A. Complex Normal Distribution

Complex normal is the distribution of a complex random variable whose imaginary and real parts are jointly Gaussian.

Standard complex normal distribution. A random variable $Z = X + \mathrm{i}Y$ where $X, Y \in \mathbb{R}$ has standard complex normal distribution represented by $\mathcal{CN}(0,1)$ if

$$X, Y \sim \mathcal{N}(0, 1/2), \quad X \perp\!\!\!\perp Y.$$

General complex Gaussian distribution. A random vector $\mathbf{Z} = \mathbf{X} + \mathrm{i}\mathbf{Y}$ where $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^n$ has complex Gaussian distribution $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \mathbf{C})$ if $\mathbf{X}$ and $\mathbf{Y}$ are jointly Gaussian with

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{Z}], \tag{42}$$
$$\boldsymbol{\Gamma} = \mathbb{E}[(\mathbf{Z} - \boldsymbol{\mu})(\mathbf{Z} - \boldsymbol{\mu})^{\mathsf{H}}], \tag{43}$$
$$\mathbf{C} = \mathbb{E}[(\mathbf{Z} - \boldsymbol{\mu})(\mathbf{Z} - \boldsymbol{\mu})^{\mathsf{T}}]. \tag{44}$$

The parameters $\boldsymbol{\mu}, \boldsymbol{\Gamma}$, and $\mathbf{C}$ are called mean vector, covariance matrix, and relation matrix respectively. Alternatively, if we define

$$\mathbf{C}_{XX} = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{X} - \boldsymbol{\mu}_x)^{\mathsf{T}}], \quad \boldsymbol{\mu}_X = \mathbb{E}[\mathbf{X}],$$
$$\mathbf{C}_{YY} = \mathbb{E}[(\mathbf{Y} - \boldsymbol{\mu}_Y)(\mathbf{Y} - \boldsymbol{\mu}_Y)^{\mathsf{T}}], \quad \boldsymbol{\mu}_X = \mathbb{E}[\mathbf{Y}],$$
$$\mathbf{C}_{XY} = \mathbf{C}_{YX}^{\mathsf{T}} = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{Y} - \boldsymbol{\mu}_Y)^{\mathsf{T}}],$$

then $\mathbf{X}, \mathbf{Y}$ are jointly Gaussian with distribution

$$(\mathbf{X}, \mathbf{Y}) \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y\end{bmatrix}, \begin{bmatrix}\mathbf{C}_{XX} & \mathbf{C}_{XY} \\ \mathbf{C}_{YX} & \mathbf{C}YY\end{bmatrix}\right).$$

The matrices $\boldsymbol{\Gamma}$ and $\mathbf{C}$ are related to covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$ through the following equations:

$$\boldsymbol{\Gamma} = \mathbf{C}_{XX} + \mathbf{C}_{YY} + \mathrm{i}(\mathbf{C}_{YX} - \mathbf{C}_{XY}),$$
$$\boldsymbol{\Gamma} = \mathbf{C}_{XX} - \mathbf{C}_{YY} + \mathrm{i}(\mathbf{C}_{YX} + \mathbf{C}_{XY}).$$

B. Empirical Convergence of Vector Sequences

Here we review some definitions that are standard to in papers that use approximate message passing.

Definition 1 (Pseudo Lipschitz Continuity)**.** A function $\mathbf{f}$ is called pseudo-Lipschitz continuous of order p with constant C if for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathrm{dom}(\mathbf{f})$

$$\|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2)\| \leq C \|\mathbf{x}_1 - \mathbf{x}_2\| (1 + \|\mathbf{x}_1\|^{p-1} + \|\mathbf{x}_2\|^{p-1}). \tag{45}$$

Note that for $p = 1$ this definition is a equivalent to the definition of the standard Lipschitz-continuity.

