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

Gabriel Arpino   GA442@CAM.AC.UK
Xiaoqi Liu   XL394@CAM.AC.UK
Julia Gontarek   JG991@CANTAB.AC.UK
Ramji Venkataramanan   RV285@CAM.AC.UK
Department of Engineering, University of Cambridge
Cambridge, CB2 1PZ, United Kingdom

Abstract

We consider the problem of localizing change points in a generalized linear model (GLM), a model that covers many widely studied problems in statistical learning including linear, logistic, and rectified linear regression. We propose a novel and computationally efficient approximate message passing (AMP) algorithm for estimating both the signals and the change point locations, and rigorously characterize its performance in the high-dimensional limit where the number of parameters p is proportional to the number of samples n. This characterization is in terms of a state evolution recursion, which allows us to precisely compute performance measures such as the asymptotic Hausdorff error of our change point estimates, and allows us to tailor the algorithm to take advantage of any prior structural information of the signals and change points. Moreover, we show how our AMP iterates can be used to efficiently compute a Bayesian posterior distribution over the change point locations in the high-dimensional limit. We validate our theory via numerical experiments, and demonstrate the favorable performance of our estimators on both synthetic and real data in the settings of linear, logistic, and rectified linear regression.

Keywords: change point detection, high-dimensional regression, generalized linear models, multi-index models, approximate message passing, data segmentation

1. Introduction

Heterogeneity is a common feature of large, high-dimensional data sets. 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 data set 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 et al., 2020; Liu et al., 2021; Li et al., 2024); covariance structures (Cho and Fryzlewicz, 2014; Wang et al., 2021b); graphs (Londschien et al., 2023; Bhattacharjee et al., 2020; Fan and 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 and Kirch, 2012), causality (He et al., 2022), and economics (Andreou and Ghysels, 2002).

In this paper, we consider high-dimensional generalized linear models (GLMs) 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

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

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

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

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, for example, via confidence sets or a posterior distribution. As we describe below, the model (1) covers many widely studied regression models including linear, logistic, and rectified linear regression.

1.1 Linear Regression with Change Points

In this model, the data $(y_i, \boldsymbol{X}_i) \in \mathbb{R} \times \mathbb{R}^p$ are generated as follows:

$$y_i = (\boldsymbol{X}_i)^\top \boldsymbol{\beta}^{(i)} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2}$$

Notice that this corresponds to model (1) with $q(z, v) := z + v$, where for $i \in [n]$ the projected signal $(\boldsymbol{X}_i)^\top \boldsymbol{\beta}^{(i)}$ is observed in additive noise ε_i. When $L^* = 1$, this reduces to standard linear regression.

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, such as Lee et al. (2015); Leonardi and Bühlmann (2016); Kaul et al. (2019); Rinaldo et al. (2021); Xu et al. (2024); Li et al. (2023); Bai and Safikhani (2024). 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 such as dynamic programming. The work of Gao and Wang (2022) assumes sparsity on the difference between signals across a change point, and Cho et al. (2025); Liu et al. (2026) consider general non-sparse signals. The recent works of Liu et al. (2026); Yang et al. (2025) propose change point detection methods for high-dimensional linear models that do not rely on strict sub-Gaussian error assumptions, with performance comparable to that of Xu et al. (2024) in the case of light-tailed additive noise distributions.

1.2 Logistic Regression with Change Points

For the logistic model, we first define the function $\zeta(z) := \log(1 + e^z)$. The data $(y_i, \boldsymbol{X}_i) \in \{0, 1\} \times \mathbb{R}^p$ are then generated according to the following model:

$$\mathbb{P}\left[y_i = 1 \middle| (\boldsymbol{X}_i)^\top \boldsymbol{\beta}^{(i)}\right] = \frac{e^{(\boldsymbol{X}_i)^\top \boldsymbol{\beta}^{(i)}}}{1 + e^{(\boldsymbol{X}_i)^\top \boldsymbol{\beta}^{(i)}}} = \zeta'((\boldsymbol{X}_i)^\top \boldsymbol{\beta}^{(i)}), \quad i = 1, \ldots, n. \tag{3}$$

We may view this as an instance of (1) with $\varepsilon_1, \ldots, \varepsilon_n \overset{i.i.d}{\sim} U[0, 1]$ and $q(z, v) = \mathbb{1}\{v \leq \zeta'(z)\}$, so that $y_i = q((\boldsymbol{X}_i)^\top \boldsymbol{\beta}^{(i)}, \varepsilon_i) := \mathbb{1}\{\varepsilon_i \leq \zeta'((\boldsymbol{X}_i)^\top \boldsymbol{\beta}^{(i)})\}$ for each $i \in [n]$. When $L^* = 1$, model (3) corresponds to the standard logistic model.

Logistic models are widely used for classification, and logistic regression with change points has been studied in number of contexts. In epidemiology, for example, logistic regression with unknown change points has been used to model the relationship between the continuous exposure variable and disease risk (Pastor-Barriuso et al., 2003). In medicine, it has been used to identify relevant immune response biomarkers in patients with potentially infectious diseases (Fong et al., 2015).

1.3 Rectified Linear Regression with Change Points

In the rectified linear regression model, the data $(y_i, \boldsymbol{X}_i) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^p$ are generated as follows:

$$y_i = \max\{(\boldsymbol{X}_i)^\top \boldsymbol{\beta}^{(i)}, 0\} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{4}$$

This is an instance of (1) with $q(z, v) := \max\{z, 0\} + v$. When $L^* = 1$, the model (4) is referred to as the ReLU (rectified linear unit) with additive noise. The output function acts as a threshold that maps negative values of its first input to zero.

Rectified linear regression is an important primitive in the theory of deep learning, and the problem without change points has been extensively studied in machine learning in recent years, both from the perspective of designing new algorithms and from that of computational hardness, as seen in the works of Soltanolkotabi (2017); Diakonikolas et al. (2021).

Challenges in the High-Dimensional Setting Change point detection in generic, low-dimensional, GLMs was studied by Hofrichter (2007), and Wang et al. (2023) recently proposed a method for detecting change points in high-dimensional GLMs with sparse regression vectors. Assuming s-sparse signals, they propose a change point estimator that combines an ℓ_1-penalized estimator with a partitioning technique, and show that it is consistent under the sparsity requirement $s = o(\sqrt{n}/\log p)$.

Although existing procedures for high-dimensional change point regression such as Wang et al. (2023) incorporate sparsity-based constraints, they cannot be easily adapted to take advantage of other kinds of signal priors. Moreover, they are not well-equipped to exploit prior information on the change point locations. Bayesian approaches to change point detection have been studied in several works, such as those of 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 $\{\eta_\ell\}_{\ell=1}^{L^*-1}$ in model (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 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 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 a limiting 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, 4).

In Section 4 we present a range of numerical experiments that demonstrate the superior performance of AMP compared to other state-of-the-art algorithms for high-dimensional regression with change points. We consider synthetic data as well as a real Myocardial Infarctions (MI) data set sorted by age. In the latter, we identify a change point that is consistent with medical literature, and further validate this finding using standard logistic regression analysis.

Although our theoretical results do not explicitly need assumptions on the separation between adjacent change points, they are most interesting when the separation is of order n, that is, $\Delta := \min_{\ell \in [L]}(\eta_\ell - \eta_{\ell-1}) = O(n)$. 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 with s sufficiently small, 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 (Wang et al., 2021c; Li et al., 2023; Wang et al., 2023). In contrast, we do not assume signal sparsity that is sublinear in n, so the change point estimation error will not converge to zero unless $n/p \to \infty$. We therefore quantify the AMP performance via precise asymptotics for the estimation error and the limiting 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 and Rangan, 2014; Maillard et al., 2020; Mondelli and Venkataramanan, 2021) and low-rank matrix estimation (Fletcher and Rangan, 2018; Lesieur et al., 2017; Montanari and 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 and Montanari, 2014; Donoho et al., 2013; Barbier et al., 2019). We refer the reader to Appendix G for further background on AMP for Generalized Linear Models.

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 heterogeneous dependence structure caused by change points, and allow for precise uncertainty quantification around possible change point locations. Our main state evolution result (Theorem 1) leverages recent results by Berthier et al. (2019) and Gerbelot and Berthier (2023) for AMP with non-separable denoisers. Non-separable denoisers for AMP have been been studied for generalized linear models and for matrix tensor product models, to exploit the dependence within the signal (Som and Schniter, 2012; Metzler et al., 2016; Ma et al., 2019; Rossetti and Reeves, 2023) or between the covariates (Zhang et al., 2024). In our setting, non-separable denoisers are required to exploit the heterogeneous dependence structure caused by change points.

Although we assume i.i.d. Gaussian covariates, based on recent AMP universality results (Wang et al., 2024), 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 and Ping, 2017; Rangan et al., 2019; Takeuchi, 2020; Pandit et al., 2020).

2. Preliminaries

Notation   We let $[n] = \{1, 2, \ldots, n\}$. 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, that is, $\lim_{n \to \infty} \mathbb{P}(|X_n - Y_n| > \epsilon) = 0$ 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, including 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. Given a matrix $\boldsymbol{B} \in \mathbb{R}^{p \times L}$, the empirical distribution of the entries of $\boldsymbol{B}$ is the measure $\mu_p(A) := \frac{1}{p} |\{\boldsymbol{B}_j \in A : j \in [p]\}|$ for any measurable set $A \subseteq \mathbb{R}^L$. We say that the empirical distribution of the entries of $\boldsymbol{B}$ converges weakly to the law $\mathbb{P}_{\bar{\boldsymbol{B}}}$ if $\frac{1}{p} \sum_{j=1}^p f(\boldsymbol{B}_j) \to \mathbb{E}_{\mathbb{P}_{\bar{\boldsymbol{B}}}}[f(\bar{\boldsymbol{B}})]$ as $p \to \infty$ for all bounded continuous functions $f : \mathbb{R}^L \to \mathbb{R}$.

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   We will state our results 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

$$\frac{\|\phi(\boldsymbol{x}) - \phi(\tilde{\boldsymbol{x}})\|_F}{\sqrt{m}} \le C \left(1 + \left(\frac{\|\boldsymbol{x}\|_F}{\sqrt{n}}\right)^{r-1} + \left(\frac{\|\tilde{\boldsymbol{x}}\|_F}{\sqrt{n}}\right)^{r-1}\right) \frac{\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_F}{\sqrt{n}}, \quad \forall \boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathbb{R}^{n \times q}.$$

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 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, 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 to form the design matrix $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]^\top \in \mathbb{R}^{n \times p}$. Similarly, we stack the observations and noise elements to form $\boldsymbol{y} := [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$ and $\boldsymbol{\varepsilon} := [\varepsilon_1, \ldots, \varepsilon_n]^\top \in \mathbb{R}^n$, respectively. Let $\boldsymbol{\eta} := [\eta_1, \ldots, \eta_{L^*}, \ldots, \eta_L]$ be the vector containing the true change points, and let

$$\boldsymbol{B} := [\boldsymbol{\beta}^{(\eta_1)}, \ldots, \boldsymbol{\beta}^{(\eta_{L^*})}, \ldots, \boldsymbol{\beta}^{(\eta_L)}] \in \mathbb{R}^{p \times L}$$

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 there is a one-to-one correspondence between $\boldsymbol{\Psi}$ and the change point vector $\boldsymbol{\eta}$. With a slight abuse of notation, we can rewrite the response vector $\boldsymbol{y}$ in a more general form:

$$\boldsymbol{y} =: q(\boldsymbol{XB}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) \in \mathbb{R}^n, \tag{5}$$

where q is the known function in model (1), expanded to act row-wise on matrix inputs and to incorporate $\boldsymbol{\Psi}$, where $(q(\boldsymbol{XB}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))_i := q(\boldsymbol{X}\boldsymbol{\beta}^{(\Psi_i)}, \varepsilon_i)$ for $i \in [n]$. We note that the mixed linear regression model (Yi et al., 2014; Zhang et al., 2022; Tan and Venkataramanan, 2023) can also be written in the form in (5), with a crucial difference. In mixed regression, the components of $\boldsymbol{\Psi}$ are assumed to be drawn independently from some distribution on $[L]$, that is, 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 indices $\eta_1, \ldots, \eta_{(L^*-1)}$.

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 $\boldsymbol{\Theta} := \boldsymbol{XB}$, 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:

$$\begin{aligned}
\boldsymbol{\Theta}^t &= \boldsymbol{X}\hat{\boldsymbol{B}}^t - \hat{\boldsymbol{R}}^{t-1}(\boldsymbol{F}^t)^\top, \quad \hat{\boldsymbol{R}}^t = g^t\left(\boldsymbol{\Theta}^t, \boldsymbol{y}\right), \\
\boldsymbol{B}^{t+1} &= \boldsymbol{X}^\top \hat{\boldsymbol{R}}^t - \hat{\boldsymbol{B}}^t (\boldsymbol{C}^t)^\top, \quad \hat{\boldsymbol{B}}^t = f^t\left(\boldsymbol{B}^t\right),
\end{aligned} \tag{6}$$

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 $\partial_i g_i^t (\boldsymbol{\Theta}, \boldsymbol{y})$ 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 (6) 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, that is, $g : \mathbb{R}^{m \times L} \to \mathbb{R}^{m \times L}$ 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 heterogeneous dependence structure created by the change points. For example, if there was one change point uniformly 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 and Venkataramanan, 2023), where under standard model assumptions, it suffices to consider separable denoisers.

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 (6) 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 $\boldsymbol{V}_{\boldsymbol{\Theta}}^t$ and $\boldsymbol{V}_{\boldsymbol{B}}^{t+1}$ respectively, where

$$\boldsymbol{V}_{\boldsymbol{\Theta}}^t := \boldsymbol{Z} \boldsymbol{\rho}^{-1} \boldsymbol{\nu}_{\boldsymbol{\Theta}}^t + \boldsymbol{G}_{\boldsymbol{\Theta}}^t \in \mathbb{R}^{n \times L}, \tag{7}$$

$$\boldsymbol{V}_{\boldsymbol{B}}^{t+1} := \boldsymbol{B} \boldsymbol{\nu}_{\boldsymbol{B}}^{t+1} + \boldsymbol{G}_{\boldsymbol{B}}^{t+1} \in \mathbb{R}^{p \times L}. \tag{8}$$

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

Given the initializer $\boldsymbol{B}^0$ for our algorithm (6), we initialize state evolution by setting $\boldsymbol{\nu}_{\boldsymbol{\Theta}}^0 := \boldsymbol{0}$, and

$$\boldsymbol{\rho} := \frac{1}{\delta} \lim_{p \to \infty} \frac{1}{p} \boldsymbol{B}^\top \boldsymbol{B}, \qquad \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{0,0} := \frac{1}{\delta} \lim_{p \to \infty} \frac{1}{p} f^0(\boldsymbol{B}^0)^\top f^0(\boldsymbol{B}^0). \tag{9}$$

Let $\tilde{g}_i^t(\boldsymbol{Z}, \boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) := g_i^t(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))$ and let $\partial_{1i} \tilde{g}_i^t$ 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:

$$\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1} := \lim_{n \to \infty} \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^{n} \partial_{1i} \tilde{g}_i^t (\boldsymbol{Z}, \boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) \right], \tag{10}$$

$$\boldsymbol{\kappa}_{\boldsymbol{B}}^{s+1,t+1} := \lim_{n \to \infty} \frac{1}{n} \mathbb{E} \left[g^s \left(\boldsymbol{V}_{\boldsymbol{\Theta}}^s, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) \right)^\top g^t \left(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) \right) \right], \tag{11}$$

$$\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1} := \frac{1}{\delta} \lim_{p \to \infty} \frac{1}{p} \mathbb{E} \left[\boldsymbol{B}^\top f^{t+1} (\boldsymbol{V}_{\boldsymbol{B}}^{t+1}) \right], \tag{12}$$

$$\boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{s+1,t+1} := \frac{1}{\delta} \lim_{p \to \infty} \frac{1}{p} \mathbb{E} \left[\left(f^{s+1} (\boldsymbol{V}_{\boldsymbol{B}}^{s+1}) - \boldsymbol{B} \boldsymbol{\rho}^{-1} \boldsymbol{\nu}_{\boldsymbol{\Theta}}^{s+1} \right)^\top \left(f^{t+1} (\boldsymbol{V}_{\boldsymbol{B}}^{t+1}) - \boldsymbol{B} \boldsymbol{\rho}^{-1} \boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1} \right) \right]. \tag{13}$$

The expectations above are taken with respect to $\boldsymbol{Z}, \boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{V}_{\boldsymbol{\Theta}}^s, \boldsymbol{V}_{\boldsymbol{B}}^{t+1}$ and $\boldsymbol{V}_{\boldsymbol{B}}^{s+1}$, and depend on $g^t, f^t, \boldsymbol{B}, \boldsymbol{\varepsilon}$, and $\boldsymbol{\Psi}$. The limits in (10)–(13) 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 parametric 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 (9)–(13) are used to define the random variables $(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{V}_{\boldsymbol{B}}^{t+1})$ in (7),(8). 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 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 $(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{V}_{\boldsymbol{B}}^{t+1})$. 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: $f^t, \tilde{g}^t, \mathrm{d}_j f_j^t, \partial_{1i} \tilde{g}_i^t$.

(A3) For $s, t \geq 0$, the limits in (9)–(13) exist and are finite.

Assumptions $(A1)-(A3)$ are natural extensions of those required for classical AMP results in settings with separable denoising functions. They are similar to those required by existing works on non-separable AMP Berthier et al. (2019); Gerbelot and Berthier (2023), and generalize these to the model (5), with a matrix signal $\boldsymbol{B}$ and an auxiliary vector $\boldsymbol{\Psi} \in [L]^n$.

Theorem 1 *Consider the model (5) and the AMP algorithm in (6), 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 $\boldsymbol{\varphi}_p(\cdot\,; \boldsymbol{B}) : \mathbb{R}^{p \times (L(t+1))} \to \mathbb{R}$ and $\boldsymbol{\varphi}_n(\cdot\,; \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) : \mathbb{R}^{n \times (L(t+2))} \to \mathbb{R}$,*

$$\boldsymbol{\varphi}_n(\boldsymbol{\Theta}^0, \dots, \boldsymbol{\Theta}^t, \boldsymbol{X}\boldsymbol{B}; \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) \stackrel{\mathbb{P}}{\simeq} \mathbb{E}_{\boldsymbol{V}_{\boldsymbol{\Theta}}^0, \dots, \boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{Z}} \{ \boldsymbol{\varphi}_n(\boldsymbol{V}_{\boldsymbol{\Theta}}^0, \dots, \boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{Z}; \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) \}, \tag{14}$$

$$\boldsymbol{\varphi}_p(\boldsymbol{B}^1, \dots, \boldsymbol{B}^{t+1}; \boldsymbol{B}) \stackrel{\mathbb{P}}{\simeq} \mathbb{E}_{\boldsymbol{V}_{\boldsymbol{B}}^1, \dots, \boldsymbol{V}_{\boldsymbol{B}}^{t+1}} \{ \boldsymbol{\varphi}_p(\boldsymbol{V}_{\boldsymbol{B}}^1, \dots, \boldsymbol{V}_{\boldsymbol{B}}^{t+1}; \boldsymbol{B}) \}, \tag{15}$$

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 (7), (8).

The proof of the theorem is given in Appendix A. It involves reducing the AMP in (6) to a variant of the symmetric AMP iteration analyzed in (Gerbelot and Berthier, 2023, Lemma 14). We use a generalized version of their iteration which allows for the auxiliary quantities $\boldsymbol{XB}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}$ to be included. Theorem 1 implies that any pseudo-Lipschitz function $\boldsymbol{\varphi}_n(\boldsymbol{\Theta}^t, \boldsymbol{XB}; \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) := \tilde{\boldsymbol{\varphi}}_n(\boldsymbol{\Theta}^t, q(\boldsymbol{XB}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})) = \tilde{\boldsymbol{\varphi}}_n(\boldsymbol{\Theta}^t, \boldsymbol{y})$ will converge in probability to a quantity involving an expectation over $(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{Z})$. An analogous statement holds for $(\boldsymbol{B}^{t+1}, \boldsymbol{B})$.

Evaluating Performance Metrics Using Theorem 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 limiting value of the RHS (as $p \to \infty$) 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 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 and Venkataramanan, 2023).

