Learning Kernels with Random Features

Aman Sinha[1] **John Duchi**[1,2]
Departments of [1]Electrical Engineering and [2]Statistics
Stanford University
{amans,jduchi}@stanford.edu

Abstract

Randomized features provide a computationally efficient way to approximate kernel machines in machine learning tasks. However, such methods require a user-defined kernel as input. We extend the randomized-feature approach to the task of learning a kernel (via its associated random features). Specifically, we present an efficient optimization problem that learns a kernel in a supervised manner. We prove the consistency of the estimated kernel as well as generalization bounds for the class of estimators induced by the optimized kernel, and we experimentally evaluate our technique on several datasets. Our approach is efficient and highly scalable, and we attain competitive results with a fraction of the training cost of other techniques.

1 Introduction

An essential element of supervised learning systems is the representation of input data. Kernel methods [27] provide one approach to this problem: they implicitly transform the data to a new feature space, allowing non-linear data representations. This representation comes with a cost, as kernelized learning algorithms require time that grows at least quadratically in the data set size, and predictions with a kernelized procedure require the entire training set. This motivated Rahimi and Recht [24, 25] to develop randomized methods that efficiently approximate kernel evaluations with *explicit* feature transformations; this approach gives substantial computational benefits for large training sets and allows the use of simple linear models in the randomly constructed feature space.

Whether we use standard kernel methods or randomized approaches, using the "right" kernel for a problem can make the difference between learning a useful or useless model. Standard kernel methods as well as the aforementioned randomized-feature techniques assume the input of a user-defined kernel—a weakness if we do not *a priori* know a good data representation. To address this weakness, one often wishes to learn a good kernel, which requires substantial computation. We combine kernel learning with randomization, exploiting the computational advantages offered by randomized features to learn the kernel in a supervised manner. Specifically, we use a simple pre-processing stage for selecting our random features rather than jointly optimizing over the kernel and model parameters. Our workflow is straightforward: we create randomized features, solve a simple optimization problem to select a subset, then train a model with the optimized features. The procedure results in lower-dimensional models than the original random-feature approach for the same performance. We give empirical evidence supporting these claims and provide theoretical guarantees that our procedure is consistent with respect to the limits of infinite training data and infinite-dimensional random features.

1.1 Related work

To discuss related work, we first describe the supervised learning problem underlying our approach. We have a cost $c : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, where $c(\cdot, y)$ is convex for $y \in \mathcal{Y}$, and a reproducing kernel Hilbert space (RKHS) of functions $\mathcal{F}$ with kernel K. Given a sample $\{(x^i, y^i)\}_{i=1}^n$, the usual ℓ_2-regularized

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

learning problem is to solve the following (shown in primal and dual forms respectively):

$$\underset{f\in\mathcal{F}}{\text{minimize}} \quad \sum_{i=1}^{n} c(f(x^i), y^i) + \frac{\lambda}{2}\|f\|_2^2, \quad \text{or} \quad \underset{\alpha\in\mathbb{R}^n}{\text{maximize}} \quad -\sum_{i=1}^{n} c^*(\alpha_i, y^i) - \frac{1}{2\lambda}\alpha^T G\alpha, \quad (1)$$

where $\|\cdot\|_2$ denotes the Hilbert space norm, $c^*(\alpha, y) = \sup_z\{\alpha z - c(z, y)\}$ is the convex conjugate of c (for fixed y) and $G = [K(x^i, x^j)]_{i,j=1}^n$ denotes the Gram matrix.

Several researchers have studied kernel learning. As noted by Gönen and Alpaydın [14], most formulations fall into one of a few categories. In the supervised setting, one assumes a base class or classes of kernels and either uses heuristic rules to combine kernels [2, 23], optimizes structured (e.g. linear, nonnegative, convex) compositions of the kernels with respect to an alignment metric [9, 16, 20, 28], or jointly optimizes kernel compositions with empirical risk [17, 20, 29]. The latter approaches require an eigendecomposition of the Gram matrix or costly optimization problems (e.g. quadratic or semidefinite programs) [10, 14], but these models have a variety of generalization guarantees [1, 8, 10, 18, 19]. Bayesian variants of compositional kernel search also exist [12, 13]. In un- and semi-supervised settings, the goal is to learn an embedding of the input distribution followed by a simple classifier in the embedded space (e.g. [15]); the hope is that the input distribution carries the structure relevant to the task. Despite the current popularity of these techniques, especially deep neural architectures, they are costly, and it is difficult to provide guarantees on their performance.

Our approach optimizes kernel compositions with respect to an alignment metric, but rather than work with Gram matrices in the original data representation, we work with randomized feature maps that approximate RKHS embeddings. We learn a kernel that is structurally different from a user-supplied base kernel, and our method is an efficiently (near linear-time) solvable convex program.

2 Proposed approach

At a high level, we take a feature mapping, find a distribution that aligns this mapping with the labels y, and draw random features from the learned distribution; we then use these features in a standard supervised learning approach.

For simplicity, we focus on binary classification: we have n datapoints $(x^i, y^i) \in \mathbb{R}^d \times \{-1, 1\}$. Letting $\phi : \mathbb{R}^d \times \mathcal{W} \to [-1, 1]$ and Q be a probability measure on a space $\mathcal{W}$, define the kernel

$$K_Q(x, x') := \int \phi(x, w)\phi(x', w)dQ(w). \quad (2)$$

We want to find the "best" kernel K_Q over all distributions Q in some (large, nonparametric) set $\mathcal{P}$ of possible distributions on random features; we consider a kernel alignment problem of the form

$$\underset{Q\in\mathcal{P}}{\text{maximize}} \quad \sum_{i,j} K_Q(x^i, x^j)y^i y^j. \quad (3)$$

We focus on sets $\mathcal{P}$ defined by divergence measures on the space of probability distributions. For a convex function f with $f(1) = 0$, the f-divergence between distributions P and Q is $D_f(P\|Q) = \int f(\frac{dP}{dQ})dQ$. Then, for a base (user-defined) distribution P_0, we consider collections $\mathcal{P} := \{Q : D_f(Q\|P_0) \le \rho\}$ where $\rho > 0$ is a specified constant. In this paper, we focus on divergences $f(t) = t^k - 1$ for $k \ge 2$. Intuitively, the distribution Q maximizing the alignment (3) gives a feature space in which pairwise distances are similar to those in the output space $\mathcal{Y}$. Unfortunately, the problem (3) is generally intractable as it is infinite dimensional.

Using the randomized feature approach, we approximate the integral (2) as a discrete sum over samples $W^i \overset{\text{iid}}{\sim} P_0, i \in [N_w]$. Defining the discrete approximation $\mathcal{P}_{N_w} := \{q : D_f(q\|\mathbf{1}/N_w) \le \rho\}$ to $\mathcal{P}$, we have the following empirical version of problem (3):

$$\underset{q\in\mathcal{P}_{N_w}}{\text{maximize}} \quad \sum_{i,j} y^i y^j \sum_{m=1}^{N_w} q_m \phi(x^i, w^m)\phi(x^j, w^m). \quad (4)$$

Using randomized features, matching the input and output distances in problem (4) translates to finding a (weighted) set of points among $w^1, w^2, ..., w^{N_w}$ that best "describe" the underlying dataset, or, more directly, finding weights q so that the kernel matrix matches the correlation matrix yy^T.

Given a solution $\widehat{q}$ to problem (4), we can solve the primal form of problem (1) in two ways. First, we can apply the Rahimi and Recht [24] approach by drawing D samples $W^1, \ldots, W^D \overset{\text{iid}}{\sim} \widehat{q}$, defining features $\phi^i = [\phi(x^i, w^1) \ \cdots \ \phi(x^i, w^D)]^T$, and solving the risk minimization problem

$$\widehat{\theta} = \operatorname*{argmin}_{\theta} \left\{ \sum_{i=1}^{n} c\left(\tfrac{1}{\sqrt{D}} \theta^T \phi^i, y^i \right) + r(\theta) \right\} \tag{5}$$

for some regularization r. Alternatively, we may set $\phi^i = [\phi(x^i, w^1) \ \cdots \ \phi(x^i, w^{N_w})]^T$, where $w^1, \ldots, w^{N_w}$ are the original random samples from P_0 used to solve (4), and directly solve

$$\widehat{\theta} = \operatorname*{argmin}_{\theta} \left\{ \sum_{i=1}^{n} c(\theta^T \operatorname{diag}(\widehat{q})^{\frac{1}{2}} \phi^i, y^i) + r(\theta) \right\}. \tag{6}$$

Notably, if $\widehat{q}$ is sparse, the problem (6) need only store the random features corresponding to non-zero entries of $\widehat{q}$. Contrast our two-phase procedure to that of Rahimi and Recht [25], which samples $W^1, \ldots, W^D \overset{\text{iid}}{\sim} P_0$ and solves the minimization problem

$$\underset{\alpha \in \mathbb{R}^{N_w}}{\text{minimize}} \quad \sum_{i=1}^{n} c\left(\sum_{m=1}^{D} \alpha_m \phi(x^i, w^m), y^i \right) \ \text{ subject to } \ \|\alpha\|_\infty \leq C/N_w, \tag{7}$$

where C is a numerical constant. At first glance, it appears that we may suffer both in terms of computational efficiency and in classification or learning performance compared to the one-step procedure (7). However, as we show in the sequel, the alignment problem (4) can be solved very efficiently and often yields sparse vectors $\widehat{q}$, thus substantially decreasing the dimensionality of problem (6). Additionally, we give experimental evidence in Section 4 that the two-phase procedure yields generalization performance similar to standard kernel and randomized feature methods.

2.1 Efficiently solving problem (4)

The optimization problem (4) has structure that enables efficient (near linear-time) solutions. Define the matrix $\Phi = [\phi^1 \ \cdots \ \phi^n] \in \mathbb{R}^{N_w \times n}$, where $\phi^i = [\phi(x^i, w^1) \ \cdots \ \phi(x^i, w^{N_w})]^T \in \mathbb{R}^{N_w}$ is the randomized feature representation for x^i and $w^m \overset{\text{iid}}{\sim} P_0$. We can rewrite the optimization objective as

$$\sum_{i,j} y^i y^j \sum_{m=1}^{N_w} q_m \phi(x^i, w^m) \phi(x^j, w^m) = \sum_{m=1}^{N_w} q_m \left(\sum_{i=1}^{n} y^i \phi(x^i, w^m) \right)^2 = q^T \left((\Phi y) \odot (\Phi y) \right),$$

where $\odot$ denotes the Hadamard product. Constructing the linear objective requires the evaluation of Φy. Assuming that the computation of ϕ is $O(d)$, construction of Φ is $O(nN_w d)$ on a single processor. However, this construction is trivially parallelizable. Furthermore, computation can be sped up even further for certain distributions P_0. For example, the Fastfood technique can approximate Φ in $O(nN_w \log(d))$ time for the Gaussian kernel [21].