Definition 2 (Uniform Lipschitz-continuity)**.** A function $\mathbf{f}$ on $\mathcal{X} \times \mathcal{W}$ is *uniformly Lipschitz-continuous* in $\mathbf{x}$ at $\bar{\boldsymbol{\omega}}$ if there exists constants $L_1, L_2 \geq 0$ and an open neighborhood U of $\bar{\boldsymbol{\omega}}$ such that for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}, \boldsymbol{\omega} \in U$

$$\|\mathbf{f}(\mathbf{x}_1, \boldsymbol{\omega}) - \mathbf{f}(\mathbf{x}_2, \boldsymbol{\omega})\| \leq L_1 \|\mathbf{x}_1 - \mathbf{x}_2\|, \tag{46}$$

and for all $\mathbf{x} \in \mathcal{X}, \boldsymbol{\omega}_1, \boldsymbol{\omega}_2 \in \mathcal{W}$

$$\|\mathbf{f}(\mathbf{x}, \boldsymbol{\omega}_1) - \mathbf{f}(\mathbf{x}, \boldsymbol{\omega}_2)\| \leq L_2(1 + \|\mathbf{x}\|) \|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|. \tag{47}$$

Definition 3 (Empirical convergence of sequences)**.** Consider a sequence of vectors $\mathbf{x}(N) = \{\mathbf{x}_n(N)\}_{n=1}^N$ with $\mathbf{x}_n(N) \in \mathbb{R}^d$, i.e. each $\mathbf{x}(N)$ is a block vector with a total of Nd components. For a finite $p \geq 1$, we say that the vector sequence $\mathbf{x}(N)$ converges empirically with pth order moments if there exists a random variable $X \in \mathbb{R}^d$ such that

- $\mathbb{E}\|X\|_p^p < \infty$;
- for any $f : \mathbb{R}^d \to \mathbb{R}$ that is pseudo-Lipschitz of order p,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x}_n(N)) = \mathbb{E}[f(X)]. \tag{48}$$

With some abuse of notation, we represent this with

$$\lim_{N\to\infty} \mathbf{x}_n \overset{PL(p)}{=} X, \tag{49}$$

where we have omitted the dependence on N to ease the notation. In this definition the sequence $\{\mathbf{x}(N)\}$ can be random or deterministic. If it is random we require the equality in (48) to hold almost surely. In particular, if the sequence $\{\mathbf{x}_n\}$ is i.i.d. with $\mathbf{x}_n \sim p_X(\cdot)$, with $\mathbb{E}\|X\|_p^p < \infty$, then $\{\mathbf{x}_n\}$ converges empirically to X with pth order. The extension of this definition to sequence of matrices and higher order tensors is straightforward.

Definition 4 (Convergence in distribution)**.** A sequence of random vectors $\mathbf{x}_n \in \mathbb{R}^d$ converges in distribution (also known as weak convergence) to $\mathbf{x}$ if for all bounded functions $f : \mathbb{R}^d \to \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{E}f(\mathbf{x}_n) = \mathbb{E}f(\mathbf{x}). \tag{50}$$

PL(p) convergence is equivalent to convergence in distribution plus convergence of the pth moment (Bayati and Montanari, 2011).

Definition 5 (Wasserstein-p distance)**.** Wasserstein-p distance between two probability measures μ, ν on Euclidean space $\mathbb{R}^d$ is

$$W_p(\mu, \nu) = \inf_{\gamma \in \Gamma} \left(\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma} \|\mathbf{x} - \mathbf{y}\|_p^p\right)^{\frac{1}{p}}, \tag{51}$$

where Γ is the set of all probability measures on the product space $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν.

PL(p) convergence is also equivalent to convergence the empirical measure of the sequence $\mathbf{x}_n$ to probability measure of X in Wasserstein-p distance (Villani, 2008).

