Statistics and Probability Letters 213 (2024) 110182

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

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

Alexis Derumigny [a], Johannes Schmidt-Hieber [b,*]

[a] TU Delft, Mekelweg 5, Delft, 2628 CD, The Netherlands
[b] University of Twente, Drienerlolaan 5, Enschede, 7522 NB, The Netherlands

ARTICLE INFO

MSC:
62C20
62G05
62C05

Keywords:
Bias–variance trade-off
Mean absolute deviation
Minimax estimation
Nonparametric estimation

ABSTRACT

In nonparametric statistics, rate-optimal estimators typically balance bias and stochastic error. The recent work on overparametrization raises the question whether rate-optimal estimators exist that do not obey this trade-off. In this work we consider pointwise estimation in the Gaussian white noise model with regression function f in a class of β-Hölder smooth functions. Let 'worst-case' refer to the supremum over all functions f in the Hölder class. It is shown that any estimator with worst-case bias $\lesssim n^{-\beta/(2\beta+1)} =: \psi_n$ must necessarily also have a worst-case mean absolute deviation that is lower bounded by $\gtrsim \psi_n$. To derive the result, we establish abstract inequalities relating the change of expectation for two probability measures to the mean absolute deviation.

1. Motivation

Inspired by recent claims that overparametrization challenges the traditional view on the bias–variance trade-off, see for instance (Belkin et al., 2019; Neal et al., 2018; Neal, 2019), we aim to quantify the extend to which the trade-off between bias and stochastic error in nonparametric and highdimensional statistics is universal. The recent work (Derumigny and Schmidt-Hieber, 2023) derives lower bounds for the bias–variance trade-off covering standard nonparametric and high-dimensional statistical models. In this work, we take this one step further by deriving a universal lower bound for the trade-off between bias and mean absolute deviation. Such universal lower bounds immediately translate into universal lower bounds for the bias–variance trade-off and are thus stronger.

Another motivation for our work is that for constructing confidence bands with small diameter in function estimation problems, one needs to find an upper bound for the bias. The bias is hard to estimate from data, see e.g. Hall and Horowitz (2013). To obtain a small confidence bands, it is therefore desirable to find rate-optimal estimators with negligible bias. Universal lower bounds on the trade-off between bias and stochastic error can be a tool to show that this is impossible in the sense that decreasing the bias necessarily increases the stochastic error.

2. Summary of previous work on universal lower bounds for the bias–variance trade-off

The previous work (Derumigny and Schmidt-Hieber, 2023) derives universal lower bounds for the bias–variance trade-off. For nonparametric function estimation, evaluating an estimator either via the squared pointwise risk or the mean integrated squared error, it is shown that there exists a universal bias–variance trade-off that also cannot be overcome by fitting overparametrized models.

* Corresponding author.
E-mail address: a.j.schmidt-hieber@utwente.nl (J. Schmidt-Hieber).

https://doi.org/10.1016/j.spl.2024.110182
Received 21 March 2023; Received in revised form 3 June 2024; Accepted 4 June 2024
Available online 13 June 2024
0167-7152/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

For estimation of a high-dimensional sparse vector in the Gaussian sequence model, the situation is different and the bias–variance trade-off does not always hold. Derumigny and Schmidt-Hieber (2023) shows that there are estimation problems driven by the worst-case bias. While the convergence rate of the worst-case variance cannot be arbitrarily fast, it can be considerably faster than the minimax estimation rate.

These lower bounds on the bias–variance trade-off rely on a number of abstract inequalities that all relate the variance to the changes that occur if expectations are taken with respect to different probability measures. To outline the idea, we recall one of these change of expectation inequalities that we modify later on. Let P and Q be two probability distributions on the same measurable space. Denote by E_P and Var_P the expectation and variance with respect to P and let E_Q and Var_Q be the expectation and variance with respect to Q. The squared Hellinger distance is defined by $H(P,Q)^2 := \frac{1}{2}\int(\sqrt{p(\omega)}-\sqrt{q(\omega)})^2\,d\nu(\omega)$ with ν a measure dominating both P and Q and p,q the respective ν-densities of P and Q. It can be checked that the Hellinger distance does not depend on the choice of ν.

