Supplement Material for 'Efficient Designs of SLOPE Penalty Sequences in Finite Dimension'

A Introduction to MMSE AMP

We firstly introduce the procedure for general AMP procedure.

$$\begin{aligned} s^{(t+1)} &= X^\top \boldsymbol{Z}^{(t)} + \boldsymbol{\beta}^{(t)} \\ \boldsymbol{\beta}^{(t+1)} &= \eta^{(t+1)}(s^{(t+1)}) \\ \boldsymbol{Z}^{(t+1)} &= y - X\boldsymbol{\beta}^{(t+1)} + \frac{1}{n}\boldsymbol{Z}^{(t)}[\nabla \eta^{(t)}(s^{(t)})] \end{aligned} \tag{A.1}$$

Different η functions give different AMP, e.g. the soft-thresholding η gives the Lasso AMP; the SLOPE proximal operator η gives the SLOPE AMP.

The MMSE AMP adopts the following denoiser $\eta^{(t)}$ [1]

$$\eta_i^{(t)}(s) = \mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta} + \tau_t \boldsymbol{z} = s_i] \quad i = 1, \ldots, p$$

with $\boldsymbol{z} \sim \mathcal{N}(0,1)$. In above, using the state evolution [6], τ_t^2 can be calculated iteratively as:

$$\tau_t^2 = \sigma_\omega^2 + \frac{1}{\delta}\mathbb{E}[(\eta^{(t-1)}(\boldsymbol{\beta} + \tau_{t-1}\boldsymbol{z}) - \boldsymbol{\beta})^2]$$

Assume that the measurement matrix X has i.i.d. $\mathcal{N}(0, 1/n)$ entries. In many scenarios, the denoiser $\eta^{(t)}$ might be hard to calculate. Here we provide a derivation about calculating $\eta^{(t)}$ in the Bernoulli-Gaussian case: we assume that true signal $\boldsymbol{\beta} \overset{i.i.d.}{\sim} \boldsymbol{B}$ where $\boldsymbol{B}$ is a Bernoulli-Gaussian distribution, i.e. $\boldsymbol{\beta}_i = 0$ with probability $e \in [0,1]$, otherwise $\boldsymbol{\beta}_i \sim \mathcal{N}(0, \sigma_{\boldsymbol{B}}^2)$.

$$\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i] = \mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta} \neq 0, \boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i]\mathbb{P}(\boldsymbol{\beta} \neq 0|\boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i) \tag{A.2}$$

It's straightforward to see that, with f denoting the corresponding probability density function,

$$\mathbb{P}(\boldsymbol{\beta} \neq 0|\boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i) = \frac{f(\boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i|\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_B^2))(1-e)}{f(\boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i|\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_B^2))(1-e) + f(\tau_t\boldsymbol{z} = s_i)e} \tag{A.3}$$

Meanwhile. we have

$$\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta} \neq 0, \boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i] = \mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_B^2), \boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i]$$

since $\boldsymbol{\beta} + \tau_t\boldsymbol{z} \sim \mathcal{N}(0, \sigma_B^2 + \tau_t^2)$, conditional expectation on joint normal distribution yields

$$\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_B^2), \boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i] = \frac{\sigma_B^2}{\sigma_B^2 + \tau_t^2}s_i \tag{A.4}$$

(A.3) and (A.4) give a simple way to calculate the denoiser using (A.2).

B Analysis of Gradient in PGD for α

Proof of Theorem 1. Minimizing the estimation error is equivalent to minimizing τ. Since the AMP algorithms are working on the finite dimension, we analyze the finite-size approximation of the state evolution [6, Equation (2.5)]:

$$\tau^2 = \sigma_w^2 + \frac{1}{\delta p}\mathbb{E}\left\|\text{prox}_{J_{\boldsymbol{\alpha}\tau}}\left(\boldsymbol{\beta}+\tau\boldsymbol{Z}\right)-\boldsymbol{\beta}\right\|^2$$

