Hybrid ARQ optimizations for wireless networks

Philippe Ciblat

Joint works with F. Bassi, P. Duhamel, A. Khreis, N. Ksairi, A. Leduc, C. Le Martret, X. Leturc

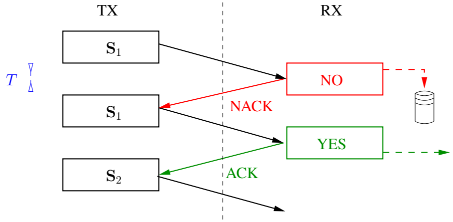
Outline

- A very short introduction to Hybrid ARQ (HARQ)
- HARQ **improvement**:
  - *superposition coding*
- HARQ **parameters' optimization**:
  - *adaptive modulation and coding scheme*
- **Resource allocation optimization** for HARQ based system:
  - *energy efficiency with Ricean channel*

Part 1 : Introduction to HARQ

From ARQ (*Automatic ReQuest*) ...

Let $\mathbf{S} = [s_0, \cdots, s_{N-1}]$ be a packet composed by N uncoded symbols



- Pros: Adaptability to the real propagation states (noise, channel)
    - Robustness to no instantaneous Channel State Information at the Transmitter (CSIT)
    - Diversity if time-varying channel
    - High granularity with adapted Modulation and Coding Scheme (MCS) related to instantaneous channel behavior
- Pros: Cheap feedback link (one bit)
- Cons: High latency, Buffer size

... Towards Hybrid ARQ (HARQ): Type-I HARQ

Remark

Retransmission does not contradict forward error coding (FEC)

Type-I HARQ: packet **S** is composed by coded symbols s_n

- first packet is more protected
- there is less retransmission
- transmission delay is reduced

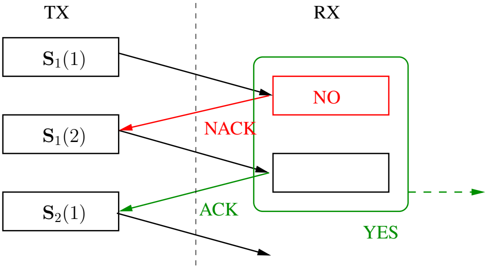
- Efficiency is upper-bounded by the code rate

Drawbacks

- Each received packet is treated independently
- Mis-decoded packet is thrown in the trash

Type-II HARQ

Memory at RX side is considered ⇒ Type-II HARQ



Main examples:

- *Chase Combining* (CC)
- *Incremental Redundancy* (IR)

Examples: CC-HARQ and IR-HARQ

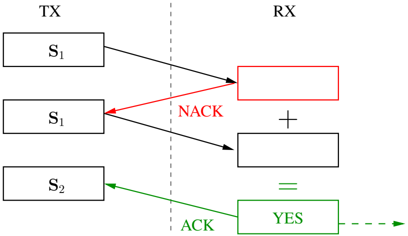
CC

$$\begin{aligned} Y_1 &= S_1 + N_1 \\ Y_2 &= S_1 + N_2 \end{aligned}$$

then detection on

$$Y = (Y_1 + Y_2)/2$$

SNR-Gain equal to 3dB



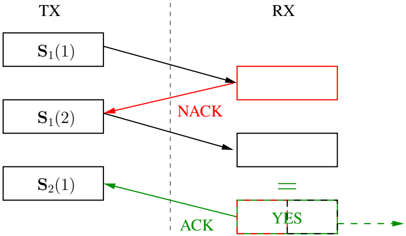
IR

$$\begin{aligned} Y_1 &= S_1(1) + N_1 \\ Y_2 &= S_1(2) + N_2 \end{aligned}$$

then detection on

$$Y = [Y_1, Y_2]$$

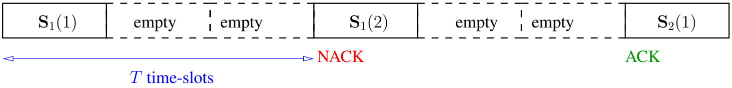
Coding gain



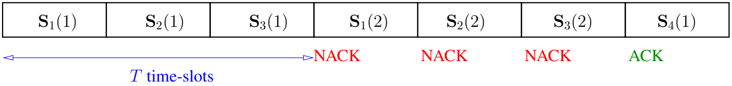
Delayed feedback management

Management for T:

