Relevance-based Radio Resource Management for Machine Learning Units

Ph.D. Defense of Afsaneh Gharouni

University of Kaiserslautern-Landau, in cooperation with Nokia Munich

December 14th, 2023

Chairman: Prof. Dr.-Ing. Norbert When

Examiners: Prof. Dr.-Ing. Hans D. Schotten

Prof. Dr.-Ing. Peter Rost (Karlsruhe Institute of Technology)

Part 1

1. Introduction

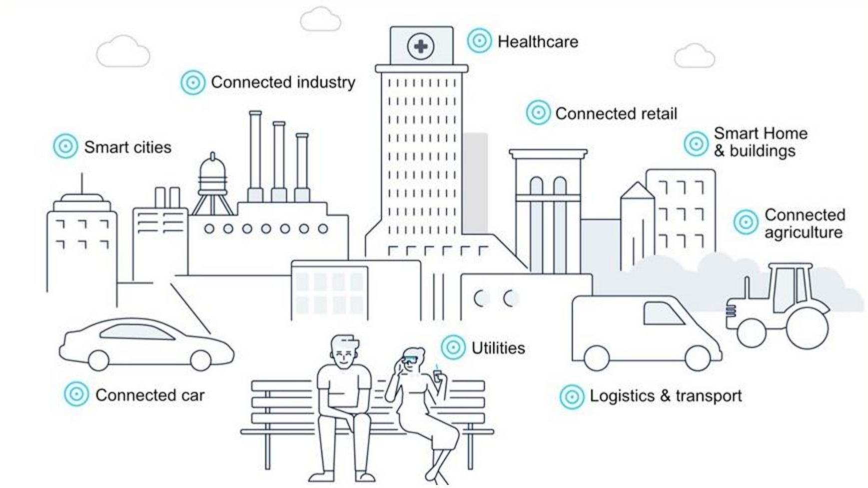
2. Relevance-based Bit Allocation
3. Relevance-based Wireless Resource Allocation
4. Relevance-based Time Domain Signal Overhead Reduction
5. Conclusion and Outlook

- **My contributions**
- Chapter 1 of the dissertation

RPTU NOKIA

Introduction
Motivation

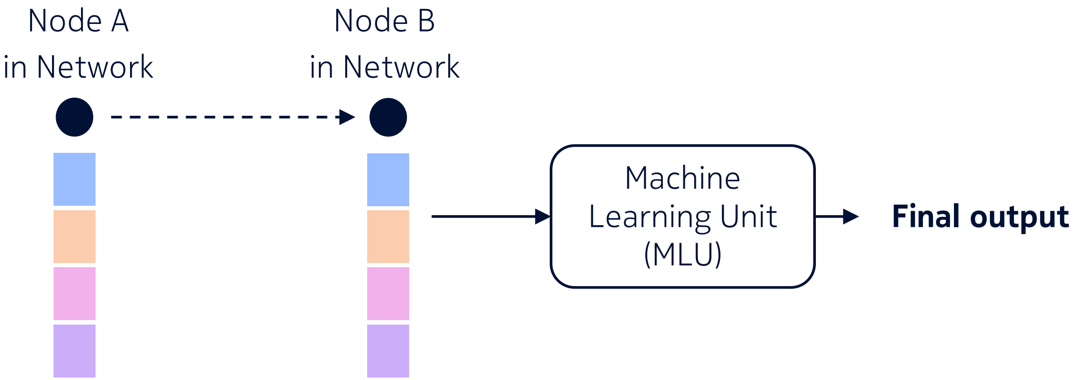
- Traditional communications systems designed to support of **human-to-human communication**.
- Machine Learning (ML) expected to play a key role in 6G.
- Many ML units (MLUs) in the network create burdens and new requirements (e.g. on data transmission and storage).
- Future communications systems designed to support **communication of MLUs**.



