Correlated random features for fast semi-supervised learning

Brian McWilliams
ETH Zürich, Switzerland
brian.mcwilliams@inf.ethz.ch

David Balduzzi
ETH Zürich, Switzerland
david.balduzzi@inf.ethz.ch

Joachim M. Buhmann
ETH Zürich, Switzerland
jbuhmann@inf.ethz.ch

Abstract

This paper presents Correlated Nyström Views (XNV), a fast semi-supervised algorithm for regression and classification. The algorithm draws on two main ideas. First, it generates two views consisting of computationally inexpensive random features. Second, multiview regression, using Canonical Correlation Analysis (CCA) on unlabeled data, biases the regression towards useful features. It has been shown that CCA regression can substantially reduce variance with a minimal increase in bias if the views contains accurate estimators. Recent theoretical and empirical work shows that regression with random features closely approximates kernel regression, implying that the accuracy requirement holds for random views. We show that XNV consistently outperforms a state-of-the-art algorithm for semi-supervised learning: substantially improving predictive performance and reducing the variability of performance on a wide variety of real-world datasets, whilst also reducing runtime by orders of magnitude.

1 Introduction

As the volume of data collected in the social and natural sciences increases, the computational cost of learning from large datasets has become an important consideration. For learning non-linear relationships, kernel methods achieve excellent performance but naïvely require operations cubic in the number of training points.

Randomization has recently been considered as an alternative to optimization that, surprisingly, can yield comparable generalization performance at a fraction of the computational cost [1, 2]. Random features have been introduced to approximate kernel machines when the number of training examples is very large, rendering exact kernel computation intractable. Among several different approaches, the Nyström method for low-rank kernel approximation [1] exhibits good theoretical properties and empirical performance [3–5].

A second problem arising with large datasets concerns obtaining *labels*, which often requires a domain expert to manually assign a label to each instance which can be very expensive – requiring significant investments of both time and money – as the size of the dataset increases. Semi-supervised learning aims to improve prediction by extracting useful structure from the unlabeled data points and using this in conjunction with a function learned on a small number of labeled points.

Contribution. This paper proposes a new semi-supervised algorithm for regression and classification, Correlated Nyström Views (XNV), that addresses both problems simultaneously. The method

consists in essentially two steps. First, we construct two "views" using random features. We investigate two ways of doing so: one based on the Nyström method and another based on random Fourier features (so-called kitchen sinks) [2, 6]. It turns out that the Nyström method almost always outperforms Fourier features by a quite large margin, so we only report these results in the main text.

The second step, following [7], uses Canonical Correlation Analysis (CCA, [8, 9]) to bias the optimization procedure towards features that are correlated across the views. Intuitively, if both views contain accurate estimators, then penalizing uncorrelated features reduces variance without increasing the bias by much. Recent theoretical work by Bach [5] shows that Nyström views can be expected to contain accurate estimators.

We perform an extensive evaluation of XNV on 18 real-world datasets, comparing against a modified version of the SSSL (simple semi-supervised learning) algorithm introduced in [10]. We find that XNV outperforms SSSL by around 10-15% on average, depending on the number of labeled points available, see §3. We also find that the performance of XNV exhibits dramatically less variability than SSSL, with a typical reduction of 30%.

We chose SSSL since it was shown in [10] to outperform a state of the art algorithm, Laplacian Regularized Least Squares [11]. However, since SSSL does not scale up to large sets of unlabeled data, we modify SSSL by introducing a Nyström approximation to improve runtime performance. This reduces runtime by a factor of $\times 1000$ on $N = 10,000$ points, with further improvements as N increases. Our approximate version of SSSL outperforms kernel ridge regression (KRR) by $> 50\%$ on the 18 datasets on average, in line with the results reported in [10], suggesting that we lose little by replacing the exact SSSL with our approximate implementation.

Related work. Multiple view learning was first introduced in the co-training method of [12] and has also recently been extended to unsupervised settings [13, 14]. Our algorithm builds on an elegant proposal for multi-view regression introduced in [7]. Surprisingly, despite guaranteeing improved prediction performance under a relatively weak assumption on the views, CCA regression has not been widely used since its proposal – to the best of our knowledge this is first empirical evaluation of multi-view regression's performance. A possible reason for this is the difficulty in obtaining naturally occurring data equipped with multiple views that can be shown to satisfy the multi-view assumption. We overcome this problem by constructing random views that satisfy the assumption by design.

2 Method

This section introduces XNV, our semi-supervised learning method. The method builds on two main ideas. First, given two equally useful but sufficiently different views on a dataset, penalizing regression using the canonical norm (computed via CCA), can substantially improve performance [7]. The second is the Nyström method for constructing random features [1], which we use to construct the views.

