1 Broader Impact Statement

The research presented in this paper on covert or learner-private optimization has significant broader impacts across multiple areas, particularly in privacy-preserving machine learning, cybersecurity, and social media moderation.

1. *Privacy-Preserving Machine Learning*: The proposed methods enhance the confidentiality of gradient-based learning processes in distributed environments. By ensuring that a learner can hide its learning activity from malicious eavesdroppers, the research contributes to the field of privacy-preserving machine learning. This is crucial for applications where sensitive data is involved, such as healthcare, finance, and personal data management, where privacy breaches can have severe consequences.

2. *Cybersecurity*: The framework for covert optimization protects against adversaries who might exploit learning processes to gather insights or reverse-engineer models. This is especially important in scenarios where sensitive intellectual property or strategic data processing methods are at risk. The ability to choose between learning and hiding strengthens systems against such malicious activities, contributing to a more secure digital environment.

3. *Social Media and Content Moderation*: The practical application demonstrated in the hate speech classification task shows the potential impact on social media platforms and content moderation systems. By preventing eavesdroppers from accurately learning the model used for detecting toxic content, the proposed methods help in stopping attempts to create and spread harmful material that can evade automated detection systems. This contributes to creating safer online communities and reducing the spread of toxic and harmful content.

4. *Federated Learning*: In the context of federated learning, where multiple distributed clients work together to train a model without sharing raw data, the proposed covert optimization techniques ensure that individual contributions remain private. This can enhance user trust and participation in federated learning projects, leading to stronger and more representative models, especially in fields requiring high privacy standards.

5. *Algorithmic Fairness and Ethical AI*: By introducing methods that can limit the information leakage through model queries, the research addresses concerns around algorithmic fairness and ethical AI. Ensuring that sensitive or important information is not inadvertently exposed through model updates helps in maintaining the integrity and fairness of machine learning systems, preventing potential misuse or biased exploitation.

6. *Future Research Directions*: The structure and policy identified in this paper open new opportunities for future research in controlling and optimizing learning processes. The development of efficient algorithms without needing transition probabilities is a significant step forward, potentially influencing a wide range of applications in decision-making under uncertainty.

Submission Checklist

1. For all authors…

   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes**

   (b) Did you describe the limitations of your work? **Yes**

   (c) Did you discuss any potential negative societal impacts of your work? **Yes**

   (d) Did you read the ethics review guidelines and ensure that your paper conforms to them? `https://2022.automl.cc/ethics-accessibility/` **Yes**

2. If you ran experiments…

   (a) Did you use the same evaluation protocol for all methods being compared (e.g., same benchmarks, data (sub)sets, available resources)? **Yes**

   (b) Did you specify all the necessary details of your evaluation (e.g., data splits, pre-processing, search spaces, hyperparameter tuning)? **Yes**

   (c) Did you repeat your experiments (e.g., across multiple random seeds or splits) to account for the impact of randomness in your methods or data? **Yes**

   (d) Did you report the uncertainty of your results (e.g., the variance across random seeds or splits)? **Yes**

   (e) Did you report the statistical significance of your results? **No**

   (f) Did you use tabular or surrogate benchmarks for in-depth evaluations? **No**

   (g) Did you compare performance over time and describe how you selected the maximum duration? **Yes**

   (h) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **No**

   (i) Did you run ablation studies to assess the impact of different components of your approach? **Yes**

3. With respect to the code used to obtain your results…

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results, including all requirements (e.g., `requirements.txt` with explicit versions), random seeds, an instructive `README` with installation, and execution commands (either in the supplemental material or as a URL)? **Yes**

   (b) Did you include a minimal example to replicate results on a small subset of the experiments or on toy data? **Yes**

   (c) Did you ensure sufficient code quality and documentation so that someone else can execute and understand your code? **Yes**

   (d) Did you include the raw results of running your experiments with the given code, data, and instructions? **Yes**

   (e) Did you include the code, additional data, and instructions needed to generate the figures and tables in your paper based on the raw results? **Yes**

4. If you used existing assets (e.g., code, data, models)…

(a) Did you cite the creators of used assets? **Yes**

(b) Did you discuss whether and how consent was obtained from people whose data you're using/curating if the license requires it? **Yes**

(c) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Yes**

5. If you created/released new assets (e.g., code, data, models)…

   (a) Did you mention the license of the new assets (e.g., as part of your code submission)? **Yes**

   (b) Did you include the new assets either in the supplemental material or as a URL (to, e.g., GitHub or Hugging Face)? **Yes**

6. If you used crowdsourcing or conducted research with human subjects…

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **Yes**

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **NA**

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **NA**

7. If you included theoretical results…

   (a) Did you state the full set of assumptions of all theoretical results? **Yes**

   (b) Did you include complete proofs of all theoretical results? **Yes**