A. Empirical Convergence of Vector Sequences

The LSL model in Section 2 and our main result in Section 4 require certain technical definitions.

Definition 1 (Pseudo-Lipschitz continuity)**.** For a given $p \geq 1$, a function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ is called Pseudo-Lipschitz of order p if

$$\|\mathbf{f}(\boldsymbol{x}_1) - \mathbf{f}(\boldsymbol{x}_2)\| \leq C\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \left(1 + \|\boldsymbol{x}_1\|^{p-1} + \|\boldsymbol{x}_2\|^{p-1}\right) \quad (36)$$

for some constant $C > 0$.

Observe that for $p = 1$, the pseudo-Lipschitz is equivalent to the standard definition of Lipschitz continuity.

Definition 2 (Uniform Lipschitz continuity)**.** Let $\boldsymbol{\phi}(\boldsymbol{x}, \theta)$ be a function on $\mathbf{r} \in \mathbb{R}^d$ and $\theta \in \mathbb{R}^s$. We say that $\boldsymbol{\phi}(\boldsymbol{x}, \theta)$ is *uniformly Lipschitz continuous* in $\boldsymbol{x}$ at $\theta = \overline{\theta}$ if there exists constants $L_1, L_2 \geq 0$ and an open neighborhood U of $\overline{\theta}$ such that

$$\|\boldsymbol{\phi}(\boldsymbol{x}_1, \theta) - \boldsymbol{\phi}(\boldsymbol{x}_2, \theta)\| \leq L_1\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \quad (37)$$

for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$ and $\theta \in U$; and

$$\|\boldsymbol{\phi}(\boldsymbol{x}, \theta_1) - \boldsymbol{\phi}(\boldsymbol{x}, \theta_2)\| \leq L_2 \left(1 + \|\boldsymbol{x}\|\right) \|\theta_1 - \theta_2\|, \quad (38)$$

for all $\boldsymbol{x} \in \mathbb{R}^d$ and $\theta_1, \theta_2 \in U$.

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

(i) $\mathbb{E}\|X\|_p^p < \infty$; and

(ii) for any $f : \mathbb{R}^d \to \mathbb{R}$ that is pseudo-Lipschitz continuous of order p,

$$\lim_{N\to\infty} \tfrac{1}{N} \sum_{n=1}^{N} f(\boldsymbol{x}_n(N)) = \mathbb{E}\left[f(X)\right]. \quad (39)$$

In this case, with some abuse of notation, we will write

$$\lim_{N\to\infty} \{\boldsymbol{x}_n\} \overset{PL(p)}{=} X, \quad (40)$$

where we have omitted the dependence on N in $\boldsymbol{x}_n(N)$. We note that the sequence $\{\boldsymbol{x}(N)\}$ can be random or deterministic. If it is random, we will require that for every pseudo-Lipschitz function $f(\cdot)$, the limit (39) holds almost surely. In particular, if $\boldsymbol{x}_n \sim X$ are i.i.d. and $\mathbb{E}\|X\|_p^p < \infty$, then $\boldsymbol{x}$ empirically converges to X with p^{th} order moments.

$PL(p)$ convergence is equivalent to weak convergence plus convergence in p moment (Bayati & Montanari, 2011), and hence $PL(p)$ convergence is also equivalent to convergence in Wasserstein-p metric (See Chapter 6. (Villani, 2008)). We use this fact later in proving Theorem 1.

B. ML-VAMP Denoisers Details

Related to $\mathbf{S}_{\text{mp}}$ and $\mathbf{s}_{\text{mp}}$ from equation (11), we need to define two quantities $\mathbf{s}_{\text{mp}}^+ \in \mathbb{R}^N$ and $\mathbf{s}_{\text{mp}}^- \in \mathbb{R}^p$ that are zero-padded versions of the singular values $\mathbf{s}_{\text{mp}}$, so that for $n > \min\{N, p\}$, we set $s_{\text{mp},n}^{\pm} = 0$. Observe that $(\mathbf{s}_{\text{mp}}^+)^2$ are eigenvalues of $\mathbf{U}\mathbf{U}^{\mathsf{T}}$ whereas $(\mathbf{s}_{\text{mp}}^-)^2$ are eigenvalues of $\mathbf{U}^{\mathsf{T}}\mathbf{U}$. Since $\mathbf{s}_{\text{mp}}$ empirically converges to S_{mp} as given in (12), the vector $\mathbf{s}_{\text{mp}}^+$ empirically converges to random variable S_{mp}^+ whereas the vector $\mathbf{s}_{\text{mp}}^-$ empirically converges to random variable S_{mp}^-, where a mass is placed at 0 appropriately. Specifically, S_{mp}^+ has a point mass of $(1-\beta)_+\delta_{\{0\}}$ when $\beta < 1$, whereas S_{mp}^- has a point mass of $(1-\frac{1}{\beta})_+\delta_{\{0\}}$, when $\beta > 1$. In Appendix H (eqn. (113)), we provide the densities over positive parts of S_{mp}^+ and S_{mp}^-.

A key property of our analysis will be that the non-linear functions (20) and the denoisers $\mathbf{g}_\ell^{\pm}(\cdot)$ have simple forms.

Non-linear functions $\boldsymbol{\phi}_\ell(\cdot)$: The non-linear functions all act *componentwise*. For example, for $\boldsymbol{\phi}_1(\cdot)$ in (20), we have

$$\mathbf{z}_1 = \boldsymbol{\phi}_1(\mathbf{p}_0, \mathbf{s}_{\text{tr}}) = \text{diag}(\mathbf{s}_{\text{tr}})\mathbf{p}_0 \iff z_{1,n} = \phi_1(p_{0,n}, s_{\text{tr},n}),$$

where $\phi_1(\cdot)$ is the scalar-valued function,

$$\phi_1(p_0, s) = sp_0. \quad (41)$$

Similarly, for $\boldsymbol{\phi}_2(\cdot)$,

$$\mathbf{z}_2 = \boldsymbol{\phi}_2(\mathbf{p}_1, \mathbf{s}_{\text{mp}}^+) \iff z_{2,n} = \phi_2(\overline{p}_{1,n}, s_{\text{mp},n}^+), \quad n < N$$

where $\overline{\mathbf{p}}_1 \in \mathbb{R}^N$ is the zero-padded version of $\mathbf{p}_1$, and

$$\phi_2(p_1, s) = s\, p_1. \quad (42)$$

Finally, the function $\boldsymbol{\phi}_3(\cdot)$ in (20) acts componentwise with

$$\phi_3(p_2, d) = \phi_{\text{out}}(p_2, d). \quad (43)$$

Input denoiser $\mathbf{g}_0^+(\cdot)$: Since $F_0(\mathbf{z}_0) = F_{\text{in}}(\mathbf{z}_0)$, and $F_{\text{in}}(\cdot)$ given in (6), the denoiser (25a) acts *componentwise* in that,

$$\widehat{\mathbf{z}}_0 = \mathbf{g}_0^+(\mathbf{r}_0^-, \gamma_0^-) \iff \widehat{z}_{0,n} = g_0^+(r_{0,n}^-, \gamma_0^-),$$

where $g_0^+(\cdot)$ is the scalar-valued function,

$$g_0^+(r_0^-, \gamma_0^-) := \underset{z_0}{\text{argmin}}\ f_{\text{in}}(z_0) + \frac{\gamma_0^-}{2}(z_0 - r_0^-)^2. \quad (44)$$

Thus, the vector optimization in (25a) reduces to a set of scalar optimizations (44) on each component.

Output denoiser $\mathbf{g}_3^-(\cdot)$: The output penalty $F_3(\mathbf{p}_2, \boldsymbol{y}) = F_{\rm out}(\mathbf{p}_2, \boldsymbol{y})$ where $F_{\rm out}(\mathbf{p}_2, \boldsymbol{y})$ has the separable form (6). Thus, similar to the case of $\mathbf{g}_0(\cdot)$, the denoiser $\mathbf{g}_3(\cdot)$ in (25b) also acts componentwise with the function,

$$g_3^-(r_2^+, \gamma_2^+, y) := \operatorname*{argmin}_{p_2} \; f_{\rm out}(p_2, y) + \tfrac{\gamma_2^+}{2}(p_2 - r_2^+)^2. \tag{45}$$

Linear denoiser $\mathbf{g}_1^{\pm}(\cdot)$: The expressions for both denoisers $g_1^{\pm}$ and $g_2^{\pm}$ are very similar and can be explained together. The penalty $F_1(\cdot)$ restricts $\mathbf{z}_1 = \mathbf{S}_{\rm tr}\mathbf{p}_0$, where $\mathbf{S}_{\rm tr}$ is a square matrix. Hence, for $\ell = 1$, the minimization in (27) is given by,

$$\widehat{\mathbf{p}}_0 := \operatorname*{argmin}_{\mathbf{p}_0} \; \tfrac{\gamma_0^+}{2}\|\mathbf{p}_0 - \mathbf{r}_0^+\|^2 + \tfrac{\gamma_1^-}{2}\|\mathbf{S}_{\rm tr}\mathbf{p}_0 - \mathbf{r}_1^-\|^2, \tag{46}$$

and $\widehat{\mathbf{z}}_1 = \mathbf{S}_{\rm tr}\widehat{\mathbf{p}}_0$. This is a simple quadratic minimization and the components of $\widehat{\mathbf{p}}_0$ and $\widehat{\mathbf{z}}_1$ are given by

$$\begin{aligned}\widehat{p}_{0,n} &= g_1^-(r_{0,n}^+, r_{1,n}^-, \gamma_0^+, \gamma_1^-, s_{{\rm tr},n})\\ \widehat{z}_{1,n} &= g_1^+(r_{0,n}^+, r_{1,n}^-, \gamma_0^+, \gamma_1^-, s_{{\rm tr},n}),\end{aligned}$$

where

$$g_1^-(r_0^+, r_1^-, \gamma_0^+, \gamma_1^-, s) := \frac{\gamma_0^+ r_0^+ + s\gamma_1^- r_1^-}{\gamma_0^+ + s^2\gamma_1^-} \tag{47a}$$

$$g_1^+(r_0^+, r_1^-, \gamma_0^+, \gamma_1^-, s) := \frac{s(\gamma_0^+ r_0^+ + s\gamma_1^- r_1^-)}{\gamma_0^+ + s^2\gamma_1^-} \tag{47b}$$

Linear denoiser $\mathbf{g}_2^{\pm}(\cdot)$: This denoiser is identical to the case $\mathbf{g}_1^{\pm}(\cdot)$ in that we need to impose the linear constraint $\mathbf{z}_2 = \mathbf{S}_{\rm mp}\mathbf{p}_1$. However $\mathbf{S}_{\rm mp}$ is in general a rectangular matrix and the two resulting cases of $\beta \lessgtr 1$ needs to be treated separately.

