Aggregating estimates by convex optimization

Anatoli Juditsky, Arkadi Nemirovski

Aggregating estimates by convex optimization

Anatoli Juditsky * Arkadi Nemirovski †‡

Abstract

We discuss the approach to estimate aggregation and adaptive estimation based upon (nearly optimal) testing of convex hypotheses. We show that in the situation where the observations stem from *simple observation schemes* [27] and where set of unknown signals is a finite union of convex and compact sets, the proposed approach leads to aggregation and adaptation routines with nearly optimal performance. As an illustration, we consider application of the proposed estimates to the problem of recovery of unknown signal known to belong to a union of ellitopes [25, 27] in Gaussian observation scheme. The proposed approach can be implemented efficiently when the number of sets in the union is "not very large." We conclude the paper with a small simulation study illustrating practical performance of the proposed procedures in the problem of signal estimation in the single-index model.

1 Introduction

We address the problem of data-driven selection of estimators from a given collection. A simplified version of the problem considered in this paper is as follows.

Problem I. We are given in advance N nonempty convex compact signal sets $X_j \subset \mathbf{R}^n$ and $m \times n$ sensing matrices A_j, $1 \le j \le N$. Given access to M independent observations

$$\omega^M = (\omega_1, ..., \omega_M) : \omega_k = Ax + \sigma\xi_k,\ 1 \le k \le M, \quad \xi_k \sim \mathcal{N}(0, I_m) \tag{1}$$

we want to recover the signal $x \in \mathbf{R}^n$ in the situation when it is known *a priori* that $x \in X_j$ and $A = A_j$ for some (unknown!) $j \le N$. Given reliability tolerance $\epsilon \in (0, 1)$, we quantify the performance of a candidate estimate $\widehat{x}(\omega^M) : \mathbf{R}^{mM} \to \mathbf{R}^n$ by its worst-case ϵ-risk—the radius of the smallest ball, in a given seminorm[1] $\|\cdot\|$, around $\widehat{x}$ which contains the signal x underlying observations with probability at least $1 - \epsilon$, that is, by the quantity

$$\mathrm{Risk}_{\epsilon,M}^{\overline{1,N}}[\widehat{x}|X] := \min\left\{\rho : \mathrm{Prob}_{\omega^M \sim p_j^M}\{\|\widehat{x}(\omega^K) - x\| > \rho\} \le \epsilon\ \forall (j \le N, x \in X_j)\right\}$$

*LJK, Université Grenoble Alpes, 700 Avenue Centrale, 38401 Domaine Universitaire de Saint-Martin-d'Hères, France, anatoli.juditsky@univ-grenoble-alpes.fr

†ISyE, Georgia Institute of Technology, Atlanta, Georgia 30332, USA, arkadi.nemirovski@isye.gatech.edu.

‡Research of the authors was supported by Multidisciplinary Institute in Artificial intelligence MIAI @ Grenoble Alpes (ANR-19-P3IA-0003).

[1]Recall that a seminorm on $\mathbf{R}^n$ satisfies exactly the same requirements as a norm, with positivity outside the origin replaced with nonnegativity. A standard example of a seminorm is $\|x\| = \pi(Bx)$ where $\pi(\cdot)$ is a norm on some $\mathbf{R}^m$ and $B \in \mathbf{R}^{m\times n}$ has a nontrivial kernel.

where p_j is the normal distribution $\mathcal{N}(A_j x, \sigma^2 I_m)$ and $p_j^M = \overbrace{p_j \times ... \times p_j}^{M}$. We intend to estimate the signal by aggregating N selected in advance "preliminary" estimates $\widetilde{x}_j$, $j = 1, ..., N$, jth of them associated with jth observation model in which x is known to belong to X_j and $A = A_j$. Specifically, we split M available observations into "pilot sample" $\omega_1, ..., \omega_{\overline{K}}$ used to build points $x_j = \widetilde{x}_j(\omega^{\overline{K}})$, and use the remaining $K = M - \overline{K}$ observations to "assemble" x_j into the resulting estimate $\widehat{x}$ of the signal.

A related problem is that of constructing an estimate which is *adaptive*—such that its risk is "as close as possible" to the maximal risk of the jth estimate under the jth observation model, $1 \leq j \leq N$.

In this work, our focus is on the aggregation step, thus, for the most of the exposition below, estimates $x_j = \widetilde{x}_j$, $j = 1, ..., N$, are regarded as known fixed points in $\mathbf{R}^n$. The above problem is closely related to another fundamental statistical problem, that of aggregation and, in particular, to "model selection" version of the problem in which the objective is to select the "nearly $\|\cdot\|$-closest" to x point among given points $x_1, ..., x_N$. Both problems have received a lot of attention in the statistical literature. The adaptive estimation problem, in its general form which is relevant for us, has been stated in O. Lepski's pioneering works [33, 34, 35, 36] (for the setting in which $\{X_j\}$ is an injected family of sets), then substantially generalized in [20, 19, 21, 31], giving rise to the celebrated Lepski's and Goldenshluger-Lepski's adaptation schemes put to use in various contexts and by various authors. A remarkable progress has also been achieved when solving the aggregation problem, in particular, in the context of L_2-estimation in the white noise model where exact oracle inequalities were derived for collections of arbitrary estimators. Specifically, the notion of optimal rates of aggregation has been introduced in [40], and aggregation procedures attaining the risk which approaches the risk of the best point among x_i with the smallest possible, in the minimax sense, remainder term have been introduced (see also [1, 42, 6, 39, 29, 14] and references therein). Aggregation of estimators with respect to other loss functions has also been studied extensively. The problem of aggregating estimates with the Kullback–Leibler divergence as a loss function has been studied in [11, 41] in the problem of density estimation and in [38] for generalized linear models. Aggregation w.r.t. L_1-risk in the context of density estimation has been studied in [43, 15, 16, 37]; that approach has been extended to the regression setup in [23]. Finally, one of our principal motivations comes from [17] where a general aggregation scheme which applies to wide variety of the risk measures have been proposed. In this paper, we aim at extending adaptive estimation and estimate aggregation framework in several directions. Specifically, we propose adaptive estimation and aggregation routines for problems where indirect observations are available under general convex constraints on unknown signal.[2]

The underlying idea of the proposed routines is that of pairwise comparison of candidate estimates: to decide if estimate $\widetilde{x}_i$ is better/worse than $\widetilde{x}_j$, $i \neq j$, we replace the relation "risk of $\widetilde{x}_i$ is less than risk of $\widetilde{x}_j$" with a pair of convex hypotheses about x. To see how this reduction operates, consider the situation of Problem I with $N = 2$, where we want to choose between just two estimates $\widetilde{x}_1$ and $\widetilde{x}_2$, associated with models indexed by $j \in \{1, 2\}$, assuming that ϵ-risk of $\widetilde{x}_j$ is bounded with $\mathfrak{r}_j$ under the jth model. For the sake of definiteness, assume that $\mathfrak{r}_1 \leq \mathfrak{r}_2$, and that the realization of noise in the preliminary observation belongs to the subset of the corresponding probability space of probability

[2]We should mention here a special status of the problem of adaptive estimation of general linear functionals of unknown signal: in a separate line of research [9, 10] the minimax affine estimator was used as "working horse" to build the near-optimal estimator of a linear functional over a finite union X of convex compact sets in Gaussian observation scheme. A different general construction for nearly minimax optimal estimation of linear functionals over union of convex sets in simple observation schemes has been developed in [26].

$1-\epsilon$ such that $\|\widetilde{x}_{j_*}-x\|\le\mathfrak{r}_{j_*}$, where j_* is the index of the "true" observation model, that is, $x\in X_{j_*}$ and $A=A_{j_*}$. In this case, if $j_*=1$, we have $x\in X_1$ with $\|x-\widetilde{x}_1\|\le\mathfrak{r}_1$ with probability $1-\epsilon$ and $A=A_1$, i.e.,

$$x\in B_1:=\{x\in X_1,\ \|x-\widetilde{x}_1\|\le\mathfrak{r}_1\},$$

so that A_1x belongs to the convex and compact set $Y_1=A_1B_1$. When $j_*=2$, we have

$$x\in B_2:=\{x\in X_2,\ \|x-\widetilde{x}_2\|\le\mathfrak{r}_2\}$$

and $A_2x\in Y_2=A_2B_2$. Now, assuming that we do have $x\in B_j$ when the actual model if the jth one, $j=1,2$, given observation ω^K, consider the problem of testing "convex hypotheses"

$$H_1:\ A_1x\in Y_1\ \ \text{and}\ \ H_2:\ A_2x\in Y_2.$$

As it is well known (see, e.g., [13, 7, 8]), when Y_1 and Y_2 do not intersect, the optimal test (that with the smallest maximal risk) deciding on H_1 against H_2 in Gaussian o.s. is the likelihood ratio test of simple hypotheses

$$\overline{H}_1:\ A_1x=\overline{y}_1\ \ \text{and}\ \ \overline{H}_2:\ A_2x=\overline{y}_2$$

where

$$(\overline{y}_1,\overline{y}_2)\in\operatorname*{Argmin}_{y_1\in Y_1,y_2\in Y_2}\|y_1-y_2\|_2.$$

Thus, assuming that two hypotheses can be separated with maximal risk $\le\epsilon$, when the first model is true, $x\in B_1$ and H_1 holds, the test will accept it (and reject H_2) with probability $1-\epsilon$, implying that the 2ϵ-risk of the estimate $\widehat{x}(\omega^K)=\widetilde{x}_1$ is bounded with $\mathfrak{r}_1$, and "symmetric" bound holds when the second model is true. On the other hand, in the case the hypotheses cannot be separated $(1-\epsilon)$-reliably, selecting $\widehat{x}=\widetilde{x}_1$ results in the ϵ-risk of $\widehat{x}$ bounded with $\mathfrak{r}_1$ when the first model is true, and with $\mathfrak{r}_1+2\mathfrak{r}_2+2\mathfrak{r}_{12}$ where

$$\mathfrak{r}_{12}=\min\left\{\tfrac{1}{2}\|x_1-x_2\|:x_1\in B_1,x_2\in B_2\right\}$$

in the case of the true second model. A simple calculation shows (cf. e.g., Theorem 5 in Section 6) that in the latter case the quantity $\mathfrak{r}_{12}$ is upper-bounded by the maximal risk of estimation over $X=X_1\cup X_2$. Note that if "separation" $\mathfrak{r}_{12}$ is majorated by $\mathfrak{r}_2$, estimate $\widehat{x}$ is adaptive in the sense of [33]—when $x\in X_1$ the ϵ-risk of $\widehat{x}$ is bounded with $\mathfrak{r}_1$, and when $x\in X_2$ its risk is bounded with $\mathfrak{r}_1+2\mathfrak{r}_2+2\mathfrak{r}_{12}$ which is the same as $\mathfrak{r}_2$, up to a moderate absolute factor. On the other hand, if $\mathfrak{r}_{12}\gg\mathfrak{r}_2$, the corresponding bound is the best one can achieve under the circumstances. More generally, reducing the problem of risk minimization to that of pairwise testing of convex hypotheses makes the problem amenable to the machinery of nearly-optimal testing of convex hypotheses developed in [18].

The proposed approach shares its motivation with another construction of estimates based on testing multiple hypotheses—the T-estimators developed in [2, 3, 4]. When applied to Problem I, the latter approach amounts to building a net of points $\{x_\tau\}$, $\tau\in\mathcal{T}$, in X and selecting the estimate by applying pairwise tests to small Euclidean balls around images of x_τ, $x_{\tau'}$ in the observation space. Note that, typically, T-estimators cannot be obtained in a computationally efficient fashion and are usually considered as a theoretical tool to explore the properties of statistical problems. Despite obvious similarities with T-estimates (e.g., great flexibility shared by the both approaches), adaptive and aggregation estimates we discuss in this paper are of a different nature. Our approach can be seen as an "operational counterpart" to that of [2] leading to adaptive estimates which are efficiently

computable provided the data of the problem—sets X_j and norm $\|\cdot\|$—are computationally tractable, and N is moderate (hundreds, perhaps, thousands). As the price to pay for generality of the proposed constructions, our estimates and their risks (provably near-optimal, as we shall see, under natural assumptions) are given by efficient computation rather than in a closed analytic form.[3] This is hardly a problem in application where efficient computation usually is not inferior to a formula.

What is ahead. In what follows we discuss two adaptive estimates: a "generic" selection procedure in the situation where $\|\cdot\|$ is an arbitrary seminorm, and a special aggregation routine for the problem setting in which $\|\cdot\|$ is a Euclidean seminorm. Our principal contribution (cf., e.g., Theorem 1 and Corollary 1 of Section 3.2 in the case of general seminorm), as applied to Problem I above, may be summarized as follows.

Let a real $\theta \geq 1$, an integer $\overline{K} \geq 1$ and $\epsilon \in (0, 1/2)$ be fixed. Assume that we are given preliminary $\overline{K}$-observation estimates $\widetilde{x}_j(\cdot)$ along with reals $\mathfrak{r}_j$ such that

$$\mathrm{Risk}^{\{j\}}_{\epsilon,\overline{K}}[\widetilde{x}_j|X_j] \leq \mathfrak{r}_j \leq \theta \mathrm{RiskOpt}^{\{j\}}_{\epsilon,\overline{K}}[X_j], \quad j = 1, ..., N,$$

where for a nonempty $\mathcal{J} \subset \overline{1,N}$ and a K-observation estimate $\widehat{x}(\cdot)$,

$$\mathrm{Risk}^{\mathcal{J}}_{\epsilon,K}[\widehat{x}|\cup_{j\in\mathcal{J}} X_j] = \min\left\{\rho : \mathrm{Prob}_{\omega^K\sim p_j^K}\left\{\|\widehat{x}(\omega^K) - x\| > \rho\right\} \leq \epsilon \ \forall (j \in \mathcal{J}, x \in X_j)\right\}$$

is the risk of estimate "on the union of models with indexes from $\mathcal{J}$," and

$$\mathrm{RiskOpt}^{\mathcal{J}}_{\epsilon,K}[\cup_{j\in\mathcal{J}} X_j] = \inf_{\widehat{x}} \mathrm{Risk}^{\mathcal{J}}_{\epsilon,K}[\widehat{x}|\cup_{j\in\mathcal{J}} X_j]$$

is the corresponding minimax risk.

Now, suppose that given $M \geq O(1)\frac{\ln(N/\epsilon)}{\ln(1/\epsilon)}\overline{K}$ independent observations (1), we utilize the first $\overline{K}$ observations to build "preliminary" estimates $x_j = \widetilde{x}_j(\omega^{\overline{K}})$, $j = 1, ..., N$, and then proceed with selection procedure of Section 3.2 using $K = M - \overline{K}$ remaining observations to aggregate points x_j into an adaptive estimate $\widehat{x}^{(a)}(\omega^M)$. Then $\widehat{x}^{(a)}(\omega^M)$ satisfies

$$\mathrm{Risk}^{\overline{1,N}}_{2\epsilon,M}[\widehat{x}^{(a)}|X] \leq O(1)\theta \mathrm{RiskOpt}^{\overline{1,N}}_{\epsilon,\overline{K}}[X]$$

In other words, modulo logarithmic in N increase in observation count and reliability parameter ϵ of the risk replaced with 2ϵ, estimate $\widehat{x}^{(a)}(\omega^M)$ is minimax optimal on X within factor $O(1)\theta$.

Furthermore, the "overall" minimax risk $\mathrm{RiskOpt}^{\overline{1,N}}$ is nearly upper-bounded by the maximum of pairwise minimax risks $\mathrm{RiskOpt}^{\{i,j\}}$. Specifically, with M as above,

$$\mathrm{RiskOpt}^{\overline{1,N}}_{2\epsilon,M}[\cup_{j\leq N} X_j] \leq O(1) \max_{i\neq j} \mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j].$$

Finally, suppose that N models in Problem I are ordered, so that bounds $\mathfrak{r}_i$ for partial risks of estimates $\widetilde{x}_j(\omega^{\overline{K}})$ satisfy

$$\mathfrak{r}_1 \leq \mathfrak{r}_2 \leq ... \leq \mathfrak{r}_N,$$

[3] We believe that in our setting, allowing for arbitrary sensing matrices and general convex parameter sets, closed form results are just impossible.

and that minimax risks of estimation over "pairwise unions" $\mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j]$, $1 \le i, j \le N$, are dominated by the pairwise maxima of the corresponding partial risks, i.e.,

$$\max_{1\le j\le i} \mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j] \le O(1)\mathrm{RiskOpt}^{\{i\}}_{\epsilon,\overline{K}}[X_i], \quad i = 1, ..., N.$$

Then we also have

$$\mathrm{Risk}^{\{i\}}_{2\epsilon,M}[\widehat{x}^{(a)}|X_i] \le O(1)\theta\mathrm{RiskOpt}^{\{i\}}_{\epsilon,\overline{K}}[X_i], \quad \forall i = 1, ..., N,$$

i.e., estimate $\widehat{x}^{(a)}$ is (again, up to logarithmic in N increase in observation count and reliability parameter ϵ of the risk replaced with 2ϵ) minimax adaptive, within factor $O(1)\theta$, in the sense of [33, 34] over considered family of observation models.

Our results are not restricted to the Gaussian observation scheme (1) and deal with *simple observation schemes*[4] (o.s.'s), as defined in [18, 27]. Aside of Gaussian o.s., important examples of simple o.s. are

- *Poisson o.s.*, where ω_k are independent across k identically distributed vectors with independent across $i \le m$ entries $[\omega_k]_i \sim \mathrm{Poisson}(a_i^T x)$, and
- *Discrete o.s.*, where ω_k are independent across k realizations of discrete random variable taking values $1, ..., m$ with probabilities affinely parameterized by x.

The presentation is organized in two parts. In the first part, we consider the problem of adaptive and minimax estimation over the sets which are unions of convex sets—a generalization of the setting of Problem I to the case of simple o.s.. We start with stating the general estimation problem and provide an "operational summary" of results on testing in simple observation schemes in Section 2. Section 3.2 deals with adaptation in the case of a general seminorm $\|\cdot\|$, and Section 3.3 with the special case of Euclidean seminorm. The second part of the paper deals with the problem of model selection aggregation. Although closely related to the problem of adaptive estimation, this problem calls for different notion of optimality with respect to which estimation routines discussed in Section 3 may be heavily suboptimal. The second part begins with a description of two "abstract" aggregation routines utilizing pairwise tests in Section 4, which we specify for aggregation in simple o.s. in Section 5. We consider next the application of these routines to signal recovery in the situation described in Section 3.1. We conclude the paper (Section 6) detailing how the proposed approach can be used to build nearly minimax estimates in the problem of signal recovery in Gaussian o.s. when the signal set X is a union of ellitopes (cf. [25]); these results are accompanied by a small simulation study illustrating numerical performance of the proposed estimates in that problem.

Proofs of the results are postponed till the appendix.

2 Preliminaries: testing convex hypotheses in simple observation schemes

2.1 Simple observation schemes: definitions

All developments to follow make use of the notion of a simple observation scheme, see [27]. To make the presentation self-contained we start with explaining this notion here.

[4] Our results can be easily extended to the more general case of *simple families*—families of distributions specified in terms of upper bounds on their moment-generating functions, see [27] for details. Restricting the framework to the case of simple observation schemes is aimed at streamlining the presentation.

Formally, a *simple observation scheme* (o.s.) is a collection $\mathcal{SO} = ((\Omega, \Pi), \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F})$, where

- (Ω, Π) is an *observation space*: Ω is a Polish (complete metric separable) space, and Π is a σ-finite σ-additive Borel reference measure on Ω, such that Ω is the support of Π;
- $\{p_\mu(\cdot) : \mu \in \mathcal{M}\}$ is a parametric family of probability densities, specifically, $\mathcal{M}$ is a convex relatively open set in some $\mathbf{R}^M$, and for $\mu \in \mathcal{M}$, $p_\mu(\cdot)$ is a probability density, taken w.r.t. Π, on Ω. We assume that the function $p_\nu(\omega)$ is positive and continuous in $(\mu, \omega) \in \mathcal{M} \times \Omega$;
- $\mathcal{F}$ is a finite-dimensional linear subspace in the space of continuous functions on Ω. We assume that $\mathcal{F}$ contains constants and all functions of the form $\ln(p_\mu(\cdot)/p_\nu(\cdot))$, $\mu, \nu \in \mathcal{M}$, and that the function
$$\Phi_{\mathcal{SO}}(\phi; \mu) = \ln\left(\int_\Omega \mathrm{e}^{\phi(\omega)} p_\mu(\omega)\Pi(d\omega)\right) \tag{2}$$
is real-valued on $\mathcal{F} \times \mathcal{M}$ and is *concave* in $\mu \in \mathcal{M}$; note that this function is automatically convex in $\phi \in \mathcal{F}$. From real-valuedness, convexity-concavity and the fact that both $\mathcal{F}$ and $\mathcal{M}$ are convex and relatively open, it follows that Φ is continuous on $\mathcal{F} \times \mathcal{M}$.

2.1.1 Examples of simple observation schemes

As shown in [27] (and can be immediately verified), the following o.s.'s are simple:

1. *Gaussian o.s.*, where Π is the Lebesgue measure on $\Omega = \mathbf{R}^d$, $\mathcal{M} = \mathbf{R}^d$, $p_\mu(\omega)$ is the density of the Gaussian distribution $\mathcal{N}(\mu, I_d)$ (mean μ, unit covariance), and $\mathcal{F}$ is the family of affine functions on $\mathbf{R}^d$. Gaussian o.s. with μ linearly parameterized by signal x underlying observations is the standard observation model in signal processing;
2. *Poisson o.s.*, where Π is the counting measure on the nonnegative integer d-dimensional lattice $\Omega = \mathbf{Z}^d_+$, $\mathcal{M} = \mathbf{R}^d_{++} = \{\mu = [\mu_1; ...; \mu_d] > 0\}$, p_μ is the density, taken w.r.t. Π, of random d-dimensional vector with independent Poisson(μ_i) entries, $i = 1, .., d$, and $\mathcal{F}$ is the family of all affine functions on Ω. Poisson o.s. with μ affinely parameterized by signal x underlying observation is the standard observation model in *Poisson imaging*;
3. *Discrete o.s.*, where Π is the counting measure on the finite set $\Omega = \{1, 2, ..., d\}$, $\mathcal{M}$ is the set of positive d-dimensional probabilistic vectors $\mu = [\mu_1; ...; \mu_d]$, $p_\mu(\omega) = \mu_\omega$, $\omega \in \Omega$, is the density, taken w.r.t. Π, of a probability distribution μ on Ω, and $\mathcal{F} = \mathbf{R}^d$ is the space of all real-valued functions on Ω;
4. *Direct product of simple o.s.'s.* Given K simple o.s.'s $\mathcal{SO}_t = ((\Omega_t, \Pi_t), \{p_{t,\mu} : \mu \in \mathcal{M}_r\}, \mathcal{F}_t)$, $t = 1, ..., K$, we can build from them a new (direct product) o.s. $\mathcal{SO}_1 \times .... \times \mathcal{SO}_K$ with observation space $\Omega_1 \times ... \times \Omega_K$, reference measure $\Pi_1 \times ... \times \Pi_K$, family of probability densities $\{p_\mu(\omega_1, ..., \omega_K) = \prod_{t=1}^K p_{t,\mu_t}(\omega_t) : \mu = [\mu_1; ...; \mu_K] \in \mathcal{M}_1 \times ... \times \mathcal{M}_K\}$, and $\mathcal{F} = \{\phi(\omega_1, ..., \omega_K) = \sum_{t=1}^K \phi_t(\omega_t) : \phi_t \in \mathcal{F}_t, t \le K\}$. In other words, the direct product of o.s.'s $\mathcal{SO}_t$ is the observation scheme in which we observe collections $\omega^K = (\omega_1, ..., \omega_K)$ with independent across t components ω_t yielded by o.s.'s $\mathcal{SO}_t$.
When all factors $\mathcal{SO}_t$, $t = 1, ..., K$, are identical to each other, we can reduce the direct product $\mathcal{SO}_1 \times ... \times \mathcal{SO}_K$ to its "diagonal," referred to as *Kth power* $\mathcal{SO}^K$, or *stationary K-repeated version*, of $\mathcal{SO} = \mathcal{SO}_1 = ... = \mathcal{SO}_K$. Just as in the direct product case, the observation space and reference measure in $\mathcal{SO}^K$ are $\Omega^K = \underbrace{\Omega \times ... \times \Omega}_{K}$ and $\Pi^K = \underbrace{\Pi \times ... \times \Pi}_{K}$, the family of densities is $\{p^K_\mu(\omega^K) = \prod_{t=1}^K p_\mu(\omega_t) : \mu \in \mathcal{M}\}$, and the family $\mathcal{F}$ is $\{\phi^{(K)}(\omega_1, ..., \omega_K) = \sum_{t=1}^K \phi(\omega_t) : \phi \in \mathcal{F}\}$.

Informally, $\mathcal{SO}^K$ is the observation scheme we arrive at when passing from a single observation drawn from a distribution p_μ, $\mu \in \mathcal{M}$, to K independent observations drawn from the same distribution p_μ.
It is immediately seen that direct product of simple o.s.'s, same as power of simple o.s., are themselves simple o.s.

