Overfitting Behaviour of Gaussian Kernel Ridgeless Regression: Varying Bandwidth or Dimensionality

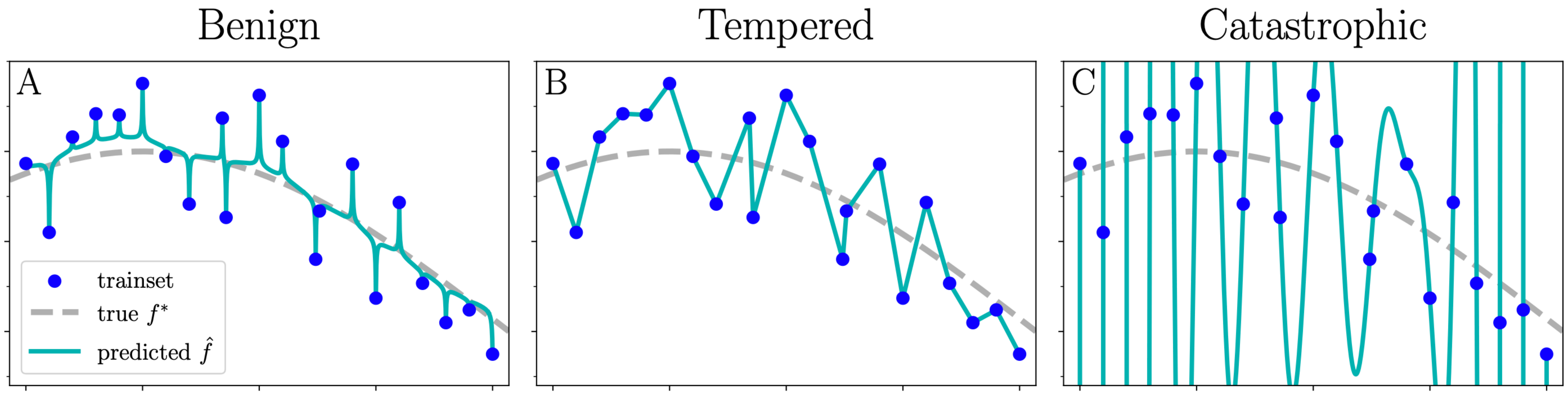
Marko Medvedev [1], Gal Vardi [2], Nathan Srebro [3]

[1]The University of Chicago [2]Weizmann Institute of Science [3]TTIC

NeurIPS 2024

Introduction

We study the overfitting behavior of Kernel Ridge(less) Regression (KRR) with Gaussian Kernel: the behavior of the **limiting test error** when training on noisy data as the number of samples tends to infinity by insisting on interpolation



(Simon et al. 2021) Illustration for three types of overfitting.

Introduction

- When the input dimension and bandwidth are fixed, the overfitting behavior is known to be "catastrophic"
  - This is not how Gaussian KRR is typically used in practice
  - In fixed dimension, the bandwidth τ_m is tuned, that is decreased, when sample size increases
- We also study the behavior when input dimension increases with sample size
  - Previous studies considered polynomial increasing dimension (i.e. dimension ∝ sample size^a, for $0 \leq a \leq 1$) but not subpolynomial scaling

Contribution

- We provide a more comprehensive picture of overfitting with Gaussian KRR by studying **the overfitting behavior with varying bandwidth or arbitrarily varying dimension**
    - For fixed dimension, we show that even with varying bandwidth, the interpolation learning is never consistent and generally not better than the null predictor
    - For increasing dimension, we show the first example of subpolynomially scaling dimension that achieves benign overfitting for (Gaussian) KRR.
    - Additionally, we show that KRR with a class of dot-product kernels on the sphere (including the Gaussian kernel) is inconsistent when the dimension scales logarithmically with sample size.

