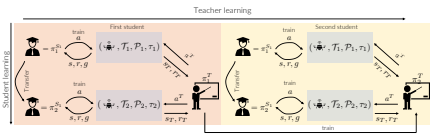
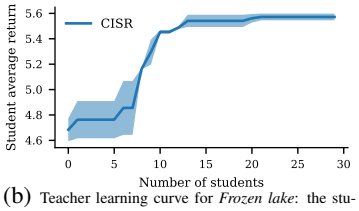
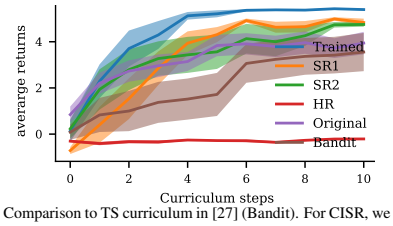


(a) Overview figure



(b) Teacher learning curve for *Frozen lake*: the student return induced by the teaching policy at the end of the curriculum improves as CISR trains more students.



(c) Comparison to TS curriculum in [27] (Bandit). For CISR, we evaluate a teacher policy trained w/30 students on new *test students*, while Bandit learns by explore-exploit for each student as [27] can't learn from previous students. This results in weak performance, esp. with HR proposed at the start of training resulting in poor student policies.

Common. Thank you for your helpful comments!

*Synthetic experiments **(R2, R4)**.* As you noted, the main contribution of this work is conceptual. As such, our experiments are in line with other conceptual CL-for-RL and safe-RL papers (e.g., [2,6,9,16,26,33)]), which evaluate on the same or similar problems due to their illustrative benefits. While synthetic, they are difficult especially from the safety standpoint. We agree it would be great to apply CISR to, say, an autonomous driving setting, but doing so would require motivating so many application-specific engineering choices that it is best done in a separate paper (e.g., [23]).

*Using multiple students **(R1, R2)**.* Using multiple students enables CISR's key novelty – allowing the teacher to *learn* a curriculum policy in a data-driven way. In contrast, in single-student CL such as [21,27] the teacher continually estimates the student's partially observed internal state and *heuristically applies* interventions based on these estimates, but the state estimates $\rightarrow$ interventions mapping – the curriculum policy – is fixed, encoded into the teacher's algorithm. Among other things, this lets CISR produce curriculum policies that are robust to student diversity (see Table 1 caption, last 3 lines). *This makes CISR applicable,e.g., in a flavor of sim-to-real transfer where a curriculum policy is learned in a crude simulator and then deployed for training real-world agents in safety-sensitive settings such as robotics.*

*Empirical benefits of multiple students and comparison to prior work **(R1, R2, R4)**.* The reviews gave great ideas for improving these aspects, and we ran additional experiments, to be included in the revised paper version. Fig. (c) shows a comparison to [27] and Fig. (b) shows how the teacher improves with multiple students.

*Proposition 2 and Assumption 2 **(R2, R3)**.* Prop 2 only says that *if* interventions are absolutely safe *then* CISR ensures absolute training safety. Assumption 2 is for conceptual simplicity, but can hold in reality: systems such as aircraft stall prevention and collision avoidance guarantee near-absolute safety. Even in the absence thereof, CISR, informally, keeps the student as safe during training as teacher's interventions allow. In Lunar Lander experiments, intervention safety is not absolute ("Train fail" column in Table 1), but CISR still improves training safety by 1000x over existing approaches.

*Clarity **(R1, R2, R4)**.* We'll add signposts, including Fig. (a) above, examples, and rework the interventions explanation.

R1: While simpler heuristics might be possible for small K, note that, e.g., in Frozen Lakes the curriculum space is large even for $K = 2$ and just 3 interventions. With 10 curriculum steps per student, there are $9 \cdot 10/2 = 45$ choices for 2 switching points, each with $3! = 6$ intervention orderings. Thus, we have at least 270 possible curricula. The fact that CISR determines a good one after only 10 students attests to its learning ability.

R2: Please see the **common** responses above. We'll add the related work you mentioned.

R3: *Teacher's dynamics knowledge.* The teacher doesn't need it because it just applies pre-designed controllers in pre-specified states. Designing such controllers may require knowing *local* dynamics around dangerous states, but this is still far less restrictive than a full dynamics knowledge assumption.

"Good" behavior of controllers. We also don't assume the interventions to be "good" in the sense of reward performance. E.g,. in reality, emergency breaking may induce undesirable behavior. Their main role is just to keep the student safe, so that it can eventually learn to avoid triggering these interventions in the first place.

Teacher can't violate dynamics. What we mean is that the reset distribution $\mathcal{T}(\cdot, s)$ should be realizable based on the CMDP's dynamics and the teacher's primitive actions, e.g., by using an option. E.g., a parent helping a child stay upright when riding a bicycle doesn't violate physics, just applies actions unavailable to the child.

Teacher's reward knowledge. We only assume that the teacher has *a* reward notion – which may not match the student's – in order to guide the learning process.

Proposition 1 is correct because the constraint over $\mathcal{D}$ is also present in the intervention CMDP in Eq (2).

R4: *The threshold array* is part of the teaching policy parameters and it's learned via Bayesian optimization, see Fig (b).

Interventions: Under intervention i, the student triggers the teacher's help whenever it enters $\mathcal{D}_i$ and transitions according to $\mathcal{T}_i$. The features for the teacher depend on the constraint violation (l. 297) and, thus, on what is dangerous in the environment according to the interventions (learning the interventions is not our focus but it has been done [15]).

We'll incorporate remaining reviewer comments, including related work, into the paper as well.