M-Estimation under High-Dimensional Asymptotics

DLD, Andrea Montanari

2014-05-01
ROBUST ESTIMATION OF A LOCATION PARAMETER

By Peter J. Huber

University of California, Berkeley

1. Introduction and summary. This paper contains a new approach toward a theory of robust estimation; it treats in detail the asymptotic theory of estimating a location parameter for contaminated normal distributions, and exhibits estimators—intermediaries between sample mean and sample median—that are asymptotically most robust (in a sense to be specified) among all translation invariant estimators. For the general background, see Tukey (1960) (p. 448 ff.).

Annals of Mathematical Statistics 1964

* Richard Olshen
M-estimation Basics

Location model

\[ Y_i = \theta + Z_i, \quad i = 1, \ldots, n \]

Errors: \( Z_i \sim F \), not necessarily Gaussian.
“Loss” Function \( \rho(t) \) eg \( t^2, |t|, -\log(f(t)), \ldots \)

\[
(M) \quad \min_{\theta} \sum_{i=1}^{n} \rho(Y_i - \theta)
\]

Asymptotic Distribution

\[
\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow_D N(0, V), \quad n \to \infty.
\]

Asymptotic Variance: \( \psi = \rho' \):

\[
V(\psi, F) = \frac{\int \psi^2 \, dF}{(\int \psi' \, dF)^2}
\]

Information Bound

\[
V(\psi, F) \geq \frac{1}{I(F)}
\]
One-Step Huber Estimates in the Linear Model

P. J. BICKEL*

Simple "one-step" versions of Huber's (M) estimates for the linear model are introduced. Some relevant Monte Carlo results obtained in the Princeton project [1] are singled out and discussed. The large sample behavior of these procedures is examined under very mild regularity conditions.

1. INTRODUCTION

In 1964 Huber [7] introduced a class of estimates (referred to as (M)) in the location problem, studied their asymptotic behavior and identified robust members of the group. These procedures are the solutions \( \hat{\theta} \) of equations of the form,

\[
\sum_{i=1}^{n} \psi(X_i - \hat{\theta}) = 0 ,
\]

(1.1)

where \( X_1 = \theta + E_1, \ldots, X_n = \theta + E_n \) and \( E_1, \ldots, E_n \) are unknown independent, identically distributed errors which have a distribution \( F \) which is symmetric about 0. If \( F \) has a density \( f \) which is smooth and if \( f \) is known, then maximum likelihood estimates if they exist satisfy (1.1) with \( \psi = -f'/f \).

Under successively milder regularity conditions on \( \psi \) and \( F \), Huber showed in [7] and [8] that such \( \hat{\theta} \) were equivalence holds in the more general context of the linear model for general \( \psi \).

Typically the estimates obtained from (1.1) are not scale equivariant.1 To obtain acceptable procedures a scale equivariant and location invariant estimate of scale \( \sigma \) must be calculated from the data and \( \hat{\theta} \) be obtained as the solution of

\[
\sum_{j=1}^{n} \psi_2(X_j - \hat{\theta}) = 0 ,
\]

(1.4)

where

\[
\psi_2(x) = \psi(x/\sigma) .
\]

(1.5)

The resulting \( \hat{\theta} \) is then both location and scale equivariant. The estimate \( \hat{\sigma} \) can be obtained simultaneously with \( \hat{\theta} \) by solving a system of equations such as those of Huber's Proposal 2 [8, p. 96] or the "likelihood equations"

\[
\sum_{j=1}^{n} \psi\left(\frac{X_j - \hat{\theta}}{\hat{\sigma}}\right) = 0 ,
\]

(1.6)

\[
\sum_{j=1}^{n} x\left(\frac{X_j - \hat{\theta}}{\hat{\sigma}}\right) = 0 ,
\]
Regression M-estimation: the One-Step Viewpoint

Regression model

\[ Y_i = X_i' \theta + Z_i, \quad Z_i \sim_{iid} F, \ i = 1, \ldots, n \]

Objective function of (M):

\[ R(\vartheta) = \sum_{i=1}^{n} \rho(Y_i - X_i' \vartheta) \]

\[ (M) \min_{\vartheta} R(\vartheta) \]

One-step estimate: \( \tilde{\theta}_n \) any \( \sqrt{n} \)-consistent estimate of \( \theta \):

\[ \hat{\theta}^1 = \tilde{\theta}_n - [\text{Hess} \ R|_{\tilde{\theta}_n}]^{-1} \nabla R|_{\tilde{\theta}_n}. \]

Effectiveness: \( \hat{\theta} \) true solution of M-equation:

\[ E(\hat{\theta}^1 - \hat{\theta})(\hat{\theta}^1 - \hat{\theta})' = o(n^{-1}) \]
The M-estimate is asymptotically equivalent to a single step of Newton’s method for finding a zero of $\nabla R$ starting at the true underlying parameter.

