HDI Lab research projects 2023/24

Arthur Goldman, Evgeny Frolov, Egor Kosov, Eugene Lagutin,
Ilya Levin, Fedor Noskov, Nikita Puchkin, Sergey Samsonov,
Alexandra Senderovich, Daniil Tiapkin

Contents

1. MCMC, Markov Chains, and related topics

1.1. Adaptive Metropolis-Hastings techniques without analytic expression for proposal density

Contact persons: Sergey Samsonov and Eugene Lagutin,
svsamsonov@hse.ru, lagutin.em@phystech.edu

Suppose that we wish to sample from the distribution π on $\mathbb{R}^d$ with density $\pi(x)$. For some reasons we prefer to use MCMC approach, that is, we aim at constructing $(X_k)_{k=0}^{\infty}$ - ergodic Markov chain with stationary distribution π. Then we estimate $\pi(f)$ by

$$\pi_n(f) = n^{-1} \sum_{k=0}^{n-1} f(X_k).$$

Popular family of the algorithms is the one based on the Langevin dynamics. Consider the following Itô SDE:

$$\mathrm{d}X_t = -\nabla U(X_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \tag{1}$$

where U is some smooth function and $(W_t)_{t\geq 0}$ is a Wiener process. Under some regularity condition, the unique invariant distribution of (1) is given by $\pi(x) \propto e^{-U(x)}$. Hence it makes sense to consider first-order discretization of (1) with step size γ. This leads to the Unadjusted Langevin Algorithm (ULA):

$$X_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma}\xi_{k+1}, \quad i.i.d.\ \xi_k \sim \mathrm{N}(0, \mathrm{I}_d). \tag{2}$$

One can also consider Metropolis-adjusted Langevin Algorithm (MALA), which takes ULA iterate as a proposal and then apply Metropolis-Hastings correction. Check https://chi-feng.github.io/mcmc-demo for demonstration.

At the same time, in many generative models one does not have the closed form density expression of the proposal distribution. For example, this is the case of VAE's where only the estimate of the density likelihood is available (see e.g. [Burda et al., 2015] and [Thin et al., 2021]). The project aim is to study both the theoretical and empirical properties of the Metropolis-Hastings type algorithms with (unbiased) proposal density estimates.

Task:

1. Study the paper [Thin et al., 2021];
2. Implement the algorithm proposed in [Thin et al., 2021] and apply it as a post-processing technique for fine-tuning VAEs;
3. Implement the algorithm for the setting of GANs treated as energy-based models [Samsonov et al., 2022].

1.2. Adaptive MCMC with diffusion-based models

Contact persons: Sergey Samsonov and Eugene Lagutin,
svsamsonov@hse.ru, lagutin.em@phystech.edu

MCMC methods often rely on adaptive tuning of the so called *proposal distribution* in order to achieve better performance. In particular, the key problem is to design effective non-local moves and global proposals, and combine them with the local exploration steps. One can naturally combine the variational autoencoders and/or normalizing flows as a proposal generators for MCMC, see e.g. Gabrié et al. [2022], Thin et al. [2021]. It will be natural to generalize this adaptive approach to the diffusion modeling.

Task:

1. Study the paper [Hunt-Smith et al., 2023];
2. Study the papers on adaptive MCMC, e.g. [Samsonov et al., 2022], [Gabrié et al., 2022];
3. Combine the diffusion-based proposal with i-SIR type MCMC following [Samsonov et al., 2022]

1.3. Stein Variational Adaptive Importance Sampling

Contact person: Eugene Lagutin, lagutin.em@phystech.edu

Suppose that we wish to sample from the distribution π on $\mathbb{R}^d$ with density $\pi(x) \propto e^{-U(x)}$. For some reasons we prefer to use MCMC approach, that is, we aim at constructing $(X_k)_{k=0}^{\infty}$ - ergodic Markov chain with stationary distribution π. Then we estimate $\pi(f)$ by

$$\pi_n(f) = n^{-1} \sum_{k=0}^{n-1} f(X_k).$$

An Importance Sampling approach (specifically Self-Normalized Importace Sampling) aims to estimate $\pi_n(f)$ by

$$\hat{\pi}_n(f) = \sum_{k=0}^{n-1} f(X_k)\tilde{w}_{n,k},$$

where $X_k \overset{i.i.d.}{\sim} q$, $w(X_k) = \frac{e^{-U(X_k)}}{q(X_k)}$, $\tilde{w}_{n,k} = \frac{w(X_k)}{\sum_{k=0}^{n-1} w(X_k)}$. Distribution q with density $q(x)$ is called proposal distribution.
However finding good proposal distribution especially with large dimension d is a tough task. In this project the technique for constructing the approximation of the target density called Stein Variational Gradient Descent is proposed. The core idea of this method is to iteratively construct a sequence of transformations $T_l(x)$ in a way that the distribution $q_l = (T_l \circ T_{l-1} \cdots \circ T_1)\sharp q_0$ would be closer than q_0 to the target π. More precisely given a set of particles $\{X_k^l\}_{k=0}^{n-1}$ and density $q_l(x)$ at the step l the update step is

$$\begin{aligned} X_k^{l+1} &= X_k^l + \varepsilon \phi(X_k^l) \\ q_{l+1}(X_k^{l+1}) &= q_l(X_k^l)|\det(\mathrm{I} + \varepsilon \nabla \phi_{l+1}(\mathrm{X}_k^l)|^{\mathrm{k}-1}, \end{aligned}$$

where $\phi_{l+1}(X_k^l) = n^{-1}\sum_{k=0}^{n-1}[-\nabla_{Y_l^k}U(Y_l^k)k(Y_l^k,X_l^k)+\nabla_{Y_l^k}k(Y_l^k,X_l^k)]$ for $\{Y_k^l\}_{k=0}^{n-1}$ being independent copy of $\{X_k^l\}_{k=0}^{n-1}$ and $k(\cdot,\cdot)$ kernel function.

Task:

1. Study the paper [Han and Liu, 2017].
2. Implement the algorithm proposed in [Han and Liu, 2017] and reproduce experiments on Gaussian Mixture Model.
3. Implement and analyze the algorithm for constructing Iterative Sampling Importance Resampling algorithm with Stein Variational Adaptive proposal.

1.4. Extensive study of practical performance of CV VR method

Contact person: Artur Goldman, art-gold1579@yandex.ru

Recently, a big portion of Variance reduction methods has appeared, which are based on Control Variates. However, throughout different papers, there was barely any accurate performance study and comparison of methods between each other. In this project it is suggested to implement various methods, perform hyperparameter search and compare their performance on different cases (various applications in different dimensions). The closest paper to the suggested study is [Si et al., 2021].

Experiments should compare:

- Loss function: EV and ESV as in [Belomestny et al., 2020], EV with floating mean and other regularisations as in [Si et al., 2021, Sun et al., 2023]
- Type of Markov chain: ULA, MALA, NUTS. Dimension of data: the bigger the better
- Type of CV. For Stein CV: form $\Delta\varphi + \langle\nabla\log\pi,\nabla\varphi\rangle$ vs $\mathrm{div}(\varphi) + \langle\nabla\log\pi,\varphi\rangle$. Underlying class: polynomial functions, kernel functions, neural networks [Si et al., 2021].

Metrics to track: Training time, loss value, variance reduction rate, bias.

Possible datasets to test on:

- Synthetic datasets [Samsonov et al., 2022]
- Standard Bayesian inference examples [Si et al., 2021, Belomestny et al., 2020]
- https://github.com/stan-dev/posteriordb [Wang et al., 2023]

Good start to study code: see implementations of [Samsonov et al., 2022, Sun et al., 2023]

Task:

1. Implement experiments. Ideally develop a useful library with clean code which could be later used later in research by other groups.

