Theory of Transformer-Based In-Context Learning for Channel Access Optimization

Presenter: Shugang Hao

Pillar of Engineering Systems and Design
Singapore University of Technology and Design

24 November 2025

Acknowledgment

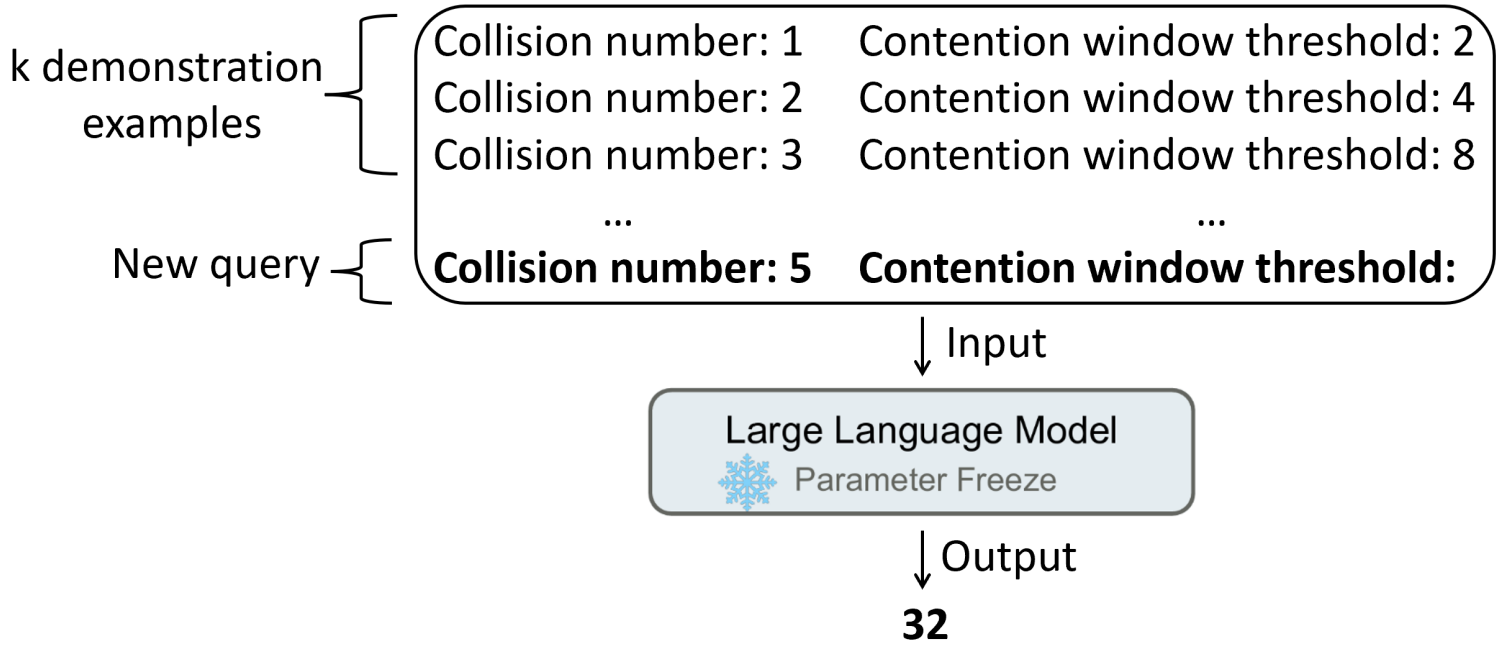
A joint work with Dr. Hongbo Li and Prof. Lingjie Duan.

Shugang Hao, Hongbo Li*, and Lingjie Duan, "To Theoretically Understand Transformer-Based In-Context Learning for Optimizing CSMA," In Proc. of ACM MobiHoc 2025, 261–270, https://doi.org/10.1145/3704413.3764423.*

Outline

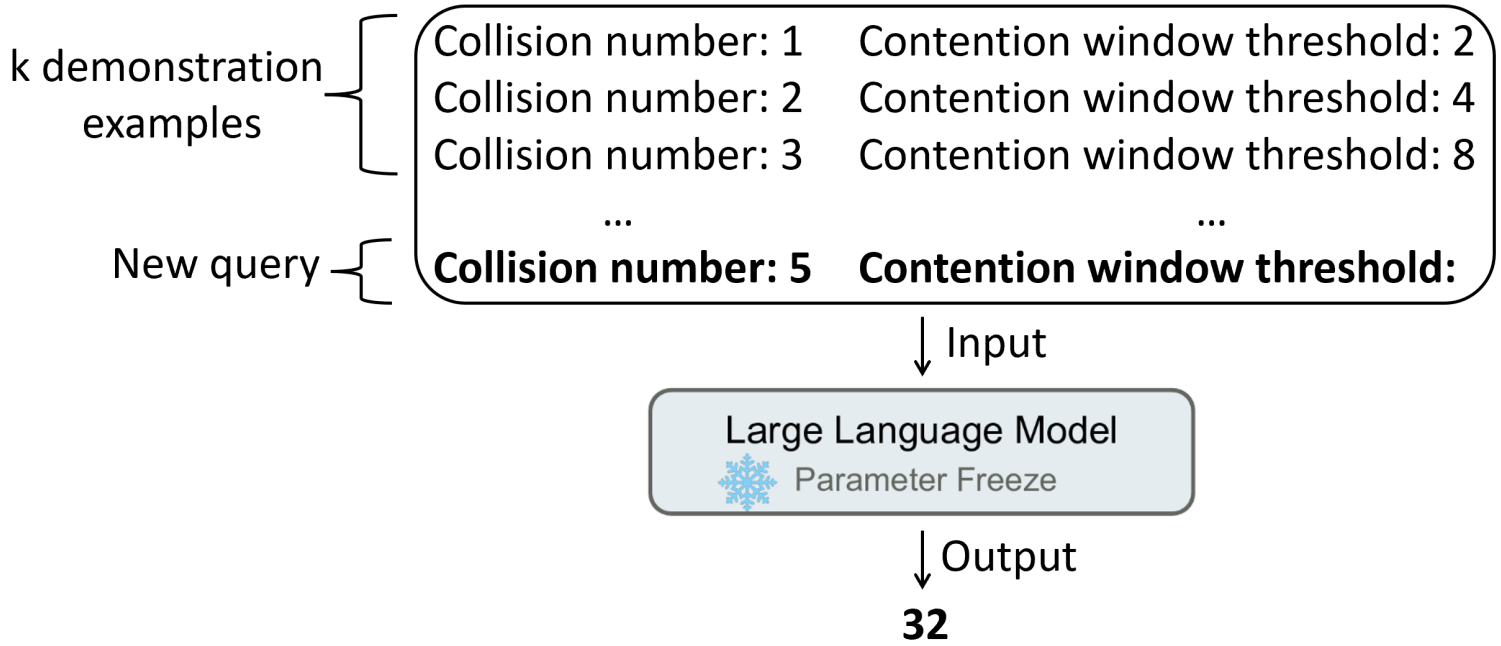
1. Background and Motivation
2. Our Transformer-Based ICL Optimizer Design
3. Theoretical Analysis of Our Transformer-Based ICL Optimizer
4. Experiments

Background. In-Context Learning (ICL)



An LLM can infer binary exponential backoff without fine-tuning.

