NUS National University of Singapore | IMS Institute for Mathematical Sciences

SCIENTIFIC REPORTS

The Mathematics of Data

02 Jan 2024–26 Jan 2024

Organizing Committee

Afonso S. Bandeira
ETH Zurich

Subhroshekhar Ghosh
National University of Singapore

Philippe Rigollet
Massachusetts Institute of Technology

Hong T.M. Chu
National University of Singapore

CONTENTS PAGE

ROBUST ESTIMATION AND REGRESSION WITH MMD

PIERRE ALQUIER

Classification AMS 2020: 62F10, 62F35, 62J02, 68T05, 46E22.

Keywords: universal estimation; robust statistics; kernel methods; minimum distance estimation.

Popular estimation methods in statistics, such as the maximum likelihood estimator (MLE), or the method of moments, require strong assumptions on the statistical model and the data-generating process to converge. When these conditions are not met, these estimators can become very unstable. We are interested in universal estimators, that would converge without assumptions on the model. For example, [15] proved that so-called minimum distance estimators converge under far more general assumptions than the MLE. More recently, the ρ-estimators defined in [4] are not only convergent, but minimax-optimal, in a very large class of models. However, there are still a few limitating assumptions in [15, 4], and the computation of these estimators might be difficult in practice.

This talk will summarize a recent line of work on a variant of minimum distance estimators based on the so-called maximum mean discrepancy (MMD). In the case of i.i.d. data, these estimators converge without *any* assumption on the model nor on the data generating process. Moreover, relatively efficient algorithms are available to compute these estimators.

Let $X_1, \dots, X_n$ be $\mathcal{X}$-valued random variables i.i.d. from a probability distribution P^0. Let $\mathcal{M} = (P_\theta, \theta \in \Theta)$ be a statistial model. Note that we don't assume $P^0 \in \mathcal{M}$. Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) equipped with a scalar product $\langle \cdot, \cdot \rangle$, its associated norm $\| \cdot \|$ and a kernel k: there is a map $\varphi : \mathcal{X} \to \mathcal{H}$ with $k(x,y) = \langle \varphi(x), \varphi(y) \rangle$. The kernel mean embedding (KME) is defined for any probability P on $\mathcal{X}$ by $\mu(P) = \mathbb{E}_{X \sim P}[\varphi(X)]$. We refer the reader to [11] for more details on this construction. In particular: if k is *bounded*, then the KME is indeed well-defined for any P; if k is *characteristic* (see [11] for a definition), μ is one-to-one. Thus, if k is both bounded and characteristic, $D_k(P, P') = \|\mu(P) - \mu(P')\|$ defines a metric on probabilities over $\mathcal{X}$. Finally, [11] provides examples of kernels that are indeed bounded and characteristic: for example, when $\mathcal{X} = \mathbb{R}^d$, the Gaussian kernel $k(x,y) = \exp(-\|x-y\|^2/\gamma)$.

Definition 0.1. *We define the MMD-minimum distance estimator, or MMD-MDE, as*

$$\hat{\theta} = \arg\min_{\theta \in \Theta} D_k(P_\theta, \hat{P}_n)$$

where $\hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ *is the empirical distribution of the sample* $X_1, \dots, X_n$.

In the talk, I will prove the following result:

Theorem 0.2 (Theorem 3.1 in [6]). *Assume $k \leq 1$, then*

$$\mathbb{E}\big[D_k(P_{\hat{\theta}}, P^0)\big] \leq \inf_{\theta\in\Theta} D_k(P_\theta, P^0) + \frac{2}{\sqrt{n}}.$$

The fact that there are no assumptions on P^0 nor on $\mathcal{M} = (P_\theta, \theta \in \Theta)$ in the theorem is not due to lack of space: this is actually a nice feature of $\hat{\theta}$! Variants of Theorem 0.2, including a bound that holds with large probability rather than in expectation, can be found in [5, 6].

Various topics on $\hat{\theta}$ will be covered in this talk:

- the MMD distance can be written in terms of expectations, and thus, the minimization in Definition 0.1 can be done with a stochastic gradient method. Details are discussed in [5, 7].
- Theorem 1 leads to convergence: if the model is well-specified, $P^0 \in \mathcal{M}$, then $\mathbb{E}[D_k(P_{\hat{\theta}}, P^0)] \leq 2/\sqrt{n} \to 0$ when $n \to \infty$. It also leads to robustness as defined by Huber: if $P^0 = (1-\varepsilon)P_{\theta^0} + \varepsilon Q$ for an arbitrary contamination Q and a small ε, then $\mathbb{E}[D_k(P_{\hat{\theta}}, P^0)] \leq 4\varepsilon + 2/\sqrt{n}$. This is proven in [7], where we also prove more difficult robustness results under adversarial contamination of the data.
- the asymptotic normality of $\hat{\theta}$ is studied in [5] (this requires assumptions).
- an extention of Theorem 0.2 to non-i.i.d. observations (time series) is provided in [7], under a new mixing condition. We also prove that this mixing condition is less restrictive than the standard β-mixing condition.
- the term $2/\sqrt{n}$ in Theorem 0.2 cannot be improved *in general*. However, under assumptions on the variance of P^0, it can actually be improved: variance-aware versions of Theorem 0.2 are proven in [14].
- we can define Bayesian-flavored estimators by using the MMD to define a pseudo-likelihood. We studied such estimators in [7] and proved their convergence using tools from the PAC-Bayes theory [1]. Other Bayesian-inspired variants of $\hat{\theta}$ are based on sampling [8] and Approximate Bayesian Computation or ABC [10].
- this estimation strategy was successfully implemented in a wide range of applications: generative artificial intelligence [9], quantisation and clustering [13], estimation of copulas [2], estimation of parameters in stochastic PDEs [5] etc.

A large part of the talk will be dedicated to parametric regression (linear, or not). In this case, the observations are pairs input-output: $X_i = (Z_i, Y_i)$. Theorem 0.2 guarantees that we can estimate the joint distribution of these pairs. But this is not relevant in regression, where the objective is rather the estimation of the conditional distribution of Y_i given Z_i.

The problem is that, while the theory of KME looks simple and elegant, a rigorous definition of conditional KME turns out to be far more difficult and cumbersome! We refer the reader to [12] for a recent account and a general approach to solve the problem.

In our recent paper [3], we proved that, in standard regression models: linear regression with Gaussian noise, logistic regression, Poisson regression, etc., the conditions for the existence of conditional KMEs are met. This can be used to define consistent and robust estimation of the regression parameters in the spirit of

Definition 0.1 above. These estimators turn out to perform extremely when compared to existing robust regression procedures.

REFERENCES

[1] Pierre Alquier. User-friendly Introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2), 174–203, 2024.

[2] Pierre Alquier, Badr-Eddine Chérief-Abdellatif, Alexis Derumigny and Jean-David Fermanian. Estimation of copulas via maximum mean discrepancy. *Journal of the American Statistical Association*, 118(543), 1997-2012, 2023.

[3] Pierre Alquier and Mathieu Gerber. Universal robust regression via maximum mean discrepancy. *Biometrika*, 111(1), 71–92, 2024.

[4] Yannick Baraud, Lucien Birgé and Mathieu Sart. A new method for estimation and model selection: ρ-estimation. *Inventiones mathematicae*, 207(2), 425–217, 2017.

[5] Francois-Xavier Briol, Alessandro Barp, Andrew B. Duncan, and Mark Girolami. *Statistical inference for generative models with maximum mean discrepancy*, arXiv preprint arXiv:1906.05944, 2019.

[6] Badr-Eddine Chérief-Abdellatif and Pierre Alquier. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. *Symposium on Advances in Approximate Bayesian Inference*, Proceedings in Machine Learning Research 118, 1–21, 2020.

[7] Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence. *Bernoulli*, 28(1), 181–213, 2022.

[8] Charita Dellaporta, Jeremias Knoblauch, Theodoros Damoulas and Francois-Xavier Briol. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. *International Conference on Artificial Intelligence and Statistics*, Proceedings in Machine Learning Research 151, 943–970, 2022.

[9] Gintare Karolina Dzuigaite, Daniel Roy and Zoubin Ghahramani. *Training generative neural networks via maximum mean discrepancy optimization*, arXiv preprint arXiv:1505.03906, 2015.

[10] Sirio Legramanti, Daniele Durante, and Pierre Alquier. *Concentration and robustness of discrepancy-based ABC via Rademacher complexity*, arXiv preprint arXiv:2206.06991, 2022.

[11] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur and Bernhard Schölkopf. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends® in Machine Learning*, 10(1–2), 1–141, 2017.

[12] Junhyung Park and Krikamol Muandet A measure-theoretic approach to kernel conditional mean embeddings. *Advances in neural information processing systems* 33, 21247–21259, 2020.

[13] Onur Teymur, Jackson Gorham, Marina Riabiz, and Chris Oates. Optimal quantisation of probability measures using maximum mean discrepancy. *International Conference on Artificial Intelligence and Statistics*, Proceedings in Machine Learning Research 130, 1027–1035, 2021.

[14] Geoffrey Wolfer and Pierre Alquier. *Variance-aware estimation of kernel mean embedding*, arXiv preprint arXiv:2210.06672, 2022.

[15] Yannis Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13(2), 768–774, 1985.

ESSEC BUSINESS SCHOOL, ASIA-PACIFIC CAMPUS, 5 NEPAL PARK, 139408 SINGAPORE
Email address: alquier@essec.edu

FORMULATION AND RESOLUTION OF INVERSE PROBLEMS IN SIGNAL AND IMAGE PROCESSING - FROM CLASSICAL METHODS TO HYBRID AI

CAROLINE CHAUX

Classification AMS 2020: 94A08, 68T07

Keywords: Unrolled algorithms, parameter estimation, maximum a posteriori, wavelets, sparsity, low-rank, deconvolution, robust PCA.

We are interested here in inverse problems arising in signal and image processing. Solving such problems firstly consists in formalising the direct problem by understanding the physics behind and secondly, solving the associated inverse problem which results in solving an optimization problem. Two main ingredients are involved in the optimization functional to be minimized: a data fidelity term accounting for the model and a regularisation term (possibly several ones) accounting for prior information on the targeted solution. Regularisation parameters come into the play to guarantee the best trade-off between the two quantities.

Classical optimization-based approaches consist in, once the optimization problem has been formulated, proposing iterative procedures (e.g. proximal algorithms [1]) converging to a solution of the considered inverse problem. In such a case, most of the time, the parameters (either regularisation or algorithm's hyperparamaters) are fixed empirically. More recently, unrolled or unfolded neural networks have been proposed [2]. They combine optimization and learning, constitute interpretable networks, and integrate information about the direct model.

Inverse problems are encountered in many scientific domains such as biology, medical imaging, chemistry, audio signal processing for which, different tasks must be tackled such as deconvolution, restoration, unmixing, missing data reconstruction, etc.

We presented in this talk a physics-informed unrolled network [3] to automatically choose regularisation parameters in image deconvolution when the regularisation is performed in the wavelet domain. The resulting optimisation problem is solved by using an unrolled version of FISTA algorithm [4]. Indeed, hyper-parameter tuning, and especially regularisation parameter estimation, is a challenging but essential task when solving inverse problems. We proposed to perform their estimation under an unrolled strategy together with the inverse problem solving. The resulting network is trained while incorporating information on the model through Maximum a Posteriori estimation which drastically decreases the amount of data needed for the training and results in better estimation results.

We also presented a physics-informed unrolled network [5] to solve the problem of Robust Principal Component Analysis (RPCA) which consists in decomposing a matrix into the sum of a low rank matrix and a sparse matrix. We proposed a deep unrolled algorithm based on an accelerated alternating projection algorithm [6] which aims to solve RPCA in its nonconvex form and where hyperparameters are automatically learnt. We demonstrated the unrolled algorithm's effectiveness on synthetic datasets and also

on a face modelling problem, where it leads to both better numerical and visual performances.

These works have been done in collaboration with Vincent Tan, Emmanuel Soubiès, Pascal Nguyen and Elisabeth Tan.

REFERENCES

[1] P. Combettes and J.-C. Pesquet, Proximal Splitting Methods in Signal Processing. *Springer Optimization and Its Applications* , Springer New York, pp. 185-212, 2011.

[2] V. Monga, Y. Li, and Y. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, 2021.

[3] P. Nguyen, E. Soubies, C. Chaux. MAP-informed Unrolled Algorithms for Hyper-parameter Estimation. *Proc. Int. Conf. Image Process.*, Kuala Lumpur, Malaysia, Oct. 8-11, 2023.

[4] A. Beck and M. Teboulle. A fast iterative shrinkage thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[5] E. Z. C. Tan, C. Chaux, E. Soubies, and V. Y. F. Tan. Deep Unrolling for Nonconvex Robust Principal Component Analysis. *IEEE Int. Workshop Mach. Learn. Signal Process.*, Rome, Italy, Sep. 17-20, 2023.

[6] H. Cai, J.-F. Cai, and K. Wei. Accelerated alternating projections for robust principal component analysis. *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 685–717, Jan 2019.

IPAL AND CNRS IRL 2955, CNRS@CREATE, 1 CREATE WAY, #08-01 CREATE TOWER, SINGAPORE 138602

Email address: caroline.chaux@cnrs.fr

MINI-COURSE ON OPTIMAL TRANSPORT AND HIGH-DIMENSIONAL PROBABILITY

SINHO CHEWI AND YAIR SHENFELD

The mini-course focused on the foundations and applications of optimal transport and high-dimensional probability. The course was composed of 8 lectures whose outline is given below.

Lecture 1. The notion of transportation of measures was introduced, followed by the Monge problem of transporting between two measures in an optimal way. Since the Monge problem is difficult to solve, a better behaved relaxation, known as the Kantorovich problem, was introduced, which involves the notion of couplings as substitutes for transport maps.

Lecture 2. The Kantorovich problem is a linear programming problem where duality plays an important. This duality was proved, yielding characterizations of the primal and dual solutions, as well as Brenier's theorem on the original problem of Monge. The important concept of cyclical monotonicity was introduced and used in the proof.

Lecture 3. The Kantorovich problem leads to the notion of the Wasserstein distance over the space of probability measures. It was shown that the Wasserstein distance is indeed a metric, which is compatible with weak convergence. The formula for the linearization of the Wasserstein distance was also derived.

