Journal of Statistical Mechanics:
Theory and Experiment

PAPER • OPEN ACCESS

# Generalization error rates in kernel regression: the crossover from the noiseless to noisy regime*

To cite this article: Hugo Cui *et al J. Stat. Mech.* (2022) 114004

OPEN ACCESS

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: ML 2022

# Generalization error rates in kernel regression: the crossover from the noiseless to noisy regime*

**Hugo Cui[1],**, Bruno Loureiro[2], Florent Krzakala[2] and Lenka Zdeborová[1]**

[1] SPOC, EPFL, Switzerland
[2] IDePHICS Lab., EPFL, Switzerland
E-mail: hugo.cui@epfl.ch

Received 31 May 2022
Accepted for publication 5 June 2022
Published 24 November 2022

Online at stacks.iop.org/JSTAT/2022/114004
https://doi.org/10.1088/1742-5468/ac9829

**Abstract.** In this manuscript we consider kernel ridge regression (KRR) under the Gaussian design. Exponents for the decay of the excess generalization error of KRR have been reported in various works under the assumption of power-law decay of eigenvalues of the features co-variance. These decays were, however, provided for sizeably different setups, namely in the noiseless case with constant regularization and in the noisy optimally regularized case. Intermediary settings have been left substantially uncharted. In this work, we unify and extend this line of work, providing characterization of all regimes and excess error decay rates that can be observed in terms of the interplay of noise and regularization. In particular, we show the existence of a transition in the noisy setting between the noiseless exponents to its noisy values as the sample complexity is increased. Finally, we illustrate how this crossover can also be observed on real data sets.

**Keywords:** machine learning, learning theory

*This article is an updated version of: Cui H, Loureiro B, Krzakala F and Zdeborová L 2021 Generalization error rates in kernel regression: the crossover from the noiseless to noisy regime *Advances in Neural Information Processing Systems* vol 34 ed M Ranzato, A Beygelzimer, Y Dauphin, P S Liang and J Wortman Vaughan (New York: Curran Associates) pp 10131–43.

**Author to whom any correspondence should be addressed.

1742-5468/22/114004+34$33.00

## Contents

https://doi.org/10.1088/1742-5468/ac9829

## 1. Introduction

Kernel methods are among the most popular models in machine learning. Despite their relative simplicity, they define a powerful framework in which non-linear features can be exploited without leaving the realm of convex optimisation. Kernel methods in machine learning have a long and rich literature dating back to the 60s [1, 2], but have recently made it back to the spotlight as a proxy for studying neural networks in different regimes, e.g. the infinite width limit [3–6] and the lazy regime of training [7]. Despite being defined in terms of a non-parametric optimisation problem, kernel methods can be mathematically understood as a standard parametric linear problem in a (possibly infinite) Hilbert space spanned by the kernel eigenvectors (a.k.a. *features*). This dual picture fully characterizes the asymptotic performance of kernels in terms of a trade-off between two key quantities: the relative decay of the eigenvalues of the kernel (a.k.a. its *capacity*) and the coefficients of the target function when expressed in feature space (a.k.a. the *source*). Indeed, a sizeable body of work has been devoted to understanding the decay rates of the excess error as a function of these two relative decays, and investigated whether these rates are attained by algorithms such as stochastic gradient descent [8, 9].

Rigorous optimal rates for the excess generalization error in kernel ridge regression (KRR) are well-known since the seminal works of [10, 11]. However, recent interesting works [12, 13] surprisingly reported very different—and actually better—rates supported by numerical evidences. These papers appeared to either not comment on this discrepancy [13], or to attribute this apparent contradiction to a difference between typical and worse-case analysis [12]. As we shall see, the key difference between these works stems instead from the fact that most of classical works considered *noisy* data and fine-tuned regularization, while [12, 13] focused on noiseless data sets. This observation raises a number of questions: is there a connection between both sets of exponents? Are Gaussian design exponents actually different from worst-case ones? What about intermediary setups (for instance noisy labels with generic regularization, noiseless labels with varying regularization) and regimes (intermediary sample complexities)? How does infinitesimal noise differ from no noise at all?

### 1.1. Main contributions

In this manuscript, we answer all the above questions, and redeem the apparent contradiction by reconsidering the Gaussian design analysis. We provide a unifying picture of the decay rates for the excess generalization error, along a more exhaustive characterization of the regimes in which each is observed, evidencing the interplay of the role of regularization, noise and sample complexity. We show in particular that typical-case analysis with a Gaussian design is actually in perfect agreement with the statistical learning worst-case data-agnostic approach. We also show how the optimal excess error decay can transition from the recently reported noiseless value to its well known noisy value as the number of samples is increased. We illustrate this crossover from the *noiseless* regime to the *noisy* regime also in a variety of KRR experiments on real data.

### 1.2. Related work

The analysis for kernel methods and ridge regression is a classical topic in statistical learning theory [10, 11, 14–17]. In this classical setting, decay exponents for optimally regularized *noisy* linear regression on features with power-law co-variance spectrum have been provided. Interestingly, it has been shown that such optimal rates can be obtained in practice by stochastic gradient descent, without explicit regularization, with single-pass [18, 19] or multi-pass [8], as well as by randomized algorithms [20]. Closed-form bounds for the prediction error have been provided in a number of worst-case analyses [16, 20]. We show how the decay rates given in the present paper can also be alternatively deduced therefrom in appendix E.

The recent line of work on the noiseless setting includes contributions from statistical learning theory [9, 21] and statistical physics [12, 13]. This much more recent second line of work proved decay rates for a given, constant regularization. An example of noise-induced crossover is furthermore mentioned in [9]. The interplay between noisy and noiseless regimes has also been investigated in the related Gaussian process literature [22].

The study of ridge regression with Gaussian design is also a classical topic. Reference [23] considered a model in which the covariates are isotropic Gaussian in $\mathbb{R}^p$, and computed the exact asymptotic generalization error in the high-dimensional asymptotic regime $p, n \to \infty$ with dimension-to-sample-complexity ratio $p/n$ fixed. This result was generalised to arbitrary co-variances [24, 25] using fundamental results from random matrix theory [26]. Non-asymptotic rates of convergence for a related problems were given in [27]. Previous results also existed in the statistical physics literature, e.g. [28–31]. Gaussian models for regression have seen a surge of popularity recently, and have been used in particular to study over-parametrization and the double-descent phenomenon, e.g. in [17, 32–44].

## 2. Setting

Consider a data set $\mathcal{D} = \{x^\mu, y^\mu\}_{\mu=1}^n$ with $n$ independent samples from a probability measure $\nu$ on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ is the input and $\mathcal{Y} \subset \mathbb{R}$ the response space. Let $K$

https://doi.org/10.1088/1742-5468/ac9829

be a kernel and $\mathcal{H}$ denote its associated reproducing kernel Hilbert space (RKHS). KRR corresponds to the following non-parametric minimisation problem:

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{\mu=1}^{n}\left(f(x^\mu)-y^\mu\right)^2+\lambda\|f\|_{\mathcal{H}}^2. \tag{1}$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm associated with the scalar product in $\mathcal{H}$, and $\lambda \geqslant 0$ is the regularisation. The convenience of KRR is that it admits a dual representation in terms of a standard parametric problem. Indeed, the kernel $K$ can be diagonalized in an orthonormal basis $\{\phi_k\}_{k=1}^{\infty}$ of $L^2(\mathcal{X})$:

$$\int_{\mathcal{X}} \nu_x(dx')K(x,x')\phi_k(x') = \eta_k\phi_k(x) \tag{2}$$

where $\{\eta_k\}_{k=1}^{\infty}$ are the corresponding (non-negative) kernel eigenvalues and $\nu_x$ is the marginal distribution over $\mathcal{X}$. Note that the kernel $\{\phi_k\}_{k=1}^{\infty}$ eigenvectors form an orthonormal basis of $L^2(\mathcal{X})$. It is convenient to define the re-scaled basis of *kernel features* $\psi_k(x) = \sqrt{\eta_k}\phi_k(x)$ and to work in matrix notation in feature space: define $\phi(x) \equiv \{\phi_k(x)\}_{k=1}^{p}$ (with $p$ possibly infinite)

$$\psi(x) = \Sigma^{\frac{1}{2}}\phi(x) \quad \mathbb{E}_{x\sim\nu_x}\left[\phi(x)\phi(x)^\top\right] = \mathbb{1}_p, \quad \mathbb{E}_{x'\sim\nu_x}[K(x,x')\phi(x')] = \Sigma\phi(x), \tag{3}$$

where $\Sigma \equiv \mathbb{E}_{x\sim\nu_x}\left[\psi(x)\psi(x)^\top\right] = \mathrm{diag}(\eta_1,\eta_2,\ldots,\eta_p)$ is the features co-variance (a diagonal operator in feature space). In this notation, the RKHS $\mathcal{H}$ can be formally written as $\mathcal{H} = \{f = \psi^\top\theta : \theta\in\mathbb{R}^p, \quad \|\theta\|_2 < \infty\}$, i.e. the space of functions for which the coefficients in the feature basis are square summable. With this notation, we can rewrite equation (4) in feature space as a standard parametric problem for the following empirical risk:

$$\hat{\mathcal{R}}_n(w) = \frac{1}{n}\sum_{\mu=1}^{n}\left(w^\top\psi(x^\mu)-y^\mu\right)^2+\lambda w^\top w. \tag{4}$$

Our main results concern the typical averaged performance of the KRR estimator, as measured by the typical prediction (out-of-sample) error

$$\epsilon_{\mathrm{g}} = \mathbb{E}_{\mathcal{D}}\mathbb{E}_{(x,y)\sim\nu}(\hat{f}(x)-y)^2, \tag{5}$$

where the first average is over the data $\mathcal{D} = \{x^\mu, y^\mu\}$ and the second over a fresh sample $(x, y) \sim \nu$.

In what follows we assume the labels $y^\mu \in \mathcal{Y}$ were generated, up to an independent additive Gaussian noise with variance $\sigma^2$, by a target function $f^\star$ (not necessarily belonging to $\mathcal{H}$):

$$y^\mu \overset{d}{=} f^\star(x^\mu) + \sigma\mathcal{N}(0,1), \tag{6}$$

and we denote by $\theta^\star$ the coefficients of the target function in the features basis $f^\star(x) = \psi(x)^\top\theta^\star$. As we will characterize below, whether the target function $f^\star$ belongs or not to $\mathcal{H}$ depends on the relative decay coefficients $\theta^\star$ with respect to the eigenvalues of the kernel. We often refer to $\theta^\star$ as the *teacher*. While the present results and discussion are provided for additive Gaussian noise for simplicity, our method are not restricted to this particular noise, and a more complete extension of the results for other noise settings is left for future work.

We are then interested in the evolution of the *excess error* $\epsilon_{\mathrm{g}} - \sigma^2$ as the number of samples $n$ is increased.

### 2.1. Capacity and source coefficients

Motivated by the discussion above, we focus on ridge regression in an infinite dimensional ($p \to \infty$) space $\mathcal{H}$ with Gaussian design $u^\mu \overset{\mathrm{def}}{=} \psi(x^\mu) \overset{d}{=} \mathcal{N}(0, \Sigma)$ with (without loss of generality) diagonal co-variance $\Sigma = \mathrm{diag}(\eta_1, \eta_2, \ldots)$. We expect however the results of this manuscript to be universal for a large class of distribution beyond the Gaussian one. In particular, we anticipate the Gaussianity assumption should be amenable to being relaxed to sub-gaussians [45] or even any concentrated distribution [46, 47].

Following the statistical learning terminology, we introduce two parameters $\alpha > 1, r \geqslant 0$, herefrom referred to as the *capacity* and *source* conditions [14], to parametrize the difficulty of the target function and the learning capacity of the kernel

$$\mathrm{tr}\,\Sigma^{\frac{1}{\alpha}} < \infty, \qquad \|\Sigma^{\frac{1}{2}-r}\theta^\star\|_{\mathcal{H}} < \infty. \tag{7}$$

As in [9, 12, 13, 25], we consider the particular case where both the spectrum of $\Sigma$ and the teacher components $\theta^\star_k$ have exactly a power-law form satisfying the limiting source/capacity conditions (7):

$$\eta_k = k^{-\alpha}, \qquad \theta^\star_k = k^{-\frac{1+\alpha(2r-1)}{2}}. \tag{8}$$

The power law ansatz (8) is empirically observed to be a rather good approximation for some real simple datasets and kernels, see figure 7 in appendix C. The parameters $\alpha, r$ introduced in (8) control the complexity of the data the teacher respectively. A large $\alpha$ can be loosely seen as characterizing a effectively low dimensional (and therefore easy to fit) data distribution. By the same token, a large $r$ signals a good alignment of the teacher with the important directions in the data covariance, and therefore an *a priori* simpler learning task.

The regularization $\lambda$ is allowed to vary with $n$ according to a power-law $\lambda = n^{-\ell}$. This very general form allows us to encompass both the zero regularization case (corresponding to $\ell = \infty$) and the case where $\lambda = \lambda^\star$ is optimized, with some optimal decay rate $\ell^\star$. Note that this power law form implies that $\lambda$ is assumed positive. While this is indeed the assumption of [10, 14] with which we intend to make contact,Wu and Xu [39] have shown that the optimal $\lambda$ may in some settings be negative. Some numerical experiments suggest that removing the positivity constraint on $\lambda$ while optimizing does not affect the results presented in this manuscript. A more detailed investigation is left to future work.

## 3. Main results

Depending on the regularization decay strength $\ell$, capacity $\alpha$, source $r$ and noise variance $\sigma^2$, four regimes can be observed. The derivation of these decays from the asymptotic solution of the Gaussian design problem is sketched in section 4 and detailed in appendix A, and here we concentrate on the key results. The different observable decays for the excess error $\epsilon_g - \sigma^2$ are summarized in figure 1, and are given by:

- If $\ell \geqslant \alpha$ (weak regularization $\lambda = n^{-\ell}$),

$$\epsilon_{\mathrm{g}} - \sigma^2 = \mathcal{O}\big(\max\big(\sigma^2, n^{-2\alpha\ \min(r,1)}\big)\big). \tag{9}$$

  The excess error transitions from a fast decay $2\alpha\ \min(r,1)$ (green region in figure 1 and green dashed line in figure 2) to a plateau (red region in figure 1 and red dashed line in figure 2) with no decay as $n$ increases. This corresponds to a crossover from the green region to the red region in the phase diagram figure 1.