In finite dimensions, the expectation is taken with respect to $\boldsymbol{Z}$. Differentiating both sides of the state evolution with respect to α_i and denoting $\tau' = \frac{\partial \tau}{\partial \alpha_i}$ gives:

$$\begin{aligned}2\tau\tau' &= \frac{\partial}{\partial\alpha_i}\left(\sigma_w^2 + \frac{1}{\delta p}\mathbb{E}\|\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{\beta}+\tau\boldsymbol{Z})-\boldsymbol{\beta}\|^2\right)\\ &= \frac{1}{n}\frac{\partial}{\partial\alpha_i}\sum_{j=1}^{p}\mathbb{E}\left([\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{\beta}+\tau\boldsymbol{Z})]_j-\boldsymbol{\beta}_j\right)^2 \end{aligned} \tag{B.1}$$

Recall η_j represents the j-th element of $\boldsymbol{\eta} := \text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{\beta}+\tau\boldsymbol{Z})$. By chain rule

$$2\tau\tau' = \frac{2}{n}\sum_{j=1}^{p}\mathbb{E}(\eta_j-\beta_j)\frac{\partial\eta_j}{\partial\alpha_i} = \frac{2}{n}\sum_{j=1}^{p}\mathbb{E}(\eta_j-\beta_j)\left[\sum_{k=1}^{p}\frac{d\eta_j}{da_k}\frac{\partial a_k}{\partial\alpha_i}+\frac{d\eta_j}{db_k}\frac{\partial b_k}{\partial\alpha_i}\right]$$

where we define $a_k := \beta_k + \tau Z_k, b_k := \alpha_k\tau$. To calculate the derivatives, we pause to discuss forms of general derivatives of $\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})$. Define

$$\partial_1\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b}) := \text{diag}\Big[\frac{\partial}{\partial a_1},\frac{\partial}{\partial a_2},\dots,\frac{\partial}{\partial a_p}\Big]\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b}) \tag{B.2}$$

$$\partial_2\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b}) := \text{diag}\Big[\frac{\partial}{\partial b_1},\frac{\partial}{\partial b_2},\dots,\frac{\partial}{\partial b_p}\Big]\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b}). \tag{B.3}$$

According to [17, Proof of Fact 3.4] and [6, Proof of Theorem 1], we have

$$[\partial_1\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_j = \frac{1}{\#\{1\le k\le p : |[\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_k| = |[\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_j|\}}$$

and that

$$\frac{d}{da_k}[\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_j = \mathbb{I}\{|\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})|_j = |\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})|_k\}\,\text{sign}(\boldsymbol{\eta}_j\boldsymbol{\eta}_k)[\partial_1\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_j$$

for the derivative regardng the first variable. Recall that the permutation $\sigma : \{1,\dots,p\}\to\{1,\dots,p\}$ is the inverse mapping for ranking of indices such that $|\boldsymbol{\eta}|_{(i)} = |[\boldsymbol{\eta}]_{\sigma(i)}|$. Similarly, according to [6, Proof of Theorem 1]:

$$\begin{aligned}\frac{d}{db_k}[\eta(\boldsymbol{a},\boldsymbol{b})]_j &= -\,\text{sign}([\eta(\boldsymbol{a},\boldsymbol{b})]_{\sigma(k)})\frac{d}{da_{\sigma(k)}}[\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_j\\ &= \mathbb{I}\big\{|\eta(\boldsymbol{a},\boldsymbol{b})|_j = |\eta(\boldsymbol{a},\boldsymbol{b})|_{\sigma(k)}\big\}\,\text{sign}\left(\eta_j\right)\big[\partial_1\eta(\boldsymbol{a},\boldsymbol{b})\big]_j.\end{aligned} \tag{B.4}$$

In addition to I_j defined in Section 2, we let $K_j := \{k : |\eta_{\sigma(k)}| = |\eta_j|\}$, which is the set of ranking indices whose corresponding entries share the same absolute value with η_j. This notion will be used to replace the indicator term $\mathbb{I}\big\{|\eta(\boldsymbol{a},\boldsymbol{b})|_j = |\eta(\boldsymbol{a},\boldsymbol{b})|_{\sigma(k)}\big\}$ above. We can rewrite (B.2) as

