Statement of Changes and Response to Reviews for L-CSS Submission

Adit Jain*, Vikram Krishnamurthy

May 3, 2024

Editor,
Control System Letters

We thank the associated editor and the reviewers for providing insightful feedback on our submission titled "Structured Reinforcement Learning for Incentivized Stochastic Covert Optimization" submitted to the control system letters (L-CSS) journal. We appreciate the opportunity to revise the manuscript and resubmit, incorporating changes to address the comments. Attached is our response to the comments by the editor and reviewers.

1 Response to the Associated Editor

In this letter, we summarize our responses to the three key issues highlighted by the reviewers and address the reviewers' comments individually, along with the changes made. The reviews raised concerns regarding the following issues,

1. Motivation and Relevance: Our submitted work solves the problem of dynamic covert optimization by formulating a Markov decision process whose optimal policy is approximated using a stochastic approximation algorithm, both of which are central to the theme of control theory, and therefore, we believe that the paper is of interest to the audience of control system letters. Our method focuses on how stochastic control can switch between two stochastic gradient descents to achieve an objective (covert optimization in this case). Our submitted work complements well with existing work on controlling and analyzing distributed optimization setup in recent control system letters editions [8, 3, 9]. Although the proposed motivating example in hate-speech classification is not control-theoretic, as the reviewer pointed out, it can indeed be adapted to a control-theoretic example, like a production-inventory system.

We have incorporated the reviewer's suggestion and have mentioned in the main text how covert optimization can be used to privately optimize inventory using a distributed setup [4, 14].

2. Related Work and Experimental Benchmarking: The related work section has been made prominent in the revised submission, along with a clear difference between the state-of-the-art and the previous work. To reiterate, the current state of the art is focused on deriving bounds on the query complexity. It does not consider an oracle whose stochasticity can be exploited, and hence, a random obfuscation policy would suffice for the framework considered in the previous work.

In contrast, [5] was the first work to look at dynamically obfuscating the eavesdropper in the presence of a Bayesian eavesdropper, our submitted manuscript improves theoretically upon [5] by a) relaxing the assumptions of supermodularity with interval dominance b) considering an incentivized oracle and c) an explicit cost function based on the eavesdropper belief. We have rerun the experiments and have added a benchmark with [5] in the numerical section of the paper using a constant incentive and binary action space. As discussed in the manuscript, the bandit algorithm is prohibitively expensive to use in practice due to the exponential dependence on the state dimension, however we have demostrated the numerical performance on a toy example, mentioned below. Due to space constraints, we have included these numerical benchmarks and additional figures in an online supplementary document available at Link. We have attached the same at the end of this letter and as a separate file for the convenience of the editor and reviewers.

3. Exposition and Clarity of Problem and Proposed Solution: We have substantially improved the introduction, clearly explaining the problem statement and the control variables, and have combined the two stochastic gradients into a single controlled and vectorized equation (2). Equation (3) has been added to describe how the query is generated based on the control variable. We have introduced an algorithm that discusses the stochastic control approach of the learner and has also explained the threshold result and its implication more clearly. We have checked for typos and ensured all the notations are clearly explained. For the numerical experiments, we have clarified how the structural results' assumptions are satisfied. We have cleaned our code base to guide the interested reader and commented on most simulation files.

We now respond to each of the comments in the reviews and highlight the corresponding changes in the text

2 Response to Reviewer 29

2.1 *The repository containing the experiments requires major modifications to guide users through the various files and their usage. For this purpose, please adhere to clean code practices and strive to provide reproducible experiments.*

Response: We have updated the codebase, accessible on **www.github.com/aditj/CovertOptimization**. In particular, we have ensured that the code is organized into logical modules, introduced a README with exact instructions on pre-processing

the data and running the code, specified the random seed, and added docstrings to most of the functions and classes. We have tried our best to make the code reproducible and adhere to standard coding practices.

2.2 The paper does not adequately motivate the problem of social learners as part of future work. To strengthen the argument, the author should provide more context about the problem's Relevance and how it explicitly connects to the paper's central topic.