Lemma 2.1 (*Lemma 2.1 in Derumigny and Schmidt-Hieber (2023)*). *For any random variable X,*

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

To derive a lower bound on the bias–variance trade-off from such an inequality, consider a statistical model $(P_\theta : \theta\in\Theta)$ with P_θ the distribution of the data for parameter θ and $\Theta\subseteq\mathbb{R}$ the parameter space. Choosing two parameters $\theta,\theta'\in\Theta$, inequality (1) shows that for any estimator $\widehat\theta$,

$$\frac{(E_\theta[\widehat\theta]-E_{\theta'}[\widehat\theta])^2}{4-2r^2(\theta,\theta')}\Big(\frac{1}{r(\theta,\theta')}-r(\theta,\theta')\Big)^2 \le \mathrm{Var}_\theta\big(\widehat\theta\big)+\mathrm{Var}_{\theta'}\big(\widehat\theta\big), \tag{2}$$

with $r(\theta,\theta') := H(P_\theta,P_{\theta'})$, $E_\theta := E_{P_\theta}$, and $\mathrm{Var}_\theta := \mathrm{Var}_{P_\theta}$. Introducing the bias $\mathrm{Bias}_\theta(\widehat\theta)=\theta-E_\theta[\widehat\theta]$, one can now rewrite the difference of the expectations as $E_\theta[\widehat\theta]-E_{\theta'}[\widehat\theta]=\theta-\theta'-\mathrm{Bias}_\theta(\widehat\theta)+\mathrm{Bias}_{\theta'}(\widehat\theta)$. If we assume that the bias is smaller than some value, say B, and take the parameters θ,θ' sufficiently far apart, such that $|\theta-\theta'|\ge 4B$, reverse triangle inequality yields $|E_\theta[\widehat\theta]-E_{\theta'}[\widehat\theta]|\ge\frac{1}{2}|\theta-\theta'|$ and (2) becomes

$$\frac{\frac{1}{4}(\theta-\theta')^2}{4-2r^2(\theta,\theta')}\Big(\frac{1}{r(\theta,\theta')}-r(\theta,\theta')\Big)^2 \le 2\sup_{\theta\in\Theta}\mathrm{Var}_\theta\big(\widehat\theta\big).$$

The left hand side of this inequality does not depend on the estimator $\widehat\theta$ anymore. Therefore, this inequality provides us with a lower bound on the worst-case variance of an arbitrary estimator.

While this applies to a one-dimensional parameter, the same procedure can immediately be extended to derive lower bounds on the worst-case variance for pointwise estimation of a function value $f(x_0)$ in a nonparametric statistical model with unknown regression function f. As shown in Derumigny and Schmidt-Hieber (2023), one can extend these ideas moreover to derive lower bounds for the integrated variance and for (high-dimensional) parameter vectors.

Rephrasing the argument leads moreover to lower bounds on the worst-case bias given an upper bound for the worst-case variance. Taking a suitable asymptotics $\theta'\to\theta$ and imposing standard regularity conditions, it can be shown moreover that (2) converges to the Cramér-Rao lower bound (Theorem A.4 in Derumigny and Schmidt-Hieber (2023)).

3. Lower bounds for bias-MAD trade-off

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 $E\big[|X-u|\big]$, where the centering point u is either the mean or the median of X. If centered at the mean, the MAD is upper bounded by $\sqrt{\mathrm{Var}(X)}$, but compared to the variance, less weight is given to large outcomes of X. For a statistical model $(P_\theta : \theta\in\Theta)$, 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 first result provides an abstract inequality that can be used to relate $m(\theta)-\theta$ and $E_\theta[|\widehat\theta-m(\theta)|]$, for any centering $m(\theta)$. It can be viewed as an analogue of (1).

