New Analysis and Algorithm for Learning with Drifting Distributions

Mehryar Mohri[1,2] and Andres Muñoz Medina[1]

[1] Courant Institute of Mathematical Sciences, New York, NY.
[2] Google Research, New York, NY.

Abstract. We present a new analysis of the problem of learning with drifting distributions in the batch setting using the notion of discrepancy. We prove learning bounds based on the Rademacher complexity of the hypothesis set and the discrepancy of distributions both for a drifting PAC scenario and a tracking scenario. Our bounds are always tighter and in some cases substantially improve upon previous ones based on the L_1 distance. We also present a generalization of the standard on-line to batch conversion to the drifting scenario in terms of the discrepancy and arbitrary convex combinations of hypotheses. We introduce a new algorithm exploiting these learning guarantees, which we show can be formulated as a simple QP. Finally, we report the results of preliminary experiments demonstrating the benefits of this algorithm.

Keywords: Drifting environment, generalization bound, domain adaptation.

1 Introduction

In the standard PAC model [1] and other similar theoretical models of learning [2], the distribution according to which training and test points are drawn is fixed over time. However, for many tasks such as spam detection, political sentiment analysis, financial market prediction under mildly fluctuating economic conditions, or news stories, the learning environment is not stationary and there is a continuous drift of its parameters over time.

There is a large body of literature devoted to the study of related problems both in the on-line and the batch learning scenarios. In the on-line scenario, the target function is typically assumed to be fixed but no distributional assumption is made, thus input points may be chosen adversarially [3]. Variants of this model where the target is allowed to change a fixed number of times have also been studied [3, 4, 5, 6]. In the batch scenario, the case of a fixed input distribution with a drifting target was originally studied by Helmbold and Long [7]. A more general scenario was introduced by Bartlett [8] where the joint distribution over the input and labels could drift over time under the assumption that the L_1 distance between the distributions in two consecutive time steps was bounded. Both generalization bounds and lower bounds have been given for this scenario [9, 10]. In particular, Long [9] showed that if the L_1 distance between two consecutive distributions is at most Δ, then a generalization error of $O((d\Delta)^{1/3})$ is achievable and Barve and Long [10] proved this bound to be tight. Further improvements were presented by Freund and Mansour [14] under the assumption of a constant

rate of change for drifting. Other settings allowing arbitrary but infrequent changes of the target have also been studied [15]. An intermediate model of drift based on a `near` relationship was also recently introduced and analyzed by [16] where consecutive distributions may change arbitrarily, modulo the restriction that the region of disagreement between nearby functions would only be assigned limited distribution mass at any time.

This paper deals with the analysis of learning in the presence of drifting distributions in the batch setting. We consider both the general drift model introduced by [8] and a related drifting PAC model that we will later describe. We present new generalization bounds for both models (Sections 3 and 4). Unlike the L_1 distance used by previous authors to measure the distance between distributions, our bounds are based on a notion of *discrepancy* between distributions generalizing the definition originally introduced by [17] in the context of domain adaptation. The L_1 distance used in previous analyses admits several drawbacks: in general, it can be very large, even in favorable learning scenarios; it ignores the loss function and the hypothesis set used; and it cannot be accurately and efficiently estimated from finite samples (see for example lower bounds on the sample complexity of testing closeness by [18]). In contrast, the discrepancy takes into consideration both the loss function and the hypothesis set.

The learning bounds we present in Sections 3 and 4 are tighter than previous bounds both because they are given in terms of the discrepancy which lower bounds the L_1 distance, and because they are given in terms of the Rademacher complexity instead of the VC-dimension. Additionally, our proofs are often simpler and more concise. We also present a generalization of the standard on-line to batch conversion to the scenario of drifting distributions in terms of the discrepancy measure (Section 5). Our guarantees hold for convex combinations of the hypotheses generated by an on-line learning algorithm. These bounds lead to the definition of a natural meta-algorithm which consists of selecting the convex combination of weights in order to minimize the discrepancy-based learning bound (Section 6). We show that this optimization problem can be formulated as a simple QP and report the results of preliminary experiments demonstrating its benefits. Finally we will discuss the practicality of our algorithm in some natural scenarios.

2 Preliminaries

In this section, we introduce some preliminary notation and key definitions, including that of the *discrepancy* between distributions, and describe the learning scenarios we consider.

Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the output space. We consider a loss function $L\colon \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_+$ bounded by some constant $M > 0$. For any two functions $h, h'\colon \mathcal{X} \to \mathcal{Y}$ and any distribution D over $\mathcal{X}\times\mathcal{Y}$, we denote by $\mathcal{L}_D(h)$ the expected loss of h and by $\mathcal{L}_D(h, h')$ the expected loss of h with respect to h':

$$\mathcal{L}_D(h) = \mathop{\mathrm{E}}_{(x,y)\sim D}[L(h(x), y)] \qquad \text{and} \qquad \mathcal{L}_D(h, h') = \mathop{\mathrm{E}}_{x\sim D^1}[L(h(x), h'(x))], \quad (1)$$

where D^1 is the marginal distribution over $\mathcal{X}$ derived from D. We adopt the standard definition of the empirical Rademacher complexity, but we will need the following sequential definition of a Rademacher complexity, which is related to that of [19].

Definition 1. *Let G be a family of functions mapping from a set $\mathcal{Z}$ to $\mathbb{R}$ and $S = (z_1, \ldots, z_T)$ a fixed sample of size T with elements in $\mathcal{Z}$. The* empirical Rademacher complexity *of G for the sample S is defined by:*

$$\widehat{\mathfrak{R}}_S(G) = \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{g \in G} \frac{1}{T} \sum_{t=1}^{T} \sigma_t g(z_t)\right], \tag{2}$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_T)^\top$, with σ_ts independent uniform random variables taking values in $\{-1, +1\}$. The Rademacher complexity *of G is the expectation of $\widehat{\mathfrak{R}}_S(G)$ over all samples $S = (z_1, \ldots, z_T)$ of size T drawn according to the product distribution $D = \bigotimes_{t=1}^{T} D_t$:*

$$\mathfrak{R}_T(G) = \mathop{\mathrm{E}}_{S \sim D}[\widehat{\mathfrak{R}}_S(G)]. \tag{3}$$

Note that this coincides with the standard Rademacher complexity when the distributions D_t, $t \in [1, T]$, all coincide.

A key question for the analysis of learning with a drifting scenario is a measure of the difference between two distributions D and D'. The distance used by previous authors is the L_1 distance. However, the L_1 distance is not helpful in this context since it can be large even in some rather favorable situations. Moreover, the L_1 distance cannot be accurately and efficiently estimated from finite samples and it ignores the loss function used. Thus, we will adopt instead the *discrepancy*, which provides a measure of the dissimilarity of two distributions that takes into consideration both the loss function and the hypothesis set used, and that is suitable to the specific scenario of drifting.

