Model updating after interventions paradoxically introduces bias

James Liley[1,2,⋆], Samuel R. Emerson[3], Bilal A. Mateen[1,4,5], Catalina A. Vallejos[1,2,⋆], Louis J. M. Aslett[1,3,⋆], Sebastian J. Vollmer[1,6,⋆]

[1] Alan Turing Institute, London, UK; [2] MRC Human Genetics Unit, Univ. of Edinburgh, UK; [3] Department of Mathematical Sciences, Durham Univ., UK; [4] Kings College Hospital, London, UK; [5] Wellcome Trust, London, UK; [6] Warwick Mathematics Institute, Univ. of Warwick, UK ⋆ Contact: james.liley@igmm.ed.ac.uk, catalina.vallejos@igmm.ed.ac.uk, louis.aslett@durham.ac.uk, svollmer@turing.ac.uk

1. Model updating with interventions

Introduction. Suppose we predict an outcome $Y \in \{0,1\}$ given predictors X, aiming to anticipate and avoid $Y = 1$ guided by predictions $\hat{\mathbb{E}}(Y|X)$. If such a model-intervention system is implemented (sometimes referred to as 'performative prediction' [2]), subsequent model updates can become 'victims of their own success' [1], in that they capture the effects of existing scores on data and outcomes, which change when the model is updated.

In particular, if a model is naively refitted to observations (X, Y) following intervention on a predictive score derived from an existing model, the refitted model cannot replace the old.

Contributions. Our main contributions in this work are:

1. Introduce a general causal framework under which this phenomenon can be quantitatively studied.
2. Use this framework to establish the hazards of naive model replacement, especially when it occurs repeatedly, in the context of a generalised ultimate aim of the predictive score.
3. Describe three broad strategies for avoiding these hazards.

4. Resolutions

Naive updating is only appropriate if no interventions are being made. Alternative strategies are:

More complex modelling. Completely model the causal setting by also observing covariates at $t = 1$ [3]. This enables an unbiased estimate of f_e by regressing Y on measured $X_e(1)$

Hold out set. Retain a set of samples for which ρ_e is not calculated. For such samples, $X_e(0) = X_e(1)$, so an unbiased estimate of f_e can be made by regressing Y on $X_e(1)$ in these samples. A problem is that any benefit of the risk score is lost for held-out samples.

Control interventions. Interventions g_e^ℓ and g_e^a may be directly specifiable. This enables directly solving equations (1), (2) for g^a, g^ℓ, and consequently unbiased estimation of f_e.

2. Model

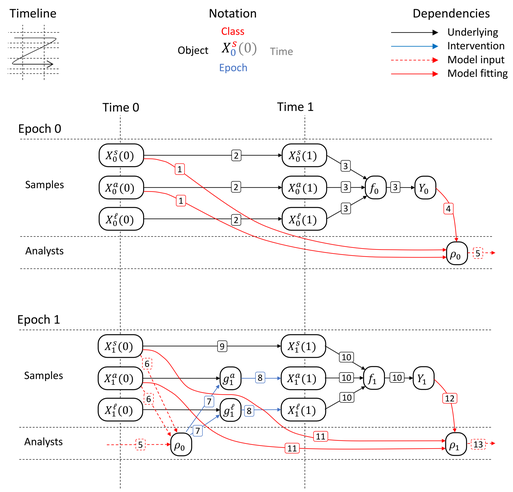

Figure 1: Causality diagram

Causal setup. We partition covariates X into

X^s : 'set'; observed, unchangeable

X^a : 'actionable'; observed, changeable

X^ℓ : 'latent'; unobserved, changeable

and associate with $X = X_e(t)$ an 'epoch' $e \in \mathbb{N}_0$ in which a new model is fitted and a 'time' $t \in \{0,1\}$ within e (Figure 1).

We suppose a predictive score $\rho_e(X^s, X^a)$ is produced at each e. Within epoch e,

At $t = 0$: ρ_e is calculated, $X_e^s(0)$, $X_e^a(0)$ are recorded

$t = 0$ **to** $t = 1$: X^a and X^ℓ are modified according to some functions $f^a(\rho_e, X^a(0))$, $f^\ell(\rho_e, X^\ell(0))$;

At $t = 1$: Y_e is determined on the basis of these covariates via a function f_e; values of $X_e^s(0)$, $X_e^a(0)$, Y_e are then used to determine ρ_{e+1}

Aim. We may presume the ultimate goal of the score/intervention system is to minimise

$$\mathbb{E}[Y_e] = \mathbb{E}_{X_e(0)}[Y_e|X_e(1)] \quad (1)$$
$$= \mathbb{E}_{X_e(0)}\left[f_e(X^s, g_e^a(\rho, X_e^a(0)), g_e^\ell(\rho, X_e^\ell(0)))\right]$$

with a cost constraint

$$\mathbb{E}\left[\text{COST}\left(X_e(0), g_e^a, g_e^\ell, \rho_e\right)\right] \leq C \quad (2)$$

We may either consider g_e^a and g_e^ℓ as fixed, and choose an optimal ρ_e or consider ρ_e fixed and choose optimal g_e^a, g_e^ℓ. If both are optimised, this becomes a generalised problem of resource allocation.