2.2 Testing pairs of convex hypotheses in simple o.s.

What follows is a summary of results of [27] which are relevant to our current needs.
Assume that $\omega^K = (\omega_1, ..., \omega_K)$ is a stationary K-repeated observation in a simple o.s. $\mathcal{SO} = ((\Omega, \Pi), \{p_\mu : \mu \in \mathcal{M}\}, \mathcal{F})$, so that $\omega_1, ..., \omega_K$ are, independently of each other, drawn from a distribution p_μ with some $\mu \in \mathcal{M}$. Given ω^K we want to decide on the hypotheses H_1 and H_2, with H_χ, $\chi = 1, 2$, stating that $\omega_t \sim p_\mu$ for some $\mu \in M_\chi$, where M_χ is a nonempty convex compact subset of $\mathcal{M}$. In the sequel, we refer to hypotheses of this type, parameterized by nonempty convex compact subsets of $\mathcal{M}$, as to *convex* hypotheses in the simple o.s. in question.

The principal "building block" of our subsequent constructions is a *simple* test[5] $\mathcal{T}^K$ for this problem which is as follows:

- Given convex compact sets M_χ, $\chi = 1, 2$, we solve the optimization problem

$$\mathrm{Opt} = \max_{\mu \in M_1, \nu \in M_2} \ln\Big(\underbrace{\int_\Omega \sqrt{p_\mu(\omega) p_\nu(\omega)}\Pi(d\omega)}_{=:\varrho(\mu,\nu)}\Big) \tag{3}$$

  It is shown in [18] that in the case of simple o.s., problem (3) is a convex problem (convexity meaning that the objective to be maximized is a concave continuous function of μ, ν) and an optimal solution exists.

  Note that for basic simple o.s.'s problem (3) reads

$$\mathrm{Opt} = \max_{\mu \in M_1, \nu \in M_2} \begin{cases} -\frac{1}{8}\|\mu - \nu\|_2^2, & \text{Gaussian o.s.} \\ -\frac{1}{2}\sum_{i=1}^d [\sqrt{\mu_i} - \sqrt{\nu_i}]^2, & \text{Poisson o.s.} \\ \ln\left(\sum_{i=1}^d \sqrt{\mu_i \nu_i}\right), & \text{Discrete o.s.} \end{cases} \tag{4}$$

- An optimal solution μ_*, ν_* to (3) induces *detectors*

$$\begin{array}{rcl} \phi_*(\omega) & = & \frac{1}{2}\ln(p_{\mu_*}(\omega)/p_{\nu_*}(\omega)) : \ \Omega \to \mathbf{R}, \\ \phi_*^{(K)}(\omega^K) & = & \sum_{t=1}^K \phi_*(\omega_t) : \ \Omega \times ... \times \Omega \to \mathbf{R} \end{array} \tag{5}$$

  Given a stationary K-repeated observation ω^K, the test $\mathcal{T}^K$ accepts hypothesis H_1 and rejects hypothesis H_2 whenever $\phi_*^{(K)}(\omega^K) \geq 0$, otherwise the test rejects H_1 and accepts H_2. The *risk* of $\mathcal{T}^K$ – the maximal probability to reject a hypothesis when it is true – does not exceed $\epsilon_\star^K$, where

$$\epsilon_\star = \exp(\mathrm{Opt}).$$

  In other words, whenever observation ω^K stems from a distribution p_μ with $\mu \in M_1 \cup M_2$,

[5] A test deciding on a pair of hypotheses is called simple, if given an observation, it always accepts exactly one of the hypotheses and rejects the other one.

– the p_μ-probability to reject H_1 when the hypothesis is true (i.e., when $\mu \in M_1$) is at most $\epsilon_\star^K$, and
– the p_μ-probability to reject H_2 when the hypothesis is true (i.e., when $\mu \in M_2$) is at most $\epsilon_\star^K$.

The test $\mathcal{T}^K$ possesses the following optimality properties:

A. *The associated detector $\phi_*^{(K)}$ and the risk $\epsilon_\star^K$ form an optimal solution and the optimal value in the optimization problem*

$$\min_{\phi} \max\left[\max_{\mu\in M_1} \int_{\Omega^K} \epsilon^{-\phi(\omega^K)} p_\mu^K(\omega^K)\Pi^K(d\omega^K), \max_{\nu\in M_2} \int_{\Omega^K} \epsilon^{\phi(\omega^K)} p_\nu^K(\omega^K)\Pi^K(d\omega^K)\right],$$
$$\left[\Omega^K = \underbrace{\Omega\times ...\times \Omega}_{K},\ \ p_\mu^K(\omega^K) = \textstyle\prod_{t=1}^K p_\mu(\omega_t),\ \right]$$

where the minimum is taken w.r.t. all Borel functions $\phi(\cdot): \Omega^K \to \mathbf{R}$;

B. *Let $\epsilon \in (0, 1/2)$, and suppose that there exists a test which, using a stationary $\overline{K}$-repeated observation, decides on the hypotheses H_1, H_2 with risk $\le \epsilon$. Then*

$$\epsilon_\star \le [2\sqrt{\epsilon(1-\epsilon)}]^{1/\overline{K}} \tag{6}$$

and the test $\mathcal{T}^K$ with[6]

$$K = \left\lceil \frac{2\ln(1/\epsilon)}{\ln\left([4\epsilon(1-\epsilon)]^{-1}\right)}\overline{K}\right\rceil$$

decides on the hypotheses H_1, H_2 with risk $\le \epsilon$ as well. Note that $K = 2(1+o(1))\overline{K}$ as $\epsilon \to +0$.[7]

In what follows we augment the test $\mathcal{T}^K$ to address the situation where one or both hypotheses are empty. When one of the hypotheses is empty, $\mathcal{T}^K$, by convention, accepts the nonempty hypothesis. When both hypotheses are empty, $\mathcal{T}^K$ accepts, say, the first of them. Because the true hypothesis cannot be empty, the risk of $\mathcal{T}^K$ vanishes in this case.

2.3 Testing multiple hypotheses in simple o.s.

As shown in [18], near-optimal pairwise tests deciding on pairs of convex hypotheses in simple o.s.'s outlined in Section 2.2 can be used as building blocks when constructing near-optimal tests deciding on multiple convex hypotheses. In the sequel, we use one of these constructions, namely, as follows.

Assume that we are given a simple o.s. $\mathcal{SO} = ((\Omega, P), \{p_\mu : \mu \in \mathcal{M}\}, \mathcal{F})$ and two finite collections of nonempty convex compact subsets $B_1, ..., B_b$ ("blue sets") and $R_1, ..., R_r$ ("red sets") of $\mathcal{M}$. Our objective is, given a stationary K-repeated observation ω^K stemming from a distribution p_μ, $\mu \in \mathcal{M}$, to infer the color of μ, that is, to decide on the hypothesis $H_B : \mu \in B := B_1 \cup ... \cup B_b$ vs. the alternative $H_R : \mu \in R := R_1 \cup ... \cup R_r$. To this end we act as follows:

1. For every pair i, j with $i \le b$ and $j \le r$, we solve the problem (4) with B_i in the role of M_1 and R_j in the role of M_2; we denote Opt_{ij} the associated optimal values. The corresponding optimal solutions μ_{ij} and ν_{ij} give rise to the detectors

$$\phi_{ij}(\omega) = \tfrac{1}{2}\ln\left(p_{\mu_{ij}}(\omega)/p_{\nu_{ij}}(\omega)\right) : \Omega \to \mathbf{R},\ \ \phi_{ij}^{(K)}(\omega^K) = \textstyle\sum_{t=1}^K \phi_{ij}(\omega_t) : \Omega^K \to \mathbf{R} \tag{7}$$

[6] Here $\lceil a \rceil$ stands for the "upper" integer part—the smallest integer greater or equal to a.

[7] It is worth mentioning that in the Gaussian o.s. test $\mathcal{T}^K$ optimal—it is the test minimizing the maximal risk of testing of H_1 vs H_2 among all tests; the corresponding optimal risk is $\epsilon = 1 - \Phi\left(\frac{1}{2}\|\mu_* - \nu_*\|_2\sqrt{K}\right)$ where Φ is the standard normal c.d.f. and $[\mu_*; \nu_*]$ is an optimal solution to (3).

(cf. (5)) and *risks*

$$\epsilon_{ij} = \exp(\mathrm{Opt}_{ij}) = \int_\Omega \sqrt{p_{\mu_{ij}}(\omega)p_{\nu_{ij}}(\omega)}P(d\omega). \tag{8}$$

2. We build the entrywise positive $b \times r$ matrix $E^{(K)} = [\epsilon_{ij}^K]_{\substack{1\le i\le b\\ 1\le j\le r}}$ and symmetric entrywise nonnegative $(b+r)\times(b+r)$ matrix $E_K = \left[\begin{array}{c|c} & E^{(K)} \\ \hline [E^{(K)}]^T & \end{array}\right]$. Let ϵ_K be the spectral norm of the matrix $E^{(K)}$ (equivalently, spectral norm of E_K), and let $e = [g; h]$[8] be the Perron-Frobenius eigenvector of E_K, so that e is a nontrivial nonnegative vector such that $E_K e = \epsilon_K e$. Note that from entrywise positivity of $E^{(K)}$ it immediately follows that $e > 0$, so that the quantities

$$\alpha_{ij} = \ln(h_j/g_i), \quad 1 \le i \le b, \ 1 \le j \le r$$

are well defined. We set

$$\psi_{ij}^{(K)}(\omega^K) = \phi_{ij}^{(K)}(\omega^K) - \alpha_{ij} = \sum_{t=1}^K \phi_{ij}(\omega_t) - \alpha_{ij} : \Omega^K \to \mathbf{R}, \ 1 \le i \le b, 1 \le j \le r \tag{9}$$

3. Let now $\mathcal{T}^K$ be the test which given observation $\omega^K \in \Omega^K$ with ω_t, $t = 1, ..., K$, drawn, independently of each other, from a distribution p_μ, claims that μ is blue (equivalently, $\mu \in B$), if there exists $i \le b$ such that $\psi_{ij}(\omega^K) \ge 0$ for all $j = 1, ..., r$, and claims that μ is red (equivalently, $\mu \in R$) otherwise.

The main result about the just described "color inferring" test is as follows

Proposition 1. [18, Propositions 3.2] *Let the components ω_t of ω^K be drawn, independently of each other, from distribution p_μ, $\mu \in B \cup R$. Then the just defined test for every ω^K assigns μ with exactly one color, blue or red, depending on the observation. Moreover,*

- *when μ is blue (i.e., $\mu \in B$), the test makes correct inference "μ is blue" with p_μ^K-probability at least $1 - \epsilon_K$;*
- *similarly, when μ is red (i.e., $\mu \in R$), the test makes correct inference "μ is red" with p_μ^K-probability at least $1 - \epsilon_K$.*

Now, suppose that $\overline{\mathcal{T}}$ is some color inferring test with maximal risk $\le \epsilon \in (0, \frac{1}{2})$. Obviously, $\overline{\mathcal{T}}$ gives rise to a straightforward test of hypotheses $H_{B_i} : \mu \in B_i$, $i \le b$ vs $H_{R_j} : \mu \in R_j$, $j \le r$ with maximal risk bounded with ϵ. This simple observation implies the following corollary of Proposition 1 (cf. [18, Proposition 3.4]).

Proposition 2. *In the just described situation, given $\epsilon \in (0, \frac{1}{2})$, assume that in nature there exists test $\overline{\mathcal{T}}$, based on $\overline{K}$-repeated observation $\omega^{\overline{K}} \sim p_\mu^{\overline{K}}$ and deciding on blue and red hypotheses, and such that $\overline{\mathcal{T}}$ never accepts more than one hypothesis and has risk $\le \epsilon$, meaning that whenever $\mu \in B$ (whenever $\mu \in R$), H_b (resp., H_r) will be accepted with $p_\mu^{\overline{K}}$- probability $\ge 1 - \epsilon$. Then risk of detector-based test $\mathcal{T}^K$ utilizing K-repeated observation ω^K does not exceed $\varepsilon \in (0, 1)$ provided that*[9]

$$K \ge \left\lceil \frac{2\ln\left(\max[b, r]\varepsilon^{-1}\right)}{\ln\left([4\epsilon(1-\epsilon)]^{-1}\right)}\overline{K} \right\rceil.$$

[8] We use "Matlab notation" $[a; b]$ for vertical and $[a, b]$ for horizontal concatenation of matrices a, b of appropriate dimensions.

[9] The case of unique observation may be of interest when the considered o.s. is Gaussian. The corresponding near-optimality result admits the following reformulation in this case: suppose that in a Gaussian o.s. in nature there exists test $\overline{\mathcal{T}}$ deciding with risk $\le \epsilon$ on hypotheses H_B and H_R using (unique) observation $\omega \sim \mathcal{N}(\mu, \bar{\sigma}^2 I_d)$. Then detector-based

3 Adaptive estimation by testing

3.1 Estimation over unions of convex sets in simple o.s.: problem setting

Problem setup. In the sequel, we deal with the situation as follows. Given are:

1. simple o.s. $\mathcal{SO} = ((\Omega, \Pi), \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F})$,
2. a collection of $N \geq 2$ convex compact sets $X_j \subset \mathbf{R}^n$, giving rise to the set $X = \bigcup_{j=1}^N X_j$,
3. affine mappings $x \mapsto \mathcal{A}_j(x)$ such that $\mathcal{A}_j(X_j) \subset \mathcal{M}$, $j = 1, ..., N$,
4. a seminorm $\|\cdot\|$ on $\mathbf{R}^n$,
5. reliability tolerance $\epsilon \in (0, 1/2)$,

Risks. Given a nonempty subset $\mathcal{J} = \{j_1 < ... < j_s\}$ of $\{1, 2, ..., N\}$, set $Y \subset \bigcup_{j=1}^N X_j$ and $\varepsilon \in (0, 1)$, we define the ε-risk of an M-observation estimate $\widehat{x}(\omega^M) : \Omega^M \to \mathbf{R}^n$ on Y as

$$\mathrm{Risk}^{\mathcal{J}}_{\varepsilon,M}[\widehat{x}|Y] = \min\left\{\rho : \mathrm{Prob}_{\omega^M \sim p^M_{\mathcal{A}_j(x)}}\left\{\|\widehat{x}(\omega^M) - x\| > \rho\right\} \leq \varepsilon \ \forall (j \in \mathcal{J}, x \in Y \cap X_j)\right\},$$

and the associated minimax risk as

$$\mathrm{RiskOpt}^{\mathcal{J}}_{\varepsilon,M}[Y] = \inf_{\widehat{x}(\cdot)} \mathrm{Risk}^{\mathcal{J}}_{\varepsilon,M}[\widehat{x}|Y]$$

where the infimum is taken over all estimates utilizing M-repeated observation ω^M.

We assume that, in addition to the above setup, we are given

6. positive integers $\overline{K}$ and K such that $M = \overline{K} + K$ and N "preliminary" $\overline{K}$-observation estimates $\widetilde{x}_i(\cdot) : \Omega^{\overline{K}} \to \mathbf{R}^n$, along with reals $\mathfrak{r}_i = \mathfrak{r}_i^{\overline{K}}(\epsilon)$, $1 \leq i \leq N$—upper bounds for the partial ϵ-risks of $\widetilde{x}_i(\cdot)$:
$$\mathrm{Risk}^{\{i\}}_{\epsilon,\overline{K}}[\widetilde{x}_i|X_i] \leq \mathfrak{r}_i^{\overline{K}}(\epsilon), \quad 1 \leq i \leq N. \tag{10}$$

Goal and strategy. Assume that we are given M independent across k observations

$$\omega_k \sim p_{\mathcal{A}_{\ell_*}(x_*)}, \ 1 \leq k \leq M$$

(using the terminology of Section 2.1—a stationary M-repeated observation $\omega^M = (\omega_1, ..., \omega_M)$), stemming from an unknown pair (ℓ_*, x_*) with $1 \leq \ell_* \leq N$ and $x_* \in X_{\ell_*}$. Our goal is to build an estimate $\widehat{x}$ of x_* with the least possible risk. To this end we intend to use collection $\omega^{\overline{K}} = (\omega_1, ..., \omega_{\overline{K}})$ of the first $\overline{K}$ observations to compute points

$$x_i = \widetilde{x}_i(\omega^{\overline{K}}).$$

Our goal is to use the remaining—secondary—K observations $\omega^K = (\omega_{\overline{K}+1}, ..., \omega_{\overline{K}+K})$ to "aggregate" these points into an estimate $\widehat{x}$ of $x_* \in X$. We are going to achieve this goal via techniques for convex hypothesis testing developed in [18, 27].

---

coloring inference $\mathcal{T}$ utilizing (unique) observation ω, $\omega \sim \mathcal{N}(\mu, \sigma^2 I_d)$ with

$$\sigma \leq \frac{q_{\mathcal{N}}(1-\epsilon)}{q_{\mathcal{N}}\left(1 - \frac{\varepsilon}{\max[b,r]}\right)}\bar{\sigma}$$

has its risk bounded with ε. Here $q_{\mathcal{N}}(p)$ is the p-quantile of $\mathcal{N}(0,1)$: $\mathrm{Prob}_{s\sim\mathcal{N}(0,1)}\{s \leq q_{\mathcal{N}}(p)\} = p$, $0 \leq p \leq 1$.

Notational conventions. We denote by $\mathcal{O}$ and $\mathcal{U}$ the sets of all ordered pairs (i,j) (resp., unordered pairs $\{i,j\}$) with $1 \le i,j \le N$ and $j \ne i$.

In the sequel we fix ℓ_* and $x_* \in X_{\ell_*}$ and, in accordance with what was said above, deal with repeated observations with i.i.d. components $\omega_k \sim p_{\mathcal{A}_{\ell_*}(x_*)}$. We denote by $\widetilde{\Omega}^{\overline{K}}$ the set of all realizations of the "preliminary" (pilot) observation $\omega^{\overline{K}}$ such that

$$\|x_* - \widetilde{x}_{\ell_*}(\omega^{\overline{K}})\| \le \mathfrak{r}_{\ell_*} := \mathfrak{r}_{\ell_*}^{\overline{K}}. \tag{11}$$

Due to (10) the $p^L_{\mathcal{A}_{\ell_*}(x_*)}$-probability of $\widetilde{\Omega}^{\overline{K}}$ is at least $1-\epsilon$.

Note: *From now on we fix a realization $\widetilde{\omega}^{\overline{K}} \in \widetilde{\Omega}^{\overline{K}}$ of the preliminary observation $\omega^{\overline{K}}$*; in what follows, ω^K is the secondary (post-pilot) K-repeated observation, $\omega^K = (\omega_{\overline{K}+1}, ..., \omega_{\overline{K}+K})$. For notational convenience, in the sequel, we suppress explicit reference to $\widetilde{\omega}^{\overline{K}}$ when defining/denoting subsequent entities which in fact depend on $\widetilde{\omega}^{\overline{K}}$ as parameter.

3.2 Case of general seminorm

3.2.1 Construction

The aggregation routine is as follows.

1. For $1 \le i \ne j \le N$ we put

$$\begin{aligned} x_i &= \widetilde{x}_i(\widetilde{\omega}^{\overline{K}}), \\ B_i &= B_i(\widetilde{\omega}^{\overline{K}}) = \big\{x \in X_i : \|x - x_i\| \le \mathfrak{r}_i := \mathfrak{r}_i^{\overline{K}}(\epsilon)\big\}, \\ \delta_{ij} &= \delta_{ij}(\widetilde{\omega}^{\overline{K}}) = \tfrac{1}{2} \min_{x \in B_i, y \in B_j} \|x - y\|, \end{aligned} \tag{12}$$

with the standard convention that minimum over an empty set is $+\infty$.
We specify hypotheses $H_i = H_i\big(B_i(\widetilde{\omega}^{\overline{K}})\big)$ "the observations stem from a pair (i,x) with $x \in B_i(\widetilde{\omega}^{\overline{K}})$" (equivalently: H_i states that the distribution of independent across $k \le K$ observations $\omega_{\overline{K}+k}$ belongs to the set $M_i = \{\mathcal{A}_i(x) : x \in B_i\}$). Note that sets $M_i = M_i(\widetilde{\omega}^{\overline{K}})$ are convex and compact subsets of $\mathcal{M}$.

Note: Everywhere in the sequel we assume w.l.o.g. that all hypotheses H_i, $i = 1, ..., N$, are nonempty (i.e., from the start, we reject all empty hypotheses and update accordingly N and the indexes of remaining points x_i and sets X_j.
Given a pair $(i,j) \in \mathcal{O}$, it may happen that there is a simple detector-based K-observation test $\mathcal{T}_{\{i,j\}}$ as built in Section 2.2, which decides on H_i vs H_j with risk bounded with $\epsilon/(N-1)$; in such case, we say that pairs (i,j) and (j,i) are K-good, and say that these pairs are K-bad otherwise. We skip the prefix "K-" when the value of K is clear from the context.

2. Let for $i \le N$ $\mathcal{J}_i$ be the set of $j \le N$, $j \ne i$, such that the pair (i,j) is good; note that $j \in \mathcal{J}_i$ if and only if $i \in \mathcal{J}_j$. For all $i \le N$ and $j \in \mathcal{J}_i$ we run tests $\mathcal{T}_{\{i,j\}}$. We call index i admissible if hypothesis H_i was never rejected by corresponding tests (i.e., all tests $\mathcal{T}_{\{i,j\}}$ (if any) with $j \in \mathcal{J}_i$

accepted H_i; in particular, i is admissible, if no pair (i,j) with $j \neq i$ is good). We denote $\mathcal{I} = \mathcal{I}(\omega^K)$ the set of all admissible i's.

The output of the procedure—the aggregated estimate $\widehat{x} = \widehat{x}(\omega^K)$—is selected as $x_{\widehat{i}}$ where $\widehat{i} = \widehat{i}(\omega^K)$ is the smallest of admissible i's when set $\mathcal{I}$ is not empty, and selected as, say, x_1 otherwise.

We have the following straightforward bound for the error of $\widehat{x}$.

Proposition 3. *In the situation described in Section 3.1, let $\overline{\Omega}^K$ be the set of all ω^K satisfying the condition*

> *All tests $\mathcal{T}_{\{\ell_*,j\}}$ in good pairs (ℓ_*, j), as applied to observation ω^K, accept the hypothesis H_{ℓ_*}.*

Then (ℓ_*, x_*)-probability[10] of $\overline{\Omega}^K$ is at least $1-\epsilon$, and for all $\omega^K \in \overline{\Omega}^K$ the set $\bigcup_{i\in\mathcal{I}} B_i$ covers x_*. Furthermore, for such ω^K one has

$$\|x_* - \widehat{x}(\omega^K)\| \le \|x_* - x_{\ell_*}\| + \max_{j\in\mathcal{I}^-_{\ell_*}} \|x_j - x_{\ell_*}\|, \qquad \mathcal{I}^-_{\ell_*} = \{j \in \mathcal{I}, j < \ell_*\} \tag{13}$$

(by convention, the maximum over an empty set is zero). Moreover,

$$\begin{array}{rcl} \max\limits_{j\in\mathcal{I}^-_{\ell_*}} \|x_j - x_{\ell_*}\| & \le & \max\limits_{j\in J^-_{\ell_*}} \|x_j - x_{\ell_*}\| \le \mathfrak{r}_{\ell_*} + \max\limits_{j\in J^-_{\ell_*}} (2\delta_{\ell_* j} + \mathfrak{r}_j), \\ J^-_{\ell_*} & = & \{j < \ell_* : \ (\ell_*, j) \ \textit{is K-bad}\}, \end{array} \tag{14}$$

and

$$\begin{array}{rcl} \max\limits_{j\in\mathcal{I}^-_{\ell_*}} \|x_j - x_{\ell_*}\| & \le & \max\limits_{j\in\mathcal{I}} \|x_j - x_{\ell_*}\| \le \max\limits_{j\in J_{\ell_*}} \|x_j - x_{\ell_*}\| \le \max\limits_{(i,j)\in\overline{J}} \|x_j - x_i\| \\ & \le & 2\max\limits_i \mathfrak{r}_i + 2 \max\limits_{(i,j)\in\overline{J}} \delta_{ij} \end{array} \tag{15}$$

$$J_{\ell_*} = \{j \neq \ell_* : (\ell_*, j) \ \textit{is K-bad}\}, \ \ \overline{J} = \{(i,j) \in \mathcal{O} : (i,j) \ \textit{is K-bad}\}.$$

3.2.2 Risk analysis

Given a pair $(i,j) \in \mathcal{O}$ and $\varepsilon \in (0, 1/2)$ consider the quantity

$$\mathfrak{r}^K_{ij}(\varepsilon) = \tfrac{1}{2} \max_{x\in X_i, y\in X_j} \left\{ \|x - y\| : \ \varrho(\mathcal{A}_i(x), \mathcal{A}_j(y)) \ge \varepsilon^{1/K} \right\} \tag{16}$$

where $\varrho(\cdot,\cdot)$ is as defined in (3) (here, as before, the maximum over an empty set is 0, by definition). In what follows we refer to $\mathfrak{r}^K_{ij}(\varepsilon)$ as *separation ε-risk* over X_i, X_j.