Recall the definitions of vectors $\mathbf{s}_{\rm mp}^+$ and $\mathbf{s}_{\rm mp}^-$ at the beginning of this section. Then, for $\ell = 2$, with the penalty $F_2(\mathbf{p}_1, \mathbf{z}_2) = \delta_{\{\mathbf{z}_2 = \mathbf{S}_{\rm mp}\mathbf{p}_1\}}$, the solution to (27) has components,

$$\widehat{p}_{1,n} = g_2^-(r_{1,n}^+, r_{2,n}^-, \gamma_1^+, \gamma_2^+, s_{{\rm mp},n}^-) \tag{48a}$$

$$\widehat{z}_{2,n} = g_2^+(r_{1,n}^+, r_{2,n}^-, \gamma_1^+, \gamma_2^+, s_{{\rm mp},n}^+), \tag{48b}$$

with the identical functions $g_2^- = g_1^-$ and $g_2^+ = g_1^+$ as given by (47a) and (47b). Note that in (48a), $n = 1, \ldots, p$ and in (48b), $n = 1, \ldots, N$.

C. State Evolution Analysis of ML-VAMP

A key property of the ML-VAMP algorithm is that its performance in the LSL can be exactly described by a *scalar equivalent system*. In the scalar equivalent system, the vector-valued outputs of the algorithm are replaced by scalar random variables representing the typical behavior of the components of the vectors in the large-scale-limit (LSL). Each of the random variables are described by a set of parameters, where the parameters are given by a set of deterministic equations called the *state evolution* or SE.

The SE for the general ML-VAMP algorithm are derived in (Pandit et al., 2019) and the special case of the updates for ML-VAMP for GLM learning are shown in Algorithm 2 with details of functions $\mathbf{g}_\ell^{\pm}$ in Appendix B. We see that the SE updates in Algorithm 2 parallel those in the ML-VAMP algorithm Algo. 1, except that vector quantities such as $\widehat{\mathbf{z}}_{k\ell}$, $\widehat{\mathbf{p}}_{k\ell}$, $\mathbf{r}_{k\ell}^+$ and $\mathbf{r}_{k\ell}^-$ are replaced by scalar random variables such as $\widehat{Z}_{k\ell}$, $\widehat{P}_{k\ell}$, $R_{k\ell}^+$ and $R_{k\ell}^-$. Each of these random variables are described by the deterministic parameters such as $\mathbf{K}_{k\ell} \in \mathbb{R}_{\succ 0}^{2\times 2}$, and $\tau_\ell^0, \tau_{k\ell}^- \in \mathbb{R}_+$.

The updates in the section labeled as "Initial", provide the scalar equivalent model for the true system (18). In these updates, Ξ_ℓ represent the limits of the vectors $\boldsymbol{\xi}_\ell$ in (19). That is,

$$\Xi_1 := S_{\rm tr}, \quad \Xi_2 := S_{\rm mp}^+, \quad \Xi_3 := D. \tag{49}$$

Due to assumptions in Section 2, we have that the components of $\boldsymbol{\xi}_\ell$ converge empirically as,

$$\lim_{N\to\infty}\{\xi_{\ell,i}\} \overset{PL(2)}{=} \Xi_\ell, \tag{50}$$

So, the Ξ_ℓ represent the asymptotic distribution of the components of the vectors $\boldsymbol{\xi}_\ell$.

The updates in sections labeled "Forward pass" and "Backward pass" in the SE equations in Algorithm 2 parallel those in Algorithm 1. The key quantities in these SE equations are the error variables,

$$\mathbf{p}_{k\ell}^+ := \mathbf{r}_{k\ell}^+ - \mathbf{p}_\ell^0, \quad \mathbf{q}_{k\ell}^- := \mathbf{r}_{k\ell}^- - \mathbf{z}_\ell^0,$$

which represent the errors of the estimates to the inputs of the denoisers. We will also be interested in their transforms,

$$\mathbf{q}_{k\ell}^+ = \mathbf{V}_\ell^{\mathsf{T}}\mathbf{p}_{k,\ell+1}^+, \quad \mathbf{p}_{k\ell}^- = \mathbf{V}_\ell\mathbf{q}_{k\ell}^-.$$

The following Theorem is an adapted version of the main result from (Pandit et al., 2019) to the iterates of Algorithms 1 and 2.

Theorem 2. *Consider the outputs of the ML-VAMP for GLM Learning Algorithm under the assumptions of Section 2. Assume the denoisers satisfy the continuity conditions in Assumption 1. Also, assume that the outputs of the SE satisfy*

$$\overline{\alpha}_{k\ell}^{\pm} \in (0, 1),$$

Algorithm 2 SE for ML-VAMP for GLM Learning

// `Initial`
Initialize $\overline{\gamma}_{0\ell}^- = \gamma_{0\ell}^-$ from Algorithm 1.
$Q_{0\ell}^- \sim \mathcal{N}(0, \tau_{0\ell}^-)$ for some $\tau_{0\ell}^- > 0$ for $\ell = 0, 1, 2$
$Z_0^0 = W^0$
for $\ell = 0, \ldots, L-1$ **do**
  $P_\ell^0 = \mathcal{N}(0, \tau_\ell^0), \quad \tau_\ell^0 = \mathrm{var}(Z_\ell^0)$
  $Z_{\ell+1}^0 = \phi_{\ell+1}(P_\ell^0, \Xi_{\ell+1})$
end for

for $k = 0, 1, \ldots$ **do**
  // `Forward Pass`
  for $\ell = 0, \ldots, L-1$ **do**
    if $\ell = 0$ **then**
      $R_{k0}^- = Z_\ell^0 + Q_{k0}^-$
      $\widehat{Z}_{k0} = g_0^+(R_{k0}^-, \overline{\gamma}_{k0}^-)$
    else
      $R_{k,\ell-1}^+ = P_{\ell-1}^0 + P_{k,\ell-1}^+, R_{k\ell}^- = Z_\ell^0 + Q_{k\ell}^-$
      $\widehat{Z}_{k\ell} = g_\ell^+(R_{k,\ell-1}^+, R_{k\ell}^-, \overline{\gamma}_{k,\ell-1}^+, \overline{\gamma}_{k\ell}^-, \Xi_\ell)$
    end if
    $\overline{\alpha}_{k\ell}^+ = \mathbb{E}\partial\widehat{Z}_{k\ell}/\partial Q_{k\ell}^-$
    $Q_{k\ell}^+ = \dfrac{\widehat{Z}_{k\ell} - Z_\ell^0 - \overline{\alpha}_{k\ell}^+ Q_{k\ell}^-}{1 - \overline{\alpha}_{k\ell}^+}$
    $\overline{\gamma}_{k\ell}^+ = (\frac{1}{\overline{\alpha}_{k\ell}^+} - 1)\overline{\gamma}_{k\ell}^-$
    $(P_\ell^0, P_{k\ell}^+) \sim \mathcal{N}(0, \mathbf{K}_{k\ell}^+), \ \mathbf{K}_{k\ell}^+ = \mathrm{cov}(Z_\ell^0, Q_{k\ell}^+)$
  end for