Setup

Let $\mathscr{D}$ be an unknown distribution over $\mathscr{X} \times \mathscr{Y} \subset \mathbb{R}^d \times \mathbb{R}$ and let $\{(x_i, y_i)\}_{i=1}^m \sim \mathscr{D}^m$ be a dataset consisting of m samples. We want to understand the limiting behavior $\lim_{m\to\infty} R(\hat{f}_0)$ of the test error $R(f) = \mathbb{E}_{\mathscr{D}}\left(f - f^*\right)^2$ of the minimum norm interpolating solution $\hat{f}_0 =$ argmin$_{\hat{R}(f)=0; f\in\mathscr{H}_K} \|f\|_K^2$. We will focus on the Gaussian kernel $K(x,t) = \exp\left(-\frac{\|x-t\|^2}{\tau_m^2}\right)$. We use taxonomy of benign, tempered, and catastrophic overfitting from (Mallinar et al. 2022), which indicates whether $\lim_{m\to\infty} R(\hat{f}_0)$ is the Bayes (optimal) error, a non-optimal but constant error, or infinity.

Assumption (Gaussian design ansatz)

When sampling $(x, \cdot\,) \sim \mathcal{D}$, we have that the Gaussian universality holds for the **eigenfunctions** ϕ in the sense that the expected risk is unchanged if we replace ϕ with $\tilde{\phi}$, where $\tilde{\phi}$ is Gaussian with appropriate parameters, i.e. $\tilde{\phi} \sim \mathcal{N}(0,\text{diag}\{\lambda_i\})$.

Under this assumption, the eigenframework gives a closed form of the test risk in terms of kernel eigenstructure.

Given a positive semi-definite kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we can decompose it as $K(x_1, x_2) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x_1) \phi_k(x_2)$, where λ_k and ϕ_k are the eigenvalues and eigenfunctions of the integral operator associated to K.

Closed form of the test risk

We can write the target function in the basis of $\{\phi_k\}$ from $K(x_1, x_2) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x_1)\phi_k(x_2)$, $f^*(x) = \sum_{i=1}^{\infty} \beta_i \phi_i(x)$. Let κ_δ be the *effective regularization*, i.e. the solution to $\sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \kappa_\delta} + \frac{\delta}{\kappa_\delta} = m$, and let $\mathscr{L}_{i,\delta} = \frac{\lambda_i}{\lambda_i + \kappa_\delta}$ and $\mathscr{E}_\delta = \frac{m}{m - \sum_{i=1}^{\infty} \mathscr{L}_{i,\delta}^2}$. Then the *predicted risk* of $\hat{f}_0$ is given by

$$\tilde{R}(\hat{f}_0) = \mathscr{E}_0 \left(\sum_{i=1}^{\infty} \left(1 - \mathscr{L}_{i,0}\right)^2 \beta_i^2 + \sigma^2 \right)$$

Closed form of the test risk

$\sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \kappa_\delta} + \frac{\delta}{\kappa_\delta} = m$, $\mathscr{L}_{i,\delta} = \frac{\lambda_i}{\lambda_i + \kappa_\delta}$ and $\mathscr{E}_\delta = \frac{m}{m - \sum_{i=1}^{\infty} \mathscr{L}_{i,\delta}^2}$. Then the *predicted risk* of $\hat{f}_\delta$ is given

$$\tilde{R}(\hat{f}_\delta) = \mathscr{E}_\delta \left(\sum_{i=1}^{\infty} \left(1 - \mathscr{L}_{i,\delta}\right)^2 \beta_i^2 + \sigma^2 \right)$$

Formally we will prove results about $\tilde{R}(\hat{f}_\delta)$ but as ample empirical evidence suggests, treating $\tilde{R}(\hat{f}_\delta) \approx R(\hat{f}_\delta)$ is sufficient for understanding the behavior of KRR.

Fixed dimension: Gaussian Kernel with varying bandwidth

We will assume that the source distribution is uniform on a d dimensional sphere, i.e. $x \sim$ Unif$(\mathbb{S}^{d-1})$. We also assume that the marginal $\mathscr{Y}$ distribution is given by a target function $f^* \in L_{\mathscr{D}_\mathscr{X}}(\mathbb{S}^{d-1})$ and noise ξ with mean zero and variance $\sigma^2 > 0$ as $y \sim f^*(x) + \xi$.