Parametric Dependence of State Evolution on $\boldsymbol{B}, \boldsymbol{\varepsilon}, \boldsymbol{\Psi}$ The parametric dependence of the state evolution parameters in (10)-(13) 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 parametric dependence of the state evolution matrices 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., 2024) 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 (6) 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 (7)–(8), we define the matrices

$$\tilde{\boldsymbol{Z}}^t := \boldsymbol{V}_{\boldsymbol{\Theta}}^t (\rho^{-1}\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t)^{-1} = \boldsymbol{Z} + \boldsymbol{G}_{\boldsymbol{\Theta}}^t (\rho^{-1}\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t)^{-1}, \tag{16}$$

$$\tilde{\boldsymbol{B}}^{t+1} := \boldsymbol{V}_{\boldsymbol{B}}^{t+1} (\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1})^{-1} = \boldsymbol{B} + \boldsymbol{G}_{\boldsymbol{B}}^{t+1} (\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1})^{-1}, \tag{17}$$

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 (16) and (17), given by

$$\mathrm{Tr}\left(\frac{1}{n} \sum_{i=1}^n \mathrm{Cov}\left(\tilde{\boldsymbol{Z}}_i^t - \boldsymbol{Z}_i\right)\right) = \mathrm{Tr}\left(((\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t)^{-1}\boldsymbol{\rho})^\top \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t,t}(\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t)^{-1}\boldsymbol{\rho}\right), \tag{18}$$

$$\mathrm{Tr}\left(\frac{1}{p} \sum_{j=1}^p \mathrm{Cov}\left(\tilde{\boldsymbol{B}}_i^{t+1} - \boldsymbol{B}_i\right)\right) = \mathrm{Tr}\left([(\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1})^{-1}]^\top \boldsymbol{\kappa}_{\boldsymbol{B}}^{t+1,t+1}(\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1})^{-1}\right), \tag{19}$$

where we recall from (10)–(13) that $\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t, \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t,t}$ are defined by f^t, and $\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1}, \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t+1,t+1}$ 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 (18) and (19). 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 parametric 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 $\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^0 := \boldsymbol{0}$, $\bar{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{0,0} := \lim_{p\to\infty} \frac{1}{\delta p} f^0(\boldsymbol{B}^0)^\top f^0(\boldsymbol{B}^0)$, for $t \geq 0$ define:

$$\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1} := \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\partial_1 \tilde{g}_i^t(\boldsymbol{Z}_1, (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1, \bar{\Psi}_i, \bar{\varepsilon})\right] \tag{20}$$

$$\bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1} := \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[g_i^t\left((\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon})\right) g_i^t\left((\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon})\right)^\top\right], \tag{21}$$

$$\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^{t+1} := \frac{1}{\delta} \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^p \mathbb{E}\left[\bar{\boldsymbol{B}} f_j^{t+1}(\bar{\boldsymbol{V}}_{\boldsymbol{B}}^{t+1})^\top\right], \tag{22}$$

$$\bar{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{t+1,t+1} := \frac{1}{\delta} \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^p \mathbb{E}\left[\left(f_j^{t+1}(\bar{\boldsymbol{V}}_{\boldsymbol{B}}^{t+1}) - (\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^{t+1})^\top \boldsymbol{\rho}^{-1}\bar{\boldsymbol{B}}\right)\right.$$
$$\left.\left(f_j^{t+1}(\bar{\boldsymbol{V}}_{\boldsymbol{B}}^{t+1}) - (\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^{t+1})^\top \boldsymbol{\rho}^{-1}\bar{\boldsymbol{B}}\right)^\top\right], \tag{23}$$

where

$$\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t := \boldsymbol{Z}\boldsymbol{\rho}^{-1}\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t + \bar{\boldsymbol{G}}_{\boldsymbol{\Theta}}^t \in \mathbb{R}^{n \times L}, \tag{24}$$

$$\bar{\boldsymbol{V}}_{\boldsymbol{B}}^{t+1} := (\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^\top \bar{\boldsymbol{B}} + (\bar{\boldsymbol{G}}_{\boldsymbol{B}}^{t+1})_1 \in \mathbb{R}^L, \tag{25}$$

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

When $\pi_{\bar{\boldsymbol{\Psi}}}$ is a unit mass on the true configuration $\boldsymbol{\Psi}$, (20)–(23) reduce to the simplified state evolution in Appendix B. The limits in (20)–(23) 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 (18)–(19):

$$\mathrm{Tr}\left(\left((\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^{-1} \boldsymbol{\rho} \right)^\top \bar{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{t,t} (\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^{-1} \boldsymbol{\rho} \right), \tag{26}$$

$$\mathrm{Tr}\left([(\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^{-1}]^\top \bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1} (\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^{-1} \right), \tag{27}$$

where the deterministic matrices $\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t, \bar{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{t,t}, \bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1}, \bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1} \in \mathbb{R}^{L \times L}$ are defined in (20)–(23).

Proposition 2 *Assume the limits in* (20)–(23) *exist. Then, for $t \geq 1$:*

1. *Given $\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^t, \bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t,t}$, the quantity* (26) *is minimized when*

$$f_j^t(\boldsymbol{U}) = f_j^{*t}(\boldsymbol{U}) := \mathbb{E}[\bar{\boldsymbol{B}}|\bar{\boldsymbol{V}}_{\boldsymbol{B}}^t = \boldsymbol{U}_j], \tag{28}$$

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

2. *Given $\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t, \bar{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{t,t}$, the quantity* (27) *is minimized when*

$$g_i^t(\boldsymbol{V}, \boldsymbol{u}) = g_i^{*t}(\boldsymbol{V}, \boldsymbol{u}) :=$$
$$\left[\mathrm{Cov}\left(\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i \right) \right]^{-1} \left(\mathbb{E}[\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i] - \mathbb{E}[\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i] \right), \tag{29}$$

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

The proof of Proposition 2, given in Appendix C.1, is similar to the derivation of the optimal denoisers for mixed regression in Tan and 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 (28)–(29) reduce to the optimal denoisers in Tan and 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 (see Appendix C.2 of Arpino et al. (2024b)). In Appendix C.2 we show how g^{*t} can be efficiently computed for the linear and rectified linear models with additive Gaussian noise, as well as for the logistic model. 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 $\boldsymbol{\Psi}$ in $[L]^n$, and without loss of generality can account for duplicate signals by increasing L. 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 1 states that $(\boldsymbol{\Theta}^t, \boldsymbol{y})$ converges in a specific sense to the random variable $(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))$, whose distribution crucially captures information about $\boldsymbol{\Psi}$. Hence, it is natural to consider point estimators for $\boldsymbol{\eta}$ of the form $\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y})$. In the linear model, for instance, one example is an estimator that searches for a signal configuration vector $\boldsymbol{\psi} \in \mathcal{X}$ such that $\boldsymbol{\Theta}^t$ indexed along $\boldsymbol{\psi}$ is closest in ℓ_2 distance to the observed vector $\boldsymbol{y}$. That is,

$$\hat{\boldsymbol{\Psi}}(\boldsymbol{\Theta}^t, \boldsymbol{y}) = \arg\min_{\boldsymbol{\psi} \in \mathcal{X}} \sum_{i=1}^{n} (y_i - (\boldsymbol{\Theta}^t)_{i, \psi_i})^2, \tag{30}$$

and $\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}) = U^{-1}(\hat{\boldsymbol{\Psi}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))$. In (35) we propose a general form for $\hat{\boldsymbol{\Psi}}$ that is suitable for a generic nonlinear model. A common metric for evaluating the accuracy of change point estimators is the Hausdorff distance (Wang and Samworth, 2017; Xu et al., 2024; Li et al., 2023). 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 *Consider the AMP in* (6). *Suppose the model assumptions in Section 2 as well as* $(A1) - (A3)$ *are satisfied. Let* $\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}) = \hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, q(\boldsymbol{XB}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))$ *be an estimator such that* $(\boldsymbol{V}, \boldsymbol{z}) \mapsto U(\hat{\boldsymbol{\eta}}(\boldsymbol{V}, q(\boldsymbol{z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})))$ *is uniformly pseudo-Lipschitz. Then:*

$$\frac{d_H(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))}{n} \stackrel{\mathbb{P}}{\simeq} \mathbb{E}_{\boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{Z}} \frac{d_H(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, U(\boldsymbol{\eta}), \boldsymbol{\varepsilon})))}{n}. \tag{31}$$

Moreover, if $(\boldsymbol{V}, \boldsymbol{z}) \mapsto |\hat{\boldsymbol{\eta}}(\boldsymbol{V}, q(\boldsymbol{z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))|$ *is uniformly pseudo-Lipschitz, then:*

$$|\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y})| \stackrel{\mathbb{P}}{\simeq} \mathbb{E}_{\boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{Z}} |\hat{\boldsymbol{\eta}}(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, U(\boldsymbol{\eta}), \boldsymbol{\varepsilon}))|. \tag{32}$$

The proof is given in Appendix D.1. To prove (31), we show that $d_H(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))/n$ is uniformly pseudo-Lipschitz. For each $\boldsymbol{\eta}$, Proposition 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 $\boldsymbol{V}_{\boldsymbol{\Theta}}^t = \boldsymbol{Z}\boldsymbol{\rho}^{-1}\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t + \boldsymbol{G}_{\boldsymbol{\Theta}}^t$ in (7), combined with an observation of the form $q(\boldsymbol{Z}, \boldsymbol{\Psi}, \bar{\boldsymbol{\varepsilon}})$, yields a recipe for constructing a posterior distribution over $\boldsymbol{\Psi}$. Recalling that $\boldsymbol{\Psi}$ is the unknown ground-truth signal configuration, we let $\bar{\boldsymbol{\Psi}}$ denote a random signal configuration drawn according to density $\pi_{\bar{\boldsymbol{\Psi}}}$, and $\boldsymbol{\psi}$ a realization of $\bar{\boldsymbol{\Psi}}$. Using the prior $\pi_{\bar{\boldsymbol{\Psi}}}$, the posterior is:

$$p_{\bar{\boldsymbol{\Psi}}|\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{\varepsilon}})}(\boldsymbol{\psi}|\boldsymbol{V}, \boldsymbol{u}) = \frac{\pi_{\bar{\boldsymbol{\Psi}}}(\boldsymbol{\psi}) \mathcal{L}(\boldsymbol{V}, \boldsymbol{u}|\boldsymbol{\psi})}{\sum_{\tilde{\boldsymbol{\psi}}} \pi_{\bar{\boldsymbol{\Psi}}}(\tilde{\boldsymbol{\psi}}) \mathcal{L}(\boldsymbol{V}, \boldsymbol{u}|\tilde{\boldsymbol{\psi}})}, \tag{33}$$

where $\boldsymbol{V} \in \mathbb{R}^{n \times L}$, $\boldsymbol{u} \in \mathbb{R}^n$. Here, $\mathcal{L}$ is the likelihood of $(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\varepsilon}))$ given $\bar{\boldsymbol{\Psi}} = \boldsymbol{\psi} \in \mathcal{X}$, where the state evolution parameters $\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t, \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^t$ associated to $\boldsymbol{V}_{\boldsymbol{\Theta}}^t$ are computed as in (10)–(13) with $\boldsymbol{\Psi}$ replaced by $\boldsymbol{\psi}$. The likelihood $\mathcal{L}$ can be computed in closed form for the linear and rectified linear models under the assumption of additive and independent Gaussian noise; for the logistic model, the likelihood can be well approximated. The computations are given in Appendix C.3.

Since Theorem 1 states that $(\boldsymbol{\Theta}^t, \boldsymbol{y})$ converges in a specific sense to $(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \varepsilon))$, we can obtain an uncertainty estimate over $\boldsymbol{\Psi}$ by plugging in $(\boldsymbol{\Theta}^t, \boldsymbol{y})$ for $(\boldsymbol{V}, \boldsymbol{u})$ in (33). It then follows from our theory that this uncertainty estimate converges point-wise in probability to a faithful posterior distribution over the change points in the high-dimensional limit.

Proposition 4 *Let $(\boldsymbol{\Theta}^t, \boldsymbol{y})$ be as in (6), $(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \varepsilon))$ be as in (7), and $\boldsymbol{\psi} \in \mathcal{X}$. Suppose the model assumptions in Section 2 as well as (A1)–(A3) are satisfied. Assume that $(\boldsymbol{V}, \boldsymbol{z}) \mapsto p_{\bar{\boldsymbol{\Psi}}|\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\varepsilon})}(\boldsymbol{\psi}|\boldsymbol{V}, q(\boldsymbol{z}, \boldsymbol{\Psi}, \varepsilon)) =: p(\boldsymbol{\psi}|\boldsymbol{V}, q(\boldsymbol{z}, \boldsymbol{\Psi}, \varepsilon))$ is uniformly pseudo-Lipschitz. Then:*

$$p(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y}) \overset{\mathbb{P}}{\simeq} p(\boldsymbol{\psi}|\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \varepsilon)). \tag{34}$$

Given $\boldsymbol{\psi}$ and ground-truth variables $\boldsymbol{\Psi}, \varepsilon$, the right-hand side of (34) can be computed by sampling $(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{Z})$, where the state evolution parameters $\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t, \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^t$ associated to $\boldsymbol{V}_{\boldsymbol{\Theta}}^t$ are computed exactly as in (10)–(13), and plugging these into $p(\boldsymbol{\psi}|\cdot, q(\cdot, \boldsymbol{\Psi}, \varepsilon))$. The proof, given in Appendix D.2, is an application of Theorem 1. The state evolution predictions on the RHS of (31) and (34) 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, with both synthetic and real data. For synthetic data, for $i \in [n]$, we use i.i.d. Gaussian covariates $\boldsymbol{X}_i \overset{i.i.d}{\sim} \mathsf{N}(0, \boldsymbol{I}_p/n)$, and we let $\varepsilon_i \overset{i.i.d}{\sim} \mathbb{P}_{\bar{\varepsilon}} = \mathsf{N}(0, \sigma^2)$ in the linear and rectified linear models. The denoisers $\{g^t, f^{t+1}\}_{t \geq 0}$ in the AMP algorithm are chosen according to Proposition 2. The computation of f^t is standard (outlined in Appendix C.2 of Arpino et al. (2024b)) and the computation of g^t is described in Appendix C.2. The Jacobians of these denoisers are computed using Automatic Differentiation in Python JAX (Bradbury et al., 2018). In all the experiments, 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 one standard deviation, and all experiments are the result of at least 10 independent trials with $t \leq 15$. Posterior densities are smoothed with a Gaussian kernel of lengthscale 0.5 for interpretability.

For a given point estimator $\hat{\boldsymbol{\Psi}}(\boldsymbol{\Theta}^t, \boldsymbol{y})$ of $\boldsymbol{\Psi}$, the iterate $\boldsymbol{\Theta}^t$ is obtained by running the AMP algorithm in (6) and we let $\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y}) := U^{-1}(\hat{\boldsymbol{\Psi}}(\boldsymbol{\Theta}^t, \boldsymbol{y}))$. As discussed in Section 3.3, the point estimation guarantees of Proposition 3 hold for a large class of sufficiently regular point estimators of $\boldsymbol{\Psi}$. For linear models we could use the point estimator in (30), and for nonlinear models, the mean of the posterior distribution $p_{\bar{\boldsymbol{\Psi}}|\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\varepsilon})}(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y})$ defined in (33). The latter, although justified by the limiting posterior distribution in Proposition 4, involves tabulating the state evolution parameters $(\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t, \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^t)$ from (10)–(13) (with $\boldsymbol{\Psi}$
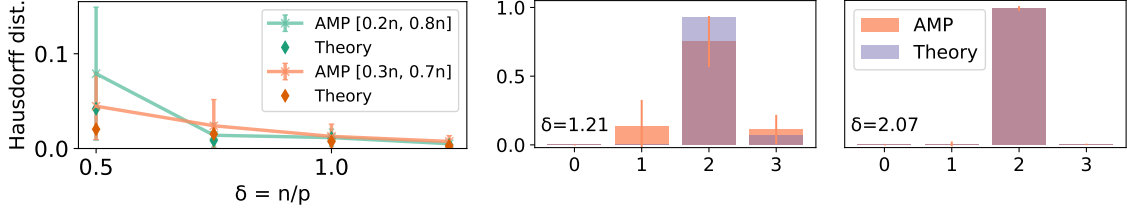
Figure 1: Performance of AMP for estimating change points in the linear model under two ground-truth configurations (left), approximate posterior density over the number of change points (two plots on the right).

replaced by $\boldsymbol{\psi}$), since these are required to compute the likelihood $\mathcal{L}$ of $(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{\varepsilon}}))$ given $\bar{\boldsymbol{\Psi}} = \boldsymbol{\psi}$. In the following experiments, we use an alternative point estimate of $\boldsymbol{\Psi}$ that is computationally faster, by considering instead the likelihood $\bar{\mathcal{L}}$ of $(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{\varepsilon}}))$ given $\bar{\boldsymbol{\Psi}} = \boldsymbol{\psi}$. We recall that $\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t$ is defined in (24) through the $\boldsymbol{\psi}$-independent parameters $(\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t, \bar{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^t)$. The point estimator of interest is then defined as follows:

$$\hat{\boldsymbol{\Psi}}(\boldsymbol{\Theta}^t, \boldsymbol{y}) = \underset{\boldsymbol{\psi} \in \mathcal{X}}{\arg\min}\, p_{\bar{\boldsymbol{\Psi}}|\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{\varepsilon}})}(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y}), \tag{35}$$

where we call $p_{\bar{\boldsymbol{\Psi}}|\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{\varepsilon}})}(\boldsymbol{\psi}|\boldsymbol{V}, \boldsymbol{u}) := \frac{\pi_{\bar{\boldsymbol{\Psi}}}(\boldsymbol{\psi}) \bar{\mathcal{L}}(\boldsymbol{V}, \boldsymbol{u}|\boldsymbol{\psi})}{\sum_{\tilde{\boldsymbol{\psi}}} \pi_{\bar{\boldsymbol{\Psi}}}(\tilde{\boldsymbol{\psi}}) \bar{\mathcal{L}}(\boldsymbol{V}, \boldsymbol{u}|\tilde{\boldsymbol{\psi}})}$ the *approximate posterior*. The explicit form of the likelihood $\bar{\mathcal{L}}$ for each of the linear, logistic, and rectified linear models is presented in Appendix C.3. Full implementation details are provided in Appendix E. A Python implementation of our algorithm and code to run the experiments is available at (Arpino et al., 2024a).

4.1 Linear Model with Change Points

Figure 1 (left) 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 closely matches the performance of AMP, verifying (31) in Proposition 3.

The two right-most plots in Figure 1 display the approximate posterior over the *number* of change points, that is, $\sum_{\boldsymbol{\psi} \in \mathcal{S}} p_{\bar{\boldsymbol{\Psi}}|\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{\varepsilon}})}(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y})$ where $\mathcal{S}$ contains all configurations with a specified number of change points. We use $p = 200$, $\mathbb{P}_{\bar{\boldsymbol{B}}} = N(\boldsymbol{0}, \boldsymbol{I})$, $\sigma = 0.1$, $\Delta = n/10$, $L^* = 3$, $w = 4$, and a uniform prior over the number of change points (zero to three). We observe that the approximate posterior concentrates around the ground truth for moderately large δ.

Figure 2 (left) plots the estimated posterior $p_{\bar{\boldsymbol{\Psi}}|\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{\varepsilon}})}(\cdot|\boldsymbol{\Theta}^t, \boldsymbol{y})$ labelled 'AMP', and the limiting *exact* posterior $p_{\bar{\boldsymbol{\Psi}}|\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{\varepsilon}})}(\cdot|\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))$ labelled 'Theory', averaged over 30 trials with true change point locations at $n/3$ and $8n/15$. The posterior is computed over the grid $[0, 1]$ subsampled by a factor of ten. The experiment uses $p = 800$, $\mathbb{P}_{\bar{\boldsymbol{B}}} =$
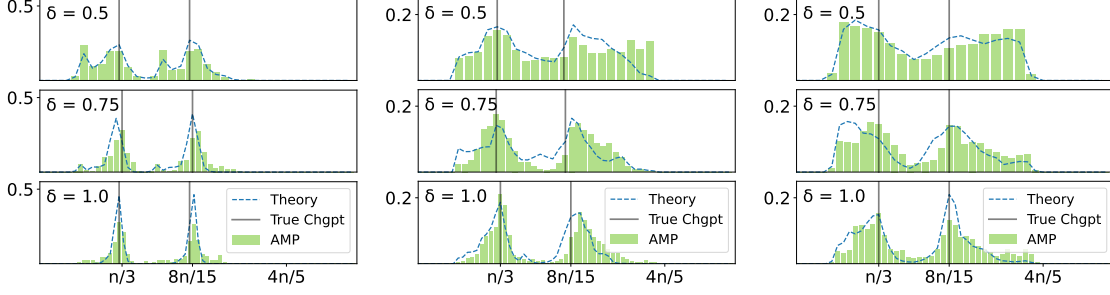
Figure 2: Estimated posterior $p(\cdot|\Theta^t, \boldsymbol{y})$ plotted against the limiting exact posterior $p(\cdot|\boldsymbol{V}_{\Theta}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))$ as per (34) for the linear (left), logistic (middle), rectified linear (right) models.

