On lower bounds for the bias-variance trade-off

Alexis Derumigny and Johannes Schmidt-Hieber

Department of Applied Mathematics, University of Twente

Abstract

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. To highlight possible extensions of the proposed framework, we moreover briefly discuss the trade-off between bias and mean absolute deviation.

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], p.45), the very recent work on overparametrization in machine learning has cast some doubt on the necessity to balance squared bias and variance [2, 29].

Although the bias-variance trade-off is omnipresent whenever estimation in complex statistical models is considered, in most cases it is unknown whether methods exist that avoid such a trade-off by having for instance a bias of negligible order. In some instances, small bias is possible. An important example is the rather subtle de-biasing of the LASSO for a class of functional in the high-dimensional regression model [45, 42, 9]. This shows that the occurrence of the bias-variance trade-off is a highly non-trivial phenomenon.

It is thus surprising that so little theoretical work has been done on lower bounds for the interplay between bias and variance. The major contribution is due to Mark Low [27] 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 [31], also considering estimation of functionals but not necessarily in the Gaussian white noise model. It is shown that estimators satisfying an asymptotic unbiasedness property must have unbounded variance.

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.

While non-trivial minimax rates exist for parametric and non-parametric problems alike, the bias-variance trade-off phenomenon occurs 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 theory for lower bounds on the minimax estimation rate share some similarities. A 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. 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 measure including the Kullback-Leibler divergence. The difference between both lower bound techniques becomes most apparent for lower bounds involving more than two parameter values. While for minimax lower bounds the parameters are chosen by a local packing of the parameter space, for bias-variance trade-off lower bounds the contribution of each of the selected parameters is measured by an orthogonality-type relation of the corresponding distribution with respect to either the Hellinger distance or the χ^2-divergence. We encode this orthogonality relation in an information matrix that we call the χ^2-divergence matrix or the Hellinger affinity matrix, depending on whether we work with the χ^2-divergence or the Hellinger distance. For lower bounds on the bias-variance trade-off it is then sufficient to control the largest eigenvalue of this matrix. As examples for the information matrix approach, we consider sparse recovery in the sequence model and the high-dimensional linear regression model.

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.

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 some parameter, 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 some parameter will necessarily inflate the variance for all parameters that are (in a suitable sense) away from the boundary of the parameter space.

In these cases, 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 a different phenomenon occurs. For estimators with bias bounded by constant$\times$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.

Summarizing the results, for all of the considered models a non-trivial 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 non-adaptive setups are considered. Adaptation to either smoothness or sparsity induces additional bias.

As mentioned above, the main motivation for this work is to test whether the new regimes found in modern machine learning could avoid the classical bias-variance trade-off. Besides that, there are many other good reasons why a better understanding of the bias-variance trade-off is relevant for statistical practice. Even in non-adaptive 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.

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 [4, 3], lower and upper bounds for estimation under privacy constraints [16, 36, 1], and parallelizable estimators under communication constraints [46, 39].

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 using these inequalities. The considered setups range from pointwise estimation in the Gaussian white noise model (Section 3) and a boundary estimation problem (Section 4) to high-dimensional models in Section 5. In Section 6 the approach via change of expectation inequalities is combined with a reduction scheme that reduces the complexity of the underlying model and the class of candidate estimators. This approach is illustrated by studying lower bounds for the trade-off between integrated squared bias and integrated variance in the Gaussian white noise model. Section 7 serves as an outlook to generalizations of the bias-variance trade-off. Specifically, we study the mean absolute deviation and derive a lower bound for the trade-off between bias and mean absolute deviation considering again pointwise estimation in the Gaussian white noise model. Most proofs are deferred to the Supplement.

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 $\mathbb{V}\mathrm{ar}_P$ for the expectation and variance with respect to P, respectively. For probability measures P_θ depending on a parameter θ, E_θ and $\mathbb{V}\mathrm{ar}_\theta$ denote the corresponding expectation and variance. If a random variable X is not square integrable with respect to P, we assign the value $+\infty$ to $\mathbb{V}\mathrm{ar}_P(X)$. For any finite number of measures $P_1, \dots, P_M$, 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 defined as $\mathrm{TV}(P,Q) := \frac{1}{2}\int |p(\omega) - q(\omega)|\, d\nu(\omega)$. 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 $\chi^2(P,Q) := \int (p(\omega)/q(\omega) - 1)^2\, d\nu(\omega)$. 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

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 $V(\theta) := \mathbb{V}\mathrm{ar}_\theta(\widehat{\theta})$, 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 semi-parametric settings ([31]). Although this seems to provide a straightforward path to lower bounds on the bias-variance trade-off, there are many good reasons why this approach is problematic for nonparametric and high-dimensional models. A major obstacle is the proper definition of a nonparametric Fisher information. It is moreover unclear how to interpret the Fisher information for parameter spaces that are not open sets such as for instance the space of all sparse vectors.

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 the 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 below, the Cramér-Rao lower bound reappears as a limit of these inequalities, but they are much better suited for nonparametric problems as no notion of differentiability of the distribution with respect to the parameter is required.

Lemma 2.1. *Let P and Q be two probability distributions on the same measurable space. Denote by E_P and $\mathbb{V}ar_P$ the expectation and variance with respect to P and let E_Q and $\mathbb{V}ar_Q$ be the expectation and variance with respect to Q. Then, for any random variable X,*

$$\frac{(E_P[X]-E_Q[X])^2}{2}\Big(\frac{1}{\mathrm{TV}(P,Q)}-1\Big) \leq \mathbb{V}ar_P(X)+\mathbb{V}ar_Q(X), \tag{1}$$

$$\frac{(E_P[X]-E_Q[X])^2}{4}\Big(\frac{1}{H(P,Q)}-H(P,Q)\Big)^2 \leq \mathbb{V}ar_P(X)+\mathbb{V}ar_Q(X), \tag{2}$$

$$(E_P[X]-E_Q[X])^2\Big(\frac{1}{\mathrm{KL}(P,Q)+\mathrm{KL}(Q,P)}-\frac{1}{4}\Big) \leq \mathbb{V}ar_P(X)\vee\mathbb{V}ar_Q(X), \tag{3}$$

$$(E_P[X]-E_Q[X])^2 \leq \chi^2(Q,P)\mathbb{V}ar_P(X)\wedge\chi^2(P,Q)\mathbb{V}ar_Q(X). \tag{4}$$

A proof is provided in Supplement B. If any 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 $(\chi^2(Q,P))^2$.

To obtain lower bounds for the variance, these 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 inequality should balance these two aspects.

To illustrate these bounds for a specific example, consider multivariate normal distributions $P=\mathcal{N}(\theta,I)$ and $Q=\mathcal{N}(\theta',I)$, for vectors θ,θ' and I the identity matrix. In this case, closed-form expressions for all four information measures exist. Denote by Φ the c.d.f. of the normal distribution. Since $\mathrm{TV}(P,Q)=1-P(dQ/dP>1)-Q(dP/dQ\geq 1)=1-2\Phi(-\frac{1}{2}\|\theta-\theta'\|_2^2)$, $H^2(P,Q)=1-\exp(-\frac{1}{8}\|\theta-\theta'\|_2^2)$, $\mathrm{KL}(P,Q)=$

$\mathrm{KL}(Q,P)=\frac{1}{2}\|\theta-\theta'\|_2^2$, and $\chi^2(P,Q)=\exp(\|\theta-\theta'\|_2^2)-1$, the inequalities (1)-(4) become

$$\begin{aligned}\big(E_\theta[X]-E_{\theta'}[X]\big)^2\frac{\Phi(-\frac{1}{2}\|\theta-\theta'\|_2^2)}{1-2\Phi(-\frac{1}{2}\|\theta-\theta'\|_2^2)}&\leq \mathbb{V}\mathrm{ar}_\theta(X)+\mathbb{V}\mathrm{ar}_{\theta'}(X)\\ \big(E_\theta[X]-E_{\theta'}[X]\big)^2\frac{\frac{1}{4}\exp(-\frac{1}{4}\|\theta-\theta'\|_2^2)}{1-\exp(-\frac{1}{8}\|\theta-\theta'\|_2^2)}&\leq \mathbb{V}\mathrm{ar}_\theta(X)+\mathbb{V}\mathrm{ar}_{\theta'}(X)\\ \big(E_\theta[X]-E_{\theta'}[X]\big)^2\left(\frac{1}{\|\theta-\theta'\|_2^2}-\frac{1}{4}\right)&\leq \mathbb{V}\mathrm{ar}_\theta(X)+\mathbb{V}\mathrm{ar}_{\theta'}(X)\\ \big(E_\theta[X]-E_{\theta'}[X]\big)^2\leq \big(\exp\big(\|\theta-\theta'\|_2^2\big)-1\big)&\big(\mathbb{V}\mathrm{ar}_\theta(X)\wedge\mathbb{V}\mathrm{ar}_{\theta'}(X)\big).\end{aligned}\tag{5}$$

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 produces non-trivial 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 [41]).

Except for the total variation version, all derived bounds are generalizations of the Cramér-Rao lower bound. As this is not the main focus of the work, we give an informal argument without stating the exact regularity conditions. In the above inequalities, take P and Q to be P_θ and $P_{\theta+\Delta}$ and let Δ tend to zero. Recall that $B'(\theta)$ is the derivative of the bias at θ and $F(\theta)$ denotes the Fisher information. For any estimator $\widehat{\theta}$, we have for small Δ, $(E_{P_\theta}[\widehat{\theta}]-E_{P_{\theta+\Delta}}[\widehat{\theta}])^2\approx\Delta^2(1+B'(\theta))^2$. Using that $(\sqrt{x}-\sqrt{y})^2=(x-y)^2/(\sqrt{x}+\sqrt{y})^2$ shows that $H^2(P_\theta,P_{\theta+\Delta})^2\approx\Delta^2F(\theta)/8$. Moreover, $\mathrm{KL}(P,Q)+\mathrm{KL}(Q,P)=\int\log(p/q)(p-q)$ and a first order Taylor expansion of $\log(x)$ shows that $\mathrm{KL}(P_\theta,P_{\theta+\Delta})+\mathrm{KL}(P_{\theta+\Delta},P_\theta)\approx\Delta^2F(\theta)$. Similarly, we find $\chi^2(P_{\theta+\Delta},P_\theta)\approx\Delta^2F(\theta)$. Replacing the divergences by their approximations then shows that for the Hellinger, Kullback-Leibler and χ^2 versions, the Cramér-Rao lower bound can be retrieved taking the limit $\Delta\to 0$.

Inspired by this limit, we can derive for parametric models the following lemma, proved in Supplement B.

Lemma 2.2. *Given a family of probability measures $(P_t)_{t\in[0,1]}$. For simplicity write E_t and $\mathbb{V}ar_t$ for E_{P_t} and $\mathbb{V}ar_{P_t}$, respectively.*
(i): If $\kappa_H:=\liminf_{\delta\to0}\delta^{-1}\sup_{t\in[0,1-\delta]}H(P_t,P_{t+\delta})$ is finite, then for any random variable X,

$$\big(E_1[X]-E_0[X]\big)^2\leq 8\kappa_H^2\sup_{t\in[0,1]}\mathbb{V}ar_t(X).\tag{6}$$

(ii): If $\kappa_K^2:=\liminf_{\delta\to0}\delta^{-2}\sup_{t\in[0,1-\delta]}\mathrm{KL}(P_t,P_{t+\delta})+\mathrm{KL}(P_{t+\delta},P_t)$ is finite, then for any random variable X,

$$\big(E_1[X]-E_0[X]\big)^2\leq \kappa_K^2\sup_{t\in[0,1]}\mathbb{V}ar_t(X).\tag{7}$$

(iii): If $\kappa_\chi^2:=\liminf_{\delta\to0}\delta^{-2}\sup_{t\in[0,1-\delta]}\chi^2(P_t,P_{t+\delta})$ is finite, then for any random variable X,

$$\big(E_1[X]-E_0[X]\big)^2\leq \kappa_\chi^2\sup_{t\in[0,1]}\mathbb{V}ar_t(X).\tag{8}$$

As an example, consider the family $P_t = \mathcal{N}(t\theta + (1-t)\theta', I)$ $t \in [0,1]$. Then, $(i) - (iii)$ all lead to the inequality $(E_\theta[X] - E_{\theta'}[X])^2 \leq \|\theta - \theta'\|_2^2 \sup_{t\in[0,1]} \mathbb{V}\mathrm{ar}_t(X)$. In (5), the bounds for the Hellinger distance and the χ^2-divergence grow exponentially in $\|\theta - \theta'\|_2^2$ and the Kullback-Leibler bound only provides a non-trivial lower bound if $\|\theta - \theta'\|_2^2 \leq 4$. Lemma 2.2 leads thus to much sharper constants if $\|\theta - \theta'\|_2$ is large. On the other hand, compared to the earlier bounds, Lemma 2.2 results in a weaker statement on the bias-variance trade-off as it only produces a lower bound for the largest of all variances $\mathbb{V}\mathrm{ar}_t(X)$, $t \in [0,1]$.

Information matrices: 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 lower bounds, we need instead to reduce the problem to a multiple testing problem with potentially a large number of tests.

A similar phenomenon occurs also for lower bounds on the bias-variance trade-off. Given $M+1$ probability measures $P_0, P_1, \ldots, P_M$, the χ^2-version of Lemma 2.1 states that for any $j = 1, \ldots, M$, $(E_{P_j}[X] - E_{P_0}[X])^2/\chi^2(P_j, P_0) \leq \mathbb{V}\mathrm{ar}_{P_0}(X)$. If $P_1, \ldots, P_M$ 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, $\sum_{j=1}^M (E_{P_j}[X] - E_{P_0}[X])^2/\chi^2(P_j, P_0) \leq \mathbb{V}\mathrm{ar}_{P_0}(X)$. In a next step, two notions of information matrices are introduced, measuring to which extent $P_1, \ldots, P_M$ represent different directions around P_0.

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

$$\chi^2(P_0, \ldots, P_M)_{j,k} := \int \frac{dP_j}{dP_0} dP_k - 1. \tag{9}$$

Observe that the j-th diagonal entry coincides with the χ^2-divergence $\chi^2(P_j, P_0)$. The χ^2-divergence matrix is also the covariance matrix of the random vector $(dP_1/dP_0(X), \ldots, dP_M/dP_0(X))^\top$ under P_0 and hence symmetric and positive semi-definite. For any vector $v = (v_1, \ldots, v_M)^\top \in \mathbb{R}^M$, we have the identity $v^\top \chi^2(P_0, \ldots, P_M) v = \int (\sum_{j=1}^M v_j(dP_j/dP_0 - 1))^2 dP_0$. This shows that for non-negative weights v_j summing to one, $v^\top \chi^2(P_0, \ldots, P_M) v$ coincides with the χ^2-divergence of the mixture distribution $\sum_{j=1}^M v_j P_j$ and P_0, that is, $\chi^2(\sum_{j=1}^M v_j P_j, P_0)$. Another consequence of this identity is that the χ^2-divergence matrix is invertible if and only if P_0 cannot be expressed as a linear combination of $P_1, \ldots, P_M$. Indeed, a vector v lies in the kernel of the χ^2-divergence matrix if and only if $\sum_{j=1}^M v_j(P_j - P_0) = 0$.

We also introduce an information matrix based on the Hellinger distance. The $M \times M$ Hellinger affinity matrix is defined entrywise by

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

Here and throughout the article, we implicitly assume that the distributions $P_0, \ldots, P_M$ 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.

distribution	$\chi^2(P_0,\dots,P_M)_{j,k}$	$\rho(P_0\|P_1,\dots,P_M)_{j,k}$
$P_j = \mathcal{N}(\theta_j,\sigma^2 I_d)$, $\theta_j \in \mathbb{R}^d$, I_d identity	$\exp\Big(\frac{\langle \theta_j-\theta_0,\theta_k-\theta_0\rangle}{\sigma^2}\Big)-1$	$\exp\Big(\frac{\langle \theta_j-\theta_0,\theta_k-\theta_0\rangle}{4\sigma^2}\Big)-1$
$P_j = \otimes_{\ell=1}^d \operatorname{Pois}(\lambda_{j\ell})$, $\lambda_{j\ell}>0$	$\exp\Big(\sum_{\ell=1}^d \frac{(\lambda_{j\ell}-\lambda_{0\ell})(\lambda_{k\ell}-\lambda_{0\ell})}{\lambda_{0\ell}}\Big)-1$	$\exp\Big(\sum_{\ell=1}^d \big(\sqrt{\lambda_{j\ell}}-\sqrt{\lambda_{0\ell}}\big)\big(\sqrt{\lambda_{k\ell}}-\sqrt{\lambda_{0\ell}}\big)\Big)-1$
$P_j = \otimes_{\ell=1}^d \operatorname{Exp}(\beta_{j\ell})$, $\beta_{j\ell}>0$	$\prod_{\ell=1}^d \frac{\beta_{j\ell}\beta_{k\ell}}{\beta_{0\ell}\big(\beta_{j\ell}+\beta_{k\ell}-\beta_{0\ell}\big)}-1$	$\prod_{\ell=1}^d \frac{(\beta_{j\ell}+\beta_{0\ell})(\beta_{k\ell}+\beta_{0\ell})}{2\beta_{0\ell}(\beta_{j\ell}+\beta_{k\ell})}-1$
$P_j = \otimes_{\ell=1}^d \operatorname{Ber}(\theta_{j\ell})$, $\theta_{j\ell}\in(0,1)$	$\prod_{\ell=1}^d \Big(\frac{(\theta_{j\ell}-\theta_{0\ell})(\theta_{k\ell}-\theta_{0\ell})}{\theta_{0\ell}(1-\theta_{0\ell})}+1\Big)-1$	$\prod_{\ell=1}^d \frac{r(\theta_{j\ell},\theta_{k\ell})}{r(\theta_{j\ell},\theta_{0\ell})r(\theta_{k\ell},\theta_{0\ell})}-1$ with $r(\theta,\theta') := \sqrt{\theta\theta'}+\sqrt{(1-\theta)(1-\theta')}$

Table 1: Closed-form expressions for the χ^2-divergence and Hellinger affinity matrix for some distributions. Proofs can be found in Supplement A.

Expanding the square in the integral, we find that for any vector $v=(v_1,\dots,v_M)^\top$,

$$v^\top \rho(P_0|P_1,\dots,P_M)v = \int \Big(\sum_{j=1}^M \Big(\frac{\sqrt{p_j}}{\int \sqrt{p_j p_0}\,d\nu} - \sqrt{p_0}\Big)v_j\Big)^2 d\nu \geq 0. \tag{10}$$

The Hellinger affinity matrix is hence symmetric and positive semi-definite. It can also be seen that it is singular if and only if $\int \sqrt{p_0}\,d\nu = 1$ and there exist numbers $w_1,\dots,w_M$, such that $\sum_{j=1}^M w_j\sqrt{p_j}$ is constant ν-almost everywhere.

For a number of distributions, closed-form expressions for the χ^2-divergence and the Hellinger affinity matrix are reported in Table 1. As mentioned before, these information matrices quantify to which extent the measures $P_1,\dots,P_M$ represent different directions around P_0. From these explicit formulas, it can be seen what this means in terms of the parameters. For the multivariate normal distribution, for instance, the χ^2-divergence matrix and the Hellinger affinity matrix are diagonal if and only if the vectors $\theta_j-\theta_0$ are pairwise orthogonal.

The formulas reveal a lot of similarity between the χ^2-divergence matrix and the Hellinger affinity matrix. It can also be shown that the diagonal elements of the χ^2-divergence matrix are entrywise larger than the diagonal elements of the Hellinger affinity matrix. To see this, observe that Hölder's inequality with $p=3/2$ and $q=3$ gives for any non-negative function f, $1=\int p_j \leq (\int f^p p_j)^{1/p}(\int f^{-q}p_j)^{1/q}$. The choice $f=(p_0/p_j)^{1/3}$ yields $1\leq (\int\sqrt{p_jp_0})^2\int p_j^2/p_0$. Rewriting this expression yields the claim.

