Learning with Random Projections

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

Seminar talk – UCL, November 14, 2014.

Outline

- Introduction to compressive learning
- Compressive classification
- Compressive regression
- Ensembles
- Summary and future work

Thanks

Thanks to my co-author Bob Durrant.

This talk is based on the following papers:

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

R.J. Durrant and A. Kabán. Random Projections as Regularizers: Learning a Linear Discriminant from Fewer Observations than Dimensions. **Machine Learning**, In Press.
(conference version won the best paper award at **ACML 2013**)

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

Introduction

- High dimensional data is everywhere... Random projection (RP) – a computationally cheap dimensionality reduction method, amenable to analysis
- Too little is known about its effect on the generalisation performance of machine learning methods.
- Most analyses of compressive learning leverage the Johnson-Lindenstrauss lemma (JLL) for the finite set of training points, or Compressed Sensing (CS) results for sparse data.
- But JLL and CS have stronger aims and hence strong assumptions – not clear if the same is needed for compressed learning tasks?

Previous work / Background (1)

- JLL 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 Johnson-Lindenstrauss lemma so the bound unnaturally loosens with increasing the sample size
  - Similar approach in [Maillard & Munos, NIPS'09] for compressive OLS regression

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].

Previous work / 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

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 $\hat{x} = \arg\min_x \{\|x\|_1 : y = Rx\}$.

Previous work / Background (3)

Works that look at preservation / non-preservation of margin after a random projection:

- Large margin implies low dimension [Balcan & Blum, MLJ 2006]

- Is margin preserved? [Shi, Shen, Hill & Hengel, ICML 2012]

Large margin is known to be a structure that implies good generalisation if it is preserved. However it is unclear whether this is required something weaker would suffice on the original data space?

Compressed Learning

Most recent of these ideas is Compressed Learning:
If the data are s-sparse then a RP with $k \in \mathcal{O}(s \log d)$ did not lose any info. Hence, one should be able to do machine learning in the k-dim space ($k << d$) instead of d-dim space.

RIP is a uniform distance preservation guarantee over all sparse vectors, so the problems faced by JLL when applied to the training set go away: k no longer grows with N.

So, sparse representation seems like another good structure for compressive learning? – Or is it just a technically convenient one?

- Perfect reconstruction is not the same as good generalisation. Intuition says the latter should need less.

Our Questions / Aims

Can we discover what kinds of structural characteristics of the problem favours successful leaning from compressive data? (e.g. for regression? for classification?)

Approach: Start from worst case guarantees in the RP space and derive tight bounds in terms of quantities of the original data space.

What can we achieve by constructing an ensemble or compressive learners? (in terms of gains in performance & understanding)

Approach: Look at a specific case simple enough to analyse in detail.

Part I – Compressive classification

Our aim is to obtain generalisation bound for a generic linear classifier trained by ERM on randomly projected data. Make no restrictive assumptions other than the original data is drawn i.i.d from some unknown distribution $\mathcal{D}$.

Nice idea in (Garg et al, 02) where they use RP as an analytic tool to bound the error of classifiers trained in the original data space.

The idea is to quantify the effect of RP by how it changes class label predictions for projected points relative to the predictions of the data space classifier. (Garg et al, 02) derived a rather loose bound on this 'flipping probability', yielding numerically trivial bound.

Main result We derive the exact expression for the 'flipping probability', and turn around the high-level approach of (Garg et al. 02) to obtain non-trivial bounds for RP-classifiers. The ingredients of our proof then also feed back to improve the bound of (Garg et al, 02).

Notations

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:

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

Problem setting

Consider the case when d is too large. Many dimensionality reduction methods; RP is computationally cheap and non-adaptive.

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

Side note: r_{ij} can be subgaussian in general. Part of this work is specific to Gaussian R as we exploit rotation-invariance.

We are interested in the generalisation error of the ERM linear classifier trained on $\mathcal{T}_R^N$ rather than $\mathcal{T}^N$.

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:

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

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

Theorem [Risk bounds] 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 $\mathcal{T}^N$,

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

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.

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:

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

Now, bound S from $\mathsf{E}_R[S]$ w.h.p, w.r.t. the random choice of R. Dependent sum, each term depends on the same R. Can use Markov inequality or a 'Chernoff bound for dependent variables' (Siegel, 95) - numerically tighter for small δ - and take min of these.

The terms of $\mathsf{E}_R[S]$ is what we call the 'flipping probabilities' – this encodes what structure makes RP linear classification learning perform well!

Theorem [Flipping Probability]

Let $h, x \in \mathbb{R}^d$ be two vectors with angle $\theta \in [0, \pi/2]$ between them. Without loss of generality take $\|h\| = \|x\| = 1$.

Let $R \in \mathcal{M}_{k \times d}$, $k < d$, be a random projection matrix with entries $r_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$ and let $Rh,\ Rx \in \mathbb{R}^k$ be the images of h, x under R with angular separation θ_R.

