Reducing the Environmental Impact of Wireless Communication via Probabilistic Machine Learning

A. Ryo Koblitz, Lorenzo Maggi, Matthew Andrews
Nokia Bell Labs
ryo.koblitz@nokia-bell-labs.com, lorenzo.maggi@nokia-bell-labs.com, matthew.andrews@nokia-bell-labs.com

Abstract

Machine learning methods are increasingly adopted in communications problems, particularly those arising in next generation wireless settings. Though seen as a key climate mitigation and societal adaptation enabler, communications related energy consumption is high and is expected to grow in future networks in spite of anticipated efficiency gains in 6G due to exponential communications traffic growth. To make meaningful climate mitigation impact in the communications sector, a mindset shift away from maximizing throughput at all cost and towards prioritizing energy efficiency is needed. Moreover, this must be adopted in both existing (without incurring further embodied carbon costs through equipment replacement) and future network infrastructure, given the long development time of mobile generations. To that end, we present summaries of two such problems, from both current and next generation network specifications, where probabilistic inference methods were used to great effect: using Bayesian parameter tuning we are able to safely reduce the energy consumption of existing hardware on a live communications network by 11% whilst maintaining operator specified performance envelopes; through spatiotemporal Gaussian process surrogate modeling we reduce the overhead in a next generation hybrid beamforming system by over 60%, greatly improving the networks' ability to target highly mobile users such as autonomous vehicles. The Bayesian paradigm is itself helpful in terms of energy usage, since training a Bayesian optimization model can require much less computation than, say, training a deep neural network.

Introduction

The sixth assessment report of the Intergovernmental Panel on Climate Change (IPCC) makes for sobering reading: even in the best case scenario we are likely to exceed +1.5 °C global warming in the near future with far-reaching impacts spanning human and ecosystems (Pörtner et al. 2022). The information and communication technology (ICT) industry—in particular, the massive bandwidth and low latency promised by next-generation network architecture—is seen as a key enabler for both societal adaptation and climate mitigation, e.g. through expanded online services in healthcare, education, and work, and through decentralized renewables based microgrids (I and Bian 2020).

However, as the popular adage goes, "there ain't no such thing as free lunch". The ICT industry already accounts for 3% of global energy consumption (GSMA 2019) and an estimated 2-4% of the world's greenhouse gas emissions (GHG) (Freitag et al. 2021). Networking makes up a significant proportion of the ICT sector, with radio access networks (RAN) alone accounting for over half of its energy consumption (Lopez-Perez et al. 2021). All the major RAN vendors (Ericsson, Huawei, Nokia, and Samsung) have committed to ambitious net-zero targets—in many cases exceeding the voluntary standard by aiming for net-zero by as early as 2030 (Ericsson 2021)—and are investing heavily in developing next generation wireless as a climate mitigation enabler (Huawei 2020; Nokia 2022; Ericsson 2022). Their customer base (mobile network operators), for whom energy costs make up a bulk of their operating expenditure (GSMA 2019), are likewise prioritizing energy efficiency. Standards bodies are also bringing energy efficiency to the forefront: 3GPP[1] have set a 90% energy reduction goal for the next generation architecture New Radio (NR), and machine learning will play an increasingly important role in unlocking next generation wireless efficiencies (Elbir et al. 2023). Regional and national future communication development initiatives have likewise placed sustainability at the forefront (National Science Foundation 2021; Department for Science, Innovation and Technology and Department for Digital, Culture, Media & Sport 2022; Uusitalo et al. 2021).

Even though 5G and beyond will confer substantial energy efficiency in terms of Joules per bit transferred, in absolute terms consumption is predicted to increase (Freitag et al. 2021). This is predicated on an anticipated rebound effect—where increased efficiency and bandwidth give way to a surge in demand—intrinsic higher network element density compared to LTE (of e.g. mobile base stations, user devices), the need for network operators to support mixed frequency sites for years to come, and embodied carbon in the production of new RAN infrastructure. Moreover, given the long development time (typically 10 years between generations) and disparate rollouts (5G will not be the dominant service

[1]Consortium of major standards organizations which develop mobile telecommunications protocols

until after 2028, and even then, many regions—notably sub-Saharan Africa—will still rely heavily on previous generation hardware (Jonsson 2022)), it is imperative that both existing and future infrastructure is optimized for efficiency. This will require a mindset shift away from maximum performance to sustainable performance, where throughput and energy consumption are jointly optimized. Energy efficiency may be addressed at various layers in the RAN architecture, from physical layer processing, to network-scale radio resource control processing. Here we focus on higher level radio resource management (RRM) applications, for examples of ML applied to the physical layer see e.g. Korpi et al. (2021) and references therein. In this paper we will present summaries of two applications of probabilistic machine learning to increase energy efficiency in RAN. Crucially, we consider both current generation 3GPP LTE and next generation 3GPP New Radio (NR). Both settings utilize Bayesian inference which is itself helpful in terms of energy usage as it typically requires far less computation than, say, training a deep neural network.