Lecture 4. The Wasserstein space, the space of probability measures endowed with the Wasserstein distance, can viewed as a formal infinite-dimensional Riemannian manifold. This perspective was explained by introducing the dynamic formulation of optimal transport, known as the Benamou–Brenier principle. Analogies to fluid dynamics were drawn. This led to the development of the Otto calculus over the Wasserstein space.

Lecture 5. The Otto calculus over the Wasserstein space allows for the development of a theory of gradient flows of functionals of probability measures. Formulas for gradients and gradient flows were derived, emphasizing the roles of important functionals such as entropy, as well as the notion of displacement convexity.

Lecture 6. Functional inequalities capture the convergence to equilibrium of gradient flows, and some important examples of these inequalities were introduced: the log-Sobolev inequality, Talagrand's inequality, and the HWI inequality. It was shown how the notion of displacement convexity can be used to prove functional inequalities, and how in turn, the property of log-concavity of a probability measure is closely connected with the displacement convexity of the relative entropy.

Lecture 7. Functional inequalities can also be used to establish the concentration of measure phenomenon. The following notions were presented: sub-Gaussian variables, the entropy method attributed to Herbst, the Bobkov–Götze theorem, tensorization, and transport inequalities.

Lecture 8. The perspective of viewing flows of probability measures as gradient flows in Wasserstein space ultimately leads to a powerful set of tools and insights from optimization which can be applied to the analysis of sampling algorithms. This idea was developed in the context of providing a non-asymptotic convergence analysis for the Euler–Maruyama discretization of the Langevin diffusion.

School of Mathematics, Institute for Advanced Study, Princeton, NJ, USA
E-mail address: schewi@ias.edu

Division of Applied Mathematics, Brown University, Providence, RI, USA
E-mail address: Yair_Shenfeld@Brown.edu

PARSITY, FEATURE SELECTION & THE SHAPLEY-FOLKMAN THEOREM.

ALEX D'ASPREMONT

Classification AMS 2020: 65K05

Keywords: Duality, Shapley-Folkman

Due to its linear complexity, naive Bayes classification remains an attractive supervised learning method, especially in very large-scale settings. We propose a sparse version of naive Bayes, which can be used for feature selection. This leads to a combinatorial maximum-likelihood problem, for which we provide an exact solution in the case of binary data, or a bound in the multinomial case. We prove that our convex relaxation bounds becomes tight as the marginal contribution of additional features decreases, using a priori duality gap bounds dervied from the Shapley-Folkman theorem. We show how to produce primal solutions satisfying these bounds. Both binary and multinomial sparse models are solvable in time almost linear in problem size, representing a very small extra relative cost compared to the classical naive Bayes. Numerical experiments on text data show that the naive Bayes feature selection method is as statistically effective as state-of-the-art feature selection methods such as recursive feature elimination, l_1-penalized logistic regression and LASSO, while being orders of magnitude faster [1]

CNRS - ENS PARIS.
Email address: aspremon@ens.fr

[1] A python implementation can be found at https://github.com/aspremon/NaiveFeatureSelection

Gradient flows for empirical Bayes in high-dimensional linear models

Zhou Fan*, Leying Guan†, Yandi Shen*, Yihong Wu*

Introduced by Robbins [Rob51, Rob56] in the 1950s, empirical Bayes (EB) is a powerful framework for large-scale inference that learns and adapts to latent structure in data. In this work, we consider empirical Bayes estimation in a linear model with Gaussian noise and i.i.d. prior,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \qquad \theta_j \overset{iid}{\sim} g_*, \qquad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2). \tag{0.1}$$

Assuming that the sample size n and dimension p are both large, we study estimation of the effect size prior g_* from the regression data $(\mathbf{X}, \mathbf{y})$. Many methods based upon variational inference or Monte Carlo EM have been proposed for this problem in the context of inferring genetic architectures of complex traits [ZQPC18, O'C21, ZZ21, SSAAP22, MCW+23], but the problem has only recently been the subject of theoretical study [MSS23].

We propose and study a new gradient-flow procedure for nonparametric estimation of g_* via the nonparametric maximum likelihood estimator (NPMLE). For a user-specified parameter $\tau^2 > 0$, reparametrizing the model by a smoothed regression vector $\boldsymbol{\varphi} = \boldsymbol{\theta} + \mathcal{N}(0, \tau^2 \,\mathrm{Id})$ and modified residual $\widetilde{\boldsymbol{\varepsilon}} \sim \mathcal{N}(0,\, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \sigma^2 \,\mathrm{Id} - \tau^2 \mathbf{X}\mathbf{X}^\top$, the negative log-likelihood $\bar{F}_n(g)$ admits a Gibbs variational representation

$$\begin{aligned}
\bar{F}_n(g) =& \min_q F_n(q, g), \\
F_n(q, g) :=& \frac{1}{p} \int \left[\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\varphi})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\varphi}) - \sum_{j=1}^p \log[\mathcal{N}_\tau * g](\varphi_j) + \log q(\boldsymbol{\varphi}) \right] q(\boldsymbol{\varphi}) \mathrm{d}\boldsymbol{\varphi} + \text{constant}.
\end{aligned}$$

We propose to jointly optimize $F_n(q, g)$ over posterior densities q on $\mathbb{R}^p$ and prior densities g on a bounded support $[-M, M]$, via the coupled gradient flows

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} q_t(\boldsymbol{\varphi}) &= -p \cdot \mathrm{grad}_q^{W_2} F_n(q_t, g_t)[\boldsymbol{\varphi}] \\
&= \nabla \cdot \left[q_t(\boldsymbol{\varphi}) \left(\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} (\mathbf{X}\boldsymbol{\varphi} - \mathbf{y}) - \left(\frac{[\mathcal{N}_\tau * g_t]'(\varphi_j)}{[\mathcal{N}_\tau * g_t](\varphi_j)} \right)_{j=1}^p \right) \right] + \Delta q_t(\boldsymbol{\varphi}) \qquad (0.2) \\
\frac{\mathrm{d}}{\mathrm{d}t} g_t(\theta) &= -\alpha \cdot \mathrm{grad}_g^{\mathrm{FR}} F_n(q_t, g_t)[\theta] = \alpha \, g_t(\theta) \left(\left[\mathcal{N}_\tau * \frac{\bar{q}_t}{\mathcal{N}_\tau * g_t} \right](\theta) - 1 \right) \qquad (0.3)
\end{aligned}$$

where $\mathrm{grad}_q^{W_2}$ denotes the Wasserstein-2 gradient in q, $\mathrm{grad}_g^{\mathrm{FR}}$ denotes the Fisher-Rao gradient in g, $\mathcal{N}_\tau * g_t$ is the convolution of the $\mathcal{N}(0, \tau^2)$ density with g_t, and $\bar{q}_t$ is the univariate averaged marginal

*Department of Statistics and Data Science, Yale University
†Department of Biostatistics, Yale University
zhou.fan@yale.edu, leying.guan@yale.edu, yandi.shen@yale.edu, yihong.wu@yale.edu

density of coordinates under q_t. The q-flow (0.2) is simulated via a Langevin diffusion [JKO98]

$$\mathrm{d}\boldsymbol{\varphi}_t = \left(-\mathbf{X}^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}\boldsymbol{\varphi}_t - \mathbf{y}) + \left(\frac{[\mathcal{N}_\tau * g_t]'(\varphi_{t,j})}{[\mathcal{N}_\tau * g_t](\varphi_{t,j})} \right)_{j=1}^{p} \right) \mathrm{d}t + \sqrt{2}\, \mathrm{d}\mathbf{B}_t \tag{0.4}$$

with time-evolving drift. This is coupled with the g-flow (0.3) which is simulated by a fixed-grid discretization of the domain $[-M, M]$ and an empirical estimate of $\bar{q}_t$ from the coordinates of the Langevin sample $\boldsymbol{\varphi}_t$. The reparametrization of the model by $\boldsymbol{\varphi}$ rather than $\boldsymbol{\theta}$ ensures that the Langevin diffusion targets a continuous and sufficiently regular posterior density, even if the prior estimate g_t approaches a discrete measure. This idea of bivariate optimization of a Gibbs variational representation for maximum likelihood inference can be attributed to [NH98], and has also been proposed and studied recently in [KLJ23, ACG+23] in different contexts.

Our results consist of (1) a statistical consistency guarantee for the NPMLE in this problem, extending recent work of [MSS23], (2) a new log-Sobolev inequality for mixing of Langevin dynamics in the linear model, deriving from an insight of [BB19], and (3) a local convergence guarantee for the above gradient flow algorithm.

For consistency, we show the following result.

Theorem 0.1. *Suppose g_* is supported on a known interval $[-M, M]$. As $n, p \to \infty$, under a deterministic condition for the design matrix $\mathbf{X}$, which in particular assumes the existence of test vectors $\mathbf{z}_j \in \mathbb{R}^n$ for sufficiently many columns $\mathbf{x}_j$ of $\mathbf{X}$ such that*

$$\mathbf{z}_j^\top \mathbf{x}_j \to 1, \qquad |\mathbf{z}_j^\top \mathbf{x}_k|^{2+\varepsilon} \to 0,$$

any approximate NPMLE $\widehat{g}$ over $[-M, M]$ satisfies $\widehat{g} \Rightarrow g_*$ almost surely in the sense of weak convergence.

Proposition 0.2. *Suppose $n/p \geq \gamma$ for any constant $\gamma > 0$, and $\mathbf{X} \in \mathbb{R}^{n \times p}$ has i.i.d. sub-Gaussian rows $\{\frac{1}{\sqrt{n}} x^{(i)}\}_{i=1}^n$ where $x^{(i)}$ has mean 0, covariance $\boldsymbol{\Sigma}_X \in \mathbb{R}^{p \times p}$ satisfying*

$$c \leq \lambda_{\min}(\boldsymbol{\Sigma}_X) \leq \lambda_{\max}(\boldsymbol{\Sigma}_X) \leq C,$$

and bounded sub-Gaussian norm. Then the assumptions for $\mathbf{X}$ needed in the above theorem hold almost surely as $n, p \to \infty$.

For mixing of the Langevin diffusion, let $\nu[g]$ be the posterior density of $\boldsymbol{\varphi}$ given $(\mathbf{X}, \mathbf{y})$ under the prior g for $\boldsymbol{\theta}$. We prove the following uniform log-Sobolev inequality, which implies exponential contraction in KL-divergence for the Langevin diffusion with $\nu[g]$ as its stationary law.

Theorem 0.3. *Let g be supported on $[-M, M]$ and fix a constant $\delta > 0$. Suppose that $\sigma^2/\|\mathbf{X}\|_{\mathrm{op}}^2 > M^2 + \delta$. Then for any $\tau^2 \in (0, \sigma^2/\|\mathbf{X}\|_{\mathrm{op}}^2 - \delta)$, $\nu[g]$ satisfies a LSI with constant $C = C(M, \tau, \delta) > 0$ that is uniform over all priors g on $[-M, M]$.*

Finally, for local convergence of the bivariate gradient flow, we show the following.

Theorem 0.4. *Suppose that $\nu[g]$ satisfies a LSI with some constant $C = C(M, \tau) > 0$, uniformly over priors g supported on $[-M, M]$. Let $\{(q_t, g_t)\}_{t \geq 0}$ be the solution to the bivariate gradient flow (0.2–0.3).*

If $\bar{F}_n$ is convex over the sub-level set $\{g : \bar{F}_n(g) \leq F_n(q_0, g_0)\}$ and additional (mild) technical conditions hold for the initialization (g_0, q_0), then for any $\varepsilon > 0$, some constant $C(\varepsilon) > 0$, and some time $t \leq C(\varepsilon)(n + p)$, it holds that $\bar{F}_n(g_t) - \bar{F}_n(g_*) \leq \varepsilon$.

We leave for future work several interesting open questions, including a better theoretical characterization of the optimization landscape of $\bar{F}_n(g)$ and conditions for global convergence, a theoretical analysis of time discretization and consistency of estimating $\bar{q}_t$ within the empirical implementation of the g-flow equation (0.3), and extensions to more complex models and sampling algorithms.

References

[ACG+23] Ö Deniz Akyildiz, Francesca Romana Crucinio, Mark Girolami, Tim Johnston, and Sotirios Sabanis, *Interacting particle Langevin algorithm for maximum marginal likelihood estimation*, arXiv preprint arXiv:2303.13429 (2023).

[BB19] Roland Bauerschmidt and Thierry Bodineau, *A very simple proof of the LSI for high temperature spin systems*, J. Funct. Anal. **276** (2019), no. 8, 2582–2588.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto, *The variational formulation of the Fokker-Planck equation*, SIAM J. Math. Anal. **29** (1998), no. 1, 1–17.

[KLJ23] Juan Kuntz, Jen Ning Lim, and Adam M Johansen, *Particle algorithms for maximum likelihood training of latent variable models*, International Conference on Artificial Intelligence and Statistics, PMLR, 2023, pp. 5134–5180.

[MCW+23] Fabio Morgante, Peter Carbonetto, Gao Wang, Yuxin Zou, Abhishek Sarkar, and Matthew Stephens, *A flexible empirical Bayes approach to multivariate multiple regression, and its improved accuracy in predicting multi-tissue gene expression from genotypes*, PLoS Genetics **19** (2023), no. 7, e1010539.

[MSS23] Sumit Mukherjee, Bodhisattva Sen, and Subhabrata Sen, *A mean field approach to empirical bayes estimation in high-dimensional linear regression*, arXiv preprint arXiv:2309.16843 (2023).

[NH98] Radford M Neal and Geoffrey E Hinton, *A view of the EM algorithm that justifies incremental, sparse, and other variants*, Learning in graphical models, Springer, 1998, pp. 355–368.

[O'C21] Luke J O'Connor, *The distribution of common-variant effect sizes*, Nature Genetics **53** (2021), no. 8, 1243–1249.

[Rob51] Herbert Robbins, *Asymptotically subminimax solutions of compound statistical decision problems*, Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, The Regents of the University of California, 1951.

[Rob56] ———, *An empirical Bayes approach to statistics*, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, The Regents of the University of California, 1956.

[SSAAP22] Jeffrey P Spence, Nasa Sinnott-Armstrong, Themistocles L Assimes, and Jonathan K Pritchard, *A flexible modeling and inference framework for estimating variant effect sizes from GWAS summary statistics*, BioRxiv (2022), 2022–04.

[ZQPC18] Yan Zhang, Guanghao Qi, Ju-Hyun Park, and Nilanjan Chatterjee, *Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits*, Nature Genetics **50** (2018), no. 9, 1318–1326.