Response: To be more explicit about how social learning can be incorporated in future work and relating it to the central topic of the paper, we have added the following paragraph in the *Conclusion* section of the text,

Changes To Text: "In future work, the problem of obfuscating sequential eavesdroppers can be formulated as a Bayesian social learning problem, where initially the eavesdropper is obfuscated maximally to make it stop participating, and its departure indicates the subsequent eavesdroppers that the learner is obfuscating. Hence, the eavesdroppers can eventually be made to herd, forming an information cascade so that they don't eavesdrop anymore regardless of whether the learner is learning or not."

2.3 To enhance readability, please rework the footnotes into the main body of the paper. Their current placement is confusing, but the information they contain is important.

Response: We acknowledge that some of the footnotes were not placed appropriately, and we have rearranged many. Specifically, we have rearranged footnotes 2, 4, and 7 in the main text.

2.4 The paper would benefit from plotting the validation accuracy curves for Learner and Eavesdropper, as done in [5]. These curves could establish useful learning speed baselines for future work.

We have generated the validation accuracy curve similar to our previous work [5]; however, due to space constraints; we have put it in the online supplementary material available at add link, and also have attached them at the end of the document (Fig. **??**).

2.5 Please ensure consistent use of the acronym 'MDP' for 'Markov Decision Process' throughout the text. Avo d switching between the full term and the acronym

Response: We have replaced Markov Decision Process in the subsection title and the conclusion with MDP, and we have ensured that no other acronyms are repeated.

2.6 Instead of using 'arg min,' consider using a clearer term to explain the eavesdropper's objective. For example, you could say, "...as an approximation for the estimate of the function the eavesdropper wants to minimize..." This would make the problem easier to understand.

Response: Thanks for this feedback; we have changed the introduction to describe the eavesdropper objective as,

Changes To Text: The eavesdropper aims to estimate $\hat{x}$ as an approximation to the minimizer of the function the eavesdropper is interested in optimizing.

2.7 Equation (8) uses the function d without explicitly defining it. Given the context, it's likely d represents the initial cost function. Specifying its definition would enhance reader understanding.

Response: We have improved the text by explicitly highlighting the definition of queue cost before the cumulative cost equation,

Changes To Text: After N queries, the learner pays a terminal queue cost computed using the function $d : \mathcal{Y} \to \mathbb{R}$.

2.8 As there is no definition for the regret function, add a reference.

Response: We have added the reference [2] pointing to the regret definition, which is the same as the reference to the regret result, and have made the following change in the Multi-Armed Bandit subsection,

Changes To Text: For brevity, we omit the definition of regret and the exact upper bound, both of which can be found in Chapter 2 of [2].

3 Response to Reviewer 25

3.1 The motivation example is not control-related, as expected for publication in L-CSS. It follows that the paper is out of scope as is. However, I believe it can be improved by going in the direction of the lines that follow the "Motivation" paragraph. Ind ed, by adapting the motivation to some production-inventory system, the paper could be made in scope.

Response: We believe that the paper is in the scope of L-CSS since the paper formulates a *Markov decision process* for dynamic covert optimization and proposes a *stochastic approximation algorithm* to estimate the optimal policy with a threshold structure both of which are central to the theme of control theory. The motivating example presented in the text was not control-related, and a motivating example related to distributed optimization for production inventory management would be more apt for L-CSS. We have incorporated the reviewer's feedback and made the following changes in the text to make pricing optimization and inventory management the main motivating examples in the introduction, both of which use distributed optimization in industrial applications.

Changes To Text: One motivating example is in pricing optimization and inventory management, where the learner (e.g., e-retailer) queries the distributed oracle (e.g., customers) pricing and product preferences to estimate the optimal price and quantity of a product to optimize the profit function [10, 1]. A competitor could spoof as a customer and use the optimal price and quantity for their competitive advantage. Our numerical experiment illustrates another application in federated learning. We aimed to present an interesting application of stochastic control and structural results in federated learning, a form of large-scale privacy-preserving distributed optimization. We'd also like to highlight another reason the paper interests the L-CSS audience: it provides a structured approach to controlling distributed optimization by formulating an MDP whose optimal policy has a threshold structure. Our proposed stochastic methods can be extended to other objectives, including fairness, robustness, and communication efficiency in controlling distributed optimization.

