

# A Concise (though Heuristic) Derivation of AMP

Weijia Zheng <sup>1</sup>

*Department of Information Engineering  
The Chinese University of Hong Kong*

wjzheng@link.cuhk.edu.hk

May 15, 2025

---

<sup>1</sup>I mainly took reference to: "A Simple Derivation of AMP and its State Evolution via First-Order Cancellation" by P. Schniter. This is a very readable file on this topic.

# Overview

- 1 Introduction
- 2 Onsager Correction Derivation
- 3 State Evolution

# Overview

## 1 Introduction

## 2 Onsager Correction Derivation

## 3 State Evolution

# High dimensional linear regression problem

## Liner regression formulation

Consider a problem of the form  $\mathbf{y} = \mathbf{A}\beta_0 + \mathbf{w}$ . We want to reconstruct  $\beta_0$  from  $\mathbf{y}$ .

$$\begin{array}{c} \textcolor{blue}{m} \end{array} \left[ \begin{array}{c} \textcolor{green}{A} \end{array} \right] \begin{array}{c} \textcolor{red}{N} \\ \textcolor{red}{\beta_0} \end{array} + \left[ \begin{array}{c} w \end{array} \right] = \left[ \begin{array}{c} \textcolor{blue}{y} \end{array} \right]$$

2

<sup>2</sup>Figure copied from "Approximate Message Passing for Statistical Inference and Estimation" (good lecture slides with Youtube video recording)

# High dimensional linear regression problem

## Linear regression formulation

Consider a problem of the form  $\mathbf{y} = \mathbf{A}\beta_0 + \mathbf{w}$ . We want to reconstruct  $\beta_0$  from  $\mathbf{y}$ .

$$\begin{array}{c} \textcolor{blue}{m} \end{array} \begin{array}{c} \textcolor{green}{A} \end{array} \begin{array}{c} \textcolor{red}{N} \\ \textcolor{red}{\beta_0} \end{array} + \begin{array}{c} \textcolor{black}{w} \end{array} = \begin{array}{c} \textcolor{blue}{y} \end{array}$$

2

$\mathbf{y}$ : an observed length- $m$  measurement vector

$\mathbf{w}$ : an unknown length- $m$  noise. Assume  $w \sim_{iid} \mathcal{N}(0, \tau_w)$

$\mathbf{A}$ : a known big  $m \times N$  (normalized) matrix with  $m < N$ ,  $\frac{m}{N} \rightarrow \delta \in \Omega(1)$

$\beta_0$ : a length- $N$  signal vector to find

<sup>2</sup>Figure copied from "Approximate Message Passing for Statistical Inference and Estimation" (good lecture slides with Youtube video recording)

The diagram illustrates a linear regression model. A green matrix  $A$  is shown with a vertical double-headed arrow to its left labeled  $m$  and a horizontal double-headed arrow above it labeled  $N$ . To the right of  $A$  is a red vector  $\beta_0$ , followed by a plus sign, a black vector  $w$ , an equals sign, and a blue vector  $y$ . All vectors are enclosed in square brackets.

$$A \beta_0 + w = y$$

Prior knowledge on  $\beta_0$ : sparsity

People may assume sparsity of  $\beta_0$ . That is, it has only  $K \ll N$  nonzero entries.

$$\begin{matrix}
 & \xrightarrow{N} & \\
 m \updownarrow & \left[ \begin{matrix} & \\ & \\ & \end{matrix} \right] & \left[ \begin{matrix} \\ \\ \end{matrix} \right] \\
 & A & \beta_0
 \end{matrix}
 + \begin{bmatrix} \\ \\ \end{bmatrix} w = \begin{bmatrix} \\ \\ \end{bmatrix} y$$

### Prior knowledge on $\beta_0$ : sparsity

People may assume sparsity of  $\beta_0$ . That is, it has only  $K \ll N$  nonzero entries.

## NP-hardness of sparse recovery in general

Assume  $K$ -sparse, the problem: for any given  $\mathbf{A}$ , find  $\arg \min_{\beta} \|\mathbf{A}\beta - \mathbf{y}\|^2$  is NP-hard. In fact, even if we know the entries' values of the ground-truth  $\beta_0$ , the problem is still NP-hard. <sup>a</sup>

<sup>a</sup>For the NP-hardness: one can do reduction using Exact Cover by 3-Sets (X3C).

