TRIPODS+X:RES: Investigations at the Interface of Data Science and Neuroscience

Project Summary

The research in this proposal lies at the interface between neuroscience and data science. The underlying theme is to develop a two-way channel between data science and cellular and cognitive neuroscience. In one direction, we will investigate how computational principles of data science can be used to understand recent empirical findings in neuroscience, associated with measurements at the cellular level in fruit flies, and brain imaging studies in humans. In the reverse direction, the project will view the processes and mechanisms of vision and cognition underlying these findings as a source for new mathematical frameworks for data analysis. The research will focus on four interrelated objectives:

Objective 1: Distributed processing. Recent work in machine learning has studied the effect of communication constraints and parallelization in distributed estimation. There is a close parallel in vision, where any given input is sensed by multiple parts of the retina and an accurate percept needs to be constructed. The project will consider different models for distributed processing, motivated by learning and perception in both lower-level organisms (visual processing in fruit flies) and higher-level cognition (visual cognition in humans).

Objective 2: Data representation. The brain stores the same information in several different ways, each emphasizing different dimensions of the input. Inspired by current understanding of representation of information in different regions of the cortex, the project will investigate how parallel lossy representations of the same inputs along different dimensions can be used for statistically and computationally efficient learning algorithms. In the other direction, we will investigate how embedding algorithms from machine learning might be used as mechanisms for processing massive cellular and brain imaging data.

Objective 3: Attentional filtering. The project will develop attention-based models in statistical learning, based on the use of lower-dimensional traces or curves through a high-dimensional input space. Attention curves have analogues in human cognition, where input dimensions are processed based on their inherent salience and relevance to a person's goals. The project will explore mathematical, computational and empirical models of attention. Experiments will focus on a public dataset of subjects watching episodes of Sherlock while being scanned with fMRI, using a frame-by-frame annotation of several dimensions of the movie as the basis for attention-based models.

Objective 4: Memory capacity. Evidence from studies of human behavior suggests that people store information about objects and events in long-term memory with incredible detail. How is this possible? We will consider cognitive studies and a current understanding of possible memory architectures in natural systems in order to inform approaches for reducing and sharing memory in artificial learning algorithms. A framework will be developed for establishing lower bounds on the risk of machine learning algorithms under memory constraints. Insights from this mathematical theory will be considered in the context of memory of complex organisms.

The ***intellectual merit*** of the proposed research includes the transfer of ideas between data science and neuroscience, with the goal of advancing knowledge in both domains. The ***broader impact*** of the research includes development of software that implements the advanced data science and machine learning algorithms, the development of labs for an undergraduate course in data science with examples drawn from neuroscience, and a series of workshops hosted at Yale and Brown Universities that expand the scope of the original TRIPODS effort at Brown.

TRIPODS+X:RES: Investigations at the Interface of Data Science and Neuroscience

Project Description

1. Introduction

This project will develop new statistical theory and methodology that accounts for constraints that are present in many modern data analysis problems, but which have not been thoroughly studied in traditional statistical approaches. We will explore theory, algorithms, and applications of nonparametric and parametric statistical procedures, with constraints imposed on the storage, computational complexity, shape, or physics of the estimators. The project will develop practical models, statistical theory, and provably correct algorithms that can help scientists to conduct more effective data analysis. Application areas include astronomy, modeling consumer preferences, energy management in operating systems, and large scale convex optimization.

Classical statistical theory studies the rate at which the error in an estimation problem decreases as the sample size increases. Methodology for a particular problem is developed to make estimation efficient, and lower bounds establish how quickly the error can decrease in principle. Asymptotically matching upper and lower bounds together yield the minimax rate of convergence

$$R_n(\mathcal{F}) = \inf_{\widehat{f}} \sup_{f\in\mathcal{F}} R(\widehat{f}, f).$$

This is the worst-case error in estimating an element of a model class $\mathcal{F}$, where $R(\widehat{f}, f)$ is the risk or expected loss, and $\widehat{f}$ is an estimator constructed on a data sample of size n. The corresponding sample complexity of the estimation problem is $n(\epsilon, \mathcal{F}) = \min\{n : R_n(\mathcal{F}) < \epsilon\}$.

In the classical setting, the infimum is over all estimators. In modern data analysis, it is increasingly of interest to understand how error depends on computation. The use of heuristics and approximation algorithms may make computation more efficient, but it is important to understand the loss in statistical efficiency that this incurs. In the minimax framework, this can be formulated by placing computational constraints on the estimator:

$$R_n(\mathcal{F}, B_n) = \inf_{\widehat{f}: C(\widehat{f}) \leq B_n} \sup_{f\in\mathcal{F}} R(\widehat{f}, f).$$

Here $C(\widehat{f}) \leq B_n$ indicates that the computation $C(\widehat{f})$ used to construct $\widehat{f}$ is required to fall within a "computational budget" B_n. Minimax lower bounds on the risk as a function of the computational budget thus determine a feasible region for computation-constrained estimation, and a Pareto-optimal tradeoff for error versus computation.

The general theme of the research problems outlined in this proposal is to explore constraints in statistical estimation and inference. In a minimax formulation, constraints are placed on the estimator, $C(\widehat{f}) \leq B$. Such constraints might entail computational cost in terms of flops. But they could also involve storage constraints, specifically the number of bits used to represent the estimator. Alternatively, the constraints might come from energy restrictions, as will be discussed below in the context of power management for operating systems.

Constraints might also be placed on the model class $\mathcal{F}$, rather than on the estimators $\widehat{f}_n$. Classical model spaces include Sobolev spaces and Besov spaces, which make smoothness assumptions on the model. But in modeling a physical system, the time-dependent behavior might be required to obey a conservation of energy principle, evolve according to one of a large family of partial differential equations, or be governed by the laws of mechanics, as for example in the the motion of an exoplanet around a star. As described in the following sections, the focus of our investigations will be a series of challenges that incorporate constraints in these and other ways.

2. Statistical Inference Under Communication Constraints (Aim 1)

Background and Preliminary Results. One important measure of computation is storage, or the space used by a procedure. In particular, we may wish to limit the number of bits used to represent an estimator. The fundamental statistical problem is to understand how the excess risk depends on any such storage constraint.

This problem is naturally motivated by certain applications. For instance, the Kepler telescope collected flux data for approximately 150,000 stars (Jenkins et al., 2010); see Figure 4. The central statistical task is to estimate the lightcurve of each star nonparametrically, in order to denoise and detect planet transits. If this estimation is done onboard the telescope, the estimated function values may need to be sent back to Earth for further analysis. To limit communication costs, the estimates can be quantized. The fundamental question is, what is lost in terms of statistical risk in quantizing the estimates? Or, in a cloud computing environment (such as Amazon EC2), a large number of nonparametric estimates might be constructed over a cluster of compute nodes and then stored (for example in Amazon S3) for later analysis. To limit the storage costs, which could dominate the compute costs in many scenarios, it is of interest to quantize the estimates. How much is lost in terms of risk, in principle, by using different levels of quantization?

With such applications as motivation, we have obtained preliminary results on the problem of risk-storage tradeoffs in the normal means model of nonparametric estimation. The normal means model is a centerpiece of nonparametric estimation. It arises naturally when representing an estimator in terms of an orthogonal basis (Johnstone, 2002; Tsybakov, 2008). Our first result is a sharp characterization of the Pareto-optimal tradeoff curve for quantized estimation of a normal means vector, in the minimax sense. We consider the case of a Euclidean ball of unknown radius in $\mathbb{R}^n$. This case exhibits many of the key technical challenges that arise in nonparametric estimation over richer spaces, including the Stein phenomenon and the problem of adaptivity.

This problem is intimately related to classical rate distortion theory (Gallager, 1968). Indeed, our approach is based on a marriage of minimax theory and rate distortion ideas. We thus build on the fundamental connection between function estimation and lossy source coding that was elucidated in Donoho's 1998 Wald Lectures (Donoho, 2000). Our result for the Euclidean ball (Zhu and Lafferty, 2015) is shown in Figure 1. Zhang et al. (2013) also consider minimax bounds with communication constraints. However, their analysis is focused on distributed parametric estimation, where the data are distributed between several machines.

