Inferring Change Points in High-Dimensional Linear Regression via Approximate Message Passing

Gabriel Arpino [1] **Xiaoqi Liu** [1] **Ramji Venkataramanan** [1]

Abstract

We consider the problem of localizing change points in high-dimensional linear regression. We propose an Approximate Message Passing (AMP) algorithm for estimating both the signals and the change point locations. Assuming Gaussian covariates, we give an exact asymptotic characterization of its estimation performance in the limit where the number of samples grows proportionally to the signal dimension. Our algorithm can be tailored to exploit any prior information on the signal, noise, and change points. It also enables uncertainty quantification in the form of an efficiently computable approximate posterior distribution, whose asymptotic form we characterize exactly. We validate our theory via numerical experiments, and demonstrate the favorable performance of our estimators on both synthetic data and images.

1. Introduction

Heterogeneity is a common feature of large, high-dimensional datasets. When the data are ordered by time, a simple form of heterogeneity is a change in the data generating mechanism at certain unknown instants of time. If these 'change points' were known, or estimated accurately, the dataset could be partitioned into homogeneous subsets, each amenable to analysis via standard statistical techniques (Fryzlewicz, 2014). Models with change points have been studied in a variety of statistical contexts, such as the detection of changes in: signal means (Wang & Samworth, 2018; Wang et al., 2020; Liu et al., 2021; Li et al., 2023a); covariance structures (Cho & Fryzlewicz, 2015; Wang et al., 2021b); graphs (Londschien et al., 2023; Bhattacharjee et al., 2020; Fan & Guan, 2018); dynamic networks (Wang et al., 2021a); and functionals (Madrid Padilla et al., 2022). Change point models have found application in a range of fields including genomics (Braun et al., 2000), neuroscience (Aston & Kirch, 2012), and economics (Andreou & Ghysels, 2002).

In this paper, we consider high-dimensional linear regression with change points. We are given a sequence of data $(y_i, \boldsymbol{X}_i) \in \mathbb{R} \times \mathbb{R}^p$, for $i \in [n]$, from the model

$$y_i = (\boldsymbol{X}_i)^\top \boldsymbol{\beta}^{(i)} + \varepsilon_i, \quad i = 1, \dots, n. \tag{1}$$

Here, $\boldsymbol{\beta}^{(i)} \in \mathbb{R}^p$ is the unknown regression vector for the ith sample, $\boldsymbol{X}_i \in \mathbb{R}^p$ is the (known) covariate vector, and ε_i is additive noise. We denote the unknown change points, i.e., the sample indices where the regression vector changes, by $\eta_1, \dots, \eta_{L^*-1}$. Specifically, we have

$$1 = \eta_0 < \eta_1 < \cdots < \eta_{L^*} = n,$$

with $\boldsymbol{\beta}^{(i)} \neq \boldsymbol{\beta}^{(i-1)}$ if and only if $i \in \{\eta_\ell\}_{\ell=1}^{L^*-1}$. We note that L^* is the number of distinct signals in the sequence $\{\boldsymbol{\beta}^{(i)}\}_{i=1}^n$, and (L^*-1) is the number of change points. The number of change points is not known, but an upper bound L on the value of L^* is available. The goal is to estimate the change point locations as well as the L^* signals. We would also like to quantify the uncertainty in these estimates, e.g., via confidence sets or a posterior distribution.

Linear regression with change points in the high-dimensional regime (where the dimension p is comparable to, or exceeds, the number of samples n) has been studied in a number of recent works, e.g. (Lee et al., 2016; Leonardi & Bühlmann, 2016; Kaul et al., 2019; Rinaldo et al., 2021; Xu et al., 2022; Li et al., 2023b; Bai & Safikhani, 2023). Most of these papers consider the setting where the signals are sparse (the number of non-zero entries in $\boldsymbol{\beta}^{(i)} \in \mathbb{R}^p$ is $o(p)$), and analyze procedures that combine the LASSO estimator (or a variant) with a partitioning technique, e.g., dynamic programming. The recent work of Gao & Wang (2022) assumes sparsity on the difference between signals across a change point, and Cho et al. (2024) consider general non-sparse signals. Although existing procedures for high-dimensional change point regression incorporate sparsity-based constraints, they cannot be easily adapted to take advantage of other kinds of signal priors. Moreover, they are

[1]University of Cambridge. Correspondence to: Gabriel Arpino <ga442@cam.ac.uk>, Xiaoqi Liu <xl394@cam.ac.uk>, Ramji Venkataramanan <rv285@cam.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

not well-equipped to exploit prior information on the change point locations. Bayesian approaches to change point detection have been studied in several works, e.g. (Fearnhead, 2006; Lungu et al., 2022), however they mainly focus on (low-dimensional) time-series.

Main Contributions We propose an Approximate Message Passing (AMP) algorithm for estimating the signals $\{\boldsymbol{\beta}^{(i)}\}_{i=1}^n$ and the change point locations $\{\eta_\ell\}_{\ell=1}^{L^*-1}$. Under the assumption that the covariates are i.i.d. Gaussian, we give an exact characterization of the performance of the algorithm in the limit as both the signal dimension p and the number of samples n grow, with n/p converging to a constant δ (Theorem 3.1). The AMP algorithm is iterative, and defined via a pair of 'denoising' functions for each iteration. We show how these functions can be tailored to take advantage of any prior information on the signals and the change points (Proposition 3.2). We then show how the change points can be estimated using the iterates of the AMP algorithm, and how the uncertainty can be quantified via an approximate posterior distribution (Section 3.3). Our theory enables asymptotic guarantees on the accuracy of the change point estimator and on the posterior distribution (Propositions 3.3, 3.4). In Section 4, we present experiments on both synthetic data and images, demonstrating the superior performance of AMP compared to other state-of-the-art algorithms for linear regression with change points.

Although our results do not explicitly need assumptions on the separation between adjacent change points, they are most interesting when the separation is of order n, i.e., $\Delta := \min_{\ell\in[L]}(\eta_\ell - \eta_{\ell-1}) = O(n)$. This separation is natural in our regime, where the number of samples n is proportional to p and the number of degrees of freedom in the signals also grows linearly in p. Existing results on high-dimensional change point regression usually assume signals that are s-sparse, and demonstrate change point estimators that are consistent when $\Delta = \omega\big(s \log p/\kappa^2\big)$, where κ is a constant determined by the separation between the signals (Rinaldo et al., 2021; Wang et al., 2021c; Li et al., 2023b). In contrast, we do not assume signal sparsity that is sublinear in n, so the change point estimation error will not tend to zero unless $n/p \to \infty$. We therefore quantify the AMP performance via precise asymptotics for the estimation error and the approximate posterior distribution.

Approximate Message Passing AMP, a family of iterative algorithms first proposed for linear regression (Kabashima, 2003; Donoho et al., 2009; Krzakala et al., 2012), has been applied to a variety of high-dimensional estimation problems including estimation in generalized linear models (Rangan, 2011; Schniter & Rangan, 2014; Maillard et al., 2020; Mondelli & Venkataramanan, 2021) and low-rank matrix estimation (Fletcher & Rangan, 2018; Lesieur et al., 2017; Montanari & Venkataramanan, 2021; Barbier et al., 2020). An attractive feature of AMP algorithms is that under suitable model assumptions, their performance in the high-dimensional limit can be characterized by a succinct deterministic recursion called *state evolution*. The state evolution characterization has been used to show that AMP achieves Bayes-optimal performance for some models (Deshpande & Montanari, 2014; Donoho et al., 2013; Barbier et al., 2019).

An important feature of our AMP algorithm, in contrast to the above works, is the use of *non-separable* denoising functions. (We say a function $g : \mathbb{R}^{m\times L} \to \mathbb{R}^{m\times L}$ is separable if it acts row-wise identically on the input matrix.) Even with simplifying assumptions on the signal and noise distributions, non-separable AMP denoisers are required to handle the temporal dependence caused by the change points, and allow for precise uncertainty quantification around possible change point locations. Our main state evolution result (Theorem 3.1) leverages recent results by Berthier et al. (2019) and Gerbelot & Berthier (2023) for AMP with non-separable denoisers. Non-separable denoisers for AMP have been been studied for linear and generalized linear models, to exploit the dependence within the signal (Som & Schniter, 2012; Metzler et al., 2016; Ma et al., 2019) or between the covariates (Zhang et al., 2023). Here non-separable denoisers are required to exploit the dependence in the observations, caused by the change points.

Although we assume i.i.d. Gaussian covariates, based on recent AMP universality results (Wang et al., 2022), we expect the results apply to a broad class of i.i.d. designs. An interesting direction for future work is to generalize our results to rotationally invariant designs, a much broader class for which AMP-like algorithms have been proposed for regression without change points (Ma & Ping, 2017; Rangan et al., 2019; Takeuchi, 2020; Pandit et al., 2020).

2. Preliminaries

Notation We let $[n] = \{1, 2, \dots, n\}$. We use boldface notation for matrices and vectors. For vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we write $\boldsymbol{x} \le \boldsymbol{y}$ if $x_i \le y_i$ for all $i \in [n]$, and let $[\boldsymbol{x}, \boldsymbol{y}] := \{\boldsymbol{v} \in \mathbb{R}^n : x_i \le v_i \le y_i \ \forall i \in [n]\}$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times L}$ and $i \in [m], j \in [L]$, we let $\boldsymbol{A}_{[i,:]}$ and $\boldsymbol{A}_{[:,j]}$ denote its ith row and jth column respectively. Similarly, for a vector $\boldsymbol{\psi} \in [L]^n$, we let $\boldsymbol{A}_{[:,\boldsymbol{\psi}]} \in \mathbb{R}^m$ denote the vector whose i-th entry is $\boldsymbol{A}_{i,\psi_i}$.

For two sequences (in n) of random variables X_n, Y_n, we write $X_n \overset{\mathbb{P}}{\simeq} Y_n$ when their difference converges in probability to 0, i.e., $\lim_{n\to\infty} \mathbb{P}(|X_n - Y_n| > \epsilon) = 0$ for any $\epsilon > 0$. Denote the covariance matrix of random vector $\boldsymbol{Z} \in \mathbb{R}^q$ as $\mathrm{Cov}(\boldsymbol{Z}) \in \mathbb{R}^{q\times q}$. We refer to all random elements, in-

cluding vectors and matrices, as random variables. When referring to probability densities, we include probability mass functions, with integrals interpreted as sums when the distribution is discrete.

Model Assumptions In model (1), we assume independent Gaussian covariate vectors $\boldsymbol{X}_i \overset{i.i.d}{\sim} \mathsf{N}(0, \boldsymbol{I}_p/n)$ for $i \in [n]$. We consider the high-dimensional regime where $n, p \to \infty$ and $\frac{n}{p}$ converges to a constant $\delta > 0$. Following the change point literature, we assume the number of change points $(L^* - 1)$ is fixed and does not scale with n, p.

Pseudo-Lipschitz Functions Our results are stated in terms of uniformly *pseudo-Lipschitz* functions (Berthier et al., 2019). For $C > 0$ and $r \in [1, \infty)$, let $PL_{n,m,q}(r, C)$ be the set of functions $\phi : \mathbb{R}^{n\times q} \to \mathbb{R}^{m\times q}$ such that $\frac{\|\phi(\boldsymbol{x})-\phi(\tilde{\boldsymbol{x}})\|_F}{\sqrt{m}} \leq C\left(1 + \left(\frac{\|\boldsymbol{x}\|_F}{\sqrt{n}}\right)^{r-1} + \left(\frac{\|\tilde{\boldsymbol{x}}\|_F}{\sqrt{n}}\right)^{r-1}\right)\frac{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_F}{\sqrt{n}}$ for all $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathbb{R}^{n\times q}$. Note that $\cup_{C>0} PL_{n,m,q}(r_1, C) \subseteq \cup_{C>0} PL_{n,m,q}(r_2, C)$ for any $1 \leq r_1 \leq r_2$. A function $\phi \in PL_{n,m,q}(r, C)$ is called pseudo-Lipschitz of order r. A family of pseudo-Lipschitz functions is said to be *uniformly* pseudo-Lipschitz if all functions of the family are pseudo-Lipschitz with the same order r and the same constant C. For functions $f : \mathbb{R}^{n\times m} \times \mathbb{R}^{n\times M} \to \mathbb{R}$ of the form $f(\boldsymbol{a}, \boldsymbol{b}) = c$, we say f is *uniformly* pseudo-Lipschitz with respect to $\boldsymbol{a}$ if the family $\{f_n(\cdot, \boldsymbol{b}) : n \in \mathbb{N}, \boldsymbol{b} \in \mathbb{R}^{n\times M}\}$ is uniformly pseudo-Lipschitz. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the mean squared error $\phi(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y}\rangle/n$ and the normalized squared correlation $\phi(\boldsymbol{x}, \boldsymbol{y}) = |\langle \boldsymbol{x}, \boldsymbol{y}\rangle|/n$ are examples of uniformly pseudo-Lipschitz functions.

3. AMP Algorithm and Main Results

We stack the feature vectors, observations and noise elements, respectively, to form $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]^\top \in \mathbb{R}^{n\times p}$, $\boldsymbol{y} := [y_1, \ldots, y_n] \in \mathbb{R}^n$ and $\boldsymbol{\varepsilon} := [\varepsilon_1, \ldots, \varepsilon_n]^\top \in \mathbb{R}^n$. Let $\boldsymbol{\eta} := [\eta_1, \ldots, \eta_{L^*}, \ldots, \eta_L]$ be the vector containing the true change points, and $\boldsymbol{B} := [\boldsymbol{\beta}^{(\eta_1)}, \ldots, \boldsymbol{\beta}^{(\eta_{L^*})}, \ldots, \boldsymbol{\beta}^{(\eta_L)}] \in \mathbb{R}^{p\times L}$ be the matrix containing the true signals. Since the algorithm assumes no knowledge of L^*, other than $L^* \leq L$, the columns $L^* + 1, \ldots, L$ of $\boldsymbol{B}$ can all be taken to be zero. Similarly, $\eta_{L^*} = \eta_{L^*+1} = \cdots = \eta_L = n$.

Recall that for $i \in [n]$, each observation y_i is generated from model (1). Let $\boldsymbol{\Psi} \in [L]^n$ be the *signal configuration vector*, whose i-th entry stores the index of the signal underlying observation y_i. That is, for $i \in [n]$ and $\ell \in [L]$, let $\Psi_i = \ell$ if and only if $\boldsymbol{\beta}^{(i)}$ equals the ℓth column of the signal matrix $\boldsymbol{B}$. We note that that there is a one-to-one correspondence between $\boldsymbol{\Psi}$ and the change point vector $\boldsymbol{\eta}$. We can then rewrite $\boldsymbol{y}$ in a more general form:

$$\boldsymbol{y} = (\boldsymbol{X}\boldsymbol{B})_{[:,\boldsymbol{\Psi}]} + \boldsymbol{\varepsilon} := q(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) \in \mathbb{R}^n, \qquad (2)$$

where q acts row-wise on matrix inputs. We note that the mixed linear regression model (Yi et al., 2014; Zhang et al., 2022; Tan & Venkataramanan, 2023) can also be written in the form in (2), with a crucial difference. In mixed regression, the components of $\boldsymbol{\Psi}$ are assumed to be drawn independently from some distribution on $[L]$, i.e., each y_i is independently generated from one of the L signals. In the change point setting, the entries of $\boldsymbol{\Psi}$ are *dependent*, since they change value only at entries $\eta_1, \ldots, \eta_{(L^*-1)}$.

AMP Algorithm We now describe the AMP algorithm for estimating $\boldsymbol{B}$ and $\boldsymbol{\eta}$. In each iteration $t \geq 1$, the algorithm produces an updated estimate of the signal matrix $\boldsymbol{B}$, which we call $\boldsymbol{B}^t$, and of the linearly transformed signal $\boldsymbol{\Theta} := \boldsymbol{X}\boldsymbol{B}$, which we call $\boldsymbol{\Theta}^t$. These estimates have distributions that can be described by a deterministic low-dimensional matrix recursion called *state evolution*. In Section 3.3, we show how the estimate $\boldsymbol{\Theta}^t$ can be combined with $\boldsymbol{y}$ to infer $\boldsymbol{\eta}$ with precisely quantifiable error.

Starting with an initializer $\boldsymbol{B}^0 \in \mathbb{R}^{p\times L}$ and defining $\hat{\boldsymbol{R}}^{-1} := \boldsymbol{0}_{n\times L}$, for $t \geq 0$ the algorithm computes:

$$\begin{aligned} \boldsymbol{\Theta}^t &= \boldsymbol{X}\hat{\boldsymbol{B}}^t - \hat{\boldsymbol{R}}^{t-1}(\boldsymbol{F}^t)^\top, \quad \hat{\boldsymbol{R}}^t = g^t\left(\boldsymbol{\Theta}^t, \boldsymbol{y}\right), \\ \boldsymbol{B}^{t+1} &= \boldsymbol{X}^\top\hat{\boldsymbol{R}}^t - \hat{\boldsymbol{B}}^t(\boldsymbol{C}^t)^\top, \quad \hat{\boldsymbol{B}}^t = f^t\left(\boldsymbol{B}^t\right), \end{aligned} \qquad (3)$$

where the denoising functions $g^t : \mathbb{R}^{n\times L} \times \mathbb{R}^n \to \mathbb{R}^{n\times L}$ and $f^t : \mathbb{R}^{p\times L} \to \mathbb{R}^{p\times L}$ are used to define the matrices $\boldsymbol{F}^t$, $\boldsymbol{C}^t$ as follows:

$$\boldsymbol{C}^t = \frac{1}{n}\sum_{i=1}^{n} \partial_i g_i^t\left(\boldsymbol{\Theta}^t, \boldsymbol{y}\right), \quad \boldsymbol{F}^t = \frac{1}{n}\sum_{j=1}^{p} \mathrm{d}_j f_j^t(\boldsymbol{B}^t).$$

Here $\partial_i g_i^t\left(\boldsymbol{\Theta}, \boldsymbol{y}\right)$ is the $L \times L$ Jacobian of g_i^t w.r.t. the ith row of $\boldsymbol{\Theta}$. Similarly, $\mathrm{d}_j f_j^t(\boldsymbol{B}^t)$ is the $L \times L$ Jacobian of f_j^t with respect to the j-th row of its argument. The time complexity of each iteration in (3) is $O(npL + r_n)$, where r_n is the time complexity of computing f^t, g^t.

Crucially, the denoising functions g^t and f^t are not restricted to being separable. (Recall that a separable function acts row-wise identically on its input, i.e., $g : \mathbb{R}^{m\times L} \to \mathbb{R}^{m\times L}$ is separable if for all $\boldsymbol{U} \in \mathbb{R}^{m\times L}$ and $i \neq j$, we have $[g(\boldsymbol{U})]_i = g_i(\boldsymbol{U}_i) = g_j(\boldsymbol{U}_i)$.) Hence, in general, we have

$$g^t(\boldsymbol{\Theta}, \boldsymbol{y}) = \begin{bmatrix} g_1^t(\boldsymbol{\Theta}, \boldsymbol{y}) \\ \vdots \\ g_n^t(\boldsymbol{\Theta}, \boldsymbol{y}) \end{bmatrix}, \quad f^t(\boldsymbol{B}) = \begin{bmatrix} f_1^t(\boldsymbol{B}) \\ \vdots \\ f_n^t(\boldsymbol{B}) \end{bmatrix}.$$

Our non-separable approach is required to handle the temporal dependence created by the change points. For example,

if there was one change point uniformly in distributed in $[n]$, then $g^t(\boldsymbol{\Theta}, \boldsymbol{y})$ should take into account that y_i is more likely to have come from the the first signal for i close to 1, and from the second signal for i close to n. This is in contrast to existing AMP algorithms for mixed regression (Tan & Venkataramanan, 2023), where under standard model assumptions, it suffices to consider separable denoisers.

In Appendix G, we review the AMP algorithm and state evolution for the generalized linear model *without* change points, i.e., the model in (2) without the vector $\boldsymbol{\Psi}$. This is useful background for the state evolution characterization described below for the change point setting.

