HUN REN | SZTAKI

Conceptual and Mathematical Foundations of

AI: Artificial Intelligence

Balázs Csanád Csáji

Institute for Computer Science and Control (SZTAKI)

Institute of Mathematics, Eötvös Loránd University (ELTE)

PBK AI Seminar, Budapest, January 26, 2026

Part I

What is Artificial Intelligence?

What is Artificial Intelligence?

- A High-Level Expert Group (HLEG) on Artificial Intelligence has been appointed by the European Commission (2018).
- AI Watch is the internal evidence-gathering and analysis initiative of EC on AI. Aim: implement the European strategy on AI (2018).

> *"AI has become an area of strategic importance and been identified as a potential key driver of economic development as underlined in the European strategy on AI."*
>
> *"Despite the increased interest in AI by the academia, industry and public institutions, there is no standard definition of what AI actually involves."*

(AI Watch, JRC Technical Report, European Commission, 2020)

(Note: ChatGPT was released on November 30, 2022 by OpenAI)

Can Machines Think?

- René Descartes (1637): a machine may perform some tasks better than a human, but only because of the specific "disposition of its organs". However, the human reason is a "universal instrument".
- Gottfried Wilhelm Leibniz (1714): perception and consciousness could never be explained by physical parts alone, "mill" argument.
- Julien Offray de La Mettrie (1747): the human body is a "machine which winds its own springs" (monist, materialist, determinist).
- Alan Turing (1950): if a machine acts indistinguishably (cf. test) from a thinking being, then for all practical purposes, it is thinking.
- John Searle (1980): syntax (following rules) is not sufficient for semantics (knowing the meaning), "Chinese room" argument.
- David Chalmers (1995): even a perfect physical explanation of the human brain seems incomplete, "philosophical zombie" argument.

What is Artificial Intelligence?

Thinking Humanly	Thinking Rationally
"*The exciting new effort to make computers think... machines with minds, in the full and literal sense.*" (Haugeland, 1985) "*[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning...*" (Bellman, 1978)	"*The study of mental faculties through the use of computational models.*" (Charniak and McDermott, 1985) "*The study of the computations that make it possible to perceive, reason, and act.*" (Winston, 1992)
Acting Humanly	**Acting Rationally**
"*The art of creating machines that perform functions that require intelligence when performed by people.*" (Kurzweil, 1990) "*The study of how to make computers do things at which, at the moment, people are better.*" (Rich and Knight, 1991)	"*Computational Intelligence is the study of the design of intelligent agents.*" (Poole et al., 1998) "*AI... is concerned with intelligent behavior in artifacts.*" (Nilsson, 1998)

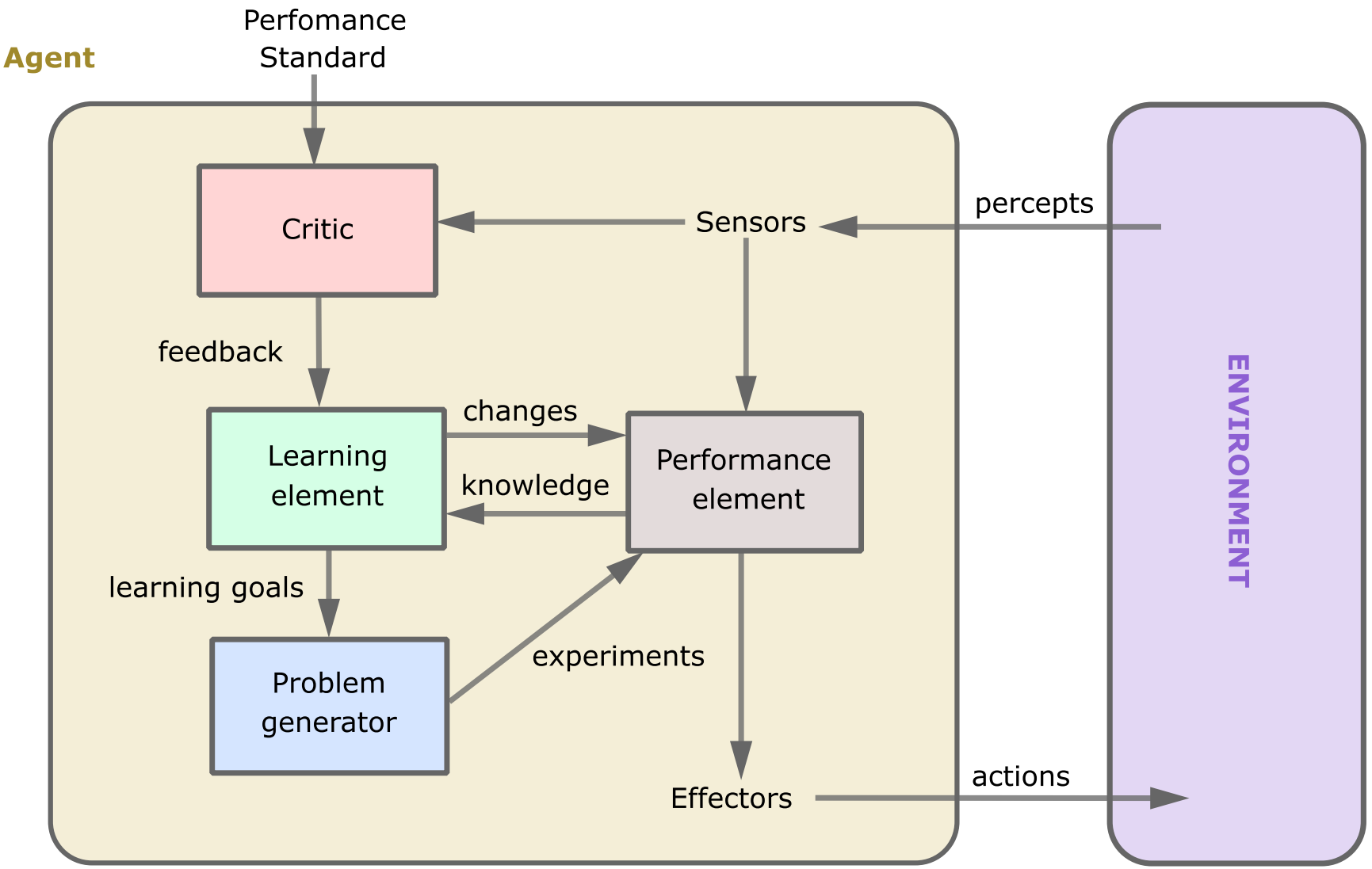
(S. Russell & P. Norvig: Artificial Intelligence: A Modern Approach, 3rd ed., Pearson, 2014)