Proposed Research A: Quantized estimation over Sobolev ellipsoids. The standard white noise model for nonparametric regression is

$$dY(t) = f(t)dt + \varepsilon dW(t)$$

where f lies in the periodic Sobolev space $\widetilde{W}_m(c) = \{f \in L_2[0,1] \;:\; \{\theta_j\} \in \Theta(m,c)\}$ where $\theta_j = \langle f, \varphi_j \rangle$ for the trigonometric basis, and $\Theta(m,c)$ is the ellipsoid $\Theta(m,c) = \left\{\theta \;:\; \sum_{j=1}^{\infty} j^{2m}\theta_j^2 \le \frac{c^2}{\pi^{2m}}\right\}$. We observe data $Y_j = \int_0^1 \varphi_j(t)dY(t) = \theta_j + \varepsilon\xi_j$ where $\xi_j \sim N(0,1)$. The minimax risk at noise level ε is given by $R_\varepsilon(m,c) = \inf_{\widehat{\theta}} \sup_{\theta\in\Theta(m,c)} \mathbb{E}\|\widehat{\theta} - \theta\|^2$, and the classical Pinsker minimax bound is

$$R(m,c) = \liminf_{\varepsilon\to 0}\; \varepsilon^{\frac{-4m}{2m+1}}\, R_\varepsilon(m,c) \ge \left(\frac{c}{\pi^m}\right)^{\frac{2}{2m+1}} (2m+1)^{\frac{1}{2m+1}} \left(\frac{m}{m+1}\right)^{\frac{2m}{2m+1}}$$

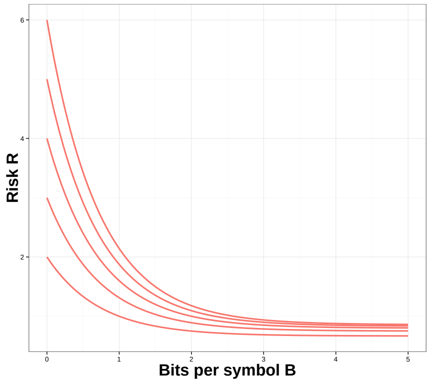

Figure 1. Our result establishes the Pareto-optimal tradeoff in the nonparametric normal means problem for risk versus number of bits:

$$R(\sigma^2, c^2, B) = \frac{c^2\sigma^2}{\sigma^2 + c^2} + \frac{c^4 2^{-2B}}{\sigma^2 + c^2}.$$

Curves for five signal sizes are shown, $c^2 = 2, 3, 4, 5, 6$. The noise level is $\sigma^2 = 1$. With zero bits, the rate is c^2, the highest point on the risk curve. The rate for large B approaches the Pinsker bound $\sigma^2 c^2/(\sigma^2 + c^2)$.

which defines the *Pinsker constant* $P_m(c)$ on the righthand side. We have obtained preliminary results that lower bound the quantized minimax risk. Specifically, four regimes are defined. For many bits $B\varepsilon^{2/(2m+1)} \to \infty$, we have the usual minimax rate with the Pinsker constant. However, with fewer bits $B\varepsilon^{2/(2m+1)} \to b$, the lower bound has the same rate of convergence, but a different constant,

$$\liminf_{\varepsilon \to 0} \varepsilon^{-\frac{4m}{2m+1}} R_\varepsilon(c, B) \geq Q_m(c, b)$$

where $Q_m(c, b)$ is the solution of a cubic equation in b. If $B\varepsilon^{2/(2m+1} \to 0$ and $B \to \infty$ then

$$\liminf_{\varepsilon \to 0} B^{2m} R_\varepsilon(c, B) \geq \frac{c^2}{\pi^{2m}} m^{2m}.$$

Finally, if B is constant then the asymptotic risk is

$$\liminf_{\varepsilon \to 0} R_\varepsilon(c, B) \geq \frac{c^2}{\pi^{2m}} \left(\exp\left(\frac{B}{m}\right) \ell! \right)^{2m/\ell}$$

where ℓ is the integer satisfying

$$\frac{\ell^\ell}{\ell!} < \exp\left(\frac{B}{m}\right) \leq \frac{(\ell+1)^{(\ell+1)}}{(\ell+1)!}.$$

We expect to be able to prove achievability of these lower bounds for the specified constants. Our approach will combine methods based on adaptive estimation over Sobolev spaces and randomized coding schemes from rate distortion theory.

Proposed Research B: Efficient coding algorithms. To make quantized estimation practical, it is important to design computationally efficient quantized estimators. One possible method is to divide the variables into smaller blocks and quantize them separately. A more interesting and promising approach is to leverage recent advances in computationally efficient, near-optimal lossy compression using sparse regression algorithms (Venkataramanan et al., 2013). These compression algorithms came on the heels of sparse regression methods for channel coding over the AWGN channel (Barron and Joseph, 2010; Joseph and Barron, 2012). We will explore modifications of greedy regression for quantized nonparametric estimation, with the goal of achieving practical algorithms. It will be necessary to trade off a worse error exponent in the convergence rate to the optimal quantized minimax risk.

Proposed Research C: Quantized testing. We have so far described nonparametric estimation under storage or communication constraints. It is also important to study inference and testing under such constraints.

Consider again the Kepler telescope data. After the lightcurves are detrended, the exoplanet inference task is a testing problem: Are the data zero mean Gaussian noise (the null), or is there a sparse, periodic signal indicating planet transits (the alternative)? Now, suppose that the residuals are quantized. How much is lost, say in terms of the excess minimax risk, when a budget of B bits is available?

Working in the white noise model, we have $dX_\varepsilon(t) = f(t)dt + \varepsilon dW(t)$, and wish to test H_0 that $f \in \Theta_0$ against H_1 that $f \in \Theta_1$. The risk of the problem is defined as $\psi_t = \inf \psi \left(t\alpha(\psi, \Theta_0) + \beta(\psi, \Theta_1)\right)$ where ψ is any test statistic, α is the Type I error rate and β is the Type II error rate. In a constrained procedure we place computational restrictions (such as storage) on the procedures ψ. To begin, as we have done for estimation in Zhu and Lafferty (2015), we will consider the nonparametric testing problem $H_0 : \theta = 0$ against $H_1 : \|\theta\| > \rho_\varepsilon$, imposing the constraint that the test statistic has storage/communication complexity of B bits. We will then consider the Sobolev and Besov cases, for which detailed results are known classically (Ingster, 1993; Ingster and Suslina, 2000, 2003).

Such analysis will be of interest theoretically, and has the potential to inform resource constrained data collection and analysis at very large scales.

Proposed Research D: Storage/risk tradeoffs for other models. We will study communication-constrained estimation for other nonparametric models. Wavelet methods are well known to be adaptive to local smoothness. Specifically, wavelet methods applied to the Besov space $B^{\varsigma}_{p,q}(c)$ achieve the minimax rate, and no linear estimator can achieve this adaptation. A natural line of investigation is to consider quantized wavelet techniques, where the coefficients are coded at each level of the multiresolution analysis. Quantized estimation using overcomplete dictionaries and deep networks (Hinton et al., 2012; Wan et al., 2013) will be another line of investigation.

We will also study quantized nonparametric estimation for additive models. Apart from minimax concerns, it will be interesting to consider practical algorithms for quantized estimation of an additive model. Backfitting is the standard approach to fitting an additive model (Hastie and Tibshirani, 1999). The challenge will be to consider iterative coding algorithms that refine the codewords for a given function based on the current quantization for the other functions. In related recent work, we have considered a different type of constraint in multivariate additive models, where we constrain the rank of the estimators (Foygel et al., 2012).

Quantized estimation for covariance matrices is another natural and important problem. We will consider this in the model of covariance matrices studied by Bickel and Levina (2008), which we have adopted in some of our other recent work (Shender and Lafferty, 2013). Finally, we will explore quantized Bayesian inference. One possible approach is to quantize the data $x_{1:n}$, which will incur error in the resulting posterior distribution. Formulating this precisely, in terms of an optimality principle, is an interesting challenge.

3. Computation-Risk Tradeoffs (Aim 2)

Background and Preliminary Results. Modern data sets requiring statistical analysis are often large and high dimensional. The computation required to construct standard estimators for such data may be prohibitive. We would like to understand the fundamental tradeoffs between statistical accuracy and computational scalability—tolerating increased predictive error, or risk, in exchange for more favorable computational requirements. While several heuristics for reduced computation are often possible, including dimension reduction, sampling, and greedy algorithms, little is known about this problem from first principles.

Sparse PCA is one problem that has been studied from the perspective of trading off computation for sample complexity. In the stylized setting of a sparse rank one covariance corrupted by noise, where the principal eigenvector of dimension p has only k nonzero entries, a simple thresholding algorithm has been shown to have sample complexity $O(k^2 \log(p-k))$ with computational complexity $O(np + p\log p)$ (Johnstone and Lu, 2004). In contrast, a more expensive semidefinite relaxation algorithm is known to have smaller sample complexity $O(k\log(p-k))$ at the expense of greater computational complexity $O(np^2 + p^4 \log p)$ (d'Aspremont et al., 2004; Amini and Wainwright, 2009). These analyses assume, however, that the solution has rank one; Berthet and Rigollet (2012) show that this cannot generally be achieved in polynomial time. This problem has also been studied by Chandrasekaran and Jordan (2013). In the setting of online learning, Agarwal et al. (2012) consider model selection under a computational budget constraint, assuming that computation grows linearly with sample size.