Required skills: MLOps, PyTorch, Parallel programming, Probability theory

2. Reinforcement Learning and Stochastic Approximation

2.1. Reinforcement Learning with Tensor Decomposition

Contact person: Daniil Tiapkin, dtyapkin@hse.ru

The central object of the reinforcement learning is Markov Decision Process (MDP). An MDP is defined by a tuple $(\mathscr{S}, \mathscr{A}, p, r, H, s_1)$, where $p\colon \mathscr{S} \times \mathscr{A} \to \Delta(\mathscr{S})$ is a transition kernel, $r\colon \mathscr{S} \times \mathscr{A} \to [0,1]$ is a reward function, H is a finite horizon (for simplicity), s_1 is an initial state. The interaction protocol of the agent with an MDP is following. Agent starts at state s_1. For each $h \in [H]$ agent stays at state s_h, play action a_h, then the environment gives an agent noisy reward $r_h = r(s_h, a_h) + \xi$, where ξ is a zero-mean noise, and next state s_{h+1}. The goal of the agent is to find a policy $\pi_h^\star\colon \mathscr{S} \to \Delta(\mathscr{A})$ i.e. the mapping for each state to a distribution over next taken actions, that maximizes the expect sum of rewards. For a general exposition on theoretical RL we refer to [Agarwal et al., 2019].

From the theoretical point of view the main problem is that the sample complexity (i.e. minimal number of samples to achieve ε-optimal policy) must depend on number of states at least linearly: $\Omega(H^2SA/\varepsilon^2)$[1]. This bound is unimprovable for general finite MDPs, and this effect is called *curse of dimensionality* for RL. The only way to fight it is to impose additional structural assumptions.

In the case of multi-armed bandits ($S = H = 1$) the authors of [Zhou et al., 2022, Shi et al., 2023] suggest the following idea. Let us assume that the action space forms a product structure $\mathscr{A} = \mathscr{A}_1 \times \dots \mathscr{A}_d$. In this case the true reward function $r(a)$ could be considered as a multi-dimensional tensor $R = r(a_1, \dots, a_d)$ that in general have size q^d, where $q = |\mathscr{A}_i|$. However, we can assume that the tensor R have a small tensor rank in sense of Tucker decomposition. In the aforementioned papers this allows to reduce the dependence in a number of actions from initial q^d to $q^{d/2}r^{d/2}$ [Zhou et al., 2022] and even $q^2 r^{d-2}$ [Shi et al., 2023], where for constant q and r it allows to resolve the curse of dimensionality. There is a several possible directions in this problem.

The goal of this project is to generalize these result to the reinforcement learning problem by assuming similar structural assumption on the state-space and transition kernel $p(s,a)$. Alternatively, the student is welcomed to stick on the bandit problem and suggest another type of tensor decompositions, such as tensor train (TT) decomposition [Oseledets, 2011].

Prerequisites: Probability Theory, Matrix Computations, Lack of Fear of Math.

1. Study the minimal required part of theory of multi-armed bandits [Lattimore and Szepesvári, 2020] and reinforcement learning [Agarwal et al., 2019];
2. Study papers on low-rank tensor bandits [Zhou et al., 2022, Shi et al., 2023];
3. Implement proposed algorithm and compare them on simple environments;

4 (A). Propose a similar structural assumptions in the RL setting and propose modification of `UCBVI` algorithm [Azar et al., 2017];

[1] In step-homogeneous setting, where p and r is not changed each step.

4 (B). Propose an algorithms for bandits that uses TT-decomposition of reward function, and obtain improved rates.

5. Implement the algorithm from one of the previous steps on simple environments.

2.2. Two timescale linear stochastic approximation

Contact person: Sergey Samsonov, svsamsonov@hse.ru We consider the applications of the general LSA results to the problems of reinforcement learning. We use the notations introduced in **?**, which slightly differs from the ones introduced in the previous section. Namely, we focus on the discounted infinite-horizon setting instead of the episodic one.

We consider a problem of estimating the policy π in a discounted MDP (Markov Decision Process) given by a tuple $(\mathscr{S}, \mathscr{A}, \mathbb{P}, r, \gamma)$. Here, $\mathscr{S}$ and $\mathscr{A}$ stand for the respective (finite) state and action spaces, i.e., $|\mathscr{S}| < \infty$, $|\mathscr{A}| < \infty$, and $\gamma \in (0,1)$ is a discount factor. $\mathbb{P}$ stands for the transition kernel $\mathbb{P}(s'|s,a)$, which determines the probability of moving from state s to state s' when action a is performed. The reward function $r : \mathscr{S} \times \mathscr{A} \to [0,1]$ is assumed to be deterministic. The policy $\pi(\cdot|s)$ is a distribution over the action space $\mathscr{A}$ corresponding to the agent's action preferences in state $s \in \mathscr{S}$. Our goal is to estimate the agent's value function, which is defined as

$$V^{\pi}(s) = \mathbb{E}[\textstyle\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) | s_0 = s] \, ,$$

where $a_k \sim \pi(\cdot|s_k)$, and $s_{k+1} \sim \mathbb{P}(\cdot|s_k, a_k)$, for any $k \in \mathbb{N}$. We also define the transition matrix

$$\mathbb{P}^{\pi}(s'|s) = \textstyle\sum_{a \in \mathscr{A}} \mathbb{P}(s'|s,a)\pi(a|s) \, , \tag{3}$$

which corresponds to the transition probability from s to s' under policy π. Since the dimension of the state space $\mathscr{S}$ can be extremely large, we instead consider the linear approximation of the true value function $V^{\pi}(s)$, which is defined for $s \in \mathscr{S}$, $\theta \in \mathbb{R}^d$, and $\varphi : \mathscr{S} \to \mathbb{R}^d$ as

$$V_{\theta}^{\pi}(s) = \varphi^{\top}(s)\theta \, ,$$

where the feature dimension d is typically chosen such that $d \ll |\mathscr{S}|$. In this case $V_{\theta} = (V_{\theta}^{\pi}(s))_{s \in \mathscr{S}} \in \mathbb{R}^{|\mathscr{S}|}$ is the approximation vector to the true value function, and can be written as $V_{\theta} = \Phi\theta$, where $\Phi = [\varphi(s_1)|\ldots|\varphi(s_{|\mathscr{S}|})]^{\top} \in \mathbb{R}^{|\mathscr{S}| \times d}$ is the *feature matrix*.

We aim to estimate the *sample complexity* of estimating the optimal approximation to the value function.

1. Study the recent papers on non-asymptotic analysis of TD learning Patil et al. [2023], Li et al. [2023];

2. Study the papers on Polyak-Ruppert averaged LSA Durmus et al. [2022] and linear two-timescale SA Kaledin et al. [2020]

3. Generalize the results of Kaledin et al. [2020] for TDC (offline TD) algorithm with error expansion technique from Durmus et al. [2022].

Prerequisites: Probability theory, Linear algebra.

3. Audio Processing

3.1. Audio Clustering

Contact person: Alexandra Senderovich, asenderovich@hse.ru

Essentially, the task of clustering objects can be divided into two parts: obtaining good representations of data and applying some clustering algorithm. For the task of clustering objects with complex structure, such as images or audio, representations are usually taken from last layers of some neural network. The simplest choice would be to take a network pre-trained on the classification task. However, this solution might suffer from domain shift: we might want to cluster data that significantly differs from the pre-training set.

