

Apprentissage et données non IID : quelques résultats
Séminaire I3S, 11/06/09
Sophia Antipolis

Liva Ralaivola

LIF, CNRS, Aix-Marseille Universités
39, rue F. Joliot Curie, F-13013 Marseille, France



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?

Outline

Overview

The Usual SLT Framework and Beyond

Concrete Examples

Formalizations

PAC Bayes Bounds for Interdependent Data

Tool 1: Classical IID Pac-Bayes Bound

Tool 2: Fractional Chromatic Number

New Bounds

Conclusion and Outlooks

Beyond the Common Statistical Learning Framework

Supervised Learning Setting

- ▶ \mathcal{X} input space, \mathcal{Y} target space, D distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- ▶ $S_{\text{train}} = \{(X_i, Y_i)\}_{i=1}^n$ sample of n **independently and identically (IID)** r.v. distributed according to D
- ▶ A loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$

Goal: minimizing the true risk

From S_{train} , learn a mapping $f : \mathcal{Z} \rightarrow \mathcal{Y}$ such that

$$R(f) = \int_{\mathcal{X} \times \mathcal{Y}} \ell(f(x), y) dD(x, y) = \min_h R(h)$$

For instance, in classification, we wish to minimize:

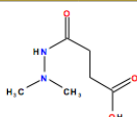


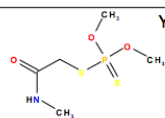
$$\mathbb{P}_{X, Y \sim D}(f(X) \neq Y)$$

Noticeable assumptions (that will not necessarily hold later on)

- ▶ Training and test data are IID
- ▶ They are distributed according to the same distribution

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**

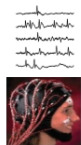
Learning f

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

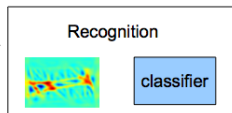
Concrete Examples

Brain computer Interface: P300 speller

EEG Signals



BCI



(from A. Rakotomamonjy)

Goal

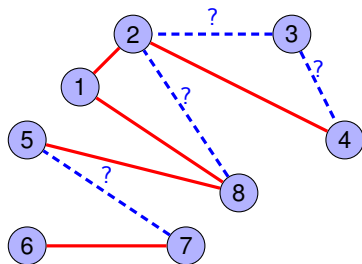
Detect P300's in EEG signal.

Nature of non-IIDness

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

Concrete Examples

Edge prediction, relational learning, etc.

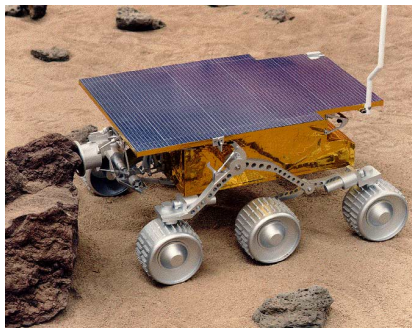


Interdependencies

- ▶ 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)

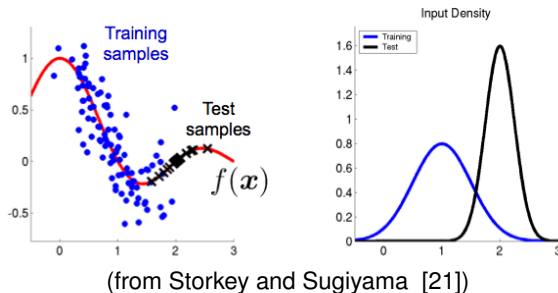
- ▶ The robot has to make a decision (e.g. {stop, right, left, forward}) at each time step t according to its environment X_t
- ▶ X_t depends on the past $X_{t'}$'s ($t' < t$) with a fading influence between the X_t 's over time (cf. mixing processes)

Covariate Shift

Pitch

“Learning when training and test distributions are different”

(NIPS 06 workshop)



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 [21, 18, 19]

Mixing processes

Setting

- ▶ $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{+\infty}$ *stationary*: for any t and $m, k \geq 0$, the random subsequences (Z_t, \dots, Z_{t+m}) and $(Z_{t+k}, \dots, Z_{t+m+k})$ 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 $\varphi(k) \rightarrow 0$ as $k \rightarrow \infty$

Recent results

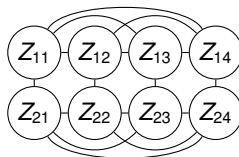
- ▶ Stability bound for β - and ϕ -mixing processes [12]
- ▶ Rademacher complexity for β -mixing processes [13]
- ▶ Consistency of learning in α -mixing non stationary processes [20]
- ▶ ...

