Sequential Counterfactual Risk Minimization

Houssam Zenati [1 2] **Eustache Diemert** [1] **Matthieu Martin** [1] **Julien Mairal** [2] **Pierre Gaillard** [2]

Abstract

Counterfactual Risk Minimization (CRM) is a framework for dealing with the logged bandit feedback problem, where the goal is to improve a logging policy using offline data. In this paper, we explore the case where it is possible to deploy learned policies multiple times and acquire new data. We extend the CRM principle and its theory to this scenario, which we call "Sequential Counterfactual Risk Minimization (SCRM)." We introduce a novel counterfactual estimator and identify conditions that can improve the performance of CRM in terms of excess risk and regret rates, by using an analysis similar to restart strategies in accelerated optimization methods. We also provide an empirical evaluation of our method in both discrete and continuous action settings, and demonstrate the benefits of multiple deployments of CRM.

1. Introduction

Counterfactual reasoning in the logged bandit problem has become a common task for practitioners in a wide range of applications such as recommender systems (Swaminathan & Joachims, 2015a), ad placements (Bottou et al., 2013) or precision medicine (Kallus & Zhou, 2018). Such a task typically consists in learning an optimal decision policy from logged contextual features and partial feedbacks induced by predictions from a logging policy. To do so, the logged data is originally obtained from a randomized data collection experiment. However, the success of counterfactual risk minimization is highly dependent on the quality of the logging policy and its ability to sample meaningful actions.

Counterfactual reasoning can be challenging due to large variance issues associated with counterfactual estimators (Swaminathan & Joachims, 2015b). Additionally, as pointed out by Bottou et al. (2013), confidence intervals obtained from counterfactual estimates may not be sufficiently accurate to select a final policy from offline data (Dai et al., 2020). This can occur when the logging policy does not sufficiently explore the action space. To address this, one option is to simply collect additional data from the same logging system to increase the sample size. However, it may be more efficient to use already collected data to design a better data collection experiment through a sequential design approach (Bottou et al., 2013, see Section 6.4). It is thus appealing to consider successive policy deployments when possible.

We tackle this sequential design problem and are interested in multiple deployments of the CRM setup of Swaminathan & Joachims (2015a), which we call sequential counterfactual risk minimization (SCRM). SCRM performs a sequence of data collection experiments by determining at each round a policy using data samples collected during previous experiments. The obtained policy is then deployed for the next round to collect additional samples. Such a sequential decision making system thus entails designing an adaptive learning strategy that minimizes the excess risk and expected regret of the learner. In contrast to the conservative learning strategy in CRM, the exploration induced by sequential deployments of enhanced logging policies should allow for improved excess risk and regret guarantees. Yet, obtaining such guarantees is nontrivial and we address it in this work.

In order to accomplish this, we first propose a new counterfactual estimator that controls the variance and analyze its convergence guarantees. Specifically, we obtain an improved dependence on the variance of importance weights between the optimal and logging policy. Second, leveraging this estimator and a weak assumption on the concentration of this variance term, we show how the error bound sequentially concentrates through CRM rollouts. This allows us to improve the excess risk bounds convergence rate as well as the regret rate. Our analysis employs methods similar to restart strategies in acceleration methods (Nesterov, 2012) and optimization for strongly convex functions (Boyd & Vandenberghe, 2004). We also conduct numerical experiments to demonstrate the effectiveness of our method in both discrete and continuous action settings, and how it improves upon CRM and other existing methods in the literature.

[1]Criteo AI Lab [2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France. Correspondence to: Houssam Zenati <housszenati@gmail.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

2. Related Work

Counterfactual learning from logged feedback (Bottou et al., 2013) uses only past interactions to learn a policy without interacting with the environment. Counterfactual risk minimization methods (Swaminathan & Joachims, 2015a;b) propose learning formulations using a variance penalization as in (Maurer & Pontil, 2009) to find policies with minimal variance. Even so, counterfactual methods remain prone to large variance issues (Dudík et al., 2014). These problems may arise when the logging policy under-explores the action space, making it difficult to use importance sampling tehcniques (Owen, 2013) that are key to counterfactual reasoning. While one could collect additional data to counter this problem, our method focuses on sequential deployments (Bottou et al., 2013, see Section 6.4) to collect data obtained from adaptive policies to explore the action space. Note also that the original motivation is related but different from the support deficiency problem (Sachdeva et al., 2020) where the support of the logging policy does not cover the support of the optimal policy.

Another related literature to our framework is batch bandit methods. Originally introduced by Perchet et al. (2015) and then extended by Gao et al. (2019) in the multi-arm setting, batch bandit agent take decisions and only observe feedback in batches. This therefore differs from the classic bandit setting (Auer et al., 2002; Audibert et al., 2007) where rewards are observed after each action taken by an agent. Extensions to the contextual case have been proposed by Han et al. and could easily be kernelized (Valko et al., 2013). The sequential counterfactual risk minimization problem is thus closely related to this setting. However, major differences can be noted. First, SCRM does not leverage any problem structure as in stochastic contextual bandits (Li et al., 2010) by assuming a linear reward function (Chu et al., 2011; Goldenshluger & Zeevi, 2013; Han et al.) nor uses regression oracles as (Foster & Rakhlin, 2020; Simchi-Levi & Xu, 2020). Second, deterministic decision rules taken by bandit agents (Lattimore & Szepesvari, 2019) do not allow for counterfactual reasoning or causal inference (Peters et al., 2017), unlike our framework which performs sequential randomized data collection. Third, unlike gradient based methods used in counterfactual methods with parametric policies, batch bandit methods use zero-order methods to learn from data and necessitate approximations to be scalable (Calandriello et al., 2020; Zenati et al., 2022).

