



Deciding When a Safety Assessment Needs Review



Figure 1. Simulation is too slow and a black-box predictor unreliable off-distribution, so L2T turns the question into a statistical check on new contexts alone — **focusing expert attention where the assessment needs review**.

Problem Setup

Constrained stochastic dynamics

$$\frac{d}{dt} s(t) = f_{\theta}(s(t), a(t); X), \quad 0 = g_{\theta}(s(t), a(t); X),$$

with unknown θ , algebraic variable $a(\cdot)$, and random context $X \in \mathcal{X}$ inducing the trajectory $s_{\theta}(\cdot; x)$.

Stability as a chance constraint

For a stability functional η (e.g. $\eta(s) = \max_{t \leq T} |s(t)|$), the system is (τ, α) -stable under P if $\mathbb{P}_{X \sim P}[\eta(s_{\theta}(\cdot; X)) \leq \tau] \geq 1 - \alpha$.

The **safe regime** $\mathcal{P}_0 = \{P : (\tau, \alpha)\text{-stable under } P\}$ is a set of *distributions*, and is unknown.

What is observed — and what is asked

Baseline: $\{(X_i, \mathcal{S}_i)\}_{i=1}^n$, $X_i \sim P_0 \in \mathcal{P}_0$ — contexts *and* trajectories, from one certified domain. Deployment: $\{X_j\} \sim P_1$, **contexts only**.

$$H_0 : P_1 \in \mathcal{P}_0 \quad \text{vs.} \quad H_1 : P_1 \notin \mathcal{P}_0.$$