2.1 Multi-view regression

Suppose we have data $T = \big((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\big)$ for $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$, sampled according to joint distribution $P(\mathbf{x}, y)$. Further suppose we have two *views* on the data

$$\mathbf{z}^{(\nu)} : \mathbb{R}^D \longrightarrow \mathcal{H}^{(\nu)} = \mathbb{R}^M : \mathbf{x} \mapsto \mathbf{z}^{(\nu)}(\mathbf{x}) =: \mathbf{z}^{(\nu)} \quad \text{for } \nu \in \{1, 2\}.$$

We make the following assumption about linear regressors which can be learned on these views.

Assumption 1 (Multi-view assumption [7])**.** *Define mean-squared error loss function $\ell(g, \mathbf{x}, y) = (g(\mathbf{x}) - y)^2$ and let $\mathrm{loss}(g) := \mathbb{E}_P \ell(g(\mathbf{x}), y)$. Further let $L(Z)$ denote the space of linear maps from a linear space Z to the reals, and define:*

$$f^{(\nu)} := \operatorname*{argmin}_{g \in L(\mathcal{H}^{(\nu)})} \mathrm{loss}(g) \text{ for } \nu \in \{1, 2\} \quad \text{and} \quad f := \operatorname*{argmin}_{g \in L(\mathcal{H}^{(1)} \oplus \mathcal{H}^{(2)})} \mathrm{loss}(g).$$

The multi-view assumption is that

$$\mathrm{loss}\left(f^{(\nu)}\right) - \mathrm{loss}(f) \leq \epsilon \quad \text{for } \nu \in \{1, 2\}. \tag{1}$$

In short, the best predictor in each view is within ϵ of the best overall predictor.

Canonical correlation analysis. Canonical correlation analysis [8, 9] extends principal component analysis (PCA) from one to two sets of variables. CCA finds bases for the two sets of variables such that the correlation between projections onto the bases are maximized.

The first pair of canonical basis vectors, $\left(\mathbf{b}_1^{(1)}, \mathbf{b}_1^{(2)}\right)$ is found by solving:

$$\underset{\mathbf{b}^{(1)}, \mathbf{b}^{(2)} \in \mathbb{R}^M}{\operatorname{argmax}} \operatorname{corr}\left(\mathbf{b}^{(1)\top}\mathbf{z}^{(1)}, \mathbf{b}^{(2)\top}\mathbf{z}^{(2)}\right). \tag{2}$$

Subsequent pairs are found by maximizing correlations subject to being orthogonal to previously found pairs. The result of performing CCA is two sets of bases, $\mathbf{B}^{(\nu)} = \left[\mathbf{b}_1^{(\nu)}, \ldots, \mathbf{b}_M^{(\nu)}\right]$ for $\nu \in \{1, 2\}$, such that the projection of $\mathbf{z}^{(\nu)}$ onto $\mathbf{B}^{(\nu)}$ which we denote $\bar{\mathbf{z}}^{(\nu)}$ satisfies

1. *Orthogonality:* $\mathbb{E}_T\left[\bar{\mathbf{z}}_j^{(\nu)\top}\bar{\mathbf{z}}_k^{(\nu)}\right] = \delta_{jk}$, where δ_{jk} is the Kronecker delta, and
2. *Correlation:* $\mathbb{E}_T\left[\bar{\mathbf{z}}_j^{(1)\top}\bar{\mathbf{z}}_k^{(2)}\right] = \lambda_j \cdot \delta_{jk}$ where w.l.o.g. we assume $1 \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq 0$.

λ_j is referred to as the j^{th} *canonical correlation coefficient*.

Definition 1 (canonical norm)**.** *Given vector $\bar{\mathbf{z}}^{(\nu)}$ in the canonical basis, define its* canonical norm *as*

$$\|\bar{\mathbf{z}}^{(\nu)}\|_{CCA} := \sqrt{\sum_{j=1}^{D} \frac{1-\lambda_j}{\lambda_j}\left(\bar{z}_j^{(\nu)}\right)^2}.$$

Canonical ridge regression. Assume we observe n pairs of views coupled with real valued labels $\left\{\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)}, y_i\right\}_{i=1}^n$, canonical ridge regression finds coefficients $\widehat{\boldsymbol{\beta}}^{(\nu)} = \left[\widehat{\beta}_1^{(\nu)}, \ldots, \widehat{\beta}_M^{(\nu)}\right]^\top$ such that

$$\widehat{\boldsymbol{\beta}}^{(\nu)} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{\beta}^{(\nu)\top}\bar{\mathbf{z}}_i^{(\nu)}\right)^2 + \|\boldsymbol{\beta}^{(\nu)}\|_{CCA}^2. \tag{3}$$

The resulting estimator, referred to as the *canonical shrinkage estimator*, is

$$\widehat{\beta}_j^{(\nu)} = \frac{\lambda_j}{n}\sum_{i=1}^{n}\bar{\mathbf{z}}_{i,j}^{(\nu)} y_i. \tag{4}$$