$\mathsf{N}(\boldsymbol{0}, \boldsymbol{I}), \sigma = 0.1, \Delta = n/5, L = L^* = 3$, with a uniform prior over all valid configurations. We observe that the estimated posterior closely matches the limiting posterior and concentrates around the ground truth as δ increases.

Figures 3a and 3b 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., 2024); divide and conquer dynamic programming (DCDP) (Li et al., 2023); and the complementary-sketching-based algorithm `charcoal` from Gao and Wang (2022). Hyperparameters are chosen using cross validation (CV), as outlined in Section A.1 of Li et al. (2023). 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 3a uses $p = 200, \sigma = 0.1, \Delta = n/10, L^* = L = 3$ and a sparse Bernoulli-Gaussian signal prior $\mathbb{P}_{\bar{\boldsymbol{B}}} = 0.5\mathsf{N}(\boldsymbol{0}, \delta\boldsymbol{I}) + 0.5\delta_{\boldsymbol{0}}$. 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 3b uses $p = 300, \Delta = n/10, L^* = L = 3$ and a Gaussian sparse difference prior with sparsity level 0.5 (described in (121)). 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 3a and 3b show that AMP consistently achieves the lowest Hausdorff distance among all algorithms and outperforms most algorithms in runtime. Figures 8 and 9 in Appendix E show results from an additional set of experiments comparing AMP with DCDP (the fastest algorithm in Figure 3a). Figure 8 shows the performance of AMP with different change point priors $\pi_{\bar{\boldsymbol{\Psi}}}$ and suboptimal denoising functions such as soft thresholding. Figure 9 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 4, we consider noiseless compressed sensing, where the signals $\{\boldsymbol{\beta}^{(i)}\}_{i \in [n]}$ are rotated versions of a $(255, 255)$ sparse
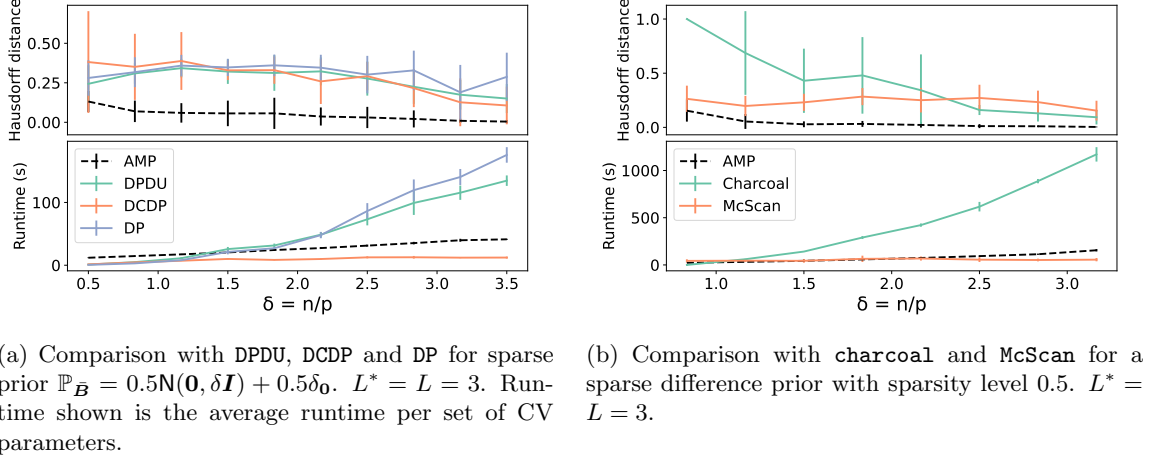
(a) Comparison with DPDU, DCDP and DP for sparse prior $\mathbb{P}_{\bar{\boldsymbol{B}}} = 0.5\mathsf{N}(\boldsymbol{0}, \delta\boldsymbol{I}) + 0.5\delta_{\boldsymbol{0}}$. $L^* = L = 3$. Runtime shown is the average runtime per set of CV parameters.

(b) Comparison with charcoal and McScan for a sparse difference prior with sparsity level 0.5. $L^* = L = 3$.

Figure 3: Linear model with sparse or sparse difference signal prior.
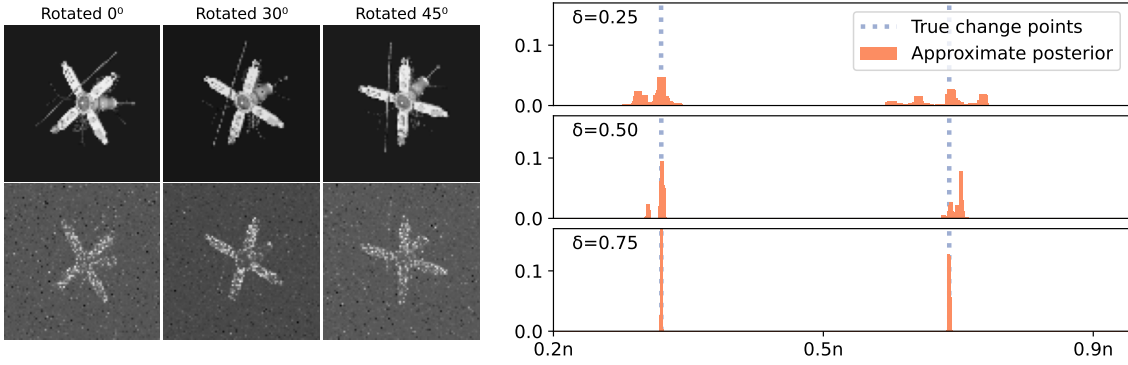
Figure 4: Top: Ground truth images (first row) and reconstruction from AMP (second row) at $\delta = 0.75$. Bottom: Approximate posterior vs. fractional change point locations for different δ.

grayscale image used by Schniter and 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 4 shows image reconstructions along with the approximate posterior $p_{\bar{\boldsymbol{\Psi}}|\bar{\boldsymbol{V}}_{\bar{\boldsymbol{\Theta}}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{\varepsilon}})}(\cdot|\boldsymbol{\Theta}^t, \boldsymbol{y})$. 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 3a ($p = 7225$ vs $p = 200$).

16

4.2 Rectified Linear Regression with Change Points

For the rectified linear model in (4) with a single change point at $(0.6n)$, Figure 5 (right) plots the Hausdorff distance divided by n as a function of δ. We let $p = 600$, $L = 3$, $\mathbb{P}_{\bar{B}} = \mathsf{N}(\mathbf{0}, \delta \boldsymbol{I})$, $\sigma = 0.3$, and average over ten trials. Similarly, Table 1 displays the performance of AMP in estimating the number of change points in the same setup, where we vary σ from 0.5 to 0.3. We observe a close match between the AMP performance and that dictated by theory in Proposition 3. Figure 2 (right) plots the estimated posterior $p_{\bar{\boldsymbol{\Psi}}|V_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\varepsilon})}(\cdot|\boldsymbol{\Theta}^t, \boldsymbol{y})$ against its the limiting posterior $p_{\bar{\boldsymbol{\Psi}}|V_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\varepsilon})}(\cdot|V_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \varepsilon))$, for the rectified linear model with $p = 600, \sigma = 0.5$, true change point locations at $n/3$ and $8n/15$, and $L = 3$ with a uniform prior over all configurations with two change points. We empirically observe a match between the estimated and limiting posteriors as per Proposition 4.

4.3 Logistic Regression with Change Points

For the logistic model in (3) with change points at $0.35n$ and $0.7n$, Figure 5 (left) plots the Hausdorff distance normalized by n for varying δ. We choose $p = 600$, $\Delta = n/5$, and $L = 3$ with a uniform prior over all valid configurations. We let $\mathbb{P}_{\bar{B}} = \mathsf{N}\left(\mathbf{0}, \begin{bmatrix} 15 & 0.2 & 0 \\ 0.2 & 15 & 0 \\ 0 & 0 & 15 \end{bmatrix}\right)$. We observe a close match between the AMP performance and that dictated by the state evolution theory in (31).

Table 1 displays AMP's estimated number of change points $|\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y})|$ in a variety of different settings. For the logistic model we set $\boldsymbol{\rho}$ to be $50\boldsymbol{I}$ and $75\boldsymbol{I}$, achieved by respectively setting $\mathbb{P}_{\bar{B}} = \mathsf{N}(\mathbf{0}, 50 \cdot \delta \boldsymbol{I})$ and $\mathbb{P}_{\bar{B}} = \mathsf{N}(\mathbf{0}, 75 \cdot \delta \boldsymbol{I})$ in our experiment. We let $p = 600$, $L = 3$, and let the true (single) change point location be at $0.6n$ ($L^* = 2$). We observe that the average number of change points predicted by AMP is close to the truth value of 1, and closely matches the theory from (32).

Figure 2 (middle) shows the estimated posterior $p(\cdot|\boldsymbol{\Theta}^t, \boldsymbol{y})$ from (34) and the limiting (exact) posterior $p(\cdot|V_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))$ in the logistic model with two true change points at $n/3$ and $8n/15$. We let $p = 600$, $\mathbb{P}_{\bar{B}} = \mathsf{N}(\mathbf{0}, 20\boldsymbol{I})$, and subsample the domain $[0, 1]$ by a factor of eight. We let $L = 3$ and place uniform prior over all possible two-changepoint configurations. We observe a close match with the theory in Proposition 4, and concentration around the true change point locations as δ increases.

Figure 6 compares AMP with the binary segmentation algorithm proposed in Wang et al. (2023) for sparse logistic regression. We use $p = 300, \Delta = 0.15n$, and a sparse Bernoulli-Gaussian signal prior $\mathbb{P}_{\bar{B}} = 0.5\mathsf{N}(\mathbf{0}, 15\delta \boldsymbol{I}) + 0.5\delta_{\mathbf{0}}$. The true change points are at $n/3$ and $8n/15$ with $L^* = 3$. The hyperparameters for binary segmentation are chosen as suggested in the source code of Wang et al. (2023). Both algorithms can take advantage of prior information on the number of change points. In Fig. 6a, both algorithms assume a maximum of $L = 3$ or 4 signals while $L^* = 3$. The algorithms are observed to achieve similar Hausdorff distance, with AMP enjoying a much lower runtime when assuming $L = 3$. Fig. 6b compares the two algorithms by assuming there are *exactly* 3 signals. The performance of AMP improves significantly with larger δ, both in terms of Hausdorff distance and runtime. We recall that the algorithm in Wang et al. (2023) is tailored to sparse logistic regression, whereas AMP can account for a broader range of signal priors.

	Logistic $\rho = 50I$			Logistic $\rho = 75I$			ReLU $\sigma = 0.50$			ReLU $\sigma = 0.30$		
	Theory	Mean	Std	Theory	Mean	Std	Theory	Mean	Std	Theory	Mean	Std
$\delta = 0.75$		–			–		1.10	1.10	0.70	1.00	1.10	0.30
$\delta = 1.25$	1.00	1.10	0.30	1.00	0.90	0.30	0.80	1.30	0.46	1.00	1.00	0.00
$\delta = 1.75$	1.00	1.30	0.46	1.00	1.20	0.40	1.00	1.20	0.60	1.00	1.00	0.00
$\delta = 2.25$	1.00	1.10	0.30	1.00	1.10	0.30		–			–	

Table 1: AMP performance when estimating the number of change points in the logistic and rectified linear (ReLU) models. The true (single) change point location is $0.6n$ ($L^* = 2$), AMP is run with $L = 3$.
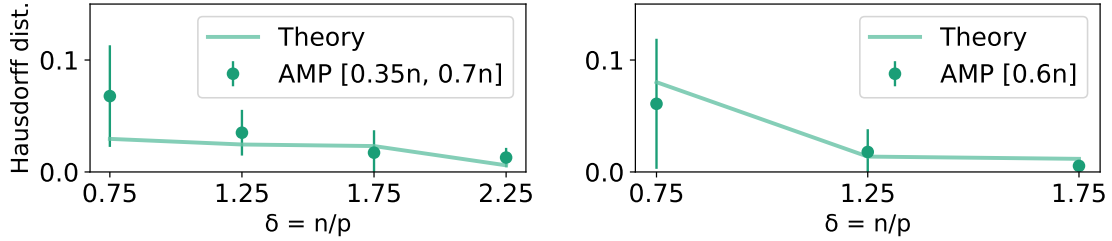
Figure 5: AMP performance in terms of the normalized Hausdorff distance plotted against the theoretical predictions of Proposition 3 for the logistic (left) and rectified linear (right) models.
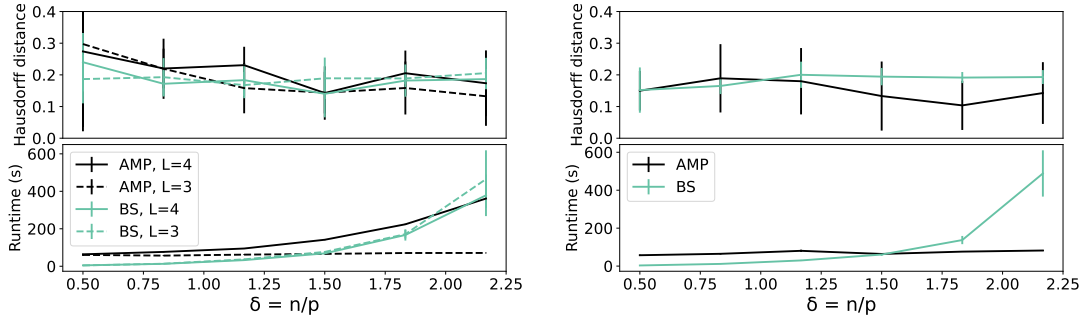
(a) Algorithms assume a maximum of 3 or 4 signals (that is, $L = 3$ or 4).

(b) Algorithms assume exactly 3 signals.

Figure 6: AMP vs. the binary segmentation (BS) algorithm in Wang et al. (2023) for logistic regression. The black plots correspond to AMP and the green plots correspond to BS. $\mathbb{P}_{\bar{B}} = 0.5\mathsf{N}(\mathbf{0}, 15\delta\mathbf{I}) + 0.5\delta_{\mathbf{0}}$, $L^* = 3$.

Inferring Change Points in Myocardial Infarction Data We consider the Myocardial Infarction (MI) complications data set from Golovenkin et al. (2020), which contains the medical information of $n = 1700$ patients (samples) with MI complications. Each sample has $p = 111$ medical features such as age, sex, heredity, and the presence of diabetes. The data set also contains 12 binary response variables for each patient relating to the state of the patient's overall heart health, indicating the presence of complications such as 'Atrial Fibration' and 'Chronic Heart Failure' (CHF). We investigate the relation between the binary CHF response variable and the features of each patient, using the logistic model. We sort the data by age in the range $26 - 92$ and exclude age as a feature, with the aim of determining whether the relationship between the response and the features changes markedly at a certain age. We whiten the age-ordered $n \times (p - 1)$ feature matrix, replacing missing values with interpolated estimates. We set $L = 2, \Delta = n/500$, and select the signal prior $\mathbb{P}_{\bar{B}} = \mathsf{N}\left(\mathbf{0}, \begin{bmatrix} 1 & 3/4 \\ 3/4 & 1 \end{bmatrix}\right)$ independently of the data. For $\pi_{\bar{\Psi}}$ we first specify a uniform prior over the number of change points (0 or 1 change points with probability $\frac{1}{2}$ each); we then define a uniform prior on the change point location for the case of 1 change point. We compute the estimated posterior $p(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y})$ from (34), reporting the existence of a change point with posterior probability close to 1. Assuming one change point, we plot the estimated posterior over all possible single change point locations in Figure 7 (left). The plot displays strong concentration of the posterior within the age range $55 - 75$, attaining its maximum value at age 66. This indicates the existence of a change point in the age range $55 - 75$ in how the presence of CHF relates to various medical features such as heredity, sex, and diabetes. This is consistent with findings in the medical literature, where older patients with the presence of CHF (aged > 75 years) were found to be more often female, have fewer cardiovascular diseases, and have fewer associated risk factors than patients with CHF aged 55 years or less (Stein et al., 2012; Azad and Lemay, 2014; Tromp et al., 2021).

We further validate our findings by running ℓ_2-regularised logistic regression separately on the data samples before and after the age of 66. The resulting estimated parameter vectors are named 'Estimated Vector 1' and 'Estimated Vector 2' in Figure 7 (right). We apply class weighting to the loss function to mitigate the effect of class imbalance, and use 5-fold cross-validation to select the regularisation constant. We observe from the plot in Figure 7 (right) that the found regression vectors differ significantly, with a correlation coefficient of 0.09, yielding further evidence for a possible change in the underlying data generating mechanism before and after age 66. Moreover, we observe that the squared sum of the coefficients of 'Estimated Vector 2' (1.9) is significantly smaller than that of 'Estimated Vector 1' (5.9), a finding that is consistent with the observation that traditional risk factors provide less combined evidence of heart failure risk in elderly patients (Tromp et al., 2021).

5. Discussion

In this work we proposed an AMP algorithm for regression with change points. The algorithm can efficiently incorporate prior information on the change points, and its state evolution characterization gives rigorous asymptotic guarantees on estimation accuracy. It
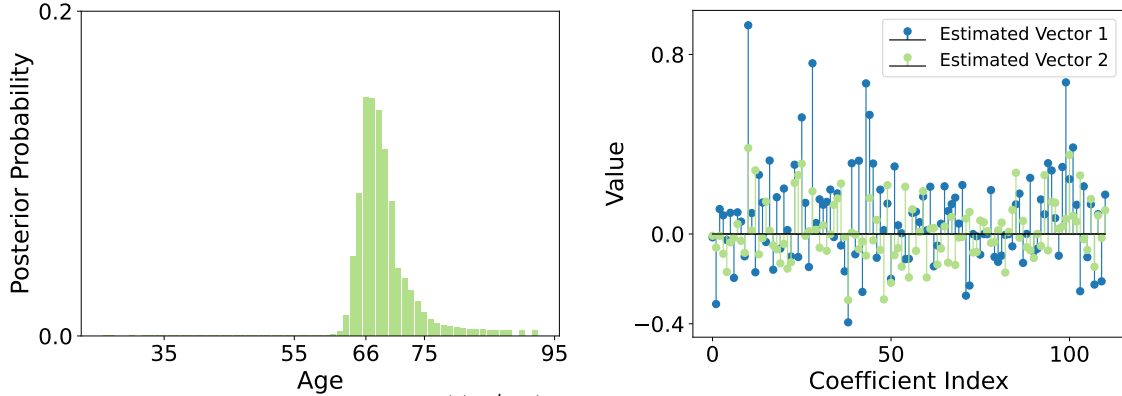
Figure 7: Change point posterior $p(\cdot|\boldsymbol{\Theta}^t, \boldsymbol{y})$ with respect to age in the Myocardial Infarction data set (left), coefficients of both regression vectors estimated using samples before and after the age of 66 (right).

also allows to quantify uncertainty in the change point estimates via a computable posterior distribution.

In principle, the choice of change point prior could significantly affect the performance of our proposed algorithm—for example, the performance would be poor if we choose a prior that assigns very small probability mass to the true change-point locations. However, our experimental results indicate that an uninformative prior (uniform on both the number and locations of the change points) consistently gives good performance across a range of settings. The algorithm also uses priors on the signal $\boldsymbol{B}$ and the measurement noise $\boldsymbol{\varepsilon}$ to tailor the AMP denoising functions f^t, g^t. A misspecified prior on $\boldsymbol{\varepsilon}$ or $\boldsymbol{B}$ would lead to choices on f^t, g^t that are suboptimal, but we emphasize the algorithm and the theoretical guarantees remain valid: the performance of the algorithm is characterized by state evolution parameters, which in turn depend on the denoising functions used within AMP. Figures 3, 4, 6, 7, 8 illustrate the robustness of our algorithm to lack of exact knowledge about signal or noise priors.

