Sparse Estimation: An MMSE Approach

Dedicated to Ron DeVore on the occasion of his 80th birthday

Tongyao Pang [*1] and Zuowei Shen [†1]

[1]Department of Mathematics, National University of Singapore, Singapore 119076

Abstract

The objective of this paper is to estimate parameters with a sparse prior via the minimum mean square error (MMSE) approach. We model the sparsity by the Bernoulli-uniform prior. The MMSE estimator gives the posterior mean of the parameter to be estimated. However, its computation involves multiple integrations of many variables that is hard to implement numerically.

In order to overcome this difficulty, we develop a coordinate minimization algorithm to approximate the MMSE estimator for any arbitrary given prior. We connect this algorithm to a variational model and establish a comprehensive convergence analysis. The algorithm converges to a special stationary point of the variational model, which attains the minimum of the mean square error at each coordinate when others are fixed. Then, this general algorithm is applied to the Bernoulli-uniform sparse prior and leads to a stable estimator that provides a good balance between sparsity and unbiasedness. The advantages of our sparsity model and algorithm over other approaches (e.g., the maximum *a posteriori* approaches) are analysed in detail and further demonstrated by numerical simulations. The applications of the general theory and algorithm developed here go beyond sparse estimation.

Keywords: minimum mean square error, Bayesian method, coordinate minimization, variational model, nonconvex optimization

1 Introduction

This paper develops an algorithm as well as its associated model and theory on Bayesian estimation. Consider a linear regression model:

$$\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{1.1}$$

[*]matpt@nus.edu.sg

[†]matzuows@nus.edu.sg

where $\boldsymbol{z} \in \mathbb{R}^n$ is the observation, $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ is the design matrix, $\boldsymbol{x} \in \mathbb{R}^p$ is the parameter of interest, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the Gaussian noise with mean zero and variance $\sigma^2\boldsymbol{I}$. The objective of linear regression is to estimate the parameter $\boldsymbol{x}$ from the given observation $\boldsymbol{z}$.

In Bayesian statistics, the underlying truth is represented as a random variable that follows some prior distribution. In our paper, we assume that x_i $(i = 1, 2, \cdots, p)$ are independent and identically distributed (i.i.d.) after some change of coordinates, following the distribution $p_{\mathrm{x}}(x)$. The sparse prior is popular in regression models as it helps to reduce the model complexity. It is also used in many other scenarios beyond regression, such as, compressed sensing [10, 16], image analysis and restorations [6, 9]. One of the key applications of sparse prior model is to purse a sparse approximation of the solution of (1.1). The theoretic base of sparse approximation is the non linear approximation or N-term approximation (see e.g. [13, 14]).

The sparse prior distribution chosen in our paper takes the form of

$$p_{\mathrm{x}}(x) = p_0\delta(x) + \frac{1-p_0}{2(U-L)}\mathbb{I}_{[-U,-L]\cup[L,U]}(x), \tag{1.2}$$

where $\delta(x)$ stands for the Dirac-delta function, and $\mathbb{I}_S(x)$ means the indicator function of the set S. It is a mixture distribution that consists of a discrete and continuous part, i.e. a Bernoulli distribution and an uniform distribution on $[-U,-L]\cup[L,U]$, and called the Bernoulli-uniform sparse prior. The Bernoulli-mixture prior has been used for modelling sparsity in many literatures. The Bernoulli-Gaussian sparse prior is studied by [30] and the Bernoulli-Laplace as well as Bernoulli-Cauchy sparse prior by [27]. The Bernoulli-mixture prior is also referred to as the spike and slab prior in some literatures, e.g. [25]. For the Bernoulli-uniform sparse prior, a larger value of p_0 represents a sparser assumption on the truth. The uniform distribution on $[-U,-L]\cup[L,U]$ reflects a high degree of uncertainty in the non-zero values and the boundedness of the truth, which is the case for the majority of applications. Finally, when $L > 0$, it means that the non-zero coefficients are significant to some degree.

Now, our aim is to find an efficient estimator $\mathcal{T}(\boldsymbol{z}) : \mathbb{R}^n \to \mathbb{R}^p$, for the parameter $\boldsymbol{x}$ from the contaminated observation $\boldsymbol{z}$. Minimizing different Bayesian risk functions results in a variety of estimators. We will concentrate on the minimum mean square error (MMSE) estimator which minimizes a quadratic cost function:

$$\mathcal{T}^{\mathrm{MMSE}} = \arg\min_{g:\mathbb{R}^n\to\mathbb{R}^p} \mathbb{E}_{(\mathbf{x},\mathbf{z})}\|\boldsymbol{x} - g(\boldsymbol{z})\|_2^2. \tag{1.3}$$

The MMSE estimator gains its popularity since it is stable with respect to the observation and balances the trade-off between variance and bias well. In addition, the MMSE estimator is given in an explicit form that is the mean of the posterior distribution:

$$\mathcal{T}^{\mathrm{MMSE}}(\boldsymbol{z}) = \mathbb{E}_{\mathbf{x}}(\boldsymbol{x}|\boldsymbol{z}) = \frac{\int \boldsymbol{x}\prod_{j=1}^p p(x_j)\exp(-\frac{\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z}\|_2^2}{2\sigma^2})d\boldsymbol{x}}{\int \prod_{j=1}^p p(x_j)\exp(-\frac{\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z}\|_2^2}{2\sigma^2})d\boldsymbol{x}}. \tag{1.4}$$

However, the posterior mean is usually difficult to compute due to the involvement of multiple integrals of many variables. In this paper, we propose an iterative algorithm to approximate it. We start from the univariate case

$$z = x + \varepsilon, \tag{1.5}$$

where $x, z \in \mathbb{R}$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The corresponding univariate MMSE estimator is given by

$$\mathcal{S}_\sigma^{\mathrm{MMSE}} = \mathbb{E}_{\mathrm{x}}(x|z), \tag{1.6}$$

which only involves single integrals and is easy to compute. The univariate MMSE estimator can be viewed as a thresholding operator under the Bernoulli-uniform sparse prior; see Figure 1 for an illustration. Furthermore, when the columns of $\boldsymbol{A}$ are orthogonal, the MMSE estimator (1.4) is separable and reduces to the univariate case:

$$\begin{aligned}
\hat{x}_i^{\mathrm{MMSE}} &= \frac{\int x_i \prod_{j=1}^p p(x_j)\exp(-\frac{\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z}\|_2^2}{2\sigma^2})d\boldsymbol{x}}{\int \prod_{j=1}^p p(x_j)\exp(-\frac{\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z}\|_2^2}{2\sigma^2})d\boldsymbol{x}} \\
&= \frac{\int x_i \prod_{j=1}^p \left\{p(x_j)\exp(-\frac{(x_j-\boldsymbol{a}_j^{\mathrm{T}}\boldsymbol{z}/\|\boldsymbol{a}_j\|_2^2)^2}{2\sigma^2/\|\boldsymbol{a}_j\|_2^2})\right\}d\boldsymbol{x}}{\int \prod_{j=1}^p \left\{p(x_j)\exp(-\frac{(x_j-\boldsymbol{a}_j^{\mathrm{T}}\boldsymbol{z}/\|\boldsymbol{a}_j\|_2^2)^2}{2\sigma^2/\|\boldsymbol{a}_j\|_2^2})\right\}d\boldsymbol{x}} \\
&= \frac{\int x_i p(x_i)\exp(-\frac{(x_i-\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{z}/\|\boldsymbol{a}_i\|_2^2)^2}{2\sigma^2/\|\boldsymbol{a}_i\|_2^2})dx_i}{\int p(x_i)\exp(-\frac{(x_i-\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{z}/\|\boldsymbol{a}_i\|_2^2)^2}{2\sigma^2/\|\boldsymbol{a}_i\|_2^2})dx_i} \\
&= \mathcal{S}_{\sigma/\|\boldsymbol{a}_i\|}^{\mathrm{MMSE}}(\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{z}/\|\boldsymbol{a}_i\|_2^2),\ i = 1, 2, \ldots, p,
\end{aligned}$$

which avoids multiple integrations and can be computed easily. However, in the general case, the MMSE estimator (1.4) is still hard to compute. Instead of calculating the involved multiple integrals directly, we convert the approximation of the MMSE estimator to the problem of finding the minimizer of

$$\min_{g:\mathbb{R}^n \to \mathbb{R}^p} \mathbb{E}_{(\mathbf{x},\mathbf{z})}\|\boldsymbol{x} - g(\boldsymbol{z})\|_2^2. \tag{1.7}$$

Next, we develop an iterative algorithm that only refines one component each time to reduce the mean square error in (1.7). Given an index i_k at iteration k, we minimize the mean square error along the direction of $x_{i_k}^k$ by fixing the remaining components:

$$\begin{aligned}
x_{i_k}^{k+1} &= \arg\min_{g:\mathbb{R}^n \to \mathbb{R}} \mathbb{E}_{(x_{i_k},\mathbf{z}|x_{j\neq i_k}^k)}(x_{i_k} - g(\boldsymbol{z}))^2 \\
&= \mathbb{E}_{x_{i_k}}(x_{i_k}|\boldsymbol{z}, x_{j\neq i_k}^k) = \mathcal{S}_{\sigma/\|\boldsymbol{a}_{i_k}\|_2}^{\mathrm{MMSE}}(\boldsymbol{a}_{i_k}^{\mathrm{T}}(\boldsymbol{z} - \sum_{j\neq i_k} \boldsymbol{a}_j x_j^k)/\|\boldsymbol{a}_{i_k}\|_2^2),
\end{aligned} \tag{1.8}$$

where $\boldsymbol{a}_j$ is the j-th column of matrix $\boldsymbol{A}$. This iterative algorithm is called the coordinate minimization algorithm in optimization. For the Bernoulli-uniform

sparse prior, it performs a thresholding operation at each iteration, which only involves single integration and is easy to implement.

The remaining question is whether this algorithm converges. It is hard to analyze the convergence of the coordinate minimization algorithm from the perspective of minimizing mean square error in (1.7). The univariate MMSE estimator is shown to be the proximity operator of a function $\varphi_\sigma^{\mathrm{MMSE}}(x)$ (see [22, 24, 23]), i.e., the solution of a penalized least-squares variational problem:

$$\mathcal{S}_\sigma^{\mathrm{MMSE}}(z) = \mathrm{Prox}_{\varphi_\sigma^{\mathrm{MMSE}}}(z) =: \arg\min_{x\in\mathbb{R}} \frac{1}{2}(x-z)^2 + \varphi_\sigma^{\mathrm{MMSE}}(x). \tag{1.9}$$

Note that the hard and soft thresholding are also proximal operators for the ℓ_0-norm and ℓ_1-norm penalized variational model respectively. Motivated by (1.9), we use the following variational model

$$\min_{\boldsymbol{x}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z}\|_2^2 + \sum_{i=1}^{p} \|\boldsymbol{a}_i\|_2^2 \varphi_{\sigma/\|\boldsymbol{a}_i\|_2}^{\mathrm{MMSE}}(x_i) \tag{1.10}$$

to analyze the convergence of our algorithm. As we will show, the coordinate minimization algorithm for solving (1.10) is the same as the iteration (1.8). Using the available theory developed in the field of optimization recently, the variational model (1.10) helps to prove the convergence of (1.8). Basically, we will prove that the sequences generated by (1.8) converges to a stationary point of the variational model (1.10). Moreover, it attains the minimum of the mean square error in (1.7) along each coordinate when the others are fixed. Specifically, the algorithm converges to the MMSE estimator when the columns of $\boldsymbol{A}$ are orthogonal.

Another way for a computable Bayesian estimate is to find the maximum *a posteriori* (MAP) point estimate instead:

$$\boldsymbol{x}^{\mathrm{MAP}} = \mathcal{T}^{\mathrm{MAP}}(z) := \arg\max_{\boldsymbol{x}\in\mathbb{R}^p} p(\boldsymbol{x}|\boldsymbol{z}) = \arg\max_{\boldsymbol{x}\in\mathbb{R}^p} \frac{p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{z})} \tag{1.11}$$

$$= \arg\max_{\boldsymbol{x}\in\mathbb{R}^p} \log p(\boldsymbol{z}|\boldsymbol{x}) + \log p(\boldsymbol{x}) \tag{1.12}$$

$$= \arg\min_{\boldsymbol{x}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z}\|_2^2 + \varphi(\boldsymbol{x}), \tag{1.13}$$

where $\varphi(\boldsymbol{x}) = -\sigma^2 \log p(\boldsymbol{x})$ is called the penalty function. The discarding of the term $p(\boldsymbol{z})$ is because $\boldsymbol{z}$ has been observed and thus $p(\boldsymbol{z})$ is constant. Actually, the MAP estimator $\mathcal{T}^{\mathrm{MAP}}$ can also be treated as

$$\mathcal{T}^{\mathrm{MAP}} = \arg\min_{g:\mathbb{R}^n\to\mathbb{R}^p} \mathbb{E}_{(\mathbf{x},\mathbf{z})}\big(-\delta(\boldsymbol{x}-g(\mathbf{z}))\big). \tag{1.14}$$

That is, the MAP estimator results from a minus delta loss function while the MMSE estimator from a quadratic loss function. The MAP estimator may be unstable for some prior as it gives a single point estimate while the MMSE estimator is more stable as it averages all potential points under the posterior

distribution. The advantage of MAP method is the existence of the variational model (1.13). Benefiting from that, the MAP estimator can be solved by optimization methods, such as coordinate descent method [21], proximal gradient method [1] and linearized Bregman method [9, 8]. All these popular algorithms to find sparse solutions of (1.1) are all essentially iterative thresholding algorithms. The thresholding keeps the solution sparse. The choice of thresholding operators is crucial to the success of the algorithms. The first iterative thresholding algorithm for high resolution image reconstruction was proposed in [11].

