From Random Projections to Learning Theory and Back

Ata Kabán
School of Computer Science
The University of Birmingham
Birmingham B15 2TT, UK
http://www.cs.bham.ac.uk/∼axk

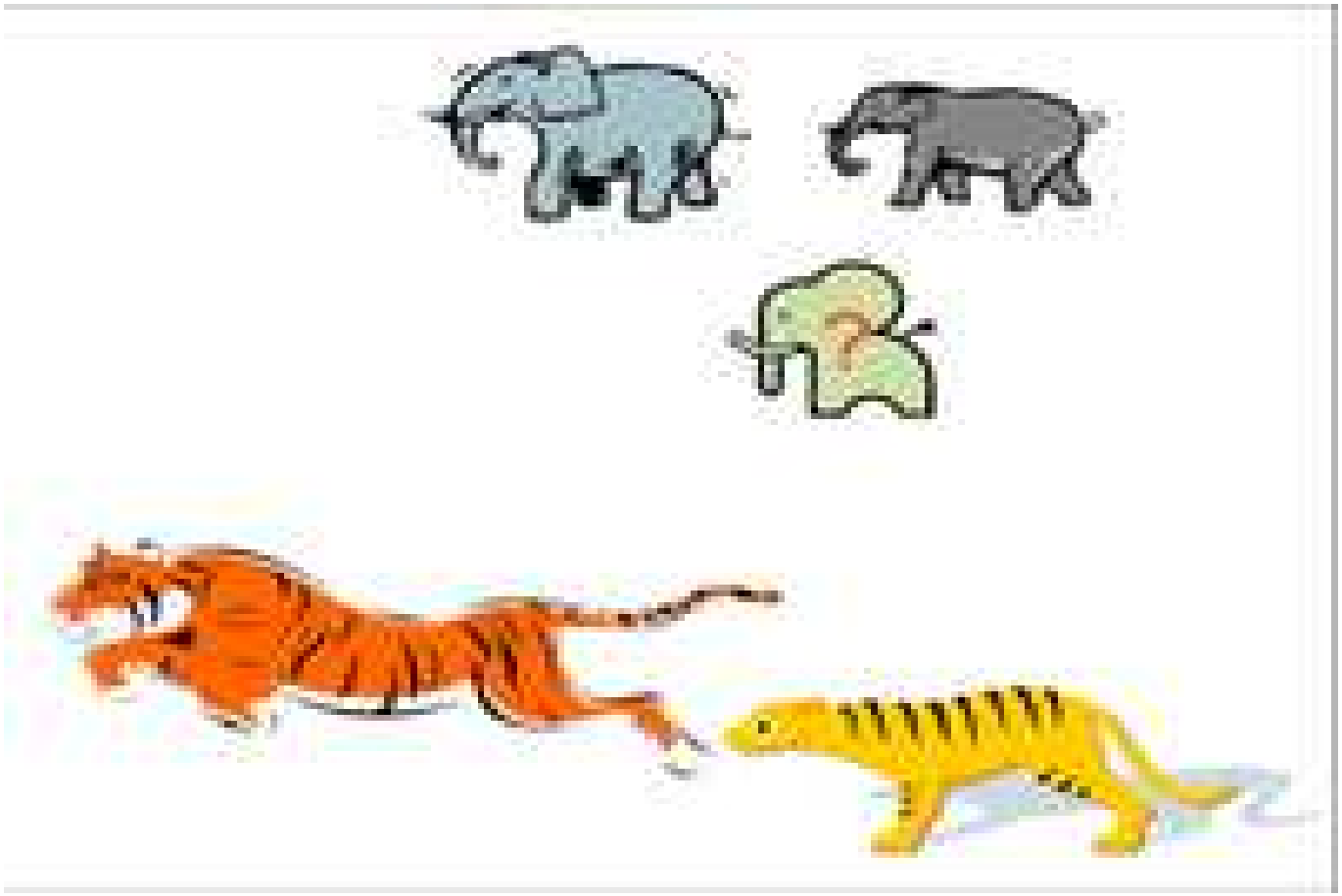
Seminar Talk, Aberystwyth, 5-th June 2017.

Motivation

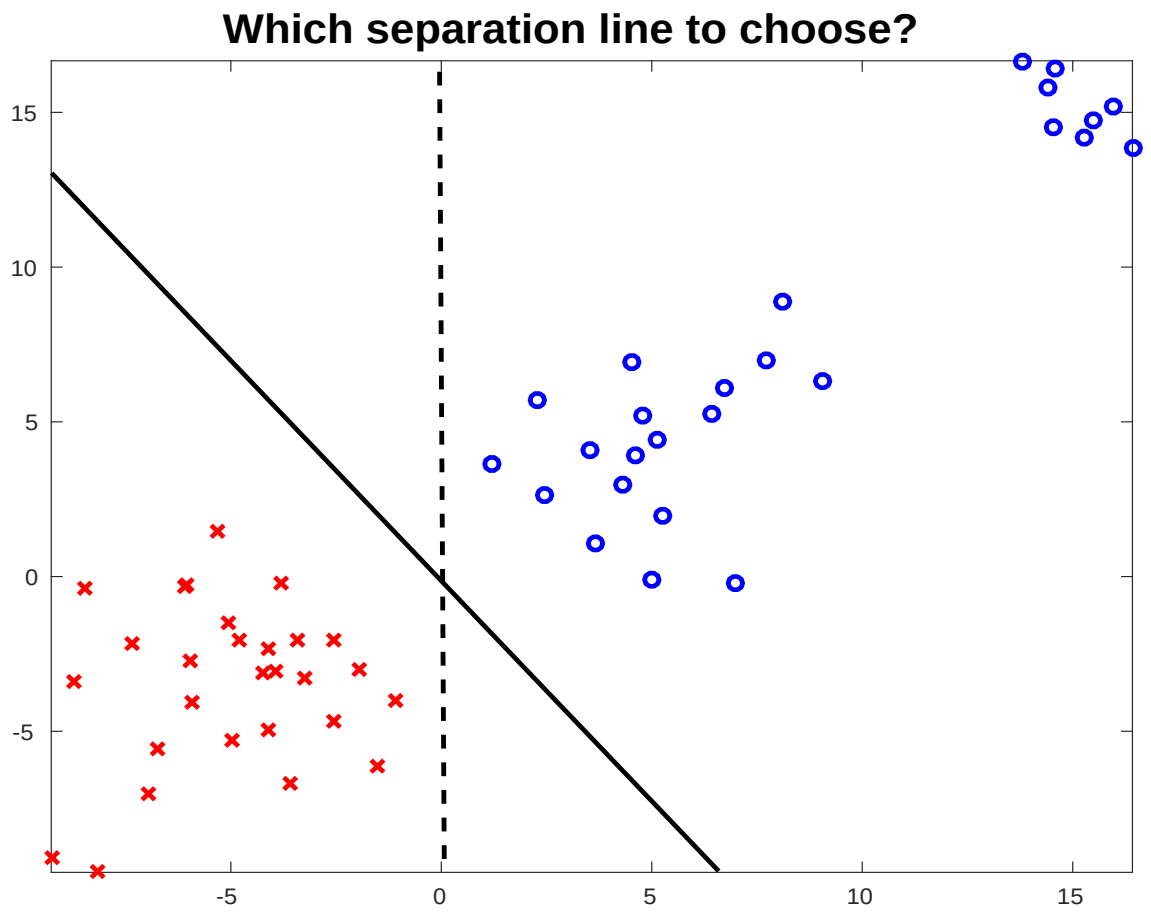
- Old Q: Given a machine learning task, what kinds of data distributions make it easier or harder?

- New Q: Given a high dimensional learning task, when can we solve it from a few random projections of the data with good-enough approximation?

Illustration for Old Q



But we knew this for classification...



Really?

