On lower bounds for the bias-variance trade-off

Derumigny, Alexis; Schmidt-Hieber, Johannes

DOI
10.1214/23-AOS2279

Publication date
2023

Document Version
Final published version

Published in
Annals of Statistics

Citation (APA)
Derumigny, A., & Schmidt-Hieber, J. (2023). On lower bounds for the bias-variance trade-off. *Annals of Statistics*, *51*(4), 1510-1533. https://doi.org/10.1214/23-AOS2279

The Annals of Statistics
2023, Vol. 51, No. 4, 1510–1533
https://doi.org/10.1214/23-AOS2279

ON LOWER BOUNDS FOR THE BIAS-VARIANCE TRADE-OFF

BY ALEXIS DERUMIGNY[1,a] AND JOHANNES SCHMIDT-HIEBER[2,b]

[1]*Department of Applied Mathematics, Delft University of Technology,* [a]*a.f.f.derumigny@tudelft.nl*
[2]*Department of Applied Mathematics, University of Twente,* [b]*a.j.schmidt-hieber@utwente.nl*

It is a common phenomenon that for high-dimensional and nonparametric statistical models, rate-optimal estimators balance squared bias and variance. Although this balancing is widely observed, little is known whether methods exist that could avoid the trade-off between bias and variance. We propose a general strategy to obtain lower bounds on the variance of any estimator with bias smaller than a prespecified bound. This shows to which extent the bias-variance trade-off is unavoidable and allows to quantify the loss of performance for methods that do not obey it. The approach is based on a number of abstract lower bounds for the variance involving the change of expectation with respect to different probability measures as well as information measures such as the Kullback–Leibler or χ^2-divergence. Some of these inequalities rely on a new concept of information matrices. In a second part of the article, the abstract lower bounds are applied to several statistical models including the Gaussian white noise model, a boundary estimation problem, the Gaussian sequence model and the high-dimensional linear regression model. For these specific statistical applications, different types of bias-variance trade-offs occur that vary considerably in their strength. For the trade-off between integrated squared bias and integrated variance in the Gaussian white noise model, we propose to combine the general strategy for lower bounds with a reduction technique. This allows us to reduce the original problem to a lower bound on the bias-variance trade-off for estimators with additional symmetry properties in a simpler statistical model. In the Gaussian sequence model, different phase transitions of the bias-variance trade-off occur. Although there is a non-trivial interplay between bias and variance, the rate of the squared bias and the variance do not have to be balanced in order to achieve the minimax estimation rate.

1. Introduction.

Can the bias-variance trade-off be avoided, for instance by using machine learning methods in the overparametrized regime? This is currently debated in machine learning. While older work on neural networks mention that “the fundamental limitations resulting from the bias-variance dilemma apply to all nonparametric inference methods, including neural networks” ([18], page 45), the very recent work on overparametrization in machine learning has cast some doubt on the necessity to balance squared bias and variance [1, 26]. While for fixed and moderate growth, the number of parameters in the method (e.g., the number of network parameters in a neural network) can be associated to the bias and the variance of the procedure, resulting in the well-known U-shaped curves for the statistical risk (see, e.g., Figure 2.11 in [20]), such a link cannot be made in the overparametrized regime. But this does not mean that the bias-variance trade-off disappears. In this work, we prove that for standard estimation problems in nonparametric and high-dimensional statistics, there are universal bias-variance trade-offs that cannot be circumvented by any method.

Besides the debate about overparametrization, there are many other good reasons why a better understanding of the bias-variance trade-off is relevant for statistical practice. Even in

Received June 2020; revised March 2023.
MSC2020 subject classifications. 62G05, 62C05, 62C20.
Key words and phrases. Bias-variance decomposition, Cramér–Rao inequality, high-dimensional statistics, minimax estimation, nonparametric estimation.

nonadaptive settings, confidence sets in nonparametric statistics require control on the bias of the centering estimator and often use a slight undersmoothing to make the bias negligible compared to the variance. If rate-optimal estimators with negligible bias would exist, such troubles could be overcome. In some instances, small bias is possible. An important example is the rather subtle debiasing of the LASSO for a class of functionals in the high-dimensional regression model [7, 38, 40]. This shows that the occurrence of the bias-variance trade-off is a highly nontrivial phenomenon.

Finite-dimensional parametric models do typically not exhibit a bias-variance trade-off and there may exist unbiased estimators with finite variance. On the contrary, our results show that for high-dimensional and infinite-dimensional statistical models, unbiased estimators with finite variance are in almost all of the considered settings impossible. The fundamental difference lies in the amount of information per parameter: for parametric models of dimension p, the sample size n is by definition, of a larger order than p, and the statistician has a budget of n/p observations per parameter; on the contrary, for nonparametric models, we have $p > n$ or even $p = +\infty$ and there is simply not enough data to estimate each parameter well using a n/p-fraction of the observations. For example, in the Gaussian white noise model, we observe the process $(Y_x)_x$ satisfying $dY_x = f(x)\,dx + n^{-1/2}\,dW_x$ for an unknown function f. If the regression function f lies in a nonparametric class, it is impossible to transform the data into the form $f(x_0)+$"noise." Instead, one has to rely here on the similarity of the regression function in a small vicinity around x_0, which leads to an unavoidable bias.

Only few theoretical articles exist on lower bounds for the interplay between bias and variance. The major contribution is due to Mark Low [24] proving that the bias-variance trade-off is unavoidable for estimation of functionals in the Gaussian white noise model. The approach relies on a complete characterization of the bias-variance trade-off phenomenon in a parametric Gaussian model via the Cramér–Rao lower bound; see also Section 3 for a more in-depth discussion. Another related result is [29], also considering estimation of functionals but not necessarily in the Gaussian white noise model. It is shown that for any functional κ, a lower bound on the asymptotic deviation probability $\lim_{u\to 0}\liminf_{n\to\infty} P_0^n(c_n|\widehat{\kappa} - \kappa(P_0)| \leq u)$ implies an asymptotic lower bound on variance-like measures of the estimator $\widehat{\kappa}_n$ of $\kappa(P_0)$. In this article, we do not consider such deviation probability and establish direct and nonasymptotic trade-offs between bias and variance. [23] introduces a notion of singular functional estimation problems and proves that for such singular problems, no unbiased estimators with finite variance exist. In the same spirit, [9] shows that the supremum of the variance of an unbiased estimator is infinite if a singular point belongs to the closure of the parameter set. Moreover, it is shown that the difference between biases is lower bounded if the worst-case variance is upper bounded.

In this article, we propose a general strategy to derive lower bounds for the bias-variance trade-off. The key ingredient are general inequalities bounding the change of expectation with respect to different distributions by the variance and information measures such as the total variation, Hellinger distance, Kullback–Leibler divergence and the χ^2-divergence.

As examples, we consider nonparametric estimation in the Gaussian white noise model as well as sparse recovery in the sequence model and the high-dimensional linear regression model. By applying the lower bounds to different statistical models, it is surprising to see different types of bias-variance trade-offs occurring. The weakest type are worst-case scenarios stating that if the bias is small for all parameters, then there exists a potentially different parameter in the parameter space with a large variance and vice versa. For the pointwise estimation in the Gaussian white noise model, the derived lower bounds imply also a stronger version proving that small bias for all parameters will necessarily inflate the variance for all parameters that are in a suitable sense separated away from the boundary of the parameter space.

We also study lower bounds for the trade-off between integrated squared bias and integrated variance in the Gaussian white noise model. In this case, a direct application of the multiple parameter lower bound is rather tricky and we propose instead a two-fold reduction first. The first reduction shows that it is sufficient to prove a lower bound on the bias-variance trade-off in a related sequence model. The second reduction states that it is enough to consider estimators that are constrained by some additional symmetry property. After the reductions, a few lines argument applying the information matrix lower bound is enough to derive a matching lower bound for the trade-off between integrated squared bias and integrated variance.

For function estimation in the Gaussian white noise model, the variance blows up if the estimator is constrained to have a bias decreasing faster than the minimax rate. In the sparse sequence model and the high-dimensional regression model with sparsity $\ll \sqrt{n}$, a different phenomenon occurs. For estimators with bias bounded by constant × minimax rate, the derived lower bounds show that a sufficiently small constant already enforces that the variance must be larger than the minimax rate by a polynomial factor in the sample size. Interestingly, for an estimator achieving the minimax estimation rate, the rate of the variance can be of a smaller order than the rate of the squared bias and, therefore, variance and squared bias do not need to be balanced.

Summarizing the results, for all of the considered models a nontrivial bias-variance trade-off could be established. For some estimation problems, the bias-variance trade-off only holds in a worst-case sense, and on subsets of the parameter space, rate-optimal methods with negligible bias exist. It should also be emphasized that for this work only nonadaptive setups are considered. Adaptation to either smoothness or sparsity induces additional bias. The bias-variance trade-off problem can also be rephrased by asking for the optimal estimation rate if only estimators with, for instance, small bias are allowed. In this sense, the work contributes to the growing literature on optimal estimation rates under constraints on the estimators. So far, major theoretical work has been done for polynomial time computable estimators [2, 3], lower and upper bounds for estimation under privacy constraints [16, 17, 34] and parallelizable estimators under communication constraints [36, 41].

The paper is organized as follows. In Section 2, we provide a number of new abstract lower bounds, where we distinguish between inequalities bounding the change of expectation for two distributions and inequalities involving an arbitrary number of expectations. The subsequent sections of the article study lower and upper bounds for the bias-variance trade-off based on these inequalities. The considered setups range from pointwise estimation in the Gaussian white noise model (Section 3 and Section 5) and a boundary estimation problem (Section 4) to high-dimensional models in Section 6. Section 7 discusses some aspects underlying a formal definition of the bias-variance trade-off and the connection between the approach in this work and minimax lower bounds. All proofs are deferred to the Supplementary Material.

Notation. Whenever the domain D is clear from the context, we write $\|\cdot\|_p$ for the $L^p(D)$-norm. Moreover, $\|\cdot\|_2$ denotes also the Euclidean norm for vectors. We denote by $A^\top$ the transpose of a matrix A. For mathematical expressions involving several probability measures, it is assumed that those are defined on the same measurable space. If P is a probability measure, we write E_P and Var_P for the expectation and variance with respect to P, respectively. For probability measures P_θ depending on a parameter θ, E_θ and $\operatorname{Var}_\theta$ denote the corresponding expectation and variance. Throughout the article, we consider estimators $\widehat{\theta}$ for which the expectation $E_\theta[\widehat{\theta}]$ exists and is finite for all parameters θ in the parameter space. This guarantees that the bias is always well defined. If a random variable X is not square integrable with respect to P, we assign the value $+\infty$ to $\operatorname{Var}_P(X)$. For any finite number of measures $P_1, \ldots, P_M$, defined on the same measurable space, we can find a measure ν dominating all of them (e.g., $\nu := \frac{1}{M}\sum_{j=1}^M P_j$). Henceforth, ν will always denote a dominating measure and p_j stands for the ν-density of P_j. The total variation is $\mathrm{TV}(P,Q) := \frac{1}{2}\int |p(\omega) - q(\omega)|\, d\nu(\omega)$. The squared Hellinger distance is defined as $H(P,Q)^2 := \frac{1}{2}\int(\sqrt{p(\omega)} - \sqrt{q(\omega)})^2\, d\nu(\omega)$ (in the literature sometimes also defined without the factor $1/2$). If P is dominated by Q, the Kullback–Leibler divergence is defined as $\mathrm{KL}(P,Q) := \int \log(p(\omega)/q(\omega))p(\omega)\, d\nu(\omega)$ and the χ^2-divergence is defined as $\chi^2(P,Q) := \int (p(\omega)/q(\omega) - 1)^2 q(\omega)\, d\nu(\omega)$. If P is not dominated by Q, both Kullback–Leibler and χ^2-divergence are assigned the value $+\infty$.

