graph model for contention resolution

- Users are represented by variable nodes, slots by check nodes.
- A user node u_i is connected to slot node s_j if user i transmits in slot j.
- Decoding process is identical to the peeling decoder for erasure channel.
- If users select repetition codes, this is known as Contention Resolution Diversity Slotted ALOHA (CRDSA).

Coded Slotted ALOHA: Decoding Example

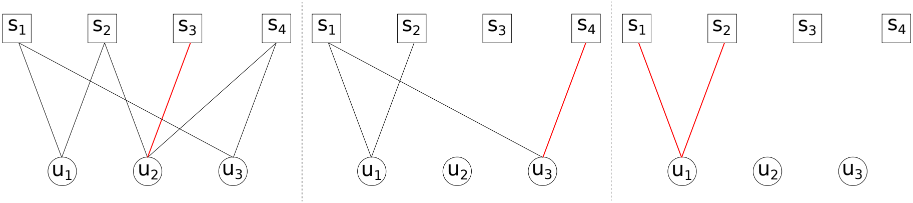
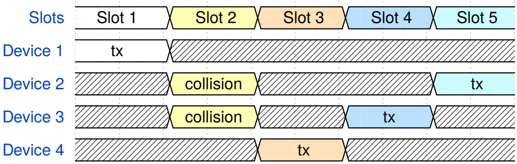


Figure: Peeling decoding for CRDSA on a bipartite graph.

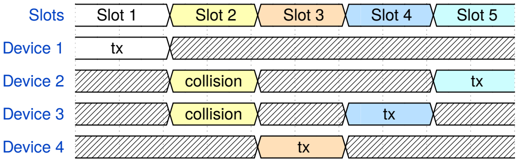
- Decoding procedure for CRDSA is similar to Fountain code or LT code.
- This connection allows us to show that the optimal user-node degree distribution is the soliton distribution [Narayanan-Pfister'12].
- With this degree distribution, the throughput $\triangleq \frac{\#\text{ of decoded users}}{\#\text{ of slots}} \to 1$ asymptotically as the number of users and slots go to infinity.

Contention vs. Scheduling



- Slotted ALOHA based schemes all involve contention and collision resolution
  - Multiple transmissions increase power consumption.
  - Collision resolution increases delay.
  - Practical coding schemes operate at less than optimal throughput.
- Scheduling is an alternative approach to contention.
- Contention-based schemes are often justified based on the assumption that the cost of coordination is too great.

Contention vs. Scheduling



Slots	Slot 1	Slot 2	Slot 3	Slot 4	Slot 5
Device 1	tx				
Device 2		collision			tx
Device 3		collision		tx	
Device 4			tx		

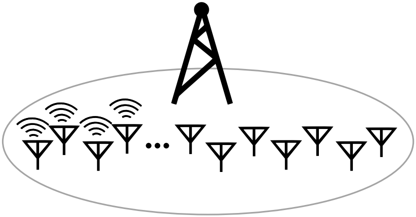
- Slotted ALOHA based schemes all involve contention and collision resolution
  - Multiple transmissions increase power consumption.
  - Collision resolution increases delay.
  - Practical coding schemes operate at less than optimal throughput.
- Scheduling is an alternative approach to contention.
- Contention-based schemes are often justified based on the assumption that the cost of coordination is too great.

What is the cost of coordination?

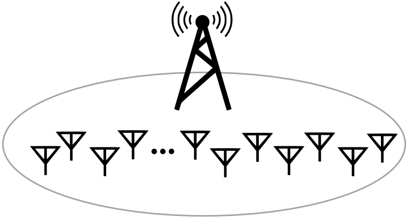
Coordinated Random Access

There are not enough *orthogonal* pilots for every user. Instead, we generate pilots at random and assign unique *non-orthogonal* pilots to all the N potential users.

.



Phase 1 (Activity Detection):
The K active users ($K \ll N$) send their pilots synchronuously to the BS.

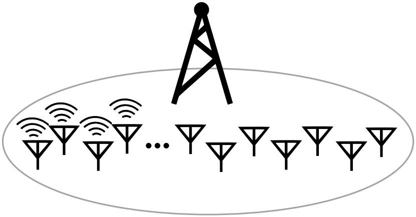


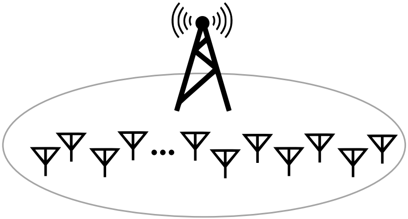
Phase 2 (Downlink Feedback): BS sends a *common* feedback message to schedule the data transmissions of K active users.

Coordinated Random Access

There are not enough *orthogonal* pilots for every user. Instead, we generate pilots at random and assign unique *non-orthogonal* pilots to all the N potential users.

.



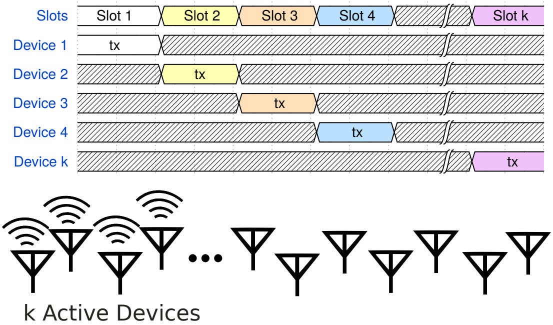


Phase 1 (Activity Detection):
The K active users ($K \ll N$) send their pilots synchronuously to the BS.

Phase 2 (Downlink Feedback): BS sends a *common* feedback message to schedule the data transmissions of K active users.

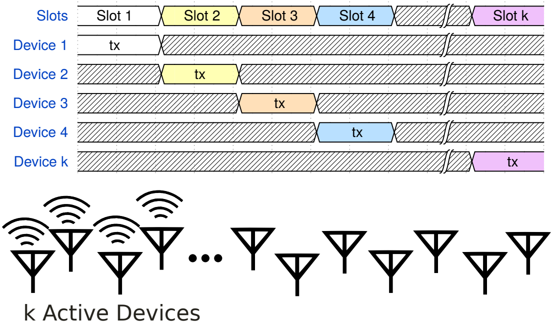
How to perform sparse activity detection?

Feedback-Based Scheduling for Random Access



Phase 3 (Uplink Payload Transmission): The K active users transmit their payload in the K slots based on the schedule provided by the BS, while avoiding collision.

Feedback-Based Scheduling for Random Access

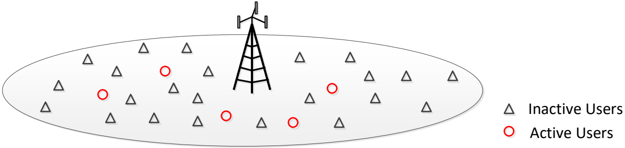


Phase 3 (Uplink Payload Transmission): The K active users transmit their payload in the K slots based on the schedule provided by the BS, while avoiding collision.

What is the minimum feedback needed to ensure collision-free scheduling?