The problem (4) is also efficiently solvable via bisection over a scalar dual variable. Using $\lambda \geq 0$ for the constraint $D_f(Q \| P_0) \leq \rho$, a partial Lagrangian is

$$\mathcal{L}(q, \lambda) = q^T \left((\Phi y) \odot (\Phi y) \right) - \lambda \left(D_f(q \| \mathbf{1}/N_w) - \rho \right).$$

The corresponding dual function is $g(\lambda) = \sup_{q \in \Delta} \mathcal{L}(q, \lambda)$, where $\Delta := \{q \in \mathbb{R}_+^{N_w} : q^T \mathbf{1} = 1\}$ is the probability simplex. Minimizing $g(\lambda)$ yields the solution to problem (4); this is a convex optimization problem in one dimension so we can use bisection. The computationally expensive step in each iteration is maximizing $\mathcal{L}(q, \lambda)$ with respect to q for a given λ. For $f(t) = t^k - 1$, we define $v := (\Phi y) \odot (\Phi y)$ and solve

$$\underset{q \in \Delta}{\text{maximize}} \ q^T v - \lambda \frac{1}{N_w} \sum_{m=1}^{N_w} (N_w q_m)^k. \tag{8}$$

This has a solution of the form $q_m = \left[v_m / \lambda N_w^{k-1} + \tau \right]_+^{\frac{1}{k-1}}$, where τ is chosen so that $\sum_m q_m = 1$. We can find such a τ by a variant of median-based search in $O(N_w)$ time [11]. Thus, for any $k \geq 2$, an ϵ-suboptimal solution to problem (4) can be found in $O(N_w \log(1/\epsilon))$ time (see Algorithm 1).

Algorithm 1 Kernel optimization with $f(t) = t^k - 1$ as divergence

INPUT: distribution P_0 on $\mathcal{W}$, sample $\{(x^i, y^i)\}_{i=1}^n$, $N_w \in \mathbb{N}$, feature function ϕ, $\epsilon > 0$
OUTPUT: $q \in \mathbb{R}^{N_w}$ that is an ϵ-suboptimal solution to (4).
SETUP: Draw N_w samples $w^m \overset{\text{iid}}{\sim} P_0$, build feature matrix Φ, compute $v := (\Phi y) \odot (\Phi y)$.
Set $\lambda_u \leftarrow \infty, \;\; \lambda_l \leftarrow 0, \;\; \lambda_s \leftarrow 1$
while $\lambda_u = \infty$
  $q \leftarrow \operatorname{argmax}_{q \in \Delta} \mathcal{L}(q, \lambda_s)$ *// (solution to problem (8))*
  if $D_f\left(q \| \mathbf{1}/N_w\right) < \rho$ **then** $\lambda_u \leftarrow \lambda_s$ **else** $\lambda_s \leftarrow 2\lambda_s$
while $\lambda_u - \lambda_l > \epsilon \lambda_s$
  $\lambda \leftarrow (\lambda_u + \lambda_l)/2$
  $q \leftarrow \operatorname{argmax}_{q \in \Delta} \mathcal{L}(q, \lambda)$ *// (solution to problem (8))*
  if $D_f\left(q \| \mathbf{1}/N_w\right) < \rho$ **then** $\lambda_u \leftarrow \lambda$ **else** $\lambda_l \leftarrow \lambda$

3 Consistency and generalization performance guarantees

Although the procedure (4) is a discrete approximation to a heuristic kernel alignment problem, we can provide guarantees on its performance as well as the generalization performance of our subsequent model trained with the optimized kernel.

Consistency First, we provide guarantees that the solution to problem (4) approaches a population optimum as the data and random sampling increase ($n \to \infty$ and $N_w \to \infty$, respectively). We consider the following (slightly more general) setting: let $S : \mathcal{X} \times \mathcal{X} \to [-1, 1]$ be a bounded function, where we intuitively think of $S(x, x')$ as a similarity metric between labels for x and x', and denote $S_{ij} := S(x^i, x^j)$ (in the binary case with $y \in \{-1, 1\}$, we have $S_{ij} = y^i y^j$). We then define the alignment functions

$$T(P) := \mathbb{E}[S(X, X') K_P(X, X')], \qquad \widehat{T}(P) := \frac{1}{n(n-1)} \sum_{i \neq j} S_{ij} K_P(x^i, x^j),$$

where the expectation is taken over S and the independent variables X, X'. Lemmas 1 and 2 provide consistency guarantees with respect to the data sample (x^i and S_{ij}) and the random feature sample (w^m); together they give us the overall consistency result of Theorem 1. We provide proofs in the supplement (Sections A.1, A.2, and A.3 respectively).

Lemma 1 (Consistency with respect to data)**.** *Let $f(t) = t^k - 1$ for $k \geq 2$. Let P_0 be any distribution on the space $\mathcal{W}$, and let $\mathcal{P} = \{Q : D_f\left(Q \| P_0\right) \leq \rho\}$. Then*

$$\mathbb{P}\left(\sup_{Q \in \mathcal{P}} \left| \widehat{T}(Q) - T(Q) \right| \geq t \right) \leq \sqrt{2} \exp\left(-\frac{n t^2}{16(1+\rho)} \right).$$

Lemma 1 shows that the empirical quantity $\widehat{T}$ is close to the true T. Now we show that, independent of the size of the training data, we can consistently estimate the optimal $Q \in \mathcal{P}$ via sampling (i.e. $Q \in \mathcal{P}_{N_w}$).

Lemma 2 (Consistency with respect to sampling features)**.** *Let the conditions of Lemma 1 hold. Then, with $C_\rho = \frac{2(\rho+1)}{\sqrt{1+\rho}-1}$ and $D_\rho = \sqrt{8(1+\rho)}$, we have*

$$\left| \sup_{Q \in \mathcal{P}_{N_w}} \widehat{T}(Q) - \sup_{Q \in \mathcal{P}} \widehat{T}(Q) \right| \leq 4 C_\rho \sqrt{\frac{\log(2N_w)}{N_w}} + D_\rho \sqrt{\frac{\log \frac{2}{\delta}}{N_w}}$$

with probability at least $1 - \delta$ over the draw of the samples $W^m \overset{\text{iid}}{\sim} P_0$.

Finally, we combine the consistency guarantees for data and sampling to reach our main result, which shows that the alignment provided by the estimated distribution $\widehat{Q}$ is nearly optimal.

Theorem 1. *Let $\widehat{Q}_w$ maximize $\widehat{T}(Q)$ over $Q \in \mathcal{P}_{N_w}$. Then, with probability at least $1 - 3\delta$ over the sampling of both (x, y) and W, we have*

$$\left| T(\widehat{Q}_w) - \sup_{Q \in \mathcal{P}} T(Q) \right| \leq 4 C_\rho \sqrt{\frac{\log(2N_w)}{N_w}} + D_\rho \sqrt{\frac{\log \frac{2}{\delta}}{N_w}} + 2 D_\rho \sqrt{\frac{2 \log \frac{2}{\delta}}{n}}.$$

Generalization performance The consistency results above show that our optimization procedure nearly maximizes alignment $T(P)$, but they say little about generalization performance for our model trained using the optimized kernel. We now show that the class of estimators employed by our method has strong performance guarantees. By construction, our estimator (6) uses the function class

$$\mathcal{F}_{N_w} := \Big\{ h(x) = \sum_{m=1}^{N_w} \alpha_m \sqrt{q_m} \phi(x, w^m) \mid q \in \mathcal{P}_{N_w}, \|\alpha\|_2 \leq B \Big\},$$

and we provide bounds on its generalization via empirical Rademacher complexity. To that end, define $\mathcal{R}_n(\mathcal{F}_{N_w}) := \frac{1}{n}\mathbb{E}[\sup_{f \in \mathcal{F}_{N_w}} \sum_{i=1}^n \sigma_i f(x^i)]$, where the expectation is taken over the i.i.d. Rademacher variables $\sigma_i \in \{-1, 1\}$. We have the following lemma, whose proof is in Section A.4.

Lemma 3. *Under the conditions of the preceding paragraph,* $\mathcal{R}_n(\mathcal{F}_{N_w}) \leq B\sqrt{\frac{2(1+\rho)}{n}}$.

Applying standard concentration results, we obtain the following generalization guarantee.

Theorem 2 ([8, 18])**.** *Let the true misclassification risk and ν-empirical misclassification risk for an estimator h be defined as follows:*

$$R(h) := \mathbb{P}(Yh(X) < 0), \qquad \widehat{R}_\nu(h) := \frac{1}{n}\sum_{i=1}^n \min\left\{1, \left[1 - yh(x^i)/\nu\right]_+\right\}.$$

Then $\sup_{h \in \mathcal{F}_{N_w}} \{R(h) - \widehat{R}_\nu(h)\} \leq \frac{2}{\nu}\mathcal{R}_n(\mathcal{F}_{N_w}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$ *with probability at least* $1 - \delta$.

The bound is independent of the number of terms N_w, though in practice we let B grow with N_w.

4 Empirical evaluations

We now turn to empirical evaluations, comparing our approach's predictive performance with that of Rahimi and Recht's randomized features [24] as well as a joint optimization over kernel compositions and empirical risk. In each of our experiments, we investigate the effect of increasing dimensionality of the randomized feature space D. For our approach, we use the χ^2-divergence ($k = 2$ or $f(t) = t^2 - 1$). Letting $\widehat{q}$ denote the solution to problem (4), we use two variants of our approach: when $D < \mathrm{nnz}(\widehat{q})$ we use estimator (5), and we use estimator (6) otherwise. For the original randomized feature approach, we relax the constraint in problem (7) with an ℓ_2 penalty. Finally, for the joint optimization in which we learn the kernel and classifier together, we consider the kernel-learning objective, i.e. finding the best Gram matrix G in problem (1) for the soft-margin SVM [14]:

$$\begin{array}{lll} \text{minimize}_{q \in \mathcal{P}_{N_w}} & \sup_\alpha & \alpha^T \mathbf{1} - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y^i y^j \sum_{m=1}^{N_w} q_m \phi(x^i, w^m)\phi(x^j, w^m) \\ & \text{subject to} & \mathbf{0} \preceq \alpha \preceq C\mathbf{1}, \;\; \alpha^T y = 0. \end{array} \tag{9}$$

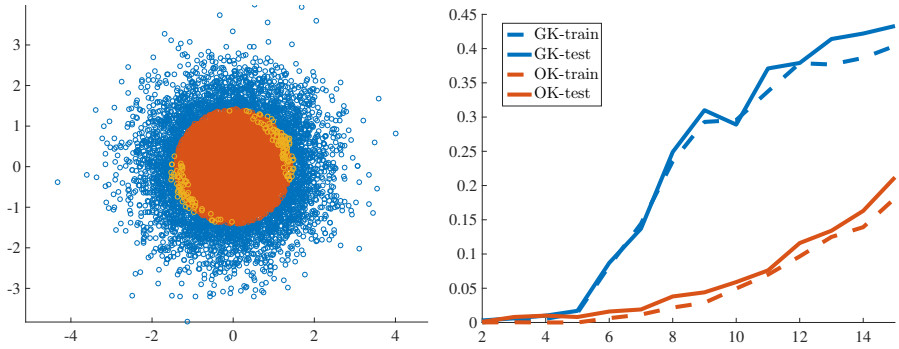
We use a standard primal-dual algorithm [4] to solve the min-max problem (9). While this is an expensive optimization, it is a convex problem and is solvable in polynomial time.