Theorem 1. *In the situation described in Section 3.1, the just built adaptive estimate $\widehat{x}^{(a)}$ (as function of pilot observation $\omega^{\overline{K}}$ and independent (secondary) observation ω^K) satisfies*

$$\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X_i] \le 2\mathfrak{r}^{\overline{K}}_i(\epsilon) + \max_{j<i} \left[\mathfrak{r}^{\overline{K}}_j(\epsilon) + 2\mathfrak{r}^K_{ij}\big(\epsilon/(N-1)\big)\right] \quad \forall i \le N. \tag{17}$$

[10] From now on, for $j \le N$ and $x \in X_j$ "(j,x)-probability" of an event is its $p^K_{\mathcal{A}_j(x)}$-probability.

Moreover, whenever $K > \overline{\vartheta}^{-1}\overline{K}$ *where*

$$\overline{\vartheta} := \frac{\ln(4\epsilon(1-\epsilon))}{2\ln(\epsilon/(N-1))} \le 1,$$

one has

$$\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X_i] \le 2\mathfrak{r}_i^{\overline{K}}(\epsilon) + \max_{j<i}\left[\mathfrak{r}_j^{\overline{K}}(\epsilon) + 2\mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j]\right] \quad \forall i \le N. \tag{18}$$

In addition, in the special case where for every pair i, j *there exists* $x_{ij} \in X_i \cap X_j$ *such that* $\mathcal{A}_i(x_{ij}) = \mathcal{A}_j(x_{ij})$ *one has for all* $K \ge \overline{K}$ *and* $i \le N$*:*

$$\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X_i] \;\le\; 2\mathfrak{r}_i^{\overline{K}}(\epsilon) + \max_{j<i}\left[\mathfrak{r}_j^{\overline{K}}(\epsilon) + 2\overline{\vartheta}^{-1}\mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j]\right]. \tag{19}$$

Theorem 1 has the following straightforward corollary.

Corollary 1. *Under the premise of Theorem 1, suppose that upper bounds* $\mathfrak{r}_i^{\overline{K}}(\epsilon)$ *on partial risks of estimates* $\widetilde{x}_i(\omega^{\overline{K}})$ *are within factor* θ *of the respective* $\overline{K}$*-observation minimax risks, i.e.,*

$$\mathrm{RiskOpt}^{\{i\}}_{\epsilon,\overline{K}}[X_i] \le \mathfrak{r}_i^{\overline{K}}(\epsilon) \le \theta\mathrm{RiskOpt}^{\{i\}}_{\epsilon,\overline{K}}[X_i].$$

Then the risk of estimate $\widehat{x}^{(a)}$ *is within a moderate factor of the minimax* $\overline{K}$*-observation risk. For instance, whenever* $K \ge \overline{\vartheta}^{-1}\overline{K}$ *one has*

$$\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X_i] \le (2+3\theta)\max_{j\le i}\mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j] \quad \forall i \le N, \tag{20}$$

and

$$\begin{aligned}\mathrm{RiskOpt}^{\overline{1,N}}_{2\epsilon,\overline{K}+K}[X] \;&\le\; \mathrm{Risk}^{\overline{1,N}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X] \le [2+3\theta]\max_{j,i\le N}\mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j] \\ &\le\; (2+3\theta)\mathrm{RiskOpt}^{\overline{1,N}}_{\epsilon,\overline{K}}[X].\end{aligned} \tag{21}$$

In the case where for every pair i, j *there exists* $x_{ij} \in X_i \cap X_j$ *such that* $\mathcal{A}_i(x_{ij}) = \mathcal{A}_j(x_{ij})$ *one has for all* $K \ge \overline{K}$ *and* $i \le N$*:*

$$\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X_i] \le (3\theta + 2\overline{\vartheta}^{-1})\max_{j\le i\le N}\mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j],$$

so that

$$\mathrm{Risk}^{\overline{1,N}}_{2\epsilon,2\overline{K}}[\widehat{x}^{(a)}|X] \le \max_{i,j\le N}(3\theta + 2\overline{\vartheta}^{-1})\mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j] \le (3\theta + 2\overline{\vartheta}^{-1})\mathrm{RiskOpt}^{\overline{1,N}}_{\epsilon,\overline{K}}[X]. \tag{22}$$

Discussion. Bounds (20), (21) imply that under the premise of the corollary, the minimax risk $\mathrm{RiskOpt}^{\overline{1,N}}_{2\epsilon,\overline{K}+K}[X]$ of estimation over union X of sets X_i, $i = 1, ..., , N$, is similar, modulo logarithmic factors, to the maximal "pairwise" minimax risk $\max_{j,i\le N}\mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j]$ of estimation over *pairwise unions* $X_i \cup X_j$ of sets. Furthermore, the upper bound (20) on the maximal over X_i risk $\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X_i]$ of adaptive estimate $\widehat{x}^{(a)}$ is also similar, in the same sense, to the maximal risk

$\max_{j\le i}\mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i\cup X_j]$ of estimation over *pairwise unions* $X_j\cup X_i$ with $j\le i$ and depends on the selected ordering of X_i's. In particular, when this order is chosen so that partial risks of estimation over X_i satisfy

$$\mathfrak{r}_1^{\overline{K}}(\epsilon)\le\mathfrak{r}_2^{\overline{K}}(\epsilon)\le....\le\mathfrak{r}_N^{\overline{K}}(\epsilon)$$

and pairwise separation risks are dominated by partial risks, i.e.,

$$\mathfrak{r}_{ij}^{K}\big(\epsilon/(N-1)\big)\le C\mathfrak{r}_i^{\overline{K}}(\epsilon)\quad\forall(i,j,\ 1\le j<i\le N),\tag{23}$$

one has

$$\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X_i]\le C'\mathfrak{r}_i^{\overline{K}}(\epsilon)\quad\forall i\le N,$$

and estimate $\widehat{x}^{(a)}$ is adaptive in the sense of [33, 34]. On the other hand, when relations (23) do not hold, adaptation in the above sense is impossible what can be seen already when $N=2$. Similar comments are applicable to bound (22).

3.3 Estimate aggregation, case of Euclidean seminorm

We continue to consider the situation described in Section 3.1. However, from now on we assume that $\|\cdot\|$ is a Euclidean seminorm such that $\|x\|=\|Bx\|_2$ where $B\in\mathbf{R}^{\nu\times n}$ is a given matrix. We build an adaptive estimate of the signal $x_*\in X_{\ell_*}$ underlying our observations: $\omega_k\sim p_{\mathcal{A}_{\ell_*}(x_*)}$ by aggregating preliminary estimates —selecting the closest to x_* point among $x_i=\widetilde{x}_i^{\overline{K}}(\widetilde{\omega}^{\overline{K}})$, $1\le i\le N$, where, same as before, $\widetilde{\omega}^{\overline{K}}\in\widetilde{\Omega}^{\overline{K}}$ is fixed.

3.3.1 Construction

We are given the number K of observations and tolerance parameters $\epsilon\in(0,1)$ and $\underline{\delta}>0$; we put $\overline{N}=2N(N-1)$.

Preliminaries

• Denote $W_i=BX_i$, $i=1,...,N$, with $W=BX$. Assuming, for the sake of simplicity, that all points $w_i=Bx_i$, $i=1,...,N$, are distinct, we associate with each pair $(i,j)\in\mathcal{O}$ the quantities

$$r_{ij}=\tfrac{1}{2}\|w_i-w_j\|_2,$$

vectors $\psi_{ij}=\frac{w_j-w_i}{\|w_j-w_i\|_2}$, $w_{ij}=\frac{1}{2}(w_i+w_j)$, and for $\delta>0$ consider sets

$$W_{ij}^{\ell-}=\{v\in W_\ell:\ \psi_{ij}^T(v-w_{ij})\le0\},\ \ W_{ij}^{\ell+}(\delta)=\{v\in W_\ell:\psi_{ij}^T(v-w_{ij})\ge\delta\},\ \ell=1,...,N.$$

Observe that $W_{ij}^{\ell-}$ is exactly the set of $v\in W_\ell$ such that $\|v-w_i\|_2\le\|v-w_j\|_2$, while $W_{ji}^{\ell+}(\delta)$ is the set of $v\in W_\ell$ such that

$$\|v-w_j\|_2^2\le\|v-w_i\|_2^2-2\delta\|w_i-w_j\|_2.$$

• Let us fix a quadruple $(i,j;\ell,\ell')$, $1\le i\ne j\le N$ and $1\le\ell,\ell'\le N$. We denote $H_{ij}^{\ell-}$ (resp., $H_{ij}^{\ell'+}(\delta)$) the hypothesis stating that observation ω^K stems from (ℓ,x) with $x\in X_\ell$ such that $w=Bx\in W_{ij}^{\ell-}$ (resp., such that observation ω^K stems from (ℓ',x) with $x\in X_{\ell'}$ and $w\in W_{ij}^{\ell'+}(\delta)$). We say that $\delta\in(0,r_{ij}]$ is $(i,j;\ell,\ell')$-good if there exists a detector-based test $\mathcal{T}_{ij}^{\ell\ell'}$ deciding on hypothesis $H_{ij}^{\ell-}$ vs

$H_{ij}^{\ell'+}(\delta)$ with risk $\leq \varepsilon = \epsilon/\overline{N}$. When good δ's exist, we say that the quadruple $(i,j;\ell,\ell')$ itself is (ε-)good. It is obvious that if $\delta' \in [0, r_{ij}]$ is good, so are all $\delta \in [\delta', r_{ij}]$. Note that goodness of δ can be checked efficiently, i.e., when $(i,j;\ell,\ell')$ is good one can efficiently find, e.g., by bisection, the value $\delta_{ij}^{\ell\ell'}$ such that $\delta_{ij}^{\ell\ell'}$ is good while $\delta_{ij}^{\ell\ell'} - \underline{\delta}$ is not. When $\delta = r_{ij}$ is not $(i,j;\ell,\ell')$-good we say that the corresponding quadruple is bad and set $\delta_{ij}^{\ell\ell'} = r_{ij}$.

Aggregation procedure

The output of the procedure are two aggregated estimates $\widehat{x}$ and $\widetilde{x}$.

1. For each $(i,j;\ell,\ell')$, $1 \leq i \neq j \leq N$ and $1 \leq \ell, \ell' \leq N$, we act as follows:
   • we reject the alternative $H_{ij}^{\ell'+}(r_{ij})$ if the quadruple in question is bad;
   • when $(i,j;\ell,\ell')$ is good we apply to observation ω^K test $\mathcal{T}_{ij}^{\ell\ell'}$ of hypothesis $H_{ij}^{\ell-}$ against $H_{ij}^{\ell\ell'+} = H_{ij}^{\ell'+}(\delta_{ij}^{\ell\ell'})$.
   We say that pair $(i;\ell)$ is *admissible* if corresponding hypotheses $H_{ij}^{\ell-}$ were never rejected by the above procedure. The result of this step is the set $\mathcal{I} = \mathcal{I}(\omega^K)$ of all admissible pairs $(i;\ell)$.
2. If $\mathcal{I}(\omega^K) = \emptyset$ we select the aggregated solution as one of x_i, e.g., $\widehat{x} = x_1$; when $\mathcal{I}(\omega^K)$ contains pairs corresponding to a unique index $\widehat{i} = \widehat{i}(\omega^K)$, we output $\widehat{x}(\omega^K) = x_{\widehat{i}}$ as aggregated solution. Otherwise,
   • we select $\widehat{i} = \widehat{i}(\omega^K)$ as (e.g., the smallest) i-component corresponding to admissible pairs $(i;\ell)$ with the smallest value of (the second index) ℓ and define the estimate $\widehat{x}(\omega^K) = x_{\widehat{i}}$.
   • To build the estimate $\widetilde{x}$ we find among w_i corresponding to admissible i's (that is, i-components of admissible pairs $(i;\ell)$) points $w_{\bar{i}}$, $w_{\bar{j}}$ with the maximal length $\|w_{\bar{i}} - w_{\bar{j}}\|_2$ of the connecting segment and select as aggregated solution $\widetilde{x}(\omega^K) = \frac{1}{2}(x_{\bar{i}} + x_{\bar{j}})$ (or choose any $\widetilde{x} \in \mathbf{R}^n$ such that $B\widetilde{x} = w_{\bar{i}\bar{j}}$).

Proposition 4. *Suppose that observation ω^K stems from the pair (ℓ_*, x_*), $x_* \in X_{\ell_*}$. Let i_* be the index of one of the $\|\cdot\|$-closest to x points among $x_1, ..., x_N$, and let $\overline{\Omega}^K$ be the set of realizations ω^K such that as applied to ω^K, all tests $\mathcal{T}_{i_*j}^{\ell_*\ell}$ and $\mathcal{T}_{ji_*}^{\ell\ell_*}$ accept the true, if any, of the hypotheses from the corresponding pair.*[11] *Then the (ℓ_*, x_*)-probability of $\overline{\Omega}^K$ is at least $1-\epsilon$, and for all $\omega^K \in \overline{\Omega}^K$ it holds*

$$\|\widehat{x} - x_*\| \leq \|x_* - x_{i_*}\| + 2\widehat{\delta}_{i_*}(\omega^K) \tag{24}$$

where $\widehat{\delta}_{i_*}^{\ell_*}(\omega^K) = \max_{j \neq i_*, \ell \leq \ell_*, (i;\ell) \in \mathcal{I}(\omega^K)} \delta_{ji_*}^{\ell\ell_*}$. *Furthermore, one has*

$$\|\widetilde{x} - x_*\|^2 \leq \|x_* - x_{i_*}\|^2 + 4\widetilde{\delta}_{i_*}^{\ell_*}(\omega^K)^2 \tag{25}$$

where $\widetilde{\delta}_{i_*}^{\ell_*}(\omega^K) = \max_{j \neq i_*, (j;\ell) \in \mathcal{I}(\omega^K)} \delta_{ji_*}^{\ell\ell_*}$.

In statistical literature, the bound (25) for prediction loss (in Problem I discussed in the introduction, this corresponds to the seminorm $\|x\| = \|Ax\|_2$) is typically obtained utilizing exponential weights (see, e.g., [1, 6, 29, 40]) or Q-aggregation [14, 30]. When risk of aggregation is measured by a Euclidean seminorm, using the aggregation procedure described above this type of results can be painlessly extended to aggregation problems with convex constraints on unknown signals and aggregation from indirect observations.

[11] In other words, as applied to ω^K, test $\mathcal{T}_{i_*j}^{\ell_*\ell}$ accepts $H_{i_*j}^{\ell_*-}$ (recall that $H_{i_*j}^{\ell_*-}$ is the "true hypothesis" in this case), while test $\mathcal{T}_{ji_*}^{\ell\ell_*}$ rejects $H_{ji_*}^{\ell-}$ and accepts $H_{ji_*}^{\ell\ell_*+}$ if $w \in W_{ji_*}^{\ell+}(\delta_{ji_*}^{\ell\ell_*})$.

3.3.2 Risk analysis

Theorem 2. *In the situation of this section, estimate* $\widehat{x}^{(a)}(\omega^{\overline{K}+K}) = \widehat{x}(\omega^K)$ *satisfies for all* $i \le N$*:*

$$\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X_i] \le \mathfrak{r}_i^{\overline{K}}(\epsilon) + 2\left[\max_{j<i} \mathfrak{r}_{ij}^K\big(\epsilon/\overline{N}\big) + \underline{\delta}\right]. \tag{26}$$

Furthermore, for $\widetilde{x}^{(a)}(\omega^{\overline{K}+K}) = \widetilde{x}(\omega^K)$ *one has*

$$\left[\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widetilde{x}^{(a)}|X_i]\right]^2 \le [\mathfrak{r}_i^{\overline{K}}(\epsilon)]^2 + 4\left[\max_{j\ne i, j\le N} \mathfrak{r}_{ij}^K\big(\epsilon/\overline{N}\big) + \underline{\delta}\right]^2 \qquad \forall i \le N \tag{27}$$

(as before, quantities $\mathfrak{r}_{ij}^K(\varepsilon)$ *are given by (16)).*

Consequently, when $K > \overline{\vartheta}^{-1}\overline{K}$ *where*

$$\overline{\vartheta} := \frac{\ln([4\epsilon(1-\epsilon)])}{2\ln(\epsilon/\overline{N})} \le 1,$$

one has

$$\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X_i] \le \mathfrak{r}_i^{\overline{K}}(\epsilon) + 2\max_{j<i} \mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j] + \underline{\delta} \quad \forall i \le N, \tag{28}$$

and

$$\left[\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widetilde{x}^{(a)}|X_i]\right]^2 \le [\mathfrak{r}_i^{\overline{K}}(\epsilon)]^2 + 4\left[\max_{j<i} \mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j] + \underline{\delta}\right]^2 \qquad \forall i \le N. \tag{29}$$

In the case where for every pair i, j *there exists* $x_{ij} \in X_i \cap X_j$ *such that* $\mathcal{A}_i(x_{ij}) = \mathcal{A}_j(x_{ij})$ *one has for all* $K \ge \overline{K}$ *and* $i \le N$*:*

$$\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X_i] \quad \le \quad \mathfrak{r}_i^{\overline{K}}(\epsilon) + 2\overline{\vartheta}^{-1}\max_{j<i} \mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j] + \underline{\delta} \tag{30}$$

and

$$\left[\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widetilde{x}^{(a)}|X_i]\right]^2 \le [\mathfrak{r}_i^{\overline{K}}(\epsilon)]^2 + 4\left[\overline{\vartheta}^{-1}\max_{j<i} \mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j] + \underline{\delta}\right]^2 \qquad \forall i \le N. \tag{31}$$

4 "Generic" test-based aggregation

4.1 Setup

We consider the situation as follows: we are given

- *observation space* Ξ,
- a compact set $X \subset \mathbf{R}^n$, with every $x \in X$ associated with a family $\mathcal{P}_x$ of probability distributions on Ξ; we refer to observations distributed according to $P \in \mathcal{P}_x$ as to observations *stemming* from x,
- a seminorm $\|\cdot\|$ on $\mathbf{R}^n$,
- N distinct points $x_i \in \mathbf{R}^n$, $i = 1, ..., N$.

Given observation $\xi \sim P$ stemming from unknown signal $x \in X$, our objective is to aggregate x_i's—to find among x_i's a point which is "as close to x as the closest to x point among x_i's." Here closeness is measured by the seminorm $\|\cdot\|$.

Note that as far as our goal is concerned, we lose nothing when assuming from now on that $\|x_i - x_j\| > 0$ whenever $i \ne j$.

Conventions

- In the sequel we say that an event (a set in the space of observations) takes place with x-probability at most (or at least) p for some $x \in X$ if this is the case for probability w.r.t. any distribution from $\mathcal{P}_x$.
- With a subset Y of X we associate *hypothesis* $H(Y)$ on the distribution of observation ξ; the hypothesis states that the observation stems from a signal $x \in Y$. Given $Y^1, Y^2 \subset X$, a *test* for the pair of hypotheses $H(Y^1), H(Y^2)$ is a procedure which, given on input an observation ξ, *accepts* exactly one of these two hypotheses (informally: claims that ξ is drawn from a distribution obeying the accepted hypothesis) and rejects the other. We say that such a test has *risk* $\leq \delta$, if "the probability to accept the true hypothesis is at least $1-\delta$," specifically, for $\chi = 1, 2$, when the observation stems from a signal $x \in Y^\chi$, the x-probability for the test to accept $H(Y^\chi)$ is at least $1-\delta$. Note that we allow for Y_1, or Y_2, or both, to be empty; whenever this is the case, the test which always accepts a nonempty hypothesis, if any, and accepts a whatever one of the hypotheses when both Y_1 and Y_2 are empty, has zero risk.
- For $(i, j) \in \mathcal{O}$ and $\delta \geq 0$ we set

$$\rho_i = \min_{x \in X} \|x - x_i\|,\ 1 \leq i \leq N, \tag{32}$$

and

$$r_{ij} = \tfrac{1}{2}\|x_i - x_j\|,\ \ X_{ij}(\delta) = \{z \in X : \|z - x_i\| \leq r_{ij} - \delta\}. \tag{33}$$

Note that $r_{ij} = r_{ji}$.

4.2 Aggregation in general seminorm

4.2.1 The setup

The setup of the general aggregation scheme is given by

1. "reliability tolerance" $\epsilon \in (0, 1)$,
2. a collection $\mathcal{C}$ of pairs $\{i, j\} \in \mathcal{U}$ with each pair $\{i, j\} \in \mathcal{C}$ associated with
   - thresholds $\Delta_{ij} = \Delta_{ji} \in [0, r_{ij}]$, giving rise to sets $X_{ij}(\Delta_{ij})$, $X_{ji}(\Delta_{ij})$ and hypotheses $H_{ij} = H(X_{ij}(\Delta_{ij}))$, $H_{ji} = H(X_{ji}(\Delta_{ij}))$, along with
   - a test $\mathcal{T}_{\{i,j\}}$ deciding on the hypotheses H_{ij} and H_{ji} with risk $\leq \epsilon/(N-1)$.

   When $\{i, j\} \in \mathcal{C}$, we say that i and j are comparable (same as "j is comparable to i" and "i is comparable to j").
3. For pairs $\{i, j\} \in \mathcal{U}$ with incomparable to each other i and j (i.e., $\{i, j\} \in \mathcal{U} \backslash \mathcal{C}$) we set

$$\Delta_{ij} = \Delta_{ji} = \max[0, r_{ij} - \max[\rho_i, \rho_j]].$$

4.2.2 Aggregation routine

Aggregation routine associated with the just described setup is as follows

1. Given observation ξ, for every pair $(i, j) \in \mathcal{O}$ we "compare i to j" according to the following rule:
   - when i, j are comparable, we run the test $\mathcal{T}_{\{i,j\}}$ on observation ξ and claim that i looses to j when the test accepts the hypothesis H_{ji}, and claim that j looses to i otherwise

- when i, j are incomparable, i looses to j whenever $\rho_j \le \rho_i$, otherwise j looses to i.

2. For $i \le N$, we denote by $\mathcal{I}_i = \mathcal{I}_i(\xi)$ the set of indices $j \ne i$ such that i looses to j, set

$$d_i(\xi) = \max_{j \in \mathcal{I}_i(\xi)} \|x_j - x_i\|,$$

and define the aggregated estimate as $\widehat{x}(\xi) = x_{\widehat{i}(\xi)}$ where

$$\widehat{i} = \widehat{i}(\xi) \in \operatorname*{Argmin}_i d_i(\xi).$$

We have the following simple statement.

Proposition 5. *Suppose that the observation stems from a signal $x_* \in X$, and let x_{i_*} be one of the $\|\cdot\|$-closest to x_* points among $x_1, ..., x_N$. Let $\overline{\Xi}$ be the set of realizations ξ satisfying the following condition:*

For every $j \ne i_*$ such that i_* and j are comparable and $x_* \in X_{i_*j}(\Delta_{i_*j})$, test $\mathcal{T}_{\{i_*,j\}}$ as applied to observation ξ accepts the hypothesis H_{i_*j}.

Then the x_*-probability of $\overline{\Xi}$ is at least $1-\epsilon$, and for all $\xi \in \overline{\Xi}$

$$\|x_* - \widehat{x}(\xi)\| \le 3\|x_{i_*} - x_*\| + 2\overline{\Delta}_{i_*}(\xi) \tag{34}$$

where

$$\overline{\Delta}_{i_*}(\xi) = \begin{cases} 0, & \mathcal{I}_{i_*}(\xi) = \emptyset, \\ \max_{j \in \mathcal{I}_{i_*}(\xi)} \Delta_{i_*j}, & \mathcal{I}_{i_*}(\xi) \ne \emptyset. \end{cases}$$

Remarks. The above construction is inspired by the aggregation procedure of [17] which itself generalizes the results on density estimation with ℓ_1-loss from [43, 37, 16]; it can also be seen as a refinement of the selection procedure of [27, Section 2.5.3].

The question of (near-)optimality of the accuracy bound (34) for the proposed routine is more involved in the considered here general framework than in the direct observation setting of [17]; we postpone the corresponding analysis till Section 5.2. Note, however, that there are in fact two questions—that of optimality of the factor "3" in front of the minimal loss $\|x_{i_*} - x\|$ which is related to problem geometry (and is independent of the observation scheme), and that of the size of the additive term $\overline{\Delta}$. It appears that in the problem of aggregation of densities when $\|\cdot\|$ is the ℓ_1-norm the factor 3 in front of the minimal error is (in a certain precise sense, cf. [5]) unimprovable even for problems with $N = 2$. On the other hand, when allowing for "improper aggregation," i.e., when removing the limitation of the aggregated solution to be one of given points [37] supplies a randomized algorithm which attains the factor 2 when $N = 2$, and factor 2 is, in a certain sense, unimprovable in the latter setting, see [12]. However, known to us attempts to generalize this kind of result to the case of $N > 2$ (cf. [5]) result in the inflation of the additive term which is too important in the situation of small minimal loss we are mainly interested here. There is however a situation where the factor 3 can be removed rather painlessly (at the price of a moderate increase of $\overline{\Delta}$)—this is the case of Euclidean seminorm $\|\cdot\|$, and this is the situation we consider next.

4.3 Aggregation in Euclidean seminorm

Now assume that $\|\cdot\|$ is a Euclidean seminorm: $\|x\| = \|Bx\|_2$ for a given matrix B. For $\delta \ge 0$ and $(i, j) \in \mathcal{O}$ we define

$$\mathcal{X}_{ij}(\delta) = \{z \in X : \|z - x_j\| \ge \delta + \|z - x_i\|\}. \tag{35}$$

Aggregation procedures presented below are refined versions of the routine from [24].