Sparse User Activity Detection

User Activity Detection and Channel Estimation via Pilots



- BS equipped with M antennas
- N single-antenna devices, K of which are active at a time
- Each device is associated with a length-L unique signature sequence $\mathbf{s}_n$
- Channel $\mathbf{h}_n$ of user n is assumed to be fixed during the L symbols.
- For single-cell system, received signal $\mathbf{Y} \in \mathbb{C}^{L \times M}$ at the BS is

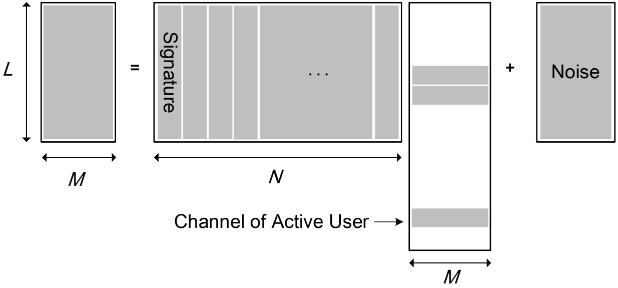
$$\mathbf{Y} = \sum_{n=1}^{N} \alpha_n \mathbf{s}_n \mathbf{h}_n^T + \mathbf{Z} = \mathbf{S}\mathbf{X} + \mathbf{Z}, \tag{2}$$

where

- $\alpha_n \in \{1, 0\}$ activity indicator; $\mathbf{Z} \in \mathbb{C}^{L \times M}$ Gaussian noise with variance σ^2
- $\mathbf{S} \triangleq [\mathbf{s}_1, \ldots, \mathbf{s}_N] \in \mathbb{C}^{L \times N}$; $\mathbf{X} \triangleq [\alpha_1 \mathbf{h}_1, \cdots, \alpha_N \mathbf{h}_N]^T \in \mathbb{C}^{N \times M}$

User Activity Detection via Compressed Sensing

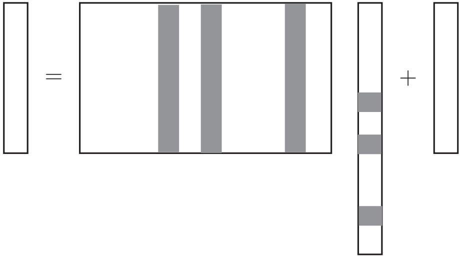
Aim to identify the K non-zero rows of $\mathbf{X}$ from $\mathbf{Y} = \mathbf{S}\mathbf{X} + \mathbf{Z}$.



- Multiple measurement vector (MMV) problem in compressed sensing
  - Columns of $\mathbf{X}$ share the same sparsity pattern, i.e., row sparsity
- Efficiently solved by the approximate message passing (AMP) algorithm [Donoho-Maleki-Montanari'09]

Single-Antenna Case

Identify the columns that correspond to non-zero elements in $\mathbf{x}$ via



Instead of minimizing $\|\mathbf{x}\|_0$, we use a convex relaxation reformulation (LASSO):

$$\hat{\mathbf{x}} = \arg\min \frac{1}{2}\|\mathbf{y} - \mathbf{S}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{3}$$

Soft Thresholding Function

Consider a special case of a single measurement of a scalar, LASSO becomes

$$\hat{x} = \arg\min \frac{1}{2}|y - x|_2^2 + \lambda|x|_1 \tag{4}$$

The solution is explicitly given by

$$\hat{x} = \eta(y; \lambda), \tag{5}$$

where η is a soft thresholding function as

$$\eta(y;\theta) = \begin{cases} y - \theta, & y > \theta \\ 0, & -\theta \leq y \leq \theta \\ y + \theta, & y < -\theta \end{cases} \tag{6}$$

Soft Thresholding Function

$$\eta(y;\theta) = \begin{cases} y-\theta, & y > \theta \\ 0, & -\theta \le y \le \theta \\ y+\theta, & y < -\theta \end{cases}$$

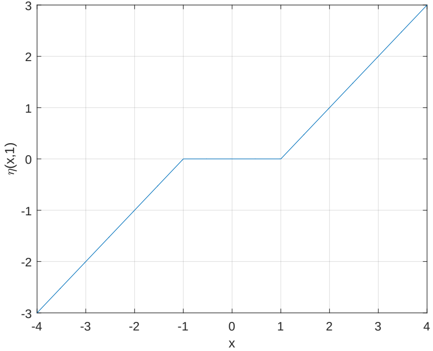
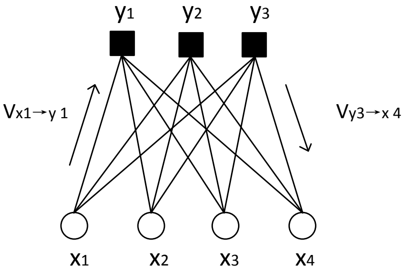


Figure: Soft thresholding function with $\theta = 1$

AMP via Graphical Model

Graphical model with message passing [Donoho-Maleki-Montanari'09]



Main features:

- Soft thresholding emerges in a minimax solution.
- State evolution describes the progress in iteration.
- Better denoiser design is possible by accounting for channel statistics.

AMP Algorithm

Algorithm: Correlate with the residual, denoise, then iterate

$$\mathbf{x}^{t+1} = \eta(\mathbf{x}^t + \mathbf{S}^T\mathbf{r}^t; \lambda + \gamma^t) \tag{7}$$

$$\mathbf{r}^t = \mathbf{y} - \mathbf{S}\mathbf{x}^t + \frac{1}{L}\mathbf{r}^{t-1}\|\mathbf{x}^t\|_0, \tag{8}$$

where the threshold satisfies

$$\gamma^{t+1} = \frac{\lambda + \gamma^t}{L}\|\mathbf{x}^{t+1}\|_0 \tag{9}$$

Note: the threshold γ^{t+1} is fixed by the recursion.

The "Onsager term" comes from statistical physics analysis.
Without it, this becomes the classical iterative soft thresholding.

Designing Denoiser to Match Channel Statistics

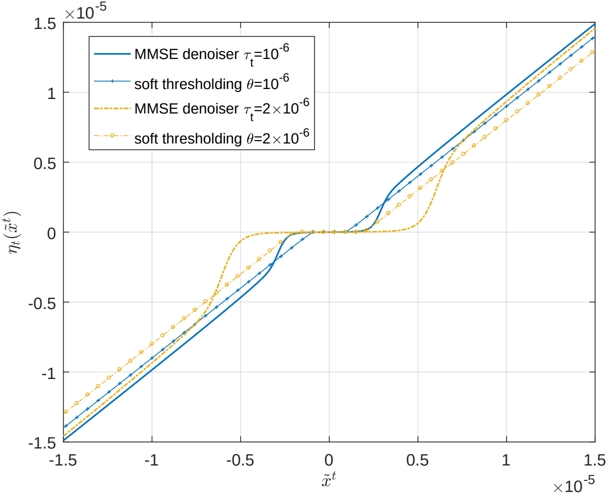


Figure: Soft thresholding denoiser and MMSE denoiser

State Evolution of AMP

The performance of AMP at each iteration can be predicted in the asymptotic regime where $L \to \infty, N \to \infty$ with fixed $\frac{L}{N}$

- $\mathbf{S}^*\mathbf{r}^t + \mathbf{x}^t$ can be modeled as signal plus noise, i.e., $\mathbf{x} + \mathbf{v}^t$
- $\mathbf{v}^t$ is i.i.d. Gaussian noise with variance τ_t tracked by state evolution equation

$$\tau_{t+1}^2 = \sigma_w^2 + \frac{1}{\delta}\mathbb{E}|\eta_t(X + \tau_t W) - X|^2 \tag{10}$$

  - X: random variable following the same distribution as $\mathbf{x}$
  - W: random variable following $\mathcal{CN}(0, 1)$
  - initialization: $\tau_0 \triangleq \sigma_w^2 + \frac{1}{\delta}\mathbb{E}|X|^2$

- Interpretation of state evolution: vector estimation $\mathbf{y} = \mathbf{S}\mathbf{x} + \mathbf{w}$ is reduced to uncoupled scalar estimation $(\mathbf{x}^t + (\mathbf{S}^*\mathbf{r}^t)_i = x_i + v_i^t$

User Activity Detection

The signal-plus-noise model in state evolution of AMP: $(\mathbf{S}^*\mathbf{r}^t + \mathbf{x}^t)_i = x_i + v_i^t$, can be re-expressed as $\tilde{X}^t = X + \tau_t W$ via random variables $\tilde{X}^t, X, W$

Consider the hypothesis testing problem

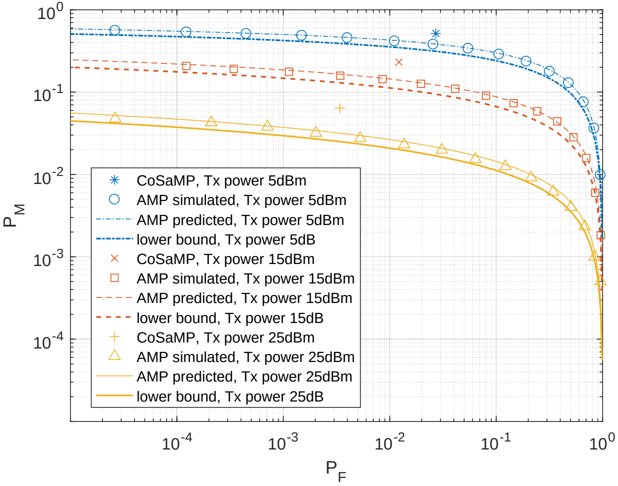
$$\begin{cases} H_0 : X = 0, \text{ user is inactive} \\ H_1 : X \neq 0, \text{ user is active} \end{cases} \tag{11}$$

The optimal decision rule

$$LLR = \log \left(\frac{p_{\tilde{X}^t|X}(\tilde{x}^t|x \neq 0)}{p_{\tilde{X}^t|X}(\tilde{x}^t|x = 0)} \right) \underset{H_1}{\overset{H_0}{\lesseqgtr}} l_{th} \tag{12}$$

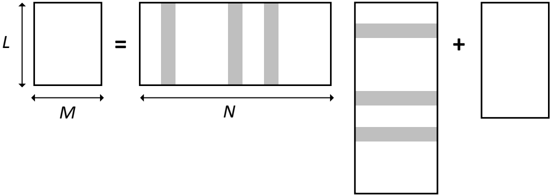
- l_{th}: decision threshold determined by the detection criterion.
- State evolution of AMP gives an analysis of detection error probabilities.