In Section 4.1, we visualize a particular problem that illustrates the effectiveness of our approach when the user-defined kernel is poor. Section 4.2 shows how learning the kernel can be used to quickly find a sparse set of features in high dimensional data, and Section 4.3 compares our performance with unoptimized random features and the joint procedure (9) on benchmark datasets. The supplement contains more experimental results in Section C.

4.1 Learning a new kernel with a poor choice of P_0

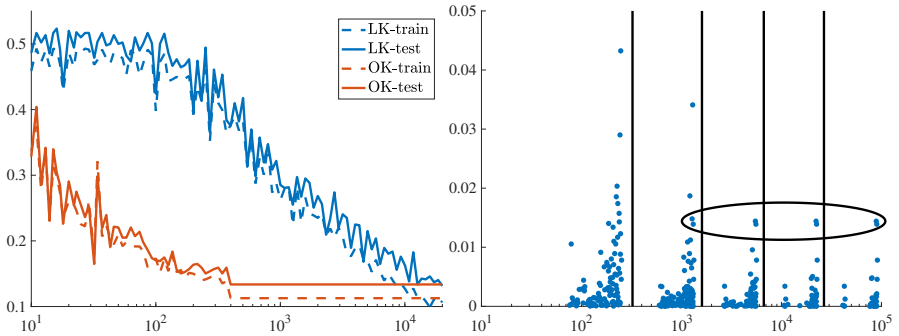
For our first experiment, we generate synthetic data $x^i \overset{\text{iid}}{\sim} \mathsf{N}(0, I)$ with labels $y^i = \mathrm{sign}(\|x\|_2 - \sqrt{d})$, where $x \in \mathbb{R}^d$. The Gaussian kernel is ill-suited for this task, as the Euclidean distance used in this kernel does not capture the underlying structure of the classes. Nevertheless, we use the Gaussian kernel, which corresponds [24] to $\phi(x, (w, v)) = \cos((x, 1)^T(w, v))$ where $(W, V) \sim \mathsf{N}(0, I) \times \mathsf{Uni}(0, 2\pi)$, to showcase the effects of our method. We consider a training set of size $n = 10^4$ and a test set of size 10^3, and we employ logistic regression with $D = \mathrm{nnz}(\widehat{q})$ for both our technique as well as the original random feature approach.[1]

[1] For $2 \leq d \leq 15$, $\mathrm{nnz}(\widehat{q}) < 250$ when the kernel is trained with $N_w = 2 \cdot 10^4$, and $\rho = 200$.



(a) Training data & optimized features for $d = 2$ (b) Error vs. d

Figure 1. Experiments with synthetic data. (a) Positive and negative training examples are blue and red, and optimized randomized features (w^m) are yellow. All offset parameters v^m were optimized to be near 0 or π (not shown). (b) Misclassification error of logistic regression model vs. dimensionality of data. GK denotes random features with a Gaussian kernel, and our optimized kernel is denoted OK.



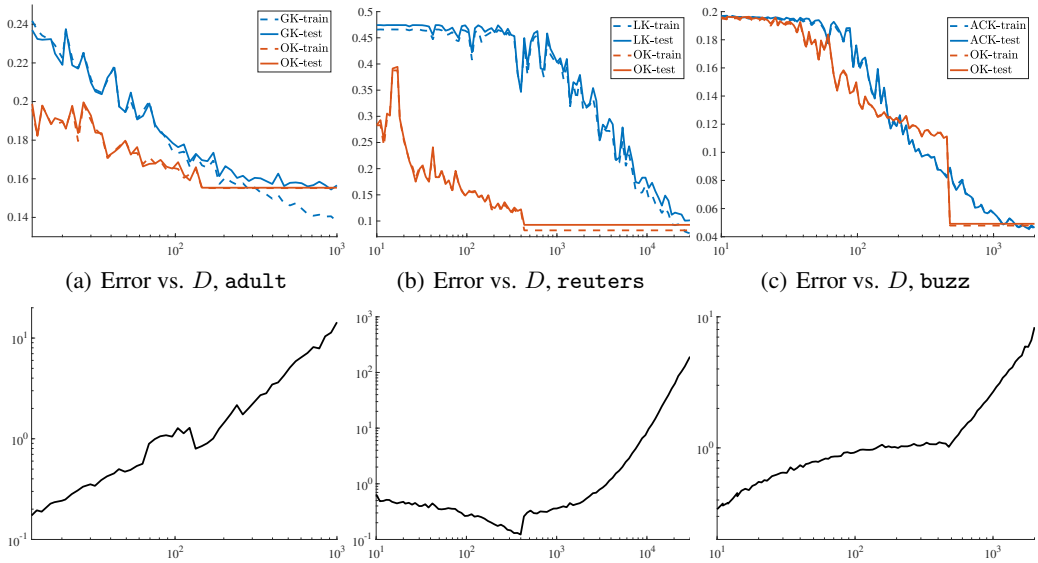
(a) Error vs. D (b) $\widehat{q}_i$ vs. i

Figure 2. Feature selection in sparse data. (a) Misclassification error of ridge regression model vs. dimensionality of data. LK denotes random features with a linear kernel, and OK denotes our method. Our error is fixed above $D = \mathrm{nnz}(\widehat{q})$ after which we employ estimator (6). (b) Weight of feature i in optimized kernel (q_i) vs. i. Vertical bars delineate separations between k-grams, where $1 \le k \le 5$ is nondecreasing in i. Circled features are prefixes of `GGTTG` and `GTTGG` at indices 60–64.

Figure 1 shows the results of the experiments for $d \in \{2, \ldots, 15\}$. Figure 1(a) illustrates the output of the optimization when $d = 2$. The selected kernel features w^m lie near $(1, 1)$ and $(-1, -1)$; the offsets v^m are near 0 and π, giving the feature $\phi(\cdot, w, v)$ a parity flip. Thus, the kernel computes similarity between datapoints via neighborhoods of $(1, 1)$ and $(-1, -1)$ close to the classification boundary. In higher dimensions, this generalizes to neighborhoods of pairs of opposing points along the surface of the d-sphere; these features provide a coarse approximation to vector magnitude. Performance degradation with d occurs because the neighborhoods grow exponentially larger and less dense (due to fixed N_w and n). Nevertheless, as shown in Figure 1(b), this degradation occurs much more slowly than that of the Gaussian kernel, which suffers a similar curse of dimensionality due to its dependence on Euclidean distance. Although somewhat contrived, this example shows that even in situations with poor base kernels our approach learns a more suitable representation.

4.2 Feature selection and biological sequences

In addition to the computational advantages rendered by the sparsity of q after performing the optimization (4), we can use this sparsity to gain insights about important features in high-dimensional datasets; this can act as an efficient filtering mechanism before further investigation. We present one example of this task, studying an aptamer selection problem [6]. In this task, we are given $n = 2900$ nucleotide sequences (aptamers) $x^i \in \mathcal{A}^{81}$, where $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and labels y^i indicate (thresholded) binding affinity of the aptamer to a molecular target. We create one-hot encoded forms of k-grams of the sequence, where $1 \le k \le 5$, resulting in $d = \sum_{k=1}^{5} |\mathcal{A}|^k (82 - k) = 105{,}476$



(a) Error vs. D, `adult` (b) Error vs. D, `reuters` (c) Error vs. D, `buzz`

(d) Speedup vs. D, `adult` (e) Speedup vs. D, `reuters` (f) Speedup vs. D, `buzz`

Figure 3. Performance analysis on benchmark datasets. The top row shows training and test misclassification rates. Our method is denoted as OK and is shown in red. The blue methods are random features with Gaussian, linear, or arc-cosine kernels (GK, LK, or ACK respectively). Our error and running time become fixed above $D = \text{nnz}(\widehat{q})$ after which we employ estimator (6). The bottom row shows the speedup factor of using our method over regular random features (speedup $= x$ indicates our method takes $1/x$ of the time required to use regular random features). Our method is faster at moderate to large D and shows better performance than the random feature approach at small to moderate D.

Table 1: Best test results over benchmark datasets

Dataset	n,	n_{test}	d	Model	Our error (%), time(s)		Random error (%), time(s)	
`adult`	32561,	16281	123	Logistic	15.54,	3.6	15.44,	43.1
`reuters`	23149,	781265	47236	Ridge	9.27,	0.8	9.36,	295.9
`buzz`	105530,	35177	77	Ridge	4.92,	2.0	4.58,	11.9

features. We consider the linear kernel, i.e. $\phi(x, w) = x_w$, where $w \sim \text{Uni}(\{1, \ldots, d\})$. Figure 2(a) compares the misclassification error of our method with that of random k-gram features, while Figure 2(b) indicates the weights q_i given to features by our method. In under 0.2 seconds, we whittle down the original feature space to 379 important features. By restricting random selection to just these features, we outperform the approach of selecting features uniformly at random when $D \ll d$. More importantly, however, we can derive insights from this selection. For example, the circled features in Figure 2(b) correspond to k-gram prefixes for the 5-grams `GGTTG` and `GTTGG` at indices 60 through 64; `G`-complexes are known to be relevant for binding affinities in aptamers [6], so this is reasonable.

4.3 Performance on benchmark datasets

We now show the benefits of our approach on large-scale datasets, since we exploit the efficiency of random features with the performance of kernel-learning techniques. We perform experiments on three distinct types of datasets, tracking training/test error rates as well as total (training + test) time. For the `adult`[2] dataset we employ the Gaussian kernel with a logistic regression model, and for the `reuters`[3] dataset we employ a linear kernel with a ridge regression model. For the `buzz`[4] dataset we employ ridge regression with an arc-cosine kernel of order 2, i.e. $P_0 = \mathcal{N}(0, I)$ and $\phi(x, w) = H(w^T x)(w^T x)^2$, where $H(\cdot)$ is the Heavyside step function [7].

[2]https://archive.ics.uci.edu/ml/datasets/Adult

[3]http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm. We consider predicting whether a document has a CCAT label.

[4]http://ama.liglab.fr/data/buzz/classification/. We use the Twitter dataset.

Table 2: Comparisons with joint optimization on subsampled data

Dataset	Our training / test error (%), time(s)			Joint training / test error (%), time(s)		
`adult`	16.22	/ 16.36,	1.8	14.88	/ 16.31,	198.1
`reuters`	7.64	/ 9.66,	0.6	6.30	/ 8.96,	173.3
`buzz`	8.44	/ 8.32,	0.4	7.38	/ 7.08,	137.5

Comparison with unoptimized random features Results comparing our method with unoptimized random features are shown in Figure 3 for many values of D, and Table 1 tabulates the best test error and corresponding time for the methods. Our method outperforms the original random feature approach in terms of generalization error for small and moderate values of D; at very large D the random feature approach either matches our surpasses our performance. The trends in speedup are opposite: our method requires extra optimizations that dominate training time at extremely small D; at very large D we use estimator (6), so our method requires less overall time. The nonmonotonic behavior for `reuters` (Figure 3(e)) occurs due to the following: at $D \lesssim \mathrm{nnz}(\widehat{q})$, sampling indices from the optimized distribution takes a non-neglible fraction of total time, and solving the linear system requires more time when rows of Φ are not unique (due to sampling).