$$\begin{aligned}2\tau\tau' &= \frac{2}{n}\sum_{j=1}^{p}\mathbb{E}(\eta_j-\beta_j)\left[\sum_{k\in I_j}\frac{d\eta_j}{da_k}\frac{\partial a_k}{\partial\alpha_i}+\sum_{k\in K_j}\frac{d\eta_j}{db_k}\frac{\partial b_k}{\partial\alpha_i}\right]\\ &= \frac{2}{n}\sum_{j=1}^{p}\mathbb{E}(\eta_j-\beta_j)\,\text{sign}(\eta_j)\left[\frac{1}{|I_j|}\sum_{k\in I_j}\text{sign}(\eta_k)\frac{\partial a_k}{\partial\alpha_i}-\frac{1}{|K_j|}\sum_{k\in K_j}\frac{\partial b_k}{\partial\alpha_i}\right]\\ &= \frac{2}{n}\sum_{j=1}^{p}\mathbb{E}(\eta_j-\beta_j)\,\text{sign}(\eta_j)\left[\frac{1}{|I_j|}\sum_{k\in I_j}\text{sign}(\eta_k)Z_k\tau'-\frac{1}{|K_j|}\sum_{k\in K_j}(\alpha_k\tau'+\mathbb{I}\{k=i\}\tau)\right]\end{aligned}$$

Merging the terms containing the derivative τ' on one side gives

$$\begin{aligned}&\frac{1}{n}\sum_{j\in I_{\sigma(i)}}\mathbb{E}(\eta_j-\beta_j)\operatorname{sign}(\eta_j)/|K_j|\\&=\frac{1}{n}\sum_{j=1}^{p}\mathbb{E}(\eta_j-\beta_j)\operatorname{sign}(\eta_j)\left[\frac{1}{|I_j|}\sum_{k\in I_j}\operatorname{sign}(\eta_k)Z_k\tau'-\frac{1}{|K_j|}\sum_{k\in K_j}\alpha_k\tau'\right]-\tau\tau'\end{aligned}$$

Notice that $|I_j|=|K_j|$ due to σ being a permutation, we can simplify above as

$$\frac{\partial\tau}{\partial\alpha_i}=\mathbb{E}\frac{1}{|I_{\sigma(i)}|D(\boldsymbol{\alpha},\tau)}\sum_{j\in I_{\sigma(i)}}(\eta_j-\beta_j)\operatorname{sign}(\eta_j)\tau \tag{B.5}$$

where $D(\boldsymbol{\alpha},\tau)$ in the denominator is

$$D(\boldsymbol{\alpha},\tau)=-n\tau+\sum_{j=1}^{p}\mathbb{E}\frac{1}{|I_j|}(\eta_j-\beta_j)\operatorname{sign}(\eta_j)\sum_{k\in I_j}(\operatorname{sign}(\eta_k)Z_k-\alpha_{\sigma^{-1}(k)})$$

We next show that $D(\boldsymbol{\alpha},\tau)$ is always negative. Firstly observe from (2.3) that

$$\tau^2>\frac{1}{n}\sum_{j=1}^{p}\mathbb{E}(\eta_j-\beta_j)^2 \tag{B.6}$$

Now for the set I_i with a fixed index i,

$$\sum_{j\in I_i}(\eta_j-\beta_j)^2\geq\frac{1}{|I_i|}(\sum_{j\in I_i}|\eta_j-\beta_j|)^2 \tag{B.7}$$

$$\geq\frac{1}{|I_i|}(\sum_{j\in I_i}(\eta_j-\beta_j)\operatorname{sign}(\eta_j))^2 \tag{B.8}$$

$$=\frac{1}{|I_i|}\sum_{j\in I_i}(\eta_j-\beta_j)\operatorname{sign}(\eta_j)\sum_{k\in I_i}\tau Z_k\operatorname{sign}(\eta_k)-\alpha_{\sigma^{-1}(k)}\tau \tag{B.9}$$

$$\geq\frac{\tau}{|I_i|}\sum_{j\in I_i}(\eta_j-\beta_j)\operatorname{sign}(\eta_j)\sum_{k\in I_j}Z_k\operatorname{sign}(\eta_k)-\alpha_{\sigma^{-1}(k)} \tag{B.10}$$

This in turn implies that

$$\sum_{j=1}^{p}(\eta_j-\beta_j)^2=\sum_{j=1}^{p}\frac{1}{|I_j|}\sum_{k\in I_j}(\eta_k-\beta_k)^2\geq\sum_{j=1}^{p}\frac{\tau}{|I_j|}(\eta_j-\beta_j)\operatorname{sign}(\eta_j)\sum_{k\in I_j}Z_k\operatorname{sign}(\eta_k)-\alpha_{\sigma^{-1}(k)} \tag{B.11}$$

Combining with (B.6) yields $D<0$.