A limitation of our method is that the AMP algorithm and its theoretical guarantees are derived assuming that the covariates are i.i.d. Gaussian. We discuss below how this assumption can be relaxed in certain settings, along with other directions for future work.

Non-Gaussian Design Based on recent universality results for AMP (Wang et al., 2024), we expect that our algorithm and theoretical guarantees will apply to a broad class of i.i.d. designs. We also expect that our approach can be generalized to the much broader class of rotationally invariant designs, for which AMP-like algorithms have been proposed for regression without change points (Ma and Ping, 2017; Rangan et al., 2019; Takeuchi, 2020; Pandit et al., 2020).

Gaussian Design with General Covariance In the setting where $\boldsymbol{X}_i \overset{i.i.d}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ for $i \in [n]$ and for some covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, we can write $\boldsymbol{X}_i = \boldsymbol{\Sigma}^{1/2} \tilde{\boldsymbol{X}}_i$, where $\tilde{\boldsymbol{X}}_i \overset{i.i.d}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ for $i \in [n]$. The model (1) can then be equivalently written as:

$$y_i = q\big((\tilde{\boldsymbol{X}}_i)^\top (\boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}^{(i)}), \, \varepsilon_i\big), \quad i = 1, \dots, n,$$

where $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n \times p}$ is now our effective design matrix composed of stacked rows $(\tilde{\boldsymbol{X}}_i)^{\top}$, and $\boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}^{(i)}$ is our effective signal. If $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ is known or can be approximated with reasonable accuracy, the user can run the AMP iteration (6) with $\tilde{\boldsymbol{X}}$ in place of $\boldsymbol{X}$. Since the entries of $\tilde{\boldsymbol{X}}$ are i.i.d. Gaussian, we can apply Theorem 1 to this iteration to obtain guarantees on the accuracy of change point estimation (via Proposition 3) and guarantees on the recovery of $\boldsymbol{\Sigma}^{1/2} \boldsymbol{B}$ (via Theorem 1). Because $\boldsymbol{\Sigma}$ is known, this also gives us guarantees on the recovery of $\boldsymbol{B}$.

The more realistic setting of unknown $\boldsymbol{\Sigma}$ is challenging, and developing effective AMP-like algorithms for this case is an interesting open question. One idea is to use weighted versions of standard convex estimators for high-dimensional regression such as LASSO or M-estimators, where the weights can be chosen to encode prior information on change point locations. By leveraging the connection between AMP and convex optimization (Section 4.4 of Feng et al. 2022), it is possible to analyze such convex estimators under unknown covariance $\boldsymbol{\Sigma}$, as in (Zhao et al., 2022).

Design with Temporal Dependence   Xu et al. (2024) recently studied the problem of change point detection in linear regression with temporally dependent data, where the rows of the design matrix $\boldsymbol{X}$ form a high-dimensional time-series. An AMP framework has recently been proposed for linear regression when the data are generated from a high-dimensional time series (Tieplova et al., 2025). Using this framework to generalize our change point AMP to temporal dependent data is a promising direction for future work.

Acknowledgments

Appendix A. Proof of State Evolution

The proof of Theorem 1 relies on a generalization of Lemma 14 in Gerbelot and Berthier (2023), presented as Lemma 5 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 and 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 5 generalizes Lemma 14 of Gerbelot and 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 $\boldsymbol{X}^0 \in \mathbb{R}^{N \times Q}$ and $\boldsymbol{M}^{-1} := \boldsymbol{0} \in \mathbb{R}^{N \times Q}$:

$$\boldsymbol{X}^{t+1} = \boldsymbol{A}\boldsymbol{M}^t - \boldsymbol{M}^{t-1}(\boldsymbol{b}^t)^\top \qquad \in \mathbb{R}^{N \times Q}, \qquad (36)$$

$$\boldsymbol{M}^t = \tilde{f}^t(\varphi(\boldsymbol{A}\boldsymbol{W}_0, \boldsymbol{\Xi}), \boldsymbol{X}^t) \qquad \in \mathbb{R}^{N \times Q}, \qquad (37)$$

$$\boldsymbol{b}^t = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \tilde{f}_i^t}{\partial \boldsymbol{X}_i^t}(\varphi(\boldsymbol{A}\boldsymbol{W}_0, \boldsymbol{\Xi}), \boldsymbol{X}^t) \qquad \in \mathbb{R}^{Q \times Q}. \qquad (38)$$

Our result in Lemma 5 presents an asymptotic characterization of (36)–(38) via the following state evolution recursion:

$$\boldsymbol{\nu}^0, \hat{\boldsymbol{\nu}}^0 = \boldsymbol{0}, \qquad (39)$$

$$\boldsymbol{\kappa}^{0,0} = \lim_{N \to \infty} \frac{1}{N} \tilde{f}^0(\boldsymbol{X}^0)^\top \tilde{f}^0(\boldsymbol{X}^0), \qquad (40)$$

$$\boldsymbol{\nu}^{t+1} = \lim_{N \to \infty} \frac{1}{N} \mathbb{E}\left[\boldsymbol{W}_0^\top \tilde{f}^t\left(\varphi(\boldsymbol{Z}_{\boldsymbol{W}_0}, \boldsymbol{\Xi}), \boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1}\boldsymbol{\nu}^t + \boldsymbol{W}_0\hat{\boldsymbol{\nu}}^t + \boldsymbol{Z}^t\right)\right], \qquad (41)$$

$$\hat{\boldsymbol{\nu}}^{t+1} = \lim_{N \to \infty} \frac{1}{N} \mathbb{E}\left[\sum_{i=1}^{N} \partial_{1i}\bar{f}_i^t\left(\boldsymbol{Z}_{\boldsymbol{W}_0}, \boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1}\boldsymbol{\nu}^t + \boldsymbol{W}_0\hat{\boldsymbol{\nu}}^t + \boldsymbol{Z}^t, \boldsymbol{\Xi}\right)\right], \qquad (42)$$

$$\boldsymbol{\kappa}^{t+1,s+1} = \boldsymbol{\kappa}^{s+1,t+1}$$

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

$$\left(\tilde{f}^t\left(\varphi(\boldsymbol{Z}_{\boldsymbol{W}_0}, \boldsymbol{\Xi}), \boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1}\boldsymbol{\nu}^t + \boldsymbol{W}_0\hat{\boldsymbol{\nu}}^t + \boldsymbol{Z}^t\right) - \boldsymbol{W}_0\boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1}\boldsymbol{\nu}^{t+1}\right)\right], \qquad (43)$$

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 $(\boldsymbol{Z}_{\boldsymbol{W}_0})_i \overset{i.i.d}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{\rho}_{\boldsymbol{W}_0})$. For $i \in [n], 0 \le s, r \le t$, we have that $(\boldsymbol{Z}_{\boldsymbol{W}_0})_i$ is independent from $(\boldsymbol{Z}^s)_i \overset{i.i.d}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{\kappa}^{s,s})$ with $\mathrm{Cov}((\boldsymbol{Z}^s)_i, (\boldsymbol{Z}^r)_i) = \boldsymbol{\kappa}^{s,r}$. In (42), 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, that is, $\boldsymbol{A} = \mathbf{G} + \mathbf{G}^\top$ 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 5 *Consider the AMP iteration* (36)–(38) *and the state evolution recursion* (39)–(43). *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* $(\mathbf{X}^1, \ldots, \mathbf{X}^t, \boldsymbol{V}) \mapsto \Phi_N(\mathbf{X}^1, ..., \mathbf{X}^t, \boldsymbol{V}; \boldsymbol{W}_0, \boldsymbol{\Xi})$ *is uniformly pseudo-Lipschitz, we have that:*

$$\Phi_N \left(\boldsymbol{X}^0, \boldsymbol{X}^1, ..., \boldsymbol{X}^t, \boldsymbol{A} \boldsymbol{W}_0; \boldsymbol{W}_0, \boldsymbol{\Xi} \right) \overset{\mathbb{P}}{\simeq}$$
$$\mathbb{E}_{\boldsymbol{Z}^1, ..., \boldsymbol{Z}^t, \boldsymbol{Z}_{\boldsymbol{W}_0}} \left[\Phi_N \left(\boldsymbol{X}^0, \boldsymbol{Z}_{\boldsymbol{W}_0} \boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1} \boldsymbol{\nu}^1 + \boldsymbol{W}_0 \hat{\boldsymbol{\nu}}^t + \boldsymbol{Z}^1, \ldots, \right. \right.$$
$$\left. \left. \boldsymbol{Z}_{\boldsymbol{W}_0} \boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1} \boldsymbol{\nu}^t + \boldsymbol{W}_0 \hat{\boldsymbol{\nu}}^t + \boldsymbol{Z}^t, \boldsymbol{Z}_{\boldsymbol{W}_0}; \boldsymbol{W}_0, \boldsymbol{\Xi} \right) \right]. \quad (44)$$

Proof The main differences between the AMP result in Gerbelot and 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 $\boldsymbol{A} \boldsymbol{W}_0$, and on the fixed parameters $\boldsymbol{W}_0$ and $\boldsymbol{\Xi}$. Assumption (B2) guarantees that $\tilde{f}^t$ maintains the same required convergence properties despite modification 1).

We now address modification 2). Since $\boldsymbol{\Xi}$ and $\boldsymbol{W}_0$ are fixed and $(\boldsymbol{X}^0, \ldots, \boldsymbol{X}^t, \boldsymbol{V}) \mapsto \Phi_N(\mathbf{X}^0, \mathbf{X}^1, ..., \mathbf{X}^t, \boldsymbol{V}; \boldsymbol{W}_0, \boldsymbol{\Xi})$ is assumed to be uniformly pseudo-Lipschitz, these can be

included in Φ_N and the right hand side of (44) is unaffected. We now present the argument for why $\boldsymbol{AW}_0$ can be included as in input to Φ_N on the LHS of (44), and yield a corresponding input $\boldsymbol{Z}_{\boldsymbol{W}_0}$ on the RHS of (44). Consider the following "augmented" AMP iteration, initialized with $\mathbf{X}^{aug,0} := \begin{bmatrix} \boldsymbol{X}^0 & \boldsymbol{X}^0 \end{bmatrix} \in \mathbb{R}^{N \times 2Q}$ and $\boldsymbol{M}^{aug,-1} := \boldsymbol{0} \in \mathbb{R}^{N \times 2Q}$:

$$\boldsymbol{X}^{aug,t+1} = \boldsymbol{A}\boldsymbol{M}^{aug,t} - \boldsymbol{M}^{aug,t-1}(\boldsymbol{b}^{aug,t})^\top \qquad \in \mathbb{R}^{N \times 2Q}, \qquad (45)$$

$$\boldsymbol{M}^{aug,t} = \tilde{f}^{aug,t}(\varphi^{aug}\left(\boldsymbol{AW}_0^{aug}, \boldsymbol{\Xi}\right), \boldsymbol{X}^{aug,t}) \qquad \in \mathbb{R}^{N \times 2Q}, \qquad (46)$$

$$\boldsymbol{b}^{aug,t} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \tilde{f}_i^{aug,t}}{\partial \boldsymbol{X}_i^{aug,t}}(\varphi^{aug}\left(\boldsymbol{AW}_0^{aug}, \boldsymbol{\Xi}\right), \boldsymbol{X}^{aug,t}) \qquad \in \mathbb{R}^{2Q \times 2Q}. \qquad (47)$$

In the above iteration, we define $\boldsymbol{W}_0^{aug} := \begin{bmatrix} \boldsymbol{W}_0 & \boldsymbol{W}_0 \end{bmatrix} \in \mathbb{R}^{N \times 2Q}, \varphi^{aug}\left(\boldsymbol{AW}_0^{aug}, \boldsymbol{\Xi}\right) := \begin{bmatrix} \varphi\left(\boldsymbol{AW}_0, \boldsymbol{\Xi}\right) & \varphi\left(\boldsymbol{AW}_0, \boldsymbol{\Xi}\right) \end{bmatrix} \in \mathbb{R}^{N \times 2Q}$. Letting $\boldsymbol{X}^{t+1}, \tilde{\boldsymbol{X}}^{t+1} \in \mathbb{R}^{N \times Q}$ take on the definition

$$\begin{bmatrix} \boldsymbol{X}^{t+1} & \tilde{\boldsymbol{X}}^{t+1} \end{bmatrix} := \boldsymbol{X}^{aug,t+1}$$

and defining:

$$\tilde{f}^{aug,t}(\varphi^{aug}\left(\boldsymbol{AW}_0^{aug}, \boldsymbol{\Xi}\right), \boldsymbol{X}^{aug,t}) := \begin{bmatrix} \tilde{f}^t(\varphi\left(\boldsymbol{AW}_0, \boldsymbol{\Xi}\right), \boldsymbol{X}^t) & \xi^t(\tilde{\boldsymbol{X}}^t) \end{bmatrix}$$

for some uniformly pseudo-Lipschitz function $\xi^t(\tilde{\boldsymbol{X}}^{t+1})$ with $\xi^0(\tilde{\boldsymbol{X}}^0) := \boldsymbol{W}_0$, we obtain that:

$$\boldsymbol{M}^{aug,t} =: \begin{bmatrix} \boldsymbol{M}^t & \tilde{\boldsymbol{M}}^t \end{bmatrix} \qquad (48)$$

$$\boldsymbol{b}^{aug,t} = \begin{bmatrix} \boldsymbol{b}^t & \boldsymbol{0} \\ \boldsymbol{0} & * \end{bmatrix}, \qquad (49)$$

where $\boldsymbol{M}^t$ and $\boldsymbol{b}^t$ are quantities from the initial iteration (36)–(38), and $*$ denotes an unspecified matrix in $\mathbb{R}^{Q \times Q}$. We therefore have that the first Q columns of the matrix iteration (45)–(47) are equivalent to the original iteration (36)–(38). The last Q columns of the augmented iteration will be used to incorporate $\boldsymbol{AW}_0$.

For $t \geq 1$, let $(\boldsymbol{X}^{aug,t})_{[:,:Q]}, (\boldsymbol{X}^{aug,t})_{[:,Q:]}$ denote the $N \times Q$ matrices composed of the first Q columns and of the last Q columns of $\boldsymbol{X}^{aug,t}$, respectively. Let $\boldsymbol{\rho} := \frac{1}{N}\boldsymbol{W}_0^\top \boldsymbol{W}_0$, and for $t \geq 1$ define the following random variables, independent of all else: $(\boldsymbol{Z}_{\boldsymbol{W}_0})_i \stackrel{i.i.d}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{\rho})$ and $(\boldsymbol{Z}^t)_i \stackrel{i.i.d}{\sim} \mathsf{N}(0, \boldsymbol{\kappa}^{t,t})$ for $i \in [n]$. We compute the state evolution equations associated with the augmented AMP iteration for $t = 1$, outlined in $(C.25) - (C.28)$ of Gerbelot and

24

Berthier (2023), to obtain:

$$
\boldsymbol{\nu}^{aug,1} = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\begin{bmatrix} \boldsymbol{W}_0^\top \\ \boldsymbol{W}_0^\top \end{bmatrix} \begin{bmatrix} \tilde{f}^0(\boldsymbol{X}^0) & \xi^0(\tilde{\boldsymbol{X}}^0) \end{bmatrix} \right] = \begin{bmatrix} \boldsymbol{\nu}^1 & \boldsymbol{\rho} \\ \boldsymbol{\nu}^1 & \boldsymbol{\rho} \end{bmatrix},
$$

$$
\hat{\boldsymbol{\nu}}^{aug,1} = \begin{bmatrix} \hat{\boldsymbol{\nu}}^1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},
$$

$$
\boldsymbol{\kappa}^{aug,1,1} = \lim_{N\to\infty} \frac{1}{N}
$$

$$
\mathbb{E}\left[\left(\begin{bmatrix} \tilde{f}^0(\varphi(\boldsymbol{Z}_{\boldsymbol{W}_0}, \boldsymbol{\Xi}), \boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}^{-1}\boldsymbol{\nu}^1 + \boldsymbol{W}_0\hat{\boldsymbol{\nu}}^1 + \boldsymbol{Z}^1) \\ \boldsymbol{W}_0 \end{bmatrix} - \begin{bmatrix} \boldsymbol{W}_0 \\ \boldsymbol{W}_0 \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\rho}^{-1}\boldsymbol{\nu}^1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{Q\times Q} \end{bmatrix} \right)^\top \right.
$$

$$
\left. \left(\begin{bmatrix} \tilde{f}^0(\varphi(\boldsymbol{Z}_{\boldsymbol{W}_0}, \boldsymbol{\Xi}), \boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}^{-1}\boldsymbol{\nu}^1 + \boldsymbol{W}_0\hat{\boldsymbol{\nu}}^1 + \boldsymbol{Z}^1) \\ \boldsymbol{W}_0 \end{bmatrix} - \begin{bmatrix} \boldsymbol{W}_0 \\ \boldsymbol{W}_0 \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\rho}^{-1}\boldsymbol{\nu}^1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{Q\times Q} \end{bmatrix} \right) \right]
$$

$$
= \begin{bmatrix} \boldsymbol{\kappa}^{1,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.
$$

Applying Lemma 14 in Gerbelot and Berthier (2023) to the pseudo-Lipschitz function

$$
\tilde{\Phi}_N\left(\boldsymbol{X}^{aug,1}, \ldots, \boldsymbol{X}^{aug,t}\right) := \Phi_N\left((\boldsymbol{X}^{aug,1})_{[:,:Q]}, \ldots, (\boldsymbol{X}^{aug,t})_{[:,:Q]}, (\boldsymbol{X}^{aug,1})_{[:,Q:]}\right),
$$

we obtain:

$$
\Phi_N\left(\boldsymbol{X}^1, \ldots, \boldsymbol{X}^t, \boldsymbol{A}\boldsymbol{W}_0\right)
$$

$$
\Phi_N\left((\boldsymbol{X}^{aug,1})_{[:,:Q]}, \ldots, (\boldsymbol{X}^{aug,t})_{[:,:Q]}, (\boldsymbol{X}^{aug,1})_{[:,Q:]}\right)
$$

$$
= \tilde{\Phi}_N\left(\boldsymbol{X}^{aug,1}, \ldots, \boldsymbol{X}^{aug,t}\right)
$$

$$
\stackrel{\mathbb{P}}{\simeq} \mathbb{E}\tilde{\Phi}_N\left(\begin{bmatrix} \boldsymbol{Z}_{\boldsymbol{W}_0} & \boldsymbol{Z}_{\boldsymbol{W}_0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\rho}^{-1}\boldsymbol{\nu}^1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{Q\times Q} \end{bmatrix} + \begin{bmatrix} \boldsymbol{W}_0 & \boldsymbol{W}_0 \end{bmatrix} \hat{\boldsymbol{\nu}}^{aug,1} + \begin{bmatrix} \boldsymbol{Z}^1 & \mathbf{0} \end{bmatrix}, \right.
$$

$$
\left. \ldots, \begin{bmatrix} \boldsymbol{Z}_{\boldsymbol{W}_0} & \boldsymbol{Z}_{\boldsymbol{W}_0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\rho}^{-1}\boldsymbol{\nu}^t & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{Q\times Q} \end{bmatrix} + \begin{bmatrix} \boldsymbol{W}_0 & \boldsymbol{W}_0 \end{bmatrix} \hat{\boldsymbol{\nu}}^{aug,t} + \begin{bmatrix} \boldsymbol{Z}^t & * \end{bmatrix} \right)
$$

$$
= \mathbb{E}\tilde{\Phi}_N\left(\begin{bmatrix} \boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}^{-1}\boldsymbol{\nu}^1 + \boldsymbol{W}_0\hat{\boldsymbol{\nu}}^1 + \boldsymbol{Z}^1 & \boldsymbol{Z}_{\boldsymbol{W}_0} \end{bmatrix}, \ldots, \right.
$$

$$
\left. \begin{bmatrix} \boldsymbol{Z}_{\boldsymbol{W}_0} & \boldsymbol{Z}_{\boldsymbol{W}_0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\rho}^{-1}\boldsymbol{\nu}^t & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{Q\times Q} \end{bmatrix} + \begin{bmatrix} \boldsymbol{W}_0 & \boldsymbol{W}_0 \end{bmatrix} \hat{\boldsymbol{\nu}}^{aug,t} + \begin{bmatrix} \boldsymbol{Z}^t & * \end{bmatrix} \right)
$$

$$
= \mathbb{E}\Phi_N\left(\boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}^{-1}\boldsymbol{\nu}^1 + \boldsymbol{W}_0\hat{\boldsymbol{\nu}}^1 + \boldsymbol{Z}^1, \cdots, \boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}^{-1}\boldsymbol{\nu}^t + \boldsymbol{W}_0\hat{\boldsymbol{\nu}}^t + \boldsymbol{Z}^t, \boldsymbol{Z}_{\boldsymbol{W}_0} \right),
$$

and the result follows. ∎

We next present the main reduction, mapping the AMP algorithm proposed in this work (6) to the symmetric one outlined in (36)–(38).