Support of MLUs in mobile network **during inference**

Introduction

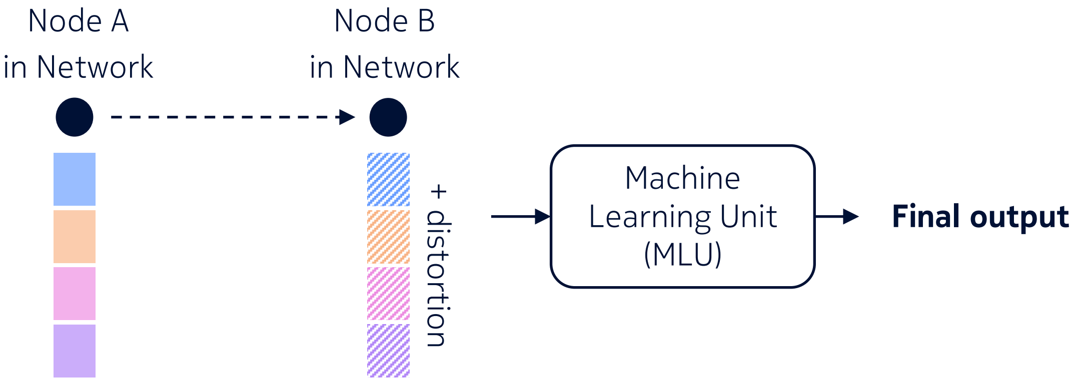
Motivation



~~Communications goal: Delivering **syntax** from A to B~~

Introduction

Motivation



~~Communications goal: Delivering **syntax** from A to B~~

ML can handle distortion at its input → **less distortion** tolerance, **more relevant input** attributes

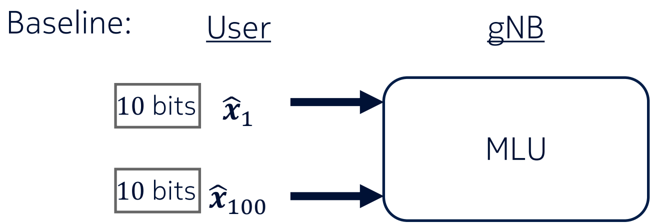
How to measure the MLU input **relevance** such that it can be used by the network?

Introduction

High-level Problem Formulation & Solution

How to measure the MLU input relevance such that it can be used by the network?

A compression technique to map continuous values to discrete values



- High-resolution quantization
- Total: 1000
- Accuracy: 90%

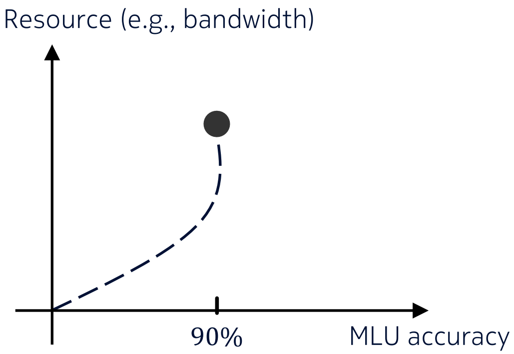
My proposed solution:

- Relevance measurement → quantization **bit allocation**
- Find quantization bit allocations that deliver **sufficient relevant information** to the MLU.

How relevance-based bit allocations can be used by the network?

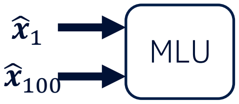
Introduction

Bit Allocation Use A) for Improved Resource Utilization



● 10-bits quantization

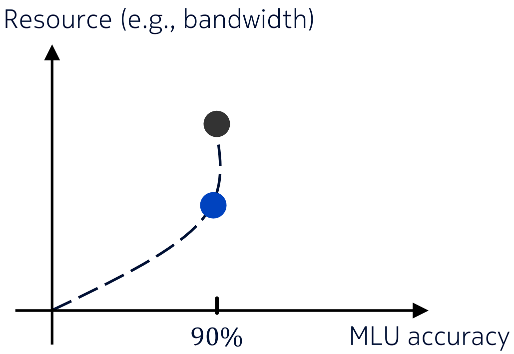
Solution:



- Bit allocation:
  - 1st 50: 10 bits
  - 2nd 50: 10 bit
- 1000 bits
- 90% acc.