- Stop-and-Wait



- Parallel Stop-and-Wait/Selective Repeat



Standard Assumption:

- No error on feedback
- No delay ($T = 1$)

Performance metrics

- **Packet Error Rate** (PER):

$$\text{PER} = \text{Prob(message is not decoded)}$$

- **Efficiency** (*Throughput*/*Goodput*/etc):

$$\eta = \frac{\text{information bits received without error}}{\text{transmitted bits}}$$

- **(Mean) delay**:

$$d = \#\ \text{transmitted packets when message is correctly received}$$

- **Jitter**:

$$\sigma_d = \text{delay standard deviation}$$

Quality of Service (QoS)

- Data: PER and efficiency
- Voice on IP: delay
- Video Streaming: efficiency and jitter

Closed-form expressions for metrics

$$\text{PER} = 1 - \sum_{k=1}^{L} p(k)$$

$$\eta \propto \frac{\sum_{k=1}^{L} p(k)}{L(1 - \sum_{k=1}^{L} p(k)) + \sum_{k=1}^{L} kp(k)}$$

$$d = \frac{\sum_{k=1}^{L} kp(k)}{\sum_{k=1}^{L} p(k)}$$

$$\sigma_d = \sqrt{\frac{\sum_{k=1}^{L} k^2 p(k)}{\sum_{k=1}^{L} p(k)} - d^2}$$

with [Leduc12]

- $p(k)$ probability to receive information packet in exactly k transmissions
- L maximum number of transmissions per message

Example: Type-I HARQ

- Let π_0 be the probability the message is not well decoded with one transmission. Then

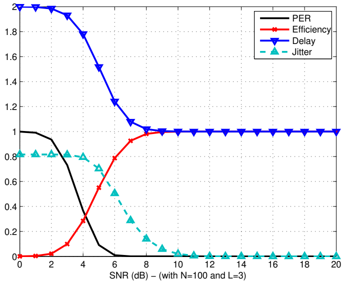
$$p(k) = (1-\pi_0)\pi_0^{k-1}$$

- Let a message be composed by N BPSK uncoded symbols. Then, on Gaussian channel,

$$\pi_0 = 1 - \left(1 - Q\left(\sqrt{2\mathrm{SNR}}\right)\right)^N$$

Results

$$\begin{aligned}
\mathrm{PER} &= \pi_0^L \\
\eta &= 1-\pi_0 \\
d &= L + \frac{1}{1-\pi_0} - \frac{L}{1-\pi_0^L} \\
\sigma_d^2 &= \frac{\pi_0+\pi_0^{2L+1}-\pi_0^L(L^2+\pi_0^2(1+\pi_0)^2)}{(1-\pi_0)^2(1-\pi_0^L)^2} \\
&- \frac{2\pi_0^{L+1}(L^2-1)}{(1-\pi_0)^2(1-\pi_0^L)^2}
\end{aligned}$$



Part 2: HARQ optimizations

2.1 HARQ improvement
2.2 HARQ parameters' optimization
2.3 Resource allocation optimization for HARQ based system

2.1 - HARQ improvement with delayed feedback

In practice, $T \neq 1$ ($T = 8$ in LTE)

Idea for parallel Stop-and-Wait

Send redundant packets **in advance** between pre-assigned time-slots and superpose them with packets related to other messages

- Similar idea when $T = 1$ [Shamai08,Assimi09,Szczecinski14]
- In SotA, with perfect CSIT or past CSIT (multi-bit feedback)
- Why could it work? non-orthogonal transmission with potential of Multiple Access Channel decoding

Expected gains

- Lower latency
- Higher reliability

2.1 - Proposed protocol

Let $\mathbf{p}_k(\ell)$ be the ℓ-th packet/chunk associated with the message k
We do a transmission with two layers