3.2 It is said that the learner's goal is to estimate a critical point of a given function unknown to the learner. This point is denoted x^*. Then, the authors mention a query q_k. Wha is is? Is t the current estimate of x^*? Then the authors mention the obfuscation task. Is he query q_k the simulated estimate z when using "obfuscate" as an action?

Response: The query q_k is a point at which the learner wishes to know the approximate gradient. Ind ed, for our obfuscation scheme, this query is either the current learning stochastic gradient estimate $\hat{x}_k$ when learning or the obfuscating stochastic gradient estimate $\hat{z}_k$. We have added an equation in the introduction to make this clearer and have made it reiterated in the text while discussing the obfuscation strategy to make sure this point is more precise,

Changes To Text: In summary, the learner obfuscates and learns by dynamically choosing the query q_k, as the current estimate $\hat{x}_{k-1}$ from the controlled stochastic gradient step of (1) or as the estimate $\hat{z}_{k-1}$ of parallel SGD of (**??**)

Also, in the algorithm we have added for the approach, we highlight the step to choose the query based on the chosen stochastic gradient algorithm.

3.3 Overall, the proposed strategy is quite difficult to follow. It is hard to get how all the agents interact, what information they exchange exactly, and, overall, what is the final objective.

Response: We have improved the clarity of the text and exposition of the problem and proposed a solution by making several changes to the text. We provide a detailed list of changes to ensure that the proposed strategy is easier to follow,

1. *Rewritten Introduction.* The introduction clearly explains the three agents: the learner, the eavesdropper, and the oracle. The learner sends a query to the oracle, the oracle replies stochastically with a noisy gradient, and an eavesdropper listens to the queries. The learner aims to optimize the function and obfuscate the eavesdropper simultaneously.

   We have rewritten the introduction to explain dynamic covert optimization as a control of two stochastic gradient algorithms. The learner aims to perform stochastic control to minimize the expected finite horizon cost.

   Changes To Text: This paper addresses the question: *Suppose the learner uses a stochastic gradient (SG) algorithm to obtain an estimate $\hat{x}$. How can the learner control the SG to hide $\hat{x}$ from an eavesdropper?*

   Our proposed approach is to switch between two stochastic gradient algorithms dynamically. Let $a_k \in \{0 = \texttt{Obfuscate SG}, 1 = \texttt{Learn SG}\}$ denote the chosen SG at time k. The first SG minimizes function f and updates the learner estimate $\hat{x}_k$. The second SG is for obfuscation and confusing the eavesdropper with estimates $\hat{z}_k$. The equation gives the update of both SGs,

$$\begin{bmatrix} \hat{x}_{k+1} \\ \hat{z}_{k+1} \end{bmatrix} = \begin{bmatrix} \hat{x}_k \\ \hat{z}_k \end{bmatrix} - \mu_k \begin{bmatrix} \mathbb{1}(a_k = 1) & 0 \\ 0 & \mathbb{1}(a_k = 0) \end{bmatrix} \begin{bmatrix} r_k \\ \bar{r}_k \end{bmatrix}, \tag{1}$$

   where μ_k is the step size, $\bar{r}_k$ is a synthetic gradient response discussed later and a_k controls the SG to update.

2. We have now added Algorithm 1 in the main text, which clearly explains the exact sequence of actions by the learner.

3. We have updated the Fig. 1 caption so that it better explains the interaction of agents and the objective,

   Changes To Text: Learner sends query q_k and incentive i_k to oracle in state o_k. The oracle evaluates noisy gradient of f at q_k, r_k according to (**??**). An eavesdropper observes q_k and i_k and aims to approximate the learner's estimate. The learner needs to control the incentive i_k and type of SG (a_k) to query using (**??**) to achieve the learning objective of (**??**) and obfuscate the eavesdropper with belief (**??**).

4. We have also included an additional figure (Fig. 1) illustrating the stochastic control approach in the online supplementary document available at link.

3.4 What is, in practice, the incentive? What does it represent?

Response: The incentive is the cost the learner bears to improve the changes of the stochastic gradients being of sufficiently good quality. In practice, it could be a monetary cost or equivalent cost that the learner pays to the distributed oracle to increase the computational resources used to compute the gradient or compute the gradient on a specific sampled datum. We have included the following line in the motivation to make this point clear in the text,

Changes To Text: An incentive that the learner pays is motivated by the learner's cost for obtaining a gradient evaluation of desired quality, e.g., a monetary compensation the learner pays to participating clients of the distributed oracle [13, 11].

