Non-IID: non-stationnarité et dépendances

Liva Ralaivola, QARMA

EPAT'14: École de Printemps sur l'Apprentissage arTificiel

11 juin 2014

Outline

Overview and Motivating Examples

Recall: the Blessings of IIDness
- Setting
- A Control on the Generalization Error
  - Warming up: $|\mathcal{H}| < +\infty$
  - Rademacher-based Generalization Bound
- Beyond IIDness

Non-Stationarity
- (Non-)assumptions
- Quick Reminder on Kernels and RKHS
- Forgetting is Nice when Online Learning with Kernels
- Sequential Rademacher Complexity

Non-Independence
- Mixing Processes
- Dependent Data

Conclusion

Outline

Problems

Learning from non-IID data

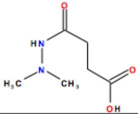
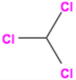
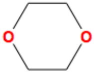
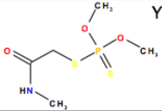
- Bipartite ranking and pairwise classification
- Similarity learning
- Classification of sequence data (mixing processes)
- Classification of connected webpages
- Active learning
- Covariate Shift
- ...

Questions

- Algorithmic: how to deal with non-IIDness?
- Theoretical: what statistical guarantees can be exhibited?
- Algorithmic and theoretical: may theoretical results motivate new algorithms? vice versa?

Concrete Examples

Virtual Screening

ID		Toxic?
1		No
2		No
3		Yes
4		Yes

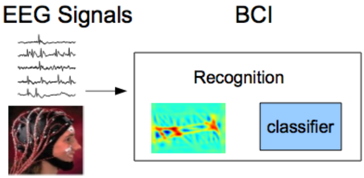
- A scoring function $f : \mathcal{M} \rightarrow \mathbb{R}$ that gives higher scores to toxic molecules
- Maximization of the Auc

Learning f

A usual strategy is to learn a pairwise binary classifier on (toxic, non toxic) pairs (with default class $+1$)

Concrete Examples

Brain computer Interface: P300 speller



(from A. Rakotomamonjy)

Goal

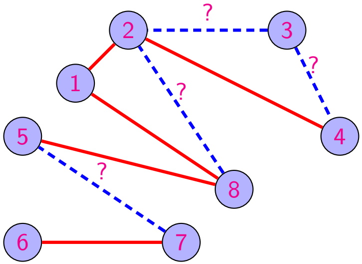
Detect P300's in EEG signal.

Nature of non-IIDness

- Drifting distribution (patient adaptation)
- Change of sampling distribution (covariate shift)

Concrete Examples

Edge prediction, relational learning, etc.

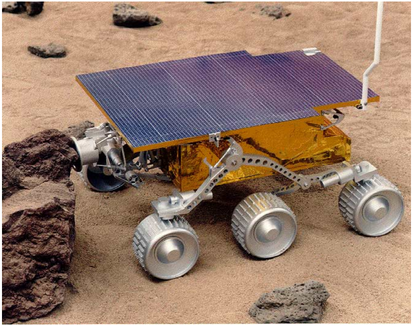


Interdependencies

- In training data
- In test data
- In general: a problem not obvious to formalize in the statistical learning framework

Concrete Examples

Robot navigation



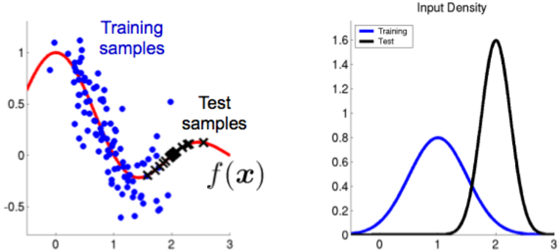
Temporal dependencies (cf. mixing processes)

- The robot has to make a decision (e.g. {`stop`, `right`, `left`, `forward`}) at each time step t according to its environment X_t
- X_t depends on the past $X_{t'}$'s ($t' < t$) with a fading influence between the X_t's over time (cf. mixing processes)

Concrete Examples

Covariate Shift

"Learning when training and test distributions are different" (NIPS 06 wshp)



(from Storkey and Sugiyama [Storkey and Sugiyama, 2007])

Results: $\mathbb{P}_{\text{train}}(Y|x) = \mathbb{P}_{\text{test}}(Y|x)$ and $p_{\text{train}}(X) \neq p_{\text{test}}(X)$

Learning setting: $S_{\text{train}} = \{(X_i, Y_i)\}_{i=1}^{n}$, $S_{\text{test}} = \{X_i\}_{i=1}^{m}$

- Importance Sampling (reweighting examples) by an estimation of $\beta(X) = p_{\text{test}}(X)/p_{\text{train}}(X)$
- Algorithmic and consistency results [Storkey and Sugiyama, 2007, Shimodaira, 2000, Smola et al., 2006]

Outline

IID Setting (supervised learning)

Notation

- $\mathcal{X}$: input space $\mathbb{R}^d$
- $\mathcal{Y}$: target space $\{-1, +1\}$
- $\mathcal{T}$: output space $\mathcal{T} = \mathbb{R}$
- D: probability distribution over $\mathcal{X} \times \mathcal{Y}$ (fixed and unknown)
- $S = \{(X_i, Y_i)\}_{i=1}^n$ IID sample $\sim D$
- $\mathcal{H} \subseteq \mathcal{T}^{\mathcal{X}}$: function class

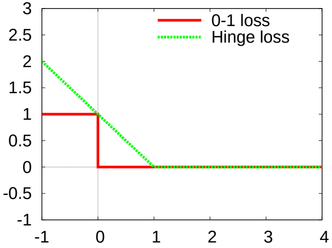
Loss function and risks

- $\ell : \mathcal{Y} \times \mathcal{T} \to \mathbb{R}$
- Empirical risk of h
$$\hat{R}_\ell(h, S) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, h(X_i))$$
- True risk of h
$$R_\ell(h, D) = \mathbb{E}_{X,Y \sim D} \ell(Y, h(X))$$

IID Setting (supervised learning)

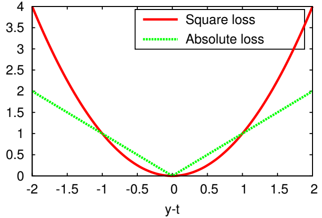
Example (Classification)

- 0-1 loss: $\ell(y, t) = \mathbb{I}[yt < 0]$
- hinge loss: $\ell(y, t) = |1 - yt|_+$



Example (Regression)

- Square loss: $\ell(y, t) = (y - t)^2$
- Absolute loss: $\ell(y, t) = |y - t|$



IID Setting (supervised learning)

Ultimate goal

Find a predictor with smallest risk within $\mathcal{H}$

$$h^* = \arg\min_{h\mathcal{H}} R_\ell(h, D)$$

Key ingredients to devise and analyze learning procedures

- Identical distribution

$$R_\ell(h, D) = \mathbb{E}_S R_\ell(h, S) \quad (= \mathbb{E}_S \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, h(X_i)))$$

- Relevant concentration inequality (usually requires some form of indepedence)
- Capacity measure of $\mathcal{H}$ or of the class of hypotheses generated by the learning algorithm (cf. sample compression schemes, stability, robustness, ...)

A Control on the Generalization Error

Targeted result

$\forall \delta \in (0,1],$, with probability at least $1-\delta$ over the draw of S:

$$\forall h \in \mathcal{H}, \qquad \mathbb{E}_{XY}\ell(h,X,Y) \leq \frac{1}{n}\sum_{i=1}^{n}\ell(Y_i,h(X_i)) + \varepsilon\left(\frac{1}{\delta},\frac{1}{n},\ldots\right).$$

For binary classification ($\ell = \ell_{0-1}$): with prob. $1-\delta$

$$\forall h \in \mathcal{H}, \qquad \mathbb{P}_{XY}(h(X) \neq Y) \leq \hat{R}(h,S) + \varepsilon\left(\frac{1}{\delta},\frac{1}{n},\ldots\right).$$

where $\hat{R}(h,S) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[h(X_i) \neq Y_i\right]$

On ε

- decreases when n increases and when δ increases
- usually contains something related to the *capacity* of $\mathcal{H}$

A Control on the Generalization Error

Targeted result

$\forall \delta \in (0,1],$, with probability at least $1-\delta$ over the draw of S:

$$\forall h \in \mathcal{H}, \qquad \mathbb{E}_{XY}\ell(h,X,Y) \leq \frac{1}{n}\sum_{i=1}^{n}\ell(Y_i,h(X_i)) + \varepsilon\left(\frac{1}{\delta},\frac{1}{n},\ldots\right).$$

For binary classification ($\ell = \ell_{0-1}$): with prob. $1-\delta$

$$\forall h \in \mathcal{H}, \qquad \mathbb{P}_{XY}(h(X) \neq Y) \leq \hat{R}(h,S) + \varepsilon\left(\frac{1}{\delta},\frac{1}{n},\ldots\right).$$

where $\hat{R}(h,S) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[h(X_i) \neq Y_i\right]$

Many ways to get generalization bounds

- VC dimension-based arguments [Vapnik, 1998]
- PAC-Bayesian theory [McAllester, 1999]
- Algorithmic stability theory [Bousquet and Elisseeff, 2002]
- Rademacher-complexity based arguments (our focus) [Bartlett and Mendelson, 2002]
- ...

