Journal of Machine Learning Research 16 (2015) 1879-1892 Submitted 6/15; Published 9/15

Sharp Oracle Bounds for Monotone and Convex Regression Through Aggregation

Pierre C. Bellec PIERRE.BELLEC@ENSAE.FR
Alexandre B. Tsybakov ALEXANDRE.TSYBAKOV@ENSAE.FR
ENSAE,
3 avenue Pierre Larousse
92240 Malakoff, France

Editor: Alex Gammerman and Vladimir Vovk

Abstract

We derive oracle inequalities for the problems of isotonic and convex regression using the combination of Q-aggregation procedure and sparsity pattern aggregation. This improves upon the previous results including the oracle inequalities for the constrained least squares estimator. One of the improvements is that our oracle inequalities are sharp, i.e., with leading constant 1. It allows us to obtain bounds for the minimax regret thus accounting for model misspecification, which was not possible based on the previous results. Another improvement is that we obtain oracle inequalities both with high probability and in expectation.

Keywords: aggregation, shape constraints, isotonic regression, convex regression, minimax regret, sharp oracle inequalities, model misspecification

1. Introduction

Assume that we have the observations

$$Y_i = \mu_i + \xi_i, \qquad i = 1, ..., n, \tag{1}$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T \in \mathbb{R}^n$ is unknown, $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^T$ is a noise vector with n-dimensional Gaussian distribution $\mathcal{N}(0, \sigma^2 I_{n\times n})$ where $\sigma > 0$. We observe $\mathbf{y} = (Y_1, ..., Y_n)^T$ and we want to estimate $\boldsymbol{\mu}$. We can interpret μ_i as the values $f(X_i)$ of an unknown regression function $f : \mathcal{X} \to \mathbb{R}$ at given non-random points $X_i \in \mathcal{X}$, $i = 1, \dots, n$, where $\mathcal{X}$ is an abstract set. Then, the equivalent setting is that we observe $\mathbf{y}$ along with $(X_1, \dots, X_n)$ but the values of X_i are of no interest and can be replaced by their indices if we measure the loss in a discrete norm. Namely, for any $\boldsymbol{u} \in \mathbb{R}^n$ we consider the scaled (or the empirical) norm $\|\cdot\|$ defined by

$$\|\boldsymbol{u}\|^2 = \frac{1}{n}\sum_{i=1}^{n} u_i^2. \tag{2}$$

We will measure the error of an estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ by the distance $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|$. Let $\mathcal{S}^{\uparrow}$ be the set of all non-decreasing sequences:

$$\mathcal{S}^{\uparrow} := \{\boldsymbol{u} = (u_1, ..., u_n) \in \mathbb{R}^n : u_i \le u_{i+1}, \quad i = 1, ..., n-1\}. \tag{3}$$

For a subset $\mathcal{S}$ of $\mathcal{S}^{\uparrow}$, and any $\boldsymbol{\mu} \in \mathbb{R}^n$ the quantity $\min_{\boldsymbol{u}\in\mathcal{S}} \|\boldsymbol{u} - \boldsymbol{\mu}\|$ is the smallest approximation error achievable by a sequence in the set $\mathcal{S}$. This quantity defines a benchmark

or oracle performance on $\mathcal{S}$. The accuracy of an estimator $\hat{\boldsymbol{\mu}}$ with respect to the oracle for any $\boldsymbol{\mu}$, not necessarily $\boldsymbol{\mu} \in \mathcal{S}$, can be characterized by the excess loss $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \min_{\boldsymbol{u}\in\mathcal{S}} \|\boldsymbol{u} - \boldsymbol{\mu}\|$. This is a measure of performance of $\boldsymbol{\mu}$ under model misspecification. One can also consider the expected quantities $R_1(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \min_{\boldsymbol{u}\in\mathcal{S}} \|\boldsymbol{u} - \boldsymbol{\mu}\|$ or $R_2(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \min_{\boldsymbol{u}\in\mathcal{S}} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2$ known under the name of regret measures. Here, $\mathbb{E}_{\boldsymbol{\mu}}$ denotes the expectation with respect to the distribution of $\mathbf{y}$ satisfying (1). The minimax regret is defined as $\min_{\hat{\boldsymbol{\mu}}} \max_{\boldsymbol{\mu}\in\mathbb{R}^n} R_i(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ for $i = 1, 2$, where $\min_{\hat{\boldsymbol{\mu}}}$ denotes the minimum over all estimators. We can characterize the performance of an estimator $\tilde{\boldsymbol{\mu}}$ by the closeness of its maximal regret $\max_{\boldsymbol{\mu}\in\mathbb{R}^n} R_i(\tilde{\boldsymbol{\mu}}, \boldsymbol{\mu})$ to the minimax regret. This approach to measure the performance of estimators under model misspecification was pioneered by Vapnik and Chervonenkis who called it the criterion of minimax of the loss (Vapnik and Chervonenkis, 1974, Chapter 6). In this paper, we follow this approach and establish non-asymptotic bounds for the maximal regret for some classes $\mathcal{S}$ of monotone and convex functions.

When the model is well-specified, i.e., the true function $\boldsymbol{\mu}$ belongs to the class $\mathcal{S}$, the approximation error vanishes and instead of the minimax regret it is natural to consider the minimax risk defined either as $\min_{\hat{\boldsymbol{\mu}}} \max_{\boldsymbol{\mu}\in\mathcal{S}} \mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|$ or as $\min_{\hat{\boldsymbol{\mu}}} \max_{\boldsymbol{\mu}\in\mathcal{S}} \mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ (the minimax squared risk). It is easy to see that the minimax risk is not greater than the minimax regret. A classical problem in nonparametric statistics is to study the behavior of minimax risks for different classes $\mathcal{S}$. In particular, there exist results concerning the minimax risks for classes of monotone and convex functions in our setting. We review some of them below. The behavior of the minimax regret is much less studied. For a recent overview and some general results we refer to Rakhlin et al. (2013) where it is shown that the rate of minimax regret can be different from that of the minimax risk. Note that Rakhlin et al. (2013) studies the prediction problem with i.i.d. observations, which is a setting different from ours.

A well-studied estimator under the monotonicity and convexity assumptions is the least squares estimator

$$\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}) \in \operatorname*{argmin}_{\boldsymbol{u}\in\mathcal{S}} \|\mathbf{y} - \boldsymbol{u}\|^2 . \tag{4}$$

In Nemirovski et al. (1985) it was shown that $\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S})$ attains, up to logarithmic factors, the rates $n^{-2/3}$ and $n^{-4/5}$ of the mean squared risk for classes $\mathcal{S}$ of monotone and convex functions respectively and that these rates are optimal up to logarithmic factors when the minimax squared risk is used as a criterion. Under monotonicity constraints, the rate $n^{-2/3}$ was later observed in different settings, see for instance Banerjee and Wellner (2001); Balabdaoui and Wellner (2007).