2. General lower bounds on the variance.

2.1. *Lower bounds based on two distributions.* Given an upper bound on the bias, the goal is to find a lower bound on the variance. For parametric models, the natural candidate is the Cramér–Rao lower bound. Given a statistical model with real parameter $\theta \in \Theta \subseteq \mathbb{R}$, and an estimator $\widehat{\theta}$ with bias $B(\theta) := E_\theta[\widehat{\theta}] - \theta$, variance $V(\theta) := \mathrm{Var}_\theta(\widehat{\theta})$ and Fisher information $F(\theta)$, the Cramér–Rao lower bound states that $V(\theta) \geq \frac{(1+B'(\theta))^2}{F(\theta)}$, where $B'(\theta)$ denotes the derivative of the bias with respect to θ. The basic idea is that if the bias is small, we cannot have $B'(\theta) \leq -1/2$ everywhere, so there must be a parameter θ^* such that $V(\theta^*) \geq 1/(4F(\theta^*))$. The constant $-1/2$ could be replaced of course by any other number in $(-1, 0)$. There are various extensions of the Cramér–Rao lower bound to multivariate and semiparametric settings [29]. Although the Cramér–Rao lower bound seems to provide a straightforward path to lower bounds on the bias-variance trade-off, the imposed regularity conditions make this approach problematic for nonparametric and high-dimensional models. For example, when the parameter space is the set of s-sparse vectors, this is not an open set and it is unclear how to define the gradient of the bias function or the Fisher information.

Instead of trying to fix the shortcomings of the Cramér–Rao lower bound for complex statistical models, we derive a number of inequalities that bound the change of expectation with respect to two different distributions by the variance and one of the four standard divergence measures: total variation, Hellinger distance, Kullback–Leibler divergence and the χ^2-divergence. As we will see later, these inequalities are much better suited for nonparametric problems as no notion of differentiability of the distribution with respect to the parameter is required. Moreover, the Cramér–Rao lower bound reappears by taking a suitable limit.

LEMMA 2.1. *Let P and Q be two probability distributions on the same measurable space. Denote by E_P and Var_P the expectation and variance with respect to P and let E_Q and Var_Q be the expectation and variance with respect to Q. Then, for any random variable X,*

$$\frac{(E_P[X] - E_Q[X])^2}{2}\left(\frac{1}{\mathrm{TV}(P,Q)} - 1\right) \leq \mathrm{Var}_P(X) + \mathrm{Var}_Q(X), \tag{1}$$

$$\frac{(E_P[X] - E_Q[X])^2}{4 - 2H^2(P,Q)}\left(\frac{1}{H(P,Q)} - H(P,Q)\right)^2 \leq \mathrm{Var}_P(X) + \mathrm{Var}_Q(X), \tag{2}$$

$$\big(E_P[X] - E_Q[X]\big)^2\left(\frac{1}{\mathrm{KL}(P,Q) + \mathrm{KL}(Q,P)} - \frac{1}{4}\right) \leq \mathrm{Var}_P(X) \vee \mathrm{Var}_Q(X), \tag{3}$$

$$\begin{aligned}\big(E_P[X] - E_Q[X]\big)^2 \leq{}& \chi^2(Q,P)\,\mathrm{Var}_P(X) \\ &\wedge \chi^2(P,Q)\,\mathrm{Var}_Q(X).\end{aligned} \tag{4}$$

The inequality in (4) is known [27], Lemma 2, and can also be viewed as a consequence of the Hammersley–Chapman–Robbins inequality [22], Example 5.2. [29], Lemma 5.3, derives analogous formulas for (2) and (4) with the variance replaced by the second moment. Inequality (2) is derived from [28], Theorem 1. To the best of our knowledge, the inequalities in (1) and (3) have not been stated yet in the literature. A proof is provided in Section A of the Supplementary Material [13].

If one of the information measures is zero, the left-hand side of the corresponding inequality should be assigned the value zero as well. The inequalities are based on different decompositions for $E_P[X] - E_Q[X] = \int X(\omega)(dP(\omega) - dQ(\omega))$. All of them involve an application of the Cauchy–Schwarz inequality. For deterministic X, both sides of the inequalities are zero, and hence we have equality. For (4), the choice $X = dQ/dP$ yields equality and in this case, both sides are $(\chi^2(Q, P))^2$. Another line of related inequalities bound the change of expectations in terms of f-divergences, without involving the variance; see, for instance, [10, 19].

To obtain lower bounds for the variance, our inequalities can be applied similarly as the Cramér–Rao inequality. Indeed, small bias implies that $E_\theta[\widehat{\theta}]$ is close to θ and $E_{\theta'}[\widehat{\theta}]$ is close to θ'. If θ and θ' are sufficiently far from each other, we obtain a lower bound for $|E_\theta[\widehat{\theta}] - E_{\theta'}[\widehat{\theta}]|$ and a fortiori a lower bound for the variance. This argument suggests that the lower bound becomes stronger by picking parameters θ and θ' that are as far as possible away from each other. But then, also the information measures of the distributions P_θ and $P_{\theta'}$ are typically larger, making the lower bounds worse. This shows that an optimal application of the inequalities should balance these two aspects.

Example A.1 of the Supplementary Material [13] illustrates these inequalities in the case of the Gaussian distribution. For other distributions, one of these four divergence measures might be easier to compute and the four inequalities can lead to substantially different lower bounds. For instance, if the measures P and Q are not dominated by each other, the Kullback–Leibler and χ^2-divergence are both infinite but the Hellinger distance and total variation version still produce nontrivial lower bounds. This justifies deriving for each divergence measure a separate inequality. It is also in line with the formulation of the theory on minimax lower bounds (see, for instance, Theorem 2.2 in [37]).

Except for the total variation version, all derived inequalities in Lemma 2.1 are generalizations of the Cramér–Rao lower bound. The Cramér–Rao lower bound appears by taking P and Q to be P_θ and $P_{\theta+\Delta}$ and letting Δ tend to zero. A proof and a variation of Lemma 2.1 for a family of distributions $(P_t)_{t\in[0,1]}$ (Lemma A.2) can be found in Section A of the Supplementary Material [13].

2.2. *Information matrices and lower bound based on multiple distributions.* For minimax lower bounds based on hypotheses tests, it has been observed that lower bounds based on two hypotheses are only rate-optimal in specific settings such as for some functional estimation problems. If the local alternatives surrounding a parameter θ spread over many different directions, estimation of θ becomes much harder. To capture this in the minimax lower bounds, we need instead to reduce the problem to a multiple testing problem involving potentially a large number of tests.

A similar phenomenon occurs also for bias-variance trade-off lower bounds. Given $M + 1$ probability measures $P_0, P_1, \dots, P_M$, the χ^2-version of Lemma 2.1 states that for any $j = 1, \dots, M$, $(E_{P_j}[X] - E_{P_0}[X])^2/\chi^2(P_j, P_0) \le \mathrm{Var}_{P_0}(X)$. If $P_1, \dots, P_M$ describe different directions around P_0 in a suitable information theoretic sense, one would hope that in this case a stronger inequality holds with the sum on the left-hand side, that is, $\sum_{j=1}^M (E_{P_j}[X] - E_{P_0}[X])^2/\chi^2(P_j, P_0) \le \mathrm{Var}_{P_0}(X)$. In a next step, two notions of information matrices are introduced, measuring to which extent $P_1, \dots, P_M$ represent different directions around P_0.

If P_0 dominates $P_1, \dots, P_M$, the χ^2-divergence matrix $\chi^2(P_0, \dots, P_M)$ is defined as the $M \times M$ matrix with (j,k)th entry

$$\chi^2(P_0, \dots, P_M)_{j,k} := \int \frac{dP_j}{dP_0} dP_k - 1.$$

The $M \times M$ Hellinger affinity matrix is defined entrywise by

$$\rho(P_0|P_1, \dots, P_M)_{j,k} := \frac{\int \sqrt{p_j p_k}\, d\nu}{\int \sqrt{p_j p_0}\, d\nu \int \sqrt{p_k p_0}\, d\nu} - 1, \quad j,k = 1, \dots, M.$$

Here and throughout the article, we implicitly assume that the distributions $P_0, \dots, P_M$ are chosen such that the Hellinger affinities $\int \sqrt{p_j p_0}\, d\nu$ are positive and the Hellinger affinity matrix is well defined. This condition is considerably weaker than assuming that P_0 dominates the other measures (which is necessary for finiteness of the χ^2-divergence matrix). These two notions of information matrices are studied in more detail in [12].

For a matrix A, the Moore–Penrose inverse A^+ always exists and satisfies the property $AA^+A = A$ and $A^+AA^+ = A^+$. We can now state the generalization of (4) to an arbitrary number of distributions. The following theorem is proved in Section A of the Supplementary Material [13].

THEOREM 2.2. *For $M \geq 1$, let $P_0, P_1, \dots, P_M$ be probability measures defined on the same probability space, and X be a random variable.*

(i) *Set $\Delta := (E_{P_1}[X] - E_{P_0}[X], \dots, E_{P_M}[X] - E_{P_0}[X])^\top$. If $P_j \ll P_0$ for all $j = 1, \dots, M$, then $\Delta^\top \chi^2(P_0, \dots, P_M)^+ \Delta \leq \operatorname{Var}_{P_0}(X)$, where $\chi^2(P_0, \dots, P_M)^+$ denotes the Moore–Penrose inverse of the χ^2-divergence matrix.*

(ii) *Let $A_\ell := \rho(P_\ell | P_1, \dots, P_{\ell-1}, P_{\ell+1}, \dots, P_M)$. Then, for $M \geq 2$,*

$$\begin{aligned} 2M \sum_{j=1}^M \left(E_j[X] - \frac{1}{M} \sum_{\ell=1}^M E_\ell[X] \right)^2 \\ = \sum_{j,k=1}^M (E_j[X] - E_k[X])^2 \leq 4 \max_{\ell=1,\dots,M} \lambda_1(A_\ell) \sum_{k=1}^M \operatorname{Var}_{P_k}(X), \end{aligned}$$

where $\lambda_1(A_\ell)$ denotes the largest eigenvalue (spectral norm) of the positive semidefinite Hellinger affinity matrix A_ℓ.

Instead of using a finite number of probability measures, it is in principle possible to extend Theorem 2.2 to families of probability measures. The divergence matrices become then operators and the sums have to be replaced by integral operators.

If the χ^2-divergence matrix is diagonal with positive entries on the diagonal, we obtain that $\sum_{j=1}^M (E_{P_j}[X] - E_{P_0}[X])^2 / \chi^2(P_j, P_0) \leq \operatorname{Var}_{P_0}(X)$. It should be observed that because of the sum, this inequality produces better lower bounds than (4).

Theorem 2.2(i) contains the multivariate Cramér–Rao lower bound as a special case; see Section A.3 of the Supplementary Material [13]. The connection to the Cramér–Rao inequality suggests that for a given statistical problem with a p-dimensional parameter space, one should apply Theorem 2.2 with $M = p$. It turns out that for the high-dimensional models discussed in Section 6 below, the number of distributions M will be chosen as $\binom{p-1}{s-1}$ with p the number of parameters and s the sparsity. Depending on the sparsity, this can be much larger than p.

We are aware of two existing inequalities that are related to Theorem 2.2(i). [39, Section 3], in our notation, states that $\sum_{j=1}^M \operatorname{Var}_{P_j}(X) \geq (\sum_{j=1}^M E_{P_j}[X] - E_{P_0}[X])^2 /$

$\sum_{j=1}^{M}\chi^2(P_j,P_0)$ and [30], pageg 330, states that for any $p\geq 1$ and any distributions P, Q on $\mathbb{R}^p$, $\chi^2(P,Q)\geq (E_P[X]-E_Q[X])^\top \operatorname{Cov}_Q(X)^{-1}(E_P[X]-E_Q[X])$, where $\operatorname{Cov}_Q(X)$ denotes the covariance matrix of X under Q. The concept of Fisher Φ-information also generalizes the Fisher information using information measures; see [8, 31]. It is worth mentioning that this notion is not comparable with our approach and only applies to Markov processes.