[ZZ21] Geyu Zhou and Hongyu Zhao, *A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics*, PLoS genetics **17** (2021), no. 7, e1009697.

ON THE EMERGENCE OF CLUSTERS IN SELF-ATTENTION DYNAMICS

BORJAN GESHKOVSKI

Classification AMS 2020: 34D05, 34D06, 35Q83

Keywords: Transformers, self-attention, interacting particle systems, clustering, gradient flows.

1. INTRODUCTION

The introduction of Transformers in 2017 marked a milestone in the development of neural network architectures. Central to this is *self-attention*, a novel mechanism which distinguishes Transformers from traditional architectures, and which plays a substantial role in their superior practical performance. We present a mathematical framework for analyzing Transformers based on their interpretation as interacting particle systems, in which time plays the role of layers. Our analysis reveals that clusters emerge in long time, confirming previous empirical findings, and shedding light on the role of the attention mechanism.

2. MAIN RESULT

As first done in [3, 8], we define an idealized model of the Transformer architecture that consists in viewing the discrete layer indices as a continuous time variable, and which focuses exclusively on two key components of the Transformers architecture: *self-attention* and *layer normalization*[1]. This results in the dynamics

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\frac{1}{Z_{\beta,i}(t)} \sum_{j=1}^{n} e^{\beta \langle x_i(t), x_j(t) \rangle} x_j(t) \right) \tag{SA}$$

for $i \in [n]$ and $t \geq 0$, where

$$Z_{\beta,i}(t) = \sum_{k=1}^{n} e^{\beta \langle x_i(t), x_k(t) \rangle} \tag{2.1}$$

and $\mathbf{P}^{\perp}_x = I_d - xx^{\top}$ is the orthogonal projector to $\mathsf{T}_x \mathbb{S}^{d-1}$. We prove the following result (see [1, 2]).

Theorem 2.1. *Let $d, n \geq 2$ and $\beta \geq 0$, and suppose that either $d \geq n$, or $\beta \gtrsim_d n^2$, or $\beta \lesssim n^{-1}$. Consider the unique solution $(x_i(\cdot))_{i\in[n]} \in C^0(\mathbb{R}_{\geq 0}; (\mathbb{S}^{d-1})^n)$ to the Cauchy problem for* (SA)*, corresponding to an initial sequence of points $(x_i(0))_{i\in[n]} \in (\mathbb{S}^{d-1})^n$ distributed uniformly at random. Then almost surely there exists $x^* \in \mathbb{S}^{d-1}$ such that*

$$\lim_{t\to+\infty} x_i(t) = x^*$$

[1]Strictly speaking, we are using root-mean-square (RMS) norm. To the best of our knowledge, the original layer normalization, which consists in standardizing every component of every token at every layer has not yet been addressed in the continuous-time formulation of Transformers.

for all $i \in [n]$.

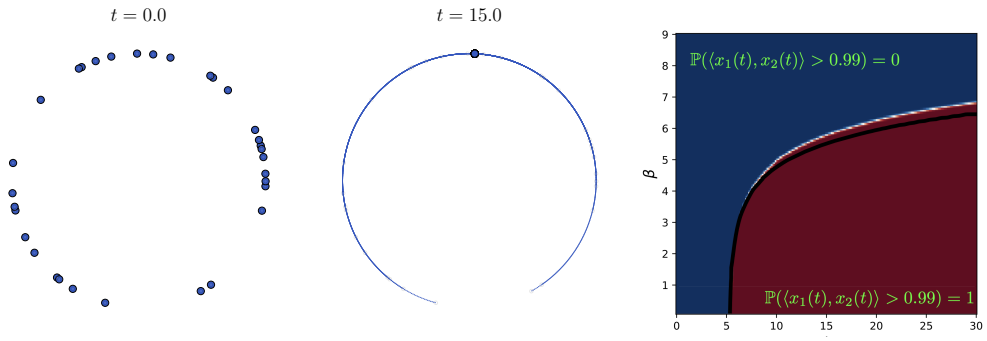

FIGURE 1. The evolution of trajectories can be fully described in very large dimension ($d \gg$ poly n), beyond solely long time asymptotics with rates, as seen in the phase diagram above. See [2].

Note that when $d = 2$ and $\beta = 0$, we recover the celebrated Kuramoto model of oscillators [7] (see [6] for mathematical details in the case $d \geq 2$).

Analyzing the self-attention model with a view toward understanding fine-grained and empirically observed phenomena in Transformers has borne fruit to several interesting results in the past year. Stability of the asymptotic results presented in [1] with respect to perturbations of the parameters has been addressed in [5]. A mean-field version of the masked attention mechanism, as well as a thorough study of the Lipschitz constant of attention is presented in [4]. Steering ensembles of empirical measures to ensembles of empirical measures via parametrized self-attention dynamics has been done in [9].

REFERENCES

[1] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, "The emergence of clusters in self-attention dynamics," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[2] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, "A mathematical perspective on Transformers," *arXiv preprint arXiv:2312.10794*, 2023.

[3] M. E. Sander, P. Ablin, M. Blondel, and G. Peyré, "Sinkformers: Transformers with doubly stochastic attention," in *International Conference on Artificial Intelligence and Statistics*, 2022, pp. 3515–3530, PMLR.

[4] V. Castin, P. Ablin, and G. Peyré, "Understanding the Regularity of Self-Attention with Optimal Transport," *arXiv preprint arXiv:2312.14820*, 2023.

[5] H. Koubbi, M. Boussard, and L. Hernandez, "The Impact of LoRA on the Emergence of Clusters in Transformers," *arXiv preprint arXiv:2402.15415*, 2024.

[6] A. Frouvelle and J.-G. Liu, "Long-time dynamics for a simple aggregation equation on the sphere," in *Stochastic Dynamics Out of Equilibrium: Institut Henri Poincaré, Paris, France, 2017*, Springer, 2019, pp. 457–479.

[7] Y. Kuramoto, "Self-entrainment of a population of coupled non-linear oscillators," in *International Symposium on Mathematical Problems in Theoretical Physics: January 23–29, 1975, Kyoto University, Kyoto/Japan*, Springer, 1975, pp. 420–422.

[8] Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*, 2019.

[9] Agrachev, Andrei, and Cyril Letrouit. *Generic controllability of equivariant systems and applications to particle systems and neural networks*. arXiv preprint arXiv:2404.08289 (2024).

Inria & Laboratoire Jacques-Louis Lions, Sorbonne Université, 4 Place Jussieu, 75005 Paris
Email address: borjan@mit.edu

THE LAPLACE APPROXIMATION IN HIGH-DIMENSIONAL BAYESIAN INFERENCE

ANYA KATSEVICH

Classification AMS 2020: 62E17, 62F15

Keywords: Laplace approximation, high dimensional Bayesian inference

Developing cheap and accurate computational techniques for Bayesian inference is an important goal, as Bayesian inference tasks can be very computationally intensive. These tasks include constructing posterior credible sets, computing the posterior mean and covariance, and computing the posterior predictive distribution of a new data point. Computing all of these quantities involves either sampling from the posterior π, or taking integrals $\int g d\pi$ against the posterior. When the dimensionality of the parameter of interest is large, these tasks can be very expensive.

The most common approach in Bayesian inference to both generate samples from, and integrate against π is Markov Chain Monte Carlo (MCMC) [3]. In principle, MCMC schemes can be made arbitrarily accurate by tuning parameters, e.g. by extending the simulation time of the Markov chain or by decreasing step sizes. However, MCMC is computationally intensive in high dimensions and it also has other disadvantages. For example, it can be difficult to identify clear-cut stopping criteria for the algorithm [5].

Another popular approach is to find a simple distribution $\hat{\gamma}$ which approximates π, and to use this distribution as a proxy for π to do all of one's inference tasks. In the ideal scenario, many integrals against $\hat{\gamma}$ are computable in closed form, and it is cheap to sample from $\hat{\gamma}$. The idea of using an approximation $\hat{\gamma}$ to π is at the heart of approximate Bayesian inference methods such as variational inference [2, 16], expectation propagation [12], and the Laplace approximation, a Gaussian approximation first introduced in the Bayesian inference context by [14].

The Laplace approximation (LA) exploits large sample properties of the posterior which have been established in the *Bernstein von-Mises* theorem (BvM). The BvM is a fundamental result which in its classical form states that if the model is well-specified, then the posterior contracts around the ground truth parameter and becomes asymptotically normal in the large sample limit [15, Section 10.2]. Despite the theoretical and philosophical importance of the BvM (it has been used as a justification of Bayesian procedures from the perspective of frequentist inference), this result does not give an implementable Gaussian approximation to the posterior. This is where the LA comes in.

To explain the LA construction, consider a posterior π whose mass concentrates in a small neighborhood of the mode, which we call $\hat{x}$. When the conditions of the BvM are satisfied, this highest mode should be unique; otherwise, concentration cannot occur. Since most of the mass of π is near $\hat{x}$, we should incur only a small error by replacing the log posterior with its second order Taylor expansion about $\hat{x}$. This gives rise to the LA,

Date: May 17, 2024.

the Gaussian density $\hat{\gamma}$ which is given by

$$\hat{\gamma} = \mathcal{N}\left(\hat{x},\ \nabla^2 V(\hat{x})^{-1}\right), \quad \hat{x} = \arg\min_{x\in\Theta} V(x) \tag{0.1}$$

where $\pi \propto e^{-V}$ is a density on $\Theta \subseteq \mathbb{R}^d$. The LA has proved to be an invaluable tool for Bayesian inference in applications ranging from deep learning [6] to inverse problems [4] to variable selection in high-dimensional regression [1].

Unlike methods such as MCMC, which can be made arbitrarily accurate, there are no parameters to tune in the LA — it incurs some fixed approximation error. Quantifying the LA's error as a function of dimension d, sample size n, and model parameters, is a worthy task given its widespread use. It is also a challenging theoretical endeavor when dimension d is large, and currently a very active research area. Major contributions have been made e.g. by [9, 13, 8]. We also study the LA error in this work.

But arguably, it is even more important to go *beyond* the LA to develop new, more accurate approximations which better capture the complexity of the posterior π. For example, a known downside of the LA $\hat{\gamma}$ is that it is symmetric about the mode and therefore cannot capture skewness of π. Instead of constructing an entirely new kind of approximation, a natural idea is to correct the LA in some way to get a higher-order accuracy approximation. Non-rigorous skew normal approximations have been developed in [17], but we are aware of only a single work [7] that rigorously derives a higher-order accurate LA. However, the corrected LA obtained by [7] is only shown to be accurate in constant dimension d. In modern applications involving very high-dimensional parameters, d cannot be considered constant relative to sample size n. So far, no prior work has obtained a higher-accuracy LA which is rigorously justified in high dimensions.

In this work, we develop a powerful technique to analyze the Laplace approximation more precisely than was possible before. This technique leads us to derive the *first ever correction to the LA which provably improves its accuracy by an order of magnitude, in high dimensions*. Our approach allows us to prove error bounds on this corrected LA in terms of a variety of error metrics discussed below. It also enables us to prove both tighter upper bounds and the first ever lower bounds on the standard LA in high dimensions. In particular, we prove that $d^2 \ll n$ is in general necessary for accuracy of the LA. Finally, we apply our theory in two example models: a Dirichlet posterior arising from a multinomial observation, and logistic regression with Gaussian design. In the latter setting, we prove high probability bounds on the accuracy of the LA and skew-corrected LA in powers of $d/\sqrt{n}$ alone.

REFERENCES

[1] Rina Foygel Barber, Mathias Drton, and Kean Ming Tan. Laplace approximation in high-dimensional Bayesian regression. In *Statistical Analysis for High-Dimensional Data: The Abel Symposium 2014*, pages 15–36. Springer, 2016.

[2] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[3] William M Bolstad. *Understanding computational Bayesian statistics*, volume 644. John Wiley & Sons, 2009.

[4] Tan Bui-Thanh, Carsten Burstedde, Omar Ghattas, James Martin, Georg Stadler, and Lucas C. Wilcox. Extreme-scale uq for Bayesian inverse problems governed by PDEs. In *SC '12: Proceedings of the*

International Conference on High Performance Computing, Networking, Storage and Analysis, pages 1–11, 2012.

[5] Mary Kathryn Cowles and Bradley P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

[6] Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless Bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103, 2021.

[7] Daniele Durante, Francesco Pozza, and Botond Szabo. Skewed Bernstein-von Mises theorem and skew-modal approximations. *arXiv preprint arXiv:2301.03038*, 2023.

[8] Tapio Helin and Remo Kretschmann. Non-asymptotic error estimates for the Laplace approximation in Bayesian inverse problems. *Numerische Mathematik*, 150(2):521–549, 2022.

[9] Mikolaj J Kasprzak, Ryan Giordano, and Tamara Broderick. How good is your Gaussian approximation of the posterior? Finite-sample computable error bounds for a variety of useful divergences. *arXiv preprint arXiv:2209.14992*, 2022.

[10] Bas JK Kleijn and Aad W van der Vaart. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

[11] Jeffrey W Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(168):1–53, 2021.

[12] Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[13] Vladimir Spokoiny. Dimension free nonasymptotic bounds on the accuracy of high-dimensional laplace approximation. *SIAM/ASA Journal on Uncertainty Quantification*, 11(3):1044–1068, 2023.

[14] Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

[15] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[16] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[17] Jackson Zhou, Clara Grazian, and John T Ormerod. Tractable skew-normal approximations via matching. *Journal of Statistical Computation and Simulation*, 94(5):1016–1034, 2024.

77 MASSACHUSETTS AVENUE, BUILDING 2, ROOM 350B, CAMBRIDGE MA 02139
Email address: akasevi@mit.edu

INFORMATION-THEORETIC THRESHOLDS FOR PLANTED DENSE CYCLES

CHENG MAO, ALEXANDER S. WEIN, AND SHENDUO ZHANG

Classification AMS 2020: 62B10, 05C82

Keywords: Planted dense cycle, small-world network, information-theoretic thresholds

The *Watts–Strogatz small-world model* has been an influential random graph model since its proposal in 1998 due to the ubiquity of the small-world phenomenon in complex networks [12, 11]. In this model, there are n vertices with latent positions on a circle, and the vertices are more likely to be connected to their k-nearest geometric neighbors than to more distant vertices. In other words, a denser cycle of length n and width k is "planted" in the sparser ambient random graph on n vertices.

