Supplemental Material for: Towards Non-Parametric Drift Detection via Dynamic Adapting Window Independence Drift Detection (DAWIDD)

A. Proofs

In this section we provide proofs for the theorems and lemmas given in our paper. We also include all definitions, section headlines and the statements that are to be proven. The numbering is the same as in the paper.

A.1. Concept Drift Definition

Definition 1. A *drift process* (p_t, P_T) is a probability measure P_T on $[0,1]$ together with a collection of probability measures p_t on $\mathbb{R}^d$ with $t \in [0,1]$, such that $t \mapsto p_t(A)$ is measurable for every measurable $A \subset \mathbb{R}^d$.

Definition 2. Let (p_t, P_T) be a drift process. We say that p_t has no *drift* if $p_t \neq p_s$ holds on a P_T null set only, i.e. $(P_T \times P_T)(\{(s,t) \in [0,1]^2 \mid p_t \neq p_s\}) = 0$.

Lemma 1. *Let (p_t, P_T) be a drift process. The following are equivalent:*

1. *p_t has no drift*
2. *there exists a probability measure P_X such that $p_t = P_X$ for P_T-a.s. all $t \in [0,1]$*
3. *there exists a probability measure P_X such that $p_t \otimes P_T = P_X \times P_T$*

Furthermore, if P_X exists, it is uniquely determined and it holds $P_X = \int p_t P_T(\mathrm{d}t)$.

Proof. Show ”1. $\Leftarrow$ 2.”: Denote by $C = \{t | p_t = P_X\}$ and by $D = \{(t,s) | p_t = p_s\}$. Obviously it holds $C \times C \subset D$. Since $P_T(C) = 1$ we have

$$1 = (P_T \times P_T)(C \times C) \leq (P_T \times P_T)(D).$$

Show ”1. $\Rightarrow$ 2.”: Since P_T is finite we may write $(P_T \times P_T)(A) = \int P_T(A^x) P_T(\mathrm{d}x)$, where $A^x = \{y | (x,y) \in A\}$. This implies that $P_T(\{s \in T | p_s = p_t\}) = 1$ for P_T-a.s. all $t \in [0,1]$. Therefore there exists a $t_0 \in [0,1]$ such that $p_t = p_{t_0}$ for P_T-a.s. all $t \in [0,1]$; so we may choose $P_X = p_{t_0}$.

Show ”2. $\iff$ 3.”: Since $\mathfrak{B}(\mathbb{R}^d)$ has a intersection stable countable generator the statement follows by the fact that in this case it holds $p_t \otimes P = q_t \otimes Q \Leftrightarrow P = Q$ and $p_t = q_t$ P-a.s. for probability measures P, Q and Markov kernels p_t, q_t.

Show uniqueness: For all $A \in \mathfrak{B}(\mathbb{R}^d)$ we have

$$\begin{aligned} P'_X(A) &= (P'_X \times P_T)(A \times [0,1]) \\ &= (p_t \otimes P_T)(A \times [0,1]) \\ &= (P_X \times P_T)(A \times [0,1]) = P_X(A). \end{aligned}$$

Show representation: Choose $t_0 \in [0,1]$ as before. Then it holds $\int p_t \mathrm{d}P_T = \int p_{t_0} \mathrm{d}P_T = p_{t_0} \int 1 \mathrm{d}P_T = P_X$. □

A.2. Change of Loss as Indicator for Drift

Definition 3. Let $\mathcal{H}$ be a hypothesis class and $\mathfrak{X}$ be a measure space (usually $\mathfrak{X} = \mathbb{R}^d$). An *empirical loss function* is a map $\hat{\ell} : \mathcal{H} \times (\coprod_{n=0}^{\infty} \prod_{i=1}^{n} \mathfrak{X}) \to \mathbb{R}$, such that for every set of $\mathfrak{X}$-valued random variables $X_1, ..., X_n$ and hypothesis $h \in \mathcal{H}$ we obtain a measurable map $\hat{\ell}(h | X_1, ..., X_n) : \Omega \to \mathbb{R}$ which measures the error of h on the the random samples delivered by $X_1, ..., X_n$.

We say that an empirical loss function $\hat{\ell}$ *decomposes into sums* for $X_1, X_2, ..., X_N$ (with $N \in \mathbb{N} \cup \{\infty\}$) if $\hat{\ell}(h | X_1, ..., X_n) = \frac{1}{n} \sum_{i=1}^{n} \hat{\ell}(h | X_i)$ holds for all $n \leq N$.

We say that an empirical loss function is *uniformly bounded* if there exists an $K < \infty$ such that $|\hat{\ell}(h | x_1, ..., x_n)| < K$ for all $x_1, ..., x_n \in \mathfrak{X}$ and $h \in \mathcal{H}$.