4.3.1 The setup

The setup for the Euclidean aggregation is given by

1. thresholds Δ_{ij}, $(i,j) \in \mathcal{O}$, such that
$$\Delta_{ij} = \Delta_{ji} \geq 0$$
2. tests $\mathcal{T}_{\{i,j\}}$, $\{i,j\} \in \mathcal{U}$, with $\mathcal{T}_{\{i,j\}}$ testing the hypothesis $\mathcal{H}_{ij} = H(\mathcal{X}_{ij}(\Delta_{ij}))$ vs. the alternative $\mathcal{H}_{ji} = H(\mathcal{X}_{ji}(\Delta_{ji}))$ such that
   - if both hypotheses $\mathcal{H}_{ij}$ and $\mathcal{H}_{ji}$ are empty, $\mathcal{T}_{\{i,j\}}$ accepts both hypotheses;
   - if exactly one of the hypotheses $\mathcal{H}_{ij}$, $\mathcal{H}_{ji}$ is empty, the test always accepts the nonempty hypothesis and rejects the empty one;
   - if hypotheses $\mathcal{H}_{ij}$ and $\mathcal{H}_{ji}$ are nonempty (in this case, we refer to the pairs (i,j) and (j,i) as *good*) the test accepts exactly one of these hypotheses, and the risk of the test does not exceed $\epsilon/\overline{N}$, $\overline{N} = \frac{N(N-1)}{2}$.

4.3.2 Aggregation routine

Aggregation routine associated with the above setup is as follows: given observation ξ, we run tests $\mathcal{T}_{\{i,j\}}$, $\{i,j\} \in \mathcal{U}$, and for every $1 \leq i \leq N$ record the "score of i"—the number $s_i(\xi)$ of those $j \neq i$, $j \leq N$ for which the test $\mathcal{T}_{\{i,j\}}$ rejects $\mathcal{H}_{ij}$. We put

$$\widehat{i}(\xi) \in \operatorname*{Argmin}_{1 \leq i \leq N} s_i(\xi)$$

and define aggregated solution as $\widehat{x}(\xi) = x_{\widehat{i}(\xi)}$.

Proposition 6. *Suppose that the observation stems from a signal $x_* \in X$. Let $\overline{\Xi}$ be the set of realizations of ξ such that*

> *as applied to observation ξ, each test $\mathcal{T}_{\{i,j\}}$ associated with a good pair (i,j) does not reject the "true hypothesis," if any (i.e., as applied to ξ, the test accepts $\mathcal{H}_{ij}$ when $x_* \in \mathcal{X}_{ij}(\Delta_{ij})$, and accepts $\mathcal{H}_{ji}$ when $x_* \in \mathcal{X}_{ji}(\Delta_{ij})$).*

Then the x_*-probability of $\overline{\Xi}$ is at least $1-\epsilon$, and for all $\xi \in \overline{\Xi}$ one has

$$\|x_* - \widehat{x}(\xi)\| \leq \|x_{i_*} - x_*\| + 2\overline{\Delta}$$

where x_{i_*} is one of the $\|\cdot\|$-closest to x_* points among $x_1, ..., x_N$ and $\overline{\Delta} = \max_{i \neq j} \Delta_{ij}$.

5 Test-based aggregation in simple observation schemes

5.1 Problem setting

In the sequel, we deal with the situation as follows. Given are:

1. simple o.s. $\mathcal{SO} = ((\Omega, \Pi), \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F})$,
2. a collection of J convex compact sets $X_\nu \subset \mathbf{R}^n$, giving rise to the set $X = \cup_{\nu=1}^J X_\nu$,
3. affine mappings $x \mapsto \mathcal{A}_\nu(x)$ such that $\mathcal{A}_\nu(X_\nu) \subset \mathcal{M}$, $\nu = 1, ..., J$,
4. a seminorm $\|\cdot\|$ on $\mathbf{R}^n$,

5. N points $x_i \in \mathbf{R}^n$, $i = 1, ..., N$.

Our objective is given a stationary repeated observation $\omega^K = (\omega_1, ..., \omega_K)$ with

$$\omega_k \sim p_{\mathcal{A}_\nu(x)},\ k = 1, 2, ..., K,$$

for some *unknown* pair (ν, x) with $\nu \le J$ and $x \in X_\nu$, to recover one of the $\|\cdot\|$-closest to x points among $x_1, ..., x_N$.

Note that we are in the situation of Section 4.1, with

- Ω^K in the role of observation space Ξ, and K-repeated observation ω^K in the role of observation ξ,
- the signal set $X = \bigcup_{\nu=1}^J X_j \subset \mathbf{R}^n$,
- the family $\mathcal{P}_x := \mathcal{P}_x^K$ of probability distributions of observations stemming from a signal $x \in X$ being comprised of all distributions with densities $p^K_{\mathcal{A}_\nu(x)}(\omega^K)$ and ν satisfying $x \in X_\nu$, the densities being taken w.r.t. the reference measure Π^K.

Thus, by convention taken in Section 4.1, claim that an event takes place with x-probability at most (or at least) p, $x \in X$, means that this is the case for probability w.r.t. any density $p^K_{\mathcal{A}_\nu(x)}$ of observation ω^K with ν satisfying $x \in X_\nu$.

We are about to achieve our goal of recovering one of the $\|\cdot\|$-closest to x points among $x_1, ..., x_N$ via techniques developed in Section 4, and in what follows we use terminology and notation from that section.

5.2 Aggregation in general seminorm

Our current objective is to describe an implementation of the aggregation procedure of Section 4.2 in the present setting.

5.2.1 Preliminaries

Given number K of observations and $\epsilon \in (0, 1)$, in order to build for $(i, j) \in \mathcal{O}$ the quantities Δ_{ij} and tests $\mathcal{T}_{i,j\}}$, as required by construction from Section 4.2, we act as follows.

• Let us associate with $\delta \ge 0$ and $(i, j) \in \mathcal{O}$ sets $X_{ij}(\delta)$, $X_{ji}(\delta)$ (see (33)) and hypotheses

$$H_{ij}[\delta] = H(X_{ij}(\delta)),\ \ H_{ji}[\delta] = H(X_{ji}(\delta)).$$

• Let a pair $(i, j) \in \mathcal{O}$ be fixed. Given $\delta \ge 0$ such that $X_{ij}(\delta) \ne \emptyset$ and $X_{ji}(\delta) \ne \emptyset$, or, which is the same, such that

$$0 \le \delta \le \overline{\delta}^{ij} := r_{ij} - \max[\rho_i, \rho_j] \tag{36}$$

where ρ_s are defined by (32), observe that $X_{ij}(\delta)$ is a finite union of convex compact sets:

$$X_{ij}(\delta) = \bigcup_{\nu=1}^J \Big\{X_\nu \cap \{z :\ \|z - x_i\| \le r_{ij} - \delta\}\Big\}.$$

We specify collection $\mathcal{R}_{ij}(\delta) = \{R^s_{ij}(\delta) : 1 \le s \le J^{\delta_r}_{ij}\}$ of "red" nonempty convex compact sets as the collection of all *nonempty* sets of the form

$$R_{ij\nu}(\delta) = \{\mathcal{A}_\nu(x) :\ x \in X_\nu,\ \|x - x_i\| \le r_{ij} - \delta\},\ 1 \le \nu \le J.$$

Similarly, we specify the collection $\mathcal{B}_{ij}(\delta) = \{B_{ij}^s(\delta) : 1 \le s \le J_{ij}^{\delta_b}\}$ of "blue" nonempty convex compact sets as the collection of all nonempty sets of the form

$$B_{ij\mu}(\delta) = \{\mathcal{A}_\mu(x) : x \in X_\mu,\ \|x - x_j\| \le r_{ij} - \delta\},\ 1 \le \mu \le J.$$

When applying to $\mathcal{R}_{ij}(\delta)$ and $\mathcal{B}_{ij}(\delta)$ the color inferring test from Section 2.3, depending on δ it may happen that the risk bound ϵ_K of the inference as defined in Section 2.3 satisfies $\epsilon_K \le \epsilon/(N-1)$. Let us refer to δ as (i,j)-*appropriate*, if $0 \le \delta \le \overline{\delta}^{ij}$ and $\epsilon_K \le \epsilon/(N-1)$.

• Given i, j and δ satisfying (36), we can check efficiently whether δ is (i,j)-appropriate—to this end we should compute the spectral norm of a $[J_{ij}^{\delta_r} + J_{ij}^{\delta_b}] \times [J_{ij}^{\delta_r} + J_{ij}^{\delta_b}]$ symmetric matrix filled with optimal values of $J_{ij}^{\delta_r} J_{ij}^{\delta_b}$ explicit convex optimization problems. Clearly, if δ is (i,j)-appropriate and $\delta' \in [\delta, \overline{\delta}^{ij}]$, then δ' is (i,j)-appropriate along with δ.

• Let us call (i,j) *appropriate*, if $\overline{\delta}^{ij}$ is nonnegative and (i,j)-appropriate. In this case the infimum $\underline{\delta}^{ij}$ of (i,j)-appropriate $\delta \in [0, \overline{\delta}^{ij}]$ is well defined, and bisection in δ allows to obtain rapidly (i,j)-appropriate upper bounds on $\underline{\delta}^{ij}$ to whatever high accuracy. The bottom line is that one can efficiently check whether the pair $(i,j) \in \mathcal{O}$ is appropriate, and whenever it is the case, the quantity

$$\underline{\delta}^{ij} \in \left[0, \overline{\delta}^{ij} = r_{ij} - \max[\rho_i, \rho_j]\right]$$

is efficiently computable, and whenever

$$\Delta_{ij} = \Delta_{ij} = \widetilde{\delta}_{ij} \in [0, \overline{\delta}_{ij}]$$

with an (i,j)-appropriate $\widetilde{\delta}_{ij}$, we can point out K-observation test $\mathcal{T}_{ij}$ which decides on the hypotheses $H_{ij}(\Delta_{ij})$, $H_{ji}(\Delta_{ji})$ with risk $\le \epsilon/(N-1)$. Under the circumstances, the latter means that as applied to observation ω^K, test $\mathcal{T}_{ij}$ accepts at most one of the hypotheses $H_{ij}(\Delta_{ij})$, $H_{ji}(\Delta_{ij})$, and whenever $\omega^k \sim p^K_{A_\nu(x)}$ for ν and x such that $x \in X_\nu$, the probability for $\mathcal{T}_{ij}$ to accept $H_{ij}(\Delta_{ij})$ is at least $1 - \epsilon/(N-1)$ when $\|x - x_i\| \le r_{ij} - \Delta_{ij}$, and the probability for $\mathcal{T}_{ij}$ to accept $H_{ji}(\Delta_{ij})$ is at least $1 - \epsilon/(N-1)$ when $\|x - x_j\| \le r_{ij} - \Delta_{ij}$. Note that for an appropriate pair (i,j), the above $\widetilde{\delta}_{ij}$ can be made arbitrarily close to $\underline{\delta}_{ij}$.

• As is immediately seen, a pair (i,j) is appropriate if and only if so is the pair j, i, and because $\overline{\delta}_{ij}$ and $\underline{\delta}_{ij}$ are symmetric in i, j, δ is (i,j)-appropriate if and only if δ is j,i-appropriate, which allows to restrict ourselves to $\widetilde{\delta}_{ij}$ which are symmetric in i, j as well. Consequently, appropriateness of a pair, appropriateness of a δ for this pair, and the parameters $\overline{\delta}_{ij}$, $\underline{\delta}_{ij}$, $\widetilde{\delta}_{ij}$ are attributes of *unordered* pair $\{i,j\}$ rather than of the ordered pair (i,j). For appropriate $\{i,j\}$, let us set $\overline{i} = \min[i,j]$, $\overline{j} = \max[i,j]$ and $\mathcal{T}_{\{i,j\}} = \mathcal{T}_{\overline{i}\overline{j}}$, so that $\mathcal{T}_{\{i,j\}}$ decides on H_{ij} vs. H_{ji} with risk $\le \epsilon/(N-1)$.

5.2.2 Aggregation routine

Consider the following procedure.

- We specify the set $\mathcal{G}$ of all appropriate pairs $\{i,j\} \in \mathcal{U}$ along with the related quantities $\widetilde{\delta}_{ij}$ (the smaller the better) and tests $\mathcal{T}_{\{i,j\}}$. Next, we declare a whatever subset $\mathcal{C}$ of the set $\mathcal{G}$ to be the set of comparable pairs of indices as defined in Section 4.2, and set

$$\Delta_{ij} = \begin{cases} \widetilde{\delta}_{ij}, & \{i,j\} \in \mathcal{C} \\ \max[0; r_{ij} - \max[\rho_i, \rho_j]], & \text{otherwise} \end{cases}$$

With the thresholds Δ_{ij} just defined, K-repeated observation ω^K in the role of ξ, and with $\mathcal{C}$ in the role of the set of comparable pairs and associated tests, we arrive at the aggregation setup as described in Section 4.2, satisfying all the requirements from that section.

- Given observation ω^K, we apply the aggregation procedure associated with the above setup, resulting in the aggregated estimate $\widehat{x}(\omega^K)$.

Results of Propositions 5 and 1 imply the following property of the resulting estimate.

Proposition 7. *In the situation of this section, suppose that the just described routine is applied to observation ω^K stemming from $x_* \in X$, so that $\omega^K \sim p^K_{\mathcal{A}_\nu(x_*)}$ for some $\nu \le J$ such that $x_* \in X_\nu$. Let also i_* be the index of one of the $\|\cdot\|$-closest to x_* points among $x_1, ..., x_N$. Finally, let $\overline{\Omega}$ be the set of all ω^K satisfying the condition (cf. Proposition 5)*

> *For every $j \neq i_*$ such that i_* and j are comparable and $x_* \in X_{i_*j}(\Delta_{i_*j})$, test $\mathcal{T}_{\{i_*,j\}}$ as applied to observation ω^K accepts the hypothesis $H_{i_*j}(\Delta_{i_*j})$*

Then the $p^K_{\mathcal{A}_\nu(x_*)}$-probability of $\overline{\Omega}$ is at least $1-\epsilon$, and the aggregated solution $\widehat{x}(\omega^K)$ satisfies

$$\omega^K \in \overline{\Omega} \Rightarrow \|x_* - \widehat{x}(\omega^K)\| \le 3\|x_{i_*} - x_*\| + 2\overline{\Delta}_{i_*},$$

where

$$\overline{\Delta}_{i_*} = \begin{cases} 0, & \mathcal{I}_{i_*}(\omega^K) = \emptyset, \\ \max_{j \in \mathcal{I}_{i_*}(\omega^K)} \Delta_{i_*j}, & \mathcal{I}_{i_*}(\omega^K) \neq \emptyset. \end{cases}$$

(for notation, see the description of aggregation in Section 4.2).

5.2.3 Characterizing performance

Theorem 3. *In the setting described in Section 5.1, assume that $x_i \in X$, $1 \le i \le N$, and that for some positive integer $\overline{K}$, $\epsilon \in (0, 1/2)$ and $(\gamma, \delta) \ge 0$ there exists inference $\omega^{\overline{K}} \mapsto \bar{x}(\omega^{\overline{K}}) \in \mathbf{R}^n$ such that*

$$\mathrm{Prob}_{\omega^{\overline{K}} \sim p^{\overline{K}}_{\mathcal{A}_\nu(x)}} \left\{ \|x - \bar{x}(\omega^{\overline{K}})\| \le \gamma \|x - x_{i_*}\| + \delta \right\} \ge 1 - \epsilon \ \ \forall (\nu \le J,\, x \in X_\nu)$$

where x_{i_*} is one of the $\|\cdot\|$-closest to x point among $x_1, ..., x_N$. Now let

$$\gamma' > \gamma$$

and let K satisfy the relation

$$K \ge \left\lceil \frac{2\ln\left(J(N-1)/\epsilon\right)}{\ln\left([4\epsilon(1-\epsilon)]^{-1}\right)} \overline{K} \right\rceil. \tag{37}$$

Then, with K-repeated observations ω^K, all pairs $\{i, j\}$ with $r_{ij} > \delta$ are appropriate, and specifying these pairs as comparable, the aggregation procedure described in this section with properly selected $\widetilde{\delta}_{ij}$ ensures that the resulting aggregated estimate $\widehat{x}(\omega^K)$ satisfies

$$\mathrm{Prob}_{\omega^K \sim p^K_{\mathcal{A}_\nu(x)}} \left\{ \|\widehat{x}(\omega^K) - x\| > (3 + 2\gamma')\|x - x_{i_*}\| + 2\delta \right\} \le \epsilon \ \ \forall (\nu \in J,\, x \in X_\nu). \tag{38}$$

5.3 Aggregation in Euclidean seminorm

We now consider the special case of situation described in Section 5.1 where $\|\cdot\|$ is an Euclidean seminorm: $\|x\| = \|Bx\|_2$ where $B \in \mathbf{R}^{q\times n}$ is a given matrix. In this case, the sets $\mathcal{X}_{ij}(\delta)$ as defined in (35) are finite unions of convex compact sets, and we can apply the "near-optimal" inferring color machinery to build the tests required by aggregation scheme from Section 4.3.

We assume to be given the number of observations K along with tolerance parameters $\epsilon \in (0,1)$ and a "negligibly small" $\underline{\delta} > 0$ (say, 10^{-100}); we put $\overline{N} = \frac{1}{2}N(N-1)$.

5.3.1 Preliminaries

Given a pair $(i,j) \in \mathcal{O}$ and $\delta > 0$, it may happen that one or both of the sets $\mathcal{X}_{ij}(\delta)$ and $\mathcal{X}_{ji}(\delta)$ as defined in (35) is/are empty, in which case we qualify δ as (i,j)-good. Now let i,j,δ be such that both of the sets $\mathcal{X}_{ij}(\delta)$ and $\mathcal{X}_{ji}(\delta)$ are nonempty. In this case we build the collection $\mathcal{R}_{ij}(\delta) = \{R_{ij}^s(\delta) : 1 \le s \le J_{ij}^{\delta_{\mathrm{r}}}\}$ of nonempty convex compact "red" sets comprised of all nonempty sets of the form

$$R_{ij\nu}(\delta) = \{\mathcal{A}_\nu(x) : x \in X_\nu,\ \|x - x_i\| \le \|x - x_j\| - \delta\},\ 1 \le \nu \le J.$$

Similarly, we build the collection $\mathcal{B}_{ij}(\delta) = \{B_{ij}^s(\delta) : 1 \le s \le J_{ij}^{\delta_{\mathrm{b}}}\}$ of nonempty convex compact "blue" sets comprised of all nonempty sets of the form

$$B_{ij\nu}(\delta) = \{\mathcal{A}_\nu(x) : x \in X_\nu,\ \|x - x_j\| \le \|x - x_i\| - \delta\},\ 1 \le \nu \le J.$$

Applying to the collections $\mathcal{R}_{ij}$, $\mathcal{B}_{ij}$ the K-observation color inferring procedure from in Section 2.3, depending on δ it may happen that the resulting risk bound ϵ_K satisfies $\epsilon_K \le \epsilon/\overline{N}$. In this case we say that δ is (i,j)-*good*, and that it is (i,j)-bad otherwise.

Clearly, whenever δ is (i,j)-good, so is $\delta' \ge \delta$. Similarly to the case of general seminorm, given $(i,j) \in \mathcal{O}$ and $\delta > 0$, we can check efficiently whether δ is or is not (i,j)-good. Given $(i,j) \in \mathcal{O}$, large enough δ definitely are (i,j)-good, since the corresponding sets $\mathcal{X}_{ij}(\delta)$ are empty. Applying Bisection, we can rapidly find the value Δ_{ij} of δ such that Δ_{ij} is (i,j)-good, and either $\Delta_{ij} \le \underline{\delta}$, or $\Delta_{ij} - \underline{\delta}$ is not (i,j)-good.

Same as in the case of general seminorm, it is immediately seen that δ is (i,j)-good if and only if δ is (j,i)-good. As a result, we can select the above Δ_{ij} to be symmetric: $\Delta_{ij} = \Delta_{ji}$. Note that as a result, every pair $\{i,j\} \in \mathcal{U}$ is assigned threshold $\Delta_{ij} = \Delta_{ji}$ which is (i,j)-good. Besides this, we can equip this pair with K-observation test $\mathcal{T}_{\{i,j\}}$ deciding on the hypotheses $\mathcal{H}_{ij}(\Delta_{ij}) := H(\mathcal{X}_{ij}(\Delta_{ij}))$ and $\mathcal{H}_{ji}(\Delta_{ij}) := H(\mathcal{X}_{ji}(\Delta_{ij}))$, specifically, the test as follows:

— when both hypotheses are empty, the test accepts both hypotheses,

— when exactly one of the hypotheses is nonempty, the test accepts this nonempty hypothesis and rejects the empty one,

— when both hypotheses are nonempty, $\mathcal{T}_{\{i,j\}}$ is the above color inferring test associated with ((i,j)-good!) Δ_{ij}, so that it accepts exactly one of the hypotheses, and its risk does not exceed $\epsilon/\overline{N}$.

5.3.2 Aggregation routine

Aggregation routine is the procedure from Section 4.3 as applied to the K-repeated observation ω^K in the role of ξ and the just defined $\Delta_{ij} = \Delta_{ij}$, $\mathcal{T}_{\{i,j\}}$; as we have seen, these entities meet all the

requirements of the setup of Section 4.3. Denoting by $\widehat{i}(\omega^K) \in \{1, ..., N\}$ the output of our aggregation, the observation being ω^K, and applying Proposition 6, we arrive at the following result:

Proposition 8. *In the situation of this section, suppose that the just described aggregation routine is applied to observation ω^K stemming from $x_* \in X$. Then*

$$\mathrm{Prob}_{\omega^K \sim P}\left\{\|x_* - x_{\widehat{i}(\omega^K)}\| \le \|x_{i_*} - x_*\| + 2\overline{\Delta}\right\} \ge 1 - \epsilon \qquad \forall P \in \mathcal{P}^K_{x_*} \tag{39}$$

where x_{i_*} is one of the closest to x_* points among $x_1, ..., x_N$ and, same as in Proposition 6, $\overline{\Delta} = \max_{j \neq i} \Delta_{ij}$.

5.3.3 Characterizing performance

Theorem 4. *In the situation under consideration, assume that for some positive integer $\overline{K}$, $\epsilon \in (0, 1/2)$ and real $\bar{\delta} > 0$, for every pair $(i, j) \in \mathcal{O}$ there exists inference $\omega^{\overline{K}} \mapsto \iota_{ij}(\omega^{\overline{K}}) \in \{i, j\}$ such that for every $x_* \in X$ and $P \in \mathcal{P}^{\overline{K}}_{x_*}$ one has*

$$\mathrm{Prob}_{\omega^{\overline{K}} \sim P}\left\{\|x_* - x_{\iota_{ij}(\omega^{\overline{K}})}\| < \min[\|x_* - x_i\|, \|x_* - x_j\|] + \bar{\delta}\right\} \ge 1 - \epsilon. \tag{40}$$

Then whenever

$$K \ge \left\lceil \frac{2\ln\left(J\overline{N}/\epsilon\right)}{\ln\left([4\epsilon(1-\epsilon)]^{-1}\right)}\overline{K} \right\rceil \tag{41}$$

the aggregated estimate $x_{\widehat{i}(\omega^K)}$ yielded by the above aggregation procedure as applied to K-repeated observation ω^K for every $x_* \in X$ satisfies

$$\mathrm{Prob}_{\omega^K \sim P}\left\{\|x_{\widehat{i}(\omega^K)} - x_*\| \ge \|x_* - x_{i_*}\| + 2(\bar{\delta} + \underline{\delta})\right\} \le \epsilon \qquad \forall P \in \mathcal{P}^K_{x_*}, \tag{42}$$

x_{i_*} being one of the $\|\cdot\|$-closest to x_* point among $x_1, ..., x_N$.

5.4 Application: adaptive estimation over unions of convex sets

It is clear that just developed aggregation routines may be applied to the problem of adaptive estimation over unions of convex sets defined in Section 3.1. Our next objective is to discuss this application in more detail and derive corresponding accuracy bounds. From now on, notation and entities such as reliability tolerance ϵ, number $\overline{K}$ of pilot observations, pilot $\overline{K}$-observation estimates $\widetilde{x}_i(\omega^{\overline{K}})$, risks $\mathrm{Risk}^{\mathcal{J}}_{\epsilon,M}[\widehat{x}|Y]$, and upper bounds $\mathfrak{r}_j = \mathfrak{r}^{\overline{K}}_j(\epsilon)$ on $\mathrm{Risk}^{\{j\}}_{\epsilon,\overline{K}}[\widetilde{x}_j|X_j]$, are as defined in that section.

5.4.1 Estimation over unions using point aggregation

The quantities $J = N$, $X = \cup_j X_j$ and points $x_i = \widetilde{x}_i(\omega^{\overline{K}})$ taken together with the mappings $\mathcal{A}_j(\cdot)$ and the seminorm $\|\cdot\|$ form the data meeting the requirements of the setup of Section 5.1. Given (post-pilot) K-repeated observation ω^K with $\omega_k \sim p_{\mathcal{A}_{j_*}}(x_*)$, $k = 1, ..., K$, with $x_* \in X_{j_*}$, we can use the routines in Sections 5.2 and 5.3 to aggregate points x_i into an estimate $\widehat{x}$ of x_*.