Lower bounds based on an arbitrary number of distributions: 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.

Theorem 2.3. *Let $P_0,\dots,P_M$ be probability measures defined on the same probability space such that $P_j \ll P_0$ for all $j=1,\dots,M$. Let X be a random variable and set $\Delta := (E_{P_1}[X]-E_{P_0}[X],\dots,E_{P_M}[X]-E_{P_0}[X])^\top$.*

Then,

$$\Delta^\top \chi^2(P_0,\ldots,P_M)^+\Delta \leq \mathbb{V}ar_{P_0}(X),$$

where $\chi^2(P_0,\ldots,P_M)^+$ *denotes the Moore-Penrose inverse of the* χ^2*-divergence matrix.*

Proof of Theorem 2.3. Write a_j for the j-th entry of the vector $\Delta^\top\chi^2(P_0,\ldots,P_M)^+$ and E_j for the expectation E_{P_j}. Observe that for any $j,k=1,\ldots,M$, $\chi^2(P_0,\ldots,P_M)_{j,k} = \int (dP_j/dP_0)dP_k - 1 = E_0[(dP_j/dP_0 - 1)(dP_k/dP_0-1)]$. Using the Cauchy-Schwarz inequality and the fact that for a Moore-Penrose inverse A^+ of A, $A^+AA^+ = A^+$, we find

$$\begin{aligned}
\Big(\Delta^\top \chi^2(P_0,\ldots,P_M)^+\Delta\Big)^2 &= \Big(\sum_{j=1}^M a_j\big(E_{P_j}[X]-E_{P_0}[X]\big)\Big)^2 \\
&= E_0^2\Big[\sum_{j=1}^M a_j\Big(\frac{dP_j}{dP_0}-1\Big)\big(X-E_0[X]\big)\Big] \\
&\leq E_0\Big[\Big(\sum_{j=1}^M a_j\Big(\frac{dP_j}{dP_0}-1\Big)\Big)^2\Big]\mathbb{V}\mathrm{ar}_{P_0}(X) \\
&= \Big(\sum_{j,k=1}^M a_j\chi^2(P_0,\ldots,P_M)_{j,k}a_k\Big)\mathbb{V}\mathrm{ar}_{P_0}(X) \\
&= \Delta^\top\chi^2(P_0,\ldots,P_M)^+\chi^2(P_0,\ldots,P_M)\chi^2(P_0,\ldots,P_M)^+\Delta\mathbb{V}\mathrm{ar}_{P_0}(X) \\
&= \Delta^\top\chi^2(P_0,\ldots,P_M)^+\Delta\mathbb{V}\mathrm{ar}_{P_0}(X).
\end{aligned}$$

For $\Delta^\top\chi^2(P_0,\ldots,P_M)^+\Delta = 0$, the asserted inequality is trivially true. For $\Delta^\top\chi^2(P_0,\ldots,P_M)^+\Delta > 0$ the claim follows by dividing both sides by $\Delta^\top\chi^2(P_0,\ldots,P_M)^+\Delta$. □

In particular, if the χ^2-divergence matrix is diagonal, we obtain $\sum_{j=1}^M (E_{P_j}[X]-E_{P_0}[X])^2/\chi^2(P_j,P_0) \leq \mathbb{V}\mathrm{ar}_{P_0}(X)$. It should be observed that because of the sum, this inequality produces better lower bounds than (4). As mentioned above, we know that a vector $v=(v_1,\ldots,v_M)$ lies in the kernel of the χ^2-divergence matrix if and only if $\sum_{j=1}^M v_j(P_j-P_0)=0$. This shows that such a v and the vector Δ must be orthogonal. Thus, Δ is orthogonal to the kernel of $\chi^2(P_0,\ldots,P_M)$ and

$$\sum_{j=1}^M \big(E_{P_j}[X]-E_{P_0}[X]\big)^2 \leq \lambda_1\big(\chi^2(P_0,\ldots,P_M)\big)\mathbb{V}\mathrm{ar}_{P_0}(X), \tag{11}$$

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

$$\sum_{j=1}^M \big(E_{P_j}[X]-E_{P_0}[X]\big)^2 \leq \big\|\chi^2(P_0,\ldots,P_M)\big\|_{1,\infty}\mathbb{V}\mathrm{ar}_{P_0}(X). \tag{12}$$

Whatever variation of Theorem 2.3 is applied to derive lower bounds on the bias-variance trade-off, the key problem is the computation of the χ^2-divergence matrix for given probability measures P_{θ_j}, $j = 0, \ldots, M$ 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 $(Q_\theta : \theta \in \Theta)$. Formally, this means that $P_\theta = KQ_\theta$ for all $\theta \in \Theta$ with K a Markov kernel that is independent of θ. Then by applying the data processing inequality below, we have the matrix inequality $\chi^2(P_{\theta_0}, \ldots, P_{\theta_M}) \leq \chi^2(Q_{\theta_0}, \ldots, Q_{\theta_M})$. We therefore can apply the previous theorem with $\chi^2(P_{\theta_0}, \ldots, P_{\theta_M})$ replaced by $\chi^2(Q_{\theta_0}, \ldots, Q_{\theta_M})$.

Theorem 2.4 (Data processing / entropy contraction)**.** *If K is a Markov kernel and $Q_0, \ldots, Q_M$ are probability measures such that Q_0 dominates $Q_1, \ldots, Q_M$, then,*

$$\chi^2\big(KQ_0, \ldots, KQ_M\big) \leq \chi^2\big(Q_0, \ldots, Q_M\big),$$

where $\leq$ denotes the order on the set of positive semi-definite matrices.

In particular, the χ^2-divergence matrix is invariant under invertible transformations. A specific application for the combination of general lower bounds and the data processing inequality is given in Section 5.

Theorem 2.3 contains the multivariate Cramer-Rao lower bound as a special case. Indeed, for a parameter vector $\theta \in \mathbb{R}^p$ and an estimator $\widehat{\theta}$, consider the measures $P_0 := P_\theta$ and $P_i = P_{\theta + he_i}$ for $i = 1, \ldots, p$ with $(e_i)_i$ the canonical basis of $\mathbb{R}^p$. Up to scaling by h, the matrix $\widetilde{\Delta} := (E_\theta[\widehat{\theta}_i] - E_{\theta+he_i}[\widehat{\theta}_j])_{1 \leq i,j \leq p}$ can be viewed as a discretized version of the Jacobian matrix $\operatorname{Jac}_\theta(E_\theta[\widehat{\theta}]) := (\partial_{\theta_i} E_\theta[\widehat{\theta}_j])_{1 \leq i,j \leq p}$, that is, $\widetilde{\Delta}/h \to \operatorname{Jac}_\theta(E_\theta[\widehat{\theta}])$ as $h \to 0$. Denote by $\operatorname{Cov}_\theta(\widehat{\theta})$ the covariance matrix of the vector $\widehat{\theta}$. Applying Theorem 2.3 with $X = t^\top \widehat{\theta}$ for a vector $t \in \mathbb{R}^d$ and using the linearity of the expectation, we get $t^\top \widetilde{\Delta}^\top \chi^2(P_0, \ldots, P_M)^+ \widetilde{\Delta} t \leq t^\top \operatorname{Cov}_\theta(\widehat{\theta}) t$. Under suitable regularity conditions (mainly, inversion of integral and derivative signs), the matrix $\chi^2(P_0, \ldots, P_M)/h^2$ tends to the inverse of the Fisher information matrix $F(\theta)$ as $h \to 0$. For $h \to 0$, the previous inequality yields $t^\top \operatorname{Jac}_\theta(E_\theta[\widehat{\theta}])^\top F(\theta)^{-1} \operatorname{Jac}_\theta(E_\theta[\widehat{\theta}]) t \leq t^\top \mathbb{V}\mathrm{ar}_\theta(\widehat{\theta}) t$. As this is true for any vector t, the multivariate Cramér-Rao inequality $\operatorname{Jac}_\theta(E_\theta[\widehat{\theta}])^\top F(\theta)^{-1} \operatorname{Jac}_\theta(E_\theta[\widehat{\theta}]) \leq \mathbb{V}\mathrm{ar}_\theta(\widehat{\theta})$ follows, where $\leq$ denotes the order on the set of positive semi-definite matrices. The concept of Fisher Φ-information also generalizes the Fisher information using information measures, see [11, 32]. It is worth mentioning that this notion is not comparable with our approach and only applies to Markov processes.

The connection to the Cramér-Rao inequality suggests that for a given statistical problem with a p-dimensional parameter space, one should apply Theorem 2.3 with $M = p$. It turns out that for the high-dimensional models discussed in Section 5 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.

There exists also an analogue of Theorem 2.3 for the Hellinger affinity matrix. This is stated and proved next.

Theorem 2.5. *Given probability measures $P_1, \ldots, P_M$, let $A_\ell := \rho(P_\ell | P_1, \ldots, P_{\ell-1}, P_{\ell+1}, \ldots, M)$. Then, for* *any random variable* X*,*

$$2M\sum_{j=1}^M\Big(E_j[X]-\frac{1}{M}\sum_{\ell=1}^M E_\ell[X]\Big)^2=\sum_{j,k=1}^M (E_j[X]-E_k[X])^2\leq 4\max_\ell \lambda_1(A_\ell)\sum_{k=1}^M \mathbb{V}ar_{P_k}(X).$$

Proof of Theorem 2.5. The first identity is elementary and follows from expansion of the squares. It therefore remains to prove the inequality. To keep the mathematical expressions readable, we agree to write $E_j := E_{P_j}[X]$ and $V_j := \mathbb{V}\mathrm{ar}_{P_j}(X)$. Furthermore, we omit the integration variable as well as the differential in the integrals. Rewriting, we find that for any real number $\alpha_{j,k}$,

$$\big(E_j-E_k\big)\int\sqrt{p_kp_j}=\int(X-E_k)\sqrt{p_k}\big(\sqrt{p_j}-\alpha_{j,k}\sqrt{p_k}\big)+\int(X-E_j)\sqrt{p_j}\big(\alpha_{k,j}\sqrt{p_j}-\sqrt{p_k}\big). \tag{13}$$

From now on, we choose $\alpha_{j,k}$ to be $\int\sqrt{p_jp_k}$. Observe that for this choice, the term $\alpha_{j,k}\sqrt{p_k}$ is the L^2-projection of $\sqrt{p_j}$ on $\sqrt{p_k}$. Dividing by $\int\sqrt{p_kp_j}$, summing over (j,k), interchanging the role of j and k for the second term in the second line, applying the Cauchy-Schwarz inequality first to each of the M integrals and then also to bound the sum over k, and using (10) gives

$$\begin{aligned}
&\sum_{j,k=1}^M \big(E_j-E_k\big)^2\\
&=\sum_{k=1}^M\int(X-E_k)\sqrt{p_k}\sum_{j=1}^M\Big(\frac{\sqrt{p_j}}{\int\sqrt{p_kp_j}}-\sqrt{p_k}\Big)(E_j-E_k)+\sum_{j=1}^M\int(X-E_j)\sqrt{p_j}\sum_{k=1}^M\Big(\sqrt{p_j}-\frac{\sqrt{p_k}}{\int\sqrt{p_kp_j}}\Big)(E_j-E_k)\\
&=2\sum_{k=1}^M\int(X-E_k)\sqrt{p_k}\sum_{j=1}^M\Big(\frac{\sqrt{p_j}}{\int\sqrt{p_kp_j}}-\sqrt{p_k}\Big)(E_j-E_k)\\
&\leq 2\sum_{k=1}^M\sqrt{V_k\int\Big(\sum_{j=1}^M\Big(\frac{\sqrt{p_j}}{\int\sqrt{p_kp_j}}-\sqrt{p_k}\Big)(E_j-E_k)\Big)^2}\\
&\leq 2\sqrt{\sum_{r=1}^M V_r}\sqrt{\sum_{k=1}^M\int\Big(\sum_{j=1}^M\Big(\frac{\sqrt{p_j}}{\int\sqrt{p_kp_j}}-\sqrt{p_k}\Big)(E_j-E_k)\Big)^2}\\
&\leq 2\sqrt{\sum_{r=1}^M V_r}\sqrt{\sum_{k=1}^M\lambda_1(A_k)\sum_{j=1}^M(E_j-E_k)^2}\\
&\leq 2\max_\ell\lambda_1(A_\ell)\sqrt{\sum_{r=1}^M V_r}\sqrt{\sum_{k,j=1}^M(E_j-E_k)^2}.
\end{aligned}$$

Squaring both sides and dividing by $\sum_{k,j=1}^M(E_j-E_k)^2$ yields the claim. ☐

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

Before discussing a number of specific statistical models, it is worth mentioning that 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 very large. Whenever Θ is a class of functions on $[0, 1]$, the lower bounds below 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 $\mathcal{A} = \mathbb{R}^p$.

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

$$dY_x = f(x)\,dx + n^{-1/2}\,dW_x, \tag{14}$$

where W is an unobserved 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 6, 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. The likelihood ratio in the Gaussian white noise model is given by Girsanov's formula $dP_f/dP_0 = \exp(n\int_0^1 f(t)dY_t - \frac{n}{2}\|f\|_2^2)$ whenever $f \in L^2([0, 1])$. In particular, under P_f and for a function $g : [0, 1] \to \mathbb{R}$, we have that

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

with W a standard Brownian motion and $\xi \sim \mathcal{N}(0, 1)$ (under P_f). 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 $\lfloor\beta\rfloor$ the largest integer that is strictly smaller than β. On a domain $D \subseteq \mathbb{R}$, define the β-Hölder norm $\|f\|_{\mathscr{C}^\beta(D)} = \sum_{\ell\le\lfloor\beta\rfloor}\|f^{(\ell)}\|_{L^\infty(D)} + \sup_{x,y\in D,x\neq y}|f^{(\lfloor\beta\rfloor)}(x) - f^{(\lfloor\beta\rfloor)}(y)|/|x - y|^{\beta-\lfloor\beta\rfloor}$, with $L^\infty(D)$ the supremum norm on D. For $D = [0, 1]$, let $\mathscr{C}^\beta(R) := \{f : [0, 1] \to \mathbb{R} : \|f\|_{\mathscr{C}^\beta([0,1])} \le R\}$ 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 $\widehat{f}(x_0) = (2h)^{-1}\int_{x_0-h}^{x_0+h} dY_t$. Assume that x_0 is not at the boundary such that $0 \le x_0 - h$ and $x_0 + h \le 1$. Bias and variance for this estimator are

$$\mathrm{Bias}_f\big(\widehat{f}(x_0)\big) = \frac{1}{2h}\int_{x_0-h}^{x_0+h}\big(f(u) - f(x_0)\big)\,du, \quad \mathbb{V}\mathrm{ar}_f\big(\widehat{f}(x_0)\big) = \frac{1}{2nh}.$$

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

$$\sup_{f\in\mathscr{C}^\beta(R)} |\operatorname{Bias}_f(\widehat{f}(x_0))|^{1/\beta} \inf_{f\in\mathscr{C}^\beta(R)} \mathbb{V}\mathrm{ar}_f(\widehat{f}(x_0)) \gtrsim \frac{1}{n}. \tag{15}$$

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 [27] 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.

[27] 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 $w(\varepsilon) := \sup\big\{|L(f-g)| : \|f-g\|_{L^2[0,1]} \leq \varepsilon, f, g \in \Theta\big\}$ is the so-called modulus of continuity, Theorem 2 in [27] rewritten in our notation states that, if Θ is closed and convex and $\lim_{\varepsilon\downarrow 0} w(\varepsilon) = 0$, then

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

and

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

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 C.1, $g \in \mathscr{C}^\beta(R)$ 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 $w(\varepsilon) \geq (\varepsilon/\|K\|_2)^{\beta/(\beta+1/2)}K(0)$. Some calculations show then that

$$\inf_{\widehat{f}(x_0):\sup_{f\in\mathscr{C}^\beta(R)} \mathbb{V}\mathrm{ar}_f(\widehat{f}(x_0))\leq V} \ \sup_{f\in\mathscr{C}^\beta(R)} \operatorname{Bias}_f\big(\widehat{f}(x_0)\big)^2 \gtrsim \frac{1}{(nV)^{2\beta}}$$

and

$$\inf_{\widehat{f}(x_0):\sup_{f\in\mathscr{C}^\beta(R)} |\operatorname{Bias}_f(\widehat{f}(x_0))|\leq B} \ \sup_{f\in\mathscr{C}^\beta(R)} \mathbb{V}\mathrm{ar}_f\big(\widehat{f}(x_0)\big) \gtrsim \frac{1}{nB^{1/\beta}}.$$

The worst-case bias and the worst-case variance are thus both lower-bounded. The result is comparable to (15) 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

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

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. *Let $\beta,R>0$, $x_0\in[0,1]$ and let $\gamma(R,\beta)$ and $\overline{\gamma}(R,\beta,C,f)$ be the constants defined above. Assign to $(+\infty)\cdot 0$ the value $+\infty$.*
(i): If $\mathcal{T}=\{\widehat{f}:\sup_{f\in\mathscr{C}^\beta(R)}\big|\operatorname{Bias}_f\big(\widehat{f}(x_0)\big)\big|<1\}$, then,

$$\inf_{\widehat{f}\in\mathcal{T}}\sup_{f\in\mathscr{C}^\beta(R)}\big|\operatorname{Bias}_f\big(\widehat{f}(x_0)\big)\big|^{1/\beta}\sup_{f\in\mathscr{C}^\beta(R)}\mathbb{V}ar_f\big(\widehat{f}(x_0)\big)\geq\frac{\gamma(R,\beta)}{n}. \tag{16}$$

(ii): Let $\mathcal{S}(C):=\{\widehat{f}:\sup_{f\in\mathscr{C}^\beta(R)}|\operatorname{Bias}_f(\widehat{f}(x_0))|<(C/n)^{\beta/(2\beta+1)}\}$, then, for any $C>0$,

$$\inf_{\widehat{f}\in\mathcal{S}(C)}\sup_{f\in\mathscr{C}^\beta(R)}\big|\operatorname{Bias}_f\big(\widehat{f}(x_0)\big)\big|^{1/\beta}\inf_{f\in\mathscr{C}^\beta(R)}\frac{\mathbb{V}ar_f(\widehat{f}(x_0))}{\overline{\gamma}(R,\beta,C,\|f\|_{\mathscr{C}^\beta})}\geq\frac{1}{n}. \tag{17}$$

Both statements can be easily derived from the abstract lower bounds in Section 2. A full proof is given in Supplement C. 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. Compared with (15), the lower bound depends on the $\mathscr{C}^\beta$-norm of f through $\overline{\gamma}(R,\beta,C,\|f\|_{\mathscr{C}^\beta})$. This quantity deteriorates if f is on the boundary of the Hölder ball. A direct consequence of (ii) is the uniform bound

$$\inf_{\widehat{f}\in\mathcal{S}(C)}\sup_{f\in\mathscr{C}^\beta(R)}\big|\operatorname{Bias}_f\big(\widehat{f}(x_0)\big)\big|^{1/\beta}\inf_{f\in\mathscr{C}^\beta(a)}\operatorname{Var}_f\big(\widehat{f}(x_0)\big)\geq\frac{\inf_{b\leq a}\overline{\gamma}(R,\beta,C,b)}{n},$$

providing a non-trivial lower bound if $a<R-1$.