# (Sparsity inspired) LASSO

## LASSO and ISTA

$$\min_{\beta} \underbrace{\frac{1}{2} \|\mathbf{y} - \mathbf{A}\beta\|^2}_{\triangleq g(\beta)} + \lambda \|\beta\|_1.$$

Iterative Soft-Thresholding Algorithm (ISTA) can solve this:

$$\begin{aligned} \mathbf{v} &= \mathbf{y} - \mathbf{A}\beta^t \\ \beta^{t+1} &= \text{soft}(\beta^t + s\mathbf{A}^T \mathbf{v}^t; s\lambda). \end{aligned}$$

Writing it into a more intuitive form:

$$\beta^{t+1} = \underbrace{\text{soft}\left(\underbrace{\beta^t - s \nabla g(\beta^t)}_{\text{grad. descent}}; s\lambda\right)}_{\text{impose sparsity}}.$$

$$\nabla g(\beta) = \mathbf{A}^T (\mathbf{A}\beta - \mathbf{y}).$$

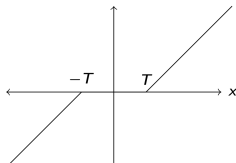


Figure:  $\text{soft}(x, T)$ .

One can tune the parameter  $\lambda$  to control the sparsity, and  $s$  here works as a stepsize.

LASSO is convex in  $\beta$ .



# AMP Framework

LASSO & ISTA are great, but...

LASSO is motivated by **sparsity** alone, and it does not consider the signal's prior distribution, which may sometimes be available. Thus, people want to integrate the knowledge of  $\beta_0 \sim_{iid} p_\beta$  into inference of  $\beta_0$ .<sup>a</sup>

---

<sup>a</sup>AMP does not explicitly assume sparsity of  $\beta_0$ .

The problem then changes to find an  $\hat{\beta}$  for  $\mathbf{y} = \mathbf{A}\beta_0 + \mathbf{w}$ , while  $\beta_0 \sim_{iid} p_\beta$ .

# AMP Framework

LASSO & ISTA are great, but...

LASSO is motivated by **sparsity** alone, and it does not consider the signal's prior distribution, which may sometimes be available. Thus, people want to integrate the knowledge of  $\beta_0 \sim_{iid} p_\beta$  into inference of  $\beta_0$ .<sup>a</sup>

<sup>a</sup>AMP does not explicitly assume sparsity of  $\beta_0$ .

The problem then changes to find an  $\hat{\beta}$  for  $\mathbf{y} = \mathbf{A}\beta_0 + \mathbf{w}$ , while  $\beta_0 \sim_{iid} p_\beta$ .

We compare the procedure of AMP and ISTA at below.

Approximate message passing (AMP)

$$\mathbf{v}^t = \mathbf{y} - \mathbf{A}\beta^t + \underbrace{\frac{\mathbf{v}^{t-1}}{m} \sum_{j=1}^N \eta'_{t-1}(r_j^{t-1})}_{\text{Onsager correction term}}$$

$$\beta^{t+1} = \eta_t(\underbrace{\beta^t + s\mathbf{A}^T \mathbf{v}^t}_{\triangleq \mathbf{r}^t}).$$

Iterative Soft Thresholding Algo.  
(ISTA)

$$\begin{aligned}\mathbf{v}^t &= \mathbf{y} - \mathbf{A}\beta^t \\ \beta^{t+1} &= \text{soft}(\beta^t + s\mathbf{A}^T \mathbf{v}^t; s\lambda).\end{aligned}$$

There are some requirements in  $\mathbf{A}$ , the sensing/measurement matrix. In general, assume  $\mathbf{A}$  to be entry-wisely iid generated with  $\mathbb{E}a_{ij} = 0$  and  $\mathbb{E}(a_{ij}^2) = \frac{1}{m}$  suffices.

In fact, in the paper we will go through, they assume  $a_{ij} \in \mathcal{U}\{\pm \frac{1}{\sqrt{m}}\}$ . But this is mainly to simplify the proof, and it can be extended to more general cases.

There are some requirements in  $\mathbf{A}$ , the sensing/measurement matrix. In general, assume  $\mathbf{A}$  to be entry-wisely iid generated with  $\mathbb{E}a_{ij} = 0$  and  $\mathbb{E}(a_{ij}^2) = \frac{1}{m}$  suffices.

In fact, in the paper we will go through, they assume  $a_{ij} \in \mathcal{U}\{\pm \frac{1}{\sqrt{m}}\}$ . But this is mainly to simplify the proof, and it can be extended to more general cases.

