Learning from Noisy Data under Distributional Assumptions

Nicolò Cesa-Bianchi
Università degli Studi di Milano
cesa-bianchi@dsi.unimi.it

Shai Shalev Shwartz
The Hebrew University
shais@cs.huji.ac.il

Ohad Shamir
Microsoft Research
ohadsh@microsoft.com

Abstract

We study the framework of online learning, when individual examples are corrupted by random noise, and both examples and noise type can be chosen adversarially. Previous work has shown that without knowledge of the noise distribution, it is possible to learn using a random, potentially unbounded number of independent noisy copies of each example. Moreover, it is generally impossible to learn with just one noisy copy per example. In this paper, we explore the consequences of being given some side information on the noise distribution. We consider several settings, and show how one can learn linear and kernel-based predictors using just one or two noisy views of each example, depending on the side information provided.

1 Introduction

In a wide variety of machine learning applications, ranging from remote sensing through bioinformatics to medical tests, one has to deal with data known to be noisy and distorted. This is particularly common in physically measured data, due to factors such as sensor quality, communication constraints, or various other physical limitations. In other cases, factors such as privacy concerns may force us deal with intentionally distorted data. In all such cases, the learner trains on a distorted version of the actual "target" data, which is where the learner's predictive ability is actually evaluated.

In [4], a general technique was introduced to deal with learning linear or kernel-based predictors from noisy data, where virtually nothing is known about the noise, except possibly an upper bound on its variance. Moreover, this was achieved in the challenging online framework, where an all-powerful adversary is able to choose the noise distribution, and change it on each and every round. This technique works assuming the learner is allowed to query for independent noisy copies of each example more than once. Moreover, the number of queries needed is a random quantity. Although the distribution of this random number does not depend on the scale of the problem, and the number is bounded with high probability, this is obviously a non-trivial requirement. Unfortunately, multiple queries cannot be avoided in general: in [4], it was shown that without seeing more than one noisy copy of each instance, learning becomes impossible, even in a stochastic, non-adversarial setting. However, this was under the crucial assumption that the learner has no prior knowledge about the noise distribution it needs to cope with.

While this is sometimes the case, in many applications the learner may actually know something about the noise distribution, either a-priori or based on previous measurements. In such cases, the impossibility result of [4] no longer holds. This leads to the central question dealt with in this paper: can we learn with a smaller, fixed number of queries per example, if we know something about the noise distribution?

Focusing on the squared loss, we discuss three different settings, reflecting different levels of knowledge about the noise distribution:

1. **Known variance bound** (this is the setting discussed in [4]). We show that one can learn linear predictors with two independent noisy copies of each instance $\mathbf{x}_t$, and one noisy copy of each target value y_t.
2. **Known covariance structure.** We show that one can learn linear predictors with only one noisy copy of $\mathbf{x}_t$ and y_t. This implies that the impossibility result [4] indeed does not hold in general, once we know the covariance structure of the noise.
3. **Gaussian distribution with known covariance matrix.** We show that one can even learn kernel-based predictors, using two independent noisy copies of each $\mathbf{x}_t$, and one copy of y_t. We focus on Gaussian kernels, but also show how the result can be extended, in a certain sense, to general radial kernels.

Thus, the positive learning results get stronger the more we can assume about the noise distribution. We emphasize that although all algorithms are based on a similar online gradient descent approach, our results include techniques which are very different than those of [4], and are not just an extension —see Section 3 for more details.

2 Framework and Notation

We consider a setting where the goal is to predict values $y \in \mathbb{R}$ based on instances $\mathbf{x} \in \mathbb{R}^d$. We focus on predictors which are either linear (i.e. of the form $\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$), or kernel-based —i.e., of the form $\mathbf{x} \mapsto \langle \mathbf{w}, \Psi(\mathbf{x}) \rangle$ where Ψ is a feature mapping into some reproducing kernel Hilbert space, with associated kernel $k(\cdot, \cdot)$.

We begin by recalling the standard online learning setting, which for linear predictors and squared loss is defined as follows (with the obvious changes for kernel-based predictors): at each round t, the learner picks a predictor $\mathbf{w}_t$. The adversary then picks an example $(\mathbf{x}_t, y_t)$, the learner suffers a loss $\ell(\langle \mathbf{w}_t, \mathbf{x}_t \rangle, y_t) = (\langle \mathbf{w}_t, \mathbf{x}_t \rangle - y_t)^2$, and the example $(\mathbf{x}_t, y_t)$ is revealed to the learner. The goal of the learner is to minimize his regret with respect to a fixed convex set of hypotheses $\mathcal{W}$, namely

$$\sum_{t=1}^{T} \ell(\langle \mathbf{w}_t, \mathbf{x}_t \rangle, y_t) - \min_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^{T} \ell(\langle \mathbf{w}, \mathbf{x}_t \rangle, y_t) \ .$$

Typically, one wishes to find a strategy for the learner, such that no matter what is the adversary's strategy for choosing examples, the expression above is sublinear in T (implying that the average regret per round vanishes with T).

Following [4], we now make the following twist, which limits the information available to the learner: In each round, the adversary also picks a vector-valued random variable $\mathbf{n}_t^x$ and random variable n_t^y. Instead of receiving $(\mathbf{x}_t, y_t)$, the learner is given access to one or two independent copies of $\tilde{\mathbf{x}}_t$ and $\tilde{y}_t$, where $\tilde{\mathbf{x}}_t = \mathbf{x}_t + \mathbf{n}_t^x$, and $\tilde{y}_t = y_t + n_t^y$. In other words, the adversary forces the learner to see only a noisy version of the data, where the noise distribution can change from round to round in a possibly adversarial manner. We will assume throughout the paper that $\mathbf{n}_t^x$ and n_t^y are zero-mean and independent. If $\mathbb{E}[\mathbf{n}_t^x]$ and $\mathbb{E}[n_t^y]$ are not zero, and are known to the learner, we can always reduce to the zero-mean case by deducting $\mathbb{E}[\mathbf{n}_t^x]$ and $\mathbb{E}[n_t^y]$ from $\tilde{\mathbf{x}}_t$ and $\tilde{y}_t$. The independence can be relaxed to uncorrelation or even disposed of entirely in some of the discussed settings, at the cost of some added technical complexity in the algorithms and proofs.

In [4], it was assumed that nothing is known about $\mathbf{n}_t^x$ and n_t^y, other than an upper bound on the variance. In this paper, we consider settings where some side information on the distribution of $\mathbf{n}_t^x, n_t^y$ is provided to the learner.

In this framework, our goal is to minimize the expected regret in hindsight with respect to the *unperturbed* data, namely

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\langle \mathbf{w}_t, \mathbf{x}_t \rangle, y_t) - \min_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^{T} \ell(\langle \mathbf{w}, \mathbf{x}_t \rangle, y_t)\right] \tag{1}$$

where the expectation is with respect to the noise introduced into the data. Using online-to-batch conversion techniques (e.g., [1]), one can easily convert an online algorithm with a sublinear regret in Eq. (1), to a learning algorithm in the batch statistical setting, in which the training set is corrupted by noise. Namely, the learner receives corrupted versions of examples $\{\mathbf{x}_t, y_t\}_{t=1}^T$ sampled i.i.d. from some unknown distribution, and we wish to find a predictor minimizing the expected loss $\mathbb{E}_{\mathbf{x},y}[\ell(\langle \mathbf{w}, \mathbf{x}\rangle, y)]$ over $\mathcal{W}$, in spite of the noise in the training data. Also, we note that under suitable conditions, using Azuma's inequality, one can obtain high-probability bounds on Eq. (1) without the expectation, and with logarithmic dependence on the probability. However, this complicates the presentation of the theorems and their proofs, without leading to any new insights, so we will ignore this potential improvement.

For simplicity, we will focus on the finite-horizon setting, where the total number of rounds T is assumed to be known beforehand. However, all our results can be easily extended to the infinite-horizon setting, where the learner needs to achieve sub-linear regret for all T simultaneously.

3 Overview of Techniques

We begin by presenting a high-level and informal overview of the techniques we use to get our results. Formal statements and proofs appear later in the paper and in the supplementary.

A fundamental method we use in all our settings is the online gradient descent algorithm due to Zinkevich [10]. At its heart is the following observation: for any set of vectors $\nabla_1, \ldots, \nabla_T$, suppose we define $\mathbf{w}_1 = 0$ and $\mathbf{w}_{t+1} = P(\mathbf{w}_t - \eta_t \nabla_t)$, where $P(\cdot)$ is a projection operator on a convex set $\mathcal{W}$, and η_t is a suitably chosen scalar value. Then for any $\mathbf{u} \in \mathcal{W}$, it holds that

$$\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{u}, \nabla_t \rangle = O(\sqrt{T}) \tag{2}$$

where the $O(\cdot)$ notation hides dependencies on the norm of $\mathbf{u}$ and the norms of ∇_t. In particular, suppose that we let ∇_t be the gradient of $\ell(\mathbf{w}_t, \mathbf{x}_t, y_t)$ w.r.t. $\mathbf{w}_t$. Then by convexity, the left-hand side of Eq. (2) is lower bounded by $\sum_{t=1}^T \ell(\langle \mathbf{w}_t, \mathbf{x}_t\rangle, y_t) - \sum_{t=1}^T \ell(\langle \mathbf{u}, \mathbf{x}_t\rangle, y_t)$. Thus, if we are provided with $(\mathbf{x}_t, y_t)$ after each round, we can compute ∇_t, perform the update as above, and get an algorithm with sublinear regret with respect to any predictor $\mathbf{u}$ of bounded norm.

3.1 "Stochastic" Online Gradient Descent

In our setting of noisy data, the algorithm described above is irrelevant, because $(\mathbf{x}_t, y_t)$ is unknown and we cannot compute ∇_t. However, one can deal with it as follows. Suppose that instead of ∇_t, we pick *random* vectors $\tilde{\nabla}_t$. It turns out that based on Eq. (2), one can still show that

$$\mathbb{E}\left[\sum_{t=1}^{T} \left\langle \mathbf{w}_t - \mathbf{u}, \tilde{\nabla}_t \right\rangle\right] = O(\sqrt{T}) \tag{3}$$

where the $O(\cdot)$ notation hides dependencies on the norm of $\mathbf{u}$ and $\mathbb{E}[\|\tilde{\nabla}_t\|^2]$ for all t. In particular, in our noisy data setting, we cannot compute ∇_t, but suppose we can use the noisy data that we do have, in order to construct a random vector $\tilde{\nabla}_t$, such that $\mathbb{E}[\tilde{\nabla}_t] = \nabla_t$. In that case, the left-hand side can be shown to equal $\mathbb{E}\big[\sum_{t=1}^T \langle \mathbf{w}_t - \mathbf{u}, \nabla_t\rangle\big]$. The expectation here is again w.r.t. the noisy examples (note that $\mathbf{w}_t$ is a random vector that depends on the noisy examples). Applying the same convexity trick as before, we get an $O(\sqrt{T})$ upper bound on the expected regret $\mathbb{E}\big[\sum_{t=1}^T \ell(\langle \mathbf{w}_t, \mathbf{x}_t\rangle, y_t) - \sum_{t=1}^T \ell(\langle \mathbf{u}, \mathbf{x}_t\rangle, y_t)\big]$. Thus, by doing updates using $\tilde{\nabla}_t$, we get an algorithm with a sublinear regret bound. The only technical issue is how to construct unbiased estimates $\tilde{\nabla}_t$ using the noisy data.