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



Results

- ▶ Adapted concentration inequalities [7, 22]
- ▶ Chromatic PAC Bayes Bounds (LR, M. Szafranski, G. Stempfel, 09)
- ▶ ...

PAC Bayes Bounds for non-IID Data

What ?

A tight generalization bound that applies in the non-IID setting: with high probability

$$\underbrace{R(f)}_{\text{true risk}} \leq \underbrace{\hat{R}(f, \mathbf{Z})}_{\text{empirical risk}} + \underbrace{\varepsilon(m, \dots)}_{\rightarrow 0}$$

Expected features

- ▶ Genericity
- ▶ Tightness
- ▶ Easily computable
- ▶ Motivates new algorithms

Motivating example: bipartite ranking and bounding the AUC

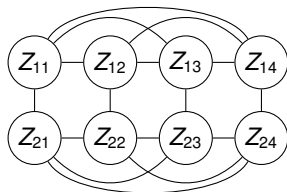
Setting (Input space $\bar{\mathcal{X}}$, target space $\bar{\mathcal{Y}} = \{-1, +1\}$, $\bar{\mathcal{Z}} = \bar{\mathcal{X}} \times \bar{\mathcal{Y}}$)

- ▶ Training set: $S = \{(\bar{X}_i, \bar{Y}_i)\}_{i=1}^{\ell} \in \bar{\mathcal{Z}}^{\ell}$, i.i.d according to D
 - ▶ ℓ^+ : number of positive data in S
 - ▶ ℓ^- : number of negative data in S
- ▶ Empirical and true ranking risks

$$\hat{R}^{\text{rank}}(f, S) = \frac{1}{\ell^+ \ell^-} \sum_{i: Y_i=1, j: Y_j=-1} \mathbb{I}_{f(X_i) < f(X_j)} \quad (= 1 - \text{AUC}(f, S))$$
$$R^{\text{rank}}(f) = \mathbb{E}_S \hat{R}^{\text{rank}}(f, S) = \mathbb{P}_{X^{\pm} \sim D^{\pm}} (f(X^+) < f(X^-))$$

Non-IIDness

$\hat{R}^{\text{rank}}(f, S)$ uses **non-independent** r.v. $Z_{ij} := (X_i^+, X_j^-) \in \mathcal{Z} := \overbrace{\bar{\mathcal{X}} \times \bar{\mathcal{X}}}^{=: \mathcal{X}} \times \overbrace{\bar{\mathcal{Y}}}^{=: \mathcal{Y}}$



Motivating example: bipartite ranking and bounding the AUC

Setting (Input space $\bar{\mathcal{X}}$, target space $\bar{\mathcal{Y}} = \{-1, +1\}$, $\bar{\mathcal{Z}} = \bar{\mathcal{X}} \times \bar{\mathcal{Y}}$)

- ▶ Training set: $S = \{(\bar{X}_i, \bar{Y}_i)\}_{i=1}^{\ell} \in \bar{\mathcal{Z}}^{\ell}$, i.i.d according to D
 - ▶ ℓ^+ : number of positive data in S
 - ▶ ℓ^- : number of negative data in S
- ▶ Empirical and true ranking risks

$$\begin{aligned}\hat{R}^{\text{rank}}(f, S) &= \frac{1}{\ell^+ \ell^-} \sum_{i: Y_i=1, j: Y_j=-1} \mathbb{I}_{f(X_i) < f(X_j)} && (= 1 - \text{AUC}(f, S)) \\ R^{\text{rank}}(f) &= \mathbb{E}_S \hat{R}^{\text{rank}}(f, S) = \mathbb{P}_{X^{\pm} \sim D^{\pm}}(f(X^+) < f(X^-))\end{aligned}$$