Generalization bound when $|\mathcal{H}| < +\infty$

Bound

With prob. at least $1 - \delta$,

$$\forall h \in \mathcal{H},\ R(h) \leq \hat{R}(h, S) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2n}}$$

Generalization bound when $|\mathcal{H}| < +\infty$

Bound

With prob. at least $1 - \delta$,

$$\forall h \in \mathcal{H},\ R(h) \leq \hat{R}(h, S) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2n}}$$

Proof.

The proof hinges on Chernoff/Hoeffding concentration inequality: for $Z_1, \ldots, Z_n$ *independent* (and identically distributed) variables with range $[0; 1]$

$$\mathbb{P}\left(\mathbb{E}Z_1 - \frac{1}{n}\sum_{i=1}^{n} Z_i \geq \varepsilon\right) \leq \exp(-2n\varepsilon^2)$$

□

Generalization bound when $|\mathcal{H}| < +\infty$

Bound

With prob. at least $1-\delta$,

$$\forall h \in \mathcal{H},\ R(h) \le \hat{R}(h,S) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2n}}$$

Proof.

The proof hinges on Chernoff/Hoeffding concentration inequality: for $Z_1, \ldots, Z_n$ *independent* (and identically distributed) variables with range $[0;1]$

$$\mathbb{P}\left(\mathbb{E}Z_1 - \frac{1}{n}\sum_{i=1}^{n} Z_i \ge \varepsilon\right) \le \exp(-2n\varepsilon^2)$$

So, for $h \in \mathcal{H}$ fixed (set $Z_i = \mathbb{I}\left[h(X_i) \ne Y_i\right]$)

$$\mathbb{P}\left(R(h) - \hat{R}(h,S) \ge \varepsilon\right) \le \exp(-2n\varepsilon^2)$$

□

Generalization bound when $|\mathcal{H}| < +\infty$

Bound

With prob. at least $1 - \delta$,

$$\forall h \in \mathcal{H},\ R(h) \leq \hat{R}(h, S) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2n}}$$

Proof.

The proof hinges on Chernoff/Hoeffding concentration inequality: for $Z_1, \ldots, Z_n$ *independent* (and identically distributed) variables with range $[0; 1]$

$$\mathbb{P}\left(\mathbb{E}Z_1 - \frac{1}{n}\sum_{i=1}^{n} Z_i \geq \varepsilon\right) \leq \exp(-2n\varepsilon^2)$$

So, for $h \in \mathcal{H}$ fixed (set $Z_i = \mathbb{I}\left[h(X_i) \neq Y_i\right]$)

$$\mathbb{P}\left(R(h) - \hat{R}(h, S) \geq \varepsilon\right) \leq \exp(-2n\varepsilon^2)$$

and, by the union bound ($\mathbb{P}(A_1 \vee \ldots \vee A_m) \leq \sum_{i=1}^{m} \mathbb{P}(A_i)$),

$$\mathbb{P}\left(\exists h \in \mathcal{H} : R(h) - \hat{R}(h, S) \geq \varepsilon\right) \leq |\mathcal{H}| \exp(-2n\varepsilon^2).$$

□

Generalization bound when $|\mathcal{H}| < +\infty$

Bound

With prob. at least $1 - \delta$,

$$\forall h \in \mathcal{H},\ R(h) \leq \hat{R}(h, S) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2n}}$$

Proof.

The proof hinges on Chernoff/Hoeffding concentration inequality: for $Z_1, \ldots, Z_n$ *independent* (and identically distributed) variables with range $[0; 1]$

$$\mathbb{P}\left(\mathbb{E}Z_1 - \frac{1}{n}\sum_{i=1}^{n} Z_i \geq \varepsilon\right) \leq \exp(-2n\varepsilon^2)$$

So, for $h \in \mathcal{H}$ fixed (set $Z_i = \mathbb{I}\left[h(X_i) \neq Y_i\right]$)

$$\mathbb{P}\left(R(h) - \hat{R}(h, S) \geq \varepsilon\right) \leq \exp(-2n\varepsilon^2)$$

and, by the union bound ($\mathbb{P}(A_1 \vee \ldots \vee A_m) \leq \sum_{i=1}^{m} \mathbb{P}(A_i)$),

$$\mathbb{P}\left(\exists h \in \mathcal{H} : R(h) - \hat{R}(h, S) \geq \varepsilon\right) \leq |\mathcal{H}| \exp(-2n\varepsilon^2).$$

Solving for the upper bound to be equal to δ gives the result. □

Generalization bound when $|\mathcal{H}| < +\infty$

Bound

With prob. at least $1 - \delta$,

$$\forall h \in \mathcal{H},\ R(h) \leq \hat{R}(h, S) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2n}}$$

Keys

- Identical distribution: relation between R and $\hat{R}$
- Independence: concentration inequality
- Finite number of hypotheses

Rademacher-based Generalization Bound

Theorem (Rademacher generalization bound [Bartlett and Mendelson, 2002, Shawe-Taylor and Cristianini, 2004])

$\forall \delta \in [0,1)$, *with probability at least* $1-\delta$, $\forall h \in \mathcal{H}$,

$$\mathbb{P}_{XY}(Yh(X) \leq 0) \leq \hat{R}(h,S) + \frac{\hat{\mathcal{R}}(\mathcal{H},S)}{2} + c\sqrt{\frac{\ln 4/\delta}{2n}}$$

where $c > 0$ *and* $\hat{\mathcal{R}}(\mathcal{H},S) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{h\in\mathcal{H}} \frac{2}{n}\sum_{i=1}^{n} \sigma_i h(X_i)$ *is the* empirical Rademacher complexity *of* $\mathcal{H}$ *with respect to* S.

Rademacher-based Generalization Bound

Theorem (Rademacher generalization bound [Bartlett and Mendelson, 2002, Shawe-Taylor and Cristianini, 2004])

$\forall \delta \in [0,1)$, with probability at least $1-\delta$, $\forall h \in \mathcal{H}$,

$$\mathbb{P}_{XY}(Yh(X) \leq 0) \leq \hat{R}(h,S) + \frac{\hat{\mathcal{R}}(\mathcal{H},S)}{2} + c\sqrt{\frac{\ln 4/\delta}{2n}}$$

where $c > 0$ and $\hat{\mathcal{R}}(\mathcal{H},S) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{h\in\mathcal{H}} \frac{2}{n}\sum_{i=1}^{n} \sigma_i h(X_i)$ is the empirical Rademacher complexity *of $\mathcal{H}$ with respect to S.*

Theorem (Bounded Difference Inequality [McDiarmid, 1989])

Assume that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies

$$\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}'_i \in \mathcal{X}} \left| f(\mathbf{x}_1,\ldots,\mathbf{x}_n) - f(\mathbf{x}_1,\ldots,\mathbf{x}_{i-1},\mathbf{x}'_i,\mathbf{x}_{i+1},\ldots,\mathbf{x}_n) \right| \leq c_i,\ \forall i = 1,\ldots,n$$

If $X_1,\ldots,X_n$ are independent *r.v.'s taking values in $\mathcal{X}$, then, for every $t > 0$,*

$$\mathbb{P}\left\{\mathbb{E}f(X_1,\ldots,X_n) - f(X_1,\ldots,X_n) \geq t\right\} \leq \exp\left(-2t^2 / \sum_{i=1}^{n} c_i^2\right)$$

$$\mathbb{P}\left\{f(X_1,\ldots,X_n) - \mathbb{E}f(X_1,\ldots,X_n) \geq t\right\} \leq \exp\left(-2t^2 / \sum_{i=1}^{n} c_i^2\right).$$

Rademacher complexity of $\mathcal{H}$

Definition (*Rademacher complexity* of $\mathcal{H}$)

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{S\boldsymbol{\sigma}} \sup_{h \in \mathcal{H}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i h(X_i),$$

where $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_n\}$, and $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2.$

On $\mathcal{R}_n$

- It measures the richness of the class $\mathcal{H}$
- Says how well $\mathcal{H}$ is capable of correlating with randomly assigned labels
- The marginal distribution over $\mathcal{X}$ is directly taken into account
- It cannot be directly computed...

Rademacher complexity of $\mathcal{H}$

Definition (*Rademacher complexity* of $\mathcal{H}$)

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{S\boldsymbol{\sigma}} \sup_{h\in\mathcal{H}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i h(X_i),$$

where $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_n\}$, and $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$.

Definition (Empirical Rademacher complexity $\hat{\mathcal{R}}(\mathcal{H}, S)$)

$$\hat{\mathcal{R}}(\mathcal{H}, S) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{h\in\mathcal{H}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i h(X_i)$$

Concentration of $\hat{\mathcal{R}}(\mathcal{H}, S)$

Using McDiarmid inequality, with prob. at least $1-\delta$

$$\mathcal{R}(\mathcal{H}) \leq \hat{\mathcal{R}}(\mathcal{H}, S) + c\sqrt{\frac{\log 2/\delta}{2n}}$$

Proof of the Rademacher-based bound

For all h (simultaneously), the following trivially holds

$$R(h) - \hat{R}(h, S) \leq \sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}(h, S) \right) \quad = \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I} \left[h(X_i) \neq Y_i \right] \right)$$

and we may want to take care of the upper bound.