What is Artificial Intelligence?

The HLEG of the European Commission (EC) defined AI as (2019):

> *"Artificial intelligence systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions."*

General Structure of a Learning Agent



Categorizing AI Risks and Challenges

Speculative risks (currently unfounded):

- Machine self-awareness (Skynet scenario), AI rebellion
- AI singularity: autonomous, accelerating self-improvement

Slightly well-founded fears:

- Significant labor market shifts and short-term unemployment

Real risks:

- Faulty systems due to human error and extreme competition
- Misuse: social control, scams, weapons and adaptive malware
- Ethical issues: privacy, freedom, security and algorithmic bias

Unresolved challenges:

- Policy lag, e.g., legal liability and science-policy readiness
- Social impact (e.g., deepfakes) and modernization of education
- Transparency (explainability) and robustness (trustworthiness)

The EU Artificial Intelligence Act

- The AI Act is a European regulation (2024/1689) that governs how AI systems are developed, deployed, and used across the EU.
- The four types of AI Risk regulated by the AI Act:

1. Unacceptable risk: systems which threaten the fundamental rights or safety. For example, social scoring by governments, or emotion recognition in workplaces / schools. These are prohibited.
2. High risk: systems that can significantly impact a person's life or safety, e.g., credit scoring, medical devices or law enforcement. They must include human oversight and logging for traceability.
3. Limited risk: AI chatbots and text / image / video generators. Users must be clearly informed they are interacting with an AI.
4. Minimal / no risk: most AI applications, e.g., video games, spam filters, and industrial optimization. There are no new obligations.

A Taxonomy for Artificial Intelligence

The taxonomy proposed by the AI Watch Report (2020):

- Knowledge representation
- Automated reasoning
- Common sense reasoning
- Planning and scheduling
- Natural language processing
- Optimization
- Searching
- Machine learning
- Computer vision
- Audio processing
- Multi-agent systems
- Robotics and automation
- Connected and automated vehicles
- AI services
- AI ethics
- Philosophy of AI

(the red/blue subdomains are classified as the "core" parts of AI)

ELLIS Members Working at Hungarian Institutions

European Laboratory for Learning and Intelligent Systems

András Benczúr
Member
Institute for Computer Science and Control (SZTAKI)

Csaba Kerepesi
Member
HUN-REN SZTAKI

Karolina Pircs
Member
Semmelweis University (HU)

Richárd Farkas
Member
University of Szeged

András Lőrincz
Member
Eötvös Loránd University (ELTE)

Daniel Barath
Member
ETH Zurich
HUN-REN SZTAKI

Levente Hajder
Member
Eötvös Loránd University (ELTE)

Tamás Szirányi
Member
Institute for Computer Science and Control (SZTAKI)

Balázs Csanád Csáji
Member
HUN-REN SZTAKI

Gergő Orbán
Member
HUN-REN Wigner Research Centre for Physics

Márk Jelasity
Member
University of Szeged

Zoltan Kato
Member
University of Szeged

Csaba Benedek
Member
Institute for Computer Science and Control (SZTAKI)

István Csabai
Member
Eötvös Loránd University (ELTE)

Peter Horvath
Member
Biological Research Centre of the Hungarian Academy of Sciences (BRC-HAS)

Zsolt Zombori
Member
Alfred Renyi Institute of Mathematics

What is Machine Learning?

– According to Tom Mitchell (Carnegie Mellon University, 1997):

"The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience. A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

– According to AI Watch report prepared for the EC (2020):

"By learning, we refer to the ability of systems to automatically learn, decide, predict, adapt and react to changes, improving from experience, without being explicitly programmed. ML is widely included in the vast majority of efforts to identify AI categories, as the basic algorithmic approach to achieve AI."

Machine Learning vs Mathematical Statistics

- According to Michael Jordan (University of California, Berkeley, 2014):

  "*I personally don't make the distinction between statistics and machine learning [...] the "ML community" realized that their ideas had had a lengthy pre-history in statistics. Decision trees, nearest neighbor, logistic regression, kernels, PCA, canonical correlation, graphical models, K means and discriminant analysis come to mind, and also many general methodological principles (e.g., method of moments, which is having a mini-renaissance, Bayesian inference methods of all kinds, M estimation, bootstrap, cross-validation, EM, ROC, and of course stochastic gradient descent, whose pre-history goes back to the 50s and beyond), and many many theoretical tools (large deviations, concentrations, empirical processes, Bernstein-von Mises, U statistics, etc).*"

- Robert Tibshirani (Stanford University): ML is "*glorified statistics*"

- Objections, e.g., deterministic ML problems, and computational aspects

Branches of Machine Learning

I. Supervised Learning

Learning from a sample of (noisy) input-output data.
Problems, e.g., classification, regression and experiment design.