Definition 6. The empirical distribution of a sequence of vectors $\{\mathbf{x}_i\}_{i=1}^N$ with $\mathbf{x}_i \in \mathbb{R}^d$ (or $\mathbb{C}^d$) is denoted by $\mathbb{P}_n$ and is defined as

$$\mathbb{P}_n(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i),$$

where $\delta(\cdot)$ is the Dirac measure.

Proposition 1 (Equivalence of PL(p) convergence and convergence in Wasserstein-p metric (Villani, 2008))**.** The empirical convergence defined in Definition 3 is equivalent to the following convergence in Wasserstein-p metric

$$\lim_{N\to\infty} W_p(\mathbb{P}_N, \mathbb{P}) = 0,$$

where $\mathbb{P}_N$ is the empirical distribution of $\mathbf{x}(N)$ and $\mathbb{P}$ is the distribution of the random variable X in Definition 3 to which the sequence is converging empirically.

C. 1D Convolution Operators in Matrix Form

In this section we derive the matrix form of 1D convolution operators to show how these operators look like in time domain. As we will see, convolution operators in time domain can be represented as a *doubly block circulant matrix*. Because of this structure, approximate message passing (AMP) (discussed in Appendix D) cannot be directly used to obtain estimation error of ridge regression for convolutional inverse problem in time domain. This is due to the assumption in AMP that the measurement matrix has i.i.d. entries. If this assumption can be relaxed, we can analyze estimators other than ridge, and compute error metrics other than MSE. We hope to follow this direction in a future work.

First assume that in the convolutional model in (1), $n_x = n_y = 1$, i.e. the input and output both have one channel. Also for a matrix $\mathbf{Z} \in \mathbb{R}^{m\times n}$, let $\vec{(\mathbf{Z})} \in \mathbb{R}^{nm}$ represent the vector constructed by stacking $\mathbf{X}$ in a vector row by row. To simplify the notation, we zero pad the convolution kernel which in this case is a vector of size k, so that it will have size T and we still use $\mathbf{K}$ to represent the zero-padded kernel to simplify the notation. In this case, the convolution operator $\mathbf{K} : \mathbf{X} \mapsto \mathbf{K} * \mathbf{X}$ can be represented as a circulant matrix $\mathbf{C} : \mathrm{vec}(\mathbf{X}) \mapsto \mathbf{C}\,\mathrm{vec}(\mathbf{X})$

$$\mathbf{C} = \begin{bmatrix} K_1 & K_2 & K_3 & \dots & K_T \\ K_T & K_1 & K_2 & \dots & K_{T-1} \\ K_{T-1} & K_T & K_1 & \dots & K_{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ K_2 & K_3 & K_4 & \dots & K_1 \end{bmatrix} \tag{52}$$

When the number of input channels and output channels are n_x and n_y respectively, the convolution can be represented in matrix form as matrix with blocks of circulant matrices

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} & \dots & \mathbf{C}1, n_x \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \dots & \mathbf{C}2, n_x \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{n_y,1} & \mathbf{C}_{n_y,2} & \dots & \mathbf{C}n_y, n_x \end{bmatrix}, \tag{53}$$

where each $\mathbf{C}_{ij}$ is a circulant matrix of the form (52) constructed from $\mathbf{K}_{ij*}$. Since the adjoint of a circulant matrix is also a circulant matrix, one can see that the adjoint of a 1D convolution (with stride 1) is also a convolution with respect to another kernel.

D. Approximate Message Passing

In this section we briefly describe the approximate message passing (AMP) algorithm for linear inverse problems (Bayati and Montanari, 2011). Consider the problem of estimating $\mathbf{x}^0$ from linear observations

$$\mathbf{y} = \mathbf{A}\mathbf{x}^0 + \boldsymbol{\xi}, \tag{54}$$

where $\mathbf{A} \in \mathbb{R}^{n_y \times n_x}$ is a known matrix and $\boldsymbol{\xi}$ is i.i.d. zero-mean Gaussian noise with variance σ^2. Approximate message passing is an iterative algorithm to solve this problem