Our definition of discrepancy is a generalization to the drifting context of the one introduced by [17] for the analysis of domain adaptation. Observe that for a fixed hypothesis $h \in H$, the quantity of interest with drifting distributions is the difference of the expected losses $\mathcal{L}_{D'}(h) - \mathcal{L}_D(h)$ for two consecutive distributions D and D'. A natural distance between distributions in this context is thus one based on the supremum of this quantity over all $h \in H$.

Definition 2. *Given a hypothesis set H and a loss function L, the* $\mathcal{Y}$-discrepancy $\mathrm{disc}_{\mathcal{Y}}$ *between two distributions D and D' over $\mathcal{X} \times \mathcal{Y}$ is defined by:*

$$\mathrm{disc}_{\mathcal{Y}}(D, D') = \sup_{h \in H} \left|\mathcal{L}_{D'}(h) - \mathcal{L}_D(h)\right|. \tag{4}$$

In a deterministic learning scenario with a labeling function f, the previous definition becomes

$$\mathrm{disc}_{\mathcal{Y}}(D, D') = \sup_{h \in H} \left|\mathcal{L}_{D'^1}(f, h) - \mathcal{L}_{D^1}(f, h)\right|, \tag{5}$$

where D'^1 and D^1 are the marginal distributions associated to D and D' defined over $\mathcal{X}$. The target function f is unknown and could match any hypothesis h'. This leads to the following definition [17].

Definition 3. *Given a hypothesis set H and a loss function L, the* discrepancy disc *between two distributions D and D' over $\mathcal{X} \times \mathcal{Y}$ is defined by:*

$$\mathrm{disc}(D, D') = \sup_{h, h' \in H} \left|\mathcal{L}_{D'^1}(h', h) - \mathcal{L}_{D^1}(h', h)\right|. \tag{6}$$

An important advantage of this last definition of discrepancy, in addition to those already mentioned, is that it can be accurately estimated from finite samples drawn from D'^1 and D^1 when the loss is bounded and the Rademacher complexity of the family of functions $L_H = \{x \mapsto L(h'(x), h(x)) \colon h, h' \in H\}$ is in $O(1/\sqrt{T})$, where T is the sample size; in particular when L_H has a finite pseudo-dimension [17]. The discrepancy is by definition symmetric and verifies the triangle inequality for any loss function L. In general, it does not define a *distance* since we may have $\mathrm{disc}(D, D') = 0$ for $D' \neq D$. However, in some cases, for example for kernel-based hypothesis sets based on a Gaussian kernel, the discrepancy has been shown to be a distance [20].

We will present our learning guarantees in terms of the $\mathcal{Y}$-discrepancy $\mathrm{disc}_{\mathcal{Y}}$, that is the most general definition since guarantees in terms of the discrepancy disc can be straightforwardly derived from them. The advantage of the latter bounds is the fact that the discrepancy can be estimated in that case from unlabeled finite samples.

We will consider two different scenarios for the analysis of learning with drifting distributions: the *drifting PAC scenario* and the *drifting tracking scenario*.

The drifting PAC scenario is a natural extension of the PAC scenario, where the objective is to select a hypothesis h out of a hypothesis set H with a small expected loss according to the distribution D_{T+1} after receiving a sample of $T \geq 1$ instances drawn from the product distribution $\bigotimes_{t=1}^{T} D_t$. Thus, the focus in this scenario is the performance of the hypothesis h with respect to the environment distribution after receiving the training sample.

The drifting tracking scenario we consider is based on the scenario originally introduced by [8] for the zero-one loss and is used to measure the performance of an algorithm $\mathcal{A}$ (as opposed to any hypothesis h). In that learning model, the performance of an algorithm is determined based on its average predictions at each time for a sequence of distributions. We will generalize its definition by using the notion of discrepancy and extending it to other loss functions. The following definitions are the key concepts defining this model.

Definition 4. *For any sample $S = (x_t, y_t)_{t=1}^{T}$ of size T, we denote by $h_{T-1} \in H$ the hypothesis returned by an algorithm $\mathcal{A}$ after receiving the first $T-1$ examples and by $\widehat{M}_T$ its loss or mistake on x_T: $\widehat{M}_T = L(h_{T-1}(x_T), y_T)$. For a product distribution $D = \bigotimes_{t=1}^{T} D_t$ on $(\mathcal{X} \times \mathcal{Y})^T$ we denote by $M_T(D)$ the expected mistake of $\mathcal{A}$:*

$$M_T(D) = \mathop{\mathrm{E}}_{S \sim D}[\widehat{M}_T] = \mathop{\mathrm{E}}_{S \sim D}[L(h_{T-1}(x_T), y_T)].$$

Definition 5. *Let $\Delta > 0$ and let $\widetilde{M}_T$ be the supremum of $M_T(D)$ over all distribution sequences $D = (D_t)$, with $\mathrm{disc}_{\mathcal{Y}}(D_t, D_{t+1}) < \Delta$. Algorithm $\mathcal{A}$ is said to* (Δ, ϵ)-track *H if there exists t_0 such that for $T > t_0$ we have $\widetilde{M}_T < \inf_{h \in H} \mathcal{L}_{D_T}(h) + \epsilon$.*

An analysis of the tracking scenario with the L_1 distance used to measure the divergence of distributions instead of the discrepancy was carried out by Long [9] and Barve and Long [10], including both upper and lower bounds for $\widetilde{M}_T$ in terms of Δ. Their analysis makes use of an algorithm very similar to empirical risk minimization, which we will also use in our theoretical analysis of both scenarios.

3 Drifting PAC scenario

In this section, we present guarantees for the drifting PAC scenario in terms of the discrepancies of D_t and D_{T+1} , $t \in [1, T]$, and the Rademacher complexity of the hypothesis set. We start with a generalization bound in this scenario and then present a bound for the agnostic learning setting.

Let us emphasize that learning bounds in the drifting scenario should of course not be expected to converge to zero as a function of the sample size but depend instead on the divergence between distributions.