- If $\ell \leqslant \alpha$ (strong regularization $\lambda = n^{-\ell}$),

$$\epsilon_{\mathrm{g}} - \sigma^2 = \mathcal{O}\left(\max\left(\sigma^2, n^{1-2\ell\ \min(r,1)-\frac{\ell}{\alpha}}\right) n^{\frac{\ell-\alpha}{\alpha}}\right). \tag{10}$$

  The excess error transitions from a fast decay $2\ell\ \min(r,1)$ (blue region in figure 1) to a slower decay $(\alpha-\ell)/\alpha$ (orange region in figure 1) as $n$ is increased and the effect of the additive noise kicks in, see figure 3. The crossover disappears for too slow decays $l \leqslant \alpha/(1+2\alpha\min(r,1))$, as the regularization $\lambda$ is always sufficiently large to completely mitigate the effect of the noise. This corresponds to the max in (10) being realized by its second argument for all $n$.

Given these four different regimes as depicted in figure 1, one may wonder about the optimal learning solution when the regularization is fine tuned to its best value. To answer this question, we further define the *asymptotically optimal* regularization decay $\ell^\star$ as the value leading to fastest decay of the typical excess error $\epsilon_{\mathrm{g}} - \sigma^2$. We find that two different optimal rates exist, depending on the quantity of data available.

- If $n \ll n_1^* \approx \sigma^{-\frac{1}{\alpha\ \min(r,1)}}$, any $\ell^\star \in (\alpha, \infty)$ yields excess error decay

$$\epsilon_{\mathrm{g}}^\star - \sigma^2 \sim n^{-2\alpha\ \min(r,1)}. \tag{11}$$

- If $n \gg n_2^* \approx \sigma^{-\max\left(2,\frac{1}{\alpha\ \min(r,1)}\right)}$,

$$\epsilon_{\mathrm{g}}^\star - \sigma^2 \sim n^{\frac{1}{1+2\alpha\ \min(r,1)}-1}, \text{ by choosing } \quad \lambda^\star \sim n^{-\frac{\alpha}{1+2\alpha\ \min(r,1)}}. \tag{12}$$

  The optimal decay for the excess error $\epsilon_{\mathrm{g}}^\star - \sigma^2$ thus transitions from a fast decay $2\alpha\ \min(r,1)$ when $n \ll n_1^*$—corresponding to, effectively, the optimal rates expected in a 'noiseless' situation—to a slower decay $2\alpha\ \min(r,1)/(1+2\alpha\ \min(r,1))$ when $n \gg n_2^*$ corresponding to the classical 'noisy' optimal rate, depicted with the purple point in figure 1. This is illustrated in figure 4 where the two rates are observed in succession for the same data as the number of points is increased.

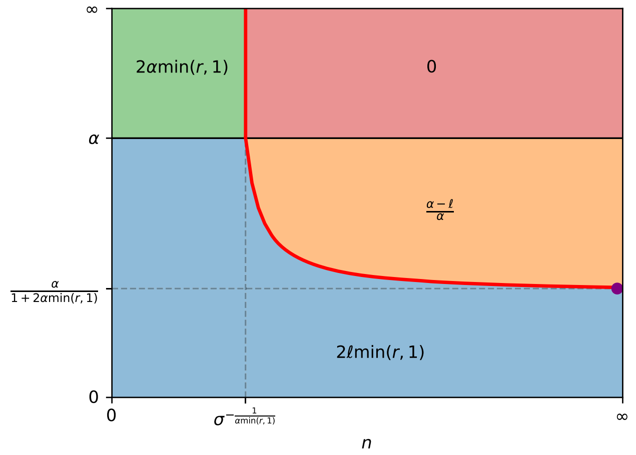

**Figure 1.** Different decays for the excess generalization error $\epsilon_g - \sigma^2$ for different values of $n$ and different decays $\ell$ of the regularization $\lambda \sim n^{-\ell}$, at given noise variance $\sigma$. The red solid line represents the noise-induced crossover line, separating the effectively noiseless regime (green and blue) on its left from the effectively noisy regime (red and orange) on its right. Any KRR experiment at fixed regularization decay $\ell$ (corresponding to drawing a horizontal line at ordinate $\ell$) crosses the crossover line if $\ell > \alpha/(1 + 2\alpha \min(r, 1))$. The corresponding learning curve will accordingly exhibit a crossover from a fast decay (noiseless regime) to a slow decay (noisy regime).

We can now finally clarify the apparent discrepancy in the recent literature discussed in the introduction. The exponent recently reported in [12, 13] actually corresponds to the 'noiseless' regime. In contrast, the rate described in (12) is the classical result [10] for the non-saturated case $r < 1$ for generic data. We see here that the same rate is also achieved with Gaussian design, and that there are no differences between fixed and Gaussian design as long as the capacity and source condition are matching. We unveiled, however, the existence of two possible sets of optimal rate exponents depending on the number of data samples.

All setups (effectively non-regularized KRR (9), effectively regularized KRR (10) or optimally regularized KRR (11) and (12)) can therefore exhibit a *crossover* from an effectively *noiseless* regime (green or blue in figure 1), to an effectively *noisy* regime (red, orange in figure 1) depending on the quantity of data available. We stress that while the noise is indeed present in the green and blue 'noiseless' regimes, its presence is effectively not felt, and noiseless rates are observed. In fact, if the noise is small, one will not observed the classical noisy rates unless an astronomical amount of data is available. This can be intuitively understood as follows: for small sample size $n$, low-variance dimensions are used to overfit the noise, while the spiked subspace of large-variance dimensions is well fitted. In noiseless regions, the excess error is thus characterized by a fast decay. This phenomenon, where the noise variance is diluted over the dimensions of lesser importance, is connected to the *benign overfitting* discussed by [17, 45]. Benign overfitting is possible due to the decaying structure of the co-variance spectrum (8). As

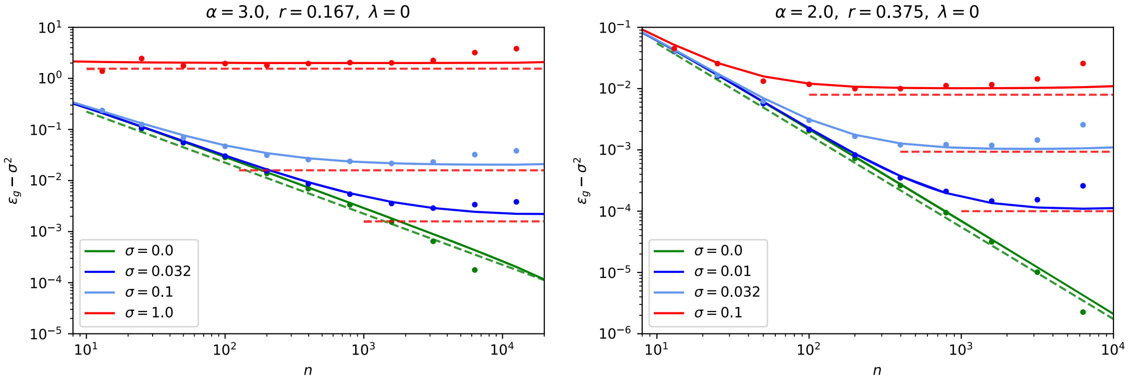

**Figure 2.** KRR on synthetic data sets with capacity $\alpha$ and source coefficient $r$ i.e. the idealized Gaussian setting (8), with no regularization $\lambda = 0$. Solid lines correspond to the theoretical prediction of equation (14) using the `GCM` package associated with [27]. Points are simulations conducted using the python `scikit-learn KernelRidge` package [48], where the feature space dimension has been cut off to $p = 10^4$ for the simulations, and to $10^5$ for the theoretical curves. Dashed lines represent the slopes predicted by equation (9), with the color (red and green) in correspondence to the regime from figure 1.

more samples are accessed, further decrease of the excess error requires good generalization also over the low-variance subspace, and the overfitting of the noise results in a slower decay.

While our analysis is for the optimal full-batch learning, we note that a similar crossover in the case of SGD in the effectively non-regularized case (from green to red) has been discussed in [9, 21]. It would be interesting to further explore how SGD can behave in the different regimes discussed here.

When $\lambda = \lambda_0 n^{-\ell}$ for a prefactor $\lambda_0$ that is allowed to be very small, a *regularization-induced* crossover, similar to the one reported in [13], can also be observed on top of the noise-induced crossover which is the focus of the present work. This setting is detailed in appendix. D.

## 4. Sketch of the derivation

We provide in this section the main ideas underlying the derivation of the main results exposed in section 3 and summarized in figure 1. A more detailed discussion is presented in appendix A.

### 4.1. Closed-form solution for Gaussian design

Closed-form, rigorous solution of the risk of ridge regression with Gaussian data of arbitrary co-variance in the high-dimensional asymptotic regime have been studied in [25, 39, 40]. We shall use here the equivalent notations of [27], who have the advantage of having rigorous non-asymptotic rates guarantees. Using these characterizations as a

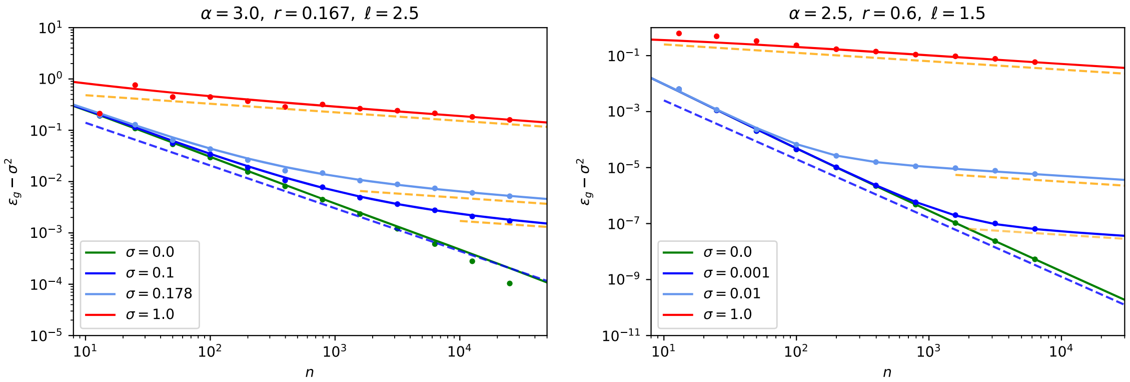

**Figure 3.** KRR on synthetic data sets with capacity $\alpha$ and source coefficient $r$, with regularization $\lambda = n^{-\ell}$. Solid lines correspond to the theoretical prediction of equation (14) using the `GCM` package associated with [27]. Points are simulations conducted using the python `scikit-learn KernelRidge` package [48], where the feature space dimension has been cut off to $p = 10^4$ for the simulations, and to $10^5$ for the theoretical curves. Dashed lines represent the slopes predicted by equation (10), with the color (blue and orange) in correspondence to the regime from figure 1.

starting point, we shall sketch how the crossover phenomena (9)–(12), which are the main contribution of this paper, can be derived. Within the framework of [27], with high-probability when $n, p$ are large the excess prediction error is expressed as

$$\epsilon_{\mathrm{g}} - \sigma^2 = \rho - 2m^\star + q^\star, \tag{13}$$

with $\rho = \theta^{\star\top}\Sigma\theta^\star$, and $(m^\star, q^\star)$ are the unique fixed-points of the following self-consistent equations:

$$\begin{cases} \hat{V} = \dfrac{\frac{n}{p}}{1+V} \\ \hat{q} = \dfrac{n}{p}\dfrac{\rho + q - 2m + \sigma^2}{(1+V)^2} \end{cases}, \quad \begin{cases} q = p\displaystyle\sum_{k=1}^{p} \frac{\hat{q}\,\eta_k^2 + \theta_k^{\star 2}\eta_k^2\,\hat{m}^{\,2}}{(n\lambda + p\,\hat{V}\,\eta_k)^2} \\ m = p\,\hat{V}\displaystyle\sum_{k=1}^{p} \frac{\theta_k^{\star 2}\eta_k^2}{n\lambda + p\,\hat{V}\,\eta_k} \end{cases}, \quad \left\{ V = \frac{1}{p}\sum_{k=1}^{p} \frac{p\eta_k}{n\lambda + p\,\hat{V}\,\eta_k} \right. . \tag{14}$$