Penalizing with the canonical norm biases the optimization towards features that are highly correlated across the views. Good regressors exist in both views by Assumption 1. Thus, intuitively, penalizing uncorrelated features significantly reduces variance, without increasing the bias by much. More formally:

Theorem 1 (canonical ridge regression, [7])**.** *Assume $\mathbb{E}[y^2|\mathbf{x}] \leq 1$ and that Assumption 1 holds. Let $f_{\widehat{\boldsymbol{\beta}}}^{(\nu)}$ denote the estimator constructed with the canonical shrinkage estimator, Eq.* (4)*, on training set T, and let f denote the best linear predictor across both views. For $\nu \in \{1, 2\}$ we have*

$$\mathbb{E}_T[\operatorname{loss}(f_{\widehat{\boldsymbol{\beta}}}^{(\nu)})] - \operatorname{loss}(f) \leq 5\epsilon + \frac{\sum_{j=1}^{M}\lambda_j^2}{n}$$

where the expectation is with respect to training sets T sampled from $P(\mathbf{x}, y)$.

The first term, 5ϵ, bounds the bias of the canonical estimator, whereas the second, $\frac{1}{n}\sum\lambda_j^2$ bounds the variance. The $\sum\lambda_j^2$ can be thought of as a measure of the "intrinsic dimensionality" of the unlabeled data, which controls the rate of convergence. If the canonical correlation coefficients decay sufficiently rapidly, then the increase in bias is more than made up for by the decrease in variance.

2.2 Constructing random views

We construct two views satisfying Assumption 1 in expectation, see Theorem 3 below. To ensure our method scales to large sets of unlabeled data, we use random features generated using the Nyström method [1].

Suppose we have data $\{\mathbf{x}_i\}_{i=1}^N$. When N is very large, constructing and manipulating the $N \times N$ Gram matrix $[\mathbf{K}]_{ii'} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_{i'}) \rangle = \kappa(\mathbf{x}_i, \mathbf{x}_{i'})$ is computationally expensive. Where here, $\phi(\mathbf{x})$ defines a mapping from $\mathbb{R}^D$ to a high dimensional feature space and $\kappa(\cdot, \cdot)$ is a positive semi-definite kernel function.

The idea behind random features is to instead define a lower-dimensional mapping, $\mathbf{z}(\mathbf{x}_i) : \mathbb{R}^D \to \mathbb{R}^M$ through a random sampling scheme such that $[\mathbf{K}]_{ii'} \approx \mathbf{z}(\mathbf{x}_i)^\top \mathbf{z}(\mathbf{x}_{i'})$ [6, 15]. Thus, using random features, non-linear functions in $\mathbf{x}$ can be learned as linear functions in $\mathbf{z}(\mathbf{x})$ leading to significant computational speed-ups. Here we give a brief overview of the Nyström method, which uses random subsampling to approximate the Gram matrix.

The Nyström method. Fix an $M \ll N$ and randomly (uniformly) sample a subset $\mathcal{M} = \{\hat{\mathbf{x}}_i\}_{i=1}^M$ of M points from the data $\{\mathbf{x}_i\}_{i=1}^N$. Let $\widehat{\mathbf{K}}$ denote the Gram matrix $[\widehat{\mathbf{K}}]_{ii'}$ where $i, i' \in \mathcal{M}$. The Nyström method [1, 3] constructs a low-rank approximation to the Gram matrix as

$$\mathbf{K} \approx \tilde{\mathbf{K}} := \sum_{i=1}^N \sum_{i'=1}^N \left[\kappa(\mathbf{x}_i, \hat{\mathbf{x}}_1), \ldots, \kappa(\mathbf{x}_i, \hat{\mathbf{x}}_M)\right] \widehat{\mathbf{K}}^\dagger \left[\kappa(\mathbf{x}_{i'}, \hat{\mathbf{x}}_1), \ldots, \kappa(\mathbf{x}_{i'}, \hat{\mathbf{x}}_M)\right]^\top, \tag{5}$$

where $\widehat{\mathbf{K}}^\dagger \in \mathbb{R}^{M \times M}$ is the pseudo-inverse of $\widehat{\mathbf{K}}$. Vectors of random features can be constructed as

$$\mathbf{z}(\mathbf{x}_i) = \widehat{\mathbf{D}}^{-1/2} \widehat{\mathbf{V}}^\top \left[\kappa(\mathbf{x}_i, \hat{\mathbf{x}}_1), \ldots, \kappa(\mathbf{x}_i, \hat{\mathbf{x}}_M)\right]^\top,$$

where the columns of $\widehat{\mathbf{V}}$ are the eigenvectors of $\widehat{\mathbf{K}}$ with $\widehat{\mathbf{D}}$ the diagonal matrix whose entries are the corresponding eigenvalues. Constructing features in this way reduces the time complexity of learning a non-linear prediction function from $O(N^3)$ to $O(N)$ [15].

