Proceedings of Machine Learning Research vol 313:1–30, 2026 37th International Conference on Algorithmic Learning Theory

# Relative Information Gain and Gaussian Process Regression

**Hamish Flynn** HAMISHFLYNN.GM@GMAIL.COM
*Universitat Pompeu Fabra, Barcelona, Spain*

**Editors:** Matus Telgarsky and Jonathan Ullman

## Abstract

The sample complexity of estimating or maximising an unknown function in a reproducing kernel Hilbert space is known to be linked to both the effective dimension and the information gain associated with the kernel. While the information gain has an attractive information-theoretic interpretation, the effective dimension typically results in better rates. We introduce a new quantity called the relative information gain, which measures the sensitivity of the information gain with respect to the observation noise. We show that the relative information gain smoothly interpolates between the effective dimension and the information gain, and that the relative information gain has the same growth rate as the effective dimension. In the second half of the paper, we prove a new PAC-Bayesian excess risk bound for Gaussian process regression. The relative information gain arises naturally from the complexity term in this PAC-Bayesian bound. We prove bounds on the relative information gain that depend on the spectral properties of the kernel. When these upper bounds are combined with our excess risk bound, we obtain minimax-optimal rates of convergence.

**Keywords:** Gaussian processes, kernel methods, PAC-Bayesian bounds

## 1. Introduction

We consider the model

$$y_i = f^\star(x_i) + \varepsilon_i \,, \tag{1}$$

with inputs $x_1, \ldots, x_n \in \mathcal{X}$ and real-valued responses $y_1, \ldots, y_n$. The target function $f^\star : \mathcal{X} \to \mathbb{R}$ is an unknown function in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The noise variables are assumed to be independent, centred and $\sigma$-sub-Gaussian.

The model in (1) has been extensively studied in the regression setting, in which a sample $(x_i, y_i)_{i=1}^n$ is used to estimate either $f^\star$ or the vector of function values $\mathbf{f}_n^\star := [f^\star(x_1), \ldots, f^\star(x_n)]^\top$ (Györfi et al., 2002; Tsybakov, 2009). Kernel ridge regression and Gaussian process regression are among the most popular approaches for the regression problem. For a regularisation parameter $\eta > 0$, the kernel ridge regression estimate is defined as

$$\widehat{f} := \operatorname*{argmin}_{f \in \mathcal{H}} \left\{ \sum_{i=1}^n (f(x_i) - y_i)^2 + \frac{1}{\eta} \|f\|_{\mathcal{H}}^2 \right\} .$$

It is well-known that there is a closed-form expression for $\widehat{f}$ (Schölkopf and Smola, 2002), which is

$$\widehat{f}(x) = \mathbf{k}_n^\top(x)(\mathbf{K}_n + \tfrac{1}{\eta}\mathbf{I})^{-1}\mathbf{y}_n \,.$$

Here, $\mathbf{k}_n(x) := [k(x, x_1), \ldots, k(x, x_n)]^\top$ is the vector of kernel comparisons between $x$ and the inputs $x_1, \ldots, x_n$, $\mathbf{K}_n := \{k(x_i, x_j)\}_{i,j=1}^n$ is the usual $n \times n$ kernel matrix and $\mathbf{y}_n := [y_1, \ldots, y_n]^\top$ is the response vector. Let us define

$$\widehat{\mathbf{f}}_n := [\widehat{f}(x_1), \ldots, \widehat{f}(x_n)]^\top = \mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{\eta}\mathbf{I})^{-1}\mathbf{y}_n$$

© 2026 H. Flynn.

as the vector of fitted values. Since each fitted value is a linear combination of the responses $y_1, \ldots, y_n$, kernel ridge regression is an example of a linear smoother. This means that we can write $\widehat{\mathbf{f}}_n = \mathbf{L}_n \mathbf{y}_n$. In this case, $\mathbf{L}_n = \mathbf{K}_n(\mathbf{K}_n + \frac{1}{\eta}\mathbf{I})^{-1}$. The matrix $\mathbf{L}_n$ is called the smoothing matrix, and its trace, denoted by $\mathrm{tr}(\mathbf{L}_n)$, is called the effective degrees of freedom (Wasserman, 2006). We will revisit this quantity shortly.

Gaussian process regression is usually presented as a Bayesian method. It is assumed that $f^\star \sim \mathcal{GP}(m(x), k(x, x'))$ is a random draw from a Gaussian process prior with mean function $m : \mathcal{X} \to \mathbb{R}$ and kernel (or covariance) function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. This means that for any fixed sequence of inputs $x_1, \ldots, x_n$, the random vector $\mathbf{f}_n^\star$ is Gaussian, and in particular, $\mathbf{f}_n^\star \sim \mathcal{N}(\mathbf{m}_n, \mathbf{K}_n)$, where $\mathbf{m}_n := [m(x_1), \ldots, m(x_n)]^\top$. If the noise variables are Gaussian with variance $1/\eta$, then the Bayesian posterior (say $Q$) is also a Gaussian process (Williams and Rasmussen, 2006). In particular, if $f^\star$ is drawn from a zero-mean prior, i.e. $f^\star \sim \mathcal{GP}(0, k(x, x'))$, then

$$Q = \mathcal{GP}(\mathbf{k}_n^\top(x)(\mathbf{K}_n + \tfrac{1}{\eta}\mathbf{I})^{-1}\mathbf{y}_n, k(x, x') - \mathbf{k}_n^\top(x)(\mathbf{K}_n + \tfrac{1}{\eta}\mathbf{I})^{-1}\mathbf{k}_n(x'))\,.$$

Note that the mean function of $Q$ is identical to the kernel ridge regression estimate. It turns out that the covariance function of $Q$, or rather, the covariance matrix of the marginal distribution of $Q$ at the points $x_1, \ldots, x_n$, is closely related to the effective degrees of freedom $\mathrm{tr}(\mathbf{L}_n)$ (cf. Lemma 15). Depending on the context in which it appears, the parameter $\eta$ is given various names. Throughout the rest of the paper, we will refer to $\eta$ as a learning rate.

The model in (1) has also been studied in the bandit setting (Srinivas et al., 2010; Valko et al., 2013). In kernelised bandits (also known as Gaussian process bandits), $f^\star$ is called the reward function. The objective is to sequentially maximise the reward function by querying it at a sequence of points $x_1, x_2, \ldots$ and receiving the random rewards $y_1, y_2, \ldots$, which can be used to inform the selection of future query points. Since the reward function is initially unknown, it must be estimated and maximised simultaneously. For this reason, many kernelised bandit algorithms use kernel ridge regression or Gaussian process regression as a subroutine for estimation.

In both kernel regression and kernelised bandits, there are two widely used notions of complexity, called the effective dimension and the information gain, which are designed to characterise the sample complexity of estimating or maximising the function $f^\star$. For any $n \geq 1$ and any learning rate $\eta > 0$, the effective dimension (Zhang, 2005) is

$$d_n(\eta) := \mathrm{tr}(\mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{\eta}\mathbf{I})^{-1})\,.$$

Note that $d_n(\eta)$ is identical to the effective degrees of freedom of the kernel ridge regression estimate. For any $n \geq 1$ and any learning rate $\eta \geq 0$, the information gain is defined as

$$\gamma_n(\eta) := \frac{1}{2}\log\det(\eta\mathbf{K}_n + \mathbf{I})\,.$$

As its name would suggest, the information gain has an information-theoretic interpretation. If in (1), $f^\star \sim \mathcal{GP}(0, k(x, x'))$ and $\varepsilon_i \sim \mathcal{N}(0, 1/\eta)$, then $\gamma_n(\eta)$ is equal to the mutual information $I(\mathbf{y}_n; f^\star)$ between $f^\star$ and the response vector $\mathbf{y}_n$. This version of the model in (1), in which both $f^\star$ and $\epsilon_1, \ldots, \epsilon_n$ are random, is a special case of a Gaussian channel (cf. Chapter 9 in Cover and Thomas, 2006). In the study of Gaussian channels, $\gamma_n(\eta)$ is an object of considerable interest.

It is known that effective dimension and the information gain are within logarithmic factors of each other (Calandriello et al., 2019; Zenati et al., 2022). In particular, the growth-rate (in $n$) of

the effective dimension is never larger than that of the information gain, whereas the growth-rate of the information gain can be larger than that of the effective dimension by a factor of $\log(n)$. Since both quantities can be used to characterise the sample complexity of estimating or maximising $f^\star$, it is natural to ask whether there are any other connections between them. One can also ask if there is a single quantity that has both the growth-rate of the effective dimension and an attractive information-theoretic interpretation like that of the information gain. We investigate these questions.

### 1.1. Contributions

We introduce a new quantity, called the relative information gain, which is the difference between the information gain at two different learning rates. As the smaller of the two learning rates is varied, a scaled version of the relative information gain smoothly interpolates between the effective dimension and the information gain, recovering each of them at the extremes. Moreover, the relative information gain matches the growth-rate of the effective dimension.

We demonstrate that the relative information gain is a reasonably natural notion of complexity for the model in (1), and not just an artificial quantity that is designed to interpolate between the effective dimension and the information gain. In particular, we derive a localised PAC-Bayesian excess risk bound for Gaussian process regression (cf. Theorem 5) and we find that the relative information gain arises naturally from the complexity term in this PAC-Bayesian bound.

Finally, we show that if $k$ is a Mercer kernel, then the relative information gain can be upper bounded based on the rate at which its eigenvalues decay. When these upper bounds are combined with our excess risk bound in Theorem 5, we obtain minimax-optimal rates of convergence.

### 1.2. Outline

The rest of this paper is organised as follows. Section 2 describes some related work on rates of convergence for kernel ridge regression and Gaussian process regression, and on PAC-Bayesian bounds for Gaussian processes. In Section 3, we define the relative information gain and establish some of its properties. In Section 4, we state and sketch the proof of a PAC-Bayesian excess risk bound. In Section 5, we provide upper bounds on the relative information gain and then use them to obtain rates of convergence with explicit dependence on the sample size.

## 2. Related Work

### 2.1. Rates of Convergence for Kernel Ridge Regression and Gaussian Process Regression

For the setting we consider, in which $f^\star$ is a fixed function in the RKHS $\mathcal{H}$, the optimal rates of convergence for kernel ridge regression are well-understood. The rates of convergence that we obtain are the same as those in (Caponnetto and De Vito, 2007; Steinwart et al., 2009; Dicker et al., 2015, 2017). These works impose similar conditions on the eigenvalues and eigenfunctions of the kernel. The rates of convergence for Gaussian process regression are also well-understood in this setting. The rate at which the (Bayesian or Gibbs) Gaussian process posterior contracts has been the subject of intense study (Seeger et al., 2008; Castillo, 2008; van der Vaart and van Zanten, 2008, 2011; Suzuki, 2012; Castillo, 2014; Pati et al., 2015; Nickl and Söhl, 2017). Several of these results have been used to derive rates of convergence for the excess risk (van der Vaart and van Zanten, 2011; Suzuki, 2012). In summary, it is already understood that both kernel ridge regression and Gaussian process

regression can achieve the minimax optimal rate of convergence for the setting that we consider. Our work provides a new way to prove optimal excess risk bounds using the relative information gain.

### 2.2. PAC-Bayesian Bounds for Gaussian Processes

PAC-Bayesian bounds originate from work by Shawe-Taylor and Williamson (1997) and McAllester (1998). We refer the reader to Alquier (2024) or Hellström et al. (2025) for a recent overview. Early work on PAC-Bayesian analysis of Gaussian processes focused on classification (Seeger, 2002, 2003). More recently, PAC-Bayesian bounds were used to fit Gaussian process predictors for either classification or regression (Reeb et al., 2018). The main focus of the aforementioned works was on obtaining numerically tight risk certificates, as opposed to obtaining the best rates of convergence. Suzuki (2012) derived PAC-Bayesian excess risk bounds for Gaussian process regression which match the optimal rate of convergence under certain conditions. Despite having a similar rate of convergence, these bounds are quite different to ours. In particular, the rate of convergence is determined by the concentration function used in van der Vaart and van Zanten (2008, 2011), as opposed to a notion of effective dimension or information gain. Alquier and Ridgway (2020) used PAC-Bayesian bounds to derive rates of convergence for Variational Bayes approximations of Gibbs distributions (such as Gaussian processes). When applied to nonparametric regression over Sobolev ellipsoids, the rate of convergence is optimal up to logarithmic factors. More recent work (Khribch and Alquier, 2024) has shown that optimal rates of convergence for a range of statistical models (including the Gaussian sequence model) can be obtained via localised PAC-Bayesian bounds. Khribch and Alquier (2024) highlight the connection between localised PAC-Bayesian bounds and mutual information bounds, although the relative information gain does not make an appearance.

## 3. Effective Dimension, Information Gain and Relative Information Gain

We describe a connection between the effective dimension and the information gain, and we introduce the relative information gain. First, we notice that both the effective dimension and the information gain can be expressed in terms of the eigenvalues $(\lambda_i)_{i=1}^n$ of $\mathbf{K}_n$. In particular, by Lemma 10 and Lemma 11,

$$d_n(\eta) = \sum_{i=1}^n \frac{\eta\lambda_i}{1+\eta\lambda_i}\,, \qquad \gamma_n(\eta) = \frac{1}{2}\sum_{i=1}^n \log(1+\eta\lambda_i)\,. \tag{2}$$

Starting from these expressions, one can easily verify that the effective dimension is related to the derivative of the information gain.

**Proposition 1** *For all $n \geq 1$ and $\eta \geq 0$,*

$$d_n(\eta) = 2\eta\gamma_n'(\eta)\,.$$

This identity is equivalent to a known result for Gaussian channels, which states that the derivative of the information gain with respect to the signal-to-noise ratio is equal to half the minimum mean squared error (Guo et al., 2005). In our formulation, $\eta$ is the signal-to-noise ratio, and it turns out that the minimum mean squared error is equal to $d_n(\eta)/\eta$ (cf. Appendix B). Thus Proposition 1 can be obtained from Theorem 1 in Guo et al. (2005).

The identity in Proposition 1 gives us an information-theoretic interpretation of the effective dimension. Namely, the effective dimension is a measure of how sensitive the information gain is to

the learning rate (or the variance of the noise in the responses). In addition, Proposition 1 suggests that we can obtain a reasonable approximation of the effective dimension by taking the difference between the information gain at two different learning rates. For a sample size $n \geq 1$, and two learning rates $\eta > \beta \geq 0$, we define the relative information gain as

$$\gamma_n(\eta, \beta) := \gamma_n(\eta) - \gamma_n(\beta)\,.$$

The relative information gain can be interpreted as the additional information that would be gained about $f^\star$ if the variance of the noise in the responses was reduced from $1/\beta$ to $1/\eta$. The bottom toast inequality (the bottom half of the sandwich inequality) in Proposition 5 of Calandriello et al. (2019) states that $d_n(\eta)$ is bounded by $2\gamma_n(\eta)$. We will now show that a scaled version of the relative information gain smoothly interpolates between these two quantities. Using Proposition 1, and the definition of the derivative, we can express the effective dimension as (the limit of) a scaled version of the relative information gain. In particular, for any learning rates $\eta > \beta > 0$,

$$d_n(\eta) = \lim_{\eta_0 \to \eta^-} \frac{2\eta}{\eta - \eta_0}\gamma_n(\eta, \eta_0) \approx \frac{2\eta}{\eta - \beta}\gamma_n(\eta, \beta)\,. \tag{3}$$

If $\beta = 0$, then the right-hand side of (3) is equal to twice the information gain, although the approximation of the limit is rather crude in this case. This establishes that the scaled information gain recovers both the effective dimension and the information gain as special cases. The following proposition shows that these are the two extreme cases.