We recall the reader that $\lambda > 0$ is the regularisation strength and $\{\eta_k\}_{k=1}^p$ are the kernel eigenvalues. The next step is thus to insert the power-law decay (8) for the eigenvalues into (14), and to take the limit $n, p \to \infty$. We note, however, that this last step is not completely justified rigorously. Indeed, [25] assumes $p/n = O(1)$ as $n, p \to \infty$ while here we first send $p \to \infty$ and then take the large $n$ limit, thus working effectively with $p/n \to 0$. While the non-asymptotic rates guarantees of [27] are reassuring in this respect, a finer control of the limit would be needed for a fully rigorous justification. Nevertheless, we observed in our experiments that the agreement between theory and numerical simulations for the excess prediction error (5) is perfect (see figures 2–4). In the large $n$ limit, one can finally close the equation for the excess prediction error into

https://doi.org/10.1088/1742-5468/ac9829

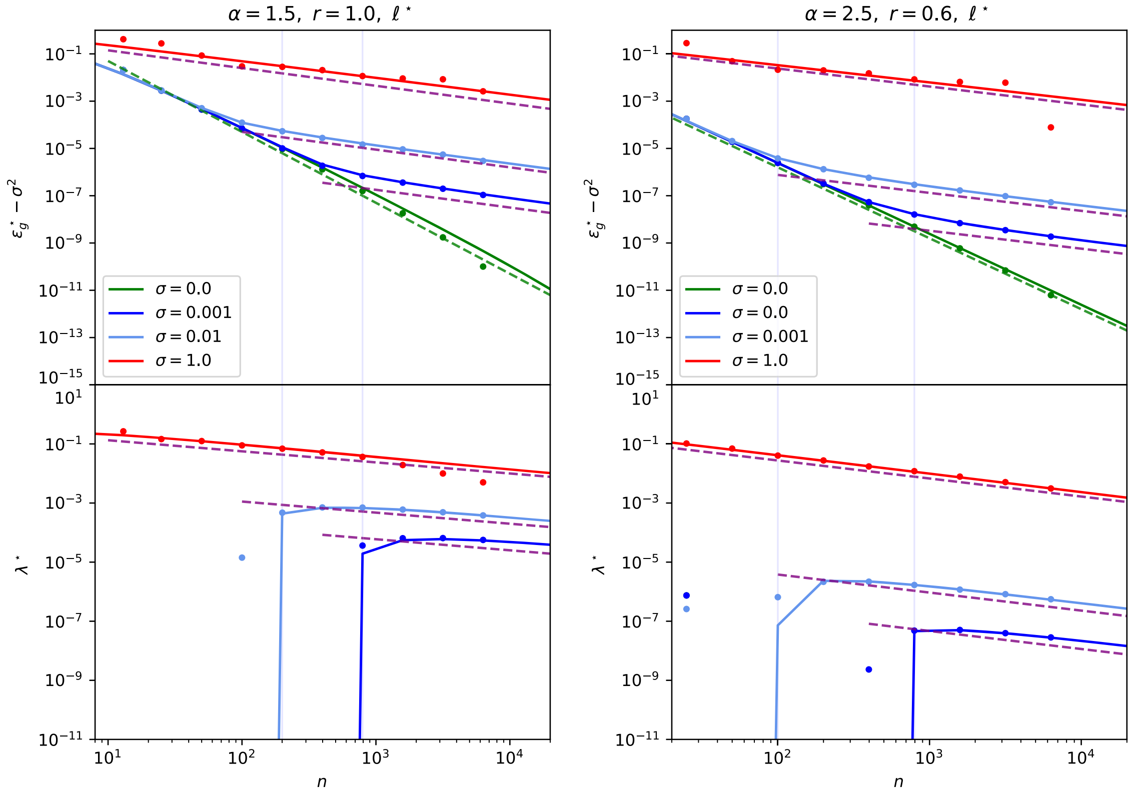

**Figure 4.** KRR on synthetic data sets with capacity $\alpha$ and source coefficient $r$. The regularization $\lambda$ is chosen as the one minimizing the theoretical prediction for the excess generalization error, deduced from equation (14) using the `GCM` package associated with [27]. Solid lines correspond to the theoretical prediction of equation (14). Points are simulations conducted with the python `scikit-learn KernelRidge` package [48], where the feature space dimension has been cut off to $p = 10^4$ for the simulations, and to $10^5$ for the theoretical curves. In simulations, the best $\lambda^\star$ was determined using python `scikit-learn GridSearchCV` cross validation package [48]. Note that because cross validation is not adapted to small training sets, a few discrepancies are observed for smaller $n$. Dashed lines represent the slopes predicted by theory, with the colors in correspondence to the regimes in figure 1, purple for the purple point in figure 1. (Top) Excess error. (Bottom) Optimal $\lambda^\star$. Note the noiseless case has $\lambda^* = 0$.

$$\epsilon_{\mathrm{g}} - \sigma^2 = \frac{\sum_{k=1}^{\infty} \frac{k^{-1-2r\alpha}}{(1+nz^{-1}k^{-\alpha})^2}}{1 - \frac{n}{z^2}\sum_{k=1}^{\infty} \frac{k^{-2\alpha}}{(1+nz^{-1}k^{-\alpha})^2}} + \sigma^2 \frac{\frac{n}{z^2}\sum_{k=1}^{\infty} \frac{k^{-2\alpha}}{(1+nz^{-1}k^{-\alpha})^2}}{1 - \frac{n}{z^2}\sum_{k=1}^{\infty} \frac{k^{-2\alpha}}{(1+nz^{-1}k^{-\alpha})^2}}. \tag{15}$$

with $z$ being a solution of

$$z \approx n\lambda + \left(\frac{z}{n}\right)^{1-\frac{1}{\alpha}} \int_{\left(\frac{z}{n}\right)^{1/\alpha}}^{\infty} \frac{dx}{1 + x^\alpha}. \tag{16}$$

The detailed derivation is provided in appendix A. We note that this equation was observed with heuristic arguments from statistical physics (using the non-rigorous cavity method) in [49].

The different regimes of excess generalization error rates discussed in section 3 are derived from this self-consistent equation. Note that the excess error (15) decomposes over a sum of two contributions, respectively accounting for the sample variance and the noise-induced variance. In contrast to a typical bias-variance decomposition, the effect of the bias introduced in the task for non-vanishing $\lambda$ is subsumed in both terms.

### 4.2. Derivation of the four regimes

If the second term in (16) dominates, then $z \sim n^{1-\alpha}$, which is self consistent if $\ell \geqslant \alpha$. This is the *effectively non-regularized regime*, where the regularization $\lambda$ is not sensed, and corresponds to the green and red regimes in the phase diagram in figure 1. This scaling of $z$ can then be used to estimate the asymptotic behaviour of the sample and noise induced variance in the decomposition on the excess error (15), yielding

$$\epsilon_{\mathrm{g}} - \sigma^2 = \mathcal{O}(n^{-2\alpha\ \min(r,1)}) + \sigma^2\mathcal{O}(1), \tag{17}$$

which can be rewritten more compactly as (9). Therefore, for small sample sizes the sample variance drives the decay of the excess prediction error, while for larger samples sizes the noise variance dominates and causes the error to plateau. The crossover happens when both variance terms in (17) are balanced, around

$$n \sim \sigma^{-\frac{1}{\alpha\ \min(r,1)}}, \tag{18}$$

which corresponds to the vertical part of the crossover line in figure 1.

If the first term $n\lambda$ dominates in (16), then $z \sim n\lambda$, which is consistent provided that $\ell < \alpha$. This is the *effectively regularized regime* (blue, orange regions in figure 1). The two variances in (15) are found to asymptotically behave like

$$\epsilon_{\mathrm{g}} - \sigma^2 = \mathcal{O}(n^{-2\ell\ \min(r,1)}) + \sigma^2\mathcal{O}\left(n^{\frac{\ell-\alpha}{\alpha}}\right), \tag{19}$$

which can be rewritten more compactly as (10). If the decay of the noise variance term $(\alpha - \ell)/\alpha$ is faster than the $2\ell\ \min(r, 1)$ decay of the sample variance term, then the latter always dominates and no crossover is observed. This is the case for $\ell < \alpha/(1 + 2\alpha \min(r, 1))$. If on the contrary the decay of the noise variance term is the slowest, then this term dominates at larger $n$, with a crossover when both terms in (19) are balanced, around

$$n \sim \sigma^{\frac{2}{1-\frac{\ell}{\alpha}(1+2\alpha\ \min(r,1))}}. \tag{20}$$

Equations (17) and (19) are respectively equivalent to (9) and (10), and completely define the four regimes observable in figure 1. Equations (20) and (18) give the expression for the crossover line in figure 1.

### 4.3. Asymptotically optimal regularization

Determining the asymptotically optimal $\ell^\star$ is a matter of finding the $\ell$ leading to fastest excess error decay. We focus on the far left part and the far right part of the phase diagram figure 1.

In the $n \gg n_2^\star \approx \sigma^{-\max\left(2, \frac{1}{\alpha\ \min(r,1)}\right)}$ limit where the crossover line confounds itself with its $\ell = \alpha/(1 + 2\alpha \min(r, 1))$ asymptote, this is tantamount to solving the maximization problem

$$\ell^\star = \underset{\ell}{\mathrm{argmax}} \left( 2\ell \min(r, 1) \mathbb{1}_{0<\ell<\frac{\alpha}{(1+2\alpha\ \min(r,1))}} + \frac{\alpha - \ell}{\alpha} \mathbb{1}_{\frac{\alpha}{(1+2\alpha\ \min(r,1))}<\ell<\alpha} + 0 \times \mathbb{1}_{\alpha<\ell} \right) \tag{21}$$

which admits as solution (12). In the $n \ll n_1^\star \approx \sigma^{-\frac{1}{\alpha\ \min(r,1)}}$ range, the maximization of the excess error decay reads

$$\ell^\star = \underset{\ell}{\mathrm{argmax}} \, (2\ell\ \min(r, 1) \mathbb{1}_{0<\ell<\alpha} + 2\alpha\ \min(r, 1) \mathbb{1}_{\alpha<\ell}), \tag{22}$$

and admits as solution (11).

## 5. Illustration on simple real data sets

In this section we show that the derived decay rates can indeed be observed in real data sets with labels artificially corrupted by additive Gaussian noise. For real data, the decay model in equation (8) is idealized, and in practice there is no firm reason to expect a power-law decay. However, we do find that for some of the data sets and kernels we investigated, the power law fit is reasonable and can be used to estimate the exponents $\alpha$ and $r$, see appendix C for details. For those cases, we compare the theoretically predicted exponents, equations (9)–(12) with the empirically measured learning curve, and obtain a very good agreement. We stress that the decay rates are not obtained by fitting the learning curves, but rather by fitting the exponents $\alpha$ and $r$ from the data. We also observe the crossover from the noiseless (blue, green in figure 1) to the noisy (orange, red in figure 1) regime given by the theory. Here we illustrate this with the learning curves for the following three data sets:

- MNIST even versus odd, a data set of $7 \times 10^4$ $28 \times 28$ images of handwritten digits. Even (odd) digits were assigned label $y = 1 + \sigma\mathcal{N}(0, 1)$ ($y = -1 + \sigma\mathcal{N}(0, 1)$).
- Fashion MNIST t-shirts versus coats, a data set of 14 70228 × 28 images of clothes from an online shopping platform [50]. T-shirts (coats) were assigned label $y = 1 + \sigma\mathcal{N}(0, 1)$ ($y = -1 + \sigma\mathcal{N}(0, 1)$).
- Superconductivity [51], a data set of 81 attributes of 21 263 superconducting materials. The target $y^\mu$ corresponds to the critical temperature of the material, corrupted by additive Gaussian noise.

Learning curves are illustrated for a radial basis function (RBF) kernel $K(x, x') = \mathrm{e}^{-\frac{\gamma}{2}\|x-x'\|^2}$ with parameter $\gamma = 10^{-4}$ and a degree 5 polynomial kernel $K(x, x') = (1 + \gamma\langle x, x'\rangle)^5$ with parameter $\gamma = 10^{-3}$. In figure 5 the regularization $\lambda$ was set

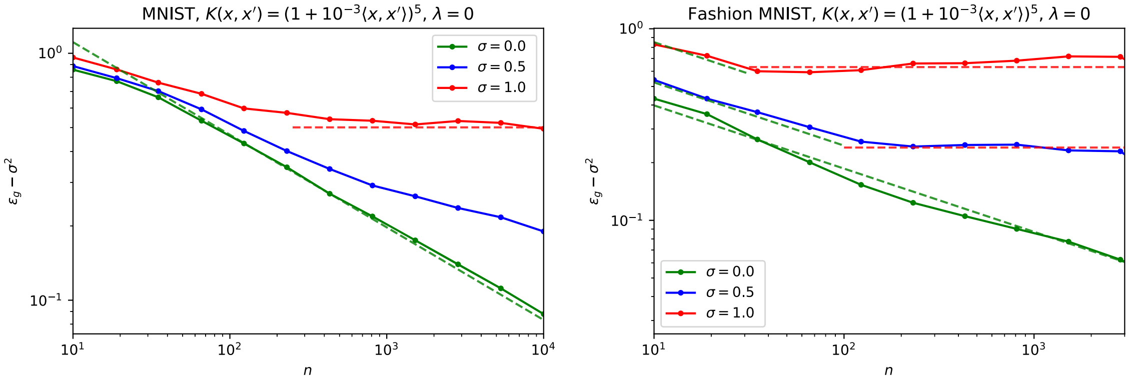

**Figure 5.** Excess error for MNIST odd versus even (above) and Fashion MNIST t-shirt versus coat (below) with labels corrupted by noise of variance $\sigma^2$. The kernel used is indicated in the title. Solid lines with points come from numerical experiments with zero regularization. Dashed lines are the slopes $-2\alpha r$ (as $r < 1$) or 0, predicted by the theory from the empirical values of $\alpha, r$ measured from the Gram matrix spectrum and the teacher for each data set, see table 1. Colors of the dashed lines (green and red) indicate the regimes in figure 1.

to 0, while in figure 6 $\lambda$ was optimized for each sample size $n$ using the python `scikit-learn GridSearchCV` package [48]. KRR was carried out using the `scikit-learn KernelRidge` package [48]. The values of $\alpha, r$ were independently measured (see appendix C) for each data set, and the estimated values summarized in table 1. From these values the theoretical decays (9), (11) and (12) were computed, and compared with the simulations with very good agreement. Since for real data the power-law form (8) does not exactly hold (see figure 7 in the appendix), the estimates for $\alpha, r$ slightly vary depending on how the power-law is fitted. The precise procedure employed is described in appendix C. Overall this variability does not hurt the good agreement with the simulated learning curves in figures 5 and 6.

When $\lambda = 0$ (figure 5) the characteristic plateau for large label noises is observed for both MNIST and Fashion MNIST. For polynomial kernel regression on Fashion MNIST (figure 5 right), the crossover between noiseless (slope $-2\alpha r$ as $r < 1$) and noisy (slope 0) regimes is apparent on the same learning curve at noise levels $\sigma = 0.5, 1$. For MNIST, the $\sigma = 0$ ($\sigma = 1$) curve is in the noiseless (noisy) regime for larger $n$, while at intermediary noise $\sigma = 0.5$, and small $n$ for $\sigma = 1$, the curve is in the crossover regime between noiseless and noisy, consequently displaying in-between decay. Our results for the decays for $\sigma = 0$ agree with simulations for RBF regression on MNIST provided in [12].