An alternative perspective on the Nyström approximation, that will be useful below, is as follows. Consider integral operators

$$L_N[f](\cdot) := \frac{1}{N} \sum_{i=1}^N \kappa(\mathbf{x}_i, \cdot) f(\mathbf{x}_i) \quad \text{and} \quad L_M[f](\cdot) := \frac{1}{M} \sum_{i=1}^M \kappa(\mathbf{x}_i, \cdot) f(\mathbf{x}_i), \tag{6}$$

and introduce Hilbert space $\hat{\mathcal{H}} = \text{span}\{\hat{\varphi}_1, \ldots, \hat{\varphi}_r\}$ where r is the rank of $\widehat{\mathbf{K}}$ and the $\hat{\varphi}_i$ are the first r eigenfunctions of L_M. Then the following proposition shows that using the Nyström approximation is equivalent to performing linear regression in the feature space ("view") $\mathbf{z} : \mathcal{X} \to \hat{\mathcal{H}}$ spanned by the eigenfunctions of linear operator L_M in Eq. (6):

Proposition 2 (random Nyström view, [3])**.** *Solving*

$$\min_{\mathbf{w} \in \mathbb{R}^r} \frac{1}{N} \sum_{i=1}^N \ell(\mathbf{w}^\top \mathbf{z}(\mathbf{x}_i), y_i) + \frac{\gamma}{2} \|\mathbf{w}\|_2^2 \tag{7}$$

is equivalent to solving

$$\min_{f \in \hat{\mathcal{H}}} \frac{1}{N} \sum_{i=1}^N \ell(f(\mathbf{x}_i), y_i) + \frac{\gamma}{2} \|f\|_{\mathcal{H}_\kappa}^2. \tag{8}$$

2.3 The proposed algorithm: Correlated Nyström Views (XNV)

Algorithm 1 details our approach to semi-supervised learning based on generating two views consisting of Nyström random features and penalizing features which are weakly correlated across views. The setting is that we have labeled data $\{\mathbf{x}_i, y_i\}_{i=1}^n$ and a large amount of unlabeled data $\{\mathbf{x}_i\}_{i=n+1}^N$.

Step 1 generates a set of random features. The next two steps implement multi-view regression using the randomly generated views $\mathbf{z}^{(1)}(\mathbf{x})$ and $\mathbf{z}^{(2)}(\mathbf{x})$. Eq. (9) yields a solution for which unimportant

Algorithm 1 `Correlated Nyström Views (XNV).`

Input: Labeled data: $\{\mathbf{x}_i, y_i\}_{i=1}^n$ and unlabeled data: $\{\mathbf{x}_i\}_{i=n+1}^N$

1: ***Generate features.*** Sample $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{2M}$ uniformly from the dataset, compute the eigendecompositions of the sub-sampled kernel matrices $\hat{\mathbf{K}}^{(1)}$ and $\hat{\mathbf{K}}^{(2)}$ which are constructed from the samples $1, \ldots, M$ and $M+1, \ldots, 2M$ respectively, and featurize the input:

$$\mathbf{z}^{(\nu)}(\mathbf{x}_i) \leftarrow \hat{\mathbf{D}}^{(\nu),-1/2}\hat{\mathbf{V}}^{(\nu)\top}\left[\kappa(\mathbf{x}_i, \hat{\mathbf{x}}_1), \ldots, \kappa(\mathbf{x}_i, \hat{\mathbf{x}}_M)\right]^\top \text{ for } \nu \in \{1, 2\}.$$

2: ***Unlabeled data.*** Compute CCA bases $\mathbf{B}^{(1)}, \mathbf{B}^{(2)}$ and canonical correlations $\lambda_1, \ldots, \lambda_M$ for the two views and set $\bar{\mathbf{z}}_i \leftarrow \mathbf{B}^{(1)}\mathbf{z}^{(1)}(\mathbf{x}_i)$.

3: ***Labeled data.*** Solve

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^n \ell\left(\boldsymbol{\beta}^\top \bar{\mathbf{z}}_i, y_i\right) + \|\boldsymbol{\beta}\|_{CCA}^2 + \gamma\|\boldsymbol{\beta}\|_2^2 \,. \tag{9}$$

Output: $\widehat{\boldsymbol{\beta}}$

features are heavily downweighted in the CCA basis *without* introducing an additional tuning parameter. The further penalty on the ℓ_2 norm (in the CCA basis) is introduced as a practical measure to control the variance of the estimator $\widehat{\boldsymbol{\beta}}$ which can become large if there are many highly correlated features (i.e. the ratio $\frac{1-\lambda_j}{\lambda_j} \approx 0$ for large j). In practice most of the shrinkage is due to the CCA norm: cross-validation obtains optimal values of γ in the range $[0.00001, 0.1]$.