By relaxing the rigorousness on the treatment of Dirac-delta functions (which will be shown in section 2.3.3), the MAP estimator under the Bernoulli-uniform sparse prior for the univariate problem (1.5) is the simple hard-thresholding. As pointed out by many (see e.g. [19]), the hard-thresholding suffers from the drawback of instability. To remedy this drawback, the continuous soft-thresholding [15] is proposed. However, it introduces bias for large observations. The hard-thresholding is unbiased but unstable, while the soft one is stable but biased. Consequently, many efforts have been made to combine the advantages of the hard and soft thresholding while avoid their disadvantages. The SCAD thresholding [19] and MCP thresholding [35] are two examples of these efforts. The key idea is to construct thresholding by interpolating the hard and soft thresholding so that it is stable and unbiased for large observations. The constructed thresholding is raised to a variational model that can be interpreted as a MAP model with a different prior from the Bernoulli-uniform sparse prior. The SCAD and MCP approaches are to refine the discontinuous point of the hard-thresholding to balance "sparsity", "continuity/stability" and "unbiasedness". However, it is unclear in which sense an optimal balanced point between the sparsity and unbiasedness can be reached for given continuity of the thresholding in these two approaches. It is interesting to find a mathematical formulation for an optimal solution to smooth the hard thresholding operator.

To exploit the balance between "sparsity " and "unbiasedness" for given continuity of the thresholding operator, it requires to reduce the variance and bias. The reason is that for small observations, thresholding is desired as it reduces the variance without increasing bias too much, while for large observations, it achieves better balance between variance and bias by keeping them due to the high signal to noise ratios . Minimizing the variance and bias leads us to use the MMSE estimator as the thresholding operator for the Bernoulli-uniform sparse prior. This is the best thresholding operator smoothing the hard thresholding in the sense that it minimizes the sum of variance and squared bias. Compared to SCAD and MCP, the MMSE estimator has a clear optimization goal for the variance and bias balance. See Figure 1 for the comparison of these thresholding operators.

For the general problem (1.1) with the sparse prior, all of SCAD, MCP and our method adopt iterative thresholding algorithm relating to the corresponding variational models. We will show that our algorithm (1.8) always converges to an estimate that attains the minimum of the mean square error along single coordinate while the iterative algorithms for the SCAD and MCP variational model may diverge as their models are not coercive. Even if they converge, it is

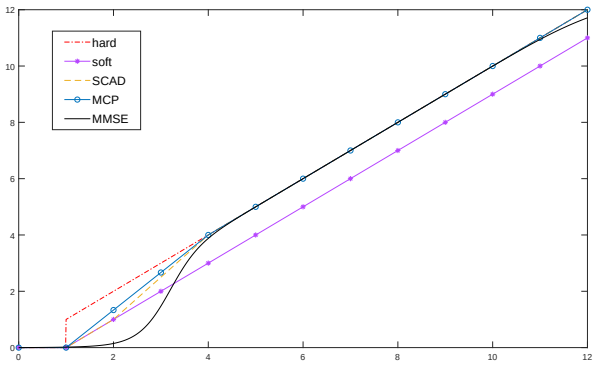

Figure 1: Comparison of different thresholding operators.

still unclear how the obtained estimates are connected to the goal of maximizing the original posterior probability density under the Bernoulli-uniform sparse prior. Finally, the numerical results demonstrate the advantages of our method over SCAD and MCP methods in terms of various criteria.

Although there are other works, e.g. [27], that also use the posterior mean or median estimator with independent observations under the Bernoulli-mixture prior, our emphasis is developing an iterative algorithm with convergence analysis for the general multi-variate case where the observations are correlated. The approaches in [12, 32] are to separate the estimation of the support and the non-zero values. Once the support has been selected correctly, the estimation of the non-zero values can be obtained easily via either MAP or MMSE estimator. However, the support selection is as hard as the sparse approximation problem itself and the algorithms developed usually do not have convergence guarantee. Some variants of support selection and applications in image processing can be found in [7, 26].

In summary, we propose an optimal thresholding operator, that smooths the hard-thresholding, in the sense of minimizing the summation of variance and squared bias by using the univariate MMSE estimator under the Bernoulli-uniform prior. We develop an iterative algorithm to approximate the MMSE estimator for solving (1.1) and prove the convergence. Furthermore, the limiting point attains the minimum of the mean square error along each coordinate when the others are fixed. The developed algorithm and analysis is also applicable to a wide variety of prior distributions. In section 2.3, we will study two additional distributions besides the Bernoulli-uniform prior, namely, the Gaussian prior and the Gaussian mixture prior. The Gaussian prior is commonly-used since it is simple and natural. The Gaussian mixture prior goes beyond the Gaussian prior and can approximate any continuous distribution with arbitrary accuracy.

The rest of the paper is organized as follows. In section 2, we first introduce

MMSE estimator from a univariate case and analyse its good properties. We next build up a connection between the univariate MMSE estimator and some variational model. At the end of this section, we give three examples of the prior distributions, i.e., the Gaussian, Gaussian mixture and Bernoulli-uniform sparse prior for illustration. In section 3, we propose to approximate the MMSE estimator by the cyclic coordinate minimization algorithm in the general case and give the main theorem that establishes the convergence of our algorithm. The numerical experiments follow to show the efficiency of our algorithm. The detailed proof of our main theorem is given in section 4. Finally, the paper concludes in section 5.

2 The Univariate MMSE estimator

For the general regression problem (1.1), the iterative scheme (1.8) approximates the MMSE estimator by performing a univariate MMSE estimation at each iteration. In order to analyse the convergence of (1.8), we ultimately need to understand the various properties of the univariate MMSE estimator and its corresponding variational model.

In our paper, we denote the random variable with a lower case letter in plain typeface (e.g., x is a random variable) and the values it can take on with lower case script letters (e.g., x_1 is a possible value of x). The probability density function for the random variable x is denoted by $p_{\mathrm{x}}(\cdot)$ and the subscript may be omitted for simplicity unless necessary. We use the notation $\delta(x)$ to denote the Dirac delta function and $\mathbb{I}_S(x)$ the indicator function of the set S.

2.1 The univariate MMSE estimator and its properties

To set up our platform, we assume the prior distribution to be a mixture of discrete and continuous distributions, with a non-zero weight on the continuous part. More specifically, it satisfies the following conditions.

Assumption 1. *Assume the probability distribution $p_{\mathrm{x}}(\cdot)$ is a mixture of a finite number of Diracs and an absolutely continuous distribution. By a slight but clear abuse of notations, we denote*

$$p_{\mathrm{x}}(x) = \sum_{i=1}^{d} w_i \delta(x - c_i) + \big(1 - \sum_{i=1}^{d} w_i\big) h(x), \tag{2.1}$$

where

$$w_i \geq 0, \ \sum_{i=1}^{d} w_i < 1,$$

and $h(x) : \mathbb{R} \to \mathbb{R}$ is a probability density function with respect to Lebesgue measure.

For the problem (1.5), when the prior distribution $p_{\mathrm{x}}(\cdot)$ satisfies Assumption 1, it is well known that the univariate MMSE estimator is (1.6) (see [28]). Since our theory and algorithm start from this basic fact, we summarize the derivation in the following.

The univariate MMSE estimator for (1.5) refers to the estimator which attains the minimum of the mean square error among all possible maps from $\mathbb{R}$ to $\mathbb{R}$:

$$\begin{aligned}\mathcal{S}_\sigma^{\mathrm{MMSE}}: \quad &= \arg\min_{g:\mathbb{R}\to\mathbb{R}} \mathbb{E}_{(\mathrm{x},\mathrm{z})}(x-g(z))^2 \\ &= \arg\min_{g:\mathbb{R}\to\mathbb{R}} \int\int (x-g(z))^2 p(x,z)dxdz \\ &= \arg\min_{g:\mathbb{R}\to\mathbb{R}} \int p_{\mathrm{z}}(z)\left(\int (x^2-2xg(z)+g(z)^2)p(x|z)dx\right)dz,\end{aligned} \tag{2.2}$$

where the subscript σ is the standard variance of the noise and $p_{\mathrm{z}} := p_{\mathrm{x}} * p_\epsilon$ denotes the probability density function of z. It is equivalent to finding $\mathcal{S}_\sigma^{\mathrm{MMSE}}(z) = \arg\min_{y\in\mathbb{R}} \int (x^2-2xy+y^2)p(x|z)dx$ for every observation z. As the noise follows the Gaussian distribution, the conditional distribution $p(x|z)$ takes the form of

$$p(x|z) = \frac{p_{\mathrm{x}}(x)\exp(-\frac{(x-z)^2}{2\sigma^2})}{\int p_{\mathrm{x}}(x)\exp(-\frac{(x-z)^2}{2\sigma^2})dx}.$$

When Assumption 1 is satisfied, the integral $\int x^n p(x|z)dx$ exists and is finite for any $n \in \mathbb{N}$. Then, by the dominated convergence theorem, we obtain

$$\begin{aligned}&\frac{d}{dy}\int (x^2-2xy+y^2)p(x|z)dx = 2\int (y-x)p(x|z)dx \\ &= 2y\int p(x|z)dx - 2\int xp(x|z)dx = 2y - 2\int xp(x|z)dx.\end{aligned}$$

Setting the differential to zero, we get a closed form of the univariate MMSE estimator:

$$\mathcal{S}_\sigma^{\mathrm{MMSE}}(z) = \mathbb{E}_{\mathrm{x}}(x|z) = \int xp(x|z)dx = \frac{\int xp_{\mathrm{x}}(x)\exp(-\frac{(x-z)^2}{2\sigma^2})dx}{\int p_{\mathrm{x}}(x)\exp(-\frac{(x-z)^2}{2\sigma^2})dx}. \tag{2.3}$$

The single integrals in (2.3) are easy to compute for every observed z, and so is the univariate MMSE estimator $\mathcal{S}_\sigma^{\mathrm{MMSE}}(z)$. In contrast, the MMSE estimator (1.4) for the general problem (1.1), which can be derived in a similar way as the univariate MMSE estimator, is much harder to compute numerically due to the involvement of the multiple integrals. Consequently, we propose an iterative algorithm (1.8) to approximate the MMSE estimator (1.4) by a sequence of univariate MMSE estimators. For the convergence analysis of the algorithm (1.8), we discuss the properties of the univariate MMSE estimator in the following.

The properties of the conditional mean, i.e. the MMSE estimator, have been studied in some literatures, e.g. [22, 24, 23]. While they mainly consider continuous prior distributions that admit probability density functions with respective to Lebesgue measure, some of their conclusions can be generalized to the case where Dirac-delta distributions are involved. We first list some

properties of the univariate conditional mean that have been proved in [22] and establish our analysis based on them.

Theorem 2.1 ([22]). *Consider $z = x+\epsilon$, where $\epsilon \sim N(0,\sigma^2)$. Assume x admits a Lebesgue measurable probability density function p_{x}.*

1. *The univariate MMSE estimator $\mathcal{S}_\sigma^{\mathrm{MMSE}}(\cdot)$ is one-to-one and C^∞ from $\mathbb{R}$ onto $Im\mathcal{S}_\sigma^{\mathrm{MMSE}}$. Its reciprocal $r_\sigma^{\mathrm{MMSE}}(\cdot) : Im\mathcal{S}_\sigma^{\mathrm{MMSE}} \to +\infty$ is also C^∞.*
2. $\frac{d\mathcal{S}_\sigma^{\mathrm{MMSE}}(x)}{dx} > 0$.
3. *For every $z \in \mathbb{R}$, the value $\mathcal{S}_\sigma^{\mathrm{MMSE}}(z)$ is the unique local minimum of the function $\frac{1}{2}\|z-x\|^2 + \varphi_\sigma^{\mathrm{MMSE}}(x)$, where*

$$\varphi_\sigma^{\mathrm{MMSE}}(x) := \begin{cases} -\frac{1}{2}(\nabla \log p_{\mathrm{z}}\big(r_\sigma^{\mathrm{MMSE}}(x)\big)^2 - \log p_{\mathrm{z}}\big(r_\sigma^{\mathrm{MMSE}}(x)\big), & x \in Im\mathcal{S}^{\mathrm{MMSE}}, \\ +\infty, & x \notin Im\mathcal{S}^{\mathrm{MMSE}}. \end{cases}$$

4. *If $\widetilde{\varphi}$ satisfies that $\mathcal{S}_\sigma^{\mathrm{MMSE}}(z) = \arg\min \frac{1}{2}\|z-x\|^2 + \widetilde{\varphi}(x)$, then there is a constant $c \in \mathbb{R}$ such that $\widetilde{\varphi}(x) = \varphi_\sigma^{\mathrm{MMSE}}(x) + c$ for all $x \in Im\mathcal{S}^{\mathrm{MMSE}}$.*

The first item in Theorem 2.1 indicates that the univariate MMSE estimator provides a stable estimation for the parameter of interest. To take it one step further, we prove that the univariate MMSE estimator and its reciprocal are analytic in the following proposition.