3.5 Several notations are used without being introduced. (Γ in equation (1), w.p. in O2, ...).

Response: We apologize and have now carefully proofread the manuscript to ensure no notation remains unexplained.

3.6 In equation (1), please do not use the empty set symbol if you do not mean empty set. r is not a set but a vector.

Response: Thanks for pointing this out; we have tried to make sure there are no other typos. Indeed, the response r_k is a zero-vector when the response is non-informative, and this was a typo that has now been rectified.

4 Response to Reviewer 14

4.1 The discussion about related work is not satisfactory. The last but one paragraph in Section I introduces the related works about convert optimization. However, it does not clarify the difference between these works and the submitted paper. Thus, the state-of-the-art works and algorithms in the literature are unclear.

Response: We agree that the discussion regarding the related work had been limited due to space constraints. How ver we have now reformulated the related work and contribution sections to have a coherent flow. We want to highlight that the previous work looked at covert optimization from a static oracle; hence, even a random obfuscation policy would suffice. The previous work is focused on deriving statistical guarantees and doesn't propose any particular algorithm. Our recent work [5] is the first to consider a stochastic oracle; therefore, the application of stochastic control is novel.

The revised manuscript has the following updated paragraph, which clearly highlights how the current work and [5] differ from the existing work in covert optimization. It urther details the two key theoretical differences betweent the current work and [5].

Changes To Text: **Related Works.** The current literature on covert optimization has been focused on deriving upper and lower bounds on the query complexity for a given obfuscation level [12]. Que y complexity for binary and convex covert optimization with a Bayesian eavesdropper has been studied in [10, 12]. The bounds assume a static oracle and a random querying policy can be used to obfuscate and learn randomly. In contrast, the authors have looked at dynamic covert optimization where stochastic control is used to query a stochastic oracle optimally [5]. This is starkly different than the current literature since a stochastic oracle models situations where the quality of gradient responses may vary (e.g., due to Markovian client participation). The success of a response can be determined by the learner (e.g., based on gradient quality [5]) or by the oracle (e.g., based on computational resources availability).

Contributions and differences from previous work [5] To prove that the optimal policy has a monotone threshold structure, [5] requires supermodularity conditions. This paper proves results under more relaxed conditions using interval dominance [7] in Theorem **??**, which can incorporate convex cost functions and more general transition probabilities [6].

⋮

The action space in this work includes an incentive the learner provides to the oracle. An ncentive that the learner pays is motivated by the learner's cost for obtaining a gradient evaluation of desired quality, e.g., a monetary compensation the learner pays to participating clients of the distributed oracle [13, 11].

4.2 I would suggest the authors include a figure or an algorithm that clearly demonstrates the entire algorithm with all components, especially clarify the relationship between two stochastic gradient updates (3) and (4), how to switch them, and different behaviors of these two updates, to avoid any confusion.

Response: We appreciate the feedback and have included the Algorithm 1 (next page) in the main text. This algorithm highlights how the learner uses a threshold policy to dynamically switch between two stochastic gradients and poses queries from the estimates of the respective gradients' descents. We apologize for the inability to include an Figure in the main text due to space constraints, however we have included the Figure **??** in our online supplementary document.

Changes To Text:

Algorithm 1 Stochastic Control for Covert Optimization

Input: Policy π, Queries N, Successful Gradient Steps M
Initialize learner queue state $b_N = M$
for k in $1, \ldots, N$ **do**
  Obtain type of SG and incentive, $(a_k, i_k) = \pi(o_k, b_k)$
  Incur cost $c((a_k, i_k), (o_k, b_k))$ from (2)
  Query oracle using query q_k (3) and incentive i_k
  Receive response r_k and success of reply s_k
  Update estimates of the two SGs using (1).
  if s_k=1 **then** $b_{k+1} = b_k - s_k$
  Oracle state evolves, $o_{k+1} \sim \Delta(\cdot|o_k)$
end for
Incur terminal cost $d(b_0)$

4.3 Also, it is necessary to clarify how the authors "suitably simulating an oracle" in (4).

Response: We have now detailed a few ways of suitably simulating an oracle for obfuscation. Please note that our stochastic control approach is amenable to changes in how the oracle is simulated as long as the assumptions for the eavesdropper are satisfied. Following is the updated excerpt from the text in the obfuscation strategy section.