Background. In-Context Learning (ICL)

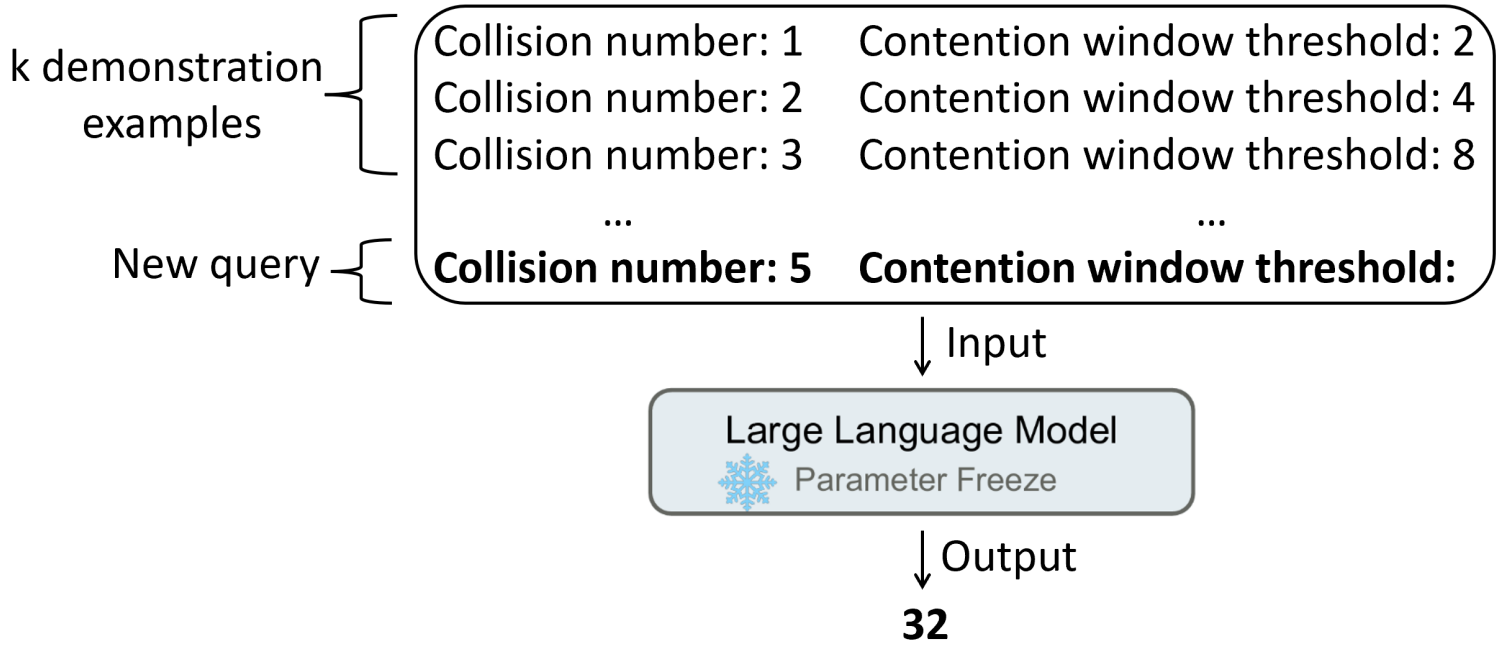


An LLM can infer binary exponential backoff without fine-tuning.

In-Context Learning

i) A language model learns new tasks by a few examples in its prompt.
ii) Leverage learned patterns and attentions without fine-tuning.

Background. In-Context Learning (ICL)



An LLM can infer binary exponential backoff without fine-tuning.

In-Context Learning

i) A language model learns new tasks by a few examples in its prompt.
ii) Leverage learned patterns and attentions without fine-tuning.

Q. How to leverage ICL for resource allocation?

Channel Throughput Maximization Problem

Consider the following channel throughput maximization problem

$$\max_{\{W_k\}_{k=0}^{K}} \; U(\{W_k\}_{k=0}^{K}).$$

- k: current collision number; W_k: contention window (CW) threshold,
- K: maximum collision number for each packet transmission.

-Nicola Cordeschi et al. "Optimal Back-Off Distribution for Maximum Weighted Throughput in CSMA," IEEE TON, 2024.

-Lin Dai, "A theoretical framework for random access: Stability regions and transmission control," IEEE TON, 2022.

-Yayu Gao et al., "When Aloha and CSMA coexist: Modeling, fairness, and throughput optimization," IEEE TWC, 2022.

Channel Throughput Maximization Problem

Consider the following channel throughput maximization problem

$$\max_{\{W_k\}_{k=0}^{K}} \; U(\{W_k\}_{k=0}^{K}).$$

- k: current collision number; W_k: contention window (CW) threshold,
- K: maximum collision number for each packet transmission.

Model-Based CSMA (*Cordeschi et al. (2024), Dai (2022), Gao et al. (2022)*)

- assume a fixed and known node density N, and
- derive the exact throughput formulation to obtain the CW thresholds.

-Nicola Cordeschi et al. "Optimal Back-Off Distribution for Maximum Weighted Throughput in CSMA," IEEE TON, 2024.

-Lin Dai, "A theoretical framework for random access: Stability regions and transmission control," IEEE TON, 2022.

-Yayu Gao et al., "When Aloha and CSMA coexist: Modeling, fairness, and throughput optimization," IEEE TWC, 2022.

Channel Throughput Maximization Problem

Consider the following channel throughput maximization problem

$$\max_{\{W_k\}_{k=0}^{K}} \; U(\{W_k\}_{k=0}^{K}).$$

- k: current collision number; W_k: contention window (CW) threshold,
- K: maximum collision number for each packet transmission.

