# Higher-Order Total Variation Classes on Grids: Minimax Theory and Trend Filtering Methods

**Authored by:**

Yu-Xiang Wang
James L. Sharpnack
Veeranjaneyulu Sadhanala
Ryan J. Tibshirani

**Abstract**

We consider the problem of estimating the values of a function over $n$ nodes of a $d$-dimensional grid graph (having equal side lengths $n{1/d}$) from noisy observations. The function is assumed to be smooth, but is allowed to exhibit different amounts of smoothness at different regions in the grid. Such heterogeneity eludes classical measures of smoothness from nonparametric statistics, such as Holder smoothness. Meanwhile, total variation (TV) smoothness classes allow for heterogeneity, but are restrictive in another sense: only constant functions count as perfectly smooth (achieve zero TV). To move past this, we define two new higher-order TV classes, based on two ways of compiling the discrete derivatives of a parameter across the nodes. We relate these two new classes to Holder classes, and derive lower bounds on their minimax errors. We also analyze two naturally associated trend filtering methods; when $d=2$, each is seen to be rate optimal over the appropriate class.

## 1 Paper Body

In this work, we focus on estimation of a mean parameter defined over the nodes of a d-dimensional grid graph G = (V, E), with equal side lengths N = n1/d . Let us enumerate V = {1, . . . , n} and E = {e1 , . . . , em }, and consider data y = (y1 , . . . , yn ) ? Rn observed over V , distributed as yi ? N (?0,i , ? 2 ),

independently, for i = 1, . . . , n,

(1)

where ?0 = (?0,1 , . . . , ?0,n ) ? Rn is the mean parameter to be estimated, and ? 2 ¿ 0 the common noise variance. We will assume that ?0 displays some kind of regularity or smoothness over G, and are specifically interested in notions of regularity built around on the total variation (TV) operator X kD?k1 = —?i ? ?j —, (2) (i,j)?E

defined with respect to G, where D ? Rm?n is the edge incidence matrix of G, which has ‘th row D‘ = (0, . . . , ?1, . . . , 1, . . . , 0), with ?1 in location i and 1 in location j, provided that the ‘th edge is e‘ = (i, j) with i ¡ j. There is an extensive literature on estimators based on TV regularization, both in Euclidean spaces and over graphs. Higher-order TV regularization, which, loosely speaking, considers the TV of derivatives of the parameter, is much less understood, especially over graphs. In this paper, we develop statistical theory for higher-order TV smoothness classes, and we analyze 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

associated trend filtering methods, which are seen to achieve the minimax optimal estimation error rate over such classes. This can be viewed as an extension of the work in [22] for the zeroth-order TV case, where by ?zeroth-order?, we refer to the usual TV operator as defined in (2). Motivation. TV denoising over grid graphs, specifically 1d and 2d grid graphs, is a well-studied problem in signal processing, statistics, and machine learning, some key references being [20, 5, 26]. Given data y ? Rn as per the setup described above, the TV denoising or fused lasso estimator over the grid G is defined as 1 ?? = argmin ky ? ?k22 + ?kD?k1 , (3) n 2 ??R where ? ? 0 is a tuning parameter. The TV denoising estimator generalizes seamlessly to arbitrary graphs. The problem of denoising over grids, the setting we focus on, is of particular relevance to a number of important applications, e.g., in time series analysis, and image and video processing. A strength of the nonlinear TV denoising estimator in (3)?where by ?nonlinear?, we mean that ?? is nonlinear as a function of y?is that it can adapt to heterogeneity in the local level of smoothness of the underlying signal ?0 . Moreover, it adapts to such heterogeneity at an extent that is beyond what linear estimators are capable of capturing. This principle is widely evident in practice and has been championed by many authors in the signal processing literature. It is also backed by statistical theory, i.e., [8, 16, 27] in the 1d setting, and most recently [22] in the general d-dimensional setting. Note that the TV denoising estimator ?? in (3) takes a piecewise constant structure by design, i.e., at many adjacent pairs (i, j) ? E we will have ??i = ??j , and this will be generally more common for larger ?. For some problems, this structure may not be ideal and we might instead seek a piecewise smooth estimator, that is still able to cope with local changes in the underlying level of smoothness, but offers a richer structure (beyond a simple constant structure) for the base trend. In a 1d setting, this is accomplished by trend filtering methods, which move from piecewise constant to piecewise polynomial structure, via TV regularization of discrete derivatives of the parameter [24, 13, 27]. An extension of trend filtering to general graphs was developed in [31]. In what follows, we study the statistical properties of this graph trend filtering method over grids, and we propose and analyze a more specialized trend filtering estimator for grids based on the idea that something like a Euclidean coordinate system is available at any (interior) node. See Figure 1 for a motivating illustration. Related work. The literature on TV denoising is enormous and we cannot give a comprehensive review, but only some brief highlights. Important methodological and computational contributions are found in [20, 5, 26, 4, 10, 6, 28, 15, 7, 12, 1, 25], and notable