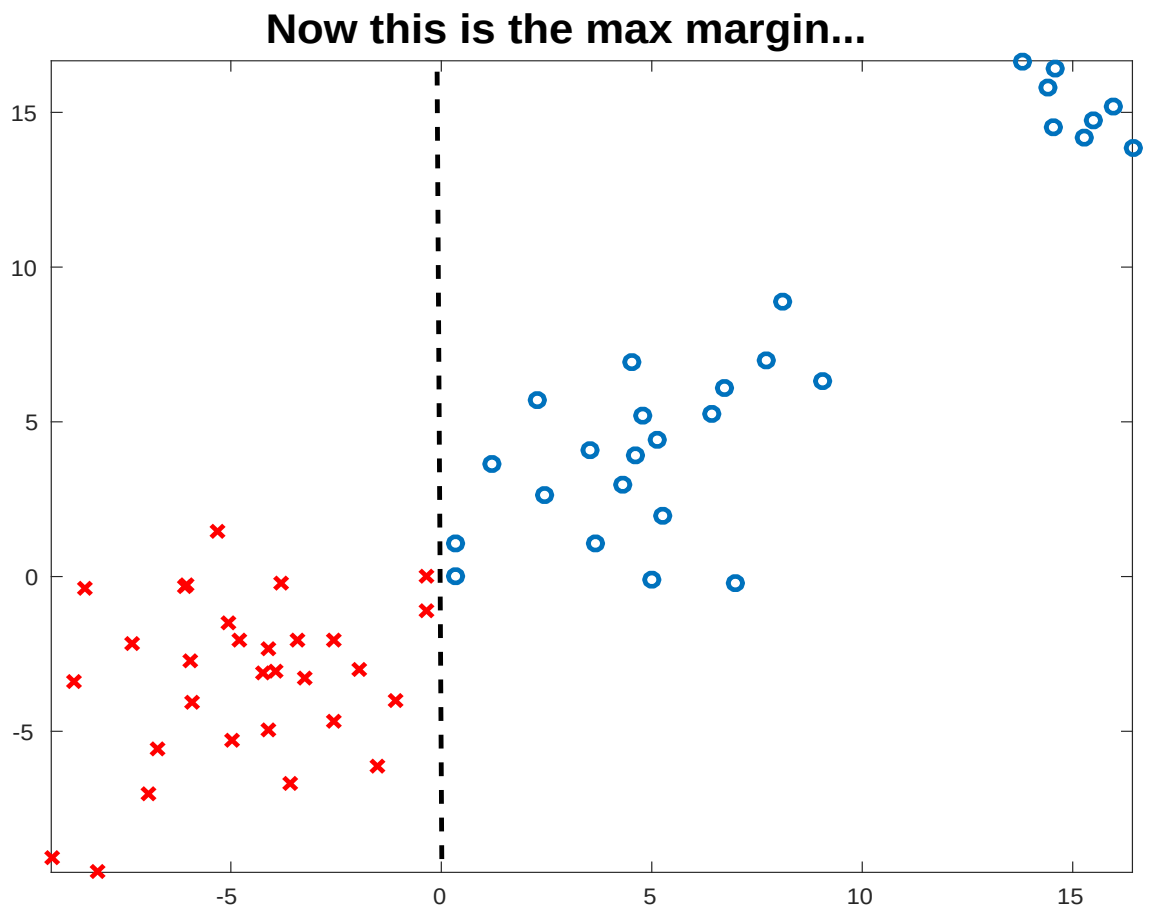


Illustration for New Q

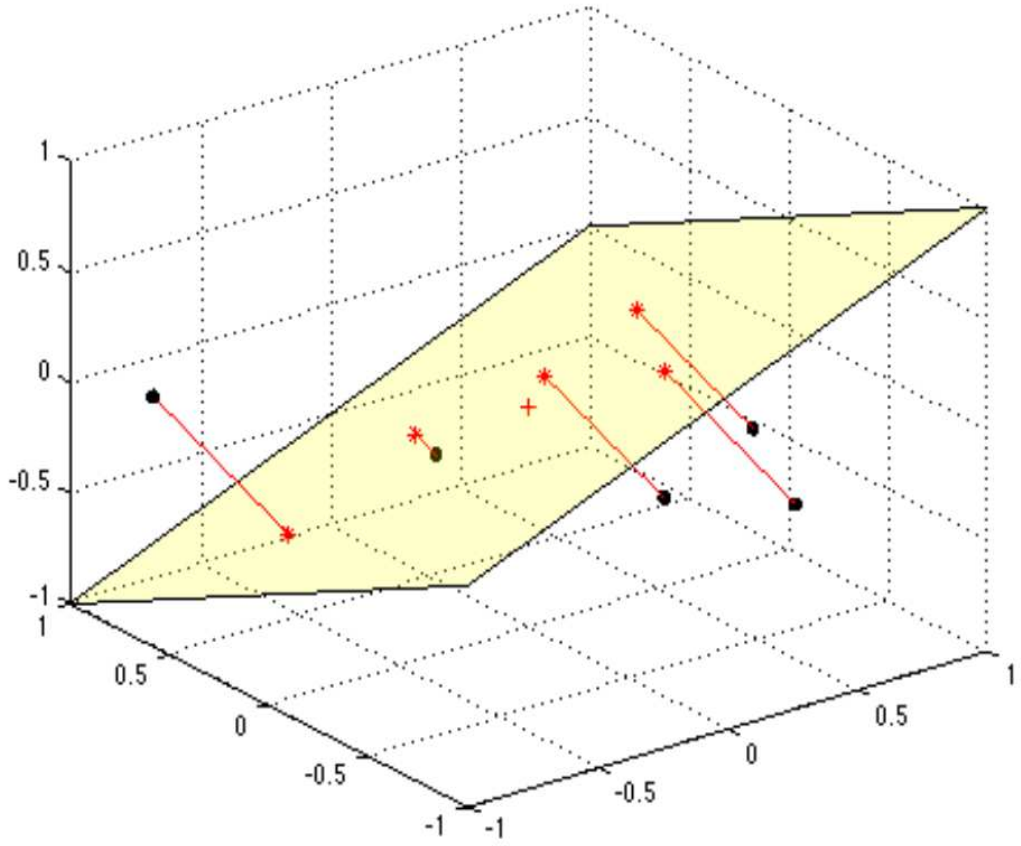
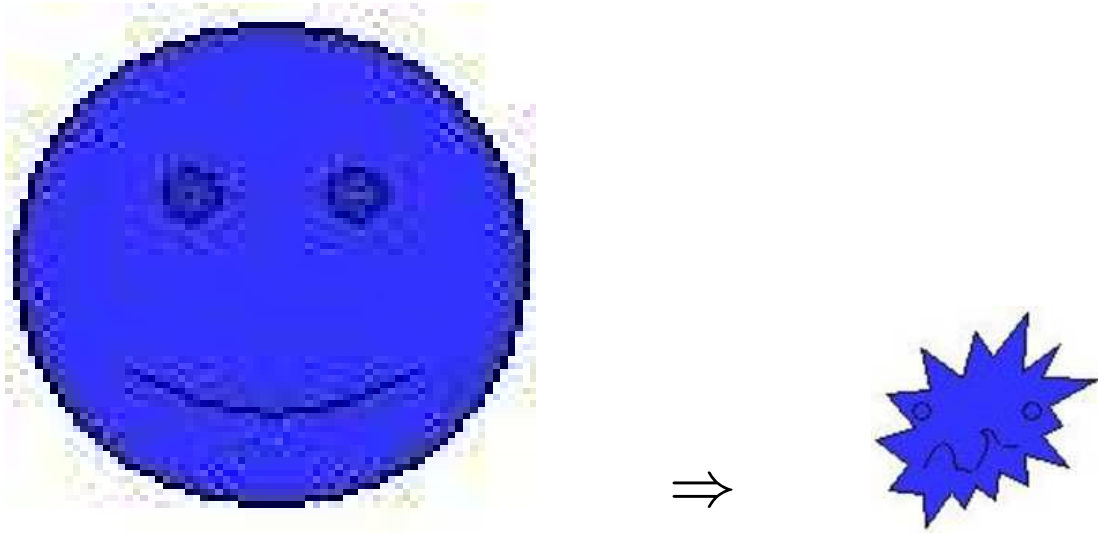


Illustration for New Q



High dimensional data **Random projection**

Background (1)

- Johnson-Lindenstrauss Lemma ensures that all pairwise Euclidean distances are preserved up to small distortion, if the reduced dimension $\geq \mathcal{O}(\log(\text{nr. points}))$.
  - Early seminal work by (Arriaga & Vempala 1999) on RP-perceptron relies on JLL
  - Similar approach in [Maillard & Munos, NIPS'09] for compressive OLS regression
  - BUT: For learning, a bound that loosens with $\geq \mathcal{O}(\log(\text{nr. points}))$ is unnatural.

Johnson-Lindenstrauss Lemma

The JLL is the following rather surprising fact:
Theorem[Johnson & Lindenstrauss, 1984] Let $\epsilon \in (0,1)$. Let $N, k \in \mathbb{N}$ such that $k \geq C\epsilon^{-2} \log N$, for a large enough absolute constant C. Let $V \subseteq \mathbb{R}^d$ be a set of N points. Then there exists a linear mapping $R : \mathbb{R}^d \to \mathbb{R}^k$, such that for all $u, v \in V$:

$$(1-\epsilon)\|u-v\|_{\ell_2^d}^2 \leq \|Ru - Rv\|_{\ell_2^k}^2 \leq (1+\epsilon)\|u-v\|_{\ell_2^d}^2$$

- With high probability *random projection* satisfies JLL [Dasgupta & Gupta '02] (proof by Chernoff bounding).

- The bound on k is essentially tight: $\forall N$, $\exists V$ s.t. $k \in \Omega(\epsilon^{-2} \log N / \log \epsilon^{-1})$ is required [Alon '03].

Background (2)

- The Restricted Isometry Property (RIP) in Compressed Sensing ensures that data that has a sparse representation can be recovered exactly from just a few of its random projections, if the reduced dimension $\geq \mathcal{O}(\text{nr. of nonzeros})$.
  - Compressive OLS regression for data that has a sparse representation [Fard et al, 2012]
  - Compressive SVM [Calderbank et al. 2009] - similar approach, bound holds only if data has sparse representation
  - BUT: Is sparse representation needed? In [K. AISTATS'2014] new bound for cOLS without sparse requirement.

Restricted Isometry Property

Definition. Let R be a $k \times d, k < d$ matrix and s an integer. The matrix R satisfies the RIP of order (s, δ) provided that, for all s-sparse vectors $x \in \mathbb{R}^d$:

$$(1 - \delta)\|x\|_2^2 \leqslant \|Rx\|_2^2 \leqslant (1 + \delta)\|x\|_2^2$$

One can show [Baraniuk '07] that random projection matrices satisfying the JLL w.h.p also satisfy the RIP w.h.p provided that $k \in \mathcal{O}(s \log d)$. (Proof: JLL + covering + union bound over subspaces of dimension k)

Theorem[Candès & Tao, 2006] If $x \in \mathbb{R}^d$ has a sparse representation with s non-zeros and R satisfies RIP of order $(2s, \delta_{2s})$, then $y := Rx$ one can recover x exactly by $\hat{x} = \arg\min_x \{\|x\|_1 : y = Rx\}$.

Background (3)

Work that looks at preservation / non-preservation of margin after a random projection:

- Large margin implies 'low dimension of the problem' [Balcan & Blum, MLJ 2006]
- Is margin preserved? [Shi, Shen, Hill & Hengel, ICML 2012]
  - The doubt / controvery on preservation of obtuse angles is now resolved [K. KDD2015]
- BUT: Is there anything more general than margin?

Dot product under RP

Theorem [K. KDD'2015] Let $x, y \in \mathbb{R}^d$. Let $R \in \mathcal{M}^{k \times d}$, $k < d$, be a random projection matrix having i.i.d. 0-mean subgaussian entries with parameter $\sigma^2 = 1/k$, and let Rx, $Ry \in \mathbb{R}^k$ be the images of x, y under R. Then, $\forall \epsilon \in (0, 1)$:

$$\Pr\{(Rx)^T Ry < x^T y - \epsilon \cdot \|x\| \cdot \|y\|\} < \exp\left(-\frac{k\epsilon^2}{8}\right) \quad (1)$$

$$\Pr\{(Rx)^T Ry > x^T y + \epsilon \cdot \|x\| \cdot \|y\|\} < \exp\left(-\frac{k\epsilon^2}{8}\right) \quad (2)$$

Corollaries: Clarifying the role of angle

Corollary [K. KDD2015] Denote by θ the angle between the vectors $x, y \in \mathbb{R}^d$. Then we have the following:

1. Relative distortion bound: Assume $x^T y \neq 0$. Then,

$$\Pr\left\{ |\frac{x^T R^T R y}{x^T y} - 1| > \epsilon \right\} < 2 \exp\left(-\frac{k}{8} \epsilon^2 \cos^2(\theta) \right) \tag{3}$$