Lemma 3.1. *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*

$$\frac{1}{5}\big(1-H^2(P,Q)\big)^2|u-v| \le E_P\Big[\big|X-u\big|\Big]\vee E_Q\Big[\big|X-v\big|\Big], \tag{3}$$

Proof. Applying the triangle inequality and the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&\big(1-H^2(P,Q)\big)\big|u-v\big| \\
&=\int\Big|X(\omega)-u-X(\omega)+v\Big|\sqrt{p(\omega)q(\omega)}\,d\nu(\omega) \\
&\le\int\Big|X(\omega)-u\Big|\sqrt{p(\omega)q(\omega)}\,d\nu(\omega)+\int\Big|X(\omega)-v\Big|\sqrt{p(\omega)q(\omega)}\,d\nu(\omega) \\
&\le\sqrt{E_P\Big[\big|X-u\big|\Big]E_Q\Big[\big|X-u\big|\Big]}+\sqrt{E_P\Big[\big|X-v\big|\Big]E_Q\Big[\big|X-v\big|\Big]}.
\end{aligned} \tag{4}$$

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 $a \geq bd^2/5$. ☐

The derived inequality does not directly follow from the triangle inequality $|u-v| \leq |x-u| + |x-v|$ as the expectations on the right-hand side of (4) are taken with respect to different measures P and Q. Equality up to a constant multiple is attained if $H(P,Q) < 1$ and $X = v$ with probability 1.

An important special case of the previously derived inequality is

$$\frac{1}{5}\big(1-H^2(P,Q)\big)^2\Big|E_P[X]-E_Q[X]\Big| \leq E_P\Big[\Big|X-E_P[X]\Big|\Big] \vee E_Q\Big[\Big|X-E_Q[X]\Big|\Big]. \tag{5}$$

Let us now compare this to the change of expectation inequalities involving the variance in the regime where the measures P and Q are close. As mentioned above, $E_P[|X-E_P[X]|] \leq \sqrt{\mathrm{Var}_P(X)}$. Moreover, $E_P[|X-E_P[X]|]$ and $\sqrt{\mathrm{Var}\, P(X)}$ are typically of the same magnitude. The Hellinger lower bound for the variance (1) is

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

Compared to (5), 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 next result shows that improving in this regime the inequality (5) requires that the likelihood ratio is uniformly close to one. This is much stronger. For instance if P_θ denotes the distribution of $\mathcal{N}(\theta,1)$, then, $H(P_\theta,P_{\theta'}) = 1-e^{-\frac{1}{8}(\theta-\theta')^2}$, and $H(P_\theta,P_{\theta'}) \to 0$ if $\theta-\theta' \to 0$. However, the likelihood ratio $dP_\theta/dP_{\theta'}$ is unbounded whenever $\theta \neq \theta'$.

Lemma 3.2. *Define $0/0$ as 0. If p and q are the respective ν-densities of P and Q, then*

$$\frac{1-H^2(P,Q)}{\|\frac{p-q}{p\wedge q}\|_{L^\infty}}\Big|E_P[X]-E_Q[X]\Big| \leq E_P\Big[\Big|X-E_P[X]\Big|\Big] \vee E_Q\Big[\Big|X-E_Q[X]\Big|\Big]. \tag{6}$$

Moreover, if P,Q are defined on a finite probability space, then there exists a random variable X^*, such that

$$E_P\Big[\Big|X^*-E_P[X^*]\Big|\Big] \vee E_Q\Big[\Big|X^*-E_Q[X^*]\Big|\Big] \leq \frac{1}{\|\frac{p-q}{p\vee q}\|_{L^\infty}}\Big|E_P[X^*]-E_Q[X^*]\Big|. \tag{7}$$

Proof. Using that $\int(X(\omega)-E_P[X])p(\omega)\,d\nu(\omega) = 0$ and $\int(X(\omega)-E_Q[X])q(\omega)\,d\nu(\omega) = 0$, we have the identity

$$\begin{aligned}
&\int \sqrt{p(\omega)q(\omega)}\,d\nu(\omega)\,\big(E_P[X]-E_Q[X]\big)\\
&= \int \big(X(\omega)-E_P[X]\big)\Big(p(\omega)-\sqrt{p(\omega)q(\omega)}\Big)d\nu(\omega)\\
&\quad + \int \big(X(\omega)-E_Q[X]\big)\Big(\sqrt{p(\omega)q(\omega)}-q(\omega)\Big)d\nu(\omega).
\end{aligned}$$