One class of monotone functions we will be interested in here is defined as

$$\mathcal{S}^{\uparrow}(V) = \{\boldsymbol{\mu} \in \mathcal{S}^{\uparrow} : \ V(\boldsymbol{\mu}) \le V\}$$

where $V(\boldsymbol{\mu}) = \mu_n - \mu_1$ for any $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n) \in \mathcal{S}^{\uparrow}$, and $V > 0$ is a given constant. In Meyer and Woodroofe (2000); Zhang (2002) it was shown that for any $\boldsymbol{\mu} \in \mathcal{S}^{\uparrow}$ we have

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \le c \max\left(\left(\frac{\sigma^2 V(\boldsymbol{\mu})}{n} \right)^{2/3}, \frac{\sigma^2 \log n}{n} \right) \tag{5}$$

for $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}^{\uparrow})$ and some absolute constant $c > 0$. This immediately implies an upper bound on the minimax risk on $\mathcal{S}^{\uparrow}(V)$. A recent paper Chatterjee et al. (2015) establishes the oracle inequality

$$\mathbb{E}_{\boldsymbol{\mu}} \left\| \hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}^{\uparrow}) - \boldsymbol{\mu} \right\|^2 \leq C_* \min_{\boldsymbol{u} \in \mathcal{S}^{\uparrow}} \left(\|\boldsymbol{\mu} - \boldsymbol{u}\|^2 + \frac{c_* \sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})} \right) \tag{6}$$

valid for all $\boldsymbol{\mu} \in \mathcal{S}^{\uparrow}$ where either $C_* = 6, c_* = 1$ (Chatterjee et al., 2015, inequality (18)) or $C_* = 4, c_* = 4$ (Chatterjee et al., 2015, inequality (30)). Here, $k(\boldsymbol{u}) \geq 1$ for $\boldsymbol{u} = (u_1, \dots, u_n) \in \mathcal{S}^{\uparrow}$ is the integer such that $k(\boldsymbol{u}) - 1$ is the number of inequalities $u_i \leq u_{i+1}$ that are strict for $i = 1, \dots, n-1$ (number of jumps of $\boldsymbol{u}$). Inequality (6) implies (up to a logarithmic factor) a bound as in (5) and also gives some more insight into the problem. For example, (6) shows that the fast rate $\frac{\log n}{n}$ is achieved if $\boldsymbol{\mu}$ has only one jump or a fixed, independent of n, number of jumps. This is not granted by (5).

Along with the least squares estimator, one may consider estimation of monotone functions via penalized least squares with total variation penalty. The corresponding estimator $\hat{\boldsymbol{\mu}}^{TV}$ is defined as

$$\hat{\boldsymbol{\mu}}^{TV} \in \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^n} \left(\frac{1}{2} \|\boldsymbol{u} - \mathbf{y}\|^2 + \lambda \sum_{i=1}^{n-1} |u_{i+1} - u_i| \right), \tag{7}$$

where $\lambda > 0$ is a tuning parameter. Statistical properties of this estimator were first studied in Mammen and van de Geer (1997) where it was shown that $\|\hat{\boldsymbol{\mu}}^{TV} - \boldsymbol{\mu}\|$ attains the optimal rate $n^{-1/3}$ in probability on the class of functions of bounded variation (and thus on $\mathcal{S}^{\uparrow}(V)$). Recently, the performance of $\hat{\boldsymbol{\mu}}^{TV}$ was analyzed in Dalalyan et al. (2014) by considering $\hat{\boldsymbol{\mu}}^{TV}$ as a special instance of the Lasso estimator. If $\boldsymbol{\mu}^{\uparrow}$ is the projection of $\boldsymbol{\mu}$ onto $\mathcal{S}^{\uparrow}$, $\delta \in (0,1)$ is a constant, and the tuning parameter λ is given by

$$\lambda = \sigma \sqrt{\frac{\log(n/\delta)}{k^* n}} \qquad \text{where } k^* = \left(\frac{V(\boldsymbol{\mu}^{\uparrow})^2 n \log(n/\delta)}{\sigma^2} \right)^{1/3}, \tag{8}$$

the estimator $\hat{\boldsymbol{\mu}}^{TV}$ satisfies with probability greater than $1 - 2\delta$ the following oracle inequality (Dalalyan et al., 2014, Proposition 6):

$$\begin{aligned} \left\| \hat{\boldsymbol{\mu}}^{TV} - \boldsymbol{\mu} \right\|^2 \;\leq\; & \left\| \boldsymbol{\mu}^{\uparrow} - \boldsymbol{\mu} \right\|^2 + 6 \left(\frac{\sigma^2 V(\boldsymbol{\mu}^{\uparrow}) \sqrt{\log(n/\delta)}}{n} \right)^{2/3} \\ & + \frac{2\sigma^2 (1 + 2\log(1/\delta))}{n} \end{aligned} \tag{9}$$

for all $\boldsymbol{\mu} \in \mathbb{R}^n$. It follows from (9) that if the tuning parameter is chosen correctly, the estimator $\hat{\boldsymbol{\mu}}^{TV}$ achieves, up to a logarithmic factor, the minimax rate $n^{-2/3}$ in probability on the class $\mathcal{S}^{\uparrow}(V)$. Also, (9) implies a bound for the excess losses $\|\hat{\boldsymbol{\mu}}^{TV} - \boldsymbol{\mu}\|^i - \min_{\boldsymbol{u} \in \mathcal{S}^{\uparrow}(V)} \|\boldsymbol{u} - \boldsymbol{\mu}\|^i$, $i = 1, 2$, corresponding to the class $\mathcal{S}^{\uparrow}(V)$. However, (9) does not allow us to evaluate the expected regrets $R_i(\hat{\boldsymbol{\mu}}^{TV}, \boldsymbol{\mu})$ since $\hat{\boldsymbol{\mu}}^{TV}$ depends on δ. It is also shown in (Dalalyan et al., 2014, Proposition 4) that if $\lambda = 2\sigma\sqrt{(2/n)\log(n/\delta)}$, the estimator $\hat{\boldsymbol{\mu}}^{TV}$ satisfies

$$\left\| \hat{\boldsymbol{\mu}}^{TV} - \boldsymbol{\mu} \right\|^2 \leq \min_{\boldsymbol{u} \in \mathbb{R}^n} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{4\sigma^2 k(\boldsymbol{u}) \log(n/\delta)}{n} r_n(\boldsymbol{u}) \right) \tag{10}$$

with probability greater than $1-2\delta$, where $k(\boldsymbol{u})-1$ for $\boldsymbol{u} \in \mathbb{R}^n$ is the number of jumps of $\boldsymbol{u}$, i.e., the cardinality of the set $\{i \in \{1, ..., n-1\} : u_i \neq u_{i+1}\}$, $r_n(\boldsymbol{u}) = 3 + 256(\log(n) + (n/\Delta(\boldsymbol{u})))$ and $\Delta(\boldsymbol{u})$ is the minimum distance between two jumps in the sequence $\boldsymbol{u}$:

$$\Delta(\boldsymbol{u}) = \min\{d \geq 1 : \ \exists k \in \{1, ..., n\} \text{ with } u_{k+1} \neq u_k \text{ and } u_{k+d+1} \neq u_{k+d}\}.$$

