New Risk Bounds for 2D Total Variation Denoising

Sabyasachi Chatterjee and Subhajit Goswami

***Abstract*—2D Total Variation Denoising (TVD) is a widely used technique for image denoising. It is also an important nonparametric regression method for estimating functions with heterogenous smoothness. Recent results have shown the TVD estimator to be nearly minimax rate optimal for the class of functions with bounded variation. In this paper, we complement these worst case guarantees by investigating the adaptivity of the TVD estimator to functions which are piecewise constant on axis aligned rectangles. We rigorously show that, when the truth is piecewise constant with few pieces, the ideally tuned TVD estimator performs better than in the worst case. We also study the issue of choosing the tuning parameter. In particular, we propose a fully data driven version of the TVD estimator which enjoys similar worst case risk guarantees as the ideally tuned TVD estimator.**

***Index Terms*—Nonparametric regression, total variation denoising, tuning free estimation, estimation of piecewise constant functions, tangent cone, gaussian width, recursive partitioning.**

I. Introduction

Total variation denoising (TVD) is a standard technique to do noise removal in images. This technique was first proposed in [33] and has since then been heavily used in the image processing community. It is well known that TVD gets rid of unwanted noise and also preserves edges in the image (see [37]). For a survey of this technique from an image analysis point of view; see [6] and references therein.

The success of the TVD technique as a denoising mechanism motivates us to revisit this problem from a statistical perspective. In this paper, we are interested in the following statistical estimation problem. Consider observing $y = \theta^* + \sigma Z$ where $y \in \mathbb{R}^{n \times n}$ is a noisy matrix/image, θ^* is the true underlying matrix/image, Z is a noise matrix consisting of independent standard Gaussian entries and σ is an unknown standard deviation of the noise entries. Thus, in this setting, the image denoising problem is cast as a Gaussian mean estimation problem. Before defining the TVD estimator in this context, let us define total variation of an arbitrary matrix.

Let us denote the $n \times n$ two dimensional *grid graph* by L_n and denote its edge set by E_n. More precisely, the vertices in L_n correspond to the pairs $(i,j) \in [n] \times [n]$ and its edge set E_n consists of:

$$\text{all } ((i,j),(k,\ell)) \in L_n \times L_n \text{ with } |i-j| + |k-\ell| = 1\,.$$

We will use L_n interchangeably for the graph as well as the underlying set of vertices. Now, thinking of $\theta \in \mathbb{R}^{n \times n}$ as a function on L_n let us define

$$\mathrm{TV}_{\mathrm{norm}}(\theta) := \frac{1}{n} \sum_{(u,v) \in E_n} |\theta_u - \theta_v| = \frac{1}{n} \|D\theta\|_1 \tag{I.1}$$

where D is the usual edge vertex incidence matrix of size $2n(n-1) \times n^2$. The $1/n$ factor is just a normalizing factor so that if $\theta_{ij} = f(i/n, j/n)$ for some underlying differentiable function on the unit square then $\mathrm{TV}_{\mathrm{norm}}(\theta)$ is precisely the discretized Riemann approximation for $\int_{[0,1]^2} \left|\frac{\partial f(x,y)}{\partial x}\right| + \left|\frac{\partial f(x,y)}{\partial y}\right|$. This $1/n$ scaling is termed as the *canonical scaling* in [34]. The above notion of total variation extends the definition of *variation* from differentiable functions on the unit square to arbitrary matrices. We can now define the TVD estimator, which is our main object of study.

$$\hat{\theta}_{\mathbf{V}} := \operatorname*{argmin}_{\theta \in \mathbb{R}^{n \times n} : \mathrm{TV}_{\mathrm{norm}}(\theta) \leq \mathbf{V}} \|y - \theta\|^2$$

where $\|.\|$ throughout this paper will denote the usual Frobenius norm for matrices. The TVD estimator is actually a family of estimators indexed by the tuning parameter $\mathbf{V} > 0$. We will measure the performance of our estimator in terms of its normalized mean squared error (MSE) defined as

$$\mathrm{MSE}(\hat{\theta}_{\mathbf{V}}, \theta^*) := \mathbb{E}_{\theta^*} \frac{\|\hat{\theta}_{\mathbf{V}} - \theta^*\|^2}{N}$$

where throughout this paper we denote $N = n^2$.

We defined the TVD estimator in its constrained form, however the penalized version is also popular in the literature, which is defined as follows:

$$\hat{\theta}_{\lambda} := \operatorname*{argmin}_{\theta \in \mathbb{R}^{n \times n}} \|y - \theta\|^2 + \lambda\, \mathrm{TV}_{\mathrm{norm}}(\theta)$$

where $\lambda > 0$ is a tuning parameter. In this paper, we focus on the analysis of the constrained version.

A. Background and Motivation

The 1D version of this problem is a well studied problem (see, e.g. [38]) in nonparametric regression. In this setting, we again have $y = \theta^* + \sigma Z$ as before, where y, θ^*, Z are now vectors instead of matrices. The total variation of a vector $v \in \mathbb{R}^n$ can now be defined as

$$\mathrm{TV}(v) := \sum_{i=1}^{n-1} |v_{i+1} - v_i|.$$

Manuscript received October 12, 2019; revised January 22, 2021; accepted January 30, 2021. Date of publication February 16, 2021; date of current version May 20, 2021. This work was supported in part by the NSF grant and in part by the IDEX grant from Paris-Saclay. *(Corresponding author: Sabyasachi Chatterjee.)*

Sabyasachi Chatterjee is with the Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820 USA (e-mail: sc1706@illinois.edu).

Subhajit Goswami is with the School of Mathematics, Tata Institute of Fundamental Research, Mumbai 400005, India (e-mail: goswami@math.tifr.res.in).

Communicated by M. Rodrigues, Associate Editor for Signal Processing.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2021.3059657.

Digital Object Identifier 10.1109/TIT.2021.3059657

Again the above definition can be seen as a discrete Riemann approximation to $\int_{[0,1]} \left|f^{'}(x)\right|dx$ when $v_i = f(i/n)$ for some differentiable function f. The constrained and the penalized versions of the TVD estimator can now be defined analogously. The penalized form seems to be more popular in the existing literature; in this case the TVD estimator is often referred to as fused lasso (see [38], [32]). In this 1D setting, it is known (see, e.g. [13], [26]) that the TVD estimator is minimax rate optimal on the class of all bounded variation signals $\{\theta : \mathrm{TV}(\theta) \leq \mathbf{V}\}$ for $\mathbf{V} > 0$. It is also shown in [13] that no estimator, which is a linear function of y, can attain this minimax rate.

It is also worthwhile to mention here that TVD in the 1D setting has been studied as part of a general family of estimators which penalize discrete derivatives of different orders. These estimators have been studied in [36], [39] and by [21] who coined the name *trend filtering*. A continuous version of these estimators, where discrete derivatives are replaced by continuous derivatives, was proposed much earlier in the statistics literature by [26] under the name *locally adaptive regression splines*.

Total variation of a signal can actually be defined over an arbitrary graph as the sum of absolute differences of the signal across edges of the graph. Trend Filtering on general graphs has been a popular research topic in the recent past; see [41], [25]. A more recent paper, [27], studies TVD on tree graphs. The 1D setting corresponds to the chain graph on n vertices whereas the 2D setting corresponds to the 2D lattice graph on $n^2 = N$ vertices.

The 2D TVD problem, while being much less studied than in its 1D counterpart, has enjoyed a recent surge of interest. Worst case performance of the TVD estimator has been studied in [19], [34], [29]. These results show that like in the 1D setting, the 2D TVD estimator is nearly minimax rate optimal over the class $\{\theta \in \mathbb{R}^{n \times n} : \mathrm{TV}_{\mathrm{norm}}(\theta) \leq \mathbf{V}\}$ of bounded variation signals. In fact, [34] also generalize the result of [13] and prove that no linear function of y can attain the minimax rate in the 2D setting as well. A representative of the state of the art risk bound for the TVD estimator in 2D setting is due to [19] (see also [29]). They studied the penalized form of the TVD estimator and proved that there exist universal constants C, $c > 0$ such that by setting $\lambda = c\sigma \log n$ (where σ is known), one gets

Theorem I.1 (Hütter and Rigollet):

$$\mathrm{MSE}(\hat{\theta}_\lambda, \theta^*) \leq C(\log n)^2 \ A$$

where

$$A = \min\{\sigma \frac{\mathrm{TV}_{\mathrm{norm}}(\theta^*)}{\sqrt{N}}, \sigma^2 \frac{\|D\theta^*\|_0}{N}\}$$

and $\|\cdot\|_0$ is the usual ℓ_0 norm.

For convenience, we will henceforth use the usual $O(\cdot)$ notation to compare sequences. We write $a_n = O(b_n)$ if there exists a constant $C > 0$ such that $a_n \leq C\ b_n$ for all sufficiently large n. We also use $a_n = \tilde{O}(b_n)$ to denote $a_n = O(b_n (\log n)^C)$ for some $C > 0$.

In words, the bound in Theorem I.1 is a minimum of two terms. The first term gives the ℓ_1 rate scaling like $O(1/\sqrt{N})$ for bounded variation functions. The second one is the ℓ_0 rate which can be much faster than the $O(1/\sqrt{N})$ rate if $|D\theta^*|_0$ is small enough. In spite of the above works, there are still a couple of unexplored aspects regarding 2D TVD, specifically its adaptivity to piecewise constant signals and minimax optimality without tuning, which are the focus of the present paper. We discuss them now.

1) Adaptivity to Piecewise Constant Signals: Observe that the total variation semi norm is a convex relaxation for the number of times the true signal θ^* changes values along the neighbouring vertices. This fact suggests that the TV estimator *might* perform very well if the true signal is indeed piecewise constant. This phenomenon is now fairly well understood in the 1D setting. In this setting, suppose that the true vector θ^* is piecewise constant with $k+1$ contiguous pieces or blocks. Given data $y \sim N_n(\theta^*, \sigma^2\ I_n)$, an oracle estimator, which knows the locations of the jumps, would just estimate the signal θ^* by the mean of the data vector y within each block. It can be easily checked that the oracle estimator will have MSE bounded by $\sigma^2(k+1)/n$. Recent works (see [12], [25]) studied the penalized TVD estimator and showed that if the *minimum length* of the blocks where θ^* is constant is not too small (scales like $O(n/k)$) and if the tuning parameter λ is set to be equal to an appropriate function of the unknown σ and n, then an oracle risk $O(k/n)$ could be achieved up to some additional logarithmic factors in k and n. In [17], this adaptive behaviour was established for the ideally tuned constrained form of the estimator with slightly better log factors. Thus, we can say that in the 1D setting, the TVD estimator is optimally adaptive to piecewise constant signals.

This motivates us to wonder whether similar adaptivity holds in the 2D setting. In this paper, we investigate adaptivity to signals/matrices which are *piecewise constant on* $k << N$ *axis aligned rectangles*. Such adaptivity of the 2D TVD estimator has not been explored at all in the literature. Estimation of functions which are piecewise constant on axis aligned rectangles are naturally motivated by methodologies such as CART (see e.g [3]) which produce outputs of the same form. Recently, adaptation to piecewise constant structure on rectangles has been of interest in the nonparametric shape constrained function estimation literature also (see Theorem 2.3 in [10] and Theorems 2 and 5 in [18]). See Section III-E where we discuss some even more recent (which appeared after we uploaded this paper) works about estimating piecewise constant functions on axis aligned rectangles. Here is the main question that we address in this paper.

Q1: If the underlying θ^* *is piecewise constant on at most* $k << N$ *axis aligned rectangles; can the ideally tuned TVD estimator attain a faster rate of convergence than the* $\tilde{O}(1/\sqrt{N})$ *rate*?

Basically we are asking the question whether the ideally tuned TVD estimator adapts to truths which are piecewise constant on a few axis aligned rectangles, which is a different notion of sparsity than the sparsity constraint of $\|D\theta^*\|_0$ being small. As a simple instance of θ^* being piecewise constant on rectangles, consider θ^* to be of the following form:

$$\theta^* = \left[\ \mathbf{0}_{n \times n/2} \quad \mathbf{1}_{n \times n/2}\ \right]$$

In this case, we have $\|D\theta^*\|_0 = O(\sqrt{N})$ and $\mathrm{TV}_{\mathrm{norm}}(\theta^*) = O(1)$. Note that the ℓ_0 bound of [19] will give us an upper bound on the MSE scaling like $\tilde{O}(1/\sqrt{N})$ which is already given by the ℓ_1 bound. Thus, the result of [19] does not help in answering our question and suggests there is no adaptation. In Theorem II.2 of this paper, we show that the ideally tuned TVD estimator indeed adapts to piecewise constant matrices on axis aligned rectangles and provably attains a rate of convergence scaling like $\tilde{O}(1/N^{3/4})$ which is strictly faster than the ℓ_1 rate $\tilde{O}(1/\sqrt{N})$. However, we also show that this $\tilde{O}(1/N^{3/4})$ rate is tight and thus the TVD estimator is not able to attain the $\tilde{O}(1/N)$ parametric rate that would be achieved by an oracle estimator. This is the main contribution of this paper and is the first result of its type in the literature as far as we are aware.

2) Minimax Rate Optimality Without Tuning: Existing results such as Theorem I.1, along with minimax lower bounds shown in [34], show that the $\tilde{O}(\frac{\mathbf{V}}{\sqrt{N}})$ rate attained by the penalized TVD estimator is near minimax rate optimal. Thus we can say that the penalized TVD estimator is near minimax rate optimal over the parameter space $\{\theta \in \mathbb{R}^{n\times n} : \mathrm{TV}_{\mathrm{norm}}(\theta) \leq \mathbf{V}\}$, *simultaneously* over $\mathbf{V}$ and N. However, this penalized TVD estimator needs to set a tuning parameter λ which depends on the unknown σ and an implicit constant C which can be potentially difficult to set in practice. This naturally raises a question which is unresolved in the literature so far as we are aware:

Q2: Does there exist a completely data driven estimator which does not depend on any unknown parameters of the problem and yet achieves MSE scaling like $\tilde{O}(\frac{\mathbf{V}}{\sqrt{N}})$, thus being simultaneously minimax rate optimal over $\mathbf{V}$ and N?

In Theorem II.7 of this paper we answer this question in the affirmative by constructing such a fully data driven estimator.

The rest of the paper is organised as follows. In Section II, we state our main results. Then in Section III, we discuss connections of our results with some recent works and also present simulation studies which support and verify our main theorems. The proofs of our main results involve sharp bounds on the *Gaussian widths* (see (II.1) in Section II-B1) for some special classes of matrices. We obtain these bounds based on a generic approach which we detail in Section IV. The next five sections describe the proofs of our main theorems and intermediate results. Section IX-B is the appendix which contains proofs of some auxiliary results.

Instructions for the reader

In all the proofs of our results from Section V onwards, we will use $\mathrm{TV}(\cdot)$ to denote the *unnormalized* version of (I.1). More precisely, for a $n \times n$ matrix θ we denote

$$\mathrm{TV}(\theta) := \sum_{(u,v)\in E_n} |\theta_u - \theta_v|. \tag{I.2}$$

We adopt this convention because we believe it is easier to read and interpret the proofs with the unnormalized definition while it is instructive to use the normalized version for our theorems to facilitate interpretation of the risk bounds as a function of the sample size $N = n^2$. Also we will generically use V to denote the unnormalized total variation whereas in Sections I–III we use *bold* $\mathbf{V}$ to denote the *corresponding* normalized total variation. In all our theorems presented in the next section we use *bold* $\mathbf{V}^*$ to denote $\mathrm{TV}_{\mathrm{norm}}(\theta^*)$ where θ^* is the underlying true matrix and in all our proofs we use $V^* = \mathrm{TV}(\theta^*)$ for the corresponding unnormalized version.

II. Main Results

A. Constrained TVD

Our first result states a risk bound of $\hat{\theta}_{\mathbf{V}}$ under the bounded variation constraint.

Theorem II.1: Let θ^* be an arbitrary $n \times n$ matrix and $N = n^2$. Suppose the tuning parameter is chosen such that $\mathbf{V} \geq \mathbf{V}^*$. Then the following risk bound is true for a universal constant $C > 0$:

$$\mathrm{MSE}(\hat{\theta}_{\mathbf{V}}, \theta^*) \leq C\big(\sigma\frac{\mathbf{V}}{\sqrt{N}}\,(\log \mathrm{e}N)^{5/2} + \frac{\sigma^2}{N}\big).$$

Remark II.1: The above result is similar to the ℓ_1 bound of [19], the difference being the above risk bound holds for the constrained TVD estimator while the existing result of [19] holds for the penalized estimator. For any sequence of $\mathbf{V} > 0$ (possibly growing with n although the canonical scaling is when $\mathbf{V} = O(1)$), the minimax lower bound results (mentioned earlier) of [34] now imply the minimax rate optimality (up to log factors) of the constrained TVD estimator $\hat{\theta}_{\mathbf{V}}$ over the parameter space $\{\theta \in \mathbb{R}^{n\times n} : \mathrm{TV}_{\mathrm{norm}}(\theta) \leq \mathbf{V}\}$.

Remark II.2: As is made clear in Section V, our technique for proving Theorem II.1 is completely different from the technique used to prove the result of [19]. While they analyze the properties of the pseudo-inverse of the edge incidence matrix D, our proof relies on computing relevant Gaussian widths by recursive partitioning. Moreover, ingredients and ideas from this proof are also used crucially in the proofs of our other results.

B. Adaptive Risk Bound

Now we come to the main result of this paper which is about proving adaptive risk bounds for θ^* which are piecewise constant on at most k axis aligned rectangles where k is a positive integer much smaller than N. We call a subset $R \subset L_n$ a (axis aligned) rectangle if it is a product of two intervals. For a generic rectangle $R = ([a,b]\cap\mathbb{N})\times([c,d]\cap\mathbb{N})$, we define $\mathrm{n}_{\mathrm{row}}(R)$ and $\mathrm{n}_{\mathrm{col}}(R)$ to be the cardinalities of $[c,d]\cap\mathbb{N}$ and $[a,b]\cap\mathbb{N}$ respectively. In words, $\mathrm{n}_{\mathrm{row}}(R)$ and $\mathrm{n}_{\mathrm{col}}(R)$ are simply the numbers of rows and columns of R respectively if one views R as a two-dimensional array of points. Then we define its aspect ratio to be $A(R) := \max\{\frac{\mathrm{n}_{\mathrm{row}}(R)}{\mathrm{n}_{\mathrm{col}}(R)}, \frac{\mathrm{n}_{\mathrm{col}}(R)}{\mathrm{n}_{\mathrm{row}}(R)}\}$. For a given matrix $\theta \in \mathbb{R}^{n\times n}$ we define $k(\theta)$ to be the cardinality of the minimal partition of L_n into rectangles $R_1,\dots,R_{k(\theta)}$ such that θ is constant on each of the rectangles. Next we state our main result for the 2D TVD estimator.

Theorem II.2: Let $\theta^* \in \mathbb{R}^{n\times n}$ be the underlying true matrix with $\mathrm{TV}_{\mathrm{norm}}(\theta^*) > 0$ and $R_1^*,\dots,R_{k(\theta^*)}^*$ be its rectangular level sets which form a partition of the 2D grid L_n. In addition, suppose the rectangles R_i^* have bounded aspect ratio, that is there exists a constant $c > 0$ such that $\max_{i\in[k(\theta^*)]} A(R_i^*) \leq c$. Then we have the following risk bound:

$$\mathrm{MSE}(\hat{\theta}_{\mathbf{V}}, \theta^*) \leq CA$$

where

$$A = \big[(\mathbf{V} - \mathbf{V}^*)^2 + \sigma^2 (\log en)^9 \frac{k(\theta^*)^{5/4}}{N^{3/4}}\big]$$

and C is a constant that only depends on c.

Remark II.3: Theorem II.2 is really a statement about an ideally tuned constrained TVD estimator. One way to interpret it is that if the tuning parameter $\mathbf{V}$ is chosen such that $(\mathbf{V} - \mathbf{V}^*)^2 \leq C\sigma^2 (\log en)^9 \frac{k(\theta^*)^{5/4}}{N^{3/4}}$ then the $\tilde{O}(\frac{k(\theta^*)^{5/4}}{N^{3/4}})$ rate of convergence holds.

Remark II.4: One consequence of the above theorem is that when $k(\theta^*) = O(1)$ then the ideally tuned TVD estimator attains a $\tilde{O}(N^{-3/4})$ rate. This rate is faster than the $\tilde{O}(N^{-1/2})$ rate that is available in the literature. Our main focus here has been to attain the right exponent for N. The exponent of $k(\theta^*)$ and $\log n$ may not be optimal. Since the current proof of this theorem is fairly involved technically, obtaining the best possible exponents of $k(\theta^*)$ and $\log n$ is left for future research endeavors. See Section III for more discussions about the proof of the above theorem and comparisons with existing results.

Remark II.5: We think a bounded aspect ratio condition would actually be necessary for the $O(N^{-3/4})$ rate to hold in the above theorem; see Section III-D for more on this issue.

A natural question is whether our upper bound in Theorem II.2 is tight. Our next theorem says that, in the low σ limit, the $N^{-3/4}$ rate is not improvable even if $k(\theta^*) = 2$.

Theorem II.3: Let $\theta^*_{ij} = 1$ if $j > n/2$ and 0 otherwise. Thus, θ^* is of the following form:

$$\theta^* = \left[\begin{array}{cc} \mathbf{0}_{n \times n/2} & \mathbf{1}_{n \times n/2} \end{array}\right]$$

Clearly $k(\theta^*) = 2$. In this case, we have a lower bound to the risk of the ideally constrained TVD estimator.

$$\lim_{\sigma \to 0} \frac{1}{\sigma^2} \mathrm{MSE}(\hat{\theta}_{\mathbf{V}^*}, \theta^*) \geq \frac{c}{N^{3/4}}\,.$$

Here $c > 0$ is a universal constant.

1) Gaussian Width Bounds: Proving Theorem II.2 and Theorem II.3 requires upper and lower bounds on the *Gaussian width* of a certain family of matrices as we now explain. The Gaussian width of a set $K \subset \mathbb{R}^n$ is defined as

$$\mathcal{GW}(K) = \mathbb{E} \sup_{\theta \in K} \langle Z, \theta \rangle \qquad \text{(II.1)}$$

where $Z = Z_n \sim N(0_n, I)$ and $\langle \cdot\,, \cdot \rangle$ is the usual Euclidean inner product between two vectors. We use $B_{m,n}(t)$ to denote the usual Euclidean ball of radius t in $\mathbb{R}^{m \times n}$. For any $A \subset \mathbb{R}^{n \times n}$ we denote the smallest cone containing A by $\mathrm{Cone}(A)$ and the closure of A by $\mathrm{Closure}(A)$. The *tangent cone* $T_{K^*}(\theta^*) \subset \mathbb{R}^{n \times n}$ *at* θ^* with respect to the closed convex set $K^* := \{\theta \in \mathbb{R}^{n \times n} : \mathrm{TV}_{\mathrm{norm}}(\theta) \leq \mathbf{V}^*\}$ is defined as follows:

$$\begin{aligned} &T_{K^*}(\theta^*) := \\ &\mathrm{Closure}(\,\mathrm{Cone}(\{\theta \in \mathbb{R}^{n \times n} : \theta^* + \theta \in K^*\})\,)\,. \qquad \text{(II.2)} \end{aligned}$$

By definition, $T_{K^*}(\theta^*)$ is a closed convex cone. Informally, $T_{K^*}(\theta^*)$ represents all directions in which one can move infinitesimally from θ^* while still remaining in K^*.

Roughly speaking, the problem of bounding the MSE from both directions is equivalent to bounding the square of $\mathcal{GW}\big(T_{K^*}(\theta^*) \cap B_{n \times n}(1)\big)$ when θ^* is a piecewise constant matrix on rectangles. The precise connection of MSE to Gaussian widths is detailed in Section VI where the proofs of Theorem II.2 and Theorem II.3 are also given. This connection prompts us to investigate how these tangent cones look like in the first place. *The major technical contribution of this paper is to give upper and lower bounds on the Gaussian width of the tangent cone at a piecewise constant matrix which we encapsulate in the following two results.*

Proposition II.4: Let $\theta \in \mathbb{R}^{n \times n}$ be a given matrix and $R_1, \dots, R_{k(\theta)}$ be its rectangular level sets which form a partition of the 2D grid L_n. In addition, let us assume that the rectangles R_i have bounded aspect ratio, that is there exists a constant $c > 0$ such that $\max_{i \in [k]} A(R_i) \leq c$. Let $K := \{v \in \mathbb{R}^{n \times n} : \mathrm{TV}_{\mathrm{norm}}(v) \leq \mathrm{TV}_{\mathrm{norm}}(\theta)\}$ and $T_K(\theta)$ be the tangent cone at θ with respect to K. Then there is a universal constant $C > 0$ such that

$$\mathcal{GW}(T_K(\theta) \cap B_{n,n}(1)) \leq C (\log n)^{4.5} k(\theta)^{5/8} n^{1/4}.$$

Proposition II.5: Consider θ^* which is piecewise constant on two rectangles and is of the following form:

$$\theta^* = \left[\begin{array}{cc} \mathbf{0}_{n \times n/2} & \mathbf{1}_{n \times n/2} \end{array}\right]$$

Then, there exists a universal constant $c > 0$ such that we have the following lower bound:

$$\mathcal{GW}(T_{K^*}(\theta^*) \cap B_{n \times n}(1)) \geq c n^{1/4}\,.$$

The proofs of Proposition II.4 and Proposition II.5 are given in Sections VIII and VII respectively.

It should be mentioned here that bounding the Gaussian width of the tangent cone is a fundamental task in a different but related problem of signal recovery from a given number of measurements; see [7] and [1]. Matrix recovery using 2D Total Variation has been studied in the signal processing literature; see for instance [4], [16] and [20]. Our bounds on the Gaussian widths given in Proposition II.4, Proposition II.5 and Theorem II.6 (see below) appear to be new and are potentially of independent interest as stand alone results. Especially our use of *optimized* partitioning schemes (see Section VIII-F for details) in the proof of Proposition II.4 can be a useful strategy to attack other problems of similar flavor. See also Section III-B for further discussion on the novelty of our proof.

2) Impossibility of Adaptation to Non Rectangular Level Sets: Theorem II.2 shows that the $O(N^{-3/4})$ rate is achievable when θ^* is piecewise constant on a few rectangles. A question arises here as to what rate is achievable when θ^* is piecewise constant but the level sets are not rectangular. The following theorem says that for a simple matrix θ^* whose level sets are triangular, the $\tilde{O}(N^{-1/2})$ rate cannot be improved. Below and in the rest of the paper we use $\mathbb{I}\{\cdot \in S\}$ to denote the indicator function for the set S (often stated as a condition defining its elements), i.e., it takes the value 1 when its argument lies in the set S and is 0 otherwise.

Theorem II.6: Consider the signal matrix $\theta^* := \mathbb{I}\{i + j > n\}$. Then, there exists a universal constant $c > 0$ such that we have the following lower bound:

$$\mathcal{GW}(T_{K^*}(\theta^*) \cap B_{n \times n}(1)) \geq c n^{1/2}\,.$$

Further, this implies a lower bound to the risk of the ideally constrained TVD estimator as follows:

$$\lim_{\sigma \to 0} \frac{1}{\sigma^2} \text{MSE}(\hat{\theta}_{\mathbf{V}^*}, \theta^*) \geq \frac{c}{N^{1/2}}\,.$$

Here $c > 0$ is a universal constant.

Remark II.6: The proof of the above theorem should be extendable when θ^* is the indicator of a circle or a regular n sided ($n > 4$) polygon or any other shape which is sufficiently non rectangular. See Remark VII.1 for more on this issue. Therefore, it seems that the rectangular shape of the level sets is crucial for the faster $\tilde{O}(N^{-3/4})$ rate to hold.

C. Tuning Free TVD

We now state our final result which relates to the question we posed about removing the tuning parameter and still retaining a risk bound which is essentially the same as in Theorem II.1. Choosing the tuning parameter is an important issue in applying the TVD methodology for denoising. The usual way out is to do some form of cross validation. There are some proposals available in the literature; see [35], [30], [22]. Soon after we uploaded our paper, a different tuning parameter free method appeared in [29] which also achieves the optimal worst case $\tilde{O}(\mathbf{V}/\sqrt{N})$ rate of convergence. See Section III-C for a comparison of our method with the one proposed in [29].

Our goal here is to construct a tuning parameter free estimator of θ^* which adapts to the true value of $\text{TV}_{\text{norm}}(\theta^*)$. The inspiration for this task comes from [8] where the author gives a general recipe to construct tuning parameter free estimators in Gaussian mean estimation problems when the truth is known to have small value of some known norm. Even though the total variation functional is not a norm but a seminorm, the general idea in [8] can be extended as we will show. However, the estimator of [8] is a randomized estimator whereas in our case we construct a non randomized version. The following is a description of our tuning free estimator.

Let $\mathbf{1}$ denote the $n \times n$ matrix consisting solely of ones. For any matrix $\theta \in \mathbb{R}^{n \times n}$, we define $\overline{\theta} := \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \theta[i,j]$ to be the mean of θ. Define the estimator

$$\hat{\theta}_{\text{notuning}} := \overline{y}\,\mathbf{1} + \mathop{\text{argmin}}_{\{v \in \mathbb{R}^{n \times n} : \overline{v} = 0,\, \|y - \overline{y}\,\mathbf{1} - v\|^2 \leq (n^2 - 1)\hat{\sigma}^2\}} \text{TV}_{\text{norm}}(v) \tag{II.3}$$

where $\hat{\sigma}$ is an estimator of σ defined as follows:

$$\hat{\sigma} := \frac{\text{TV}_{\text{norm}}(y)}{\mathbb{E}\,\text{TV}_{\text{norm}}(Z)} = \frac{\sqrt{\pi}\,\text{TV}_{\text{norm}}(y)}{4\,(n-1)}\,. \tag{II.4}$$

The intuition behind the estimator defined above is as follows. The estimation of θ^* is done by estimating the two orthogonal parts $\overline{\theta^*}\,\mathbf{1}$ and $\theta^* - \overline{\theta^*}\mathbf{1}$ separately. The first part is estimated by $\overline{y}\,\mathbf{1}$. To estimate $\theta^* - \overline{\theta^*}\mathbf{1}$, we use a Dantzig Selector type (see [5]) version of the TVD estimator, which computes a zero mean matrix with the least total variation subject to being within a Euclidean ball of a suitable radius around the centered data matrix $y - \overline{y}\,\mathbf{1}$. A good choice of this radius actually depends on the true σ and hence as an intermediate step, we have to estimate σ in the process which is denoted by $\hat{\sigma}$. The main idea behind our construction of $\hat{\sigma}$ here is the fact that $\text{TV}_{\text{norm}}(\theta^*)$ is small compared to $\text{TV}_{\text{norm}}(Z)$ and hence $\text{TV}_{\text{norm}}(y) = \text{TV}_{\text{norm}}(\theta^* + \sigma Z)$ approximately equals $\sigma \text{TV}_{\text{norm}}(Z)$. We can then use concentration properties of the $\text{TV}_{\text{norm}}(Z)$ statistic to show that $\frac{\text{TV}_{\text{norm}}(Z)}{\mathbb{E}\text{TV}_{\text{norm}}(Z)}$ is approximately equal to 1. The following theorem supplies a risk bound for $\hat{\theta}_{\text{notuning}}$.

