Submitted to the Annals of Statistics

SUPPLEMENT TO "ADAPTIVE PIECEWISE POLYNOMIAL ESTIMATION VIA TREND FILTERING"

By Ryan J. Tibshirani

Carnegie Mellon University

We provide proofs and supplementary technical results for "Adaptive Piecewise Polynomial via Trend Filtering". We use the prefix "TF" when referring to equations, lemmas, etc., in the former paper, as in equation (TF-2) or Lemma TF-2 (this stands for *T*rend *F*iltering).

1. Lasso and continuous-time representations.

1.1. *Proof of Lemma TF-2.* Consider making the variable transformation $\alpha = n^k/k! \cdot D\beta$ in (TF-2), with $D \in \mathbb{R}^{n \times n}$ defined as

$$D = \begin{bmatrix} D_1^{(0)} \\ \vdots \\ D_1^{(k)} \\ D^{(k+1)} \end{bmatrix},$$

where $D_1^{(i)} \in \mathbb{R}^{1 \times n}$ denotes the first row of the ith discrete difference operator $D^{(i)}$, for $i = 0, \ldots k$ (and $D^{(0)} = I$ by convention). We first show that $D^{-1} = M$, where $M = M^{(0)} \cdot \ldots \cdot M^{(k)}$ and

$$M^{(i)} = \begin{bmatrix} I_{i \times i} & 0 \\ 0 & L_{(n-i) \times (n-i)} \end{bmatrix} \quad \text{for } i = 0, \ldots k.$$

Here $I_{i \times i}$ is the $i \times i$ identity matrix, and $L_{(n-i) \times (n-i)}$ is the $(n-i) \times (n-i)$ lower triangular matrix of 1s. In particular, we prove that $M^{-1} = D$ by induction on k. When $k = 0$, that

$$\begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 1 & \ldots & 0 \\ \vdots & & & & \\ 1 & 1 & 1 & \ldots & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix}$$

can be seen directly by inspection. Assume now that the statement holds for $k-1$. Then

$$\begin{aligned}\Big(M^{(0)} \cdot \ldots \cdot M^{(k)}\Big)^{-1} &= (M^{(k)})^{-1}\Big(M^{(0)} \cdot \ldots \cdot M^{(k-1)}\Big)^{-1} \\ &= \begin{bmatrix} I & 0 \\ 0 & L^{-1} \end{bmatrix} \begin{bmatrix} D_1^{(0)} \\ \vdots \\ D_1^{(k-1)} \\ D^{(k)} \end{bmatrix} \\ &= \begin{bmatrix} D_1^{(0)} \\ \vdots \\ D_1^{(k-1)} \\ L^{-1}D^{(k)} \end{bmatrix},\end{aligned}$$

where we have abbreviated $I = I_{k\times k}$ and $L = L_{(n-k)\times(n-k)}$, and in the second equality we used the inductive hypothesis. Moreover,

$$L^{-1}D^{(k)} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix} D^{(k)} = \begin{bmatrix} D_1^{(k)} \\ D^{(k+1)} \end{bmatrix},$$

completing the inductive proof. Therefore, substituting $\alpha = n^k/k! \cdot D\beta$ in (TF-2) yields the problem

$$\hat{\alpha} = \underset{\alpha\in\mathbb{R}^n}{\operatorname{argmin}} \; \frac{1}{2}\Big\|y - \frac{k!}{n^k} M\alpha\Big\|_2^2 + \lambda \sum_{j=k+2}^{n} |\alpha_j|.$$

It is now straightforward to check that the last $n-k-1$ columns of $(k!/n^k)M$ match those of H, as defined in (TF-24). Furthermore, the first $k+1$ columns of $(k!/n^k)M$ have the same linear span as the first $k+1$ columns of H, which is sufficient because the ℓ_1 penalty above [and in (TF-23)] only applies to the last $n-k-1$ coefficients. □

1.2. *Proof of Lemma TF-3.* By inspection of (TF-22) and (TF-24), we have

$$G = H = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 0 \\ \vdots & & & \\ 1 & 1 & \ldots & 1 \end{bmatrix} \text{ if } k = 0,$$

$$G = H = \frac{1}{n} \cdot \begin{bmatrix} n & 1 & 0 & 0 & \ldots & 0 \\ n & 2 & 0 & 0 & \ldots & 0 \\ n & 3 & 1 & 0 & \ldots & 0 \\ n & 4 & 2 & 1 & \ldots & 0 \\ \vdots & & & & & \\ n & n & n-2 & n-3 & \ldots & 1 \end{bmatrix} \text{ if } k = 1,$$

and $G \neq H$ if $k \geq 2$ [in this case G and H differ by more than a scalar factor, so problems (TF-20) and (TF-23) cannot be equated by simply modifying the tuning parameters]. □