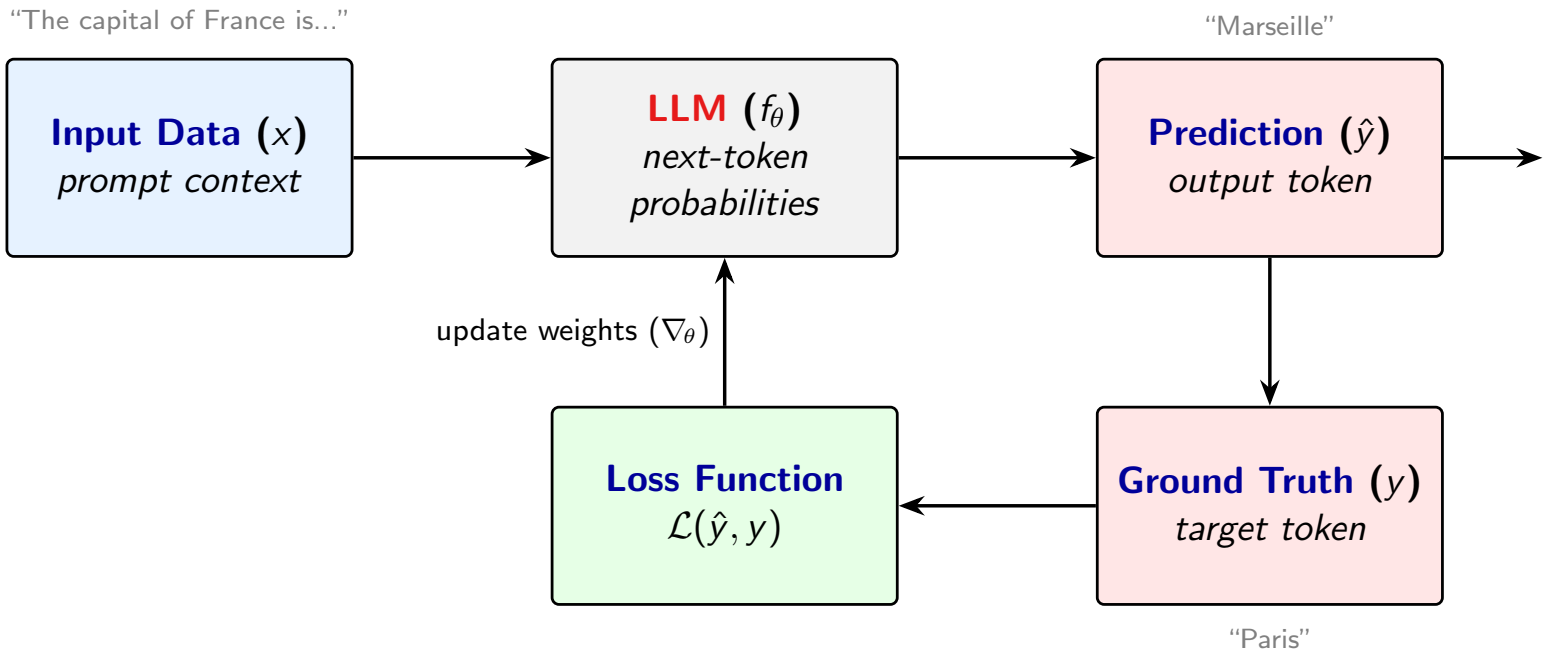
II. Unsupervised Learning

Learning from a sample of unlabelled data (no outputs).
Problems, e.g., clustering, dim. reduction and anomaly detection.

III. Reinforcement Learning

Learning via interactions with an uncertain dynamic environment.
Problems, e.g., stochastic shortest paths and multi-armed bandits.