This is the basic learning mechanism in the first two settings we consider, in Sections 4 and 5. This technique already appears in [4] (as well as previous work in other settings, e.g.,

[2]), and our main contribution for these two settings is the observation that it can be done with one or two noisy copies of each example, under appropriate distributional assumptions.

The third setting we consider, for kernel-based predictors (Section 6), is where the main technical novelty of this paper lies, as it requires a rather different approach than that of [4]. This approach is discussed below.

3.2 "Parallel Worlds" Online Gradient Descent

With kernels, one assumes that the predictors lie in some reproducing kernel Hilbert space (RKHS), in which inner products can be efficiently computed using a kernel function. Instances are mapped into that space using a feature mapping Ψ. In [4], learning was achieved by constructing an unbiased estimate $\tilde{\nabla}_t$ of ∇_t in the RKHS, but that required a random number of noisy copies of each example. Here, we take the following very different approach: we construct a *surrogate* RKHS, with a *surrogate* feature mapping $\hat{\Psi}$, such that for any noisy copy $\tilde{\mathbf{x}}_t$ of $\mathbf{x}_t$, and any fixed instance $\mathbf{a}$, it holds that

$$\mathbb{E}\left[\langle \hat{\Psi}(\mathbf{a}), \hat{\Psi}(\tilde{\mathbf{x}}_t)\rangle\right] = \langle \Psi(\mathbf{a}), \Psi(\mathbf{x}_t)\rangle \tag{4}$$

where the expectation is with respect to the noise. Thus, "noisy" inner products in the surrogate RKHS correspond (in expectation) to "clean" inner products in the original RKHS. This allows us to use the noisy data in order to construct vectors $\hat{\nabla}_t$ *in the surrogate RKHS* with the following interesting property: if we apply Zinkevich's algorithm on $\hat{\nabla}_1, \ldots, \hat{\nabla}_T$ (using kernels), to get predictors $\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_T$ in the RKHS of $\hat{\Psi}$, then for any $\hat{\mathbf{u}}$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \left\langle \hat{\mathbf{w}}_t - \hat{\mathbf{u}}, \hat{\nabla}_t \right\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{u}, \nabla_t \rangle\right]$$

where $\mathbf{w}_t$ and $\mathbf{u}$ are certain mappings of each $\hat{\mathbf{w}}_t$ and $\hat{\mathbf{u}}$ to the RKHS of Ψ, and ∇_t are the gradients w.r.t. the unperturbed examples $(\mathbf{x}_t, y_t)$. But the left-hand side is $O(\sqrt{T})$ by Eq. (3), since we applied Zinkevich's algorithm in the surrogate RKHS. Thus, we get that $\mathbb{E}\left[\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{u}, \nabla_t \rangle\right]$ is $O(\sqrt{T})$, which implies a sublinear regret bound for $\mathbf{w}_1, \ldots, \mathbf{w}_T$. We emphasize that unlike the approach used for the other settings, $\mathbb{E}[\hat{\nabla}_t]$ *is not equal* to ∇_t. Indeed, they live in different mathematical spaces! Thus, this is a different approach rather than an extension of the technique discussed in Subsec. 3.1.

A technical issue which needs addressing is that the norm of $\hat{\mathbf{u}}$ has to be related to the norm of the actual predictor $\mathbf{u}$ we compare ourselves with. While this cannot be always done, such a relation does hold if $\mathbf{u}$ is reasonably "nice", in a sense which will be formalized later.

Constructing a surrogate RKHS as in Eq. (4) can be done when the original RKHS corresponds to a Gaussian kernel. Nevertheless, we can extend our results, in a certain sense, to more general radial kernels. The basic tool we use is Schoenberg's theorem, which implies that any radial kernel can be written as an integral of Gaussian kernels of differing width. Using this result, we can show that one can still construct a surrogate RKHS, which has the property of Eq. (4) w.r.t. an approximate version of our original radial kernel.

4 Setting 1: Upper bound on the Variance

We begin with the simplest setting, which is when we only know that $\mathbb{E}[\|\tilde{\mathbf{x}}_t\|^2] \leq B_{\tilde{\mathbf{x}}}^2$ and $\mathbb{E}[\tilde{y}_t]^2 \leq B_{\tilde{y}}^2$ for some parameters[1] $B_{\tilde{\mathbf{x}}}, B_{\tilde{y}}$. While this is a similar setting to that discussed in [4], the results there focus on learning with general kernels and loss functions, using a random number of independent noisy copies of each example. Here, we present an algorithm for learning linear predictors, using exactly two independent noisy copies of the instance $\mathbf{x}_t$ and one noisy copy of the target value y_t. As discussed in Section 3, the algorithm is based

[1] Strictly speaking, this is an upper bound on the raw second moment, not the variance. However, it is easily implied by assuming a bound on the noise variance, and a bound on the norm of the unperturbed data.

on an adaptation of the online convex optimization algorithm due to [10], and the main requirement is to construct an unbiased estimate of the gradient ∇_t. This follows from the following lemma, whose proof appears in the supplementary material.

Lemma 1 *Let $\nabla_t = 2(\langle \mathbf{w}_t, \mathbf{x}_t \rangle - y_t)\mathbf{x}_t$ be the gradient of $\ell(\langle \mathbf{w}_t, \mathbf{x}_t \rangle, y_t)$ at $\mathbf{w}_t$. Let $\tilde{\mathbf{x}}'_t$ be an additional independent copy of $\tilde{\mathbf{x}}_t$, and denote $\tilde{\nabla}_t = 2(\langle \mathbf{w}_t, \tilde{\mathbf{x}}_t \rangle - \tilde{y}_t)\tilde{\mathbf{x}}'_t$. Then under the assumptions above, if $\|\mathbf{w}_t\| \le B_{\mathbf{w}}$, then $\mathbb{E}[\tilde{\nabla}_t] = \nabla_t$ and $\mathbb{E}[\|\tilde{\nabla}_t\|^2] \le G$, where $G = 4(B_{\mathbf{w}}^2 B_{\tilde{\mathbf{x}}}^2 + B_{\tilde{y}}^2)B_{\tilde{\mathbf{x}}}^2$.*

Algorithm 1 Linear Regression with Noise.

Upper bound on variance

```
PARAMETERS: $\eta, B_{\mathbf{w}}$.
INITIALIZE: $\mathbf{w}_1 = \mathbf{0}$.
For $t = 1, 2, \ldots, T$
  Receive $(\tilde{\mathbf{x}}_t, \tilde{y}_t)$
  Receive another independent copy $\tilde{\mathbf{x}}'_t$
  $\tilde{\nabla}_t = 2(\langle \mathbf{w}_t, \tilde{\mathbf{x}}_t \rangle - \tilde{y}_t)\tilde{\mathbf{x}}'_t$
  $\mathbf{w}' = \mathbf{w}_t - \eta\tilde{\nabla}_t$
  $\mathbf{w}_{t+1} = \min\{1, B_{\mathbf{w}}/\|\mathbf{w}'\|\}\,\mathbf{w}'$
```

Known covariance

```
PARAMETERS: $\eta, B_{\mathbf{w}}$.
INITIALIZE: $\mathbf{w}_1 = \mathbf{0}$.
For $t = 1, 2, \ldots, T$
  Receive $(\tilde{\mathbf{x}}_t, \tilde{y}_t)$

  $\tilde{\nabla}_t = 2(\langle \mathbf{w}_t, \tilde{\mathbf{x}}_t \rangle - \tilde{y}_t)\tilde{\mathbf{x}}_t - \Sigma_t\mathbf{w}_t$
  $\mathbf{w}' = \mathbf{w}_t - \eta\tilde{\nabla}_t$
  $\mathbf{w}_{t+1} = \min\{1, B_{\mathbf{w}}/\|\mathbf{w}'\|\}\,\mathbf{w}'$
```

The following theorem provides a regret bound for Algorithm 1 (left). The proof is provided in the supplementary material.

Theorem 1 *Assume that $\mathbb{E}[\|\tilde{\mathbf{x}}_t\|^2] \le B_{\tilde{\mathbf{x}}}^2$, $\mathbb{E}[\tilde{y}_t]^2 \le B_{\tilde{y}}^2$, and that $\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}'_t, \tilde{y}_t$ are mutually independent. Then if we run Algorithm 1 (left) with parameters $B_{\mathbf{w}}$, $\eta = B_{\mathbf{w}}/\sqrt{GT}$ (where G is defined in Lemma 1), and compute $\tilde{\nabla}_t$ as in Lemma 1, we have*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\langle \mathbf{w}_t, \mathbf{x}_t \rangle, y_t) - \min_{\mathbf{w}:\|\mathbf{w}\| \le B_{\mathbf{w}}} \sum_{t=1}^{T} \ell(\langle \mathbf{w}, \mathbf{x}_t \rangle, y_t)\right] \le B_{\mathbf{w}}\sqrt{GT}.$$

5 Setting 2: Known Covariance

We now turn to the case where rather than an upper bound on the variance, we actually know the covariance matrix of the noise at each round, which we denote as Σ_t. We assume that $\|\Sigma_t\| \le B_\Sigma$ for all t, where $\|\cdot\|$ denotes the spectral norm. As to $\tilde{y}_t$, we can still assume we only have an upper bound $B_{\tilde{y}}^2$ on $\mathbb{E}[\tilde{y}_t^2]$ (with our algorithmic approach, knowing $\mathbb{E}[\tilde{y}_t^2]$ does not help much).

In this setting, we show it is possible to learn linear predictors, using just a single noisy copy $(\tilde{\mathbf{x}}_t, \tilde{y}_t)$. This is opposed to Section 4, where we needed an additional independent copy of $\tilde{\mathbf{x}}_t$. The idea is that if we use just one noisy copy in our gradient estimate, we need to deal with bias terms. When the covariance structure is known, we can calculate and remove these bias terms, allowing an online gradient descent similar to Algorithm 1 (left) to work. As in Algorithm 1 (left), the basic building block is a construction of an unbiased estimate of the gradient ∇_t at each iteration.