1.3. *Proof of Lemma TF-4.* Define $s_i^{(0)} = 1$ for all i, and

$$s_i^{(k)} = \sum_{j=1}^{i-1} s_j^{(k-1)} \quad \text{for } k = 1, 2, 3, \ldots,$$

i.e., $s_i^{(k)}$ is the kth order cumulative sum of $(1, 1, \ldots 1) \in \mathbb{R}^i$, with lag 1. We will prove that

$$\frac{(x-k)(x-k+1)\cdot \ldots \cdot (x-1)}{k!} = s_x^{(k)} \quad \text{for all } x = 1, 2, 3, \ldots \text{ and } k = 1, 2, 3, \ldots, \tag{1}$$

by induction on k. Note that this would be sufficient to prove the result in the lemma, as it would show that the bottom right $(n-k-1) \times (n-k-1)$ sub-block of H in (TF-24), which can be expressed as

$$\frac{k!}{n^k} \cdot \begin{bmatrix} s_{k+1}^{(k)} & 0 & \ldots & 0 \\ s_{k+2}^{(k)} & s_{k+1}^{(k)} & \ldots & 0 \\ \vdots & & & \\ s_{n-1}^{(k)} & s_{n-2}^{(k)} & \ldots & s_{k+1}^{(k)} \end{bmatrix},$$

is equal to that in (TF-27). We now give the inductive argument for (1). As for the base case, $k = 1$: clearly $x - 1 = s_x^{(1)}$ for all x. Assume that the inductive hypothesis holds for k. Then for any x,

$$\begin{aligned} s_x^{(k+1)} &= \sum_{i=1}^{x-1} \frac{(i-k)(i-k+1)\cdot \ldots \cdot (i-1)}{k!} \\ &= \sum_{i=1}^{x-1} \sum_{j=1}^{i-1} \frac{(j-k+1)(j-k+2)\cdot \ldots \cdot (j-1)}{(k-1)!}, \end{aligned}$$

by the inductive hypothesis. Switching the order of the summations,

$$\begin{aligned} s_x^{(k+1)} &= \sum_{j=1}^{x-2} \sum_{i=j+1}^{x-1} \frac{(j-k+1)(j-k+2)\cdot \ldots \cdot (j-1)}{(k-1)!} \\ &= \sum_{j=1}^{x-2} \frac{(j-k+1)(j-k+2)\cdot \ldots \cdot (j-1)}{(k-1)!} \cdot (x-j-1) \\ &= \sum_{j=1}^{x-2} \frac{(j-k+1)(j-k+2)\cdot \ldots \cdot (j-1)}{(k-1)!} \cdot (x-k-1-j+k). \end{aligned}$$

Grouping terms and again applying the inductive hypothesis,

$$s_x^{(k+1)} = \frac{(x-k-1)(x-k-2)\cdot \ldots \cdot (x-2)}{k!} \cdot (x-k-1) - s_{x-1}^{(k+1)} \cdot k.$$

Noting that $s_x^{(k+1)} = s_{x-1}^{(k+1)} + (x-k-1)\cdot\ldots\cdot(x-2)/k!$, and rearranging terms finally gives

$$s_{x-1}^{(k+1)} = \frac{(x-k-2)(x-k-1)\cdot\ldots\cdot(x-2)}{(k+1)!}.$$

Since x was arbitrary, this completes the inductive step, and hence the proof. □

2. Bounding the difference in lasso fitted values.

2.1. *Lasso problems in standard form.* Consider two lasso problems sharing the same outcome vector $y \in \mathbb{R}^n$,

$$\min_{\theta\in\mathbb{R}^p} \frac{1}{2}\|y - X\theta\|_2^2 + \lambda\|\theta\|_1, \tag{2}$$

$$\min_{\alpha\in\mathbb{R}^p} \frac{1}{2}\|y - Z\alpha\|_2^2 + \lambda'\|\alpha\|_1, \tag{3}$$

where $X, Z \in \mathbb{R}^{n\times p}$, and $\lambda, \lambda' \geq 0$. One might ask: if λ, λ' are chosen appropriately, can we bound the difference in the fitted values $X\hat{\theta}$ and $Z\hat{\alpha}$ of (2) and (3), respectively, in terms of the difference between X and Z? This question can be answered by first deriving a "basic inequality", much like the one used for bounding lasso prediction error.

LEMMA 1. *For fixed λ, λ', solutions $\hat{\theta}$ and $\hat{\alpha}$ of lasso problems* (2) *and* (3) *satisfy*

$$\frac{1}{2}\|X\hat{\theta} - Z\hat{\alpha}\|_2^2 \leq \langle y - X\hat{\theta}, Z\hat{\alpha}\rangle + (\lambda' - \lambda)\|\hat{\theta}\|_1 - \lambda'\|\hat{\alpha}\|_1 + R, \tag{4}$$

where $R = \frac{1}{2}\|(X-Z)\hat{\theta}\|_2^2 + \langle y - X\hat{\theta}, (X-Z)\hat{\theta}\rangle$.

PROOF. Note that by optimality,

$$\begin{aligned}\frac{1}{2}\|y - Z\hat{\alpha}\|_2^2 + \lambda'\|\hat{\alpha}\|_1 &\leq \frac{1}{2}\|y - Z\hat{\theta}\|_2^2 + \lambda'\|\hat{\theta}\|_1 \\ &= \frac{1}{2}\|y - X\hat{\theta}\|_2^2 + \lambda'\|\hat{\theta}\|_1 + R,\end{aligned}$$

where $R = \frac{1}{2}\|y - Z\hat{\theta}\|_2^2 - \frac{1}{2}\|y - X\hat{\theta}\|_2^2$. We can rearrange the above inequality as

$$\frac{1}{2}\|Z\hat{\alpha}\|_2^2 - \frac{1}{2}\|X\hat{\theta}\|_2^2 \leq \langle y, Z\hat{\alpha} - X\hat{\theta}\rangle + \lambda'\|\hat{\theta}\|_1 - \lambda'\|\hat{\alpha}\|_1 + R.$$