Proposition 2.1. *Consider $z = x + \epsilon$, where $\epsilon \sim N(0,\sigma^2)$. Assume* x *follows $p_{\mathrm{x}}(x)$ that satisfies Assumption 1. The univariate MMSE estimator $\mathcal{S}_\sigma^{\mathrm{MMSE}}(\cdot)$ is one-to-one and analytic. Its reciprocal $r_\sigma^{\mathrm{MMSE}}(\cdot) : Im\mathcal{S}_\sigma^{\mathrm{MMSE}} \to +\infty$ is also analytic.*

Proof. The proof of the one-to-one correspondence follows that of Theorem 2.1 in the literature [22], as the integrals involved in their proof still hold true for the mixture distribution in Assumption 1. Next, we prove the analyticity of $\mathcal{S}_\sigma^{\mathrm{MMSE}}(\cdot)$, and the analyticity of its reciprocal can then be obtained by the Lagrange inversion theorem.

Since the denominator of $\mathcal{S}_\sigma^{\mathrm{MMSE}}(z)$ in (2.3) is non-zero, it is enough to show both the numerator

$$\int x p_{\mathrm{x}}(x) \exp(-\frac{(x-z)^2}{2\sigma^2}) dx$$

and the denominator

$$\int p_{\mathrm{x}}(x) \exp(-\frac{(x-z)^2}{2\sigma^2}) dx$$

are analytic in terms of z. We only prove the numerator is analytic in the following, as the proof for the denominator is similar.

Let

$$f(z) = \int x p_{\mathrm{x}}(x) \exp(-\frac{(x-z)^2}{2\sigma^2}) dx.$$

It can be rewritten as

$$f(z) = \exp(-\frac{z^2}{2\sigma^2})h(z),\ h(z) = \int xp_{\mathrm{x}}(x)\exp(-\frac{x^2}{2\sigma^2})\exp(\frac{xz}{\sigma^2})dx.$$

The Taylor series for $\exp(\frac{xz}{\sigma^2})$ with respect to x is

$$\exp(\frac{xz}{\sigma^2}) = \sum_{k=1}^{+\infty} \frac{z^k}{\sigma^{2k}k!}x^k,$$

which converges for all $x \in \mathbb{R}$. Denote the first n terms of the Taylor series by $g_n(x) = \sum_{k=1}^{n} \frac{z^k}{\sigma^{2k}k!}x^k$. We have

$$\begin{aligned}|\exp(\frac{xz}{\sigma^2}) - g_n(x)| &= \Big|\sum_{k=n+1}^{\infty} \frac{z^k}{\sigma^{2k}k!}x^k\Big| \le \sum_{k=n+1}^{\infty} \frac{|z|^k}{\sigma^{2k}k!}|x|^k \\ &\le \sum_{k=1}^{\infty} \frac{|z|^k}{\sigma^{2k}k!}|x|^k = \exp(\frac{|xz|}{\sigma^2}).\end{aligned}$$

As $p_x(x)$ satisfies Assumption 1, the integral $\int xp_{\mathrm{x}}(x)\exp(-\frac{x^2}{2\sigma^2})\exp(\frac{|xz|}{\sigma^2})dx$ exists and is finite. By dominated convergence theorem, we can obtain

$$\begin{aligned}\int xp_{\mathrm{x}}(x)\exp(-\frac{x^2}{2\sigma^2})\exp(\frac{xz}{\sigma^2})dx &= \int xp_{\mathrm{x}}(x)\exp(-\frac{x^2}{2\sigma^2}) \lim_{n\to+\infty} g_n(x)dx \\ &= \lim_{n\to+\infty} \int xp_{\mathrm{x}}(x)\exp(-\frac{x^2}{2\sigma^2})g_n(x)dx \\ &= \sum_{k=1}^{+\infty} \Big(\frac{1}{\sigma^{2k}k!}\int x^{k+1}p_{\mathrm{x}}(x)\exp(-\frac{x^2}{2\sigma^2})dx\Big)z^k.\end{aligned}$$

for all $z \in \mathbb{R}$. By the definition of an analytic function, we have $h(z)$ is analytic. Therefore, $f(z)$ is also analytic. □

2.2 The MMSE variational model

The second item in Theorem 2.1 establishes a connection between the univariate MMSE estimator (2.3) and the variational model

$$\min_{x\in\mathbb{R}} \frac{1}{2}(x-z)^2 + \varphi_\sigma^{\mathrm{MMSE}}(x). \tag{2.4}$$

Variational models are widely used in many applications due to their numerical efficiency.

While we have already obtained the univariate MMSE estimator, its corresponding variational model is built up for the purpose of the convergence analysis of the iteration (1.8).

Furthermore, we prove that the objective function of the variational model (2.4), denoted by $h(x)$ in the following, is strongly convex on any compact set.

Proposition 2.2. *Assume $p_{\mathrm{x}}(x)$ satisfies Assumption 1. Let $\mathcal{C} \subseteq Im\mathcal{S}_\sigma^{\mathrm{MMSE}}$ be an arbitrary bounded convex set and $\widetilde{\mathcal{C}}$ be the range of r_σ^{MMSE} on set $\mathcal{C}$. The objective function of the univariate variational model (2.4) is Δ-strongly convex on $\mathcal{C}$, where $\Delta = 1/\sup_{x\in\widetilde{\mathcal{C}}}\{\frac{d\mathcal{S}_\sigma^{\mathrm{MMSE}}(x)}{dx}\}$.*

Proof. Denote the objective function of (2.4) as $h(x)$. Since $\mathcal{S}_\sigma^{\mathrm{MMSE}}(z)$ is the unique solution of (2.4), it holds that

$$\nabla h(\mathcal{S}_\sigma^{\mathrm{MMSE}}(z)) = \mathcal{S}_\sigma^{\mathrm{MMSE}}(z) - z + \nabla\varphi_\sigma^{\mathrm{MMSE}}\big(\mathcal{S}_\sigma^{\mathrm{MMSE}}(z)\big) = 0.$$

Substituting $x = \mathcal{S}_\sigma^{\mathrm{MMSE}}(z)$ into the above equation, we can obtain

$$\nabla\varphi_\sigma^{\mathrm{MMSE}}(x) = r_\sigma^{MMSE}(x) - x. \tag{2.5}$$

Denote $h_\Delta(x) = h(x) - \frac{\Delta}{2}x^2$. Then we have

$$h''_\Delta(x) = \frac{dr_\sigma^{MMSE}(x)}{dx} - \Delta \geq 0,$$

on $\mathcal{C}$. Since $\mathcal{S}_\sigma^{\mathrm{MMSE}}(x)$ is analytic, $\frac{d\mathcal{S}_\sigma^{\mathrm{MMSE}}(x)}{dx}$ is bounded on the bounded set $\widetilde{\mathcal{C}}$. On the other hand, it is also proved that $\frac{d\mathcal{S}_\sigma^{\mathrm{MMSE}}(x)}{dx} > 0$ in Theorem 2.1. Thus, Δ is positive. Then we have $h(x)$ is Δ-strongly convex on $\mathcal{C}$. □

Remark. According to (2.5), $\varphi_\sigma^{\mathrm{MMSE}}$ can also be rewritten in the form of

$$\varphi_\sigma^{\mathrm{MMSE}}(x) = \begin{cases} \int_0^x (r_\sigma^{\mathrm{MMSE}}(u) - u)du + c, & x \in Im\mathcal{S}^{\mathrm{MMSE}}, \\ +\infty, & x \notin Im\mathcal{S}^{\mathrm{MMSE}}. \end{cases} \tag{2.6}$$

for some constant $c \in \mathbb{R}$.

2.3 Examples

We focus on three examples with special prior, namely, the Gaussian, Gaussian mixture and Bernoulli-uniform sparse prior (1.2), which all satisfy Assumption 1. They are widely used in many applications. Besides the MAP estimator, another way to obtain a tractable Bayesian estimator is to constrain the estimator to be linear, that is, to minimize the mean square error among all possible linear maps. The resulting estimator is called the linear minimum mean square error (LMMSE) estimator. The LMMSE estimator for the problem (1.1) is summarized by Bayesian Gauss-Markov. In this section, We compare the MMSE estimator with the LMMSE and MAP estimator for the three examples.

2.3.1 The Gaussian prior

The first example is one of the most commonly used prior, the Gaussian prior distribution

$$p_{\mathrm{x}}(x) = \frac{1}{\sqrt{2\pi\sigma_s^2}}\exp(-\frac{(x-\mu_s)^2}{2\sigma_s^2}). \tag{2.7}$$

The posterior distribution $p(x|z)$ is also Gaussian, as claimed in [28]. In this form the univariate MMSE estimator is found as

$$\hat{x}^{\mathrm{MMSE}} = \alpha z + (1-\alpha)\mu_s,\ \alpha = \frac{\sigma_s^2}{\sigma_s^2 + \sigma^2}. \tag{2.8}$$

This estimation is also known as the Wiener filtering. As (2.8) is linear, the linear MMSE estimator coincides with the MMSE estimator. Moreover, since the mode of the Gaussian distribution is the same as its mean, the MAP estimator is also identical to the MMSE estimator.

2.3.2 The Gaussian mixture prior

The second example goes beyond the Gaussian setting and extends to the Gaussian mixture prior distribution

$$p_{\mathrm{x}}(x) = \sum_{i=1}^{k} p_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-\frac{(x-\mu_i)^2}{2\sigma_i^2}),\ \sum_{i=1}^{k} p_i = 1. \tag{2.9}$$

The Gaussian mixture prior is widely applicable due to several properties. Firstly, Gaussian mixture distributions can approximate any continuous density with arbitrary accuracy (see [33]). Secondly, the Gaussian mixture prior together with the Gaussian noise setting gives a Gaussian mixture posterior distribution. The MMSE estimator, i.e., the posterior mean, takes the form of

$$x^{\mathrm{MMSE}}(z) = \sum_{i=1}^{k} \beta_i(z)(\alpha_i z + (1-\alpha_i)\mu_i), \tag{2.10}$$

where

$$\alpha_i = \frac{\sigma_i^2}{\sigma_i^2 + \sigma^2},\ \beta_i(z) = \frac{p_i \exp(-\frac{(z-\mu_i)^2}{2(\sigma_i^2+\sigma^2)})/\sqrt{2\pi(\sigma_i^2+\sigma^2)}}{\sum_{i=1}^{k} p_i \exp(-\frac{(z-\mu_i)^2}{2(\sigma_i^2+\sigma^2)})/\sqrt{2\pi(\sigma_i^2+\sigma^2)}}.$$

By Bayesian Gauss-Markov theorem (Theorem 12.1 of [28]), the LMMSE estimator is

$$\boldsymbol{x}^{\mathrm{LMMSE}} = \hat{\alpha} z + (1-\hat{\alpha})\hat{\mu}, \tag{2.11}$$

where

$$\hat{\mu} = \sum_{i=1}^{k} p_i \mu_i,\ \hat{\sigma}^2 = \sum_{i=1}^{k} p_i \sigma_i^2 + \sum_{i=1}^{k} p_i \mu_i^2 - \hat{\mu}^2,\ \hat{\alpha} = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \sigma^2}. \tag{2.12}$$

Actually, $\hat{\mu}$ and $\hat{\sigma}^2$ are the mean and variance of (2.9). Comparing with (2.8), we find that the LMMSE estimator (2.11) behaves as if to minimize the mean square error under the assumption of the Gaussian prior with mean $\hat{\mu}$ and variance $\hat{\sigma}^2$.

Compared to the MMSE and LMMSE estimator, the MAP estimator has no explicit form and is not easy to compute. It seeks the maximum of the Gaussian mixture posterior probability density function while the Gaussian mixture distribution has multiple modes.