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 $R\leq 1$ and consider the estimator $\widehat{f}(x_0)=0$. Notice that for any

$f \in \mathscr{C}^\beta(R)$, $|\operatorname{Bias}_f(\widehat{f}(x_0))| = |f(x_0)| \leq \|f\|_\infty \leq 1$ and $\mathbb{V}\mathrm{ar}_f(\widehat{f}(x_0)) = 0$. The left-hand side of the inequality (16) 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 $\widehat{f}(x_0) \in T$, there exists an $f \in \mathscr{C}^\beta(R)$ with $|\operatorname{Bias}_f(\widehat{f}(x_0))|^{1/\beta}\mathbb{V}\mathrm{ar}_f(\widehat{f}(x_0)) \geq \gamma(R,\beta)/n$ and thus for such an f,

$$\operatorname{MSE}_f\left(\widehat{f}(x_0)\right) = \operatorname{Bias}_f\left(\widehat{f}(x_0)\right)^2 + \mathbb{V}\mathrm{ar}_f\left(\widehat{f}(x_0)\right) \geq \left(\frac{\gamma(R,\beta)}{n\mathbb{V}\mathrm{ar}_f\left(\widehat{f}(x_0)\right)}\right)^{2\beta} + \frac{\gamma(R,\beta)}{n|\operatorname{Bias}_f\left(\widehat{f}(x_0)\right)|^{1/\beta}}.$$

This shows that small bias or small variance increases the mean squared error. It also implies that the minimax rate $n^{-2\beta/(2\beta+1)}$ can only be achieved for estimators balancing the worst-case squared bias and the worst-case variance.

For nonparametric problems, an estimator can be superefficient for many parameters simultaneously, see [7]. 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. This can be viewed as a bias-variance formulation of the constrained risk problem, see Section B in [6]. Statement (ii) of the previous theorem shows that such constructions are impossible if R is large enough.

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 $(Y_1,\ldots,Y_n)$ with $Y_i = f(i/n) + \varepsilon_i$, $i = 1,\ldots,n$ 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 $\operatorname{KL}(P_f,P_g) = \frac{n}{2}\|f-g\|_n^2$ 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 Pointwise estimation of the boundary

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 boundary estimation model.

Consider the model, where we observe a Poisson point process (PPP) $N = \sum_i \delta_{(X_i,Y_i)}$ with intensity $\lambda_f(x,y) = n\mathbf{1}(f(x) \leq y)$ 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 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 [28].

For a parametric estimation problem, we can typically achieve the estimation rate n^{-1} in this model. This 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 $n^{-2\beta/(2\beta+1)}$.

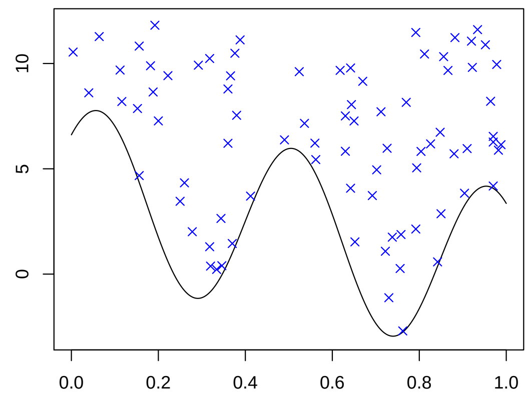

Figure 1: Generated data (blue) and support boundary (black) for PPP model.

The information measures in this model are governed by the L^1-geometry. If P_f denotes the distribution of the data for support boundary f, then it can be shown that P_f is dominated by P_g if and only if $g \leq f$ pointwise. If indeed $g \leq f$, the likelihood ratio is given by $dP_f/dP_g = \exp(n\int_0^1 (f(x)-g(x))\,dx)\mathbf{1}(\forall i: f(X_i) \leq Y_i)$, see Lemma 2.1 in [33]. In particular, we have for $g \leq f$, $\alpha > 0$, and $\|\cdot\|_1$ the $L^1([0,1])$-norm, $E_g[(dP_f/dP_g)^\alpha] = \exp(n\|f-g\|_1(\alpha-1))E_g[dP_f/dP_g] = \exp(n\|f-g\|_1(\alpha-1))$ and so $H(P_f,P_g) = 1-\exp(-\frac{n}{2}\|f-g\|_1)$ and $\chi^2(P_f,P_g) = \exp(n\|f-g\|_1)-1$.

Since $\mathrm{KL}(P_f,P_g)+\mathrm{KL}(P_g,P_f) = \infty$ whenever $f \neq g$, the Kullback-Leibler version of Lemma 2.1 is not applicable in this case. Also we argued earlier that for regular models, we can retrieve the Cramér-Rao lower bound from the lower bounds in Lemma 2.1 by choosing $P = P_\theta$, $Q = P_{\theta+\Delta}$ and letting Δ tend to 0. As no Fisher information exists in the support boundary model, it is of interest to study the abstract lower bounds in Lemma 2.1 under the limit $\Delta \to 0$. For this, consider constant support boundaries $f_\theta = \theta$. It is then natural to evaluate the lower bounds for $X = \min_i Y_i$, which can be shown to be a sufficient statistic for θ. Moreover, under P_{f_θ}, $X-\theta$ follows an exponential distribution with rate parameter n (see also Proposition 3.1 in [35] and Section 4.1 in [34]). With $P = P_{f_\theta}$ and $Q = P_{f_{\theta+\Delta}}$, $(H^{-1}(P,Q)-H(P,Q))^{-2} = e^{n\Delta}(1-e^{-n\Delta/2})$ and $\chi^2(P,Q)\wedge\chi^2(Q,P) = e^{n\Delta}-1$. Since $E_P[X] = \theta+1/n$, $E_Q[X] = \theta+\Delta+1/n$, and $\mathbb{V}\mathrm{ar}_P(X) = \mathbb{V}\mathrm{ar}_Q(X) = 1/n^2$, we find that the Hellinger lower bound (2) can be rewritten as $\Delta^2 \leq 8e^{n\Delta}(1-e^{-n\Delta/2})/n^2$ and the χ^2-divergence lower bound (4) becomes $\Delta^2 \leq (e^{n\Delta}-1)/n^2$. In both inequalities the upper bound is of the order Δ^2 if $\Delta \asymp 1/n$. Otherwise the inequalities are suboptimal. In particular, the Cramér-Rao asymptotics $\Delta \to 0$ for fixed n does not yield anything useful here. For the bias-variance trade-off this asymptotic regime is, however, less important and we still can obtain rate-optimal bounds by applying the inequalities in the regime $\Delta \asymp 1/n$.

Theorem 4.1. *Let $0 < \beta < 1$, $C > 0$ and $R > \kappa := 2\inf_{K\in L^2(\mathbb{R})}\{\|K\|_{\mathscr{C}^\beta(\mathbb{R})} : K(0) = 1, K \geq 0\}$. For any estimator $\widehat{f}$ with*

$$\sup_{f\in\mathscr{C}^\beta(R)} \mathrm{MSE}_f\left(\widehat{f}(x_0)\right) < \left(\frac{C}{n}\right)^{\frac{2\beta}{\beta+1}},$$

there exist positive constants $c := c(\beta, C, R)$ and $c' := c'(\beta, C, R)$ such that

$$\sup_{f\in\mathscr{C}^\beta(R)} \mathrm{Bias}_f\left(\widehat{f}(x_0)\right)^2 \geq cn^{-\frac{2\beta}{\beta+1}} \tag{18}$$

and

$$\mathbb{V}ar_f\big(\widehat{f}(x_0)\big) \geq c'n^{-\frac{2\beta}{\beta+1}}, \quad \text{for all } f \in \mathscr{C}^\beta\big((R-\kappa)/2\big). \tag{19}$$

The theorem is proved in Supplement C. It states 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 MSE) is always lower-bounded by $\gtrsim n^{-2\beta/(\beta+1)}$. Another consequence of the theorem is that $n^{-2\beta/(\beta+1)}$ is a lower bound for the mean squared error. The smoothness constraint $\beta \leq 1$ is fairly common in the literature on support boundary estimation, see [35].

5 The bias-variance trade-off for high-dimensional models

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 $(\theta_1, \ldots, \theta_n)$ with at most s non-zero components. For any estimator $\widehat{\theta}$, the bias-variance decomposition of the mean squared error of $\hat{\theta}$ is

$$E_\theta\big[\big\|\widehat{\theta}-\theta\big\|^2\big] = \big\|E_\theta\big[\widehat{\theta}\big]-\theta\big\|^2 + \sum_{i=1}^n \mathbb{V}\mathrm{ar}_\theta\big(\widehat{\theta}_i\big), \tag{20}$$

where the first term on the right-hand side plays the role of the 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. The most extreme variance reduction occurs for the case of a completely black signal, that is, $\theta = (0, \ldots, 0)^\top$. 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 5.1. *Consider the Gaussian sequence model under sparsity. Let $n \geq 4$ and $0 < s \leq \sqrt{n}/2$. Given an estimator $\widehat{\theta}$ and a real number γ such that $4\gamma + 1/\log(n/s^2) \leq 0.99$ and*

$$\sup_{\theta\in\Theta(s)} \big\|E_\theta\big[\widehat{\theta}\big]-\theta\big\|^2 \leq \gamma s \log\left(\frac{n}{s^2}\right),$$

then, for all sufficiently large n,

$$\sum_{i=1}^{n} \mathbb{V}ar_0\big(\widehat{\theta}_i\big) \geq \frac{(1-(1/2)^{0.01})}{25e\log(n/s^2)} n\Big(\frac{s^2}{n}\Big)^{4\gamma},$$

where $\mathbb{V}ar_0$ denotes the variance for parameter vector $\theta = (0, \ldots, 0)^\top$.

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 $s \leq n^{1/2-\delta}$, with δ a small positive number, the lower bound tends to $n/\log(n)$ if γ is made small (since everything is non-asymptotic, we can even allow γ to depend on n). As a consequence of the bias-variance decomposition (20), the maximum quadratic risk of such an estimator is lower-bounded by a rate that, for small γ, is close to $n/\log(n)$. Thus, already reducing the constant of the bias will necessarily lead to estimators with highly suboptimal risk.

The proof of Theorem 5.1 applies the χ^2-divergence lower bound (12) by comparing the data distribution induced by the zero vector to the $\binom{n}{s}$ many distributions corresponding to s-sparse vectors with non-zero entries $\sqrt{4\gamma\log(n/s^2)+1}$. By Table (1), 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 non-zero. The whole problem reduces then to a combinatorial counting argument. The key observation is that if we fix a s-sparse vector, say θ^*, there are of the order n/s^2 more s-sparse vectors that have exactly $r-1$ non-zero components in common with θ^* than s-sparse vectors that that have exactly r non-zero 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^*.

One might wonder whether the proposed lower bound technique can be extended for sparsity $s \gg \sqrt{n}$. The fact that $\sqrt{n}$ appears as an upper bound on the sparsity might be related to the testing theory in the Gaussian sequence model. It is well-known that for sparse models with sparsity $s \ll \sqrt{n}$, we cannot consistently test for signal in the sparse Gaussian mixture formulation. On the contrary, for any $s = n^{1/2+\delta}$ with $\delta > 0$ this is indeed possible, see [21, 13, 10].

The lower bound in Theorem 5.1 can be extended to several related problems by invoking the data processing inequality in Theorem 2.4. As an example suppose that we observe only $X_1^2, \ldots, X_n^2$ with $(X_1, \ldots, X_n)$ the data from the Gaussian sequence model. As parameter space, consider the class $\Theta_+(s)$ of s-sparse vectors with non-negative entries. Since the proof of Theorem 5.1 only uses parameters in $\Theta_+(s)$, the same lower bound as in Theorem 5.1 holds also in this modified setting.

Proof of Theorem 5.1. For each $i = 1, \ldots, n$, we derive a lower bound for $\mathbb{V}\mathrm{ar}_0(\widehat{\theta}_i)$ applying (12). Denote by P_θ the distribution of the data in the Gaussian sequence model for the parameter vector $\theta = (\theta_1, \ldots, \theta_n)$. Fix an integer $i \in \{1, \ldots, n\}$. There are $M := \binom{n-1}{s-1}$ distinct vectors $\theta_1^{(i)}, \ldots, \theta_M^{(i)}$ with exactly s non-zero entries, having a non-zero entry at the i-th position and all non-zero entries equal to $\sqrt{\alpha\log(n/s^2)}$, where $\alpha := 4\gamma + 1/\log(n/s^2)$. To indicate also the dependence on i, for each $j \in \{1, \ldots, M\}$ write $P_{ji} := P_{\theta_j^{(i)}}$ and $P_0 = P_{(0,\ldots,0)}$.

As mentioned in Table 1 (see also (32)), we have that $\chi^2(P_0,\dots,P_M)_{j,k} = \exp(\langle \theta_j^{(i)}, \theta_k^{(i)}\rangle) - 1$. For fixed j, there are $b(n,s,r) := \binom{s-1}{r-1}\binom{n-s}{s-r}$ among the M vectors $\theta_1^{(i)},\dots,\theta_M^{(i)}$ with exactly r non-zero components with $\theta_j^{(i)}$ in common, that is, $\langle \theta_j^{(i)}, \theta_k^{(i)}\rangle = \alpha r \log(n/s^2)$. Hence,

$$\big\|\chi^2(P_0, P_{1i}\dots, P_{Mi})\big\|_{1,\infty} = \sum_{r=1}^s b(n,s,r)\Big[\Big(\frac{n}{s^2}\Big)^{r\alpha} - 1\Big].$$

Since $s \le \sqrt{n}/2$, we have for $r = 0,\dots,s-1$,

$$b(n,s,r+1) = \frac{(s-r)^2}{r(n-2s+r+1)} b(n,s,r) \le \frac{s^2}{n(1-n^{-1/2})} b(n,s,r).$$

Recall that $\alpha = 4\gamma + 1/\log(n/s^2) \le 0.99$. Thus, for all sufficiently large n, $(1-n^{-1/2})^{-1}(s^2/n)^{1-\alpha} \le (1-n^{-1/2})^{-1}(1/4)^{0.01} \le (1/2)^{0.01}$. Combined with the recursion formula for $b(n,s,r)$ and the formula for the geometric sum, we obtain

$$\begin{aligned}\big\|\chi^2(P_0, P_{1i}\dots, P_{Mi})\big\|_{1,\infty} &\le b(n,s,1)\Big(\frac{n}{s^2}\Big)^{\alpha} \sum_{q=0}^{s-1} \frac{1}{(1-n^{-1/2})^q}\Big(\frac{s^2}{n}\Big)^{q(1-\alpha)} \\ &\le b(n,s,1)\frac{1}{1-(1/2)^{0.01}}\Big(\frac{n}{s^2}\Big)^{\alpha},\end{aligned}$$

where the last inequality holds for all sufficiently large n. We must have that $M = \sum_{r=1}^s b(n,s,r)$ and so $b(n,s,1) \le M$. Let $\widehat{\theta} = (\widehat{\theta}_1,\dots,\widehat{\theta}_n)$ be an arbitrary estimator for θ. Applying Theorem 2.3 to the random variable $\widehat{\theta}_i$ yields

$$\sum_{j=1}^M \big(E_{P_{ji}}[\widehat{\theta}_i] - E_{P_0}[\widehat{\theta}_i]\big)^2 \le M\frac{1}{1-(1/2)^{0.01}}\Big(\frac{n}{s^2}\Big)^{\alpha}\mathbb{V}\mathrm{ar}_0\big(\widehat{\theta}_i\big). \tag{21}$$

Let $\mathcal{M}$ be the set of all $\binom{n}{s}$ distributions $P \sim \mathcal{N}(\theta, I_n)$, where the mean vector θ has exactly s non-zero entries and all non-zero entries equal to $\sqrt{\alpha\log(n/s^2)}$. For a $P \in \mathcal{M}$ denote by $S(P)$ the support (the location of the non-zero entries) of the corresponding mean vector θ. For $S \subset \{1,\dots,n\}$, define moreover $\widehat{\theta}_S := (\widehat{\theta}_j)_{j\in S}$. Summing over i in (21) yields then,

$$\sum_{P\in\mathcal{M}} \big\|E_P\big[\widehat{\theta}_{S(P)}\big] - E_{P_0}\big[\widehat{\theta}_{S(P)}\big]\big\|_2^2 \le M\frac{1}{1-(1/2)^{0.01}}\Big(\frac{n}{s^2}\Big)^{\alpha}\sum_{i=1}^n \mathbb{V}\mathrm{ar}_0\big(\widehat{\theta}_i\big). \tag{22}$$

For any $P \in \mathcal{M}$ with $P = \mathcal{N}(\theta, I_d)$, we obtain using triangle inequality, $\theta_0 = 0$, $\|\theta\|_2 = \|\theta_{S(P)}\|_2$, the bound on the bias, and $\alpha = 4\gamma + 1/\log(n/s^2) \le 1$ combined with $\sqrt{\alpha} - 2\sqrt{\gamma} = (\alpha-4\gamma)/(\sqrt{\alpha}+2\sqrt{\gamma}) \ge (\alpha-4\gamma)/5 = 1/(5\log(n/s^2))$,

$$\begin{aligned}\big\|E_P\big[\widehat{\theta}_{S(P)}\big] - E_{P_0}\big[\widehat{\theta}_{S(P)}\big]\big\|_2^2 &\ge \|\theta\|_2 - \big\|E_{P_j}[\widehat{\theta}] - \theta\big\|_2 - \big\|E_{P_0}[\widehat{\theta}]\big\|_2 \\ &\ge \sqrt{s\alpha\log(n/s^2)} - 2\sqrt{\gamma s\log(n/s^2)} \\ &\ge \sqrt{\frac{s}{25\log(n/s^2)}}.\end{aligned}$$

Observe that $(n/s^2)^{\alpha} = (n/s^2)^{4\gamma}e$ and that the cardinality of $\mathcal{M}$ is $\binom{n}{s} = \frac{n}{s}\binom{n-1}{s-1} = \frac{n}{s}M$. Combining this with (22) yields

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

completing the proof. $\square$

We now establish an upper bound. 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. The bound on the bias dictates the largest possible truncation level. One can further reduce the variance at zero if large observations are also shrunk as much as possible to zero. To obtain matching upper bounds, this motivates then to study the soft-thresholding estimator

$$\widehat{\theta}_i = \operatorname{sign}(X_i)\Big(|X_i| - \sqrt{\gamma \log(n/s^2)}\Big), \quad i = 1, \ldots, n. \tag{23}$$

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 see that the squared bias $\|E_\theta[\widehat{\theta}] - \theta\|_2^2$ is bounded by $\gamma s \log(n/s^2)$.

Lemma 5.2. *For the soft-thresholding estimator $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)^\top$ defined in* (23)*, we have*

$$\sum_{i=1}^n \mathbb{V}ar_0\big(\widehat{\theta}_i\big) \leq \frac{\sqrt{2}}{\sqrt{\pi \gamma^3 \log^3(n/s^2)}} n \Big(\frac{s^2}{n}\Big)^{\frac{\gamma}{2}}.$$

Proof of Lemma 5.2. Let $T := \sqrt{\gamma \log(n/s^2)}$ denote the truncation value. Since $\widehat{\theta}$ is unbiased under $\theta = (0, \ldots, 0)^\top$, we have for any $i = 1, \ldots, n$ using substitution twice

$$\begin{aligned}\mathbb{V}\mathrm{ar}_0\big(\widehat{\theta}_i\big) &= \sqrt{\frac{2}{\pi}} \int_T^\infty (x - T)^2 e^{-\frac{x^2}{2}}\, dx = \sqrt{\frac{2}{\pi}} \int_0^\infty x^2 e^{-\frac{(x+T)^2}{2}}\, dx \\ &\leq \sqrt{\frac{2}{\pi}} e^{-\frac{T^2}{2}} \int_0^\infty x^2 e^{-xT}\, dx = \frac{\sqrt{2}}{\sqrt{\pi} T^3} e^{-\frac{T^2}{2}} \int_0^\infty y^2 e^{-y}\, dy = \frac{\sqrt{2}}{\sqrt{\pi} T^3} e^{-\frac{T^2}{2}}.\end{aligned}$$

Summing over i and inserting the expression for T yields the result. $\square$

The upper and lower bound have the same structure. Key differences are that the exponent is 4γ in the lower bound and $\gamma/2$ in the lower bound. As discussed already, this 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 by relying on the Hellinger version based on Theorem 2.5. 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 $\sum_{i=1}^n \mathbb{V}\mathrm{ar}_0(\widehat{\theta}_i)$ is of order $n(s^2/n)^{\gamma/2}$.