Proof [Proof of Theorem 1] Consider the change point linear regression model (1), and recall that it can be rewritten as (5). We reduce the algorithm (6) to (36)–(38), following the alternating technique of Javanmard and 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 (36)–(38) 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 (6) 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

$$
\boldsymbol{A} = \sqrt{\frac{\delta}{\delta+1}} \begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{X} \\ \boldsymbol{X}^\top & \boldsymbol{D}_2 \end{bmatrix} \sim \mathrm{GOE}(N), \quad \boldsymbol{W}_0 = \begin{bmatrix} \boldsymbol{0}_{n \times L} \\ \boldsymbol{B} \end{bmatrix} \in \mathbb{R}^{N \times L}, \quad \boldsymbol{X}^0 = \begin{bmatrix} \boldsymbol{0}_{n \times L} \\ \boldsymbol{B}^0 \end{bmatrix} \quad (50)
$$

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

$$
\varphi : (\boldsymbol{W}, \boldsymbol{\Xi}) \mapsto q\left(\sqrt{\frac{\delta+1}{\delta}}\, \boldsymbol{W}_{[n]}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}\right) \in \mathbb{R}^N, \quad (51)
$$

for any $\boldsymbol{W} \in \mathbb{R}^{N \times L}$. We therefore have $\varphi(\boldsymbol{A}\boldsymbol{W}_0, \boldsymbol{\Xi}) = q(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})$. Let

$$
\tilde{f}^t : \mathbb{R}^{N \times L} \times \mathbb{R}^N \to \mathbb{R}^{N \times L}
$$

such that

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

$$
\tilde{f}^{2t}(\varphi(\boldsymbol{A}\boldsymbol{W}_0, \boldsymbol{\Xi}), \boldsymbol{U}) = \sqrt{\frac{\delta+1}{\delta}} \begin{bmatrix} \boldsymbol{0}_{n \times L} \\ f^t(\boldsymbol{U}_{[-p]}) \end{bmatrix},
$$

$$(52)$$

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

Next, consider the AMP iteration (36)–(38) with $\boldsymbol{A}, \boldsymbol{W}_0, \boldsymbol{X}^0, \boldsymbol{\Xi}, \varphi, \tilde{f}^t$ defined as in (50)–(52). Note that the assumptions $(B1) - (B6)$ are satisfied by construction, and hence the state evolution result in Lemma 5 holds for the iteration (36)–(38). We will now show that the state evolution equations decompose into those of interest, (10)–(13). First, note that

$$\boldsymbol{X}^{2t+1}$$

$$
= \begin{bmatrix} \boldsymbol{X} f^t(\boldsymbol{X}^{2t}_{[-p]}) - g^{t-1}(\boldsymbol{X}^{2t-1}_{[n]}, q(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{\Xi})) \cdot \overbrace{\frac{1}{n}\left(\sum_{i=1}^{p} \frac{\partial f^t(\boldsymbol{X}^{2t}_{[-p]})_i}{\partial \boldsymbol{X}^{2t}_{[-p],i}}\right)^\top}^{=(\boldsymbol{F}^t)^\top} \\ \boldsymbol{D}_2 f^t(\boldsymbol{X}^{2t}_{[-p]}) \end{bmatrix}, \quad (53)
$$

$$\boldsymbol{X}^{2t}$$

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

$$(54)$$

and hence observe that $\boldsymbol{X}_{[n]}^{2t+1}$ and $\boldsymbol{X}_{[-p]}^{2t}$ are equal to $\boldsymbol{\Theta}^t$ and $\boldsymbol{B}^t$ in (6) respectively. Define the main iterate $\boldsymbol{Q}^t = \boldsymbol{Z}_{\boldsymbol{W}_0}\boldsymbol{\rho}_{\boldsymbol{W}_0}^{-1}\boldsymbol{\nu}^t + \boldsymbol{W}_0\hat{\boldsymbol{\nu}}^t + \boldsymbol{Z}^t \in \mathbb{R}^{N\times L}$. For $i \in [n]$, let $\tilde{g}_i^t : (\boldsymbol{Z}, \boldsymbol{V}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) \mapsto g_i^t(\boldsymbol{V}, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))$ and let $\partial_{1i}\tilde{g}_i^t$ be the partial derivative (Jacobian) w.r.t. the ith row of the first argument. Following (39)–(43), we then have that

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

(55)

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

(56)

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

(57)

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

(58)

$$\boldsymbol{\kappa}^{2t,2s}$$

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

$$\left. \cdot g^{s-1}\left(\boldsymbol{Q}_{[n]}^{2s-1}, q\left(\sqrt{\frac{\delta+1}{\delta}}\boldsymbol{Z}_{\boldsymbol{W}_0,[n]}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}\right)\right)\right]$$

(59)

$$\boldsymbol{\kappa}^{2t+1,2s+1}$$

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

(60)

where in (60) 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

$$\boldsymbol{\nu}_{\boldsymbol{\Theta}}^t = \sqrt{\frac{\delta+1}{\delta}}\boldsymbol{\nu}^{2t+1}, \qquad \boldsymbol{\nu}_{\boldsymbol{B}}^t = \hat{\boldsymbol{\nu}}^{2t}, \qquad \boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t,s} = \boldsymbol{\kappa}^{2t+1,2s+1}, \quad \boldsymbol{\kappa}_{\boldsymbol{B}}^{t,s} = \boldsymbol{\kappa}^{2t,2s}, \qquad (61)$$

$$\boldsymbol{Z} \overset{d}{=} \sqrt{\frac{\delta+1}{\delta}}\boldsymbol{Z}_{\boldsymbol{W}_0,[n]}, \qquad \boldsymbol{G}_{\boldsymbol{\Theta}}^t \overset{d}{=} \boldsymbol{Z}_{[n]}^{2t+1}, \quad \boldsymbol{G}_{\boldsymbol{B}}^t \overset{d}{=} \boldsymbol{Z}_{[-p]}^{2t}, \qquad (62)$$

$$\boldsymbol{V}_{\boldsymbol{\Theta}}^t \overset{d}{=} \boldsymbol{Q}_{[n]}^{2t+1}, \qquad \boldsymbol{V}_{\boldsymbol{B}}^t \overset{d}{=} \boldsymbol{Q}_{[-p]}^{2t}. \qquad (63)$$

Substituting the change of variables (61)–(63) into (55)–(60), we obtain (10)–(13). Moreover, substituting (53)–(54) and (63) into Lemma 5 yields Theorem 1. ∎

Appendix B. State Evolution Limits and Simplifications

B.1 Parametric Dependence of State Evolution on Signal and Noise

In this section, we give a set of sufficient conditions for the existence of the limits in the state evolution equations (10)–(13), which also allow for removing the parametric dependence of the state evolution on $\boldsymbol{B}, \boldsymbol{\varepsilon}$. Assume (S0) on p.9, 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 heterogeneous dependence structure induced by change points, we require g_i^t to depend on i, for $i \in [n]$. However, Proposition 2 shows that it can be chosen to act row-wise, that is, $g_i^{*t}(\boldsymbol{\Theta}^t, \boldsymbol{y}) = g_i^{*t}(\boldsymbol{\Theta}_i^t, y_i)$. 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 (10)–(13) reduce to:

$$\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1} = \sum_{\ell=0}^{L-1} \mathbb{E}\left[\check{g}_{\eta_\ell}^t(\boldsymbol{Z}_1, (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, \Psi_{\eta_\ell}, \bar{\varepsilon})\right], \tag{64}$$

$$\boldsymbol{\kappa}_{\boldsymbol{B}}^{s+1,t+1} = \sum_{\ell=0}^{L-1} \mathbb{E}\left[\hat{g}_{\eta_\ell}^s((\boldsymbol{V}_{\boldsymbol{\Theta}}^s)_1, q_1(\boldsymbol{Z}_1, \Psi_{\eta_\ell}, \bar{\varepsilon}))\, \hat{g}_{\eta_\ell}^t((\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, q_1(\boldsymbol{Z}_1, \Psi_{\eta_\ell}, \bar{\varepsilon}))^\top\right], \tag{65}$$

$$\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1} = \frac{1}{\delta}\mathbb{E}\left[\bar{\boldsymbol{B}} f_1^{t+1}((\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1})^\top \bar{\boldsymbol{B}} + (\boldsymbol{G}_{\boldsymbol{B}}^{t+1})_1)^\top\right], \tag{66}$$

$$\boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{s+1,t+1} = \frac{1}{\delta}\mathbb{E}\left[\left(f_1^{s+1}((\boldsymbol{\nu}_{\boldsymbol{B}}^{s+1})^\top \bar{\boldsymbol{B}} + (\boldsymbol{G}_{\boldsymbol{B}}^{s+1})_1) - (\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{s+1})^\top \boldsymbol{\rho}^{-1} \bar{\boldsymbol{B}}\right)\right.$$
$$\left.\left(f_1^{t+1}((\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1})^\top \bar{\boldsymbol{B}} + (\boldsymbol{G}_{\boldsymbol{B}}^{t+1})_1) - (\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1})^\top \boldsymbol{\rho}^{-1} \bar{\boldsymbol{B}}\right)^\top\right], \tag{67}$$

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

B.2 Removing Dependencies on Noise on the RHS of (31) and (34)

Under assumptions (S0)–(S3) above, the parametric dependence of $\boldsymbol{V}_{\boldsymbol{\Theta}}^t$ on the RHS of (31)–(32) on $\boldsymbol{B}, \boldsymbol{\varepsilon}$ can be removed. Moreover, the parametric 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 parametric dependence on $\boldsymbol{\varepsilon}$ is via a normalized inner product. For example the estimator in (30) can be expressed as:

$$U(\hat{\boldsymbol{\eta}}(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, U(\boldsymbol{\eta}), \boldsymbol{\varepsilon}))) = \arg\max_{\psi \in \mathcal{X}} \frac{\langle (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_{[:,\psi]}, \boldsymbol{Z}_{[:,\psi]} \rangle}{n} + \frac{\langle (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_{[:,\psi]}, \boldsymbol{\varepsilon} \rangle}{n}. \qquad (68)$$

Indeed, the numerical experiments in Section 4 for chosen estimators demonstrate a strong agreement between the left-hand and right-hand sides of (31) and of (34), when $\boldsymbol{\varepsilon}$ on the right-hand side is substituted by an independent copy with the same limiting moments.

Appendix C. Proof and Computation of Optimal Denoisers f^t, g^t and likelihood $\mathcal{L}$

C.1 Proof of Proposition 2

This proof is similar to the derivation of the optimal denoisers for mixed regression in Tan and Venkataramanan (2023), with a few differences in the derivation of g^{*t}. We include both parts here for completeness. The proof relies on Lemma 6 and Lemma 7, 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 6 (Matrix Cauchy-Schwarz inequality, Lemma 2 in Lavergne (2008))
Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times L}$ 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

$$\mathbb{E}[\boldsymbol{B}^\top \boldsymbol{B}] - \mathbb{E}[\boldsymbol{B}^\top \boldsymbol{A}](\mathbb{E}[\boldsymbol{A}^\top \boldsymbol{A}])^{-1} \mathbb{E}[\boldsymbol{A}^\top \boldsymbol{B}] \succcurlyeq 0. \qquad (69)$$

Lemma 7 (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*

$$\mathbb{E}\left[(\boldsymbol{x} - \boldsymbol{\mu})h(\boldsymbol{x})^\top\right] = \boldsymbol{\Sigma} \, \mathbb{E}[h'(\boldsymbol{x})]^\top, \qquad (70)$$

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

The proof of Lemma 7 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 (28), we can rewrite $\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t$ in (22) as:

$$\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t = \frac{1}{\delta} \lim_{p \to \infty} \frac{1}{p} \sum_{j=1}^p \mathbb{E}\left[\bar{\boldsymbol{B}}\left(f_j^t(\bar{\boldsymbol{V}}_{\boldsymbol{B}}^t)\right)^\top\right] = \frac{1}{\delta}\mathbb{E}\left[\bar{\boldsymbol{B}}\left(f_j^t(\bar{\boldsymbol{V}}_{\boldsymbol{B}}^t)\right)^\top\right] = \frac{1}{\delta}\mathbb{E}\left[\mathbb{E}\left[\bar{\boldsymbol{B}}\left(f_j^t(\bar{\boldsymbol{V}}_{\boldsymbol{B}}^t)\right)^\top \big| \bar{\boldsymbol{V}}_{\boldsymbol{B}}^t\right]\right]$$

$$= \frac{1}{\delta}\mathbb{E}\left[\mathbb{E}\left[\bar{\boldsymbol{B}} | \bar{\boldsymbol{V}}_{\boldsymbol{B}}^t\right] \left(f_j^t(\bar{\boldsymbol{V}}_{\boldsymbol{B}}^t)\right)^\top\right] = \frac{1}{\delta}\mathbb{E}\left[f_j^{*t}(f_j^t)^\top\right], \qquad (71)$$

where we used the shorthand $f_j^t \equiv f_j^t(\bar{V}_B^t)$ and $f_j^{*t} \equiv f_j^{*t}(\bar{V}_B^t) = \mathbb{E}[\bar{B}|\bar{V}_B^t]$. Applying Lemma 6 yields

$$\mathbb{E}\left[f_j^{*t}(f_j^{*t})^\top\right] - \mathbb{E}\left[f_j^{*t}(f_j^t)^\top\right]\mathbb{E}\left[f_j^t(f_j^t)^\top\right]^{-1}\mathbb{E}\left[f_j^t(f_j^{*t})^\top\right] \succcurlyeq 0. \tag{72}$$

Since (23) can be simplified into

$$\bar{\kappa}_{\Theta}^{t,t} = \frac{1}{\delta}\mathbb{E}\left[f_j^t(f_j^t)^\top\right] - (\bar{\nu}_{\Theta}^t)^\top \rho^{-1}\nu_{\Theta}^t, \tag{73}$$

using (71) and (73) in (72), we obtain that

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

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

$$\rho^\top\left((\bar{\nu}_{\Theta}^t)^{-1}\right)^\top \bar{\kappa}_{\Theta}^{t,t}(\bar{\nu}_{\Theta}^t)^{-1}\rho - \underbrace{\rho^\top\left((\bar{\nu}_{\Theta}^t)^{-1}\right)^\top(\bar{\kappa}_{\Theta}^{t,t} - \delta)(\bar{\nu}_{\Theta}^t)^{-1}\rho}_{=:\Gamma_{\Theta}^t} \succcurlyeq 0, \tag{74}$$

which implies

$$\mathrm{Tr}\left(\rho^\top\left((\bar{\nu}_{\Theta}^t)^{-1}\right)^\top \bar{\kappa}_{\Theta}^{t,t}(\bar{\nu}_{\Theta}^t)^{-1}\rho\right) \geq \mathrm{Tr}\left(\Gamma_{\Theta}^t\right). \tag{75}$$

Recall from (26) that the LHS of (75) is the objective we wish to minimise via optimal f^t. Indeed, setting $f^t = f^{*t}$, (75) is satisfied with equality, proving part 1 of Proposition 2.

Remark 8 *When $f^t = f^{*t}$, (71) and (73) reduce to $\bar{\kappa}_{\Theta}^{t,t} = \bar{\nu}_{\Theta}^t - (\bar{\nu}_{\Theta}^t)^\top \rho^{-1}\bar{\nu}_{\Theta}^t$.*

Proof of part 2 (optimal g^t) Recall from (20) that

$$\bar{\nu}_B^{t+1} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\partial_{1i}\tilde{g}_i^t(Z_1, (\bar{V}_{\Theta}^t)_1, \bar{\Psi}_i, \bar{\varepsilon})\right].$$

We can rewrite the transpose of each summand as follows:

$$
\mathbb{E}\left[\partial_{1i}\tilde{g}_i^t(\boldsymbol{Z}_1,(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1,\bar{\Psi}_i,\bar{\varepsilon})\right]^\top
$$

$$
\stackrel{(a)}{=}\mathbb{E}_{(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1}\left\{\mathbb{E}_{\boldsymbol{Z}_1,\bar{\varepsilon}}\left[\partial_{1i}\tilde{g}_i^t\left(\boldsymbol{Z}_1,(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1,\bar{\Psi}_i,\bar{\varepsilon}\right)\Big|(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1\right]\right\}^\top
$$

$$
\stackrel{(b)}{=}\mathbb{E}_{(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1}\left\{\text{Cov}\left(\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1\right)^{-1}\right.
$$

$$
\left.\cdot\mathbb{E}_{\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon}}\left[\left(\boldsymbol{Z}_1-\mathbb{E}[\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1]\right)\tilde{g}_i^t\left(\boldsymbol{Z}_1,(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1,\bar{\Psi}_i,\bar{\varepsilon}\right)^\top\Big|(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1\right]\right\}
$$

$$
\stackrel{(c)}{=}\mathbb{E}_{(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1}\left\{\text{Cov}\left(\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1\right)^{-1}\right.
$$

$$
\left.\cdot\mathbb{E}_{q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})}\left[\mathbb{E}_{\boldsymbol{Z}_1}\left[\left(\boldsymbol{Z}_1-\mathbb{E}[\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1]\right)g_i^t\left((\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1,q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right)^\top\Big|(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1,q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right]\right]\right\}
$$

$$
=\mathbb{E}_{(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1}\left\{\text{Cov}\left(\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1\right)^{-1}\right.
$$

$$
\left.\cdot\mathbb{E}_{q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})}\left[\left(\mathbb{E}[\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1,q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})]-\mathbb{E}[\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1]\right)g_i^t\left((\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1,q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right)^\top\right]\right\}
$$

$$
\stackrel{(d)}{=}\mathbb{E}\Big\{\overbrace{\text{Cov}\left(\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1\right)^{-1}\left(\mathbb{E}\left[\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1,q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right]-\mathbb{E}[\boldsymbol{Z}_1|(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1]\right)}^{=g_i^{*t}((\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1,q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon}))}
$$

$$
\cdot g_i^t\left((\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1,q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right)^\top\Big\}
$$

$$
=\mathbb{E}\left[g_i^{*t}(g_i^t)^\top\right],\tag{76}
$$

where (a) and (c) follow from the law of total expectation; (b) uses Lemma 7; and (d) uses (29). In the last line of (76) we have used the shorthand $g_i^t\equiv g_i^t\left((\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1,q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right)$ and $g_i^{*t}\equiv g_i^{*t}\left((\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1,q(\boldsymbol{Z}_1,\bar{\Psi}_i,\bar{\varepsilon})\right)$. Substituting (76) in (20) yields

$$
\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1}=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\mathbb{E}[g_i^t(g_i^{*t})^\top]=\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}(g^t)^\top g^{*t}\right].\tag{77}
$$

Note that since $\pi_{\bar{\Psi}_i}$ differs across i, g_i^t differs across i and the sum in (77) cannot be reduced. Lemma 6 implies that

$$
\mathbb{E}\left[\frac{1}{n}(g^{*t})^\top g^{*t}\right]-\mathbb{E}\left[\frac{1}{n}(g^{*t})^\top(g^t)\right]\mathbb{E}\left[\frac{1}{n}(g^t)^\top g^t\right]^{-1}\mathbb{E}\left[\frac{1}{n}(g^t)^\top g^{*t}\right]\succcurlyeq 0.\tag{78}
$$

Recalling from (20)-(21) that the limiting parameters $\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1}=\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}(g^t)^\top g^{*t}\right]$ and $\bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1}=\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}(g^t)^\top g^t\right]$ exist, we can take limits to obtain

$$
\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}(g^{*t})^\top g^{*t}\right]
$$

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

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

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

(80)

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

$$[(\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^{-1}]^{\top}\bar{\boldsymbol{\kappa}}_{\boldsymbol{B}}^{t+1,t+1}(\bar{\boldsymbol{\nu}}_{\boldsymbol{B}}^{t+1})^{-1}-\boldsymbol{\Gamma}_{\boldsymbol{B}}^{t+1}\succcurlyeq 0,\tag{81}$$

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 (27) that g^t is optimized 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 9 *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 the Optimal Denoiser g^{*t}

In this section, we compute the optimal denoiser g^{*t} for the linear, logistic, and rectified linear models. We shall use $\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ to denote the probability density function (PDF) of $\mathsf{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ evaluated at $\boldsymbol{x}$. We use $\phi(x)$ to denote the PDF of a standard scalar Gaussian $\mathsf{N}(0,1)$ evaluated at x, and $\Phi(x)$ to denote its cumulative density function (CDF) up to x. For a generic random vector $\boldsymbol{U}$, with a slight abuse of notation we will use $\mathbb{P}(\boldsymbol{U}=\boldsymbol{u})$ to denote its density evaluated at $\boldsymbol{u}$. We begin with the following standard lemmas as useful preliminaries.