2. Dot product under random sign projection: Assume $x^T y \neq 0$. Then,

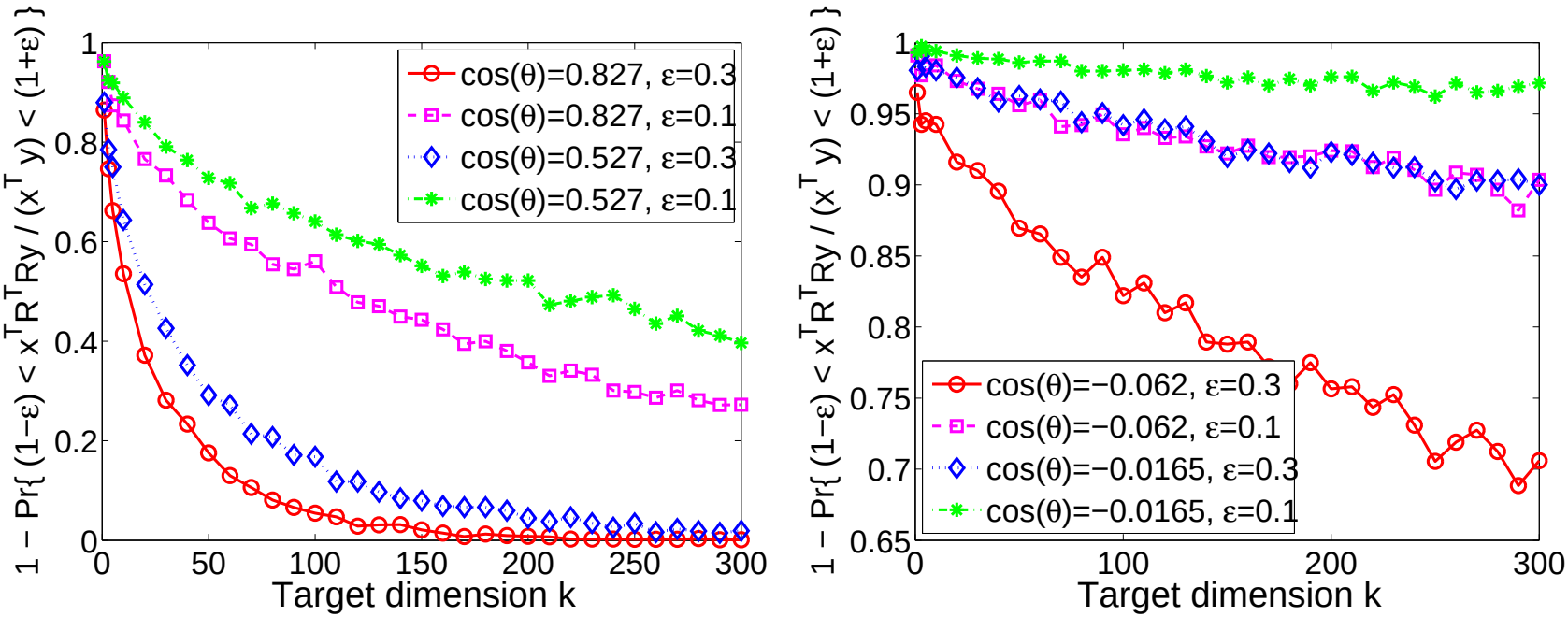
$$\Pr\left\{ \frac{x^T R^T R y}{x^T y} < 0 \right\} < \exp\left(-\frac{k}{8} \cos^2(\theta) \right) \tag{4}$$

Experimental corroboration

We will compute empirical estimates of the following probabilities, from 2000 independently drawn instances of the RP. The target dimension varies from 1 to the original dimension $d = 300$.

- Rejection probability for dot product preservation = Probability that the relative distortion of the dot product after RP falls outside the allowed error tolerance ϵ:

$$1 - Pr\left\{(1-\epsilon) < \frac{(Rx)^T Ry}{x^T y} < (1+\epsilon)\right\} \tag{5}$$



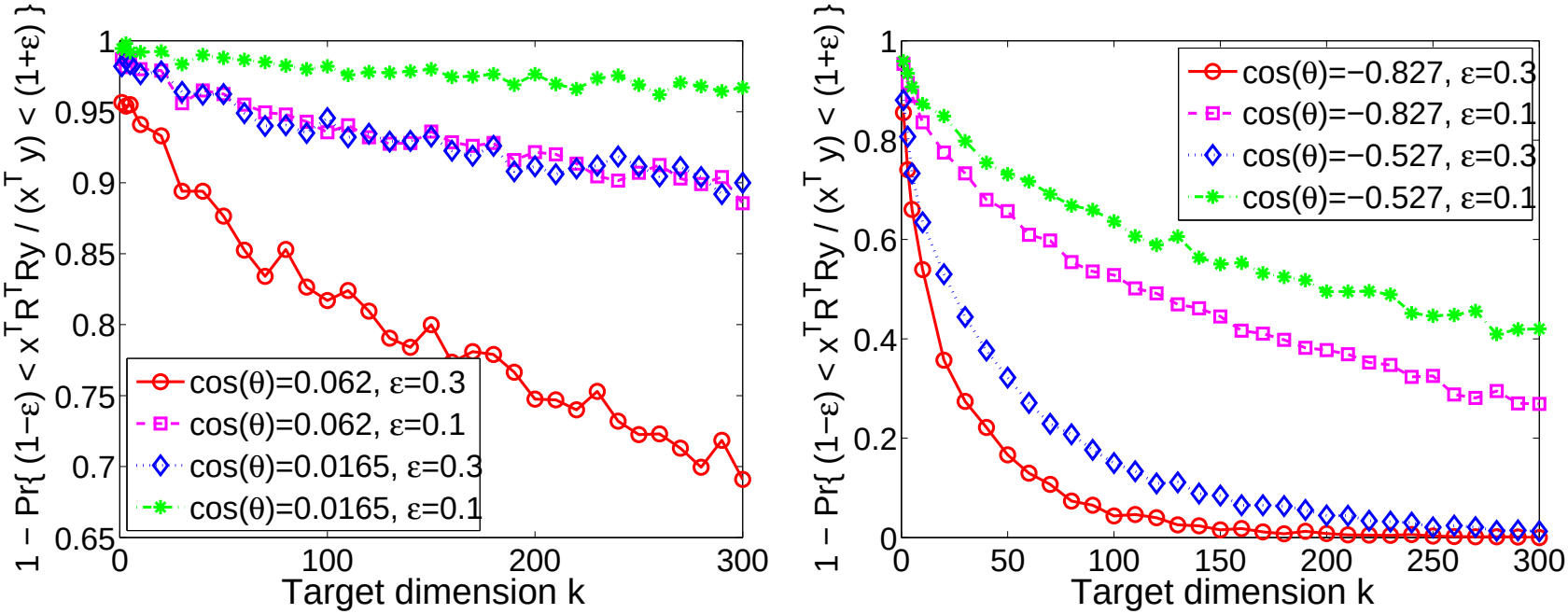
Replicating the results in [Shi et al, ICML'12].

Left: Two acute angles; *Right*: Two obtuse angles.

Preservation of these obtuse angles looks indeed worse...

...but not because they are obtuse (see next slide!).

Now take the angles symmetrical around $\pi/2$ and observe the opposite behaviour. – this is why the previous result in [Shi et al, ICML'12] has been misleading.



Left: Two acute angles; *Right*: Two obtuse angles.

Roadmap for rest of the talk

- Discovering benign structures & learning from RP data as two sides of the same coin
  - linear classification
  - unconstrained nonparametric classification
- Addendum: Ensembles of compressive classifiers

Compressive Linear classification

Training set $\mathcal{T}^N = \{(x_i, y_i)\}_{i=1}^N$; $(x_i, y_i) \overset{\text{i.i.d}}{\sim} \mathcal{D}$ over $\mathbb{R}^d \times \{0, 1\}$.
Let $\hat{h} \in \mathcal{H}$ be the ERM linear classifier. So $\hat{h} \in \mathbb{R}^d$, and w.l.o.g. we take it to pass through the origin, and can take that all data lies on $S^{d-1} \subseteq \mathbb{R}^d$ and $\|\hat{h}\| = 1$.

For an unlabelled query point x_q the label returned by $\hat{h}$ is then:

$$\hat{h}(x_q) = \mathbf{1}\left\{\hat{h}^T x_q > 0\right\}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function.

The risk (generalisation error) of $\hat{h}$ is defined as $\mathsf{E}_{(x_q,y_q)\sim\mathcal{D}}[\mathcal{L}(\hat{h}(x_q), y_q)]$, and we use the $(0, 1)$-loss:

$$\mathcal{L}_{(0,1)}(\hat{h}(x_q), y_q) = \begin{cases} 0 & \text{if } \hat{h}(x_q) = y_q \\ 1 & \text{otherwise.} \end{cases}$$

Random projection: $R \in \mathcal{M}_{k\times d}$, $k \ll d$, with entries $r_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. Pre-multiply the data points with it: $\mathcal{T}_R^N = \{(Rx_i, y_i)\}_{i=1}^N$.

Denote the trained classifier by $\hat{h}_R \in \mathbb{R}^k$ (possibly not through the origin, but translation does not affect our proof technique) The label returned by $\hat{h}_R$ is therefore:

$$\hat{h}_R(Rx_q) = \mathbf{1}\left\{\hat{h}_R^T Rx_q + b > 0\right\}$$

where $b \in \mathbb{R}$.

We want to estimate the generalisation error of the ERM linear classifier trained on $\mathcal{T}_R^N$ rather than $\mathcal{T}^N$:

$$\mathsf{E}_{(x_q,y_q)\sim\mathcal{D}}\left[\mathcal{L}_{(0,1)}(\hat{h}_R(Rx_q), y_q)\right] = \mathsf{Pr}_{(x_q,y_q)\sim\mathcal{D}}\left\{\hat{h}_R(Rx_q) \neq y_q\right\}$$

with high probability w.r.t the random choice of $\mathcal{T}_N$ and R.

Theorem [D-K, ICML'2013] For all $\delta \in (0, 1]$, with probability at least $1 - 2\delta$,

$$\Pr_{x_q, y_q}\{\hat{h}_R(Rx_q) \neq y_q\} \leq \hat{E}(\mathcal{T}^N, \hat{h}) + \frac{1}{N}\sum_{i=1}^{N} f_k(\theta_i)$$

$$+ \quad \min\left\{\sqrt{3\log\frac{1}{\delta}}\sqrt{\frac{1}{N}\sum_{i=1}^{N} f_k(\theta_i)}, \frac{1-\delta}{\delta}\cdot\frac{1}{N}\sum_{i=1}^{N} f_k(\theta_i)\right\} + 2\sqrt{\frac{(k+1)\log\frac{2eN}{k+1} + \log\frac{1}{\delta}}{N}}$$

where $f_k(\theta_i) := \Pr_R\{\text{sign}\,(\hat{h}R^T Rx_i) \neq \text{sign}\,(\hat{h}^T x_i)\}$ is the flipping probability for x_i with θ_i the principal angle between $\hat{h}$ and x_i, and $\hat{E}(\mathcal{T}^N, \hat{h})$ is the empirical risk of the data space classifier.