Goes back to Fisher, ‘Method of Scoring’ for MLE.
Derivation of Asymptotic Variance Formula

Approximation to One-Step:

$$\hat{\theta}^1 = \theta + \frac{1}{B(\psi, F)}(X'X)^{-1}X' (\psi(Z_i)) + o_p(n^{-1/2})$$

where $B(\psi, F) = \int \psi' dF$. Observe that

$$\text{Var}((X'X)^{-1}X' (\psi(Z_i))) \sim (X'X)^{-1}A(\psi, F)$$

where $A(\psi, F) = \int \psi^2 dF$. Hence if $X_{i,j} \sim N(0, \frac{1}{n})$

$$\text{Var}(\hat{\theta}_i - \theta_i) \rightarrow \frac{A(\psi, F)}{B(\psi, F)^2} = V(\psi, F)$$
THE 1972 WALD MEMORIAL LECTURES

ROBUST REGRESSION: ASYMPTOTICS, CONJECTURES AND MONTE CARLO

BY PETER J. HUBER
Swiss Federal Institute of Technology, Zürich

Maximum likelihood type robust estimates of regression are defined and their asymptotic properties are investigated both theoretically and empirically. Perhaps the most important new feature is that the number $p$ of parameters is allowed to increase with the number $n$ of observations. The initial terms of a formal power series expansion (essentially in powers of $p/n$) show an excellent agreement with Monte Carlo results, in most cases down to 4 observations per parameter.
We intend to build an asymptotic theory for $n \to \infty$; but there are several possibilities for the concomitant behavior of $p$. In particular, with decreasing restrictiveness:

(a) $\lim \sup p < \infty$
(b) $\lim p^2/n = 0$
(c) $\lim p^2/n = 0$
(d) $\lim p/n = 0$
(e) $\lim \sup p/n < 1$
(f) $\lim n - p = \infty$.

Case (a) has been treated by Relles (1968). The generalization to case (b) is relatively straightforward. Cases (d) and (e), possibly also (f) seem to be the interesting ones for the practical applications. I may quote a crystallographer's recommendation that there should be at least 5 observations per parameter, i.e., $p/n \leq 0.2$ (Hamilton (1970)). It will become clear in the next section why (e) and (f) are unlikely to yield to a reasonably simple asymptotic theory; then we shall attack what are, essentially, cases (b) to (d). Theoretical results and conjectures are summarized near the end of Sections 3 and 6; Monte Carlo results are summarized in Section 9.

ON THE DISTRIBUTION OF THE RESIDUALS FROM
A FITTED LINEAR MODEL

by

Peter Bloomfield

Technical Report 56, Series 2
Department of Statistics
Princeton University
January 1974
However, the expected value of this is the average of the marginal distributions, and is thus approximately

$$G(x) = F(x) + p/n\left\{ \frac{\psi(x)}{E(\psi')} f(x) + \frac{\text{var} \psi}{2E(\psi')^2} f'(x) - \theta f(x) \right\} \quad (2.5)$$

Study of $G$ thus shows what we should see in the empirical distribution function in the long run.
On robust regression with high-dimensional predictors

Noureddine El Karoui\textsuperscript{a,1}, Derek Bean\textsuperscript{a}, Peter J. Bickel\textsuperscript{a,1}, Chinghway Lim\textsuperscript{b}, and Bin Yu\textsuperscript{a}

\textsuperscript{a}Department of Statistics, University of California, Berkeley, CA 94720; and \textsuperscript{b}Department of Statistics and Applied Probability, Faculty of Science, National University of Singapore, 119077

Contributed by Peter J. Bickel, April 25, 2013 (sent for review March 1, 2012)

We study regression M-estimates in the setting where $p$, the number of covariates, and $n$, the number of observations, are both large, but $p \leq n$. We find an exact stochastic representation for the distribution of $\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^n \rho(Y_i - X_i^T \beta)$ at fixed $p$ and $n$ under various assumptions on the objective function $\rho$ and our statistical model. A scalar random variable whose deterministic limit $r_p(\kappa)$ can be studied when $p/n \to \kappa > 0$ plays a central role in this representation. We discover a nonlinear system of two deterministic equations that characterizes $r_p(\kappa)$. Interestingly, the system shows that $r_p(\kappa)$ depends on $\rho$ through proximal mappings of $\rho$ as well as various aspects of the statistical model underlying our study. Several surprising results emerge. In particular, we show that, when $p/n$ is large enough, least squares becomes preferable to least absolute deviations for double-exponential errors.

simulations. (While the paper was under review, we have managed to obtain rigorous proofs for many of our assertions. They will be presented elsewhere because they are very long and technical.) We give several results for covariates that are Gaussian or derived from Gaussian but present grounds that the behavior holds much more generally—the key being concentration of certain quadratic forms involving the vectors of covariates. We also investigate the sensitivity of our results to the geometry of the design matrix. [Further results with different designs can be found in our work (5).]