Proof of the Rademacher-based bound

For all h (simultaneously), the following trivially holds

$$R(h) - \hat{R}(h, S) \leq \sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}(h, S) \right) \quad = \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i \right] \right)$$

and we may want to take care of the upper bound.
Let us define $H : (\mathcal{X} \times \mathcal{Y})^n \to [0; 1]$ as

$$H((x_1, y_1), \ldots, (x_n, y_n)) = \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(x_i) \neq y_i \right] \right)$$

Proof of the Rademacher-based bound

For all h (simultaneously), the following trivially holds

$$R(h) - \hat{R}(h,S) \leq \sup_{h\in\mathcal{H}} \Big(R(h) - \hat{R}(h,S) \Big) \quad = \sup_{h\in\mathcal{H}} \left(R(h) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right)$$

and we may want to take care of the upper bound.
Let us define $H : (\mathcal{X}\times\mathcal{Y})^n \to [0;1]$ as

$$H((x_1,y_1),\ldots,(x_n,y_n)) = \sup_{h\in\mathcal{H}} \left(R(h) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[h(x_i) \neq y_i\right] \right)$$

Note that, for $i \in \{1,\ldots,n\}$ and g realizing the sup of $H((x_1,y_1),\ldots,(x_n,y_n))$

$$H((x_1,y_1),\ldots,(x_n,y_n)) - H((x_1,y_1),\ldots,(x_i',y_i')\ldots,(x_n,y_n))$$
$$= \left(R(g) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[g(x_i) \neq y_i\right] \right) - \sup_{h\in\mathcal{H}} \left(R(h) - \frac{1}{n}\sum_{j\neq i}^{n} \mathbb{I}\left[h(x_j) \neq y_j\right] - \frac{1}{n}\mathbb{I}\left[h(x_i') \neq y_i'\right] \right)$$

Proof of the Rademacher-based bound

For all h (simultaneously), the following trivially holds

$$R(h) - \hat{R}(h,S) \leq \sup_{h\in\mathcal{H}} \Big(R(h) - \hat{R}(h,S)\Big) \quad = \sup_{h\in\mathcal{H}} \left(R(h) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right]\right)$$

and we may want to take care of the upper bound.
Let us define $H : (\mathcal{X}\times\mathcal{Y})^n \to [0;1]$ as

$$H((x_1,y_1),\ldots,(x_n,y_n)) = \sup_{h\in\mathcal{H}} \left(R(h) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[h(x_i) \neq y_i\right]\right)$$

Note that, for $i \in \{1,\ldots,n\}$ and g realizing the sup of $H((x_1,y_1),\ldots,(x_n,y_n))$

$$\begin{aligned}
&H((x_1,y_1),\ldots,(x_n,y_n)) - H((x_1,y_1),\ldots,(x_i',y_i')\ldots,(x_n,y_n)) \\
&= \left(R(g) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[g(x_i)\neq y_i\right]\right) - \sup_{h\in\mathcal{H}}\left(R(h) - \frac{1}{n}\sum_{j\neq i}^{n}\mathbb{I}\left[h(x_j)\neq y_j\right] - \frac{1}{n}\mathbb{I}\left[h(x_i')\neq y_i'\right]\right) \\
&\leq \left(R(g) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[g(x_i)\neq y_i\right]\right) - \left(R(g) - \frac{1}{n}\sum_{j\neq i}^{n}\mathbb{I}\left[g(x_j)\neq y_j\right] - \frac{1}{n}\mathbb{I}\left[g(x_i')\neq y_i'\right]\right)
\end{aligned}$$

Proof of the Rademacher-based bound

For all h (simultaneously), the following trivially holds

$$R(h) - \hat{R}(h,S) \leq \sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}(h,S) \right) \quad = \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i \right] \right)$$

and we may want to take care of the upper bound.
Let us define $H : (\mathcal{X} \times \mathcal{Y})^n \to [0;1]$ as

$$H((x_1,y_1),\dots,(x_n,y_n)) = \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(x_i) \neq y_i \right] \right)$$

Note that, for $i \in \{1,\dots,n\}$ and g realizing the sup of $H((x_1,y_1),\dots,(x_n,y_n))$

$$\begin{aligned}
& H((x_1,y_1),\dots,(x_n,y_n)) - H((x_1,y_1),\dots,(x_i',y_i')\dots,(x_n,y_n)) \\
&= \left(R(g) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[g(x_i) \neq y_i \right] \right) - \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{j \neq i}^{n} \mathbb{I}\left[h(x_j) \neq y_j \right] - \frac{1}{n} \mathbb{I}\left[h(x_i') \neq y_i' \right] \right) \\
&\leq \left(R(g) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[g(x_i) \neq y_i \right] \right) - \left(R(g) - \frac{1}{n} \sum_{j \neq i}^{n} \mathbb{I}\left[g(x_j) \neq y_j \right] - \frac{1}{n} \mathbb{I}\left[g(x_i') \neq y_i' \right] \right) \\
&= \frac{1}{n} \left(\mathbb{I}\left[g(x_i') \neq y_i' \right] - \mathbb{I}\left[g(x_i) \neq y_i \right] \right) \leq \frac{1}{n}
\end{aligned}$$

Proof of the Rademacher-based bound

For all h (simultaneously), the following trivially holds

$$R(h) - \hat{R}(h,S) \leq \sup_{h\in\mathcal{H}} \left(R(h) - \hat{R}(h,S)\right) \quad = \sup_{h\in\mathcal{H}} \left(R(h) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right]\right)$$

and we may want to take care of the upper bound.
Let us define $H : (\mathcal{X} \times \mathcal{Y})^n \to [0;1]$ as

$$H((x_1,y_1),\ldots,(x_n,y_n)) = \sup_{h\in\mathcal{H}} \left(R(h) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[h(x_i) \neq y_i\right]\right)$$

We thus have

$$\left|H((x_1,y_1),\ldots,(x_n,y_n)) - H((x_1,y_1),\ldots,(x_i',y_i')\ldots,(x_n,y_n))\right| \leq \frac{1}{n}$$

and we may use McDiarmid's concentration inequality:

$$\mathbb{P}\left(H(S) - \mathbb{E}_S H(S) \geq \varepsilon\right) \leq \exp(-2n\varepsilon^2)$$

or, with probability $1-\delta$

$$H(S) \leq \mathbb{E}_S \sup_{h\in\mathcal{H}} \left(R(h) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right]\right) + \sqrt{\frac{\log 1/\delta}{2n}}$$

Proof of the Rademacher-based bound (we don't back down)

We now have

$$R(h) - \hat{R}(h, S) \leq_\delta \mathbb{E}_S \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) + \sqrt{\frac{\log 1/\delta}{2n}}$$

and the crux is, again, to work out the upper bound.

Proof of the Rademacher-based bound (we don't back down)

We now have

$$R(h) - \hat{R}(h, S) \leq_\delta \mathbb{E}_S \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) + \sqrt{\frac{\log 1/\delta}{2n}}$$

and the crux is, again, to work out the upper bound.

$$\mathbb{E}_S \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right)$$

Proof of the Rademacher-based bound (we don't back down)

We now have

$$R(h) - \hat{R}(h,S) \leq_\delta \mathbb{E}_S \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) + \sqrt{\frac{\log 1/\delta}{2n}}$$

and the crux is, again, to work out the upper bound.

$$\begin{aligned} &\mathbb{E}_S \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \\ &= \mathbb{E}_S \sup_{h \in \mathcal{H}} \left(\mathbb{E}_{S'} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i') \neq Y_i'\right] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \end{aligned}$$

Proof of the Rademacher-based bound (we don't back down)

We now have

$$R(h) - \hat{R}(h, S) \leq_\delta \mathbb{E}_S \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) + \sqrt{\frac{\log 1/\delta}{2n}}$$

and the crux is, again, to work out the upper bound.

$$\begin{aligned}
&\mathbb{E}_S \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \\
&= \mathbb{E}_S \sup_{h \in \mathcal{H}} \left(\mathbb{E}_{S'} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i') \neq Y_i'\right] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \\
&\leq \mathbb{E}_{SS'} \frac{1}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \left(\mathbb{I}\left[h(X_i') \neq Y_i'\right] - \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \qquad \text{(convexity of sup)}
\end{aligned}$$

Proof of the Rademacher-based bound (we don't back down)

We now have

$$R(h) - \hat{R}(h, S) \leq_\delta \mathbb{E}_S \sup_{h\in\mathcal{H}} \left(R(h) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) + \sqrt{\frac{\log 1/\delta}{2n}}$$

and the crux is, again, to work out the upper bound.

$$\begin{aligned}
&\mathbb{E}_S \sup_{h\in\mathcal{H}} \left(R(h) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \\
&= \mathbb{E}_S \sup_{h\in\mathcal{H}} \left(\mathbb{E}_{S'} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[h(X_i') \neq Y_i'\right] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \\
&\leq \mathbb{E}_{SS'} \frac{1}{n} \sup_{h\in\mathcal{H}} \sum_{i=1}^{n} \left(\mathbb{I}\left[h(X_i') \neq Y_i'\right] - \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) && \text{(convexity of sup)} \\
&= \mathbb{E}_{SS'\boldsymbol{\sigma}} \frac{1}{n} \sup_{h\in\mathcal{H}} \sum_{i=1}^{n} \sigma_i \left(\mathbb{I}\left[h(X_i') \neq Y_i'\right] - \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) && \text{(identical distribution)}
\end{aligned}$$

Proof of the Rademacher-based bound (we don't back down)

We now have

$$R(h) - \hat{R}(h,S) \leq_\delta \mathbb{E}_S \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) + \sqrt{\frac{\log 1/\delta}{2n}}$$

and the crux is, again, to work out the upper bound.