  // `Backward Pass`
  for $\ell = L, \ldots, 1$ **do**
    if $\ell = L$ **then**
      $R_{k,L-1}^+ = P_{L-1}^0 + P_{k,L-1}^+$
      $\widehat{P}_{k,L-1} = g_L^-(R_{k,L-1}^+, \overline{\gamma}_{k,L-1}^+, Z_L^0)$
    else
      $R_{k,\ell-1}^+ = P_{\ell-1}^0 + P_{k,\ell-1}^+, R_{k+1,\ell}^- = Z_\ell^0 + Q_{k+1,\ell}^-$
      $\widehat{P}_{k,\ell-1} = g_\ell^-(R_{k,\ell-1}^+, R_{k+1,\ell}^-, \overline{\gamma}_{k,\ell-1}^+, \overline{\gamma}_{k+1,\ell}^-, \Xi_\ell)$
    end if
    $\overline{\alpha}_{k,\ell-1}^- = \mathbb{E}\partial\widehat{P}_{k,\ell-1}/\partial P_{k,\ell-1}^+$
    $P_{k+1,\ell-1}^- = \dfrac{\widehat{P}_{k,\ell-1} - P_{\ell-1}^0 - \overline{\alpha}_{k,\ell-1}^- P_{k,\ell-1}^+}{1 - \overline{\alpha}_{k,\ell-1}^-}$
    $\overline{\gamma}_{k+1,\ell-1}^- = (\frac{1}{\overline{\alpha}_{k,\ell-1}^-} - 1)\overline{\gamma}_{k,\ell-1}^+$
    $Q_{k+1,\ell-1}^- \sim \mathcal{N}(0, \tau_{k+1,\ell-1}^-), \tau_{k,\ell-1}^- = \mathbb{E}(P_{k+1,\ell-1}^-)^2$
  end for
end for

for all k and ℓ. Suppose Algo. 1 is initialized so that the following convergence holds

$$\lim_{N\to\infty} \{\mathbf{r}_{0\ell}^- - \mathbf{z}_\ell^0\} \overset{PL(2)}{=} Q_{0\ell}^-$$

where $(Q_{00}^-, Q_{01}^-, Q_{02}^-)$ are independent zero-mean Gaussians, independent of $\{\Xi_\ell\}$. Then,

(a) For any fixed iteration $k \geq 0$ in the forward direction and layer $\ell = 1, \ldots, L-1$, we have that, almost surely,

$$\lim_{N\to\infty} (\gamma_{k,\ell-1}^+, \gamma_{k\ell}^-) = (\overline{\gamma}_{k,\ell-1}^+, \overline{\gamma}_{k\ell}^-), \quad \text{and,} \tag{51}$$

$$\lim_{N\to\infty} \{(\widehat{\mathbf{z}}_{k\ell}^+, \mathbf{z}_\ell^0, \mathbf{p}_{\ell-1}^0, \mathbf{r}_{k,\ell-1}^+, \mathbf{r}_\ell^-)\} \overset{PL(2)}{=} (\widehat{Z}_{k\ell}^+, Z_\ell^0, P_{\ell-1}^0, R_{k,\ell-1}^+, R_\ell^-) \tag{52}$$

where the variables on the right-hand side are from the SE equations (51) and (52) are the outputs of the SE equations in Algorithm 2. Similar equations hold for $\ell = 0$ with the appropriate variables removed.

(b) Similarly, in the reverse direction, For any fixed iteration $k \geq 0$ and layer $\ell = 1, \ldots, L-2$, we have that, almost surely,

$$\lim_{N\to\infty} (\gamma_{k,\ell-1}^+, \gamma_{k+1,\ell}^-) = (\overline{\gamma}_{k,\ell-1}^+, \overline{\gamma}_{k+1,\ell}^-), \quad \text{and} \tag{53}$$

$$\lim_{N\to\infty} \{(\widehat{\mathbf{p}}_{k+1,\ell-1}^+, \mathbf{z}_\ell^0, \mathbf{p}_{\ell-1}^0, \mathbf{r}_{k,\ell-1}^+, \mathbf{r}_{k+1,\ell}^-)\} \overset{PL(2)}{=} (\widehat{P}_{k+1,\ell-1}^+, Z_\ell^0, P_{\ell-1}^0, R_{k,\ell-1}^+, R_{k+1,\ell}^-). \tag{54}$$

Furthermore, $(P_{\ell-1}^0, P_{k\ell-1}^+)$ and $Q_{k\ell}^-$ are independent.

Proof. This is a direct application of Theorem 3 from (Pandit et al., 2019) to the iterations of Algorithm 1. The convergence result in (Pandit et al., 2019) requires the uniform Lipschitz continuity of functions $\mathbf{g}_\ell^\pm(\cdot)$. Assumption 1 provides the required uniform Lipschitz continuity assumption on $\mathbf{g}_0^+(\cdot)$ and $\mathbf{g}_3^-(\cdot)$. For the "middle" layers, $\ell = 1, 2$, the denoisers $\mathbf{g}_\ell^\pm(\cdot)$ are linear and the uniform continuity assumption is valid since we have made the additional assumption that the terms $\mathbf{s}_{\mathrm{tr}}$ and $\mathbf{s}_{\mathrm{mp}}$ are bounded almost surely. □

A key use of the Theorem is to compute asymptotic empirical limits. Specifically, for a componentwise function $\psi(\cdot)$, let $\langle\psi(\boldsymbol{x})\rangle$ denotes the average $\frac{1}{N}\sum_{n=1}^N \psi(x_n)$ The above theorem then states that for any componentwise pseudo-Lipschitz function $\psi(\cdot)$ of order 2, as $N \to \infty$, we have the following two properties

$$\lim_{N\to\infty} \langle\psi(\widehat{\mathbf{z}}_{k\ell}^+, \mathbf{z}_\ell^0, \mathbf{p}_{\ell-1}^0, \mathbf{r}_{k,\ell-1}^+, \mathbf{r}_\ell^-)\rangle = \mathbb{E}\,\psi(\widehat{Z}_{k\ell}^+, Z_\ell^0, P_{\ell-1}^0, R_{k,\ell-1}^+, R_\ell^-)$$

$$\lim_{N\to\infty} \langle\psi(\widehat{\mathbf{p}}_{k+1,\ell-1}^+, \mathbf{z}_\ell^0, \mathbf{p}_{\ell-1}^0, \mathbf{r}_{k,\ell-1}^+, \mathbf{r}_{k+1,\ell)}^-)\rangle = \mathbb{E}\,\psi(\widehat{P}_{k+1,\ell-1}^+, Z_\ell^0, P_{\ell-1}^0, R_{k,\ell-1}^+, R_{k+1,\ell}^-).$$

That is, we can compute the empirical average over components with the expected value of the random variable limit. This convergence is key to proving Theorem 1.

D. Empirical Convergence of Fixed Points

A consequence of Assumption 2 is that we can take the limit $k \to \infty$ of the random variables in the SE algorithm. Specifically, let $\boldsymbol{x}_k = \boldsymbol{x}_k(N)$ be any set of d outputs from the ML-VAMP for GLM Learning Algorithm under the assumptions of Theorem 2. Under Assumption 2, for each N, there exists a vector

$$\boldsymbol{x}(N) = \lim_{k\to\infty} \boldsymbol{x}_k(N), \tag{55}$$

representing the limit over k. For each k, Theorem 2 shows there also exists a random vector limit,

$$\lim_{N\to\infty} \{\boldsymbol{x}_{k,i}(N)\} \overset{PL(2)}{=} X_k, \tag{56}$$

representing the limit over N. The following proposition shows that we can take the limits of the random variables X_k.

Proposition 1. *Consider the outputs of the ML-VAMP for GLM Learning Algorithm under the assumptions of Theorem 2 and Assumption 2. Let $\boldsymbol{x}_k = \boldsymbol{x}_k(N)$ be any set of d outputs from the algorithm and let $\boldsymbol{x}(N)$ be its limit from (55) and let X_k be the random variable limit (56). Then, there exists a random variable $X \in \mathbb{R}^d$ such that, for any pseudo-Lipschitz continuous $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\lim_{k\to\infty} \mathbb{E} f(X_k) = \mathbb{E} f(X) = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} f(x_i(N)). \tag{57}$$

In addition, the SE parameter limits $\overline{\alpha}_{k\ell}^{\pm}$ and $\overline{\gamma}_{k\ell}^{\pm}$ converge to limits,

$$\lim_{k\to\infty} \overline{\alpha}_{k\ell}^{\pm} = \overline{\alpha}_{\ell}^{\pm}, \quad \lim_{k\to\infty} \overline{\gamma}_{k\ell}^{\pm} = \overline{\gamma}_{\ell}^{\pm}. \tag{58}$$

The proposition shows that, under the convergence assumption, Assumption 2, we can take the limits as $k \to \infty$ of the random variables from the SE. To prove the proposition we first need the following simple lemma.

Lemma 1. *If α_N and $\beta_k \in \mathbb{R}$ are sequences such that*

$$\lim_{k\to\infty} \lim_{N\to\infty} |\alpha_N - \beta_k| = 0, \tag{59}$$

then, there exists a constant C such that,

$$\lim_{N\to\infty} \alpha_N = \lim_{k\to\infty} \beta_k = C. \tag{60}$$

In particular, the two limits in (60) exist.

