Massive Random Access with Massive MIMO: Sparse Activity Detection

Wei Yu

Joint Work with Zhilin Chen, Foad Sohrabi, Liang Liu

University of Toronto

2021

Introduction and Outline

Massive device connectivity is a key requirement for 5G 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

This talk is about how to design such a network:

- Sparsity device activity detection algorithms
- Massive connectivity with massive MIMO
- Scheduling and feedback in massive random access

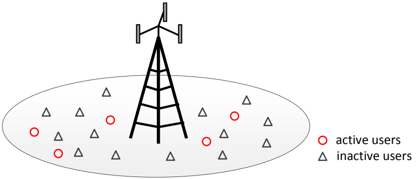
Main Messages

To support massive connectivity:

- The use of non-orthogonal pilots is inevitable.
- Compressed sensing techniques are indispensible for device detection.
- Massive MIMO can significantly enhance device activity detection.
- Channel estimation is the main bottleneck.
- Cooperative detection across multiple cells further improves performance.
- Scheduing and feedback are superior to uncoordinated random access.

Massive Random Access

- Cellular system with N users, but only K of which are active.



- BS needs to detect which users are active, then their messages.
- User activity pattern carries information. [Chen-Guo'14]

$$R + H(A) \leq I(X; Y) \tag{1}$$

- We also need to take the cost of channel estimation into account.

Fundamental Limit of Massive Random Access

For a massive device communications scenario $Y = HAX + Z$, the achievable sum rate of data transmission across all the users is approximately bounded by

$$R \lesssim I(X; Y|HA) - H(A) - I(HA; Y|X). \tag{2}$$

Interpretation:

- $I(X; Y|HA)$: Transmission rate with known channels and activity pattern;
- $H(A)$: Information content of device activity pattern;
- $I(HA; Y|X)$: Channel estimation and user activity detection.

Why? We see that $R + H(A) \leq I(X; Y)$.

$$\begin{aligned} I(HA, X; Y) &= I(X; Y) + I(HA; Y|X) && (3) \\ &= I(HA; Y) + I(X; Y|HA) && (4) \end{aligned}$$

Note that the $I(HA; Y)$ term is negligible.

Cost of User Activity Detection

Traditional MIMO system [Zheng-Tse'02, Lozano-Heath-Andrew'12]:

$$R \lesssim I(X;Y|H) - I(H;Y|X). \tag{5}$$

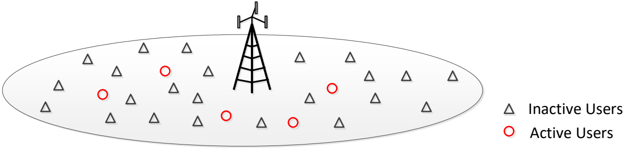
Massive connectivity system:

$$R \lesssim I(X;Y|HA) - H(A) - I(HA;Y|X). \tag{6}$$

where the cost of user activity detection is:

$$H(A) = Nh\left(\frac{K}{N}\right) \approx \log\begin{pmatrix} N \\ K \end{pmatrix} \approx K\log(N/K). \tag{7}$$

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 $\boldsymbol{s}_n$
- Channel $\boldsymbol{h}_n$ of user n is assumed to be fixed during the L symbols.
- For single-cell system, received signal $Y \in \mathbb{C}^{L \times M}$ at the BS is

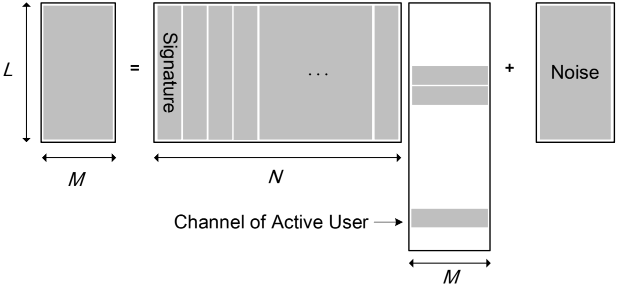
$$\boldsymbol{Y} = \sum_{n=1}^{N} \alpha_n \boldsymbol{s}_n \boldsymbol{h}_n^T + \boldsymbol{Z} = \boldsymbol{S}\boldsymbol{X} + \boldsymbol{Z}, \tag{8}$$

where

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

User Activity Detection via Compressed Sensing

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



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

Practical Detector Design

Device Identification via Non-Orthogonal Pilots:

- Due to large number of potential devices N, orthogonal pilot is not feasible.
- Natural choice of pilot sequences: i.i.d. Gaussian signature
- Approximate Message Passing (AMP) [Donoho-Maleki-Montanari'09]

Prior work on compressed sensing for massive connectivity:

- Without channel estimation [Fletcher-Rangan-Goyal'09, Zhang-Luo-Guo'13]
- Joint user activity detection and channel estimation: Orthogonal matching pursuit [Schepker-Bockelmann-Dekorsy'13, Wunder-Jung-Ramadan'15, Wunder-Boche-Strohmer-Jung'15], Baysian [Xu-Rao-Lau'15]
- AMP is used for device detection in [Hannak-Mayer-Jung-Matz-Goertz'15].

