RESEARCH STATEMENT

AMELIA PERRY

I design and study algorithms to analyze very noisy data. My graduate work centers a family of noisy geometric problems arising across the sciences, including structural biology, robotics, signals processing, and clustering in networks. Surprisingly, once we pare away the domain-specific language of these disparate problems, we find that they reduce to the same fundamental and challenging core data tasks.

The explosion of data in recent years has led to an explosion of attempts to algorithmically extract conclusions and insight out of data. The harder challenge is to design algorithms that are powerful, robust, and make sense. Convex optimization is the core of my algorithm design process, but also a perspective on algorithms in general: most algorithms that we really understand well, like PCA, can be viewed in this lens. Algorithm design is blind without an understanding of how algorithms perform, and why, and how to improve them; I obtain such insight using probability theory and statistics, most particularly with the rich theory of random matrices and random networks. In fact, problems such as cluster detection in networks exhibit sharp phase transitions, as weak data become stronger or numerous enough to suddenly coalesce into meaningful conclusions. I draw from the statistical physics of disordered systems to both locate phase transitions and algorithmically exploit these phenomena as well as possible. Lastly, to exploit the rich geometry in key problems such as cryo-electron microscopy (cryo-EM), I draw from the algebraic study of symmetry to form novel, non-obvious algorithmic tools, finding great pleasure in spanning abstract algebra, statistical physics, and practical microscopy within a single algorithm. My progression from the theoretical edge of mathematics into applied algorithm design equips me with an unusually broad toolkit to tackle the core challenges in data science, and the perspective to keep abstractions relevant to the reality of applications.

Energized by my recent work on cryo-EM, I next plan to dive deeper into the world of computational biology. I hope to play the role of cross-fertilizing the machine learning and biology communities, translating rich research problems from biology into a form amenable to theoretical study, and bringing the advanced tools of modern machine learning to bear on meaningful scientific problems.

1. Symmetry and synchronization

Cryo-electron microscopy (cryo-EM) is among the most impactful new developments in structural biology, as evidenced by the 2017 Nobel Prize in Chemistry. By cooling many copies of a biomolecule to cryogenic temperatures and imaging them in an electron microscope, it is in principle possible to stitch the resulting images together into a 3D reconstruction of the molecule. This algorithmic task is highly challenging. The microscopy images tend to be overwhelmingly noisy, giving only the barest hint of the molecule shape; moreover, each image captures the molecule at a different, unknown angle.

This problem is usually addressed by posing a guess of the molecule structure, then estimating the angle of each image based on matching it to the guess, then using the aligned images to improve the guess, and repeating. This approach has a fatal flaw: it is unclear how much of the estimated structure is simply bias from the initial guess, and how much is truly driven from the data. Indeed, incorrect guesses have lead to incorrect structures [Cohen, 2013]. Much of my work is in developing novel techniques for guess-free, *de novo* estimation of molecule structures from cryo-EM images.

My initial work follows the approach of [Singer and Shkolnisky, 2011], by estimating the angle of each image; once the angles are known, reconstruction becomes a classical tomography problem. We first focus on each pair of images; testing correlations along conjectural lines of overlap gives some information about the relative rotation $R_i R_j^{-1}$ from the angle R_j of one image to the angle R_i of another. Reconciling noisy estimates of all these pairwise relative rotations into clean estimates of the rotations R_i is the task of *synchronization*.

Synchronization may be considered over any abstract symmetry group. A similar image synchronization task over the Euclidean group $SE(3)$ is the basis of simultaneous localization and mapping (SLAM) for autonomous robots [Rosen et al., 2016]. Synchronization over the 2D rotations $SO(2)$ or a discretization $\mathbb{Z}/L$ is the core of certain signals processing problems, as well as of clustering in networks [Bandeira, 2015].