The expressions on the right hand sides of (6) and (10) are small if the unknown sequence $\boldsymbol{\mu}$ is well approximated by a piecewise constant sequence with not too many pieces. In this regard, the two bounds have some similarity to sparsity oracle inequalities in high-dimensional linear regression (cf. Rigollet and Tsybakov, 2011, 2012; Tsybakov, 2014). This similarity can be easily explained as follows. Write (1) in the equivalent form

$$\mathbf{y} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\xi},$$

with the matrix $\mathbb{X} = (X_{ij})_{i=1,...,n,\ j=1,...,n}$ where $X_{ij} = 1$ if $j \leq i$ and $X_{ij} = 0$ otherwise, and $\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_n^*)$ where $\beta_1^* = \mu_1$ and $\beta_i^* = \mu_i - \mu_{i-1}$ for $i = 2, \dots, n$. With this notation, $k(\boldsymbol{\mu}) \in \{|\boldsymbol{\beta}^*|_0, 1 + |\boldsymbol{\beta}^*|_0\}$, where $|\boldsymbol{\beta}^*|_0$ denotes the number of non-zero components of $\boldsymbol{\beta}^*$. The value $k(\boldsymbol{\mu})$ is small when $\boldsymbol{\beta}^*$ is sparse. Thus, the problem of estimation of piecewise constant sequence $\boldsymbol{\mu}$ with small number of pieces can be considered as the problem of prediction in sparse linear regression with a specific design matrix $\mathbb{X}$. Similarly, we may write $\boldsymbol{u} = \mathbb{X}\boldsymbol{\beta}$, for $\boldsymbol{\beta}$ with components $\beta_1 = u_1$ and $\beta_i = u_i - u_{i-1}$ for $i = 2, \dots, n$. These remarks suggest that we can apply the theory of sparsity oracle inequalities, in particular, sparsity pattern aggregation (cf. Rigollet and Tsybakov, 2011, 2012; Tsybakov, 2014) in the context of monotone estimation described above. Similar observation is valid for estimation under convexity constraints (see Section 3 below). In the present paper, we develop this argument using as a building block the Q-aggregation procedures Rigollet (2012); Dai et al. (2012, 2014); Bellec (2014). In particular, we construct an estimator $\hat{\boldsymbol{\mu}}$ such that

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in \mathcal{S}^{\uparrow}} \left(\|\boldsymbol{\mu} - \boldsymbol{u}\|^2 + \frac{c\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})} \right), \quad \forall\ \boldsymbol{\mu} \in \mathbb{R}^n, \tag{11}$$

for some absolute constant $c > 0$. Note that (11) is a sharp oracle inequality (i.e., an inequality with leading constant 1). It improves upon the oracle inequality (6) for the least squares estimator where the leading constant C_* is noticeably greater than 1 and the bound is valid only for $\boldsymbol{\mu} \in \mathcal{S}^{\uparrow}$. The advantage of having leading constant 1 and arbitrary $\boldsymbol{\mu}$ in (11) is that it allows us to derive bounds on the excess risk and on the minimax regret, which was not possible based on the previous results. We also obtain sharp oracle inequalities with high probability for the same estimator. In addition, we show that it satisfies stronger sharp inequalities with the minimum $\min_{\boldsymbol{u} \in \mathcal{S}^{\uparrow}}$ on the right hand side of (11) replaced by $\min_{\boldsymbol{u} \in \mathbb{R}^n}$. This implies that our results are invariant to the direction of monotonicity; they remain valid if we replace everywhere monotone increasing by monotone decreasing functions. Finally, we derive similar results for the problem of estimation under the convexity constraints improving an oracle inequality obtained in Guntuboyina and Sen (2013).

2. Sparsity Pattern Aggregation for Piecewise Constant Sequences

For any non-empty set $J \subseteq \{1, ..., n-1\}$, let $|J|$ denote the cardinality of J and define

$$\pi_J \coloneqq \frac{\exp(-|J|)}{H \binom{n-1}{|J|}}, \qquad H \coloneqq \sum_{i=0}^{n-1} \exp(-i). \tag{12}$$

Let $P_J \in \mathbb{R}^{n\times n}$ be the projector on the linear subspace V_J of $\mathbb{R}^n$ defined by

$$V_J \coloneqq \Big\{ \boldsymbol{u} \in \mathbb{R}^n : \ \forall i \in \{1, ..., n-1\} \setminus J, \ u_{i+1} = u_i \Big\}. \tag{13}$$

In words, V_J is the space of all piecewise constant sequences that have jumps only at points in J. Given a vector $\mathbf{y}$ of observations and $\boldsymbol{\theta} = (\theta_J)_{J\subseteq\{1,...,n-1\}}$ where each $\theta_J \in \mathbb{R}$, let

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \sum_{J\subseteq\{1,...,n-1\}} \theta_J P_J \mathbf{y}. \tag{14}$$

Finally, let

$$\hat{\boldsymbol{\mu}}^Q = \boldsymbol{\mu}_{\hat{\boldsymbol{\theta}}} \tag{15}$$

where $\hat{\boldsymbol{\theta}}$ is the solution of the optimization problem

$$\min_{\boldsymbol{\theta}\in\Lambda} \quad \|\boldsymbol{\mu}_{\boldsymbol{\theta}} - \mathbf{y}\|^2 + \sum_{J\subseteq\{1,...,n-1\}} \theta_J \left(\frac{2\sigma^2|J|}{n} + \frac{1}{2}\|\boldsymbol{\mu}_{\boldsymbol{\theta}} - P_J\mathbf{y}\|^2 + \frac{46\sigma^2}{n} \log\frac{1}{\pi_J} \right)$$

where

$$\Lambda = \left\{ \boldsymbol{\theta} : \ \theta_J \geq 0 \text{ for all } J \subseteq \{1, ..., n-1\}, \text{ and } \sum_{J\subseteq\{1,...,n-1\}} \theta_J = 1 \right\}.$$

This optimization problem is a convex quadratic program with a simplex constraint. It performs aggregation of the linear estimators $(P_J\mathbf{y})_{J\subseteq\{1,...,n-1\}}$ using the Q-aggregation procedure Dai et al. (2012, 2014); Bellec (2014) with the prior weights (12). As the size of this quadratic program is of order 2^n, it is a computationally hard problem. The estimator $\hat{\boldsymbol{\mu}}^Q$ satisfies the following sharp oracle inequalities.

Theorem 1 *Let $\boldsymbol{\mu} \in \mathbb{R}^n$, $n \geq 2$, and assume that the noise vector $\boldsymbol{\xi}$ has distribution $\mathcal{N}(0, \sigma^2 I_{n\times n})$. There exist absolute constants $c, c' > 0$ such that for all $\delta \in (0, 1/3)$, the estimator $\hat{\boldsymbol{\mu}}^Q$ satisfies with probability at least $1 - 3\delta$,*

$$\left\|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\right\|^2 \leq \min_{\boldsymbol{u}\in\mathbb{R}^n} \left(\|\boldsymbol{\mu} - \boldsymbol{u}\|^2 + \frac{c\sigma^2 k(\boldsymbol{u}))}{n} \log\frac{en}{k(\boldsymbol{u})} \right) + \frac{c\sigma^2 \log(1/\delta)}{n}, \tag{16}$$

and

$$\mathbb{E}_{\boldsymbol{\mu}} \left\|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\right\|^2 \leq \min_{\boldsymbol{u}\in\mathbb{R}^n} \left(\|\boldsymbol{\mu} - \boldsymbol{u}\|^2 + \frac{c'\sigma^2 k(\boldsymbol{u})}{n} \log\frac{en}{k(\boldsymbol{u})} \right). \tag{17}$$