The Question

How to bound the true risk $R^{\text{rank}}(f)$ of f based on $\hat{R}^{\text{rank}}(f, S)$ given that there is some non-IIDness in the data used to evaluate $\hat{R}^{\text{rank}}(f, S)$.

$$R^{\text{rank}}(f) \leq \hat{R}^{\text{rank}}(f, S) + \varepsilon(\ell, \dots)$$

Tool 1: IID Pac-Bayes Bound

Theorem (IID PAC-Bayes Bound, [10, 17])

$\forall m, \forall D, \forall \mathcal{H}, \forall \delta \in (0, 1], \forall P$, with probability at least $1 - \delta$ over the random draw of $\mathbf{Z} \sim \mathbf{D}_m = D^m$, the following holds:

$$\forall Q, kl(\hat{e}_Q || e_Q) \leq \frac{1}{m} \left[KL(Q || P) + \ln \frac{m+1}{\delta} \right],$$

where, for $Z = (X, Y)$, $r(h, Z) = \mathbb{I}_{h(X) \neq Y}$, and,

$$\hat{e}_Q = \mathbb{E}_{h \sim Q} \frac{1}{m} \sum_{i=1}^m r(h, Z_i) = \mathbb{E}_{h \sim Q} \hat{R}(h, \mathbf{Z})$$

$$e_Q = \mathbb{E}_{\mathbf{Z} \sim \mathbf{D}_m} \hat{e}_Q = \mathbb{E}_{\substack{Z \sim D \\ h \sim Q}} r(h, Z) = \mathbb{E}_{h \sim Q} R(h).$$

Tool 1: IID Pac-Bayes Bound

Theorem (IID PAC-Bayes Bound, [10, 17])

$\forall m, \forall D, \forall \mathcal{H}, \forall \delta \in (0, 1], \forall P$, with probability at least $1 - \delta$ over the random draw of $\mathbf{Z} \sim \mathbf{D}_m = D^m$, the following holds:

$$\forall Q, kl(\hat{e}_Q || e_Q) \leq \frac{1}{m} \left[KL(Q || P) + \ln \frac{m+1}{\delta} \right],$$

Comments

- ▶ Implies: $e_Q \leq \hat{e}_Q + \sqrt{\frac{2}{m} [KL(Q || P) + \ln \frac{m+1}{\delta}]}$
- ▶ Bounds the risk of the *Gibbs classifier*
- ▶ P and Q *structure* the hypothesis space
- ▶ When P and Q are specialized, the bound can be tight and computable

Tool 1: IID Pac-Bayes Bound

Theorem (IID PAC-Bayes Bound, [10, 17])

$\forall m, \forall D, \forall \mathcal{H}, \forall \delta \in (0, 1], \forall P$, with probability at least $1 - \delta$ over the random draw of $\mathbf{Z} \sim \mathbf{D}_m = D^m$, the following holds:

$$\forall Q, kl(\hat{e}_Q || e_Q) \leq \frac{1}{m} \left[KL(Q || P) + \ln \frac{m+1}{\delta} \right],$$

Elements of proof: extremely simple

Message r.v. $e^{mkl(\hat{R}(h, \mathbf{Z}) || R(h))}$ softly

- ▶ Bound its expectation (wrt to \mathbf{Z} and h) and use Markov's inequality
- ▶ Use Jensen's inequality to introduce Q and to retrieve \hat{e}_Q and e_Q

Tool 2: 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 $\{1, \dots, m\}$ and:

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

Definition (Fractional Covers, [16])

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

- ▶ 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

Tool 2: 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 $\{1, \dots, m\}$ and:

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

Property on $\chi(\Gamma)$ and $\chi^*(\Gamma)$ [16]

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

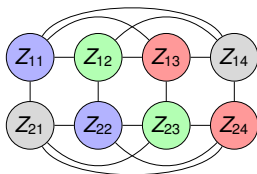
Tool 2: Graph Fractional Chromatic Number

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 $\{1, \dots, m\}$ 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

Chromatic Pac-Bayes Bounds

Theorem (LR, M. Szafranski, G. Stempfel [14])

$\forall m, \forall \mathbf{D}_m, \forall \mathcal{H}, \forall \delta \in (0, 1], \forall P$, with probability at least $1 - \delta$ over the random draw of $\mathbf{Z} \sim \mathbf{D}_m$, the following holds

$$\forall Q, kl(\hat{e}_Q \| e_Q) \leq \frac{\chi^*}{m} \left[KL(Q \| P) + \ln \frac{m + \chi^*}{\delta \chi^*} \right], \quad (1)$$

where χ^* is the fractional chromatic number of $\Gamma(\mathbf{D}_m)$.