Model-Based CSMA (*Cordeschi et al. (2024), Dai (2022), Gao et al. (2022)*)

- assume a fixed and known node density N, and
- derive the exact throughput formulation to obtain the CW thresholds.

Q. How bad can model-based CSMA be under unknown node densities?

-Nicola Cordeschi et al. "Optimal Back-Off Distribution for Maximum Weighted Throughput in CSMA," IEEE TON, 2024.

-Lin Dai, "A theoretical framework for random access: Stability regions and transmission control," IEEE TON, 2022.

-Yayu Gao et al., "When Aloha and CSMA coexist: Modeling, fairness, and throughput optimization," IEEE TWC, 2022.

Certain Throughput Loss by Model-Based CSMA

Theorem 1. Certain Throughput Loss by Model-Based CSMA

Suppose that $\bar{W} \geq 2\bar{N} - 1$. In a dynamic channel environment with an unknown node density $N(t) \leq \bar{N}$, the model-based approach with inaccurate estimation $\hat{N}(t)$ of $N(t)$ leads to a certain throughput loss

$$U(N(t)) - U(\hat{N}(t)) \geq \Theta\left(\frac{T_\sigma}{\bar{N}K^2\bar{W}^3T_c^2}\right)|N(t) - \hat{N}(t)|.$$

Certain Throughput Loss by Model-Based CSMA

Theorem 1. Certain Throughput Loss by Model-Based CSMA

Suppose that $\bar{W} \geq 2\bar{N} - 1$. In a dynamic channel environment with an unknown node density $N(t) \leq \bar{N}$, the model-based approach with inaccurate estimation $\hat{N}(t)$ of $N(t)$ leads to a certain throughput loss

$$U(N(t)) - U(\hat{N}(t)) \geq \Theta\left(\frac{T_\sigma}{\bar{N}K^2\bar{W}^3T_c^2}\right)|N(t) - \hat{N}(t)|.$$

Need an adaptive design for dynamic wireless environments!

Related Work. DRL-Based CSMA

DRL-Based CSMA

i) To propose deep reinforcement learning (DRL) based approaches.
ii) To solve the CW thresholds under unknown channel environments.

-Witold Wydmański and Szymon Szott, "Contention window optimization in IEEE 802.11 ax networks with deep reinforcement learning,"IEEE WCNC 2021.

-Rong Yan et al., "Deep Reinforcement Learning Based Contention Window Optimization for IEEE 802.11 bn," In IEEE VTC2024-Spring.

-Chanho Lee et al., "Dynamic-Persistent CSMA: A Reinforcement Learning Approach for Multi-User Channel Access," IEEE Access 2024.

Related Work. DRL-Based CSMA

DRL-Based CSMA

i) To propose deep reinforcement learning (DRL) based approaches.
ii) To solve the CW thresholds under unknown channel environments.

Wydmanski et al. in WCNC (2021), Yan et al. in VTC (2024), Lee et al. (2024)

- need to retrain from scratch once the channel environment changes,
- inherits heavy training and inference costs,
- hard to implement on resource-constrained wireless devices.

-Witold Wydmański and Szymon Szott, "Contention window optimization in IEEE 802.11 ax networks with deep reinforcement learning,"IEEE WCNC 2021.

-Rong Yan et al., "Deep Reinforcement Learning Based Contention Window Optimization for IEEE 802.11 bn," In IEEE VTC2024-Spring.

-Chanho Lee et al., "Dynamic-Persistent CSMA: A Reinforcement Learning Approach for Multi-User Channel Access," IEEE Access 2024.

Related Work. DRL-Based CSMA

DRL-Based CSMA

i) To propose deep reinforcement learning (DRL) based approaches.
ii) To solve the CW thresholds under unknown channel environments.

Wydmanski et al. in WCNC (2021), Yan et al. in VTC (2024), Lee et al. (2024)

- need to retrain from scratch once the channel environment changes,
- inherits heavy training and inference costs,
- hard to implement on resource-constrained wireless devices.

Q. How to achieve efficient and adaptive channel throughput optimization?

-Witold Wydmański and Szymon Szott, "Contention window optimization in IEEE 802.11 ax networks with deep reinforcement learning,"IEEE WCNC 2021.

-Rong Yan et al., "Deep Reinforcement Learning Based Contention Window Optimization for IEEE 802.11 bn," In IEEE VTC2024-Spring.

-Chanho Lee et al., "Dynamic-Persistent CSMA: A Reinforcement Learning Approach for Multi-User Channel Access," IEEE Access 2024.

Related Work. Theoretical Guarantees on ICL

There are few studies on theoretical guarantees of ICL prediction.

Zhang et al. in JMLR (2024), Huang et al. in ICML (2024), Li et al. in ICML (2024)

- consider either binary outputs/labels,
- or a linear mapping function between the input and the label.

-Ruiqi Zhang et al., "Trained transformers learn linear models in-context," JMLR, 2024.

-Yu Huang et al., "In-context convergence of transformers," ICML 2024.

-Hongkang Li et al., "How Do Nonlinear Transformers Learn and Generalize in In-Context Learning?" ICML 2024.

Related Work. Theoretical Guarantees on ICL

There are few studies on theoretical guarantees of ICL prediction.

Zhang et al. in JMLR (2024), Huang et al. in ICML (2024), Li et al. in ICML (2024)

- consider either binary outputs/labels,
- or a linear mapping function between the input and the label.

However, in channel throughput optimization,

- CW thresholds are non-binary integers,
- mapping between collision number and its CW threshold is non-linear.

-Ruiqi Zhang et al., "Trained transformers learn linear models in-context," JMLR, 2024.

-Yu Huang et al., "In-context convergence of transformers," ICML 2024.

-Hongkang Li et al., "How Do Nonlinear Transformers Learn and Generalize in In-Context Learning?" ICML 2024.

Related Work. Theoretical Guarantees on ICL

There are few studies on theoretical guarantees of ICL prediction.

Zhang et al. in JMLR (2024), Huang et al. in ICML (2024), Li et al. in ICML (2024)

- consider either binary outputs/labels,
- or a linear mapping function between the input and the label.

However, in channel throughput optimization,

- CW thresholds are non-binary integers,
- mapping between collision number and its CW threshold is non-linear.

The applicability of ICL to real-world CSMA scenarios remains unclear.

-Ruiqi Zhang et al., "Trained transformers learn linear models in-context," JMLR, 2024.

-Yu Huang et al., "In-context convergence of transformers," ICML 2024.