While there has been extensive literature on small-world networks and geometric graphs in general, the associated statistical problems, such as detection and recovery of the latent geometry from the observed random graph, have only gained attention more recently. The information-theoretic thresholds and efficient algorithms for the small-world model are studied in [2], but there remain several gaps between upper and lower bounds that are unknown to be inherent or not. Much sharper characterizations of the recovery thresholds are given in [1, 3] but only for a small bandwidth parameter $k = n^{o(1)}$. From the perspective of graphon estimation, there have also been algorithms and statistical analyses introduced for related models recently [6, 10, 5]. Moreover, the study of random geometric graphs has received broad interests in recent years; see the survey [4] and the works [7, 8] for high-dimensional random geometric graphs with edge noise similar to what we consider.

Since the model of interest consists of a hidden dense cycle planted in a sparser random graph, the recent work [9] refers to it as the *planted dense cycle* problem, following the etymology of planted clique and planted dense subgraph problems. The work [9] studies the problem in the framework of *low-degree polynomial algorithms*, a framework that has proved to be successful at predicting *computational* thresholds. However, the more fundamental *information-theoretic* (or *statistical*) thresholds—where no constraints are placed on computation time—remain open, and we aim to address them in this work. More specifically, suppose that in an ambient random graph on n vertices with edge density q, there is a hidden cycle with expected width $k = n\tau$ and edge density p. We find the information-theoretic thresholds for detecting the presence of the cycle and for recovering the location of the cycle, in terms of the parameters n, τ, p, q. In particular, the information-theoretic thresholds for detection and recovery both differ from the computational thresholds given in [9], justifying the existence of *statistical-to-computational gaps* for this problem.

Problem setup. The planted dense cycle model can be described as follows. For any $a, b \in [0, 1]$, define $\mathfrak{d}(a, b) := \min\{|a-b|, 1-|a-b|\}$. In other words, $\mathfrak{d}(a, b)$ is the distance between a and b on a circle of circumference 1. Throughout the paper, we consider the

setting where the number of vertices n grows, and other parameters $p, q, r, \tau \in [0,1]$ may depend on n. In the models to be defined, p and q will be the average edge densities on and off the planted cycle respectively, r is the average edge density of the entire graph, and $n\tau$ is the bandwidth of the cycle, satisfying

$$n \to \infty, \qquad 0 < q < r < p \le 1, \qquad 0 < \tau < 1/2, \qquad r = \tau p + (1-\tau)q. \tag{1}$$

Definition 1 (Model $\mathcal{P}$, planted dense cycle)**.** *Let $z \in [0,1]^n$ be a latent random vector whose entries $z_1, \dots, z_n$ are i.i.d. $\mathrm{Unif}([0,1])$ variables. Let $X_{ij} := \mathbb{1}_{\mathfrak{d}(z_i, z_j) \le \tau/2}$ for all $(i,j) \in \binom{[n]}{2}$. For concreteness, we let $X_{ii} = 0$ and $X_{ji} = X_{ij}$ for $i \neq j$, so that $X \in \mathbb{R}^{n \times n}$ is the adjacency matrix of the underlying cycle. We observe an undirected graph with adjacency matrix $A \in \mathbb{R}^{n \times n}$ whose edges, conditional on $z_1, \dots, z_n$, are independently sampled as follows: $A_{ij} \sim \mathrm{Bern}(p)$ if $X_{ij} = 1$ and $A_{ij} \sim \mathrm{Bern}(q)$ if $X_{ij} = 0$ for $(i,j) \in \binom{[n]}{2}$. We write $A \sim \mathcal{P}_A$ and $(A, X) \sim \mathcal{P}$ (or, equivalently, $(A, z) \sim \mathcal{P}$).*

Definition 2 (Model $\mathcal{Q}$, Erdős–Rényi graph)**.** *We observe a $G(n, r)$ Erdős–Rényi graph with adjacency matrix $A \in \mathbb{R}^{n \times n}$. We write $A \sim \mathcal{Q}$.*

We now formulate the detection and recovery problems of interest.

Definition 3 (Detection)**.** *Let $\mathcal{P}$ and $\mathcal{Q}$ be the models from Definitions 1 and 2 respectively, with parameters in* (1)*. Observing the adjacency matrix $A \in \mathbb{R}^{n \times n}$ of a graph, we test $H_1 : A \sim \mathcal{P}_A$ against $H_0 : A \sim \mathcal{Q}$. We say that a test Φ, a $\{0,1\}$-valued measurable function of the observation A, achieves*

- *strong detection, if $\lim_{n\to\infty}[\mathcal{P}\{\Phi(A) = 0\} + \mathcal{Q}\{\Phi(A) = 1\}] = 0$;*
- *weak detection, if $\limsup_{n\to\infty}[\mathcal{P}\{\Phi(A) = 0\} + \mathcal{Q}\{\Phi(A) = 1\}] < 1$.*

Definition 4 (Recovery)**.** *Let $\mathcal{P}$ be the model from Definition 1 with parameters in* (1)*. For $(A, X) \sim \mathcal{P}$, observing A, we aim to estimate X with an estimator $\hat{X} \in \mathbb{R}^{n \times n}$ that is measurable with respect to A. Consider the mean squared error $R(\hat{X}, X) := \sum_{1 \le i < j \le n} \mathbb{E}[(\hat{X}_{ij} - X_{ij})^2]$, where the expectation is with respect to $(A, X) \sim \mathcal{P}$. We say that an estimator $\hat{X}$ achieves*

- *strong recovery, if $\lim_{n\to\infty} \frac{R(\hat{X}, X)}{\binom{n}{2}\tau(1-\tau)} = 0$;*
- *weak recovery, if $\limsup_{n\to\infty} \frac{R(\hat{X}, X)}{\binom{n}{2}\tau(1-\tau)} < 1$.*

Note that each X_{ij} is marginally a $\mathrm{Bern}(\tau)$ random variable, so estimating X_{ij} by its mean τ for all $(i,j) \in \binom{[n]}{2}$ yields a trivial mean squared error $\binom{n}{2}\tau(1-\tau)$, which justifies the above definition.

Main results. Our main results are summarized in the following theorem.

Theorem 5 (Information-theoretic thresholds)**.** *Consider the detection and recovery problems in Definitions 3 and 4 respectively, with parameters n, τ, p, q, r in* (1)*. Furthermore, suppose that $(\log n)^3 \le n\tau \le \frac{n}{(\log n)^2}$ and $\frac{\log n}{n} \le r \le \frac{1}{2}$. Define $\lambda := \frac{(p-q)^2}{r(1-r)}$ which can be seen as the signal-to-noise ratio of the problem. Then we have:*

- *If $n\tau\lambda \to 0$ as $n \to \infty$, then no test achieves weak detection, and no estimator achieves weak recovery.*
- *If $\frac{n\tau\lambda}{\log n} \to \infty$ and $\frac{n\tau(p-r)}{\log n} \to \infty$ as $n \to \infty$, then there is a test that achieves strong recovery, and there is an estimator that achieves strong recovery.*

Our conditions for positive and negative results match up to a logarithmic factor in most cases. The threshold for both detection and recovery is $n\tau\lambda = \tilde{\Theta}(1)$, where $n\tau$ is the bandwidth of the planted dense cycle, and λ is the edge-wise signal-to-noise ratio.

Remark 6 (Statistical-to-computational gap)**.** *Our information-theoretic results for the detection and recovery of a planted dense cycle complement the work* [9] *which studies low-degree polynomial algorithms for the same model. Suppose that the parameters in* (1) *satisfy $Cq \le p \le C'q$ for some constants $C' > C > 1$. As shown in* [9]*, the detection threshold for the class of low-degree polynomial algorithms is $n^3p^3\tau^4 = n^{o(1)}$, and the recovery threshold for the class of low-degree polynomial algorithms is $np\tau^2 = n^{o(1)}$. In this regime, the signal-to-noise ratio $\lambda = \frac{(p-q)^2}{r(1-r)}$ has the same order as p, so the information-theoretic threshold from Theorem 5 can be expressed as $np\tau = \tilde{\Theta}(1)$. Therefore, Theorem 5 and* [9] *together suggest that there are statistical-to-computational gaps for both detection and recovery of a planted dense cycle.*

REFERENCES

[1] Vivek Bagaria, Jian Ding, David Tse, Yihong Wu, and Jiaming Xu. Hidden hamiltonian cycle recovery via linear programming. *Operations research*, 68(1):53–70, 2020.

[2] Tony Cai, Tengyuan Liang, and Alexander Rakhlin. On detection and structural reconstruction of small-world random networks. *IEEE Transactions on Network Science and Engineering*, 4(3):165–176, 2017.

[3] Jian Ding, Yihong Wu, Jiaming Xu, and Dana Yang. Consistent recovery threshold of hidden nearest neighbor graphs. In *Conference on Learning Theory*, pages 1540–1553. PMLR, 2020.

[4] Quentin Duchemin and Yohann De Castro. Random geometric graph: Some recent developments and perspectives. *arXiv preprint arXiv:2203.15351*, 2022.

[5] Christophe Giraud, Yann Issartel, and Nicolas Verzelen. Localization in 1d non-parametric latent space models from pairwise affinities. *Electronic Journal of Statistics*, 17(1):1587–1662, 2023.

[6] Jeannette Janssen and Aaron Smith. Reconstruction of line-embeddings of graphons. *Electronic Journal of Statistics*, 16(1):331–407, 2022.

[7] Suqi Liu and Miklós Z Rácz. Phase transition in noisy high-dimensional random geometric graphs. *Electronic Journal of Statistics*, 17(2):3512–3574, 2023.

[8] Suqi Liu and Miklós Z Rácz. A probabilistic view of latent space graphs and phase transitions. *Bernoulli*, 29(3):2417–2441, 2023.

[9] Cheng Mao, Alexander S. Wein, and Shenduo Zhang. Detection-recovery gap for planted dense cycles. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2440–2481, 12–15 Jul 2023.

[10] Amine Natik and Aaron Smith. Consistency of spectral seriation. *arXiv preprint arXiv:2112.04408*, 2021.

[11] Duncan J Watts. *Small worlds: the dynamics of networks between order and randomness*, volume 36. Princeton university press, 2004.

[12] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

SCHOOL OF MATHEMATICS, GEORGIA INSTITUTE OF TECHNOLOGY, ATLANTA, GA, USA
Email address: cheng.mao@math.gatech.edu

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, DAVIS, CA, USA
Email address: aswein@ucdavis.edu

SCHOOL OF MATHEMATICS, GEORGIA INSTITUTE OF TECHNOLOGY, ATLANTA, GA, USA
Email address: szhang705@gatech.edu

GEOMETRY AND DYNAMICS OF THE DEEP LINEAR NETWORK

GOVIND MENON

Classification AMS 2020: 68Q32, 68T05, 93E35

Keywords: Deep learning, gradient flows, entropy of overparametrization.

1. GRADIENT DYNAMICS IN DEEP LEARNING

These lectures study the training dynamics of the deep linear network (DLN) using the geometric theory of dynamical systems. The DLN is deep learning restricted to *linear* functions. However, it retains two essential features of deep learning: *overparametrization* and *degenerate loss functions.* Overparametrization provides a foliation of phase space by invariant manifolds. Of these, there is a fundamental invariant manifold, the *balanced manifold*, which is itself foliated by group orbits. This geometric structure allows us to define a natural Boltzmann entropy (the logarithm of the volume of a group orbit) that may be computed explicitly. Our approach unifies the work of several authors into a thermodynamic framework.

2. THE MODEL

We fix two positive integer d and N referred to as the width and depth of the network. The state space for the DLN is $\mathbb{M}_d^N$, where $\mathbb{M}_d$ denotes the space of $d \times d$ real matrices. Each point $\mathbf{W} \in \mathbb{M}_d^N$ is denoted by $\mathbf{W} = (W_N, W_{N-1}, \dots, W_1)$. We equip $\mathbb{M}_d$ with the Frobenius norm so that $\mathbb{M}_d^N$ is Euclidean with the norm $\|\mathbf{W}\|^2 = \sum_{p=1}^N \operatorname{Tr}\left(W_p^T W_p\right)$. We define the projection $\phi : \mathbb{M}_d^N \to \mathbb{M}_d$ and *end-to-end* matrix X by

$$\phi(\mathbf{W}) = W_N W_{N-1} \cdots W_1 =: W. \tag{2.1}$$

We assume given an energy $E : \mathbb{M}_d \to \mathbb{R}$. The training dynamics are described by the gradient flow in $\mathbb{M}_d^N$ of the 'lifted' loss function $L = E \circ \phi$

$$\dot{\mathbf{W}} = -\nabla_{\mathbf{W}} L(\mathbf{W}). \tag{2.2}$$

A computation then reduces equation (2.2) to a collection of N equations in $\mathbb{M}_d$

$$\dot{W}_p = -(W_N \cdots W_{p+1})^T E'(X)(W_{p-1} \cdots W_1)^T, \quad p = 1, \dots, N. \tag{2.3}$$

Here $E'(X)$ denotes the $d \times d$ matrix with entries

$$E'(X)_{ij} = \frac{\partial E}{\partial X_{jk}}, \quad 1 \le j, k \le d. \tag{2.4}$$

Date: May 20, 2024.

3. Overparametrization and degenerate loss functions

Let $f_p : \mathbb{R}^d \to \mathbb{R}^d$ define the linear function $f_p(x) = W_p x$ for $1 \leq p \leq N$. Then the linear function $f : \mathbb{R}^d \to \mathbb{R}^d$ corresponding to the matrix $f(x) = Xx$ is $f = f_N \circ f_{N-1} \ldots \circ f_1$. The output function f depends only on the end-to-end matrix X. However, the same function f may be represented by Nd^2 choices of training parameters $\mathbf{W}$. Thus, despite the absence of the nonlinear activation element and shifts, the choice of variables in the DLN models *overparametrization* in deep learning.

Natural learning tasks, such as matrix completion, give rise to *degenerate loss functions*. Assume given a subset $S \subset \{(i,j)\}_{1 \leq i,j \leq d}$ and assume given the values of X_{ij}, for $(i,j) \in S$, say $X_{ij} = a_{ij}$. The task in matrix completion task is to obtain a principled answer to the question: how do we reconstruct X from the partial observations X_{ij} for $(i,j) \in S$?