Changes To Text: The synthetic responses can be generated by suitably simulating an oracle; for e.g., the learner can train a neural network separately with an unbalanced subset of as was done in [5]. If the learner is sure that the eavesdropper has no public dataset to validate, the learner can take mirrored gradients with (**??**).

4.4 The number of successful gradient steps M is used as the queue state of the learner, which requires an exact value. However, in Theorem 1, it is only shown in an approximate format on the order of O(...). How do the authors compute M in the algorithm?

Response: Theorem 1 ensures a worst-case type complexity bound on M to ensure that the learner estimate is an ϵ critical point in expectation. This guarantees the existence of a finite state space for the MDP formulation. In case the lower bound f^* and the Lipschtiz constant are known (for example, many loss functions have a minimum possible value of 0 (for example, MSE), then the exact value of M can be computed using the step sizes with the following exact expression (from proof of Theorem 1 from [5]),

$$M = \max\left(\frac{4F\gamma}{\epsilon}, \frac{8F\gamma\sigma^2}{\epsilon^2}\right),$$

where $F = (\mathbb{E}f(x_0) - f^*)$. However, it is standard in practice to choose M heuristically or by tuning it as a hyper-parameter (and fixing it using the `max iter` argument in gradient algorithms).

In the MDP setup, we assume that the M is known either by exact computation or by heuristically choosing it and is finite, which is ensured by Theorem 1. Also, Theorem 1 helps us characterize the dependence of number of successful steps to achieve the learner objective in terms of the learning objective parameter ϵ. For the numerical experiments, by independent experimentation, we take $M = 50$ as the number of successful gradient steps that need to be taken.

We have further made this clear in the main text,

Changes To Text:

4.5 In the paragraph before (6), the authors mentioned that "denote the normalized probability vector for the transition of the learner state b given (o', o, u)." Is this a typo? I think it should be the transition given (o, b, u), no?

Response: Yes, the previous formulation was more convoluted, and we have now corrected it to the following and also checked for any other typos,

Changes To Text: Let $\mathcal{P}_b^{o,o'}(u) = \mathbb{P}((o',\cdot)|o,b,u)$ denote the normalized probability vector for the transition of the buffer state with a future oracle state o' given (o,b,u).

4.6 What is the expression of ψ_1 and ψ_2 in (7)? And what is the usage of them?

Response: We have added and improved the following explanation for ψ_2 and ψ_1, in the cost section of the MDP formulation. For completeness, we have included a slightly longer excerpt from the edited text,

Changes To Text: " We consider the following learning cost, which is proportional to the logarithm of the improvement in the eavesdropper's estimate ($\propto \log(\delta_n/\delta_{n+1})$) and is given by,

$$\begin{aligned} c_n(y_n,u_n) = \frac{\psi_1(b_n)}{\psi_2(o_n)} \log\left(\frac{I_n + {}^{i_n}/_{\delta_n}}{I_n + i_n}\right) \mathbb{1}(a_n) \\ + \frac{\psi_2(o_n)}{\psi_1(b_n)} \log\left(\frac{I_n}{I_n + i_n}\right)(1 - \mathbb{1}(a_n)) \end{aligned}, \tag{2}$$

where $\psi_1 : \mathcal{Y}^B \to \mathbb{R}^+$ and $\psi_2 : \mathcal{Y}^O \to \mathbb{R}^+$ are positive, convex and increasing cost functions, $I_n = \sum_{k=N-1}^{n+1} i_k$ is the sum of the previous incentives and δ_n is the eavesdropper's estimate of the trajectory $\mathcal{J}_1$ being the true trajectory computed using (**??**). ψ_1 and ψ_2 are used to incorporate the cost w.r.t. the state spaces, e.g., the functions ψ_1, and ψ_2 are considered quadratic in the respective states in the experiments. The form of the fractions ensures the structure as discussed next."

4.7 The definition of threshold structure is unclear. As t is the key property considered in this paper, the meaning of threshold structure should be stated or defined clearly to readers.