To avoid this problem, one should use methods for unsupervised learning. Some of these methods are specifically designed for solving clustering task. For images, the task of learning representations for clustering has been extensively studied. One of the simplest ideas (DeepCluster, [Caron et al., 2018]) is to iterate between steps of clustering data and training a classification model based on the classes obtained on the previous step. While this approach is interpretable and easy to implement, it does not produce good results in terms of accuracy. State-of-the-art method for image clustering is SPICE, proposed in [Niu and Wang, 2021]. It significantly reduced the gap between unsupervised and supervised classification: the difference between classification accuracy of Cifar-10 with this approach and the baseline is only 2%.

However, the task of audio clustering remains underexplored. Authors of one of the recent papers [Ghosh et al., 2021] tried to adapt the DeepCluster approach to the task of audio clustering. While their work is a good baseline to compare to, it also does not produce good accuracy. The idea of this project is to try to adapt the SPICE paper to the task of audio clustering:

1. Study the paper [Niu and Wang, 2021].
2. Identify the changes that have to be introduced in order to apply the same pipeline to audio (mostly changing augmentations).
3. Implement this algorithm and apply to different audio datasets.

3.2. Knowledge Distillation with applications in audio processing

Contact person: Ilya Levin, ivlevin@hse.ru

Knowledge distillation is a very hot topic in recent time, for example, two year old survey [Gou et al., 2021] has more than a 1k citations. Knowledge distillation is a method of NN-based model compression which transfers the perfomance of a bigger model(teacher) into a smaller one(student). This approach is especially helpful in Automatic Speaker Recognition(ASR) problem. In signal processing it is important to have small and fast models in order to inference them on mobile devices. Ensembling method applied to the teacher network can enhance the quality of the solution [Yang et al., 2023]. Also,

in Neural Language Processing knowledge distillation is also widely used. For example, there is a work [Chang et al., 2022] where authors propose an ensembling of student networks. Thus, this approach can be studied within ASR problem. There are several steps for it:

1. Study recent approaches for knowledge distillation in ASR;
2. Choose and apply the ensembling method for student network;
3. Compare with baselines according to the chosen scores.

4. Enumerating 2-distance sets in $\mathbb{R}^8$

Contact person: Fedor Noskov, fnoskov@hse.ru

Say a finite set $S \subset \mathbb{R}^d$ is 2-distance if there are two positive numbers a, b such that for any distinct points $x, y \in S$ we have $\|x - y\|_2 \in \{a, b\}$. In [Lisoněk, 1997], Petr Lisonek proposed an algorithm for enumeration of all 2-distance sets in $\mathbb{R}^d$.

In [Bondarenko, 2013], Andriy Bondarenko constructed a 2-distance set in $\mathbb{R}^{64}$ to disprove Borsuk's conjecture for 64-dimensional Euclidian space. Thus, we may try to find a suitable 2-distance set in $\mathbb{R}^8$ to check whether Bondarenko's method works in $\mathbb{R}^8$. The problem is that a naive realization of Lisonek's algorithm takes too much time for enumeration of all 2-distance sets. The goal of this project is to accelerate this algorithm with modern computing techniques. We expect the following steps of the research:

1. Study paper [Lisoněk, 1997].
2. Implement Lisonek's algorithm using C++ or other (fast enough) programming languages. The student can use an implementation in Sage provided us by Danilo Radchenko [Radchenko, 2021].
3. Check papers on enumeration of isomorphism classes of graphs with GPU.
4. Try to accelerate Lisonek's algorithm.
5. Enumerate 2-distance sets in $\mathbb{R}^8$ if possible.

5. Optimization and related questions

5.1. Finite-time deviation bounds for variational inequalitites

Contact person: Sergey Samsonov, svsamsonov@hse.ru

Number of recent papers study the finite-time analysis of min-max optimization problems. Most popular algorithms for such types of problems are Stochastic Extragradient (SEG) and Stochastic Gradient Descent Ascent (SGDA). People often study the constant-step size versions of the mentioned algorithms together with the trajectory average. Such type of techniques is known as Polyak-Ruppert averaging and can be shown to be optimal in various stochastic problems.

In recent paper [Vlatakis-Gkaragkounis et al., 2023] SEG-type dynamics is studied using the Markov chains technique. The authors manages to get asymptotic results, such as the law of large numbers and central limit theorem. At the same time, available techniques in Markov chains literature allows to obtain the results, which characterize the finite-time behavior of the algorithm.

1. Study the paper [Vlatakis-Gkaragkounis et al., 2023];
2. Study the paper [Durmus et al., 2023] for the technique required to deal with the deviation bounds for Markov chains;
3. Generalize the results of [Vlatakis-Gkaragkounis et al., 2023] for the case of finite-time p-moment error bounds of the last iterated and trajectory avergae, respectively.

Prerequisites: Probability theory, experience with optimization or Markov chains is desirable, but not mandatory.

5.2. Optimal decentralized algorithms on time-varying graphs

Contact person: Darina Dvinskikh, dmdvinskikh@gmail.com

For the current moment of time, there has been a significant increase in interest in decentralized optimization. This is because modern problems that arise, e.g., in machine learning, are high-dimensional optimization problems requiring processing huge amounts of data and training high-dimensional parameters. Thus, almost every modern problem cannot be solved numerically without distributed computing, e.g. training large neural networks. Moreover, the motivation for using decentralized optimization can be local data storage, which makes centralized collection prohibitive due to its privacy or other reasons. Over the past 10 years, great progress has been made in decentralized optimization over time-static graphs [Gorbunov et al., 2022]. For instance, the need to solve optimization problems on time-varying graphs comes from wireless networks, where computing devices may periodically disconnect from the distributed network because of poor connection or other reasons. However, decentralized optimization on time varying graphs, is significantly less studied [Rogozin et al., 2022], and a large number of open questions remain. Particularly, an optimal algorithm (optimality is in terms of the number of communications and the number of oracle calls) has not yet been proposed for non-smooth problems of (strongly) convex decentralized (stochastic) optimization time-varying graphs.

1. Study papers [Kovalev et al., 2021, Rogozin et al., 2022, Gorbunov et al., 2022]
2. Implement main decentralized algorithms
3. Try to accelerate these algorithms based on aforementioned papers and [Allen-Zhu and Hazan, 2016]

6. Statistics and statistical learning theory topics

6.1. Beyond realizable setting in the empirical risk minimizaton with dependent data

Contact person: Sergey Samsonov, svsamsonov@hse.ru

It is well-known, that learning from dependent data is more complicated then from i.i.d. observation. This fact holds true both for optimization and statistical learning problems, see e.g. Beznosikov et al. [2023]. The same holds true for the concentration bounds for additive functional, see e.g. Adamczak [2008], Durmus et al. [2023]. At the same time, there are recent papers, showing that the bounds might be not so pessimistic.

1. Study the paper [Ziemann et al., 2023];
2. Generalize the results for the realizable setting in RL (e.g. TD learning) using the framework of Durmus et al. [2023];

6.2. Optimal estimation in Mixed-Membership Stochastic Block Model

Contact person: Fedor Noskov, fnoskov@hse.ru

The simplest parametric model in network analysis is the Erdős-Rényi model [Erdos and Renyi, 1960], which assumes that edges in a network are generated independently with a fixed probability p, the single parameter of the model. The stochastic block model (SBM; [Holland et al., 1983]) is a more flexible parametric model that allows for communities or groups within a network. In this model, the network nodes are partitioned into K communities, and the probability p_{ij} of an edge between nodes i and j depends on only what communities these nodes belong to. The mixed-membership stochastic block model (MMSB; [Airoldi et al., 2009]) is a stochastic block model generalization, allowing nodes to belong to multiple communities with varying degrees of membership. This model is characterized by a set of community membership vectors, representing the probability of a node belonging to each community.