Theorem II.7: We have the following risk bound for our tuning free estimator:

$$\text{MSE}(\hat{\theta}_{\text{notuning}}, \theta^*) \leq CA$$

where

$$A = \big(\sigma \frac{\mathbf{V}^*}{\sqrt{N}} \log(\mathrm{e}n)^{5/2} + \frac{(\mathbf{V}^*)^2}{N} + \frac{\sigma^2}{\sqrt{N}}\big)$$

and C is a universal constant.

Remark II.7: Note that the above bound is meaningful only when $\lim_{N \to \infty} \frac{\mathbf{V}^*}{\sqrt{N}} = 0$. Therefore in this regime, $\frac{(\mathbf{V}^*)^2}{N}$ is a lower order term. Thus, Theorem II.7 basically says that the MSE of $\hat{\theta}_{\text{notuning}}$, up to multiplicative log factors and an additive factor $\frac{\sigma}{\sqrt{N}}$, scales like $\frac{\mathbf{V}^*}{\sqrt{N}}$. In light of Remark II.1 we can say that $\hat{\theta}_{\text{notuning}}$ is minimax rate optimal (up to log factors) over $\{\theta \in \mathbb{R}^{n \times n} : \text{TV}_{\text{norm}}(\theta) \leq \mathbf{V}\}$, simultaneously for any sequence of $\mathbf{V}$ (depending on n) which is bounded below by a constant and above by $\sqrt{N}$. To the best of our knowledge, this is the first result demonstrating such an estimator which is completely tuning free.

III. Comparison With Existing Results, Simulation Studies and Discussions

To place our theorems in context, it is worthwhile to compare and relate our results with a couple of recent papers. We also discuss some issues related to our results.

A. Comparison With [19]

Let us compare our risk bound in Theorem II.2 to the adaptive risk bound (Theorem I.1) of [19] when the truth θ^* is piecewise constant on a few axis aligned rectangles. Both of these theorems prove statements about tuned TVD estimators. Considering the very simple case when θ^* is of the following form:

$$\theta^* = \begin{bmatrix} \mathbf{0}_{n \times n/2} & \mathbf{1}_{n \times n/2} \end{bmatrix}$$

we have already mentioned in Section I that $\|D\theta^*\|_0 = O(\sqrt{N})$. Thus, Theorem I.1 gives us an upper bound on the MSE scaling like $\tilde{O}(1/\sqrt{N})$ whereas our Theorem II.2 gives a faster rate of convergence scaling like $\tilde{O}(1/N^{3/4})$. More generally, if θ^* is piecewise constant on k axis aligned rectangles with bounded aspect ratio and roughly equal size, it can be checked that $\|D\theta^*\|_0 \approx \sqrt{k^* N}$. This means that Theorem I.1 gives us an upper bound on the MSE scaling like $\tilde{O}(\sqrt{k^*/N})$. Compare this to Theorem II.2 which gives a rate of convergence scaling like $\tilde{O}((k^*)^{5/4}/N^{3/4})$. Thus, in the small k^* regime when $k^* < N^{1/3}$, Theorem II.2 provides a faster rate of convergence. This is the main contribution of this paper and to the best of our knowledge is the first of its kind in the literature.

B. Comparison With [17]

As mentioned in Section I, one of our motivating factors behind investigating adaptivity of the 2D TVD estimator was its success in optimally estimating piecewise constant vectors in the 1D setting. Theorem 2.2 in [17] gives a $\tilde{O}(k^*/n)$ rate for the ideally tuned constrained 1D TVD estimator when the truth θ^* is piecewise constant with k^* pieces or blocks and each block satisfies a certain minimum length condition. In a sense, our Theorem II.2 is a natural successor, giving the corresponding result in the 2D setting. Our bounded aspect ratio condition is the 2D version of the minimum length condition. A consequence of Theorem II.2 and Theorem II.3 is that, in contrast to the 1D setting, the ideally tuned constrained TVD estimator can no longer obtain the oracle rate of convergence $\tilde{O}(k^*/n)$ in the 2D setting.

The proof of Theorem 2.2 in [17] was done by bounding the Gaussian widths of certain tangent cones. Our proof of Theorem II.2 also adopts the same strategy and precisely characterizes the tangent cone $T_{K(V^*)}(\theta^*)$ (defined in (II.2)) for piecewise constant θ^* and then bounds its Gaussian width. The main idea in [17] was to observe that any unit norm element of the tangent cone is nearly made up of two monotonic blocks in each constant block of θ^*. Then the available metric entropy bounds for monotone vectors were used to bound the Gaussian width. A crucial ingredient in this proof is the well-known fact that any univariate function of bounded variation has a canonical representation as a difference of two monotonic functions. However, it is not clear at all how to adapt such a strategy to the 2D setting. In particular, it is not nearly as natural and convenient to express a matrix of bounded variation as a difference of two bi-monotone matrices. Our computation of Gaussian width of the tangent cone is therefore essentially *two dimensional* and involves judicious recursive partitioning in both dimensions. We believe that our Gaussian width computations, especially the proof of Proposition VIII.9, consist of new techniques and are potentially useful for problems of similar flavor.

C. Comparison With [29]

At the latter stages of preparation of this manuscript we became aware of an independent work by [29] which is related to our manuscript. In [29], the authors give a general technique to derive slow (ℓ_1) and fast (ℓ_0) rates for penalized TVD estimators and its square root version on general graphs. Thus, there seems to be two routes for obtaining fast rates for TVD. One goes through the route of bounding Gaussian width of an appropriate tangent cone to derive fast rates for the constrained TVD estimator; as done here in this manuscript as well as in [17]. The other route; followed by [19] and generalized by [29] is based on bounding the so called *compatibility factor*. [29] show how to bound this compatibility factor for specific graphs such as the $1d$ grid graph and the $1d$ cycle graph. To the best of our knowledge, bounding the compatibility factor for piecewise constant functions on axis aligned rectangles for a $2d$ grid remains an open problem. Thus, as far as we are aware, the work in this manuscript proving fast rates of convergence on $2d$ grid graph for piecewise constant functions on axis aligned rectangles is the first of its type in the literature.

The work in [29] also proposes a general technique to obtain slow rates for a square root version of the TVD estimator. Similar to our paper, [29] also considers the case when the noise variables are i.i.d. Gaussian. The advantage of using this square root version is that the tuning parameter does not need to depend on the unknown parameter σ. While the theoretically recommended choice of the tuning parameter λ in [29, Corollary 4.13] does not depend on the noise variance σ^2, there is however the presence of an unspecified large universal constant C. It is not clear to us whether this C can be explicitly specified. On the other hand, our tuning free estimator is explicitly specified and involves no unknown constants. We think the analysis of our tuning free estimator is also reasonably clean with the sources of the various possible errors made transparent in the proof. This is why we believe that our tuning free estimator provides a theoretically valid and possibly useful alternative to the square root regularization approach. Just to be clear, we are not claiming any optimality of our tuning free method, our intention is to demonstrate one theoretically valid way to obtain a minimax rate optimal tuning free estimator.

D. Necessity of Bounded Aspect Ratio Condition in Theorem II.2

We think a bounded aspect ratio condition would actually be necessary for the $O(N^{-3/4})$ rate to hold in Theorem II.2. For instance, consider the sequence of matrices θ^* such that $\theta^*[i,j] = \mathbb{I}\{j = n\}$. Clearly, the rectangular level sets of the sequence of matrices θ^* do not satisfy the bounded aspect ratio condition. By an argument similar to the one used to prove our lower bound results in Theorem II.3 and Theorem II.6, one can show that the $\mathrm{MSE}(\hat{\theta}_{\mathbf{V}^*}, \theta^*) \geq cN^{-1/2}$. We have also verified this scaling of the MSE in our numerical experiments.

The bounded aspect ratio condition says that the rectangular level sets of θ^* should not be too skinny or too long. In our proof, the bounded aspect ratio is needed for similar reasons as a minimum length condition is needed for the length of the pieces in the 1D setting; see [17], [12].

E. On Obtaining the Oracle Rate for Piecewise Constant Signals

In light of Theorem II.2 and Theorem II.3 we can say the following statement. When the truth θ^* is piecewise constant on k^* axis aligned rectangles, the TVD estimator *cannot* attain the oracle rate of convergence scaling like $O(k^*/N)$. The question that now arises is whether there exists *any* estimator which attains the $\tilde{O}(k^*/N)$ rate of convergence for all piecewise constant truths *as well as* the minimax rate $\tilde{O}(\mathbf{V}^*/\sqrt{N})$? Furthermore, can this estimator be chosen so that it is computationally efficient? These questions led us to examine decision tree estimators which are different from the TVD type estimators. We would like to point out here that in the paper [9] we have been able to demonstrate computationally efficient estimators which attain both the aforementioned goals.

Apart from [9], some other recent papers have also sprung up which target piecewise rectangular signals. The papers [28] and [15] study a different version of the TVD estimator which is also termed as the *Hardy Krause estimator*. As far as we understand, this estimator is well suited for estimating piecewise rectangular signals as it actually fits piecewise rectangular estimates. For a general signal with k^* rectangular pieces (with some regularity conditions on the rectangular pieces), the rate proved by [28] is $\tilde{O}((k^*)^{3/2}/N)$ which is better than $\tilde{O}\big((k^*)^{5/4}/N^{3/4}\big)$. Notice that the $\tilde{O}((k^*)^{3/2}/N)$ rate still does not match the near oracle rate $\tilde{O}(k^*/N)$ which has been obtained in [9].

F. Constrained Vs Penalized

In this paper, we have focussed on the constrained version of the 2D TVD estimator. As mentioned in the introduction, the penalized version is also quite popular. In the low σ limit, it can be proved that the constrained estimator $\hat{\theta}_{\mathbf{V}}$ with $\mathbf{V} = \mathbf{V}^* = \mathrm{TV}_{\mathrm{norm}}(\theta^*)$ is better than the penalized estimator for every deterministic choice of λ. More precisely, we have for all $\lambda \geq 0$,

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2}\, \mathrm{MSE}(\hat{\theta}_{\mathbf{V}^*}, \theta^*) < \lim_{\sigma \downarrow 0} \frac{1}{\sigma^2}\, \mathrm{MSE}(\hat{\theta}_{\lambda}, \theta^*).$$

The above inequality follows from the results of [31] as described in Section 5.2 in [17]. Since our main question here is whether faster/adaptive rates are possible for piecewise constant matrices, it is therefore natural to first study the constrained version with ideal tuning. A possible next step is to investigate whether a similar $N^{-3/4}$ rate is atttained by the penalized TVD estimator and if so, for what range of the tuning parameter λ.

G. Simulation Studies

We consider three distinct sequences of matrices to facilitate comparison. We consider the simplest piecewise constant matrix $\theta^{\mathrm{two}} \in \mathbb{R}^{n \times n}$ where $\theta^{\mathrm{two}} := \mathbb{I}\{j > n/2\}$. Hence θ^{two} just takes two distinct values. The next matrix θ^{four} is a block matrix with four constant blocks.

$$\theta^{\mathrm{four}} := \begin{bmatrix} 1_{n/2 \times n/2} & 2_{n/2 \times n/2} \\ 0_{n/2 \times n/2} & 1_{n/2 \times n/2} \end{bmatrix}$$

Finally, we also consider a $n \times n$ matrix $\theta^{\mathrm{worst}} := \mathbb{I}\{i + j > n\}$. Clearly, θ^{worst} does not have a block constant structure. For the matrix θ^{worst} we incur the worst case rate $\tilde{O}(N^{-1/2})$ as shown in Theorem II.6; hence the name. The noise variance has been set to 1 for all the numerical experiments reported in this section.

The dependence of the MSE with $N = n^2$ can be experimentally checked as follows. We can estimate the MSE for a fixed n by Monte Carlo repetitions and then iterate this for a grid of n values. We then plot log of the estimated MSE with log N and fit a least squares line to the plot. The slope of the least squares line then gives an indication of the correct exponent of N in the MSE. Figure 1 is such a plot for the ideally tuned constrained TVD estimator.

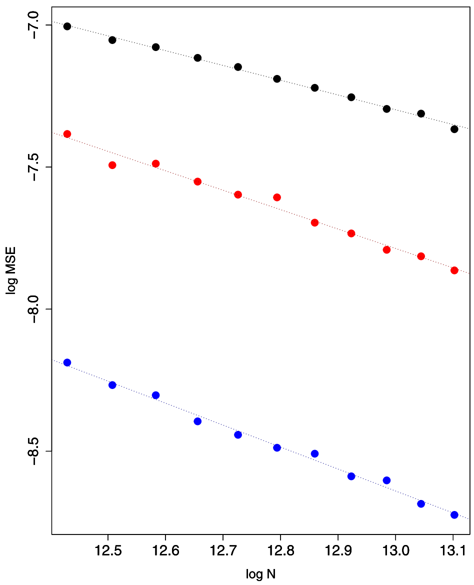

Fig. 1. The MSE of the ideally tuned TVD estimator $\hat{\theta}_{\mathbf{V}^*}$ is estimated with 50 Monte carlo repetitions for a grid of $n = \sqrt{N}$ ranging from 500 to 700 in increments of 20. The true matrices were taken to be θ^{two} (blue), θ^{four} (red) and θ^{worst} (black). In each case, we have chosen the ideal tuning parameter to allow fair comparison. We plot log of estimated MSE versus log N where log is taken in base e. The points are the estimated log MSE and the dotted lines are the least squares line fitted to the points. The least squares slope for θ^{two} is -0.73 and for θ^{four} is -0.68 which is considerably lower than the slope for the matrix θ^{worst} which is -0.52..

In Figure 1, the risk is seen to be minimum for θ^{two} followed by θ^{four} and then θ^{worst}. The slope for θ^{two} and θ^{four} came out to be -0.73 and -0.68. This agrees well with Theorem II.2 and Theorem II.3 which says that the MSE decays at the rate $n^{-0.75}$ upto log factors. For the matrix θ^{worst} the slope turned out be -0.52 which is in agreement with the worst case $\tilde{O}(N^{-1/2})$ rate given in Theorem II.6.

To investigate the dependence of MSE with the number of rectangular level sets $k(\theta^*)$, we took four matrices. The first two are $\theta^{\mathrm{two}}, \theta^{\mathrm{four}}$ and the last two are obtained by further binary division so that the number of rectangular level sets is $8, 16$ respectively. We normalized the matrices such that $\mathbf{V}^* = 1$. We fixed $n = 800$ and did 50 iterations of Monte Carlo simulations for each of the four matrices. We then plotted $\log \mathrm{MSE}$ versus $\log_2\, k$ (see Figure 2) where $k = 2, 4, 8, 16$. The slope of the least squares line we got is 0.81. This suggests that our exponent of k $(= 1.25)$ in the risk bound in Theorem II.2 may not be optimal.

To assess the risk of our fully data driven estimator $\hat{\theta}_{\mathrm{notuning}}$, we again consider the three matrices $\theta^{\mathrm{two}}, \theta^{\mathrm{four}}$ and θ^{worst} respectively. Figure 3 is a plot of log MSE versus $\log n$.

The simulations in Figure 3 suggest that our estimator has MSE decaying at a $O(1/\sqrt{N})$ rate for all three matrices. The slope of all three least squares lines are reasonably close to -0.5. This matches the rate given in Theorem II.7. However, our tuning free estimator does not seem to be

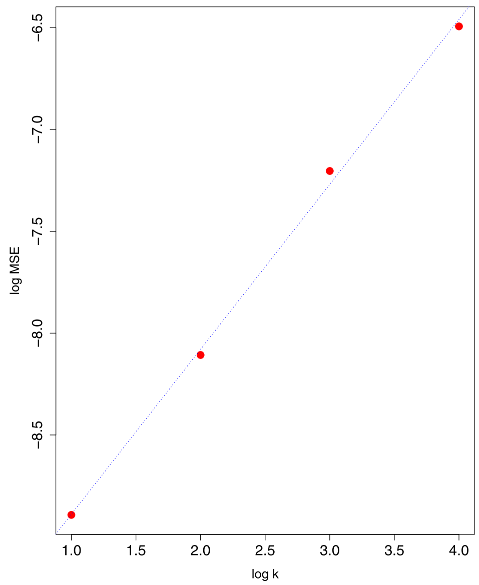

Fig. 2. The MSE of the ideally tuned TVD estimator is estimated with 50 Monte carlo repetitions when $n = 800$. The true matrices were taken to be such that the number of rectangular level sets is $2, 4, 8, 16$. In each case, we have chosen the ideal tuning parameter to allow fair comparison. We have also normalized the matrices so that $\mathbf{V}^* = 1$. We plot log of estimated MSE versus $\log_2\ k$ where $k = 2, 4, 8, 16$. The points are the estimated log MSE and the dotted lines are the least squares line fitted to the points. The least squares slope is 0.81..

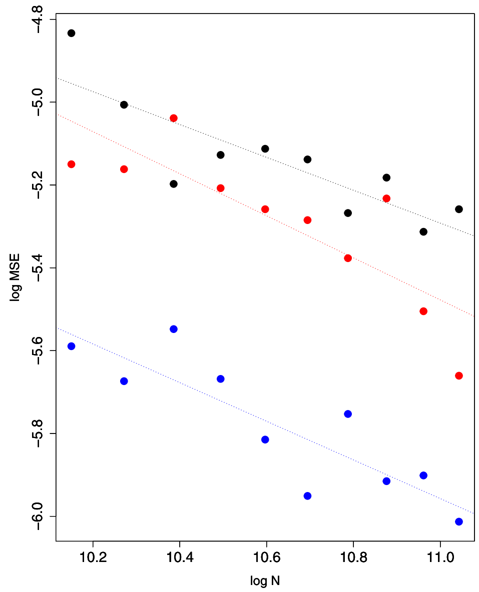

Fig. 3. The MSE of our tuning free estimator is estimated with 50 Monte carlo repetitions for a grid of $n = \sqrt{N}$ ranging from 160 to 250 in increments of 10. The true matrices were taken to be θ^{two} (blue), θ^{four} (red), and θ^{worst} (black). We plot log of estimated MSE versus $\log\ N$ where log is taken in base e. The circular points are the estimated log MSE and the dotted lines are the least squares line fitted to the points. The slopes of the least squares lines are -0.47,-0.51,-0.40 for θ^{two}, θ^{four}, θ^{worst} respectively.

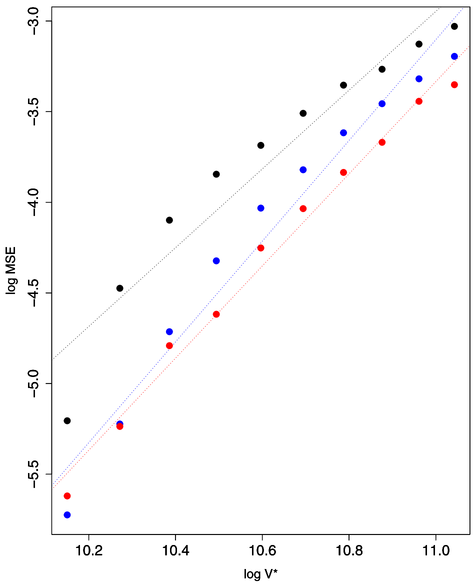

Fig. 4. The MSE of the tuning free TVD estimator is estimated with 50 Monte carlo repetitions for a grid of $\mathbf{V}^* \in [10]$ and $n = 200$. The true matrices were taken to be θ^{two} (blue), θ^{four} (red) and θ^{worst} (black) properly normalized. We plot log of estimated MSE versus $\log\ N$ where log is taken in base e. The points are the estimated log MSE and the dotted lines are the least squares line fitted to the points. The least squares slope for θ^{two} is 1.16, for θ^{four} is 1.07 and for the matrix θ^{worst} it is 0.94..

adaptive to piecewise constant structure like the constrained TVD estimator with ideal tuning.

To investigate the dependence of the risk of our tuning free estimator on $\mathbf{V}^*$, for each of the three matrices θ^{two}, θ^{four}, θ^{worst}, we normalized the matrix such that $\mathbf{V}^* = 1, 2, \ldots, 10$. We fixed $n = 200$ and did 50 iterations for each $\mathbf{V}^*$ and each matrix. We then plotted log MSE versus log $\mathbf{V}^*$ (see Figure 4) and fitted a least squares line. The slopes for each of these three matrices came out to be $1.16, 1.07, 0.94$ respectively. This suggests that the right exponent of V^* is 1 and our risk bound has the right dependence on $\mathbf{V}^*$.

IV. A Generic Approach Towards Bounding Gaussian Widths

Let us recall from Section II-B1 that the Gaussian width of a set $K \subset \mathbb{R}^n$ is defined as

$$\mathcal{GW}(K) = \mathbb{E} \sup_{\theta \in K} \langle Z, \theta \rangle$$

where $Z = Z_n \sim N(0_n, I)$ and $\langle \cdot\,, \cdot \rangle$ is the usual Euclidean inner product between two vectors. Our principal result in this section provides an upper bound on $\mathcal{GW}(K)$ in terms of the numbers and dimensions of its covering (linear) subspaces. This result executes and adapts the idea of chaining (see, e.g., [40, Theorem 5.24]) to the case when the covering sets are linear subspaces of $\mathbb{R}^n$. To this end let us define, for any $\epsilon > 0$, an ϵ *subspace cover* of K to be any finite collection $\mathcal{S}$ of linear subspaces of $\mathbb{R}^n$ such that

$$\sup_{\theta \in K} \operatorname{dist}\Big(\theta, \bigcup_{S \in \mathcal{S}} S\Big) \leq \epsilon$$

where $\operatorname{dist}(A, B)$ denotes the Euclidean distance between the sets A and B. We denote by $\operatorname{diam}(K)$ the diameter of K which we assume to be finite. Also for any $t > 0$, we denote by $B_n(t)$ the t-Euclidean ball $\{\theta \in \mathbb{R}^n : \|\theta\| \leq t\}$ where $\|.\|$ is the Euclidean norm. We will often drop the subscript n and just write $B(t)$ when the dimension is clear from the context.

Proposition IV.1 (Gaussian Width Bound): For every $\epsilon \in (0, \mathrm{diam}(K))$, let $\mathcal{S}_\epsilon$ be an ϵ subspace cover of K. Also let $k_1 > k_0$ be integers with k_0 being the smallest integer satisfying $2^{-k_0} \geq \mathrm{diam}(K)$. Then we have

$$\mathcal{GW}(K) \leq \sqrt{n}\, 2^{-k_1} \; + \\ 3 \sum_{k=k_0+1}^{k_1} 2^{-k} \big[\max_{S \in \mathcal{S}_{2^{-k}}} \sqrt{2\, \dim(S)} + 2\sqrt{\log |\mathcal{S}_{2^{-k}}|} + 1 \big].$$

Proof: For any $\theta \in K$ and integer k such that $2^{-k} < \mathrm{diam}(K)$, let θ_k denote a point in $\mathcal{N}_k := \cup_{S \in \mathcal{S}_{2^{-k}}} S$ such that

$$\mathrm{dist}(\theta, \theta_k) = \mathrm{dist}(\theta, \mathcal{N}_k).$$

Such a point always exists since $\mathcal{N}_k$ is a finite union of linear subspaces. When $\mathrm{diam}(K) \geq 2^{-k}$, on the other hand, we simply choose θ_k to be some fixed but arbitrary point θ_K in K. By definition, we thus have

$$\|\theta - \theta_k\| \leq 2^{-k} \tag{IV.1}$$

for all $k \in \mathbb{Z}$. Let us now write for every $\theta \in K$,

$$\theta = \theta_K + \sum_{k=k_0+1}^{k_1} (\theta_k - \theta_{k-1}) + (\theta - \theta_{k_1})$$

so that

$$\mathcal{GW}(K) = \mathbb{E} \sup_{\theta \in K} \langle Z, \theta \rangle \leq \mathbb{E} \langle Z, \theta_K \rangle + \\ \sum_{k=k_0+1}^{k_1} \mathbb{E} \sup_{\theta \in K} \langle Z, \theta_k - \theta_{k-1} \rangle + \mathbb{E} \sup_{\theta \in K} \langle Z, \theta - \theta_{k_1} \rangle.$$

The first term on the right hand side above is 0, whereas the third time is bounded by $\sqrt{n} 2^{-k_1}$ in view of the Cauchy-Schwarz inequality, display (IV.1) and the standard bound $\mathbb{E}\|Z\| \leq \sqrt{n}$. Therefore we can conclude the proof if we can show

$$\mathbb{E} \sup_{\theta \in K} \langle Z, \theta_k - \theta_{k-1} \rangle \leq \\ 3 \cdot 2^{-k} \big[\max_{S \in \mathcal{S}_{2^{-k}}} \sqrt{2\, \dim(S)} + 2\sqrt{\log |\mathcal{S}_{2^{-k}}|} \,\big]$$

for every integer k satisfying $2^{-k} < \mathrm{diam}(K)$. To this end observe that

$$\|\theta_k - \theta_{k-1}\| \leq \|\theta_k - \theta\| + \|\theta_{k-1} - \theta\| \leq 3 \cdot 2^{-k}$$

in view of (IV.1) and $\theta_k - \theta_{k-1} \in \mathcal{M}_k$ where $\mathcal{M}_k := \{S_1 + S_2 : S_1 \in \mathcal{N}_{2^{-k}}, S_2 \in \mathcal{N}_{2^{-(k-1)}}\}$ is another finite collection of linear subspaces of $\mathbb{R}^n$. It is also clear from the definition that $|\mathcal{M}_k| \leq |\mathcal{N}_{2^{-k}}||\mathcal{N}_{2^{-(k-1)}}|$. All these observations bring us to the setting of:

Lemma IV.2 (Gaussian Width for Union of Subspaces): Let $\mathcal{S}$ be a finite collection of linear subspaces of $\mathbb{R}^n$ and $\Theta = \cup_{S \in \mathcal{S}} S \subset \mathbb{R}^n$. In words, Θ is the union of subspaces in $\mathcal{S}$. Then we have

$$\mathcal{GW}(\Theta \cap B(t)) \leq t \,\big[\max_{S \in \mathcal{S}} \sqrt{\dim(S)} + \sqrt{2\, \log |\mathcal{S}|} + 1 \big].$$

Using Lemma IV.2, we can immediately deduce that

$$\mathbb{E} \sup_{\theta \in K} \langle Z, \theta_k - \theta_{k-1} \rangle \leq 3 \cdot 2^{-k} M$$

where

$$M := \big[\max_{S_1 \in \mathcal{S}_{2^{-k}}, S_2 \in \mathcal{S}_{2^{-(k-1)}}} \sqrt{\dim(S_1) + \dim(S_2)} + \\ \sqrt{2\, \log |\mathcal{S}_{2^{-k}}| + 2\, \log |\mathcal{S}_{2^{-(k-1)}}|} + 1 \big].$$

Now we can assume without any loss of generality that $|\mathcal{S}_{2^{-(k-1)}}| \leq |\mathcal{S}_{2^{-k}}|$ as well as

$$\max_{S \in \mathcal{S}_{2^{-(k-1)}}} \dim(S) \leq \max_{S \in \mathcal{S}_{2^{-k}}} \dim(S),$$

which finishes the proof of the proposition.

Let us now return to the proof of Lemma IV.2. Since $\Theta = t\,\Theta$, it follows from the definition of Gaussian widths that $\mathcal{GW}(\Theta \cap B(t)) = t\,\mathcal{GW}(\Theta \cap B(1))$ meaning we only need to work with $t = 1$. We will use the following lemma involving only one linear subspace:

Lemma IV.3: For any linear subspace S of $\mathbb{R}^n$ and $u \geq 0$, we have with probability at least $1 - \exp(-\frac{u^2}{2})$,

$$\sup_{\theta \in S \cap B(1)} \langle Z, \theta \rangle \leq \sqrt{\dim(S)} + u. \tag{IV.2}$$

Proof: We will use the well-known concentration inequality for Lipschitz functions of a Gaussian vector (see, e.g. [23, Theorem 7.1]). First of all notice that the random variable $f(Z) := \sup_{\theta \in S \cap B(1)} \langle Z, \theta \rangle$ is a Lipschitz function of Z with Lipschitz constant 1. It follows from the observation that, for any $z, z' \in \mathbb{R}^n$ and $\theta \in B(1)$,

$$\langle z, \theta \rangle - \langle z', \theta \rangle = \langle z - z', \theta \rangle \leq \|z - z'\| \|\theta\| \leq \|z - z'\|$$

where in the last but one step we used the Cauchy-Schwarz inequality. Therefore by the Gaussian concentration inequality mentioned in the beginning, we have for any $u \geq 0$

$$\mathbb{P}(f(Z) - \mathbb{E}f(Z) \geq u) \leq \exp(-\frac{u^2}{2}).$$

Hence we can deduce the lemma upon showing that $\mathbb{E}f(Z) \leq \sqrt{\dim(S)}$. To this end notice that $f(Z) = \|P_S Z\|$ where P_S is the orthogonal projector onto the subspace S. Therefore, $f(Z)^2$ is a chi squared random variable whose degree of freedom equals $\dim(S)$ whence we get

$$\mathbb{E}f(Z) \leq \sqrt{\mathbb{E}f(Z)^2} \leq \sqrt{\dim(S)}. \qquad \square$$

Now, using a union bound followed by Lemma IV.3 we get

$$\mathbb{P}\big(\sup_{\theta \in \Theta \cap B(1)} \langle Z, \theta \rangle \geq \max_{S \in \mathcal{S}} \sqrt{\dim(S)} + u \big) \\ \leq \sum_{S \in \mathcal{S}} \mathbb{P}\big(\sup_{\theta \in S \cap B(1)} \langle Z, \theta \rangle \geq \max_{S \in \mathcal{S}} \sqrt{\dim(S)} + u \big) \\ \leq |\mathcal{S}| \exp(-\frac{u^2}{2}).$$

Plugging in $u = \sqrt{2\, \log |\mathcal{S}| + v^2}$ we obtain

$$\mathbb{P}\big(\sup_{\theta \in \Theta \cap B(1)} \langle Z, \theta \rangle \geq \max_{S \in \mathcal{S}} \sqrt{\dim(S)} + \sqrt{2\, \log |\mathcal{S}|} + \\ v \big) \leq \exp(-\frac{v^2}{2}).$$

Integrating the above tail bound finishes the proof of Lemma IV.2. □

Remark IV.1: A general and perhaps more standard way of bounding the Gaussian width of a set is through Dudley's entropy integral inequality (see [14]). In this approach one first finds a "good" covering set corresponding to any given radius r for the underlying set to obtain upper bounds on covering numbers which then enter an integral (after being transformed appropriately) bounding the Gaussian width. Proposition IV.1 provides an alternative way when the covering sets are contained in finite unions of linear subspaces with comparable dimensions. For the purpose of the current article, this approach would save us some extraneous log factors in our bounds.