$$\mathbf{x}^{t+1} = \boldsymbol{\eta}_t(\mathbf{A}^\mathsf{T}\mathbf{z}^t + \mathbf{x}^t) \tag{55}$$

$$\mathbf{z}^t = \mathbf{y} - \mathbf{A}\mathbf{x}^t + \underbrace{\frac{1}{\delta}\mathbf{z}^{t-1}\langle \boldsymbol{\eta}'_{t-1}(\mathbf{A}^\mathsf{T}\mathbf{z}^{t-1} + \mathbf{x}^{t-1})\rangle}_{\text{Onsager correction}}, \tag{56}$$

where $\boldsymbol{\eta}_t(\cdot)$ is a denoiser that acts component-wise, and $\langle\cdot\rangle$ is the empirical averaging operator.

The key property of AMP algorithm is that when the sensing matrix $\mathbf{A}$ is large with i.i.d. sub-Gaussian entries with $\mathbb{E}\mathbf{A}_{ij}^2 = 1/n_y$, the behavior of the algorithm at each iteration can be exactly characterized via a *scalar* recursive equation called the *state evolution* (SE)

$$\tau_{t+1}^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\left[(\eta_t(X_0 + \tau_t Z) - X_0)^2\right], \tag{57}$$

where $X_0 \sim p_{X_0}$ independent of $Z \sim \mathcal{N}(0,1)$. Here p_{X_0} is the distribution to which the components of $\mathbf{x}^0$ are converging empirically. See Appendix B for background on empirical convergence of sequences and some definitions we would use throughout this paper. Given τ_t, as $n_x, n_y \to \infty$ with fixed ratio $\delta := n_y/n_x$ we have

$$\begin{bmatrix} \mathbf{x}_0 \\ \mathbf{x}^t \end{bmatrix} \overset{PL(2)}{=} \begin{bmatrix} X_0 \\ \eta_{t-1}(X_0 + \tau_{t-1}Z) \end{bmatrix}, \tag{58}$$

where as in SE we have $X_0 \sim p_{X_0}$ independent of $Z \sim \mathcal{N}(0,1)$.

This convergence allows us to compute the estimation error. If we define the mean squared error of the estimate at iteration t to be $\text{MSE} = 1/n_x \left\|\mathbf{x}^0 - \mathbf{x}^t\right\|_2^2$, then in the large system limit almost surely

$$\text{MSE} = \mathbb{E}\left[(\boldsymbol{\eta}_{t-1}(X_0 + \tau_{t-1}Z) - X_0)^2\right], \tag{59}$$

where the expectation is over X_0 and Z.

D.1. AMP for ridge regression

In this section we show how AMP can be used to derive asymptotic error of ridge regression

$$\widehat{\mathbf{x}}_{\text{ridge}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_2^2. \tag{60}$$

The solution to this optimization problem is

$$\widehat{\mathbf{x}}_{\text{ridge}} = (\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda \mathbf{I})^{-1}\mathbf{A}^\mathsf{T}\mathbf{y}. \tag{61}$$

Next, consider the AMP recursion in (55) and (56) with a fixed denoiser $\boldsymbol{\eta}_t(\mathbf{x}) = \alpha\mathbf{x}$

$$\mathbf{x}^{t+1} = \alpha(\mathbf{A}^\mathsf{T}\mathbf{z}^t + \mathbf{x}^t), \tag{62}$$

$$\mathbf{z}^t = \mathbf{y} - \mathbf{A}\mathbf{x}^t + \frac{\alpha}{\delta}\mathbf{z}^{t-1}. \tag{63}$$

The next lemma shows that this recursion solves the ridge regression for a specific regularization parameter $lambda$.

Lemma 7. The fixed point of AMP algorithm with $\boldsymbol{\eta}_t(\mathbf{x}) = \alpha\mathbf{x}$ is the solution of ridge regression with

$$\lambda = \frac{(1-\alpha)(1-\alpha/\delta)}{\alpha}, \tag{64}$$

where $\delta = n_y/n_x$.