2.3.3 The Bernoulli-uniform sparse prior

The third example is the Bernoulli-uniform sparse prior given in (1.2). It assumes that most of the parameters can be set to zero without substantially affecting the fitting of model. Substituting the Bernoulli-uniform sparse prior and the Gaussian noise distribution into (2.2), we get an explicit form of the univariate MMSE estimator as follows

$$\mathcal{S}_{\sigma}^{\mathrm{MMSE}}(z) = \frac{\frac{1-p_0}{2(U-L)} \cdot \int_{[-U,-L]\cup[L,U]} \left(x\phi_{\sigma}(x-z)\right) dx}{p_0\phi_{\sigma}(z) + \frac{1-p_0}{2(U-L)} \cdot \int_{[-U,-L]\cup[L,U]} \phi_{\sigma}(x-z)dx},$$

where $\phi_{\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{x^2}{2\sigma^2})$. Let $\Psi_{\sigma}(t)$ denote the cumulative distribution function of the Gaussian distribution $\mathcal{N}(0,\sigma^2)$ and

$$C = \frac{2p_0(U-L)}{1-p_0}, \tag{2.13}$$

$$G_1 = \Psi_{\sigma}(U-z) + \Psi_{\sigma}(-L-z) - \Psi_{\sigma}(L-z) - \Psi_{\sigma}(-U-z), \tag{2.14}$$

$$G_2 = \sigma^2\big(\phi_{\sigma}(L-z) + \phi_{\sigma}(-U-z) - \phi_{\sigma}(U-z) - \phi_{\sigma}(-L-z)\big). \tag{2.15}$$

Then the univariate MMSE estimator can be rewritten as

$$\mathcal{S}_{\sigma}^{\mathrm{MMSE}}(z) = \frac{zG_1 + G_2}{C\phi_{\sigma}(z) + G_1}. \tag{2.16}$$

Figure 1 displays the univariate MMSE estimator obtained by (2.16).

As suggested by [19], a good sparse estimator should satisfy the following three conditions:

1. *continuity*: the estimator is continuous in data z to avoid instability in model prediction;
2. *unbiasedness*: the estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modeling bias;
3. *sparsity*: the estimator sets small estimated coefficients to zero to reduce model complexity.

The univariate MMSE estimator is a good sparse estimator. Firstly, the univariate MMSE estimator is analytic by Proposition 2.1. Secondly, it is unbiased since

$$\mathbb{E}_{\mathrm{z}}\mathcal{S}_{\sigma}^{\mathrm{MMSE}}(z) = \mathbb{E}_{\mathrm{z}}\mathbb{E}_{\mathrm{x}}(x|z) = x.$$

Thirdly, it is also sparse. Specifically, when z is very small, G_1 in (2.14) and G_2 in (2.15) is much smaller than the constant value C in (2.13); hence, the estimate is very close to zero.

Next, we compare the MMSE estimator with the LMMSE estimator. The LMMSE estimator under the Bernoulli-uniform sparse prior is a shrinkage operator

$$x^{\mathrm{LMMSE}} = \frac{\sigma_s^2}{\sigma^2 + \sigma_s^2} z, \tag{2.17}$$

where $\sigma_s = \frac{1}{3}(1-p_0)(L^2+U^2+LU)$ is the variance of the Bernoulli-uniform sparse prior. The LMMSE estimator (2.17) is continuous, but it is biased and loses the sparsity.

Finally, we compare the MMSE estimator with the MAP estimator. From (1.13), the MAP model for the univariate problem (1.5) with Gaussian noise takes the form of

$$\min_{x\in\mathbb{R}} \frac{1}{2}(x-z)^2 + \varphi(x), \tag{2.18}$$

where $\varphi(x) = -\sigma^2 \log p_{\mathrm{x}}(x)$. Note that the above MAP model is developed for continuous random variables and seeks the maximum of the posterior probability density function. However, the Bernoulli-uniform distribution density function is not continuous and has an infinite value at zero. Nevertheless, we relax the rigorousness in analysis and write $\varphi(x)$ formally by assuming $p(0)$ is arbitrarily large:

$$\varphi^{\ell_0}(x) = \begin{cases} a, & x = 0, \\ b, & x \in [-U,-L]\cup[L,U] \text{ and } x \neq 0, \\ +\infty, & \text{otherwise}, \end{cases} \tag{2.19}$$

where

$$a = -\sigma^2 \log p(0),\ b = -\sigma^2 \log \frac{1-p_0}{2(U-L)}.$$

The solution of (2.18) with $\varphi(x) = \varphi^{\ell_0}(x)$ can be solved formally. Since the Bernoulli-uniform sparse prior assumes the underlying truth is bounded, the MAP estimate for $z > U$ (resp. $z < -U$) is either 0 or U (resp. $-U$). In the following, we only consider the case $|z| \leq U$. Let $\lambda = \sqrt{2(b-a)} = \sigma\sqrt{2\log\frac{2p(0)(U-L)}{1-p_0}}$. When $\lambda > U$, the resulting MAP estimator for the Bernoulli-uniform sparse prior is always zero. When $L \leq \lambda \leq U$, the MAP estimator is the hard-thresholding

$$\boldsymbol{x}^{\mathrm{hard}} = \mathcal{S}_\lambda^h(z) := z\mathbb{I}_{\{|z|>\lambda\}}.$$

Since the hard-thresholding is a good sparse estimator, it justifies the validity of the Bernoulli-uniform sparse prior. In contrast, the Bernoulli-Gaussian prior, which is the combination of a Dirac-delta distribution and a Gaussian distribution, results in an additional ℓ_2-norm penalization on non-zero values and only gives the hard-thresholding estimator in the limiting case where the variance of the Gaussian distribution goes to infinity (see [34]). We note that one also has to relax the rigorousness when using the Bernoulli-Gaussian prior as a sparse prior for the MAP model. Strictly speaking, there is no prior distribution in the MAP setting corresponding to the hard thresholding.

The hard-thresholding is sparse and nearly unbiased for large coefficients, but it is discontinuous. As shown in (1.14), its discontinuity results from the instability of the minus delta loss function of the MAP model. To remedy the drawback of the hard-thresholding, the continuous soft-thresholding [18] is proposed:

$$\boldsymbol{x}^{\mathrm{soft}} = \mathcal{S}_\lambda^s(z) := \mathrm{sign}(z)(|z|-\lambda)_+,$$

which corresponds to the ℓ_1-norm penalty (the ℓ_1-norm penalty is the convex relaxation of the ℓ_0-norm penalty and assumes a Laplace prior distribution). However, the continuity of the soft-thresholding comes at the price of shifting the large coefficients by λ and introducing larger bias. Under the spirit of "continuity", "unbiasedness" and "sparsity", Fan et al. [19] proposed a *Smoothly Clipped Absolute Deviation Penalty* (SCAD) thresholding:

$$\boldsymbol{x}^{\mathrm{SCAD}} = \mathcal{S}^{\mathrm{scad}}_{\gamma,\lambda}(z) := \begin{cases} \mathrm{sign}(z)(|z|-\lambda)_+, & |z| \le 2\lambda, \\ \frac{(\gamma-1)z-\mathrm{sign}(z)\gamma\lambda}{\gamma-2}, & 2\lambda < |z| < \gamma\lambda, \\ z, & |z| \ge \gamma\lambda. \end{cases}$$

with $\gamma \ge 2$. The corresponding SCAD penalty of the MAP model (2.18) is given by

$$\varphi^{\mathrm{scad}}_{\gamma,\lambda}(x) = \begin{cases} \lambda|x|, & |x| \le \lambda, \\ \frac{\gamma+1}{2}\lambda^2 - \frac{(\gamma\lambda-|x|)^2_+}{2(\gamma-1)}, & |x| > \lambda, \end{cases}$$

The SCAD-thresholding is continuous compared to the hard-thresholding and avoids the large bias introduced by the soft-thresholding. Actually, it can be interpreted as a linear interpolation between the soft-thresholding and the hard-thresholding in the interval $(2\lambda, \gamma\lambda)$. The SCAD-thresholding also assumes a different prior from the Bernoulli-uniform sparse prior. The relating prior of SCAD smooths the discontinuity of the Bernoulli-uniform sparse prior around the origin, but still keeps sparsity and is uniform for large parameters. The MCP-thresholding [35] is proposed under a similar spirit as SCAD while the corresponding penalty of MCP is less concave than that of SCAD. The MCP penalty term takes the form of

$$\varphi^{\mathrm{mcp}}_{\gamma,\lambda}(x) = \begin{cases} \lambda|x| - \frac{|x|^2}{2\gamma}, & |x| \le \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & |x| > \gamma\lambda, \end{cases}$$

and the MCP thresholding is

$$\boldsymbol{x}^{\mathrm{MCP}} = \mathcal{S}^{\mathrm{mcp}}_{\gamma,\lambda}(z) := \begin{cases} 0, & |z| \le \lambda, \\ \mathrm{sign}(z)\frac{\gamma(|z|-\lambda)}{\gamma-1}, & \lambda < |z| < \gamma\lambda, \\ z, & |z| \ge \gamma\lambda, \end{cases}$$

with $\gamma \ge 1$.

By proposing the three conditions of "sparsity", "unbiasedness" and "continuity", Fan and Li [19] actually means to find a stable way to balance the variance and bias. In Bayesian estimation, the prior knowledge influences the parameter estimation by shifting its posterior probability mass density towards the region preferred by the prior, e.g., the sparse prior shifts the density towards zero. As seen, this influence has two effects: reducing the variance of the estimate (or the noise effect), but gaining bias simultaneously. The "sparsity" and "unbiasedness" is the result of negotiation between the variance and bias in different ranges of observations. The "sparsity" means omitting the small

observations while the "unbiasedness" means maintaining the large ones. Since the small observations contain more noise than signal information, omitting them reduces much variance and gains little bias. On the contrary, the large observations contain strong signal information such that the noise effect can be neglected; and thus maintaining them is wiser.

The mean square error can be decomposed as the sum of the variance and square bias of the estimator

$$\begin{aligned} \text{MSE} &= \mathbb{E}_{(\mathrm{x},\mathrm{z})}(x - g(z))^2 \\ &= \mathbb{E}_{\mathrm{x}}(x - \mathbb{E}_{\mathrm{z}}g(z))^2 + \mathbb{E}_{\mathrm{z}}\big(g(z) - \mathbb{E}_{\mathrm{z}}g(z)\big)^2 \\ &= \text{Bias}^2(g(z)) + \text{Var}(g(z)). \end{aligned} \tag{2.20}$$

So the univariate MMSE estimator under the Bernoulli-uniform sparse prior naturally provides a balance between the variance and bias, i.e., a balance between "sparsity" and "unbiasedness". Moreover, it is smooth according to Proposition 2.1. In short, the univariate MMSE estimator balances the variance and bias in a stable way and satisfies the three conditions of "continuity", "unbiasedness" and "sparsity".

Finally, to demonstrate the efficiency of the univariate MMSE estimator in the sparse case, we implement a parameter estimation test for the univariate model (1.5) and evaluate the performance by the square error $|\hat{x} - x|^2$. The results are reported in Table 1. The parameter x is constructed to follow the Bernoulli-uniform sparse prior with $p_0 = 0.8, L = 3, U = 6$ in (1.2) and the noise variance σ^2 is 1. The experiments are repeated 1000 times. The parameters λ and γ in the related thresholding vary in $[0.5, 3]$ and $[1, 3]$ respectively. Only the least square error and the corresponding parameters are reported. It comes at no surprise that the univariate MMSE estimator outperforms the others.

Table 1: Parameter estimation results by the thresholding operators.

thresholding	hard	soft	SCAD	MCP	MMSE
λ	2.27	0.90	1.46	1.72	–
γ	–	–	1.00	1.94	–
$\lvert\hat{x} - x\rvert^2$	22.06	23.30	20.99	21.08	**19.82**

2.4 Risk analysis

In the above, we study some specific probability distributions and analyse the corresponding MMSE estimators. However, in reality, it is rare to know the probability distribution of complex signals. For example, a common prior about natural images is that they have piecewise regularity and belong to the space of bounded total variation. This prior information defines a signal set Θ but does not specify the probability distribution of signals over Θ. Thus, to control the risk of an estimation g for any signal $\boldsymbol{x}$ in Θ, one consider the minimax risk

$$R(\Theta) = \inf_{g} \sup_{\boldsymbol{x}\in\Theta} \mathbb{E}_{\mathbf{y}} \|g(\boldsymbol{y}) - \boldsymbol{x}\|_2^2.$$

The minimax risk over the ℓ_p-ball $\Theta_{p,n}(r) := \{\boldsymbol{x} : \sum_{i=1}^n |x_i|^p \leq r^p\}$ has been studied extensively by Donoho and Johnstone in [17]. The ℓ_p-ball prior can arise naturally in various applications. The ℓ_2-balls correspond to a space of bounded energy. As $p \to 0$, the ℓ_p-balls assume sparsity that only a small number of components can be non-zero.

Here we include some minimax risk of our proposed MMSE estimators with specific prior probability distributions, especially the Bernoulli-uniform sparse prior to verify that they still give valuable estimation over certain set Θ even though the prior probability may not model the truth accurately. To simplify the study, we only consider the case where the design matrix $\boldsymbol{A}$ is identity. Due to the i.i.d. structure of $\boldsymbol{x}$ and $\boldsymbol{y}$ when $\boldsymbol{A}$ is identity, the minimax risk $\Theta_{p,n}(r)$ reduces to the the univariate Bayes-minimax problem, for which it is enough to study the MMSE error $\mathbb{E}_{(\mathrm{x}\sim\mathrm{F},\mathrm{y})}|\mathcal{S}(y) - x|^2$ for the least favourable distribution F over $\mathbb{R}$ that satisfies the p-th moments constraint $\mathbb{E}_{\mathrm{x}\sim\mathrm{F}}|x|^p \leq r^p/n$.