In the MMSB model, for each node $i \in [n]$, we assume that there exists a vector $\boldsymbol{\theta}_i \in [0,1]^K$ drawn from the $(K-1)$-dimensional simplex that determines the community membership probabilities for the given node. Then, a symmetric matrix $\mathbf{B} \in [0,1]^{K\times K}$ determines the relations inside and between communities. According to the model, the probability of obtaining the edge between nodes i and j is $\boldsymbol{\theta}_i^\top \mathbf{B}\boldsymbol{\theta}_j$. Importantly, in the considered model, we allow for self-loops.

More precisely, let us observe the adjacency matrix of the undirected unweighted graph $\mathbf{A} \in \{0,1\}^{n\times n}$. Under MMSB model $\mathbf{A}_{ij} = Bern(\mathbf{P}_{ij})$ for $1 \le i \le j \le n$, where $\mathbf{P}_{ij} = \boldsymbol{\theta}_i^\top \mathbf{B}\boldsymbol{\theta}_j = \rho\, \boldsymbol{\theta}_i^\top \mathbf{B}\boldsymbol{\theta}_j$. Here we denote $\mathbf{B} = \rho\overline{\mathbf{B}}$ with $\overline{\mathbf{B}} \in [0,1]^{K\times K}$ being a matrix with the maximum value equal to 1 and $\rho \in (0,1]$ being the sparsity parameter that is crucial for the properties of this model. Stacking vectors $\boldsymbol{\theta}_i$ into matrix Θ, $\Theta_i = \boldsymbol{\theta}_i^\top$, we get the following formula for the matrix of edge probabilities $\mathbf{P}$:

$$\mathbf{P} = \Theta\mathbf{B}\Theta^\top = \rho\,\Theta\overline{\mathbf{B}}\Theta^\top.$$

Recently, Noskov and Panov, 2023 constructed the optimal estimator of $\mathbf{B}$ under the assumption that for each community $k \in [K]$ there exist $\Omega(n)$ nodes S_k such that $\boldsymbol{\theta}_i = \mathbf{e}_k, i \in S_k$. If this assumption does not hold, one can improve the lower bound of estimation of $\mathbf{B}$.

The key steps of the project are as follows:

1. Study the paper by Noskov and Panov, 2023.
2. Construct a large enough family of hypotheses $(\Theta_i, \mathbf{B}_i)$ such that no estimator can distinguish them with high probability.
3. Obtain lower bounds for degree-corrected MMSB model.

6.3. Approximation properties of quantized neural networks

Contact person: Nikita Puchkin, npuchkin@hse.ru

The huge empirical success of neural networks attracted attention of many scientists. In particular, a lot of papers study approximation power of deep neural networks (see, for instance, [Yarotsky, 2017, Yarotsky and Zhevnerchuk, 2020, Belomestny et al., 2023] and references therein). As largest IT-companies try to incorporate neural network based technologies into mobile devices, the question of memory usage plays a larger role. To simplify neural network training and save memory, the engineers use neural networks with quantized weights (that is, weights with values from a fixed discrete set). The goal of this project is to study some properties of neural networks with weights from $\{-1, 0, 1\}$.

1. Study the papers on approximation properties of neural networks [Yarotsky, 2017, Yarotsky and Zhevnerchuk, 2020, Belomestny et al., 2023].
2. Consider deep neural networks with sigmoid activations and weights from $\{-1, 0, 1\}$. Try to approximate univariate linear functions $f(x) = ax + b$, $a, b \in \mathbb{R}$, and the quadratic function $g(x) = x^2$ with these neural networks.

*3. Using the results obtained on the previous step, derive an upper bound on the complexity of approximation of a smooth multivariate function on $[0, 1]^d$ with quantized neural networks.

6.4. Online change point detection

Contact person: Nikita Puchkin, npuchkin@hse.ru

Detecting a structural break in a time series is a long-standing statistical problem. Though the first works on this topic were published in the middle of the 20th century, this problem is still of interest of many researchers due to its practical importance, and the number of papers studying change point detection grows constantly (see, for example, our recent paper [Puchkin and Shcherbakova, 2023] for a brief exposition). The goal of the present project is to suggest a new sequential change point detection procedure based on the approaches from prediction with expert advice.

1. Get familiar with the problem of prediction with expert advice [Cesa-Bianchi and Lugosi, 2006, Section 2].

2. Study the properties of the exponentially weighted average forecaster [Cesa-Bianchi and Lugosi, 2006, Sections 2.1, 2.2] and the fixed share algorithm [György et al., 2005].

3. Consider a parametric change point detection problem as a problem of prediction with expert advice where an expert indexed by $\theta \in \Theta$ suffers a loss $(-\log p_\theta(X_t))$ on each round.

4. Assuming the parameter set Θ finite, suggest a change point detection procedure which uses the difference between the cumulative losses of exponential weighting and fixed-share algorithm as a test statistic. Note that in the stationary regime the losses of two forecasters should be close to each other, while in the presence of a change point the fixed-share algorithm should be significantly better.

5. Perform numerical experiments to illustrate the performance of the procedure.

*6. Extend the algorithm to the case of an infinite set Θ. The paper [Cao et al., 2018], where the authors consider the change point detection problem through the lens of online convex optimization, may be useful.

6.5. Covariance estimation via Bures-Wasserstein barycenters

Contact person: Nikita Puchkin, npuchkin@hse.ru

Let $X_1, \dots, X_n$ be i.i.d. random vectors in $\mathbb{R}^d$ with zero mean and covariance matrix $\Sigma = \mathbb{E}X_1X_1^\top$. The most common estimator for Σ is the sample covariance $\widehat{\Sigma}$, defined as

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_iX_i^\top = \operatorname*{argmin}_{S \succeq O} \frac{1}{n}\sum_{i=1}^n \left\| S - X_iX_i^\top \right\|_{\mathrm{F}}^2.$$

However, the Frobenius norm is not the only way to define a metric on the space of positive semidefinite matrices. For instance, one may consider the Bures-Wasserstein distance, given by

$$d_{\mathrm{BW}}^2(\Sigma_1, \Sigma_2) = \operatorname{Tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right) = W_2^2\big(\mathcal{N}(0,\Sigma_1), \mathcal{N}(0,\Sigma_2)\big),$$

where $W_2\big(\mathcal{N}(0,\Sigma_1), \mathcal{N}(0,\Sigma_2)\big)$ is the Kantorovich-Wasserstein optimal transport distance between two Gaussian measures. If Σ_1 and Σ_2 commute (i.e. $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$), then $d_{\mathrm{BW}}(\Sigma_1, \Sigma_2) = \|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|_{\mathrm{F}}$. The goal of the project is to study theoretical properties of the estimator

$$\widetilde{\Sigma} \in \operatorname*{argmin}_{S \succeq O} \frac{1}{n}\sum_{i=1}^n d_{\mathrm{BW}}^2(S, X_iX_i^\top),$$

also referred to as *empirical barycenter*. In the existing literature, the authors consider barycenters over sets of Gaussian measures with nondegenerate covariance matrices (in fact, they often require the covariances to have a bounded condition number). This is not applicable in our case, because the matrices $X_1X_1^\top, \dots, X_nX_n^\top$ are clearly degenerate.

1. Get familiar with the papers on Bures-Wasserstein barycenters, e.g., [Chewi et al., 2020, Kroshnin et al., 2021, Maunu et al., 2023].