Case of general seminorm. Let us start with the aggregation procedure described in Section 5.2. In our present setting its implementation is as follows. For $(i,j) \in \mathcal{O}$ we set

$$r_{ij} = \tfrac{1}{2}\|x_i - x_j\|,\ X_{ij}(\delta) = X \cap \{z : \|z - x_i\| \le r_{ij} - \delta\}, \tag{43}$$

and consider hypotheses $H_{ij}(\delta)$ and $H_{ji}(\delta)$ stating, respectively, that observations stem from a signal $x \in X_{ij}(\delta)$ and $x \in X_{ji}(\delta)$. Same as before, we say that δ is (i,j)-appropriate, if the risk of the K-observation test $\mathcal{T}_{\{i,j\}}$, yielded by the machinery from Section 2.3, deciding on $H_{ij}(\delta)$ vs $H_{ji}(\delta)$ does not exceed $\epsilon/(N-1)$. We define parameters Δ_{ij} and tests $\mathcal{T}_{\{i,j\}}$ as prescribed by the construction in Section 5.2 and utilize the resulting entities in the aggregation procedure from Section 4.2 thus arriving at the aggregated estimate $\widehat{x}^{(a)}(\omega^{\overline{K}+K}) = \widehat{x}(\omega^K)$ of x_*.

By Proposition 7, for all $j_* \le N$ and $x_* \in X_{j_*}$ aggregation $\widehat{x}(\omega^K)$ satisfies

$$\mathrm{Prob}_{\omega^K \sim p^K_{\mathcal{A}_{j_*}(x_*)}} \left\{ \|x_* - \widehat{x}(\omega^K)\| \le 3\|x_{i_*} - x_*\| + 2\overline{\Delta}_{i_*} \right\} \ge 1 - \epsilon \tag{44}$$

where x_{i_*} is one of the $\|\cdot\|$-closest to x_* points among $x_1, ..., x_N$, and $\overline{\Delta}_{i_*} \le \max_{j \ne i_*} \Delta_{i_* j}$ is defined in Proposition 7. Note that due to $\omega^{\overline{K}} \in \widetilde{\Omega}^{\overline{K}}$ we also have $\|x_* - x_{j_*}\| \le \mathfrak{r}_{j_*} = \mathfrak{r}^{\overline{K}}_{j_*}(\epsilon)$, which combines with (44) and $\|x_* - x_{i_*}\| \le \|x_* - x_{j_*}\|$ to imply that the x_*-probability for ω^K to satisfy

$$\|x_* - \widehat{x}(\omega^K)\| \le 3\mathfrak{r}_{j_*} + 2\overline{\Delta}_{i_*} \le 3\overline{\mathfrak{r}} + 2\overline{\Delta},\ \overline{\mathfrak{r}} = \max_i \mathfrak{r}_i,\ \overline{\Delta} = \max_{i \le N} \overline{\Delta}_i, \tag{45}$$

is at least $1 - \epsilon$.

Proposition 9. *In the situation described in Section 3.1, suppose that we are given a positive integer $\overline{K}$, tolerances $\epsilon \in (0, 1/2)$ and $\kappa > 0$, and K such that*

$$K \ge \left\lceil \frac{2\ln\left(N(N-1)/\epsilon\right)}{\ln\left([4\epsilon(1-\epsilon)]^{-1}\right)} \overline{K} \right\rceil. \tag{46}$$

Then estimate $\widehat{x}(\omega^K)$ yielded by the procedure described above with properly selected Δ_{ij} as applied to observation ω^K satisfies

$$\mathrm{Prob}_{\omega^K \sim p^K_{\mathcal{A}_{j_*}(x_*)}} \left\{ \|\widehat{x}(\omega^K) - x_*\| > 3 \max_i \mathfrak{r}^{\overline{K}}_i(\epsilon) + 2\mathrm{RiskOpt}^{\overline{1,N}}_{\epsilon,\overline{K}}[X] + \kappa \right\} \le \epsilon.$$

In particular, when the upper bounds $\mathfrak{r}^{\overline{K}}_i(\epsilon)$ on the risks $\mathrm{Risk}^{\{i\}}_{\epsilon,\overline{K}}[\widetilde{x}_i|X_i]$ of estimates $\widetilde{x}_i(\omega^{\overline{K}})$ are within factor θ of the respective $\overline{K}$-observation minimax risks, i.e.,

$$\mathrm{RiskOpt}^{\{i\}}_{\epsilon,\overline{K}}[X_i] \le \mathfrak{r}^{\overline{K}}_i(\epsilon) \le \theta \mathrm{RiskOpt}^{\{i\}}_{\epsilon,\overline{K}}[X_i]$$

the risk $\mathrm{Risk}^{\overline{1,N}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X]$ of the estimate $\widehat{x}^{(a)}(\omega^{\overline{K}+K}) = \widehat{x}(\omega^K)$ (as function of pilot observation $\omega^{\overline{K}}$ and independent observation ω^K) is within a moderate factor from the minimax $\overline{K}$-observation risk $\mathrm{RiskOpt}^{\overline{1,K}}_{\epsilon,\overline{K}}[X]$:

$$\mathrm{Risk}^{\overline{1,N}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X] \le [2 + 3\theta]\mathrm{RiskOpt}^{\overline{1,N}}_{\epsilon,\overline{K}}[X] + \kappa.$$

Case of Euclidean seminorm. When $\|\cdot\|$ is a Euclidean seminorm, we can utilize the aggregation procedure described in Section 5.3 to build the "two-stage" estimate $\widehat{x}^{(a)}(\omega^{\overline{K}+K}) = \widehat{x}(\omega^K)$. Specifically, given $\epsilon \in (0, \frac{1}{2})$, "negligibly small" $\underline{\delta} > 0$, and $\delta \geq 0$, consider sets

$$\mathcal{X}_{ij}(\delta) = \{z \in X : \|z - x_j\| \geq \delta + \|z - x_i\|\},\ \mathcal{X}_{ji}(\delta) = \{z \in X : \|z - x_i\| \geq \delta + \|z - x_j\|\}.$$

We apply the construction of Section 5.3 to compute for every $(i, j) \in \mathcal{O}$ (i, j)-good quantities $\Delta_{ij} = \Delta_{ji}$ such that either $\Delta_{ij} \leq \underline{\delta}$, or $\Delta_{ij} > \underline{\delta}$ and $\Delta_{ij} - \underline{\delta}$ is not (i, j)-good, and proceed as explained in that section, ending up with the aggregated estimate $\widehat{x}(\omega^K)$. Invoking Proposition 8, we have

$$\mathrm{Prob}_{\omega^K \sim p^K_{\mathcal{A}_{j_*}(x_*)}} \left\{\|x_* - \widehat{x}(\omega^K)\| \leq \|x_{i_*} - x_*\| + 2\overline{\Delta}\right\} \geq 1 - \epsilon,\ \overline{\Delta} = \max_{i,j} \Delta_{ij},$$

where x_{i_*} is one of the closest to x_* points among $x_1, x_2, ..., x_N$. We have the following analog of Proposition 9 in this case.

Proposition 10. *Let $\|\cdot\|$ be a Euclidean seminorm. In the situation described in Section 3.1, suppose that we are given a positive integer $\overline{K}$, tolerances $\epsilon \in (0, 1/2)$ and $\varkappa > 0$, and K satisfying*

$$K \geq \left\lceil \frac{2\ln\left(N^2(N-1)/(2\epsilon)\right)}{\ln\left([4\epsilon(1-\epsilon)]^{-1}\right)} \overline{K} \right\rceil. \tag{47}$$

Then estimate $\widehat{x}(\omega^K)$ yielded by the above procedure with properly selected parameters as applied to observation ω^K satisfies

$$\mathrm{Prob}_{\omega^K \sim p^K_{\mathcal{A}_{j_*}(x_*)}} \left\{\|\widehat{x}(\omega^K) - x_*\| > \max_i \mathfrak{r}_i^{\overline{K}}(\epsilon) + 4\mathrm{RiskOpt}^{\overline{1,N}}_{\epsilon,\overline{K}}[X] + \varkappa\right\} \leq \epsilon.$$

In particular, when the upper bounds $\mathfrak{r}_i^{\overline{K}}(\epsilon)$ on the partial risks $\mathrm{Risk}^{\{i\}}_{\epsilon,\overline{K}}[\widetilde{x}_i|X_i]$ of $\overline{K}$-observation estimates $\widetilde{x}_i(\cdot)$ are within a factor θ of the respective $\overline{K}$-observation minimax risks, i.e.,

$$\mathrm{RiskOpt}^{\{i\}}_{\epsilon,\overline{K}}[X_i] \leq \mathfrak{r}_i^{\overline{K}}(\epsilon) \leq \theta \mathrm{RiskOpt}^{\{i\}}_{\epsilon,\overline{K}}[X_i],$$

the maximal risk $\mathrm{Risk}^{\overline{1,N}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X]$ of aggregated estimate $\widehat{x}^{(a)}(\omega^{\overline{K}+K}) := \widehat{x}(\omega^K)$ (considered as function of the pilot observation $\omega^{\overline{K}}$ and independent observation ω^K) is within a moderate factor from the minimax $\overline{K}$-observation risk $\mathrm{RiskOpt}^{\overline{1,N}}_{\epsilon,\overline{K}}[X]$:

$$\mathrm{Risk}^{\overline{1,N}}_{2\epsilon,K+\overline{K}}[\widehat{x}|X] \leq [4 + \theta]\mathrm{RiskOpt}^{\overline{1,N}}_{\epsilon,\overline{K}}[X] + \varkappa.$$

6 Adaptive estimation over unions of ellitopes

6.1 Ellitopic setup

Ellitopes, as introduced in [25, 27], are symmetric w.r.t. the origin convex and compact sets. In this section we consider the special case of the estimation problem described in Section 3.1 in which

1. observation scheme is Gaussian, i.e., observations $\omega_k \in \mathbf{R}^m$ stemming from (j, x), $x \in X_j$, are normal, $\omega_k \sim \mathcal{N}(\mathcal{A}_j(x), I_m)$ where $\mathcal{A}_j(x)$ are *linear*, rather than affine, mappings: $\mathcal{A}_j(x) = A_j x$, where $A_j \in \mathbf{R}^{m\times n}$, $j = 1, ..., N$, are given matrices;

2. sets X_j $j = 1, ..., N$, are *basic ellitopes*:

$$X_j = \{x \in \mathbf{R}^n : \exists r \in \mathcal{R}_j : x^T R_{j\tau} x \le r_\ell, \tau \le L\}, \quad j = 1, ..., N,$$

3. seminorm $\|\cdot\|$ is of the form $\|x\| = \pi(Bx)$ where B is a $q \times n$ matrix and the unit ball $\mathcal{B}_*$ of the conjugate to $\pi(\cdot)$ norm $\pi_*(\cdot)$ is an ellitope

$$\mathcal{B}_* = \{y \in \mathbf{R}^q : y \in \mathbf{R}^q : \exists s \in \mathcal{S} : x^T S_\tau x \le r_\tau, \tau \le L'\}.$$

   Here

   - $\mathcal{R}_j \subset \mathbf{R}^L_+$, $j = 1, ..., N$, and $\mathcal{S} \subset \mathbf{R}^{L'}_+$ are computationally tractable convex compact sets intersecting with $\operatorname{int} \mathbf{R}^L_+$ which are *monotone*.[12]
   - $R_{j\tau}$, $1 \le j \le N$, $1 \le \tau \le L$, are $n \times n$ matrices with $R_{j\tau} \succeq 0$ and $\sum_\tau R_{j\tau} \succ 0$; S_τ are $q \times q$ matrices such that $S_\tau \succeq 0$ and $\sum_\tau S_\tau \succ 0$.

   We refer to L and L' as *sizes* of corresponding ellitopes.

Particular choices of sets X_j and seminorm $\|\cdot\|$ encompass a variety of situations.

- When $L = 1$, $\mathcal{R}_j = [0, 1]$ and $R_1^{(j)} \succ 0$, X_j is an ellipsoid.
- When $L \ge 1$, $\mathcal{R}_j = [0, 1]^L$, X_j is an intersection of ellipsoids and elliptic cylinders centered at the origin, $\bigcap_{\tau \le L} \{z : z^T R_{j\tau} z \le 1\}$.
- When $U = [u_1, ..., u_L] \in \mathbf{R}^{n \times L}$, $\mathrm{Rank}[U] = n$, $\mathcal{R}_j = [0, 1]^L$, and $R_{j\tau} = u_\tau u_\tau^T$, X_j is a symmetric w.r.t. the origin polytope $\{z : \|U^T z\|_\infty \le 1\}$.
- When for $p \ge 2$, $\mathcal{S} = \left\{s \in \mathbf{R}^{L'}_+ : \sum_\tau [s]_\tau^{p/2} \le 1\right\}$ and as, in the previous example $U = [u_1, ..., u_{L'}] \in \mathbf{R}^{q \times L'}$, $\mathrm{Rank}[U] = q$, and $S_\tau = u_\tau u_\tau^T$, $\mathcal{B}_*$ is the set $\{y : \|U^T y\|_p \le 1\}$ and the seminorm $\|\cdot\|$ is $\|w\| = \|UBw\|_{p/(p-1)}$.

The family of ellitopes admits simple and fully algorithmic "calculus" demonstrating that this family is closed w.r.t. nearly all operations preserving convexity and symmetry w.r.t. the origin (e.g., taking finite intersections, direct products, linear images, and inverse images under linear embeddings; for details, see [27, Section 4.6]).

We are about to show that in the present situation, *estimates yielded by the approach described in Section 3.2 are nearly optimal in the minimax sense.*[13] Moreover, in this case we are able to provide "reasonably good" bounding of minimax risks of recovery over pairwise unions $X_i \cup X_j$ of ellitopes implying that tight bounds for the minimax risk of estimation over $X = \cup_{i=1}^N X_i$ can be efficiently computed.

6.2 Near-optimality of the aggregated estimate

Let $\overline{K}$ and K be positive integers, and let us assume that in the situation described in Section 6.1 we are given $\epsilon \in (0, 1/8)$ and $M = \overline{K} + K \ge 2$ independent observations ω_k stemming from unknown pair (ℓ_*, x_*), $x_* \in X_{\ell_*}$, $1 \le \ell_* \le N$. To build an M-observation estimate $\widehat{x}^{(a)}(\omega^M)$ of x_* we proceed as explained in Section 3.2:

- we split the observation sample into two observations: a $\overline{K}$-repeated observation $(\omega_1, ..., \omega_{\overline{K}})$ (preliminary observation) and $\omega^K = (\omega_{\overline{K}+1}, ..., \omega_{\overline{K}+K})$ (secondary observation).

[12] Here monotonicity of $V \subset \mathbf{R}^k_+$ means that if $0 \le v' \le v$ and $v \in V$ then also $v' \in V$.

[13] An analog of the results below in the special case where $\|\cdot\|$ is a Euclidean seminorm can be obtained by applying construction of Section 3.3.

- Preliminary observation is averaged to build observation

$$\widehat{\omega} = \frac{1}{\overline{K}} \sum_{k=1}^{\overline{K}} \omega_k$$

  which is then used to compute N *polyhedral* estimates $\widetilde{x}_i(\widehat{\omega})$ following the recipe in [28] and [27, Section 5.1.5].
- Finally, we apply the aggregation routine from Section 3.2 to assemble points $x_i = \widetilde{x}_i(\widehat{\omega})$ into estimate $\widehat{x}(\omega^K)$ obtaining as a result adaptive estimate $\widehat{x}^{(a)}(\omega^M) = \widehat{x}(\omega^K)$ of x_*.

Recall (cf. [27, Proposition 5.10]) that polyhedral estimates $\widetilde{x}_i$ satisfy $\widetilde{x}_i(\widehat{\omega}) \in X_i$ and

$$\mathrm{Risk}^{\{i\}}_{\epsilon,1}[\widetilde{x}_i|X_i] \le \mathfrak{r}_i(\epsilon) \le \mathfrak{C}_1 \ln(L+L')\sqrt{\ln[m/\epsilon]}\,\overline{\mathrm{RiskOpt}}^{\{i\}}_{\frac{1}{8},1}[X_i] \tag{48}$$

where bound $\mathfrak{r}_i(\epsilon)$ for maximal risk of estimation under the ith observation model is efficiently computable and $\overline{\mathrm{RiskOpt}}^{\{i\}}_{\epsilon,1}[X_i]$ is the minimax ϵ-risk of recovering $x \in X_i$ from single "averaged" observation $\widehat{\omega}$ stemming from i.i.d. $\omega_k \sim \mathcal{N}(A_j x, I_m)$; From now on, $\mathfrak{C}_i$ stand for appropriate *absolute* constants.

Given $(i,j) \in \mathcal{O}$, positive integer K, and $\delta \in (0,1/2)$ we (re-)define the notion of δ-separation risk (cf. (16)) in the present situation according to

$$\mathfrak{g}^K_{ij}(\delta) = \tfrac{1}{2} \max_{x \in X_i, y \in X_j} \left\{ \|x-y\| : \|\mathcal{A}_i(x) - \mathcal{A}_j(y)\|_2 \le \frac{2}{\sqrt{K}} q_{\mathcal{N}}(1-\delta) \right\} \tag{49}$$

where $q_{\mathcal{N}}(p)$ is the p-quantile of the standard normal distribution: $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{q_{\mathcal{N}}(p)} \mathrm{e}^{-s^2/2} ds = p$. Note that (49) is feasible, and therefore solvable, due to $0 \in X_j$ and $\mathcal{A}_j(0) = 0$ for all j.

For the sake of simplicity, from now on we restrict ourselves to the case

$$K \le \overline{K} \tag{50}$$

The next statement provides a refined version of results of Section 3.2 in the present setting:

Theorem 5. *In the situation of this section, assuming (50) and $0 < \epsilon < 1/16$, the just built estimate $\widehat{x}^{(a)}$ (as function of pilot observation $\omega^{\overline{K}}$ and secondary observation ω^K) satisfies*

$$\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X_i] \le 2\mathfrak{r}_i(\epsilon) + \max_{j<i} \left[\mathfrak{r}_j(\epsilon) + 2\mathfrak{g}^K_{ij}(\varepsilon)\right] \quad \forall i \le N, \tag{51}$$

with $\varepsilon = \frac{\epsilon}{N-1}$. Moreover, setting

$$\overline{\vartheta} := \frac{\sqrt{K} q_{\mathcal{N}}(1-\epsilon)}{\sqrt{\overline{K}} q_{\mathcal{N}}(1-\varepsilon)}$$

one has $\overline{\vartheta} \le 1$ and

$$\mathrm{Risk}^{\{i\}}_{2\epsilon,\overline{K}+K}[\widehat{x}^{(a)}|X_i] \le 2\mathfrak{r}_i(\epsilon) + \max_{j<i} \left[\mathfrak{r}_j(\epsilon) + 2\overline{\vartheta}^{-1} \mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j]\right] \quad \forall i \le N, \tag{52}$$

whence, in particular,

$$\mathrm{Risk}^{\{i\}}_{2\epsilon,2\overline{K}}[\widehat{x}^{(a)}|X_i] \le \max_{1 \le j \le i} \left[3\mathfrak{r}_j(\epsilon) + 2\frac{q_{\mathcal{N}}(1-\varepsilon)}{q_{\mathcal{N}}(1-\epsilon)} \mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j]\right] \qquad \forall i \le N.$$

Besides this, one has

$$\mathrm{Risk}^{\{i\}}_{2\epsilon,2\overline{K}}[\widehat{x}^{(a)}|X_i] \le \mathfrak{C}_2\Big(\ln(L+L')\sqrt{\ln[m/\epsilon]}+\sqrt{\ln[N/\epsilon]}\Big)\max_{1\le j\le i}\mathrm{RiskOpt}^{\{i,j\}}_{\frac{1}{16},\overline{K}}[X_i\cup X_j] \qquad \forall i\le N,$$

so that

$$\mathrm{Risk}^{\overline{1,N}}_{2\epsilon,2\overline{K}}[\widehat{x}^{(a)}|X] \le \mathfrak{C}_3\Big(\ln(L+L')\sqrt{\ln[m/\epsilon]}+\sqrt{\ln[N/\epsilon]}\Big)\mathrm{RiskOpt}^{\overline{1,N}}_{\frac{1}{16},\overline{K}}[X]. \tag{53}$$

6.3 Bounding the maximal risk of estimation

Our current objective is to provide efficient bounding for separation risks $\mathfrak{g}^K_{ij}(\varepsilon)$; taken together with bounds $\mathfrak{r}_j(\epsilon)$ for partial risks this would allow to bound the minimax risk of estimation over X. Under the premise of Theorem 5, let (i,j), $1\le i,j\le N$, and $K\le\overline{K}$ be fixed, let, same as in (51), $\varepsilon=\epsilon/(N-1)$, and let $\delta=2K^{-1/2}q_{\mathcal{N}}(1-\varepsilon)$. Observe that

$$\begin{aligned}
\mathfrak{g}^K_{ij}(\varepsilon) &= \tfrac{1}{2}\max_{x\in X_i,y\in X_j}\{\|x-y\|:\ \|A_ix-A_jy\|_2\le\delta\}\\
&= \tfrac{1}{2}\max_{x\in X_i,y\in X_j}\left\{\pi(B(x-y)):\ \|A_ix-A_jy\|_2^2\le\delta^2\right\}\\
&= \tfrac{1}{2}\max_{[x;y;u]}\left\{u^TB(x-y):\ u\in\mathcal{B}_*,\ x\in X_i,\ y\in X_j,\ \|A_ix-A_jy\|_2^2\le\delta^2\right\}.
\end{aligned}\tag{54}$$

Because the direct product of ellitopes $\mathcal{B}_*\times X_i\times X_j$ is an ellitope of the size not exceeding $L'+2L$ (cf. [27, Section 4.6]), when writing $u^TB(x-y)=[u;x;y]^T\overline{B}[u;x;y]$ and $\|A_ix-A_jy\|_2^2=[u;x;y]^TQ_{ij}[u;x;y]$ with

$$\overline{B}=\left[\begin{array}{c|c|c}0_{q\times q}&\frac{1}{2}B&-\frac{1}{2}B\\\hline\frac{1}{2}B^T&0_{n\times n}&0_{n\times n}\\\hline-\frac{1}{2}B^T&0_{n\times n}&0_{n\times n}\end{array}\right],\quad Q_{ij}=\left[\begin{array}{c|c|c}0_{q\times q}&0_{q\times n}&0_{q\times n}\\\hline0_{n\times q}&A_i^TA_i&-A_i^TA_j\\\hline0_{n\times q}&-A_j^TA_i&A_j^TA_j\end{array}\right]$$

we conclude that the quantity $\mathfrak{g}^K_{ij}(\varepsilon)$ is the maximum of a homogeneous quadratic form over an ellitope of size at most $\overline{D}=L'+2L+1$. Therefore, it can be upper-bounded by an efficiently computable quantity $\overline{\mathfrak{g}}^K_{ij}(\varepsilon)$ within factor $2\ln\overline{D}+2\sqrt{\ln\overline{D}}+1$ (see, e.g., [25, Proposition 3.3]) using semidefinite relaxation.