Missed Detection vs. False Alarm Probabilities



$N = 4000$ users, 200 active, Tx powers $5, 15, 25$dBm, cell radius 1000m. $L = 800$.

Multiple Antennas at the BS



- Multiple measurement vector (MMV) problem
  - Better performance than single measurement vector (SMV)
- Asymptotic analysis: Fix M, let $N, K, L \to \infty$, $\epsilon = \frac{K}{N}$, $\delta = \frac{L}{N}$,
- In theory, perfect user detection is possible when $M \to \infty$!
- In practice, the multi-antenna case is also more challenging:
  (i) Convergence is slower;
  (ii) Channel estimation is poor due to non-orthogonal pilots.

AMP Algorithm for the MIMO Case

- The AMP algorithm expressed in matrix form:

$$\mathbf{X}^{t+1} = \eta_t(\mathbf{S}^H\mathbf{R}^t + \mathbf{X}^t), \tag{13}$$

$$\mathbf{R}^{t+1} = \mathbf{Y} - \mathbf{S}\mathbf{X}^{t+1} + \frac{N}{L}\mathbf{R}^t\langle\eta_t'(\mathbf{S}^H\mathbf{R}^t + \mathbf{X}^t)\rangle, \tag{14}$$

  where
  - $\mathbf{X}^{t+1}$, estimate at iteration $t+1$;
  - $\mathbf{R}^{t+1}$, residual at iteration $t+1$;
  - $\eta_t(\cdot)$, a non-linear function known as denoiser that performs on each row
  - $\langle\cdot\rangle$, sample averaging operation
- Works well if M is fixed, and L, N, $K \to \infty$.
- Complexity: $\mathcal{O}(NLM)$ + complexity of $\eta_t(\cdot)$ per iteration

  When M is large: AMP becomes increasingly difficult to converge.

Quick Recap

AMP is a practical sparse user activities detection algorithm:

- State evolution provides accurate detector performance analysis.
- Denoiser should be designed to match channel characteristics.
- Detection becomes accurate with massive MIMO but convergence is slower.

Implications for network design:

- Non-orthogonal pilots should be used for massive random access.
- Massive MIMO needs to be deployed for good detection performance.
- AMP provides a channel estimation, but the estimation error can be large due to pilot contamination from *non-orthogonal pilots*.

Quick Recap

AMP is a practical sparse user activities detection algorithm:

- State evolution provides accurate detector performance analysis.
- Denoiser should be designed to match channel characteristics.
- Detection becomes accurate with massive MIMO but convergence is slower.

Implications for network design:

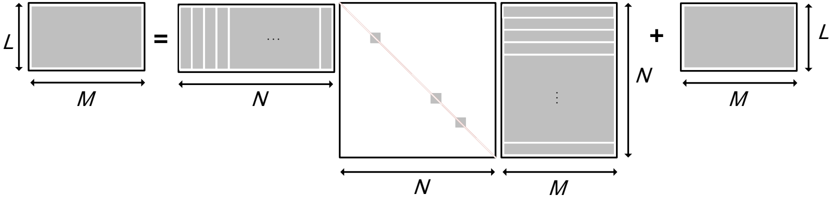
- Non-orthogonal pilots should be used for massive random access.
- Massive MIMO needs to be deployed for good detection performance.
- AMP provides a channel estimation, but the estimation error can be large due to pilot contamination from *non-orthogonal pilots*.

Can we do better?

Activity Detection Without Channel Estimation

Reformulate sparse activity detection as a large-scale-fading estimation problem:

$$\mathbf{Y} = \sum_{n=1}^{N} \alpha_n \mathbf{s}_n \mathbf{h}_n^T + \mathbf{Z} \triangleq \mathbf{S}\boldsymbol{\Gamma}^{\frac{1}{2}}\tilde{\mathbf{H}} + \mathbf{Z} \tag{15}$$

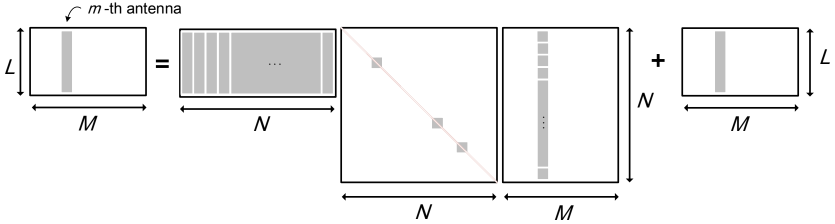


- $\mathbf{S} \triangleq [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_N] \in \mathbb{C}^{L\times N}$, signature matrix
- $\boldsymbol{\Gamma} \triangleq \operatorname{diag}\{\alpha_1\beta_1, \alpha_2\beta_2, \cdots, \alpha_N\beta_N\} \in \mathbb{R}^{N\times N}$, where β_n is large-scale fading
- $\tilde{\mathbf{H}} \triangleq \left[\mathbf{h}_1/\sqrt{\beta_1}, \mathbf{h}_2/\sqrt{\beta_2}, \cdots, \mathbf{h}_N/\sqrt{\beta_N}\right]^T \in \mathbb{C}^{N\times M}$, normalized channel

Statistics of the Received Signal

Let $\mathbf{y}_m$ be the received signal at the m-th antenna, and let $\tilde{\mathbf{h}}_m$ be the normalized channel and $\mathbf{z}_m$ be the noise. Then, $\mathbf{y}_m$ can be expressed as

$$\mathbf{y}_m = \mathbf{S}\mathbf{\Gamma}^{\frac{1}{2}}\tilde{\mathbf{h}}_m + \mathbf{z}_m \tag{16}$$



- Model: Small-scale fading is i.i.d. Rayleigh across M received antennas.
- Then, $\tilde{\mathbf{h}}_m$ follows $\mathcal{CN}(\mathbf{0}, \mathbf{I})$. Also, $\mathbf{z}_m$ follows $\mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$.
- Therefore, given $\mathbf{\Gamma}$, $\mathbf{y}_m$ is i.i.d. across m as $\mathcal{CN}(\mathbf{0}, \mathbf{\Sigma})$, $\mathbf{\Sigma} = \mathbf{S}\mathbf{\Gamma}\mathbf{S}^H + \sigma^2\mathbf{I}$.