Theorem 1. *Assume that the loss function L is bounded by M. Let $D_1, \ldots, D_{T+1}$ be a sequence of distributions and let $H_L = \{(x, y) \mapsto L(h(x), y) \colon h \in H\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in H$:*

$$\mathcal{L}_{D_{T+1}}(h) \leq \frac{1}{T}\sum_{t=1}^{T} L(h(x_t), y_t) + 2\mathfrak{R}_T(H_L) + \frac{1}{T}\sum_{t=1}^{T} \mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1}) + M\sqrt{\frac{\log \frac{1}{\delta}}{2T}}.$$

Proof. We denote by D the product distribution $\bigotimes_{t=1}^{T} D_t$. Let Φ be the function defined over any sample $S = ((x_1, y_1), \ldots, (x_T, y_T)) \in (\mathcal{X} \times \mathcal{Y})^T$ by

$$\Phi(S) = \sup_{h \in H} \mathcal{L}_{D_{T+1}}(h) - \frac{1}{T}\sum_{t=1}^{T} L(h(x_t), y_t).$$

Let S and S' be two samples differing by one labeled point, say (x_t, y_t) in S and (x'_t, y'_t) in S', then:

$$\Phi(S') - \Phi(S) \leq \sup_{h \in H} \frac{1}{T}\Big[L(h(x'_t), y'_t) - L(h(x_t), y_t)\Big] \leq \frac{M}{T}.$$

Thus, by McDiarmid's inequality, the following holds:[3]

$$\Pr_{S \sim D}\Big[\Phi(S) - \mathop{\mathrm{E}}_{S \sim D}[\Phi(S)] > \epsilon\Big] \leq \exp(-2T\epsilon^2/M^2).$$

We now bound $\mathrm{E}_{S \sim D}[\Phi(S)]$ by first rewriting it, as follows:

$$\begin{aligned}
&\mathrm{E}\Big[\sup_{h \in H} \mathcal{L}_{D_{T+1}}(h) - \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_{D_t}(h) + \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_{D_t}(h) - \frac{1}{T}\sum_{t=1}^{T} L(h(x_t), y_t)\Big] \\
&\leq \mathrm{E}\Big[\sup_{h \in H} \mathcal{L}_{D_{T+1}}(h) - \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_{D_t}(h)\Big] + \mathrm{E}\Big[\sup_{h \in H} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_{D_t}(h) - \frac{1}{T}\sum_{t=1}^{T} L(h(x_t), y_t)\Big] \\
&\leq \mathrm{E}\Big[\frac{1}{T}\sum_{t=1}^{T} \sup_{h \in H}\big(\mathcal{L}_{D_{T+1}}(h) - \mathcal{L}_{D_t}(h)\big) + \sup_{h \in H} \frac{1}{T}\sum_{t=1}^{T}\big(\mathcal{L}_{D_t}(h) - L(h(x_t), y_t)\big)\Big] \\
&\leq \frac{1}{T}\sum_{t=1}^{T} \mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1}) + \mathrm{E}\Big[\sup_{h \in H} \frac{1}{T}\sum_{t=1}^{T}\big(\mathcal{L}_{D_t}(h) - L(h(x_t), y_t)\big)\Big].
\end{aligned}$$

[3] Note that McDiarmid's inequality does not require points to be drawn according to the same distribution but only that they would be drawn independently.

It is not hard to see, using a symmetrization argument as in the non-sequential case, that the second term can be bounded by $2\mathfrak{R}_T(H_L)$. □

For many commonly used loss functions, the empirical Rademacher complexity $\mathfrak{R}_T(H_L)$ can be upper bounded in terms of that of the function class H. In particular, for the zero-one loss it is known that $\mathfrak{R}_T(H_L) = \mathfrak{R}_T(H)/2$ and when L is the L_q loss for some $q \geq 1$, that is $L(y, y') = |y' - y|^q$ for all $y, y' \in \mathcal{Y}$, then $\mathfrak{R}_T(H_L) \leq qM^{q-1}\mathfrak{R}_T(H)$. Indeed, since $x \mapsto |x|^q$ is qM^{q-1}-Lipschitz over $[-M, +M]$, by Talagrand's contraction lemma, $\mathfrak{R}_T(H_L)$ is bounded by $qM^{q-1}\widehat{\mathfrak{R}}_T(G)$ with $G = \{(x, y) \mapsto (h(x) - y) \colon h \in H\}$. Furthermore, $\widehat{\mathfrak{R}}_T(G)$ can be analyzed as follows:

$$\widehat{\mathfrak{R}}_T(G) = \frac{1}{T}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in H}\sum_{t=1}^{T}\sigma_t(h(x_t) - y_t)\right]$$
$$= \frac{1}{T}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in H}\sum_{t=1}^{T}\sigma_t h(x_t)\right] + \frac{1}{T}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T}-\sigma_t y_t\right] = \widehat{\mathfrak{R}}_T(H),$$

since $\mathrm{E}_{\boldsymbol{\sigma}}[\sum_{t=1}^{T} -\sigma_t y_t] = 0$. Taking the expectation of both sides yields a similar inequality for Rademacher complexities. Thus, in the statement of the previous theorem, $\mathfrak{R}_T(H_L)$ can be replaced with $qM^{q-1}\mathfrak{R}_T(H)$ when L is the L_q loss.

Observe that the bound of Theorem 1 is tight as a function of the divergence measure (discrepancy) we are using. Consider for example the case where $D_1 = \ldots = D_T$, then a standard Rademacher complexity generalization bound holds for all $h \in H$:

$$\mathcal{L}_{D_T}(h) \leq \frac{1}{T}\sum_{t=1}^{T} L(h(x_t), y_t) + 2\mathfrak{R}_T(H_L) + O(1/\sqrt{T}).$$

Now, our generalization bound for $\mathcal{L}_{D_{T+1}}(h)$ includes only the additive term $\mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1})$, but by definition of the discrepancy, for any $\epsilon > 0$, there exists $h \in H$ such that the inequality $|\mathcal{L}_{D_{T+1}}(h) - \mathcal{L}_{D_T}(h)| < \mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1}) + \epsilon$ holds.

Next, we present PAC learning bounds for empirical risk minimization. Let h_T^* be a best-in class hypothesis in H, that is one with the best expected loss. By a similar reasoning as in theorem 1, we can show that with probability $1 - \frac{\delta}{2}$ we have

$$\frac{1}{T}\sum_{t=1}^{T} L(h_T^*(x_t), y_t) \leq \mathcal{L}_{D_{T+1}}(h_T^*) + 2\mathfrak{R}_T(H_L) + \frac{1}{T}\sum_{t=1}^{T}\mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1}) + 2M\sqrt{\frac{\log\frac{2}{\delta}}{2T}}.$$

Let h_T be a hypothesis returned by empirical risk minimization (ERM). Combining this inequality with the bound of theorem 1 while using the definition of h_T and using the union bound, we obtain that with probability $1 - \delta$ the following holds:

$$\mathcal{L}_{D_{T+1}}(h_T) - \mathcal{L}_{D_{T+1}}(h_T^*) \leq 4\mathfrak{R}_T(H_L) + \frac{2}{T}\sum_{t=1}^{T}\mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1}) + 2M\sqrt{\frac{\log\frac{2}{\delta}}{2T}}. \quad (7)$$

This learning bound indicates a trade-off: larger values of the sample size T guarantee smaller first and third terms; however, as T increases, the average discrepancy term is

likely to grow as well, thereby making learning increasingly challenging. This suggests an algorithm similar to empirical risk minimization but limited to the last m examples instead of the whole sample with $m < T$. This algorithm was previously used in [10] for the study of the tracking scenario. We will use it here to prove several theoretical guarantees in the PAC learning model.