Lemma 2 *Let $\nabla_t = 2(\langle \mathbf{w}_t, \mathbf{x}_t \rangle - y_t)\mathbf{x}_t$ be the gradient of $\ell(\langle \mathbf{w}_t, \mathbf{x}_t \rangle, y_t)$ at $\mathbf{w}_t$. Denote $\tilde{\nabla}_t = 2(\langle \mathbf{w}_t, \tilde{\mathbf{x}}_t \rangle - \tilde{y}_t)\tilde{\mathbf{x}}_t - \Sigma_t\mathbf{w}_t$. Then under the assumptions above, if $\|\mathbf{w}_t\| \le B_{\mathbf{w}}$, $\mathbb{E}[\|\tilde{\mathbf{x}}_t\|^2] \le B_{\tilde{\mathbf{x}}}^2$, and $\mathbb{E}[\|\tilde{\mathbf{x}}_t\|^4] \le B_{\tilde{\mathbf{x}}}'^4$, then $\mathbb{E}[\tilde{\nabla}_t] = \nabla_t$ and $\mathbb{E}[\|\tilde{\nabla}\|_t^2] \le G$, where $G = 8B_{\mathbf{w}}^2 B_{\tilde{\mathbf{x}}}'^4 + 8B_{\tilde{y}}^2 B_{\tilde{\mathbf{x}}}^2 + 4B_{\mathbf{w}}^2 B_{\tilde{\mathbf{x}}}^2 B_\Sigma + B_\Sigma^2 B_{\mathbf{w}}^2$.*

The proof is a slightly more involved version of the proof of Lemma 1, and appears in the supplementary material. See Algorithm 1 (right) for the pseudocode.

Theorem 2 *Assume that $\mathbb{E}[\|\tilde{\mathbf{x}}_t\|^2] \le B_{\tilde{\mathbf{x}}}$, $\mathbb{E}[\|\tilde{\mathbf{x}}_t\|^4] \le B'^4_{\tilde{\mathbf{x}}}$, $\mathbb{E}[\|\tilde{y}_t\|^2] \le B^2_{\tilde{y}}$. Suppose the noise introduced into $\mathbf{x}_t$ is independent of that introduced into $\mathbf{y}_t$, and has a known covariance matrix Σ_t such that $\|\Sigma_t\| \le B_\Sigma$ for all t. Then if we run Algorithm 1 (right) with parameters $B_\mathbf{w}$ and $\eta = B_\mathbf{w}/\sqrt{T}$, where G is defined in Lemma 2, we have*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\langle \mathbf{w}_t, \mathbf{x}_t\rangle, y_t) - \min_{\mathbf{w}:\|\mathbf{w}\|\le B_\mathbf{w}} \sum_{t=1}^{T} \ell(\langle \mathbf{w}, \mathbf{x}_t\rangle, y_t)\right] \le \frac{G+1}{2} B_\mathbf{w}\sqrt{T}.$$

The proof is similar to the proof of Theorem 1, with Lemma 2 replacing Lemma 1. We note that if G is known, then by picking $\eta = B_\mathbf{w}/\sqrt{GT}$, one can improve the bound to $B_\mathbf{w}\sqrt{GT}$.

6 Setting 3: Gaussian Distribution

The final and most complex setting we consider is when the noise is assumed to have a Gaussian distribution $\mathcal{N}(0,\Sigma)$. Clearly, if we know the distribution, then we can derive upper bounds on the moments of $\tilde{\mathbf{x}}_t$ (assuming bounds are known on the original instances $\mathbf{x}_t$). Thus, the results of Section 5 carry through to our setting, and we can learn linear predictors. However, when we also know the noise has a specific Gaussian distribution, we can learn the much more powerful hypothesis class of kernel-based predictors.

The basic premise of kernel learning is that we map our data (originally in $\mathbb{R}^d$) to some reproducing kernel Hilbert space (RKHS), via a feature mapping $\Psi(\mathbf{x})$, and learn a linear predictor in that space. In our original space, this corresponds to learning a non-linear function. Using the well-known kernel trick, inner products $\langle\Psi(\mathbf{x}),\Psi(\mathbf{x}')\rangle$ in the RKHS (which might be infinite-dimensional) can be easily computed via a kernel function $k(\mathbf{x},\mathbf{x}')$.

While there are many possible kernel functions, perhaps the most popular one is the Gaussian kernel, defined as $k(\mathbf{x},\mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|^2/s^2)$ for some s^2 (the kernel width). This corresponds to the inner product $\langle\Psi(\mathbf{x}),\Psi(\mathbf{x}')\rangle$ in an appropriate RKHS. We we will show below how to learn from noisy data with Gaussian kernels. In Subsection 6.1 and the supplementary material, we show how this can be extended, in a certain sense, to general *radial* kernels, i.e., kernels of the form $k(\mathbf{x},\mathbf{x}') = f(\|\mathbf{x}-\mathbf{x}'\|)$ for an appropriate real function f.

In this section, we assume that the noise distribution is fixed for all t. Hence, we may assume w.l.o.g. that Σ is a diagonal matrix, with element σ_i^2 at row/column i. To see why, notice that there always exists a rotation matrix R, such that $R\,\tilde{\mathbf{x}}_t$ has a Gaussian distribution with diagonal covariance matrix. Therefore, instead of learning w.r.t. $\{(\tilde{\mathbf{x}}_t, y_t)\}_{t=1}^T$, we can just learn w.r.t. $\{(R\,\tilde{\mathbf{x}}_t, y_t)\}_{t=1}^T$, and predict on any instance $\mathbf{x}$ by pre-rotating it using R. Since a uniform rotation in the data space does not change norms and distances, all our guarantees will hold. As to $\tilde{y}_t$, similar to the previous settings, we will only need to assume that $\mathbb{E}[\tilde{y}_t^2] \le B^2_{\tilde{y}}$ for some parameter $B_{\tilde{y}}$.

The algorithm that we present (Algorithm 2) is based on being able to receive two independent copies of each instance $\tilde{\mathbf{x}}_t$, as well as a single independent copy of $\tilde{\mathbf{y}}_t$. As in the linear case, the learning algorithm that we use is based on the online gradient descent procedure due to [10], with the main twist being that instead of using a Gaussian kernel of width s^2, we use a surrogate kernel, as discussed in Section 3.

To define the surrogate kernel that we use, consider the RKHS corresponding to the kernel

$$\hat{k}(\mathbf{x},\mathbf{x}') = R^2_{\Sigma,s,d}\exp\left(-\sum_{i=1}^{d}\frac{(x_i-x_i')^2}{s^2-2\sigma_i^2}\right), \quad \text{where } R_{\Sigma,s,d} = \left(\prod_{i=1}^{d}\frac{s^2}{s^2-2\sigma_i^2}\right)^{1/4}. \tag{5}$$

This can be shown to be a kernel by standard results (see for instance [8]). Note that we generally assume that $2\|\Sigma\| = 2\max_i \sigma_i^2 < s^2$ for all i. Moreover, $R_{\Sigma,s,d}$ can be bounded by a constant when $\sigma_i = O(1)$ for all i (constant noise) and $s^2 = \Theta(d)$ —plausible when the feature values of instances $\mathbf{x}$ provided are of order $\Theta(1)$.

Denote $\hat{\Psi}$ to be the feature mapping corresponding to this RKHS. The pseudocode of our algorithm is presented below. Formally speaking, it is just applying Zinkevich's algorithm, using kernels, in the surrogate RKHS that we constructed. However, it is crucial to note that the actual output are elements $\mathbf{w}_1, \mathbf{w}_2, \ldots$ in the RKHS corresponding to Ψ.

Algorithm 2 Kernel Learning Algorithm with Gaussian Noise $\mathcal{N}(0, \Sigma)$

```
PARAMETERS: $W, \eta$
INITIALIZE:
  $\alpha_i := 0$ for all $i = 1 \ldots T$
For $t = 1, \ldots, T$:
  Define $\hat{\mathbf{w}}_t = \sum_{i=1}^{t-1} \alpha_i \hat{\Psi}(\tilde{\mathbf{x}}_i)$
  Define $\mathbf{w}_t = \sum_{i=1}^{t-1} \alpha_i \Psi(\tilde{\mathbf{x}}_i)$
  Receive $\tilde{y}_t$, $\tilde{\mathbf{x}}_t$, and independent copy $\tilde{\mathbf{x}}'_t$
  Let $\tilde{g}_t := 2\left(\sum_{i=1}^{t-1} \alpha_i \hat{k}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}'_t) - \tilde{y}_t\right)$    //$\tilde{g}_t$ is gradient length w.r.t. $\hat{\Psi}(\tilde{\mathbf{x}}'_t)$ at $\hat{\mathbf{w}}_t$
  Let $\alpha_t := -\eta \tilde{g}_t$
  Let $r_t := \sum_{i=1}^{t} \sum_{j=1}^{t} \alpha_i \alpha_j \hat{k}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$
    If $r_t > W^2$    // If $\|\hat{\mathbf{w}}_t\|^2 > W^2$, then project
      Let $\alpha_i := \alpha_i \frac{W}{\sqrt{r_t}}$ for all $i = 1, \ldots, t$
```

Before stating the bound for Algorithm 2 we need an auxiliary definitions. Suppose that $\mathbf{u}$ is any element in the RKHS of Ψ, which can be written as $\sum_{t=1}^{T} a_t \Psi(\mathbf{x}_t)$ for some $a_1, \ldots, a_m \geq 0$ (for example, this includes $\operatorname{argmin}_{\mathbf{w}:\|\mathbf{w}\| \leq B_{\mathbf{w}}} \sum_{t=1}^{T} \ell(\langle \mathbf{w}_t, \Psi(\mathbf{x}_t)\rangle, y_t)$ for any $B_{\mathbf{w}} > 0$ by the representer theorem). Define $\beta_{\mathbf{u}}$ to be the angle between $\sum_{t:a_t>0} a_t \Psi(\mathbf{x}_t)$ and $-\sum_{t:a_t<0} a_t \Psi(\mathbf{x}_t)$ (in other words, the angle between the component due to positive support vectors, and the component due to the negative support vectors). If one of the components is zero, define $\beta_{\mathbf{u}}$ to be $\pi/2$. The main theorem, whose proof is presented in the supplementary material, is the following.