**Proposition 2** *For all $n \geq 1$ and $\eta > \beta \geq 0$,*

$$d_n(\eta) \leq \frac{2\eta}{\eta - \beta}\gamma_n(\eta, \beta) \leq 2\gamma_n(\eta)\,.$$

*The first inequality is sharp in the limit as $\beta$ tends to $\eta$ from below and the second inequality is sharp when $\beta = 0$.*

**Proof** The sharpness of each inequality follows from (3). For any $\lambda \geq 0$, the mapping $\beta \mapsto \frac{1}{\eta-\beta}\log\frac{1+\eta\lambda}{1+\beta\lambda}$ is decreasing on $[0, \eta)$ (cf. Lemma 12). From this and (2), it follows that $\beta \mapsto \frac{2\eta}{\eta-\beta}\gamma_n(\eta, \beta)$ is also decreasing on $[0, \eta)$. Thus the scaled information gain is bounded between $d_n(\eta)$ and $2\gamma_n(\eta)$ for all $\beta$ in the interval $[0, \eta)$. ■

The full sandwich inequality in Proposition 5 of Calandriello et al. (2019) shows that the effective dimension and the information gain are within logarithmic factors of each other. In particular,

$$d_n(\eta) \leq 2\gamma_n(\eta) \leq (1 + \log(1 + \eta\lambda_{\max}))d_n(\eta)\,,$$

where $\lambda_{\max} = \max_i \lambda_i$. Since $\lambda_{\max}$ is at most of order $n$, this implies that $\gamma_n(\eta) = \mathcal{O}(d_n(\eta)\log(n))$. We prove that the effective dimension and the relative information gain can form a thinner sandwich.

**Proposition 3** *For all $n \geq 1$ and $\eta > \beta > 0$,*

$$d_n(\eta) \leq \frac{2\eta}{\eta - \beta}\gamma_n(\eta, \beta) \leq \frac{\eta}{\beta}d_n(\eta)$$

**Proof** The first inequality follows from Proposition 2. Using the inequality $\log(1+x) \leq x$ for $x \geq 0$, we obtain

$$\frac{\eta}{\eta-\beta}\sum_{i=1}^{n}\log\frac{1+\eta\lambda_i}{1+\beta\lambda_i} \leq \sum_{i=1}^{n}\frac{\eta\lambda_i}{1+\beta\lambda_i} = \sum_{i=1}^{n}\frac{1+\eta\lambda_i}{1+\beta\lambda_i}\frac{\eta\lambda_i}{1+\eta\lambda_i} \leq \frac{\eta}{\beta}\sum_{i=1}\frac{\eta\lambda_i}{1+\eta\lambda_i}\,.$$

The second inequality now follows from (2). ■

Proposition 3 tells us that the effective dimension and the relative information gain have the same growth-rate whenever $\beta = c\eta$, for some constant $c \in (0,1)$.

### 3.1. Illustrative Examples

We demonstrate that the information gain and the relative information gain really can have different growth rates. First, we consider a simple example in which we can easily verify that $\gamma_n(\eta,\beta) = \mathcal{O}(1)$ and $\gamma_n(\eta) = \Omega(\log(n))$. We assume that the inputs $x_1,\ldots,x_n \in \mathcal{X}$ are elements of a finite set $\mathcal{X}$ with cardinality $|\mathcal{X}| = k < n$, and that the kernel function satisfies $k(x_i,x_i) = 1$ for all $i \in [n]$. First, we upper bound the effective dimension. Suppose that $\mathbf{K}_n$ has $m$ non-zero eigenvalues, and let us assume that it is the first $m$ eigenvalues that are non-zero. Then

$$d_n(\eta) = \sum_{i=1}^{m}\frac{\eta\lambda_i}{1+\eta\lambda_i} \leq \sum_{i=1}^{m}\frac{\eta\lambda_i}{\eta\lambda_i} = m\,.$$

Since $|\mathcal{X}| \leq k$, the rank of the kernel matrix $\mathbf{K}_n$ is at most $k$, which means that $d_n(\eta) \leq k$. By Proposition 3, this also means that $\gamma_n(\eta,\beta) \leq \frac{\eta-\beta}{2\beta}k$. If for some $c \in (0,1)$, $\beta = c\eta$, then $\gamma_n(\eta,\beta) = \mathcal{O}(1)$. Next, we lower bound the information gain. Since every term in the expansion of $\prod_{i=1}^{n}(1+\eta\lambda_i)$ is non-negative,

$$\gamma_n(\eta) = \frac{1}{2}\log\prod_{i=1}^{n}(1+\eta\lambda_i) \geq \frac{1}{2}\log\left(1+\sum_{i=1}^{n}\eta\lambda_i\right) = \frac{1}{2}\log(1+\eta n)\,.$$

Here, we used the fact that since $k(x_i,x_i) = 1$ for all $i \in [n]$, $\mathrm{tr}(\mathbf{K}_n) = n$. We conclude that whenever $\eta$ does not depend on $n$, $\gamma_n(\eta) = \Omega(\log(n))$.

Next, we consider a somewhat less artificial example, and estimate the growth-rates of the information gain and the relative information gain via simulation. This time, the inputs $x_1,\ldots,x_n$ are drawn independently and uniformly at random from the interval $[0,1]$, the kernel function is the squared exponential kernel $k(x,y) = \exp(-(x-y)^2/2)$ and we take $\eta = 1$ and $\beta = 1/2$. For several values of $n$ between 1 and 3000, and 20 random draws of $x_1,\ldots,x_n$, we record the values of all three quantities that appear in Proposition 2. Namely, we record $d_n(\eta)$, $\frac{2\eta}{\eta-\beta}\gamma_n(\eta,\beta)$ and $2\gamma_n(\eta)$. We plot the average values of these three quantities in Figure 1, and obtain the estimates

$$\mathbb{E}[\gamma_n(1)] \approx 0.267\cdot\log^{8/5}(n) + 0.347\,,\quad \mathbb{E}[\gamma_n(1,1/2)] \approx 0.113\cdot\log(n) + 0.144\,.$$

We see that the information gain grows at a rate that is faster by a factor of approximately $\log^{3/5}(n)$.

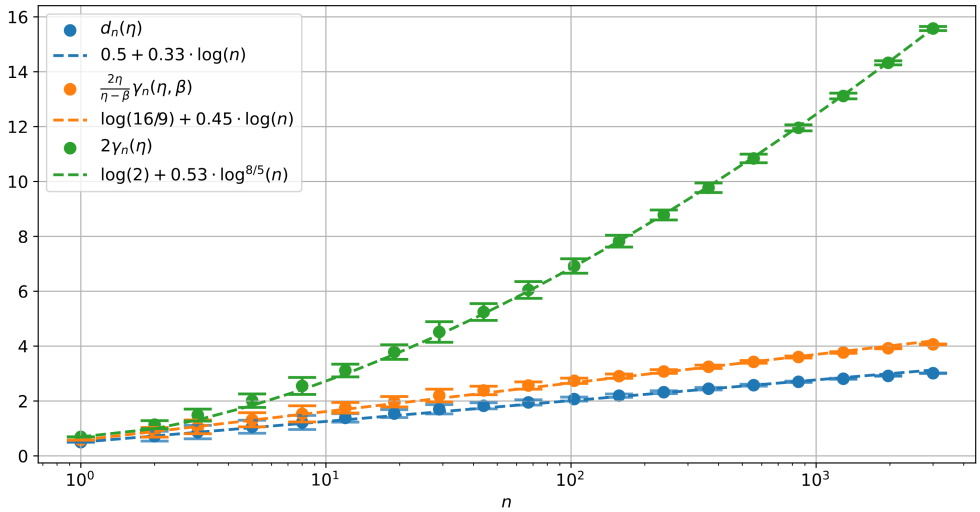


Figure 1: Estimates of the effective dimension, the information gain and the relative information gain. The inputs are drawn uniformly at random from $[0, 1]$, the kernel function is $k(x, y) = \exp(-(x - y)^2/2)$ and the learning rates are $\eta = 1$ and $\beta = 1/2$. We plot the average value plus/minus three times the standard deviation over 20 repetitions.

## 4. Excess Risk Bounds for Gaussian Process Regression

We demonstrate that the relative information gain arises naturally from a localised PAC-Bayesian bound for Gaussian process regression. We use the zero-mean prior $P_\alpha = \mathcal{GP}(0, \alpha k(x, x'))$ with a scale parameter $\alpha > 0$. For a learning rate $\eta > 0$, we will consider Gibbs distributions $Q_{n,\eta,\alpha}$ defined by

$$\frac{\mathrm{d}Q_{n,\eta,\alpha}}{\mathrm{d}P_\alpha}(f) \propto \exp\left( -\eta \sum_{i=1}^{n} (f(x_i) - y_i)^2 \right).$$

It is known that $Q_{n,\eta,\alpha}$ is also a Gaussian process (Williams and Rasmussen, 2006). In particular,

$$Q_{n,\eta,\alpha} = \mathcal{GP}(\mathbf{k}_n^\top(x)(\mathbf{K}_n + \tfrac{1}{2\eta\alpha}\mathbf{I})^{-1}\mathbf{y}_n, \alpha k(x, x') - \alpha \mathbf{k}_n^\top(x)(\mathbf{K}_n + \tfrac{1}{2\eta\alpha}\mathbf{I})^{-1}\mathbf{k}_n(x')).$$

Note that if we set the learning rate to $\eta = 1/(2\sigma^2)$, then $Q_{n,\eta,\alpha}$ coincides with the Bayesian posterior for the model in (1). We consider the problem of regression with fixed design, which means that the inputs $x_1, \ldots, x_n$ are deterministic and our objective is to estimate the function values $f^\star(x_1), \ldots, f^\star(x_n)$. In this setting, we only need to consider the marginal distributions of $Q_{n,\eta,\alpha}$ and $P_\alpha$ on the inputs $x_1, \ldots, x_n$, which are $n$-dimensional Gaussians. By a small abuse of notation, we will also use $Q_{n,\eta,\alpha}$ and $P_\alpha$ to refer to these marginal distributions. It is easy to check that $P_\alpha = \mathcal{N}(0, \alpha\mathbf{K}_n)$ and $Q_{n,\eta,\alpha} = \mathcal{N}(\mathbf{m}_{n,\eta,\alpha}, \mathbf{K}_{n,\eta,\alpha})$, where

$$\mathbf{m}_{n,\eta,\alpha} = \mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{2\eta\alpha}\mathbf{I})^{-1}\mathbf{y}_n\,, \quad \mathbf{K}_{n,\eta,\alpha} = \alpha\mathbf{K}_n - \alpha\mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{2\eta\alpha}\mathbf{I})^{-1}\mathbf{K}_n\,.$$

We assume that the noise variables $\varepsilon_1, \ldots, \varepsilon_n$ are independent and $\sigma$-sub-Gaussian.

**Assumption 1** *The noise variables $\varepsilon_1, \dots, \varepsilon_n$ are independent. For all $i \in \{1, \dots, n\}$ and all $\eta \in \mathbb{R}$, $\mathbb{E}[\exp(\eta\varepsilon_i)] \leq \exp(\sigma^2\eta^2/2)$.*

For simplicity, we also assume that the inputs $x_1, \dots, x_n$ and the kernel $k$ are such that the kernel matrix $\mathbf{K}_n$ is strictly positive-definite. This assumption is quite mild, since it is satisfied by typical kernels whenever $x_1, \dots, x_n$ are distinct. It is used to prove some of the auxiliary lemmas in Appendix C.1, but we expect that if one has the desire, this assumption can be dropped.

**Assumption 2** *The kernel matrix $\mathbf{K}_n$ is strictly positive-definite.*

In what follows, let us use $\mathbf{g} := [g_1, \dots, g_n]^\top$ to denote an arbitrary vector (of function values) in $\mathbb{R}^n$. We define the *excess risk* $R_n : \mathbb{R}^n \to \mathbb{R}$ and the *empirical risk* $r_n : \mathbb{R}^n \to \mathbb{R}$ as

$$R_n(\mathbf{g}) := \frac{1}{n}\|\mathbf{g} - \mathbf{f}_n^\star\|_2^2\,, \quad r_n(\mathbf{g}) := \frac{1}{n}\|\mathbf{g} - \mathbf{y}_n\|_2^2\,.$$

We want to upper bound the average excess risk $\int R_n(\mathbf{g})\mathrm{d}Q_{n,\eta,\alpha}(\mathbf{g})$ of $Q_{n,\eta,\alpha}$. Using a PAC-Bayesian bound for sums of independent random variables (cf. Proposition 2.1 in Catoni and Giulini, 2017), one can derive the following PAC-Bayesian bound. For any $n \geq 1$, any $\delta \in (0, 1]$, any $\eta \in (0, \frac{1}{2\sigma^2})$ and any $\alpha > 0$, with probability at least $1 - \delta$,

$$\int R_n(\mathbf{g})\mathrm{d}Q_{n,\eta,\alpha}(\mathbf{g}) \leq \inf_Q \left\{ \frac{\int \eta(r_n(\mathbf{g}) - r_n(\mathbf{f}_n^\star))\mathrm{d}Q(\mathbf{g})}{\eta - 2\sigma^2\eta^2} + \frac{D_{\mathrm{KL}}(Q||P_\alpha) + \log\frac{1}{\delta}}{n(\eta - 2\sigma^2\eta^2)} \right\}. \tag{4}$$

We consider this to be a fairly standard PAC-Bayesian bound for the excess risk. Indeed, one can obtain a bound similar to the one in (4) from Theorem 2.1 in Alquier and Ridgway (2020). However, this bound is not completely satisfactory. Lemma 13 tells us that the infimum on the right-hand side is achieved when $Q = Q_{n,\eta,\alpha}$. For this choice of $Q$, the KL divergence $D_{\mathrm{KL}}(Q_{n,\eta,\alpha}||P_\alpha)$ contains the log-determinant term $\frac{1}{2}\log(\det(\alpha\mathbf{K}_n)/\det(\mathbf{K}_{n,\eta,\alpha}))$. Due to Corollary 16,

$$\frac{1}{2}\log\frac{\det(\alpha\mathbf{K}_n)}{\det(\mathbf{K}_{n,\eta,\alpha})} = \frac{1}{2}\log\det(2\eta\alpha\mathbf{K}_n + \mathbf{I}) = \gamma_n(2\eta\alpha)\,.$$

Therefore, one would expect that the bound in (4) is at best of the order $\gamma_n(2\eta\alpha)/n$. In fact, one can show that this bound is of the order $[\gamma_n(2\eta\alpha) + \|f^\star\|_{\mathcal{H}}^2/\alpha]/n$. Sadly, plugging in the bounds on the information gain from Vakili et al. (2021) and then optimising the value of $\alpha$ does not yield minimax-optimal rates of convergence. The extra (up to) logarithmic factor in the growth rate of $\gamma_n(\eta)$ (compared to $d_n(\eta)$ or $\gamma_n(\eta, \beta)$) results in unnecessary log factors in the rate of convergence. To fix this, we use Catoni's localisation technique (Catoni, 2007), which allows one to prove PAC-Bayesian bounds with data-dependent priors. In Proposition 4, the prior is the Gibbs distribution $Q_{n,\beta,\alpha}$, with a learning rate $\beta < \eta$.