Writing $y = X\hat{\theta} + (y - X\hat{\theta})$ within the inner product on the right-hand side above, and bringing the term involving $X\hat{\theta}$ over to the left-hand side, we have

$$\begin{aligned}\frac{1}{2}\|X\hat{\theta} - Z\hat{\alpha}\|_2^2 &\leq \langle y - X\hat{\theta}, Z\hat{\alpha} - X\hat{\theta}\rangle + \lambda'\|\hat{\theta}\|_1 - \lambda'\|\hat{\alpha}\|_1 + R, \\ &= \langle y - X\hat{\theta}, Z\hat{\alpha}\rangle + (\lambda' - \lambda)\|\hat{\theta}\|_1 - \lambda'\|\hat{\alpha}\|_1 + R,\end{aligned}$$

where in the last line we used the fact that $\langle y - X\hat{\theta}, X\hat{\theta}\rangle = \lambda\|\hat{\theta}\|_1$, from the KKT conditions for problem (2). Lastly, we rewrite

$$\begin{aligned} R &= \frac{1}{2}\|Z\hat{\theta}\|_2^2 - \frac{1}{2}\|X\hat{\theta}\|_2^2 + \langle y, X\hat{\theta} - Z\hat{\theta}\rangle \\ &= \frac{1}{2}\|X\hat{\theta} - Z\hat{\theta}\|_2^2 + \langle y - X\hat{\theta}, X\hat{\theta} - Z\hat{\theta}\rangle, \end{aligned}$$

which completes the proof. □

In order to bound $\|X\hat{\theta} - Z\hat{\alpha}\|_2$, the goal now is to determine conditions under which the right-hand side in (4) is small. Note that both terms in R involves the difference $X - Z$, which will have small entries if X and Z are close. The second term in (4) can be controlled by taking λ' and λ to be close. As for the first term in (4), we can rewrite

$$\langle y - X\hat{\theta}, Z\hat{\alpha}\rangle = \langle y - X\hat{\theta}, (Z - X)\hat{\alpha}\rangle + \langle X^T(y - X\hat{\theta}), \hat{\alpha}\rangle;$$

above, the first term again involves the difference $X - Z$, and the second term can be balanced by the term $-\lambda\|\hat{\alpha}\|_1$ appearing in (4) if λ and λ' are chosen carefully. These ideas are all made precise in the next lemma.

LEMMA 2. *Consider a sequence of lasso problems* (2), (3) *(all quantities $p, y, X, Z, \lambda, \lambda'$ considered as functions of n), such that $\lambda' = (1+\delta)\lambda$ for some fixed $\delta > 0$. Assume that*

$$\sqrt{p}\|X - Z\|_\infty\|\hat{\theta}\|_1 = O\Big(\sqrt{\lambda\|\hat{\theta}\|_1}\Big) \quad \text{and} \quad \frac{\sqrt{p}\|X - Z\|_\infty\|y - X\hat{\theta}\|_2}{\lambda} \to 0 \quad \text{as } n \to \infty. \tag{5}$$

Then any solutions $\hat{\theta}, \hat{\alpha}$ of (2), (3) *satisfy*

$$\|X\hat{\theta} - Z\hat{\alpha}\|_2 = O\Big(\sqrt{\lambda\|\hat{\theta}\|_1}\Big). \tag{6}$$

PROOF. As suggested in the discussion before the lemma, we rewrite the term $\langle y - X\hat{\theta}, Z\hat{\alpha}\rangle$ in the right-hand side of (4) as

$$\begin{aligned} \langle y - X\hat{\theta}, Z\hat{\alpha}\rangle &= \langle y - X\hat{\theta}, (Z - X)\hat{\alpha}\rangle + \langle X^T(y - X\hat{\theta}), \hat{\alpha}\rangle \\ &\le \|y - X\hat{\theta}\|_2\|(X - Z)\hat{\alpha}\|_2 + \lambda\|\hat{\alpha}\|_1 \\ &\le \sqrt{p}\|y - X\hat{\theta}\|_2\|X - Z\|_\infty\|\hat{\alpha}\|_1 + \lambda\|\hat{\alpha}\|_1, \end{aligned}$$

where in the second line we used Hölder's inequality and the fact that $\|X^T(y - X\hat{\theta})\|_\infty \le \lambda$ from the KKT conditions for (2), and in the third line we used the bound $\|Ax\|_2 \le \sqrt{p}\|A\|_\infty\|x\|_1$ for a matrix $A \in \mathbb{R}^{n\times p}$, where $\|A\|_\infty$ denotes the maximum element of A in absolute value. The assumption that $\sqrt{p}\|X - Z\|_\infty\|y - X\hat{\theta}\|_2/\lambda \to 0$ now implies that, for large enough n,

$$\langle y - X\hat{\theta}, Z\hat{\alpha}\rangle \le \delta\lambda\|\hat{\alpha}\|_1 + \lambda\|\hat{\alpha}\|_1 = (1+\delta)\lambda\|\hat{\alpha}\|_1.$$

Plugging this into the right-hand side of (4), and using $\lambda' = (1+\delta)\lambda$, we see that

$$\frac{1}{2}\|X\hat{\theta} - Z\hat{\alpha}\|_2^2 \le \delta\lambda\|\hat{\theta}\|_1 + R.$$

Finally,

$$R \leq \frac{1}{2}\Big(\sqrt{p}\|X - Z\|_\infty \|\hat{\theta}\|_1\Big)^2 + \sqrt{p}\|X - Z\|_\infty \|y - X\hat{\theta}\|_2 \|\hat{\theta}\|_1,$$

and using both conditions in (5), we have $R = O(\lambda\|\hat{\theta}\|_1)$, completing the proof. □

Remark. Had we instead chosen $\lambda' = \lambda$, which may seem like more of a natural choice for pairing the two problems (2), (3), the same arguments would have yielded the final bound

$$\|X\hat{\theta} - Z\hat{\alpha}\|_2 = O\Big(\sqrt{\lambda \max\{\|\hat{\theta}\|_1, \|\hat{\alpha}\|_1\}}\Big).$$

For some purposes, this may be just fine. However, the envisioned main use case of this lemma is one in which some (desirable) theoretical properties are known for the lasso problem (2) with a particular predictor matrix X, and analogous results for a lasso problem (3) with similar predictor matrix Z are sought (e.g., this is the usage for locally adaptive regression splines and trend filtering); in such a case, a bound of the form (6) is preferred as it does not depend at all on the output of problem (3).