Theorem 1. *Let $\hat{\ell}$ be an empirical loss function on a hypothesis class $\mathcal{H}$ which is uniformly bounded by some $K < \infty$. Let $X_1, ..., X_n$ be independent and $X'_1, ..., X'_m$ be independent random variables for which $\hat{\ell}$ decomposes into sums. Then for all $h \in \mathcal{H}$ and $\varepsilon > 0$ it holds*

$$\begin{aligned} &\mathbb{P}[|\hat{\ell}(h | X_1, ..., X_n) - \hat{\ell}(h | X'_1, ..., X'_m)| \geq \varepsilon] \\ &\leq \frac{K}{\varepsilon} \sqrt{\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)^2 + \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_{X_i} - \frac{1}{m} \sum_{i=1}^{m} \mathbb{P}_{X'_i} \right\|_{\text{TV}}^2}, \end{aligned}$$

where $\|\cdot\|_{\text{TV}}$ denotes the total variation norm.

Proof. Let Z be a real-valued random variable, then by using that $\text{var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$, $x \leq \sqrt{x}$ for $x \in [0,1]$

and Markov's inequality it follows that for all $\varepsilon > 0$ it holds

$$\mathbb{P}[|Z| \geq \varepsilon] = \mathbb{P}[Z^2 \geq \varepsilon^2] \leq \frac{1}{\varepsilon}\sqrt{\mathrm{var}(Z) + \mathbb{E}[Z]^2}. \qquad (1)$$

Now consider the case where $Z = \hat{\ell}(h|X_1, ..., X_n) - \hat{\ell}(h|X'_1, ..., X'_m)$; we have to show that a) $\mathrm{var}(\hat{\ell}(h|X_1, ..., X_n) - \hat{\ell}(h|X'_1, ..., X'_m)) \leq K^2(n^{-1/2} + m^{-1/2})^2$ and b) $\mathbb{E}[\hat{\ell}(h|X_1, ..., X_n) - \hat{\ell}(h|X'_1, ..., X'_m)] \leq K \left\| \frac{1}{n}\sum_{i=1}^n \mathbb{P}_{X_i} - \frac{1}{m}\sum_{i=1}^m \mathbb{P}_{X'_i} \right\|_{\mathrm{TV}}$.

Start with a): First note that

$$\begin{aligned} \mathrm{var}(Z - Z') &\leq \mathrm{var}(Z) + \mathrm{var}(Z') + 2|\mathrm{cov}(Z, Z')| \\ &\leq (\sqrt{\mathrm{var}(Z)} + \sqrt{\mathrm{var}(Z')})^2, \end{aligned} \qquad (2)$$

where we used that $\mathrm{cov}(Z, Z')^2 \leq \mathrm{var}(Z)\mathrm{var}(Z')$. Since $\hat{\ell}$ decomposes into sums, with each summat being independent and bounded by K we have

$$\mathrm{var}(\hat{\ell}(h|X_1, ..., X_n)) = \frac{1}{n^2}\sum_{i=1}^n \underbrace{\mathrm{var}(X_i)}_{\leq K^2} \leq K^2 \frac{1}{n} \qquad (3)$$

plugging this in we obtain the stated bound for a).

Now prove b): Recall that $\mathbb{E}[f(Z)] = \int f \mathrm{d}\mathbb{P}_Z$. Now by using that the integral is a bilinear map and the fact that $\hat{\ell}$ decomposes into sums it follows that

$$\begin{aligned} &\mathbb{E}[\hat{\ell}(h|X_1, ..., X_n) - \hat{\ell}(h|X'_1, ..., X'_m)] \\ &= \int \hat{\ell}(h|x)\mathrm{d}\left(\frac{1}{n}\sum_i \mathbb{P}_{X_i}\right) \\ &\quad - \int \hat{\ell}(h|x)\mathrm{d}\left(\frac{1}{m}\sum_i \mathbb{P}_{X'_i}\right) \\ &= K\left(\int K^{-1}\hat{\ell}(h|x)\mathrm{d}\left(\frac{1}{n}\sum_i \mathbb{P}_{X_i}\right)\right. \\ &\quad \left. - \int K^{-1}\hat{\ell}(h|x)\mathrm{d}\left(\frac{1}{m}\sum_i \mathbb{P}_{X'_i}\right)\right) \\ &\leq K \sup_{\substack{f:\mathbb{R}\to[-1,1] \\ \text{measurable}}} \left(\int f(x)\mathrm{d}\left(\frac{1}{n}\sum_i \mathbb{P}_{X_i}\right)\right. \\ &\quad \left. - \int f(x)\mathrm{d}\left(\frac{1}{m}\sum_i \mathbb{P}_{X'_i}\right)\right) \\ &= K\left\|\frac{1}{n}\sum_i \mathbb{P}_{X_i} - \frac{1}{m}\sum_i \mathbb{P}_{X'_i}\right\|_{\mathrm{TV}}. \end{aligned} \qquad (4)$$