Regression Perspective of Large Language Models

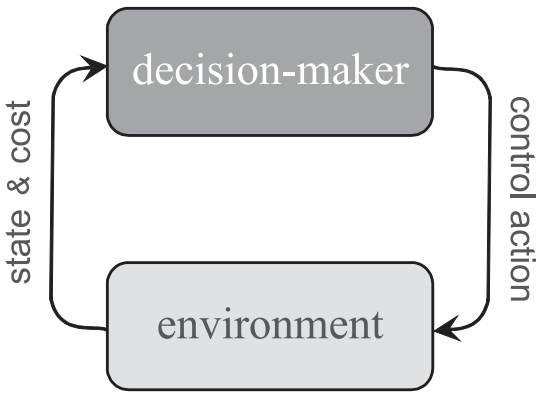


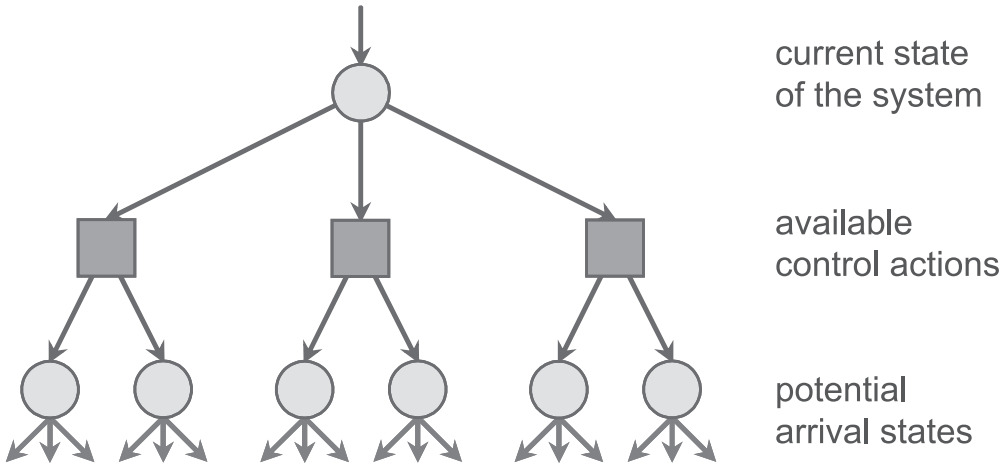
The objective is to minimize the expected loss $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f_\theta(x), y)]$ where the dataset $\mathcal{D}$ consists of trillions of tokens (e.g., from the web).

Tokens are the "atomic units" of LLMs. A token is typically a word ("apple"), a sub-word ("un", "able"), or a space / mark ("!", ",").

Reinforcement Learning

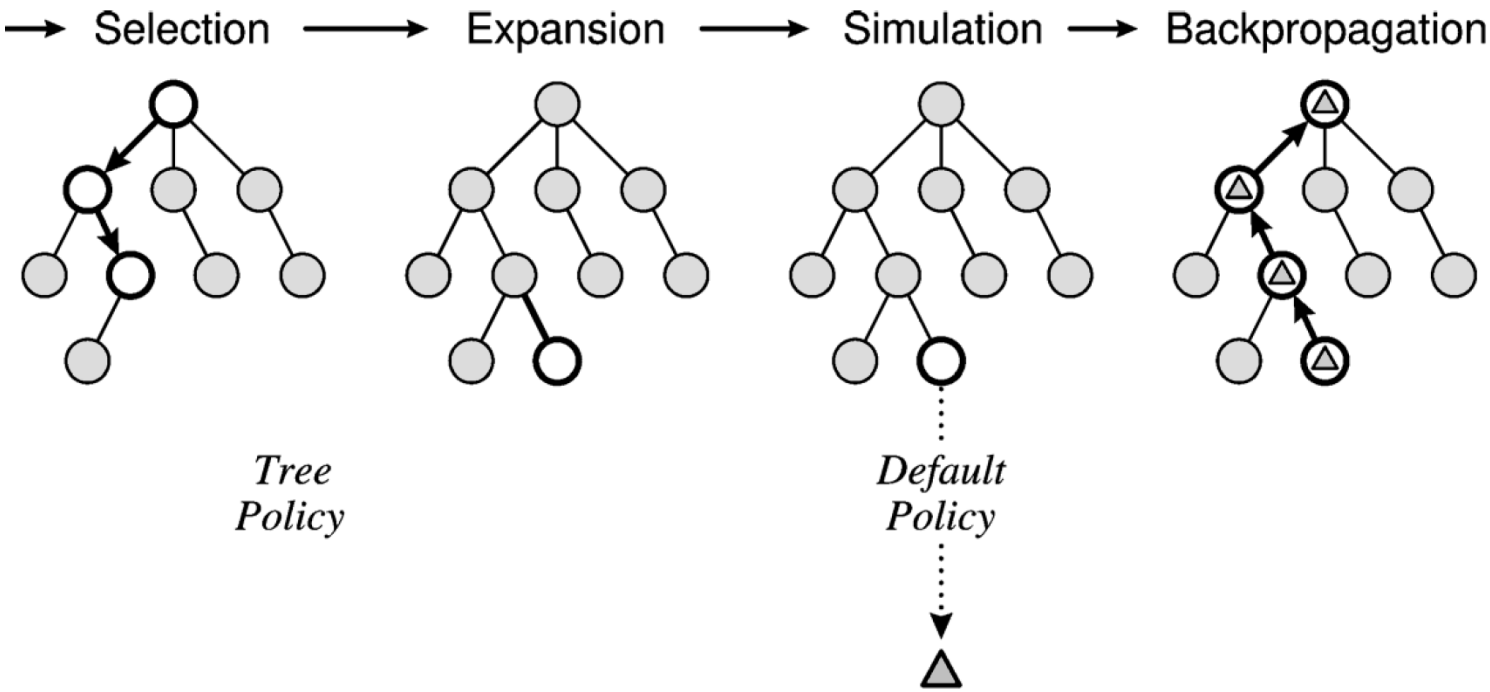
- Reinforcement learning (RL) is one of the main branches of machine learning to learn from interactions with a system based on feedbacks.
- An interpretation: consider an agent acting in an potentially unknown uncertain environment and receiving information about states and costs.
- The aim is to learn an efficient behavior (control policy), such that applying this strategy minimizes the cumulative costs in the long run.





Example Application: Monte Carlo Tree Search

– A well-known application of RL is the UCT (Upper Confidence Bounds for Trees) algorithm that is a powerful Monte Carlo Tree Search method.



Source: C. Browne *et al.* "A Survey of Monte Carlo Tree Search Methods", IEEE Transactions on Computational Intelligence and AI in Games, 2012.