The sequential designs that we use are adaptive data collection experiments, which have been studied by Bakshy et al. (2018); Kasy & Sautmann (2021). Closely related to our method is policy learning from adaptive data that has been studied by Zhan et al. (2021) and Bibaut et al. (2021) in the online setting. In contrast, we consider a batch setting and our analysis achieve fast rates in more general conditions. Zhan et al. (2021) use a doubly robust estimator and provide regret guarantees but assume a deterministic lower bound on the propensity score to control the variance. Instead, our novel counterfactual estimator does not require such an assumption. Bibaut et al. (2021) propose a novel maximal inequality and derive thereof fast rate regret guarantees under an additional margin condition that can only hold for finite action sets. Our work instead uses a different assumption on the expected risk, which is similar to Hölderian error bounds in acceleration methods (d'Aspremont et al., 2021) that are known to be satisfied for a broad class of subanalytic functions (Bolte et al., 2007).

In the reinforcement learning literature (Sutton & Barto, 1998), off-policy methods (Harutyunyan et al., 2016; Munos et al., 2016) evaluate and learn a policy using actions sampled from a behavior (logging) policy, which is therefore closely related to our setting. Among methods that have shown to be empirically successful are the PPO (Schulman et al., 2017) and TRPO (Schulman et al., 2015) algorithms which learn policies using a Kullback-Leibler distributional constraint to ensure robust learning, which can be compared to our learning strategy that improves the logging policy at each round. However reinforcement learning models transitions in the states (contexts) induced by the agent's actions while bandit problems like ours assume that actions do not influence the context distribution. This enables to design algorithms that exploit the problem structure, have theoretical guarantees and can achieve better performance in practice.

Finally, our method is related to acceleration methods (d'Aspremont et al., 2021) where current iterates are used as new initial points in the optimization of strongly convex functions (Boyd & Vandenberghe, 2004). While different schemes use fixed (Powell, 1977) or adaptive (Nocedal & Wright, 2006; Becker et al., 2011; Nesterov, 2012; Bolte et al., 2007; Gaillard & Wintenberger, 2018) strategies, our method differs in that it does not consider the same original setting, does not require the same assumptions nor provides the same guarantees. Eventually, while current models are also used as new starting points, additional data is effectively collected in our setting unlike those previous works that do not assume partial feedbacks as in our case.

3. Sequential Counterfactual Risk Minimization

In this section, we introduce the (CRM) framework and motivate the use of sequential designs for (SCRM).

Notations For random variables $x \sim \mathcal{P}_{\mathcal{X}}$, $a \sim \pi_\theta(\cdot|x)$ and $y \sim \mathcal{P}_{\mathcal{Y}}(\cdot|x,a)$, we write the expectation $\mathbb{E}_{x,\theta,y}[\cdot] = \mathbb{E}_{x\sim\mathcal{P}_{\mathcal{X}},a\sim\pi_\theta(\cdot|x),y\sim\mathcal{P}_{\mathcal{Y}}(\cdot|x,a)}[\cdot]$ and do the same for the variance $\mathrm{Var}_{x,\theta,y}$. Moreover, $\lesssim$ denotes approximate inequalities up to universal multiplicative terms.

3.1. Background

In the counterfactual risk minimization (CRM) problem, we are given n logged observations $(x_i, a_i, y_i)_{i=1,\dots,n}$ where contexts $x_i \in \mathcal{X}$ are sampled from a stochastic environment distribution $x_i \sim \mathcal{P}_{\mathcal{X}}$, actions $a_i \sim \pi_{\theta_0}(\cdot|x_i)$ are drawn from a logging policy π_{θ_0} with a model θ_0 in a parameter space Θ. The losses are drawn from a conditional distribution $y_i \sim \mathcal{P}_{\mathcal{Y}}(\cdot|x_i, a_i)$. We note $\pi_{0,i} = \pi_{\theta_0}(a_i|x_i)$ the associated propensities and assume them to be known. We will assume that the policies in $\pi_\theta, \theta \in \Theta$ admit densities so that the propensities will denote the density function of the logging policy on the actions given the contexts. The expected risk of a model θ is defined as:

$$L(\theta) = \mathbb{E}_{x,\theta,y}\left[y\right]. \tag{1}$$

Counterfactual reasoning uses the logged data sampled from the logging policy associated to θ_0 to estimate the risk of any model $\theta \in \Theta$ with importance sampling:

$$L(\theta) = \mathbb{E}_{x,\theta_0,y}\left[y\frac{\pi_\theta(a|x)}{\pi_{\theta_0}(a|x)}\right], \tag{2}$$

under the common support assumption (the support of π_θ support is included in the support of π_{θ_0}). The goal in CRM is to find a model $\theta \in \Theta$ with minimal risk by minimizing

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \mathcal{L}_0(\theta), \tag{3}$$

where $\mathcal{L}_0(\theta) = \hat{L}_0(\theta) + \lambda\sqrt{\frac{\hat{V}_0(\theta)}{n}}$ uses the sample variance penalization principle (Maurer & Pontil, 2009) on samples from θ_0 with counterfactual estimates of the expected risk $\hat{L}_0$, an empirical variance $\hat{V}_0$ and $\lambda > 0$. Specifically, in the (CRM) framework, multiple estimators are derived from the IPS method (Horvitz & Thompson, 1952) that uses the following clipped importance sampling estimator of the risk of a model θ by Bottou et al. (2013); Swaminathan & Joachims (2015a):

$$\hat{L}_0^{\text{IS}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} y_i \min\left(\frac{\pi_\theta(a_i|x_i)}{\pi_i}, \alpha\right), \tag{4}$$

where α is a clipping parameter. Writing $\chi_i(\theta) = y_i \min(\frac{\pi_\theta(a_i|x_i)}{\pi_{0,i}}, \alpha)$ and $\bar{\chi}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\chi_i(\theta)$ the empirical variance estimator becomes:

$$\hat{V}_0^{\text{IS}}(\theta) = \frac{1}{n-1}\sum_{i=1}^{n}(\chi_i(\theta) - \bar{\chi}(\theta))^2. \tag{5}$$

Other estimators aim at controlling the variance of the estimator with self-normalized estimators (Swaminathan & Joachims, 2015b) or with direct methods (Dudik et al., 2011; Dudík et al., 2014) in doubly robust estimators. Even so, the performance of counterfactual learning is harmed when the logging policy under-explores the action space (Owen, 2013). Likewise, counterfactual estimates obtained from a first round of randomized data collection may not suffice (Bottou et al., 2013) to select a model $\hat{\theta}$. In those cases, it could be natural to consider collecting additional samples. While it is possible to use the same logging model θ_0 to do so, we will present a framework for designing an improved sequential data collection strategy, following the intuition of sequential designs of Bottou et al. (2013).