To apply Theorem 2.2(i), we now introduce several variations. A vector $v=(v_1,\dots,v_M)$ lies in the kernel of the χ^2-divergence matrix if and only if $\sum_{j=1}^{M}v_j(P_j-P_0)=0$ (see Section 3 of [12]). This shows that such a v and the vector Δ must be orthogonal. Thus, Δ is orthogonal to the kernel of $\chi^2(P_0,\dots,P_M)$ and

$$\sum_{j=1}^{M}\big(E_{P_j}[X]-E_{P_0}[X]\big)^2\leq \lambda_1\big(\chi^2(P_0,\dots,P_M)\big)\operatorname{Var}_{P_0}(X), \tag{5}$$

where $\lambda_1(\chi^2(P_0,\dots,P_M))$ denotes the largest eigenvalue (spectral norm) of the χ^2-divergence matrix. Given a symmetric matrix $A=(a_{ij})_{i,j=1,\dots,M}$, the maximum row sum norm is defined as $\|A\|_{1,\infty}:=\max_{i=1,\dots,M}\sum_{j=1}^{M}|a_{ij}|$. For any eigenvalue λ of A with corresponding eigenvector $v=(v_1,\dots,v_M)^\top$ and any $i\in\{1,\dots,M\}$, we have that $\lambda v_i=\sum_{j=1}^{M}a_{ij}v_j$ and, therefore, $|\lambda|\max_{i=1,\dots,M}|v_i|\leq \max_{i=1,\dots,M}\sum_{j=1}^{M}|a_{ij}|\|v\|_\infty$. Therefore, $\|A\|_{1,\infty}$ is an upper bound for the spectral norm and

$$\sum_{j=1}^{M}\big(E_{P_j}[X]-E_{P_0}[X]\big)^2\leq \big\|\chi^2(P_0,\dots,P_M)\big\|_{1,\infty}\operatorname{Var}_{P_0}(X). \tag{6}$$

Whatever variation of Theorem 2.2 is applied to derive lower bounds on the bias-variance trade-off, the key problem is the computation of the information matrix for given probability measures P_{θ_j}, $j=0,\dots,M$ in the underlying statistical model $(P_\theta:\theta\in\Theta)$. Suppose there exists a more tractable statistical model $(Q_\theta:\theta\in\Theta)$ with the same parameter space such that the data in the original model can be obtained by a transformation of the data generated from $(Q_\theta:\theta\in\Theta)$. Theorem 4.1 in the companion paper [12] states a data processing inequality for χ^2-divergence matrices. In the setting considered above, this data processing inequality can be written as a matrix inequality

$$\chi^2(P_{\theta_0},\dots,P_{\theta_M})\leq \chi^2(Q_{\theta_0},\dots,Q_{\theta_M}), \tag{7}$$

where $\leq$ is understood with respect to the partial order on the set of positive semidefinite matrices. We therefore can apply the upper bounds (5) and (6) with $\chi^2(P_{\theta_0},\dots,P_{\theta_M})$ replaced by $\chi^2(Q_{\theta_0},\dots,Q_{\theta_M})$. In Theorem 2.2(i), $\chi^2(P_{\theta_0},\dots,P_{\theta_M})^+$ can be replaced by $\chi^2(Q_{\theta_0},\dots,Q_{\theta_M})^+$ if the matrix $\chi^2(P_{\theta_0},\dots,P_{\theta_M})$ is invertible. A specific application for the combination of general lower bounds and the data processing inequality is given in Section 6.

For various distributions, closed-form expression for the information matrices are derived in [12]. In particular, if $P_j=\mathcal{N}(\theta_j,\sigma^2 I_d)$ with $\theta_j\in\mathbb{R}^d$ and $\sigma>0$, then

$$\chi^2(P_0,P_1,\dots,P_M)_{j,k}=\exp\left(\frac{\langle\theta_j-\theta_0,\theta_k-\theta_0\rangle}{\sigma^2}\right)-1. \tag{8}$$

3. The bias-variance trade-off for pointwise estimation in the Gaussian white noise model.

In the Gaussian white noise model, we observe a random function $Y=(Y_x)_{x\in[0,1]}$, with

$$dY_x=f(x)\,dx+n^{-1/2}\,dW_x, \tag{9}$$

where W is an unobserved standard Brownian motion. The aim is to recover the regression function $f : [0, 1] \to \mathbb{R}$ from the data Y. In this section, the bias-variance trade-off for estimation of $f(x_0)$ with fixed $x_0 \in [0, 1]$ is studied. In Section 5, we will also derive a lower bound for the trade-off between integrated squared bias and integrated variance.

Denote by $\|\cdot\|_2$ the $L^2([0, 1])$-norm. For $f \in L^2([0, 1])$, the likelihood ratio in the Gaussian white noise model is given by Girsanov's formula $dP_f/dP_0(Y) = \exp(n\int_0^1 f(t)\,dY_t - \frac{n}{2}\|f\|_2^2)$. In particular, for $Y \sim P_f$ and for any function $g \in L^2([0, 1])$, we have that

$$\begin{aligned}\frac{dP_f}{dP_g}(Y) &= \exp\Bigl(n\int \bigl(f(x) - g(x)\bigr)\,dY_x - \frac{n}{2}\|f\|_2^2 + \frac{n}{2}\|g\|_2^2\Bigr)\\ &= \exp\Bigl(\sqrt{n}\int \bigl(f(x) - g(x)\bigr)\,dW_x + \frac{n}{2}\|f - g\|_2^2\Bigr)\\ &= \exp\Bigl(\sqrt{n}\|f - g\|_2\xi + \frac{n}{2}\|f - g\|_2^2\Bigr),\end{aligned}$$

with W a standard Brownian motion and $\xi \sim \mathcal{N}(0, 1)$. From this representation, we can easily deduce that $1 - H^2(P_f, P_g) = E_f[(dP_f/dP_g)^{-1/2}] = \exp(-\frac{n}{8}\|f - g\|_2^2)$, $\mathrm{KL}(P_f, P_g) = E_f[\log(dP_f/dP_g)] = \frac{n}{2}\|f-g\|_2^2$ and $\chi^2(P_f, P_g) = E_f[dP_f/dP_g]-1 = \exp(n\|f-g\|_2^2)-1$.

Let $R > 0$, $\beta > 0$ and denote by $\lfloor\beta\rfloor$ the largest integer that is strictly smaller than β. On a domain $D \subseteq \mathbb{R}$, we define the β-Hölder norm by $\|f\|_{\mathscr{C}^\beta(D)} = \sum_{\ell\leq\lfloor\beta\rfloor}\|f^{(\ell)}\|_{L^\infty(D)} + \sup_{x,y\in D, x\neq y}|f^{(\lfloor\beta\rfloor)}(x) - f^{(\lfloor\beta\rfloor)}(y)|/|x - y|^{\beta-\lfloor\beta\rfloor}$, with $L^\infty(D)$ the supremum norm on D and $f^{(\ell)}$ denoting the ℓth (strong) derivative of f for $\ell \leq \lfloor\beta\rfloor$. For $D = [0, 1]$, let $\mathscr{C}^\beta(R) := \{f : [0, 1] \to \mathbb{R} : \|f\|_{\mathscr{C}^\beta([0,1])} \leq R\}$ be the ball of β-Hölder smooth functions $f : [0, 1] \to \mathbb{R}$ with radius R. We also write $\mathscr{C}^\beta(\mathbb{R}) := \{K : \mathbb{R} \to \mathbb{R} : \|K\|_{\mathscr{C}^\beta(\mathbb{R})} < \infty\}$.

To explore the bias-variance trade-off for pointwise estimation in more detail, consider for a moment the kernel smoothing estimator, defined by $\widehat{f}(x_0) = (2h)^{-1}\int_{x_0-h}^{x_0+h} dY_t$. Assume that x_0 is not at the boundary such that $0 \leq x_0 - h$ and $x_0 + h \leq 1$. Bias and variance for this estimator are

$$\begin{aligned}\mathrm{Bias}_f\bigl(\widehat{f}(x_0)\bigr) &= \frac{1}{2h}\int_{x_0-h}^{x_0+h}\bigl(f(u) - f(x_0)\bigr)\,du,\\ \mathrm{Var}_f\bigl(\widehat{f}(x_0)\bigr) &= \frac{1}{2nh}.\end{aligned}$$

While the variance is independent of f, the bias vanishes for large subclasses of f such as, for instance, any function f satisfying $f(x_0 - v) = -f(x_0 + v)$ for all $0 \leq v \leq h$. The largest possible bias over this parameter class is of the order h^β and it is attained for functions that lie on the boundary of $\mathscr{C}^\beta(R)$. Because of this asymmetry between bias and variance, the strongest lower bound on the bias-variance trade-off that we can hope for is that any estimator $\widehat{f}(x_0)$ satisfies an inequality of the form

$$\sup_{f\in\mathscr{C}^\beta(R)}\bigl|\mathrm{Bias}_f\bigl(\widehat{f}(x_0)\bigr)\bigr|^{1/\beta}\inf_{f\in\mathscr{C}^\beta(R)}\mathrm{Var}_f\bigl(\widehat{f}(x_0)\bigr) \gtrsim \frac{1}{n}. \tag{10}$$

Since for fixed x_0, $f \mapsto f(x_0)$ is a linear functional, pointwise reconstruction is a specific linear functional estimation problem. This means in particular that the theory in [24] for arbitrary linear functionals in the Gaussian white noise model applies. We now summarize the implications of this work on the bias-variance trade-off and state the new lower bounds based on the change of expectation inequalities derived in the previous section afterwards.

[24] shows that the bias-variance trade-off for estimation of functionals in the Gaussian white noise model can be reduced to the bias-variance trade-off for estimation of a bounded

mean in a normal location family. If $f \mapsto Lf$ denotes a linear functional, $\widehat{Lf}$ stands for an estimator of Lf, Θ is the parameter space and $w(\varepsilon) := \sup\{|L(f-g)| : \|f-g\|_{L^2[0,1]} \leq \varepsilon,\, f, g \in \Theta\}$ is the so-called modulus of continuity, Theorem 2 in [24] rewritten in our notation states that, if Θ is closed and convex and $\lim_{\varepsilon \downarrow 0} w(\varepsilon) = 0$, then

$$\inf_{\widehat{Lf}: \sup_{f\in\Theta} \operatorname{Var}_f(\widehat{Lf}) \leq V} \sup_{f\in\Theta} \operatorname{Bias}_f(\widehat{Lf})^2 = \frac{1}{4} \sup_{\varepsilon>0} \big(w(\varepsilon) - \sqrt{nV}\varepsilon\big)_+^2, \quad \text{and,}$$

$$\inf_{\widehat{Lf}: \sup_{f\in\Theta} |\operatorname{Bias}_f(\widehat{Lf})| \leq B} \sup_{f\in\Theta} \operatorname{Var}_f(\widehat{Lf}) = \frac{1}{n} \sup_{\varepsilon>0} \varepsilon^{-2} \big(w(\varepsilon) - 2B\big)_+^2,$$

with $(x)_+ := \max(x, 0)$. Moreover, an affine estimator $\widehat{Lf}$ can be found attaining these bounds. For pointwise estimation on Hölder balls, $Lf = f(x_0)$ and $\Theta = \mathscr{C}^\beta(R)$. To find a lower bound for the modulus of continuity in this case, choose $K \in \mathscr{C}^\beta(\mathbb{R})$, $f = 0$ and $g = h^\beta K((x - x_0)/h)$. By Lemma B.1 of the Supplementary Material [13], $g \in \mathscr{C}^\beta(R)$ whenever $R \geq \|K\|_{\mathscr{C}^\beta(\mathbb{R})}$ and by substitution, $\|f - g\|_2 = \|g\|_2 \leq h^{\beta+1/2}\|K\|_2 \leq \varepsilon$ for $h = (\varepsilon/\|K\|_2)^{1/(\beta+1/2)}$. This proves $w(\varepsilon) \geq (\varepsilon/\|K\|_2)^{\beta/(\beta+1/2)} K(0)$. In Section B.1 of the Supplementary Material [13], we show that this further implies

$$\inf_{\widehat{f}(x_0)} \sup_{f\in\mathscr{C}^\beta(R)} \big|\operatorname{Bias}_f\big(\widehat{f}(x_0)\big)\big|^{1/\beta} \sup_{f\in\mathscr{C}^\beta(R)} \operatorname{Var}_f\big(\widehat{f}(x_0)\big) \geq \frac{\gamma_{\mathrm{Low}}(R,\beta)}{n}, \tag{11}$$

where

$$\gamma_{\mathrm{Low}}(R,\beta) := \sup_{K\in\mathscr{C}^\beta(\mathbb{R}): R \geq \|K\|_{\mathscr{C}^\beta(\mathbb{R})}} \frac{(2\beta)^2}{2^{1/\beta}(2\beta+1)^{2+1/\beta}} \frac{K(0)^{2+1/\beta}}{\|K\|_2^2}.$$

The result is comparable to (10) with a supremum instead of an infimum in front of the variance.

