Interpolating Classifiers Make Few Mistakes

Tengyuan Liang*[1] and Benjamin Recht†[2]

[1]University of Chicago
[2]University of California, Berkeley

January 22, 2021

Abstract

This paper provides elementary analyses of the regret and generalization of minimum-norm interpolating classifiers (MNIC). The MNIC is the function of smallest Reproducing Kernel Hilbert Space norm that perfectly interpolates a label pattern on a finite data set. We derive a mistake bound for MNIC and a regularized variant that holds for all data sets. This bound follows from elementary properties of matrix inverses. Under the assumption that the data is independently and identically distributed, the mistake bound implies that MNIC generalizes at a rate proportional to the norm of the interpolating solution and inversely proportional to the number of data points. This rate matches similar rates derived for margin classifiers and perceptrons. We derive several plausible generative models where the norm of the interpolating classifier is bounded or grows at a rate sublinear in n. We also show that as long as the population class conditional distributions are sufficiently separable in total variation, then MNIC generalizes with a fast rate.

Keywords— Generalization, Regret, Mistake Bound, Interpolation, Least-squares classification.

1 Introduction

Using a squared loss for classification problems remains a controversial topic in machine learning. Though popular for regression, the squared loss makes little intuitive or semantic sense for classification problems. A squared loss appears to penalize models for being "too correct." Moreover, the labels in classification are arbitrary, so the minimum square error doesn't convey much information about the latent regression problem. Similarly, minimum-norm interpolation, the limiting solution of ridge regression as the regularization parameter tends to zero, appears to be a curious choice for a classifier. Forcing a function to be equal to an arbitrary label set could be unnecessarily aggressive if we only aim to have our classifier predict the correct sign on our data.

Despite these reservations, minimum-norm interpolation classifiers (MNIC) and regularized least-squares classification (RLSC) work exceptionally well in practice. Recent investigations by Shankar et al. [2020] demonstrated that there was no advantage to using more sophisticated classification techniques on popular contemporary data sets. From a theoretical perspective, stochastic gradient descent on a variety of empirical risk minimization problems will converge to the MNIC solution under mild assumptions. Jacot

*tengyuan.liang@chicagobooth.edu.
†brecht@berkeley.edu.

et al. [2018] and Heckel and Soltanolkotabi [2020] have shown that neural nets trained with stochastic gradient descent approximate the minimum-norm solutions of appropriate Reproducing Kernel Hilbert Spaces.

On top of these connections and empirical results, least-squares methods have many attractive features that argue in their favor. Their solutions can be written out algebraically, and we can lean on powerful linear algebra tools to compute them at large scale (see Avron et al. [2017], Ma and Belkin [2017], Rudi et al. [2017], Wang et al. [2019], Shankar et al. [2021], for example). Many auxiliary quantities such as the leave-one-out error can be also be computed in closed form. If these methods are at all competitive, the machine learning community should encourage their use.

In this paper we provide an elementary analysis of MNIC, partially helping to explain the method's success. We first present a mistake bound for minimum-norm interpolation classification that holds unconditionally of how the data was generated. The proof follows immediately from two applications of the matrix inversion lemma. Moreover, we present two simple geometric and algorithmic proofs that shed further insights into the properties of MNIC and RLSC and the structure of QR decompositions.

Adding the assumption that the data is generated i.i.d., our mistake bound and an application of Markov's inequality implies that MNIC generalizes with a rate of B_n^2/n where B_n denotes the expected norm of the MNIC when interpolating n i.i.d. sampled data points. This compares favorably with the generalization bounds of Vapnik and Chervonenkis for the perceptron which show the generalization error scales as $\frac{1}{M_{n+1}^2 n}$ where M_{n+1} denotes the expected size of the margin for a sample of $n+1$ data points Vapnik and Chervonenkis [1974]. Since for the maximum margin solution, the margin is the inverse of the norm of the separating hyperplane, our bound tracks a similar scaling as these classic margin-based results.

A key feature of MNIC and RLSC is that the expected norm can be estimated for a variety of probabilistic models of data. By leveraging concentration inequalities and random matrix tools, we can analyze a variety of plausible data generation schemes and show that as long as there is certain separation between the distributions of the two classes, then we can expect the interpolant norm to not grow too quickly.

2 Related Work

Using least-squares loss for classification has a long history in machine learning and statistics. Its modern popularization was by Suykens and Vandewalle [1999] and connections to general kernel machines were studied by Rifkin et al. [2003]. While many authors use least-squares classification as their default tool (for example, Rudi et al. [2017], Shankar et al. [2020]), it is still not common practice to use a squared loss for classification. Theory bounds for regression with bounded labels do apply to least-squares classification, and such results exist. For example, Lee et al. [1998] showed agnostic learning is possible with a squared loss. The results in this work focus on a particular class of interpolators that are amenable to a simpler analysis.

While we focus on interpolating classifiers, our inspiration comes from work on maximum margin classifiers. Novikoff [1962] famously derived a mistake bound for the perceptron, and Vapnik and Chervonenkis [1974] used this bound and a leave-one-out argument to turn the perceptron mistake bound into a generalization bound. Vapnik and Chapelle [2000] prove a similar generalization bound for the maximum-margin classifier using a similar argument, though their proof requires a strong assumption about support vectors not changing when data is resampled. A more general theory of generalization for margin classifiers using Rademacher complexity was developed by Koltchinskii and Panchenko [2002]. These bounds yielded suboptimal rates in the case when the data was separable, and Srebro et al. [2010] provided an argument to yield fast rates. All margin bounds require the size of the margin to not shrink too quickly as the number of data points increases.

Many recent papers have used generative models to understand how such margins scale (see, for example, Deng et al. [2019], Montanari et al. [2019], Liang and Sur [2020], Chatterji and Long [2020]). In this work, we study scaling on the norm of the interpolating function using similar ideas. Leveraging recent developments in non-linear random matrices, Liang and Rakhlin [2020] and later in Liang et al. [2020]

studied the generalization of kernel ridgeless regression. Bartlett et al. [2020] studied the generalization of minimum-norm interpolants in the context of infinite-dimensional Gaussian process regression. Hastie et al. [2019] applied precise asymptotic tools to analyze random non-linear feature regression. Though our generalization results can be extended to the case of regression, we focus here on generative models of classification, and show these generalize with mild separation assumptions on the data generating process.

3 A Mistake Bound for MNIC

Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space (RKHS) with embedding functions φ_x so that $k(x,z) = \langle \varphi_x, \varphi_z \rangle_K$. Suppose that we are given a data set $S_n = \{(x_i, y_i)\}_{i=1}^n$ and consider the optimization problem

$$\begin{array}{ll} \text{minimize} & \|f\|_K \\ \text{subject to} & f(x_i) = y_i, \quad i = 1, \ldots, n. \end{array} \tag{3.1}$$

Let $\widehat{f_{S_n}}$ denote the optimal solution of this infinite-dimensional optimization problem.

The key to our analysis will be controlling the error of $\widehat{f_{S_i}}$ on the next data point in the sequence, x_{i+1}. To do so, we use a simple identity that follows from linear algebra. First, let's fix some notation. Given n data points, let

$$s_i := \operatorname{dist}\left(\operatorname{span}(\varphi_{x_1}, \ldots, \varphi_{x_{i-1}}), \varphi_{x_i}\right).$$

and

$$s_{S_n \setminus i} := \operatorname{dist}\left(\operatorname{span}(\varphi_{x_1}, \ldots, \varphi_{x_{i-1}}, \varphi_{x_{i+1}}, \ldots, \varphi_{x_n}), \varphi_{x_i}\right).$$

Here the distance measures the RKHS norm of the embedding function. Let S_n denote a set of n data points, and let S_i denote the first i points in the sequence. Let $\widehat{f_{S_i}}$ denote the norm of the optimal solution to Problem (3.1) applied to S_i. Then we have the following

Lemma 3.1. *For any $i \in \mathbb{N}_+$, and any data sequence $\{(x_i, y_i)\}_{i \in \mathbb{N}_+}$, $\widehat{f}$ in Problem (3.1) satisfies*

$$\left(y_i - \widehat{f_{S_{i-1}}}(x_i)\right)^2 = s_i^2 \left(\|\widehat{f_{S_i}}\|_K^2 - \|\widehat{f_{S_{i-1}}}\|_K^2\right).$$

We hold off on proving this Lemma until the next section where we give several proofs that highlight multiple aspects of minimum norm interpolation. With the Lemma in hand, however, we can now state and prove our two main results. Our first result is a deterministic mistake bound that makes no assumptions about the data.

Theorem 3.2. *Let $R_n^2 = \max_{1 \le i \le n} \|\varphi_{x_i}\|_K^2$ and $r_n^2 = \min_{1 \le i \le n} s_{S_n \setminus i}^2$. Then we have the following regret bound*

$$r_n^2 \cdot \frac{\|\widehat{f_{S_n}}\|_K^2}{n} \le \frac{1}{n} \sum_{i=1}^n \left(y_i - \widehat{f_{S_{i-1}}}(x_i)\right)^2 \le R_n^2 \cdot \frac{\|\widehat{f_{S_n}}\|_K^2}{n} \quad , \tag{3.2}$$

In the binary classification setting with $\mathcal{Y} = \{\pm 1\}$, we further have

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}\left\{y_i \widehat{f_{S_{i-1}}}(x_i) \le 0\right\} \le R_n^2 \cdot \frac{\|\widehat{f_{S_n}}\|_K^2}{n} \quad . \tag{3.3}$$

We emphasize that R_n^2 is uniformly bounded over n as long as $\sup_{x \in \mathcal{X}} K(x,x) < \infty$, since $R_n^2 = \max_i K(x_i, x_i)$.