2. Study the properties of empirical risk minimizers [Boucheron et al., 2005].

3. Derive an upper bound on $\mathbb{E}d^2_{\mathrm{BW}}(\widetilde{\Sigma}, \Sigma)$ with explicit dependence on the sample size n.

*4. Derive a dimension-free high-probability upper bound on $d^2_{\mathrm{BW}}(\widetilde{\Sigma}, \Sigma)$ with explicit dependence on the sample size n.

7. Probability and related topics

7.1. Rosenthal-type inequalities for random matrices with Markovian dependence

Contact person: Sergey Samsonov, svsamsonov@hse.ru

1. Study the paper Durmus et al. [2023] with Rosenthal inequalities for Markov chains in scalar case;
2. Study the paper Neeman et al. [2023] for the Bernstein-type bounds in Markovian setting;
3. Generalize the techniques from Durmus et al. [2023] to deal with the matrix-valued setting and get the precise variance term under the assumptions imposed in Neeman et al. [2023];

7.2. Poincaré and log-Sobolev inequalities from the Gardner-Zvavitch theorem

Contact person: Nikita Puchkin, npuchkin@hse.ru

Poincaré and log-Sobolev inequalities are important tools in statistical inference, concentration of measure, and sampling. The starting point in their proof is the Brunn-Minkowski inequality (see, for instance, [Bobkov and Ledoux, 2000]). Recently, Cordero-Erausquin and Rotem [2022] proved a counterpart of the Brunn-Minkowski inequality for rotation-invariant log-concave measures. In particular, this includes the standard Gaussian measure in $\mathbb{R}^n$ and resolves the Gardner-Zvavitch conjecture [Gardner and Zvavitch, 2010] positively. The goal of this project is to derive analogs of Poincaré and log-Sobolev inequalities based on the Brunn-Minkowski inequality for the Gaussian measure.

1. Study the paper [Bobkov and Ledoux, 2000].
2. Suggest and prove counterparts of Poincaré and log-Sobolev inequalities, starting from the Brunn-Minkowski inequality for the Gaussian measure.

7.3. Projection property

Contact person: Egor Kosov, ked_2006@mail.ru

Let f be a functional on the (sub)set of all d-dimensional probability distributions. With some abuse of notation, for a d-dimensional random vector X, let $f(X)$ be the same as the value of the functional f on the distribution of the vector X. As an example, we can consider the functional $f_{max}(X) = \sup_{x\in\mathbb{R}^d} \rho_X(x)$ where ρ_X is the density of X. This functional is defined on the subset of all absolutely continuous distributions. We say that a functional f satisfies **the projection property** if there is a numerical constant $C > 0$ such that, for every $n \in \mathbb{N}$, for every collection of i.i.d random vectors $X_1, \dots, X_n$, satisfying the assumption $\max_{1\le j\le n} f(X_j) \le 1$, one has $f(a_1X_1 + \dots + a_nX_n) \le C$ for every $a_1, \dots, a_n \in \mathbb{R}$, $a_1^2 + \dots + a_n^2 = 1$. It is known, that functional f_{max} satisfies the projection property for

all $d \in \mathbb{N}$ (see Bobkov and Chistyakov [2012], Rudelson and Vershynin [2015], Madiman et al. [2017]). For the derivative functional

$$f_1(X) := \int_{\mathbb{R}^d} |\nabla \rho_X(x)|\, dx = \mathbb{E}|p_X(X)|, \quad p_X(x) := \frac{\nabla \rho_X(x)}{\rho_X(x)} = \nabla \log \rho_X(x),$$

the projection property was established in Kosov [2022a] only for the case $d = 1$. For the Fisher information functional $f_2(X) := \mathbb{E}|p_X(X)|^2$ and functionals $f_{2k} := \mathbb{E}|p_X(X)|^{2k}$, $k \in \mathbb{N}$, the projection property was established in Bobkov [2019] (for $d = 1$).

There are two main questions:

1. Are there any other functionals that satisfy the projection property? For example, one may try to answer this question for the functionals

$$f_{L^p}(X) := \Big(\int_{\mathbb{R}^d} |\rho_X(x)|^p\, dx \Big)^{1/p}.$$

2. What is going with the already mentioned functionals in the case $d > 1$?

We also mention that one can study the projection property not only for one dimensional projections $a = (a_1, \dots, a_n)$, but for k-dimensional ones. In that case in place of a vector a there will be a matrix with k-dimensional image.

One can start with the following plan:

1. Study the papers Bobkov [2019], Madiman et al. [2017], Kosov [2022a].
2. Try to answer the question 2 above for f_{2k} and f_1.

7.4. Distances between norms of Gaussian vectors

Contact person: Egor Kosov, ked_2006@mail.ru

Let X and Y be two Gaussian random vectors with zero mean and covariance matrices Σ_X and Σ_Y respectively. Let λ_{jX} be the eigenvalues of the matrix Σ_X counting multiplicities, and ordered in descending order. Let $\Lambda_{kX}^2 := \sum\limits_{j=k}^{\infty} \lambda_{jX}^2$. Let λ_{jY} and Λ_{kY} be defined similarly for the vector Y. In Götze et al. [2019] the following bound was established

$$d_{\mathrm{Kol}}(|X|, |Y-a|) \le C\Big(\frac{1}{\sqrt{\Lambda_{1X}\Lambda_{2X}}} + \frac{1}{\sqrt{\Lambda_{1Y}\Lambda_{2Y}}} \Big) \Big(\|\Sigma_X - \Sigma_Y\|_{(1)} + |a|^2 \Big), \tag{4}$$

where $\|\cdot\|_{(1)}$ is the nuclear norm of a matrix and where d_{Kol} is the Kolmogorov distance:

$$d_{\mathrm{Kol}}(\xi, \eta) := \sup_{t \in \mathbb{R}} \big| P(\xi \le t) - P(\eta \le t) \big|.$$

In Kosov [2022b] the following bound was proved:

$$d_{\mathrm{TV}}\big(|X|, |Y-a|\big) \le \frac{160}{\sqrt{\lambda_{1X} \cdot \lambda_{2X}}} \big(\|\Sigma_X - \Sigma_Y\|_{HS} + |\mathrm{tr}\Sigma_X - \mathrm{tr}\Sigma_Y| + |a|^2 + |\Sigma_Y^{1/2} a| \big), \tag{5}$$

where $\|\cdot\|_{HS}$ is the Hilbert–Shmidt (Frobenius) norm of a matrix and where

$$d_{\mathrm{TV}}(\xi,\eta) := \sup\Big\{\mathbb{E}\big[\varphi(\xi)-\varphi(\eta)\big],\ \varphi\in C_0^\infty(\mathbb{R}),\ \|\varphi\|_\infty\le 1\Big\}.$$

In the right hand side of (5) the factor $\frac{1}{\sqrt{\lambda_{1X}\cdot\lambda_{2X}}}$ is bigger than $\frac{1}{\sqrt{\Lambda_{1X}\Lambda_{2X}}}$ from (4). Moreover, the expression $|\Sigma_Y^{1/2}a|$ depends only linearly on $|a|\to 0$. On the other hand, the expression

$$\|\Sigma_X-\Sigma_Y\|_{HS}+|\mathrm{tr}\Sigma_X-\mathrm{tr}\Sigma_Y|$$

is sharper than the nuclear norm $\|\Sigma_X-\Sigma_Y\|_{(1)}$ (e.g. in the case $\mathrm{tr}\Sigma_X=\mathrm{tr}\Sigma_Y$).

Taking into account these two inequalities, one may ask the following three questions:

The total variation distance is stronger than the Kolmogorov distance and thus, here appears the following question.