Learning Considerations

RRM problems are challenging to optimize, with objectives that tend to be expensive to evaluate either through simulations or via point estimates of user quality of service (Zhang et al. 2023). Many approaches leverage deep neural network based techniques, with the goal of developing generalized models for wide deployment. These approaches can have good performance at inference time with relatively low energy requirements. However, they are typically expensive (from a computational and therefore energy point of view) to train. Another practical consideration is model life-cycle management: periodic re-training/tuning and model transfer over-the-air from base station to user equipment are challenges that should not be underestimated.

Both applications considered in this paper share characteristics–namely, low dimensionality and information scarcity–making them amenable to online Bayesian optimization. Bayesian optimization is a class of stochastic optimization methods suitable for derivative-free optimization of expensive to evaluate black-box objective functions, see Frazier (2018) for an excellent introduction. There are two core aspects: a surrogate model, developed using Bayesian statistics, of the true objective, and an acquisition policy, which uses the surrogate model to infer where to sample the objective next. Because we forego generalization across wide-ranging network deployments, and our problem dimensions are small, the costly Gaussian process parameter fitting[2] is abated, making this a low-cost alternative to wide deployment generalized neural network based techniques.

Results and Discussion

We now describe the cellular energy efficiency problems in more detail. For current networks we are typically faced with a situation where traffic at a cell site fluctuates during the day. In this case it is wasteful to keep all carriers (i.e. frequencies) and their associated power amplifiers active at all time. We therefore want to learn a strategy that reacts to the changing traffic and switches off carriers to save energy whenever this can be done without violating customer performance metrics.

For future networks, large antenna arrays improve efficiency by focusing energy at the intended receiver. For this mechanism to work, the cell must track the best beam for each mobile user, and it is prohibitively expensive to measure every beam at every time instant. However, if the wrong beam is selected then energy is wasted because the focused beam "misses" its target. We therefore apply Bayesian techniques to learn the best beam with a minimal number of measurements.

Current Generation: Cell Switch-Off

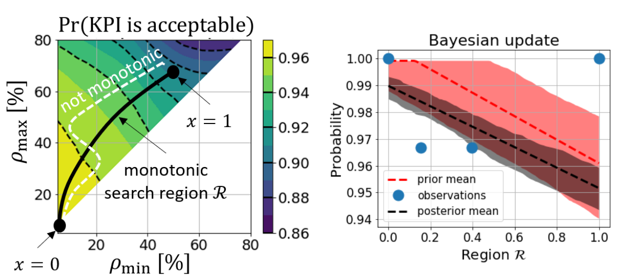

Figure 1: Left: Adjusting load thresholds (ρ) to effect carrier switch-off is transformed to a Bayesian root finding problem. The search space (density plot of example quality of service metric) is reduced to a monotonic region $\mathcal{R}$. Right: The parameterized probability function of our root finding problem is refined using Bayesian posterior updating. Figure adapted from Maggi et al. (2023b) © 2023 IEEE.

Power amplifiers (PA) account for over 65% of a base stations energy consumption (Lopez-Perez et al. 2021). Switching off PAs without (overly) degrading network performance is therefore a good route to curbing energy consumption.

In Maggi et al. (2023b) we develop an approach whereby carriers are shut down according to a hysteresis mechanism: the next carrier in line is switched off (on) if the traffic load on the sector is lower (greater) than a certain threshold $\rho_{\min}$ ($\rho_{\max}$). The key is dynamic selection of the thresholds such that key performance indicators (KPIs) meet quality of service (QoS) constraints, with a desired likelihood. We formulate this as a root finding problem, in which the solution is mapped to a one-dimensional space that is iteratively searched using a Bayesian approach, see Maggi et al. (2023b) for derivation and implementation details. Briefly, we parameterize a probability function of our threshold, describing the probability of satisfying service constraints, and maintain a probabilistic belief over the parameter space. We convolve this belief with a Markovian transition law, allowing traffic changes to be learned, and update this belief with new observations via Bayesian posterior updating.