Applications of Reinforcement Learning

- Robot Control
- Dispatching & Scheduling
- Optimal Stopping
- Routing
- Maintenance and Repair
- Recommender Systems
- Inventory Control
- Optimal Control of Queues
- Strategic Asset Pricing
- Dynamic Options
- Insurance Risk Management
- Web System Configuration
- Bidding and Advertising
- Traffic Light Control
- Logic Games
- Communication Networks
- Dynamic Channel Allocation
- Power Grid Management
- Supply-Chain Management
- Fault Detection
- Sequential Clinical Trials
- PageRank Optimization

Part II

A Fundamental Problem Class for Artificial Intelligence

Mathematical Foundations of Machine Learning

- ML researches creatively built on existing mathematical theories, but there is wide range of open questions and unsolved hard problems.
- Several ML problems need novel mathematical approches, foundations.
- Other parts of mathematics benefit from ML, e.g., Vapnik-Chervonenkis dimension (used in param. complexity, comp. geometry, probability, etc).
- Analogy: mathematical concepts inspired by physics, such as, derivatives, vector products, tensors, operator algebras, and the Fourer transform.
- Other benefits: social usefulness (wide range of vital ML applications), improved employement (e.g., for students) and funding opportunities.
- ML approaches with good mathemacial foundations: statistical learning, kernel methods, reinforcement learning, Bayesian methods (e.g., GPR).
- ML approaches with weaker foundations: artificial neural networks and deep learning, genetic algorithms, and deep reinforcement learning, etc.

A Fundamental Problem Class for ML

The framework is known under many names in various disciplines:

- Online Learning (machine learning)
- Stochastic Approximation (probability theory and statistics)
- Stochastic Optimization (operations research)
- Adaptive Algorithms (control engineering)
- Stochastic Iterative Algorithms (computer science)
- Stochastic Recursive Algorithms (computer science)

It is very important for many fields, such as reinforcement learning, deep learning, adaptive filtering, recursive estimation of time-series models, etc.

Typical stoch. approximation (SA) problems include finding a root, a fixed point or an extremum of an unknown function based only on noisy queries.

Stochastic Approximation

Stochastic Approximation (SA)

$$\underbrace{\theta_{n+1}}_{\text{next estimate}} \doteq \underbrace{\theta_n}_{\text{current estimate}} + \underbrace{\gamma_n}_{\text{learning rate}} \underbrace{H(\theta_n, X_{n+1})}_{\text{update operator}}$$

- $\theta_n \in \Theta$ is the estimate at time n.
- $\gamma_n \in [0, \infty)$ is the learning rate at time n.
- $X_n \in \mathcal{X}$ is the new data available at time n.
- $H : \Theta \times \mathcal{X} \to \Theta$ is the update operator.

Note: Θ and $\mathcal{X}$ are typically Euclidean or (separable) Hilbert spaces.

Reminder: Strong Law of Large Numbers

Recall the law of large numbers: that the empirical mean (of i.i.d. variables) converges (with probability one) to the expected value.

Strong Law of Large Numbers (SLLN)

Let $\{X_t\}$ be i.i.d. (real) random variables with $\mathbb{E}\big[X_t\big] = \theta^*$ and

$$\theta_n \doteq \frac{1}{n}\sum_{t=1}^{n} X_t.$$

Then, we have that $\theta_n \xrightarrow{a.s.} \theta^*$, as $n \to \infty$. In other words,

$$\mathbb{P}\Big(\lim_{n\to\infty} \theta_n = \theta^*\Big) = 1.$$

(Note that the sample mean is typically denoted by $\overline{X}_n$, here θ_n is used instead to make an easier connection with SA algorithms.)

Recursive Averaging

The averages $\{\theta_n\}$ can be computed recursively, $\theta_0 \doteq 0$ and

Recursive Averaging

$$\underbrace{\theta_{n+1}}_{\substack{\text{next}\\ \text{estimate}}} = \underbrace{\theta_n}_{\substack{\text{current}\\ \text{estimate}}} + \underbrace{\frac{1}{n+1}}_{\text{step-size}} \underbrace{\Big(X_{n+1} - \theta_n\Big)}_{\text{correction term}}$$

$$\begin{aligned}\theta_{n+1} &= \frac{1}{n+1}\sum_{i=1}^{n+1} X_i = \frac{n}{n+1}\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n+1}X_{n+1}\\ &= \frac{n}{n+1}\theta_n + \frac{1}{n+1}X_{n+1} = \theta_n + \frac{1}{n+1}\Big(X_{n+1} - \theta_n\Big).\end{aligned}$$

Root Finding Perspective

– Recursive averaging can be reformulated as root finding.

Recursive Averaging as Root Finding

$$\theta_{n+1} = \theta_n + \gamma_n H\big(\theta_n, X_{n+1}\big),$$

where $\gamma_n = 1/(n+1)$ is called the step-size or learning rate and

$$H(\theta_n, X_{n+1}) \doteq X_{n+1} - \theta_n = \theta^* - \theta_n + \varepsilon_n$$
$$= h(\theta_n) + \varepsilon_n,$$

is the update operator, where $\varepsilon_n \doteq X_{n+1} - \theta^*$, so $\mathbb{E}\big[\varepsilon_n\big] = 0$.

– Therefore, we have noisy observations of a decreasing function $h(\theta) \doteq \theta^* - \theta$, and we iteratively search for its root, $h(\theta^*) = 0$.