theoretical contributions are found in [16, 19, 9, 23, 11, 22, 17]. The literature on higher-order TV-based methods is more sparse and more concentrated on the 1d setting. Trend filtering methods in 1d were pioneered in [24, 13], and analyzed statistically in [27], where they were also shown to be asymptotically equivalent to locally adaptive regression splines of [16]. An extension of trend filtering to additive models was given in [21]. A generalization of trend filtering that operates over an arbitrary graph structure was given in [31]. Trend filtering is not the only avenue for higher-order TV regularization: the signal processing community has also studied higher-order variants of TV, see, e.g., [18, 3]. The construction of the discrete versions of these higher-order TV operators is somewhat similar to that in [31] as well our Kronecker trend filtering proposal, however, the focus of the work is quite different. Summary of contributions. An overview of our contributions is given below. ? We propose a new method for trend filtering over grid graphs that we call Kronecker trend filtering (KTF), and compare its properties to the more general graph trend filtering (GTF) proposal of [31]. ? For 2d grids, we derive estimation error rates for GTF and KTF, each of these rates being a function of the regularizer evaluated at the mean ?0 . ? For d-dimensional grids, we derive minimax lower bounds for estimation over two higherorder TV classes defined using the operators from GTF and KTF. When d = 2, these lower bounds match the upper bounds in rate (apart from log factors) derived for GTF and KTF, ensuring that each method is minimax rate optimal (modulo log factors) for its own notion of regularity. Also, the KTF class contains a Holder class of an appropriate order, and KTF is seen to be rate optimal (modulo log factors) for this more homogeneous class as well. 2

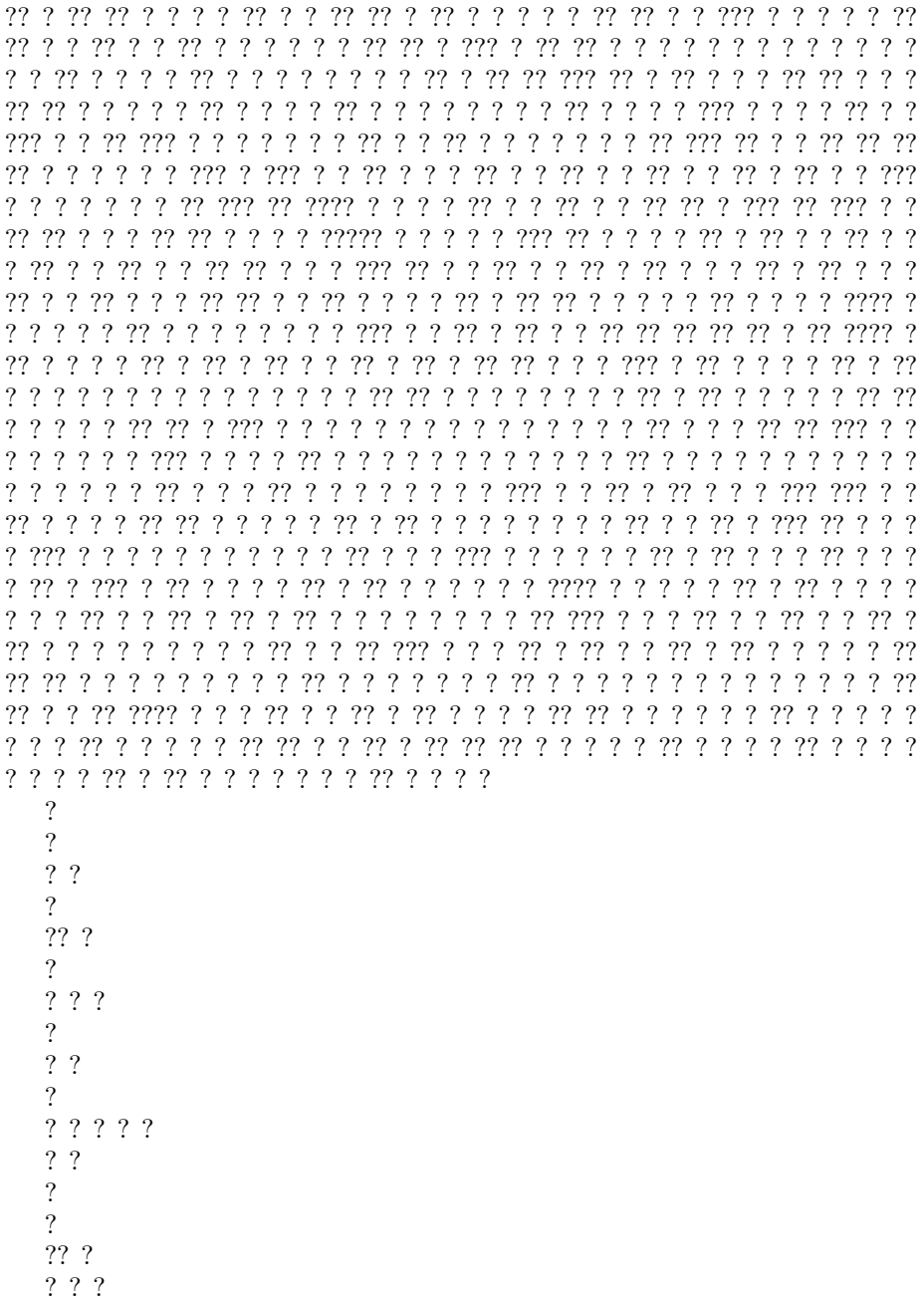


Figure 1: Top left: an underlying signal ?0 and associated data y (shown as black points). Top middle and top right: Laplacian smoothing fit to y, at large and small tuning parameter values, respectively. Bottom left, middle, and right: TV denoising (3), graph trend filtering (5), and Kronecker trend