State Evolution The memory terms $-\hat{\boldsymbol{R}}^{t-1}(\boldsymbol{F}^t)^\top$ and $-\hat{\boldsymbol{B}}^t(\boldsymbol{C}^t)^\top$ in our AMP algorithm (3) debias the iterates $\boldsymbol{\Theta}^t$ and $\boldsymbol{B}^{t+1}$, and enable a succinct distributional characterization. In the high-dimensional limit as $n, p \to \infty$ (with $n/p \to \delta$), the empirical distributions of $\boldsymbol{\Theta}^t$ and $\boldsymbol{B}^{t+1}$ are quantified through the random variables $\boldsymbol{V}^t_{\boldsymbol{\Theta}}$ and $\boldsymbol{V}^{t+1}_{\boldsymbol{B}}$ respectively, where

$$\boldsymbol{V}^t_{\boldsymbol{\Theta}} := \boldsymbol{Z}\boldsymbol{\rho}^{-1}\boldsymbol{\nu}^t_{\boldsymbol{\Theta}} + \boldsymbol{G}^t_{\boldsymbol{\Theta}} \in \mathbb{R}^{n\times L}, \tag{4}$$

$$\boldsymbol{V}^{t+1}_{\boldsymbol{B}} := \boldsymbol{B}\boldsymbol{\nu}^{t+1}_{\boldsymbol{B}} + \boldsymbol{G}^{t+1}_{\boldsymbol{B}} \in \mathbb{R}^{p\times L}. \tag{5}$$

The matrices $\boldsymbol{\rho}, \boldsymbol{\nu}^t_{\boldsymbol{\Theta}}, \boldsymbol{\nu}^{t+1}_{\boldsymbol{B}} \in \mathbb{R}^{L\times L}$ are deterministic and defined below. The random matrices $\boldsymbol{Z}$, $\boldsymbol{G}^t_{\boldsymbol{\Theta}}$, and $\boldsymbol{G}^{t+1}_{\boldsymbol{B}}$ are independent of $\boldsymbol{X}$, and have i.i.d. rows following a Gaussian distribution. Namely, for $i \in [n]$ we have $\boldsymbol{Z}_i \overset{i.i.d}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{\rho})$. For $i \in [n], j \in [p]$ and $s, r \geq 0$, $(\boldsymbol{G}^t_{\boldsymbol{\Theta}})_i \overset{i.i.d}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{\kappa}^{t,t}_{\boldsymbol{\Theta}})$ with $\mathrm{Cov}((\boldsymbol{G}^r_{\boldsymbol{\Theta}})_i, (\boldsymbol{G}^s_{\boldsymbol{\Theta}})_i) = \boldsymbol{\kappa}^{r,s}_{\boldsymbol{\Theta}}$. Similarly, $(\boldsymbol{G}^t_{\boldsymbol{B}})_j \overset{i.i.d}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{\kappa}^{t,t}_{\boldsymbol{B}})$ with $\mathrm{Cov}((\boldsymbol{G}^r_{\boldsymbol{B}})_j, (\boldsymbol{G}^s_{\boldsymbol{B}})_j) = \boldsymbol{\kappa}^{r,s}_{\boldsymbol{B}}$. The $L \times L$ deterministic matrices $\boldsymbol{\nu}^t_{\boldsymbol{\Theta}}, \boldsymbol{\kappa}^{r,s}_{\boldsymbol{\Theta}}, \boldsymbol{\nu}^t_{\boldsymbol{B}}$, and $\boldsymbol{\kappa}^{r,s}_{\boldsymbol{B}}$ are defined below via the *state evolution* recursion.

Given the initializer $\boldsymbol{B}^0$ for our algorithm (3), we initialize state evolution by setting $\boldsymbol{\nu}^0_{\boldsymbol{\Theta}} := \boldsymbol{0}$, and

$$\boldsymbol{\rho} := \frac{1}{\delta}\lim_{p\to\infty}\frac{1}{p}\boldsymbol{B}^\top\boldsymbol{B},\ \boldsymbol{\kappa}^{0,0}_{\boldsymbol{\Theta}} := \frac{1}{\delta}\lim_{p\to\infty}\frac{1}{p}f^0(\boldsymbol{B}^0)^\top f^0(\boldsymbol{B}^0). \tag{6}$$

Let $\tilde{g}^t_i(\boldsymbol{Z}, \boldsymbol{V}^t_{\boldsymbol{\Theta}}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) := g^t_i(\boldsymbol{V}^t_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))$ and let $\partial_{1i}\tilde{g}^t_i$ be the partial derivative (Jacobian) w.r.t. the ith row of the first argument. Then, the state evolution matrices are defined recursively as follows:

$$\boldsymbol{\nu}^{t+1}_{\boldsymbol{B}} := \lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \partial_{1i}\tilde{g}^t_i(\boldsymbol{Z}, \boldsymbol{V}^t_{\boldsymbol{\Theta}}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})\right], \tag{7}$$

$$\boldsymbol{\kappa}^{s+1,t+1}_{\boldsymbol{B}} := \lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[g^s\left(\boldsymbol{V}^s_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})\right)^\top g^t\left(\boldsymbol{V}^t_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})\right)\right], \tag{8}$$

$$\boldsymbol{\nu}^{t+1}_{\boldsymbol{\Theta}} := \frac{1}{\delta}\lim_{p\to\infty}\frac{1}{p}\mathbb{E}\left[\boldsymbol{B}^\top f^{t+1}(\boldsymbol{V}^{t+1}_{\boldsymbol{B}})\right], \tag{9}$$

$$\boldsymbol{\kappa}^{s+1,t+1}_{\boldsymbol{\Theta}} := \frac{1}{\delta}\lim_{p\to\infty}\frac{1}{p}\mathbb{E}\left[\left(f^{s+1}(\boldsymbol{V}^{s+1}_{\boldsymbol{B}}) - \boldsymbol{B}\boldsymbol{\rho}^{-1}\boldsymbol{\nu}^{s+1}_{\boldsymbol{\Theta}}\right)^\top \left(f^{t+1}(\boldsymbol{V}^{t+1}_{\boldsymbol{B}}) - \boldsymbol{B}\boldsymbol{\rho}^{-1}\boldsymbol{\nu}^{t+1}_{\boldsymbol{\Theta}}\right)\right]. \tag{10}$$

The expectations above are taken with respect to $\boldsymbol{Z}, \boldsymbol{V}^t_{\boldsymbol{\Theta}}, \boldsymbol{V}^s_{\boldsymbol{\Theta}}, \boldsymbol{V}^{t+1}_{\boldsymbol{B}}$ and $\boldsymbol{V}^{s+1}_{\boldsymbol{B}}$, and depend on g^t, f^t, $\boldsymbol{B}$, $\boldsymbol{\varepsilon}$, and $\boldsymbol{\Psi}$. The limits in (7)–(10) exist under suitable regularity conditions on f_t, g_t, and on the limiting empirical distributions on $\boldsymbol{B}, \boldsymbol{\varepsilon}$; see Appendix B. The dependence on $\boldsymbol{B}$, $\boldsymbol{\varepsilon}$ can also be removed under these conditions – this is discussed in the next subsection.

3.1. State Evolution Characterization of AMP Iterates

Recall that the matrices (6)–(10) are used to define the random variables $(\boldsymbol{V}^t_{\boldsymbol{\Theta}}, \boldsymbol{V}^{t+1}_{\boldsymbol{B}})$ in (4),(5). Through these quantities, we now give a precise characterization of the AMP iterates $(\boldsymbol{\Theta}^t, \boldsymbol{B}^{t+1})$ in the high-dimensional limit. Theorem 3.1 below shows that any real-valued pseudo-Lipschitz function of $(\boldsymbol{\Theta}^t, \boldsymbol{B}^{t+1})$ converges in probability to its expectation under the limiting random variables $(\boldsymbol{V}^t_{\boldsymbol{\Theta}}, \boldsymbol{V}^{t+1}_{\boldsymbol{B}})$. In addition to the model assumptions in Section 2, we make the following assumptions:

(A1) The following limits exist and are finite almost surely: $\lim_{p\to\infty}\|\boldsymbol{B}^0\|_F/\sqrt{p}$, $\lim_{p\to\infty}\|\boldsymbol{B}^\top\boldsymbol{B}\|_F/p$, and $\lim_{n\to\infty}\|\boldsymbol{\varepsilon}\|_2/\sqrt{n}$.

(A2) For each $t \geq 0$, let $\tilde{g}^t : (\boldsymbol{u}, \boldsymbol{z}) \mapsto g^t(\boldsymbol{u}, q(\boldsymbol{z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))$, where $(\boldsymbol{u}, \boldsymbol{z}) \in \mathbb{R}^{n\times L} \times \mathbb{R}^{n\times L}$. For each $i \in [n], j \in [p]$, the following functions are uniformly pseudo-Lipschitz: $f^t, \tilde{g}^t, \mathrm{d}_j f^t_j, \partial_{1i}\tilde{g}^t_i$.

(A3) For $s, t \geq 0$, the limits in (6)–(10) exist and are finite.

Assumptions $(A1)-(A3)$ are natural extensions of classical AMP results which assume separable signals and denoising functions. They are similar to those required by existing works on non-separable AMP (Berthier et al., 2019; Gerbelot & Berthier, 2023), and generalize these to the model (2), with a matrix signal $\boldsymbol{B}$ and an auxiliary vector $\boldsymbol{\Psi} \in [L]^n$.

Theorem 3.1. *Consider the AMP in* (3) *for the Change Point Regression model in* (2)*, with the model assumptions in Section 2 as well as* $(A1) - (A3)$*. Then for* $t \geq 0$

and any sequence of uniformly pseudo-Lipschitz functions $\boldsymbol{\varphi}_n : \mathbb{R}^{n\times(L(t+1)+2)} \to \mathbb{R}$ *and* $\boldsymbol{\varphi}_p : \mathbb{R}^{p\times(L(t+2))} \to \mathbb{R}$,

$$\boldsymbol{\varphi}_n(\boldsymbol{\Theta}^0, \ldots, \boldsymbol{\Theta}^t, \boldsymbol{y}, \boldsymbol{\Psi}) \overset{\mathbb{P}}{\simeq} \mathbb{E}_{\boldsymbol{V}_{\boldsymbol{\Theta}}^0, \ldots, \boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{Z}}\{\boldsymbol{\varphi}_n(\boldsymbol{V}_{\boldsymbol{\Theta}}^0, \ldots, \boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}), \boldsymbol{\Psi})\}, \tag{11}$$

$$\boldsymbol{\varphi}_p(\boldsymbol{B}^1, \ldots, \boldsymbol{B}^{t+1}, \boldsymbol{B}) \overset{\mathbb{P}}{\simeq} \mathbb{E}_{\boldsymbol{V}_{\boldsymbol{B}}^1, \ldots, \boldsymbol{V}_{\boldsymbol{B}}^{t+1}}\{\boldsymbol{\varphi}_p(\boldsymbol{V}_{\boldsymbol{B}}^1, \ldots, \boldsymbol{V}_{\boldsymbol{B}}^{t+1}, \boldsymbol{B})\}, \tag{12}$$

as $n, p \to \infty$ *with* $n/p \to \delta$*, where the random variables* $\boldsymbol{Z}, \boldsymbol{V}_{\boldsymbol{\Theta}}^t$ *and* $\boldsymbol{V}_{\boldsymbol{B}}^{t+1}$ *are defined in* (4), (5).

The proof of the theorem is given in Appendix A. It involves reducing the AMP in (3) to a variant of the symmetric AMP iteration analyzed in (Gerbelot & Berthier, 2023, Lemma 14). We use a generalized version of their iteration which allows for the auxiliary quantities $\boldsymbol{\Psi}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})$ to be included. Theorem 3.1 implies that any pseudo-Lipschitz function $\boldsymbol{\varphi}_n$ of $(\boldsymbol{\Theta}^t, \boldsymbol{y})$ will converge in probability to a quantity involving an expectation over $(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{Z})$. An analogous statement holds for $(\boldsymbol{B}^{t+1}, \boldsymbol{B})$.

Useful Choices of $\boldsymbol{\varphi}_n, \boldsymbol{\varphi}_p$ Using Theorem 3.1, we can evaluate performance metrics such as the mean squared error between the signal matrix $\boldsymbol{B}$ and the estimate $\hat{\boldsymbol{B}}^t = f^t(\boldsymbol{B}^t)$. Taking $\varphi_p(\boldsymbol{B}^t, \boldsymbol{B}) = \|f^t(\boldsymbol{B}^t) - \boldsymbol{B}\|_F^2/p$ leads to $\|f^t(\boldsymbol{B}^t) - \boldsymbol{B}\|_F^2/p \overset{\mathbb{P}}{\simeq} \mathbb{E}[\|f^t(\boldsymbol{V}_{\boldsymbol{B}}^t) - \boldsymbol{B}\|_F^2/p]$, where the RHS can be precisely computed under suitable assumptions. In Section 3.3, we will choose φ_n to capture metrics of interest for estimating change points.

Special Cases Theorem 3.1 recovers two known special cases of (separable) AMP results, with complete convergence replaced by convergence in probability: linear regression when $L = 1$ (Feng et al., 2022), and mixed linear regression where $L > 1$ and the empirical distributions of the rows of $\boldsymbol{B}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}$ converge weakly to laws of well-defined random variables (Tan & Venkataramanan, 2023).

Dependence of State Evolution on $\boldsymbol{B}, \boldsymbol{\varepsilon}, \boldsymbol{\Psi}$ The dependence of the state evolution parameters in (7)-(10) on $\boldsymbol{B}, \boldsymbol{\varepsilon}$ can be removed under reasonable assumptions. A standard assumption in the AMP literature (Feng et al., 2022) is:

(S0) As $n, p \to \infty$, the empirical distributions of $\{\boldsymbol{B}_j\}_{j\in[p]}$ and $\{\varepsilon_i\}_{i\in[n]}$ converge weakly to laws $\mathbb{P}_{\bar{\boldsymbol{B}}}$ and $\mathbb{P}_{\bar{\varepsilon}}$, respectively, with bounded second moments.

In Appendix B, we give conditions on f_t, g_t, which together with (S0), allow the state evolution equations to be simplified and written in terms of $\bar{\boldsymbol{B}} \sim \mathbb{P}_{\bar{\boldsymbol{B}}}$ and $\bar{\varepsilon} \sim \mathbb{P}_{\bar{\varepsilon}}$ instead of $(\boldsymbol{B}, \boldsymbol{\varepsilon})$. We believe that the dependence of the state evolution on the signal configuration vector $\boldsymbol{\Psi}$ is fundamental. Since the entries of $\boldsymbol{\Psi}$ change value only at a finite number of change points, the state evolution parameters will depend on the limiting fractional values of these change points; see (S1) in Appendix B. This is also consistent with recent change point regression literature, where the limiting distribution of the change point estimators in (Xu et al., 2022) is shown to be a function of the data generating mechanism.

3.2. Choosing the Denoising Functions f^t, g^t

The performance of the AMP algorithm (3) is determined by the functions $\{f^{t+1}, g^t\}_{t\geq 0}$. We now describe how these functions can be chosen based on the state evolution recursion to maximize estimation performance. Using $(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{V}_{\boldsymbol{B}}^{t+1})$ defined in (4)–(5), we define the matrices

$$\tilde{\boldsymbol{Z}}^t := \boldsymbol{V}_{\boldsymbol{\Theta}}^t(\boldsymbol{\rho}^{-1}\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t)^{-1} = \boldsymbol{Z} + \boldsymbol{G}_{\boldsymbol{\Theta}}^t(\boldsymbol{\rho}^{-1}\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t)^{-1}, \tag{13}$$

$$\tilde{\boldsymbol{B}}^{t+1} := \boldsymbol{V}_{\boldsymbol{B}}^{t+1}(\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1})^{-1} = \boldsymbol{B} + \boldsymbol{G}_{\boldsymbol{B}}^{t+1}(\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1})^{-1}, \tag{14}$$

where for $i \in [n], j \in [p]$ we have $\boldsymbol{Z}_i \overset{i.i.d}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{\rho})$, $\boldsymbol{G}_{\boldsymbol{\Theta},i}^t \overset{i.i.d}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t,t})$, and $\boldsymbol{G}_{\boldsymbol{B},j}^t \overset{i.i.d}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{\kappa}_{\boldsymbol{B}}^{t,t})$. (If the inverse doesn't exist we post-multiply by the pseudo-inverse). A natural objective is to minimize the trace of the covariance of the "noise" matrices in (13) and (14), given by

$$\mathrm{Tr}\left(\frac{1}{n}\sum_{i=1}^n \mathrm{Cov}\left(\tilde{\boldsymbol{Z}}_i^t - \boldsymbol{Z}_i\right)\right) = \mathrm{Tr}\left(\left((\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t)^{-1}\boldsymbol{\rho}\right)^\top \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t,t}(\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t)^{-1}\boldsymbol{\rho}\right), \tag{15}$$

$$\mathrm{Tr}\left(\frac{1}{p}\sum_{j=1}^p \mathrm{Cov}\left(\tilde{\boldsymbol{B}}_i^{t+1} - \boldsymbol{B}_i\right)\right) = \mathrm{Tr}\left([(\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1})^{-1}]^\top \boldsymbol{\kappa}_{\boldsymbol{B}}^{t+1,t+1}(\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1})^{-1}\right), \tag{16}$$

where we recall from (7)–(10) that $\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t, \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t,t}$ are defined by f^t, and $\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1}, \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t+1,t+1}$ are defined by g^t.

For $t \geq 1$, we would like to iteratively choose the denoisers f^t, g^t to minimize the quantities on the RHS of (15) and (16). However, as discussed above $\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t, \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t,t}, \boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1}, \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t+1,t+1}$ depend on the unknown $\boldsymbol{\Psi}$, so any denoiser construction based quantities cannot be executed in practice.

Ensemble State Evolution To remove the dependence on $\boldsymbol{\Psi}$ in the state evolution equations, we can postulate randomness over this variable and take expectations accordingly. Indeed, assume that (S0) holds, and postulate a prior distribution $\pi_{\bar{\boldsymbol{\Psi}}}$ over $[L]^n$. For example, $\pi_{\bar{\boldsymbol{\Psi}}}$ may be the uniform distribution over all the signal configuration vectors with two change points that are at least $n/10$ apart. We emphasize that our theoretical results do not assume that the true $\boldsymbol{\Psi}$ is drawn according to $\pi_{\bar{\boldsymbol{\Psi}}}$. Rather, the prior $\pi_{\bar{\boldsymbol{\Psi}}}$ allows us to encode any knowledge we may have about the change

point locations, and use it to define efficient AMP denoisers f^t, g^t. This is done via the following *ensemble state evolution* recursion, defined in terms of the independent random variables $\bar{\boldsymbol{\Psi}} \sim \mathbb{P}_{\bar{\boldsymbol{\Psi}}}$, $\bar{\boldsymbol{B}} \sim \mathbb{P}_{\bar{\boldsymbol{B}}}$, and $\bar{\varepsilon} \sim \mathbb{P}_{\bar{\varepsilon}}$.

Starting with initialization $\bar{\boldsymbol{\nu}}^0_{\boldsymbol{\Theta}} := \mathbf{0}$, $\bar{\boldsymbol{\kappa}}^{0,0}_{\boldsymbol{\Theta}} := \lim_{p\to\infty} \frac{1}{\delta p} f^0(\boldsymbol{B}^0)^\top f^0(\boldsymbol{B}^0)$, for $t \geq 0$ define:

$$\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}} := \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\partial_1 \tilde{g}^t_i(\boldsymbol{Z}_1, (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, \bar{\Psi}_i, \bar{\varepsilon})\right] \quad (17)$$

$$\bar{\boldsymbol{\kappa}}^{t+1,t+1}_{\boldsymbol{B}} := \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\Big[g^t_i\left((\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon})\right) g^t_i\left((\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon})\right)^\top\Big], \quad (18)$$

$$\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{\Theta}} := \frac{1}{\delta} \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \mathbb{E}\left[\bar{\boldsymbol{B}} f^{t+1}_j(\bar{\boldsymbol{V}}^{t+1}_{\boldsymbol{B}})^\top\right], \quad (19)$$

$$\bar{\boldsymbol{\kappa}}^{t+1,t+1}_{\boldsymbol{\Theta}} := \frac{1}{\delta} \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \mathbb{E}\Big[\left(f^{t+1}_j(\bar{\boldsymbol{V}}^{t+1}_{\boldsymbol{B}}) - (\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{\Theta}})^\top \boldsymbol{\rho}^{-1} \bar{\boldsymbol{B}}\right) \left(f^{t+1}_j(\bar{\boldsymbol{V}}^{t+1}_{\boldsymbol{B}}) - (\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{\Theta}})^\top \boldsymbol{\rho}^{-1} \bar{\boldsymbol{B}}\right)^\top\Big], \quad (20)$$

where

$$\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}} := \boldsymbol{Z}\boldsymbol{\rho}^{-1}\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}} + \bar{\boldsymbol{G}}^t_{\boldsymbol{\Theta}} \in \mathbb{R}^{n\times L}, \quad (21)$$

$$\bar{\boldsymbol{V}}^{t+1}_{\boldsymbol{B}} := (\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}})^\top \bar{\boldsymbol{B}} + (\bar{\boldsymbol{G}}^{t+1}_{\boldsymbol{B}})_1 \in \mathbb{R}^L, \quad (22)$$

and for $i \in [n], j \in [p]$ we have that $\boldsymbol{Z}_i \overset{i.i.d}{\sim} \mathsf{N}(\mathbf{0}, \boldsymbol{\rho})$, $\bar{\boldsymbol{G}}^t_{\boldsymbol{\Theta},i} \overset{i.i.d}{\sim} \mathsf{N}(\mathbf{0}, \bar{\boldsymbol{\kappa}}^{t,t}_{\boldsymbol{\Theta}})$ and $\bar{\boldsymbol{G}}^t_{\boldsymbol{B},j} \overset{i.i.d}{\sim} \mathsf{N}(\mathbf{0}, \bar{\boldsymbol{\kappa}}^{t,t}_{\boldsymbol{B}})$.

When $\pi_{\bar{\boldsymbol{\Psi}}}$ is a unit mass on the true configuration $\boldsymbol{\Psi}$, (17)–(20) reduce to the simplified state evolution in Appendix B. The limits in (17)–(20) exist under suitable regularity conditions on f^t, g^t, such as those in Appendix B.