Single-Antenna Case

Single-Antenna Case

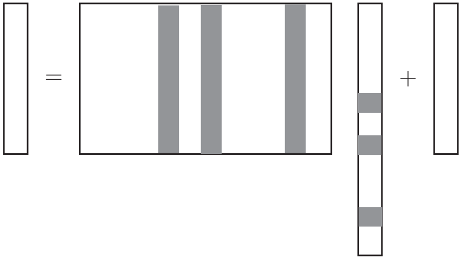
A single-antenna BS, N devices randomly located in a cell of radius R,

$$\mathbf{y} = \sum_{n=1}^{N} h_n \alpha_n \mathbf{s}_n + \mathbf{w} \triangleq \mathbf{S}\mathbf{x} + \mathbf{z} \tag{9}$$

- $h_n \in \mathbb{C}$: channel coefficient between user n and BS, including path-loss fading, shadowing and Rayleigh fading static within each block;
- $\alpha_n \in \{1, 0\}$: indicating whether user n is active
- $\mathbf{x} \triangleq [h_1\alpha_1, h_2\alpha_2, \cdots, h_N\alpha_N]^T \in \mathbb{C}^{N\times 1}$
- $\mathbf{s}_n \in \mathbb{C}^{L\times 1}$: signature sequence of user n generated as i.i.d. $\mathcal{CN}(0, 1/L)$
- $\mathbf{S} \triangleq [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_N]^T \in \mathbb{C}^{L\times N}$
- $\mathbf{z} \in \mathbb{C}^{L\times 1}$: effective noise following i.i.d. $\mathcal{CN}(0, \sigma^2)$

Sparse Recovery Problem

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



LASSO formulation:

$$\hat{\boldsymbol{x}} = \arg\min \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{S}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 \tag{10}$$

CoSaMP is computationally complex: Not scalable at $N = 10^5$.

Soft Thresholding Function

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

$$\hat{x} = \arg\min \frac{1}{2}|y - x|_2^2 + \lambda|x|_1 \tag{11}$$

The solution is explicitly given by

$$\hat{x} = \eta(y; \lambda), \tag{12}$$

where η is a soft thresholding function as

$$\eta(y;\theta) = \begin{cases} y - \theta, & y > \theta \\ 0, & -\theta \leq y \leq \theta \\ y + \theta, & y < -\theta \end{cases} \tag{13}$$

Soft Thresholding Function

This denoiser is nearly minimax optimal:

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

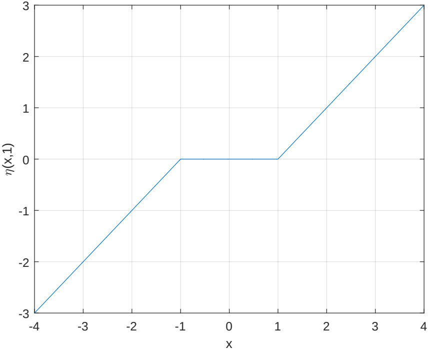
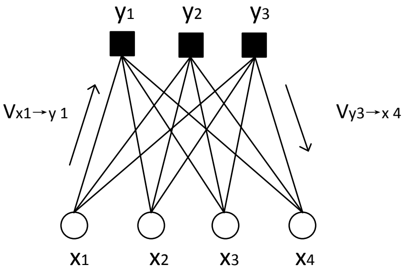


Figure: Soft thresholding function with $\theta = 1$

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, denoise, then iterate with the residual

$$\boldsymbol{x}^{t+1} = \eta(\boldsymbol{x}^t + \boldsymbol{S}^T \boldsymbol{r}^t; \lambda + \gamma^t) \tag{14}$$

$$\boldsymbol{r}^t = \boldsymbol{y} - \boldsymbol{S}\boldsymbol{x}^t + \frac{1}{L}\boldsymbol{r}^{t-1}\|\boldsymbol{x}^t\|_0, \tag{15}$$

where the threshold satisfies

$$\gamma^{t+1} = \frac{\lambda + \gamma^t}{L}\|\boldsymbol{x}^{t+1}\|_0 \tag{16}$$

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

Without the last "Onsager term", this is the classical iterative soft thresholding.

AMP Algorithm – General Form

Recover $\boldsymbol{x}$ from $\boldsymbol{y}$ via AMP algorithm (complex case)

$$\boldsymbol{x}^{t+1} = \eta_t(\boldsymbol{S}^*\boldsymbol{r}^t + \boldsymbol{x}^t)$$
$$\boldsymbol{r}^{t+1} = \boldsymbol{y} - \boldsymbol{S}\boldsymbol{x}^{t+1} + \frac{\boldsymbol{r}^t}{\delta}\langle \eta_t'(\boldsymbol{S}^*\boldsymbol{r}^t + \boldsymbol{x}^t)\rangle$$

- $\boldsymbol{x}^t$: estimate of $\boldsymbol{x}$ at iteration t
- $\boldsymbol{r}^t$: residual at iteration t
- $\eta_t(\cdot)$: for soft thresholding, $\eta_t(\cdot) = \eta(\cdot, \theta\frac{1}{\sqrt{L}}\|\boldsymbol{r}^t\|_2)$, where θ is free parameter
- $\eta_t'(\cdot)$: first order derivative of $\eta_t(\cdot)$
- $\delta \triangleq \frac{L}{N}$: undersampling ratio
- $\langle\cdot\rangle$: averaging operation over all entries of a vector

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 $\frac{L}{N}$

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

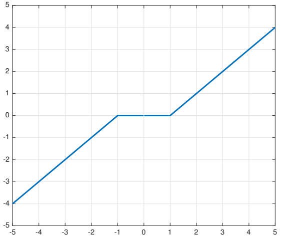
$$\tau_{t+1}^2 = \sigma_w^2 + \frac{1}{\delta}\mathbb{E}|\eta_t(X + \tau_t W) - X|^2 \tag{17}$$

  - X: random variable following the same distribution as $\boldsymbol{x}$
  - 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 $\boldsymbol{y} = \boldsymbol{S}\boldsymbol{x} + \boldsymbol{w}$ is reduced to uncoupled scalar estimation $(\boldsymbol{x}^t + (\boldsymbol{S}^* \boldsymbol{r}^t)_i = x_i + v_i^t$

Denoiser for AMP

Complex soft thresholding denoiser:

$$\eta_t^{\mathrm{soft}}(\tilde{x}^t) \triangleq \left(\tilde{x}^t - \theta\tau_t \frac{\tilde{x}^t}{|\tilde{x}^t|}\right) \mathbb{I}(|\tilde{x}^t| > \theta\tau_t) \quad (18)$$



- θ: threshold control parameter
- τ_t: noise variance, estimated by $\hat{\tau}_t = \frac{1}{\sqrt{L}}\|\boldsymbol{r}^t\|_2$
- $\mathbb{I}(\cdot)$: indicator function

The above is the classical minimax denoiser based on soft thresholding.