For optimal regularization $\lambda = \lambda^\star$ (figure 6), as the measured $r < 1$ we have exponents $-2r\alpha$ for the noiseless regime and $-2r\alpha/(1 + 2r\alpha)$ for noisy. Since the measured value of $2r\alpha$ is rather small the difference between the two rates is less prominent. Nevertheless, it seems that in our experiments the noisy regime is observed for polynomial and RBF kernels on MNIST and $\sigma = 0.5, 1$. For Superconductivity, the green and purple decay have close values and it is difficult to clearly identify the regime. For Fashion MNIST only the noiseless rate is observable in the considered noise range and sample range.

https://doi.org/10.1088/1742-5468/ac9829

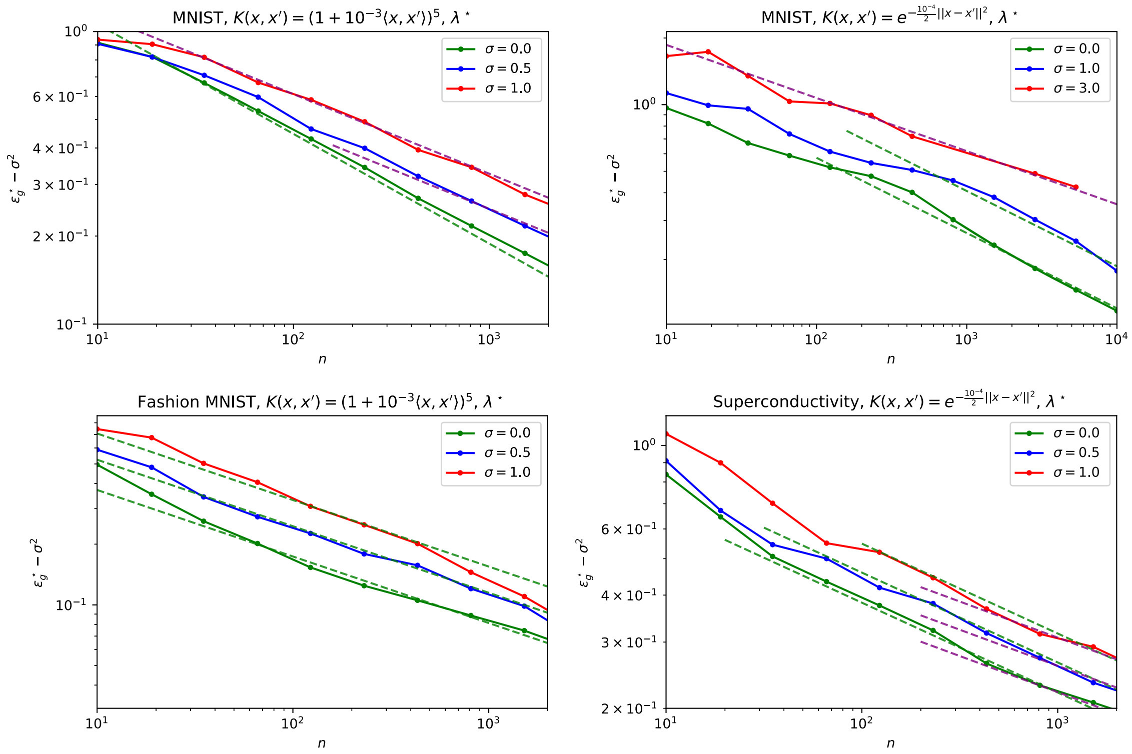

**Figure 6.** Excess error for MNIST odd versus even, and Fashion MNIST t-shirt versus coat, and the critical temperature regression. The kernel used is indicated in the title. Solid lines with dots come from numerical experiments with the regularization optimized using the python `scikit-learn GridSearchCV` package [48]. Dashed lines are the slopes predicted by the theory, from the empirical values of $\alpha, r$ measured from the Gram matrix spectrum and the teacher for each data set, see table 1. Colors of the dashed lines indicate the regime in figure 1.

**Table 1.** Values of the source and capacity coefficients (7) as estimated from the data sets. The details on the estimation procedure can be found in appendix C.

| Dataset | Kernel | $\alpha$ | $r$ |
|---|---|---|---|
| Fashion MNIST | $K(x,x') = (1+10^{-3}\langle x,x'\rangle)^5$ | 1.3 | 0.13 |
| MNIST | $K(x,x') = (1+10^{-3}\langle x,x'\rangle)^5$ | 1.2 | 0.15 |
| MNIST | $K(x,x') = \exp(-10^{-4}\|x-x'\|^2/2)$ | 1.65 | 0.097 |
| Superconductivity | $K(x,x') = \exp(-10^{-4}\|x-x'\|^2/2)$ | 2.7 | 0.046 |

## 6. Conclusion

To conclude, we unify hitherto disparate lines of work, and give a comprehensive study of observable regimes, along the associated decay rates for the excess error, for KRR with features having power-law co-variance spectrum. We show that the effect of the noise only kicks in at larger sample complexity, meaning, in particular, that the KRR transitions from a *noiseless* regime with fast error decay to a *noisy* regime with slower decay. This crossover is shown to happen for zero, decaying and optimized regularization, and is observed on a variety of real data sets corrupted with label noise.

## Acknowledgments

We warmly thank Loucas Pillaud-Vivien for discussions, and for his help in navigating the literature on KRR. We also thank Matthieu Wyart and Cedric Gerbelot for discussions. We acknowledge funding from the ERC under the European Union's Horizon 2020 Research and Innovation Programme Grant Agreement 714608-SMiLe, and from the French National Research Agency Grants ANR-17-CE23-0023-01 PAIL.

## Appendix A. Derivation of the decays

### A.1. Equations for Gaussian design

In this appendix we discuss the derivation of equations (13) and (14) describing the excess prediction error for the ridge regression problem with generic covariance matrix. Exact asymptotic formulas for the excess prediction error of least-squares and ridge regression are a classic result in high-dimensional statistics, and have been derived in many different works [25, 34, 52, 53]. In this manuscript, we follow the presentation given in [27], which is particularly adapted to our derivation and has the advantage to hold rigorously at large but finite number of samples $n$ and features $p$. We start by reviewing the formulas in [27]. Consider the ridge regression problem on $n$ independent $p$-dimensional samples $\{u^\mu, y^\mu\}_{\mu=1}^n$, defined by a minimisation of the following empirical risk:

$$\hat{\mathcal{R}}_n(w) = \sum_{\mu=1}^{n} \left( \frac{w \cdot u^\mu}{\sqrt{p}} - y^\mu \right)^2 + \lambda \|w\|_2^2. \tag{A1}$$

Assume a Gaussian design $u^\mu \overset{d}{=} \mathcal{N}(0, \Sigma)$ with diagonal covariance $\Sigma = \mathrm{diag}(\eta_1, \ldots, \eta_p)$ and labels $y^\mu$ generated from a teacher/target/oracle $\theta^\star \in \mathbb{R}^p$:

$$y^\mu = \frac{\theta^\star \cdot u^\mu}{\sqrt{p}} + \sigma \mathcal{N}(0, 1). \tag{A2}$$

Under the assumptions

(A1) $n \gg 1, p \gg 1, n/p = \mathcal{O}(1)$,

(A2) $0 < \|\theta^\star\|^2/p < \infty$,

there exists constants $C$, $c$, $c' > 0$ such that for all $0 < \epsilon < c'$,

$$\mathbb{P}\left(|\epsilon_{\mathrm{g}} - \sigma^2 - (\rho - 2m^\star + q^\star)| > \epsilon\right) < \frac{C}{\epsilon}\,\mathrm{e}^{-cn\epsilon^2}. \tag{A3}$$

where $\rho = \theta^\star \cdot \Sigma \cdot \theta^\star / p$, and $(m^\star, q^\star)$ are fixed-points of the following *self-consistent equations*

$$\begin{cases} \hat{V} = \hat{m} = \dfrac{\frac{n}{p}}{1+V} \\ \hat{q} = \dfrac{n}{p}\dfrac{\rho + q - 2m + \sigma^2}{(1+V)^2} \end{cases}, \begin{cases} V = \dfrac{1}{p}\displaystyle\sum_{k=1}^{p} \dfrac{\eta_k}{\lambda + \hat{V}\,\eta_k} \\ q = \dfrac{1}{p}\displaystyle\sum_{k=1}^{p} \dfrac{\hat{q}\eta_k^2 + \theta_k^{\star 2}\eta_k^2\,\hat{m}^{\,2}}{(\lambda + \hat{V}\,\eta_k)^2} \\ m = \dfrac{\hat{m}}{p}\displaystyle\sum_{k=1}^{p} \dfrac{\theta_k^{\star 2}\eta_k^2}{\lambda + \hat{V}\,\eta_k} \end{cases}. \tag{A4}$$

Note that the risk considered in equation (A1) slightly differs from equation (4) by: (a) a $1/n$ factor multiplying the sum, (b) additional $\sqrt{p}$ scalings and (c) the fact that it is written for finite $p$. Accounting for these differences, we can rewrite theorem 1 of [27] in our setting as:

$$\epsilon_{\mathrm{g}} - \sigma^2 = \lim_{p\to\infty} (\rho - 2m^\star + q^\star), \tag{A5}$$

with $\rho = \theta^{\star\top}\Sigma\theta^\star$, and $(m^\star, q^\star)$ fixed-points of

$$\begin{cases} \hat{V} = \dfrac{\frac{n}{p}}{1+V} \\ \hat{q} = \dfrac{n}{p}\dfrac{\rho + q - 2m + \sigma^2}{(1+V)^2} \end{cases}, \begin{cases} V = \dfrac{1}{p}\displaystyle\sum_{k=1}^{p} \dfrac{p\eta_k}{n\lambda + p\,\hat{V}\,\eta_k} \\ q = p\displaystyle\sum_{k=1}^{p} \dfrac{\hat{q}\eta_k^2 + \theta_k^{\star 2}\eta_k^2\,\hat{m}^{\,2}}{(n\lambda + p\,\hat{V}\,\eta_k)^2} \\ m = p\,\hat{V}\displaystyle\sum_{k=1}^{p} \dfrac{\theta_k^{\star 2}\eta_k^2}{n\lambda + p\,\hat{V}\,\eta_k} \end{cases}. \tag{A6}$$

Note, however, that rescaling from (A4) to (A6), sending $p \to \infty$ while keeping $n$ finitely large, and further allowing $\lambda$ to scale with $n$ all break the initial assumptions of theorem 1 [27], thereby losing the control in equation (A3). Therefore, strictly speaking the results derived hereafter are not rigorous, and we assume that the typical excess error can still be computed from equation (A5). In fact, this is well-justified by comparing the results obtained from extrapolating the theory with finite instance simulation, e.g. figures 2–4.

### A.2. Self-consistent equations for the excess prediction error

Defining $z = \frac{n^2}{p}\frac{\lambda}{\hat{V}}$, the equation (A6) allow to write

$$z = n\lambda + \frac{z}{n}\sum_{k=1}^{p} \frac{\eta_k}{\frac{z}{n} + \eta_k}. \tag{A7}$$

An expression for the excess error $\epsilon_{\mathrm{g}} - \sigma^2$ can be obtained combining (A5) with (A6):

$$\epsilon_{\mathrm{g}} - \sigma^2 \underset{(\mathrm{a})}{=} \lim_{p\to\infty} \frac{1}{p}\sum_{k=1}^{p}\left[\theta_k^{\star 2} p\eta_k + \frac{\hat{q}p^2\eta_k^2 + \theta_k^{\star 2}p^2\eta_k^2\,\hat{m}^{\,2}}{(n\lambda + \hat{V}\,p\eta_k)^2} - \frac{2\,\hat{m}\,\theta_k^{\star 2}p^2\eta_k^2}{n\lambda + \hat{V}\,p\eta_k}\right]$$

$$\underset{(\mathrm{b})}{=} \lim_{p\to\infty}\sum_{k=1}^{p} \frac{\theta_k^{\star 2}\eta_k\Big(n\lambda + \hat{V}\,p\eta_k\Big)^2 + \frac{p^2}{n}\eta_k^2\,\hat{V}^{\,2}\epsilon_g + \hat{V}^{\,2}\theta_k^{\star 2}p\eta_k^2 - 2\theta_k^{\star 2}\,\hat{V}\,p\eta_k^2\Big(n\lambda + \hat{V}\,p\eta_k\Big)}{(n\lambda + \hat{V}\,p\eta_k)^2} \tag{A8}$$

$$= \lim_{p\to\infty}\sum_{k=1}^{p}\frac{\frac{p^2}{n}\eta_k^2\,\hat{V}^{\,2}\epsilon_{\mathrm{g}} + n^2\lambda^2\theta_k^{\star 2}\eta_k}{(n\lambda + \hat{V}\,p\eta_k)^2}, \tag{A9}$$

thus

$$\epsilon_{\mathrm{g}} = \lim_{p\to\infty}\frac{\frac{z^2}{n^2}\sum_{k=1}^{p}\frac{\theta_k^{\star 2}\eta_k}{\left(z\frac{1}{n}+\eta_k\right)^2} + \sigma^2}{1 - \frac{1}{n}\sum_{k=1}^{p}\frac{\eta_k^2}{\left(z\frac{1}{n}+\eta_k\right)^2}}. \tag{A10}$$

Therefore, for the excess prediction error:

$$\epsilon_{\mathrm{g}} - \sigma^2 = \lim_{p\to\infty}\frac{\frac{z^2}{n^2}\sum_{k=1}^{p}\frac{\theta_k^{\star 2}\eta_k}{\left(z\frac{1}{n}+\eta_k\right)^2} + \frac{\sigma^2}{n}\sum_{k=1}^{p}\frac{\eta_k^2}{\left(z\frac{1}{n}+\eta_k\right)^2}}{1 - \frac{1}{n}\sum_{k=1}^{p}\frac{\eta_k^2}{\left(z\frac{1}{n}+\eta_k\right)^2}}. \tag{A11}$$

We now assume power-law form for the covariance spectrum and the teacher coordinates (8)

$$\eta_k = k^{-\alpha}, \qquad \theta_k^{\star 2}\eta_k = k^{-1-2r\alpha}, \tag{A12}$$

Then equation (A10) can be simplified to

$$\epsilon_{\mathrm{g}} - \sigma^2 = \lim_{p\to\infty}\frac{\frac{z^2}{n^2}\sum_{k=1}^{p}\frac{k^{-1-2r\alpha}}{\left(z\frac{1}{n}+k^{-\alpha}\right)^2} + \frac{\sigma^2}{n}\sum_{k=1}^{p}\frac{k^{-2\alpha}}{\left(z\frac{1}{n}+k^{-\alpha}\right)^2}}{1 - \frac{1}{n}\sum_{k=1}^{p}\frac{k^{-2\alpha}}{\left(z\frac{1}{n}+k^{-\alpha}\right)^2}}, \tag{A13}$$

which has a meaningful limit as $p \to \infty$ (with $n$, $\lambda$ kept fixed):

$$\epsilon_{\mathrm{g}} - \sigma^2 = \frac{\sum_{k=1}^{\infty}\frac{k^{-1-2r\alpha}}{(1+nz^{-1}k^{-\alpha})^2} + \frac{\sigma^2 n}{z^2}\sum_{k=1}^{\infty}\frac{k^{-2\alpha}}{1+nz^{-1}k^{-\alpha})^2}}{1 - \frac{n}{z^2}\sum_{k=1}^{\infty}\frac{k^{-2\alpha}}{(1+nz^{-1}k^{-\alpha})^2}}. \tag{A14}$$

Therefore, the excess prediction error suggestively decomposes into two terms, the first accounting for the variance due to sampling, while the second reflects the additional variance entailed by the label noise. Unlike a typical bias-variance decomposition, the effect of the bias (as manifested by the $\lambda$-dependent $z$ term) is subsumed in both terms. For simplicity, the first term in the numerator shall be referred to in the rest of the derivation as the *sample variance term*, and the second sum in the numerator as the *noise variance term*.