Linear regression is a workhorse method for many statistical problems. But without special assumptions, the method has quadratic computational cost $O(np^2)$ in the dimension p, when the sample size n is larger than p. This may be prohibitive when p is large. Even in this simplest setting, it is important to consider computational tradeoffs. A growing body of work is investigating algorithms for scaling regression to large data sets, notably using coresets (Drineas et al., 2006; Dasgupta et al., 2009; Clarkson and Woodruff, 2012) and sketching (Raskutti and Mahoney, 2014). Recently, Clarkson and Woodruff (2013) proposed a new algorithm for generating subspace embedding matrices, which yields a regression algorithm running in $O(m(X)) + \widetilde{O}(p^3\varepsilon^{-2})$ time, with $m(X)$ denoting the number of nonzeros of X.

We have studied a method that sparsifies the sample covariance in order to enable fast algorithms for solving the linear system (Shender and Lafferty, 2013). The standard ridge regression estimator is

$$\widehat{\beta}_\lambda = \Big(\frac{1}{n}\mathbb{X}^T\mathbb{X} + \lambda_n I\Big)^{-1}\frac{1}{n}\mathbb{X}^T Y \tag{1}$$

$$= (\mathbb{S} + \lambda_n I)^{-1}\, b_n \tag{2}$$

where $\mathbb{X}$ is the $n \times p$ design matrix, $\mathbb{S} = \frac{1}{n}\mathbb{X}^T\mathbb{X}$ is the sample covariance, and $b_n = \frac{1}{n}\mathbb{X}^T Y$ is the sample marginal correlation for data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, assuming for convenience that the data are scaled to have mean zero and variance one. We consider the family of estimators $\widetilde{\beta}_{t,\lambda} = \big(\mathbb{S}_t + \lambda_n I\big)^{-1} b_n$ where $\mathbb{S}_t$ is a sparsified version of the sample covariance obtained by hard thresholding, to zero out the small entries. That is, $\mathbb{S}_t = T_t(\mathbb{S})$ where $T_t([m_{ij}]) = \big[m_{ij}\mathbb{1}(|m_{ij}| > t)\big]$. The basic intuition is that as we increase the threshold t, so that the matrix $\mathbb{S}_t$ becomes more sparse, the model degrades, but the estimator can be obtained with less computation. For sufficiently large regularization level λ_n and sparsity threshold t, the linear system $\big(\mathbb{S}_t + \lambda_n I\big)\beta = b_n$ is sparse and symmetric diagonally dominant (SDD). Recent research in algorithms and scientific computation has led to a breakthrough in fast solvers for such systems. In particular, work of Spielman and Teng (2009) and Koutis et al. (2012) shows that sparse SDD systems can be solved in near linear time in the number of nonzero entries in the matrix. Since calculation of $\mathbb{S}_t$ is parallelizable in a simple and direct manner, and the cost of the computation can be amortized over different regressions, we adopt a computational model in which the sparsification $\mathbb{S}_t$ is not included in the computational cost. Our main result is a combination of the computational analysis with a statistical analysis of the predictive risk for this family of linear models, making precise the tradeoff between computation and error in this setting.

Proposed Research A: Graduated dropout and subspace projections. We have begun to study a procedure to iteratively "drop out" entries in the data matrix, drawing inspiration from a recent heuristic for training large multi-layer neural networks for image processing (Hinton et al., 2012; Wan et al., 2013). The dropout method in deep learning is used as a form of regularization to avoid over-fitting. A formal correspondence with regularization was established in Wager et al. (2013). In contrast, our rationale for employing the dropout is to enable more efficient computation of an estimator, while controlling the excess risk that this data corruption incurs. To achieve this we will make use of recent algorithmic advances in "subspace embedding" techniques (Clarkson and Woodruff, 2012; Nelson and Nguyen, 2012), mentioned

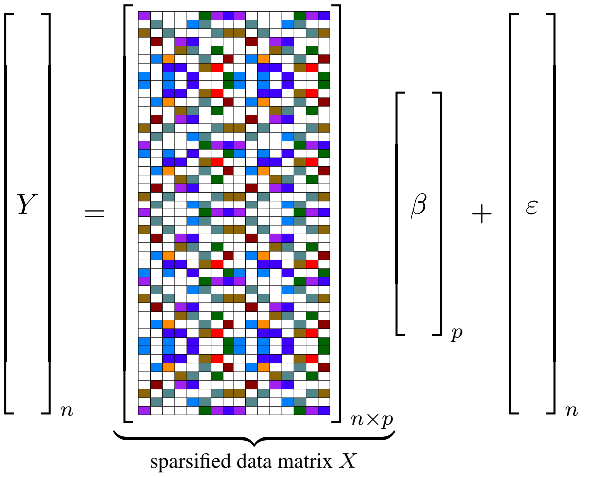

Figure 2. Drawing inspiration from the "dropout" in deep learning, we have begun to study data sparsification as a way or regulating the computation/risk tradeoff in large scale linear regression. We selectively zero out entries in the design matrix. This incurs error, but allows the estimator to be computed efficiently using subspace embedding methods developed recently in the theoretical computer science literature (Clarkson and Woodruff, 2012).

above, which enable fast algorithms for low rank approximation and least squares estimation from sparse data matrices. Our preliminary simulations with these algorithms suggest that when the data have fat tails, as in a t-distribution, sparsification followed by subspace embedding gives a favorable tradeoff compared to subsampling—which simply discards rows of the data matrix. We will study this further and obtain probability bounds on the error of the dropout method for a generalized ridge regression problem.

The key to subspace embedding algorithms lies in leverage scores. Leverage scores represent the relative importance of rows and columns. When rows and columns are subsampled according to the leverage scores, the effective sample size and/or dimension of the problem is reduced. This induces a computation/risk tradeoff, as recently studied by Ma et al. (2013). We will study the use of the dropout method—data sparsification—to allow fast approximate computation of the leverage scores, thereby improving the computation/risk tradeoff.

We will then apply this thinking and analysis to the setting where many of the variables are irrelevant. In this case, intuitively, we wish to drop out irrelevant variables at a high rate, and drop out important predictors at a low rate. Clearly, if the irrelevant variables were known, the data could be optimally processed by dropping out all of them completely. The challenge is to separate the relevant and irrelevant variables while controlling computation. Toward this goal, we will develop a version of the dropout inspired by our work on "the rodeo" (Lafferty and Wasserman, 2008). In particular, we will develop a dropout version of the alternating direction method of multipliers (ADMM) algorithm for approximate ℓ_1-regularized least squares (lasso) (Boyd et al., 2011). Each stage of the ADMM algorithm for the lasso involves a form of ridge regression. Leveraging our analysis of the dropout for this case, we gradually decrease the dropout rate for variables identified as relevant in an ADMM iteration.

Proposed Research B: Sparsified stochastic gradient descent. In a related direction, we will study dropping out entries from the data in stochastic gradient descent. Each step of stochastic gradient can be accelerated by dropping out components of the vector, resulting in a computation-risk tradeoff. This is similar to ideas of budgeted learning studied by Cesa-Bianchi et al. (2011) For sparsified stochastic gradient descent we will also develop lower bounds in the Nemirovsky-Yudin model (Nemirovsky and Yudin, 1983). In this model, an algorithm to minimize a convex function can make queries to a first-order oracle, and the complexity is defined as the minimum number of queries needed, in the worst case, to drive the error below some desired level. Specifically, the oracle is queried with an input point $x \in \mathcal{C}$ from a convex domain $\mathcal{C}$, and returns a pair $(f(x) + \varepsilon, v + \delta)$ where $v \in \partial f(x)$ is a subgradient vector to the function f at x, and ε and δ denote mean-zero stochastic noise. After T calls to the oracle, an algorithm $\mathcal{A}_T$ returns a value $X_{T+1} \in \mathcal{C}$, which is a random variable due to the stochastic nature of the oracle (and possibly also due to randomness in the algorithm). The Nemirovski-Yudin analysis reveals that, in the worst case, the number of calls to the oracle required to drive the expected error $\mathbb{E}(f(X_{T+1})) - \inf_{x \in \mathcal{C}} f(x)$ below ε scales as $T = O(1/\varepsilon)$ for the class of strongly convex functions, and as $T = O(1/\varepsilon^2)$ for the class of Lipschitz convex functions. When elements of the gradient are "dropped out," it should be possible to carry out a Nemirovski-Yudin style analysis to give larger lower bounds, quantifying the excess risk that comes from the reduced computation of the dropout.