Proposition 1. *Let $\Delta \geq 0$. Assume that $(D_t)_{t\geq 0}$ is a sequence of distributions such that $\mathrm{disc}_{\mathcal{Y}}(D_t, D_{t+1}) \leq \Delta$ for all $t \geq 0$. Fix $m \geq 1$ and let h_T denote the hypothesis returned by the algorithm $\mathcal{A}$ that minimizes $\sum_{t=T-m}^{T} L(h(x_t), y_t)$ after receiving $T > m$ examples. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following learning bound holds:*

$$\mathcal{L}_{D_{T+1}}(h_T) - \inf_{h\in H} \mathcal{L}_{D_{T+1}}(h) \leq 4\mathfrak{R}_m(H_L) + (m+1)\Delta + 2M\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \quad (8)$$

Proof. The proof is straightforward. Notice that the algorithm discards the first $T-m$ examples and considers exactly m instances. Thus, as in inequality 7, we have:

$$\mathcal{L}_{D_{T+1}}(h_T) - \mathcal{L}_{D_{T+1}}(h_T^*) \leq 4\mathfrak{R}_m(H_L) + \frac{2}{m}\sum_{t=T-m}^{T} \mathrm{disc}(D_t, D_{T+1}) + 2M\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Now, we can use the triangle inequality to bound $\mathrm{disc}(D_t, D_{T+1})$ by $(T+1-m)\Delta$. Thus, the sum of the discrepancy terms can be bounded by $(m+1)\Delta$. □

To obtain the best learning guarantee, we can select m to minimize the bound just presented. This requires the expression of the Rademacher complexity in terms of m. The following is the result obtained when using a VC-dimension upper bound of $O(\sqrt{d/m})$ for the Rademacher complexity.

Corollary 1. *Fix $\Delta > 0$. Let H be a hypothesis set with VC-dimension d such that for all $m \geq 1$, $\mathfrak{R}_m(H_L) \leq \frac{C}{4}\sqrt{\frac{d}{m}}$ for some constant $C > 0$. Assume that $(D_t)_{t>0}$ is a sequence of distributions such that $\mathrm{disc}_{\mathcal{Y}}(D_t, D_{t+1}) \leq \Delta$ for all $t \geq 0$. Then, there exists an algorithm $\mathcal{A}$ such that for any $\delta > 0$, the hypothesis h_T it returns after receiving $T > \left[\frac{C+C'}{2}\right]^{\frac{2}{3}}(\frac{d}{\Delta^2})^{\frac{1}{3}}$ instances, where $C' = 2M\sqrt{\frac{\log(\frac{2}{\delta})}{2d}}$, satisfies the following with probability at least $1-\delta$:*

$$\mathcal{L}_{D_{T+1}}(h_T) - \inf_{h\in H} \mathcal{L}_{D_{T+1}}(h) \leq 3\left[\frac{C+C'}{2}\right]^{2/3}(d\Delta)^{1/3} + \Delta. \quad (9)$$

Proof: Fix $\delta > 0$. Replacing $\mathfrak{R}_m(H_L)$ by the upper bound $\frac{C}{4}\sqrt{\frac{d}{m}}$ in (8) yields

$$\mathcal{L}_{D_{T+1}}(h_T) - \inf_{h\in H} \mathcal{L}_{D_{T+1}}(h) \leq (C+C')\sqrt{\frac{d}{m}} + (m+1)\Delta.$$

Choosing $m = (\frac{C+C'}{2})^{\frac{2}{3}}(\frac{d}{\Delta^2})^{\frac{1}{3}}$ to minimize the right-hand side gives exactly (9). □

When H has finite VC-dimension d, it is known that $\mathfrak{R}_m(H_L)$ can be bounded by $C\sqrt{d/m}$ for some constant $C > 0$, by using a chaining argument [11, 12, 13]. Thus, the assumption of the corollary holds for many loss functions L, when H has finite VC-dimension.

4 Drifting Tracking scenario

In this section, we present a simpler proof of the bounds given by [9] for the agnostic case demonstrating that using the discrepancy as a measure of the divergence between distributions leads to tighter and more informative bounds than using the L_1 distance.

Proposition 2. *Let $\Delta > 0$ and let $(D_t)_{t\geq 0}$ be a sequence of distributions such that $\mathrm{disc}_{\mathcal{Y}}(D_t, D_{t+1}) \leq \Delta$ for all $t \geq 0$. Let $m > 1$ and let h_T be as in proposition 1. Then,*

$$\mathop{\mathrm{E}}_{D}[\widehat{M}_{T+1}] - \inf_h \mathcal{L}_{D_{T+1}}(h) \leq 4\mathfrak{R}_m(H_L) + 2M\sqrt{\frac{\pi}{m}} + (m+1)\Delta. \tag{10}$$

Proof. Let $D = \bigotimes_{t=1}^{T+1} D_t$ and $D' = \bigotimes_{t=1}^{T} D_t$. By Fubini's theorem we can write:

$$\mathop{\mathrm{E}}_{D}[\widehat{M}_{T+1}] - \inf_h \mathcal{L}_{D_{T+1}}(h) = \mathop{\mathrm{E}}_{D'}\Big[\mathcal{L}_{D_{T+1}}(h_T) - \inf_h \mathcal{L}_{D_{T+1}}(h)\Big]. \tag{11}$$

Now, let $\phi^{-1}(\delta) = 4\mathfrak{R}_m(H_L) + (m+1)\Delta + 2M\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$, then, by (8), for $\beta > 4\mathfrak{R}_m(h) + (m+1)\Delta$, the following holds:

$$\Pr_{D'}[\mathcal{L}_{D_{T+1}}(h_T) - \inf_h \mathcal{L}_{D_{T+1}}(h) > \beta] < \phi(\beta).$$

Thus, the expectation on the right-hand side of (11) can be bounded as follows:

$$\mathop{\mathrm{E}}_{D'}\Big[\mathcal{L}_{D_{T+1}}(h_T) - \inf_h \mathcal{L}_{D_{T+1}}(h)\Big] \leq 4\mathfrak{R}_m(H_L) + (m+1)\Delta + \int_{4\mathfrak{R}_m(H_L)+(m+1)\Delta}^{\infty} \phi(\beta)d\beta.$$

The last integral can be rewritten as $2M\int_0^2 \frac{d\delta}{\sqrt{m\log\frac{2}{\delta}}} = 2M\sqrt{\frac{\pi}{m}}$ using the change of variable $\delta = \phi(\beta)$. This concludes the proof. □

The following corollary can be shown using the same proof as that of corollary 1.