V. Proof of Theorem II.1

We first set up some notations which would henceforth be used throughout the paper. For a positive integer n, we will denote the subset of positive integers $\{1,\dots,n\}$ by $[n]$. Recall that in all the proofs of our results, we are going to use TV to denote the *unnormalized* version of (I.1) as defined in (I.2). Also we will use V for the unnormalized total variation instead of the *bold* $\mathbf{V}$ used for the *corresponding* normalized version.

Let us recall that the estimator $\hat{\theta}_V$ is the least squares estimator on the set

$$K_n(V) := \{\theta \in \mathbb{R}^{n\times n} : \mathrm{TV}(\theta) \le V\}. \tag{V.1}$$

We will often drop the subscript n and just write $K(V)$ when the dimension is clear from the context. Below we adopt the standard approach of using the basic inequality defining least squares estimators to reduce our problem to controlling Gaussian widths.

Lemma V.1: Under the same conditions as in the statement of Theorem II.1 we have

$$\mathbb{E}\|\hat{\theta}_V - \theta^*\|^2 \le 2\,\sigma\,\mathbb{E}\sup_{\theta:\mathrm{TV}(\theta)\le 2V,\,\overline{\theta}=0}\langle Z,\theta\rangle + 2\,\sigma^2.$$

Proof: Since $V \ge V^*$ we have the basic inequality $\|y - \hat{\theta}_V\|^2 \le \|y-\theta^*\|^2$. Substituting $y = \theta^* + \sigma Z$ gives us

$$\begin{aligned}\|\theta^* - \hat{\theta}_V\|^2 &\le 2\langle \hat{\theta}_V - \theta^*, y - \theta^*\rangle = 2\langle \hat{\theta}_V - \theta^*, \sigma\, Z\rangle\\ &= 2\,\sigma\,\langle \hat{\theta}_V - \overline{y}\mathbf{1} - (\theta^* - \overline{\theta^*}\mathbf{1}), Z\rangle + 2\,\sigma\,\langle \overline{y}\mathbf{1} - \overline{\theta^*}\mathbf{1}, Z\rangle\\ &\le 2\,\sigma \sup_{v:\mathrm{TV}(v)\le 2V,\,\overline{v}=0}\langle Z, v\rangle + 2\,\sigma\,\langle \overline{y}\mathbf{1} - \overline{\theta^*}\mathbf{1}, Z\rangle.\end{aligned}$$

where the last inequality follows because $\overline{\hat{\theta}_V} = \overline{y}$ and $\mathbf{1}$ refers to the $n\times n$ matrix whose all elements equal 1. Now taking expectation on both sides of the above display and noting that

$$\mathbb{E}\langle \overline{y}\mathbf{1} - \overline{\theta^*}\mathbf{1}, Z\rangle = \sigma\, n^2 \mathbb{E}\overline{Z}^2 = \sigma$$

finishes the proof. □

Let us define

$$K^0(V) = K^0_n(V) := \\ \{\theta \in \mathbb{R}^{n\times n} : \mathrm{TV}(\theta) \le V,\, \overline{\theta} = 0\}\,.$$

In view of Lemma V.1, all we need is to evaluate the Gaussian width of the set $K^0(2V)$ to which end we will use Proposition IV.1. But for that we need to find "efficient" subspace covers of the set $K^0(V)$ corresponding to any distance ϵ. Our next proposition will be crucial for this purpose. Below we denote, for any rectangular partition P of L_n, the linear subspace of $\mathbb{R}^{n\times n}$ comprising only matrices that are constant on each (rectangular) block of P by S_P.

Proposition V.2: For every $\eta, V > 0$, there exist a set of rectangular partitions $\mathcal{P}(V,n,\eta)$ of L_n (recall the definition from Section II-B) and a universal constant $C > 0$ such that

- For any $\theta \in K(V)$ (recall the definition from (V.1)), there exists a partition $P \in \mathcal{P}(V,n,\eta)$ satisfying
$$(\mathrm{dist}(\theta, S_p))^2 \le V\eta\log n + \eta^2.$$
- Any partition $P \in \mathcal{P}(V,n,\eta)$ has number of (rectangular) blocks bounded by
$$|P| \le 1 + C\frac{V}{\eta}\log n.$$
- The cardinality of $\mathcal{P}(V,n,\eta)$ is bounded as
$$\log|\mathcal{P}(V,n,\eta)| \le C\frac{V}{\eta}(\log n)^2.$$

Before we prove this proposition, let us finish the proof of Theorem 2.1 assuming it.

Proof of Theorem II.1: Throughout this proof, we will use C to denote an unspecified but universal positive constant whose exact value may change from one line to the next. For any $\epsilon, V > 0$, let $\eta_\epsilon := \min(\frac{\epsilon^2}{2V\log n}, \frac{\epsilon}{\sqrt{2}})$ and consider the set of rectangular partitions $\mathcal{P}(V,n,\eta_\epsilon)$ given by Proposition V.2. Next define a collection $\mathcal{S}_\epsilon$ of linear subspaces of $\mathbb{R}^{n\times n}$ as follows:

$$\mathcal{S}_\epsilon := \{S_P : P \in \mathcal{P}(V,n,\eta_\epsilon)\}.$$

By Proposition V.2 it can be seen that $\mathcal{S}_\epsilon$ forms an ϵ subspace cover of $K(V)$ and hence of $K^0(V)$ as well. Also, from the second and third properties of $\mathcal{P}(V,n,\eta)$ we get

$$\max_{S\in\mathcal{S}_\epsilon}\dim(S) \le 1 + C\max\big(\frac{V^2(\log n)^2}{\epsilon^2}, \frac{V\log n}{\epsilon}\big)$$

and

$$\log|\mathcal{S}_\epsilon| \le C\max\big(\frac{V^2(\log n)^3}{\epsilon^2}, \frac{V(\log n)^2}{\epsilon}\big).$$

We now have all the ingredients to apply Proposition IV.1 except for an upper bound on the diameter of $K^0(V)$. To this end we use Proposition V.3 — which we are going to state in the next subsection — to deduce that $t := \mathrm{diam}(K^0(V)) \le CV$. We thus obtain from Proposition IV.1, with $k_0 = \lfloor -\log_2\, t\rfloor$ and $k_1 = \lceil \log_2(\frac{n}{V})\rceil$,

$$\begin{aligned}&\mathcal{GW}(K^0(V))\\ &\le C\sum_{k=k_0+1}^{k_1} 2^{-k}\big(\frac{V(\log n)^{3/2}}{2^{-k}} + (\log \mathrm{e}n)^{1/2}\big) + n\,2^{-k_1}\\ &\le C\,V(k_1-k_0)(\log n)^{3/2} + C\cdot 2^{-k_0}(\log \mathrm{e}n)^{1/2} + V\\ &\le C\big(\log\big(\frac{tn}{V}\vee \mathrm{e}\big)V(\log \mathrm{e}n)^{3/2} + t(\log \mathrm{e}n)^{1/2}\big)\\ &\le CV(\log \mathrm{e}n)^{5/2}.\end{aligned} \tag{V.2}$$

Theorem II.1 now follows immediately from Lemma V.1 □

A. Proof of Proposition V.2

Given any partition P of L_n into rectangles, it is clear that the orthogonal projection of θ onto S_P, i.e., the unique matrix $\hat{\theta}_P \in S_P$ satisfying $\|\theta - \hat{\theta}_P\| = \mathrm{dist}(\theta, S_P)$, is constant on every rectangle R of P with the common value being the mean of $\theta_{|R}$ — the restriction of θ to R. Therefore, with $\bar{\theta}_{|R}$ denoting the mean of $\theta_{|R}$,

$$\mathrm{dist}(\theta, S_P)^2 = \|\theta - \hat{\theta}_P\|^2 = \sum_{R\in P} \|\theta_{|R} - \bar{\theta}_{|R}1_R\|^2 \leq |P| \max_{R\in P} \|\theta_{|R} - \bar{\theta}_{|R}1_R\|^2 \qquad \text{(V.3)}$$

where $1_R \in \mathbb{R}^R$ consists only of 1's. Our next result provides a way to bound the squared Frobenius distance between $\theta_{|R}$ and $\bar{\theta}_{|R}1_R$ in terms of the total variation of $\theta_{|R}$. This result, which is a discrete analogue of the Gagliardo-Nirenberg-Sobolev inequality for compactly supported smooth functions, will be crucial for deriving the first condition stipulated in Proposition V.2 for the particular partitioning scheme we are going to propose in this regard.

Proposition V.3 (Discrete Gagliardo-Nirenberg-Sobolev Inequality): Let $\theta \in \mathbb{R}^{m\times n}$ and $\overline{\theta} := \sum_{i=1}^m \sum_{j=1}^n \theta[i,j]/mn$ be the average of the elements of θ. Then we have, with $a \wedge b$ denoting the minimum of the (real) numbers a and b,

$$\sum_{i=1}^m \sum_{j=1}^n (\theta[i,j] - \overline{\theta})^2 \leq (5 + \frac{4mn}{n^2 \wedge m^2})\mathrm{TV}(\theta)^2 \,.$$

So in particular when $m = n$, we have

$$\sum_{i=1}^n \sum_{j=1}^n (\theta[i,j] - \overline{\theta})^2 \leq 9\mathrm{TV}(\theta)^2 \,.$$

Remark V.1: Although the Gagliardo-Nirenberg-Sobolev inequality is classical for Sobolev spaces (see, e.g., Chapter 12 in [24]), we are not aware of any discrete version in the literature that applies to arbitrary matrices of finite size. Also it is not clear if the inequality in this exact form follows directly from the classical version.

Now we give a scheme for subdividing θ in multiple steps until the total variation of each of the resulting submatrices is bounded above by η.

A greedy partitioning scheme: For convenience of exposition we will assume that n is an integer power of 2. The general n can then be accommodated from the following observation. For any $t > 0$, let $B_{n,n}(t)$ denote the t-Euclidean (Frobenius) ball in $\mathbb{R}^{n\times n}$ and consider $\theta \in K_n(V) \cap B_{n,n}(t)$ (recall from our proof of Theorem II.1 that we actually bound $\mathcal{GW}(K_n(V) \cap B_{n,n}(CV))$ for some universal constant C). Now let n' denote the smallest integer power of 2 that is larger than or equal to n and partition θ as

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix}$$

where $\theta_{22} \in \mathbb{R}^{(n'-n)\times(n'-n)}$. Also define a $n' \times n'$ matrix $f(\theta)$ as

$$f(\theta) = \begin{bmatrix} \theta_{11} & \theta_{12} & \overleftarrow{\theta_{12}} \\ \theta_{21} & \theta_{22} & \overleftarrow{\theta_{22}} \\ \theta_{21}\uparrow & \theta_{22}\uparrow & \overleftarrow{\theta_{22}}\uparrow \end{bmatrix}$$

where $\overleftarrow{M}$, for any matrix M, denotes the matrix obtained by reversing the order of its columns whereas $M\uparrow$ is obtained by reversing the order of its rows. It is clear from the definition that $f(\theta) \in K_{n'}(4V) \cap B_{n',n'}(t)$ and also

$$\begin{aligned} &\mathcal{GW}(K_n(V) \cap B_{n,n}(t)) \\ &\leq \mathbb{E} \sup_{\theta\in K_n(V):\|\theta\|\leq t} \langle Z_{n',n'}, f(\theta)\rangle \\ &\leq \mathcal{GW}(K_{n'}(4V) \cap B_{n',n'}(2t)) \end{aligned}$$

where $Z_{n',n'} \sim N(0_{n'\times n'}, I)$.

Let us now describe the scheme which is of the same flavor as the *breadth-first exploration* of a quaternary tree. The root node of the tree represents L_n and the nodes at any level (or depth) $i \in [\log_2\ n]$ represent (disjoint) rectangles of side-length $n2^{-i}$ with the property that the leaves of the tree truncated at level i form a partition of L_n. Given level $i-1$, the i-th level is constructed (or explored) as follows. For every leaf, i.e., rectangle R at level $i-1$ satisfying $\mathrm{TV}(\theta_{|R}) > \eta$, we add four children of R, namely R_{11}, R_{12}, R_{21} and R_{22}, to the tree where

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

and $\mathrm{n_{row}}(R_{11}) = \mathrm{n_{col}}(R_{11}) = \mathrm{n_{row}}(R)/2$. If the set of such leaves is empty or if $i - 1 = \log_2\ n$, we stop.

Let us denote the final rectangular partition of L_n obtained by applying the (TV, η) scheme to η as $P_{\theta,\eta}$ and the set of partitions $\{P_{\theta,\eta} : \theta \in \mathbb{R}^{n\times n}, \mathrm{TV}(\theta) \leq V\}$ as $\mathcal{P}(V, n, \eta)$. In our next result we verify that $\mathcal{P}(V, n, \eta)$ satisfy the last two properties stipulated in Proposition V.2.

Lemma V.4: There exists a universal constant $C > 0$ such that for any $\theta \in \mathbb{R}^{n\times n}$ and $\eta > 0$, we have

$$|P_{\theta,\eta}| \leq 1 + C\,\mathrm{TV}(\theta)\eta^{-1}\log n.$$

Furthermore, for any $V > 0$ we have

$$\log|\mathcal{P}(V, n, \eta)| \leq C\,V\eta^{-1}(\log n)^2 \,.$$

Proof: The basic idea of the proof hinges on super-additivity of the TV functional over disjoint rectangles. Let n_i denote the number of leaves in the tree formed by the scheme truncated at level i. In other words, n_i is the cardinality of the partition P_i formed by the rectangles corresponding to the leaves of the tree truncated at level i. Also let s_i denote the number of leaves R at level i satisfying $\mathrm{TV}(\theta_{|R}) > \eta$. Clearly $n_0 = 1$ and $n_{i+1} = n_i + 3\ s_i$. Notice that, due to super-additivity of the TV functional, we must have

$$s_i \leq \frac{\mathrm{TV}(\theta)}{\eta}. \qquad \text{(V.4)}$$

This implies in particular that

$$n_i \leq 1 + 3i\frac{\mathrm{TV}(\theta)}{\eta}. \qquad \text{(V.5)}$$

Since $i \leq \log_2\ n$ by construction, it then follows

$$|P_{\theta,\eta}| \leq 1 + 3\log_2\ n\frac{\mathrm{TV}(\theta)}{\eta}.$$

Next we bound the number of possible partitions $P_{\theta,\eta}$ when $\mathrm{TV}(\theta) \leq V$. The number of distinct ways of adding leaves at level $i+1$ is at most $\left(1 + 3\ i\frac{V}{\eta}\right)^{\frac{V}{\eta}}$ in light of the displays (V.4) and (V.5). Therefore

$$\log |\mathcal{P}(V, n, \epsilon)| \leq C\frac{V}{\eta}(\log n)^2$$

for some universal constant $C > 0$. □

With Lemma V.4 and Proposition V.3 in hand, we are now in a position to finish the proof of Proposition V.2.

Proof of Proposition V.2: For any given $\theta \in K(V)$, run the (TV, η) greedy scheme to obtain the partition $P_{\theta,\eta}$. Within every rectangle of the partition $P_{\theta,\eta}$ the total variation of θ is at most η. Also, the number of rectangles in $P_{\theta,\eta}$ is at most $1 + C\frac{V}{\eta}\log n$. Then by Proposition V.3 and (V.3) we can conclude

$$\|\tilde{\theta} - \theta\|^2 \leq CV\eta \log n + \eta^2.$$

Also, by Lemma V.4, as θ varies in $K(V)$, the number of distinct partitions $P_{\theta,\eta}$ that can be obtained is bounded by $\frac{V}{\eta}\log n$. This finishes the proof. □

Finally it remains to give the proof of Proposition V.3.

Proof of Proposition V.3: For any $(i, j) \in [m] \times [n]$, we have

$$(\theta[i,j] - \overline{\theta})^2 \leq \sum_{j' \in [j]} |\theta[i,j'] - \theta[i,j'-1]| \sum_{i' \in [i]} |\theta[i',j] - \theta[i'-1,j]|$$

where $\theta[i,0] = \theta[0,j] = \overline{\theta}$ for all $(i,j) \in [m] \times [n]$. Summing this over all i and j we get

$$\begin{aligned}
&\sum_{i\in[m],j\in[n]} (\theta[i,j] - \overline{\theta})^2 \\
&\leq \sum_{i'\in[m],j'\in[n]} \sum_{i \geq i', j \geq j'} |\theta[i,j'] - \theta[i,j'-1]| \times |\theta[i',j] - \theta[i'-1,j]| \\
&\leq \sum_{i'\in[m],j'\in[n]} \sum_{i\in[m],j\in[n]} |\theta[i,j'] - \theta[i,j'-1]| \times |\theta[i',j] - \theta[i'-1,j]| \\
&= \sum_{i\in[m],j\in[n]} |\theta[i,j] - \theta[i,j-1]| \sum_{i\in[m],j\in[n]} |\theta[i,j] - \theta[i-1,j]| \\
&= \big(\mathrm{TV}_{\mathrm{row}}(\theta) + \sum_{i\in[m]} |\theta[i,1] - \overline{\theta}|\big)\big(\mathrm{TV}_{\mathrm{col}}(\theta) + \sum_{j\in[n]} |\theta[1,j] - \overline{\theta}|\big)\,.
\end{aligned} \tag{V.6}$$

Here the total variation $\mathrm{TV}_{\mathrm{row}}(\theta)$ along rows is defined as

$$\mathrm{TV}_{\mathrm{row}}(\theta) \coloneqq \sum_{i\in[m]} \sum_{j\in[n-1]} |\theta[i,j+1] - \theta[i,j]|$$

and $\mathrm{TV}_{\mathrm{col}}(\theta) \coloneqq \mathrm{TV}_{\mathrm{row}}(\theta^T)$. Now let us try to bound $|\theta[i,1] - \overline{\theta}|$.

$$\begin{aligned}
&|\theta[i,1] - \overline{\theta}| \\
&\quad = \frac{1}{mn} \sum_{i'\in[m],j'\in[n]} |\theta[i,1] - \theta[i',j']| \\
&\quad \leq \frac{1}{mn} \sum_{i'\in[m],j'\in[n]} (|\theta[i,1] - \theta[i,j']| + |\theta[i,j'] - \theta[i',j']|) \\
&\quad \leq \frac{1}{n} \sum_{j'\in[n]} |\theta[i,1] - \theta[i,j']| + \frac{1}{mn} \sum_{i'\in[m],j'\in[n]} |\theta[i,j'] - \theta[i',j']| \\
&\quad \leq \mathrm{TV}(\theta[i,]) + \frac{1}{mn} \sum_{j'\in[n]} \sum_{i'\in[m]} \mathrm{TV}(\theta[,j']) \\
&\quad \leq \mathrm{TV}(\theta[i,]) + \frac{1}{n} \sum_{j'\in[n]} \mathrm{TV}(\theta[,j']) = \mathrm{TV}(\theta[i,]) + \frac{1}{n}\mathrm{TV}(\theta).
\end{aligned}$$

Hence

$$\sum_{i\in[m]} |\theta[i,1] - \overline{\theta}| \leq (1 + \frac{m}{n})\mathrm{TV}(\theta)\,.$$

Similarly

$$\sum_{j\in[n]} |\theta[1,j] - \overline{\theta}| \leq (1 + \frac{n}{m})\mathrm{TV}(\theta)\,.$$

Plugging these bounds into the last expression in (V.6), we get $sum_{i\in[m],j\in[n]}(\theta[i,j] - \overline{\theta})^2$ is at most

$$(2 + \frac{m}{n})(2 + \frac{n}{m})\mathrm{TV}(\theta)^2 \leq (5 + \frac{4mn}{n^2 \wedge m^2})\mathrm{TV}(\theta)^2\,.$$ □

VI. Proofs of Theorem II.2 and Theorem II.3

We first describe the precise connection between MSE and Gaussian widths. Recall that use $B_{m,n}(t)$ to denote the usual Euclidean ball of radius t in $\mathbb{R}^{m\times n}$. The *statistical dimension* of a closed convex cone $K \subset \mathbb{R}^N = \mathbb{R}^{n\times n}$ is defined as

$$\delta(K) := \mathbb{E}\|\Pi_K(Z)\|^2 \quad \text{where } Z \sim N(0, I)$$

and $\Pi_K(Z) := \operatorname*{argmin}_{u\in K} \|Z - u\|^2$ is the Euclidean projection of Z onto K. The terminology of statistical dimension is due to [1] and we refer the reader to this paper for many properties of the statistical dimension. The statistical dimension $\delta(K)$ is closely related to the Gaussian width of $K \cap B_{n,n}(1)$. It has been shown in [1, Proposition 10.2] that

$$\begin{aligned}
&\left[\mathcal{GW}(K \cap B_{n,n}(1))\right]^2 \leq \\
&\delta(K) \leq \left[\mathcal{GW}(K \cap B_{n,n}(1))\right]^2 + 1
\end{aligned} \tag{VI.1}$$

for every closed convex cone K.

The connection of the statistical dimension of tangent cones to the risk of $\hat{\theta}$ is the content of the following result due to [2, Corollary 2.2].

Theorem VI.1 ([2]): Suppose $Y \sim N(\theta^*, \sigma^2\ I)$ for some $\theta^* \in \mathbb{R}^N$. Then

$$\mathrm{MSE}(\hat{\theta}_V, \theta^*) \leq A$$

where

$$A = \inf_{\theta\in K(V)} \left[\frac{1}{N}\|\theta - \theta^*\|^2 + \frac{\sigma^2}{N}\delta(T_{K(V)}(\theta))\right].$$

Another result that is of use to us is the following result of [31] (Theorem 2.1). It says that the upper bound provided in Theorem VI.1 is essentially tight. Recall from Section II-B1 that $K^* = \{\theta \in \mathbb{R}^{n\times n} : \mathrm{TV}(\theta) \leq \mathrm{TV}(\theta^*)\}$.

Theorem VI.2 ([31]):

$$\begin{aligned}
&\lim_{\sigma\to 0} \frac{1}{\sigma^2}\mathrm{MSE}(\hat{\theta}_{V^*}, \theta^*) = \frac{1}{N}\delta(T_{K^*}(\theta^*)) \geq \\
&\frac{1}{N}\left[\mathcal{GW}(T_{K^*}(\theta^*) \cap B_{n,n}(1))\right]^2.
\end{aligned}$$

Remark VI.1: To clarify, Theorem 2.1 in [31] actually says that

$$\lim_{\sigma \to 0} \frac{1}{\sigma^2} \mathrm{MSE}(\hat{\theta}_{\mathbf{V}^*}, \theta^*) = \mathbb{E}\, \mathrm{dist}^2(Z, \mathrm{Polar}(T_{K^*}(\theta^*)))\,.$$

Here Z, as usual, refers to a matrix of independent $N(0,1)$ entries, $\mathrm{Polar}(T_{K^*}(\theta^*))$ refers to the Polar Cone of $T_{K^*}(\theta^*)$ and dist refers to the Euclidean Distance between two sets. Letting K denote a general cone and Π_K denote the Euclidean projection operator onto K, the standard Pythagorean Theorem for cones implies

$$\mathrm{dist}^2(Z, \mathrm{Polar}(K)) = \|\Pi_K(Z)\|^2\,.$$

Also, it holds that $\|\Pi_K(Z)\| = \sup_{\theta \in K : \|\theta\| \le 1} \langle Z, \theta \rangle$. A proof of the above fact is available in Lemma $\bar{A}$.3 in [11]. Theorem VI.2 now follows from applying the above facts to Theorem 2.1 in [31] and then using the elementary inequality $\mathbb{E}X^2 \geq (\mathbb{E}X)^2$.

In light of the above facts and armed with Proposition II.4 and Proposition II.5 we are now ready to prove Theorem II.2 and Theorem II.3 respectively.

Proof of Theorem II.2: Theorem VI.1 along with (VI.1) gives us

$$\mathrm{MSE}(\hat{\theta}_V, \theta^*) \leq \inf_{\theta \in K(V)} \left[\frac{1}{N}\|\theta - \theta^*\|^2 + \frac{\sigma^2}{N} + \frac{\sigma^2}{N} \big[\mathcal{GW}(T_{K(V)}(\theta))\big]^2 \right]. \tag{VI.2}$$

With $V^* = \mathrm{TV}(\theta^*) > 0$, define

$$\theta := \overline{\theta^*}\mathbf{1} + \frac{V}{V^*}\big(\theta^* - \overline{\theta^*}\mathbf{1}\big).$$

By definition, $\mathrm{TV}(\theta) = V$ and θ is piecewise constant on the same partition of L_n as is θ^*. By Proposition V.3 we can assert that

$$\|\theta - \theta^*\|^2 = (V - V^*)^2 \frac{\|\theta^* - \overline{\theta^*}\mathbf{1}\|^2}{(V^*)^2} \leq 9(V - V^*)^2.$$

Therefore, in view of (VI.2), we obtain

$$\mathrm{MSE}(\hat{\theta}_V, \theta^*) \leq \frac{9}{N}(V - V^*)^2 + \frac{\sigma^2}{N} + \frac{\sigma^2}{N} C(\log en)^9\ k(\theta^*)^{5/4} N^{1/4}$$

where we have also used Proposition II.4 and the fact that $k(\theta) = k(\theta^*)$. □

Proof of Theorem II.3: The proof of Theorem II.3 is immediate once we use Theorem VI.2 along with Proposition II.5. □

VII. Proofs of Proposition II.5 and Theorem II.6

A. Tangent Cone Characterization

We fix a $\theta^* \in \mathbb{R}^{n \times n}$ and proceed to investigate the tangent cone $T_{K^*}(\theta^*)$. Notice that K^* is same as $K(V^*)$ defined in Section II-B1 (see (V.1)). Let $\mathcal{R}^* = (R_1^*, R_2^*, \ldots, R_{k^*}^*)$ be a partition of $[n] \times [n]$ into k^* rectangles where $k^* = k(\theta^*)$.

Recall that the vertices in the grid graph L_n correspond to the pairs $(i,j) \in [n] \times [n]$ and its edge set E_n consists of:

$$\text{all } ((i,j),(k,\ell)) \in L_n \times L_n \text{ such that } |i - j| + |k - \ell| = 1\,.$$

For any edge $e \in E_n$, we denote by e^+ and e^- the vertices associated with e with respect to the natural partial order. For any $\theta \in \mathbb{R}^{L_n}$, we will use $\Delta_e \theta$ as a shorthand notation for the (discrete) *edge gradient* $\theta(e^+) - \theta(e^-)$. Thus $\mathrm{TV}(\theta) = \sum_{e \in E_n} |\Delta_e \theta|$. For a general rectangle $R := ([a_1, a_2] \times [b_1, b_2]) \cap \mathbb{Z}^2 \subset L_n$, we define its right boundary as follows:

$$\partial^{\mathrm{right}}(R) := \{(i,j) \in R : j = b_2\}\,.$$

While defining the above set, we are using the matrix convention for indexing the vertices of L_n. Thus, the top-left vertex in the two-dimensional array L_n is indexed by $(1,1)$ and the bottom-right vertex by (n,n). Similarly we define the left, top and bottom boundaries of R and denote them by $\partial^{\mathrm{left}}(R)$, $\partial^{\mathrm{top}}(R)$ and $\partial^{\mathrm{bottom}}(R)$ respectively. The boundary of R, denoted by ∂R, is defined as

$$\partial R := \partial^{\mathrm{right}}(R) \cup \partial^{\mathrm{left}}(R) \cup \partial^{\mathrm{top}}(R) \cup \partial^{\mathrm{bottom}}(R)\,.$$

1) Starting From the Definition: The tangent cone $T_{K(V^*)}(\theta^*)$ is the smallest closed, convex cone containing all the elements θ in $\mathbb{R}^{n \times n}$ such that $\theta^* + \theta \in K(V^*)$ for $V^* = \mathrm{TV}(\theta^*)$. Let $A^* := \{e \in E_n : |\Delta_e \theta^*| > 0\}$ and $(A^*)^c = E_n \setminus A$. Observe that $|\Delta_e(\theta^* + \theta)| - |\Delta_e(\theta^*)| = |\Delta_e \theta| - 0 = |\Delta_e \theta|$ for every edge e in $(A^*)^c$. Thus in order for $\theta^* + \theta \in K(V^*)$, the increments in the absolute edge gradients of $\theta^* + \theta$ from the edges in $(A^*)^c$ must be compensated by an equal or greater amount of decrease in the absolute edge gradients for the edges in A^*. The precise statement is the content of

Lemma VII.1: We have the following set equality:

$$T_{K(V^*)}(\theta^*) = \Big\{\theta \in \mathbb{R}^{n \times n} : \sum_{e \in (A^*)^c} |\Delta_e \theta| \leq - \sum_{e \in A^*} sgn(\Delta_e \theta^*) \Delta_e \theta\Big\} \tag{VII.1}$$

Here, $sgn(x) := \mathbb{I}\{x > 0\} - \mathbb{I}\{x < 0\}$ is the usual sign function.