Taking the absolute value inside the integrals gives

$$\begin{aligned}
&\int \sqrt{p(\omega)q(\omega)}\,d\nu(\omega)\,\Big|E_P[X]-E_Q[X]\Big|\\
&\leq \int \Big|X(\omega)-E_P[X]\Big|p(\omega)d\nu(\omega)\,\Big\|1-\sqrt{\frac{q(\omega)}{p(\omega)}}\Big\|_{L^\infty}\\
&\quad + \int \Big|X(\omega)-E_Q[X]\Big|q(\omega)d\nu(\omega)\,\Big\|\sqrt{\frac{p(\omega)}{q(\omega)}}-1\Big\|_{L^\infty}.
\end{aligned}$$

By definition of the Hellinger distance, $\int\sqrt{p(\omega)q(\omega)} = 1-H^2(P,Q)$. Moreover, for $a,b \geq 0$, we have $|1-\sqrt{a/b}| = |(b-a)/((\sqrt{a}+\sqrt{b})\sqrt{b})| \leq |(b-a)/(b\wedge a)|$. Combining these arguments gives

$$\big(1-H^2(P,Q)\big)\Big|E_P[X]-E_Q[X]\Big| \leq \Big(E_P\Big[\Big|X-E_P[X]\Big|\Big] \vee E_Q\Big[\Big|X-E_Q[X]\Big|\Big]\Big)\Big\|\frac{p-q}{p\wedge q}\Big\|_{L^\infty},$$

proving the first claim.

For the second claim, recall that the probability space $\Omega = \{\omega_j, j = 1,2,\ldots,M\}$ is assumed to be finite and denote by $p_j := P(\{\omega_j\})$, $q_j := Q(\{\omega_j\})$, $j = 1,2,\ldots,M$ the respective probability mass functions of P and Q. Define the random variable $X_j(\omega) := \mathbf{1}(\omega = \omega_j)$. Then, $E_P[X_j]-E_Q[X_j] = p_j - q_j$, $E_P[|X_j-E_P[X_j]|] = p_j(1-p_j)$, and $E_Q[|X_j-E_Q[X_j]|] = q_j(1-q_j)$. For $j^* \in \mathrm{argmax}_{\ell=1,\ldots,M}\,|p_\ell - q_\ell|/(p_\ell \vee q_\ell)$,

$$E_P[|X_{j^*}-E_P[X_{j^*}]|] \vee E_Q[|X_{j^*}-E_Q[X_{j^*}]|] \leq p_{j^*} \vee q_{j^*} = \frac{1}{\max_\ell|\frac{p_\ell-q_\ell}{p_\ell\vee q_\ell}|}\underbrace{|p_{j^*}-q_{j^*}|}_{=|E_P[X_{j^*}]-E_Q[X_{j^*}]|}. \quad ☐$$

For the application to statistics, the random variable X is an estimator. Thus, given X, a related question is to find a random variable X' with $E_P[X'] = E_P[X]$ and $E_Q[X'] = E_Q[X]$, but smaller mean absolute deviations $E_P[|X' - E_P[X']|] < E_P[|X - E_P[X]|]$ and $E_Q[|X' - E_Q[X']|] < E_Q[|X - E_Q[X]|]$. In particular, for the trade-off between bias and mean absolute deviation, it does not seem favorable that X attains large values, as this mainly increases the mean absolute deviation. If for a measurable set A, the conditional means are the same, that is, $E_P[X|X \in A] = E_Q[X|X \in A]$, then, $X' = X\mathbf{1}(X \in A^c) + E_P[X|X \in A]\mathbf{1}(X \in A)$ satisfies $E_P[X'] = E_P[X]$, $E_Q[X'] = E_Q[X]$,

$$\begin{aligned}
&E_P[|X' - E_P[X']|] \\
&= E_P[|X' - E_P[X]|] \\
&= E_P\Big[|X - E_P[X]| \,\Big|\, X \in A^c\Big] P(A^c) + E_P\Big[|E_P[X|X \in A] - E_P[X]| \,\Big|\, X \in A\Big] P(A) \\
&= E_P\Big[|X - E_P[X]| \,\Big|\, X \in A^c\Big] P(A^c) + \Big|E_P[X|X \in A] - E_P[X]\Big| P(A) \\
&= E_P\Big[|X - E_P[X]| \,\Big|\, X \in A^c\Big] P(A^c) + \Big|E_P\Big[X - E_P[X] \,\Big|\, X \in A\Big]\Big| P(A) \\
&\le E_P\Big[|X - E_P[X]| \,\Big|\, X \in A^c\Big] P(A^c) + E_P\Big[\Big|X - E_P[X]\Big| \,\Big|\, X \in A\Big] P(A) \\
&= E_P[|X - E_P[X]|],
\end{aligned}$$

and similarly $E_Q[|X' - E_Q[X']|] \le E_Q[|X - E_Q[X]|]$. The argument can be viewed as a variation of the convex loss version of the Rao-Blackwell theorem.

