**CMSC 35900 (Spring 2009) Large Scale Learning** **Lecture: 3**

# Dimensionality Reduction and Learning

*Instructors: Sham Kakade and Greg Shakhnarovich*

## 1 $L_2$ Supervised Methods and Dimensionality Reduction

The theme of these two lectures is that for $L_2$ methods we need not work in infinite dimensional spaces. In particular, we can unadaptively find and work in a low dimensional space and achieve about as good results. These results question the need for explicitly working in infinite (or high) dimensional spaces for $L_2$ methods. In contrast, for sparsity based methods (including $L_1$ regularization), such non-adaptive projection methods significantly loose predictive power.

## 2 Margin Based Classification

For now, assume we have a distribution over $X \in \mathcal{X} \subset \mathbb{R}^d$ and $Y \in \{-1, 1\}$. Assume that there exists a weight vector $\beta$ such that $sign(\beta \cdot X) = Y$, with probability one. Hence, the distribution is separable. Furthermore, let us scale the distribution so that it is separable at margin 1, i.e.:

$$Y(\beta \cdot X) \geq 1$$

What learning algorithm should we use? The The VC dimension of halfspaces is $\Omega(D)$, so naively minimizing the $0/1$ loss in $D$ dimensions may not be lead to good generalization properties (and it's not clear how to do this anyways). Instead, maximizing the margin can be shown to provide good generalization properties — however computationally, this may be a little cumbersome (even though it is polytime).

Let us say say we have a training set $T = \{(X_i, Y_i)\}_i$.

Often, what is done, is that the perceptron algorithm is run on the training set. The perceptron algorithm run on any sequence of points $\{(X_i, Y_i)\} \subset T$ sampled from this distribution makes at most:

$$M \leq \|\mathcal{X}\|^2 \|\beta\|^2$$

mistakes (regardless of the length of the sequence) where $\|\mathcal{X}\| = \max_{X \in \mathcal{X}} \|X\|$. Hence, if we repeatedly cycle through the dataset, then eventually we will no longer make mistakes.

But what about generalization? Naively using this perceptron predictor does not necessarily lead to good generalization behavior since the VC dimension of halfspaces is $\Omega(D)$ (and no bound is known for this convergent point of the perceptron).

### 2.1 Random Projections and Margin Preservation

Now let us project $\beta$ and $X$ by $P = \frac{1}{k}A$, where $A \in \mathbb{R}^{k \times d}$ and each entry in $A$ is sample independently from $N(0, 1)$. Is separability preserved under our training set?

**Lemma 2.1.** *Assume $\|\mathcal{X}\| \leq 1$. If $k = O(\|\beta\|^2 \|\mathcal{X}\|^2 \log \frac{n}{\delta})$, then with probability greater than $1 - \delta$ for all $i$*

$$P\beta \cdot PX_i \geq \frac{1}{2}$$

*and*

$$\frac{1}{2}\|X_i\|^2 \leq \|PX_i\|^2 \leq 2\|X_i\|^2, \quad \frac{1}{2}\|\beta\|^2 \leq \|P\beta\|^2 \leq 2\|\beta\|^2$$

*Proof.* Choose $\epsilon = \frac{1}{2\|\beta\|\|\mathcal{X}\|}$ and apply the inner product preserving lemma, which implies that for any particular $X_i$ and $\beta$, that $|P\beta \cdot PX_i - \beta \cdot X_i| \geq \frac{1}{2}$, so that:

$$|Y_i P\beta \cdot PX_i - Y_i \beta \cdot X_i| \leq \frac{1}{2}$$

For $O(n^2)$ events, we use $O(\delta/n^2)$ so the total error probability is $\delta$. The final claim follows from the norm preserving lemma. □

### 2.2 Generalization

If we run the perceptron algorithm, on the training set, then the total number of mistakes made is:

$$M_t \leq O(\|\beta\|^2 \|\mathcal{X}\|^2)$$

Note that this implies that after $O(\|\beta\|^2\|\mathcal{X}\|^2)$ iteration the perceptron will stabilize to a constant solution, which has zero error.