The second condition in (5) involves the quantity $y - X\hat{\theta}$, and so may appear more complicated than necessary. Indeed, under weak assumptions on y and $X\hat{\theta}$, this condition can be simplified.

COROLLARY 1. *Consider again a sequence of lasso problems* (2), (3) *such that* $\lambda' = (1+\delta)\lambda$ *for some fixed* $\delta > 0$. *Assume that the outcome vector* y *is drawn from the regression model*

$$y = \mu + \epsilon,$$

where $\epsilon_1, \ldots \epsilon_n$ *are i.i.d. with* $\mathbb{E}[\epsilon_i^4] < \infty$, *and assume that* $\|\mu - X\hat{\theta}\|_2 = O_{\mathbb{P}}(\sqrt{n})$. *Further assume*

$$\sqrt{p}\|X - Z\|_\infty \|\hat{\theta}\|_1 = O_{\mathbb{P}}\Big(\sqrt{\lambda\|\hat{\theta}\|_1}\Big) \quad \textit{and} \quad \sqrt{np}\|X - Z\|_\infty/\lambda \to 0 \quad \textit{as } n \to \infty.$$

Then any solutions $\hat{\theta}, \hat{\alpha}$ *of* (2), (3) *satisfy*

$$\|X\hat{\theta} - Z\hat{\alpha}\|_2 = O_{\mathbb{P}}\Big(\sqrt{\lambda\|\hat{\theta}\|_1}\Big).$$

PROOF. Note that

$$\|y - X\hat{\theta}\|_2 \leq \|\epsilon\|_2 + \|\mu - X\hat{\theta}\|_2.$$

Both terms on the right-hand side above are $O_{\mathbb{P}}(\sqrt{n})$; for the second term, this is true by assumption, and for the first term, we can use the fact that $\epsilon_1, \ldots \epsilon_n$ are i.i.d. with finite fourth moment to argue

$$\mathbb{P}\left(\frac{\sum_{i=1}^n \epsilon_i^2}{n} - \mathbb{E}[\epsilon_i^2] > 1\right) \leq \frac{\mathrm{Var}(\epsilon_i^2)}{n} \to 0,$$

where we used Chebyshev's inequality. Therefore $\|y - X\hat{\theta}\|_2 = O_{\mathbb{P}}(\sqrt{n})$, and to show that $\sqrt{p}\|X - Z\|_\infty \|y - X\hat{\theta}\|_2/\lambda \to 0$ in probability, it suffices to show $\sqrt{np}\|X - Z\|_\infty/\lambda \to 0$. □

Remark. The fourth moment condition on the errors, $\mathbb{E}[\epsilon_i^4] < \infty$, is not a strong one. E.g., any sub-Gaussian distribution [which was our distributional assumption in (TF-31) for the theoretical work in Section TF-5] has finite moments of all orders. Moreover, the assumption $\|\mu - X\hat{\theta}\|_2 = O_{\mathbb{P}}(\sqrt{n})$ is also quite weak; we only maintain that the average prediction error $\|\mu - X\hat{\theta}\|_2/\sqrt{n}$ is bounded in probability, and not even that it converges to zero.

2.2. *Lasso problems in nonstandard form.* Now consider two lasso problems

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{2}\|y - X\theta\|_2^2 + \lambda\|\theta_2\|_1, \tag{7}$$

$$\min_{\alpha \in \mathbb{R}^p} \frac{1}{2}\|y - Z\alpha\|_2^2 + \lambda'\|\alpha_2\|_1, \tag{8}$$

where the coefficient vectors decompose as $\theta = (\theta_1, \theta_2)$, $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^{p_1+p_2}$, and we partition the columns of the predictor matrices accordingly,

$$X = [X_1\ X_2],\ Z = [Z_1\ Z_2] \in \mathbb{R}^{n \times (p_1+p_2)}.$$

In words, the ℓ_1 penalties in (7), (8) only apply to part of the coefficient vectors. We have the same goal of the last section: to bound $\|X\hat{\theta} - Z\hat{\alpha}\|_2$ in terms of the differences between X and Z.

Simply transforming (7), (8) into standard form lasso problems (by partially solving for θ_1, α_1) will not generically provide a tight bound, because the predictor matrices of the resulting lasso problems have restricted images.[1] Therefore, we must rederive the analogues of Lemmas 1 and 2, and Corollary 1. For the upcoming bounds in Lemma 4 and Corollary 2, it is critical that $X_1 = Z_1$, i.e., the predictor variables corresponding to unpenalized coefficients in (7), (8) are identical. We state the results below, but omit the proofs because they follow essentially the same arguments as those in the last section.