4. Application to 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

$$dY_x = f(x)\,dx + \frac{1}{\sqrt{n}}\,dW_x, \tag{8}$$

where W is an unobserved standard Brownian motion. The aim is to recover the unobserved, real-valued regression function $f \in L^2([0,1])$ from the data Y. Below, we study the bias-MAD trade-off for estimation of $f(x_0)$ with fixed $x_0 \in [0,1]$.

Concerning upper bounds for the MAD risk in this setting, optimal convergence rates are obtained in Tsybakov (1986) and the first order asymptotics of the mean absolute deviation risk for Lipschitz functions is derived in Fan and Hall (1994).

To obtain lower bounds for the bias-MAD trade-off, denote by P_f the data distribution of the Gaussian white noise model with regression function f. It is known that the Hellinger distance is

$$H^2(P_f, P_g) = 1 - \exp\Big(-\frac{n}{8}\|f - g\|_2^2\Big), \tag{9}$$

whenever $f, g \in L^2([0,1])$, see Derumigny and Schmidt-Hieber (2023) for a reference and a derivation. This means that the inequality (3) becomes

$$\frac{1}{5}\exp\Big(-\frac{n}{4}\|f - g\|_2^2\Big)|u - v| \le E_f\Big[\Big|X - u\Big|\Big] \vee E_g\Big[\Big|X - v\Big|\Big]. \tag{10}$$

As commonly done in nonparametric statistics, we impose an Hölder smoothness condition on the regression function f. Let $R > 0$, $\beta > 0$ and denote by $\lfloor\beta\rfloor$ the largest integer that is strictly smaller than β. On a domain $D \subseteq \mathbb{R}$, we define the β-Hölder norm by $\|f\|_{\mathscr{C}^\beta(D)} = \sum_{\ell\le\lfloor\beta\rfloor} \|f^{(\ell)}\|_{L^\infty(D)} + \sup_{x,y\in D, x\neq y} |f^{(\lfloor\beta\rfloor)}(x) - f^{(\lfloor\beta\rfloor)}(y)|/|x-y|^{\beta-\lfloor\beta\rfloor}$, with $L^\infty(D)$ the supremum norm on D and $f^{(\ell)}$ denoting the ℓ-th (strong) derivative of f for $\ell \le \lfloor\beta\rfloor$. 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 $\mathscr{C}^\beta(\mathbb{R}) := \{K : \mathbb{R} \to \mathbb{R} : \|K\|_{\mathscr{C}^\beta(\mathbb{R})} < \infty\}$.

Theorem 4.1. *Consider the Gaussian white noise model (8) 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*

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

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

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

Explicit expressions for c and N can be derived from the proof. If $\mathrm{Med}_f[\widehat{f}(x_0)]|]$ denotes the median of $\widehat{f}(x_0)$, then the same holds if $\mathrm{Bias}_f(\widehat{f}(x_0))$ and $E_f[|\widehat{f}(x_0) - E_f[\widehat{f}(x_0)]|]$ are replaced by $\mathrm{Med}_f[\widehat{f}(x_0)] - f(x_0)$ and $E_f[|\widehat{f}(x_0) - \mathrm{Med}_f[\widehat{f}(x_0)]|]$, respectively.