Proof. Using the bounds $r_n^2 \le s_i^2 \le R_n^2$ for all $1 \le i \le n$ and Lemma 3.1, we have

$$r_n^2 \left(\|\widehat{f_{S_i}}\|_K^2 - \|\widehat{f_{S_{i-1}}}\|_K^2\right) \le \left(y_i - \widehat{f_{S_{i-1}}}(x_i)\right)^2 \le R_n^2 \left(\|\widehat{f_{S_i}}\|_K^2 - \|\widehat{f_{S_{i-1}}}\|_K^2\right).$$

Adding these inequalities together for $1 \le i \le n$ and dividing by n proves the inequalities in Equation (3.2). Equation (3.3) follows because the square loss upper bounds the zero-one loss. ∎

This theorem immediately gives the following corollary, a generalization bound for MNIC in the case where the $\{(x_i, y_i)\}_{i=1}^n$ are sampled *i.i.d.* from some distribution.

Theorem 3.3. *Suppose that S_n consists of i.i.d. samples from $\mathcal{D}$ and that y denotes a class label in $\{-1,1\}$. Let $(\mathbf{x},\mathbf{y})$ denote a new random draw from the same $\mathcal{D}$.*

1. *We have*
$$\min_{1\le i\le n} \frac{n\cdot \mathbb{P}[\mathbf{y}\widehat{f_{S_i}}(\mathbf{x})<0]}{\mathbb{E}[R_n^2 B_n^2]} \le 1\,.$$

2. *If either $\mathbb{E}[(1-\mathbf{y}\widehat{f_{S_i}}(\mathbf{x}))^2]$ or $\mathbb{P}[\mathbf{y}\widehat{f_{S_i}}(\mathbf{x})\le 0]$ is a non-increasing sequence indexed by i, we have*
$$\mathbb{P}[\mathbf{y}\widehat{f_{S_n}}(\mathbf{x})\le 0] \le \frac{\mathbb{E}[R_n^2 B_n^2]}{n}\,.$$

3. *If we denote the Polyak average of $\widehat{f_{S_i}}$ by*
$$\widetilde{f_n} = \frac{1}{n}\sum_{i=1}^n \widehat{f_{S_i}}$$
then we have
$$\mathbb{P}\left[\mathbf{y}\widetilde{f_n}(\mathbf{x})\le 0\right] \le \frac{\mathbb{E}[R_{n+1}^2 B_{n+1}^2]}{n+1}\,.$$

Proof. Taking expected values in Equation (3.2), we can drop the subscripts on (x_i, y_i) as they are identically distributed to $(\mathbf{x},\mathbf{y})\sim\mathcal{D}$ and is independent of S_{i-1}. This yields the bound

$$\sum_{i=1}^n \mathbb{E}[(1-\mathbf{y}\widehat{f_{S_{i-1}}}(\mathbf{x}))^2] = \sum_{i=1}^n \mathbb{E}[(1-y_i\widehat{f_{S_{i-1}}}(x_i))^2] \le \mathbb{E}[R_n^2 B_n^2]. \tag{3.4}$$

Inequality (3.4) immediately proves

$$\min_{0\le i\le n-1} \mathbb{E}[(1-\mathbf{y}\widehat{f_{S_i}}(\mathbf{x}))^2] \le \frac{\mathbb{E}[R_n^2 B_n^2]}{n}\,.$$

Markov's inequality implies that

$$\mathbb{P}[\mathbf{y}\widehat{f_{S_i}}(\mathbf{x})\le 0] \le \mathbb{E}[(1-\mathbf{y}\widehat{f_{S_i}}(\mathbf{x}))^2]\,.$$

It is clear that $\min_{1\le i\le n}\mathbb{P}[\mathbf{y}\widehat{f_{S_i}}(\mathbf{x})\le 0] \le \min_{1\le i\le n-1}\mathbb{P}[\mathbf{y}\widehat{f_{S_i}}(\mathbf{x})\le 0]$. Also, the second minimum value stays the same over $0\le i\le n-1$ as $\widehat{f_{S_0}}=0$. Therefore the first part of the theorem is proved. Similarly, assuming the sequence is non-increasing means that the minimum summand of (3.4) is $\mathbb{E}[(1-\mathbf{y}\widehat{f_n}(\mathbf{x}))^2]$ (or respectively $\mathbb{P}[\mathbf{y}\widehat{f_{S_i}}(\mathbf{x})\le 0]$). This proves the second part of the theorem. The third part of the theorem follows by applying Jensen's inequality to lower bound the left-hand-side of inequality (3.4) (with $n+1$ substituting n), applying Markov's inequality, and then rescaling by and then rescaling by $\frac{n+1}{n}$. ∎

Finally, note that we could have derived a result similar to Theorem 3.3 for *regression*, replacing probability of error with expected squared loss. For example, the same argument would show

$$\min_{1\le i\le n} \mathbb{E}\left[(\mathbf{y}-\widehat{f_{S_i}}(\mathbf{x}))^2\right] \le \frac{\mathbb{E}[R_n^2 B_n^2]}{n}\,.$$

However, we note that the norm bound B_n may grow rapidly for such regression problems. Whereas we will see several examples in the sequel where the norm bound B_n grows slowly with n for classification, we leave study of conditions under which these norms grow sublinearly in n for regression to future work.

4 Proofs of Lemma 3.1: Algorithmic and Geometric Perspectives

Lemma 3.1 is the heart of our analysis and has a several simple proofs. We present these varied views as they provide many useful algebraic, geometric, and algorithmic insights into least-squares classification and interpolation.

4.1 Algebraic proof

The algebraic proof of Lemma 3.1 follows from formulae for matrix inverses. Let K denote the kernel matrix for S_i. Partition K as

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$$

where K_{11} is $(i-1)\times(i-1)$ and K_{22} is a scalar equal to $K(x_i, x_i)$. Note that under this partitioning,

$$\widehat{f}_{S_{i-1}}(x_i) = K_{21} K_{11}^{-1} y_{1:i-1}\,,$$

where $y_{1:i-1}$ slices all but the last element of y.

First, note that

$$s_i^2 = K_{22} - K_{21} K_{11}^{-1} K_{12}\,.$$

Next, using the formula for inverting partitioned matrices, we have that

$$K^{-1} = \begin{bmatrix} (K_{11} - K_{12} K_{22}^{-1} K_{21})^{-1} & s_i^{-2} K_{11}^{-1} K_{12} \\ s_i^{-2} (K_{11}^{-1} K_{12})^\top & s_i^{-2} \end{bmatrix}.$$

By the Woodbury formula we have

$$(K_{11} - K_{12} K_{22}^{-1} K_{21})^{-1} = K_{11}^{-1} + s_i^{-2} \left(K_{21} K_{11}^{-1}\right)^\top \left(K_{21} K_{11}^{-1}\right).$$

Hence,

$$\|\widehat{f}_{S_i}\|_K^2 = y^\top K^{-1} y = s_i^{-2}(y_i - 2 y_i \widehat{f}_{S_{i-1}}(x_i) + \widehat{f}^2_{S_{i-1}}(x_i)) + y_{1:i-1}^\top K_{11}^{-1} y_{1:i-1}\,.$$

Rearranging terms proves the lemma.

4.2 Geometric Proof

We can sketch a geometric proof Lemma 3.1 which gives more light into the sequential nature of our regret bound. This proof naturally gives rise to an online algorithm for solving MNIC, demonstrating that at each step we only need to increment our function in a direction orthogonal to all of the previously seen examples.

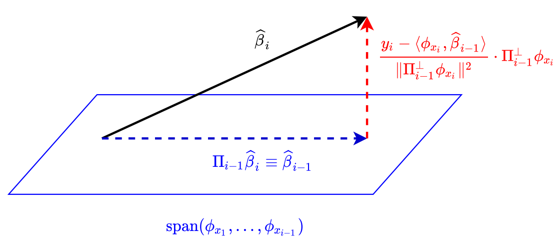

Figure 1: Illustration of the geometric proof.

Recall that $\widehat{f}_{S_i}(x)$ is a linear function in φ_x

$$\widehat{f}_{S_i}(x) = \langle \varphi_x, \widehat{\beta}_i \rangle_K \quad , \tag{4.1}$$

where $\widehat{\beta_i}$ lies in the span$(\varphi_{x_1},\dots,\varphi_{x_i})$. Let us compare $\widehat{\beta_i}$ and $\widehat{\beta}_{i-1}$. Let Π_{i-1} denote the orthogonal projection onto span$(\varphi_{x_1},\dots,\varphi_{x_{i-1}})$. Observe that one must have

$$\Pi_{i-1}\widehat{\beta_i} = \widehat{\beta}_{i-1} \quad . \tag{4.2}$$

This is because the function $g_1(x) := \langle \varphi_x, \Pi_{i-1}\widehat{\beta_i}\rangle_K$ has to interpolate the data set S_{i-1} as

$$g_1(x_j) \equiv \widehat{f}_{S_i}(x_j)$$

for $1 \le j \le i-1$. On the span of Π_{i-1}, there is a unique interpolating solution to S_{i-1}, namely $\widehat{\beta}_{i-1}$, thus proving the observation. Therefore,

$$\widehat{\beta_i} = \widehat{\beta}_{i-1} + \Pi_{i-1}^{\perp}\widehat{\beta_i} \tag{4.3}$$

$$= \widehat{\beta}_{i-1} + c\cdot\Pi_{i-1}^{\perp}\varphi_{x_i} \tag{4.4}$$

with some constant c. Since $\widehat{\beta_i}$ lies in the span of $\{\varphi_{x_1},\dots,\varphi_{x_{i-1}}\}\cup\{\varphi_{x_i}\}$. To evaluate c, let us apply the function to the data point (x_i, y_i)

$$y_i = \widehat{f}_{S_i}(x_i) = \langle \varphi_{x_i}, \widehat{\beta_i}\rangle_K = \langle \varphi_{x_i}, \widehat{\beta}_{i-1}\rangle_K + c\cdot\langle \varphi_{x_i}, \Pi_{i-1}^{\perp}\varphi_{x_i}\rangle_K \tag{4.5}$$

$$= \widehat{f}_{S_{i-1}}(x_i) + c\cdot s_i^2 \quad . \tag{4.6}$$

Thus, we have proven

$$\widehat{\beta_i} = \widehat{\beta}_{i-1} + \frac{y_i - \widehat{f}_{S_{i-1}}(x_i)}{s_i^2}\cdot\Pi_{i-1}^{\perp}\varphi_{x_i} \quad . \tag{4.7}$$

Evaluating the norm on both sides of the equation will conclude the proof.