**Proposition 4** *Suppose that Assumption 1 is satisfied. For any $n \geq 1$, any $\delta \in (0, 1]$, any $\eta \in (0, \frac{1}{2\sigma^2})$, any $\beta > 0$ such that $\eta - 2\eta^2\sigma^2 - \beta - 2\beta^2\sigma^2 > 0$ and any $\alpha > 0$, w.p. at least $1 - \delta$,*

$$\int R_n(\mathbf{g})\mathrm{d}Q_{n,\eta,\alpha}(\mathbf{g}) \leq \inf_Q \left\{ \frac{\int(\eta - \beta)(r_n(\mathbf{g}) - r_n(\mathbf{f}_n^\star))\mathrm{d}Q(\mathbf{g})}{\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2} + \frac{D_{\mathrm{KL}}(Q||Q_{n,\beta,\alpha}) + 2\log\frac{1}{\delta}}{n(\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2)} \right\}.$$

Using Lemma 13 again, it can be shown that the infimum on the right-hand side is still achieved when $Q = Q_{n,\eta,\alpha}$ (cf. Appendix C.3). This time, the KL divergence $D_{\mathrm{KL}}(Q_{n,\eta,\alpha}||Q_{n,\beta,\alpha})$ contains the log-determinant term $\frac{1}{2}\log(\det(\mathbf{K}_{n,\beta,\alpha}))/\det(\mathbf{K}_{n,\eta,\alpha}))$. Using Corollary 16 again,

$$\frac{1}{2}\log\frac{\det(\mathbf{K}_{n,\beta,\alpha}))}{\det(\mathbf{K}_{n,\eta,\alpha})} = \frac{1}{2}\log\det\frac{2\eta\alpha\mathbf{K}_n + \mathbf{I}}{2\beta\alpha\mathbf{K}_n + \mathbf{I}} = \gamma_n(2\eta\alpha, 2\beta\alpha)\,.$$

One might expect that the bound in Proposition 4 is of the order $[\gamma_n(2\eta\alpha, 2\beta\alpha) + \|f^\star\|^2_{\mathcal{H}}/\alpha]/n$. We will see shortly that this is indeed the case. We sketch the main ideas of the proof of Proposition 4 here, and refer the interested reader to Appendix C. A key idea is to use a prior that assigns higher probability to functions (or vectors of function values) for which the excess risk is small. We define the distribution-dependent prior $\overline{Q}_{n,\beta,\alpha}$ by

$$\frac{\mathrm{d}\overline{Q}_{n,\beta,\alpha}}{\mathrm{d}P_\alpha}(f) \propto \exp\left(-(\beta + 2\sigma^2\beta^2)\sum_{i=1}^n (f(x_i) - f^\star(x_i))^2\right).$$

Note that $\overline{Q}_{n,\beta,\alpha}$ depends on the distribution of $y_1, \ldots, y_n$ through $x_1, \ldots, x_n$ and $f^\star$, but it does not depend on the random draw of $y_1, \ldots, y_n$, and is therefore a valid prior. It is not important for the proof, but one can also notice that $\overline{Q}_{n,\beta,\alpha}$ is another Gaussian process and that its marginal distribution at the inputs $x_1, \ldots, x_n$ is

$$\overline{Q}_{n,\beta,\alpha} = \mathcal{N}(\mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{2\alpha(\beta+2\sigma^2\beta^2)}\mathbf{I})^{-1}\mathbf{f}^\star_n, \alpha\mathbf{K}_n - \alpha\mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{2\alpha(\beta+2\sigma^2\beta^2)}\mathbf{I})^{-1}\mathbf{K}_n)\,.$$

In the bound we want to prove, the prior is $Q_{n,\beta,\alpha}$. It turns out that for all $Q \in \Delta(\mathbb{R}^n)$, the difference between $D_{\mathrm{KL}}(Q||\overline{Q}_{n,\beta,\alpha})$ and $D_{\mathrm{KL}}(Q||Q_{n,\beta,\alpha})$ can be upper bounded with high probability. This is made possible by the following concentration inequality, which is proved using Assumption 1 (cf. Proposition 20). For any $\eta \in \mathbb{R}$ and any $\delta \in (0, 1]$, with probability at least $1 - \delta$,

$$\sup_Q \left\{\int \eta(r_n(\mathbf{f}^\star_n) - r_n(\mathbf{g})) + (\eta - 2\sigma^2\eta^2)R_n(\mathbf{g})\mathrm{d}Q(\mathbf{g}) - D_{\mathrm{KL}}(Q||\overline{Q}_{n,\beta,\alpha})\right\} \le \log\tfrac{1}{\delta}\,.$$

One can then show that, with probability at least $1 - \delta$, for all $Q \in \Delta(\mathbb{R}^n)$,

$$-D_{\mathrm{KL}}(Q||Q_{n,\beta,\alpha}) - \int \beta(r_n(\mathbf{f}^\star_n) - r_n(\mathbf{g})) + (\beta + 2\sigma^2\beta^2)R_n(\mathbf{g})\mathrm{d}Q(\mathbf{g}) \le -D_{\mathrm{KL}}(Q||\overline{Q}_{n,\beta,\alpha}) + \log\tfrac{1}{\delta}\,.$$

By the union bound, these inequalities can be used in succession to deduce that, with probability at least $1 - \delta$,

$$\begin{aligned}
&\sup_Q\left\{\int (\eta - \beta)(r_n(\mathbf{f}^\star_n) - r_n(\mathbf{g})) + (\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2)R_n(\mathbf{g})\mathrm{d}Q(\mathbf{g}) - D_{\mathrm{KL}}(Q||Q_{n,\beta,\alpha})\right\} \\
&\quad\le \sup_Q\left\{\int \eta(r_n(\mathbf{f}^\star_n) - r_n(\mathbf{g})) + (\eta - 2\sigma^2\eta^2)R_n(\mathbf{g})\mathrm{d}Q(\mathbf{g}) - D_{\mathrm{KL}}(Q||\overline{Q}_{n,\beta,\alpha})\right\} + \log\tfrac{2}{\delta} \\
&\quad\le 2\log\tfrac{2}{\delta}\,.
\end{aligned}$$

This inequality can be rearranged into the one in Proposition 4. The $2\log(2/\delta)$ term can be improved to $2\log(1/\delta)$ by replacing the union bound with Cauchy-Schwarz (cf. Appendix C.2). Using Proposition 4, we obtain the main result of this section.

**Theorem 5** *Suppose that Assumption 1 and Assumption 2 are satisfied. For any $n \geq 1$, any $\delta \in (0, 1]$, any $\eta \in (0, \frac{1}{2\sigma^2})$, any $\beta > 0$ such that $\eta - 2\eta^2\sigma^2 - \beta - 2\beta^2\sigma^2 > 0$ and any $\alpha > 0$, with probability at least $1 - \delta$,*

$$\int R_n(\mathbf{g})\mathrm{d}Q_{n,\eta,\alpha}(\mathbf{g}) \leq \frac{2\gamma_n(2\eta\alpha, 2\beta\alpha) + \frac{1}{2\alpha}\|f^\star\|_{\mathcal{H}}^2 + 2\log\frac{1}{\delta}}{n(\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2)}\,.$$

We prove Theorem 5 by evaluating the bound in Proposition 4 at a particular choice of $Q$, and then upper bounding each term by either $\gamma_n(2\eta\alpha, 2\beta\alpha)$ or $\|f^\star\|_{\mathcal{H}}^2$. The gory details can be found in Appendix C.4. By Jensen's inequality, $R_n(\mathbf{m}_{n,\eta,\alpha}) \leq \int R_n(\mathbf{g})\mathrm{d}Q_{n,\eta,\alpha}(\mathbf{g})$. Since $\mathbf{m}_{n,\eta,\alpha}$ is identical to the vector of fitted values for kernel ridge regression (with learning rate $2\eta\alpha$), the excess risk bound in Theorem 5 also applies to kernel ridge regression.

## 5. Rates of Convergence for Mercer Kernels

We provide worst-case bounds on the relative information gain, which, when combined with Theorem 5, can be used to provide explicit rates of convergence for the excess risk of $Q_{n,\eta,\alpha}$.

### 5.1. Bounds on the Relative Information Gain

Aside from one or two tricks, the method with which we bound the relative information gain is identical to the method that Vakili et al. (2021) used to bound the information gain. As a result, we require the same regularity assumptions. A positive-definite kernel $k$ on a set $\mathcal{X}$ is called a Mercer kernel if $\mathcal{X}$ is a compact metric space and the kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous. Due to Mercer's theorem (cf. Appendix D.1), any Mercer kernel can be expressed as an infinite sum of the form

$$k(x, x') = \sum_{i=1}^{\infty} \xi\phi_i(x)\phi_i(x')\,,$$

where $(\xi_i)_{i=1}^{\infty}$ and $(\phi_i)_{i=1}^{\infty}$ are the (non-negative) eigenvalues and eigenfunctions of the kernel.

**Assumption 3** *Assume that $k$ is a Mercer kernel, and that $\sup_{i\in\mathbb{N}} \|\phi_i\|_\infty \leq \psi$.*

The rate at which the eigenvalues $\xi_1, \xi_2, \ldots$ of the kernel decay to zero determines the complexity of the corresponding RKHS. The two eigenvalue decay conditions studied by Vakili et al. (2021) (and many others) are defined as follows.

**Assumption 4** *The polynomial eigenvalue decay condition is satisfied if there exist $C_p > 0$ and $\beta_p > 1$ such that*

$$\xi_i \leq C_p i^{-\beta_p}.$$

*The exponential eigenvalue decay condition is satisfied if there exist $C_{e_1}, C_{e_2} > 0$ and $\beta_e \in (0, 1]$ such that*

$$\xi_i \leq C_{e_1}\exp(-C_{e_2} i^{\beta_e}).$$

If a kernel is known to satisfy one of these eigenvalue decay conditions, then upper bounds on $\gamma_n(\eta)$ with explicit dependence on $n$ can be given (cf. Corollary 1 in Vakili et al., 2021). In particular, under the polynomial eigenvalue decay condition, $\gamma_n(\eta) = \mathcal{O}((n\eta)^{1/\beta_p} \log^{1-1/\beta_p}(n\eta))$. Under the exponential eigenvalue decay condition, $\gamma_n(\eta) = \mathcal{O}(\log^{1+1/\beta_e}(n\eta))$. A central idea in the method used by Vakili et al. (2021) is to separate the kernel function into $k_{\|}(x,x') := \sum_{i=1}^{D} \xi_i \phi_i(x)\phi_i(x')$ and $k_{\perp}(x,x') := \sum_{D+1}^{\infty} \xi_i \phi_i(x)\phi_i(x')$. It can be seen that $k_{\|}$ is the reproducing kernel of the subspace of $\mathcal{H}$ spanned by $\phi_1, \dots, \phi_D$ and that $k_{\|}(x,x')$ is the reproducing kernel of the subspace of $\mathcal{H}$ which is orthogonal to $\phi_1, \dots, \phi_D$. Let us write $\mathbf{K}_n^{\|}$ and $\mathbf{K}_n^{\perp}$ for the corresponding kernel matrices. We define

$$\delta_D := \sum_{i=D+1}^{\infty} \xi_i \psi^2 \,.$$

The first step is to re-write the information gain as a term depending on the rank $D$ kernel matrix $\mathbf{K}_n^{\|}$ and another term depending on both $\mathbf{K}_n^{\|}$ and $\mathbf{K}_n^{\perp}$. Vakili et al. (2021) show that

$$\gamma_n(\eta) = \frac{1}{2}\log\det(\eta\mathbf{K}_n^{\|} + \mathbf{I}) + \frac{1}{2}\log\det(\mathbf{I} + \eta(\mathbf{I} + \eta\mathbf{K}_n^{\|})^{-1}\mathbf{K}_n^{\perp}) \,. \tag{5}$$

From here, Vakili et al. (2021) show that the first term is bounded is by $\frac{1}{2}D\log(1 + \frac{\overline{k}n\eta}{D})$, where $\overline{k} = \sup_x |k(x,x)|$, and that the second term is bounded by $\frac{1}{2}n\eta\delta_D$. One then has the upper bound

$$\gamma_n(\eta) \leq \frac{1}{2}D\log\left(1 + \frac{\overline{k}n\eta}{D}\right) + \frac{1}{2}n\eta\delta_D \,.$$

The relative information gain satisfies a similar, but slightly tighter bound.

**Proposition 6** *For all $\eta > \beta > 0$ and any integer $D \geq 1$,*

$$\gamma_n(\eta,\beta) \leq \frac{1}{2}D\log\frac{\eta}{\beta} + \frac{1}{2}n\eta\delta_D \,.$$

The full proof of Proposition 6 can be found in Appendix D.3. From (5), it follows that the relative information gain can be re-written as

$$\gamma_n(\eta,\beta) = \frac{1}{2}\log\frac{\det(\eta\mathbf{K}_n^{\|} + \mathbf{I})}{\det(\beta\mathbf{K}_n^{\|} + \mathbf{I})} + \frac{1}{2}\log\frac{\det(\mathbf{I} + \eta(\mathbf{I} + \eta\mathbf{K}_n^{\|})^{-1}\mathbf{K}_n^{\perp})}{\det(\mathbf{I} + \beta(\mathbf{I} + \beta\mathbf{K}_n^{\|})^{-1}\mathbf{K}_n^{\perp})} \,.$$

Since $\log\det(\mathbf{I} + \beta(\mathbf{I} + \beta\mathbf{K}_n^{\|})^{-1}\mathbf{K}_n^{\perp}) \geq 0$ (cf. Lemma 25), the second term is still upper bounded by $\frac{1}{2}n\eta\delta_D$. In Lemma 24, we show that the first term can be bounded by $\frac{1}{2}D\log\frac{\eta}{\beta}$, saving a factor of $\log(n)$. Using Proposition 6, one can upper bound the relative information gain based on the spectral decay of the kernel via $\delta_D$.

**Proposition 7** *If the polynomial eigenvalue decay condition is satisfied, then*

$$\gamma_n(\eta,\beta) \leq (n\eta C_p\psi^2)^{1/\beta_p}\log^{1-1/\beta_p}(\tfrac{\eta}{\beta}) + \log\tfrac{\eta}{\beta} \,.$$

*If the exponential eigenvalue decay condition is satisfied, then*

$$\gamma_n(\eta,\beta) \leq \frac{1}{2}\left(\frac{2}{C_{e,2}}\log(n\eta C_{\beta_e})\right)^{1/\beta_e}\log\frac{\eta}{\beta} + \frac{1}{2}\log\frac{e\eta}{\beta} \,,$$

*where $C_{\beta_e} = \frac{C_{e,1}\psi^2}{C_{e,2}}$ if $\beta_e = 1$, and $C_{\beta_e} = \frac{2C_{e,1}\psi^2}{C_{e,2}\beta_e}(\frac{2-2\beta_e}{C_{e,2}\beta_e})^{1/\beta_e - 1}\exp(\frac{1-\beta_e}{\beta_e})$ if $\beta_e \in (0,1)$.*

If $\beta \propto \eta$, then the bound for polynomial decay can be simplified to $\gamma_n(\eta, \beta) = \mathcal{O}((n\eta)^{1/\beta_e})$. In this case, the bound for exponential decay can be simplified to $\gamma_n(\eta, \beta) = \mathcal{O}(\log^{1/\beta_e}(n\eta))$. The proof is very similar to that of Corollary 1 in Vakili et al. (2021), and can be found in Appendix D.4.