The DLN is used to study this question as follows. Introduce the quadratic loss function

$$E(X) = \frac{1}{2} \sum_{(i,j) \in S} |X_{ij} - a_{ij}|^2, \tag{3.1}$$

and seek the limit of $X(t) = \phi(\mathbf{W}(t))$ as $t \to \infty$ when $\mathbf{W}(t)$ solves the gradient flow (2.3). The loss function $E(X)$ is degenerate because it does not depend on X_{ij} when $(i,j) \notin S$. Thus, E is minimized on the affine subspace

$$\mathcal{S} = \{X \in \mathbb{M}_d : X_{ij} = a_{ij}, \quad (i,j) \in S\}.$$

In particular, E does not have compact sublevel sets, and we cannot apply La Salle's invariance principle. However, a new mathematical structure emerges.

4. Geometric features of training dynamics

(1) *Invariant varieties.* Equation (2.3) shows that each $\dot{W}_p$ is obtained by a linear transformation of $E'(X)$. Thus, the space of gradients of $L = E \circ \phi$ at $\mathbf{W}$ has only d^2 dimensions, whereas the dimension of the tangent space $T_{\mathbf{W}}\mathbb{M}_d^N$ is Nd^2. A simple calculation then yields $(N-1)d(d+1)/2$ conserved quantities [1, 2, 3]

$$G_p = W_{p+1}^T W_{p+1} - W_p W_p^T, \quad 1 \leq p \leq N-1. \tag{4.1}$$

The solution set to these equations is a conic section in $\mathbb{M}_d^N$ parametrized by $N-1$ symmetric matrices. We call these the $\mathbf{G}$-balanced varieties, denoted $\mathcal{M}_{\mathbf{G}}$, where $\mathbf{G} = (G_{N-1}, \ldots, G_1)$. Each of these varieties is invariant under the flow (2.2).

(2) *Riemannian submersion of $\mathcal{M}$* When $\mathbf{G} = 0$ and X has full rank, we obtain a fundamental invariant manifold termed the *balanced manifold*, $\mathcal{M}$. The dynamics of $X(t) \in \mathbb{M}_d$ are slaved to the dynamics of $\mathbf{W}(t) \in \mathcal{M}_{\mathbf{G}}$. However, when $\mathbf{W}(t)$ lies on $\mathcal{M}$, then $X(t)$ satisfies the Riemannian gradient flow [4]

$$\dot{X} = -\mathrm{grad}_{g^N} E(X), \tag{4.2}$$

where the metric g^N is obtained by Riemannian submersion of the metric on $\mathbb{M}_d^N$ restricted to $\mathcal{M}$.

(3) *Group orbits and an entropy formula* Given $X \in \mathbb{M}_d$, we may lift it to an $O(d)^{N-1}$ orbit $\mathcal{O}_X \in \mathbb{M}_d^N$, such that $\phi(\mathbf{W}) = X$ for each $\mathbf{W} \in \mathcal{O}_X$. We may then quantify overparametrization with a Boltzmann entropy of the form $S(X) = \log \mathrm{vol}(O_X)$ [7], improving the main result in [5].

(4) *From gradient descent to the Riemannian Langevin equation (RLE)*. We augment the gradient dynamics with natural stochastic dynamics, using the geometry of overparametrization to add noise in the 'null directions'. This extends the idea of optimization by the DLN to that of Gibbs sampling, providing a thermodynamic framework in which the energy $E(X)$ is replaced by a Helmholtz free energy $F(X) = E(X) - \beta^{-1}S(X)$ at inverse temperature $\beta > 0$ [6, 7].

The volume formula is as follows. Given $X \in \mathbb{M}_d$ with full rank, let its SVD be $X = Q_N \Sigma Q_0^T$, where $Q_N, Q_0 \in \mathrm{O}(d), \Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_d)$. Then

$$\mathrm{vol}(\mathcal{O}_X) = c_d^{N-1}\sqrt{\frac{\mathrm{van}(\Sigma^2)}{\mathrm{van}(\Sigma^{\frac{2}{N}})}} = c_d^{N-1}\prod_{1\le i<j\le d}\sqrt{\frac{\sigma_i^2-\sigma_j^2}{\sigma_i^{\frac{2}{N}}-\sigma_j^{\frac{2}{N}}}} \tag{4.3}$$

where $c_d = \mathrm{vol}(\mathrm{O}(d)) = 2^{\frac{1}{2}d(d+3)}\prod_{r=1}^{d}\frac{\pi^{\frac{r}{2}}}{\Gamma(\frac{r}{2})}$ is the volume of $\mathrm{O}(d)\mathrm{van}(\Lambda)$ denotes the Vandermonde determinant associated to a diagonal matrix Λ.

The formulation of the RLE, especially the definition of the noise in the null directions, requires greater detail and is contained in the forthcoming work [6, 7].

REFERENCES

[1] Arora, Sanjeev and Cohen, Nadav and Hazan, Elad. On the optimization of deep networks: Implicit acceleration by overparameterization. *International conference on machine learning*, 244-253, 2018.

[2] Arora, Sanjeev and Cohen, Nadav and Hu, Wei and Luo, Yuping. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[3] Arora, Sanjeev and Cohen, Nadav and Hu, Wei and Luo, Yuping. A convergence analysis of gradient descent for deep linear neural networks. *arXiv:1810.02281*, 2018.

[4] Bah, Bubacarr and Rauhut, Holger and Terstiege, Ulrich and Westdickenberg, Michael. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11, 307–353, 2022.

[5] Cohen, Nadav and Menon, Govind and Veraszto, Zsolt. Deep Linear Networks for Matrix Completion—an Infinite Depth Limit. *SIAM Journal on Applied Dynamical Systems*, 22, 3208–3232, 2023.

[6] Menon, Govind. The geometry of the deep linear network. *Preprint*, 2024.

[7] Menon, Govind and Yu, Tianmin. An entropy formula for the deep linear network. *Preprint*, 2024.

BROWN UNIVERSITY, 182 GEORGE ST., PROVIDENCE, RI 02912, USA

Email address: govind_menon@brown.edu

Deep Learning based algorithm for nonlinear PDEs in finance & gradient descent type algorithm for non-convex stochastic optimization with ReLU neural networks

Ariel Neufeld

Nanyang Technological University

Joint work with

Christian Beck, Sebastian Becker, Patrick Cheridito, Arnulf Jentzen

and

Dong-Young Lim, Sotirios Sabanis, Ying Zhang

- typically, financial derivatives depend on several underlying assets
- Prices of derivatives are expressed by partial differential equations (PDE)
  $\Longrightarrow$ high-dimensional PDE
- In idealized markets without friction
  $\Longrightarrow$ PDE is linear
  $\Longrightarrow$ Efficient Algorithms to solve PDE

Want: More sophisticated models

$\Longrightarrow$ PDE becomes nonlinear
$\Longrightarrow$ Complexity of Algorithms grows exponentially in dimension
$\Longrightarrow$ Algorithms are non-efficient in high-dimensions!

Idea: Train neural network (NN) to approx. solve nonlinear PDE

Mathematical Definition of m-layered neural network $\mathcal{N}^{(m,\theta)}$

$$\mathcal{N}_\sigma^{(m,\theta)}(x) := f_m^{A^{(m)},b^{(m)}} \circ \sigma^{(m)} \circ f_{m-1}^{A^{(m-1)},b^{(m-1)}} \circ \cdots \circ \sigma^{(1)} \circ f_0^{A^{(0)},b^{(0)}}(x)$$

where for u=0,...,m

- **affine function:** $f_u^{A^{(u)},b^{(u)}}(v) := A^{(u)}v + b^{(u)}$, where
- **weights:** $A^{(u)} \in \mathbb{R}^{\ell_{u+1}\times\ell_u}$,
- **bias:** $b^{(u)} \in \mathbb{R}^{\ell_{u+1}}$
- **activation function** $(v_1,\dots v_{\ell_u}) \mapsto \sigma^{(u)}(v) := (\sigma(v_1),\dots\sigma(v_{\ell_u})) \in \mathbb{R}^{\ell_u}$

  typically:

  -$\sigma(x) := \max\{x,0\}$ "rectified linear unit (ReLU)"

  -$\sigma(x) := \frac{1}{1+e^{-x}}$ "Sigmoid function"

$\Longrightarrow$ Need to specify $\sum_{u=0}^{m} \ell_u\ell_{u+1} + \ell_{u+1}$ many parameters

Universal Approximation Theorem (for sigmoid) by Cybenko

Let X be compact, then set of 1-layered neural networks is dense in $C(X)$.

Other approximation results: Hornik, Maiorov, Pinkus, Mhaskar, Yarotsky...

Idea: Use Neural networks to approx. solve high-dim. nonlinear PDEs

Questions:

- How can we use universal approximation property of neural networks?
- How can we "learn" approximatively the solution of PDE?

Training of neural network involves solving optimization problems:

- How to formulate solution of PDE as solution of an optimization problem?

Quadratic Minimization

For any $x \in [a,b]^d$ let X_x be suff. int. random variable, $x \mapsto E[X_x]$ contin. Then

$$\int_{[a,b]^d} \mathbb{E}\big[|X_x - \mathbb{E}[X_x]|^2\big]\, dx = \min_{u \in C([a,b]^d,\mathbb{R})} \Big(\int_{[a,b]^d} \mathbb{E}\big[|X_x - u(x)|^2\big]\, dx \Big)$$

Moral: Expectation is solution of quadratic minimization problem!

Note: Similar results for conditional expectation

$$\Longrightarrow \text{Write solution of PDE as Expectation}$$

Feynman-Kac Representation

For any $t \in [0,T]$, $x \in \mathbb{R}^d$ consider

$$X_s^{t,x} = x + \int_t^s \mu(X_u^x)\, du + \int_t^s \sigma(X_u^x)\, dW_u$$

Then, the solution of the PDE

$$-\tfrac{\partial}{\partial t} v(t,x) - \langle \mu(x), (\nabla_x v)(t,x) \rangle - \tfrac{1}{2} \mathrm{Trace}\big(\sigma(x)\sigma(x)^T (\mathrm{Hess}_x v)(t,x)\big) = 0$$
$$v(T,x) = \varphi(x)$$

satisfies the following representation

$$v(t,x) = \mathbb{E}[\varphi(X_T^{t,x})]$$

Moral: Can indeed write solution of PDE as Expectation!

Solution of linear PDE is solution of minimization problem

For each t, x the solution $v(t,x)$ of the PDE

$$-\tfrac{\partial}{\partial t}v(t,x) - \langle \mu(x), (\nabla_x v)(t,x)\rangle - \tfrac{1}{2}\mathrm{Trace}\big(\sigma(x)\sigma(x)^T(\mathrm{Hess}_x v)(t,x)\big) = 0$$
$$v(T,x) = \varphi(x)$$

satisfies

$$\mathbb{E}\big[|\varphi(X_T^{t,x}) - v(t,x)|^2\big] = \inf_{u \in \mathrm{cont.}} \mathbb{E}\big[|\varphi(X_T^{t,x}) - u(x)|^2\big]$$

Idea: Use neural network to approx. solve minimization problem (RHS)

How to get from linear to semilinear parabolic PDEs

We are interested in semilinear parabolic PDEs

$$-\tfrac{\partial}{\partial t}v - f(x, v, \nabla_x v) - \langle \mu(x), (\nabla_x v)\rangle - \tfrac{1}{2}\mathrm{Trace}\big(\sigma(x)\sigma(x)^T(\mathrm{Hess}_x v)\big) = 0$$
$$v(T,x) = \varphi(x)$$

Existing literature (to name but a few:)

E, Fujii, Jaafari, Han, Henry-Labordere, Huré, Hutzenthaler, Grohs, Jentzen, Long, Mikael, Pham, Privault, Sirignano, Spiliopoulos, Takahashi, Warin, Yu.....

Most cases:

use nonlinear Feynman-Kac representation via BSDE to write solution of PDE as solution of (squared) minimization problem

Our approach:

Treat separately linear and nonlinear part (splitting method)

Deep splitting method: Derivation

1. Time discretization $0 = t_0 \le t_1 \le \cdots \le t_N = T$

- Set $V_T(x) = v(T,x) = \varphi(x)$
- Define recursively for each $n = N-1, \ldots, 0$ and $t \in [t_n, t_{n+1})$

$$V_t(x) = V_{t_{n+1}}(x) + f(x, V_{t_{n+1}}(x), \nabla_x V_{t_{n+1}}(x))(t_{n+1} - t_n) + \int_t^{t_{n+1}} \langle \mu(x), (\nabla_x V_s)(x) \rangle + \tfrac{1}{2}\mathrm{Trace}\big(\sigma(x)\sigma(x)^T (\mathrm{Hess}_x V_s)(x)\big)\, ds$$

Consequences: For each $n = N-1, \ldots, 0$

$$-\tfrac{\partial}{\partial t} V_t(x) - \langle \mu(x), (\nabla_x V_t)(x) \rangle - \tfrac{1}{2}\mathrm{Trace}\big(\sigma(x)\sigma(x)^T (\mathrm{Hess}_x V_t)(x)\big) = 0$$

$$v(t_n, x) \approx V_{t_n}(x)$$

Approximated 1 nonlinear PDE on [0,T] by N linear PDEs on $[t_n, t_{n+1})$, $n = N-1, \ldots, 0$

2. Feynman-Kac type of representation

Consider for some ξ

$$X_t = \xi + \int_0^t \mu(X_u)\, du + \int_0^t \sigma(X_u)\, dW_u$$

By *Itô's formula* and *PDE equation* for $V_t(x)$, we see for all $t \in [t_n, t_{n+1})$

$$V_{t_n}(X_{t_n}) = \mathbb{E}\big[V_t(X_t) \,\big|\, \mathbb{F}_{t_n}\big] = \mathbb{E}\big[V_t(X_t) \,\big|\, \sigma(X_{t_n})\big]$$

Using the splitting up discretization, we see that when $t \to t_{n+1}$

$$V_{t_n}(X_{t_n}) = \mathbb{E}\big[V_{t_{n+1}}(X_{t_{n+1}}) + f(x, V_{t_{n+1}}(X_{t_{n+1}}), \nabla_x V_{t_{n+1}}(X_{t_{n+1}}))(t_{n+1} - t_n) \,\big|\, \sigma(X_{t_n})\big]$$

3. From the property of (conditional) expectation

$$V_{t_n}(x) = \arg\min_{u \in C(\mathbb{R}^d)} \mathbb{E}\big[|V_{t_{n+1}}(X_{t_{n+1}}) + f(x, V_{t_{n+1}}(X_{t_{n+1}}), \nabla_x V_{t_{n+1}}(X_{t_{n+1}}))(t_{n+1} - t_n) - u(x)|^2\big]$$