## AMP iteration

$$\mathbf{v}^t = \mathbf{y} - \mathbf{A}\boldsymbol{\beta}^t + \underbrace{\frac{\mathbf{v}^{t-1}}{m} \sum_{j=1}^N \eta'_{t-1}(r_j^{t-1})}_{\text{Onsager correction term}}$$

$$\boldsymbol{\beta}^{t+1} = \eta_t(\underbrace{\boldsymbol{\beta}^t + s\mathbf{A}^T \mathbf{v}^t}_{\triangleq \mathbf{r}^t}).$$

$\mathbf{r}^t \in \mathbb{R}^N$ , termed "effective observation"

$\eta_t(\cdot)$  is called a "denoising function"

$$[\eta_t(\mathbf{r})]_j = \eta_t(r_j)$$

In ISTA,  $\eta_t = \text{soft}()$  and we do not consider any correction term

## Main purpose of the paper

Simply to understand: why there is such an "Onsager term", and what should be chosen as the denoising function  $\eta_t(\cdot)$ .

# Overview

1 Introduction

2 Onsager Correction Derivation

3 State Evolution

# Why the Onsager Correction?

## AMP Iteration Recap

For  $\mathbf{y} = \mathbf{A}\beta_0 + \mathbf{w}$ , AMP iterates:

$$\mathbf{v}^t = \mathbf{y} - \mathbf{A}\beta^t + \underbrace{\frac{\mathbf{v}^{t-1}}{m} \sum_{j=1}^N \eta'_{t-1}(r_j^{t-1})}_{\text{Onsager correction term}},$$

$$\beta^{t+1} = \eta_t (\beta^t + \mathbf{A}^T \mathbf{v}^t) \triangleq \eta_t(\mathbf{r}^t).$$

# Why the Onsager Correction?

## AMP Iteration Recap

For  $\mathbf{y} = \mathbf{A}\beta_0 + \mathbf{w}$ , AMP iterates:

$$\mathbf{v}^t = \mathbf{y} - \mathbf{A}\beta^t + \underbrace{\frac{\mathbf{v}^{t-1}}{m} \sum_{j=1}^N \eta'_{t-1}(r_j^{t-1})}_{\text{Onsager correction term}},$$

$$\beta^{t+1} = \eta_t(\beta^t + \mathbf{A}^T \mathbf{v}^t) \triangleq \eta_t(\mathbf{r}^t).$$

Partial goal: Ensure  $\mathbf{r}^t - \beta_0 \approx$  Gaussian noise. (See next page.)

Onsager term adjusts  $\mathbf{v}^t$  to cancel error correlations.

## Our focus soon

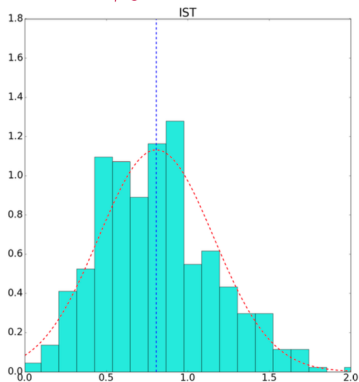
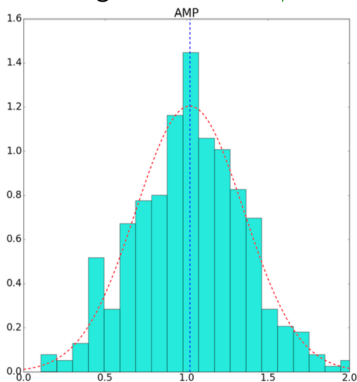
Derive the Onsager term by analyzing the difference between the effective observation and ground-truth signal  $\mathbf{e}^t = \mathbf{r}^t - \beta_0$ .

# Correction terms work

Correction terms push the effective observation to  $\beta_0 +$  some (tiny) Gaussian.  
Comparing with vanilla IST, such effect is decisive.

