The goal of my research is to develop a better understanding of how and when machine learning succeeds, and the broader societal implications of data driven systems. My work lies at the intersection of machine learning, optimization, and statistics, including topics such as high dimensional learning and algorithmic fairness.

The remarkable and continuing empirical success of modern machine learning, especially deep learning, has opened up several foundational questions about our understanding of why learning models succeed. Answering these questions has the potential to expand the theoretical and algorithmic frontiers of learning and optimization. At the same time, as machine learning proliferates into the real world, it is of profound importance to critically evaluate the broad societal impacts of these systems. I am interested in developing a theoretical understanding of learning models that will lead to principled methods for building systems with positive social impact.

My recent research spans two lines of work that contribute to these broad themes. My primary line of work builds new theory towards understanding the role of optimization in the success of modern machine learning models, especially deep neural networks. My second line of work is on studying the challenges in building machine learning systems that avoid *discrimination* against protected population groups.

Inductive bias from optimization in learning. In typical machine learning systems, the parameters of a prediction model are estimated by minimizing some loss over data samples (training). In the modern practice of machine learning, especially in deep learning, many successful models are highly overparameterized with far more trainable parameters than the number of training examples. Consequently, the resulting optimization objective has many minimizers that will simply overfit to the training data and perform poorly on new examples. In practice though, when such overparameterized objectives are minimized using simple algorithms like (stochastic) gradient descent ((S)GD), the *specific minimizers* returned by these algorithms have remarkably good performance on new examples. Indeed, large scale neural networks learned in this way continue to outperform traditional non-overparameterized models in many applications. A natural question then is:

Why do models learned using (S)GD work so well even with overparameterized objectives?

By picking some specific minimizer of the training loss, the optimization algorithm introduces additional *inductive bias* into learning, which is not explicitly specified in the objective. Studying the nature of models learned with this inductive bias is crucial for understanding the success and limitations of deep learning and would lead to new algorithms for practice. *My work establishes new foundational theory on formalizing the "special" type of minimizers that common algorithms yield when optimizing overparameterized objectives.*

Learning non-discriminatory predictors. As machine learning begins to play a role in socially consequential domains, for ethical and legal reasons, there is a pressing need to develop tools for measuring and mitigating objectionable discrimination by these automated systems. My second line of research looks at formalisms of *non-discrimination* against protected groups, and broadly explores the question of:

When is it feasible to learn accurate and non-discriminatory prediction models?

In algorithmic terms, learning non-discriminatory predictors often amounts to enforcing non-trivial constraints on the learning problem, *e.g.,* equal error rates across different groups. *My work identifies key bottlenecks in terms of the statistical feasibility and the computational tractability of learning with such non-discrimination constraints.*

Research agenda. My interests are broadly driven by statistical, algorithmic, and societal aspects of machine learning. My research combines tools from a range of mathematical areas including optimization, learning theory, and dynamical systems, while staying relevant to practical deployment. The projects and their extensions described in the remainder of the statement constitute my immediate research agenda. Aside from these topics, in the past, I have also worked on high dimensional estimation, matrix completion, and ranking problems.

Going forward, I plan to advance my research on developing strong theoretical foundations for responsible and efficient use of machine learning. I envision that these theoretical insights would inform practical guidelines for developing new and advanced systems with less trial and error than is present in current practice.

Inductive bias from optimization in learning

When optimizing overparameterized training objectives, optimization algorithms like (S)GD bias the final solution to some special minimizers of the training loss. In learning deep neural networks, much of the inductive bias of the learned model comes from such implicit bias of optimization (*e.g.,* [18, 15, 22]). My work described below gives precise answers to which specific minimizers different algorithms yield for different overparameterized models. These results lay the foundation for a theoretical understanding of the modern practice of deep learning.

Implicit regularization in matrix factorization

In my joint work with Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro (NIPS 2017 [4]), we studied the dynamics of gradient descent on a simple non-convex objective. This was one of the early work that rigorously established how the inductive bias from optimization can indeed lead to useful and interesting models even in overparameterized problems. Consider a least squares loss $\|\mathcal{X}(\mathbf{W})-\mathbf{y}\|^2$ over the matrix variable $\mathbf{W}$, where $\mathcal{X}$ is some linear measurement operator and $\mathbf{y}$ are the observations. If the number of observations is small compared to matrix dimensions, this objective is ill-posed with many matrices achieving zero loss. For example, in matrix completion, the objective is merely the squared loss over the observed entries, which is trivially minimized by any imputation of the missing values. Thus, in order to get useful completions, we typically need to impose additional constraints, such as low-rankness, on the matrix variables [11].