Proof: Let T be the set on the right side of (VII.1). Let us first prove that $T_{K(V^*)}(\theta^*) \subset T$. An important feature of T is that it is a closed convex cone. Hence it suffices to show that $\theta \in T$ whenever $\theta^* + \theta \in K(V^*)$. To this end let θ be such that $\mathrm{TV}(\theta^* + \theta) \leq \mathrm{TV}(\theta^*)$. Since $K(V^*)$ is a convex set, we have

$$\mathrm{TV}(\theta^* + c\theta) = \mathrm{TV}\big(c(\theta^* + \theta) + (1 - c)\theta^*\big) \leq \mathrm{TV}(\theta^*)$$

for any $0 \leq c \leq 1$. Now observing that

$$\mathrm{TV}(\theta^* + c\theta) = \sum_{e \in A^*} |\Delta_e \theta^* + c\Delta_e \theta| + c \sum_{e \in (A^*)^c} |\Delta_e \theta|\,,$$

we can write

$$\mathrm{TV}(\theta^* + c\theta) = \sum_{e \in A^*} \big[sgn(\Delta_e \theta^*)\Delta_e \theta^* + c\, sgn(\Delta_e \theta^*)\Delta_e \theta\big] + c \sum_{e \in (A^*)^c} |\Delta_e \theta| \leq \mathrm{TV}(\theta^*) \tag{VII.2}$$

whenever c is small enough satisfying $sgn(\Delta_e \theta^* + c\Delta_e \theta) = sgn(\Delta_e \theta^*)$ for all $e \in A^*$. By definition,

$$\mathrm{TV}(\theta^*) = \sum_{e \in A^*} sgn(\Delta_e \theta^*)\Delta_e \theta^*$$

which together with (VII.2) gives us $\theta \in T$.

It remains to show that $T \subset T_{K(V^*)}(\theta^*)$. It suffices to show that for any $\theta \in T$ there exists a small enough $c > 0$ such that $\mathrm{TV}(\theta^* + c\theta) \leq V^*$. This can be shown using the same reasoning given after (VII.2). □

With the above characterization of the tangent cone, we are now ready to prove our lower bound to the risk given in Theorem II.3.

B. Proof of Proposition II.5

Recall that here we consider θ^* which is piecewise constant on two rectangles and is of the following form:

$$\theta^* = \begin{bmatrix} \mathbf{0}_{n \times n/2} & \mathbf{1}_{n \times n/2} \end{bmatrix}$$

Proof: Consider n to be even and a perfect square (i.e., $\sqrt{n}$ is an integer) for simplicity of exposition. Also for a generic $n \times n$ matrix θ we will denote $\theta^{(1)}$ to be the submatrix formed by the first $n/2 - 1$ columns, $v^{(\theta)}$ to be the $n/2$-th column and $\theta^{(2)}$ to be the submatrix formed by the last $n/2$ columns. Also, for two matrices θ and $\theta^{'}$ with the same number of rows, we will denote $[\theta : \theta^{'}]$ to be the matrix obtained by concatenating the columns of θ and $\theta^{'}$.

We can now use Lemma VII.1 to characterize the tangent cone $T_{K(V^*)}(\theta^*)$.

$$T_{K(V^*)}(\theta^*) = \big\{\theta \in \mathbb{R}^{n \times n} : \mathrm{TV}([\theta^{(1)} : v^{(\theta)}]) + \mathrm{TV}(\theta^{(2)}) \leq \sum_{i=1}^{n} \theta[i, n/2] - \theta[i, n/2 + 1]\big\}$$

In this proof, we will actually lower bound the Gaussian width of a convenient subset of $T_{K(V^*)}(\theta^*)$. To this end, for constants $c_1, c_2 \in (0, 1)$ to be specified later, let us define

$$S := \big\{\theta \in T_{K(V^*)}(\theta^*) : \theta^{(1)} = \frac{c_1}{n}\mathbf{1}_{n \times (n/2-1)}, \theta^{(2)} = \mathbf{0}_{n \times n/2},\ v^{(\theta)} \in \{c_1/n, c_2/\sqrt{n}\}^n\big\}.$$

In words, for $\theta \in S$, the first $n/2 - 1$ columns are all equal to c_1/n, the last $n/2$ columns of θ are 0 and the entries in the $n/2$-th column can take two values; either $c_2/\sqrt{n}$ or c_1/n. Also, for any *such* matrix θ,

$$\mathrm{TV}([\theta^{(1)} : v^{(\theta)}]) \leq \sum_{i=1}^{n} v_i^{(\theta)} \iff \theta \in S. \tag{VII.3}$$

Before going further, let us define the set of indices $B_j := \{(j-1)\sqrt{n} + 1, (j-1)\sqrt{n} + 2, \ldots, j\sqrt{n}\}$ for $j \in [\sqrt{n}]$. In words, we divide $[n]$ into $\sqrt{n}$ many equal contiguous blocks and B_j refers to the jth block. Now, for any realization of a random Gaussian matrix Z, let us define the matrix ν so that $\nu^{(1)} := \frac{c_1}{n}\mathbf{1}_{n \times (n/2-1)}$ and $\nu^{(2)} := \mathbf{0}_{n \times n/2}$. Moreover, we define $v^{(\nu)}$ as follows:

$$v_i^{(\nu)} := \sum_{j \in [\sqrt{n}] : B_j \ni i} \big(\mathbb{I}\{\textstyle\sum_{k \in B_j} Z[k, n/2] > 0\}\frac{c_2}{\sqrt{n}} + \mathbb{I}\{\textstyle\sum_{k \in B_j} Z[k, n/2] < 0\}\frac{c_1}{n}\big).$$

In words, the vector $v^{(\nu)}$ is defined so that it is constant on each of the blocks B_j. If $\sum_{i \in B_j} Z[i, n/2] > 0$, the value on B_j is $\frac{c_2}{\sqrt{n}}$, otherwise the value is $\frac{c_1}{n}$. Now we claim that the following are true for some appropriate choice of c_1 and c_2:

a) $\nu \in S$ for any Z.

b) $\|\nu\| \leq 1$.

Taking the above claims to be true we can write

$$\begin{aligned} &\mathcal{GW}\big(T_{K(V^*)}(\theta^*) \cap B_{n,n}(1)\big) \geq \mathcal{GW}\big(S \cap B_{n,n}(1)\big) \\ &= \mathbb{E} \sup_{\theta \in S \cap B_{n,n}(1)} \langle \theta, Z\rangle \geq \mathbb{E}\langle \nu, Z\rangle \\ &= \mathbb{E}\langle \nu^{(1)}, Z^{(1)}\rangle + \mathbb{E}\langle \nu^{(2)}, Z^{(2)}\rangle + \mathbb{E}\sum_{i=1}^{n} v_i^{(\nu)} Z[i, n/2] \\ &= \mathbb{E}\sum_{i=1}^{n} v_i^{(\nu)} Z[i, n/2] \end{aligned}$$

where we used the fact that $\nu^{(1)}, \nu^{(2)}$ are constant matrices and Z has mean zero entries.

Now let us denote $\mathcal{Z}_j := \sum_{i \in B_j} Z[i, n/2]$. Note that $(\mathcal{Z}_1, \ldots, \mathcal{Z}_{\sqrt{n}})$ are independent mean zero Gaussians with standard deviation $n^{1/4}$. Therefore

$$\begin{aligned} &\mathbb{E}\sum_{i=1}^{n} v_i^{(\nu)} Z[i, n/2] = \\ &\sum_{j \in [\sqrt{n}]} \mathbb{E}\big(\mathbb{I}\{\mathcal{Z}_j > 0\}\mathcal{Z}_j\frac{c_2}{\sqrt{n}} + \mathbb{I}\{\mathcal{Z}_j < 0\}\mathcal{Z}_j\frac{c_1}{n}\big) \\ &= \sum_{j \in [\sqrt{n}]} (\frac{c_2}{\sqrt{n}} - \frac{c_1}{n})n^{1/4}\phi = (c_2 - \frac{c_1}{\sqrt{n}})\phi n^{1/4} \end{aligned}$$

where for a standard Gaussian random variable z, we denote $\phi = \mathbb{E}\, z\mathbb{I}\{z > 0\}$.

It remains to choose c_1, c_2 so that the two claims hold as well as $c_2 - \frac{c_1}{\sqrt{n}}$ is positive. To this end notice that for validating the first claim it suffices to show, in view of the definition of ν, that the first inequality in (VII.3) holds for ν, i.e., the following is true

$$\mathrm{TV}([\nu^{(1)} : v^{\nu}]) \leq \sum_{i=1}^{n} v_i^{\nu}\,. \tag{VII.4}$$

Now entries of v^{ν} can take two values, either $\frac{c_2}{\sqrt{n}}$ or $\frac{c_1}{n}$. In either case it can be checked that when $c_2 \geq \frac{c_1}{\sqrt{n}}$ we have for each row index $i \in [n]$

$$v_i^{\nu} - \mathrm{TV}(\nu^{(1)}[i, 1], \ldots, \nu^{(1)}[i, n-1], v_i^{\nu}) = \frac{c_1}{n}. \tag{VII.5}$$

Along with the fact that

$$\mathrm{TV}([\nu^{(1)} : v^{\nu}]) = \sum_{i=1}^{n} \mathrm{TV}(\nu^{(1)}[i,1], \dots, \nu^{(1)}[i,n-1], v_i^{\nu}) + \mathrm{TV}(v^{\nu}),$$

(VII.5) implies that in order to verify (VII.4) it suffices to show $\mathrm{TV}(v^{\nu}) \le c_1$. But v^{ν} is a piecewise constant vector with at most $\sqrt{n}$ jumps of size $\frac{c_2}{\sqrt{n}}$. Thus we have $\mathrm{TV}(v^{\nu}) \le c_2$. Hence ensuring $c_2 \le c_1$ is sufficient to obtain the first claim. The second claim is trivially satisfied if $c_2 \le \sqrt{1 - c_1^2}$. Thus, choosing $c_1 = c_2 = 1/\sqrt{2}$ we can satisfy both claims as well as $c_2 - \frac{c_1}{\sqrt{n}} = \frac{1}{\sqrt{2}}(1 - 1/\sqrt{n}) > 0$ for all $n \ge 2$. □

The task now is to obtain a "matching" upper bound on the gaussian width, which would eventually lead to the proof of Theorem II.2 in view of Theorem VI.1. Since the proof is lengthy and somewhat technical, for the benefit of the reader we first provide an informal roadmap of the proof before starting it formally in Section VIII.

C. Proof of Theorem II.6

Proof: Consider the signal matrix $\theta^* := \mathbb{I}\{i + j > n\}$. From the characterization of the tangent cone given by Lemma VII.1, we have

$$T_{K(V^*)}(\theta^*) = \Big\{\theta \in \mathbb{R}^{n \times n} : \sum_{e \in (A^*)^c} |\Delta_e \theta| \le -\sum_{e \in A^*} \Delta_e \theta\Big\}$$

where every edge e in A^* is either of the form $((i, n-i), (i, n-i+1))$ or $((i, n-i), (i+1, n-i))$ for some $i \in [n-1]$.

Now consider the family T^* of matrices defined below:

$$T^* := \{\theta \in \mathbb{R}_+^{n \times n} : \theta[i,j] = 0 \ \forall\, (i,j) \text{ satisfying } i + j \ne n\}.$$

It is not difficult to check that $T^* \subseteq T_{K(V^*)}(\theta^*)$. It is also clear that T^* is (linearly) isomorphic to $\mathbb{R}_+^{n-1}$. Therefore

$$\begin{aligned}\mathcal{GW}(T_{K(V^*)}(\theta^*) \cap B_{n \times n}(1)) &\ge \mathcal{GW}(T^* \cap B_{n \times n}(1)) \\ &= \mathcal{GW}(\mathbb{R}_+^{n-1} \cap B_{n-1}(1)) \ge c\sqrt{n}.\end{aligned}$$

where $B_m(r)$ denotes the usual Euclidean ball of radius r in $\mathbb{R}^m$ and $c > 0$ is a universal constant. Now an application of Theorem VI.2 along with the above Gaussian width lower bound also furnishes a lower bound to the limiting MSE. □

Remark VII.1: The vertex boundary of a set $A \subset L_n$ is defined to be the set of vertices which share an edge with A^c. Consider the level sets of θ^* which are the sets $A = \{(i,j) \in L_n : i + j > n\}$ and A^c. The simple argument presented in the proof of Theorem II.6 relies crucially on the fact that the vertex boundary of the level sets A and A^c are not connected in the graph L_n. One can now consider other signals of the form $\theta^* = \mathbb{I}\{A\}$ for a general subset $A \subset L_n$. One can check that if A is of the shape of a circle or a square rotated by 45 degrees then also the vertex boundary of the level sets will contain $O(n)$ connected components which are singletons. Therefore, a similar argument will give a $O(n)$ lower bound to $\mathcal{GW}(T_{K(V^*)}(\theta^*)$. We believe that it might be possible to formalize the intuition that whenever A is sufficiently far from being a rectangle, $\mathcal{GW}(T_{K(V^*)}(\theta^*)$ is lower bounded by $O(n)$.

VIII. Proof of Proposition II.4

A. Informal Roadmap

The proof of Proposition II.4 can be divided into three major steps which we now describe. Recall that θ^* is the true signal which is piecewise constant on axis aligned rectangles $R_1^*, \dots, R_{k(\theta^*)}^*$ which partition L_n.

Step 1: We have to bound $\mathcal{GW}(T_{K(V^*)}(\theta^*) \cap B_{n \times n}(1))$. To do this, we show that if a matrix θ is in $T_{K(V^*)}(\theta^*) \cap B_{n \times n}(1)$ then each rectangular submatrix $\theta_{R_i^*}$ satisfies the property that $\mathrm{TV}(\theta_{R_i^*})$ is at most the ℓ_1 norm of its four boundaries plus a small wiggle room $\delta > 0$. Such matrices are denoted later in (VIII.2) as $\mathcal{M}^4$. This fact then reduces our problem to bounding the Gaussian width for the class of matrices $\mathcal{M}^4$. Corollary VIII.1, Lemma VIII.2 and Lemma VIII.3 are part of this step.

Step 2: Before starting the Gaussian width calculations, we found it convenient to further simplify the class of matrices $\mathcal{M}^4$. In this step, we show that if a matrix θ lies in $\mathcal{M}^4$ then we can subdivide it further into several submatrices which now satisfy a simpler property. The property is that the total variation of these submatrices are at most the ℓ_1 norm of only one or none of its boundaries (instead of four) plus an appropriately small "wiggle room" $\delta > 0$. These sets of matrices are denoted by $\mathcal{M}^1$ and $\mathcal{M}^0$ respectively and are defined just before Lemma VIII.4. Along with Lemma VIII.4, Lemmata VIII.5–VIII.7 are also parts of this step.

Step 3: This is the step where we actually compute the metric entropies of the classes of matrices $\mathcal{M}^1$ and $\mathcal{M}^0$ and finally bring all the pieces together. Proposition VIII.9 and Lemmata VIII.8–VIII.14 are all parts of this step.

B. Towards Simplifying the Tangent Cone

We first want to split θ into submatrices each of which satisfies a separate constraint. This and the next subsection are devoted to this goal. Let us revisit Lemma VII.1. Since θ^* is constant on each rectangle $R_i^* \in \mathcal{R}^*$ it follows that

$$A^* = \{e \in E_n : e^+ \in R_i^* \text{ and } e^- \in R_j^* \text{ for some } i \ne j \in [k^*]\}.$$

As a consequence we get the following corollary:

Corollary VIII.1: Fix $\theta^* \in \mathbb{R}^{n \times n}$. We have

$$T_{K(V^*)}(\theta^*) \subset \Big\{\theta \in \mathbb{R}^{n \times n} : \sum_{i \in [k^*]} \mathrm{TV}(\theta_{R_i^*}) \le \sum_{i \in [k^*]} \sum_{u \in \partial R_i^*} |\theta(u)|\Big\}.$$

The first step towards obtaining a decomposition where each submatrix satisfies some constraint is to separate the constraints for R_i^*'s. More precisely we would like

$$\mathrm{TV}(\theta_{R_i^*}) \le \sum_{u \in \partial R_i^*} |\theta(u)| \qquad \text{(VIII.1)}$$

for each $i \in [k^*]$. As we will see below that this is "almost" the truth when we consider matrices in the tangent cone which are of unit norm.

Let us make precise the notion of an "almost" version of (VIII.1). To this end we introduce for any δ, $t > 0$:

$$\begin{aligned}&\mathcal{M}^4(m', n', \delta, t) := \\ &\{\theta \in \mathbb{R}^{m' \times n'} : \mathrm{TV}(\theta) \le \|\theta_{\rm left}\|_1 + \|\theta_{\rm right}\|_1 + \\ &\|\theta_{\rm top}\|_1 + \|\theta_{\rm bottom}\|_1 + \delta,\ \|\theta\| \le t\}\,,\end{aligned} \qquad \text{(VIII.2)}$$

where $\theta_{\rm left} := \theta[\ ,\ 1]$, $\theta_{\rm right} := \theta[\ ,\ n']$, $\theta_{\rm top} := \theta[1,\ ]$ and $\theta_{\rm bottom} := \theta[m',\ ]$. In plain words, $\mathcal{M}^4(m', n', \delta, t)$ consists of matrices of norm at most t whose total variation is bounded by the total ℓ_1 norm of its four boundaries plus an extra *wiggle room* $\delta > 0$. In our next result we show that for any θ in $T_{K(V^*)}(\theta^*)$ intersected with the unit Euclidean ball $B_{n\times n}(1)$, the restriction $\theta_{|R_i^*}$ of θ to R_i^* lies in $\mathcal{M}^4(m_i, n_i, \delta_i, t_i)$ for each $i \in [k]$ with $m_i := \mathrm{n_{row}}(R_i^*)$, $n_i := \mathrm{n_{col}}(R_i^*)$ and t_i's and δ_i's satisfying some upper bounds on their ℓ_2 and ℓ_1-norms respectively.

Lemma VIII.2: We have the set inclusion

$$\begin{aligned}&T_{K(V^*)}(\theta^*) \cap B_{n\times n}(1) \subset \\ &\bigcup_{\boldsymbol{\delta} \in S_{k^*, \Delta(\theta^*)}} \ \bigcup_{\boldsymbol{t^2} \in S_{k^*,1}} \{\theta \in \mathbb{R}^{n\times n} : \\ &\theta_{|R_i^*} \in \mathcal{M}^4(m_i, n_i, \delta_i, t_i),\ \forall i \in [k^*]\}\end{aligned}$$

where $S_{k^*,r} := \{a \in \mathbb{R}_+^{k^*} : \sum_{i\in[k^*]} a_i \le r\}$ is the non negative simplex with radius $r > 0$, $\boldsymbol{t^2}$ is the vector $(t_1^2, \ldots, t_{k^*}^2) \in \mathbb{R}_+^{k^*}$ and

$$\Delta(\theta^*) = \sqrt{2\sum_{i\in[k^*]} \Big(\frac{m_i}{n_i} + \frac{n_i}{m_i}\Big)}.$$

Remark VIII.1: By virtue of Lemma VIII.2, we achieve our objective of obtaining a characterization of $T_{K(V^*)}(\theta^*)$ where we have separate constraints for each $R_i^* \in \mathcal{R}^*$. The constraints are now coupled together by the wiggle room vector $\boldsymbol{\delta} \in S_{k^*, \Delta(\theta^*)}$ and the (squared) ℓ_2-norm vector $\boldsymbol{t^2}$.

Proof: We will start with a claim.

Claim VIII.1: Let $\theta \in T_{K(V^*)}(\theta^*) \cap B_{n\times n}(1)$. Then for each $i \in [k^*]$ and any fixed choice of rows and columns r_i, c_i in R_i^*, we have $\theta_{|R_i^*} \in \mathcal{M}^4(m_i, n_i, \delta_i, t_i)$ where $\sum_{i\in[k^*]} t_i^2 \le 1$ and $(\delta_1, \ldots, \delta_{k^*}) =: \boldsymbol{\delta} \in \mathbb{R}_+^{k^*}$ satisfies

$$\|\boldsymbol{\delta}\|_1 \ \le\ 2\sum_{i\in[k^*]} \Big(\big\|\theta_{|c_i}\big\|_1 + \big\|\theta_{|r_i}\big\|_1\Big)\,,$$

where $\theta_{|c_i}$ (or $\theta_{|r_i}$) is the vector obtained by restricting θ to the row c_i (respectively the column r_i).

Let us first deduce the lemma assuming our claim. Consider a $\theta \in T_{K(V^*)}(\theta^*)$ such that $\|\theta\|_2 \le 1$ and for each $i \in [k^*]$, let r_i and c_i denote the rows and columns such that the ℓ_1 norms of $\theta_{|r_i}$ and $\theta_{|c_i}$ are minimum. Then by Claim VIII.1, each $\theta_{|R_i^*} \in \mathcal{M}^4(m_i, n_i, \delta_i, t_i)$ with $\boldsymbol{t^2} \in S_{k^*,1}$ and $\boldsymbol{\delta}$ satisfying

$$\|\boldsymbol{\delta}\|_1 \ \le\ 2\sum_{i\in[k^*]} \min_{\substack{c:\, c \text{ is a column of } R_i^*,\\ r:\, r \text{ is a row of } R_i^*}} \Big(\big\|\theta_{|c}\big\|_1 + \big\|\theta_{|r}\big\|_1\Big). \qquad \text{(VIII.3)}$$

Now for each $i \in [k^*]$, we have

$$\begin{aligned}&\min_{c:\, c \text{ is a column of } R_i^*} \big\|\theta_{|c}\big\|_1 \le \\ &\sqrt{m_i} \min_{c: c \text{ is a column of } R_i^*} \big\|\theta_{|c}\big\|_2 \le \sqrt{\frac{m_i}{n_i}}\|\theta_{|R_i^*}\|_2\,.\end{aligned}$$

The first inequality is an application of the Cauchy-Schwarz inequality and the second inequality follows from the "minimum is less than the average" principle. Similarly, one can obtain the row version of these inequalities and together they give us

$$\begin{aligned}&\min_{\substack{c:\, c \text{ is a column of } R_i^*,\\ r:\, r \text{ is a row of } R_i^*}} \Big(\big\|\theta_{|c}\big\|_1 + \big\|\theta_{|r}\big\|_1\Big) \le \\ &\Big(\sqrt{\frac{m_i}{n_i}} + \sqrt{\frac{n_i}{m_i}}\Big)\|\theta_{|R_i^*}\|_2.\end{aligned}$$

Summing the above inequality over all $i \in [k^*]$ and subsequently using the Cauchy-Schwarz inequality as well as the fact that $\|\theta\|_2 \le 1$, we get in view of (II.2)

$$\|\boldsymbol{\delta}\|_1 \le \sqrt{2\sum_{i\in[k^*]} \Big(\frac{m_i}{n_i} + \frac{n_i}{m_i}\Big)} = \Delta(\theta^*)\,,$$

thus yielding the lemma.

Proof of Claim VIII.1. The constraint on t_i's is clear and therefore all we need to show is the constraint on δ_i's. Recall from the definition in (VIII.2) that δ_i can be chosen, for any $i \in [k^*]$, as

$$\delta_i := \Big(\mathrm{TV}(\theta_{|R_i^*}) - \sum_{u\in\partial R_i^*} |\theta(u)|\Big)_+ \qquad \text{(VIII.4)}$$

where $a_+ := \max\{a, 0\}$ for any $a \in \mathbb{R}$. Now fix $i \in [k^*]$ and consider a generic row r_i of R_i^*. Treating r_i as a horizontal path in the graph L_n, let us denote its two end-vertices by u and w with $u \in \partial^{\rm left}(R_i^*)$ and $w \in \partial^{\rm right}(R_i^*)$. Now denoting the vertex in $r_i \cap c_i$ by v, we see that v occurs between the vertices u and w in the row r_i. Therefore we can write

$$\mathrm{TV}(\theta_r) \ge |\theta(u)| + |\theta(w)| - 2|\theta(v)|\,.$$

Summing the above inequality for every row in the rectangle R_i^* gives us

$$\begin{aligned}&\mathrm{TV_{row}}(\theta_{|R_i^*}) \ge \\ &\sum_{u\in\partial^{\rm left}(R_i^*)} |\theta(u)| + \sum_{w\in\partial^{\rm right}(R_i^*)} |\theta(w)| - 2\,\|\theta_{c_i}\|_1\,.\end{aligned}$$

By a similar argument applied to the columns of R we obtain

$$\begin{aligned}&\mathrm{TV_{col}}(\theta_{|R_i^*}) \ge \\ &\sum_{u\in\partial^{\rm top}(R_i^*)} |\theta(u)| + \sum_{w\in\partial^{\rm bottom}(R_i^*)} |\theta(w)| - 2\,\|\theta_{r_i}\|_1\,.\end{aligned}$$

Summing the previous two displays we get the following inequality:

$$\mathrm{TV}(\theta_{|R_i^*}) \ge \sum_{u\in\partial R_i^*} |\theta(u)| - 2\,\|\theta_{r_i}\|_1 - 2\,\|\theta_{c_i}\|_1\,.$$

Now if $\theta \in T_{K(V^*)}(\theta^*)$, then as a consequence of Corollary VIII.1 we also have

$$\sum_{i \in [k^*]} \mathrm{TV}(\theta_{|R_i^*}) \leq \sum_{i \in [k^*]} \sum_{u \in \partial R_i^*} |\theta(u)| \,.$$

Hence an application of Lemma A.1 (stated and proved in the appendix) to $f_i = \mathrm{TV}(\theta_{|R_i^*})$, $g_i = \sum_{u \in \partial R_i^*} |\theta(u)|$, $h_i = 2(\|\theta_{r_i}\|_1 + \|\theta_{c_i}\|_1)$ and $w_i = \delta = 0$, would give us the claim in view of (VIII.4). □

With the help of Lemma VIII.2 we can now deduce the following lemma.

Lemma VIII.3: With the notation described in this section, we have the following upper bound:

$$\begin{aligned} &\mathcal{GW}(T_{K(V^*)}(\theta^*) \cap B_{n,n}(1)) \leq \\ &\max_{\Delta(\theta^*)\boldsymbol{\delta}: \boldsymbol{\delta} \in S_{k^*,2} \cap H_{k^*}} \; \max_{\boldsymbol{t^2} \in S_{k^*,2} \cap H_{k^*}} \Big[\\ &\sum_{i \in [k^*]} \mathcal{GW}(\mathcal{M}^4(m_i, n_i, \Delta(\theta^*)\delta_i, t_i)) \Big] \\ &+ \; C\sqrt{k^*}. \end{aligned}$$

where $H_{k^*} := \{\frac{1}{k^*}, \frac{2}{k^*}, \ldots, 1\}^{k^*}$ and $C > 0$ is a universal constant.

Proof: Using Lemma VIII.2 we can write

$$\begin{aligned} &\mathbb{E} \sup_{\theta \in T_{K(V^*)}(\theta^*): \|\theta\| \leq 1} \langle Z, \theta \rangle \leq \\ &\mathbb{E} \sup_{\boldsymbol{\delta} \in S_{k^*,\Delta(\theta^*)}} \; \sup_{\boldsymbol{t^2} \in S_{k^*,1}} \Big[\\ &\sum_{i \in [k^*]} \sup_{\theta \in T_{K(V^*)}(\theta^*): \|\theta\| \leq 1} \langle Z_{|R_i^*}, \theta_{|R_i^*} \rangle \Big] \\ &\leq \mathbb{E} \sup_{\boldsymbol{\delta} \in S_{k^*,\Delta(\theta^*)}} \; \sup_{\boldsymbol{t^2} \in S_{k^*,1}} \Big[\\ &\sum_{i \in [k^*]} \sup_{\theta_i \in \mathcal{M}^4(m_i, n_i, \delta_i, t_i)} \langle Z, \theta_i \rangle \Big] \end{aligned} \tag{VIII.5}$$

where, by a slight abuse of notation, Z always refers to a matrix of independent standard normals with appropriate number of rows and columns.

At this point, we would like to convert the supremum over $\boldsymbol{\delta}, \boldsymbol{t^2}$ (or, equivalently $\boldsymbol{t}$) in the non negative simplex to a maximum over a finite net of $\boldsymbol{\delta}, \boldsymbol{t}$ We can accomplish this by the following trick. Fix any $\boldsymbol{\delta} \in S_{k^*,\Delta(\theta^*)}$. Then we can define a vector $\boldsymbol{q} = \boldsymbol{q}(\boldsymbol{\delta}) \in \mathbb{R}^{k^*}$ such that

$$q_i := \frac{1}{k^*} \lceil \frac{k^* \delta_i}{\Delta(\theta^*)} \rceil.$$

It is clear that $\mathbf{q} \in H_{k^*} \cap S_{k^*,2}$. It is also clear that $\boldsymbol{\delta} \leq \boldsymbol{q}\Delta(\theta^*)$ element-wise. Due to similar reason, for any $\boldsymbol{t^2} \in S_{k^*,1}$ there exists $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{t}) \in H_{k^*} \cap S_{k^*,2}$ such that $\boldsymbol{t^2} \leq \boldsymbol{w}$ element-wise. Since the collections $\mathcal{M}^4(m_i, n_i, \delta_i, t_i)$ are increasing in (δ_i, t_i) (with respect to set inclusion), it follows from the previous discussion that

$$\begin{aligned} &\mathbb{E} \sup_{\boldsymbol{\delta} \in S_{k^*,\Delta(\theta^*)}} \; \sup_{\boldsymbol{t^2} \in S_{k^*,1}} \sum_{i \in [k^*]} \sup_{\theta_i \in \mathcal{M}^4(m_i, n_i, \delta_i, t_i)} \langle Z, \theta_i \rangle \\ &\leq \mathbb{E} \max_{\Delta(\theta^*)\boldsymbol{\delta}: \boldsymbol{\delta} \in S_{k^*,2} \cap H_{k^*}} \; \max_{\boldsymbol{t^2} \in S_{k^*,2} \cap H_{k^*}} \sum_{i \in [k^*]} \Big[\\ &\sup_{\theta_i \in \mathcal{M}^4(m_i, n_i, \Delta(\theta^*)\delta_i, t_i)} \langle Z, \theta_i \rangle \Big]. \end{aligned} \tag{VIII.6}$$

Since Z is a matrix with i.i.d $N(0,1)$ entries, the first two maximums in the right hand side of the above display can actually be taken outside the expectation upto an additive term. This follows from the well known concentration properties of suprema of gaussian random variables. In particular, we now apply Lemma A.2 (stated in the appendix), true for suprema of gaussians, to obtain for a universal constant C,

$$\begin{aligned} &\mathbb{E} \max_{\Delta(\theta^*)\boldsymbol{\delta}: \boldsymbol{\delta} \in S_{k^*,2} \cap H_{k^*}} \; \max_{\boldsymbol{t^2} \in S_{k^*,2} \cap H_{k^*}} \Big[\\ &\sum_{i \in [k^*]} \sup_{\theta_i \in \mathcal{M}^4(m_i, n_i, \Delta(\theta^*)\delta_i, t_i)} \langle Z, \theta_i \rangle \Big] \\ &\leq \max_{\Delta(\theta^*)\boldsymbol{\delta}: \boldsymbol{\delta} \in S_{k^*,2} \cap H_{k^*}} \; \max_{\boldsymbol{t^2} \in S_{k^*,2} \cap H_{k^*}} \Big[\\ &\sum_{i \in [k^*]} \mathbb{E} \sup_{\theta_i \in \mathcal{M}^4(m_i, n_i, \Delta(\theta^*)\delta_i, t_i)} \langle Z, \theta_i \rangle \Big] + \\ &C\sqrt{\log |H_{k^*} \cap S_{k^*,2}|}. \end{aligned} \tag{VIII.7}$$

To bound the log cardinality $\log |H_{k^*} \cap S_{k^*,2}|$, note that for any positive integer k^*, the cardinality $|H_{k^*} \cap S_{k^*,2}|$ is the same as the number of k^* tuples of positive integers summing up to at most $2k^*$. By standard combinatorics, we have

$$|H_{k^*} \cap S_{k^*,2}| = \sum_{s=k^*}^{2k^*} \binom{s-1}{k^*-1}.$$

Since

$$\frac{\binom{s}{k^*-1}}{\binom{s-1}{k^*-1}} = \frac{s}{s - k^* + 1} \geq \frac{2k^* - 1}{k^*}$$

for all $s \in \{k^*, \ldots, 2k^* - 1\}$, it follows that

$$|H_{k^*} \cap S_{k^*,2}| \leq 3\binom{2k^*-1}{k^*-1} \leq C\mathrm{e}^{Ck^*}$$

for some positive absolute constant C.

