Precise Asymptotics in High-Dimensional Statistics using Random Matrix Theory and Statistical Physics

Yihan Zhang*

March 30, 2025

My research focuses on deriving precise asymptotics for problems in high-dimensional statistics with the aim of understanding their computational / statistical limits and potential gaps between them.

Consider the problem of estimation from a *generalized linear model* (GLM) $y = \phi(X\beta^\star)$ specified by a given nonlinearity[1] $\phi\colon \mathbb{R} \to \mathbb{R}$. Given a random design matrix $X \in \mathbb{R}^{n\times d}$ and the response $y \in \mathbb{R}^n$, the statistician seeks an estimate $\widehat{\beta} \equiv \widehat{\beta}(X, y) \in \mathbb{R}^d$ of the regression coefficients $\beta^\star \in \mathbb{R}^d$ that maximizes the asymptotic overlap, i.e.,

$$\mathcal{O} \coloneqq \lim_{d\to\infty} \sup_{\widehat{\beta}} \mathbb{E}\left[\frac{|\langle\widehat{\beta}, \beta^\star\rangle|^2}{\|\widehat{\beta}\|_2^2\|\beta^\star\|_2^2}\right] \in [0, 1].$$

Research over the last decade or so reveals that as $n, d \to \infty$, the solution to this problem undergoes a phase transition as $\delta \coloneqq \lim n/d$ varies. That is, if δ is at most a critical value $\delta^\star$, $\mathcal{O}$ is zero, indicating a complete failure in estimation due to a shortage of data; if δ exceeds $\delta^\star$, $\mathcal{O}$ becomes positive, indicating the possibility of nontrivial estimation. The abrupt change in estimation performance at $\delta^\star$ is reminiscent of the transformation between different states of a large particle system, e.g., water freezes at temperature 0°C, transitioning discontinuously from liquid to solid.

I am interested in characterizing, for various problems in high-dimensional statistics, the precise asymptotic value of fundamental quantities such as the overlap, generalization error, input-output mutual information, etc. An appealing feature of results of this type is that the characterization becomes increasingly accurate as the system size grows, distinguishing itself from many existing bounds in non-asymptotic statistics whose accuracy typically degrades with dimensions.

Deriving precise asymptotics requires insights and tools from recent advances in random matrix theory and statistical physics.

Random matrix theory. Modern machine learning practice operates on high-dimensional datasets in which the number of features is comparable to the number of samples. Random matrix theory offers a suite of powerful tools for assessing high-dimensional data through their spectral statistics including the distributions of singular values and singular vectors. In statistical inference, characterization of estimation / generalization error may require

- studying limiting spectral distribution of sum / product of "asymptotically free"[2] matrices,
- computing spherical integrals against the Haar measure over orthogonal matrices,
- concentrating the product of multiple correlated resolvent matrices, etc.

To these ends, input from random matrix theory is handy.

*University of Bristol. Email: zephyr.z798@gmail.com.

[1] By notational convention, the function ϕ is applied component-wise to its vector argument.

[2] Informally, *asymptotic freeness* of random matrices is akin to *independence* of random variables.

Statistical physics. Recent research reveals intimate connections between precise asymptotics in high-dimensional statistics and mean-field approximation for spin glasses — a certain disordered system extensively studied in statistical physics. A useful tool that emerges from these connections is an abstract family of algorithms known as *approximate message passing* (AMP). AMP formally resembles nonlinear power iteration in numerical linear algebra, with the crucial distinction of an additional one-step memory term that finds its root in Onsager correction in Thouless–Anderson–Palmer equation for mean-field spin glasses. An important utility associated with AMP is the so-called state evolution that precisely characterizes the empirical distribution of the iterates in the high-dimensional limit. Again, this mirrors the replica saddle point equation in spin glasses. Besides being an efficient algorithm that finds numerous applications and is conjectured optimal among a large family of algorithms, AMP can also be employed as a proof technique for analyzing problems with random data.

I seek motivated students with backgrounds in statistics / mathematics / theoretical physics / computer science to work on the following aspects of the research program outlined above.

Beyond i.i.d. Gaussian data. Existing theory of precise asymptotics is largely confined to data that are entry-wise i.i.d. and/or normally distributed. A significant portion of my current research addresses problems involving *orthogonally invariant* data. In the prototypical GLM, this means that the distribution of the design matrix X remains unchanged under conjugation of Haar orthogonal matrices, thereby modelling generic singular spaces yet completely general singular value spectrum.

Computational limits. In high dimensions, a perfect understanding of *statistical* optimality does not transfer to that of *computational* optimality. Models with structures such as sparsity can exhibit an intriguing phase where the problem is information theoretically solvable but not (believed to be) computationally efficiently so. The frameworks of *statistical query model* and *low-degree polynomial* propose to probe the computational power of general polynomial-time algorithms through specific families of abstract algorithms. It is of great interest to derive formal evidence to computational bottlenecks that precisely match the asymptotic performance of best-known efficient algorithms.

Debiasing in the inconsistency regime. As the saying goes, "all models are wrong". Practically, our statistical model may not be exactly specified and may contain unknown nuisance parameters. Consider again the example of GLM and suppose that X has i.i.d. rows from $\mathcal{N}(0_d, \Sigma)$ (a.k.a. correlated Gaussian design). The covariance matrix Σ here cannot be estimated consistently, if unknown (hence the name "inconsistency regime"). Though Σ is not of interest for estimation, missing its knowledge can cause bias to popular estimation procedures such as LASSO. Developing debiasing procedures and understanding optimality guarantees in the inconsistency regime constitute another project.