Aligning Model Properties via Conformal Risk Control

William Overman
Jacqueline Jil Vallon
Mohsen Bayati

Problem Motivation

Background: AI models often **misalign** with user requirements due to biases in training data, underspecified objectives, or reward misspecification.[1,2]

Challenges: Many existing methods for alignment require costly retraining or human feedback and are mainly applicable to generative models.[3]

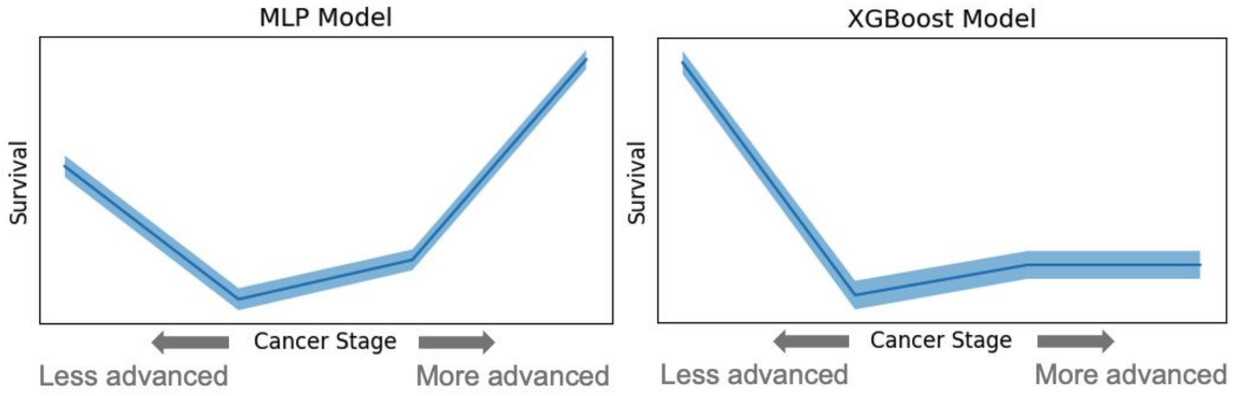
Goal: Align pre-trained models to user-desired properties, especially in non-generative contexts, using a post-processing technique.

1. Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective, 2024.
2. D'Amour et al.. Underspecification Presents Challenges for Credibility in Modern Machine Learning. JMLR, 2022.
3. Ji et al. AI Alignment: A Comprehensive Survey. 2023.

Non-generative settings, prostate cancer example

Modern clinical risk prediction models do not always effectively capture patterns required by the clinician

Ex. Predict 10 year survival probability of prostate cancer patient.[1]



Clinical stakeholder would likely not be happy with this

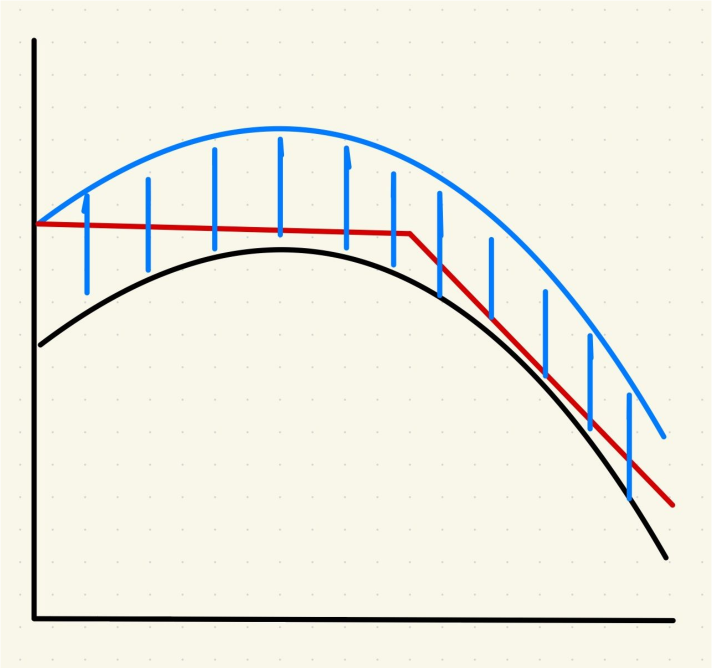
1. J. Vallon, W. Overman, W. Xu, N. Panjwani, X. Ling, S. Vij, H. Bagshaw, S. Srinivas, A. Fan, S. Shah, G. Sonn, J. Leppert, E. Pollom, M. Buyyounouski, M. Bayati. On Aligning Prediction Models with Clinical Experiential Learning: A Prostate Cancer Case Study

Think of nonincreasing behavior as a **property** of the model

Property testing in CS theory involves designing algorithms to check if a function has a certain property or is far from having it, using only a small, randomized sample of its input/output pairs.[1]