Maximum Likelihood Estimate of Γ

The sparse device activity is included in the diagonal matrix $\mathbf{\Gamma}$, which can be estimated using the maximum likelihood estimation (MLE) as:

$$
\begin{aligned}
\min_{\mathbf{\Gamma}\geq\mathbf{0}} \quad f(\mathbf{\Gamma}) :&= -\frac{1}{M}\log p(\mathbf{Y}|\mathbf{\Gamma}) \qquad \longleftarrow \text{minimization of negative log-likelihood} \\
&= -\frac{1}{M}\sum_{m=1}^{M}\log p(\mathbf{y}_m|\mathbf{\Gamma}) \qquad \longleftarrow \text{i.i.d. over antennas} \\
&= -\frac{1}{M}\sum_{m=1}^{M}\log\left(\frac{1}{|\pi\mathbf{\Sigma}|}\exp\left(-\mathbf{y}_m^H\mathbf{\Sigma}^{-1}\mathbf{y}_m\right)\right) \\
&= -\frac{1}{M}\sum_{m=1}^{M}\log\left(\frac{1}{|\pi\mathbf{\Sigma}|}\right) - \frac{1}{M}\sum_{m=1}^{M}\log\left(\exp\left(-\mathbf{y}_m^H\mathbf{\Sigma}^{-1}\mathbf{y}_m\right)\right) \\
&= \log|\mathbf{\Sigma}| + \frac{1}{M}\sum_{m=1}^{M}\operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{y}_m\mathbf{y}_m^H\right) + const. \\
&= \log|\mathbf{\Sigma}| + \operatorname{tr}\left(\mathbf{\Sigma}^{-1}\frac{1}{M}\sum_{m=1}^{M}\mathbf{y}_m\mathbf{y}_m^H\right) + const. \qquad (17)
\end{aligned}
$$

Sample Covariance as a Sufficient Statistic

Define the sample covariance matrix of the received signal as

$$\hat{\boldsymbol{\Sigma}} \triangleq \frac{1}{M}\sum_{m=1}^{M} \mathbf{y}_m \mathbf{y}_m^H = \frac{1}{M}\mathbf{Y}\mathbf{Y}^H. \tag{18}$$

With the sample covariance matrix, the MLE of $\boldsymbol{\Gamma}$ can be expressed as

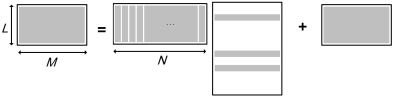
$$\begin{aligned} \min_{\boldsymbol{\Gamma}\geq \mathbf{0}} \quad f(\boldsymbol{\Gamma}) := & \log|\boldsymbol{\Sigma}| + \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\right) + const. \\ = & \log|\mathbf{S}\boldsymbol{\Gamma}\mathbf{S}^H + \sigma^2\mathbf{I}| + \operatorname{tr}\left((\mathbf{S}\boldsymbol{\Gamma}\mathbf{S}^H + \sigma^2\mathbf{I})^{-1}\hat{\boldsymbol{\Sigma}}\right) + const. \end{aligned} \tag{19}$$

- Ideally, we would like to match the covariance, i.e., make $\mathbf{S}\boldsymbol{\Gamma}\mathbf{S}^H + \sigma^2\mathbf{I} = \hat{\boldsymbol{\Sigma}}$.
- $\hat{\boldsymbol{\Sigma}}$ is computed by averaging over different antennas (not time slots).
- $\hat{\boldsymbol{\Sigma}}$ is a sufficient statistics since $f(\boldsymbol{\Gamma})$ depends on $\mathbf{Y}$ only through $\hat{\boldsymbol{\Sigma}}$.
- The size of the MLE problem depends on N, L only, not M.

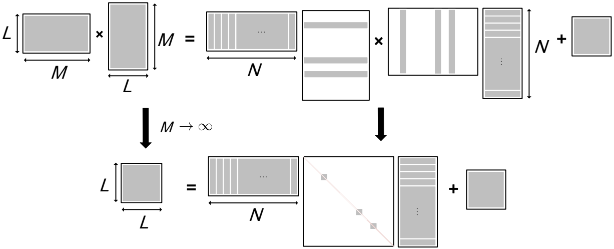
A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire: "Non-Bayesian Activity Detection, Large-Scale Fading Coefficient Estimation, and Unsourced Random Access with a Massive MIMO Receiver", *IEEE Trans. Inf. Theory*, May 2021.

Covariance Based Sparse Activity Detection

Instead of jointly estimating the channel, i.e., the non-zero rows in $\mathbf{X}$ based on $\mathbf{Y}$:



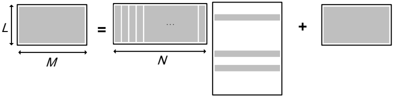
We now estimate large-scale fading $\mathbf{\Gamma}$ based on $\hat{\mathbf{\Sigma}} = \frac{1}{M}\mathbf{Y}\mathbf{Y}^H$:



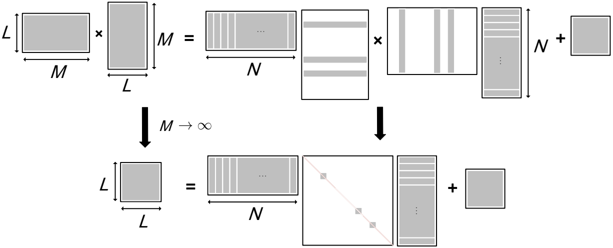
In the massive MIMO regime, i.e., if we let $M \to \infty$, this can be thought of detecting a diagonal sparse matrix from the sample covariance.

Covariance Based Sparse Activity Detection

Instead of jointly estimating the channel, i.e., the non-zero rows in $\mathbf{X}$ based on $\mathbf{Y}$:



We now estimate large-scale fading $\mathbf{\Gamma}$ based on $\hat{\mathbf{\Sigma}} = \frac{1}{M}\mathbf{Y}\mathbf{Y}^H$:



- AMP: Detect KM variables based on LM observations, so $K = O(L)$.
- Covariance: Detect K variables based on L^2 observations, so $K = O(L^2)$.

Coordinate Descent for Solving the MLE problem

Let γ_n be the n-th diagonal entry of $\mathbf{\Gamma}$. The MLE can be expressed as

$$\min_{\gamma_1,\ldots,\gamma_N \geq 0} \log\left|\sum_{n=1}^{N}\gamma_n \mathbf{s}_n\mathbf{s}_n^H + \sigma^2\mathbf{I}\right| + \operatorname{tr}\left(\left(\sum_{n=1}^{N}\gamma_n \mathbf{s}_n\mathbf{s}_n^H + \sigma^2\mathbf{I}\right)^{-1}\hat{\mathbf{\Sigma}}\right). \quad (20)$$

- **Basic Idea**: Update the coordinates $\gamma_1, \ldots, \gamma_N$ alternatively
- **Coordinate update**: Let $\hat{\gamma}_n, \forall n$ be the current estimates. Update $\hat{\gamma}_k$ with other $\hat{\gamma}_n, n \neq k$ fixed at a time. Let $\hat{\gamma}_k + d$ be the update. Determine d by

$$\min_{d \geq -\hat{\gamma}_k} \quad \log\left(1 + d\mathbf{s}_k^H\tilde{\mathbf{\Sigma}}^{-1}\mathbf{s}_k\right) - \frac{d\mathbf{s}_k^H\tilde{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Sigma}}\tilde{\mathbf{\Sigma}}^{-1}\mathbf{s}_k}{1 + d\mathbf{s}_k^H\tilde{\mathbf{\Sigma}}^{-1}\mathbf{s}_k}. \quad (21)$$

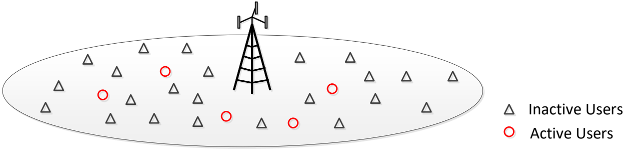
  - $\tilde{\mathbf{\Sigma}} = \sum_{n=1}^{N}\hat{\gamma}_n\mathbf{s}_n\mathbf{s}_n^H + \sigma^2\mathbf{I}$ is the current value of the covariance based on $\hat{\gamma}_n$.
  - The constraint $d \geq -\hat{\gamma}_k$ ensures the new $\hat{\gamma}_k + d$ is always non-negative.
  - By taking the derivative of the objective in (21), a closed-form solution is

$$d = \max\left\{\frac{\mathbf{s}_k^H\tilde{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Sigma}}\tilde{\mathbf{\Sigma}}^{-1}\mathbf{s}_k - \mathbf{s}_k^H\tilde{\mathbf{\Sigma}}^{-1}\mathbf{s}_k}{(\mathbf{s}_k^H\tilde{\mathbf{\Sigma}}^{-1}\mathbf{s}_k)^2}, -\hat{\gamma}_k\right\}. \quad (22)$$

- **Advantages**: Efficient due to closed-form solution; empirically performs well.