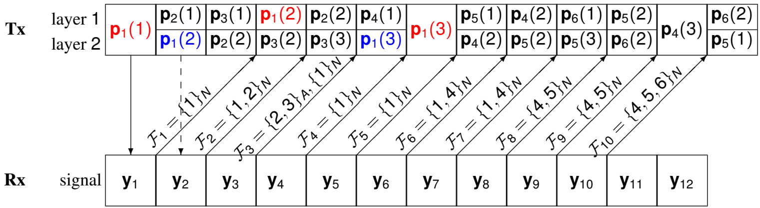
- Layer 1: parallel Stop-and-Wait HARQ
- Layer 2: superposed redundant packets

$$\mathbf{p}_k(\ell) \quad \text{without superposition}$$

$$\sqrt{\alpha}\,\underbrace{\mathbf{p}_k(\ell)}_{layer\ 1}+\sqrt{1-\alpha}\,\underbrace{\mathbf{p}_{k'}(\ell')}_{layer\ 1} \quad \text{with superposition}$$

How do we choose the superposed redundant packets?

- Superpose packets of the most recent messages ⇒ Low latency
- Superpose unsent redundant packets ⇒ High reliability



2.1 - How to decode?

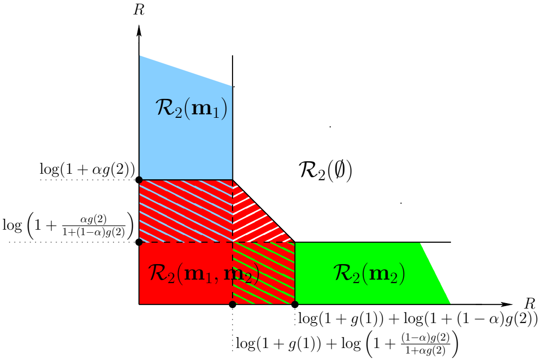
Received signal until time-slot 2

$$\mathbf{y}_1 = h(1)\mathbf{p}_1(1) + \mathbf{w}(1)$$
$$\mathbf{y}_2 = h(2)\sqrt{\alpha}\mathbf{p}_2(1) + h(2)\sqrt{1-\alpha}\mathbf{p}_1(2) + \mathbf{w}(2)$$

Equivalent to a MIMO-MAC

Decoders:

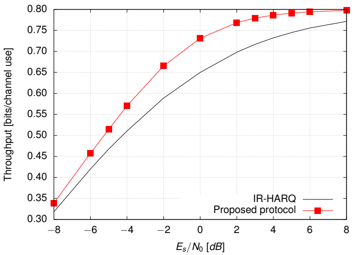
- Multi-message based Decoder
- Single-message based Decoder

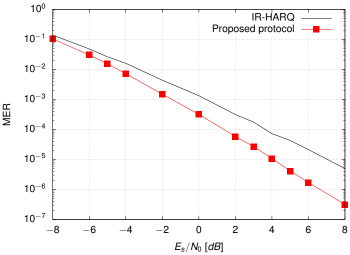


$(g(\ell) = |h(\ell)|^2)$

2.1 - Numerical results: Throughput and MER

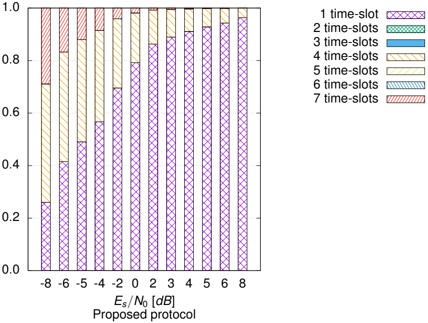
- **HARQ protocol** : IR-HARQ with $L = 3$, $R = 0.8$, best α
- **Feedback delay** : $T = 3$ time-slots
- **Transmit energy** : E_s (per symbol)

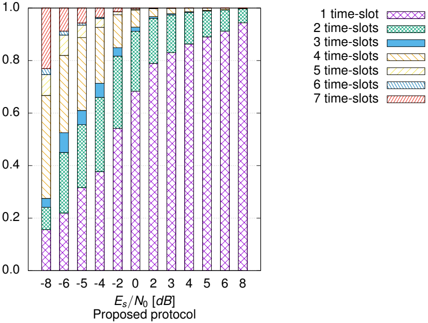




- ○ Around 2dB-gain at moderate SNR
- ○ 10% throughput gain at 0dB
- ○ Diversity gain due to multi-layer transmission