3.2. Sequential Designs

In this section we present a design of data collections that sequentially learn a policy from logged data in order to deploy it and learn from the newly collected data. Specifically, we assume that at a round $m \in \{1, \dots M\}$, a model $\theta_m \in \Theta$ is deployed and a set s_m of n_m observations $s_m = (x_{m,i}, a_{m,i}, y_{m,i}, \pi_{m,i})_{i=1,\dots,n_m}$ is collected thereof, with propensities $\pi_{m,i} = \pi_{\theta_m}(a_{m,i}|x_{m,i})$ to learn a new model θ_{m+1} and reiterate. In this work, we assume that the loss y is bounded in $[-1, 0]$ as in (Swaminathan & Joachims, 2015a) (note however that this assumption could be relaxed to bounded losses) and follows a fixed distribution $\mathcal{P}_{\mathcal{Y}}$. Next, we will introduce useful definitions.

Definition 3.1 (Excess Risk and Expected Regret)**.** *Given an optimal model $\theta^* \in \arg\min_{\theta\in\Theta} L(\theta)$, we write for each rollout m the excess risk:*

$$\Delta_m = L(\theta_m) - L(\theta^*), \tag{6}$$

and define the expected regret as:

$$R_n = \sum_{m=0}^{M} \Delta_m n_{m+1}. \tag{7}$$

The objective is now to find a sequence of models $\{\theta_m\}_{m=1\dots M}$ that have an excess risk and an expected regret R_n that improve upon CRM guarantees. To do so, we define a sequence of minimization problems for $m \in \{1, \dots M\}$:

$$\hat{\theta}_{m+1} \in \arg\min_{\theta\in\Theta} \mathcal{L}_m(\theta), \tag{8}$$

where $\mathcal{L}_m$ is an objective function that we define in Section 4.2. Note that in the setting we consider, samples are i.i.d inside a rollout m but dependencies exist between different sets of observations. From a causal inference perspective (Peters et al., 2017), this does not incur an additional bias because of the successive conditioning on past observations. We provide detailed explanations in Appendix A.1 on this matter. Note also that the main intuition and motivation of our work is to shed light on how learning intermediate models θ_m to adaptively collect data can improve upon sampling from the same logging system by using the same total sample size $n = \sum_{i=0}^{m} n_m$. To illustrate the learning benefits of SCRM we now provide a simple example.

Algorithm 1 Sequential Counterfactual Risk Minimization

Input: Logged observations $(x_{0,i}, a_{0,i}, y_{0,i}, \pi_{0,i})_{i=1,\dots,n_0}$, parameter $\lambda > 0$
for $m = 1$ *to* M **do**
  Build $\mathcal{L}_m$ from observations s_m using Eq. (11)
  Learn θ_{m+1} using Eq. (8)
  Deploy the model θ_{m+1} and collect observations $s_{m+1} = (x_{m+1,i}, a_{m+1,i}, y_{m+1,i}, \pi_{m+1,i})_{i=1,\dots,n_{m+1}}$
end

Example 3.1 (Gaussian policies with quadratic loss)**.** *Let us consider Gaussian parametrized policies $\pi_\theta = \mathcal{N}(\theta, \sigma^2)$ and a loss $l_t(a) = (a - y_t)^2 - 1$ where $y_t \sim \mathcal{N}(\theta^*, \sigma^2)$. We illustrate in Figure 1 the evolution of the losses of learned models θ_m through 15 rollouts with either i) Batch CRM learning on aggregation of data, being generated by the unique initial logging policy θ_0 or ii) Sequential CRM learning with models $\theta_0, \dots, \theta_{m-1}$ deployed adaptively, with data being generated by the last learned model θ_{m-1} for the batch m. We see that the models learned with SCRM take larger optimization steps than the ones with CRM.*

We summarize our (SCRM) framework in Algorithm 1 with the different blocks exposed previously. We provide an additional graphical illustration of SCRM compared to CRM in Appendix A.1. In the next section we will define counterfactual estimators from the observations s_m at each round and define a learning strategy $\mathcal{L}_m$.

4. Variance-Dependent Convergence Guarantees

In this part we aim at providing convergence guarantees of counterfactual learning. We show how we can obtain a dependency of the excess risk on the variance of importance weights between the logging model and the optimal model.

4.1. Implicit exploration and controlled variance

We first introduce a new counterfactual estimator. For this, we will require a common support assumption as in importance sampling methods (Owen, 2013). We will assume that the policies π_θ for $\theta \in \Theta$ have all the same support. We then consider the following estimator of the risk of a model θ:

$$\hat{L}_m^{\text{IPS-IX}}(\theta) = \frac{1}{n_m} \sum_{i=1}^{n_m} \frac{\pi_{\theta,i}}{\pi_{m,i} + \alpha \pi_{\theta,i}} y_{m,i}, \qquad (9)$$

where $\pi_{\theta,i} = \pi_\theta(a_{m,i}|x_{m,i})$ and α is like a clipping parameter which ensures that the modified propensities $\pi_{m,i} + \alpha\pi_\theta(a_{m,i}|x_{m,i})$ are lower bounded. Noting $\zeta_i(\theta) = \left(\frac{\pi_{\theta,i}}{\pi_{m,i}+\alpha\pi_{\theta,i}} - 1\right) y_{m,i}$, $\bar{\zeta}(\theta) = \frac{1}{n_m}\sum_{i=1}^{n_m} \zeta_i(\theta)$ we can write the empirical variance estimator as:

$$\hat{V}_m^{\text{IPS-IX}}(\theta) = \frac{1}{n_m - 1} \sum_{i=1}^{n_m} (\zeta_i(\theta) - \bar{\zeta}(\theta))^2. \qquad (10)$$

Here, the empirical variance uses a control variate since it uses the expression of $\zeta_i(\theta)$ above instead of $y_{m,i}\frac{\pi_{\theta,i}}{\pi_{m,i}+\alpha\pi_{\theta,i}}$. This allows to improve the dependency on the variance in the excess risk provided in Proposition 4.2. Note also that our estimator resembles the implicit exploration estimator in the EXP3-IX algorithm (Lattimore & Szepesvari, 2019), as our motivation is to improve the control of the variance.