Better MMSE denoiser can be designed while accounting for channel distribution:

$$\eta_t^{\mathrm{mmse}}(\tilde{x}^t) \triangleq \mathbb{E}(X|\tilde{X}^t = \tilde{x}^t) \quad (19)$$

where $\tilde{X}^t$ is the random variable defined as $\tilde{X}^t \triangleq X + \tau_t W$.

Comparison of Denoisers

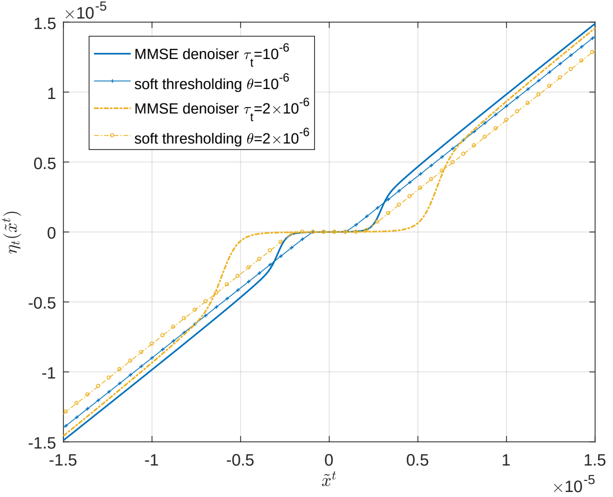


Figure: Soft thresholding denoiser and MMSE denoiser

User Activity Detection

Recall the signal plus noise model in AMP: $(\boldsymbol{S}^*\boldsymbol{r}^t + \boldsymbol{x}^t)_i = x_i + v_i^t$, which 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

$$\begin{cases} H_0 : X = 0, \text{ user is inactive} \\ H_1 : X \neq 0, \text{ user is active} \end{cases} \tag{20}$$

The optimal decision rule

$$LLR = \log\left(\frac{p_{\tilde{X}^t|X}(\tilde{x}^t|x \neq 0)}{p_{\tilde{X}^t|X}(\tilde{x}^t|x = 0)}\right) \underset{H_1}{\overset{H_0}{\lesseqgtr}} l_{th} \tag{21}$$

- LLR: log-likelihood ratio
- l_{th}: decision threshold determined by the detection criterion.

Analysis of Detection Error Probability

By state evolution, the likelihood distribution given X can be derived as:

$$p_{\tilde{X}^t|X}(\tilde{x}^t|x=0) = \frac{1}{\pi\tau_t^2}\exp\left(-\frac{|\tilde{x}^t|^2}{\tau_t^2}\right) \quad (22)$$

$$p_{\tilde{X}^t|X}(\tilde{x}^t|x\neq 0) = a\int_0^\infty \frac{\mathrm{erfc}(b\ln z + c)}{z^\gamma(z^2+\tau_t^2)}\exp\left(\frac{-|\tilde{x}^t|^2}{z^2+\tau_t^2}\right)dz \quad (23)$$

The log-likelihood ratio is computed as

$$LLR = \log\int_0^\infty \frac{a\pi\tau_t^2 z^{-\gamma}}{z^2+\tau_t^2}\mathrm{erfc}(b\ln z + c)\exp(|\tilde{x}^t|^2\Delta)dz \quad (24)$$

- $\Delta \triangleq \frac{1}{\tau_t^2} - \frac{1}{z^2+\tau_t^2}$
- LLR is monotonic in $|\tilde{x}^t|$

Missed Detection vs. False Alarm Probabilities

Based on the monotonicity, we simplify the decision rule as

$$|\tilde{x}^t| \underset{H_1}{\overset{H_0}{\lessgtr}} l'_{th} \tag{25}$$

The false alarm and missed detection probabilities:

$$P_F^t = \int_{|\tilde{x}^t|>l'_{th}} p_{\tilde{X}^t|X}(\tilde{x}^t|x=0)d\tilde{x}^t \tag{26}$$

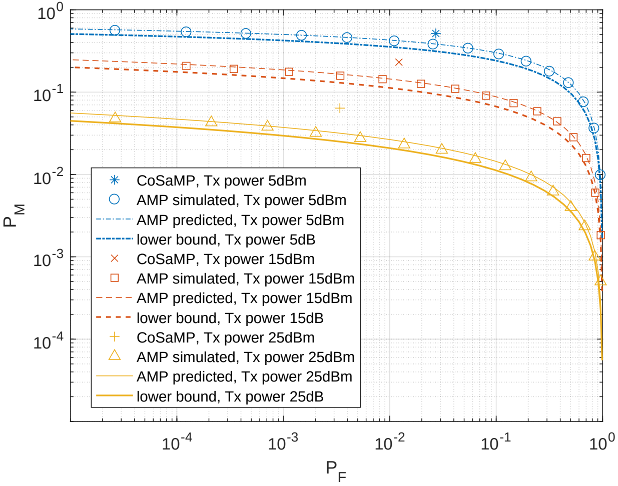
$$P_M^t = \int_{|\tilde{x}^t|<l'_{th}} p_{\tilde{X}^t|X}(\tilde{x}^t|x\neq 0)d\tilde{x}^t \tag{27}$$