Performance improvements also depend on the kernel choice for a dataset. Namely, our method provides the most improvement, in terms of training time for a given amount of generalization error, over random features generated for the linear kernel on the `reuters` dataset; we are able to surpass the best results of the random feature approach 2 orders of magnitude faster. This makes sense when considering the ability of our method to sample from a small subset of important features. On the other hand, random features for the arc-cosine kernel are able to achieve excellent results on the `buzz` dataset even without optimization, so our approach only offers modest improvement at small to moderate D. For the Gaussian kernel employed on the `adult` dataset, our method is able to achieve the same generalization performance as random features in roughly $1/12$ the training time.

Thus, we see that our optimization approach generally achieves competitive results with random features at lower computational costs, and it offers the most improvements when either the base kernel is not well-suited to the data or requires a large number of random features (large D) for good performance. In other words, our method reduces the sensitivity of model performance to the user's selection of base kernels.

Comparison with joint optimization Despite the fact that we do not choose empirical risk as our objective in optimizing kernel compositions, our optimized kernel enjoys competitive generalization performance compared to the joint optimization procedure (9). Because the joint optimization is very costly, we consider subsampled training datasets of 5000 training examples. Results are shown in Table 2, where it is evident that the efficiency of our method outweighs the marginal gain in classification performance for joint optimization.

5 Conclusion

We have developed a method to learn a kernel in a supervised manner using random features. Although we consider a kernel alignment problem similar to other approaches in the literature, we exploit computational advantages offered by random features to develop a much more efficient and scalable optimization procedure. Our concentration bounds guarantee the results of our optimization procedure closely match the limits of infinite data ($n \to \infty$) and sampling ($N_w \to \infty$), and our method produces models that enjoy good generalization performance guarantees. Empirical evaluations indicate that our optimized kernels indeed "learn" structure from data, and we attain competitive results on benchmark datasets at a fraction of the training time for other methods. Generalizing the theoretical results for concentration and risk to other $f-$divergences is the subject of further research. More broadly, our approach opens exciting questions regarding the usefulness of simple optimizations on random features in speeding up other traditionally expensive learning problems.

Acknowledgements This research was supported by a Fannie & John Hertz Foundation Fellowship and a Stanford Graduate Fellowship.

References

[1] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

[2] A. Ben-Hur and W. S. Noble. Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(suppl 1):i38–i46, 2005.

[3] A. Ben-Tal, D. den Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[4] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[5] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: a Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[6] M. Cho, S. S. Oh, J. Nie, R. Stewart, M. Eisenstein, J. Chambers, J. D. Marth, F. Walker, J. A. Thomson, and H. T. Soh. Quantitative selection and parallel characterization of aptamers. *Proceedings of the National Academy of Sciences*, 110(46), 2013.

[7] Y. Cho and L. K. Saul. Kernel methods for deep learning. In *Advances in neural information processing systems*, pages 342–350, 2009.

[8] C. Cortes, M. Mohri, and A. Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 247–254, 2010.

[9] C. Cortes, M. Mohri, and A. Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828, 2012.

[10] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor. On kernel target alignment. In *Innovations in Machine Learning*, pages 205–256. Springer, 2006.

[11] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[12] D. Duvenaud, J. R. Lloyd, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint arXiv:1302.4922*, 2013.

[13] M. Girolami and S. Rogers. Hierarchic bayesian models for kernel learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 241–248. ACM, 2005.

[14] M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

[15] G. E. Hinton and R. R. Salakhutdinov. Using deep belief nets to learn covariance kernels for gaussian processes. In *Advances in neural information processing systems*, pages 1249–1256, 2008.

[16] J. Kandola, J. Shawe-Taylor, and N. Cristianini. Optimizing kernel alignment over combinations of kernel. 2002.

[17] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. Lp-norm multiple kernel learning. *The Journal of Machine Learning Research*, 12:953–997, 2011.

[18] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, pages 1–50, 2002.

[19] V. Koltchinskii, D. Panchenko, et al. Complexities of convex combinations and bounding the generalization error in classification. *The Annals of Statistics*, 33(4):1455–1496, 2005.

[20] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

[21] Q. Le, T. Sarlós, and A. Smola. Fastfood-computing hilbert space expansions in loglinear time. In *Proceedings of the 30th International Conference on Machine Learning*, pages 244–252, 2013.

[22] D. Luenberger. *Optimization by Vector Space Methods*. Wiley, 1969.

[23] S. Qiu and T. Lane. A framework for multiple kernel support vector regression and its applications to sirna efficacy prediction. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 6(2): 190–199, 2009.

[24] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, 2007.

[25] A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems 21*, 2008.

[26] P. Samson. Concentration of measure inequalities for Markov chains and ϕ-mixing processes. *Annals of Probability*, 28(1):416–461, 2000.

[27] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[28] Y. Ying, K. Huang, and C. Campbell. Enhanced protein fold recognition through a novel data integration approach. *BMC bioinformatics*, 10(1):1, 2009.

[29] A. Zien and C. S. Ong. Multiclass multiple kernel learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1191–1198. ACM, 2007.

A Proofs of major results

Before proving our results, we provide a few technical lemmas to which we refer in the sequel, and we also give a few definitions. The first is the standard definition of sub-Gaussian random variables.

Definition 1. *A random variable X is σ^2-sub-Gaussian if*

$$\mathbb{E}\left[\exp(\lambda(X - \mathbb{E}[X]))\right] \le \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$

for all $\lambda \in \mathbb{R}$.

We enumerate a few standard consequences of sub-Gaussianity [5]. If X_i are independent and σ^2-sub-Gaussian, then $\sum_{i=1}^n X_i$ is $n\sigma^2$-sub-Gaussian. Moreover, we have the standard concentration guarantee

$$\max\{\mathbb{P}(X \ge \mathbb{E}[X] + t), \mathbb{P}(X \le \mathbb{E}[X] - t)\} \le 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

for all $t \ge 0$ if X is σ^2-sub-Gaussian, and if there are bounds $a \le X \le b$, then X is $\frac{(b-a)^2}{4}$-sub-Gaussian. Moreover, if X is mean-zero and σ^2-sub-Gaussian, then

$$\mathbb{E}\left[\exp(\lambda X^2)\right] \le \frac{1}{\left[1 - 2\lambda\sigma^2\right]_+^{\frac{1}{2}}} = \exp\left(-\frac{1}{2}\log\left[1 - 2\lambda\sigma^2\right]_+\right). \tag{10}$$

Throughout our proofs, for a given $k \in [1, \infty]$, we use $k_* = \frac{k}{k-1}$, so that $1/k + 1/k_* = 1$, to denote the conjugate to k.

The technical lemmas that we shall need follow. The first is an essentially standard duality result.

Lemma 4 (Ben-Tal et al. [3])**.** *Let f be any closed convex function with domain $\operatorname{dom} f \subset [0, \infty)$, and let $f^*(s) = \sup_{t \ge 0}\{ts - f(t)\}$ be its conjugate. Then for any distribution P and any function $g : \mathcal{W} \to \mathbb{R}$ we have*

$$\sup_{Q : D_f(Q\|P) \le \rho} \int g(w) dQ(w) = \inf_{\lambda \ge 0, \eta}\left\{\lambda \int f^*\left(\frac{g(w) - \eta}{\lambda}\right) dP(w) + \rho\lambda + \eta\right\}.$$

See Section B.1 for a proof of this lemma. Note that as an immediate consequence of this result, we have an expectation upper bound on empirical versions of $\sup_{Q : D_f(Q\|P) \le \rho} \int g(w) dQ(w)$. Indeed, let $Z_1, \ldots, Z_{N_w}$ be drawn i.i.d. from a base distribution P_0. To simplify algebra, we work with a scaled version of the f-divergence: $f(t) = \frac{1}{k}(t^k - 1)$, so the population and empirical constraint sets we consider are defined by

$$\mathcal{P} = \left\{Q : D_f\left(Q\|P_0\right) \le \frac{\rho}{k}\right\} \text{ and } \mathcal{P}_{N_w} := \left\{q : D_f\left(q\|\mathbf{1}/N_w\right) \le \frac{\rho}{k}\right\}.$$

Then by Lemma 4, we obtain

$$\begin{aligned}
\mathbb{E}\left[\sup_{Q \in \mathcal{P}_{N_w}} \mathbb{E}_Q[Z]\right] &= \mathbb{E}_{P_0}\left[\inf_{\lambda \ge 0, \eta} \frac{1}{N}\sum_{i=1}^N \lambda f^*\left(\frac{Z_i - \eta}{\lambda}\right) + \eta + \frac{\rho}{k}\lambda\right] \\
&\le \inf_{\lambda \ge 0, \eta} \mathbb{E}_{P_0}\left[\frac{1}{N}\sum_{i=1}^N \lambda f^*\left(\frac{Z_i - \eta}{\lambda}\right) + \eta + \frac{\rho}{k}\lambda\right] \\
&= \inf_{\lambda \ge 0, \eta}\left\{\mathbb{E}_{P_0}\left[\lambda f^*\left(\frac{Z - \eta}{\lambda}\right)\right] + \frac{\rho}{k}\lambda + \eta\right\} \\
&= \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[Z].
\end{aligned} \tag{11}$$

The second lemma provides a lower bound on the expectation of certain robust quantities, and we provide a proof of the lemma in Section B.2.

Lemma 5. *Let $Z = (Z_1, \ldots, Z_{N_w})$ be a random vector of independent random variables $Z_i \stackrel{\text{iid}}{\sim} P_0$, where $|Z_i| \leq M$ with probability 1. Let $k \in [2, \infty]$ and define $C_{\rho,k} = \frac{2(1+\rho)}{(1+\rho)^{\frac{1}{k_*}} - 1} \leq C_\rho = \frac{2(\rho+1)}{\sqrt{1+\rho}-1}$. Let $f(t) = \frac{1}{k}(t^k - 1)$. Then*

$$\mathbb{E}\left[\sup_{Q \in \mathcal{P}_{N_w}} \mathbb{E}_Q[Z]\right] \geq \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[Z] - 4C_\rho M \sqrt{\frac{\log(2N_w)}{N_w}}$$

and

$$\mathbb{E}\left[\sup_{Q \in \mathcal{P}_{N_w}} \mathbb{E}_Q[Z]\right] \leq \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[Z].$$

A.1 Proof of Lemma 1

The result follows from a dual formulation of the expression on the left hand side as well as standard concentration results for sub-Gaussian random variables. Define

$$\widehat{e}_n(w) := \frac{1}{n(n-1)} \sum_{i \neq j} S_{ij} \phi(x^i, w) \phi(x^j, w) - \mathbb{E}[S(X, X') \phi(X, w) \phi(X', w)] \tag{12}$$

to be the error in the kernel estimate at the kernel parameter w. We give our argument by duality, noting that the lemma is equivalent to proving

$$\mathbb{P}\left(\sup_{Q \in \mathcal{P}} \left| \int \widehat{e}_n(w) dQ(w) \right| \geq t\right) \leq \sqrt{2} \exp\left(-\frac{nt^2}{16(\rho+1)}\right).$$

Before continuing, we note the following useful result, whose proof we provide in Section B.3.