Classic approaches to synchronization include PCA (eigenvector methods), which throws away much of the structure of the problem and performs far from optimally. Indeed, an offshoot of my work [Perry et al., 2016c, Perry et al., 2016a] explores how optimal or sub-optimal PCA is at similar tasks. More sophisticated approaches [Bandeira et al., 2015] include semidefinite programming, which unfortunately is slow enough to be difficult to execute or empirically test. Reaching for a fast, iterative algorithm like PCA that better exploits structure, we devised in [Perry et al., 2016b] a form of Approximate Message Passing algorithm, an approach derived from the Thouless–Anderson–Palmer equations in statistical physics. We tailored this approach to the geometry and symmetry of a synchronization problem, through heavy use of the representation theory of compact groups—that is, the study of all ways that rotations can manifest in linear algebra. The result is a highly efficient and perhaps optimally powerful algorithm for synchronization over (essentially) any group, thus simultaneously addressing cryo-EM, robotics problems, and more.

My current work [Bandeira et al., 2017] aims to entirely circumvent estimating the imaging angles. Instead, we compute various rotationally-invariant statistics of the images, subtract the influence of noise to obtain statistics of the molecule structure, and attempt to solve for the molecule structure that fits those statistics. The choice of statistics draws from the invariant theory of rotation groups, thus bringing more abstract algebra to bear on cryo-EM.

2. Robustness

One core data task across all domains is the detection of communities within networks. This is the analogue of clustering for network data, which is increasingly prevalent, for instance as protein–protein interaction networks or gene regulatory networks in biology, as social networks, or as networks of interconnected infrastructure. A plethora of algorithms have been suggested, based on jargon such as Girvan–Newman modularity, random walks, spectral clustering, MCMC, semidefinite programming, and many other techniques [Abbe et al., 2015]. It is well-motivated then to compare algorithms and understand their strengths and weaknesses, when and why they fail, and to secure mathematical guarantees that they work under certain conditions. The *stochastic block model* (SBM) [Holland et al., 1983] is perhaps the most popular statistical model for networks structured roughly into clusters, and it provides a basis on which to compare algorithms.

An extensive literature exists on this topic, especially on reaching the sharp statistical limits of clustering in the SBM, as predicted through statistical physics by [Decelle et al., 2011]. This literature hits a pitfall: it invents algorithms which are fabulous on SBM networks, but which entirely fail on real-world networks, which often have many features that SBM networks lack, such as a range of node roles from major hubs down to minor leaves, following a power-law distribution [Faloutsos et al., 1999]. Some algorithms for the SBM crucially exploit its particular statistical properties to eke out all the performance possible, while losing sight of the scientific problem of understanding real networks. To illustrate: once a few 'tetrahedral' micro-clusters of four vertices are planted in an SBM network, a leading theoretical algorithm for the SBM breaks down entirely [Ricci-Tersenghi et al., 2016].

My work addresses this disconnect between theory and practice by replacing the SBM with a version that promotes robust, realistic algorithms. The *semirandom stochastic block model* begins with an SBM network, but allows an 'adversary' to alter the network by adding edges arbitrarily within clusters, and removing edges arbitrarily between clusters. These edits should only make the cluster structure stronger, which ought to help algorithms; and yet, these edits can break the fragile statistical properties of the SBM that many theoretical algorithms rely on, for instance by adding tetrahedral micro-clusters within communities, or adding edges to form hubs and simulate a power-law degree distribution. By raising the bar and requiring algorithms that cope with these edits, my work encourages robustness and practicality in algorithm design, bringing the theoretical discourse closer to actual scientific needs.

Specifically, in [Perry and Wein, 2017], we introduce a robust algorithm that recovers the clusters perfectly from the semirandom SBM whenever this is statistically possible, adapting gracefully to the edits described above. In [Moitra et al., 2016], we show that these edits can in fact move the phase transition between

clustering being impossible versus approximable, suggesting that 'optimal' algorithms that reach the classical phase transition must *necessarily* over-tune to the SBM, giving up robustness and practicality in doing so. Such a phenomenon seems to have never before been reported across the algorithms literature, and it tells an important story of how the theory community can go astray from practical needs.