Also, if h^* is the optimal linear classifier in $\mathbb{R}^d$ then $\forall \delta \in (0,1]$, w.p. at least $1-2\delta$, denoting $\theta^*_x = \angle(x, h^*)$:

$$\mathsf{Pr}_{x_q,y_q}\{\hat{h}_R(Rx_q) \neq y_q\} \leq \mathsf{Pr}_{x_q,y_q}\{h^*(x_q) \neq y_q\} + \mathsf{E}_{x_q}[f_k(\theta^*_x)]$$

$$+ \quad \min\left\{\sqrt{3\log\frac{1}{\delta}}\sqrt{\mathsf{E}_{x_q}[f_k(\theta^*_{x_q})]}, \frac{1-\delta}{\delta}\cdot\mathsf{E}_{x_q}[f_k(\theta^*_{x_q})]\right\} + 4\sqrt{\frac{(k+1)\log\frac{2eN}{k+1} + \log\frac{1}{\delta}}{N}}$$

Proof.(sketch) For a fixed instance of R, from classical VC theory we have $\forall \delta \in (0,1)$ w.p. $1-\delta$ over $\mathcal{T}^N$,

$$\Pr_{x_q,y_q}\{\hat{h}_R(Rx_q) \neq y_q\} \leq \hat{E}(\mathcal{T}_R^N, \hat{h}_R) + 2\sqrt{\frac{(k+1)\cdot\log(2eN/(k+1)) + \log(1/\delta)}{N}}$$

where $\hat{E}(\mathcal{T}_R^N, \hat{h}_R) = \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{\hat{h}_R(Rx_i) \neq y_i\}$ the empirical error. We see RP reduces the complexity term but will increase the empirical error. We bound the latter further:

$$\hat{E}(\mathcal{T}_R^N, \hat{h}_R) \leq \ldots \leq \underbrace{\frac{1}{N}\sum_{i=1}^N \mathbf{1}\{sign((R\hat{h})^T Rx_i) \neq sign(\hat{h}^T x_i)\}}_{S} + \hat{E}(\mathcal{T}^N, \hat{h})$$

Now, bound S from $\mathrm{E}_R[S]$ w.h.p, w.r.t. the random choice of R.

The terms $\mathrm{E}_R[S]$ represent the probability of label flipping.

Theorem [Upper Bound on Generalisation Error in Data Space]

Let $\mathcal{T}^{2N} = \{(x_i, y_i)\}_{i=1}^{2N}$ be a set of d-dimensional labelled training examples drawn i.i.d. from some data distribution $\mathcal{D}$, and let $\hat{h}$ be a linear classifier estimated from $\mathcal{T}^{2N}$ by ERM. Let $k \in \{1, 2, \ldots, d\}$ be an integer and let $R \in \mathcal{M}_{k \times d}$ be a random projection matrix, with entries $r_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. Then for all $\delta \in (0, 1]$, with probability at least $1 - 4\delta$ w.r.t. the random draws of $\mathcal{T}^{2N}$ and R the generalisation error of $\hat{h}$ w.r.t the (0,1)-loss is bounded above by:

$$\text{Pr}_{x_q, y_q}\{\hat{h}^T x_q \neq y_q\} \leq \hat{E}(\mathcal{T}^{2N}, \hat{h}) + 2 \cdot \min_k \left\{ \frac{1}{N} \sum_{i=1}^{2N} f_k(\theta_i) \ldots \right.$$

$$+ \quad \min \left\{ \sqrt{3 \log \frac{1}{\delta}} \sqrt{\frac{1}{N} \sum_{i=1}^{2N} f_k(\theta_i)}, \frac{1-\delta}{\delta} \cdot \frac{1}{N} \sum_{i=1}^{2N} f_k(\theta_i) \right\} + \left. \sqrt{\frac{(k+1) \log \frac{2eN}{k+1} + \log \frac{1}{\delta}}{2N}} \right\}$$

Theorem [Sign Flipping Probability - case of Gaussian RP]
Let R be a RP matrix with entries $r_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$, let $h, x \in \mathbb{R}^d$, and let θ be the angle between them. Let $Rh, Rx \in \mathbb{R}^k$ be the images of x, y under R.

1. Exact form:

$$\Pr\{(Rh)^T Rx < 0\} = \frac{\Gamma(k)}{(\Gamma(k/2))^2} \int_0^{\psi} \frac{z^{(k-2)/2}}{(1+z)^k} \mathrm{d}z \tag{6}$$

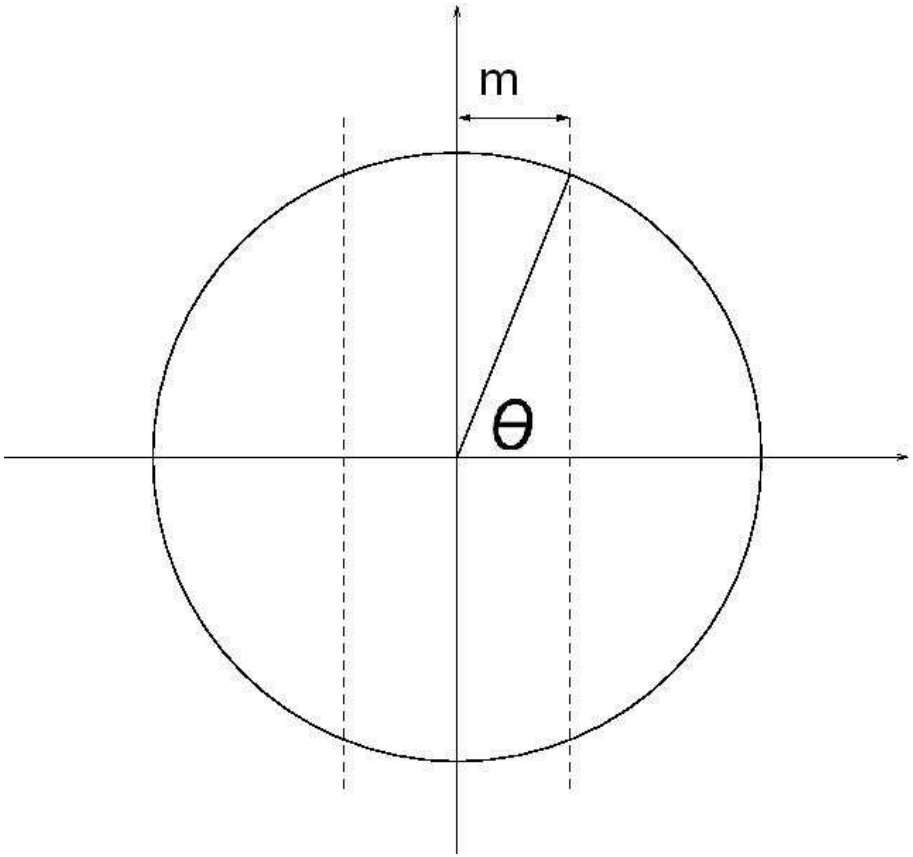
   where $\psi = (1 - \cos(\theta))/(1 + \cos(\theta))$.

2. If $\cos(\theta) > 0$, we have the following upper-bound:

$$\Pr\{(Rh)^T Rx < 0 | h^T x > 0\} \leq \exp(-k \cos^2(\theta)/2) \tag{7}$$

Using [K., KDD'2015] it is possible to replace with computationally cheaper sub-Gaussians for the slightly worse constant 8 replacong the 2 in eq.(7).

Relation of Sign Flipping Probability vs Margin



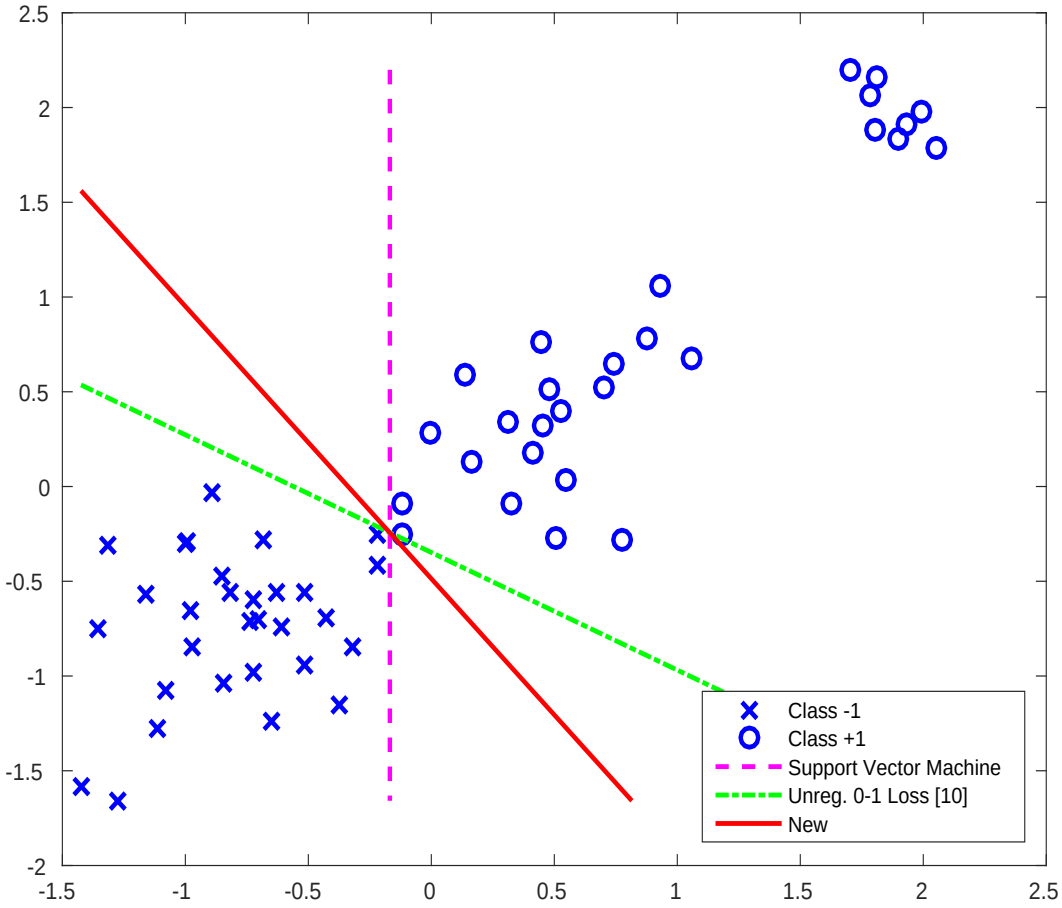
Flip probability and Margins

$f_k(\theta) \leq \exp(-\frac{1}{8}k\cos^2(\theta))$

$\cos(\theta) = m$

Hence, low flip probabilities imply large margins of points.

Optimising a RP-based dataspace bound



Optimising a RP-based dataspace bound

Data set	New	SVM	p-value
Australian	**0.137± 0.015**	0.148± 0.013	0.0002
German	**0.260± 0.018**	0.280± 0.016	< 0.0001
Haberman	**0.265± 0.025**	0.285± 0.050	0.0136
Parkinsons	**0.141± 0.032**	0.221± 0.049	< 0.0001
PlanningRelax	**0.285± 0.029**	0.361± 0.166	0.0024
Sonar	0.256± 0.045	0.271± 0.036	0.0689

Test error rates $\pm$ std for the bound optimizer in comparison with the SVM. Bold font indicates significant improvement at the 0.05 level cf. a paired t-test.