filtering (5) fit to y, respectively (the latter two are of order k = 2, with penalty operators as described in Section 2). In order to capture the larger of the two peaks, Laplacian smoothing must significantly undersmooth throughout; with more regularization, it undersmooths throughout. TV denoising is able to adapt to heterogeneity in the smoothness of the underlying signal, but exhibits ?staircasing? artifacts, as it is restricted to fitting piecewise constant functions. Graph and Kronecker trend filtering overcome this, while maintaining local adaptivity.

Notation. For deterministic sequences an , bn we write an = O(bn ) to denote that an /bn is upper ?1 bounded for large enough n, and an bn to denote that both an = O(bn ) and a?1 n = O(bn ). For random sequences An , Bn , we write An = OP (Bn ) to denote that An /Bn is bounded in probability. Given a d-dimensional grid G = (V, E), where V = {1, . . . , n}, as before, we will sometimes index a parameter ? ? Rn defined over the nodes in the following convenient way. Letting N = n1/d and Zd = {(i1 /N, . . . , id /N ) : i1 , . . . , id ? {1, . . . , N }} ? [0, 1]d , we will index the components of ? by their lattice positions, denoted ?(x), x ? Zd . Further, for each j = 1, . . . , d, we will define the discrete derivative of ? in the jth coordinate direction at a location x by

?(x + ej /N ) ? ?(x) if x, x + ej /N ? Zd , (Dxj ?)(x) = (4) 0 else. Naturally, we denote by Dxj ? ? Rn the vector with components (Dxj ?)(x), x ? Zd . Higher-order discrete derivatives are simply defined by repeated application of the above definition. We use abbreviations (Dx2j ?)(x) = (Dxj (Dxj ?))(x), for j = 1, . . . , d, and (Dxj ,x‘ ?)(x) = (Dxj (Dx‘ ?))(x), for j, ‘ = 1, . . . , d, and so on. Given an estimator ?? of the mean parameter ?0 in (1), and K ? Rn , two quantities of interest are:

? ?0 ) = 1 k?? ? ?0 k2 and R(K) = inf sup E MSE(?, ? ?0 ) . MSE(?, 2 ? n ? ?0 ?K ? ?0 )] The first quantity here is called the mean squared error (MSE) of ?; we will also call E[MSE(?, ? the risk of ?. The second quantity is called the minimax risk over K (the infimum being taken over ? all estimators ?). 3

2

Trend filtering methods

Review: graph trend filtering. To review the family of estimators developed in [31], we start by introducing a general-form estimator called the generalized lasso signal approximator [28], 1 ?? = argmin ky ? ?k22 + ?k??k1 , 2 ??Rn

(5)

for a matrix ? ? Rr?n , referred to as the penalty operator. For an integer k ? 0, the authors [31] defined the graph trend filtering (GTF) estimator of order k by (5), with the penalty operator being

DLk/2 for k even, (k+1) ? = (6) L(k+1)/2 for k odd. Here, as before, we use D for the edge incidence matrix of G. We also use L = DT D for the graph Laplacian matrix of G. The intuition behind the above definition is that ?(k+1) ? gives something roughly like the (k + 1)st order discrete derivatives of ? over the graph G. Note that the GTF estimator reduces to TV denoising in (3) when k = 0. Also, like TV denoising, GTF applies to arbitrary graph structures; see [31] for more details and for the study of GTF over general graphs. Our interest is of course its behavior over grids, and we will now use the notation

introduced in (4), to shed more light on the GTF penalty P operator in (6) over a d-dimensional grid. For any signal ? ? Rn , we can write k?(k+1) ?k1 = x?Zd dx , where at all points x ? Zd (except for those close to the boundary), ?

d d

X X ? ?