Lemma 6. *For each fixed w, the random variable $\widehat{e}_n(w)$ is mean-zero and $\frac{4}{n}$-sub-Gaussian.*

To simplify the algebra, we work with a scaled version of the f-divergence: $f(t) = \frac{1}{k}(t^k - 1)$, so the equivalent constraint sets are $\mathcal{P} := \left\{Q : D_f\left(Q \| P_0\right) \leq \frac{\rho}{k}\right\}$ and $\mathcal{P}_{N_w} := \{q : D_f\left(q \| \mathbf{1}/N_w\right) \leq \frac{\rho}{k}\}$. In this rescaled form, the convex conjugate of $f(t)$ is $f^*(s) = \frac{1}{k_*}\left[s\right]_+^{k_*} + \frac{1}{k}$, where we recall the definition that $\frac{1}{k} + \frac{1}{k_*} = 1$.

Using Lemma 4, we obtain

$$\begin{aligned}
\sup_{Q \in \mathcal{P}} \left| \int \widehat{e}_n(w) dQ(w) \right| &\leq \sup_{Q \in \mathcal{P}} \int \left|\widehat{e}_n(w)\right| dQ(w) \\
&\leq \inf_{\lambda \geq 0} \left\{ \frac{1}{k_*} \mathbb{E}_{P_0}[|\widehat{e}_n(W)|^{k_*}] \lambda^{1-k_*} + \frac{\rho+1}{k} \lambda \right\} \\
&= (\rho+1)^{\frac{1}{k}} \mathbb{E}_{P_0}[|\widehat{e}_n(W)|^{k_*}]^{1/k_*} \\
&\leq \sqrt{\rho+1} \mathbb{E}_{P_0}[\widehat{e}_n(W)^2]^{\frac{1}{2}},
\end{aligned}$$

where the second inequality follows by using $\eta = 0$ in Lemma 4 and the last inequality follows from the fact that $k \geq 2$ and $k_* \leq 2$. The expectation $\mathbb{E}_{P_0}$ is with respect to the variable W for a fixed $\widehat{e}_n$. We now see that to prove the theorem, it suffices to show that

$$\mathbb{P}\left(\int \widehat{e}_n(w)^2 dP_0(w) \geq \frac{t^2}{\rho+1}\right) \leq \sqrt{2} \exp\left(-\frac{nt^2}{16(\rho+1)}\right).$$

By Lemma 6, $\widehat{e}_n$ is $4/n$-sub-Gaussian, whence $\mathbb{E}\left[\exp\left(\lambda \widehat{e}_n(w)^2\right)\right] \leq \exp\left(-\frac{1}{2}\log\left(1 - \frac{8\lambda}{n}\right)\right)$ for $\lambda \leq \frac{n}{8}$ (recall inequality (10) above). Integrating over w, we find that for any distribution P_0 we have by the Chernoff bound technique that for $\lambda \leq \frac{n}{8}$,

$$\begin{aligned}
\mathbb{P}\left(\int \widehat{e}_n(w)^2 dP_0(w) \geq \frac{t^2}{\rho+1}\right) &\leq \mathbb{E}\left[\exp\left(\lambda \int \widehat{e}_n(w)^2 dP(w)\right)\right] \exp\left(-\lambda \frac{t^2}{\rho+1}\right) \\
&\leq \int \mathbb{E}\left[\exp\left(\lambda \widehat{e}_n(w)^2\right)\right] dP(w) \exp\left(-\lambda \frac{t^2}{\rho+1}\right) \\
&\leq \exp\left(-\frac{1}{2}\log\left(1 - \frac{8\lambda}{n}\right)\right) \exp\left(-\lambda \frac{t^2}{\rho+1}\right).
\end{aligned}$$

Note that $-\log(1-t) \leq t \log 4$ for $t \leq \frac{1}{2}$, and take $\lambda = n/16$ to get the result.

A.2 Proof of Lemma 2

Let $F : \mathcal{W} \to [-\|F\|_\infty, \|F\|_\infty]$ be a function of the random W. In our setting, this map is equal to

$$F(w) = \frac{1}{n(n-1)} \sum_{i \neq j} S_{ij} \phi(x^i, w) \phi(x^j, w),$$

where we treat the S_{ij} and x^i as fixed and work conditionally; that is, only W is random. We consider the convergence of

$$\sup_{Q \in \mathcal{P}_{N_w}} \mathbb{E}_Q[F(W)] \text{ to } \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[F(W)].$$

In the sequel, we suppress dependence on W for notational convenience, and for a sample $W_1, \ldots, W_{N_w}$ of random vectors W_k, we let

$$F_k = \frac{1}{n(n-1)} \sum_{i \neq j} S_{ij} \phi(x^i, W_k) \phi(x^j, W_k)$$

for shorthand, so that the F_k are bounded indepenent random variables.

Treating $F = (F_1, \ldots, F_{N_w})$ as a vector, the mapping $F \mapsto \sup_{Q \in \mathcal{P}_{N_w}} \mathbb{E}_Q[F]$ is a Lipschitz convex function of independent bounded random variables. Indeed, letting $q \in \mathbb{R}_+^{N_w}$ be the empirical probability mass function associated with $Q \in \mathcal{P}_{N_w}$ and recalling that $\|x\|_2 \leq n^{\frac{k-2}{2k}} \|x\|_k$ for $x \in \mathbb{R}^n$ and $k \geq 2$, we have $\frac{1}{N_w} \sum_{i=1}^{N_w} (N_w q_i)^k \leq \rho + 1$, which is equivalent to

$$\|q\|_2 \leq {N_w}^{\frac{k-2}{2k}} \|q\|_k \leq {N_w}^{\frac{k-2}{2k}} (\rho+1)^{\frac{1}{k}} {N_w}^{1/k-1} = (\rho+1)^{\frac{1}{k}} {N_w}^{-\frac{1}{2}}. \tag{13}$$

That is, the function $(F_1, \ldots, F_{N_w}) \mapsto \sup_{Q \in \mathcal{P}_{N_w}} \mathbb{E}_Q[F]$ is an $L_{N_w} = \sqrt{\rho+1}/\sqrt{N_w}$-Lipschitz and convex function of bounded random variables. Using Samson's sub-Gaussian concentration inequality [26] for Lipschitz convex functions of bounded random variables, we have with probability at least $1 - \delta$ that

$$\sup_{Q \in \mathcal{P}_{N_w}} \mathbb{E}_Q[F] \in \mathbb{E}\left[\sup_{Q \in \mathcal{P}_{N_w}} \mathbb{E}_Q[F]\right] \pm 2\sqrt{2} \|F\|_\infty \sqrt{\frac{(1+\rho)\log\frac{2}{\delta}}{N_w}}. \tag{14}$$

By the containment (14), we need consider only the convergence of the expectation

$$\mathbb{E}\left[\sup_{Q \in \mathcal{P}_{N_w}} \mathbb{E}_Q[F]\right] \text{ to } \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[F].$$

But of course, this convergence is described precisely by Lemma 5. Thus, combining Lemma 5 with containment (14) gives

$$\left| \sup_{Q \in \mathcal{P}_{N_w}} \mathbb{E}_Q[F] - \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[F] \right| \leq 4C_\rho \|F\|_\infty \sqrt{\frac{\log(2N_w)}{N_w}} + 2\sqrt{2} \|F\|_\infty \sqrt{\frac{(1+\rho)\log\frac{2}{\delta}}{N_w}}$$

Now, since $\|F\|_\infty = 1$ we can simplify this to get the result.

A.3 Proof of Theorem 1

We can write

$$\begin{aligned}
\left| T(\widehat{Q}_w) - \sup_{Q \in \mathcal{P}} T(Q) \right| &\leq \left| \sup_{Q \in \mathcal{P}} T(Q) - \sup_{Q \in \mathcal{P}} \widehat{T}(Q) \right| + \left| \sup_{Q \in \mathcal{P}} \widehat{T}(Q) - \widehat{T}(\widehat{Q}_w) \right| + \left| \widehat{T}(\widehat{Q}_w) - T(\widehat{Q}_w) \right| \\
&\leq \sup_{Q \in \mathcal{P}} \left| T(Q) - \widehat{T}(Q) \right| + \left| \sup_{Q \in \mathcal{P}} \widehat{T}(Q) - \widehat{T}(\widehat{Q}_w) \right| + \sup_{Q \in \mathcal{P}_{N_w}} \left| \widehat{T}(Q) - T(Q) \right|
\end{aligned}$$

Now apply Lemma 1 to the first and third terms, apply Lemma 2 to the second term, and use a union bound to get the result.

A.4 Proof of Lemma 3

We define define the "dual" representation of the feature matrix: let $\Psi = \Phi^T = [\psi^1 \ \cdots \ \psi^{N_w}]$, with columns given by $\psi^m := [\phi(x^1, w^m) \ \cdots \ \phi(x^n, w^m)]^T \in \mathbb{R}^n$. Mimicking the proof of Proposition 1 of [8], we have

$$\mathcal{R}_n(\mathcal{F}_{N_w}) = \frac{B}{n}\mathbb{E}\left[\sup_{q \in \mathcal{P}_{N_w}} \sqrt{\sigma^T \left(\sum_{k=1}^{N_w} q_k \psi^k (\psi^k)^T\right) \sigma}\right], \tag{15}$$

where $\sigma_i \in \{-1, 1\}$ are iid. Rademacher variables. By the bound (13), the containment $q \in \mathcal{P}_{N_w}$ implies the bound $\|q\|_2 \leq \sqrt{(1+\rho)/N_w}$, so

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}_{N_w}) &\leq \frac{B}{n}\mathbb{E}\left[\sqrt{\sqrt{\frac{1+\rho}{N_w}} \sum_{k=1}^{N_w} \frac{(\sigma^T \psi^k)^4}{\sqrt{\sum_{a=1}^{N_w} (\sigma^T \psi^a)^4}}}\right] \\
&= \frac{B}{n}\mathbb{E}\left[\left(\frac{1+\rho}{N_w} \sum_{k=1}^{N_w} (\sigma^T \psi^k)^4\right)^{\frac{1}{4}}\right] \\
&\leq \frac{B}{n}\left(\mathbb{E}\left[\frac{1+\rho}{N_w} \sum_{k=1}^{N_w} (\sigma^T \psi^k)^4\right]\right)^{\frac{1}{4}},
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality and the second inequality is Jensen's inequality. As $\psi_i \in [-1, 1]$, we have

$$\begin{aligned}
\mathbb{E}\left[(\sigma^T \psi)^4\right] &\leq \mathbb{E}\left[\left(\sum_{i=1}^{n} \sigma_i\right)^4\right] \\
&= 3n^2 - 2n \leq 3n^2.
\end{aligned}$$

Then

$$\mathcal{R}_n(\mathcal{F}_{N_w}) \leq \frac{B}{n}\left(3(1+\rho)n^2\right)^{\frac{1}{4}} \leq B\sqrt{\frac{2(1+\rho)}{n}}$$

as desired.