-Hongkang Li et al., "How Do Nonlinear Transformers Learn and Generalize in In-Context Learning?" ICML 2024.

Technical Contributions

New theory of transformer-based ICL for optimizing channel access:

- adaptive and NOT requiring the node-density information.

Transformer-based ICL design with performance guarantee:

- guarantee convergence to the optimum within limited training steps.

NS-3 experimental comparisons under unknown node densities:

- our approach's fast convergence and near-optimal throughput.

Technical Contributions

New theory of transformer-based ICL for optimizing channel access:

- adaptive and NOT requiring the node-density information.

Transformer-based ICL design with performance guarantee:

- guarantee convergence to the optimum within limited training steps.

NS-3 experimental comparisons under unknown node densities:

- our approach's fast convergence and near-optimal throughput.

Technical Contributions

New theory of transformer-based ICL for optimizing channel access:

- adaptive and NOT requiring the node-density information.

Transformer-based ICL design with performance guarantee:

- guarantee convergence to the optimum within limited training steps.

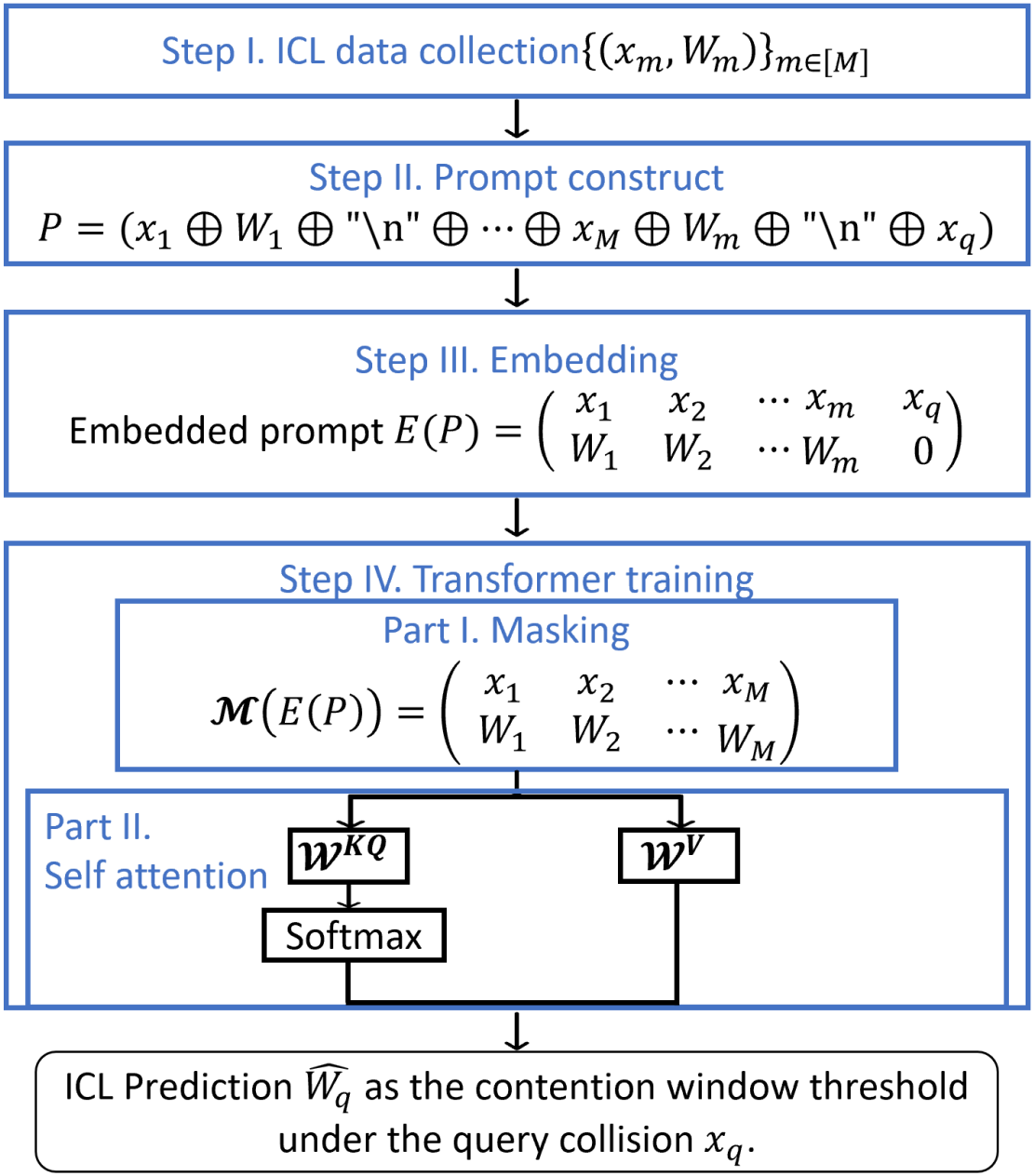
NS-3 experimental comparisons under unknown node densities:

- our approach's fast convergence and near-optimal throughput.

Outline

1. Background and Motivation
2. Our Transformer-Based ICL Optimizer Design
3. Theoretical Analysis of Our Transformer-Based ICL Optimizer
4. Experiments

Overview of Transformer-Based ICL Optimizer



Step I. ICL Data Collection

Step I. ICL data collection $\{(x_m, W_m)\}_{m\in[M]}$

We construct M collision-threshold examples $\{(x_m, W_m)\}_{m\in[M]}$, where

- x_m: a collision vector to contain the current collision number m,
- W_m: the corresponding optimal CW threshold for x_m.

These examples are obtained through historical data.

Step II. Prompt Construction

Step II. Prompt construct

$$P = (x_1 \oplus W_1 \oplus \text{"\textbackslash n"} \oplus \cdots \oplus x_M \oplus W_m \oplus \text{"\textbackslash n"} \oplus x_q)$$

- $x_m \oplus W_m$, $m \in [M]$: a collision-threshold example,
- x_q: new query collision case.