We now derive the lower bounds on the bias-variance trade-off for the pointwise estimation problem, that are based on the general framework developed in the previous section. Define

$$\gamma(R,\beta) := \sup_{K\in\mathscr{C}^\beta(\mathbb{R}): K(0)=1} \left(\|K\|_2^{-1}\left(1 - \frac{\|K\|_{\mathscr{C}^\beta(\mathbb{R})}}{R}\right)_+\right)^2.$$

For fixed $\beta > 0$, this quantity is positive if and only if $R > 1$. Indeed, if $R \leq 1$, for any function K satisfying $K(0) = 1$, we have $R \leq 1 \leq \|K\|_\infty \leq \|K\|_{\mathscr{C}^\beta(\mathbb{R})}$ and, therefore, $\|K\|_{\mathscr{C}^\beta(\mathbb{R})}/R \geq 1$, implying $\gamma(R,\beta) = 0$. On the contrary, when $R > 1$, we can take for example $K(x) = \exp(-x^2/A)$ with A large enough such that $1 \leq \|K\|_{\mathscr{C}^\beta(\mathbb{R})} < R$. This shows that $\gamma(R,\beta) > 0$ in this case.

If C is a positive constant and $a \in [0, R)$, define moreover

$$\overline{\gamma}(R,\beta,C,a) := \sup_{K\in\mathscr{C}^\beta(\mathbb{R}): K(0)=1} \left(\|K\|_2^{-1}\left(1 - \frac{\|K\|_{\mathscr{C}^\beta(\mathbb{R})}}{R-a}\right)_+\right)^2 \times \exp\left(-C(R-a)^2 \frac{\|K\|_2^2}{\|K\|_{\mathscr{C}^\beta(\mathbb{R})}^2}\right).$$

Arguing as above, for fixed $\beta > 0$, this quantity is positive if and only if $a + 1 < R$. We can now state the main result of this section.

THEOREM 3.1. *Given $\beta, R, C > 0$ and $x_0 \in [0,1]$, let $\gamma(R,\beta)$ and $\overline{\gamma}(R,\beta,C,a)$ be the constants defined above. Assign to $(+\infty)\cdot 0$ the value $+\infty$.*

(i) *If* $\mathcal{T}=\{\widehat{f}:\sup_{f\in\mathscr{C}^\beta(R)}|\operatorname{Bias}_f(\widehat{f}(x_0))|<1\}$, *then*

$$\inf_{\widehat{f}\in\mathcal{T}}\sup_{f\in\mathscr{C}^\beta(R)}\bigl|\operatorname{Bias}_f\bigl(\widehat{f}(x_0)\bigr)\bigr|^{1/\beta}\sup_{f\in\mathscr{C}^\beta(R)}\operatorname{Var}_f\bigl(\widehat{f}(x_0)\bigr)\ge\frac{\gamma(R,\beta)}{n}. \tag{12}$$

(ii) *Let* $\mathcal{S}(C):=\{\widehat{f}:\sup_{f\in\mathscr{C}^\beta(R)}|\operatorname{Bias}_f(\widehat{f}(x_0))|<(C/n)^{\beta/(2\beta+1)}\}\cap\mathcal{T}$, *then*

$$\inf_{\widehat{f}\in\mathcal{S}(C)}\sup_{f\in\mathscr{C}^\beta(R)}\bigl|\operatorname{Bias}_f\bigl(\widehat{f}(x_0)\bigr)\bigr|^{1/\beta}\inf_{f\in\mathscr{C}^\beta(R)}\frac{\operatorname{Var}_f(\widehat{f}(x_0))}{\overline{\gamma}(R,\beta,C,\|f\|_{\mathscr{C}^\beta})}\ge\frac{1}{n}. \tag{13}$$

Both statements can be easily derived from the abstract lower bounds in Section 2. A full proof is given in Section B of the Supplementary Material [13] where statement (i) is derived from Lemma A.2 and statement (ii) is derived from Lemma 2.1. The first statement quantifies a worst-case bias-variance trade-off that must hold for any estimator. The case that $\sup_{f\in\mathscr{C}^\beta(R)}|\operatorname{Bias}_f(\widehat{f}(x_0))|$ exceeds one is not covered. As it leads to inconsistent mean squared error it is of little interest and, therefore, omitted. The second statement restricts attention to estimators with minimax rate-optimal bias. Because of the infimum, we obtain a lower bound on the variance for any function f. Note that this statement is much stronger than (11) or (12) as it holds for the best-case variance instead of the worst-case variance. Compared with (10), the lower bound depends on the $\mathscr{C}^\beta$-norm of f through $\overline{\gamma}(R,\beta,C,\|f\|_{\mathscr{C}^\beta})$. This quantity becomes large if f is close to the boundary of the Hölder ball. A consequence of (ii) is the uniform bound

$$\inf_{\widehat{f}\in\mathcal{S}(C)}\sup_{f\in\mathscr{C}^\beta(R)}\bigl|\operatorname{Bias}_f\bigl(\widehat{f}(x_0)\bigr)\bigr|^{1/\beta}\inf_{f\in\mathscr{C}^\beta(a)}\operatorname{Var}_f\bigl(\widehat{f}(x_0)\bigr)\ge\frac{\inf_{b\le a}\overline{\gamma}(R,\beta,C,b)}{n}, \tag{14}$$

providing a nontrivial lower bound if $a<R-1$; see Section B.3 of the Supplementary Material [13] for a proof. The established lower bound requires that the radius of the Hölder ball R is sufficiently large. Such a condition is necessary. To see this, suppose $R\le 1$ and consider the estimator $\widehat{f}(x_0)=0$. Notice that for any $f\in\mathscr{C}^\beta(R)$, $|\operatorname{Bias}_f(\widehat{f}(x_0))|=|f(x_0)|\le\|f\|_\infty\le 1$ and $\operatorname{Var}_f(\widehat{f}(x_0))=0$. The left-hand side of the inequality (12) is hence zero and even such a worst-case bias-variance trade-off does not hold.

Thanks to the bias-variance decomposition of the mean squared error, for every estimator $\widehat{f}(x_0)\in\mathcal{T}$,

$$\begin{aligned}\sup_{f\in\mathscr{C}^\beta(R)}\operatorname{MSE}_f\bigl(\widehat{f}(x_0)\bigr)\ge{}&\left(\frac{\gamma(R,\beta)}{n\sup_{f\in\mathscr{C}^\beta(R)}\operatorname{Var}_f(\widehat{f}(x_0))}\right)^{2\beta}\\&\wedge\frac{\gamma(R,\beta)}{n\sup_{f\in\mathscr{C}^\beta(R)}|\operatorname{Bias}_f(\widehat{f}(x_0))|^{1/\beta}},\end{aligned}$$

showing that, in a worst case sense, small bias or small variance increases the mean squared error.

COROLLARY 3.2 (Classical unconstrained minimax rates). *Under the same conditions as Theorem 3.1, we have*

$$\inf_{\widehat{f}}\sup_{f\in\mathscr{C}^\beta(R)}\operatorname{MSE}_f\bigl(\widehat{f}(x_0)\bigr)\ge\left(\frac{\gamma(R,\beta)}{n}\right)^{2\beta/(2\beta+1)}\wedge 1,$$

where the infimum is over all measurable estimators. Moreover, the minimax estimation rate $n^{-2\beta/(2\beta+1)}$ can only be achieved for estimators balancing the rate of the worst-case squared bias and the rate of the worst-case variance.

For nonparametric problems, an estimator can be superefficient for many parameters simultaneously; see [5]. Based on that, one might wonder whether it is possible to take for instance a kernel smoothing estimator and shrink small values to zero such that the variance for the regression function $f = 0$ is of a smaller order but the order of the variance and bias for all other parameters remains the same. Statement (ii) of Theorem 3.1 shows that such constructions are impossible if the Hölder radius R is large enough. This question can be viewed as a bias-variance formulation of the constrained risk problem. In the constrained risk problem, we wonder whether an estimator achieving a faster rate for a fixed parameter will have necessarily a suboptimal rate for some other parameter in the parameter space. For pointwise estimation in nonparametric regression, this was studied in Section B of [4].

The proof of Theorem 3.1 depends on the Gaussian white noise model only through the Kullback–Leibler divergence and χ^2-divergence. This indicates that an analogous result can be proved for other nonparametric models with a similar likelihood geometry. As an example consider the Gaussian nonparametric regression model with fixed and uniform design on [0, 1], that is, we observe $(Y_1, \dots, Y_n)$ with $Y_i = f(i/n) + \varepsilon_i$, $i = 1, \dots, n$ and $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Again, f is the (unknown) regression function and we write P_f for the distribution of the observations with regression function f. By evaluating the Gaussian likelihood, we obtain the well-known explicit expressions $\mathrm{KL}(P_f, P_g) = \frac{n}{2}\|f - g\|_n^2$ and $\chi^2(P_f, P_g) = \exp(n\|f - g\|_n^2) - 1$ where $\|h\|_n^2 := \frac{1}{n}\sum_{i=1}^n h(i/n)^2$ is the empirical $L^2([0, 1])$-norm. Compared to the Kullback–Leibler divergence and χ^2-divergence in the Gaussian white noise model, the only difference is that the $L^2([0, 1])$-norm is replaced here by the empirical $L^2([0, 1])$-norm. These norms are very close for functions that are not too spiky. Thus, by following exactly the same steps as in the proof of Theorem 3.1, a similar lower bound can be obtained for the pointwise loss in the nonparametric regression model.

4. The bias-variance trade-off for support boundary recovery.

Compared to approaches using the Cramér–Rao lower bound, the abstract lower bounds based on information measures have the advantage to be applicable also for irregular models. This is illustrated in this section by deriving lower bounds on the bias-variance trade-off for a support boundary estimation problem.

Consider the model, where we observe a Poisson point process (PPP) $N = \sum_i \delta_{(X_i, Y_i)}$ with intensity $\lambda_f(x, y) = n\mathbf{1}(f(x) \le y)$ in the plane $(x, y) \in [0, 1] \times \mathbb{R}$. Differently speaking, the Poisson point process has intensity n on the epigraph of the function f and zero intensity on the subgraph of f. The unknown function f appears therefore as a boundary if the data are plotted; see Figure 1. Throughout the following, n plays the role of the sample size and we refer to (X_i, Y_i) as the support points of the PPP. Estimation of f is also known as support

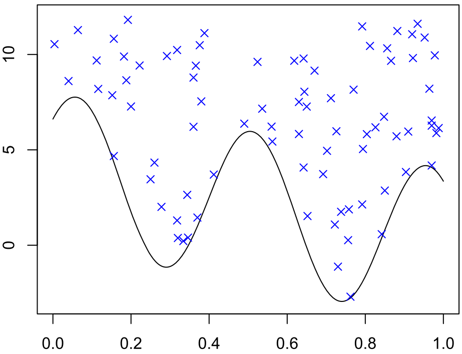

FIG. 1. *Generated data (blue) and support boundary (black) for PPP model*

boundary recovery problem. Similarly, as the Gaussian white noise model is a continuous analogue of the nonparametric regression model with Gaussian errors, the support boundary problem arises as a continuous analogue of the nonparametric regression model with one-sided errors; see [25].

For a parametric estimation problem, we can typically achieve the estimation rate n^{-1} in this model. For squared loss, this becomes n^{-2}. The n^{-1} rate is to be contrasted with the classical $n^{-1/2}$ rate in regular parametric models. Also, for nonparametric problems, faster rates can be achieved. If β denotes the Hölder smoothness of the support boundary f, the optimal MSE for estimation of $f(x_0)$ is $n^{-2\beta/(\beta+1)}$, which can be considerably faster than the typical nonparametric rate $n^{-2\beta/(2\beta+1)}$, [32]. The following theorem is proved in Section C of the Supplementary Material [13] applying the χ^2-divergence version of Lemma 2.1.

THEOREM 4.1. *Let $0 < \beta < 1$, $C > 0$ and $R > \kappa := 2\inf\{\|K\|_{\mathscr{C}^\beta(\mathbb{R})} : K \in L^2(\mathbb{R}), K(0) = 1, K \geq 0\}$.*

For any estimator $\widehat{f}$ with $\sup_{f\in\mathscr{C}^\beta(R)} \mathrm{MSE}_f(\widehat{f}(x_0)) < (C/n)^{2\beta/(\beta+1)}$, there exist positive constants $c := c(\beta, C, R)$ and $c' := c'(\beta, C, R)$ such that

$$\sup_{f\in\mathscr{C}^\beta(R)} \mathrm{Bias}_f\big(\widehat{f}(x_0)\big)^2 \geq cn^{-\frac{2\beta}{\beta+1}}, \quad \textit{and} \tag{15}$$

$$\mathrm{Var}_f\big(\widehat{f}(x_0)\big) \geq c'n^{-\frac{2\beta}{\beta+1}}, \quad \textit{for all } f \in \mathscr{C}^\beta\big((R-\kappa)/2\big). \tag{16}$$

The result shows that any estimator achieving the optimal $n^{-2\beta/(\beta+1)}$ MSE rate must also have worst-case squared bias of the same order. Moreover, no superefficiency is possible for functions that are not too close to the boundary of the Hölder ball. Indeed the variance and, therefore, also the mean squared error, is always lower-bounded by $\gtrsim n^{-2\beta/(\beta+1)}$. The smoothness constraint $\beta \leq 1$ is fairly common in the literature on support boundary estimation; see [33].