? ? Dxj1 ,x2j ,...,x2j ? (x) for k even, where q = k/2,

? q ? 2

j =1 j ,...,j =1

dx = 1 d 2 q (7)

? ? X

? ? Dx2j ,x2j ,...,x2j ? (x) for k odd, where q = (k + 1)/2. ? q ? 1 2

j1 ,...,jq =1

Written in this form, it appears that the GTF operator ?(k+1) aggregates derivatives in somewhat of an unnatural way. But we must remember that for a general graph structure, only first derivatives and divergences have obvious discrete analogs?given by application of D and L, respectively. Hence, GTF, which was originally designed for general graphs, relies on combinations of D and L to produce something like higher-order discrete derivatives. This explains the form of the aggregated derivatives in (6), which is entirely based on divergences. Kronecker trend filtering. There is a natural alternative to the GTF penalty operator that takes advantage of the Euclidean-like structure available at the (interior) nodes of a grid graph. At a point x ? Zd (not close to the boundary), consider using dx =

d X

D k+1 ? (x) x j

(8)

j=1

as a basic building block for penalizing derivatives, rather than (7). This gives rise to a method we call Kronecker trend filtering (KTF), which for an integer order k ? 0 is defined by (5), but now with the choice of penalty operator ? ? (k+1) D1d ? I ? ??? ? I ? ? (k+1) ? I ? D1d ? ??? ? I ? (k+1) e ? ?. ? =? (9) .. ? ? ? . (k+1) I ? I ? ? ? ? ? D1d (k+1)

Here, D1d ? R(N ?k?1)?N is the 1d discrete derivative operator of order k + 1 (e.g., as used in univariate trend filtering, see [27]), I ? RN ?N is the identity matrix, and A ? B is the Kronecker product of matrices A, B. Each group of rows in (9) features a total of d ? 1 Kronecker products. KTF reduces to TV denoising in (3) when k = 0, and thus also to GTF with k = 0. But for k ? 1, GTF and KTF are different estimators. A look at the action of their penalty operators, as displayed in 4

(7), (8) reveals some of their differences. For example, we see that GTF considers mixed derivatives of total order k + 1, but KTF only considers directional derivatives of order k + 1 that are parallel to the coordinate axes. Also, GTF penalizes aggregate derivatives (i.e., sums of derivatives), whereas KTF penalizes individual ones. More subtle differences between GTF and KTF have to do with the structure of their estimates, as we discuss next. Another subtle

difference lies in how the GTF and KTF operators (6), (9) relate to more classical notions of smoothness, particularly, Holder smoothness. This is covered in Section 4. Structure of estimates. It is straightforward to see that the GTF operator (6) has a 1-dimensional null space, spanned by 1 = (1, . . . , 1) ? Rn . This means that GTF lets constant signals pass through unpenalized, but nothing else; or, in other words, it preserves the projection of y onto the space of constant signals, y?1, but nothing else. The KTF operator, meanwhile, has a much richer null space. Lemma 1. The null space of the KTF operator (9) has dimension (k + 1)d , and it is spanned by a polynomial basis made up of elements p(x) = xa1 1 xa2 2 ? ? ? xadd ,

x ? Zd ,

where a1 , . . . , ad ? {0, . . . , k}. The proof is elementary and (as with all proofs in this paper) is given in the supplement. The lemma shows that KTF preserves the projection of y onto the space of polynomials of max degree k, i.e., lets much more than just constant signals pass through unpenalized. Beyond the differences in these base trends (represented by their null spaces), GTF and KTF admit estimates with similar but generally different structures. KTF has the advantage that this structure is more transparent: its estimates are piecewise polynomial functions of max degree k, with generally fewer pieces for larger ?. This is demonstrated by a functional representation for KTF, given next. Lemma 2. Let hi : [0, 1] ? R, i = 1, . . . , N be the (univariate) falling factorial functions [27, 30] of order k, defined over knots 1/N, 2/N, . . . , N : hi (t) =

i?1 Y

(t ? t‘ ),

t ? [0, 1], i = 1, . . . , k + 1,

‘=1

k Y i+‘ i+k hi+k+1 (t) = t? ?1 t¿ , N N

(10) t ? [0, 1], i = 1, . . . , N ? k ? 1.

‘=1

(For k = 0, our convention is for the empty product to equal 1.) Let Hd be the space spanned by all d-wise tensor products of falling factorial functions, i.e., Hd contains f : [0, 1]d ? R of the form f (x) =

N X

?i1 ,...,id hi1 (x)hi2 (x2 ) ? ? ? hid (xd ),

x ? [0, 1]d ,

i1 ,...,id =1

for coefficients ? ? Rn (whose components we index by ?i1 ,...,id , for i1 , . . . , id = 1, . . . , N ). Then the KTF estimator defined in (5), (9) is equivalent to the functional optimization problem k

d X X 2 1 X ? f (?, x?j ) ? f = argmin y(x) ? f (x) + ? TV , (11) 2 ?xkj f ?Hd j=1 x ?Z x?Z ?j

d

d?1

where f (?, x?j ) denotes f as function of the jth dimension with all other dimensions fixed at x?j , ? k /?xkj (?) denotes the kth partial weak derivative

operator with respect to xj , for j = 1, . . . , d, and TV(?) denotes the total variation operator. The discrete (5), (9) and functional (11) representations are equivalent in that f? and ?? match at all grid locations x ? Zd . Aside from shedding light on the structure of KTF solutions, the functional optimization problem in ? which (11) is of practical importance: the function f? is defined over all of [0, 1]d (as opposed to ?, is of course only defined on the grid Zd ) and thus we may use it to interpolate the KTF estimate to non-grid locations. It is not clear to us that a functional representation as in (11) (or even a sensible interpolation strategy) is available for GTF on d-dimensional grids. 5