LEMMA 3. *For fixed λ, λ', solutions $\hat{\theta}$ and $\hat{\alpha}$ of lasso problems* (7) *and* (8) *satisfy*

$$\frac{1}{2}\|X\hat{\theta} - Z\hat{\alpha}\|_2^2 \leq \langle y - X\hat{\theta}, Z\hat{\alpha}\rangle + (\lambda' - \lambda)\|\hat{\theta}_2\|_1 - \lambda'\|\hat{\alpha}_2\|_1 + R,$$

where $R = \frac{1}{2}\|(X - Z)\hat{\theta}\|_2^2 + \langle y - X\hat{\theta}, (X - Z)\hat{\theta}\rangle$.

LEMMA 4. *Consider a sequence of problems* (7), (8) *(all quantities $p_1, p_2, y, X, Z, \lambda, \lambda'$ considered as functions of n), such that $X_1 = Z_1$ and $\lambda' = (1 + \delta)\lambda$ for some fixed $\delta > 0$. Assume that*

$$\sqrt{p_2}\|X_2 - Z_2\|_\infty\|\hat{\theta}_2\|_1 = O\Big(\sqrt{\lambda\|\hat{\theta}_2\|_1}\Big) \quad \text{and} \quad \frac{\sqrt{p_2}\|X_2 - Z_2\|_\infty\|y - X\hat{\theta}\|_2}{\lambda} \to 0 \quad \text{as } n \to \infty.$$

Then any solutions $\hat{\theta}, \hat{\alpha}$ of (2), (3) *satisfy*

$$\|X\hat{\theta} - Z\hat{\alpha}\|_2 = O\Big(\sqrt{\lambda\|\hat{\theta}_2\|_1}\Big).$$

[1] In more detail, solving for $X_1\hat{\theta}_1 = P_{X_1}(y - X_2\hat{\theta}_2)$ and $Z_1\hat{\theta}_1 = P_{Z_1}(y - Z_2\hat{\alpha}_2)$ [where $P_A = A(A^TA)^+A^T$ denotes the projection matrix onto $\mathrm{col}(A)$, the column space of a matrix A], yields "new" standard form lasso problems with predictor matrices $(I - P_{X_1})X_2$ and $(I - P_{Z_1})Z_2$. This is problematic because we require a bound in X_2 and Z_2.

COROLLARY 2. *Consider again a sequence of lasso problems* (7), (8) *with* $X_1 = Z_1$ *and* $\lambda' = (1+\delta)\lambda$ *for some fixed* $\delta > 0$. *Assume that the outcome vector* y *is drawn from the model*

$$y = \mu + \epsilon,$$

where $\epsilon_1, \ldots \epsilon_n$ *are i.i.d. with* $\mathbb{E}[\epsilon_i^4] < \infty$, *and assume that* $\|\mu - X\hat{\theta}\|_2 = O_{\mathbb{P}}(\sqrt{n})$. *Further assume*

$$\sqrt{p_2}\|X_2 - Z_2\|_\infty \|\hat{\theta}_2\|_1 = O_{\mathbb{P}}\Big(\sqrt{\lambda\|\hat{\theta}_2\|_1}\Big) \quad \text{and} \quad \sqrt{np_2}\|X_2 - Z_2\|_\infty/\lambda \to 0 \quad \text{as } n \to \infty. \tag{9}$$

Then any solutions $\hat{\theta}, \hat{\alpha}$ *of* (7), (8) *satisfy*

$$\|X\hat{\theta} - Z\hat{\alpha}\|_2 = O_{\mathbb{P}}\Big(\sqrt{\lambda\|\hat{\theta}_2\|_1}\Big).$$

3. Convergence of trend filtering and locally adaptive regression spline basis matrices.

LEMMA 5. *For any integer* $k \geq 0$, *consider the matrices* $G_2, H_2 \in \mathbb{R}^{n\times(n-k-1)}$, *the last* $n-k-1$ *columns of the trend filtering and locally adaptive regression spline basis matrices in (TF-22), (TF-27). With evenly spaced inputs on* $[0,1]$, $\{x_1, x_2, \ldots x_n\} = \{1/n, 2/n, \ldots 1\}$, *we have*

$$\|G_2 - H_2\|_\infty = O(1/n).$$

PROOF. For $k = 0, 1$ the result is vacuous, as $G_2 = H_2$ according to Lemma TF-3 in Section TF-3. Hence we assume $k \geq 2$, and without a loss of generality, we assume that k is even, since the case for k odd follows from essentially the same arguments. By Lemma 6, for large enough n,

$$\|G_2 - H_2\|_\infty = \frac{1}{n^k}\left(\prod_{i=1}^{k}(n-1-i) - \left(n-1-\frac{k+2}{2}\right)^k\right),$$

i.e., for large enough n,

$$n\|G_2 - H_2\|_\infty = \underbrace{\frac{\left(n-1-\frac{k+2}{2}\right)^k}{n^k}}_{a_n} \cdot \underbrace{n\left(\frac{\prod_{i=1}^{k}(n-1-i)}{\left(n-1-\frac{k+2}{2}\right)^k} - 1\right)}_{b_n},$$

We investigate the convergence of the sequence $a_n \cdot b_n$ as defined above. It is clear that $a_n \to 1$ as $n \to \infty$. Hence it remains to bound b_n.