Proof Let $J \subseteq \{1, ..., n-1\}$. Denote by $d = |J| + 1$ the dimension of the subspace V_J. Then, the projection estimator $P_J \mathbf{y}$ satisfies with probability at least $1 - \delta$ (see, for example, Hsu et al. (2012)):

$$\begin{aligned} \|P_J \mathbf{y} - \boldsymbol{\mu}\|^2 &\le \|P_J \boldsymbol{\mu} - \boldsymbol{\mu}\|^2 + \frac{d + 2\sqrt{d \log(1/\delta)} + 2\log(1/\delta)}{n} \\ &\le \min_{\boldsymbol{u} \in V_J} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{2(|J|+1) + 3\log(1/\delta)}{n}. \end{aligned} \tag{18}$$

The sharp oracle inequality from Bellec (2014) yields that with probability at least $1 - 2\delta$ for all $J \subseteq \{1, ..., n-1\}$ we have

$$\left\| \hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu} \right\|^2 \le \|P_J \mathbf{y} - \boldsymbol{\mu}\|^2 + C\sigma^2 \log \frac{1}{\pi_J} + C\sigma^2 \log(1/\delta), \tag{19}$$

for some absolute constant $C > 0$. Combining (18) and (19) with the union bound and the inequality (cf. (Rigollet and Tsybakov, 2012, (5.4))) $\log(1/\pi_J) \le 2(|J|+1)\log(en/(|J|+1)) + 1/2$, we find that with probability at least $1 - 3\delta$,

$$\begin{aligned} \left\| \hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu} \right\|^2 \le \min_{J \subseteq \{1,...,n-1\}} \min_{\boldsymbol{u} \in V_J} \Bigg(\|\boldsymbol{\mu} - \boldsymbol{u}\|^2 + \frac{c\sigma^2(|J|+1)}{n} \log\left(\frac{en}{|J|+1}\right) \Bigg) \\ + c\sigma^2 \log(1/\delta) \end{aligned}$$

where $c > 0$ is an absolute constant. Since we have that $|J| + 1 = k(\boldsymbol{u})$ for all $\boldsymbol{u} \in V_J$ and also that $\min_{J \subseteq \{1,...,n-1\}} \min_{\boldsymbol{u} \in V_J} = \min_{\boldsymbol{u} \in \mathbb{R}^n}$, the bound (16) follows. Finally, (17) is obtained from (16) by integration. ■

We now discuss some corollaries of Theorem 1. First, it follows that (11) is satisfied for $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^Q$, so the remarks after (11) apply. Next, in view of (17), for the class of monotone sequences with at most k jumps $\mathcal{S}_k^{\uparrow} = \{\boldsymbol{u} \in \mathcal{S}^{\uparrow} : k(\boldsymbol{u}) \le k\}$ we have the following bounds for the maximal expected regrets

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^n} \left(\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\| - \min_{\boldsymbol{u} \in \mathcal{S}_k^{\uparrow}} \|\boldsymbol{u} - \boldsymbol{\mu}\| \right) \le c\sqrt{\frac{\sigma^2 k}{n} \log\left(\frac{en}{k}\right)}, \tag{20}$$

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^n} \left(\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\|^2 - \min_{\boldsymbol{u} \in \mathcal{S}_k^{\uparrow}} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 \right) \le \frac{c\sigma^2 k}{n} \log\left(\frac{en}{k}\right), \tag{21}$$

where $c > 0$ is an absolute constant. The same bounds hold for the minimax risks over $\mathcal{S}_k^{\uparrow}$ since the minimax risk is smaller than the minimax regret. Theorem 4 below shows that the bounds (20) and (21) are optimal up to logarithmic factors.

Finally, consider the consequences of Theorem 1 for the class $\mathcal{S}^{\uparrow}(V)$. To this end, define the integer k^* such that

$$k^* = \min \left\{ m \in \mathbb{N} : m \ge \left(\frac{V(\boldsymbol{\mu})^2 n}{\sigma^2 \log(en)} \right)^{1/3} \right\}$$

if the set $\left\{m \in \mathbb{N} : m \geq \left(\frac{V(\boldsymbol{\mu})^2 n}{\sigma^2 \log(en)}\right)^{1/3}\right\}$ is non-empty, and $k^* = 1$ otherwise. We will need the following lemma.

Lemma 2 *Let $\boldsymbol{\mu} \in \mathcal{S}^{\uparrow}$ and let $1 \leq k \leq n$ be an integer. Then there exists a sequence $\bar{\boldsymbol{u}} \in \mathcal{S}_k^{\uparrow}$ such that*

$$\|\bar{\boldsymbol{u}} - \boldsymbol{\mu}\| \leq \frac{V(\boldsymbol{\mu})}{2k}. \tag{22}$$

Next, there exists a sequence $\bar{\boldsymbol{u}} \in \mathcal{S}_{k^*}^{\uparrow}$ such that

$$\|\bar{\boldsymbol{u}} - \boldsymbol{\mu}\|^2 \leq \frac{1}{4} \max\left(\left(\frac{\sigma^2 V(\boldsymbol{\mu}) \log(en)}{n}\right)^{2/3}, \frac{\sigma^2 \log(en)}{n}\right). \tag{23}$$

In addition,

$$\frac{\sigma^2 k^*}{n} \log \frac{en}{k^*} \leq 2 \max\left(\left(\frac{\sigma^2 V(\boldsymbol{\mu}) \log(en)}{n}\right)^{2/3}, \frac{\sigma^2 \log(en)}{n}\right). \tag{24}$$

Proof To construct the sequence $\bar{\boldsymbol{u}}$, consider the k intervals

$$I_j = \Big[\mu_1 + \frac{j-1}{k} V(\boldsymbol{\mu}), \mu_1 + \frac{j}{k} V(\boldsymbol{\mu})\Big), \qquad j = 1, ..., k-1, \tag{25}$$

and $I_k = [\mu_1 + \frac{k-1}{k} V(\boldsymbol{\mu}), \mu_n]$. For all $j = 1, ..., k$, let

$$J_j = \{i = 1, ..., n : \quad \mu_i \in I_j\}. \tag{26}$$

For any $i \in \{1, ..., n\}$ there exists a unique $j \in \{1, ..., k\}$ such that $i \in I_j$. Let $\bar{u}_i = \mu_1 + \frac{j-1/2}{k} V(\boldsymbol{\mu})$ for all $i \in I_j$. Then the sequence $\bar{\boldsymbol{u}} = (\bar{u}_1, \ldots, \bar{u}_n)$ is non-decreasing, it has at most k pieces, i.e., $k(\bar{\boldsymbol{u}}) \leq k$, and $|\bar{u}_i - \mu_i| \leq \frac{V(\boldsymbol{\mu})}{2k}$ for $i = 1, ..., n$. Thus (22) follows. Next, note that if $k^* = 1$, then $V(\boldsymbol{\mu})^2 \leq \sigma^2 \log(en)/n$. If $k^* > 1$, then by definition of k^*, $V(\boldsymbol{\mu})^2/(k^*)^2 \leq (\sigma^2 V(\boldsymbol{\mu}) \log(en)/n)^{2/3}$. Thus, (23) follows. The bound (24) is straightforward by studying the cases $k^* = 1$ and $k^* > 1$ separately. ■

We can now derive the following corollary of Theorem 1.