3. Future directions

Energized by my recent work on cryo-EM, I am excited to dive further into biological application. My ideal role would be as a bridge between the biology and academic machine learning communities. One side of this is in bringing non-obvious machinery to bear on biological problems, much as in bringing abstract algebra and statistical physics to bear on cryo-EM. Indeed, simple but often sub-optimal methods like PCA are prevalent throughout the sciences, leaving a great deal of room for more tailored approaches to extract more meaning from large datasets. In group dynamics, my role recently has been often to communicate techniques from abstract algebra to applied mathematicians; indeed, in my current collaboration with algebraist Ben Blum-Smith, I have often taken the group role of translating his algebraic language to a format that my applied coworkers are comfortable with. I believe that these skills of interdisciplinary communication will aid me in communicating sophisticated machine learning techniques in a biological setting.

The flip side of this role is in maintaining a connection with the academic ML community and in presenting it with research problems from biology, so as to broaden collaboration between the fields. Much of the challenge here is in finding the right models and abstractions, so as to pose clean ML problems requiring a minimum of biology background, while keeping these abstractions relevant to application needs; my experience with robustness in network clustering has taught me how easily this can go awry, and how the right choice of model can help keep the field on track. I also enjoy coding algorithms and testing them empirically; I am eager for opportunities to work with actual biological data, both for the excitement of seeing meaningful results, and as an opportunity to stay anchored in what is really effective in practice, and to reconcile theory with that. While I undoubtedly have a lot to learn in the biology domain, I am very eager for this opportunity for scientific impact.

References

[Abbe et al., 2015] Abbe, E., Bandeira, A. S., and Hall, G. (2015). Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*.

[Bandeira, 2015] Bandeira, A. S. (2015). *Convex Relaxations for Certain Inverse Problems on Graphs*. PhD thesis, Princeton University.

[Bandeira et al., 2017] Bandeira, A. S., Blum-Smith, B., Perry, A., Weed, J., and Wein, A. S. (2017). Estimation under group actions: recovering orbits from invariants. *Working draft, available at* `http://math.mit.edu/%7Eawein/orbit-recovery.pdf`.

[Bandeira et al., 2015] Bandeira, A. S., Chen, Y., and Singer, A. (2015). Non-unique games over compact groups and orientation estimation in cryo-EM. *arXiv:1505.03840*.

[Cohen, 2013] Cohen, J. (2013). Is high-tech view of HIV too good to be true? *Science*, 341(6145):443–444.

[Decelle et al., 2011] Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106.

[Faloutsos et al., 1999] Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, volume 29, pages 251–262. ACM.

[Holland et al., 1983] Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

[Moitra et al., 2016] Moitra, A., Perry, W., and Wein, A. S. (2016). How robust are reconstruction thresholds for community detection? In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing (STOC 2016)*, pages 828–841. ACM.

[Perry and Wein, 2017] Perry, A. and Wein, A. S. (2017). A semidefinite program for unbalanced multisection in the stochastic block model. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 64–67.

[Perry et al., 2016a] Perry, A., Wein, A. S., and Bandeira, A. S. (2016a). Statistical limits of spiked tensor models. *arXiv:1612.07728*.

[Perry et al., 2016b] Perry, A., Wein, A. S., Bandeira, A. S., and Moitra, A. (2016b). Message-passing algorithms for synchronization problems over compact groups. *Communications on Pure and Applied Mathematics*. To appear.

[Perry et al., 2016c] Perry, A., Wein, A. S., Bandeira, A. S., and Moitra, A. (2016c). Optimality and sub-optimality of PCA I: spiked random matrix models. *Annals of Statistics*. To appear.

[Ricci-Tersenghi et al., 2016] Ricci-Tersenghi, F., Javanmard, A., and Montanari, A. (2016). Performance of a community detection algorithm based on semidefinite programming. In *Journal of Physics: Conference Series*, volume 699, page 012015. IOP Publishing.

[Rosen et al., 2016] Rosen, D. M., Carlone, L., Bandeira, A. S., and Leonard, J. J. (2016). A certifiably correct algorithm for synchronization over the special euclidean group. *arXiv:1611.00128*.

[Singer and Shkolnisky, 2011] Singer, A. and Shkolnisky, Y. (2011). Three-dimensional structure determination from common lines in cryo-EM by eigenvectors and semidefinite programming. *SIAM Journal on Imaging Sciences*, 4(2):543–572.