4. Approximation by Neural Networks

- Set $\mathbb{V}_N^{\Theta}(x) := V_T(x) = v(T,x) = \varphi(x)$

- Recursively for each $n = N-1, \dots, 0$ use neural network $\mathbb{V}_n^{\Theta}(x)$ to learn $V_{t_n}(x)$

More precisely: Recursively for each $n = N-1, \dots, 0$ do:

Learn optimal parameter $\hat{\Theta}_n$ for $\mathbb{V}_n^{\Theta}(x)$ given already trained NN $\mathbb{V}^{\hat{\Theta}_{n+1}}(\cdot)$ s.t.

$$\mathbb{E}\big[|\mathbb{V}_{n+1}^{\hat{\Theta}_{n+1}}(X_{t_{n+1}}) + f(x, \mathbb{V}_{n+1}^{\hat{\Theta}_{n+1}}(X_{t_{n+1}}), \nabla_x \mathbb{V}_{n+1}^{\hat{\Theta}_{n+1}}(X_{t_{n+1}}))(t_{n+1} - t_n) - \mathbb{V}_n^{\hat{\Theta}_n}(x)|^2\big]$$
$$\approx \min_u \mathbb{E}\big[|\mathbb{V}_{n+1}^{\hat{\Theta}_{n+1}}(X_{t_{n+1}}) + f(x, \mathbb{V}_{n+1}^{\hat{\Theta}_{n+1}}(X_{t_{n+1}}), \nabla_x \mathbb{V}_{n+1}^{\hat{\Theta}_{n+1}}(X_{t_{n+1}}))(t_{n+1} - t_n) - u(x)|^2\big]$$

Consequence: For each $n = N-1, \dots, 0$

$$\mathbb{V}_n^{\hat{\Theta}_n}(\cdot) \approx V_n(\cdot) \approx v(t_n, \cdot)$$

Deep Splitting Algorithm

- **Define** $\mathbb{V}_N^{\Theta}(x) = v(T,x) = \varphi(x)$

- **Recursively** for each $n = N-1, \dots, 0$ learn optimal parameter $\hat{\Theta}_n \equiv \Theta_n^{(M)}$ by:

For each m:= 1, 2, ...

1. Simulate (for some ξ) paths for $X = \xi + \int_0^{\cdot} \mu(X_u)\, du + \int_0^{\cdot} \sigma(X_u)\, dW_u$
   using *Euler-Maruyama approximation scheme*, denoted by $(\mathcal{X}_n^m)_{n=0,\dots,N}$

2. Recursively for each $n = N-1, \dots, 0$ learn optimal parameter $\hat{\Theta}_n \equiv \Theta_n^{(M)}$
   by (stochastic) **gradient descent method**:

$$\Theta_n^{(m+1)} = \Theta_n^{(m)} - \gamma \cdot \nabla_\theta(\phi_n^{(m)})(\Theta_n^{(m)}), \qquad \text{where}$$
$$\phi_n^{(m)}(\theta) = |\mathbb{V}_{n+1}^{\hat{\Theta}_{n+1}}(\mathcal{X}_{n+1}^m) + f(x, \mathbb{V}_{n+1}^{\hat{\Theta}_{n+1}}(\mathcal{X}_{n+1}^m), \nabla_x \mathbb{V}_{n+1}^{\hat{\Theta}_{n+1}}(\mathcal{X}_{n+1}^m))(t_{n+1} - t_n) - \mathbb{V}_n^{\theta}(\mathcal{X}_n^m)|^2$$

Consequence: For M large enough, we have for each $n = N-1, \dots, 0$

$$\mathbb{V}_n^{\Theta_n^{(M)}}(\cdot) \approx v(t_n, \cdot)$$

Nonlinear Black-Scholes equation with default risks

- Start with Black-Scholes model with interest rate $r > 0$
- Include default risk for claim:
  If default occurs, you only get fraction $\delta \in [0,1)$ of current value
- model default by first time of Poisson process with intensity $Q(\cdot)$,
  where $Q(\cdot)$ is decreasing function of the current value
- Choose $Q(\cdot)$ piecewise linear with three regions with $(u^h < u^l, \gamma^h > \gamma^l)$

$$Q(y) = 1_{(-\infty,u^h)}(y)\,\gamma^h + 1_{[u^l,\infty)}(y)\,\gamma^l + 1_{[u^h,u^l)}(y)\left[\frac{(\gamma^h-\gamma^l)}{(u^h-u^l)}\,(y-u^h) + \gamma^h\right]$$

$\Longrightarrow$ price of claim is solution of semilinear PDE ([Bender et al., 2017])

$$-\frac{\partial}{\partial t}v(t,x) - rx\cdot\nabla v(t,x) - \frac{\sigma^2}{2}\sum_{i=1}^{d}|x_i|^2\frac{\partial^2}{\partial x_i^2}v(t,x) + (1-\delta)Q(v(t,x))v(t,x) + rv(t,x) = 0$$

$$v(T,x) = \varphi(x)$$

Information about our implementation

- **multi-layered neural network** with
  - d-dimensional input layer
  - Two $(d+10)$-dimensional hidden layers
  - 1-dimensional output layer
- 10 independent runs of the Algorithm
- Used **TensorFlow** on a NVIDIA GeForce GTX 1080 GPU with 1974 MHz core clock and 8 GB GDDR5X memory with 1809.5 MHz clock rate, where underlying system consists of Intel Core i7-6800K 3.4 GHz CPU with 64 GB DDR4-2133 memory

- $\varphi(x) = \min_{i \in \{1,\dots,d\}} x_i$
- **Calculated** $\mathbb{V}_0^{\hat{\Theta}_0}(x)$
- **approximate reference solution** obtained **by** multilevel Picard method [E, Hutzenthaler, Jentzen, Kruse, 2016]

d	Mean	Stdev	Ref. value	rel. L^1-error	Stdev rel. error	avg. runtime
10	40.6553107	0.1000347132	40.7611353	0.0029624273	0.0019393471	858.3129092
50	37.421057	0.0339765334	37.5217732	0.0026842068	0.0009055151	975.3706101
100	36.3498646	0.027989905	36.4084035	0.0016078403	0.000768776	1481.5484843
200	35.374638	0.035236816	35.4127342	0.0012857962	0.0006625744	951.2294598
300	34.8476466	0.0225350305	34.8747946	0.0008818254	0.0004762554	953.3183895
500	34.2206181	0.0081072294	34.2357988	0.0004701552	0.0001701012	956.0106124
1000	33.4058827	0.0050161752	33.4358163	0.0008952555	0.000150024	1039.5774061
5000	31.7511529	0.0048508218	31.7906594	0.0012427078	0.0001525864	7229.6752827
10000	31.1215014	0.0031131196	31.1569116	0.0011365119	0.00009991746	23593.212019

Open problem when training neural networks (NN)

- Training of NN involves stochastic gradient descent (SGD): Recall

$$\Theta_n^{(m+1)} = \Theta_n^{(m)} - \gamma \cdot \nabla_\theta(\phi_n^{(m)})(\Theta_n^{(m)}), \quad \textbf{but} \quad \theta \mapsto \phi_n^{(m)}(\theta) \quad \text{not convex!}$$

Remark: for strong L^p-convergence of SGD under convexity assumptions:

A. Jentzen, B. Kuckuck, A. Neufeld, P. von Wurstemberger:
Strong error analysis for stochastic gradient descent optimization algorithms
IMA Journal of Numerical Analysis, Vol. 41, No. 1, pp. 455-492, 2021

Want to analyze non-convex stochastic optimization problem

$$\text{minimize} \qquad \mathbb{R}^d \ni \theta \mapsto u(\theta) := \mathbb{E}[U(\theta, X)] \in \mathbb{R}$$

- $U\colon \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ non-convex
- X is $\mathbb{R}^m$-valued random variable.

Goal: Find Estimator $\widehat{\theta}$ to (approx.) minimize $\qquad \mathbb{E}[u(\widehat{\theta})] - \inf_{\theta \in \mathbb{R}^d} u(\theta)$

Focus on regularized optimization problem involving ReLU neural networks

$$\text{minimize} \quad \mathbb{R}^d \ni \theta \mapsto u(\theta) := \mathbb{E}\big[(f(Z) - \mathcal{N}_\sigma^{(1,\theta)}(Z))^2\big] + \eta \frac{\|\theta\|^{2(r+1)}}{2(r+1)} \qquad (1)$$

- $f : \mathbb{R}^{m_1} \to \mathbb{R}^{m_2}$ has polynomial growth
- $\sigma = \text{ReLU}, \quad r > 0, \quad \eta > 0$
- $X \equiv (f(Z), Z)$ is $\mathbb{R}^m$-valued random variable $(m = m_1 + m_2)$

A priori difficulties:

- $\mathbb{R}^d \ni \theta \mapsto u(\theta)$ non-convex
- $\mathbb{R}^d \ni \theta \mapsto \mathcal{N}_\sigma^{(1,\theta)}(Z)$ not differentiable, due to ReLU

However:

- If Z has a bounded density, then one can show that

$$\mathbb{R}^d \ni \theta \mapsto u(\theta) \qquad \text{is continuously differentiable!}$$

- Dissipativity condition for $h := \nabla u$:

$$\exists a, b > 0 \text{ s.t. } \forall \theta \in \mathbb{R}^d\colon \langle \theta, h(\theta) \rangle \geq a\|\theta\|^2 - b$$

$\Rightarrow \exists$ minimizer $\theta^* \in \mathbb{R}^d$, i.e. $u(\theta^*) = \inf_{\theta \in \mathbb{R}^d} u(\theta)$ (as u is coercive).

- One-sided Lipschitz continuity of h

$$\exists L > 0\colon \qquad \langle \theta - \theta', h(\theta) - h(\theta') \rangle \geq -L\|\theta - \theta'\|^2$$

Langevin stochastic differential equation ($h := \nabla u$)

$$dS_t = -h(S_t)\,dt + \sqrt{2\beta^{-1}}\,dW_t \qquad (\beta > 0 \text{ fixed parameter})$$

- SDE admits unique invariant measure π_β.
- π_β concentrates around minimizers of u (see [Dalalyan])

Tamed unadjusted stochastic Langevin algorithm (TUSLA):

$$\theta_0^{(\lambda,\beta)} := S_0, \qquad \theta_{n+1}^{(\lambda,\beta)} = \theta_n^{(\lambda,\beta)} - \lambda \frac{H(\theta_n^{(\lambda,\beta)}, X_{n+1})}{1 + \sqrt{\lambda}\|\theta_n^{(\lambda)}\|^{2r}} + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}$$

- $(X_n)_{n\in\mathbb{N}_0}$ i.i.d. copies of X defined in optimization problem (1)
- $(\xi_n)_{n\in\mathbb{N}_0}$ i.i.d. standard d-dim. Gaussian indep. of $(X_n)_{n\in\mathbb{N}_0}$.
- $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ such that $\nabla u(\theta) =: h(\theta) = E[H(\theta, X)]$, $\theta \in \mathbb{R}^d$
- stepsize $\lambda > 0$, inverse temperature parameter $\beta > 0$.

- $u(\theta) := \mathbb{E}\big[(f(Z) - \mathcal{N}_\sigma^{(1,\theta)}(Z))^2\big] + \eta\frac{\|\theta\|^{2(r+1)}}{2(r+1)}$
- $\theta_{n+1}^{(\lambda,\beta)} = \theta_n^{(\lambda,\beta)} - \lambda\frac{H(\theta_n^{(\lambda,\beta)}, X_{n+1})}{1+\sqrt{\lambda}\|\theta_n^{(\lambda)}\|^{2r}} + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}$, $\theta_0^{(\lambda,\beta)} = S_0$, **(TUSLA)**

Theorem 1: Lim, N., Sabanis, Zhang

Let Z have a bounded density and let Z and S_0 be sufficiently integrable. Then there exist (explicit) $C \equiv C(r) > 0$, $\lambda_{max} \equiv \lambda_{max}(r) > 0$ such that for every $n \in \mathbb{N}_0$, $\beta > 0$, $0 < \lambda < \lambda_{max}$

$$\mathbb{E}\big[u(\theta_n^{(\lambda,\beta)})\big] - \inf_{\theta\in\mathbb{R}^d} u(\theta) \le Ce^{-C\lambda n} + C\lambda^{1/4} + \frac{C}{\beta}$$

Related results: Raginsky et al. (2017), Cheng et al. & Xu et al. (2018), Chau et al. (2019), Lovas et al. (2020)...

Idea of the proof of Theorem 1

- $\forall \beta > 0$ let $S_\infty^{(\beta)}$ be d-dim. r.v. with $\mathcal{L}(S_\infty^{(\beta)}) = \pi_\beta$ (invar. meas. of SDE)
- $\mathbb{E}\big[u(\theta_n^{(\lambda,\beta)})\big] - \inf_\theta u(\theta) \leq \mathbb{E}\big[u(\theta_n^{(\lambda,\beta)})\big] - \mathbb{E}[u(S_\infty^{(\beta)})] + \mathbb{E}[u(S_\infty^{(\beta)})] - \inf_\theta u(\theta)$

Theorem 2: Lim, N., Sabanis, Zhang

Let Z have a bounded density and let Z and S_0 be sufficiently integrable. Then there exist (explicit) $C \equiv C(r) > 0$, $\lambda_{max} \equiv \lambda_{max}(r) > 0$ such that for every $n \in \mathbb{N}_0$, $\beta > 0$, $0 < \lambda < \lambda_{max}$

$$W_2(\mathcal{L}(\theta_n^{(\lambda,\beta)}), \pi_\beta) \leq Ce^{-C\lambda n} + C\lambda^{1/4}$$

- **By using Theorem 2:** $\mathbb{E}\big[u(\theta_n^{(\lambda,\beta)})\big] - \mathbb{E}[u(S_\infty)] \leq Ce^{-C\lambda n} + C\lambda^{1/4}$

Proposition 3: Lim, N., Sabanis, Zhang, see also [Raginsky et al.]