The result is considerably weaker than the earlier derived lower bounds for the bias–variance trade-off for pointwise estimation. This is due to the fact that (3) is less sharp. Nevertheless, the conclusion provides still more information than the minimax lower bound for the absolute value loss. To see this, observe that by the triangle inequality,

$$\sup_{f\in\mathscr{C}^\beta(R)} E_f[|\hat f(x_0)-E_f[\hat f(x_0)]|] \geq \sup_{f\in\mathscr{C}^\beta(R)} E_f[|\hat f(x_0)-f(x_0)|] - |\operatorname{Bias}_f(\hat f(x_0))|$$
$$\geq \sup_{f\in\mathscr{C}^\beta(R)} E_f[|\hat f(x_0)-f(x_0)|] - \sup_{f\in\mathscr{C}^\beta(R)} |\operatorname{Bias}_f(\hat f(x_0))|.$$

Thus, the conclusion of Theorem 4.1 can be deduced from the minimax lower bound $\sup_{f\in\mathscr{C}^\beta(R)} E_f[|\hat f(x_0)-f(x_0)|] \geq (K/n)^{\beta/(2\beta+1)}$, as long as $C < K$. Arguing via the minimax rate, nothing, however, can be said if $C > K$, that is, the bias is of the optimal order with a potentially large constant. Indeed, if we would change the role of the worst-case bias and the worst-case risk in the previous display, we get the lower bound

$$\sup_{f\in\mathscr{C}^\beta(R)} E_f[|\hat f(x_0)-E_f[\hat f(x_0)]|] \geq \sup_{f\in\mathscr{C}^\beta(R)} |\operatorname{Bias}_f(\hat f(x_0))| - \sup_{f\in\mathscr{C}^\beta(R)} E_f[|\hat f(x_0)-f(x_0)|].$$

Since $\hat f(x_0)$ is an arbitrary estimator, we cannot exclude the possibility that

$$\sup_{f\in\mathscr{C}^\beta(R)} \left|\operatorname{Bias}_f(\hat f(x_0))\right| \approx \sup_{f\in\mathscr{C}^\beta(R)} E_f\left[|\hat f(x_0)-f(x_0)|\right].$$

However, Theorem 4.1 shows that even in the case $C > K$, the worst-case variance cannot converge faster than $n^{-\beta/(2\beta+1)}$.

Proof of Theorem 4.1. For any function $K \in \mathscr{C}^\beta(\mathbb{R})$ satisfying $K(0)=1$ and $\|K\|_2 < +\infty$, define $V := R/\|K\|_{\mathscr{C}^\beta(\mathbb{R})}$, $r_n := (2/V)^{1/\beta}(C/n)^{1/(2\beta+1)}$, and

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

By Lemma B.1 in Derumigny and Schmidt-Hieber (2023), we have for $0 < h \leq 1$, $\|h^\beta K((\cdot - x_0)/h)\|_{\mathscr{C}^\beta(\mathbb{R})} \leq \|K\|_{\mathscr{C}^\beta(\mathbb{R})}$. Since $r_n \leq 1$ for all sufficiently large n, taking $h = r_n$, we find $\|f_\theta\|_{\mathscr{C}^\beta([0,1])} \leq |\theta| V \|K\|_{\mathscr{C}^\beta(\mathbb{R})} \leq R$ for all $\theta \in [-1,1]$. Thus, $\mathscr{F} \subseteq \mathscr{C}^\beta(R)$ whenever $r_n \leq 1$. We can now apply (10) to the random variable $\hat f(x_0)$ choosing $P = P_{f_{\pm 1}}$, $Q = P_0$ and centering $u = E_{f_{\pm 1}}[\hat f(x_0)]$, $v = E_0[\hat f(x_0)]$,

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

Using substitution and the definition $r_n = (2/V)^{1/\beta}(C/n)^{1/(2\beta+1)}$, we find

$$\|f_{\pm 1}\|_2^2 \leq V^2 r_n^{2\beta} \int_{\mathbb{R}} K^2\left(\frac{x-x_0}{r_n}\right) dx = V^2 r_n^{2\beta+1}\|K\|_2^2 = \frac{1}{n} 2^{2+1/\beta} V^{-1/\beta} C\|K\|_2^2$$

and so,

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

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

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

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

5. Further extensions of the bias–variance trade-off

A natural follow-up question is to wonder about other concepts to measure systematic and stochastic error of an estimator. This section is intended as an overview of related concepts.

