Transactions of A. Razmadze
Mathematical Institute
Vol. **174** (2020), issue 2, 217–233

LOGISTIC REGRESSION WITH TOTAL VARIATION REGULARIZATION

SARA VAN DE GEER

A paper devoted to the 75th *birthday of Estate Khmaladze*

Abstract. We study logistic regression with total variation penalty on the canonical parameter and show that the resulting estimator satisfies a sharp oracle inequality: the excess risk of the estimator is adaptive to the number of jumps of the underlying signal or an approximation thereof. In particular, when there are finitely many jumps, and jumps up are sufficiently separated from jumps down, then the estimator converges with a parametric rate up to a logarithmic term $\log n/n$, provided the tuning parameter is chosen appropriately of order $1/\sqrt{n}$. Our results extend earlier results for quadratic loss to logistic loss. We do not assume any a priori known bounds on the canonical parameter, but instead only make use of the local curvature of the theoretical risk.

1. Introduction

In this paper we consider logistic regression with a total variation penalty on the canonical parameter. Total variation based de-noising was introduced in [15]. Our aim here is to develop theoretical results that show that the estimator adapts to the number of jumps in the signal.

For $i = 1, \ldots, n$, let $Y_i \in \{0, 1\}$ be independent binary observations. Write the unknown probability of success as $\theta_i^0 := P(Y_i = 1)$, and let $f_i^0 := \log(\theta_i^0/(1-\theta_i^0))$ be the log-odds ratio, $i = 1, \ldots, n$. Define the total variation of a vector $f \in \mathbb{R}^n$ as

$$\mathrm{TV}(f) := \sum_{i=2}^n |f_i - f_{i-1}|.$$

We propose to estimate the unknown vector f^0 of log-odds ratios applying logistic regression with total variation regularization. The estimator is

$$\hat{f} := \arg \min_{f \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^n \biggl(-Y_i f_i + \log(1 + \mathrm{e}^{f_i}) \biggr) + \lambda \mathrm{TV}(f) \right\}.$$

Our goal is to derive oracle inequalities for this estimator. The approach we take shares some ideas with [4, 11] and [13]. These papers deal with least squares loss, whereas the current paper studies logistic loss. Moreover, instead of using the projection arguments of the previous mentioned papers, we use entropy bounds. This allows us to remove a redundant logarithmic term: we show that the excess risk of estimator $\hat{f}$ converges under certain conditions with rate $(s + 1) \log n/n$, where s is the number of jumps of f^0 or of an oracle approximation thereof (see Theorem 2.1). This extends the result in [6] to logistic loss and to a sharp oracle inequality.

To arrive at the results of this paper we require that $\|\hat{f}\|_\infty$ stays bounded with high probability. In Theorem 3.1 we show that this requirement holds assuming that both $\|f^0\|_\infty$ and $\mathrm{TV}(f^0)$ remain bounded.

The theory for a total variation regularization with the least squares loss (the fused Lasso) has been developed in a series of papers [4, 8, 14, 16, 17, 21, 22] including higher dimensional extensions [3, 5, 7, 12] and higher order total variation [6, 13, 18, 19].

Logistic regression with ℓ_1-regularization has many applications. When there are co-variables, the penalty is on the total variation of the coefficients. In [25], logistic regression with the fused Lasso is applied to spectral data, and in [9] to gene expression data, whereas [1] applies it to time-varying

2020 *Mathematics Subject Classification.* 62J07, 62G05.

Key words and phrases. Logistic regression; Oracle inequality; Total variation penalty.

networks. In [20] the penalty alternatively takes links between variables into account using a quadratic penalty. The papers [26] and [10] present algorithms for the fused Lasso. In [2], a Bayesian approach with the fused Lasso is presented.

This paper is organized as follows. In Section 2 we state the oracle inequality for $\hat{f}$ (Theorem 2.1). Section 3 derives a bound for $\|\hat{f}\|_\infty$ (Theorem 3.1). The remainder of the paper is devoted to proofs. Section 4 states some standard tools to this end, Section 5 contains a proof of Theorem 2.1 and Section 6 a proof of Theorem 3.1.

2. A Sharp Oracle Inequality

The empirical risk in this paper is given by the normalized minus log-likelihood

$$R_n(f) := \frac{1}{n}\sum_{i=1}^n \biggl(-Y_i f_i + \log(1+\mathrm{e}^{f_i}) \biggr),\ f \in \mathbb{R}^n.$$

The theoretical risk is

$$R(f) := \mathbb{E} R_n(f),\ f \in \mathbb{R}^n$$

and $R(f) - R(f^0)$ is called the "excess risk". For $f \in \mathbb{R}^n$, we write $\dot{R}_n(f) := \partial R_n(f)/\partial f$ and $\dot{R}(f) := \mathbb{E}\dot{R}_n(f)$. These are column vectors in $\mathbb{R}^n$. Most of the arguments that follow go through for general convex differentiable loss functions. We do use, however, that or all $f \in \mathbb{R}^n$, $R_n(f) - R(f) = -\epsilon^T f/n$ where $\epsilon = Y - \mathbb{E}Y$ is the noise. In other words, f is the canonical parameter. In the case where the entries of the response vector Y are in $\{0,1\}$, the entries of a noise vector ϵ are bounded by 1. More generally, our theory would need that ϵ has sub-exponential entries. To avoid digressions, we simply restrict ourselves to logistic loss.

Fix a vector $\mathbf{f} \in \mathbb{R}^n$. This vector will play the role of the "oracle" as we will see in Theorem 2.1. We let $S := \{t_1, \ldots, t_s\}$ $(1 < t_1 < \cdots < t_s < n)$ be the location of its jumps:

$$\mathbf{f}_1 = \cdots = \mathbf{f}_{t_1 - 1} \neq \mathbf{f}_{t_1} = \cdots = \mathbf{f}_{t_2 - 1} \neq \mathbf{f}_{t_2} \cdots \mathbf{f}_{t_s - 1} \neq \mathbf{f}_{t_s} = \cdots = \mathbf{f}_n.$$

Let $d_j := t_j - t_{j-1}$ be the distance between jumps, $j = 1, \ldots, r$, where $r = s+1$, $t_r := n+1$ and $t_0 = 1$. Define $d_{\max} := \max_{1 \le j \le r} d_j$.

The quantities Δ_n^2, $\delta_n^2(t)$, $\lambda_n(t)$ and $\Gamma_n^2(t)$ we are about to introduce all depend on $\mathbf{f}$ although we do not express this in our notation. Moreover, being non-asymptotic, these quantities are somewhat involved. After the explicit expressions for Δ_n^2, $\delta_n^2(t)$ and $\lambda_n(t)$ we will give their asymptotic order of magnitude. The asymptotic order of magnitude for $\Gamma_n^2(t)$ depends on the situation. We discuss a special case after the statement of Theorem 2.1.

We let

$$\Delta_n^2 := \frac{4\sum_{j \in [1:r]:\ d_j \ge 1} (\log(d_j - 1) + 1)}{n} + \frac{s}{n},$$

and define for $t > 0$

$$\begin{aligned} \delta_n^2(t) := & \biggl(4\nu A_0 \Delta_n + 8\sqrt{\frac{1+t+\log(3+2\log_2 n)}{n}} \biggr)^2 \\ & + \biggl(\frac{2}{\nu} + 4\sqrt{\frac{A_0 \Delta_n}{n}} + \frac{4\sqrt{1+t+\log(3+2\log_2 n)}}{n} \biggr) \\ & \times \biggl(\Delta_n + 2\sqrt{\frac{s}{n}} \biggr)^2, \end{aligned}$$

and

$$\lambda_n(t) := \frac{1}{\sqrt{n}} \biggl(\frac{4}{\nu} + 8\sqrt{\frac{A_0 \Delta_n}{n}} + \frac{8\sqrt{1+t+\log(3+2\log_2 n)}}{n} \biggr).$$

One can see that

$$\Delta_n^2 = \mathcal{O}\biggl(\frac{(s+1)\log n}{n} \biggr).$$

Furthermore, for $\nu = 1$ (say) and each fixed t

$$\delta_n^2(t) = \mathcal{O}\bigg(\frac{(s+1)\log n}{n}\bigg),\ \lambda_n(t) = \mathcal{O}\bigg(\frac{1}{\sqrt{n}}\bigg),$$

assuming $n^{-1}\sqrt{(s+1)\log n/n} = \mathcal{O}(1)$ which is certainly true under the standard sparsity assumption $(s+1)\log n/n = o(1)$.