Using (VIII.5), (VIII.6), (VIII.7) and the above cardinality bound, we can finally finish the proof by writing

$$\begin{aligned} &\mathcal{GW}(T_{K(V^*)}(\theta^*) \cap B_{n,n}(1)) = \\ &\mathbb{E} \sup_{\theta \in T_{K(V^*)}(\theta^*): \|\theta\| \leq 1} \langle Z, \theta \rangle \\ &\leq \max_{\Delta(\theta^*)\boldsymbol{\delta}: \boldsymbol{\delta} \in S_{k^*,2} \cap H_{k^*}} \; \max_{\boldsymbol{t^2} \in S_{k^*,2} \cap H_{k^*}} \Big[\\ &\sum_{i \in [k^*]} \mathbb{E} \sup_{\theta_i \in \mathcal{M}^4(m_i, n_i, \Delta(\theta^*)\delta_i, t_i)} \langle Z, \theta_i \rangle \Big] \\ &+ \; C\sqrt{k^*}\,. \end{aligned}$$

□

Operationally, the above lemma reduces the task of upper bounding the Gaussian width of $T_{K(V^*)(\theta^*)} \cap B_{n,n}(1)$ to upper bounding the Gaussian width of $\mathcal{M}^4$ with appropriate parameters. However, it would be convenient for us to bound the Gaussian width when the number of boundaries involved in the constraint is at most one instead of four. The results in the next subsection makes this possible.

C. Further Simplification: From Four Boundaries to One

We now proceed to the second step, i.e., reducing the number of boundaries involved in the constraints from four to one (or zero). Thus, we will keep on subdividing each $\theta_{|R_i}$ until we obtain submatrices satisfying constraints similar to (VIII.2), albeit with the ℓ_1-norm of at most one boundary vector appearing on the right hand side of the bound on total variation. This is the content of this subsection.

Taking the cue from the the previous subsection, let us define

$$\mathcal{M}^{\mathrm{top}}(m',n',\delta,t) := \{\theta \in \mathbb{R}^{m'\times n'} : \mathrm{TV}(\theta) \le \|\theta_{\mathrm{top}}\|_1 + \delta,\ \|\theta\| \le t\}.$$

We can define $\mathcal{M}^{\mathrm{bottom}}(m',n',\delta,t)$, $\mathcal{M}^{\mathrm{left}}(m',n',\delta,t)$ and $\mathcal{M}^{\mathrm{right}}(m',n',\delta,t)$ in a similar fashion. Notice that the constraint satisfied by the total variation of the members of $\mathcal{M}^{\mathrm{right}}(m',n',\delta,t)$ is "almost" identical to (VII.3). *By abuse of notation we will refer to any of the four families of matrices described above by a generic notation which is* $\mathcal{M}^1(m',n',\delta,t)$. The reason behind this is that our ultimate concerns would be the Gaussian widths of these families which, for m' and n' close enough to each other, are expected to be of similar order by symmetry. Using a single notation for them would thus minimize the notational clutter. In a similar vein we define

$$\mathcal{M}^0(m',n',\delta,t) := \{\theta \in \mathbb{R}^{m'\times n'} : \mathrm{TV}(\theta) \le \delta,\ \|\theta\| \le t\}.$$

Having defined the relevant families of matrices, we can now state our main result for this subsection.

Lemma VIII.4: Fix positive integers m,n and positive numbers δ,t. Define for each integer $j \ge 1$,

$$\delta^{(j)} := \delta + 16\,(j+1)\ t\,\Big(\sqrt{\frac{m}{n}} + \sqrt{\frac{n}{m}}\Big). \qquad \text{(VIII.8)}$$

Then we have the following bound for a universal constant $C > 0$,

$$\mathcal{GW}(\mathcal{M}^4(m,n,\delta,t)) \le C\,\Big(\sum_{j=1}^{K}\big(\mathcal{GW}(\mathcal{M}^1(\frac{m}{2^j},\frac{n}{2^j},\delta^{(j)},t)) + \mathcal{GW}(\mathcal{M}^0(\frac{m}{2^j},\frac{n}{2^j},\delta^{(j)},t))\Big).$$

Here, to simplify notations, we use $m/2^j$, for $m,j\in\mathbb{N}$, to denote any (but fixed in any given context) integer m' between $m2^{-(j+1)}$ and $m2^{-j}$. The similar definition for n instead of m is denoted by $n/2^j$. K equals the number of binary divisions of $[m]\times[n]$ on both axes that are possible and equals $\min\{\log_2\ m, \log_2\ n\}$ up to a universal constant.

The above lemma bounds the Gaussian width of $\mathcal{M}^4$ in terms of Gaussian widths of simpler classes of matrices $\mathcal{M}^1$ and $\mathcal{M}^0$. We devote the next subsection to its proof.

D. Proof of Lemma VIII.4

We need some intermediate lemmas. We start with the following lemma. The notation convention is same as in Lemma VIII.4.

Lemma VIII.5: There exists a rectangular partition $\mathcal{R}$ of $[m]\times[n]$ with the following property. For any $\theta \in \mathcal{M}^4(m,n,\delta,t)$, there exists non negative real numbers t_R for every rectangle $R \in \mathcal{R}$ such that:

- $\mathcal{R} = \bigcup_{j\in[K],\,k\in[2]} \mathcal{R}_{j,k}$ where $\mathcal{R}_{j,k}$'s are disjoint sets of rectangles and all the rectangles in $\mathcal{R}_{j,k}$ are of size $m_i/2^j \times n_i/2^j$.
- $|\mathcal{R}_{j,1}| \le 8$ and for any $R \in \mathcal{R}_{j,1}$ we have $\theta_{|R} \in \mathcal{M}^1(m/2^j, n/2^j, \delta^{(j)}, t_R)$.
- $|\mathcal{R}_{j,2}| \le 4$ and for any $R \in \mathcal{R}_{j,2}$ we have $\theta_{|R} \in \mathcal{M}^0(m/2^j, n/2^j, \delta^{(j)}, t_R)$.
- $\sum_{R\in\mathcal{R}} t_R^2 = t^2$

Proof of Lemma VIII.4: The proof of Lemma VIII.4 follows directly from Lemma VIII.5 and the sub-additivity of the Gaussian width functional. □

The task now is to prove Lemma VIII.5. The proof of Lemma VIII.5 is divided into two steps where we state and prove two intermediate lemmas. In the first step we reduce the number of "active" boundaries, i.e., the number of boundary vectors involved in the bound on total variation, from four to two and in the second step we reduce them from two to one or zero. The main idea of the proofs is essentially same as that of Lemma VIII.2.

Remark VIII.2: While lemma VIII.5 is true for any integers m,n, the reader can safely read on as if m,n are powers of 2. The essential aspects of the proof of Lemma VIII.5 all go through in this case. Writing the general case would make the notations messy. For the sake of clean exposition, we thus write the entire proof when m and n are powers of 2. At the end, we mention the modifications needed when m,n are not powers of 2.

Four to two boundaries.

In order to state this result let us define for any $\delta > 0$,

$$\mathcal{M}^{\mathrm{topright}}(m,n,\delta,t) := \{\theta \in \mathbb{R}^{m\times n} : \mathrm{TV}(\theta) \le \|\theta_{\mathrm{right}}\|_1 + \|\theta_{\mathrm{top}}\|_1 + \delta,\ \|\theta\| \le t\}.$$

Similarly we can define the families $\mathcal{M}^{\mathrm{topleft}}(\cdots)$, $\mathcal{M}^{\mathrm{bottomleft}}(\cdots)$ and $\mathcal{M}^{\mathrm{bottomright}}(\cdots)$. Likewise $\mathcal{M}^1(m,n,\delta,t)$, we will refer generically to any of these four families of matrices by $\mathcal{M}^2(m,n,\ \delta,t)$. Below we call a partitioning of a matrix $\theta \in \mathbb{R}^{m\times n}$ as an *equal dyadic partitioning* if each submatrix lies in $\mathbb{R}^{m/2\times n/2}$ and is formed by adjacent rows and columns of θ as $\theta^{\mathrm{topleft}}$, $\theta^{\mathrm{topright}}$, $\theta^{\mathrm{bottomleft}}$ and $\theta^{\mathrm{bottomright}}$ in the obvious order.

Lemma VIII.6: Take any $\theta \in \mathcal{M}^4(m,n,\delta,t)$. Let us denote the four submatrices obtained by an equal dyadic partitioning of θ. Then the submatrix θ^{ab}, where $a \in \{\mathrm{top},\mathrm{bottom}\}$ and $b \in \{\mathrm{left},\mathrm{right}\}$, itself satisfies

$$\mathrm{TV}(\theta^{ab}) \le \|\theta^{ab}_a\|_1 + \|\theta^{ab}_b\|_1 + \delta + 16\ t\Big(\sqrt{\frac{m}{n}} + \sqrt{\frac{n}{m}}\Big).$$

In words, if a matrix $\theta \in \mathcal{M}^4(m,n,\delta,t)$ is dyadically partitioned into four equal sized submatrices, each of these four submatrices lies in $\mathcal{M}^2(m/2, n/2, \delta', t)$ where $\delta' := \delta + 16\; t(\sqrt{\frac{m}{n}} + \sqrt{\frac{n}{m}})$; furthermore the boundaries that are active for these submatrices are the ones that they share with θ.

Proof: Since $\|\theta\|_2 \le t$, there exists $1 \le i \le m/2 < i' \le m$ and $1 \le j \le n/2 < j' \le n$ such that

$$\max\{\|\theta[i,]\|\,, \|\theta[i',]\|\} \le \frac{2t}{\sqrt{m}}$$
$$\max\{\|\theta[, j]\|, \|\theta[, j']\|\} \le \frac{2t}{\sqrt{n}}\,.$$

The previous display and the Cauchy-Schwarz inequality together imply

$$\max\{\|\theta[i,]\|_1, \|\theta[i',]\|_1\} \le 2t\sqrt{\frac{n}{m}}$$
$$\max\{\|\theta[, j]\|_1, \|\theta[, j']\|_1\} \le 2t\sqrt{\frac{m}{n}}\,. \qquad \text{(VIII.9)}$$

Now consider the submatrix θ^{topleft} for which we have

$$\text{TV}_{\text{row}}(\theta^{\text{topleft}}) \ge \|\theta^{\text{topleft}}[\,, 1]\|_1 - \|\theta^{\text{topleft}}[\,, j]\|_1 = \|\theta^{\text{topleft}}_{\text{left}}\|_1 - \|\theta[1 : m/2\,, j]\|_1\,,$$

where in the last step we used the fact that $\theta^{\text{topleft}}[\,, j] = \theta[1 : m/2, j]$. A similar argument gives us

$$\text{TV}_{\text{col}}(\theta^{\text{topleft}}) \ge \|\theta^{\text{topleft}}_{\text{top}}\|_1 - \|\theta[i\,, 1 : n/2\,]\|_1\,.$$

Analogous lower bounds for TV_{row} and TV_{col} of the other three submatrices can be derived involving the ℓ_1 norms of appropriate boundaries and (partial) rows or columns of θ. Adding all these together and using (VIII.9), we obtain

$$\sum_{\substack{a\in\{\text{top},\text{bottom}\},\\ b\in\{\text{left},\text{right}\}}} (\text{TV}_{\text{row}}(\theta^{ab}) + \text{TV}_{\text{col}}(\theta^{ab})) \ge \sum_{\substack{a\in\{\text{top},\text{bottom}\},\\ b\in\{\text{left},\text{right}\}}} (\|\theta^{ab}_a\|_1 + \|\theta^{ab}_b\|_1) - 16\; t\big(\sqrt{\frac{m}{n}} + \sqrt{\frac{n}{m}}\big).$$

On the other hand, since $\theta \in \mathcal{M}^4(m, n, \delta, t)$ we have

$$\sum_{\substack{a\in\{\text{top},\text{bottom}\},\\ b\in\{\text{left},\text{right}\}}} (\text{TV}_{\text{row}}(\theta^{ab}) + \text{TV}_{\text{col}}(\theta^{ab})) \le \text{TV}(\theta) \le \sum_{c\in\{\text{top},\text{bottom},\text{left},\text{right}\}} \|\theta_c\|_1 + \delta = \sum_{\substack{a\in\{\text{top},\text{bottom}\},\\ b\in\{\text{left},\text{right}\}}} (\|\theta^{ab}_a\|_1 + \|\theta^{ab}_b\|_1) + \delta\,.$$

An application of Lemma A.1 now finishes the proof of the lemma from the previous two displays. □

Two to one or zero boundary.

Let us start by stating the following lemma which one can think of as a version of Lemma VIII.6 applied to an element of $\mathcal{M}^2(m, n, \delta, t)$. The proof is very similar and we leave it to the reader to verify.

Lemma VIII.7: Let $\theta \in \mathcal{M}^{ab}(m, n, \delta, t)$ for some $a \in \{\text{top}, \text{bottom}\}$ and $b \in \{\text{left}, \text{right}\}$. We can partition θ into equal sized four submatrices $\theta^{\text{topleft}}, \theta^{\text{topright}}, \theta^{\text{bottomleft}}$ and $\theta^{\text{bottomright}}$ in the obvious manner such that the submatrix θ^{cd}, where $c \in \{\text{top, bottom}\}$ and $d \in \{\text{left}, \text{right}\}$, satisfies

$$\text{TV}(\theta^{cd}) \le \|\theta^{cd}_c\|_1 \mathbb{I}\{a = c\} + \|\theta^{cd}_d\|_1 \mathbb{I}\{b = d\} + \delta + 16\, t\big(\sqrt{\frac{m}{n}} + \sqrt{\frac{n}{m}}\big)\,.$$

In words, if a matrix $\theta \in \mathcal{M}^2(m, n, \delta, t)$ is dyadically partitioned into four equal sized submatrices, then each of these four submatrices has at most two active boundaries and a wiggle room of at most $\delta + 16t(\sqrt{\frac{m}{n}} + \sqrt{\frac{n}{m}})$; furthermore the active boundaries are the ones that they share with one of the active boundaries of θ.

We are now ready to conclude the proof of Lemma VIII.5.

Proof of Lemma VIII.5: Recall that we are assuming m, n are powers of 2 for simplicity of exposition.

Step 0: Partition $[m] \times [n]$ dyadically into four equal rectangles so that for any such rectangle S, $\theta_{|S} \in \mathcal{M}^2(m/2, n/2, \delta^{(0)}, \|\theta_{|S}\|)$ by Lemma VIII.6 where

$$\delta^{(0)} = \delta + 16\, t\big(\sqrt{\frac{m}{n}} + \sqrt{\frac{n}{m}}\big)\,.$$

Step 1: Let S (there are four of them) be a generic rectangle obtained from the previous step. Using Lemma VIII.7, we now partition $\theta_{|S}$ into four equal parts (rectangles). We then get two matrices in $\mathcal{M}^1(m/4, n/4, \delta^{(1)}, t)$, one matrix in $\mathcal{M}^0(m/4, n/4, \delta^{(1)}, t)$ and *the remaining one from* $\mathcal{M}^2(m/4, n/4, \delta^{(1)}, t)$. Here,

$$\delta^{(1)} = \delta^{(0)} + 16\, t(\sqrt{\frac{m}{n}} + \sqrt{\frac{n}{m}}) = \delta + 32\, t(\sqrt{\frac{m}{n}} + \sqrt{\frac{n}{m}})\,.$$

Steps $j \ge 2$: From the last step we get exactly one matrix in $\mathcal{M}^2(m/4, n/4, \delta^{(1)}, t)$, for each of the 4 rectangles S. For each S, we now recursively use Lemma VIII.7 by partitioning this matrix again into four exactly equal parts in a dyadic fashion and continue the same procedure with the matrix obtained in each step with two active boundaries until we end up with matrices only with 0 or 1 active boundary. Observe that in the very last step we arrive at a submatrix with exactly one row or column in place of the one with two active boundaries.

For each $j \ge 1$, define $\mathcal{R}_{j,1}$ as the collection of rectangles R obtained in step j such that $\theta_{|R}$ has exactly 1 active boundary. From Lemma VIII.7, we know that there are exactly two such rectangles for any given S (from step 0) and therefore $|\mathcal{R}_{j,1}| \le 8$. For any $j \ge 1$, and any rectangle $R \in \mathcal{R}_{j,1}$, repeated application of Lemma VIII.7 yields that $\theta_{|R} \in \mathcal{M}^1(m/2^j, n/2^j, \delta^{(j)}, \|\theta_{|R}\|)$ where

$$\delta^{(j)} = \delta + 16(j + 1)\, t(\sqrt{\frac{m}{n}} + \sqrt{\frac{n}{m}}).$$

Now defining $\mathcal{R}_{j,2}$ as the collection of rectangles R obtained in step j such that $\theta_{|R}$ has no active boundary, we can deduce in a similar way that $|\mathcal{R}_{j,2}| \le 4$. Also for such rectangles R and $j \ge 1$ we have $\theta_{|R} \in \mathcal{M}^0(m/2^j, n/2^j, \delta^{(j)}, \|\theta_{|R}\|)$. Finally, notice that

$$\sum_{j\ge1} \sum_{R\in\mathcal{R}_{j,1}\cup\mathcal{R}_{j,2}} \|\theta_{|R}\|^2 = \|\theta\|^2 \le t^2\,.$$

Thus the collection of rectangles $\{\mathcal{R}_{j,k} : j \geq 1, k \in [2]\}$ satisfies all the conditions of Lemma VIII.5.

□

Remark VIII.3: For the statement of Lemma VIII.4 to hold, the important thing in the proof of Lemma VIII.5 is that in every step $1 \leq j \leq K$, the aspect ratio of the submatrices does not change significantly. The reader can check that at every step, both the number of rows and columns halve, thus keeping the aspect ratio constant. At every step, the dimensions of the submatrices halve and thus decrease geometrically, while the allowable wiggle room increases additively by the factor (does not change with j) $16\,t(\sqrt{\frac{m}{n}} + \sqrt{\frac{n}{m}})$.

Remark VIII.4: Let us discuss the case when m, n are not necessarily powers of 2 in the proof of Lemma VIII.5. The first step of reducing the number of active boundaries from four to two, by applying Lemma VIII.6, can be carried out in the same way by splitting at the point $\lfloor m/2 \rfloor$ and $\lfloor n/2 \rfloor$. Next, we come to the stage when we are applying Lemma VIII.7 to reduce the number of active boundaries from two to one, on the four submatrices obtained from the previous step. Let us denote the dimensions of these 4 submatrices generically by m', n'. Recall, in the first step of subdivision, we get exactly one submatrix with 2 active boundaries. The others have 1 or 0 active boundaries. **At this step, we can subdivide such that the submatrix with two active boundaries has dimensions which are exactly powers of** 2. For instance, we can split at the unique power of 2 between $m'/4$ and $m'/2$ on one dimension and do the exact same thing for the other dimension. Once we have this submatrix with two active boundaries to have dimensions which are exactly powers of 2, we can carry out the rest of the steps as in the proof of Lemma VIII.5. It can be checked that, in this case, all the inequalities we deduce while proving Lemma VIII.4 goes through with the possible mutiplication of a universal constant.

E. Upper Bounds on Gaussian Widths and the Proof of Proposition II.4

Now that we have reduced the problem of bounding the gaussian width of $T_{K(V^*)(\theta^*)} \cap B_{n,n}(1)$ to that of $\mathcal{M}^0(m, n, \delta, t)$ and $\mathcal{M}^1(m, n, \delta, t)$, we need to obtain upper bounds on these quantities in order to conclude the proof of Theorem II.2. Our next lemma provides an upper bound on the gaussian width of $\mathcal{M}^0(m, n, \delta, t)$ which we henceforth denote as $\mathcal{GW}^0(m, n, \delta, t)$.

Lemma VIII.8: Fix $\delta > 0$ and $t \in (0, 1]$. For positive integers m and n such that $\max\{m/n, n/m\} \leq c$ for some $c > 0$, we have the following upper bound on the Gaussian width:

$$\mathcal{GW}^0(m, n, \delta, t) \leq$$
$$C\big(\log\big(\frac{tn}{\delta} \vee \mathrm{e}\big)\delta(\log \mathrm{e}n)^{3/2} + t(\log \mathrm{e}n)^{1/2}\big)$$

where C is a constant depending only on c.

Proof: Since m and n are of the same order, the bound computed in (V.2) from Section V remains valid in this case.

□

In our next proposition, we provide an upper bound on $\mathcal{GW}^1(m, n, \delta, t)$, i.e., the gaussian width of $\mathcal{M}^1(m, n, \delta, t)$. This is the main result in this subsection and one of the main technical contributions of this paper.

Proposition VIII.9: Fix $\delta \in (0, n]$ and $t \in (0, 1]$. Then for positive integers m, n satisfying the conditions of the previous lemma, we have the following upper bound on the Gaussian width:

$$\mathcal{GW}^1(m, n, \delta, t) \leq C(\log \mathrm{e}n)^{9/2} n^{1/4}\sqrt{(t+\delta)^{2\downarrow}} +$$
$$C\big((\log \mathrm{e}n)^4 t + n^{-9}\big)\,. \qquad \text{(VIII.10)}$$

Here $x^{2\downarrow} := x + x^2$ and $C > 0$ is a constant depending solely on c.

We will prove the above proposition slightly later. Lemma VIII.8 and Proposition VIII.9 together with Lemma VIII.4 now imply (with $\mathcal{GW}^4(m, n, \delta, t)$ denoting the gaussian width of $\mathcal{M}^4(m, n, \delta, t)$)

Lemma VIII.10: Under the same condition as in the previous proposition, we have

$$\mathcal{GW}^4(m, n, \delta, t) \leq C(\log \mathrm{e}n)^{9/2} n^{1/4}(\sqrt{t} + \sqrt{\delta} + \delta)$$
$$+ C\big((\log \mathrm{e}n)^5 t + n^{-9}\log \mathrm{e}n\big)$$

where $C > 0$ is a universal constant.

The proof just involves collecting all the relevant terms and adding them up. The reader can safely skip the proof in the first reading.

Proof: In this proof, we write $a \lesssim b$ to mean $a \leq C\,b$ for some positive constant C — depending at most on the aspect ratio c — whose exact value can change from line to line. Recall that Lemma VIII.4 implies for $K \lesssim \log n$,

$$\mathcal{GW}^4(m, n, \delta, t) \lesssim$$
$$\sum_{j=1}^{K}\big[\mathcal{GW}^1(\frac{m}{2^j}, \frac{n}{2^j}, \delta^{(j)}, t) + \mathcal{GW}^0(\frac{m}{2^j}, \frac{n}{2^j}, \delta^{(j)}, t)\big]\,.$$

First we compute, in view of Proposition VIII.9,

$$\sum_{j=1}^{K}\mathcal{GW}^1(\frac{m}{2^j}, \frac{n}{2^j}, \delta^{(j)}, t) \lesssim$$
$$(\log \mathrm{e}n)^{9/2}\sum_{j=1}^{K}(\frac{n}{2^j})^{1/4}\big(\sqrt{t+\delta^{(j)}} + t + \delta^{(j)}\big)$$
$$+ (\log \mathrm{e}n)^5 t + n^{-9}\log \mathrm{e}n$$
$$\lesssim (\log \mathrm{e}n)^{9/2}\sum_{j=1}^{K}(\frac{n}{2^j})^{1/4}\big(\sqrt{t} + \sqrt{\delta^{(j)}} + t + \delta^{(j)}\big)$$
$$+ (\log \mathrm{e}n)^5 t + n^{-9}\log n \lesssim (\log \mathrm{e}n)^{9/2}$$
$$\sum_{j=1}^{K}(\frac{n}{2^j})^{1/4}\big(\sqrt{t} + \sqrt{\delta + jt} + t + \delta + jt\big)$$
$$+ (\log n)^5 t + n^{-9}\log \mathrm{e}n$$
$$\lesssim (\log \mathrm{e}n)^{9/2} n^{1/4}\big(\sqrt{t} + \sqrt{\delta} + \delta\big)$$
$$+ (\log \mathrm{e}n)^5 t + n^{-9}\log \mathrm{e}n$$

where we have repeatedly used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and in the last inequality we have summed up the geometric series.

On the other hand, Lemma VIII.8 implies

$$\sum_{j=1}^{K} \mathcal{GW}^0\Big(\frac{m}{2^j}, \frac{n}{2^j}, \delta^{(j)}, t\Big) \lesssim$$
$$(\log n)^{5/2} \sum_{j=1}^{K} (\delta + jt) \lesssim \delta (\log \mathrm{e}n)^{7/2} + (\log \mathrm{e}n)^{9/2} t\,.$$

We can now deduce the lemma from the last two displays. □

With the help of the above lemma we can now conclude the proof of Proposition II.4.

Proof of Proposition II.4: Throughout this proof we will use the notation C to denote some positive constant — depending at most on the aspect ratio c like in the previous proof — whose exact value may change from one line to the next. Also we will use "$a \lesssim b$" to mean "$a \lesssim Cb$". Recall that by Lemma VIII.3, $\mathcal{GW}(T_{K(V^*)}(\theta^*) \cap B_{n,n}(1))$ is at most

$$\max_{\Delta(\theta^*)\delta : \delta \in S_{k^*,2} \cap H_{k^*}} \; \max_{\boldsymbol{t^2} \in S_{k^*,2} \cap H_{k^*}} \sum_{i \in [k^*]} \Big[$$
$$\mathcal{GW}^4(m_i, n_i, \Delta(\theta^*)\delta_i, t_i)\Big] + \; C\sqrt{k^*}. \qquad \text{(VIII.11)}$$

Now we plug in the bound from Lemma VIII.10 to obtain a bound on the sum inside the two maximums in the above display:

$$\sum_{i \in [k^*]} \mathcal{GW}^4(m_i, n_i, \Delta(\theta^*)\delta_i, t_i)$$
$$\lesssim (\log n)^{9/2} \sum_{i \in [k^*]} n_i^{1/4}\big(\sqrt{t_i} + (\Delta(\theta^*)\delta_i)^{1/2} + \Delta(\theta^*)\delta_i\big)$$
$$+ (\log n)^5 \sum_{i \in [k^*]} t_i + k^* n^{-9} \log n\,.$$

Since the aspect ratios of each of the rectangular level sets of θ^* are bounded by a constant, we have $\sum_{i=1}^{k^*} n_i{}^2 \lesssim n^2$. This can be seen as follows:

$$\sum_{i=1}^{k^*} n_i^2 \lesssim \sum_{i=1}^{k^*} n_i m_i \lesssim mn \lesssim n^2.$$

Therefore, we can repeatedly apply the Cauchy-Schwarz inequality to deduce for δ, $\boldsymbol{t^2} \in S_{k^*,2}$,

$$\sum_{i=1}^{k^*} n_i^{1/4} \sqrt{t_i} \lesssim (k^*)^{5/8} n^{1/4}\,,$$
$$\sum_{i=1}^{k^*} n_i^{1/4} \delta_i^{1/2} \lesssim (k^*)^{3/8} n^{1/4}\,,$$
$$\sum_{i=1}^{k^*} n_i^{1/4} \delta_i \lesssim n^{1/4} \text{ and } \sum_{i=1}^{k^*} t_i \lesssim \sqrt{k^*}\,.$$

Also because of constant aspect ratio, we have

$$\Delta(\theta^*) = \sqrt{\sum_{i=1}^{k^*} 2\Big(\frac{m_i}{n_i} + \frac{n_i}{m_i}\Big)} \lesssim \sqrt{k^*}\,.$$

Combining the last two displays we notice that $(\log n)^{9/2} (k^*)^{5/8} n^{1/4}$ emerges as the dominant term and hence

$$\sum_{i \in [k^*]} \mathcal{GW}^4(m_i, n_i, \Delta(\theta^*)\delta_i, t_i) \lesssim$$
$$(\log n)^{9/2} (k^*)^{5/8} n^{1/4}\,.$$

Together with (VIII.11) this finishes the proof. □

All that remains towards the proof of Proposition II.4 is Proposition VIII.9. The proof of this proposition is fairly involved. The rest of this section is devoted to its proof.

F. Proof of Proposition VIII.9

By symmetry, it is enough to bound $\mathcal{GW}(\mathcal{M}^{\text{right}}(m,n,\delta,t))$. To this end, let us introduce a new class of matrices as follows:

$$\mathcal{A}(m,n,u,v,t) :=$$
$$\{\theta \in \mathbb{R}^{m \times n} : \mathrm{TV}_{\text{row}}(\theta) \le u, \mathrm{TV}_{\text{col}}(\theta) \le v, \|\theta\| \le t\}$$

where, let us recall, that the total variation $\mathrm{TV}_{\text{row}}(\theta)$ along rows is defined as

$$\mathrm{TV}_{\text{row}}(\theta) := \sum_{i \in [m]} \sum_{j \in [n-1]} |\theta[i, j+1] - \theta[i, j]|$$

and $\mathrm{TV}_{\text{col}}(\theta) := \mathrm{TV}_{\text{row}}(\theta^T)$.