$$\begin{aligned}
&\mathbb{E}_S \sup_{h \in \mathcal{H}} \left(R(h) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \\
&= \mathbb{E}_S \sup_{h \in \mathcal{H}} \left(\mathbb{E}_{S'} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i') \neq Y_i'\right] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \\
&\leq \mathbb{E}_{SS'} \frac{1}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \left(\mathbb{I}\left[h(X_i') \neq Y_i'\right] - \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \qquad \text{(convexity of sup)} \\
&= \mathbb{E}_{SS'\boldsymbol{\sigma}} \frac{1}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \left(\mathbb{I}\left[h(X_i') \neq Y_i'\right] - \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \qquad \text{(identical distribution)} \\
&= \mathbb{E}_{SS'\boldsymbol{\sigma}} \sup_{h \in \mathcal{H}} \left(\sum_{i=1}^{n} \sigma_i \mathbb{I}\left[h(X_i') \neq Y_i'\right] - \sum_{i=1}^{n} \sigma_i \mathbb{I}\left[h(X_i) \neq Y_i\right] \right)
\end{aligned}$$

Proof of the Rademacher-based bound (we don't back down)

We now have

$$R(h) - \hat{R}(h,S) \leq_\delta \mathbb{E}_S \sup_{h\in\mathcal{H}} \left(R(h) - \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) + \sqrt{\frac{\log 1/\delta}{2n}}$$

and the crux is, again, to work out the upper bound.

$$\begin{aligned}
&\mathbb{E}_S \sup_{h\in\mathcal{H}} \left(R(h) - \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \\
&= \mathbb{E}_S \sup_{h\in\mathcal{H}} \left(\mathbb{E}_{S'} \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left[h(X_i') \neq Y_i'\right] - \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \\
&\leq \mathbb{E}_{SS'} \frac{1}{n} \sup_{h\in\mathcal{H}} \sum_{i=1}^n \left(\mathbb{I}\left[h(X_i') \neq Y_i'\right] - \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \qquad \text{(convexity of sup)} \\
&= \mathbb{E}_{SS'\boldsymbol{\sigma}} \frac{1}{n} \sup_{h\in\mathcal{H}} \sum_{i=1}^n \sigma_i \left(\mathbb{I}\left[h(X_i') \neq Y_i'\right] - \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \qquad \text{(identical distribution)} \\
&= \mathbb{E}_{SS'\boldsymbol{\sigma}} \sup_{h\in\mathcal{H}} \left(\sum_{i=1}^n \sigma_i \mathbb{I}\left[h(X_i') \neq Y_i'\right] - \sum_{i=1}^n \sigma_i \mathbb{I}\left[h(X_i) \neq Y_i\right] \right) \\
&\leq \mathbb{E}_{S\boldsymbol{\sigma}} \frac{2}{n} \sup_{h\in\mathcal{H}} \sum_{i=1}^n \sigma_i \mathbb{I}\left[h(X_i) \neq Y_i\right]
\end{aligned}$$

Proof of the Rademacher-based bound (almost there)

We are at the point where

$$R(h) - \hat{R}(h,S) \leq_\delta \mathbb{E}_{S\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \mathbb{I}\left[h(X_i) \neq Y_i\right] + \sqrt{\frac{\log 1/\delta}{2n}}$$

and the upper bound might be tamed as follows.

Proof of the Rademacher-based bound (almost there)

We are at the point where

$$R(h) - \hat{R}(h, S) \leq_\delta \mathbb{E}_{S\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \mathbb{I}\left[h(X_i) \neq Y_i\right] + \sqrt{\frac{\log 1/\delta}{2n}}$$

and the upper bound might be tamed as follows.

$$\mathbb{E}_{\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \mathbb{I}\left[h(X_i) \neq Y_i\right] = \mathbb{E}_{\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i (1 - Y_i h(X_i))/2$$

Proof of the Rademacher-based bound (almost there)

We are at the point where

$$R(h) - \hat{R}(h,S) \leq_\delta \mathbb{E}_{S\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \mathbb{I}\left[h(X_i) \neq Y_i\right] + \sqrt{\frac{\log 1/\delta}{2n}}$$

and the upper bound might be tamed as follows.

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \mathbb{I}\left[h(X_i) \neq Y_i\right] &= \mathbb{E}_{\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i (1 - Y_i h(X_i))/2 \\ &= \mathbb{E}_{\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i Y_i h(X_i)/2\end{aligned}$$

Proof of the Rademacher-based bound (almost there)

We are at the point where

$$R(h) - \hat{R}(h, S) \leq_\delta \mathbb{E}_{S\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \mathbb{I}\left[h(X_i) \neq Y_i\right] + \sqrt{\frac{\log 1/\delta}{2n}}$$

and the upper bound might be tamed as follows.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \mathbb{I}\left[h(X_i) \neq Y_i\right] &= \mathbb{E}_{\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i (1 - Y_i h(X_i))/2 \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i Y_i h(X_i)/2 \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i h(X_i)/2
\end{aligned}$$

Proof of the Rademacher-based bound (almost there)

We are at the point where

$$R(h) - \hat{R}(h, S) \leq_\delta \mathbb{E}_{S\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \mathbb{I}\left[h(X_i) \neq Y_i\right] + \sqrt{\frac{\log 1/\delta}{2n}}$$

and the upper bound might be tamed as follows.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \mathbb{I}\left[h(X_i) \neq Y_i\right] &= \mathbb{E}_{\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i (1 - Y_i h(X_i))/2 \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i Y_i h(X_i)/2 \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i h(X_i)/2
\end{aligned}$$

and, therefore,

$$\mathbb{E}_{S\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \mathbb{I}\left[h(X_i) \neq Y_i\right] = \mathbb{E}_{S\boldsymbol{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i h(X_i)/2 = \frac{\mathcal{R}_n(\mathcal{H})}{2}$$

Proof of the Rademacher-based bound (we are done)

Previous calculations amount to

$$R(h) - \hat{R}(h, S) \leq_\delta \frac{\mathcal{R}_n(\mathcal{H})}{2} + \sqrt{\frac{\log 1/\delta}{2n}}$$

Proof of the Rademacher-based bound (we are done)

Previous calculations amount to

$$R(h) - \hat{R}(h, S) \leq_\delta \frac{\mathcal{R}_n(\mathcal{H})}{2} + \sqrt{\frac{\log 1/\delta}{2n}}$$

This finally gives, using the concentration of $\hat{\mathcal{R}}(\mathcal{H}, S)$

$$R(h) - \hat{R}(h, S) \leq_\delta \frac{\hat{\mathcal{R}}(\mathcal{H}, S)}{2} + \sqrt{\frac{\log 2/\delta}{2n}}$$

Proof of the Rademacher-based bound (we are done)

Previous calculations amount to

$$R(h) - \hat{R}(h, S) \leq_\delta \frac{\mathcal{R}_n(\mathcal{H})}{2} + \sqrt{\frac{\log 1/\delta}{2n}}$$

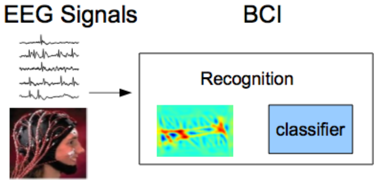
This finally gives, using the concentration of $\hat{\mathcal{R}}(\mathcal{H}, S)$

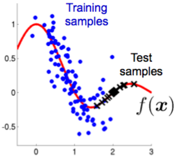
$$R(h) - \hat{R}(h, S) \leq_\delta \frac{\hat{\mathcal{R}}(\mathcal{H}, S)}{2} + \sqrt{\frac{\log 2/\delta}{2n}}$$

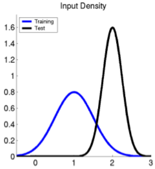
Critical observations

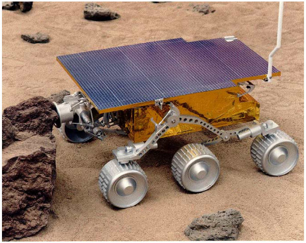
- Identical distributions is pivotal to relate $\mathbb{E}_S$ to $\mathbb{E}_{XYY}$
- It is important as well for the double-sample trick
- Independence is a necessary condition for the proof (even though there are concentration inequalities for dependent data)
- On a side note:
  - $\hat{\mathcal{R}}(\mathcal{H}, S)$ can be computed from data
  - there are *local* versions of Rademacher complexities [Bartlett et al., 2005]

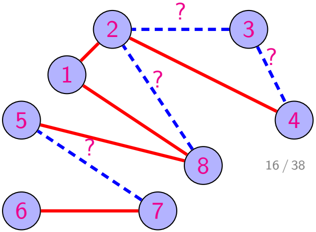
Beyond IIDness

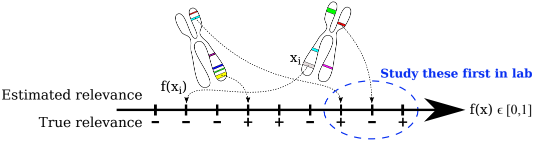












Outline

Non-stationarity

(Non-)assumptions

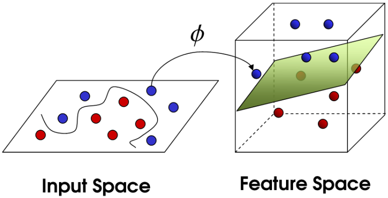
- Training data: $Z_1, \ldots, Z_n$ observations not identically distributed;
- Test data: $Z'_1, \ldots, Z'_m$ observations not identically distributed.