For generalization, we are now working with a space of dimension $O(\|\beta\|^2 \log \frac{n}{\delta})$.

There are other methods to obtain generalization but the important point here is that under the margin assumption, we are essentially working in a finite dimensional space (and this subspace can be determine non-adaptively from the labels $\{Y_i\}$).

## 3 Ridge Regression and Dimensionality Reduction

Let us now consider the 'normal means' problem, sometimes referred to as the fixed design setting. Here, we have a set of $n$ points $\mathcal{X} = \{X_i\} \subset \mathbb{R}^d$, and let $X$ denote the $\mathbb{R}^{n \times d}$ matrix where the $i$ row of $X$ is $X_i$. We also observe a output vector $Y \in \mathbb{R}^n$. We desire to learn $\mathbb{E}[Y]$. In particular, we seek to predict $\mathbb{E}[Y]$ as $X\hat{\beta}$.

The square loss of an estimator $w$ is:

$$L(w) = \frac{1}{n}\mathbb{E}_Y\|Y - Xw\|^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i - X_i w)^2$$

where the expectation is with respect to $Y$. Let $\beta$ be the optimal predictor:

$$\beta = \text{argmin}_w L(w)$$

The risk of an estimator $\hat{\beta}$ is defined as:

$$R(\hat{\beta}) = L(\hat{\beta}) - L(\beta) = \frac{1}{n}\|X\hat{\beta} - X\beta\|^2$$

(which is the fixed design risk). Denoting,

$$\Sigma := \frac{1}{n}X^\top X$$

we can write the risk as:

$$R(\hat{\beta}) = (\hat{\beta} - \beta)^\top \Sigma (\hat{\beta} - \beta) := \|\hat{\beta} - \beta\|_\Sigma^2$$

Another interpretation of the risk is how well we accurately learn the parameters of the model.

Assume that $\hat{\beta}(Y)$ is an estimator constructed with the outcome $Y$ — we drop the explicit $Y$ dependence as this is clear from context. Let $\overline{\beta} = \mathbb{E}_Y\hat{\beta}$ be expected weight. We can decompose the expected risk as:

$$\begin{aligned}\mathbb{E}_Y[R(\hat{\beta})] &= \frac{1}{n}\mathbb{E}_Y\|X\hat{\beta} - X\overline{\beta}\|^2 + \frac{1}{n}\|X\overline{\beta} - X\beta\|^2 \\ &= \mathbb{E}_Y\|\hat{\beta} - \overline{\beta}\|_\Sigma^2 + \|\overline{\beta} - \beta\|_\Sigma^2\end{aligned}$$

where we have that:

$$\text{(average) variance} = \frac{1}{n}\mathbb{E}_Y\|X\hat{\beta} - X\overline{\beta}\|^2$$

and

$$\text{prediction bias vector} = X\overline{\beta} - X\beta$$

which shows a certain bias/variance decomposition of the error.

## 3.1 Risk Bounds for Ridge Regression

The ridge regression estimator using an outcome $Y$ is just:

$$\hat{\beta}_\lambda = \text{argmin}_w \frac{1}{n}\|Y - Xw\|^2 + \lambda\|w\|^2$$

The estimator is then:

$$\hat{\beta}_\lambda = (\Sigma + \lambda I)^{-1}(\frac{1}{n}X^\top Y) = (\Sigma + \lambda I)^{-1}(\frac{1}{n}\sum Y_i X_i^\top)$$

For simplicity, let us rotate $X$ such that:

$$\Sigma := \frac{1}{n}X^\top X = diag(\lambda_1, \lambda_2, \ldots \lambda_d)$$

(note this rotation does not alter the predictions of rotationally invariant algorithms). With this choice, we have that:

$$[\hat{\beta}_\lambda]_j = \frac{\frac{1}{n}\sum_{i=1}^n Y_i [X_i]_j}{\lambda_j + \lambda}$$

It is straightforward to see that:

$$\beta = E[\hat{\beta}_0]$$

and it follows that:

$$[\overline{\beta}_\lambda]_j := \mathbb{E}[\hat{\beta}_\lambda]_j = \frac{\lambda_j}{\lambda_j + \lambda}\beta_j$$

by just taking expectations.