Plugging (2)-(4) into (1) the statement follows. □

Lemma 2. *Let $\hat{\ell}$ be an empirical loss function and $X_1, ..., X_n$ be random variable for which $\hat{\ell}$ decomposes into sums. Denote by $F_{\hat{\ell}(h|X_1,...,X_n)}(x)$ the empirical cumulative distribution over $\hat{\ell}(h|X_1, ..., X_n)$, i.e.*

$$F_{\hat{\ell}(h|X_1,...,X_n)}(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{(\hat{\ell}(h|X_i),\infty)}(x).$$

Then for every $x \in \mathbb{R}$ we have that $F_{\hat{\ell}(h|X_1,...,X_n)}(x)$ is again an empirical loss function that decomposes into sums with $K = 1$.

Proof. Obvious □

Corollary 1. *Let (p_t, P_T) be a drift process and $\hat{\ell}$ be an empirical loss function on a hypothesis class $\mathcal{H}$ which is uniformly bounded by some $K < \infty$. Let $(X_1, T_1), ..., (X_n, T_n) \sim p_t \otimes P_T$ and $(X'_1, T'_1), ..., (X'_n, T'_n) \sim p_t \otimes P_T$ be independent random variables. Then for all $h \in \mathcal{H}$, $A, B \subset [0, 1]$ measurable with $P_T(A), P_T(B) > 0$ and $\varepsilon > 0$ it holds*

$$\begin{aligned} &\mathbb{P}[|\hat{\ell}(h|\mathbf{X}) - \hat{\ell}(h|\mathbf{X}')| \geq \varepsilon | \mathbf{T} \in A, \mathbf{T}' \in B] \\ &\leq \frac{K}{\varepsilon}\sqrt{(n^{-1/2} + m^{-1/2})^2 + \|p_A - p_B\|_{\mathrm{TV}}^2}, \end{aligned}$$

where $p_A = P_T(A)^{-1}\int_A p_t(\cdot)P_T(\mathrm{d}t)$ and p_B analogous and we used the short hands $\hat{\ell}(h|\mathbf{X}) = \hat{\ell}(h|X_1, ..., X_n)$ and $\mathbf{T} \in A :\Longleftrightarrow T_1 \in A, ..., T_n \in A$ and analogous for $\mathbf{X}'$ and $\mathbf{T}'$.

Proof. The conditional version of Markov's inequality states that $\mathbb{P}[X \geq Y|\mathcal{F}] \leq Y^{-1}\mathbb{E}[X|\mathcal{F}]$ for a $\mathcal{F}$ adapted, positive random variable Y. Now by redoing the proof of theorem 1 using that $X_i|T_i \in A \sim P_T(A)^{-1}\int_A p_t \mathrm{d}P_T(t) = p_A$ for $i = 1, ..., n$ and analogous for $X'_i|T'_i \in B$ with $i = 1, ..., m$ the statement follows. □

Lemma 3. *Let p_t be a drift process. Then we may find a model h and an empirical loss function $\hat{\ell}$ such that*

$$|\hat{\ell}(h|X_1, ..., X_n) - \hat{\ell}(h|X'_1, ..., X'_n)| \xrightarrow{a.s.} \|p_A - p_B\|_{\mathrm{TV}},$$

with $X_1, X_2, ... \sim p_A$, $X'_1, X'_2, ... \sim p_B$ independent.

Proof. Choose $\mathcal{H} = \mathfrak{B}(\mathbb{R}^d)$ and $\hat{\ell}(h|x) = 2\mathbb{I}_h(x)$, $\hat{\ell}(h|x_1, ..., x_n) = \frac{1}{n}\sum_{i=1}^n \hat{\ell}(h|x_i)$. Now by the law of large numbers we have

$$\begin{aligned} &|\hat{\ell}(h|X_1, ..., X_n) - \hat{\ell}(h|X'_1, ..., X'_n)| \\ &= \quad |\hat{\ell}(h|X_1, ..., X_n) - \mathbb{E}[\hat{\ell}(h|X_1)]| \\ &\quad + |\mathbb{E}[\hat{\ell}(h|X_1)] - \mathbb{E}[\hat{\ell}(h|X'_1)]| \\ &\quad + |\mathbb{E}[\hat{\ell}(h|X'_1)] - \hat{\ell}(h|X'_1, ..., X'_n)| \\ &\xrightarrow{\text{a.s.}} |\mathbb{E}[\hat{\ell}(h|X_1)] - \mathbb{E}[\hat{\ell}(h|X'_1)]| \\ &= 2|p_A(h) - p_B(h)|. \end{aligned}$$

Using that

$$\|p_A - p_B\|_{\mathrm{TV}} = 2 \sup_{h\in\mathfrak{B}(\mathbb{R}^d)} |p_A(h) - p_B(h)|$$

and that such h may actually be found (Hahn-Banach) the statement follows. □

Definition 4. We say that a drift process (p_t, P_T) has *model drift* iff there exists measurable sets $A, B \subset [0,1]$ with $P_T(A), P_T(B) > 0$, such that $p_A \neq p_B$ or equivalent $\|p_A - p_B\|_{\mathrm{TV}} > 0$, with $p_A = P_T(A)^{-1}\int_A p_t(\cdot)P_T(\mathrm{d}t)$ and analogous for p_B.