Let S_0 (i.e.starting point of SDE) be sufficiently integrable. Then there exist (explicit) $C \equiv C(r) > 0$, s.t. for every $\beta > 0$

$$\mathbb{E}[u(S_\infty^{(\beta)})] - \inf_{\theta \in \mathbb{R}^d} u(\theta) \leq \frac{C}{\beta}$$

Summary & References

- Introduced deep learning based algorithm for semilinear parabolic PDE
- Idea is to use splitting up method and (lin.) Feynman-Kac Representation
- Discussed challenges of training neural networks
- Introduced TUSLA algo for nonconvex optimization with ReLU NN

C. Beck, S. Becker, P. Cheridito, A. Jentzen, A. Neufeld:
Deep splitting method for parabolic PDEs
SIAM Journal on Scientific Computing, Vol. 43, No. 5, pp. A3135-A3154, 2021

D.-Y. Lim, A. Neufeld, S. Sabanis, Y. Zhang:
Non-asymptotic estimates for TUSLA algorithm for non-convex learning with applications to neural networks with ReLU activation function
IMA Journal of Numerical Analysis, 2023

https://personal.ntu.edu.sg/ariel.neufeld/

OPTIMAL TRANSPORT MAP ESTIMATION IN GENERAL FUNCTION SPACES

JONATHAN NILES-WEED

Classification AMS 2020: 62G05

Keywords: Optimal transportation, Brenier maps, empirical processes

We study the problem of estimating a function T given independent samples from a distribution P and from the pushforward distribution $T_\sharp P$. This setting is motivated by applications in the sciences, where T represents the evolution of a physical system over time, and in machine learning, where, for example, T may represent a transformation learned by a deep neural network trained for a generative modeling task.

To ensure identifiability, we assume that $T = \nabla\phi_0$ is the gradient of a convex function, in which case T is known as an *optimal transport map*. The estimation of such maps was inaugurated by Hütter and Rigollet [2], and has been subsequently studied in a number of works [2, 1, 3, 4, 5]. These works all study estimation of T under the assumption that it lies in a Hölder class, but general theory is lacking. We present a unified methodology for obtaining rates of estimation of optimal transport maps in general function spaces. Our assumptions are significantly weaker than those appearing in the literature: we require only that the source measure P satisfy a Poincaré inequality and that the optimal map be the gradient of a smooth convex function that lies in a space whose metric entropy can be controlled. As a special case, we recover known estimation rates for Hölder transport maps, but also obtain nearly sharp results in many settings not covered by prior work. For example, we provide the first statistical rates of estimation when P is the normal distribution, between log-smooth and strongly log-concave distributions, and when the transport map is given by an infinite-width shallow neural network.

REFERENCES

[1] N. Deb, P. Ghosal, and B. Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *arXiv preprint arXiv:2107.01718*, 2021.

[2] J.-C. Hütter and P. Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.

[3] T. Manole, S. Balakrishnan, J. Niles-Weed, and L. Wasserman. Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*, 2021.

[4] B. Muzellec, A. Vacher, F. Bach, F.-X. Vialard, and A. Rudi. Near-optimal estimation of smooth transport maps with kernel sums-of-squares. *arXiv preprint arXiv:2112.01907*, 2021.

[5] A.-A. Pooladian and J. Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

COURANT INSTITUTE FOR MATHEMATICAL SCIENCES, NEW YORK UNIVERSITY, NEW YORK, NY
Email address: jnw@cims.nyu.edu

WHEN A SYSTEM OF REAL QUADRATIC EQUATIONS HAS A SOLUTION

MARK RUDELSON

Classification AMS 2020: Primary: 14P05. Secondary: 14Q30, 90C22.

Keywords: positive semidefinite relaxation, quadratic equations, algorithms

This talk is based on a joint paper with Alexander Barvinok [1]. It discussed a problem of providing a computationally efficient certificate for the existence of a real solution of a system of m real quadratic equations with n variables. The question of feasibility of this type of a system of equations systems appears in various computer science contexts, see, for example, [2], [3].

If m and n are both allowed to grow, the problem becomes computationally hard. Unless the computational complexity hierarchy collapses, no polynomial time algorithm provides a necessary and sufficient condition for the existence of solutions of such a system. In fact, testing the feasibility of an arbitrary system of real polynomial equations can be easily reduced to testing the feasibility of a system quadratic ones. First, we gradually reduce the degree of polynomials by repeatedly introducing new variables and equations of the type $\xi_{ij} - \xi_i\xi_j = 0$, which allows us to replace the product $\xi_i\xi_j$ of old variables by a single new variable ξ_{ij}, and hence eventually reduce a given polynomial system to a system

$$q_i(x) = 0 \quad \text{for} \quad i = 1, \dots, m,$$

where q_i are quadratic, not necessarily homogeneous, polynomials. Then we introduce another, more delicate change of variables which is based on a semi-definite relaxation. This semi-definite relaxation problem can be efficiently solved in polynomial time. It allows to further reduce the system of general quadratic equations to a system

$$q_i(x) = \langle Q_i x, x\rangle = \mathrm{tr}(Q_i) \quad \text{for} \quad i = 1, \dots, m, \tag{0.1}$$

where Q_i are $n \times n$ symmetric matrices and

$$\langle x, y\rangle = \sum_{i=1}^{n} \xi_i\eta_i \quad \text{for} \quad x = (\xi_1, \dots, \xi_n) \quad \text{and} \quad y = (\eta_1, \dots, \eta_n)$$

is the standard inner product in $\mathbb{R}^n$ and $\mathrm{tr}(A)$ denotes the trace of a matrix A.

We present computationally simple sufficient criterion for (0.1), to have a solution. First, note that any linear combination of the equations of the form (0.1) has the same form. Using this observation, we can choose the simplest set of matrices $Q_1, \dots, Q_m$ which generates the same system of equations. To this end, we introduce the standard inner product of matrices

$$\langle A, B\rangle = \mathrm{tr}(A^\top B).$$

Then without loss of generality, we can assume that the matrices $Q_1, \dots, Q_m$ form an orthonormal system with respect to this product. Computing such an orthonormal system is also efficient and can be done, for example, by the Gram-Schmidt procedure.

For an $n \times n$ real symmetric matrix Q, we denote by $\|Q\|_{\mathrm{op}}$ the operator norm of Q, that is, the largest absolute value of an eigenvalue of Q.

Now, we can formulate the main result.

Theorem 0.1. *There is an absolute constant $\eta > 0$ such that the following holds. Let $Q_1, \ldots, Q_m$, $m \geq 3$, be linearly independent $n \times n$ symmetric matrices. Suppose that*

$$\left\| \sum_{i=1}^m A_i^2 \right\|_{\mathrm{op}} \leq \frac{\eta}{m}$$

for some (equivalently, for any) orthonormal basis $A_1, \ldots, A_m$ of the linear subspace spanned by $Q_1, \ldots, Q_m$. Then the system of quadratic equations

$$\langle Q_i x, x \rangle = \mathrm{tr}(Q_i) \quad \textit{for} \quad i = 1, \ldots, m$$

has a solution $x \in \mathbb{R}^n$.

We also show that while the choice of the basis $A_1, \ldots, A_m$ depends on the orthogonalization procedure, the key quantity $\|\sum_{i=1}^m A_i^2\|_{\mathrm{op}}$ is independent of it. Moreover, since this quantity is the largest eigenvalue of a positive definite matrix, it can be computed efficiently.

To illustrate the applicability of Theorem 0.1, we show that if $Q_1, \ldots, Q_m$ are symmetric matrices with independent identically distributed entries, then with high probability, the system is feasible provided that $m \leq c\sqrt{n}$ for some absolute constant $c > 0$.

While the condition we obtain is of an algebraic nature, the proof relies on analytic tools including Fourier analysis and measure concentration.

REFERENCES

[1] A. Barvinok, M. Rudelson. When a system of real quadratic equations has a solution. *Adv. Math.* , 403 (2022), Paper No. 108391, 38 pp.

[2] D. Bienstock. A note on polynomial solvability of the CDT problem. *SIAM Journal on Optimization*, 26 (2016), no. 1, 488–498.

[3] L. Liberti, C. Lavor, N. Maculan and A. Mucherino. Euclidean distance geometry and applications. *SIAM Review* , 56 (2014), no. 1, 3–69.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MICHIGAN, ANN ARBOR, MI 48109-1043, USA
Email address: rudelson@umich.edu

APPROXIMATELY HADAMARD MATRICES AND RIESZ BASES IN RANDOM FRAMES

MARK RUDELSON

Classification AMS 2020: 60B20.

Keywords: random frames, random matrices, Hadamard matrices.

This talk was based on a joint paper with Xiaoyu Dong [4]. We considered a problem which originated in signal processing. This problem was quickly reduced to a question about sub-matrices of a random matrix. We showed that this probabilistic question can be further reduced to a completely deterministic problem about the existence of approximately Hadamard matrices. The solution of the last problem was achieved by a combination of number theoretic and probabilistic tools, thus going back to the realm of probability.

Let $n < N$ be natural numbers. A set of vectors $X_1, \dots, X_N \in \mathbb{R}^n$ is called a *frame* if

$$K(n,N)\,\|x\|_2^2 \le \sum_{j=1}^{N} \langle x, X_j\rangle^2 \le RK(n,N)\,\|x\|_2^2 \tag{0.1}$$

for all $x \in \mathbb{R}^n$. Here $R \ge 1$ is called the frame constant, and $K(n,N) > 0$ is some function of n and N. The frame is considered to be good if its frame constant is relatively small. The notation $\|x\|_2$ stands for the Euclidean norm of the vector $x = (x_1, \dots, x_n)$.

In the last 40 years, frame theory became a well-developed area of applied mathematics, see [1], [2], [3], and the references therein. A frame can intuitively be regarded as overcomplete basis in $\mathbb{R}^n$. Because of this property, frames became a valuable tool in signal transmission. A signal which is viewed as an n-dimensional vector can be encoded by the sequence of its inner products with the frame vectors. If this sequence is transmitted over a communication line, then the original signal can be reconstructed even if part of the coefficients is lost or corrupted in the process of transmission. Moreover, this encoding is robust, which means that if the inner products are evaluated with some noise, then the reconstructed version will be close to the original one with the error depending on the noise magnitude.

One of the most popular classes of frames in algorithmic applications is the set of random frames. Such frames became also the method of choice in compressed sensing where one needs to reconstruct a low complexity signal from a small number of linear measurements, see, e.g., [5]. For example, if complexity is measured as the size of the support, and the support itself is unknown, the random frames provide robust recovery with optimal or almost optimal theoretical guarantees.

We consider a problem when a random frame contains a copy, or many copies of a "nice" basis. This problem can be conveniently translated to the language of random matrices. Define the condition number of a matrix A as

$$\kappa(A) = \frac{\max_{\|x\|_2=1} \|Ax\|_2}{\min_{\|x\|_2=1} \|Ax\|_2}.$$

With this notation, the frame property (0.1) can be rewritten as $\kappa(A_{n,N}) \leq C$ where $A_{n,N}$ is the $n \times N$ matrix with columns $X_1, \ldots, X_N$. Thus, the problem of existence of a 'nice" basis in a random frame can be recast as the question of existence of one or many well-conditioned square $n \times n$ sub-matrices of an $n \times N$ random matrix $A_{n,N}$ with i.i.d. entries. Our main result shows that the probability of finding such a sub-matrix undergoes a phase transition when N is exponential in terms of n. Since the upper and the lower bound hold under somewhat different assumptions, we formulate them separately.

Denote by $[N]$ the set $\{1, \ldots, N\}$. Let A be an $n \times N$ matrix. If $I \subset [N]$, denote by A_I the sub-matrix of A whose columns belong to I. The following theorem shows that if N is exponential in n, then with high probability, the $n \times N$ random matrix has many square submatrices with uniformly bounded condition numbers. In the language of frames, it means that a random frame with exponentially many vectors contains a large number of bases whose frame constants are uniformly bounded.

Theorem 0.1. *Let A be an $n \times N$ matrix with i.i.d. symmetric non-degenerate entries. Then there exist constants $c, C, \alpha, \beta > 0$ depending on the distribution of entries of A with the following property.*
Assume that $N \geq \exp(Cn)$. Then there exists $L \geq \exp(cn)$ such that

$$\mathbb{P}\left(\text{exist disjoint subsets } I_1, \ldots, I_L \text{ of } [N] \text{ with } |I_j| = n \text{ and } \kappa(A_{I_j}) < \alpha \text{ for all } j \in [L]\right) \geq 1 - \exp\left(-\exp(\beta n)\right).$$

The strategy of proving Theorem 0.1 relies on finding columns of A which are close to the columns of a certain deterministic $n \times n$ matrix V having a bounded condition number. The key to this strategy is a successful choice of the pattern matrix V. The requirement that a column of A can be close to a column of V with a non-negligible probability forces us to look for a matrix V which is a scaled copy of a matrix with ± 1 entries. Such matrices are known in some cases. For instance, the condition number of any Hadamard matrix is one. An $n \times n$ matrix H is called Hadamard if $n^{-1/2}H$ is an isometry. Hadamard matrices is a well-studied subject, and a number of constructions of such matrices are available. Yet, the dimensions in which Hadamard matrices were constructed are rare, so we have to use *approximately Hadamard matrices* instead.

The existence of an approximately Hadamard matrix in any dimension is established in the next theorem.

Theorem 0.2. *There exists a constant $C \geq 1$ such that for any $n \in \mathbb{N}$, one can find an $n \times n$ matrix V with ± 1 entries satisfying*

$$\kappa(V) \leq C.$$

The proof of Theorem 0.2 relies on Vinogradov's theorem from analytic number theory and combines number-theoretic and probabilistic ideas.

The conclusion of Theorem 0.1 holds under minimal assumptions on the distribution of entries. If we assume that the entries of the matrix are sub-gaussian, then the bound of Theorem 0.1 becomes sharp. Recall that a random variable X is called subgaussian if there is $a > 0$ such that $\mathbb{E} \exp\left(\frac{X^2}{a^2}\right) \leq 2$. Subgaussian random variables form a large family containing many naturally arising ones, see, e.g. [5].

The next theorem shows that finding a submatrix with a bounded condition number requires an exponential number of columns for matrices with subgaussian entries.

Theorem 0.3. *Let X be a centered subgaussian random variable. Then there exist $C, c, \tilde{c}, t_0 > 0$ with the following property. Let $t > t_0$, and assume that*

$$N \leq \exp\left(\frac{\tilde{c}}{t^4} n\right).$$

Let A be an $n \times N$ matrix whose entries are independent copies of X. Then

$$\mathbb{P}\left(\exists I \subset [N] \; |I| = n \text{ and } \kappa(A_I) < t\right) \leq \exp\left(-c\frac{n^2}{t^4}\right).$$

REFERENCES

[1] P. Casazza, G. Kutyniok, F.Philipp. Introduction to finite frame theory. Finite frames, 1–53, Appl. Numer. Harmon. Anal., Birkhäuser/Springer, New York, 2013.