4.2. Learning strategy

Next, we aim in this part to provide a learning objective strategy $\mathcal{L}_m$, as referred to in Eq. (8). Our approach, like the (CRM) framework, uses the sample variance penalization principle (Maurer & Pontil, 2009) to learn models that have low expected risk with high probability. To do so, we first provide an assumption to be used in our generalization error bound.

Assumption 4.1 (Bounded importance weights)**.** *For any models $\theta, \theta' \in \Theta$ and any $(x, a) \in \mathcal{X} \times \mathcal{A}$, we assume $\pi_\theta(a|x)/\pi_{\theta'}(a|x) \leq W$, for some $W > 0$.*

This assumption has been made in previous works (Kallus & Zhou, 2018; Zenati et al., 2020) and is reasonable when we consider a bounded parameter space Θ. Next, we state an error bound for our estimator.

Proposition 4.1 (Generalization Error Bound)**.** *Let $\hat{L}_m^{\text{IPS-IX}}$ and $\hat{V}_m^{\text{IPS-IX}}$ be the empirical estimators defined respectively in Eq. (9) and Eq. (10). Let $\theta \in \Theta$, $\delta \in (0, 1)$, and $n_m \geq 2$. Then, under Ass. 4.1, for $\lambda_m = \sqrt{18(C_m(\Theta) + \log(2/\delta))}$, with probability at least $1 - \delta$:*

$$L(\theta) \leq \hat{L}_m^{\text{IPS-IX}}(\theta) + \lambda_m \sqrt{\frac{\hat{V}_m^{\text{IPS-IX}}(\theta)}{n_m}} + \frac{2\lambda_m^2 W}{n_m} + \delta_m,$$

where $C_m(\Theta)$ is a metric entropy complexity measure defined in App. B.1 and $\delta_m = \sqrt{\log(2/\delta)/(2n_m)}$.

This Proposition is proved in Appendix B.2 and essentially uses empirical bounds (Maurer & Pontil, 2009). By minimizing the latter high-probability upper bound, we can find models θ with guarantees of minimizing the expected risk. Therefore, at each round, we minimize the following loss:

$$\mathcal{L}_m(\theta) = \hat{L}_m^{\text{IPS-IX}}(\theta) + \lambda_m \sqrt{\frac{\hat{V}_m^{\text{IPS-IX}}(\theta)}{n_m}}, \qquad (11)$$

where $\lambda_m > 0$ is a positive parameter. Unlike deterministic decision rules used for example in UCB-based algorithms (Lattimore & Szepesvari, 2019), the exploration is naturally guaranteed by the stochasticity of the policies we use.

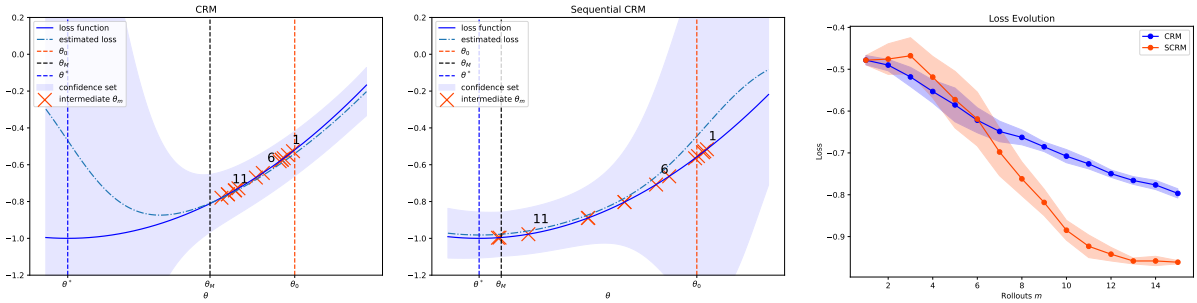

Figure 1: Comparison of CRM and SCRM on a simple setting described in Example 3.1. The models learned through CRM using re-deployments of θ_0 (left) reach θ^* slower than SCRM (center) that uses intermediate deployments $\theta_1, \dots, \theta_M$ indicated with 'x' markers and rollout numbers. The comparison of the evolution of averaged losses (right) over 10 random runs also shows SCRM converges faster. Here $\theta^* = 1$, $\sigma = 0.3$ and we take $M = 15$ total rollouts with batches m of size $n_m = 100 \times 2^m$. The parameter λ is set to its theoretical value.

4.3. Excess risk upper bound

Eventually, we establish an upper bound on the excess risk of the IPS-IX estimator for counterfactual risk minimization using the learning strategy that we just defined. For this, we require an assumption on the complexity measure.

Assumption 4.2. *We assume that the set Θ is compact and that there exists $d > 0$ such that $C_m(\Theta) \leq d \log(n_m)$.*

This assumption states that the complexity grows logarithmically with the sample size. It holds for parametric policies so long as the propensities are lower bounded, which is verified using our estimator. We now state our variance-dependent excess risk bound.