4.3 An Online Interpolation Algorithm

The geometric proof of Lemma 3.1 describes the following online algorithm for MNIC:

Note that the function update can be written equivalently in the kernel form: define $g_i(x)$

$$g_i(x) = K(x,x_i) - K(x,X_{i-1})[K(X_{i-1},X_{i-1})]^{-1}K(X_{i-1},x_i) \tag{4.8}$$

with X_{i-1} concatenating $\{x_1,\dots,x_{i-1}\}$, then

$$\widehat{f}_{S_i}(x) = \widehat{f}_{S_{i-1}}(x) + \frac{\epsilon_i}{g_i(x_i)}g_i(x) \quad . \tag{4.9}$$

Theorem 3.2 can be interpreted as deriving a regret bound for this online algorithm with minimal assumptions.

4.4 Connections to the QR decomposition

The online algorithm defined by Algorithm 1 turns out to be equivalent to solving MNIC with a QR factorization. Recall that the QR factorization of a matrix A writes $A = QR$ where Q has orthonormal columns and R is upper triangular. If A is a data matrix with each column a data point, then Q is an orthonormal basis for the span of the data, and hence we can compute the projection matrices Π directly from Q. We formalize this observation in the following Proposition.

Proposition 4.1. *Suppose φ maps into a p dimensional space and $K(x,x') = \langle\varphi_x,\varphi_{x'}\rangle$ with the inner product in $\mathbb{R}^p$. Define the $p\times n$ dimensional matrix Φ to have ith column equal to φ_{x_i}. Let $\Phi = QR$ be a QR decomposition of Φ and denote the ith column of Q by q_i. Let $z = R^{-\top}y$. Then $\widehat{\beta}_k = \sum_{i=1}^{k} q_i z_i$.*

Algorithm 1: Online Minimum-Norm Interpolation Algorithm.

Data: A sequence of data $\{z_i\}_{i=1}^n$. A feature embedding $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ that corresponds to an RKHS.

Result: A min-norm interpolation function $\widehat{f_{S_{i-1}}} : \mathcal{X} \to \mathcal{Y}$ on all past data S_{i-1} at each time stamp i.

Initialization: Set $\widehat{\beta_0} = \mathbf{0} \in \mathbb{R}^p$ and $\widehat{f_{S_0}} = 0$, and $i = 0$;

while $i < n$ **do**

Set $i = i + 1$ and receive a new data $z_i = (x_i, y_i)$;

Make prediction based on the previous interpolator $\widehat{f_{S_{i-1}}}$ and record the error $\epsilon_i = y_i - f_{S_{i-1}}(x_i)$. Update the vector $\widehat{\beta_i}$

$$\widehat{\beta_i} = \widehat{\beta}_{i-1} + \frac{\epsilon_i}{\|\Pi_{i-1}^{\perp} \varphi_{x_i}\|^2} \cdot \Pi_{i-1}^{\perp} \varphi_{x_i} \quad ,$$

and the corresponding min-norm interpolation function $\widehat{f_{S_i}}(x) = \langle \varphi_x, \widehat{\beta_i} \rangle$;

end

We defer the proof of this proposition to Appendix A. Proposition 4.1 also allows us to compute the average of $\widehat{\beta_i}$ from the QR decomposition. Let D denote the diagonal matrix with kth diagonal entry $1 - \frac{k-1}{n}$. Then Proposition 4.1 immediately implies that

$$\frac{1}{n} \sum_{i=1}^{n} \widehat{\beta_i} = QDR^{-\top} y. \tag{4.10}$$

This implies that computing the average of $\widehat{\beta}$ merely requires only n floating point operations more than solving for the least-norm solution of $\Phi^\top \beta = y$.

We can also kernelize the computation of the average. With slight abuse of notation, we can still let Φ denote the semiinfinite dimensional data matrix consisting of concatenation of φ_{x_i}. This Φ will have a QR decomposition. The kernel matrix $K = \Phi^\top \Phi$ will have Cholesky decomposition $R^\top R$ where R is the same matrix as in the QR decomposition of Φ. Our goal is to write $\widehat{\beta_k}$ as $\sum_{i=1}^n c_i \varphi_{x_i}$. Let P_k denote the diagonal matrix that projects onto the first k standard coordinates in $\mathbb{R}^n$. Then we have $\widehat{\beta_k} = QP_k z$ and

$$Kc = \Phi^\top Q P_k z = R^\top P_k z.$$

Hence, $\widehat{\beta_k} = \Phi c$ with $c = R^{-1} P_k R^{-\top} y$ and $\frac{1}{n} \sum_{i=1}^n \widehat{\beta_k} = \Phi \bar{c}$ with $\bar{c} = R^{-1} D R^{-\top} y$.

5 Generalization to Ridge Regression

By replacing the matrix K with $K + \lambda I$ in Lemma 3.1, all of the results derived so far immediately generalize to kernel ridge regression and RLSC. Recall that ridge regression solves the optimization problem

$$\sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_K^2 .$$

The optimal solution of the problem is

$$\widehat{f_{S_n, \lambda}} = \sum_{i=1}^{n} \alpha_i K(x_i, x), \text{ where } \alpha = (K + \lambda I)^{-1} y .$$

Thus, the only thing distinguishing the solution of the minimum norm interpolation problem (3.1) is the addition of λI. Hence, Lemma 3.1 holds, with slightly different complexity measures. We need to replace the norm of $\widehat{f}_{S_n,\lambda}$ with the complexity measure

$$B_{n,\lambda}^2 := y^\top (K+\lambda I)^{-1} y.$$

This complexity measure upper bounds the norm of the ridge regression solution. To see this, observe $y^T(K+\lambda I)^{-1}y \geq y^T(K+\lambda I)^{-1}K(K+\lambda I)^{-1}y = \alpha^T K\alpha = \|\widehat{f}_{S_n,\lambda}\|_K^2$. We additionally need to replace s_i^2 with $s_{i,\lambda}^2 := \min_{c\in\mathbb{R}^{i-1}} \left\|\varphi_{x_i} - \sum_{j=1}^{i-1} c_j\varphi_{x_j}\right\|_K^2 + \lambda\|c\|_2^2 + \lambda$. With these substitutions, immediate analogs of Theorems 3.2 and Theorem 3.3 follow without modification of the proofs.

6 Some Plausible Scenarios for Slow Norm Growth

We now turn to understanding the magnitude of the norm $\|\widehat{f}_{S_n}\|_K^2$ under different generative models. Provided there is some separation between the probability distributions of the two classes, we expect the norm to grow slowly as a function of the sample size. As we shall see, we need not assume that the data realizations are well separated or that there is no noise on the labels. Rather, we will only need that $p(x|y=1)$ and $p(x|y=-1)$ are sufficiently distinct at the population level.

We begin with a simple over-parametrized Gaussian mixture model in Section 6.1. In this case, the interpolant's norm stays bounded almost surely, implying an $O(1)$ regret with probability one. This example is further extended to more general mixture models in Section 6.2. This concerns a much broader range of scenarios where a $o(n)$ regret (sublinear) is possible. In Section 7, we study general non-linear kernels with considerably weaker distributional assumptions. There, we relate the magnitude of the interpolant's norm to the Total Variation distance between the class conditional distributions. We defer all proofs to the Appendix.

6.1 Over-parametrized Gaussian Mixture Model

The first case we consider is a linear kernel with $\varphi_x = x \in \mathbb{R}^d$ and $K(x,x') = \langle x, x'\rangle$. Assume

$$x_i = \rho y_i \cdot \vartheta_\star + \epsilon_i \tag{6.1}$$

where ρ parametrizes the strength of the signal, $\vartheta_\star$ the direction with $\|\vartheta_\star\|_2 = 1$, and $\epsilon_i \sim \mathcal{N}(0, 1/d\cdot I_d)$ the noise (independent of y_i). We consider the over-parametrized regime with $d(n)/n = \psi > 1$.

Lemma 6.1. *For i.i.d. data generated from the model defined by (6.1), $R_n^2 \leq (\rho+1)^2$ a.s. and*

$$\limsup_{n\to\infty} \ \|\widehat{f}_{S_n}\|_K^2 \leq \frac{\psi}{(\psi-1)\rho^2}, \ \ a.s. \tag{6.2}$$

This Lemma and Theorem 3.2 immediately imply that for the Gaussian mixture model (6.1),

$$\limsup_{n\to\infty} \sum_{i=1}^n \mathbb{1}\left\{y_i \widehat{f}_{S_{i-1}}(x_i) \leq 0\right\} \leq (\rho+1)^2 \frac{\psi}{(\psi-1)\rho^2} \ \ , \ \text{a.s.} \tag{6.3}$$

In other words, we make at most a constant number of errors with probability one in the infinite sequence of problems.

We can also establish a lower bound on r_n^2, which appears in the lower bound of Theorem 3.2. This model shows that our upper bound in Theorem 3.2 cannot be significantly improved.

Lemma 6.2. *The following bound holds under the same setting as in Lemma 6.1*

$$r_n^2 \geq (\rho^2+1)\frac{1}{1+C_{\rho,\psi}\cdot n}, \ \ a.s. \tag{6.4}$$

where $C_{\rho,\psi} > 0$ is a constant that does not depend on n.

6.2 Generalizations of the Mixture Model

We study a generalization to the mixture model considered before. Consider $\varphi_{x_i} \in \mathbb{R}^p$ and $K(x,x') = \langle \varphi_x, \varphi_{x'} \rangle$ with the inner product in $\mathbb{R}^p$. Suppose the following generative structure holds with a positive-definite covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$

$$y_i \cdot \varphi_{x_i} = \mu \cdot \vartheta_\star + \Sigma^{1/2}\epsilon_i \ , \ \text{with } \mu \cdot \vartheta_\star := \mathbb{E}[\mathbf{y}\varphi_{\mathbf{x}}] \tag{6.5}$$

where $\mu > 0$ is the signal strength and $\vartheta_\star$ is a unit vector in $\mathbb{R}^p$. Let $\lambda_1(\Sigma), \dots, \lambda_p(\Sigma) > 0$ denote the non-increasing sequence of eigenvalues of Σ.