Proof. For any $\epsilon > 0$, the limit (59) implies that there exists a $k_\epsilon (\uparrow \infty$ as $\epsilon \downarrow 0)$ such that for all $k > k_\epsilon$,

$$\lim_{N\to\infty} |\alpha_N - \beta_k| < \epsilon,$$

from which we can conclude,

$$\liminf_{N\to\infty} \alpha_N > \beta_k - \epsilon$$

for all $k > k_\epsilon$. Therefore,

$$\liminf_{N\to\infty} \alpha_N \geq \sup_{k\geq k_\epsilon} \beta_k - \epsilon.$$

Since this is true for all $\epsilon > 0$, it follows that

$$\liminf_{N\to\infty} \alpha_N \geq \limsup_{k\to\infty} \beta_k. \tag{61}$$

Similarly, $\limsup_{N\to\infty} \alpha_N \leq \inf_{k>k_\epsilon} \beta_k + \epsilon$, whereby

$$\limsup_{N\to\infty} \alpha_N \leq \liminf_{k\to\infty} \beta_k. \tag{62}$$

Equations (61) and (62) together show that the limits in (60) exists and are equal. □

Proof of Proposition 1 Let $f : \mathbb{R}^d \to \mathbb{R}$ be any pseudo-Lipschitz function of order 2, and define,

$$\alpha_N = \frac{1}{N}\sum_{i=1}^{N} f(x_i(N)), \quad \beta_k = \mathbb{E} f(X_k). \tag{63}$$

Their difference can be written as,

$$\alpha_N - \beta_k = A_{N,k} + B_{N,k}, \tag{64}$$

where

$$A_{N,k} := \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{x}_i(N)) - f(\boldsymbol{x}_{k,i}(N)), \tag{65}$$

$$B_{N,k} := \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{x}_{k,i}(N)) - \mathbb{E} f(X_k). \tag{66}$$

Since $\{x_{k,i}(N)\}$ converges $PL(2)$ to X_k, we have,

$$\lim_{N\to\infty} B_{N,k} = 0. \tag{67}$$

For the term $A_{N,k}$,

$$\begin{aligned}
|A_{N,k}| &\overset{(a)}{\leq} \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} |f(\boldsymbol{x}_i(N)) - f(\boldsymbol{x}_{k,i}(N))| \\
&\overset{(b)}{\leq} \lim_{N\to\infty} \frac{C}{N}\sum_{i=1}^{N} a_{ki}(N)(1 + a_{ki}(N)) \\
&\overset{(c)}{\leq} C \lim_{N\to\infty} \sqrt{\frac{1}{N}\sum_{i=1}^{N} a_{ki}^2(N)} + \frac{1}{N}\sum_{i=1}^{N} a_{ki}^2(N) \\
&= C \lim_{N\to\infty} \epsilon_k(N)(1 + \epsilon_k(N)),
\end{aligned} \tag{68}$$

where (a) follows from applying the triangle inequality to the definition of $A_{N,k}$ in (65); (b) follows from the definition of pseudo-Lipschitz continuity in Definition 1, $C > 0$ is the Lipschitz contant and

$$a_{ki}(N) := \|\boldsymbol{x}_{k,i}(N) - \boldsymbol{x}_i(N)\|_2,$$

and (c) follows from the RMS-AM inequality:

$$\left(\frac{1}{N}\sum_{i=1}^{N} a_{ki}(N)\right)^2 \le \frac{1}{N}\sum_{i=1}^{N} a_{ki}^2(N) =: \epsilon_k^2(N).$$

By (29), we have that,

$$\lim_{k\to\infty}\lim_{N\to\infty} \epsilon_k(N) = 0.$$

Hence, from (68), it follows that,

$$\lim_{k\to\infty}\lim_{N\to\infty} A_{N,k} = 0. \tag{69}$$

Substituting (67) and (69) into (64) show that α_N and β_k satisfy (59). Therefore, applying Lemma 1 we have that for any pseudo-Lipschitz function $f(\cdot)$, there exists a limit $\Phi(f)$ such that,

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} f(x_i(N)) = \lim_{k\to\infty}\mathbb{E}f(X_k) = \Phi(f). \tag{70}$$

In particular, the two limits in (70) exists. When restricted to the continuous, bounded functions with the $\|f\|_\infty$ norm, it is easy verified that $\Phi(f)$ is a positive, linear, bounded function of f, with $\Phi(1) = 1$. Therefore, by the Riesz representation theorem, there exists a random variable X such that $\Phi(f) = \mathbb{E}f(X)$. This fact in combination with (70) shows (57).

It remains to prove the parameter limits in (58). We prove the result for the parameter $\overline{\alpha}_{k\ell}^+$. The proof for the other parameters are similar. Using Stein's lemma, it is shown in (Pandit et al., 2019) that

$$\overline{\alpha}_{k\ell}^+ = \frac{\mathbb{E}(\widehat{Z}_{k\ell}Q_{k\ell}^-)}{\mathbb{E}(Q_\ell^-)^2}. \tag{71}$$

Since the numerator and denominator of (71) are $PL(2)$ functions we have that the limit,

$$\begin{aligned}\overline{\alpha}_\ell^+ &:= \lim_{k\to\infty}\overline{\alpha}_{k\ell}^+ = \lim_{k\to\infty}\frac{\mathbb{E}(\widehat{Z}_{k\ell}Q_{k\ell}^-)}{\mathbb{E}(Q_{k\ell}^-)^2}\\ &= \frac{\mathbb{E}(\widehat{Z}_\ell Q_\ell^-)}{\mathbb{E}(Q_\ell^-)^2},\end{aligned} \tag{72}$$

where $\widehat{Z}_\ell$ and Q_ℓ^- are the limits of $\widehat{Z}_{k\ell}$ and $Q_{k\ell}^-$. This completes the proof. □

E. Proof of Theorem 1

From Assumption 2, we know that for every N, every group of vectors $\boldsymbol{x}_k$ converge to limits, $\boldsymbol{x} := \lim_{k\to\infty}\boldsymbol{x}_k$. The parameters, $\gamma_{k\ell}^\pm$, also converge to limits $\overline{\gamma}_\ell^\pm := \lim_{k\to\infty}\gamma_{k\ell}^\pm$ for all ℓ. By the continuity assumptions on the functions $\mathbf{g}_\ell^\pm(\cdot)$, the limits $\boldsymbol{x}$ and $\overline{\gamma}_\ell^\pm$ are fixed points of the algorithms.

A proof similar to that in (Pandit et al., 2019) shows that the fixed points $\widehat{\mathbf{z}}_\ell$ and $\widehat{\mathbf{p}}_\ell$ satisfy the KKT condition of the constrained optimization (22). This proves part (a).

The estimate $\widehat{\boldsymbol{w}}$ is the limit,

$$\widehat{\boldsymbol{w}} = \widehat{\mathbf{z}}_0 = \lim_{k\to\infty}\widehat{\mathbf{z}}_{k0}.$$

Also, the true parameter is $\mathbf{z}_0^0 = \boldsymbol{w}^0$. By Proposition 1, we have that the $PL(2)$ limits of these variables are

$$\lim_{N\to\infty}\{(\widehat{\boldsymbol{w}}, \boldsymbol{w}_0)\} \overset{PL(2)}{=} (\widehat{W}, W_0) := (\widehat{Z}_0, Z_0^0).$$

From line 2 of the SE Algorithm 2, we have

$$\widehat{W} = \widehat{Z}_0 = g_0^+(R_0^-, \overline{\gamma}_0^-) = \mathrm{prox}_{f_{\mathrm{in}}/\overline{\gamma}_0^-}(W^0 + Q_0^-).$$

This proves part (b).

To prove part (c), we use the limit

$$\lim_{N\to\infty}\{p_{0,n}^0, \widehat{p}_{0,n}\} \overset{PL(2)}{=} (P_0^0, \widehat{P}_0). \tag{73}$$

Since the fixed points are critical points of the constrained optimization (22), $\widehat{\mathbf{p}}_0 = \mathbf{V}_0\widehat{\boldsymbol{w}}$. We also have $\mathbf{p}_0^0 = \mathbf{V}_0\boldsymbol{w}^0$. Therefore,

$$\begin{aligned}\begin{bmatrix} z_{\mathrm{ts}}^{(N)} & \widehat{z}_{\mathrm{ts}}^{(N)}\end{bmatrix} &:= \mathbf{u}^\mathsf{T}\,\mathrm{Diag}(\mathbf{s}_{\mathrm{ts}})\mathbf{V}_0[\boldsymbol{w}^0\ \widehat{\boldsymbol{w}}]\\ &= \mathbf{u}^\mathsf{T}\,\mathrm{Diag}(\mathbf{s}_{\mathrm{ts}})[\mathbf{p}_0^0\ \widehat{\mathbf{p}}_0].\end{aligned} \tag{74}$$

Here, (N) in the subscript denotes the dependence on N. Since $\mathbf{u} \sim \mathcal{N}(0, \frac{1}{p}\mathbf{I})$, $[z_{\mathrm{ts}}^{(N)}\ \widehat{z}_{\mathrm{ts}}^{(N)}]$ is a zero-mean bivariate Gaussian with covariance matrix

$$\mathbf{M}^{(N)} = \tfrac{1}{p}\sum_{n=1}^{p}\begin{bmatrix} s_{\mathrm{ts},n}^2 p_{0,n}^0 p_{0,n}^0 & s_{\mathrm{ts},n}^2 p_{0,n}^0 \widehat{p}_{0,n} \\ s_{\mathrm{ts},n}^2 p_{0,n}^0 \widehat{p}_{0,n} & s_{\mathrm{ts},n}^2 \widehat{p}_{0,n}\widehat{p}_{0,n}\end{bmatrix}$$

The empirical convergence (73) yields the following limit,

$$\lim_{N\to\infty}\mathbf{M}^{(N)} = \mathbf{M} := \mathbb{E}\,S_{\mathrm{ts}}^2\begin{bmatrix} P_0^0 P_0^0 & P_0^0\widehat{P}_0 \\ P_0^0\widehat{P}_0 & \widehat{P}_0\widehat{P}_0\end{bmatrix}. \tag{75}$$

It suffices to show that the distribution of $[z_{\mathrm{ts}}^{(N)}\ \widehat{z}_{\mathrm{ts}}^{(N)}]$ converges to the distribution of $[Z_{\mathrm{ts}}\ \widehat{Z}_{\mathrm{ts}}]$ in the Wasserstein-2 metric as $N \to \infty$. (See the discussion in Appendix A on the equivalence of convergence in Wasserstein-2 metric and PL(2) convergence.)