To this end, consider the term

$$\prod_{i=1}^{k}(n-1-i) = \frac{(n-1)!}{(n-k-1)!}.$$

We use Stirling's approximation to both the numerator and denominator, writing

$$\frac{(n-1)!}{(n-k-1)!} = \underbrace{\frac{(n-1)! \Big/ \left((n-1)^{n-1/2} e^{-n+1}\sqrt{2\pi}\right)}{(n-k-1)! \Big/ \left((n-k-1)^{n-k-1/2} e^{-n+k+1}\sqrt{2\pi}\right)}}_{c_n} \cdot \frac{(n-1)^{n-1/2}}{(n-k-1)^{n-k-1/2}} \cdot e^{-k}.$$

Therefore

$$\begin{aligned}\frac{\prod_{i=1}^k (n-1-i)}{\left(n-1-\frac{k+2}{2}\right)^k} &= c_n \cdot \frac{\left((n-1)\,/\left(n-1-\frac{k+2}{2}\right)\right)^{n-1/2}}{\left((n-k-1)\,/\left(n-1-\frac{k+2}{2}\right)\right)^{n-k-1/2}} \cdot e^{-k} \\ &= c_n \cdot \underbrace{\frac{\left(1+\frac{(k-2)/2}{n-k-1}\right)^{n-k-1/2} e^{-(k-2)/2}}{\left(1-\frac{(k+2)/2}{n-1}\right)^{n-1/2} e^{(k+2)/2}}}_{d_n}.\end{aligned}$$

At this point, we have expressed $b_n = n(c_n d_n - 1)$. Note that $c_n \to 1$ by Stirling's formula, and $d_n \to 1$ by the well-known limit for e^x,

$$e^x = \lim_{t\to\infty} \left(1+\frac{x}{t}\right)^t. \tag{10}$$

The question is of course how fast these two sequences c_n, d_n converge; if the remainder $c_n d_n - 1$ is $O(1/n)$, then $b_n = O(1)$ and indeed $\|G_2 - H_2\|_\infty = O(1/n)$.

First we address c_n. It is known that Stirling's approximation satisfies [e.g., see Nemes (2010)]

$$\frac{n!}{n^{n+1/2}e^{-n}\sqrt{2\pi}} = e^{\gamma_n}, \quad \text{where } \frac{1}{12n+1} \le \gamma_n \le \frac{1}{12n}.$$

Hence

$$c_n = \exp(\gamma_{n-1} - \gamma_{n-k-1}) \le \exp\left(\frac{1}{12(n-1)}\right).$$

Next we address d_n. Lemma 7 derives the following error bound for the exponential limit in (10):

$$\left(1+\frac{x}{n}\right)^n e^{-x} = e^{\delta_{x,n}}, \quad \text{where } \frac{-x^2}{n+x} \le \delta_{x,n} \le 0,$$

for sufficiently large n. Therefore

$$\begin{aligned} d_n &= \exp\left(\delta_{(k-2)/2,n-k-1} - \delta_{-(k+2)/2,n-1}\right) \cdot \left(\frac{1+\frac{(k-2)/2}{n-k-1}}{1-\frac{(k+2)/2}{n-1}}\right)^{1/2} \\ &= \exp\left(\delta_{(k-2)/2,n-k-1} - \delta_{-(k+2)/2,n-1}\right) \cdot \left(\frac{n-1}{n-k-1}\right)^{1/2} \\ &\le \exp\left(\frac{(k+2)^2}{4n-2(k+4)}\right) \cdot \left(\frac{n-1}{n-k-1}\right)^{1/2}. \end{aligned}$$

We can simplify, for large enough n,

$$\frac{1}{12(n-1)} + \frac{(k+2)^2}{4n - 2(k+4)} \leq \frac{(k+2)^2}{n},$$

and putting this all together, we have

$$b_n = n(c_n d_n - 1) \leq n \left(\exp\left(\frac{(k+2)^2}{n}\right) \cdot \left(\frac{n-1}{n-k-1}\right)^{1/2} - 1 \right).$$

An application of l'Hôpital's rule shows that the bound on the right-hand hand side above converges to a positive constant. This completes the proof. □

LEMMA 6. *Let $k \geq 2$. If k is even, then*

$$\|G_2 - H_2\|_\infty = \frac{1}{n^k} \left(\prod_{i=1}^{k} (n-1-i) - \left(n - 1 - \frac{k+2}{2}\right)^k \right),$$

for sufficiently large n; if k is odd, then

$$\|G_2 - H_2\|_\infty = \frac{1}{n^k} \left(\prod_{i=1}^{k} (n-1-i) - \left(n - 1 - \frac{k+1}{2}\right)^k \right),$$

for sufficiently large n.

PROOF. We prove the result for k even; the proof for k odd is similar. Consider first the sequence

$$a_n = \prod_{i=1}^{k} (n-i) - \left(n - \frac{(k+2)}{2}\right)^k.$$

Note that

$$a_n = \left(\frac{k(k+2)}{2} - \frac{k(k+1)}{2} \right) \cdot n^{k-1} + O(n^{k-2}).$$

Because the coefficient of the leading term is positive, we know that $a_n \to \infty$ as $n \to \infty$; further, for large enough n, this convergence is monotone (since a_n is polynomial in n). Now recall that G_2, H_2 are the last $n-k-1$ columns of G, H, in (TF-22), (TF-27). Hence

$$(H_2 - G_2)_{ij} = \frac{1}{n^k} \cdot \begin{cases} 0 & \text{if } i \leq j + (k+2)/2 \\ -(i - j - (k+2)/2)^k & \text{if } j + (k+2)/2 < i \leq j + k \\ a_{i-j} & \text{if } i > j + k, \end{cases}$$

given by taking $j+k+1$ in place of j in (TF-22), (TF-27). For $i - j \leq k$, the term $(i - j - (k+2)/2)^k$ is bounded by k^k. And since $a_n \uparrow \infty$, as argued above, we conclude that $\|G_2 - H_2\|_\infty = a_{n-1}/n^k$ for sufficiently large n. □