Corollary 3 *Under the assumptions of Theorem 1, there exists an absolute constant $c > 0$ such that, for any $\boldsymbol{\mu} \in \mathcal{S}^{\uparrow}$,*

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\|^2 \leq c \max\left(\left(\frac{\sigma^2 V(\boldsymbol{\mu}) \log n}{n}\right)^{2/3}, \frac{\sigma^2 \log n}{n}\right). \tag{27}$$

In addition, for any $V > 0$ and any $\boldsymbol{\mu} \in \mathbb{R}^n$ the expected regret of $\hat{\boldsymbol{\mu}}^Q$ satisfies

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\| - \min_{\boldsymbol{u} \in \mathcal{S}^{\uparrow}(V)} \|\boldsymbol{u} - \boldsymbol{\mu}\| \ \leq c \max\left(\left(\frac{\sigma^2 V \log n}{n}\right)^{1/3}, \sigma\sqrt{\frac{\log n}{n}}\right) \tag{28}$$

where $c > 0$ is an absolute constant.

Proof Inequality (27) is straightforward in view of (17), (23), and (24). To prove (28), fix any $\boldsymbol{\mu} \in \mathbb{R}^n$ and consider

$$\boldsymbol{\mu}^* \in \operatorname*{argmin}_{\boldsymbol{\mu}' \in \mathcal{S}^\uparrow(V)} \|\boldsymbol{\mu}' - \boldsymbol{\mu}\|.$$

From (17) and the fact that the function $x \mapsto x \log\left(\frac{en}{x}\right)$ is increasing for $1 \le x \le n$ we get

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\| &\le \min_{\boldsymbol{u} \in \mathcal{S}^\uparrow_{k^*}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\| + \sqrt{c' \frac{\sigma^2 k^*}{n} \log\left(\frac{en}{k^*}\right)} \right) \\
&\le \min_{\boldsymbol{u} \in \mathcal{S}^\uparrow_{k^*}} \|\boldsymbol{u} - \boldsymbol{\mu}^*\| + \|\boldsymbol{\mu}^* - \boldsymbol{\mu}\| + \sqrt{c' \frac{\sigma^2 k^*}{n} \log\left(\frac{en}{k^*}\right)} \\
&\le \|\boldsymbol{\mu}^* - \boldsymbol{\mu}\| + c'' \max\left(\left(\frac{\sigma^2 V \log n}{n} \right)^{1/3}, \sigma \sqrt{\frac{\log n}{n}} \right)
\end{aligned}$$

for an absolute constant $c'' > 0$ where the last inequality follows from (23) and (24). ■

The estimator $\hat{\boldsymbol{\mu}}^Q$ shown in Theorem 1 satisfies the sharp oracle inequalities both in expectation and with high probability. Previous results for the least squares estimator Chatterjee et al. (2015) were only obtained in expectation and the results on the ℓ_1-penalized estimator (7) are only obtained with high probability.

Finally, the following result shows that the upper bounds (20) and (21) are optimal up to logarithmic factors.

Proposition 4 *Let $n \ge 2, V > 0$ and $\sigma > 0$. There exist absolute constants $c, c' > 0$ such that for any positive integer $k \le n$ satisfying $k^3 \le 16nV^2/\sigma^2$ we have*

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathcal{S}^\uparrow_k \cap \mathcal{S}^\uparrow(V)} \mathbb{P}_{\boldsymbol{\mu}} \left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \ge \frac{c\sigma^2 k}{n} \right) > c', \tag{29}$$

where $\mathbb{P}_{\boldsymbol{\mu}}$ denotes the distribution of $\mathbf{y}$ satisfying (1) and $\inf_{\hat{\boldsymbol{\mu}}}$ is the infimum over all estimators.

For $k = 1, ..., n$, take any $V > 0$ large enough to satisfy $k^3 \le 16nV^2/\sigma^2$. Then, Theorem 4 and Markov's inequality yield the following lower bounds on the minimax risks over the class $\mathcal{S}^\uparrow_k$:

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathcal{S}^\uparrow_k} \mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \ge c \sqrt{\frac{c'\sigma^2 k}{n}}, \quad \inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathcal{S}^\uparrow_k} \mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \ge \frac{cc'\sigma^2 k}{n}. \tag{30}$$

As the minimax risk is smaller than the minimax regret, (30) also provides lower bounds for the corresponding minimax regrets over $\mathcal{S}^\uparrow_k$. Combining this with (20) and (21) we find that the estimator $\hat{\boldsymbol{\mu}}^Q$ achieves up to logarithmic factors the optimal rate with respect to the minimax regret.

Next, Proposition 4 implies the following lower bound on the minimax deviation risk on $\mathcal{S}^\uparrow(V)$.

Corollary 5 *Let $n \geq 2, V > 0$ and $\sigma > 0$. There exist absolute constants $c, c' > 0$ such that*

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathcal{S}^{\uparrow}(V)} \mathbb{P}_{\boldsymbol{\mu}} \left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \geq c \max \left\{ \left(\frac{\sigma^2 V}{n} \right)^{2/3}, \frac{\sigma^2}{n} \right\} \right) > c'. \tag{31}$$

To prove this corollary it is enough to note that if $16nV^2/\sigma^2 \geq 1$, by choosing k in Proposition 4 as the integer part of $(16nV^2/\sigma^2)^{1/3}$, we obtain the lower bound corresponding to $\left(\frac{\sigma^2 V}{n}\right)^{2/3}$ under the maximum in (31). On the other hand, if $16nV^2/\sigma^2 < 1$ the term $\frac{\sigma^2}{n}$ is dominant, so that we need to have the lower bound of the order $\frac{\sigma^2}{n}$, which is trivial (it follows from a reduction to the bound for the class composed of two constant functions).

It follows from (31) and (27) that the estimator $\hat{\boldsymbol{\mu}}^Q$ achieves, up to logarithmic factors, the optimal rate with respect to the minimax risk on the class $\mathcal{S}^{\uparrow}(V)$. Using (28) and the fact that the minimax risk is smaller than the minimax regret, we conclude that it is also the optimal rate up to logarithmic factors for the minimax regret.

Proof [Proof of Theorem 4] We assume for simplicity that n is a multiple of k. The general case is treated analogously. For any $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \{0,1\}^k$, let $d_H(\boldsymbol{\omega}, \boldsymbol{\omega}') = |\{i = 1, ..., k : \omega_i \neq \omega_i'\}|$ be the Hamming distance between $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$. By the Varshamov-Gilbert bound (Tsybakov, 2009, Lemma 2.9), there exists a set $\Omega \subset \{0,1\}^k$ such that

$$\mathbf{0} = (0, ..., 0) \in \Omega, \quad \log(|\Omega| - 1) \geq k/8, \quad \text{and} \quad d_H(\boldsymbol{\omega}, \boldsymbol{\omega}') > k/8 \tag{32}$$

for any two distinct $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$. For each $\boldsymbol{\omega} \in \Omega$, define a vector $\boldsymbol{u}^{\boldsymbol{\omega}} \in \mathbb{R}^n$ with components

$$u_i^{\boldsymbol{\omega}} = \frac{\lfloor (i-1)k/n \rfloor \, V}{2k} + \gamma \omega_{\lfloor (i-1)k/n \rfloor + 1}, \qquad i = 1, ..., n,$$