Corollary 2. *Fix $\Delta > 0$. Let H be a hypothesis set with VC-dimension d such that for all $m > 1$, $4\mathfrak{R}_m(H_L) \leq C\sqrt{\frac{d}{m}}$. Let $(D_t)_{t>0}$ be a sequence of distributions over $\mathcal{X} \times \mathcal{Y}$ such that $\mathrm{disc}_{\mathcal{Y}}(D_t, D_{t+1}) \leq \Delta$. Let $C' = 2M\sqrt{\frac{\pi}{d}}$ and $K = 3\left[\frac{C+C'}{2}\right]^{2/3}$. Then, for $T > \left[\frac{C+C'}{2}\right]^{\frac{2}{3}}(\frac{d}{\Delta^2})^{\frac{1}{3}}$, the following inequality holds:*

$$\mathop{\mathrm{E}}_{D}[\widehat{M}_{T+1}] - \inf_h \mathcal{L}_{D_{T+1}}(h) < K(d\Delta)^{1/3} + \Delta.$$

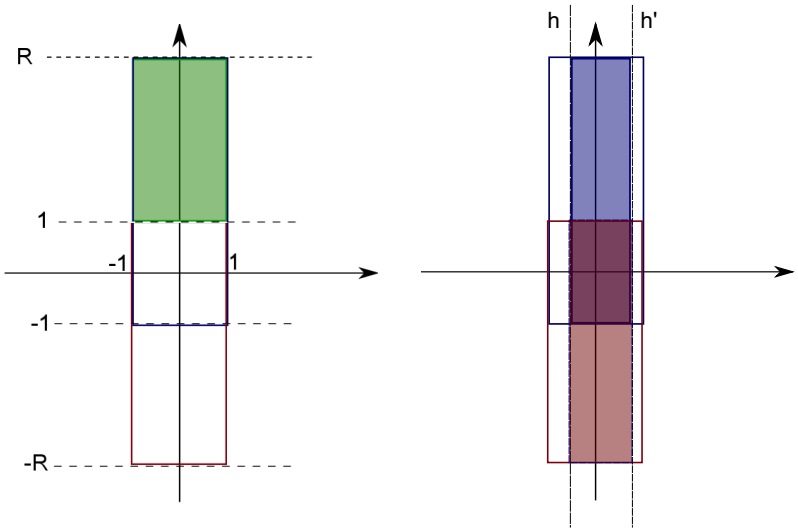

Fig. 1. Figure depicting the difference between the L_1 distance and the discrepancy. In the left figure, the L_1 distance is given by twice the area of the green rectangle. In the right figure, $P(h(x) \neq h'(x))$ is equal to the area of the blue rectangle and $Q(h(x) \neq h'(x))$ is the area of the red rectangle. The two areas are equal, thus $\mathrm{disc}(P, Q) = 0$.

In terms of definition 5, this corollary shows that algorithm $\mathcal{A}$ $(\Delta, K(d\Delta)^{1/3} + \Delta)$-tracks H. This result is similar to a result of [9] which states that given $\epsilon > 0$ if $\Delta = O(d\epsilon^3)$ then $\mathcal{A}$ (Δ, ϵ)-tracks H. However, in [9], Δ is an upper bound on the L_1 distance and not the discrepancy. Our result provides thus a tighter and more general guarantee than that of [9], the latter because this result is applicable to any loss function and not only the zero-one loss, the former because our bound is based on the Rademacher complexity instead of the VC-dimension and more importantly because it is based on the discrepancy, which is a finer measure of the divergence between distributions than the L_1 distance. Indeed, for any $t \in [1, T]$,

$$\begin{aligned}
\mathrm{disc}_{\mathcal{Y}}(D_t, D_{t+1}) &= \sup_{h \in H} \left| \mathcal{L}_{D_t}(h) - \mathcal{L}_{D_{t+1}}(h) \right| \\
&= \sup_{h \in H} \Big| \sum_{x,y} (D_t(x,y) - D_{t+1}(x,y)) L(h(x), y) \Big| \Big| \\
&\leq M \sup_{h \in H} \sum_{x,y} |D_t(x,y) - D_{t+1}(x,y)| = M L_1(D_t, D_{t+1}).
\end{aligned}$$

Furthermore, when the target function f is in H, then the $\mathcal{Y}$-discrepancies can be bounded by the discrepancies $\mathrm{disc}(D_t, D_{T+1})$, which, unlike the L_1 distance, can be accurately estimated from finite samples.

It is important to emphasize that even though our analysis was based on a particular algorithm, that of "truncated" empirical risk minimization, the bounds obtained here cannot be improved upon in the general scenario of drifting distributions, as shown by [10] in the case of binary classification.

We now illustrate the difference between the guarantees we present and those based on the L_1 distance by presenting a simple example for the zero-one loss where the L_1

distance can be made arbitrarily close to 2 while the discrepancy is 0. In that case, our bounds state that the learning problem is as favorable as in the absence of any drifting, while a learning bound with the L_1 distance would be uninformative. Consider measures P and Q in $\mathbb{R}^2$. Where P is uniform in the rectangle R_1 defined by the vertices $(-1, R)$, $(1, R)$, $(1, -1)$, $(-1, -1)$ and Q is uniform in the rectangle R_2 spanned by $(-1, -R)$, $(1, -R)$, $(-1, 1)$, $(1, 1)$. The measures are depicted in figure 1. The L_1 distance of these probability measures is given by twice the difference of measure in the green rectangle, i.e, $|P - Q| = 2\frac{(R-1)}{R+1}$ this distance goes to 2 as $R \to \infty$. On the other hand consider the zero-one loss and the hypothesis set consisting of threshold functions on the first coordinate, i.e. $h(x, y) = 1$ iff $h < x$. For any two hypotheses $h < h'$ the area of disagreement of this two hypotheses is given by the stripe $S = \{x\colon h < x < h'\}$. But it is trivial to see that $P(S) = P(S \cap R_1) = (h - h')/2$, but also $Q(S) = Q(S \cap R_2) = (h - h')/2$, since this is true for any pair of hypotheses we conclude that $\mathrm{disc}(P, Q) = 0$. This example shows that the learning bounds we presented can be dramatically more favorable than those given in the past using the L_1 distance.