The quantity $\delta_n^2(t)$ will be a part of the bound for the excess risk of $\hat{f}$, and $\lambda_n(t)$ can be thought of as the "noise level" to be overruled by the penalty (see Theorem 2.1). The constant A_0 is the (universal) constant appearing when bounding the entropy of the class of functions with both $\|\cdot\|_\infty$ and $\mathrm{TV}(\cdot)$ bounded by 1 (see Lemma 4.3). The free parameter $t > 0$ determines the confidence level of our statements. Both $\delta_n(t)$ and $\lambda_n(t)$ depend on a further free parameter $\nu > 0$ which we do not express in our notation as one can simply choose $\nu = 1$. It is, however, an option to choose ν larger than 1, possibly growing with n: larger ν relaxes the requirement on the tuning parameter λ, but results in larger bounds for the excess risk.

Finally, we present a bound $\Gamma_n^2(t)$ for the so-called "effective sparsity" as introduced in [13] (see also Definition 5.1). The effective sparsity may be seen as a substitute for the sparsity, which is defined as the number of active parameters of the oracle, which is $s+1$. The effective sparsity will, in general, be larger, than $s+1$. Without going into details, we remark that this is due to correlations in the dictionary X when writing $f = Xb$, with dictionary $X \in \mathbb{R}^{n\times n}$ and coefficients $b_1 := f_1$, $b_k := f_k - f_{k-1}$, $k \in [2:n]$.

Let $q_{t_j} := \mathrm{sign}(\mathbf{f}_{t_j})$, $j = 1,\dots,s$. We write $J_{\text{monotone}} := \{2 \le j \le s :\ q_{t_{j-1}} = q_{t_j}\}$ and $J_{\text{change}} := [1:r]\backslash J_{\text{monotone}}$. Thus J_{monotone} are jumps with the same sign as the previous one, and J_{change} are jumps that change sign. We count the first jump as well as the endpoint $t_r = n+1$ as a sign change. Our bound for the effective sparsity is now

$$\Gamma_n^2(t) := \frac{\lambda_n^2(t)}{\lambda^2}\sum_{j\in J_{\text{monotone}}} 8(\log(d_j)+1) + \sum_{j\in J_{\text{change}}} \frac{8n(\log(d_j)+2)}{d_j}.$$

The following theorem presents an oracle inequality for $\hat{f}$. Its proof can be found in Section 5.

Theorem 2.1. *Let $\mathcal{F}$ be a convex subset of $\mathbb{R}^n$ (possibly $\mathcal{F} = \mathbb{R}^n$) and*

$$\hat{f} := \arg\min_{f\in\mathcal{F}}\bigg\{R_n(f) + \lambda\mathrm{TV}(f)\bigg\}.$$

Assume $\mathbf{f} \in \mathcal{F}$ satisfies $\|\mathbf{f}\|_\infty \le B$ for some constant B and define

$$\kappa := \frac{(1+\mathrm{e}^B)^2}{\mathrm{e}^B}.$$

Take

$$\lambda \ge \lambda_n(t)\sqrt{\frac{d_{\max}}{2n}}.$$

Then with probability at least $\mathbb{P}(\|\hat{f}\|_\infty \le B) - \exp[-t]$, we have

$$R(\hat{f}) - R(\mathbf{f}) \le 4\kappa\delta_n^2(t) + \frac{\lambda^2}{4}\Gamma_n^2(t).$$

Keeping the constant B fixed, this theorem tells us that

$$R(\hat{f}) - R(\mathbf{f}) = \mathcal{O}_{\mathbf{P}}\bigg(\frac{\sum_{j=1}^r(\log(d_j)+1)}{n} + \lambda^2\Gamma_n^2\bigg),$$

where we recall that $r = s+1$. If the jumps of $\mathbf{f}$ are roughly equidistant we see that $d_j \asymp d_{\max} \sim n/r$. Taking $\lambda \asymp \lambda_n(t)/\sqrt{r} \asymp \sqrt{1/(nr)}$, the bound for the effective sparsity $\Gamma_n^2(t)$ is in the worst case (where the jumps of $\mathbf{f}$ have alternating signs) of order $r^2\log(n/r)$. In other words, in that case the rate is $R(\hat{f}) - R(\mathbf{f}) = \mathcal{O}_{\mathbf{P}}(r\log(n/r)/n)$, which for least squares loss is the minimax rate (see [8]).

If $\mathbf{f}$ is monotone, with $\lambda \asymp \sqrt{d_{\max}}/n$, we get

$$\lambda^2 \Gamma_n^2 \asymp \frac{\sum_{j=2}^s \log(d_j) + 1}{n} + \frac{1}{n}\biggl(\frac{\log(d_1) d_{\max}}{d_1} + \frac{\log(d_r) d_{\max}}{d_r} \biggr).$$

In other words, the first jump of $\mathbf{f}$ should not occur to early, and the last jump not too late, relative to the distance between the jumps.

We note that the choice $\lambda \asymp \lambda_n(t)\sqrt{d_{\max}/n}$ depends on the oracle $\mathbf{f}$. Thus, if the tuning parameter λ is given, the choice of $\mathbf{f}$ depends on λ.

We assumed that $\|\mathbf{f}\|_\infty \le B$. We do not assume $\|f^0\|_\infty$ to be bounded by the same constant B, but we do hope for a good approximation $\mathbf{f}$ of f^0 with $\|\mathbf{f}\|_\infty \le B$. Nevertheless, Theorem 2.1 presents a sharp oracle inequality directly comparing $R(\hat{f})$ with $R(\mathbf{f})$: it does not require that the excess risk $R(\mathbf{f}) - R(f^0)$ is small in any sense. In the same spirit, the theorem requires that $\|\hat{f}\|_\infty \le B$ with high probability. This can be accomplished by taking $\mathcal{F} := \{f \in \mathbb{R}^n : \ \|f\|_\infty \le B\}$ (or some convex subset thereof). Theorem 2.1 holds for any B, i.e., it is a free parameter. However, one may not want to force $\hat{f}$ to be bounded by a given constant, but let the data decide for a bound on $\hat{f}$. This is a reason why we establish Theorem 3.1 given in the next section.

3. Showing that $\|\hat{f}\|_\infty$ is Bounded (Instead of Assuming this)

Since f^0 minimizes $R(f)$, a two-term Taylor expansion around f^0 gives

$$R(f) - R(f^0) = \frac{1}{2}(f - f^0)^T \ddot{R}(\tilde{f})(f - f^0)$$

where $\tilde{f}_i$ lies between f_i and f_i^0, $i = 1, \ldots, n$. It follows that

$$R(f) - R(f_0) \ge \frac{1}{2K_f^2} \|f - f^0\|_{Q_n}^2,$$

where

$$\| \cdot \|_{Q_n} = \| \cdot \|_2 / \sqrt{n}$$

and where (for logistic loss)

$$K_f^2 := \frac{(1 + \mathrm{e}^{\|f\|_\infty \vee \|f^0\|_\infty})^2}{\mathrm{e}^{\|f\|_\infty \vee \|f^0\|_\infty}}.$$

Thus, if both $\|f\|_\infty$ and $\|f^0\|_\infty$ stay within the bounds, we have a standard quadratic curvature of $R(\cdot)$ at f^0. Otherwise, the constant K_f grows exponentially fast. We will therefore assume that $\|f^0\|_\infty$ stays bounded and our task is then to show that $\|\hat{f}\|_\infty$ stays bounded, as well. The following theorem (where we have not been very careful with the constants) is derived in Section 6.

Theorem 3.1. *Let* $\mathrm{TV}(f^0) \le M_0$ *for some constant* $M_0 \ge 1$. *Define*

$$K := \frac{(1 + \mathrm{e}^{1 + 2^4 M_0 + \|f^0\|_\infty})^2}{\mathrm{e}^{1 + 2^4 M_0 + \|f^0\|_\infty}}.$$

Suppose

$$\begin{aligned} \lambda \le & \Bigl(2^4 (2K^2) M_0 \Bigr)^{-1}, \\ \lambda \ge & 2^8 n^{-2/3} A_0^{2/3} (2K^2)^{1/3}, \\ \lambda \ge & 2^8 (2K^2) \frac{1+t}{n}, \end{aligned}$$

where the last inequality holds for some $t > 0$, *and where in the second last inequality* A_0 *is the constant appearing when bounding the entropy of the class of functions with both* $\| \cdot \|_\infty$ *and* $\mathrm{TV}(\cdot)$ *bounded by* 1 (*see Lemma* 4.3). *Then with probability at least* $1 - \exp[-t]$ *it holds that*

$$\frac{\|\hat{f} - f^0\|_{Q_n}^2}{2K^2} + \lambda \mathrm{TV}(\hat{f} - f^0) \le 4\lambda M_0$$

and

$$\|\hat{f} - f^0\|_\infty \leq \frac{1 + 8M_0}{2}.$$

One may object that the conditions on the tuning parameter λ depend on f^0 via the bounds on $\|f^0\|_\infty$ and $\mathrm{TV}(f^0)$. On the other hand, the choice of λ in Theorem 2.1 will be of larger order than $n^{-2/3}$ if one aims at adaptive results, and it will need to tend to zero. For such λ and for $\|f^0\|_\infty$ and $\mathrm{TV}(f^0)$ remaining bounded, the conditions of Theorem 3.1 will be met for all n sufficiently large.