3

Upper bounds on estimation error

In this section, we assume that d = 2, and derive upper bounds on the estimation error of GTF and KTF for 2d grids. Upper bounds for generalized lasso estimators were studied in [31], and we will leverage one of their key results, which is based on what these authors call incoherence of the left singular vectors of the penalty operator ?. A slightly refined version of this result is stated below. Theorem 1 (Theorem 6 in [31]). Suppose that ? ? Rr?n has rank q, and denote by ?1 ? . . . ? ?q its nonzero singular values. Also let u1 , . . . , uq be the corresponding left singular vectors. Assume that these vectors, except for the first i0 , are incoherent, meaning that for a constant ? ? 1, ? kui k? ? ?/ n, i = i0 + 1, . . . , q, q Pq Then for ? ? (log r/n) i=i0 +1 ?i?2 , the generalized lasso estimator ?? in (5) satisfies v ! u q u log r X i ? 1 nullity(?) 0 ? ?0 ) = OP ? k??0 k1 . MSE(?, + + t n n n n i=i +1 ?i2 0

For GTF and KTF, we will apply this result, an appropriate choice of i0 with the partial Pbalancing q sum of reciprocal squared singular values i=i0 +1 ?i?2 . The main challenge, as we will see, is in establishing incoherence of the singular vectors. Error bounds for graph trend filtering. The authors in [31] have already used Theorem 1 (their Theorem 6) in order to derive error rates for GTF on 2d grids. However, their resultsP (specifically, q their Corollary 8) can be refined using a tighter upper bound for the partial sum term i=i0 +1 ?i?2 . No real further tightening is possible, since, as we show later, the results below match the minimax lower bound in rate, up to log factors. Theorem 2. Assume that d = 2. For k = 0, Cn = k?(1) ?0 k1 (i.e., Cn equal to the TV of ?0 , as in (2)), and ? log n, the GTF estimator in (5), (6) (i.e., the TV denoising estimator in (3)) satisfies

1 log n ? MSE(?, ?0 ) = OP + Cn . n n k

1

?

k

For any integer k ? 1, Cn = k?(k+1) ?0 k1 and ? n k+2 (log n) k+2 Cn k+2 , GTF satisfies

2 2 1 1 ? k+2 k+2 ? k+2 +n . MSE(?, ?0 ) = OP (log n) Cn n Remark 1. The result for k = 0 in Theorem 2 was essentially already established by [11] (a small difference is that the above rate is sharper by a factor of log n; though to be fair, [11] also take into account ‘0 sparsity). It is interesting to note that the case k = 0 appears to be quite special, in that the GTF estimator, i.e., TV

denoising estimator, is adaptive to the underlying smoothness parameter Cn (the prescribed choice of tuning parameter ? log n does not depend on Cn ). Pq The technique for upper bounding i=i0 +1 ?i?2 in the proof of Theorem 2 can be roughly explained as follows. The GTF operator ?(k+1) on a 2d grid has squared singular values:

k+1 ?(i1 ? 1) ?(i2 ? 1) 4 sin2 + 4 sin2 , i1 , i2 = 1, . . . , N. 2N 2N We can upper bound the sum of squared reciprocal singular values with a integral over [0, 1]2 , make use of the identity sin x ? x/2 for small enough x, and then switch to polar coordinates to calculate the integral (similar to [11], in analyzing TV denoising). The arguments to verify incoherence of the left singular vectors of ?(k+1) are themselves somewhat delicate, but were already given in [31]. Error bounds for Kronecker trend filtering. In comparison to the GTF case, the application of Theorem 1 to KTF is a much more difficult task, because (to the best of our knowledge) the KTF e (k+1) does not admit closed-form expressions for its singular values and vectors. This operator ? is true in any dimension (i.e., even for d = 1, where KTF reduces to univariate trend filtering). As it turns out, the singular values can be handled with a relatively straightforward application of the Cauchy interlacing theorem. It is establishing the incoherence of the singular vectors that proves to be the real challenge. This is accomplished by leveraging specialized approximation bounds for the eigenvectors of Toeplitz matrices from [2]. 6

Theorem 3. Assume that d = 2. For k = 0, since KTF reduces to the GTF with k = 0 (and to TV denoising), it satisfies the result stated in the first part of Theorem 2. k 1 ? k e (k+1) ?0 k1 and ? n k+2 For any integer k ? 1, Cn = k? (log n) k+2 Cn k+2 , the KTF estimator in (5), (9) satisfies

2 1 2 1 ? k+2 k+2 ? k+2 (log n) Cn MSE(?, ?0 ) = OP +n . n The results in Theorems 2 and 3 match, in terms of their dependence on n, k, d and the smoothness parameter Cn . As we will see in the next section, the smoothness classes defined by the GTF and KTF operators are similar, though not exactly the same, and each GTF and KTF is minimax rate optimal with respect to its own smoothness class, up to log factors. Beyond 2d? To analyze GTF and KTF on grids of dimension d ? 3, we would need to establish incoherence of the left singular vectors of the GTF and KTF operators. This should be possible by extending the arguments given in [31] (for GTF) and in the proof of Theorem 3 (for KTF), and is left to future work.