[2] O. Christensen. Frames and bases. An introductory course. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, Inc., Boston, MA, 2008.

[3] O. Christensen. An introduction to frames and Riesz bases. Second edition. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, 2016.

[4] X. Dong, M. Rudelson. Approximately Hadamard matrices and Riesz bases in random frames. *Int. Math. Res. Not.*, (2024), no. 3, 2044–2065.

[5] R. Vershynin. High-dimensional probability. An introduction with applications in data science. With a foreword by Sara van de Geer. Cambridge Series in Statistical and Probabilistic Mathematics, 47. Cambridge University Press, Cambridge, 2018.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MICHIGAN, ANN ARBOR, MI 48109-1043, USA
Email address: rudelson@umich.edu

RECENT DEVELOPMENTS IN GROUP TESTING: FUNDAMENTAL LIMITS AND ALGORITHMS

JONATHAN SCARLETT

Classification AMS 2020:

Keywords: Group testing, sparsity, information-theoretic limits, discrete algorithms

The group testing problem concerns discovering a small number of defective items within a large population by performing tests on pools of items. A test is positive if the pool contains at least one defective, and negative if it contains no defectives. This is a sparse inference problem with a combinatorial flavor, with applications in medical testing, biology, multi-access communication, database systems, and more.

In this talk, I reviewed recent advances in the mathematics of group testing, including both information-theoretic limits and performance bounds for practical algorithms, with an emphasis on the following defining features:

- Non-adaptive testing (all tests must be designed in advance) vs. adaptive testing (tests are designed sequentially based on previous outcomes)
- Noiseless testing (tests are perfectly reliable) vs. noisy tests (some test outcomes are corrupted)

These two features lead to 4 distinct settings with varying degrees of difficulty. The over-arching goal is to determine the smallest possible number of tests while maintaining reliable recovery of the defective set. This question has been studied from a wide variety of perspectives, including high-dimensional statistics, information theory, discrete algorithms, combinatorics, error-correcting codes, and others.

The results that I surveyed in this talk are summarized as follows:

- The noisy adaptive setting has long been very well-understood, with a prominent approach being Hwang's adaptive binary splitting algorithm [9].
- The noiseless non-adaptive setting was extensively studied over the last decade or so (by myself and others), eventually leading to exact information-theoretic limits and a method for matching them in with an efficient algorithm [1, 12, 2, 10, 6, 7].
- The noisy non-adaptive setting has had analogous developments and generally lagged behind [4, 13, 14], but very recent works (including ours) works have substantially closed these gaps [5, 8].
- The noisy adaptive setting was seemingly overlooked for a long time, and my works showed that it can offer significant reductions in the number of tests compared to non-adaptive methods [11, 15].

A detailed survey covering all of these settings can be found in [3].

REFERENCES

[1] M. Aldridge, L. Baldassini, and O. Johnson, "Group testing algorithms: Bounds and simulations," *IEEE Trans. Inf. Theory*, vol. 60, no. 6, pp. 3671–3687, June 2014.

[2] M. Aldridge, "The capacity of Bernoulli nonadaptive group testing," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7142–7148, 2017.

[3] M. Aldridge, O. Johnson, and J. Scarlett, "Group testing: An information theory perspective," *Found. Trend. Comms. Inf. Theory*, vol. 15, no. 3–4, pp. 196–392, 2019.

[4] C. L. Chan, P. H. Che, S. Jaggi, and V. Saligrama, "Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms," in *Allerton Conf. Comm., Ctrl., Comp.*, Sep. 2011, pp. 1832–1839.

[5] J. Chen and J. Scarlett, "Exact thresholds for noisy non-adaptive group testing," *arXiv preprint arXiv:2401.04884*, 2024.

[6] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick, "Information-theoretic and algorithmic thresholds for group testing," in *Int. Colloq. Aut., Lang. and Prog. (ICALP)*, 2019.

[7] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick, "Optimal group testing," in *Conf. Learn. Theory (COLT)*, 2020.

[8] A. Coja-Oghlan, M. Hahn-Klimroth, L. Hintze, D. Kaaser, L. Krieg, M. Rolvien, and O. Scheftelowitsch, "Noisy group testing via spatial coupling," *arXiv preprint arXiv:2402.02895*, 2024.

[9] F. Hwang, "A method for detecting all defective members in a population by group testing," *J. Amer. Stats. Assoc.*, vol. 67, no. 339, pp. 605–608, 1972.

[10] O. Johnson, M. Aldridge, and J. Scarlett, "Performance of group testing algorithms with near-constant tests-per-item," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 707–723, Feb. 2019.

[11] J. Scarlett, "Noisy adaptive group testing: Bounds and algorithms," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3646–3661, June 2019.

[12] J. Scarlett and V. Cevher, "Phase transitions in group testing," in *ACM-SIAM Symp. Disc. Alg. (SODA)*, 2016.

[13] J. Scarlett and V. Cevher, "Near-optimal noisy group testing via separate decoding of items," *IEEE J. Sel. Topics Sig. Proc.*, vol. 2, no. 4, pp. 625–638, 2018.

[14] J. Scarlett and O. Johnson, "Noisy non-adaptive group testing: A (near-) definite defectives approach," *IEEE Trans. Inf. Theory*, vol. 66, no. 6, pp. 3775–3797, 2020.

[15] B. Teo and J. Scarlett, "Noisy adaptive group testing via noisy binary search," *IEEE Trans. Inf. Theory*, vol. 68, no. 5, pp. 3340–3353, 2022.

NUS SCHOOL OF COMPUTING (OFFICE COM3-02-57), 13 COMPUTING DRIVE, 117417
Email address: scarlett@comp.nus.edu.sg

SAMPLING FROM CONVEX SETS WITH A COLD START USING MULTISCALE DECOMPOSITIONS

HARIHARAN NARAYANAN, AMIT RAJARAMAN, AND PIYUSH SRIVASTAVA

ABSTRACT. A standard approach for sampling approximately uniformly from a convex body $K \subseteq \mathbb{R}^n$ is to run a random walk within K. The requirement is that starting from a suitable initial distribution, the random walk should "mix rapidly", i.e., after a number of steps that is polynomial in n and the aspect ratio R/r (here, K is assumed to contain a ball of radius r and to be contained within a ball of radius R), the distribution of the random walk should come close to the uniform distribution π_K on K. Different random walks differ in aspects such as the ease of implementation of each step, or suitability for a specific class of convex bodies. Therefore, the rapid mixing of a wide variety of random walks on convex bodies has been studied.

Many proofs of rapid mixing of such random walks however require that the initial distribution of the random walk is not too different from the target distribution π_K. In particular, they require that the probability density function of the initial distribution with respect to the uniform distribution π_K on K must be bounded above by $\mathrm{poly}(n)$: this is called a *warm start*. Achieving such a warm start often requires a non-trivial pre-processing step before the random walk can be started. This motivates the problem of proving rapid mixing from "cold starts", i.e., when the density of the initial distribution with respect to π_K can be as high as $\exp(\mathrm{poly}(n))$. In contrast to warm starts, a cold start is usually trivial to achieve. However, rapid mixing from a cold start may not hold for every random walk, e.g., the well-known "ball walk" does not have rapid mixing from an arbitrary cold start. On the other hand, for the "hit-and-run" random walk, Lovász and Vempala proved rapid mixing from a cold start. For the related *coordinate* hit-and-run (CHR) random walk, which has been found to be promising in computational experiments, a rapid mixing result starting from a warm start was proven only recently, while the question of whether CHR mixes rapidly from a cold start remained open.

In this paper, we construct a family of Markov chains inspired by classical multiscale decompositions of subsets of $\mathbb{R}^n$ into countably many axis-aligned cubes. We show that even with a cold start, the mixing times of these chains are bounded by a polynomial in n and the aspect ratio of the body. Our main technical ingredient is an isoperimetric inequality for K for a metric that magnifies distances between points that are close to the boundary of K. As a byproduct of the analysis of this new family of chains, we show that the coordinate hit-and-run (CHR) random walk also mixes rapidly from a cold start, and also from any point that is not too close to the boundary of the body.

Classification AMS 2020: 60J22, 68Q25

Keywords: Convex bodies, Markov chains, Isoperimetric inequalities.

A full version of the paper is available at arXiv:2211.04439. An extended abstract appears in *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC)*, June 2023, pp. 117–130.

RANDOM PERTURBATION OF LOW-RANK MATRICES

KE WANG

Classification AMS 2020: 60B20, 62H25, 62H30.

Keywords: Singular vector perturbation, singular subspace perturbation, low-rank structures, random matrices.

We consider an unknown $N \times n$ data matrix A. Suppose we cannot observe A directly but instead have access to a corrupted version $\widetilde{A}$ given by

$$\widetilde{A} := A + E,$$

where E represents the noise matrix. A classical question is to estimate the leading singular values and singular vectors (alternatively, eigenvalues and eigenvectors) of A using $\widetilde{A}$. This problem is crucial across various fields including engineering, statistics, machine learning, computer science, and mathematics. Generally, it is assumed that A is not arbitrary but exhibits specific structural characteristics, such as low-rank, which is a common assumption in numerous applications.

Assume that A has rank $r \geq 1$. The singular value decomposition (SVD) of A takes the form $A = U\Sigma V^T$, where $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_r)$ is a diagonal matrix containing the non-zero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ of A; the columns of the matrices $U = (u_1, \dots, u_r)$ and $V = (v_1, \dots, v_r)$ are the orthonormal left and right singular vectors of A, respectively. In other words, u_i and v_i are the left and right singular vectors corresponding to σ_i. It follows that $U^T U = V^T V = I_r$, where I_r is the $r \times r$ identity matrix. For convenience we will take $\sigma_{r+i} = 0$ for all $i \geq 1$. Denote the SVD of $\widetilde{A}$ similarly by $\widetilde{A} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T$, where the diagonal entries of $\widetilde{\Sigma}$ are the singular values $\widetilde{\sigma}_1 \geq \widetilde{\sigma}_2 \geq \cdots \geq \widetilde{\sigma}_{\min\{N,n\}} \geq 0$, and the columns of $\widetilde{U}$ and $\widetilde{V}$ are the orthonormal left and right singular vectors, denoted by $\widetilde{u}_i$ and $\widetilde{v}_i$, respectively.

The famous Davis–Kahan bound addresses this question for the eigenvectors of deterministic real symmetric matrices. An analogous version for the singular vectors of non-square matrices was established by Wedin. The key parameters in this bound are the gap (or separation) δ_1 between the largest singular values of A given by $\delta_1 := \sigma_1 - \sigma_2$ and the spectral norm of E defined by $\|E\| := \max_{\|u\|=1} \|Eu\|$, where $\|u\|$ denotes the Euclidean norm of the vector u. The classical Wedin's bound gives

$$\sin \angle(u_1, \widetilde{u}_1) \leq C \frac{\|E\|}{\delta_1},$$

where $\angle(u_1, \widetilde{u}_1)$ is the acute angle between u_1 and $\widetilde{u}_1$ taken in $[0, \pi/2]$ and $C > 0$ is an absolute constant. The same bound holds for $\sin \angle(v_1, \widetilde{v}_1)$.

For the case when E is a random matrix, an earlier paper by O'Rourke, Vu and myself [1] proved a version of bounds of the form

$$\sin \angle(u_1, \widetilde{u}_1) \lesssim \frac{r^{1/\alpha}}{\delta_1} + \frac{\|E\|}{\sigma_1} + \frac{\|E\|^2}{\sigma_1 \delta_1}, \tag{0.1}$$

which holds with high probability. Here, $\alpha > 0$ is a parameter that depends on the distribution of the random matrix E. When the rank r of A is sufficiently small compared to the dimensions and σ_1, δ_1 are sufficiently large, this bound improves upon the bound in Wedin's theorem. The first term on the right-hand side of (0.1) was conjectured as a consequence of the true dimension being actually r. The second term represents the signal-to-noise ratio. However, the third term on the right-hand side of (0.1) appears unnatural and unnecessary. By a completely different method, which uses tools from random matrix theory, in a recent joint work with O'Rourke and Vu [2], we removed the third term from (0.1) when the entries of E are independent and identically distributed (i.i.d.) copies of a standard normal random variable. In particular, we obtained that with high probability

$$\sin \angle(u_1, \widetilde{u}_1) \lesssim \frac{r\sqrt{\log(N+n)}}{\delta_1} + \frac{\|E\|}{\sigma_1}.$$

My talk is concerned with the most recent progress in this direction of research. In [3], we have enhanced the r-dependence in the bounds obtained from our previous work [2] and have eased some technical assumptions. For instance, for the largest singular vector, we have obtained that

$$\sin \angle(u_1, \widetilde{u}_1) \lesssim \frac{\sqrt{r+\log(N+n)}}{\delta_1} + \frac{\|E\|}{\sigma_1}.$$

Furthermore, we have presented a comprehensive extension of the Davis-Kahan-Wedin $\sin\Theta$ theorem. This extension applies to any unitarily invariant norm and operates under the assumption that E contains i.i.d. standard normal entries. Moreover, we have derived precise ℓ_∞ bounds for the perturbed singular vectors and the $\ell_{2,\infty}$ bounds for the perturbed singular subspaces of $A + E$. Beyond these specific bounds, we have also established results pertaining to the generalized components - also known as linear and bilinear forms - of the perturbed singular vectors and singular subspaces. We further investigate the $\ell_{2,\infty}$ bounds on the perturbed singular vectors, taking into account the weighting by their respective singular values. These fine-grained analysis is motivated by the substantial impact and wide-ranging applications these analyses offer in statistics and machine learning.

REFERENCES

[1] Sean O'Rourke, Van Vu and Ke Wang. Random perturbation of low rank matrices: improving classical bounds. Linear Algebra Appl., 540:26–59, 2018.

[2] Sean O'Rourke, Van Vu and Ke Wang. Matrices with Gaussian noise: optimal estimates for singular subspace perturbation. IEEE Transactions on Information Theory, 70(3):1978–2002, 2024.

[3] Ke Wang. Analysis of singular subspaces under random perturbations. arXiv:2403.09170.

DEPARTMENT OF MATHEMATICS, HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY, HONG KONG
E-mail address: kewang@ust.hk