Distribution Shift Taxonomy

Level 1: Surface Variations

Same computation, different appearance

- Image noise/blur
- TTA: Style transfer Sensor variations

Level 2: Statistical Shifts

Same process, different statistics

- Object frequencies
- Demographic shifts Marginal Seasonal Variations

Level 3: Mechanism Changes

New computation required

• g → g(t)

TTA: X Degrades (238%)

Physics System Examples

Two-Ball System

Training: g = 9.8Test: $g(t) = 9.8 + 2\sin(0.1t)$

Pendulum System

Training: $L = L_0$ Test: $L(t) = L_0(1 + 0.2\sin(0.1t))$

Key Insight

Test-Time Adaptation (TTA) methods optimize self-supervised objectives that align with true error for Level 1-2 shifts but become misaligned Gradient Alignatenesshifts, causing systematic degradate formance Impact:

Level 1-2: $cos(\theta) > 0$

Level 3: $cos(\theta) < 0$

Two-ball: 238% worse

Pendulum: 12-18x worse