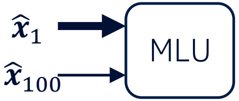
Introduction

Bit Allocation Use A) for Improved Resource Utilization



- 10-bits quantization
- New bit allocation #1

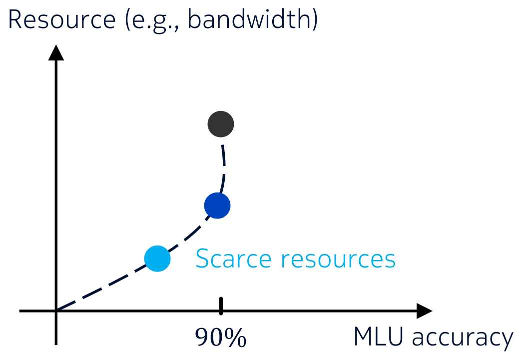
Solution:



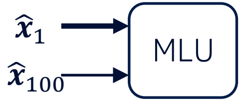
- New bit allocation #1:
  - 1st 50: 10 bits
  - 2nd 50: 1 bit
- 550 bits → 45% gain
- 90% acc.

Introduction

Bit Allocation Use B) for Scarce Resource Utilization



Solution:



- New bit allocation #2:
  - 1st 50: 5 bits
  - 2nd 50: 1 bit
- 300 bits → 70% gain
- 84% acc. → 6% loss

- 10-bits quantization
- New bit allocation #1
- New bit allocation #2 → **best effort performance**

How to find such bit allocations and employ them for use A) and B)?

Introduction

Overview

To address, “How to find such bit allocations and employ them for use A) and B)” and more:

MLU Input Relevance captured in (three domains):

Quantization Bit Allocation

Case study:
Indoor environment classification

Radio Resource Allocation

Case study:
A network of inverted pendulums on carts

Time Domain
Signal Overhead Reduction

Case study:
Conditional handover preparation

Introduction

Overview

To address, “How to find such bit allocations and employ them for use A) and B)” and more:

MLU Input Relevance captured in (three domains):

Quantization Bit Allocation
Part 2

Case study:
Indoor environment classification

Radio Resource Allocation
Part 3

Case study:
A network of inverted pendulums on carts

Time Domain
Signal Overhead Reduction
Part 4

Case study:
Conditional handover preparation

How to find such bit allocations and employ them for improved resource utilization?

Part 2

1. Introduction
2. **Relevance-based Bit Allocation**
3. Relevance-based Wireless Resource Allocation
4. Relevance-based Time Domain Signal Overhead Reduction
5. Conclusion and Outlook

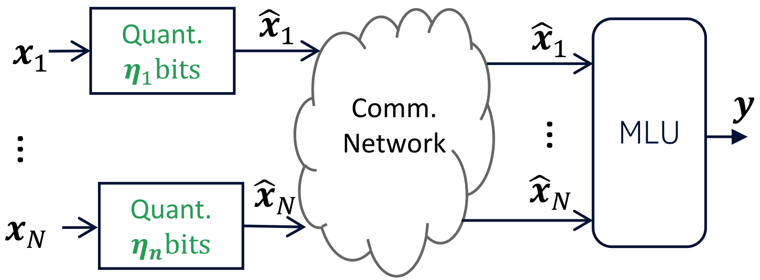
- Chapters 3 and 4 of the dissertation
- Publications [1] and [2]

Quantization Bit Allocation

Indoor environment classification

Relevance-based Bit Allocation

Problem Formulation



Problem:

$$\eta^* = \operatorname*{argmin}_{\eta} d_{\text{rel}}(\hat{x}, y)$$

Subject to constraints on BW, $\eta_n > 0$, …

Assumption: ML parameters are not changing.

$$\cancel{d_{\text{rel}} \to \mathbb{E}\{(x_n - \hat{x}_n)^2\}\ ?}$$

Objectives:

- MLU: black-box
- Applicability: Multiterminal, No Gaussian distribution and independency assumptions, …
- Relevance consideration

$d_{\text{rel}} \to ?$

Relevance-based Bit Allocation

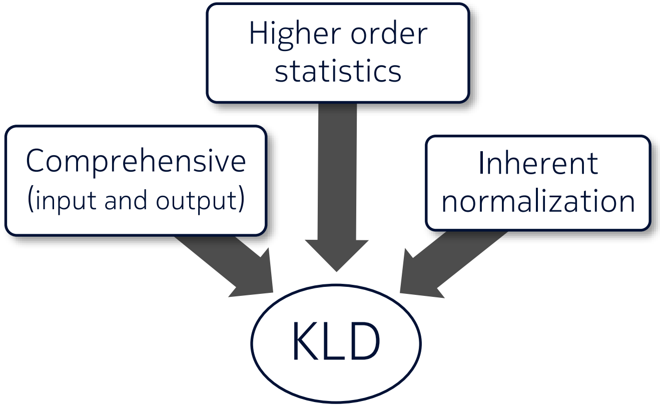
KLD-based Solution (1/3)