The following lemma gives an upper bound of $\mathcal{GW}(\mathcal{M}^{\text{right}})$ in terms of the Gaussian widths of $\mathcal{A}$ with appropriate parameters.

Lemma VIII.11: Let k denote the smallest integer satisfying $(1 + 2 + \ldots 2^k) \ge n$. Then we have the following inequality:

$$\mathcal{GW}^{\text{right}}(m,n,\delta,t) \le$$
$$\sum_{j \in [k]} \mathcal{GW}(\mathcal{A}(m,\ n_j,\ 2t\sqrt{m/n_j} +\ \delta,\ t\sqrt{m/n} + \delta,\ t))\,,$$

where $n_j = 2^j$ for $j \in [k-1]$ and $n = \sum_{j \in [k]} n_j$.

Proof: The proof proceeds by dividing the n columns into blocks of geometrically increasing length and showing that for any $\theta \in \mathcal{M}^{\text{right}}(m,n,\delta,t)$ the submatrices defined by the blocks live in $\mathcal{A}$ with appropriate parameters. Let $\theta \in \mathcal{M}^{\text{right}}(m,n,\delta,t)$ and subdivide θ into submatrices $[\theta^{(k)}|\theta^{(k-1)}|\cdots|\theta^{(1)}]$ where $\theta^{(j)}$ has n_j many columns. Therefore it suffices to prove that

$$\theta^{(j)} \in \mathcal{A}(m,\ n_j,\ t\sqrt{m/n_{j-1}} +\ \delta,\ t\sqrt{m/n} + \delta,\ t)$$

for all $j \in [k]$ as $\sqrt{n_j/n_{j-1}} < 2$. Since $\|\theta^{(j)}\| \le \|\theta\| \le t$, we only need to verify the required bounds on $\mathrm{TV}_{\text{col}}(\theta^{(j)})$ and $\mathrm{TV}_{\text{row}}(\theta^{(j)})$.

Verifying the bound on $\mathrm{TV}_{\text{col}}(\theta^{(j)})$. We will prove the stronger statement $\mathrm{TV}_{\text{col}}(\theta) \le t\sqrt{m/n} + \delta$. Since $\|\theta\| \le t$ and $\|\theta\|^2 = \sum_{\ell \in [n]} \|\theta[\,,\,\ell]\|^2$, it follows that $\|\theta[\,,\,\ell^*]\| \le t/\sqrt{n}$ for some $\ell^* \in [n]$ and hence $\|\theta[\,,\,\ell^*]\|_1 \le t\sqrt{m/n}$ by the Cauchy-Schwartz inequality. Now using the condition that $\mathrm{TV}(\theta) \le \|\theta[\,,\,n]\|_1 + \delta$ (from the definition of

$\mathcal{M}^{\text{right}}(m, n, \delta, t)$), we get

$$\begin{aligned}&\|\theta[\,,\, n]\|_1 + \delta - \mathrm{TV}_{\mathrm{col}}(\theta) \geq \mathrm{TV}(\theta) - \mathrm{TV}_{\mathrm{col}}(\theta) = \\ &\mathrm{TV}_{\mathrm{row}}(\theta) \geq \|\theta[\,,\, n] - \theta[\,,\, \ell^*]\|_1 \\ &\geq \|\theta[\,,\, n]\|_1 - \|\theta[\,,\, \ell^*]\|_1 \geq \|\theta[\,,\, n]\|_1 - t\sqrt{m/n}\,.\end{aligned} \tag{VIII.12}$$

Thus $\mathrm{TV}_{\mathrm{col}}(\theta) \leq \delta + t\sqrt{m/n}$.

Verifying the bound on $\mathrm{TV}_{\mathrm{row}}(\theta^{(j)})$. Let us start with $\theta^{(1)}$. By the Cauchy-Schwartz inequality, $\|\theta[\,,\, n]\|_1 \leq \sqrt{m}\,\|\theta[\,,\, n]\|_2 \leq t\sqrt{m}$ and thus

$$\mathrm{TV}_{\mathrm{row}}(\theta^{(1)}) \leq \mathrm{TV}(\theta) \leq \|\theta[\,,\, n]\|_1 \leq t\sqrt{m}\,.$$

Next consider $\theta^{(j)}$ for some $j \geq 2$. Since $\left\|\theta^{(j-1)}\right\|_2^2 \leq t$ and it has n_{j-1} columns, there is a column of $\theta^{(j-1)}$ whose ℓ_2-norm is at most $t/\sqrt{n_{j-1}}$. Suppose this column is $\theta[\,,\, a]$. Then a calculation similar to (VIII.12) yields,

$$\begin{aligned}&\|\theta[\,,\, n]\|_1 + \delta \geq \mathrm{TV}(\theta) \geq \mathrm{TV}_{\mathrm{row}}(\theta) \geq \\ &\mathrm{TV}_{\mathrm{row}}([\theta^{(j-1)}|\theta^{(j-2)}|\cdots|\theta^{(1)}]) + \mathrm{TV}_{\mathrm{row}}(\theta^{(j)}) \\ &\geq\; \|\theta[\,,\, n] - \theta[\,,\, a]\|_1 + \mathrm{TV}_{\mathrm{row}}(\theta^{(j)}) \geq \\ &\|\theta[\,,\, n]\|_1 - \|\theta[\,,\, a]\|_1 + \mathrm{TV}_{\mathrm{row}}(\theta^{(j)}).\end{aligned}$$

But this implies, along with the Cauchy-Schwartz inequality, that

$$\begin{aligned}&\mathrm{TV}_{\mathrm{row}}(\theta^{(j)}) \leq \|\theta[\,,\, a]\|_1 + \delta \leq \\ &\sqrt{m}\,\|\theta[\,,\, a]\|_2 + \delta \leq t\sqrt{m/n_j} + \delta\,.\end{aligned}$$ □

It therefore suffices, in view of the previous lemma, to bound the gaussian width of each $\mathcal{A}(m, n_j,\, 2t\sqrt{m/n_j} + \delta,\; t\sqrt{m/n} + \delta, t)$ from above in order to bound $\mathcal{GW}^{\text{right}}(m, n, \delta, t)$. Defining $a = \sqrt{m/n_j}$, we can write

$$\begin{aligned}&\mathcal{A}(m, n_j, 2t\sqrt{m/n_j} +\; \delta,\; t\sqrt{m/n} + \delta, t) = \\ &\mathcal{A}(m, m/a^2, 2ta +\; \delta,\; t\sqrt{m/n} + \delta, t) =: \mathcal{A}_a\,.\end{aligned}$$

Notice that we suppressed the dependence on m, n, δ and t which henceforth refer to the corresponding parameters in Proposition VIII.9.

In our next result, which is crucial for the proof of Proposition VIII.9, we give a subspace cover for the set $\mathcal{A}_a$ corresponding to any distance τ between $1/m$ and 1.

Lemma VIII.12: Let $t \leq 1$, $\tau \in [1/m, 1]$ and $a \geq c$ be such that m/a^2 is a positive integer between 1 and n. Here c is from the statement of Proposition VIII.9. Then there exists a τ subspace cover $\mathcal{S}_\tau$ of $\mathcal{A}$, depending on m, n, a, δ and t in addition to τ, and a constant $C > 0$ depending solely on c such that

$$\begin{aligned}&\max(\log|S_\tau|,\, \max_{S \in S_\tau} \dim(S)) \leq \\ &C(\log(\mathrm{e}m))^3 \mathcal{L}_m\big(1 + \frac{\sqrt{m}C_{m,n,\delta,t}}{\tau^2}\big)\end{aligned}$$

where $\mathcal{L}(x) := x\log(\mathrm{e}\log(\mathrm{e}m)^2\; x)$ and

$$C_{m,n,\delta,t} := \log(\mathrm{e}m)\big(t\sqrt{\frac{m}{n}} + t + \delta\big)^{2\downarrow}$$

(recall that $x^{2\downarrow} := x + x^2$).

Remark VIII.5: Notice that $\mathcal{L}(x)$ is linear in x ignoring the log factors. Thus it is helpful to read the above bound as scaling like $\frac{\sqrt{m}}{\tau^2}$ up to log factors and the lower order terms. This $\sqrt{m}$-scaling is crucial for us in order to derive the $1/4$ exponent of n in Proposition VIII.9 and subsequently the correct exponent of n in Theorem II.2.

Remark VIII.6: The reason for assuming a polynomial lower bound (in m) on τ is that we want $\log(1/\tau)$ to be at most $O(\log m)$. Hence the bounds of Lemma VIII.12 remain valid, with appropriate changes in C, as long as $\tau \geq 1/m^c$ for some universal constant $c > 0$.

With Lemma VIII.12 we can now finish the proof of Proposition VIII.9.

Proof of Proposition VIII.9: An important feature of the bounds in Lemma VIII.12 is that it does not depend on a. Hence an application of Proposition IV.1 would yield the same bound on each Gaussian width appearing inside the summation in the statement of Lemma VIII.11. From this we can deduce Proposition VIII.9 in a straightforward manner. The detailed computation is given below. In the remainder of the proof we will use C to denote any positive constant depending *at most* on c whose exact value may change from one line to the next.

Applying Proposition IV.1 with $k_0 = \lfloor -\log_2 2t \rfloor$ and $k_1 = -\lceil \log_2 \nu \rceil$ where $\nu = t/m \vee m^{-10}$ and using Lemma VIII.12 subsequently to bound the relevant terms (see Remark VIII.5), we get

$$\begin{aligned}&\mathcal{GW}(\mathcal{A}_a) \leq \\ &C\sum_{k=k_0+1}^{k_1} 2^{-k}(\log(\mathrm{e}m))^{1.5}\sqrt{\mathcal{L}_m\big(1 + 2^{2k}\sqrt{m}\,C_{m,n,\delta,t}\big)} \\ &+\; \sqrt{mn}\,\nu\,.\end{aligned}$$

Now recalling the definition of $\mathcal{L}_m(\cdot)$, we can write

$$\begin{aligned}&\mathcal{L}_m\big(1 + 2^{2k}\sqrt{m}C_{m,n,\delta,t}\big) = \\ &\big(1{+}2^{2k}\sqrt{m}C_{m,n,\delta,t}\big)\big(1{+}\log\log(em)^2{+}\log 2^{2k}\sqrt{m}C_{m,n,\delta,t}\big) \\ &\leq \big(1{+}2^{2k}\sqrt{m}C_{m,n,\delta,t}\big)\big(1{+}\log\log(em)^2{+}\log(m^{21}C_{m,n,\delta,t})\big) \\ &\leq C\log(\mathrm{e}m(1+\delta))\,\big(1 + 2^{2k}\sqrt{m}C_{m,n,\delta,t}\big)\end{aligned}$$

where in the last inequality we used the fact that $C_{m,n,\delta,t} \leq C(1+\delta)\log(\mathrm{e}m)$ since $t \leq 1$ and m/n is assumed to be bounded by a constant. The last two displays therefore imply

$$\begin{aligned}&\mathcal{GW}(\mathcal{A}_a) \leq C(\log(\mathrm{e}m))^{1.5}\sqrt{\log(\mathrm{e}m(1+\delta))} \\ &(t + m^{1/4}\log(\mathrm{e}m)\sqrt{C_{m,n,\delta,t}}\,) \\ &+ \sqrt{tn/m} + \sqrt{n}/m^{20} \\ &\leq C(\log\mathrm{e}n)^{3.5}n^{1/4}\sqrt{(t+\delta)^{2\downarrow}} + \\ &C(\log\mathrm{e}n)^3\; t + Cn^{-9.5}\end{aligned}$$

where in the final step we used the fact that $\delta \in (0, n]$ as well as $\max\{m/n, n/m\} \leq c$. The proposition now follows from summing this bound over k as in Lemma VIII.11. □

The thing that remains to be done is the proof of Lemma VIII.12. An important ingredient is the following weaker analogue for the general case.

Lemma VIII.13: Let k, m, n be positive integers with $1 \le k \le m$ (not to be confused with the parameters in Lemmata VIII.11 – VIII.12). Also let $t \le 1$ and $u, v, \tau > 0$. Then there exists a τ subspace cover S_τ of $\mathcal{A}(m, n, u, v, t)$, depending on m, n, k, u, v and t in addition to τ, and a universal constant $C > 0$ such that

$$\max(\log |S_\tau|, \max_{S \in S_\tau} \dim(S)) \le$$
$$C\big(J_k + \sqrt{J_k}\frac{v\sqrt{m}}{\tau\sqrt{k}}\big) \log\big(emJ_k + em\sqrt{J_k}\frac{v\sqrt{m}}{\tau\sqrt{k}}\big)$$

when $k < m$, whereas for $k = m$

$$\max(\log |S_\tau|, \max_{S \in S_\tau} \dim(S)) \le CJ_k \log(emJ_k).$$

Here

$$J_k := C\log(en)\big(k + \frac{u\sqrt{nk}}{\tau}\big).$$

Remark VIII.7: Lemma VIII.13, by itself, is not sufficient to prove Lemma VIII.12. To see this, let us plug in $n = m/a^2$ and $u = 2ta$ in the expression for J_k. One can easily check that while this makes J_k free from a, the principal terms in the bounds on the dimension and cardinality do not attain the required $\sqrt{m}$-scaling for any choice of k.

In the course of proving Lemma VIII.13, we will repeatedly use a subdivision scheme based on the value of either $\mathrm{TV}_{\mathrm{row}}$ or $\mathrm{TV}_{\mathrm{col}}$. We will also use it in the proof of Lemma VIII.12 and therefore describe it here in a general setting. Let us point out that a very similar scheme was described in Section V-A in the context of proving Theorem II.1.

A greedy partitioning scheme: Consider a set $\mathcal{S}$ and a function $T : \cup_{n \in \mathbb{N}} \mathcal{S}^n \mapsto \mathbb{R}_{\ge 0}$ satisfying $T(AB) \ge T(A) + T(B)$ for all $A, B \in \cup_{n \in \mathbb{N}} \mathcal{S}^n$ where AB denotes the concatenation of A and B. Also suppose for any singleton $s \in S$, the function T satisfies $T(s) = 0$. To relate this to a concrete example, the reader may consider the case where $\mathcal{S} = \mathbb{R}^m$ so that $\mathcal{S}^n \equiv \mathbb{R}^{m \times n}$ and T is the function $\mathrm{TV}_{\mathrm{row}}$. Now for any $\epsilon > 0$, the (T, ϵ) *scheme* subdivides an element U of $\cup_{n \in \mathbb{N}} \mathcal{S}^n$ as $U_1 U_2 \cdots U_K$ such that $T(U_i) \le \epsilon$ for all $i \in [K]$. This is achieved in several steps of binary division as follows. In the first step, we check whether $T(U) \le \epsilon$. If so, then stop and output U. Else, divide U as $U_1' U_2'$ into two almost equal parts. This means $|U_1'| = \lfloor |U|/2 \rfloor$ and $|U_2'| = |U'| - |U_1'|$. In each step, we have a representation of U of the form $U_1' U_2' \cdots U_{K'}'$. We consider each $i \in [K']$ such that $T(U_i') > \epsilon$ and subdivide U_i' into two almost equal parts. We repeat this procedure until each part U' in the current representation satisfies $T(U') \le \epsilon$.

Suppose that $|U| = n$. The subdivision of U produced by the (T, ϵ) scheme corresponds to a partition of $[n]$ into contiguous blocks, say, $P_{U;T,\epsilon}$. Let $|P_{U;T,\epsilon}|$ denote the number of blocks of the partition $P_{U;T,\epsilon}$. Now for $t > 0$, let $\mathcal{P}(t, n, \epsilon, T)$ denote the set of partitions $\{P_{U;T,\epsilon} : U \in \mathcal{S}^n, T(C) \le t\}$. A key ingredient in the proof of Lemma VIII.13 (and subsequently Lemma VIII.12) is the following universal upper bound on the cardinality of $P_{U;T,\epsilon}$.

Lemma VIII.14: Then for the (T, ϵ) division scheme we have

$$\max_{P \in \mathcal{P}(t,n,\epsilon,T)} |P_{U;T,\epsilon}| \le \log_2(4n)(1 + \frac{t}{\epsilon}).$$

The proof of Lemma VIII.14 is very similar to that of Lemma V.4. Nevertheless, for the sake of completeness, we provide its proof in the appendix (see Section A-B). We also defer the proof of Lemma VIII.13 to the end of this subsection and finish the proof of Lemma VIII.12 assuming it.

Proof of Lemma VIII.12: Take any $\theta \in \mathcal{A}_a$ and fix $\epsilon \in (0, 1)$ whose precise value based on τ would be chosen later. Let us denote the $m' \times n'$ two dimensional grid (graph) by $L_{m',n'}$ and subdivide $L_{m,m/a^2}$ as

$$L_{m,m/a^2} = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_K \end{bmatrix} \tag{VIII.13}$$

where $\mathrm{TV}_{\mathrm{col}}(\theta_{|R_i}) \le \epsilon$ for all $i \in [K]$ and $K \le \log_2(4m)(1 + \mathrm{TV}_{\mathrm{col}}(\theta)\epsilon^{-1})$. We achieve this by applying the $(\mathrm{TV}_{\mathrm{col}}, \epsilon)$ division scheme to the rows of θ (see Lemma VIII.14). Denoting the set of all possible partitions of $L_{m,m/a^2}$ obtained in this manner by $\mathcal{P}$, we deduce

$$|\mathcal{P}| \le m^{\log_2(4m)(1+(t\sqrt{\frac{m}{n}}+\delta)\epsilon^{-1})}. \tag{VIII.14}$$

Corresponding to the partition $P = P(\theta)$ in (VIII.13), let $S_{P,\mathrm{row}}$ denote the linear subspace of $\mathbb{R}^{m \times m/a^2}$ comprising only matrices having identical rows in each R_i. It is clear that the orthogonal projection of θ onto $S_{P,\mathrm{row}}$ is given by

$$\hat{\theta}_{S_{P,\mathrm{row}}} = \tilde{\theta} = \begin{bmatrix} \tilde{\theta}_1 \\ \tilde{\theta}_2 \\ \vdots \\ \tilde{\theta}_K \end{bmatrix}$$

where each row of $\tilde{\theta}_i := \tilde{\theta}_{|R_i}$ is equal to the average row of θ_{R_i}. By repeated application of Lemma A.3 (stated and proved in the appendix), we obtain

$$\mathrm{dist}(\theta, S_{P,\mathrm{row}}) = \|\theta - \tilde{\theta}\|_2 \le \sqrt{m}\epsilon. \tag{VIII.15}$$

Also by standard properties of orthogonal projections, it follows that $\|\tilde{\theta}\|_2 \le \|\theta\|_2 \le t$. We further claim that $\tilde{\theta} \in \mathcal{A}_a \equiv \mathcal{A}(m, \frac{m}{a^2}, 2\ ta +\ \delta, t\sqrt{\frac{m}{n}} + \delta, t)$. Hence to establish this claim we only need to show that $\mathrm{TV}_{\mathrm{row}}(\tilde{\theta}) \le \mathrm{TV}_{\mathrm{row}}(\theta)$ and $\mathrm{TV}_{\mathrm{col}}(\tilde{\theta}) \le \mathrm{TV}_{\mathrm{col}}(\theta)$. We can obtain the first inequality as follows:

$$\begin{aligned} &\mathrm{TV}_{\mathrm{row}}(\tilde{\theta}) = \\ &\sum_{i \in [K]} \mathrm{n}_{\mathrm{row}}(R_i) \sum_{\ell \in [\mathrm{n}_{\mathrm{col}}(R_i)-1]} |\tilde{\theta}_i[1, \ell+1] - \tilde{\theta}_i[1, \ell]| \\ &= \sum_{i \in [K]} \mathrm{n}_{\mathrm{row}}(R_i) \sum_{\ell \in [\mathrm{n}_{\mathrm{col}}(R_i)-1]} \Big| \\ &\mathrm{n}_{\mathrm{row}}(R_i)^{-1} \sum_{i' \in [\mathrm{n}_{\mathrm{row}}(R_i)]} \big(\tilde{\theta}_i[i', \ell+1] - \tilde{\theta}_i[i', \ell]\big)\Big| \end{aligned}$$

$$\leq \sum_{i\in[K]} \sum_{\ell\in[\mathrm{n}_{\mathrm{col}}(R_i)-1]} \sum_{i'\in[\mathrm{n}_{\mathrm{row}}(R_i)]} |\tilde{\theta}_i[i', \ell+1] - \tilde{\theta}_i[i', \ell]|$$
$$= \mathrm{TV}_{\mathrm{row}}(\tilde{\theta})\,. \tag{VIII.16}$$

For the second inequality we just apply Lemma A.4 (stated and proved in the appendix section) to each column of θ.

In the rest of the article we call a subset of $L_{m'\times n'}$ a *subgrid* if it is a product of subsets (as opposed to only subintervals) of $[1, m']\cap\mathbb{N}$ and $[1, n']\cap\mathbb{N}$ respectively. We will now regroup R_i's into several subgrids. For any positive integer ℓ such that $2^\ell \leq 2m$, define the set $S_\ell := \{i \in [K] : 2^{\ell-1} \leq \mathrm{n}_{\mathrm{row}}(R_i) < 2^\ell\}$ and let B_ℓ be the vector which is the sorted version of S_ℓ. Now consider the *subgrid* of $L_{m,m/a^2}$

$$R^\ell := \begin{bmatrix} R_{B_\ell(1)} \\ R_{B_\ell(2)} \\ \vdots \\ R_{B_\ell(K_\ell)} \end{bmatrix}$$

where $K_\ell := |B_\ell|$. In words, R^ℓ comprises the rectangles R_i, in order, whose number of rows lies between $2^{\ell-1}$ and 2^ℓ. It is clear that $R^1, R^2, \ldots, R^L$ are disjoint subgrids of $L_{m,n}$ where $L \leq \log_2(2\ m)$. Let us also denote $\tilde{\theta}_{|R^\ell}$ by $\tilde{\theta}^\ell$. Notice that if the matrices $\hat{\theta}^1, \hat{\theta}^2, \ldots, \hat{\theta}^L$ satisfy $\|\tilde{\theta}^\ell - \hat{\theta}^\ell\| \leq \sqrt{m}\epsilon$ for all $\ell \in [L]$ and $\hat{\theta} \in \mathbb{R}^{m\times m/a^2}$ is such that $\hat{\theta}_{|R^\ell} = \hat{\theta}^\ell$ for all $\ell \in [L]$, then we have

$$\|\theta - \hat{\theta}\| \leq \|\theta - \tilde{\theta}\| + \|\tilde{\theta} - \hat{\theta}\| \overset{\text{(VIII.15)}}{\leq}$$
$$\sqrt{m\epsilon^2} + \sqrt{m\log_2(2m)\epsilon^2} \leq \sqrt{2\,m\log_2(4m)}\,\epsilon\,. \tag{VIII.17}$$

We now choose ϵ by requiring this approximation error to be τ, i.e., by setting $\epsilon = \tau/\sqrt{2\,m\log_2(4m)}$ (notice that $1/4m^2 \leq \epsilon \leq 1/\sqrt{m}$ when $\tau \in [1/m, 1]$). Therefore if $\mathcal{S}^\ell_{\tau,P}$ is a $\sqrt{m}\epsilon$ subspace cover for the family $\mathcal{A}^*_{\ell,P}$ (say) of matrices $\tilde{\theta}^\ell$ corresponding to $P \in \mathcal{P}$ and $\ell \in [L]$, we can immediately obtain a τ subspace cover S_τ for $\mathcal{A}_a$ satisfying:

$$\max_{S\in S_\tau} \dim(S) \leq \max_{P\in\mathcal{P}} \sum_{\ell\in[L]} \max_{S\in\mathcal{S}^\ell_{\tau,P}} \dim(S) \tag{VIII.18}$$

and

$$|\mathcal{S}_\tau| \leq |\mathcal{P}|\,.\max_{P\in\mathcal{P}} \prod_{\ell\in[L]} |\mathcal{S}^\ell_{\tau,P}|. \tag{VIII.19}$$

Now fix a $P \in \mathcal{P}$ and let Θ^ℓ denote the matrix formed by the first (or any) rows of $\tilde{\theta}_{B_\ell(1)}, \tilde{\theta}_{B_\ell(2)}, \ldots, \tilde{\theta}_{B_\ell(k_\ell)}$ in order, i.e., the rows of $\tilde{\theta}^\ell$ that are *potentially* distinct. We claim that

$$\Theta^\ell \in \mathcal{A}\big(K_\ell,\, \frac{m}{a^2},\, \frac{2\ ta + \delta}{2^{\ell-1}},\, t\sqrt{\frac{m}{n}} + \delta,\, t\big)$$
$$=: \mathcal{A}_{a,\ell}\,(= \mathcal{A}_{\ell,P})\,. \tag{VIII.20}$$

The constraints on the number of rows and columns of Θ^ℓ as well as $\|\Theta^\ell\|$ are clear. For the remaining constraints first observe that $\tilde{\theta}^\ell \in \mathcal{A}(\mathrm{n}_{\mathrm{row}}(R^\ell), \frac{m}{a^2}, 2\ ta +\ \delta,\ t\sqrt{\frac{m}{n}} + \delta, t)$ (the only non-obvious part is the bound on $\mathrm{TV}_{\mathrm{col}}(\tilde{\theta}^\ell)$ which follows from the triangle inequality). From the definition of Θ^ℓ it is immediate that

$$\mathrm{TV}_{\mathrm{col}}(\Theta^\ell) = \mathrm{TV}_{\mathrm{col}}(\tilde{\theta}^\ell) \text{ and}$$
$$\mathrm{TV}_{\mathrm{row}}(\Theta^\ell) \leq \frac{\mathrm{TV}_{\mathrm{row}}(\tilde{\theta}^\ell)}{\min_{i\in[K_\ell]} \mathrm{n}_{\mathrm{row}}(R_{B_\ell(i)})}\,.$$

Therefore the bounds on $\mathrm{TV}_{\mathrm{col}}(\Theta^\ell)$ and $\mathrm{TV}_{\mathrm{row}}(\Theta^\ell)$ follow from the similar bounds for $\tilde{\theta}^\ell$ and the fact that $\mathrm{n}_{\mathrm{row}}(R_{B_\ell(i)}) \geq 2^{\ell-1}$ for each $i \in [K_\ell]$.

Further notice that since $\mathrm{n}_{\mathrm{row}}(R_{B_\ell(i)}) < 2^\ell$ for each $i \in [K_\ell]$, we have $\|\tilde{\theta}^\ell - \hat{\theta}^\ell\| \leq 2^{\ell/2}\|\Theta^\ell - \hat{\Theta}^\ell\|$ where $\hat{\theta}^\ell$ comprises repetitions of the rows of $\hat{\Theta}^\ell$ in the same way as $\tilde{\theta}^\ell$ comprises repetitions of the rows of $\tilde{\Theta}^\ell$. Therefore any $2^{-\ell/2}\sqrt{m}\epsilon$ subspace cover $\mathcal{S}^\ell_\epsilon$ for $\mathcal{A}_{a,\ell}$ induces a $\sqrt{m}\epsilon$ subspace cover $\mathcal{S}^\ell_{\tau,P}$ for $\mathcal{A}^*_{\ell,P}$. Our next claim is about a *uniform* upper bound on $\max_{S\in\mathcal{S}^\ell_\epsilon} \dim(S)$ and $|\mathcal{S}^\ell_\epsilon|$ for some particular choice of $\mathcal{S}^\ell_\epsilon$ and hence that of $\max_{S\in\mathcal{S}^\ell_{\tau,P}} \dim(S)$ and $|\mathcal{S}^\ell_{\tau,P}|$ as well.

Claim VIII.2: There is a choice of $\mathcal{S}^\ell_\epsilon$ for any $\ell \in \mathbb{N}_{>0}$ and $\epsilon \in [1/m^2, 1/\sqrt{m}]$ such that for some universal constant $C > 0$,

$$\max(\log|\mathcal{S}^\ell_\epsilon|,\ \max_{S\in\mathcal{S}^\ell_\epsilon} \dim(S)) \leq$$
$$C\log(\mathrm{e}m)^2\mathcal{L}_m\big(1 + \frac{1}{\sqrt{m}\epsilon^2}\big(t\sqrt{\frac{m}{n}} + t + \delta\big)^{2\downarrow}\big)$$

where we recall from the statement of Lemma VIII.12 that $\mathcal{L}(x) = x\log(\mathrm{e}\log(\mathrm{e}m)^2\ x)$ and $x^{2\downarrow} = x + x^2$.

Claim VIII.2 follows directly from Lemma VIII.13 when we choose k in an *appropriate* manner. The complete proof is given after the current proof.