Proposition 4.2 (Excess Risk Bound)**.** *Let $n_m \geq 1$ and $\theta_m \in \Theta$. Let s_m be a set of n_m samples collected with policy π_{θ_m}. Then, under Assumptions 4.1 and 4.2, a minimizer θ_{m+1} of Eq. (11) on the samples s_m satisfies the excess risk upper-bound: w.p. $1-\delta$*

$$\Delta_{m+1} = L(\theta_{m+1}) - L(\theta^*) \lesssim \sqrt{\nu_m^2 \frac{d \log n_m - \log \delta}{n_m}} + \frac{W^2 + W(d \log n_m - \log \delta)}{n_m},$$

where $\nu_m^2 = \mathrm{Var}_{x,\theta_m}\left(\frac{\pi_{\theta^*}(a|x)}{\pi_{\theta_m}(a|x)}\right)$.

The proof is postponed to Appendix B.2. The modified propensities in IPS-IX as well as the control variate used in the variance estimator allow us to improve the dependency in ν_m^2, compared to ν_m^2+1 obtained in previous work (Zenati et al., 2020). This turns out to be a crucial point to use these error bounds sequentially as in acceleration methods since $\nu_m \to 0$ if $\theta_m \to \theta^*$, as explained in the next section.

5. SCRM Analysis

In this section we provide the main theoretical result of this work on the excess risk and regret analysis of SCRM. We start by stating an assumption that is common in acceleration methods (d'Aspremont et al., 2021) with restart strategies (Becker et al., 2011; Nesterov, 2012) that we will require to achieve the benefits of sequential designs.

Assumption 5.1 (Hölderian Error Bound)**.** *We assume that there exist $\gamma > 0$ and $\beta > 0$ such that for any $\theta \in \Theta$, there exists $\theta^* \in \arg\min_{\theta \in \Theta} L(\theta)$ such that*

$$\gamma \mathrm{Var}_{x,\theta}\left(\frac{\pi_{\theta^*}(x|a)}{\pi_\theta(x|a)}\right) \leq (L(\theta) - L(\theta^*))^\beta .$$

Typically, in acceleration methods, Hölderian error bounds (Bolte et al., 2007) are of the form:

$$\gamma d(\theta, S_\Theta^*) \leq (L(\theta) - L(\theta^*))^\beta$$

for some $\gamma, \beta > 0$ and where $d(\theta, S_\Theta^*)$ is some distance to the optimal set ($S_\Theta^* = \arg\min_{\theta \in \Theta} L(\theta)$). This bound is akin to a local version of strong convexity ($\beta = 1$) or a bounded parameter space ($\beta = 0$) if d is the Euclidean distance. When $\beta \in [0, 1]$, this has also been referred to as the Łojasiewicz assumption introduced in (Łojasiewicz, 1963; 1993). Notably, it has been used in online learning (Gaillard & Wintenberger, 2018) to obtain fast rates with restart strategies. This assumption holds for instance for Example 3.1 with $\beta = 1$ (see App C.1). We also discuss this assumption for distributions in the exponential family in Appendix C.2 notably for distributions that have been used practice (Swaminathan & Joachims, 2015b; Kallus & Zhou, 2018; Zenati et al., 2020). Next we state our main result that is the acceleration of the excess risk convergence rate and the regret upper bound of SCRM.

Proposition 5.1. *Let $n_0, n \geq 2$ and $\theta^* \in \arg\min_\theta L(\theta)$. Let $n_m = n_0 2^m$ for $m = 0, \dots, M = \lfloor \log_2(1 + \frac{n}{n_0}) \rfloor$. Then, under Assumptions 4.1, 4.2 and 5.1 with $\beta > 0$, the SCRM procedure (Alg. 1) satisfies the excess risk upper-bound*

$$\Delta_M = L(\theta_M) - L(\theta^*) \leq O\Big(n^{-\frac{1}{2-\beta}} \log n\Big).$$

Moreover, the expected regret is bounded as follows:

$$R_n = \sum_{m=0}^{M} \Delta_m n_{m+1} \leq O\Big(n^{\frac{1-\beta}{2-\beta}} \log(n)^2\Big).$$

The proof of our result is detailed in Appendix B.3.

Discussion This result illustrates that an excess risk of order $O\big(\frac{\log(n)}{n}\big)$ may be obtained when $\beta = 1$ (which is implied by a local version of strong convexity assumption in acceleration methods). When $\beta = 0$, which merely accounts that the variance of importance weights are bounded, we simply recover the original rate of CRM of order $O(\log(n)/\sqrt{n})$. The SCRM procedures thus improves the excess risk rate whenever $\beta > 0$. It is worth to emphasize that the knowledge of β is not needed by Alg. 1. We also note that our assumption seems related to the Bernstein condition (Bartlett & Mendelson, 2006, see Def 2.6), and (van Erven et al., 2015, see Def 5.1) that bounds a variance term by an excess risk term to the power. In empirical risk minimization, this implies the same excess risk rate and regret rate (van Erven & Koolen, 2016), which are exactly the same rates as ours (up to logs).

6. Empirical Evaluation

In this section we perform numerical experiments to validate our method in practical settings. We present the experimental setup as well as experiments comparing SCRM to related approaches and internal details of the method.

6.1. Experimental setup

As our method is able to handle both discrete and continuous actions we experiment in both settings. We now provide a brief description of the setups, with extensive details available in Appendix D.2. [1]

Continuous actions We perform evaluation on synthetic problems pertaining to personalized pricing problems from (Demirer et al., 2019) (*Pricing*) and advertising from (Zenati et al., 2020) (*Advertising*). We consider Gaussian policies $\pi_\theta(\cdot|x) = \mathcal{N}(\mu_\theta(x), \sigma^2)$ with linear contextual parametrization $\mu_\theta(x) = \theta^\top x$ and fixed variance σ^2 that corresponds to the exploration budget allowed in the original randomized experiment. The features are up to 10 dimensions and the actions are one-dimensional. We keep the original logging baselines from the settings and compare results to a skyline supervised model trained on the whole training data with full information.