Lemma 10 (Conditioning property of multivariate Gaussian) *Let $\boldsymbol{x}\in\mathbb{R}^n$ and $\boldsymbol{y}\in\mathbb{R}^m$ be jointly Gaussian, with*

$$\begin{bmatrix}\boldsymbol{x}\\\boldsymbol{y}\end{bmatrix}\sim\mathsf{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x\\\boldsymbol{\mu}_y\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_x&\boldsymbol{\Sigma}_{xy}\\\boldsymbol{\Sigma}_{xy}^{\top}&\boldsymbol{\Sigma}_y\end{bmatrix}\right).$$

Then $(\boldsymbol{x}\mid\boldsymbol{y}=\tilde{\boldsymbol{y}})\sim\mathsf{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu}=\boldsymbol{\mu}_x+\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}(\tilde{\boldsymbol{y}}-\boldsymbol{\mu}_y),\quad\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_x-\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{xy}^{\top}.$$

Lemma 11 (Conditioning property of Gaussian mixtures) *Let $\Gamma\sim\pi_\Gamma$ be a categorical variable supported on $[1,2,\ldots,K]$. With $\boldsymbol{x}\in\mathbb{R}^n$ and $\boldsymbol{y}\in\mathbb{R}^m$, suppose*

$$\begin{bmatrix}\boldsymbol{x}\\\boldsymbol{y}\end{bmatrix}\sim\mathsf{N}\left(\begin{bmatrix}\boldsymbol{\mu}_{k,x}\\\boldsymbol{\mu}_{k,y}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{k,x},\boldsymbol{\Sigma}_{k,xy}\\\boldsymbol{\Sigma}_{k,xy}^{\top},\boldsymbol{\Sigma}_{k,y}\end{bmatrix}\right)\qquad if\quad\Gamma=k.\tag{82}$$

Then, with Γ is independent of $\boldsymbol{x},\boldsymbol{y}$, we have

$$\mathbb{E}[\boldsymbol{x}\mid\boldsymbol{y}=\boldsymbol{y}_0]=\frac{\sum_{k=1}^K\pi_\Gamma(k)\phi(\boldsymbol{y}_0;\boldsymbol{\mu}_{k,y},\boldsymbol{\Sigma}_{k,y})\mathbb{E}[\boldsymbol{x}\mid\boldsymbol{y}=\boldsymbol{y}_0,\Gamma=k]}{\sum_{\tilde{k}=1}^K\pi_\Gamma(\tilde{k})\phi(\boldsymbol{y}_0;\boldsymbol{\mu}_{\tilde{k},y},\boldsymbol{\Sigma}_{\tilde{k},y})}.\tag{83}$$

C.2.1 Computation of g^{*t} for Linear Regression

Assume Gaussian noise $\bar{\varepsilon} \sim \mathsf{N}(0, \sigma^2)$. Using (24), we have

$$\begin{bmatrix} \boldsymbol{Z}_1 \\ (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 \end{bmatrix} \sim \mathsf{N}_{2L}(\boldsymbol{0}, \boldsymbol{\Gamma}^t), \quad \text{where} \quad \boldsymbol{\Gamma}^t := \begin{bmatrix} \boldsymbol{\rho} & \bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t \\ (\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^\top & \boldsymbol{\Sigma}_V^t \end{bmatrix}, \quad \boldsymbol{\Sigma}_V^t = (\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^\top \boldsymbol{\rho}^{-1} \bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t + \bar{\boldsymbol{\kappa}}_{\boldsymbol{\Theta}}^{t,t}. \quad (84)$$

Lemma 10 then gives the formulas for $\mathbb{E}[\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i]$ and $\mathrm{Cov}(\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i)$, the first and last terms of g_i^{*t} in (29):

$$\begin{aligned} \mathbb{E}[\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i] &= \bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t \left(\boldsymbol{\Sigma}_V^t\right)^{-1} \boldsymbol{V}_i, \\ \mathrm{Cov}(\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i) &= \boldsymbol{\rho} - (\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^\top \left(\boldsymbol{\Sigma}_V^t\right)^{-1} \bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t. \end{aligned} \quad (85)$$

When $f^t = f^{*t}$, we can use the simplifications in Remark 8 to obtain:

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

We now calculate the middle term $\mathbb{E}[\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i]$ of g_i^{*t} in (29). Recalling from (5) that $q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = (Z_1)_{\bar{\Psi}_i} + \bar{\varepsilon}$ where $\boldsymbol{Z}_1 \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{\rho})$, we have

$$\begin{aligned} \mathbb{E}[\boldsymbol{Z}_1 q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) \, | \bar{\Psi}_i = \ell] &= \mathbb{E}[\boldsymbol{Z}_1 (Z_1)_\ell] = \boldsymbol{\rho}_{[:,\ell]}, \\ \mathbb{E}[q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon})^2 \, | \bar{\Psi}_i = \ell] &= \mathbb{E}[(Z_1)_\ell^2 + \bar{\varepsilon}^2] = \rho_{\ell,\ell} + \sigma^2, \\ \mathbb{E}[(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) \, | \bar{\Psi}_i = \ell] &= \mathbb{E}[(\boldsymbol{Z}\boldsymbol{\rho}^{-1}\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)_1 (Z_1)_\ell] = (\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)^\top \boldsymbol{\rho}^{-1} \boldsymbol{\rho}_{[:,\ell]} = \left((\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)_{[\ell,:]}\right)^\top, \end{aligned} \quad (86)$$

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

$$\begin{bmatrix} (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 \\ q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) \end{bmatrix} \sim \mathsf{N}_{L+1}\left(\boldsymbol{0}, \boldsymbol{\Sigma}^t\right), \quad \text{where} \quad \boldsymbol{\Sigma}^t := \begin{bmatrix} \boldsymbol{\Sigma}_V^t & \left((\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)_{[\ell,:]}\right)^\top \\ (\bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t)_{[\ell,:]} & \rho_{\ell,\ell} + \sigma^2 \end{bmatrix}. \quad (87)$$

Conditioned on $\Psi_i = \ell$ the random variables $(\boldsymbol{Z}_1, (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1, q_1(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}))$ are jointly Gaussian with zero mean and covariance matrix determined by (84)-(87). Applying Lemmas 10–11 on these jointly Gaussian variables, we obtain:

$$\mathbb{E}\left[\boldsymbol{Z}_1 \mid (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, \, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i\right] = \frac{\sum_{\ell=1}^L \pi_{\bar{\Psi}_i}(\ell) \, \boldsymbol{\lambda}^t(\boldsymbol{V}_i, u_i, \ell) \, \mathcal{L}_1(\boldsymbol{V}_i, u_i | \ell)}{\sum_{\tilde{\ell}=1}^L \pi_{\bar{\Psi}_i}(\tilde{\ell}) \, \mathcal{L}_1(\boldsymbol{V}_i, u_i | \tilde{\ell})}, \quad (88)$$

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

$$\mathcal{L}_1(\boldsymbol{V}_i, u_i | \ell) := \phi\left(\begin{bmatrix} \boldsymbol{V}_i \\ u_i \end{bmatrix}; \boldsymbol{0}, \boldsymbol{\Sigma}^t\right), \quad (89)$$

$$\begin{aligned} \boldsymbol{\lambda}^t(\boldsymbol{V}_i, u_i, \ell) &:= \mathbb{E}[\boldsymbol{Z}_1 \mid (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, \, q_1(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i, \, \bar{\Psi}_i = \ell] \\ &= \begin{bmatrix} \bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}}^t & \boldsymbol{\rho}_{[:,\ell]} \end{bmatrix} [\boldsymbol{\Sigma}^t]^{-1} \begin{bmatrix} \boldsymbol{V}_i \\ u_i \end{bmatrix} \in \mathbb{R}^L. \end{aligned} \quad (90)$$

C.2.2 Computation of g^{*t} for Logistic Regression

The only term in the expression for g^{*t} in (29) that differs from the linear regression case, and therefore needs to be re-computed, is $\mathbb{E}[\boldsymbol{Z}_1 \mid (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i]$. For $k \in [L]$, we can write

$$\mathbb{E}\Big[(Z_1)_k \mid (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i\Big] = \frac{\sum_{\ell=1}^L \pi_{\bar{\Psi}_i}(\ell)\, \widetilde{\lambda}^t(\boldsymbol{V}_i, u_i, \ell)}{\sum_{\tilde{\ell}=1}^L \pi_{\bar{\Psi}_i}(\tilde{\ell})\, \mathcal{L}_1(\boldsymbol{V}_i, u_i | \tilde{\ell})}\,,$$

where

$$\mathcal{L}_1(\boldsymbol{V}_i, u_i | \ell) := \mathbb{P}\left((\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i \mid \bar{\Psi}_i = \ell\right), \tag{91}$$

$$\widetilde{\lambda}^t(\boldsymbol{V}_i, u_i, \ell) := \int_{-\infty}^{\infty} z\, \mathbb{P}\left((Z_1)_k = z, (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i \mid \bar{\Psi}_i = \ell\right) \mathrm{d}z. \tag{92}$$

In the rest of this subsection, we provide expressions for $\mathcal{L}_1(\boldsymbol{V}_i, u_i | \ell)$ and $\widetilde{\lambda}^t(\boldsymbol{V}_i, u_i, \ell)$. As preliminaries, we adopt the following shorthand for $\ell \in [L]$:

$$\mu_\ell \equiv \mu_\ell(\boldsymbol{V}_i) := \left(\mathbb{E}[\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i]\right)_\ell, \quad \sigma_\ell \equiv \sigma_\ell(\boldsymbol{V}_i) := \sqrt{\left[\mathrm{Cov}(\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i)\right]_{\ell,\ell}}. \tag{93}$$

where the quantities on the right are defined in (85). We also recall from (85) that conditioned on $(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i$,

$$\begin{bmatrix} (Z_1)_k \\ (Z_1)_\ell \end{bmatrix} \sim \mathsf{N}\left(\begin{bmatrix} \mu_k \\ \mu_\ell \end{bmatrix}, \mathrm{Cov}\left((Z_1)_k, (Z_1)_\ell\right)\right),$$

where

$$\mathrm{Cov}\left((Z_1)_k, (Z_1)_\ell\right) = \begin{bmatrix} \rho_{kk} & \rho_{k\ell} \\ \rho_{k\ell} & \rho_{\ell\ell} \end{bmatrix} - \begin{bmatrix} \bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}[k,:]}^t \\ \bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}[\ell,:]}^t \end{bmatrix} (\boldsymbol{\Sigma}_{\boldsymbol{V}}^t)^{-1} \begin{bmatrix} \bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}[k,:]}^t \\ \bar{\boldsymbol{\nu}}_{\boldsymbol{\Theta}[\ell,:]}^t \end{bmatrix}^\top =: \begin{bmatrix} \sigma_k^2 & \lambda_{k\ell} \\ \lambda_{k\ell} & \sigma_\ell^2 \end{bmatrix}.$$

Applying Lemma 10, this implies that conditioned on $(\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i$,

$$(Z_1)_k \mid (Z_1)_\ell \sim \mathsf{N}\left(\mu_{k|\ell}, \sigma_{k|\ell}^2\right),$$

where

$$\mu_{k|\ell} = \mu_k + \frac{\lambda_{k\ell}}{\sigma_\ell^2}\left((Z_1)_\ell - \mu_\ell\right), \qquad \sigma_{k|\ell}^2 = \sigma_k^2 - \frac{\lambda_{kl}^2}{\sigma_\ell^2}. \tag{94}$$

We will also use Lemmas 12–13 below to avoid intractable integrals in the sequel.

Lemma 12 (Logistic approximation, Section 4.5.2 in Bishop (2006)) *The logistic function $\zeta'(z)$ in (3) can be approximated by a standard Gaussian CDF as:*

$$\zeta'(z) \approx \Phi\left(\gamma z\right), \quad \text{where} \quad \gamma = \sqrt{\frac{\pi}{8}}\,. \tag{95}$$

Lemma 13 (Useful Gaussian Identities, Table 1 in Owen (1980)) *It holds that*

$$\int_{-\infty}^{\infty} \Phi(a + bx)\phi(x)\mathrm{d}x = \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right), \tag{96}$$

$$\int_{-\infty}^{\infty} x\Phi(a + bx)\phi(x)\mathrm{d}x = \frac{b}{\sqrt{1 + b^2}}\, \phi\left(\frac{a}{\sqrt{1 + b^2}}\right). \tag{97}$$

We will compute $\widetilde{\lambda}^t(\boldsymbol{V}_i, u_i, \ell)$ in the $k = \ell$ and $k \neq \ell$ cases separately.

Case 1: $k = \ell$. We can rewrite the probability inside the integral in (92) as follows:

$$
\begin{aligned}
\mathbb{P}\left((Z_1)_\ell = z, (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i \mid \bar{\Psi}_i = \ell\right) \\
= \mathbb{P}\left((\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i\right) \mathbb{P}\left((Z_1)_\ell = z \mid (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i\right) \mathbb{P}\left(q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i \mid (Z_1)_\ell = z, \bar{\Psi}_i = \ell\right) \\
= \phi(\boldsymbol{V}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_V^t)\phi(z; \mu_\ell, \sigma_\ell^2) \left[\mathbb{1}\{u_i = 1\}\zeta'(z) + \mathbb{1}\{u_i = 0\}(1 - \zeta'(z))\right],
\end{aligned}
\tag{98}
$$

where we recalled from (3) that $\mathbb{P}\left(q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = 1 \mid (Z_1)_\ell = z, \bar{\Psi}_i = \ell\right) = \zeta'(z)$. Substituting (98) into (92), and applying Lemmas 12–13 yields

$$
\begin{aligned}
\widetilde{\lambda}^t(\boldsymbol{V}_i, u_i, \ell) \approx \phi(\boldsymbol{V}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_V^t) \big\{ \mathbb{1}\{u_i = 1\} \left[\tau_\ell \sigma_\ell^2 \phi(\tau_\ell \mu_\ell) + \mu_\ell \Phi(\tau_\ell \mu_\ell)\right] + \\
\mathbb{1}\{u_i = 0\} \left[\mu_\ell - \tau_\ell \sigma_\ell^2 \phi(\tau_\ell \mu_\ell) - \mu_\ell \Phi(\tau_\ell \mu_\ell)\right] \big\},
\end{aligned}
$$

where $\tau_\ell = \gamma/\sqrt{1 + \gamma^2 \sigma_\ell^2}$ with $\gamma = \sqrt{\pi/8}$ as defined in (95).

Case 2: $k \neq \ell$. We rewrite the density inside the integral in (92) as follows:

$$
\begin{aligned}
\mathbb{P}\left((Z_1)_k = z, (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i \mid \bar{\Psi}_i = \ell\right) \\
= \int_{-\infty}^{\infty} \mathbb{P}\left((\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i\right) \mathbb{P}\left((Z_1)_\ell = s \mid (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i\right) \mathbb{P}\left(q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i \mid (Z_1)_\ell = s, \bar{\Psi}_i = \ell\right) \\
\mathbb{P}\left((Z_1)_k = z \mid (Z_1)_\ell = s, (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i\right) \mathrm{d}s \\
= \phi(\boldsymbol{V}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_V^t) \int_{-\infty}^{\infty} \phi(s; \mu_\ell, \sigma_\ell^2) \left[\mathbb{1}\{u_i = 1\}\zeta'(s) + \mathbb{1}\{u_i = 0\}(1 - \zeta'(s))\right] \phi\left(z; \mu_{k|\ell}, \sigma_{k|\ell}^2\right) \mathrm{d}s.
\end{aligned}
\tag{99}
$$

Substituting (99) into (92) and applying Lemmas 12–13 yields

$$
\begin{aligned}
\widetilde{\lambda}^t(\boldsymbol{V}_i, u_i, \ell) \approx \phi(\boldsymbol{V}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_V^t) \big\{ \mathbb{1}\{u_i = 1\} \left[\lambda_{k\ell} \tau_\ell \phi(\tau_\ell \mu_\ell) + \mu_k \Phi(\tau_\ell \mu_\ell)\right] + \\
\mathbb{1}\{u_i = 0\} \left[\mu_k - \lambda_{k\ell} \tau_\ell \phi(\tau_\ell \mu_\ell) - \mu_k \Phi(\tau_\ell \mu_\ell)\right] \big\}.
\end{aligned}
$$

Finally, to compute $\mathcal{L}_1(\boldsymbol{V}_i, u_i | \ell)$, we sum over z in (98) and apply Lemma 12 to obtain

$$
\mathcal{L}_1(\boldsymbol{V}_i, u_i | \ell) \approx \phi(\boldsymbol{V}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_V^t) \left\{\mathbb{1}\{u_1 = 1\}\Phi(\tau_\ell \mu_\ell) + \mathbb{1}\{u_1 = 0\}[1 - \Phi(\tau_\ell \mu_\ell)]\right\}.
\tag{100}
$$

C.2.3 COMPUTATION OF g^{*t} FOR RECTIFIED LINEAR REGRESSION

Similar to Appendix C.2.2, the only term in the expressions for g^{*t} that differs from linear regression, and therefore needs to be re-computed, is $\mathbb{E}[\boldsymbol{Z}_1 | (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i]$. We will express this term using a binary auxiliary variable

$$
\Omega := \mathbb{1}\{(Z_1)_\ell \geq 0\},
$$

which indicates whether the data are generated through the sloped part or flat part of the rectified linear unit, conditioned on $\bar{\Psi}_i = \ell$. In many parts of the derivation, variables depending on Ω exhibit truncated Gaussian distributions. We begin with some useful results on these distributions.

Lemma 14 (Useful truncated Gaussian identities) *Consider the truncated Gaussian variables X^+ and X^- with probability densities*

$$\phi^+(x; \mu, \sigma^2) := \mathbb{1}\{x \geq 0\} \frac{\phi(x; \mu, \sigma^2)}{\Phi(\mu/\sigma)}, \qquad \phi^-(x; \mu, \sigma^2) := \mathbb{1}\{x < 0\} \frac{\phi(x; \mu, \sigma^2)}{\Phi(-\mu/\sigma)}.$$

It holds that

$$\mathbb{E}[X^+] = \mu + \sigma \frac{\phi(0; \mu, \sigma^2)}{\Phi(\mu/\sigma)}, \qquad \mathbb{E}[X^-] = \mu - \sigma \frac{\phi(0; \mu, \sigma^2)}{\Phi(-\mu/\sigma)},$$

and it follows that:

$$\phi^+(x; \mu, \sigma^2)\phi(z; x, \tau^2) = \phi(z; \mu, \sigma^2 + \tau^2) \frac{\Phi(\mu_*/\sigma_*)}{\Phi(\mu/\sigma)} \phi^+(x; \mu_*, \sigma_*^2),$$

where $\mu_* = \frac{z\sigma^2 + \mu\tau^2}{\sigma^2 + \tau^2}$ and $\sigma_*^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}$.