In the same limit, the equation defining $z$ (A7) is amenable to being rewritten:

$$z = n\lambda + \frac{z}{n}\sum_{k=1}^{\infty} \frac{1}{1+\frac{z}{n}k^{\alpha}}, \tag{A15}$$

or, approximating the Riemann sum by an integral

$$z \approx n\lambda + \left(\frac{z}{n}\right)^{1-\frac{1}{\alpha}} \int_{\left(\frac{z}{n}\right)^{1/\alpha}}^{\infty} \frac{dx}{1+x^{\alpha}}. \tag{A16}$$

### A.3. Infinite sample limit and the scaling of the generalisation error

Consider now the limit $n \gg 1$ with $\lambda$ scaling with $n$

$$\lambda \sim n^{-\ell}. \tag{A17}$$

Note that the scalings of $z$ with respect to $n$ differ according to the regularisation $\lambda$, depending on which of the two terms on the right-hand side of equation (A15) dominates. If the first $n\lambda$ term dominates, then (A15) simplifies to $z \approx n\lambda$. For this to be self-consistent, we must have $(z/n)^{1-\frac{1}{\alpha}} \approx \lambda^{1-\frac{1}{\alpha}} \ll n\lambda$, i.e. $n \gg \lambda^{-\frac{1}{\alpha}}$. In the converse case where the second term in (A15) dominates, $z \sim n^{1-\alpha}$. For this to consistently hold, one needs $(z/n)^{1-\frac{1}{\alpha}} \approx n^{1-\alpha} \gg n\lambda$, i.e. $n \ll \lambda^{-\frac{1}{\alpha-1}}$. Depending on which term dominates in (A15), two regime may be distinguished:

- In the *effectively non-regularized* $\ell > \alpha$ regime, $n \ll \lambda^{-\frac{1}{\alpha}}$ so $z \sim n^{1-\alpha}$. In this regime the regularization totally disappears from the analysis and KRR behaves just as if $\lambda = 0$.
- In the *effectively regularized* $\ell < \alpha$ regime, $n \gg \lambda^{-\frac{1}{\alpha}}$ regime, $z \approx n\lambda$.

### A.4. Effectively non-regularized regime

*A.4.1. Sample variance term.* As before, depending on $1 + 2r\alpha, \alpha$, it is sometimes possible to rewrite the sample variance term in integral form. If $r < 1$,

$$\sum_{k=1}^{\infty} \frac{k^{-1-2r\alpha}}{(1+nz^{-1}k^{-\alpha})^2} \sim n^{-2r\alpha}\sum_{k=1}^{\infty} \frac{\left(\frac{k}{n}\right)^{-1-2r\alpha}}{\left(1+\left(\frac{k}{n}\right)^{-\alpha}\right)^2}\frac{1}{n} \sim n^{-2r\alpha}\int_0^{\infty} \frac{x^{-1+2(1-r)\alpha}}{(1+x^{\alpha})^2} = \mathcal{O}(n^{-2r\alpha}). \tag{A18}$$

If $r > 1$, it is no longer possible to write the Riemann sum as an integral, and

$$\sum_{k=1}^{\infty} \frac{k^{-1-2r\alpha}}{(1+nz^{-1}k^{-\alpha})^2} = \sum_{k=1}^{n} \frac{k^{-1-2r\alpha}}{(1+n^{\alpha}k^{-\alpha})^2} + n^{-2r\alpha}\sum_{k=n}^{\infty} \frac{\left(\frac{k}{n}\right)^{-1-2r\alpha}}{\left(1+\left(\frac{k}{n}\right)^{-\alpha}\right)^2}\frac{1}{n} = \mathcal{O}(n^{-2\alpha}). \tag{A19}$$

*A.4.2. Noise variance term.* It is possible to similarly decompose the sum in the noise variance term to find

$$\frac{n\sigma^2}{z^2}\sum_{k=1}^{\infty}\frac{k^{-2\alpha}}{(1+nz^{-1}k^{-\alpha})^2}=\mathcal{O}(\sigma^2). \tag{A20}$$

From this, it follows that:

- for $n \ll \sigma^{-\frac{2}{2\alpha\ \min(r,1)}}$ the sample variance term dominates the numerator, and

$$\epsilon_{\mathrm{g}}-\sigma^2=\mathcal{O}(n^{-2\alpha\ \min(r,1)}) \tag{A21}$$

- for $n \gg \sigma^{-\frac{2}{2\alpha\ \min(r,1)}}$ the noise variance term dominates the numerator, and determines the decay of the excess prediction error

$$\epsilon_{\mathrm{g}}-\sigma^2=\mathcal{O}(\sigma^2) \tag{A22}$$

These two subregimes are amenable to being written in the more compact form (9):

$$\epsilon_{\mathrm{g}}-\sigma^2=\mathcal{O}\big(\max\big(\sigma^2,n^{-2\alpha\ \min(r,1)}\big)\big) \tag{A23}$$

### A.5. Effectively regularized regime

*A.5.1. Sample variance term.* By the same token, in the second $\ell<\alpha$ regularized regime, provided $r<1$, one can write the sample variance term as a Riemann sum (since $\lambda\sim n^{-\ell}=o(1)$):

$$\sum_{k=1}^{\infty}\frac{k^{-1-2r\alpha}}{(1+nz^{-1}k^{-\alpha})^2}\sim\lambda^{2r}\sum_{k=1}^{\infty}\frac{\left(k\lambda^{\frac{1}{\alpha}}\right)^{-1+2(1-r)\alpha}}{\left(\left(k\lambda^{\frac{1}{\alpha}}\right)^{\alpha}+1\right)^2}\lambda^{\frac{1}{\alpha}}\sim\lambda^{2r}\int_0^{\infty}\frac{x^{-1+2(1-r)\alpha}}{(1+x^{\alpha})^2}$$
$$=\mathcal{O}(n^{-2\ell r}). \tag{A24}$$

In the $r>1$ case,

$$\sum_{k=1}^{\infty}\frac{k^{-1-2r\alpha}}{(1+nz^{-1}k^{-\alpha})^2}=\sum_{k=1}^{n^{\frac{\ell}{\alpha}}}\frac{k^{-1-2r\alpha}}{\left(1+\frac{1}{\lambda}k^{-\alpha}\right)^2}+\lambda^{\frac{-2r\alpha}{\alpha}}\sum_{k=n^{\frac{\ell}{\alpha}}}^{\infty}\frac{\left(k\lambda^{\frac{1}{\alpha}}\right)^{-1+2(1-r)\alpha}}{\left(\left(k\lambda^{\frac{1}{\alpha}}\right)^{\alpha}+1\right)^2}\lambda^{\frac{1}{\alpha}}. \tag{A25}$$

Upper and lower bounds can be straightforwardly found for the first sum and the following equivalence established

$$\sum_{k=1}^{n^{\frac{\ell}{\alpha}}}\frac{k^{-1-2r\alpha}}{\left(1+\frac{1}{\lambda}k^{-\alpha}\right)^2}\sim n^{-2\ell}\sum_{k=1}^{n^{\frac{\ell}{\alpha}}}k^{-1+2(1-r)\alpha}=\mathcal{O}(n^{-2\ell}), \tag{A26}$$

while the second sum is a Riemann sum of order $\mathcal{O}(n^{\frac{(-2r\alpha)\ell}{\alpha}}) = o(n^{-2\ell})$. Therefore, the first sum in the numerator scales like

$$\sum_{k=1}^{\infty} \frac{k^{-1-2r\alpha}}{(1+nz^{-1}k^{-\alpha})^2} = \mathcal{O}\big(n^{-2\,\ell \min(r,1)}\big) \tag{A27}$$

*A.5.2. Noise variance term:.* The scaling of the noise variance term is found along similar lines to be

$$\frac{n\sigma^2}{z^2}\sum_{k=1}^{\infty} \frac{k^{-2\alpha}}{(1+nz^{-1}k^{-\alpha})^2} = \mathcal{O}\Big(\sigma^2 n^{\frac{\ell-\alpha}{\alpha}}\Big). \tag{A28}$$

If the noise variance term decays faster in $n$, then the sample variance term always dominates (since $\sigma^2$ is at most $\mathcal{O}(1)$). This is the case when

$$0 < \ell < \frac{\alpha}{2\alpha\ \min(r,1)+1} \tag{A29}$$

and then the generalization excess prediction error scales like

$$\epsilon_{\mathrm{g}} - \sigma^2 = \mathcal{O}\big(n^{-2\ell\ \min(r,1)}\big). \tag{A30}$$

In the case where $\alpha > \ell > \frac{\alpha}{2\alpha\ \min(r,1)+1}$ there exist two regimes depending on how $n$ compares with the noise strength

- if $n \ll \sigma^{\frac{2}{1-\frac{\ell}{\alpha}(1+2\alpha\ \min(r,1))}}$ the sample variance term dominates and we recover the noiseless case

$$\epsilon_{\mathrm{g}} - \sigma^2 = \mathcal{O}\big(n^{-2\ell\ \min(r,1)}\big). \tag{A31}$$

- if $n \gg \sigma^{\frac{2}{1-\frac{\ell}{\alpha}(1+2\alpha\ \min(r,1))}}$ the noise variance term dominates and

$$\epsilon_{\mathrm{g}} - \sigma^2 = \mathcal{O}\Big(\sigma^2 n^{\frac{\ell-\alpha}{\alpha}}\Big). \tag{A32}$$

All those regimes can be written more compactly as (10)

$$\epsilon_{\mathrm{g}} - \sigma^2 = \mathcal{O}\Big(\max\Big(\sigma^2, n^{1-2\ell\ \min(r,1)-\frac{\ell}{\alpha}}\Big) n^{\frac{\ell-\alpha}{\alpha}}\Big). \tag{A33}$$

*A.5.3. Case $\ell < 0$.* We give here for completeness the case in which the regularization grows with $n$. Then the sample variance term scales like

$$\sum_{k=1}^{\infty} \frac{k^{-1-2r\alpha}}{(1+nz^{-1}k^{-\alpha})^2} = \mathcal{O}(1). \tag{A34}$$

To see this, use a lower and upper bound starting from $0 \leqslant nz^{-1}k^{-\alpha} \sim n^{\ell}k^{-\alpha} \leqslant 1$ for all $k \geqslant 1$ and all $n$. The noise variance term scales like

$$\frac{\sigma^2 n}{z^2}\sum_{k=1}^{\infty}\frac{k^{-2\alpha}}{(1+nz^{-1}k^{-\alpha})^2}\sim\sigma^2 n^{2\ell-1}=o(1), \tag{A35}$$

meaning

$$\epsilon_{\mathrm{g}}-\sigma^2=\mathcal{O}(1) \tag{A36}$$

### A.6. Continuity across the regularization crossover line

The $\ell=\alpha$ is actually comprised in the $\ell>0$ case of the $\ell<\alpha$ regimes. On the $\ell=\alpha$ separation line, there is no discontinuity between the non-regularized exponents and the regularized exponents, since

$$\max\left(\sigma^2, n^{1-2\ell\ \min(r,1)-\frac{\ell}{\alpha}}\right)n^{\frac{\ell-\alpha}{\alpha}}\overset{\ell=\alpha}{=}\max\left(\sigma^2, n^{-2\alpha\ \min(1,r)}\right). \tag{A37}$$

### A.7. Asymptotically optimal regularization

The derivation in appendices A.4 and A.5 effectively delimit the four regimes in figure 1: the effectively non-regularized noiseless green regime, the effectively regularized noiseless blue regime, the effectively non-regularized noisy red regime, and the effectively regularized noisy orange regime.

For any given $n$, we define the *asymptotically optimal* $\ell$ as the regularization decay yielding fastest decay of the excess prediction error. This corresponds to finding the $\ell$ with maximal excess error decay along a vertical line at abscissa $n$ in the phase diagram figure 1.