1. Oded Goldreich. Introduction to Property Testing. Cambridge University Press, 2017.

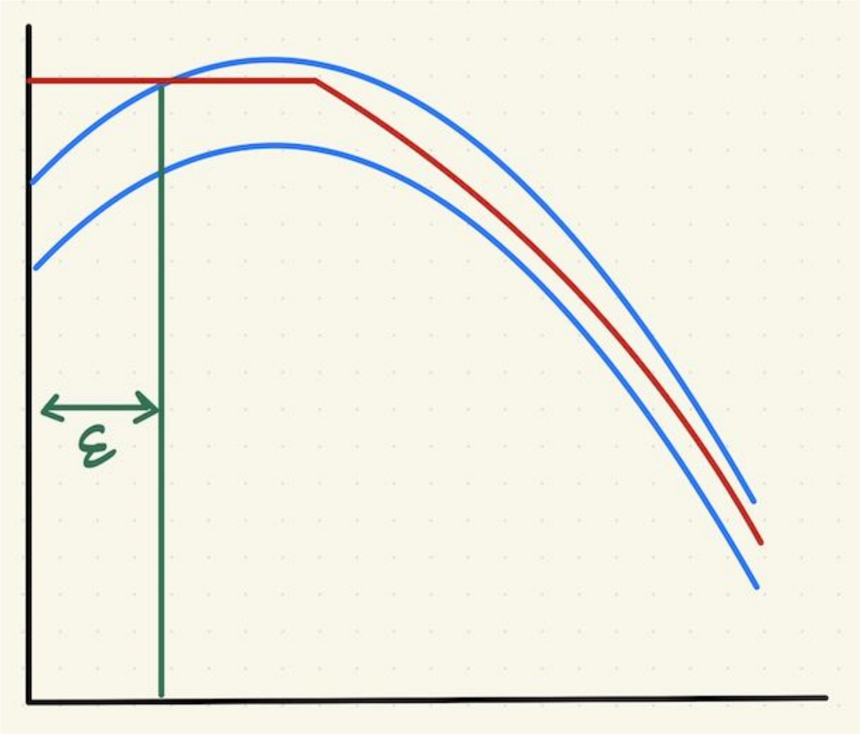
We say that a set-valued function F **accommodates** a property P if we can find a point $\hat{y} \in C(x)$ for each x such that the resulting function $g(x) = \hat{y}$ satisfies P.



- Original function
- Set-valued function
- Monotone function

The original function is not monotone, but we can fit a monotone function in the blue set-valued function

We say that F is ε-Faraway from P if any function g satisfying P falls outside of F on at least ε fraction of the data



The blue set-valued function is ε-Faraway from P

A **proximity oblivious tester** (POT) is a randomized algorithm such that

1. If F accommodates P, then Pr{M(F) = Accept} = 1
2. If F is ε-Faraway from P, then Pr{M(F) = Reject} >= p(ε)

Where p(ε) is monotone increasing in ε.

Algorithm 1 POT $\mathcal{T}$ for property $\mathcal{P}$ of monotonically decreasing in dimension k

Sample $X_1 \sim \mathcal{D}$. Let $X_1 = (x_1, x^{-k})$

Sample x_2 from the marginal distribution of $\mathcal{D}$ in dimension k. Set $X_2 = (x_2, x^{-k})$

if $x_1 < x_2$ and $\max F(X_1) < \min F(X_2)$ **then**

  return Reject

else if $x_2 < x_1$ and $\max F(X_2) < \min F(X_1)$ **then**

  return Reject

end if

return Accept

Conformal Risk Control

Procedure for converting point predictions of any black box model f(x) into set-valued predictions F(x) given a calibration set of n points (x_i, y_i)

Extends **conformal prediction** to notions of error beyond miscoverage.[1]

Main object is a parameter λ that controls our level of conservativeness, larger λ gives a larger prediction set $F_\lambda(x)$ around each point

We have a loss function L that is a function of both $F_\lambda(x_i)$ and the true label y_i

1. Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, Tal Schuster. Conformal Risk Control. In The Twelfth International Conference on Learning Representations, 2024.