As a result, given a pair (i,j), $i\ne j$, we can upper-bound the 2ϵ-minimax risk $\mathrm{RiskOpt}^{\{i,j\}}_{2\epsilon,2\overline{K}}[X_i\cup X_j]$ with efficiently computable quantity

$$\overline{\mathfrak{r}}_{ij}(\epsilon)=3\max[\mathfrak{r}_i(\epsilon),\mathfrak{r}_j(\epsilon)]+2\overline{\mathfrak{g}}^{\overline{K}}_{ij}(\varepsilon)$$

such that

$$\begin{aligned}
\overline{\mathfrak{r}}_{ij}(\epsilon) &\le \mathfrak{C}_6\ln[\overline{D}]\sqrt{\ln[m/\epsilon]}\max\left(\mathrm{RiskOpt}^{\{i\}}_{\frac{1}{16},\overline{K}}[X_i],\mathrm{RiskOpt}^{\{j\}}_{\frac{1}{16},\overline{K}}[X_j]\right)+\mathfrak{C}_7\ln[\overline{D}]\overline{\mathfrak{g}}^{\overline{K}}_{ij}(\varepsilon)\\
&\le \mathfrak{C}_8\ln[\overline{D}][\sqrt{\ln[m/\epsilon]}+\sqrt{\ln[N/\epsilon]}]\mathrm{RiskOpt}^{\{i,j\}}_{\frac{1}{16},\overline{K}}[X_i\cup X_j]
\end{aligned}$$

(we have used (64)). Similarly, 2ϵ-minimax risk $\mathrm{RiskOpt}^{\overline{1,N}}_{2\epsilon,2\overline{K}}[X]$ of estimation over X can be bounded with efficiently computable quantity

$$\overline{\mathfrak{r}}(\epsilon)=\max_{i,j\le N}\left[3\mathfrak{r}_i(\epsilon)+2\overline{\mathfrak{g}}^{\overline{K}}_{ij}(\varepsilon)\right]$$

such that

$$\begin{aligned}\overline{\mathfrak{r}}(\epsilon) &\leq \mathfrak{C}_9 \max_{i,j\leq N}\left[\ln[\overline{D}]\sqrt{\ln[m/\epsilon]}\mathrm{RiskOpt}^{\{i\}}_{\frac{1}{16},\overline{K}}[X_i] + \mathfrak{C}_{10}\ln[\overline{D}]\mathfrak{g}^{\overline{K}}_{ij}(\varepsilon)\right]\\ &\leq \mathfrak{C}_{11}\ln[\overline{D}]\left(\sqrt{\ln[m/\epsilon]}+\sqrt{\ln[N/\epsilon]}\right)\mathrm{RiskOpt}^{\overline{1,N}}_{\frac{1}{16},\overline{K}}[X].\end{aligned}$$

6.4 Numerical illustration: application to estimation in the single-index model

In this section, we apply the proposed adaptive estimate to a toy problem of estimation in the simple single index model in which

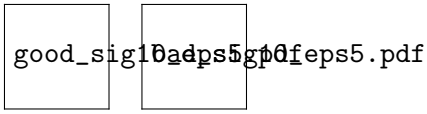
- "Unknown signal" x is a vector of coefficients of one-dimensional spline $s(t)$ on $[-1,1]$ split into 10 equal segments. In each segment, s is a quadratic polynomial, and its derivative $s'(t)$ is continuous on the entire $[-1,1]$, making the number of degrees of freedom in the spline—dimension of the parameter vector x—equal to 12. Signal vector x is restricted to have $\|\cdot\|_2$-norm not exceeding 1, thus, the signal set X is the unit Euclidean ball in $\mathbf{R}^{12}$.
- We consider the situation in which all signal sets X_j, $j=1,...,N=64$, are equal to X, but there are N different encodings $\mathcal{A}_j(\cdot)=A_j\in\mathbf{R}^{1024\times 12}$ built as follows: for $j=1,...,J=64$, we specify e_j as unit vector in $\mathbf{R}^2$ at angle $2\pi(j-1)/N$ with the first basis vector. Specifying Γ as a set of 1024 points sampled from a uniform distribution on $\{u\in\mathbf{R}^2:\|u\|_\infty\leq 1\}$, A_jx is the restriction onto Γ of the function $f_{j,x}(u)=s(e_j^Tu)$.
  Note: for $u\in\Gamma$, e_j^Tu can be outside of $[-1,1]$, and when defining $s(e_j^Tu)$, we extend s from $[-1,1]$ onto the entire real axis in such a way that the extended function is continuously differentiable and is affine to the left of -1 and to the right of 1.
- Observations A_jx are corrupted by white Gaussian noise $\xi\sim\mathcal{N}(0,\sigma^2I)$.
- We deal with $\overline{K}=K=1$ and split our actual observation into two independent unbiased Gaussian observations—pilot $\widetilde{\omega}$ and secondary ω—of variance $2\sigma^2$ each.

It is worth mentioning that the considered situation differs from the "classical" setting of the single index estimation problem: here our objective is neither to estimate the index—unit vector e corresponding to the "orientation" in $\mathbf{R}^2$ of the univariate function underlying observations, nor to estimate the bivariate regression function $f_{i,x}(\cdot)$,[14] but to recover vector x of spline coefficients of $s(\cdot)$, the norm $\|\cdot\|$ being the Euclidean norm. As such, the problem we consider is that of recovery from noisy indirect observations, the latter being equivalent to estimating univariate function $s(\cdot)$, estimation error being measured in the L_2-norm on $[-1,1]$. We consider two implementations of the recovery procedure; in both implementations we utilize polyhedral estimate of [28] to build pilot estimates $\widetilde{x}_i(\widetilde{\omega})$, $i=1,...,N$. The first recovery, we denote it $\widehat{x}^{(I)}$, utilizes the aggregated estimate described in Sections 5.3, 5.4; $\widehat{x}^{(II)}$ is the adaptive estimate of Section 3.3; finally, estimate $\widehat{x}^{(III)}$ is the slightly modified adaptive estimate of Section 3.2 in which, when the set $\mathcal{I}(\omega)$ of admissible estimates contains more than 1 point, instead of selecting the admissible estimate with the smallest index i, adaptive estimate $\widehat{x}$ is obtained by aggregating admissible points $\widetilde{x}_i$, $i\in\mathcal{I}(\omega)$, as the optimal solution to the optimization problem

$$\widehat{x}=\mathop{\mathrm{argmin}}_u\max_{i\in\mathcal{I}(\omega)}\|u-\widetilde{x}_i\|_2.$$

To see how the error of recovery depends on the noise variance σ^2, for each value of the variance we sample $K=100$ realizations of the signal x_k from the uniform distribution on the unit sphere

[14] For "state of art" adaptive estimates of regression function f in a general d-dimensional single index model under $L_2([-1,1]^d)$-losses see, e.g., [22]; see also [32] for adaptation w.r.t pointwise and general L_p-risks.



(a) (b)

Figure 1: Typical graphs of the true function $s(\cdot)$ (solid line) and its recoveries utilizing estimate $\widehat{x}^{(I)}$ (dotted line), estimate $\widehat{x}^{(II)}$ (dash-dot line), and estimate $\widehat{x}^{(III)}$ (dashed line).

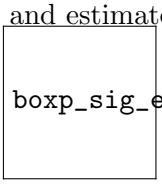


Figure 2: Error distribution of recoveries $\widehat{x}^{(I)}$, $\widehat{x}^{(II)}$, and $\widehat{x}^{(III)}$ for different values of noise variance σ^2: from left to right, box plots for $\sigma = 0.1$, $\sigma = 0.05$, and $\sigma = 0.02$.

along with directions e_k from the uniform distribution on the unit circle. Results of these experiments are presented in Figure 2 (note the logarithmic scale of the y-axis); the red bar over each box plot represents the upper bound $\max_j \mathfrak{r}_j^1(\epsilon)$ of partial ϵ-risks of preliminary estimates $\widetilde{x}_j(\widetilde{\omega})$.

The reliability parameter of the recoveries being set to 95% (i.e., $\epsilon = 0.05$), upper bounds $\mathfrak{r}_i^1(\varepsilon)$ exceed 1 for $\sigma > 0.15$. We present in Figure 1, for $\sigma = 0.1$, typical graphs of the true signal $s(\cdot)$ and its recoveries.

- Plot (a): set $\mathcal{I}(\omega)$ of admissible estimates for recoveries $\widehat{x}^{(II)}$ and $\widehat{x}^{(III)}$ is a singleton (in this case, $\|x_* - \widehat{x}^{(I)}\|_2 = 0.0949$, $\|x_* - \widehat{x}^{(II)}\|_2 = 0.0616$, and $\|x_* - \widehat{x}^{(III)}\|_2 = 0.0710$).
- Plot (b): cardinality $|\mathcal{I}(\omega)| = 3$ of the set of admissible estimates for recovery $\widehat{x}^{(III)}$, $\mathcal{I}(\omega)$ is a singleton for recovery $\widehat{x}^{(II)}$ (in this case, $\|x_* - \widehat{x}^{(I)}\|_2 = 0.0752$, $\|x_* - \widehat{x}^{(II)}\|_2 = 0.0846$, and $\|x_* - \widehat{x}^{(III)}\|_2 = 0.1508$).

For $\sigma \leq 0.05$ in all simulations the set $\mathcal{I}(\omega)$ of admissible estimates was a singleton for all recoveries. Moreover, in these simulations, selected indices j of encodings A_j were the same for both recoveries, corresponding to the closest to the "true direction e" element e_j of the "grid of directions." When $\sigma = 0.1$, corresponding admissible sets for recovery $\widehat{x}^{(I)}$ were singletons, with corresponding direction being the closest to true e in 93/100 simulations (and second close in remaining 7/100); admissible set for recovery $\widehat{x}^{(II)}$ was a singleton corresponding to the closest direction in 56/100 experiments, in the remaining 44/100 the admissible set contained two closest to e directions. "Population" of admissible sets of recovery $\widehat{x}^{(III)}$ is represented in Figure 3; admissible i's obtained in each simulation are "centered" w.r.t. the "index" $j_e = \frac{N\theta}{2\pi} + 1$ where θ is the angle between random vector e underlying the observation and the first basis vector of $\mathbf{R}^2$. For $\sigma = 0.1$ we also present in Figure 4 typical plot of the bound $\overline{\mathfrak{g}}_{ij}^1(\varepsilon)$, $\varepsilon = \epsilon/(N-1)$, $i = 33$, for separation risk $\mathfrak{g}_{ij}^1(\varepsilon)$ along with the lower bound computed by Monte-Carlo simulations.

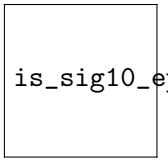


Figure 3: Admissible sets $\mathcal{I}(\omega)$ of recovery $\widehat{x}^{(III)}$ (blue dots), those of recovery $\widehat{x}^{(II)}$ (green crosses), and selected $\widehat{i}$'s of recovery $\widehat{x}^{(I)}$ (solid red line).

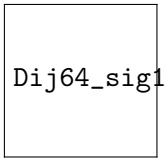


Figure 4: Bounding $\mathfrak{g}^1_{ij}(\varepsilon)$: solid line—upper bound $\overline{\mathfrak{g}}^1_{ij}(\varepsilon)$ by semidefinite relaxation, dash-dot line—lower bound by Monte-Carlo simulations.

A Proofs

A.1 Proofs for Section 3

A.1.1 Proof of Proposition 3

The fact that the (ℓ_*, x_*)-probability of $\overline{\Omega}^K$ is at least $1-\epsilon$ is readily given by the union bound and the fact that for good pairs (ℓ_*, j) the (ℓ_*, x_*)-probability for test $\mathcal{T}_{\{\ell_*,j\}}$ to accept H_{ℓ_*} is at least $1-\epsilon/(N-1)$. Furthermore, because the preliminary observation belongs to $\widetilde{\Omega}^{\overline{K}}$ we have

$$\|x_* - x_{\ell_*}\| \le \mathfrak{r}_{\ell_*}. \tag{55}$$

Let now $\omega^K \in \overline{\Omega}^K$ be fixed; then the set $\mathcal{I} = \mathcal{I}(\omega^K)$ is not empty because $\ell_* \in \mathcal{I}$; indeed, when $\omega^K \in \overline{\Omega}^K$ "true" hypothesis H_{ℓ_*} is never rejected. Consequently, if $i' \in \mathcal{I}$ differs from ℓ_*, then (i', ℓ_*) is bad, since otherwise $i' \in \mathcal{J}_{\ell_*}$ and the test $\mathcal{T}_{\{i',\ell_*\}}$ would reject the true hypothesis H_{ℓ_*} (otherwise $i' \in \mathcal{I}$ would be impossible), which contradicts $\omega^K \in \overline{\Omega}^K$. As a byproduct of the just made observation, $\mathcal{I}^-_{\ell_*} \subset J^-_{\ell_*}$. Now, since we are in the case of $\ell_* \in \mathcal{I}$, either $\widehat{\imath} = \ell_*$, or $\mathcal{I} \ni \widehat{\imath} < \ell_*$. In the first case, (13) is evident, in the second $\widehat{\imath} \in \mathcal{I}^-_{\ell_*}$, and therefore (13) holds true as well. (13) is proved. Next, the first inequality in (14) is trivially true due to already proved inclusion $\mathcal{I}^-_{\ell_*} \subset J^-_{\ell_*}$; the second inequality is evident from the definitions of $\mathfrak{r}_i$'s and δ_{ij}'s (recall that we have assumed all B_i to be nonempty). Finally, (15) is an immediate consequence of inclusions $\mathcal{I}^-_{\ell_*} \subset \mathcal{I}\backslash\{\ell_*\} \subset J_{\ell_*} \subset \overline{J}$ (the first and the third are evident, the second has been proved) and the definition of $\mathfrak{r}$'s and δ_{ij}'s. □

A.1.2 Proof of Theorem 1

1^o. Let $\varepsilon = \epsilon/(N-1)$. Consider a pair $(i,j) \in \mathcal{O}$ which is bad. In this case, one has $\delta_{ij} \le \mathfrak{r}^K_{ij}(\varepsilon)$. Indeed, consider optimization problem

$$\max_{x \in B_i, y \in B_j} \varrho(\mathcal{A}_i(x), \mathcal{A}_j(y)). \tag{56}$$

Observe that problem (56) is solvable, and its optimal solution $x' \in B_i$, $y' \in B_j$ satisfies $\|x'-y'\| \ge 2\delta_{ij}$. On the other hand, the optimal value $\bar\rho$ of (56) is greater than $\varepsilon^{1/K}$ because, otherwise, the risk of a K-observation test $\mathcal{T}_{\{i,j\}}$ deciding on hypotheses H_i and H_j, as discussed in Section 2.2, would be bounded by $\bar\rho^K = \varepsilon$, implying that pair (i,j) is good what is not the case. We conclude that $\mathfrak{r}^K_{ij}(\varepsilon)$, as defined in (16), satisfies $\mathfrak{r}^K_{ij}(\varepsilon) \ge \frac{1}{2}\|x'-y'\| \ge \delta_{ij}$. Combined with (55) and the bounds (13) and (14), the latter relation implies (17).

2^o. To show (18) we need the following statement.

Lemma 1. *Given (i,j) with $1 \le j < i \le N$, let $\bar{\varrho}^*_{ij} = \mathrm{RiskOpt}^{\{i,j\}}_{\epsilon,\overline{K}}[X_i \cup X_j]$ be the minimax $\overline{K}$-observation ϵ-risk of estimation over $X_i \cup X_j$. Suppose that $\varepsilon \in (0,1/2)$ is such that*

$$\widetilde{\vartheta} = \frac{\ln\left(4\epsilon(1-\epsilon)\right)}{2\ln(\varepsilon)} \le 1$$

(i) *Assume that $K > \widetilde{\vartheta}^{-1}\overline{K}$. Then $\mathfrak{r}^K_{ij}(\varepsilon) \le \bar{\varrho}^*_{ij}$.*

(ii) *In addition, if $\mathcal{A}_i(x_{ij}) = \mathcal{A}_j(x_{ij})$ for some $x_{ij} \in X_i \cap X_j$, one has*

$$\mathfrak{r}^K_{ij}(\varepsilon) \le \widetilde{\vartheta}^{-1}\bar{\varrho}^*_{ij}$$

whenever $K \ge \overline{K}$.

Proof. When problem (16) is infeasible, we have $\mathfrak{r}^K_{ij}(\varepsilon) = 0$, and the claims in Lemma are trivially true. Now let (16) be feasible. Then the problem is solvable; let $(\bar{x}, \bar{y})$, $\bar{x} \in X_i, \bar{y} \in X_j$, be an optimal solution. Suppose that $\bar{\varrho}^*_{ij} < \frac{1}{2}\|\bar{x} - \bar{y}\|$. This would imply existence of a $\overline{K}$-observation estimate $\tilde{x}(\cdot)$ with maximal ϵ-risk over $X_i \cup X_j$ which is smaller than $\frac{1}{2}\|\bar{x} - \bar{y}\|$, meaning that there is a simple $\overline{K}$-observation test deciding on hypothesis $H_{\bar{x}}$: "observation $\omega^{\overline{K}}$ stems from $(i,\bar{x})$" against $H_{\bar{y}}$: "observation $\omega^{\overline{K}}$ stems from $(j,\bar{y})$" with risk bounded with ϵ, namely, the test which accepts $H_{\bar{x}}$ whenever $\|\tilde{x} - \bar{x}\| \le \|\tilde{x} - \bar{y}\|$ and accepts $H_{\bar{y}}$ otherwise. By what we know about testing in simple observation schemes, this means that Hellinger affinity $\varrho(\mathcal{A}_i(\bar{x}), \mathcal{A}_j(\bar{y}))$ between the corresponding distributions of observations satisfies (cf. (6)) $\varrho(\mathcal{A}_i(\bar{x}), \mathcal{A}_j(\bar{y})) \le [4\epsilon(1-\epsilon)]^{1/(2\overline{K})} < \varepsilon^{1/K}$ contradicting the fact that, by construction of $\bar{x}$ and $\bar{y}$, $\varrho(\mathcal{A}_i(\bar{x}), \mathcal{A}_j(\bar{y}) \ge \varepsilon^{1/K}$. (i) is proved.

Next, to prove (ii), for $\vartheta \in [0,1]$, let $x(\vartheta) = x_{ij} + \vartheta(\bar{x} - x_{ij})$ and $y(\vartheta) = x_{ij} + \vartheta(\bar{y} - x_{ij})$; observe that $\tilde{\rho}(\vartheta) := \ln \varrho(\mathcal{A}_i(x(\vartheta)), \mathcal{A}_j(y(\vartheta)))$ is a concave function of ϑ with $\tilde{\rho}(1) \ge K^{-1}\ln\varepsilon$ and $\tilde{\rho}(0) = 0$. Thus, for any $\vartheta < \widetilde{\vartheta}$,

$$\tilde{\rho}(\vartheta) \ge \vartheta K^{-1}\ln\varepsilon > \widetilde{\vartheta}K^{-1}\ln\varepsilon \ge \tfrac{1}{2}K^{-1}\ln[4\epsilon(1-\epsilon)].$$

As we already know, this means that there is no K-observation test capable of deciding between hypotheses $H_{x(\vartheta)}$ and $H_{y(\vartheta)}$ with risk bounded with ϵ, implying in its turn that

$$\bar{\varrho}^*_{ij} > \tfrac{1}{2}\|x(\vartheta) - y(\vartheta)\| = \tfrac{1}{2}\vartheta\|\bar{x} - \bar{y}\| = \vartheta\mathfrak{r}^K_{ij}(\varepsilon). \qquad \square$$

Setting $\varepsilon = \epsilon/(N-1)$ (which results in $\widetilde{\vartheta} = \overline{\vartheta}$) and substituting into (17) bounds of Lemma 1 we arrive at (18) and (19). $\square$

A.1.3 Proof of Proposition 4

1^o. Observe first that the "true hypothesis" $H^{\ell_*-}_{i_*j}$ in quadruple $(i_*, j; \ell_*, \ell)$ is never empty because $w_* = Bx_* \in W^{\ell_*-}_{i_*j}$ for all $j \ne i_*$. Furthermore, whenever one of the hypotheses $H^{\ell-}_{ij}$, $H^{\ell\ell'+}_{ij}$ is true in a good quadruple $(i,j;\ell,\ell')$, test $\mathcal{T}^{\ell\ell'}_{ij}$ will accept it with probability at least $1 - \epsilon/\overline{N}$. Indeed, we have assumed that this is the case if both hypotheses are nonempty; since a hypothesis, when true, cannot be empty, the only other case to be considered is that of the other hypothesis in the pair being empty. It remains to recall that in this case the test always accepts the nonempty hypothesis. Thus, the (ℓ_*, x_*)-probability of $\overline{\Omega}^K$ is $\ge 1 - \epsilon$ due to the union bound.

2^o. From now on, let $\omega^K \in \overline{\Omega}^K$; in this case we have $(i_*;\ell_*) \in \mathcal{I}(\omega^K)$, implying that $\mathcal{I}(\omega^K) \neq \emptyset$; thus, if all pairs $(i,\ell) \in \mathcal{I}(\omega^K)$ share the same i-component $\widehat{i} = \widehat{i}(\omega^K)$ we clearly have $\widehat{i}(\omega^K) = i_*$. Next, suppose that $(i';\ell') \in \mathcal{I}(\omega^K)$ with $i' \neq i_*$. Observe that for all $j \neq i_*$ one has

$$\|x_* - x_j\| \leq \|x_* - x_{i_*}\| + \|x_{i_*} - x_j\| \leq \|x_* - x_{i_*}\| + 2r_{ij}.$$

We conclude that whenever quadruple $(i', i_*; \ell', \ell_*)$ is bad one has

$$\|x_* - x_{i'}\| \leq \|x_* - x_{i_*}\| + 2\delta_{i'i_*}^{\ell'\ell_*}.$$

Let us now fix a good quadruple $(i', i_*; \ell', \ell_*)$. We have

$$0 \leq \psi_{i'i_*}^T(w_* - w_{i'i_*}) < \delta_{i'i_*}^{\ell'\ell_*}$$

where the first inequality is due to

$$\|w_* - w_{i'}\|_2 = \|x_* - x_{i'}\| \geq \|x_* - x_{i_*}\| = \|w_* - w_{i_*}\|_2,$$

while the second one is due to the fact that were it false, the hypothesis $H_{i'i_*}^{\ell'+}(\delta_{i'i_*}^{\ell'\ell_*})$ would be true and thus $H_{i'i_*}^{\ell'-}$ would be rejected by test $\mathcal{T}_{i'i_*}^{\ell'\ell_*}$ (recall that $\omega^K \in \overline{\Omega}^K$), which is not the case because $(i';\ell') \in \mathcal{I}(\omega^K)$. Denoting by $\pi_{i_*i'}$ the projection of w_* onto the line passing through $w_{i'}$ and w_{i_*}, let $\tau_* = \tau(w_{i_*})$, $\tau_\pi = \tau(\pi_{i_*i'})$, and $\tau' = \tau(w_{i'})$ be coordinates of w_{i_*}, $\pi_{i_*i'}$, and $w_{i'}$ on this line, the origin on the line being the midpoint $w_{i_*i'}$ of the segment $[w_{i_*}, w_{i'}]$, its orientation given by $\psi_{i'i_*}$. One has $\tau_* = r_{i_*i'}$, $\tau' = -r_{i_*i'}$, and $\tau_\pi \leq \delta_{i'i_*}^{\ell'\ell_*}$, and so

$$\begin{array}{rcl}\|x_* - x_{i'}\|^2 - \|x_* - x_{i_*}\|^2 &=& \|w_* - w_{i'}\|_2^2 - \|w_* - w_{i_*}\|_2^2 = \|\pi_{i_*i'} - w_{i'}\|_2^2 - \|\pi_{i_*i'} - w_{i_*}\|_2^2 \\ &=& (\tau_\pi + r_{i_*i'})^2 - (\tau_\pi - r_{i_*i'})^2 = 4r_{i_*i'}\tau_\pi \leq 4r_{i_*i'}\delta_{i'i_*}^{\ell'\ell_*}.\end{array}$$

We conclude that

$$\|x_* - x_{i'}\| - \|x_* - x_{i_*}\| = \frac{\|x_* - x_{i'}\|^2 - \|x_* - x_{i_*}\|^2}{\|x_* - x_{i'}\| + \|x_* - x_{i_*}\|} \leq \frac{\|x_* - x_{i'}\|^2 - \|x_* - x_{i_*}\|^2}{2r_{i_*i'}} \leq 2\delta_{i'i_*}^{\ell'\ell_*}$$

implying (24).

3^o. Denote

$$R = \tfrac{1}{2} \max_{(i;\ell),(j;\ell') \in \mathcal{I}(\omega^K)} \|w_i - w_j\|_2,$$

and let $w_{\bar{i}}$ and $w_{\bar{j}}$ be the endpoints of a maximizing segment with $w_{\bar{i}\bar{j}} = \frac{1}{2}(w_{\bar{i}} + w_{\bar{j}})$ being its midpoint and $\widehat{x}(\omega^K) = \frac{1}{2}(x_{\bar{i}} + x_{\bar{j}})$ being the aggregated solution yielded by our algorithm. W.l.o.g. assume that $\|w_{\bar{i}} - w_*\|_2 \leq \|w_{\bar{j}} - w_*\|_2$, implying that $(w_* - w_{\bar{i}\bar{j}})^T(w_{\bar{j}} - w_{\bar{i}\bar{j}}) \leq 0$. We have $\|w_{\bar{j}} - w_{i_*}\|_2 \leq 2R$, whence, as we have just established,

$$\|w_{\bar{j}} - w_*\|_2^2 - \|w_{i_*} - w_*\|_2^2 \leq 2\|w_{\bar{j}} - w_{i_*}\|_2 \max_{\ell:(j;\ell)\in\mathcal{I}(\omega^K)} \delta_{ji_*}^{\ell'\ell_*} \leq 4R\widetilde{\delta}_{i_*}^{\ell_*}(\omega^K).$$

On the other hand,

$$\|w_{\bar{j}} - w_*\|_2^2 - \|w_{\bar{i}\bar{j}} - w_*\|_2^2 = 2(w_{\bar{i}\bar{j}} - w_*)^T(w_{\bar{j}} - w_{\bar{i}\bar{j}}) + \|w_{\bar{j}} - w_{\bar{i}\bar{j}}\|_2^2 \geq \|w_{\bar{j}} - w_{\bar{i}\bar{j}}\|_2^2 = R^2,$$

and we conclude that

$$\begin{array}{rcl}\|\widetilde{x}(\omega^K) - x_*\|^2 &=& \|w_{\bar{i}\bar{j}} - w_*\|_2^2 \leq \|w_{\bar{j}} - w_*\|_2^2 - R^2 \leq \|w_{i_*} - w_*\|_2^2 + 4R\widetilde{\delta}_{i_*}^{\ell_*}(\omega^K) - R^2 \\ &\leq& \|w_{i_*} - w\|_2^2 + 4\widetilde{\delta}_{i_*}^{\ell_*}(\omega^K)^2.\end{array}$$

what is (25). □

A.1.4 Proof of Theorem 2

1^o. Suppose that quadruple $(i,j;\ell,\ell')$ is bad. Let us verify that in this case one has $r_{ij} \le \mathfrak{r}^K_{\ell\ell'}(\varepsilon)$ where (cf. (16), (26))

$$\varepsilon = \epsilon/\overline{N},\ \ \mathfrak{r}^K_{\ell\ell'}(\varepsilon) = \tfrac{1}{2}\max_{x\in X_\ell, y\in X_{\ell'}}\left\{\|x-y\| :\ \varrho(\mathcal{A}_\ell(x),\mathcal{A}_{\ell'}(y)) \ge \varepsilon^{1/K}\right\}. \tag{57}$$

To this end, consider optimization problem

$$\begin{array}{c}\max_{x\in X_{ij}^{\ell-},y\in X_{ij}^{\ell'+}(\delta)}\varrho(\mathcal{A}_\ell(x),\mathcal{A}_{\ell'}(y)),\\ X_{ij}^{\ell-}=\{x\in X_\ell:\ Bx\in W_{ij}^{\ell-}\},\ \ X_{ij}^{\ell'+}(\delta)=\{x\in X_{\ell'}:\ Bx\in W_{ij}^{\ell'+}(\delta)\}\end{array} \tag{58}$$

for $\delta = r_{ij}$. Note that $X_{ij}^{\ell-}$ and $X_{ij}^{\ell'+}(r_{ij})$ are nonempty (otherwise the corresponding quadruple would be ε-good) convex and compact sets. Thus, problem (58) is solvable, and its optimal solution $x'\in X_{ij}^{\ell-}, y'\in X_{ij}^{\ell'+}(\delta)$ satisfies $\|x'-y'\| = \|Bx'-By'\|_2 \ge \delta = r_{ij}$. On the other hand, optimal value $\bar\rho$ of (58) is greater than $\varepsilon^{1/K}$ because, otherwise, the risk of a K-observation test $\mathcal{T}_{ij}^{\ell\ell'}$ deciding on hypothesis $H_{ij}^{\ell-}$ against $H_{ij}^{\ell'+}(r_{ij})$, as built in Section 2.2, would be bounded by $\bar\rho^K = \varepsilon$, so the quadruple $(i,j;\ell,\ell')$ would be ε-good what is not the case. In other words, x',y' is a feasible solution to the maximization problem in (57) with the value of the objective $\ge r_{ij}$, implying the desired inequality $r_{ij}\le \mathfrak{r}^K_{\ell\ell'}(\varepsilon)$.