- Decision is based on the amplitude of $\tilde{x}$
- Trade-off between P_F^t and P_M^t is achieved by adjusting l'_{th}
- P_F^t and P_M^t depend on noise variance τ_t (τ_∞ after converging), which can be tracked via the AMP state evolution

Simulation Parameters

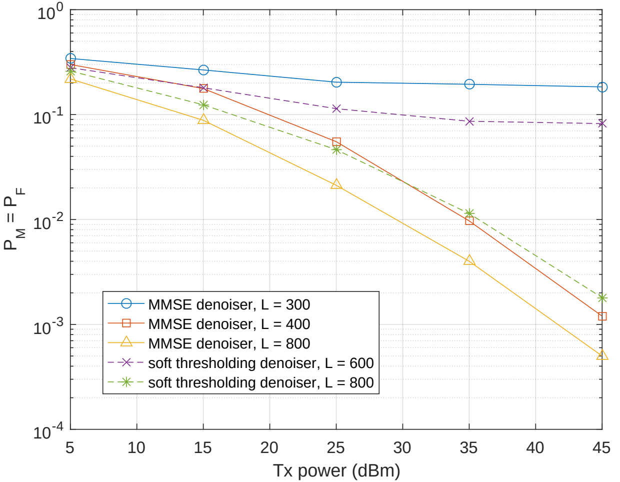
User number N	4000
Cell radius R	1000m
Activity probability ϵ	0.05
Signature sequence length L	800
Pathloss parameter α	15.3
Pathloss parameter β	37.6
Shadowing parameter σ_{SF}	8 dB
Background noise power	-99 dBm
Transmission power	5, 15, 25 dBm

Missed Detection vs. False Alarm



Small mismatch of predicted vs simulated curves due to neglecting shadowing

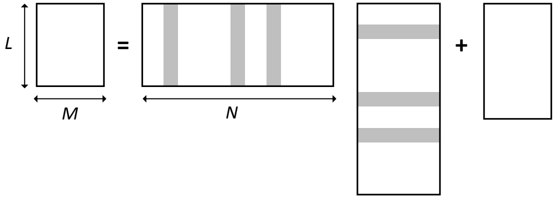
AMP Performance vs. SNR



Threshold for MMSE denoiser is better than soft thresholding denoiser.

Multi-Antenna Case

Multiple Antennas at the BS



- Multiple measurement vector (MMV) problem
  - Better performance than single measurement vector (SMV)
- Asymptotic analysis: Fix M, let $N, K, L \to \infty$, $\epsilon = \frac{K}{N}$, $\delta = \frac{L}{N}$,
- Main insight: Perfect user detection is possible when $M \to \infty$!
- But, the multi-antenna case is also more challenging:
  (i) convergence is slower;
  (ii) channel estimation error.

Two-Phase Transmission

- Pilot Transmission Phase
  - Due to large number of devices, non-orthogonal pilots are inevitable.
  - The same pilots can be used for both activity detection and channel estimation.
- Data Transmission Phase
  - The achievable rates are limited by the channel estimation error.

Signal Model in Pilot Phase

Received signal in pilot phase:

$$\boldsymbol{Y} = \sqrt{\xi} \sum_{n=1}^{N} \alpha_n \boldsymbol{s}_n \boldsymbol{h}_n^T + \boldsymbol{Z} \triangleq \sqrt{\xi} \boldsymbol{S} \boldsymbol{X} + \boldsymbol{Z} \quad (28)$$

- $\xi = \rho^{\mathrm{pilot}} L$: total transmit energy in pilot phase
- $\boldsymbol{h}_n \in \mathbb{C}^{M \times 1}$: channel coefficient between user n and BS, including path-loss fading, shadowing and Rayleigh fading static within each block;
- $\alpha_n \in \{1, 0\}$: indicating whether user n is active
- $\boldsymbol{X} \triangleq [\boldsymbol{h}_1 \alpha_1, \boldsymbol{h}_2 \alpha_2, \cdots, \boldsymbol{h}_N \alpha_N]^T \in \mathbb{C}^{N \times M}$
- $\boldsymbol{s}_n \in \mathbb{C}^{L \times 1}$: signature sequence of user n following i.i.d. $\mathcal{CN}(0, 1/L)$
- $\boldsymbol{S} \triangleq [\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_N]^T \in \mathbb{C}^{L \times N}$
- $\boldsymbol{Z} \in \mathbb{C}^{L \times M}$: effective noise following i.i.d. $\mathcal{CN}(0, \sigma^2 \boldsymbol{I})$

Vector Approximate Message Passing

- Vector generalization of AMP works iteratively as follows:

$$\boldsymbol{x}_n^{t+1} = \eta_{t,n}((\boldsymbol{R}^t)^H \boldsymbol{s}_n + \boldsymbol{x}_n^t)$$

$$\boldsymbol{R}^{t+1} = \boldsymbol{Y} - \boldsymbol{S}\boldsymbol{X}^{t+1} + \frac{1}{\delta}\boldsymbol{R}^t \sum_{n=1}^{N} \frac{\eta'_{t,n}((\boldsymbol{R}^t)^H \boldsymbol{s}_n + \boldsymbol{x}_n^t)}{N}$$

  where $\boldsymbol{R}^t = [\boldsymbol{r}_1^t, \cdots, \boldsymbol{r}_L^t]^T \in \mathbb{C}^{L \times M}$ is the residual