Concluding the proof. In the remainder of the proof we will use C to denote any positive, universal constant whose exact value may change from one line to the next. Using Claim VIII.2 let us first bound

$$\max_{P\in\mathcal{P}} \sum_{\ell\leq\log_2 2\ m} \max(\log|\mathcal{S}^\ell_{\tau,P}|,\ \max_{S\in\mathcal{S}^\ell_{\tau,P}} \dim(S)) \leq$$
$$C(\log(\mathrm{e}m))^3\mathcal{L}_m\big(1 + \frac{1}{\sqrt{m}\epsilon^2}\big(t\sqrt{\frac{m}{n}} + t + \delta\big)^{2\downarrow}\big)$$
$$\leq C(\log(\mathrm{e}m))^3\mathcal{L}_m\big(1 + \frac{\sqrt{m}\log(\mathrm{e}m)}{\tau^2}\big(t\sqrt{\frac{m}{n}} + t + \delta\big)^{2\downarrow}\big)$$
$$= C(\log(\mathrm{e}m))^3\mathcal{L}_m\big(1 + \frac{\sqrt{m}C_{m,n,\delta,t}}{\tau^2}\big)$$

where we used the fact that $\frac{1}{\sqrt{m}\epsilon^2} = \frac{\sqrt{m}\log_2(4m)}{\tau^2}$ (recall the choice of ϵ after (VIII.17) and also the definition of $C_{m,n,\delta,t}$ from the statement of Lemma VIII.12). On the other hand, since $\epsilon \leq 1/\sqrt{m}$, we can bound $\log|\mathcal{P}|$ in view of (VIII.14) as

$$\log_2(4m)(1 +\ (t\sqrt{\tfrac{m}{n}} + \delta)\frac{1}{\epsilon})\log m \leq$$
$$\log_2(4m)(1 +\ (t\sqrt{\tfrac{m}{n}} + \delta)\frac{1}{\sqrt{m}\epsilon^2})\log m$$
$$\leq C\log(\mathrm{e}m)^2\big(1 + \frac{\sqrt{m}C_{m,n,\delta,t}}{\tau^2}\big)\,.$$

Since $\mathcal{L}(x) \ge x$ for all $x \ge 1$, we deduce by combining the previous two displays and subsequently plugging them into (VIII.18)–(VIII.19):

$$\max(\log|\mathcal{S}_\tau|, \max_{S\in\mathcal{S}_\tau}\dim(S)) \le C(\log(\mathrm{e}m))^3\mathcal{L}_m\big(1+\frac{\sqrt{m}C_{m,n,\delta,t}}{\tau^2}\big)\,. \qquad \square$$

Proof of Claim VIII.2: The "main" contribution in the bounds on $\log|\mathcal{S}_\epsilon^\ell|$ and $\max_{S\in\mathcal{S}_\epsilon^\ell}\dim(S)$ given by Lemma VIII.13 comes from

$$J^*_{k,\ell} := \big(J_{k,\ell} + \sqrt{J_{k,\ell}}\frac{(t\sqrt{\frac{m}{n}}+\delta)\sqrt{K_\ell}}{2^{-\ell/2}\sqrt{m}\epsilon\sqrt{k}}\mathbb{I}\{k<K_\ell\}\big)$$

where

$$J_{k,\ell} = C\log(\mathrm{e}m/a^2)\big(k + \frac{2^{-\ell}(2ta+\delta)\sqrt{mk}}{a2^{-\ell/2}\sqrt{m}\epsilon}\big) \overset{a\ge c}{\le} C\log(\mathrm{e}m)\big(k+\frac{2^{-\ell}(2t+\delta)\sqrt{mk}}{2^{-\ell/2}\sqrt{m}\epsilon}\big) \qquad \text{(VIII.21)}$$

(recall the statement of Lemma VIII.13 and (VIII.20)). Therefore, as already mentioned in the proof of Lemma VIII.12, we will apply Lemma VIII.13 for some $k \in [K_\ell]$ so that $J^*_{k,\ell}$ has a small value. In the rest of the proof we will use C to denote an unspecified but universal positive constant whose value may change from one instant to the next. Using the simple fact $\sqrt{x+y} \le \sqrt{x}+\sqrt{y}$, we can bound $J^*_{k,\ell}$ as follows:

$$J^*_{k,\ell} \le J_{k,\ell} + \mathbb{I}\{k<K_\ell\}\,C_{\epsilon,k,\ell}\,, \qquad \text{(VIII.22)}$$

where

$$C_{\epsilon,k,\ell} := C\sqrt{\log(\mathrm{e}m)}(t\sqrt{\frac{m}{n}}+\delta)\,\sqrt{K_\ell}\,\big(\frac{1}{2^{-\ell/2}\sqrt{m}\epsilon} + \frac{2^{-\ell/2}\sqrt{(2t+\delta)}m^{1/4}}{(2^{-\ell/2}\sqrt{m}\epsilon)^{3/2}k^{1/4}}\big)\,. \qquad \text{(VIII.23)}$$

Now let us consider two cases separately based on whether $\sqrt{K_\ell}2^{-\ell/2}\sqrt{m}\epsilon$ is smaller or larger than 1. Recall that $2^{-\ell/2}\sqrt{m}\epsilon$ is the covering radius in question, and the condition above is equivalent to K_ℓ being smaller or larger than the inverse of the covering radius squared.

Case 1: $K_\ell \le \frac{2^\ell}{m\epsilon^2}$. In this case we choose $k = K_\ell$ so that Lemma VIII.13 and (VIII.21) together give us

$$\max(\log|\mathcal{S}^\ell_\epsilon|, \max_{S\in\mathcal{S}^\ell_\epsilon}\dim(S)) \le CJ_{k,\ell}\log(\mathrm{e}K_\ell J_{k,\ell}) \qquad \text{(VIII.24)}$$

where

$$J_{k,\ell}\le C\log(\mathrm{e}m)\big(K_\ell + \frac{2^{-\ell}(2t+\delta)\sqrt{mK_\ell}}{2^{-\ell/2}\sqrt{m}\epsilon}\big)\,. \qquad \text{(VIII.25)}$$

Now using

$$K_\ell \le K \le C\log(\mathrm{e}m)\big(1+(t\sqrt{\frac{m}{n}}+\delta)\epsilon^{-1}\big) \qquad \text{(VIII.26)}$$

for the first term inside the parenthesis in (VIII.25) (recall the definition of K_ℓ and K from the proof of Lemma VIII.12) and using $K_\ell \le \frac{2^\ell}{m\epsilon^2}$ for the second, we get

$$J_{k,\ell}\le C(\log(\mathrm{e}m))^2 + C(\log(\mathrm{e}m))^2\big(t\sqrt{\frac{m}{n}}+t+\delta\big)\big(\frac{1}{\epsilon}+\frac{1}{\sqrt{m}\epsilon^2}\big)\,.$$

Further noticing that $\epsilon\le 1/\sqrt{m}$, so that $\frac{1}{\epsilon}\le\frac{1}{\sqrt{m}\epsilon^2}$, we obtain

$$J_{k,\ell}\le C(\log(\mathrm{e}m))^2\Big(1+\frac{1}{\sqrt{m}\epsilon^2}\big(t\sqrt{\frac{m}{n}}+t+\delta\big)^{2\downarrow}\Big) \qquad \text{(VIII.27)}$$

(recall that $x^{2\downarrow} = x+x^2$). On the other hand we have $K_\ell = k \le J_{k,\ell}$ for $C > 1$. Plugging these bounds into the right hand side of (VIII.24) and rewriting the expression in terms of $\mathcal{L}_m(x) = x\log(\mathrm{e}\log(\mathrm{e}m)^2\ x)$ we obtain

$$\max(\log|\mathcal{S}^\ell_\epsilon|, \max_{S\in\mathcal{S}^\ell_\epsilon}\dim(S)) \le C(\log(\mathrm{e}m))^2\mathcal{L}_m\big(1+\frac{1}{\sqrt{m}\epsilon^2}\big(t\sqrt{\frac{m}{n}}+t+\delta\big)^{2\downarrow}\big)\,. \qquad \text{(VIII.28)}$$

where we used the fact that $\log(C\mathrm{e}\log(\mathrm{e}m)^2\ x) \le C\log(\mathrm{e}\log(\mathrm{e}m)^2\ x)$ for all $x\ge1$ and large enough C.

Case 2: $K_\ell \ge \frac{2^\ell}{m\epsilon^2}$. Notice that in this case we can choose $k = \lfloor\frac{2^\ell}{m\epsilon^2}\rfloor$ and Lemma VIII.13 gives us

$$\max(\log|\mathcal{S}^\ell_\epsilon|, \max_{S\in\mathcal{S}^\ell_\epsilon}\dim(S)) \le CJ^*_{k,\ell}\log(\mathrm{e}K_\ell J^*_{k,\ell})\,. \qquad \text{(VIII.29)}$$

We will show below that the right hand side of (VIII.27) also serves as an upper bound for $J^*_{k,\ell}$ and K_ℓ, and consequently the upper bound in (VIII.28) holds in this case as well, thus proving the claim. To this end we will use the bounds (VIII.22) and (VIII.23). First observe that the bound on $J_{k,\ell}$ is same as in the previous case since the only bounds we used there were $k \le K_\ell$ and $k \le \frac{2^\ell}{m\epsilon^2}$, both of which are valid in this case. On the other hand, $C_{\epsilon,k,\ell}$ can be bounded by

$$C\sqrt{\log(\mathrm{e}m)}(t\sqrt{\frac{m}{n}}+\delta)\,\big(\frac{2^{\ell/2}\sqrt{K_\ell}}{\sqrt{m}\epsilon}+\frac{\sqrt{(2t+\delta)K_\ell}}{m^{1/4}\epsilon}\big)\,. \qquad \text{(VIII.30)}$$

Since $K_\ell \ge \frac{2^\ell}{m\epsilon^2}$ and $\epsilon\le 1/\sqrt{m}$, we have

$$\begin{aligned}\frac{\sqrt{K_\ell}}{2^{-\ell/2}\sqrt{m}\epsilon}\le K_\ell &\overset{\text{(VIII.26)}}{\le} C\log(\mathrm{e}m)\big(1+\,(t\sqrt{\frac{m}{n}}+\delta)\frac{\sqrt{m}\epsilon}{\sqrt{m}\epsilon^2}\big)\\ &\le C\log(\mathrm{e}m)+C\log(\mathrm{e}m)\,(t\sqrt{\frac{m}{n}}+\delta)\frac{1}{\sqrt{m}\epsilon^2}\end{aligned}$$

(cf. the right hand side of (VIII.27)). Similarly we can bound

$$\begin{aligned}\frac{\sqrt{(2t+\delta)K_\ell}}{m^{1/4}\epsilon} &\le C\sqrt{\log(\mathrm{e}m)}\sqrt{t+\delta}\,\big(\frac{1}{m^{1/4}\epsilon}+\frac{\sqrt{t\sqrt{\frac{m}{n}}+\delta}}{m^{1/4}\epsilon^{3/2}}\big)\\ &\le C\sqrt{\log(\mathrm{e}m)}\sqrt{t+\delta}\,\Big(1+\sqrt{t\sqrt{\frac{m}{n}}+\delta}\Big)\frac{1}{m^{1/4}\epsilon^{3/2}}\\ &= C\sqrt{\log(\mathrm{e}m)}\sqrt{t+\delta}\,\Big(1+\sqrt{t\sqrt{\frac{m}{n}}+\delta}\Big)\frac{\sqrt{\sqrt{m}\epsilon}}{\sqrt{m}\epsilon^2}\\ &\le C\sqrt{\log(\mathrm{e}m)}\sqrt{t+\delta}\,\Big(1+\sqrt{t\sqrt{\frac{m}{n}}+\delta}\Big)\frac{1}{\sqrt{m}\epsilon^2}\,.\end{aligned}$$

Plugging these bounds into the (VIII.30) we get

$$C_{\epsilon,k,\ell}\le C(\log(\mathrm{e}m))^2\Big(1+\frac{1}{\sqrt{m}\epsilon^2}\big(t\sqrt{\frac{m}{n}}+t+\delta\big)^{2\downarrow}\Big)\,.$$

where used the simple fact that $x^{3/2} \leq x^{2\downarrow}$. Combined with (VIII.27) and the discussion preceding the display (VIII.30), this yields us a similar upper bound for $J^*_{k,\ell}$. □

We are only left with the proof of Lemma VIII.13.

Proof of Lemma VIII.13: The proof is split into two parts. In the first part we try to construct, for any given $\theta \in \mathcal{A}(m,n,u,v,t)$, another matrix $\hat{\theta}$ satisfying $\|\theta - \hat{\theta}\| \leq \tau$ such that $\hat{\theta}$ is piecewise constant on rectangles with as few blocks as possible. These blocks define a partition P of $L_{m,n}$ and let $\mathcal{P}$ denote the set of all such partitions. It is then clear that $S_\tau := \{S_P : P \in \mathcal{P}\}$ forms a τ subspace cover of $\mathcal{A}(m,n,u,v,t)$ (see the proof of Theorem II.1 in Section V for the notation and similar notions). In the second and the final part we bound $\max_{P \in \mathcal{P}} |P|$ and $|\mathcal{P}|$ which, in view of the definition above, yield the desired upper bounds on $\max_{S \in \mathcal{S}_\tau} \dim(S)$ and $|\mathcal{S}_\tau|$.

Approximating θ by a piecewise constant matrix. This part consists of three steps. In the "zeroth" step, we divide θ equally into k submatrices by *horizontal divisions*. We do not choose, a priori, any specific value of k which is the reason why our final bound depends on k. Then in step 1, each of these submatrices is divided into submatrices by *vertical divisions* which are again subdivided in step 2 by horizontal divisions. The rectangles corresponding to these submatrices will be the final level sets of $\hat{\theta}$. We now elaborate the steps.

Step 0: Horizontal Divisions. Fix a positive integer $1 \leq k \leq m$ and divide $L_{m,n}$ into k submatrices as follows:

$$L_{m,n} = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_k \end{bmatrix}$$

where each R_i has either $\lceil m/k \rceil$ or $\lfloor m/k \rfloor$ many rows. We want to stress that we use the same partitioning for every θ in this step.

Step 1: Vertical Divisions. Next we want to subdivide each R_i (where $i \in [k]$) by making j_i many vertical divisions:

$$R_i = [R_{i,1} | R_{i,2} | \ldots | R_{i,j_i}]$$

such that $\mathrm{TV}_{\mathrm{row}}(\theta_{|R_{i,j}}) \leq \tau_k$ for all $j \in [j_i]$ and some $\tau_k > 0$ to be chosen shortly. We can do this by the $(\mathrm{TV}_{\mathrm{row}}, \tau_k)$ scheme applied to the columns of θ_i so that Lemma VIII.14 gives us the bounds

$$j_i \leq \log_2(4n)\big(1 + \frac{\mathrm{TV}_{\mathrm{row}}(\theta_i)}{\tau_k}\big)\,. \tag{VIII.31}$$

Replacing each element in every row of $\theta_{i,j} := \theta_{|R_{i,j}}$ with the corresponding row mean, we then obtain a new matrix

$$\tilde{\theta}_i = [\tilde{\theta}_{i,1} | \tilde{\theta}_{i,2} | \ldots | \tilde{\theta}_{i,j_i}]\,.$$

By construction, each $\tilde{\theta}_{i,j}$ has identical columns. Finally, let us define

$$\tilde{\theta} = \begin{bmatrix} \tilde{\theta}_1 \\ \tilde{\theta}_2 \\ \vdots \\ \tilde{\theta}_k \end{bmatrix}$$

From the Cauchy-Schwarz inequality, it is clear that $\|\tilde{\theta}\| \leq \|\theta\|$. One important observation we need make at this point is that while this averaging procedure might increase the value of $\mathrm{TV}_{\mathrm{col}}(\tilde{\theta})$, it does not increase the value of $\mathrm{TV}_{\mathrm{col}}(\tilde{\theta}_{i,j})$ for any i and j. Indeed by a computation exactly similar to that performed in (VIII.16) we get

$$\mathrm{TV}_{\mathrm{col}}(\tilde{\theta}_{i,j}) \leq \mathrm{TV}_{\mathrm{col}}(\theta_{i,j})\,. \tag{VIII.32}$$

Let us now try to bound $\|\theta - \tilde{\theta}\|$. To this end notice that

$$\begin{aligned} &\|\theta - \tilde{\theta}\|_2^2 \\ &\quad = \sum_{i \in [k], j \in [j_i]} \sum_{i' \in [\mathrm{n}_{\mathrm{row}}(R_i)]} \|\theta_{i,j}[i',\,] - \tilde{\theta}_{i,j}[i',\,]\|_2^2 \\ &\quad \leq \sum_{i \in [k], j \in [j_i]} \mathrm{n}_{\mathrm{col}}(R_{i,j}) \sum_{i' \in [\mathrm{n}_{\mathrm{row}}(R_i)]} \mathrm{TV}(\theta_{i,j}[i',\,])^2 \end{aligned} \tag{VIII.33}$$

where in the final step we used Lemma A.3. Since $\mathrm{TV}_{\mathrm{row}}(\theta_{i,j}) \leq \tau_k$, we can then deduce

$$\begin{aligned} &\|\theta - \tilde{\theta}\|_2^2 \\ &\quad \leq \sum_{i \in [k], j \in [j_i]} \mathrm{n}_{\mathrm{col}}(R_{i,j}) \big(\sum_{i' \in [\mathrm{n}_{\mathrm{row}}(\theta_i)]} \mathrm{TV}(\theta_{i,j}[i',\,]) \big)^2 \\ &\quad \leq \sum_{i \in [k], j \in [j_i]} \mathrm{n}_{\mathrm{col}}(R_{i,j}) \tau_k^2 = nk\tau_k^2\,. \end{aligned} \tag{VIII.34}$$

Setting $\tau_k = \tau/2\sqrt{nk}$, we get $\|\theta - \tilde{\theta}\|_2 \leq \tau/2$.

Step 2: Horizontal Divisions. In this step, we are going to make horizontal divisions within each $R_{i,j}$ obtained from step 1 so that the total variation of columns of $\tilde{\theta}_{i,j}$ restricted to each subdivision is smaller than some fixed, small number. To this end fix $\tau'_k > 0$ whose exact value will be chosen later. Now use the $(\mathrm{TV}_{\mathrm{col}}, \tau'_k)$ scheme applied to the rows of $R_{i,j}$ to obtain the following subdivision:

$$R_{i,j} = \begin{bmatrix} R_{i,1;j} \\ R_{i,2;j} \\ \vdots \\ R_{i,\ell_{i,j};j} \end{bmatrix}$$

where, with $\tilde{\theta}_{i,\ell;j} := \tilde{\theta}_{|R_{i,\ell;j}}$, $\mathrm{TV}_{\mathrm{col}}[\tilde{\theta}_{i,\ell;j}] \leq \tau'_k$ for all $\ell \in [\ell_{i,j}]$. From Lemma VIII.14 we can deduce

$$\ell_{i,j} \leq \log_2(4m)\big(1 + \frac{\mathrm{TV}_{\mathrm{col}}(\tilde{\theta}_{i,j})}{\tau'_k}\big)\,. \tag{VIII.35}$$

Like in the definition of $\tilde{\theta}_{i,j}$, we now replace every element in each column of $\tilde{\theta}_{i,\ell;j}$ (recall at this point that $\tilde{\theta}_{i,j}$ and hence $\tilde{\theta}_{i,\ell;j}$ has identical columns) with the corresponding column mean and obtain a new matrix

$$\hat{\theta}_{i,j} = \begin{bmatrix} \hat{\theta}_{i,1;j} \\ \hat{\theta}_{i,2;j} \\ \vdots \\ \hat{\theta}_{i,\ell_{i,j};j} \end{bmatrix}$$

By construction, $\hat{\theta}_{i,\ell;j}$ is a constant matrix. Let $\hat{\theta} \in \mathbb{R}^{m \times n}$ be such that $\hat{\theta}_{|R_{i,\ell;j}} = \hat{\theta}_{i,\ell;j}$. By the Cauchy-Schwarz inequality we have $\|\hat{\theta}\| \leq \|\tilde{\theta}\| \leq \|\theta\|$.

We now want to bound the distance between $\tilde{\theta}$ and $\hat{\theta}$. Notice that, since the columns of $\tilde{\theta}_{i,\ell;j}$ are identical, we get from Lemma A.3

$$\|\tilde{\theta}_{i,\ell;j}[\,,j'] - \hat{\theta}_{i,\ell;j}[\,,j']\|_2^2 \le \mathrm{n}_{\mathrm{row}}(\tilde{\theta}_{i,\ell;j})(\tau'_k/\mathrm{n}_{\mathrm{col}}(\tilde{\theta}_{i,j}))^2$$

for every $j' \in \mathrm{n}_{\mathrm{col}}(\tilde{\theta}_{i,j}) = \mathrm{n}_{\mathrm{col}}(\tilde{\theta}_{i,\ell;j})$. Summing over i, j, ℓ and j', we then deduce

$$\begin{aligned}\|\tilde{\theta} - \hat{\theta}\|_2^2 &\le \sum_{i\in[k],j\in[j_i],\ell\in[\ell_{i,j}]} \frac{\mathrm{n}_{\mathrm{row}}(R_{i,\ell;j})}{\mathrm{n}_{\mathrm{col}}(R_{i,j})}\tau_k'^2 \\ &= \tau_k'^2 \sum_{i\in[k],j\in[j_i]} \frac{\mathrm{n}_{\mathrm{row}}(R_{i,j})}{\mathrm{n}_{\mathrm{col}}(R_{i,j})} \\ &\le \frac{2\tau_k'^2 m}{k} \sum_{i\in[k],j\in[j_i]} \frac{1}{\mathrm{n}_{\mathrm{col}}(R_{i,j})}\,.\end{aligned}$$

Let us choose

$$\tau_k'^{\,2} = \frac{\tau^2 k}{8m \sum\limits_{i\in[k],j\in[j_i]} \frac{1}{\mathrm{n}_{\mathrm{col}}(R_{i,j})}}\,, \tag{VIII.36}$$

so that $\|\tilde{\theta} - \hat{\theta}\|_2 \le \tau/2$ and hence

$$\|\theta - \hat{\theta}\|_2 \le \|\theta - \tilde{\theta}\|_2 + \|\tilde{\theta} - \hat{\theta}\|_2 \le \tau/2 + \tau/2 = \tau\,.$$

Counting the number of possible partitions for any θ. Fix any vertical division of θ obtained in step 1. Now summing (VIII.35) over all i and j we get

$$\sum_{i\in[k],j\in[j_i]} \ell_{i,j} \le \log_2(4m)\,(\sum_{i\in[k]} j_i + v/\tau'_k) \tag{VIII.37}$$

where we used the following fact

$$\begin{aligned}\sum_{i\in[k],j\in[j_i]} \mathrm{TV}_{\mathrm{col}}(\tilde{\theta}_{i,j}) &\overset{\text{(VIII.32)}}{\le} \sum_{i\in[k],j\in[j_i]} \mathrm{TV}_{\mathrm{col}}(\theta_{i,j}) \\ &\le \mathrm{TV}_{\mathrm{col}}(\theta) \le v\,.\end{aligned}$$

On the other hand (VIII.36) allows us to deduce a naive lower bound on τ'_k as follows:

$$\tau'_k \ge \frac{\tau\sqrt{k}}{4\sqrt{2m}\sqrt{\sum_{i\in[k]} j_i}}\,.$$

Plugging this into (VIII.37) we get for a universal constant $C > 0$,

$$\begin{aligned}n_{piece}(\hat{\theta}) &:= \sum_{i\in[k],j\in[j_i]} \ell_{i,j} \\ &\le C\log(\mathrm{e}m)\big(J + \sqrt{J}\frac{v\sqrt{m}}{\tau\sqrt{k}}\big)\end{aligned} \tag{VIII.38}$$

where $n_{piece}(\hat{\theta})$ is the total number of rectangular level sets of $\hat{\theta}$ and $J := \sum_{i\in[k]} j_i$. From now onwards we will implicitly assume that C is a positive, universal constant whose exact value may vary from one line to the next.

Therefore the number of tuples $(\ell_{1,1}, \ell_{1,2}, \ldots, \ell_{k,j_k})$ satisfying (VIII.38) is at most

$$(C\log(em))^J\big(J + \sqrt{J}\frac{v\sqrt{m}}{\tau\sqrt{k}}\big)^J\,. \tag{VIII.39}$$

Similarly, in order to bound J we sum (VIII.31) over all i to obtain

$$\begin{aligned}J = \sum_{i\in[k]} j_i &\le \log_2(4n)\big(k + \frac{1}{\tau_k}\sum_{i\in[k]} \mathrm{TV}_{\mathrm{row}}(\theta_i)\big) \\ &= \log_2(4n)\big(k + \frac{1}{\tau_k}\mathrm{TV}_{\mathrm{row}}(\theta)\big) \\ &\le C\log(en)\big(k + \frac{u}{\tau_k}\big) \\ &\le C\log(en)\big(k + \frac{u\sqrt{nk}}{\tau}\big) =: J_k\,,\end{aligned} \tag{VIII.40}$$

where in the final step we used $\tau_k = \tau/2\sqrt{nk}$ (see the end of step 1 in the previous part).

It remains to count the number of possible vertical divisions in step 1. To this end let us fix a tuple $(j_1, j_2, \ldots, j_k)$ satisfying $\sum_{i\in[k]} j_i \le J_k$. The number of possible vertical divisions in this case is bounded by $\prod_{i\in[k]} n^{j_i} = n^{J_k}$. On the other hand, in view of (VIII.31) and (VIII.40), the number of tuples $(j_1, j_2, \ldots, j_k)$ is bounded by the number of nonnegative integral solutions to the inequality $\sum_{i\in[k]} j_i \le J_k$ which in turn is bounded by $(J_k)^k$. Putting all of these together with (VIII.39) and (VIII.40), we can now deduce the following upper bound on the total number of possible partitions for any $\theta \in \mathcal{A}(m,n,u,v,t)$:

$$(J_k)^k n^{J_k}(C\log(em))^{J_k}\big(J_k + \sqrt{J_k}\frac{v\sqrt{m}}{\tau\sqrt{k}}\big)^{J_k}\,. \tag{VIII.41}$$

From this and (VIII.38) we can derive the bound for any $1 \le k < m$. For the second bound, that is when $k = m$, recall that the second summand in the right hand side of (VIII.37) comes from the horizontal division conducted in step 2. Since this step becomes void for $k = m$, the required bound follows in exactly similar fashion with J_k replacing $J_k + \sqrt{J_k}\frac{v\sqrt{m}}{\tau\sqrt{k}}$. □

IX. Proof of Theorem II.7

To prove Theorem II.7 we apply the general machinery developed in [8] with suitable modifications. Let us define $w = y - \overline{y}$ to be the centered data matrix, $w^* = \theta^* - \overline{\theta^*}$ to be the centered ground truth matrix and let

$$\hat{w} := \operatorname*{argmin}_{v:\,\overline{v}=0,\,\|w-v\|^2\le(n^2-1)\hat{\sigma}^2} \mathrm{TV}(v)\,. \tag{IX.1}$$

Also, for any $V \ge 0$, let $\hat{w}_V$ denote the Euclidean projection of w onto the convex set $K_n^0(V)$. Recall that $K_n^0(V) := \{\theta \in \mathbb{R}^{n\times n} : \mathrm{TV}(\theta) \le V,\, \overline{\theta} = 0\}$.

A. Sketch of Proof

To show that $\hat{\theta}_{\mathrm{notuning}}$ is a good estimator of θ^* it clearly suffices to show that $\hat{w}$ is a good estimator of w^*. If we knew $\mathrm{TV}(\theta^*) = \mathrm{TV}(w^*) = V^*$, a similar argument as in the proof of Theorem II.1 would tell us that $\hat{w}_{V^*}$ attains the $\tilde{O}(\frac{V^*}{\sqrt{N}})$ rate that we desire. Of course, the aim here is to get the same rate without knowing V^* and σ. One part of our proof deals with showing that using $\hat{\sigma}$ in the definition of our estimator is not much worse than if we knew σ and used it in defining our estimator. This is shown by showing that $\hat{\sigma} \approx \sigma$ using a concentration of measure argument where

"$\approx$" is a somewhat informal notation conveying the meaning of approximately equal to.

To analyze the risk of $\hat{w}$, a natural first step is to decompose the risk as follows:

$$\|\hat{w} - w^*\|^2 \leq 2\|\hat{w}_{V^*} - w^*\|^2 + 2\|\hat{w} - \hat{w}_{V^*}\|^2 .$$

Here we used the elementary inequality $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. The above decomposition has a natural interpretation as twice the sum of the ideal risk (achievable when V^* is known) and an excess risk due to not knowing V^* and σ. The main task therefore is to upper bound the excess risk term $\|\hat{w} - \hat{w}_{V^*}\|^2$.

We now need to look at two different cases. The first case is when $\hat{w} \neq \mathbf{0}$. In this case we first show that the minimum of the optimization problem defined in (IX.1) is attained on the boundary. This would mean we have $\|\hat{w} - w\|^2 = (n^2 - 1)\hat{\sigma}^2 \approx (n^2 - 1)\sigma^2$. Letting $\hat{V} = \mathrm{TV}(\hat{w})$, a simple geometric argument also shows that $\hat{w}_{\hat{V}} = \hat{w}$. Thus, both $\hat{w}_{V^*}$ and $\hat{w}$ are Euclidean projections onto $K_n^0(V)$ for two possibly different choices of V. Thus, we can now use standard characterizations of Euclidean projections onto convex sets (content of Lemma IX.1) for both $\hat{w}_{V^*}$ and $\hat{w}$ to obtain a bound on the excess risk as follows:

$$\|\hat{w} - \hat{w}_{V^*}\|^2 \leq \left|\|\hat{w} - w\|^2 - \|w - \hat{w}_{V^*}\|^2\right| .$$

Since $\|\hat{w} - w\|^2 \approx (n^2 - 1)\sigma^2$ we can then conclude

$$\left|\|\hat{w} - w\|^2 - \|w - \hat{w}_{V^*}\|^2\right| \approx \left|(n^2 - 1)\sigma^2 - \|w - \hat{w}_{V^*}\|^2\right| .$$

Further, since $\hat{w}_{V^*}$ is known to be a good estimator of w^* we can write

$$\begin{aligned}\|w - \hat{w}_{V^*}\|^2 &\approx \|w - w^*\|^2 = \|Z - \overline{Z}\mathbf{1}\|^2\sigma^2 \\ &\approx (n^2 - 1)\sigma^2 .\end{aligned}$$

where the last approximation is again by a simple concentration of measure argument. The last three displays then suggest that $\hat{w}$ is close to $\hat{w}_{V^*}$. Quantifying the last three displays gives us the desired upper bound on the excess risk.

The second case is when $\hat{w} = \mathbf{0}$. By definition we have $\|\hat{w}\|^2 \leq (n^2 - 1)\hat{\sigma}^2 \approx (n^2 - 1)\sigma^2$. Since $\mathbf{0} \in K_n^0(V^*)$ and $\hat{w}_{V^*}$ is the projection of w onto $K_n^0(V^*)$, a standard fact about Euclidean projections onto convex sets gives $\langle w - \hat{w}_{V^*}, \hat{w}_{V^*}\rangle \geq 0$. This implies

$$\begin{aligned}\|\hat{w} - \hat{w}_{V^*}\|^2 = \|\hat{w}_{V^*}\|^2 &\leq \|w\|^2 - \|w - \hat{w}_{V^*}\|^2 \\ &\lesssim (n^2 - 1)\sigma^2 - \|w - \hat{w}_{V^*}\|^2 .\end{aligned}$$

The rest of the proof then follows similarly as in the previous case.

B. Full Proof

While proving Theorem II.7 we will prove a few intermediate results. Our first lemma is a basic fact about Euclidean projections onto $K_n^0(V)$ for two different choices of V. This also appears as Lemma 5.1 in [8]. For the sake of completeness, we give a proof below.

