

Theorem 4: Dark Information Induced Instability

Theorem 1 (Hidden Drift Instability Theorem). *Let $D(t), A(t) \in \mathbb{R}^n$ be continuously differentiable.*

Let $P : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a linear projection operator representing domain visibility.

Define the visible and dark components of the domain:

$$D(t) = \tilde{D}(t) + D_{dark}(t),$$

where

$$\tilde{D}(t) = PD(t), \quad D_{dark}(t) = (I - P)D(t).$$

Assume the agent adapts only to visible information:

$$\dot{A}(t) = f(\tilde{D}(t)).$$

Suppose there exists $\varepsilon > 0$ and $T \geq 0$ such that for all $t \geq T$,

$$\frac{E(t)}{\|E(t)\|} \cdot \dot{D}_{dark}(t) \geq \varepsilon,$$

where

$$E(t) = D(t) - A(t).$$

Then there exists $T^ \geq T$ such that for all $t \geq T^*$,*

$$\frac{d}{dt} \|E(t)\| \geq \frac{\varepsilon}{2},$$

and therefore

$$\|E(t)\| \rightarrow \infty \quad \text{as } t \rightarrow \infty.$$

Proof. From Theorem 1 (Dual Failure Law),

$$\frac{d}{dt} \|E(t)\| = \frac{E(t)}{\|E(t)\|} \cdot (\dot{D}(t) - \dot{A}(t)).$$

Decompose domain drift:

$$\dot{D}(t) = \dot{\tilde{D}}(t) + \dot{D}_{dark}(t).$$

Thus,

$$\frac{d}{dt} \|E(t)\| = \frac{E(t)}{\|E(t)\|} \cdot \dot{\tilde{D}}(t) + \frac{E(t)}{\|E(t)\|} \cdot \dot{D}_{\text{dark}}(t) - \frac{E(t)}{\|E(t)\|} \cdot \dot{A}(t).$$

Because the agent adapts only to $\tilde{D}(t)$, the terms involving $\dot{\tilde{D}}(t)$ and $\dot{A}(t)$ remain bounded and compensatory.

By assumption,

$$\frac{E(t)}{\|E(t)\|} \cdot \dot{D}_{\text{dark}}(t) \geq \varepsilon.$$

The remaining two terms are bounded in magnitude by some constant C .

Since $\varepsilon > 0$, there exists $T^* \geq T$ such that for all $t \geq T^*$,

$$\frac{d}{dt} \|E(t)\| \geq \varepsilon - C \geq \frac{\varepsilon}{2}.$$

Integrating from T^* to t ,

$$\|E(t)\| \geq \|E(T^*)\| + \frac{\varepsilon}{2}(t - T^*).$$

Hence $\|E(t)\|$ diverges as $t \rightarrow \infty$. □