- Use MMSE denoise that accounts for channel distribution: (Here $\delta = \frac{L}{N}$)
  - $\eta_{t,n}(\cdot)$: denoiser that depends on β_n
  - $\eta'_{t,n}(\cdot)$: first-order derivative of $\eta_{t,n}(\cdot)$

State Evolution

- Performance analysis by state evolution for $K, N, L \to \infty$ with $\frac{L}{N} = \delta$ [Bayati-Montanari'11], [Kim-Chang-Jung-Baron-Ye'11], [Rangan'11]:

$$\Sigma_{t+1} = \frac{\sigma^2}{\xi}\boldsymbol{I} + \frac{1}{\delta}\mathbb{E}\left[(\eta_{t,\beta}(\boldsymbol{X}_\beta + \Sigma_t^{\frac{1}{2}}\boldsymbol{V}) - \boldsymbol{X}_\beta)(\eta_{t,\beta}(\boldsymbol{X}_\beta + \Sigma_t^{\frac{1}{2}}\boldsymbol{V}) - \boldsymbol{X}_\beta)^H\right] \tag{29}$$

- AMP is statistically equivalent to applying the denoiser to

$$\hat{\boldsymbol{x}}_{t,n} = \boldsymbol{x}_n + \Sigma_t^{\frac{1}{2}}\boldsymbol{v}_n = \alpha_n\boldsymbol{h}_n + \Sigma_t^{\frac{1}{2}}\boldsymbol{v}_n \tag{30}$$

- MMSE denoiser:

$$\eta_{t,n}(\hat{\boldsymbol{x}}_{t,n}) = \phi_{t,n}\beta_n(\beta_n\boldsymbol{I} + \Sigma_t)^{-1}\hat{\boldsymbol{x}}_{t,n} \tag{31}$$

$$\phi_{t,n} = \frac{1}{1 + \frac{1-\varepsilon}{\varepsilon}\exp\left(-\frac{M}{2}\left(\pi_{t,n} - \psi_{t,n}\right)\right)} \tag{32}$$

$$\pi_{t,n} = \frac{\hat{\boldsymbol{x}}_{t,n}^H(\Sigma_t^{-1} - (\Sigma_t + \beta_n\boldsymbol{I})^{-1})\hat{\boldsymbol{x}}_{t,n}}{M} \tag{33}$$

$$\psi_{t,n} = \frac{\log\det(\boldsymbol{I} + \beta_n\Sigma_t^{-1})}{M} \tag{34}$$

Simplified MMSE Denoiser

- With i.i.d. fading, Σ_{t+1} is a diagonal matrix with identical diagonal entries

$$\Sigma_t = \tau_t^2 \boldsymbol{I}$$

- MMSE denoiser reduces to

$$\eta_{t,n}(\hat{\boldsymbol{x}}_{t,n}) = {\color{red}\phi_{t,n}} \frac{\beta_n}{\beta_n + \tau_t^2} \hat{\boldsymbol{x}}_{t,n} \tag{35}$$

$$\phi_{t,n} = \frac{1}{1 + \frac{1-\varepsilon}{\varepsilon} \exp\left(-\frac{M}{2}\left(\pi_{t,n} - \psi_{t,n}\right)\right)} \tag{36}$$

$$\pi_{t,n} = \frac{\left(\frac{1}{\tau_t^2} - \frac{1}{\tau_t^2 + \beta_n}\right) \hat{\boldsymbol{x}}_{t,n}^H \hat{\boldsymbol{x}}_{t,n}}{M} \tag{37}$$

$$\psi_{t,n} = \log\left(1 + \frac{\beta_n}{\tau_t^2}\right) \tag{38}$$

- Asymptotically as $M \to \infty$, $\phi_{t,n}$ is either 0 or 1 depending on whether device n is active or not.

Massive MIMO Guarantees Perfect Activity Detection

Theorem: A massive MIMO system can detect device activities perfectly, i.e.,

$$\lim_{M\to\infty} P_M^{t,n}(M) = \lim_{M\to\infty} P_F^{t,n}(M) = 0$$

Proof: By strong law of large numbers:

$$\pi_{t,n} \to \begin{cases} \beta_n/\tau_t^2, & \text{if } \alpha_n = 1 \\ \beta_n/(\beta_n + \tau_t^2), & \text{if } \alpha_n = 0 \end{cases}$$

The proof follows as $a > \log(1+a) > \frac{a}{1+a}$ for all $a > 0$.

What is the cost of massive connectivity?

Channel Estimation Error

- Covariance of estimated channel $\hat{\boldsymbol{h}}_{t,k}$: $\mathrm{Cov}(\hat{\boldsymbol{h}}_{t,k}, \hat{\boldsymbol{h}}_{t,k}) = \upsilon_{t,k}(M)\boldsymbol{I}$
- Covariance of channel estimation error $\Delta\boldsymbol{h}_{t,k} = \boldsymbol{h}_{t,k} - \hat{\boldsymbol{h}}_{t,k}$

$$\mathrm{Cov}(\Delta\boldsymbol{h}_{t,k}, \Delta\boldsymbol{h}_{t,k}) = \Delta\upsilon_{t,k}(M)\boldsymbol{I} \tag{39}$$

- As $M \to \infty$

$$\lim_{M\to\infty} \upsilon_k(M) = \frac{\beta_k^2}{\beta_k + \tau_\infty^2} \tag{40}$$

$$\lim_{M\to\infty} \Delta\upsilon_k(M) = \frac{\beta_k \tau_\infty^2}{\beta_k + \tau_\infty^2} \tag{41}$$

where τ_∞^2 is the fixed-point solution to state evolution: ($\epsilon = \frac{K}{N}$, $\delta = \frac{L}{N}$)

$$\tau_{t+1}^2 = \frac{\sigma^2}{\xi} + \frac{\epsilon}{\delta}\mathbb{E}_\beta\left[\frac{\beta\tau_t^2}{\beta + \tau_t^2}\right] \tag{42}$$

Assuming $L > K$ and high SNR, then $\tau_\infty^2 \to \frac{\sigma^2}{\xi(1-\frac{\epsilon}{\delta})}$.