where $\gamma = (1/8)\sqrt{\sigma^2 k/n}$, and $\lfloor x \rfloor$ denotes the maximal integer smaller than x. For any $\boldsymbol{\omega} \in \Omega$, $\boldsymbol{u}^{\boldsymbol{\omega}}$ is a piecewise constant sequence with $k(\boldsymbol{u}^{\boldsymbol{\omega}}) \leq k$, $\boldsymbol{u}^{\boldsymbol{\omega}}$ is a non-decreasing sequence because $\gamma \leq V/(2k)$, and by construction $V(\boldsymbol{u}^{\boldsymbol{\omega}}) \leq V$. Thus, $\boldsymbol{u}^{\boldsymbol{\omega}} \in \mathcal{S}_k^{\uparrow} \cap \mathcal{S}^{\uparrow}(V)$ for all $\boldsymbol{\omega} \in \Omega$. Moreover, for any $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$,

$$\|\boldsymbol{u}^{\omega} - \boldsymbol{u}^{\omega'}\|^2 = \frac{\gamma^2}{k} d_H(\boldsymbol{\omega}, \boldsymbol{\omega}') \geq \frac{\gamma^2}{8} = \frac{\sigma^2 k}{512 n}. \tag{33}$$

Set for brevity $P_{\boldsymbol{\omega}} = \mathbb{P}_{\boldsymbol{u}^{\omega}}$. The Kullback-Leibler divergence $K(P_{\boldsymbol{\omega}}, P_{\boldsymbol{\omega}'})$ between $P_{\boldsymbol{\omega}}$ and $P_{\boldsymbol{\omega}'}$ is equal to $\frac{n}{2\sigma^2}\|\boldsymbol{u}^{\omega} - \boldsymbol{u}^{\omega'}\|^2$ for all $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$. Thus,

$$K(P_{\boldsymbol{\omega}}, P_{\mathbf{0}}) = \frac{\gamma^2 n d_H(\mathbf{0}, \boldsymbol{\omega})}{2k\sigma^2} \leq \frac{k}{128} \leq \frac{\log(|\Omega| - 1)}{16}. \tag{34}$$

Applying (Tsybakov, 2009, Theorem 2.7) with $\alpha = 1/16$ completes the proof. ■

3. Estimation of Convex Sequences by Aggregation

Assume that $n \geq 3$ and define the set of convex sequences $\mathcal{S}^{\mathrm{C}}$ as follows:

$$\mathcal{S}^{\mathrm{C}} = \{\boldsymbol{u} = (u_1, \ldots, u_n) \in \mathbb{R}^n : \; 2u_i \leq u_{i+1} + u_{i-1}, \; i = 2, \ldots, n-1\}. \tag{35}$$

For any $\boldsymbol{u} \in \mathbb{R}^n$, we introduce the integer $q(\boldsymbol{u}) \geq 1$ such that $q(\boldsymbol{u}) - 1$ is the cardinality of the set $\{i = 1, ..., n-1 : 2u_i \neq u_{i+1} + u_{i-1}\}$. If $\boldsymbol{u} \in \mathcal{S}^{\mathrm{C}}$, $q(\boldsymbol{u}) - 1$ is the number of inequalities $2u_i \leq u_{i+1} + u_{i-1}$ that are strict for $i = 2, ..., n-1$. The value $q(\boldsymbol{u})$ is small if $\boldsymbol{u}$ is a piecewise linear sequence with a small number of pieces.

The performance of the least squares estimator over convex sequences $\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}^{\mathrm{C}})$ has been recently studied in Guntuboyina and Sen (2013). If the unknown vector $\boldsymbol{\mu}$ belongs to the set $\mathcal{S}^{\mathrm{C}}$, Guntuboyina and Sen (2013) shows that the estimator $\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}^{\mathrm{C}})$ satisfies the risk bound

$$\mathbb{E}_{\boldsymbol{\mu}} \left\| \hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}^{\mathrm{C}}) - \boldsymbol{\mu} \right\|^2 \leq c \log(en)^{5/4} \left(\frac{\sigma^2 \sqrt{R(\boldsymbol{\mu})}}{n} \right)^{4/5},$$

where $R(\boldsymbol{\mu}) = \max(1, \min\{\|\boldsymbol{\tau} - \boldsymbol{\mu}\|^2, \boldsymbol{\tau} \text{ is affine}\})$ and $c > 0$ is an absolute constant. It is proved in (Chatterjee et al., 2015, Example 2.3) that the least squares estimator $\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}^{\mathrm{C}})$ satisfies the oracle inequality

$$\mathbb{E}_{\boldsymbol{\mu}} \left\| \hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}^{\mathrm{C}}) - \boldsymbol{\mu} \right\|^2 \leq 6 \min_{\boldsymbol{u} \in \mathcal{S}^{\mathrm{C}}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{c\sigma^2 q(\boldsymbol{u}) \log\left(\frac{en}{q(\boldsymbol{u})}\right)^{5/4}}{n} \right), \tag{36}$$

where $c > 0$ is an absolute constant. The right hand side of (36) is small if the unknown vector $\boldsymbol{\mu}$ can be well approximated by a piecewise linear sequence in $\mathcal{S}^{\mathrm{C}}$ with not too many pieces.

The leading constant in (36) is 6. We will show that sparsity pattern aggregation achieves a substantially better performance. We obtain the sharp oracle inequality (39) below, improving upon (36) not only in the fact that the leading constant is 1 but also in the rate of the remainder term; we will see that the exponent 5/4 of the logarithmic factor is reduced to 1.

For any set $J \subseteq \{2, ..., n-1\}$, define

$$\nu_J \coloneqq \frac{\exp(-|J|)}{H_C \binom{n-2}{|J|}}, \qquad H_C \coloneqq \sum_{i=0}^{n-2} \exp(-i). \tag{37}$$

Let $Q_J \in \mathbb{R}^{n \times n}$ be the projector on the linear subspace W_J of $\mathbb{R}^n$ given by

$$W_J \coloneqq \Big\{ \boldsymbol{u} \in \mathbb{R}^n : \ \forall i \in \{2, ..., n-1\} \setminus J, \ 2u_i = u_{i+1} + u_{i-1} \Big\}.$$

Given a vector $\mathbf{y}$ of observations and $\boldsymbol{\theta} = (\theta_J)_{J \subseteq \{2,...,n-1\}}$ where each θ_J belongs to $\mathbb{R}$, let

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \sum_{J \subseteq \{2,...,n-1\}} \theta_J Q_J \mathbf{y}.$$

Finally, let

$$\hat{\boldsymbol{\mu}}^{Q-conv} = \boldsymbol{\mu}_{\hat{\boldsymbol{\theta}}}$$

where $\hat{\boldsymbol{\theta}}$ is the solution of the optimization problem

$$\min_{\boldsymbol{\theta} \in \Lambda'} \ \|\boldsymbol{\mu}_{\boldsymbol{\theta}} - \mathbf{y}\|^2 + \sum_{J \subset \{2,...,n-1\}} \theta_J \left(\frac{2\sigma^2 |J|}{n} + \frac{1}{2} \|\boldsymbol{\mu}_{\boldsymbol{\theta}} - Q_J \mathbf{y}\|^2 + \frac{46\sigma^2}{n} \log \frac{1}{\nu_J} \right)$$

where

$$\Lambda' = \left\{ \boldsymbol{\theta} : \ \theta_J \geq 0 \text{ for all } J \subseteq \{2, ..., n-1\}, \text{ and } \sum_{J \subseteq \{2,...,n-1\}} \theta_J = 1 \right\}.$$

The structure of this minimization problem is the same as of its analog introduced in Section 2. This is a quadratic program that aggregates the linear estimators $(Q_J \mathbf{y})_{J \subseteq \{2,...,n-1\}}$ using the Q-aggregation procedure Dai et al. (2012, 2014); Bellec (2014) with the prior weights (37).