### 5.2. Rates of Convergence

For the case of polynomial eigenvalue decay, by plugging the bound in Proposition 7 into Theorem 5 and then choosing suitable values of $\eta$ and $\alpha$, we obtain the following rate of convergence.

**Corollary 8** *Suppose that the polynomial eigenvalue decay condition is satisfied,* $\eta = \frac{1}{4\sigma^2}$ *and* $\alpha = n^{-\frac{1}{1+\beta_p}}$. *For any* $\delta \in (0, 1]$, *with probability at least* $1 - \delta$,

$$\int R_n(\mathbf{g}) \mathrm{d}Q_{n,\eta,\alpha}(\mathbf{g}) = \mathcal{O}\left(n^{-\frac{\beta_p}{1+\beta_p}}\right).$$

The proof can be found in Appendix E.1. The rate of convergence in Corollary 8 is minimax-optimal (Dicker et al., 2017). For instance, for kernels satisfying the polynomial decay condition with $\beta_p = 2q$, the unit ball of the corresponding RKHS $\mathcal{H}$ is a Sobolev space of $q - 1$ times absolutely continuous and differentiable functions. In this case, we recover the standard rate of $n^{-\frac{2q}{1+2q}}$ for nonparametric regression (Tsybakov, 2009). By tuning $\alpha$ according to $\sigma$, we can also match the dependence on $\sigma$ in the minimax lower bound for generalised Sobolev ellipsoids in Example 15.23 of Wainwright (2019). Note, however, that this example considers regression with random design rather than fixed design. For exponential decay, we obtain the following rate of convergence.

**Corollary 9** *Suppose that the exponential eigenvalue decay condition is satisfied,* $\eta = \frac{1}{4\sigma^2}$ *and* $\alpha = 1$. *For any* $\delta \in (0, 1]$, *with probability at least* $1 - \delta$,

$$\int R_n(\mathbf{g}) \mathrm{d}Q_{n,\eta,\alpha}(\mathbf{g}) = \mathcal{O}\left(\frac{\log^{1/\beta_e}(n)}{n}\right).$$

The proof can be found in Appendix E.2. This rate of convergence for exponential eigenvalue decay is also minimax-optimal (Dicker et al., 2017).

## 6. Discussion

We have introduced a new quantity called the relative information gain, which measures the sensitivity of the information gain with respect to the variance of the noise in the responses. We demonstrated that the relative information gain arises naturally from the complexity term of a PAC-Bayesian excess risk bound for Gaussian process regression. Finally, we proved bounds on the relative information gain. When we combined these bounds with the excess risk bound in Theorem 5, we recovered minimax-optimal rates of convergence.

In future work, the extent to which the PAC-Bayesian bounds in this paper can be generalised to other regression models could be invesitgated. For instance, one could consider regression with random design, misspecified regression or regression with sparse or other approximate Gaussian process predictors.

## Acknowledgments

We would like to thank David Reeb for numerous helpful discussions about and comments on an early version of this work. Hamish was funded by the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme (grant agreement 950180).

## References

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024.

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *Annals of Statistics*, 48(3):1475–1497, 2020.

Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Conference on Learning Theory*, pages 533–557. PMLR, 2019.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Ismaël Castillo. Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics*, 2:1281–1299, 2008.

Ismaël Castillo. On Bayesian supremum norm contraction rates. *Annals of Statistics*, 42(5):2058–2091, 2014.

Olivier Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. *Lecture Notes-Monograph Series. IMS*, 1277, 2007.

Olivier Catoni and Ilaria Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2006.

Lee H Dicker, Dean P Foster, and Daniel Hsu. Kernel methods and regularization techniques for nonparametric regression: Minimax optimality and adaptation. *Rutgers Univ., New Brunswick, NJ, USA, Tech. Rep*, 2015.

Lee H Dicker, Dean P Foster, and Daniel Hsu. Kernel ridge vs. principal component regression: minimax bounds and adaptability of regularization operators. *Electronic Journal of Statistics*, 11 (1):1022–1047, 2017.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time—iii. *Communications on pure and applied Mathematics*, 29(4): 389–461, 1976.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning, 2009.

Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj, Maxim Raginsky, et al. Generalization bounds: Perspectives from information theory and PAC-Bayes. *Foundations and Trends® in Machine Learning*, 18(1):1–223, 2025.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

EL Mahdi Khribch and Pierre Alquier. Convergence of statistical estimators via mutual information bounds. *arXiv preprint arXiv:2412.18539*, 2024.

David A McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234, 1998.

Richard Nickl and Jakob Söhl. Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *Annals of Statistics*, 45(4):1664–1693, 2017.

Debdeep Pati, Anirban Bhattacharya, and Guang Cheng. Optimal Bayesian estimation in random covariate design with a rescaled Gaussian process prior. *Journal of Machine Learning Research*, 16(87):2837–2851, 2015.

David Reeb, Andreas Doerr, Sebastian Gerwinn, and Barbara Rakitsch. Learning Gaussian processes by minimizing PAC-Bayesian generalization bounds. *Advances in Neural Information Processing Systems*, 31, 2018.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Matthias Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3(Oct):233–269, 2002.

Matthias Seeger. *Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations*. PhD thesis, University of Edinburgh, 2003.

Matthias W Seeger, Sham M Kakade, and Dean P Foster. Information consistency of nonparametric gaussian process methods. *IEEE Transactions on Information Theory*, 54(5):2376–2382, 2008.

John Shawe-Taylor and Robert C Williamson. A PAC analysis of a Bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 2–9, 1997.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proc. International Conference on Machine Learning (ICML)*, 2010.

Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In *COLT*, pages 79–93, 2009.

Taiji Suzuki. PAC-Bayesian bound for Gaussian process regression and multiple kernel additive model. In *Conference on Learning Theory*, pages 8–1. JMLR Workshop and Conference Proceedings, 2012.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*, volume 1. Springer New York, NY, 2009.

Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in Gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 82–90. PMLR, 2021.

Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nelo Cristianini. Finite-time analysis of kernelised contextual bandits. In *Uncertainty in Artificial Intelligence*, 2013.

Aad van der Vaart and Harry van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Annals of Statistics*, 36(3):1435–1463, 2008.

Aad van der Vaart and Harry van Zanten. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12(6), 2011.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Larry Wasserman. *All of nonparametric statistics*. Springer, 2006.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

David Williams. *Probability with martingales*. Cambridge university press, 1991.

Houssam Zenati, Alberto Bietti, Eustache Diemert, Julien Mairal, Matthieu Martin, and Pierre Gaillard. Efficient kernelized UCB for contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 5689–5720. PMLR, 2022.

Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural computation*, 17(9):2077–2098, 2005.

## Appendix A. Auxiliary Lemmas for Section 3

First, we find an expression for the effective dimension in terms of the eigenvalues of the kernel matrix. This expression is well-known (see e.g. , Chapter 2.6 in Williams and Rasmussen, 2006 or Chapter 5.4 in Hastie et al., 2009).

**Lemma 10** *Let $(\lambda_i)_{i=1}^n$ be the eigenvalues of $\mathbf{K}_n$. For all $n \geq 1$ and $\eta > 0$,*

$$d_n(\eta) = \sum_{i=1}^{n} \frac{\eta\lambda_i}{1+\eta\lambda_i} \,.$$

**Proof** Let $(\mathbf{v}_i)_{i=1}^n$ be the eigenvectors of $\mathbf{K}_n$. First, we notice that for all $i \in [n]$,

$$(\mathbf{K}_n + \tfrac{1}{\eta}\mathbf{I})\mathbf{v}_i = (\lambda_i + 1/\eta)\mathbf{v}_i \,.$$

It follows that

$$\mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{\eta}\mathbf{I})^{-1}\mathbf{v}_i = \tfrac{1}{\lambda_i + 1/\eta}\mathbf{K}_n\mathbf{v}_i = \tfrac{\eta\lambda_i}{1+\eta\lambda_i}\mathbf{v}_i \,. \tag{6}$$

Since the trace of a matrix is equal to the sum of its eigenvalues, this concludes the proof. ■

**Lemma 11** *Let $(\lambda_i)_{i=1}^n$ be the eigenvalues of $\mathbf{K}_n$. For all $n \geq 1$ and $\eta \geq 0$,*

$$\gamma_n(\eta) = \frac{1}{2}\sum_{i=1}^n \log(1 + \eta\lambda_i) \,.$$

**Proof** Let $(\mathbf{v}_i)_{i=1}^n$ be the eigenvectors of $\mathbf{K}_n$. We notice that for all $i \in [n]$,

$$(\eta\mathbf{K}_n + \mathbf{I})\mathbf{v}_i = (1 + \eta\lambda_i)\mathbf{v}_i \,.$$

Since the determinant of a matrix is equal to the product of its eigenvalues,

$$\gamma_n(\eta) = \frac{1}{2}\log\det(\eta\mathbf{K}_n + \mathbf{I}) = \frac{1}{2}\log\prod_{i=1}^n(1 + \eta\lambda_i) = \frac{1}{2}\sum_{i=1}^n \log(1 + \eta\lambda_i) \,.$$

This concludes the proof. ■

An obvious consequence of Lemma 11 is that for all $n \geq 1$ and $\eta > \beta \geq 0$,

$$\gamma_n(\eta, \beta) = \frac{1}{2}\sum_{i=1}^n \log\frac{1 + \eta\lambda_i}{1 + \beta\lambda_i} \,.$$

**Lemma 12** *For any $\lambda \geq 0$, the function $f(\beta) := \frac{1}{\eta-\beta}\log\frac{1+\eta\lambda}{1+\beta\lambda}$ is decreasing on $[0, \eta)$.*

**Proof** One can verify that the derivative of $f$ is

$$f'(\beta) = \frac{1}{(\eta-\beta)^2}\log\frac{1+\eta\lambda}{1+\beta\lambda} - \frac{\lambda}{(\eta-\beta)(1+\beta\lambda)} \,.$$

Using the inequality $\log(1 + x) \leq x$ for $x \geq 0$, we obtain

$$\frac{1}{(\eta-\beta)^2}\log\frac{1+\eta\lambda}{1+\beta\lambda} = \frac{1}{(\eta-\beta)^2}\log\left(1 + \frac{(\eta-\beta)\lambda}{1+\beta\lambda}\right) \leq \frac{\lambda}{(\eta-\beta)(1+\beta\lambda)} \,.$$

It follows that $f'(\beta) \leq 0$, and so $f$ must be decreasing on $[0, \eta)$. ■

## Appendix B. Effective Dimension and Minimum Mean Squared Error

We explain how Proposition 1 can be derived from Theorem 1 in Guo et al. (2005). We write the model in (1) as

$$\mathbf{y}_n = \mathbf{f}_n^\star + \boldsymbol{\varepsilon}_n \,,$$

where $\mathbf{f}_n^\star \sim \mathcal{N}(0, \mathbf{K}_n)$ and $\boldsymbol{\varepsilon}_n \sim \mathcal{N}(0, (1/\eta)\mathbf{I})$. For an estimate $f(\mathbf{y}_n)$ of $\mathbf{f}_n^\star$, the mean squared error is $\mathbb{E}[\|f(\mathbf{y}_n) - \mathbf{f}_n^\star\|_2^2]$. The minimum achievable mean squared error (w.r.t. the choice of $f$) depends on $\eta$. Let us therefore define the function

$$\mathrm{MMSE}(\eta) := \min_f \mathbb{E}[\|f(\mathbf{y}_n) - \mathbf{f}_n^\star\|_2^2] \,.$$

Theorem 1 in Guo et al. (2005) states that

$$\frac{1}{2}\mathrm{MMSE}(\eta) = \gamma_n'(\eta) \,.$$

If we can show that $\mathrm{MMSE}(\eta) = d_n(\eta)/\eta$, then this matches the identity in Proposition 1. By the least-squares property of the conditional expectation (see e.g. , Chapter 9.4 in Williams, 1991), the minimum mean squared error w.r.t. $f$ is achieved when

$$f(\mathbf{y}_n) = \mathbb{E}[\mathbf{f}_n^\star | \mathbf{y}_n] = \mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{\eta}\mathbf{I})^{-1}\mathbf{y}_n \,.$$

Thus we can express the minimum mean squared error as

$$\mathrm{MMSE}(\eta) = \mathbb{E}[\|\mathbb{E}[\mathbf{f}_n^\star|\mathbf{y}_n] - \mathbf{f}_n^\star\|_2^2] = \mathbb{E}[\|\mathbf{f}_n^\star\|_2^2 - \|\mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{\eta}\mathbf{I})^{-1}\mathbf{y}_n\|_2^2] \,.$$

Since $\mathbf{K}_n(\mathbf{K}_n + \frac{1}{\eta}\mathbf{I})^{-1}\mathbf{y}_n \sim \mathcal{N}(0, \mathbf{K}_n(\mathbf{K}_n + \frac{1}{\eta}\mathbf{I})^{-1}\mathbf{K}_n)$, we obtain

$$\begin{aligned}
\mathbb{E}[\|\mathbf{f}_n^\star\|_2^2 - \|\mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{\eta})^{-1}\mathbf{y}_n\|_2^2] &= \mathrm{tr}(\mathbf{K}_n) - \mathrm{tr}(\mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{\eta})^{-1}\mathbf{K}_n) \\
&= \mathrm{tr}(\mathbf{K}_n(\mathbf{I} - (\mathbf{K}_n + \tfrac{1}{\eta})^{-1}\mathbf{K}_n)) \\
&= \tfrac{1}{\eta}\mathrm{tr}(\mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{\eta}\mathbf{I})^{-1}) \,.
\end{aligned}$$

We conclude that $\mathrm{MMSE}(\eta) = d_n(\eta)/\eta$.

## Appendix C. PAC-Bayesian Bounds

### C.1. Auxiliary Lemmas

We will use the following variational representation of the KL divergence, which was proved by Donsker and Varadhan (1976).

**Lemma 13** *For any measurable function $h : \mathbb{R}^n \to \mathbb{R}$ and any probability measure $P \in \Delta(\mathbb{R}^n)$ such that $\int \exp(h(\mathbf{g}))\mathrm{d}P(\mathbf{g}) < \infty$,*

$$\sup_{Q \in \Delta(\mathbb{R}^n)} \left\{ \int h(\mathbf{g})\mathrm{d}Q(\mathbf{g}) - D_{\mathrm{KL}}(Q||P) \right\} = \log \int \exp(h(\mathbf{g}))\mathrm{d}P(\mathbf{g}) \,.$$

*If $h$ is bounded above, then the supremum is achieved when*

$$\frac{\mathrm{d}Q}{\mathrm{d}P}(\mathbf{g}) \propto \exp(h(\mathbf{g})) \,.$$

By rearranging the statement involving the supremum, we see that

$$\inf_{Q\in\Delta(\mathbb{R}^n)}\left\{\int h(\mathbf{g})\mathrm{d}Q(\mathbf{g})+D_{\mathrm{KL}}(Q||P)\right\}=-\log\int\exp(-h(\mathbf{g}))\mathrm{d}P(\mathbf{g})\,.$$

If $h$ is bounded below, then the infimum is achieved when

$$\frac{\mathrm{d}Q}{\mathrm{d}P}(\mathbf{g})\propto\exp(-h(\mathbf{g}))\,.$$

**Lemma 14** *The function* $f(\zeta):=\|(\mathbf{K}_n(\mathbf{K}_n+\zeta\mathbf{I})^{-1}-\mathbf{I})\mathbf{y}_n\|_2^2$ *is increasing on* $[0,\infty)$.