We show that based on how the bandwidth τ_m changes, the minimum norm interpolating solution $\hat{f}_0$ exhibits either tempered or catastrophic overfitting, and we argue that it is almost always worse than the null predictor.

Theorem (Overfitting behavior of Gaussian KRR in fixed dimension)

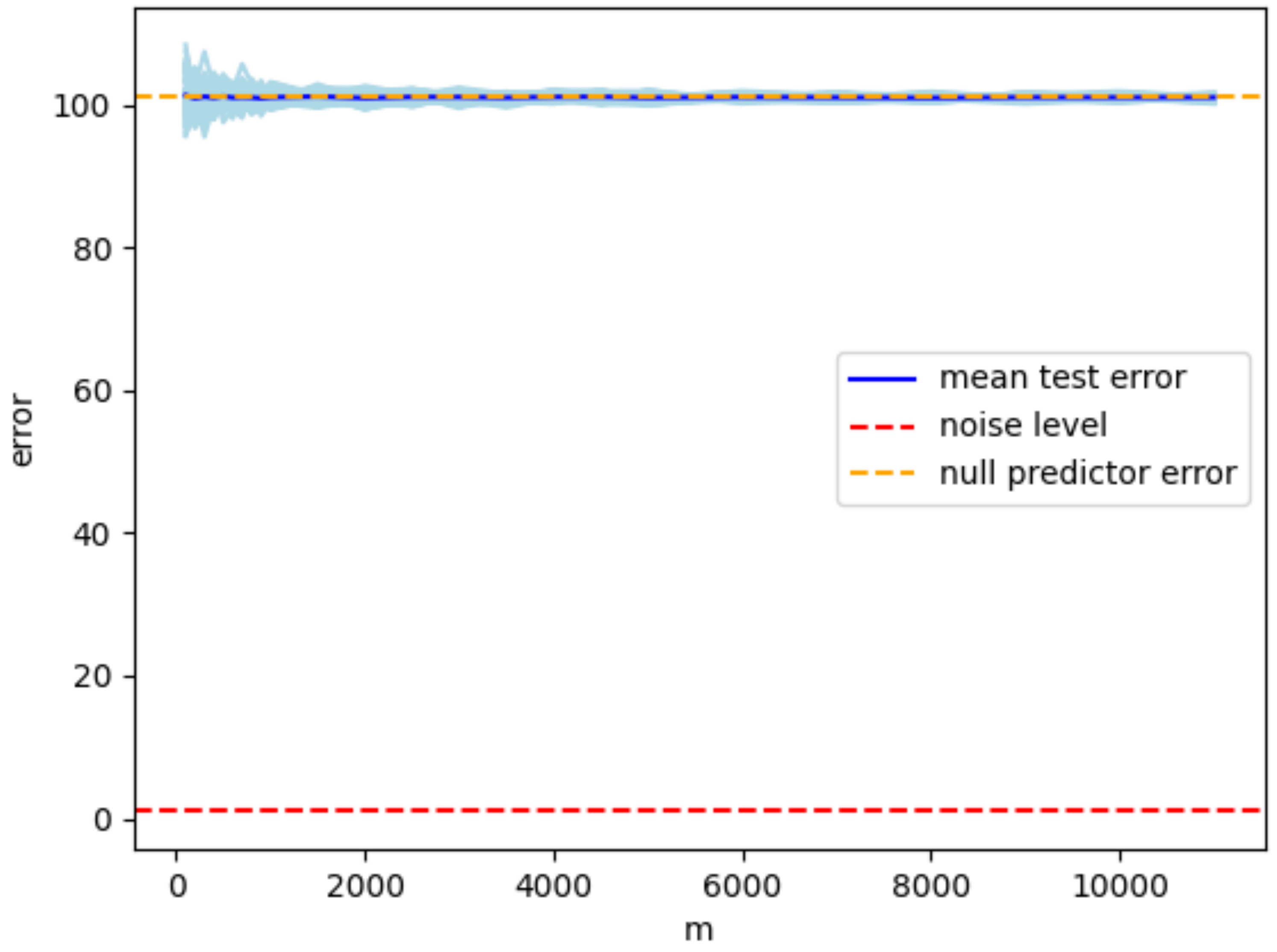
The following bounds hold for the predicted risk $\tilde{R}(\hat{f}_0)$ of the minimum norm interpolating solution of Gaussian KRR:

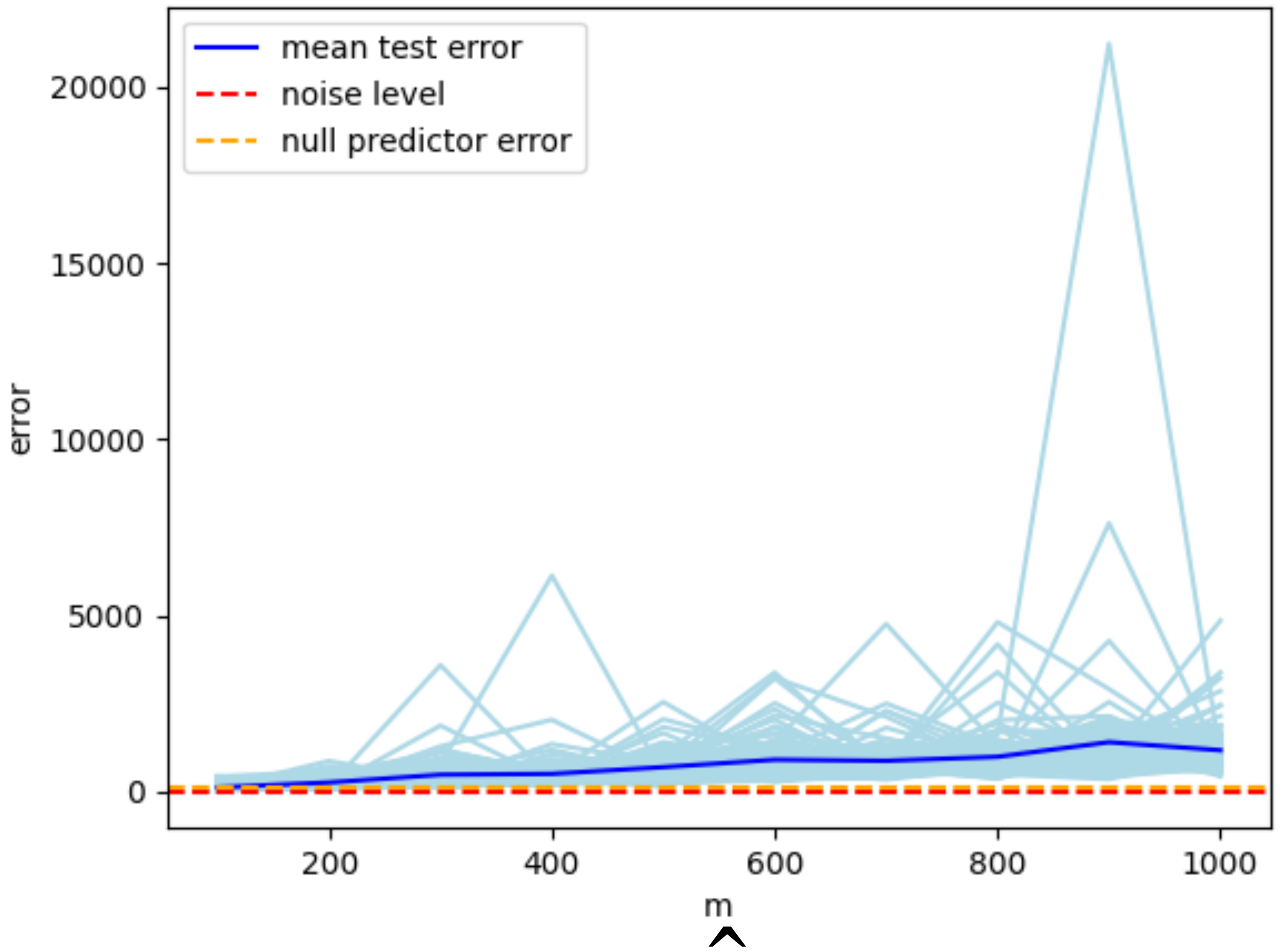
1. If $\tau_m = o(m^{-\frac{1}{d-1}})$, then $\tilde{R}(0) \leq \liminf_{m\to\infty} \tilde{R}(\hat{f}_0) \leq \limsup_{m\to\infty} \tilde{R}(\hat{f}_0) < \infty$. More precisely, if $\tau_m \leq m^{-\frac{1}{d-1}} t(m)$ where $t(m) \to 0$ as $m \to \infty$, then there is a scalar c_d that depends only on the dimension and m_0 that depends on $t(m)$ such that for all $m > m_0$ we have $\tilde{R}(\hat{f}_0) > \sigma^2 + (1 - c_d t(m)^{\frac{d-1}{2}})\|f^*\|^2$.