Discrete actions We adapt the setup of (Swaminathan & Joachims, 2015a) that transforms a multilabel classification task into a contextual bandit problem with discrete, combinatorial action space. We keep the original modeling (akin to CRF) with categorical policies $\pi_\theta(a|x) \propto \exp(\theta^\top(x \bigotimes a))$. The baseline (resp. skyline) is a supervised, full information model with identical parameter space than CRM methods trained on 5% (resp. 100%) of the training data. We consider the class of probabilistic policies that satisfy Assumption 5.1 by predicting actions in an Epsilon Greedy fashion (Sutton & Barto, 1998)): $\pi_\theta^\varepsilon(a, x) = (1 - \varepsilon)\pi_\theta(a, x) + \varepsilon/|\mathcal{A}|$ where $\varepsilon = .1$. Real-world datasets include *Scene*, *Yeast* and *TMC2007* with feature space up to 30,438 dimensions and action space up to 2^{22}. To account for this combinatorial action space we allow a model θ_m to be learned using data from all past rollouts $\{s_l\}_{l<m}$ for better sample efficiency and therefore adjust variance estimation in Appendix A.2 to take into account sequential dependencies.

6.2. SCRM compared to CRM and related methods

We first compare SCRM to CRM and existing methods in the literature.

Comparison between SCRM and CRM First, we provide insights on the performance that SCRM can achieve compared to classical CRM with increasing sample sizes. The key difference between CRM/SCRM is that for each sample size n_m CRM learns from samples generated by the logging model $s_m^{CRM} \leftarrow \theta_0$ (see Alg. 2) whilst SCRM learns from samples generated by a series of optimized models $s_m^{SCRM} \leftarrow \theta_m$ (see Alg. 1). For each sample size we select a posteriori the best λ for both methods based on test set loss value. We report in Figure 2 over $M = 10$ rollouts the mean test loss depending on sample size up to 2^{10}, with standard deviation estimated over 10 random runs. We observe that SCRM converges very fast, often within the first rollouts. Conversely, CRM needs more samples and the variance is higher. We conclude that there is a striking benefit to use a sequential design in order to achieve near optimal loss with much fewer samples and better confidence compared to CRM. Complementary results on other datasets are available in Appendix E.1.

Moreover, to further illustrate this benefit of efficient learning we also report in Table 1 the sample size needed to

[1] All the code to reproduce the empirical results is available at: https://github.com/criteo-research/sequential-conterfactual-risk-minimization

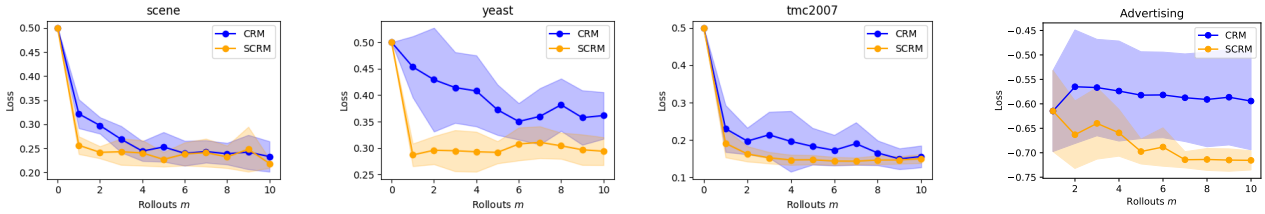

Figure 2: Test loss as a function of sample size on *Scene*, *Yeast*, *TMC2007*, *Advertising*, (from left to right). SCRM (in orange) converges faster and with less variance than CRM (in blue).

Percentage p	0.7	0.8	0.9
CRM	100×2^{10}	100×2^{16}	$> 100 \times 2^{22}$
SCRM (ours)	100×2^{8}	100×2^{9}	100×2^{11}

Table 1: Needed sample size to achieve test loss $L(\theta) \leq p * L(\theta^*)$ on the setting in Example 3.1 over the average of 10 random runs. SCRM needs way less data to converge to near optimal solution. λ is set to its theoretical value.

	Pricing	*Advertising*	*Yeast*	*TMC2007*
λ'	$-5.353 \pm .178$	$-.716 \pm .020$	$.294 \pm .026$	$.146 \pm .012$
$\hat{\lambda}$	$-5.575 \pm .036$	$-.726 \pm .001$	$.299 \pm .039$	$.164 \pm .021$

Table 2: Test loss after 10 rollouts when choosing λ by a posteriori selection (λ') or with proposed heuristic ($\hat{\lambda}$). Our heuristic is competitive with the a posteriori selection of a fixed λ'.

attain near optimal performance when θ^* is known as in Example 3.1, where we also observe that SCRM reaches optimal performances faster than CRM. This corroborates the benefits of improved excess risk rates for SCRM.

Hyper-parameter selection for SCRM In our experiments, hyperparameter selection consists in choosing a value for λ. We describe a simple heuristic and evaluate its performance on different datasets. We propose to select $\hat{\lambda}_m$ by estimating the non-penalized CRM loss (eq. 3) using offline cross-validation on past data $s_{t<m}$. We report in Table 2 the test loss obtained when choosing a fixed λ a posteriori (λ') or with this heuristic ($\hat{\lambda}$). We observe that loss confidence intervals for both methods intersect for all discrete datasets, except on *TMC2007* where the degradation shows only at the 3rd digit. On continuous datasets, the heuristic actually improves upon the fixed a posteriori selection. We conclude that this heuristic is usable in practice.

Comparison with other methods In this paragraph we compare our SCRM to related methods to explore practical implications of existing methods in our setting. We first consider batch bandits methods and implement the stochastic sequential batch pure exploitation (SBPE) algorithm in (Han et al.) and a batch version of kernel UCB (Valko et al., 2013) algorithm (BKUCB) with an optimized library (see implementations details in Appendix D.3). We also experiment with off-policy RL methods PPO (Schulman et al., 2017) and TRPO (Schulman et al., 2015) from the Stable-Baselines library (Raffin et al., 2021) (see Appendix D.3). Indeed, such methods model more general state transitions based on past actions, but they could be used in our setting. To fairly compare all methods (in particular those for which no heuristic existing for hyper-parameter selection) we report the mean and standard deviation over 10 random runs of the best test loss a posteriori over hyperparameter grids of the same size. First, we observe that SCRM beats CRM on all datasets, illustrating the benefit of the sequential design. Second, on discrete tasks (where we the combinatorial action space is large) we observe that SCRM achieves nearly the best test loss in all tasks, while RL methods have difficulties maintaining good performances. Third, batch bandits algorithms can achieve good performances in practice because of their deterministic decision rules. However, they involve an $O(n^3)$ matrix inversion and therefore did not finish (DNF) in 24h (per single run) on a 46 CPU / 500G RAM machine in most of our settings with large sample size n, which make them unpractical for large scale experiments. We conclude that SCRM is an effective learning paradigm and that it scales successfully on a variety of settings.