Remarks

- ▶ The bound is “cover independent”
- ▶ Comes down to the IID Pac-Bayes bound when data are IID ($\chi^* = \chi = 1$)

A high-level view

- ▶ Break the dependency graph into independent subsets
- ▶ Apply the IID PAC-Bayes bound to each subset
- ▶ Gather everything

Ranking bound with linear classifiers

Theorem (Linear Ranking/AUC Pac-Bayes Bound)

$\forall \ell, \forall \bar{D}$ over $\bar{\mathcal{X}} \times \bar{\mathcal{Y}}, \forall \delta \in (0, 1],$ the following holds with probability at least $1 - \delta$ over the draw of $\mathbf{S} \sim \bar{D}^\ell$:

$$\forall \mathbf{w}, \mu > 0, kl(\hat{R}_{Q_{\mathbf{w}, \mu}}^{\text{rank}} || R_{Q_{\mathbf{w}, \mu}}^{\text{rank}}) \leq \frac{1}{\ell_{\min}} \left[\frac{\mu^2}{2} + \ln \frac{\ell_{\min} + 1}{\delta} \right].$$

Where the prior P is $\mathcal{N}(0, I)$, the posterior $Q_{\mathbf{w}, \mu}$ is $\mathcal{N}(\mu, 1)$ in the direction of \mathbf{w} and $\mathcal{N}(0, 1)$ in all other directions.

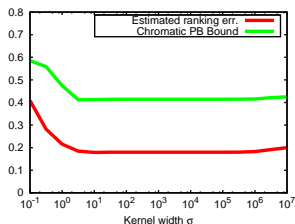
Comments

- ▶ For rotational invariant distributions wrt \mathbf{w} , \mathbf{w} is Bayes equivalent to the voting classifier
- ▶ The bound is tightened by adjusting μ
- ▶ Yet tighter bounds can be obtained by using an appropriate prior P

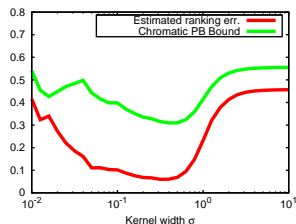
Ranking bound with linear classifiers

Ranking experiments, preliminary results (UCI datasets, [15])

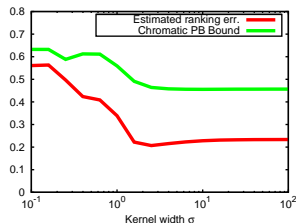
Gaussian kernel: $k(\mathbf{x}, \mathbf{x}') = \exp(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2})$. Homemade ranker.



Diabetis (268 / 500)



Banana (217 / 183)



German (211 / 489)

Related Work

Fractional Covers/Concentration Inequalities/Generalization bounds

- ▶ Fractional Graph Theory [16]
- ▶ Generalization of Chernoff bound, Svante Janson [7] (+ refs therein)
- ▶ Generalization McDiarmid's inequality, Fractional Rademacher Complexity, Usunier et al [22]

PAC Bayes bounds

- ▶ Base results, McAllester [10], Seeger [17], Langford [8], Catoni [5]
- ▶ Prior learning and bound tightening, Ambroladze et al. [3]
- ▶ Averaging classifiers, Langford et al. [9], Meir et al. [11]

Other bounds/results

- ▶ AUC: Agarwal [1, 2], Clemencon [6]
- ▶ Mixing processes: Mohri [12, 13], Steinwart [20]
- ▶ Occam's Hammer, Blanchard and Fleuret [4]
- ▶ ...