The soft-thresholding estimator does not produce an s-sparse model. Indeed, from the tail decay of the Gaussian distribution, one expects that the sparsity of the reconstruction is $n(s^2/n)^{\gamma/2}$ which can be considerable bigger than s. 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.

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. As common, we assume that the diagonal coefficients of the Gram matrix $X^\top X$ are standardized such that $(X^\top X)_{i,i} = n$ for all $i = 1, \ldots, p$ (see for instance also Section 6 in [8]). Define the mutual coherence condition number by $\mathrm{mc}(X) := \max_{1 \leq i \neq j \leq n} (X^\top X)_{i,j}/(X^\top X)_{i,i}$. This notion goes back to [14]. Below, we work under the restriction $\mathrm{mc}(X) \leq 1/(s^2 \log(p/s^2))$. This is stronger than the mutual coherence bound of the form const.$/s$ 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.

Theorem 5.3. *Consider the sparse high-dimensional regression model with Gaussian noise. Let* $p \geq 4$, $0 < s \leq \sqrt{p}/2$, *and* $\mathrm{mc}(X) \leq 1/(s^2 \log(p/s^2))$. *Given an estimator* $\widehat{\beta}$ *and a real number* γ *such that* $4\gamma + 1/\log(p/s^2) \leq 0.99$ *and*

$$\sup_{\beta \in \Theta(s)} \left\| E_\beta\big[\widehat{\beta}\big] - \beta \right\|^2 \leq \gamma \frac{s}{n} \log\Big(\frac{p}{s^2}\Big),$$

then, for all sufficiently large p,

$$\sum_{i=1}^{p} \mathbb{V}ar_0\big(\widehat{\beta}_i\big) \geq \frac{(1 - (1/2)^{0.01})}{25 e^2 \log(p/s^2)} \frac{p}{n} \Big(\frac{s^2}{p}\Big)^{4\gamma},$$

where $\mathbb{V}ar_0$ *denotes the variance for parameter vector* $\beta = (0, \ldots, 0)^\top$.

Proof of Theorem 5.3. The proof is a variation of the proof of Theorem 5.1 with n replaced by p. To comply with standard notation, the parameter vectors are denoted by β and therefore all the symbols θ in the proof of Theorem 5.1 have to be replaced by β. In particular the vectors θ_j are now denoted by β_j. Because of the standardization of the diagonal entries in the Gram matrix, we need to choose the non-zero components of β_j as $\sqrt{\alpha \log(p/s^2)/n}$. Compared with the proof of Theorem 5.1, the main difference is that the entries of the χ^2-divergence matrix are bounded as

$$\begin{aligned}
\chi^2(P_0, \ldots, P_M)_{j,k} &= \exp(\beta_j^\top X^\top X \beta_k) - 1 \\
&\leq \exp(n \beta_j^\top \beta_k + n s^2 \,\mathrm{mc}(X) \|\beta_j\|_\infty \|\beta_k\|_\infty) \\
&\leq \exp(n \beta_j^\top \beta_k + \alpha) \\
&\leq \exp(n \beta_j^\top \beta_k + 1),
\end{aligned}$$

where the first inequality follows from separating the diagonal from the off-diagonal entries and exploiting that the vectors are s-sparse and the second inequality uses that the maximum entry norm $\|\cdot\|_\infty$ is bounded by construction of the vectors β_j, β_k by $\sqrt{\alpha \log(p/s^2)/n}$. Thus, following exactly the same steps as in the proof of Theorem 5.1, we can derive that in analogy with (21),

$$\sum_{j=1}^{M} \big(E_{P_{ji}}[\widehat{\beta}_i] - E_{P_0}[\widehat{\beta}_i]\big)^2 \leq M \frac{e}{1 - (1/2)^{0.01}} \Big(\frac{p}{s^2}\Big)^{\alpha} \mathbb{V}\mathrm{ar}_0\big(\widehat{\beta}_i\big).$$

The remainder of the proof is also nearly the same as the one for Theorem 5.1. The only real difference is that $\|\beta_j\|_2^2$ and the upper bound on the bias are smaller by a factor $1/n$, which consequently also occurs in the lower bound on the variance. $\square$

6 Lower bounds based on reduction

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. Firstly, 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 tries to constraint 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 (14). Recall that the mean integrated squared error (MISE) can be decomposed as

$$\begin{aligned}\mathrm{MISE}_f\,\big(\widehat{f}\big) := E_f\big[\|\widehat{f}-f\|_{L^2[0,1]}^2\big] &= \int_0^1 \mathrm{Bias}_f^2\,\big(\widehat{f}(x)\big)\,dx + \int_0^1 \mathbb{V}\mathrm{ar}_f\big(\widehat{f}(x)\big)\,dx \\ &=: \mathrm{IBias}_f^2(\widehat{f}) + \mathrm{IVar}_f\,\big(\widehat{f}\big).\end{aligned} \tag{24}$$

To establish minimax lower bounds for the MISE is substantially more difficult than for pointwise loss as it requires a multiple testing approach together with a careful selection of the hypotheses (see Section 2.6.1 in [41]). We identified this also as a particularly hard problem to prove bias-variance lower bounds. In particular, we cannot obtain a lower bound on IBias^2 and IVar by integrating 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 $\|f\|_{S^\beta([0,1])} \le R$, where for a general domain D, $\|f\|_{S^\beta(D)}^2 := \|f\|_{L^2(D)}^2 + \|f^{(\beta)}\|_{L^2(D)}^2$. Define

$$\Gamma_\beta := \inf\Big\{\|K\|_{S^\beta} : \|K\|_{L^2(\mathbb{R})} = 1, \operatorname{supp} K \subset [-1/2, 1/2]\Big\}. \tag{25}$$

Theorem 6.1. *Consider the Gaussian white noise model* (14) *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,*

$$\inf_{\widehat{f}\in T}\ \sup_{f\in S^\beta(R)} \big|\,\mathrm{IBias}_f(\widehat{f})\big|^{1/\beta} \sup_{f\in S^\beta(R)} \mathrm{IVar}_f\,\big(\widehat{f}\big) \ge \frac{1}{8n}, \tag{26}$$

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 (24), for every estimator $\widehat{f} \in T$ the following lower bound on the MISE holds

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

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. Moreover, this rate is attained if the worst-case integrated squared bias and the worst-case integrated variance are balanced to be of the same order. If applied to functions, recall that $\|\cdot\|_p$ denotes the $L^p([0,1])$-norm. Since $\|\cdot\|_2^2 \leq \|\cdot\|_p^p$, another direct consequence of the previous theorem is

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

for any $p \geq 2$ and any estimator with $\sup_{f\in S^\beta(R)} \|E[\widehat{f}] - f\|_{L_p} \leq 1$.

We now sketch the main reduction steps in the proof of Theorem 6.1. The first step is a model reduction to a Gaussian sequence model

$$X_i = \theta_i + \frac{1}{\sqrt{n}}\varepsilon_i, \quad i = 1, \dots, m \tag{27}$$

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

$$E_\theta\big[\|\widehat{\theta} - \theta\|_2^2\big] = \big\|E_\theta\big[\widehat{\theta}\big] - \theta\big\|_2^2 + \sum_{i=1}^m \mathbb{V}\mathrm{ar}\big(\widehat{\theta}_i\big) =: \mathrm{IBias}_\theta^2\left(\widehat{\theta}\right) + \mathrm{IVar}_\theta\left(\widehat{\theta}\right),$$

recalling that $\|\cdot\|_2$ denotes the Euclidean norm if applied to vectors. As this leads to more appealing formulas below, we have chosen to define also in the sequence model the bias and variance term by IBias and IVar, respectively.

Proposition 6.2. *Let m be a positive integer. Then, for any estimator $\widehat{f}$ of the regression function f in the Gaussian white noise model* (14) *with parameter space $S^\beta(R)$, there exists a non-randomized 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*

$$\sup_{\theta\in\Theta_m^\beta(R)} \mathrm{IBias}_\theta^2(\widehat{\theta}) \leq \sup_{f\in S^\beta(R)} \mathrm{IBias}_f^2(\widehat{f}), \quad \textit{and} \quad \sup_{\theta\in\Theta_m^\beta(R)} \mathrm{IVar}_\theta\left(\widehat{\theta}\right) \leq \sup_{f\in S^\beta(R)} \mathrm{IVar}_f\left(\widehat{f}\right).$$

A proof is given in Supplement E. 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 6.2 reduces the original problem to deriving lower bounds on the bias-variance trade-off in the sequence model (27) 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 could suggest that the reduction to the Gaussian sequence model would destroy the original bias-variance trade-off and therefore would not be suitable for deriving lower bounds. This is, however, not true as the bias will be induced through the choice of m. Indeed, to prove Theorem 6.1, 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 (26).

To obtain lower bounds 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 $X=(X_1,\ldots,X_m)^\top$ for $m>2$. While its risk $E_\theta[\|\widehat{\theta}-\theta\|_2^2] = \mathrm{IBias}_\theta^2(\widehat{\theta})+\mathrm{IVar}_\theta(\widehat{\theta})$ is upper bounded by m/n for all $\theta\in\mathbb{R}^m$, the risk for the zero vector $\theta=(0,\ldots,0)^\top$ is bounded by the potentially much smaller value $2/n$ (see Proposition 2.8 in [24]). Thus, for the zero parameter vector both $\mathrm{IBias}_\theta^2(\widehat{\theta})$ and $\mathrm{IVar}_\theta(\widehat{\theta})$ are simultaneously small. Furthermore, for any parameter vector θ^* there exists an estimator $\widehat{\theta}$ with small bias and variance at θ^*. This suggests that fixing a parameter 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. 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 $D\in\mathcal{O}_m$, $f(x)=D^{-1}f(Dx)$. We say that an estimator $\widehat{\theta}=\widehat{\theta}(X)$ is spherically symmetric if $X\mapsto\widehat{\theta}(X)$ is spherically symmetric. In the seminal work by Stein [38], it has been shown that any minimax estimator in the sequence model (27) with $\Theta=\mathbb{R}^m$ has to be spherically symmetric. By extending this argument we show that this is also true for the worst-case bias-variance trade-off.

Proposition 6.3. *Consider the sequence model* (27) *with parameter space* $\Theta_m^\beta(R)$. *For any estimator* $\widehat{\theta}$ *there exists a spherically symmetric estimator* $\widetilde{\theta}$ *such that*

$$\sup_{\theta\in\Theta_m^\beta(R)}\mathrm{IBias}_\theta^2(\widetilde{\theta})\le\sup_{\theta\in\Theta_m^\beta(R)}\mathrm{IBias}_\theta^2(\widehat{\theta})\quad\text{and}\quad\sup_{\theta\in\Theta_m^\beta(R)}\mathrm{IVar}_\theta\big(\widetilde{\theta}\big)\le\sup_{\theta\in\Theta_m^\beta(R)}\mathrm{IVar}_\theta\big(\widehat{\theta}\big).$$

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 [38] that any spherically symmetric function h is of the form

$$h(x)=r(\|x\|_2)x, \tag{28}$$

for some real-valued function r. In Lemma E.1 in the supplement, 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

$$\begin{aligned}E_{D\theta}\big[\widetilde{\theta}(X)\big]&=\int s(\|x\|_2)x\phi(\|x-D\theta\|_2)\,dx\\&=\int s(\|D^{-1}x\|_2)x\phi(\|D^{-1}x-\theta\|_2)\,dx\\&=\int s(\|y\|_2)Dy\phi(\|y-\theta\|_2)\,dy\\&=DE_\theta\big[\widetilde{\theta}(X)\big].\end{aligned}\tag{29}$$

We can now prove Theorem 6.1.

Proof of Theorem 6.1. Fix an estimator $\widehat{f}$ in the Gaussian white noise model (14) and set $B := \sup_{\theta \in S^\beta(R)} \mathrm{IBias}^2_\theta(\widetilde{f})$. Consider first the case that $B > 0$. Choose $m_* := \lfloor B^{-1/\beta} \rfloor$ and observe that since $B < 2^{-\beta}$, we must have $m_* \geq 2$. Also $2m_* \geq m_* + 1 \geq B^{-1/\beta}$ and so $m_* \geq B^{-1/\beta}/2$. Applying Proposition 6.2 and Proposition 6.3, there exists a spherically symmetric estimator $\widetilde{\theta}$ in the Gaussian sequence model with $m = m_*$ satisfying

$$\sup_{\theta \in \Theta^\beta_{m_*}(R)} \mathrm{IBias}^2_\theta(\widetilde{\theta}) \leq B \quad \text{and} \quad \sup_{\theta \in \Theta^\beta_{m_*}(R)} \mathrm{IVar}_\theta(\widetilde{\theta}) \leq \sup_{f \in S^\beta(R)} \mathrm{IVar}_f\big(\widehat{f}\big).$$

Below we will construct a θ_0 for which

$$\mathrm{IVar}_{\theta_0}\big(\widetilde{\theta}\big) B^{1/\beta} \geq \frac{1}{8n}. \tag{30}$$

This proves then the result.

By (29), we know that $\theta \mapsto g(\theta) := E_\theta[\widetilde{\theta}(X)]$ is a spherically symmetric function. Using Lemma E.1 we can write $g(\theta) = t(\|\theta\|_2)\theta$ for some real value function t. Since by assumption $R \geq 2\Gamma_\beta$, for any θ with $\|\theta\|_2 = R/(\Gamma_\beta m_*^\beta)$, we have that $\|\theta\|_2 \geq 2B$ and consequently,

$$\mathrm{IBias}^2_\theta\big(\widetilde{\theta}\big) = \|g(\theta) - \theta\|_2^2 = \|\theta\|_2^2 \big(t(\|\theta\|_2) - 1\big)^2 \geq 4B^2 \big(t(\|\theta\|_2) - 1\big)^2.$$

As B is an upper bound for the bias, $|t(\|\theta\|_2) - 1| \leq 1/2$ and thus $t(\|\theta\|_2) \geq 1/2$.

Let $0 < \Delta \leq 1/2$ and consider $\theta_0 := R/(\Gamma_\beta m_*^{\beta+1/2})(1, \ldots, 1)^\top$ and $\theta_i = (\theta_{ij})^\top_{j=1,\ldots,m_*}$ with $\theta_{ii} := \sqrt{1+\Delta} R/(\Gamma_\beta m_*^{\beta+1/2})$ and $\theta_{ij} := \sqrt{1 - \Delta/(m_* - 1)} R/(\Gamma_\beta m_*^{\beta+1/2})$ for $j \neq i$. By construction $\|\theta_i\|_2 = R/(\Gamma_\beta m_*^\beta)$ and in particular $\theta_i \in \Theta^\beta_{m_*}(R)$ for all $i = 0, 1, \ldots, m_*$. Using that $\sqrt{1+u} - 1 = u/2 + O(u^2)$ for $u \to 0$, we have for $i = 1, \ldots, m_*$ and $\Delta \to 0$,

$$\begin{aligned} \Delta^{-2} \|\theta_i - \theta_0\|_2^2 &= \frac{R^2}{\Delta^2 \Gamma_\beta^2 m_*^{2\beta+1}} \left[(m_* - 1)\Big(\sqrt{1 - \frac{\Delta}{m_* - 1}} - 1\Big)^2 + \Big(\sqrt{1+\Delta} - 1\Big)^2 \right] \\ &= \frac{R^2}{4\Gamma_\beta^2 m_*^{2\beta+1}} \Big(1 + \frac{1}{m_* - 1}\Big) + O(\Delta). \end{aligned}$$

Similarly, for $i, j = 1, \ldots, m_*$, $i \neq j$ and $\Delta \to 0$,

$$\begin{aligned} \frac{1}{\Delta^2} \langle \theta_i - \theta_0, \theta_j - \theta_0 \rangle &= \frac{R^2}{\Delta^2 \Gamma_\beta^2 m_*^{2\beta+1}} \left[(m_* - 2)\Big(\sqrt{1 - \frac{\Delta}{m_* - 1}} - 1\Big)^2 + 2\Big(\sqrt{1+\Delta} - 1\Big)\Big(\sqrt{1 - \frac{\Delta}{m_* - 1}} - 1\Big) \right] \\ &= -\frac{R^2}{4\Gamma_\beta^2 m_*^{2\beta+1}(m_* - 1)} \Big(1 + \frac{1}{m_* - 1}\Big) + O(\Delta). \end{aligned}$$

Recall that $\|\theta_i\|_2 = \|\theta_j\|_2$ by construction. Applying (12) and (32) yields

$$\frac{1}{2} \sum_{i=1}^{m_*} \|\theta_i - \theta_0\|_2^2 \leq \sum_{i=1}^{m_*} t(\|\theta_i\|_2) \|\theta_i - \theta_0\|_2^2 = \sum_{i=1}^{m_*} \big(E_{\theta_i}[\widetilde{\theta}] - E_{\theta_0}[\widetilde{\theta}]\big)^2 \leq \max_i \sum_{j=1}^{m_*} |e^{n\langle \theta_i - \theta_0, \theta_j - \theta_0 \rangle} - 1| \, \mathrm{IVar}_{\theta_0}\big(\widetilde{\theta}\big).$$

Multiplying both sides of the inequality with Δ^{-2}, using the expressions for $\Delta^{-2}\|\theta_i - \theta_0\|_2^2$ and $\Delta^{-2}\langle \theta_i - \theta_0, \theta_j - \theta_0 \rangle$, and letting Δ tend to zero yields $\mathrm{IVar}_{\theta_0}(\widetilde{\theta}) \geq m_*/(4n)$. As remarked above $m_* \geq B^{-1/\beta}/2$ and this shows finally (30) proving the theorem for $B > 0$.

If $B = 0$ we consider the estimator $\widehat{f}_\delta := \widehat{f} + \delta$ for an arbitrary deterministic $\delta > 0$ that is sufficiently small such that $\widehat{f}_\delta \in T$. Observe that $\mathrm{IVar}_f(\widehat{f}_\delta) = \mathrm{IVar}_f(\widehat{f})$. We can now apply the result from the first part and let δ tend to zero to verify that $\mathrm{IVar}_f(\widehat{f})$ must be unbounded in this case. The result follows since $0 \cdot (+\infty)$ is interpreted as $+\infty$. $\square$

The proof strategy carries over to the nonparametric regression model with fixed uniform design on $[0, 1]$. The discretization effects result in a slight heteroscedasticity of the noise in the Gaussian sequence model which make the computations considerably more technical. It is unclear to which extent a similar approach could be used for lower bounds on the bias-variance trade-off for nonparametric density estimation.

7 Lower bounds for the trade-off between bias and mean deviation

So far, we have studied the bias-variance trade-off for a range of statistical models. One might wonder whether similar results can be obtained for other concepts to measure systematic and stochastic error of an estimator. This section is intended as an overview of related concepts. In a second part, we derive an inequality allowing to link mean deviation error to the bias and apply this to pointwise estimation in the Gaussian white noise model (14)

A large chunk of literature on variations of the bias-variance trade-off is concerned with extensions to classification under 0-1 loss [25, 5, 40, 23]. These approaches have been compared in [37]. [26] proposes an extension to the multi-class setting. In a Bayesian framework, [43] argues that the bias-variance trade-off becomes a bias-covariance-covariance trade-off, where a covariance correction is added. For relational domains, [30] propose to separate the bias and the variance due to the learning process from the bias and the variance due to the inference process. Bias-variance decompositions for the Kullback-Leibler divergence and for the log-likelihood are studied in [20]. Somehow related, [44] introduces the Kullback-Leibler bias and the Kullback-Leibler variance, and shows, using information theory, that a similar decomposition is valid. [12] propose generalized definitions of bias and variance for a general loss, but without showing a bias-variance decomposition. For several exponential families [19] shows that there exist a loss L such that a bias-variance decomposition of L is possible. [22] studied a bias-variance decomposition for arbitrary loss functions, comparing different ways of defining the bias and the variance in such cases.

