

On lower bounds for the bias-variance trade-off

Johannes Schmidt-Hieber

joint work with Alexis Derumigny

feel free to ask questions during the talk!

August 2023

On lower bounds for the bias-variance trade-off

Alexis Derumigny, Johannes Schmidt-Hieber

Author Affiliations +

Ann. Statist. 51(4): 1510-1533 (August 2023). DOI: 10.1214/23-AOS2279



Statistics & Probability Letters

Volume 213, October 2024, 110182



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

Alexis Derumigny^a, Johannes Schmidt-Hieber^b  

Information Geometry (2024) 7:253–282
<https://doi.org/10.1007/s41884-024-00135-2>

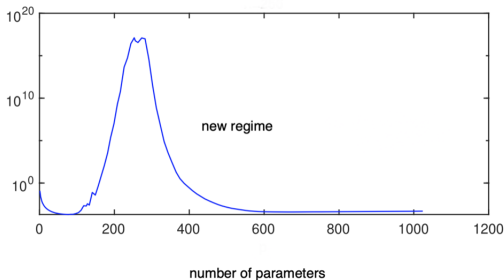
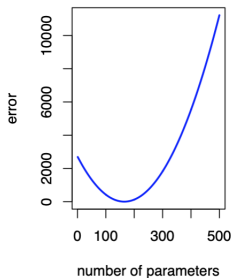
RESEARCH PAPER



Codivergences and information matrices

Alexis Derumigny¹  · Johannes Schmidt-Hieber²

double descent and implicit regularization



overparametrization generalizes well \rightsquigarrow implicit regularization

can we defy the bias-variance trade-off?

Geman et al. '92: "the fundamental limitations resulting from the bias-variance dilemma apply to all nonparametric inference methods, including neural networks"

Because of the double descent phenomenon, there is some doubt whether this statement is true. Recent work includes

Statistics > Machine Learning

[Submitted on 28 Dec 2018 (v1), last revised 10 Sep 2019 (this version, v2)]

Reconciling modern machine learning practice and the bias-variance trade-off

Mikhail Belkin, Daniel Hsu, Siyuan Ma, Soumik Mandal

Computer Science > Machine Learning

[Submitted on 19 Oct 2018 (v1), last revised 18 Dec 2019 (this version, v4)]

A Modern Take on the Bias-Variance Tradeoff in Neural Networks

Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, Ioannis Mitliagkas

lower bounds on the bias-variance trade-off

Similar to minimax lower bounds we want to establish a general mathematical framework to derive lower bounds on the bias-variance trade-off that hold for all estimators.

given such bounds we can answer many interesting questions

- are there methods (e.g. deep learning) that can defy the bias-variance trade-off?
- lower bounds for the U -shaped curve of the classical bias-variance trade-off

related literature

- Low '95 provides complete characterization of bias-variance trade-off for functionals in the Gaussian white noise model
- Pfanzagl '99 shows that estimators of functionals satisfying an asymptotic unbiasedness property must have unbounded variance

No general treatment of lower bounds for the bias-variance trade-off yet.

overview

- ① abstract lower bounds for bias-variance trade-off
- ② applications to standard nonparametric and high-dimensional models

Cramér-Rao inequality

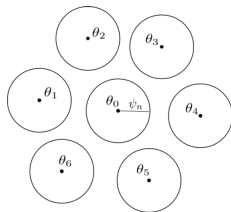
for parametric problems:

$$V(\theta) \geq \frac{(1 + B'(\theta))^2}{F(\theta)}$$

- $V(\theta)$ the variance
- $B'(\theta)$ the derivative of the bias
- $F(\theta)$ the Fisher information

abstract lower bounds on bias-variance trade-off

the well-known hypotheses testing reduction to prove lower bounds for the minimax risk relies on a careful selection of probability measures P_0, \dots, P_M together with bounds on information distances such as KL, Hellinger, \dots



- to prove lower bounds for the bias-variance trade-off, we want to follow a similar approach
- similar as for minimax lower bounds we need to distinguish the cases $M = 1$ (linear functionals) and $M > 1$

change of expectation inequalities

for two measures P, Q ($M = 1$), we derived inequalities for the most common information measures that link the change of expectation and the variance

Hellinger version: For all random variables X ,

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

with $H(P, Q)$ the Hellinger distance.