Computational complexity. `XNV` is extremely fast. Nyström sampling, step 1, reduces the $O(N^3)$ operations required for kernel learning to $O(N)$. Computing the CCA basis, step 2, using standard algorithms is in $O(NM^2)$. However, we reduce the runtime to $O(NM)$ by applying a recently proposed randomized CCA algorithm of [16]. Finally, step 3 is a computationally cheap linear program on n samples and M features.

Performance guarantees. The quality of the kernel approximation in (5) has been the subject of detailed study in recent years leading to a number of strong empirical and theoretical results [3–5, 15]. Recent work of Bach [5] provides theoretical guarantees on the quality of Nyström estimates in the fixed design setting that are relevant to our approach.[1]

Theorem 3 (Nyström generalization bound, [5])**.** *Let $\xi \in \mathbb{R}^N$ be a random vector with finite variance and zero mean, $\mathbf{y} = [y_1, \ldots, y_N]^\top$, and define smoothed estimate $\hat{\mathbf{y}}_{kernel} := (\mathbf{K} + N\gamma\mathbf{I})^{-1}\mathbf{K}(\mathbf{y} + \xi)$ and smoothed Nyström estimate $\hat{\mathbf{y}}_{Nyström} := (\tilde{\mathbf{K}} + N\gamma\mathbf{I})^{-1}\tilde{\mathbf{K}}(\mathbf{y} + \xi)$, both computed by minimizing the MSE with ridge penalty γ. Let $\eta \in (0, 1)$. For sufficiently large M (depending on η, see [5]), we have*

$$\mathbb{E}_{\mathcal{M}}\mathbb{E}_\xi\left[\|\mathbf{y} - \hat{\mathbf{y}}_{Nyström}\|_2^2\right] \le (1 + 4\eta) \cdot \mathbb{E}_\xi\left[\|\mathbf{y} - \hat{\mathbf{y}}_{kernel}\|_2^2\right]$$

where $\mathbb{E}_{\mathcal{M}}$ refers to the expectation over subsampled columns used to construct $\tilde{\mathbf{K}}$.

In short, the best smoothed estimators in the Nyström views are close to the optimal smoothed estimator. Since the kernel estimate is consistent, $\text{loss}(f) \to 0$ as $n \to \infty$. Thus, Assumption 1 holds in expectation and the generalization performance of `XNV` is controlled by Theorem 1.

Random Fourier Features. An alternative approach to constructing random views is to use Fourier features instead of Nyström features in Step 1. We refer to this approach as Correlated Kitchen Sinks (`XKS`) after [2]. It turns out that the performance of `XKS` is consistently worse than `XNV`, in line with the detailed comparison presented in [3]. We therefore do not discuss Fourier features in the main text, see §SI.3 for details on implementation and experimental results.

[1] Extending to a random design requires techniques from [17].

Table 1: Datasets used for evaluation.

Set	Name	Task	N	D	Set	Name	Task	N	D
1	abalone[2]	C	2,089	6	10	elevators[4]	R	8,752	18
2	adult[2]	C	32,561	14	11	HIVa[3]	C	21,339	1,617
3	ailerons[4]	R	7,154	40	12	house[4]	R	11,392	16
4	bank8[4]	C	4,096	8	13	ibn Sina[3]	C	10,361	92
5	bank32[4]	C	4,096	32	14	orange[3]	C	25,000	230
6	cal housing[4]	R	10,320	8	15	sarcos 1[5]	R	44,484	21
7	census[2]	R	18,186	119	16	sarcos 5[5]	R	44,484	21
8	CPU[2]	R	6,554	21	17	sarcos 7[5]	R	44,484	21
9	CT[2]	R	30,000	385	18	sylva[3]	C	72,626	216

2.4 A fast approximation to SSSL

The SSSL (simple semi-supervised learning) algorithm proposed in [10] finds the first s eigenfunctions ϕ_i of the integral operator L_N in Eq. (6) and then solves

$$\operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^s} \sum_{i=1}^{n} \left(\sum_{j=1}^{s} w_j \phi_k(\mathbf{x}_i) - y_i \right)^2, \tag{10}$$

where s is set by the user. SSSL outperforms Laplacian Regularized Least Squares [11], a state of the art semi-supervised learning method, see [10]. It also has good generalization guarantees under reasonable assumptions on the distribution of eigenvalues of L_N. However, since SSSL requires computing the full $N \times N$ Gram matrix, it is extremely computationally intensive for large N. Moreover, tuning s is difficult since it is discrete.

We therefore propose SSSL_M, an approximation to SSSL. First, instead of constructing the full Gram matrix, we construct a Nyström approximation by sampling M points from the labeled and unlabeled training set. Second, instead of thresholding eigenfunctions, we use the easier to tune ridge penalty which penalizes directions proportional to the inverse square of their eigenvalues [18].