We find that (i) estimates of coordinates and contrasts that have coefficients independent of the observed covariates continue to be unbiased and asymptotically normal; and (ii) as in the fixed $p$ case, this happens at scale $n^{-1/2}$, at least when the minimal and maximal eigenvalues of the covariance of the predictors stay bounded away from 0 and $\infty$, respectively.*

These findings are obtained by (i) using leave-one-out per-
M-estimation
Our Paper
Isometry Between (M)-estimation & Lasso

Classical M-estimation
Big Data M-estimation

Bickel, Yu, El Karoui
High-Dimensional Asymptotics (HDA)

- \( n, p_n \to \infty \).
- \( X_{i,j} \sim iid \ N(0, \frac{1}{n}) \)
- \( p_n^{-1} \| \theta_{0,n} \|_2^2 \to \tau_0^2 \).
- \( Y = X \theta_0 + Z \)
- \( n/p_n \to \delta \in (1, \infty) \).
- Meaning of \( \delta \): “\# of observations per parameter to be estimated”
Emergent Phenomena under HDA

- Classical setting - random design, $p$ fixed, $n \to \infty$.

$$\text{var}(\hat{\theta}_i) \to V(\psi, F), \quad n \to \infty.$$ 

- HDA setting - for $n/p_n \to \delta \in (1, \infty)$
  - Effective Score $\tilde{\Psi} = \tilde{\Psi}_{\delta, \psi, F}$ (to be described... )
  - Effective Error Distribution

$$\tilde{F} = F * N(0, \tau_{\infty}^2)$$

Extra Gaussian noise: $\tau_{\infty} = \tau_{\infty}(\delta, \psi, F)$.

- Asymptotic Variance under HDA

$$\text{var}(\hat{\theta}_i) \to V(\tilde{\psi}, \tilde{F}), \quad n, p_n \to \infty.$$ 

- Classical Correspondence

$$\tilde{\Psi}_{\delta, \psi, F} \to \psi, \quad \tilde{F}_{\delta, \psi, F} \to F, \quad \delta \to \infty.$$
Immediate Implications

1. Classical formulas for confidence statements about M-estimates are \textit{overoptimistic} under high dimensional asymptotics (even dramatically so).

2. Maximum likelihood estimates are \textit{inefficient} under high-dimensional asymptotics.
   Score $\psi = (-\log f_W)'$ does not yield an efficient estimator.

3. The usual Fisher Information bound is \textit{not attainable}, as $I(\hat{F}) < I(F)$. 
Sample implication of DLD & Montanari (2013)

**Corollary.** For an M-estimator under HDA with errors $Z_i \ iid \ F$:

$$\lim_{n \to \infty} \text{Ave}_i \text{Var}(\hat{\theta}_i) \geq \frac{1}{1 - 1/\delta} \cdot \frac{1}{I(F)}$$

- Effect of HDA parameter $\delta \in (1, \infty)$ evident
M-estimation

Our Paper

Isometry Between (M)-estimation & Lasso

AMP Algorithm

State Evolution

Correctness of State Evolution

Convergence of AMP to $\hat{\theta}$
Regularized $\rho$

**Definition**

$\rho : \mathbb{R} \rightarrow \mathbb{R}$ is *smooth* if $C^1$ with absolutely continuous derivative $\psi = \rho'$ having a bounded a.e. derivative $\psi'$.

- Excludes $\ell_1$: $\rho_{\ell_1}(z) = |z|$.
- Allows Huber: $\rho_H(z) = \min(z^2/2, \lambda|z| - \lambda^2/2)$

**Regularized $\rho$:**

$$\rho_b(z) \equiv \min_{x \in \mathbb{R}} \left\{ b\rho(x) + \frac{1}{2}(x - z)^2 \right\}, \quad (1)$$

Min-convolution of $\rho$ with a square loss.
Regularized Score Function

\[ \Psi(z; b) = \rho'_b(z). \]

Example: Huber loss \( \rho_H(z; \lambda) \), with score function
\[ \psi(z; \lambda) = \min(\max(-\lambda, z), \lambda). \]

\[ \Psi(z; b) = b\psi\left(\frac{z}{1 + b}; \lambda\right). \]

\( \Psi \) ‘like’ \( \psi \), but central slope \( \|\Psi'(\cdot; b)\|_\infty = \frac{b}{1+b} < 1 \).