When $p = 2$, the least favorable distribution is Gaussian and the corresponding MMSE estimator (2.8) is minimax over ℓ_2-ball. As pointed out by Donoho and Johnstone in [17], when $p < 2$, the least favorable distribution in the asymptotically case is the discrete priors putting mass only at zero and two opposite numbers. Given in our case, the Bernoulli-uniform distribution degenerates to that least favorable distribution by setting $L = U$ in (1.2). Thus, the corresponding univariate MMSE estimator (2.16) is also asymptotically minimax over ℓ_p-balls ($p < 2$) at small signal-to-noise ratios. Additionally, Donoho and Johnstone showed that the ℓ_2-losses of the soft-thresholding and hard-thresholding are close to that of the MMSE estimator for the least favorable distribution in the asymptotic case, and thus the soft-thresholding and hard-thresholding are also nearly minimax. In all, we showed that the univariate MMSE estimator (2.16) are asymptotically minimax with ℓ_2-loss over ℓ_p-balls when p is less than 2 and the signal-to-noise ratio is low. Our arguments follow the proofs of Donoho and Johnstone in [17] besides that they need an additional step to compare the ℓ_2-loss of the MMSE estimator (2.16) and thresholding rules.

3 Approximation of the MMSE estimator

In this section, we will propose the cyclic coordinate minimization algorithm based on the iteration (1.8) and establish a theorem on its convergence. The numerical experiments will follow to show the efficiency of our algorithm.

3.1 The algorithm and the main theorem

Recall the regression model

$$\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where x_i ($i = 1, 2, \cdots, p$) are i.i.d. drawn from p_{x} and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. More generally, we can consider the coefficients of $\boldsymbol{x}$ under some orthonormal basis $\boldsymbol{D}$: $\boldsymbol{u}_i = (\boldsymbol{D}\boldsymbol{x})_i$ ($i = 1, 2, \cdots, p$), to be i.i.d., and follow some prior distribution.

Once an estimate of $\boldsymbol{u}$ is obtained, $\boldsymbol{x}$ can be recovered by $\boldsymbol{x} = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{u}$. So we focus on the estimation of $\boldsymbol{u}$ from

$$\boldsymbol{z} = \boldsymbol{A}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{u} + \boldsymbol{\varepsilon}. \tag{3.2}$$

Nevertheless, we can build the orthonormal matrix $\boldsymbol{D}$ into the design matrix $\boldsymbol{A}$ such that the form of (3.2) reduces to that of (3.1). So we only consider the problem (3.1) in the following.

To approximate the MMSE estimator (1.4) for (3.1), we propose a cyclic coordinate minimization (CCM) algorithm, which minimizes the mean square error with respect to a single coordinate in a sequential manner. More specifically, for the current estimate $\boldsymbol{x}^k$, an index $i_k \in \{1, 2, \ldots, p\}$ is selected and then the i_k-th component of $\boldsymbol{x}^k$ is adjusted to minimize the mean square error:

$$x_{i_k}^{k+1} = \arg\min_{g:\mathbb{R}^n \to \mathbb{R}} \mathbb{E}_{(\mathrm{x}_{i_k}, \boldsymbol{\mathrm{z}}|\boldsymbol{x}_{j\neq i_k}^k)}(x_{i_k} - g(\boldsymbol{z}))^2. \tag{3.3}$$

It is equivalent to finding the MMSE estimator for the following linear regression model

$$\boldsymbol{z} - \sum_{j\neq i_k} \boldsymbol{a}_j x_j^k = \boldsymbol{a}_{i_k} x_{i_k} + \boldsymbol{\varepsilon}, \tag{3.4}$$

where $\boldsymbol{a}_i$ is the i-th column of $\boldsymbol{A}$, $x_{i_k} \in \mathbb{R}$ is the parameter to be estimated and $\boldsymbol{z} - \sum_{j\neq i_k} \boldsymbol{a}_j x_j^k \in \mathbb{R}^n$ is the vector of observations. The MMSE estimator for (3.4) is the posterior mean:

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}_{i_k}}(x_{i_k}|\boldsymbol{z} - \sum_{j\neq i_k} \boldsymbol{a}_j x_j^k) &= \frac{\int x_{i_k} p(x_{i_k}) \exp\big(\frac{-\|\boldsymbol{a}_{i_k} x_{i_k} - (\boldsymbol{z} - \sum_{j\neq i_k} \boldsymbol{a}_j x_j^k)\|_2^2}{2\sigma^2}\big) dx_{i_k}}{\int p(x_{i_k}) \exp\big(\frac{-\|\boldsymbol{a}_{i_k} x_{i_k} - (\boldsymbol{z} - \sum_{j\neq i_k} \boldsymbol{a}_j x_j^k)\|_2^2}{2\sigma^2}\big) dx_{i_k}} \\
&= \frac{\int x_{i_k} p(x_{i_k}) \exp\big(\frac{-(x_{i_k} - \boldsymbol{a}_{i_k}^{\mathrm{T}}(\boldsymbol{z} - \sum_{j\neq i_k} \boldsymbol{a}_j x_j^k)/\|\boldsymbol{a}_{i_k}\|_2^2)^2}{2\sigma^2/\|\boldsymbol{a}_{i_k}\|_2^2}\big) dx_{i_k}}{\int p(x_{i_k}) \exp\big(\frac{-(x_{i_k} - \boldsymbol{a}_{i_k}^{\mathrm{T}}(\boldsymbol{z} - \sum_{j\neq i_k} \boldsymbol{a}_j x_j^k)/\|\boldsymbol{a}_{i_k}\|_2^2)^2}{2\sigma^2/\|\boldsymbol{a}_{i_k}\|_2^2}\big) dx_{i_k}} \\
&= \mathcal{S}^{\mathrm{MMSE}}_{\sigma/\|\boldsymbol{a}_{i_k}\|_2}(\boldsymbol{a}_{i_k}^{\mathrm{T}}(\boldsymbol{z} - \sum_{j\neq i_k} \boldsymbol{a}_j x_j^k)/\|\boldsymbol{a}_{i_k}\|_2^2).
\end{aligned}$$

That is, the solution of (3.3) is obtained as

$$\boldsymbol{x}_{i_k}^{k+1} = \mathcal{S}^{\mathrm{MMSE}}_{\sigma/\|\boldsymbol{a}_{i_k}\|_2}(\boldsymbol{a}_{i_k}^{\mathrm{T}}(\boldsymbol{z} - \sum_{j\neq i_k} \boldsymbol{a}_j x_j^k)/\|\boldsymbol{a}_{i_k}\|_2^2). \tag{3.5}$$

At each iteration, the computation of (3.5) involves only single integration and is easy to perform. The coordinate minimization method can be deterministic or randomized depending on the choice of the update coordinates. Here, we select the coordinate index i_k in a cyclic fashion from the set $\{1, 2, \cdots, p\}$ so that every coordinate is modified once in every cycle of p iterations. The resulting algorithm is called the cyclic coordinate minimization (CCM) algorithm. Given an permutation $\{i_0, i_1, \cdots, i_{p-1}\}$ of sequence $\{1, 2, \cdots, p\}$, we choose the index i_k as

$$i_k = i_{\bar{k}}, \text{ where } \bar{k} = [k \bmod p]. \tag{3.6}$$

Algorithm 1 The Cyclic Coordinate Minimization (CCM) algorithm

Set $k \leftarrow 0$ and choose $\boldsymbol{x}^0 \in \mathbb{R}^p$;
while termination test is not satisfied **do**
  choose index i_k as (3.6)
  $x_{i_k}^{k+1} = \mathcal{S}^{\mathrm{MMSE}}_{\sigma/\|\boldsymbol{a}_{i_k}\|_2}(\boldsymbol{a}_{i_k}^{\mathrm{T}}(\boldsymbol{z} - \sum_{j\neq i_k} \boldsymbol{a}_j x_j^k)/\|\boldsymbol{a}_{i_k}\|_2^2)$
  $k \leftarrow k+1$;
end while

Finally, the CCM algorithm is summarized as follows.

We note that once Algorithm 1 converges (which is guaranteed by the following Theorem 3.1), the limiting point is a a fixed point of the iteration (3.3) that satisfies

$$\hat{x}_{i_k} = \arg \min_{g(\boldsymbol{z}):\mathbb{R}^n\to\mathbb{R}} \mathbb{E}_{(\mathrm{x}_{i_k},\mathbf{z}|\hat{\boldsymbol{x}}_{j\neq i_k})}(x_{i_k} - g(\boldsymbol{z}))^2. \tag{3.7}$$

The above equation implies that $\hat{\boldsymbol{x}}$ attains the minimum of the mean square error with respect to arbitrary single coordinate by fixing the remaining ones. For any point that satisfies (3.7), we call it a coordinatewise minimum mean square error (CMMSE) estimate.

Before giving the main theorem on the convergence of Algorithm 1, we need the following assumption.

Assumption 2. *Assume the prior distribution $p_{\mathrm{x}}(x)$ satisfies Assumption 1. Furthermore, there exist $K > 0$, $b > 0$ and a function $a(x) \geq 0$ which monotonously decreases to zero when $|x|$ goes to infinity, such that $p_{\mathrm{x}}(x)$ can be represented as*

$$p_{\mathrm{x}}(x) = a(x)\exp(-b|x|), \ \textit{when } |x| > K. \tag{3.8}$$

Note that Assumption 2 is not strong. The Gaussian, Gaussian mixture and Bernoulli-uniform sparse prior all satisfy Assumption 2. For example, since the Bernoulli-uniform sparse prior is compactly supported, it satisfies (3.8) as we can choose $a(x) = 0$ for a sufficiently large K. Next we state the main theorem of this paper and its proof will be given in section 4.

Theorem 3.1. *Assume the prior distribution $p_{\mathrm{x}}(x)$ satisfies Assumption 2. Let $\{\boldsymbol{x}^k\}$ be the sequence generated by Algorithm 1. Then $\{\boldsymbol{x}^k\}$ converges to a coordinatewise minimum mean square error (CMMSE) estimate that satisfies* (3.7).

3.2 Numerical results

Finally, we show the efficiency of our algorithm by the three examples discussed in section 2.3, i.e., the Gaussian, Gaussian mixture and Bernoulli-uniform sparse prior. The case where $n = 100$, $p = 200$ is tested. The columns of matrix $\boldsymbol{A}$ in (3.1) are highly correlated, which are generated from a Gaussian distribution

with zero mean and variance $\boldsymbol{\Sigma}_{i,j} = 0.3^{|i-j|}$ $(1 \le i \le n,\ 1 \le j \le p)$. We stop the iteration for Algorithm 1 if a maximum number of 300 iterations is reached, or the difference of the estimates between two successive cycles is smaller than 10^{-8}, i.e., $\|\boldsymbol{x}^{(l+1)p} - \boldsymbol{x}^{lp}\|_2 \le 10^{-8}$. For comparison, the MMSE, LMMSE and MAP estimator are also computed when they are available and easy to compute.

The built-in univariate MMSE estimator in Algorithm 1 for each prior distribution has been given in section 2.3 accordingly. In the implementation, we can calculate the values of the univariate MMSE estimator, i.e. $\mathbb{E}(x|z)$, at a given set of sample points of z in advance. Then the univariate MMSE estimator can be approximated by piecewise linear functions with arbitrary accuracy as long as the set of sample points is large enough. When iterating, we only need to call that piecewise linear function to update the sequence and avoid calculating the single integrals every time such that the computation of the algorithm may be faster.