1. Is it true that

$$d_{\mathrm{Kol}}(|X|,|Y-a|)\le C\Big(\frac{1}{\sqrt{\Lambda_{1X}\Lambda_{2X}}}+\frac{1}{\sqrt{\Lambda_{1Y}\Lambda_{2Y}}}\Big)\Big(\|\Sigma_X-\Sigma_Y\|_{HS}+|\mathrm{tr}\Sigma_X-\mathrm{tr}\Sigma_Y|+|a|^2\Big)?$$

2. Is it true that

$$d_{\mathrm{TV}}\big(|X|,|Y-a|\big)\le\frac{C}{\sqrt{\Lambda_{1X}\cdot\Lambda_{2X}}}\big(\|\Sigma_X-\Sigma_Y\|_{HS}+|\mathrm{tr}\Sigma_X-\mathrm{tr}\Sigma_Y|+|a|^2+|\Sigma_Y^{1/2}a|\big)?$$

3. Is it true that

$$d_{\mathrm{TV}}\big(|X|,|Y-a|\big)\le\frac{C}{\sqrt{\Lambda_{1X}\cdot\Lambda_{2X}}}\big(\|\Sigma_X-\Sigma_Y\|_{HS}+|\mathrm{tr}\Sigma_X-\mathrm{tr}\Sigma_Y|+|a|^2\big)?$$

8. Recommender systems projects

Contact person: Evgeny Frolov, evgeny.frolov@skoltech.ru

8.1. Alternative negative sampling schemes for training RL models in RecSys

This is a subproject from a larger ongioing project at HSE led by Sergey Samsonov.

The current baseline in the project is an ensemble of 50+ models trained using bagging technique. This renders such model inefficient in dynamic recsys environments. The task is to take a single baseline RL model and improve its learning capabilities by properly tweaking the way data is exposed to the model during the training.

8.2. Improved learning for neural collaborative filtering

The main goal of the project is to derive intuition about the training behavior of typical architectures used in recommender systems and try to improve learning capabilities of those architectures based on thatknowledge.

This is a project in collaboration with Dmitry Vetrov. It is based on the toolset developed by D. Vetrov's team for analyzing global properties of surrogate loss functions but in application to specific scenarios in recommender systems and specific architectures based on Autoencoders.

Tasks:

- Study the techniques proposed by D. Vetrov, a good start is to watch Vetrov [2023].
- Adapt the techniques for the recommender systems domain.
- Analyze the possibility to improve learning behavior of recsys architectures.

8.3. Towards better understanding of latent factor models utility in top-n recommendation tasks

The main goal of the project is to establish the connection of the quality of matrix factorization-based models in top-n recommendation task to the spectral properties of their factor matrices. For example, link the properties of random matrices spectrum to the task of "flattening" of real collaborative filtering matrices spectrum.

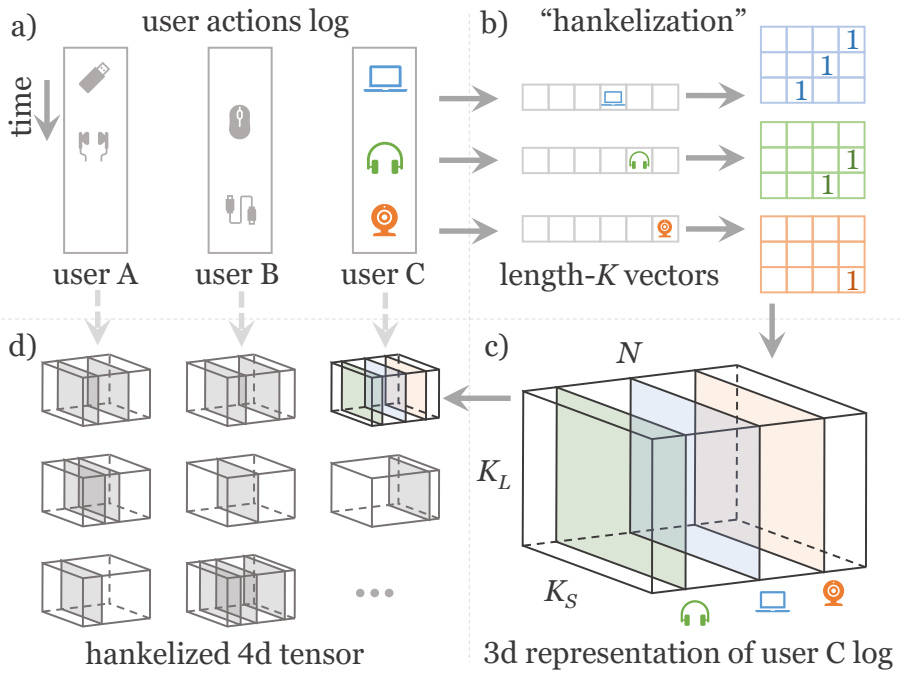
This is an exploratory research with high degree of uncertainty. Possible collaboration with Alexey Naumov and Maxim Rakhuba.

8.4. Hankelized tensor factorization for session-based recommenders

The goal is to adapt the existing approach of hankelized representations for sequential learning with tensor factorization to a particular case of the session-based recommendations. The project is based on the prior work Frolov and Oseledets [2023].

8.5. Positional Tensor Factorization with padding-induced sparse-dense fiber structure

The project is also based on the prior work Frolov and Oseledets [2023] and is devoted to sequential learning using tensor factorization models. The main idea is to pad sequences



that are shorter than the predefined dimensionality of sequential coordinate of the tensor. This will form dense fibers in otherwise extremely sparse data. The main goal is to utilize this special sparse-dense structure to derive efficient training algorithm and also to analyze how the padding affects the quality of recommendations.

8.6. Unlimited history positional Tensor factorization

This is another project on sequential learning based on the prior work Frolov and Oseledets [2023]. In this representation we will introduce one additional slice along the positional coordinate of the tensor that will encode all previous user interactions beyond the predefined sequence length. The main goal is to adapt existing algorithms to this format and analyze whether it allows to improve the quality of recommendations.

8.7. Hankelized tensor factorization in planetary-scale soil moisture analysis

The main goal of the project is to adapt the machinery of Hankelized representations in tensor factorization to the task of real-time analysis of climate data at a planetary scale. You will work with high-dimensional timeseries data coming from field sensors. The project is based on the prior work Frolov and Oseledets [2023]. It will be held in collaboration with the Center of Agricultural Technologies at Skoltech.

8.8. FPMC model with shared-factors Tensor Factorization

This project is aimed at testing the method of shared-factors tensor factorization, recently proposed by Maxim Rakhuba, to the existing tensor models in the recommender systems that conform the idea of having the same objects encoded in several modes of the same tensor. On of such model is a well-known Factorized Personalized Markov Chain model (FPMC) Rendle et al. [2010].

8.9. Knowledge Graph Link Prediction with Tensor Factorization

This project is devoted to an attempt to derive an alternative training scheme for a tensor with shared factors based on the combination of the ideas from RESCAL model Nickel and Tresp [2013] and low-rank representation. A useful related reading on knowledge graphs and the task of learning accurate representation of multirelational data: Chekalina et al. [2022].

8.10. Asymmetric tensor factorization for recommendations

Given a 3D tensor of user-item-context interactions data (e.g., user-item-position Frolov and Oseledets [2023] or user-item-rating Frolov and Oseledets [2016]), we typically learn a 3-factor model U, V, W that embeds objects along each mode of the tensor into a lower-dimensional space. On the other hand, for the task of item recommendations, the factor matrix of user embeddings U is not required, i.e. in the prediction matrix R for some user one has:

$$R = VV^\top PWW^\top,$$

where P is a "one-hot" matrix representing a user preferences (a slice of the tensor along the user mode), and matrices V, W encode item and "context" embeddings. Considering the fact that there can be millions or even billions of users, having to train the model that explicitly builds matrix U can be a challenging and cost-ineffective task.