Proposed Research C: Tradeoffs for nonparametric estimation. Kernel regression estimates the value of the regression function at a new point as a weighted average of nearby points. Naively, this requires n distance computations to first determine which points are nearby. A large literature exists on methods for both exact and approximate nearest neighbor search. Tree-based methods, such as kd-trees or spill-trees, build a data structure by partitioning the space, allowing depth-first search on the tree.

Another approach uses locality sensitive hashing. LSH uses hash functions that map nearby points to the same value with high probability, and map distant points to different values with high probability. The hash functions are used to design data structures that efficiently retrieve points that are approximately within a fixed distance h of a query point. The LSH parameters are independent of the ambient dimension, and the search time grows sublinearly in n. The preprocessing and storage requirements are sub-quadratic, and can be amortized over multiple queries.

We propose to develop nonparametric estimation algorithms using LSH to quickly find approximate near neighbors of a query point. A threshold M on the number of near neighbors returned acts as a tuning parameter for the risk-computation tradeoff. The algorithm sets the LSH parameters, increasing the size of the data structure with M so that the threshold is reached with high probability. The points returned by the LSH algorithm are then used to compute the kernel regression estimate at the query point. As M increases, the variance of the kernel regression estimate decreases, but the LSH data structure grows, increasing both the preprocessing and query time. This yields a fine-grained tradeoff between computation time and predictive risk. We will rigorously analyze the tradeoff between computation and statistical accuracy that such an algorithm provides. This approach can be adapted to many different nonparametric estimation problems, not only regression.

Proposed Research D: Upper bounds using statistical dimension with covariates. Recent work by Chandrasekaran and Jordan (2013) studies the relationship between predictive risk, sample size, and computational efficiency. This paper focuses on the normal means family of denoising problems. The main advance is a result that lower bounds the sample size sufficient to yield a specified level of risk for a given convex relaxation in terms of a geometric quantity of the convex set. Specifically, the sample size is bounded below by the Gaussian squared complexity of the tangent cone of the set at the target signal,

$$n \geq \sigma^2 g\left(T_{\mathcal{C}}(x^*) \cup B_{\ell_2}^p\right).$$

Here $g(\cdot)$ denotes the Gaussian squared complexity, and $T_{\mathcal{C}}(x^*)$ is the tangent cone of the convex constraint set $\mathcal{C}$; $B_{\ell_2}^p$ is the unit ball. This makes intuitive geometric sense. If the convex set is "pointy" at the true signal, then the approximation is relatively tight and the Gaussian complexity is small. Examples are given of tradeoffs obtained using this approach of convex relaxations: denoising signed matrices, ordering variables to get a banded covariance matrix, analyzing sparse principal components, and estimating matchings.

We will build on this work by considering problems of estimation with covariates, rather than denoising problems. We will thus construct "data-dependent liftings" where the geometry of the convex cone is dependent on the actual observed data. Intuitively, many of the constraints will be inactive for a given problem instance, and relaxing the inactive constraints will have no impact, even though they may affect a crude analysis of run time.

4. Estimation and Inference Under Shape Restrictions (Aim 3)

Background and Preliminary Results. Shape restrictions such as monotonicity, convexity, and concavity provide a natural way of limiting the complexity of many statistical estimation problems. Shape-constrained estimation is not as well understood as more traditional nonparametric estimation involving smoothness constraints. For instance, the minimax rate of convergence for multivariate convex regression has yet to be rigorously established in full generality, although covering and bracketing number bounds have been recently established (Kim and Samworth, 2014). Even the one-dimensional case is challenging, and has been of recent interest (Guntuboyina and Sen, 2013).

Estimation of convex functions arises naturally in several applications. Examples include geometric programming (Boyd and Vandenberghe, 2004), computed tomography (Prince and Willsky, 1990), target reconstruction (Lele et al., 1992), image analysis (Goldenshluger and Zeevi, 2006) and circuit design (Hannah and Dunson, 2012). Other applications include queuing theory (Chen and Yao, 2001) and economics, where it is of interest to estimate concave utility functions (Meyer and Pratt, 1968). Beyond cases where the assumption of convexity is natural, the convexity assumption can be attractive as a tractable, nonparametric relaxation of the linear model.

The convex regression problem is naturally formulated using finite dimensional convex optimization, with no tuning parameters for smoothness. We have recently begun to study the problem of variable selection in high dimensional multivariate convex regression. Assuming that the regression function is convex and sparse, our goal is to identify the relevant variables. We have shown that it suffices to estimate a sum of one-dimensional convex functions—an additive model—leading to significant computational and statistical advantages (Xu et al., 2014). This is in contrast to general nonparametric regression, where fitting an additive model can result in false negatives.

To briefly explain, the infinite-dimensional nonparametric convex regression $\min_{f \text{ convex}} \sum_i (y_i - f(x_i))^2$ is equivalent to the finite dimensional quadratic program $\min_{f,\beta} \sum_i (y_i - f_i)^2$ subject to the subgradient constraints $f_j \geq f_i + \beta_i^T (x_j - x_i)$. This optimization is subject to the statistical curse of dimensionality. To carry out scalable variable selection, we have devised a two-stage quadratic programming procedure. In the first stage, we fit a convex additive model, imposing a sparsity penalty. In the second stage, we fit a concave function on the residual for each variable. This non-intuitive second stage is in general necessary. We call this the AC/DC algorithm:

1. *AC Stage*: Estimate an additive convex model

$$\{\widehat{f}_k\}, \widehat{\mu} = \mathop{\arg\min}_{f_1,\ldots,f_p \in \mathcal{C}^1,\, \mu \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \mu - \sum_{k=1}^{p} f_k(x_{ik}) \Big)^2 + \lambda \sum_{k=1}^{p} \|f_k\|_\infty .$$

2. *DC Stage*: If $\|\widehat{f}_k\|_\infty = 0$, estimate a decoupled concave function

$$\widehat{g}_k = \mathop{\arg\min}_{g_k \in \text{-}\mathcal{C}^1} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \widehat{\mu} - \sum_{k'} \widehat{f}_{k'}(x_{ik'}) - g_k(x_{ik}) \Big)^2 + \lambda \|g_k\|_\infty .$$

3. Estimated support $\widehat{S}_n = \{k : \|\widehat{f}_k\|_\infty > 0 \text{ or } \|\widehat{g}_k\|_\infty > 0\}$.

We have proven that this procedure is faithful in the population setting, meaning that it results in no false negatives, under mild assumptions on the density of the covariates. Our second result is a finite sample statistical analysis of the procedure, where we upper bound the statistical rate of convergence. Specifically,

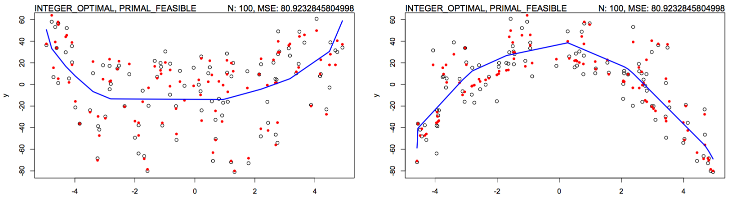

Figure 3. Simple illustration of convexity pattern estimation. We estimate a two-dimensional additive model, where each component is either convex or concave; there are 2^p possible convexity patterns.

"sparsistent" variable selection is achievable with sample complexity n satisfying $n^{4/5} \geq Cs^5\sigma^2 \log^2 p$ where s is the number of true relevant variables (Xu et al., 2014). The proof of this result exploits recent bounds on bracketing numbers for convex function classes due to Kim and Samworth (2014).

Proposed Research A: Convexity pattern estimation. Suppose we have an additive model with a sum of convex and concave functions. Then estimation is a quadratic program, with no smoothing parameters. But what if we don't know the ***convexity pattern***—which functions are convex and which are concave? Can it be learned? Solving this problem will lead to a new, powerful approach to nonparametric modeling.