Although this may be viewed as a trivial illustrative example, the discrepancy and the L_1 distance can greatly differ in more complex but realistic cases.

5 On-line to batch conversion

In this section, we present learning guarantees for drifting distributions in terms of the regret of an on-line learning algorithm $\mathcal{A}$. The algorithm processes a sample $(x_t)_{t\geq 1}$ sequentially by receiving a sample point $x_t \in \mathcal{X}$, generating a hypothesis h_t, and incurring a loss $L(h(x_t), y_t)$, with $y_t \in \mathcal{Y}$. We denote by R_T the regret of algorithm $\mathcal{A}$ after processing $T \geq 1$ sample points:

$$R_T = \sum_{t=1}^{T} L(h(x_t), y_t) - \inf_{h \in H} \sum_{t=1}^{T} L(h(x_t), y_t).$$

The standard setting of on-line learning assumes an adversarial scenario with no distributional assumption. Nevertheless, when the data is generated according to some distribution, the hypotheses returned by an on-line algorithm $\mathcal{A}$ can be combined to define a hypothesis with strong learning guarantees in the distributional setting when the regret R_T is in $O(\sqrt{T})$ (which is attainable by several regret minimization algorithms) [21, 22]. Here, we extend these results to the drifting scenario and the case of a convex combination of the hypotheses generated by the algorithm. The following lemma will be needed for the proof of our main result.

Lemma 1. *Let $\mathcal{S} = (x_t, y_t)_{t=1}^T$ be a sample drawn from the distribution $D = \bigotimes D_t$ and let $(h_t)_{t=1}^T$ be the sequence of hypotheses returned by an on-line algorithm sequentially processing $\mathcal{S}$. Let $\mathbf{w} = (w_1, \ldots, w_t)^\top$ be a vector of non-negative weights verifying $\sum_{t=1}^T w_t = 1$. If the loss function L is bounded by M then, for any $\delta > 0$,*

with probability at least $1-\delta$*, each of the following inequalities hold:*

$$\sum_{t=1}^{T} w_t \mathcal{L}_{D_{T+1}}(h_t) \leq \sum_{t=1}^{T} w_t L(h_t(x_t), y_t) + \bar{\Delta}(\mathbf{w}, T) + M\|\mathbf{w}\|_2 \sqrt{2\log\frac{1}{\delta}}$$
$$\sum_{t=1}^{T} w_t L(h_t(x_t), y_t) \leq \sum_{t=1}^{T} w_t \mathcal{L}_{D_{T+1}}(h_t) + \bar{\Delta}(\mathbf{w}, T) + M\|\mathbf{w}\|_2 \sqrt{2\log\frac{1}{\delta}},$$

where $\bar{\Delta}(\mathbf{w}, T)$ *denotes the average discrepancy* $\sum_{t=1}^{T} w_t \mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1})$.

Proof. Consider the random process: $Z_t = w_t L(h_t(x_t), y_t) - w_t \mathcal{L}(h_t)$ and let $\mathcal{F}_t$ denote the filtration associated to the sample process. We have: $|Z_t| \leq M w_t$ and

$$\mathop{\mathrm{E}}_{D}[Z_t|\mathcal{F}_{t-1}] = \mathop{\mathrm{E}}_{D}[w_t L(h_t(x_t), y_t)|\mathcal{F}_{t-1}] - \mathop{\mathrm{E}}_{D_t}[w_t L(h_t(x_t), y_t)] = 0$$

The second equality holds because h_t is determined at time $t-1$ and x_t, y_t are independent of $\mathcal{F}_{t-1}$. Thus, by Azuma-Hoeffding's inequality, for any $\delta > 0$, with probability at least $1-\delta$ the following holds:

$$\sum_{t=1}^{T} w_t \mathcal{L}_{D_t}(h_t) \leq \sum_{t=1}^{T} w_t L(h(x_t), y_t) + M\|\mathbf{w}\|_2 \sqrt{2\log\frac{1}{\delta}}. \tag{12}$$

By definition of the discrepancy, the following inequality holds for any $t \in [1, T]$:

$$\mathcal{L}_{D_{T+1}}(h_t) \leq \mathcal{L}_{D_t}(h_t) + \mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1}).$$

Summing up these inequalities and using (12) to bound $\sum_{t=1}^{T} w_t \mathcal{L}_{D_t}(h_t)$ proves the first statement. The second statement can be proven in a similar way. □

The following theorem is the main result of this section.

Theorem 2. *Assume that* L *is bounded by* M *and convex with respect to its first argument. Let* $h_1, \ldots, h_T$ *be the hypotheses returned by* $\mathcal{A}$ *when sequentially processing* $(x_t, y_t)_{t=1}^{T}$ *and let* h *be the hypothesis defined by* $h = \sum_{t=1}^{T} w_t h_t$*, where* $w_1, \ldots, w_T$ *are arbitrary non-negative weights verifying* $\sum_{t=1}^{T} w_t = 1$*. Then, for any* $\delta > 0$*, with probability at least* $1-\delta$*,* h *satisfies each of the following learning guarantees:*

$$\mathcal{L}_{D_{T+1}}(h) \leq \sum_{t=1}^{T} w_t L(h_t(x_t), y_t) + \bar{\Delta}(\mathbf{w}, T) + M\|\mathbf{w}\|_2 \sqrt{2\log\frac{1}{\delta}}$$
$$\mathcal{L}_{D_{T+1}}(h) \leq \inf_{h \in H} \mathcal{L}(h) + \frac{R_T}{T} + \bar{\Delta}(\mathbf{w}, T) + M\|\mathbf{w} - \mathbf{u}_0\|_1 + 2M\|\mathbf{w}\|_2 \sqrt{2\log\frac{2}{\delta}},$$

where $\mathbf{w} = (w_1, \ldots, w_T)^\top$*,* $\bar{\Delta}(\mathbf{w}, T) = \sum_{t=1}^{T} w_t \mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1})$*, and* $\mathbf{u}_0 \in \mathbb{R}^T$ *is the vector with all its components equal to* $1/T$.

Observe that when all weights are all equal to $\frac{1}{T}$, the result we obtain is similar to the learning guarantee obtained in theorem 1 when the Rademacher complexity of H_L is $O(\frac{1}{\sqrt{T}})$. Also, if the learning scenario is i.i.d., then the first sum of the bound vanishes and it can be seen straightforwardly that to minimize the RHS of the inequality we need to set $w_t = \frac{1}{T}$, which results in the known i.i.d. guarantees for on-line to batch conversion [21, 22].