Assumption 6.3 (Weak Moment). *$\epsilon_{ij}, 1 \le i \le n, 1 \le j \le p(n)$ are i.i.d. entries from a distribution with, zero first-moment, unit second-moment and uniformly-bounded m-th moment, with $m \ge 8$.*

The zero first-moment and unit second-moment can be assumed without loss of generality. The real assumptions made here are (1) bounded m-moments and (2) i.i.d. entries in ϵ_i. We leave as future work to further relax this assumption either using the small-ball analysis or more sophisticated random matrix theory.

Assumption 6.4 (Over-parametrization). *For sufficiently large $C > 0$, we have $\frac{p}{\log p} > C \cdot n$.*

This assumption is mild, in other words, it ensures that the feature map is sufficiently over-parametrized such that the empirical kernel matrix is of full rank n.

With the above, the following non-asymptotic bound holds

Theorem 6.5 (General Covariance). *Under the Assumptions 6.3 and 6.4, the following bound holds almost surely,*

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_{S_{i-1}}(x_i)\right)^2 \le R_n^2 \frac{\|\widehat{f}_{S_n}\|_K^2}{n} \le c_1 \cdot \left(\mu^2 + \mathrm{tr}(\Sigma) + \gamma_p\right) \frac{\frac{1}{\lambda_p(\Sigma)}\frac{n}{p}\left(1 + c_2 \cdot \frac{\vartheta_\star^\top \Sigma \vartheta_\star}{\lambda_p(\Sigma)}\frac{n}{p}\right)}{1 + c_3 \cdot \frac{\left(\mu^2 + \vartheta_\star^\top \Sigma \vartheta_\star\right)}{\lambda_1(\Sigma)}\frac{n}{p}} \cdot \frac{1}{n} \tag{6.6}$$

where $\gamma_p = p^{\frac{1}{2}+\frac{2}{m}}(\log p)^{0.51}$ and $c_j, 1 \le j \le 3$ are universal constants.

Corollary 6.6 (General Covariance: Ridge). *Under the same assumptions and notations as in Theorem 6.5, now consider the ridge regularized predictor $\widehat{f}_{S_i,\lambda_\star}$ in Section 5, with an explicit regularization $\lambda_\star \ge 0$. The following bound holds almost surely,*

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_{S_{i-1},\lambda_\star}(x_i)\right)^2 \le R_{n,\lambda_\star}^2 \frac{y^\top \left[K + \lambda_\star I\right]^{-1} y}{n}$$

$$\le c_1 \cdot \left(\lambda_\star + \mu^2 + \mathrm{tr}(\Sigma) + \gamma_p\right) \frac{\frac{1}{\lambda_\star/p + \lambda_p(\Sigma)}\frac{n}{p}\left(1 + c_2 \cdot \frac{\vartheta_\star^\top \Sigma \vartheta_\star}{\lambda_\star/p + \lambda_p(\Sigma)}\frac{n}{p}\right)}{1 + c_3 \cdot \frac{(\mu^2 + \vartheta_\star^\top \Sigma \vartheta_\star)}{\lambda_\star/p + \lambda_1(\Sigma)}\frac{n}{p}} \cdot \frac{1}{n} \ . \tag{6.7}$$

Let's unpack the bound in Theorem 6.5 by exploring some examples, highlighting the impact of the signal strength μ, over-parametrization p, and error covariance Σ on the regret. Consider a range of problem instances with sub-Gaussian ϵ_{ij} where

1. $\lambda_i(\Sigma) = i^{-\alpha}, 1 \le i \le p$ with some $\alpha \in (0, 1/2)$,
2. $\mu^2 = n^x$ for $x \in (0, \infty)$,
3. $p = n^y$ for $y \in (1, \infty)$.

For any fixed $\alpha \in (0, 1/2)$, Figure 2 illustrates the region on the x, y domain when the upper bounds in Theorem 6.5 in Corollary 6.6 are $o(1)$, asserting the plausible scenarios for slow norm growth. The detailed calculations are deferred to Appendix D.

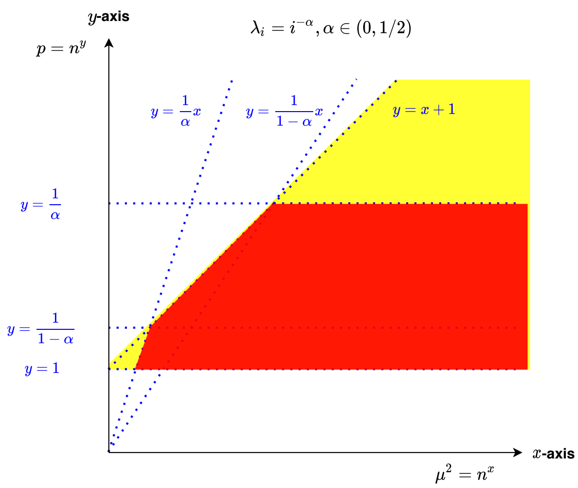

Figure 2: Illustration diagram on when slow norm growth is possible. x-axis corresponds to $\log(\mu^2)/\log(n)$, and y-axis corresponds to $\log(p)/\log(n)$. Red region shows when the upper bound in Theorem 6.5 is $o(1)$, the extended Yellow region demonstrates that with a properly chosen $\lambda_\star$, the upper bound in Corollary 6.6 is $o(1)$.

7 Separation and Slow Norm Growth

We close by considering binary classification with general non-linear kernels. Let $\eta_\star$ denote the conditional expectation of y given x

$$\eta_\star(x) = \mathbb{P}(\mathbf{y} = +1|\mathbf{x} = x) - \mathbb{P}(\mathbf{y} = -1|\mathbf{x} = x) \ . \tag{7.1}$$

We assume that with some $B_\star > 0$

$$\|\eta_\star\|_K^2 \leq B_\star^2 \ .$$

Lemma 7.1 (Expected Norm Bound Conditional on Design)**.** *Consider S_n consists of i.i.d. data drawn from some distribution $\mathcal{D}$. Denote X as the concatenated data matrix of $\{x_i\}_{i=1}^n$. Conditional on the design X, the following bound holds,*

$$\mathbb{E}\left[\|\widehat{f}_{S_n}\|_K^2 \mid X\right] \leq \|\eta_\star\|_K^2 + \frac{1}{r_n^2}\sum_{i=1}^n \left(1 - \eta_\star^2(x_i)\right) \ . \tag{7.2}$$

Lemma 7.1 and Theorem 3.2 together imply

$$\frac{1}{n}\sum_{i=1}^n \mathbb{P}\left[y_i \widehat{f}_{S_{i-1}}(x_i) \leq 0 \mid X\right] \leq \frac{R_n^2}{r_n^2} \cdot \frac{1}{n}\sum_{i=1}^n (1 - \eta_\star^2(x_i)) + \frac{R_n^2 B_\star^2}{n} \ . \tag{7.3}$$

The second term here is a generalization error that tends to zero with n. The first term is proportional to the *Bayes error*

$$\mathcal{E}(X) = \frac{1}{n}\sum_{i=1}^n (1 - \eta_\star^2(x_i))$$

Note that $\mathcal{E}(X)$ is nonnegative and measures a form of accuracy of the regression function on the sample. $\mathcal{E}(X)$ is equal to zero when the regression function makes correct assignments for all data points and equal to 1 if all of the points are impossible to distinguish under the generative model. In order for the bound (7.3) to imply low pediction error, we need the Bayes error to be small.

Two natural conditions under which the Bayes error can be analyzed are the Mammen-Tsybakov and Massart noise conditions. The Mammen-Tsybakov condition asserts that there exists some $\alpha \in [0,1]$ and $C_0 > 0$ such that

$$\mathbb{P}\Big[|\eta_\star(\mathbf{x})| \leq t\Big] \leq C_0 \cdot t^{\frac{\alpha}{1-\alpha}} \quad \forall t \in [0,1] \ . \tag{7.4}$$

Under this noise condition, we have the following upper bound on the expected Bayes Error

$$\mathbb{E}\Big[\mathcal{E}(X)\Big] = 1 - \mathbb{E}[|\eta_\star(\mathbf{x})|^2] = 1 - \int_0^1 \mathbb{P}(|\eta_\star(\mathbf{x})| > \sqrt{t})\mathrm{d}t \tag{7.5}$$

$$\leq 1 - \int_0^1 \Big(1 - C_0 \cdot t^{\frac{\alpha}{2(1-\alpha)}}\Big)\mathrm{d}t \leq C_0 \cdot \frac{2(1-\alpha)}{2-\alpha} \ . \tag{7.6}$$

Note that when $\alpha = 1$ (i.e., under the Massart noise condition), the expected Bayes Error is 0.

We can construct more general conditions under which we expect $\mathcal{E}(X)$ to be small. Let P_+ denote the conditional distribution of x given $y = +1$, and correspondingly define P_-. Let $d_{\mathrm{TV}}(P_+, P_-)$ denote the total variation distance between these two distributions.

Lemma 7.2 (Sufficient Separation)**.** *For any $\epsilon > 0$,*

$$\mathbb{P}[\mathcal{E}(X) \leq 4\epsilon] \geq 1 - \tfrac{\epsilon^{-1}+1}{2} n(1 - d_{\mathrm{TV}}(P_+, P_-)) \ . \tag{7.7}$$

Lemma 7.2 gives a general way to bound the expected Bayes error by lower bounding the total variation between the class conditional distributions. For example, in the Gaussian mixture model in Section 6.1, we can choose $\epsilon = n^{-1}$ and have with probability at least $1 - n^2 \exp(-d\rho^2/2) = 1 - o(1)$ that the error is $O(1/n)$.