1. Denote by $f_k(\theta)$ the 'flipping probability' $f_k(\theta) := \text{Pr}\{(Rh)^T Rx < 0 | h^T x > 0\}$. Then:

$$f_k(\theta) = \frac{\Gamma(k)}{(\Gamma(k/2))^2} \int_0^{\psi} \frac{z^{(k-2)/2}}{(1+z)^k} \, \text{d}z \tag{1}$$

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

2. The expression above can be rewritten as the quotient of the surface area of a hyperspherical cap with an angle of 2θ by the surface area of the corresponding hypersphere, namely:

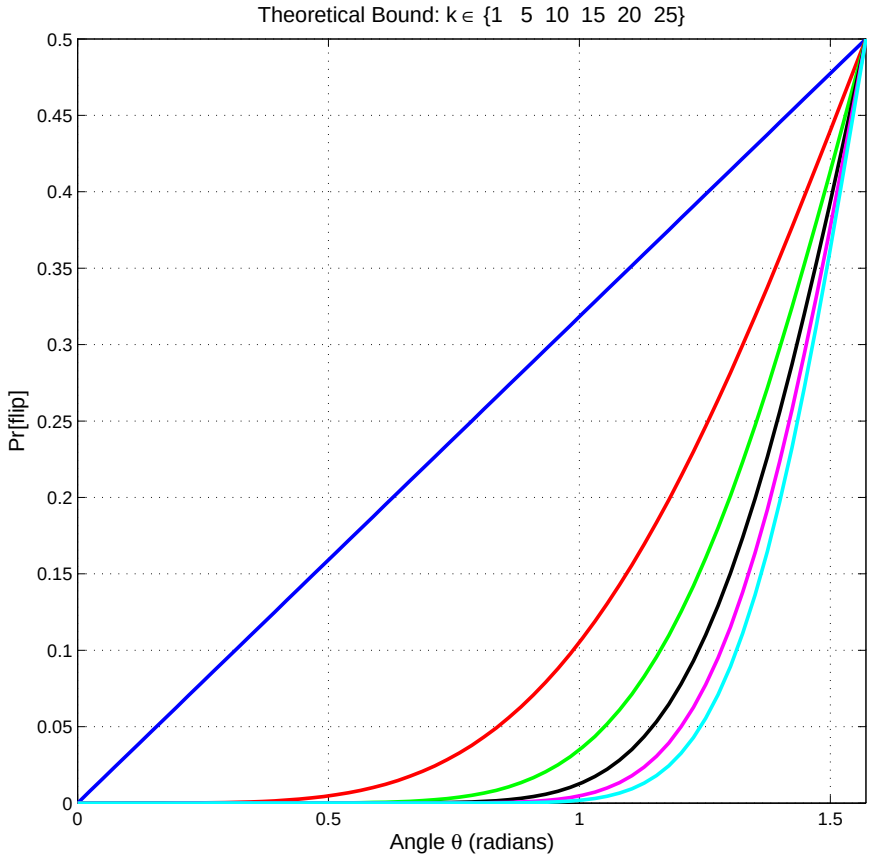
$$f_k(\theta) = \frac{\int_0^\theta \sin^{k-1}(\phi)\ \mathrm{d}\phi}{\int_0^\pi \sin^{k-1}(\phi)\ \mathrm{d}\phi} \quad (2)$$

$$\leq \exp(-\frac{k\cos^2(\theta)}{2}) \quad (3)$$

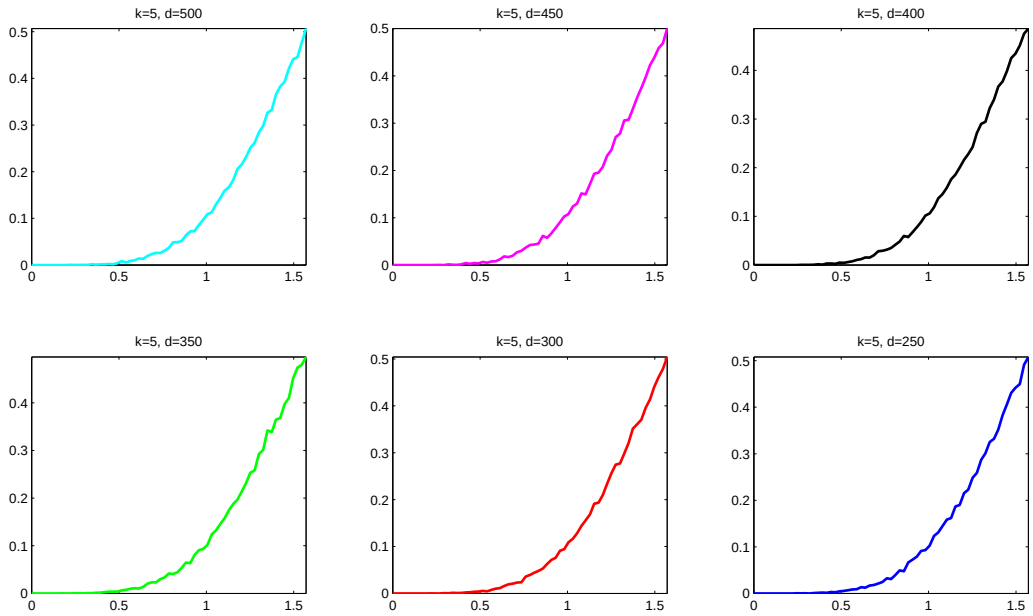
Proof is in the paper [Durrant & Kabán, ICML 2013]