Now, Wassestein-2 distance between between two probability measures ν_1 and ν_2 is defined as

$$W_2(\nu_1,\nu_2) = \left(\inf_{\gamma\in\Gamma}\mathbb{E}_\gamma \|X_1 - X_2\|^2\right)^{1/2}, \tag{76}$$

where Γ is the set of probability distributions on the product space with marginals consistent with ν_1 and ν_2. For Gaussian measures $\nu_1 = \mathcal{N}(\mathbf{0},\Sigma_1)$ and $\nu_2 = \mathcal{N}(\mathbf{0},\Sigma_2)$ we have (Givens et al., 1984)

$$W_2^2(\nu_1,\nu_2) = \mathrm{tr}(\Sigma_1 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2} + \Sigma_2).$$

Therefore, for Gaussian distributions $\nu_1^{(N)} = \mathcal{N}(\mathbf{0},\mathbf{M}^{(N)})$, and $\nu_2 = \mathcal{N}(\mathbf{0},\mathbf{M})$, the convergence (75) implies $W_2(\nu_1^{(N)},\nu_2)\to 0$, i.e., convergence in Wasserstein-2 distance. Hence,

$$(z_{\rm ts}^{(N)}, \widehat{z}_{\rm ts}^{(N)}) \xrightarrow{W_2} (Z_{\rm ts},\widehat{Z}_{\rm ts}) \sim \mathcal{N}(0,\mathbf{M}),$$

where $\mathbf{M}$ is the covariance matrix in (75). Hence the convergence holds in the PL(2) sense (see discussion in Appendix A on the equivalence of convergence in W_2 and PL(2) convergence).

Hence the asymptotic generalization error (17) is

$$\begin{aligned}\mathcal{E}_{\rm ts} &:= \lim_{N\to\infty}\mathbb{E} f_{\rm ts}(\widehat{y}_{\rm ts}, y_{\rm ts})\\ &\overset{(a)}{=} \lim_{N\to\infty}\mathbb{E} f_{\rm ts}(\phi_{\rm out}(z_{\rm ts}^{(N)}, D), \phi(\widehat{z}_{\rm ts}^{(N)}))\\ &\overset{(b)}{=} \mathbb{E} f_{\rm ts}(\phi_{\rm out}(Z_{\rm ts}, D), \phi(\widehat{Z}_{\rm ts})),\end{aligned} \tag{77}$$

where (a) follows from (3); and step (b) follows from continuity assumption in Assumption 1(b) along with the definition of PL(2) convergence in Def. 3. This proves part (c).

F. Formula for M

For the special cases in the next Appendix, it is useful to derive expressions for the entries the covariance matrix $\mathbf{M}$ in (75). For the term m_{11},

$$m_{11} = \mathbb{E}\, S_{\rm ts}^2 (P_0^0)^2 = \mathbb{E}\, S_{\rm ts}^2 \mathbb{E}(P_0^0)^2 = \mathbb{E}\, S_{\rm ts}^2 \cdot k_{11}, \tag{78}$$

where we have used the fact that $P_0^0 \perp\!\!\!\perp (S_{\rm ts}, S_{\rm tr})$. Next, $m_{12} = \mathbb{E}\, S_{\rm ts}^2\, P_0^0 \widehat{P}_0$. where,

$$\begin{aligned}\widehat{P}_0 &= g_1^-(P_0^0 + P_0^+, Z_1^0 + Q_1^-, \overline{\gamma}_0^+, \overline{\gamma}_1^-, S_{\rm tr}^-)\\ &= \frac{\overline{\gamma}_0^+ P_0^+ + S_{\rm tr}\overline{\gamma}_1^- Q_1^-}{\overline{\gamma}_0^+ + S_{\rm tr}^2\overline{\gamma}_1^-} + P_0^0,\end{aligned} \tag{79}$$

where (P_0^0, P_0^+, Q_0^-) are independent of $(S_{\rm tr}, S_{\rm ts})$. Hence,

$$\begin{aligned}m_{12} &= \mathbb{E}\, S_{\rm ts}^2 \cdot \mathbb{E}(P_0^0)^2 + \mathbb{E}\frac{S_{\rm ts}^2\overline{\gamma}_0^+}{\overline{\gamma}_0^+ + S_{\rm tr}^2\overline{\gamma}_1^-}\mathbb{E}[P_0^0 P_0^+]\\ &= m_{11} + \mathbb{E}\left(\frac{S_{\rm ts}^2\overline{\gamma}_0^+}{\overline{\gamma}_0^+ + S_{\rm tr}^2\overline{\gamma}_1^-}\right)\cdot k_{12},\end{aligned} \tag{80}$$

since $\mathbb{E}[P_0^0 Q_1^-] = 0$ and $\mathbf{K}_0^+$ is the covariance matrix of (P_0^0, P_0^+) from line 2.

Finally for m_{22} we have,

$$\begin{aligned}m_{22} &= \mathbb{E}\, S_{\rm ts}^2 \widehat{P}_0 \widehat{P}_0\\ &= \mathbb{E}\left(\frac{S_{\rm ts}\overline{\gamma}_0^+}{\overline{\gamma}_0^+ + S_{\rm tr}^2\overline{\gamma}_1^-}\right)^2 \mathbb{E}(P_0^+)^2\\ &+ \mathbb{E}\left(\frac{S_{\rm ts}S_{\rm tr}\overline{\gamma}_1^-}{\overline{\gamma}_0^+ + S_{\rm tr}^2\overline{\gamma}_1^-}\right)^2 \mathbb{E}(Q_1^-)^2\\ &+ \mathbb{E}\, S_{\rm ts}^2\mathbb{E}(P_0^0)^2 + 2\mathbb{E}\frac{\overline{\gamma}_0^+ S_{\rm ts}^2}{\overline{\gamma}_0^+ + \overline{\gamma}_1^- S_{\rm tr}^2}\cdot \mathbb{E}\, P_0^0 P_0^+\\ &= k_{22}\mathbb{E}\left(\frac{S_{\rm ts}\overline{\gamma}_0^+}{\overline{\gamma}_0^+ + S_{\rm tr}^2\overline{\gamma}_1^-}\right)^2\\ &+ \tau_1^-\mathbb{E}\left(\frac{S_{\rm ts}S_{\rm tr}\overline{\gamma}_1^-}{\overline{\gamma}_0^+ + S_{\rm tr}^2\overline{\gamma}_1^-}\right)^2 - m_{11} + 2m_{12}.\end{aligned} \tag{81}$$

G. Special Cases

G.1. Linear Output with Square Error

In this section we examine a few special cases of the GLM problem (2). We first consider a linear output with additive Gaussian noise and a squared error training and test loss. Specifically, consider the model,

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{w}^0 + \mathbf{d} \tag{82}$$

We consider estimates of $\boldsymbol{w}^0$ such that:

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\mathrm{argmin}}\ \tfrac{1}{2}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{w}\|^2 + \tfrac{\lambda}{2\beta}\|\boldsymbol{w}\|^2 \tag{83}$$

The factor β is added above since the two terms scale with a ratio of β. It does not change analysis. Consider the ML-VAMP GLM learning algorithm applied to this problem. The following corollary follows from the Main result in Theorem 1.

Corollary 1 (Squared error)**.** *For linear regression,* i.e., $\phi(t) = t$, $\phi_{\rm out}(t,d) = t + d$, $f_{\rm ts}(y,\widehat{y}) = (y_{\rm ts} - \widehat{y}_{\rm ts})^2$, $F_{\rm out}(\mathbf{p}_2) = \frac{1}{N}\|\boldsymbol{y} - \mathbf{p}_2\|^2$, *we have*

$$\mathcal{E}_{\rm ts}^{\mathsf{LR}} = \mathbb{E}\left(\frac{\overline{\gamma}_0^+ S_{\rm ts}}{\overline{\gamma}_0^+ + S_{\rm tr}^2\overline{\gamma}_1^-}\right)^2 k_{22} + \mathbb{E}\left(\frac{\overline{\gamma}_1^- S_{\rm tr}S_{\rm ts}}{\overline{\gamma}_0^+ + S_{\rm tr}^2\overline{\gamma}_1^-}\right)^2 \tau_1^- + \sigma_d^2.$$

The quantities k_{22}, $\tau_1^-, \overline{\gamma}_0^+, \overline{\gamma}_1^-$ depend on the choice of regularizer λ and the covariance between features.

Proof. This follows directly from the following observation:

$$\begin{aligned}\mathcal{E}_{\rm ts}^{\mathsf{SLR}} &= \mathbb{E}(Z_{\rm ts} + D - \widehat{Z}_{\rm ts})^2 = \mathbb{E}(Z_{\rm ts} - \widehat{Z}_{\rm ts})^2 + \mathbb{E}\, D^2\\ &= m_{11} + m_{22} - 2m_{12} + \sigma_d^2.\end{aligned}$$

Substituting equation (81) proves the claim. □

G.2. Ridge Regression with i.i.d. Covariates

We next the special case when the input features are independent, i.e., (83) where rows of $\mathbf{X}$ corresponding to the training data has i.i.d Gaussian features with covariance $\mathbf{P}_{\text{train}} = \frac{\sigma_{\text{tr}}^2}{p}\mathbf{I}$ and $S_{\text{tr}} = \sigma_{\text{tr}}$.