4. Some Standard Results Useful for Both Theorem 2.1 and Theorem 3.1

Lemma 4.1. *For all vectors $g \in \mathbb{R}^n$, we have*

$$\mathbb{P}(\epsilon^T g \geq \|g\|_2 \sqrt{2t}) \leq \exp[-t], \ \forall \ t > 0.$$

Proof. The entries in ϵ have mean zero, are bounded by 1, and are independent. This means we can apply Hoeffding's inequality to $\epsilon^T g/n$. □

For $\mathbf{Q}$ a probability measure on $\{1, \ldots, n\}$ and a set $\mathcal{G} \subset \mathbb{R}^n$ we let $H(\cdot, \mathcal{G}, \mathbf{Q})$ be the entropy[1] of $\mathcal{G}$ endowed with the metric induced by the $L_2(\mathbf{Q})$-norm

Lemma 4.2. *Let $\mathcal{G} \subset \mathbb{R}^n$ be a set with diameter*

$$R := \sup_{g \in \mathcal{G}} \|g\|_{Q_n}.$$

Suppose

$$J(R) := 2 \int_0^R \sqrt{2H(u, \mathcal{G}, Q_n)} du$$

exists. Then for all $t > 0$, with probability at least $1 - \exp[-t]$ it holds that

$$\sup_{g \in \mathcal{G}} \epsilon^T g/n \leq \frac{J(R)}{\sqrt{n}} + 4R\sqrt{\frac{1+t}{n}}.$$

Proof. We can apply Hoeffding's inequality to $\epsilon^T g/n$ for each g fixed (see Lemma 4.1). The result of the current lemma is thus essentially applying Dudley's entropy integral. The constants are taken from Theorem 17.3 in [23]. □

Lemma 4.3. *Let $\mathcal{G} := \{g \in \mathbb{R}^n : \ \|g\|_\infty \leq 1, \ \mathrm{TV}(g) \leq 1\}$. It holds for any probability measure $\mathbf{Q}$*

$$H(u, \mathcal{G}, \mathbf{Q}) \leq \frac{A_0}{u} \ \forall \ u > 0,$$

where A_0 is a universal constant.

Proof. See [24], Theorem 2.7.5. □

5. Proof of Theorem 2.1.

5.1. The main body of the proof of Theorem 2.1.

The following lemma is Lemma 7.1 in [23]. We present a proof for completeness.

Lemma 5.1. *Let $\mathcal{F}$ be a convex subset of $\mathbb{R}^n$ (possibly $\mathcal{F} = \mathbb{R}^n$) and*

$$\hat{f} := \arg\min_{f \in \mathcal{F}} \Big\{ R_n(f) + \lambda \mathrm{TV}(f) \Big\}.$$

Then for all $f \in \mathcal{F}$,

$$-\dot{R}_n(\hat{f})^T (f - \hat{f}) \leq \lambda \mathrm{TV}(f) - \lambda \mathrm{TV}(\hat{f}).$$

[1] For $u > 0$ the u-covering number $N(u)$ of a metric space $(\mathcal{V}, d)$ is the smallest N such that there exists $\{v_j\}_{j=1}^N \subset \mathcal{V}$ with $\sup_{v \in \mathcal{V}} \min_{1 \leq j \leq N} d(v, v_j) \leq u$. The entropy is $H(\cdot) := \log N(\cdot)$.

Proof of Lemma 5.1. Define, for $0 < \alpha < 1$, $\hat{f}_\alpha := (1-\alpha)\hat{f} + \alpha f$. Then, using the convexity of $\mathcal{F}$

$$\begin{aligned} R_n(\hat{f}) + \lambda \mathrm{TV}(\hat{f}) \le & R_n(\hat{f}_\alpha) + \lambda \mathrm{TV}(\hat{f}_\alpha) \\ = & R_n(\hat{f}_\alpha) + (1-\alpha)\lambda \mathrm{TV}(\hat{f}) + \alpha \lambda \mathrm{TV}(f). \end{aligned}$$

Thus,

$$\frac{R_n(\hat{f}) - R_n(\hat{f}_\alpha)}{\alpha} \le \lambda \mathrm{TV}(f) - \lambda \mathrm{TV}(\hat{f}).$$

The result now follows by letting $\alpha \downarrow 0$. □

Lemma 5.2. *Let $\mathcal{F}$ be a convex subset of $\mathbb{R}^n$ and*

$$\hat{f} := \arg\min_{f \in \mathcal{F}} \biggl\{ R_n(f) + \lambda \mathrm{TV}(f) \biggr\}.$$

Then for all $f \in \mathcal{F}$,

$$R(\hat{f}) - R(f) + \mathrm{rem}(f, \hat{f}) \le \epsilon^T(\hat{f} - f)/n + \lambda \mathrm{TV}(f) - \lambda \mathrm{TV}(\hat{f}),$$

where

$$\mathrm{rem}(f, \hat{f}) = R(f) - R(\hat{f}) - \dot{R}(\hat{f})^T(f - \hat{f}).$$

Proof of Lemma 5.2. By Lemma 5.1,

$$-\dot{R}_n(\hat{f})^T(f - \hat{f}) \le \lambda \mathrm{TV}(f) - \lambda \mathrm{TV}(\hat{f}).$$

So,

$$\begin{aligned} R(\hat{f}) - R(f) + \mathrm{rem}(f, \hat{f}) = & - \dot{R}(\hat{f})^T(f - \hat{f}) \\ = & (\dot{R}_n(\hat{f}) - \dot{R}(\hat{f}))^T(f - \hat{f}) - \dot{R}_n(\hat{f})^T(f - \hat{f}) \\ = & \epsilon^T(\hat{f} - f)/n - \dot{R}_n(\hat{f})^T(f - \hat{f}) \\ \le & \epsilon^T(\hat{f} - f)/n + \lambda \mathrm{TV}(f) - \lambda \mathrm{TV}(\hat{f}). \end{aligned}$$

□

One sees from Lemma 6.4 that we need appropriate bounds for the empirical process $\{\epsilon^T f/n : \ f \in \mathbb{R}^n\}$. These will be established in the next two subsections, Subsections 5.2 and 5.3. In Subsection 5.2 we announce the final result, and Subsection 5.3 presents the technicalities that lead to this result.

5.2. The empirical process $\{\epsilon^T f/n : \ f \in \mathbb{R}^n\}$.

We consider the weights[2]

$$w_k^2 := \begin{cases} \left(\frac{k - t_{j-1}}{d_j}\right)\left(\frac{t_j - k}{n}\right), & t_{j-1} + 1 \le k \le t_j - 1, \ j \in [1:r] \\ \frac{1}{n}, & k = t_j, \ j \in [1:s] \end{cases}.$$

For a vector $f \in \mathbb{R}^n$ we define $(Df)_k := f_k - f_{k-1}$ ($k = [2:n]$) so that $\|Df\|_1 = \mathrm{TV}(f)$. Let $w = (w_1, \ldots, w_n)^T$ be the vector of weights and $w^{-1} := (1/w_1, \ldots, 1/w_n)$. Write

$$w_{-S}(Df)_{-S} := \{w_k (Df)_k\}_{k \notin S}.$$

[2] These weights are inspired by the following. Let $\mathcal{V}_S$ be the linear space of functions that are piecewise constant with jumps at S and Π_S be the projection operator on the space $\mathcal{V}_S$. Then

$$\epsilon^T f/n = \epsilon^T \Pi_S f/n + \epsilon^T(I - \Pi_S)f/n,$$

and one can verify that

$$\epsilon^T(I - \Pi_S f)/n = \sum_{k \notin S} V_k (f_k - f_{k-1})$$

where $V_{-S} = \{V_k\}_{k \notin S}$ is a vector of random variables with $\mathrm{var}(V_k) = w_k^2$, $k \notin S$.