6.3. Details on SCRM

Next, we provide additional empirical evaluations of details of our method.

Evaluation of IPS-IX To understand the bias-variance trade-off that IPS-IX can achieve in practice compared to other counterfactual estimators we consider a policy evaluation experiment. The task we consider uses sinusoidal losses $y(a) = \cos(a)$ and evaluated policies are shifted Gaussians $\{\pi_i = \mathcal{N}(i * \pi/4, 1)\}_{i=0,4}$, with π_0 being the logging policy. Evaluated policies with large shifts with π_0 therefore simulate the setting where the logging policy under-explores the action space. The estimators we consider include IPS, SNIPS (Swaminathan & Joachims, 2015b), clipped IPS (eq. 4) with heuristic from (Bottou et al., 2013) and IPS-IX (eq. 9) with $\alpha = 1/n$. All methods therefore use

$n/\lvert\mathcal{A}\rvert/dim(\mathcal{X})\rvert$	*Pricing* $10^5/\infty/10$	*Advertising* $10^5/\infty/2$	*Scene* $2.10^3/2^6/295$	*Yeast* $2.10^3/2^{14}/104$	*TMC2007* $3.10^4/2^{22}/3.10^4$
Baseline	$-3.414 \pm .162$	$-.431 \pm .120$	$.353 \pm .009$	$.478 \pm .014$	$.511 \pm .003$
SBPE	DNF	DNF	**.179** $\pm .001$	$.302 \pm .003$	DNF
BKUCB	DNF	DNF	$.236 \pm .014$	$.303 \pm .004$	DNF
TRPO	**-5.750** $\pm .020$	$-.670 \pm .030$	$.376 \pm .001$	$.434 \pm .001$	$.396 \pm .001$
PPO	$-5.274 \pm .200$	$-.637 \pm .015$	$.206 \pm .001$	$.463 \pm .001$	$.263 \pm .001$
CRM	$-5.325 \pm .068$	$-.594 \pm .100$	$.233 \pm .031$	$.362 \pm .044$	$.158 \pm .034$
SCRM (ours)	$-5.575 \pm .036$	**-.726** $\pm .020$	$.219 \pm .009$	**.294** $\pm .026$	**.146** $\pm .012$
Skyline	$-5.830 \pm .020$	$-.739 \pm .002$	$.179 \pm .002$	$.312 \pm .003$	$.142 \pm .001$

Table 3: Test loss $\pm$ stddev of different methods after 10 rollouts. SCRM achieves optimal or near optimal performance in all datasets. Batch bandit methods did not finish (DNF) on large scale settings, and RL methods perform overall poorly on discrete settings with large action space.

their respective heuristics to set hyperparameters. We report in Figure 3 the bias and variance of estimators for each shift $\mu_0 - \mu = i * \pi/4$ for $i = 0, \ldots, 4$. We observe that IPS-IX shows an empirical bias comparable to IPS, lower than SNIPS and clipped IPS while maintaining a lower variance. Moreover its variance is only slightly higher than clipped IPS which introduced a large bias. We conclude that besides being a key component of our analysis IPS-IX also controls the variance with a better trade-off in practice. More details are available in Appendix E.2.

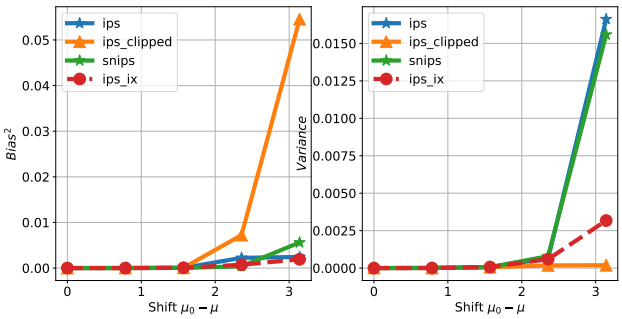

Figure 3: Comparison of counterfactual estimators on policy evaluation. Bias (left), Variance (right). IPS-IX shows a low bias and compares favorably to IPS and SNIPS in terms of variance.

When is SCRM useful is a natural question of interest when choosing the method to be used on a given logged bandit feedback problem. Intuitively one can imagine that SCRM will be most useful when the logging policy under-explores the action space, for example when the distance (in parameter space) between the logging and optimal parameters is large. To study this question we proceed to the following experiment on the setup of Example 3.1 with Gaussian distributions $\mathcal{N}(\theta, \sigma)$ and fixed loss variance $\sigma^* = \mathrm{Var}_y(y)$. We vary the distance $\delta_0 = \|\theta^* - \theta_0\|$ between the optimal

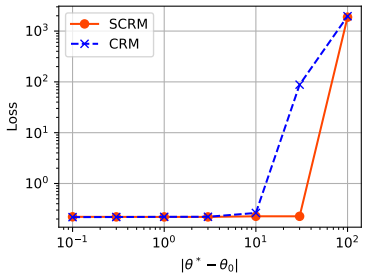

Figure 4: Best final loss when varying $\delta_0 = \|\theta^* - \theta_0\|$. SCRM achieves better losses especially for larger δ_0.

model θ^* and the logging model θ_0. Since the ideal exploration level may be task dependent we choose a posteriori the best σ on a grid, for both CRM and SCRM. We report in Figure 4 the best final loss for both CRM and SCRM for a range of values of δ_0. We observe in particular that SCRM achieves better final losses for larger distances δ_0 than CRM. With the same number of rollouts M, SCRM can extend the exploration to further areas while CRM fails for any exploration level in those cases, which advocates for using sequential deployments.