If $n\gg n_1^{\star}\approx\sigma^{-\frac{1}{\alpha\ \min(r,1)}}$ (effectively noisy regime), the noise-induced crossover line is crossed for

$$\ell_c\approx\left(1-2\frac{\ln\,\sigma}{\ln\,n}\right)\frac{\alpha}{1+2\alpha\ \min(r,1)}. \tag{A38}$$

The asymptotically optimal $\ell^{\star}$ is found as

$$\ell^{\star}=\underset{\ell}{\mathrm{argmax}}\left(2\ell\ \min(r,1)\mathbb{1}_{0<\ell<\ell_c}+\frac{\alpha-\ell}{\alpha}\mathbb{1}_{\ell_c<\ell<\alpha}+0\times\mathbb{1}_{\alpha<\ell}\right). \tag{A39}$$

Since the argument of the argmax is an increasing function of $\ell$ on $(0,\ell_c)$ and a decreasing function on $(\ell_c,\infty)$ the maximum is found for $\ell^{\star}=\ell_c$. The corresponding decay for the excess error is

$$\max\left(2\ell^{\star}\ \min(r,1),\frac{\alpha-\ell^{\star}}{\alpha}\right)=2\ell^{\star}\ \min(r,1)\approx\frac{2\alpha\ \min(r,1)}{1+2\alpha\ \min(r,1)}\left(1-2\frac{\ln\,\sigma}{\ln\,n}\right). \tag{A40}$$

It is nonetheless ill-defined to talk about an asymptotically optimal rate for the regularization that continuously varies with $n$ when $n$ is comparable with $\sigma^{-2}$, since (A40)

**Table 2.** Dictionary between different notations previously used in the KRR literature.

| Reference | $\alpha$ [8] | $r$ [8] |
|---|---|---|
| [13] | $b$ | $\frac{a-1}{2b}$ |
| [12] | $\frac{\alpha_S}{d}$ | $\frac{1}{2}(\frac{\alpha_T}{\alpha_S} - d)$ |
| [9] | $\beta$ | $\frac{2\delta+\beta-1}{2\beta}$ |
| [10, 14] | $b$ | $\frac{c}{2}$ |
| [11, 15] | $\frac{1}{p}$ | $\frac{\beta}{2}$ |
| [20] | $b$ | $\beta$ |

means that the excess error is not even a power law in this region. An asymptotic statement can be however made. For $n \gg n_2^\star \approx \max(n_1^\star, \sigma^{-2})$,

$$\ell^\star \approx \frac{\alpha}{1 + 2\alpha \ \min(r, 1)}, \tag{A41}$$

and the excess error decays like (12)

$$\epsilon_g^\star - \sigma^2 = \mathcal{O}\Big(n^{-\frac{2\alpha \ \min(r,1)}{1+2\alpha \ \min(r,1)}}\Big). \tag{A42}$$

For $n \ll \sigma^{-\frac{2}{2\alpha \ \min(r,1)}}$ (effectively noiseless regime), we have

$$\ell^\star = \underset{\ell}{\operatorname{argmax}} \, (2\ell \ \min(r, 1)\mathbb{1}_{0<\ell<\alpha} + 2\alpha \ \min(r, 1)\mathbb{1}_{\alpha<\ell}), \tag{A43}$$

which means that any $\ell^\star \in (\alpha, \infty)$ is optimal (in particular, vanishing regularization is optimal), and we recover (11)

$$\epsilon_g^\star - \sigma^2 = \mathcal{O}\big(n^{-2\alpha \ \min(r,1)}\big). \tag{A44}$$

## Appendix B. A dictionary of notation in the literature

While the capacity and source conditions are assumed in almost all works concerned with the decay rates of kernel methods, the actual notations for the capacity and source terms $\alpha, r$ greatly vary. We provide in this appendix a table summarizing notations for the references [8–15, 20] (table 2).

## Appendix C. Details on real data sets

### C.1. Feature map to diagonal covariance for real datasets

In the general case where the data $x$ is drawn from a generic distribution $\rho_x$, we remind the equations defining the feature map $\psi$ (3):

$$\psi(x) = \Sigma^{\frac{1}{2}}\phi(x) \tag{C1}$$

$$\mathbb{E}_{x\sim\rho_x}\left[\phi(x)\phi(x)^{\mathrm{T}}\right] = \mathbb{1}_p \tag{C2}$$

$$\mathbb{E}_{x'\sim\rho_x}[K(x,x')\phi(x')] = \Sigma\phi(x) \tag{C3}$$

In the of a real dataset $\mathcal{D} = \{x^\mu, y^\mu\}_{\mu=1}^{n_{\text{tot}}}$ from which both the train and test set are uniformly drawn, the distribution is then the empirical uniform distribution over $\mathcal{D}$,

$$\rho_x(\cdot) = \frac{1}{n_{\text{tot}}}\sum_{\mu=1}^{n_{\text{tot}}}\delta(\cdot - x^\mu). \tag{C4}$$

Defining the Gram matrix $(K_{\mu\nu})_{\mu,\nu=1}^{n_{\text{tot}}} \overset{\text{def}}{=} (K(x^\mu,x^\nu))_{\mu,\nu=1}^{n_{\text{tot}}} \in \mathbb{R}^{n_{\text{tot}}\times n_{\text{tot}}}$, the equations defining the feature map (3) can be rewritten in the simpler matricial form

$$\psi = \phi\Sigma^{\frac{1}{2}}, \qquad \frac{1}{n_{\text{tot}}}\phi^{\mathrm{T}}\phi = \mathbb{1}_{n_{\text{tot}}}, \qquad \frac{1}{n_{\text{tot}}}K\phi = \phi\Sigma \tag{C5}$$

where $\phi, \psi, \lambda, K \in \mathbb{R}^{n_{\text{tot}}\times n_{\text{tot}}}$, and the feature space is of dimension $p = n_{\text{tot}}$, with the $\mu^{\text{th}}$ line of $\psi$ (resp. $\phi$) corresponding to $\psi(x^\mu)$ (resp. $\phi(x^\mu)$). To access the coordinates $\theta^\star_k$ in the basis of the features $\psi$, remember $\psi\theta^\star = y$, hence

$$\theta^\star = \frac{1}{n_{\text{tot}}}\Sigma^{-1}\psi^{\mathrm{T}}y \tag{C6}$$

### C.2. Estimation of source and capacity

The capacity and source terms $\alpha, r$ can be empirically estimated for the dataset $\mathcal{D}$ from the eigenvalues $\{\lambda_k\}_{k=1}^{n_{\text{tot}}}$ of the Gram matrix $K$ and the components $\{\theta^\star_k\}_{k=1}^{n_{\text{tot}}}$ of the teacher vector. Supposing decays like (8), the cumulative functions read:

$$\sum_{k'=k}^{n_{\text{tot}}}\lambda_{k'} \sim k^{1-\alpha}, \qquad \sum_{k'=k}^{n_{\text{tot}}}\lambda_{k'}\theta^{\star 2}_{k'} \sim k^{-2r\alpha}. \tag{C7}$$

These functions are plotted in figure 7 and the terms $\alpha, r$ estimated therefrom. The use of the cumulative functions, rather than a direct estimation from the coordinates, allows the integration to smoothen out the curves and get a more consistent estimation. The values of $\alpha, r$ thereby measured are summarized in table 1. Note that the power-law form (8) and the assumption $p = \infty$ fail to hold for real data, and the series (C7) have power-law form only on a range of indices $k$, before a sharp drop due to the finite dimensionality $n_{\text{tot}}$ of the feature space, see figure 7. The range of indices $k$ where the power-law form (8) seems to hold was qualitatively assessed, and linear regression run thereon to estimate $\alpha, r$. Since there is no clear objective way to determine the range the fit should be conducted on, the estimates slightly vary depending on the precise choice of the regression range, without however overly hurting the qualitative agreement with simulations figures 6 and 5.

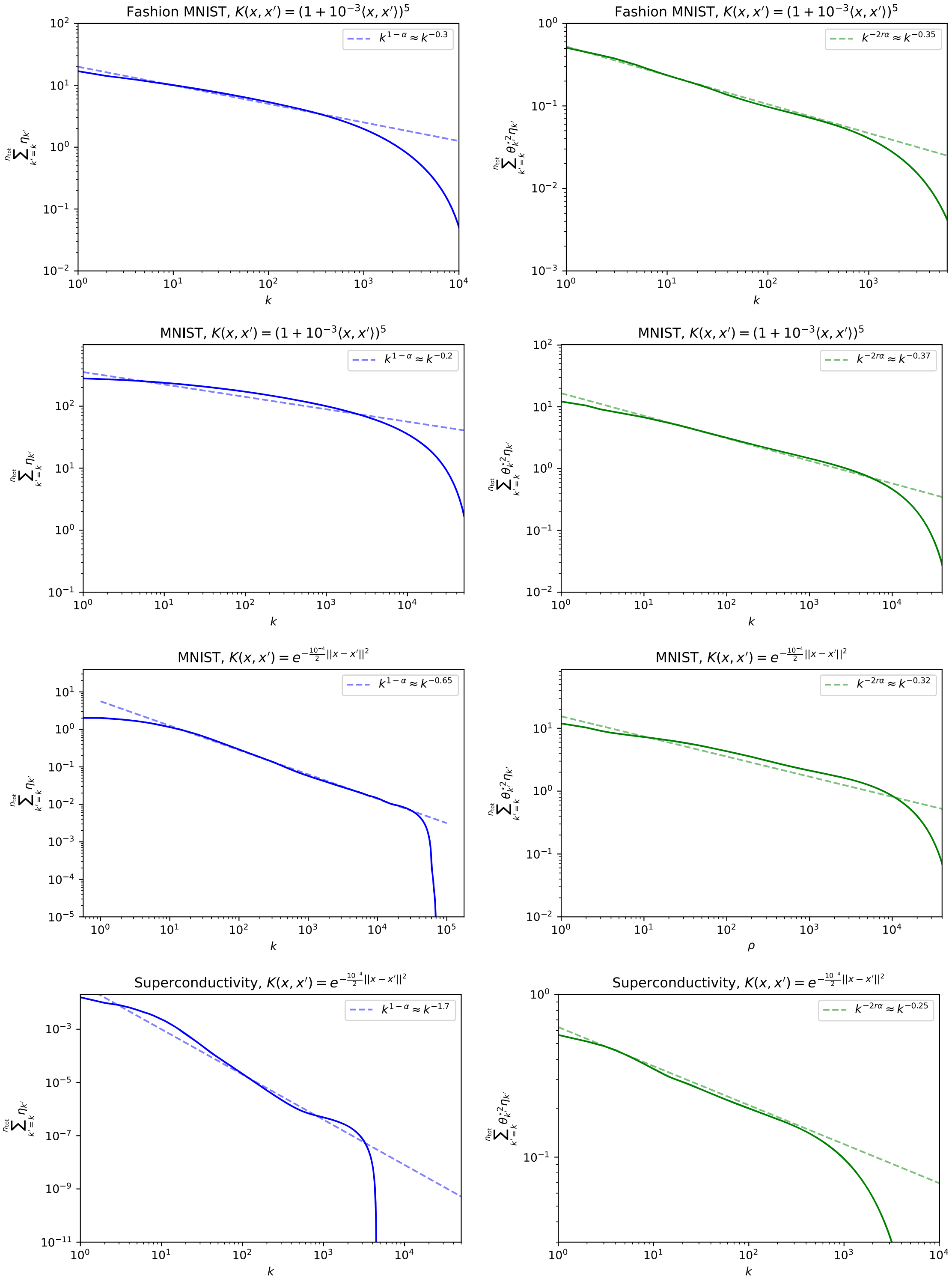

**Figure 7.** Measurement of empirical values for capacity and source $\alpha, r$ for real datasets (Fashion MNIST t-shirt and coat, MNIST) and RBF, polynomial kernels. Because the feature space is of finite dimension $n_{\rm tot}$ all the curves exhibit a sharp drop at $n_{\rm tot}$. A power-law was fitted on the functions (C7) on a range of $k$ where these looked reasonably like power-laws.

https://doi.org/10.1088/1742-5468/ac9829

### C.3. Details on simulations

We close this appendix by providing further detail on the simulations on real data (figures 6 and 5).

For each simulation at sample size $n$ a train set was created by subsampling $n$ samples from the total available dataset $\mathcal{D}$. To mitigate the effect of spurious correlations due to sampling a finite dataset, the whole dataset $\mathcal{D}$ has been used as a test set, following [27, 49]. A kernel ridge regressor was fitted on the train set with the help of the `scikit-learn KernelRidge` package [48]. For figure 6, the best regularization $\lambda$ was estimated using the `scikit-learn GridSearchCV` [48] default five-fold cross-validation procedure on the Grid $\lambda \in \{0\} \cup (10^{-10}, 10^{5})$, with logarithmic step size $\delta \log \lambda = 0.026$. The excess test error was averaged over 10 independent samplings of the train set and noise realizations.

## Appendix D. More crossovers

### D.1. Regularization-induced crossover

On top of the distinction between effectively noiseless regimes (green, blue regions in figure 1) and effectively noisy regimes (red, orange in figure 1), the four regimes can also be classified in effectively non-regularized (green, red) and effectively regularized (blue, orange), see also the discussion in section 4. In figure 1, the non-regularized regions lie above the horizontal separation line $\ell = \alpha$, while the regularized ones lie below. In this appendix, we discuss a more generic setting for which this separation line ceases to be horizontal, thereby creating *a new crossover line*. Similarly to the noise-type crossover line discussed in the main text, a learning curve that crosses this regularization-induced crossover line transitions from an non-regularized regime (green, red) to a regularized one (blue, orange), characterized by differing decays for the excess error $\epsilon_g - \sigma^2$. In figures 8 and 9, the noise-type crossover line is depicted in red, while the regularization-type crossover line is in blue.

In this section we thus detail the more general setting

$$\lambda = \lambda_0 n^{-\ell}, \tag{D1}$$

with, compared to the setup studies in the main text and appendix A, an additional prefactor $\lambda_0$ to the regularization $\lambda$ that is allowed to be $\ll 1$. The particular case $\ell = 0, \lambda_0$ small, has been studied in [13], and has been shown to give rise to a crossover due to the regularization, on top on the one evidenced in the present work due to the noise.