Theorem 6 *Let $\boldsymbol{\mu} \in \mathbb{R}^n$, $n \geq 3$, and assume that the noise vector $\boldsymbol{\xi}$ has distribution $\mathcal{N}(0, \sigma^2 I_{n \times n})$. There exist absolute constants $c, c' > 0$ such that for all $\delta \in (0, 1/3)$, the estimator $\hat{\boldsymbol{\mu}}^{Q-conv}$ satisfies with probability at least $1 - 3\delta$,*

$$\left\| \hat{\boldsymbol{\mu}}^{Q-conv} - \boldsymbol{\mu} \right\|^2 \leq \min_{\boldsymbol{u} \in \mathbb{R}^n} \left(\|\boldsymbol{\mu} - \boldsymbol{u}\|^2 + \frac{c\sigma^2 q(\boldsymbol{u})}{n} \log \frac{en}{q(\boldsymbol{u})} \right) + \frac{c\sigma^2 \log(1/\delta)}{n}, \tag{38}$$

and we have

$$\mathbb{E}_{\boldsymbol{\mu}} \left\| \hat{\boldsymbol{\mu}}^{Q-conv} - \boldsymbol{\mu} \right\|^2 \leq \min_{\boldsymbol{u} \in \mathbb{R}^n} \left(\|\boldsymbol{\mu} - \boldsymbol{u}\|^2 + \frac{c'\sigma^2 q(\boldsymbol{u})}{n} \log \frac{en}{q(\boldsymbol{u})} \right). \tag{39}$$

The proof of this theorem is the same as that of Theorem 1 with the only difference that J is now a subset of $\{2, ..., n-1\}$ rather than that of $\{1, ..., n-1\}$, and we replace the notation P_J and V_J by Q_J and W_J respectively.

The leading constant of the oracle inequality (39) is 1, and the remainder term is proportional to $q(\boldsymbol{u}) \log(en/q(\boldsymbol{u}))$. These are two improvements upon (36), where the leading constant is 6 and the remainder term is proportional to $q(\boldsymbol{u}) \log(en/q(\boldsymbol{u}))^{5/4}$.

In view of (39), for the class of piecewise linear convex sequences with at most q linear pieces, $\mathcal{S}_q^{\mathrm{C}} = \{\boldsymbol{u} \in \mathcal{S}^{\mathrm{C}} : q(\boldsymbol{u}) \leq q\}$ we have the following bounds for the maximal expected regrets

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^n} \left(\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\| - \min_{\boldsymbol{u} \in \mathcal{S}_q^{\mathrm{C}}} \|\boldsymbol{u} - \boldsymbol{\mu}\| \right) \leq c \sqrt{\frac{\sigma^2 q}{n} \log\left(\frac{en}{q}\right)}, \tag{40}$$

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^n} \left(\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\|^2 - \min_{\boldsymbol{u} \in \mathcal{S}_q^{\mathrm{C}}} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 \right) \leq \frac{c\sigma^2 q}{n} \log\left(\frac{en}{q}\right), \tag{41}$$

where $c > 0$ is an absolute constant. The same bounds hold for the minimax risks over $\mathcal{S}_q^{\mathrm{C}}$ since the minimax risk is smaller than the minimax regret.

The following proposition shows that the rates of convergence in (40) and (41) are optimal up to logarithmic factors. We omit the discussion since it is similar to that after Theorem 4.

Proposition 7 *Let $n \geq 3$. There exist absolute constants $c, c' > 0$ such that, for any positive integer $q \leq n$,*

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathcal{S}_q^{\mathrm{C}}} \mathbb{P}_{\boldsymbol{\mu}} \left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \geq \frac{c\sigma^2 q}{n} \right) > c', \tag{42}$$

where the infimum is taken over all estimators.

Proof Assume that $q \geq 2$ since for $q = 1$ the result is trivial. We also assume for simplicity that n is a multiple of q. Let $m = n/q$ and $\gamma = (1/8)\sqrt{\sigma^2 q/n}$. Set $\beta_0 = 0, \alpha_0 = 0$ and define, for all integers $j \geq 1$,

$$\beta_j = \beta_{j-1} + \gamma + m\alpha_{j-1}, \qquad \alpha_j = 2\gamma + \alpha_{j-1}. \tag{43}$$

By the Varshamov-Gilbert bound (Tsybakov, 2009, Lemma 2.9) there exists $\Omega \subset \{0,1\}^q$ such that (32) is satisfied, with k replaced by q. For each $\boldsymbol{\omega} \in \Omega$, define a vector $\boldsymbol{u}^{\boldsymbol{\omega}} \in \mathbb{R}^n$ with components

$$u^{\boldsymbol{\omega}}_{jm+i} = \omega_{j+1}\gamma + \alpha_j(i-1) + \beta_j, \qquad j = 0, ..., q-1, \quad i = 1, ..., m.$$

The sequence $\boldsymbol{u}^{\boldsymbol{\omega}}$ is piecewise linear. It is linear with slope α_j on the set $\{jm+1, ..., (j+1)m\}$ for any $j = 0, ..., q-1$. Thus, $q(\boldsymbol{u}^{\boldsymbol{\omega}}) = q$. Next, we prove that $\boldsymbol{u}^{\boldsymbol{\omega}} \in \mathcal{S}^{\mathrm{C}}$ for all $\boldsymbol{\omega} \in \Omega$. It is enough to check the convexity condition at the endpoints of the linear pieces:

$$2u^{\boldsymbol{\omega}}_{jm} \leq u^{\boldsymbol{\omega}}_{jm-1} + u^{\boldsymbol{\omega}}_{jm+1}, \qquad 2u^{\boldsymbol{\omega}}_{jm+1} \leq u^{\boldsymbol{\omega}}_{jm} + u^{\boldsymbol{\omega}}_{jm+2}, \tag{44}$$

for all $j = 1, ..., q-1$. Using (43) we get that, for all $j = 1, ..., q-1$,

$$\begin{aligned} u^{\boldsymbol{\omega}}_{jm+1} - u^{\boldsymbol{\omega}}_{jm} &= \omega_{j+1}\gamma + \beta_j - (\omega_j\gamma + \alpha_{j-1}(m-1) + \beta_{j-1}), \\ &= (\omega_{j+1} - \omega_j + 1)\gamma + \alpha_{j-1}, \\ &= (\omega_{j+1} - \omega_j - 1)\gamma + \alpha_j. \end{aligned}$$

Hence, $\alpha_{j-1} \leq u^{\boldsymbol{\omega}}_{jm+1} - u^{\boldsymbol{\omega}}_{jm} \leq \alpha_j$. Since also $\alpha_{j-1} = u^{\boldsymbol{\omega}}_{jm} - u^{\boldsymbol{\omega}}_{jm-1}$ and $\alpha_j = u^{\boldsymbol{\omega}}_{jm+2} - u^{\boldsymbol{\omega}}_{jm+1}$, it follows that the two inequalities (44) hold, for all $j = 1, ..., q-1$. Thus, $\boldsymbol{u}^{\boldsymbol{\omega}} \in \mathcal{S}^{\mathrm{C}}$. In summary, we have proved that $\boldsymbol{u}^{\boldsymbol{\omega}} \in \mathcal{S}^{\mathrm{C}}_q$ for all $\boldsymbol{\omega} \in \Omega$.