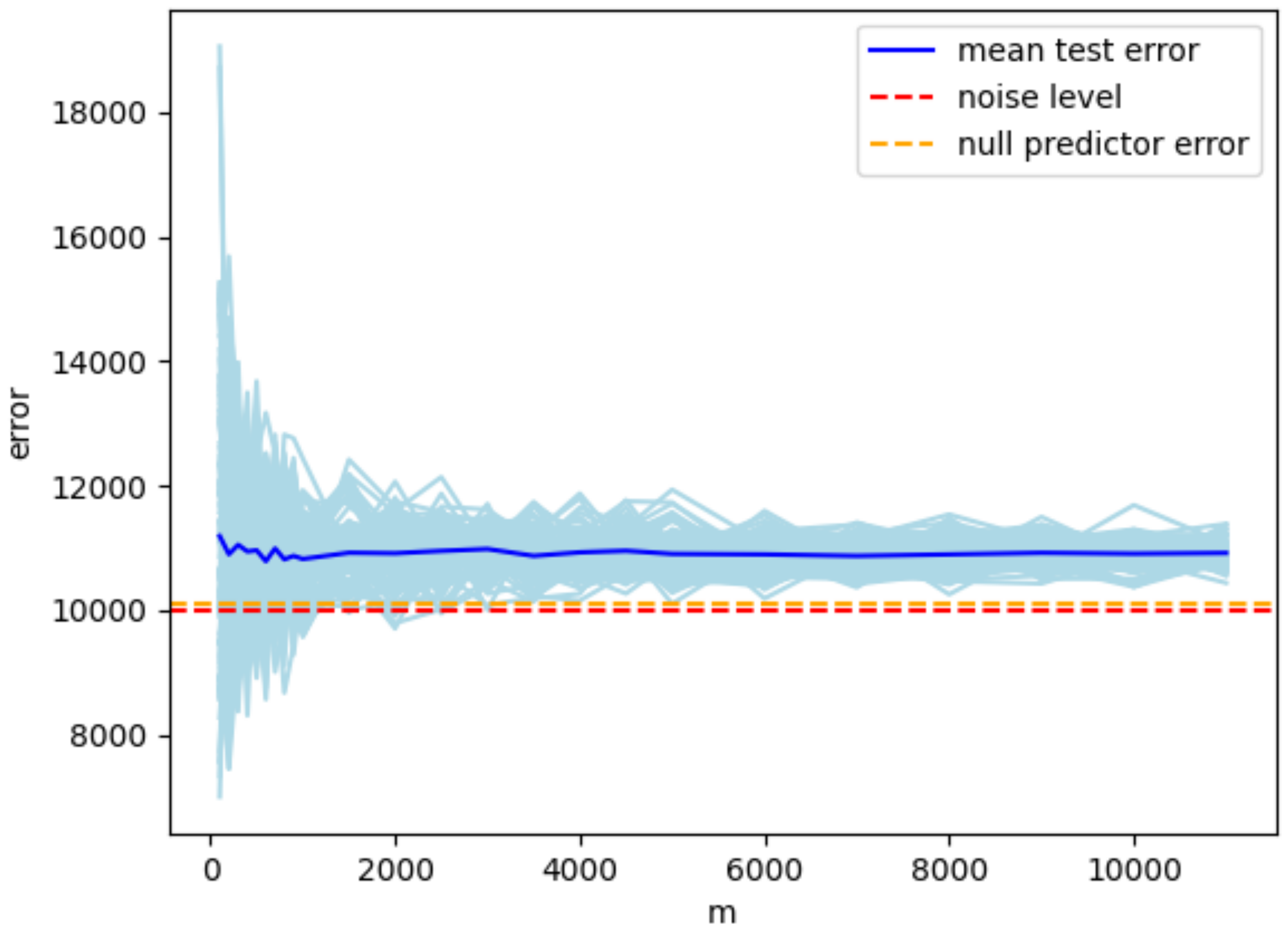
2. If $\tau_m = \omega(m^{-\frac{1}{d-1}})$, then $\lim_{m\to\infty} \tilde{R}(\hat{f}_0) = \infty$, so for large m we have $\tilde{R}(\hat{f}_0) > \tilde{R}(0)$.