□

C Analysis of Projection in PGD for α

C.1 Characterization of projection on $\mathcal{S}$

We firstly prove that Algorithm 1 indeed finds the projection. To do so we firstly provide a detailed characterization of the projection, then prove that the output of Algorithm 1 matches the form of projection. We start by defining *blocks* and *segmentation blocks*, upon which our proof highly relies. Suppose $\boldsymbol{\gamma}=\{\gamma_1,\dots,\gamma_p\}$, *blocks* are subsequences defined as $B(\boldsymbol{\gamma},u):=\{\gamma_u,\dots,\gamma_{u+L(\boldsymbol{\gamma},u)-1}\}$ where length $L(\boldsymbol{\gamma},u)$ is defined as

$$L(\boldsymbol{\gamma},u)=\begin{cases}L^* & \text{if } L^*\neq\emptyset\\ p & \text{otherwise}\end{cases} \tag{C.1}$$

where

$$L^* \triangleq \min\left\{1 \le L \le p-u \middle| \forall 0 \le k \le p-u-L, \frac{1}{k+1}\sum_{i=0}^{k}\gamma_{u+L+i} < \frac{1}{L}\sum_{i=0}^{L-1}\gamma_{u+i}\right\}$$

Roughly speaking, $L(\boldsymbol{\gamma}, u)$ is the minimum value of a finite set (truncated at p when the set is empty). For each element L in this set, the average value in sequence $\{\gamma_u, \dots, \gamma_{u+L-1}\}$ is always larger than that of arbitrary sequence $\{\gamma_{u+L}, \dots, \gamma_{u+L+k}\}$ whose left start is γ_{u+L}. With such definition of blocks, we can now segment $\boldsymbol{\gamma}$ into $q \le p$ blocks:

$$\boldsymbol{\gamma} = \{B(\boldsymbol{\gamma}, 1), B(\boldsymbol{\gamma}, L(\boldsymbol{\gamma}, 1)+1), B(\boldsymbol{\gamma}, L(\boldsymbol{\gamma}, L(\boldsymbol{\gamma}, 1)+1)+L(\boldsymbol{\gamma}, 1)+1), \dots\} \triangleq \{B_1, \dots, B_q\}$$

We call $B_1, \dots, B_q$ *segmentation blocks* for vector $\boldsymbol{\gamma}$. It's straightforward to see that $B_k = B(\boldsymbol{\gamma}, L_k)$ where L_k satisfies $L_1 = L(\boldsymbol{\gamma}, 1)$ and

$$L_k = L(\boldsymbol{\gamma}, \sum_{i=1}^{k-1} L_i + 1)$$

Our result shows that for input vector γ, its projection vector $\Pi_{\mathcal{S}}(\boldsymbol{\gamma})$ takes identical values inside each of the segmentation blocks. Before formally stating the theorem, We first highlight the following fact that will be frequently used in the proof of the theorem.

Fact C.1. *For two sequences of length p: $\{a_i\}$ and $\{b_i\}$, if $\sum a_i = \sum b_i$, then function $g(C) := \sum(b_i - a_i + C)^2$ is monotonically increasing with respect to $|C|$.*

Proof. Notice that

$$\sum(b_i - a_i + C)^2 = \sum(b_i - a_i)^2 + \sum 2C(b_i - a_i) + pC^2 = pC^2 + \sum(b_i - a_i)^2$$

Hence $g(C)$ is is monotonically increasing with respect to $|C|$. □

Theorem 3. *Let B denote the segmentation block that contains γ_i, then*

$$(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i = \max\left\{\frac{1}{|B|}\sum_{\gamma_j \in B}\gamma_j, 0\right\}$$

Proof. The proof consists of two steps. In the first step, we prove that for each segmentation block B, the projection of each coordinates share the same value. That is, $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i = \mathcal{C}(B)$ as long as $\gamma_i \in B$. In the second step, we show that this constant is the mean of the block truncated at 0: $\mathcal{C}(B) = \max\left\{\frac{1}{|B|}\sum_{\gamma_j \in B}\gamma_j, 0\right\}$.