AMP vs. Covariance Method

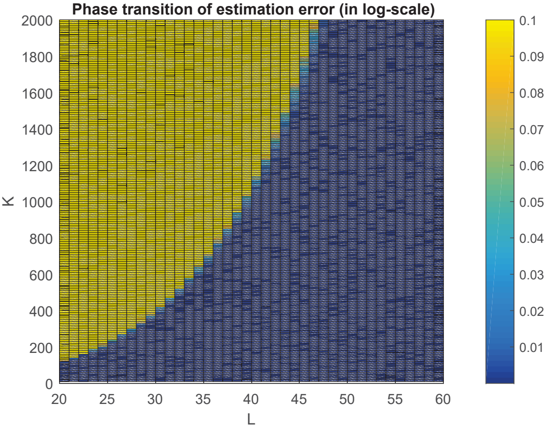
- Sparse user activity detection with channel $\alpha_n \mathbf{h}_n \sim \alpha_n \sqrt{\beta}_n \mathcal{CN}(\mathbf{0}, \mathbf{I})$:



- If channel estimate is needed for subsequent data transmission:
  - We can use AMP, which gives a rough estimate of the instantaneous $\mathbf{h}_n$.
- If only user activities (α_n) are needed and large-scale fading is not known:
  - We can estimate large-scale fading $(\alpha_n \beta_n)$ using the covariance method.
- Computational complexity
  - AMP is more efficient when $K < L$ and for small M.
  - Covariance method is more effective in exploiting large M.

Scaling Law of the Covariance Approach

Assume high SNR and perfect sampled covariance matrix $\hat{\boldsymbol{\Sigma}}$ ($M \to \infty$), we plot the estimation error of $\boldsymbol{\Gamma}$ under different (K, L) with $N = 2000$



Analysis of Covariance Approach via Fisher Info Matrix

Recall the MLE formulation, and let γ denote the diagonal entries of $\mathbf{\Gamma}$

$$\begin{aligned}\min_{\boldsymbol{\gamma}\geq\mathbf{0}} \quad f(\boldsymbol{\gamma}) :&= -\frac{1}{M}\log p(\mathbf{Y}|\boldsymbol{\gamma}) = -\frac{1}{M}\sum_{m=1}^{M}\log p(\mathbf{y}_m|\boldsymbol{\gamma}) \\ &= \log|\mathbf{S}\mathbf{\Gamma}\mathbf{S}^H+\sigma^2\mathbf{I}| + \mathrm{tr}\left((\mathbf{S}\mathbf{\Gamma}\mathbf{S}^H+\sigma^2\mathbf{I})^{-1}\hat{\mathbf{\Sigma}}\right) + const. \end{aligned} \tag{23}$$

- Analyzing the solution to (23) under coordinate descent is hard.
- Instead, let's analyze the true optimum of (23), i.e., MLE solution $\hat{\boldsymbol{\gamma}}^{ML}$.
- Investigate asymptotic property of $\hat{\boldsymbol{\gamma}}^{ML}$ in the massive MIMO regime.
- The Fisher information matrix, denoted by $\mathbf{J}(\boldsymbol{\gamma})$, plays a critical role in the asymptotic analysis. The (i,j)-th entry of $\mathbf{J}(\boldsymbol{\gamma})$ is defined as

$$[\mathbf{J}(\boldsymbol{\gamma})]_{ij} = \mathbb{E}\left[\frac{\partial\log p(\mathbf{Y}|\boldsymbol{\gamma})}{\partial\gamma_i}\frac{\partial\log p(\mathbf{Y}|\boldsymbol{\gamma})}{\partial\gamma_j}\right]. \tag{24}$$

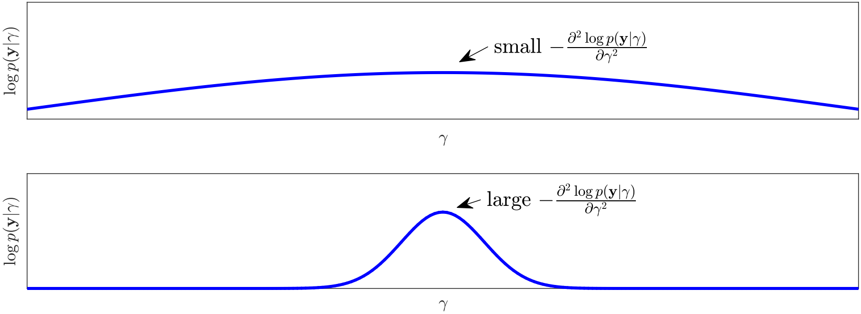
- Key assumption for the analysis: $M\to\infty$.

Fisher Information Matrix

- The Fisher Information matrix can be also written as the negative expected second derivative of the log-likelihood function

$$[\mathbf{J}(\boldsymbol{\gamma})]_{ij} = \mathbb{E}\left[\frac{\partial \log p(\mathbf{Y}|\boldsymbol{\gamma})}{\partial \gamma_i}\frac{\partial \log p(\mathbf{Y}|\boldsymbol{\gamma})}{\partial \gamma_j}\right] = -\mathbb{E}\left[\frac{\partial^2 \log p(\mathbf{Y}|\boldsymbol{\gamma})}{\partial \gamma_i \partial \gamma_j}\right] \qquad (25)$$

- **Intuition**: Fisher information matrix measures how informative the likelihood function is, and how effective the MLE can be



Cramér-Rao Bound and Asymptotic Property of MLE

Fisher information matrix plays a critical role in classic estimation theory.

- **Cramér-Rao bound**: Let $\boldsymbol{\gamma}$ be a parameter, and let $\hat{\boldsymbol{\gamma}}$ be an unbiased estimator of $\boldsymbol{\gamma}$. Then the covariance of estimation error is lower bounded by

$$\mathbb{E}\left[(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})^T\right] \geq \mathbf{J}^{-1}(\boldsymbol{\gamma}) \tag{26}$$

- **Asymptotic properties of the MLE**: Let $\hat{\boldsymbol{\gamma}}^{ML}$ be the maximum likelihood estimator of $\boldsymbol{\gamma}$. Then, under certain regularity conditions, as $M \to \infty$

$$\text{Consistency:} \qquad \hat{\boldsymbol{\gamma}}^{ML} \xrightarrow{P} \boldsymbol{\gamma} \tag{27}$$

$$\text{Asymptotic normality:} \qquad \sqrt{M}(\hat{\boldsymbol{\gamma}}^{ML}-\boldsymbol{\gamma}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, M\mathbf{J}^{-1}(\boldsymbol{\gamma})) \tag{28}$$

It means that the maximum likelihood estimator $\hat{\boldsymbol{\gamma}}^{ML}$ is asymptotically unbiased and asymptotically attains the Cramér-Rao bound, i.e., asymptotically efficient.

Regularity Conditions

- The regularity conditions for consistency and asymptotic normality include
  - The true parameter $\boldsymbol{\gamma}^0$ is identifiable, i.e.,, there exists no other $\boldsymbol{\gamma}' \neq \boldsymbol{\gamma}^0$ with

$$p(\mathbf{Y}|\boldsymbol{\gamma}^0) = p(\mathbf{Y}|\boldsymbol{\gamma}'). \tag{29}$$

  - The true parameter should be in the interior of the feasible region, as otherwise $\hat{\boldsymbol{\gamma}}^{ML} - \boldsymbol{\gamma}^0$ cannot be Gaussian distributed.
- These conditions are usually reasonable and mild.
- But, these conditions are NOT always satisfied for sparse activity detection.
  - The identifiability may not be guaranteed when

$$N \gg L^2, \tag{30}$$

    i.e., when the dimension of $\boldsymbol{\gamma}^0$ is larger than the dimensions of the sample covariance $\hat{\boldsymbol{\Sigma}}$, there are too many parameters to estimate.
  - The true parameter $\boldsymbol{\gamma}^0$ in fact always lies on the boundary of its parameter space $[0, \infty)^N$, because most of the entries of $\boldsymbol{\gamma}^0$ are zero.

We can understand the phase transition based on these regularity conditions!