Summing up linear classification

For linear classifiers trained by ERM on i.i.d. traiing sample (and no other assumptions a-priori), the use of RP revealed:

- The task is solvable in a random linear subspace (i.e. with performance guarantees) if the label flipping probabilities under a RP are small. This requirement is more general than large margin.
- The dataspace ERM classifier's error is small under the same conditions.
- Note, we did not require any sparse representation for our bounds to hold, as usually compressed learning approaches do.

Roadmap for rest of the talk

- Discovering benign structures & learning from RP data as two sides of the same coin
  - linear classification
  - unconstrained nonparametric classification
- Addendum: Ensembles of compressive classifiers

An unconstrained classifier: Nearest Neighbour

NN performs general learning of an unrestricted function class. Because of this, some sort of smoothness of the label posterior probability is known to be needed for learnability.

Let $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ be a training set drawn i.i.d. from some unknown distribution $\mathcal{D}$ over the input-output domain $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} = \{0, 1\}$ for classification problems, and we take $\mathcal{X} = [-1, 1]^d$.

Denote by $\eta : \mathbb{R}^d \to \mathbb{R}$ the true conditional probability of the labels, i.e. $\eta(x) = \Pr(Y = 1 | X = x)$. Since we consider unconstrained general learning of an unconstrained function class,

some form of Lipschitz-like assumption is known to be needed on $\eta(\cdot)$ for learnability.

Nearest neighbours classifier of S will be denoted as h_S. Given an input point $x \in \mathcal{X}$ it looks up its nearest neighbour, denoted $N(x) \in S$ it returns its label, $h_S(x) = Y_{N(x)}$.

The generalisation error of h_S is defined as
$err(h_S) = \mathsf{E}_{(x,y)\sim\mathcal{D}}[h_S(x) \neq y]$,
where (x, y) is a query point drawn independently from and identically distributed as the training points.

The Bayes-optimal classifier will be denoted as h^*.

Known result on Sample Complexity of NN

Theorem[Shalev-Schwarz & Ben-David 2014] Let $\mathcal{X} = [0,1]^d, \mathcal{Y} = \{0,1\}$, and $\mathcal{D}$ a distribution over $\mathcal{X} \times \mathcal{Y}$ for which the conditional probability function is L-Lipschitz. Let h_S denote the nearest neighbour rule applied to the training set $S \sim \mathcal{D}^N$. Then,

$$\mathsf{E}_S[err(h_S)] \leq 2err(h^*) + 4L\sqrt{d}N^{-\frac{1}{d+1}} \tag{8}$$

This implies the sample complexity

$$N \geq \left(\frac{4L\sqrt{d}}{\epsilon}\right)^{d+1} \in \tilde{\Omega}(\exp(d)) \tag{9}$$

to guarantee $\mathsf{E}_S[err(h_S)] \leq 2err(h^*) + \epsilon$.

Tools

Definition[Packing number] Let $(T, \|\cdot\|)$ be a totally bounded pseudo metric space. Let $\alpha > 0$. We say that T is α-separated if $\forall a, b \in T, a \neq b, \|a, b\| \geq \alpha$.
The *α-packing number of T is defined as the maximum cardinality of the α-separated subsets of T, denoted as:*
$N_{\|\cdot\|}(\alpha, T) = \max\{|T'| : T' \text{ is } \alpha\text{-separable}, T' \subset T\}$. When the pseudometric is clear from the context we can omit the subscript.

Definition[α-entropy number] The α-entropy number of T is defined as the log of the packing number, $H(\alpha, T) = \log N(\alpha, T)$.

Definition[Metric entropy] The function $H(\cdot, T)$ is called the metric entropy of T.

Theorem [Klartag & Mendelson '06] Let $\mathcal{X} \subset \mathbb{R}^d$. Let R be a $k \times d, k < d$ random projection matrix with i.i.d. Gaussian or Rademacher entries with mean 0 and variance σ^2. Consider the set of all normalised chords of $\mathcal{X}$: $T = \left\{ \frac{a-b}{\|a-b\|} : a, b \in \mathcal{X} \right\}$, with $\|\cdot\|$ being the Euclidean distance, and define the metric entropy integral

$$\gamma(T) = \int_0^1 \sqrt{H(\alpha, T)} d\alpha \tag{10}$$

where $H(\alpha, T)$ is the α-entropy number of T w.r.t. the Euclidean distance.

Then, $\exists c$ absolute constant s.t. $\forall \zeta, \delta \in (0, 1)$, if

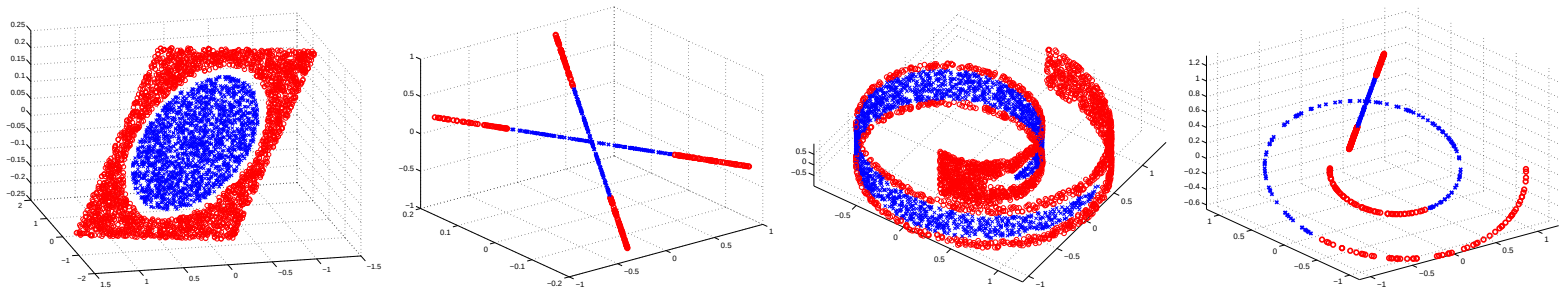
$$k \geq c\zeta^{-2}(\gamma^2(T) + \log(2/\delta)) \tag{11}$$

then R is an ζ-isometry on $\mathcal{X}$ with high probability, i.e. with probability at least $1-\delta$ we have:

$$(1-\zeta)k\sigma^2\|x-x'\|^2 \leq \|Rx-Rx'\|^2 \leq (1+\zeta)k\sigma^2\|x-x'\|^2, \forall x, x' \in \mathcal{X}$$

- Generalises of the Johnson-Lindenstrauss lemma to infinte sets of points $\mathcal{X}$ as long as $\gamma(T)$ is finite.
  - When $\mathcal{X}$ is a finite set of N points it recovers Johnson-Lindenstrauss: $\gamma^2(T) \in \mathcal{O}(\log(N))$.
  - Also recovers the Restricted Isometry Property from Compressed Sensing as a special case: for s-sparse vectors $\gamma^2(T) \leq 2s\log(d/(2s))$.
  - Other low complexity structures include certain smooth manifolds, and metric spaces with finite doubling dimension.

- Examples of low complexity input domains: linear subspace of the input domain, certain smooth nonlinear subspaces, domains that have a sparse representation, etc.



- We shall see that such structures (including sparsity) do help NN.

Compressive NN

Let R be a $k \times d, k < d$ random matrix with i.i.d. subgaussian entries, e.g. a Gaussian or a Rademacher distribution.

Let $S_R = \{(Rx_1, y_i), \ldots, (Rx_N, y_N))\}$ (compressive training set).

Compressive NN receives S_R, and will be denoted by $h^R_{S_R}$.

We are interested in the distribution of its expected generalisation error:

$$\mathsf{E}_{S\sim\mathcal{D}^N}[err(h^R_{S_R})] = \mathsf{E}_S[\mathsf{E}_{(x,y)\sim\mathcal{D}}[h^R_{S_R}(Rx) \neq y]]$$

as a random function of R.

Generalisation of compressive-NN

Theorem [K. ACML'2015] Let $\mathcal{X} = \mathcal{B}(0, \rho) \subset \mathbb{R}^d$ the ball of radius ρ centered at 0, and $\mathcal{D}$ a distribution over $\mathcal{X} \times \mathcal{Y}$ for which the conditional probability function is L-Lipschitz. Let R be a $k \times d$ RP matrix; let $h^R_{S_R}$ the nearest neighbour rule on RP data S_R where $S \sim \mathcal{D}^N$. Then $\forall \delta, \zeta \in (0, 1)$, with probability at least $1 - \delta$ over the random draws of R, the sample size required to guarantee $\mathsf{E}_S[err(h^R_{S_R})] \leq 2err(h^*) + \epsilon$ w.p. $1 - \delta$ is

$$\begin{aligned} N &\geq \frac{1}{e}\left(\frac{2\sqrt{2}\sqrt{k}}{\epsilon}\right)^{k+1}\left(L\rho\sqrt{\frac{1+\zeta}{1-\zeta}}\right)^{k} \\ &= \tilde{\Omega}(exp(\gamma(T)) \end{aligned}$$

provided that $k \in \Omega(\zeta^{-2}(\gamma^2(T) + \log(2/\delta)))$.