Conclusion and Outlooks

New PAC-Bayes Bounds

- ▶ Based on fractional covers of graphs
- ▶ A generalization of the IID PAC Bayes bound
- ▶ New bounds for bipartite ranking
- ▶ Encouraging preliminary simulation results
- ▶ New bounds for ϕ -mixing processes (not shown here)

Outlooks

- ▶ More numerical simulations, with learned priors
- ▶ New algorithms
- ▶ Connection to variational approximation
- ▶ Connection with other approaches to handle non-IID data (weakly dependent variables)
- ▶ Bayesian prediction
- ▶ Use of spectral graph theory



Learning from non-IID data: Theory, Algorithms and Practice

ECML 2009 LNIID Workshop

7 September 2009, Bled, Slovenia

<http://www-connex.lip6.fr/~amini/ecml-wk-lniid.html>

by M. Amini, A. Habrard, LR, N. Usunier

References I

- [1] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization Bounds for the Area Under the ROC Curve. *J. of Machine Learning Research*, 6:393–425, 2005.
- [2] S. Agarwal and P. Niyogi. Stability and generalization of bipartite ranking algorithms. In *COLT*, pages 32–47, 2005.
- [3] A. Ambroladze, E. Parrado-Hernandez, and J. Shawe-Taylor. Tighter PAC-Bayes Bounds. In *Adv. in Neural Information Processing Systems 19*, pages 9–16, 2007.
- [4] G. Blanchard and F. Fleuret. Occam's hammer. In *COLT*, pages 112–126, 2007.
- [5] O. Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56. 2007.
- [6] S. Cléménçon, G. Lugosi, and N. Vayatis. Ranking and scoring using empirical risk minimization. In *COLT*, pages 1–15, 2005.
- [7] S. Janson. Large Deviations for Sums of Partly Dependent Random Variables. *Random Structures Algorithms*, 24:234–248, 2004.
- [8] J. Langford. Tutorial on Practical Theory for Classification. *J. of Machine Learning Research*, pages 273–306, 2005.
- [9] J. Langford, M. Seeger, and N. Megiddo. An improved predictive accuracy bound for averaging classifiers. In *In Proceeding of the Eighteenth International Conference on Machine Learning*, pages 290–297, 2001.
- [10] D. McAllester. Some PAC-Bayesian Theorems. *Machine Learning*, 37:355–363, 1999.
- [11] R. Meir, T. Zhang, T. Graepel, and R. Herbrich. Generalization error bounds for Bayesian mixture algorithms. *J. of Machine Learning Research*, 4:2003, 2003.
- [12] M. Mohri and A. Rostamizadeh. Stability Bounds for Non-i.i.d. Processes. In *Adv. in Neural Information Processing Systems 20*, pages 1025–1032, 2008.

References II

- [13] M. Mohri and A. Rostamizadeh. Rademacher Complexity Bounds for Non-I.I.D. Processes. In *Adv. in Neural Information Processing Systems 21*, pages 1025–1032, 2009.
- [14] L. Ralaivola, M. Szafranski, and G. Stempfel. Chromatic Pac-Bayes Bounds for non-IID Data. In *JMLR Workshop and Conference Proc.*, volume 5, 2009.
- [15] G. Rätsch, T. Onoda, and K.-R. Müller. Soft Margins for AdaBoost. *Machine Learning*, 42:287–320, 2001.
- [16] E.R. Schreinerman and D.H. Ullman. *Fractional graph theory: A rational approach to the theory of graphs*. Wiley Interscience Series in Discrete Math., 1997.
- [17] M. Seeger. PAC-Bayesian generalization bounds for gaussian processes. *J. of Machine Learning Research*, 3:233–269, 2002.
- [18] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227244, 2000.
- [19] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert Space Embedding for Distributions. In *Proc. of Int. Conf. on Algorithmic Learning Theory*, 2006.
- [20] I. Steinwart, D. Hush, and C. Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009.
- [21] A. Storkey and M. Sugiyama. Mixture regression for covariate shift. In *Adv. in Neural Information Processing Systems*, volume 19, 2007.
- [22] N. Usunier, M.-R. Amini, and P. Gallinari. Generalization Error Bounds for Classifiers Trained with Interdependent Data. In *Adv. in Neural Information Processing Systems 18*, pages 1369–1376, 2006.