Stability Assessment Becomes Statistical Testing

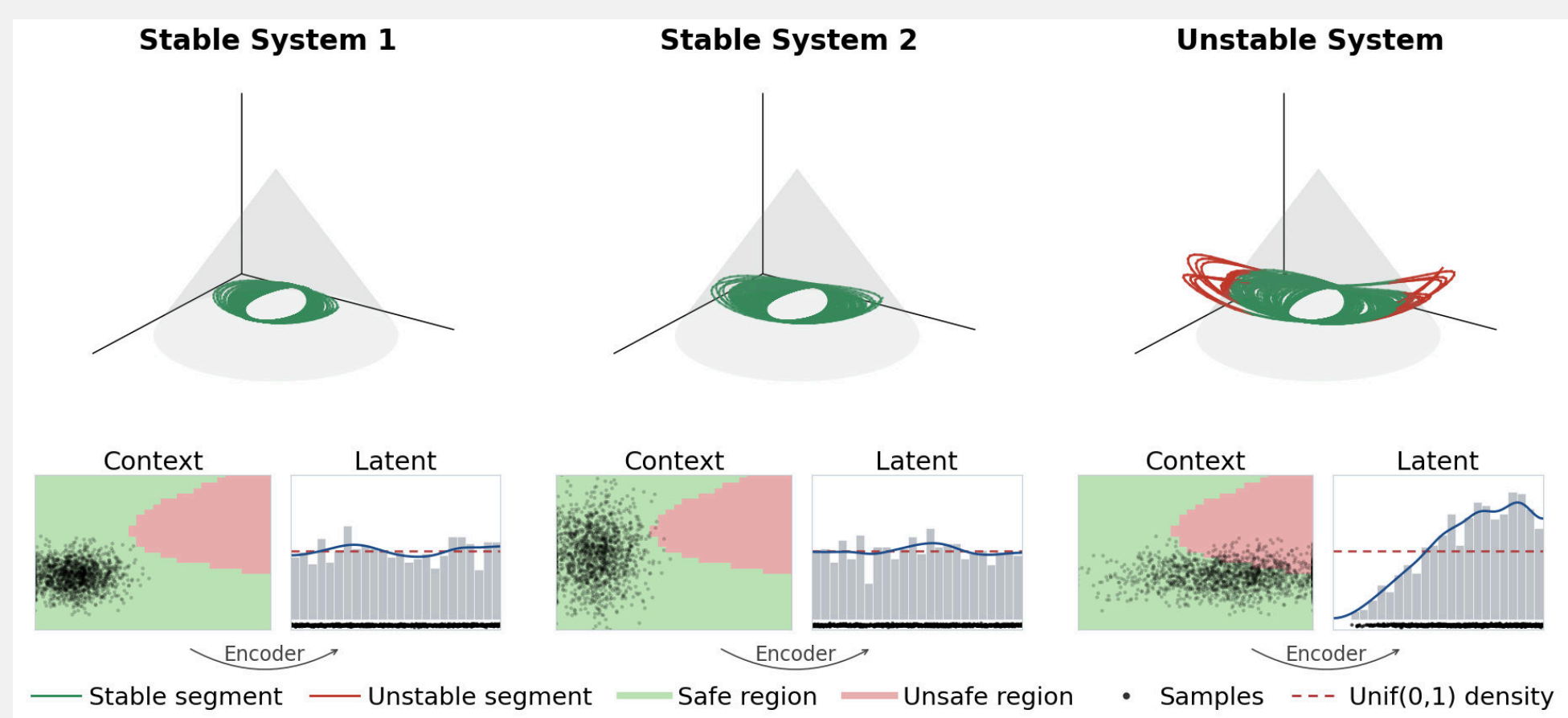
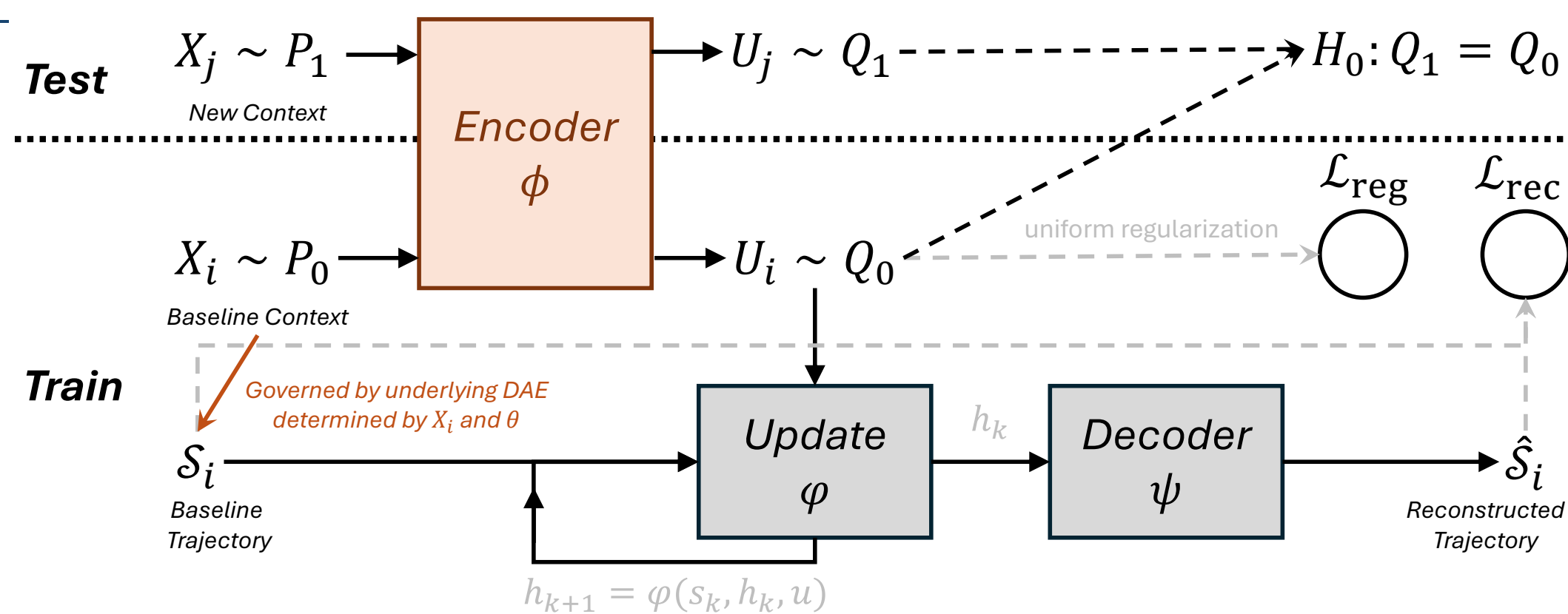


Figure 2. Cartesian pendulum, two stable regimes and one unstable. Contexts inside the safe region give an approximately **uniform** latent law; contexts leaking into the unsafe region make it **depart from uniformity** — instability becomes something we can **test for**.