**Proof** We notice that

$$\begin{aligned}\mathbf{K}_n(\mathbf{K}_n+\zeta\mathbf{I})^{-1}-\mathbf{I}&=(\mathbf{K}_n+\zeta\mathbf{I})(\mathbf{K}_n+\zeta\mathbf{I})^{-1}-\mathbf{I}-\zeta(\mathbf{K}_n+\zeta\mathbf{I})^{-1}\\&=-\zeta(\mathbf{K}_n+\zeta\mathbf{I})^{-1}\,.\end{aligned}$$

Thus from (6), we see that the eigenvalues of $(\mathbf{K}_n(\mathbf{K}_n+\zeta\mathbf{I})^{-1}-\mathbf{I})^2$ are $(\zeta^2/(\zeta+\lambda_i)^2)_{i=1}^n$. Fix $\zeta_2\geq\zeta_1\geq 0$. For each $i\in[n]$,

$$\zeta_1^2(\zeta_2+\lambda_i)^2=\zeta_1^2\zeta_2^2+2\zeta_1^2\zeta_2\lambda_i+\zeta_1^2\lambda_i^2\leq\zeta_1^2\zeta_2^2+2\zeta_1\zeta_2^2\lambda_i+\zeta_2^2\lambda_i^2=\zeta_2^2(\zeta_1+\lambda_i)^2\,.$$

It follows that each of the functions $g_i(\zeta):=\zeta^2/(\zeta+\lambda_i)^2$ is increasing on $[0,\infty)$. Hence,

$$(\mathbf{K}_n(\mathbf{K}_n+\zeta_1\mathbf{I})^{-1}-\mathbf{I})^2\preccurlyeq(\mathbf{K}_n(\mathbf{K}_n+\zeta_2\mathbf{I})^{-1}-\mathbf{I})^2\,.$$

This concludes the proof. ■

**Lemma 15** *For all* $n\geq 1$, $\eta>0$ *and* $\alpha\geq 0$,

$$\mathbf{K}_{n,\eta,\alpha}=\alpha\mathbf{K}_n(2\eta\alpha\mathbf{K}_n+\mathbf{I})^{-1}\,.$$

**Proof** By adding zero to $\mathbf{K}_{n,\eta,\alpha}$, we obtain

$$\begin{aligned}\mathbf{K}_{n,\eta,\alpha}&=\alpha\mathbf{K}_n-\alpha\mathbf{K}_n(\mathbf{K}_n+\tfrac{1}{2\eta\alpha}\mathbf{I})^{-1}(\mathbf{K}_n+\tfrac{1}{2\eta\alpha}\mathbf{I}-\tfrac{1}{2\eta\alpha}\mathbf{I})\\&=\tfrac{1}{2\eta}\mathbf{K}_n(\mathbf{K}_n+\frac{1}{2\eta\alpha}\mathbf{I})^{-1}\\&=\alpha\mathbf{K}_n(2\eta\alpha\mathbf{K}_n+\mathbf{I})^{-1}\,.\end{aligned}$$

This concludes the proof. ■

As a result of Lemma 15, $\mathrm{tr}(\mathbf{K}_{n,\eta,\alpha})=\frac{1}{2\eta}d_n(2\eta\alpha)$.

**Corollary 16** *For all* $n\geq 1$, $\eta>0$ *and* $\alpha\geq 0$,

$$\alpha\mathbf{K}_{n,\eta,\alpha}^{-1}\mathbf{K}_n=2\eta\alpha\mathbf{K}_n+\mathbf{I}\,. \tag{7}$$

**Proof** From Lemma 15, we have the identity

$$\alpha \mathbf{K}_{n,\eta,\alpha}^{-1} \mathbf{K}_n (2\eta\alpha \mathbf{K}_n + \mathbf{I})^{-1} = \mathbf{I} \,.$$

Post-multiplying both sides by $2\eta\alpha\mathbf{K}_n + \mathbf{I}$ gives the desired result. ■

**Corollary 17** *For all $n \geq 1$, $\eta > \beta > 0$ and $\alpha \geq 0$,*

$$\frac{1}{2} \log \frac{\det(\mathbf{K}_{n,\beta,\alpha})}{\det(\mathbf{K}_{n,\eta,\alpha})} = \gamma_n(2\eta\alpha, 2\beta\alpha) \,.$$

**Proof** Using Lemma 15 and standard properties of determinants,

$$\frac{1}{2} \log \frac{\det(\mathbf{K}_{n,\beta,\alpha})}{\det(\mathbf{K}_{n,\eta,\alpha})} = \frac{1}{2} \log \frac{\det((2\beta\alpha\mathbf{K}_n + \mathbf{I})^{-1})}{\det((2\eta\alpha\mathbf{K}_n + \mathbf{I})^{-1})} = \frac{1}{2} \log \frac{\det(2\eta\alpha\mathbf{K}_n + \mathbf{I})}{\det(2\beta\alpha\mathbf{K}_n + \mathbf{I})} \,.$$

This concludes the proof. ■

**Corollary 18** *For all $n \geq 1$, $\eta > \beta > 0$ and $\alpha > 0$,*

$$\operatorname{tr}(\mathbf{K}_{n,\eta,\alpha}) \leq \frac{1}{\eta - \beta} \gamma_n(2\eta\alpha, 2\beta\alpha) \,.$$

**Proof** From Lemma 15 and Proposition 2, we obtain

$$\operatorname{tr}(\mathbf{K}_{n,\eta,\alpha}) = \frac{1}{2\eta} d_n(2\eta\alpha) \leq \frac{1}{2\eta} \frac{4\eta\alpha}{2\eta\alpha - 2\beta\alpha} \gamma_n(2\eta\alpha, 2\beta\alpha) = \frac{1}{\eta - \beta} \gamma_n(2\eta\alpha, 2\beta\alpha) \,.$$

This concludes the proof. ■

**Lemma 19** *For all $n \geq 1$, $\eta > \beta > 0$ and $\alpha > 0$,*

$$\operatorname{tr}(\mathbf{K}_{n,\beta,\alpha}^{-1} \mathbf{K}_{n,\eta,\alpha}) \leq n \,.$$

**Proof** Let $(\lambda_i)_{i=1}^n$ and $(\mathbf{v}_i)_{i=1}^n$ be the eigenvalues and eigenvectors of $\mathbf{K}_n$. From Lemma 15 and (6), it follows that, for all $i \in [n]$,

$$\mathbf{K}_{n,\eta,\alpha} \mathbf{v}_i = \frac{\alpha\lambda_i}{2\eta\alpha\lambda_i + 1} \mathbf{v}_i \,.$$

As long as $\mathbf{K}_n$ is positive-definite, this implies that, for all $i \in [n]$,

$$\mathbf{K}_{n,\beta,\alpha}^{-1} \mathbf{v}_i = \frac{2\beta\alpha\lambda_i + 1}{\alpha\lambda_i} \mathbf{v}_i \,.$$

In particular, the eigenvalues of $\mathbf{K}_{n,\beta,\alpha}^{-1} \mathbf{K}_{n,\eta,\alpha}$ are $(\frac{2\beta\alpha\lambda_i + 1}{2\eta\alpha\lambda_i + 1})_{i=1}^n$. Therefore,

$$\operatorname{tr}(\mathbf{K}_{n,\beta,\alpha}^{-1} \mathbf{K}_{n,\eta,\alpha}) = \sum_{i=1}^n \frac{2\beta\alpha\lambda_i + 1}{2\eta\alpha\lambda_i + 1} \leq \sum_{i=1}^n \frac{2\eta\alpha\lambda_i + 1}{2\eta\alpha\lambda_i + 1} = n \,.$$

This concludes the proof. ■

### C.2. A PAC-Bayesian Bound for Sums of Independent Random Variables

Proposition 4 is a special case of a localised PAC-Bayesian bound for collections of sums of independent random variables. We consider a collection of random variables $(Z_i(\mathbf{g}))_{i\in[n],\mathbf{g}\in\mathbb{R}^n}$, such that for each $\mathbf{g}\in\mathbb{R}^n$, $(Z_i(\mathbf{g}))_{i=1}^n$ is a sequence of independent random variables. We let $\mu_i(\mathbf{g}) := \mathbb{E}[Z_i(\mathbf{g})]$ and $\psi_i(\mathbf{g},\eta) := \log\mathbb{E}[\exp(\eta(Z_i(\mathbf{g})-\mu_i(\mathbf{g}))]$ denote the mean and the cumulant generating function of $Z_i(\mathbf{g})$. We assume that for all $\eta\in\mathbb{R}$, $i\in[n]$ and $\mathbf{g}\in\mathbb{R}^n$, $\mathbb{E}[\exp(\eta(Z_i(\mathbf{g})-\mu_i(\mathbf{g}))]<\infty$. For an arbitrary distribution $P\in\Delta(\mathbb{R}^n)$, we define the localised prior $Q_{\mu,\beta}$ by

$$\frac{\mathrm{d}Q_{\mu,\beta}}{\mathrm{d}P}(\mathbf{g}) \propto \exp\left(\sum_{i=1}^n \beta\mu_i(\mathbf{g}) - \psi_i(\mathbf{f}_n,-\beta)\right).$$

Note that while $Q_{\mu,\beta}$ depends the unobserved quantities $\mu_i(\mathbf{g})$ and $\psi_i(\mathbf{g},-\beta)$, it does not depend on the random draw of $(Z_i(\mathbf{g}))_{i\in[n],\mathbf{g}\in\mathbb{R}^n}$. We define the empirical approximation $Q_{Z,\beta}$ of the localised prior by

$$\frac{\mathrm{d}Q_{Z,\beta}}{\mathrm{d}P}(\mathbf{g}) \propto \exp\left(\sum_{i=1}^n \beta Z_i(\mathbf{g})\right)$$

In contrast to $Q_{\mu,\beta}$, $Q_{Z,\beta}$ does depend on the random draw of $(Z_i(\mathbf{g}))_{i\in[n],\mathbf{g}\in\mathbb{R}^n}$, but does not depend on any unobservable quantities. The following proposition combines a well-known PAC-Bayesian bound for sums of independent random variables (cf. Proposition 2.1 in Catoni and Giulini, 2017) with Catoni's localisation technique (cf. Section 1.3.4 in Catoni, 2007).

**Proposition 20** *For any $n\geq 1$, any $\delta\in(0,1]$, any $\eta>0$, any $\beta\in[0,\eta)$ and any $P\in\Delta(\mathbb{R}^n)$, with probability at least $1-\delta$, $\forall Q\in\Delta(\mathbb{R}^n)$,*

$$\int\left[\sum_{i=1}^n(\eta-\beta)(Z_i(\mathbf{g})-\mu_i(\mathbf{g})) - \psi_i(\mathbf{g},\eta) - \psi_i(\mathbf{g},-\beta)\right]\mathrm{d}Q(\mathbf{g}) - D_{\mathrm{KL}}(Q||Q_{\beta Z_n}) \leq 2\log\tfrac{1}{\delta}\,.$$

The proof is a fairly straightforward combination of the proofs of Proposition 2.1 from Catoni and Giulini (2017) and some of the derivations in Section 1.3.4 in Catoni (2007).

**Proof** Fix an arbitrary $Q\in\Delta(\mathbb{R}^n)$. We begin by finding a relationship between $D_{\mathrm{KL}}(Q||Q_{\mu,\beta})$ and $D_{\mathrm{KL}}(Q||Q_{Z,\beta})$. From the definitions of $Q_{\mu,\beta}$ and $Q_{Z,\beta}$, we obtain

$$\begin{aligned}
D_{\mathrm{KL}}(Q||Q_{\mu,\beta}) &= \int \log\frac{\mathrm{d}Q}{\mathrm{d}Q_{Z,\beta}}(\mathbf{g})\frac{\mathrm{d}Q_{Z,\beta}}{\mathrm{d}Q_{\mu,\beta}}(\mathbf{g})\mathrm{d}Q(\mathbf{g}) \qquad (8)\\
&= D_{\mathrm{KL}}(Q||Q_{Z,\beta})\\
&+ \int \log\frac{\exp(\sum_{i=1}^n\beta Z_i(\mathbf{g}))\int\exp(\sum_{i=1}^n\beta\mu_i(\mathbf{g})-\psi_i(\mathbf{g},-\beta))\mathrm{d}P(\mathbf{g})}{\exp(\sum_{i=1}^n\beta\mu_i(\mathbf{g})-\psi_i(\mathbf{g},-\beta))\int\exp(\sum_{i=1}^n\beta Z_i(\mathbf{g}))\mathrm{d}P(\mathbf{g})}\mathrm{d}Q(\mathbf{g})\\
&= D_{\mathrm{KL}}(Q||Q_{Z,\beta}) + \int\left[\sum_{i=1}^n\beta Z_i(\mathbf{g}) - \beta\mu_i(\mathbf{g}) + \psi_i(\mathbf{g},-\beta)\right]\mathrm{d}Q(\mathbf{g})\\
&+ \log\frac{\int\exp\left(\sum_{i=1}^n\beta\mu_i(\mathbf{g})-\psi_i(\mathbf{g},-\beta)\right)\mathrm{d}P(\mathbf{g})}{\int\exp\left(\sum_{i=1}^n\beta Z_i(\mathbf{g})\right)\mathrm{d}P(\mathbf{g})}\,.
\end{aligned}$$

Using Lemma 13, Tonelli's theorem and independence, for any $\eta > 0$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\sup_{Q\in\Delta(\mathbb{R}^n)}\left\{\int\sum_{i=1}^n\left[\eta(Z_i(\mathbf{g})-\mu_i(\mathbf{g}))-\psi_i(\mathbf{g},\eta)\right]\mathrm{d}Q(\mathbf{g})-D_{\mathrm{KL}}(Q||Q_{\mu,\beta})\right\}\right)\right] \qquad (9)\\
&=\mathbb{E}\left[\int\exp\left(\sum_{i=1}^n\eta(Z_i(\mathbf{g})-\mu_i(\mathbf{g}))-\psi_i(\mathbf{g},\eta)\right)\mathrm{d}Q_{\mu,\beta}(\mathbf{g})\right]\\
&=\int\mathbb{E}\left[\exp\left(\sum_{i=1}^n\eta(Z_i(\mathbf{g})-\mu_i(\mathbf{g}))-\psi_i(\mathbf{g},\eta)\right)\right]\mathrm{d}Q_{\mu,\beta}(\mathbf{g})\\
&=\int\prod_{i=1}^n\mathbb{E}\left[\exp\left(\eta(Z_i(\mathbf{g})-\mu_i(\mathbf{g}))-\psi_i(\mathbf{g},\eta)\right)\right]\mathrm{d}Q_{\mu,\beta}(\mathbf{g})\\
&=\int\prod_{i=1}^n\frac{\exp(\psi_i(\mathbf{g},\eta))}{\exp(\psi_i(\mathbf{g},\eta))}\mathrm{d}Q_{\mu,\beta}(\mathbf{g})=1\,.
\end{aligned}$$