General Learning Rates

Recursive Weighted Averaging

Let $\{X_n\}$ be a sequence of i.i.d. $\mathbb{R}$-valued random variables with bounded variance and $\mathbb{E}[X_n] = \theta^*$. Consider the recursion

$$\theta_{n+1} \doteq \theta_n + \gamma_n (X_{n+1} - \theta_n),$$

where $\theta_0 \in \mathbb{R}$ is fixed and $\{\gamma_n\}$ are nonnegative and satisfy (a.s.)

$$\sum_{n=0}^{\infty} \gamma_n = \infty, \qquad \text{and} \qquad \sum_{n=0}^{\infty} \gamma_n^2 < \infty.$$

Then, we have that $\theta_n \xrightarrow{a.s.} \theta^*$, as $n \to \infty$.

- Terminology: we typically say that $\{\theta_n\}$ are strongly consistent.
- SLLN is a special case, assuming variables with bounded variance.

Robbins-Monro Algorithm (1)

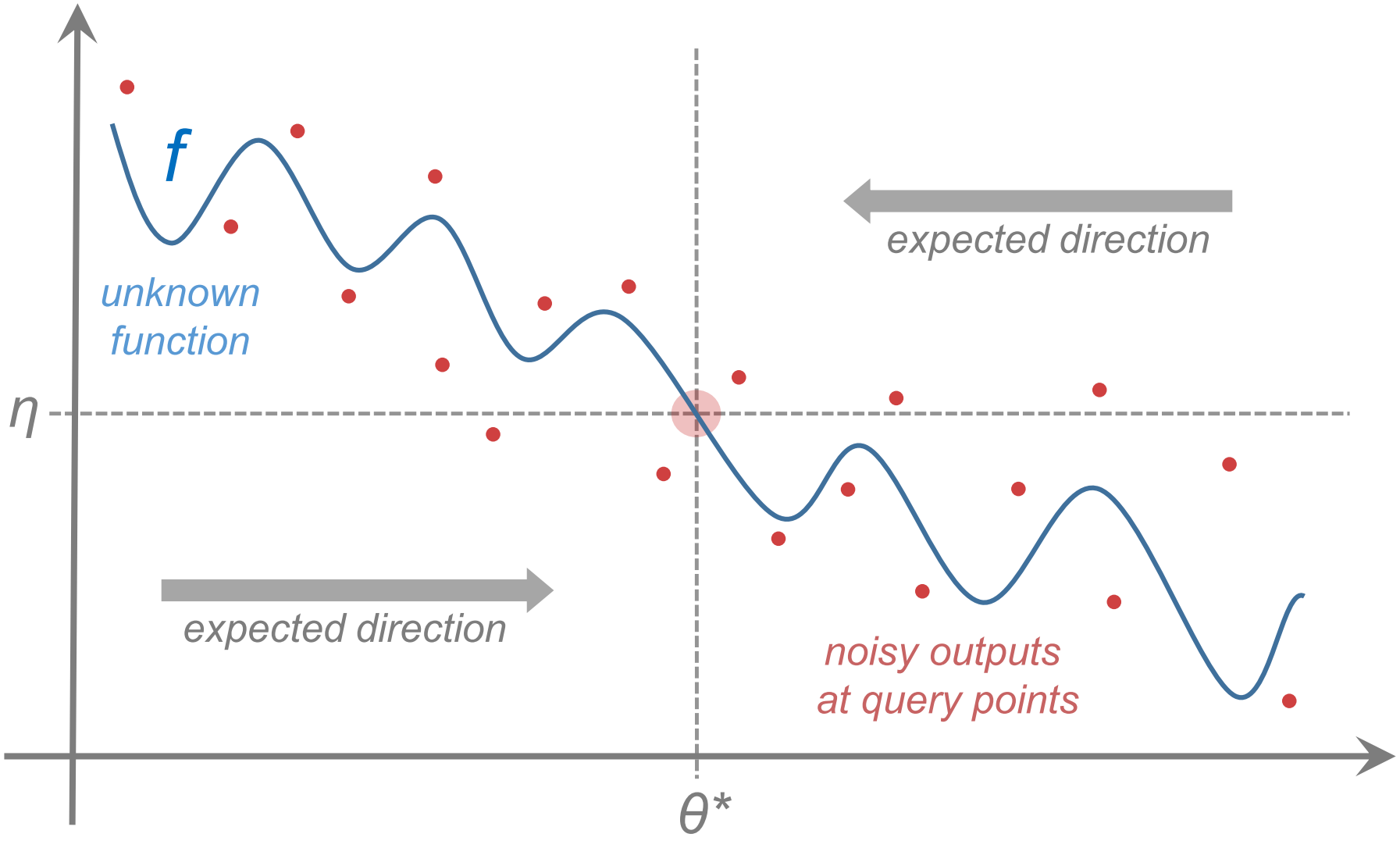
- We can make noisy queries of $f : \mathbb{R} \to \mathbb{R}$ and searching for a θ^* such that $f(\theta^*) = \eta$ for a known η (e.g., $\eta = 0$ is root finding).
- If we query f at point θ at time n, we observe $F_n(\theta) \doteq f(\theta) + \varepsilon_n$, where $\{\varepsilon_n\}$ are i.i.d. variables with $\mathbb{E}[\varepsilon_n] = 0$ and $\mathbb{E}[\varepsilon_n^2] < \infty$.