As justification, note that Proposition 2 states that the Nyström approximation to kernel regression actually solves a ridge regression problem in the span of the eigenfunctions of $\hat{L}_M$. As M increases, the span of $\hat{L}_M$ tends towards that of L_N [15]. We will also refer to the Nyström approximation to SSSL using $2M$ features as SSSL_{2M}. See experiments below for further discussion of the quality of the approximation.

3 Experiments

Setup. We evaluate the performance of XNV on 18 real-world datasets, see Table 1. The datasets cover a variety of regression (denoted by R) and two-class classification (C) problems. The sarcos dataset involves predicting the joint position of a robot arm; following convention we report results on the 1st, 5th and 7th joint positions.

The SSSL algorithm was shown to exhibit state-of-the-art performance over fully and semi-supervised methods in scenarios where few labeled training examples are available [10]. However, as discussed in §2.2, due to its computational cost we compare the performance of XNV to the Nyström approximations SSSL_M and SSSL_{2M}.

We used a Gaussian kernel for all datasets. We set the kernel width, σ and the ℓ_2 regularisation strength, γ, for each method using 5-fold cross validation with 1000 labeled training examples. We trained all methods using a squared error loss function, $\ell(f(\mathbf{x}_i), y_i) = (f(\mathbf{x}_i) - y_i)^2$, with $M = 200$ random features, and $n = 100, 150, 200, \ldots, 1000$ randomly selected training examples.

[2]Taken from the UCI repository http://archive.ics.uci.edu/ml/datasets.html

[3]Taken from http://www.causality.inf.ethz.ch/activelearning.php

[4]Taken from http://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html

[5]Taken from http://www.gaussianprocess.org/gpml/data/

Runtime performance. The SSSL algorithm of [10] is not computationally feasible on large datasets, since it has time complexity $O(N^3)$. For illustrative purposes, we report run times[6] in seconds of the SSSL algorithm against SSSL_M and XNV on three datasets of different sizes.

runtimes	bank8	cal housing	sylva
SSSL	72s	2300s	-
SSSL_{2M}	0.3s	0.6s	24s
XNV	0.9s	1.3s	26s

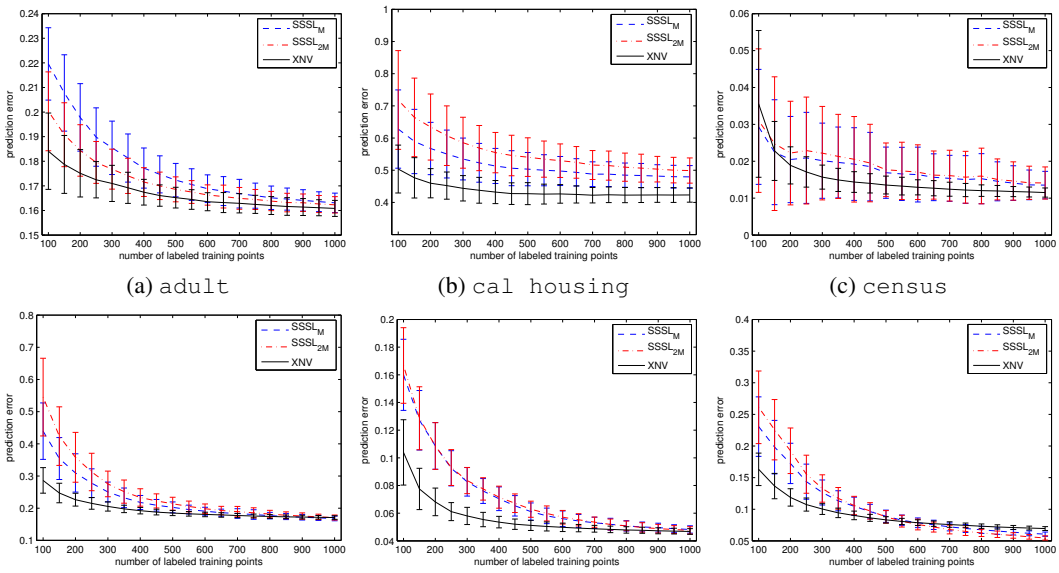
For the cal housing dataset, XNV exhibits an almost $1800\times$ speed up over SSSL. For the largest dataset, sylva, exact SSSL is computationally intractable. Importantly, the computational overhead of XNV over SSSL_{2M} is small.

Generalization performance. We report on the prediction performance averaged over 100 experiments. For regression tasks we report on the mean squared error (MSE) on the testing set normalized by the variance of the test output. For classification tasks we report the percentage of the test set that was misclassified.

The table below shows the improvement in performance of XNV over SSSL_M and SSSL_{2M} (taking whichever performs better out of M or $2M$ on each dataset), averaged over all 18 datasets. Observe that XNV is considerably more accurate and more robust than SSSL_M.