5. The trade-off between integrated bias and integrated variance in the Gaussian white noise model.

All lower bounds so far are based on change of expectation inequalities. In this section, we combine this with a different proving strategy for bias-variance lower bounds based on two types of reduction. First, one can in some cases relate the bias-variance trade-off in the original model to the bias-variance trade-off in a simpler model. We refer to this as model reduction. The second type of reduction constraints the class of estimators by showing that it is sufficient to consider estimators satisfying additional symmetry properties.

To which extent such reductions are possible is highly dependent on the structure of the underlying problem. In this section, we illustrate the approach deriving a lower bound on the trade-off between the integrated squared bias (IBias^2) and the integrated variance (IVar) in the Gaussian white noise model (9). Recall that the mean integrated squared error (MISE) can be decomposed as

$$\begin{aligned} \mathrm{MISE}_f(\widehat{f}) &:= E_f\big[\|\widehat{f} - f\|_{L^2[0,1]}^2\big] \\ &= \int_0^1 \mathrm{Bias}_f^2\big(\widehat{f}(x)\big)\,dx + \int_0^1 \mathrm{Var}_f\big(\widehat{f}(x)\big)\,dx \\ &=: \mathrm{IBias}_f^2(\widehat{f}) + \mathrm{IVar}_f(\widehat{f}). \end{aligned} \tag{17}$$

To establish a trade-off between integrated bias and integrated variance, turns out to be a hard problem. In particular, we cannot simply integrate the pointwise lower bounds. Below we explain the major reduction steps to prove a lower bound. To avoid unnecessary technicalities involving the Fourier transform, we only consider integer smoothness $\beta = 1, 2, \ldots$

and denote by $S^\beta(R)$ the ball of radius R in the L^2-Sobolev space with index β on $[0,1]$, that is, all L^2-functions satisfying $\|f\|_{S^\beta([0,1])} \leq R$, where for a general domain D, $\|f\|_{S^\beta(D)}^2 := \|f\|_{L^2(D)}^2 + \|f^{(\beta)}\|_{L^2(D)}^2$. Define

$$\Gamma_\beta := \inf\{\|K\|_{S^\beta} : \|K\|_{L^2(\mathbb{R})} = 1, \operatorname{supp} K \subset [-1/2, 1/2]\}. \tag{18}$$

THEOREM 5.1. *Consider the Gaussian white noise model* (9) *with parameter space $S^\beta(R)$ and β a positive integer. If $R > 2\Gamma_\beta$ and $0 \cdot (+\infty)$ is assigned the value $+\infty$, then*

$$\inf_{\widehat{f} \in T} \sup_{f \in S^\beta(R)} \left|\mathrm{IBias}_f(\widehat{f})\right|^{1/\beta} \sup_{f \in S^\beta(R)} \mathrm{IVar}_f(\widehat{f}) \geq \frac{1}{8n}, \tag{19}$$

with $T := \{\widehat{f} : \sup_{f \in S^\beta(R)} \mathrm{IBias}_f^2(\widehat{f}) < 2^{-\beta}\}$.

As in the pointwise case, estimators with larger bias are of little interest as they will lead to procedures that are inconsistent with respect to the MISE. Thanks to the bias-variance decomposition of the MISE (18), for every estimator $\widehat{f} \in T$ the following lower bound on the MISE holds:

$$\sup_{f \in S^\beta(R)} \mathrm{MISE}_f(\widehat{f}) \geq \left(\frac{1}{8n \sup_{f \in S^\beta(R)} \mathrm{IVar}_f(\widehat{f})}\right)^{2\beta} \vee \frac{1}{8n \sup_{f \in S^\beta(R)} |\mathrm{IBias}_f(\widehat{f})|^{1/\beta}}.$$

Small worst-case bias or variance will therefore automatically enforce a large MISE. This provides a lower bound for the widely observed U-shaped bias-variance trade-off and shows in particular that $n^{-2\beta/(2\beta+1)}$ is a lower bound for the minimax estimation rate with respect to the MISE.

COROLLARY 5.2 (Classical unconstrained minimax rates). *Under the same conditions as Theorem 5.1, we have*

$$\inf_{\widehat{f}} \sup_{f \in S^\beta(R)} \mathrm{MISE}_f(\widehat{f}) \geq \left(\frac{1}{8n}\right)^{2\beta/(2\beta+1)} \wedge 1,$$

where the infimum is over all measurable estimators. Moreover, the minimax rate $n^{-2\beta/(2\beta+1)}$ can only be achieved for estimators balancing the rates of the worst-case integrated squared bias and the worst-case integrated variance.

If applied to functions, recall that $\|\cdot\|_p$ denotes the $L^p([0,1])$-norm. Let $p \geq 2$. Since $\|\cdot\|_2 \leq \|\cdot\|_p$, another direct consequence of the previous theorem is

$$\sup_{f \in S^\beta(R)} \|E_f[\widehat{f}] - f\|_p^{1/\beta} \sup_{f \in S^\beta(R)} E_f[\|\widehat{f} - E_f[\widehat{f}]\|_p]^2 \geq \frac{1}{8n},$$

for any estimator with $\sup_{f \in S^\beta(R)} \|E_f[\widehat{f}] - f\|_p < 2^{-\beta}$.

We now sketch the main reduction steps in the proof of Theorem 5.1. The first step is a model reduction to a Gaussian sequence model

$$X_i = \theta_i + \frac{1}{\sqrt{n}}\varepsilon_i, \quad i = 1, \ldots, m \tag{20}$$

with independent noise $\varepsilon_i \sim \mathcal{N}(0,1)$. For any estimator $\widehat{\theta}$ of the parameter vector $\theta = (\theta_1, \dots, \theta_m)^\top$, we have the bias-variance type decomposition

$$E_\theta\big[\|\widehat{\theta} - \theta\|_2^2\big] = \big\|E_\theta[\widehat{\theta}] - \theta\big\|_2^2 + \sum_{i=1}^m \operatorname{Var}_\theta(\widehat{\theta}_i)$$

recalling that $\|\cdot\|_2$ denotes the Euclidean norm if applied to vectors.

PROPOSITION 5.3. *Let m be a positive integer and let Γ_β be defined as in (18). Then, for any estimator $\widehat{f}$ of the regression function f in the Gaussian white noise model (9) with parameter space $S^\beta(R)$, there exists a nonrandomized estimator $\widehat{\theta}$ in the Gaussian sequence model with parameter space $\Theta_m^\beta(R) := \{\theta : \|\theta\|_2 \leq R/(\Gamma_\beta m^\beta)\}$, such that*

$$\sup_{\theta \in \Theta_m^\beta(R)} \big\|E_\theta[\widehat{\theta}] - \theta\big\|_2^2 \leq \sup_{f \in S^\beta(R)} \operatorname{IBias}_f^2(\widehat{f}), \quad \textit{and}$$

$$\sup_{\theta \in \Theta_m^\beta(R)} \sum_{i=1}^m \operatorname{Var}_\theta(\widehat{\theta}_i) \leq \sup_{f \in S^\beta(R)} \operatorname{IVar}_f(\widehat{f}).$$

A proof is given in Section D of the Supplementary Material [13]. The rough idea is to restrict the parameter space $S^\beta(R)$ to a suitable ball in an m-dimensional subspace. Denoting the m parameters in this subspace by $\theta_1, \dots, \theta_m$, every estimator $\widehat{f}$ for the regression function induces an estimator for $\theta_1, \dots, \theta_m$ by projection on this subspace. It has then to be checked that the projected estimator can be identified with an estimator $\widehat{\theta}$ in the sequence model and that the projection does not increase squared bias and variance.

Proposition 5.3 reduces the original problem to deriving lower bounds on the bias-variance trade-off in the sequence model (20) with parameter space $\Theta_m^\beta(R)$. Observe that $X = (X_1, \dots, X_m)$ is an unbiased estimator for θ. The existence of unbiased estimators suggests that the reduction to the Gaussian sequence model is unsuitable for deriving lower bounds as it destroys the original bias-variance trade-off. This is, however, not true as the bias will be induced through the choice of m. Indeed, to prove Theorem 5.1, m is chosen such that $m^{-\beta}$ is proportional to the worst-case bias and it is shown that the worst-case variance in the sequence model is lower-bounded by m/n. Rewriting m in terms of the bias yields finally a lower bound of form (19).

To obtain bias-variance lower bounds in the sequence model (20) is, however, still a very difficult problem as superefficient estimators exist with simultaneously small bias and variance for some parameters. An example is the James–Stein estimator $\widehat{\theta}_{\mathrm{JS}} := (1 - (m-2)/(n\|X\|_2^2))X$ with $X = (X_1, \dots, X_m)^\top$ for $m > 2$. While its risk $E_\theta[\|\widehat{\theta} - \theta\|_2^2] = \|E_\theta[\widehat{\theta}] - \theta\|_2^2 + \sum_{i=1}^m \operatorname{Var}_\theta(\widehat{\theta}_i)$ is upper bounded by m/n for all $\theta \in \mathbb{R}^m$, the risk for the zero vector $\theta = (0, \dots, 0)^\top$ is bounded by the potentially much smaller value $2/n$ (see Proposition 2.8 in [21]). Thus, for the zero parameter vector both $\|E_\theta[\widehat{\theta}] - \theta\|_2^2$ and $\sum_{i=1}^m \operatorname{Var}_\theta(\widehat{\theta}_i)$ are simultaneously small. Furthermore, for any parameter vector θ^* there exists an estimator $\widehat{\theta}$ with small bias and variance at θ^*. For instance, the shifted James–Stein estimator $\widehat{\theta}_{\mathrm{JS},\theta^*} := (1 - (m-2)/(n\|X - \theta^*\|_2^2))(X - \theta^*) + \theta^*$ has this property. This suggests that fixing a number of parameters in the neighborhood of some θ^* and applying an abstract lower bound that applies to all estimators $\widehat{\theta}$ will always lead to a suboptimal rate in this lower bound.

Instead, we will first show that it is sufficient to study a smaller class of estimators with additional symmetry properties. Denote by $\mathcal{O}_m$ the class of $m \times m$ orthogonal matrices. For any $D \in \mathcal{O}_m$, $D\theta \in \Theta_m^\beta(R)$ and $DX \sim \mathcal{N}(D\theta, I_m/n)$. Therefore, the model is rotation-invariant [22], Chapter 3. Following Stein [35], we say that a function $f : \mathbb{R}^m \to \mathbb{R}^m$ is spherically symmetric if for any $x \in \mathbb{R}^m$ and any $D \in \mathcal{O}_m$, $f(x) = D^{-1} f(Dx)$. An estimator $\widehat{\theta} = \widehat{\theta}(X)$ is called spherically symmetric if $X \mapsto \widehat{\theta}(X)$ is spherically symmetric. In particular, the James–Stein estimator $\widehat{\theta}_{\mathrm{JS}}$ is spherically symmetric but, unless $\theta^* = 0$, the shifted James–Stein estimator $\widehat{\theta}_{\mathrm{JS},\theta^*}$ is not. The discussion above suggests that if we can reduce the class of estimators to spherically symmetric estimators, all parameters with both small bias and variance must be close to the origin. We can then apply one of the abstract lower bounds to probability measures $P_{\theta_0}, \dots, P_{\theta_M}$ with $\theta_0, \dots, \theta_M$ suitably chosen parameter vectors in the neighborhood of some θ^* that is far enough away from the origin.

This proof strategy works. In a first step, we show the reduction to spherically symmetric estimators.

PROPOSITION 5.4. *Consider the sequence model* (20) *with parameter space* $\Theta_m^\beta(R)$. *For any estimator* $\widehat{\theta}$, *there exists a spherically symmetric estimator* $\widetilde{\theta}$ *such that*

$$\sup_{\theta \in \Theta_m^\beta(R)} \|E_\theta[\widetilde{\theta}] - \theta\|_2^2 \le \sup_{\theta \in \Theta_m^\beta(R)} \|E_\theta[\widehat{\theta}] - \theta\|_2^2, \quad \textit{and},$$

$$\sup_{\theta \in \Theta_m^\beta(R)} \sum_{i=1}^m \mathrm{Var}_\theta(\widetilde{\theta}_i) \le \sup_{\theta \in \Theta_m^\beta(R)} \sum_{i=1}^m \mathrm{Var}_\theta(\widehat{\theta}_i).$$

The main idea of the proof is to define $\widetilde{\theta}$ as a spherically symmetrized version of $\widehat{\theta}$.