We can also show, when R has 0-mean subgaussian entries, then $f_k(\theta) \leq \exp(-\frac{k\cos^2(\theta)}{8})$.

Flip Probability - Illustration



Flip Probability - d-invariance

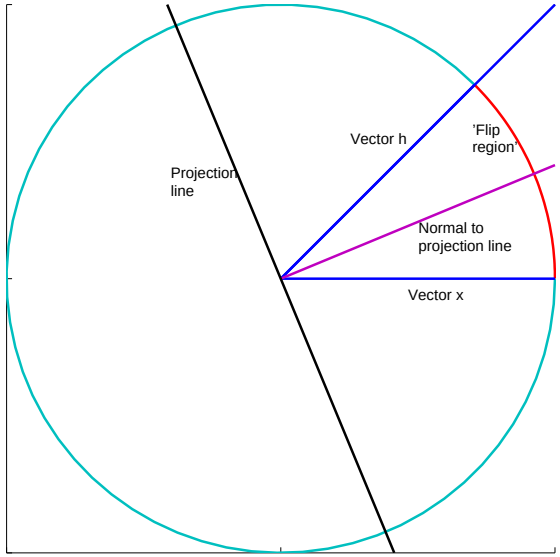


MC trials showing flip probability invariance w.r.t d. Here $k = 5$, $d \in \{500, 450, \ldots, 250\}$.

Geometric interpretation of Flipping Probability

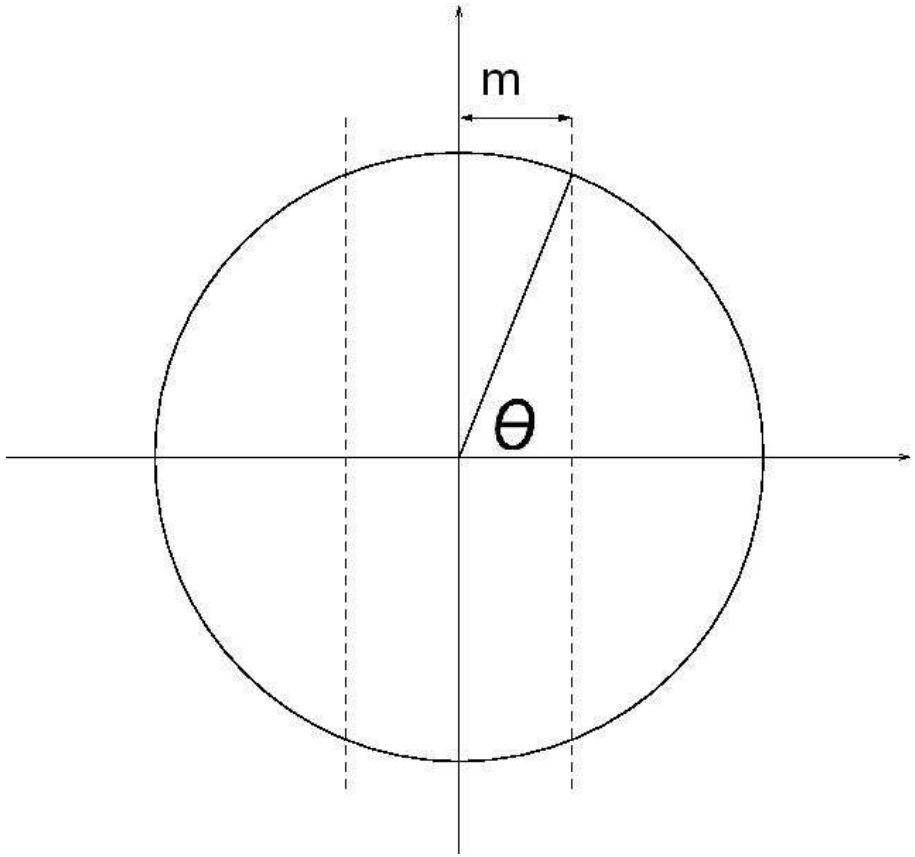
The flip probability $f_k(\theta)$ recovers the known result for $k = 1$, namely θ/π, as given in (Goemans & Williamson, 1995, Lemma 3.2.).

Geometrically, when $k = 1$ the flipping probability is the quotient of the length of the arc with angle 2θ by the circumference of the unit circle which is 2π.