Although the solution to (83) exists in closed form $(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y}$, we can study the effect of the regularization parameter λ on the generalization error $\mathcal{E}_{\text{ts}}$ as detailed in the result below.

Corollary 2. *Consider the ridge regression problem (83) with regularization parameter $\lambda > 0$. For the squared loss* i.e., $f_{\text{ts}}(y, \widehat{y}) = (y - \widehat{y})^2$, *i.i.d Gaussian features without train-test mismatch,* i.e., $S_{\text{tr}} = S_{\text{ts}} = \sigma_{\text{tr}}$, *the generalization error $\mathcal{E}_{\text{ts}}^{\mathsf{RR}}$ is given by Corollary 1, with constants*

$$k_{22} = \text{Var}(W^0), \qquad \overline{\gamma}_0^+ = \lambda/\beta,$$
$$\overline{\gamma}_1^- = \begin{cases} \frac{1}{G} - \frac{\lambda}{\sigma_{\text{tr}}^2} & \beta < 1 \\ \frac{\frac{\lambda}{\sigma_{\text{tr}}^2\beta}(\frac{1}{G} - \frac{\lambda}{\sigma_{\text{tr}}^2\beta})}{\frac{\beta-1}{G} + \frac{\lambda}{\sigma_{\text{tr}}^2\beta}} & \beta > 1 \end{cases}$$

where $G = G_{\text{mp}}(-\frac{\lambda}{\sigma_{\text{tr}}^2\beta})$, with G_{mp} given in Appendix H, and $\tau_1^- = \mathbb{E}(P_1^-)^2$ where P_1^- is given in equation (95) in the proof.

Proof of Corollary 2. We are interested in identifying the following constants appearing in Corollary 1:

$$\mathbf{K}_0^+, \tau_1^-, \overline{\gamma}_0^+, \overline{\gamma}_1^-.$$

These quantities are obtained as fixed points of the State Evolution Equations in Algo. 2. We explain below how to obtain expressions for these constants. Since these are fixed points we ignore the subscript k corresponding to the iteration number in Algo. 2.

In the case of problem (83), the maps $\text{prox}_{f_{\text{in}}}$ and $\text{prox}_{f_{\text{out}}}$, i.e., g_0^+ and g_3^- respectively, can be expressed as closed-form formulae. This leads to simplification of the SE equations as explained below.

We start by looking at the *forward pass* (finding quantities with superscript '+') of Algorithm 2 for different layers and then the *backward pass* (finding quantities with superscript '-') to get the parameters $\{\mathbf{K}_\ell^+, \tau_\ell^-, \overline{\alpha}_\ell^\pm, \overline{\gamma}_\ell^\pm\}$ for $\ell = 0, 1, 2$.

To begin with, notice that $f_{\text{in}}(w) = \frac{\lambda}{2}w^2$, and therefore the denoiser $g_0^+(\cdot)$ in (44) is simply,

$$g_0^+(r_0^-, \gamma_0^-) = \frac{\gamma_0^-}{\gamma_0^- + \lambda/\beta} r_0^-, \quad \text{and} \quad \frac{\partial g_0^+}{\partial r_0^-} = \frac{\gamma_0^-}{\gamma_0^- + \lambda/\beta}$$

Using the random variable R_0^- and substituting in the expression of the denoiser to get $\widehat{Z}_0$, we can now calculate $\overline{\alpha}_0^+$ using lines 2 and 2,

$$\overline{\alpha}_0^+ = \frac{\overline{\gamma}_0^-}{\overline{\gamma}_0^- + \lambda/\beta}, \qquad \overline{\gamma}_0^+ = \lambda/\beta. \tag{84}$$

Similarly, we have $f_{\text{out}}(p_2) = \frac{1}{2}(p_2 - y)^2$, whereby the output denoiser $g_3^-(\cdot)$ in the last layer for ridge regression is given by,

$$g_3^-(r_2^+, \gamma_2^+, y) = \frac{\gamma_2^+ r_2^+ + y}{\gamma_2^+ + 1}. \tag{85}$$

By substituting this denoiser in line 2 of the algorithm we get $\widehat{P}_2^-$ and thus, following the lines 2-2 of the algorithm we have

$$\overline{\alpha}_2^- = \frac{\overline{\gamma}_2^+}{\overline{\gamma}_2^+ + 1}, \qquad \text{whereby} \quad \overline{\gamma}_2^- = 1. \tag{86}$$

Having identified these constants $\overline{\alpha}_0^+, \overline{\gamma}_0^+, \overline{\alpha}_2^-, \overline{\gamma}_2^-$, we will now sequentially identify the quantities

$$(\overline{\alpha}_0^+, \overline{\gamma}_0^+) \to \mathbf{K}_0^+ \to (\overline{\alpha}_1^+, \overline{\gamma}_1^+) \to \mathbf{K}_1^+ \to (\overline{\alpha}_2^+, \overline{\gamma}_2^+) \to \mathbf{K}_2^+$$

in the forward pass, and then the quantities

$$\tau_0^- \leftarrow (\overline{\alpha}_0^-, \overline{\gamma}_0^-) \leftarrow \tau_1^- \leftarrow (\overline{\alpha}_1^-, \overline{\gamma}_1^-) \leftarrow \tau_2^- \leftarrow (\overline{\alpha}_2^-, \overline{\gamma}_2^-)$$

in the backward pass.

We also note that we have

$$\overline{\alpha}_\ell^+ + \overline{\alpha}_\ell^- = 1 \tag{87}$$

Forward Pass: Observe that $\mathbf{K}_0^+ = \text{Cov}(Z_0, Q_0^+)$. Now, from line 2, on simplification we get $Q_0^+ = -W_0^0$ whereby,

$$\mathbf{K}_0^+ = \text{var}(W^0)\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \tag{88}$$

Notice that from line 2, the pair (P_0^0, P_0^+) is jointly Gaussian with covariance matrix $\mathbf{K}_0^+$. But the above equation means that $P_0^+ = -P_0^0$, whereby $R_0^+ = 0$ from line 2.

Now, the linear denoiser $g_1^+(\cdot)$ is defined as in (47a). Note that since we are considering i.i.d Gaussian features for this problem, the random variable S_{tr} in this layer is a constant σ_{tr}. Therefore, similar to layer $\ell = 0$ by evaluating lines 2-2 of the algorithm we get $Q_1^+ = -Z_1^0$, whereby

$$\overline{\alpha}_1^+ = \frac{\sigma_{\text{tr}}^2\overline{\gamma}_1^-}{\overline{\gamma}_0^+ + \sigma_{\text{tr}}^2\overline{\gamma}_1^-}, \ \overline{\gamma}_1^+ = \frac{\overline{\gamma}_0^+}{\sigma_{\text{tr}}^2} = \frac{\lambda}{\sigma_{\text{tr}}^2\beta}, \ \mathbf{K}_1^+ = \sigma_{\text{tr}}^2\mathbf{K}_0^+. \tag{89}$$

Observe that this means

$$P_1^+ = -P_1^0. \tag{90}$$

Backward Pass: Since $Y = \phi_{\text{out}}(P_2^0, D) = P_2^0 + D$, line 2 of algorithm on simplification yields $P_2^- = D$, whereby we can get τ_2^-,

$$\tau_2^- = \mathbb{E}(P_2^-)^2 = \mathbb{E}[D^2] = \sigma_d^2. \tag{91}$$

Next, to calculate the terms $(\overline{\alpha}_1^-, \overline{\gamma}_1^-)$, we use the decoiser g_2^- defined in (47a) for line 2 of the algorithm to get $\widehat{P}_1$.

$$\widehat{P}_1 = \frac{\overline{\gamma}_1^+ R_1^+ + S_{\text{mp}}^- \overline{\gamma}_2^- R_2^-}{\overline{\gamma}_1^+ + (S_{\text{mp}}^-)^2 \overline{\gamma}_2^-} = \frac{S_{\text{mp}}^-(S_{\text{mp}}^+ P_1^0 + Q_2^-)}{\overline{\gamma}_1^+ + (S_{\text{mp}}^-)^2}, \tag{92}$$

where we have used $\overline{\gamma}_2^- = 1$, $R_1^+ = P_1^0 + P_1^+ = 0$ due to (90), and $R_2^- = Z_2^0 + Q_2^- = S_{\text{mp}}^+ P_1^0 + Q_2^-$ from lines 2, 2 and 2 respectively.

Then, we calculate $\overline{\alpha}_1^-$ and $\overline{\gamma}_1^-$ as $\overline{\alpha}_1^- = \mathbb{E}\frac{\partial g_2^-}{\partial P_1^+} = \mathbb{E}\frac{\overline{\gamma}_1^+}{\overline{\gamma}_1^+ + (S_{\text{mp}}^-)^2}$. This gives,

$$\overline{\alpha}_1^- = \begin{cases} \frac{\lambda}{\sigma_{\text{tr}}^2 \beta} G & \beta < 1 \\ (1 - \frac{1}{\beta}) + \frac{1}{\beta}\frac{\lambda}{\sigma_{\text{tr}}^2 \beta} G & \beta \geq 1 \end{cases} \tag{93}$$

Here, in the overparameterized case ($\beta > 1$), the denoiser g_2^- outputs R_1^+ with probability $1 - \frac{1}{\beta}$ and $\frac{\lambda}{\sigma_{\text{tr}}^2 \beta} G$ with probability $\frac{1}{\beta}$.