We now propose a construction of f^t, g^t based on minimizing the following alternative objectives to (15)–(16):

$$\mathrm{Tr}\left(\left((\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})^{-1}\boldsymbol{\rho}\right)^\top \bar{\boldsymbol{\kappa}}^{t,t}_{\boldsymbol{\Theta}} (\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})^{-1}\boldsymbol{\rho}\right), \quad (23)$$

$$\mathrm{Tr}\left([(\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}})^{-1}]^\top \bar{\boldsymbol{\kappa}}^{t+1,t+1}_{\boldsymbol{B}} (\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}})^{-1}\right), \quad (24)$$

where the deterministic matrices $\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}}$, $\bar{\boldsymbol{\kappa}}^{t,t}_{\boldsymbol{\Theta}}$, $\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}}$, $\bar{\boldsymbol{\kappa}}^{t+1,t+1}_{\boldsymbol{B}} \in \mathbb{R}^{L\times L}$ are defined in (17)–(20).

Proposition 3.2. *Assume the limits in* (17)–(20) *exist. Then, for* $t \geq 1$*:*

1. *Given* $\bar{\boldsymbol{\nu}}^t_{\boldsymbol{B}}$, $\bar{\boldsymbol{\kappa}}^{t,t}_{\boldsymbol{B}}$, *the quantity* (23) *is minimized when*

$$f^t_j(\boldsymbol{U}) = f^{*t}_j(\boldsymbol{U}) := \mathbb{E}[\bar{\boldsymbol{B}} | \bar{\boldsymbol{V}}^t_{\boldsymbol{B}} = \boldsymbol{U}_j], \quad (25)$$

for $\boldsymbol{U} \in \mathbb{R}^{p\times L}, j \in [p]$.

2. *Given* $\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}}, \bar{\boldsymbol{\kappa}}^{t,t}_{\boldsymbol{\Theta}}$, *the quantity* (24) *is minimized when*

$$g^t_i(\boldsymbol{V}, \boldsymbol{u}) = g^{*t}_i(\boldsymbol{V}, \boldsymbol{u}) := \left[\mathrm{Cov}\left(\boldsymbol{Z}_1 | \bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta},1}\right)\right]^{-1} \cdot \left(\mathbb{E}[\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1 = \boldsymbol{V}_i, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i] - \mathbb{E}[\boldsymbol{Z}_1 | \bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta},1} = \boldsymbol{V}_1]\right), \quad (26)$$

for $\boldsymbol{V} \in \mathbb{R}^{n\times L}, \boldsymbol{u} \in \mathbb{R}^n, i \in [n]$.

The proof of Proposition 3.2, given in Appendix C.1, is similar to the derivation of the optimal denoisers for mixed regression in (Tan & Venkataramanan, 2023), with a few key differences in the derivation of g^{*t}, which is not separable in the change point setting. With a product distribution on $\pi_{\bar{\boldsymbol{\Psi}}}$, we recover mixed regression and (25)–(26) reduce to the optimal denoisers in (Tan & Venkataramanan, 2023).

The denoiser f^{*t} is separable and can be easily computed for sufficiently regular distributions $\mathbb{P}_{\bar{\boldsymbol{B}}}$ such as discrete, Gaussian, or Bernoulli-Gaussian distributions. In Appendix C.2, we show how g^{*t} can also be efficiently computed for Gaussian $\bar{\varepsilon}$. As detailed in Appendix F, f^{*t} and g^{*t} can be computed in $O(nL^3)$ time, yielding a total computational complexity of $O(npL^3)$ for AMP with these denoisers.

3.3. Change Point Estimation and Inference

We now show how the AMP algorithm can be used for estimation and inference of the change points $\{\eta_1, \ldots, \eta_{L^*-1}\}$. We first define some notation. Let $\mathcal{X} \subset [L]^n$ be the set of all piece-wise constant vectors with respect to $i \in [n]$ with at most $(L-1)$ jumps. This set includes all possible instances of the signal configuration vector $\boldsymbol{\Psi}$. Let the function $U : \boldsymbol{\eta} \mapsto \boldsymbol{\Psi}$ denote the one-to-one mapping between change point vectors $\boldsymbol{\eta}$ and signal configuration vectors $\boldsymbol{\Psi}$. For a vector $\hat{\boldsymbol{\eta}}$, we let $|\hat{\boldsymbol{\eta}}|$ denote its dimension (number of elements).

Change Point Estimation Theorem 3.1 states that $(\boldsymbol{\Theta}^t, \boldsymbol{y})$ converges in a specific sense to the random variables $(\boldsymbol{V}^t_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})) = (\boldsymbol{Z}\boldsymbol{\rho}^{-1}\boldsymbol{\nu}^t_{\boldsymbol{\Theta}} + \boldsymbol{G}^t_{\boldsymbol{\Theta}}, \boldsymbol{Z}_{[:,\boldsymbol{\Psi}]} + \boldsymbol{\varepsilon})$, whose distribution crucially captures information about $\boldsymbol{\Psi}$. Hence, it is natural to consider estimators for $\boldsymbol{\eta}$ of the form $\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y})$, one example being an estimator that searches for a signal configuration vector $\boldsymbol{\psi} \in \mathcal{X}$ such that $\boldsymbol{\Theta}^t$ indexed along $\boldsymbol{\psi}$ has the largest correlation with the observed vector $\boldsymbol{y}$. That is,

$$\hat{\boldsymbol{\Psi}}(\boldsymbol{\Theta}^t, \boldsymbol{y}) = \arg\max_{\boldsymbol{\psi}\in\mathcal{X}} \sum_{i=1}^{n} (\boldsymbol{\Theta}^t)_{i,\psi_i} \cdot y_i, \quad (27)$$

and $\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}) = U^{-1}(\hat{\boldsymbol{\Psi}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))$. A common metric for evaluating the accuracy of change point estimators is the Hausdorff distance (Wang & Samworth, 2018; Xu et al., 2022; Li et al., 2023b). The Hausdorff distance between

two non-empty subsets X, Y of $\mathbb{R}$ is

$$d_H(X,Y) = \max\left\{\sup_{x\in X} d(x,Y),\ \sup_{y\in Y} d(X,y)\right\},$$

where $d(x,Y) := \min_{y\in Y}\|x-y\|_2$. The Hausdorff distance is a metric, and can be viewed as the largest of all distances from a point in X to its closest point in Y and vice versa. We interpret the Hausdorff distance between $\boldsymbol{\eta}$ and an estimate $\hat{\boldsymbol{\eta}}$ as the Hausdorff distance between the sets formed by their elements. The following theorem states that any well-behaved estimator $\hat{\boldsymbol{\eta}}$ produced using the AMP iterate $\boldsymbol{\Theta}^t$ admits a precise asymptotic characterization in terms of Hausdorff distance and size.

Proposition 3.3. *Consider the AMP in (3). Suppose the model assumptions in Section 2 as well as (A1) − (A3) are satisfied. Let $\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y})$ be an estimator such that $U(\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))$ is uniformly pseudo-Lipschitz. Then:*

$$\frac{d_H(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))}{n} \overset{\mathbb{P}}{\simeq} \mathbb{E}_{\boldsymbol{V}^t_{\boldsymbol{\Theta}}, \boldsymbol{Z}} \frac{d_H(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}(\boldsymbol{V}^t_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, U(\boldsymbol{\eta}), \boldsymbol{\varepsilon})))}{n} \tag{28}$$

Moreover, if the nth component of $U(\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))$ is uniformly pseudo-Lipschitz, then:

$$|(L^* - 1) - |\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y})|| \overset{\mathbb{P}}{\simeq} \mathbb{E}_{\boldsymbol{V}^t_{\boldsymbol{\Theta}}, \boldsymbol{Z}}\left|(L^* - 1) - |\hat{\boldsymbol{\eta}}(\boldsymbol{V}^t_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, U(\boldsymbol{\eta}), \boldsymbol{\varepsilon}))|\right|. \tag{29}$$

The proof is given in Appendix D.1. The proof of (28) involves showing that $d_H(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))/n$ is uniformly pseudo-Lipschitz. For each $\boldsymbol{\eta}$, Proposition 3.3 precisely characterizes the asymptotic Hausdorff distance and size errors for a large class of estimators $\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y})$.

Uncertainty Quantification The random variable $\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}} = \boldsymbol{Z}\boldsymbol{\rho}^{-1}\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}} + \bar{\boldsymbol{G}}^t_{\boldsymbol{\Theta}}$ in (21), combined with an observation of the form $q(\boldsymbol{Z}, \boldsymbol{\Psi}, \bar{\boldsymbol{\varepsilon}}) = \boldsymbol{Z}_{[:,\boldsymbol{\Psi}]} + \bar{\boldsymbol{\varepsilon}}$, yields a recipe for constructing a posterior distribution over $\boldsymbol{\Psi}$. Using the prior $\pi_{\bar{\boldsymbol{\Psi}}}$, the posterior is:

$$p_{\boldsymbol{\Psi}|\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \bar{\boldsymbol{\varepsilon}})}(\boldsymbol{\psi}|\boldsymbol{V}, \boldsymbol{u}) = \frac{\pi_{\bar{\boldsymbol{\Psi}}}(\boldsymbol{\psi})\mathcal{L}(\boldsymbol{V}, \boldsymbol{u}|\boldsymbol{\psi})}{\sum_{\tilde{\boldsymbol{\psi}}} \pi_{\bar{\boldsymbol{\Psi}}}(\tilde{\boldsymbol{\psi}})\mathcal{L}(\boldsymbol{V}, \boldsymbol{u}|\tilde{\boldsymbol{\psi}})}, \tag{30}$$

where $\boldsymbol{V} \in \mathbb{R}^{n\times L}$, $\boldsymbol{u} \in \mathbb{R}^n$, and $\mathcal{L}$ represents the likelihood of $(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, \boldsymbol{\psi}, \bar{\boldsymbol{\varepsilon}}))$ given $\boldsymbol{\psi} \in \mathcal{X}$. Under the assumption that $\bar{\varepsilon}_i \overset{i.i.d}{\sim} \mathsf{N}(0, \sigma^2)$ for some $\sigma > 0$, the likelihood can be computed in closed form using the Gaussian likelihood $\mathcal{L}_1$ in (92), with $\mathcal{L}(\boldsymbol{V}, \boldsymbol{u}|\boldsymbol{\psi}) = \prod_{i=1}^n \mathcal{L}_1(\boldsymbol{V}_i, u_i|\psi_i)$.

Since Theorem 3.1 states that $(\boldsymbol{\Theta}^t, \boldsymbol{y})$ converges in a specific sense to $(\boldsymbol{V}^t_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))$, we can obtain an approximate posterior density over $\boldsymbol{\Psi}$ by plugging in $(\boldsymbol{\Theta}^t, \boldsymbol{y})$ for $(\boldsymbol{V}, \boldsymbol{u})$ in (30). Our theory then allows us to characterize the asymptotic form of this approximate posterior distribution.

Proposition 3.4. *Suppose the model assumptions in Section 2 as well as (A1) − (A3) are satisfied. Assume that the posterior density $p(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y}) := p_{\boldsymbol{\Psi}|\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \bar{\boldsymbol{\varepsilon}})}(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y})$ in (30) is uniformly pseudo-Lipschitz with respect to $\boldsymbol{\Theta}^t, \boldsymbol{y}$, for each $\boldsymbol{\psi} \in \mathcal{X}$. Then, for $\boldsymbol{\psi} \in \mathcal{X}$, we have that:*

$$p(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y}) \overset{\mathbb{P}}{\simeq} \mathbb{E}_{\boldsymbol{V}^t_{\boldsymbol{\Theta}}, \boldsymbol{Z}}\left[p(\boldsymbol{\psi}|\boldsymbol{V}^t_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))\right]. \tag{31}$$

The proof, given in Appendix D.2, is a direct application of Theorem 3.1. The state evolution predictions on the RHS of (28) and (31) can be computed under reasonable assumptions, as outlined in Appendix B.

4. Experiments

In this section, we demonstrate the estimation and inference capabilities of the AMP algorithm in a range of settings. After running AMP, the signal configuration is estimated using the approximate Maximum a Posteriori (MAP) estimate:

$$\hat{\boldsymbol{\Psi}} = \arg\max_{\boldsymbol{\psi}\in\mathcal{X}} p_{\boldsymbol{\Psi}|\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \bar{\boldsymbol{\varepsilon}})}(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y}).$$

and $\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}) = U^{-1}(\hat{\boldsymbol{\Psi}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))$. A Python implementation of our algorithm and code to run the experiments is available at (Arpino & Liu, 2024).

For all experiments, we use i.i.d. Gaussian covariates: $\boldsymbol{X}_i \overset{i.i.d}{\sim} \mathsf{N}(0, \boldsymbol{I}_p/n)$ for $i \in [n]$, and study the performance on both synthetic and image signals. For synthetic data, the noise distribution is chosen to be $\mathbb{P}_{\bar{\varepsilon}} = \mathsf{N}(0, \sigma^2)$. The denoisers $\{g^t, f^{t+1}\}_{t\geq 0}$ in the AMP algorithm are chosen according to Proposition 3.2, unless otherwise stated. We use a uniform prior $\pi_{\bar{\Psi}}$ over all configurations with change points at least Δ apart, for some Δ that is a fraction of n. Error bars represent $\pm$ one standard deviation. Full implementation details are provided in Appendix E.

Figure 1 plots the Hausdorff distance normalized by n for varying δ, for two different change point configurations $\boldsymbol{\Psi}$. We choose $p = 600$, $\mathbb{P}_{\bar{\boldsymbol{B}}} = \mathsf{N}(\boldsymbol{0}, \boldsymbol{I})$, $\sigma = 0.1$, $\Delta = n/5$ and fix two true change points, whose locations are indicated in the legend. The algorithm uses $L = L^* = 3$. The state evolution prediction of Hausdorff distance closely matches the performance of AMP, verifying (28) in Proposition 3.3.

Figure 2 shows the approximate posterior on the change point locations, computed using $p_{\boldsymbol{\Psi}|\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \bar{\boldsymbol{\varepsilon}})}(\cdot|\boldsymbol{\Theta}^t, \boldsymbol{y})$. We observe strong agreement with the state evolution prediction, validating Proposition 3.4. The experiment uses $p = 400$, $\mathbb{P}_{\bar{\boldsymbol{B}}} = \mathsf{N}(\boldsymbol{0}, \boldsymbol{I})$, $\sigma = 0.1$, $\Delta = n/5$, the true change points are at $n/3$ and $8n/15$ (i.e., $L^* = 3$), and the algorithm uses $L = 3$. As δ increases, we observe that the approximate posterior concentrates around the ground truth.

Figure 3 shows the approximate posterior on the *number* of change points, i.e., $\sum_{\boldsymbol{\psi}\in\mathcal{S}} p_{\boldsymbol{\Psi}|\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \bar{\boldsymbol{\varepsilon}})}(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y})$ where

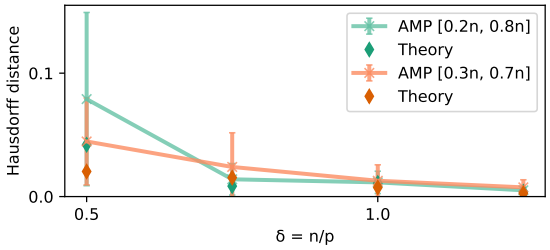


Figure 1. Hausdorff error vs. $\delta = n/p$ for two configurations.

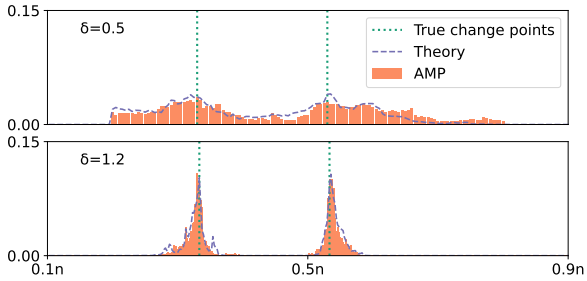


Figure 2. Approximate posterior vs. change point locations.

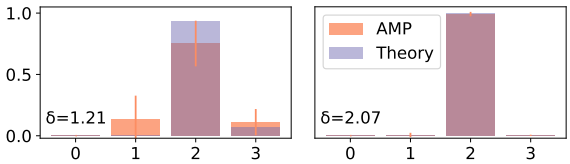


Figure 3. Approximate posterior on the number of change points $(L^* - 1)$. $L^* = 3, L = 4$.

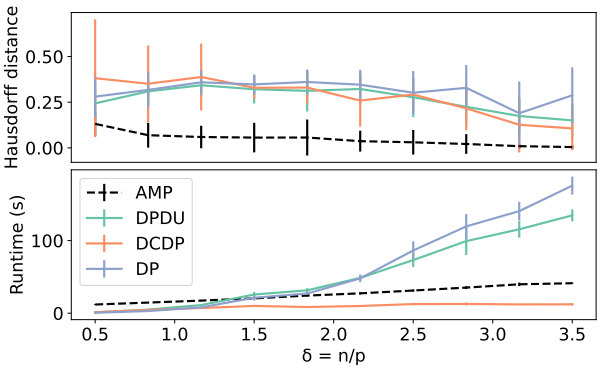


Figure 4. Comparison with DPDU, DCDP and DP for sparse prior $\mathbb{P}_{\bar{\boldsymbol{B}}} = 0.5\mathsf{N}(\mathbf{0}, \delta \boldsymbol{I}) + 0.5\delta_{\mathbf{0}}$. $L^* = L = 3$. Runtime shown is the average runtime per set of CV parameters.

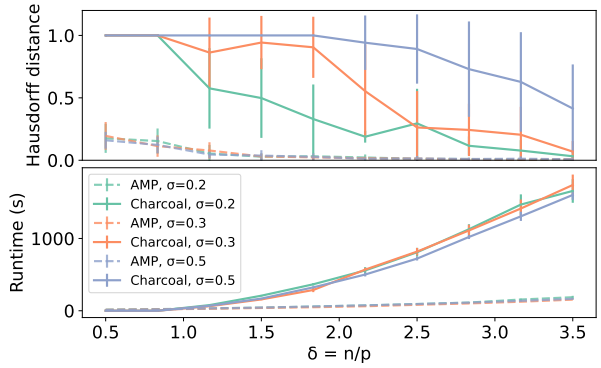


Figure 5. Comparison with `charcoal` for a sparse difference prior with sparsity level 0.5. $L^* = L = 3$.

$\mathcal{S}$ contains all configurations with a specified number of change points. We use $p = 200, \mathbb{P}_{\bar{\boldsymbol{B}}} = N(\mathbf{0}, \boldsymbol{I}), \sigma = 0.1, \Delta = n/10, L^* = 3, L = 4$, and a uniform prior over the number of change points (zero to three). We observe that the posterior concentrates around the ground truth for moderately large δ.

Figures 4 and 5 compare the performance of AMP against four state-of-the-art algorithms: the dynamic programming (DP) approach in (Rinaldo et al., 2021); dynamic programming with dynamic updates (DPDU) (Xu et al., 2022); divide and conquer dynamic programming (DCDP) (Li et al., 2023b); and a complementary-sketching-based algorithm called `charcoal` (Gao & Wang, 2022). Hyperparameters are chosen using cross validation (CV), as outlined in Section A.1 of (Li et al., 2023b). The first three algorithms, designed for sparse signals, combine LASSO-type estimators with partitioning techniques based on dynamic programming. The `charcoal` algorithm is designed for the setting where the difference $\boldsymbol{\beta}^{(\eta_\ell)} - \boldsymbol{\beta}^{(\eta_{\ell+1})}$ between adjacent signals is sparse. None of these algorithms uses a prior on the change point locations, unlike AMP which can flexibly incorporate both priors via $\mathbb{P}_{\bar{\boldsymbol{B}}}$ and $\pi_{\bar{\boldsymbol{\Psi}}}$.

Figure 4 uses $p = 200, \sigma = 0.1, \Delta = n/10, L^* = L = 3$ and a sparse Bernoulli-Gaussian signal prior $\mathbb{P}_{\bar{\boldsymbol{B}}} = 0.5\mathsf{N}(\mathbf{0}, \delta \boldsymbol{I}) + 0.5\delta_{\mathbf{0}}$. AMP assumes no knowledge of the true sparsity level 0.5 and estimates the sparsity level using CV over a set of values not including the ground truth (details in Appendix E). Figure 5 uses $p = 300, \Delta = n/10, L^* = L = 3$ and a Gaussian sparse difference prior with sparsity level 0.5 (described in (84)-(86)). AMP is run assuming a mismatched sparsity level of 0.9 and a mismatched magnitude for the sparse difference vector (details in Appendix E). Figures 4 and 5 show that AMP consistently achieves the lowest Hausdorff distance among all algorithms and outperforms most algorithms in runtime. Figures 7 and 8 in Appendix E show results from an additional set of experiments comparing AMP with DCDP (the fastest algorithm in Figure 4). Figure 7 shows the performance of AMP with different change point priors $\pi_{\bar{\boldsymbol{\Psi}}}$ and suboptimal denoising functions, e.g., soft threshold-

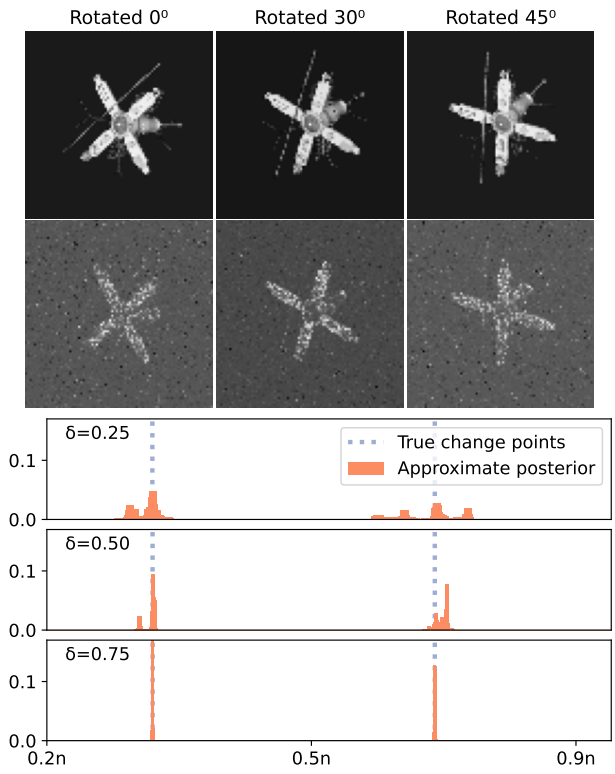


Figure 6. Top: Ground truth images (first row) and reconstruction from AMP (second row) at $\delta = 0.75$. Bottom: Approximate posterior vs. fractional change point locations for different δ.

ing. Figure 8 demonstrates the favourable runtime scaling of AMP over DCDP with respect to p, due to the LASSO computations involved in DCDP.