B Technical lemmas

B.1 Proof of Lemma 4

Let $L \geq 0$ satisfy $L(w) = dQ(w)/dP(w)$, so that L is the likelihood ratio between Q and P. Then we have

$$\begin{aligned}
\sup_{Q: D_f(Q\|P) \leq \rho} \int g(w) dQ(w) &= \sup_{\int f(L) dP \leq \rho, \mathbb{E}_P[L]=1} \int g(w) L(w) dP(w) \\
&= \sup_{L \geq 0} \inf_{\lambda \geq 0, \eta} \left\{\int g(w) L(w) dP(w) - \lambda\left(\int f(L(w)) dP(w) - \rho\right) - \eta\left(\int L(w) dP(w) - 1\right)\right\} \\
&= \inf_{\lambda \geq 0, \eta} \sup_{L \geq 0} \left\{\int g(w) L(w) dP(w) - \lambda\left(\int f(L(w)) dP(w) - \rho\right) - \eta\left(\int L(w) dP(w) - 1\right)\right\},
\end{aligned}$$

where we have used that strong duality obtains because the problem is strictly feasible in its non-linear constraints (take $L \equiv 1$), so that the extended Slater condition holds [22, Theorem 8.6.1 and Problem 8.7]. Noting that L is simply a positive (but otherwise arbitrary) function, we obtain

$$\begin{aligned}
\sup_{Q: D_f(Q\|P) \leq \rho} \int g(w) dQ(w) &= \inf_{\lambda \geq 0, \eta} \int \sup_{\ell \geq 0} \left\{(g(w) - \eta)\ell - \lambda f(\ell)\right\} dP(w) + \lambda\rho + \eta \\
&= \inf_{\lambda \geq 0, \eta} \int \lambda f^*\left(\frac{g(w) - \eta}{\lambda}\right) dP(w) + \eta + \rho\lambda.
\end{aligned}$$

Here we have used that $f^*(s) = \sup_{t \geq 0}\{st - f(t)\}$ is the conjugate of f and that $\lambda \geq 0$, so that we may take divide and multiply by λ in the supremum calculation.

B.2 Proof of Lemma 5

We remark that the upper bound in the lemma is immediate from the argument for inequality (11). Thus we focus only on the lower bound claimed in the lemma.

Before beginning the proof proper, we state a useful lemma lower bounding expectations of various moments of random variables. (See Section B.4 for a proof.)

Lemma 7. *Let $Z \geq 0, Z \not\equiv 0$ be a random variable with finite $2p$-th moment for $1 \leq p \leq \infty$. Then we have the following inequalities:*

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}}\right] \geq \|Z\|_p - \begin{cases} \frac{p-1}{p}\sqrt{\frac{2}{n}}\sqrt{\mathrm{Var}(Z^p/\mathbb{E}[Z^p])}\|Z\|_2, & \text{if } p \leq 2 \\ 2\min\left(\frac{p-1}{p}\sqrt{\frac{1}{n}}\sqrt{\mathrm{Var}(Z^p/\mathbb{E}[Z^p])}\|Z\|_p, \frac{1}{n}\left(\frac{p-1}{p}\right)^2 \frac{\mathrm{Var}(Z^p)}{\|Z\|_p^{2p-1}}\right) & \text{if } p \geq 2. \end{cases} \tag{16a}$$

and if $\|Z\|_\infty \leq C$, then

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}}\right] \geq \|Z\|_p - \begin{cases} C\frac{p-1}{p}\sqrt{\frac{2}{n}}, & \text{if } p \leq 2 \\ 2C\left(\frac{1}{n}\right)^{\frac{1}{p}} & \text{if } p > 2 \end{cases} \tag{16b}$$

For convenience in the proof to follow, we define the shorthand

$$S_{N_w}(\eta) := (1+\rho)^{1/k}\left(\frac{1}{N_w}\sum_{i=1}^{N_w}[Z_i - \eta]_+^{k_*}\right)^{\frac{1}{k_*}} + \eta.$$

We also rescale ρ to ρ/k for algebraic convenience. For the function $f(t) = \frac{1}{k}(t^k - 1)$, we have $f^*(s) = \frac{1}{k_*}[s]_+^{k_*} + \frac{1}{k}$, so that the duality result in Lemma 4 shows that (after taking an infimum over $\lambda \geq 0$)

$$\sup_{Q \in \mathcal{P}_{N_w}} \mathbb{E}_Q[Z] = \inf_\eta \left\{(1+\rho)^{1/k}\left(\frac{1}{N_w}\sum_{i=1}^{N_w}[Z_i - \eta]_+^{k_*}\right)^{\frac{1}{k_*}} + \eta\right\}.$$

Because $|Z_i| \leq M$ for all i, we claim that any η minimizing the preceding expression must satisfy

$$\eta \in \left[-\frac{1 + (1+\rho)^{\frac{1}{k_*}}}{(1+\rho)^{\frac{1}{k_*}} - 1}, 1\right] \cdot M. \tag{17}$$

Indeed, it is clear that $\eta \leq M$, because otherwise we would have $S_{N_w}(\eta) > M \geq \inf_\eta S_{N_w}(\eta)$. The lower bound on η is somewhat less trivial. Let $\eta = -cM$ for some $c > 1$. Taking derivatives of the objective $S_{N_w}(\eta)$ with respect to η, we have

$$\begin{aligned} S'_{N_w}(\eta) = 1 - (1+\rho)^{1/k}\frac{\frac{1}{N_w}\sum_{i=1}^{N_w}[Z_i - \eta]_+^{k_*-1}}{\left(\frac{1}{N_w}\sum_{i=1}^{N_w}[Z_i - \eta]_+^{k_*}\right)^{1-\frac{1}{k_*}}} &\leq 1 - (1+\rho)^{1/k}\left(\frac{(c-1)M}{(c+1)M}\right)^{k_*-1} \\ &= 1 - (1+\rho)^{1/k}\left(\frac{c-1}{c+1}\right)^{k_*-1}. \end{aligned}$$

Defining the constant $c_{\rho,k} := \frac{(1+\rho)^{\frac{1}{k_*}}+1}{(1+\rho)^{\frac{1}{k_*}}-1}$, we see that for any $c > c_{\rho,k}$, the preceding display is negative, so we must have $\eta \geq -c_{\rho,k}M$ (since the derivative is 0 at optimality). For the remainder of the proof, we thus define the interval

$$U := [-Mc_{\rho,k}, M], \;\; c_{\rho,k} = \frac{(1+\rho)^{\frac{1}{k_*}}+1}{(1+\rho)^{\frac{1}{k_*}}-1},$$

and we assume w.l.o.g. that $\eta \in U$.

Again applying the duality result of Lemma 4, we have that

$$\mathbb{E}\left[\sup_{Q \in \mathcal{P}_{N_w}} \mathbb{E}_Q[Z]\right] = \mathbb{E}\left[\inf_{\eta \in U} S_{N_w}(\eta)\right] = \mathbb{E}\left[\inf_{\eta \in U}\{S_{N_w}(\eta) - \mathbb{E}[S_{N_w}(\eta)] + \mathbb{E}[S_{N_w}(\eta)]\}\right]$$
$$\geq \inf_{\eta \in U} \mathbb{E}[S_{N_w}(\eta)] - \mathbb{E}\left[\sup_{\eta \in U} |S_{N_w}(\eta) - \mathbb{E}[S_{N_w}(\eta)]|\right]. \tag{18}$$

To bound the first term in expression (18), note that $[Z - \eta]_+ \in [0, 1 + c_{\rho,k}]M$ and $(1+\rho)^{1/k}(1 + c_{\rho,k}) = C_{\rho,k}$. Thus, by Lemma 7 we obtain that

$$\mathbb{E}[S_{N_w}(\eta)] \geq (1+\rho)^{1/k}\mathbb{E}\left[[Z-\eta]_+^{k_*}\right]^{1/k_*} + \eta - C_{\rho,k}M\frac{k_*-1}{k_*}\sqrt{\frac{2}{N_w}}.$$

Using that $\frac{k_*-1}{k_*} = \frac{1}{k}$, taking the infimum over η on the right hand side and using duality yields

$$\inf_{\eta} \mathbb{E}[S_{N_w}(\eta)] \geq \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[Z] - C_{\rho,k}\frac{M}{k}\sqrt{\frac{2}{N_w}}.$$

To bound the second term in expression (18), we use concentration results for Lipschitz functions. First, the function $\eta \mapsto S_{N_w}(\eta)$ is $\sqrt{1+\rho}$-Lipschitz in η. To see this, note that for $1 \leq k^\star \leq 2$ and $X \geq 0$, by Jensen's inequality,

$$\frac{\mathbb{E}[X^{k^\star - 1}]}{(\mathbb{E}[X^{k^\star}])^{1-1/k^\star}} \leq \frac{\mathbb{E}[X]^{k^\star-1}}{(\mathbb{E}[X^{k^\star}])^{1-1/k^\star}} \leq \frac{\mathbb{E}[X]^{k^\star-1}}{\mathbb{E}[X]^{k^\star-1}} = 1,$$

so $S'_{N_w}(\eta) \in [1 - (1+\rho)^{\frac{1}{k}}, 1]$ and therefore S_{N_w} is $(1+\rho)^{1/k}$-Lipschitz in η. Furthermore, the mapping $T : z \mapsto (1+\rho)^{\frac{1}{k}}(\frac{1}{N_w}\sum_{i=1}^{N_w}[z_i - \eta]_+^{k_*})^{\frac{1}{k_*}}$ for $z \in \mathbb{R}^{N_w}$ is convex and $(1+\rho)^{\frac{1}{k}}/\sqrt{N_w}$-Lipschitz. This is verified by the following:

$$|T(z) - T(z')| \leq (1+\rho)^{1/k}\left|\left(\frac{1}{N_w}\sum_{i=1}^{N_w}\left|[z_i-\eta]_+ - [z'_i - \eta]_+\right|^{k_*}\right)^{\frac{1}{k_*}}\right|$$
$$\leq \frac{(1+\rho)^{1/k}}{{N_w}^{1/k_*}}\left|\left(\sum_{i=1}^{N_w}|z_i - z'_i|^{k_*}\right)^{\frac{1}{k_*}}\right|$$
$$\leq \frac{(1+\rho)^{1/k}}{\sqrt{N_w}}\|z - z'\|_2,$$

where the first inequality is Minkowski's inequality and the third inequality follows from the fact that for any vector $x \in \mathbb{R}^n$, we have $\|x\|_p \leq n^{\frac{2-p}{2p}}\|x\|_2$ for $p \in [1,2]$, where these denote the usual vector norms. Thus, the mapping $Z \mapsto S_{N_w}(\eta)$ is $(1+\rho)^{1/k}/\sqrt{N_w}$-Lipschitz continuous with respect to the ℓ_2-norm on Z. Again applying Samson's sub-Gaussian concentration result for convex Lipschitz functions, we have

$$\mathbb{P}\left(|S_{N_w}(\eta) - \mathbb{E}[S_{N_w}(\eta)]| \geq \delta\right) \leq 2\exp\left(-\frac{N_w\delta^2}{2C_{\rho,k}^2M^2}\right)$$

for any fixed $\eta \in \mathbb{R}$ and any $\delta \geq 0$. Now, let $\mathcal{N}(U,\epsilon) = \{\eta_1, \ldots, \eta_{N(U,\epsilon)}\}$ be an ϵ cover of the set U, which we may take to have size at most $N(U,\epsilon) \leq M(1+c_{\rho,k})\frac{1}{\epsilon}$. Then we have

$$\sup_{\eta \in U}|S_{N_w}(\eta) - \mathbb{E}[S_{N_w}(\eta)] \leq \max_{i \in \mathcal{N}(U,\epsilon)}|S_{N_w}(\eta_i) - \mathbb{E}[S_{N_w}(\eta_i)]| + \epsilon(1+\rho)^{1/k}.$$