More specifically, the model is

$$Y = \sum_{j=1}^{p} z_j\, f_j(x_j) + \varepsilon$$
$$z_j \in \{-1, 1\}, \quad f_j \text{ convex}$$

and our objective is to decode $z = (z_1, \ldots, z_p) \in \{-1, 1\}^p$ from observed data $\{(X_i, Y_i)\}_{i=1}^n$, $X_i \in \mathbb{R}^p$, $Y_i \in \mathbb{R}$. One approach is to formulate a mixed integer second-order cone program, minimizing the squared error for an additive model $\sum_j (f_j(x_j) + g_j(x_j))$ where f_j is convex and g_j is concave. We then impose second order cone constraints $\|f_j\| \leq z_j B$ and $\|g_j\| \leq w_j B$ with the constraint $z_j + w_j \leq 1$. When we impose the integer constraints $z_j, w_j \in \{0, 1\}$, this results in a mixed integer SOCP. We have experimented with this using the R package `Rmosek` to call the Mosek mixed integer SOCP code. The procedure works well; however, the run time will in general be exponential, and the lack of a duality theory for such optimization makes analysis difficult.

A better, convex approach is the following. Consider the optimization

$$\begin{aligned} \min_{f,g,\beta,\gamma,z,w} \quad & \sum_{i=1}^{n} \Big(Y_i - \sum_{j=1}^{p} (f_{ij} + g_{ij}) \Big)^2 \\ \text{subject to} \quad & \text{convexity constraints on } f_j \\ & \text{concavity constraints on } g_j \\ & \sum_{j=1}^{p} \left\{ \beta_{(n)j} - \beta_{(1)j} + \gamma_{(1)j} - \gamma_{(n)j} \right\} \leq L \end{aligned}$$

where $\beta_{(1)j}, \beta_{(n)j}, \gamma_{(1)j}, \gamma_{(n)j}$ are the first and last subgradient vectors of f_j and g_j. This can be thought of as a nonstandard type of lasso. This works very well in simulation. However, the analysis will require new advances beyond the standard primal-dual witness approach. We will fully develop the theory and algorithms for this approach.

Proposed Research B: Estimation of SOS-convex densities. Log-concavity is often an appropriate shape constraint for density estimation. A density $f(x)$ is said to be log-concave if $f(x) = \exp(-s(x))$ where $s(x)$ is a convex function. Cule et al. (2010) have studied nonparametric log-concave density estimation. They show that with probability one there exists a unique optimal solution $\widehat{f}_n$ of the nonparametric log-likelihood $\ell(f) = \sum_{i=1}^n \log f(X_i)$ and that $\log \widehat{f}_n$ is a piecewise affine "tent function" on the convex hull C_n of the data. Thus, $\widehat{f}_n$ is not in general smooth. Computing $\widehat{f}_n$ is formulated as a non-smooth convex optimization problem with is solved through Shor's r-algorithm. Unfortunately, this does not scale to dimensions much larger than five with current technology.

We will study a different approach, restricting our attention to the class of log-SOS-concave functions $f(x) = \exp(-s(x))$ where $s(x)$ is an SOS-convex polynomial (Lasserre, 2009). While checking a polynomial's convexity is in general strongly NP-hard, as shown by Ahmadi (2013), an algebraic sum of squares (SOS) is a sufficient condition for convexity that can be certified by solving a semidefinite program. We propose to exploit this construction to develop algorithms based on projected stochastic gradient descent, where an SDP is solved in each gradient step to project onto the cone of SOS-concave functions. This requires MCMC to estimate expectations. Since our algorithm is an iterative procedure employing stochastic gradient descent, we will leverage recent results on sequential sampling for log-concave densities (Narayanan and Rakhlin, 2013).

Proposed Research C: Utility function estimation. One natural application of convex regression is modeling utility functions in economics and marketing. Many human behaviors can be modeled as a consumer selecting one item from among a set of alternatives. Examples include buying a product on Amazon, choosing the bus or car for commuting (Ortuzar and Willumsen, 1994), deciding where to buy a house (Nechyba and Strauss, 1998), and even choosing where to commit a crime (Bernasco and Block, 2009). The discrete choice model (DCM) originated in econometrics (McFadden, 1974) as a general method to model such finite choice problems. The DCM measures the attractiveness of item i to consumer n by a utility function $f(\mathbf{x}_i, \mathbf{s}_n)$ where $\mathbf{x}_i, \mathbf{s}_n$ are feature vectors of the item and the consumer, respectively. The consumer is more likely to pick item i over the alternatives if the utility $f(\mathbf{x}_i, \mathbf{s}_n)$ is higher. The AI and machine learning communities have in recent years rediscovered the DCM as a form of *preference learning* (Fürnkranz and Hüllermeier, 2010; Chu and Ghahramani, 2005).

Our work on variable selection in shape-constrained regression can be applied to the DCM. This is pertinent since people tend to make decisions based on a few important cues or factors (Shah and Oppenheimer, 2008). Good variable selection methods can give insight into how consumers make choices. While estimation of a low dimensional concave utility function for the DCM is studied by Matzkin using parametric distributional assumptions (Matzkin, 1991), we are unaware of previous results on variable selection in the DCM in the high dimensional nonparametric context. We will also study how a *mixture* of utility functions can be estimated, specializing the model to groups of consumers. For this goal, we will investigate nonparametric versions of the method of moments for mixtures (Chaganty and Liang, 2013; Anandkumar et al., 2012). We are interacting with faculty and students in Chicago's Booth School of Business on initial research on these topics.

5. Estimation Under Physical Constraints (Aim 4)

Background and Preliminary Results. As we have discussed, traditional statistical methods and theory are oriented toward constraints on properties such as smoothness. The statistical, machine learning, and signal processing communities have more recently invested heavily in notions such as sparsity, manifold structure, and low rank assumptions. On the other hand, classical applied mathematics was driven by physical problems that can be tightly modeled by PDEs reflecting the underlying physical laws. A large gap remains between these traditions. In particular, we have only the most rudimentary understanding of how to

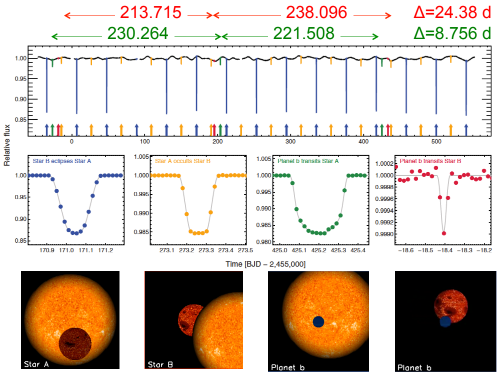

Figure 4. Geometry of planet transits in a lightcurve from the Kepler data (Fabrycky et al., 2012). The top panel shows the detrended lightcurve, normalized flux. The transit times for the planet and binary star are shown as color arrows at the bottom of the panel. The shape of planet/binary star transits are different, shown in the middle panel for the physical configuration shown in the bottom panel.

combine physical models with data driven approaches.

Indeed, this is precisely the perspective of Andrew Majda, one of the leaders of classical PDE approaches. As expressed in Majda and Harlim (2013),

> A central issue in contemporary science is the development of data driven statistical-dynamical models for times series of a partial subset of observed values $u_I(t) \in \mathbb{R}^{N_1}$, which arise from observations of nature or from an extremely complex physical model [. . .] purely data driven ad-hoc regression models are developed through various criteria to fit the data but by design, do not respect the underlying physical dynamics of the partially observed system or the causal processes in the dynamics.

In this paper, Majda and Harlim introduce a family of nonlinear regression models. In a simplified form, the models can be expressed in terms of the stochastic differential equation

$$du_t = (Lu_t + B(u_t, u_t) + f_t)\, dt + \sigma dW_t$$

where $u_t \in \mathbb{R}^N$ and f_t is a known forcing function, and dW_t is the increment of a Brownian motion. The key assumption is that the quadratic interactions $B(u_t, u_t)$ are subject to the *physical constraints of conservation of energy*, specified by

$$u_t \cdot B(u_t, u_t) = 0. \tag{3}$$

The conservation of energy constraint (3) is the distinguishing feature of the model; otherwise it is a standard quadratic regression model.

As another example of natural physical constraints, let us return to the problem of detection of exoplanets in lightcurves, considered in Aim 1. After detrending and computing residuals, the detection problem can be considered as a sparse normal means testing problem. Much is known about optimal testing of sparse normal means or simple geometric structures (Donoho and Jin, 2004, 2008; Arias-Castro et al., 2006). But how can we test for signals that are governed by known physical laws? For instance, the shape of binary star and planet transits is governed by simple physical and geometrical properties; see Figure 4.

Finally, as another example of statistical modeling with physical constraints, we mention recent work on the long-term goal of building smarter operating systems under power/performance constraints. This is joint work with University of Chicago Computer Science Professor Hank Hoffmann, merging our work in high dimensional statistics with a computer systems perspective.