3. If $\tau_m = \Theta(m^{-\frac{1}{d-1}})$, then $\limsup_{m\to\infty} R(\hat{f}_0) < \infty$. Moreover, suppose that $C_1 m^{-\frac{1}{d-1}} \leq \tau_m \leq C_2 m^{-\frac{1}{d-1}}$ for some constants C_1 and C_2, then there exist $\eta, \mu > 0$ that depend only on d, C_1, and C_2, such that for all m we have $\tilde{R}(\hat{f}_0) > \mu\|f^*\|^2 + (1+\eta)\sigma^2$. Consequently, $\tilde{R}(\hat{f}_0) > \tilde{R}(0)$ as long as $\sigma^2 > \dfrac{1-\mu}{\eta}\|f^*\|^2$.

Empirical validation (Overfitting behavior of Gaussian KRR in fixed dimension)







We plot the dependence of the test error $R(\hat{f}_0)$ on the sample size m for Gaussian

KRR with $x \sim$ Unif($\mathbb{S}^{d-1}$), $f^* = 10$, dimensions $d = 6{,}4{,}6$, and noise level $\sigma^2 = 1{,}10{,}1000$

respectively. We compare mean test error (blue) with noise level (red) and null predictor error (yellow). We also plot test errors for each of the runs (light blue).

Increasing dimension

Consider learning a sequence of distributions $\mathscr{D}^{(d)}$ over $\mathscr{X} \times \mathscr{Y} = \mathbb{R}^d \times \mathbb{R}$ given by $y \sim f_d^*(x) + \xi_d$ using a sequence of kernels $K^{(d)}$ where ξ_d is an independent noise with mean 0 and variance $\sigma^2 > 0$. We assume that the projections of f_d^* onto the eigenfunction of $\phi_k^{(d)}$of the kernels $K^{(d)}$ are uniformly bounded.

We show a generic upper and lower bound on the test risk of KRR in increasing dimension, for any scaling of the dimension and sample size. We use a few assumptions:

- The sum of eigenvalues is bounded as $\sum_{i=1}^{\infty} \lambda_i \leq A$.
- The eigenvalues are not too small, or
- The eigenvalues don't decay too quickly

These hold for the Gaussian kernel and other dot-product kernels on the sphere.

Theorem (Upper bound for increasing dimension)

Let $N(k)$ be the multiplicity of k-th eigenvalue corresponding to K and let $N_k = N(1) + \ldots + N(k)$. Let $k_m = \max\{k \in \mathbb{N} \,|\, N_k < m\}$, $L_m = N_{k_m}$, and $U_m = N_{k_m+1}$.