In each prompt P, all $\{(x_m, W_m)_{m\in[M]}\}$ follow the same mapping function

$$f : x_m \to W_m.$$

Step III. Embedding & Step IV.1. Masking

Step III. Embedding

$$\text{Embedded prompt } E(P) = \begin{pmatrix} x_1 & x_2 & \cdots & x_m & x_q \\ W_1 & W_2 & \cdots & W_m & 0 \end{pmatrix}$$

Step IV. Transformer training

Part I. Masking

$$\mathcal{M}\big(E(P)\big) = \begin{pmatrix} x_1 & x_2 & \cdots & x_M \\ W_1 & W_2 & \cdots & W_M \end{pmatrix}$$

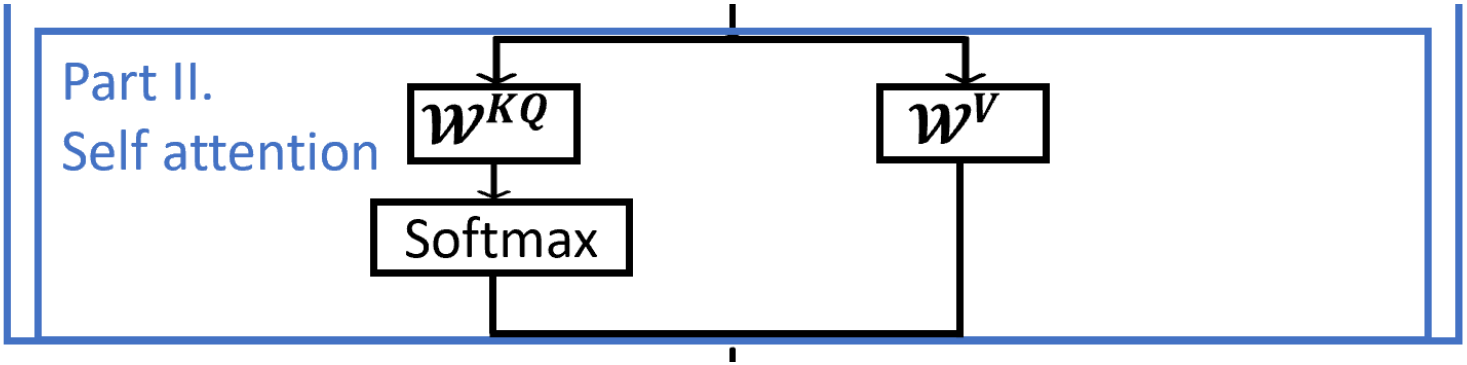
Embedding:

- set the label of new query x_q as 0,

Masking:

- removes the last column to prevent query from attending to itself.

Step IV.2 Self-Attention and ICL Prediction



We consider a one-layer transformer with the self-attention mechanism.

The self-attention mechanism in the parameter $\theta := (1, Q)$ is

$$F_{\mathrm{SA}}(E(P);\theta) = \mathcal{M}(E^{W}(P)) \cdot \mathrm{softmax}\left(\mathcal{M}(E^{x}(P))^{\top} Q E^{x}(P)\right),$$

$E^{x}(P)$: the first d rows of $E(P)$; $E^{W}(P)$: the last row of $E(P)$.

The ICL prediction for the query collision x_q is the last entry of F_{SA}:

$$\hat{W}_q = \hat{W}_q(E(P);\theta) = [F_{\mathrm{SA}}(E(P);\theta)]_{(M+1)}.$$

Highly Non-Convex Transformer Training Problem

We aim to minimize the following squared loss of the prediction error:

$$\mathcal{L}(\theta) = \mathbb{E}_{f \in \mathcal{F}}[(\hat{W}_q - W_q)^2].$$

- $\hat{W}_q$: our ICL prediction,
- W_q: ground-truth/optimal CW threshold for query collision x_q.

Non-linear softmax function couples attention weights across input tokens,

- making $\mathcal{L}(\theta)$ highly non-convex and interdependent.

We adopt a simple algorithm to train the transformer

$$\theta^{(t+1)} = \theta^{(t)} - \eta \cdot \nabla_\theta \mathcal{L}(\theta^{(t)}).$$

Q. How to guarantee convergence to the optimum within limited training steps?

Outline

1. Background and Motivation
2. Our Transformer-Based ICL Optimizer Design
3. Theoretical Analysis of Our Transformer-Based ICL Optimizer
4. Experiments

Non-Linear Mapping Functions

Different from the ICL literature assuming linearity,

- we consider general non-linear mapping functions.

Assumption 1. Non-Degenerate L-Lipschitz Functions

Each mapping function f is L-Lipschitz continuous, i.e.,

$$|f(x) - f(x')| \leq L\|x - x'\|, \forall \|x - x'\| = \Theta(\delta_0),$$

where $L > 0$ and $\delta_0 = \mathcal{O}(1)$. Moreover, for every collision vector $x_k \in \mathcal{X}$, there exists some other collision vector $x_{k'} \in \mathcal{X}$ with $k' \neq k$, such that

$$|f(x_k) - f(x_{k'})| = \Theta(L) \cdot \|x_k - x_{k'}\|.$$

Reflects the practical requirement in wireless systems that

- adjustments to the CW respond gradually to changes in collisions.
- different CW thresholds remain distinguishable.

Attention Scores

To derive the optimal ICL prediction for the query collision, we need to

- build its connection with any other collision example in the prompt.

The attention score for the m-th collision input x_m^s is defined as

$$\mathrm{attn}_m^{(t)}(\theta^{(t)}; E(P)) := \left[\mathrm{softmax}\left(\mathcal{M}(E^x(P))^\top Q E^x(P)\right)\right]_m.$$

$\mathcal{X}_k(P) \subset [M]$: the index set of collision input $x_m = x_k$ for $m \in \mathcal{X}_k(P)$.

Then the attention score for the k-th collision case is given by

$$\mathrm{Attn}_k^{(t)}(\theta^{(t)}; E(P)) := \sum_{m \in \mathcal{X}_k(P)} \mathrm{attn}_m^{(t)}(\theta^{(t)}; E(P)).$$

Attention Scores and ICL Prediction

Attention scores measure

- how much the new query collision case x_q of P attends
- or relates to other input collision parameter vectors $\{x_m\}_{m\in[M]}$.

For simplicity, we represent

- $\mathrm{attn}_m^{(t)}(\theta^{(t)}; E^s(P^s))$ as $\mathrm{attn}_m^{(t)}$,
- $\mathrm{Attn}_k^{(t)}(\theta^{(t)}; E^s(P^s))$ as $\mathrm{Attn}_k^{(t)}$.