The goal is to let the OS determine how to allocate power among multiple applications. For example, when running a coupled simulation, one simulation may be slower than another. Ideally, we would recognize that these two jobs should run at the same rate, and then shift power from the faster one to the slower one, reducing overall runtime while enforcing the power budget. The required nonlinear optimization is difficult, as different jobs and even processes within a job will have different resource needs and different power/performance tradeoffs. Acquiring this knowledge is complicated by the fact that these power/performance tradeoffs are often application or input dependent. Statistical techniques are needed to accurately estimate the application-dependent properties online.

We have developed a prototype system called LEO—Learning for Energy Optimization. We assume that there is some set of applications for which the power and performance tradeoffs are gathered offline. LEO uses that set of known applications to form prior beliefs about the probability distributions of the power and performance achievable in different system configurations. LEO then takes a small number of observations of the target application and uses a hierarchical Bayesian model to estimate the power and performance for that application in all the other configurations. For example, if LEO has previously seen an application that only scales to 8 cores, it can use that information to quickly determine if the current application will be limited in its scaling. Preliminary results are promising (**?**), and we are continuing to develop this approach.

Proposed Research A: Develop minimax theory for physics-constrained regression and detection. How can we understand the minimax complexity of physics-constrained regression? Apart from energy conservation, what are other tractable and physically meaningful constraints? In our work on quantized nonparametric regression (Zhu and Lafferty, 2015), we quantify the loss in the minimax rate of convergence due to quantization. Note that for classical minimax theory, the estimators are not required to adhere to the same smoothness constraints as the function class. For example, kernel estimators will typically not lie in the Sobolev space assumed of the true function. For physically constrained regression, it seems natural to impose the same constraints both on the function class and on the estimators—the world obeys the constraints, and our estimator should too. How does this effect the minimax performance? Can we quantify the affect of introducing various physical constraints into the problem? We expect that this will require new techniques to complement the usual suite of Fano methods.

Proposed Research B: Develop estimation and detection algorithms for physics-constrained regression. How can we construct efficient estimators and detectors that incorporate physical constraints? If a regression model is required to satisfy a physical invariant, how is that constraint imposed on the smoothing procedure? Or, how are physical constraints to be incorporated into a computationally efficient approximation to a likelihood ratio test?

For example, in the Kepler data, to test for planet transits the procedure used by physicists is a simple "windowing" technique whereby different possible transit periods and phases are matched against the detrended signal. This corresponds to a crude form of a scan statistic in a likelihood ratio test. As the testing problems become more ambitious, the physics becomes more complex. For instance, my colleague Prof. Dan Fabrycky at the University of Chicago is beginning to study the distribution of multiplanet systems. One approach is based on the radial velocity method, where one computes the correlation of planets of different masses. But the gravitational interactions of multiple planets on a star are complex, and the signal-to-noise ratio may be increased by incorporating physical constraints. Microlensing is another recent approach. We will consider these problems as testbeds for incorporating physical constraints into testing. An advantage of such problems is that accurate simulation studies can be carried out because the physics is fairly well understood.

Proposed Research C: Power-performance control in application runtime environments. We will also continue our work on modeling energy/performance tradeoffs for runtime systems. In this case, the challenges are to fold physical constraints into the regression models, and to incorporate the physics into a control algorithm that uses the predictions made by the regression models of power usage for different applications. The current decision making mechanism of LEO is a control problem that is set up as a sequential linear program. This allows the system to quickly match the behavior of the current application to a subset of the previously observed applications, and solve the control problem to dynamically select the optimal mixture of system configurations. Improvements are needed in both the control algorithm, and in the hierarchical models used to predict performance and energy usage. The predictions are made using a high dimensional vector of covariates that includes statistics of cache performance, attributes of the processors and memory usage, and many others. We expect that incorporating known physical relationships between system components can increase statistical efficiency.

6. Nonconvex estimation for large scale convex optimization (Aim 5)

Background and Preliminary Results. Optimization plays a central role in modern applied and theoretical statistics. Convex optimization is particularly important, as the uniqueness properties, duality theory, and KKT conditions can enable theoretical guarantees and the development of algorithms. But convex optimization is a relatively new introduction to the arsenal of statisticians. Even ten years ago, little expertise and understanding existed in the statistical and machine learning communities on how to approach the large scale quadratic programs that arise in the lasso and other constrained regression problems.

A flurry of work now regularly introduces more sophisticated optimization problems—both convex and nonconvex—into methodology for data analysis. Conic programming methods are particularly flexible and important. For example, semidefinite relaxations are useful for a range of problems from sparse PCA (d'Aspremont et al., 2004; Amini and Wainwright, 2009) to SOS-convexity for density estimation and regression, as we have indicated above in Aim 3. A "dirty secret" in this research is that while semidefinite programs (SDPs) are often advertised as offering practical algorithms because they have polynomial runtime guarantees, current optimization algorithms based on interior point methods can only handle relatively small problems. More scalable algorithms for semidefinite programming, and conic programming more generally, are needed.

A parallel development is the surprising effectiveness of simple classical procedures such as stochastic gradient descent for large scale problems, as explored in the recent machine learning literature (Bach and Moulines, 2011; Bach, 2014; Hoffman et al., 2013). We propose to leverage stochastic gradient descent procedures for solving large scale semidefinite programs.

Our motivation comes from recent work on phase recovery from randomized measurements (Candès et al., 2014). The problem is to reconstruct a complex signal $z \in \mathbb{C}^n$ from phaseless measurements $b_r^2 = |\langle a_r, z\rangle|^2$, for m random measurement vectors $a_r \in \mathbb{C}^n$, $r = 1 \ldots, m$. This leads to a nonconvex quadratic program. The physical motivation is the reconstruction of an object given diffraction patterns consisting of measured light without phase, collected using different filters. Candès et al. (2014) have shown how a variant on stochastic gradient descent leads to a remarkably effective algorithm for solving this nonconvex optimization that scales to large problems. Specifically, with the loss function $\ell(z) = \frac{1}{2m}\sum_{r=1}^m (b_r^2 - a_r^* z z^* a_r)^2$ the stochastic gradient descent algorithm ("Wirtinger flow") iteratively updates

$$z_{t+1} \longleftarrow z_t - \frac{\mu_t}{\|z_0\|^2}\nabla \ell(z_t).$$

If the step size μ_t is chosen appropriately and the initial "guess" z_0 is sufficiently good, theory and experiment confirm that this procedure successfully performs phase retrieval.

The SDP connection is based on the Goemans-Williamson SDP relaxation for max-cut (Goemans and Williamson, 1995), which results in an SDP relaxation of this nonconvex optimization (Fajwel Fogel and d'Aspremont, 2014; Candès et al., 2013).

Proposed Research A: Wirtinger flows for families of SDPs. We propose to turn this around, and use the idea behind Wirtinger flows for phase retrieval to solve—or approximately solve—large scale SDPs. The

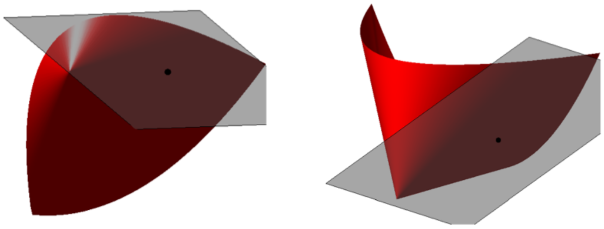

Figure 5. Illustration of the positive-semidefinite cone of 2×2 matrices (red) intersected with the affine space $\mathrm{Tr}(X) = 1$; visualization from Candès et al. (2013). The geometry of the intersection determines properties of approximation algorithms.

standard form SDP is

$$\begin{aligned} \min\ & \mathrm{Tr}(CX) \\ & \mathrm{Tr}(A_i X) = b_i,\ i = 1\ldots, m \\ & X \succeq 0. \end{aligned}$$

Under the assumption that $C \succ 0$ is positive semidefinite, a calculation shows that we can re-express this as an equivalent SDP in standard form with $C = I$. Suppose that the solution is expected to be of rank one. Then, motivated from Wirtinger flows, an attractive approach is stochastic gradient descent on the objective

$$\ell(x) = \frac{1}{2m} \sum_{i=1}^{m} \big(b_i - \mathrm{Tr}(A_i x x^T)\big)^2.$$

If a rank two solution holds, then this leads to the optimization of $\ell(x, y) = \frac{1}{2m} \sum_{i=1}^{m} \big(b_i - \mathrm{Tr}(A_i x x^T) - \mathrm{Tr}(A_i y y^T)\big)^2$ and so on. This suggests adopting a forward stagewise type of procedure, or the incorporation of a sparsity penalty—methods that have enjoyed great success in high dimensional statistics. Thus, we are attacking hard convex optimization problems with very scalable and flexible nonconvex regression techniques. We know from Candès et al. (2014) that this will work for certain families of SDPs. We will extensively experiment with and rigorously study the use of this technique for more general SDPs.