Now, from the Varshamov-Gilbert bound, cf. (32), for $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$ we have

$$\|\boldsymbol{u}^{\omega} - \boldsymbol{u}^{\omega'}\|^2 = \frac{\gamma^2}{q} d_H(\boldsymbol{\omega}, \boldsymbol{\omega}') \geq \frac{\gamma^2}{8} = \frac{\sigma^2 q}{512n}, \tag{45}$$

where $d_H(\cdot,\cdot)$ is the Hamming distance. Finally, similarly to (34), the Kullback-Leibler divergence between $P_{\boldsymbol{\omega}}$ and $P_{\mathbf{0}}$ satisfies $K(P_{\boldsymbol{\omega}}, P_{\mathbf{0}}) \leq \frac{\log(|\Omega|-1)}{16}$. Applying (Tsybakov, 2009, Theorem 2.7) with $\alpha = 1/16$ completes the proof. ■

4. Concluding Remarks and Discussion

In this short note, we have shown that the estimators $\hat{\boldsymbol{\mu}}^{Q}$ and $\hat{\boldsymbol{\mu}}^{Q-conv}$ based on sparsity pattern aggregation (in its Q-aggregation version) achieve oracle inequalities that improve on some previous results for isotonic and convex regression.

One of the improvements is that oracle inequalities (17) and (39) are sharp, i.e., with leading constant 1 and they are valid for all $\boldsymbol{\mu} \in \mathbb{R}^n$. It allows us to obtain bounds for the minimax regret under arbitrary model misspecification, which was not possible based on the previous results. We show that these bounds are rate optimal up to logarithmic factors.

The question on whether the least squares estimators under monotonicity and convexity constraints can achieve sharp oracle inequalities with correct rates remains open.

Another improvement is that we obtain oracle inequalities both with high probability and in expectation, which was not the case in the previous work.

An advantage of the least squares estimator is that it requires no tuning parameters. In particular, the knowledge of σ^2 is not needed to construct the estimators $\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}^{\uparrow})$ and $\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}^{\mathrm{C}})$. This is in contrast to the ℓ_1 penalized estimator (7) and the estimators $\hat{\boldsymbol{\mu}}^{Q}$ and $\hat{\boldsymbol{\mu}}^{Q-conv}$; their construction requires the knowledge of σ^2. For the ℓ_1 penalized estimator (7), the issue may be addressed by using a scale-free version of the Lasso Belloni et al. (2014); Sun and Zhang (2012). For the Q-aggregation estimators $\hat{\boldsymbol{\mu}}^{Q}$ and $\hat{\boldsymbol{\mu}}^{Q-conv}$, we can treat the issue of unknown σ as in Bellec (2014). Namely, it is shown in Bellec (2014) that the oracle inequalities for Q-aggregation procedures are essentially preserved after plugging in an estimator $\hat{\sigma}^2$ of σ^2 that satisfies $|\hat{\sigma}^2/\sigma^2 - 1| \leq 1/8$ with high probability, which is even weaker than consistency.

Finally, note that instead of Q-aggregation we could have used sparsity pattern aggregation by the Exponential Screening procedure of Rigollet and Tsybakov (2011). This would lead to sharp oracle inequalities in expectation of the form (17) and (39) but not to inequalities with high probability such as (16) and (38). This is the reason why we have opted for Q-aggregation rather than for Exponential Screening in this paper. On the other hand, Exponential Screening estimators are computationally more attractive than Q-aggregation since they can be successfully approximated by MCMC algorithms (see Rigollet and Tsybakov (2011, 2012) for details).

Acknowledgement. This work was supported by GENES and by the French National Research Agency (ANR) under the grants IPANEMA (ANR-13-BSH1-0004-02), and Labex ECODEC (ANR - 11-LABEX-0047). It was also supported by the "Chaire Economie et Gestion des Nouvelles Données", under the auspices of Institut Louis Bachelier, Havas-Media and Paris-Dauphine.

References

Fadoua Balabdaoui and Jon A. Wellner. Estimation of a k-monotone density: Limit distribution theory and the spline connection. *Annals of Statistics*, 35:2536–2564, 2007.

Moulinath Banerjee and Jon A. Wellner. Likelihood ratio tests for monotone functions. *Annals of Statistics*, 29:1699–1731, 2001.

Pierre C. Bellec. Optimal bounds for aggregation of affine estimators. *arXiv:1410.0346*, 2014.

A. Belloni, V. Chernozhukov, and L. Wang. Pivotal estimation via square-root lasso in nonparametric regression. *Annals of Statistics*, 42:757–788, 2014.

S. Chatterjee, A. Guntuboyina, and B. Sen. On risk bounds in isotonic and other shape restricted regression problems. *Annals of Statistics*, 43:1774–1800, 2015.

D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy Q-aggregation. *Annals of Statistics*, 40:1878–1905, 2012.

D. Dai, P. Rigollet, Xia L., and Zhang T. Aggregation of affine estimators. *Electronic J. Statist.*, 8:302–327, 2014.

A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the lasso. *arXiv:1402.1700*, 2014.

A. Guntuboyina and B. Sen. Global risk bounds and adaptation in univariate convex regression. *arXiv:1305.1648*, 2013. To appear in Probability Theory and Related Fields.

D. Hsu, S. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6, 2012.

E. Mammen and S. van de Geer. Locally adaptive regression splines. *Annals of Statistics*, 25:387–413, 1997.

M. Meyer and M. Woodroofe. On the degrees of freedom in shape-restricted regression. *Annals of Statistics*, 28:1083–1104, 2000.

A.M. Nemirovski, B.T. Polyak, and Tsybakov A.B. Rate of convergence of nonparametric estimators of maximum-likelihood type. *Problems of Information Transmission*, 21: 258–272, 1985.

A. Rakhlin, K. Sridharan, and A.B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *arXiv:1308.1147*, 2013. To appear in Bernoulli.

P. Rigollet. Kullback–Leibler aggregation and misspecified generalized linear models. *Annals of Statistics*, 40:639–665, 2012.

P. Rigollet and A.B. Tsybakov. Exponential screening and optimal rates of sparse estimation. *Annals of Statistics*, 39:731–771, 2011.

P. Rigollet and A.B. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 27:558–575, 2012.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99:879–898, 2012.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

A.B. Tsybakov. Aggregation and minimax optimality in high dimensional estimation. *Proceedings of International Congress of Mathematicians (Seoul, 2014)*, 3:225–246, 2014.

V. N. Vapnik and A. Y. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

C.-H. Zhang. Risk bounds in isotonic regression. *Annals of Statistics*, 30:528–555, 2002.