2.1 - Numerical results: Latency

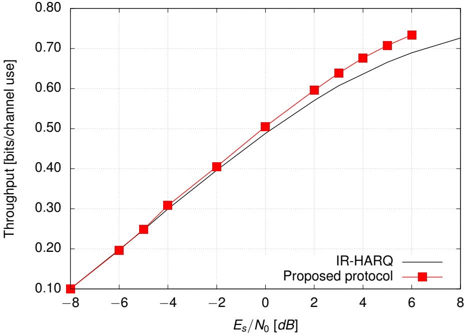




- More packets served with small delays ($<$ 4 time-slots)
- but average delay close to each other

2.1 - Numerical results: Practical scheme

- IR-HARQ with $L = 3$
- RCPC whose successive rates are 0.8, 0.4, and 0.26.
- BPSK modulated symbols
- Decoding of message k:
  - Combining observations samples sharing the same packet $\mathbf{p}_k(\ell)$
  - Calculating LLR for each observation sample
  - Computing Soft Viterbi's algorithm



2.2 - Modulation and Coding scheme optimization

Main goal

Selecting the Modulation and Coding Scheme (MCS) per packet

- when IR-HARQ is used (packet=chunk)
- based on the available partially-outdated CSI
- Why is it of interest?
  - if BPSK: a few redundant bits sent but well protected
  - if QAM: a lot of redundant bits sent but not well protected

LTE context

- Downlink from a base station (BS) to K mobile users
- Transmission done by Resource Block (RB) $=$ Q channel uses
  - B assigned RBs per frame and TX power per user constant during long duration, e.g., the so-called "semi-persistent scheduling" mode
  - MCS constant within 1 frame but adjustable frame by frame

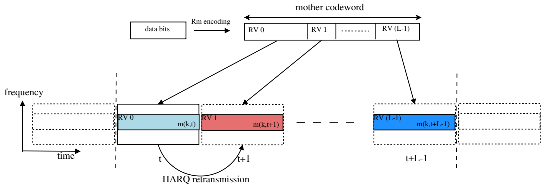
$\rightarrow$ $\mathrm{MCS}_{k,t} \stackrel{\text{def}}{=} (m_{k,t}, R_{k,t})$: the MCS during frame t

2.2 - IR-HARQ structure

- 1 HARQ round/transmission per frame
- When new transmission
  - Choose (by our algo.) a MCS with $2^{m_{k,t}}$-QAM and coding rate $R_{k,t}$
  - Thus $D_{k,t} = m_{k,t} R_{k,t}.(QB)$ information bits to send
  - Apply a mother code of rate R_0 on these information bits
  - Then pick up $m_{k,t}.(QB)$ coded bits to send
- When retransmission (after NACK): $D_{k,t}$ already fixed, and

$$R_{k,t} = \frac{D_{k,t}}{QB \sum_{j=t-\ell_{k,t}}^{t} m_{k,j}},$$

so choose $m_{k,t}$ only, such that, $R_{k,t} \geq R_0$



2.2 - Channel model between BTS and user k

- **block fading and frequency selective:**
  - channel impulse response $\mathbf{h}_{k,t} = [h_{k,t}(0), \ldots, h_{k,t}(M-1)]^{\mathrm{T}}$ constant on frame t
  - taps are independent and Rayleigh distributed
- **time correlated:**
  - $(\mathbf{h}_{k,t})_t$ is a first-order Gauss-Markov process

$$\mathbf{h}_{k,t} = \alpha \mathbf{h}_{k,t-1} + \sqrt{1-\alpha^2}\mathbf{w}_{k,t}(m),\ t \geq 0\,.$$

  - $\alpha \in (0,1)$ is the temporal fading coefficient from Jakes' model
- **frequency response:** N-FFT of $\mathbf{h}_{k,t}$

$$\mathbf{H}_{k,t} = [H_{k,t}(0), \ldots, H_{k,t}(N-1)]^{\mathrm{T}}$$

2.2 - Performance metrics: Block Error Rate (BLER)