Compressed Sensing with Change Points In Figure 6, we consider noiseless compressed sensing, where the signals $\{\boldsymbol{\beta}^{(i)}\}_{i\in[n]}$ are rotated versions of a $(255, 255)$ sparse grayscale image used by Schniter & Rangan (2014). The fraction of nonzero components in the image is 8645/50625. We downsample the original image by a factor of three and flatten, yielding an operation dimension of $p = 85^2 = 7225$. We set $\{\boldsymbol{\beta}^{(i)}\}_{i=1}^{0.3n}$ to be the image, $\{\boldsymbol{\beta}^{(i)}\}_{i=0.3n}^{0.7n}$ to be a $30°$ rotated version, and $\{\boldsymbol{\beta}^{(i)}\}_{i=0.7n}^{n}$ to be a $45°$ rotated version. We run AMP with a Bernoulli-Gaussian prior $\mathbb{P}_{\bar{\boldsymbol{B}}}$, $\Delta = n/4$ and $L = L^* = 3$. Figure 6 shows image reconstructions along with the approximate posterior $p_{\boldsymbol{\Psi}|\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}},q(\boldsymbol{Z},\boldsymbol{\Psi},\bar{\boldsymbol{\varepsilon}})}(\cdot|\boldsymbol{\Theta}^t, \boldsymbol{y})$. The approximate posterior concentrates around the true change point locations as δ increases, even when the image reconstructions are approximate. The experiment took one hour to complete on an Apple M1 Max chip, whereas competing algorithms did not return an output within 2.5 hours, due to the larger signal dimension compared to Figure 4 ($p = 7225$ vs $p = 200$).

Acknowledgements

We would like to thank anonymous reviewers for useful comments, and Cédric Gerbelot for pointing to their work on graph-based approximate message passing iterations. G. Arpino is supported by a Cambridge Trust Scholarship. X. Liu is supported by a Schlumberger Cambridge International Scholarship.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Andreou, E. and Ghysels, E. Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics*, 2002.

Arpino, G. and Liu, X. AMP implementation for change point inference in high-dimensional linear regression. https://github.com/gabrielarpino/AMP_chgpt_lin_reg, 2024.

Aston, J. A. D. and Kirch, C. Evaluating stationarity via change-point alternatives with applications to fmri data. *The Annals of Applied Statistics*, 2012.

Bai, Y. and Safikhani, A. A unified framework for change point detection in high-dimensional linear models. *Statistica Sinica*, 2023.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Barbier, J., Macris, N., and Rush, C. All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Bayati, M. and Montanari, A. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57: 764–785, 2011.

Berthier, R., Montanari, A., and Nguyen, P.-M. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 2019.

Bhattacharjee, M., Banerjee, M., and Michailidis, G. Change Point Estimation in a Dynamic Stochastic Block Model. *Journal of Machine Learning Research*, 2020.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018. 0.3.13.

Braun, J. V., Braun, R., and Müller, H.-G. Multiple change-point fitting via quasilikelihood, with application to dna sequence segmentation. *Biometrika*, 2000.

Cho, H. and Fryzlewicz, P. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2015.

Cho, H., Kley, T., and Li, H. Detection and inference of changes in high-dimensional linear regression with non-sparse structures, 2024. arXiv:2402.06915.

Deshpande, Y. and Montanari, A. Information-theoretically optimal sparse PCA. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 2197–2201, 2014.

Donoho, D. L., Maleki, A., and Montanari, A. Message passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106:18914–18919, 2009.

Donoho, D. L., Javanmard, A., and Montanari, A. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Transactions on Information Theory*, 59(11):7434–7464, Nov. 2013.

Fan, Z. and Guan, L. Approximate ℓ_0-penalized estimation of piecewise-constant signals on graphs. *The Annals of Statistics*, 46(6B):3217 – 3245, 2018.

Fearnhead, P. Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and computing*, 16:203–213, 2006.

Feng, O. Y., Venkataramanan, R., Rush, C., and Samworth, R. J. A Unifying Tutorial on Approximate Message Passing. *Foundations and Trends® in Machine Learning*, 15 (4):335–536, 2022.

Fletcher, A. K. and Rangan, S. Iterative reconstruction of rank-one matrices in noise. *Information and Inference: A Journal of the IMA*, 7(3):531–562, 2018.

Fryzlewicz, P. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6): 2243–2281, 2014.

Gao, F. and Wang, T. Sparse change detection in high-dimensional linear regression, 2022. arXiv:2208.06326.

Gerbelot, C. and Berthier, R. Graph-based approximate message passing iterations. *Information and Inference: A Journal of the IMA*, 2023.

Javanmard, A. and Montanari, A. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2013.

Kabashima, Y. A CDMA multiuser detection algorithm on the basis of belief propagation. *Journal of Physics A: Mathematical and General*, 36(43):11111–11121, Oct 2003.

Kaul, A., Jandhyala, V. K., and Fotopoulos, S. B. An efficient two step algorithm for high dimensional change point regression models without grid search. *J. Mach. Learn. Res.*, 2019.

Krzakala, F., Mézard, M., Sausset, F., Sun, Y., and Zdeborová, L. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(8), 2012.

Lavergne, P. A Cauchy-Schwarz inequality for expectation of matrices. Discussion Papers dp08-07, Department of Economics, Simon Fraser University, November 2008.

Lee, S., Seo, M. H., and Shin, Y. The Lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2016.

Leonardi, F. and Bühlmann, P. Computationally efficient change point detection for high-dimensional regression, 2016. arXiv:1601.03704.

Lesieur, T., Krzakala, F., and Zdeborová, L. Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7): 073403, 2017.

Li, J., Fearnhead, P., Fryzlewicz, P., and Wang, T. Automatic Change-Point Detection in Time Series via Deep Learning, 2023a. arXiv:2211.03860.

Li, W., Wang, D., and Rinaldo, A. Divide and Conquer Dynamic Programming: An Almost Linear Time Change Point Detection Methodology in High Dimensions. In *Proceedings of the 40th International Conference on Machine Learning*, 2023b.

Liu, H., Gao, C., and Samworth, R. J. Minimax rates in sparse, high-dimensional change point detection. *The Annals of Statistics*, 49(2), 2021.

Londschien, M., Bühlmann, P., and Kovács, S. Random Forests for Change Point Detection. *Journal of Machine Learning Research*, 2023.

Lungu, V., Papageorgiou, I., and Kontoyiannis, I. Change-point detection and segmentation of discrete data using Bayesian context trees, 2022. arXiv: 2203.04341.

Ma, J. and Ping, L. Orthogonal AMP. *IEEE Access*, 5: 2020–2033, 2017.

Ma, Y., Rush, C., and Baron, D. Analysis of approximate message passing with non-separable denoisers and markov random field priors. *IEEE Transactions on Information Theory*, 65(11):7367–7389, 2019.

Madrid Padilla, O. H., Yu, Y., Wang, D., and Rinaldo, A. Optimal nonparametric multivariate change point detection and localization. *IEEE Transactions on Information Theory*, 68(3):1922–1944, 2022.

Maillard, A., Loureiro, B., Krzakala, F., and Zdeborová, L. Phase retrieval in high dimensions: Statistical and computational phase transitions. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Metzler, C. A., Maleki, A., and Baraniuk, R. G. From denoising to compressed sensing. *IEEE Transactions on Information Theory*, 62(9):5117–5144, 2016.

Mondelli, M. and Venkataramanan, R. Approximate message passing with spectral initialization for generalized linear models. *International Conference on Artificial Intelligence and Statistics*, pp. 397–405, 2021.

Montanari, A. and Venkataramanan, R. Estimation of low-rank matrices via approximate message passing. *Annals of Statistics*, 45(1):321–345, 2021.

Pandit, P., Sahraee-Ardakan, M., Rangan, S., Schniter, P., and Fletcher, A. K. Inference with deep generative priors in high dimensions. *IEEE Journal on Selected Areas in Information Theory*, 1(1):336–347, 2020.

Rangan, S. Generalized approximate message passing for estimation with random linear mixing. *IEEE International Symposium on Information Theory*, 2011.

Rangan, S., Schniter, P., and Fletcher, A. K. Vector approximate message passing. *IEEE Transactions on Information Theory*, 65(10):6664–6684, 2019.

Rinaldo, A., Wang, D., Wen, Q., Willett, R., and Yu, Y. Localizing Changes in High-Dimensional Regression Models. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Schniter, P. and Rangan, S. Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing*, 63(4):1043–1055, 2014.

Som, S. and Schniter, P. Compressive imaging using approximate message passing and a Markov-tree prior. *IEEE Transactions on Signal Processing*, 60(7):3439–3448, 2012.

Takeuchi, K. Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements. *IEEE Transactions on Information Theory*, 66(1):368–386, 2020.

Tan, N. and Venkataramanan, R. Mixed Regression via Approximate Message Passing. *Journal of Machine Learning Research*, 24(317):1–44, 2023.

Wang, D., Yu, Y., and Rinaldo, A. Univariate mean change point detection: Penalization, CUSUM and optimality. *Electronic Journal of Statistics*, pp. 1917 – 1961, 2020.

Wang, D., Yu, Y., and Rinaldo, A. Optimal change point detection and localization in sparse dynamic networks. *The Annals of Statistics*, 2021a.

Wang, D., Yu, Y., and Rinaldo, A. Optimal covariance change point localization in high dimensions. *Bernoulli*, 2021b.

Wang, D., Zhao, Z., Lin, K. Z., and Willett, R. Statistically and computationally efficient change point localization in regression settings. *Journal of Machine Learning Research*, 22(248):1–46, 2021c.

Wang, T. and Samworth, R. J. High Dimensional Change Point Estimation via Sparse Projection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2018.

Wang, T., Zhong, X., and Fan, Z. Universality of approximate message passing algorithms and tensor networks, 2022. arXiv:2206.13037.

Xu, H., Wang, D., Zhao, Z., and Yu, Y. Change point inference in high-dimensional regression models under temporal dependence, 2022. arXiv:2207.12453.

Yi, X., Caramanis, C., and Sanghavi, S. Alternating minimization for mixed linear regression. *International Conference on Machine Learning*, pp. 613–621, 2014.

Zhang, Y., Mondelli, M., and Venkataramanan, R. Precise asymptotics for spectral methods in mixed generalized linear models, 2022. arXiv:2211.11368.

Zhang, Y., Ji, H. C., Venkataramanan, R., and Mondelli, M. Spectral estimators for structured generalized linear models via approximate message passing, 2023. arXiv:2308.14507.

A. Proof of State Evolution

The proof of Theorem 3.1 relies on a generalization of Lemma 14 in (Gerbelot & Berthier, 2023), presented as Lemma A.1 below. Let $\boldsymbol{W}_0 \in \mathbb{R}^{N\times Q}$ be a matrix such that $\|\boldsymbol{W}_0^\top \boldsymbol{W}_0\|_F/N$ converges to a finite constant as $N \to \infty$, and let $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ be a symmetric GOE(N) matrix, independent of $\boldsymbol{W}_0$. Then Lemma 14 in (Gerbelot & Berthier, 2023) gives a state evolution result involving an AMP iteration whose denoising function takes as input the output of the generalized linear model $\varphi(\boldsymbol{A}\boldsymbol{W}_0)$, where $\varphi : \mathbb{R}^{N\times Q} \to \mathbb{R}^N$.

Lemma A.1 generalizes Lemma 14 of Gerbelot & Berthier (2023) in two ways: i) it allows for the inclusion of an auxiliary matrix $\boldsymbol{\Xi} \in \mathbb{R}^{N\times L_{\boldsymbol{\Xi}}}$ so that the generalized linear model in question is of the form $\varphi(\boldsymbol{A}\boldsymbol{W}_0, \boldsymbol{\Xi})$, and ii) the state evolution convergence result holds for pseudo-Lipschitz test functions taking both the auxiliary matrix $\boldsymbol{\Xi}$ and $\boldsymbol{A}\boldsymbol{W}_0$ as inputs.

We extend this lemma by considering an independent random matrix $\boldsymbol{\Xi} \in \mathbb{R}^{N\times L_{\boldsymbol{\Xi}}}$, serving as input to a set of pseudo-Lipschitz functions $\varphi : \mathbb{R}^{N\times(Q+L_{\boldsymbol{\Xi}})} \to \mathbb{R}^N$. We then analyze a similar AMP iteration whose denoising function $\tilde{f}^t$ takes $\varphi(\boldsymbol{A}\boldsymbol{W}_0, \boldsymbol{\Xi})$ as input instead of $\varphi(\boldsymbol{A}\boldsymbol{W}_0)$, initialized with an independent initializer $\mathbf{X}^0 \in \mathbb{R}^{N\times Q}$:

$$\boldsymbol{X}^{t+1} = \boldsymbol{A}\boldsymbol{M}^t - \boldsymbol{M}^{t-1}(\boldsymbol{b}^t)^\top \qquad \in \mathbb{R}^{N\times Q}, \tag{32}$$

$$\boldsymbol{M}^t = \tilde{f}^t(\varphi(\boldsymbol{A}\boldsymbol{W}_0, \boldsymbol{\Xi}), \boldsymbol{X}^t) \qquad \in \mathbb{R}^{N\times Q}, \tag{33}$$

$$\boldsymbol{b}^t = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \tilde{f}_i^t}{\partial \boldsymbol{X}_i^t}(\varphi(\boldsymbol{A}\boldsymbol{W}_0, \boldsymbol{\Xi}), \boldsymbol{X}^t) \qquad \in \mathbb{R}^{Q\times Q}. \tag{34}$$

Our result in Lemma A.1 presents an asymptotic characterization of (32)–(34) via the following state evolution recursion:

$$\boldsymbol{\nu}^0, \hat{\boldsymbol{\nu}}^0 = \mathbf{0}, \tag{35}$$

$$\boldsymbol{\kappa}^{0,0} = \lim_{N\to\infty} \frac{1}{N}\tilde{f}^0(\boldsymbol{X}^0)^\top \tilde{f}^0(\boldsymbol{X}^0), \tag{36}$$

$$\boldsymbol{\nu}^{t+1} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\boldsymbol{W}_0^\top \ \tilde{f}^t\left(\varphi(\boldsymbol{Z}_{\boldsymbol{W}_0}, \boldsymbol{\Xi}), \boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1}\boldsymbol{\nu}^t + \boldsymbol{W}_0\hat{\boldsymbol{\nu}}^t + \boldsymbol{Z}^t\right)\right], \tag{37}$$

$$\hat{\boldsymbol{\nu}}^{t+1} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} \partial_{1i}\bar{f}_i^t\left(\boldsymbol{Z}_{\boldsymbol{W}_0}, \boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1}\boldsymbol{\nu}^t + \boldsymbol{W}_0\hat{\boldsymbol{\nu}}^t + \boldsymbol{Z}^t, \boldsymbol{\Xi}\right)\right], \tag{38}$$

$$\begin{aligned}\boldsymbol{\kappa}^{t+1,s+1} = \boldsymbol{\kappa}^{s+1,t+1} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\Bigg[&\left(\tilde{f}^s\left(\varphi(\boldsymbol{Z}_{\boldsymbol{W}_0}, \boldsymbol{\Xi}), \boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1}\boldsymbol{\nu}^s + \boldsymbol{W}_0\hat{\boldsymbol{\nu}}^s + \boldsymbol{Z}^s\right) - \boldsymbol{W}_0\boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1}\boldsymbol{\nu}^{s+1}\right)^\top \\ &\left(\tilde{f}^t\left(\varphi(\boldsymbol{Z}_{\boldsymbol{W}_0}, \boldsymbol{\Xi}), \boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1}\boldsymbol{\nu}^t + \boldsymbol{W}_0\hat{\boldsymbol{\nu}}^t + \boldsymbol{Z}^t\right) - \boldsymbol{W}_0\boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1}\boldsymbol{\nu}^{t+1}\right)\Bigg],\end{aligned} \tag{39}$$

where $\boldsymbol{\rho}_{\boldsymbol{W}_0} = \lim_{N\to\infty} \frac{1}{N}\boldsymbol{W}_0^\top \boldsymbol{W}_0$, and for $i \in [n]$ we have $(\boldsymbol{Z}_{\boldsymbol{W}_0})_i \overset{i.i.d}{\sim} \mathsf{N}(\mathbf{0}, \boldsymbol{\rho}_{\boldsymbol{W}_0})$. For $i \in [n], 0 \le s, r \le t$, we have that $(\boldsymbol{Z}_{\boldsymbol{W}_0})_i$ is independent from $(\boldsymbol{Z}^s)_i \overset{i.i.d}{\sim} \mathsf{N}(\mathbf{0}, \boldsymbol{\kappa}^{s,s})$ with $\mathrm{Cov}((\boldsymbol{Z}^s)_i, (\boldsymbol{Z}^r)_i) = \boldsymbol{\kappa}^{s,r}$. In (38), we let $\bar{f}^t : (\boldsymbol{z}, \boldsymbol{u}, \boldsymbol{v}) \mapsto \tilde{f}^t(\varphi(\boldsymbol{z}, \boldsymbol{v}), \boldsymbol{u})$ and we let $\partial_{1i}\bar{f}_i^t$ denote the partial derivative of $\bar{f}_i^t$ with respect to the i-th row of its first argument.

We list the necessary assumptions for characterizing this AMP iteration, followed by the result:

Assumptions.

(B1) $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is a GOE(N) matrix, i.e., $\boldsymbol{A} = \mathbf{G} + \mathbf{G}^\top$ for $\mathbf{G} \in \mathbb{R}^{N\times N}$ with i.i.d. entries $G_{ij} \sim \mathsf{N}(0, 1/(2N))$.

(B2) For each $t \in \mathbb{N}_{>0}$, $\bar{f}^t : (\boldsymbol{z}, \boldsymbol{u}, \boldsymbol{\Xi}) \mapsto \tilde{f}^t(\varphi(\boldsymbol{z}, \boldsymbol{\Xi}), \boldsymbol{u})$ is uniformly pseudo-Lipschitz. For each $t \in \mathbb{N}_{>0}$ and for any $1 \leqslant i \leqslant N$, $(\boldsymbol{u}, \boldsymbol{z}) \mapsto \frac{\partial \tilde{f}_i^t}{\partial \boldsymbol{X}_i}(\varphi(\boldsymbol{z}, \boldsymbol{\Xi}), \boldsymbol{u})$ is uniformly pseudo-Lipschitz. The function $\tilde{f}^0 : \mathbb{R}^{N\times Q} \to \mathbb{R}^{N\times Q}$ is uniformly pseudo-Lipschitz.

(B3) The initialization $\boldsymbol{X}^0$ is deterministic, and $\|\boldsymbol{X}^0\|_F/\sqrt{N}$, $\|\boldsymbol{W}_0^\top \boldsymbol{W}_0\|_2/N$, $\|\boldsymbol{\Xi}\|_F/\sqrt{N}$ converge almost surely to finite constants as $N \to \infty$.

(B4) The following limits exist and are finite:

$$\lim_{N\to\infty} \frac{1}{N}\tilde{f}^0(\boldsymbol{X}^0)^\top \tilde{f}^0(\boldsymbol{X}^0), \quad \lim_{N\to\infty} \frac{1}{N}\boldsymbol{W}_0^\top \tilde{f}^0(\boldsymbol{X}^0).$$

(B5) For any $t \in \mathbb{N}_{>0}$ and any $\boldsymbol{\kappa} \in \mathcal{S}_Q^+$, the following limit exists and is finite:

$$\lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\tilde{f}^0(\boldsymbol{X}^0)^\top \tilde{f}^t(\varphi(\boldsymbol{Z},\boldsymbol{\Xi}),\boldsymbol{Z})\right]$$

where $\boldsymbol{Z} \in \mathbb{R}^{N\times Q}$, $\boldsymbol{Z} \sim \mathsf{N}(0, \boldsymbol{\kappa}\otimes \mathbf{I}_N)$.