Robbins-Monro (RM) Algorithm (1951)

$$\theta_{n+1} \doteq \theta_n + \gamma_n\big(F_n(\theta_n) - \eta\big).$$

- Assume that $f(\theta) < \eta$ for $\theta > \theta^*$, $f(\theta) > \eta$ for $\theta < \theta^*$; that $\frac{\partial f}{\partial \theta}$ is strictly negative and is bounded in a neighborhood of θ^*; and that $|f(\theta)| < A|\theta| + B < \infty$, for all θ and for suitable A and B.
- Assume the learning rates $\{\gamma_n\}$ satisfy the previous assumptions.
- Then, for any $\theta_0 \in \mathbb{R}$, $\{\theta_n\}$ converges (a.s.) to θ^*, as $n \to \infty$.

Robbins-Monro Algorithm (2)



Robbins-Monro Algorithm (3)

– For a given $n \in \mathbb{N}$ and $\Delta > 0$ let us chose $m = m(n, \Delta)$ such that

$$\gamma_n + \gamma_{n+1} + \cdots + \gamma_m \approx \Delta.$$

– Then, the change in θ from time n to $n+m$ is approximately

$$\theta_{n+m} - \theta_n \approx \Delta(f(\theta_n) - \eta) + \sum_{k=n}^{m} \gamma_k \varepsilon_k,$$

where the variance of the (zero mean) "error" term is

$$\mathbb{E}\left[\sum_{k=n}^{m} \gamma_k \varepsilon_k\right]^2 = \mathbb{E}\left[\sum_{k=n}^{m} \gamma_k^2 \varepsilon_k^2\right] = \sum_{k=n}^{m} O(\gamma_k^2) = O(\Delta)\gamma_n$$

– The asymptotic behavior can be approximated by the mean ODE

$$\dot{\theta} = f(\theta) - \eta,$$

as $\theta_n + \Delta(f(\theta_n) - \eta)$ can be interpreted as an Euler method.

Kiefer-Wolfowitz Algorithm (1)

- We can query $f \in \mathcal{C}^1(\mathbb{R}, \mathbb{R})$ with noise: $F_n(\theta) \doteq f(\theta) + \varepsilon_n$, where $\{\varepsilon_n\}$ are i.i.d. variables with $\mathbb{E}\left[\varepsilon_n\right] = 0$ and $\mathbb{E}\left[\varepsilon_n^2\right] < \infty$ ($\mathcal{C}^1(\mathbb{X}, \mathbb{Y})$ is the class of cont. differentiable $\mathbb{X} \to \mathbb{Y}$ functions).
- We are searching for a (local) minimum point θ^* of function f.
- Assume the learning rates $\{\gamma_n\}$ satisfy the usual assumptions.

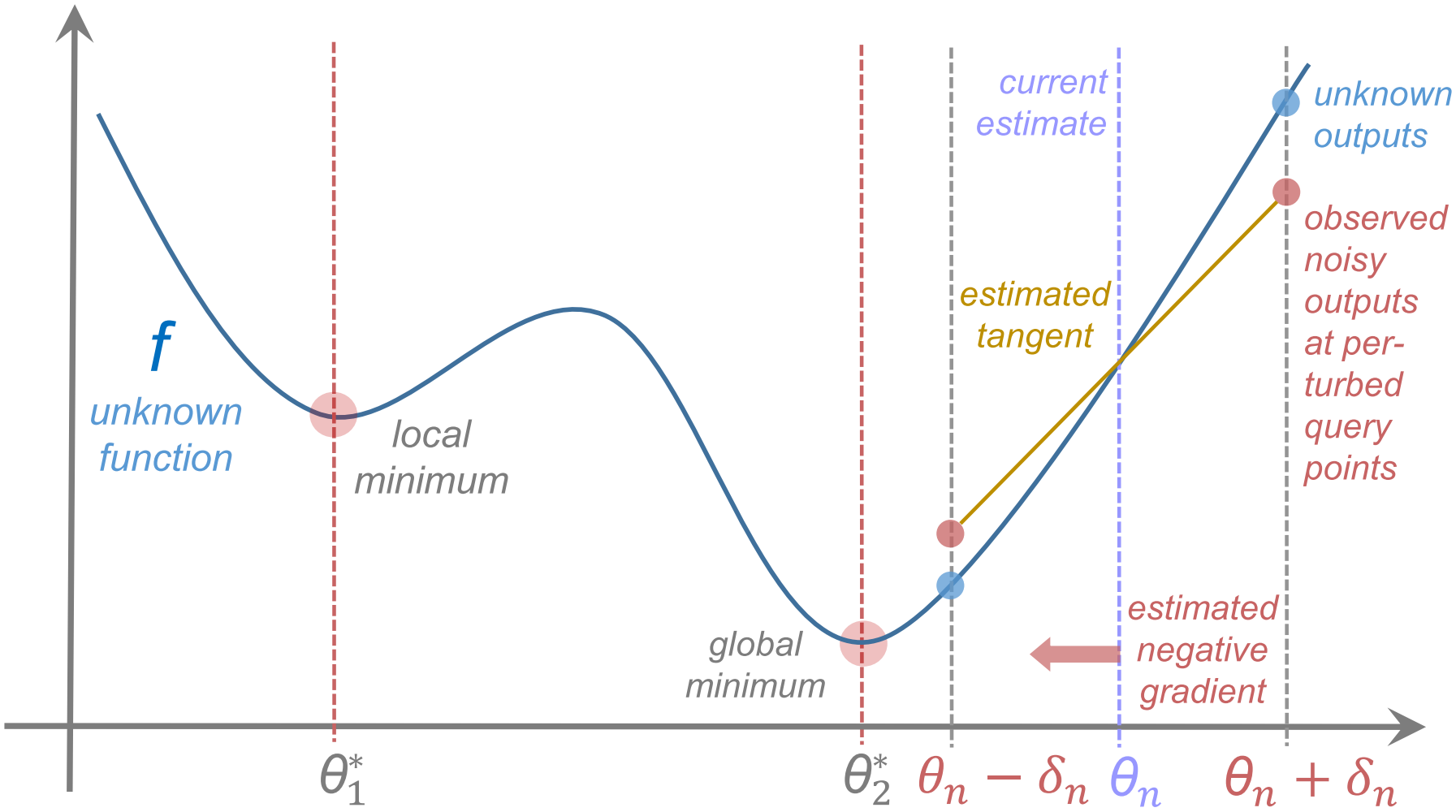
Kiefer-Wolfowitz (KF) Algorithm (1952)