To measure the stochastic error of an estimator, a competitor of the variance is the mean absolute deviation (MAD). For a random variable X, the MAD is defined as $E[|X - u|]$, where u is either the mean or the median of X. If centered at the mean, the MAD is upper bounded by $\sqrt{\mathbb{V}\mathrm{ar}(X)}$, but compared to the variance, less weight is given to large values of X. For $(P_\theta : \theta \in \Theta)$, the most natural extension seems therefore to study the trade-off between $m(\theta) - \theta$ and $E_\theta[|\widehat{\theta} - m(\theta)|]$, where again $m(\theta)$ is either the mean or the median of the estimator $\widehat{\theta}$ under P_θ. The next result provides an analogue of (2) that applies to both versions.

Lemma 7.1. *Let P, Q be two probability distributions on the same measurable space and write E_P, E_Q for the expectations with respect to P and Q. Then for any random variable X and any real numbers u, v, we have*

$$\frac{1}{5}\big(1-H^2(P,Q)\big)^2|u-v| \le E_P\big[|X-u|\big] \vee E_Q\big[|X-v|\big], \tag{31}$$

Notice that the inequality does not directly follow from the triangle inequality $|u-v| \le |x-u| + |x-v|$ as the expectations on the right-hand side of (51) are taken with respect to different measures P and Q. Equality up to a constant multiple is attained if $H(P,Q) < 1$ and $X = v$ a.e.

As mentioned above, $E_P[|X - E_P[X]|] \le \sqrt{\mathbb{V}\mathrm{ar}_P(X)}$. Moreover, $E_P[|X - E_P[X]|]$ and $\sqrt{\mathbb{V}\mathrm{ar}P(X)}$ are typically of the same order. It is thus instructive to compare the lower bound for the mean absolute deviation centered at $u = E_P[X]$, $v = E_Q[X]$ with the in (2) derived Hellinger lower bound for the variance, that is,

$$\frac{1}{2}(1-H^2(P,Q))\frac{|E_P[X]-E_Q[X]|}{H(P,Q)} \le \sqrt{\mathbb{V}\mathrm{ar}_P(X) + \mathbb{V}\mathrm{ar}_Q(X)}.$$

The variance lower bound also includes a term $H(P,Q)^{-1}$ on the left hand side that improves the inequality if the distributions P and Q are close. The comparison suggests that the factor $H(P,Q)^{-1}$ should also appear in the lower bound for the absolute mean deviation. But if P now tends to Q and u and v are fixed, the lower bound would then tend to infinity. This is impossible and therefore a lower bound of the form $\gtrsim |u-v|/H(P,Q)$ can only hold for special choices of u and v such as $u = E_P[X]$, $v = E_Q[X]$.

We now apply this inequality to pointwise estimation in the Gaussian white noise model, see Section 3 for notation and definitions. Concerning upper bounds in this setting, the first order asymptotics of the mean absolute deviation risk for Lipschitz functions was derived in [17]. Recall that $\mathscr{C}^\beta(R)$ denotes the Hölder ball with smoothness index β and radius R.

Theorem 7.2. *Consider the Gaussian white noise model* (14) *with parameter space $\mathscr{C}^\beta(R)$. Let $C > 0$ be a positive constant. If $\widehat{f}(x_0)$ is an estimator for $f(x_0)$ satisfying*

$$\sup_{f\in\mathscr{C}^\beta(R)} \big|\operatorname{Bias}_f(\widehat{f}(x_0))\big| < \Big(\frac{C}{n}\Big)^{\beta/(2\beta+1)},$$

then, there exist positive constants $c = c(C,R)$ and $N = N(C,R)$, such that

$$\sup_{f\in\mathscr{C}^\beta(R)} E_f\big[\big|\widehat{f}(x_0) - E_f[\widehat{f}(x_0)]\big|\big] \ge cn^{-\beta/(2\beta+1)}, \quad \textit{for all } n \ge N.$$

The same holds if $\operatorname{Bias}_f(\widehat{f}(x_0))$ *and* $E_f[|\widehat{f}(x_0) - E_f[\widehat{f}(x_0)]|]$ *are replaced by* $\operatorname{Med}_f[\widehat{f}(x_0)] - f(x_0)$ *and* $E_f[|\widehat{f}(x_0) - \operatorname{Med}_f[\widehat{f}(x_0)]|]$*, respectively.*

The result is considerably weaker than the lower bounds for the bias-variance trade-off for pointwise estimation. This is due to the fact that (31) is less sharp. Nevertheless, the conclusion provides still more information than the minimax lower bound. To see this, observe that by the triangle inequality, $E_f[|\widehat{f}(x_0) - E_f[\widehat{f}(x_0)]|] \ge E_f[|\widehat{f}(x_0) - f(x_0)|] - |\operatorname{Bias}_f(\widehat{f}(x_0))|$. Thus, the conclusion of Theorem 7.2 follows from the minimax lower bound $\sup_{f\in\mathscr{C}^\beta(R)} E_f[|\widehat{f}(x_0)-f(x_0)|] \geq (K/n)^{\beta/(2\beta+1)}$ as long as $C<K$. Arguing via the minimax rate, nothing, however, can be said if $C>K$. This is still an interesting case, where the bias is of the optimal order with a potentially large constant. Theorem 7.2 shows that even in this case, the worst case variance cannot converge faster than $n^{-\beta/(2\beta+1)}$. As we believe that more refined versions of Lemma 7.1 are obtainable, this approach has, moreover, the potential to lead to a complete characterization of the interplay between bias and mean absolute deviation.

Acknowledgements

We are grateful to Ming Yuan for helpful discussions during an early stage of the project. The project has received funding from the Dutch Science Foundation (NWO) via a Vidi grant.

A Explicit expressions for the χ^2-divergence and the Hellinger affinity matrix

In this section we provide proofs for the explicit formulas of the χ^2-divergence and the Hellinger affinity matrix in Table 1. We also derive a closed-form formula for the case of Gamma distributions and discuss a first order approximation of it.

Multivariate normal distribution: Suppose $P_j=\mathcal{N}(\theta_j,\sigma^2 I_d)$ for $j=0,\ldots,M$. Here $\theta_j=(\theta_{j1},\ldots,\theta_{jd})^\top$ are vectors in $\mathbb{R}^d$ and I_d denotes the $d\times d$ identity matrix. Then,

$$\chi^2(P_0,\ldots,P_M)_{j,k}=\exp\Big(\frac{\langle\theta_j-\theta_0,\theta_k-\theta_0\rangle}{\sigma^2}\Big)-1. \tag{32}$$

and

$$\rho(P_0|P_1,\ldots,P_M)_{j,k}=\exp\Big(\frac{\langle\theta_j-\theta_0,\theta_k-\theta_0\rangle}{4\sigma^2}\Big)-1. \tag{33}$$

Proof. To verify (32), write

$$\int\frac{dP_j}{dP_0}\,dP_k=\frac{1}{(2\pi\sigma^2)^{d/2}}\int_{\mathbb{R}^d}\exp\Big(-\frac{1}{2\sigma^2}\|x-\theta_j\|_2^2-\frac{1}{2\sigma^2}\|x-\theta_k\|_2^2+\frac{1}{2\sigma^2}\|x-\theta_0\|_2^2\Big)\,dx.$$

Substituting $y=x-\theta_0$ shows that it is enough to prove that for θ_0 the zero vector, $\int\frac{dP_j}{dP_0}\,dP_k=\exp(\frac{\langle\theta_j,\theta_k\rangle}{\sigma^2})$. We have that $-\|x-\theta_j\|_2^2-\|x-\theta_k\|_2^2+\|x\|_2^2=-\|x-(\theta_j+\theta_k)\|_2^2+\|\theta_j+\theta_k\|_2^2-\|\theta_j\|_2^2-\|\theta_k\|_2^2$. Identifying the first term as the p.d.f. of a normal distribution with mean $\theta_j+\theta_k$, we can evaluate the integral to obtain

$$\int\frac{dP_j}{dP_0}\,dP_k=\exp\Big(\frac{\|\theta_j+\theta_k\|_2^2-\|\theta_j\|_2^2-\|\theta_k\|_2^2}{2\sigma^2}\Big)=\exp\Big(\frac{\langle\theta_j,\theta_k\rangle}{\sigma^2}\Big).$$

To check (33), we use that if p and q are densities of the $\mathcal{N}(\mu,\sigma^2 I)$ and $\mathcal{N}(\mu',\sigma^2 I)$ distribution, respectively, applying the parallelogram identity yields,

$$\int\sqrt{p(x)q(x)}\,dx=\frac{1}{(2\pi\sigma^2)^{d/2}}\int\exp\Big(-\frac{1}{4\sigma^2}\|x-\mu\|_2^2-\frac{1}{4\sigma^2}\|x-\mu'\|_2^2\Big)\,dx$$

$$
\begin{aligned}
&= \frac{1}{(2\pi\sigma^2)^{d/2}} \int \exp\Big(-\frac{1}{2\sigma^2}\Big\|x - \frac{\mu+\mu'}{2}\Big\|_2^2 - \frac{1}{8\sigma^2}\|\mu-\mu'\|_2^2\Big)\, dx \\
&= \exp\Big(-\frac{1}{8\sigma^2}\|\mu-\mu'\|_2^2\Big).
\end{aligned}
$$

Rewriting $\theta_j - \theta_k = (\theta_j - \theta_0) - (\theta_k - \theta_0)$, this shows that

$$
\begin{aligned}
\frac{\int \sqrt{p_j p_k}\, d\nu}{\int \sqrt{p_j p_0}\, d\nu \int \sqrt{p_k p_0}\, d\nu} - 1 &= \exp\Bigg(-\frac{1}{8\sigma^2}\|\theta_j - \theta_k\|_2^2 + \frac{1}{8\sigma^2}\|\theta_k - \theta_0\|_2^2 + \frac{1}{8\sigma^2}\|\theta_j - \theta_0\|^2\Bigg) - 1 \\
&= \exp\Bigg(\frac{\langle \theta_j - \theta_0, \theta_k - \theta_0\rangle}{4\sigma^2}\Bigg) - 1.
\end{aligned}
$$

□

Poisson distribution: Suppose $P_j = \otimes_{\ell=1}^d \operatorname{Pois}(\lambda_{j\ell})$ for $j = 0, \ldots, M$ and $\lambda_{j\ell} > 0$ for all j, ℓ. Here $\operatorname{Pois}(\lambda)$ denotes the Poisson distribution with intensity $\lambda > 0$. Then,

$$
\chi^2(P_0, \ldots, P_M)_{j,k} = \exp\Big(\sum_{\ell=1}^d \frac{(\lambda_{j\ell} - \lambda_{0\ell})(\lambda_{k\ell} - \lambda_{0\ell})}{\lambda_{0\ell}}\Big) - 1 \tag{34}
$$

and

$$
\rho(P_0|P_1, \ldots, P_M)_{j,k} = \exp\Big(\sum_{\ell=1}^d \big(\sqrt{\lambda_{j\ell}} - \sqrt{\lambda_{0\ell}}\big)\big(\sqrt{\lambda_{k\ell}} - \sqrt{\lambda_{0\ell}}\big)\Big) - 1. \tag{35}
$$

Proof. To verify (34) assume that for $i = 1, 2, 3$, p_i is the p.m.f. of a Poisson distributed random variable with intensity $\lambda_i > 0$. Then,

$$
\sum_{k=0}^\infty \frac{p_1(k)p_2(k)}{p_0(k)} = e^{-\lambda_1 - \lambda_2 + \lambda_0} \sum_{k=0}^\infty \frac{(\lambda_1\lambda_2/\lambda_0)^k}{k!} = e^{-\lambda_1-\lambda_2+\lambda_0+\lambda_1\lambda_2/\lambda_0} = \exp\Big(\frac{(\lambda_1 - \lambda_0)(\lambda_2-\lambda_0)}{\lambda_0}\Big).
$$

Taking product measures, (34) follows. For (35), the Hellinger affinity of two Poisson distributed random variables is given by

$$
\sum_{k=0}^\infty \sqrt{p_1(k)p_2(k)} = \exp\Big(-\frac{\lambda_1+\lambda_2}{2}\Big)\sum_{k=0}^\infty \frac{(\sqrt{\lambda_1\lambda_2})^k}{k!} = \exp\Big(-\frac{\lambda_1+\lambda_2}{2} + \sqrt{\lambda_1\lambda_2}\Big) = \exp\Big(-\frac{1}{2}\big(\sqrt{\lambda_1} - \sqrt{\lambda_2}\big)^2\Big).
$$

The proof of (35) can be completed by arguing as for (33). □

Bernoulli distribution: Suppose $P_j = \otimes_{\ell=1}^d \operatorname{Ber}(\theta_{j\ell})$ for $j = 0, \ldots, M$ and $\theta_{j\ell} \in (0,1)$ for all j, ℓ. Here $\operatorname{Ber}(\theta)$ denotes the Bernoulli distribution with parameter $\theta \in (0,1)$. Then,

$$
\chi^2(P_0, \ldots, P_M)_{j,k} = \prod_{\ell=1}^d \Big(\frac{(\theta_{j\ell} - \theta_{0\ell})(\theta_{k\ell} - \theta_{0\ell})}{\theta_{0\ell}(1-\theta_{0\ell})} + 1\Big) - 1, \tag{36}
$$

and

$$
\rho(P_0|P_1, \ldots, P_M)_{j,k} = \prod_{\ell=1}^d \frac{r(\theta_{j\ell}, \theta_{k\ell})}{r(\theta_{j\ell}, \theta_{0\ell})r(\theta_{k\ell}, \theta_{0\ell})} - 1, \tag{37}
$$

with $r(\theta, \theta') := \sqrt{\theta\theta'} + \sqrt{(1-\theta)(1-\theta')}$.

Proof. To check (36), note that

$$\int \frac{dP_j}{dP_0}\, dP_k = \prod_{\ell=1}^{d} \left(\frac{\theta_{j\ell}\theta_{k\ell}}{\theta_{0\ell}} + \frac{(1-\theta_{j\ell})(1-\theta_{k\ell})}{1-\theta_{0\ell}} \right) = \prod_{\ell=1}^{d} \left(\frac{(\theta_{j\ell}-\theta_{0\ell})(\theta_{k\ell}-\theta_{0\ell})}{\theta_{0\ell}(1-\theta_{0\ell})} + 1 \right),$$

where the last step is a purely algebraic manipulation. To prove (37), note that when P and Q are two Bernoulli distributions with parameters θ and θ', we have $\int \sqrt{p(x)q(x)}\, d\nu(x) = \sqrt{\theta\theta'} + \sqrt{(1-\theta)(1-\theta')} = r(\theta, \theta')$. □

Gamma distribution: Suppose $P_j = \otimes_{\ell=1}^{d} \Gamma(\alpha_{j\ell}, \beta_{j\ell})$, where $\Gamma(\alpha, \beta)$ denotes the Gamma distribution with shape $\alpha > 0$ and inverse scale $\beta > 0$. The entries of the χ^2-divergence matrix are all finite if and only if $\alpha_{0\ell} \leq 2\alpha_{j\ell}$ and $\beta_{0\ell} \leq 2\beta_{j\ell}$ for all $j, \ell = 1, \ldots, M$. If the entries of the χ^2-divergence matrix are finite, they are given by the formula

$$\chi^2(P_0, \ldots, P_M)_{j,k} = \prod_{\ell=1}^{d} \frac{\Gamma(\alpha_{0\ell})\Gamma(\alpha_{j\ell}+\alpha_{k\ell}-\alpha_{0\ell})}{\Gamma(\alpha_{j\ell})\Gamma(\alpha_{k\ell})} \frac{\beta_{j\ell}^{\alpha_{j\ell}} \beta_{k\ell}^{\alpha_{k\ell}}}{\beta_{0\ell}^{\alpha_{0\ell}} \big(\beta_{j\ell}+\beta_{k\ell}-\beta_{0\ell}\big)^{(\alpha_{j\ell}+\alpha_{k\ell}-\alpha_{0\ell})}} - 1. \tag{38}$$

The entries of the Hellinger affinity matrix are

$$\rho(P_0|P_1, \ldots, P_M)_{j,k} = \prod_{\ell=1}^{d} \frac{\Gamma(\alpha_{0\ell})\Gamma(\alpha_{j\ell}/2+\alpha_{k\ell}/2)}{\Gamma(\alpha_{j\ell}/2+\alpha_{0\ell}/2)\Gamma(\alpha_{k\ell}/2+\alpha_{0\ell}/2)} \frac{(\beta_{j\ell}+\beta_{0\ell})^{\alpha_{j\ell}/2+\alpha_{0\ell}/2}(\beta_{k\ell}+\beta_{0\ell})^{\alpha_{k\ell}/2+\alpha_{0\ell}/2}}{2^{\alpha_{0\ell}}\beta_{0\ell}^{\alpha_{0\ell}}(\beta_{j\ell}+\beta_{k\ell})^{\alpha_{j\ell}/2+\alpha_{k\ell}/2}} - 1. \tag{39}$$

Proof. For Equation (38) and if the integrals are finite,

$$\begin{aligned}
\int \frac{dP_j}{dP_0}\, dP_k &= \int_{\mathbb{R}^d} \prod_{\ell=1}^{d} \beta_{j\ell}^{\alpha_{j\ell}} x_\ell^{\alpha_{j\ell}-1} e^{-\beta_{j\ell}x_\ell} \Gamma(\alpha_{j\ell})^{-1} \beta_{k\ell}^{\alpha_{k\ell}} x_\ell^{\alpha_{k\ell}-1} e^{-\beta_{k\ell}x_\ell} \Gamma(\alpha_{k\ell})^{-1} \beta_{0\ell}^{-\alpha_{0\ell}} x_\ell^{-\alpha_{0\ell}+1} e^{\beta_{0\ell}x_\ell} \Gamma(\alpha_{0\ell})\, dx \\
&= \prod_{\ell=1}^{d} \frac{\Gamma(\alpha_{0\ell})\beta_{j\ell}^{\alpha_{j\ell}}\beta_{k\ell}^{\alpha_{k\ell}}}{\Gamma(\alpha_{j\ell})\Gamma(\alpha_{k\ell})\beta_{0\ell}^{\alpha_{0\ell}}} \int_{\mathbb{R}} x_\ell^{\alpha_{j\ell}+\alpha_{k\ell}-\alpha_{0\ell}-1} e^{(\beta_{0\ell}-\beta_{j\ell}-\beta_{k\ell})x_\ell}\, dx_\ell \\
&= \prod_{\ell=1}^{d} \frac{\Gamma(\alpha_{0\ell})\beta_{j\ell}^{\alpha_{j\ell}}\beta_{k\ell}^{\alpha_{k\ell}}}{\Gamma(\alpha_{j\ell})\Gamma(\alpha_{k\ell})\beta_{0\ell}^{\alpha_{0\ell}}} \frac{\Gamma(\alpha_{j\ell}+\alpha_{k\ell}-\alpha_{0\ell})}{\big(\beta_{j\ell}+\beta_{k\ell}-\beta_{0\ell}\big)^{(\alpha_{j\ell}+\alpha_{k\ell}-\alpha_{0\ell})}}.
\end{aligned}$$

It is straightforward to see that the integrals are all finite if and only if $\alpha_{0\ell} \leq 2\alpha_{j\ell}$ and $\beta_{0\ell} \leq 2\beta_{j\ell}$ for all $j, \ell = 1, \ldots, M$. For the closed-form formula of the Hellinger distance, write p and q for the density of a $\Gamma(\alpha, \beta)$ and a $\Gamma(\alpha', \beta')$ distribution and observe that the Hellinger affinity is

$$\int \sqrt{pq} = \sqrt{\frac{\beta^\alpha(\beta')^{\alpha'}}{\Gamma(\alpha)\Gamma(\alpha')}} \int_0^\infty x^{\frac{\alpha+\alpha'}{2}-1} e^{-\frac{1}{2}(\beta+\beta')x}\, dx = \sqrt{\frac{\beta^\alpha(\beta')^{\alpha'}}{\Gamma(\alpha)\Gamma(\alpha')}} \frac{\Gamma((\alpha+\alpha')/2)}{(\beta/2+\beta'/2)^{(\alpha+\alpha')/2}}.$$

Together with the definition of the Hellinger affinity matrix, Equation (39) follows. □

A formula for the exponential distribution can be obtained as a special case setting $\alpha_{j\ell} = 1$ for all j, ℓ. For the families of distributions discussed above, the formulas for the χ^2-divergence and the Hellinger affinity matrix encode an orthogonality relation on the parameter vectors. This is less visible in the expressions for the Gamma distribution but can be made more explicit using the first order approximation that we state next. It shows that even for the Gamma distribution these matrix entries can be written in leading order as a term involving a weighted inner product of $\beta_j - \beta_0$ and $\beta_k - \beta_0$, where β_r denotes the vector $(\beta_{r\ell})_{1\leq\ell\leq d}$.