4

Lower bounds on estimation error

We present lower bounds on the minimax estimation error over smoothness classes defined by the operators from GTF (6) and KTF (9), denoted Tdk (Cn ) = {? ? Rn : k?(k+1) ?k1 ? Cn }, e (k+1) ?k1 ? Cn }, Tedk (Cn ) = {? ? Rn : k?

(12) (13)

respectively (where the subscripts mark the dependence on the dimension d of the underlying grid graph). Before we derive such lower bounds, we examine embeddings of (the discretization of) the class of Holder smooth functions into the GTF and KTF classes, both to understand the nature of these new classes,

and to define what we call a ?canonical? scaling for the radius parameter Cn . Embedding of Holder spaces and canonical scaling. Given an integer k ? 0 and L ¿ 0, recall that the Holder class H(k + 1, L; [0, 1]d ) contains k times differentiable functions f : [0, 1]d ? R, such that for all integers ?1 , . . . , ?d ? 0 with ?1 + ? ? ? + ?d = k,

? k f (x)

? k f (z) d ?

?x 1 ? ? ? ?x?d ? ?x?1 ? ? ? ?x?d ? Lkx ? zk2 , for all x, z ? [0, 1] . 1 1 d d To compare Holder smoothness with the GTF and KTF classes defined in (12), (13), we discretize the class H(k + 1, L; [0, 1]d ) by considering function evaluations over the grid Zd , defining

Hdk+1 (L) = ? ? Rn : ?(x) = f (x), x ? Zd , for some f ? H(k + 1, L; [0, 1]d ) . (14) Now we ask: how does the (discretized) Holder class in (14) compare to the GTF and KTF classes in (12), (13)? Beginning with a comparison to KTF, fix ? ? Hdk+1 (L), corresponding to evaluations of f ? H(k + 1, L; [0, 1]d ), and consider a point x ? Zd that is away from the boundary. Then the KTF penalty at x is

D k+1 ? (x) = Dxk ? (x + ej /N ) ? Dxk ? (x) xj j j k

?

?k + N k ?(N ) ? N k k f (x + ej /N ) ? f (x)

k ?xj ?xj ? LN k?1 + cLN k?1 .

(15)

In the second line above, we define ?(N ) to be the sum of absolute errors in the discrete approximations to the partial derivatives (i.e., the error in approximating ? k f (x)/?xkj by (Dxkj ?)(x)/N k , and similarly at x + ej /N ). In the third line, we use Holder smoothness to upper bound the first term, and we use standard numerical analysis (details in the supplement) for the second term to ensure that ?(N ) ? cL/N for a constant c ¿ 0 depending only on k. Summing the bound in (15) over x ? Zd as appropriate gives a uniform bound on the KTF penalty at ?, and leads to the next result. 7

Lemma 3. For any integers k ? 0, d ? 1, the (discretized) Holder and KTF classes defined in (14), (13) satisfy Hdk+1 (L) ? Tedk (cLn1?(k+1)/d ), where c ¿ 0 is a constant depending only on k. This lemma has three purposes. First, it provides some supporting evidence that the KTF class is an interesting smoothness class to study, as it shows the KTF class contains (discretizations of) Holder smooth functions, which are a cornerstone of classical nonparametric regression theory. In fact, this containment is strict and the KTF class contains more heterogeneous functions in it as well. Second, it leads us to define what we call the canonical scaling Cn n1?(k+1)/d for the radius of the KTF class (13). This will be helpful for interpreting our minimax lower bounds in what follows; at this scaling, note that we have Hdk+1 (1) ? Tedk (Cn ). Third and finally, it gives us an easy way to establish lower bounds on the minimax estimation error over KTF classes, by invoking well-known results on minimax rates for Holder classes. This will be described shortly. As for GTF, calculations similar to (15) are possible, but complications ensue for x on the boundary of the grid Zd . Importantly, unlike the KTF penalty, the GTF penalty includes discrete deriva-

tives at the boundary and so these complications have serious consequences, as stated next. Lemma 4. For any integers k, d ? 1, there are elements in the (discretized) Holder class Hdk+1 (1) in (14) that do not lie in the GTF class Tdk (Cn ) in (12) for arbitrarily large Cn . This lemma reveals a very subtle drawback of GTF caused by the use of discrete derivatives at the boundary of the grid. The fact that GTF classes do not contain (discretized) Holder classes makes them seem less natural (and perhaps, in a sense, less interesting) than KTF classes. In addition, it means that we cannot use standard minimax theory for Holder classes to establish lower bounds for the estimation error over GTF classes. However, as we will see next, we can construct lower bounds for GTF classes via another (more purely geometric) embedding strategy; interestingly, the resulting rates match the Holder rates, suggesting that, while GTF classes do not contain all (discretized) Holder functions, they do contain ?enough? of these functions to admit the same lower bound rates. Minimax rates for GTF and KTF classes. Following from classical minimax theory for Holder classes [14, 29], and Lemma 3, we have the following result for the minimax rates over KTF classes. Theorem 4. For any integers k ? 0, d ? 1, the KTF class defined in (13) has minimax estimation error satisfying 2d