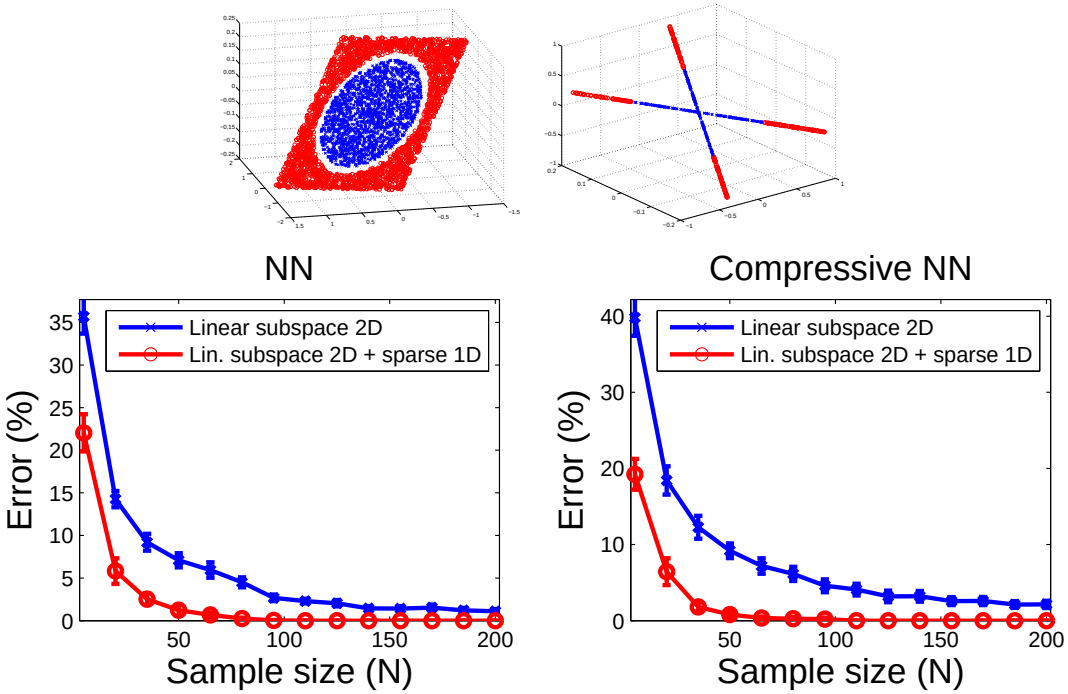
Implication: Dataspace NN is no worse

Corollary Let $\mathcal{X} = \mathcal{B}(0, \rho) \subset \mathbb{R}^d, \mathcal{Y} = \{0, 1\}$, and $\mathcal{D}$ a distribution over $\mathcal{X} \times \mathcal{Y}$ for which the conditional probability function is L-Lipschitz. h_S denote the nearest neighbour rule, where $S \sim \mathcal{D}^N$. For any $\delta, \zeta \in (0, 1)$, The sample size required to guarantee $\mathsf{E}_S[err(h_S)] \leq 2err(h^*) + \epsilon$ w.p. $1 - \delta$ is

$$\begin{aligned} N &\geq \frac{1}{e}\left(\frac{2\sqrt{2}\sqrt{k}}{\epsilon}\right)^{k+1}\left(L\rho\sqrt{\frac{1+\zeta}{1-\zeta}}\right)^{k} \\ &= \tilde{\Omega}(exp(\gamma(T)) \end{aligned}$$

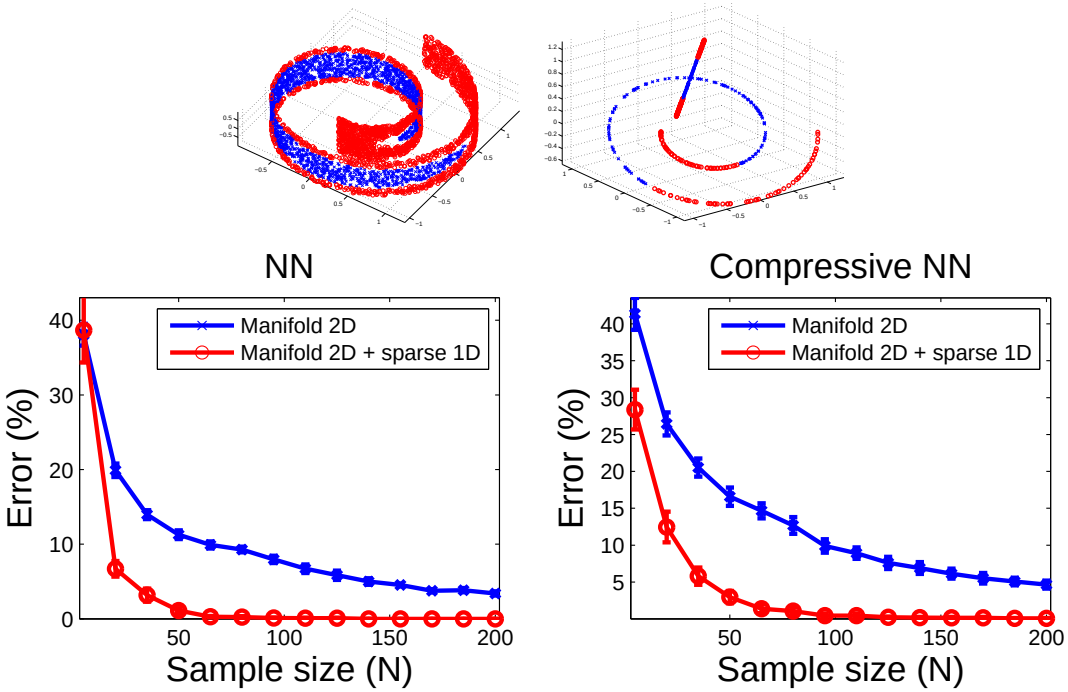
Proof [of Corollary] $\|x - N(x)\|_2 \leq \|x - N_R(x)\|_2$, where $N(x) \in S$ is the NN of x, and $N_R(x)$ is the $x' \in S$ s.t. Rx' is the NN of Rx after RP. $\square$

Empirical results - 100-D data on 2D linear subspace / & sparse



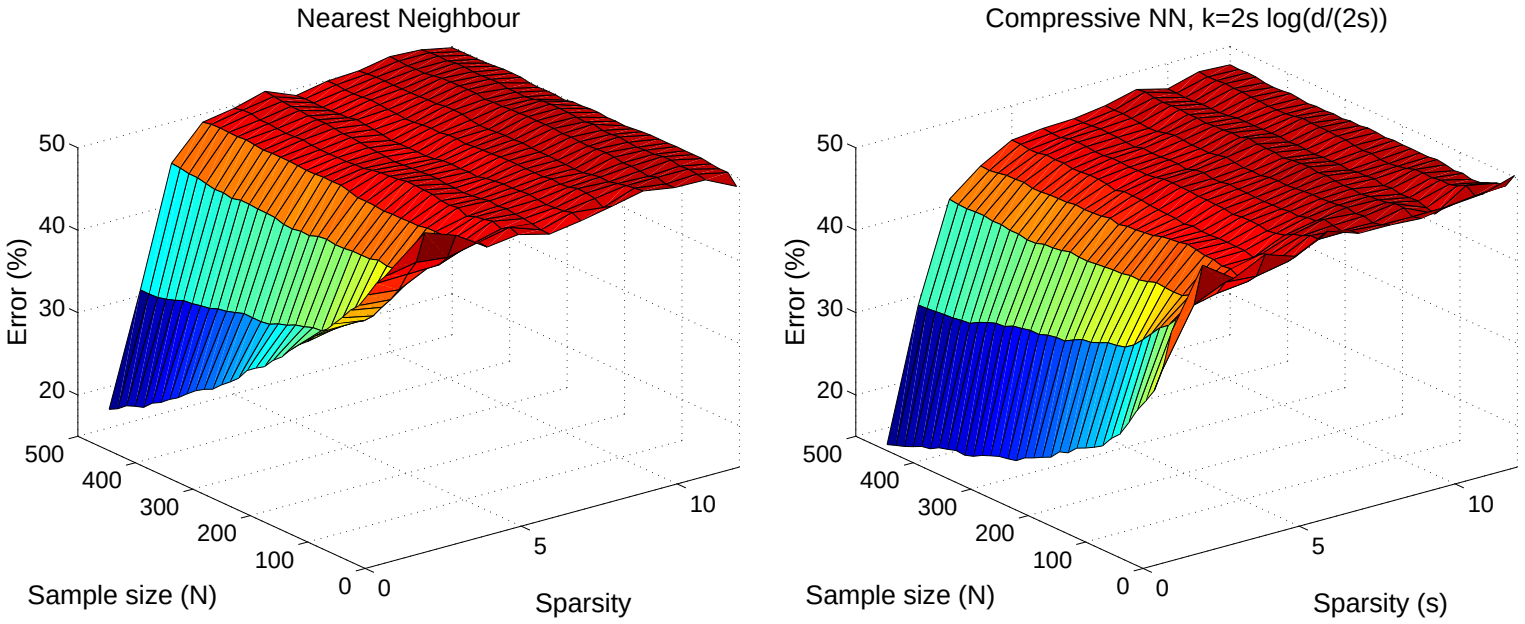
NN and Compressive NN have very similar error behaviour; Sparse representation of the input data lowers the error.

Empirical results - 100-D data on 2D nonlinear subspace / & sparse



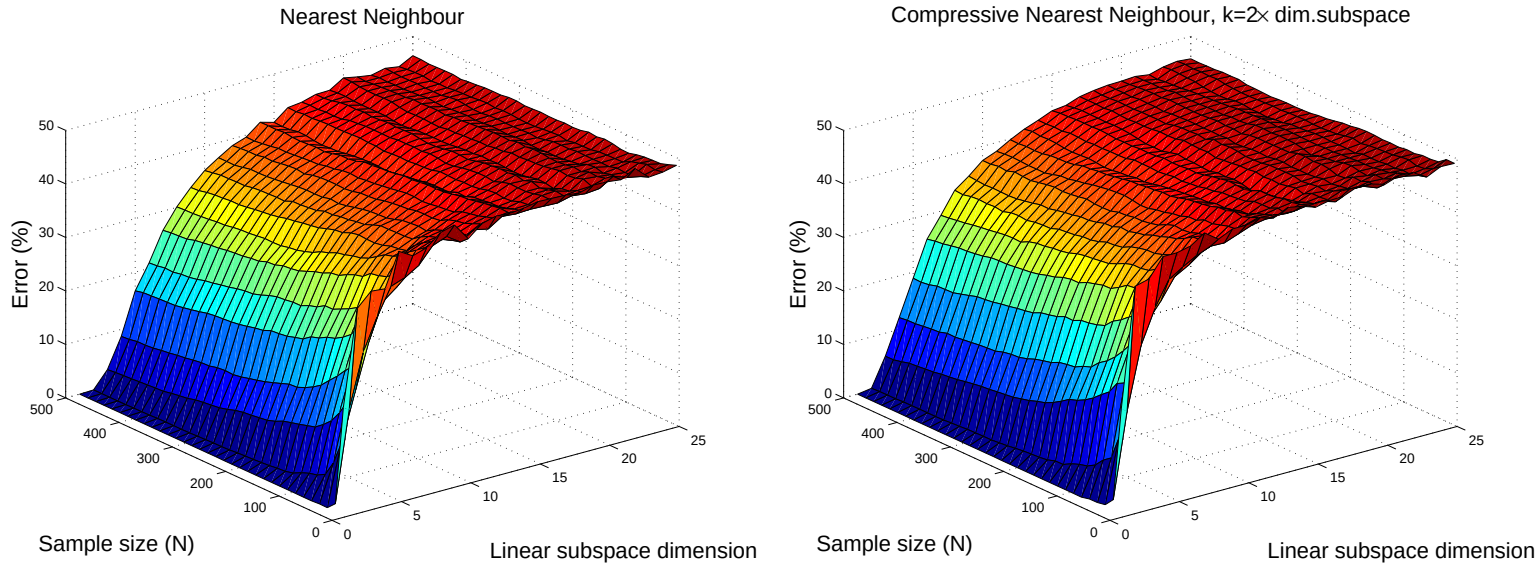
NN and Compressive NN have very similar error behaviour; Sparse representation of the input data lowers the error.