Effective score: for special choice \( b_* \),
\[ \tilde{\Psi} = \Psi(z; b_*) \].
Approximate Message Passing Algorithm

Initialization $\hat{\theta}^0 = 0$.

Adjusted residuals.

$$R^t = Y - X\hat{\theta}^t + \Psi(R^{t-1}; b_{t-1});$$ \hspace{1cm} (2)

Effective Score. Choose $b_t > 0$ achieving empirical average slope $p/n \in (0, 1)$.

$$\frac{p}{n} = \frac{1}{n} \sum_{i=1}^{n} \Psi'(R^t_i; b).$$ \hspace{1cm} (3)

Scoring. Apply effective score function $\Psi(R^t; b_t)$:

$$\hat{\theta}^{t+1} = \hat{\theta}^t + \delta X^T \Psi(R^t; b_t).$$ \hspace{1cm} (4)
Determining Regularization Parameter $b_t$

Example: $\rho = \rho_H(\cdot; 3)$, $F = 0.95N(0, 1) + 0.05\delta_{10}$. 

![Graph of Determination of $b^3$ and History of $b_t$](image)
Connection of AMP to M-estimation

Lemma
Let \((\hat{\theta}_*, R_*, b_*)\) be a fixed point of the AMP iteration (2), (3), (4) having \(b_* > 0\). Then \(\hat{\theta}_*\) is a minimizer of the problem \((M)\). Viceversa, any minimizer \(\hat{\theta}_*\) of the problem \((M)\) corresponds to an AMP fixed point of the form \((\hat{\theta}_*, R_*, b_*)\).
Example:

Size: $n = 1000$, $p = 200$, so $\delta = 5$. 

Truth: $\theta_0$ random $\|\theta_0\|_2 = 6\sqrt{p}$. 

Errors: $F = \text{CN}(0.05, 10)$, i.e. $F = 0.95\Phi + 0.05H_{10}$ 

$H_x$ denotes unit mass at $x$. 

Loss: $\rho = \rho_H(z; \lambda)$ with $\lambda = 3$. 

Iterations: Run Amp for 20 iterations. 

Comparison: Use CVX to obtain Huber estimator directly.
Convergence of AMP in Example

Convergence of AMP $\hat{\theta}^t$ to $\theta_0$

Convergence of AMP $\hat{\theta}^t$ to $\hat{\theta}$
Contrast AMP w/ Traditional Scoring

Traditional residual: \( Z = Y - X'\tilde{\theta} \)

Traditional Scoring:

\[
\hat{\theta}^1 = \tilde{\theta} + \frac{1}{\frac{1}{n} \sum_{i=1}^{n} \psi'(Z_i)} (X'X)^{-1} X'(\psi(Z_i)) , \tag{5}
\]

AMP:

\[
\hat{\theta}^{t+1} = \hat{\theta}^t + \delta n^{-1} \cdot X' \psi(R_t^t; b_t) .
\]

<table>
<thead>
<tr>
<th>Object</th>
<th>Scoring</th>
<th>AMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Slope</td>
<td>Traditional ((Z_i))</td>
<td>[\sum_{i=1}^{n} \psi'(R_i^t; b_t) / n ]</td>
</tr>
</tbody>
</table>
| Gram         | \((X'X)^{-1}\)       | \[\delta^{-1} \]
| Residual     | True \(\theta\)      | \(1\)                          |
| Score        | \(\psi(\cdot)\)     | Adjusted \((R_i^t)\)           |
| Basepoint    | 1                     | \(\psi(\cdot; b_t)\)           |
| Iteration    | \(t \geq 1\)         | Current Iterate               |
Basic contrasts between HDA and Classical Asymptotics

Under the High-Dimensional Asymptotic, $\delta \in (1, \infty)$,

*The heuristic “expand around $\theta$” is incorrect; it understates the true uncertainty.*

*The heuristic “errors equal the residuals” is incorrect; the residuals are noisier.*

*The heuristic “take one step” is incorrect; must take many steps.*
Extra Variance at Initialization of AMP

- Initialize with $\hat{\theta}^0 = 0$, $R^{-1} = 0$.
- Initial residual $R^1 = Y - X\hat{\theta}^0 = Z + X(\theta_0 - \hat{\theta}^0)$.
- Terms $Z$ and $X(\theta_0 - \hat{\theta}^0)$ are independent.
- $X(\theta_0 - \hat{\theta}^0)$ is Gaussian with variance $\tau_0^2 = \|\hat{\theta}^0 - \theta_0\|^2_2/n = \text{MSE}(\hat{\theta}^0, \theta_0)/\delta$.