(B6) For any $s,t \in \mathbb{N}_{>0}$ and any $\boldsymbol{\kappa} \in \mathcal{S}_{2Q}^+$, the following limit exists and is finite:

$$\lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\tilde{f}^s(\varphi(\boldsymbol{Z}^s,\boldsymbol{\Xi}),\boldsymbol{Z}^s)^\top \tilde{f}^t(\varphi(\boldsymbol{Z}^t,\boldsymbol{\Xi}),\boldsymbol{Z}^t)\right]$$

where $(\boldsymbol{Z}^s,\boldsymbol{Z}^t) \in (\mathbb{R}^{N\times Q})^2$,$(\boldsymbol{Z}^s,\boldsymbol{Z}^t) \sim \mathsf{N}(0, \boldsymbol{\kappa}\otimes \mathbf{I}_N)$.

Lemma A.1. *Consider the AMP iteration* (32)–(34) *and the state evolution recursion* (35)–(39). *Assume* (B1) – (B6). *Then for any sequence of functions* $\Phi_N : (\mathbb{R}^{N\times Q})^{\otimes(t+3)} \times \mathbb{R}^{N\times L_{\boldsymbol{\Xi}}} \to \mathbb{R}$ *such that* $(\boldsymbol{X}^0,\dots,\boldsymbol{X}^t,\boldsymbol{V}) \mapsto \Phi_N(\mathbf{X}^0,\mathbf{X}^1,...,\mathbf{X}^t,\boldsymbol{V},\boldsymbol{W}_0,\boldsymbol{\Xi})$ *is uniformly pseudo-Lipschitz, we have that:*

$$\begin{aligned}&\Phi_N\left(\mathbf{X}^0,\mathbf{X}^1,...,\mathbf{X}^t,\boldsymbol{A}\boldsymbol{W}_0,\boldsymbol{W}_0,\boldsymbol{\Xi}\right)\\ &\qquad \overset{\mathbb{P}}{\simeq} \mathbb{E}_{\boldsymbol{Z}^0,\dots,\boldsymbol{Z}^t,\boldsymbol{Z}_{\boldsymbol{W}_0}}\left[\Phi_N\left(\boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1}\boldsymbol{\nu}^0+\boldsymbol{W}_0\hat{\boldsymbol{\nu}}^0+\boldsymbol{Z}^0,\dots,\boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1}\boldsymbol{\nu}^t+\boldsymbol{W}_0\hat{\boldsymbol{\nu}}^t+\boldsymbol{Z}^t,\boldsymbol{Z}_{\boldsymbol{W}_0},\boldsymbol{W}_0,\boldsymbol{\Xi}\right)\right]. \qquad (40)\end{aligned}$$

Proof. The main differences between the AMP result in (Gerbelot & Berthier, 2023) and our characterization are: 1) the φ function is allowed to depend on an auxiliary random variable $\boldsymbol{\Xi}$, 2) the test function Φ_N is allowed to depend additionally on $\boldsymbol{A}\boldsymbol{W}_0$, $\boldsymbol{W}_0$, and $\boldsymbol{\Xi}$. Assumption (B2) guarantees that $\tilde{f}^t$ maintains the same required coverage properties despite modification 1).

We now address modification 2). Since $\boldsymbol{\Xi}$ and $\boldsymbol{W}_0$ are fixed and $\tilde{\Phi}_N : (\boldsymbol{X}^0,\dots,\boldsymbol{X}^t,\boldsymbol{V}) \mapsto \Phi_N(\mathbf{X}^0,\mathbf{X}^1,...,\mathbf{X}^t,\boldsymbol{V},\boldsymbol{W}_0,\boldsymbol{\Xi})$ is assumed to be uniformly pseudo-Lipschitz, these can be included in Φ_N and the right hand side of (40) is unaffected. Further, note that by Proposition 2 in (Gerbelot & Berthier, 2023) (a standard upper bound on the operator norm of $\boldsymbol{A} \sim GOE(N)$) we have that $\|\boldsymbol{A}\boldsymbol{W}_0\|_F/\sqrt{N} < \infty$ as $N \to \infty$. We can therefore include $\boldsymbol{A}\boldsymbol{W}_0$ in the definition of Φ_N as well, and the uniform pseudo-Lipschitzness of $\tilde{\Phi}_N$ guarantees convergence to the result. □

We next present the main reduction, mapping the AMP algorithm proposed in this work (3) to the symmetric one outlined in (32)–(34).

Proof of Theorem 3.1. Consider the Change Point Linear Regression model (1), and recall that it can be rewritten as (2). We reduce the algorithm (3) to (32)–(34), following the alternating technique of (Javanmard & Montanari, 2013). The idea is to define a symmetric GOE matrix with $\boldsymbol{X}$ and $\boldsymbol{X}^\top$ on the off-diagonals. With a suitable initialization, the iteration (32)–(34) then yields $\boldsymbol{B}^{t+1}$ in the even iterations and for $\boldsymbol{\Theta}^t$ in the odd iterations.

Let $N = n+p$. For a matrix $\boldsymbol{E} \in \mathbb{R}^{N\times L}$, we use $\boldsymbol{E}_{[n]}$ and $\boldsymbol{E}_{[-p]}$ to denote the first n rows and the last p rows of $\boldsymbol{E}$ respectively. Recall from (3) that $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{B},\boldsymbol{B}^t \in \mathbb{R}^{p\times L}$ and $\boldsymbol{\Theta} = \boldsymbol{X}\boldsymbol{B}$, $\boldsymbol{\Theta}^t \in \mathbb{R}^{n\times L}$, and $n/p \to \delta$ as $n,p \to \infty$. We let

$$\boldsymbol{A} = \sqrt{\frac{\delta}{\delta+1}}\begin{bmatrix}\boldsymbol{D}_1 & \boldsymbol{X}\\ \boldsymbol{X}^\top & \boldsymbol{D}_2\end{bmatrix} \sim \mathrm{GOE}(N), \quad \boldsymbol{W}_0 = \begin{bmatrix}\mathbf{0}_{n\times L}\\ \boldsymbol{B}\end{bmatrix} \in \mathbb{R}^{N\times L}, \quad \boldsymbol{X}^0 = \begin{bmatrix}\mathbf{0}_{n\times L}\\ \boldsymbol{B}^0\end{bmatrix} \qquad (41)$$

where $\boldsymbol{D}_1 \sim \mathrm{GOE}(n)$ and $\sqrt{\delta}\boldsymbol{D}_2 \sim \mathrm{GOE}(p)$ are independent of each other and of $\boldsymbol{X}$. Let $\boldsymbol{\Xi} := \begin{bmatrix}\boldsymbol{\Psi} & \boldsymbol{\varepsilon}\end{bmatrix} \in \mathbb{R}^{n\times 2}$ and define

$$\varphi : (\boldsymbol{W},\boldsymbol{\Xi}) \mapsto q\left(\sqrt{\frac{\delta+1}{\delta}}\boldsymbol{W}_{[n]},\boldsymbol{\Psi},\boldsymbol{\varepsilon}\right) \in \mathbb{R}^N, \qquad (42)$$

for any $\boldsymbol{W} \in \mathbb{R}^{N\times L}$. We therefore have $\varphi(\boldsymbol{A}\boldsymbol{W}_0,\boldsymbol{\Xi}) = q\left(\boldsymbol{X}\boldsymbol{B},\boldsymbol{\Psi},\boldsymbol{\varepsilon}\right)$. Let $\tilde{f}^t : \mathbb{R}^{N\times L} \times \mathbb{R}^N \to \mathbb{R}^{N\times L}$ such that

$$\tilde{f}^{2t+1}(\varphi(\boldsymbol{A}\boldsymbol{W}_0,\boldsymbol{\Xi}),\boldsymbol{U}) = \sqrt{\frac{\delta+1}{\delta}}\begin{bmatrix}g^t(\boldsymbol{U}_{[n]},q(\boldsymbol{X}\boldsymbol{B},\boldsymbol{\Psi},\boldsymbol{\varepsilon}))\\ \mathbf{0}_{p\times L}\end{bmatrix} \quad \text{and} \quad \tilde{f}^{2t}(\varphi(\boldsymbol{A}\boldsymbol{W}_0,\boldsymbol{\Xi}),\boldsymbol{U}) = \sqrt{\frac{\delta+1}{\delta}}\begin{bmatrix}\mathbf{0}_{n\times L}\\ f^t(\boldsymbol{U}_{[-p]})\end{bmatrix}, \qquad (43)$$

for any $\boldsymbol{U}$ in $\mathbb{R}^{N\times L}$.

Next, consider the AMP iteration (32)–(34) with $\boldsymbol{A}, \boldsymbol{W}_0, \boldsymbol{X}^0, \boldsymbol{\Xi}, \varphi, \tilde{f}^t$ defined as in (41)–(43). Note that the assumptions $(B1)-(B6)$ are satisfied by construction, and hence the state evolution result in Lemma A.1 holds for the iteration (32)–(34). We will now show that the state evolution equations decompose into those of interest, (7)–(10). First, note that

$$\boldsymbol{X}^{2t+1} = \begin{bmatrix} \boldsymbol{X} f^t(\boldsymbol{X}^{2t}_{[-p]}) - g^{t-1}(\boldsymbol{X}^{2t-1}_{[n]}, q(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{\Xi})) \cdot \overbrace{\frac{1}{n}\left(\sum_{i=1}^{p} \frac{\partial f^t(\boldsymbol{X}^{2t}_{[-p]})_i}{\partial \boldsymbol{X}^{2t}_{[-p],i}}\right)^\top}^{=(\boldsymbol{F}^t)^\top} \\ \boldsymbol{D}_2 f^t(\boldsymbol{X}^{2t}_{[-p]}) \end{bmatrix}, \tag{44}$$

$$\boldsymbol{X}^{2t} = \begin{bmatrix} \boldsymbol{D}_1 g^{t-1}(\boldsymbol{X}^{2t-1}_{[n]}, q(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{\Xi})) \\ \boldsymbol{X}^\top g^{t-1}(\boldsymbol{X}^{2t-1}_{[n]}, q(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{\Xi})) - f^{t-1}(\boldsymbol{X}^{2t-2}_{[-p]}) \cdot \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial g^{t-1}(\boldsymbol{X}^{2t-1}_{[n]}, q(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{\Xi}))}{\partial \boldsymbol{X}^{2t-1}_{[n],i}}\right)^\top}_{=(\boldsymbol{C}^{t-1})^\top} \end{bmatrix}, \tag{45}$$

and hence observe that $\boldsymbol{X}^{2t+1}_{[n]}$ and $\boldsymbol{X}^{2t}_{[-p]}$ are equal to $\boldsymbol{\Theta}^t$ and $\boldsymbol{B}^t$ in (3) respectively. Define the main iterate $\boldsymbol{Q}^t = \boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}^{-1}_{\boldsymbol{W}_0}\boldsymbol{\nu}^t + \boldsymbol{W}_0\hat{\boldsymbol{\nu}}^t + \boldsymbol{Z}^t \in \mathbb{R}^{N\times L}$. For $i \in [n]$, let $\tilde{g}^t_i : (\boldsymbol{Z}, \boldsymbol{V}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) \mapsto g^t_i(\boldsymbol{V}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))$ and let $\partial_{1i}\tilde{g}^t_i$ be the partial derivative (Jacobian) w.r.t. the ith row of the first argument. Following (35)–(39), we then have that

$$\boldsymbol{\nu}^{2t+1} = \lim_{N\to\infty} \frac{1}{N}\sqrt{\frac{\delta+1}{\delta}}\mathbb{E}\left[\begin{bmatrix}\boldsymbol{0}_{L\times n} & \boldsymbol{B}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{0}_{n\times L} \\ f^t(\boldsymbol{Q}^{2t}_{[-p]})\end{bmatrix}\right] = \sqrt{\frac{\delta}{\delta+1}} \lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\boldsymbol{B}^\top f^t(\boldsymbol{Q}^{2t}_{[-p]})\right], \tag{46}$$

$$\boldsymbol{\nu}^{2t} = \sqrt{\frac{\delta+1}{\delta}} \lim_{N\to\infty}\frac{1}{N}\mathbb{E}\left[\begin{bmatrix}\boldsymbol{0}_{L\times n} & \boldsymbol{B}^\top\end{bmatrix}\begin{bmatrix} g^{t-1}(\boldsymbol{Q}^{2t-1}_{[n]}, \boldsymbol{Y}) \\ \boldsymbol{0}_{p\times L}\end{bmatrix}\right] = \boldsymbol{0}_{L\times L}, \tag{47}$$

$$\hat{\boldsymbol{\nu}}^{2t+1} = \boldsymbol{0}_{L\times L} \tag{48}$$

$$\hat{\boldsymbol{\nu}}^{2t} = \lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\partial_{1i}\tilde{g}^{t-1}_i\left(\sqrt{\frac{\delta+1}{\delta}}\boldsymbol{Z}_{\boldsymbol{W}_0,[n]}, \boldsymbol{Q}^{2t-1}_{[n]}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}\right)\right], \tag{49}$$

$$\boldsymbol{\kappa}^{2t,2s} = \lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[g^{t-1}\left(\boldsymbol{Q}^{2t-1}_{[n]}, q\left(\sqrt{\frac{\delta+1}{\delta}}\boldsymbol{Z}_{\boldsymbol{W}_0,[n]}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}\right)\right)^\top g^{s-1}\left(\boldsymbol{Q}^{2s-1}_{[n]}, q\left(\sqrt{\frac{\delta+1}{\delta}}\boldsymbol{Z}_{\boldsymbol{W}_0,[n]}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}\right)\right)\right] \tag{50}$$

$$\boldsymbol{\kappa}^{2t+1,2s+1} = \lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\left(f^t(\boldsymbol{Q}^{2t}_{[-p]}) - \sqrt{\frac{\delta+1}{\delta}}\boldsymbol{B}\boldsymbol{\rho}^{-1}_{\boldsymbol{B}}\boldsymbol{\nu}^{2t+1}\right)^\top\left(f^s(\boldsymbol{Q}^{2s}_{[-p]}) - \sqrt{\frac{\delta+1}{\delta}}\boldsymbol{B}\boldsymbol{\rho}^{-1}_{\boldsymbol{B}}\boldsymbol{\nu}^{2s+1}\right)\right], \tag{51}$$

where in (51) we used $\boldsymbol{\rho}_{\boldsymbol{W}_0} = \frac{1}{N}\boldsymbol{B}^\top\boldsymbol{B} = \frac{\delta}{\delta+1}\boldsymbol{\rho}_{\boldsymbol{B}}$. Hence, we can associate

$$\boldsymbol{\nu}^t_{\boldsymbol{\Theta}} = \sqrt{\frac{\delta+1}{\delta}}\boldsymbol{\nu}^{2t+1}, \tag{52}$$

$$\boldsymbol{\nu}^t_{\boldsymbol{B}} = \hat{\boldsymbol{\nu}}^{2t}, \tag{53}$$

$$\boldsymbol{\kappa}^{t,s}_{\boldsymbol{\Theta}} = \boldsymbol{\kappa}^{2t+1,2s+1}, \tag{54}$$

$$\boldsymbol{\kappa}^{t,s}_{\boldsymbol{B}} = \boldsymbol{\kappa}^{2t,2s}. \tag{55}$$

$$\boldsymbol{Z}_{\boldsymbol{B}} \overset{d}{=} \sqrt{\frac{\delta+1}{\delta}}\boldsymbol{Z}_{\boldsymbol{W}_0,[n]}, \tag{56}$$

$$\boldsymbol{G}_{\boldsymbol{\Theta}}^t \stackrel{d}{=} \boldsymbol{Z}_{[n]}^{2t+1}, \tag{57}$$

$$\boldsymbol{G}_{\boldsymbol{B}}^t \stackrel{d}{=} \boldsymbol{Z}_{[-p]}^{2t}, \tag{58}$$

$$\boldsymbol{V}_{\boldsymbol{\Theta}}^t \stackrel{d}{=} \boldsymbol{Q}_{[n]}^{2t+1}, \tag{59}$$

$$\boldsymbol{V}_{\boldsymbol{B}}^t \stackrel{d}{=} \boldsymbol{Q}_{[-p]}^{2t}. \tag{60}$$

Substituting the change of variables (52)–(60) into (46)–(51), we obtain (7)–(10). Moreover, substituting (44)–(45) and (59)–(60) into Lemma A.1 yields Theorem 3.1. □

B. State Evolution Limits and Simplifications

B.1. State Evolution Dependence on $\boldsymbol{B}, \boldsymbol{\varepsilon}$

In this section, we outline how the limits in (7)–(10) can be checked, and how the dependence of (7)–(10) on $\boldsymbol{B}, \boldsymbol{\varepsilon}$ can be removed. We give a set of sufficient conditions for the existence of the limits in (7)–(10), which allow for removing the dependency of the state evolution (7)–(10) on $\boldsymbol{B}, \boldsymbol{\varepsilon}$. Assume (S0) on p.5, and also that the following assumptions hold:

(S1) As $n \to \infty$, the entries of the normalized change point vector $\boldsymbol{\eta}/n$ converge to constants $\alpha_0, \ldots, \alpha_L$ such that $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_{L^*} = \ldots = \alpha_L = 1$.

(S2) For $t \geq 0$, f^t is separable, and g^t acts row-wise on its input. (We recall that a separable function acts row-wise *and* identically on each row.)

(S3) For $\ell \in [L^* - 1]$, the empirical distributions of $\{\tilde{g}_i^t(\boldsymbol{Z}_1, (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, \Psi_{\eta_\ell}, \varepsilon_i)\}_{i \in [\eta_\ell, \eta_{\ell+1})}$ and $\{\partial_1 \tilde{g}_i^t(\boldsymbol{Z}_1, (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, \Psi_{\eta_\ell}, \varepsilon_i)\}_{i \in [\eta_\ell, \eta_{\ell+1})}$ converge weakly to the laws of random variables $\hat{g}_{\eta_\ell}^t(\boldsymbol{Z}_1, (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, \Psi_{\eta_\ell}, , \bar{\varepsilon}))$ and $\check{g}_{\eta_\ell}^t(\boldsymbol{Z}_1, (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, \Psi_{\eta_\ell}, \bar{\varepsilon})$, respectively, where ∂_1 denotes Jacobian with respect to the first argument.

The assumption (S1) is natural in the regime where the number of samples n is proportional to p, and the number of degrees of freedom in the signals also grows linearly in p. Without change points, f^t, g^t can both be assumed separable without loss of optimality (due to (S0)). To handle the temporal dependence created by change points, we require g_i^t to depend on i, for $i \in [n]$. However, Proposition 3.2 shows that it can be chosen to act row-wise, i.e., $g_i^{*t}(\boldsymbol{\Theta}^t, \boldsymbol{y}) = g_i^{*t}(\boldsymbol{\Theta}_i^t, y_i)$. This justifies (S2). When g_i^t is chosen to be g_i^{*t} for $i \in [n]$, the condition (S3) can be translated into regularity conditions on the prior marginals $\pi_{\bar{\Psi}_i}$ and distributional convergence conditions on the noise ε_i, as these are the only quantities that differ along the elements of the sets $\{\tilde{g}_i^t(\boldsymbol{Z}_1, (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, \Psi_{\eta_\ell}, \varepsilon_i)\}_{i \in [\eta_\ell, \eta_{\ell+1})}$ and $\{\partial_1 \tilde{g}_i^t(\boldsymbol{Z}_1, (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, \Psi_{\eta_\ell}, \varepsilon_i)\}_{i \in [\eta_\ell, \eta_{\ell+1})}$ for $\ell \in [L^* - 1]$.

Under assumptions (S0)-(S3), the state evolution equations in (7)–(10) reduce to:

$$\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1} = \sum_{\ell=0}^{L-1} \mathbb{E}\left[\check{g}_{\eta_\ell}^t(\boldsymbol{Z}_1, (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, \Psi_{\eta_\ell}, \bar{\varepsilon})\right], \tag{61}$$

$$\boldsymbol{\kappa}_{\boldsymbol{B}}^{s+1,t+1} = \sum_{\ell=0}^{L-1} \mathbb{E}\left[\hat{g}_{\eta_\ell}^s((\boldsymbol{V}_{\boldsymbol{\Theta}}^s)_1, q_1(\boldsymbol{Z}_1, \Psi_{\eta_\ell}, \bar{\varepsilon}))\, \hat{g}_{\eta_\ell}^t((\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, q_1(\boldsymbol{Z}_1, \Psi_{\eta_\ell}, \bar{\varepsilon}))^\top\right], \tag{62}$$

$$\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1} = \frac{1}{\delta}\mathbb{E}\left[\bar{\boldsymbol{B}} f_1^{t+1}((\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1})^\top \bar{\boldsymbol{B}} + (\boldsymbol{G}_{\boldsymbol{B}}^{t+1})_1)^\top\right], \tag{63}$$

$$\boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{s+1,t+1} = \frac{1}{\delta}\mathbb{E}\Big[\big(f_1^{s+1}((\boldsymbol{\nu}_{\boldsymbol{B}}^{s+1})^\top \bar{\boldsymbol{B}} + (\boldsymbol{G}_{\boldsymbol{B}}^{s+1})_1) - (\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{s+1})^\top \boldsymbol{\rho}^{-1} \bar{\boldsymbol{B}}\big) \big(f_1^{t+1}((\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1})^\top \bar{\boldsymbol{B}} + (\boldsymbol{G}_{\boldsymbol{B}}^{t+1})_1) - (\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1})^\top \boldsymbol{\rho}^{-1} \bar{\boldsymbol{B}}\big)^\top\Big], \tag{64}$$

where (S2) has allowed us to reduce the matrix products into sums of vector outer-products, and assumptions (S0)–(S3) have removed the dependence on $\boldsymbol{B}$ and $\boldsymbol{\varepsilon}$ due to the law of large numbers argument in Lemma 4 of (Bayati & Montanari, 2011).