Example 1. ***The Gaussian prior.*** *Under the Gaussian prior, all of the MMSE, LMMSE and MAP estimators for the problem* (1.1) *take the same form of*

$$\hat{\boldsymbol{x}} = \mu\boldsymbol{e} + \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} + \frac{\sigma^2}{\sigma_s^2}\boldsymbol{I})^{-1}(\boldsymbol{z} - \mu\boldsymbol{A}\boldsymbol{e}), \tag{3.9}$$

where $\boldsymbol{I}$ *is the identity matrix, and* $\boldsymbol{e}$ *is the vector of all ones. Our variational model* (1.10) *is the same as the MAP model under the Gaussian prior, which is convex and has the unique stationary point* (3.9)*. Then, our algorithm also converges to the estimate* (3.9)*.*

Example 2. ***The Gaussian Mixture prior.*** *Similar to the univariate case, the posterior distribution with the Gaussian mixture prior and Gaussian noise is also a Gaussian mixture distribution for the general problem* (3.1)*. The MMSE estimator for the general problem* (3.1) *under the Gaussian mixture prior is given by*

$$\boldsymbol{x}^{\mathrm{MMSE}} = \sum_{i=1}^{k} \boldsymbol{\beta}_i(\boldsymbol{z})\big(\mu_i\boldsymbol{e} + \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} + \frac{\sigma^2}{\sigma_i^2}\boldsymbol{I})^{-1}(\boldsymbol{z} - \mu_i\boldsymbol{A}\boldsymbol{e})\big),\ \boldsymbol{\beta}_i(\boldsymbol{z}) = \frac{\boldsymbol{\phi}_i(\boldsymbol{z})}{\sum_{i=1}^{k}\boldsymbol{\phi}_i(\boldsymbol{z})}$$

where $\boldsymbol{\phi}_i(\boldsymbol{z})$ *is the Gaussian probability density function (PDF) in* $\boldsymbol{z}$ *with mean*

$$\hat{\boldsymbol{\mu}}_i = \mu_i\boldsymbol{A}\boldsymbol{e},$$

and variance

$$\boldsymbol{C}_i = \sigma_i^2\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} + \sigma^2\boldsymbol{I}.$$

From [28], the LMMSE estimator is

$$\boldsymbol{x}^{\mathrm{LMMSE}} = \hat{\mu}\boldsymbol{e} + \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} + \frac{\sigma^2}{\hat{\sigma}^2}\boldsymbol{I})^{-1}(\boldsymbol{z} - \hat{\mu}\boldsymbol{A}\boldsymbol{e}).$$

where

$$\hat{\mu} = \sum_{i=1}^{k} p_i\mu_i,\ \hat{\sigma}^2 = \sum_{i=1}^{k} p_i\sigma_i^2 + \sum_{i=1}^{k} p_i\mu_i^2 - \hat{\mu}^2.$$

The MAP estimator is hard to compute due to the multi-mode issue of the Gaussian mixture posterior distribution, so we do not compare it here.

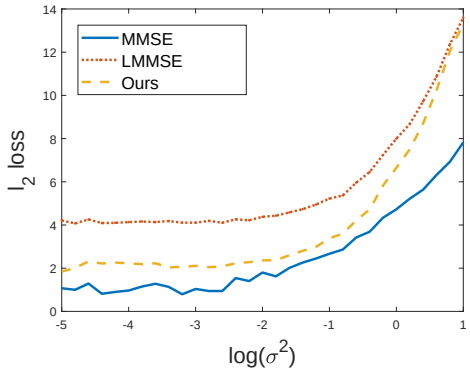

Figure 2: The average ℓ_2-norm error of the estimates for the general problem (3.1) under the Gaussian mixture prior.

We select the Gaussian mixture prior with five components. The component means are

$$(-1, -0.5, 0, 1, 3),$$

and the variances are all 1. We vary the logarithm of the noise variance from -5 to 1 equally. Algorithm 1 is initialized with the LMMSE estimate. The average ℓ_2-norm error of the obtained estimates over 50 independent trials are plotted in Figure 2. Clearly, Algorithm 1 provides a better approximation of the MMSE estimator than the LMMSE estimator.

Example 3. ***The Bernoulli-uniform sparse prior.*** *Under the Bernoulli-uniform sparse prior, the MAP models with the ℓ_1-norm, SCAD and MCP penalty for the general problem (3.1) are solved to compare with our algorithm. We choose $\gamma = 3.7$ in the SCAD penalty and $\gamma = 2$ in the MCP penalty. The ℓ_1-norm penalized model is solved by the ADMM algorithm [5], while the SCAD or MCP penalized model is solved by the LLA algorithm proposed by [20]. The LLA algorithm computes a sequential weighted ℓ_1-norm penalized problems. It is initialized with zero and terminated if a maximum number of 300 iterations is reached or the ℓ_2 norm of the difference between two successive updates is smaller than 10^{-8}. In addition, we compute the oracle solution for comparison. The oracle solution is the solution obtained by knowing the true support:*

$$\boldsymbol{x}^{\text{oracle}} = \arg \min_{\boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{x}_{S^c}^{\text{oracle}} = 0} \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{z}\|_2^2,$$

where S is the support set of the truth $\boldsymbol{x}$ and S^c is the complementary set of S. The LMMSE estimator is not compared here because it fails to enhance sparsity. Our algorithm is initialized with the solution of the the ℓ_1-norm penalized model.

Table 2: The numerical results of different methods for the problem (3.1) under the Bernoulli-uniform sparse prior. The results are averaged over 50 independent trials with the standard error shown in the parenthesis.

Method	ℓ_1 error	ℓ_2 error	#FP	#FN
Oracle	5.4407	1.4980	0	0
	(0.9543)	(0.2411)	(0)	(0)
ℓ_1-norm penalty	25.4007	4.6593	35.88	0.00
	(7.3596)	(1.2588)	(4.58)	(0.00)
SCAD	5.7361	1.6021	0.36	0.00
	(1.1306)	(0.3327)	(0.66)	(0.00)
MCP	5.4622	1.5075	0.00	0.00
	(0.9834)	(0.2568)	(0.00)	(0.00)
Algorithm 1	**4.7092**	**1.3060**	0.00	0.00
	(0.7872)	**(0.2076)**	(0.00)	(0.00)

The estimation accuracy is evaluated by the average ℓ_1-norm error and ℓ_2-norm error. We treat the estimated parameter as zero if its absolute value is less than 10^{-6}. The selection accuracy of zero parameters is measured by the average counts of false positive (#FP) and false negative (#FN). The average performance over 50 independent trials is reported in Table 2 with the standard error shown in the parenthesis. The experiment results show that our method performs the best in any way of evaluation.

4 Proof of Theorem 3.1

It is hard to analyse the convergence of Algorithm 1 from the aspect of minimizing the mean square error sequentially. Inspired by the connection between the univariate MMSE estimator and the univariate variational model, we consider the following variational model for the general problem (3.1)

$$\min_{x\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z}\|_2^2 + \sum_{i=1}^{p} \|\boldsymbol{a}_i\|_2^2 \varphi^{\mathrm{MMSE}}_{\sigma/\|\boldsymbol{a}_i\|_2}(x_i), \tag{4.1}$$

which has an ℓ_2 fidelity for Gaussian noise and a separable MMSE penalty. We denote the objective function of (4.1) as $H(\boldsymbol{x})$ in the following. Applying the coordinate minimization algorithm to (4.1), we update a single coordinate and fix the remaining ones to minimize $H(\boldsymbol{x})$ at each iteration. Specifically, at

iteration k, we only update x_{i_k} as follows

$$\begin{aligned} x_{i_k}^{k+1} &= \arg\min_{x_{i_k}\in\mathbb{R}} H(\boldsymbol{x}|_{x_j=x_j^k \text{ for } j\neq i_k}) \\ &= \arg\min_{x_{i_k}\in\mathbb{R}} \frac{1}{2}\|\boldsymbol{a}_{i_k}\|_2^2\big(x_{i_k} - \frac{1}{\|\boldsymbol{a}_{i_k}\|_2^2}\boldsymbol{a}_{i_k}^{\mathrm{T}}(\boldsymbol{z} - \sum_{j\neq i_k}\boldsymbol{a}_j x_j^k)\big)^2 + \|\boldsymbol{a}_{i_k}\|_2^2\varphi^{\mathrm{MMSE}}_{\sigma/\|\boldsymbol{a}_{i_k}\|_2}(x_{i_k}). \end{aligned} \tag{4.2}$$

Since the values of x_j^k where $j \neq i_k$ are fixed, (4.2) reduces to the univariate variational model with respect to x_{i_k}. By Theorem 2.1, the solution is

$$x_{i_k}^{k+1} = \mathcal{S}^{\mathrm{MMSE}}_{\sigma/\|\boldsymbol{a}_{i_k}\|_2}(\boldsymbol{a}_{i_k}^{\mathrm{T}}(\boldsymbol{z} - \sum_{j\neq i_k}\boldsymbol{a}_j x_j^k)/\|\boldsymbol{a}_{i_k}\|_2^2), \tag{4.3}$$

which is the same as the iteration (3.5). To conclude, we have the following lemma.

Lemma 4.1. *Applying the coordinate minimization algorithm to the the variational problem* (4.1)*, we obtain the iteration* (3.5)*.*

Moreover, choosing the index i_k as (3.6) in a cyclic manner, we obtain the cyclic coordinate minimization algorithm for solving the variational model (4.1) that is the same as Algorithm 1. Thus we can prove the convergence of Algorithm 1 by analysing the properties of the objective function $H(\boldsymbol{x})$ of the variational model (4.1) and employing the tools developed in optimization.

In order to prove the convergence of Algorithm 1, we require the objective function $H(\boldsymbol{x})$ to satisfy the Kurdyka-Łojasiewicz property and to be coercive. Once these two requirements are satisfied, Theorem 3.1 can be proved by applying Theorem 2.9 in [2].

4.1 The Kurdyka-Łojasiewicz property and coerciveness of $H(\boldsymbol{x})$

We start by giving a definition of the Kurdyka-Łojasiewicz (KL) property, which plays an important role in the following analysis. We say a real extended valued function $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is proper if it is not positive infinity identically.

Definition 1 (Kurdyka-Łojasiewicz Property)**.** *A proper real extended valued function $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is said to have the* Kurdyka-Łojasiewicz property *at $\bar{z} \in \operatorname{dom}\partial f = \{z \in \mathbb{R}^p : \partial f(z) \neq \emptyset\}$ if there exist $\zeta \in (0, +\infty]$, a neighborhood U of $\bar{z}$, and a continuous concave function $\psi : [0, \zeta) \to \mathbb{R}_+$ such that*

1. *$\psi(0) = 0$;*
2. *$\psi(0)$ is C^1 on $(0, \zeta)$;*
3. *for all $s \in (0, \zeta)$, $\psi'(s) > 0$;*

4. *for all* $z \in U$ *satisfying* $f(\bar{z}) < f(z) < f(\bar{z}) + \zeta$, *the Kurdyka-Łojasiewicz inequality holds:*

$$\psi'(f(z) - f(\bar{z}))\,\mathrm{dist}(0, \partial f(z)) \geq 1.$$

where $\mathrm{dist}(0, \partial f(z)) = \min\{\|v\| : v \in \partial f(z)\}$,

The foundational works on the KL property are due to Łojasiewicz [31] and Kurdyka [29]. See [3, 1, 2, 4] for the development and applications of KL property in optimization problems. The KL property is a generalization of the Łojasiewicz inequality to nonsmooth subanalytic functions, while it is enough for our analysis to employ the Łojasiewicz inequality which says that for real analytic funciton $f : U \to \mathbb{R}$, there exists $\theta \in [\frac{1}{2}, 1)$ such that $|f - f(a)|^{\theta}\|\nabla f\|^{-1}$ remains bounded for a critical point $a \in U$. That is, f has the KL property at any $a \in U$ with $\psi(x) = c|f(x) - f(a)|^{1-\theta}$ for some constant c.

Let $\mathcal{E}_i$ be the range of $\mathcal{S}^{\mathrm{MMSE}}_{\sigma/\|\boldsymbol{a}_i\|_2}$, then $\varphi^{\mathrm{MMSE}}_{\sigma/\|\boldsymbol{a}_i\|_2}$ is analytic on $\mathcal{E}_i$ according to Lemma 2.1. Denote $\mathcal{E} = \prod_{i=1}^{p} \mathcal{E}_i$. Since $H(\boldsymbol{x})$ is analytic on $\mathcal{E}$, it has the KL property at any point in $\mathcal{E}$.

Proposition 4.1. *Assume* $p_{\mathrm{x}}(x)$ *satisfies Assumption 1. The objective function* $H(\boldsymbol{x})$ *of the variational model* (4.1) *has the KL property at any point in* $\mathcal{E}$.

The coerciveness of the objective function guarantees the boundedness of the generated sequence if the objective function value is finite. We give the definition of coerciveness as follows.

Definition 2 (Coerciveness)**.** *A real extended valued function* $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ *is called coercive iff* $f(\boldsymbol{x}) \to +\infty$ *as* $\|\boldsymbol{x}\| \to +\infty$.

Finally, we prove $H(\boldsymbol{x})$ is coercive with Assumption 2 in the following lemma.

Lemma 4.2. *Assume the prior* $p_{\mathrm{x}}(x)$ *satisfies Assumption 2. The objective function* $H(\boldsymbol{x})$ *of the variational model* (4.1) *is coercive.*

Proof. Note that $H(\boldsymbol{x})$ includes a separable MMSE penalty which is the sum of the univariate MMSE penalty for every coordinate. Once the coerciveness of the univariate MMSE penalty $\varphi^{\mathrm{MMSE}}_{\sigma}(x)$ is proved, we can obtain that $H(\boldsymbol{x})$ is coercive. Next, we focus on the proof of the coerciveness of $\varphi^{\mathrm{MMSE}}_{\sigma}(x)$. When the domain of $r^{\mathrm{MMSE}}_{\sigma}$ is not $\mathbb{R}$, $\varphi^{\mathrm{MMSE}}_{\sigma}(x)$ is defined to be positive infinity outside. So we only need to prove the case where the domain of $r^{\mathrm{MMSE}}_{\sigma}$ is $\mathbb{R}$ and

$$\varphi^{\mathrm{MMSE}}_{\sigma}(x) = \int_0^x (r^{\mathrm{MMSE}}_{\sigma}(u) - u)du + c,\ x \in \mathbb{R}.$$

We only show that $\varphi^{\mathrm{MMSE}}_{\sigma}(x) \to +\infty$ as $x \to +\infty$, because the case is similar when x goes to $-\infty$.