How can this be used to prove lower bounds for the bias-variance trade-off?

recovering the Cramér-Rao inequality

If $P = P_\theta$ and $Q = P_{\theta+\Delta}$, then, the inequality

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

becomes (under suitable regularity conditions) the Cramér-Rao inequality as $\Delta \downarrow 0$ since

$$\frac{H(P_\theta, P_{\theta+\Delta})^2}{\Delta^2} = \frac{1}{8} \text{Fisher info}(\theta) + o(\Delta).$$

why we need to generalize

- with two measures, we can only study perturbation in one direction
- not enough to derive rate optimal lower bounds for bias-variance trade-off
- we derived two change of expectation inequalities to deal with multiple probability measures

change of expectation for multiple measures

χ^2 -version:

- probability measures P_0, \dots, P_M
- $\chi^2(P_0, \dots, P_M)$ the matrix with entries

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

- any random variable X
- $\Delta := (E_{P_1}[X] - E_{P_0}[X], \dots, E_{P_M}[X] - E_{P_0}[X])^\top$

then,

$$\Delta^\top \chi^2(P_0, \dots, P_M)^{-1} \Delta \leq \text{Var}_{P_0}(X)$$

some examples for χ^2 -matrix

distribution	$\chi^2(P_0, \dots, P_M)_{j,k}$
$P_j = \mathcal{N}(\theta_j, \sigma^2 I_d),$ $\theta_j \in \mathbb{R}^d, I_d$ identity	$\exp\left(\frac{\langle \theta_j - \theta_0, \theta_k - \theta_0 \rangle}{\sigma^2}\right) - 1$
$P_j = \otimes_{\ell=1}^d \text{Pois}(\lambda_{j\ell}),$ $\lambda_{j\ell} > 0$	$\exp\left(\sum_{\ell=1}^d \frac{(\lambda_{j\ell} - \lambda_{0\ell})(\lambda_{k\ell} - \lambda_{0\ell})}{\lambda_{0\ell}}\right) - 1$
$P_j = \otimes_{\ell=1}^d \text{Exp}(\beta_{j\ell}),$ $\beta_{j\ell} > 0$	$\prod_{\ell=1}^d \frac{\beta_{j\ell} \beta_{k\ell}}{\beta_{0\ell}(\beta_{j\ell} + \beta_{k\ell} - \beta_{0\ell})} - 1$
$P_j = \otimes_{\ell=1}^d \text{Ber}(\theta_{j\ell}),$ $\theta_{j\ell} \in (0, 1)$	$\prod_{\ell=1}^d \left(\frac{(\theta_{j\ell} - \theta_{0\ell})(\theta_{k\ell} - \theta_{0\ell})}{\theta_{0\ell}(1 - \theta_{0\ell})} + 1\right) - 1$

χ^2 -matrix decodes a form of linear dependence of the measures
 $P_1 - P_0, \dots, P_M - P_0$

overview

- ① abstract lower bounds for bias-variance trade-off
- ② applications

pointwise estimation

Gaussian white noise model: We observe $(Y_x)_x$ with

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

- estimate $f(x_0)$ for a fixed x_0
- $\mathcal{C}^\beta(R)$ denotes ball of Hölder β -smooth functions
- for any estimator $\hat{f}(x_0)$, we obtain the **bias-variance lower bound**

$$\inf_{\hat{f}} \sup_{f \in \mathcal{C}^\beta(R)} |\text{Bias}_f(\hat{f}(x_0))|^{1/\beta} \sup_{f \in \mathcal{C}^\beta(R)} \text{Var}_f(\hat{f}(x_0)) \gtrsim \frac{1}{n}$$