Proof. Let $\mathbf{x}^*$ and $\mathbf{y}^*$ denote the fixed points of the AMP recursion. Then we have

$$\mathbf{x}^* = \alpha(\mathbf{A}^\mathsf{T}\mathbf{z}^* + \mathbf{x}^*), \tag{65}$$

$$\mathbf{z}^* = \mathbf{y} - \mathbf{A}\mathbf{x}^* + \frac{\alpha}{\delta}\mathbf{z}^*. \tag{66}$$

Therefore,

$$\mathbf{z}^* = \frac{1}{1-\alpha/\delta}(\mathbf{y} - \mathbf{A}\mathbf{x}). \tag{67}$$

Plugging this back to Equation (65) we get

$$\mathbf{x}^* = \left(\mathbf{A}^\mathsf{T}\mathbf{A} + \frac{(1-\alpha)(1-\alpha/\delta)}{\alpha}\mathbf{I}\right)^{-1}\mathbf{A}^\mathsf{T}\mathbf{y}. \tag{68}$$

Comparing this to (61) proves the result. □

Given a λ, one can solve the quadratic equation (64) to find the α that satisfies the equation. This is a quadratic equation that has two solutions. As we show in Section D.2, so long as the regularization parameter λ is non-negative, this quadratic equation always has two real and positive solutions. But only for the smaller solution the AMP recursions for solving ridge regression converges, and hence only the smaller one is valid.

Having found the α we can use the state evolution (57) to find its fixed point. For ridge regression, this can be done in closed form. When $\mathbb{E}\mathbf{A}_{ij}^2 = 1/n_y$ the state evolution for ridge regression can be written as

$$\tau_{t+1}^2 = \sigma^2 + \frac{1}{\delta}((1-\alpha)^2\sigma_X^2 + \alpha^2\tau_t^2), \tag{69}$$

If we define the fixed point value $\tau := \lim_{t\to\infty}\tau_t$ we have that it should satisfy

$$\tau^2 = \sigma^2 + \frac{1}{\delta}((1-\alpha)^2\sigma_X^2 + \alpha^2\tau^2), \tag{70}$$

from which we obtain

$$\tau^2 = \frac{\sigma^2 + \frac{1}{\delta}(1-\alpha^2)\sigma_X^2}{1 - \frac{\alpha^2}{\delta}}. \tag{71}$$

The mean squared error then can be obtained as

$$\frac{1}{n_x}\left\|\widehat{\mathbf{x}}_{\text{ridge}} - \mathbf{x}_0\right\|_2^2 = \mathbb{E}\left[(\alpha(X_0 + \tau Z) - X_0)^2\right] \tag{72}$$

$$= (\alpha - 1)^2\mathbb{E}X_0^2 + \alpha^2\tau^2. \tag{73}$$

Adjusting for variance of $\mathbf{A}_{ij}$. When $\mathbf{A}_{ij} \sim \mathcal{N}(0, \sigma_A^2/n_y)$ instead of having $\mathbb{E}\mathbf{A}_{ij}^2 = 1/n_y$, either we have to slightly modify the state evolution or rescale the inverse problem to adjust for the variance. Here we use the latter approach. Assume that in Equation (54) we have $\mathbb{E}\mathbf{A}_{ij}^2 = \sigma_A^2/n_y$. Then we can divide both sides by $1/\sigma_A$ to correct for the variance. This rescales the noise variance and the ridge regression parameter λ as well. Putting everything together we get that the asymptotic error of ridge estimator in (60) when $\mathbb{E}\mathbf{A}_{ij}^2 = \sigma_A^2/n_y$ can be found as found as follows:

1. Find the smaller solution of the quadratic equation

$$\frac{\lambda}{\sigma_A^2} = \frac{(1-\alpha)(1-\alpha/\delta)}{\alpha}. \tag{74}$$