Lemma A.1. *Suppose $P_j = \otimes_{\ell=1}^d \Gamma(\alpha_\ell, \beta_{j\ell})$. Let $\Delta := \max_j |\beta_{j\ell} - \beta_{0\ell}|/\beta_{0\ell}$ and $A := \sum_{\ell=1}^d \alpha_\ell$. Denote by Σ the $d \times d$ diagonal matrix with entries $\beta_{0\ell}^2/\alpha_\ell$. Then,*

$$\chi^2(P_0, \ldots, P_M)_{j,k} = \exp\Big(-(\beta_j - \beta_0)^\top \Sigma^{-1}(\beta_j - \beta_0) + o(A\Delta^2)\Big) - 1 \tag{40}$$

and

$$\rho(P_0|P_1, \ldots, P_M)_{j,k} = \exp\Big(-\frac{1}{4}(\beta_j - \beta_0)^\top \Sigma^{-1}(\beta_j - \beta_0) + o(A\Delta^2)\Big) - 1. \tag{41}$$

Proof. Using that α_ℓ does not depend on j, identity (38) simplifies and a second order Taylor expansion of the logarithm (since all first order terms cancel) yields

$$\begin{aligned}
\int \frac{dP_j}{dP_0}\, dP_k &= \prod_{\ell=1}^d \Big(\frac{\beta_{j\ell}\beta_{k\ell}}{\beta_{0\ell}\beta_{j\ell} + \beta_{0\ell}\beta_{k\ell} - \beta_{0\ell}^2}\Big)^{\alpha_\ell} = \exp\Big(\sum_{\ell=1}^d \alpha_\ell\big(\log(\beta_{j\ell}\beta_{k\ell}) - \log(\beta_{0\ell}\beta_{j\ell} + \beta_{0\ell}\beta_{k\ell} - \beta_{0\ell}^2)\big)\Big) \\
&= \exp\left(\sum_{\ell=1}^d \alpha_\ell\left(\log\left(1 + \frac{\beta_{j\ell} - \beta_{0\ell}}{\beta_{0\ell}}\right) + \log\left(1 + \frac{\beta_{k\ell} - \beta_{0\ell}}{\beta_{0\ell}}\right) - \log\left(1 + \frac{\beta_{j\ell} - \beta_{0\ell} + \beta_{k\ell} - \beta_{0\ell}}{\beta_{0\ell}}\right)\right)\right) \\
&= \exp\left(\sum_{\ell=1}^d \alpha_\ell\left(-\frac{(\beta_{j\ell} - \beta_{0\ell})^2}{2\beta_{0\ell}^2} - \frac{(\beta_{k\ell} - \beta_{0\ell})^2}{2\beta_{0\ell}^2} + \frac{(\beta_{j\ell} - \beta_{0\ell} + \beta_{k\ell} - \beta_{0\ell})^2}{2\beta_{0\ell}^2} + o(\Delta^2)\right)\right) \\
&= \exp\left(\sum_{\ell=1}^d \frac{\alpha_\ell(\beta_{j\ell} - \beta_{0\ell})(\beta_{k\ell} - \beta_{0\ell})}{\beta_{0\ell}^2} + o\big(A\Delta^2\big)\right).
\end{aligned}$$

This shows (40). For (41), we can argue in a similar way to find

$$\begin{aligned}
&\frac{\int \sqrt{p_j p_k}\, d\nu}{\int \sqrt{p_j p_0}\, d\nu \int \sqrt{p_k p_0}\, d\nu} \\
&= \prod_{\ell=1}^d \frac{(\beta_{j\ell} + \beta_{0\ell})^{\alpha_\ell}(\beta_{k\ell} + \beta_{0\ell})^{\alpha_\ell}}{(2\beta_{0\ell})^{\alpha_{0\ell}}(\beta_{j\ell} + \beta_{k\ell})^{\alpha_\ell}} \\
&= \prod_{\ell=1}^d \exp\left(\alpha_\ell\left(\log\left(1 + \frac{\beta_{j\ell} - \beta_{0\ell}}{2\beta_{0\ell}}\right) + \log\left(1 + \frac{\beta_{k\ell} - \beta_{0\ell}}{2\beta_{0\ell}}\right) - \log\left(1 + \frac{\beta_{j\ell} - \beta_{0\ell} + \beta_{k\ell} - \beta_{0\ell}}{2\beta_{0\ell}}\right)\right)\right).
\end{aligned}$$

This is the same expression as for the χ^2-divergence matrix with an additional factor 1/2 in each of the logarithms. Following the same steps as in the χ^2-divergence case leads to the desired result. □

B Proofs for Section 2

Proof of Lemma 2.1. We first prove (1). Applying the Cauchy-Schwarz inequality, for any real number a,

$$\begin{aligned}
\big|E_P[X] - E_Q[X]\big| &= \Big|\int (X(\omega) - a)(p(\omega) - q(\omega))\, d\nu(\omega)\Big| \\
&\leq \Big(\int (X(\omega) - a)^2 |p(\omega) - q(\omega)|\, d\nu(\omega)\Big)^{1/2} \sqrt{2\,\mathrm{TV}(P, Q)}.
\end{aligned}$$

We can bound $|p(\omega) - q(\omega)| \leq p(\omega) + q(\omega)$ and $E_P[(X - a)^2] = \mathbb{V}\mathrm{ar}_P(X) + (E_P[X] - a)^2$ (which holds for all P and all a), to deduce that for $a_* := (E_P[X] + E_Q[X])/2$,

$$\int (X(\omega) - a_*)^2 |p(\omega) - q(\omega)|\, d\nu(\omega) \leq \mathbb{V}\mathrm{ar}_P(X) + \mathbb{V}\mathrm{ar}_Q(X) + 2\Big(\frac{E_P[X] - E_Q[X]}{2}\Big)^2.$$

This shows that

$$\big(E_P[X]-E_Q[X]\big)^2 \le \Big(\mathbb{V}\mathrm{ar}_P(X)+\mathbb{V}\mathrm{ar}_Q(X)+\frac{(E_P[X]-E_Q[X])^2}{2}\Big)2\,\mathrm{TV}(P,Q).$$

Rearranging the inequality yields (1).

We now prove (2). Using $H^2(P,Q)=1-\int\sqrt{pq}$, triangle inequality and Cauchy-Schwarz twice, we find

$$\begin{aligned}
&\big|E_P[X]-E_Q[X]\big| \\
&=\Big|\int (X(\omega)-E_P[X])\sqrt{p(\omega)}(\sqrt{p(\omega)}-\sqrt{q(\omega)})\,d\nu(\omega)+\int (X(\omega)-E_Q[X])\sqrt{q(\omega)}(\sqrt{p(\omega)}-\sqrt{q(\omega)})\,d\nu(\omega) \\
&\quad+\big(E_P[X]-E_Q[X]\big)H^2(P,Q)\Big| \\
&\le \Big(\mathbb{V}\mathrm{ar}_P(X)^{1/2}+\mathbb{V}\mathrm{ar}_Q(X)^{1/2}\Big)\sqrt{2}H(P,Q)+\big|E_P[X]-E_Q[X]\big|H^2(P,Q).
\end{aligned}$$

Squaring, rearranging the terms and using that for any positive real numbers u,v, $(\sqrt{u}+\sqrt{v})^2\le 2u+2v$ yields (2).

To prove (3), it is enough to consider the case that $K(P,Q)+K(Q,P)<\infty$. This implies in particular that the Radon-Nikodym derivatives dP/dQ and dQ/dP both exist. Set $h(t,\omega):=\exp(t\log p+(1-t)\log q(\omega))$. Observe that $p(\omega)-q(\omega)=\int_0^1\log(p(\omega)/q(\omega))h(t,\omega)\,dt$. Due to the concavity of the logarithm, we also have that $h(t,\omega)\le tp(\omega)+(1-t)q(\omega)$. Choosing again $a_*:=(E_P[X]+E_Q[X])/2$, and using $E_P[(X-a^*)^2]=\mathbb{V}\mathrm{ar}_P(X)+(E_P[X]-E_Q[X])^2/4$ and $E_Q[(X-a^*)^2]=\mathbb{V}\mathrm{ar}_Q(X)+(E_P[X]-E_Q[X])^2/4$, we therefore have that

$$\begin{aligned}
\int\big(X(\omega)-a^*\big)^2h(t,\omega)d\omega &\le t\mathbb{V}\mathrm{ar}_P(X)+(1-t)\mathbb{V}\mathrm{ar}_Q(X)+\frac{(E_P[X]-E_Q[X])^2}{4} \\
&\le \big(\mathbb{V}\mathrm{ar}_P(X)\vee\mathbb{V}\mathrm{ar}_Q(X)\big)+\frac{(E_P[X]-E_Q[X])^2}{4}.
\end{aligned}$$

Also notice that

$$\int\log^2\Big(\frac{p(\omega)}{q(\omega)}\Big)\int_0^1 h(t,\omega)\,dt\,d\omega=\int\log\Big(\frac{p(\omega)}{q(\omega)}\Big)\big(p(\omega)-q(\omega)\big)\,d\omega=\mathrm{KL}(P,Q)+\mathrm{KL}(Q,P).$$

Changing the order of integration and applying the properties of the $h(t,\omega)$ function, the Cauchy-Schwarz inequality, and Jensen's inequality, we find

$$\begin{aligned}
\big|E_P[X]-E_Q[X]\big| &=\Big|\int\big(X(\omega)-a^*\big)\big(p(\omega)-q(\omega)\big)\,d\omega\Big| \\
&=\Big|\int_0^1\Big(\int\big(X(\omega)-a^*\big)\sqrt{h(t,\omega)}\log\Big(\frac{p(\omega)}{q(\omega)}\Big)\sqrt{h(t,\omega)}\,d\omega\,dt\Big| \\
&\le\int_0^1\Big(\int\big(X(\omega)-a^*\big)^2h(t,\omega)\,d\omega\Big)^{1/2}\Big(\int\log^2\Big(\frac{p(\omega)}{q(\omega)}\Big)h(t,\omega)\,d\omega\Big)^{1/2}dt \\
&\le\Big(\big(\mathbb{V}\mathrm{ar}_P(X)\vee\mathbb{V}\mathrm{ar}_Q(X)\big)+\frac{(E_P[X]-E_Q[X])^2}{4}\Big)^{1/2}\Big(\int\log^2\Big(\frac{p(\omega)}{q(\omega)}\Big)\int_0^1h(t,\omega)\,dt\,d\omega\Big)^{1/2} \\
&=\Big(\big(\mathbb{V}\mathrm{ar}_P(X)\vee\mathbb{V}\mathrm{ar}_Q(X)\big)+\frac{(E_P[X]-E_Q[X])^2}{4}\Big)^{1/2}\Big(\mathrm{KL}(P,Q)+\mathrm{KL}(Q,P)\Big)^{1/2}.
\end{aligned}$$

Squaring and rearranging the terms yields (3).

The proof for (4) combines change of measure and the Cauchy-Schwarz inequality via

$$\big|E_P[X]-E_Q[X]\big| = \Big|E_P\Big[\Big(\frac{dQ}{dP}-1\Big)\big(X-E_P[X]\big)\Big]\Big| \le \sqrt{\chi^2(Q,P)\mathbb{V}\mathrm{ar}_P(X)}.$$

Squaring and interchanging the role of P and Q completes the proof. ☐

Proof of Lemma 2.2. Rewriting $E_1[X]-E_0[X]$ as the telescoping sum $\sum_{j=1}^K E_{j/K}[X]-E_{(j-1)/K}[X]$ and taking the limit $K\to\infty$ over a subset converging to the lim inf, we find that

$$(E_1[X]-E_0[X])^2 \le \liminf_{K\to\infty} K^2 \max_j (E_{j/K}[X]-E_{(j-1)/K}[X])^2.$$

Applying (2), (3) and (4) to $(E_{j/K}[X]-E_{(j-1)/K}[X])^2$, bounding $\mathbb{V}\mathrm{ar}_{j/K}(X)$ and $\mathbb{V}\mathrm{ar}_{(j-1)/K}(X)$ always by $\sup_{t\in[0,1]}\mathbb{V}\mathrm{ar}_t(X)$, and taking the limit $K\to\infty$ yields the three inequalities. ☐

Proof of Theorem 2.4. Because of the identity $v^\top\chi^2(Q_0,\ldots,Q_M)v = \int(\sum_{j=1}^M v_j(dQ_j/dQ_0-1))^2 dQ_0$, it is enough to prove that for any arbitrary vector $v=(v_1,\ldots,v_M)^\top$,

$$\int\Big(\sum_{j=1}^M v_j\Big(\frac{dKQ_j}{dKQ_0}-1\Big)\Big)^2 dKQ_0 \le \int\Big(\sum_{j=1}^M v_j\Big(\frac{dQ_j}{dQ_0}-1\Big)\Big)^2 dQ_0. \tag{42}$$

Let ν be a dominating measure for $Q_0,\ldots,Q_M$ and μ a dominating measure for $KQ_0,\ldots,KQ_M$. Write Q_j for the ν-density of Q_j. Then, $dKQ_j(y)=\int_X k(y,x)q_j(x)\,d\nu(x)\,d\mu(y)$ for $j=1,\ldots,M$ and a suitable non-negative kernel function k satisfying $\int k(y,x)\,d\mu(y)=1$ for all x. Applying the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\Big(\sum_{j=1}^M v_j\Big(\frac{dKQ_j}{dKQ_0}(y)-1\Big)\Big)^2 &= \Big(\frac{\int k(y,x)[\sum_{j=1}^M v_j(q_j(x)-q_0(x))]\,d\nu(x)}{\int k(y,x')q_0(x')\,d\nu(x')}\Big)^2\\ &\le \frac{\int k(y,x)\big(\sum_{j=1}^M v_j\frac{(q_j(x)-q_0(x))}{q_0(x)}\big)^2 q_0(x)\,d\nu(x)}{\int k(y,x')q_0(x')\,d\nu(x')}.\end{aligned}$$

Inserting this in (42), rewriting $dKQ_0(y)=\int_X k(y,x)q_0(x)\,d\nu(x)\,d\mu(y)$, interchanging the order of integration using Fubini's theorem, and applying $\int k(y,x)\,d\mu(y)=1$, yields

$$\begin{aligned}\int\Big(\sum_{j=1}^M v_j\Big(\frac{dKQ_j}{dKQ_0}-1\Big)\Big)^2 dKQ_0 &\le \iint k(y,x)\Big(\sum_{j=1}^M v_j\frac{(q_j(x)-q_0(x))}{q_0(x)}\Big)^2 q_0(x)\,d\nu(x)\,d\mu(y)\\ &= \int\Big(\sum_{j=1}^M v_j\Big(\frac{q_j(x)}{q_0(x)}-1\Big)\Big)^2 q_0(x)\,d\nu(x)\\ &= \int\Big(\sum_{j=1}^M v_j\Big(\frac{dQ_j}{dQ_0}-1\Big)\Big)^2 dQ_0.\end{aligned}$$

☐

C Proofs for Section 3

Lemma C.1. *For* $0<h\le 1$,

$$\Big\|h^\beta K\Big(\frac{\cdot - x_0}{h}\Big)\Big\|_{\mathscr{C}^\beta} \le \|K\|_{\mathscr{C}^\beta(\mathbb{R})}.$$

Proof. Set $f(x) := h^\beta K((x-x_0)/h)$. Then,

$$\begin{aligned}\|f\|_{\mathscr{C}^\beta} &= \sum_{\ell \le \lfloor\beta\rfloor} \big\|f^{(\ell)}\big\|_\infty + \sup_{x,y\in[0,1]} \frac{|f^{(\lfloor\beta\rfloor)}(x) - f^{(\lfloor\beta\rfloor)}(y)|}{|x-y|^{\beta-\lfloor\beta\rfloor}} \\ &= \sum_{\ell \le \lfloor\beta\rfloor} h^{\beta-\ell}\big\|K^{(\ell)}\big\|_\infty + \sup_{x,y\in[0,1]} \frac{|K^{(\lfloor\beta\rfloor)}((x-x_0)/h) - K^{(\lfloor\beta\rfloor)}((y-x_0)/h)|}{|(x-y)/h|^{\beta-\lfloor\beta\rfloor}} \\ &\le \|K\|_{\mathscr{C}^\beta(\mathbb{R})}.\end{aligned}$$

□

Proof of Theorem 3.1. (i): Given an estimator $\widehat{f}$, let $B := \sup_{f\in\mathscr{C}^\beta(R)} |\operatorname{Bias}_f(\widehat{f}(x_0))|$. It is sufficient to show that for an arbitrary estimator and any $K \in \mathscr{C}^\beta(\mathbb{R}) \cap L^2(\mathbb{R})$ satisfying $K(0) = 1$,

$$B^{1/\beta} \sup_{f\in\mathscr{C}^\beta(R)} \mathbb{V}\mathrm{ar}_f\big(\widehat{f}(x_0)\big) \ge \frac{1}{n}\|K\|_2^{-2}\bigg(1 - \frac{\|K\|_{\mathscr{C}^\beta(\mathbb{R})}}{R}\bigg)_+^2. \tag{43}$$

Assume first that $B > 0$. For $K \in \mathscr{C}^\beta(\mathbb{R})$ any function satisfying $K(0) = 1$ and $\|K\|_2 < +\infty$, define $V := R/\|K\|_{\mathscr{C}^\beta(R)}$ and

$$\mathscr{F} := \Big\{f_\theta(x) = \theta V B K\Big(\frac{x-x_0}{B^{1/\beta}}\Big) : |\theta| \le 1\Big\}.$$

Using Lemma C.1 and $B < 1$, we have that $\|f_\theta\|_{\mathscr{C}^\beta} \le |\theta| V \|K\|_{\mathscr{C}^\beta(R)} \le R$ for all $\theta \in [-1,1]$. This implies $\mathscr{F} \subseteq \mathscr{C}^\beta(R)$. As explained at the beginning of Section 3, $\mathrm{KL}(P_f, P_g) = \frac{n}{2}\|f-g\|^2_{L^2[0,1]}$. We will apply Lemma 2.2 (ii) to the family of distributions $(P_{f_\theta})_{\theta\in[0,1]}$ and $(P_{f_\theta})_{\theta\in[-1,0]}$. Due to

$$\begin{aligned}\mathrm{KL}\left(P_{f_\theta}, P_{f_{\theta+\delta}}\right) = \mathrm{KL}\left(P_{f_{\theta+\delta}}, P_{f_\theta}\right) &= \frac{n}{2}\big\|f_\theta - f_{\theta+\delta}\big\|^2_{L^2[0,1]} \\ &= \frac{n}{2}\Big\|\delta V B K\Big(\frac{x-x_0}{B^{1/\beta}}\Big)\Big\|^2_{L^2[0,1]} \le \frac{n}{2}\delta^2 V^2 B^{2+1/\beta}\|K\|^2_{L^2(\mathbb{R})},\end{aligned} \tag{44}$$

the constant κ_K^2 in the statement of Lemma 2.2 (ii) is bounded by $nV^2B^{2+1/\beta}\|K\|^2_{L^2(\mathbb{R})}$. Now (7) applied to the random variable $\widehat{f}(x_0)$ gives

$$\big(E_{f_{\pm1}}\big[\widehat{f}(x_0)\big] - E_{f_0}\big[\widehat{f}(x_0)\big]\big)^2 \le nV^2B^{2+1/\beta}\|K\|^2_{L^2(\mathbb{R})} \sup_{|\theta|\le 1} \mathbb{V}\mathrm{ar}_{f_\theta}\big(\widehat{f}(x_0)\big),$$

where $E_{f_{\pm1}}$ stand for either E_{f_1} or $E_{f_{-1}}$. Recall that $K(0) = 1$ and notice that it is enough to prove the result for $V \ge 1$. Therefore, $\operatorname{Bias}_{f_\theta}(\widehat{f}(x_0)) = E_{f_\theta}[\widehat{f}(x_0)] - \theta V B$ as well as $E_{f_1}[\widehat{f}(x_0)] \ge (V-1)B$ and $E_{f_{-1}}[\widehat{f}(x_0)] \le -(V-1)B$. Choosing for the lower bound f_1 if $E_{f_0}[\widehat{f}(x_0)]$ is negative and f_{-1} if $E_{f_0}[\widehat{f}(x_0)]$ is positive, we find

$$(V-1)^2B^2 \le nV^2B^{2+1/\beta}\|K\|^2_{L^2(\mathbb{R})} \sup_{|\theta|\le 1} \mathbb{V}\mathrm{ar}_{f_\theta}\big(\widehat{f}(x_0)\big).$$

Dividing both sides by $nV^2B^2\|K\|^2_{L^2(\mathbb{R})}$ yields (43).