Step 1 Without loss of generality, we consider $B = B(\boldsymbol{\gamma}, u)$. We know from definition of blocks that $\forall 1 \le l \le L-1$, $\exists k_l$ s.t. $\frac{1}{k_l}\sum_{i=1}^{k_l}\gamma_{u+l-1+i} \ge \frac{1}{l}\sum_{i=1}^{l}\gamma_{u+i-1}$. We use induction to prove that $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u = (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l}$, $\forall 1 \le l \le L(\gamma, u) - 1$. For $l = 1$, assume $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u > (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+1}$. Consider two cases: (i) $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u > \gamma_u$. (ii) $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u \le \gamma_u$. We now show that both cases lead to contradiction and hence do not hold. In case (i), we consider

$$(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i = \begin{cases}\max\{\gamma_u, (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+1}\} & \text{if } i = u\\ (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i & \text{otherwise}\end{cases}$$

then obviously,

$$\left|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_u - \gamma_u\right| < |(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u - \gamma_u|$$

which leads to that $\frac{1}{2}\|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2 < \frac{1}{2}\|(\Pi_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2$. This contradicts to the definition of projection. In case (ii), from definition of blocks we have that $\exists k_0 \ge 1$ s.t. $\frac{1}{k_0}\sum_{i=1}^{k_0}\gamma_{u+i} \ge \gamma_u$. Consider

$$(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i = \begin{cases}(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u & \text{if } i \in \{u+1, \dots, u+k_0\}\\ (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i & \text{otherwise}\end{cases}$$

Notice that $\frac{1}{k_0}\sum_{i=1}^{k_0}\gamma_{u+i} \geq \gamma_u \geq (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u > (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+1}$, we have for $i \in \{u+1,\ldots,u+k_0\}$, $\left|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i - (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i\right|$ is a constant independent of i and that

$$\left|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i - \frac{1}{k_0}\sum_{i=1}^{k_0}\gamma_{u+i}\right| < \left|(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i - \frac{1}{k_0}\sum_{i=1}^{k_0}\gamma_{u+i}\right|$$

According to Fact C.1, we define substitution for $i \in \{u+1,\ldots,u+k_0\}$: $b_i = \frac{1}{k_0}\sum_{i=1}^{k_0}\gamma_{u+i}$, $a_i = \gamma_{u+i}$, $b_i + C_1 = (\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i$ and $b_i + C_2 = (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i$. Then since $|C_1 < C_2|$, we have $\frac{1}{2}\|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2 < \frac{1}{2}\|(\Pi_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2$, which contradicts to the definition of projection.

Now assume the statement holds for $1 \leq l \leq l_0 - 1$, that is $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u = \cdots = (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l_0-1}$, we want to prove that $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u = (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l_0}$. Since the projection is on $\mathcal{S}$, by definition we know $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u$ can never be smaller than $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l_0}$. We now assume $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u > (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l_0}$ and consider two cases: (i) $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u > \frac{1}{l_0}\sum_{i=0}^{l_0-1}\gamma_{u+i}$. (ii) $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u \leq \frac{1}{l_0}\sum_{i=0}^{l_0-1}\gamma_{u+i}$. To complete the proof, it suffices for us to show that neither of the cases can hold without contradictions. In case (i), we consider

$$(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i = \begin{cases} \max\{\frac{1}{l_0}\sum_{j=0}^{l_0-1}\gamma_{u+j}, (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l_0}\} \\ \qquad\qquad \text{if } i \in \{u,\ldots,u+l_0-1\} \\ (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i \qquad\qquad \text{otherwise} \end{cases}$$

then obviously for $i \in \{u,\ldots,u+l_0-1\}$, $\left|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i - (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i\right|$ is a constant independent of i and that

$$\left|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i - \frac{1}{l_0}\sum_{i=0}^{l_0-1}\gamma_{u+i}\right| < \left|(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i - \frac{1}{l_0}\sum_{i=0}^{l_0-1}\gamma_{u+i}\right|$$