We use the notation $\|\cdot\|_{Q_n} := \|\cdot\|_2/\sqrt{n}$ for the normalized Euclidean norm. For $t > 0$, let

$$\begin{aligned}\delta_n^2(t) \geq & \left(\frac{4\nu A_0\|w^{-1}\|_{Q_n}}{\sqrt{n}} + 8\sqrt{\frac{1+t+\log(3+2\log_2 n)}{n}}\right)^2 \\ & + \left(\frac{1}{2\nu} + 4\sqrt{\frac{A_0\|w^{-1}\|_{Q_n}/\sqrt{n}}{n}} + \frac{4\sqrt{1+t+\log(3+2\log_2 n)}}{n}\right) \\ & \times \left(\|Dw\|_2 + 2\sqrt{\frac{s}{n}}\right)^2,\end{aligned}$$

and

$$\lambda_n(t) \geq \frac{1}{\sqrt{n}}\left(\frac{4}{\nu} + 8\sqrt{\frac{A_0\|w^{-1}\|_{Q_n}/\sqrt{n}}{n}} + \frac{8\sqrt{1+t+\log(3+2\log_2 n)}}{n}\right).$$

After establishing the material of Subsection 5.3 we are able show the following result:

Theorem 5.1. *Let $\mu > 0$ and $t > 0$ be arbitrary. With probability at least* $1 - \exp[-t]$

$$\epsilon^T f/n \leq \mu\delta_n^2(t) + \frac{\|f\|_{Q_n}^2}{\mu} + \lambda_n(t)\|w_{-S}(Df)_{-S}\|_1,$$

uniformly for all $f \in \mathbb{R}^n$.

Proof of Theorem 5.1. This follows from combining Lemma 5.7 with Lemma 5.6 (see Corollary 5.2). □

5.3. Material for the result for the empirical process $\{\epsilon^T f/n : \ f \in \mathbb{R}^n\}$ in Theorem 5.1.

For all $f \in \mathbb{R}^n$, let

$$\gamma_f := \frac{\sum_{j=1}^n f_j/w_j}{\|w^{-1}\|_2^2}$$

and let

$$f_{\text{P}} := \Pi_{w^{-1}} f := w^{-1}\gamma_f$$

be the projection of f onto the vector w^{-1}. Define the anti-projection $f_{\text{A}} := (I - \Pi_{w^{-1}})f$.

We let

$$wf := \{w_k f_k\}_{k=1}^n.$$

We start with some preliminary bounds.

Lemma 5.3. *For all $f \in \mathbb{R}^n$,*

$$\|wf - \gamma_f\|_\infty \leq \text{TV}(wf)$$

holds, and

$$\frac{\|f_{\text{A}}\|_\infty}{\text{TV}(wf)} \leq \sqrt{n}.$$

Proof of Lemma 5.3. For all $i \in [1:n]$,

$$\begin{aligned}w_i f_i - \gamma_f = & w_i f_i - \frac{\sum_{k=1} f_k/w_k}{\|w^{-1}\|_2^2} \\ = & \frac{\sum_{k=1}^n (w_i f_i - w_k f_k)/w_k^2}{\|w^{-1}\|_2^2} \leq \text{TV}(wf),\end{aligned}$$

or $\|wf - \gamma_f\|_\infty \leq \text{TV}(wf)$. Since, when $g = wf$,

$$f_{\text{A}} = w^{-1}(g - \gamma_f),$$

we see that

$$\|f_{\text{A}}\|_\infty \leq \|w^{-1}\|_\infty \text{TV}(g) = \|w^{-1}\|_\infty \text{TV}(wf).$$

Since $\|w^{-1}\|_\infty = \sqrt{n}$, we conclude that

$$\|f_{\text{A}}\|_\infty \leq \sqrt{n}\text{TV}(wf).$$

□

We use Dudley's entropy integral to bound the empirical process over $\{f: \ \|f_{\rm A}\|_{Q_n} \le R, \ {\rm TV}(wf) \le 1\}$ with the radius R some fixed value.

Lemma 5.4. *Let $R > 0$ be arbitrary. For all $t > 0$, with probability at least $1 - \exp[-t]$,*

$$\sup_{\|f_{\rm A}\|_{Q_n} \le R, \ {\rm TV}(wf) \le 1} \epsilon^T f/n \quad \le \quad 4\sqrt{\frac{2A_0 \|w^{-1}\|_{Q_n} R}{n}} + 4R\sqrt{\frac{1+t}{n}}.$$

Proof of Lemma 5.4. Let $\mathbf{Q}_w$ be the discrete probability measure that puts mass $w_i^{-2}/\|w^{-1}\|_2^2$ on i, $(i \in [1:n])$. Denote the $L_2(\mathbf{Q}_w)$-norm by $\| \cdot \|_{\mathbf{Q}_w}$. For $\mathcal{G} \subset \mathbb{R}^n$, we let $\mathcal{H}(\cdot, \mathcal{G}, \mathbf{Q}_w)$ denote the entropy of $\mathcal{G}$ for the metric induced by $\| \cdot \|_{\mathbf{Q}_w}$. By Lemma 5.3,

$$\|wf - \gamma_f\|_\infty \le {\rm TV}(wf).$$

Thus by Lemma 4.3, with A_0 the constant given there,

$$\mathcal{H}(u, \{wf - \gamma_f : \quad {\rm TV}(wf) \le 1\}, \mathbf{Q}_w) \le \frac{A_0}{u} \ \forall \ u > 0.$$

For $f \in \mathbb{R}^n$, we have

$$\|f_{\rm A}\|_{Q_n}^2 = \frac{1}{n}\sum_{i=1}^n (w_i f_i - \gamma_f)^2 / w_i^2 = \|wf - \gamma_f\|_{\mathbf{Q}_w}^2 \|w^{-1}\|_{Q_n}^2.$$

Therefore,

$$\mathcal{H}(u, \{f_{\rm A}, \ {\rm TV}(wf) \le 1\}, Q_n) \le \frac{A_0 \|w^{-1}\|_{Q_n}}{u} \ \forall \ u > 0.$$

The entropy integral may therefore be bounded as follows:

$$2\int_0^R \sqrt{2\mathcal{H}(u, \{f_{\rm A} : \|f_{\rm A}\|_{Q_n} \le R, \ {\rm TV}(wf) \le 1\}, Q_n)} du$$
$$\le 4\sqrt{2A_0 \|w^{-1}\|_{Q_n} R}.$$

By Lemma 4.2 the result follows. □

The next lemma invokes Lemma 5.4 and the peeling device to obtain a result for the weighted empirical process.

Lemma 5.5. *For all $t > 0$, with probability at least $1 - \exp[-t]$,*

$$\epsilon^T f_{\rm A}/n \le 8\sqrt{\frac{A_0 \|w^{-1}\|_{Q_n}}{n}} \biggl(\sqrt{\|f_{\rm A}\|_{Q_n} {\rm TV}(wf)} \vee \frac{{\rm TV}(wf)}{n^{3/4}} \biggr)$$
$$+ 8\biggl(\|f_{\rm A}\|_{Q_n} \vee \frac{{\rm TV}(wf)}{n^{3/2}} \biggr) \sqrt{\frac{1 + t + \log(2 + 2\log_2 n)}{n}}$$

holds uniformly over all f.

Proof of Lemma 5.5. Let $t > 0$ and let $\mathcal{A}$ be the event

$$\biggl\{ \epsilon^T f_{\rm A}/n \ge 8\sqrt{\frac{A_0 \|w^{-1}\|_{Q_n}}{n}} \sqrt{\|f\|_{Q_n} \vee \frac{1}{n^{3/2}}}$$
$$+ 8\biggl(\|f\|_{Q_n} \vee \frac{1}{n^{3/2}} \biggr) \sqrt{\frac{1 + t + \log(2 + 2\log_2 n)}{n}},$$
$$\text{for some } f \text{ with } \|f\|_{Q_n} \le \sqrt{n} \text{ and } {\rm TV}(wf) \le 1 \biggr\}.$$

Let $\mathcal{A}_0$ be the event

$$\begin{aligned}\Bigg\{ \sup_{\|f_{\rm A}\|_{Q_n} \le \frac{1}{n^{3/2}},\ {\rm TV}(wf)\le 1} \epsilon^T f_{\rm A}/n \le & 8\sqrt{\frac{A_0 \|w^{-1}\|_{Q_n}}{n}} \sqrt{\frac{1}{n^{3/2}}} \\ & + \frac{8}{n^{3/2}} \sqrt{\frac{1+t+\log(2+2\log_2 n)}{n}} \Bigg\}.\end{aligned}$$

Let $N \in \mathbb{N}$ satisfy $2\log_2 n \le N \le 1 + 2\log_2 n$ and for $j \in [1:N]$ let $\mathcal{A}_j$ be the event

$$\begin{aligned}\Bigg\{ \sup_{\frac{2^{j-1}}{n^{3/2}} < \|f_{\rm A}\|_{Q_n} \le \frac{2^j}{n^{3/2}},\ {\rm TV}(wf)\le 1} \epsilon^T f_{\rm A}/n \le & 8\sqrt{\frac{A_0 \|w^{-1}\|_{Q_n}}{n}} \sqrt{\frac{2^{j-1}}{n^{3/2}}} \\ & + \frac{82^{j-1}}{n^{3/2}} \sqrt{\frac{1+t+\log(2+2\log_2 n)}{n}} \Bigg\}.\end{aligned}$$