$$\text{Var}(R_i^1) = \text{Var}(Z) + \text{Var}(X(\theta_0 - \hat{\theta}^0)) = \text{Var}(Z) + \text{MSE}(\hat{\theta}^0, \theta_0)/\delta.$$  
- Extra Gaussian noise $\tau_0^2 = \text{MSE}(\hat{\theta}^0, \theta_0)/\delta$. 
Evolution of Extra Variance at later AMP iterations

Pretend $R^t \approx Z + \tau_t \cdot W$ with $W$ an independent standard normal

Define variance map

$$\mathcal{V}(\tau^2, b; \delta, F) = \delta \mathbb{E}\left\{ \psi(Z + \tau \cdot W; b)^2 \right\},$$

Define slope par. $b = b(\tau; \delta, F)$: smallest solution $b \geq 0$ to

$$\frac{1}{\delta} = \mathbb{E}\left\{ \psi'(Z + \tau \cdot W; b) \right\}$$ (6)

Definition

State Evolution dynamical system $\{\tau^2_t\}_{t \geq 0}$, starting at $\tau^2_0 \in \mathbb{R}_{\geq 0}$ by

$$\tau^2_{t+1} = \mathcal{V}(\tau^2_t, b(\tau_t)) = \mathcal{V}(\tau_t, b(\tau_t; \delta, F); \delta, F).$$ (7)
Variance Map, in running example

\[ F = 0.95N(0, 1) + 0.05H_{10}, \rho_H \text{ with } \lambda = 3. \]
State Evolution, in running example

\[ F = 0.95N(0, 1) + 0.05H_{10}, \rho_H \text{ with } \lambda = 3. \]
Operating Characteristics from State Evolution

Definition

The *state evolution formalism* assigns predictions $E$ under states $S = (\tau, b, \delta, F)$

Functions $\xi$ of $\hat{\theta} - \theta_0$. Predict $p^{-1} \sum_{i \in p} \xi(\hat{\theta}_i - \theta_0, i)$ by

$$E(\xi(\hat{\theta} - \vartheta)|S) \equiv \mathbb{E}\{\xi(\sqrt{\delta} \tau Z)\},$$

Functions $\xi_2$ of Residual, Error. Predict $n^{-1} \sum_{i=1}^{n} \xi_2(R_i, Z_i)$ by

$$E(\xi_2(R, Z)|S) \equiv \mathbb{E}\xi_2(Z + \tau W, Z)$$

where $W \sim \mathcal{N}(0, 1)$ and $Z \sim F$ is independent of $W$. 

Example of State Evolution Predictions

- **MSE at iteration $t$.** We let $S_t = (\tau_t, b(\tau_t), \delta, F)$ denote the state of AMP at iteration $t$, and predict

$$\text{MSE}(\hat{\theta}^t, \theta_0) \approx \mathcal{E}((\hat{\vartheta} - \vartheta)^2 | S_t) = \mathbb{E}\left\{ (\sqrt{\delta} \tau_t Z)^2 \right\} = \delta \tau_t^2.$$  

- **MSE at convergence.** With $\tau_* > 0$ the limit of $\tau_t$, let $S_* = (\tau_*, b(\tau_*), \delta, F)$ denote the equilibrium state of AMP

$$\text{MSE}(\hat{\theta}_*, \theta_0) \approx \mathcal{E}((\hat{\vartheta} - \vartheta)^2 | S_*) = \mathbb{E}\left\{ (\sqrt{\delta} \tau_* Z)^2 \right\} = \delta \tau_*^2.$$  

- **Ordinary residuals $Y - X\hat{\theta}^*$ at AMP convergence.** Setting $\eta(z; b) = z - \Psi(z; b)$

$$Y - X\hat{\theta}^* \Rightarrow_D \eta(Z + \tau_* W; b_*).$$
**Figure**: Experimental means from 10 simulations compared with State Evolution predictions under $CN(0.05, 10)$, with Huber $\psi$, $\lambda = 3$. Upper Left: $\hat{\tau}_t = \|\hat{\theta}_t - \theta_0\|_2 / \sqrt{n}$. Upper Right: $\hat{b}_t$. Lower Left: MSE, Mean Squared Error. Lower Right: MAE, Mean Absolute Error. Blue ‘+’ symbols: Empirical means of AMP observables. Green Curve: Theoretical predictions by SE.
Lower Bounds on State Evolution

**Lemma.** Suppose that $F$ has a well-defined Fisher information $I(F)$. Then for any $t > 0$

$$
\tau_t^2 \geq \frac{1}{\delta I(F)}.
$$

**Lemma.** (uses Barron-Madiman, 2007)

$$
I(F \ast N(0, \tau^2)) \leq \frac{I(F)}{1 + \tau^2 I(F)}.
$$

**Corollary.** For $t > k$

$$
\tau_t^2 \geq \frac{1 + \frac{1}{\delta} + \frac{1}{\delta^2} + \cdots + \frac{1}{\delta^k}}{\delta I(F)}.
$$