Theorem 3 *Assume that the noise introduced into $\mathbf{x}_t$ has a known distribution $\mathcal{N}(0, \Sigma)$, with Σ diagonal and $2\|\Sigma\| \leq s^2$ (where s^2 is the kernel width), and that the noise introduced into y_t is independent with $\mathbb{E}[\tilde{y}_t^2] \leq B_{\tilde{y}}^2$. Let $B_{\mathbf{w}} > 0, \beta \in (0, \pi/2]$ be fixed. Then if we run Algorithm 2 with any $W \geq \frac{\sqrt{5} B_{\mathbf{w}} R_{\Sigma,s,d}}{\sin(\beta)}$, and $\eta = W/2R\sqrt{(W^2 R_{\Sigma,s,d}^2 + B_{\tilde{y}}^2)T}$, then for all $\mathbf{u}$ with $\|\mathbf{u}\| \leq B_{\mathbf{w}}, \beta_{\mathbf{u}} \geq \beta$, it holds that*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\langle \mathbf{w}_t, \Psi(\mathbf{x}_t)\rangle, y_t) - \ell(\langle \mathbf{u}, \Psi(\mathbf{x}_t)\rangle, y_t)\right] \leq 2WR_{\Sigma,s,d}\sqrt{(W^2 R_{\Sigma,s,d}^2 + B_{\tilde{y}}^2)T}$$

In particular, if $s^2 = \Omega(d)$, $\|\Sigma\| = O(1)$, and $B_{\tilde{y}} = O(1)$, then the right-hand side is $O(W^2\sqrt{T})$.

The intuition for $\beta_{\mathbf{u}}$ is that it measures how well separated are the training examples: if the "positive" and "negative" example groups are not too close together, then the angle between $\sum_{t:a_t>0} a_t \Psi(\mathbf{x}_t)$ and $-\sum_{i:a_t<0} a_t \Psi(\mathbf{x}_t)$ will be large, and the bound will be small. Note that in the RKHS corresponding to a Gaussian kernel, $\beta_{\mathbf{u}}$ is *always* between 0 and $\pi/2$, since the inner product between any two elements $\Psi(\mathbf{x})$ and $\Psi(\mathbf{x}')$ is positive. In addition, $\beta_{\mathbf{u}}$ can be shown to be exactly zero if and only if the positive and negative examples exactly coincide. Overall, on realistic datasets, assuming there exist some good predictor $\mathbf{u}$ with $\beta_{\mathbf{u}}$ not too small is a pretty mild assumption, if something interesting can be learned even on the unperturbed data.

6.1 Extension to General Radial Kernels

The Gaussian kernel we discussed previously is a member of the family of *radial kernels*, that is kernels on $\mathbf{x}, \mathbf{x}'$ which can be written as a function of $\|\mathbf{x} - \mathbf{x}'\|$. Although the Gaussian kernel is the most popular member of this family, there are many other radial kernels, such as $\exp(-\|\mathbf{x} - \mathbf{x}'\|/s)$ and $(1 + \|\mathbf{x} - \mathbf{x}'\|^2/m)^{-\alpha}$ for appropriate parameters s, m, α. Thus, a reasonable question is whether Algorithm 2 and its analysis can be extended to general radial kernels. The extension we are able to show is in the following sense: for any radial kernel $g(\mathbf{x}, \mathbf{x}')$, there exists another radial kernel $k(\mathbf{x}, \mathbf{x}')$, which approximates $g(\mathbf{x}, \mathbf{x}')$ well, for which one can extend Algorithm 2 and its analysis. The approximation parameter is user-defined, but the flip side is that the regret bound depends on this parameter and deteriorates as the approximation gets better.

Recall from Subsec. 3.2 that the heart of our approach is constructing a surrogate RKHS, with surrogate kernel $\hat{k}$, such that $\mathbb{E}[\hat{k}(\mathbf{a}, \tilde{\mathbf{x}})] = k(\mathbf{a}, \mathbf{x})$. In the Gaussian kernel case, the required surrogate RKHS corresponds to the kernel defined in Eq. (5). To deal with other kernels, constructing an appropriate surrogate kernel becomes trickier. Luckily, we can still reduce the problem, in some sense, to the case of Gaussian kernels. The key technical result is the following theorem due to Schoenberg ([7], see also [5]), slightly paraphrased and adapted to our purposes[2]:

Theorem 4 (Schoenberg's Theorem) *A function $g(\cdot, \cdot)$ is a radial kernel corresponding to a valid RKHS, if and only if there exists a finite nonnegative measure μ on $[0, \infty)$, such that for any $\mathbf{x}, \mathbf{a} \in \mathbb{R}^d$,*

$$g(\mathbf{x}, \mathbf{a}) = \int_{u=0}^{\infty} \exp\left(-u \left\|\mathbf{x} - \mathbf{a}\right\|^2\right) \mu(u) du.$$

This result asserts that, up to normalization factors, radial kernels can be characterized as Laplace transforms of probability measures on the positive reals. Schoenberg's Theorem has been used by Micchelli et al. [3] to prove universality of radial kernels and by Scovel et al. [9] to establish approximation error bounds. A related result is Bochner's theorem (see, e.g., [6]), which characterizes the more general class of shift-invariant kernels as Fourier transforms of multivariate distributions on $\mathbb{R}^d$.

The above theorem implies that we can write inner products in our RKHS using the approximate kernel

$$k(\mathbf{x}, \mathbf{a}) = \int_{u=0}^{c} k_u(\mathbf{x}, \mathbf{a}) \mu(u) du \tag{6}$$

where $c > 0$ is a parameter and k_u is the Gaussian kernel $k_u(\mathbf{x}, \mathbf{a}) = \exp(-u \left\|\mathbf{x} - \mathbf{a}\right\|^2)$ with kernel width $1/u$. Note that this is a valid kernel by the reverse direction of Theorem 4. If c is chosen not too small, then $k(\mathbf{x}, \mathbf{a})$ is an excellent approximation to $g(\mathbf{x}, \mathbf{a})$ for all $\mathbf{x}, \mathbf{a}$. The reason why we must settle for approximations of the radial kernel, rather than the kernel itself, is the following: for each k_u in the above integral, we construct a surrogate kernel $\hat{k}_u$ such that $\mathbb{E}_{\tilde{\mathbf{x}}}[\hat{k}_u(\tilde{\mathbf{x}}, \mathbf{a})] = k_u(\mathbf{x}, \mathbf{a})$. The surrogate kernel $\hat{k}_u$ is based on subtracting certain constants from the kernel width $1/u$ along each dimension, and this cannot be done if u is larger than those constants.

By Fubini's theorem, we can write Eq. (6) as

$$k(\mathbf{x}, \mathbf{a}) = \int_{u=0}^{c} \mathbb{E}\left[\hat{k}_u(\tilde{\mathbf{x}}, \mathbf{a})\right] \mu(u) du = \mathbb{E}\left[\int_{u=0}^{c} \hat{k}_u(\tilde{\mathbf{x}}, \mathbf{a}) du\right].$$

It turns out that the integral inside the expectation corresponds to an inner product, in a valid RKHS, between the noisy instance $\tilde{\mathbf{x}}$ and $\mathbf{a}$. This will be our surrogate kernel for k.

Of course, there are many details to fill in, and the analysis depends on the radial kernel chosen. In the supplementary material, we present a full analysis of the regret for the concrete radial kernel

$$k(\mathbf{x}, \mathbf{x}') = \left(1 + \frac{1}{d} \left\|\mathbf{x} - \mathbf{x}'\right\|^2\right)^{-1}.$$

Note that the scaling factor $1/d$ is the reasonable one to take, when we assume that the attribute values in the instances are on the order of $\Theta(1)$.

[2] To be precise, the theorem here is a corollary of Schoenberg's theorem, which discusses necessary and sufficient conditions for $k(\cdot, \cdot)$ to be positive definite, and Mercer's theorem (see [8]), which asserts that such a function is a kernel of a valid RKHS.

References

[1] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, September 2004.

[2] A. Flaxman, A. Tauman Kalai, and H. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of SODA*, pages 385–394, 2005.

[3] C.A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

[4] Shai Shalev-Shwartz Nicolò Cesa-Bianchi and Ohad Shamir. Online learning of noisy data with kernels. In *Proceedings of COLT*, 2010.

[5] P. Ressel. A short proof of Schoenberg's theorem. *Proceedings of the AMS*, 57(1):66–68, May 1976.

[6] W. Rudin. *Fourier Analysis on Groups (reprint edition)*. Wiley Classics Library, 1994.

[7] I. Schoenberg. Metric spaces and completely monotone functions. *The Annals of Mathematics*, 39(4):811–841, October 1938.

[8] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

[9] C. Scovel, D. Hush, I. Steinwart, and J. Theiler. Radial kernels and their Reproducing Kernel Hilbert Spaces. Technical Report LA-UR-09-01900, Los Alamos National Laboratory, March 2010.

[10] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

A Proof of Lemma 1

Proof Because of the independence assumption, we have that

$$\mathbb{E}[\tilde{\nabla}_t] = 2\,\mathbb{E}[\langle \mathbf{w}_t, \tilde{\mathbf{x}}_t\rangle - \tilde{y}_t]\,\mathbb{E}[\tilde{\mathbf{x}}'_t] = 2(\langle \mathbf{w}_t, \mathbf{x}_t\rangle - y_t)\mathbf{x}_t = \nabla_t \ .$$

For the second claim, we have by the independence assumption that

$$\begin{aligned}
\mathbb{E}[\|\tilde{\nabla}_t\|^2] &= 4\,\mathbb{E}\left[(\langle \mathbf{w}_t, \tilde{\mathbf{x}}_t\rangle - \tilde{y}_t)^2\right]\mathbb{E}[\|\tilde{\mathbf{x}}'_t\|]^2 \\
&= 4\left(\mathbb{E}\left[\langle \mathbf{w}_t, \tilde{\mathbf{x}}_t\rangle^2\right] + \mathbb{E}[\tilde{y}_t^2]\right)\mathbb{E}[\|\tilde{\mathbf{x}}'_t\|]^2 \\
&\leq 4\left(B_{\mathbf{w}}^2 B_{\tilde{\mathbf{x}}}^2 + B_{\tilde{y}}^2\right) B_{\tilde{\mathbf{x}}}^2.
\end{aligned}$$

■

B Proof of Theorem 1

First, we use the following lemma that can be easily adapted from [10].