Proof. Since L is convex with respect to its first argument, by Jensen's inequality, we have $\mathcal{L}_{D_{T+1}}(\sum_{t=1}^{T} w_t h_t) \leq \sum_{t=1}^{T} w_t \mathcal{L}_{D_{T+1}}(h_t)$. Thus, by Lemma 1, for any $\delta > 0$, the following holds with probability at least $1 - \delta$:

$$\mathcal{L}_{D_{T+1}}\left(\sum_{t=1}^{T} w_t h_t\right) \leq \sum_{t=1}^{T} w_t L(h_t(x_t), y_t) + \bar{\Delta}(\mathbf{w}, T) + M\|\mathbf{w}\|_2 \sqrt{2 \log \frac{1}{\delta}}. \quad (13)$$

This proves the first statement of the theorem. To prove the second claim, we will bound the empirical error in terms of the regret. For any $h^* \in H$, we can write using $\inf_{h \in H} \frac{1}{T} \sum_{t=1}^{T} L(h(x_t), y_t) \leq \frac{1}{T} \sum_{t=1}^{T} L(h^*(x_t), y_t)$:

$$\begin{aligned}
&\sum_{t=1}^{T} w_t L(h_t(x_t), y_t) - \sum_{t=1}^{T} w_t L(h^*(x_t), y_t) \\
&= \sum_{t=1}^{T} \Big(w_t - \frac{1}{T}\Big)[L(h_t(x_t), y_t) - L(h^*(x_t), y_t)] + \frac{1}{T} \sum_{t=1}^{T} [L(h_t(x_t), y_t) - L(h^*(x_t), y_t)] \\
&\leq M\|\mathbf{w} - \mathbf{u}_0\|_1 + \frac{1}{T} \sum_{t=1}^{T} L(h_t(x_t), y_t) - \inf_h \frac{1}{T} \sum_{t=1}^{T} L(h(x_t), y_t) \\
&\leq M\|\mathbf{w} - \mathbf{u}_0\|_1 + \frac{R_T}{T}.
\end{aligned}$$

Now, by definition of the infimum, for any $\epsilon > 0$, there exists $h^* \in H$ such that $\mathcal{L}_{D_{T+1}}(h^*) \leq \inf_{h \in H} \mathcal{L}_{D_{T+1}}(h) + \epsilon$. For that choice of h^*, in view of (13), with probability at least $1 - \delta/2$, the following holds:

$$\mathcal{L}_{D_{T+1}}(h) \leq \sum_{t=1}^{T} w_t L(h^*(x_t), y_t) + M\|\mathbf{w}_- \mathbf{u}_0\|_1 + \frac{R_T}{T} + \bar{\Delta}(\mathbf{w}, T) + M\|\mathbf{w}\|_2 \sqrt{2 \log \frac{2}{\delta}}.$$

By the second statement of Lemma 1, for any $\delta > 0$, with probability at least $1 - \delta/2$,

$$\sum_{t=1}^{T} w_t L(h^*(x_t), y_t) \leq \mathcal{L}_{D_{T+1}}(h^*) + \bar{\Delta}(\mathbf{w}, T) + M\|\mathbf{w}\|_2 \sqrt{2 \log \frac{2}{\delta}}.$$

Combining these last two inequalities, by the union bound, with probability at least $1 - \delta$, the following holds with $B(\mathbf{w}, \delta) = M\|\mathbf{w}_- \mathbf{u}_0\|_1 + \frac{R_T}{T} + 2M\|\mathbf{w}\|_2 \sqrt{2 \log \frac{2}{\delta}}$:

$$\begin{aligned}
\mathcal{L}_{D_{T+1}}(h) &\leq \mathcal{L}_{D_{T+1}}(h^*) + 2\bar{\Delta}(\mathbf{w}, T) + B(\mathbf{w}, \delta) \\
&\leq \inf_{h \in H} \mathcal{L}_{D_{T+1}}(h) + \epsilon + 2\bar{\Delta}(\mathbf{w}, T) + B(\mathbf{w}, \delta).
\end{aligned}$$

The last inequality holds for all $\epsilon > 0$, therefore also for $\epsilon = 0$ by taking the limit. □

6 Algorithm

The results of the previous section suggest a natural algorithm based on the values of the discrepancy between distributions. Let $(h_t)_{t=1}^T$ be the sequence of hypotheses generated by an on-line algorithm. Theorem 2 provides a learning guarantee for any convex combination of these hypotheses. The convex combination based on the weight vector $\mathbf{w}$ minimizing the bound of Theorem 2 benefits from the most favorable guarantee. This leads to an algorithm for determining $\mathbf{w}$ based on the following convex optimization problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \lambda \|\mathbf{w}\|_2^2 + \sum_{t=1}^{T} w_t \left(\mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1}) + L(h_t(x_t), y_t) \right) \\ \text{subject to:} \quad & \Big(\sum_{t=1}^{T} w_t = 1 \Big) \wedge (\forall t \in [1, T], w_t \geq 0), \end{aligned} \tag{14}$$

where $\lambda \geq 0$ is a regularization parameter. This is a standard QP problem that can be efficiently solved using a variety of techniques and available software.

In practice, the discrepancy values $\mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1})$ are not available since they require labeled samples. But, in the deterministic scenario where the labeling function f is in H, we have $\mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1}) \leq \mathrm{disc}(D_t, D_{T+1})$. Thus, the discrepancy values $\mathrm{disc}(D_t, D_{T+1})$ can be used instead in our learning bounds and in the optimization (14). This also holds approximately when f is not in H but is close to some $h \in H$.

As shown in [17], given two (unlabeled) samples of size n from D_t and D_{T+1}, the discrepancy $\mathrm{disc}(D_t, D_{T+1})$ can be estimated within $O(1/\sqrt{n})$, when $\mathfrak{R}_n(H_L) = O(1/\sqrt{n})$. In many realistic settings, for tasks such as spam filtering, the distribution D_t does not change within a day. This gives us the opportunity to collect an independent *unlabeled* sample of size n from each distribution D_t. If we choose $n \gg T$, by the union bound, with high probability, all of our estimated discrepancies will be within $O(1/\sqrt{T})$ of their exact counterparts $\mathrm{disc}(D_t, D_{T+1})$.

Additionally, in many cases, the distributions D_t remain unchanged over some longer periods (cycles) which may be known to us. This in fact typically holds for some tasks such as spam filtering, political sentiment analysis, some financial market prediction problems, and other problems. For example, in the absence of any major political event such as a debate, speech, or a prominent measure, we can expect the political sentiment to remain stable. In such scenarios, it should be even easier to collect an unlabeled sample from each distribution. More crucially, we do not need then to estimate the discrepancy for all $t \in [1, T]$ but only once for each cycle.