The ICL prediction $\hat{W}_q^s$ for the current query case x_q^s can be rewritten as

$$\hat{W}_q = \sum_{m\in[M]} \mathrm{attn}_m^{(t)} W_m = \sum_{k\in[K]} \mathrm{Attn}_k^{(t)} f(x_k).$$

An Explicit Formulation of Prediction Loss

Lemma 2. An Explicit Formulation of Prediction Loss

Given constants $L, \Delta > 0$, for any $t \in [T]$, the prediction loss can be expressed as:

$$\mathcal{L}(\theta) = \sum_{k=1}^{K} \mathbb{E}\Big[\mathbf{1}\{x_q = x_k\}\Big(1 - \mathrm{Attn}_k^{(t)}\Big)^2 \Theta(L^2\Delta^2)\Big],$$

where $\mathbf{1}\{x_q = x_k\}$ is the indicator function that equals 1 if $x_q = x_k$.

If the query collsion $x_q = x_k$,

- it is equivalent to prove $\mathrm{Attn}_k^{(t)} \to 1$ for convergence to the optimum.

Convergence to The Optimum

T^*: convergence time.

Proposition 1. A Necessary and Sufficient Condition for Convergence

Given $\delta_0 = \mathcal{O}(1)$, our Algorithm 1 achieves convergence if and only if, for any $t > T^*$ with query collision case $x_q = x_k$, the attention score with respect to x_k satisfies

$$1 - \text{Attn}_k^{(t)} = \mathcal{O}(\delta_0).$$

Theorem 2. Convergence of Our ICL Optimizer

By our gradient descent method, with at most $T^* = \Theta(\frac{K \log(K\epsilon^{-1})}{\eta \delta_0^2 L^2 \Delta^2})$ iterations, we obtain $\text{Attn}_k^{(T^*)} = \Omega(\frac{1}{1+\delta_0 \epsilon})$ and the prediction loss satisfies $\mathcal{L}\left(\theta^{(T^*)}\right) = \mathcal{O}(\epsilon^2)$.

Adaptiveness to Unknown Node Densities

Corollary 1. Adaptiveness to Unseen Node Densities

After the prediction loss $\mathcal{L}(\theta)$ converges to $\mathcal{L}(\theta) = \mathcal{O}(\epsilon^2)$, for any unseen mapping function $f^s(x) \in \mathcal{F}$ under a node density s in training, if query collision satisfies $x_q^s = x_k$, we have $(\hat{W}_q^s - W_q^s)^2 = \mathcal{O}(\epsilon^2)$.

Our ICL optimizer can adapt to unknown node densities in practice.

Theoretical Guarantees on Channel Throughput

We define the throughput loss ΔU as the expected difference between

- the throughput U^* under the optimal CW threshold W_q
- and $\hat{U}$ of the ICL prediction $\hat{W}_q$ for all the prompts:

$$\Delta U := U^* - \hat{U} = \mathbb{E}_{f \in \mathcal{F}}\big[U(W_q) - U(\hat{W}_q)\big].$$

Theorem 3. Theoretical Guarantees on Channel Throughput

The throughput loss ΔU of our transformer-based ICL approach from the optimum is upper bounded as follows:

$$\Delta U \leq \mathcal{O}\left(\frac{T_P \bar{N} \epsilon}{8 T_\sigma}\right).$$

Extension to Erroneous Data Input

In practice, it is difficult to collect the optimal CWT for each collision case.

Then, we allow each W_m to be erroneous for collision x_m for evaluation.

Theorem 4. (Informal) Theoretical Guarantee with Erroneous Data

If in-context example number

$$M \geq \max\left\{\frac{(\ln\frac{1}{q})(16\ell^2)(\ln^2\beta)}{KL^2(\mathbb{P}_f,\mathbb{P}_{f^*})}, \frac{-2\ln(\frac{\epsilon}{2(1-\frac{c}{2})^{-1}\alpha^{-2}\beta^{-T}\gamma^{-1}})}{\min_f KL(\mathbb{P}_f,\mathbb{P}_{f^*})+8\ln(\alpha\beta)}\right\},$$

we have the throughput loss ΔU due to an erroneous prompt is:

$$Pr\left(\Delta U \leq \mathcal{O}\left(\frac{T_P\bar{N}}{8T_\sigma}\epsilon^{\frac{1}{2}}\bar{W}\right)\right) \geq 1-q.$$

Outline

1. Background and Motivation
2. Our Transformer-Based ICL Optimizer Design
3. Theoretical Analysis of Our Transformer-Based ICL Optimizer
4. Experiments

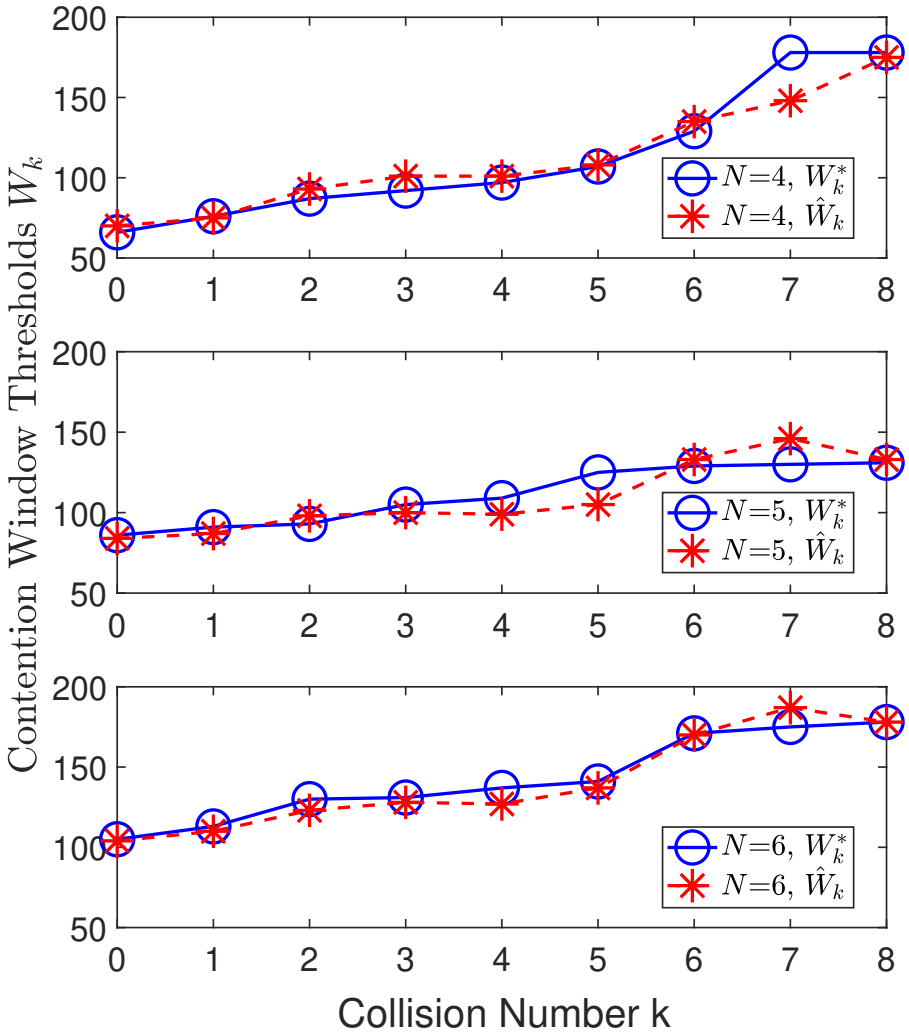
Simulation Settings

We fix the CBR as 1Mbps and use network settings (unit of μs):