Lemma 3 *Let $\mathbf{v}_1, \ldots, \mathbf{v}_T$ be a sequence of vectors. Let $\mathbf{w}_1 = \mathbf{0}$ and for $t \geq 1$ let $\mathbf{w}_t = P(\mathbf{w}_t - \eta \mathbf{v}_t)$, where $P(\cdot)$ is the projection operator on an origin-centered ball of radius $B_{\mathbf{w}}$. Then, for all $\mathbf{u}$ s.t. $\|\mathbf{u}\| \leq B_{\mathbf{w}}$ we have*

$$\sum_{t=1}^{m} \langle \mathbf{v}_t, \mathbf{w}_t - \mathbf{u}\rangle \leq \frac{B_{\mathbf{w}}^2}{2\eta} + \frac{\eta \sum_{t=1}^{T} \|\mathbf{v}_t\|^2}{2} \ .$$

Applying Lemma 3 with $\tilde{\nabla}_t$ as defined in Lemma 1 we obtain:

$$\sum_{t=1}^{T} \left\langle \tilde{\nabla}_t, \mathbf{w}_t - \mathbf{u}\right\rangle \leq \frac{B_{\mathbf{w}}^2}{2\eta} + \frac{\eta \sum_{t=1}^{T} \|\tilde{\nabla}_t\|^2}{2} \ .$$

Taking expectation of both sides and using again Lemma 1 we obtain that

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \nabla_t, \mathbf{w}_t - \mathbf{u}\rangle\right] \leq \frac{B_{\mathbf{w}}^2}{2\eta} + \frac{\eta T G}{2} \ ,$$

Now, using convexity we get that $\langle \nabla_t, \mathbf{w}_t - \mathbf{u}\rangle \geq (\langle \mathbf{w}_t, \mathbf{x}_t\rangle - y_t)^2 - (\langle \mathbf{w}_t, \mathbf{u}\rangle - y_t)^2$, which gives

$$\mathbb{E}\left[\sum_{t=1}^{T} (\langle \mathbf{w}_t, \mathbf{x}_t\rangle - y_t)^2\right] \leq \sum_{t=1}^{T} (\langle \mathbf{u}, \mathbf{x}_t\rangle - y_t)^2 + \frac{B_{\mathbf{w}}^2}{2\eta} + \frac{\eta T G}{2} \ .$$

Picking η as in the theorem statement concludes our proof.

C Proof of Lemma 2

Using the zero-mean and independence assumptions on $\mathbf{n}_t^x, n_t^y$, we have

$$\mathbb{E}[\tilde{\nabla}_t] + \Sigma_t \mathbf{w}_t = 2\,\mathbb{E}[(\langle \mathbf{w}_t, \mathbf{x}_t + \mathbf{n}_t^x\rangle - y_t - n_t^y)(\mathbf{x}_t + \mathbf{n}_t^x)] \ = \ 2(\langle \mathbf{w}_t, \mathbf{x}_t\rangle - y_t)\mathbf{x}_t + \mathbb{E}[\mathbf{n}_t^x \langle \mathbf{w}_t, \mathbf{n}_t^x\rangle],$$

which equals $\nabla_t + \Sigma_t \mathbf{w}_t$. So we have $\mathbb{E}[\tilde{\nabla}_t] = \nabla_t$. As to the second claim, using the well-known inequality $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$, we have

$$\begin{aligned}
&\mathbb{E}[\|\tilde{\nabla}_t\|^2] = \mathbb{E}[\|2(\langle \mathbf{w}_t, \tilde{\mathbf{x}_t} - \tilde{y}_t\rangle)\tilde{\mathbf{x}}_t - \Sigma_t \mathbf{w}_t\|^2] \\
&= \mathbb{E}\,\|2(\langle \mathbf{w}_t, \tilde{\mathbf{x}}_t\rangle - \tilde{y}_t)\tilde{\mathbf{x}}_t\|^2 - 4\mathbf{w}_t^\top \Sigma_t\,\mathbb{E}[(\langle \mathbf{w}_t, \tilde{\mathbf{x}}_t\rangle - y_t)\tilde{\mathbf{x}}_t] + \|\Sigma_t \mathbf{w}_t\|^2 \\
&\leq 8\,\mathbb{E}[\|\langle \mathbf{w}_t, \tilde{\mathbf{x}}_t\rangle\,\tilde{\mathbf{x}}_t\|^2] + 8\,\mathbb{E}[\|\tilde{y}_t \tilde{\mathbf{x}}_t\|^2] - 4\mathbf{w}_t^\top \Sigma_t\,\mathbb{E}[\langle \mathbf{w}_t, \tilde{\mathbf{x}}_t\rangle\,\mathbf{x}_t] + \|\Sigma_t \mathbf{w}_t\|^2 \\
&\leq 8\,\|\mathbf{w}_t\|^2\,\mathbb{E}[\|\tilde{\mathbf{x}}_t\|^4] + 8\,\mathbb{E}[\tilde{y}_t^2]\,\mathbb{E}[\left\|\tilde{\mathbf{x}}_t^2\right\|] + 4\,\|\mathbf{w}_t\|^2\,\mathbb{E}[\|\tilde{\mathbf{x}}_t\|^2]\,\|\Sigma_t\| + \|\Sigma_t\|^2\,\|\mathbf{w}_t\|^2 \\
&\leq 8B_{\mathbf{w}}^2 B_{\tilde{\mathbf{x}}}^{\prime 4} + 8B_{\tilde{y}}^2 B_{\tilde{\mathbf{x}}}^2 + 4B_{\mathbf{w}}^2 B_{\tilde{\mathbf{x}}}^2 B_{\Sigma} + B_{\Sigma}^2 B_{\mathbf{w}}^2.
\end{aligned}$$

D Proof of Theorem 3

D.1 Auxiliary Lemmas

To prove the theorem, we will need a few auxiliary lemmas. In particular, Lemma 4 is a key technical lemma, which will prove crucial in connecting the RKHS w.r.t. $\Psi(\cdot), k(\cdot,\cdot)$, and the RKHS w.r.t. $\hat{\Psi}(\cdot), \hat{k}(\cdot,\cdot)$. Lemma 6 connects between the norms of elements in the two RKHS's.

To state the lemmas and proofs conveniently, recall the shorthand

$$R_{\Sigma,s,d} = \left(\prod_{i=1}^{d} \frac{s^2}{s^2 - 2\sigma_i^2}\right)^{1/4}.$$

Lemma 4 *For any $\mathbf{a}, \mathbf{x} \in \mathbb{R}^d$, if we let $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{n}$ where $\mathbf{n} \sim \mathcal{N}(0, \Sigma)$ is a Gaussian random vector with covariance matrix Σ, then it holds that*

$$\mathop{\mathbb{E}}_{\mathbf{n}} \left\langle \hat{\Psi}(\mathbf{a}), \hat{\Psi}(\tilde{\mathbf{x}}) \right\rangle = \langle \Psi(\mathbf{a}), \Psi(\mathbf{x}) \rangle .$$

Proof The expectation in the lemma can be written as

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{n}} \left[R^2_{\Sigma,s,d} \exp\left(-\sum_{i=1}^{d} \frac{(a_i - x_i - n_i)^2}{s^2 - 2\sigma_i^2} \right) \right] \\ &= \prod_{i=1}^{d} R^{2/d}_{\Sigma,s,d} \int_{z_i} (2\pi s^2)^{-1/2} \exp\left(-\frac{z_i^2}{2\sigma_i^2} - \frac{(a_i - x_i - z_i)^2}{s^2 - 2\sigma_i^2} \right) dz_i \qquad (7) \end{aligned}$$

A purely technical integration exercise reveals that each element i in this product equals $\exp(-(a_i - x_i)^2/s^2)$. Therefore, Eq. (7) equals

$$\prod_{i=1}^{d} \exp\left(-\frac{(a_i - x_i)^2}{s^2} \right) = \exp\left(-\frac{\|\mathbf{a} - \mathbf{x}\|^2}{s^2} \right),$$

which is exactly $\langle \Psi(\mathbf{x}), \Psi(\mathbf{a}) \rangle$. ■

Lemma 5 *Let $\Psi(\cdot)$ denote a feature mapping to an arbitrary RKHS. Let $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T$ be vectors in $\mathbb{R}^d$, and $a_1, a_2, \ldots, a_T$ scalars, such that $\left\|\sum_i a_i \Psi(\mathbf{z}_i)\right\|^2 \leq B^2$ for some $B > 0$. Then it holds that*

$$\left\| \sum_{i:a_i>0} a_i \Psi(\mathbf{z}_i) \right\|^2 + \left\| \sum_{i:a_i<0} a_i \Psi(\mathbf{z}_i) \right\|^2 \leq \frac{5B^2}{\sin^2(\beta)},$$

where β is the angle between $\sum_{i:a_i>0} a_i \Psi(\mathbf{z}_i)$ and $\sum_{i:a_i<0} a_i \Psi(\mathbf{z}_i)$ in the RKHS (or $\beta = \pi/2$ if one of these elements is zero).

We remark that this bound is designed for readability - it is not the tightest upper bound possible.

Proof The bound trivially holds if $\sum_{i:a_i>0} a_i \Psi(\mathbf{z}_i)$ or $\sum_{i:a_i<0} a_i \Psi(\mathbf{z}_i)$ are zero, so we will assume w.l.o.g. that they are both non-zero.

To simplify notation, let

$$\mathbf{w} = \sum_i a_i \Psi(\mathbf{z}_i), \qquad \mathbf{w}_+ = \sum_{i:a_i>0} a_i \Psi(\mathbf{z}_i), \qquad \mathbf{w}_- = \sum_{i:a_i<0} -a_i \Psi(\mathbf{z}_i).$$

and notice that $\mathbf{w} = \mathbf{w}_+ - \mathbf{w}_-$. By the cosine theorem and the fact that $\mathbf{w} = \mathbf{w}_+ - \mathbf{w}_-$, we have that

$$\|\mathbf{w}\|^2 = \|\mathbf{w}_+\|^2 + \|\mathbf{w}_-\|^2 - 2\|\mathbf{w}_+\|\|\mathbf{w}_-\|\cos(\beta).$$

Solving for $\|\mathbf{w}_-\|$ and taking the larger root in the resulting quadratic equation, we have that

$$\|\mathbf{w}_-\| \leq \|\mathbf{w}_+\|\cos(\beta) + \sqrt{\|\mathbf{w}\|^2 - \|\mathbf{w}_+\|^2\sin^2(\beta)} \tag{8}$$

(it is easy to verify that the term in the square root is always non-negative). Therefore

$$\begin{aligned}\|\mathbf{w}_+\|^2 + \|\mathbf{w}_-\|^2 &\leq \|\mathbf{w}_+\|^2 + \left(\|\mathbf{w}_+\|\cos(\beta) + \sqrt{\|\mathbf{w}\|^2 - \|\mathbf{w}_+\|^2\sin^2(\beta)}\right)^2 \\ &\leq \|\mathbf{w}_+\|^2 + \left(\|\mathbf{w}_+\|\,|\cos(\beta)| + \|\mathbf{w}\|\right)^2.\end{aligned}$$

From straightforward geometric arguments, we must have $\|\mathbf{w}_+\|^2 \leq \|\mathbf{w}\|^2 / \sin^2(\beta)$ (this is the same reason the term in the square root in Eq. (8) is non-negative). Plugging this into the right hand side of the inequality above, we get an upper bound of the form

$$\frac{\|\mathbf{w}\|^2}{\sin^2(\beta)} + \left(\|\mathbf{w}\|\left|\frac{\cos(\beta)}{\sin(\beta)}\right| + \|\mathbf{w}\|\right)^2 \leq \|\mathbf{w}\|^2\left(1 + \frac{2}{|\sin(\beta)|} + \frac{2}{\sin^2(\beta)}\right),$$

where we used the fact that $|\cos(\beta)| \leq 1$. A straightforward upper bounding leads to the lemma statement. ■

The following lemma is basically a corollary of Lemma 5.