**Lemma 3.1.** *(Risk Bound) If* $\text{Var}(Y_i) \leq 1$*, we have that:*

$$\mathbb{E}_Y[R(\hat{\beta}_\lambda)] \leq \frac{1}{n}\sum_j (\frac{\lambda_j}{\lambda_j + \lambda})^2 + \sum_j \beta_j^2 \frac{\lambda_j}{(1 + \lambda_j/\lambda)^2}$$

*This holds with equality if* $\text{Var}(Y_i) = 1$.

*Proof.* For the variance term, we have:

$$\begin{aligned}
\mathbb{E}_Y\|\hat{\beta}_\lambda - \overline{\beta}_\lambda\|_\Sigma^2 &= \sum_j \lambda_j \mathbb{E}_Y([\hat{\beta}_\lambda]_j - [\overline{\beta}_\lambda]_j)^2 \\
&= \sum_j \frac{\lambda_j}{(\lambda_j + \lambda)^2}\frac{1}{n^2}\mathbb{E}[\sum_{i=1}^n (Y_i - E[Y_i])[X_i]_j \sum_{i'=1}^n (Y_{i'} - E[Y_{i'}])[X_{i'}]_j] \\
&= \sum_j \frac{\lambda_j}{(\lambda_j + \lambda)^2}\frac{1}{n}\sum_{i=1}^n \text{Var}(Y_i)[X_i]_j^2 \\
&\leq \sum_j \frac{\lambda_j}{(\lambda_j + \lambda)^2}\frac{1}{n}\sum_{i=1}^n [X_i]_j^2 \\
&= \frac{1}{n}\sum_j \frac{\lambda_j^2}{(\lambda_j + \lambda)^2}
\end{aligned}$$

This holds with equality if $\mathrm{Var}(Y_i) = 1$. For the bias term,

$$\begin{aligned}\|\overline{\beta}_\lambda - \beta\|_\Sigma^2 &= \sum_j \lambda_j ([\overline{\beta}_\lambda]_j - [\beta]_j)^2 \\ &= \sum_j \beta_j^2 \lambda_j (\frac{\lambda_j}{\lambda_j + \lambda} - 1)^2 \\ &= \sum_j \beta_j^2 \lambda_j (\frac{\lambda}{\lambda_j + \lambda})^2\end{aligned}$$

and the result follows from algebraic manipulations. □

There following bound characterizes the risk for two natural settings for $\lambda$.

**Corollary 3.2.** *Assume* $\mathrm{Var}(Y_i) \le 1$

- *(Finite Dims) For* $\lambda = 0$*,*

$$\mathbb{E}_Y[R(\hat{\beta}_\lambda)] \le \frac{d}{n}$$

  *And if* $Var(Y_i) = 1$*, then* $\mathbb{E}_Y[R(\hat{\beta}_\lambda)] = \frac{d}{n}$*.*

- *(Infinite Dims) For* $\lambda = \frac{\sqrt{\|\Sigma\|_{trace}}}{\|\beta\|\sqrt{n}}$*, then:*

$$\mathbb{E}_Y[R(\hat{\beta}_\lambda)] \le \frac{\|\beta\|\sqrt{\|\Sigma\|_{trace}}}{2\sqrt{n}} = \frac{\|\beta\|\sqrt{\frac{1}{n}\sum_i ||X_i||^2}}{2\sqrt{n}} \le \frac{\|\beta\|\|\mathcal{X}\|}{2\sqrt{n}}$$