A large chunk of literature on variations of the bias–variance trade-off is concerned with extensions to classification under 0-1 loss, see Kohavi and Wolpert (1996), Breiman (1996), Tibshirani (1996), James and Hastie (1997). These approaches have been compared in Rozmus (2007). Le Borgne (2005) proposes an extension to the multi-class setting. In a Bayesian framework, Wolpert (1997) argues that the bias–variance trade-off becomes a bias-covariance–covariance trade-off, where a covariance correction is added. For relational domains, Neville and Jensen (2007) 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 Heskes (1998). Somehow related, Wu and Vos (2012) introduces the Kullback–Leibler bias and the Kullback–Leibler variance, and shows, using information theory, that a similar decomposition is valid. Domingos (2000) propose generalized definitions of bias and variance for a general loss, but without showing a bias–variance decomposition. For several exponential families (Hansen and Heskes, 2000) shows that there exist a loss L such that a bias–variance decomposition of L is possible. James (2003) studied a bias–variance decomposition for arbitrary loss functions, comparing different ways of defining the bias and the variance in such cases.

Data availability

No data was used for the research described in the article.

Acknowledgments

The project has received funding from the Dutch Research Council (NWO) via the Vidi grant VI.Vidi.192.021.

References

Belkin, M., Hsu, D., Ma, S., Mandal, S., 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proc. Natl. Acad. Sci. 116 (32), 15849–15854. http://dx.doi.org/10.1073/pnas.1903070116, URL: https://www.pnas.org/content/116/32/15849.

Breiman, L., 1996. Bias, Variance, and Arcing Classifiers. Technical Report 460, Statistics Department, University of California, Berkeley, CA, USA.

Derumigny, A., Schmidt-Hieber, J., 2023. On lower bounds for the bias-variance trade-off. Ann. Statist. 51 (4), 1510–1533. http://dx.doi.org/10.1214/23-aos2279.

Domingos, P., 2000. 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. AAAI Press, pp. 564–569.

Fan, J., Hall, P., 1994. On curve estimation by minimizing mean absolute deviation and its implications. Ann. Statist. 22 (2), 867–885. http://dx.doi.org/10.1214/aos/1176325499.

Hall, P., Horowitz, J., 2013. A simple bootstrap method for constructing nonparametric confidence bands for functions. Ann. Statist. 41 (4), 1892–1921. http://dx.doi.org/10.1214/13-AOS1137.

Hansen, J.V., Heskes, T., 2000. 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, Vol. 2. ICPR-2000, IEEE, pp. 207–210.

Heskes, T., 1998. Bias/variance decompositions for likelihood-based estimators. Neural Comput. 10 (6), 1425–1433. http://dx.doi.org/10.1162/089976698300017232.

James, G., 2003. Variance and bias for general loss functions. Mach. Learn. 51, 115–135. http://dx.doi.org/10.1023/A:1022899518027.

James, G., Hastie, T., 1997. Generalizations of the Bias/variance Decomposition for Prediction Error. Tech. Rep., Dept. Statistics, Stanford Univ., Stanford, CA.

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

Le Borgne, Y., 2005. Bias-Variance Trade-Off Characterization in a Classification Problem: What Differences with Regression. Technical Report, Machine Learning Group, Univ. Libre de Bruxelles, Belgium.

Neal, B., 2019. On the bias-variance tradeoff: Textbooks need an update. http://dx.doi.org/10.48550/arXiv.1912.08286, arXiv e-prints arXiv:1912.08286.

Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S., Mitliagkas, I., 2018. A modern take on the bias-variance tradeoff in neural networks. arXiv e-prints, arXiv:1810.08591.

Neville, J., Jensen, D., 2007. Bias/variance analysis for relational domains. In: International Conference on Inductive Logic Programming. Springer, pp. 27–28.

Rozmus, D., 2007. Methods of classification error decompositions and their properties. Acta Univ. Lodziensis. Folia Oeconomica.

Tibshirani, R., 1996. Bias, Variance and Prediction Error for Classification Rules. Technical Report, Statistics Department, University of Toronto.

Tsybakov, A.B., 1986. Robust reconstruction of functions by a local approximation method. Probl. Pereda. Inf. 22 (2), 69–84.

Wolpert, D.H., 1997. On bias plus variance. Neural Comput. 9 (6), 1211–1243.

Wu, Q., Vos, P., 2012. Decomposition of Kullback-Leibler risk and unbiasedness for parameter-free estimators. J. Statist. Plann. Inference 142 (6), 1525–1536. http://dx.doi.org/10.1016/j.jspi.2012.01.002.