**Corollary.** For every accumulation point $\tau_*$ of State Evolution

$$
\tau_*^2 \geq \frac{1}{\delta - 1} \cdot \frac{1}{I(F)}.
$$

**Corollary.** For an M-estimator under HDA with errors $Z_i \ iid \ F$:

$$
\lim_{n \to \infty} \text{Var}(\hat{\theta}_i) \geq \frac{1}{1 - 1/\delta} \cdot \frac{1}{I(F)}.
$$
Correctness of State Evolution 1

Basic Assumptions:

A1 Discrepancy function $\rho$ is convex and smooth;
A2 Matrices $\{X(n)\}_n$ are $\sim_{iid} N(0, \frac{1}{n})$
A3 $\theta_0, \hat{\theta}^0 = 0$ are deterministic sequences such that $\text{AMSE}(\theta_0, \hat{\theta}^0) = \delta \tau_0^2$.
A4 $F$ has finite second moment.

Terminology:

- Let $\{\tau_t^2\}_{t \geq 0}$ denote the state evolution sequence with initial condition $\tau_0^2$.
- Let $\{\hat{\theta}^t, R^t\}_{t \geq 0}$ be the AMP trajectory with parameters $b_t$.

Definition: A function $\xi: \mathbb{R}^k \rightarrow \mathbb{R}$ is pseudo-Lipschitz if there exists $L < \infty$ such that, for all $x, y \in \mathbb{R}^k$, $|\xi(x) - \xi(y)| \leq L(1 + \|x\|_2 + \|y\|_2) \|x - y\|_2$. 
Correctness of State Evolution 2


Let \( \xi : \mathbb{R} \rightarrow \mathbb{R} \), \( \xi_2 : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \) be pseudo-Lipschitz functions. Then, for any \( t > 0 \), we have, for \( W \sim \mathcal{N}(0, 1) \) independent of \( Z \sim F \)

\[
\lim_{n \to \infty} \frac{1}{p} \sum_{i=1}^{p} \xi(\hat{\theta}_i^t - \theta_{0,i}) =_{a.s.} \mathbb{E}\left\{ \xi(\sqrt{\delta} \tau_t Z) \right\}, \tag{8}
\]

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \xi_2(R_i^t, Z_i) =_{a.s.} \mathbb{E}\left\{ \xi_2(Z + \tau_t W, Z) \right\}. \tag{9}
\]
Corollary of Correctness

Under HDA, can define

$$\text{AMSE}(\hat{\theta}^t, \theta_0) = a.s \lim_{n,p \to \infty} ||\hat{\theta}^t - \theta_0||_2^2/p;$$

and

$$\text{AMSE}(\hat{\theta}^t, \theta_0) = \delta \tau_t^2.$$

and

$$\lim_{t \to \infty} \text{AMSE}(\hat{\theta}^t, \theta_0) = \delta \tau_*^2.$$
Convergence of AMP to M-Estimator, 2

**Theorem.** Assume A1-A4 and that \( \rho \) is strongly convex and \( \delta > 1 \). Let \((\tau_*, b_*)\) be a solution of the two equations

\[
\tau^2 = \delta \mathbb{E}\left\{ \psi(Z + \tau W; b)^2 \right\}, \tag{10}
\]

\[
\frac{1}{\delta} = \mathbb{E}\left\{ \psi'(Z + \tau W; b) \right\}. \tag{11}
\]

Assume that \( \text{AMSE}(\hat{\theta}^0, \theta_0) = \delta \tau_*^2 \). Then

\[
\lim_{t \to \infty} \text{AMSE}(\hat{\theta}^t, \hat{\theta}) = 0. \tag{12}
\]
Main Result: Asymptotic Variance Formula under High-Dimensional Asymptotics.

**Corollary.** Assume setting of previous Theorem.

\[
\lim_{n,p \to \infty} \text{Ave}_{i \in [p]} \text{Var}(\hat{\theta}_i) =_{a.s} V(\tilde{\Psi}, \tilde{F}),
\]

where the effective score \(\tilde{\Psi}\) is

\[
\tilde{\Psi}(\cdot) = \psi(\cdot; b^*),
\]

while the effective noise distribution \(\tilde{F}\) is

\[
\tilde{F} = F \ast N(0, \tau^2).$

Here \((\tau^*, b^*)\) are the unique solutions of the equations (10)-(11).
Driving Idea of High-Dimensional Asymptotics, 1

- M-estimate asy. equivalent to limit of many steps of AMP.
- First step of AMP contains extra Gaussian noise.
- Extra Gaussian noise caused by errors in initial parameter estimates.
- Extra Gaussian noise declines with AMP iterations but does not go away completely; declines to level defined by the fixed point equations of state evolution.
- Extra Gaussian noise in M-estimate characterized by fixed point equations of state evolution.
Driving Idea of High-Dimensional Asymptotics, 2

