Validation Geometry for Reward Optimization under Proxy Misspecification

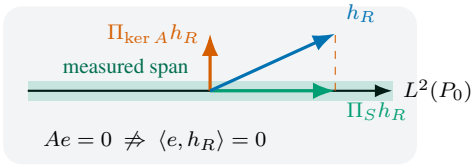
Mu Chen · chenm369@mail2.sysu.edu.cn · woody-sudo.github.io

Research question. What can a finite validation protocol certify when a reward proxy is misspecified and a specified optimizer acts on it? A proxy can pass every measured check while retaining residual errors aligned with the distribution shift induced by optimization.

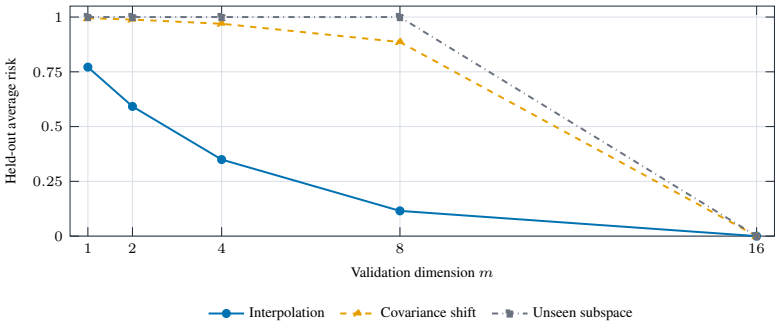
Endpoint bridge and sharp consequence. Let P_0 be the reference distribution, P_R the endpoint returned after optimizing the observed proxy R, and $e = R - U$ the unknown residual relative to utility U. For any fixed endpoint $P_R \ll P_0$, define

$$h_R = \frac{dP_R}{dP_0} - 1. \qquad B_\Omega(R,U) = \langle e, h_R \rangle, \qquad \sup_{Ae=0,\, \|e\| \le \varepsilon} |B_\Omega| = \varepsilon \|\Pi_{\ker A} h_R\|.$$

The endpoint bias vanishes for every validation-invisible residual exactly when $h_R \in \mathrm{Range}(A^*)$. The projection step is classical. The modeling contribution is the exact likelihood-ratio sensitivity that connects a nonlinear, iterative, or proxy-dependent optimizer endpoint to validation geometry.



Validation protects only the optimizer-sensitive directions that it spans.



A validator optimized on training directions transfers under interpolation but fails under rotated covariance and a genuinely unseen optimizer subspace at low measurement dimension.

Validation-design consequences. Leading eigenspaces are optimal for a frozen average-risk distribution. Kolmogorov widths instead characterize worst-case coverage over a declared optimizer class. When measurements come from a finite dictionary with cost and noise, neither result solves the operational design problem. In the frozen real measurement dictionary, the best feasible one-measurement design reduces held-out average squared radius from 1.23 to 0.76, while the unrestricted eigenspace lower bound is 0.28. The difference measures the implementability gap.

Evidence and scope. Exact kernel witnesses attain the sharp radius, and the phase transitions match sensitivity-span coverage. Noisy certificates fail when the sampling radius is removed. Four preregistered language-model-math cells retain null and reverse results and show strongly proxy-specific signed alignment. The result is an optimizer-conditioned diagnostic certificate. It does not learn true reward, cover support mismatch, or guarantee the absence of reward hacking.

Research direction. I am interested in extending this framework to adaptive measurements, online optimizer shift, structured reward uncertainty, and robust control under support change.

Full manuscript attached · Project Homepage.