References

Haim Avron, Kenneth L Clarkson, and David P Woodruff. Faster kernel ridge regression using sketching and preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1116–1138, 2017.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *arXiv preprint arXiv:2004.12019*, 2020.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*, 2019.

Noureddinel El Karoui. The spectrum of kernel random matrices. *Annals of Statistics*, 38(1):1–50, February 2010. ISSN 0090-5364, 2168-8966. doi: 10.1214/08-AOS648.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Reinhard Heckel and Mahdi Soltanolkotabi. Compressive sensing with un-trained neural networks: Gradient descent finds the smoothest approximation. 2020. arXiv preprint arXiv:2005.03991.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8580–8589, 2018.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Wee Sun Lee, Peter L Bartlett, and Robert C Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "Ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, June 2020. doi: 10.1214/19-AOS1849.

Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*, 2020.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2683–2711. PMLR, July 2020.

Siyuan Ma and Mikhail Belkin. Diving into the shallows: a computational perspective on large-scale shallow learning. In *Advances in Neural Information Processing Systems*, pages 3778–3787, 2017.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Albert B. J. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, pages 615–622, 1962.

Ryan Rifkin, Gene Yeo, and Tomaso Poggio. Regularized least squares classification. In J.A.K. Suykens, I. Horvath, S. Basu, C. Micchell, and J. Vandewalle, editors, *Advances in Learning Theory: Methods, Models and Applications*, volume 190 of *NATO Science Series, III: Computer and Systems Sciences*, chapter 7, pages 131–154. IOS Press, 2003.

Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Vaishaal Shankar, Alex Fang, Wenshuo Guo, Sara Fridovich-Keil, Jonathan Ragan-Kelley, Ludwig Schmidt, and Benjamin Recht. Neural kernels without tangents. In *International Conference on Machine Learning (ICML)*, 2020.

Vaishaal Shankar, Karl Krauth, Qifan Pu, Eric Jonas, Shivaram Venkataraman, Ion Stoica, Benjamin Recht, and Jonathan Ragan-Kelley. Numpywren: Serverless linear algebra. *Proceedings of ACM Symposium on Cloud Computing*, 2021. Preprint available at arXiv:1810.09679.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23:2199–2207, 2010.

Johan A. K. Suykens and Joose Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

Vladimir Vapnik and Olivier Chapelle. Bounds on error expectation for support vector machines. *Neural computation*, 12(9):2013–2036, 2000.

Vladimir Vapnik and Alexey Chervonenkis. *Theory of Pattern Recognition: Statistical Learning Problems*. Nauka, Moscow, 1974. In Russian.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Ke Wang, Geoff Pleiss, Jacob Gardner, Stephen Tyree, Kilian Q Weinberger, and Andrew Gordon Wilson. Exact gaussian processes on a million data points. In *Advances in Neural Information Processing Systems*, pages 14648–14659, 2019.

A Proof of Proposition 4.1

First note that by construction,

$$\varphi_{x_k} = \sum_{i=1}^{k} R_{ik} q_i \,.$$

Hence, $\Pi_{k-1}^{\perp} \varphi_{x_k} = R_{kk} q_k$. From this same calculation, we have that $s_k = R_{kk}$.

Consider the vector $z = R^{-\top} y$. We have

$$z_k = \frac{y_k - \sum_{i=1}^{k-1} R_{ik} z_i}{R_{kk}} \,.$$

Now let us proceed by induction. Certainly, for the base case, we have

$$\widehat{\beta}_1 = \frac{y_1}{s_1^2} \varphi_{x_i} = \frac{y_1}{R_{11}^2} R_{11} q_1 = z_1 q_1 \,.$$

Now, suppose the formula holds for $i < k$ and consider

$$\widehat{f}_{S_{k-1}}(x_k) = \langle \varphi_{x_k}, \widehat{\beta}_{k-1} \rangle = \left\langle \sum_{j=1}^{k} R_{jk} q_j, \sum_{\ell=1}^{k-1} q_\ell z_\ell \right\rangle = \sum_{j=1}^{k-1} R_{jk} z_j \,.$$

Thus,

$$z_k = \frac{y_k - \widehat{f}_{S_{k-1}}(x_k)}{s_k} \,.$$

and we have

$$\widehat{\beta}_{k+1} - \widehat{\beta}_k = \frac{y_k - \widehat{f}_{S_{k-1}}(x_k)}{s_k^2} \Pi_{k-1}^{\perp} \varphi_{x_k} = z_k q_k \,,$$

which completes the proof.

B Probabilistic Tools for Analyzing Generative Models

We will need two results from the literature. First, the following is Theorem 5.41 in Vershynin [2010].

Proposition B.1. *Let A be an $p \times n$ random matrix whose rows A_i are independent isotropic random vectors in $\mathbb{R}^n$. Assume that $\|A_i\|_2 \leq \sqrt{N}$ almost surely for all i. Then for every $t \geq 0$, one has the following bound on the singular values of A*

$$\sqrt{p} - t\sqrt{N} \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{p} + t\sqrt{N} \tag{B.1}$$

with probability at least $1 - 2n \cdot \exp(-ct^2)$, where $c > 0$ is an absolute constant.

The next proposition follows from Lemma A.3 in El Karoui [2010].

Proposition B.2. *Let $\{z_i\}_{i=1}^n$ be i.i.d. random vectors in $\mathbb{R}^p$, whose entries are i.i.d., mean 0, variance 1 and have bounded m-absolute moments ($m > 4$). Suppose that $\{\Sigma_p\}$ is a sequence of positive semi-definite matrices whose operator norms are uniformly bounded and n/p is asymptotically bounded. We have,*

$$\max_{1 \leq i \leq n} \left| z_i^\top \Sigma_p z_i - \operatorname{tr}(\Sigma_p) \right| \leq p^{\frac{1}{2} + \frac{2}{m}} (\log p)^{0.51} \quad a.s. \tag{B.2}$$

The following propositions are useful calculations needed for our proofs. The following proposition follows from rotational invariance of Gaussian and standard concentration of χ^2 random variables.

Proposition B.3. *Let $G \in \mathbb{R}^{n\times d}$ with each entry i.i.d. $\mathcal{N}(0,1/d)$, then for any fixed vector $v \in \mathbb{R}^n$*

$$v^\top [GG^\top]^{-1} v \overset{dist.}{=} \|v\|^2 \cdot \frac{d}{\chi^2_{d-n+1}} \ . \tag{B.3}$$

In the case with $d/n = \psi > 1$, with probability at least $1-2\exp(-t)$,

$$(1-\psi^{-1}) - 2\sqrt{\psi^{-1}(1-\psi^{-1})\frac{t}{n}} \le \frac{\|v\|^2}{v^\top[GG^\top]^{-1}v} \le (1-\psi^{-1}) + 2\sqrt{\psi^{-1}(1-\psi^{-1})\frac{t}{n}} + 2\psi^{-1}\frac{t}{n} \ . \tag{B.4}$$

Proof. Due to rotational invariance, $v^\top[GG^\top]^{-1}v \overset{dist.}{=} \|v\|^2 K_{11}$ where K_{11} is the top-left element of the matrix $[GG^\top]^{-1}$. It is easy to see that K_{11} follows the same distribution as $d/\chi^2_{d-(n-1)}$. The proof completes using the standard concentration of χ^2 random variables (see for instance, Massart-Laurent). ∎

The following proposition controls some matrix quantities needed in our analysis.

Proposition B.4. *Let $E \in \mathbb{R}^{n\times p}$ be random matrix with i.i.d. entries, zero-mean and bounded-m moments. $E_{i,\cdot} \in \mathbb{R}^p$ denotes the i-th row of E, and $E_{\cdot,j} \in \mathbb{R}^n$ denotes the j-th column of E. The following bounds hold almost surely if $m > 4$*

$$\max_{1\le i\le n} \left| E_{i,\cdot}^\top \Sigma_p E_{i,\cdot} - \mathrm{tr}(\Sigma_p)\right| \le p^{\frac{1}{2}+\frac{2}{m}} (\log p)^{0.51} \ , \tag{B.5}$$

$$\max_{1\le j\le p} \left| E_{\cdot,j}^\top E_{\cdot,j} - n\right| \le p^{\frac{1}{2}+\frac{2}{m}} (\log p)^{0.51} \ , \tag{B.6}$$

$$\lambda_1(E\Sigma_p E^\top) \le \lambda_1(\Sigma_p)\cdot\left(\sqrt{p} + \sqrt{n(\log n)} + \sqrt{p^{\frac{1}{2}+\frac{2}{m}}(\log p)^{0.51}(\log n)}\right)^2 \le 1.5\lambda_1(\Sigma_p)\cdot p \ , \tag{B.7}$$

$$\lambda_n(E\Sigma_p E^\top) \ge \lambda_p(\Sigma_p)\cdot\left(\sqrt{p} - \sqrt{n(\log n)} - \sqrt{p^{\frac{1}{2}+\frac{2}{m}}(\log p)^{0.51}(\log n)}\right)^2 \ge 0.5\lambda_p(\Sigma_p)\cdot p \ . \tag{B.8}$$

Proof. This is a direct implication of Proposition B.1 and Proposition B.2 with the choice $t = C(\log n)^{0.5}$ ($C > 0$ large enough) and $N = n + p^{\frac{1}{2}+\frac{2}{m}}(\log p)^{0.51}$. Notice here that $p^{\frac{1}{2}+\frac{2}{m}}(\log p)^{0.51} \ll p$ for $m > 4$. ∎

Proposition B.5. *Consider the same setting as in Proposition B.4. If assumed in addition that $m \ge 8$, then the following bounds hold almost surely for a fixed $v \in \mathbb{R}^p$*

$$\left|\sum_{i=1}^n \langle v, E_{i,\cdot}\rangle^2 - n\cdot\|v\|^2\right| \le n^{0.8}\cdot\|v\|^2 \ , \tag{B.9}$$

$$\left|\sum_{i=1}^n \langle v, E_{i,\cdot}\rangle\right| \le n^{0.8}\cdot\|v\| \ . \tag{B.10}$$

of Proposition B.5. We start with the second statement. Define i.i.d. random variables $\mathbf{w}_i := \langle v, E_{i,\cdot}\rangle$, then $\mathbb{E}[\mathbf{w}_i] = 0$ and

$$\mathbb{P}\left(\left|\sum_{i=1}^n \mathbf{w}_i\right| \ge t\right) \le \frac{\mathbb{E}\left[\left(\sum_{i=1}^n \mathbf{w}_i\right)^4\right]}{t^4} \le C\frac{n\,\mathbb{E}[\mathbf{w}_1^4] + n^2(\mathbb{E}[\mathbf{w}_1^2])^2}{t^4} \tag{B.11}$$

$$\mathbb{P}\left(\left|\sum_{i=1}^n \mathbf{w}_i\right| \ge n^{0.8}\cdot\|v\|\right) \le C'\frac{n^2\|v\|^4}{t^4} \le C'\frac{1}{n^{1.2}} \ , \tag{B.12}$$

with choice $t = n^{0.8}\|v\|$. Since the tail probability is summable w.r.t. n, Borell-Cantelli Lemma proved the almost surely statement.