- For small $n$, KRR focuses on fitting the spiked subspace comprising large variance dimensions, and satisfies the norm constraint introduced by the regularization on the lower importance subspace. This phenomenon can be loosely regarded as the bias version of the benign overfitting for noise variance [17, 45]: the bias induced by the loss of expressivity due to the norm constraint is effectively diluted over less important dimensions, thereby not impacting the generalization.

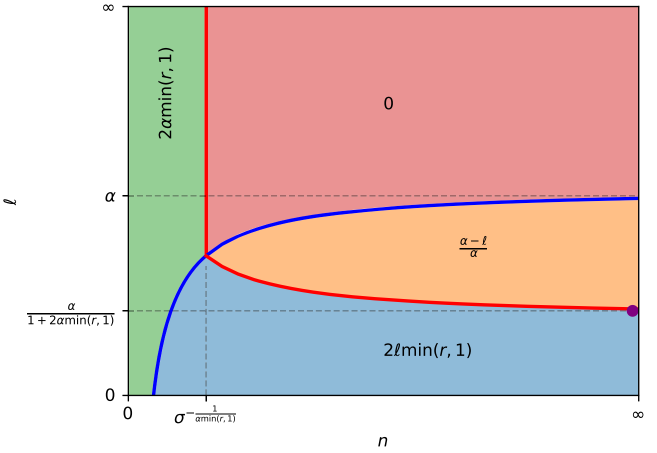
∞
2αmin(r, 1)
0
ℓ
α
$\frac{\alpha-\ell}{\alpha}$
$\frac{\alpha}{1+2\alpha\min(r,1)}$
2ℓmin(r, 1)
0
0
$\sigma^{-\frac{1}{\alpha\min(r,1)}}$
∞
n

**Figure 8.** Phase diagram for $\lambda_0 \ll 1$ and $\sigma < \lambda_0^{\min(r,1)+\frac{1}{\alpha}}$. As for figure 1. The solid red line corresponds to the noise-type crossover line, while the blue line indicates the regularization-type crossover line. Note that for low enough values of the regularization, the two crossover lines can be intercepted by a same horizontal line. This means that a double crossover is in theory observable (green-blue-orange), the first induce by regularization (see also [13]) and the second being noise-induced.

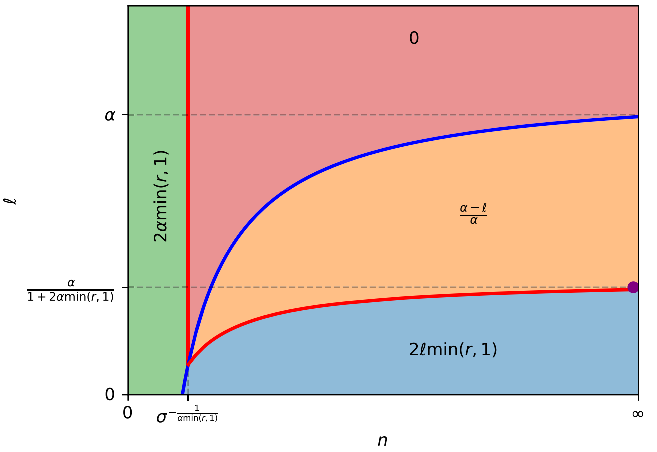
0
α
ℓ
2αmin(r, 1)
$\frac{\alpha-\ell}{\alpha}$
$\frac{\alpha}{1+2\alpha\min(r,1)}$
2ℓmin(r, 1)
0
0
$\sigma^{-\frac{1}{\alpha\min(r,1)}}$
∞
n

**Figure 9.** Phase diagram for $\lambda_0 \ll 1$ and $\sigma > \lambda_0^{\min(r,1)+\frac{1}{\alpha}}$. As for figure 1, the asymptotically optimal decays $\ell^{\star}$ are indicated in solid red. The solid red line corresponds to the noise-type crossover line, while the blue line indicates the regularization-type crossover line. Note that for low enough values of the regularization, the blue crossover line can be intercepted by a horizontal line, alongside the red crossover line twice. Consequently a triple crossover is in theory observable (green-red-orange-blue), with two noise-induced and one regularization-induced.

- For larger $n$, decreasing the excess error $\epsilon_{\mathrm{g}} - \sigma^2$ requires a good KRR fit also on the subspace of lesser importance, and the regularization effect is felt. In a noiseless green-blue crossover, this results in a slower decay because of the bias introduced by regularizing. On a noisy red-orange crossover, the regularization conversely helps to mitigate the noise and enables the excess risk to decay again.

### D.2. Outline of the computation

The derivation for the general case (D1) follows very closely appendix A.

- If $n \ll \lambda_0^{-\frac{1}{\alpha-\ell}}$ or $\ell > \alpha$, $n \ll \lambda^{-\frac{1}{\alpha}}$ so $z \sim n^{1-\alpha}$,
- If $n \gg \lambda_0^{-\frac{1}{\alpha-\ell}}$ and $\ell < \alpha$, $n \gg \lambda^{-\frac{1}{\alpha}}$ and $z \approx \lambda$.
- If $n = \lambda_0^{-\frac{1}{\alpha-\ell}}$ regime and $\ell < \alpha$, $n \sim \lambda^{-\frac{1}{\alpha}}$ so $z \sim \lambda \sim n^{1-\alpha} \sim n^{-\ell}$,

Note that the introduction of $\lambda_0 \ll 1$ means the limits between regularized and non-regularized regime are now involving both $n, \ell$ as opposed to just $\ell$ in appendix A (see also figure 1). In the first $n \ll \lambda_0^{-\frac{1}{\alpha-\ell}}$ regime, the regularization effect is not sensed and the computation is identical to the $\lambda_0 = 1$ case in appendix A. In the regularized $n \gg \lambda_0^{-\frac{1}{\alpha-\ell}}$, keeping track of the prefactors yields

$$\epsilon_{\mathrm{g}} - \sigma^2 = \mathcal{O}\left(\lambda_0^{2\,\min(r,1)} n^{-2\ell\,\min(r,1)}\right) + \mathcal{O}\left(\sigma^2 n^{\frac{\ell-\alpha}{\alpha}} \lambda_0^{\frac{-1}{\alpha}}\right), \tag{D2}$$

so the excess risk decays like

$$\epsilon_{\mathrm{g}} - \sigma^2 = \mathcal{O}\left(\lambda_0^{2\,\min(r,1)} n^{\frac{\ell-\alpha}{\alpha}} \max\left(\left(\frac{\sigma}{\lambda_0^{\min(r,1)+\frac{1}{\alpha}}}\right)^2, n^{-2\ell\,\min(r,1)+1-\frac{\ell}{\alpha}}\right)\right). \tag{D3}$$

Depending on whether the maximum in (D3) is always realized by one of its two arguments, or by one then the other as $n$ is varied, there may be a noise-induced crossover.

- If $\sigma < \lambda_0^{\min(r,1)+\frac{1}{\alpha}}$ and $\ell \leqslant \frac{\alpha}{1+2\alpha\,\min(r,1)}$, the second argument of the maximum in (D3) dominates for all $n \geqslant 1$ so no crossover is to be observed (see figure 8), and

$$\epsilon_{\mathrm{g}} - \sigma^2 = \mathcal{O}\left(\lambda_0^{2\,\min(r,1)} n^{-2\ell\,\min(r,1)}\right). \tag{D4}$$

- If $\sigma > \lambda_0^{\min(r,1)+\frac{1}{\alpha}}$ and $\ell \geqslant \frac{\alpha}{1+2\alpha\,\min(r,1)}$, the first argument of the maximum in (D3) dominates for all $n \leqslant 1$ so no crossover is to be observed (see figure 9), and

$$\epsilon_{\mathrm{g}} - \sigma^2 = \mathcal{O}\left(\sigma^2 n^{\frac{\ell-\alpha}{\alpha}} \lambda_0^{\frac{-1}{\alpha}}\right). \tag{D5}$$

- In any other case, a crossover between the decays (D4) and (D5) is observed, at a sample size

$$n_1^\star = \left(\frac{\sigma}{\lambda_0^{\min(r,1)+\frac{1}{\alpha}}}\right)^{\frac{2}{1-\frac{\ell}{\alpha}(1+2\alpha\,\min(r,1))}}. \tag{D6}$$

The crossover is from (D4) to (D5) if $\ell \geqslant \frac{\alpha}{1+2\alpha\ \min(r,1)}$ an in the other order if $\ell \leqslant \frac{\alpha}{1+2\alpha\ \min(r,1)}$.

The determination of the asymptotically optimal decays carries through as appendix A, with the same conclusions. The four regimes and their respective limit, as well as the optimal $\ell$ at very large $n$ (purple point), are summarized in figures 8 and 9.

### D.3. Double and triple crossovers

We therefore recover the regularization induced crossover reported in [13] for the special case $\ell = 0, \sigma = 0$. It corresponds to the green-to-blue transition for the lowest $\ell$ in figures 8 and 9. We stress that such a mechanism is entirely due to the regularization, and hence happens *on top* of the noise-induced crossover studied in the present work. It is therefore possible in theory to observe both crossovers in succession.

We detail as an example a double green-to-blue-to-orange crossover (see blue curves in figure 10). For small $n$ (non-regularized noiseless green regime), KRR fits the heavy dimensions. Both noise overfitting and bias are benign. As $n$ is increased, the blue regularization type crossover line in figure 8 is crossed and the regularized noiseless blue region entered. More of the less important dimensions need to be fitted well: bias is felt and entails a slower decay, but the noise overfitting is diluted over even less important dimensions and remains benign. As the red noise-type crossover line is passed into the regularized noisy orange region, the overfitting ceases to be benign and hurts the decay rate.

## Appendix E. Relation to worst-case bounds

In this section, we sketch informally how the blue and orange exponents (10) can also be derived from the worst case bounds [16, 20]. Note that the recovery from worst case bounds of the exponents (10), which were here derived in the Gaussian design setting, suggests that for these regimes the worst case exponents are also equal to the typical case exponents. We remind the reader that this has also already been shown to be the case for the asymptotically optimal lambda, see section 3, exponents (12) and [10, 14].

### E.1. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces [16]

To relate the notations employed in [16] to ours, the correspondences

$$\gamma \in ]0,1] = \frac{1}{\alpha} \qquad \zeta \in [0,1] = r \qquad \theta \in [0,1] = 1-\ell \tag{E1}$$

$$\mathcal{L} = \Sigma \qquad f_H = f^\star \tag{E2}$$

should be used, see also appendix B. With respect to their equation (18) defining the source condition, the setting considered in the present work corresponds to the special case $\phi(u) = u^\zeta$. Note that the assumption $\ell \leqslant 1$ is slightly more restrictive than those

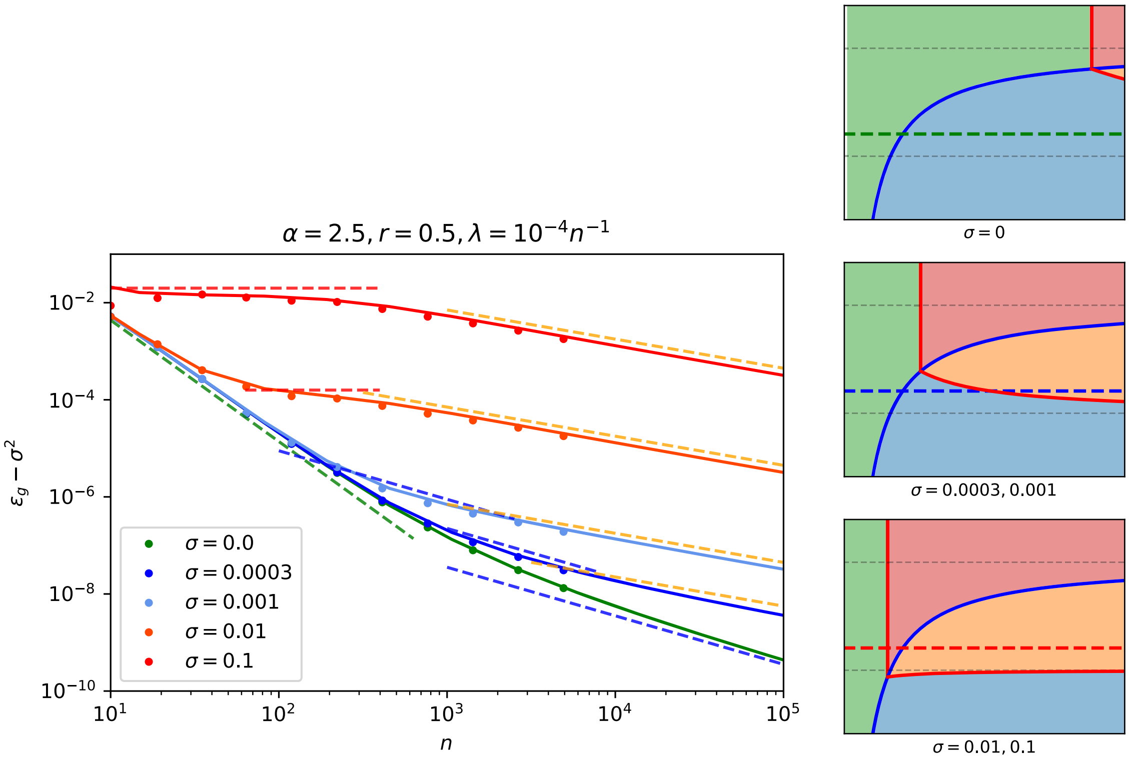

**Figure 10.** Excess risk learning curves for $\alpha = 2.5, r = 0.5, \lambda_0 = 10^{-4}, \ell = 1$. The noise level $\sigma$ is varied and the corresponding phase diagrams given on the right. For $\sigma = 0$ (green curve and top diagram on the right), a simple regularization-induced crossover (green-to-blue) is observed. For $\sigma = 3 \times 10^{-4}$ and $\sigma = 10^{-3}$ (blue curves on the left, middle diagram on the right) a double crossover green-to-blue-to-orange is observed. The first is regularization induced, while the second is due to noise. For $\sigma = 10^{-2}$ and $\sigma = 10^{-1}$ (orange, red curves and bottom diagram), a double green-to-red-to-orange crossover is observed, respectively noise and regularization induced.

