Aryeh (Leonid) Kontorovich

Research Statement

I work in the field of machine learning, which I view as lying in the intersection of probability, statistics, and computer science. In recent years, I have focused on three main areas: metric spaces, Markov chains, and sample compression. I will detail my past, present and future research in these areas below.

Learning in metric spaces

Traditionally, much of the machinery for classification and regression algorithms, as well as finite-sample generalization bounds, depended strongly on the data residing in a Hilbert (or at least a Banach) space. For some important applications, this structural assumption was unrealistic, as it is often the case that the inherent geometry of the data is strongly non-Euclidean, or not even vectorial. In particular, consider the case of images. Here, the earthmover distance is commonly used [30] — yet it does not embed into the Euclidean (or, more generally, any ℓ_p) norm without a large distortion [26], which is liable to corrupt the data geometry before the learning process has even begun. The case of text strings is similar: the edit distance is known to be strongly non-Euclidean [2].

Over the past eight years or so, my collaborators and I have been pursuing a program of developing an efficient framework for classification and regression in metric spaces of low intrinsic dimensionality [7, 8, 10, 9, 23, 11]. We had raised fundamental questions, such as:

- What is the best performance of an efficient sample compression algorithm for nearest-neighbor classification?
- Can efficient (e.g., near-linear training and logarithmic evaluation time) regression be performed in a metric space?
- Is there a metric analogue of data-dependent dimensionality reduction (such as PCA)?
- How does one learn in a semimetric space (i.e., without the triangle inequality)?
- Do such methods yield Bayes-consistent learning procedures? Can they be combined with active learning to significantly improve sample complexity?

Encouraged by our success in largely resolving many of these basic questions, we continue to push for greater generality, better generalization bounds, and faster algorithmic runtimes.

A central feature of the previous work was its *pessimistic* (worst-case) analysis, which failed to recognize distributions with "nice" global behavior and to leverage this characteristic in the learning algorithm and analysis. Instead, these approaches required everywhere nice local behavior. I am therefore interested in how benign global distribution properties affect algorithmic runtimes and generalization error — in particular, when viewed through the lens of sample compression.

For a brief illustration, let us compare the "worst-case" and the "average-case" margins. Consider a sample of points in a metric space, annotated with binary labels. Its *worst-case margin* $\gamma_{\min}$ is the closest distance between any pair of opposite-labeled points. In a metric space with intrinsic dimension d, such samples can be compressed down to $(1/\gamma_{\min})^{O(d)}$ points; this quantity also controls the generalization error. Although this measure is worst-case optimal [11, 10], it is quite pessimistic: a single point pair can significantly degrade the margin of an otherwise well-separated sample. Instead, I would like to identify various notions of *average* margin $\bar{\gamma}$, which measure the well-separatedness of the sample as a whole, without being overly sensitive to occasional local irregularities. This average margin must also satisfy the desideratum that the more optimistic quantity $(1/\bar{\gamma})^{O(d)}$ continues to control the compression size and error rate. This has turned out to be surprisingly challenging, but recently, L. Gottlieb and I seem to have found the right notion and obtained some promising preliminary results.

My near-term agenda for continuing this work includes:

1. **Average margin.**
   - (a) Can problem-specific notions of average margin be defined for classification and regression problems in metric spaces which yield (i) better compression than via the worst-case (i.e., minimum) margin and (ii) efficient learning algorithms?
   - (b) Is there a notion of average margin for polytope learning in $\mathbb{R}^d$ [6] which allows for efficient learning and improved generalization bounds?
   - (c) The hard-SVM classifier in $\mathbb{R}^d$ optimizes the worst-case (geometric) margin, which is also used to derive generalization bounds. Can the worst-case margin be replaced by the average of the geometric margins at each sample point?
   - (d) Can a notion of average margin be connected to common distributional assumptions, such as Tsybakov noise?
2. **Bayes consistency.** A recently proposed metric-space supervised classification algorithm, based on taking majority vote on Voronoi cells induced by γ-nets (and referred to as KSU) [21, 22], has been shown to be Bayes-consistent on a strict superset of the problems for which k-NN succeeds. Can this majority-net algorithm be made to work in the average margin framework? Is there a Bayes-consistent version for regression? What about the *activized* majority-net classifier — is it (or a close variant) Bayes-consistent? How do these methods perform in the various adversarial-noise models?
3. **Non-metric spaces.** Can the above approaches be extended to non-metric (such as semimetric — i.e., those without the triangle inquality [10]) spaces? What about quasimetric spaces (i.e., those with an asymmetric distance function)?