For generic k, our flip probability equals the ratio of the surface area in $\mathbb{R}^{k+1}$ of a hyperspherical cap with angle 2θ, to the surface area of the unit hypersphere, This ratio is known to be upperbounded by $\exp(-\frac{1}{2}k\cos^2(\theta))$ (Ball, 1997).

Straightforward Corollaries



Flip probability and Margins

$f_k(\theta) \leq \exp(-\frac{1}{2}k\cos^2(\theta))$
(Ball, 1997,Lemma 2.2.)

$\cos(\theta) = m$

Hence, low flip probabilities imply large margins of points.

Corollary 1 [Margin Distribution Bound]

If the conditions of Theorem hold and m_i is the margin of point x_i in the data space, then for all $\delta \in (0,1)$ with probability at least $1-2\delta$ we have:

$$\mathsf{Pr}_{x_q,y_q}\{\hat{h}_R(Rx_q) \neq y_q\} \leq \hat{E}(\mathcal{T}^N, \hat{h}) + \sum_{i=1}^{N} \exp\left(-\frac{1}{2}km_i^2\right)$$

$$+ \min\left\{\sqrt{3\log\frac{1}{\delta}} \cdot \sqrt{\sum_{i=1}^{N} \exp\left(-\frac{1}{2}km_i^2\right)}, \frac{1-\delta}{\delta} \cdot \sum_{i=1}^{N} \exp\left(-\frac{1}{2}km_i^2\right)\right\}$$

$$+ \quad 2\sqrt{\frac{(k+1)\log\frac{2eN}{k+1} + \log\frac{1}{\delta}}{N}}$$

Corollary 2 [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:

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

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

Summing up Part I

- We derived the exact probability of 'label flipping' as a result of Gaussian random projection & used this to give sharp upper bounds on the generalisation error of a randomly-projected classifier. This term bounds the added error due to RP in terms of dataspace quantities, hence it encodes the info on what kind of structure ensures that the RP-classifier succeeds.
- Margins of all points matter, and we saw that large margins imply low flipping probabilities; identifying other structures that imply low flipping probabilities would be interesting.
- Further tightening of the complexity term is possible using the results in (Bartlett & Mendelson '02).

Part II - A New Analysis of Compressive Ordinary Least Square Regression

It was in fact intuitive that in linear classification not all details of the data geometry matters, and hence indeed we did not need to appeal to either JLL nor RIP to preserve all of the pairwise distances in our training set.

But how about regression? – continuous-valued targets, so it is less intuitively immediate whether we can do better than to preserve all distances.

[Kabán, AISTATS 2014]

Notations

We consider ordinary linear least squares regression in the *fixed design setting*. Given a set of N input-target pairs, $S = \{(x_1, y_1), ..., (x_N, y_N)\}$ where $x_n \in \mathbb{R}^d, y_n \in \mathbb{R}, n = 1, ..., N$, the goal is to learn an estimator v so that $x_n^T v$ approximates $x_n^T w$, under the linear model assumption:

$$y_n = x_n^T w + \gamma_n, n = 1, ..., N \tag{4}$$

where $w \in \mathbb{R}^d, \gamma_n$ is i.i.d. Gaussian noise as $\mathcal{N}(0, \sigma^2)$.

The **square loss** of an estimator v is:

$$L(v) = \frac{1}{N} \sum_{n=1}^{N} \mathsf{E}_{y_n}[(y_n - x_n^T v)^2] = \frac{1}{N} \mathsf{E}_Y[\|Y - X^T v\|^2] \tag{5}$$

The **true minimiser of the loss**: $w = \arg\min_u L(u)$

The excess risk of an estimator v: $R(v) = L(v) - L(w)$

The empirical square loss of an estimator v: $\hat{L}(v) = \frac{1}{N}\|Y - X^T v\|^2$

The **ordinary least square (OLS) estimator** is the minimiser of the empirical square loss:

$$\hat{w} = \arg\min_u \hat{L}(u) \tag{6}$$

A well-known result:
Lemma 1 Let $\sigma^2 = \mathrm{Var}(y_i)$ $(i = 1, ..., N)$, $\Sigma = XX^T/N$ fixed and invertible, w the optimal OLS as above, and $\hat{w}$ the OLS estimator. Then the expected risk $\mathrm{E}[R(\hat{w})]$ equals:

$$\mathrm{E}[L(\hat{w})] - L(w) = \sigma^2 \frac{d}{N} \tag{7}$$

where the expectation is w.r.t. $\hat{w}$ that is a function of the random vector Y.

Problem setting: Excess risk of compressive OLS

Let k be the dimension of a randomly oriented subspace that we project our input points to.

Let R be the $k \times d$ 'random projection matrix', with entries drawn i.i.d. from a 0-mean symmetric distribution with finite first 4 moments.

Let $S_R = \{(Rx_1, y_1), ..., (Rx_N, y_N)\}$ be the random-projected training set; $Rx_n \in \mathbb{R}^k$.