Next, using Lemma 13 and Jensen's inequality, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\frac{\int\exp\left(\sum_{i=1}^n\beta\mu_i(\mathbf{g})-\psi_i(\mathbf{g},-\beta)\right)\mathrm{d}P(\mathbf{g})}{\int\exp\left(\sum_{i=1}^n\beta Z_i(\mathbf{g})\right)\mathrm{d}P(\mathbf{g})}\right] \qquad (10)\\
&=\mathbb{E}\left[\frac{\exp(\sup_{Q\in\Delta(\mathbb{R}^d)}\{\int\sum_{i=1}^n\beta\mu_i(\mathbf{g})-\psi_i(\mathbf{g},-\beta)\mathrm{d}Q(\mathbf{g})-D_{\mathrm{KL}}(Q||P)\})}{\exp(\sup_{Q\in\Delta(\mathbb{R}^d)}\{\int\sum_{i=1}^n\beta Z_i(\mathbf{g})\mathrm{d}Q(\mathbf{g})-D_{\mathrm{KL}}(Q||P)\})}\right]\\
&\le\mathbb{E}\left[\exp\left(\int\sum_{i=1}^n-\beta(Z_i(\mathbf{g})-\mu_i(\mathbf{g}))-\psi_i(\mathbf{g},-\beta)\mathrm{d}Q_{\mu,\beta}(\mathbf{g})\right)\right]\\
&\le\mathbb{E}\left[\int\exp\left(\sum_{i=1}^n-\beta(Z_i(\mathbf{g})-\mu_i(\mathbf{g}))-\psi_i(\mathbf{g},-\beta)\right)\mathrm{d}Q_{\mu,\beta}(\mathbf{g})\right]=1\,.
\end{aligned}$$

The last step follows from (9) with $\eta=-\beta$. Using (8), then the Cathy-Schwarz inequality, and then (9) and (10), we obtain

$$\begin{aligned}
&\mathbb{E}\Big[\exp\left(\frac{1}{2}\sup_{Q\in\Delta(\mathbb{R}^d)}\left\{\int\sum_{i=1}^n(\eta-\beta)(Z_i(\mathbf{g})-\mu_i(\mathbf{g}))-\psi_i(\mathbf{g},\eta)-\psi_i(\mathbf{g},-\beta)\mathrm{d}Q(\mathbf{g})-D_{\mathrm{KL}}(Q||Q_{Z,\beta})\right\}\right)\Big]\\
&=\mathbb{E}\Big[\exp\left(\frac{1}{2}\sup_{Q\in\Delta(\mathbb{R}^d)}\left\{\int\sum_{i=1}^n\eta(Z_i(\mathbf{g})-\mu_i(\mathbf{g}))-\psi_i(\mathbf{g},\eta)\mathrm{d}Q(\mathbf{g})-D_{\mathrm{KL}}(Q||Q_{\mu,\beta})\right\}\right)\\
&\qquad\times\exp\left(\frac{1}{2}\log\frac{\int\exp\left(\sum_{i=1}^n\beta\mu_i(\mathbf{g})-\psi_i(\mathbf{g},-\beta)\right)\mathrm{d}P(\mathbf{g})}{\int\exp\left(\sum_{i=1}^n\beta Z_i(\mathbf{g})\right)\mathrm{d}P(\mathbf{g})}\right)\Big]\\
&\le\mathbb{E}\Big[\exp\left(\sup_{Q\in\Delta(\mathbb{R}^d)}\left\{\int\sum_{i=1}^n\eta(Z_i(\mathbf{g})-\mu_i(\mathbf{g}))-\psi_i(\mathbf{g},\eta)\mathrm{d}Q(\mathbf{g})-D_{\mathrm{KL}}(Q||Q_{\mu,\beta})\right\}\right)\Big]^{1/2}\\
&\times\mathbb{E}\left[\frac{\int\exp\left(\sum_{i=1}^n\beta\mu_i(\mathbf{g})-\psi_i(\mathbf{g},-\beta)\right)\mathrm{d}P(\mathbf{g})}{\int\exp\left(\sum_{i=1}^n\beta Z_i(\mathbf{g})\right)\mathrm{d}P(\mathbf{g})}\right]^{1/2}\\
&\le 1\,.
\end{aligned}$$

The statement now follows from Markov's inequality. ■

Notice that if every occurence of $\psi_i(\mathbf{g}, \eta)$ (and $\psi_i(\mathbf{g}, -\beta)$) is replaced with an upper bound on $\psi_i(\mathbf{g}, \eta)$ (or $\psi_i(\mathbf{g}, -\beta)$), then the proof still goes through. This includes the occurences of $\psi_i(\mathbf{g}, -\beta)$ in the localised prior $Q_{\mu,\beta}$.

### C.3. Proof of Proposition 4

**Proof of Proposition 4** We set

$$Z_i(\mathbf{g}) = (f^\star(x_i) - y_i)^2 - (g_i - y_i)^2 = -(g_i - f^\star(x_i))^2 + 2(g_i - f^\star(x_i))\varepsilon_i \,.$$

With this choice of $Z_i(\mathbf{g})$, $\mu_i(\mathbf{g}) = -(g_i - f^\star(x_i))^2$, and for all $\eta \in \mathbb{R}$,

$$\psi_i(\mathbf{g}, \eta) = \log \mathbb{E}[\exp(2\eta(g_i - f^\star(x_i))\varepsilon_i)] \leq 2\sigma^2\eta^2(g_i - f^\star(x_i))^2 \,. \tag{11}$$

In addition, we choose $P = P_\alpha$. From the definition of $Z_i(\mathbf{g})$, $Q_{Z,\beta}$ is given by

$$\frac{\mathrm{d}Q_{Z,\beta}}{\mathrm{d}P_\alpha}(\mathbf{g}) \propto \exp(n\beta(r_n(\mathbf{f}_n^\star) - r_n(\mathbf{g}))) \propto \exp(-n\beta r_n(\mathbf{g})) \,.$$

Namely, $Q_{Z,\beta} = Q_{n,\beta,\alpha}$. Even though we will not need it, we can also determine the expression for the localised prior $Q_{\mu,\beta}$ (or rather, for $\mathrm{d}Q_{\mu,\beta}/\mathrm{d}P$). From the bound in (11), it follows that $Q_{\mu,\beta}$ is given by

$$\frac{\mathrm{d}Q_{\mu,\beta}}{\mathrm{d}P_\alpha}(\mathbf{g}) \propto \exp(-n(\beta + 2\sigma^2\beta^2)R_n(\mathbf{g})) \,.$$

One can notice that $Q_{\mu,\beta}$ assigns higher density to vectors $\mathbf{g}$ for which the excess risk is small, whereas $Q_{Z,\beta}$ assigns higher density to vectors $\mathbf{g}$ for which the empirical risk is small. Next, we substitute the definitions of/bounds on $Z_i(\mathbf{g})$, $\mu_i(\mathbf{g})$ and $\psi_i(\mathbf{g}, \eta)$ into Proposition 20. In particular, for any $n \geq 1$, any $\delta \in (0, 1]$, any $\eta > \beta \geq 0$ and any $P \in \Delta(\mathbb{R}^n)$, with probability at least $1 - \delta$, for all $Q \in \Delta(\mathbb{R}^n)$,

$$\int n(\eta-\beta)(r_n(\mathbf{f}_n^\star)-r_n(\mathbf{g}))+n(\eta-2\sigma^2\eta^2-\beta-2\sigma^2\beta^2)R_n(\mathbf{g})\mathrm{d}Q(\mathbf{g})-D_{\mathrm{KL}}(Q||Q_{n,\beta,\alpha}) \leq 2\log\tfrac{1}{\delta} \,.$$

If $\eta$ and $\beta$ are chosen such that $\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2 > 0$, then we can arrange this inequality to obtain

$$\int R_n(\mathbf{g})\mathrm{d}Q(\mathbf{g}) \leq \frac{\int(\eta - \beta)(r_n(\mathbf{g}) - r_n(\mathbf{f}_n^\star))\mathrm{d}Q(\mathbf{g})}{\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2} + \frac{D_{\mathrm{KL}}(Q||Q_{n,\beta,\alpha}) + 2\log\frac{1}{\delta}}{n(\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2)} \,. \tag{12}$$

All that remains is to show that $Q_{n,\eta,\alpha}$ minimises the right-hand side of this inequality w.r.t. $Q$. Since $r_n$ is bounded below (by 0), Lemma 13 tells us that the infimum of the right-hand side is achieved when

$$\frac{\mathrm{d}Q}{\mathrm{d}Q_{n,\beta,\alpha}}(\mathbf{g}) \propto \exp(-n(\eta - \beta)r_n(\mathbf{g})) \,.$$

For $Q = Q_{n,\eta,\alpha}$,

$$\begin{aligned}\frac{\mathrm{d}Q_{n,\eta,\alpha}}{\mathrm{d}Q_{n,\beta,\alpha}}(\mathbf{g}) &= \frac{\mathrm{d}Q_{n,\eta,\alpha}}{\mathrm{d}P_\alpha}(\mathbf{g})\frac{\mathrm{d}P_\alpha}{\mathrm{d}Q_{n,\beta,\alpha}}(\mathbf{g}) \\ &\propto \exp(-n\eta r_n(\mathbf{g}))\exp(n\beta r_n(\mathbf{g})) \\ &\propto \exp(-n(\eta - \beta)r_n(\mathbf{g})) \,.\end{aligned}$$

Therefore $Q_{n,\eta,\alpha}$ is indeed a minimiser of the right-hand side of (12). Since (12) holds simultaneously for all $Q \in \Delta(\mathbb{R}^n)$, under the same conditions as before, with probability at least $1-\delta$,

$$\int R_n(\mathbf{g})\mathrm{d}Q_{n,\eta,\alpha}(\mathbf{g}) \leq \inf_Q \left\{ \frac{\int(\eta-\beta)(r_n(\mathbf{g})-r_n(\mathbf{f}_n^\star))\mathrm{d}Q(\mathbf{g})}{\eta-2\sigma^2\eta^2-\beta-2\sigma^2\beta^2} + \frac{D_{\mathrm{KL}}(Q||Q_{n,\beta,\alpha})+2\log\frac{1}{\delta}}{n(\eta-2\sigma^2\eta^2-\beta-2\sigma^2\beta^2)} \right\}.$$

This concludes the proof. ■

### C.4. Proof of Theorem 5

**Proof of Theorem 5** We define the constrained least squares estimate

$$\widehat{f} := \operatorname*{argmin}_{f\in\mathcal{H},\|f\|_{\mathcal{H}}\leq\|f^\star\|_{\mathcal{H}}} \left\{ \sum_{i=1}^n (f(x_i)-y_i)^2 \right\}.$$

For some $\zeta \geq 0$, one can express $\widehat{f}$ via

$$\widehat{f}(x) = \mathbf{k}_n^\top(x)(\mathbf{K}_n+\zeta\mathbf{I})^{-1}\mathbf{y}_n\,.$$

In addition, the vector of fitted values $\widehat{\mathbf{f}}_n := [\widehat{f}(x_1),\ldots,\widehat{f}(x_n)]^\top$ is

$$\widehat{\mathbf{f}}_n = \mathbf{K}_n(\mathbf{K}_n+\zeta\mathbf{I})^{-1}\mathbf{y}_n\,.$$

From the definition of $\widehat{f}$, we have $r_n(\widehat{\mathbf{f}}_n) \leq r_n(\mathbf{f}_n^\star)$ and

$$\mathbf{y}_n^\top(\mathbf{K}_n+\zeta\mathbf{I})^{-1}\mathbf{K}_n(\mathbf{K}_n+\zeta\mathbf{I})^{-1}\mathbf{y}_n = \|\widehat{f}\|_{\mathcal{H}}^2 \leq \|f^\star\|_{\mathcal{H}}^2\,. \tag{13}$$

We want to find an upper bound for

$$\inf_{Q\in\Delta(\mathbb{R}^n)} \left\{ \frac{\int(\eta-\beta)(r_n(\mathbf{g})-r_n(\mathbf{f}_n^\star))\mathrm{d}Q(\mathbf{g})}{\eta-2\sigma^2\eta^2-\beta-2\sigma^2\beta^2} + \frac{D_{\mathrm{KL}}(Q||Q_{n,\beta,\alpha})+2\log\frac{1}{\delta}}{n(\eta-2\sigma^2\eta^2-\beta-2\sigma^2\beta^2)} \right\}.$$

We consider two cases. First, suppose that $\frac{1}{2\beta\alpha} \leq \zeta$. We set $\widehat{Q} = \mathcal{N}(\mathbf{m}_{n,\beta,\alpha}, \mathbf{K}_{n,\eta,\alpha})$. Since the empirical risk $r_n(\mathbf{f}_n^{(\zeta)})$ of the $\mathbf{f}_n^{(\zeta)} := \mathbf{K}_n(\mathbf{K}_n+\zeta\mathbf{I})^{-1}\mathbf{y}_n$ is increasing in $\zeta$ (cf. Lemma C.1), and since $\mathbf{m}_{n,\beta,\alpha} = \mathbf{f}_n^{(1/(2\beta\alpha))}$, it follows that

$$r_n(\mathbf{m}_{n,\beta,\alpha}) \leq r_n(\widehat{\mathbf{f}}_n) \leq r_n(\mathbf{f}_n^\star)\,.$$

Using this inequality, along with Corollary 18, we obtain

$$\begin{aligned}\int(\eta-\beta)(r_n(\mathbf{g})-r_n(\mathbf{f}_n^\star))\mathrm{d}\widehat{Q}(\mathbf{g}) &= (\eta-\beta)(r_n(\mathbf{m}_{n,\beta,\alpha})-r_n(\mathbf{f}_n^\star)) + \frac{\eta-\beta}{n}\mathrm{tr}(\mathbf{K}_{n,\eta,\alpha}) \\ &\leq \frac{1}{n}\gamma_n(2\eta\alpha, 2\beta\alpha)\,.\end{aligned}$$

Using Lemma 19, Corollary 17 and the expression for the KL divergence between two Gaussians (with the same mean), we see that

$$D_{\mathrm{KL}}(\widehat{Q}||Q_{n,\beta,\alpha}) = \frac{1}{2}\left(\mathrm{tr}(\mathbf{K}_{n,\beta,\alpha}^{-1}\mathbf{K}_{n,\beta,\alpha}) - n + \log\frac{\det(\mathbf{K}_{n,\beta,\alpha})}{\det(\mathbf{K}_{n,\eta,\alpha})}\right) \leq \gamma_n(2\eta\alpha, 2\beta\alpha)\,.$$