$$\boldsymbol{\eta}^* = \underset{\boldsymbol{\eta}}{\operatorname{argmin}}\, D_{\mathrm{KL}}\left(p_{\hat{\mathbf{X}},\mathbf{Y}}(\hat{\boldsymbol{x}},\boldsymbol{y}) || q_{\hat{\mathbf{X}},\mathbf{Y}}(\hat{\boldsymbol{x}},\boldsymbol{y})\right)$$

$D_{\mathrm{KL}}(\cdot\ ||\ \cdot)$: Kullback-Leibler Divergence (KLD)

$p_{\hat{\mathbf{X}},\mathbf{Y}}(\hat{\boldsymbol{x}},\boldsymbol{y})$: Baseline distribution

$q_{\hat{\mathbf{X}},\mathbf{Y}}(\hat{\boldsymbol{x}},\boldsymbol{y})$: The distribution for a bit allocation $\boldsymbol{\eta} = \{\eta_m\}$



My contributions:

- **Different KLD estimators for regression**
- **Issues addressed, e.g.,**
    - The problematic condition on subset containment → with data smoothing.
    - Simplifying the estimation for systems with feedback loop.

Relevance-based Bit Allocation

KLD-based Solution (2/3)

$$\boldsymbol{\eta}^* = \underset{\boldsymbol{\eta}}{\operatorname{argmin}}\, D_{\mathrm{KL}}\left(p_{\widehat{\boldsymbol{X}},\boldsymbol{Y}}(\widehat{\boldsymbol{x}},\boldsymbol{y})||q_{\widehat{\boldsymbol{X}},\boldsymbol{Y}}(\widehat{\boldsymbol{x}},\boldsymbol{y})\right)$$

Works for **low-dimensional** input,
I noticed improvement is needed for **high-dimensional** input:

Syntax → Relevance

$$D_{\mathrm{KL}}(p_{\widehat{\boldsymbol{X}},\boldsymbol{Y}}(\widehat{\boldsymbol{x}},\boldsymbol{y})||q_{\widehat{\boldsymbol{X}},\boldsymbol{Y}}(\widehat{\boldsymbol{x}},\boldsymbol{y})) = \underbrace{D_{\mathrm{KL}}\left(p_{\boldsymbol{Y}|\widehat{\boldsymbol{X}}}(\boldsymbol{y}|\widehat{\boldsymbol{x}})||q_{\boldsymbol{Y}|\widehat{\boldsymbol{X}}}(\boldsymbol{y}|\widehat{\boldsymbol{x}})\right)}_{\text{Relevance-based}} + \underbrace{\cancel{D_{\mathrm{KL}}(p_{\widehat{\boldsymbol{X}}}(\widehat{\boldsymbol{x}})||q_{\widehat{\boldsymbol{X}}}(\widehat{\boldsymbol{x}}))}}_{\text{Syntax-based}}$$

Dominant for
high-dimensional input

$$\approx \mathbb{E}_{\boldsymbol{x}_j}\{d \log(\frac{R_q(\boldsymbol{x}_j)}{R_p(\boldsymbol{x}_j)})\}$$

Relevance-based Bit Allocation

KLD-based Solution (3/3)

$$\boldsymbol{\eta}^* = \underset{\boldsymbol{\eta}}{\operatorname{argmin}}\, D_{\mathrm{KL}}\left(p_{\widehat{\boldsymbol{X}},\boldsymbol{Y}}(\widehat{\boldsymbol{x}},\boldsymbol{y}) || q_{\widehat{\boldsymbol{X}},\boldsymbol{Y}}(\widehat{\boldsymbol{x}},\boldsymbol{y})\right)$$

Works for **low-dimensional** input.