employed in this paper. The main result of [16] is their theorem 4.2, which in our notations translates to

**Theorem 1. (Theorem 4.2 1)** [16]. *With probability* $1 - \delta$, *there exist constants* $\tilde{C}_1$, $\tilde{C}_2$, $\tilde{C}_3$ *s.t.*

$$(\epsilon_{\mathrm{g}} - \sigma^2)^{\frac{1}{2}} = \|f^\star - \hat{f}\|_{\mathcal{H}} = \|\theta^\star - \hat{w}\|_{\mathcal{L}^2} \leqslant \left(\tilde{C}_1 n^{-\max\left(\frac{1}{2}, 1-r\right)} + \tilde{C}_2 n^{-\frac{1}{2}} \lambda^{\frac{-1}{2\alpha}} + \tilde{C}_3 \lambda^r\right) \times \ln\frac{6}{\delta}\left(\ln\frac{6}{\delta} + \frac{\max\left(\frac{1}{1-l}, \ln\, n\right)}{\alpha}\right) \qquad \text{(E3)}$$

*viz expliciting the scalings*

$$(\epsilon_{\mathrm{g}}-\sigma^2)^{\frac{1}{2}} \leqslant \left( \tilde{C}_1 n^{-1+\min\left(\frac{1}{2},\min(r,1)\right)} + \tilde{C}_2 n^{-\frac{1}{2}\frac{\alpha-\ell}{\alpha}} + \tilde{C}_3 n^{-\ell\min(\mathrm{r},1)}\right) \\ \times \ln\frac{6}{\delta}\left(\ln\frac{6}{\delta}+\frac{\max\left(\frac{1}{1-l},\ln\, n\right)}{\alpha}\right) \tag{E4}$$

*we replaced $\zeta$ by $\min(r,1)$ since [16] work under the assumption $\zeta = r \in [0,1]$ in order to make contact with the exponents in the present paper. Up to logarithmic corrections, one recognizes the blue ($\tilde{C}_3$ term in (E4)) and orange ($\tilde{C}_2$ term in (E4)) exponents, the effectively unregularized red and green regimes (9) being inaccessible in this setting because of the restriction $\ell \leqslant 1$. One can further show that the first $\tilde{C}_1$ term in (E4) is always subdominant, since:*

- *if $r \geqslant \frac{1}{2}$, $n^{-\frac{1}{2}+\frac{\ell}{2\alpha}} \gg n^{-\frac{1}{2}}$ and the $\tilde{C}_2$ term dominates the $\tilde{C}_1$ term*
- *if $r \leqslant \frac{1}{2}$, $n^{-1+\min(1,r)} \ll n^{-\frac{1}{2}} \ll n^{-\ell\,\min(r,1)}$ since $\ell r \leqslant r \leqslant \frac{1}{2}$ and the $\tilde{C}_3$ term dominates $\tilde{C}_1$ term*

*The relative competition between the $\tilde{C}_{2,3}$ contributions in (E4) determine the blue to orange crossover. This suggests in particular that typical and worst case coincide within these regimes.*

### E.2. Kernel truncated randomized ridge regression: optimal rates and low noise acceleration [20]

Similar to [16], the notations translate to

$$\beta \in \left[0,\frac{1}{2}\right] = r \quad b\in[0,1] = \frac{1}{\alpha}. \tag{E5}$$

Note that in [20], it is further assumed that the labels are bounded by a constant $Y$ while this only holds with high probability in our setting. The theorem 3 in [20] then reads:

**Theorem 2. (Informal).** *The error gap given by the KTR$^3$ algorithm [20] is approximately bounded by, for any $\epsilon_r, \epsilon_\alpha > 0$, for the power law ansatz (A12):*

$$\epsilon_{\mathrm{g}}-\sigma^2 \leqslant \lambda^{2r-2\epsilon_r}\frac{1}{2\alpha\epsilon_r} + \min\left[\frac{4Y^2}{\alpha\epsilon_\alpha\lambda^{\frac{1}{\alpha}+\epsilon_\alpha}n}\min\left(\ln\left(1+\frac{1}{\lambda}\right)^{1-\frac{1}{\alpha}-\epsilon_\alpha},\frac{\alpha}{1+\alpha\epsilon_\alpha}\right),\ \frac{\lambda^{2r-2\epsilon_r-1}}{2\alpha\epsilon_r n}+\frac{\sigma^2}{\lambda n}\right], \tag{E6}$$

*so in the particular setting $\lambda = n^{-\ell}$*

$$\epsilon_{\mathrm{g}}-\sigma^2 \leqslant n^{-2r\ell+2\ell\epsilon_r}\frac{1}{2\alpha\epsilon_r} + \min\left[\frac{4Y^2}{\alpha\epsilon_\alpha}n^{-\frac{\alpha-\ell}{\alpha}+\epsilon_\alpha\ell} \\ \times\min\left(\ln\left(1+n^\ell\right)^{1-\frac{1}{\alpha}-\epsilon_\alpha},\frac{\alpha}{1+\alpha\epsilon_\alpha}\right),\frac{n^{-2r\ell+2\ell\epsilon_r+\ell-1}}{2\alpha\epsilon_r}+\sigma^2 n^{-1+\ell}\right]. \tag{E7}$$

If $\sigma \neq 0$, the $\sigma^2 n^{-1+\ell}$ term dominates in the second argument of the min and the min is realized by its first argument, leading to

$$\epsilon_{\mathrm{g}} - \sigma^2 = \mathcal{O}(n^{-2\ell r}) + \mathcal{O}\left(n^{-\frac{\alpha-\ell}{\alpha}}\right) \tag{E8}$$

namely the blue/orange crossover (10). If $\sigma = 0$ the bound is necessarily looser than $\mathcal{O}(n^{-2\ell r})$ which is coherent since in the noiseless setting only the blue exponent can be observed.

## References

[1] Nadaraja È A 1964 On a regression estimate *Teor. Verojatnost. i Primenen.* **9** 157–9

[2] Watson G S 1964 Smooth regression analysis *Sankhyā: Indian J. Stat.* A **26** 359–72

[3] Neal R M 1996 Priors for infinite networks *Bayesian Learning for Neural Networks* (New York: Springer) pp 29–53

[4] Williams C K I 1996 Computing with infinite networks *Proc. 9th Int. Conf. Neural Information Processing Systems, NIPS'96* (Cambridge, MA: MIT Press) pp 295–301

[5] Jacot A, Gabriel F and Hongler C 2018 Neural tangent kernel: convergence and generalization in neural networks *Advances in Neural Information Processing Systems* pp 8571–80

[6] Lee J, Bahri Y, Novak R, Schoenholz S S, Pennington J and Sohl-Dickstein J 2018 Deep neural networks as Gaussian processes *ICLR 2018*

[7] Chizat L, Oyallon E and Bach F 2019 On lazy training in differentiable programming *Advances in Neural Information Processing Systems* vol 32 ed H Wallach, H Larochelle, A Beygelzimer, F d' Alché-Buc, E Fox and R Garnett (Curran Associates, Inc.)

[8] Pillaud-Vivien L, Rudi A and Bach F 2018 Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes *Advances in Neural Information Processing Systems* vol 32 pp 8114–24

[9] Berthier R, Bach F and Gaillard P 2020 Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model (arXiv:2006.08212)

[10] Caponnetto A and Vito E D 2005 Fast rates for regularized least-squares algorithm *CBCL Paper 248/AI Memo 2005-013* Massachussets Institute of Technology, Cambridge, MA

[11] Steinwart I, Hush D R and Scovel C 2009 Optimal rates for regularized least squares regression *COLT* pp 79–93

[12] Spigler S, Geiger M and Wyart M 2020 Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm *J. Stat. Mech.* 124001

[13] Bordelon B, Canatar A and Pehlevan C 2020 Spectrum dependent learning curves in kernel regression and wide neural networks *Int. Conf. Machine Learning* (PMLR) pp 1024–34

[14] Caponnetto A and De Vito E 2007 Optimal rates for the regularized least-squares algorithm *Found. Comput. Math.* **7** 331–68

[15] Fischer S and Steinwart I 2020 Sobolev norm learning rates for regularized least-squares algorithms *J. Mach. Learn. Res.* **21** 1–38

[16] Lin J, Rudi A, Rosasco L and Cevher V 2018 Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces *Appl. Comput. Harmon. Anal.* **48** 868–90

[17] Bartlett P, Long P M, Lugosi G and Tsigler T 2020 Benign overfitting in linear regression *Proc. Natl Acad. Sci. USA* **117** 30063–70

[18] Polyak B T and Juditsky A B 1992 Acceleration of stochastic approximation by averaging *SIAM J. Control Optim.* **30** 838–55

[19] Nemirovskij A S and Yudin D B 1983 *Problem Complexity and Method Efficiency in Optimization* (New York: Wiley)

[20] Jun K-S, Cutkosky A and Orabona F 2019 Kernel truncated randomized ridge regression: optimal rates and low noise acceleration *NeurIPS*

[21] Varre A, Pillaud-Vivien L and Flammarion N 2021 Last iterate convergence of SGD for least-squares in the interpolation regime (arXiv:2102.03183)

[22] Kanagawa M, Hennig P, Sejdinovic D and Sriperumbudur B K 2018 Gaussian processes and kernel methods: a review on connections and equivalences (arXiv:1805.08845)

[23] Dicker L H 2016 Ridge regression and asymptotic minimax estimation over spheres of growing dimension *Bernoulli* **22** 1–37

[24] Hsu D, Kakade S M and Zhang T Random design analysis of ridge regression 2012 *Proc. 25th Annual Conf. Learning Theory* (Proceedings of Machine Learning Research vol 23) (Edinburgh, Scotland 25–27 June 2012) ed S Mannor, N Srebro and R C Williamson pp 9.1–9.24

[25] Dobriban E and Wager S 2018 High-dimensional asymptotics of prediction: ridge regression and classification *Ann. Stat.* **46** 247–79

[26] Ledoit O and Péché S 2011 Eigenvectors of some large sample covariance matrix ensembles *Probab. Theor. Relat. Field* **151** 233–64

[27] Loureiro B, Gerbelot C, Cui H, Goldt S, Krzakala F, Mézard M and Zdeborová L 2021 Capturing the learning curves of generic features maps for realistic data sets with a teacher–student model (arXiv:2102.08127)

[28] Dietrich R, Opper M and Sompolinsky H 1999 Statistical mechanics of support vector networks *Phys. Rev. Lett.* **82** 2975

[29] Opper M and Kinzel W 1996 Statistical mechanics of generalization *Models of Neural Networks III* (Berlin: Springer) pp 151–209

[30] Opper M and Urbanczik R 2001 Universal learning curves of support vector machines *Phys. Rev. Lett.* **86** 4410–3

[31] Kabashima Y 2008 Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels *J. Phys.: Conf. Ser.* **95** 012001

[32] Advani M S, Saxe A M and Sompolinsky H 2020 High-dimensional dynamics of generalization error in neural networks *Neural Netw.* **132** 428–46

[33] Belkin M, Hsu D and Xu J 2020 Two models of double descent for weak features *SIAM J. Math. Data Sci.* **2** 1167–80

[34] Hastie T, Montanari A, Rosset S and Tibshirani R J 2019 Surprises in high-dimensional ridgeless least squares interpolation (arXiv:1903.08560)

[35] Song M and Montanari A 2019 The generalization error of random features regression: precise asymptotics and double descent curve (arXiv:1908.05355)

[36] Gerace F, Loureiro B, Krzakala F, Mézard M and Zdeborová L 2020 Generalisation error in learning with random features and the hidden manifold model *37th Int. Conf. Machine Learning*

[37] Ghorbani B, Mei S, Misiakiewicz T and Montanari A 2019 Limitations of lazy training of two-layers neural network *Advances in Neural Information Processing Systems* vol 32 ed H Wallach, H Larochelle, A Beygelzimer, F dÁlché-Buc, E Fox and R Garnett pp 9111–21

[38] Kobak D, Lomond J and Sanchez B 2020 The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization *J. Mach. Learn. Res.* **21** 1–16

[39] Wu D and Xu J 2020 On the optimal weighted $\ell_2$ regularization in overparameterized linear regression *Advances in Neural Information Processing Systems* vol 33

[40] Richards D, Mourtada J and Rosasco L 2021 Asymptotics of ridge (less) regression under general source condition *Int. Conf. Artificial Intelligence and Statistics* (PMLR) pp 3889–97

[41] Liao Z, Couillet R and Mahoney M W 2020 A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent *Advances in Neural Information Processing Systems* vol 33

[42] Arthur J, Şimşek B, Spadaro F, Hongler C and Gabriel F 2020 Kernel alignment risk estimator: risk prediction from training data (arXiv:2006.09796)

[43] Ghorbani B, Mei S, Misiakiewicz T and Montanari A 2020 When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems* vol 33

[44] Liu F, Liao Z and Suykens J A K 2020 Kernel regression in high dimension: refined analysis beyond double descent (arXiv:2010.02681)

[45] Tsigler A and Bartlett P 2020 Benign overfitting in ridge regression (arXiv:2009.14286)

[46] Talagrand M 1994 Concentration of measure and isoperimetric inequalities in product spaces *Publ. Math. Inst. Hautes Études Sci.* **81** 73–205

[47] Louart C, Liao Z and Couillet R 2017 A random matrix approach to neural networks (arXiv:1702.05419)

[48] Pedregosa F *et al* 2011 Scikit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825–30

[49] Canatar A, Bordelon B and Pehlevan C 2021 Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks *Nat. Commun.* **12** 2914

[50] Xiao H, Rasul K and Vollgraf R 2017 Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms (arXiv:1708.07747)

[51] Hamidieh K 2018 A data-driven statistical model for predicting the critical temperature of a superconductor *Comput. Mater. Sci.* **154** 346–54

[52] Karoui N E 2013 Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results (arXiv:1311.2445)

[53] Thrampoulidis C, Abbasi E and Hassibi B 2018 Precise error analysis of regularized $m$-estimators in high dimensions *IEEE Trans. Inf. Theory* **64** 5592–628

https://doi.org/10.1088/1742-5468/ac9829