$\pi_{k,t}$: Block Error Rate (BLER) based on the so-far received frames

$$\pi_{k,t} = 1 - d e^{-c_F \bar{\epsilon}_{k,t}^F - \cdots - c_1 \bar{\epsilon}_{k,t}}$$

- $F \in \mathbb{N}^*$: approximation order
- $d, c_1, \ldots, c_F$: curve-fitting parameters
- $\bar{\epsilon}_{k,t}$: average physical-layer Bit Error Rate (BER) associated with the hard-decision made on the so-far received coded bits [Vandendorpe09]

$$\bar{\epsilon}_{k,t} = \frac{\sum_{j \in \text{HARQ rounds}} m_{k,t} \sum_{n \in \text{assigned channel uses}} 0.2 e^{-1.6 \frac{E_k |H_{k,j}(n)|^2}{(2^{m_{k,j}} - 1) N_0}}}{\#\text{of assigned channel uses} \times \sum_{j \in \text{HARQ rounds}} m_{k,j}}$$

2.2 - Mathematical Goal

Our objectives

- For each frame, determine
  - the modulation scheme $m_{k,t}$
  - the coding rate $R_{k,t}$ (when new transmission)
- based on delayed CSI
  - $\mathbf{h}_{k,t-1}$ or $\bar{\epsilon}_{k,t-1}$
- to maximize the average throughput

Appropriate tool: Markov Decision Process (MDP)

2.2 - MDP definition

MDP framework

- state space $\mathcal{S}$: a set of states
- action space $\mathcal{A}$: a set of *actions*
  - $\mathcal{A}(s)$: admissible actions for state $s \in \mathcal{S}$
- state transition distribution $Q(.|s, a)$: $a \in \mathcal{A}(s)$
- reward: a function $r : \mathcal{S} \to \mathbb{R}$

A deterministic policy

A function $f : \mathcal{S} \to \mathcal{A}$

- such that $f(s) \in \mathcal{A}(s), \forall s \in \mathcal{S}$

Goal

Find out a policy maximizing an **average reward**

2.2 - Our MDP

- State:

  $\left(\ell_{k,t}, \mathbf{h}_{k,t-1}, \bar{\epsilon}_{k,t-1}, \#\text{of data bits in the codeword}, \#\text{of so-far received coded bits}\right) \in \mathcal{S}$

  with $\ell_{k,t}$ corresponds to the number of previous transmissions but
  - $\ell_{k,t} = 0$: a new transmission after an ACK
  - $\ell_{k,t} = L$: a new transmission after L NACKs
- Action:
  - for new transmission: $(D_{k,t} = \#\text{of data bits}, m_{k,t})$ (29 values in LTE)
  - for retransmission: $m_{k,t}$ (3 values in LTE)
- Reward:

$$r(s) \stackrel{\text{def}}{=} \begin{cases} \frac{\#\text{of data bits in } s}{T_{\text{frame}}}, & \text{if } \ell = 0, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\eta_k^{f_k}(s_{k,0}) = \liminf_{t \to \infty} \frac{1}{t} \mathbb{E}\left[\sum_{j=0}^{t-1} r(s_{k,j})\right] \text{bits/sec}$$

2.2 - Optimal policy

Main result

Under mild assumptions, the problem is solvable, i.e., it exists an optimal policy f_k^* such that

$$\eta_k^{f_k^*} = \eta_k^* \quad \text{with} \quad \eta_k^* \stackrel{\text{def}}{=} \sup_{f_k \in \mathcal{F}} \eta_k^{f_k}(s_{k,0})$$

Algorithm (Value Iteration)

$$f_{k,t}(s) = \arg \max_{a \in \mathcal{A}(s)} \left[r(s) + \int_{\mathcal{S}} v_{k,t-1}(y) Q(dy|s,a) \right]$$