Response: In the introduction of the paper, we have added the following description of a threshold policy for an action space with cardinality 2. This addition intuitively explains how the policy has a threshold structure with respect to the number of informative learning steps. The complete mathematical form for the threshold incentivized policy (cardinality of action space > 2) is given in (14). We have also made the explanation of the implication of the threshold structure after the structural result of Theorem 2 more intuitive.

Changes To Text:

In Introduction: " The optimal policy π^* solving the MDP is shown to have a threshold structure (Theorem **??**) of the form,

$$\pi^*(b,o,n) = \begin{cases} a = 0\ (\texttt{obfuscate}), & b \leq \bar{b}(o,n) \\ a = 1\ (\texttt{learn}), & b > \bar{b}(o,n) \end{cases},$$

where b is the number of informative learning steps left, n is the number of queries left and $\bar{b}$ is the threshold function dependent on the oracle state o and n. Note that the exact dependence with the incentive is discussed later."

After Theorem 2: Theorem 2 implies that the policy is threshold in the learner queue state; hence, the learner learns more aggressively when the number of successful gradient steps (Def. 1) left is more. This intuitively makes sense from an obfuscation perspective since the learner should ideally spend more time obfuscating when it is closer to the minimizer (the queue state is small).

4.8 For experiments, as the algorithm is based on many assumptions, such as R1-R6, does the considered experiment satisfy all these assumptions considered in the paper?

Response: We have added an explanation in the Numerical Results sections detailing how the cost and transition probabilities function considered in the experiment satisfy the assumptions,

Changes To Text: "The functions ψ_1 and ψ_2 in (2) are quadratic in b and o, respectively. This satisfies assumptions R1, R2, R3. The emperical success probabilities along with the resulting cost function of (2) ensure that R4, R5, R6 are satisfied for $\alpha_{b,b',u} = \beta_{b,b',u} \leq 1$."

4.9 This work is based on [5], and some difference is listed in the last paragraph of Section I. It s then necessary to at least compare the algorithm of this work to [5] in the experiment to show the benefit of the introduced difference.

Response: We have added an experimental result in our previous row showcasing how a threshold policy with constant incentivization achieves a performance similar to the the optimal policy of this work, however has a much higher incentive utilization. Note that benchmarking the current approach as is with the algorithm of our previous work [5] since the previous setting does not include incentivization by the learner. The efore we fix the incentive as the maximum possible incentive for both the actions (learn and obfuscate).

We have updated the results table in the main text to reflect this change (we rerun the experiments with updated parameters and fixed seed and have updated the corresponding parameters in the manuscript),

Type of Policy	Learner Acc.	Eavesdropper Acc.	Incentive
Optimal Policy	87%	52%	242
Optimal Policy from [5]	87%	53%	270
Greedy Policy	89%	89%	300
Random Policy	48%	51%	214

Table 1: The optimal stationary policy with a threshold structure outperforms greedy policy by 37% on eavesdropper accuracy and random policy by 39% on learner accuracy.

4.10 This work uses two methods to approximate the optimal stationary policy: Simultaneous Perturbation Stochastic Approximation and Multi-armed bandit approach. Which one is used in the experiment? Bot methods should be tested in the experiments.

Response: We have mentioned in the Numerical Section manuscript that the stationary policy is approximated using the simultaneous perturbation stochastic approximation method.

As mentioned in the manuscript, the multi-armed bandit approach is computationally infeasible since it has an exponential dependency on the state space. For example, for the numerical experiment presented, the number of bandit arms is roughly $\sim (5)^{6\times 3} \approx 3 \times 10^{30}$ arms, which makes it infeasible for it to optimize using any of the known methods. The objective of putting the bandit algorithm

We have performed the comparison between the cost of the policies approximated by the SPSA and bandit algorithms on a toy example ($M = 10, |\mathcal{U}| = 2, |\mathcal{Y}_O| = 2$) and present it below as well for the convenience of the reviewer. Due to space restrictions, the example is in the supplementary document and is available online at supplementary document link.

Cost Function	LP	Stochastic Approximate		Multi Armed Bandit	
	Cost	Thresholds	Cost	Thresholds	Costs
		-0.49, 2.72		0, 3	
		-0.52, 6.59		0, 7	
		-0.72, 8.92		0, 9	

Table 2:

References

[1] C. Bersani, H. Dagdougui, C. Roncoli, and R. Sacile. Distributed Product Flow Control in a Network of Inventories With Stochastic Production and Demand. *IEEE Access*, 7:22486–22494, 2019.