According to Fact C.1, using the same substitution as that in analysis of $l = 1$, we have that $\frac{1}{2}\|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2 < \frac{1}{2}\|(\Pi_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2$, which makes contradiction to the definition of projection. In case (ii), from definition of blocks we have that $\exists k_0 \geq 1$ s.t. $\frac{1}{k_0}\sum_{i=1}^{k_0}\gamma_{u+l_0-1+i} \geq \frac{1}{l_0}\sum_{i=0}^{l_0-1}\gamma_{u+i}$. Now we consider

$$(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i = \begin{cases} (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u & \text{if } i \in \{u+l_0,\ldots,u+l_0-1+k_0\} \\ (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i & \text{otherwise} \end{cases}$$

Notice that $\frac{1}{k_0}\sum_{i=1}^{k_0}\gamma_{u+l_0-1+i} \geq \frac{1}{l_0}\sum_{i=0}^{l_0-1}\gamma_{u+i} \geq (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u > (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l_0}$, we have for $i \in \{u+l_0,\ldots,u+l_0-1+k_0\}$, $\left|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i - (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i\right|$ is a constant independent of i and that

$$\left|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i - \frac{1}{k_0}\sum_{i=0}^{k_0-1}\gamma_{u+l_0+i}\right| < \left|(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i - \frac{1}{k_0}\sum_{i=0}^{k_0-1}\gamma_{u+l_0+i}\right|$$

Again according to Fact C.1, we have $\frac{1}{2}\|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2 < \frac{1}{2}\|(\Pi_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2$, which contradicts to the definition of projection. This implies that it can never happen that $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u > (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l_0}$, which completes the induction. We have proved that $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u = \cdots = (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+L(\boldsymbol{\gamma},u)-1} \triangleq \mathcal{C}(B(u))$ for each segmentation block $B(u)$ of vector $\boldsymbol{\gamma}$.

Step 2 Now we already know that inside each segmentation block, the projection of each coordinate is a constant $\mathcal{C}(B)$, we now optimize the sequence $\{\mathcal{C}(B_i)\}_{i=1}^q$. According to Fact C.1, inside each B_i, the optimal constant (i.e. constant gives smallest ℓ_2 error $\operatorname{argmin}_{C\geq 0}\frac{1}{2}\sum_{\gamma_j\in B_i}(\gamma_j - C)^2$) is : $\max\left\{\frac{1}{|B_i|}\sum_{\gamma_j\in B_i}\gamma_j, 0\right\}$. Meanwhile, it's feasible to set

$$(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i = \max\left\{\frac{1}{|B|}\sum_{\gamma_j\in B}\gamma_j, 0\right\}$$

since we have that $\max\left\{\frac{1}{|B_i|}\sum_{\gamma_j\in B_i}\gamma_j, 0\right\} \geq \max\left\{\frac{1}{|B_{i+1}|}\sum_{\gamma_j\in B_{i+1}}\gamma_j, 0\right\}$ by definition of blocks. This wraps up the proof.

□

C.2 Proof of Theorem 2

We next prove the validity of Algorithm 1.

Proof. Suppose $\boldsymbol{\gamma}$ has segmentation blocks $B_1, \ldots, B_q$, we firstly prove that $(\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_i = (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i$ for $i \le |B_1|$. We let $\gamma_j(t)$ denote the value of γ_j at the moment i was assigned from t to $t+1$ in Algorithm 1 (i.e. the time when first t iterations are finished). We also let $\gamma_j(0)$ denote the initial value of γ_j in the input. Then clearly $(\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_j = \max\{\gamma_j(p), 0\}$. During the value-averaging step, the algorithm is constantly transporting values from elements with larger index to those with smaller. Hence it's straightforward to see that

$$\sum_{j=1}^{J} \gamma_j(t) \ge \sum_{j=1}^{J} \gamma_j(t-1) \tag{C.2}$$

for arbitrary $J, t \in \{1, \ldots, p\}$. First assume $\gamma_1(p) = \cdots = \gamma_{\widetilde{L}_1}(p) > \gamma_{\widetilde{L}_1+1}(p)$. Since Algorithm 1 only involves averaging values among subsequences, we have that $\sum_{j=1}^{p} \gamma_j(p) = \sum_{j=1}^{p} \gamma_j$. Moreover since $\gamma_{\widetilde{L}_1}(p) > \gamma_{\widetilde{L}_1+1}(p)$, there's no value-averaging steps between any one of the first $\widetilde{L}_1$ elements and one of the rest elements. This implies

$$\sum_{j=1}^{\widetilde{L}_1} \gamma_j(p) = \sum_{j=1}^{\widetilde{L}_1} \gamma_j \tag{C.3}$$