From S_R, we seek to learn an estimator $\hat{w}_R$ so that $x_n^T R^T \hat{w}_R$ approximates $x_n^T w$.

We are interested in the expected excess risk of $\hat{w}_R$ with respect to the optimal OLS in the original data space, w.

Definitions are analogous, notations use subscript R in the projected space:

The **square loss** of an estimator v_R: $L_R(v_R) = \frac{1}{N}\mathsf{E}_{Y|R}[\|Y - XR^T v_R\|^2]$

The **optimal OLS achievable in the random subspace** defined by R: $w_R = \arg\min_{u_R} L(u_R)$

The **empirical square loss** of an estimator v_R: $\hat{L}_R(v_R) = \frac{1}{N}\|Y - X^T R^T v_R\|^2$

and the **OLS estimator** in the randomly projected space is

$$\hat{w}_R = \arg\min_{u_R} \hat{L}_R(u_R) \tag{8}$$

Finally, our quantity of interest is:

$$\mathsf{E}_{R,Y}[L_R(\hat{w}_R)] - L(w) \tag{9}$$

i.e. the **expected excess risk** of compressive OLS.

We are also interested in $\mathsf{E}_{Y|R}[L_R(\hat{w}_R)] - L(w)$ w.h.p. with respect to the random draw of R.

Main Result

Theorem 1. Let $\sigma^2 = \mathrm{Var}(y_i)$, and $\Sigma = X^T X/N$ fixed. Let w be the optimal OLS in $\mathbb{R}^d$, and $\hat{w}_R$ the OLS estimator in the random projection space $\mathbb{R}^k$ defined by the $k \times d$ random matrix R with entries drawn i.i.d. from a zero-mean symmetric distribution with variance $1/k$ and excess kurtosis $\kappa = \frac{\mathsf{E}[R_{ij}^4]}{\mathsf{E}[R_{ij}^2]^2} - 3$. Then, the following holds:

$$\mathsf{E}_{R,Y}[L_R(\hat{w}_R)] - L(w) \leq \sigma^2 \frac{k}{N} + \frac{1}{k} \cdot \|w\|^2_{\Sigma + (1+\kappa)_+ Tr(\Sigma) I_d} \tag{10}$$

where $\|u\|^2_M = u^T M u$ stands for the squared Mahalanobis norm, I_d is the d-dimensional identity matrix, $(\cdot)_+ = \max\{0, \cdot\}$.

The first term is the variance of the estimator (much reduced from $\sigma^2 d/N$), the second term is the expected bias (the price for the reduced variance).

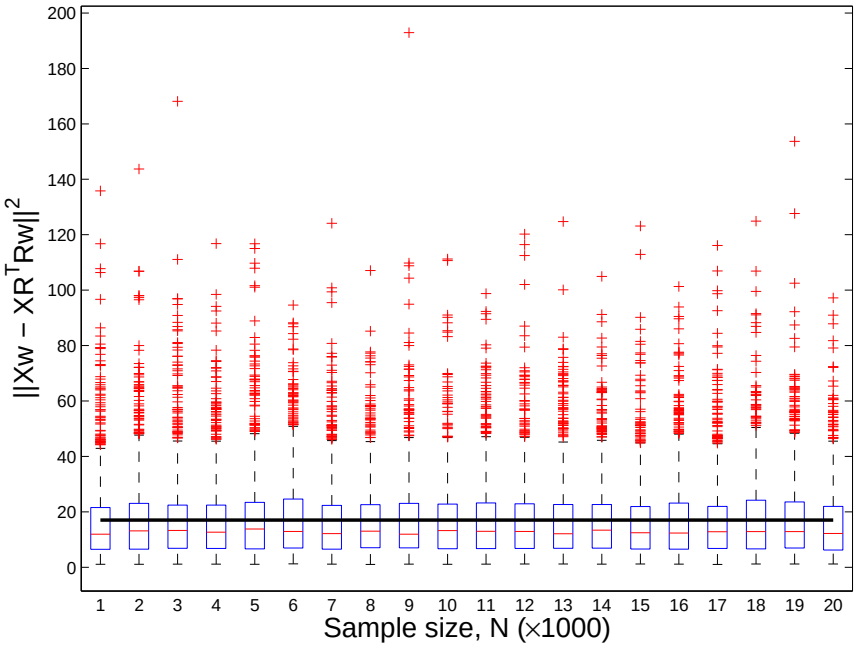
The main ingredient of the proof of Theorem 1 – this generalises a random matrix theory result of [Marzetta, 2011]:

Lemma 2 Let R be a $k \times d$ random matrix, $k < d$, with entries drawn i.i.d. from a symmetric distribution with 0-mean and finite first four moments. Let Σ be a $d \times d$ fixed positive semi-definite matrix with eigenvalues $\lambda_1, ..., \lambda_d$. Then,
$\mathsf{E}[R^T R \Sigma R^T R] = ...$

$$k \cdot \mathsf{E}[R_{ij}^2]^2 \left[(k+1)\Sigma + \mathsf{Tr}(\Sigma) I_d + \left(\frac{\mathsf{E}[R_{ij}^4]}{\mathsf{E}[R_{ij}^2]^2} - 3 \right) \sum_{i=1}^{d} \lambda_i A_i \right] \quad (11)$$

where A_i are $d \times d$ diagonal matrices: $A_i = \begin{bmatrix} \sum_{a=1}^{d} U_{ai}^2 U_{a1}^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sum_{a=1}^{d} U_{ai}^2 U_{ad}^2 \end{bmatrix}$,
and U_{ai} is the a-th entry of the i-th eigenvector of Σ.

Numerical validation



Empirical verification that the bias term of RP-OLS, $\frac{1}{N}\|X^T w - X^T R^T R w\|^2$, has no dependence on N. The entries of R are drawn i.i.d. from $\mathcal{N}(0, 1/k)$. The straight line is $\mathrm{E}_R[\frac{1}{N}\|X^T w - X^T R^T R w\|^2]$ as computed using Lemma 2.

On the distribution of the entries of R

Throughout so far we did not require R to have the Johnson-Lindenstrauss property, or to have RIP property. In fact, the moments of the entries of R higher than 4 do not affect the matrix expectation in Lemma 2.

Bound of Theorem 1 depends on the excess kurtosis of entries of R.

- negative for platikurtic distributions - bound no worse that for Gaussian
- positive for leptokurtic distributions – bound gets looser
- Examples of distributions with excess kurtosis $\kappa = 0$:
  - $r_{ij} \overset{iid}{\sim} \mathcal{N}(0, \omega^2)$
  - The following sparse random matrix, originally proposed for computational efficiency in [Achlioptas, 2003]:

$$r_{ij} \overset{iid}{\sim} \begin{cases} -\omega\sqrt{3} & \text{w.p.} 1/6 \\ 0, & \text{w.p.} 2/3 \\ \omega\sqrt{3} & \text{w.p.} 1/6 \end{cases}$$

Comparison with previous results

The bound on the bias term of compressive OLS obtained in [Maillard & Munos, NIPS'09] under assumption that R satisfies the Johnson-Lindenstrauss property (i.e. sub-Gaussian R):

$$\frac{1}{N}\|X^T w - X^T R^T R w\|^2 \leq \|w\|^2 \cdot Tr(\Sigma) \cdot \frac{8}{k} \log \frac{4N}{\delta} \qquad (12)$$

with probability $1 - \delta$.

- Eq. (12) has an extra $\log(N)$ factor that comes from the union bound after N applications of JLL for dot-products.
- Our bound has a similar flavour but eliminates the spurious $\mathcal{O}(\log N)$ factor, and holds under more general assumptions on R.
- On the other hand, our bound is in expectation whereas eq. (12) is a bound on the tails.

From our Theorem 1, we trivially obtain a tail bound that eliminates the $\log N$ factor at the expense of $1/\delta$ dependence (use Markov ineq.): With probability $1-\delta$:

$$\mathsf{E}_{Y|R}[L_R(\hat{w}_R)] - L(w) \leqslant \sigma^2 \frac{k}{N} + \frac{1}{\delta} \cdot \frac{1}{k} \cdot \|w\|^2_{\Sigma+(1+\kappa)Tr(\Sigma)I_d} \qquad (13)$$

It is open question whether the dependence on δ can be improved without extra restrictions on R.

For sub-Gaussian R, we are able to obtain upper and lower tail bounds on the term is question, $\frac{1}{N}\|X^T w - X^T R^T R w\|^2$, with $\sqrt{\log(1/\delta)}$ dependence, that also eliminate the $\log N$ term from eq.(12):

Theorem 2 Let $X \in \mathbb{R}^{d\times N}, \Sigma = XX^T/N$ fixed, and denote $\rho =$ rank(Σ). Let $R \in \mathbb{R}^{k\times d}$ be a random matrix with i.i.d. 0-mean subgaussian entries with variance $1/k$. Then, for any $\delta > 0$, the following upper and lower bounds hold simultaneously w.p. $1-\delta$:

$$\begin{aligned}
\frac{1}{N}\|X^Tw - X^TR^TRw\|^2 &\leqslant \left(1+2\sqrt{\frac{2\log(4/\delta)}{k}}\right)\left(\sqrt{\frac{\rho}{k}}+C+\sqrt{\frac{2\log(4/\delta)}{ck}}\right)^2\|w\|^2\lambda_{\max}(\Sigma) \\
&- w^T\Sigma w + 4\sqrt{\frac{2\log(4/\delta)}{k}}\|w\|^2\lambda_{\max}(\Sigma) \\
\frac{1}{N}\|X^Tw - X^TR^TRw\|^2 &\geq \left(1-2\sqrt{\frac{2\log(4/\delta)}{k}}\right)_+\left(\sqrt{\frac{\rho}{k}}-C-\sqrt{\frac{2\log(4/\delta)}{ck}}\right)_+^2\|w\|^2\lambda_{\min\neq 0}(\Sigma) \\
&- w^T\Sigma w - 4\sqrt{\frac{2\log(4/\delta)}{k}}\|w\|^2\lambda_{\min\neq 0}(\Sigma)
\end{aligned}$$