$A : m \times N = 2000 \times 4000$ ;  $\beta_0$  has 500 non-zeros  $\sim$  iid unif  $\pm 1$

Histogram of  $A^T r^t + \beta^t$  at indices where  $\beta_0 = +1$  at  $t = 10$





# Error Analysis

## Define the Error

Recall  $\mathbf{r}^t = \boldsymbol{\beta}^t + \mathbf{A}^T \mathbf{v}^t$ . The error is:  $\mathbf{e}^t = \mathbf{r}^t - \boldsymbol{\beta}_0 = (\boldsymbol{\beta}^t + \mathbf{A}^T \mathbf{v}^t) - \boldsymbol{\beta}_0$ .  
 Substitute  $\mathbf{v}^t = \mathbf{y} - \mathbf{A}\boldsymbol{\beta}^t + \mathbf{u}^t$ , where  $\mathbf{u}^t$  denotes (any) correction term:

$$\mathbf{r}^t = \boldsymbol{\beta}^t + \mathbf{A}^T (\mathbf{y} - \mathbf{A}\boldsymbol{\beta}^t + \mathbf{u}^t).$$

Since  $\mathbf{y} = \mathbf{A}\boldsymbol{\beta}_0 + \mathbf{w}$ ,  $\mathbf{A}^T \mathbf{y} = \mathbf{A}^T (\mathbf{A}\boldsymbol{\beta}_0 + \mathbf{w})$ . Then

$$\mathbf{e}^t = (\mathbf{I} - \mathbf{A}^T \mathbf{A})(\boldsymbol{\beta}^t - \boldsymbol{\beta}_0) + \mathbf{A}^T (\mathbf{w} + \mathbf{u}^t).$$

# Error Analysis

## Define the Error

Recall  $\mathbf{r}^t = \beta^t + \mathbf{A}^T \mathbf{v}^t$ . The error is:  $\mathbf{e}^t = \mathbf{r}^t - \beta_0 = (\beta^t + \mathbf{A}^T \mathbf{v}^t) - \beta_0$ .  
Substitute  $\mathbf{v}^t = \mathbf{y} - \mathbf{A}\beta^t + \mathbf{u}^t$ , where  $\mathbf{u}^t$  denotes (any) correction term:

$$\mathbf{r}^t = \beta^t + \mathbf{A}^T (\mathbf{y} - \mathbf{A}\beta^t + \mathbf{u}^t).$$

Since  $\mathbf{y} = \mathbf{A}\beta_0 + \mathbf{w}$ ,  $\mathbf{A}^T \mathbf{y} = \mathbf{A}^T (\mathbf{A}\beta_0 + \mathbf{w})$ . Then

$$\mathbf{e}^t = (\mathbf{I} - \mathbf{A}^T \mathbf{A})(\beta^t - \beta_0) + \mathbf{A}^T (\mathbf{w} + \mathbf{u}^t).$$

## Zooming in the $l$ -th entry

We focus more on  $\beta^t$ .  $\beta^t = \eta_{t-1}(\mathbf{r}^{t-1})$ . Its  $l$ -th entry is  $\beta_l^t = \eta_{t-1}(r_l^{t-1})$ .  
And we decompose  $r_l^{t-1}$  as:

$$r_l^{t-1} = \beta_l^{t-1} + \sum_k a_{kl} v_k^{t-1} = \underbrace{\beta_l^{t-1} + \sum_{k \neq i} a_{kl} v_k^{t-1}}_{\triangleq r_{l \setminus i}^{t-1}, \text{ assumed indep. of } \{a_{ij}\}_j} + a_{il} v_i^{t-1}. \quad (1)$$

## Error decomposition

Then the error (its  $j$ -th entry) becomes

$$\begin{aligned}
 e_j^t &= \sum_i a_{ij} \left[ \sum_{l \neq j} a_{il} (\beta_{0,l} - \beta_l^t) + w_i + u_i^t \right] \\
 &= \underbrace{\sum_i a_{ij} \sum_{l \neq j} a_{il} [\beta_{0,l} - \eta_{t-1}(r_{l \setminus i}^{t-1})]}_{\text{Independence + CLT} \Rightarrow \sim_d \text{ Gaussian}} + \underbrace{\sum_i a_{ij} \left[ u_i^t + \sum_{l \neq j} -\frac{v_i^{t-1}}{m} \eta'_{t-1}(r_{l \setminus i}^{t-1}) \right]}_{\triangleq T_1, \text{ want to make it small when } m \text{ is large}}
 \end{aligned}$$

$a_{ij}$  and  $v_i^{t-1}$  are coupled!

In the above, we used Taylor expansion:

$$\beta_l^t = \eta_{t-1}(r_l^{t-1}) = \eta_{t-1}(r_{l \setminus i}^{t-1} + a_{il} v_i^{t-1}) \approx \eta_{t-1}(r_{l \setminus i}^{t-1}) + a_{il} v_i^{t-1} \eta'_{t-1}(r_{l \setminus i}^{t-1}).$$

And, we used  $a_{il}^2 = \frac{1}{m}$ .