For **high-dimensional** input, I propose using the conditional KLD:

$$\boldsymbol{\eta}^* = \underset{\boldsymbol{\eta}}{\operatorname{argmin}}\, D_{\mathrm{KL}}\left(p_{\boldsymbol{Y}|\widehat{\boldsymbol{X}}}(\boldsymbol{y}|\widehat{\boldsymbol{x}}) || q_{\boldsymbol{Y}|\widehat{\boldsymbol{X}}}(\boldsymbol{y}|\widehat{\boldsymbol{x}})\right)$$

Relevance-based Bit Allocation

Indoor Environment Classification

Selected results:

Various ML hypotheses, codebook designs, benchmarks, etc. investigated.

- The proposed approach → **best** results in **all** studies.
- Significant gains, dependency, e.g., on resource availability
  - **19% gain** in classification accuracy with 13 bits
- No additional sensitivity but higher robustness to packet loss by using the more compressed KLD-based quantization.

Bit allocation	Accuracy (13 bits)	Accuracy (16 bits)
KLD-based (proposed)	$\approx \mathbf{88\%}$	$\approx \mathbf{91\%}$
Equal bits	$\approx 74\%$	$\approx 86\%$
MSE-based	$\approx 69\%$	$\approx 82\%$

*Full-resolution accuracy: 99.5%

How to use KLD−based bit allocations for radio resource allocation **with changing channel quality**?

Part 3

1. Introduction
2. Relevance-based Bit Allocation
3. **Relevance-based Wireless Resource Allocation**
4. Relevance-based Time Domain Signal Overhead Reduction
5. Conclusion and Outlook

- Chapter 6 of the dissertation
- Publication [3]

Radio Resource Allocation

↓

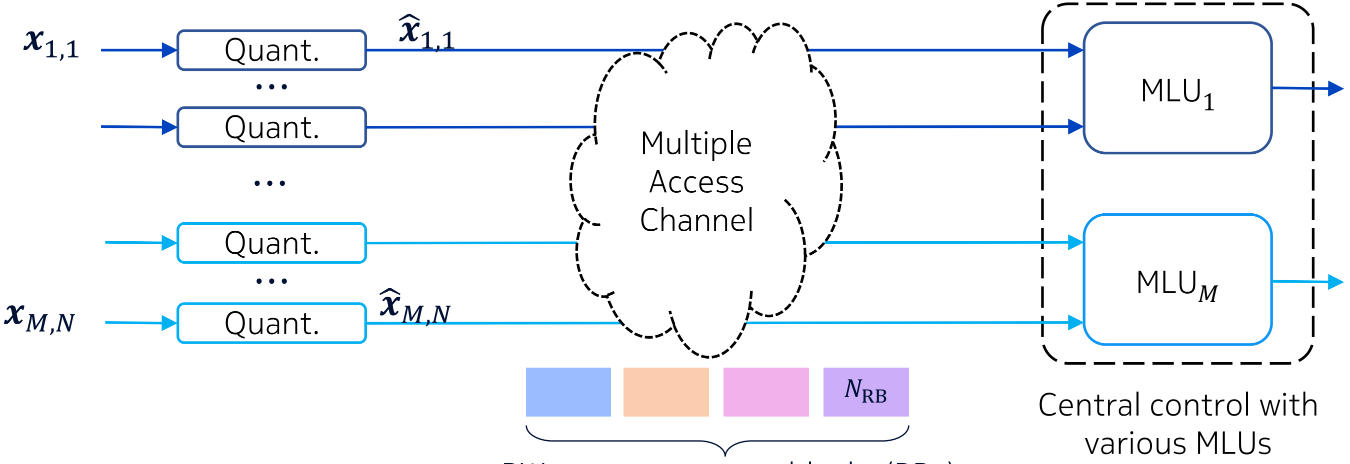
A network of pendulums on carts with a central control system

RPTU NOKIA

Relevance-based Wireless Resource Allocation

Problem

How to use KLD−based bit allocations for radio resource allocation with changing channel quality?