Theorem 2. *Let (p_t, P_T) be a drift process. Then it holds that p_t has drift if and only if p_t has model drift.*

Proof. Show ”$\Leftarrow$”: Let $A, B \subset [0,1]$ measurable, $P_T(A), P_T(B) > 0$ with $p_A \neq p_B$. Assume that p_t has no drift. By lemma 1 we have $p_t \otimes P_T = P_X \times P_T$ and hence $p_A = P_X = p_B$ which is a contradiction.

Show ”$\Rightarrow$”: Assume p_t has no model drift. Then for all $A, B \subset [0,1]$ measurable, $P_T(A), P_T(B) > 0$ it holds $p_A = p_B$ so $P_X := p_A$ is well defined. Now it holds

$$\begin{aligned}(p_t \otimes P_T)(B\times C) &\overset{\text{def. } p_B}{=} \begin{cases} P_T(B)p_B(C) & \text{if } P_T(B) > 0 \\ 0 & \text{if } P_T(B) = 0 \end{cases}\\ &= P_T(B)P_X(C)\\ &= (P_T \times P_X)(B \times C)\end{aligned}$$

for all $B \in \mathfrak{B}([0,1]), C \in \mathfrak{B}(\mathbb{R}^d)$, which is a intersection stable generator of $\mathfrak{B}(\mathbb{R}^d \times [0,1])$. So we have $p_t \otimes P_T = P_X \times P_T$ which by lemma 1 implies that p_t has no drift. This is a contradiction. □

A.3. Drift as Dependency between Data and Time

Definition 5. Let (p_t, P_T) be a drift process and let $(X, T) \sim p_t \otimes P_T$ a pair of random variables. We say that p_t has *dependency drift* if X and T are statistically dependent, i.e. are not independent random variables.

Theorem 3. *Let (p_t, P_T) be a drift process. Then p_t has drift if and only if it has dependency drift.*

Proof. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space, i.e. X and T are measurable maps from Ω to $\mathbb{R}^d$ resp. $[0,1]$. X and T are independent if and only if

$$(p_t \otimes P_T)(A \times B) = \mathbb{P}_{X,T}(A \times B) = \mathbb{P}_X(A)\mathbb{P}_T(B)$$

holds for all $A \in \mathfrak{B}(\mathbb{R}^d), B \in \mathfrak{B}([0,1])$. By setting $A = \mathbb{R}^d$ we obtain $\mathbb{P}_T = P_T$ and therefore $p_t \otimes P_T = \mathbb{P}_X \times P_T$ which, by lemma 1, holds if and only if p_t has no drift. □

A.4. Drift Detection via Independence Tests on Dynamically Adapted Windows

Lemma 4. *Let (p_t, P_T) and (q_t, Q_T) be drift processes. Suppose $P_T(A) = 0 \Rightarrow Q_T(A) = 0$ for all measurable $A \in \mathfrak{B}([0,1])$ and that $p_t = q_t$ for P_T-a.s. all $t \in [0,1]$. Then it holds: if p_t has no drift then q_t has no drift.*

Proof. Denote by $f = \frac{\mathrm{d}Q_T}{\mathrm{d}P_T}$ the Radon-Nikodym density. Then it holds

$$\begin{aligned}(q_t \otimes Q_T)(A \times B) &= \int_A q_t(B)\mathrm{d}Q_T(t)\\ &= \int_A q_t(B)f(t)\mathrm{d}P_T(t)\\ &\overset{\text{Cauchy-Schwarz}}{=} \int_A p_t(B)f(t)\mathrm{d}P_T(t)\\ &\overset{\text{lemma 1}}{=} P_X(B)\int_A f(t)\mathrm{d}P_T(t)\\ &= P_X(B)Q_T(A)\end{aligned}$$

for all $A \in \mathfrak{B}([0,1]), B \in \mathfrak{B}(\mathbb{R}^d)$, which is a intersection stable generator of $\mathfrak{B}(\mathbb{R}^d \times [0,1])$. So we have $q_t \otimes Q_T = P_X \times Q_T$ which by lemma 1 implies that q_t has no drift. □