2. Find the fixed point of state evolution

$$\tau^2 == \frac{\frac{\sigma^2}{\sigma_A^2} + \frac{1}{\delta}(1-\alpha^2)\sigma_X^2}{1 - \frac{\alpha^2}{\delta}}.$$

3. The mean squared error would the same as in (73)

$$\frac{1}{n_x}\left\|\widehat{\mathbf{x}}_{\text{ridge}} - \mathbf{x}_0\right\|_2^2 = \mathbb{E}\left[(\alpha(X_0 + \tau Z) - X_0)^2\right]$$

$$= (\alpha - 1)^2\mathbb{E}X_0^2 + \alpha^2\tau^2.$$

D.2. Convergence of AMP

As mentioned in the previous section, when we use AMP to find the solution of ridge regression, we first need to find an α that satisfies Equation (64). This is a quadratic equation that has two solutions. In theory, the solution of ridge regression with a given λ is the fixed points of AMP iterations for both values of α. However, we should also note that the results of AMP are only valid if the iterations converge to a fixed point. This is equivalent to stability of the dynamics corresponding to AMP recursion. We saw in Lemma 7 that a linear denoiser $\eta_t(\mathbf{x}) = \alpha\mathbf{x}$ can be used to solve for a ridge regression with regularization parameter λ. Recall that the AMP iterations for this denoiser are

$$\mathbf{x}^{t+1} = \alpha(\mathbf{A}^\mathsf{T}\mathbf{z}^t + \mathbf{x}^t) \tag{75}$$

$$\mathbf{z}^t = \mathbf{y} - \mathbf{A}\mathbf{x}^t + \frac{\alpha}{\delta}\mathbf{z}^{t-1}. \tag{76}$$

Plugging Equation (76) in Equation (75) we get

$$\mathbf{x}^{t+1} = \alpha(\mathbf{I} - \mathbf{A}^\mathsf{T}\mathbf{A})\mathbf{x}^t + \frac{\alpha^2}{\delta}\mathbf{z}^{t-1} + \alpha\mathbf{A}^\mathsf{T}\mathbf{y}, \tag{77}$$

$$\mathbf{z}^t = \mathbf{y} - \mathbf{A}\mathbf{x}^t + \frac{\alpha}{\delta}\mathbf{z}^{t-1}. \tag{78}$$

These equations correspond to a linear time invariant system with state matrix

$$\mathcal{A} = \begin{bmatrix} \alpha(\mathbf{I} - \mathbf{A}^\mathsf{T}\mathbf{A}) & \frac{\alpha^2}{\delta}\mathbf{A}^\mathsf{T} \\ -\mathbf{A} & \frac{\alpha}{\delta}\mathbf{I} \end{bmatrix}. \tag{79}$$

The system is stable if and only if all the eigenvalues of $\mathcal{A}$ lie inside the unit circle. We use results from random matrix theory and classical control theory to show that as $n_x \to \infty$, this system is almost surely stable. Let $[\mathbf{u}^\mathsf{T}, \mathbf{v}^\mathsf{T}]^\mathsf{T}$ be an eigenvector of $\mathcal{A}$ corresponding to the eigenvalue ρ

$$\begin{bmatrix} \alpha(\mathbf{I} - \mathbf{A}^\mathsf{T}\mathbf{A}) & \frac{\alpha^2}{\delta}\mathbf{A}^\mathsf{T} \\ -\mathbf{A} & \frac{\alpha}{\delta}\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \rho \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}. \tag{80}$$

From this we get

$$\mathbf{v} = \frac{1}{\alpha/\delta - \rho}\mathbf{A}\mathbf{u}.$$

Plugging this back in (80) we obtain

$$\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{u} = \frac{(\rho - \alpha)(\alpha/\delta - \rho)}{\alpha\rho}\mathbf{u}. \tag{81}$$