To establish lower bounds, it is therefore sufficient to consider spherically symmetric estimators. It has been mentioned in [35] that any spherically symmetric function h is of the form $h(x) = r(\|x\|_2)x$, for some real-valued function r. In Lemma D.1 in the Supplementary Material, we provide a more detailed proof of this fact. Using this property, we can then also show that if $\widetilde{\theta}(X)$ is a spherically symmetric estimator, the expectation map $\theta \mapsto E_\theta[\widetilde{\theta}(X)]$ is a spherically symmetric function. To see this, rewrite $\widetilde{\theta}(X) = s(\|X\|_2)X$ and define $\phi(u) := (2\pi/n)^{-m/2} \exp(-nu^2/2)$. Substituting $y = D^{-1}x$ and noticing that the determinant of the Jacobian matrix of this transformation is one since D is orthogonal, we obtain

$$\begin{aligned} E_{D\theta}[\widetilde{\theta}(X)] &= \int s(\|x\|_2)x\phi(\|x - D\theta\|_2)\,dx \\ &= \int s(\|D^{-1}x\|_2)x\phi(\|D^{-1}x - \theta\|_2)\,dx \\ &= \int s(\|y\|_2)Dy\phi(\|y - \theta\|_2)\,dy = DE_\theta[\widetilde{\theta}(X)]. \end{aligned} \tag{21}$$

Together with Lemma D.1 of the Supplementary Material [13], this implies that there exists a function t such that for any θ, $E_\theta[\widetilde{\theta}(X)] = t(\|\theta\|_2)\theta$, and hence

$$\|E_\theta[\widetilde{\theta}(X)] - \theta\|_2^2 = \|t(\|\theta\|_2)\theta - \theta\|_2^2 = \|\theta\|_2^2(t(\|\theta\|_2) - 1)^2. \tag{22}$$

Based on these reductions, we can now prove Theorem 5.1 by applying the change of expectation inequality in Theorem 2.2(i). The details can be found in Section D of the Supplementary Material [13].

6. The bias-variance trade-off for high-dimensional models with sparsity constraints.

In the Gaussian sequence model, we observe n independent random variables $X_i \sim \mathcal{N}(\theta_i, 1)$. The space of s-sparse signals $\Theta(s)$ is the collection of all vectors $(\theta_1, \dots, \theta_n)$ with at most

s nonzero components. For any estimator $\widehat{\theta}$, the bias-variance decomposition of the mean squared error of $\hat{\theta}$ is

$$E_\theta\big[\|\widehat{\theta}-\theta\|_2^2\big] = \big\|E_\theta[\widehat{\theta}]-\theta\big\|_2^2 + \sum_{i=1}^n \operatorname{Var}_\theta(\widehat{\theta}_i), \tag{23}$$

where the first term on the right-hand side plays the role of the squared bias. For this model, it is known that the exact minimax risk is $2s\log(n/s)$ up to smaller-order terms and that the risk is attained by a soft-thresholding estimator [15]. This estimator exploits the sparsity by shrinking small values to zero. Shrinkage obviously causes some bias but at the same time reduces the variance for sparse signals. We now show that there is indeed a non-trivial bias-variance trade-off both for estimation of the full vector θ and for estimation of the quadratic functional $\theta \mapsto \|\theta\|_2^2$. The two main results of this section are stated next.

THEOREM 6.1. *Consider the Gaussian sequence model with sparsity $s \ll \sqrt{n}$. Any estimator $\widehat{\theta}$ that attains the minimax estimation rate $s\log(n)$ with respect to the worst case risk $\sup_{\theta\in\Theta(s)} E_\theta[\|\widehat{\theta}-\theta\|_2^2]$ also satisfies for all sufficiently large n,*

$$\sup_{\theta\in\Theta(s)} \big\|E_\theta[\widehat{\theta}]-\theta\big\|_2^2 \asymp s\log(n), \quad \textit{and} \quad \sup_{\theta\in\Theta(s)} \sum_{i=1}^n \operatorname{Var}_\theta(\widehat{\theta}_i) \geq \frac{s}{2}.$$

Moreover, if $s \leq n^{1/2-\delta}$ for some $0<\delta<1/2$, then there exists an estimator attaining the minimax estimation rate with $\sup_{\theta\in\Theta(s)} \sum_{i=1}^n \operatorname{Var}_\theta(\widehat{\theta}_i) \lesssim s$.

The result shows that for a minimax rate optimal estimator, squared bias and variance do not necessarily need to be of the same order and the rate of the variance can be slower by at most a $\log(n)$-factor.

One might wonder whether the proposed lower bound technique can be extended for sparsity $s \gg \sqrt{n}$. While this question remains open, we now prove that for estimation of the quadratic functional a phase transition occurs if the sparsity is of the order $\sqrt{n}$. For sparsity $s \ll \sqrt{n},$ the bias-variance trade-off is nontrivial, but for sparsity $s \gtrsim \sqrt{n}$, we can find an unbiased estimator achieving the minimax estimation rate.

For estimation of the quadratic functional, consider the parameter space

$$\Theta_n^2(s) := \Theta(s) \cap \left\{\theta : \sum_{i=1}^n \theta_i^2 \leq 2s\log\left(1+\frac{\sqrt{n}}{s}\right)\right\}. \tag{24}$$

Those are all s-sparse vectors with squared Euclidean norm bounded by $2s\log(1+\sqrt{n}/s)$. We have chosen this specific threshold as it leads to the most unusual behavior of the bias-variance trade-off. For this parameter space, the minimax estimation rate for the functional $\theta \mapsto \|\theta\|_2^2$ with respect to the MSE is

$$s^2\log^2\left(1+\frac{\sqrt{n}}{s}\right) \asymp s^2\log^2\left(\frac{n}{s^2}\right) \vee n \tag{25}$$

as stated in [11], Corollary 1. See also Section E of the Supplementary Material [13] for more details about (25).

THEOREM 6.2. *Consider estimation of the functional $\theta \mapsto \|\theta\|_2^2$ in the Gaussian sequence model with sparsity s and parameter space $\Theta_n^2(s)$.*

(i) *If $s \ll \sqrt{n}$, then the minimax estimation rate is $s^2 \log^2(n/s^2)$ and any estimator $\widehat{\|\theta\|_2^2}$ attaining the minimax optimal estimation rate must satisfy*

$$\sup_{\theta \in \Theta_n^2(s)} \big(E_\theta[\widehat{\|\theta\|_2^2}] - \|\theta\|_2^2\big)^2 \asymp s^2 \log^2\Big(\frac{n}{s^2}\Big)$$

for all sufficiently large n. Moreover, if $s \le n^{1/2-\delta}$ for some $0 < \delta < 1/2$, then there exists a minimax rate optimal estimator $\widehat{\|\theta\|_2^2}$ with $\sup_{\theta \in \Theta_n^2(s)} \operatorname{Var}_\theta(\widehat{\|\theta\|_2^2}) \lesssim s \log(n/s^2)$.

(ii) *If $s \gtrsim \sqrt{n}$, then there exists a minimax rate optimal estimator that is unbiased.*

For sparsity of the order $o(\sqrt{n})$, every minimax rate optimal estimator will have necessarily a worst case squared bias that is of the same order as the minimax rate. But worst case squared bias and variance do not have to be of the same order if $s \to \infty$. Indeed, the second part of (i) shows existence of a minimax rate optimal estimator with variance $s \log(n/s^2) \ll s^2 \log^2(n/s^2) =$ minimax estimation rate.

Surprisingly, there is a phase transition if s is of the order $\sqrt{n}$. If $s \gtrsim \sqrt{n}$, suddenly unbiased estimation is possible, which means that now the variance is dominating the risk.

That typically either squared bias or variance dominates seems to be symptomatic for estimation of functionals. For instance, for estimation of the squared functional $f \mapsto \int f^2$ in the Gaussian white noise model, we conjecture that if the Hölder smoothness of f is below $1/4$, the squared bias will dominate, whereas for smoothness indices above $1/4$, the convergence rate is driven in first order by the variance.

Below we analyze the two main results above in more detail. Since the bias-variance lower bounds are very different from the ones in the previous chapters, we discuss the lower bounds on the variance and the lower bounds on the bias in separate subsections. All proofs of this section are deferred to Section E of the Supplementary Material [13].

Lower bounds on the variance. Using the lower bound technique based on multiple probability distributions, we can derive a lower bound for the variance at zero of any estimator that satisfies a bound on the bias.

THEOREM 6.3. *Consider the Gaussian sequence model with sparsity $0 < s \le \sqrt{n}/2$. Given an estimator $\widehat{\theta}$ and a real number γ such that $4\gamma + 1/\log(n/s^2) \le 0.99$ and*

$$\sup_{\theta \in \Theta(s)} \|E_\theta[\widehat{\theta}] - \theta\|_2^2 \le \gamma s \log\Big(\frac{n}{s^2}\Big),$$

then, for all sufficiently large n,

$$\sum_{i=1}^n \operatorname{Var}_0(\widehat{\theta}_i) \ge \frac{(1 - (1/2)^{0.01})}{25e \log(n/s^2)} n \Big(\frac{s^2}{n}\Big)^{4\gamma},$$

where Var_0 denotes the variance for parameter vector $\theta = (0, \dots, 0)^\top$.

Compared to pointwise estimation, the result shows a different type of bias-variance trade-off. Decreasing the constant γ in the upper bound for the bias, increases the rate in the lower bound for the variance. For instance, in the regime $s \le n^{1/2-\delta}$, with $0 < \delta < 1/2$, we can find for any $\rho > 0$ a sufficiently small constant γ, such that the lower bound is of the form constant $\times n^{1-\rho}$. As a consequence of the bias-variance decomposition (23), the maximum quadratic risk of such an estimator in this regime is also lower bounded by $\gtrsim n^{1-\rho}$. Reducing the constant of the bias will therefore necessarily lead to estimators with highly suboptimal estimation risk.

The proof of Theorem 6.3 applies the χ^2-divergence lower bound (6) by comparing the data distribution induced by the zero vector to the $\binom{n}{s}$ many distributions corresponding to s-sparse vectors with nonzero entries $\sqrt{4\gamma\log(n/s^2)+1}$. By (8), the size of the (j,k)th entry of the χ^2-divergence matrix is completely described by the number of components on which the corresponding s-sparse vectors are both nonzero. The whole problem reduces then to a combinatorial counting argument. The key observation is that if we fix an s-sparse vector, say θ^*, there are of the order n/s^2 more s-sparse vectors that have exactly $r-1$ nonzero components in common with θ^* than s-sparse vectors that that have exactly r nonzero components in common with θ^*. This means that as long as $s \ll \sqrt{n}$, most of the s-sparse vectors are (nearly) orthogonal to θ^*.

The lower bound in Theorem 6.3 can be extended to several related problems by invoking the data processing inequality (7). As an example, suppose that we observe only $X_1^2, \dots, X_n^2$ with $(X_1, \dots, X_n)$ the data from the Gaussian sequence model. As parameter space, consider the class $\Theta_+(s)$ of s-sparse vectors with nonnegative entries. This choice is natural as the parameter θ is not identifiable in this model over the full space of s-sparse vectors $\Theta(s)$. Since the proof of Theorem 6.3 only uses parameters in $\Theta_+(s)$, the same lower bound as in Theorem 6.3 holds also in this modified setting. The next result shows an analogous version of Theorem 6.3 for estimation of the functional $\theta \mapsto \|\theta\|_2^2$.

THEOREM 6.4. *Consider the Gaussian sequence model with parameter space* $\Theta_n^2(s)$ *defined in* (24) *and sparsity* $0 < s \le \sqrt{n}/2$. *Given an estimator* $\widehat{\|\theta\|_2^2}$ *of* $\|\theta\|_2^2$ *and a real number* γ *such that* $2\gamma + 1/\log(n/s^2) \le 0.99$ *and*

$$\sup_{\theta\in\Theta_n^2(s)} \left|\mathrm{Bias}_\theta\big(\widehat{\|\theta\|_2^2}\big)\right| \le \gamma s \log\left(\frac{n}{s^2}\right),$$

then, for all sufficiently large n,

$$\mathrm{Var}_0\big(\widehat{\|\theta\|_2^2}\big) \ge \frac{1-(1/2)^{0.01}}{e} n \left(\frac{s^2}{n}\right)^{2\gamma},$$

where Var_0 *denotes the variance for parameter vector* $\theta = (0, \dots, 0)^\top$.