B.2. Removing dependencies on $\boldsymbol{\varepsilon}$ on the RHS of (28)–(29)

Under assumptions (S0)–(S3) above, the dependence of $\boldsymbol{V}_{\boldsymbol{\Theta}}^t$ on the RHS of (28)–(29) on $\boldsymbol{B}, \boldsymbol{\varepsilon}$ can be removed. Moreover, the dependence of the RHS on $\boldsymbol{\varepsilon}$ through $\hat{\boldsymbol{\eta}}$ can be removed on a case-by-case basis. We expect this to hold, for example,

under (S0)–(S3) for estimators whose dependence on $\boldsymbol{\varepsilon}$ is via a normalized inner product such as (27):

$$U(\hat{\boldsymbol{\eta}}(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, U(\boldsymbol{\eta}), \boldsymbol{\varepsilon}))) = \arg\max_{\psi \in \mathcal{X}} \frac{\langle (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_{[:,\psi]}, \boldsymbol{Z}_{[:,\psi]} \rangle}{n} + \frac{\langle (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_{[:,\psi]}, \boldsymbol{\varepsilon} \rangle}{n}. \tag{65}$$

Indeed, the numerical experiments in Section 4 for chosen estimators demonstrate a strong agreement between the left-hand and right-hand sides of (28) and of (31), when $\boldsymbol{\varepsilon}$ on the right-hand side is substituted by an independent copy with the same limiting moments.

C. Proof and Computation of Optimal Denoisers f^t, g^t

C.1. Proof of Proposition 3.2

It is similar to the derivation of the optimal denoisers for mixed regression in (Tan & Venkataramanan, 2023), with a few differences in the derivation of g^{*t}. We include both parts here for completeness. This proof relies on Lemma C.1 and Lemma C.2, Cauchy-Schwarz inequality and Stein's Lemma extended to vector or matrix random variables. Recall that we treat vectors, including rows of matrices, as column vectors, and that functions $h : \mathbb{R}^q \to \mathbb{R}^q$ have column vectors as input and output.

Lemma C.1 (Extended Cauchy-Schwarz inequality, Lemma 2 in (Lavergne, 2008))**.** *Let $\boldsymbol{A} \in \mathbb{R}^{n \times L}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times L}$ be random matrices such that $\mathbb{E}\|\boldsymbol{A}\|_F^2 < \infty, \mathbb{E}\|\boldsymbol{B}\|_F^2 < \infty$, and $\mathbb{E}[\boldsymbol{A}^\top \boldsymbol{A}]$ is non-singular, then*

$$\mathbb{E}[\boldsymbol{B}^\top \boldsymbol{B}] - \mathbb{E}[\boldsymbol{B}^\top \boldsymbol{A}](\mathbb{E}[\boldsymbol{A}^\top \boldsymbol{A}])^{-1}\mathbb{E}[\boldsymbol{A}^\top \boldsymbol{B}] \succcurlyeq 0. \tag{66}$$

Lemma C.2 (Extended Stein's Lemma)**.** *Let $\boldsymbol{x} \in \mathbb{R}^L$ and $h : \mathbb{R}^L \to \mathbb{R}^L$ be such that for $\ell \in [L]$, the function $h_\ell : x_\ell \to [h(\boldsymbol{x})]_\ell$ is absolutely continuous for Lebesgue almost every $(x_i : i \neq \ell) \in \mathbb{R}^{L-1}$, with weak derivative $\partial h_\ell(\boldsymbol{x})/\partial x_\ell : \mathbb{R}^L \to \mathbb{R}$ satisfying $\mathbb{E}|\partial h_\ell(\boldsymbol{x})/\partial x_\ell| < \infty$. If $\boldsymbol{x} \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \in \mathbb{R}^L$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{L \times L}$ positive definite, then*

$$\mathbb{E}\left[(\boldsymbol{x} - \boldsymbol{\mu})h(\boldsymbol{x})^\top\right] = \boldsymbol{\Sigma}\,\mathbb{E}[h'(\boldsymbol{x})]^\top, \tag{67}$$

where $h'(\boldsymbol{x})$ is the Jacobian matrix of h.

The proof of Lemma C.2 follows from Lemma 6.20 of (Feng et al., 2022).

Proof of part 1 (optimal f^t). Using the law of total expectation and applying (25), we can rewrite $\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t$ in (19) as:

$$\begin{aligned}\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t &= \frac{1}{\delta}\lim_{p\to\infty}\frac{1}{p}\sum_{j=1}^p \mathbb{E}\left[\bar{\boldsymbol{B}}\left(f_j^t(\bar{\boldsymbol{V}}_{\boldsymbol{B}}^t)\right)^\top\right] = \frac{1}{\delta}\mathbb{E}\left[\bar{\boldsymbol{B}}\left(f_j^t(\bar{\boldsymbol{V}}_{\boldsymbol{B}}^t)\right)^\top\right] = \frac{1}{\delta}\mathbb{E}\left[\mathbb{E}\left[\bar{\boldsymbol{B}}\left(f_j^t(\bar{\boldsymbol{V}}_{\boldsymbol{B}}^t)\right)^\top \middle| \bar{\boldsymbol{V}}_{\boldsymbol{B}}^t\right]\right] \\ &= \frac{1}{\delta}\mathbb{E}\left[\mathbb{E}\left[\bar{\boldsymbol{B}}|\bar{\boldsymbol{V}}_{\boldsymbol{B}}^t\right]\left(f_j^t(\bar{\boldsymbol{V}}_{\boldsymbol{B}}^t)\right)^\top\right] = \frac{1}{\delta}\mathbb{E}\left[f_j^{*t}(f_j^t)^\top\right],\end{aligned} \tag{68}$$

where we used the shorthand $f_j^t \equiv f_j^t(\bar{\boldsymbol{V}}_{\boldsymbol{B}}^t)$ and $f_j^{*t} \equiv f_j^{*t}(\bar{\boldsymbol{V}}_{\boldsymbol{B}}^t) = \mathbb{E}[\bar{\boldsymbol{B}}|\bar{\boldsymbol{V}}_{\boldsymbol{B}}^t]$. Applying Lemma C.1 yields

$$\mathbb{E}\left[f_j^{*t}(f_j^{*t})^\top\right] - \mathbb{E}\left[f_j^{*t}(f_j^t)^\top\right]\mathbb{E}\left[f_j^t(f_j^t)^\top\right]^{-1}\mathbb{E}\left[f_j^t(f_j^{*t})^\top\right] \succcurlyeq 0. \tag{69}$$

Since (20) can be simplified into

$$\bar{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{t,t} = \frac{1}{\delta}\mathbb{E}\left[f_j^t(f_j^t)^\top\right] - (\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^\top \boldsymbol{\rho}^{-1}\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t, \tag{70}$$

using (68) and (70) in (69), we obtain that

$$\boldsymbol{\Delta} := \frac{1}{\delta}\mathbb{E}\left[f_j^{*t}(f_j^{*t})^\top\right] - \bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t\left[\bar{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{t,t} + (\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^\top\boldsymbol{\rho}^{-1}(\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t)\right]^{-1}(\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^\top \succcurlyeq 0.$$

Adding and subtracting $\bar{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{t,t}$ on the LHS gives $\bar{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{t,t} - \bar{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{t,t} + \boldsymbol{\Delta} \succcurlyeq 0$. Left multiplying by $\boldsymbol{\rho}^\top\left((\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^{-1}\right)^\top$ and right multiplying by $(\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^{-1}\boldsymbol{\rho}$ maintains the positive semi-definiteness of the LHS and further yields

$$\boldsymbol{\rho}^\top\left((\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^{-1}\right)^\top\bar{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{t,t}(\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^{-1}\boldsymbol{\rho} - \underbrace{\boldsymbol{\rho}^\top\left((\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^{-1}\right)^\top(\bar{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{t,t} - \boldsymbol{\Delta})(\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^{-1}\boldsymbol{\rho}}_{=:\boldsymbol{\Gamma}_{\boldsymbol{\Theta}}^t} \succcurlyeq 0, \tag{71}$$

which implies

$$\operatorname{Tr}\left(\boldsymbol{\rho}^\top \left((\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})^{-1}\right)^\top \bar{\boldsymbol{\kappa}}^{t,t}_{\boldsymbol{\Theta}} (\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})^{-1}\boldsymbol{\rho}\right) \geq \operatorname{Tr}\left(\boldsymbol{\Gamma}^t_{\boldsymbol{\Theta}}\right). \tag{72}$$

Recall from (23) that the LHS of (72) is the objective we wish to minimise via optimal f^t. Indeed, by setting $f^t = f^{*t}$, (72) is satisfied with equality, which proves part 1 of Proposition 3.2.

Remark C.3. When $f^t = f^{*t}$, (68) and (70) reduce to $\bar{\boldsymbol{\kappa}}^{t,t}_{\boldsymbol{\Theta}} = \bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}} - (\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})^\top \boldsymbol{\rho}^{-1}\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}}$.

Proof of part 2 (optimal g^t) Recall from (17) that $\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\partial_{1i}\tilde{g}^t_i(\boldsymbol{Z}_1, (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, \bar{\Psi}_i, \bar{\varepsilon})\right]$, and we can rewrite the transpose of each summand as:

$$\begin{aligned}
&\mathbb{E}\left[\partial_{1i}\tilde{g}^t_i(\boldsymbol{Z}_1, (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, \bar{\Psi}_i, \bar{\varepsilon})\right]^\top && (73)\\
&\overset{(a)}{=} \mathbb{E}_{(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1}\left\{\mathbb{E}_{\boldsymbol{Z}_1,\bar{\varepsilon}}\left[\partial_{1i}\tilde{g}^t_i\left(\boldsymbol{Z}_1, (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, \bar{\Psi}_i, \bar{\varepsilon}\right)\middle|(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1\right]\right\}^\top \\
&\overset{(b)}{=} \mathbb{E}_{(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1}\left\{\operatorname{Cov}\left(\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1\right)^{-1}\mathbb{E}_{\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon}}\left[\left(\boldsymbol{Z}_1 - \mathbb{E}[\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1]\right)\tilde{g}^t_i\left(\boldsymbol{Z}_1, (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, \bar{\Psi}_i, \bar{\varepsilon}\right)^\top\middle|(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1\right]\right\} \\
&\overset{(c)}{=} \mathbb{E}_{(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1}\left\{\operatorname{Cov}\left(\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1\right)^{-1}\right. && (74)\\
&\qquad\left.\mathbb{E}_{q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})}\left[\mathbb{E}_{\boldsymbol{Z}_1}\left[\left(\boldsymbol{Z}_1 - \mathbb{E}[\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1]\right)g^t_i\left((\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right)^\top\middle|(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right]\right]\right\} \\
&= \mathbb{E}_{(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1}\left\{\operatorname{Cov}\left(\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1\right)^{-1}\mathbb{E}_{q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})}\left[\left(\mathbb{E}[\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})] - \mathbb{E}[\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1]\right)g^t_i\left((\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right)^\top\right]\right\} \\
&\overset{(d)}{=} \mathbb{E}\left\{\underbrace{\operatorname{Cov}\left(\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1\right)^{-1}\left(\mathbb{E}\left[\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right] - \mathbb{E}[\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1]\right)}_{=g^{*t}_i\left((\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right)}g^t_i\left((\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right)^\top\right\} \\
&= \mathbb{E}\left[g^{*t}_i(g^t_i)^\top\right], && (75)
\end{aligned}$$

where (a) and (c) follow from the law of total expectation; (b) uses Lemma C.2; and (d) uses (26). In the last line of (75) we have used the shorthand $g^t_i \equiv g^t_i\left((\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right)$ and $g^{*t}_i \equiv g^{*t}_i\left((\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right)$. Substituting (75) in (17) yields

$$\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \mathbb{E}[g^t_i(g^{*t}_i)^\top] = \lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}(g^t)^\top g^{*t}\right]. \tag{76}$$

Note that since $\pi_{\bar{\Psi}_i}$ differs across i, g^t_i differs across i and the sum in (76) cannot be reduced. Lemma C.1 implies that

$$\mathbb{E}\left[\frac{1}{n}(g^{*t})^\top g^{*t}\right] - \mathbb{E}\left[\frac{1}{n}(g^{*t})^\top(g^t)\right]\mathbb{E}\left[\frac{1}{n}(g^t)^\top g^t\right]^{-1}\mathbb{E}\left[\frac{1}{n}(g^t)^\top g^{*t}\right] \succcurlyeq 0. \tag{77}$$

Recalling from (17) and (18) that the limits in the state evolution iterates $\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}} = \lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}(g^t)^\top g^{*t}\right]$ and $\bar{\boldsymbol{\kappa}}^{t+1,t+1}_{\boldsymbol{B}} = \lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}(g^t)^\top g^t\right]$ exist, we can take limits to obtain

$$\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}(g^{*t})^\top g^{*t}\right] - \underbrace{\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}\left(g^{*t}\right)^\top g^t\right]}_{\left(\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}}\right)^\top}\underbrace{\left(\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}(g^t)^\top g^t\right]\right)^{-1}}_{\left(\bar{\boldsymbol{\kappa}}^{t+1,t+1}_{\boldsymbol{B}}\right)^{-1}}\underbrace{\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}(g^t)^\top g^{*t}\right]}_{\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}}} \succcurlyeq 0. \tag{78}$$

Left multiplying $\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}}\left((\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}})^{-1}\right)^\top$ and right multiplying $(\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}})^{-1}\left(\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}}\right)^\top$ on the LHS maintains the positive semi-definiteness of the LHS to give

$$\underbrace{\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}}\left((\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}})^{-1}\right)^\top\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}(g^{*t})^\top g^{*t}\right](\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}})^{-1}\left(\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}}\right)^\top}_{=:\boldsymbol{\Gamma}^{t+1}_{\boldsymbol{B}}} - \bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}}\left(\bar{\boldsymbol{\kappa}}^{t+1,t+1}_{\boldsymbol{B}}\right)^{-1}\left(\bar{\boldsymbol{\nu}}^{t+1}_{\boldsymbol{B}}\right)^\top \succcurlyeq 0, \tag{79}$$

Moreover, since for positive definite matrices $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$, $\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2 \succcurlyeq 0$ implies $\boldsymbol{\Gamma}_1^{-1} - \boldsymbol{\Gamma}_2^{-1} \preccurlyeq 0$, we have that

$$[(\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^{-1}]^\top \bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1} (\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^{-1} - \boldsymbol{\Gamma}_{\boldsymbol{B}}^{t+1} \succcurlyeq 0, \tag{80}$$

which implies

$$\mathrm{Tr}\left([(\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^{-1}]^\top \bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1} (\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^{-1}\right) \geq \mathrm{Tr}\left(\boldsymbol{\Gamma}_{\boldsymbol{B}}^{t+1}\right).$$

Recall from (24) that g^t is optimised by minimising $\mathrm{Tr}\left([(\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^{-1}]^\top \bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1} (\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^{-1}\right)$. Indeed, by choosing $g^t = g^{*t}$, the objective achieves its lower bound, which completes the proof.

Remark C.4. Note setting $g^t = g^{*t}$ leads to $\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1} = \bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1}$.

C.2. Computation of f^{*t} and g^{*t}

The computation of the optimal denoisers relies on Lemmas C.5 and C.6, which state the conditioning properties of multivariate Gaussians or mixtures of Gaussians. We compute the Jacobians of these denoisers in (3) using Automatic Differentiation in Python JAX (Bradbury et al., 2018).

Lemma C.5 (Conditioning property of multivariate Gaussian)**.** *Suppose $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^m$ are jointly Gaussian: $\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} \sim \mathsf{N}\left(\begin{bmatrix}\boldsymbol{\mu}_{\boldsymbol{x}}\\ \boldsymbol{\mu}_{\boldsymbol{y}}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{\Sigma}_{\boldsymbol{xy}}\\ \boldsymbol{\Sigma}_{\boldsymbol{xy}}^\top & \boldsymbol{\Sigma}_{\boldsymbol{y}}\end{bmatrix}\right)$, then*

$$\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y} = \tilde{\boldsymbol{y}}] = \boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{xy}}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{\mu}_{\boldsymbol{y}}), \quad \mathrm{Cov}(\boldsymbol{x} \mid \boldsymbol{y}) = \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{xy}}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{xy}}^\top.$$

Lemma C.6 (Conditioning property of Gaussian mixtures)**.** *Let $\boldsymbol{x}_a \in \mathbb{R}^{L_a}, \boldsymbol{x}_b \in \mathbb{R}^{L_b}$ with $\begin{bmatrix}\boldsymbol{x}_a\\ \boldsymbol{x}_b\end{bmatrix} \sim \sum_{k=1}^K \pi_k \mathsf{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, then*

$$\mathbb{E}[\boldsymbol{x}_a \mid \boldsymbol{x}_b = \boldsymbol{x}] = \frac{\sum_{k=1}^K \pi_k \mathbb{P}_k(\boldsymbol{x}_b = \boldsymbol{x})\mathbb{E}_k[\boldsymbol{x}_a \mid \boldsymbol{x}_b = \boldsymbol{x}]}{\sum_{k'=1}^K \pi_{k'} \mathbb{P}_{k'}(\boldsymbol{x}_b = \boldsymbol{x})} \tag{81}$$

where $\mathbb{E}_k[\boldsymbol{x}_a \mid \boldsymbol{x}_b]$ is the conditional expectation given $\begin{bmatrix}\boldsymbol{x}_a\\ \boldsymbol{x}_b\end{bmatrix} \sim \mathsf{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Computation of f^{*t} Recalling (22), f_j^{*t} in (25) can be computed by expanding the conditional expectation:

$$f_j^{*t}(\boldsymbol{U}) = f_j^{*t}(\boldsymbol{U}_j) := \mathbb{E}[\bar{\boldsymbol{B}} | \bar{\boldsymbol{V}}_{\boldsymbol{B}}^t = \boldsymbol{U}_j],$$

which permits easily computable expressions for sufficiently regular distributions $\mathbb{P}_{\bar{\boldsymbol{B}}}$ such as discrete, Bernoulli-Gaussian or Gaussian distributions. We give the formulas for these below.