Learning to Test (L2T)



An encoder $\phi : \mathcal{X} \rightarrow [0, 1]^d$ sends a context to $u = \phi(x)$, driving a neural-ODE surrogate and a decoder, $h_{k+1} = \varphi(s_k, h_k, u)$, $\hat{s}_k = \psi(h_k)$. One objective, three jobs:

$$\underbrace{\frac{1}{n} \sum_i \frac{1}{K_i} \sum_k \omega_{\text{stb}} \|s_{i,k} - \hat{s}_{i,k}\|^2}_{\text{trajectory reconstruction}} + \underbrace{\lambda \text{KL}[p_U \| \text{Unif}(0, 1)^d]}_{\text{testable latent geometry}}$$

- Reconstruction** makes u carry the DAE dynamics — this is what transfers to unseen regimes.
- Uniformity** gives a goodness-of-fit test against a *known* reference law.
- Stability reweighting** upweights unstable trajectories and sharpens the boundary.

Why the Latent Test Works

Assumption. The encoder preserves the stability-relevant information in the context:

$$\phi(X) \text{ is sufficient to distinguish stable vs. unstable regimes.}$$

Validity. If the deployment latent law stays sufficiently close to the safe reference, then the original stability requirement still holds:

$$D(Q_1, Q^*) \text{ small} \iff P_1 \in \mathcal{P}_0.$$

Power. If the deployment regime becomes unsafe, the latent law must separate:

$$D(Q_1, Q^*) \text{ large} \iff P_1 \notin \mathcal{P}_0.$$

Synthetic Systems: Validity Held, Power Retained

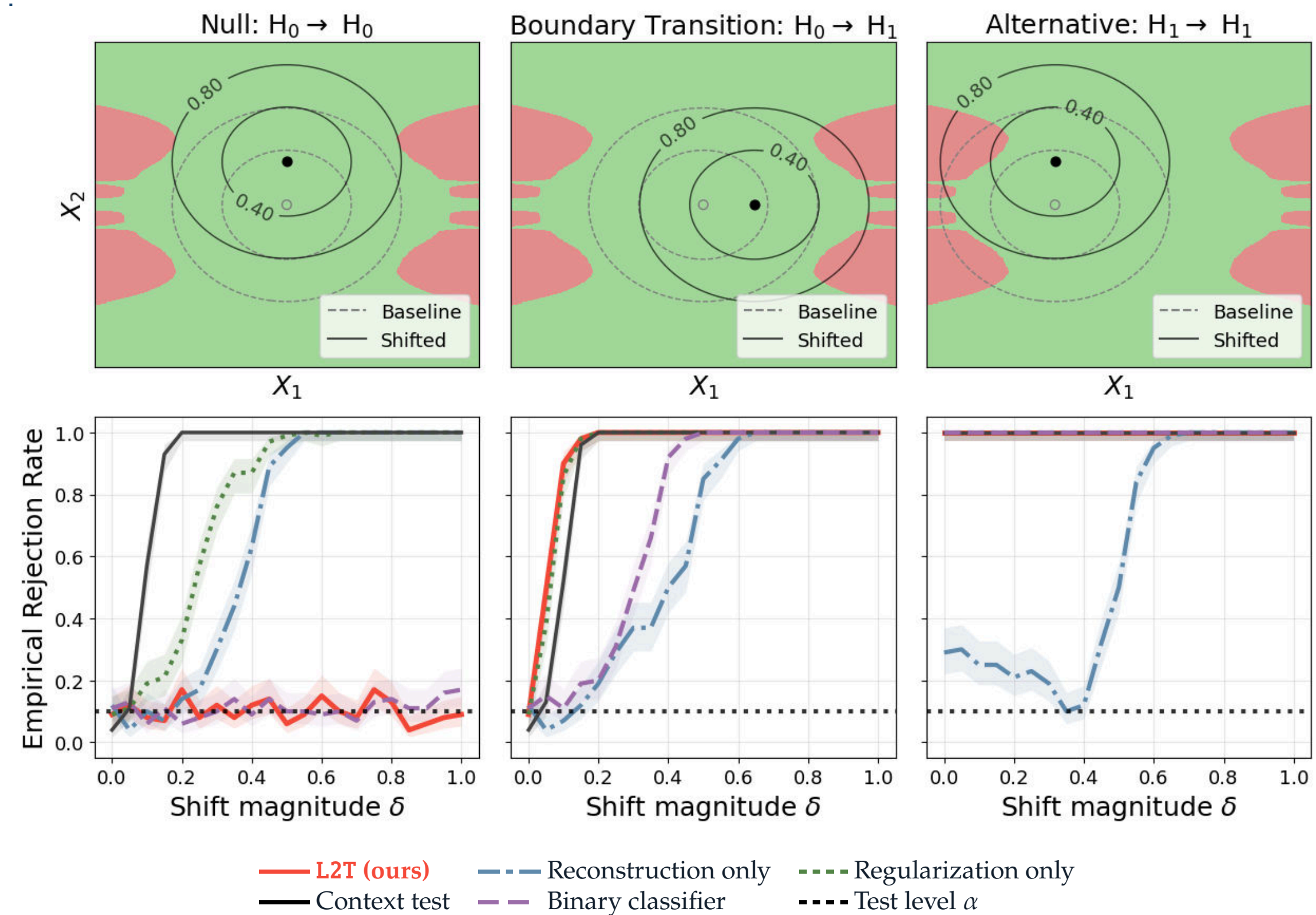


Figure 3. Mass-spring-damper. *Top*: baseline (dashed) and shifted (solid) contexts over **safe** / **unsafe** regions. *Bottom*: rejection rate vs. shift δ — only L2T holds the level *and* gains power.