Application of Lemma 5.4 gives that for all $j \ge 0$,

$$\mathbb{P}(\mathcal{A}_j) \le \exp[-(t+\log(2+2\log_2 n)].$$

Since $\mathcal{A} \subset \cup_{j=0}^N \mathcal{A}_j$, it follows that

$$\mathbb{P}(\mathcal{A}) \le \sum_{j=0}^N \mathbb{P}(\mathcal{A}_j) \le (1+N)\exp[-(t+\log(2+2\log_2 n)] \le \exp[-t].$$

The result now follows by replacing $f_{\rm A}$ by $f_{\rm A}/{\rm TV}(wf)$ and noting that

$${\rm TV}\Big(w f_{\rm A}/TV(wf) \Big) = 1,$$

and invoking from Lemma 5.3 the bound

$$\|f_{\rm A}/{\rm TV}(wf)\|_{Q_n} \le \|f_{\rm A}/{\rm TV}(wf)\|_\infty \le \sqrt{n}. \qquad \square$$

We present a corollary that applies the "conjugate inequality" $2ab \le a^2 + b^2$ (with constants a and b in $\mathbb{R}$), then gathers terms and applies the conjugate inequality again.

Corollary 5.1. *Let $\nu > 0$ and $\mu > 0$ be arbitrary. For all $t > 0$ with probability at least $1 - \exp[-t]$,*

$$\begin{aligned}
& \epsilon^T f_{\rm A}/n \\
\le & \Bigg(\frac{4\nu A_0 \|w^{-1}\|_{Q_n}}{\sqrt{n}} + 8\sqrt{\frac{1+t+\log(2+2\log_2 n)}{n}} \Bigg) \|f_{\rm A}\|_{Q_n} \\
+ & \Bigg(\frac{4}{\nu} + 8\sqrt{\frac{A_0\|w^{-1}\|_{Q_n}/\sqrt{n}}{n}} + \frac{8\sqrt{1+t+\log(2+2\log_2 n)}}{n} \Bigg) \frac{{\rm TV}(wf)}{\sqrt{n}} \\
\le & \frac{\mu}{2} \Bigg(\frac{4\nu A_0 \|w^{-1}\|_{Q_n}}{\sqrt{n}} + 8\sqrt{\frac{1+t+\log(2+2\log_2 n)}{n}} \Bigg)^2 \\
+ & \frac{\|f_{\rm A}\|_{Q_n}^2}{2\mu} \\
+ & \Bigg(\frac{4}{\nu} + 8\sqrt{\frac{A_0\|w^{-1}\|_{Q_n}/\sqrt{n}/}{n}} + \frac{8\sqrt{1+t+\log(2+2\log_2 n)}}{n} \Bigg) \frac{{\rm TV}(wf)}{\sqrt{n}},
\end{aligned}$$

uniformly for all f.

We now add the missing $f_{\rm P} = f - f_{\rm A}$.

Lemma 5.6. *For all $t > 0$ with probability at least* $1 - \exp[-t]$,

$$\begin{aligned} &\epsilon^T f/n \\ \leq &\frac{\mu}{2}\left(\frac{4\nu A_0 \|w^{-1}\|_{Q_n}}{\sqrt{n}} + 8\sqrt{\frac{1+t+\log(3+2\log_2 n)}{n}}\right)^2 \\ + &\frac{\|f\|_{Q_n}^2}{2\mu} \\ + &\left(\frac{4}{\nu} + 8\sqrt{\frac{A_0\|w^{-1}\|_{Q_n}/\sqrt{n}}{n}} + \frac{8\sqrt{1+t+\log(3+2\log_2 n)}}{n}\right)\frac{\mathrm{TV}(wf)}{\sqrt{n}}, \end{aligned}$$

uniformly for all f.

Proof of Lemma 5.6. By Pythagoras' rule, we have $\|f\|_2^2 = \|f_{\mathrm{P}}\|_2^2 + \|f_{\mathrm{A}}\|_2^2$. Moreover, by Hoeffding's inequality, with probability at least $1 - \exp[-t]$,

$$\epsilon^T f_{\mathrm{P}}/n \leq \|f_{\mathrm{P}}\|_{Q_n}\sqrt{\frac{2t}{n}} \leq \frac{\mu t}{n} + \frac{\|f_{\mathrm{P}}\|_{Q_n}^2}{2\mu}. \qquad \square$$

In Lemma 5.6, the term including $\mathrm{TV}(wf)$ is almost, but not yet quite the one to be dealt with by the penalty. We bound it by $\|w_{-S}(Df)_{-S}\|_1$ with appropriate remaining terms invoking the "chain rule". Here,

$$w_{-S}(Df)_{-S} := \{w_k(Df)_k\}_{k\notin S}.$$

Lemma 5.7. *For all $f \in \mathbb{R}^n$,*

$$\mathrm{TV}(wf) \leq \sqrt{n}\left(\|Dw\|_2 + 2\sqrt{s/n}\right)\|f\|_{Q_n} + \|w_{-S}(Df)_{-S}\|_1.$$

Proof of Lemma 5.7. We use the fact that

$$\begin{aligned} \mathrm{TV}(wf) \leq &\sum_{i=2}^n |(w_i - w_{i-1})f_{i-1}| + \sum_{i=2}^n |w_i(f_i - f_{i-1})| \\ \leq &\|Dw\|_2\|f\|_2 + \|wDf\|_1. \end{aligned}$$

Moreover,

$$\|wDf\|_1 = \|w_S(Df)_S\|_1 + \|w_{-S}(Df)_{-S}\|_1$$

with

$$w_S(Df)_S := \{w_k(Df)_k\}_{k\in S},$$

satisfying

$$\begin{aligned} \|w_S(Df)_S\|_1 = &\sum_{j=1}^s |f_{t_j+1} - f_{t_j}|/\sqrt{n} \\ \leq &\sqrt{s}\sqrt{\sum_{j=1}^s |f_{t_j+1} - f_{t_j}|^2/\sqrt{n}} \\ \leq &2\sqrt{s}\|f\|_2/\sqrt{n}. \end{aligned}$$

Thus,

$$\mathrm{TV}(wf) \leq \left(\|Dw\|_2 + 2\sqrt{s/n}\right)\|f\|_2 + \|w_{-S}(Df)_{-S}\|_1. \qquad \square$$

Corollary 5.2. *The result from Theorem* 5.1 *now follows by using*

$$\left(\|Dw\|_2 + 2\sqrt{s/n}\right)\|f\|_{Q_n} \leq \frac{\mu}{2}\left(\|Dw\|_2 + 2\sqrt{s/n}\right)^2 + \frac{\|f\|_{Q_n}^2}{2\mu},\ f \in \mathbb{R}^n.$$

5.4. **Bounds for the weights and their inverses.** So far we assumed in this section (see Subsection 5.2), that for $t > 0$, the quantities $\delta_n^2(t)$ and $\lambda_n(t)$ involved in the bound for the empirical process in Theorem 5.1 satisfy

$$\begin{aligned}\delta_n^2(t) \geq & \left(\frac{4\nu A_0\|w^{-1}\|_{Q_n}}{\sqrt{n}} + 8\sqrt{\frac{1+t+\log(3+2\log_2 n)}{n}}\right)^2 \\ + & \left(\frac{1}{2\nu} + 4\sqrt{\frac{A_0\|w^{-1}\|_{Q_n}/\sqrt{n}}{n}} + \frac{4\sqrt{1+t+\log(3+2\log_2 n)}}{n}\right) \\ \times & \left(\|Dw\|_2 + 2\sqrt{\frac{s}{n}}\right)^2,\end{aligned}$$

and

$$\lambda_n(t) \geq \frac{1}{\sqrt{n}}\left(\frac{4}{\nu} + 8\sqrt{\frac{A_0\|w^{-1}\|_{Q_n}/\sqrt{n}}{n}} + \frac{8\sqrt{1+t+\log(3+2\log_2 n)}}{n}\right),$$

involving $\|w^{-1}\|_{Q_n}$ and $\|Dw\|_2$. In this subsection, we present the bounds for these, leading to the values $\delta_n^2(t)$ and $\lambda_n(t)$ presented in Section 2.