where C and c are positive constants that only depend on the 'subgaussian norm' of the rows of R; $\lambda_{\max}(\cdot)$ and $\lambda_{\min\neq 0}(\cdot)$ denote the largest and the smallest non-zero eigenvalues respectively, and $(\cdot)_+ = \max(\cdot, 0)$.

Summing up part II

We gave improved bounds on the excess risk of compressive OLS regression in the fixed design setting, that remove a spurious factor of $\log(N)$ from a previous result – hence we see that preserving all distances was not necessary!

The bias term is the added error due to RP, hence it encodes the info on what structure makes RP-OLS succeed – this turned out to be a Mahalanobis norm of w.

The main technical ingredient was an extension of a result of [Marzetta, 2011] for computing a matrix expectation that was required in our proof – may be of independent interest, e.g. in contexts that involve dealing with singular covariances matrices.

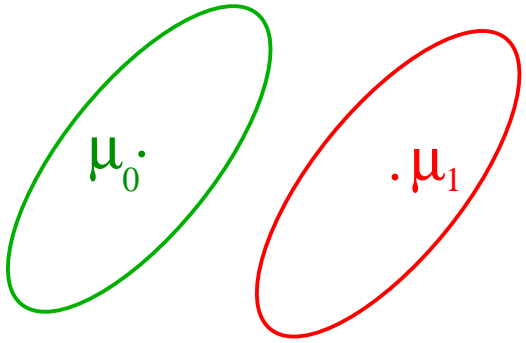
Our upper bound on the expected excess risk holds for any random projection matrix that has entries drawn i.i.d. from a symmetric distribution with finite first four moments. We also obtained high probability bounds of the same order when the random projection matrix is subgaussian.

Part III - Ensembles of compressive learners

Here 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:

$$\mathbf{1}\left\{\frac{1}{M}\sum_{i=1}^{M}(\hat{\mu}_1 - \hat{\mu}_0)^T R_i^T \left(R_i\hat{\Sigma}R_i^T\right)^{-1} R_i\left(x_q - \frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right) > 0\right\} \qquad (14)$$

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:

$$\mathsf{E}\left[R^T\left(R\hat{\Sigma}R^T\right)^{-1}R\right] = U\ \mathsf{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1}R\right]U^T \qquad (15)$$

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 (15) 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. 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 (15) for S_R^{-1} above.

Main Result: Generalisation error of the converged ensemble

Theorem. 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 $\mathcal{T}$:

$$\Pr_{x_q, y_q}(\hat{h}_{ens}(x_q) \neq y_q) \leqslant \sum_{y=0}^{1} \pi_y \Phi\left(-\left[g\left(\bar{\kappa}\left(\sqrt{2\log\frac{5}{\delta}}\right)\right) \times \ldots \right.\right. \tag{16}$$

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

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.]