XNV vs $\text{SSSL}_{M/2M}$	$n = 100$	$n = 200$	$n = 300$	$n = 400$	$n = 500$
Avg reduction in error	11%	16%	15%	12%	9%
Avg reduction in std err	15%	30%	31%	33%	30%

The reduced variability is to be expected from Theorem 1.



(a) adult (b) cal housing (c) census

(d) elevators (e) ibn Sina (f) sarcos 5

Figure 1: Comparison of mean prediction error and standard deviation on a selection of datasets.

Table 2 presents more detailed comparison of performance for individual datasets when $n = 200, 400$. The plots in Figure 1 shows a representative comparison of mean prediction errors for several datasets when $n = 100, \ldots, 1000$. Error bars represent one standard deviation. Observe that XNV almost always improves prediction accuracy and reduces variance compared with SSSL_M and SSSL_{2M} when the labeled training set contains between 100 and 500 labeled points. A complete set of results is provided in §SI.1.

Discussion of SSSL_M. Our experiments show that going from M to $2M$ does not improve generalization performance in practice. This suggests that when there are few labeled points, obtaining a

[6]Computed in Matlab 7.14 on a Core i5 with 4GB memory.

more accurate estimate of the eigenfunctions of the kernel does not necessarily improve predictive performance. Indeed, when more random features are added, stronger regularization is required to reduce the influence of uninformative features, this also has the effect of downweighting informative features. This suggests that the low rank approximation $\texttt{SSSL}_M$ to SSSL suffices.

Finally, §SI.2 compares the performance of $\texttt{SSSL}_M$ and XNV to fully supervised kernel ridge regression (KRR). We observe dramatic improvements, between 48% and 63%, consistent with the results observed in [10] for the exact SSSL algorithm.

Random Fourier features. Nyström features significantly outperform Fourier features, in line with observations in [3]. The table below shows the relative improvement of XNV over XKS:

XNV vs XKS	$n = 100$	$n = 200$	$n = 300$	$n = 400$	$n = 500$
Avg reduction in error	30%	28%	26%	25%	24%
Avg reduction in std err	36%	44%	34%	37%	36%

Further results and discussion for XKS are included in the supplementary material.

Table 2: Performance (normalized MSE/classification error rate). Standard errors in parentheses.

set	$\texttt{SSSL}_M$	$\texttt{SSSL}_{2M}$	XNV	set	$\texttt{SSSL}_M$	$\texttt{SSSL}_{2M}$	XNV
$n = 200$							
1	0.054 (0.005)	0.055 (0.006)	**0.053** (**0.004**)	10	0.309 (0.059)	0.358 (0.077)	**0.226** (**0.020**)
2	0.198 (0.014)	0.184 (0.010)	**0.175** (**0.010**)	11	0.146 (0.048)	0.072 (0.024)	**0.036** (**0.001**)
3	0.218 (0.016)	0.231 (0.020)	**0.213** (**0.016**)	12	**0.761** (**0.075**)	0.787 (0.091)	0.792 (0.100)
4	**0.558** (**0.027**)	0.567 (0.029)	0.561 (0.030)	13	0.109 (0.017)	0.109 (0.017)	**0.068** (**0.010**)
5	0.058 (0.004)	0.060 (0.005)	**0.055** (**0.003**)	14	0.019 (0.001)	0.019 (0.001)	**0.019** (**0.000**)
6	0.567 (0.081)	0.634 (0.103)	**0.459** (**0.045**)	15	0.076 (0.008)	0.078 (0.009)	**0.071** (**0.006**)
7	0.020 (0.012)	0.022 (0.014)	**0.019** (**0.005**)	16	0.172 (0.032)	0.192 (0.036)	**0.119** (**0.014**)
8	0.395 (0.395)	0.463 (0.414)	**0.263** (**0.352**)	17	0.041 (0.004)	0.043 (0.005)	**0.040** (**0.004**)
9	0.437 (0.096)	0.367 (0.060)	**0.222** (**0.015**)	18	0.036 (0.007)	0.039 (0.007)	**0.028** (**0.009**)
$n = 400$							
1	0.051 (0.003)	0.052 (0.003)	**0.050** (**0.002**)	10	0.218 (0.022)	0.233 (0.027)	**0.192** (**0.010**)
2	0.177 (0.008)	0.172 (0.006)	**0.167** (**0.005**)	11	0.051 (0.009)	0.122 (0.031)	**0.036** (**0.001**)
3	0.199 (0.011)	0.209 (0.013)	**0.193** (**0.010**)	12	**0.691** (**0.040**)	0.701 (0.051)	0.709 (0.058)
4	0.517 (0.018)	0.527 (0.019)	**0.510** (**0.016**)	13	0.070 (0.009)	0.072 (0.008)	**0.054** (**0.004**)
5	0.050 (0.003)	0.051 (0.003)	**0.050** (**0.002**)	14	0.019 (0.001)	0.019 (0.001)	**0.019** (**0.000**)
6	0.513 (0.055)	0.555 (0.063)	**0.432** (**0.036**)	15	0.059 (0.004)	0.060 (0.005)	**0.057** (**0.003**)
7	0.019 (0.010)	0.021 (0.012)	**0.014** (**0.003**)	16	0.105 (0.014)	0.106 (0.014)	**0.090** (**0.007**)
8	0.209 (0.171)	0.286 (0.248)	**0.110** (**0.107**)	17	**0.032** (**0.002**)	0.033 (0.003)	**0.032** (**0.002**)
9	0.249 (0.024)	0.304 (0.037)	**0.201** (**0.013**)	18	0.029 (0.006)	0.032 (0.005)	**0.023** (**0.006**)