Define $\mathbf{w}_i^2 = \langle v, E_{i,\cdot}\rangle^2$, then $\mathbb{E}[\mathbf{w}_i^2] = \|v\|^2$ and

$$\mathbb{E}[\mathbf{w}_i^4] = \mathbb{E}\left[\Big(\sum_{j\in[p]} v_j E_{i,j}\Big)^4\right] \le C\Big(\sum_j v_j^4 + \sum_{j\neq j'} v_j^2 v_{j'}^2\Big) \le C\|v\|^4 \ . \tag{B.13}$$

By Chebyshev's inequality

$$\mathbb{P}\left(\Big|\sum_{i=1}^n \mathbf{w}_i^2 - \mathbb{E}[\mathbf{w}_i^2]\Big| \ge t\right) \le \frac{\mathbb{E}\left[\left(\sum_{i=1}^n \mathbf{w}_i^2 - \mathbb{E}[\mathbf{w}_i^2]\right)^4\right]}{t^4} \ . \tag{B.14}$$

It is easy to see that

$$\mathbb{E}\left[\Big(\sum_{i=1}^n \mathbf{w}_i^2 - \mathbb{E}[\mathbf{w}_i^2]\Big)^4\right] \le C\left(n\,\mathbb{E}[(\mathbf{w}_i^2 - \mathbb{E}[\mathbf{w}_i^2])^4] + n^2\Big(\mathrm{Var}[\mathbf{w}_i^2]\Big)^2\right) \le C' n^2 \|v\|^8,$$

as $\mathrm{Var}[\mathbf{w}_i^2] \le \mathbb{E}[\mathbf{w}_i^4]$ and $\mathbb{E}[(\mathbf{w}_i^2 - \mathbb{E}[\mathbf{w}_i^2])^4] \le C\|v\|^8$ if E_{ij} has uniformly bounded m-th moment with $m \ge 8$. Now choose $t = n^{0.8}\|v\|^2$, we have

$$\mathbb{P}\left(\sum_{i=1}^n (z_i - \mathbb{E}[z_i]) \ge t\right) \le C' \frac{n^2\|v\|^8}{(n^{0.8}\|v\|^2)^4} \le C' \frac{1}{n^{1.2}} \tag{B.15}$$

The proof is again complete using the Borell-Cantelli Lemma. ■

C Proofs for Section 6

In this section, we use $\Phi_n \in \mathbb{R}^{n\times p}, X_n \in \mathbb{R}^{n\times d}$ to denote the embedded feature matrix and the original design matrix respectively. $Y_n \in \mathbb{R}^n$ the response vector, and $Y_n \circ X_n = [y_1 x_1, \dots, y_n x_n]^\top \in \mathbb{R}^{n\times d}$.

of Lemma 6.1. It can be verified that, based on the definition of Hadamard product

$$\|\widehat{f}_{S_n}\|_K^2 = \mathbf{1}^\top [(Y_n \circ X_n)(Y_n \circ X_n)^\top]^{-1} \mathbf{1} \ . \tag{C.1}$$

Due to the rotational invariance of Gaussian, the distribution of $\|\widehat{f}_{S_n}\|_K^2$ stays unchanged under orthogonal groups on $\mathbb{R}^d$. Therefore one can without loss of generality assume $\vartheta_\star = e_1$, then

$$Y_n \circ X_n = [\rho\mathbf{1} + g, Z], \quad g \in \mathbb{R}^n, Z \in \mathbb{R}^{n\times(d-1)} \tag{C.2}$$

where each entry of g, Z is drawn i.i.d. from $\mathcal{N}(0, 1/d)$.

Now, by the Woodbury formula

$$[(Y_n \circ X_n)(Y_n \circ X_n)^\top]^{-1} = [ZZ^\top]^{-1} - \frac{[ZZ^\top]^{-1}(\rho\mathbf{1}+g)(\rho\mathbf{1}+g)^\top [ZZ^\top]^{-1}}{1 + (\rho\mathbf{1}+g)^\top [ZZ^\top]^{-1}(\rho\mathbf{1}+g)} \ , \tag{C.3}$$

thus

$$\|\widehat{f}_{S_n}\|_K^2 \overset{dist.}{=} \mathbf{1}^\top [ZZ^\top]^{-1}\mathbf{1} - \frac{\left(\mathbf{1}^\top [ZZ^\top]^{-1}(\rho\mathbf{1}+g)\right)^2}{1 + (\rho\mathbf{1}+g)^\top [ZZ^\top]^{-1}(\rho\mathbf{1}+g)} \tag{C.4}$$

$$= \frac{\mathbf{1}^\top [ZZ^\top]^{-1}\mathbf{1} \cdot \left(1 + g^\top [ZZ^\top]^{-1} g\right) - \left(\mathbf{1}^\top [ZZ^\top]^{-1} g\right)^2}{1 + (\rho\mathbf{1}+g)^\top [ZZ^\top]^{-1}(\rho\mathbf{1}+g)} \ . \tag{C.5}$$

Using Proposition B.3, we know that with probability at least $1-4\exp(-t)$

$$(\rho\mathbf{1}+g)^\top[ZZ^\top]^{-1}(\rho\mathbf{1}+g) \tag{C.6}$$

$$\geq \left[\rho^2 n - 2\rho\sqrt{\psi^{-1}t} + \psi^{-1}(1-2\sqrt{\frac{t}{n}})\right]\left(\frac{\psi}{\psi-1}\frac{1}{1+2\sqrt{\frac{1}{\psi-1}\frac{t}{n}}+2\frac{1}{\psi-1}\frac{t}{n}}\right) \tag{C.7}$$

$$\geq n\rho^2\frac{\psi}{\psi-1}(1-O(\sqrt{t/n})) \ . \tag{C.8}$$

Similarly, with probability at least $1-5\exp(-t)$,

$$g^\top[ZZ^\top]^{-1}g \leq \psi^{-1}\left[1+2\sqrt{\frac{t}{n}}+2\frac{t}{n}\right]\left(\frac{\psi}{\psi-1}\frac{1}{1-2\sqrt{\frac{1}{\psi-1}\frac{t}{n}}}\right) \tag{C.9}$$

$$\leq \frac{1}{\psi-1}(1+O(\sqrt{t/n})) \ , \tag{C.10}$$

and

$$\mathbf{1}^\top[ZZ^\top]^{-1}\mathbf{1} \leq n\left(\frac{\psi}{\psi-1}\frac{1}{1-2\sqrt{\frac{1}{\psi-1}\frac{t}{n}}}\right) \tag{C.11}$$

$$\leq n\frac{\psi}{\psi-1}(1+O(\sqrt{t/n})) \ . \tag{C.12}$$

Putting the above three estimates together, with probability at least $1-9\exp(-t)$, we have

$$\|\widehat{f}_{S_n}\|_K^2 \leq \frac{\psi}{(\psi-1)\rho^2}(1+O(\sqrt{t/n})) \ . \tag{C.13}$$

Take $t = 11\log n$, we know that

$$\limsup_{n\to\infty} \ \|\widehat{f}_{S_n}\|_K^2 \leq \frac{\psi}{(\psi-1)\rho^2}, \ \text{a.s.} \tag{C.14}$$

by using the Borell-Cantelli Lemma. ■

of Lemma 6.2. For the lower bound, let's first condition on $x_i = x$ and project $X_{S_n\backslash i}$ to the normal vector $\bar{x} = x/\|x\|$, which denoted as $v = X_{S_n\backslash i}\Pi_x \in \mathbb{R}^n$. We know

$$x^\top x - x^\top X_{S_n\backslash i}^\top[X_{S_n\backslash i}X_{S_n\backslash i}^\top]^{-1}X_{S_n\backslash i}x \tag{C.15}$$

$$= \|x\|^2\left(1 - v^\top[vv^\top + X_{S_n\backslash i}\Pi_x^\perp X_{S_n\backslash i}^\top]^{-1}v\right) \tag{C.16}$$

$$= \|x\|^2\left(1-\left[v^\top[X_{S_n\backslash i}\Pi_x^\perp X_{S_n\backslash i}^\top]^{-1}v - \frac{(v^\top[X_{S_n\backslash i}\Pi_x^\perp X_{S_n\backslash i}^\top]^{-1}v)^2}{1+v^\top[X_{S_n\backslash i}\Pi_x^\perp X_{S_n\backslash i}^\top]^{-1}v}\right]\right) \tag{C.17}$$

$$= \|x\|^2\frac{1}{1+v^\top[X_{S_n\backslash i}\Pi_x^\perp X_{S_n\backslash i}^\top]^{-1}v} \ . \tag{C.18}$$

Now we will upper bound $v^\top[X_{S_n\backslash i}\Pi_x^\perp X_{S_n\backslash i}^\top]^{-1}v$. Define a vector u, which is the projection of the matrix $X_{S_n\backslash i}$ to the vector space

$$\tilde{\vartheta} := \vartheta_\star - \langle\vartheta_\star, \bar{x}\rangle\bar{x} \in \mathbb{R}^d \tag{C.19}$$

$$u = X_{S_n \setminus i} \Pi_{\tilde{\vartheta}} \ . \tag{C.20}$$

By construction, $\tilde{\vartheta} \perp x$. We can verify that

$$v^\top [X_{S_n\setminus i} \Pi_x^\perp X_{S_n\setminus i}^\top]^{-1} v \overset{dist.}{=} v^\top [uu^\top + ZZ^\top]^{-1} v^\top \tag{C.21}$$

$$= v^\top [ZZ^\top]^{-1} v - \frac{(v^\top [ZZ^\top]^{-1} u)^2}{1 + u^\top [ZZ^\top]^{-1} u} \tag{C.22}$$

where (v, u) is independent of Z and that $Z \in \mathbb{R}^{n\times(d-2)}$ has i.i.d. entries $Z_{ij} \sim \mathcal{N}(0, 1/d)$.