Computation of f^{*t} for Gaussian signals When $\bar{\boldsymbol{B}} \sim \mathsf{N}_L(\mathbb{E}[\bar{\boldsymbol{B}}], \delta\boldsymbol{\rho})$, we have

$$\begin{bmatrix}\bar{\boldsymbol{B}}\\ \bar{\boldsymbol{V}}_{\boldsymbol{B}}^t\end{bmatrix} \sim \mathsf{N}_{2L}\left(\boldsymbol{\mu}_{\boldsymbol{B}}, \boldsymbol{\Sigma}_{\boldsymbol{B}}^t\right), \quad \text{where} \quad \boldsymbol{\mu}_{\boldsymbol{B}}^t := \begin{bmatrix}\boldsymbol{I}_{L\times L}\\ (\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^t)^\top\end{bmatrix}\mathbb{E}[\bar{\boldsymbol{B}}], \ \boldsymbol{\Sigma}_{\boldsymbol{B}}^t := \begin{bmatrix}\boldsymbol{\rho} & \boldsymbol{\rho}\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^t\\ (\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^t)^\top\boldsymbol{\rho} & (\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^t)^\top\boldsymbol{\rho}\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^t + \bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t,t}/\delta\end{bmatrix},$$

and applying Lemma C.5 yields:

$$f_j^t(\boldsymbol{U}_j) = \mathbb{E}[\bar{\boldsymbol{B}}] + \boldsymbol{\rho}\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^t\left((\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^t)^\top\boldsymbol{\rho}\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^t + \bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t,t}/\delta\right)^{-1}(\boldsymbol{U}_j - (\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^t)^\top\mathbb{E}[\bar{\boldsymbol{B}}]).$$

Computation of f^{*t} for Bernoulli-Gaussian prior This prior is used in the experiments in Figures 4, 6 and 7. It assumes that $\bar{B}_\ell \in \{\bar{B}_\ell\}_{\ell\in[L]}$ is distributed as follows:

$$\bar{B}_\ell \overset{\text{i.i.d}}{\sim} \zeta(\alpha; \sigma_\ell) := (1-\alpha)\delta_0 + \alpha\mathsf{N}\left(0, \sigma_\ell^2\right),$$

with $\sigma_\ell^2 = \frac{\delta}{\alpha}\rho_{\ell,\ell}$. Recall that for $t \geq 0$, $\bar{\boldsymbol{V}}_{\boldsymbol{B}}^{t+1} := (\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1})^\top \bar{\boldsymbol{B}} + (\bar{\boldsymbol{G}}_{\boldsymbol{B}}^{t+1})_1$, where $\bar{\boldsymbol{B}}$ and $(\bar{\boldsymbol{G}}_{\boldsymbol{B}}^{t+1})_1$ are independent. Then, for $j \in [p]$, we have that,

$$f_j^{*t}(\boldsymbol{s}) = \mathbb{E}[\bar{\boldsymbol{B}}|\bar{\boldsymbol{V}}_{\boldsymbol{B}}^t = \boldsymbol{s}] = \frac{\int \bar{\boldsymbol{B}}\mathbb{P}[\bar{\boldsymbol{V}}_{\boldsymbol{B}}^{t+1} = \boldsymbol{s}|\bar{\boldsymbol{B}}]\prod_{\ell=1}^L \zeta(\alpha;\sigma_\ell)d\bar{\boldsymbol{B}}}{\int \mathbb{P}[\bar{\boldsymbol{V}}_{\boldsymbol{B}}^{t+1} = \boldsymbol{s}|\bar{\boldsymbol{B}}]\prod_{\ell=1}^L \zeta(\alpha;\sigma_\ell)d\bar{\boldsymbol{B}}} = \frac{\int \bar{\boldsymbol{B}}\phi(\boldsymbol{s};(\hat{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^\top\bar{\boldsymbol{B}}, \boldsymbol{\kappa}_{\boldsymbol{B}}^{t+1,t+1})\prod_{\ell=1}^L \zeta(\alpha;\sigma_\ell)d\bar{\boldsymbol{B}}}{\int \phi(\boldsymbol{s};(\hat{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^\top\bar{\boldsymbol{B}}, \boldsymbol{\kappa}_{\boldsymbol{B}}^{t+1,t+1})\prod_{\ell=1}^L \zeta(\alpha;\sigma_\ell)d\bar{\boldsymbol{B}}},$$

where $\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}^t)$ denotes the density of a zero-mean multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}^t$, evaluated at $\boldsymbol{x}$. The above can be computed numerically, or analytically using properties of Gaussian integrals. The latter approach yields $f_j^{*t}(\boldsymbol{s}) = \frac{d_1}{d_2}$, where:

$$d_2 = \sum_{\mathcal{S}\subseteq[L]} \alpha^{L-|\mathcal{S}|}(1-\alpha)^{|\mathcal{S}|}\int \phi\left(\bar{\boldsymbol{B}}_{\mathcal{S}};\mathbf{0},\boldsymbol{\Sigma}_{\mathcal{S}}\right)\cdot\phi\left(\boldsymbol{s}-(\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^\top\bar{\boldsymbol{B}};\mathbf{0},\bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1}\right)d\bar{\boldsymbol{B}}, \tag{82}$$

$$d_1 = \sum_{\mathcal{S}\subseteq[L]} \alpha^{L-|\mathcal{S}|}(1-\alpha)^{|\mathcal{S}|}\int \bar{\boldsymbol{B}}\phi\left(\bar{\boldsymbol{B}}_{\mathcal{S}};\mathbf{0},\boldsymbol{\Sigma}_{\mathcal{S}}\right)\cdot\phi\left(\boldsymbol{s}-(\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^\top\bar{\boldsymbol{B}};\mathbf{0},\bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1}\right)d\bar{\boldsymbol{B}}, \tag{83}$$

and $\bar{\boldsymbol{B}}_{\mathcal{S}}$ is the size $|\mathcal{S}|$ vector containing the entries of $\bar{\boldsymbol{B}}$ indexed by $\mathcal{S}$. Here, $\boldsymbol{\Sigma}_{\mathcal{S}}$ is the $|\mathcal{S}|\times|\mathcal{S}|$ sub-matrix of $\boldsymbol{\rho}$ indexed along $\mathcal{S}\times\mathcal{S}$. We then have that,

$$\begin{aligned}&\int \phi\left(\bar{\boldsymbol{B}}_{\mathcal{S}};\mathbf{0},\boldsymbol{\Sigma}_{\mathcal{S}}\right)\cdot\phi\left(\boldsymbol{s}-(\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^\top\bar{\boldsymbol{B}};\mathbf{0},\bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1}\right)d\bar{\boldsymbol{B}}\\ &= (2\pi)^{-\frac{L}{2}}\left|\boldsymbol{\Sigma}_{\mathcal{S}}\right|^{-\frac{1}{2}}\left|\bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1}\right|^{-\frac{1}{2}}\cdot\left|\boldsymbol{\Sigma}_c\right|^{\frac{1}{2}}\exp\left(-\frac{1}{2}\left[-\boldsymbol{\mu}_c^\top\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c+\boldsymbol{s}^\top(\bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1})^{-1}\boldsymbol{s}\right]\right),\end{aligned}$$

where $\boldsymbol{\Sigma}_c = \left[\boldsymbol{\Sigma}_{\mathcal{S}}^{-1}+\boldsymbol{\Delta}\right]^{-1}$ with $\boldsymbol{\Delta} = \bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1}(\bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1})^{-1}(\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^\top$ and $\boldsymbol{\mu}_c = \boldsymbol{\Sigma}_{\mathcal{S}}\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1}(\bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1})^{-1}\boldsymbol{s}$. Moreover,

$$\int \bar{\boldsymbol{B}}\phi\left(\bar{\boldsymbol{B}}_{\mathcal{S}};\mathbf{0}_L,\boldsymbol{\Sigma}_{\mathcal{S}}\right)\cdot\phi\left(\boldsymbol{s}-(\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^\top\bar{\boldsymbol{B}};\mathbf{0}_L,\bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1}\right)d\bar{\boldsymbol{B}} =: \boldsymbol{z}\in\mathbb{R}^L,$$

where

$$\begin{aligned}\boldsymbol{z}_{\mathcal{S}} &= (2\pi)^{-\frac{L}{2}}\left|\boldsymbol{\Sigma}_{\mathcal{S}}\right|^{-\frac{1}{2}}\left|\bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1}\right|^{-\frac{1}{2}}\cdot\left|\boldsymbol{\Sigma}_c\right|^{\frac{1}{2}}\exp\left(-\frac{1}{2}\left[-\boldsymbol{\mu}_c^\top\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c+\boldsymbol{s}^\top(\bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1})^{-1}\boldsymbol{s}\right]\right)\boldsymbol{\mu}_c,\\ \boldsymbol{z}_{[L]\setminus\mathcal{S}} &= \mathbf{0}.\end{aligned}$$

Plugging the above expressions into (82)–(83), we obtain a closed-form expression for f_j^{*t}.

Computation of f^{*t} for sparse difference prior For the experiments in Figure 5, we consider signals with sparse changes between adjacent signals. The prior takes the following form:

$$\begin{aligned}&\beta_j^{(\eta_0)} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,\kappa^2)\\ &\beta_j^{(\eta_1)} = \begin{cases}\beta_j^{(\eta_0)}, & \text{with probability } 1-\alpha\\ \gamma(\beta_j^{(\eta_0)}+w), & \text{with probability } \alpha\end{cases}\\ &\vdots\\ &\beta_j^{(\eta_L)} = \begin{cases}\beta_j^{(\eta_{L-1})}, & \text{with probability } 1-\alpha\\ \gamma(\beta_j^{(\eta_{L-1})}+w), & \text{with probability } \alpha\end{cases}\end{aligned} \tag{84}$$

where $w \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,\sigma_w^2)$ creates the sparse change between adjacent signals, and $\gamma := \sqrt{\frac{\kappa^2}{\kappa^2+\sigma_w^2}}$ is a rescaling factor that ensures uniform signal magnitude $\mathbb{E}[(\beta_j^{(\eta_0)})^2] = \cdots = \mathbb{E}[(\beta_j^{(\eta_L)})^2]$, i.e., $\mathbb{E}[(\bar{B}_1)^2] = \cdots = \mathbb{E}[(\bar{B}_L)^2]$. Note (84) can be compactly expressed as a mixture of Gaussians, for example, for $L = 3$:

$$\mathbb{P}_{\bar{\boldsymbol{B}}} = \pi_{00}\mathsf{N}\left(\mathbf{0}_3,\mathrm{Cov}^{00}\right)+\pi_{01}\mathsf{N}\left(\mathbf{0}_3,\mathrm{Cov}^{01}\right)+\pi_{10}\mathsf{N}\left(\mathbf{0}_3,\mathrm{Cov}^{10}\right)+\pi_{11}\mathsf{N}\left(\mathbf{0}_3,\mathrm{Cov}^{11}\right) \tag{85}$$

where

$$\pi_{00} = (1-\alpha)^2, \quad \pi_{01} = (1-\alpha)\alpha, \quad \pi_{10} = \alpha(1-\alpha), \quad \pi_{11} = \alpha^2,$$
$$\mathrm{Cov}^{00} = \kappa^2 \mathbf{1}_{3\times 3},\ \mathrm{Cov}^{01} = \kappa^2 \begin{bmatrix} 1 & 1 & \gamma \\ 1 & 1 & \gamma \\ \gamma & \gamma & 1 \end{bmatrix},\ \mathrm{Cov}^{10} = \kappa^2 \begin{bmatrix} 1 & \gamma & \gamma \\ \gamma & 1 & 1 \\ \gamma & 1 & 1 \end{bmatrix},\ \mathrm{Cov}^{11} = \kappa^2 \begin{bmatrix} 1 & \gamma & \gamma^2 \\ \gamma & 1 & \gamma \\ \gamma^2 & \gamma & 1 \end{bmatrix}.$$

In general, for any L, defining the shorthand $\boldsymbol{\upsilon} \equiv [\upsilon_1, \ldots, \upsilon_{L-1}]^\top \in \{0,1\}^{L-1}$, we have

$$\mathbb{P}_{\bar{\boldsymbol{B}}} = \sum_{\boldsymbol{\upsilon} \in \{0,1\}^{L-1}} \pi_{\boldsymbol{\upsilon}}\, \mathsf{N}(\mathbf{0}_L, \mathrm{Cov}^{\boldsymbol{\upsilon}}) \tag{86}$$

and $\{\pi_{\boldsymbol{\upsilon}}\}$ and $\{\mathrm{Cov}^{\boldsymbol{\upsilon}}\}$ can be defined analogously to those in (84). Using Lemma C.6, we obtain a closed-form expression for f^{*t} involving $\{\pi_{\boldsymbol{\upsilon}}\}$ and $\{\mathrm{Cov}^{\boldsymbol{\upsilon}}\}$.

Computation of g^{*t} for Gaussian noise $\bar{\varepsilon} \sim \mathsf{N}(0, \sigma^2)$ Recalling (21), we have

$$\begin{bmatrix} \boldsymbol{Z}_1 \\ (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1 \end{bmatrix} \sim \mathsf{N}_{2L}(\mathbf{0}, \boldsymbol{\Sigma}^t_{\boldsymbol{\Theta}}),\ \text{where } \boldsymbol{\Sigma}^t_{\boldsymbol{\Theta}} := \begin{bmatrix} \boldsymbol{\rho} & \bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}} \\ (\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})^\top & (\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})^\top \boldsymbol{\rho}^{-1} \bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}} + \bar{\boldsymbol{\kappa}}^{t,t}_{\boldsymbol{\Theta}} \end{bmatrix}. \tag{87}$$

Lemma C.5 then gives the formulas for $\mathbb{E}[\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1 = \boldsymbol{V}_1]$ and $\mathrm{Cov}(\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1)$, the first and last terms of g^{*t}_i in (26), as follows:

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1) &= \boldsymbol{\rho} - (\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})^\top \left((\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})^\top \boldsymbol{\rho}^{-1} \bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}} + \bar{\boldsymbol{\kappa}}^{t,t}_{\boldsymbol{\Theta}} \right)^{-1} \bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}}, \\ \mathbb{E}[\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1 = \boldsymbol{V}_1] &= \boldsymbol{\nu}^t_{\boldsymbol{\Theta}} \left((\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})^\top \boldsymbol{\rho}^{-1} \boldsymbol{\nu}^t_{\boldsymbol{\Theta}} + \boldsymbol{\kappa}^{t,t}_{\boldsymbol{\Theta}} \right)^{-1} \boldsymbol{V}_1. \end{aligned} \tag{88}$$

When $f^t = f^{*t}$, we can use the simplifications in Remark C.3 to obtain:

$$\mathrm{Cov}(\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1) = \boldsymbol{\rho} - \bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}} \quad \text{and} \quad \mathbb{E}[\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1 = \boldsymbol{V}_1] = \boldsymbol{V}_1.$$

We now calculate the middle term $\mathbb{E}[\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1 = \boldsymbol{V}_i, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i]$ of g^{*t}_i in (26). Recalling from (2) that $q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = (Z_1)_{\bar{\Psi}_i} + \bar{\varepsilon}$ where $\boldsymbol{Z}_1 \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\rho})$, we have

$$\begin{aligned} &\mathbb{E}[\boldsymbol{Z}_1 q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) \,|\bar{\Psi}_i = \ell] = \mathbb{E}[\boldsymbol{Z}_1 (Z_1)_\ell] = \boldsymbol{\rho}_{[:,\ell]}, \\ &\mathbb{E}[q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon})^2 \,|\bar{\Psi}_i = \ell] = \mathbb{E}[(Z_1)^2_\ell + \bar{\varepsilon}^2] = \rho_{\ell,\ell} + \sigma^2, \\ &\mathbb{E}[(\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1 q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) \,|\bar{\Psi}_i = \ell] = \mathbb{E}[(\boldsymbol{Z}\boldsymbol{\rho}^{-1} \bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})_1 (Z_1)_\ell] = (\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})^\top \boldsymbol{\rho}^{-1} \boldsymbol{\rho}_{[:,\ell]} = \left((\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})_{[\ell,:]} \right)^\top, \end{aligned} \tag{89}$$

which implies that conditioned on $\bar{\Psi}_i = \ell$,

$$\begin{bmatrix} (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1 \\ q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) \end{bmatrix} \sim \mathsf{N}_{L+1}\left(\mathbf{0}, \boldsymbol{\Sigma}^t\right),\ \text{where } \boldsymbol{\Sigma}^t := \begin{bmatrix} (\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})^\top \boldsymbol{\rho}^{-1} \bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}} + \bar{\boldsymbol{\kappa}}^{t,t}_{\boldsymbol{\Theta}} & \left((\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})_{[\ell,:]} \right)^\top \\ (\bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}})_{[\ell,:]} & \rho_{\ell,\ell} + \sigma^2 \end{bmatrix}. \tag{90}$$

Conditioned on $\Psi_i = \ell$ the random variables $(\boldsymbol{Z}_1, (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1, q_1(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}))$ are jointly Gaussian with zero mean and covariance matrix determined by (87)-(90). Using Lemma C.5 on these jointly Gaussian variables, we obtain:

$$\mathbb{E}[\boldsymbol{Z}_1 \mid (\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}})_1 = \boldsymbol{V}_i,\ q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i] = \frac{\sum_{\ell=1}^L \boldsymbol{\lambda}^t(\boldsymbol{V}_i, u_i, \ell)\, \pi_{\bar{\Psi}_i}(\ell)\, \mathcal{L}_1(\boldsymbol{V}_i, u_i | \ell)}{\sum_{\tilde{\ell}=1}^L \pi_{\bar{\Psi}_i}(\tilde{\ell})\, \mathcal{L}_1(\boldsymbol{V}_i, u_i | \tilde{\ell})}, \tag{91}$$

where $\pi_{\bar{\Psi}_i} : \ell \mapsto \sum_{\{\boldsymbol{\psi}: \psi_i = \ell\}} \pi_{\bar{\boldsymbol{\Psi}}}(\boldsymbol{\psi})$ denotes the marginal probability of $\bar{\Psi}_i$, and

$$\mathcal{L}_1(\boldsymbol{V}_i, u_i | \ell) := \phi\left(\begin{bmatrix} \boldsymbol{V}_i \\ u_i \end{bmatrix}; \mathbf{0}, \boldsymbol{\Sigma}^t \right), \tag{92}$$

$$\boldsymbol{\lambda}^t(\boldsymbol{V}_i, u_i, \ell) := \mathbb{E}[\boldsymbol{Z}_1 \mid \bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta},1} = \boldsymbol{V}_i,\ q_1(\boldsymbol{Z}_1, \bar{\Psi}_1, \bar{\varepsilon}) = u_i,\ \bar{\Psi}_1 = \ell] = \begin{bmatrix} \bar{\boldsymbol{\nu}}^t_{\boldsymbol{\Theta}} & \boldsymbol{\rho}_{[:,\ell]} \end{bmatrix} [\boldsymbol{\Sigma}^t]^{-1} \begin{bmatrix} \boldsymbol{V}_i \\ u_i \end{bmatrix} \in \mathbb{R}^L. \tag{93}$$

Here, $\phi(\boldsymbol{x}; \mathbf{0}, \boldsymbol{\Sigma}^t)$ denotes the density of a zero-mean multivariate Gaussian with covariance matrix $\boldsymbol{\Sigma}^t$, evaluated at $\boldsymbol{x}$.

D. Proof of Propositions 3.3 and 3.4

D.1. Proof of Proposition 3.3

Proof. Recall that a signal configuration vector $\boldsymbol{\Psi} \in \mathcal{X}$ is a vector that is piece-wise constant with respect to its indices $i \in [n]$, with jumps of size 1 occurring at the indices $\{\eta_\ell\}_{\ell=1}^{L^*-1}$. Without loss of generality, we assume $\boldsymbol{\Psi}$ is also monotone (otherwise, any non-distinct signal can be treated as a new signal having perfect correlation with the first). Recall that $\boldsymbol{\eta} \in [n]^{L^*-1}$ is the change point vector corresponding to $\boldsymbol{\Psi}$, i.e., $\boldsymbol{\eta} = U^{-1}(\boldsymbol{\Psi})$. Also recall that for $i \in [n]$, the distance of η_i from a change point estimate $\hat{\boldsymbol{\eta}} \in [n]^{L-1}$, is defined as $d(\eta_i, \{\hat{\eta}_j\}_{j=1}^{L-1}) := \min_{\hat{\eta}_j \in \{\hat{\eta}_j\}_{j=1}^{L-1}} \|\eta_i - \hat{\eta}_j\|_2$. Then, we have that $d(\eta_i, \{\hat{\eta}_j\}_{j=1}^{L-1}) \leq \|U(\boldsymbol{\eta}) - U(\hat{\boldsymbol{\eta}})\|_F^2$ for all $i \in [L^*-1]$, and similarly $d(\{\eta_i\}_{i=1}^{L^*-1}, \hat{\eta}_j) \leq \|U(\boldsymbol{\eta}) - U(\hat{\boldsymbol{\eta}})\|_F^2$ for all $j \in [L-1]$. This implies that $d_H(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) \leq \|U(\boldsymbol{\eta}) - U(\hat{\boldsymbol{\eta}})\|_F^2$.

We first prove that $(\boldsymbol{\Theta}^t, \boldsymbol{y}) \mapsto d_H(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))$ is uniformly pseudo-Lipschitz. Consider two inputs, $\boldsymbol{A}^{(1)} := ((\boldsymbol{\Theta}^t)^{(1)}, \boldsymbol{y}^{(1)})$ and $\boldsymbol{A}^{(2)} := ((\boldsymbol{\Theta}^t)^{(2)}, \boldsymbol{y}^{(2)})$. We then have that:

$$
\begin{aligned}
&\frac{1}{n}\left|d_H(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(1)})) - d_H(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(2)}))\right| \\
&\leq \frac{1}{n} d_H(\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(1)}), \hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(2)})) && (94)\\
&\leq \frac{1}{n}\|U(\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(1)})) - U(\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(2)}))\|_F^2 && (95)\\
&\leq L \cdot \frac{1}{\sqrt{n}}\|U(\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(1)})) - U(\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(2)}))\|_F && (96)\\
&\leq L \cdot C\left(1 + \left(\frac{\left\|\left[\boldsymbol{\Theta}^{(1)} \quad \boldsymbol{y}^{(1)}\right]\right\|_F}{\sqrt{n}}\right)^{r-1} + \left(\frac{\left\|\left[\boldsymbol{\Theta}^{(2)} \quad \boldsymbol{y}^{(2)}\right]\right\|_F}{\sqrt{n}}\right)^{r-1}\right)\frac{\left\|\left[\boldsymbol{\Theta}^{(1)} \quad \boldsymbol{y}^{(1)}\right] - \left[\boldsymbol{\Theta}^{(2)} \quad \boldsymbol{y}^{(2)}\right]\right\|_F}{\sqrt{n}}, && (97)
\end{aligned}
$$

for some constants $C > 0, r \geq 1$. Here (94) follows from the reverse triangle inequality for the metric d_H, (95) follows from the argument in the paragraph above, (96) follows from the fact that $\mathcal{X} \subseteq [L]^n$, and (97) follows from the pseudo-Lipschitz assumption on $U(\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))$. Applying Theorem 3.1 with $\boldsymbol{\varphi}_n(\boldsymbol{\Theta}^t, \boldsymbol{y}, \boldsymbol{\Psi}) := \frac{1}{n} d_H(U^{-1}(\boldsymbol{\Psi}), \hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))$, we obtain the first result in Proposition 3.3.