Necessary and Sufficient Condition for $\hat{\boldsymbol{\gamma}}^{ML} \to \boldsymbol{\gamma}^0$

Theorem

Let $\mathcal{I}$ be an index set corresponding to zero entries of $\boldsymbol{\gamma}^0$, i.e., $\mathcal{I} \triangleq \{i \mid \gamma_i^0 = 0\}$. We define two sets $\mathcal{N}, \mathcal{C}$ in the space $\mathbb{R}^N$, respectively, as follows

$$\mathcal{N} \triangleq \{\mathbf{x} \mid \mathbf{x}^T \mathbf{J}(\boldsymbol{\gamma}^0)\mathbf{x} = 0, \mathbf{x} \in \mathbb{R}^N\}, \tag{31}$$

$$\mathcal{C} \triangleq \{\mathbf{x} \mid x_i \geq 0, i \in \mathcal{I}, \mathbf{x} \in \mathbb{R}^N\}, \tag{32}$$

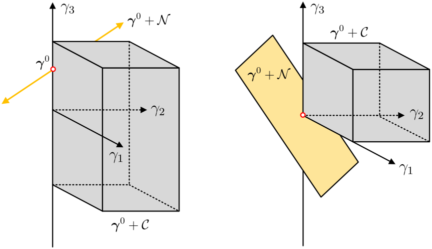
where x_i is the i-th entry of $\mathbf{x}$. Then a necessary and sufficient condition for the consistency of $\hat{\boldsymbol{\gamma}}^{ML}$, i.e., $\hat{\boldsymbol{\gamma}}^{ML} \to \boldsymbol{\gamma}^0$ as $M \to \infty$, is $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$.

$\mathcal{N}$ is the "null space" of $\mathbf{J}(\boldsymbol{\gamma}^0)$; $\mathcal{C}$ is a cone with non-negative entries indexed by $\mathcal{I}$.

This condition leads to a numerical characterization of phase transition for the covariance method, i.e., the set of (N, L, K) outside of which $\hat{\boldsymbol{\gamma}}^{ML}$ cannot approach $\boldsymbol{\gamma}^0$ even in large M limit.

Interpretation of the Condition

- $\mathcal{N}$ corresponds to all directions in $\mathbb{R}^N$ along which likelihood stays constant. In these directions, the true parameter cannot be identified via the likelihood.
- $\mathcal{C}$ is the directions along which parameters remain within constraint $\mathbb{R}_+^N$
- $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ ensures that the true parameter $\boldsymbol{\gamma}^0$ is uniquely identifiable via the likelihood in its feasible neighborhood, also termed as local identifiability



γ_3
$\boldsymbol{\gamma}^0 + \mathcal{N}$
$\boldsymbol{\gamma}^0$
γ_2
γ_1
$\boldsymbol{\gamma}^0 + \mathcal{C}$
γ_3
$\boldsymbol{\gamma}^0 + \mathcal{C}$
$\boldsymbol{\gamma}^0 + \mathcal{N}$
γ_2
γ_1

- Local identifiability is clearly necessary.
- Sufficiency due to equivalence of local and global identifiability in this case.
- A necessary condition for $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ is that $dim(\mathcal{N}) < |\mathcal{I}|$.
- Since $dim(\mathcal{N})$ is roughly $N - L^2$ and $|\mathcal{I}| = N - K$, we have $K < L^2$.

Scaling Law for MLE Formulation

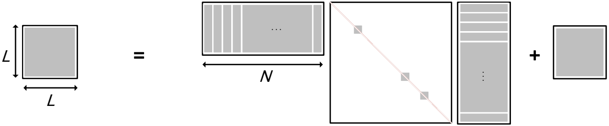
- Scaling laws in compressed sensing:
  - For $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A}$ satisfying restricted isometry property, the number of measurements needed to recover a K-sparse vector $\mathbf{x}$ of length-N is

$$L = O(K \log(N/K)).$$

  - For $\hat{\mathbf{\Sigma}} = \mathbf{S\Gamma S}^H + \sigma^2 \mathbf{I}$ with $\widehat{\mathbf{S}}$ (the Khatri-Rao product of $\mathbf{S}$) satisfying robust null space property, the number of measurements needed to recover a K-sparse diagonal matrix $\mathbf{\Gamma}$ of size $N \times N$ is

$$L^2 = O(K \log^2(N/K)).$$

- The condition $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ is closely related to robust null space property:



This allows us to show that the consistency of MLE has the same scaling law

$$L^2 = O(K \log^2(N/K)).$$

Numerical Results – Scaling Law of Covariance Approach

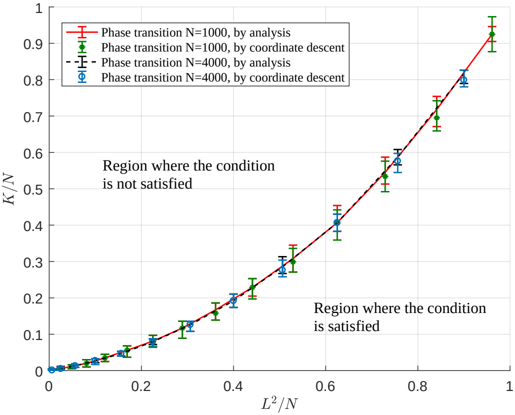


Figure: Phase transition in the space of N, L, K generated by 100 Monte Carlo simulations. All users are located at the cell-edge (1000m) with transmit power 23dBm. Path-loss is $128.1 + 37.6\log(d[\text{km}])$. Error bars indicate the range below which all 100 realizations satisfy the condition and above which none satisfies the condition.

Recap

Covariance based method is very efficient for user activity detection

- Phase transition analysis shows that $K = O(L^2)$ users can be detected.
- It does not estimate the exact channel and only the large-scale fading.

Implications of covariance method for network design:

- Non-orthogonal pilots should be used for massive random access.
- Massive MIMO is essential for detecting $K = O(L^2)$ active users.
- Channel estimation needs to be performed in subsequent stage.
- Feedback from BS to the active users is needed to coordinate scheduling and channel estimation.

Recap

Covariance based method is very efficient for user activity detection

- Phase transition analysis shows that $K = O(L^2)$ users can be detected.
- It does not estimate the exact channel and only the large-scale fading.

Implications of covariance method for network design:

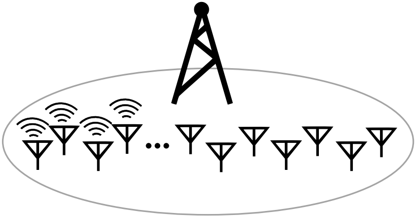
- Non-orthogonal pilots should be used for massive random access.
- Massive MIMO is essential for detecting $K = O(L^2)$ active users.
- Channel estimation needs to be performed in subsequent stage.
- Feedback from BS to the active users is needed to coordinate scheduling and channel estimation.

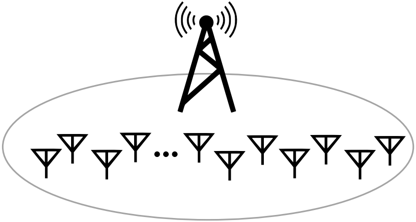
How to design efficient feedback strategy?

Feedback for Collision-Free Scheduling

Feedback-Based Scheduling for Random Access

Each of N potential users is assigned a unique non-orthogonal pilot.



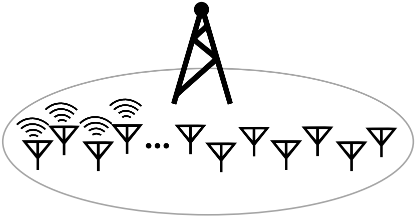


Phase 1 (Activity Detection): The K active users ($K \ll N$) send their pilots synchronously to the BS.

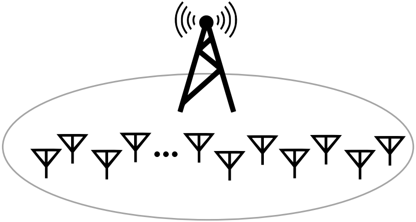
Phase 2 (Downlink Feedback): BS sends a *common* feedback message to schedule the data transmissions of K active users.

Feedback-Based Scheduling for Random Access

Each of N potential users is assigned a unique non-orthogonal pilot.