Let's analyze the distribution of each entry of $u, v \in \mathbb{R}^n$ conditioned on x. Define $r = \langle \bar{x}, \vartheta_\star \rangle$

$$v_i = \rho r y_i + \frac{1}{\sqrt{d}} g_i^{(1)}, \ \ g_i^{(1)} \sim \mathcal{N}(0,1) \ , \tag{C.23}$$

$$u_i = \frac{1}{\sqrt{1-r^2}} (\rho y_i - r^2) + \frac{1}{\sqrt{d}} g_i^{(2)}, \ \ g_i^{(2)} \sim \mathcal{N}(0,1) \ , \tag{C.24}$$

with $g^{(1)}, g^{(2)}, Z$ mutually independent. Using the Proposition B.3, we know that

$$(v^\top [ZZ^\top]^{-1} v)(u^\top [ZZ^\top]^{-1} u) - (v^\top [ZZ^\top]^{-1} u)^2 \le \frac{\rho^2 r^6}{1-r^2} n^2 \frac{\psi}{\psi - 1} (1 + O(\sqrt{\frac{\log n}{n}})) \tag{C.25}$$

and

$$u^\top [ZZ^\top]^{-1} u \ge \rho^2 r^2 n \frac{\psi}{\psi - 1} (1 - O(\sqrt{\frac{\log n}{n}})) \tag{C.26}$$

Combining with the concentration bound on $r^2 = \frac{\rho^2}{1+\rho^2}(1 + O(\sqrt{\frac{\log n}{n}}))$, we finish the proof. ∎

of Theorem 6.5. The proof proceeds in a similar way as in Lemma 6.1,

$$\|\widehat{f}_{S_n}\|_K^2 = \mathbf{1}^\top [(Y_n \circ \Phi_n)(Y_n \circ \Phi_n)^\top]^{-1} \mathbf{1} \ .$$

Define the projection matrix $\Pi_{\vartheta_\star}$ and $\Pi_{\vartheta_\star}^\perp := I_p - \Pi_{\vartheta_\star}$ and $E = [\epsilon_{ij}] \in \mathbb{R}^{n\times p}$, we know that

$$(Y_n \circ \Phi_n)(Y_n \circ \Phi_n)^\top = (\mu \mathbf{1} \vartheta_\star^\top + E\Sigma^{1/2})(\Pi_{\vartheta_\star} + \Pi_{\vartheta_\star}^\perp)(\mu \mathbf{1} \vartheta_\star^\top + E\Sigma^{1/2})^\top \tag{C.27}$$

$$= (\mu \mathbf{1} + g)(\mu \mathbf{1} + g)^\top + ZZ^\top \tag{C.28}$$

with $Z := E\Sigma^{1/2} \Pi_{\vartheta_\star}^\perp \in \mathbb{R}^{n\times p}$ and $g := E\Sigma^{1/2} \vartheta_\star \in \mathbb{R}^n$

$$\mathrm{cov}(g_i) = \vartheta_\star^\top \Sigma \vartheta_\star \tag{C.29}$$

$$\mathrm{cov}(Z_i) = \Pi_{\vartheta_\star}^\perp \Sigma \Pi_{\vartheta_\star}^\perp \tag{C.30}$$

$$\mathrm{cov}(g_i, Z_i) = \vartheta_\star^\top \Sigma \Pi_{\vartheta_\star}^\perp \tag{C.31}$$

across $1 \le i \le n$, $\{g_i, Z_i\}$ are mutually independent. Recall Equation (C.3), we know

$$\|\widehat{f}_{S_n}\|_K^2 = \mathbf{1}^\top [ZZ^\top]^{-1} \mathbf{1} - \frac{\left(\mathbf{1}^\top [ZZ^\top]^{-1} (\mu \mathbf{1} + g)\right)^2}{1 + (\mu \mathbf{1} + g)^\top [ZZ^\top]^{-1} (\mu \mathbf{1} + g)} \tag{C.32}$$

$$= \frac{\mathbf{1}^\top [ZZ^\top]^{-1} \mathbf{1} \cdot \left(1 + g^\top [ZZ^\top]^{-1} g\right) - \left(\mathbf{1}^\top [ZZ^\top]^{-1} g\right)^2}{1 + (\mu \mathbf{1} + g)^\top [ZZ^\top]^{-1} (\mu \mathbf{1} + g)} \tag{C.33}$$

$$\le \frac{\mathbf{1}^\top [ZZ^\top]^{-1} \mathbf{1} \cdot \left(1 + g^\top [ZZ^\top]^{-1} g\right)}{1 + (\mu \mathbf{1} + g)^\top [ZZ^\top]^{-1} (\mu \mathbf{1} + g)} \ . \tag{C.34}$$

Now we are going to lower bound $(\mu\mathbf{1}+g)^\top[ZZ^\top]^{-1}(\mu\mathbf{1}+g)$ and upper bound $\mathbf{1}^\top[ZZ^\top]^{-1}\mathbf{1}$ and $g^\top[ZZ^\top]^{-1}g$. We can verify that from Proposition B.4

$$1.5\lambda_1(\Sigma)p \ge \lambda_{\max}(ZZ^\top) \ge \lambda_{\min}(ZZ^\top) \ge 0.5\lambda_p(\Sigma)p \tag{C.35}$$

almost surely when $p/\log p > n$ and $m > 4$. Therefore, we have

$$\mathbf{1}^\top[ZZ^\top]^{-1}\mathbf{1} \le \frac{\|\mathbf{1}\|^2}{\lambda_{\min}(ZZ^\top)} \le c'\frac{1}{\lambda_p(\Sigma)}\frac{n}{p} \ , \tag{C.36}$$

$$g^\top[ZZ^\top]^{-1}g \le \frac{\|g\|^2}{\lambda_{\min}(ZZ^\top)} \le c'\frac{\vartheta_\star^\top\Sigma\vartheta_\star}{\lambda_p(\Sigma)}\frac{n}{p} \ , \tag{C.37}$$

$$(\mu\mathbf{1}+g)^\top[ZZ^\top]^{-1}(\mu\mathbf{1}+g) \ge \frac{\|\mu\mathbf{1}+g\|^2}{\lambda_{\max}(ZZ^\top)} = \frac{\mu^2 n + \|g\|^2 + 2\mu\langle\mathbf{1},g\rangle}{\lambda_{\max}(ZZ^\top)} \tag{C.38}$$

$$\ge c''\frac{\mu^2+\vartheta_\star^\top\Sigma\vartheta_\star}{\lambda_1(\Sigma)}\frac{n}{p} \ . \tag{C.39}$$

In the above equations, we used the fact that almost surely on g

$$\left|\|g\|^2 - n\cdot\vartheta_\star^\top\Sigma\vartheta_\star\right| \le n^{0.8}\cdot\vartheta_\star^\top\Sigma\vartheta_\star, \tag{C.40}$$

$$\left|\langle\mathbf{1},g\rangle\right| \le n^{0.8}\cdot\sqrt{\vartheta_\star^\top\Sigma\vartheta_\star} \tag{C.41}$$

by using Proposition B.5 with $g_i = \langle\Sigma^{1/2}\vartheta_\star, E_{i,\cdot}\rangle$.

By Proposition B.4, we have

$$R_n^2 \le \max_i \|y_i\varphi_{x_i}\|^2 \le 2\left(\mu^2 + \mathrm{tr}(\Sigma) + p^{\frac{1}{2}+\frac{2}{m}}(\log p)^{0.51}\right) \ . \tag{C.42}$$

Putting things together, we know

$$R_n^2\frac{\|\widehat{f}_{S_n}\|_K^2}{n} \le c_1\cdot\left(\mu^2+\mathrm{tr}(\Sigma)+p^{\frac{1}{2}+\frac{2}{m}}(\log p)^{0.51}\right)\frac{\frac{1}{\lambda_p(\Sigma)}\frac{n}{p}\left(1+c_2\cdot\frac{\vartheta_\star^\top\Sigma\vartheta_\star}{\lambda_p(\Sigma)}\frac{n}{p}\right)}{1+c_3\cdot\frac{\left(\mu^2+\vartheta_\star^\top\Sigma\vartheta_\star\right)}{\lambda_1(\Sigma)}\frac{n}{p}}\cdot\frac{1}{n} \ . \tag{C.43}$$

■

of Corollary 6.6. The proof follows exactly the same steps as in Theorem 6.5, with $\lambda_j(\Sigma)$ substituted by $\lambda_j(\Sigma)+\lambda_\star/p$ (with j being either 1 or p), and noting that $R_{n,\lambda_\star}^2 \le \lambda_\star + R_n^2$. Here we also used the fact that

$$Y_n^\top\left[\lambda_\star I_n + \Phi_n\Phi_n^\top\right]^{-1}Y_n = \mathbf{1}^\top\left[\lambda_\star I_n\circ(Y_nY_n^\top)+(Y_n\circ\Phi_n)(Y_n\circ\Phi_n)^\top\right]^{-1}\mathbf{1} \ , \tag{C.44}$$

and the Hadamard product of $\lambda_\star I_n\circ(Y_nY_n^\top) = \lambda_\star I_n$. The rest of the proof follows as in proof of Theorem 6.5 with $ZZ^\top$ replaced by $\lambda_\star I_n + ZZ^\top$. ■

D Calculations for Figure 2

Consider $m=\infty$, namely, the tail of ϵ_{ij} behaves sub-Gaussian. The diagram illustrating the following cases is in Figure 2.