$$\theta_{n+1} \doteq \theta_n + \gamma_n \frac{F_n^+(\theta_n - \delta_n) - F_n^-(\theta_n + \delta_n)}{2\delta_n},$$

where F_n^+, F_n^- are two independent queries with noises $\varepsilon_n^+, \varepsilon_n^-$.

- Terms $\{\delta_n\}$ define a finite difference interval (e.g., $\delta_n = n^{-1/4}$), i.e., the correction terms are estimates of the negative gradient of f.

Kiefer-Wolfowitz Algorithm (2)



Kiefer-Wolfowitz Algorithm (3)

– Let us assume that sequences $\{\gamma_n\}$ and $\{\delta_n\}$ are positive and

$$\sum_{n=0}^{\infty} \gamma_n = \infty, \qquad \lim_{n\to\infty} \delta_n = 0, \qquad \sum_{n=0}^{\infty} \frac{\gamma_n^2}{\delta_n^2} < \infty,$$

as well as (the unobserved, noiseless) function f satisfies

$$|f(\theta+1) - f(\theta)| < A|\theta| + B < \infty,$$

for all θ and suitably chosen constants A and B; and for all k:

$$\sup_{1/k<\theta^*-\theta<k} \frac{\partial}{\partial\theta} f(\theta) < 0, \qquad \text{and} \qquad \inf_{1/k<\theta-\theta^*<k} \frac{\partial}{\partial\theta} f(\theta) > 0.$$

– Then, for any initial estimate $\theta_0 \in \mathbb{R}$, the estimate sequence $\{\theta_n\}$ converges to θ^* both with probability one and in mean square.

– Blum (1954) proved the consistency of the $f \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ variant.

Simultaneous Perturbations

- KW in $\mathbb{R}^d$ requires $O(d)$ queries to estimate ∇f at a given point.
- Surprisingly, it can be estimated by just two, independently of d.
- This is achieved by SPSA by making queries in random directions.
- Let $\{\Delta_n\}$ be $\mathbb{R}^d$-valued random-vectors, which determine the random perturbations at each iteration for estimating the gradient.

Simultaneous Perturbation SA (SPSA) by Spall (1992)

$$\theta_{n+1,k} \doteq \theta_{n,k} + \gamma_n \frac{F_n^+(\theta_n - \delta_n\Delta_n) - F_n^-(\theta_n + \delta_n\Delta_n)}{2\delta_n\Delta_{n,k}},$$

for $k = 1, \ldots, d$, where F_n^+, F_n^- are the same as in KW.

- $\{\Delta_n\}$ are typically chosen as independent, symmetric, zero-mean, for example, i.i.d. Bernoulli with $\Delta_{n,k} = \pm 1$ with prob. $1/2$ each.

Stochastic Gradient Descent

– We want to minimize an unknown function, $f : \mathbb{R}^d \to \mathbb{R}$, based only on noisy queries about its gradient, ∇f, at selected points.

Stochastic Gradient Descent (SGD)

$$\theta_{n+1} \doteq \theta_n + \mu\left(-\nabla_\theta f(\theta_n) + \varepsilon_n\right)$$

– Polyak's heavy-ball or momentum method is defined as

SGD with Momentum Acceleration

$$\theta_{n+1} \doteq \theta_n + \mu\left(-\nabla_\theta f(\theta_n) + \varepsilon_n\right) + \gamma\left(\theta_n - \theta_{n-1}\right)$$

– The added term acts both as a smoother and an accelerator.
(The extra momentum dampens oscillations and helps us getting through narrow valleys, small humps and local minima.)

Some Challenges for Stochastic Approximation

- Non-asymptotic and distribution-free guarantees for SA.
- Guaranteeing global optimality for nonconvex problems.
- Analyzing SA methods in abstract (e.g., Hilbert) spaces.
- Combining SA with supervised learning approaches (e.g., deep learning or kernel methods), studying various representations.
- Handling the exploration-expoitation trade-off (especially in RL).
- Studying acceleration methods for the stochastic setting.
- Adapting to changing environments (changing dynamics).
- Extending the results to more general stochastic processes, under mild statistical assumptions (e.g., stationarity, ergodicity, mixing).
- Optimal choice of learning rates, generalized learning rates.
- SA in manifolds, combining SA with information geometry.

Summary

1. Machine learning (ML) is the essential part of artificial intelligence.
2. ML has a wide range of open mathematical problems and provides a great opportunity for mathematical development.
3. Stochastic approximation (SA) plays a principal role in several ML approaches. SA has many challenges to be addressed.

Thank you for your attention!

https://csaji.pages.sztaki.hu/ csaji@sztaki.hu