# A. Theoretical Analysis and Deferred Proofs

## A.1. Proofs in Sec. 2

**Lemma A.1** (Equivalence condition). ~~*If we assume (1) identical ground truth labeling function in the training and deployment $g = g'$, (2) Restricted TV distance between training and deployment $TV(\boldsymbol{P}_x, \boldsymbol{Q}_x) \leq \kappa$, then with probability $1 - 2\epsilon - \kappa$, PDD is equivalent to D-PDD.*~~

*Proof.* **~~Step 1:~~** ~~Def 1 $\rightarrow$ Def 2. If $g' = g$, it is clear~~

$$\cancel{\mathrm{err}(f, \boldsymbol{Q}_g) > \mathrm{err}(f, \boldsymbol{P}_g)}$$

~~Then we set $h = g \in \mathcal{H}$, we have Def 2.~~

**~~Step 2:~~** ~~Def 2 $\rightarrow$ Def 1. If $g' = g$, we need to prove~~

$$\cancel{\mathrm{err}(f, \boldsymbol{Q}_g) > \mathrm{err}(f, \boldsymbol{P}_g)}$$

~~Given the disagreement condition in Def 2,~~

$$\cancel{\mathrm{err}(f, \boldsymbol{Q}_h) > \mathrm{err}(f, \boldsymbol{P}_h)}$$

~~We need to demonstrate $h = g$ with high probability, given the binary risk definition and Markov inequality, we have high probability $1 - \epsilon$ such that:~~

$$\cancel{h = g, \quad f = g}$$

~~Therefore $\mathrm{err}(f, \boldsymbol{P}_g) = \mathrm{err}(f, \boldsymbol{P}_h)$ happens in probability $\boldsymbol{P}$ with $1 - \epsilon$. Then we consider this events in $\boldsymbol{Q}$, given a small TV-distance ($\kappa$) between $\boldsymbol{P}$ and $\boldsymbol{Q}$, we have~~

$$\cancel{|\boldsymbol{P}(f(x) = g(x)) - \boldsymbol{Q}(f(x) = g(x))| \leq \mathrm{TV}(\boldsymbol{P}_x, \boldsymbol{Q}_x) \leq \kappa}$$

~~Thus with high probability $1 - \kappa - \epsilon$ in $\boldsymbol{Q}$, we still have $f(x) = g(x)$. Union bounding yields the desired conclusion.~~ □

**Lemma A.2** (Equivalence condition). *Assume that (1) the ground truth distribution at training and deployment are identical, i.e. $g = g'$, and that $\mathrm{TV}(\boldsymbol{P}, \boldsymbol{Q}) \leq \kappa$, we have that when $\mathrm{err}(f, \boldsymbol{Q}_h) - \mathrm{err}(f, \boldsymbol{P}_h) \geq 2(\kappa + \epsilon)$, i.e. the disagreement gap is large enough, D-PDD and PDD are equivalent.*

*Proof.* To show PDD $\implies$ D-PDD, assume $g = g'$, i.e. identical concepts during training and deployment. Assume our base classifier $f$ is well-trained with $\mathrm{err}(f, \boldsymbol{P}_g) = \epsilon$. We have that

$$\mathrm{err}(f, \boldsymbol{Q}_g) > \mathrm{err}(f, \boldsymbol{P}_g)$$

Let our candidate auxiliary function $h \in \mathcal{H}$ be given by $h = g$. Then, $h$ satisfies all conditions for D-PDD.

We now show that D-PDD $\implies$ PDD. Assume that there is no concept shift, i.e. the ground truth distribution is identical, $g = g'$.

We transport the D-PDD condition $\exists h \in \mathcal{H}$ s.t. $\mathrm{err}(f, \boldsymbol{Q}_h) > \mathrm{err}(f, \boldsymbol{P}_h)$ to the general PDD condition $\mathrm{err}(f, \boldsymbol{Q}_g) > \mathrm{err}(f, \boldsymbol{P}_g)$ by leveraging the proximity of $h$ to $g$ on $\boldsymbol{P}$ and that the total variation between $P$ and $Q$ are constrained by $\kappa$.

We observe that for any $f, g, h \in \mathcal{H}$:

$$|\mathrm{err}(f, \boldsymbol{P}_g) - \mathrm{err}(f, \boldsymbol{P}_h)| < \epsilon$$

Indeed, this is true since:

$$\begin{aligned}
|\mathrm{err}(f, \boldsymbol{P}_g) - \mathrm{err}(f, \boldsymbol{P}_h)| &= |\boldsymbol{P}(f \neq g) - \boldsymbol{P}(f \neq h)| \\
&= |\mathbb{E}_{\boldsymbol{P}}[\mathbb{1}\{f \neq g\} - \mathbb{1}\{f \neq h\}]| \\
&\leq \mathbb{E}_{\boldsymbol{P}}[|\mathbb{1}\{f \neq g\} - \mathbb{1}\{f \neq h\}|] \\
&\leq \mathbb{E}_{\boldsymbol{P}}[|\mathbb{1}\{g \neq h\}|] \\
&= \boldsymbol{P}(g \neq h) \leq \epsilon
\end{aligned}$$

where we used Jensen's inequality, and that $|\mathbb{1}\{f \neq g\} + \mathbb{1}\{f \neq h\}| = |\mathbb{1}\{g \neq h\}|$.