## Focus on Correction Term

The last error term (the not Gaussian one) is the only term involving  $\mathbf{u}^t$ :

$$T_1 = \sum_i a_{ij} \left[ u_i^t - \sum_{l \neq j} \frac{v_i^{t-1}}{m} \eta'_{t-1}(r_{l \setminus i}^{t-1}) \right] \quad (2)$$

Note that  $\mathbf{u}^t$  is some correction term free to choose. And we wish such choice to make  $T_1$  small.

We can see why Onsager is good by observing its form:  $u_i^t \triangleq \frac{v_i^{t-1}}{m} \sum_{l=1}^N \eta'_{t-1}(r_l^{t-1})$ .

We can proceed an estimation of  $T_1$ :

$$T_1 \approx_{2 \text{ order Taylor}} \sum_i a_{ij} \left[ \frac{v_i^{t-1}}{m} \sum_{l=1}^N \eta'_{t-1}(r_l^{t-1}) - \sum_{l \neq j} \frac{v_i^{t-1}}{m} \eta'_{t-1}(r_{l \setminus i}^{t-1}) \right] \quad (3)$$

$$\approx \frac{1}{m} \sum_i a_{ij} v_i^{t-1} \left[ \eta'_{t-1}(r_j^{t-1}) + \sum_{l \neq j} a_{il} v_i^{t-1} \eta''_{t-1}(r_{l \setminus i}^{t-1}) \right] \in \mathcal{O}\left(\frac{1}{\sqrt{m}}\right) \quad (4)$$

The two terms in eq. (4) are both  $\mathcal{O}(\frac{1}{\sqrt{m}})$ . And **it is hard to design a better correction term (without sacrificing too much computational complexity.)**

# Overview

1 Introduction

2 Onsager Correction Derivation

3 State Evolution

# Completing Error Estimation

## Recap: Error Decomposition

From the previous slide, the error  $e_j^t$  is:

$$e_j^t = \underbrace{\sum_i a_{ij} \sum_{l \neq j} a_{il} \underbrace{[\beta_{0,l} - \eta_{t-1}(r_{l \setminus i}^{t-1})]}_{\triangleq \epsilon_{l \setminus i}^t}}_{=S_1} + \underbrace{\sum_i a_{ij} w_i}_{=S_2} + \underbrace{\cancel{T_1}}_{\text{ignored when } m \gg 1}$$

Independence + CLT  $\Rightarrow \sim_d$  Gaussian

$S_2$  is much easier to handle, it has zero mean and variance  $= \tau_w$ . ( $w$ 's power)

$S_1$  has mean zero. And it has variance  $\frac{1}{m^2} \sum_{i=1}^m \sum_{l \neq j} (\epsilon_{l \setminus i}^t)^2 \approx \frac{n}{m} \frac{1}{n} \sum_{l=1}^n (\epsilon_l^t)^2$ , where  $\epsilon_l^t \triangleq \beta_{0,l} - \eta_{t-1}(r_l^{t-1})$  is the **effective noise**.

People denote  $\mathcal{E}^t = \frac{1}{n} \sum_{l=1}^n (\epsilon_l^t)^2$ . Then one can write

$$e_j^t = \beta_{0,j} - r_j^t \sim_{\text{approx.}} \mathcal{N}(0, \underbrace{\delta^{-1} \mathcal{E}^t + \tau_w}_{\triangleq \tau_r^t}).$$

fixed

# State Evolution: Predicting Performance

## State Evolution Equations

Track error variance  $\tau_r^t = \text{Var}(r_j^t - \beta_{0,j})$ :

$$\tau_r^t = \delta^{-1} \mathcal{E}^t + \tau_w, \quad \mathcal{E}^{t+1} = \mathbb{E} [\eta_t(\beta_0 + Z_t) - \beta_0]^2, \quad Z_t \sim \mathcal{N}(0, \underbrace{\tau_r^t}_{\approx \frac{\|\mathbf{v}^t\|^2}{m}}).$$

$\mathcal{E}^t$ : Mean squared error (MSE) of denoiser at iteration  $t$ .

It predicts AMP's MSE without running the algorithm.

Basically,  $\mathcal{E}^t$  is what we can control in  $\tau_r^t$ . Hence we want to choose a function  $\eta_t(\cdot)$  to minimize it!

Now, we see why people choose  $\eta_t \leftarrow$  posterior mean estimator (PME).<sup>3</sup>

<sup>3</sup>One can use Tweedie's formula here, when  $\beta_0$ 's prior is known.