Classically M-estimate propagates true errors through score

\[ \hat{\theta} = \theta_0 + \frac{1}{B(\psi, F)} (X'X)^{-1} X' (\psi(Z_i)) + o_p(n^{-1/2}) \]

HDA propagates noisier effective errors through effective score

\[ \hat{\theta} = \theta_0 + \frac{1}{B(\tilde{\psi}, F)} \cdot n^{-1} \cdot X' \left( \tilde{\psi}(Z_i + \tau_* W_i) \right) + o_p(n^{-1/2}) \]

where \( W_i \perp\!\!\!\!\perp Z_i \) and both are iid, and \( \tilde{\psi} \) is the effective score.
How is Correctness of State Evolution Proved, 1?

- Simply apply existing paper:
  

  \[
  \min \| \tilde{Y} - \tilde{X}\beta \|_2^2 / 2 + \lambda \| \beta \|_1
  \]

- Setting was compressed sensing, where $\tilde{X}_{i,j} \sim iid \ N(0, \frac{1}{n})$, and
  \[p_n/n \to \delta \in (0, 1).\]

- Formalism of AMP and State Evolution was introduced, and developed in


- These papers systematically understood and used the ‘Extra Gaussian Noise’ property of High Dimensional Asymptotics.

- Generality of Bayati-Montanari treatment, easily accommodated M-estimation.
Iterative Thresholding

Soft Threshold Nonlinearity

\[ \eta(x; \lambda) = (|x| - \lambda)_+ \cdot \text{sgn}(x). \]

Iterative solution \( \beta^t, t = 0, 1, 2, 3, \ldots \)

\[
\begin{align*}
\beta^0 &= 0 \\
z^t &= \tilde{Y} - \tilde{X} \beta^t \\
\beta^{t+1} &= \eta(\beta^t + \tilde{X}^* z^t; \lambda_t)
\end{align*}
\]

Heuristic to solve LASSO: \( \min_\beta \| \tilde{Y} - \tilde{X} \beta \|_2^2/2 + \lambda \| \beta \|_1. \)
Approximate Message Passing (AMP) Iterative Thresholding

**First order Approximate Message Passing (AMP) algorithm**

\[ z^t = \tilde{Y} - \tilde{X}\beta^t + \frac{1}{\delta} z^{t-1} \langle \eta'_{t-1}(\tilde{X}^* z^{t-1} + \beta^{t-1}) \rangle. \]

\[ \beta^{t+1} = \eta_t(\beta^t + \tilde{X}^* z^t), \]

\[ \eta'_t(s) = \frac{\partial}{\partial s} \eta_t(s). \]

**Feature:** Essentially same cost as Iterative Soft Thresholding

If \( \eta = \text{soft thresholding}, \) \( \frac{1}{\delta} \langle \eta'_{t-1}(\tilde{X}^* z^{t-1} + \beta^{t-1}) \rangle = \frac{\|\beta^t\|_0}{n}. \)
How is Correctness of State Evolution Proved, 2?

Central Recursion in Bayati-Montanari 2011

- $h^t, q^t \in \mathbb{R}^N$
- $z^t, m^t \in \mathbb{R}^n$
- Initial Condition $q^0; m^{-1} = 0.$

$$
\begin{align*}
    h^{t+1} &= A^* m^t - \xi_t q^t, & m^t &= g_t(b^t, w) \\
    b^t &= A q^t - \lambda_t m^{t-1}, & q^t &= f_t(h^t, x_0)
\end{align*}
$$

- Reaction Coefficients: $\xi_t = \langle g'_t(b^t, w) \rangle; \lambda_t = \frac{1}{\delta} \langle f'_t(h^t, x^0) \rangle$
- State Evolution

$$
\tau_t^2 = E\{g_t(\sigma_t Z, W)^2\}; \quad \sigma_t^2 = \frac{1}{\delta} E\{f_t(\tau_{t-1} Z, X_0)\}
$$

where $W \sim F_W, X_0 \sim F_{X_0}$
\[ v^{t+1} = \delta X^\top \psi(W + S^t; b_t) + q_t v^t \]
\[ h^{t+1} = A^* m^t - \xi_t q^t \]

\[ S^t = -X v^t + \psi(W + S^{t-1}; b_{t-1}) \]
\[ b^t = A q^t - \lambda_t m^{t-1} \]

<table>
<thead>
<tr>
<th>(\vartheta^t)</th>
<th>(\psi(W + S^t; b_t))</th>
<th>(q_t)</th>
<th>(\vartheta^t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A^* m^t)</td>
<td>(m^t)</td>
<td>(q^t)</td>
<td>(A q^t)</td>
</tr>
<tr>
<td>(h^t)</td>
<td>(h^{t+1})</td>
<td>(1)</td>
<td>(-q^t)</td>
</tr>
</tbody>
</table>

**Table**: Correspondences between terms in the centered recursions of DLD-Montanari 2013, and recursions analyzed in Bayati-Montanari 2011.