Supposing $r^{\mathrm{MMSE}}_{\sigma}(x) > x + \xi$ for some $\xi > 0$ when x is sufficiently large, immediately we get that $\varphi^{\mathrm{MMSE}}_{\sigma}(x) \to +\infty$ as $x \to +\infty$. Since $r^{\mathrm{MMSE}}_{\sigma}(x)$ is the inverse function of $\mathcal{S}^{\mathrm{MMSE}}_{\sigma}$, the condition $r^{\mathrm{MMSE}}_{\sigma}(x) > x + \xi$ for sufficiently

large x can be guaranteed by showing that $\mathcal{S}_\sigma^{\mathrm{MMSE}}(z) < z - \xi$ for sufficiently large z. In the following, we will show there exist $\xi > 0$ and $K > 0$ such that $\mathcal{S}_\sigma^{\mathrm{MMSE}}(z) < z - \xi$ when $z > K$.

Recall

$$\mathcal{S}_\sigma^{\mathrm{MMSE}}(z) = \frac{\int x p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx}{\int p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx}, \tag{4.4}$$

where ϕ_σ is the probability density function of the Gaussian distribution with mean zero and variance σ^2. We separate the involved integrals in (4.4) into two parts

$$\mathcal{S}_\sigma^{\mathrm{MMSE}}(z) = \frac{\int_{-\infty}^{\tilde{z}} x p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx}{\int p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx} + \frac{\int_{\tilde{z}}^{\infty} x p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx}{\int p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx}.$$

If we can prove there exist $\xi > 0$, $K > 0$ and $\tilde{z}$ such that

$$\frac{\int_{-\infty}^{\tilde{z}} x p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx}{\int p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx} < z - 2\xi,\ z > K', \tag{4.5}$$

and

$$\frac{\int_{\tilde{z}}^{\infty} x p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx}{\int p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx} < \xi,\ z > K'', \tag{4.6}$$

we obtain $\mathcal{S}_\sigma^{\mathrm{MMSE}}(z) < z - \xi$ when $z > K = \max\{K', K''\}$. Next we focus on the proof of the inequality (4.5) and (4.6).

To show the inequality (4.5), we need to employ Assumption 2. It says that there exist $K' > 0$ and $b > 0$ such that

$$p_{\mathrm{x}}(x) = a(x)\exp(-bx),\ x > K',$$

where $a(x) \geq 0$ monotonously decreases to zero when $|x|$ goes to infinity. We choose

$$\xi = b\sigma^2/2,\ \tau = \exp(-\frac{z^2 - (z-2\xi)^2}{2\sigma^2}),\ \tilde{z} = 2(z-2\xi) - K',$$

and assume $\tilde{z} > K'$. Then we have

$$\begin{aligned}
&\int_{K'}^{\tilde{z}} (x - z + 2\xi) p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx \\
=&\int_{K'}^{\tilde{z}} (x - z + 2\xi) a(x)\exp(-bx)\phi_\sigma(x-z)dx \\
=&\tau \int_{K'}^{\tilde{z}} (x - z + 2\xi) a(x)\phi_\sigma(x-z+2\xi)dx \\
=&\tau \int_{K'-(z-2\xi)}^{(z-2\xi)-K'} x a(x+z-2\xi)\phi_\sigma(x)dx \\
=&\tau \int_{0}^{(z-2\xi)-K'} x\big(a(x+z-2\xi) - a(-x+z-2\xi)\big)\phi_\sigma(x)dx \\
\leq&0.
\end{aligned}$$

That is,

$$\int_{K'}^{\tilde{z}} xp_{\mathrm{x}}(x)\phi_\sigma(x-z) \le (z-2\xi)\int_{K'}^{\tilde{z}} p_{\mathrm{x}}(x)\phi_\sigma(x-z).$$

It is obvious that

$$\int_{-\infty}^{K'} xp_{\mathrm{x}}(x)\phi_\sigma(x-z) \le K'\int_{-\infty}^{K'} p_{\mathrm{x}}(x)\phi_\sigma(x-z) < (z-2\xi)\int_{-\infty}^{K'} p_{\mathrm{x}}(x)\phi_\sigma(x-z).$$

Combining the above two inequalities, we get that

$$\frac{\int_{-\infty}^{\tilde{z}} xp_{\mathrm{x}}(x)\phi_\sigma(x-z)dx}{\int_{-\infty}^{\tilde{z}} p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx} = \frac{\int_{-\infty}^{K'} xp_{\mathrm{x}}(x)\phi_\sigma(x-z)dx + \int_{K'}^{\tilde{z}} xp_{\mathrm{x}}(x)\phi_\sigma(x-z)dx}{\int_{-\infty}^{K'} p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx + \int_{K'}^{\tilde{z}} p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx} < z-2\xi.$$

Then, it is easy to obtain the inequality (4.5):

$$\frac{\int_{-\infty}^{\tilde{z}} xp_{\mathrm{x}}(x)\phi_\sigma(x-z)dx}{\int p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx} \le \min\left\{0, \frac{\int_{-\infty}^{\tilde{z}} xp_{\mathrm{x}}(x)\phi_\sigma(x-z)dx}{\int_{-\infty}^{\tilde{z}} p_{\mathrm{x}}(x)\phi_\sigma(x-z)dx}\right\} < z-2\xi.$$

Next we prove the inequality (4.6). If $a(\tilde{z}) = 0$, which means that $a(x) = 0$ for all $x \in (\tilde{z}, +\infty)$, the proof is done; otherwise, since $a(x)$ decreases when $x > K'$, we have

$$\begin{aligned}
\frac{\int_{\tilde{z}}^{\infty} xa(x)\phi_\sigma(x-z+2\xi)dx}{\int a(x)\phi_\sigma(x-z+2\xi)dx} \le & \frac{\int_{\tilde{z}}^{\infty} xa(x)\phi_\sigma(x-z+2\xi)dx}{\int_{K'}^{\tilde{z}} a(x)\phi_\sigma(x-z+2\xi)dx} \\
\le & \frac{a(\tilde{z})\int_{\tilde{z}}^{\infty} x\phi_\sigma(x-z+2\xi)dx}{a(\tilde{z})\int_{K'-(z-2\xi)}^{(z-2\xi)-K'} \phi_\sigma(x)dx} \\
= & \frac{\int_{\tilde{z}}^{\infty} x\phi_\sigma(x-z+2\xi)dx}{\int_{K'-(z-2\xi)}^{(z-2\xi)-K'} \phi_\sigma(x)dx} \\
= & \frac{\sigma^2\phi_\sigma((z-2\xi)-K') + (z-2\xi)\int_{(z-2\xi)-K'}^{\infty} \phi_\sigma(x)dx}{\int_{K'-(z-2\xi)}^{(z-2\xi)-K'} \phi_\sigma(x)dx}.
\end{aligned}$$

By L'Hòpital's rule, we obtain

$$\lim_{z\to+\infty}(z-2\xi)\int_{(z-2\xi)-K'}^{\infty} \phi_\sigma(x)dx = \lim_{z\to+\infty}(z-2\xi)^2\phi_\sigma((z-2\xi)-K') = 0.$$

Moreover, it is easy to see that

$$\lim_{z\to+\infty}\sigma^2\phi_\sigma((z-2\xi)-K') = 0, \text{ and } \lim_{z\to+\infty}\int_{K'-(z-2\xi)}^{(z-2\xi)-K'} \phi_\sigma(x)dx = 1.$$

Then there exists $K'' > 0$ such that

$$\frac{\sigma^2\phi_\sigma((z-2\xi)-K') + (z-2\xi)\int_{(z-2\xi)-K'}^{\infty} \phi_\sigma(x)dx}{\int_{K'-(z-2\xi)}^{(z-2\xi)-K'} \phi_\sigma(x)dx} < \xi, \text{ when } z > K''. \quad (4.7)$$

So we complete the proof of the inequality (4.6). □

4.2 Proof of the main theorem

In order to prove the convergence of Algorithm 1, we need to employ the following theorem that is given in [2].

Theorem 4.1 ([2]). *Let $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be a semi-continuous function. Consider a sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ that satisfies*

C1. (Sufficient decrease condition). For each $k \in \mathbb{N}$,

$$f(\boldsymbol{x}^{k+1}) + a\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2^2 \leq f(\boldsymbol{x}^k),\ a > 0;$$

C2. (Relative error condition). For each $k \in \mathbb{N}$, there exists $\boldsymbol{w}^{k+1} \in \partial f(\boldsymbol{x}^{k+1})$ such that

$$\|\boldsymbol{w}^{k+1}\|_2 \leq b\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2,\ b > 0;$$

C3. (Continuity condition). There exists a subsequence $\{\boldsymbol{x}^{k_j}\}_{j\in\mathbb{N}}$ and such that

$$\boldsymbol{x}^{k_j} \to \hat{\boldsymbol{x}} \text{ and } f(\boldsymbol{x}^{k_j}) \to f(\hat{\boldsymbol{x}}),\ \text{as } j \to +\infty.$$

If f has the KL property at $\hat{\boldsymbol{x}}$, then the sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ converges to $\hat{\boldsymbol{x}}$ as k goes to infinity, and $\hat{\boldsymbol{x}}$ is a stationary point of f .

To apply Theorem 4.1, we need to show our objective function $H(\boldsymbol{x})$ has the KL property and satisfies the conditions C1-C3. The KL property of $H(\boldsymbol{x})$ is proven in Lemma 4.1 and the coerciveness in Lemma 4.2. The condition C3 follows from the coerciveness and continuity of $H(\boldsymbol{x})$. The rest of the proof of Theorem 3.1 is given below.

Proof. Let $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1. As only one coordinate is updated between two successive iterates of $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$, we consider the subsequence $\{\boldsymbol{x}^{lp}\}_{l\in\mathbb{N}}$ in every cycle of p iterations such that all coordinates is updated once after one cycle. Then for any $k' \in \{1, 2, \ldots, p\}$, we have

$$\|\boldsymbol{x}^{lp+k'} - \boldsymbol{x}^{lp}\|_2 \leq \|\boldsymbol{x}^{(l+1)p} - \boldsymbol{x}^{lp}\|_2.$$

Thus the convergence of $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ can be guaranteed by that of the subsequence $\{\boldsymbol{x}^{lp}\}_{l\in\mathbb{N}}$. In the following, we focus on the proof of the convergence of the subsequence $\{\boldsymbol{x}^{lp}\}_{l\in\mathbb{N}}$. Since $H(\boldsymbol{x})$ is semi-continuous and satisfies the KL property (by Lemma 4.1) at each finite point of $\mathcal{E}$, to obtain the convergence of $\{\boldsymbol{x}^{lp}\}_{l\in\mathbb{N}}$ by Theorem 4.1, we only need to prove it satisfies C1 (sufficient decrease condition), C2 (relative error condition), and C3 (Continuity condition).