Therefore, $\mathbf{u}$ should be an eigenvector of $\mathbf{A}^\mathsf{T}\mathbf{A}$. Large random matrices of the form $\mathbf{A}^\mathsf{T}\mathbf{A}$ are well studied objects in random matrix theory and a lot of their properties are known. In particular, the empirical distribution of the eigenvalues this matrix converges to the Marchenko-Pastur distribution (see (Chafaı et al., 2009) for a brief overview) and in the limit of $n_x \to \infty$ we have

$$\lambda_{\max}(\mathbf{A}^\mathsf{T}\mathbf{A}) = \left(1 + \frac{1}{\sqrt{\delta}}\right)^2, \quad \text{almost surely,} \tag{82}$$

and

$$\lambda_{\min}(\mathbf{A}^\mathsf{T}\mathbf{A}) = \begin{cases} \left(1 - \frac{1}{\sqrt{\delta}}\right)^2, & \delta \geq 1 \\ 0, & \delta < 1 \end{cases}, \quad \text{almost surely.} \tag{83}$$

Therefore, the question of convergence of AMP algorithm reduces to showing that for all $\lambda_{\min}(\mathbf{A}^\mathsf{T}\mathbf{A}) \leq \mu \leq \lambda_{\max}(\mathbf{A}^\mathsf{T}\mathbf{A})$, whether the ρ satisfying the equation

$$\mu = \frac{(\rho - \alpha)(\alpha/\delta - \rho)}{\alpha\rho}$$

also satisfies $|\rho| < 1$. We can rearrange this equation to get

$$\rho^2 + \alpha(\mu - 1/\delta - 1) + \alpha^2/\delta = 0.$$

Applying Jury stability criterion (Jury, 1963), we see that roots of this quadratic equation lie inside the unit circle if and only if the following two conditions are satisfied

$$1 - \frac{\alpha^4}{\delta^2} > 0, \tag{Jury 1}$$

$$\left(1 + \frac{\alpha^2}{\delta}\right)^2 \geq \alpha^2\left(\mu - 1/\delta - 1\right)^2. \tag{Jury 2}$$

If regularization parameter $\lambda \geq 0$, solving the quadratic equation (64) (or similarly (74)) for α, it is not hard to show that it has two solutions α_1, α_2 that are always real and satisfy

$$0 < \alpha_1 \leq \min(1, \delta) \leq \max(1, \delta) \leq \alpha_2. \tag{84}$$

Clearly, α_1 satisfies the condition in (Jury 1) and α_2 fails this criterion. Therefore, for α_2 the AMP recursion is always unstable. It remains to show that α_1 also satisfies (Jury 2) and hence makes the AMP recursion for ridge regression stable.

First observe that the condition in (Jury 2) can be rewritten as

$$-\frac{1}{\alpha} - \frac{\alpha}{\delta} \leq \mu - 1 - \frac{1}{\delta} \leq \frac{1}{\alpha} + \frac{\alpha}{\delta}. \tag{85}$$

The upper bound holds because using (82)

$$\begin{aligned} \mu - 1 - \frac{1}{\delta} &\leq \left(1 + \frac{1}{\sqrt{\delta}}\right)^2 - 1 - \frac{1}{\delta} \\ &= \frac{2}{\sqrt{\delta}} \\ &\leq \frac{2}{\sqrt{\delta}} + \left(\frac{1}{\sqrt{\alpha}} - \frac{\sqrt{\alpha}}{\sqrt{\delta}}\right)^2 \\ &= \frac{1}{\alpha} + \frac{\alpha}{\delta}, \end{aligned}$$

where the first inequality holds almost surely. For $\delta \geq 1$, the lower bound can also be shown to hold similarly. When