Intriguingly, we empirically observed that even when no explicit constraints were imposed on the objective, some specific optimization techniques implicitly returned near low-rank solutions. Specifically, by representing $\mathbf{W}$ as a product of two *unconstrained* and full dimensional factor matrices $\mathbf{W}=\mathbf{U}\mathbf{V}$, we get an exact equivalent objective of $\|\mathcal{X}(\mathbf{U}\mathbf{V})-\mathbf{y}\|^2$ over $\mathbf{U},\mathbf{V}$. Gradient descent on this new ill-posed objective was then observed to return solutions with good performance on low-rank problems. At the same time, gradient descent for the original objective of $\|\mathcal{X}(\mathbf{W})-\mathbf{y}\|^2$ yielded very different solutions, which were not typically low-rank. We conjectured and provided further empirical and theoretical evidence that with infinitesimal step-size and infinitesimal initialization, gradient descent over the factor matrices converges to the minimum *nuclear norm* solution. This explains our empirical observations since minimizing nuclear norm has been well known to yield generalization and recovery in low rank problems [21, 19, 11]. In a follow up work, Li, Zhang, and Ma [17] proved the conjecture for a large class of interesting matrix estimation problems.

Relating optimization geometry to optimization bias

For an underdetermined linear least squares objective, gradient descent can be shown to yield the minimum *Euclidean norm* solution. In my work described above [4], we showed that gradient descent in matrix factorization problems is biased towards the minimum nuclear norm solution, which amounts to minimizing the Euclidean norm of the factor matrices. Since the gradient descent updates are given by local steepest descent in the geometry of Euclidean norm, these results hint at a link between the optimization bias and the optimization geometry. This link was also observed in another paper on minimizing strictly monotone losses like logistic or exponential loss for classification tasks [6]. In this case, we proved that gradient descent yields the ℓ_2 *max-margin* classifier, which maximizes the Euclidean distance of the closest training data to the decision boundary.

To understand the role of the optimization geometry further, in my joint work with Jason Lee, Daniel Soudry, and Nati Srebro (ICML 2018 [2]), we studied algorithms that work under different (non-Euclidean) geometries. This included mirror descent with respect to general potentials and steepest descent in general norms, where the potentials and the norms specify the corresponding geometry. Our results formalized strong connections between the inductive bias from optimization and the geometry of the updates. We further showed that the nature of the optimization bias is very different for strictly monotone losses like the exponential loss compared to the squared loss. For the squared loss, the solution returned by mirror descent can be robustly specified in terms of the mirror descent potential and independent of step-sizes. However, for steepest descent in general norms, the optimization bias could strongly depend on step-sizes. The situation is fundamentally different for the exponential loss, where we proved that steepest descent in any norm converges to the classifier that maximizes margin in that norm. In general, for strictly monotone losses, the asymptotic behavior of the optimization iterates are often independent of the initialization and step-sizes, which allows for simpler characterizations of the optimization bias.

Implicit bias of gradient descent in linear networks

In my more recent joint work with Jason Lee, Daniel Soudry, and Nati Srebro (NeurIPS 2018 [3]), we uncovered interesting inductive biases from optimizing over multi-layer linear networks. Multi-layer linear networks represent composition of multiple linear maps $W_1, W_2, \ldots$ as $\mathbf{x} \to W_1 \cdot W_2 \cdot \ldots \cdot W_L \cdot \mathbf{x}$. We primarily contrasted the inductive bias from gradient descent in two types of networks: (a) fully connected linear networks where the component maps $W_1, W_2, \ldots$ are unconstrained, and (b) linear convolutional networks where $W_1, W_2, \ldots$ are constrained to represent circular convolutions. Both types of networks ultimately represent the model class of linear classifiers. Thus, the resulting optimization objectives for training are entirely equivalent. Nevertheless, we proved that optimizing these differently parameterized objectives with gradient descent leads to very different solutions. For fully connected networks of any depth, we again get the maximum ℓ_2 margin classifier.

More interestingly, training linear convolutional networks with gradient descent biases towards classifiers that are sparse in the frequency domain (frequency filtering). Specifically, we proved that the linear classifier resulting from gradient descent is related to minimizing a sparsity inducing penalty on the Fourier coefficients of the classifier. Furthermore, the specific sparsity inducing penalty is the $\ell_{2/L}$ *quasi-norm* which changes with the depth L of the network—the penalty induces sparsity more aggressively for larger depths. These results showed how different architectures can lead to interesting inductive biases even for linear networks.