4 Conclusion

We have introduced the XNV algorithm for semi-supervised learning. By combining two randomly generated views of Nyström features via an efficient implementation of CCA, XNV outperforms the prior state-of-the-art, SSSL, by 10-15% (depending on the number of labeled points) on average over 18 datasets. Furthermore, XNV is over 3 orders of magnitude faster than SSSL on medium sized datasets ($N = 10,000$) with further gains as N increases. An interesting research direction is to investigate using the recently developed deep CCA algorithm, which extracts higher order correlations between views [19], as a preprocessing step.

In this work we use a uniform sampling scheme for the Nyström method for computational reasons since it has been shown to perform well empirically relative to more expensive schemes [20]. Since CCA gives us a criterion by which to measure the important of random features, in the future we aim to investigate active sampling schemes based on canonical correlations which may yield better performance by selecting the most informative indices to sample.

Acknowledgements. We thank Haim Avron for help with implementing randomized CCA and Patrick Pletscher for drawing our attention to the Nyström method.

References

[1] Williams C, Seeger M: **Using the Nyström method to speed up kernel machines**. In *NIPS* 2001.

[2] Rahimi A, Recht B: **Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning**. In *Adv in Neural Information Processing Systems (NIPS)* 2008.

[3] Yang T, Li YF, Mahdavi M, Jin R, Zhou ZH: **Nyström Method vs Random Fourier Features: A Theoretical and Empirical Comparison**. In *NIPS* 2012.

[4] Gittens A, Mahoney MW: **Revisiting the Nyström method for improved large-scale machine learning**. In *ICML* 2013.

[5] Bach F: **Sharp analysis of low-rank kernel approximations**. In *COLT* 2013.

[6] Rahimi A, Recht B: **Random Features for Large-Scale Kernel Machines**. In *Adv in Neural Information Processing Systems* 2007.

[7] Kakade S, Foster DP: **Multi-view Regression Via Canonical Correlation Analysis**. In *Computational Learning Theory (COLT)* 2007.

[8] Hotelling H: **Relations between two sets of variates**. *Biometrika* 1936, **28**:312–377.

[9] Hardoon DR, Szedmak S, Shawe-Taylor J: **Canonical Correlation Analysis: An Overview with Application to Learning Methods**. *Neural Comp* 2004, **16**(12):2639–2664.

[10] Ji M, Yang T, Lin B, Jin R, Han J: **A Simple Algorithm for Semi-supervised Learning with Improved Generalization Error Bound**. In *ICML* 2012.

[11] Belkin M, Niyogi P, Sindhwani V: **Manifold regularization: A geometric framework for learning from labeled and unlabeled examples**. *JMLR* 2006, **7**:2399–2434.

[12] Blum A, Mitchell T: **Combining labeled and unlabeled data with co-training**. In *COLT* 1998.

[13] Chaudhuri K, Kakade SM, Livescu K, Sridharan K: **Multiview clustering via Canonical Correlation Analysis**. In *ICML* 2009.

[14] McWilliams B, Montana G: **Multi-view predictive partitioning in high dimensions**. *Statistical Analysis and Data Mining* 2012, **5**:304–321.

[15] Drineas P, Mahoney MW: **On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-Based Learning**. *JMLR* 2005, **6**:2153–2175.

[16] Avron H, Boutsidis C, Toledo S, Zouzias A: **Efficient Dimensionality Reduction for Canonical Correlation Analysis**. In *ICML* 2013.

[17] Hsu D, Kakade S, Zhang T: **An Analysis of Random Design Linear Regression**. In *COLT* 2012.

[18] Dhillon PS, Foster DP, Kakade SM, Ungar LH: **A Risk Comparison of Ordinary Least Squares vs Ridge Regression**. *Journal of Machine Learning Research* 2013, **14**:1505–1511.

[19] Andrew G, Arora R, Bilmes J, Livescu K: **Deep Canonical Correlation Analysis**. In *ICML* 2013.

[20] Kumar S, Mohri M, Talwalkar A: **Sampling methods for the Nyström method**. *JMLR* 2012, **13**:981–1006.