$\delta < 1$ we have $\lambda_{\min}(\mathbf{A}^\mathsf{T}\mathbf{A}) = 0$ almost surely. Therefore, in this case we have to show that

$$0 \leq \frac{1}{\alpha} + \frac{\alpha}{\delta} - 1 - \frac{1}{\delta}.$$

The right hand side can be rewritten as

$$\frac{1}{\alpha} + \frac{\alpha}{\delta} - 1 - \frac{1}{\delta} = \frac{\alpha^2 - \alpha - \alpha\delta + \delta}{\alpha\delta}.$$

Since both α and δ are positive, we need to show that the numerator is always positive. The numerator is a quadratic function that attains its minimum at $\alpha = (\delta + 1)/2$. But for $\delta < 1$ from (84) we have $\alpha \leq \delta < 1$. Therefore, the minimum of the numerator for such α is attained at $\alpha = \delta$ which proves that

$$\alpha^2 - \alpha - \alpha\delta + \delta \geq 0.$$

Hence, for α_1, the AMP recursion is almost surely stable and converges.

As a sanity check, we can also verify that if AMP iterations for ridge regression in (75) and (76) are stable, so is the state evolution recursion. The state evolution for ridge regression is given in (69). This is a scalar linear time invariant system that is stable if and only if

$$-1 \leq \frac{\alpha^2}{\delta} \leq 1. \tag{86}$$

This is similar to (Jury 1) and clearly (84) implies that α_1 satisfies this inequality. Therefore, the stability of AMP recursions for ridge regression also implies the stability of the state evolution for ridge regression. As a result, the smaller value of α that satisfies (64) should be used to get the correct prediction of error.

D.3. AMP for complex ridge regression

Approximate message passing can also be used when the signals in (54) are complex valued. So long as the sensing matrix has i.i.d. complex normal entries $\mathbf{A}_{ij} \sim \mathcal{CN}(0, \sigma_A^2/n_y)$ (see Appendix A for a brief overview of complex normal distribution), i.e. the real and imaginary parts of each entry are i.i.d. Gaussian random variables with variance $\sigma_A^2/(2n_y)$ and independent of each other, the state evolution holds (Maleki et al., 2013). Therefore, by changing all variables to complex variables, we can use AMP exactly as in Appendix D.1 and get the asymptotic error of complex ridge regression using the state evolution almost without any changes.

E. Experiment with Gaussian AR(1) Process

As mentioned in the experiments, for an AR(1) process as in (37), the auto-correlation function derived in Equation (40) does not depend on the distribution of the noise ξ_t, but only its second moment. This is true in general for an AR(p) process that evolves as a linear time-invariant (LTI) system driven with zero-mean i.i.d. noise. For such processes the auto-correlation only depends on the second order statistics of the noise as well parameters of the linear system. Therefore, we expect to get identical results in the limit if the any zero mean noise is driving the process so long as the variances match. In Figure 2, we showed the results for the case where the noise was a scaled Rademacher random variable. Figure 3 shows the same results for the case where the noise is Gaussian with the matched variance. As expected, this plot is almost indistinguishable from Figure 2.

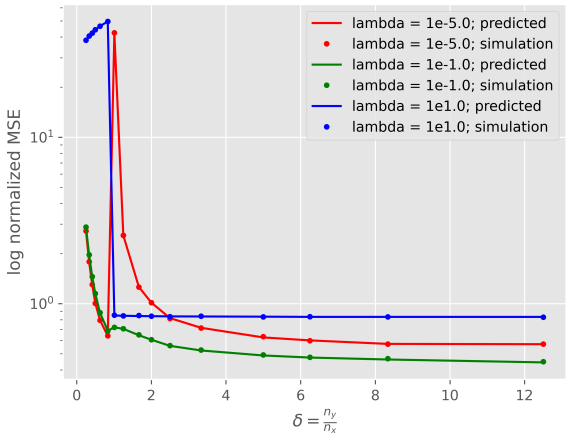

Figure 3: Log of normalized error for the AR(1) features with the process noise $\mathcal{N}(0, s^2)$, with respect to $\delta = n_y/n_x$ for three different values of λ. The figure is almost indistinguishable from Figure 2.