Formal frameworks

- Learning from noisy data: privacy learning, semi-supervised learning,. . .
- Transfer learning
- Drifting distributions
  - switching regimes
  - smoothly changing parameterized distributions
- Online learning (with adversarial oracle)

Quick Reminder on Kernels

Kernel Trick Basics



We are happy if we know $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$

Quick Reminder on Kernels

RKHS

Given a *positive kernel* k, the associated RKHS is the Hilbert space

$$\mathbb{H} = \overline{\left\{ f : f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i),\ x_i \in \mathcal{X} \right\}},$$

such that for $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$, $g = \sum_{j=1}^{m} \beta_j k(\cdot, z_j)$,

$$\langle f, g \rangle = \sum_{ij} \alpha_i \beta_j k(x_i, z_j).$$

Mapping ϕ might be thought of as $\phi(x) = k(\cdot, x)$.

Quick Reminder on Kernels

RKHS

Given a *positive kernel* k, the associated RKHS is the Hilbert space

$$\mathbb{H} = \overline{\left\{ f : f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i),\ x_i \in \mathcal{X} \right\}},$$

such that for $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$, $g = \sum_{j=1}^{m} \beta_j k(\cdot, z_j)$,

$$\langle f, g \rangle = \sum_{ij} \alpha_i \beta_j k(x_i, z_j).$$

Mapping ϕ might be thought of as $\phi(x) = k(\cdot, x)$.

Evaluation operator $k(\cdot, \mathbf{x})$

With the definition of $\mathbb{H}$, it comes that $\forall h \in \mathbb{H}$:

$$\forall x \in \mathcal{X},\ h(x) = \langle h, k(\cdot, x) \rangle.$$

Online Learning

General scheme for online learning,
[Cesa-Bianchi and Lugosi, 2006, Shalev-Shwartz, 2007]

$(x_1, y_1), \ldots, (x_t, y_t), \ldots$ data stream

- initialize h_0
- Repeat
  - predict $\hat{y}_t = h_{t-1}(x_t)$
  - receive correct target $\hat{y}_t$
  - incur loss $\ell_t = \ell(y_t, \hat{y}_t, h_{t-1}, x_t)$
  - adjust $h_{t-1} \rightarrow h_t$ using $\{\ell_t, y_t, \hat{y}_t, h_{t-1}, x_t\}$

Online Learning

General scheme for online learning,
[Cesa-Bianchi and Lugosi, 2006, Shalev-Shwartz, 2007]

$(x_1, y_1), \dots, (x_t, y_t), \dots$ data stream

- initialize h_0
- Repeat
  - predict $\hat{y}_t = h_{t-1}(x_t)$
  - receive correct target $\hat{y}_t$
  - incur loss $\ell_t = \ell(y_t, \hat{y}_t, h_{t-1}, x_t)$
  - adjust $h_{t-1} \to h_t$ using $\{\ell_t, y_t, \hat{y}_t, h_{t-1}, x_t\}$

Example (The Immortal Perceptron [Block, 1962, Novikoff, 1963])

Update of w_t when w_{t-1} errs on (x_t, y_t):

$$w_t \leftarrow w_{t-1} + y_t x_t$$

- Second-Order Perceptron [Cesa-Bianchi and Lugosi, 2006]
- Ultraconservative Algorithms [Crammer and Singer, 2003]
- Passive-Aggressive Learning [Crammer et al., 2006]
- ...

Online Learning

General scheme for online learning,
[Cesa-Bianchi and Lugosi, 2006, Shalev-Shwartz, 2007]

$(x_1, y_1), \ldots, (x_t, y_t), \ldots$ data stream

- initialize h_0
- Repeat
  - predict $\hat{y}_t = h_{t-1}(x_t)$
  - receive correct target $\hat{y}_t$
  - incur loss $\ell_t = \ell(y_t, \hat{y}_t, h_{t-1}, x_t)$
  - adjust $h_{t-1} \to h_t$ using $\{\ell_t, y_t, \hat{y}_t, h_{t-1}, x_t\}$

Example (Stochastic Optimization)

- Pegasos [Shwartz et al., 2007, Shalev-Shwartz et al., 2011]
- Stochastic Gradient Descent [Kivinen et al., 2010, Bordes et al., 2005]
- ...

Online Learning

General scheme for online learning,
[Cesa-Bianchi and Lugosi, 2006, Shalev-Shwartz, 2007]

$(x_1, y_1), \ldots, (x_t, y_t), \ldots$ data stream

- initialize h_0
- Repeat
  - predict $\hat{y}_t = h_{t-1}(x_t)$
  - receive correct target $\hat{y}_t$
  - incur loss $\ell_t = \ell(y_t, \hat{y}_t, h_{t-1}, x_t)$
  - adjust $h_{t-1} \rightarrow h_t$ using $\{\ell_t, y_t, \hat{y}_t, h_{t-1}, x_t\}$

Example (Winnow algorithm [Littlestone, 1988])

Update of w_t when w_{t-1} errs on (x_t, y_t):

$$w_t \propto w_{t-1} \exp(\eta y_t x_t)$$

Online Learning

General scheme for online learning,
[Cesa-Bianchi and Lugosi, 2006, Shalev-Shwartz, 2007]

$(x_1, y_1), \ldots, (x_t, y_t), \ldots$ data stream

- initialize h_0
- Repeat
  - predict $\hat{y}_t = h_{t-1}(x_t)$
  - receive correct target $\hat{y}_t$
  - incur loss $\ell_t = \ell(y_t, \hat{y}_t, h_{t-1}, x_t)$
  - adjust $h_{t-1} \rightarrow h_t$ using $\{\ell_t, y_t, \hat{y}_t, h_{t-1}, x_t\}$

Example (Recursive Least Squares / Optimal Control)

$$w_t \leftarrow w_{t-1} + A_t x_t$$

- Online Kernel Recursive Least Squares [Engel et al., 2003]
- Sparse Online Gaussian Processes [Csato and Opper, 2002]
- ...

Online Learning

General scheme for online learning,
[Cesa-Bianchi and Lugosi, 2006, Shalev-Shwartz, 2007]

$(x_1, y_1), \ldots, (x_t, y_t), \ldots$ data stream

- initialize h_0
- Repeat
  - predict $\hat{y}_t = h_{t-1}(x_t)$
  - receive correct target $\hat{y}_t$
  - incur loss $\ell_t = \ell(y_t, \hat{y}_t, h_{t-1}, x_t)$
  - adjust $h_{t-1} \rightarrow h_t$ using $\{\ell_t, y_t, \hat{y}_t, h_{t-1}, x_t\}$

Example (Bandits)

- Thompson Sampling [Thompson, 1933, Chapelle and Li, 2012]
- UCB, UCT and variants [Bubeck and Cesa-Bianchi, 2012, Munos, 2014]
- Exp3, and variants [Auer et al., 2002]
- ...

Forgetting is Nice when Online Learning with Kernels

Study of Online Learning with Kernels of [Kivinen et al., 2010]

- Target: a kernel classifier $h = \sum_{i=1}^{n} \alpha_i k(\cdot, x)$
- Philosophical update when processing (x_t, y_t) at time t

$$h_t \leftarrow \beta_t h_{t-1} + \alpha_t k(\cdot, x_t)$$

Forgetting is Nice when Online Learning with Kernels

Study of Online Learning with Kernels of [Kivinen et al., 2010]

- Target: a kernel classifier $h = \sum_{i=1}^{n} \alpha_i k(\cdot, x)$
- Philosophical update when processing (x_t, y_t) at time t

$$h_t \leftarrow \beta_t h_{t-1} + \alpha_t k(\cdot, x_t)$$

Drawbacks

- The kernel expansion grows with time (and so do prediction time and storage)
- There is no recovery of the algorithm to change of distribution (old examples have ‘too much weight’)

Forgetting is Nice when Online Learning with Kernels

Study of Online Learning with Kernels of [Kivinen et al., 2010]

- Target: a kernel classifier $h = \sum_{i=1}^{n} \alpha_i k(\cdot, x)$
- Philosophical update when processing (x_t, y_t) at time t

$$h_t \leftarrow \beta_t h_{t-1} + \alpha_t k(\cdot, x_t)$$

Drawbacks

- The kernel expansion grows with time (and so do prediction time and storage)
- There is no recovery of the algorithm to change of distribution (old examples have 'too much weight')

Solution

- Implement a strategy to forget old information
- Do it so the *regret* of the algorithm is controlled

Forgetting is Nice when Online Learning with Kernels

Study of Online Learning with Kernels of [Kivinen et al., 2010]

- Target: a kernel classifier $h = \sum_{i=1}^{n} \alpha_i k(\cdot, x)$
- Philosophical update when processing (x_t, y_t) at time t

$$h_t \leftarrow \beta_t h_{t-1} + \alpha_t k(\cdot, x_t)$$

Many related works

- Kernel Perceptron [Shawe-Taylor and Cristianini, 2004], Passive-Aggressive Learning [Crammer et al., 2006], Pegasos [Shwartz et al., 2007, Shalev-Shwartz et al., 2011]
- Budget online learning: Budget Perceptron [Crammer et al., 2003], Forgetron [Dekel et al., 2008], Last Recent Budget Perceptron [Cavallanti et al., 2007], Projectron [Orabona and Keshet, 2008]

NORMA [Kivinen et al., 2010]

Setting

- Stream of data $(x_1, y_1), \ldots, (x_t, y_t), \ldots$
- In the hindsight, a batch procedure

$$h = \arg\min_{f \in \mathbb{H}} \frac{\lambda}{2}\|f\|^2 + \frac{1}{n}\sum_{t=1}^{n} \ell(f(x_t), y_t),$$

where ℓ is some convex loss function and $\lambda > 0$.