Lemma 5.8. *It holds that*

$$\|w^{-1}\|_2^2 \leq 2n \sum_{d_j \geq 2} (\log(d_j - 1) + 1) + ns \leq n^2 \Delta_n^2$$

and

$$\|Dw\|_2^2 \leq 4 \sum_{d_j \geq 2} (\log(d_j - 1) + 1)/n + s/n =: \Delta_n^2.$$

Proof of Lemma 5.8. We have[3]

$$\begin{aligned}\|w^{-1}\|_2^2 = & \sum_{d_j \geq 2} \sum_{k=1}^{d_j - 1} \frac{nd_j}{k(d_j - k)} + ns \\ \leq & 2n \sum_{j=1}^{r} (\log(d_j - 1) + 1) + ns.\end{aligned}$$

Moreover, for $1 \leq k \leq d_j - 1$, $j \in [1 : r]$,

$$\begin{aligned}|\sqrt{k}\sqrt{d_j - k} - \sqrt{k-1}\sqrt{d_j - (k-1)}| & \leq \sqrt{\frac{d_j - k}{k}} + \sqrt{\frac{k-1}{d_j - k}} \\ & \leq \sqrt{\frac{d_j - 1}{k}} + \sqrt{\frac{d_j - 2}{d_j - k}} \leq \sqrt{\frac{d_j}{k}} + \sqrt{\frac{d_j}{d_j - k}},\end{aligned}$$

so that

$$\begin{aligned}& \sum_{k=1}^{d_j - 1} \frac{|\sqrt{k}\sqrt{d_j - k} - \sqrt{k-1}\sqrt{d_j - (k-1)}|^2}{nd_j} \\ \leq & \frac{2}{n} \sum_{k=1}^{d_j - 1} \left(\frac{1}{k} + \frac{1}{d_j - k}\right) \\ \leq & \frac{1}{n} \sum_{j=1}^{r} (4\log(d_j - 1) + 2).\end{aligned}$$

[3] We use $\sum_{k=1}^{d-1} \frac{d}{k(d-k)} = \sum_{k=1}^{d-1} \left(\frac{1}{k} + \frac{1}{d-k}\right) = 2\sum_{k=1}^{d-1} \frac{1}{k} \leq 2(1 + \log(d-1))$.

Finally, for $j \in [1:s]$,

$$|w_{t_j} - w_{t_j - 1}| = \left| \frac{1}{\sqrt{n}} - \sqrt{\frac{d_j - 1}{d_j}} \frac{1}{\sqrt{n}} \right| \le \frac{1}{\sqrt{n}}. \qquad \square$$

5.5. **A bound for the effective sparsity.** For all $f \in \mathbb{R}^n$, we let

$$(Df)_S := \{(Df)_k\}_{k \in S},\ (Df)_{-S} := \{(Df)_k\}_{k \notin S}.$$

and recall that

$$w_{-S}(Df)_{-S} := \{w_k (Df)_k\}_{k \notin S}.$$

Let $q_{t_j} := \text{sign}(\mathbf{f}_{t_j})$, $j \in [1:s]$. We define $q_S := \{q_{t_j}\}_{j=1}^s$.

Definition 5.1. Let $\lambda \ge \lambda_n(t)\sqrt{d_{\max}/(2n)}$. The effective sparsity at $\mathbf{f}$ is

$$\Gamma^2(\mathbf{f}, t) := \left(\min \left\{ \|f\|_{Q_n} :\ q^T (Df)_S - \|(1 - w_{-S}\lambda(t)/\lambda)(Df)_{-S}\|_1 = 1 \right\} \right)^{-2}.$$

Recall the definitions

$$J_{\text{monotone}} := \{2 \le j \le s : q_{t_j} = q_{t_{j-1}})\},\ J_{\text{change}} := [1:r] \backslash J_{\text{monotone}}.$$

Lemma 5.9. *For $\lambda \ge \lambda_n(t)\sqrt{d_{\max}/n}$ we have*

$$\Gamma(\mathbf{f}, t) \le \Gamma_n^2(t),$$

where

$$\Gamma_n^2(t) \ := \ \frac{\lambda_n^2(t)}{\lambda^2} \sum_{j \in J_{\text{monotone}}} 8(\log(d_j) + 1) + \sum_{j \in J_{\text{change}}} \frac{8n(\log(d_j) + 2)}{d_j}.$$

Proof of Lemma 5.9. The proof uses interpolating vectors $q \in \mathbb{R}^n$ as in [13], where $q = (q_1, q_{-1})^T$ is given below. We show that

$$q_S^T (Df)_S - \|(1 - w_{-S}\lambda(t)/\lambda)(Df)_{-S}\|_1 \le q_{-1}^T D(\mathbf{f} - \hat{f}).$$

The result then follows from

$$q_{-1}^T D(\mathbf{f} - f) = (D^T q_{-1})^T (\mathbf{f} - f) \le \|D^T q_{-1}\|_2 \|\mathbf{f} - f\|_2.$$

Furthermore, under the boundary conditions $q_1 = q_n = 0$ we see that $\|D^T q_{-1}\|_2 = \|Dq\|_2$. Define

$$\omega_k^2 := \begin{cases} \left(\frac{k - t_{j-1}}{d_j}\right)\left(\frac{t_j - k}{n}\right)\frac{\lambda_n(t)}{\lambda}, & t_{j-1} + 1 \le k \le t_j - 1,\ j \in J_{\text{monotone}},\ d_j \ge 2 \\ \left(\frac{k - t_{j-1}}{d_j}\right)\left(\frac{t_j - k}{d_j}\right), & t_{j-1} + 1 \le k \le t_j - 1,\ j \in J_{\text{change}} \\ 0, & k = t_j,\ j \in [1:s] \end{cases}.$$

For $j \in [1:r]$, we let $\bar{t}_j = \frac{t_{j-1} + t_j}{2}$ be the midpoints. Moreover, for $k \notin \{t_1, \ldots, t_s\}$, let

$$q_k := \begin{cases} 0 & 1 \le k < \bar{t}_1 \\ \text{sign}(\mathbf{f}_{t_1})(1 - 2\omega_k), & \bar{t}_1 \le k \le t_1 - 1 \\ \text{sign}(\mathbf{f}_{t_{j-1}})(1 - 2\omega_k), & t_{j-1} + 1 \le k < \bar{t}_j,\ j \in [2:s] \\ \text{sign}(\mathbf{f}_{t_j})(1 - 2\omega_k), & \bar{t}_j \le k \le t_j - 1,\ j \in [2:s] \\ \text{sign}(\mathbf{f}_{t_{r-1}})(1 - 2\omega_k), & t_{r-1} \le k < \bar{t}_r \\ 0 & \bar{t}_r \le k \le n \end{cases}.$$

For $\bar{t}_j - 1 \le k < \bar{t}_j$, $j \in J_1$, we get

$$|1 - 2\omega_k| \le \frac{4}{d_j}.$$

For $j \in J_{\text{monotone}}$, we see that

$$\sum_{k=1}^{d_j} |q_{t_{j-1}+k} - q_{t_{j-1}+k-1}|^2 \leq \frac{\lambda_n^2(t)}{\lambda^2} \frac{8(\log d_j + 1)}{n},$$

and for $j \in J_{\text{change}}$,

$$\sum_{k=1}^{d_j} |q_{t_{j-1}+k} - q_{t_{j-1}+k-1}|^2 \leq \frac{8(\log d_j + 2)}{d_j}.$$

Thus,

$$\|Dq\|_2^2 \leq \frac{\lambda_n^2(t)}{\lambda^2} \sum_{j \in J_{\text{monotone}}} \frac{8(\log(d_j) + 1)}{n} + \sum_{j \in J_{\text{change}}} \frac{8(\log(d_j) + 2)}{d_j}.$$

The lemma now follows from $\Gamma^2(\mathbf{f}, t) \leq n\|Dq\|_2^2$. □

5.6. **Finalizing the proof of Theorem 2.1.** By Lemma 6.4, we have

$$\begin{aligned}
&R(\hat{f}) - R(\mathbf{f}) + \text{rem}(\mathbf{f}, \hat{f}) \\
\leq&\mu\delta_n^2(t) + \frac{\|\hat{f} - \mathbf{f}\|_{Q_n}}{\mu} + \lambda_n(t)\|w_{-S}(D\hat{f})_{-S}\|_1 + \lambda\|D_S\mathbf{f}\|_1 - \lambda\|D\hat{f}\|_1 \\
=&\mu\delta_n^2(t) + \frac{\|\hat{f} - \mathbf{f}\|_{Q_n}^2}{\mu} + \lambda\bigg(\|(D\mathbf{f})_S\|_1 - \|(D\hat{f})_S\|_1 - \|(1 - \lambda_n(t)w_{-S}/\lambda)(D\hat{f})_{-S}\|_1\bigg) \\
\leq&\mu\delta_n^2(t) + \frac{\|\hat{f} - \mathbf{f}\|_{Q_n}^2}{\mu} + \lambda\Gamma_n(t)\|\hat{f} - \mathbf{f}\|_{Q_n} \\
\leq&\mu\delta_n^2(t) + \frac{2\|\hat{f} - \mathbf{f}\|_{Q_n}^2}{\mu} + \frac{\lambda^2}{4}\Gamma_n^2(t).
\end{aligned}$$