We tested this policy in a proof of concept (PoC) trial on a live customer 4G network, spanning 19 sites and 57 sec-

[2] For us this involves a matrix inversion with memory requirements that scales cubically in the number of data points.

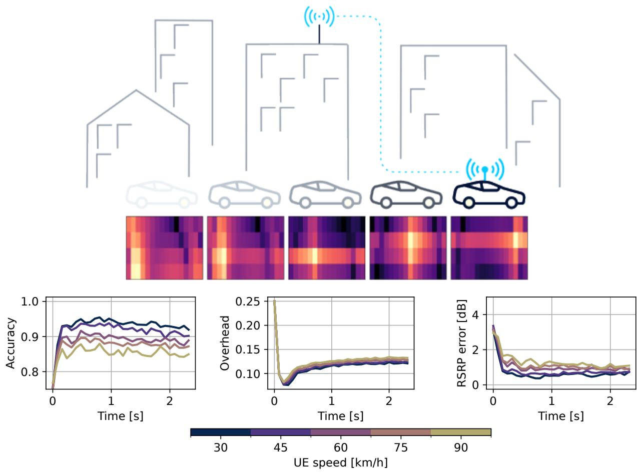

Figure 2: Highly mobile UEs, such as connected vehicles, will experience different radio environments over time, requiring tracking of the best transmission and receiver beam pairs. The schematic (top) shows a mobile UE traversing a cell sector in an urban environment, maintaining a communication link with a base station. As the UE moves from left to right, the radio environment changes (due to, for example, radio reflections), which affects the signal strength of the base station beams, as experienced by the UE. This is indicated by the contour plots below the UE (center), with brighter color indicating a stronger received signal reference power (RSRP), and therefore better beam choice for the communication link. The time series (bottom) show the efficacy of our Bayesian optimization solution in tracking the best beam for mobile UEs, with varying mobility level indicated by the color (brighter is faster): the online learning approach suffers a cold start of less than 300 ms, with minimal degradation of RSRP error and overhead metrics for increasingly mobile UEs. Figure adapted from Maggi et al. (2023a) © 2023 IEEE.

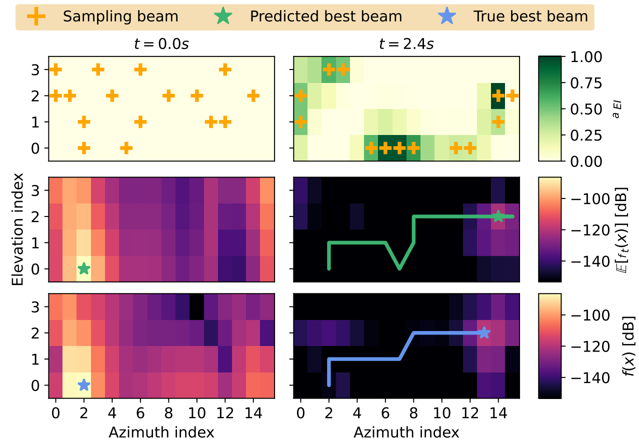

Figure 3: We visualize the salient components of our beam tracking solution at an initial and late stage time slot. Top row: with an uninformative prior, we distribute initial beam samples (orange crosses) according to a space filling sequence (left). At subsequent time slots (right) samples are distributed according to an acquisition function (darker is better). Bottom row: The ground truth signal strength (brighter is better) experienced by the UE for each transmission beam index changes as the time slots evolve. The optimal beam index is denoted with a blue star, and the solid line traces the evolution of the optimal beam index. Center row: we plot the expectation over our posterior estimate of the signal strength across beam indices. This estimate is used to predict the optimal beam index for a given time slot (green start), and once again the solid line traces the evolution of this predicted optimal beam index. Figure adapted from Maggi et al. (2023a) © 2023 IEEE.

tors. Most of the sites had 4 frequency layers (800, 1800, 2100 and 2600 MHz), with the 800 and 1800 MHz layers left active to preserve coverage. Baseline measurements were taken over periods spanning a few weeks immediately before and after the PoC trial, during which all carriers were kept active. We used two weeks of historical data to initialize the prior, and we used a Gaussian distribution with zero mean and diagonal covariance matrix as our Markovian transition rule, allowing the policy to adapt to traffic variations by gradually "forgetting" past observations. During our PoC, we were able to shut down carriers for approximately 30% of the time with negligible impacts on cell congestion and traffic volume, including cells in neighboring sites. Overall, we are able to effect an 11% reduction in energy consumption during the PoC, which is remarkable considering no hardware changes were made.