Control-theoretic formulation. We may frame the setting as a partially-observed Markov decision process (POMDP) as follows, using notation from [4] whereby we consider the POMDP as a 7-Tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \Omega, \mathcal{Z}, \gamma)$: where $\mathcal{S}, \mathcal{A}$ and Ω are spaces of states, actions and observations, $\mathcal{T}$ is a state transition kernel given $(s, a) \in (S, A)$, $\mathcal{Z}$ is a kernel for $\Omega|(s \in S)$, $r_e : S, A \to \mathbb{R}$ is a reward function, and γ is a discount factor A solution candidate is a policy

$$a_e \sim \pi\left(\{o_s, r_s, a_s\}_{s=1}^{e-1}\right)$$
$$= \arg\max_\pi \mathbb{E}\left[\sum_{e=1}^{M} \gamma^{e-1} r(s_e, a_e)\right]$$

Casting the above in this framework:

$$s_e = (X_e(0), X_e(1), Y_e) \qquad a_e = \rho_e$$
$$o_e = ((X_e^s(0), X_e^a(0)), Y_e) \qquad r_e = \mathbb{P}\left(\bar{Y}_{e+1} \mid s_e, a_e\right)$$

where $\bar{Y}$ is the rate of events in the population. With an appropriate choice of reward function, this can enable use of tools in this area.

5. Conclusions

Naive updating should generally be avoided, but all solutions are costly. This may lead to more accurately fitted models appearing to have worse performance. Strategies to update predictive scores should be designated when such scores are initially deployed. Causal formulations can help to describe and solve problems in this field, and PODMP-based formulations may enable use of reinforcement learning methods.

3. Consequences

If we intend $\rho_e(x^s, x^a)$ to be an estimator of $\mathbb{E}(Y|X_0^s = x^s, X_0^a = x^a)$ (a standard predictive score), a simple updating procedure estimates ρ_e by regressing Y_{e-1} on $X_{e-1}^s(0), X_{e-1}^a(0)$. If done repeatedly, we call this 'naive updating'. Naive updating is perilous; the following can occur:

1. Successive models which perform better in expectation in a 'fair' setting (i.e. when trained on data sets which have not been intervened on) appear to perform worse when interventions take place.
2. Successive estimates $\rho_e(x^s, x^a)$ for fixed x^s, x^a may tend towards a wide oscillation.
3. Successive estimates $\rho_e(x^s, x^a)$ may converge (a sufficient condition is slow change of g^a, g^ℓ, f; also see [2]) but the limiting value does not generally solve the constrained optimisation in equations (1), (2)

It is easily seen that these problems all worsen as the predictive score is more widely used; that is, as $g^a(\cdot, x^a)$ and $g^\ell(\cdot, x^\ell)$ move farther away from identity functions.

Examples of successive estimates of ρ_e oscillating or diverging are shown in figure 2.

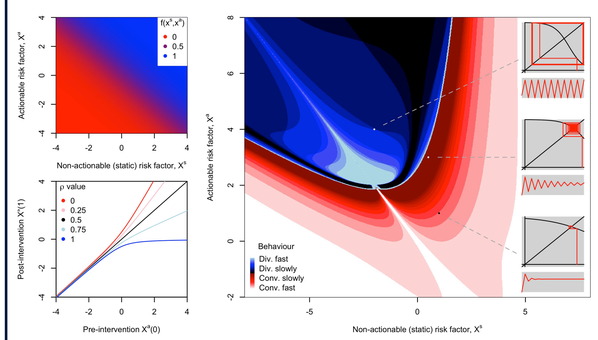

Figure 2. Convergence/divergence of typical ρ_e. We choose $f(x^s, x^a) = logit(x^s, x^a)$ (top left). We choose g^a to be 'zero-sum' (we intervene by lowering $X^a(0)$ when $\rho_e > 1/2$, but allow $X^a(0)$ to increase when $\rho_e < 1/2$; resources for intervention are redistributed rather than introduced), shown at bottom left. Depending on x^s, x^a, ρ_e converges or diverges at various rates [2]. Insets show cobweb plots and ρ_e against e.

References

[1] Matthew C Lenert, Michael E Matheny, and Colin G Walsh. Prognostic models will be victims of their own success, unless.... *Journal of the American Medical Informatics Association*, 26(12):1645–1650, 2019.
[2] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.
[3] Matthew Sperrin, Glen P Martin, Alexander Pate, Tjeerd Van Staa, Niels Peek, and Iain Buchan. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. *Statistics in Medicine*, 37(28):4142–4154, 2018.
[4] Yunbo Wang, Bo Liu, Jiajun Wu, Yuke Zhu, Simon S Du, Li Fei-Fei, and Joshua B Tenenbaum. DualSMC: Tunneling differentiable filtering and planning under continuous POMDPs. *ijcai.org*, 2019.

Funding and acknowledgements

We thank the Alan Turing Institute, Health Data Research UK (HDRUK), Wellcome Trust, Universities of Edinburgh, Durham, and Warwick, and Kings College Hospital, London for their support of the authors. We thank LJMA and Dr Ioanna Manolopoulou drawing our attention to this problem.
Funding grants were as follows. JL, CAV and LJMA: Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the "Health" theme within that grant and The Alan Turing Institute; JL, BAM, CAV, LJMA and SJV: HDRUK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England), the devolved administrations, and leading medical research charities; SRE: EPSRC doctoral training partnership (DTP), Durham University, grant reference EP/R513039/1; LJMA: Health Programme Fellowship, Alan Turing Institute; CAV: Chancellor's Fellowship, University of Edinburgh.