NORMA [Kivinen et al., 2010]

A stochastic (sub-)gradient descent procedure

General update:

$$h_t \leftarrow h_{t-1} - \eta \nabla_f |_{f=h_{t-1}} R_t(f, (x_t, y_t))$$

for

$$R_t(f, (x_t, y_t)) = \frac{\lambda}{2} \|f\|^2 + \ell(f(x_t), y_t)$$

and (x_t, y_t) 'randomly' chosen.

NORMA [Kivinen et al., 2010]

A stochastic (sub-)gradient descent procedure

General update:

$$h_t \leftarrow h_{t-1} - \eta \nabla_f|_{f=h_{t-1}} R_t(f, (x_t, y_t))$$

for

$$R_t(f, (x_t, y_t)) = \frac{\lambda}{2}\|f\|^2 + \ell(f(x_t), y_t)$$

and (x_t, y_t) 'randomly' chosen.

Working out $\nabla_f R_t(f, (x_t, y_t))$

Thanks to $\|f\|^2 = \langle f, f\rangle$ and $f(x) = \langle f, k(\cdot, x)\rangle$, we have

$$\nabla_f R_t(f, (x_t, y_t)) = \lambda f + \nabla_f \ell(\langle f, k(\cdot, x_t)\rangle, y_t)$$

where $\partial\ell$ denotes the derivative (or subderivative) of ℓ wrt its first variable

NORMA [Kivinen et al., 2010]

A stochastic (sub-)gradient descent procedure

General update:

$$h_t \leftarrow h_{t-1} - \eta \nabla_f|_{f=h_{t-1}} R_t(f, (x_t, y_t))$$

for

$$R_t(f, (x_t, y_t)) = \frac{\lambda}{2}\|f\|^2 + \ell(f(x_t), y_t)$$

and (x_t, y_t) 'randomly' chosen.

Working out $\nabla_f R_t(f, (x_t, y_t))$

Thanks to $\|f\|^2 = \langle f, f \rangle$ and $f(x) = \langle f, k(\cdot, x) \rangle$, we have

$$\begin{aligned}\nabla_f R_t(f, (x_t, y_t)) &= \lambda f + \nabla_f \ell\left(\langle f, k(\cdot, x_t) \rangle, y_t\right) \\ &= \lambda f + \partial\ell(\langle f, k(\cdot, x_t) \rangle, y_t) k(\cdot, x_t)\end{aligned}$$

where $\partial\ell$ denotes the derivative (or subderivative) of ℓ wrt its first variable

NORMA [Kivinen et al., 2010]

A stochastic (sub-)gradient descent procedure

General update:

$$h_t \leftarrow h_{t-1} - \eta \nabla_f |_{f=h_{t-1}} R_t(f, (x_t, y_t))$$

for

$$R_t(f, (x_t, y_t)) = \frac{\lambda}{2} \|f\|^2 + \ell(f(x_t), y_t)$$

and (x_t, y_t) 'randomly' chosen.

Working out $\nabla_f R_t(f, (x_t, y_t))$

Thanks to $\|f\|^2 = \langle f, f \rangle$ and $f(x) = \langle f, k(\cdot, x) \rangle$, we have

$$\begin{aligned} \nabla_f R_t(f, (x_t, y_t)) &= \lambda f + \nabla_f \ell (\langle f, k(\cdot, x_t) \rangle, y_t) \\ &= \lambda f + \partial \ell(\langle f, k(\cdot, x_t) \rangle, y_t) k(\cdot, x_t) \\ &= \lambda f + \partial \ell(f(x_t), y_t) k(\cdot, x_t) \end{aligned}$$

where $\partial \ell$ denotes the derivative (or subderivative) of ℓ wrt its first variable

NORMA [Kivinen et al., 2010]

Update

We have

$$h_t = h_{t-1} - \eta \nabla_f |_{f=h_{t-1}} R_t(f, (x_t, y_t))$$

NORMA [Kivinen et al., 2010]

Update

We have

$$
\begin{aligned}
h_t &= h_{t-1} - \eta \nabla_f |_{f=h_{t-1}} R_t(f, (x_t, y_t)) \\
&= h_{t-1} - \eta \left[\lambda h_{t-1} + \partial \ell(h_{t-1}(x_t), y_t) k(\cdot, x_t) \right]
\end{aligned}
$$

NORMA [Kivinen et al., 2010]

Update

We have

$$\begin{aligned} h_t &= h_{t-1} - \eta \nabla_f |_{f=h_{t-1}} R_t(f, (x_t, y_t)) \\ &= h_{t-1} - \eta \left[\lambda h_{t-1} + \partial \ell(h_{t-1}(x_t), y_t) k(\cdot, x_t)\right] \\ &= (1 - \lambda\eta) h_{t-1} - \eta \partial \ell(h_{t-1}(x_t), y_t) k(\cdot, x_t) \end{aligned}$$

NORMA [Kivinen et al., 2010]

Update

We have

$$\begin{aligned} h_t &= h_{t-1} - \eta \nabla_f |_{f=h_{t-1}} R_t(f, (x_t, y_t)) \\ &= h_{t-1} - \eta \left[\lambda h_{t-1} + \partial \ell(h_{t-1}(x_t), y_t) k(\cdot, x_t)\right] \\ &= (1 - \lambda\eta) h_{t-1} \underbrace{-\eta \partial \ell(h_{t-1}(x_t), y_t)}_{\alpha_t^t} k(\cdot, x_t) \end{aligned}$$

NORMA [Kivinen et al., 2010]

Update

We have

$$\begin{aligned} h_t &= h_{t-1} - \eta \nabla_f|_{f=h_{t-1}} R_t(f,(x_t,y_t)) \\ &= h_{t-1} - \eta \left[\lambda h_{t-1} + \partial\ell(h_{t-1}(x_t),y_t)k(\cdot,x_t)\right] \\ &= (1-\lambda\eta)h_{t-1} \underbrace{-\eta\partial\ell(h_{t-1}(x_t),y_t)}_{\alpha_t^t} k(\cdot,x_t) \end{aligned}$$

Compact representation

At time t,

$$h_t = \sum_{\tau=1}^{t} \alpha_\tau^t k(\cdot, x_\tau)$$

where (by induction, with $h_0 = 0$)

$$\alpha_\tau^t = \begin{cases} -\eta\partial\ell(h_{t-1}(x_t),y_t) & \text{if } \tau = t \\ (1-\eta\lambda)^{t-\tau}\alpha_\tau^\tau & \text{otherwise} \end{cases}$$

NORMA [Kivinen et al., 2010]

Compact representation

At time t,

$$h_t = \sum_{\tau=1}^{t} \alpha_\tau^t k(\cdot, x_\tau)$$

where (by induction, with $h_0 = 0$)

$$\alpha_\tau^t = \begin{cases} -\eta \partial \ell(h_{t-1}(x_t), y_t) & \text{if } \tau = t \\ (1 - \eta\lambda)^{t-\tau} \alpha_\tau^\tau & \text{otherwise} \end{cases}$$

Observations

- If $0 < 1 - \eta\lambda < 1$, the weights of old examples decrease exponentially fast
- This calls for a (smooth) *truncation* procedure motivated by
  - numerical representation purposes
  - adaptation purposes
  - compactness purposes

NORMA [Kivinen et al., 2010]

Compact representation

At time t,

$$h_t = \sum_{\tau=1}^{t} \alpha_\tau^t k(\cdot, x_\tau)$$

where (by induction, with $h_0 = 0$)

$$\alpha_\tau^t = \begin{cases} -\eta \partial \ell(h_{t-1}(x_t), y_t) & \text{if } \tau = t \\ (1-\eta\lambda)^{t-\tau} \alpha_\tau^\tau & \text{otherwise} \end{cases}$$

Theorem (Truncation error, smooth forgetting of old examples)

If the loss function is such that $|\partial_z \ell(z, y)| \leq C$, $\|k\| \leq X$ *and*

$$h_t^{trunc} = \sum_{i=\max(1,t-\tau)}^{t} \alpha_i^t k(\cdot, x_i),$$

then

$$\left\| h_t - h_t^{trunc} \right\| \leq (1-\eta\lambda)^\tau CX/\lambda$$

NORMA [Kivinen et al., 2010]

A Controlled Drifting Class of Sequence of Predictors

$$\mathcal{G}(B, D_1, D_2) = \left\{(g_1, \ldots, g_t) : \sum_{\tau} \|g_\tau - g_{\tau+1}\| \leq D_1, \sum_{\tau} \|g_\tau - g_{\tau+1}\|^2 \leq D_2, \|g_\tau\| \leq B\right\}$$

Cumulative Loss L_{cum}

$$L_{\text{cum}}(\mathbf{h}, S) = \sum_t \ell(h_{t-1}(x_t), y_t),$$

where $\mathbf{h} = (h_1, \ldots, h_t)$.