Next Generation: Beam Tracking

Next generation millimeter wave (mmWave) communications will use highly directional transmitter and receiver (Tx and Rx, respectively) beams, complicating the mutual discovery process–both in initial access (IA) and the recovery from link failure– and the ongoing beam tracking for highly mobile UEs such as connected vehicles or mobile users on mass transit. Finding a good beam pair is important not only from a QoS point of view but also energy efficiency: a suboptimal beam pair means that the Tx and Rx beams are misaligned, so energy is wasted by effectively pointing beams at empty space. Even a loss as small as 1 dB equates to approximately 20 percent power waste in transmission, so as our wireless communications become more directional and highly mobile UEs more ubiquitous efficient beam alignment will go a long way towards curbing emissions growth.

The signal received by a UE is a function of its gain, exogenous noise, the selected Tx and Rx beams, and the channel state information (CSI[3]) matrix. A good signal has high magnitude, known as the received signal reference power (RSRP). If the CSI is known, then the RSRP can be computed for all Tx-Rx beam pairs, making beam selection trivial. However, if no CSI is available, then beam pair selection must be made based on RSRP measurements. Current stan-

[3]Complex valued matrix whose rank depends on BS and UE antenna array sizes

dards employ an exhaustive search whereby all beam pairs are sampled and the best one selected. Because beam width is inversely proportional to array size, and analog beam formers may only probe one direction at a time, this exhaustive search is significantly more expensive for next generation architectures that use far larger antenna arrays.

In Maggi et al. (2023a) we developed an alignment and tracking procedure using Bayesian optimization to greatly reduce the overhead required during both IA and subsequent beam tracking, especially when dealing with highly mobile UEs. We maintain a Gaussian process surrogate model of the RSRP landscape over all beam pairs, and use this model with a modified expected improvement based acquisition function to rapidly identify the best beam pair in a given time slot. Through appropriate Gaussian Process (GP) kernel design we can cater for highly mobile UEs, modeling the RSRP evolution in both space and time. See Maggi et al. (2023a) for modeling and implementation details, including a strategy for restricting the size of the set of sampled beams with proven optimality gap.

We evaluated our approaching using a 3GPP NR-compliant system level simulator, and benchmarked its performance against exhaustive search. A full system description and complete benchmark comparisons against spatial interpolation and long short term memory based approaches may be found in (Maggi et al. 2023a). Overall, we were able to use 60% fewer measurements than the current standard, resulting in more efficient use of transmission power, without compromising alignment accuracy: RSRP errors were kept around 1dB (translating to approximately milliwatt power wastage due to beam misalignment), even for UEs travelling at 90 km/h (highway speed). Importantly, the performance was found to be stable over time, with no overhead surges or accuracy degradation evident.

Conclusion

Machine learning will play an important role in unlocking efficiencies in current and next generation wireless networks—a key climate mitigation and societal adaptation enabler. But to mitigate against growing ICT emissions an industry mindset shift is needed to jointly optimize performance and efficiency. Crucially, given the long development time of next generation wireless networks, we must include current generation architectures in this joint optimization effort to have an impact on ICT climate mitigation today.

In this paper we introduced two problems, one from the current network generation, the other from the next generation, where we applied probabilistic machine learning to strike a balance between performance and energy consumption. We sought to introduce radio resource management to the wider machine learning community, whose contribution to the adoption of machine learning to next generation wireless will be crucial.

Acknowledgments

This work is a contribution by Project REASON, a UK Government funded project under the Future Open Networks Research Challenge (FONRC) sponsored by the Department of Science Innovation and Technology (DSIT). We would like to thank the anonymous reviewers for their constructive feedback, and our colleague Mohammad Malekzadeh for commenting on an earlier version of this paper.

References

Department for Science, Innovation and Technology; and Department for Digital, Culture, Media & Sport. 2022. Future Open Network Research Challenge. https://www.gov.uk/guidance/future-open-networks-research-challenge-competition. Accessed: 2023-11-06.