In another of paper of ours, along with Pedro Savarese and Mor Nacson [5], we dove deeper into this study by calculating the rate at which the gradient descent iterates converge to the maximum margin classifiers. While with a fixed step-size, this is extremely slow at $1/\log(t)$, we also proved that the rate can be improved to $\log(t)/\sqrt{t}$ merely by using aggressive step-sizes akin to normalized gradient descent. Initial empirical results suggest that such choice of step-sizes might also be useful for training deep networks with non-linear activations.

Going forward: exploiting optimization for learning

While deep non-linear neural networks used in practice are certainly more complex than the models we have analyzed so far, our current results provide sufficient intuition for taking the next steps in understanding and exploiting optimization bias. I am pursuing two broad directions of research in this topic.

1. *Exploring the relationship between optimization path and regularization path.* In our results, we have related the models learned by specific algorithms to some special minimizers of the training loss $\mathcal{L}(\mathbf{w})$. These special minimizers, beyond fitting the training data, also minimize some additional penalty (say $\phi(\mathbf{w})$) on the parameters. In an alternate view, these special solutions can be thought of as the limit of the corresponding *regularization path* given by $\text{argmin}_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \lambda\phi(\mathbf{w})$ in the limit of $\lambda \to 0$. Based on our analysis so far, we suspect a strong link between the optimization iterates and the regularization path. For non-linear neural networks, we suspect the limit solutions of the *regularization path* as potential candidates for characterizing the inductive bias from specific algorithms. Furthermore, the regularization effect of early stopping can also be understood in this framework by establishing a fine-grained connection to the regularization path in the non-asymptotic regimes.

2. *Better optimization algorithms for neural network training.* Complementary to characterizing the optimization bias, I am also interested in developing better empirical methods for learning overparameterized models. For example, motivated by our analysis of convergence rate of normalized gradient descent [5], we already see some evidence for the effectiveness of using similar normalized gradient steps for training even non-linear networks. Along this line, I am more broadly interested in the design of models and algorithms where the uncovered inductive biases from optimization are incorporated directly through more efficient implementations or regularization.

Learning non-discriminatory predictors

In the literature on algorithmic fairness, several criteria formalizing different fairness desiderata have been proposed (see surveys in [9, 20, 8]). The results described below focuses on one representative criterion called the *equalized odds* [14]. In binary classification tasks (*e.g.*, lending), equalized odds requires achieving same false negative and false positive rates across all protected groups. More broadly, consider an underlying joint distribution over input features X, target labels Y, and additional protected attributes such as race and gender denoted as A. Common criteria for a prediction model $f(X)$ to be non-discriminatory against protected groups are specified as non-trivial restrictions on the joint distribution over A, Y, and $f(X)$ [10, 13, 14, 12, 16]. Equalized odds in particular requires the predictions $f(X)$ to be independent of the protected attribute A when conditioned on the labels Y. Learning non-discriminatory predictors in this case amounts to incorporating the distributional restriction into the learning problem. My work looks at the statistical and computational issues that arise in learning prediction models with such non-discrimination constraints.

Incorporating non-discrimination in batch learning

In my joint work with Blake Woodworth, Mesrob Ohanessian, and Nati Srebro (COLT 2017 [7]), we took the first steps towards developing a statistical and computational theory of learning non-discriminatory predictors. Suppose we have a dataset of i.i.d. examples from an underlying population distribution. The statistical question we answer is: *what is the best predictor we can learn that would be as good and as non-discriminatory as possible on the population?* The complementary computational question is: *can such a predictor be efficiently learned?*

For the statistical question, we first showed that a naïve constrained empirical risk minimizer is statistically sub-optimal. We then presented a two-step learning rule that achieves near optimal bounds on accuracy and approximate non-discrimination. However, we also proved that in there are distributions under which learning such approximate equalized odds predictors is computationally intractable. We alternatively proposed relaxations of equalized odds criteria that allow for efficient learning of non-discriminatory predictors.

Online learning of non-discriminatory predictors

In many decision tasks, availability of i.i.d. samples as training data is an unreasonable model of reality. For example, in applications like education and hiring, not only does the data distribution change over time, but due to historical biases the observed data does not also represent independent samples. Online learning is an alternate framework that captures prediction tasks with non-i.i.d. data. In online learning, samples of input-label pairs arrive sequentially in an arbitrary order and the learner makes a prediction at each instance before the true label is revealed. The goal is to design an algorithm for online prediction whose performance is comparable to the best predictor chosen in hindsight from a model class (*no-regret* property).