LEMMA 7. *For any $x \in \mathbb{R}$, and for sufficiently large $t > 0$,*

$$\left(1 + \frac{x}{t}\right)^t e^{-x} = e^{\delta_{x,t}}, \quad \textit{where } \frac{-x^2}{t+x} \leq \delta_{x,t} \leq 0.$$

PROOF. Define $f(t) = (1 + x/t)^t$. Consider

$$\log f(t) = t \cdot (\log(t + x) - \log t),$$

which is well-defined as long as $t \geq \max\{-x, 0\}$. Because log is a concave function, its tangent line is a global overestimate of the function, that is,

$$\log(t + x) \leq \log t + \frac{1}{t} \cdot x,$$

which means that $t \cdot (\log(t + x) - \log t) \leq x$, i.e., $f(t) \leq e^x$. Hence $f(t)e^{-x} = e^{\delta_{x,t}}$ where $\delta_{x,t} \leq 0$. The lower bound on $\delta_{x,t}$ follows similarly. Again by concavity,

$$\log t \leq \log(t + x) + \frac{1}{t + x} \cdot (-x),$$

so $\log(t+x)-\log t \geq x/(t+x)$, and $f(t) \geq \exp(tx/(t+x))$. Therefore $f(t)e^{-x} \geq \exp(tx/(t+x) - x) = \exp(-x^2/(t + x))$. □

4. Unevenly spaced inputs. Our implicit assumption with trend filtering has been that the inputs $x_1, \ldots x_n$ are evenly spaced. [We have been writing this assumption as $x_i = i/n$ for $i = 1, \ldots n$, but really, it is only the spacings that matter; for a common spacing of $d > 0$ between inputs, if we wanted to compare the trend filtering problem in (TF-2) with, say, the locally adaptive regression spline problem in (TF-13) across λ values, then we would simply replace the factor of n^k in (TF-2) by $1/d^k$.] How could we extend the trend filtering criterion in (TF-2) to account for arbitrarily spaced inputs? One nice feature of the continuous-time representation of trend filtering in (TF-29) is that it provides a natural answer to this question.

For arbitrary input points $x_1 < x_2 < \ldots < x_n$, consider defining the basis matrix H as in (TF-25), (TF-26), and defining the trend filtering estimate by the fitted values $H\hat{\alpha}$ of the problem in (TF-23). Aside from its connection to the continuous-time representation, this definition is supported by the fact that the trend filtering estimates continue to match those from locally adaptive regression splines for polynomial orders $k = 0$ or 1, as they did in the evenly spaced input case. [This follows from the fact that for $k = 0$ or 1, the basis functions $h_1, \ldots h_n$ defined in (TF-25) match the truncated power basis functions $g_1, \ldots g_n$ in (TF-19), with knots as in (TF-12).] Let us write the trend filtering basis matrix as $H^{(x)}$ to emphasize its dependence on the inputs $x_1, \ldots x_n$. To express the fitted values $H^{(x)}\hat{\alpha}$ in a more familiar form, we seek a matrix $D^{(x,k+1)} \in \mathbb{R}^{(n-k-1)\times n}$ so that the estimate

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^n} \frac{1}{2}\|y - \beta\|_2^2 + \frac{1}{k!} \cdot \lambda \|D^{(x,k+1)}\beta\|_1 \tag{11}$$

satisfies $\hat{\beta} = H^{(x)}\hat{\alpha}$. Note that $D^{(x,k+1)}$ is given precisely by the last $n - k - 1$ rows of $(H^{(x)})^{-1}/k!$. For $k = 0$, it is easy to see that $D^{(x,1)} = D^{(1)}$, the first difference matrix (for evenly spaced inputs) given in (TF-3). For $k \geq 1$, an inductive calculation (similar to the proofs of Lemmas TF-2, TF-4) shows that

$$D^{(x,k+1)} = D^{(1)} \cdot \operatorname{diag}\left(\frac{k}{x_{k+1} - x_1}, \frac{k}{x_{k+2} - x_2}, \ldots \frac{k}{x_n - x_{n-k}}\right) \cdot D^{(x,k)}. \tag{12}$$

[The leading $D^{(1)}$ above is the $(n-k-1) \times (n-k)$ version of the first difference matrix in (TF-3).] This parallels the definition of the $(k+1)$st difference matrix in (TF-4), and hence (as our notation would suggest), $D^{(x,k+1)}$ can still be thought of as a difference operator of order $k+1$, but adjusted to account for the unevenly spaced inputs $x_1, \ldots x_n$.

In (12), multiplication by the diagonal matrix of spacing weights does not cause any change in structure, so $D^{(x,k+1)}$ is still banded with bandwidth $k+2$. Therefore, in principle, the trend filtering problem in (11) is not any harder computationally than the original problem in (TF-2) for evenly spaced inputs, because algorithms like the primal-dual interior point method of Kim et al. (2009) and the dual path algorithm of Tibshirani & Taylor (2011) only rely on such bandedness for efficient calculations. In practice, too, these algorithms are able to efficiently handle the extension to unevenly spaced input points, in the majority of cases. However, when dealing with inputs that have highly irregular spacings, numerical accuracy can become an issue.[2] Robustifying trend filtering algorithms to handle these difficult cases is a direction for future work.