Using the fact that $\mathbb{E}[\max_{i \leq n}|X_i|] \leq \sqrt{2\sigma^2\log(2n)}$ for X_i all σ^2-sub-Gaussian, we have

$$\mathbb{E}\left[\max_{i \in \mathcal{N}(U,\epsilon)}|S_{N_w}(\eta_i) - \mathbb{E}[S_{N_w}(\eta_i)]|\right] \leq C_{\rho,k}\sqrt{2\frac{M^2}{N_w}\log 2N(U,\epsilon)}.$$

Taking $\epsilon = M(1 + c_{\rho,k})/N_w$ gives that

$$\mathbb{E}\left[\sup_{\eta \in U} |S_{N_w}(\eta) - \mathbb{E}[S_{N_w}(\eta)]\right] \leq \sqrt{2} M C_{\rho,k} \sqrt{\frac{1}{N_w} \log(2N_w)} + \frac{C_{\rho,k} M}{N_w}.$$

Then, in total we have (using $C_\rho \geq C_{\rho,k}$, $k \geq 2$, and $N_w \geq 1$),

$$\begin{aligned}
\mathbb{E}\left[\sup_{Q \in \mathcal{P}_{N_w}} \mathbb{E}_Q[Z]\right] &\geq \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[Z] - \frac{C_\rho M \sqrt{2}}{\sqrt{N_w}} \left(\frac{1}{k} + \sqrt{\log(2N_w)} + \frac{1}{\sqrt{2N_w}}\right) \\
&\geq \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[Z] - 4 C_\rho M \sqrt{\frac{\log(2N_w)}{N_w}}.
\end{aligned}$$

This gives the desired result of the lemma.

B.3 Proof of Lemma 6

The result follows from bounded differences. First, we let

$$\widehat{e}'_n(w) = \frac{1}{n(n-1)} \sum_{i \neq j} S'_{ij} \phi(x'_i; w) \phi(x'_j; w) - \mathbb{E}[S(X, X') \phi(X, w) \phi(X', w)],$$

where we assume $d_{\text{ham}}(x_{1:n}, x'_{1:n}) \leq 1$ and $S_{ij} = S'_{ij}$ except for those pairs (i, j) such that $x'_i \neq x_i$ or $x_j \neq x'_j$. Assuming (without loss of generality by symmetry) that $x_{2:n} = x'_{2:n}$, we have

$$\begin{aligned}
|\widehat{e}_n(w) - \widehat{e}'_n(w)| &\leq \frac{1}{n(n-1)} \sum_{j > 1} \left| S_{1j} \phi(x_1; w) \phi(x_j; w) - S'_{1j} \phi(x'_1; w) \phi(x_j; w) \right| \\
&\quad + \frac{1}{n(n-1)} \sum_{i > 1} |S_{i1} \phi(x_i; w) \phi(x_1; w) - S'_{i1} \phi(x'_i; w) \phi(x'_1; w)| \\
&\leq \frac{2(n-1)}{n(n-1)} + \frac{2(n-1)}{n(n-1)} = \frac{4}{n},
\end{aligned}$$

where in the last line we have used that $\max\{\|\phi\|_\infty, \|S\|_\infty\} \leq 1$. In particular, $\widehat{e}_n(w)$ has bounded differences and is mean zero, so that the usual construction with Doob martingales yields

$$\mathbb{E}\left[\exp(\lambda \widehat{e}_n(w))\right] \leq \exp\left(\frac{16\lambda^2}{8n^2}\right)^n = \exp\left(\frac{2\lambda^2}{n}\right).$$

This is the desired result.

B.4 Proof of Lemma 7

For $a > 0$, we have

$$\inf_{\lambda \geq 0} \left\{ \frac{a^p}{p\lambda^{p-1}} + \lambda \frac{p-1}{p} \right\} = a,$$

(with $\lambda = a$ attaining the infimum), and taking derivatives yields

$$\frac{a^p}{p\lambda^{p-1}} + \lambda \frac{p-1}{p} \geq \frac{a^p}{p\lambda_1^{p-1}} + \lambda_1 \frac{p-1}{p} + \frac{p-1}{p} \left(1 - \frac{a^p}{\lambda_1^p}\right)(\lambda - \lambda_1).$$

Using this in the moment expectation, by setting $\lambda_n = \sqrt[p]{\frac{1}{n} \sum_{i=1}^n Z_i^p}$, we have for any $\lambda \geq 0$ that

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}}\right] &= \mathbb{E}\left[\frac{\sum_{i=1}^n Z_i^p}{pn\lambda_n^{p-1}} + \lambda_n \frac{p-1}{p}\right] \\
&\geq \mathbb{E}\left[\frac{\sum_{i=1}^n Z_i^p}{pn\lambda^{p-1}} + \lambda \frac{p-1}{p}\right] + \frac{p-1}{p} \mathbb{E}\left[\left(1 - \frac{\sum_{i=1}^n Z_i^p}{n\lambda^p}\right)(\lambda_n - \lambda)\right].
\end{aligned}$$

Now we take $\lambda = \|Z\|_p$, and we apply the Cauchy-Schwarz inequality to obtain

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}}\right] &\geq \|Z\|_p - \frac{p-1}{p}\mathbb{E}\left[\left(1 - \frac{\frac{1}{n}\sum_{i=1}^n Z_i^p}{\|Z\|_p^p}\right)^2\right]^{\frac{1}{2}} \mathbb{E}\left[\left(\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}} - \|Z\|_p\right)^2\right]^{\frac{1}{2}} \\
&= \|Z\|_p - \frac{p-1}{p\sqrt{n}}\sqrt{\mathrm{Var}(Z^p/\mathbb{E}[Z^p])}\mathbb{E}\left[\left(\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}} - \mathbb{E}[Z^p]^{\frac{1}{p}}\right)^2\right]^{\frac{1}{2}} \\
&\geq \|Z\|_p - \frac{p-1}{p\sqrt{n}}\sqrt{\mathrm{Var}(Z^p/\mathbb{E}[Z^p])}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{2}{p}} + \mathbb{E}[Z^p]^{\frac{2}{p}}\right]^{\frac{1}{2}}.
\end{aligned} \tag{19}$$

Now, for $p \leq 2$, we have

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}}\right] \geq \|Z\|_p - \frac{p-1}{p}\sqrt{\frac{2}{n}}\sqrt{\mathrm{Var}(Z^p/\mathbb{E}[Z^p])}\|Z\|_2,$$

by Jensen, or equivalently, the fact that the norm is non-decreasing in p. For $p \geq 2$, we have by the triangle inequality applied to expression (19), followed by an application of Jensen's inequality (using that $\mathbb{E}[Y^{2/p}] \leq \mathbb{E}[Y]^{2/p}$ for $p \geq 2$),

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}}\right] \geq \|Z\|_p - 2\frac{p-1}{p}\sqrt{\frac{1}{n}}\sqrt{\mathrm{Var}(Z^p/\mathbb{E}[Z^p])}\|Z\|_p,$$

Now, we can make this tighter (for $p \geq 2$):

$$\begin{aligned}
\mathbb{E}\left[\left(\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}} - \mathbb{E}[Z^p]^{\frac{1}{p}}\right)^2\right] &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{2}{p}}\right] + \|Z\|_p^2 - 2\|Z\|_p\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}}\right] \\
&\leq 2\|Z\|_p^2 - 2\|Z\|_p\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}}\right] \\
&\leq 2\frac{p-1}{p}\frac{2}{\sqrt{n}}\sqrt{\mathrm{Var}(Z^p/\mathbb{E}[Z^p])}\|Z\|_p^2.
\end{aligned}$$

Further, we can recurse this argument. Let

$$\begin{aligned}
Y &:= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}}\right] \\
A &:= \|Z\|_p \\
B &:= \frac{p-1}{p}\sqrt{\frac{1}{n}}\sqrt{\mathrm{Var}(Z^p/\mathbb{E}[Z^p])}, \\
C &:= \mathbb{E}\left[\left(\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}} - \mathbb{E}[Z^p]^{\frac{1}{p}}\right)^2\right].
\end{aligned}$$

Then, we have three primary relationships $r : Y \geq A - BC^{\frac{1}{2}}$, $s_0 : C \leq 2A^2 - 2AY$, and $t_0 : Y \geq A - 2AB$. Recursion works as follows: for $i \geq 0$, we plug t_i into s_0 to yield a tighter inequality s_{i+1} for C, which in turn plugs in to r to yield a tighter inequality t_{i+1} for Y. In this way, we have the relations $s_i : C \leq 4A^2B^{a_{i-1}}$ for $i \geq 1$, and $t_i : Y \geq A - 2AB^{a_i}$ for $i \geq 0$, where

$a_i = 2 - 2^{-i}$. Taking $i \to \infty$, we have $Y \geq A - 2AB^2$, or

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}}\right] \geq \|Z\|_p - 2\|Z\|_p \left(\frac{p-1}{p}\right)^2 \frac{\mathrm{Var}(Z^p/\mathbb{E}[Z^p])}{n}$$
$$= \|Z\|_p - \frac{2}{n}\left(\frac{p-1}{p}\right)^2 \frac{\mathrm{Var}(Z^p)}{\|Z\|_p^{2p-1}}$$

Thus, we have

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}}\right] \geq \|Z\|_p - \begin{cases} \frac{p-1}{p}\sqrt{\frac{2}{n}}\sqrt{\mathrm{Var}(Z^p/\mathbb{E}[Z^p])}\|Z\|_2, & \text{if } p \leq 2 \\ 2\min\left(\frac{p-1}{p}\sqrt{\frac{1}{n}}\sqrt{\mathrm{Var}(Z^p/\mathbb{E}[Z^p])}\|Z\|_p, \frac{1}{n}\left(\frac{p-1}{p}\right)^2 \frac{\mathrm{Var}(Z^p)}{\|Z\|_p^{2p-1}}\right) & \text{if } p \geq 2 \end{cases}$$

In the case that we have the unifom bound $\|Z\|_\infty \leq C$, we can get tighter guarantees. To that end, we state a simple lemma.