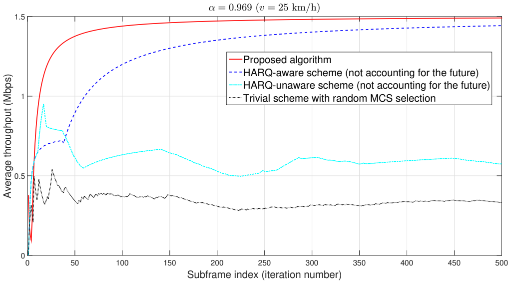
and

$$v_{k,t}(s) = r(s) + \int_{\mathcal{S}} v_{k,t-1}(y) Q(dy|s, f_{k,t}(s))$$

with $v_{k,0} = 0$

2.2 - Numerical results (LTE setup)

- In red, our proposed policy
- In blue, MCS leading to the smallest BLER $\bar{\epsilon}_{k,t}$ for the (outdated) $\mathbf{h}_{k,t-1}$ (channel correlation not taken into account)
- In cyan, MCS leading to the smallest BER for the (outdated) $\mathbf{h}_{k,t-1}$ (HARQ and channel correlation not taken into account)
- In black, random MCS



2.3 - Statistical CSIT based resource allocation

Only channel statistics known at the transmitter

- fast-varying Rayleigh/Rice fading channel
- costly to report instantaneous channel realizations
- cheap to report statistics due to its coherence time

HARQ to handle unknown channel variation

Applications

- Mobile Ad Hoc networks (MANET)
- Cellular networks with high mobility

2.3 - Communication model

- **PHY layer within a cell**
  - OFDMA: no Inter-Symbol and Multi-User interferences
  - No multi-cell interference assumption

- **Statistical channel model (for the k-th link)**
  - Let $h_k(j, m)$ be the m-th filter tap at OFDMA symbol j
    Independent but not identically distributed $\sim \mathcal{CN}(A_k \delta_{m,0}, \varsigma_{k,m}^2)$
  - Let $H_k(j, n)$ be the n-th Fourier component at OFDMA symbol j
    non-independent wrt n but identically distributed $\sim \mathcal{CN}(A_k, \varsigma_k^2)$ *with* $\varsigma_k^2 = \sum_m \varsigma_{k,m}^2$

k-th link characterization

- Subcarriers are statistically equivalent
  - γ_k: bandwidth proportion assigned to link k
  - Q_k: energy used by link k in one OFDM symbol
    - independent of subcarrier
    - $E_k = Q_k / \gamma_k$: energy of link k in entire bandwidth
- Rice fading channel

2.3 - Resource allocation optimization problem

Energy-efficiency based problem

$$\min_{\gamma,\mathbf{E}} f(\{\mathcal{E}_k(\gamma_k, E_k)\}_{k=1,\ldots,K})$$

$$\text{s.t.} \quad \mathbf{QoS}_k(\gamma_k, E_k) \geq \mathbf{QoS}_k^{(0)}, \ \forall k \in \{1,\ldots,K\}$$

$$\sum_{k=1}^{K} \gamma_k \leq 1$$

$$\gamma_k \geq 0, \ E_k \geq 0, \ \forall k \in \{1,\ldots,K\}$$

with $\mathcal{E}_k = \frac{\text{\# total amount of data correctly delivered by link } k}{\text{\# total consumed energy on link } k}$ the energy efficiency

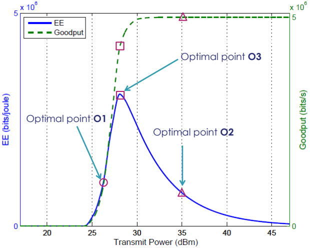
Extensions:

- Cost functions: sum-goodput (MGO), sum-power (MPO)
- QoS constraints: MER, delay, goodput

2.3 - Why Energy Efficiency?