By definition of blocks, we know that $\exists k$ such that $\frac{1}{k}\sum_{i=1}^{k} \gamma_{\widetilde{L}_1+i} \ge \frac{1}{\widetilde{L}_1}\sum_{i=1}^{\widetilde{L}_1} \gamma_i = \gamma_1(p)$. By (C.2) we have that

$$\frac{1}{k}\sum_{i=1}^{k} \gamma_{\widetilde{L}_1+i} \le \frac{1}{k}\sum_{i=1}^{k} \gamma_{\widetilde{L}_1+i}(p) \le \gamma_{\widetilde{L}_1+1}(p)$$

Together with above, this implies that $\gamma_1(p) \le \gamma_{\widetilde{L}_1+1}(p)$, which contradicts to the assumption. Hence we have that $\widetilde{L}_1 \ge L_1$.

On the other hand, if $\widetilde{L}_1 > L_1$, then at the moment i is assigned to be $\widetilde{L}_1 + 1$ in the algorithm (i.e. the time when first $\widetilde{L}_1$ iterations are finished), we must have that

$$\frac{\sum_{j=1}^{\widetilde{L}_1} \gamma_j(\widetilde{L}_1 - 1)}{\widetilde{L}_1} \ge \frac{\sum_{j=1}^{L_1} \gamma_j(\widetilde{L}_1 - 1)}{L_1}$$

This implies that

$$\frac{\sum_{j=L_1+1}^{\widetilde{L}_1} \gamma_j(\widetilde{L}_1 - 1)}{\widetilde{L}_1 - L_1} \ge \frac{\sum_{j=1}^{L_1} \gamma_j(\widetilde{L}_1 - 1)}{L_1} \tag{C.4}$$

By (C.2) we have

$$\frac{\sum_{j=1}^{L_1} \gamma_j}{L_1} \le \frac{\sum_{j=1}^{L_1} \gamma_j(\widetilde{L}_1 - 1)}{L_1} \tag{C.5}$$

Meanwhile at $t = \widetilde{L}_1 - 1$, the sum of first L_1 terms is the same as that in $\boldsymbol{\gamma}$. This implies

$$\begin{aligned} \sum_{j=L_1+1}^{\widetilde{L}_1} \gamma_j(\widetilde{L}_1 - 1) &= \sum_{j=1}^{L_1} \gamma_j + \sum_{j=L_1+1}^{\widetilde{L}_1} \gamma_j - \sum_{j=1}^{L_1} \gamma_j(\widetilde{L}_1 - 1) \\ &\le \sum_{j=L_1+1}^{\widetilde{L}_1} \gamma_j \end{aligned} \tag{C.6}$$

where the last inequality is given by (C.2). Combining (C.4), (C.5) and ((C.6)) yields

$$\frac{\sum_{j=L_1+1}^{\widetilde{L}_1} \gamma_j}{\widetilde{L}_1 - L_1} \ge \frac{\sum_{j=1}^{L_1} \gamma_j}{L_1}$$

This contradicts to definition of L_1 in (C.1). Hence we have that $\widetilde{L}_1 = L_1$. This means $\gamma_1(p) = \cdots = \gamma_{L_1}(p) > \gamma_{L_1+1}(p)$. Recall that $(\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_j = \max\{\gamma_j(p), 0\}$, this together with (C.3) yields

$$\begin{aligned}(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_1 &= \max\left\{\frac{1}{|B_1|}\sum_{j=1}^{L_1}\gamma_j, 0\right\} = (\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_1 \\ &= \cdots = (\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_{L_1} > (\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_{L_1+1}\end{aligned}$$

Now we have prove that $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i = (\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_i$ for $i \leq |B_1|$ and that there is no interaction between element in B_1 and that outside B_1. This implies that the existence of B_1 does *not* affect the rest of output values $(\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_{i>|B_1|}$. Hence we can ignore B_1 and repeat exactly the same procedure to prove that $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i = (\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_i$ when $|B_1|+1 \leq i \leq |B_2|$ and that there is no interactions between element in B_2 and that outside B_2. Iteratively we can prove $\Pi_{\mathcal{S}}(\boldsymbol{\gamma}) = \Lambda_{\mathcal{S}}(\boldsymbol{\gamma})$

□