The main goal of the project is to derive an efficient tensor factorization algorithm that will avoid materializing matrix U.

8.11. Scalable Softmax for extreme classification task in recommender systems

Recommender systems often have to deal with millions of hundreds of millions of items. One of the ubiquitously used objective function in training modern neural networks is a cross-entropy loss that captures distributions over entire item catalog. Considering the scale of rela-world recommender systems, the task becomes challenging for any reasonable batch sizes and the intermediate results may not fit into GPU memory during computations. The main goal of the project is to find good-enough heuristics on reducing the memory load during softmax computations. This project will be held in collaboration with he COmputational Intelligence Lab at Skoltech.

8.12. Convolutional Attention for Sequential Learning

This project is an attempt to take the advantages of the hankelized tensor factorization format Frolov and Oseledets [2023] and construct an alternative sequential attention architecture that will be more efficient then conventional one used in transformer networks without compromising its quality. The project will require reading and understanding recent techniques on CNN- and MLP-mixer-based techniques used in sequential learning tasks (particularly, recys and NLP).

8.13. Autoencoders with structured layers

This is an attempt to learn Autoencoder models with specially structured layer coming from a tensor format as in Marin et al. [2022]. The main promise of this approach is that it will help to learn additional hidden correlations in contextual data in recommender

systems. For example, it may help learning individual ratings scales of users in a more efficient way than standard context-aware methods.

8.14. Negative Sampling vs Hyperbolic Geometry

There is a strong connection of general data properties in recommender systems and hyperbolic geometry Mirvakhabova et al. [2020]. However, to take full advantage of the geometric approach, the corresponding neural network architecture of a recommender model must obey certain rules and restrictions. Otherwise, the learning ability of such networks falls short and they may even underperform the Eucliden counterparts. One of the components that seem to ruin the learning capabilities of the hyperbolic models is negative sampling.

The aim of the project is to verify this connection, demonstrate that depending on the negative sampling scheme the resulting hyperbolic model may or may not exhibit additional performance improvements in quality. Ideally, we want to derive a common set of recipes for constructing high-quality hyperbolic models.

8.15. Hyperbolic geometry vs popularity bias

This project is also related to the usage of hyperbolic geometry in recommender systems Mirvakhabova et al. [2020]. One of the key features of hyperbolic geometry is that it allows capturing hierarchical relations in data. In the recommender systems case, one of the major sources of hierarchy is the popularity of items. It is then natural to assume that there must exist a connection between the geometrical properties (e.g., space curvature) and the statistical or topological properties in the data. The project is aimed at establishing the connection between these two realms, which will help to build accurate hyperbolic models more effectively.

8.16. Feature selection by Mitigating Anchoring effects in RecSys

One of the major unsolved problem in recommender systems is a feature selection task. Hybrid recommender systems that utilize both behavioral information (i.e., what user purchase or rate) and side features (i.e., item description, user demographics) suffer from the "grabage in, garbage out" effect. The task of filtering out the "garbage" features is non-trivial (see e.g. Nikitin et al. [2022]) and remains largely underexplored in the literature. This project aims to utilize the recently proposed paradigm of feature anchoring for the selection task. Conceptually, the idea is to capture a noize coming from feature data in a special variational component of an architecture during the training and exclude this component during the recommendations phase.

8.17. Dynamic feature weighting in hybrid models

This project is devoted to the hybrid recommender systems as well, i.e. the aim is to effectively utilize both behavioral information (i.e., what user purchase or rate) and side features (i.e., item description, user demographics). It is a common paradigm to treat features as static, e.g. not evolving and being fully descriptive of the entity they belong to at any moment of time. On the other hand, features may depend on general trends and evolvig user interest so that the contribution of features into the prediction may change

over time. The main goal of the project is to build such an architecture that will capture feature dynamics in evolving user interests. This is an exploratory research project.

References

Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. 2008.

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pages 10–4, 2019.

Edo M Airoldi, David Blei, Stephen Fienberg, and Eric Xing. Mixed membership stochastic blockmodels. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21, pages 33–40. Curran Associates, Inc., 2009.

Zeyuan Allen-Zhu and Elad Hazan. Optimal black-box reductions between optimization objectives. *Advances in Neural Information Processing Systems*, 29, 2016.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov, and S. Samsonov. Variance reduction for markov chains with application to mcmc, 2020.

Denis Belomestny, Alexey Naumov, Nikita Puchkin, and Sergey Samsonov. Simultaneous approximation of a smooth function and its derivatives by deep neural networks with piecewise-polynomial activations. *Neural Networks*, 161:242–253, 2023.

Aleksandr Beznosikov, Sergey Samsonov, Marina Sheshukova, Alexander Gasnikov, Alexey Naumov, and Eric Moulines. First order methods with markovian noise: from acceleration to variational inequalities. *arXiv preprint arXiv:2305.15938*, 2023.

S. G. Bobkov. Moments of the scores. *IEEE Transactions on Information Theory*, 65(9): 5294–5301, 2019.

S. G. Bobkov and M. Ledoux. From Brunn-Minkowski to Brascamp-Lieb and to logarithmic Sobolev inequalities. *Geometric and Functional Analysis*, 10(5):1028–1052, 2000.

S.G. Bobkov and G. P. Chistyakov. Bounds on the maximum of the density for sums of independent random variables. *Zapiski Nauchnykh Seminarov POMI*, 408:62–73, 2012.

Andriy V. Bondarenko. On Borsuk's conjecture for two-distance sets. *arXiv:1305.2584 [math]*, August 2013. URL http://arxiv.org/abs/1305.2584. arXiv: 1305.2584.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability & Statistics*, 9:323–375, 2005.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Yang Cao, Liyan Xie, Yao Xie, and Huan Xu. Sequential change-point detection via online convex optimization. *Entropy*, 20(2):108, 2018.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, 2018.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

Xiaoqin Chang, Sophia Yat Mei Lee, Suyang Zhu, Shoushan Li, and Guodong Zhou. One-teacher and multiple-student knowledge distillation on sentiment classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7042–7052, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.614.

Viktoriia Chekalina, Anton Razzhigaev, Albert Sayapin, Evgeny Frolov, and Alexander Panchenko. Meker: memory efficient knowledge embedding representation for link prediction and question answering. *arXiv preprint arXiv:2204.10629*, 2022.

Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J. Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1276–1304. PMLR, 2020.

Dario Cordero-Erausquin and Liran Rotem. Improved log-concavity for rotationally invariant measures of symmetric convex sets. Preprint, arXiv:2111.05110, 2022.

Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Finite-time high-probability bounds for polyak-ruppert averaged iterates of linear stochastic approximation. *arXiv preprint arXiv:2207.04475*, 2022.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Marina Sheshukova. Rosenthal-type inequalities for linear statistics of markov chains. *arXiv preprint arXiv:2303.05838*, 2023.

P. Erdos and A. Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–61, 1960.

Evgeny Frolov and Ivan Oseledets. Fifty shades of ratings: How to benefit from a negative feedback in top-n recommendations tasks. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 91–98, 2016.

Evgeny Frolov and Ivan Oseledets. Tensor-based sequential learning via hankel matrix representation for next item recommendations. *IEEE Access*, 11:6357–6371, 2023.

Marylou Gabrié, Grant M Rotskoff, and Eric Vanden-Eijnden. Adaptive monte carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119(10):e2109420119, 2022.