Lemma 6 *Let $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T$ be vectors in $\mathbb{R}^d$, and $a_1, a_2, \ldots, a_T$ scalars, such that $\|\sum_i a_i\Psi(\mathbf{z}_i)\|^2 \leq B^2$. Then $\sum_i a_i\hat{\Psi}(z_i)$ is an element in the RKHS w.r.t. $\hat{\Psi}(\cdot)$, whose norm squared is at most*

$$\frac{5B^2}{\sin^2(\beta)}R^2_{\Sigma,s,d}.$$

Here, β is the angle between $\sum_{i:a_i>0} a_i\Psi(\mathbf{z}_i)$ and $-\sum_{i:a_i<0} a_i\Psi(\mathbf{z}_i)$ in the RKHS (or $\beta = \pi/2$ if one of the elements is zero).

Proof Picking some $\mathbf{z}_1, \mathbf{z}_2, \ldots$ and $a_1, a_2, \ldots$ as in the lemma statement, we have

$$\left\|\sum_i a_i\hat{\Psi}(z_i)\right\|^2 = \sum_{i,j} a_i a_j \hat{k}(\mathbf{z}_i, \mathbf{z}_j) \leq \sum_{i,j:a_ia_j>0} a_i a_j \hat{k}(\mathbf{z}_i, \mathbf{z}_j), \tag{9}$$

where the last transition is by the fact that $\hat{k}$ is always positive. Now, by definition of $k(\cdot,\cdot), \hat{k}(\cdot,\cdot)$, it holds for any $\mathbf{z}_i, \mathbf{z}_j$ that

$$\frac{\hat{k}(\mathbf{z}_i, \mathbf{z}_j)}{k(\mathbf{z}_i, \mathbf{z}_j)} \;=\; R^2_{\Sigma,s,d}\exp\left(\sum_{l=1}^d \frac{(z_{i,l} - z_{j,l})^2}{s^2} - \frac{(z_{i,l} - z_{j,l})^2}{s^2 - 2\sigma_l^2}\right) \;\leq\; R^2_{\Sigma,s,d}.$$

Therefore, we can upper bound Eq. (9) by

$$\sum_{i,j:a_ia_j>0} a_i a_j \hat{k}(\mathbf{z}_i, \mathbf{z}_j) \leq R^2_{\Sigma,s,d}\sum_{i,j:a_ia_j>0} a_i a_j k(\mathbf{z}_i, \mathbf{z}_j).$$

The lemma follows by noting that according to Lemma 5,

$$\sum_{i,j:a_ia_j>0} a_i a_j k(\mathbf{z}_i, \mathbf{z}_j) = \left\|\sum_{i:a_i>0} a_i\Psi(\mathbf{z}_i)\right\|^2 + \left\|\sum_{i:a_i<0} \Psi(\mathbf{z}_i)\right\|^2 \leq \frac{5B^2}{\sin^2(\beta)}.$$

■

With these lemmas in hand, we are now ready to prove the main theorem.

D.2 Proof of Theorem 3

To make the proof clearer, let $\alpha_{t,i}$ denote the value of α_i in algorithm 2 at the beginning of round t.

The first step of the proof consists of applying Lemma 3, since our algorithm follows the protocol outlined in that lemma, using kernels. We therefore have that for any $\hat{\mathbf{u}}$ in the RKHS corresponding to $\hat{\Psi}(\cdot)$, such that $\|\hat{\mathbf{u}}\|^2 \leq W^2$, it holds that

$$\sum_{t=1}^{T} \left\langle \hat{\mathbf{w}}_t, \tilde{g}_t \hat{\Psi}(\tilde{\mathbf{x}}_t) \right\rangle - \sum_{t=1}^{T} \left\langle \hat{\mathbf{u}}, \tilde{g}_t \hat{\Psi}(\tilde{\mathbf{x}}_t) \right\rangle \leq \frac{W^2}{2\eta} + \frac{\eta \sum_{t=1}^{T} \left\| \tilde{g}_t \hat{\Psi}(\tilde{\mathbf{x}}_t) \right\|^2}{2}. \tag{10}$$

In particular, consider $\mathbf{u} = \sum_{i=1}^{T} a_t \Psi(\mathbf{x}_t)$ from the theorem's statement, and define

$$\hat{\mathbf{u}}^* = \sum_{i=1}^{T} a_t \hat{\Psi}(\mathbf{x}_t).$$

This is an element in the RKHS corresponding to $\hat{\Psi}(\cdot)$, but it shares the same set of weights as $\mathbf{u}$, which is an element in the RKHS corresponding to $\Psi(\cdot)$. Since $\|\mathbf{u}\|^2 \leq B_{\mathbf{w}}^2$, it follows from Lemma 6 and the definition of W that $\|\hat{\mathbf{u}}^*\|^2 \leq W^2$. Therefore, Eq. (10) applies, and we get

$$\sum_{t=1}^{T} \left\langle \hat{\mathbf{w}}_t, \tilde{g}_t \hat{\Psi}(\tilde{\mathbf{x}}_t) \right\rangle - \sum_{t=1}^{T} \left\langle \hat{\mathbf{u}}^*, \tilde{g}_t \hat{\Psi}(\tilde{\mathbf{x}}_t) \right\rangle \leq \frac{W^2}{2\eta} + \frac{\eta \sum_{t=1}^{T} \left\| \tilde{g}_t \hat{\Psi}(\tilde{\mathbf{x}}_t) \right\|^2}{2}.$$

This inequality holds for any $\{\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_t', \tilde{y}_t\}_{t=1}^{T}$. In particular, it will remain valid if we take expectations of both sides with respect to the Gaussian noise injected into the unperturbed data:

$$\mathbb{E}\left[\sum_{t=1}^{T} \left\langle \hat{\mathbf{w}}_t, \tilde{g}_t \hat{\Psi}(\tilde{\mathbf{x}}_t) \right\rangle - \sum_{t=1}^{T} \left\langle \hat{\mathbf{u}}^*, \tilde{g}_t \hat{\Psi}(\tilde{\mathbf{x}}_t) \right\rangle\right] \leq \frac{W^2}{2\eta} + \frac{\eta \sum_{t=1}^{T} \mathbb{E}\left[\left\| \tilde{g}_t \hat{\Psi}(\tilde{\mathbf{x}}_t) \right\|^2\right]}{2}. \tag{11}$$

Starting with the right hand side, we note that by definition of $\tilde{g}_t$ from the algorithm's pseudocode, and the fact that $\left\|\hat{\Psi}(\tilde{\mathbf{x}}_t)\right\|^2 = \hat{k}(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) \leq R_{\Sigma,s,d}^2$ by definition of the kernel $\hat{k}$ in Eq. (5),

$$\begin{aligned}
&\mathbb{E}\left[\left\|\tilde{g}_t \hat{\Psi}(\tilde{\mathbf{x}}_t')\right\|^2\right] = 4\,\mathbb{E}\left[\left\|\left(\left\langle \hat{\mathbf{w}}_t, \hat{\Psi}(\tilde{\mathbf{x}}_t)\right\rangle - \tilde{y}_t\right) \hat{\Psi}(\tilde{\mathbf{x}}_t')\right\|^2\right] \\
&= 4\,\mathbb{E}\left[\left\|\hat{\Psi}(\tilde{\mathbf{x}}_t')\right\|^2\right] \mathbb{E}\left[\left(\left\langle \hat{\mathbf{w}}_t, \hat{\Psi}(\tilde{\mathbf{x}}_t)\right\rangle - \tilde{y}_t\right)^2\right] = 4\,\mathbb{E}\left[\left\|\hat{\Psi}(\tilde{\mathbf{x}}_t')\right\|^2\right] \mathbb{E}\left[\left(\left\langle \hat{\mathbf{w}}_t, \hat{\Psi}(\tilde{\mathbf{x}}_t)\right\rangle\right)^2 + \tilde{y}_t^2\right] \\
&\leq 4\,\mathbb{E}\left[\left\|\hat{\Psi}(\tilde{\mathbf{x}}_t')\right\|^2\right] \left(\|\hat{\mathbf{w}}_t\|^2\, \mathbb{E}\left[\left\|\hat{\Psi}(\tilde{\mathbf{x}}_t)\right\|^2\right] + \mathbb{E}[\tilde{y}_t^2]\right) = 4R_{\Sigma,s,d}^2 (W^2 R_{\Sigma,s,d}^2 + B_{\tilde{y}}^2).
\end{aligned}$$

Plugging this back into Eq. (11), and choosing $\eta = W/2R\sqrt{(W^2 R_{\Sigma,s,d}^2 + B_{\tilde{y}}^2)T}$ as in the theorem's statement, we finally get

$$\mathbb{E}\left[\sum_{t=1}^{T} \left\langle \hat{\mathbf{w}}_t, \tilde{g}_t \hat{\Psi}(\tilde{\mathbf{x}}_t) \right\rangle - \sum_{t=1}^{T} \left\langle \hat{\mathbf{u}}^*, \tilde{g}_t \hat{\Psi}(\tilde{\mathbf{x}}_t) \right\rangle\right] \leq 2W R_{\Sigma,s,d} \sqrt{(W^2 R_{\Sigma,s,d}^2 + B_{\tilde{y}}^2)T}. \tag{12}$$

We now turn to analyze the more interesting left hand side of Eq. (12). Let $\mathbb{E}_t[\cdot]$ be a shorthand for expectation over $(\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_t', \tilde{y}_t)$ conditioned on $\{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_i', \tilde{y}_i)\}_{i=1}^{t-1}$. The l.h.s. of Eq. (12) can be written as