Markov chains

My interest in Markov chains began during my Phd work, when I proved McDiarmid-type concentration inequalities for contracting Markov chains (and more generally, mixing processes [25]). In [18], I proposed a general technique for obtaining concentration from Markov contraction, and with my student Roi Weiss managed to significantly weaken the contraction assumption [24]. These results are have already found numerous applications, spanning the gamut from theoretical [1, 5, 17, 29, 20] to more applied [3, 19, 28, 36].

More recently, in collaboration with D. Hsu and Cs. Szepesvári, we obtained the first fully empirical confidence intervals on the spectral gap of a reversible ergodic Markov chain [16]; these were later sharpened together with D. Levin and Y. Peres [15].

My ongoing research with my PhD student G. Wolfer concerns estimating and testing properties of Markov chains from a single long observation sequence. Our first result was to obtain the first PAC-type minimax sample complexity rates for learning a Markov transition matrix [34]. In a recent breakthrough [33], we have managed to extend the results of [16, 15] to the non-reversible setting, via the notion of the *pseudo*-spectral gap [27]. Additionally, we have obtained a minimax result for Markov chain identity testing [35].

Sample compression

My interest in sample compression began while collaborting with L. Gottlieb [11], where we closed a long-standing gap in solutions to the Nearest Neighbor Condensing problem in metric spaces. We followed this up with the discovery [10] that sample compression continues to be a viable (and essentially, the only) learning technique in semimetric spaces. This technique, based on constructing γ-nets, was shown (by AK, S. Sabato, and R. Weiss) to be universally Bayes-consistent [22]. More dramatically, it turns out that our algorithm *optimistically universal* [13]: it succeeds whenever there is *some* Bayes-consistent learner. As such, it is strictly more general than the classic k-NN classifier, which requires the so-called Besicovitch property in order to be Bayes-consistent [4].

In a different line of work, in collaboration with S. Hanneke, we have established the tightness of some commonly invoked compression-based generalization bounds [12], and, together with my MSc student M.

Sadigurschi, given the first compression scheme for real-valued learners — both in the realizable and agnostic cases [14]. Our investigation naturally leads to some challenging open questions (including Warmuth's conjecture [32] and Simon's tight upper bound problem [31]).

References

[1] Radosław Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

[2] Alexandr Andoni and Robert Krauthgamer. The computational hardness of estimating edit distance. *SIAM J. Comput.*, 39(6):2398–2429, April 2010.

[3] M. Gheshlaghi Azar and H. J. Kappen. Asymptotic Performance Guarantee for Online Reinforcement Learning with the Least-Squares Regression. In *Pattern Analysis, Statistical modelling and ComputAtional Learning (PASCAL)*, 2011.

[4] Frédéric Cérou and Arnaud Guyader. Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics*, 10:340–355, 2006.

[5] Jean-René Chazottes and Frank Redig. Concentration inequalities for Markov processes via coupling. *Electron. J. Probab.*, 14:no. 40, 1162–1180, 2009.

[6] Lee-Ad Gottlieb, Eran Kaufman, Aryeh Kontorovich, and Gabriel Nivasch. Learning convex polytopes with margin. In *Neural Information Processing Systems (NIPS)*, 2018.

[7] Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data (extended abstract: COLT 2010). *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014.

[8] Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Adaptive metric dimensionality reduction (extended abstract: ALT 2013). *Theoretical Computer Science*, pages 105–118, 2016.

[9] Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate Lipschitz extension (extended abstract: SIMBAD 2013). *IEEE Transactions on Information Theory*, 63(8):4838–4849, 2017.

[10] Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Nearly optimal classification for semimetrics (extended abstract: AISTATS 2016). *Journal of Machine Learning Research*, 2017.