To complete the proof, it remains to consider the case $B = 0$. Let $\widehat{f}$ be an estimator such that $B = \sup_{f\in\mathscr{C}^\beta(R)} |\operatorname{Bias}_f(\widehat{f})| = 0$. Define the estimator $\widehat{f}_\delta := \widehat{f} + \delta$ with $\delta \in (0,1)$. Since δ is deterministic, $\mathbb{V}\mathrm{ar}_f(\widehat{f}_\delta(x_0)) = \mathbb{V}\mathrm{ar}_f(\widehat{f}(x_0))$. Applying the lower bound derived above gives

$$\delta^{1/\beta} \sup_{f\in\mathscr{C}^\beta(R)} \mathbb{V}\mathrm{ar}_f\big(\widehat{f}(x_0)\big) \ge \frac{1}{n}\bigg(\|K\|_2^{-1}\bigg(1 - \frac{\|K\|_{\mathscr{C}^\beta(\mathbb{R})}}{R}\bigg)_+\bigg)^2.$$

For $\delta \to 0$, we obtain $\sup_{f\in\mathscr{C}^\beta(R)} \mathbb{V}\mathrm{ar}_f(\widehat{f}(x_0)) \to \infty$ and the conclusion holds because of $(+\infty)\cdot 0 = +\infty$. This completes the proof for (i).

(ii): We use the same notation as for the proof of (i). It is sufficient to show that for an arbitrary estimator $\widehat{f}$ and any $f \in \mathscr{C}^\beta(R)$,

$$B^{1/\beta}\mathbb{V}\mathrm{ar}_f\big(\widehat{f}(x_0)\big) \geq \frac{\overline{\gamma}(R,\beta,C,f)}{n}. \tag{45}$$

Assume first that $B > 0$. For any function $K \in \mathscr{C}^\beta(\mathbb{R})$ satisfying $K(0) = 1$ and $\|K\|_2 < +\infty$, define $U := (R - \|f\|_{\mathscr{C}^\beta([0,1])})/\|K\|_{\mathscr{C}^\beta(\mathbb{R})}$ and

$$\mathscr{G} := \Big\{ f_\theta(x) = f(x) + \theta U B K\Big(\frac{x-x_0}{B^{1/\beta}}\Big) : |\theta| \leq 1 \Big\}.$$

Combining the fact that the triangle inequality holds for any norm with Lemma C.1 and $|\theta| \leq 1$, we obtain $\|f_\theta\|_{\mathscr{C}^\beta([0,1])} \leq \|f\|_{\mathscr{C}^\beta([0,1])} + U\|K\|_{\mathscr{C}^\beta(\mathbb{R})} \leq R$. Hence $\mathscr{G} \subseteq \mathscr{C}^\beta(R)$. As explained at the beginning of Section 3, the χ^2-divergence in this model is $\chi^2(P_f, P_g) = \exp(n\|f-g\|_{L^2[0,1]}^2) - 1$. By assumption, $B^{2+1/\beta} \leq C/n$. Combining this with the inequality $e^x - 1 \leq xe^x$ and arguing as in (44), we find that

$$\chi^2(P_{f_{\pm 1}}, P_{f_0}) \leq n\|f_{\pm 1} - f\|_{L^2[0,1]}^2 \exp\Big(n\|f_{\pm 1} - f\|_{L^2[0,1]}^2\Big) \leq nU^2B^{2+1/\beta}\|K\|_2^2 \exp\big(CU^2\|K\|_2^2\big).$$

Applying the χ^2-divergence version of Lemma 2.1 to the random variable $\widehat{f}(x_0)$ and using the just derived bound for the χ^2-divergence in the Gaussian white noise model yields

$$\big(E_{f_{\pm 1}}\big[\widehat{f}(x_0)\big] - E_0\big[\widehat{f}(x_0)\big]\big)^2 \leq nU^2B^{2+1/\beta}\|K\|_2^2 \exp\big(CU^2\|K\|_2^2\big)\mathbb{V}\mathrm{ar}_f\big(\widehat{f}(x_0)\big).$$

By arguing as for the proof of (i) with the constant V replaced by U, we obtain

$$(U-1)_+^2 B^2 \leq nU^2B^{2+1/\beta}\|K\|_2^2 \exp\big(CU^2\|K\|_2^2\big)\mathbb{V}\mathrm{ar}_f\big(\widehat{f}(x_0)\big).$$

Rearranging the terms and taking the supremum over all kernels $K \in \mathscr{C}^\beta(\mathbb{R})$ with $K(0) = 1$ yields (45).

The case $B = 0$ can be treated in the same way as for (i) since we can always choose a sufficiently small $\delta > 0$ such that $\widehat{f}_\delta = \widehat{f} + \delta \in \mathcal{S}$. □

D Proofs for Section 4

Proof of Theorem 4.1. We follow the same strategy as in the proof of Theorem 3.1. Let $B := \sup_{f\in\mathscr{C}^\beta(R)} |\operatorname{Bias}_f(\widehat{f}(x_0))|$. Assume first that $B > 0$. By assumption, we can find a function $K \in L^2(\mathbb{R})$ satisfying $\|K\|_{\mathscr{C}^\beta(\mathbb{R})} < (R+\kappa)/4$, $K(0) = 1$ and $K \geq 0$. For such a K, define $U := (R - \|f\|_{\mathscr{C}^\beta([0,1])})/\|K\|_{\mathscr{C}^\beta(\mathbb{R})}$ and observe that $U > 2$, whenever $f \in \mathscr{C}^\beta((R-\kappa)/2)$. Let

$$\mathscr{G} := \Big\{ f_\theta(x) = f(x) + \theta U B K\Big(\frac{x-x_0}{B^{1/\beta}}\Big) : |\theta| \leq 1 \Big\}.$$

As seen in the proof of Theorem 3.1, this defines a subset of $\mathscr{C}^\beta(R)$. As derived in Section 4, the χ^2-divergence in this model is $\chi^2(P_f, P_g) = \exp(n\|f-g\|_1) - 1$, whenever $f \geq g$. By assumption, $B^2 \leq \sup_f \mathrm{MSE}_f(\widehat{f}(x_0)) \leq$

$(C/n)^{2\beta/(\beta+1)}$. Rewriting gives $B^{1+1/\beta} = B^{(\beta+1)/\beta} \leq C/n$. Combining this with the inequality $e^x - 1 \leq xe^x$ and using that $f \leq f_1$ pointwise, we find that

$$\chi^2(P_{f_1}, P_f) \leq n\|f_1 - f\|_1 \exp\Big(n\|f_1 - f\|_1\Big) \leq nUB^{1+1/\beta}\|K\|_1 \exp\big(CU\|K\|_1\big).$$

Applying the χ^2-divergence version of Lemma 2.1 to the random variable $\widehat{f}(x_0)$ and using the just derived bound for the χ^2-divergence yields

$$\big(E_{f_1}\big[\widehat{f}(x_0)\big] - E_f\big[\widehat{f}(x_0)\big]\big)^2 \leq nUB^{1+1/\beta}\|K\|_1 \exp\big(CU\|K\|_1\big)\mathbb{V}\mathrm{ar}_f\big(\widehat{f}(x_0)\big).$$

Due to $K(0) = 1$, we have that $f_1(x_0) - f(x_0) = UB$. Since B is the supremum over the absolute value of the bias, it follows that $E_{f_1}[\widehat{f}(x_0)] - E_f[\widehat{f}(x_0)] \geq UB - 2B$ and consequently

$$(U-2)_+^2 B^2 \leq nUB^{1+1/\beta}\|K\|_1 \exp\big(2CU\|K\|_1\big)\mathbb{V}\mathrm{ar}_f\big(\widehat{f}(x_0)\big). \tag{46}$$

Recall that $U > 2$, whenever $f \in \mathscr{C}^\beta((R-\kappa)/2)$. Due to $\beta < 1$, the bound $B^2 < cn^{-2\beta/(\beta+1)}$ implies $B^{1-1/\beta}/n \geq c^{(1-1/\beta)/2}n^{-2\beta/(\beta+1)}$. By making c sufficiently small, (46) shows that eventually $\mathbb{V}\mathrm{ar}_0(\widehat{f}(x_0)) \geq (C/n)^{2\beta/(\beta+1)}$. This is a contradiction, since also $\mathbb{V}\mathrm{ar}_0(\widehat{f}(x_0)) \leq \mathrm{MSE}_0(\widehat{f}(x_0)) < (C/n)^{2\beta/(\beta+1)}$. Hence, there exists a $c = c(\beta, C, R)$, such that $B^2 \geq cn^{-2\beta/(\beta+1)}$. This proves (18).

To verify (19), we can use that $B^2 \leq \sup_{f\in\mathscr{C}^\beta(R)} \mathrm{MSE}_f(\widehat{f}(x_0)) \leq (C/n)^{2\beta/(\beta+1)}$. This gives $B^{1-1/\beta}/n \geq C^{(1-1/\beta)/2}n^{-2\beta/(\beta+1)}$ and if inserted in (46) shows the existence of a positive constant $c'(\beta, C, R)$ with $\mathbb{V}\mathrm{ar}_f(\widehat{f}(x_0)) \geq c'(\beta, C, R)n^{-2\beta/(\beta+1)}$.

Suppose now that $B = 0$ holds. Then we can add a (deterministic) positive sequence $\delta_n < \sqrt{c}n^{-\beta/(\beta+1)}$ to the estimator such that for the perturbed estimator $\widehat{f}_\delta$, we still have $\sup_{f\in\mathscr{C}^\beta(R)} \mathrm{MSE}_f(\widehat{f}_\delta(x_0)) < (C/n)^{2\beta/(\beta+1)}$. Since $B^2 < cn^{-2\beta/(\beta+1)}$, applying the argument above shows that such an estimator cannot exist. Therefore, $B = 0$ is impossible. ☐

E Proofs for Section 6

Proof of Proposition 6.2. It will be enough to prove the result for Γ_β replaced by $\|K\|_{S^\beta}$ for an arbitrary function $K \in S^\beta(\mathbb{R})$ with $\|K\|_{L^2(\mathbb{R})} = 1$ and support contained in $[-1/2, 1/2]$. Introduce

$$\mathscr{F} := \Big\{ f_\theta(x) = \sum_{i=1}^m \theta_i \sqrt{m} K\big(mx - (i-1/2)\big) : \|\theta\|_2 \leq \frac{R}{\|K\|_{S^\beta} m^\beta} \Big\}. \tag{47}$$

The support of the function $K(mx - (i-1/2))$ is contained in $[i-1, i]$. For different i and j, the dilated and scaled kernel functions have therefore disjoint support and

$$\begin{aligned}\|f_\theta\|_{S^\beta}^2 &= \int_0^1 \Big(\sum_{i=1}^m \theta_i \sqrt{m} K\big(mx - (i-1/2)\big)\Big)^2 dx + \int_0^1 \Big(\sum_{i=1}^m \theta_i m^{\beta+1/2} K^{(\beta)}\big(mx - (i-1/2)\big)\Big)^2 dx \\ &= \sum_{i=1}^m \theta_i^2 \int_0^1 mK\big(mx - (i-1/2)\big)^2 + m^{2\beta+1} K^{(\beta)}\big(mx - (i-1/2)\big)^2\, dx\end{aligned}$$

$$= \sum_{i=1}^{m} \theta_i^2 m^{2\beta} \|K\|_{S^\beta}^2 \leq R,$$

so that $\mathscr{F} \subset S^\beta(R)$, since $\|\theta\|_2 \leq R/(\|K\|_{S^\beta} m^\beta)$. It is therefore sufficient to prove Proposition 6.2 with $S^\beta(R)$ replaced by $\mathscr{F}$. We say that two statistical models are equivalent if the data can be transformed into each other without knowledge of the unknown parameters. The Gaussian white noise model (14) is by definition equivalent to observing all functionals $\int_0^1 \phi(t)\, dY_t$ with $\phi \in L^2([0,1])$. In particular, for any orthonormal $L^2([0,1])$ basis $(\phi_i)_{i=1,\ldots}$, the Gaussian white noise model is equivalent to observing $X_i := \int_0^1 \phi_i(t)\, dY_t$, $i = 1, \ldots$ The latter is the well-known sequence space formulation. The functions $\psi_i := \sqrt{m} K(m \cdot -(i-1/2))$ are orthogonal (because of the disjoint support) and L^2-normalized. Choosing $\phi_i = \psi_i$ for $i = 1, \ldots, m$ and extending this to an orthonormal basis of $L^2([0,1])$, we find that the Gaussian white noise model with parameter space $\mathscr{F}$ is equivalent to observing

$$X_i = \theta_i \mathbf{1}(i \leq m) + \frac{1}{\sqrt{n}} \varepsilon_i, \quad i = 1, \ldots$$

with independent $\varepsilon_i \sim \mathcal{N}(0,1)$. Here we have used that $\int_0^1 \phi_i(t)\, dY_t = \int_0^1 \phi_i(t) f(t)\, dt + n^{-1/2} \int_0^1 \phi_i(t)\, dW_t$ and that $\varepsilon_i := \int_0^1 \phi_i(t)\, dW_t$ are standard normal and independent.

Because of the equivalence, every estimator $\widehat{f}$ in the Gaussian white noise model with parameter space $\mathscr{F}$ can be rewritten as an estimator $\widehat{f} = \widehat{f}(X_1, \ldots)$ depending on the transformed data $X_1, X_2, \ldots$ Moreover, for any estimator $\widehat{f}$ for the regression f in the Gaussian white noise model, we can consider the estimator $\widetilde{\theta} = (\widetilde{\theta}_1, \ldots, \widetilde{\theta}_m)$ with $\widetilde{\theta}_i := \int_0^1 \widehat{f}(x) \psi_i(x)\, dx$. This is now an estimator depending on $X_1, X_2, \ldots$ Observe that $(X_1, \ldots, X_m)$ is a sufficient statistic for the vector θ. In view of the Rao-Blackwell theorem, it is then natural to eliminate the dependence on $X_{m+1}, X_{m+2}, \ldots$ by considering the estimator $\widehat{\theta}_i := E[\widetilde{\theta}_i | X_1, \ldots, X_m]$. This estimator only depends on the Gaussian sequence model $(X_1, \ldots, X_m)$.

The proof is complete if we can show that $\mathrm{IBias}_\theta^2(\widehat{\theta}) \leq \mathrm{IBias}_{f_\theta}^2(\widehat{f})$ and $\mathrm{IVar}_\theta(\widehat{\theta}) \leq \mathrm{IVar}_{f_\theta}(\widehat{f})$ for all $f_\theta \in \mathscr{F}$, or equivalently, for all $\theta \in \Theta$. First observe that $\mathrm{IBias}_\theta^2(\widehat{\theta}) = \mathrm{IBias}_\theta^2(\widetilde{\theta})$ and by using the formula for the conditional variance, we have $\mathrm{IVar}_\theta(\widehat{\theta}) = \mathrm{IVar}_\theta(\widetilde{\theta}) - E[\mathrm{IVar}_\theta(\widetilde{\theta} | X_1, \ldots, X_m)] \leq \mathrm{IVar}_\theta(\widetilde{\theta})$. It is therefore sufficient to show that $\mathrm{IBias}_\theta^2(\widetilde{\theta}) \leq \mathrm{IBias}_{f_\theta}^2(\widehat{f})$ and $\mathrm{IVar}_\theta(\widetilde{\theta}) \leq \mathrm{IVar}_{f_\theta}(\widehat{f})$ for all $f_\theta \in \mathscr{F}$.

Denote by $\mathscr{G}$ the linear span of $(\psi_i)_{i=1,\ldots,m}$ and by $\mathscr{G}^c$ the orthogonal complement of $\mathscr{G}$ in $L^2([0,1])$. Obviously, $\mathscr{G}$ is a finite-dimensional subspace of $L^2([0,1])$ and hence closed. Let $\widetilde{f} := \sum_{i=1}^m \widetilde{\theta}_i \psi_i$ with $\widetilde{\theta}_i$ as defined above. Since $\widetilde{f}$ is the L^2-projection of f on $\mathscr{G}$, it holds that $\widehat{f} - \widetilde{f} \in \mathscr{G}^c$. Consequently, $\widetilde{f}$ and $\widehat{f} - \widetilde{f}$ must be orthogonal in $L^2([0,1])$. Moreover, also $E_{f_\theta}[\widetilde{f}] \in \mathscr{G}$ and $E_{f_\theta}[\widehat{f} - \widetilde{f}] \in \mathscr{G}^c$. Therefore, for any $f_\theta \in \mathscr{F}$,

$$\begin{aligned} \mathrm{IVar}_{f_\theta}(\widehat{f}) &= \int_0^1 \mathbb{V}\mathrm{ar}_{f_\theta}(\widehat{f}(x))\, dx = \int_0^1 E_{f_\theta}\Big[\big(\widehat{f}(x)) - E_{f_\theta}[\widehat{f}(x)]\big)^2\Big]\, dx = E_{f_\theta}\Big[\int_0^1 \big(\widehat{f}(x)) - E_{f_\theta}[\widehat{f}(x)]\big)^2\, dx\Big] \\ &= E_{f_\theta}\Big[\big\|\widetilde{f} + (\widehat{f} - \widetilde{f}) - E_{f_\theta}\big[\widetilde{f} + (\widehat{f} - \widetilde{f})\big]\big\|_2^2\Big] = E_{f_\theta}\Big[\big\|\widetilde{f} - E_{f_\theta}[\widetilde{f}]\big\|_2^2\Big] + E_{f_\theta}\Big[\big\|\widehat{f} - \widetilde{f} - E_{f_\theta}[\widehat{f} - \widetilde{f}]\big\|_2^2\Big] \\ &\geq E_{f_\theta}\Big[\big\|\widetilde{f} - E_{f_\theta}[\widetilde{f}]\big\|_2^2\Big] = \mathrm{IVar}_{f_\theta}(\widetilde{f}). \end{aligned}$$

Using that the ψ_i are orthonormal with respect to $L^2([0,1])$,

$$\begin{aligned}\mathrm{IVar}_{f_\theta}(\widetilde{f}) &= \int_0^1 E_{f_\theta}\left[\left(\sum_{i=1}^m(\widetilde{\theta}_i - E_{f_\theta}[\widetilde{\theta}_i])\psi_i(x)\right)^2\right]dx = \int_0^1 E_{f_\theta}\left[\sum_{i=1}^m\left(\widetilde{\theta}_i - E_{f_\theta}[\widetilde{\theta}_i]\right)^2\psi_i^2(x)\right]dx \\ &= \sum_{i=1}^m \mathbb{V}\mathrm{ar}_\theta\big(\widetilde{\theta}_i\big) = \mathrm{IVar}_\theta\left(\widetilde{\theta}_i\right).\end{aligned}$$

Combined with the previous display, this proves that $\mathrm{IVar}_\theta(\widetilde{\theta}) \leq \mathrm{IVar}_{f_\theta}(\widehat{f})$ for all $f_\theta \in \mathscr{F}$.