Let $\mathrm{TV}(\boldsymbol{P}, \boldsymbol{Q}) \leq \kappa$ for some $\kappa > 0$. We further observe that for any $f, g \in \mathcal{H}$:

$$
\begin{aligned}
|\mathrm{err}(f, \boldsymbol{Q}_g) - \mathrm{err}(f, \boldsymbol{P}_g)| &= |\boldsymbol{Q}(f \neq g) - \boldsymbol{P}(f \neq g)| \\
&\leq \sup_A |\boldsymbol{Q}(A) - \boldsymbol{P}(A)| = \kappa
\end{aligned}
$$

Putting our two observations together yields following decomposition:

$$
\begin{aligned}
|\mathrm{err}(f, \boldsymbol{Q}_h) - \mathrm{err}(f, \boldsymbol{Q}_g)| \leq &\ |\mathrm{err}(f, \boldsymbol{Q}_h) - \mathrm{err}(f, \boldsymbol{P}_h)| \\
&+ |\mathrm{err}(f, \boldsymbol{P}_h) - \mathrm{err}(f, \boldsymbol{P}_g)| + |\mathrm{err}(f, \boldsymbol{P}_g) - \mathrm{err}(f, \boldsymbol{Q}_g)| \\
\leq &\ 2\kappa + \epsilon
\end{aligned}
$$

For PDD to hold, $\mathrm{err}(f, \boldsymbol{Q}_g)$ needs to be no less than $\mathrm{err}(f, \boldsymbol{P}_h) + \epsilon$ and at most $2\kappa + \epsilon$ less than $\mathrm{err}(f, \boldsymbol{Q}_h)$. Equating yields:

$$
\begin{aligned}
&\mathrm{err}(f, \boldsymbol{P}_h) + \epsilon \leq \mathrm{err}(f, \boldsymbol{Q}_h) - 2\kappa - \epsilon \\
\implies &\ \mathrm{err}(f, \boldsymbol{Q}_h) - \mathrm{err}(f, \boldsymbol{P}_h) \geq 2(\kappa + \epsilon)
\end{aligned}
$$

prescribing the conditions for which D-PDD implies PDD.

□

### A.2. Proofs in Sec. 4

### A.3. Preliminary quantities

**Definition A.3** (Deployed classifier error)**.** This quantifies the generalization error of the deployed base classifier $f$. This is measured on the distribution seen during training $\boldsymbol{P}_g$,

$$
\epsilon_f := \mathrm{err}(f; \boldsymbol{P}_g) \tag{8}
$$

Indeed in Def. 2.2, we want the population error to be at most $\epsilon_f$, which results in the constraint for the empirical error in the optimization problems of Algorithm 1 at most $\epsilon = \epsilon_f - \epsilon_0$, where $\epsilon_0$ is a hyper-parameter to measure the gap between the empirical and population error.

We also re-define here for convenience the VC dimensions of the hypothesis space $\mathcal{H}$ and the subset of interest $\mathcal{H}_p$ as:

$$
\mathcal{H}_p := \{h \in \mathcal{H} : \mathrm{err}(h; \boldsymbol{P}_g) \leq \epsilon_f\}, \quad d_p := \mathrm{VC}(\mathcal{H}_p), \quad d := \mathrm{VC}(\mathcal{H})
$$

Note that $d_p \leq d$. If the base classifier $f$ is well-trained ($\epsilon_f$ is low), then $d_p$ can be much smaller than $d$ i.e., $d_p \ll d$.

**Definition A.4** ($\epsilon_p, \epsilon_q$ maximum error in $\mathcal{H}_p$)**.** We define the maximum error in $\mathcal{H}_p$ for both $\boldsymbol{P}$ and $\boldsymbol{Q}$ using pseudo-labels from $f$,

$$
\epsilon_p = \max_{h \in \mathcal{H}_p} \mathrm{err}(h; \boldsymbol{P}_f), \quad \epsilon_q = \max_{h \in \mathcal{H}_p} \mathrm{err}(h; \boldsymbol{Q}_f) \tag{9}
$$

Note that empirical quantities of these are also the maximum empirical disagreement rates used in Algo. 1 and Algo. 2. Effectively, the algorithm detects $\epsilon_q - \epsilon_p > 0$ with finite samples.

**Definition A.5** ($\xi$ quantifies D-PDD)**.** We define $\xi$ to quantify the degree of D-PDD. We adopt Def. 2.2 and define $\xi$ as

$$
\xi := \max_{h \in \mathcal{H}_p} \{\mathrm{err}(h; \boldsymbol{Q}_f) - \mathrm{err}(h; \boldsymbol{P}_f)\} \tag{10}
$$

Therefore, D-PDDM detects whether $\xi > 0$. Furthermore, $\xi$ is non-negative since $f \in \mathcal{H}_p$. Hence, in case of non-deteriorating shift, $\xi = 0$.