Equipped with the definitions and lemmas above, we can write for $k \in [L]$,

$$\mathbb{E}\left[(Z_1)_k \mid (\bar{V}_\Theta^t)_1 = V_i, q(Z_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i\right] = \frac{\sum_{\ell=1}^L \sum_{\omega \in \{0,1\}} \pi_{\bar{\Psi}_i}(\ell)\, \widetilde{\lambda}^t(k, V_i, u_i, \ell, \omega)}{\sum_{\tilde{\ell}=1}^L \sum_{\tilde{\omega} \in \{0,1\}} \pi_{\bar{\Psi}_i}(\tilde{\ell})\, \mathcal{L}_1(V_i, u_i, \tilde{\omega}|\tilde{\ell})},$$

where

$$\mathcal{L}_1(V_i, u_i, \omega|\ell) := \mathbb{P}\left((\bar{V}_\Theta^t)_1 = V_i, q(Z_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i, \Omega = \omega \mid \bar{\Psi}_i = \ell\right),$$

$$\widetilde{\lambda}^t(k, V_i, u_i, \ell, \omega) = \int_{-\infty}^\infty z \mathbb{P}\left((Z_1)_k = z, (\bar{V}_\Theta^t)_1 = V_i, q(Z_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i, \Omega = \omega \mid \bar{\Psi}_i = \ell\right) \mathrm{d}z.$$

(101)

In the rest of this subsection, we provide the expression for the terms $\mathcal{L}_1(V_i, u_i, \omega|\ell)$ and $\widetilde{\lambda}^t(k, V_i, u_i, \ell, \omega)$ for $\omega = 0$ or 1, respectively.

Case 1: $\Omega = 0$. This implies that $(Z_1)_\ell < 0$, and $q(Z_1, \bar{\Psi}_i = \ell, \bar{\varepsilon}) = \max\{(Z_1)_\ell, 0\} + \bar{\varepsilon} = \bar{\varepsilon}$. Using the law of conditional probability, we have

$$\mathcal{L}_1(V_i, u_i, \Omega = 0|\ell) = \mathbb{P}\left((\bar{V}_\Theta^t)_1 = V_i\right) \mathbb{P}\left(\Omega = 0 \mid (\bar{V}_\Theta^t)_1 = V_i, \bar{\Psi}_i = \ell\right) \cdot$$
$$\mathbb{P}\left(q(Z_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i \mid (\bar{V}_\Theta^t)_1 = V_i, \Omega = 0, \bar{\Psi}_i = \ell\right)$$
$$= \phi\left(V_i; \mathbf{0}, \Sigma_V^t\right) \Phi\left(-\mu_\ell/\sigma_\ell\right) \phi\left(u_i; 0, \sigma^2\right),$$

(102)

where Σ_V^t is defined in (84), and μ_ℓ, σ_ℓ in (93). To evaluate $\widetilde{\lambda}^t(k, V_i, u_i, \ell, \omega)$, we consider the $k = \ell$ and $k \neq \ell$ cases separately.

Case 1a: $\Omega = 0, k = \ell$. The density inside the integral of (101) can be rewritten as

$$\mathbb{P}\left((Z_1)_\ell = z, (\bar{V}_\Theta^t)_1 = V_i, q(Z_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i, \Omega = 0 \mid \bar{\Psi}_i = \ell\right)$$
$$= \mathbb{P}\left((\bar{V}_\Theta^t)_1 = V_i\right) \mathbb{P}\left(\Omega = 0 \mid (\bar{V}_\Theta^t)_1 = V_i, \bar{\Psi}_i = \ell\right) \cdot$$
$$\mathbb{P}\left((Z_1)_\ell = z \mid (\bar{V}_\Theta^t)_1 = V_i, \Omega = 0, \bar{\Psi}_i = \ell\right) \cdot$$
$$\mathbb{P}\left(q(Z_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i \mid (Z_1)_\ell = z, (\bar{V}_\Theta^t)_1 = V_i, \Omega = 0, \bar{\Psi}_i = \ell\right)$$
$$= \phi\left(V_i; \mathbf{0}, \Sigma_V^t\right) \Phi\left(-\mu_\ell/\sigma_\ell\right) \phi^-(z; \mu_\ell, \sigma_\ell^2)\phi(u_i; 0, \sigma^2).$$

(103)

Substituting (103) into (101) and applying Lemma 14 yields

$$\widetilde{\lambda}^t(k = \ell, \boldsymbol{V}_i, u_i, \ell, \Omega = 0) = \phi\left(\boldsymbol{V}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{V}}^t\right) \Phi\left(-\mu_\ell/\sigma_\ell\right) \left[\mu_\ell - \sigma_\ell \frac{\phi(0; \mu_\ell, \sigma_\ell^2)}{\Phi(-\mu_\ell/\sigma_\ell)}\right] \phi(u_i; 0, \sigma^2).$$

Case 1b: $\Omega = 0, k \neq \ell$. Similar to (103), the density inside the integral of (101) can be rewritten as

$$\mathbb{P}\left((Z_1)_k = z, (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i, \Omega = 0 \mid \bar{\Psi}_i = \ell\right)$$
$$= \phi\left(\boldsymbol{V}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{V}}^t\right) \Phi\left(-\mu_\ell/\sigma_\ell\right) \mathbb{P}\left((Z_1)_k = z \mid (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, \Omega = 0, \bar{\Psi}_i = \ell\right) \phi(u_i; 0, \sigma^2). \quad (104)$$

To evaluate the second last term in (104), we apply the law of total probability to obtain

$$\mathbb{P}\left((Z_1)_k = z \mid (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, \Omega = 0, \bar{\Psi}_i = \ell\right)$$
$$= \int \mathbb{P}\left((Z_1)_\ell = s \mid (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, (Z_1)_\ell < 0\right) \mathbb{P}\left((Z_1)_k = z \mid (Z_1)_\ell = s, (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, (Z_1)_\ell < 0\right) \mathrm{d}s$$
$$= \int \phi^-(s; \mu_\ell, \sigma_\ell^2) \phi\left(z; \mu_{k|\ell}, \sigma_{k|\ell}^2\right) \mathrm{d}s, \quad (105)$$

where $\mu_{k|\ell}, \sigma_{k|\ell}^2$ are given by (94). Substituting (105) into (104) then into (101) and applying Lemma 14 gives

$$\widetilde{\lambda}^t(k \neq \ell, \boldsymbol{V}_i, u_i, \ell, \Omega = 0)$$
$$= \phi\left(\boldsymbol{V}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{V}}^t\right) \Phi\left(-\mu_\ell/\sigma_\ell\right) \phi(u_i; 0, \sigma^2) \int \phi^-(s; \mu_\ell, \sigma_\ell^2) \mu_{k|\ell} \, \mathrm{d}s$$
$$= \phi\left(\boldsymbol{V}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{V}}^t\right) \Phi\left(-\mu_\ell/\sigma_\ell\right) \phi(u_i; 0, \sigma^2) \left\{\frac{\lambda_{k\ell}}{\sigma_\ell^2}\left[\mu_\ell - \sigma_\ell \frac{\phi(0; \mu_\ell, \sigma_\ell^2)}{\Phi(-\mu_\ell/\sigma_\ell)}\right] + \mu_k - \frac{\lambda_{k\ell}}{\sigma_\ell^2}\mu_\ell\right\}.$$

Case 2: $\Omega = 1$. In this case, $(Z_1)_\ell \geq 0$, and $q(\boldsymbol{Z}_1, \bar{\Psi}_i = \ell, \bar{\varepsilon}) = (Z_1)_\ell + \bar{\varepsilon}$. Applying Lemma 14 yields

$$\mathbb{P}\left(q(\boldsymbol{Z}_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i \mid (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, \Omega = 1, \bar{\Psi}_i = \ell\right)$$
$$= \mathbb{P}\left((Z_1)_\ell + \bar{\varepsilon} = u_i \mid (\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{V}_i, (Z_1)_\ell \geq 0\right) = \phi(u_i; \mu_\ell, \sigma_\ell^2 + \sigma^2) \frac{\Phi(\mu_{\ell*}/\sigma_{\ell*})}{\Phi(\mu_\ell/\sigma_\ell)}, \quad (106)$$

where

$$\mu_{\ell*} = \frac{u_i \sigma_\ell^2 + \mu_\ell \sigma^2}{\sigma_\ell^2 + \sigma^2}, \qquad \sigma_{\ell*}^2 = \frac{\sigma_\ell^2 \sigma^2}{\sigma_\ell^2 + \sigma^2}. \quad (107)$$

Using (106)–(107) and the law of conditional probability similarly to (102), we obtain

$$\mathcal{L}_1(\boldsymbol{V}_i, u_i, \Omega = 1 | \ell) = \phi\left(\boldsymbol{V}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{V}}^t\right) \phi\left(u_i; \mu_\ell, \sigma_\ell^2 + \sigma^2\right) \Phi(\mu_{\ell*}/\sigma_{\ell*}). \quad (108)$$

Next, we evaluate $\lambda^t(k, \boldsymbol{V}_i, u_i, \ell, \Omega = 1)$ for $k = \ell$ and $k \neq \ell$ separately.

Case 2a: $\Omega = 1, k = \ell$. Analogously to (103), we apply the law of conditional probability and Lemma 14 to obtain

$$\mathbb{P}\left((Z_1)_\ell = z, (\bar{V}_{\Theta}^t)_1 = V_i, q(Z_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i, \Omega = 1 \mid \bar{\Psi}_i = \ell\right)$$
$$= \phi\left(V_i; 0, \Sigma_V^t\right) \Phi\left(\mu_\ell/\sigma_\ell\right) \phi^+\left(z; \mu_\ell, \sigma_\ell^2\right) \phi(u_i; z, \sigma^2)$$
$$= \phi\left(V_i; 0, \Sigma_V^t\right) \phi\left(u_i; \mu_\ell, \sigma_\ell^2 + \sigma^2\right) \Phi(\mu_{\ell*}/\sigma_{\ell*})\phi^+(z; \mu_{\ell*}, \sigma_{\ell*}^2). \quad (109)$$

Substituting (109) into (101) and applying Lemma 14 yields

$$\widetilde{\lambda}^t(k = \ell, V_i, u_i, \ell, \Omega = 1)$$
$$= \phi\left(V_i; 0, \Sigma_V^t\right) \phi\left(u_i; \mu_\ell, \sigma_\ell^2 + \sigma^2\right) \Phi(\mu_{\ell*}/\sigma_{\ell*}) \left[\mu_{\ell*} + \sigma_{\ell*} \frac{\phi(0; \mu_{\ell*}, \sigma_{\ell*})}{\Phi(\mu_{\ell*}/\sigma_{\ell*})}\right]. \quad (110)$$

Case 2b: $\Omega = 1, k \neq \ell$. Following the same steps as (104)–(105) and applying Lemma 14, we can write

$$\mathbb{P}\left((Z_1)_k = z, (\bar{V}_{\Theta}^t)_1 = V_i, q(Z_1, \bar{\Psi}_i, \bar{\varepsilon}) = u_i, \Omega = 1 \mid \bar{\Psi}_i = \ell\right)$$
$$= \phi\left(V_i; 0, \Sigma_V^t\right) \Phi\left(\mu_\ell/\sigma_\ell\right) \int \phi^+\left(s; \mu_\ell, \sigma_\ell^2\right) \phi\left(z; \mu_{k|\ell}, \sigma_{k|\ell}^2\right) \phi(u_i; s, \sigma^2)\mathrm{d}s$$
$$= \phi\left(V_i; 0, \Sigma_V^t\right) \phi(u_i; \mu_\ell, \sigma_\ell^2 + \sigma^2)\Phi(\mu_{\ell*}/\sigma_{\ell*}) \int \phi^+(s; \mu_{\ell*}, \sigma_{\ell*}^2)\phi\left(z; \mu_{k|\ell}, \sigma_{k|\ell}^2\right) \mathrm{d}s \quad (111)$$

where $\mu_{k|\ell}$ and $\sigma_{k|\ell}^2$ are defined in (94), while $\mu_{\ell*}$ and $\sigma_{\ell*}^2$ are defined in (107). Substituting (111) into (101) and applying Lemma 14 gives

$$\widetilde{\lambda}^t(k \neq \ell, V_i, u_i, \ell, \Omega = 1)$$
$$= \phi\left(V_i; 0, \Sigma_V^t\right) \phi(u_i; \mu_\ell, \sigma_\ell^2 + \sigma^2)\Phi(\mu_{\ell*}/\sigma_{\ell*}) \int \phi^+(s; \mu_{\ell*}, \sigma_{\ell*}^2) \int z\phi\left(z; \mu_{k|\ell}, \sigma_{k|\ell}^2\right) \mathrm{d}z\mathrm{d}s$$
$$= \phi\left(V_i; 0, \Sigma_V^t\right) \phi(u_i; \mu_\ell, \sigma_\ell^2 + \sigma^2)\Phi(\mu_{\ell*}/\sigma_{\ell*}) \int \phi^+(s; \mu_{\ell*}, \sigma_{\ell*}^2)\mu_{k|\ell} \,\mathrm{d}s$$
$$= \phi\left(V_i; 0, \Sigma_V^t\right) \phi(u_i; \mu_\ell, \sigma_\ell^2 + \sigma^2)\Phi(\mu_{\ell*}/\sigma_{\ell*})$$
$$\cdot \left\{\frac{\lambda_{k\ell}}{\sigma_\ell^2} \left[\mu_{\ell*} + \sigma_{\ell*} \frac{\phi(0; \mu_{\ell*}, \sigma_{\ell*}^2)}{\Phi(\mu_{\ell*}/\sigma_{\ell*})}\right] + \mu_k - \frac{\lambda_{k\ell}}{\sigma_\ell^2}\mu_\ell\right\}. \quad (112)$$

C.3 Computation of the Likelihood $\mathcal{L}$

Recall from (33) that $\mathcal{L}$ represents the likelihood of $(V_\Theta^t, q(Z, \bar{\Psi}, \bar{\varepsilon}))$ given $\bar{\Psi} = \psi$, where $\nu_\Theta^t, \kappa_\Theta^t$ are computed as in (10)–(13) with ψ substituted for Ψ. 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 in the case of linear regression as follows:

$$\mathcal{L}(V, u|\psi) := \prod_{i=1}^n \phi\left(\begin{bmatrix} V_i \\ u_i \end{bmatrix}; 0, \Sigma_{\psi_i}^t\right), \quad \text{where } \Sigma_{\psi_i}^t := \begin{bmatrix} (\nu_\Theta^t)^\top \rho^{-1}\nu_\Theta^t + \kappa_\Theta^{t,t} & \left((\nu_\Theta^t)_{[\psi_i,:]}\right)^\top \\ (\nu_\Theta^t)_{[\psi_i,:]} & \rho_{\psi_i,\psi_i} + \sigma^2 \end{bmatrix}.$$

In the cases of logistic regression and rectified linear regression with $\bar{\varepsilon}_i \stackrel{i.i.d}{\sim} \mathsf{N}(0, \sigma^2)$, the likelihood $\mathcal{L}$ can be computed by letting $\boldsymbol{\Sigma}_V^t := (\boldsymbol{\nu}_\Theta^t)^\top \boldsymbol{\rho}^{-1} \boldsymbol{\nu}_\Theta^t + \boldsymbol{\kappa}_\Theta^{t,t}$,

$$\mu_\ell \equiv \mu_\ell(\boldsymbol{V}_i) := \left(\mathbb{E}[\boldsymbol{Z}_1 | (\boldsymbol{V}_\Theta^t)_1 = \boldsymbol{V}_i]\right)_\ell, \qquad \sigma_\ell \equiv \sigma_\ell(\boldsymbol{V}_i) := \sqrt{\left[\mathrm{Cov}(\boldsymbol{Z}_1 | (\boldsymbol{V}_\Theta^t)_1 = \boldsymbol{V}_i)\right]_{\ell,\ell}},$$

in (100), (102), (108).

Appendix D. Proof of Propositions 3 and 4

D.1 Proof of Proposition 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 $\boldsymbol{\Psi}$, that is, $\boldsymbol{\eta} = U^{-1}(\boldsymbol{\Psi})$, and that for $i \in [n]$ the distance of η_i to a change point estimate $\hat{\boldsymbol{\eta}} \in [n]^{L-1}$ is defined as $d(\eta_i, \{\hat{\eta}_j\}_{j=1}^{L-1}) := \min_{\hat{\eta}_j \in \{\hat{\eta}_j\}_{j=1}^{L-1}} \|\eta_i - \hat{\eta}_j\|_2$. We then 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 $(\boldsymbol{\Theta}^t, \boldsymbol{XB}) \mapsto d_H(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, q(\boldsymbol{XB}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))/n$ is uniformly pseudo-Lipschitz. Consider two sets of inputs, $\boldsymbol{A}^{(1)} := \left[(\boldsymbol{\Theta}^t)^{(1)} \quad (\boldsymbol{XB})^{(1)}\right] \in \mathbb{R}^{n \times 2L}$ and $\boldsymbol{A}^{(2)} := \left[(\boldsymbol{\Theta}^t)^{(2)} \quad (\boldsymbol{XB})^{(2)}\right] \in \mathbb{R}^{n \times 2L}$, and let $\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(1)}) := \hat{\boldsymbol{\eta}}\left((\boldsymbol{\Theta}^t)^{(1)}, q((\boldsymbol{XB})^{(1)}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})\right), \hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(2)}) := \hat{\boldsymbol{\eta}}\left((\boldsymbol{\Theta}^t)^{(2)}, q((\boldsymbol{XB})^{(2)}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})\right)$. We then have that:

$$\frac{1}{n}\left| d_H(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(1)})) - d_H(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(2)})) \right|$$

$$\leq \frac{1}{n} d_H(\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(1)}), \hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(2)})) \tag{113}$$

$$\leq \frac{1}{n} \|U(\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(1)})) - U(\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(2)}))\|_F^2 \tag{114}$$

$$\leq L \cdot \frac{1}{\sqrt{n}} \|U(\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(1)})) - U(\hat{\boldsymbol{\eta}}(\boldsymbol{A}^{(2)}))\|_F \tag{115}$$

$$\leq L \cdot C \left(1 + \left(\frac{\|\boldsymbol{A}^{(1)}\|_F}{\sqrt{n}}\right)^{r-1} + \left(\frac{\|\boldsymbol{A}^{(2)}\|_F}{\sqrt{n}}\right)^{r-1}\right) \frac{\|\boldsymbol{A}^{(1)} - \boldsymbol{A}^{(2)}\|_F}{\sqrt{n}}, \tag{116}$$

for some constants $C > 0, r \geq 1$. Here (113) follows from the reverse triangle inequality for the metric d_H, (114) follows from the argument in the paragraph above, (115) follows from the fact that $\mathcal{X} \subseteq [L]^n$, and (116) follows from the pseudo-Lipschitz assumption on $(\boldsymbol{V}, \boldsymbol{z}) \mapsto U(\hat{\boldsymbol{\eta}}(\boldsymbol{V}, q(\boldsymbol{z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))$. Applying Theorem 1 with

$$\boldsymbol{\varphi}_n(\boldsymbol{\Theta}^t, \boldsymbol{XB}; \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) := \frac{1}{n} d_H(U^{-1}(\boldsymbol{\Psi}), \hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, q(\boldsymbol{XB}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})),$$

we obtain the first result in Proposition 3. The result (32) follows directly by applying Theorem 1 to $\boldsymbol{\varphi}_n(\boldsymbol{\Theta}^t, \boldsymbol{XB}; \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) := |\hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, q(\boldsymbol{XB}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))|$. $\blacksquare$

D.2 Proof of Proposition 4

Proof Let $\varphi_n(\boldsymbol{V}, \boldsymbol{z}; \boldsymbol{\Psi}, \boldsymbol{\varepsilon}) := p(\boldsymbol{\psi}|\boldsymbol{V}, q(\boldsymbol{z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))$. By assumption, φ_n is uniformly pseudo-Lipschitz with respect to $(\boldsymbol{V}, \boldsymbol{z})$. Applying Theorem 1 to φ_n, we obtain that:

$$p(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y}) = p(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, q(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})) \stackrel{\mathbb{P}}{\simeq} \mathbb{E}_{\boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{Z}}\, p(\boldsymbol{\psi}|\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})) =: \mu_p.$$