Theorem (Mistake Bound of Norma with Non-Stationary Targets)

Suppose that:

- $\ell(h(x), y) = \max(0, \rho - yh(x))$ *for* $\rho > 0$
- $\exists \mathbf{g} \in \mathcal{G}(B, D_1, D_2)$

Then there exists right choices for η *and* λ *such that*

$$\left| \{1 \leq \tau \leq t : y_\tau h_{\tau-1}(x_\tau) \leq \rho\} \right| \leq \mathcal{K}(\eta, \lambda, \rho, D_1, D_2, B)$$

NORMA [Kivinen et al., 2010]

A Controlled Drifting Class of Sequence of Predictors

$$\mathcal{G}(B, D_1, D_2) = \left\{ (g_1, \ldots, g_t) : \sum_\tau \|g_\tau - g_{\tau+1}\| \leq D_1, \sum_\tau \|g_\tau - g_{\tau+1}\|^2 \leq D_2, \|g_\tau\| \leq B \right\}$$

Cumulative Loss L_{cum}

$$L_{\text{cum}}(\mathbf{h}, S) = \sum_t \ell(h_{t-1}(x_t), y_t),$$

where $\mathbf{h} = (h_1, \ldots, h_t)$.

Theorem (Mistake Bound of Norma with Non-Stationary Targets)

Suppose that:

- $\ell(h(x), y) = \max(0, \rho - yh(x))$ *for* $\rho > 0$
- $\exists \mathbf{g} \in \mathcal{G}(B, D_1, D_2)$

Then there exists right choices for η *and* λ *such that*

$$\left| \{1 \leq \tau \leq t : y_\tau h_{\tau-1}(x_\tau) \leq \rho\} \right| \leq \mathcal{K}(\eta, \lambda, \rho, D_1, D_2, B)$$

Proof.

Just kidding □

What to take home from online learning with kernels

Algorithmically

- Many algorithms for online learning
- They implement some sort of forgetting to be able to adapt to drifting distribution.

What to take home from online learning with kernels

Algorithmically

- Many algorithms for online learning
- They implement some sort of forgetting to be able to adapt to drifting distribution.

Regret, Mistake bounds

- Natural way to analyze online learning algorithms: mistake bounds, regret
- It is known that small regret gives good generalization error under the right assumptions [Cesa-Bianchi et al., 2004]

What to take home from online learning with kernels

Algorithmically

- Many algorithms for online learning
- They implement some sort of forgetting to be able to adapt to drifting distribution.

Regret, Mistake bounds

- Natural way to analyze online learning algorithms: mistake bounds, regret
- It is known that small regret gives good generalization error under the right assumptions [Cesa-Bianchi et al., 2004]

Where is Rademacher???

Online Learning and Sequential Rademacher Complexity

Issues

- seems that the tools used to analyze online learning algorithm are very different from those for batch algorithm
- not easy to take advantage of things made in one field in the other field
- connections between the learning approach is not straightforward

Online Learning and Sequential Rademacher Complexity

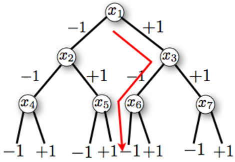
Issues

- seems that the tools used to analyze online learning algorithm are very different from those for batch algorithm
- not easy to take advantage of things made in one field in the other field
- connections between the learning approach is not straightforward

A beautiful contribution to address these issues

Work of [Rakhlin et al., 2010a, Rakhlin et al., 2010b]. One of the pivotal notion: Sequential Rademacher Complexity (where $\mathbf{x}_\tau : \{\pm 1\}^\tau \to \mathcal{X}$)

$$\mathcal{R}_t(\mathcal{H}) = \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\sup_{h \in \mathcal{H}} \sum_{\tau=1}^{t} \epsilon_\tau f(\mathbf{x}_\tau(\epsilon)) \right],$$



Example : $\epsilon = (+1, -1, -1)$

$$\sum_{t=1}^{3} \epsilon_t f(\mathbf{x}_t(\epsilon)) = +f(x_1) - f(x_3) - f(x_6)$$

(picture from Rakhlin's poster)

Outline

Mixing processes

Setting

- $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{+\infty}$ *stationary*: for any t and $m, k \geq 0$, the random subsequences $(Z_t, \ldots, Z_{t+m})$ and $(Z_{t+k}, \ldots, Z_{t+m+k})$ are identically distributed
- The dependencies are fading over time, e.g., ϕ-mixing process:

$$\varphi(k) = \sup_{n, A \in \sigma_{n+k}^{+\infty}, B \in \sigma_{-\infty}^{n}} \left| \mathbb{P}[A|B] - \mathbb{P}[A] \right| .$$

$\mathbf{Z}$ is φ-mixing if $\varphi(k) \to 0$ as $k \to 0$

Mixing processes

Setting

- $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{+\infty}$ *stationary*: for any t and $m, k \geq 0$, the random subsequences $(Z_t, \ldots, Z_{t+m})$ and $(Z_{t+k}, \ldots, Z_{t+m+k})$ are identically distributed
- The dependencies are fading over time, e.g., ϕ-mixing process:

$$\varphi(k) = \sup_{n, A \in \sigma_{n+k}^{+\infty}, B \in \sigma_{-\infty}^{n}} |\mathbb{P}[A|B] - \mathbb{P}[A]| .$$

$\mathbf{Z}$ is φ-mixing if $\varphi(k) \to 0$ as $k \to 0$

Theorem

([Kontorovich and Ramanan, 2008, Mohri and Rostamizadeh, 2008])

Let $\psi : \mathcal{U}^m \to \mathbb{R}$ be a function defined over a countable space $\mathcal{U}$, and $\underline{X}$ be a stationary φ mixing process. If ψ is l-Lipschitz with respect to the Hamming metric for some $l > 0$, then the following holds for all $t > 0$:

$$\mathbb{P}_{\underline{X}}\left[|\psi(\underline{X}) - \mathbb{E}\psi(\underline{X})| > t\right] \leq 2 \exp\left[-\frac{t^2}{2ml^2\|\Lambda_m\|_\infty^2}\right], \tag{1}$$

where $\|\Lambda_m\|_\infty \leq 1 + 2\sum_{k=1}^{m} \varphi(k)$.

Mixing processes

Setting

- $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{+\infty}$ *stationary*: for any t and $m, k \geq 0$, the random subsequences $(Z_t, \ldots, Z_{t+m})$ and $(Z_{t+k}, \ldots, Z_{t+m+k})$ are identically distributed
- The dependencies are fading over time, e.g., ϕ-mixing process:

$$\varphi(k) = \sup_{n, A \in \sigma_{n+k}^{+\infty}, B \in \sigma_{-\infty}^{n}} \left| \mathbb{P}[A|B] - \mathbb{P}[A] \right| .$$

  $\mathbf{Z}$ is φ-mixing if $\varphi(k) \to 0$ as $k \to 0$

Recent results

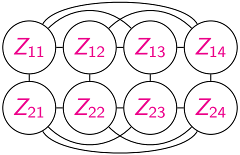
- Stability bound for β- and ϕ-mixing processes [Mohri and Rostamizadeh, 2008]
- Rademacher complexity fpr β-mixing processes [Mohri and Rostamizadeh, 2009]
- Consistency of learning in α-mixing non stationary processes [Steinwart et al., 2009]
- ...

Interdependent and Identically Distributed Data

Basic assumptions

- $\mathbf{Z}_{\text{train}} = \{Z_i\}_{i=1}^{m}$ distributed according to D_m
- $p(\mathbf{Z}_{\text{train}}) \neq \prod_{i=1}^{m} p(Z_i)$
- $p_{\text{train}}(Z_i) = p_{\text{train}}(Z) = p_{\text{test}}(Z)$ (similar to a stationarity condition)
- Goal: control the risk of a learned function wrt $p_{\text{test}}(Z)$

Illustration



Graph Fractional Chromatic Number

Definition (Dependency Graph)

Let $\mathbf{Z} = \{Z_i\}_{i=1}^m$ be a set of r.v. taking values in $\mathcal{Z}$. The *dependency graph* $\Gamma(\mathbf{Z})$ of $\mathbf{Z}$ is such that the vertices of $\Gamma(\mathbf{Z})$ are $\{1, \ldots, m\}$ and:

$$i \sim j \Leftrightarrow p(Z_i, Z_j) \neq p(Z_i)p(Z_j).$$

Definition (Fractional Covers, [Schreinerman and Ullman, 1997])

Let $\Gamma = (V, E)$ be an undirected graph, with $V = \{1, \ldots, m\}$.

- A cover $\mathbf{C} = \{C_j\}_{j=1}^n$ of Γ, with $C_j \subseteq V$, is such that no two nodes in C_j are connected
- A fractional cover $\mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n$ is a slightly refined version of a cover which assigns weights to each element of $\mathbf{C}$

Finding a minimal (fractional) cover amounts to finding a minimal coloring of Γ

$\chi(\Gamma)$ ($\chi^*(\Gamma)$) is the (fractional) chromatic number of Γ

Graph Fractional Chromatic Number

Definition (Dependency Graph)

Let $\mathbf{Z} = \{Z_i\}_{i=1}^m$ be a set of r.v. taking values in $\mathcal{Z}$. The *dependency graph* $\Gamma(\mathbf{Z})$ of $\mathbf{Z}$ is such that the vertices of $\Gamma(\mathbf{Z})$ are $\{1, \ldots, m\}$ and:

$$i \sim j \Leftrightarrow p(Z_i, Z_j) \neq p(Z_i)p(Z_j).$$

Property on $\chi(\Gamma)$ and $\chi^*(\Gamma)$[Schreinerman and Ullman, 1997]

Let $\Gamma = (V, E)$ be a graph. Let $c(\Gamma)$ be the *clique number* of Γ. Let $\Delta(\Gamma)$ be the maximum degree of a vertex in Γ. The following holds

$$1 \leq c(\Gamma) \leq \chi^*(\Gamma) \leq \chi(\Gamma) \leq \Delta(\Gamma) + 1.$$

In addition, $1 = c(\Gamma) = \chi^*(\Gamma) = \chi(\Gamma) = \Delta(\Gamma) + 1$ *if and only if* Γ is totally disconnected.