IEEE 39-Bus Transient Stability

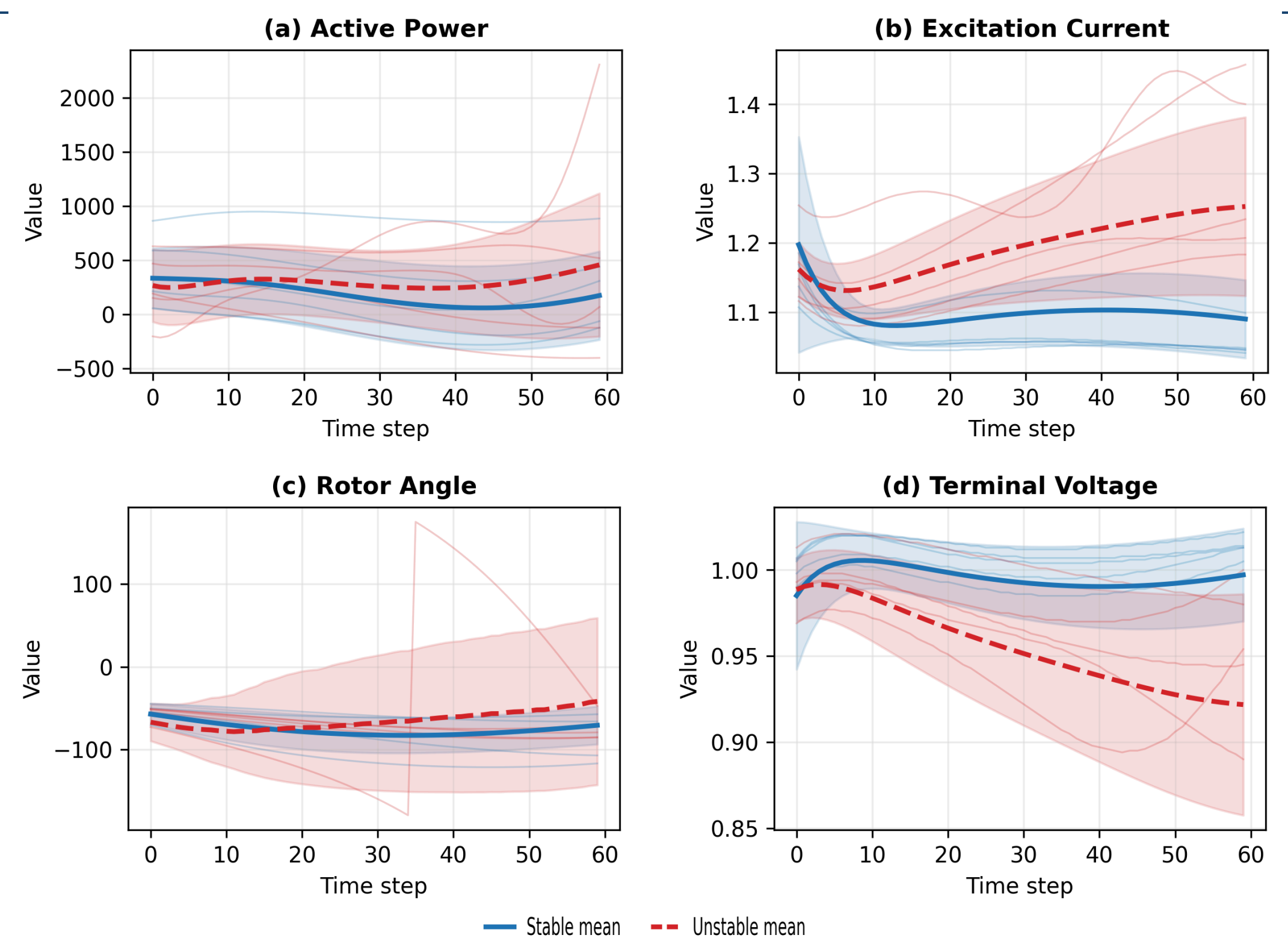


Figure 4. Generator under real faults: stable trajectories stay concentrated, unstable ones disperse.

	L2T (Flow)	L2T (MMD)	Recon. only	Regul. only	Context test	Binary clf.
Type-I @ 0.00	0.100	0.145	0.110	0.115	0.010	1.000
Type-II @ 0.20	0.005	0.435	0.000	0.710	0.175	0.000

Table 1. Type-I and Type-II error on IEEE 39-bus at level 0.1 (Type-I at 0% unstable, Type-II at 20%). Only L2T (Flow) holds the level *and* keeps power; 90% CI half-widths ≤ 0.06 .

Ablation: All Three Loss Terms Are Needed

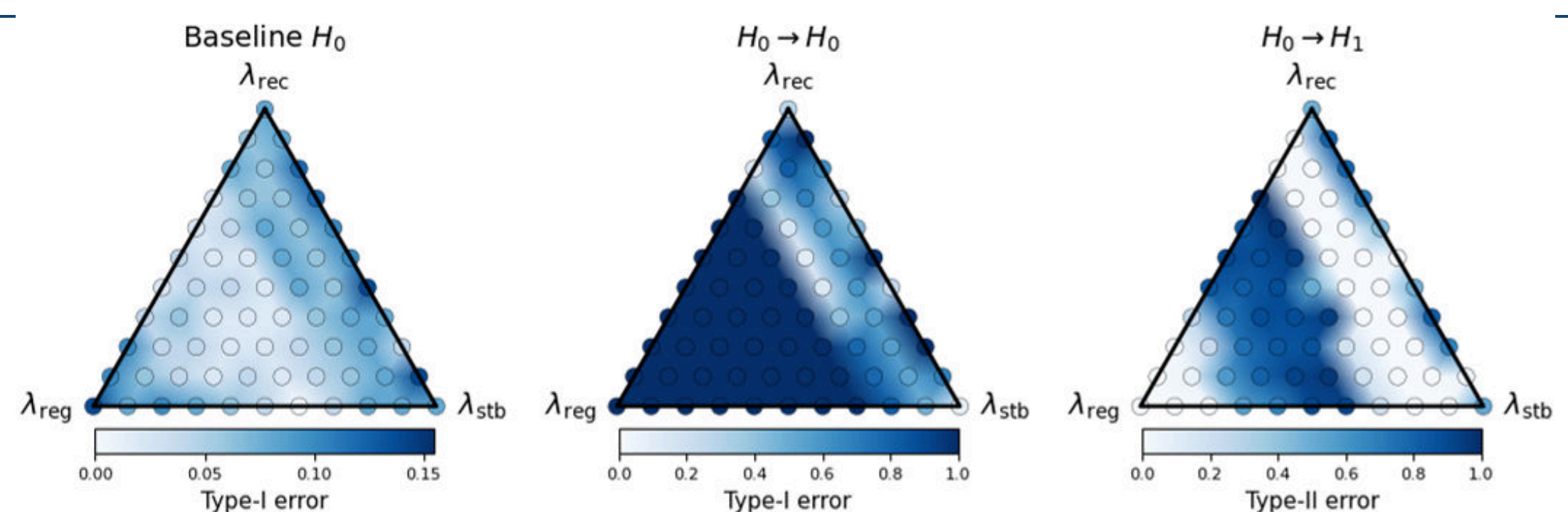


Figure 5. Error over the loss-weight simplex ($\lambda_{\text{rec}}, \lambda_{\text{reg}}, \lambda_{\text{stb}}$); darker is worse. The usable region is *interior*: no auxiliary term works in isolation.