2d R Tedk (Cn ) = ?(n? 2k+2+d Cn2k+2+d ). For GTF classes, we use a different strategy. We embed an ellipse, then rotate the parameter space and embed a hypercube, leading to the following result. Theorem 5. For any integers k ? 0, d ? 1, the GTF class defined in (12) has minimax estimation error satisfying 2d

2d R Tdk (Cn ) = ?(n? 2k+2+d Cn2k+2+d ). Several remarks are in order. Remark 2. Plugging in the canonical scaling Cn n1?(k+1)/d in Theorems 4 and 5, we see that 2k+2

R(Tedk (Cn )) = ?(n? 2k+2+d ) and

2k+2

R(Tdk (Cn )) = ?(n? 2k+2+d ),

both matching the usual rate for the Holder class Hdk+1 (1). For KTF, this should be expected, as its lower bound is constructed via the Holder embedding given in Lemma 3. But for GTF, it may come as somewhat of a surprise?despite the fact it does not embed a Holder class as in Lemma 4, we see that the GTF class shares the same rate, suggesting it still contains something like ?hardest? Holder smooth signals. Remark 3. For d = 2 and all k ? 0, we can certify that the lower bound rate in Theorem 4 is tight, modulo log factors, by comparing it to the upper bound in Theorem 3. Likewise, we can certify that the lower bound rate in Theorem 5 is tight, up to log factors, by comparing it to the upper bound in Theorem 2. For d ? 3, the lower bound rates in Theorems 4 and 5 will not be tight for some values of k. For example, when k = 0, at the canonical scaling Cn n1?1/d , the lower bound rate (given by either theorem) is n?2/(2+d) , however, [22] prove that the minimax error of the TV class scales (up to log factors) as n?1/d for d ? 2, so we see there is a departure in the rates for d ? 3. 8

GTFcl ass KTFcl ass H?l dercl ass

Figure 2: Illustration of the two higher-order TV classes, namely the GTF

and KTF classes, as they relate to the (discretized) Holder class. The horizontally/vertically checkered region denotes the part of Holder class not contained in the GTF class. As explained in Section 4, this is due to the fact that the GTF operator penalizes discrete derivatives on the boundary of the grid graph. The diagonally checkered region (also colored in blue) denotes the part of the Holder class contained in the GTF class. The minimax lower bound rates we derive for the GTF class in Theorem 5 match the well-known Holder rates, suggesting that this region is actually sizeable and contains the ?hardest? Holder smooth signals.

In general, we conjecture that the Holder embedding for the KTF class (and ellipse embedding for GTF) will deliver tight lower bound rates, up to log factors, when k is large enough compared to d. This would have interesting implications for adaptivity to smoother signals (see the next remark); a precise study will be left to future work, along with tight minimax lower bounds for all k, d. Remark 4. Again by comparing Theorems 3 and 4, as well as Theorems 2 and 5, we find that, for d = 2 and all k ? 0, KTF is rate optimal for the KTF smoothness class and GTF is rate optimal for the GTF smoothness class, modulo log factors. We conjecture that this will continue to hold for all d ? 3, which will be examined in future work. Moreover, an immediate consequence of Theorem 3 and the Holder embedding in Lemma 3 is that KTF adapts automatically to Holder smooth signals, i.e., it achieves a rate (up to log factors) of n?(k+1)/(k+2) over H2k+1 (1), matching the well-known minimax rate for the more homogeneous Holder class. It is not clear that GTF shares this property.

5

Discussion

In this paper, we studied two natural higher-order extensions of the TV estimator on d-dimensional grid graphs. The first was graph trend filtering (GTF) as defined in [31], applied to grids; the second was a new Kronecker trend filtering (KTF) method, which was built with the special (Euclidean-like) structure of grids in mind. GTF and KTF exhibit some similarities, but are different in important ways. Notably, the notion of smoothness defined using the KTF operator is somewhat more natural, and is a strict generalization of the standard notion of Holder smoothness (in the sense that the KTF smoothness class strictly contains a Holder class of an appropriate order). This is not true for the notion of smoothness defined using the GTF operator. Figure 2 gives an illustration. When d = 2, we derived tight upper bounds for the estimation error achieved by the GTF and KTF estimators?tight in the sense that these upper bound match in rate (modulo log factors) the lower bounds on the minimax estimation errors for the GTF and KTF classes. We constructed the lower bound for the KTF class by leveraging the fact that it embeds a Holder class; for the GTF class, we used a different (more geometric) embedding. While these constructions proved to be tight for d = 2 and all k ? 0, we suspect this will no longer be the case in general, when d is large enough relative to k. We will examine this in future work, along with upper bounds for GTF and KTF when d ? 3. Another important consideration for future work are the minimax linear rates over GTF and KTF classes, i.e., minimax rates when we restrict

our attention to linear estimators. We anticipate that a gap will exist between minimax linear and nonlinear rates for all k, d (as it does for k = 0, as shown in [22]). This would, e.g., provide some rigorous backing to the claim that the KTF class is larger than its embedded Holder class (the latter having matching minimax linear and nonlinear rates). Acknowledgements. We thank Sivaraman Balakrishnan for helpful discussions regarding minimax rates for Holder classes on grids. JS was supported by NSF Grant DMS-1712996. VS, YW, and RT were supported by NSF Grants DMS-1309174 and DMS-1554123. 9