Example 1:

- **O1**: minimum power with goodput constraint ($\geq$ 1Mbits/s)
- **O2**: maximum goodput with power constraint ($\leq$ 35dBm)
- **O3**: maximum energy efficiency



Example 2: Q_r battery state (%), T_t time to transmit the messages (s), N_p number of transmitted messages, and goodput (Mbits/s)

		Q_r	T_t (s)	N_p	Goodput
10^7 sent messages	EE	96	297	10^7	4.3
	MGO	85	256	10^7	5
	MPO	89	1 280	10^7	1
Full battery drain	EE	0	8 327	2.8×10^8	4.3
	MGO	0	1 800	7×10^7	5
	MPO	0	12 180	9.5×10^7	1

2.3 - Practical optimization problem

Type-I HARQ with Rice channel and minimum goodput constraints

$$\max_{\gamma,\mathbf{E}} \sum_{k=1}^{K} \frac{m_k R_k \gamma_k (1 - q_k(G_k E_k))}{\kappa_{1,k}\gamma_k E_k + \kappa_{2,k}}$$

s.t. $m_k R_k \gamma_k (1 - q_k(G_k E_k)) \geq \eta_k^{(0)}$, $\sum_{k=1}^{K} \gamma_k \leq 1$, $\gamma_k \geq 0$, $E_k \geq 0$
with

- $G_k = |A_k|^2 + \varsigma_k^2$
- q_k probability that one frame in error

$$q_k(G_k E_k) \approx a_k \left(b_k \sum_{\ell=1}^{4} c_\ell \frac{e^{-\frac{|A_k|^2 G_k E_k \theta_\ell d_k}{1+\varsigma_k^2 G_k E_k \theta_\ell d_k}}}{1 + \varsigma_k^2 G_k E_k \theta_\ell d_k} \right)^{\delta_k}$$

Remark: Real MCS instead of information-theoretic metrics (like outage probability)

2.3 - How to solve it?

- $f_k : x \mapsto 1 - q_k(G_k x)$ concave
- change of variables $(\gamma_k, E_k) \mapsto (\gamma_k, Q_k)$, then

$$(\gamma_k, Q_k) \mapsto \gamma_k(1 - q_k(G_k Q_k/\gamma_k)) = \gamma_k f_k(Q_k/\gamma_k)$$

  is concave as perspective of f_k

  Consequently, in (γ_k, Q_k)
  - Numerator: concave
  - Denominator: convex (as linear)
  - Constraints set: convex set

Results (fractional programming tool)

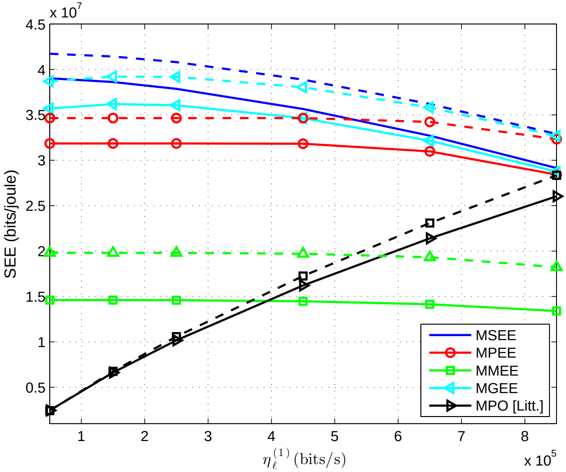
- Jong's algorithm: solve at iteration i (with the above constraints)

$$\max_{\gamma,\mathbf{Q}} \sum_{k=1}^{K} u_k^{(i)} m_k R_k \gamma_k (1 - q_k(G_k Q_k/\gamma_k)) - v_k^{(i)} \kappa_{1,k} Q_k$$

  and update $u_k^{(i)}$ and $v_k^{(i)}$ according to well-defined equations
- KKT can be written in closed-form

2.3 - Numerical results

- $K = 10$ links, Bandwidth $W = 5$ MHz
- QPSK, convolutional code of rate $1/2$
- Rician factor: 10 (dashed line), 0 (solid line)



Future works

- Multi-layer HARQ: why does it work so well?

while multi-layer single-user communications w/o feedback is useless

$$\mathbf{y} = \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{w}$$

does <u>not</u> increase the capacity

$$R = R_1 + R_2 < \log_2(1 + P_1 + P_2) = \log_2(1 + P)$$

with P the transmit power

- Design system with statistical CSI: time-varying Rician factor (joint work with French Thales company on MANET)