Notice that the upper bound in the previous result is for the bias, not the squared bias.

A lower bound for the bias. What can be said about the bias for small variance? The next result shows that if the variance is strictly smaller than $s/2$, then the worst case bias is infinite.

THEOREM 6.5. *Consider the Gaussian sequence model with sparsity* $1 \le s \le n$ *and assume that* $\widehat{\theta} = (\widehat{\theta}_1, \dots, \widehat{\theta}_n)$ *is an estimator such that* $E_\theta[\widehat{\theta}_i]$ *exists and is finite for all* $i = 1, \dots, n$ *and all* $\theta \in \Theta(s)$.

If $\sup_{\theta\in\Theta(s)} \sum_{i=1}^n \mathrm{Var}_\theta(\widehat{\theta}_i) < s/2$, *then* $\sup_{\theta\in\Theta(s)} \|E_\theta[\widehat{\theta}] - \theta\|_2 = \infty$.

Nearly matching upper bounds. To show that the rates in the derived lower bounds are nearly sharp, we now establish corresponding upper bounds. For an estimator thresholding small observations, the variance under P_0 is determined by both the probability that an observation falls outside the truncation level and the value it is then assigned to. One can further reduce the variance at zero if large observations are shrunk as much as possible to zero. The bound on the bias dictates the largest possible truncation level. To obtain matching upper bounds, this motivates then to study the soft-thresholding estimator

$$\widehat{\theta}_i = \mathrm{sign}(X_i)\big(|X_i| - \sqrt{\gamma\log(n/s^2)}\big)_+, \quad i = 1, \dots, n. \tag{26}$$

If $\theta_i = 0$, then $E_\theta[\widehat{\theta}_i] = 0$. For $\theta_i \neq 0$, one can use $|\widehat{\theta}_i - X_i| \leq \sqrt{\gamma \log(n/s^2)}$ and $E_\theta[X_i] = \theta_i$ to verify that the squared bias $\|E_\theta[\widehat{\theta}] - \theta\|_2^2$ is bounded by $\gamma s \log(n/s^2)$, uniformly over the space of s-sparse vectors $\Theta(s)$. As an estimator for the functional $\|\theta\|_2^2$, we study

$$\widehat{\|\theta\|_2^2} = \sum_{i=1}^{n} \big((X_i^2 - \gamma \log(n/s^2))_+ - E_{\xi\sim\mathcal{N}(0,1)}[(\xi^2 - \gamma \log(n/s^2))_+]\big). \tag{27}$$

LEMMA 6.6. *For the soft-thresholding estimator $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)^\top$ defined in* (26), *we have*

$$\sum_{i=1}^{n} \operatorname{Var}_0(\widehat{\theta}_i) \leq \frac{\sqrt{2}}{\sqrt{\pi\gamma^3 \log^3(n/s^2)}} n\left(\frac{s^2}{n}\right)^{\frac{\gamma}{2}}, \quad \textit{and}, \tag{28}$$

$$\textit{for any } \theta \in \Theta(s), \quad \sum_{i=1}^{n} \operatorname{Var}_\theta(\widehat{\theta}_i) \leq 4s + \frac{\sqrt{2}}{\sqrt{\pi\gamma^3 \log^3(n/s^2)}} n\left(\frac{s^2}{n}\right)^{\frac{\gamma}{2}}. \tag{29}$$

Moreover, for any n, s, γ, for which $\gamma \log(n/s^2) \geq 2$, we have for the estimator $\widehat{\|\theta\|_2^2}$ defined in (27),

$$\sup_{\theta\in\Theta(s)} \left|\operatorname{Bias}_\theta(\widehat{\|\theta\|_2^2})\right| \leq \gamma s \log\left(\frac{n}{s^2}\right), \tag{30}$$

$$\operatorname{Var}_0(\widehat{\|\theta\|_2^2}) \leq \frac{8}{\sqrt{\gamma \log(n/s^2)}} n\left(\frac{s^2}{n}\right)^{\frac{\gamma}{2}}, \quad \textit{and}, \tag{31}$$

$$\operatorname{Var}_\theta(\widehat{\|\theta\|_2^2}) \leq \|\theta\|_2^2 + 3s + \frac{8}{\sqrt{\gamma \log(n/s^2)}} n\left(\frac{s^2}{n}\right)^{\frac{\gamma}{2}}. \tag{32}$$

The constraint $\gamma \log(n/s^2) \geq 2$ holds for all sufficiently large n, whenever γ is fixed and $s \ll \sqrt{n}$.

Compared with Theorem 6.3, the corresponding upper bound (28) has the same structure. Key difference is that the exponent is 4γ in the lower bound and $\gamma/2$ in the upper bound. As discussed already, this discrepancy seems to be due to the lower bound. If instead of a tight control of the variance at zero, one is interested in a global bound on the variance over the whole parameter space, one could gain a factor 4 in the exponent by relying on the Hellinger version using Theorem 2.2(ii) instead of (i). A second difference is that there is an additional factor $1/\sqrt{\log(n/s^2)}$ in the upper bound. This extra factor tends to zero, which seems to be a contradiction. Notice, however, that this is compensated by the different exponents $(s^2/n)^{\gamma/2}$ and $(s^2/n)^{4\gamma}$. It is also not hard to see that for the hard thresholding estimator with truncation level $\sqrt{\gamma \log(n/s^2)}$, the variance $\sum_{i=1}^n \operatorname{Var}_0(\widehat{\theta}_i)$ is of order $n(s^2/n)^{\gamma/2}$.

The upper bound in (31) corresponds to the lower bound in Theorem 6.4. The differences between upper and lower bound are similarly as the ones between the upper bound (28) and Theorem 6.3 discussed in the previous paragraph.

If $s \leq n^{1/2-\delta}$ for some $0 < \delta < 1/2$, then by choosing γ large enough, one can show that (29) implies $\sup_{\theta\in\Theta(s)} \sum_{i=1}^n \operatorname{Var}_\theta(\widehat{\theta}_i) \lesssim s$. This yields then the last statement of Theorem 6.1. Similarly, one can use (32) to construct an estimator satisfying the variance bound in Theorem 6.2(i).

The soft-thresholding estimator (26) does not produce an s-sparse model. Indeed, from the tail decay of the Gaussian distribution, one expects that the sparsity of the reconstruction for

$\theta = (0, \dots, 0)$ is $n(s^2/n)^{\gamma/2}$, which can be considerably bigger than s for small values of γ. Because testing for signal is very hard in the sparse sequence model, it is unclear whether one can reduce the variance further by projecting it to an s-sparse set without inflating the bias.

Extension to high-dimensional regression. The lower bound can also be extended to a useful lower bound on the interplay between bias and variance in sparse high-dimensional regression. Suppose we observe $Y = X\beta + \varepsilon$ where Y is a vector of size n, X is an $n \times p$ design matrix, $\varepsilon \sim \mathcal{N}(0, I_n)$ and β is a vector of size p to be estimated. Again denote by $\Theta(s)$ the class of s-sparse vectors. We impose the common assumption that the diagonal coefficients of the Gram matrix $X^\top X$ are standardized such that $(X^\top X)_{i,i} = n$ for all $i = 1, \dots, p$ (see, for instance, also Section 6 in [6]). Define the mutual coherence condition number by $\mathrm{mc}(X) := \max_{1 \le i \ne j \le n} (X^\top X)_{i,j}/(X^\top X)_{i,i}$. This notion goes back to [14]. Below, we work under the restriction $\mathrm{mc}(X) \le 1/(s^2 \log(p/s^2))$. This is stronger than the mutual coherence bound of the form constant$/s$ normally encountered in high-dimensional statistics. As this is not the main point of the paper, we did not attempt to derive the theorem under the sharpest possible condition and also only provide the generalization of Theorem 6.3.

THEOREM 6.7. *Consider the sparse high-dimensional regression model with Gaussian noise. Let* $0 < s \le \sqrt{p}/2$, *and* $\mathrm{mc}(X) \le 1/(s^2 \log(p/s^2))$. *Given an estimator* $\widehat{\beta}$ *and a real number* γ *such that* $4\gamma + 1/\log(p/s^2) \le 0.99$ *and* $\sup_{\beta \in \Theta(s)} \|E_\beta[\widehat{\beta}] - \beta\|^2 \le (\gamma s/n) \log(p/s^2)$, *then, for all sufficiently large* p,

$$\sum_{i=1}^{p} \mathrm{Var}_0(\widehat{\beta}_i) \ge \frac{(1 - (1/2)^{0.01})}{25e^2 \log(p/s^2)} \frac{p}{n} \left(\frac{s^2}{p}\right)^{4\gamma},$$

where Var_0 *denotes the variance for parameter vector* $\beta = (0, \dots, 0)^\top$.

7. Discussion.

7.1. *General definition of a bias-variance trade-off.* The proper definition of the bias-variance trade-off depends on some subtleties underlying the choice of the space of values that can be attained by an estimator, subsequently denoted by $\mathcal{A}$. To illustrate this, suppose we observe $X \sim \mathcal{N}(\theta, 1)$ with parameter space $\Theta = \{-1, 1\}$. For any estimator $\widehat{\theta}$ with $\mathcal{A} = \Theta$, $E_1[\widehat{\theta}] < 1$ or $E_{-1}[\widehat{\theta}] > -1$. Thus, no unbiased estimator with $\mathcal{A} = \Theta$ exists. If the estimator is, however, allowed to take values on the real line, then $\widehat{\theta} = X$ is an unbiased estimator for θ. We believe that the correct way to derive lower bounds on the bias-variance trade-off is to allow the action space $\mathcal{A}$ to be large. Whenever Θ is a class of functions on $[0, 1]$, the derived lower bounds are over all estimators with $\mathcal{A}$ the real-valued functions on $[0, 1]$; for high-dimensional problems with $\Theta \subseteq \mathbb{R}^p$, the lower bounds are over all estimators with $\mathcal{A} = \mathbb{R}^p$. In particular, if the true parameter vector is assumed to be sparse, we do not require the estimator to be sparse.

Given a statistical model $(P_\theta)_{\theta \in \Theta}$, consider a symmetric (and nonnegative) loss function $\ell(\theta, \theta') = \ell(\theta', \theta)$. The risk of an estimator $\widehat{\theta}$ is then $E_\theta[\ell(\widehat{\theta}, \theta)]$. If $E_\theta[\widehat{\theta}]$ exists, we call $\ell(\theta, E_\theta[\widehat{\theta}])$ the deterministic error and $E_\theta[\ell(\widehat{\theta}, E_\theta[\widehat{\theta}])]$ the stochastic error. If for all estimators $\widehat{\theta}$ and all parameters θ, $E_\theta[\ell(\widehat{\theta}, \theta)] = \ell(\theta, E_\theta[\widehat{\theta}]) + E_\theta[\ell(\widehat{\theta}, E_\theta[\widehat{\theta}])]$, then we say that a (generalized) bias-variance decomposition holds and refer to the deterministic error $\ell(\theta, E_\theta[\widehat{\theta}])$ as the squared bias part and to the stochastic error $E_\theta[\ell(\widehat{\theta}, E_\theta[\widehat{\theta}])]$ as the variance part. Note that the squared bias is defined directly without introducing first a notion of bias.

A bias-variance decomposition exists if $\ell(\theta, \theta') = \|\theta - \theta'\|^2$ with $\|\cdot\|$ a Hilbert space norm. In particular, for $\ell(\theta, \theta') = (\theta - \theta')^2$, we have the classical bias-variance decomposition of the MSE. On a vector space Θ, the loss function $\ell(\theta, \theta') = \|\theta - \theta'\|_2^2$ leads to the decomposition (23). In this case, the squared bias part is $\|E_\theta[\widehat{\theta}] - \theta\|_2^2$ and the variance part is $\sum_{i=1}^n \operatorname{Var}_\theta(\widehat{\theta}_i)$. For Θ consisting of L^2-functions, the decomposition of the MISE in integrated squared bias and integrated variance in (18) is another example.