Next, from line 2 we get,

$$\overline{\gamma}_1^- = (\frac{1}{\overline{\alpha}_1^-} - 1)\overline{\gamma}_1^+ = \begin{cases} \frac{1}{G} - \frac{\lambda}{\sigma_{\text{tr}}^2 \beta} & \beta < 1 \\ \frac{\frac{\lambda}{\sigma_{\text{tr}}^2 \beta}(\frac{1}{G} - \frac{\lambda}{\sigma_{\text{tr}}^2 \beta})}{\frac{\beta - 1}{G} + \frac{\lambda}{\sigma_{\text{tr}}^2 \beta}} & \beta > 1 \end{cases} \tag{94}$$

Now from line 2 and equation (87) we get,

$$\begin{aligned} \overline{\alpha}_1^+ P_1^- &= \widehat{P}_1 - P_1^0 - \overline{\alpha}_1^- P_1^+ \overset{(a)}{=} \widehat{P}_1 - \overline{\alpha}_1^+ P_1^0 \\ &\overset{(b)}{=} \underbrace{\left(\frac{S_{\text{mp}}^- S_{\text{mp}}^+}{\frac{\lambda}{\sigma_{\text{tr}}^2 \beta} + (S_{\text{mp}}^-)^2} - \overline{\alpha}_1^+ \right)}_{A} P_1^0 + \underbrace{\frac{S_{\text{mp}}^-}{\frac{\lambda}{\sigma_{\text{tr}}^2 \beta} + (S_{\text{mp}}^-)^2}}_{B} Q_2^- \end{aligned} \tag{95}$$

where (a) follows from (90) and (87), and (b) follows from (92). From this one can obtain $\tau_1^- = \mathbb{E}(P_1^-)^2$ which can be calculated using the knowledge that P_1^0, Q_2^- are independent Gaussian with covariances $\mathbb{E}(P_1^0)^2 = \sigma_{\text{tr}}^2 \text{Var}(W^0)$, $\mathbb{E}(Q_2^-)^2 = \sigma_d^2$. Further, P_1^0, Q_2^- are independent of $(S_{\text{mp}}^+, S_{\text{mp}}^-)$.

Observe that by (95) we have

$$\tau_1^- = \frac{1}{(\overline{\alpha}_1^+)^2} \left(\mathbb{E}(A^2)\sigma_{\text{tr}}^2 \text{Var}(W^0) + \mathbb{E}(B^2)\sigma_d^2 \right). \tag{96}$$

with some simplification we get

$$\mathbb{E}(A^2) = (\frac{\lambda}{\sigma_{\text{tr}}^2 \beta})^2 G' - (\frac{\lambda}{\sigma_{\text{tr}}^2 \beta} G)^2, \tag{97a}$$

$$\mathbb{E}(B^2) = G - \frac{\lambda}{\sigma_{\text{tr}}^2 \beta} G', \tag{97b}$$

where $G = G_{\text{mp}}(-\frac{\lambda}{\sigma_{\text{tr}}^2 \beta})$, with G_{mp} given in Appendix H, and G' is the derivative of G_{mp} calculated at $-\frac{\lambda}{\sigma_{\text{tr}}^2 \beta}$.

Now consider the **under-parametrized** case ($\beta < 1$):

Let $u = -\frac{\lambda}{\sigma_{\text{tr}}^2 \beta}$ and $z = G_{\text{mp}}(u)$. In this case we have

$$\overline{\alpha}_1^+ = 1 - \frac{\lambda}{\sigma_{\text{tr}}^2 \beta} G = 1 + uz. \tag{98}$$

Note that,

$$\begin{aligned} G_{\text{mp}}^{-1}(z) = u &\overset{(a)}{\Rightarrow} R_{\text{mp}}(z) + \frac{1}{z} = u \\ &\overset{(b)}{\Rightarrow} \frac{1}{1 - \beta z} + \frac{1}{z} = u, \end{aligned} \tag{99a}$$

where $R_{\text{mp}}(.)$ is the R-transform defined in (Tulino et al., 2004) and (a) follows from the relationship between the R- and Stieltjes-transform and (b) follows from the fact that for Marchenko-Pastur distribution we have $R_{\text{mp}}(z) = \frac{1}{1 - z\beta}$. Therefore,

$$\begin{aligned} &G_{\text{mp}}(\frac{1}{1 - \beta z} + \frac{1}{z}) = z \\ &\Rightarrow G'_{\text{mp}}(\frac{1}{1 - \beta z} + \frac{1}{z}) = G' = \frac{1}{\frac{\beta}{(1 - \beta z)^2} - \frac{1}{z^2}}. \end{aligned} \tag{100}$$

For the **over-parametrized** case ($\beta > 1$) we have:

$$\overline{\alpha}_1^+ = \tfrac{1}{\beta}(1 + \tfrac{\lambda}{\sigma_{\text{tr}}^2 \beta} G) = \frac{1 - uz}{\beta}. \tag{101}$$

In this case, as mentioned in Appendix H and following the results from (Tulino et al., 2004), the measure μ_β scales with β and thus $R_{\text{mp}}(z) = \frac{\beta}{1 - z}$. Therefore, similar to (99a), z satisfies

$$\frac{\beta}{1 - z} + \frac{1}{z} = u \quad \Rightarrow G' = \frac{1}{\frac{\beta}{(1 - z)^2} - \frac{1}{z^2}}. \tag{102}$$

Now τ_1^- can be calculated as follows:

$$\tau_1^- = \eta^2 \left(u^2 z^2 \sigma_{\text{tr}}^2 var(W^0)(\kappa - 1) + \sigma_d^2 z(uz\kappa + 1) \right) \tag{103}$$

where

$$\eta = \begin{cases} \frac{1}{(1 + uz)} & \beta < 1 \\ \frac{\beta}{(1 - uz)} & \beta \geq 1 \end{cases}, \quad \kappa = \begin{cases} \frac{(1 - \beta z)^2}{\beta z^2 - (1 - \beta z)^2} & \beta < 1 \\ \frac{(1 - z)^2}{\beta z^2 - (1 - z)^2} & \beta \geq 1 \end{cases} \tag{104}$$

and z is the solution to the fixed points

$$\begin{cases} \frac{1}{1-\beta z} + \frac{1}{z} = u & \beta < 1 \\ \frac{\beta}{1-z} + \frac{1}{z} = u & \beta \geq 1 \end{cases}. \tag{105}$$

□

G.3. Ridgeless Linear Regression

Here we consider the case of Ridge regression (83) when $\lambda \to 0^+$. Note that the solution to the problem (83) is $(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y}$ remains unique since $\lambda > 0$. The following result was stated in (Hastie et al., 2019), and can be recovered using our methodology. Note however, that we calculate the generalization error whereas they have calculated the squared error, whereby we obtain an additional additive factor of σ_d^2. The result explains the double-descent phenomenon for Ridgeless linear regression.

Corollary 3. *For ridgeless linear regression, we have*

$$\lim_{\lambda \to 0^+} \mathcal{E}_{\text{ts}}^{\mathsf{RR}} = \begin{cases} \frac{1}{1-\beta}\sigma_d^2 & \beta < 1 \\ \frac{\beta}{\beta-1}\sigma_d^2 + (1-\frac{1}{\beta})\sigma_{\text{tr}}^2 \text{Var}(W^0) & \beta \geq 1 \end{cases}$$

Proof of Corollary 3. We calculate the parameters $\overline{\gamma}_0^+, \overline{\gamma}_1^-$, k_{22} and τ_1^- when $\lambda \to 0^+$. Before starting off, we note that

$$G_0 := \lim_{z \to 0^+} G_{\text{mp}}(-z) = \begin{cases} \frac{\beta}{1-\beta} & \beta < 1 \\ \frac{\beta}{\beta-1} & \beta > 1 \end{cases}, \tag{106}$$

as described in Appendix H. Following the derivations in Corollary 2, we have

$$\overline{\gamma}_0^+ = \lambda/\beta, \quad k_{22} = \text{Var}(W^0) \tag{107}$$

Now for $\lambda \to 0^+$, we have

$$1 - \overline{\alpha}_1^- = \begin{cases} 1 & \beta < 1 \\ \frac{1}{\beta} & \beta \geq 1 \end{cases}, \quad \overline{\gamma}_1^- = \begin{cases} \frac{1}{G_0} = \frac{1-\beta}{\beta} & \beta < 1 \\ \frac{\lambda}{(\beta-1)\sigma_{\text{tr}}^2 \beta} & \beta > 1 \end{cases}, \tag{108}$$

Using this in simplifying (95) for $\lambda \to 0^+$, we get

$$\tau_1^- = \mathbb{E}(P_1^-)^2 = \begin{cases} \sigma_d^2 G_0 & \beta < 1 \\ \beta\sigma_d^2 G_0 + \sigma_{\text{tr}}^2 \text{Var}(W^0)(\beta - 1) & \beta \geq 1 \end{cases}$$

where during the evaluation of $\mathbb{E}\left(\frac{S_{\text{mp}}^-}{\overline{\gamma}_1^+ + (S_{\text{mp}}^-)^2}\right)^2$, for the case of $\beta > 1$, we need to account for the point mass at 0 for S_{mp}^- with weight $1 - \frac{1}{\beta}$.