Assume that the target function has at most S_d nonzero coefficients $f_d^* = \sum_{i=1}^{S_d} \beta_i^{(d)} \phi_i^{(d)}$ with $||\beta||_\infty \leq B$ and $S_d \leq N_l$ for some $l \in \mathbb{N}$. Then, if $\tilde{\lambda}_k$ is the k-th unique eigenvalue and m and d are any sample size and dimension, the predicted test risk of minimum norm interpolating solution satisfies

$$\tilde{R}(\hat{f}_0) \leq \left(1 - \frac{L_m}{m}\right)^{-1} \left(1 - \frac{m}{U_m}\right)^{-1} \sigma^2 + B^2 \left(1 - \frac{L_m}{m}\right)^{-1} \left(1 - \frac{m}{U_m}\right)^{-1} \frac{A^2}{m^2} \left(\sum_{i=1}^{l} N(i) \frac{1}{\tilde{\lambda}_i^2} \right)$$

Theorem (Lower bound for increasing dimension)

If additionally the eigenvalues of $K^{(d)}$ are not too small, in the sense that there is a constant $b > 0$ such $\max_{i \leq k_m} \left(\frac{1}{\tilde{\lambda}_i} \right) < \frac{m - L_m}{b}$, then for the predicted test risk of KRR, it holds

$$\tilde{R}(\hat{f}_0) > \left(1 - \left(\frac{b}{b+1} \right)^2 \frac{L_m}{m} \right)^{-1} \sigma^2.$$

Note that these conditions hold for Gaussian kernel and dot-product kernels on the sphere.

Corollary (Inconsistency with dot-product kernel in logarithmically scaling dimension)

Let $K^{(d)}$ be a sequence of dot-product kernels on $\mathbb{S}^{d-1}$ that satisfy $\max_{i \le k_m} \left(\frac{1}{\tilde{\lambda}_i} \right) < \frac{m - L_m}{b}$ for some $b > 0$. Let the dimension d grow logarithmically in sample size m, $d = \log_2(m)$ (i.e. $m = 2^d$). Then, then the minimum norm interpolant cannot exhibit benign overfitting, i.e. there exist an absolute constant $\eta > 0$ such that for all m, d

$$\tilde{R}(\hat{f}_0) > (1 + \eta)\sigma^2.$$

Corollary (Benign overfitting with Gaussian kernel and subpolynomial dimension)

Let K be the Gaussian kernel on the sphere $\mathbb{S}^{d-1}$ with a fixed bandwidth, and take a sequence of dimensions d and sample sizes m that scale as $d = \exp\left(\sqrt{\log m}\right)$ (in particular, we take $l \in \mathbb{N}$ such that $d = 2^{2^l}$ and $m = 2^{2^{2l}}$ with $l = 1,2,3\ldots$). Consider learning a sequence of target functions f_d^* that have uniformly bounded projections to each eigendirection with at $S_d \leq m^{\frac{1}{4}}$ of them nonzero. Then, we have that the minimum norm interpolating solution achieves the Bayes error in the limit $(m, d) \to \infty$. In particular, for $d \geq 4$ and $m \geq 16$ we have

$$\tilde{R}(\hat{f}_0) \leq \left(1 - \frac{1}{\log m}\right)^{-1} \left(1 - \exp\left(-0.89\sqrt{\log m}\right)\right)^{-1} \sigma^2 + 2B^2 \frac{1}{m}.$$

This establishes the first case of sub-polynomially scaling dimension with benign overfitting using the Gaussian kernel.

Summary

We studied the overfitting behavior of Gaussian KRR with varying bandwidth or dimension.

- For fixed dimension, we show that even with varying bandwidth, the interpolation learning is never consistent and generally not better than the null predictor

- For increasing dimension, we show the first example of subpolynomially scaling dimension that achieves benign overfitting for (Gaussian) KRR.

- Additionally, we show that KRR with a class of dot-product kernels on the sphere (including the Gaussian kernel) is inconsistent when the dimension scales logarithmically with sample size.