Lemma 8. *For any random variable $X \geq 0$ and $a \in [1,2]$, we have*

$$\mathbb{E}[X^{ak}] \leq \mathbb{E}[X^k]^{2-a}\mathbb{E}[X^{2k}]^{a-1}$$

Proof For $c \in [0,1]$, $1/p + 1/q = 1$ and $A \geq 0$, we have by Holder's inequality,

$$\mathbb{E}[A] = \mathbb{E}[A^c A^{1-c}] \leq \mathbb{E}[A^{pc}]^{1/p}\mathbb{E}[A^{q(1-c)}]^{1/q}$$

Now take $A := X^{ak}$, $1/p = 2 - a$, $1/q = a - 1$, and $c = \frac{2}{a} - 1$. □

First, note that $\mathbb{E}[Z^{2p}] \leq C^p\mathbb{E}[Z^p]$. For $1 \leq p \leq 2$, we can take $a = 2/p$ in Lemma 8, so that we have

$$E[Z^2] \leq \mathbb{E}[Z^p]^{2-\frac{2}{p}}\mathbb{E}[Z^{2p}]^{\frac{2}{p}-1} \leq \|Z\|_p^p C^{2-p}.$$

Now, we can plug these into the expression above (using $\mathrm{Var}\, Z^p \leq \mathbb{E}[Z^{2p}] \leq C^p\|Z\|_p^p$):

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}}\right] \geq \|Z\|_p - \begin{cases} C\frac{p-1}{p}\sqrt{\frac{2}{n}}, & \text{if } p \leq 2 \\ 2\min\left(\frac{p-1}{p}\sqrt{\frac{1}{n}}\sqrt{\mathrm{Var}(Z^p/\mathbb{E}[Z^p])}\|Z\|_p, \frac{1}{n}\left(\frac{p-1}{p}\right)^2 \frac{\mathrm{Var}(Z^p)}{\|Z\|_p^{2p-1}}\right) & \text{if } p \geq 2 \end{cases}$$

In fact, we can give a somewhat sharper result by noting that $\mathbb{E}[(\frac{1}{n}\sum_{i=1}^n Z_i^p)^{1/p}] \geq 0$, and similarly, $\|Z\|_p \geq 0$. For shorthand, let $D = (\frac{p-1}{p})^2 C^p$. Then using that $\mathrm{Var}(Z^p/\mathbb{E}[Z^p]) = \mathrm{Var}(Z^p)/\|Z\|_p^{2p} \leq \mathbb{E}[Z^{2p}]/\|Z\|_p^{2p} \leq C^p/\|Z\|_p^p$, the preceding inequality, in the case that $p \geq 2$, implies

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}}\right] \geq \|Z\|_p - 2\min\left\{\sqrt{D/n}\,\|Z\|_p^{1-p/2}, (D/n)\,\|Z\|_p^{1-p}, \|Z\|_p/2\right\}$$
$$\geq \|Z\|_p - 2\min\left\{\sqrt{D/n}\,\|Z\|_p^{1-p/2}, (D/n)\,\|Z\|_p^{1-p}, \|Z\|_p\right\}.$$

But now, we note that

$$\min_{t\geq 0}\left\{\sqrt{\frac{D}{n}}t^{1-p/2}, \frac{D}{n}t^{1-p}, t\right\} = \begin{cases} t, & \text{if } t \leq (D/n)^{1/p} \\ \frac{D}{n}t^{1-p}, & \text{if } t > (D/n)^{1/p} \end{cases}$$
$$\leq (D/n)^{1/p}.$$

In particular, we have for $p \geq 2$ that

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i^p\right)^{\frac{1}{p}}\right] \geq \|Z\|_p - 2\left(\frac{1}{n}\left(\frac{p-1}{p}\right)^2 C^p\right)^{1/p} \geq \|Z\|_p - 2C\left(\frac{1}{n}\right)^{\frac{1}{p}}.$$

Finally, we note that the bound for $p \leq 2$ is tighter than the above expression for $p = 2$.

C More experiments

We present further details of the experiments shown in Section 4 as well as experiments on more datasets and kernel-learning methods. Specifically, we also show experiments with the `ads`[5], `farm`[6], `mnist`[7], and `weight`[8] datasets. When training/test splits do not already exist, we split the dataset into 75% training and 25% test sets.

Table 3 shows parameters used in our method for each dataset. The last column indicates the size of the subset of the training data used to solve problem (4). We use subsets to increase the efficiency of our approach. Furthermore, we show ρ/N_w simply because it is easier to work with this quantity rather than ρ: the value is chosen to balance fit with efficiency via cross validation. Very large ρ yields extremely sparse $\widehat{q}$ and poor fit, whereas very small ρ yields dense $\widehat{q}$ and long training times. We note that all values of ρ are less than 1000. Finally, for ridge regression models, we choose the l_2 penalty term such that we may absorb the $\sqrt{\widehat{q}_i}$ factors into θ.

Table 4 compares the accuracy of our approach (OK) with other methods: random features with 2 values for D, and two standard multiple-kernel-learning algorithms from [14]. Table 5 shows the (training + test) times of the same methods. Algorithm ABMKSVM(ratio) is a heuristic alignment-based kernel derived in problem (2) in [14] followed by an SVM. Algorithm MKSVM jointly optimizes kernel composition with empirical risk via problem (9) in [14]. For both of these methods, we consider optimizing the combination of a linear, second-order polynomial, and Gaussian kernel.

The two multiple-kernel-learning approaches require an extremely large amount of memory to build Gram matrices, so we train on subsets of data when necessary to avoid latencies introduced by swapping data from memory. For ABMKSVM(ratio) we train on $n = 17500$ for `adult` and `weight`, and $n = 10000$ for `reuters`. Similarly, we break up the test data for `reuters` into $n_{test} = 1000$ chunks, which accounts for the large amount of time taken for this dataset (training time was roughly $400s$). For MKSVM, we use a subset of size $n = 7500$ for all applicable datasets, and we use the same testing scheme as ABMKSVM(ratio) for `reuters` (training time for MKSVM was roughly $1000s$).

The performance of our method on all datasets is consistent: we improve the performance for random features at a given computational cost, and we are generally competitive with much costlier standard multiple-kernel-learning techniques. The `mnist` and `weight` datasets are slightly peculiar: both ABSVM(ratio) and MKSVM require many support vectors, indicating that the chosen kernels are poor for the task; this hypothesis is corroborated by the slightly worse performance of both our method and random features (the arc-cosine kernel is similar to polynomial and Guassian kernels). A large number of support vectors roughly translates to large $\mathrm{nnz}(\widehat{q})$, which can be achieved by increasing N_w or decreasing ρ. We can also achieve better performance by increasing the subset of training data used in problem (4). Doing the latter two options yields comparable results for our method (Table 6). For the `mnist` models, we switch to ridge regression to enhance efficiency of the larger problem. The upshot of this analysis is that our method is most effective in regimes where standard multiple-kernel-learning techniques are intractable, that is, datasets with both large n and d.

[5] http://archive.ics.uci.edu/ml/datasets/Internet+Advertisements. We use all but the first 3 features which are sometimes missing in the data.

[6] https://archive.ics.uci.edu/ml/datasets/Farm+Ads

[7] http://yann.lecun.com/exdb/mnist/. We do pairwise classifications of digits 1 vs. 7, 4 vs. 9, and 5 vs. 6.

[8] http://archive.ics.uci.edu/ml/datasets/Weight+Lifting+Exercises+monitored+with+Inertial+Measurement+Units. We neglect the first 4 features, and furthermore we only use remaining features that are not missing in any datapoint. We consider classifying the datapoint as class A or not.

Table 3: Dataset parameters

Dataset	n,	n_{test}	d	Model	Base kernel	ρ/N_w	N_w	$\%n$ in problem (4)
adult	32561,	16281	123	Logistic	Gaussian	0.0120	20000	50
reuters	23149,	781265	47236	Ridge	Linear	0.0123	47236	100
buzz	105530,	35177	77	Ridge	Arc-cosine	0.0145	2000	6.67
ads	2459,	820	1554	Ridge	Linear	0.1000	1554	100
farm	3107,	1036	54877	Ridge	Linear	0.0050	54877	100
mnist17	13007,	2163	784	Logistic	Arc-cosine	0.0300	20000	25
mnist49	11791,	1991	784	Logistic	Arc-cosine	0.0300	20000	25
mnist56	11339,	1850	784	Logistic	Arc-cosine	0.0300	20000	25
weight	29431,	9811	53	Ridge	Gaussian	0.0020	20000	50

Table 4: Test misclassification error (%)

Dataset	OK $D = \mathrm{nnz}(\widehat{q})$	Random $D = \mathrm{nnz}(\widehat{q})$	Random $D = 10\,\mathrm{nnz}(\widehat{q})$	ABMKSVM(ratio)	MKSVM
adult	15.54	17.51	16.08	15.44	16.79
reuters	9.27	46.49	23.69	9.09	10.13
buzz	4.92	8.68	4.16	3.48	3.54
ads	5.37	8.05	3.54	3.05	3.17
farm	11.58	23.36	14.58	10.81	10.23
mnist17	3.24	4.44	1.76	0.51	0.97
mnist49	6.53	21.55	4.02	1.10	1.26
mnist56	6.81	5.89	3.03	0.87	0.59
weight	13.08	15.68	2.89	0.78	1.49

Table 5: Time (s)

Dataset	OK $D = \mathrm{nnz}(\widehat{q})$	Random $D = \mathrm{nnz}(\widehat{q})$	Random $D = 10\,\mathrm{nnz}(\widehat{q})$	ABMKSVM(ratio)	MKSVM
adult	3.6	4.6	86.9	87.3	740.9
reuters	0.8	0.2	1.0	31207.4	17490.7
buzz	2.0	1.9	60.2	92.7	1035.1
ads	0.017	0.013	0.014	56.7	92.3
farm	0.27	0.05	8.3	86.3	180.0
mnist17	3.4	4.0	53.1	38.0	702.6
mnist49	3.7	4.4	78.1	27.0	602.5
mnist56	2.9	3.6	56.4	24.3	623.9
weight	1.9	1.0	65.0	83.1	695.3

Table 6: Auxiliary experiments on mnist and weight with OK

Dataset	Model	Base kernel	ρ/N_w	$\%n$ in problem (4)	Test error (%)	Time (s)
mnist17	Ridge	Arc-cosine	0.00100	50	1.06	9.1
mnist49	Ridge	Arc-cosine	0.00100	50	1.91	9.4
mnist56	Ridge	Arc-cosine	0.00100	50	1.68	8.3
weight	Ridge	Gaussian	0.00015	100	2.04	64.7