Data Transmission Phase

- Consider a system with K users transmitting to BS with M antennas.
- Received signal at the BS with estimated channel $\tilde{\boldsymbol{h}}_k$'s:

$$\boldsymbol{y} = \sum_{k\in\mathcal{K}} \boldsymbol{h}_k \sqrt{\rho^{\mathrm{data}}} u_k + \boldsymbol{z} = \sum_{k\in\mathcal{K}} \tilde{\boldsymbol{h}}_k \sqrt{\rho^{\mathrm{data}}} u_k + \sum_{k\in\mathcal{K}} \Delta\boldsymbol{h}_k \sqrt{\rho^{\mathrm{data}}} u_k + \boldsymbol{z}$$

- Maximum ratio combining:

$$\hat{u}_k = \boldsymbol{w}_k^H \tilde{\boldsymbol{h}}_k \sqrt{\rho^{\mathrm{data}}} u_k + \boldsymbol{w}_k^H \left(\sum_{n\in\mathcal{K}, n\neq k} \tilde{\boldsymbol{h}}_n \sqrt{\rho^{\mathrm{data}}} u_n + \sum_{n\in\mathcal{K}} \Delta\boldsymbol{h}_n \sqrt{\rho^{\mathrm{data}}} u_n + \boldsymbol{z} \right)$$

- The achievable rate of user k is [Hassibi-Hochwald'03]

$$R_k = \frac{T-L}{T} \mathbb{E}[\log_2(1+\gamma_k)], \quad \forall k \in \mathcal{K},$$

where

$$\gamma_k = \frac{\rho^{\mathrm{data}} |\boldsymbol{w}_k^H \hat{\boldsymbol{h}}_k|^2}{\rho^{\mathrm{data}} \sum\limits_{n\in\mathcal{K}, n\neq k} |\boldsymbol{w}_k^H \hat{\boldsymbol{h}}_n|^2 + \rho^{\mathrm{data}} \|\boldsymbol{w}_k\|^2 \sum\limits_{n\in\mathcal{K}} \frac{\beta_n \tau_\infty^2}{\beta_n + \tau_\infty^2} + \sigma^2 \|\boldsymbol{w}_k\|^2}$$

Achievable Rate with MMSE Beamforming

- With MMSE: $\boldsymbol{w}_k^{\mathrm{MMSE}} = \left(\sum_{n \in \mathcal{K}} \rho^{\mathrm{data}} \hat{\boldsymbol{h}}_n \hat{\boldsymbol{h}}_n^H + \sum_{n \in \mathcal{K}} \frac{\rho^{\mathrm{data}} \beta_n \tau_\infty^2}{\beta_n + \tau_\infty^2} \boldsymbol{I} + \sigma^2 \boldsymbol{I} \right)^{-1} \hat{\boldsymbol{h}}_k$

$$\lim_{M \to \infty} \gamma_k^{\mathrm{MMSE}} \to \frac{\beta_k^2}{\beta_k + \tau_\infty^2} \Gamma$$

where Γ is fixed-point solution to ($\mu = \frac{K}{M}$):

$$\Gamma = \frac{1}{\mu \mathbb{E}\left[\frac{\beta^2}{\beta + \tau_\infty^2 + \beta^2 \Gamma} \right] + \mu \mathbb{E}\left[\frac{\beta \tau_\infty^2}{\beta + \tau_\infty^2} \right]}$$

In the special case of perfect CSI, the above result reduces to [Tse-Hanly'99]

Cost of Massive Uncoordinated Access

- Fixed number of K users:

$$\gamma_k \to \frac{\beta_k^2}{\beta_k + \frac{\sigma^2}{\rho^{\text{pilot}} L}} \Gamma \tag{43}$$

$$\frac{1}{\Gamma} = \frac{1}{M} \sum_{n \in \mathcal{K}} \frac{\beta_n^2}{\beta_n + \frac{\sigma^2}{\rho^{\text{pilot}} L} + \beta_n^2 \Gamma} + \frac{1}{M} \sum_{n \in \mathcal{K}} \frac{\frac{\beta_n \sigma^2}{\rho^{\text{pilot}} L}}{\beta_n + \frac{\sigma^2}{\rho^{\text{pilot}} L}} \tag{44}$$

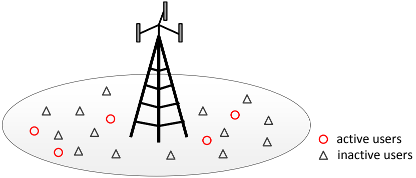
- Massive number of N potential users with K active user:

$$\gamma_k \to \frac{\beta_k^2}{\beta_k + \tau_\infty^2} \Gamma \tag{45}$$

$$\frac{1}{\Gamma} = \frac{1}{M} \sum_{n \in \mathcal{K}} \frac{\beta_n^2}{\beta_n + \tau_\infty^2 + \beta_n^2 \Gamma} + \frac{1}{M} \sum_{n \in \mathcal{K}} \frac{\beta_n \tau_\infty^2}{\beta_n + \tau_\infty^2} \tag{46}$$