R. J. Gardner and A. Zvavitch. Gaussian Brunn-Minkowski inequalities. *Transactions of the American Mathematical Society*, 362(10):5333–5353, 2010.

Sreyan Ghosh, Sandesh Katta, Ashish Seth, and Santhosh Umesh. Deep clustering for general-purpose audio representations, 10 2021.

Eduard Gorbunov, Alexander Rogozin, Aleksandr Beznosikov, Darina Dvinskikh, and Alexander Gasnikov. Recent theoretical advances in decentralized distributed convex optimization. In *High-Dimensional Optimization and Probability: With a View Towards Data Science*, pages 253–325. Springer, 2022.

F. Götze, A. Naumov, V. Spokoiny, and V. Ulyanov. Large ball probabilities, gaussian comparison and anti-concentration. *Bernoulli*, 25:2538–2563, 2019.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

András György, Tamás Linder, and Gábor Lugosi. Tracking the best of many experts. In *Learning Theory*, pages 204–216, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

Jun Han and Qiang Liu. Stein variational adaptive importance sampling. *arXiv preprint arXiv:1704.05201*, 2017.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

NT Hunt-Smith, W Melnitchouk, F Ringer, N Sato, AW Thomas, and MJ White. Accelerating markov chain monte carlo sampling with diffusion models. *arXiv preprint arXiv:2309.01454*, 2023.

Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with markovian noise. In *Conference on Learning Theory*, pages 2144–2203. PMLR, 2020.

E. D. Kosov. Regularity of linear and polynomial images of skorohod differentiable measures. *Advances in Mathematics*, 397:108193, 2022a.

E. D. Kosov. Distributions of second order polynomials in gaussian random variables. *Mathematical Notes*, 111:71–81, 2022b.

Dmitry Kovalev, Elnur Gasanov, Alexander Gasnikov, and Peter Richtarik. Lower bounds and optimal algorithms for smooth and strongly convex decentralized optimization over time-varying networks. *Advances in Neural Information Processing Systems*, 34: 22325–22335, 2021.

Alexey Kroshnin, Vladimir Spokoiny, and Alexandra Suvorikova. Statistical inference for Bures-Wasserstein barycenters. *The Annals of Applied Probability*, 31(3):1264–1298, 2021.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Gen Li, Weichen Wu, Yuejie Chi, Cong Ma, Alessandro Rinaldo, and Yuting Wei. Sharp high-probability sample complexities for policy evaluation with linear function approximation. *arXiv preprint arXiv:2305.19001*, 2023.

Petr Lisoněk. New Maximal Two-Distance Sets. *Journal of Combinatorial Theory, Series A*, 77(2):318–338, February 1997. ISSN 00973165. doi: 10.1006/jcta.1997.2749.

M. Madiman, J. Melbourne, and P. Xu. Rogozin's convolution inequality for locally compact groups. *arXiv preprint arXiv:1705.00642*, 2017.

Nikita Marin, Elizaveta Makhneva, Maria Lysyuk, Vladimir Chernyy, Ivan Oseledets, and Evgeny Frolov. Tensor-based collaborative filtering with smooth ratings scale. *arXiv preprint arXiv:2205.05070*, 2022.

Tyler Maunu, Thibaut Le Gouic, and Philippe Rigollet. Bures-wasserstein barycenters and low-rank matrix recovery. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8183–8210. PMLR, 2023.

Leyla Mirvakhabova, Evgeny Frolov, Valentin Khrulkov, Ivan Oseledets, and Alexander Tuzhilin. Performance of hyperbolic geometry models on top-n recommendation tasks. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 527–532, 2020.

Joe Neeman, Bobby Shi, and Rachel Ward. Concentration inequalities for sums of markov dependent random matrices. *arXiv preprint arXiv:2303.02150*, 2023.

Maximilian Nickel and Volker Tresp. Tensor factorization for multi-relational learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 617–621. Springer, 2013.

Artyom Nikitin, Andrei Chertkov, Rafael Ballester-Ripoll, Ivan Oseledets, and Evgeny Frolov. Are quantum computers practical yet? a case for feature selection in recommender systems using tensor networks. *arXiv preprint arXiv:2205.04490*, 2022.

Chuang Niu and Ge Wang. Spice: Semantic pseudo-labeling for image clustering, 2021.

Fedor Noskov and Maxim Panov. Optimal estimation in mixed-membership stochastic block models, 2023. URL https://arxiv.org/abs/2307.14530.

I. V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33 (5):2295–2317, 2011. doi: 10.1137/090752286. URL https://doi.org/10.1137/090752286.

Gandharv Patil, LA Prashanth, Dheeraj Nagaraj, and Doina Precup. Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation. In *International Conference on Artificial Intelligence and Statistics*, pages 5438–5448. PMLR, 2023.

Nikita Puchkin and Valeriia Shcherbakova. A contrastive approach to online change point detection. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5686–5713. PMLR, 2023.

Danilo Radchenko. Private communication, 2021.

Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 811–820, 2010.

Alexander Rogozin, Alexander Gasnikov, Aleksander Beznosikov, and Dmitry Kovalev. Decentralized optimization over time-varying graphs: a survey. *arXiv preprint arXiv:2210.09719*, 2022.

M. Rudelson and R. Vershynin. Small ball probabilities for linear images of high-dimensional distributions. *International Mathematics Research Notices*, 2015(19): 9594–9617, 2015.

Sergey Samsonov, Evgeny Lagutin, Marylou Gabrié, Alain Durmus, Alexey Naumov, and Eric Moulines. Local-global mcmc kernels: the best of both worlds. *Advances in Neural Information Processing Systems*, 35:5178–5193, 2022.

Chengshuai Shi, Cong Shen, and Nicholas D. Sidiropoulos. On high-dimensional and low-rank tensor bandits, 2023.

Shijing Si, Chris. J. Oates, Andrew B. Duncan, Lawrence Carin, and François-Xavier Briol. Scalable control variates for monte carlo methods via stochastic optimization, 2021.

Zhuo Sun, Chris J. Oates, and François-Xavier Briol. Meta-learning control variates: Variance reduction with limited data, 2023.

Achille Thin, Nikita Kotelevskii, Arnaud Doucet, Alain Durmus, Eric Moulines, and Maxim Panov. Monte carlo variational auto-encoders. In *International Conference on Machine Learning*, pages 10247–10257. PMLR, 2021.

Dmitry Vetrov. Surprising properties of loss landscape in over-parameterized models, 2023. URL https://www.youtube.com/watch?app=desktop&v=4RYeLmFeyWo. YouTube video.

Emmanouil-Vasileios Vlatakis-Gkaragkounis, Angeliki Giannou, Yudong Chen, and Qiaomin Xie. Stochastic methods in variational inequalities: Ergodicity, bias and refinements. *arXiv preprint arXiv:2306.16502*, 2023.

Congye Wang, Wilson Chen, Heishiro Kanagawa, and Chris. J. Oates. Stein π-importance sampling, 2023.

Xiaoyu Yang, Qiujia Li, Chao Zhang, and Philip C. Woodland. Knowledge distillation from multiple foundation models for end-to-end speech recognition, 2023.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 13005–13015, 2020.

Jie Zhou, Botao Hao, Zheng Wen, Jingfei Zhang, and Will Wei Sun. Stochastic low-rank tensor bandits for multi-dimensional online decision making, 2022.

Ingvar Ziemann, Stephen Tu, George J Pappas, and Nikolai Matni. The noise level in linear regression with dependent data. *arXiv preprint arXiv:2305.11165*, 2023.