With this definition of squared bias and variance, we can now define a bias-variance trade-off informally as either a restriction of the squared bias part that follows from imposing a constraint on the variance part or a restriction on the variance part that is implied by a constraint on the squared bias part. To introduce a formal definition, denote the squared bias part by $B_\theta(\widehat{\theta})^2$ and the variance part by $V_\theta(\widehat{\theta})$. The functions $\theta \mapsto B_\theta(\widehat{\theta})$ and $\theta \mapsto V_\theta(\widehat{\theta})$ belong to the space $[0, +\infty]^\Theta$. Let $\mathcal{T} \subset \Theta^{\mathcal{X}}$ be a class of estimators and let ψ be a function $\psi : [0, +\infty]^\Theta \times [0, +\infty]^\Theta \to [0, +\infty]$ increasing in both of its arguments in the sense that for $b_1(\cdot)^2 \le b_2(\cdot)^2$ and $v_1(\cdot) \le v_2(\cdot)$, $\psi(b_1^2, v_1) \le \psi(b_2^2, v_2)$. We say that ψ is a bias-variance trade-off for the class of estimators $\mathcal{T}$ if $\inf_{\widehat{\theta} \in \mathcal{T}} \psi(\theta \mapsto B_\theta(\widehat{\theta})^2, \theta \mapsto V_\theta(\widehat{\theta})) \ge 1$. All bias-variance trade-offs derived in this paper can be put into this form. For instance, for the two bias-variance trade-offs for pointwise estimation in Theorem 3.1, we can choose using the notation introduced in Section 3,

$$\psi(b^2, v) := \frac{n}{\gamma(R, \beta)} \sup_{f \in \mathscr{C}^\beta(R)} |b(f)|^{1/\beta} \sup_{f \in \mathscr{C}^\beta(R)} v(f), \quad \text{and,}$$

$$\psi(b^2, v) := n \sup_{f \in \mathscr{C}^\beta(R)} |b(f)|^{1/\beta} \inf_{f \in \mathscr{C}^\beta(R)} \frac{v(f)}{\overline{\gamma}(R, \beta, C, \|f\|_{\mathscr{C}^\beta})}.$$

7.2. *Comparison of the abstract lower bounds for the bias-variance trade-off and the hypothesis testing approach for minimax lower bounds.* While nontrivial minimax rates exist for parametric and nonparametric problems alike, the bias-variance trade-off phenomenon occurs mainly in high-dimensional and infinite-dimensional models. Despite these differences, the here proposed strategy for lower bounds on the bias-variance trade-off and the well-developed testing approach for lower bounds on the minimax estimation rate share some similarities. A clear similarity is that for both approaches, the problem is reduced in a first step by selecting a discrete subset of the parameter space. To achieve rate-optimal minimax lower bounds, it is well known that for a large class of functionals, reduction to two parameters is sufficient. On the contrary, optimal lower bounds for global loss functions, such as L^p-loss in nonparametric regression, require to pick a number of parameter values that increases with the sample size. We argued in this work that a similar distinction occurs also for bias-variance trade-off lower bounds. As in the case of the minimax estimation risk, we can relate the two-parameter lower bounds to a bound with respect to any of the commonly used information measures including the Kullback–Leibler divergence.

More pronounced differences occur in the formulation of both lower bound techniques for lower bounds involving more than two parameter values. While for minimax lower bounds the parameters correspond to several hypotheses that form a local packing of the parameter space, for bias-variance trade-off lower bounds the contribution of the selected parameters is determined by how orthogonal the corresponding distributions are. Here, the orthogonality of distributions is measured by the χ^2-divergence matrix or the Hellinger affinity matrices; see Table 1 in [12] for examples.

As shown in Section 3 of [12], $\mathbf{v}^\top \chi^2(P_0|P_1, \dots, P_M)\mathbf{v} = \chi^2(\sum_{j=1}^M v_j P_j, P_0)$, where $\sum_{j=1}^M v_j P_j$ is the mixture (signed) measure of $P_1, \dots, P_M$. This suggests to interpret the case of multiple measures $P_0, \dots, P_M$ as a two-point testing problem, where we measure

TABLE 1
Proof techniques for different examples

Framework	Theorem	Proof technique
Pointwise estimation in the Gaussian white noise model	3.1	Univariate change of expectation
Pointwise estimation of the boundary of a Poisson point process	4.1	Univariate change of expectation
Function estimation in the Gaussian white noise model with L_2-loss	5.1	2 reductions + multivariate change of expectation
Estimation of θ in the Gaussian sequence model under sparsity	6.3 and 6.5	Multivariate change of expectation (+ 1 reduction for Theorem 6.5)
Estimation of $\|\theta\|_2^2$ in the Gaussian sequence model under sparsity	6.4	Multivariate change of expectation
Sparse high-dimensional regression with Gaussian noise	6.7	Multivariate change of expectation

the information distance between P_0 and a linear combination $\sum_{j=1}^M v_j P_j$. Viewed from this perspective, the proposed approach shares some similarities with the minimax lower bounds based on two fuzzy hypothesis and Fano's lemma. For a description of these approaches, see Section 2.7 in [37].

7.3. *Bias-variance trade-off lower bounds and their proof techniques*. Table 1 states the applied proof technique for each of the lower bounds proved in this article. Here, "univariate change of expectation" refers to the use of Lemma 2.1 or Lemma A.2 of the Supplementary Material [13]; "multivariate change of expectation" refers to applying Theorem 2.2 or the variations stated in equations (5) and (6).

Acknowledgments. We are grateful to Ming Yuan for helpful discussions during an early stage of the project, to Zijian Guo for pointing us to the article [11], and to Tomohiro Nishiyama for mentioning a typo in an earlier version. We thank two reviewers and an Associate Editor for many helpful comments and suggestions that significantly improved the manuscript.

Funding. The project has received funding from the Dutch Science Foundation (NWO) via the Vidi grant VI.Vidi.192.021.

SUPPLEMENTARY MATERIAL

Supplement to "On lower bounds for the bias-variance trade-off" (DOI: 10.1214/23-AOS2279SUPP; .pdf). All proofs are given in the Supplement [13].

REFERENCES

[1] BELKIN, M., HSU, D., MA, S. and MANDAL, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. USA* **116** 15849–15854. MR3997901 https://doi.org/10.1073/pnas.1903070116

[2] BERTHET, Q. and RIGOLLET, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, *PMLR* 1046–1066.

[3] BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849 https://doi.org/10.1214/13-AOS1127

[4] Brown, L. D. and Low, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24** 2524–2535. MR1425965 https://doi.org/10.1214/aos/1032181166

[5] Brown, L. D., Low, M. G. and Zhao, L. H. (1997). Superefficiency in nonparametric function estimation. *Ann. Statist.* **25** 2607–2625. MR1604424 https://doi.org/10.1214/aos/1030741087

[6] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data. Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

[7] Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395 https://doi.org/10.1214/16-AOS1461

[8] Chafaï, D. (2004). Entropies, convexity, and functional inequalities: On Φ-entropies and Φ-Sobolev inequalities. *J. Math. Kyoto Univ.* **44** 325–363. MR2081075 https://doi.org/10.1215/kjm/1250283556

[9] Chen, J. (2004). Notes on the bias-variance trade-off phenomenon. In *A Festschrift for Herman Rubin. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **45** 207–217. IMS, Beachwood, OH. MR2126898 https://doi.org/10.1214/lnms/1196285391

[10] Chen, X., Guntuboyina, A. and Zhang, Y. (2016). On Bayes risk lower bounds. *J. Mach. Learn. Res.* **17** 219. MR3595153

[11] Collier, O., Comminges, L. and Tsybakov, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *Ann. Statist.* **45** 923–958. MR3662444 https://doi.org/10.1214/15-AOS1432

[12] Derumigny, A. and Schmidt-Hieber, J. (2023). Codivergences and information matrices. ArXiv E-prints. Available at arXiv:2303.08122.

[13] Derumigny, A. and Schmidt-Hieber, J. (2023). Supplement to "On lower bounds for the bias-variance trade-off." https://doi.org/10.1214/23-AOS2279SUPP

[14] Donoho, D. L., Elad, M. and Temlyakov, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inf. Theory* **52** 6–18. MR2237332 https://doi.org/10.1109/TIT.2005.860430

[15] Donoho, D. L., Johnstone, I. M., Hoch, J. C. and Stern, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* **54** 41–81. With discussion and a reply by the authors. MR1157714

[16] Duchi, J. C., Jordan, M. I. and Wainwright, M. J. (2013). Local privacy and statistical minimax rates. In 2013 *IEEE* 54*th Annual Symposium on Foundations of Computer Science—FOCS* 2013 429–438. IEEE Computer Soc., Los Alamitos, CA. MR3246246 https://doi.org/10.1109/FOCS.2013.53

[17] Feldman, V., Ligett, K. and Sabato, S., eds. (2021). *Algorithmic Learning Theory. Proceedings of Machine Learning Research* (*PMLR*) **132**.

[18] Geman, S., Bienenstock, E. and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Comput.* **1** 1–58.

[19] Gerchinovitz, S., Ménard, P. and Stoltz, G. (2020). Fano's inequality for random variables. *Statist. Sci.* **35** 178–201. MR4106600 https://doi.org/10.1214/19-STS716

[20] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

[21] Johnstone, I. M. (2019). Gaussian estimation: Sequence and wavelet models. Available at http://statweb.stanford.edu/~imj/GE_09_16_19.pdf.

[22] Lehmann, E. L. and Casella, G. (2006). *Theory of Point Estimation*. Springer, Berlin.

[23] Liu, R. C. and Brown, L. D. (1993). Nonexistence of informative unbiased estimators in singular problems. *Ann. Statist.* **21** 1–13. MR1212163 https://doi.org/10.1214/aos/1176349012

[24] Low, M. G. (1995). Bias-variance tradeoffs in functional estimation problems. *Ann. Statist.* **23** 824–835. MR1345202 https://doi.org/10.1214/aos/1176324624

[25] Meister, A. and Reiss, M. (2013). Asymptotic equivalence for nonparametric regression with nonregular errors. *Probab. Theory Related Fields* **155** 201–229. MR3010397 https://doi.org/10.1007/s00440-011-0396-x

[26] Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S. and Mitliagkas, I. (2018). A modern take on the bias-variance tradeoff in neural networks. ArXiv E-prints. Available at arXiv:1810.08591.

[27] Nishiyama, T. (2019). A new lower bound for Kullback–Leibler divergence based on Hammersley–Chapman–Robbins bound. arXiv e-prints. Available at arXiv:1907.00288.

[28] Nishiyama, T. (2020). A tight lower bound for the Hellinger distance with given means and variances. arXiv e-prints. Available at arXiv:2010.13548.

[29] PFANZAGL, J. (2001). A nonparametric asymptotic version of the Cramér–Rao bound. In *State of the Art in Probability and Statistics* (*Leiden*, 1999). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **36** 499–517. IMS, Beachwood, OH. MR1836577 https://doi.org/10.1214/lnms/1215090085

[30] POLYANSKIY, Y. (2021). Lecture 2: Dpi and statistics. information theoretic methods in statistics and computer science. Available at http://people.lids.mit.edu/yp/homepage/sdpi_course.html.

[31] RAGINSKY, M. (2016). Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels. *IEEE Trans. Inf. Theory* **62** 3355–3389. MR3506739 https://doi.org/10.1109/TIT.2016.2549542

[32] REISS, M. and SCHMIDT-HIEBER, J. (2020). Posterior contraction rates for support boundary recovery. *Stochastic Process. Appl.* **130** 6638–6656. MR4158798 https://doi.org/10.1016/j.spa.2020.06.005

[33] REISS, M. and SELK, L. (2017). Efficient estimation of functionals in nonparametric boundary models. *Bernoulli* **23** 1022–1055. MR3606758 https://doi.org/10.3150/15-BEJ768

[34] ROHDE, A. and STEINBERGER, L. (2020). Geometrizing rates of convergence under local differential privacy constraints. *Ann. Statist.* **48** 2646–2670. MR4152116 https://doi.org/10.1214/19-AOS1901

[35] STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 197–206. Univ. California Press, Berkeley-Los Angeles, CA. MR0084922

[36] SZABÓ, B. and VAN ZANTEN, H. (2020). Adaptive distributed methods under communication constraints. *Ann. Statist.* **48** 2347–2380. MR4134798 https://doi.org/10.1214/19-AOS1890

[37] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

[38] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

[39] WAHL, M. (2022). Van Trees inequality, group equivariance, and estimation of principal subspaces. *Ann. Inst. Henri Poincaré Probab. Stat.* **58** 1565–1589. MR4452643 https://doi.org/10.1214/21-AIHP1193

[40] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

[41] ZHANG, Y., DUCHI, J., JORDAN, M. I. and WAINWRIGHT, M. J. (2013). Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems* 26 2328–2336. Curran Associates, Red Hook.