BW → N_{RB} resource blocks (RBs)

- Time and user dependent channel coefficients,
- RB length in time ≤ the coherence time

Relevance-based Wireless Resource Allocation

Problem

- Conventional resource allocation Quality of Service (QoS) → utilities targeting sum rate
- **Relevance-based resource allocation QoS → targets MLU performance:**

$$\kappa^* = \underset{\boldsymbol{\kappa}}{\operatorname{argmin}} \sum_m e_m(\kappa)$$

subject to,

$$C_n \cap C_{n' \neq n} = \emptyset, \forall n,$$

$$\cup_n C_n \subseteq \{1, \cdots, N_{\mathrm{RB}}\},$$

$$\gamma_{n,r} \leq \gamma_{\max}, \forall n, r,$$

$e_m(\kappa) \rightarrow$ error function for MLU m given a feasible resource allocation κ; requires affordable computations and should be relevance-based.

only one source is scheduled on each RB

union of allocated RBs is a subset of available RBs

constrains on transmission power, $\gamma_{n,r}$ is the SNR at nth source on rth RB

~~$e_m(\kappa) \rightarrow$ Usual ML performance metrics~~

$e_m(\kappa) \rightarrow ?$

Relevance-based Wireless Resource Allocation

Solution (1/2)

Introduction of a lookup table per MLU in an offline process:

A lookup table with various KLD-based bit allocations for MLU m

	Total bits	Bit allocation vector/ Payload requirement	KLD
1	300	$[20, 15, \cdots, 30]$	0.01
$\dots$	$\dots$	$\dots$	$\dots$
5	60	$[2, 5, \cdots, 15]$	0.8

If a resource allocation $\boldsymbol{\kappa}$ satisfies one of the payload requirements, $e_m(\kappa) \rightarrow$ **the pre-calculated KLD value**.

$\Rightarrow$ affordable quick computation for error function

A greedy algorithm (GKLD) is proposed to operate online to solve the resource allocation optimization of last slide.

Relevance-based Wireless Resource Allocation
Solution (2/2)

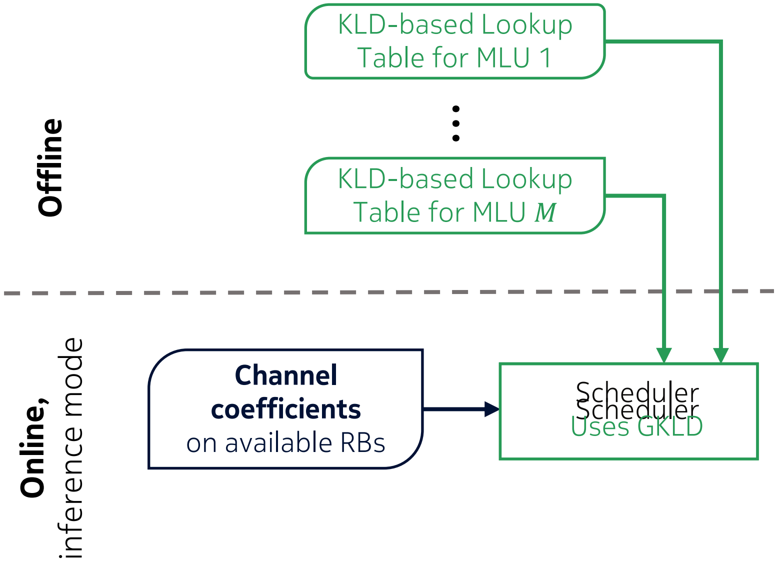
The overview of the proposed solution:

Offline part:

1. Deriving a lookup table per MLU
2. The lookup tables are input for the scheduler.

Online part:

3. The scheduler constantly gets channel coefficients of available resource blocks (RBs).
4. The GKLD uses a **novel QoS** and achieving a best effort MLU performance instead of throughput maximization.



Relevance-based Wireless Resource Allocation

Network of Inverted Pendulums on Carts

Selected results:

Here, benchmark is equal bit assignment lookup tables and scheduler maximizing sum rate.

	Number of RBs	Number of sources	Max SNR (dB)	**Overall steady state error**
Benchmark	64	32	0	0%
KLD lookup tables & GKLD	64	**40**	0	0%

Gain: Serving 8 more sources

Relevance-based Wireless Resource Allocation

Network of Inverted Pendulums on Carts

Selected results:

Here, benchmark is equal bit assignment lookup tables and scheduler maximizing sum rate.

	Number of RBs	Number of sources	Max SNR (dB)	Overall steady state error
Benchmark	64	32	0	0%
KLD lookup tables & GKLD	64	**40**	0	0%
Benchmark	8	8	9	41%
KLD lookup tables & GKLD	8	8	9	**0.25%**

Gain: $\geq 40\%$ less error probability

Relevance-based Wireless Resource Allocation

Network of Inverted Pendulums on Carts

Selected results:

Here, benchmark is equal bit assignment lookup tables and scheduler maximizing sum rate.