T_σ	T_{DIFS}	T_{SIFS}	T_δ	T_{ACK}	T_{Header}	T_{P}	T_s	T_c
50	128	28	1	240	400	8184	8982	8783

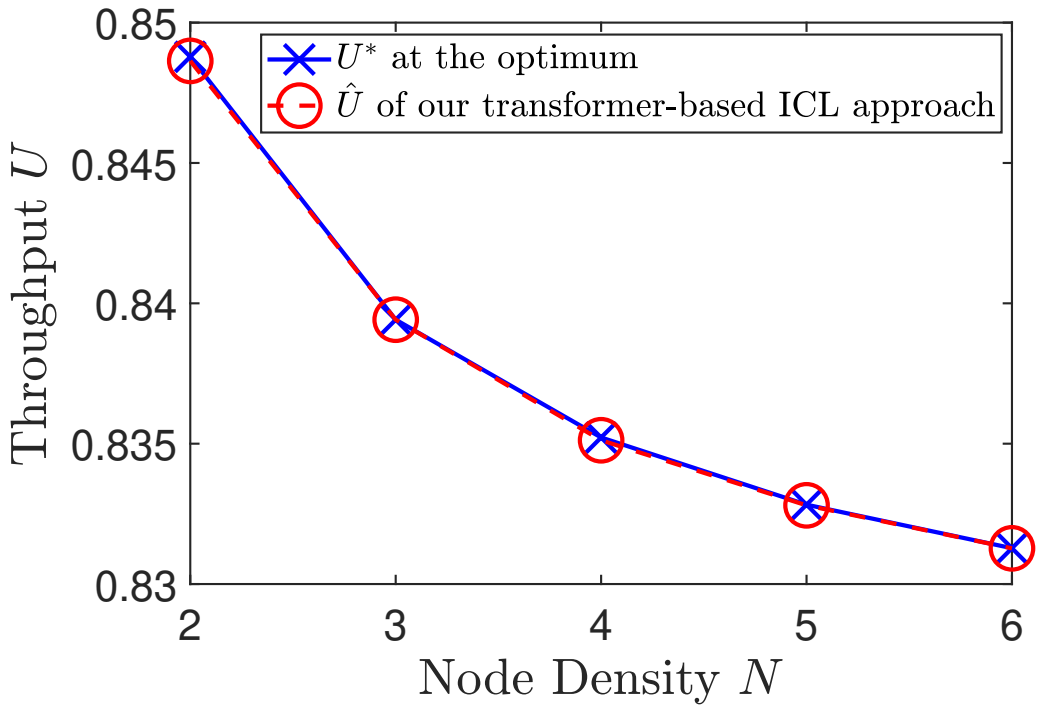
We choose $N \in \{2, 3, 4, 5, 6\}$ to construct the training prompts.

Simulation Results in the Training Stage (1)



- Near-optimal approximations for most collision cases.

Simulation Results in the Training Stage (2)

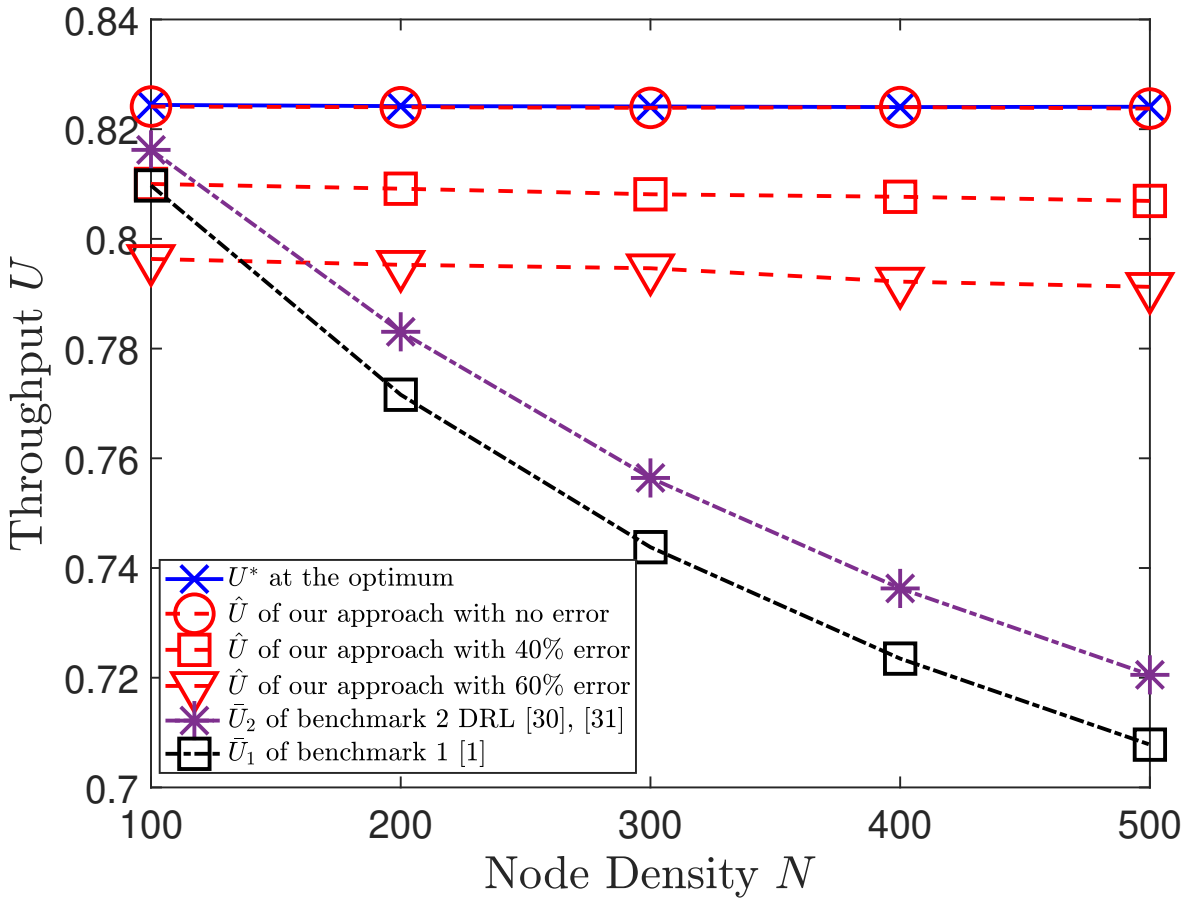


Near-optima throughput

- due to good approximations of CWs for all the collision cases.