In my joint work with Thodoris Lykouris, Nati Srebro, and Avrim Blum (NeurIPS 2018 [1]), we studied the question of: *can we design a no-regret online learning algorithm such that the cumulative statistics of predictions also satisfy equalized odds*? We considered a benign setting for binary classification where all the competing predictors individually satisfy equalized odds, and further, all the protected groups have substantial representation in the population. This makes the task trivial with i.i.d. data. Without the i.i.d. assumption though, surprisingly, we showed a strong impossibility result that even in such benign settings, one cannot have an algorithm that achieves both diminishing regret as well as diminishing deviation from equalized odds. This negative result brings to question the appropriateness of the cumulative equalized odds criterion in online learning. We further suggested the notion of *equalized error rates*, which requires equalizing the overall error rates across all the groups. We proved that this alternate criterion allows for online learning of non-discriminatory predictions.

Ongoing work and future directions on algorithmic fairness

In the completed work so far, we have pointed out significant statistical and computational bottlenecks in learning non-discriminatory predictors. These results motivate alternate definitions or relaxations that are cognizant of these fundamental challenges. One direction I am currently pursuing involves studying how the statistical and computational complexities of learning non-discriminatory models compare to the complexity of detecting non-discrimination. For example, detecting violations of a conditional independence property requires computational models for independence tests and access to samples from the data and prediction distribution. By restricting the number of such samples or by restricting the computation for detection, one can now study the applicability of a range of relaxations of the non-discrimination criterion. More broadly, in this line of work, I am interested in studying how considerations of algorithmic challenges can inform the formalism of non-discrimination criteria.

Overall, it is evident that there are no easy fixes for mitigating the social and legal concerns with automated systems. I believe there is a need for interdisciplinary and continuously evolving effort towards regulating these systems. I plan to actively contribute to such efforts in my future research and teaching.

Select Publications

[1] A. Blum, S. Gunasekar, T. Lykouris, and N. Srebro. On preserving non−discrimination when combining expert advice. In *Conference on Neural Information Processing Systems (NeurIPS) [to appear]*, 2018.

[2] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning (ICML)*, 2018.

[3] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Conference on Neural Information Processing Systems (NeurIPS) [to appear]*, 2018.

[4] S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[5] M. S. Nacson, J. Lee, S. Gunasekar, P. Savarese, N. Srebro, and D. Soudry. Convergence of gradient descent on separable data. *arXiv preprint arXiv:1803.01905*, 2018.

[6] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. In *Journal of Machine Learning Research 2018 [to appear]*, 2018.

[7] B. Woodworth, S. Gunasekar, M. Ohanessian, and N. Srebro. Learning non−discriminatory predictors. In *Conference on Learning Theory (COLT)*, 2017.

Other References

[8] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. 2017.

[9] S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, 2016.

[10] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *International Conference on Data Mining Workshops (ICDMW)*, 2009.

[11] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 2009.

[12] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 2017.

[13] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[14] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Conference on Neural Information Processing Systems*, 2016.

[15] N. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2016.

[16] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations of Theoretical Computer Science*, 2016.

[17] Y. Li, T. Ma, and H. Zhang. Algorithmic regularization in over-parameterized matrix recovery. In *Conference on Learning Theory*, 2018.

[18] B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations*, 2015.

[19] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 2010.

[20] A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *Knowledge Engineering Review*, 2014.

[21] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, 2005.

[22] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Collaborations

I have had the privilege of working with and learning from many wonderful researchers who have helped shape my research skills and agenda. My primary mentors are Joydeep Ghosh (UT) and Nati Srebro (TTIC) with whom I have worked with closely on many projects. Over the years, I am grateful to have also been mentored by Avrim Blum (TTIC), David Sontag (MIT), Arindam Banerjee (UMN), Pradeep Ravikumar (CMU), Alan Bovik (UT), and Sujay Sanghavi (UT). The projects overviewed above developed out of productive collaborations with many colleagues and students in the field. The work on *implicit bias of optimization* is largely based on joint work with long term collaborators Jason Lee (USC), Daniel Soudry (Technion), Mor Shpigel (Technion), Pedro Savarese (TTIC), and Nathan Srebro (TTIC). The projects on *algorithmic fairness* were primarily spearheaded by Thodoris Lykouris (Cornell) and Blake Woodworth (TTIC) under the mentorship of Avrim Blum and Nathan Srebro. I have also had great fun working with my other co-authors including Behnam Neyshabur (NYU), Srinadh Bhojanapalli (Google), Mesrob Ohanessian (TTIC), Sanmi Koyejo (UIUC), Shalmali Joshi (UT), Ayan Acharya (Netflix), and Sindhu Raghavan (Netflix), many of whom have become close personal friends over time.