[11] Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors (extended abstract: NIPS 2014). *IEEE Trans. Information Theory*, 64(6):4120–4128, 2018.

[12] Steve Hanneke and Aryeh Kontorovich. A sharp lower bound for agnostic learning with sample compression schemes. In *ALT*, 2019.

[13] Steve Hanneke, Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Universal consistency in metric probability spaces via a 1-NN classifier (provisional title, in preparation). 2018+.

[14] Steve Hanneke, Aryeh Kontorovich, and Menachem Sadigurschi. Agnostic sample compression for linear regression, preprint. 2018.

[15] Daniel J. Hsu, Aryeh Kontorovich, David A. Levin, Yuval Peres, and Csaba Szepesvári. Mixing time estimation in reversible Markov chains from a single sample path. *CoRR*, abs/1708.07367, 2017.

[16] Daniel J. Hsu, Aryeh Kontorovich, and Csaba Szepesvári. Mixing time estimation in reversible Markov chains from a single sample path. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1459–1467, 2015.

[17] ShuLan Hu. Transportation inequalities for hidden Markov chains and applications. *SCIENCE CHINA Mathematics*, 54:1027–1042, 2011.

[18] Aryeh Kontorovich. Obtaining measure concentration from Markov contraction. *Markov Processes and Related Fields*, 4:613–638, 2012.

[19] Aryeh Kontorovich and Anthony E. Brockwell. A Strong Law of Large Numbers for Strongly Mixing Processes. *Communications in Statistics - Theory and Methods*, 43(18):3777–3796, 2014.

[20] Aryeh Kontorovich, Boaz Nadler, and Roi Weiss. On learning parametric-output hmms. In *ICML (3)*, pages 702–710, 2013.

[21] Aryeh Kontorovich, Sivan Sabato, and Ruth Urner. Active nearest-neighbor learning in metric spaces (extended abstract: NIPS 2016). *Journal of Machine Learning Research*, 18:195:1–195:38, 2017.

[22] Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Nearest-neighbor sample compression: Efficiency, consistency, infinite dimensions. In *Advances in Neural Information Processing Systems 30*, pages 1572–1582, 2017.

[23] Aryeh Kontorovich and Roi Weiss. Maximum margin multiclass nearest neighbors. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014*, pages 892–900, 2014.

[24] Aryeh Kontorovich and Roi Weiss. Uniform Chernoff and Dvoretzky-Kiefer-Wolfowitz-type inequalities for Markov chains and related processes. *Journal of Applied Probability*, 51:1–14, 2014.

[25] Leonid (Aryeh) Kontorovich and Kavita Ramanan. Concentration Inequalities for Dependent Random Variables via the Martingale Method. *Ann. Probab.*, 36(6):2126–2158, 2008.

[26] Assaf Naor and Gideon Schechtman. Planar earthmover is not in l_1. *SIAM J. Comput.*, 37:804–826, June 2007.

[27] Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1–32, 2015.

[28] Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. *J. Mach. Learn. Res.*, 99:1927–1956, August 2010.

[29] Afshin Rostamizadeh and Mehryar Mohri. Stability bounds for non-i.i.d. processes. In *Neural Information Processing Systems (NIPS)*, 2007.

[30] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[31] Hans Ulrich Simon. Bounds on the number of examples needed for learning functions. *SIAM J. Comput.*, 26(3):751–763, 1997.

[32] Manfred K. Warmuth. Compressing to VC dimension many points. In *Proceedings of the* 16^{th} *Conference on Learning Theory*, 2003.

[33] Geoffrey Wolfer and Aryeh Kontorovich. Estimating the mixing time of ergodic Markov chains. In *COLT*, 2019.

[34] Geoffrey Wolfer and Aryeh Kontorovich. Minimax learning of ergodic Markov chains. In *ALT*, 2019.

[35] Geoffrey Wolfer and Aryeh Kontorovich. Minimax testing of identity to a reference ergodic markov chain. *CoRR*, abs/1902.00080, 2019.

[36] Bin Zou, Zong-ben Xu, and Jie Xu. Generalization bounds of ERM algorithm with Markov chain samples. *Acta Mathematicae Applicatae Sinica (English Series)*, pages 1–16. 10.1007/s10255-011-0096-4.