Similarly, we apply Lemma 19 from (Gerbelot and Berthier, 2023), a concentration result regarding pseudo-Lipschitz functions of Gaussian random variables, to $\varphi_n(\boldsymbol{V}_{\boldsymbol{\Theta}}^t, \boldsymbol{Z}; \boldsymbol{\Psi}, \boldsymbol{\varepsilon})$, to obtain $p(\boldsymbol{\psi}|\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})) \stackrel{\mathbb{P}}{\simeq} \mu_p$. Combining the aforementioned results, we obtain that for any $\epsilon > 0$:

$$\begin{aligned}
&\mathbb{P}[|p(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y}) - p(\boldsymbol{\psi}|\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}))| > \epsilon] \\
&\leq \mathbb{P}[|p(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y}) - \mu_p| > \epsilon/2] + \mathbb{P}[|p(\boldsymbol{\psi}|\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon})) - \mu_p| > \epsilon/2] \to 0.
\end{aligned}$$

∎

Appendix E. Further Implementation and Experiment Details

State Evolution Implementation Our state evolution implementation involves computing (64)–(67). 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 (64)–(67) 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:

$$\boldsymbol{\nu}_{\boldsymbol{B}}^{t+1} \approx \sum_{\ell=0}^{L-1} \frac{1}{n} \sum_{i \in [\eta_\ell, \eta_{\ell+1})} \hat{\mathbb{E}}\left[\partial_1 \tilde{g}_i^t(\boldsymbol{Z}_1, (\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, \Psi_{\eta_\ell}, \bar{\varepsilon})\right], \tag{117}$$

$$\boldsymbol{\kappa}_{\boldsymbol{B}}^{t+1,t+1} \approx \sum_{\ell=0}^{L-1} \frac{1}{n} \sum_{i \in [\eta_\ell, \eta_{\ell+1})} \hat{\mathbb{E}}\left[g_i^t\left((\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, q_1(\boldsymbol{Z}_1, \Psi_{\eta_\ell}, \bar{\varepsilon})\right)^\top g_i^t\left((\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1, q_1(\boldsymbol{Z}_1, \Psi_{\eta_\ell}, \bar{\varepsilon})\right)\right], \tag{118}$$

$$\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1} \approx \frac{1}{\delta} \hat{\mathbb{E}}\left[\bar{\boldsymbol{B}} f_1^{t+1}(\boldsymbol{V}_{\boldsymbol{B}}^{t+1})^\top\right], \tag{119}$$

$$\boldsymbol{\kappa}_{\boldsymbol{\Theta}}^{t+1,t+1} \approx \frac{1}{\delta} \hat{\mathbb{E}}\left[\left(f_1^{t+1}(\boldsymbol{V}_{\boldsymbol{B}}^{t+1}) - \bar{\boldsymbol{B}}\boldsymbol{\rho}^{-1}\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1}\right)\left(f^{t+1}(\boldsymbol{V}_{\boldsymbol{B}}^{t+1}) - \bar{\boldsymbol{B}}\boldsymbol{\rho}^{-1}\boldsymbol{\nu}_{\boldsymbol{\Theta}}^{t+1}\right)^\top\right]. \tag{120}$$

where $\hat{\mathbb{E}}$ denotes an expectation estimate via Monte Carlo. For example, in the case of (118), 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 (7), that is, $(\boldsymbol{V}_{\boldsymbol{\Theta}}^t)_1 = \boldsymbol{Z}_1 \boldsymbol{\rho}^{-1} \boldsymbol{\nu}_{\boldsymbol{\Theta}}^t + (\boldsymbol{G}_{\boldsymbol{\Theta}}^t)_1$. 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 (117) and (118) by averaging. The $\hat{\mathbb{E}}$ terms in (119)–(120) are similarly computed.

The ensemble state evolution recursion (20)–(23) is used to compute the optimal denoisers $\{g^{*t}, f^{*t+1}\}_{t \geq 0}$ according to Proposition 2. The expectation in the ensemble state evolution iterates (20)–(23) are estimated through sample averages and AD, similar to the simplified state evolution iterates.
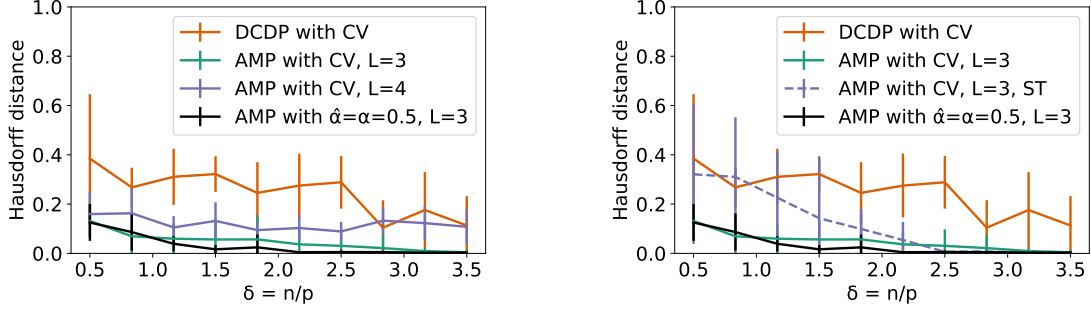
(a) Solid purple: AMP using $L = 3, 4$ and the optimal denoiser f^t with the sparsity level estimated using CV.

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

Figure 8: Comparison between AMP and DCDP in the experimental setting of Figure 3a. 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.
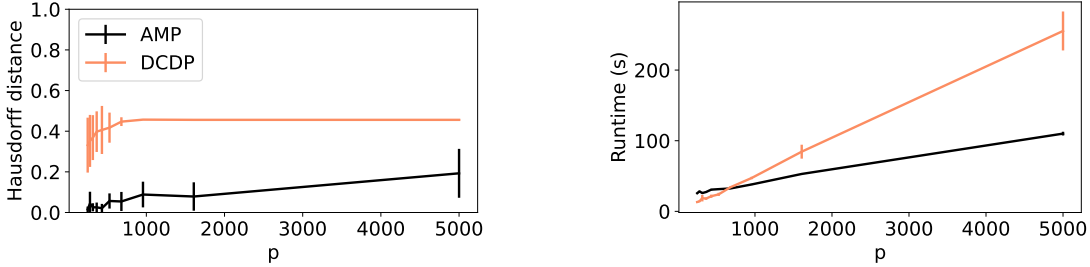
Figure 9: AMP vs. DCDP for fixed $n = 500$ and varying p, with $\sigma = 0.1$, $L^* = 3$ and sparse signal prior $\mathbb{P}_{\bar{B}} = 0.5\mathsf{N}(\mathbf{0}, \delta\mathbf{I}) + 0.5\delta_{\mathbf{0}}$. AMP uses $\Delta = n/10$ and $L = 3$.

Additional Numerical Results Figure 8 shows results from an additional set of experiments comparing AMP with DCDP (the fastest algorithm in Figure 3a) for the linear model, using different change point priors $\pi_{\bar{\Psi}}$ (Figure 8a) or denoisers f^t (Figure 8b). 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 8a, 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 8b, AMP performs slightly worse at lower δ when using a suboptimal soft thresholding (ST) denoiser. Nevertheless, in both Figures 8a and 8b, AMP largely outperforms DCDP despite the suboptimal choices of prior or denoiser.

Figure 9 compares AMP with `DCDP` for $n = 500$ and varying p, both using hyperparameters chosen via cross-validation. 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.

Experiment Details   For synthetic data experiments in Section 4, 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 (20)–(23). Figures 3a, 4, 6, 8 and 9 use Bernoulli-Gaussian priors taking the form $\mathbb{P}_{\bar{\boldsymbol{B}}} = \alpha\mathsf{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}) + (1 - \alpha)\delta_{\boldsymbol{0}}$ with $\alpha \in (0, 1)$. Figure 4 uses $\alpha = 1/6$ and $\sigma = 2.5$. Figures 3a, 6, 8 and 9 use $\alpha = 0.5$ and $\sigma^2 = \delta$ to ensure that the signal power $\mathbb{E}[(\boldsymbol{X}_i)^\top\boldsymbol{\beta}^{(i)}]^2$ is held constant for varying δ, and they use two change points at $n/3$ and $8n/15$ as the ground truth. In Figures 3a, 8 and 9, 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.

For the experiments in Figure 3b, we consider signals with sparse changes between adjacent signals. The prior takes the following form:

$$\beta_j^{(\eta_0)} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \kappa^2)$$

$$\beta_j^{(\eta_\ell)} = \begin{cases} \beta_j^{(\eta_{\ell-1})}, & \text{with probability } 1 - \alpha \\ \nu(\beta_j^{(\eta_{\ell-1})} + w_\ell), & \text{with probability } \alpha \end{cases} , \quad \ell \in [L], \quad (121)$$

where $w_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma_w^2)$ creates the sparse change between adjacent signals, and $\nu := \sqrt{\frac{\kappa^2}{\kappa^2 + \sigma_w^2}}$ is a rescaling factor that ensures uniform signal magnitude $\mathbb{E}[(\beta_j^{(\eta_0)})^2] = \cdots = \mathbb{E}[(\beta_j^{(\eta_L)})^2] = \kappa^2$. 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 3b.

Appendix F. Computational Cost

Computing g^{*t}, f^{*t+1}   For $i \in [n]$, the function g_i^{*t} in Proposition 2 only depends on $\bar{\boldsymbol{\Psi}}$ in (88) through the marginal probability of $\bar{\Psi}_i$, that is, $\pi_{\bar{\Psi}_i} : \ell \mapsto \sum_{\{\boldsymbol{\psi}:\psi_i=\ell\}} \pi_{\bar{\boldsymbol{\Psi}}}(\boldsymbol{\psi})$. 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 (6), 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 2. For example, the arg max in (30) 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 and Exact Posteriors For $\boldsymbol{\Psi} \in \mathcal{X}$, the denominator in (33) only requires evaluating a polynomial number of terms, $\binom{n}{L-1}$. Therefore, choosing f^t, g^t as in Proposition 2, the computational complexity of computing the quantity $p_{\bar{\boldsymbol{\Psi}}|\bar{\boldsymbol{V}}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\varepsilon})}(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y})$ from (35) along with AMP is $O(npL + nL^3 + n^{L-1}) = O(np + n^{L-1})$. Further, assuming the expectations in (10)–(13) can be well-approximated in $O(n)$ time, $p_{\bar{\boldsymbol{\Psi}}|\boldsymbol{V}_{\boldsymbol{\Theta}}^t, q(\boldsymbol{Z}, \bar{\boldsymbol{\Psi}}, \bar{\varepsilon})}(\boldsymbol{\psi}|\boldsymbol{\Theta}^t, \boldsymbol{y})$ from (33) along with AMP can be computed in $O(np + n^{L-1})$ time.

Appendix G. Background on AMP for Generalized Linear Models

Here we review the AMP algorithm and its state evolution characterization for the standard Generalized Linear Model (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

$$\boldsymbol{y} = q(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{\varepsilon}) \in \mathbb{R}^{n \times L}, \tag{122}$$

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 (5) is the absence of the signal configuration vector $\boldsymbol{\Psi}$. The AMP algorithm for the GLM in (122) 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 and Venkataramanan (2023). For ease of exposition, we will make the following standard assumption (see (S0) on p.9): 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 (122) is the same as the one in (6), but due to the assumption above, we can take f^t, g^t to be separable, that is, $f^t : \mathbb{R}^L \to \mathbb{R}^L$ 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 (6) 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 (6) 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 and 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.

References

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

Gabriel Arpino, Xiaoqi Liu, and Julia Gontarek. AMP implementation for change point inference in high-dimensional regression. https://github.com/gabrielarpino/AMP_chgpt_reg, 2024a.

Gabriel Arpino, Xiaoqi Liu, and Ramji Venkataramanan. Inferring change points in high-dimensional linear regression via approximate message passing. In *International Conference on Machine Learning (ICML)*, volume 235, pages 1841–1864, 2024b.

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

Nahid Azad and Genevieve Lemay. Management of chronic heart failure in the older population. *Journal of Geriatric Cardiology: JGC*, 11(4):329, 2014.

Yue Bai and Abolfazl Safikhani. A unified framework for change point detection in high-dimensional linear models. *Statistica Sinica*, 2024.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Jean Barbier, Nicolas Macris, and Cynthia Rush. All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

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

Raphaël Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79, 2019.

Monika Bhattacharjee, Moulinath Banerjee, and George Michailidis. Change point estimation in a dynamic stochastic block model. *Journal of Machine Learning Research*, 21(107):1–59, 2020.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/jax-ml/jax`. http://github.com/jax-ml/jax.

Jerome V Braun, RK Braun, and Hans-Georg Müller. Multiple changepoint fitting via quasilikelihood, with application to dna sequence segmentation. *Biometrika*, 2000.

Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(2):475–507, 2014.

Haeran Cho, Tobias Kley, and Housen Li. Detection and inference of changes in high-dimensional linear regression with nonsparse structures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2025.

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

Ilias Diakonikolas, Jongho Park, and Christos Tzamos. ReLU regression with Massart noise. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

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

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

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

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

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

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

Youyi Fong, Chongzhi Di, and Sallie Permar. Change point testing in logistic regression models with interaction term: Change point testing in logistic regression models with interaction term. *Statistics in Medicine*, 34(9):1483–1494, 2015.

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

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

Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations. *Information and Inference: A Journal of the IMA*, 12(4):2562–2628, 2023.

S.E. Golovenkin, V.A. Shulman, D.A. Rossiev, P.A. Shesternya, S.Yu. Nikulina, Yu.V. Orlova, , and V.F. Voino-Yasenetsky. Myocardial infarction complications. UCI Machine Learning Repository, 2020. DOI: https://doi.org/10.24432/C53P5M.

Yuzi He, Keith A. Burghardt, and Kristina Lerman. Leveraging change point detection to discover natural experiments in data. *EPJ Data Science*, 2022.

Johannes Hofrichter. *Change Point Detection in Generalized Linear Models*. Phd thesis, Technische Universität Graz, 2007.

Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

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

Abhishek Kaul, Venkata K. Jandhyala, and Stergios B. Fotopoulos. An efficient two step algorithm for high dimensional change point regression models without grid search. *Journal of Machine Learning Research*, 20(111):1–40, 2019.

Florent Krzakala, Marc Mézard, Fran*c*cois Sausset, Yifan Sun, and Lenka Zdeborová. Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08009, 2012.

Pascal Lavergne. A Cauchy-Schwarz inequality for expectation of matrices. Discussion Papers dp08-07, Department of Economics, Simon Fraser University, 2008. URL `https://ideas.repec.org/p/sfu/sfudps/dp08-07.html`.

Sokbae Lee, Myung Hwan Seo, and Youngki Shin. The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(1):193–210, 2015.

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

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

Jie Li, Paul Fearnhead, Piotr Fryzlewicz, and Tengyao Wang. Automatic change-point detection in time series via deep learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2024.

Wanshan Li, Daren Wang, and Alessandro Rinaldo. Divide and conquer dynamic programming: An almost linear time change point detection methodology in high dimensions. In *International Conference on Machine Learning (ICML)*, volume 202, pages 20065–20148, 2023.

Bin Liu, Zhengling Qi, Xinsheng Zhang, and Yufeng Liu. Change point detection for high-dimensional linear models: a general tail-adaptive approach. *Statistica Sinica*, 2026.

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

Malte Londschien, Peter Bühlmann, and Solt Kovács. Random forests for change point detection. *Journal of Machine Learning Research*, 24(216):1–45, 2023.

Valentinian Lungu, Ioannis Papageorgiou, and Ioannis Kontoyiannis. Bayesian change-point detection via context-tree weighting. In *IEEE Information Theory Workshop (ITW)*, pages 125–130, 2022.

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

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

Oscar Hernan Madrid Padilla, Yi Yu, Daren Wang, and Alessandro Rinaldo. Optimal nonparametric multivariate change point detection and localization. *IEEE Transactions on Information Theory*, 68(3):1922–1944, 2022.

Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: statistical and computational phase transitions. In *Neural Information Processing Systems (NeurIPS)*, volume 33, pages 11071–11082, 2020.

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

Marco Mondelli and Ramji Venkataramanan. Approximate message passing with spectral initialization for generalized linear models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

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

D. B. Owen. A table of normal integrals. *Communications in Statistics - Simulation and Computation*, 9(4):389–419, 1980.

Parthe Pandit, Mojtaba Sahraee-Ardakan, Sundeep Rangan, Philip Schniter, and Alyson K. Fletcher. Inference with deep generative priors in high dimensions. *IEEE Journal on Selected Areas in Information Theory*, 1(1):336–347, 2020.

Roberto Pastor-Barriuso, Eliseo Guallar, and Josef Coresh. Transition models for change-point estimation in logistic regression. *Statistics in Medicine*, 22(7):1141–1162, 2003.

Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *IEEE International Symposium on Information Theory (ISIT)*, 2011.

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

Alessandro Rinaldo, Daren Wang, Qin Wen, Rebecca Willett, and Yi Yu. Localizing changes in high-dimensional regression models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130, pages 2089–2097, 2021.

Riccardo Rossetti and Galen Reeves. Approximate message passing for the matrix tensor product model. arXiv:2306.15580, 2023.

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

Mahdi Soltanolkotabi. Learning ReLUs via gradient descent. In *Neural Information Processing Systems (NeurIPS)*, 2017.

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

Gideon Y. Stein, Angelika Kremer, Tzipi Shochat, Tamir Bental, Roman Korenfeld, Evgeny Abramson, Tuvia Ben-Gal, Alexander Sagie, and Shmuel Fuchs. The Diversity of Heart Failure in a Hospitalized Population: The Role of Age. *Journal of Cardiac Failure*, 18(8):645–653, 2012.

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

Nelvin Tan and Ramji Venkataramanan. Mixed regression via approximate message passing. *Journal of Machine Learning Research*, 24(317):1–44, 2023.

Daria Tieplova, Samriddha Lahiry, and Jean Barbier. Information-theoretic limits and approximate message-passing for high-dimensional time series. arXiv:2501.13625, 2025.

Jasper Tromp, Samantha M A Paniagua, Emily S Lau, Norrina B Allen, Michael J Blaha, Ron T Gansevoort, Hans L Hillege, Douglas E Lee, Daniel Levy, Ramachandran S Vasan, Pim Van Der Harst, Wiek H Van Gilst, Martin G Larson, Sanjiv J Shah, Rudolf A De Boer, Carolyn S P Lam, and Jennifer E Ho. Age dependent associations of risk factors with heart failure: pooled population based cohort study. *BMJ*, page n461, 2021.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Univariate mean change point detection: penalization, CUSUM and optimality. *Electronic Journal of Statistics*, pages 1917 – 1961, 2020.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal change point detection and localization in sparse dynamic networks. *The Annals of Statistics*, 49(1):203 – 232, 2021a.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal covariance change point localization in high dimensions. *Bernoulli*, 27(1):554–575, 2021b.

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

Tengyao Wang and Richard J. Samworth. High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):57–83, 2017.

Tianhao Wang, Xinyi Zhong, and Zhou Fan. Universality of approximate message passing algorithms and tensor networks. *The Annals of Applied Probability*, 34(4):3943–3994, 2024.

Xianru Wang, Bin Liu, Xinsheng Zhang, Yufeng Liu, and for the Alzheimer's Disease Neuroimaging Initiative. Efficient multiple change point detection for high-dimensional generalized linear models. *Canadian Journal of Statistics*, 51, 2023.

Haotian Xu, Daren Wang, Zifeng Zhao, and Yi Yu. Change-point inference in high-dimensional regression models under temporal dependence. *The Annals of Statistics*, 52(3):999–1026, 2024.

Zhi Yang, Liwen Zhang, Siyu Sun, and Bin Liu. Robust change point detection for high-dimensional linear models with tolerance for outliers and heavy tails. *Canadian Journal of Statistics*, 53(1): e11826, 2025.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning (ICML)*, 2014.

Yihan Zhang, Marco Mondelli, and Ramji Venkataramanan. Precise asymptotics for spectral methods in mixed generalized linear models. arXiv:2211.11368, 2022.

Yihan Zhang, Hong Chang Ji, Ramji Venkataramanan, and Marco Mondelli. Spectral estimators for structured generalized linear models via approximate message passing. In *Conference on Learning Theory (COLT)*, 2024.

Qian Zhao, Pragya Sur, and Emmanuel J. Candès. The asymptotic distribution of the MLE in high-dimensional logistic models: Arbitrary covariance. *Bernoulli*, 28(3):1835–1861, 2022.