Choose $\mu = 4\kappa$ to obtain

$$\frac{2\|\hat{f} - \mathbf{f}\|_{Q_n}^2}{\mu} = \frac{\|\hat{f} - \mathbf{f}\|_{Q_n}^2}{2\kappa} \leq \text{rem}(\mathbf{f}, \hat{f}),$$

whenever $\|\hat{f}\|_\infty \leq B$. □

6. Proof of Theorem 3.1

6.1. **Some lemmas used in the proof of Theorem 3.1.** The proof of Theorem 3.1 applies some auxiliary lemmas which we develop in this subsection. Define

$$\tau(f) := \|f\|_{Q_n}/(\sqrt{2}K) + (\lambda/\delta)\text{TV}(f)$$

with

$$\delta^2 := 2^4\lambda M_0, \; K^2 := \frac{(1 + \mathrm{e}^{1+2^4M_0+\|f^0\|_\infty})^2}{\mathrm{e}^{1+2^4M_0+\|f^0\|_\infty}},$$

where $M_0 \geq \text{TV}(f^0) \vee 1$. Moreover, we let

$$\hat{f}_\alpha := \alpha\hat{f} + (1 - \alpha)f^0$$

with

$$\alpha := \frac{\delta}{\delta + \tau(f - f_0)}.$$

Let $\mathcal{F}_0 := \{f : \; \tau(f) \leq \delta\}$.

Lemma 6.1. *It holds that $\hat{f}_\alpha - f^0 \in \mathcal{F}_0$, i.e., $\tau(\hat{f}_\alpha - f^0) \leq \delta$. Moreover, if in fact $\tau(\hat{f}_\alpha - f^0) \leq \delta/2$, then $\hat{f} - f^0 \in \mathcal{F}_0$, as well.*

Proof. We have

$$\tau(\hat{f}_\alpha - f^0) = \alpha\tau(\hat{f} - f^0) = \frac{\delta\tau(\hat{f} - f^0)}{\delta + \tau(\hat{f} - f_0)} \le \delta.$$

If in fact $\tau(\hat{f}_\alpha - f^0) \le \delta/2$, we have

$$\tau(\hat{f}_\alpha - f^0) = \frac{\delta\tau(\hat{f} - f^0)}{\delta + \tau(\hat{f} - f_0)} \le \delta/2$$

which gives $\tau(\hat{f} - f^0) \le \delta/2 + \tau(\hat{f} - f_0)/2$, or $\tau(\hat{f} - f^0) \le \delta$. □

Lemma 6.2. *For all $f \in \mathbb{R}^n$,*

$$\|f\|_\infty \le \|f\|_{Q_n} + \mathrm{TV}(f).$$

Moreover,

$$\mathcal{F}_0 \subset \{f :\ \|f\|_\infty \le \sqrt{2}K\delta + \delta^2/\lambda,\ \mathrm{TV}(f) \le \delta^2/\lambda\}.$$

Proof. For $f \in \mathbb{R}^n$, we denote its average by

$$\bar{f} := \frac{1}{n}\sum_{i=1}^n f_i.$$

Then

$$\|f\|_{Q_n}^2 = \bar{f}^2 + \|f - \bar{f}\|_{Q_n} \ge \bar{f}^2.$$

Moreover, for all i,

$$f_i - \bar{f} = \frac{1}{n}\sum_{j=1}^n (f_i - f_j) \le \mathrm{TV}(f).$$

It follows that

$$\|f\|_\infty \le \bar{f} + \|f - \bar{f}\|_\infty \le \|f\|_{Q_n} + \mathrm{TV}(f).$$

For $f \in \mathcal{F}_0$, we have $\|f\|_2/\sqrt{n} \le \sqrt{2}K\delta$ and $\mathrm{TV}(f) \le \delta^2/\lambda$, so that also $\|f\|_\infty \le \sqrt{2}K\delta + \delta^2/\lambda$. □

Lemma 6.3. *Let*

$$K^2 := \frac{(1 + \mathrm{e}^{1+2^4M_0+\|f^0\|_\infty})^2}{\mathrm{e}^{1+2^4M_0+\|f^0\|_\infty}}$$

and let $\delta^2 := 2^4\lambda M_0 \le 1/(2K^2)$. Then for all f with $f - f^0 \in \mathcal{F}_0$, it is true that $K_f \le K$.

Proof. Since for $f - f^0 \in \mathcal{F}_0$, $\|f - f^0\|_\infty \le \sqrt{2}K\delta + \delta^2/\lambda \le 1 + 2^4M_0$, we see that $\|f\|_\infty \le 1 + 2^4M_0 + \|f^0\|_\infty$. Therefore,

$$K_f^2 = \frac{(1 + \mathrm{e}^{\|f\|_\infty \vee \|f^0\|_\infty})^2}{\mathrm{e}^{\|f\|_\infty \vee \|f^0\|_\infty}} \le K^2.$$

□

Lemma 6.4. *We have*

$$0 \le R(\hat{f}) - R(f^0) \le \epsilon^T(\hat{f} - f^0)/n + \lambda\mathrm{TV}(f^0) - \lambda\mathrm{TV}(\hat{f}).$$

This inequality is also true with $\hat{f}$ replaced by $\hat{f}_\alpha$.

Proof. For any f,

$$\begin{aligned} 0 \le R(f) - R(f^0) =& - \left[\Big(R_n(f) - R(f)\Big) - \Big(R_n(f^0) - R(f^0)\Big)\right] \\ &+ R_n(f) - R_n(f^0) \\ =& \epsilon^T(f - f^0)/n + R_n(f) - R_n(f^0). \end{aligned}$$

Insert the basic inequality

$$R_n(\hat{f}) + \lambda\mathrm{TV}(\hat{f}) \le R_n(f^0) + \lambda\mathrm{TV}(f^0),$$

or

$$R_n(\hat{f}) - R_n(f^0) \le \lambda\mathrm{TV}(f^0) - \lambda\mathrm{TV}(\hat{f}),$$

to arrive at the first statement of the lemma. To obtain the second statement, we note that by the convexity of $f \mapsto R_n(f)$ such basic inequality is also true for $\hat{f}_\alpha$:

$$\begin{aligned}
&R_n(\hat{f}_\alpha) + \lambda \mathrm{TV}(\hat{f}_\alpha) \\
\le&\alpha R_n(\hat{f}) + \alpha\lambda\mathrm{TV}(\hat{f}) + (1-\alpha)R_n(f^0) + (1-\alpha)\lambda\mathrm{TV}(f^0) \\
\le&R_n(f^0) + \lambda\mathrm{TV}(f^0). \qquad \square
\end{aligned}$$

6.2. **Proof of Theorem 3.1.** We have for $f \in \mathcal{F}_0$, $\|f\|_\infty \le \sqrt{2}K\delta + \delta^2/\lambda \le 2\delta^2/\lambda$ as well as $\mathrm{TV}(f) \le \delta^2/\lambda \le 2\delta^2/\lambda$. It follows from Lemma 4.3 that

$$H(u, \mathcal{F}_0, Q_n) \le \frac{2A_0\delta^2}{\lambda u} \ \forall \ u > 0,$$

so that

$$\begin{aligned}
2\int_0^{\sqrt{2}K\delta} \sqrt{2H(u,\mathcal{F}_0,Q_n)}du \le& 4\sqrt{2A_0\sqrt{2}K}\frac{\delta}{\sqrt{\lambda}}\int_0^{\sqrt{2}K\delta}\frac{1}{\sqrt{u}}du \\
=&8\sqrt{\frac{2A_0\sqrt{2}K}{\lambda}}\delta^{3/2}.
\end{aligned}$$

But then, in view of Lemma 4.2, for all $t > 0$ with probability at least $1 - \exp[-t]$,

$$\sup_{f\in\mathcal{F}_0} \epsilon^T f/n \quad \le \quad 8\sqrt{\frac{2A_0\sqrt{2}K}{n\lambda}}\delta^{3/2} + 4\sqrt{2}K\delta\sqrt{\frac{1+t}{n}}.$$