  *where the trace norm is the sum of the singular values and* $\|\mathcal{X}\| = \max_i ||X_i||^2$*. Furthermore, for all* $n$ *there exists a distribution* $\Pr[Y]$ *and an* $X$ *such that the* $\inf_\lambda \mathbb{E}_Y[R(\hat{\beta}_\lambda)]$ *is* $\Omega^*(\frac{\|\beta\|\sqrt{\|\Sigma\|_{trace}}}{2\sqrt{n}})$ *(so the above bound is tight up to log factors).*

Conceptually, the second bound is 'dimension free', i.e. it does not depend explicitly on $d$, which could be infinite. And we are effectively doing regression in a large (potentially) infinite dimensional space.

*Proof.* The $\lambda = 0$ case follows directly from the previous lemma. Using that $(a+b)^2 \ge 2ab$, we can bound the variance term for general $\lambda$ as follows:

$$\frac{1}{n}\sum_j (\frac{\lambda_j}{\lambda_j + \lambda})^2 \le \frac{1}{n}\sum_j \frac{\lambda_j^2}{2\lambda_j \lambda} = \frac{\sum_j \lambda_j}{2n\lambda}$$

Again, using that $(a+b)^2 \ge 2ab$, the bias term is bounded as:

$$\sum_j \beta_j^2 \frac{\lambda_j}{(1 + \lambda_j/\lambda)^2} \le \sum_j \beta_j^2 \frac{\lambda_j}{2\lambda_j/\lambda} = \frac{\lambda}{2}||\beta||^2$$

So we have that:

$$\mathbb{E}_Y[R(\hat{\beta}_\lambda)] \le \frac{\|\Sigma\|_{\text{trace}}}{2n\lambda} + \frac{\lambda}{2}||\beta||^2$$

and using the choice of $\lambda$ completes the proof.

To see the above bound is tight, consider the following problem. Let $X_i = \sqrt{\frac{n}{i}}$ and $\beta_i = \sqrt{\frac{1}{i}}$ and let $Y = X\beta + \eta$ where $\eta$ is unit variance. Here, we have that $\lambda_i = \frac{1}{i}$ so $\sum_j \lambda_j \leq \log n$ and $\|\beta\|^2 \leq \log n$, so the upper is $\frac{\log n}{\sqrt{n}}$. Now one can write the risk as:

$$\mathbb{E}_Y[R(\hat{\beta}_\lambda)] = \frac{1}{n}\sum_j (\frac{\frac{1}{i}}{\frac{1}{i}+\lambda})^2 + \sum_j \frac{\frac{1}{i^2}}{(1+\frac{1}{i\lambda})^2} \tag{1}$$

$$= \sum_j \frac{\frac{1}{n}+\lambda^2}{(1+i\lambda)^2} \tag{2}$$

$$\geq \int_1^n \frac{\frac{1}{n}+\lambda^2}{(1+x\lambda)^2} dx \tag{3}$$

$$= (\frac{1}{n}+\lambda^2)(\frac{1}{\lambda(1+\lambda)} - \frac{1}{\lambda(1+n\lambda)}) \tag{4}$$

$$= (\frac{1}{n\lambda}+\lambda)(\frac{1}{1+\lambda} - \frac{1}{1+n\lambda}) \tag{5}$$

$$\tag{6}$$

and this is $\Omega(\sqrt{n})$, for all $\lambda$. □

However, now we show that with $L_2$ complexity, we can effectively working in finite dimensions (where the dimension is chosen as a function of $n$).