Elbir, A. M.; Mishra, K. V.; Vorobyov, S. A.; and Heath, R. W. 2023. Twenty-Five Years of Advances in Beamforming: From convex and nonconvex optimization to learning techniques. *IEEE Signal Processing Magazine*, 40(4): 118–131.

Ericsson. 2021. Sustainability and Corporate Responsibility report. https://www.ericsson.com/en/about-us/sustainability-and-corporate-responsibility/sustainability-report. Accessed: 2023-11-06.

Ericsson. 2022. On the road to breaking the energy curve. https://www.ericsson.com/en/about-us/sustainability-and-corporate-responsibility/environment/product-energy-performance. Accessed: 2023-11-06.

Frazier, P. I. 2018. A Tutorial on Bayesian Optimization.

Freitag, C.; Berners-Lee, M.; Widdicks, K.; Knowles, B.; Blair, G.; and Friday, A. 2021. The climate impact of ICT: A review of estimates, trends and regulations. arXiv:2102.02622.

GSMA. 2019. Energy Efficiency: An Overview. https://www.gsma.com/futurenetworks/wiki/energy-efficiency-2/. Accessed: 2023-11-06.

Huawei. 2020. Green Development 2030. https://www.huawei.com/en/giv/green-development-2030. Accessed: 2023-11-06.

I, C. L.; and Bian, S. 2020. Energy-efficient 5G for a greener future. *Nature Electronics*, 3: 182–184.

Jonsson, P. 2022. Ericsson Mobility Report. Technical report, Ericsson, Stockholm, Sweden.

Korpi, D.; Honkala, M.; Huttunen, J. M.; and Starck, V. 2021. DeepRx MIMO: Convolutional MIMO Detection with Learned Multiplicative Transformations. In *ICC 2021 - IEEE International Conference on Communications*, 1–7.

Lopez-Perez, D.; Domenico, A. D.; Piovesan, N.; Bao, H.; Xinli, G.; Qitao, S.; and Debbah, M. 2021. A Survey on 5G Radio Access Network Energy Efficiency: Massive MIMO, Lean Carrier Design, Sleep Modes, and Machine Learning. arXiv:2101.11246.

Maggi, L.; Koblitz, A. R.; Zhu, Q.; and Andrews, M. 2023a. Tracking the Best Beam for a Mobile User via Bayesian Optimization. In *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, 1–7.

Maggi, L.; Mihailescu, C.; Cao, Q.; Tetich, A.; Khan, S.; Aaltonen, S.; Koblitz, R.; Holma, M.; Macchi, S.; Ruggieri, M. E.; Korenev, I.; and Klausen, B. 2023b. Energy Savings under Performance Constraints via Carrier Shutdown with

Bayesian Learning. In *2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 1–6.

National Science Foundation. 2021. ”Resilient & Intelligent NextG Systems (RINGS)”. https://www.nsf.gov/pubs/2021/nsf21581/nsf21581.pdf. Accessed: 2023-11-06.

Nokia. 2022. People & Planet 2021. https://www.nokia.com/about-us/sustainability/sustainability-downloads/. Accessed: 2023-11-06.

Pörtner, H. O.; Roberts, D. C.; Tignor, M.; Poloczanska, E. S.; Mintenbeck, K.; Alegría, A.; Craig, M.; Langsdorf, S.; Löschke, S.; Möller, V.; Okem, A.; and Rama, B., eds. 2022. *Summary for Policymakers*. Cambridge, UK: Cambridge University Press. In Press.

Uusitalo, M. A.; Ericson, M.; Richerzhagen, B.; Soykan, E. U.; Rugeland, P.; Fettweis, G.; Sabella, D.; Wikström, G.; Boldi, M.; Hamon, M.-H.; Schotten, H. D.; Ziegler, V.; Strinati, E. C.; Latva-aho, M.; Serrano, P.; Zou, Y.; Carrozzo, G.; Martrat, J.; Stea, G.; Demestichas, P.; Pärssinen, A.; and Svensson, T. 2021. Hexa-X The European 6G flagship project. In *2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 580–585.

Zhang, Y.; Simeone, O.; Jose, S. T.; Maggi, L.; and Valcarce, A. 2023. Bayesian and Multi-Armed Contextual Meta-Optimization for Efficient Wireless Radio Resource Management. *IEEE Transactions on Cognitive Communications and Networking*, 9(5): 1282–1295.