We have shown that, for the case where $\frac{1}{2\beta\alpha} \leq \zeta$,

$$\begin{aligned}
&\inf_{Q \in \Delta(\mathbb{R}^n)} \left\{ \frac{\int (\eta - \beta)(r_n(\mathbf{g}) - r_n(\mathbf{f}_n^\star)) \mathrm{d}Q(\mathbf{g})}{\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2} + \frac{D_{\mathrm{KL}}(Q || Q_{n,\beta,\alpha}) + 2\log\frac{1}{\delta}}{n(\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2)} \right\} \\
&\quad \leq \frac{\int (\eta - \beta)(r_n(\mathbf{g}) - r_n(\mathbf{f}_n^\star)) \mathrm{d}\widehat{Q}(\mathbf{g})}{\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2} + \frac{D_{\mathrm{KL}}(\widehat{Q} || Q_{n,\beta,\alpha}) + 2\log\frac{1}{\delta}}{n(\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2)} \\
&\quad \leq \frac{2\gamma_n(2\eta\alpha, 2\beta\alpha) + 2\log\frac{1}{\delta}}{n(\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2)} \,.
\end{aligned}$$

Next, we consider the second case. Suppose that $\frac{1}{2\alpha\beta} > \zeta$. We set $\widehat{Q} = \mathcal{N}(\widehat{\mathbf{f}}_n, \mathbf{K}_{n,\eta,\alpha})$. Since $r_n(\widehat{\mathbf{f}}_n) \leq r_n(\mathbf{f}_n^\star)$, Corollary 18 tells us that

$$\begin{aligned}
\int (\eta - \beta)(r_n(\mathbf{g}) - r_n(\mathbf{f}_n^\star)) \mathrm{d}\widehat{Q}(\mathbf{g}) &= (\eta - \beta)(r_n(\widehat{\mathbf{f}}_n) - r_n(\mathbf{f}_n^\star)) + \frac{\eta - \beta}{n} \mathrm{tr}(\mathbf{K}_{n,\eta,\alpha}) \\
&\leq \frac{1}{n} \gamma_n(2\eta\alpha, 2\beta\alpha) \,.
\end{aligned}$$

With this choice of $\widehat{Q}$, the KL divergence $D_{\mathrm{KL}}(\widehat{Q} || Q_{n,\beta,\alpha})$ is equal to

$$\frac{1}{2}\left( \mathrm{tr}(\mathbf{K}_{n,\beta,\alpha}^{-1}\mathbf{K}_{n,\beta,\alpha}) - n + (\widehat{\mathbf{f}}_n - \mathbf{m}_{n,\beta,\alpha})^\top \mathbf{K}_{n,\beta,\alpha}^{-1}(\widehat{\mathbf{f}}_n - \mathbf{m}_{n,\beta,\alpha}) + \log\frac{\det(\mathbf{K}_{n,\beta,\alpha})}{\det(\mathbf{K}_{n,\eta,\alpha})} \right).$$

As previously, the trace (minus $n$) and log-determinant terms can be bounded using Lemma 19 and Corollary 17. After some slightly unpleasant calculation, one can show that the remaining quadratic term is upper bounded by the squared RKHS norm of $f^\star$. First, we expand the quadratic term, and obtain

$$(\widehat{\mathbf{f}}_n - \mathbf{m}_{n,\beta,\alpha})^\top \mathbf{K}_{n,\beta,\alpha}^{-1}(\widehat{\mathbf{f}}_n - \mathbf{m}_{n,\beta,\alpha}) = \widehat{\mathbf{f}}_n^\top \mathbf{K}_{n,\beta,\alpha}^{-1}\widehat{\mathbf{f}}_n - 2\widehat{\mathbf{f}}_n^\top \mathbf{K}_{n,\beta,\alpha}^{-1}\mathbf{m}_{n,\beta,\alpha} + \mathbf{m}_{n,\beta,\alpha}^\top \mathbf{K}_{n,\beta,\alpha}^{-1}\mathbf{m}_{n,\beta,\alpha} \,.$$

Using Corollary 16 and (13), we see that

$$\begin{aligned}
\widehat{\mathbf{f}}_n^\top \mathbf{K}_{n,\beta,\alpha}^{-1}\widehat{\mathbf{f}}_n &= \mathbf{y}_n^\top (\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{K}_n\mathbf{K}_{n,\beta,\alpha}^{-1}\mathbf{K}_n(\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{y}_n \\
&= 2\beta\mathbf{y}_n^\top (\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{2\beta\alpha}\mathbf{I})(\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{y}_n \\
&= 2\beta\mathbf{y}_n^\top (\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{K}_n(\mathbf{K}_n + \zeta\mathbf{I} + (\tfrac{1}{2\beta\alpha} - \zeta)\mathbf{I})(\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{y}_n \\
&= 2\beta\mathbf{y}_n^\top (\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{K}_n\mathbf{y}_n + 2\beta(\tfrac{1}{2\beta\alpha} - \zeta)\mathbf{y}_n^\top (\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{K}_n(\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{y}_n \\
&\leq 2\beta\mathbf{y}_n^\top (\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{K}_n\mathbf{y}_n + \tfrac{1}{\alpha}\|f^\star\|_{\mathcal{H}}^2 \,.
\end{aligned}$$

The second term in the quadratic expansion can be re-written using Corollary 16 again. In particular,

$$\begin{aligned}
2\widehat{\mathbf{f}}_n^\top \mathbf{K}_{n,\beta,\alpha}^{-1}\mathbf{m}_{n,\beta,\alpha} &= 2\mathbf{y}_n^\top (\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{K}_n\mathbf{K}_{n,\beta,\alpha}^{-1}\mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{2\beta\alpha}\mathbf{I})^{-1}\mathbf{y}_n \\
&= 4\beta\mathbf{y}_n^\top (\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{2\beta\alpha}\mathbf{I})(\mathbf{K}_n + \tfrac{1}{2\beta\alpha}\mathbf{I})^{-1}\mathbf{y}_n \\
&= 4\beta\mathbf{y}_n^\top (\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{K}_n\mathbf{y}_n \,.
\end{aligned}$$

The third term in the quadratic expansion can be re-written in the same way. In particular,

$$\begin{aligned}\mathbf{m}_{n,\beta,\alpha}^\top \mathbf{K}_{n,\beta,\alpha}^{-1}\mathbf{m}_{n,\beta,\alpha} &= \mathbf{y}_n^\top(\mathbf{K}_n + \tfrac{1}{2\beta\alpha}\mathbf{I})^{-1}\mathbf{K}_n\mathbf{K}_{n,\beta,\alpha}^{-1}\mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{2\beta\alpha}\mathbf{I})^{-1}\mathbf{y}_n \\ &= 2\beta\mathbf{y}_n^\top(\mathbf{K}_n + \tfrac{1}{2\beta\alpha}\mathbf{I})^{-1}\mathbf{K}_n(\mathbf{K}_n + \tfrac{1}{2\beta\alpha}\mathbf{I})(\mathbf{K}_n + \tfrac{1}{2\beta\alpha}\mathbf{I})^{-1}\mathbf{y}_n \\ &= 2\beta\mathbf{y}_n^\top(\mathbf{K}_n + \tfrac{1}{2\beta\alpha}\mathbf{I})^{-1}\mathbf{K}_n\mathbf{y}_n\,.\end{aligned}$$

We have now shown that

$$\begin{aligned}(\widehat{\mathbf{f}}_n - \mathbf{m}_{n,\beta,\alpha})^\top\mathbf{K}_{n,\beta,\alpha}^{-1}(\widehat{\mathbf{f}}_n - \mathbf{m}_{n,\beta,\alpha}) \le \tfrac{1}{\alpha}\|f^\star\|_{\mathcal{H}}^2 &+ 2\beta\mathbf{y}_n^\top(\mathbf{K}_n + \tfrac{1}{2\beta\alpha}\mathbf{I})^{-1}\mathbf{K}_n\mathbf{y}_n \\ &- 2\beta\mathbf{y}_n^\top(\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{K}_n\mathbf{y}_n\,.\end{aligned}$$

Since $\frac{1}{2\beta\alpha} > \zeta$ and the eigenvalues of $(\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{K}_n$ are decreasing in $\zeta$, it follows that $(\mathbf{K}_n + \frac{1}{2\beta\alpha}\mathbf{I})^{-1}\mathbf{K}_n \preccurlyeq (\mathbf{K}_n + \zeta\mathbf{I})^{-1}\mathbf{K}_n$. Therefore,

$$(\widehat{\mathbf{f}}_n - \mathbf{m}_{n,\beta,\alpha})^\top\mathbf{K}_{n,\beta,\alpha}^{-1}(\widehat{\mathbf{f}}_n - \mathbf{m}_{n,\beta,\alpha}) \le \tfrac{1}{\alpha}\|f^\star\|_{\mathcal{H}}^2\,.$$

For the second case where $\frac{1}{2\beta\alpha} > \zeta$, we have

$$\begin{aligned}&\inf_{Q\in\Delta(\mathbb{R}^n)}\left\{\frac{\int(\eta-\beta)(r_n(\mathbf{g}) - r_n(\mathbf{f}_n^\star))\mathrm{d}Q(\mathbf{g})}{\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2} + \frac{D_{\mathrm{KL}}(Q||Q_{n,\beta,\alpha}) + 2\log\frac{1}{\delta}}{n(\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2)}\right\} \\ &\quad\le \frac{\int(\eta-\beta)(r_n(\mathbf{g}) - r_n(\mathbf{f}_n^\star))\mathrm{d}\widehat{Q}(\mathbf{g})}{\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2} + \frac{D_{\mathrm{KL}}(\widehat{Q}||Q_{n,\beta,\alpha}) + 2\log\frac{1}{\delta}}{n(\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2)} \\ &\quad\le \frac{2\gamma_n(2\eta\alpha, 2\beta\alpha) + \frac{1}{2\alpha}\|f^\star\|_{\mathcal{H}}^2 + 2\log\frac{1}{\delta}}{n(\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2)}\,.\end{aligned}$$

This concludes the proof. ■

## Appendix D. Bounds on the Relative Information Gain

### D.1. Mercer Kernels

A positive-definite kernel $k$ on a set $\mathcal{X}$ is called a Mercer kernel if $\mathcal{X}$ is a compact metric space and the kernel function $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is continuous. Mercer's theorem provides a useful representation for Mercer kernels. Let $\rho$ be a non-degenerate Borel measure on $\mathcal{X}$ and let $L^2(\mathcal{X},\rho)$ denote the set of square integrable functions on $\mathcal{X}$. Namely,

$$L^2(\mathcal{X},\rho) := \left\{f : \mathcal{X}\to\mathbb{R} : \int_{\mathcal{X}}(f(x))^2\mathrm{d}\rho(x) < \infty\right\}\,.$$

Define the linear operator $L_k : L^2(\mathcal{X},\rho) \to L^2(\mathcal{X},\rho)$ as

$$L_k(f)(x) := \int_{\mathcal{X}} k(x,y)f(y)\mathrm{d}\rho(y)\,.$$

**Theorem 21 (Mercer's Theorem)** *If $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a Mercer kernel, then there exist non-negative eigenvalues $\xi_1 \geq \xi_2 \geq \cdots \geq 0$ and corresponding eigenfunctions $\phi_1, \phi_2, \ldots,$ such that*

$$L_k(\phi_m) = \xi_m \phi_m, \quad \textit{for all } m = 1, 2, \ldots . \tag{14}$$

*In addition, the kernel function has the eigendecomposition*

$$k(x, x') = \sum_{m=1}^{\infty} \xi_m \phi_m(x) \phi_m(x') , \tag{15}$$

*where the convergence of the infinite series is absolute for each $x, x' \in \mathcal{X}$ and uniform on $\mathcal{X} \times \mathcal{X}$.*

### D.2. Auxiliary Lemmas

Lemma 22 and Lemma 23 were extracted from the proof of Theorem 3 in Vakili et al. (2021). Lemma 22 allows us to re-write the information gain as the sum of a term depending on only $\mathbf{K}_n^{\|}$ and another term depending on both $\mathbf{K}_n^{\|}$ and $\mathbf{K}_n^{\perp}$.

**Lemma 22** *For any $\eta > 0$ and any kernel $k$ that satisfies Assumption 3,*

$$\frac{1}{2} \log \det(\eta \mathbf{K}_n + \mathbf{I}) = \frac{1}{2} \log \det(\eta \mathbf{K}_n^{\|} + \mathbf{I}) + \frac{1}{2} \log \det(\mathbf{I} + \eta(\mathbf{I} + \eta \mathbf{K}_n^{\|})^{-1} \mathbf{K}_n^{\perp}) .$$

Lemma 23 provides an upper bound on the second term in Lemma 22.

**Lemma 23** *For any $\eta > 0$ and any kernel $k$ that satisfies Assumption 3,*

$$\frac{1}{2} \log \det(\mathbf{I} + \eta(\mathbf{I} + \eta \mathbf{K}_n^{\|})^{-1} \mathbf{K}_n^{\perp}) \leq \frac{1}{2} n \eta \delta_D .$$

To bound the first term on the right-hand side of the inequality in Lemma 22, we will consider the log-determinant of a $D \times D$ gram matrix. We define a $D$-dimensional feature map $\boldsymbol{\phi}_D(x) := [\phi_1(x), \ldots, \phi_D(x)]^\top$, an $n \times D$ design matrix $\boldsymbol{\Phi}_{n,D} := [\boldsymbol{\phi}_D(x_1), \ldots, \boldsymbol{\phi}_D(x_n)]^\top$ and a diagonal $D \times D$ matrix $\boldsymbol{\Xi}_D := \operatorname{diag}(\xi_1, \ldots, \xi_D)$. Notice that $\mathbf{K}_n^{\|} = \boldsymbol{\Phi}_{n,D} \boldsymbol{\Xi}_D \boldsymbol{\Phi}_{n,D}^\top$. We define the gram matrix

$$\mathbf{G}_n := \boldsymbol{\Xi}_D^{1/2} \boldsymbol{\Phi}_{n,D}^\top \boldsymbol{\Phi}_{n,D} \boldsymbol{\Xi}_D^{1/2} .$$

Since we will deal with the relative information gain, we will need to bound the log of the ratio of the determinants of two of these gram matrices. This is handled by Lemma

**Lemma 24** *For any $n \geq 1$, $D \geq 1$ and $\eta > \beta > 0$,*

$$\log \frac{\det(\eta \mathbf{G}_n + \mathbf{I})}{\det(\beta \mathbf{G}_n + \mathbf{I})} \leq D \log \frac{\eta}{\beta} .$$

**Proof** Let $(\lambda_i)_{i=1}^D$ be the eigenvalues of $\mathbf{G}_n$, which are all real and positive. We notice that for any $\zeta \in (0, 1]$,

$$\det(\beta \mathbf{G}_n + \zeta \mathbf{I}) = \prod_{i=1}^D (\beta \lambda_i + \zeta) \leq \prod_{i=1}^D (\beta \lambda_i + 1) = \det(\beta \mathbf{G}_n + \mathbf{I}) \,.$$

Therefore, we can bound the logarithm of the ratio of determinants as

$$\begin{aligned} \log \frac{\det(\eta \mathbf{G}_n + \mathbf{I})}{\det(\beta \mathbf{G}_n + \mathbf{I})} &= D \log \tfrac{\eta}{\beta} + \log \frac{\det(\beta \mathbf{G}_n + \frac{\beta}{\eta} \mathbf{I})}{\det(\beta \mathbf{G}_n + \mathbf{I})} \\ &\leq D \log \tfrac{\eta}{\beta} + \log \frac{\det(\beta \mathbf{G}_n + \mathbf{I})}{\det(\beta \mathbf{G}_n + \mathbf{I})} \\ &= D \log \tfrac{\eta}{\beta} \,. \end{aligned}$$

This concludes the proof. ■

We will use the fact that the second term on the right-hand side of the inequality in Lemma 22 is non-negative.