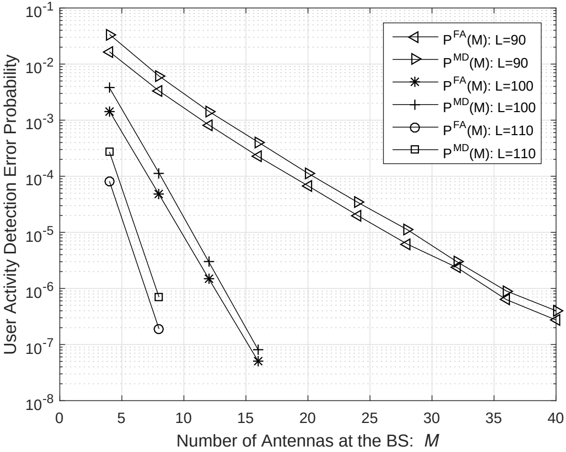
- At high SNR: $\tau_\infty^2 \approx \frac{\sigma^2}{\rho^{\text{pilot}}(L-K)}$
- Channel estimation error is increased due to the non-orthogonal pilots

Numerical Example



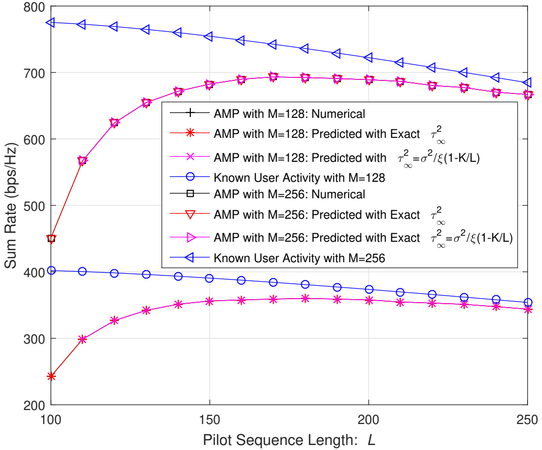
- $N = 2000$ users
- $K = 100$ active users
- Transmit power: $\rho^{\text{pilot}} = \rho^{\text{data}} = 23\text{dBm}$
- User distance to BS: $[0.5\text{km}, 1\text{km}]$
- Path loss: $\beta_n = -128.1 - 36.7 \log_{10}(d_n), \forall n$
- 100kHz bandwidth, 10ms coherence time
- $T = 1000$ symbols per coherence time

User Activity Detection



- Probabilities of missed detection and false alarm reduce as L increases
- Probabilities of missed detection and false alarm go to zero as M increases

Achievable Rate with Massive MIMO

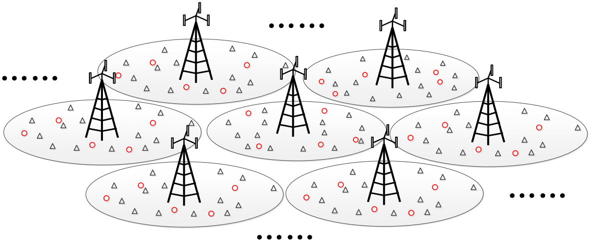


- The optimal L when user activity is unknown needs to be longer
- There is a loss in sum-rate due to the need for longer pilot

Multi-Cell Systems

User Activity Detection in Multicell Systems

- What is the impact of the inter-cell interference?



- How to overcome the inter-cell interference?

Activity Detection in Multicell Systems

- Multi-cell system with B BSs each equipped with M antennas;
- N single-antenna devices per cell, K of which are active;
- Device n in cell b is assigned a length-L unique signature sequence $\boldsymbol{s}_{bn}$;
- Received signal $\boldsymbol{Y}_b \in \mathbb{C}^{L\times M}$ at BS b is

$$\begin{aligned}\boldsymbol{Y}_b &= \sum_{n=1}^{N} \alpha_{bn}\boldsymbol{s}_{bn}\boldsymbol{h}_{bbn}^T + \sum_{j=1,j\neq b}^{j=B}\sum_{n=1}^{N}\alpha_{jn}\boldsymbol{s}_{jn}\boldsymbol{h}_{bjn}^T + \boldsymbol{Z}_b \\ &= \boldsymbol{S}_b\boldsymbol{X}_{bb} + \sum_{j=1,j\neq b}^{j=B}\boldsymbol{S}_j\boldsymbol{X}_{bj} + \boldsymbol{Z}_b, \end{aligned} \tag{47}$$

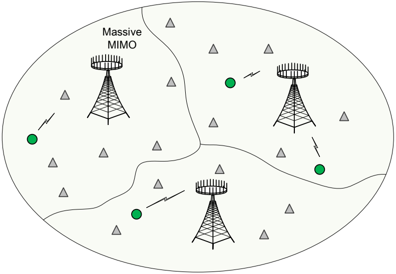
where

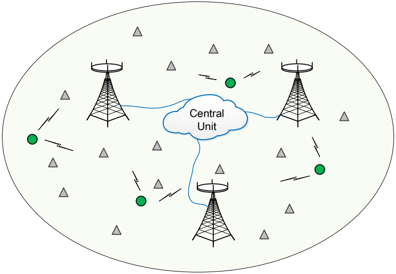
- $\alpha_{bn} \in \{1,0\}$ activity indicator; $\boldsymbol{Z}_b \in \mathbb{C}^{L\times M}$ Gaussian noise with variance σ^2.
- $\boldsymbol{h}_{bjn} \in \mathbb{C}^{M\times 1}$ is the channel from user n in cell j to BS b
- $\boldsymbol{S}_j \triangleq [\boldsymbol{s}_{j1},\cdots,\boldsymbol{s}_{jN}] \in \mathbb{C}^{L\times N}$; $\boldsymbol{X}_{bj} \triangleq [\alpha_{j1}\boldsymbol{h}_{bj1},\cdots,\alpha_{jN}\boldsymbol{h}_{bjN}]^T \in \mathbb{C}^{N\times M}$

The inter-cell interference brings performance degradation for activity detection.