## 3.2 Random Projections and Maximum Likelihood Estimation

First note that if we project to $k = O(\frac{\log n}{\epsilon^2})$ dimensions then (using $P = \frac{1}{\sqrt{k}}A$), we have that for all $i$:

$$|P\beta \cdot PX_i - \beta \cdot X_i| \leq \|\beta\|\|X_i\|\epsilon$$

Let us define the loss using only $XP^\top$ as:

$$L_P(w) = \frac{1}{n}\mathbb{E}_Y\|Y - XP^\top w\|^2$$

Let $\beta_P$ be the best fit of $Y$ with $XP^\top$, i.e.

$$\beta_P = \text{argmin}_w L_P(w)$$

and let $\hat{\beta}_P$ be the MLE fit of $Y$ with $XP^\top$ (so $\lambda = 0$). Now by the previous corollary, then:

$$\mathbb{E}_Y[L_P(\hat{\beta}_P)] - L_P(\beta_P) = \mathbb{E}_Y[\|XP^\top\hat{\beta}_P - XP^\top\beta_P\|^2] \leq \frac{k}{n}$$

Also note that:

$$\begin{aligned}
L_P(\beta_P) &\leq L_P(P\beta) \\
&= \frac{1}{n}\mathbb{E}[\|Y - XP^\top P\beta\|^2] \\
&= \frac{1}{n}\mathbb{E}[\|Y - X\beta\|^2] + \frac{1}{n}\|X\beta - XP^\top P\beta\|^2 \\
&= L(\beta) + \frac{1}{n}\sum_i (P\beta \cdot PX_i - \beta \cdot X_i)^2 \\
&\leq L(\beta) - \|\beta\|^2(\frac{1}{n}\sum_i \|X_i\|^2)\epsilon^2 \\
&= L(\beta) - \|\beta\|^2\|\Sigma\|_{\text{trace}}\epsilon^2
\end{aligned}$$

**Theorem 3.3.** *(Risk Bound after Random Projection) Assuming* $\mathrm{Var}(Y_i) \leq 1$*, and that* $P$ *is* $\epsilon$ *inner product preserving for* $k = O(\frac{\log n}{\epsilon^2})$*, then:*

$$\mathbb{E}_Y \|XP^\top \hat{\beta}_P - X\beta\|^2 = \mathbb{E}_Y[L_P(\hat{\beta}_P)] - L(\beta) \leq \frac{k}{n} + \|\beta\|^2 \|\Sigma\|_{trace}^2 \epsilon^2 = \frac{20 \log n}{n\epsilon^2} + \|\beta\|^2 \|\Sigma\|_{trace} \epsilon^2$$

*Hence, choosing* $\epsilon^2 = O(\sqrt{\frac{\log n}{n\|\beta\|^2 \|\Sigma\|_{trace}}})$*, implies that* $k = O(\|\beta\| \sqrt{\|\Sigma\|_{trace} n \log n})$ *and:*

$$\mathbb{E}_Y \|XP^\top \hat{\beta}_P - X\beta\|^2 \leq O\left( \frac{\sqrt{\log n \|\beta\|^2 \|\Sigma\|_{trace}}}{\sqrt{n}} \right)$$

*Proof.* From above we have that:

$$L(\beta) \geq L_P(\beta_P) - \|\beta\|^2 \|\Sigma\|_{\text{trace}} \epsilon^2$$

so that:

$$\mathbb{E}_Y[L_P(\hat{\beta}_P)] - L(\beta) \leq \mathbb{E}_Y[L_P(\hat{\beta}_P)] - L_P(\beta_P) + \|\beta\|^2 \|\Sigma\|_{\text{trace}} \epsilon^2 = \mathbb{E}_Y[\|XP^\top \hat{\beta}_P - XP^\top \beta_P\|^2] + \|\beta\|^2 \|\Sigma\|_{\text{trace}} \epsilon^2$$

and we have bounded the risk in the last terms as $\frac{k}{n}$. □

This matches the risk bound up to log factors. Also, our algorithm is simply an MLE estimate in $k = O(\|\beta\| \sqrt{\|\Sigma\|_{\text{trace}} n \log n})$ dimensions. Note that the number of dimensions we choose is growing as $O(\sqrt{n})$.

# References

The discussion on classification used results from Santosh Vempala's monograph on random projections. The ridge regression results, to my knowledge, are new.