Firstly, we show $\{\boldsymbol{x}^{lp}\}_{l\in\mathbb{N}}$ satisfies the condition C1. As the coordinate minimization algorithm only updates one coordinate once, we consider the univariate objective function:

$$H_{i_k}^k(x) = \|\boldsymbol{a}_{i_k}\|_2^2\big(x - \frac{1}{\|\boldsymbol{a}_{i_k}\|_2^2}\boldsymbol{a}_{i_k}^{\mathrm{T}}(\boldsymbol{z} - \sum_{j\neq i_k}\boldsymbol{a}_j x_j^k)\big)^2 + \|\boldsymbol{a}_{i_k}\|_2^2\varphi_{\sigma/\|\boldsymbol{a}_{i_k}\|_2}^{\mathrm{MMSE}}(x).$$

Because $x_{i_k}^{k+1}$ is the minimizer of $H_{i_k}^k(x)$ and $\boldsymbol{x}_{j\neq i_k}^{k+1} = \boldsymbol{x}_{j\neq i_k}^k$, it yields that

$$0 \in \partial H_{i_k}^k(x_{i_k}^{k+1}), \tag{4.8}$$

and

$$H(\boldsymbol{x}^{k+1}) \leq H(\boldsymbol{x}^k) < +\infty. \tag{4.9}$$

Since $H(\boldsymbol{x})$ is coercive by Lemma 4.2, we can get that $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ is bounded. Recall that the components of $\boldsymbol{x}^k$ are computed by

$$\boldsymbol{x}_{i_k}^{k+1} = \mathcal{S}^{\mathrm{MMSE}}_{\sigma/\|\boldsymbol{a}_{i_k}\|_2}(\boldsymbol{a}_{i_k}^{\mathrm{T}}(\boldsymbol{z} - \sum_{j\neq i_k}\boldsymbol{a}_j x_j^k)/\|\boldsymbol{a}_{i_k}\|_2^2). \tag{4.10}$$

As $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ is bounded, so is $\{\boldsymbol{a}_i^{\mathrm{T}}(\boldsymbol{z} - \sum_{j\neq i}\boldsymbol{a}_j x_j^k)/\|\boldsymbol{a}_i\|_2^2\}_{k\in\mathbb{N}}$. Let $\widetilde{\mathcal{C}}_i$ denote the compact subset of $\mathbb{R}$ which $\{\boldsymbol{a}_i^{\mathrm{T}}(\boldsymbol{z} - \sum_{j\neq i}\boldsymbol{a}_j x_j^k)/\|\boldsymbol{a}_i\|_2^2\}_{k\in\mathbb{N}}$ belongs to and

$$\widetilde{\Delta} = \min_{i\in\{1,2,\ldots,p\}} \inf_{x\in\widetilde{\mathcal{C}}_i} \|\boldsymbol{a}_i\|_2^2 / \Big\{\frac{d\mathcal{S}^{\mathrm{MMSE}}_{\sigma/\|\boldsymbol{a}_i\|_2}(x)}{dx}\Big\}.$$

Then by Proposition 2.2, we have $H_{i_k}^k(x)$ is $\widetilde{\Delta}$-strongly convex on the compact set $\widetilde{\mathcal{C}}_i$ and thus

$$H_{i_k}^k(\boldsymbol{x}_{i_k}^{k+1}) + \frac{\widetilde{\Delta}}{2}|\boldsymbol{x}_{i_k}^{k+1} - \boldsymbol{x}_{i_k}^k|^2 \leq H_{i_k}^k(\boldsymbol{x}_{i_k}^k).$$

As the remaining coordinates except i_k are fixed, it also gives that

$$H(\boldsymbol{x}^{k+1}) + \frac{\widetilde{\Delta}}{2}|\boldsymbol{x}_{i_k}^{k+1} - \boldsymbol{x}_{i_k}^k|^2 \leq H(\boldsymbol{x}^k). \tag{4.11}$$

Summing up the inequality (4.11) from $k = lp$ to $k = (l+1)p$, we obtain that $\{\boldsymbol{x}^{lp}\}_{l\in\mathbb{N}}$ satisfies the condition C1:

$$\begin{aligned}\frac{\widetilde{\Delta}}{2}\|\boldsymbol{x}^{(l+1)p} - \boldsymbol{x}^{lp}\|_2^2 = \frac{\widetilde{\Delta}}{2}\sum_{k=0}^{p-1}|\boldsymbol{x}_{i_k}^{lp+k+1} - \boldsymbol{x}_{i_k}^{lp+k}|^2 &\leq \sum_{k=0}^{p-1}\big(H(\boldsymbol{x}^{lp+k}) - H(\boldsymbol{x}^{lp+k+1})\big)\\ &= H(\boldsymbol{x}^{lp}) - H(\boldsymbol{x}^{(l+1)p}).\end{aligned} \tag{4.12}$$

Next we show the subsequence $\{\boldsymbol{x}^{lp}\}_{l\in\mathbb{N}}$ also satisfies C2. Since

$$x_{i_j}^{lp+k+1} = \begin{cases} x_{i_j}^{(l+1)p}, & \text{if } j \leq k;\\ x_{i_j}^{lp}, & \text{if } j > k,\end{cases}$$

we can get that

$$\begin{aligned}\partial_{x_{i_k}} H(\boldsymbol{x}^{(l+1)p}) &= \boldsymbol{a}_{i_k}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x}^{(l+1)p} - \boldsymbol{z}) + \partial_{x_{i_k}}\varphi^{\mathrm{MMSE}}_{\sigma/\|\boldsymbol{a}_{i_k}\|_2}(x_{i_k}^{(l+1)p})\\ &= \boldsymbol{a}_{i_k}^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}^{(l+1)p} - \boldsymbol{x}^{lp+k+1}) + \partial_{x_{i_k}} H_{i_k}^{lp+k}(x_{i_k}^{lp+k+1})\\ &= \boldsymbol{a}_{i_k}^{\mathrm{T}}\sum_{j>k}\boldsymbol{a}_{i_j}(\boldsymbol{x}_{i_j}^{(l+1)p} - \boldsymbol{x}_{i_j}^{lp}) + \partial_{x_{i_k}} H_{i_k}^{lp+k}(x_{i_k}^{lp+k+1}).\end{aligned}$$

Combining with (4.8), we have

$$\boldsymbol{a}_{i_k}^{\mathrm{T}} \sum_{j>k} \boldsymbol{a}_{i_j} (\boldsymbol{x}_{i_j}^{(l+1)p} - \boldsymbol{x}_{i_j}^{lp}) \in \partial_{x_{i_k}} H(\boldsymbol{x}^{(l+1)p}),$$

and thus

$$\boldsymbol{w}^{(l+1)p} = \boldsymbol{S}(\boldsymbol{x}^{(l+1)p} - \boldsymbol{x}^{lp}) \in \partial H(\boldsymbol{x}^{(l+1)p}), \tag{4.13}$$

where the matrix $\boldsymbol{S}$ takes the form of

$$\boldsymbol{S}_{st} = \begin{cases} (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})_{st}, & \text{if } i_t > i_s, \\ 0, & \text{othersie.} \end{cases}$$

Taking ℓ_2-norm on both sides of (4.13), we get that

$$\|\boldsymbol{w}^{(l+1)p}\|_2 \leq \rho_{\max}(\boldsymbol{S}) \|\boldsymbol{x}^{(l+1)p} - \boldsymbol{x}^{lp}\|_2. \tag{4.14}$$

That is, the subsequence $\{\boldsymbol{x}^{lp}\}_{l\in\mathbb{N}}$ satisfies C2.

Lastly, we prove the condition C3. Since we have obtained that the sequence $\{\boldsymbol{x}^{lp}\}_{l\in\mathbb{N}}$ is bounded, there exists a convergent subsequence of $\{\boldsymbol{x}^{lp}\}_{l\in\mathbb{N}}$, e.g. $\{\boldsymbol{x}^{k_j}\}_{j\in\mathbb{N}}$ that converges to $\hat{\boldsymbol{x}}$. It can be shown that $\{\boldsymbol{x}^{lp}\}_{l\in\mathbb{N}}$ belongs to the compact subset $\mathcal{C} = \prod_i \mathcal{S}^{\mathrm{MMSE}}_{\sigma/\|\boldsymbol{a}_i\|_2}(\widetilde{\mathcal{C}}_i)$, where $\mathcal{S}^{\mathrm{MMSE}}_{\sigma/\|\boldsymbol{a}_i\|_2}(\widetilde{\mathcal{C}}_i)$ stands for the range of $\mathcal{S}^{\mathrm{MMSE}}_{\sigma/\|\boldsymbol{a}_i\|_2}$ on $\widetilde{\mathcal{C}}_i$. So does $\hat{\boldsymbol{x}}$. Obviously, $H(\boldsymbol{x})$ is continuous on $\mathcal{C}$, then $\lim_{j\to+\infty} H(\boldsymbol{x}^{k_j}) = H(\hat{\boldsymbol{x}})$. The condition C3 is satisfied.

Finally, by Theorem 4.1, we can conclude that the subsequence $\{\boldsymbol{x}^{lp}\}_{l\in\mathbb{N}}$ converges to $\hat{\boldsymbol{x}}$, and $\hat{\boldsymbol{x}}$ is a stationary point of $H(\boldsymbol{x})$. Thus, the whole sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ also converges to $\hat{\boldsymbol{x}}$. Furthermore, since $H(\boldsymbol{x})$ is strongly convex along arbitrary single coordinate by fixing the remaining ones on $\mathcal{C}$, $\hat{\boldsymbol{x}} \in \mathcal{C}$ always attains the minimum of $H(\boldsymbol{x})$ with respect to arbitrary single coordinate, which means that it attains the minimum of the mean square error with respect to arbitrary single coordinate as well. □

5 Conclusion

For Bayesian method, the minimum mean square error estimator makes a good balance between bias and variance, and has the posterior mean as its explicit form. However, it is hard to compute numerically, since it involves multiple integrals of many variables. This paper proposes a simple iterative algorithm to approximate the MMSE estimator, which is easy to implement and efficient in numerical experiments. A complete convergence analysis is given. The analysis and algorithm developed here are then applied to a few given prior distributions with Gaussian noise. In particular, we give a complete analysis and implementation details for estimation under the Bernoulli-uniform sparse prior assumption. We also compare with other available approaches, e.g. the MAP method. Among many properties stated in this paper, our approach gives a stable estimator balancing unbiasedness and sparsity.

References

[1] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[2] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[3] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2006.

[4] Jérôme Bolte Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[5] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[6] Jian-Feng Cai, Bin Dong, Stanley Osher, and Zuowei Shen. Image restoration: Total variation, wavelet frames, and beyond. *Journal of the American Mathematical Society*, 25(4):1033–1089, 2012.

[7] Jian-Feng Cai, Bin Dong, and Zuowei Shen. Image restoration: A wavelet frame based model for piecewise smooth functions and beyond. *Applied and Computational Harmonic Analysis*, 41(1):94–138, July 2016.

[8] Jian-Feng Cai, Stanley Osher, and Zuowei Shen. Linearized bregman iterations for compressed sensing. *Mathematics of computation*, 78(267):1515–1536, 2009.

[9] Jian-Feng Cai, Stanley Osher, and Zuowei Shen. Linearized bregman iterations for frame-based image deblurring. *SIAM Journal on Imaging Sciences*, 2(1):226–252, 2009.

[10] E.J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[11] Raymond H. Chan, Tony F. Chan, Lixin Shen, and Zuowei Shen. Wavelet algorithms for high-resolution image reconstruction. *SIAM Journal on Scientific Computing*, 24(4):1408–1432, 2003.

[12] Robert Crandall, Bin Dong, and Ali Bilgin. Randomized iterative hard thresholding: A fast approximate mmse estimator for sparse approximations. *preprint http://bicmr.pku.edu.cn/ dongbin/Publications/RandIHT.pdf, accessed*, 15:17, 2013.

[13] Ronald A DeVore. Nonlinear approximation. *Acta numerica*, 7:51–150, 1998.

[14] Bin Dong, Zuowei Shen, and Jianbin Yang. Approximation from noisy data. *SIAM Journal on Numerical Analysis*, 59(5):2722–2745, 2021.

[15] David L Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.

[16] David L Donoho et al. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[17] David L Donoho and Iain M Johnstone. Minimax risk overl p-balls forl p-error. *Probability Theory and Related Fields*, 99(2):277–303, 1994.

[18] David L. Donoho and Jain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[19] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Mathematical Society*, 96(456):1348–1360, 2001.

[20] Jianqing Fan, Lingzhou Xue, and Hui Zou. Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics*, 42(3):819, 2014.

[21] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[22] Rémi Gribonval. Should penalized least squares regression be interpreted as maximum a posteriori estimation? *IEEE Transactions on Signal Processing*, 59(5):2405–2410, 2011.

[23] Rémi Gribonval and Mila Nikolova. A characterization of proximity operators. *Journal of Mathematical Imaging and Vision*, 62(6):773–789, 2020.

[24] Rémi Gribonval and Mila Nikolova. On bayesian estimation and proximity operators. *Applied and Computational Harmonic Analysis*, 50:49–72, 2021.

[25] Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

[26] Hui Ji, Yu Luo, and Zuowei Shen. Image recovery via geometrically structured approximation. *Applied and Computational Harmonic Analysis*, 41(1):75–93, 2016.

[27] Iain M Johnstone, Bernard W Silverman, et al. Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.

[28] S.M. Kay. *Fundamentals of Statistical Signal Processing: Detection theory.* Prentice Hall Signal Processing Series. Prentice-Hall PTR, 1998.

[29] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48(3):769–783, 1998.

[30] Marc Lavielle. Bayesian deconvolution of Bernoulli-Gaussian processes. *Signal processing*, 33(1):67–79, 1993.

[31] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles (Paris, 1962)*, pages 87–89. Éditions du Centre National de la Recherche Scientifique, Paris, 1963.

[32] Matan Protter, Irad Yavneh, and Michael Elad. Closed-form MMSE estimation for signal denoising under sparse representation modeling over a unitary dictionary. *IEEE Transactions on Signal Processing*, 58(7):3471–3484, 2010.

[33] Harold W Sorenson and Daniel L Alspach. Recursive bayesian estimation using Gaussian sums. *Automatica*, 7(4):465–479, 1971.

[34] Charles Soussen, Jérôme Idier, David Brie, and Junbo Duan. From Bernoulli–Gaussian deconvolution to sparse signal restoration. *IEEE Transactions on Signal Processing*, 59(10):4572–4584, 2011.

[35] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, April 2010.