We now prove that $(\boldsymbol{\Theta}^t, \boldsymbol{y}) \mapsto |\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y})|$ is uniformly pseudo-Lipschitz. Denoting the nth component of $U(\hat{\boldsymbol{\eta}})$ by $U(\hat{\boldsymbol{\eta}})_n$, notice that for a change point vector $\hat{\boldsymbol{\eta}} \in [n]^{L-1}$, by the monotonicity of $\boldsymbol{\Psi}$, we have that $|\hat{\boldsymbol{\eta}}| = U(\hat{\boldsymbol{\eta}})_n$. Consider two inputs $\boldsymbol{A}^{(1)} := ((\boldsymbol{\Theta}^t)^{(1)}, \boldsymbol{y}^{(1)})$, $\boldsymbol{A}^{(2)} := ((\boldsymbol{\Theta}^t)^{(2)}, \boldsymbol{y}^{(2)})$. We then have that:

$$
\begin{aligned}
&\left||\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(1)})| - |\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(2)})|\right| && (98)\\
&= \left||U(\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(1)})_n| - |U(\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(2)})_n|\right| \\
&\leq C\left(1 + \left(\frac{\left\|\left[\boldsymbol{\Theta}^{(1)} \quad \boldsymbol{y}^{(1)}\right]\right\|_F}{\sqrt{n}}\right)^{r-1} + \left(\frac{\left\|\left[\boldsymbol{\Theta}^{(2)} \quad \boldsymbol{y}^{(2)}\right]\right\|_F}{\sqrt{n}}\right)^{r-1}\right)\frac{\left\|\left[\boldsymbol{\Theta}^{(1)} \quad \boldsymbol{y}^{(1)}\right] - \left[\boldsymbol{\Theta}^{(2)} \quad \boldsymbol{y}^{(2)}\right]\right\|_F}{\sqrt{n}}, && (99)
\end{aligned}
$$

for some constants $C > 0, r \geq 1$, where (99) follows because $U(\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))_n$ is assumed to be uniformly pseudo-Lipschitz. Applying Theorem 3.1 to $\boldsymbol{\varphi}_n(\boldsymbol{\Theta}^t, \boldsymbol{y}, \boldsymbol{\Psi}) := |U(\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))_n|$, we obtain the second claim in in Proposition 3.3. □

D.2. Proof of Proposition 3.4

Proof. Fix $\boldsymbol{\psi} \in \mathcal{X}$, and define $\boldsymbol{\varphi}_n : (\boldsymbol{\Theta}^t, \boldsymbol{y}) \mapsto p_{\boldsymbol{\Psi}|\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \bar{\boldsymbol{\varepsilon}})}(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y})$. Applying Theorem 3.1 to $\boldsymbol{\varphi}_n$, we obtain the result. □

E. Further Implementation and Experiment Details

State Evolution Implementation Our state evolution implementation involves computing (61)–(64). We estimate $\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1}, \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t+1,t+1}, \boldsymbol{\nu}_{\boldsymbol{B}}^{t+1}, \boldsymbol{\kappa}_{\boldsymbol{B}}^{t+1,t+1}$ in (61)–(64) with finite n, p via empirical averages, for a given change point configuration $\{\Psi_{\eta_\ell}\}_{\ell=0}^{L-1}$. Specifically, assuming $\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t, \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t,t}, \boldsymbol{\nu}_{\boldsymbol{B}}^t, \boldsymbol{\kappa}_{\boldsymbol{B}}^{t,t}$ have been computed, we compute $\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1}, \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t+1,t+1}, \boldsymbol{\nu}_{\boldsymbol{B}}^{t+1}, \boldsymbol{\kappa}_{\boldsymbol{B}}^{t+1,t+1}$ as

follows:

$$\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1} \approx \sum_{\ell=0}^{L-1} \frac{1}{n} \sum_{i\in[\eta_\ell,\eta_{\ell+1})} \hat{\mathbb{E}}\left[\partial_1 \tilde{g}_i^t(\boldsymbol{Z}_1, (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, \Psi_{\eta_\ell}, \bar{\varepsilon})\right], \tag{100}$$

$$\boldsymbol{\kappa}_{\boldsymbol{B}}^{t+1,t+1} \approx \sum_{\ell=0}^{L-1} \frac{1}{n} \sum_{i\in[\eta_\ell,\eta_{\ell+1})} \hat{\mathbb{E}}\left[g_i^t\left((\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, q_1(\boldsymbol{Z}_1, \Psi_{\eta_\ell}, \bar{\varepsilon})\right)^\top g_i^t\left((\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, q_1(\boldsymbol{Z}_1, \Psi_{\eta_\ell}, \bar{\varepsilon})\right)\right], \tag{101}$$

$$\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1} \approx \frac{1}{\delta}\hat{\mathbb{E}}\left[\bar{\boldsymbol{B}} f_1^{t+1}(\boldsymbol{V}_{\boldsymbol{B}}^{t+1})^\top\right], \tag{102}$$

$$\boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t+1,t+1} \approx \frac{1}{\delta}\hat{\mathbb{E}}\left[\left(f_1^{t+1}(\boldsymbol{V}_{\boldsymbol{B}}^{t+1}) - \bar{\boldsymbol{B}}\boldsymbol{\rho}^{-1}\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1}\right)\left(f^{t+1}(\boldsymbol{V}_{\boldsymbol{B}}^{t+1}) - \bar{\boldsymbol{B}}\boldsymbol{\rho}^{-1}\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1}\right)^\top\right]. \tag{103}$$

where $\hat{\mathbb{E}}$ denotes an expectation estimate via Monte Carlo. For example, in the case of (101), we generate 300 to 1000 independent samples of $(\boldsymbol{Z}_1, (\boldsymbol{G}_{\boldsymbol{\Theta}}^t)_1, \bar{\varepsilon})$, with $\boldsymbol{Z}_1 \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{\rho})$, $(\boldsymbol{G}_{\boldsymbol{\Theta}}^t)_1 \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t,t})$, and $\bar{\varepsilon} \sim \mathbb{P}_{\bar{\varepsilon}}$. We form $(\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1$ according to the first row of (4), i.e., $(\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{Z}_1\boldsymbol{\rho}^{-1}\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t + (\boldsymbol{G}_{\boldsymbol{\Theta}}^t)_1$. For $\ell \in \{0, \ldots, L-1\}$, this yields a set of samples of the random variables $g_i^t\left((\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, q_1(\boldsymbol{Z}_1, \Psi_{\eta_\ell}, \bar{\varepsilon})\right)$ and $\partial_1 \tilde{g}_i^t(\boldsymbol{Z}_1, (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, \Psi_{\eta_\ell}, \bar{\varepsilon})$ (the latter function is computed using Automatic Differentiation (AD)). We then compute $\hat{\mathbb{E}}$ in (100) and (101) by averaging. The $\hat{\mathbb{E}}$ terms in (102)–(103) are similarly computed.

The ensemble state evolution recursion (17)–(20) is used to compute the optimal denoisers $\{g^{*t}, f^{*t+1}\}_{t\geq 0}$ according to Proposition 3.2. The expectation in the ensemble state evolution iterates (17)–(20) are estimated through sample averages and AD, similar to the simplified state evolution iterates.

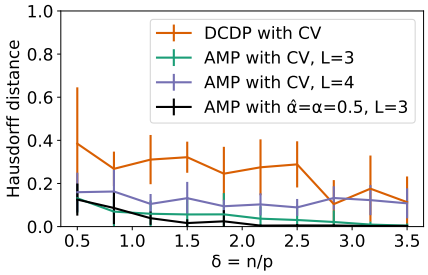
Experiment Details For the synthetic data experiments in Section 4, we consider model (2) with $\boldsymbol{\varepsilon} \overset{i.i.d}{\sim} \mathsf{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ and under the model assumptions in Section 2. For each experiment, we run AMP for $t \leq 15$ iterations, and average over 10 to 20 independent trials. We initialize AMP with $f^0(\boldsymbol{B}^0)$ sampled row-wise independently from the prior $\mathbb{P}_{\bar{\boldsymbol{B}}}$ used to define the ensemble state evolution (17)–(20). The denoisers $\{g^t, f^{t+1}\}_{t\geq 0}$ in the AMP algorithm are chosen to be the ones given by Proposition 3.2, whose computation is detailed in Appendix C.2.

Figures 4, 6, 7 and 8 use Bernoulli-Gaussian priors, as defined in Appendix C.2. Figure 6 uses $\alpha = 1/6$ and $\sigma_\ell = 2.5$ for $\ell \in [L]$. Figures 4, 7 and 8 use $\alpha = 0.5$ and $\sigma_\ell^2 = \delta$ for $\ell \in [L]$ to ensure that the signal power $\mathbb{E}[(\boldsymbol{X}_i)^\top\boldsymbol{\beta}^{(i)}]^2$ is held constant for varying δ. There are two change points at $n/3$ and $8n/15$ in the experiments in Figures 4, 7 and 8. AMP uses cross validation (CV) over 5 values of $\hat{\alpha} : \{0.1, 0.3, 0.45, 0.6, 0.9\}$ which do not contain the true $\alpha = 0.5$. `DCDP`, `DPDU`, and `DP` each have two hyperparameters: one corresponding to the ℓ_1 penalty and the other penalizing the number of change points. We run cross-validation on these hyperparameters for 12, 12, or 42 pairs of values, respectively, using the suggested values from the original papers.

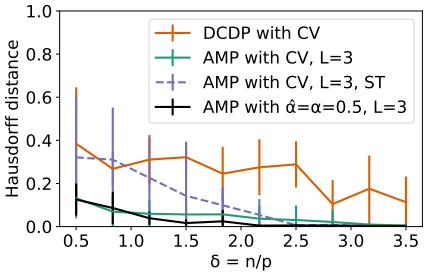
Figure 5 uses the sparse difference prior defined in Appendix C.2. We recall that, for $j \in [p]$, $\beta_j^{(\eta_0)} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \kappa^2)$ and with probability α we have that $\beta_j^{(\eta_\ell)} = \gamma(\beta_j^{(\eta_{\ell-1})} + w)$ where $w \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma_w^2)$ and $\gamma > 0$ is a normalizing constant so that $\mathbb{E}[(\beta_j^{(\eta_\ell)})^2] = \kappa^2$ for $\ell \in [L]$. We run the experiment with sparsity level $\alpha = 0.5$, variance of the entries of the first signal $\kappa^2 = 8\delta$, and perturbation to each consecutive signal of variance $\sigma_w^2 = 400\delta$, for varying δ. This ensures that the signal power $\mathbb{E}[(\boldsymbol{X}_i)^\top\boldsymbol{\beta}^{(i)}]^2$ and the signal difference $\mathbb{E}[(\boldsymbol{X}_i)^\top\boldsymbol{\beta}^{(i)} - (\boldsymbol{X}_{i+1})^\top\boldsymbol{\beta}^{(i+1)}]^2$ for $i \in \{\eta_\ell\}_{\ell=1}^{L-1}$ are held constant for varying δ. There are two change points at $n/3$ and $8n/15$. Without perfect knowledge of the magnitude σ_w^2 and sparsity level α of the sparse difference vector, AMP assumes $\hat{\sigma}_w^2 = 2500\delta$ and $\hat{\alpha} = 0.9$ in the experiments in Figure 5.

Additional numerical results Figure 7 shows results from an additional set of experiments comparing AMP with `DCDP` (the fastest algorithm in Figure 4), using different change point priors $\pi_{\bar{\boldsymbol{\Psi}}}$ (Figure 7a) or denoisers f^t (Figure 7b). In both figures, the solid red plot shows `DCDP` performance with hyperparameters chosen using CV. Solid black corresponds to AMP using the true signal prior. Solid green corresponds to AMP using $L = 3$, the optimal denoiser f^t with the sparsity level estimated using CV. In Figure 7a, AMP performs slightly worse with $L = 4$ instead of $L = 3$, because the prior assigns non-zero probability to the change point configurations with $L = 4$ signals, which is mismatched from the ground truth $L^* = 3$. In Figure 7b, AMP performs slightly worse at lower δ when using a suboptimal soft thresholding (ST) denoiser. Nevertheless, in both Figures 7a and 7b, AMP largely outperforms `DCDP` despite the suboptimal choices of prior or denoiser.

Figure 8 compares AMP with `DCDP` for $n = 500$ and varying p, both using hyperparameters chosen via cross-validation.



(a) Solid purple: AMP using $L = 3, 4$ and the optimal denoiser f^t with the sparsity level estimated using CV.



(b) Dashed purple: AMP using $L = 3$ and a suboptimal soft thresholding (ST) denoiser f^t whose threshold is selected using CV.

Figure 7. Comparison between AMP and DCDP in the same setting as in Figure 4: $p = 200, \sigma = 0.1, L^* = 3$, sparse signal prior $\mathbb{P}_{\bar{\boldsymbol{B}}} = 0.5\mathsf{N}(\mathbf{0}, \delta \boldsymbol{I}) + 0.5\delta_{\mathbf{0}}$ and AMP uses $\Delta = n/10$. Solid red: DCDP with hyperparameters chosen using CV. Solid green: AMP using $L = 3$, the optimal denoiser f^t with the sparsity level estimated using CV. Solid black: AMP using the true signal prior.

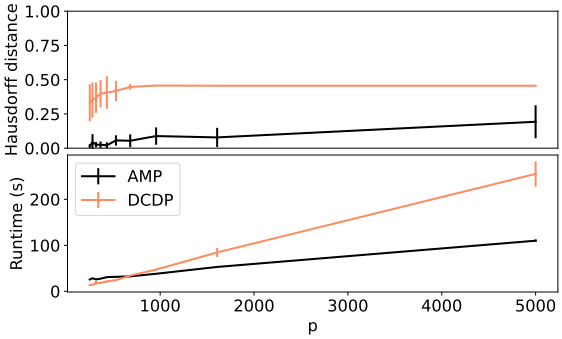


Figure 8. Comparison between AMP and DCDP for fixed $n = 500$ and varying p, with $\sigma = 0.1$, $L^* = 3$ and sparse signal prior $\mathbb{P}_{\bar{\boldsymbol{B}}} = 0.5\mathsf{N}(\mathbf{0}, \delta \boldsymbol{I}) + 0.5\delta_{\mathbf{0}}$. AMP uses $\Delta = n/10$ and $L = 3$.

The runtime shown is the average runtime per set of CV parameters. For a fixed set of hyperparameters, the time complexity of DCDP scales as LASSO(n, p) compared to $O(np)$ for AMP. The runtime of DCDP therefore grows faster with increasing p.

F. Computational Cost

Computing g^{*t}, f^{*t+1} For $i \in [n]$, the function g_i^{*t} in Proposition 3.2 only depends on $\bar{\boldsymbol{\Psi}}$ in (91) through the marginal probability of $\bar{\Psi}_i$, i.e., $\pi_{\bar{\Psi}_i} : \ell \mapsto \sum_{\{\boldsymbol{\psi}:\psi_i=\ell\}} \pi_{\bar{\boldsymbol{\Psi}}}(\boldsymbol{\psi})$. This means g_i^{*t} can be efficiently computed, involving only a sum over $\bar{\Psi}_i \in [L]$. Indeed, from the implementation details in Appendix C.2, both f_j^{*t} and g_i^{*t} can be computed in $O(L^3)$ time for each i, j. Thus f^{*t} and g^{*t} can be computed in $O(nL^3)$. The per-iteration computational cost of AMP is therefore dominated by the matrix multiplications in (3), which are $O(npL)$.

Computing change point estimators For estimators $\hat{\boldsymbol{\eta}}$ with $O(np)$ runtime, the combined AMP and estimator computation can be made to run in $O(np)$ time by selecting denoisers f^t, g^t as in Proposition 3.2. For example, the $\arg\max$ in (27) can be replaced with a greedy best-first search: search for the location of one change point at a time, conditioning on past estimates of the other change points. This will yield at most L rounds of searching over $O(n)$ elements, resulting in $O(np)$ total runtime.

Computing the approximate posterior For low-dimensional latent random variables $\boldsymbol{\Psi}$, the denominator in (30) only requires evaluating a polynomial number of terms, $\binom{n}{L}$ in the case of change point signal configurations. Therefore, choosing f^t, g^t as in Proposition 3.2, the computational complexity of computing $p_{\boldsymbol{\Psi}|\bar{\boldsymbol{V}}^t_{\boldsymbol{\Theta}}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \bar{\boldsymbol{\varepsilon}})}(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y})$ along with AMP is

$O(npL + nL^3 + n^L) = O(np + n^L)$.

G. Background on AMP for Generalized Linear Models (GLMs)

Here we review the AMP algorithm and its state evolution characterization for the GLM without change points. As in Section 3, let $\boldsymbol{B} \in \mathbb{R}^{p\times L}$ be a signal matrix and let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be a design matrix . The observation $\boldsymbol{y} \in \mathbb{R}^{n\times L}$ is produced as

$$\boldsymbol{y} = q(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{\varepsilon}) \in \mathbb{R}^{n\times L}, \tag{104}$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a noise vector, and $q : \mathbb{R}^L \to \mathbb{R}^L$ is a known output function. The only difference between this model and the one in (2) is the absence of the signal configuration vector $\boldsymbol{\Psi}$. The AMP algorithm for the GLM in (104) was derived by Rangan (2011) for the case of vector signals ($L = 1$); see also Section 4 of (Feng et al., 2022). Here we discuss the algorithm for the general case ($L \geq 1$), which can be found in (Tan & Venkataramanan, 2023). For ease of exposition, we will make the following standard assumption (see (S0) on p.5): as $n, p \to \infty$, the empirical distributions of $\{\boldsymbol{B}_j\}_{j\in[p]}$ and $\{\varepsilon_i\}_{i\in[n]}$ converge weakly to laws $\mathbb{P}_{\bar{\boldsymbol{B}}}$ and $\mathbb{P}_{\bar{\varepsilon}}$, respectively, with bounded second moments. We also recall that $n/p \to \delta$ as $n, p \to \infty$.

The AMP algorithm for the model (104) is the same as the one in (3), but due to the assumption above, we can take f^t, g^t to be separable, i.e., $f^t : \mathbb{R}^L \to \mathbb{R}^L$ and $g^t : \mathbb{R}^L \times \mathbb{R}^L \to \mathbb{R}^L$ act row-wise on their matrix inputs. Then, the matrices $\boldsymbol{F}^t$ and $\boldsymbol{C}^t$ in (3) can be simplified to $\boldsymbol{C}^t = \frac{1}{n}\sum_{i=1}^n \mathrm{d}g^t(\boldsymbol{\Theta}_i^t, y_i)$ and $\boldsymbol{F}^t = \frac{1}{n}\sum_{j=1}^p \mathrm{d}f^t(\boldsymbol{B}_j^t)$, where $\mathrm{d}g^t$ and $\mathrm{d}f^t$ denote the $L \times L$ Jacobians with respect to the first argument.

State evolution The memory terms $-\hat{\boldsymbol{R}}^{t-1}(\boldsymbol{F}^t)^\top$ and $-\hat{\boldsymbol{B}}^t(\boldsymbol{C}^t)^\top$ in (3) debias the iterates $\boldsymbol{\Theta}^t$ and $\boldsymbol{B}^{t+1}$ and enable a succinct distributional characterization, guaranteeing that their empirical distributions converge to well-defined limits as $n, p \to \infty$. Specifically, Theorem 1 in (Tan & Venkataramanan, 2023) shows that for each $t \geq 1$, the empirical distribution of the rows of $\boldsymbol{B}^t$ converges to the law of a random vector $\tilde{\boldsymbol{V}}_{\boldsymbol{B}}^t := \bar{\boldsymbol{B}}\tilde{\boldsymbol{\nu}}_{\boldsymbol{B}}^t + \tilde{\boldsymbol{G}}_{\boldsymbol{B}}^t \in \mathbb{R}^{1\times L}$, where $\tilde{\boldsymbol{G}}_{\boldsymbol{B}}^{t+1} \sim \mathsf{N}(0, \tilde{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t,t})$ is independent of $\bar{\boldsymbol{B}} \sim \mathbb{P}_{\bar{\boldsymbol{B}}}$. Similarly, recalling that $\boldsymbol{\Theta} = \boldsymbol{X}\boldsymbol{B} \in \mathbb{R}^{n\times L}$, the empirical distribution of the rows of $(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t)$ converges to the law of the random vectors $(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t + \tilde{\boldsymbol{G}}_{\boldsymbol{\Theta}}^t)$, where $\tilde{\boldsymbol{G}}_{\boldsymbol{\Theta}}^{t+1} \sim \mathsf{N}(0, \tilde{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{t,t})$ is independent of $\tilde{\boldsymbol{Z}} \sim \mathsf{N}\left(0, \delta(\mathbb{E}[\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\mathsf{T}])^{-1}\right)$. The deterministic $L \times L$ matrices $\tilde{\boldsymbol{\nu}}_{\boldsymbol{B}}^t, \tilde{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t,t}, \tilde{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^{t,t}$, and $\tilde{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{t,t}$ can be recursively computed for $t \geq 1$ via a state evolution recursion that depends on f^t, g^t and the limiting laws $\mathbb{P}_{\bar{\boldsymbol{B}}}$ and $\mathbb{P}_{\bar{\varepsilon}}$.

The state evolution characterization allows us to compute asymptotic performance measures such as the MSE of the AMP algorithm. Indeed, for each $t \geq 1$, we almost surely have $\lim_{p\to\infty} \frac{1}{p}\|f^t(\boldsymbol{B}^t) - \boldsymbol{B}\|_F^2 = \mathbb{E}[\|f^t(\tilde{\boldsymbol{V}}_{\boldsymbol{B}}^t) - \bar{\boldsymbol{B}}\|_2^2]$, where the expectation on the right can be computed using the joint law of the L-dimensional random vectors $\bar{\boldsymbol{B}}$ and $\tilde{\boldsymbol{V}}_{\boldsymbol{B}}^t = \bar{\boldsymbol{B}}\tilde{\boldsymbol{\nu}}_{\boldsymbol{B}}^t + \tilde{\boldsymbol{G}}_{\boldsymbol{B}}^t$.

In the model (3) with change points, we have an additional signal configuration vector $\boldsymbol{\Psi}$, because of which we cannot take the AMP denoising function g^t to be separable, even under Assumption (S0). This leads to a more complicated state evolution characterization, as described in Section 3.