Empirical results - Effect of sparse representation



$$k = 2s \log(d/(2s))$$

Empirical results - All points on the same linear subspace



Subspace dimension varied in the same range as the sparsity in the previous experiment.

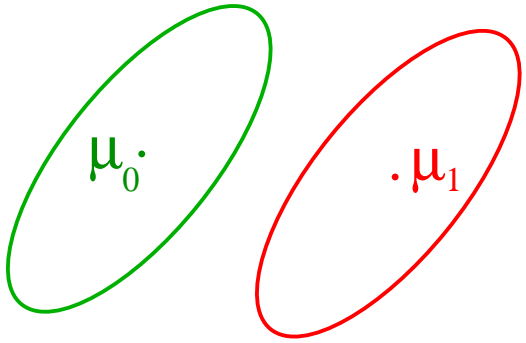
$$k = 2s$$

Addendum - Ensembles of compressive learners

We will look at a very specific ensemble, for problems with less training observations than data dimensions.
The base learners will be Fisher Linear Discriminants, and the combination rule is simple averaging.

- Can we achieve (or improve on) the classification performance in data space, using the RP FLD ensemble?
- Can we understand how the RP FLD ensemble acts to improve performance?
- Can we interpret the RP ensemble classifier parameters in terms of data space parameters?

Fisher's Linear Discriminant (FLD)



- Simple and popular linear classifier, in widespread application. Classes are modelled as identical multivariate Gaussians.
- Assign class label to any query point according to its Mahalanobis distance from the class means.
- Simple enough to allow a deeper analysis addressing our questions.

RP-FLD classifier ensemble

Training set $\mathcal{T} = \{(\mathbf{x}_i, y_i) : (\mathbf{x}, y) \in \mathbb{R}^d \times \{0, 1\}\}_{i=1}^N$ of N real-valued d-dimensional points. Two-class classification setting.
Assume that $N \ll d$, which is a common situation e.g. medical imaging, genomics, proteomics, etc.

For a single RP FLD classifier, the decision rule is given by:

$$\mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T \left(R\hat{\Sigma}R^T\right)^{-1} R\left(x_q - \frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right) > 0\right\}$$

which is the randomly projected analogue of the FLD decision rule. For the ensemble we use an equally weighted linear combination of RP FLD classifiers:

$$\mathbf{1}\left\{\frac{1}{M}\sum_{i=1}^{M}(\hat{\mu}_1 - \hat{\mu}_0)^T R_i^T \left(R_i\hat{\Sigma}R_i^T\right)^{-1} R_i\left(x_q - \frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right) > 0\right\} \qquad (12)$$

Linear combination rules are a common choice for ensembles. This rule works well in practice and it is also tractable to analysis.

Observation

We can rewrite decision rule as:

$$\mathbf{1}\left\{(\hat{\mu}_1-\hat{\mu}_0)^T\ \frac{1}{M}\sum_{i=1}^{M}R_i^T\left(R_i\hat{\Sigma}R_i^T\right)^{-1}R_i\ \left(x_q-\frac{\hat{\mu}_1+\hat{\mu}_0}{2}\right)>0\right\}$$

Then, for average case analysis with a *fixed* training set, it is enough to consider:

$$\lim_{M\to\infty}\ \frac{1}{M}\sum_{i=1}^{M}R_i^T\left(R_i\hat{\Sigma}R_i^T\right)^{-1}R_i\ =\ \mathsf{E}\left[R^T\left(R\hat{\Sigma}R^T\right)^{-1}R\right]$$

Ingredients (1)

Rows (and columns) of R drawn from a spherical Gaussian, hence for any orthogonal matrix U, $R \sim RU$. Eigendecomposing $\hat{\Sigma} = U\hat{\Lambda}U^T$ and using $UU^T = I$ we find that:

$$\mathsf{E}\left[R^T\left(R\hat{\Sigma}R^T\right)^{-1}R\right] = U\ \mathsf{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1}R\right]U^T \qquad (13)$$

Furthermore since a matrix A is diagonal if and only if $VAV^T = A$ for all *diagonal* orthogonal matrices $V = \mathrm{diag}\{\pm 1\}$ we can similarly show that the expectation on RHS is diagonal.
Now enough to evaluate the diagonal terms on RHS!

[Marzetta et al.'11] by a complicated procedure. We are more interested in how it relates to characteristics of $\hat{\Sigma}$ so we prefer simply interpretable estimates.

Ingredients (2)

Define $\rho := \text{rank}(\hat{\Lambda}) = \text{rank}(\hat{\Sigma})$.

Work with positive semidefinite ordering: $A \succeq B \iff A - B$ is positive semidefinite (p.s.d $\equiv$ symmetric with all eigenvalues $\geqslant 0$).

Upper and lower bound the diagonal matrix expectation (13) in the p.s.d ordering with spherical matrices $\alpha_{\max} \cdot I$, $\alpha_{\min} \cdot I$ to bound its condition number in terms of *data space parameters*:

$$\alpha_{\max} \cdot I \succeq \mathsf{E}\left[R^T \left(R\Lambda R^T\right)^{-1} R\right] \succeq \alpha_{\min} \cdot I$$

Where $\alpha = \alpha(k, \rho, \lambda_{\max}, \lambda_{\min \neq 0})$, k is the projected dimensionality, $\rho = \text{rank}(\hat{\Lambda}) = \text{rank}(\hat{\Sigma})$, $\lambda_{\max}$ and $\lambda_{\min \neq 0}$ are respectively the greatest and least non-zero eigenvalues of $\hat{\Sigma}$.

Results: The regularisation effect

Theorem. Let $\hat{\Sigma} \in \mathcal{M}_{d\times d}$ be a symmetric positive semi-definite matrix with rank $\rho \in \{3, ..., d-1\}$, and denote by $\lambda_{\max}(\hat{\Sigma}), \lambda_{\min\neq 0}(\hat{\Sigma}) > 0$ its greatest and least non-zero eigenvalues. Let $k < \rho - 1$ be a positive integer, and let $R \in \mathcal{M}_{k\times d}$ be a random matrix with i.i.d $\mathcal{N}(0,1)$ entries. Let $\hat{S}^{-1} := \mathsf{E}\left[R^T\left(R\hat{\Sigma}R^T\right)^{-1}R\right]$, and denote by $\kappa(\hat{S}^{-1})$ its condition number, $\kappa(\hat{S}^{-1}) = \lambda_{\max}(\hat{S}^{-1})/\lambda_{\min}(\hat{S}^{-1})$. Then:

$$\kappa(\hat{S}^{-1}) \leqslant \frac{\rho}{\rho - k - 1} \cdot \frac{\lambda_{\max}(\hat{\Sigma})}{\lambda_{\min\neq 0}(\hat{\Sigma})}$$

This theorem implies that for a large enough ensemble the condition number of the sum of random matrices $\frac{1}{M}\sum_{i=1}^{M} R_i^T\left(R_i\hat{\Sigma}R_i^T\right)^{-1}R_i$ is bounded.

Exact Generalisation error of the converged ensemble conditioned on fixed training set

Lemma [D-K, MLJ]. Let $x_q|y_q \sim \mathcal{N}(\mu_y, \Sigma)$, where $\Sigma \in \mathcal{M}_{d\times d}$ is a full rank covariance matrix. Let $R \in \mathcal{M}_{k\times d}$ be a RP matrix with i.i.d. Gaussian entries and denote $S_R^{-1} := \frac{1}{M}\sum_{i=1}^{M} R_i^T \left(R_i \hat{\Sigma} R_i^T\right)^{-1} R_i$. Then the error of the ensemble conditioned on training set equals:

$$\sum_{y=0}^{1} \pi_y \Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_{\neg y} - \hat{\mu}_y)^T S_R^{-1}(\hat{\mu}_0 + \hat{\mu}_1 - 2\mu_y)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T S_R^{-1}\Sigma S_R^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}}\right)$$

For the converged ensemble, substitute the expectation (13) for S_R^{-1} above.

Generalisation error of the converged ensemble

Theorem [D-K, MLJ]. Let $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{N}$ be a set of training data of size $N = N_0 + N_1$, subject to $N < d$ and $N_y > 1 \ \forall y$. Let x_q be a query point with Gaussian class-conditionals $x_q|y_q \sim \mathcal{N}(\mu_y, \Sigma)$, and let $\Pr\{y_q = y\} = \pi_y$. Let ρ be the rank of the maximum likelihood estimate of the covariance matrix and let $k < \rho - 1$ be a positive integer. Then for any $\delta \in (0, 1)$ we have w.p. $1 - \delta$ w,r,t, random draws of $\mathcal{T}$:

$$\Pr_{x_q, y_q}(\hat{h}_{ens}(x_q) \neq y_q) \leqslant \sum_{y=0}^{1} \pi_y \Phi\left(-\left[g\left(\bar{\kappa}\left(\sqrt{2\log\frac{5}{\delta}}\right)\right) \times \ldots\right.\right. \tag{14}$$

$$\left.\left. \ldots \left[\sqrt{\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_0)\|^2 + \frac{dN}{N_0 N_1}} - \sqrt{\frac{2N}{N_0 N_1}\log\frac{5}{\delta}}\right]_+ - \sqrt{\frac{d}{N_y}}\left(1 + \sqrt{\frac{2}{d}\log\frac{5}{\delta}}\right)\right]\right)$$

where $\bar{\kappa}(\epsilon)$ is a high probability (w.r.t draws of $\mathcal{T}$) upper bound on the condition number of $\Sigma\hat{S}^{-1}$ (given in the paper) and $g(\cdot)$ is the function $g(a) := \frac{\sqrt{a}}{1+a}$.

Experiments: Datasets

Datasets:

Name	Source	#samples	#features
colon	[Alon et al.]	62	2000
leukemia	[Golub et al.]	72	3571
leukemia large	[Golub et al.]	72	7129
prostate	[Singh et al.]	102	6033
duke	[West et al.]	44	7129

Experiments: Protocol

- Standardised features to have mean 0 and variance 1 and ran experiments on 100 independent splits. In each split took 12 points for testing, rest for training.

- For data space experiments on colon and leukaemia used ridge-regularised FLD for comparison and fitted regularisation parameter using 5-fold CV.

- For other datasets we used diagonal FLD in the data space (size, no sig. diff. in error on colon, leuk.).

- RP base learners: FLDs with full covariance and no regularisation when $k \leqslant \rho$ and pseudoinverted FLD when $k > \rho$.

- Compared performance with SVM with linear kernel as in [Fradkin et al.]