Next, assume that quadruple $(i,j;\ell,\ell')$ is good and that $\delta_{ij}^{\ell\ell'} > \underline{\delta}$. In this case set $X_{ij}^{\ell'+}(\delta_{ij}^{\ell\ell'}-\underline{\delta})$ is not empty because $\delta = \delta_{ij}^{\ell\ell'}-\underline{\delta}$ would be ε-good otherwise, and we know it is not. Same as above, we conclude that in this case $\delta_{ij}^{\ell\ell'}-\underline{\delta} \le \mathfrak{r}^K_{\ell\ell'}(\varepsilon)$, implying that whether quadruple $(i_*,j;\ell_*,\ell)$ is good or bad, one has

$$\delta_{i_*j}^{\ell_*\ell} \le \mathfrak{r}^K_{\ell_*\ell}(\varepsilon)+\underline{\delta},$$

so that

$$\widehat{\delta}_{i_*}^{\ell_*}(\omega^K) \le \max_{\ell\le\ell_*}\mathfrak{r}^K_{\ell_*\ell}(\varepsilon)+\underline{\delta}, \tag{59a}$$

$$\widetilde{\delta}_{i_*}^{\ell_*}(\omega^K) \le \max_{\ell}\mathfrak{r}^K_{\ell_*\ell}(\varepsilon)+\underline{\delta}. \tag{59b}$$

2^o. Now (59a) combined with bound (24) imply that whenever $\omega^K\in\overline{\Omega}^K$

$$\|x_*-\widehat{x}(\omega^K)\| \le \|x_*-x_{i_*}\| + 2\left[\max_{\ell\le\ell_*}\mathfrak{r}^K_{\ell_*\ell}(\varepsilon)+\underline{\delta}\right]$$

(recall that $\|x_*-x_{i_*}\|\le\|x_*-x_{\ell_*}\|$ by construction and $\|x_*-x_{\ell_*}\| = \|x_*-\widetilde{x}_{\ell_*}(\widetilde{\omega}^{\overline{K}})\| \le \mathfrak{r}^{\overline{K}}_{\ell_*}(\epsilon)$ due to $\widetilde{\omega}^{\overline{K}}\in\widetilde{\Omega}^{\overline{K}}$). Utilizing the bound in (59b) we conclude that

$$\|x_*-\widetilde{x}(\omega^K)\|^2 \le \|x_*-x_{i_*}\|^2 + 4\left[\max_{\ell}\mathfrak{r}^K_{\ell_*\ell}(\varepsilon)+\underline{\delta}\right]^2.$$

Finally, the second part of the statement of the theorem (starting with "Consequently...") is readily given by (26) and (27) combined with the result of Lemma 1 applied with $\varepsilon=\epsilon/\overline{N}$. □

A.2 Proofs for Sections 4 and 5

A.2.1 Proof of Proposition 5

1^o. The fact that the x_*-probability of $\overline{\Xi}$ is at least $1-\epsilon$ is readily given by the union bound and the fact that when i, j are comparable and $x_* \in X_{ij}(\Delta_{ij})$, the x_*-probability for $\mathcal{T}_{ij}$ (when $i<j$) or $\mathcal{T}_{ji}$ (when $i>j$) to accept H_{ij} is at least $1-\epsilon/(N-1)$.

2^o. Let us fix $\xi \in \overline{\Xi}$ and set $\widehat{x} = \widehat{x}(\xi)$, $\widehat{\imath} = \widehat{\imath}(\xi)$, $\rho_* = \|x_* - x_{i_*}\|$. We claim that whenever i_* loses to $j \neq i_*$, we have

$$\rho_* \geq r_{i_*j} - \Delta_{i_*j} \tag{60}$$

Indeed, let $j \neq i_*$ be such that i_* loses to j. If j is comparable to i_*, we have $x_* \notin X_{i_*j}(\Delta_{i_*j})$. Indeed, otherwise the test $\mathcal{T}_{\{i_*,j\}}$ would accept H_{i_*j} due to $\xi \in \overline{\Xi}$ and j would loose to i_*, which is not the case. On the other hand, $x_* \notin X_{i_*j}(\Delta_{i_*j})$ is exactly the same as $\rho_* = \|x_* - x_{i_*}\| > r_{i_*j} - \Delta_{i_*j}$. Now, let j and i_* be incomparable; in this case i_* loosing to j means that $\rho_j \leq \rho_{i_*}$, that is, $\Delta_{i_*j} = \max[0, r_{i_*j} - \rho_{i_*}]$, implying that

$$r_{i_*j} - \Delta_{i_*j} \leq \rho_{i_*} \leq \rho_*$$

(the concluding $\leq$ being given by $\rho_{i_*} = \min_{x' \in X} \|x' - x_{i_*}\|$ combined with $\rho_* = \|x_* - x_{i_*}\|$ and $x \in X$).

3^o. Note that if $\widehat{\imath} = i_*$ (34) clearly is true. Let now $i_* \neq \widehat{\imath}$. Then, if i_* loses to no j we would have $d_{i_*} = -\infty$, and since every $i \neq i_*$ loses to i_*, $d_i \geq 0$ for all $i \neq i_*$, resulting in $\widehat{\imath} = i_*$ which is not the case. Let us assume that $\widehat{\imath} \neq i_*$ and that i_* loses to some j's; let also $j_* = j_*(\xi) \in \mathcal{I}_{i_*}(\xi)$ be such that $\|x_{i_*} - x_{j_*}\| = d_{i_*}(\xi)$. There are two possibilities:

- i_* loses to $\widehat{\imath}$; when it is the case, (60) says that $\rho_* \geq r_{i_*\widehat{\imath}} - \Delta_{i_*\widehat{\imath}}$, whence

  $$\|x_* - x_{\widehat{\imath}}\| \leq \|x_* - x_{i_*}\| + \|x_{i_*} - x_{\widehat{\imath}}\| = \rho_* + 2r_{i_*\widehat{\imath}} \leq \rho_* + 2[\rho_* + \Delta_{i_*\widehat{\imath}}] = 3\rho_* + 2\Delta_{i_*\widehat{\imath}},$$

  and (34) follows.
- $\widehat{\imath}$ loses to i_*, implying that

  $$\|x_{\widehat{\imath}} - x_{i_*}\| \leq d_{\widehat{\imath}}(\xi) \leq d_{i_*}(\xi) = \|x_{i_*} - x_{j_*}\|.$$

  Since i_* loses to j_*, we have $\rho_* \geq r_{i_*j_*} - \Delta_{i_*j_*}$ due to (60), resulting in

  $$\begin{array}{rcl} \|x_* - x_{\widehat{\imath}}\| & \leq & \|x_* - x_{i_*}\| + \|x_{i_*} - x_{\widehat{\imath}}\| \leq \|x_* - x_{i_*}\| + \|x_{i_*} - x_{j_*}\| \\ & = & \rho_* + 2r_{i_*j_*} \leq \rho_* + 2[\rho_* + \Delta_{i_*j_*}], \end{array}$$

  and (34) follows. □

A.2.2 Proof of Proposition 6

The fact that x_*-probability of $\overline{\Xi}$ is at least $1-\epsilon$ is obvious (cf. the proof of Proposition 5). Now let us fix $\xi \in \overline{\Xi}$ and let $\widehat{x} = \widehat{x}(\xi)$, $\widehat{\imath} = \widehat{\imath}(\xi)$, and $\rho_* = \|x_* - x_{i_*}\|$. Consider the following "coloring" of indices $1 \leq j \leq N$:

- j is white if $\|x_* - x_j\| \leq \rho_* + \overline{\Delta}$;
- j is gray if $\rho_* + \overline{\Delta} < \|x_* - x_j\| \leq \rho_* + 2\overline{\Delta}$;

- j is black if $\|x_* - x_j\| > \rho_* + 2\overline{\Delta}$.

Let k_w, k_g, k_b be the numbers of white, gray, and black indices, respectively. Recalling that $\xi \in \overline{\Xi}$, observe that

- When j is gray or black,

$$\|x_* - x_j\| > \rho_* + \overline{\Delta} = \|x_* - x_{i_*}\| + \overline{\Delta} \geq \|x_* - x_{i_*}\| + \Delta_{i_*j},$$

  that is, $x_* \in \mathcal{X}_{i_*j}(\Delta_{i_*j})$. It follows that as applied to observation ξ, the test $\mathcal{T}_{\{i_*,j\}}$ of hypotheses $\mathcal{H}_{i_*j}$ and $\mathcal{H}_{ji_*}$ accepts $\mathcal{H}_{i_*j}$, that is, $s_{i_*}(\xi) \leq k_w - 1$.

- When index i is black and j is a white index, we have

$$\|x_* - x_i\| > \rho_* + 2\overline{\Delta} \geq \overline{\Delta} + \|x_* - x_j\|,$$

  that is, $x_* \in \mathcal{X}_{ji}(\Delta_{ij})$. As a consequence, as applied to observation ξ, the test $\mathcal{T}_{\{i,j\}}$ of hypotheses $\mathcal{H}_{ij}$ and $\mathcal{H}_{ji}$ accepts the second hypothesis, implying that $s_i(\xi) \geq k_w$.

Taken together, the above observations say that when $\xi \in \overline{\Xi}$ stems from x_*, index $\widehat{i}(\xi)$ is either white or gray, but definitely is not black, implying that

$$\|\widehat{x} - x_*\| \leq \rho_* + 2\overline{\Delta}. \qquad \square$$

A.2.3 Proof of Theorem 3

1^o. Given a pair $(i,j) \in \mathcal{O}$ such that $r_{ij} = \frac{1}{2}\|x_i - x_j\| > \delta$, let us set

$$X_{ij}(\lambda,\delta) = \Big\{z \in X : \|z - x_i\| \leq \underbrace{\lambda(r_{ij} - \delta)}_{d_{ij}}\Big\}$$

where $0 < \lambda < (1+\gamma)^{-1}$. Under the premise of Theorem, for any such pair i,j there exists a $\overline{K}$-observation test deciding with risk $\leq \epsilon$ on a pair of hypotheses $\overline{H}_{ij}$ and $\overline{H}_{ji}$ stating, respectively, that $\omega^{\overline{K}}$ stems from signal belonging to $X_{ij}(\lambda,\delta)$ and $X_{ji}(\lambda,\delta)$, and both these sets are nonempty (recall that we are in the case where $x_s \in X$, $1 \leq s \leq N$). The desired test $\mathcal{T}$ is as follows: given observation $\omega^{\overline{K}}$ we compute $\bar{x}(\omega^{\overline{K}})$ and accept $\overline{H}_{ij}$ when $\|\bar{x}(\omega^{\overline{K}}) - x_i\| \leq \|\bar{x}(\omega^{\overline{K}}) - x_j\|$, and accept $\overline{H}_{ji}$ otherwise.

Let us verify that the risk of this test is indeed at most ϵ. Suppose, first, that $\overline{H}_{ij}$ takes place, so that $\omega^{\overline{K}} \sim p^{\overline{K}}_{\mathcal{A}_\nu(x_*)}$ for some $\nu \leq J$ and $x_* \in X_{ij}(\lambda(r_{ij} - \delta))$. Then, if x_{i_*} is the closest to x_* point among $x_1, ..., x_N$, we have

$$\|x_* - x_{i_*}\| \leq \|x_* - x_i\| \leq \lambda(r_{ij} - \delta),$$

and so $p^{\overline{K}}_{\mathcal{A}_\nu(x)}$-probability of the event

$$\mathcal{E} = \{\omega^{\overline{K}} : \|\bar{x}(\omega^{\overline{K}}) - x\| \leq \gamma d_{ij} + \delta\}$$

is at least $1 - \epsilon$ due to the origin of $\bar{x}(\cdot)$. But if $\mathcal{E}$ takes place,

$$\|\bar{x}(\omega^{\overline{K}}) - x_i\| \leq \|\bar{x}(\omega^{\overline{K}}) - x_*\| + \|x_i - x_*\| \leq (\gamma+1)d_{ij} + \delta < r_{ij},$$

so that

$$\|\bar{x}(\omega^{\overline{K}}) - x_j\| \geq \|x_i - x_j\| - \|\bar{x}(\omega^{\overline{K}}) - x_i\| > r_{ij}.$$

We conclude that the $p^{\overline{K}}_{\mathcal{A}_\nu(x_*)}$-probability for $\mathcal{T}$ not to accept $\overline{H}_{ij}$ is $\le \epsilon$. By "symmetric reasoning," when $\overline{H}_{ji}$ holds true, so that $\omega^K \sim p^{\overline{K}}_{\mathcal{A}_\nu(x_*)}$ for some $\nu \le J$ and $x_* \in X_{ji}(\lambda(r_{ij}-\delta))$, $p^{\overline{K}}_{\mathcal{A}_\nu(x_*)}$-probability to reject $\overline{H}_{ji}$ is at most ϵ.

Now, testing $\overline{H}_{ji}$ against $\overline{H}_{ij}$ is equivalent to deciding between "red" set $R_{ij}(\lambda,\delta)$ and "blue" set $B_{ij}(\lambda,\delta)$ in the space $\mathcal{M}$ of parameters of distribution $p^{\overline{K}}_\mu$ of $\omega^{\overline{K}}$, each set being a union of at most J convex and compact sets:

$$R_{ij}(\lambda,\delta) = \bigcup_{\nu=1}^{J} R_{ij\nu}(\lambda,\delta),\ R_{ij\nu}(\lambda,\delta) = \{\mathcal{A}_\nu(x) : x \in X_\nu,\ \|x-x_i\| \le \lambda(r_{ij}-\delta)\},$$

and

$$B_{ij}(\lambda,\delta) = \bigcup_{\nu=1}^{J} B_{ij\nu}(\lambda,\delta),\ B_{ij\nu}(\lambda,\delta) = \{\mathcal{A}_\nu(x) : x \in X_\nu,\ \|x-x_j\| \le \lambda(r_{ij}-\delta)\},\ \nu = 1,...,J.$$

From what we know about color inferring test in simple observation schemes, the fact that the hypotheses $\overline{H}_{ij}$ and $\overline{H}_{ji}$ can be decided upon via $\overline{K}$-repeated observation $\omega^{\overline{K}} \sim p^{\overline{K}}_{\mathcal{A}_\nu(x)}$ with risk $\epsilon \in (0,1/2)$ implies (cf. Proposition 2) that when K satisfies (37), we have at our disposal test $\mathcal{T}_{ij}$ utilising K-repeated observation ω^K which decides with maximal risk not exceeding $\epsilon/(N-1)$ upon hypotheses H_{ij} and H_{ji} stating that ω^K stems from a signal x such that $x \in X_{ij}(\lambda(r_{ij}-\delta))$ (for H_{ij}) and $x \in X_{ji}(\lambda(r_{ij}-\delta)$ (for H_{ji}).

2^o. Now let us apply the aggregation procedure described in Section 5.2 to K-repeated observations, with K satisfying (37). From what we have just seen, in this case, all pairs (i,j) such that $r_{ij} > \delta$ are appropriate, and the quantity $(1-\lambda)r_{ij} + \lambda\delta$ with $\lambda \in D := (0,(1+\gamma)^{-1})$ is (i,j)-appropriate. Let us set

$$\widetilde{\lambda} = (1+\gamma')^{-1} \in D,\ \widetilde{d}_{ij} = \widetilde{\lambda}(r_{ij}-\delta),\ \widetilde{\delta}_{ij} = r_{ij} - \widetilde{d}_{ij} = (1-\widetilde{\lambda})r_{ij} + \widetilde{\lambda}\delta,$$

so that $\Delta_{ij} := \widetilde{\delta}_{ij}$ is (i,j)-appropriate, as required in the construction we are implementing. Let us define the set of comparable pairs to be exactly the set of pairs $\{i,j\} \in \mathcal{U}$ with $r_{ij} > \delta$ and equip these pairs with the tests $\mathcal{T}_{\{i,j\}} = \mathcal{T}_{\min[i,j],\max[i,j]}$. For these pairs the quantities Δ_{ij} satisfy the relations

$$\Delta_{ij} = (1-\widetilde{\lambda})r_{ij} + \widetilde{\lambda}\delta = \frac{\gamma'}{1+\gamma'}r_{ij} + \frac{1}{1+\gamma'}\delta. \tag{61}$$

Note that for incomparable pairs $(i,j) \in \mathcal{O}$ we have $\Delta_{ij} = \max[0, r_{ij} - \max[\rho_i,\rho_j]] \le r_{ij} \le \delta$.

3^o. Now let observation ω^K stem from signal $x_* \in X$, so that $\omega^K \sim p^K_{\mathcal{A}_\nu(x_*)}$ for some ν such that $x_* \in X_\nu$, and let i_* be the index of one of the $\|\cdot\|$-closest to x_* points $x_1,...,x_N$. Finally, let $\overline{\Omega}$ be the set of ω^K satisfying the condition

> whenever $j \ne i_*$ is such that $r_{i_*j} > \delta$ (i.e., i_* and j are comparable) and
>
> $$\|x_* - x_{i_*}\| \le r_{i_*j} - \Delta_{i_*j}$$
>
> (i.e., hypothesis H_{i_*j} holds true), test $\mathcal{T}_{\{i_*,j\}}$ as applied to observation ω^K accepts H_{i_*j}.

By Proposition 5, the $p^K_{\mathcal{A}_\nu(x_*)}$-probability of $\overline{\Omega}$ is at least $1-\epsilon$, and

$$\|x_* - \widehat{x}(\omega^K)\| \le 3\|x_* - x_{i_*}\| + 2\overline{\Delta}_{i_*}(\omega^K). \tag{62}$$

Next, let $\omega^K \in \overline{\Omega}$, and let $j \in \mathcal{I}_{i_*}(\omega^K)$. It may happen that i_* and j are comparable; in this case H_{i_*j} cannot be true due to $\omega^K \in \overline{\Omega}$, that is, $\|x_* - x_{i_*}\| > r_{i_*j} - \Delta_{i_*j}$, and besides this,

$$\Delta_{i_*j} \le \frac{\gamma'}{1+\gamma'} r_{i_*j} + \frac{1}{1+\gamma'}\delta$$

due to (61), implying that $r_{i_*j} \le (1+\gamma')\|x_* - x_{i_*}\| + \delta$. Hence,

$$\Delta_{i_*j} \le \frac{\gamma'}{1+\gamma'} r_{i_*j} + \frac{1}{1+\gamma'}\delta \le \gamma'\|x_* - x_{i_*}\| + \delta.$$

When i_* and j are incomparable, we have $\Delta_{i_*j} = r_{i_*j} \le \delta$. We see that when $\omega^K \in \overline{\Omega}$ (what happens with $p^K_{\mathcal{A}_\nu(x_*)}$-probability at least $1-\epsilon$), we have $\overline{\Delta}_{i_*}(\omega^K) \le \gamma'\|x_* - x_{i_*}\| + \delta$. This combines with (62) to imply that when $\omega^K \in \overline{\Omega}$, we also have $\|\widehat{x} - x_{i_*}\| \le (3+2\gamma')\|x_* - x_{i_*}\| + 2\delta$. □

A.2.4 Proof of Theorem 4

1^o. Let $\overline{\mathcal{I}}$ be the set of pairs $\{i,j\} \in \mathcal{U}$ such that both hypotheses $\mathcal{H}_{ij}(\bar{\delta}) = H(\mathcal{X}_{ij}(\bar{\delta}))$ and $\mathcal{H}_{ji}(\bar{\delta}) = H(\mathcal{X}_{ji}(\bar{\delta}))$ are nonempty. Let $\{i,j\} \in \overline{\mathcal{I}}$ be fixed, and let us show that under the premise of the theorem the simple test which, given $\omega^{\overline{K}}$, accepts $\mathcal{H}_{ij}$ when $\iota_{ij}(\omega^{\overline{K}}) = i$, and accepts $\mathcal{H}_{ji}$ otherwise has its risk bounded with ϵ. Indeed, let $\mathcal{H}_{ij}$ be true, that is, the distribution P of observation $\omega^{\overline{K}}$ satisfies $P \in \mathcal{P}_x^{\overline{K}}$ for some $x \in \mathcal{X}_{ij}(\bar{\delta})$, so that $\|x - x_j\| \ge \|x - x_i\| + \bar{\delta}$, whence $\|x - x_j\| \ge \min[\|x - x_i\|, \|x - x_j\|] + \bar{\delta}$. By (40) the P-probability of the event $\iota_{ij} = j$, that is, the probability of the test in question rejecting $\mathcal{H}_{ij}$, is $\le \epsilon$. By "symmetric" reasoning, the P-probability to reject $\mathcal{H}_{ji}(\bar{\delta})$ when the hypothesis is true is $\le \epsilon$ as well.

Now recall that testing $\mathcal{H}_{ij}(\bar{\delta})$ vs $\mathcal{H}_{ji}(\bar{\delta})$ via a $\overline{K}$-repeated observation is equivalent to deciding via this observation on "red" set $R_{ij}(\bar{\delta})$ vs. "blue" set $B_{ij}(\bar{\delta})$ in the space $\mathcal{M}$ of parameters of distribution p_μ of ω_k, and each set is a union of at most J convex and compact sets:

$$R_{ij}(\bar{\delta}) = \bigcup_{\nu=1}^{J} R_{ij\nu}(\bar{\delta}),\ R_{ij\nu}(\bar{\delta}) = \{\mathcal{A}_\nu(x) : x \in X_\nu,\ \|x - x_i\| \le \|x - x_j\| - \bar{\delta})\},$$

and

$$B_{ij}(\bar{\delta}) = \bigcup_{\nu=1}^{J} B_{ij\nu}(\bar{\delta}),\ B_{ij\nu}(\bar{\delta}) = \{\mathcal{A}_\nu(x) : x \in X_\nu,\ \|x - x_j\| \le \|x - x_i\| - \bar{\delta}\}.$$

The fact that hypotheses $\mathcal{H}_{ij}(\bar{\delta})$ and $\mathcal{H}_{ji}(\bar{\delta})$ can be decided upon via $\overline{K}$-repeated observation with risk $0 \le \epsilon < 1/2$ implies by Proposition 2 that whenever

$$K \ge \left\lceil \frac{2\ln\left(J\overline{N}/\epsilon\right)}{\ln\left([4\epsilon(1-\epsilon)]^{-1}\right)}\overline{K} \right\rceil,$$

$\overline{\delta}$ is (i,j)-good in the sense of Section 5.3.1.