AMP Based Activity Detection for Multi-cell

With AMP, we consider two strategies to deal with the inter-cell interference





- **Massive MIMO**: Each BS has a large-scale antenna array, and operates independently, while treating the inter-cell interference as noise.
- **Cooperative MIMO**: Each BS has a moderate number of antennas, and is connected to a central unit (CU), where cooperative detection is performed.

Activity Detection in Massive MIMO System

- Each BS is equipped with a large-scale antenna array, i.e., M is large.
- Each BS aims to detect the active devices within its own cell, and the inter-cell interference is treated as noise:

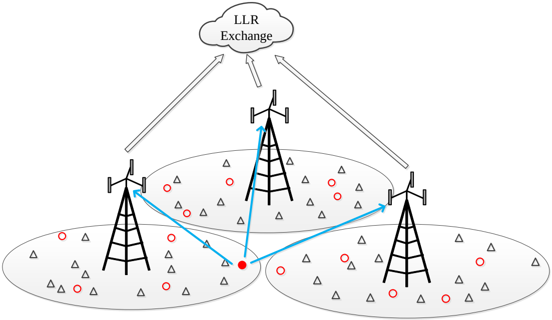
$$\begin{aligned} \boldsymbol{Y}_b &= \boldsymbol{S}_b \boldsymbol{X}_{bb} + \sum_{j \neq b} \boldsymbol{S}_j \boldsymbol{X}_{bj} + \boldsymbol{Z}_b \\ &\triangleq \boldsymbol{S}_b \boldsymbol{X}_{bb} + \boldsymbol{Z'}_b \end{aligned} \tag{48}$$

- By approximating $\boldsymbol{Z'}_b$ as a Gaussian noise, the resulting system model in multicell case is similar to that in the single-cell case.
- AMP can be used to detect the active devices in cell b by recovering the non-zero rows of $\boldsymbol{X}_{bb}$ based on $\boldsymbol{Y}_b$.

Activity Detection in Cooperative MIMO System

Potential ways to perform the cooperative detection with BSs connected to CU:

- **Centralized detection**: The received signals at the BSs $\boldsymbol{Y}_b$'s are forwarded to the CU, where a large-scale AMP is used for activity detection.
  Interference is completely avoided. However, need high-bandwidth BS-CU links.
- **Distributed detection**: Each BS performs a preliminary activity detection, and forwards the results to the CU, where an aggregation is performed.
  Forwarding the detection LLRs can save bandwidth of the BS-CU links.



Cooperative Activity Detection

- Each BS detects the active devices in all B cells using the knowledge of all signature sequences. This can be achieved by recovering the interference as

$$
\begin{aligned}
\boldsymbol{Y}_b &= \boldsymbol{S}_b \boldsymbol{X}_{bb} + \sum_{j \neq b} \boldsymbol{S}_j \boldsymbol{X}_{bj} + \boldsymbol{Z}_b \\
&= \begin{bmatrix} \boldsymbol{S}_1 & \cdots & \boldsymbol{S}_B \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_{1b} \\ \vdots \\ \boldsymbol{X}_{Bb} \end{bmatrix} + \boldsymbol{Z}_b \\
&\triangleq \boldsymbol{S}\boldsymbol{X}_b + \boldsymbol{Z}_b
\end{aligned} \tag{49}
$$

- **Preliminary detection**: The BS detects the active devices by estimating the non-zero row of $\boldsymbol{X}_b$ from $\boldsymbol{Y}_b$ using AMP. This is similar to the single-cell case.
- **Quantization and forwarding**: The detection results by AMP at each BS are quantized and sent to the CU in the form of LLRs (e.g., 3-4 bits per LLR.)
- **Aggregation**: CU aggregates the independent LLRs and declares activities.

Comparison of Massive MIMO and Cooperative MIMO

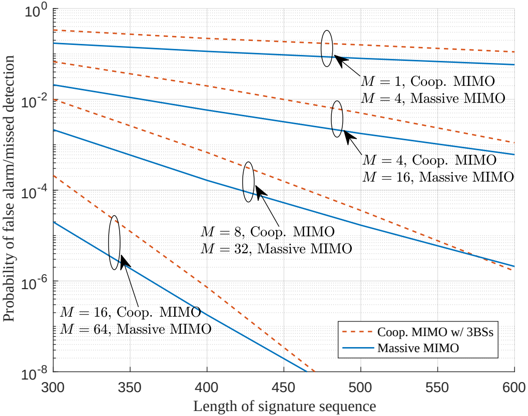


Figure: Cell-edge user performance in a network with 19 cells and 2000 devices per cell, among which 100 devices are active. To achieve comparable performance as cooperative MIMO, four times as many as antennas are required in the massive MIMO case.

Summary

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.
- Cooperation can improve the cell-edge performance.

Implications for network design:

- The use of non-orthogonal pilots is inevitable.
- Massive MIMO needs to be deployed for good detection performance.
- Multi-cell cooperation can further help.
- Channel estimation is the main bottleneck.

Further Information

Zhilin Chen, Foad Sohrabi, and Wei Yu,
"Sparse Activity Detection for Massive Connectivity",
IEEE Transactions on Signal Processing, vol. 66, no. 7, pp. 1890-1904, 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, pp. 2933-2959, June 2018.

Zhilin Chen, Foad Sohrabi, and Wei Yu,
"Multi-Cell Sparse Activity Detection for Massive Connectivity: Massive MIMO versus Cooperative MIMO",
IEEE Transactions on Wireless Communications, vol. 18, no. 8, pp. 4060-4074, August 2019.