On the (fractional) chromatic number

- Computing χ and χ^* is an NP-hard problem, but...
- we will consider instances of graphs for which they can be computed

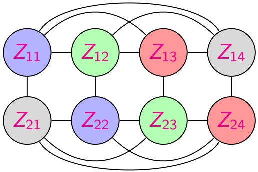
Graph Fractional Chromatic Number

Definition (Dependency Graph)

Let $\mathbf{Z} = \{Z_i\}_{i=1}^m$ be a set of r.v. taking values in $\mathcal{Z}$. The *dependency graph* $\Gamma(\mathbf{Z})$ of $\mathbf{Z}$ is such that the vertices of $\Gamma(\mathbf{Z})$ are $\{1, \ldots, m\}$ and:

$$i \sim j \Leftrightarrow p(Z_i, Z_j) \neq p(Z_i)p(Z_j).$$

Example: Bipartite Ranking



$$c = \chi^* = \chi = 4$$

Usefulness of covers

A (fractional) cover of minimal weight breaks a set of *dependent* r.v.'s into a minimal set of (large) subsets of *independent* r.v.'s

Concentration Inequalities

Theorem (McDiarmid's inequality for dependent variables)

With mild assumptions as so that $Z = f(X_1, \ldots, X_N)$ *decomposes according to a fractional cover of* $X_1, \ldots, X_N$.*, the following concentration inequalities hold:*

$$\mathbb{P}(Z - \mathbb{E}Z \geq \varepsilon) \leq \exp\left\{-\frac{N\varepsilon^2}{4\chi_f}\right\}$$

$$\mathbb{P}(\mathbb{E}Z - Z \geq \varepsilon) \leq \exp\left\{-\frac{N\varepsilon^2}{4\chi_f}\right\}$$

Theorem (Bennett's Inequality for Dependent Variables)

Suppose some mild assumptions hold, with $Z = f(X_1, \ldots, X_N)$ *which decomposes according to a fractional cover of* $X_1, \ldots, X_N$. *We have the following results:*

- *for all* $t \geq 0$

$$\mathbb{P}(Z \geq \mathbb{E}Z + t) \leq \exp\left(-\frac{v}{\chi_f} h\left(\frac{4t}{5v}\right)\right),$$

  with $h(x) = (1 + x)\log(1 + x) - x$ *and* $v \doteq (1 + b)\mathbb{E}Z + N\sigma^2$

- *for all* $t \geq 0$

$$\mathbb{P}\left(Z \geq \mathbb{E}Z + \sqrt{2cvt} + \frac{ct}{3}\right) \leq e^{-t}$$

  with $c \doteq 25\chi/16$.

Notes

- secret tool to get these concentration inequalities
- Rademacher-based bound on generalization can be obtained... and more

Outline

What to take home

IID

- Much has been done in the field of IID learning
- This assumption allows one to get strong generalization results

Non-IIDness

- No agreed-upon parametrization of non-stationarity
- A lot of work to do in online learning
- Nice tools from graph theory and concentration inequality for the dependent case

References I

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. (2002).
The nonstochastic multiarmed bandit problem.
SIAM Journal on Computing, 32(1):48–77.

Bartlett, P., Bousquet, O., and Mendelson, S. (2005).
Local rademacher complexities.
Annals of Statistics, 33(4):1497–1537.

Bartlett, P. L. and Mendelson, S. (2002).
Rademacher and gaussian complexities: Risk bounds and structural results.
Journal of Machine Learning Research, 3:463–482.

Block, H. (1962).
The perceptron: a model for brain functioning.
Reviews of Modern Physics, 34:123—135.

Bordes, A., Ertekin, S., Weston, J., and Bottou, L. (2005).
Fast kernel classifiers with online and active learning.
Journal of Machine Learning Research, 6:1579–1619.

Bousquet, O. and Elisseeff, A. (2002).
Stability and Generalization.
Journal of Machine Learning Research, 2:499–526.

References II

Bubeck, S. and Cesa-Bianchi, N. (2012).
Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems, volume 5 of *Foundation and Trends in Machine Learning*.
NOW.

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. (2007).
Tracking the best hyperplane with a simple budget perceptron.
Machine Learning, 69(2-3):143–167.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. (2004).
On the generalization ability of online learning algorithms.
IEEE Transactions on Information Theory, 50(9):2050–2057.

Cesa-Bianchi, N. and Lugosi, G. (2006).
Prediction, Learning, and Games.
Cambridge University Press, New York, NY, USA.

Chapelle, O. and Li, L. (2012).
An empirical evaluation of thompson sampling.
In *Advances in Neural Information Processing Systems 24*, pages 2249–2257.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (2006).
Online passive-aggressive algorithms.
JMLR, 7:551–585.

References III

Crammer, K., Kandola, J. S., and Singer, Y. (2003).
Online classification on a budget.
In *NIPS*. MIT Press.

Crammer, K. and Singer, Y. (2003).
Ultraconservative online algorithms for multiclass problems.
Journal of Machine Learning Research, 3:951–991.

Csato, L. and Opper, M. (2002).
Sparse Online Gaussian Processes.
Neural Computation, 14:641–668.

Dekel, O., Shalev-Shwartz, S., and Singer, Y. (2008).
The forgetron: A kernel-based perceptron on a budget.
SIAM J. Comput., 37(5):1342–1372.

Engel, Y., Mannor, S., and Meir, R. (2003).
The kernel recursive least squares algorithm.
IEEE Transactions on Signal Processing, 52:2275–2285.

Kivinen, J., Smola, A. J., and Williamson, B. (2010).
Online learning with kernels.
IEEE Transactions on Signal Processing, 100(10).

References IV

Kontorovich, L. and Ramanan, K. (2008).
Concentration inequalities for dependent random variables via the martingale method.
The Annals of Probability, 36(6):2126–2158.

Littlestone, N. (1988).
Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm.
Machine Learning, 2:285–318.

McAllester, D. (1999).
Pac-bayesian model averaging.
In *Proc. of the 12th Annual Conf. on Comp. learning theory*, pages 164–170, New York, NY, USA.

McDiarmid, C. (1989).
On the method of bounded differences.
Survey in Combinatorics, pages 148–188.

Mohri, M. and Rostamizadeh, A. (2008).
Stability Bounds for Non-i.i.d. Processes.
In *Adv. in Neural Information Processing Systems 20*, pages 1025–1032.

Mohri, M. and Rostamizadeh, A. (2009).
Rademacher Complexity Bounds for Non-I.I.D. Processes.
In *Adv. in Neural Information Processing Systems 21*, pages 1025–1032.

References V

Munos, R. (2014).

From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning, volume 7(1) of *Foundations and Trends in Machine Learning*.
NOW.

Novikoff, A. (1963).

On convergence proofs for perceptrons.

In *Proc. of the Symposium on the Mathematical Theory of Automata, Vol. 12*, pages 615—622.

Orabona, F. and Keshet, J. (2008).

The projectron: a bounded kernel-based perceptron.

In *In Proceedings of the 25th international conference on Machine learning*, pages 720–727. ACM.

Rakhlin, A., Sridharan, K., and Tewari, A. (2010a).

Online learning: Random averages, combinatorial parameters, and learnability.

In *NIPS*, pages 1984–1992.

Rakhlin, A., Sridharan, K., and Tewari, A. (2010b).

Online learning: Random averages, combinatorial parameters, and learnability.

CoRR, abs/1006.1138.

References VI

Schreinerman, E. and Ullman, D. (1997).
Fractional graph theory: A rational approach to the theory of graphs.
Wiley Interscience Series in Discrete Math.

Shalev-Shwartz, S. (2007).
Online Learning: Theory, Algorithms, and Applications.
PhD thesis, The Hebrew University of Jerusalem.

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. (2011).
Pegasos: primal estimated sub-gradient solver for svm.
Math. Program., 127(1):3–30.

Shawe-Taylor, J. and Cristianini, N. (2004).
Kernel Methods for Pattern Analysis.
Cambridge University Press.

Shimodaira, H. (2000).
Improving predictive inference under covariate shift by weighting the log-likelihood function.
Journal of Statistical Planning and Inference, 90:227–244.

Shwartz, S. S., Singer, Y., and Srebro, N. (2007).
Pegasos: Primal estimated sub-gradient solver for svm.
In *ICML '07: Proceedings of the 24th international conference on Machine learning*, New York, NY, USA.

References VII

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2006).
A Hilbert Space Embedding for Distributions.
In *Proc. of Int. Conf. on Algorithmic Learning Theory*.

Steinwart, I., Hush, D., and Scovel, C. (2009).
Learning from dependent observations.
Journal of Multivariate Analysis, 100(1):175–194.

Storkey, A. and Sugiyama, M. (2007).
Mixture regression for covariate shift.
In *Adv. in Neural Information Processing Systems*, volume 19.

Thompson, W. R. (1933).
On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.
Biometrika, 25(3-4):285—294.

Vapnik, V. (1998).
Statistical Learning Theory.
John Wiley and Sons, inc.