2^o. Now let K satisfy (41) and $\{i,j\} \in \overline{\mathcal{I}}$, that is, both $\mathcal{X}_{ij}(\bar{\delta})$ and $\mathcal{X}_{ji}(\bar{\delta})$ are nonempty. Recall that in this case in our aggregation procedure Δ_{ij} is selected to be (i,j)-good (that is, with K observations, the test yielded by the machinery from Section 2.3 decides on the hypothesis $\mathcal{H}_{ij}(\Delta_{ij})$ vs. the alternative $\mathcal{H}_{ji}(\Delta_{ij})$ with risk not exceeding $\epsilon/\overline{N}$) and either $\Delta_{ij} \le \underline{\delta}$, or $\Delta_{ij} - \underline{\delta}$ is not (i,j)-good. By item 1^o,

for our i, j, K $\delta = \bar{\delta}$ is (i,j)-good, so that the second option implies that $\Delta_{ij} - \underline{\delta} \le \bar{\delta}$ and one always has $\Delta_{ij} \le \bar{\delta} + \underline{\delta}$.

On the other hand, if $i \ne j$ and $\{i,j\} \notin \overline{\mathcal{I}}$, at least one of the sets $\mathcal{X}_{ij}(\bar{\delta})$, $\mathcal{X}_{ji}(\bar{\delta})$ is empty, implying that $\bar{\delta}$ is (i,j)-good. Consequently, in our aggregation procedure, same as in the case of $\{i,j\} \in \mathcal{I}$, one has $\Delta_{ij} \le \bar{\delta} + \underline{\delta}$. Thus, $\Delta_{ij} \le \bar{\delta} + \underline{\delta}$ for all $i \ne j$, and (42) is given by Proposition 6. □

A.2.5 Proof of Proposition 9

We start with the following observation.

Lemma 2. *Under the premise of the proposition, let $\bar{\varrho}^* = \mathrm{RiskOpt}^{\overline{1,N}}_{\epsilon,\overline{K}}[X]$ be the minimax risk of $\overline{K}$-observation estimation over X. Let also K satisfy (46) and $\{i,j\} \in \mathcal{U}$ be such that $\bar{\varrho}^* < \overline{\delta}^{ij}$ (cf. (36)). Then any δ such that $\bar{\varrho}^* < \delta \le \overline{\delta}^{ij}$ is (i,j)-appropriate.*

Proof. Under the lemma's premise, for any $\rho > \bar{\varrho}^*$ there exists an estimate $\bar{x} = \bar{x}(\omega^{\overline{K}})$ such that for every $x \in X$, the x-probability of the event $\|\bar{x} - x\| \le \rho$ is at least $1 - \epsilon$. As a result, for any $i \ne j$ and $\delta > \rho$ there exists a $\overline{K}$-observation test deciding on hypotheses $H_{ij}(\delta)$ and $H_{ji}(\delta)$ with risk bounded with ϵ, namely, test $\overline{T}_{\{i,j\}}$ accepting $H_{ij}(\delta)$ if $\|\bar{x} - x_i\| \le r_{ij}$ and accepting $H_{ji}(\delta)$ otherwise. Indeed, assuming that $H_{ij}(\delta)$ takes place, the distribution $P^{\overline{K}}$ of observation $\omega^{\overline{K}}$ stems from some $x \in X$ satisfying $\|x - x_i\| \le r_{ij} - \delta < r_{ij} - \rho$, so that when the event $\|\bar{x} - x\| \le \rho$ takes place (which happens with $P^{\overline{K}}$-probability $\ge 1 - \epsilon$), we have

$$\|\bar{x} - x_i\| \le \|\bar{x} - x\| + \|x - x_i\| < r_{ij},$$

and test $\overline{T}_{\{i,j\}}$ accepts $H_{ij}(\delta)$. Similarly, when $H_{ji}(\delta)$ takes place, the distribution $P^{\overline{K}}$ of $\omega^{\overline{K}}$ stems from some $x \in X$ satisfying $\|x - x_j\| \le r_{ij} - \delta < r_{ij} - \rho$, so that when the event $\|\bar{x} - x\| \le \rho$ takes place (which happens with $P^{\overline{K}}$-probability $\ge 1 - \epsilon$), we have $\|\bar{x} - x_j\| \le \|\bar{x} - x\| + \|x - x_j\| < \rho + r_{ij} - \rho = r_{ij}$, whence $\|\bar{x} - x_i\| > 2r_{ij} - r_{ij} > r_{ij}$, and $\overline{T}_{\{i,j\}}$ accepts $H_{ji}(\delta)$.

Recalling that X is the union of at most N convex and compact sets, we conclude that when K satisfies (46), the risk of the K-observation test deciding on $H_{ij}(\delta)$ vs $H_{ji}(\delta)$ constructed in Section 5.2 does not exceed $\epsilon/(N-1)$. □

The claim of the proposition is readily given by combining the bound (45) with the fact that by Lemma 2 the quantity $\overline{\Delta}$, as is immediately seen, can be bounded by a quantity arbitrarily close to $\bar{\varrho}^*$. □

A.2.6 Proof of Proposition 10

The statement of the proposition is readily implied by the following analog of Lemma 2.

Lemma 3. *Given a positive integer $\overline{K}$ and $\epsilon \in (0, 1/2)$, let $\bar{\varrho}^* = \mathrm{Risk}^{\overline{1,N}}_{\epsilon,\overline{K}}[X]$ be the minimax ϵ-risk of estimation over X, and let K satisfy (47). Then $\overline{\Delta} \le 2\bar{\varrho}^* + \underline{\delta}$.*

Proof. Let $\widetilde{\delta} > 2\bar{\varrho}^*$ and let $(i,j) \in \mathcal{O}$; let us show that $\widetilde{\delta}$ is (i,j)-good (for the definition of (i,j)-goodness, see Section 5.3). There is nothing to prove when at least one of the sets $\mathcal{X}_{ij}(\widetilde{\delta})$, $\mathcal{X}_{ji}(\widetilde{\delta})$ is empty. Assuming these sets nonempty, let $\rho > \bar{\varrho}^*$ be such that $2\rho < \widetilde{\delta}$. Then there is an estimate $\bar{x}(\omega^{\overline{K}})$

such that for every $x \in X$, the x-probability of the event $\|x - \bar{x}(\omega^{\overline{K}})\| \le \rho$ is $\ge 1-\epsilon$. We immediately convert this estimate into a $\overline{K}$-observation test deciding on the hypothesis $\mathcal{H}_{ij} = H(\mathcal{X}_{ij}(\widetilde{\delta}))$ vs the alternative $\mathcal{H}_{ji} = H(\mathcal{X}_{ji}(\widetilde{\delta}))$: given $\omega^{\overline{K}}$ and setting $\bar{x} = \bar{x}(\omega^{\overline{K}})$, this test accepts $\mathcal{H}_{ij}$ (and rejects $\mathcal{H}_{ji}$) when $\|\bar{x} - x_i\| \le \|\bar{x} - x_j\|$, and accepts $\mathcal{H}_{ji}$ (and rejects $\mathcal{H}_{ij}$) otherwise. Observe that the risk of this test is $\le \epsilon$. Indeed, when $\mathcal{H}_{ij}$ takes place, the distribution $P^{\overline{K}}$ of $\omega^{\overline{K}}$ stems from some $x \in \mathcal{X}_{ij}(\widetilde{\delta})$, that is, $x \in X$ and $\|x - x_i\| \le \|x - x_j\| - \widetilde{\delta}$. Therefore when $\|x - \bar{x}\| \le \rho$ (the latter happens with $P^{\overline{K}}$-probability $\ge 1-\epsilon$), we have

$$\begin{array}{rcl} \|\bar{x} - x_i\| & \le & \|x - \bar{x}\| + \|x - x_i\| \le \rho + \|x - x_j\| - \widetilde{\delta} \le \rho + \|\bar{x} - x_j\| + \|\bar{x} - x\| - \widetilde{\delta} \\ & \le & \|\bar{x} - x_j\| + 2\rho - \widetilde{\delta} < \|\bar{x} - x_j\|, \end{array}$$

and the test accepts $\mathcal{H}_{ij}$. "Symmetric" reasoning shows that when $\mathcal{H}_{ji}$ takes place, the test accepts $\mathcal{H}_{ji}$ and rejects $\mathcal{H}_{ij}$ when $\|x - \bar{x}\| \le \rho$, which happens with x-probability $\ge 1-\epsilon$, implying that the risk of the test is $\le \epsilon$.

Because X is the union of N convex compact sets, existence of pairwise $\overline{K}$-observation tests deciding with risk $\le \epsilon$ on all pairs $\mathcal{H}_{ij}$, $\mathcal{H}_{ji}$ of nonempty hypotheses with $(i,j) \in \mathcal{O}$ implies, by the results of Section 2.3, that with K as in (47) $\widetilde{\delta}$ indeed is (i,j)-good for all $(i,j) \in \mathcal{O}$.

Now, for every $(i,j) \in \mathcal{O}$ by construction the quantity Δ_{ij} is (i,j)-good and is either $\le \underline{\delta}$, or is such that $\Delta_{ij} - \underline{\delta}$ is not (i,j)-good. By the above, the second option implies that $\Delta_{ij} < \widetilde{\delta} + \underline{\delta}$ for all $(i,j) \in \mathcal{O}$, so that $\overline{\Delta} < \widetilde{\delta} + \underline{\delta}$. The latter inequality holds true whenever $\widetilde{\delta} > 2\bar{\varrho}^*$, and the conclusion of the lemma follows. □

A.3 Proof of Theorem 5

Let us verify that in a K-bad pair (i,j) δ_{ij}, as defined in (12), satisfies

$$\delta_{ij} \le \mathfrak{g}_{ij}^K(\varepsilon).$$

Indeed, consider optimization problem

$$\min_{x \in B_i, y \in B_j} \|\mathcal{A}_i(x) - \mathcal{A}_j(y)\|_2; \tag{63}$$

observe that (63) is solvable, and its optimal solution $x' \in B_i$, $y' \in B_j$ satisfies $\|x' - y'\| \ge 2\delta_{ij}$. On the other hand, the optimal value of (56) is less than $\frac{2}{\sqrt{K}} q_{\mathcal{N}}(1-\varepsilon)$ because, otherwise, the risk of K-observation test $\mathcal{T}_{\{i,j\}}$ deciding on hypotheses H_i and H_j, see (12), as yielded in Gaussian case by the machinery from Section 2.2, would be bounded by ε, implying that pair (i,j) is K-good what is not the case. We conclude that $\mathfrak{g}_{ij}^K(\varepsilon)$, as defined in (49), satisfies $\mathfrak{g}_{ij}^K(\varepsilon) \ge \frac{1}{2}\|x' - y'\| \ge \delta_{ij}$; along with the result of Proposition 3 (see (13) and (14)) this implies relation (51).

2°. Let us fix $(i,j) \in \mathcal{O}$. Let for $\upsilon \in (0,1)$ $\bar{\varrho}_{ij}^*(\upsilon) = \mathrm{RiskOpt}_{\upsilon,\overline{K}}^{\{i,j\}}[X_i \cup X_j]$; let also $(\bar{x}, \bar{y})$, $\bar{x} \in X_i, \bar{y} \in X_j$, be an optimal solution to (49) with $\delta = \varepsilon$. Note that for $\vartheta \in [0,1]$, $x(\vartheta) = \vartheta\bar{x} \in X_i$ and $y(\vartheta) = \vartheta\bar{y} \in X_j$, while $\rho(\vartheta) = \|\mathcal{A}_i(x(\vartheta)) - \mathcal{A}_j(y(\vartheta))\|_2$ is a linear function of ϑ with $\rho(1) \le \frac{2}{\sqrt{K}} q_{\mathcal{N}}(1-\varepsilon)$ and $\rho(0) = 0$. Let now

$$\widetilde{\vartheta} = \frac{\sqrt{K} q_{\mathcal{N}}(1-\upsilon)}{\sqrt{\overline{K}} q_{\mathcal{N}}\Big(1-\varepsilon\Big)} \le 1;$$

then for $\vartheta < \widetilde{\vartheta}$ one has

$$\rho(\vartheta) \le \vartheta\rho(1) < \frac{2}{\sqrt{\overline{K}}} q_{\mathcal{N}}(1-\upsilon).$$

The latter relation means that there is no $\overline{K}$-observation test capable of deciding between hypotheses $H_{x(\vartheta)} : \omega_k \sim \mathcal{N}(A_i x(\vartheta), I_n)$ and $H_{y(\vartheta)} : \omega_k \sim \mathcal{N}(A_j y(\vartheta), I_m)$ with risk bounded with υ, implying in its turn that

$$\bar{\varrho}^*_{ij}(\upsilon) > \tfrac{1}{2}\|x(\vartheta) - y(\vartheta)\| = \tfrac{1}{2}\vartheta\|\bar{x} - \bar{y}\| = \vartheta \mathfrak{g}^K_{ij}(\varepsilon).$$

Applying the latter bound to $\upsilon = \epsilon$ (recall that $\widetilde{\vartheta} \le 1$ for $\upsilon = \epsilon$ due to (50)), we obtain

$$\mathfrak{g}^K_{ij}(\varepsilon) \le \overline{\vartheta}^{-1} \bar{\varrho}^*_{ij}(\epsilon)$$

which combines with (51) to imply (52).

The same bound as applied with $\upsilon = 1/16$ and $K = \overline{K}$ (this again is possible due to $\varepsilon \le \epsilon < 1/16$) implies that

$$\mathfrak{g}^{\overline{K}}_{ij}(\varepsilon) \le \frac{q_{\mathcal{N}}(1-\varepsilon)}{q_{\mathcal{N}}(\frac{15}{16})} \bar{\varrho}^*_{ij}(\tfrac{1}{16}) \le \mathfrak{C}_4 \sqrt{\ln[N/\epsilon]} \mathrm{RiskOpt}^{\{i,j\}}_{\frac{1}{16},\overline{K}}[X_i \cup X_j]. \tag{64}$$

3^o. Thus, all we need to show the last statement of the theorem, is to bound the quantity $\overline{\mathrm{RiskOpt}}^{\{j\}}_{\frac{1}{8},\overline{K}}[X_j]$—the minimax 1/8-risk of recovering $x \in X_j$ from single "averaged" observation

$$\widehat{\omega} \sim \mathcal{N}(\mathcal{A}_j(x), \overline{K}^{-1} I_m).$$

Common sense says that $\overline{\mathrm{RiskOpt}}^{\{j\}}_{\frac{1}{8},\overline{K}}[X_j]$ is exactly the same as $\mathrm{RiskOpt}^{\{j\}}_{1/8,\overline{K}}[X_j]$, but we do not know why this would be the case.[15] Instead, we are about to establish a slightly weaker fact which is sufficient for our purposes.

Lemma 4. *Suppose that for a positive integer M, $Y \subset \mathbf{R}^n$, and $\upsilon \in (0, 1/4)$ $\mathrm{Riskopt}_{\upsilon,M}[Y]$ is the minimax over Y υ-risk of estimation given an M-repeated observation $(\omega_1, ..., \omega_M)$, $\omega_k \sim \mathcal{N}(\mathcal{A}(x), I_m)$, $x \in Y$. Then minimax over Y 2υ-risk $\overline{\mathrm{Riskopt}}_{2\upsilon,M}[Y]$ of estimation given single observation*

$$\overline{\omega} = \frac{1}{M}\sum_{i=1}^{M} \omega_k$$

satisfies

$$\overline{\mathrm{Riskopt}}_{2\upsilon,M}[Y] \le 2\mathrm{Riskopt}_{\upsilon,M}[Y]. \tag{65}$$

Proof. Note that ω_k, $k = 1, ..., M$ can be represented as $\omega_k = \eta_k + \overline{\omega}$ where $\eta_k \sim \mathcal{N}\Big(0, \frac{M-1}{M} I\Big)$ are independent of $\overline{\omega}$. This observation implies that if $\overline{\mathrm{Riskopt}}^{\mathcal{R}}_{\upsilon,M}[X_j]$ is defined in the same fashion as $\overline{\mathrm{Riskopt}}_{\upsilon,M}[X_j]$ but with candidate estimates which may be *randomized* then

$$\overline{\mathrm{Riskopt}}^{\mathcal{R}}_{\upsilon,M}[X_j] \le \mathrm{Riskopt}_{\upsilon,M}[X_j].$$

[15] Recall that $\widehat{\omega}$ is sufficient statistics when estimating functions of the mean of the Gaussian distribution—conditional distributions of $\omega^{\overline{K}}$ given $\widehat{\omega}$ is Gaussian and does not depend on x. Were the considered loss convex, the corresponding result would be readily given by the Rao-Blackwell theorem.

We claim that

$$\overline{\mathrm{Riskopt}}_{2\upsilon,M}[X_j] \le 2\overline{\mathrm{Riskopt}}^{\mathcal{R}}_{\upsilon,M}[X_j] \tag{66}$$

what obviously implies the lemma. Indeed, let $\rho > \overline{\mathrm{Riskopt}}^{\mathcal{R}}_{\upsilon,M}$, so that there exists a deterministic function $\phi(\omega,\zeta)$ taking values in $\mathbf{R}^n$ such that for every $x \in Y$ it holds

$$\mathrm{Prob}_{(\overline{\omega},\zeta)\sim P}\{\|\phi(\overline{\omega},\zeta) - x\| > \rho\} \le \upsilon,$$

where P is the distribution of $(\overline{\omega},\zeta)$ with independent of each other $\overline{\omega} \sim \mathcal{N}(\mathcal{A}(x), M^{-1}I_m)$ and $\zeta \sim U$, U being the uniform distribution over $[0,1]$. Let

$$\overline{\Omega} = \big\{\overline{\omega} \in \mathbf{R}^m : \exists y \in Y : \mathrm{Prob}_{\zeta\sim U}\{\zeta :\ \|\phi(\overline{\omega},\zeta) - y\| \le \rho\} > 1/2\big\}.$$

For every $\overline{\omega} \in \overline{\Omega}$, we can specify $\psi(\overline{\omega}) \in Y$ in such a way that

$$\mathrm{Prob}_{\zeta\sim U}\{\zeta : \|\phi(\overline{\omega},\zeta) - \psi(\overline{\omega})\| \le \rho\} > 1/2,$$

and define $\psi(\overline{\omega})$ as once for ever fixed point of Y when $\overline{\omega} \notin \overline{\Omega}$. For $x \in Y$, let also

$$\widetilde{\Omega}[x] = \big\{\overline{\omega} \in \mathbf{R}^m : \mathrm{Prob}_{\zeta\sim U}\{\zeta : \|\phi(\bar{\omega},\zeta) - x\| \le \rho\} > 1/2\big\},$$

note that $\widetilde{\Omega}[x] \subset \overline{\Omega}$. Let now $\widetilde{\Omega}^c[x]$ be the complement of $\widetilde{\Omega}[x]$; due to the origin ρ, we have for every $x \in Y$:

$$\mathrm{Prob}_{\overline{\omega}\sim\mathcal{N}(\mathcal{A}(x),M^{-1}I_m)}\widetilde{\Omega}^c[x] \le 2\upsilon.$$

On the other hand, whenever $\overline{\omega} \in \widetilde{\Omega}[x]$, both sets

$$\{\zeta : \|\phi(\overline{\omega},\zeta) - x\| \le \rho\} \ \text{ and } \ \{\zeta : \|\phi(\overline{\omega},\zeta) - \psi(\overline{\omega})\| \le \rho\}$$

are subsets of $[0,1]$ of measure $> 1/2$ and thus intersect, implying that $\|x - \psi(\overline{\omega})\| \le 2\rho$. We conclude that for every $x \in Y$ the stemming from x probability of the event $\|\psi(\overline{\omega}) - x\| > 2\rho$ is at most 2υ, that is,

$$\overline{\mathrm{RiskOpt}}_{2\upsilon,M}[X_j] \le 2\rho.$$

Because ρ may be arbitrary $> \mathrm{RiskOpt}^{\mathcal{R}}_{\upsilon,M}[Y]$, (66) follows. □

When combining (48) and (65) we conclude that because $\epsilon \le 1/16$ one has

$$\mathfrak{r}_j(\epsilon) \le \mathfrak{C}_5 \ln(L+L')\sqrt{\ln(m/\epsilon)}\mathrm{RiskOpt}_{\frac{1}{16},M}[X_j] \quad \forall j \le N.$$

Taken together with (64) the latter bound implies the last statement of the theorem. □

References

[1] J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. In *Annales de l'IHP Probabilités et statistiques*, volume 40, pages 685–736, 2004.

[2] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. In *Annales de l'IHP Probabilités et statistiques*, volume 42, pages 273–325, 2006.

[3] L. Birgé et al. Model selection for poisson processes. In *Asymptotics: particles, processes and inverse problems*, pages 32–64. Institute of Mathematical Statistics, 2007.

[4] L. Birgé et al. Robust tests for model selection. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 47–64. Institute of Mathematical Statistics, 2013.

[5] O. Bousquet, D. Kane, and S. Moran. The optimal approximation factor in density estimation. *arXiv preprint arXiv:1902.05876*, 2019.

[6] F. Bunea, A. B. Tsybakov, M. H. Wegkamp, et al. Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

[7] M. Burnashev. On the minimax detection of an imperfectly known signal in a white noise background. *Theory Probab. Appl.*, 24:107–119, 1979.

[8] M. Burnashev. Discrimination of hypotheses for gaussian measures and a geometric characterization of the gaussian distribution. *Math. Notes*, 32:757–761, 1982.

[9] T. T. Cai and M. G. Low. Minimax estimation of linear functionals over nonconvex parameter spaces. *The Annals of Statistics*, 32(2):552–576, 2004.

[10] T. T. Cai and M. G. Low. On adaptive estimation of linear functionals. *The Annals of Statistics*, 33(5):2311–2343, 2005.

[11] O. Catoni. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*, volume 1851. Springer Science & Business Media, 2004.

[12] S. O. Chan, I. Diakonikolas, R. A. Servedio, and X. Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *Advances in neural information processing systems*, pages 1844–1852, 2014.

[13] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

[14] D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy q-aggregation. *The Annals of Statistics*, 40(3):1878–1905, 2012.

[15] L. Devroye and G. Lugosi. A universally acceptable smoothing factor for kernel density estimates. *The Annals of Statistics*, pages 2499–2512, 1996.

[16] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.

[17] A. Goldenshluger. A universal procedure for aggregating estimators. *The Annals of Statistics*, pages 542–568, 2009.

[18] A. Goldenshluger, A. Juditsky, and A. Nemirovski. Hypothesis testing by convex optimization. *Electronic Journal of Statistics*, 9(2):1645–1712, 2015.

[19] A. Goldenshluger and O. Lepski. Structural adaptation via l_p-norm oracle inequalities. *Probability Theory and Related Fields*, 143(1-2):41–71, 2009.

[20] A. Goldenshluger, O. Lepski, et al. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190, 2008.

[21] A. Goldenshluger, O. Lepski, et al. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011.

[22] G. K. Golubev. Asymptotic minimax estimation of regression in the additive model. *Problems of Information Transmission*, 28:101–112, 1992.

[23] N. Hengartner and M. Wegkamp. Estimation and selection procedures in regression: an l1 approach. *Canadian Journal of Statistics*, 29(4):621–632, 2001.

[24] A. Juditsky and A. Nemirovski. Hypothesis testing via affine detectors. *Electronic journal of statistics*, 10(2):2204–2242, 2016.

[25] A. Juditsky and A. Nemirovski. Near-optimality of linear recovery in gaussian observation scheme under $\|\cdot\|_2^2$-loss. *The Annals of Statistics*, 46(4):1603–1629, 2018.

[26] A. Juditsky and A. Nemirovski. Near-optimal recovery of linear and n-convex functions on unions of convex sets. *Information and Inference: A Journal of the IMA*, 9(2):423–453, 2020.

[27] A. Juditsky and A. Nemirovski. *Statistical Inference via Convex Optimization*. Princeton University Press, 2020.

[28] A. Juditsky, A. Nemirovski, et al. On polyhedral estimation of signals via indirect observations. *Electronic Journal of Statistics*, 14(1):458–502, 2020.

[29] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

[30] G. Lecué, P. Rigollet, et al. Optimal learning with q-aggregation. *Annals of Statistics*, 42(1):211–224, 2014.

[31] O. Lepski et al. Adaptive estimation over anisotropic functional classes via oracle approach. *Annals of Statistics*, 43(3):1178–1242, 2015.

[32] O. Lepski, N. Serdyukova, et al. Adaptive estimation under single-index constraint in a regression model. *The Annals of Statistics*, 42(1):1–28, 2014.

[33] O. V. Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1990.

[34] O. V. Lepskii. Asymptotically minimax adaptive estimation: I. Upper bounds. Optimally adaptive estimates. *Theory of Probability and Applications*, 36:682–697, 1991.

[35] O. V. Lepskii. Asymptotically minimax adaptive estimation: II. Statistical model without adaptation. *Theory of Probability and Applications*, 37:433–468, 1992.

[36] O. V. Lepskii. A problem of adaptive estimation in Gaussian white noise. *Advances in Soviet Mathematics*, 12:87–106, 1992.

[37] S. Mahalanabis and D. Stefankovic. Density estimation in linear time. In *21st Annual Conference in Learning Theory - COLT 2008*, pages 503–512, 2008. arXiv preprint arXiv:0712.2869.

[38] P. Rigollet. Kullback–leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, 40(2):639–665, 2012.

[39] P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16(3):260–280, 2007.

[40] A. B. Tsybakov. Optimal rates of aggregation. In *Learning theory and kernel machines*, pages 303–313. Springer, 2003.

[41] Y. Yang. Mixing strategies for density estimation. *Annals of Statistics*, pages 75–87, 2000.

[42] Y. Yang et al. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004.

[43] Y. G. Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, pages 768–774, 1985.