**Lemma 25** *For any $\eta \geq 0$ and any kernel $k$ that satisfies Assumption 3,*

$$\frac{1}{2} \log \det(\mathbf{I} + \eta(\mathbf{I} + \eta \mathbf{K}_n^{\|})^{-1} \mathbf{K}_n^{\perp}) \geq 0 \,.$$

**Proof** Since $\mathbf{K}_n = \mathbf{K}_n^{\|} + \mathbf{K}_n^{\perp}$, and both $\mathbf{K}_n^{\|}$ and $\mathbf{K}_n^{\perp}$ are positive semi-definite, it follows that

$$\eta \mathbf{K}_n + \mathbf{I} \succcurlyeq \eta \mathbf{K}_n^{\|} + \mathbf{I} \,.$$

Due to monotonicity of the determinant with respect to the Loewner ordering (cf. Corollary 7.7.4 in Horn and Johnson, 2012),

$$\log \det(\eta \mathbf{K}_n + \mathbf{I}) \geq \log \det(\eta \mathbf{K}_n^{\|} + \mathbf{I}) \,.$$

Finally, using Lemma 22,

$$\frac{1}{2} \log \det(\mathbf{I} + \eta(\mathbf{I} + \eta \mathbf{K}_n^{\|})^{-1} \mathbf{K}_n^{\perp}) = \frac{1}{2} \log \det(\eta \mathbf{K}_n + \mathbf{I}) - \frac{1}{2} \log \det(\eta \mathbf{K}_n^{\|} + \mathbf{I}) \geq 0 \,.$$

This concludes the proof. ■

The next lemma comes from the proof of Corollary 1 in Vakili et al. (2021). It provides bounds on $\delta_D$ under each of the eigenvalue decay assumptions.

**Lemma 26** *If the polynomial eigenvalue decay condition from Assumption 4 is satisfied, then*

$$\delta_D \leq C_p D^{1-\beta_p} \psi^2 \,.$$

*If the exponential eigenvalue decay condition from Assumption 4 is satisfied and $\beta_e = 1$, then*

$$\delta_D \leq \frac{C_{e,1} \psi^2}{C_{e,2}} \exp(-C_{e,2} D) \,.$$

*If the exponential eigenvalue decay condition from Assumption 4 is satisfied and $\beta_e \in (0, 1)$, then*

$$\delta_D \leq \frac{2 C_{1,e} \psi^2}{C_{e,2} \beta_e} \left( \frac{2}{C_{e,2}} \left( \frac{1}{\beta_e} - 1 \right) \right)^{1/\beta_e - 1} \exp\left(1 - \frac{1}{\beta_e}\right) \exp\left( - C_{e,2} \frac{D^{\beta_e}}{2} \right) .$$

### D.3. Proof of Proposition 6

**Proof of Proposition 6** Using Lemma 22 and Lemma 25, we can re-write and then upper bound the relative information gain as

$$\begin{aligned}\gamma_n(\eta,\beta) &= \frac{1}{2}\log\frac{\det(\eta\mathbf{K}_n^{\|}+\mathbf{I})}{\det(\beta\mathbf{K}_n^{\|}+\mathbf{I})} + \frac{1}{2}\log\det(\mathbf{I}+\eta(\mathbf{I}+\eta\mathbf{K}_n^{\|})^{-1}\mathbf{K}_n^{\perp}) \\ &\quad - \frac{1}{2}\log\det(\mathbf{I}+\beta(\mathbf{I}+\beta\mathbf{K}_n^{\|})^{-1}\mathbf{K}_n^{\perp}) \\ &\leq \frac{1}{2}\log\frac{\det(\eta\mathbf{K}_n^{\|}+\mathbf{I})}{\det(\beta\mathbf{K}_n^{\|}+\mathbf{I})} + \frac{1}{2}\log\det(\mathbf{I}+\eta(\mathbf{I}+\eta\mathbf{K}_n^{\|})^{-1}\mathbf{K}_n^{\perp})\,.\end{aligned}$$

By the Weinstein–Aronszajn identity,

$$\frac{1}{2}\log\frac{\det(\eta\mathbf{K}_n^{\|}+\mathbf{I})}{\det(\beta\mathbf{K}_n^{\|}+\mathbf{I})} = \frac{1}{2}\log\frac{\det(\eta\mathbf{G}_n+\mathbf{I})}{\det(\beta\mathbf{G}_n+\mathbf{I})}\,.$$

Therefore, from Lemma 24 and Lemma 23, it follows that

$$\gamma_n(\eta,\beta) \leq \frac{1}{2}D\log\frac{\eta}{\beta} + \frac{1}{2}n\eta\delta_D\,.$$

This concludes the proof. ■

### D.4. Proof of Proposition 7

**Proof of Proposition 7** From Proposition 6,

$$\gamma_n(\eta,\beta) \leq \frac{1}{2}D\log\frac{\eta}{\beta} + \frac{1}{2}n\eta\delta_D\,.$$

We consider each eigenvalue decay condition separately. If the polynomial eigenvalue decay condition is satisfied, then Lemma 26 tells us that

$$\frac{1}{2}n\eta\delta_D \leq \frac{1}{2}n\eta C_p D^{1-\beta_p}\psi^2\,.$$

We choose the smallest value of $D$ such that

$$n\eta C_p D^{1-\beta_p}\psi^2 \leq D\log\frac{\eta}{\beta}\,.$$

One can verify that this inequality is satisfied if we choose

$$D = \left\lceil (n\eta C_p\psi^2)^{1/\beta_p}\log^{-1/\beta_p}(\tfrac{\eta}{\beta})\right\rceil\,.$$

With this choice of $D$, the relative information gain satisfies

$$\gamma_n(\eta,\beta) \leq D\log\tfrac{\eta}{\beta} \leq (n\eta C_p\psi^2)^{1/\beta_p}\log^{1-1/\beta_p}(\tfrac{\eta}{\beta}) + \log\tfrac{\eta}{\beta}\,.$$

If the exponential eigenvalue decay condition is satisfied and $\beta_e = 1$, then Lemma 26 tells us that

$$\frac{1}{2} n\eta\delta_D \leq \frac{1}{2} n\eta \frac{C_{e,1}\psi^2}{C_{e,2}} \exp(-C_{e,2}D)\,.$$

This time, we choose

$$D = \left\lceil \frac{1}{C_{e,2}} \log \frac{n\eta C_{e,1}\psi^2}{C_{e,2}} \right\rceil .$$

With this choice of $D$, the relative information gain satisfies

$$\gamma_n(\eta,\beta) \leq \frac{1}{2}\frac{1}{C_{e,2}} \log \frac{n\eta C_{e,1}\psi^2}{C_{e,2}} \log \frac{\eta}{\beta} + \frac{1}{2} \log \frac{e\eta}{\beta}\,.$$

If the exponential eigenvalue decay condition is satisfied and $\beta_e = 1$, then Lemma 26 tells us that

$$\frac{1}{2} n\eta\delta_D \leq \frac{1}{2} n\eta \frac{2C_{1,e}\psi^2}{C_{e,2}\beta_e} \left( \frac{2}{C_{e,2}} \left( \frac{1}{\beta_e} - 1 \right) \right)^{1/\beta_e - 1} \exp\left(1 - \frac{1}{\beta_e}\right) \exp\left( - C_{e,2} \frac{D^{\beta_e}}{2} \right).$$

If we choose

$$D = \left\lceil \left( \frac{2}{C_{e,2}} \log \left( \frac{2n\eta C_{1,e}\psi^2}{C_{e,2}\beta_e} \left( \frac{2}{C_{e,2}} \left( \frac{1}{\beta_e} - 1 \right) \right)^{1/\beta_e - 1} \exp\left(1 - \frac{1}{\beta_e}\right) \right) \right)^{1/\beta_e} \right\rceil ,$$

Then the relative information gain satisfies

$$\gamma_n(\eta,\beta) \leq \frac{1}{2} \left( \frac{2}{C_{e,2}} \log \left( \frac{2n\eta C_{1,e}\psi^2}{C_{e,2}\beta_e} \left( \frac{2}{C_{e,2}} \left( \frac{1}{\beta_e} - 1 \right) \right)^{1/\beta_e - 1} \exp\left(1 - \frac{1}{\beta_e}\right) \right) \right)^{1/\beta_e} \log \frac{\eta}{\beta} + \frac{1}{2} \log \frac{e\eta}{\beta}\,.$$

Therefore, if the exponential decay condition is satisfied with $\beta_e \in (0,1]$, the relative information gain satisfies

$$\gamma_n(\eta,\beta) \leq \frac{1}{2} \left( \frac{2}{C_{e,2}} \log(n\eta C_{\beta_e}) \right)^{1/\beta_e} \log \frac{\eta}{\beta} + \frac{1}{2} \log \frac{e\eta}{\beta}\,,$$

where $C_{\beta_e} = \frac{C_{e,1}\psi^2}{C_{e,2}}$ if $\beta_e = 1$, and $C_{\beta_e} = \frac{2C_{e,1}\psi^2}{C_{e,2}\beta_e} (\frac{2-2\beta_e}{C_{e,2}\beta_e})^{1/\beta_e - 1} \exp(\frac{1-\beta_e}{\beta_e})$ if $\beta_e \in (0,1)$. ■

## Appendix E. Rates of Convergence

### E.1. Proof of Corollary 8

**Proof of Corollary 8** When the polynomial eigenvalue decay condition is satisfied, the bound on the relative information gain in Proposition 7 reads as

$$\gamma_n(2\eta\alpha, 2\beta\alpha) \leq (2n\eta\alpha C_p \psi^2)^{1/\beta_p} \log^{1-1/\beta_p}(\tfrac{\eta}{\beta}) + \log \tfrac{\eta}{\beta}\,.$$

From this and Theorem 5, we know that (with probability at least $1-\delta$), the excess risk of $Q_{n,\eta,\alpha}$ satisfies

$$\int R_n(\mathbf{g})\mathrm{d}Q_{n,\eta,\alpha}(\mathbf{g}) \leq \frac{2(2n\eta\alpha C_p\psi^2)^{1/\beta_p}\log^{1-1/\beta_p}(\frac{\eta}{\beta}) + 2\log\frac{\eta}{\beta} + \frac{1}{2\alpha}\|f^\star\|_{\mathcal{H}}^2 + 2\log\frac{1}{\delta}}{n(\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2)}\,.$$

If we set $\eta = \frac{1}{4\sigma^2}$ and $\beta = \frac{1}{32\sigma^2}$, then this inequality becomes

$$\int R_n(\mathbf{g})\mathrm{d}Q_{n,\eta,\alpha}(\mathbf{g}) \leq \frac{512\sigma^2}{47n}\left(2\left(\frac{n\alpha C_p\psi^2}{2\sigma^2}\right)^{1/\beta_p}\log^{1-1/\beta_p}(8) + \frac{1}{2\alpha}\|f^\star\|_{\mathcal{H}}^2 + 2\log\frac{8}{\delta}\right).$$

If we then set $\alpha = n^{-\frac{1}{1+\beta_p}}$, we obtain

$$\int R_n(\mathbf{g})\mathrm{d}Q_{n,\eta,\alpha}(\mathbf{g}) \leq \frac{512\sigma^2}{47}\left(2\left(\frac{C_p\psi^2}{2\sigma^2}\right)^{1/\beta_p}\log^{1-1/\beta_p}(8) + \frac{1}{2}\|f^\star\|_{\mathcal{H}}^2\right)n^{-\frac{\beta_p}{1+\beta_p}} + \frac{1024\sigma^2\log\frac{8}{\delta}}{47n}\,,$$

and the proof is complete. ■

One can instead choose a value of $\alpha$ that results in the best dependence on $\sigma$ and $\|f^\star\|_{\mathcal{H}}$ (assuming $\|f^\star\|_{\mathcal{H}}$ is known). If

$$\alpha = \left(\tfrac{\beta_p}{2}\right)^{\frac{\beta_p}{1+\beta_p}}\left(\frac{C_p\psi^2}{2}\right)^{-\frac{1}{1+\beta_p}}\log^{\frac{1-\beta_p}{1+\beta_p}}(8)\|f^\star\|_{\mathcal{H}}^{\frac{2\beta_p}{1+\beta_p}}\sigma^{\frac{2}{1+\beta_p}}n^{-\frac{1}{1+\beta_p}}\,,$$

then the excess risk bound becomes

$$\int R_n(\mathbf{g})\mathrm{d}Q_{n,\eta.\alpha}(\mathbf{g}) \leq \tfrac{512}{47}\left(\tfrac{1+\beta_p}{\beta_p}\left(\tfrac{\beta_p C_p\psi^2}{4}\right)^{\frac{1}{1+\beta_p}}\log^{\frac{\beta_p-1}{1+\beta_p}}(8)\|f^\star\|_{\mathcal{H}}^{\frac{2}{1+\beta_p}}\sigma^{\frac{2\beta_p}{1+\beta_p}}n^{-\frac{\beta_p}{1+\beta_p}} + \frac{2\sigma^2\log\frac{8}{\delta}}{n}\right).$$

### E.2. Proof of Corollary 9

**Proof of Corollary 9** When the exponential eigenvalue decay condition is satisfied, the bound on the relative information gain in Proposition 7 reads as

$$\gamma_n(2\eta\alpha, 2\beta\alpha) \leq \frac{1}{2}\left(\frac{2}{C_{e,2}}\log(2n\eta\alpha C_{\beta_e})\right)^{1/\beta_e}\log\frac{\eta}{\beta} + \frac{1}{2}\log\frac{e\eta}{\beta}\,,$$

From this and Theorem 5, we know that (with probability at least $1-\delta$), the excess risk of $Q_{n,\eta,\alpha}$ satisfies

$$\int R_n(\mathbf{g})\mathrm{d}Q_{n,\eta,\alpha}(\mathbf{g}) \leq \frac{\left(\frac{2}{C_{e,2}}\log(2n\eta\alpha C_{\beta_e})\right)^{1/\beta_e}\log\frac{\eta}{\beta} + \log\frac{e\eta}{\beta} + \frac{1}{2\alpha}\|f^\star\|_{\mathcal{H}}^2 + 2\log\frac{1}{\delta}}{n(\eta - 2\sigma^2\eta^2 - \beta - 2\sigma^2\beta^2)}\,.$$

If we set $\eta = \frac{1}{4\sigma^2}$ and $\beta = \frac{1}{32\sigma^2}$ and $\alpha = 1$, then this inequality becomes

$$\int R_n(\mathbf{g})\mathrm{d}Q_{n,\eta,\alpha}(\mathbf{g}) \leq \frac{512\sigma^2}{47n}\left(\left(\frac{2}{C_{e,2}}\log\left(\frac{nC_{\beta_e}}{2\sigma^2}\right)\right)^{1/\beta_e}\log(8) + \frac{1}{2}\|f^\star\|_{\mathcal{H}}^2 + 2\log\frac{\sqrt{8e}}{\delta}\right),$$

and the proof is complete. ■

If $\|f^\star\|_{\mathcal{H}}$ is known, one can instead set $\alpha = \|f^\star\|_{\mathcal{H}}^2$ to obtain an excess risk bound with polylogarithmic dependence on the norm of $f^\star$.