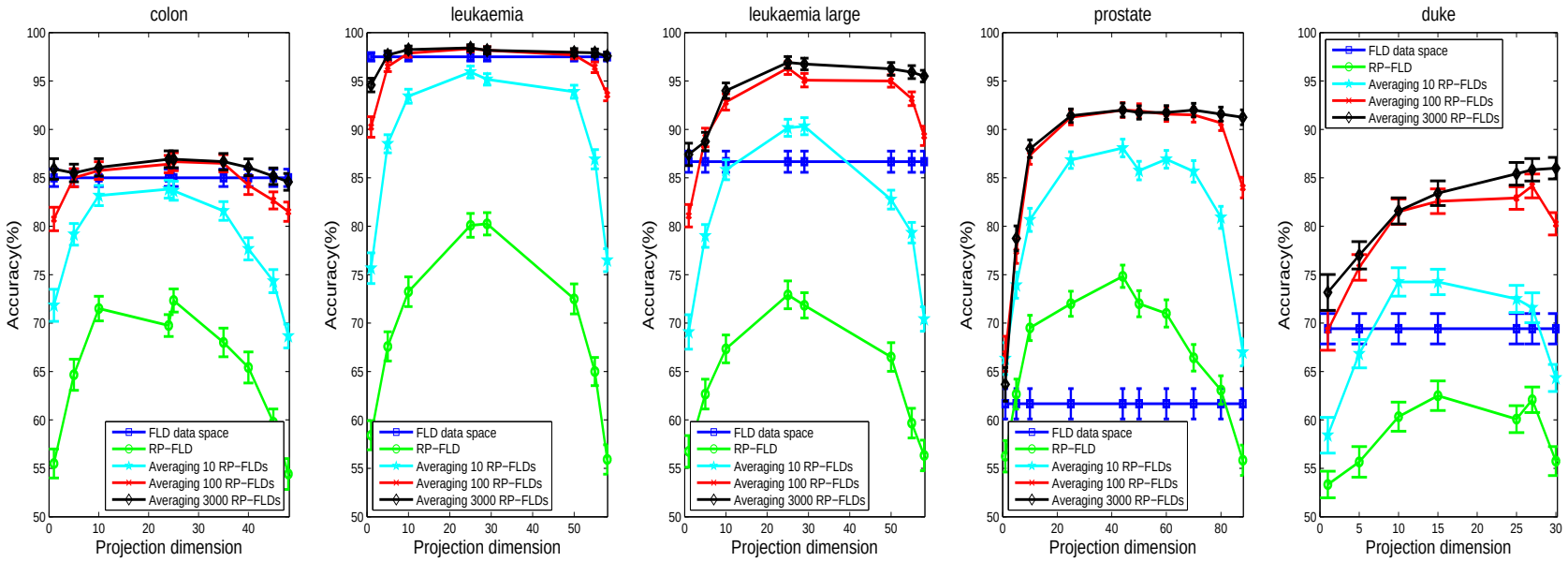
Experiments: Results for $k = \rho/2$

Mean error rates $\pm$ 1 standard error, estimated from 100 independent splits when $k = \rho/2$:

Dataset	$\rho/2$	100 RP-FLD	1000 RP-FLD	SVM
colon	24	13.58 $\pm$ 0.89	13.08 $\pm$ 0.86	16.58 $\pm$ 0.95
leuk.	29	1.83 $\pm$ 0.36	1.83 $\pm$ 0.37	1.67 $\pm$ 0.36
leuk.lg.	29	4.91$\pm$ 0.70	3.25 $\pm$ 0.60	3.50 $\pm$ 0.46
prost.	44	8.00 $\pm$ 0.76	8.00 $\pm$ 0.72	8.00 $\pm$ 0.72
duke	15	17.41 $\pm$1.27	16.58 $\pm$ 1.27	13.50 $\pm$ 1.10

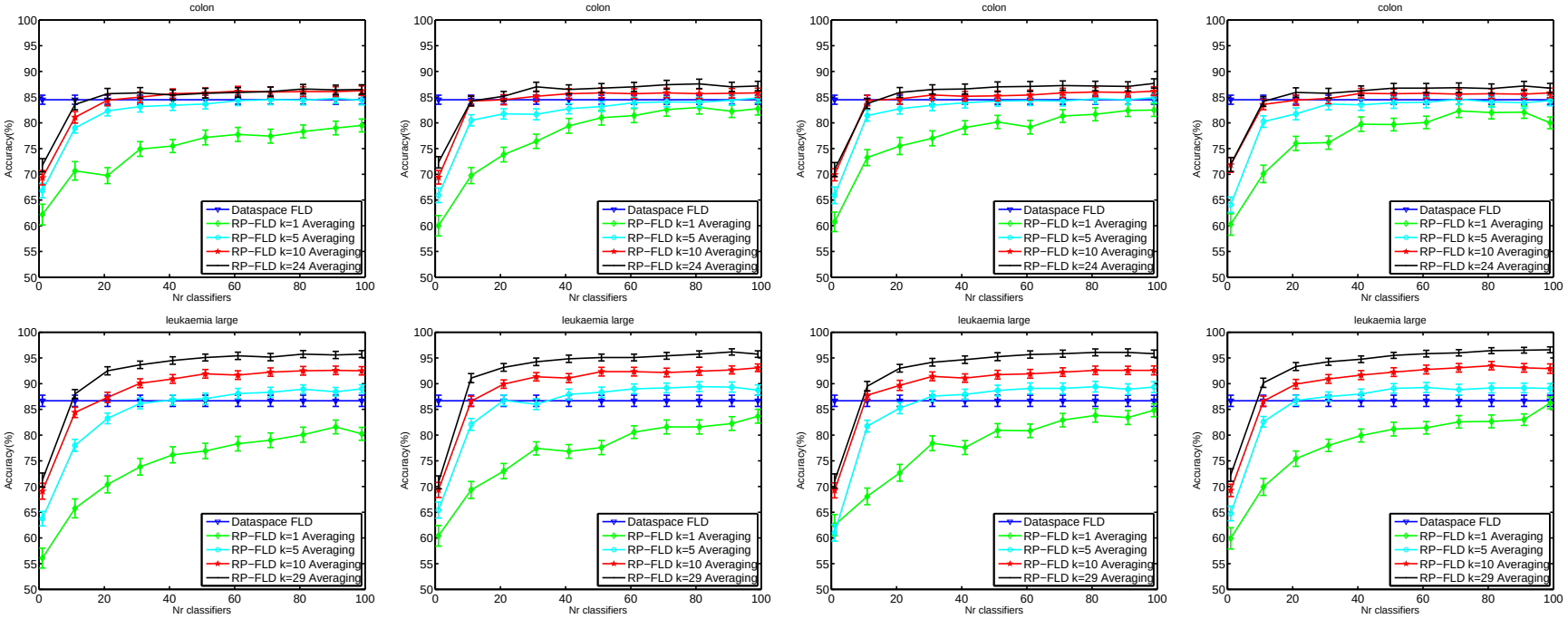
More experiments, incl. the 100,000-dimensional Dorothea data set + detailed comparisons are in the paper:
[D-K, Machine Learning, 2015]

Experiments – effect of k



Test error rates versus k and error bars mark 1 standard error estimated from 100 runs. In these experiments we used Gaussian random matrices with i.i.d $\mathcal{N}(0,1)$ entries.

Experiments – different RP matrices



Column 1: Majority Vote using Gaussian random matrices; *Column 2*: Averaging ensemble using Gaussian r.m; *Column 3*: Averaging ensemble using ± 1 random matrices. *Column 4*: Averaging ensemble using the sparse $\{-1, 0, +1\}$ random matrices from [Achlioptas '03].

Summing up compressive ensenbles

We examined a simple averaging ensemble of compressive FLD, which turns out to be interpretable in the original $\mathbb{R}^d$ as implementing a sophisticated regularisation scheme that can outperform ridge regularised dataspace FLD.

Our results on single compressive learners, as well as on ensembles, suggest that random projections may be used to uncover the structures and problem characteristics that allow effective and efficient learning for high dimensional data.

Extending the analysis to study other learning settings is subject to future work.

Selective References

R.I. Arriaga, and S. Vempala. An Algorithmic Theory of Learning: Robust Concepts and Random Projection. In 40th Annual Symposium on Foundations of Computer Science (FOCS 1999). , pp. 616–623. IEEE, 1999.

K. Ball. An Elementary Introduction to Modern Convex Geometry. Flavors of Geometry, 31: 1–58, 1997.

M-F. Bãlcan, A. Blum, and S. Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings. Machine Learning 65(1): 79-94, 2006.

P.K. Bartlett, and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. J. Machine Learning Research, 3:463–482, 2002.

R. Calderbank, S. Jafarpour, and R. Schapire. Compressed Learning: Universal Sparse Dimensionality Reduction and Learning in the Measurement Domain. Technical Report, Rice University, 2009.

E.J. Candès, and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? IEEE Transactions on Information Theory, 52, Nr. 12, 5406-5425, 2006.

S. Dasgupta, and A. Gupta. An Elementary Proof of the Johnson-Lindenstrauss Lemma. Random Structures & Algorithms 22, pp. 60-65, 2002.

R.J. Durrant and A. Kabán. Sharp Generalization Error Bounds for Randomly-projected Classifiers, ICML 2013, JMLR W&CP 28(3):693-701, 2013.

R.J. Durrant, A. Kabán. Random Projections as Regularizers: Learning a Linear Discriminant from Fewer Observations than Dimensions. Machine Learning 99(2), pp. 257-286, 2015.

M. Fard, Y. Grinberg, J. Pineau, and D. Precup. Compressed least-squares regression on sparse spaces, Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.

D. Fradkin, and D. Madigan. Experiments with random projections for machine learning. KDD 2003, pp. 522-529.

A. Garg, S. Har-Peled, and D. Roth. On Generalization Bounds, Projection Profile, and Margin Distribution. ICML 2002, pp. 171–178, 2002.

M.X. Goemans, and D.P. Williamson. Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems using Semidefinite Programming. Journal of the ACM, 42 (6):1145, 1995.

A. Kabán. New Bounds for Compressive Linear Least Squares Regression. AISTATS 2014, JMLR W&CP, 33: 448-456.

A. Kabán. Improved Bounds on the Dot Product under Random Projection and Random Sign Projection. 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD2015), 10-13 August, Sydney, pp. 487-496.

A. Kabán. A New Look at Nearest Neighbours: Identifying Benign Input Geometries via Random Projections. The 7th Asian Conference on Machine Learning (ACML 2015), Hong Kong 20-22 November 2015, Journal of Machine Learning Research-Proceedings Track 45: 65-80, 2015.

O. Maillard and R. Munos. Compressed least squares regression. NIPS 22, pp. 1213-1221, 2009.

T. Marzetta, G. Tucci, S. Simon. A random matrix theoretic approach to handling singular covariance estimates. IEEE Transactions on Information Theory, Vol. 57, Issue 9, 2011.

Q. Shi, C. Shen, R. Hill and A. Hengel. Is margin preserved after random projection? ICML 2012, pp. 591–598.

A. Siegel. Toward a Usable Theory of Chernoff bounds for Heterogeneous and Partially Dependent Random Variables. Technical Report, New York University, 1995.