Lemma IX.1: Let $y \in \mathbb{R}^{n \times n}$ and recall $K_n^0(V) := \{\theta \in \mathbb{R}^{n \times n} : \mathrm{TV}(\theta) \leq V,\ \overline{\theta} = 0\}$. Let $V_1 > V_2 \geq 0$ and let $\pi_1(y), \pi_2(y)$ be the Euclidean projection of y onto the convex sets $K_n^0(V_1), K_n^0(V_2)$ respectively. Then we have the following inequality:

$$\|\pi_1(y) - \pi_2(y)\|^2 \leq \|y - \pi_2(y)\|^2 - \|y - \pi_1(y)\|^2.$$

Proof: Since $\pi_2(y) \in K_n^0(V_1)$ by definition, the standard KKT condition for projections onto convex sets implies $\langle y - \pi_1(y), \pi_2(y) - \pi_1(y)\rangle \leq 0$. Therefore we can write

$$\begin{aligned}\|y - \pi_2(y)\|^2 &= \|y - \pi_1(y)\|^2 + \|\pi_1(y) - \pi_2(y)\|^2 \\ &\quad + 2\langle y - \pi_1(y), \pi_1(y) - \pi_2(y)\rangle \\ &\geq \|y - \pi_1(y)\|^2 + \|\pi_1(y) - \pi_2(y)\|^2.\end{aligned}$$

This finishes the proof of the lemma. □

Our next lemma is the following pointwise inequality.

Lemma IX.2: Let $w = y - \overline{y}\mathbf{1}$ be the centered version of y. For any $V \geq 0$, let $\hat{w}_V$ denote the projection of w onto the convex set $K_n^0(V)$. Let

$$\hat{w} = \operatorname*{argmin}_{v:\,\overline{v}=0,\ \|w-v\|^2 \leq (n^2-1)\hat{\sigma}^2} \mathrm{TV}(v). \tag{IX.2}$$

Then we have the following pointwise inequality;

$$\|\hat{w} - \hat{w}_{V^*}\|^2 \leq |(n^2 - 1)\hat{\sigma}^2 - \|w - \hat{w}_{V^*}\|^2|.$$

Proof: Let us first consider the case when $\hat{w} \neq \mathbf{0}$. Define $\hat{V} := \mathrm{TV}(\hat{w})$. We claim that $\hat{w}_{\hat{V}} = \hat{w}$ and further

$$\|w - \hat{w}\|^2 = (n^2 - 1)\hat{\sigma}^2. \tag{IX.3}$$

To prove the above claim, suppose $\hat{w}_{\hat{V}} \neq \hat{w}$. Then we have $\|w - \hat{w}_{\hat{V}}\|^2 < \|w - \hat{w}\|^2 \leq (n^2 - 1)\hat{\sigma}^2$ because of uniqueness of Euclidean projections onto convex sets. Therefore, we have $\|w - \hat{w}_{\hat{V}}\|^2 < (n^2 - 1)\hat{\sigma}^2$ and $\|w - \mathbf{0}\|^2 > (n^2 - 1)\hat{\sigma}^2$ by assumption. Let us now draw a line segment connecting $\hat{w}_{\hat{V}}$ to the origin and select the point which cuts the boundary of the $\sqrt{(n^2 - 1)}\,\hat{\sigma}$ ball around w and call it w^{bdry}. Then by construction we have

$$\mathrm{TV}(w^{\mathrm{bdry}}) < \mathrm{TV}(\hat{w}_{\hat{V}}) \leq \mathrm{TV}(\hat{w}). \tag{IX.4}$$

Since w has zero mean, it is not hard to see that $\hat{w}_{\hat{V}}$ has mean zero as well because $\hat{w}_{\hat{V}}$ is the Euclidean projection of w onto $K_n^0(\hat{V})$. Therefore any point falling on the line segment between $\hat{w}_{\hat{V}}$ and the origin also must have mean zero, including w^{bdry}. Thus w^{bdry} is feasible for the optimization problem defined in (IX.2). Together with (IX.4) this contradicts the definition of $\hat{w}$. Therefore $\hat{w}_{\hat{V}}$ must be equal to $\hat{w}$ and (IX.3) must hold.

Letting $V^* = \mathrm{TV}(\theta^*)$, we can now write

$$\begin{aligned}\|\hat{w} - \hat{w}_{V^*}\|^2 &= \|\hat{w}_{\hat{V}} - \hat{w}_{V^*}\|^2 \\ &\leq |\|w - \hat{w}_{\hat{V}}\|^2 - \|w - \hat{w}_{V^*}\|^2| \\ &= |(n^2 - 1)\hat{\sigma}^2 - \|w - \hat{w}_{V^*}\|^2|\end{aligned}$$

where we have applied Lemma IX.1 in the first inequality and used (IX.3) in the last equality.

Now let us consider the case when $\hat{w} = \mathbf{0}$. In this case we can write

$$\begin{aligned}\|\hat{w} - \hat{w}_{V^*}\|^2 &= \|\hat{w}_0 - \hat{w}_{V^*}\|^2 \\ &\leq |\|w\|^2 - \|w - \hat{w}_{V^*}\|^2| \\ &= \|w\|^2 - \|w - \hat{w}_{V^*}\|^2 \\ &\leq (n^2 - 1)\hat{\sigma}^2 - \|w - \hat{w}_{V^*}\|^2.\end{aligned}$$

The first inequality uses Lemma IX.1 and the second equality follows from the definition of $\hat{w}_{V^*}$ upon observing that $\mathbf{0} \in K_n^0(V^*)$. Finally the third inequality uses the fact that $\|w\|^2 \leq (n^2-1)\hat{\sigma}^2$ since $\hat{w} = \mathbf{0}$. This finishes the proof of the lemma. □

Our next result is a proposition which gives a pointwise upper bound to the squared loss.

Proposition IX.3: Let $V^* = \mathrm{TV}(\theta^*)$. Let $w = y - \overline{y}\mathbf{1}$ and $w^* = \theta^* - \overline{\theta^*}\mathbf{1}$ be the centered versions of y and θ^* respectively. Also let $\hat{w}_{V^*}$ denote the Euclidean projection of w onto $K_n^0(V^*)$. Then the following pointwise risk inequality holds:

$$\begin{aligned}\|\hat{\theta} - \theta^*\|^2 \leq\ & 8\,\sigma \sup_{v \in K_n^0(2V^*)} \langle Z, v\rangle \ +\ |\overline{y} - \overline{\theta^*}|^2\ n^2 \\ &+ 2|\|w - w^*\|^2 - (n^2-1)\sigma^2| \\ &+ 2(n^2-1)\ |\hat{\sigma}^2 - \sigma^2|.\end{aligned}$$

Proof: By definition of $\hat{\theta}$ and Pythagorean theorem we have

$$\begin{aligned}\|\hat{\theta} - \theta^*\|^2 &= \|\overline{y}\,\mathbf{1} - \overline{\theta^*}\,\mathbf{1}\|^2 + \|\hat{w} - w^*\|^2 \\ &\leq \|\overline{y}\mathbf{1} - \overline{\theta^*}\mathbf{1}\|^2 \\ &\quad + 2\|\hat{w} - \hat{w}_{V^*}\|^2 + 2\|\hat{w}_{V^*} - w^*\|^2. \qquad \text{(IX.5)}\end{aligned}$$

We can now use Lemma IX.2 and the triangle inequality to write

$$\begin{aligned}\|\hat{w} - \hat{w}_{V^*}\|^2 &\leq |(n^2-1)\hat{\sigma}^2 - \|w - \hat{w}_{V^*}\|^2| \\ &\leq (n^2-1)\ |\hat{\sigma}^2 - \sigma^2| \\ &\quad + |\ \|w - w^*\|^2 - (n^2-1)\sigma^2| \\ &\quad + |\|w - \hat{w}_{V^*}\|^2 - \|w - w^*\|^2| \qquad \text{(IX.6)}\end{aligned}$$

Let us now bound the third term above on the right side.

$$\begin{aligned}&|\|w - \hat{w}_{V^*}\|^2 - \|w - w^*\|^2| \\ &= |\|w^* - \hat{w}_{V^*}\|^2 + 2\langle w - w^*, w^* - \hat{w}_{V^*}\rangle| \\ &\leq \|w^* - \hat{w}_{V^*}\|^2 + 2 \sup_{v \in K_n^0(2V^*)} \langle w - w^*, v\rangle.\end{aligned}$$

We now observe that for any mean zero matrix v, we can write

$$\begin{aligned}\langle w - w^*, v\rangle &= \langle y - \theta^* - (\overline{y} - \overline{\theta^*})\mathbf{1}, v\rangle = \langle y - \theta^*, v\rangle \\ &= \sigma\ \langle Z, v\rangle.\end{aligned}$$

The last two displays then imply that

$$\begin{aligned}&|\|w - \hat{w}_{V^*}\|^2 - \|w - w^*\|^2| \\ &\leq \|w^* - \hat{w}_{V^*}\|^2 + 2\,\sigma \sup_{v \in K_n^0(2V^*)} \langle Z, v\rangle. \qquad \text{(IX.7)}\end{aligned}$$

Further, from the basic inequality $\|w - \hat{w}_{V^*}\|^2 \leq \|w - w^*\|^2$ we can conclude

$$\begin{aligned}\|w^* - \hat{w}_{V^*}\|^2 &\leq 2\langle \hat{w}_{V^*} - w^*, w - w^*\rangle \\ &= 2\langle \hat{w}_{V^*} - w^*, y - \theta^*\rangle \leq 2\,\sigma \sup_{v \in K_n^0(2V^*)} \langle Z, v\rangle.\end{aligned}$$

The last display along with (IX.5), (IX.6) and (IX.7) finish the proof of the proposition. □

We are now in a position to finally prove Theorem II.7.

Proof of Theorem II.7: It suffices to take expectation over the four terms which consists in the upper bound given in Proposition IX.3. We now sequentially bound the expectation of these terms. We will use C to denote a positive, universal constant whose exact value may change from one line to the next.

The first term is just 8σ times the Gaussian width of $K_n^0(2V^*)$ and we can use V.2 to upper bound it. As for the second term, it is clear that

$$n^2\mathbb{E}(\overline{y} - \overline{\theta^*})^2 = n^2\mathrm{Var}(\overline{y}) = \sigma^2.$$

Also we observe that $\frac{\|w - w^*\|^2}{\sigma^2} = \sum_{i=1}^n \sum_{j=1}^n (Z_{ij} - \overline{Z})^2 \approx \chi^2_{n^2-1}$. This is a standard fact about standard normal random variables. Therefore we can write

$$\begin{aligned}&\mathbb{E}|\|w - w^*\|^2 - (n^2-1)\sigma^2| \\ &\leq \big(\mathbb{E}|\|w - w^*\|^2 - (n^2-1)\sigma^2|^2\big)^{1/2} \\ &\leq \sigma^2\big(\mathrm{Var}(\chi^2_{n^2-1})\big)^{1/2} = \sigma^2\sqrt{2\,(n^2-1)} \leq \sqrt{2}\,\sigma^2\ n\end{aligned}$$

where the first inequality follows from the Cauchy Schwartz inequality and the last equality follows because $\mathrm{Var}(\chi^2_k)) = 2k$ for any positive integer k.

Next we bound $\mathbb{E}|\hat{\sigma}^2 - \sigma^2|$. We can write

$$|\hat{\sigma}^2 - \sigma^2| \leq |\hat{\sigma} - \sigma|^2 + 2\sigma|\hat{\sigma} - \sigma|\,. \qquad \text{(IX.8)}$$

Recalling the definition of $\hat{\sigma}$ we have

$$\begin{aligned}|\hat{\sigma} - \sigma| &= |\frac{\mathrm{TV}(\theta^* + \sigma Z) - \sigma\mathbb{E}\mathrm{TV}(Z)}{\mathbb{E}\mathrm{TV}(Z)}| \\ &\leq \frac{\mathrm{TV}(\theta^*)}{\mathbb{E}\mathrm{TV}(Z)} + \sigma\frac{|\mathrm{TV}(Z) - \mathbb{E}\mathrm{TV}(Z)|}{\mathbb{E}\mathrm{TV}(Z)}\,.\end{aligned}$$

Thus we can write

$$\begin{aligned}&|\hat{\sigma} - \sigma|^2 \\ &\leq 2\Big(\frac{V^*}{\mathbb{E}\mathrm{TV}(Z)}\Big)^2 + 2\sigma^2\Big(\frac{|\mathrm{TV}(Z) - \mathbb{E}\mathrm{TV}(Z)|}{\mathbb{E}\mathrm{TV}(Z)}\Big)^2. \qquad \text{(IX.9)}\end{aligned}$$

Now, since $\mathrm{TV}(Z)$ is a sum of $N(0,2)$ random variables it is easy to check that $\mathbb{E}\mathrm{TV}(Z) = \frac{4\,n\,(n-1)}{\sqrt{\pi}}$. Also by Lemma IX.4 we can upper bound the variance of $\mathrm{TV}(Z)$ to get

$$\mathrm{Var}(\mathrm{TV}(\mathbf{Z})) \leq Cn(n-1)\,.$$

Taking expectation on both sides of (IX.9) we obtain

$$\begin{aligned}\mathbb{E}|\hat{\sigma} - \sigma|^2 &\leq 2\Big(\frac{V^*\sqrt{\pi}}{4\,n\,(n-1)}\Big)^2 + 2\sigma^2\frac{C\,\pi\,n(n-1)}{(4\,n\,(n-1))^2} \\ &\leq C\Big(\frac{(V^*)^2}{n^4} + \frac{\sigma^2}{n^2}\Big)\,.\end{aligned}$$

Using (IX.8), the last display and the Cauchy-Schwarz inequality to bound $\mathbb{E}|\hat{\sigma} - \sigma|$, we can deduce

$$\begin{aligned}\mathbb{E}|\hat{\sigma}^2 - \sigma^2| &\leq \mathbb{E}|\hat{\sigma} - \sigma|^2 + 2\sigma\big(\mathbb{E}|\hat{\sigma} - \sigma|^2\big)^{1/2} \\ &\leq C\big(\frac{(V^*)^2}{n^4} + \frac{\sigma^2}{n^2}\big) + C\sigma\big(\frac{V^*}{n^2} + \frac{\sigma}{n}\big)\,.\end{aligned}$$

Collecting the bounds we have obtained in this proof for the four terms comprising the upper bound given in Proposition IX.3, we can conclude that

$$\begin{aligned}&\mathrm{MSE}(\hat{\theta}_{\mathrm{notuning}}, \theta^*) \\ &\leq C\big(\sigma\frac{V^*}{N}\log(en)\log(2 + 2V^* n^2) \\ &\quad + \big(\frac{V^*}{N}\big)^2 + \frac{\sigma^2}{\sqrt{N}} + \frac{\sigma^2}{N}\big)\,.\end{aligned}$$

This finishes the proof of Theorem II.7. □

It only remains to prove the following lemma.

Lemma IX.4: There exists a universal constant $C > 0$ such that

$$\mathrm{Var}(\mathrm{TV}(\mathbf{Z})) \leq Cn(n-1)\,.$$

Proof: Expanding $\mathrm{Var}(\mathrm{TV}(\mathbf{Z}))$ we get

$$\begin{aligned}\mathrm{Var}(\mathrm{TV}(\mathbf{Z})) &= \sum_{e, e' \in E_n} \mathrm{Cov}(|\Delta_e \mathbf{Z}|, |\Delta_{e'}\mathbf{Z}|) \\ &= \sum_{e \in E_n} \sum_{e' \in E_n, e' \sim e} \mathrm{Cov}(|\Delta_e \mathbf{Z}|, |\Delta_{e'}\mathbf{Z}|)\end{aligned} \tag{IX.10}$$

where in the second step we used the observation that $\mathrm{Cov}(|\Delta_e \mathbf{Z}|, |\Delta_{e'}\mathbf{Z}|) = 0$ for all non-adjacent e, e', i.e., e, e' which do not share any vertex. Here $e' \sim e$ means the edges e, e' are adjacent. Since each edge e is adjacent to finitely many edges (including e itself) we get from (IX.10) that $\mathrm{Var}(\mathrm{TV}(\mathbf{Z})) \leq C|E_n|$ for some universal constant $C > 0$. The lemma now follows by noting that $|E_n| = 2n(n-1)$. □

Appendix

A. Some Auxiliary Results

Lemma A.1: Suppose $\{f_i, g_i, h_i\}_{i=1}^n$ are non negative real numbers satisfying the following inequality for each $i \in [n]$,

$$f_i \geq g_i - h_i.$$

Let $\{w_i\}_{i=1}^m$ be some other non negative numbers. In addition, also suppose the following inequality holds for some $\delta > 0$,

$$\sum_{i=1}^n f_i + \sum_{i=1}^m w_i \leq \sum_{i=1}^n g_i + \delta.$$

Then the following is true:

$$\sum_{i=1}^n (f_i - g_i)_+ + \sum_{i=1}^m w_i \leq \delta + \sum_{i=1}^n h_i\,,$$

where $a_+ = \max\{a, 0\}$ for any $a \in \mathbb{R}$.

Proof: The first equation in the above proposition basically says $(f_i - g_i)_- \leq h_i$ for $i \in [n]$ where $a_- = (-a)_+$ for any $a \in \mathbb{R}$. Therefore we can write

$$\begin{aligned}\delta &\geq \sum_{i=1}^n (f_i - g_i) + \sum_{i=1}^m w_i \\ &= \sum_{i=1}^n (f_i - g_i)_+ - \sum_{i=1}^n (f_i - g_i)_- + \sum_{i=1}^m w_i \\ &\geq \sum_{i=1}^n (f_i - g_i)_+ - \sum_{i=1}^n h_i + \sum_{i=1}^m w_i\end{aligned}$$

which finishes the proof of the lemma. □

We state the following lemma which appears as Lemma $D.1$ in [17].

Lemma A.2 (Guntuboyina et al.): Suppose $p, n \geq 1$ and let $\Theta_1, \ldots, \Theta_p$ be subsets of $\mathbb{R}^n$ each containing the origin and contained in the closed Euclidean ball of radius $D > 0$ centered at the origin. Then for $Z \sim N(0, \sigma^2\ I)$ we have

$$\begin{aligned}&\mathbb{E}\big(\max_{i \in [p]} \sup_{\theta \in \Theta_i} \langle Z, \theta\rangle\big) \\ &\quad \leq \max_{i \in [p]} \mathbb{E}\big(\sup_{\theta \in \Theta_i} \langle Z, \theta\rangle\big) + D\sigma\big(\sqrt{2\log p} + \sqrt{\frac{\pi}{2}}\big).\end{aligned}$$

Recall that for a vector $v \in \mathbb{R}^n$ we define

$$\mathrm{TV}(v) = \sum_{i=1}^{n-1} |v_{i+1} - v_i|.$$

Lemma A.3: Let $\theta \in \mathbb{R}^n$. Let us define $\overline{\theta} = (\sum_{i=1}^n \theta_i)/n$. Then we have the following inequality:

$$\sum_{i=1}^n \big(\theta_i - \overline{\theta}\big)^2 \leq n\mathrm{TV}(\theta)^2\,.$$

Proof: Define $\alpha_1 = \theta_1, \beta_1 = 0$ and for every $i > 2$ define

$$\alpha_i = \alpha_{i-1} + (\theta_i - \theta_{i-1})_+.$$

Now define $\beta = \alpha - \theta$. Observe that as defined, α, β are monotonically non decreasing vectors. Also, we have the equality

$$\mathrm{TV}(\theta) = \mathrm{TV}(\alpha) + \mathrm{TV}(\beta) = (\alpha_n - \alpha_1) + (\beta_n - \beta_1).$$

Now we can expand:

$$\begin{aligned}\sum_{i=1}^n \big(\theta_i - \overline{\theta}\big)^2 &= \sum_{i=1}^n \big(\alpha_i - \overline{\alpha}\big)^2 + \sum_{i=1}^n \big(\beta_i - \overline{\beta}\big)^2 \\ &\quad - 2\sum_{i=1}^n (\alpha_i - \overline{\alpha})(\beta_i - \overline{\beta}) \\ &\leq \sum_{i=1}^n \big(\alpha_i - \overline{\alpha}\big)^2 + \sum_{i=1}^n \big(\beta_i - \overline{\beta}\big)^2 \\ &\quad + 2\sum_{i=1}^n |\alpha_i - \overline{\alpha}||\beta_i - \overline{\beta}| \\ &\leq n(\alpha_n - \alpha_1)^2 + n(\beta_n - \beta_1)^2 \\ &\quad + 2n(\alpha_n - \alpha_1)(\beta_n - \beta_1) \\ &= n(\alpha_n - \alpha_1 + \beta_n - \beta_1)^2 \\ &= n\mathrm{TV}(\theta)^2\,,\end{aligned}$$

thus giving us the lemma. □

Lemma A.4: Let $\alpha \in \mathbb{R}^n$ and let $B_1, B_2, \ldots, B_k$ be a partition of $[n]$ into contiguous blocks. Let α_{B_j} denote the restriction of α to the block B_j. Also let $\tilde{\alpha} \in \mathbb{R}^n$ be defined so that

$$\tilde{\alpha}_{B_j} = \frac{1}{|B_j|} \sum_{i \in B_j} \alpha_i \,.$$

In other words, $\tilde{\alpha}$ is the best Euclidean approximation to α within the subspace of all vectors which are constant on each block B_j. We then have the following inequality:

$$\mathrm{TV}(\tilde{\alpha}) \leq \mathrm{TV}(\alpha).$$

Proof: For any set of indices $i_1 \in B_1, \ldots, i_k \in B_k$, we have the following inequality:

$$\mathrm{TV}(\alpha) \geq \sum_{j=1}^{k-1} |\alpha_{i_{j+1}} - \alpha_{i_j}|.$$

Now averaging over the indices $i_j \in B_j$ and using Jensen's inequality gives us

$$\sum_{j=1}^{k-1} |\alpha_{i_{j+1}} - \alpha_{i_j}| \geq \sum_{j=1}^{k-1} |\tilde{\alpha}_{B_{j+1}} - \tilde{\alpha}_{B_j}|.$$

The last two displays finish the proof of the proposition. □

B. Proof of Lemma VIII.14

Proof: Let $P_0 = [n]$ be the initial partition. At every step we take the blocks $b_i \in P_i$ for which $T(b_i) > \epsilon$ and divide b_i into two equal parts. Let n_i be the number of blocks of the partition P_i and s_i equal the number of blocks B_i in P_i that are divided to obtain P_{i+1}. Define $s_0 = 0$. Therefore we have $n_{i+1} = n_i + s_i$. Note that, due to superadditivity of T, we must have $s_i \leq \lceil \frac{t}{\epsilon} \rceil$. This implies in particular that $n_i \leq 1 + i \lceil \frac{t}{\epsilon} \rceil$. Now the division scheme can go on for atmost $N = \lceil \log_2 \, n \rceil$ rounds. Therefore we have

$$\begin{aligned}\max_{P \in \mathcal{P}(t,n,\epsilon,T)} |P_{U;T,\epsilon}| &\leq 1 + \lceil \log_2 \, n \rceil \lceil \frac{t}{\epsilon} \rceil \\ &\leq 1 + (1 + \log_2 \, n)(1 + \frac{t}{\epsilon}) \,.\end{aligned}$$ □

Acknowledgment

The authors would to thank the anonymous referees for extremely detailed comments and suggestions about the paper. These comments and suggestions have helped us to a great extent to improve our article. The project started when both the authors were at the University of Chicago.

References

[1] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Inf. Inference*, vol. 3, no. 3, pp. 224–294, Sep. 2014.

[2] P. C. Bellec, "Sharp oracle inequalities for least squares estimators in shape restricted regression," *Ann. Statist.*, vol. 46, no. 2, pp. 745–780, Apr. 2018.

[3] L. Breiman, *Classification Regression Trees*. Evanston, IL, USA: Routledge, 2017.

[4] J.-F. Cai and W. Xu, "Guarantees of total variation minimization for signal recovery," *Inf. Inference A, J. IMA*, vol. 4, no. 4, pp. 328–353, 2015.

[5] E. Candes and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, Dec. 2007.

[6] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock, "An introduction to total variation for image analysis," *Theor. Found. Numer. Methods Sparse Recovery*, vol. 9, nos. 263–340, p. 227, 2010.

[7] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, Dec. 2012.

[8] S. Chatterjee, "High dimensional regression and matrix estimation without tuning parameters," 2015, *arXiv:1510.07294*. [Online]. Available: http://arxiv.org/abs/1510.07294

[9] S. Chatterjee and S. Goswami, "Adaptive estimation of multivariate piecewise polynomials and bounded variation functions by optimal decision trees," 2019, *arXiv:1911.11562*. [Online]. Available: http://arxiv.org/abs/1911.11562

[10] S. Chatterjee, A. Guntuboyina, and B. Sen, "On matrix estimation under monotonicity constraints," *Bernoulli*, vol. 24, no. 2, pp. 1072–1100, May 2018.

[11] S. Chatterjee and J. Lafferty, "Adaptive risk bounds in unimodal regression," *Bernoulli*, vol. 25, no. 1, pp. 1–25, Feb. 2019.

[12] A. S. Dalalyan, M. Hebiri, and J. Lederer, "On the prediction performance of the lasso," *Bernoulli*, vol. 23, no. 1, pp. 552–581, Feb. 2017.

[13] D. L. Donoho and I. M. Johnstone, "Minimax estimation via wavelet shrinkage," *Ann. Statist.*, vol. 26, no. 3, pp. 879–921, Jun. 1998.

[14] R. M. Dudley, "The sizes of compact subsets of Hilbert space and continuity of Gaussian processes," *J. Funct. Anal.*, vol. 1, no. 3, pp. 290–330, Oct. 1967.

[15] B. Fang, A. Guntuboyina, and B. Sen, "Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and hardy-krause variation," 2019, *arXiv:1903.01395*. [Online]. Available: http://arxiv.org/abs/1903.01395

[16] M. Genzel and G. M. Kutyniok März, "ℓ_1-analysis minimization and generalized (co-) sparsity: When does recovery succeed?" *Appl. Comput. Harmon. Anal.*, to be published.

[17] A. Guntuboyina, D. Lieu, S. Chatterjee, and B. Sen, "Adaptive risk bounds in univariate total variation denoising and trend filtering," 2017, *arXiv:1702.05113*. [Online]. Available: http://arxiv.org/abs/1702.05113

[18] Q. Han, T. Wang, S. Chatterjee, and R. J. Samworth, "Isotonic regression in general dimensions," 2017, *arXiv:1708.09468*. [Online]. Available: http://arxiv.org/abs/1708.09468

[19] J.-C. Hütter and P. Rigollet, "Optimal rates for total variation denoising," in *Proc. Conf. Learn. Theory*, 2016, pp. 1115–1146.

[20] M. Kabanava, H. Rauhut, and H. Zhang, "Robust analysis ℓ_1-recovery from Gaussian measurements and total variation minimization," 2014, *arXiv:1407.7402*. [Online]. Available: http://arxiv.org/abs/1407.7402

[21] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky, "ℓ_1 trend filtering," *SIAM Rev.*, vol. 51, no. 2, pp. 339–360, 2009.

[22] A. Langer, "Automated parameter selection for total variation minimization in image restoration," *J. Math. Imag. Vis.*, vol. 57, no. 2, pp. 239–268, Feb. 2017.

[23] M. Ledoux, "The concentration of measure phenomenon," in *Mathematical Surveys and Monographs*, vol. 89. Providence, RI, USA: AMS, 2001.

[24] G. Leoni, *A 1st Course Sobolev Spaces*. Providence, RI, USA: AMS, 2017.

[25] K. Lin, J. Sharpnack, A. Rinaldo, and R. J. Tibshirani, "Approximate recovery in changepoint problems, from ℓ_2 estimation error rates," 2016, *arXiv:1606.06746*. [Online]. Available: http://arxiv.org/abs/1606.06746

[26] E. Mammen and S. van de Geer, "Locally adaptive regression splines," *Ann. Statist.*, vol. 25, no. 1, pp. 387–413, Feb. 1997.

[27] F. Ortelli and S. van de Geer, "On the total variation regularized estimator over a class of tree graphs," *Electron. J. Statist.*, vol. 12, no. 2, pp. 4517–4570, 2018.

[28] F. Ortelli and S. van de Geer, "Adaptive rates for total variation image denoising," 2019, *arXiv:1911.07231*. [Online]. Available: http://arxiv.org/abs/1911.07231

[29] F. Ortelli and S. van de Geer, "Oracle inequalities for square root analysis estimators with application to total variation penalties," 2019, *arXiv:1902.11192*. [Online]. Available: http://arxiv.org/abs/1902.11192

[30] M. Osadebey, T. Adni, N. Bouguila, and D. Arnold, "Optimal selection of regularization parameter in total variation method for reducing noise in magnetic resonance images of the brain," *Biomed. Eng. Lett.*, vol. 4, no. 1, pp. 80–92, Mar. 2014.

[31] S. Oymak and B. Hassibi, "Sharp MSE bounds for proximal denoising," *Found. Comput. Math.*, vol. 16, no. 4, pp. 965–1029, 2013.

[32] A. Rinaldo, "Properties and refinements of the fused lasso," *Ann. Statist.*, vol. 37, no. 5B, pp. 2922–2952, Oct. 2009.

[33] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D, Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, Nov. 1992.

[34] V. Sadhanala and Y.-X. R. J. Wang Tibshirani, "Total variation classes beyond 1D: Minimax rates, and the limitations of linear smoothers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3513–3521.

[35] V. Solo, "Selection of regularisation parameters for total variation denoising," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 3, Mar. 1999, pp. 1653–1655.

[36] G. Steidl, S. Didas, and J. Neumann, "Splines in higher order TV regularization," *Int. J. Comput. Vis.*, vol. 70, no. 3, pp. 241–255, Dec. 2006.

[37] D. Strong and T. Chan, "Edge-preserving and scale-dependent properties of total variation regularization," *Inverse Problems*, vol. 19, no. 6, pp. S165–S187, Dec. 2003.

[38] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 67, no. 1, pp. 91–108, Feb. 2005.

[39] R. J. Tibshirani, "Adaptive piecewise polynomial estimation via trend filtering," *Ann. Statist.*, vol. 42, no. 1, pp. 285–323, Feb. 2014.

[40] R. van Handel, "Probability in high dimension," Princeton Univ., Princeton, NJ, USA, Tech. Rep., 2014.

[41] Y.-X. Wang, J. Sharpnack, A. J. Smola, and R. J. Tibshirani, "Trend filtering on graphs," *J. Mach. Learn. Res.*, vol. 17, no. 105, pp. 1–41, 2016.