Proposed Research B: Application to SOS-convex density estimation and regression. Using the approach to solving large scale SDPs described above, we will return to one of the goals of Aim 3. Recall that for log-SOS-concave density estimation, we are required to solve an iterative sequence of SDPs, which scale according to both the dimension p and order d of the polynomials used. Using standard SDP solvers this does not scale to high dimensions. We will investigate the use of stochastic gradient descent using low rank approximations to the solution, as described above, in order to carry out approximate maximum likelihood log-SOS-concave density estimation and SOS-convex regression.

7. Broader Impact

Three aspects of the proposed research will have a broad impact, extending beyond the intellectual contribution of the targeted research. First, as science becomes more data-driven, we see an acute need for powerful nonparametric methods to handle the increasing complexity of modern datasets. Adequate methodology is lacking, particularly methodology that effectively incorporates realistic constraints from the problem at hand. We expect to develop new families of flexible and principled large scale data analysis tools that can benefit many scientific domains. Second, in our previous work on some of the foundational problems of machine learning and statistics, our research group and collaborators have developed software that has been widely distributed, allowing others to build on our work. A recent example is the `huge` R package for high-dimensional undirected graph estimation (Zhao et al., 2012). We will develop and make available software for all of the methods developed in this project. Third, we expect this research to have impact across multiple communities, not only the statistics and machine learning communities, but also in the domain sciences.

During the last three years we have developed new courses at the University of Chicago that have reflected our research in nonparametric statistics and large scale data analysis. One class is "Machine Learning and Large-Scale Data Analysis" which is taught at the advanced undergraduate and beginning graduate level. Students from many different departments enroll. This course includes both standard lectures and exercises, but also a series of four large scale data projects (LSD projects). One of the LSD projects applies semi-supervised learning to a dataset of 80 million images from Google searches. Another LSD project focuses on stochastic gradient descent for fitting a logistic regression model on a streaming Twitter feed. Another involves kernel smoothing for a Poisson regression model to predict crime rates from historical data in the city of Chicago for one week during the course. (These data are not sparse.) The last concerns exoplanet finding from the Kepler telescope data, which involves fitting nonparametric regressions, and running hypothesis tests on the residuals, for the light curves of roughly 150,000 stars. All of the computation is done in the cloud using a Python interface to Amazon AWS that was developed specifically for this course. The costs are covered by an Amazon Machine Learning in Education grant. A University of Chicago undergraduate developed this infrastructure—he is now employed full time by Amazon AWS in Seattle. A second course is "Nonparametric inference" taught at the advanced undergraduate and beginning graduate level. Several undergraduates have gotten interested in research in statistics and machine learning through these courses, and have gone on to graduate school in these areas. Undergraduates have also initiated research projects with the PI following these courses. We led an REU project during the summer of 2014 on the theme of "Learning Polynomials, Graphs and Densities" (http://theorycenter.cs.uchicago.edu/REU/2014/projects.php), which has motivated part of Aim 3. Another such project currently involves four undergraduates in a collaboration with Zillow, the real estate company. A new, advanced graduate course to be offered in the Winter 2015 quarter is "High dimensional statistics." In other activity that broadens the impact of such research, the PI recently co-organized a National Academy of Sciences workshop called "Training Students to Extract Value from Big Data" (http://www.nap.edu/catalog.php?record_id=18981). Our work on high dimensional statistics informs all of these outreach activities.

The broad range of topics in this research activity has the potential to appeal to a broad range of students. The PI currently supervises eight Ph.D. students, four of whom are women. The University of Chicago Department of Statistics is always looking for opportunities to increase the diversity of the pool of students involved in statistical research, including domestic students. Our efforts are aided by the University's commitment to increase the number of domestic students from under-represented groups who enroll in and complete Ph.D. programs—either at the University of Chicago or at other national universities. The University's *Leadership Alliance* program is one way of targeting this diversity. Several other new and existing activities in the division and at the University also contribute to an overall goal to support diverse students in their Ph.D. studies. These activities include the *Master of Science in the Physical Sciences Division* (MS-PSD), the *Multicultural Graduate Community* (a new registered student organization), a new *UChicago chapter with SACNAS* (Society for Advancement of Hispanics/Chicanos and Native Americans in Science), *CMEP Week* and *Discover UChicago* (both events for prospective graduate students). Collectively, these activities demonstrate that departmental efforts to recruit, prepare, and retain diverse students in the statistical sciences will be supported by wider divisional and University efforts.

References Cited

AGARWAL, A., BARTLETT, P. L. and DUCHI, J. C. (2012). Oracle inequalities for computationally adaptive model selection. arXiv:1208.0129.

AHMADI, A. A. (2013). *Algebraic Relaxations and Hardness Results in Polynomial Optimization and Lyapunov Analysis*. Ph.D. thesis, Massachusets Institute of Technology.

AMINI, A. A. and WAINWRIGHT, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics* **37** 2877–2921.

ANANDKUMAR, A., HSU, D. and KAKADE, S. (2012). Method of moments for mixture models and hidden Markov models. In *Conference on Learning Theory (COLT)*.

ARIAS-CASTRO, E., DONOHO, D. L. and HUO, X. (2006). Adaptive multiscale detection of filamentary structures in a background of uniform random points. *Ann. Statist.* **34** 326–349.

BACH, F. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research* **15** 595–627.

BACH, F. and MOULINES, E. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*.

BARRON, A. R. and JOSEPH, A. (2010). Toward fast reliable communication at rates near capacity with Gaussian noise. arXiv:1006.3870.

BERNASCO, W. and BLOCK, R. (2009). Where offenders choose to attack: A discrete choice model of robberies in Chicago. *Criminology* **47** 93–130.

BERTHET, Q. and RIGOLLET, P. (2012). Optimal detection of sparse principal components in high dimension. arXiv:1202.5070.

BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36** 199–227.

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3** 1–122.

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge University Press.

CANDÈS, E., LI, X. and SOLTANOLKOTABI, M. (2014). Phase retrieval via Wirtinger flows: Theory and algorithms. arXiv:1407.1065.

CANDÈS, E., STROHMER, T. and VORONINSKI, V. (2013). Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics* **66** 1241–1274.

CESA-BIANCHI, N., SHALEV-SHWARTZ, S. and SHAMIR, O. (2011). Efficient learning with partially observed attributes. *The Journal of Machine Learning Research* **12** 2857–2878.

CHAGANTY, A. T. and LIANG, P. (2013). Spectral experts for estimating mixtures of linear regressions. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (S. Dasgupta and D. McAllester, eds.), vol. 28-3. JMLR Workshop and Conference Proceedings.

CHANDRASEKARAN, V. and JORDAN, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *PNAS* **110**.

CHEN, H. and YAO, D. D. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer-Verlag.

CHU, W. and GHAHRAMANI, Z. (2005). Preference learning with Gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*. ACM.

CLARKSON, K. L. and WOODRUFF, D. P. (2012). Low rank approximation and regression in input sparsity time. arXiv:1207.6365.

CLARKSON, K. L. and WOODRUFF, D. P. (2013). Low rank approximation and regression in input sparsity time. Tech. rep., IBM Almaden Research Center. arXiv:1207.6365.

CULE, M., SAMWORTH, R. and STEWART, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density (with discussion). *J. R. Stat. Soc. Ser. B, Stat. Methodol.* **72** 545–600.

DASGUPTA, A., DRINEAS, P., HARB, B., KUMAR, R. and MAHONEY, M. W. (2009). Sampling algorithms and coresets for lp regression. *SIAM J. Computing* **38** 2060–2078.

D'ASPREMONT, A., GHAOUI, L. E., JORDAN, M. I. and LANCKRIET, G. (2004). A direct formulation for sparse PCA using semidefinite programming. In *In S. Thrun, L. Saul, and B. Schoelkopf (Eds.), Advances in Neural Information Processing Systems (NIPS) 16, 2004.*

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32** 962–994.

DONOHO, D. and JIN, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences* **105** 14790–14795.

DONOHO, D. L. (2000). Wald lecture I: Counting bits with Kolmogorov and Shannon. Wald Lectures.