[2] S. Bubeck, N. Cesa-Bianchi, and S. Bubeck. *Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems*. Now Publishers, 2012. Google-Books-ID: Rl2skwEACAAJ.

[3] G. Carnevale and G. Notarstefano. Nonconvex Distributed Optimization via Lasalle and Singular Perturbations. *IEEE Control Systems Letters*, 7:301–306, 2023.

[4] X. Chao, B. Yang, and Y. Xu. Dynamic inventory and pricing policy in a capacitated stochastic inventory system with fixed ordering cost. *Operations Research Letters*, 40(2):99–107, Mar. 2012.

[5] A. Jain and V. Krishnamurthy. Controlling Federated Learning for Covertness. *Transactions on Machine Learning Research*, 2024.

[6] V. Krishnamurthy. Interval dominance based structural results for Markov decision process. *Automatica*, 153:111024, July 2023.

[7] J. K.-H. Quah and B. Strulovici. Comparative Statics, Informativeness, and the Interval Dominance Order. *Econometrica*, 77(6):1949–1992, 2009. Publisher: Wiley, The Econometric Society.

[8] X. Shi, G. Wen, and X. Yu. Finite-Time Convergent Algorithms for Time-Varying Distributed Optimization. *IEEE Control Systems Letters*, 7:3223–3228, 2023.

[9] M. T. Toghani, S. Lee, and C. A. Uribe. PARS-Push: Personalized, Asynchronous and Robust Decentralized Optimization. *IEEE Control Systems Letters*, 7:361–366, 2023.

[10] J. N. Tsitsiklis, K. Xu, and Z. Xu. Private Sequential Learning. *Operations Research*, 69(5):1575–1590, Sept. 2021.

[11] L. Witt, M. Heyer, K. Toyoda, W. Samek, and D. Li. Decentral and Incentivized Federated Learning Frameworks: A Systematic Literature Review. *IEEE Internet of Things Journal*, 10(4):3642–3663, Feb. 2023.

[12] J. Xu, K. Xu, and D. Yang. Learner-Private Convex Optimization. *IEEE Transactions on Information Theory*, 69(1):528–547, Jan. 2023.

[13] Y. Zhan, J. Zhang, Z. Hong, L. Wu, P. Li, and S. Guo. A Survey of Incentive Mechanism Design for Federated Learning. *IEEE Transactions on Emerging Topics in Computing*, 10(2):1035–1044, Apr. 2022.

[14] D. Çelebi. Inventory control in a centralized distribution network using genetic algorithms: A case study. *Computers & Industrial Engineering*, 87:532–539, Sept. 2015.

5 Supplementary Document

5.1 Benchmark using Multi Armed Bandit Approach and Computational Infeasibility Beyond Toy Problems

Cost Function	LP		Stochastic Approximate		Multi Armed Bandit	
	Thresholds	Cost	Thresholds	Cost	Thresholds	Cost
	0, 3		-0.49, 2.72		0, 3	
	0, 7		-0.52, 6.59		0, 7	
	0, 9		-0.72, 8.92		0, 9	

5.2 Illustrating Stochastic Control Approach

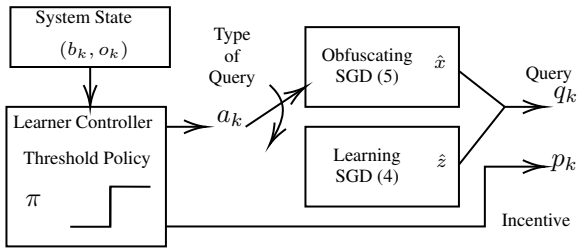

Figure 1: Extra Figure to demonstrate stochastic control for Covert Optimization: The learner uses π to dynamically switch between two stochastic gradient descents (1) and (**??**) and incentivize the oracle based on the queue state b_k and oracle state o_k. The query q_k is chosen as the last estimate from the selected SGD. The optimal policy minimizes the cumulative cost function of the finite-horizon MDP.

5.3 Convergence Curves of Learners and Eavesdropper