6.1 Experiments

Here, we report the results of preliminary experiments demonstrating the performance of our algorithm. We tested our algorithm on synthetic data in a regression setting. The testing and training data were created as follows: instances were sampled from a two-dimensional Gaussian random variables $\mathcal{N}(\boldsymbol{\mu}_t, 1)$. The objective function at each time was given by $y_t = \mathbf{w}_t \cdot \mathbf{x}_t$. The weight vectors $\mathbf{w}_t$ and mean vectors $\boldsymbol{\mu}_t$ were selected as follows: $\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \mathbf{U}$ and $\mathbf{w}_t = R_\theta \mathbf{w}_{t-1}$, where $\mathbf{U}$ is the uniform random variable

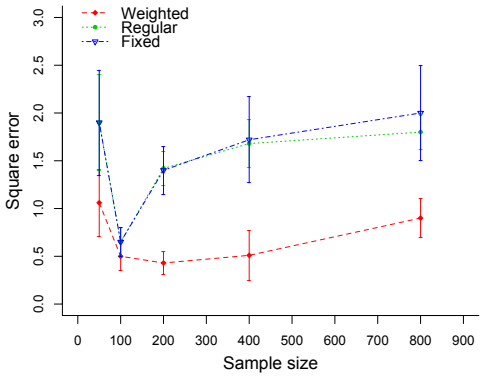

Fig. 2. Comparison of the performance of three algorithms as a function of the sample size T. `Weighted` stands for the algorithm described in this paper, `Regular` for an algorithm that averages over all the hypotheses, and `Fixed` for the algorithm that averages only over the last 100 hypotheses.

over $[-.1, +.1]^2$ and R_θ a rotation of magnitude θ distributed uniformly over $(-1, 1)$. We used the Widrow-Hoff algorithm [23] as our base on-line algorithm to determine h_t. After receiving T examples, we tested our final hypothesis on 100 points taken from the same Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_{T+1}, 1)$. We ran the experiment 50 times for different amounts of sample points and took the average performance of our classifier. For these experiments, we are considering the ideal situation where the discrepancy values are given.

We compared the performance of our algorithm with that of the algorithm that (uniformly) averages all of the hypotheses and with that of the algorithm that averages only the last 100 hypotheses generated by the perceptron algorithm. Figure 2 shows the results of our experiments in the first setting. Observe that the error increases with the sample size. While the analysis of Section 3 could provide an explanation of this phenomenon in the case of the uniform averaging algorithm, in principle, it does not explain why the error also increases in the case of our algorithm. The answer to this can be found in the setting of the experiment. Notice that the Gaussians considered are moving their center and that the squared loss grows proportional to the radius of the smallest sphere containing the sample. Thus, as the number of points increases, so does the maximum value of the loss function in the test set. Nevertheless, our algorithm still outperforms the other two algorithms. It is worth noting that the accuracy of our algorithm can drastically change of course depending on the choice of the online algorithm used.

7 Conclusion

We presented a theoretical analysis of the problem of learning with drifting distributions in the batch setting. Our learning guarantees improve upon previous ones based on the

L_1 distance, in some cases substantially, and our proofs are simpler and concise. These bounds benefit from the notion of discrepancy which seems to be the natural measure of the divergence between distributions in a drifting scenario. This work motivates a number of related studies, in particular a discrepancy-based analysis of the scenario introduced by [16] and further improvements of the algorithm we presented, in particular by exploiting the specific on-line learning algorithm used.

Acknowledgments

We thank Yishay Mansour for discussions about the topic of this paper. This work was partly funded by the NSF award IIS-1117591.

Bibliography

[1] Valiant, L.G.: A theory of the learnable. ACM Press New York, NY, USA (1984)
[2] Vapnik, V.N.: Statistical Learning Theory. J. Wiley & Sons (1998)
[3] Cesa-Bianchi, N., Lugosi, G.: Prediction, learning, and games. Cambridge University Press (2006)
[4] Herbster, M., Warmuth, M.: Tracking the best expert. Machine Learning **32** (1998) 151–78
[5] Herbster, M., Warmuth, M.: Tracking the best linear predictor. Journal of Machine Learning Research **1** (2001) 281–309
[6] Cavallanti, G., Cesa-Bianchi, N., Gentile, C.: Tracking the best hyperplane with a simple budget perceptron. Machine Learning **69** (2007) 143–167
[7] Helmbold, D.P., Long, P.M.: Tracking drifting concepts by minimizing disagreements. Machine Learning **14** (1994) 27–46
[8] Bartlett, P.L.: Learning with a slowly changing distribution. In: Proceedings of the fifth annual workshop on Computational learning theory. COLT '92, New York, NY, USA, ACM (1992) 243–252
[9] Long, P.M.: The complexity of learning according to two models of a drifting environment. Machine Learning **37** (1999) 337–354
[10] Barve, R.D., Long, P.M.: On the complexity of learning from drifting distributions. Information and Computation **138** (1997) 101–123
[11] Dudley, R.M.: A course on empirical processes. Lecture Notes in Math. **1097** (1984) 2 – 142
[12] Pollard, D.: Convergence of Stochastic Processess. Springer, New York (1984)
[13] Talagrand, M.: The Generic Chaining. Springer, New York (2005)
[14] Freund, Y., Mansour, Y.: Learning under persistent drift. In: EuroColt. (1997) 109–118
[15] Bartlett, P.L., Ben-David, S., Kulkarni, S.: Learning changing concepts by exploiting the structure of change. Machine Learning **41** (2000) 153–174
[16] Crammer, K., Even-Dar, E., Mansour, Y., Vaughan, J.W.: Regret minimization with concept drift. In: COLT. (2010) 168–180
[17] Mansour, Y., Mohri, M., Rostamizadeh, A.: Domain adaptation: Learning bounds and algorithms. In: Proceedings of COLT, Montréal, Canada, Omnipress (2009)
[18] Valiant, P.: Testing symmetric properties of distributions. SIAM J. Comput. **40** (2011) 1927–1968
[19] Rakhlin, A., Sridharan, K., Tewari, A.: Online learning: Random averages, combinatorial parameters, and learnability (2010)

[20] Cortes, C., Mohri, M.: Domain adaptation in regression. In: Proceedings of ALT 2011. (2011) 308–323
[21] Littlestone, N.: From on-line to batch learning. In: Proceedings of the second annual workshop on Computational learning theory, Morgan Kaufmann Publishers Inc. (1989) 269–284
[22] Cesa-Bianchi, N., Conconi, A., Gentile, C.: On the generalization ability of on-line learning algorithms. In: NIPS. (2001) 359–366
[23] Widrow, B., Hoff, M.E.: Adaptive switching circuits. Neurocomputing: Foundations of Research (1988)