Next, notice that

$$a := \frac{\overline{\gamma}_0^+ \sigma_{\text{tr}}}{\overline{\gamma}_0^+ + \overline{\gamma}_1^- \sigma_{\text{tr}}^2} = \begin{cases} 0 & \beta < 1 \\ (1 - \frac{1}{\beta})\sigma_{\text{tr}} & \beta \geq 1 \end{cases},$$

and,

$$b := \frac{\overline{\gamma}_1^- \sigma_{\text{tr}}^2}{\overline{\gamma}_0^+ + \overline{\gamma}_1^- \sigma_{\text{tr}}^2} = \begin{cases} 1 & \beta < 1 \\ \frac{1}{\beta} & \beta \geq 1 \end{cases},$$

Thus applying Corollary 1, we get

$$\begin{aligned} \mathcal{E}_{\text{ts}}^{\mathsf{RR}} &= a^2 k_{22} + b^2 \tau_1^- + \sigma_d^2 \\ &= \begin{cases} \frac{1}{1-\beta}\sigma_d^2 & \beta < 1 \\ \frac{\beta}{\beta-1}\sigma_d^2 + (1-\frac{1}{\beta})\sigma_{\text{tr}}^2 \text{Var}(W^0) & \beta \geq 1 \end{cases} \end{aligned}$$

This proves the claim. □

G.4. Train-Test Mismatch

Observe that our formulation allows for analyzing the effect of mismatch in the training and test distribution. One can consider arbitrary joint distributions over $(S_{\text{tr}}, S_{\text{ts}})$ that model the mismatch between training and test features. Here we give a simple example which highlights the effect of this mismatch.

Definition 4 (Bernoulli ε-mismatch)**.** $(S_{\text{ts}}, S_{\text{tr}})$ has a bivariate Bernoulli distribution with

- $\Pr\{S_{\text{tr}} = S_{\text{ts}} = 0\} = \mathbb{P}\{S_{\text{tr}} = S_{\text{ts}} = 1\} = (1-\varepsilon)/2$
- $\Pr\{S_{\text{tr}} = 0, S_{\text{ts}} = 1\} = \mathbb{P}\{S_{\text{tr}} = 1, S_{\text{ts}} = 0\} = \varepsilon/2$

Notice that the marginal distribution of the S_{tr} in the Bernoulli $\varepsilon-$mismatch model is such that $\mathbb{P}(S_{\text{tr}} \neq 0) = \frac{1}{2}$. Hence half of the features extracted by the matrix V_0 are relevant during training. Similarly, $\mathbb{P}(S_{\text{ts}} \neq 0) = \frac{1}{2}$. However the features spanned by the test data do not exactly overlap with the features captured in the training data, and the fraction of features common to both the training and test data is $1 - \varepsilon$. Hence for $\varepsilon = 0$, there is no training-test mismatch, whereas for $\varepsilon = 1$ there is a complete mismatch.

The following result shows that the generalization error increases linearly with the mismatch parameter ε.

Corollary 4 (Mismatch)**.** *Consider the problem of Linear Regression (83) under the conditions of Corollary 1. Additionally suppose we have Bernoulli ε-mismatch between the training and test distributions. Then*

$$\mathcal{E}_{\text{ts}} = \tfrac{k_{22}}{2}((1-\varepsilon)\gamma^{*2} + \varepsilon) + \tfrac{\tau_1^-}{2}(1-\gamma^*)(1-\varepsilon) + \sigma_d^2,$$

where $\gamma^* := \frac{\overline{\gamma}_0^+}{\overline{\gamma}_0^+ + \overline{\gamma}_1^-}$. The terms $k_{22}, \tau_1^-, \gamma^*$ are independent of ε.

Proof. This follows directly by calculating the expectations of the terms in Corollary 1, with the joint distribution of $(S_{\text{tr}}, S_{\text{ts}})$ given in Definition 4. □

The quantities k_{22} and τ_1^- in the result above can be calculated similar to the derivation in the proof of Corollary 2 and can in general depend on the regularization parameter λ and overparameterization parameter β.

G.5. Logistic Regression

The precise analysis for the special case of regularized logistic regression estimator with i.i.d Gaussian features is provided in (Salehi et al., 2019). Consider the logistic regression model,

$$\mathbb{P}(y_i = 1|\boldsymbol{x}_i) := \rho(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w}) \quad \text{for } i = 1, \cdots, N$$

where $\rho(x) = \frac{1}{1+e^{-x}}$ is the standard logistic function.

In this problem we consider estimates of $\boldsymbol{w}^0$ such that

$$\widehat{\boldsymbol{w}} := \underset{\boldsymbol{w}}{\operatorname{argmin}}\, \mathbf{1}^\mathsf{T}\log(1+e^{\mathbf{X}\boldsymbol{w}}) - \boldsymbol{y}^\mathsf{T}\mathbf{X}\boldsymbol{w} + F_{\text{in}}(\boldsymbol{w}).$$

where F_{in} is the reguralization function. This is a special case of optimization problem (2) where

$$F_{\text{out}}(\boldsymbol{y}, \mathbf{X}\boldsymbol{w}) = \mathbf{1}^\mathsf{T}\log(1+e^{\mathbf{X}\boldsymbol{w}}) - \boldsymbol{y}^\mathsf{T}\mathbf{X}\boldsymbol{w}. \tag{109}$$

Similar to the linear regression model, using the ML-VAMP GLM learning algorithm, we can characterize the generalization error for this model with quantities $\mathbf{K}_0^+, \tau_1^-, \overline{\gamma}_0^+, \overline{\gamma}_1^-$ given by algorithm 2. We note that in this case, the output non-linearity is

$$\phi_{\text{out}}(p_2, d) = \mathbb{1}_{\{\rho(p_2)>d\}} \tag{110}$$

where $d \sim \text{Unif}(0,1)$. Also, the denoisers g_0^+, and g_3^- can be derived as the proximal operators of F_{in}, and F_{out} defined in (25).

G.6. Support Vector Machines

The asymptotic generalization error for support vector machine (SVM) is provided in (Deng et al., 2019). Our model can also handle SVMs. Similar to logistic regression, SVM finds a linear classifier using the hinge loss instead of logistic loss. Assuming the class labels are $y = \pm 1$ the hinge loss is

$$\ell_{\text{hinge}}(y, \widehat{y}) = \max(0, 1 - y\widehat{y}). \tag{111}$$

Therefore, if we take

$$F_{\text{out}}(\boldsymbol{y}, \mathbf{X}\boldsymbol{w}) = \sum_i \max(0, 1 - y_i\mathbf{X}_i\boldsymbol{w}), \tag{112}$$

where $\mathbf{X}_i$ is the i^{th} row of the data matrix, the ML-VAMP algorithm for GLMs finds the SVM classifier. The algorithm would have proximal map of hinge loss and our theory provides exact predictions for the estimation and prediction error of SVM.

As with all other models considered in this work, the true underlying data generating model could be anything that can be represented by the graphical model in Figure 1, e.g. logistic or probit model, and our theory is able to exactly predict the error when SVM is applied to learn such linear classifiers in the large system limit.

H. Marchenko-Pastur distribution

We describe the random variable S_{mp} defined in (12) where S_{mp}^2 has a rescaled Marchenko-Pastur distribution. Notice that the positive entries of $\mathbf{s}_{\text{mp}}$ are the positive eigenvalues of $\mathbf{U}^\mathsf{T}\mathbf{U}$ (or $\mathbf{U}\mathbf{U}^\mathsf{T}$).

Observe that $U_{ij} \sim N(0, \frac{1}{p})$, whereas, the standard scaling while studying the Marchenko-Pastur distribution is for matrices $\mathbf{H}$ such that $H_{ij} \sim \mathcal{N}(0, \frac{1}{N})$ (for e.g. see equation (1.10) from (Tulino et al., 2004) and the discussion preceding it). Also notice that $\sqrt{\beta}\mathbf{U}$ has the same distribution as $\mathbf{H}$. Thus the results from (Tulino et al., 2004) apply directly to the distributions of eigenvalues of $\beta\mathbf{U}^\mathsf{T}\mathbf{U}$ and $\beta\mathbf{U}\mathbf{U}^\mathsf{T}$. We state their result below taking into account this disparity in scaling.

The positive eigenvalues of $\beta\mathbf{U}^\mathsf{T}\mathbf{U}$ have an empirical distribution which converges to the following density:

$$\mu_\beta(x) = \frac{\sqrt{(b_\beta - x)_+(x - a_\beta)_+}}{2\pi\beta x} \tag{113}$$

where $a_\beta = (1-\sqrt{\beta})^2$, $b_\beta := (1+\sqrt{\beta})^2$. Similarly the positive eigenvalues of $\beta\mathbf{U}\mathbf{U}^\mathsf{T}$ have an empirical distribution converging to the density $\beta\mu_\beta$. We note the following integral which is useful in our analysis:

$$\begin{aligned} G_0 :&= \lim_{z\to 0^-} \mathbb{E}\frac{1}{S_{\text{mp}}^2 - z}\mathbb{1}_{\{S_{\text{mp}}>0\}} \\ &= \lim_{z\to 0^-} \int_{a_\beta}^{b_\beta} \frac{1}{x/\beta - z}\mu_\beta(x)dx = \frac{\beta}{|\beta - 1|}. \end{aligned} \tag{114}$$

More generally, the Stieltjes transform of the density is given by:

$$G_{\text{mp}}(z) = \mathbb{E}\frac{1}{S_{\text{mp}}^2 - z}\mathbb{1}_{\{S_{\text{mp}}>0\}} = \int_{a_\beta}^{b_\beta} \frac{1}{x/\beta - z}\mu_\beta(x)dx \tag{115}$$