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathop{\mathbb{E}}_{t}\left[\left\langle \hat{\mathbf{w}}_t, \tilde{g}_t \hat{\Psi}(\tilde{\mathbf{x}}_t) \right\rangle\right] - \sum_{t=1}^{T} \mathop{\mathbb{E}}_{t}\left[\left\langle \hat{\mathbf{u}}^*, \tilde{g}_t \hat{\Psi}(\tilde{\mathbf{x}}_t) \right\rangle\right]\right]. \tag{13}$$

To analyze the first sum inside the expectation, recall that $\hat{\mathbf{w}}_t$ can be written as $\sum_{i=1}^{t-1} \alpha_{t,i}\hat{\Psi}(\tilde{\mathbf{x}}_i)$. Therefore, we have that

$$\mathop{\mathbb{E}}_t\left[\left\langle \hat{\mathbf{w}}_t, \tilde{g}_t\hat{\Psi}(\tilde{\mathbf{x}}_t)\right\rangle\right] = \sum_{i=1}^{t-1}\mathop{\mathbb{E}}_t\left[\alpha_{t,i}\left\langle \hat{\Psi}(\tilde{\mathbf{x}}_i), \tilde{g}_t\hat{\Psi}(\tilde{\mathbf{x}}_t)\right\rangle\right] = \sum_{i=1}^{t-1}\mathop{\mathbb{E}}_t[\tilde{g}_t]\mathop{\mathbb{E}}_t\left[\alpha_{t,i}\left\langle \hat{\Psi}(\tilde{\mathbf{x}}_i), \hat{\Psi}(\tilde{\mathbf{x}}_t)\right\rangle\right],$$

where the last transition is by the fact that $\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}'_t, \tilde{y}_t$ are mutually independent, and therefore $\tilde{g}_t$ is independent of $\hat{\Psi}(\tilde{\mathbf{x}}_t)$ conditioned on $\{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}'_i, \tilde{y}_i)\}_{i=1}^{t-1}$.

We now make two crucial observations, which are really the heart of our proof: First, by Lemma 4, we have that

$$\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_t\left[\alpha_{t,i}\left\langle \hat{\Psi}(\tilde{\mathbf{x}}_i), \hat{\Psi}(\tilde{\mathbf{x}}_t)\right\rangle\right] = \sum_{i=1}^{t-1}\alpha_{t,i}\left\langle \Psi(\tilde{\mathbf{x}}_i), \Psi(\mathbf{x}_t)\right\rangle = \left\langle \mathbf{w}_t, \Psi(\mathbf{x}_t)\right\rangle.$$

Secondly, using Lemma 4 in a similar manner, we also have

$$\mathop{\mathbb{E}}_t[\tilde{g}_t] = 2\left(\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_t\left[\alpha_{t,i}\left\langle \hat{\Psi}(\tilde{\mathbf{x}}_i), \hat{\Psi}(\mathbf{x}_t)\right\rangle - \tilde{y}_t\right]\right) = 2(\left\langle \mathbf{w}_t, \Psi(\mathbf{x}_t)\right\rangle - y_t).$$

Define this expression as g_t. Notice that it is exactly the gradient of $\mathbf{w}_t$ w.r.t. $\ell(\left\langle \mathbf{w}_t, \Psi(\mathbf{x}_t)\right\rangle, y_t)$.

As a result of these two observations, we get overall that

$$\mathop{\mathbb{E}}_t\left[\left\langle \hat{\mathbf{w}}_t, \tilde{g}_t\hat{\Psi}(\tilde{\mathbf{x}}_t)\right\rangle\right] = \sum_{i=1}^{t-1}\alpha_{t,i}\left\langle \Psi(\tilde{\mathbf{x}}_i), g_t\Psi(\mathbf{x}_t)\right\rangle = \left\langle \mathbf{w}_t, g_t\Psi(\mathbf{x}_t)\right\rangle. \tag{14}$$

Moving to the second sum in the l.h.s. of Eq. (13), recall that there exist some $a_1, \ldots, a_T$ such that $\hat{\mathbf{u}}^* = \sum_{t=1}^{T} a_t\hat{\Psi}(\mathbf{x}_t)$. Therefore,

$$\mathop{\mathbb{E}}_t\left[\left\langle \hat{\mathbf{u}}^*, \tilde{g}_t\hat{\Psi}(\tilde{\mathbf{x}}_t)\right\rangle\right] \;=\; \sum_{i=1}^{T}\mathop{\mathbb{E}}_t\left[a_i\left\langle \hat{\Psi}(\mathbf{x}_i), \tilde{g}_t\hat{\Psi}(\tilde{\mathbf{x}}_t)\right\rangle\right].$$

As before, we have by Lemma 4 that $\mathbb{E}_t\left[\left\langle \hat{\Psi}(\mathbf{x}_i), \hat{\Psi}(\tilde{\mathbf{x}}_t)\right\rangle\right] = \left\langle \Psi(\mathbf{x}_i), \Psi(\mathbf{x}_t)\right\rangle$, and that $\tilde{g}_t$ is conditionally independent with expected value $g_t = 2(\sum_{i=1}^{t-1}\alpha_{t,i}k(\tilde{\mathbf{x}}_i, \mathbf{x}_t) - y_t)$. Substituting this into the expression above, we get that it is equal to

$$\sum_{i=1}^{T} a_i\left\langle \Psi(\mathbf{x}_i), g_t\Psi(\mathbf{x}_t)\right\rangle = \left\langle \mathbf{u}, g_t\Psi(\mathbf{x}_t)\right\rangle.$$

Combining this and Eq. (14), and summing over t, we get that

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathop{\mathbb{E}}_t\left[\left\langle \hat{\mathbf{w}}_t, \tilde{g}_t\hat{\Psi}(\tilde{\mathbf{x}}_t)\right\rangle\right] - \sum_{t=1}^{T}\mathop{\mathbb{E}}_t\left[\left\langle \hat{\mathbf{u}}^*, \tilde{g}_t\hat{\Psi}(\tilde{\mathbf{x}}_t)\right\rangle\right]\right] \;=\; \mathbb{E}\left[\sum_{t=1}^{T}\left\langle \mathbf{w}_t, g_t\Psi(\mathbf{x}_t)\right\rangle - \sum_{t=1}^{T}\left\langle \mathbf{u}, g_t\Psi(\mathbf{x}_t)\right\rangle\right]. \tag{15}$$

Remarkably, this equation links between classifiers $\hat{\mathbf{w}}_t$ in the RKHS corresponding to $\Psi(\cdot)$, and the classifiers $\mathbf{w}_t$ in another RKHS, corresponding to $\Psi(\cdot)$.

Plugging Eq. (15) into Eq. (12), we get that

$$\mathbb{E}\left[\sum_{t=1}^{T}\left\langle \mathbf{w}_t, g_t\Psi(\mathbf{x}_t)\right\rangle - \sum_{t=1}^{T}\left\langle \mathbf{u}, g_t\Psi(\mathbf{x}_t)\right\rangle\right] \leq 2WR_{\Sigma,s,d}\sqrt{(W^2R_{\Sigma,s,d}^2 + B_{\tilde{y}}^2)T}.$$

Now, since $\ell(\left\langle \mathbf{w}_t, \Psi(\mathbf{x}_t)\right\rangle, y_t) = (\left\langle \mathbf{w}_t, \Psi(\mathbf{x}_t)\right\rangle - y_t)^2$ is a convex function of $\mathbf{w}_t$, and since $g_t\Psi(\mathbf{x}_t)$ is the gradient at $\mathbf{w}_t$, we can lower bound the l.h.s. as

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(\left\langle \mathbf{w}_t, \Psi(\mathbf{x}_t)\right\rangle, y_t) - \ell(\left\langle \mathbf{u}, \Psi(\mathbf{x}_t)\right\rangle, y_t)\right],$$

from which the theorem follows.

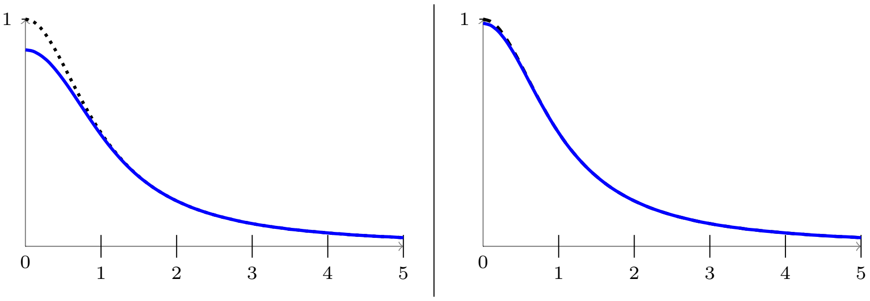

Figure 1: Comparison of $g(\mathbf{x}, \mathbf{x}')$ (solid line) and $k(\mathbf{x}, \mathbf{x}')$ (dashed line) as a function of $\|\mathbf{x} - \mathbf{x}'\|$, for $c = 2$ (left) and $c = 4$ (right).

E Extension to Radial Kernel Approximations: a Case Study

Recall that we assume a Gaussian distribution $\mathcal{N}(0, \Sigma)$ for the noise introduced into each instance, where Σ is a known matrix. Just to make our case analysis simpler to present, we assume here that $\Sigma = \sigma^2 I$ for some parameter σ^2, where $\sigma^2 = O(1)$ (this is a reasonable assumption to make when the feature values of the original data is $\Theta(1)$).

In this appendix, we adapt Algorithm 2 and Theorem 3 to an approximation of the radial kernel

$$g(\mathbf{x}, \mathbf{x}') = \left(1 + \frac{1}{d} \|\mathbf{x} - \mathbf{x}'\|^2\right)^{-1} = d \int_0^\infty \exp\left(-u \|\mathbf{x} - \mathbf{x}'\|^2 - ud\right) du$$

As discussed in the main text, the analysis we perform here can be extended to other radial kernels as well.

The approximate kernel we will consider is

$$k(\mathbf{x}, \mathbf{x}') = d \int_0^{c/d} \exp\left(-u \|\mathbf{x} - \mathbf{x}'\|^2 - ud\right) du = g(\mathbf{x}, \mathbf{x}') \left(1 - \exp\left(-\frac{c}{g(\mathbf{x}, \mathbf{x}')}\right)\right) \qquad (16)$$

where $c \in (0, d/4\sigma)$ is a user-defined parameter. This is a valid kernel by Eq. (20) and the reverse direction of Theorem 4. Note that $g(\mathbf{x}, \mathbf{x}')$ is always between 0 and 1, so $\frac{k(\mathbf{x},\mathbf{x}')}{g(\mathbf{x},\mathbf{x}')} \in [1 - \exp(-c), 1]$. Therefore, $k(\mathbf{x}, \mathbf{x}')$ is an excellent approximation of $g(\mathbf{x}, \mathbf{x}')$ for values of c not too small. As before, we let Ψ denote the feature mapping associated with the kernel k.