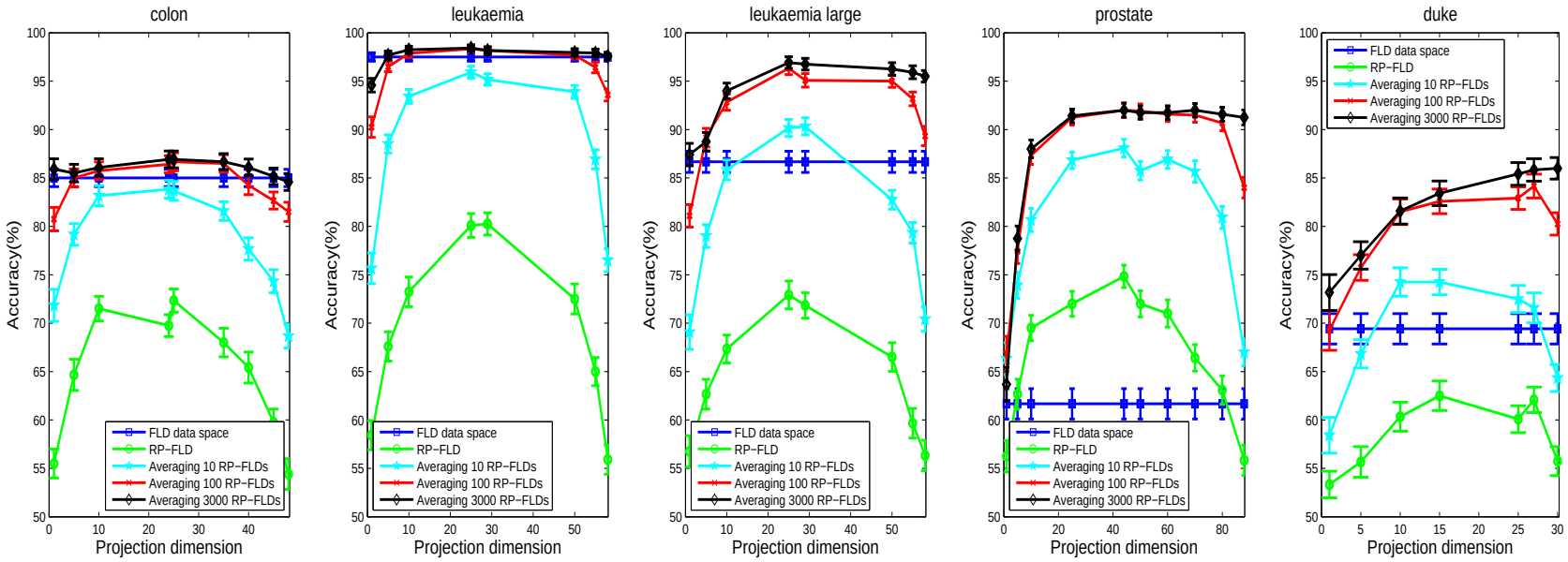
Experiments: Results for $k = \rho/2$

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

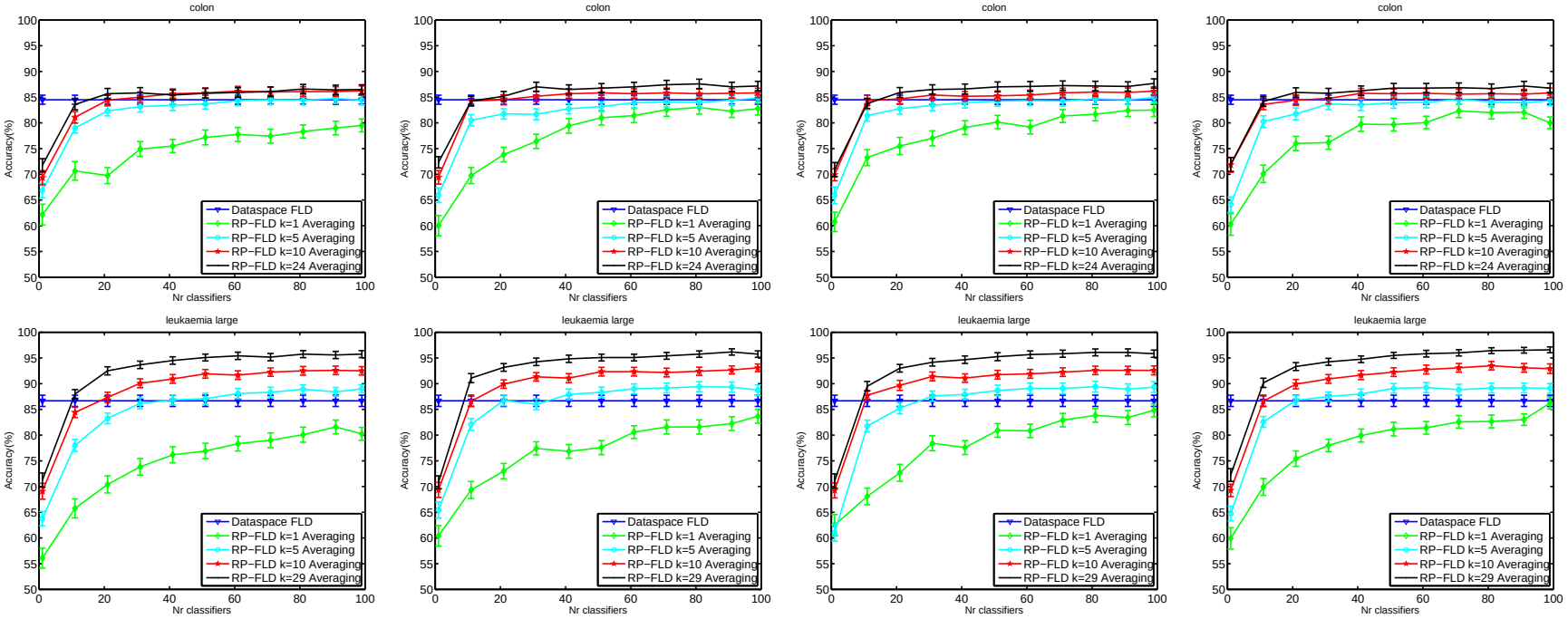
More experiments, incl. the 100,000-dimensional Dorothea data set + detailed comparisons are in the paper:
[Durrant & Kabán, Machine Learning (In Press)]

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

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.

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 and A. Kabán. Random Projections as Regularizers: Learning a Linear Discriminant from Fewer Observations than Dimensions. Machine Learning, In Press.

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.

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.