- bound is attained by most estimators
- generates U -shaped curve

pointwise estimation (ctd.)

for any estimator satisfying

$$\sup_{f \in \mathcal{C}^\beta(R)} |\text{Bias}_f(\hat{f}(x_0))| \lesssim n^{-\beta/(2\beta+1)}$$

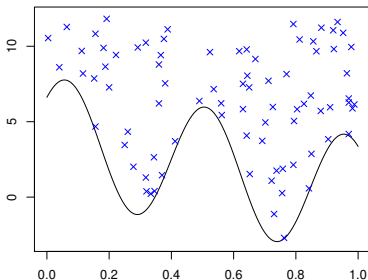
we also have that

$$\inf_{\hat{f}} \sup_{f \in \mathcal{C}^\beta(R)} |\text{Bias}_f(\hat{f}(x_0))|^{1/\beta} \inf_{f \in \mathcal{C}^\beta(R)} \text{Var}_f(\hat{f}(x_0)) \gtrsim \frac{1}{n}$$

- much stronger bias-variance trade-off
- sup over bias can never be replaced by inf!

pointwise estimation of the boundary

For an example of an irregular nonparametric problem consider support boundary recovery



\rightsquigarrow we observe Poisson point process with intensity n on the epigraph of an unknown function f and want to recover f

pointwise estimation of the boundary (ctd.)

the minimax estimation rate in this model is $n^{-2\beta/(\beta+1)}$

For any estimator \hat{f} with

$$\sup_{f \in \mathcal{C}^\beta(R)} \text{MSE}_f(\hat{f}(x_0)) \lesssim n^{-\frac{2\beta}{\beta+1}},$$

we have

$$\sup_{f \in \mathcal{C}^\beta(R)} \text{Bias}_f(\hat{f}(x_0))^2 \gtrsim n^{-\frac{2\beta}{\beta+1}}$$

and for all f in the interior of the Hölder ball

$$\text{Var}_f(\hat{f}(x_0)) \gtrsim n^{-\frac{2\beta}{\beta+1}}$$

Gaussian sequence model:

- observe independent $X_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, \dots, n$
- $\Theta(s)$ the space of s -sparse vectors (here: $s \leq \sqrt{n}/2$)
- bias-variance decomposition

$$E_{\theta}[\|\hat{\theta} - \theta\|^2] = \underbrace{\|E_{\theta}[\hat{\theta}] - \theta\|^2}_{B^2(\theta)} + \underbrace{\sum_{i=1}^n \text{Var}_{\theta}(\hat{\theta}_i)}_{V(\theta)}$$

- consider $s \ll \sqrt{n}$
- any estimator attaining the minimax rate $s \log n$ has

$$\sup_{\theta \in \Theta(s)} B^2(\theta) \asymp s \log n, \quad \sup_{\theta \in \Theta(s)} V(\theta) \geq \frac{s}{2}.$$

- lower bounds can be attained, **no strict bias-variance trade-off**

sequence model

- consider again $s \ll \sqrt{n}$
- **bias-variance lower bound:** if $B^2(\theta) \leq \gamma s \log n$, then,

$$\sum_{i=1}^n \text{Var}_0(\hat{\theta}_i) \gtrsim n \left(\frac{s^2}{n} \right)^{4\gamma}$$

- bound is matched (up to a factor in the exponent) by soft thresholding
- bias-variance trade-off more extreme than U -shape

sequence model

- estimation of quadratic functional $\theta \mapsto \|\theta\|_2^2$
- parameter space

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

sequence model

- if $s \gg \sqrt{n}$, then, unbiased minimax rate optimal estimators exist
- consider now $s \ll \sqrt{n}$
- minimax rate is $s^2 \log^2(n/s^2)$, see Collier et al. '17
- any estimator $\widehat{\|\theta\|_2^2}$ attaining the minimax optimal estimation rate must satisfy

$$\sup_{\theta \in \Theta_n^2(s)} B^2(\theta) \asymp s^2 \log^2\left(\frac{n}{s^2}\right)$$