Simulation Results in the Testing Stage

We fix the node density as $N = 50$ for benchmarks to test the adaptiveness.

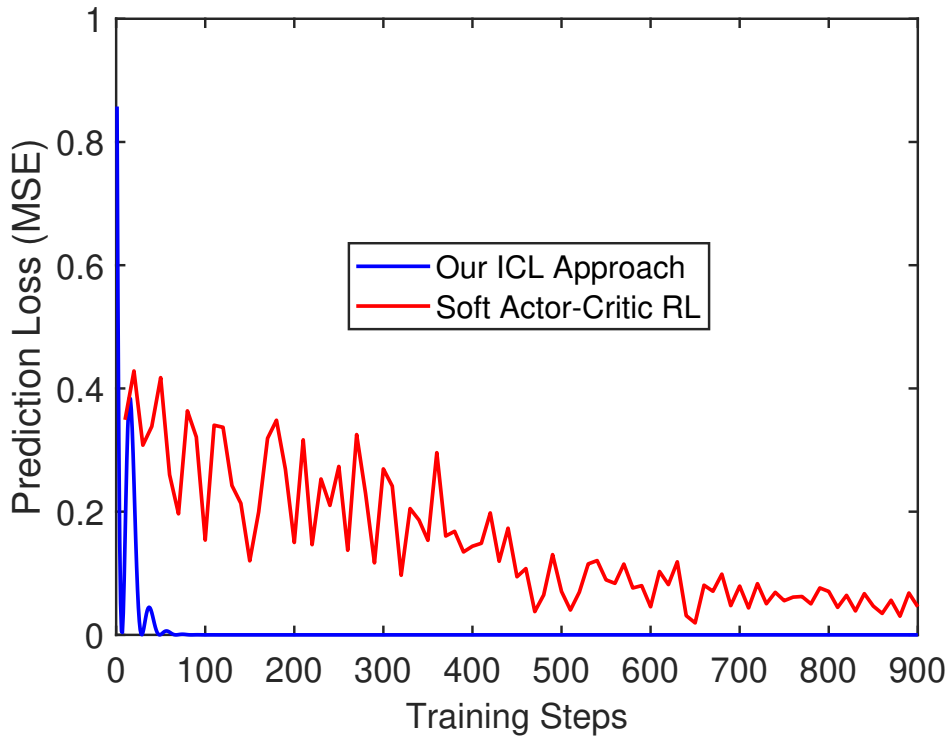


Smallest loss of our ICL optimizer even with highly erroneous prompts.

More Practical NS-3 Experiments (1)

NS-3.38: wireless devices

- are randomly placed on a 20m-radius circle around an AP,
- and send UDP packets every 1 ms to a server on the AP.

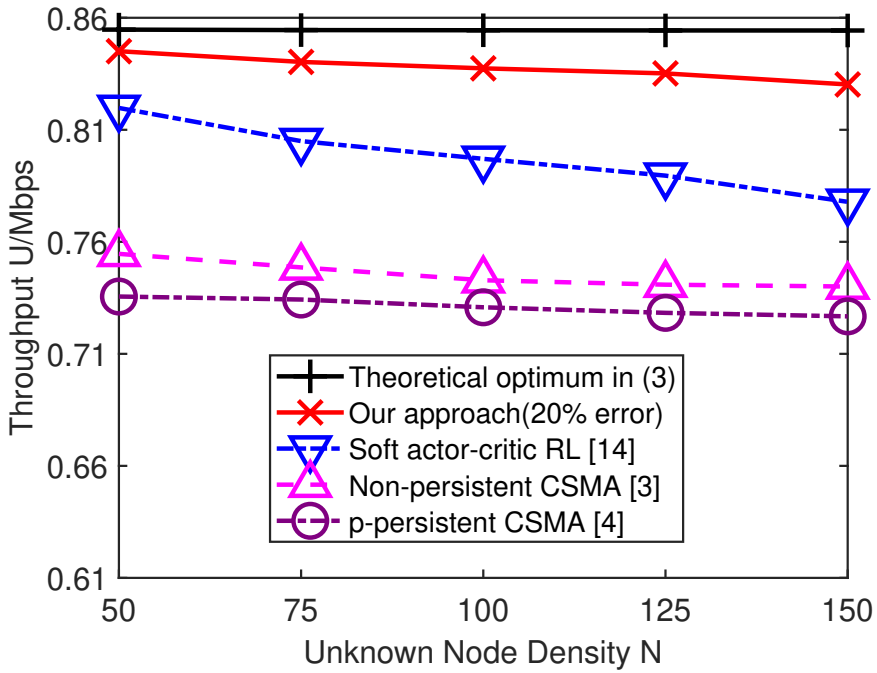


Our ICL optimizer converges within 100 training steps

- much faster than Soft Actor-Critic RL (more than 650 steps.)

More Practical NS-3 Experiments (2)

We further compare our ICL optimizer to CSMA work within 3 years.



- Smallest throughput loss of our ICL optimizer.

---

[3] Cordeschi et al. "Optimal Back-Off Distribution for Maximum Weighted Throughput in CSMA," IEEE TON, 2024.

[4] Lin Dai, "A Theoretical Framework for Random Access: Effects of Carrier Sensing on Stability," IEEE TCOM 2024.

[14] Chanho Lee et al., "Dynamic-Persistent CSMA: A Reinforcement Learning Approach for Multi-User Channel Access," IEEE Access 2024.

Conclusion

New theory of transformer-based ICL for optimizing channel access:

- adaptive and NOT requiring the node-density information.

Transformer-based ICL design with performance guarantee:

- guarantee convergence to the optimum within limited training steps.

NS-3 experimental comparisons under unknown node densities:

- our approach's fast convergence and near-optimal throughput.