On the theoretical side, the same strategy used to derive the convergence rates in Section TF-5 for evenly spaced input points can be applied to the unevenly spaced case, as well. Recall that the basic idea was to tie trend filtering estimates together asymptotically with locally adaptive regression splines, at a tight enough rate that trend filtering estimates inherit the (known) convergence rates of the latter estimators. Section 2 of this document provides the technical framework for tying together these two estimators, which can be seen as the fitted values of two lasso problems. The asymptotic bounds between the two estimators, in the current setting, depend primarily on the maximum elementwise difference between $G^{(x)}$, the truncated power basis matrix in (TF-19), (TF-17) evaluated at the inputs $x_1, \ldots x_n$, and $H^{(x)}$, the trend filtering basis matrix in (TF-25), (TF-26) evaluated at the inputs $x_1 \ldots x_n$. We state the following convergence result without proof, since it follows from similar arguments to those in Section TF-5.

THEOREM 1. *Assume that $y \in \mathbb{R}^n$ is drawn from the model (TF-30), with inputs $x_1 < \ldots < x_n \in [0,1]$ and sub-Gaussian errors (TF-31). Assume also that $f_0 \in \mathcal{F}_k(C_n)$, i.e., for a fixed integer $k \geq 0$ and $C_n > 0$ (depending on n), the true function f_0 is k times weakly differentiable and $\mathrm{TV}(f_0^{(k)}) \leq C_n$. Let $\hat{f}$ denote the kth order locally adaptive regression spline estimate in (TF-13) with tuning parameter $\lambda = \Theta(n^{1/(2k+3)} C_n^{-(2k+1)/(2k+3)})$, and let $\hat{\beta}$ denote the kth order trend filtering estimate in (TF-2) with tuning parameter $(1+\delta)\lambda$, for any fixed $\delta > 0$. Finally, if $k \geq 2$, then we must assume that the following conditions are met: C_n does not grow too quickly,*

$$C_n = O(n^{(k+2)/(2k+2)}),$$

[2]Recall that these algorithms iteratively solve linear systems in $D^{(x,k+1)}(D^{(x,k+1)})^T$; each system requires $O(n)$ operations, regardless of the inputs $x_1, \ldots x_n$. However, if $x_1, \ldots x_n$ have highly irregular spacings, with some points being very close together and some quite far apart, then $D^{(x,k+1)}$ can contain both very large and very small elements, which can cause numerical inaccuracies when solving the linear systems.

and the input points $x_1, \dots x_n$ satisfy

$$\max_{i=1,\dots n-1} (x_{i+1} - x_i) = O(n^{-(k+1)/(k(2k+3))} C_n^{-(2k+2)/(k(2k+3))}), \tag{13}$$

$$\max_{i=1,\dots n-k-1} \left| \prod_{\ell=1}^{k} (x_n - x_{i+\ell}) - (x_n - x_{i+\lfloor (k+2)/2 \rfloor})^k \right| = O(1/n). \tag{14}$$

(Above, we use $\lfloor \cdot \rfloor$ to denote the floor function.) Then

$$\frac{1}{n} \sum_{i=1}^{n} \big(\hat{\beta}_i - \hat{f}(x_i)\big)^2 = O_{\mathbb{P}}(n^{-(2k+2)/(2k+3)} C_n^{2/(2k+3)}),$$

and furthermore

$$\frac{1}{n} \sum_{i=1}^{n} \big(\hat{\beta}_i - f_0(x_i)\big)^2 = O_{\mathbb{P}}(n^{-(2k+2)/(2k+3)} C_n^{2/(2k+3)}).$$

Hence, if the kth derivative of f_0 has bounded total variation (i.e., C_n is a constant), then the trend filtering estimate converges to f_0 in probability at the minimax rate.

Remark. The conditions in Theorem 1 should all look familiar to those in Theorems TF-1 and TF-2 from Section TF-5, except for the design conditions (13), (14). The first condition (13) is needed by Mammen & van de Geer (1997) for their convergence result on locally adaptive splines for unevenly spaced inputs,

$$\frac{1}{n} \sum_{i=1}^{n} \big(\hat{f}(x_i) - f_0(x_i)\big)^2 = O_{\mathbb{P}}(n^{-(2k+2)/(2k+3)} C_n^{2/(2k+3)}).$$

In words, condition (13) maintains that no pair of adjacent inputs should be too far apart (and is clearly satisfied for evenly spaced inputs over $[0, 1]$). The second condition (14) is sufficient to imply

$$\|G_2^{(x)} - H_2^{(x)}\|_\infty = O(1/n),$$

where $G_2^{(x)}, H_2^{(x)}$ are the last $n-k-1$ columns of $G^{(x)}, H^{(x)}$ (the locally adaptive regression splines and trend filtering basis matrices, respectively). This enables us to apply Corollary 2 in Section 2 to bound the difference between the trend filtering and locally adaptive regression spline estimates. (Recall that Lemma 5 in Section 3 establishes the above condition for evenly spaced inputs.) At this point, assuming the condition (14) in Theorem 1 seems to be the most explicit way of ensuring the bound between $G^{(x)}, H^{(x)}$ that is needed for Corollary 2, but we feel that this condition can be made even more explicit, and likely, simplified. This is left for future work.

References.

Kim, S.-J., Koh, K., Boyd, S. & Gorinevsky, D. (2009), 'ℓ_1 trend filtering', *SIAM Review* **51**(2), 339–360.

Mammen, E. & van de Geer, S. (1997), 'Locally apadtive regression splines', *Annals of Statistics* **25**(1), 387–413.

Nemes, G. (2010), 'On the coefficients of the asymptotic expansion of $n!$', *Journal of Integer Sequences* **13**(6), 5.

Tibshirani, R. J. & Taylor, J. (2011), 'The solution path of the generalized lasso', *Annals of Statistics* **39**(3), 1335–1371.