# 2 References

[1] Alvaro Barbero and Suvrit Sra. Modular proximal optimization for multidimensional totalvariation regularization. arXiv: 1411.0589, 2014. [2] Johan M. Bogoya, Albrecht Bottcher, Sergei M. Grudsky, and Egor A. Maximenko. Eigenvectors of Hermitian Toeplitz matrices with smooth simple-loop symbols. Linear Algebra and its Applications, 493:606?637, 2016. [3] Kristian Bredies, Karl Kunisch, and Thomas Pock. Total generalized variation. SIAM Journal on Imaging Sciences, 3(3):492?526, 2010. [4] Antonin Chambolle and Jerome Darbon. On total variation minimization and surface evolution using parametric maximum flows. International Journal of Computer Vision, 84:288?307, 2009. [5] Antonin Chambolle and Pierre-Louis Lions. Image recovery via total variation minimization and related problems. Numerische Mathematik, 76(2):167?188, 1997. [6] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision, 40:120?145, 2011. [7] Laurent Condat. A direct algorithm for 1d total variation denoising. HAL: 00675043, 2012. [8] David L. Donoho and Iain M. Johnstone. Minimax estimation via wavelet shrinkage. Annals of Statistics, 26(8):879?921, 1998. [9] Zaid Harchaoui and Celine Levy-Leduc. Multiple change-point estimation with a total variation penalty. Journal of the American Statistical Association, 105(492):1480?1493, 2010. [10] Holger Hoefling. A path algorithm for the fused lasso signal approximator. Journal of Computational and Graphical Statistics, 19(4):984?1006, 2010. [11] Jan-Christian Hutter and Philippe Rigollet. Optimal rates for total variation denoising. Annual Conference on Learning Theory, 29:1115?1146, 2016. [12] Nicholas Johnson. A dynamic programming algorithm for the fused lasso and l0 -segmentation. Journal of Computational and Graphical Statistics, 22(2):246?260, 2013. [13] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ‘1 trend filtering. SIAM Review, 51(2):339?360, 2009. [14] Aleksandr P. Korostelev and Alexandre B. Tsybakov. Minimax Theory of Image Reconstructions. Springer, 2003. [15] Arne Kovac and Andrew Smith. Nonparametric regression on a graph. Journal of Computational and Graphical Statistics, 20(2):432?447, 2011. [16] Enno Mammen and Sara van de Geer. Locally apadtive regression splines. Annals of Statistics, 25(1):387?413, 1997. [17] Oscar Hernan Madrid Padilla, James Sharpnack, James Scott, , and Ryan J. Tibshirani. The DFS fused lasso: Linear-time denoising over general graphs. arXiv: 1608.03384, 2016. [18] Chris-

tiane Poschl and Otmar Scherzer. Characterization of minimizers of convex regularization functionals. In Frames and Operator Theory in Analysis and Signal Processing, volume 451, pages 219?248. AMS eBook Collections, 2008. [19] Alessandro Rinaldo. Properties and refinements of the fused lasso. Annals of Statistics, 37(5): 2922?2952, 2009. [20] Leonid I. Rudin, Stanley Osher, and Emad Faterni. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena, 60(1):259?268, 1992. [21] Veeranjaneyulu Sadhanala and Ryan J. Tibshirani. Additive models via trend filtering. arXiv: 1702.05037, 2017. 10

[22] Veeranjaneyulu Sadhanala, Yu-Xiang Wang, and Ryan J. Tibshirani. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. Advances in Neural Information Processing Systems, 29, 2016. [23] James Sharpnack, Alessandro Rinaldo, and Aarti Singh. Sparsistency via the edge lasso. International Conference on Artificial Intelligence and Statistics, 15, 2012. [24] Gabriel Steidl, Stephan Didas, and Julia Neumann. Splines in higher order TV regularization. International Journal of Computer Vision, 70(3):214?255, 2006. [25] Wesley Tansey and James Scott. A fast and flexible algorithm for the graph-fused lasso. arXiv: 1505.06475, 2015. [26] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B, 67(1):91?108, 2005. [27] Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. Annals of Statistics, 42(1):285?323, 2014. [28] Ryan J. Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. Annals of Statistics, 39(3):1335?1371, 2011. [29] Alexandre B. Tsybakov. Introduction to Nonparametric Estimation. Springer, 2009. [30] Yu-Xiang Wang, Alexander Smola, and Ryan J. Tibshirani. The falling factorial basis and its statistical applications. International Conference on Machine Learning, 31, 2014. [31] Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan J. Tibshirani. Trend filtering on graphs. Journal of Machine Learning Research, 17(105):1?41, 2016.

11