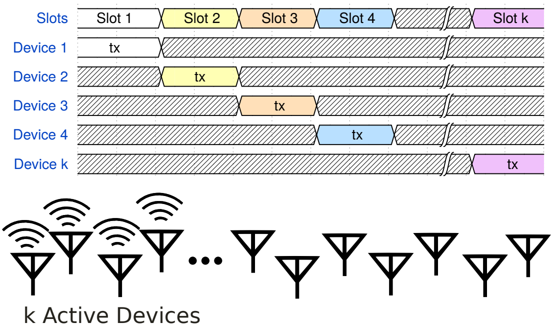
Phase 1 (Activity Detection):
The K active users ($K \ll N$) send their pilots synchronously to the BS.



Phase 2 (Downlink Feedback): BS sends a *common* feedback message to schedule the data transmissions of K active users.

What is the minimum feedback needed to ensure collision-free scheduling?

Feedback-Based Scheduling for Random Access



Phase 3 (Uplink Payload Transmission): The K active users transmit their payload in the K slots based on the schedule provided by the BS, while avoiding collision.

Straightforward Feedback Scheme

- A naive scheme to schedule K out of N users:
  - Assign a unique index to each of the N users;
  - The BS detects the K active users based on the pilots;
  - The BS lists the K users in the order in which they should transmit;
  - Each active user finds its index in the list, waits for its turn to transmit.
- The feedback overhead of this scheme is $K \log(N)$ bits.
  - When $N = 10^6$, the cost of identification is $\log(N) = 20$ bits per user.

Straightforward Feedback Scheme

- A naive scheme to schedule K out of N users:
  - Assign a unique index to each of the N users;
  - The BS detects the K active users based on the pilots;
  - The BS lists the K users in the order in which they should transmit;
  - Each active user finds its index in the list, waits for its turn to transmit.
- The feedback overhead of this scheme is $K \log(N)$ bits.
  - When $N = 10^6$, the cost of identification is $\log(N) = 20$ bits per user.

Can we do better?

Why Can We Do Better?

- The naive $K \log(N)$ feedback scheme is not optimal.
- There is flexibility in the order that users are scheduled.

  Example: Users $1, \ldots, K$ are to be scheduled. The BS can schedule according to any of the $K!$ permutations of these users, e.g. $\{1, \ldots, K\}$ or $\{K, \ldots, 1\}$.

  We can remove this extraneous cost via *enumerative source coding*.

  This still requires $\log \binom{N}{K}$ bits feedback, which scales as $O(\log(N))$.
- Each user only needs to know its own slot, and NOT the other users' slots. Removing this extraneous information is the key to further reducing feedback.

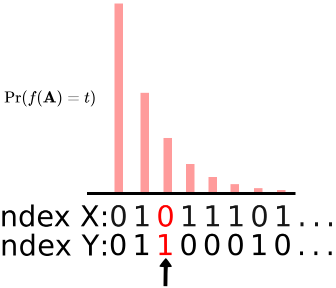
G. K. Facenda and D. Silva, "Efficient Scheduling for the Massive Random Access Gaussian Channel," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7598–7609, Aug. 2020.

Two-User Case

- Consider the index based feedback strategy for $K = 2$. Choose the first position where user indices differ.

- To code the location of any index using fixed-length code requires only $O(\log\log(N))$ bits.

- The location of the first index follows a geometric distribution.



Index X:0 1 0 1 1 1 0 1 . . .
Index Y:0 1 1 0 0 0 1 0 . . .

- The entropy of this truncated geometric distribution is:

$$R_v(N, 2) = 2 - \frac{\log(N) + 1}{N - 1}.$$

- This implies $\lim\limits_{N\to\infty} R_v^*(N, 2) \le 2$, thus if we can use variable rate code, the achievable feedback rate is bounded by 2 as N tends to infinity.

Fundamental Limits

- The problem of finding the optimal feedback fixed-rate code is related to perfect hashing or hypergraph covering [Fredman-Komlós'84, Körner-Marton'88] based on which we can show that $O(\log(\log(N)))$ scaling is optimal.

- For variable-rate code, we can prove that scaling is independent of N.

Theorem

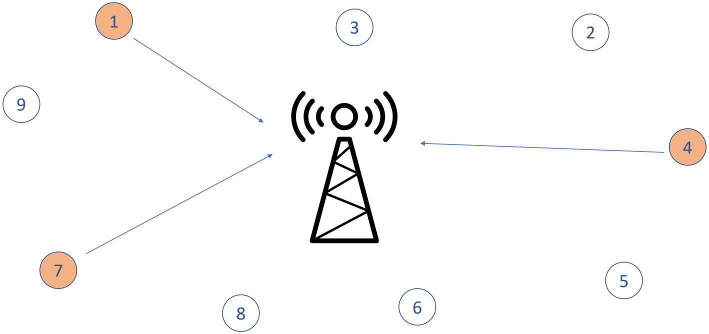
The minimum rate for variable-length collision-free feedback code for scheduling K users out of a pool of N users is bounded above and below as

$$(K+1)\log(e) \geq R_v^*(N,K) \geq K\log(e) - \log\left(\frac{N^K}{N^{\underline{K}}}\right) - \frac{1}{2}\log(2\pi K) - \frac{\log(e)}{12K}.$$

Justin Kang and Wei Yu, "Minimum Feedback for Collision-Free Scheduling in Massive Random Access", *IEEE Transactions on Information Theory*, vol. 67, no. 12, Dec. 2021.

Example

- Total of $N = 9$ users; $K = 3$ are active. Set of active users is $\{1, 4, 7\}$.
- BS learns identities of active users via activity detection.



Code Construction

- Construct an infinite codebook of random schedules.
- The BS broadcasts the index of the first codeword such that the active users have a non-colliding schedule:

		1	2	3	4	5	6	7	8	9	...
	1	2	3	1	3	1	1	2	1	1	...
	2	3	3	1	3	1	3	3	1	1	...
	3	1	2	3	3	2	1	3	1	3	...
	4	1	3	3	1	2	1	3	3	1	...
Users	5	2	1	2	1	3	1	1	2	2	...
	6	3	1	2	1	3	2	3	1	3	...
	7	1	1	3	2	3	1	2	1	1	...
	8	2	2	1	1	1	2	1	1	2	...
	9	2	3	1	1	1	2	1	2	1	...

→ Broadcast T = 4

Analysis of Rate

- The probability that a random schedule of K users happens to be non-colliding

$$p = \left(\frac{K}{K}\right)\left(\frac{K-1}{K}\right)\cdots\left(\frac{1}{K}\right) = \frac{K!}{K^K}$$

- The index of the first codeword with non-colliding schedule for a given set of K active users is geometric distributed with parameter p:

$$\Pr(T = t) = (1-p)^{t-1}p.$$

- Entropy of a geometric random variable with small p is approximately:

$$H(T) \approx \log\left(\frac{1}{p}\right) = \log\left(\frac{K^K}{K!}\right) \approx K\log(e).$$

- The rate has no dependence on N. The key is to exploit the fact that
  1. Only the active users are listening to the feedback message.
  2. Each active user only needs to learn their own scheduling slot.

.

Cost of Coordinating Collision-Free Scheduling

- Fixed-length feedback codes for collision-free scheduling of K active users among N potential users into K slots requires a rate of approximately $K \log(e)$ bits, plus a $\Theta(\log \log(N))$ term.

- Using variable-length feedback codes can reduce the required feedback rate for collision-free scheduling to $(K+1)\log(e)$ bits, independent of N.

Therefore 1.44 feedback bits per user is sufficient
to ensure collision-free scheduling!

- This strategy can be generalized to the scenario of transmitting exchangeable sources $\{X_i\}_{i=1}^{K}$ to a randomly activated set of K users out of N potential users (e.g., acknowledgement) using $H(X_1, \cdots, X_K) + O(K)$ bits.

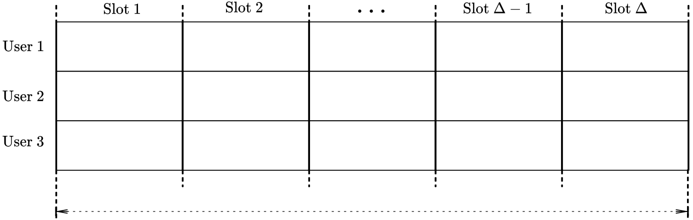
Coordinated vs Uncoordinated Massive Random Access

Random Access for Massive MIMO Systems

1. Uncoordinated Random Access for Massive MIMO
   - Channel estimation and data transmission must both be without coordination.
   - Coded ALOHA can be adapted to Massive MIMO systems to enable uncoordinated communication.
   - We consider a variant of coded ALOHA known as *Coded Pilot Access* (CPA).