With the same notation as above, we find using $f_\theta \in \mathscr{G}$,

$$\begin{aligned}\mathrm{IBias}^2_{f_\theta}\left(\widehat{f}\right) &= \int_0^1 \left(E_{f_\theta}[\widehat{f}(x)] - f_\theta(x)\right)^2 dx = \big\|E_{f_\theta}[\widehat{f}] - f_\theta\big\|_2^2 = \big\|E_{f_\theta}[\widetilde{f}] - f_\theta\big\|_2^2 + \big\|E_{f_\theta}[\widehat{f} - \widetilde{f}]\big\|_2^2 \\ &\geq \big\|E_{f_\theta}[\widetilde{f}] - f_\theta\big\|_2^2 = \mathrm{IBias}^2_{f_\theta}\left(\widetilde{f}\right)\end{aligned}$$

and

$$\begin{aligned}\mathrm{IBias}^2_{f_\theta}(\widetilde{f}) &= \int_0^1 \left(E_{f_\theta}[\widetilde{f}(x)] - f_\theta(x)\right)^2 dx = \int_0^1 \left(\sum_{i=1}^m\left(E_{f_\theta}[\widetilde{\theta}_i] - \theta_i\right)\psi_i(x)\right)^2 dx \\ &= \int_0^1 \sum_{i=1}^m \left(E_{f_\theta}[\widetilde{\theta}_i] - \theta_i\right)^2 \psi_i^2(x)\,dx = \sum_{i=1}^m\left(E_{f_\theta}[\widetilde{\theta}_i] - \theta_i\right)^2 = \mathrm{IBias}^2_\theta\left(\widetilde{\theta}\right).\end{aligned}$$

This finally proves $\mathrm{IBias}^2_\theta(\widetilde{\theta}) \leq \mathrm{IBias}^2_{f_\theta}(\widehat{f})$. The proof is complete. □

Proof of Proposition 6.3. We follow Stein [38, p.201] and denote by μ the Haar measure on the orthogonal group $\mathcal{O}_m$. In particular, $\mu(\mathcal{O}_m) = 1$. We write $\widehat{\theta}(X)$ and $\widetilde{\theta}(X)$ to highlight the dependence on the sample $X \in \mathbb{R}^m$. Given $\widehat{\theta}(X)$, define

$$\widetilde{\theta}(X) := \int D^{-1}\widehat{\theta}(DX)\,d\mu(D),$$

where the integral is over the orthogonal group. By construction, $\widetilde{\theta}(X)$ is a spherically symmetric estimator. Using Jensen's inequality, the fact that $DX \sim \mathcal{N}(D\theta, I_m/n)$ with I_m the $m \times m$ identity matrix, and $\theta = D^{-1}D\theta$ yields for any $\theta \in \Theta_m^\beta(R)$,

$$\begin{aligned}\mathrm{IBias}^2_\theta\left(\widetilde{\theta}(X)\right) &= \big\|E_\theta[\widetilde{\theta}(X)] - \theta\big\|_2^2 = \left\|E_\theta\left[\int_{D\in\mathcal{O}_m} D^{-1}\widehat{\theta}(DX)\,d\mu(D)\right] - \theta\right\|_2^2 \\ &\leq \int_{D\in\mathcal{O}_m}\left\|E_\theta[D^{-1}\widehat{\theta}(DX)] - \theta\right\|_2^2 d\mu(D) \\ &\leq \int_{D\in\mathcal{O}_m}\left\|E_{D\theta}[D^{-1}\widehat{\theta}(X)] - \theta\right\|_2^2 d\mu(D) \\ &\leq \int_{D\in\mathcal{O}_m}\left\|E_{D\theta}[\widehat{\theta}(X)] - D\theta\right\|_2^2 d\mu(D) \\ &\leq \sup_{\theta\in\Theta_m^\beta(R)} \mathrm{IBias}^2_\theta\left(\widehat{\theta}(X)\right).\end{aligned}$$

With e_i the i-th standard basis vector of $\mathbb{R}^m$, we also find using that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$, $D = (D^{-1})^\top$, and again $DX \sim \mathcal{N}(D\theta, I_m/n)$,

$$\sum_{i=1}^m \mathbb{V}\mathrm{ar}_\theta\big(\widetilde{\theta}_i(X)\big) = \int_{D\in\mathcal{O}_m}\sum_{i=1}^m \mathbb{V}\mathrm{ar}_\theta\Big(e_i^\top D^{-1}\widehat{\theta}(DX)\Big)\,d\mu(D)$$

$$
\begin{aligned}
&= \int_{D\in\mathcal{O}_m} \operatorname{Tr}\Big[\mathbb{V}\mathrm{ar}_\theta\big(D^{-1}\widehat{\theta}(DX)\big)\Big]\, d\mu(D) \\
&= \int_{D\in\mathcal{O}_m} \operatorname{Tr}\Big[D^{-1}\mathbb{V}\mathrm{ar}_\theta\big(\widehat{\theta}(DX)\big)(D^{-1})^\top\Big]\, d\mu(D) \\
&= \int_{D\in\mathcal{O}_m} \operatorname{Tr}\Big[\mathbb{V}\mathrm{ar}_\theta\big(\widehat{\theta}(DX)\big)\Big]\, d\mu(D) \\
&= \int_{D\in\mathcal{O}_m} \operatorname{Tr}\Big[\mathbb{V}\mathrm{ar}_{D\theta}\big(\widehat{\theta}(X)\big)\Big]\, d\mu(D) \\
&\le \sup_{\theta\in\Theta_m^\beta(R)} \operatorname{IVar}_\theta\big(\widehat{\theta}\big).
\end{aligned}
$$

□

Lemma E.1. *Any function $h(x)$ satisfying $h(x) = D^{-1}h(Dx)$ for all $x \in \mathbb{R}^m$ and all orthogonal transformations D must be of the form*

$$h(x) = r(\|x\|_2)x$$

for some univariate function r.

Proof. Throughout the proof, we write $\|\cdot\|$ for the Euclidean norm. In a first step of the proof, we show that

$$h(x) = \lambda(x)x \tag{48}$$

for some univariate function λ.

Fix x and consider an orthogonal basis $v_1 := x/\|x\|, v_2, \ldots, v_m$ of $\mathbb{R}^m$. The orthogonal matrix $D := \sum_{j=1}^m (-1)^{\mathbf{1}(j\neq 1)} v_j v_j^\top$ has eigenvector $v_1 = x/\|x\|$ with corresponding eigenvalue one. For all other eigenvectors the eigenvalue is always -1. Using that $h(x) = D^{-1}h(Dx)$, we find that $h(x) = D^{-1}h(x)$ which implies that $h(x)$ is a multiple of x and therefore $h(x) = \lambda(x)x$, proving (48).

Let x and y be such that $\|x\| = \|y\|$. Let $v = x - y$, and observe that $D = I - 2vv^\top/\|v\|^2$ is an orthogonal matrix. Since $\|v\|^2 = 2\|x\|^2 - 2y^\top x = 2\|y\|^2 - 2y^\top x$, we also have that $Dx = y$ and $Dy = x$. For this D, we have

$$\lambda(x)x = h(x) = h(Dy) = Dh(y) = \lambda(y)Dy = \lambda(y)x$$

which shows that $\lambda(x) = \lambda(y)$ whenever $\|x\| = \|y\|$. Differently speaking, λ only depends on y through $\|y\|$. This completes the proof. □

F Proofs for Section 7

Proof of Lemma 7.1. Applying the triangle inequality and the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\big(1 - H^2(P,Q)\big)\big|u - v\big| &= \int \big|X(\omega) - u - X(\omega) + v\big| \sqrt{p(\omega)q(\omega)}\, d\nu(\omega) &(49)\\
&\le \int \big|X(\omega) - u\big| \sqrt{p(\omega)q(\omega)}\, d\nu(\omega) + \int \big|X(\omega) - v\big| \sqrt{p(\omega)q(\omega)}\, d\nu(\omega) &(50)
\end{aligned}
$$

$$\leq \sqrt{E_P\big[|X-u|\big]E_Q\big[|X-u|\big]} + \sqrt{E_P\big[|X-v|\big]E_Q\big[|X-v|\big]}. \tag{51}$$

Bound $E_Q[|X-u|] \leq E_Q[|X-v|] + |u-v|$ and $E_P[|X-v|] \leq E_P[|X-u|] + |u-v|$. With $a := E_P[|X-v|] \vee E_Q[|X-u|]$, $b := |u-v|$ and $d := 1 - H^2(P,Q)$, we then have $db \leq 2\sqrt{a^2+ab}$ or equivalently $a^2 + ab - d^2b^2/4 \geq 0$. Since $a \geq 0$, solving the quadratic equation $a^2 + ab - d^2b^2/4 = 0$ in a gives that $a \geq b(\sqrt{1+d^2} - 1)/2$. Since $0 \leq d \leq 1$, we also have that $\sqrt{1+d^2} - 1 \geq 2d^2/5$, which can be verified by adding one to both sides and squaring. Combining the last two inequalities gives finally the desired result $a \geq bd^2/5$. □

Proof of Theorem 7.2. The proof is a variation of the proof for Theorem 3.1. For $K \in \mathscr{C}^\beta(\mathbb{R})$ any function satisfying $K(0) = 1$ and $\|K\|_2 < +\infty$, define $V := R/\|K\|_{\mathscr{C}^\beta(R)}$, $r_n := (2/V)^{1/\beta}(C/n)^{1/(2\beta+1)}$, and

$$\mathscr{F} := \Big\{ f_\theta(x) = \theta V r_n^\beta K\Big(\frac{x-x_0}{r_n}\Big) : |\theta| \leq 1 \Big\}.$$

Arguing as in the proof for Theorem 3.1 (i), we have $\mathscr{F} \subseteq \mathscr{C}^\beta(R)$ whenever $r_n \leq 1$, which holds for all sufficiently large n. As mentioned at the beginning of Section 3, $1 - H^2(P_f, P_g) = \exp(-\frac{n}{8}\|f-g\|_2^2)$. We can now apply Lemma 7.1 to the random variable $\widehat{f}(x_0)$ choosing $P = P_{f_{\pm 1}}$, $Q = P_0$ and centering $u = E_{f_{\pm 1}}[\widehat{f}(x_0)]$, $v = E_0[\widehat{f}(x_0)]$,

$$\frac{1}{5}\exp\Big(-\frac{n}{4}\|f_{\pm 1}\|_2^2\Big)\Big|E_{f_{\pm 1}}[\widehat{f}(x_0)] - E_0[\widehat{f}(x_0)]\Big| \leq E_{f_{\pm 1}}\big[\big|\widehat{f}(x_0) - E_{f_{\pm 1}}[\widehat{f}(x_0)]\big|\big] \vee E_0\big[\big|\widehat{f}(x_0) - E_0[\widehat{f}(x_0)]\big|\big],$$

Now $\|f_{\pm 1}\|_2^2 \leq V^2 r_n^{2\beta+1}\|K\|_2^2 = 2^{2+1/\beta}V^{-1/\beta}C\|K\|_2^2/n$ and so,

$$\big|E_{f_{\pm 1}}[\widehat{f}(x_0)] - E_0[\widehat{f}(x_0)]\big| \leq 5\exp\big((2/V)^{1/\beta}C\|K\|_2^2\big) \sup_{f \in \mathscr{C}^\beta(R)} E_f\big|\widehat{f}(x_0) - E_f[\widehat{f}(x_0)]\big|.$$

Due to $K(0) = 1$, we have $f_{\pm 1}(x_0) = \pm V r_n^\beta = 2(C/n)^{\beta/(2\beta+1)}$ and because of the bound on the bias, $E_{f_1}[\widehat{f}(x_0)] \geq (C/n)^{\beta/(2\beta+1)}$ and $E_{f_{-1}}[\widehat{f}(x_0)] \leq -(C/n)^{\beta/(2\beta+1)}$. Choosing for the lower bound f_1 if $E_{f_0}[\widehat{f}(x_0)]$ is negative and f_{-1} if $E_{f_0}[\widehat{f}(x_0)]$ is positive, we find

$$\frac{1}{5}\exp\big(-(2/V)^{1/\beta}C\|K\|_2^2\big)\Big(\frac{C}{n}\Big)^{\frac{\beta}{2\beta+1}} \leq \sup_{f \in \mathscr{C}^\beta(R)} E_f\big|\widehat{f}(x_0) - E_f[\widehat{f}(x_0)]\big|.$$

This shows the claim. The proof for the median centering follows exactly the same steps. □

References

[1] J. Acharya, Z. Sun, and H. Zhang. Differentially Private Assouad, Fano, and Le Cam. *arXiv e-prints*, page arXiv:2004.06830, 2020.

[2] M. Belkin, D. Hsu, S. Ma, and S. Mand al. Reconciling modern machine learning practice and the bias-variance trade-off. *arXiv e-prints*, page arXiv:1812.11118, 2018.

[3] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In S. Shalev-Shwartz and I. Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 1046–1066, Princeton, NJ, USA, 12–14 Jun 2013. PMLR.

[4] Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *Ann. Statist.*, 41(4):1780–1815, 2013.

[5] L. Breiman. Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, University of California, Berkeley, CA, USA, 1996.

[6] L. D. Brown and M. G. Low. A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.*, 24(6):2524–2535, 1996.

[7] L. D. Brown, M. G. Low, and L. H. Zhao. Superefficiency in nonparametric function estimation. *Ann. Statist.*, 25(6):2607–2625, 1997.

[8] P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[9] T. T. Cai and Z. Guo. Confidence intervals for high-dimensional linear regression: minimax rates and adaptivity. *Ann. Statist.*, 45(2):615–646, 2017.

[10] T. T. Cai, J. Jin, and M. G. Low. Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.*, 35(6):2421–2449, 2007.

[11] D. Chafaï. Entropies, convexity, and functional inequalities: on Φ-entropies and Φ-Sobolev inequalities. *J. Math. Kyoto Univ.*, 44(2):325–363, 2004.

[12] P. Domingos. A unified bias-variance decomposition for zero-one and squared loss. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, page 564–569. AAAI Press, 2000.

[13] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 2004.

[14] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18, 2006.

[15] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern. Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B*, 54(1):41–81, 1992. With discussion and a reply by the authors.

[16] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438, 2013.

[17] J. Fan and P. Hall. On curve estimation by minimizing mean absolute deviation and its implications. *Ann. Statist.*, 22(2):867–885, 1994.

[18] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[19] J. V. Hansen and T. Heskes. General bias/variance decomposition with target independent variance of error functions derived from the exponential family of distributions. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 2, pages 207–210. IEEE, 2000.

[20] T. Heskes. Bias/variance decompositions for likelihood-based estimators. *Neural Computation*, 10(6):1425–1433, 1998.

[21] Y. I. Ingster. Minimax detection of a signal for l^n-balls. *Math. Methods Statist.*, 7(4):401–428 (1999), 1998.

[22] G. James. Variance and bias for general loss functions. *Machine Learning*, 51:115–135, 2003.

[23] G. James and T. Hastie. Generalizations of the bias/variance decomposition for prediction error. *Dept. Statistics, Stanford Univ., Stanford, CA, Tech. Rep*, 1997.

[24] I. M. Johnstone. Gaussian estimation: Sequence and wavelet models, September 2019. URL: `http://statweb.stanford.edu/~imj/GE_09_16_19.pdf`.

[25] R. Kohavi and D. Wolpert. Bias plus variance decomposition for zero-one loss functions. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ICML'96, page 275–283, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc.

[26] Y. Le Borgne. Bias-variance trade-off characterization in a classification problem: What differences with regression. Technical report, Machine Learning Group, Univ. Libre de Bruxelles, Belgium, 2005.

[27] M. G. Low. Bias-variance tradeoffs in functional estimation problems. *Ann. Statist.*, 23(3):824–835, 1995.

[28] A. Meister and M. Reiß. Asymptotic equivalence for nonparametric regression with non-regular errors. *Probab. Theory Related Fields*, 155(1-2):201–229, 2013.

[29] B. Neal, S. Mittal, A. Baratin, V. Tantia, M. Scicluna, S. Lacoste-Julien, and I. Mitliagkas. A Modern Take on the Bias-Variance Tradeoff in Neural Networks. *arXiv e-prints*, page arXiv:1810.08591, Oct. 2018.

[30] J. Neville and D. Jensen. Bias/variance analysis for relational domains. In *International Conference on Inductive Logic Programming*, pages 27–28. Springer, 2007.

[31] J. Pfanzagl. A nonparametric asymptotic version of the Cramér-Rao bound. In *State of the art in probability and statistics (Leiden, 1999)*, volume 36 of *IMS Lecture Notes Monogr. Ser.*, pages 499–517. Inst. Math. Statist., Beachwood, OH, 2001.

[32] M. Raginsky. Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels. *IEEE Trans. Inform. Theory*, 62(6):3355–3389, 2016.

[33] M. Reiß and J. Schmidt-Hieber. Posterior contraction rates for support boundary recovery. *arXiv e-prints*, page arXiv:1703.08358, Mar. 2017.

[34] M. Reiß and J. Schmidt-Hieber. Nonparametric Bayesian analysis of the compound Poisson prior for support boundary recovery. *Annals of Statistics*, 2020. To appear.

[35] M. Reiß and L. Selk. Efficient estimation of functionals in nonparametric boundary models. *Bernoulli*, 23(2):1022–1055, 2017.

[36] A. Rohde and L. Steinberger. Geometrizing rates of convergence under local differential privacy constraints. *arXiv e-prints*, page arXiv:1805.01422, May 2018.

[37] D. Rozmus. Methods of classification error decompositions and their properties. *Acta Universitatis Lodziensis. Folia Oeconomica*.

[38] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceeding of the fourth Berkeley symposium on Mathematical statistics and Probability*, volume 1, pages 197–206. University of California Press, 1956.

[39] B. Szabo and H. van Zanten. Adaptive distributed methods under communication constraints. *arXiv e-prints*, page arXiv:1804.00864, Apr. 2018.

[40] R. Tibshirani. Bias, variance and prediction error for classification rules. Technical report, Statistics Department, University of Toronto, 1996.

[41] A. B. Tsybakov. *Introduction to non-parametric estimation*. Springer Series in Statistics. Springer, 2009.

[42] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014.

[43] D. H. Wolpert. On bias plus variance. *Neural Computation*, 9(6):1211–1243, 1997.

[44] Q. Wu and P. Vos. Decomposition of Kullback-Leibler risk and unbiasedness for parameter-free estimators. *J. Statist. Plann. Inference*, 142(6):1525–1536, 2012.

[45] C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(1):217–242, 2014.

[46] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2328–2336. Curran Associates, Inc., 2013.