- there exists a minimax rate optimal estimator $\widehat{\|\theta\|_2^2}$ with $\sup_{\theta \in \Theta_n^2(s)} \text{Var}_{\theta}(\widehat{\|\theta\|_2^2}) \lesssim s \log(n/s^2)$.

large gap between bias and variance

Gaussian white noise model: We observe $(Y_x)_x$ with

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

- bias-variance decomposition

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

- is there a bias-variance trade-off between $\text{IBias}_f^2(\hat{f})$ and $\text{IVar}_f(\hat{f})$?
- turns out to be a very hard problem

reductions for L^2 -loss

- direct application of abstract lower bound does not seem to work
- we propose a two-fold reduction scheme
 - reduction to a simpler model
 - reduction to a smaller class of estimators

first reduction for L^2 -loss

Consider the Gaussian sequence model

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

- $S^\beta(R)$ Sobolev space of β -smooth functions
- $\Theta_m^\beta(R) := \{\theta : \|\theta\|_2 \leq R/(\Gamma_\beta m^\beta)\}$, Γ_β a suitable constant

Reduction: For any estimator \hat{f} , there exists estimator $\hat{\theta}$, s.t.

$$\sup_{\theta \in \Theta_m^\beta(R)} \|E_\theta[\hat{\theta}] - \theta\|_2^2 \leq \sup_{f \in S^\beta(R)} \text{IBias}_f^2(\hat{f}),$$

and

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

second reduction for L^2 -loss

- a function is spherically symmetric if for any x and any orthogonal matrix D , $f(x) = f(Dx)$.
- estimator $\hat{\theta} = \hat{\theta}(X)$ is spherically symmetric if $X \mapsto \hat{\theta}(X)$ is spherically symmetric

in the second reduction step we show that the best bias-variance trade-off is attained for spherically symmetric estimators

- argument in Stein '56 shows that $\hat{\theta}$ is of the form $r(\|X\|_2)X$

L^2 -loss (ctd.)

Bias-variance lower bound: For any estimator \hat{f} ,

$$\inf_{\hat{f}} \sup_{f \in S^\beta(R)} |\text{IBias}_f(\hat{f})|^{1/\beta} \sup_{f \in S^\beta(R)} \text{IVar}_f(\hat{f}) \geq \frac{1}{8n},$$

- many estimators \hat{f} can be found with upper bound $\lesssim 1/n$

mean absolute deviation

- several extensions of the bias-variance trade-off have been proposed in the literature, e.g. for classification
- the mean absolute deviation (MAD) of an estimator $\hat{\theta}$ is

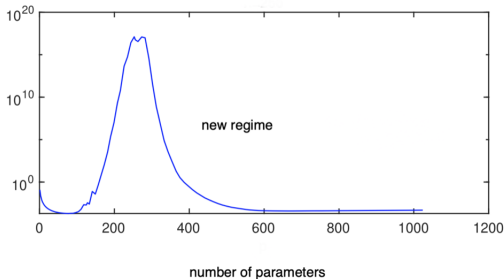
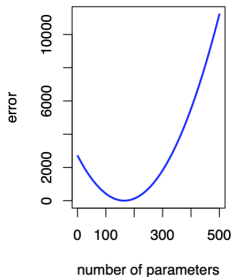
$$E_{\theta}[|\hat{\theta} - m|]$$

with m either the mean or the median of $\hat{\theta}$

can the general framework be extended to lower bounds on the trade-off between bias and MAD?

- derived change of expectation inequality
- this can be used to obtain a partial answer for pointwise estimation in the Gaussian white noise model

double descent and implicit regularization



overparametrization generalizes well \rightsquigarrow implicit regularization

summary

- general framework to derive bias-variance lower bounds
- leads to matching bias-variance lower bounds for standard models in nonparametric and high-dimensional statistics
- different types of the bias-variance trade-off occur
- can machine learning methods defy the bias-variance trade-off? **No, there are universal lower bounds that no method can avoid**