Since, by Lemma 6.1, $\hat{f}_\alpha - f^0 \in \mathcal{F}_0$ we know from Lemma 6.3 that $K_{\hat{f}_\alpha} \le K$. Thus, in view of Lemma 6.4 and the bound

$$R(\hat{f}_\alpha) - R(f^0) \ge \frac{\|\hat{f}_\alpha - f^0\|^2_{Q_n}}{2K^2},$$

we have shown that with probability at least $1 - \exp[-t]$,

$$\begin{aligned}
&\frac{\|\hat{f}_\alpha - f^0\|^2_{Q_n}}{2K^2} + \lambda\mathrm{TV}(\hat{f}_\alpha - f^0) \\
\le&2\lambda\mathrm{TV}(f^0) + 8\sqrt{\frac{2A_0\sqrt{2}K}{n\lambda}}\delta^{3/2} + 4\sqrt{2}K\delta\sqrt{\frac{1+t}{n}} \\
\le&\ 2\lambda M_0 \quad + \quad 8\sqrt{\frac{2A_0\sqrt{2}K}{n\lambda}}\delta^{3/2} + 4\sqrt{2}K\delta\sqrt{\frac{1+t}{n}}.
\end{aligned}$$

We want the three terms on the right-hand side to add up to at most $\delta^2/4$. We choose

$$\begin{aligned}
\lambda M_0 =&\delta^2/2^3, \\
8\sqrt{\frac{2A_0\sqrt{2}K}{n\lambda}}\delta^{3/2} \le&\delta^2/2^4, \\
4\sqrt{2}K\delta\sqrt{\frac{1+t}{n}} \le&\delta^2/2^4.
\end{aligned}$$

or

$$\begin{aligned}
2^4\lambda M_0 =&\delta^2, \\
\left(\frac{2^7\sqrt{2A_0\sqrt{2}K}}{\sqrt{n\lambda}}\right)^4 \le&\delta^2, \\
\left(2^6\sqrt{2}K\sqrt{\frac{1+t}{n}}\right)^2 \le&\delta^2.
\end{aligned}$$

The first one is the largest of the three. This leads to the requirement

$$2^4\lambda M_0 \geq \left(2^7\sqrt{\frac{2A_0\sqrt{2}K}{n\lambda}}\right)^4,$$

which is true for

$$\lambda \geq 2^8 n^{-2/3} A_0^{2/3}(\sqrt{2}K)^{2/3},$$

and

$$2^4\lambda M_0 \geq \left(2^6\sqrt{2}K\sqrt{\frac{1+t}{n}}\right)^2,$$

which holds for

$$\lambda \geq 2^8(2K^2)\frac{1+t}{n},$$

where we invoked for both requirements that $M_0 \geq 1$. Then with probability at least $1-\exp[-t]$,

$$\frac{\|\hat{f}_\alpha - f^0\|_{Q_n}^2}{2K^2} + \lambda \mathrm{TV}(\hat{f}_\alpha - f^0) \leq \delta^2/4.$$

For all $f \in \mathbb{R}^n$,

$$\delta\tau(f) = \frac{\delta\|f\|_{Q_n}}{\sqrt{2}K} + \lambda\mathrm{TV}(f) \leq \delta^2/4 + \frac{\|f\|_2^2/n}{2K^2} + \lambda\mathrm{TV}(f).$$

Thus we have shown that

$$\delta\tau(\hat{f}_\alpha - f^0) \leq \delta^2/4 + \delta^2/4 = \delta^2/2$$

or

$$\tau(\hat{f}_\alpha - f^0) \leq \delta/2.$$

By Lemma 6.1, this implies $\hat{f} \in \mathcal{F}_0$. We can now apply the same arguments to $\hat{f}$ as we did for $\hat{f}_\alpha$ to obtain that with probability at least $1 - 2\exp[-t]$,

$$\frac{\|\hat{f} - f^0\|_{Q_n}^2}{2K^2} + \lambda\mathrm{TV}(\hat{f} - f^0) \leq \delta^2/4 = 4\lambda M_0$$

holds. By Lemma 6.2, this implies

$$\|\hat{f} - f^0\|_\infty \leq \frac{\sqrt{2}K\delta}{2} + \frac{\delta^2}{4\lambda} \leq \frac{1+8M_0}{2}. \qquad \square$$

Acknowledgement

We are very grateful to the anonymous reviewer for the helpful comments and careful reading of the manuscript.

References

1. A. Ahmed, E. P. Xing, Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences* **106** (2009), no. 29, 11878–11883.
2. B. Betancourt, A. Rodríguez, N. Boyd, Bayesian fused lasso regression for dynamic binary networks. *J. Comput. Graph. Statist.* **26** (2017), no. 4, 840–850.
3. S. Chatterjee, S. Goswami, New risk bounds for 2d total variation denoising, (2019). arXiv preprint arXiv:1902.01215.
4. A. S. Dalalyan, M. Hebiri, J. Lederer, On the prediction performance of the Lasso. *Bernoulli* **23** (2017), no. 1, 552–581.
5. B. Fang, A. Guntuboyina, B. Sen, Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy-Krause variation, 2019. arXiv preprint arXiv:1903.01395.
6. A. Guntuboyina, D. Lieu, S. Chatterjee, B. Sen, Adaptive risk bounds in univariate total variation denoising and trend filtering. *Ann. Statist.* **48** (2020), no. 1, 205–229.
7. J.-C. Hütter, P. Rigollet, Optimal rates for total variation denoising. In: *Conference on Learning Theory*, pp. 1115–1146, 2016.
8. K. Lin, J. L. Sharpnack, A. Rinaldo, R. J. Tibshirani, A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In: *Advances in Neural Information Processing Systems*, pp. 6884–6893, 2017.
9. C. Liu, H. S. Wong, Structured Penalized Logistic Regression for Gene Selection in Gene Expression Data Analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **16** (2017), no. 1, 312–321.

10. J. Liu, L. Yuan, J. Ye, An efficient algorithm for a class of fused lasso problems. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 323–332, 2010.
11. F. Ortelli, S. van de Geer, On the total variation regularized estimator over a class of tree graphs. *Electron. J. Stat.* **12** (2018), no. 2, 4517–4570.
12. F. Ortelli, S. van de Geer, Oracle inequalities for image denoising with total variation regularization, 2019. arXiv preprint arXiv:1911.07231.
13. F. Ortelli, S. van de Geer, Prediction bounds for (higher order) total variation regularized least squares, 2019. arXiv preprint arXiv:1904.10871.
14. O. H. M. Padilla, J. Sharpnack, J. G. Scott, The DFS fused lasso: Linear-time denoising over general graphs. *The Journal of Machine Learning Research* **18** (2017), no. 1, 6410–6445.
15. L. I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms. Experimental mathematics: computational issues in nonlinear science (Los Alamos, NM, 1991). *Phys. D* **60** (1992), no. 1-4, 259–268.
16. V. Sadhanala, R. J. Tibshirani, Additive models with trend filtering. *Ann. Statist.* **47** (2019), no. 6, 3032–3068.
17. V. Sadhanala, Y.-X. Wang, R. J. Tibshirani, Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In: *Advances in Neural Information Processing Systems*, pp. 3513–3521, 2016.
18. V. Sadhanala, Y.-X. Wang, J. L. Sharpnack, R. J. Tibshirani, Higher-order total variation classes on grids: Minimax theory and trend filtering methods. In: *Advances in Neural Information Processing Systems*, pp. 5800–5810, 2017.
19. G. Steidl, S. Didas, J. Neumann, Splines in higher order TV regularization. *International Journal of Computer Vision* **70** (2006), no. 3, 241–255.
20. H. Sun, S. Wang, Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. *Bioinformatics* **28** (2012), no. 10, 1368–1375.
21. R. J. Tibshirani, Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** (2014), no. 1, 285–323.
22. R. Tibshirani, M. Saunders, S. Rosset, Z. Ji, K. Knight, Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** (2005) no. 1, 91–108.
23. S. van de Geer, *Estimation and Testing Under Sparsity: École d'Eté de Probabilités de Saint Flour XLV-2016*. Springer Science & Business Media, 2016.
24. A. W. van der Vaart, J. A. Wellner, *Weak Convergence and Empirical Processes*. Springer, 1996.
25. D. Yu, S. J. Lee, W. J. Lee, S. C. Kim, J. Lim, S. W. Kwon, Classification of spectral data using fused lasso logistic regression. *Chemometrics and Intelligent Laboratory Systems* **142** (2015), 70–77.
26. D. Yu, J.-H. Won, T. Lee, J. Lim, S. Yoon, High-dimensional fused lasso regression using majorization-minimization and parallel processing. *J. Comput. Graph. Statist.* **24** (2015), no. 1, 121–153.

(Received 06.03.2020)

Seminar for Statistics, ETH Zürich Rämistrasse 101, 8092 Zürich Switzerland
E-mail address: geer@stat.math.ethz.ch