- Age of Information (AoI): relationship between HARQ and information refreshness

Our publications devoted to HARQ: a long story

HARQ **improvement:** superposition coding, IP level

[Khreis18a] A. Khreis, P. Ciblat, F. Bassi, P. Duhamel, "Multi-Packet HARQ with delayed feedback," Proc. PIMRC, 2018

[Khreis18b] A. Khreis, P. Ciblat, F. Bassi, P. Duhamel "Throughput-efficient Relay assisted HARQ," Proc. ISWCS, 2018

[Leduc12] C. Le Martret, A. Leduc, S. Marcille, P. Ciblat, "Analytical Performance derivation of HARQ at IP layer," IEEE TCOM, May 2012.

[Leduc11] A. Le Duc, P. Ciblat, C. Le Martret "Analysis of a cross-layer HARQ: application to unequal packet protection," Proc. ICC, 2011

HARQ **parameters' optimization**: URLLC, MCS

[Avranas18] A. Avranas, M. Kountouris, P. Ciblat "Energy-Latency Tradeoff in URLLC with Retransmissions," IEEE JSAC, Nov. 2018

[Ksairi15] N. Ksairi, P. Ciblat "Modulation and Coding schemes selection of HARQ in time-correlated fading channels," IEEE SPAWC, 2015

Resource allocation optimization for HARQ:

[Leturc19] X. Leturc, P. Ciblat, C. Le Martret "Energy-Efficient Resource Allocation for HARQ under Rician channel," IEEE TWC, minor rev.

[Leturc18] X. Leturc, P. Ciblat, C. Le Martret "Energy-Efficient Resource Allocation for HARQ with Statistical CSI," IEEE TVT, Dec. 2018

[Tajan16] R. Tajan, P. Ciblat "Information-theoretic multi-user power adaptation in retransmission schemes," Proc. SPAWC, 2016

[Marcille12] S. Marcille, , P. Ciblat, C. Le Martret, "Resource Allocation for HARQ wireless ad hoc networks," IEEE WC Letters, Dec. 2012.

[Ksairi14] N. Ksairi, P. Ciblat, C. Le Martret, "Near-Optimal Resource Allocation for HARQ-based OFDMA Ad Hoc Networks under rate and power constraints," IEEE TWC, Oct. 2014

Other references

[Gunduz19] E. Ceran, D. Gündüz, A. György “Reinforcement Learning to Minimize Age of Information with an Energy Harvesting Sensor with HARQ and Sensing Cost,” AoI workshop, 2019

[Yates17] Y. Sun, E. Uysal-Biyikoglu, R. Yates, C. Koksal, N. Shroff “Update or wait: How to keep your data fresh,” IEEE TIT, Nov. 2017

[Szczecinski15] M. Jabi, A. E. Hamss, L. Szczecinski, P. Piantanida, “Multipacket HARQ: Closing Gap to the Ergodic Capacity,” IEEE TCOM, Dec. 2015

[Soljanin12] I. Andriyanova, E. Soljanin “Optimized IR-HARQ Schemes Based on Punctured LDPC Codes Over the BEC,” IEEE TIT, Oct. 2012

[Assimi09] A. Assimi, C. Poulliat, I. Fijalkow “Packet combining for multi-layer HARQ over frequency-selective fading channels,” Proc. EUSIPCO, 2009

[Vandendorpe09] L. Vandendorpe et al. , “Subchannel, bit, and power allocation in multi-user OFDM systems for goodput optimization with fairness”, IEEE Symposium WPMC, 2009

[Jindal09] P. Wu and N. Jindal, “Coding Versus ARQ in Fading Channels: How reliable should the PHY be?,” Proc. Globecom, 2009

[Puterman05] M. Puterman, “Markov decision processes,” Wiley, 2005

[Shamai08] A. Steiner and S. Shamai, “Multi-layer broadcasting HARQ strategies for block fading channels,” IEEE TWC, Jul. 2008

[Soljanin07] C. Lott, O. Milenkovic, E. Soljanin “Hybrid ARQ: Theory, state of the art and future directions,” Proc. ITW, 2007