The surrogate kernel that we will pick is defined as follows (see the proof of Lemma 8 for the construction, following the approach outlined in Subsection 6.1):

$$\hat{k}(\mathbf{x}, \mathbf{x}') = d \int_{u=0}^{c/d} (1 - 2\sigma u)^{-d/2} \exp\left(-\frac{u \|\mathbf{x} - \mathbf{x}'\|^2}{1 - 2\sigma u} - ud\right) du. \qquad (17)$$

As before, we let $\hat{\Psi}$ denote the feature mapping associated with this kernel. This is a valid kernel by the reverse direction of Theorem 4.

Our algorithm looks exactly like Algorithm 2, only that now we use the new definitions of $\hat{\Psi}, \Psi$ above. To state the bound, recall that for any $\mathbf{u} = \sum_{i=1}^t a_t \Psi(\mathbf{x}_t)$ for some $a_1, \ldots, a_m \geq 0$, we define $\beta_{\mathbf{u}}$ to be the angle between $\sum_{t:a_t>0} a_t \Psi(\mathbf{x}_t)$ and $-\sum_{t:a_t<0} a_t \Psi(\mathbf{x}_t)$. The bound takes the following form.

Theorem 5 *Assume that the noise introduced into $\mathbf{x}_t$ has a known distribution $\mathcal{N}(0, \sigma^2 I)$, and that the noise introduced into y_t is independent with $\mathbb{E}[\tilde{y}_t^2] \leq B_{\tilde{y}}^2$. Suppose we use the kernel defined in Eq. (16), with some $c \in (0, d/4\sigma)$. Let $B_{\mathbf{w}} > 0, \beta \in (0, \pi/2]$ be fixed. Then if we run Algorithm 2 with parameters $B_{\mathbf{w}}$, any $W \geq \frac{\sqrt{5} B_{\mathbf{w}} \exp(\sigma c)}{\sin(\beta)}$, and $\eta =$*

$W/2\exp(\sigma c)\sqrt{(W^2\exp(2\sigma c)+B_{\tilde{y}}^2)T}$, *then for all* $\mathbf{u}$ *with* $\|\mathbf{u}\|\leq B_{\mathbf{w}},\beta_{\mathbf{u}}\geq\beta$, *it holds that*

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(\langle\mathbf{w}_t,\Psi(\mathbf{x}_t)\rangle,y_t)-\ell(\langle\mathbf{u},\Psi(\mathbf{x}_t)\rangle,y_t)\right]\leq 2W\exp(\sigma c)\sqrt{(W^2\exp(2\sigma c)+B_{\tilde{y}}^2)T}.$$

The proof follows the same lines as the proof of Theorem 3. The changes mostly have to do with the auxiliary lemmas from Subsection D.1, which we present below. The proof of the theorem itself is virtually identical to the one appearing in Subsection D.2, and is thus skipped.

E.1 Auxiliary Lemmas

The auxiliary lemmas below modify the parallel lemmas in Subsection D.1, based on the new definitions of the feature mapping Ψ and the surrogate feature mapping $\hat{\Psi}$. But before that, we begin with a lemma which explicitly upper bounds $\|\hat{\Psi}(\mathbf{x})\|^2$ for any $\mathbf{x}$. With Gaussian kernels, this was trivial, but now we need to work a bit harder.

Lemma 7 *For any vector* $\mathbf{x}\in\mathbb{R}^d$, *we have*

$$\left\|\hat{\Psi}(\mathbf{x})\right\|^2\leq\exp(2\sigma c).$$

Proof By Eq. (17),

$$\begin{aligned}\left\|\hat{\Psi}(\mathbf{x})\right\|^2&=\hat{k}(\mathbf{x},\mathbf{x})=d\int_{u=0}^{c/d}(1-2\sigma u)^{-d/2}\exp(-ud)du\\&\leq(1-2\sigma c/d)^{-d/2}\int_{u=0}^{c/d}d\exp(-ud)du=(1-2\sigma c/d)^{-d/2}(1-\exp(-c)).\end{aligned}\tag{18}$$

Also, by a Taylor expansion of the log function, and using the fact that $2\sigma c/d<1/2$ by the assumption that $c<d/4\sigma$, we get

$$\left(1-\frac{2\sigma c}{d}\right)^d=\exp\left(d\log\left(1-\frac{2\sigma c}{d}\right)\right)\geq\exp\left(d\left(-\frac{4\log(2)\sigma c}{d}\right)\right)=\exp(-4\log(2)\sigma c).\tag{19}$$

Plugging this into Eq. (18), we get the upper bound

$$\exp(2\log(2)\sigma c)(1-\exp(-c))\leq\exp(2\sigma c).$$

■

Lemma 8 *For any* $\mathbf{a},\mathbf{x}\in\mathbb{R}^d$, *if we let* $\tilde{\mathbf{x}}=\mathbf{x}+\mathbf{n}$ *where* $\mathbf{n}\sim\mathcal{N}(0,\Sigma)$ *is a Gaussian random vector with covariance matrix* Σ, *then it holds that*

$$\mathop{\mathbb{E}}_{\mathbf{n}}\left\langle\hat{\Psi}(\tilde{\mathbf{x}}),\hat{\Psi}(\mathbf{a})\right\rangle=\langle\Psi(\mathbf{x}),\Psi(\mathbf{a})\rangle.$$

Proof On one hand, based on the definition of k in Eq. (16), it can be verified that

$$\langle\Psi(\mathbf{x}),\Psi(\mathbf{a})\rangle=k(\mathbf{x},\mathbf{a})=\int_{u=0}^{c/d}d\exp\left(-u\|\mathbf{x}-\mathbf{a}\|^2-ud\right)du.\tag{20}$$

On the other hand, using the proof of Lemma 4 and Fubini's theorem, the expectation in the lemma can be written as

$$\mathop{\mathbb{E}}_{\mathbf{n}} \left\langle \hat{\Psi}(\tilde{\mathbf{x}}), \hat{\Psi}(\mathbf{a}) \right\rangle = \mathop{\mathbb{E}}_{\mathbf{n}} \left[d \int_{u=0}^{c/d} (1-2\sigma u)^{-d/2} \exp\left(-\frac{u \left\| \tilde{\mathbf{x}} - \mathbf{a} \right\|^2}{1-2\sigma u} - ud \right) du \right]$$

$$= \int_{u=0}^{c/d} d \exp(-ud) \mathop{\mathbb{E}}_{\mathbf{n}} \left[(1-2\sigma u)^{-d/2} \exp\left(-\frac{\left\| \tilde{\mathbf{x}} - \mathbf{a} \right\|^2}{1/u - 2\sigma} \right) \right] du$$

$$= \int_{u=0}^{c/d} d \exp(-ud) \exp\left(-\frac{\left\| \mathbf{x} - \mathbf{a} \right\|^2}{1/u} \right) du$$

$$= \int_{u=0}^{c/d} d \exp\left(-u \left\| \mathbf{x} - \mathbf{a} \right\|^2 - ud \right) du.$$

■

Lemma 9 *Let $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T$ be vectors in $\mathbb{R}^d$, and $a_1, a_2, \ldots, a_T$ scalars, such that $\left\| \sum_i a_i \Psi(\mathbf{z}_i) \right\|^2 \leq B^2$. Then $\sum_i a_i \hat{\Psi}(z_i)$ is an element in the RKHS w.r.t. $\hat{\Psi}(\cdot)$, whose norm squared is at most*

$$\frac{5B^2}{\sin^2(\beta)} \exp(2\sigma c).$$

Here, β is the angle between $\sum_{i:a_i>0} a_i \Psi(\mathbf{z}_i)$ and $-\sum_{i:a_i<0} a_i \Psi(\mathbf{z}_i)$ in the RKHS (or $\beta = \pi/2$ is one of the elements is zero).

Proof Picking some $\mathbf{z}_1, \mathbf{z}_2, \ldots$ and $a_1, a_2, \ldots$ as in the lemma statement, we have

$$\left\| \sum_i a_i \hat{\Psi}(z_i) \right\|^2 = \sum_{i,j} a_i a_j \hat{k}(\mathbf{z}_i, \mathbf{z}_j) \leq \sum_{i,j: a_i a_j > 0} a_i a_j \hat{k}(\mathbf{z}_i, \mathbf{z}_j) \tag{21}$$

Now, by definition of $\hat{k}(\cdot,\cdot)$ in Eq. (17), and the representation of $k(\cdot,\cdot)$ as in Eq. (20), it holds for any $\mathbf{z}_i, \mathbf{z}_j$ that

$$\frac{\hat{k}(\mathbf{z}_i, \mathbf{z}_j)}{k(\mathbf{z}_i, \mathbf{z}_j)} = \frac{d \int_{u=0}^{c/d} (1-2\sigma u)^{-d/2} \exp\left(-\frac{u \|\mathbf{z}_i - \mathbf{z}_j\|^2}{1-2\sigma u} - ud \right) du}{d \int_{u=0}^{c/d} \exp\left(-u \left\| \mathbf{z}_i - \mathbf{z}_j \right\|^2 - ud \right) du}$$

$$\leq (1 - 2\sigma c/d)^{-d/2} \frac{\int_{u=0}^{c/d} \exp\left(-\frac{u \|\mathbf{z}_i - \mathbf{z}_j\|^2}{1-2\sigma u} - ud \right) du}{\int_{u=0}^{c/d} \exp\left(-u \left\| \mathbf{z}_i - \mathbf{z}_j \right\|^2 - ud \right) du}$$

$$\leq (1 - 2\sigma c/d)^{-d/2} \leq \exp(2\sigma c),$$

where the last transition can be verified as in Eq. (19).

Therefore, we can upper bound Eq. (21) by

$$\sum_{i,j: a_i a_j > 0} a_i a_j \hat{k}(\mathbf{z}_i, \mathbf{z}_j) \leq \exp(2\sigma c) \sum_{i,j: a_i a_j > 0} a_i a_j k(\mathbf{z}_i, \mathbf{z}_j).$$

The lemma follows by noting that according to Lemma 5,

$$\sum_{i,j: a_i a_j > 0} a_i a_j k(\mathbf{z}_i, \mathbf{z}_j) = \left\| \sum_{i:a_i>0} a_i \Psi(\mathbf{z}_i) \right\|^2 + \left\| \sum_{i:a_i<0} \Psi(\mathbf{z}_i) \right\|^2 \leq \frac{5B^2}{\sin^2(\beta)}.$$

■