We use the asymptotic notations $a(n)\lesssim b(n)$, $a(n)\gtrsim b(n)$ to denote the usual asymptotic ordering. We start with the interpolation case (no explicit regularization).

- Consider the spectral decay $\lambda_1(\Sigma) = \lambda_p(\Sigma)$ and signal strength $\mu^2 \geq \mathrm{tr}(\Sigma)$:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_{S_{i-1}}(x_i)\right)^2 \lesssim \mu^2 \frac{\frac{n}{p}}{\mu^2 \frac{n}{p}} \cdot \frac{1}{n} \lesssim \frac{1}{n}.$$

  This again confirms our $O(1)$ cumulative regret result as in Lemma 6.1.

- Consider the spectral decay $\lambda_i(\Sigma) = i^{-\alpha}, i \in [p]$ with $\alpha \in (0, 1/2)$ and signal strength $\mu^2 \gtrsim p^{1-\alpha}$ (note that $\mathrm{tr}(\Sigma) = p^{1-\alpha} \gtrsim \gamma_p = p^{\frac{1}{2}}(\log p)^{0.51}$):

  - if $p^{1-\alpha} \gtrsim n \gtrsim p^{\alpha}$, then

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_{S_{i-1}}(x_i)\right)^2 \lesssim \left(\mu^2 + p^{1-\alpha}\right)\frac{p^{\alpha}\frac{n}{p}(1+o(1))}{\mu^2\frac{n}{p}} \cdot \frac{1}{n} \lesssim \frac{p^{\alpha}}{n} = o(1)$$

  - if $p \gtrsim n \gtrsim p^{1-\alpha}$, then

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_{S_{i-1}}(x_i)\right)^2 \lesssim \left(\mu^2 + p^{1-\alpha}\right)\frac{\left(p^{\alpha}\frac{n}{p}\right)^2}{\mu^2\frac{n}{p}} \cdot \frac{1}{n} \lesssim \frac{1}{p^{1-2\alpha}} = o(1)$$

- Consider the spectral decay $\lambda_i(\Sigma) = i^{-\alpha}, i \in [p]$ with $\alpha \in (0, 1/2)$ and signal strength $\max\{p/n, p^{\alpha}\} \lesssim \mu^2 \lesssim p^{1-\alpha}$ (note that $\mathrm{tr}(\Sigma) = p^{1-\alpha} \gtrsim \gamma_p = p^{\frac{1}{2}}(\log p)^{0.51}$):

  - if $n \lesssim p^{1-\alpha}$, then

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_{S_{i-1}}(x_i)\right)^2 \lesssim p^{1-\alpha}\frac{p^{\alpha}\frac{n}{p}(1+o(1))}{\mu^2\frac{n}{p}} \cdot \frac{1}{n} \lesssim \frac{p/n}{\mu^2} = o(1)$$

  - if $n \gtrsim p^{1-\alpha}$, then

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_{S_{i-1}}(x_i)\right)^2 \lesssim p^{1-\alpha}\frac{\left(p^{\alpha}\frac{n}{p}\right)^2}{\mu^2\frac{n}{p}} \cdot \frac{1}{n} \lesssim \frac{p^{\alpha}}{\mu^2} = o(1)$$

All the above scenarios correspond to the Red region in Figure 2.

For the ridge case with explicit regularization $\lambda_\star$. First notice that the Red region certainly correspond to $o(1)$ error with the choice $\lambda_\star = 0$. Below we only consider how to extend the region with a proper choice $\lambda_\star \neq 0$.

- Consider the spectral decay $\lambda_i(\Sigma) = i^{-\alpha}, i \in [p]$ with $\alpha \in (0, 1/2)$ and signal strength $\mu^2 \gtrsim p/n$ and overparametrization $p^{\alpha} \gtrsim n$ (the upper Yellow region in Figure 2):

  - if $\mu^2 \lesssim p$, then with the choice of $\max\{n, \mu^2, p^{1-\alpha}\} \lesssim \lambda_\star \lesssim p\lambda_1(\Sigma)$

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_{S_{i-1}}(x_i)\right)^2 \lesssim \lambda_\star \frac{\frac{n}{\lambda_\star}(1+o(1))}{1+\mu^2\frac{n}{p\lambda_1(\Sigma)}} \cdot \frac{1}{n} \lesssim \frac{p/n}{\mu^2} = o(1)$$

  - if $\mu^2 \gtrsim p$, then with the choice of $\max\{n, \mu^2, p\lambda_1(\Sigma)\} \lesssim \lambda_\star \lesssim n\mu^2$

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_{S_{i-1}}(x_i)\right)^2 \lesssim \lambda_\star \frac{\frac{n}{\lambda_\star}(1+o(1))}{1+\mu^2\frac{n}{\lambda_\star}} \cdot \frac{1}{n} \lesssim \frac{\lambda_\star}{n\mu^2} = o(1)$$

- Consider the spectral decay $\lambda_i(\Sigma) = i^{-\alpha}, i \in [p]$ with $\alpha \in (0, 1/2)$ and signal strength $\mu^2 \gtrsim p/n$ and overparametrization $p^\alpha \gtrsim \mu^2$ (the lower Yellow region in Figure 2): with the choice of $p/\mu^2 \lesssim \lambda_\star \lesssim n$, we know that $\lambda_\star \gtrsim p/\mu^2 \gtrsim p^{1-\alpha}$ and $\lambda_\star \gtrsim p/\mu^2 \gtrsim \mu^2$ (due to $p^\alpha \gtrsim \mu^2$ and $\alpha < 1/2$), and therefore

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_{S_{i-1}}(x_i)\right)^2 \lesssim \lambda_\star \frac{\left(\frac{n}{\lambda_\star}\right)^2}{1+\mu^2\frac{n}{p}} \cdot \frac{1}{n} \lesssim \frac{p}{\mu^2 \lambda_\star} = o(1)$$

E Proofs for Section 7

of Lemma 7.1.

$$\mathbb{E}\left[\|\widehat{f}_{S_n}\|_K^2 \mid X_n\right] \tag{E.1}$$
$$= \mathbb{E}\left[\langle Y_n Y_n^\top, [K(X_n, X_n)]^{-1}\rangle \mid X_n\right] \tag{E.2}$$
$$= \langle \eta_\star(X_n)\eta_\star(X_n)^\top, [K(X_n,X_n)]^{-1}\rangle + \langle \mathrm{diag}\{1-\eta_\star^2(x_1), \ldots, 1-\eta_\star^2(x_n)\}, [K(X_n,X_n)]^{-1}\rangle \tag{E.3}$$
$$\le \|\eta_\star\|_K^2 + \sum_{i=1}^{n} \frac{1-\eta_\star^2(x_i)}{K(x_i,x_i) - K(x_i, X_{S_n\backslash i})[K(X_{S_n\backslash i}, X_{S_n\backslash i})]^{-1}K(X_{S_n\backslash i}, x_i)} \ . \tag{E.4}$$

Here the last line uses the Riesz representation theorem that $\eta_\star(x) = \langle \eta_\star, \varphi_x\rangle_K$, and

$$\langle \eta_\star(X_n)\eta_\star(X_n)^\top, [K(X_n,X_n)]^{-1}\rangle = \|\Pi_{\varphi_{X_n}}\eta_\star\|_K^2 \le \|\eta_\star\|_K^2 \ . \tag{E.5}$$

■

of Lemma 7.2. For the first claim, observe that

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\left(1-\eta_\star^2(\mathbf{x}_i)\right)\Big] = \int \Big[1 - \Big(\frac{\mathrm{d}P_+ - \mathrm{d}P_-}{\mathrm{d}P_+ + \mathrm{d}P_-}\Big)^2\Big]\frac{\mathrm{d}P_+ + \mathrm{d}P_-}{2} \tag{E.6}$$
$$= \int \frac{2\mathrm{d}P_+\mathrm{d}P_-}{\mathrm{d}P_+ + \mathrm{d}P_-} \le 2\int \mathrm{d}P_+ \wedge \mathrm{d}P_- \ . \tag{E.7}$$

For the second claim, let's calculate the probability of each x_i falls in the region

$$\mathcal{S}_\epsilon := \left\{x \in \mathcal{X} \mid \frac{\mathrm{d}P_+}{\mathrm{d}P_-}(x) \in [\epsilon, \epsilon^{-1}]\right\} \ . \tag{E.8}$$

We know that

$$\mathbb{P}\left[\mathbf{x} \in \mathcal{S}_\epsilon\right] = \int_{x\in\mathcal{S}_\epsilon} \frac{\mathrm{d}P_+ + \mathrm{d}P_-}{2} \le \frac{\epsilon^{-1}+1}{2}\int_{x\in\mathcal{S}_\epsilon} \mathrm{d}P_+ \wedge \mathrm{d}P_- \tag{E.9}$$
$$\le \frac{\epsilon^{-1}+1}{2}\left(1 - d_{\mathrm{TV}}(P_+, P_-)\right) \ . \tag{E.10}$$

Therefore, with union bound, we know that

$$\mathbb{P}\left[\mathbf{x}_i \notin \mathcal{S}_\epsilon, \ \ \forall 1 \le i \le n\right] \ge \left[1 - \frac{\epsilon^{-1}+1}{2}\left(1 - d_{\mathrm{TV}}(P_+, P_-)\right)\right]^n \tag{E.11}$$
$$\ge 1 - \frac{\epsilon^{-1}+1}{2}n\left(1 - d_{\mathrm{TV}}(P_+, P_-)\right) \ . \tag{E.12}$$

For $x_i \notin \mathcal{S}_\epsilon$, it is easy to verify that

$$1 - \eta_\star^2(x_i) \le 1 - \Big(\frac{1-\epsilon}{1+\epsilon}\Big)^2 \le 4\epsilon \ . \tag{E.13}$$

■