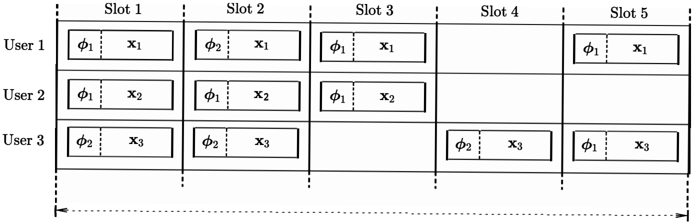
2. Scheduled Random Access (SRA) for Massive MIMO
   - Activity detection can serve as an initial step for scheduled random access.
   - A relatively small amount of feedback can be used to ensure collision-free scheduling for the users.
   - Users are assigned orthogonal pilots for channel estimation.

Slotted Random Access



- The BS is equipped with M antennas.
- There are N single-antenna devices K of which are active.
- Active users transmit across Δ temporal slots each containing L symbols.
- The channels $\mathbf{h}_{d,i} \sim \mathcal{CN}(0,1)$ is i.i.d for each user i in the d^{th} slot. We assume users apply inverse power control to compensate for large scale fading.
- The BS uses the received signal $\mathbf{Y}_d$ over Δ slots to decode the messages of K active users.

Coded Pilot Access

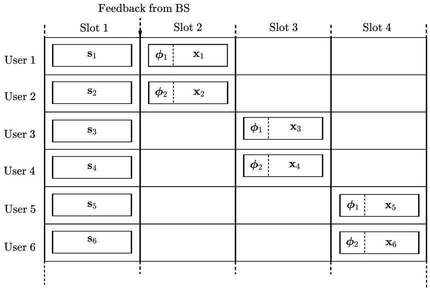


- Users transmit their payload $\mathbf{x}_i$ multiple times, each time preceded by a pilot randomly selected from a set of orthogonal pilots $\{\phi_t\}_{t=1}^{\tau}$.
- In cases with no collision, the BS can perform channel estimation and data decoding for that user.
- The data contains the location of the other slots where the user has transmitted, allowing the BS to perform SIC.

J. H. Sørensen, E. De Carvalho, Č. Stefanović, and P. Popovski, "Coded Pilot Random Access for Massive MIMO Systems", *IEEE Trans. Wireless Commun.*, vol.17, no.12, pp.8035–8046, 2018.

Scheduled Random Access for Massive MIMO

- Users first transmit non-orthogonal pilots $\mathbf{s}_i \in \mathbb{C}^L$ for activity detection.
- BS sends scheduling message.
- Each user is assigned a unique (slot, orthogonal pilot) pair based on common feedback from the BS.



Feedback from BS

	Slot 1	Slot 2	Slot 3	Slot 4
User 1	$\mathbf{s}_1$	$\boldsymbol{\phi}_1$ $\mathbf{x}_1$		
User 2	$\mathbf{s}_2$	$\boldsymbol{\phi}_2$ $\mathbf{x}_2$		
User 3	$\mathbf{s}_3$		$\boldsymbol{\phi}_1$ $\mathbf{x}_3$	
User 4	$\mathbf{s}_4$		$\boldsymbol{\phi}_2$ $\mathbf{x}_4$	
User 5	$\mathbf{s}_5$			$\boldsymbol{\phi}_1$ $\mathbf{x}_5$
User 6	$\mathbf{s}_6$			$\boldsymbol{\phi}_2$ $\mathbf{x}_6$

- The BS performs channel estimation using the orthogonal pilots, and then maximum ratio combining to reconstruct the payload.
- Each user is only required to transmit twice, in contrast to Coded ALOHA.

Scheduled Random Access vs. Coded Pilot Access

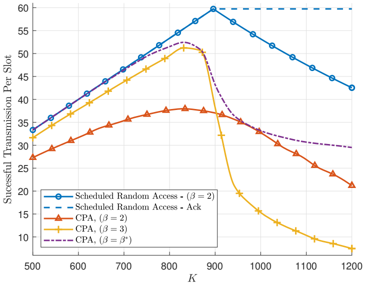


Figure: Coded Pilot Access (CPA) vs. Scheduled Random Access (SRA) with $N = 10000$, SNR $= 10$dB, $M = 400$ BS antennas, $\tau = 64$ orthogonal pilots

- Each slot consists of $L = 300$ symbols.
- Number of slots $\Delta = 15$, so can schedule up to $\tau(\Delta - 1) = 896$ users.
- Activity detection is done via covariance method over one slot for SRA.
- CPA performs comparably to SRA, if $\beta = 3$, which costs 50% more power.
- The performances of SRA is more stable than CPA, if K is close to fully loaded.
- Cost of feedback is 1.3 kbits, which is $< 1\%$ of typical payload.

Conclusions

- Massive device connectivity is a key use case in 5G/6G research agenda.

Conclusions

- Massive device connectivity is a key use case in 5G/6G research agenda.
- Classic contention-based random access does not utilize resources efficiently.
  - Coded random access can alleviate some of the loss due to collision.

Conclusions

- Massive device connectivity is a key use case in 5G/6G research agenda.
- Classic contention-based random access does not utilize resources efficiently.
    - Coded random access can alleviate some of the loss due to collision.
- Theoretical results on coordinated random access:
    - AMP for joint user activity detection and channel estimation with $K = O(L)$.
    - Covariance method for sparse activity detection with $K = O(L^2)$.
    - Feedback for collision-free scheduling using $K \log(e)$ bits independent of N.

Conclusions

- Massive device connectivity is a key use case in 5G/6G research agenda.
- Classic contention-based random access does not utilize resources efficiently.
  - Coded random access can alleviate some of the loss due to collision.
- Theoretical results on coordinated random access:
  - AMP for joint user activity detection and channel estimation with $K = O(L)$.
  - Covariance method for sparse activity detection with $K = O(L^2)$.
  - Feedback for collision-free scheduling using $K \log(e)$ bits independent of N.
- Significant performance improvement can be obtained at moderate feedback of 1.44 bits/user for collision-free scheduling.

Conclusions

- Massive device connectivity is a key use case in 5G/6G research agenda.
- Classic contention-based random access does not utilize resources efficiently.
  - Coded random access can alleviate some of the loss due to collision.
- Theoretical results on coordinated random access:
  - AMP for joint user activity detection and channel estimation with $K = O(L)$.
  - Covariance method for sparse activity detection with $K = O(L^2)$.
  - Feedback for collision-free scheduling using $K \log(e)$ bits independent of N.
- Significant performance improvement can be obtained at moderate feedback of 1.44 bits/user for collision-free scheduling.

Scheduled Random Access as a Viable Option for Massive Connectivity!

Further Information

- Zhilin Chen, Foad Sohrabi, and Wei Yu,
  "Sparse Activity Detection for Massive Connectivity",
  IEEE Transactions on Signal Processing, vol. 66, no. 7, April 2018.

- Liang Liu and Wei Yu,
  "Massive Connectivity with Massive MIMO – Part I: Device Activity Detection and Channel Estimation and Part II: Achievable Rate Characterization",
  IEEE Transactions on Signal Processing, vol. 66, no. 11, June 2018.

- Zhilin Chen, Foad Sohrabi, Ya-Feng Liu, Wei Yu,
  "Phase Transition Analysis for Covariance Based Massive Random Access with Massive MIMO",
  IEEE Transactions on Information Theory, vol. 68, no. 3, March 2022.

- Justin Kang and Wei Yu,
  "Minimum Feedback for Collision-Free Scheduling in Massive Random Access",
  IEEE Transactions on Information Theory, vol. 67, no. 12, Dec. 2021.

- Justin Kang and Wei Yu,
  "Scheduling vs. Contention for Massive Random Access in Massive MIMO Systems"
  IEEE Transactions on Communications, vol. 70, no. 9, Sept. 2022.