Gain: 9 dB

	Number of RBs	Number of sources	Max SNR (dB)	Overall steady state error
Benchmark	64	32	0	0%
KLD lookup tables & GKLD	64	**40**	0	0%
Benchmark	8	8	9	41%
KLD lookup tables & GKLD	8	8	9	**0.25%**
Benchmark	8	8	15	≤ 5%
KLD lookup tables & GKLD	8	8	**6**	≤ 5%

Part 4

1. Introduction
2. Relevance-based Bit Allocation
3. Relevance-based Wireless Resource Allocation
4. **Relevance-based Time Domain Signal Overhead Reduction**
5. Conclusion and Outlook

Time Domain

Signal overhead reduction for conditional handover preparation

- Chapter 5 of the dissertation
- Publication [4]

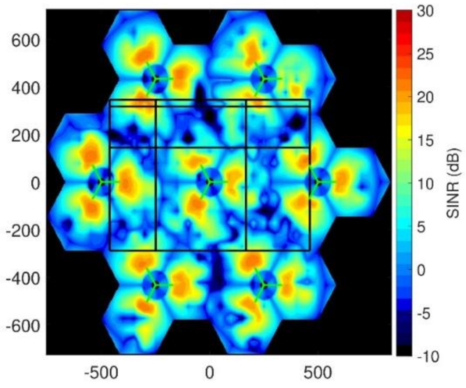
RPTU NOKIA

Relevance-based Time Domain Signal Overhead Reduction (SOR)

Conditional Handover (CHO) Preparation

Should a quantized packet of data be transmitted? → **SOR** classifier

Selected results:



Simulation Layout

	CHO prep	Successful CHOs	Ping Pong	Radio Link Failure	SOR gain
3GPP (Benchmark)	2.91	1.99	0.040	0.19	-
The **SOR** classifier & 3GPP	2.91	1.97	0.039	**0.21**	**28.5%**
The **SOR** classifier & **KLD** bit allocation	**3.51**	1.95	0.035	0.21	**53%**

Part 5

❖ Chapters 7 of the dissertation

RPTU NOKIA

Conclusion

- The proposed framework circumvents syntax and focuses on the semantics/relevance of MLU input during inference.
- Low and high levels of relevance rather than not relevant and relevant input components.
- The proposed approaches deliver the best outcome in all studies:
  - ➢ In many cases, the best outcome implies significant gains.
  - ➢ More significant gains when having limited resources.
- Higher robustness using the proposed bit allocation in presence of packet loss.

Goal achieved: More efficient MLU support by measuring MLU input relevance

Outlook

- Enhanced search algorithms to cope with adaptive scenarios, i.e., non-fixed MLUs.
- Joint optimization of bit allocation, codebook and MLU training.
- Impact of input space partitioning combined with the proposed bit allocation.
- Resource allocation with asynchronized requests

- Impact of other methods for distribution and KLD estimations
- Impact of having MLU trained for dealing with missing values
- Heterogeneous network of MLUs and various priority levels.

List of Publications

1. A. Gharouni, P. Rost, A. Maeder and H. Schotten, "*Impact of Bit Allocation Strategies on Machine Learning Performance in Rate Limited Systems*", IEEE Wireless Communications Letters, vol. 10, no. 6, pp. 1168-1172, June 2021.
2. A. Gharouni, P. Rost, A. Maeder and H. Schotten, "*Divergence-based Bit Allocation for Indoor Environment Classification*", IEEE 7th World Forum on Internet of Things (WF-IoT), pp. 639-644, 2021.
3. A. Gharouni, P. Rost, A. Maeder and H. Schotten, "*Relevance-Based Wireless Resource Allocation for a Machine Learning-Based Centralized Control System*", IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2021.
4. A. Gharouni, U. Karabulut, A. Enqvist, P. Rost, A. Maeder and H. Schotten, "*Signal Overhead Reduction for AI-Assisted Conditional Handover Preparation*", Mobile Communication - Technologies and Applications; 25th ITG-Symposium, Osnabrueck, November 2021.

Rheinland-Pfälzische
Technische Universität
Kaiserslautern
Landau