DRINEAS, P., MAHONEY, M. W. and MUTHUKRISHNAN, S. (2006). Sampling algorithms for l_2 regression and applications. In *Proc. of the 17-th Annual SODA*.

FABRYCKY, D. C., FORD, E. B., STEFFEN, J. H., ROWE, J. F., CARTER, J. A., MOORHEAD, A. V., BATALHA, N. M., BORUCKI, W. J., BRYSON, S., BUCHHAVE, L. A., CHRISTIANSEN, J. L., CIARDI, D. R., COCHRAN, W. D., ENDL, M., FANELLI, M. N., FISCHER, D., FRESSIN, F., GEARY, J., HAAS, M. R., HALL, J. R., HOLMAN, M. J., JENKINS, J. M., KOCH, D. G., LATHAM, D. W., LI, J., LISSAUER, J. J., LUCAS, P., MARCY, G. W., MAZEH, T., MCCAULIFF, S., QUINN, S., RAGOZZINE, D., SASSELOV, D. and SHPORER, A. (2012). Transit timing observations from Kepler: IV. Confirmation of 4 multiple planet systems by simple physical models. Tech. rep., University of Chicago. arXiv:1201.5415.

FAJWEL FOGEL, I. W. and D'ASPREMONT, A. (2014). Phase retrieval for imaging problems. arXiv:1304.7735.

FOYGEL, R., HORRELL, M., DRTON, M. and LAFFERTY, J. (2012). Nonparametric reduced rank regression. In *Advances in Neural Information Processing Systems 25* (P. Bartlett, F. Pereira, C. Burges, L. Bottou and K. Weinberger, eds.). NIPS Foundation, 1637–1645.

FÜRNKRANZ, J. and HÜLLERMEIER, E. (2010). *Preference learning*. Springer.

GALLAGER, R. G. (1968). *Information Theory and Reliable Communication.* John Wiley & Sons.

GOEMANS, M. X. and WILLIAMSON, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM* **42** 1115–1145.

GOLDENSHLUGER, A. and ZEEVI, A. (2006). Recovering convex boundaries from blurred and noisy observations. *Ann. Statist* **34** 1375–1394.

GUNTUBOYINA, A. and SEN, B. (2013). Global risk bounds and adaptation in univariate convex regression. *arXiv:1305.1648* .

HANNAH, L. A. and DUNSON, D. B. (2012). Ensemble methods for convex regression with applications to geometric programming based circuit design. In *International Conference on Machine Learning (ICML).*

HASTIE, T. and TIBSHIRANI, R. (1999). *Generalized Additive Models.* Chapman and Hall. New York, NY.

HINTON, G. E., SRIVASTAVA, N., KRIZHEVSKY, A., SUTSKEVER, I. and SALAKHUTDINOV, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580.

HOFFMAN, M., BLEI, D. M., WANG, C. and PAISLEY, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research* **14**.

INGSTER, Y. and SUSLINA, I. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models.* Springer-Verlag. New York, NY.

INGSTER, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. I. *Mathematical Methods of Statistics* **2** 85–114.

INGSTER, Y. I. and SUSLINA, I. A. (2000). Minimax nonparametric hypothesis testing for ellipsoids and Besov bodies. *ESAIM P&S: Probability and Statistics* **4** 53–135.

JENKINS, J. M., CALDWELL, D. A., CHANDRASEKARAN, H., TWICKEN, J. D., BRYSON, S. T., QUINTANA, E. V., CLARKE, B. D., LI, J., ALLEN, C., TENENBAUM, P., WU, H., KLAUS, T. C., MIDDOUR, C. K., COTE, M. T., MCCAULIFF, S., GIROUARD, F. R., GUNTER, J. P., WOHLER, B., SOMMERS, J., HALL, J. R., UDDIN, K., WU, M. S., BHAVSAR, P. A., CLEVE, J. V., PLETCHER, D. L., DOTSON, J. A., HAAS, M. R., GILLILAND, R. L., KOCH, D. G. and BORUCKI, W. J. (2010). Overview of the Kepler science processing pipeline. *The Astrophysical Journal Letters* **713** L87.

JOHNSTONE, I. M. (2002). Function estimation and Gaussian sequence models. Unpublished manuscript.

JOHNSTONE, I. M. and LU, A. Y. (2004). Sparse principal components analysis. Tech. rep., Stanford University.

JOSEPH, A. and BARRON, A. R. (2012). Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity. *IEEE Trans. Info. Theory* **58** 2541–2557.

KIM, A. and SAMWORTH, R. (2014). Global rates of convergence in log-concave density estimation. arXiv:1404.2298.

KOUTIS, I., MILLER, G. and PENG, R. (2012). A nearly-$m \log n$ time solver for SDD linear systems. Tech. rep., Carnegie Mellon University. ArXiv:cs/1102.4842; FOCS 2011.

LAFFERTY, J. and WASSERMAN, L. (2008). Rodeo: Sparse, greedy nonparametric regression. *The Annals of Statistics* **36** 28–63.

LASSERRE, J. B. (2009). *Moments, Positive Polynomials and Their Applications*. World Scientific.

LELE, A. S., KULKARNI, S. R. and WILLSKY, A. S. (1992). Convex-polygon estimation from support-line measurements and applications to target reconstruction from laser-radar data. *Journal of the Optical Society of America, Series A* **9** 1693–1714.

MA, P., MAHONEY, M. and YU, B. (2013). A statistical perspective on algorithmic leveraging. arXiv:1304.5362.

MAJDA, A. J. and HARLIM, J. (2013). Physics constrained nonlinear regression models for time series. *Nonlinearity* **26**.

MATZKIN, R. L. (1991). Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica: Journal of the Econometric Society* 1315–1327.

MCFADDEN, D. (1974). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics* (P. Zarembka, ed.). Academic Press.

MEYER, R. F. and PRATT, J. W. (1968). The consistent assessment and fairing of preference functions. *IEEE Trans. Systems Sci. Cybernetics* **4** 270–278.

NARAYANAN, H. and RAKHLIN, A. (2013). Efficient sampling from time-varying log-concave distributions. arXiv:1309.5977.

NECHYBA, T. J. and STRAUSS, R. P. (1998). Community choice and local public services: A discrete choice approach. *Regional Science and Urban Economics* **28** 51–73.

NELSON, J. and NGUYEN, H. L. (2012). OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. arXiv:1211.1002.

NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons.

ORTUZAR, J. D. and WILLUMSEN, L. G. (1994). *Modelling transport*. Wiley.

PRINCE, J. L. and WILLSKY, A. S. (1990). Reconstructing convex sets from support line measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12** 377–389.

RASKUTTI, G. and MAHONEY, M. (2014). A statistical perspective on randomized sketching for ordinary least-squares. arXiv:1406.5986.

SHAH, A. K. and OPPENHEIMER, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological bulletin* **134** 207.

SHENDER, D. and LAFFERTY, J. (2013). Computation-risk tradeoffs for covariance-thresholded regression. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (S. Dasgupta and D. McAllester, eds.), vol. 28-3. JMLR Workshop and Conference Proceedings.

SPIELMAN, D. A. and TENG, S.-H. (2009). Nearly-linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. Tech. rep., Yale University. ArXiv:cs/0607105.

TSYBAKOV, A. B. (2008). *Introduction to Nonparametric Estimation*. 1st ed. Springer Series in Statistics.

VENKATARAMANAN, R., SARKAR, T. and TATIKONDA, S. (2013). Lossy compression via sparse linear regression: Computationally efficient encoding and decoding. In *IEEE International Symposium on Information Theory (ISIT)*. IEEE.

WAGER, S., WANG, S. and LIANG, P. (2013). Dropout training as adaptive regularization. arXiv:1307.1493.

WAN, L., ZEILER, M., ZHANG, S., CUN, Y. L. and FERGUS, R. (2013). Regularization of neural networks using dropconnect. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (S. Dasgupta and D. Mcallester, eds.). 3, JMLR Workshop and Conference Proceedings.

XU, M., CHEN, M. and LAFFERTY, J. (2014). Faithful variable screening for high-dimensional convex regression. arXiv:1411.1711; submitted to Annals of Statistics.

ZHANG, Y., DUCHI, J., JORDAN, M. and WAINWRIGHT, M. J. (2013). Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*.

ZHAO, T., LIU, H., ROEDER, K., LAFFERTY, J. and WASSERMAN, L. (2012). The `huge` package for high-dimensional undirected graph estimation in R. *The Journal of Machine Learning Research* **13** 1059–1062.

ZHU, Y. and LAFFERTY, J. (2015). Quantized nonparametric regression. arXiv:1503.07368; submitted to Annals of Statistics.