Our Results

High Level Idea: Convert the output of a POT for a given property P into a loss function for conformal risk control

Ex. 0 loss if the POT accepts, 1 loss if the POT rejects

Conformal risk control then grows a prediction band F_λ around the pre-trained function f such that the expected loss falls below target

Our Results

Main Theorem (informal). Let T be a POT for property P. Assume we have access to a calibration dataset. If we run conformal risk control on this dataset with the loss functions generated by T then for any ε such that $p(\varepsilon)>\alpha$ the probability that the resulting function is ε -Faraway from P is at most $\alpha/p(\varepsilon)$

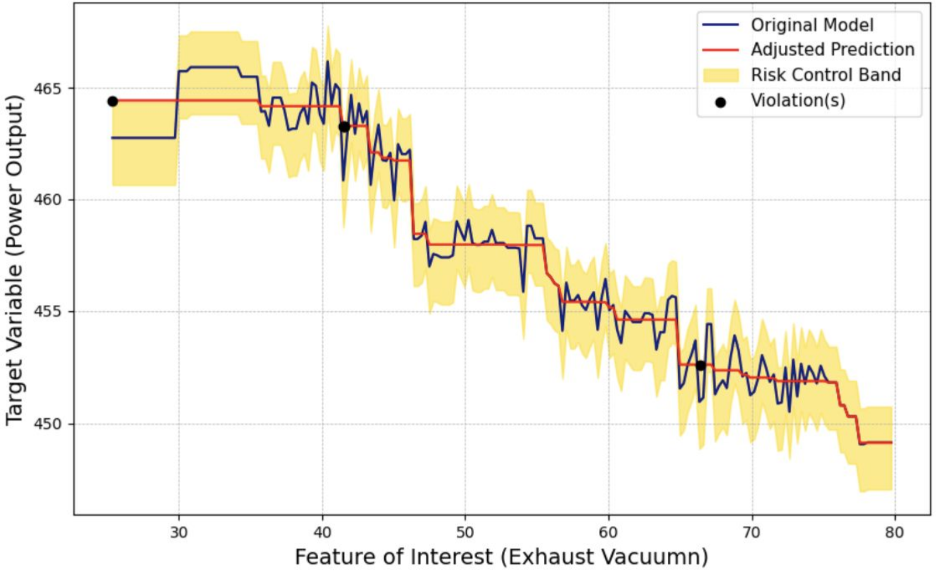
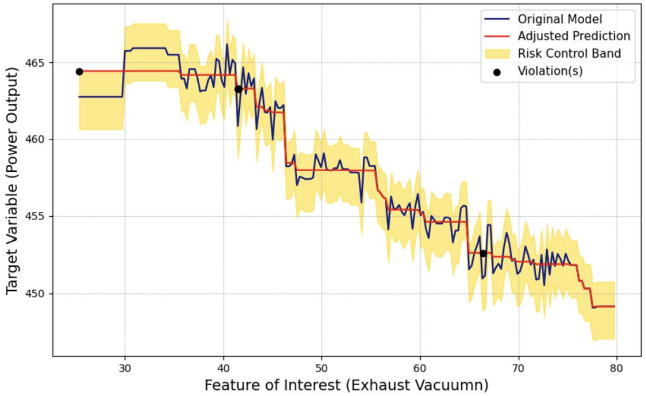


Table 1: Power Plant, $n = 9568$. Monotonically decreasing on Exhaust Vacuum. $\boldsymbol{\lambda}^{\max} = (10, 10)$.

α	$\boldsymbol{\lambda}$	Metric	Unconstrained	Adjusted	Constrained
0.1	$\lambda^+ = 0.51_{(\pm 0.24)}$	MSE	$10.19_{(\pm 0.46)}$	$10.47_{(\pm 0.46)}$	$16.21_{(\pm 0.45)}$
	$\lambda^- = 0.76_{(\pm 0.24)}$	Risk	$0.75_{(\pm 0.09)}$	$0.10_{(\pm 0.001)}$	$0.00_{(\pm 0.00)}$
0.05	$\lambda^+ = 1.09_{(\pm 0.51)}$	MSE	$10.19_{(\pm 0.46)}$	$11.42_{(\pm 0.44)}$	$16.21_{(\pm 0.45)}$
	$\lambda^- = 1.61_{(\pm 0.50)}$	Risk	$0.75_{(\pm 0.09)}$	$0.05_{(\pm 0.001)}$	$0.00_{(\pm 0.00)}$
0.01	$\lambda^+ = 2.39_{(\pm 0.82)}$	MSE	$10.19_{(\pm 0.46)}$	$14.46_{(\pm 0.48)}$	$16.21_{(\pm 0.45)}$
	$\lambda^- = 3.33_{(\pm 0.79)}$	Risk	$0.75_{(\pm 0.09)}$	$0.01_{(\pm 0.001)}$	$0.00_{(\pm 0.00)}$

Risk – violation of property
MSE– accuracy

$\lambda^+ + \lambda^-$ – size of the interval



Summary

Given a pretrained model that doesn't align with a desired behavior, we can use conformal risk control with loss functions coming from property testing to obtain a set-valued function that accommodates the desired property

Thank you!