We get exact correspondence between the two systems, provided we identify \(\delta \psi(W + S^t; b_t)\) with \(m^t = g_t(b^t; w)\) and \(-\delta h^t\) with \(f_t(h^t)\). One has, in particular, that \(\lambda_t = \frac{1}{\delta} \langle f_t'(h^t) \rangle = -1\), and that \(\xi_t = \langle g_t'(b^t, w) \rangle = \langle \delta \psi'(W + S^t; b_t) \rangle = q_t\).
Why did work on sparse signal recovery solve problem in robust regression?

_Duality of Sparsity and Robustness_

Explicit link between solutions of Lasso with $p > n$ and Huber with $p < n$.

Identity between estimating sparsely nonzero vector and uncovering outliers in a linear relation.
Let $\tilde{X}$ be a matrix with orthonormal rows such that $\tilde{X}X = 0$, i.e.

$$\text{null}(\tilde{X}) = \text{image}(X),$$  \hspace{1cm} (16)

finally, set $\tilde{Y} = \tilde{X}Y$.

**Proposition.**

*With problem instances $(Y, X)$ and $(\tilde{Y}, \tilde{X})$ related as above, the optimal values of the Lasso problem $(\text{Lasso}_\lambda)$ and the Huber problem $(\text{Huber}_\lambda)$ are identical. The solutions of the two problems are in one-one-relation. In particular, we have*

$$\hat{\vartheta} = (X^T X)^{-1} X^T (Y - \hat{\beta}).$$  \hspace{1cm} (17)

*(numerous references: e.g. Art Owen/IPOD & earlier)*
Isometry between performance of
- Lasso in $\varepsilon$-sparse regression problem, $p > n$
- Huber (M)-estimation in $\varepsilon$-contaminated data, $p < n$
Two Scalar Minimax problems

- Huber (1964) Minimax problem:

\[
v(\varepsilon) = \min_{\lambda} \sup_{H} \{ V(\psi_\lambda, F) : F = (1 - \varepsilon)\Phi + \varepsilon H \}
\]

Here $\psi_\lambda$ is Huber $\psi$ capped at $\lambda$, $\Phi$ is $N(0,1)$

- Minimax MSE under sparse means (DLD & Johnstone, 1992)

\[
m(\varepsilon) = \inf_{\lambda} \sup_{H} E_F(\eta_\lambda(X + Z) - X)^2 : X \sim (1 - \varepsilon)\nu_0 + \varepsilon H
\]

Here $\eta_\lambda$ is soft-thresholding at $\lambda$, $\nu_0$ is point mass at zero.
Full Circle, 4 (w/ Montanari, Johnstone)

- HDA regression with fraction $\varepsilon$ contaminated errors: $\delta = n/p$
  \[ V(\varepsilon, \delta) \equiv \min_{\lambda} \max_{H} \lim_{n \to \infty} \text{Ave}_i \text{Var}(\hat{\theta}_i) \]
  \[ V(\varepsilon, \delta) = \begin{cases} +\infty & v(\varepsilon) > \delta \\ v(\varepsilon) & 1 - v(\varepsilon)/\delta \end{cases} \]
  where $v(\varepsilon) \sim n/p$.

- HDA Sparse Regression, $\delta = n/p < 1$, $\tilde{Y} = \tilde{X}\beta_0 + \sigma \tilde{Z}$, $\tilde{Z} \sim_{iid} N(0, 1)$, $X_{i,j} \sim_{iid} N(0, n^{-1})$
  $\beta_0$ is $\varepsilon$-sparse
  \[ M(\varepsilon, \delta) \equiv \sup_{\sigma > 0, \lambda} \min_{\|\beta_0\|_0/n \leq \varepsilon} \lim_{n \to \infty} \text{MSE}(\hat{\beta}_\lambda, \beta_0) / \sigma^2 \]
  \[ M(\varepsilon, \delta) = \begin{cases} +\infty & m(\varepsilon) > \delta \\ m(\varepsilon) & 1 - m(\varepsilon)/\delta \end{cases} \]
Conclusions

- High-Dimensional Asymptotics imposes extra Gaussian noise in estimation, not seen classically
- Approximate Message Passing – new algorithm to understand and analyse
- State Evolution – new type of analysis to obtain properties of estimates
- New phenomena become visible, e.g. phase transitions in (M)-estimation, previously unknown.