7. Discussions

In this work, we have proposed a method to extend the CRM perspective for designing sequential data collection experiments. We have introduced a novel counterfactual estimator to improve variance control in excess risk bounds. Under a weak error bound assumption, we have sequentially applied these excess risk guarantees to achieve faster rates similarly to acceleration methods. Our method also improves upon CRM in practice and is particularly well-suited for this setting compared to existing methods in the literature. It is worth noting that, in order to avoid introducing dependencies in the excess risk bounds we analyzed, the theoretical

algorithm we have studied uses geometric sample sizes to discard previous samples. However, using all past samples has been found to be also effective in practice and developing guarantees for this case would be an interesting area for future research. Additionally, similar to online settings that involve an exploration-exploitation tradeoff, investigating the use of optimism in the face of uncertainty (OFUL) principle in SCRM would also be a promising avenue for future work.

Acknowledgements

The authors thank Alberto Bietti for the insightful early discussions on this project. The authors also thank the reviewers for their feedback on this paper. This work was supported by ANR 3IA MIAI@Grenoble-Alpes (ANR-19-P3IA0003).

References

Audibert, J.-Y., Munos, R., and Szepesvari, C. Tuning bandit algorithms in stochastic environments. In *International Conference on Algorithmic Learning Theory*, 2007. 2

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 05 2002. 2

Bakshy, E., Dworkin, L., Karrer, B., Kashin, K., Letham, B., Murthy, A., and Singh, S. Ae: A domain-agnostic platform for adaptive experimentation. 2018. 2

Bartlett, P. L. and Mendelson, S. Empirical minimization. *Probability Theory and Related Fields*, 2006. 6

Becker, S. R., Candès, E. J., and Grant, M. C. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3 (3):165–218, jul 2011. 2, 5

Bibaut, A., Kallus, N., Dimakopoulou, M., Chambaz, A., and van der Laan, M. Risk minimization from adaptively collected data: Guarantees for supervised and policy learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 19261–19273. Curran Associates, Inc., 2021. 2

Bolte, J., Daniilidis, A., and Lewis, A. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007. 2, 5

Bottou, L., Peters, J., Quiñonero Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research (JMLR)*, 14(1):3207–3260, 2013. 1, 2, 3, 7, 24

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004. 1, 2

Calandriello, D., Carratino, L., Lazaric, A., Valko, M., and Rosasco, L. Near-linear time Gaussian process optimization with adaptive batching and resparsification. In *International Conference on Machine Learning (ICML)*, 2020. 2

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In Gordon, G., Dunson, D., and Dudík, M. (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. 2

Dai, B., Nachum, O., Chow, Y., Li, L., Szepesvári, C., and Schuurmans, D. Coindice: Off-policy confidence interval estimation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. 1

Demirer, M., Syrgkanis, V., Lewis, G., and Chernozhukov, V. Semi-parametric efficient policy learning with continuous actions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 6, 23

Dudik, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning (ICML)*, 2011. 3

Dudík, M., Erhan, D., Langford, J., and Li, L. Doubly robust policy evaluation and optimization. *Statist. Sci.*, 29(4):485–511, 11 2014. doi: 10.1214/14-STS500. 2, 3

d'Aspremont, A., Scieur, D., and Taylor, A. Acceleration methods. *Foundations and Trends® in Optimization*, 5 (1-2):1–245, 2021. ISSN 2167-3888. 2, 5, 22

Faury, L., Tanielian, U., Dohmatob, E., Smirnova, E., and Vasile, F. Distributionally Robust Counterfactual Risk Minimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3850–3857, Apr. 2020. 22

Foster, D. and Rakhlin, A. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3199–3210. PMLR, 13–18 Jul 2020. 2

Gaillard, P. and Wintenberger, O. Efficient online algorithms for fast-rate regret bounds under sparsity. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 2, 5

Gao, Z., Han, Y., Ren, Z., and Zhou, Z. Batched multi-armed bandits problem. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019. 2

Goldenshluger, A. and Zeevi, A. A linear response bandit problem. *Stochastic Systems*, 3(1):230 – 261, 2013. 2

Han, Y., Zhou, Z., Zhou, Z., Blanchet, J. H., Glynn, P. W., and Ye, Y. Sequential batch learning in finite-action linear contextual bandits. doi: 10.48550/ARXIV.2004.06321. 2, 7, 25

Harutyunyan, A., Bellemare, M. G., Stepleton, T., and Munos, R. Q(λ) with off-policy corrections. In Ortner, R., Simon, H. U., and Zilles, S. (eds.), *Algorithmic Learning Theory*, pp. 305–320, Cham, 2016. Springer International Publishing. 2

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. 3

Kallus, N. and Zhou, A. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018. 1, 4, 5, 22, 23

Kasy, M. and Sautmann, A. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1): 113–132, 2021. 2

Lattimore, T. and Szepesvari, C. *Bandit Algorithms*. 2019. 2, 4, 25

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, New York, NY, USA, 2010. ACM. 2

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. In *Conference on Learning Theory (COLT)*, 2009. 2, 3, 4, 15, 16

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. G. Safe and efficient off-policy reinforcement learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 1054–1062, Red Hook, NY, USA, 2016. Curran Associates Inc. 2

Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140:125 – 161, 2012. 1, 2, 5

Nocedal, J. and Wright, S. J. *Numerical Optimization*. Springer, New York, NY, USA, 2e edition, 2006. 2

Owen, A. B. *Monte Carlo theory, methods and examples*. 2013. 2, 3, 4, 13, 14

Perchet, V., Rigollet, P., Chassang, S., and Snowberg, E. Batched bandit problems. *The Annals of Statistics*, 44, 05 2015. 2

Peters, J., Janzing, D., and Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017. 2, 3, 12

Powell, M. J. Restart procedures for the conjugate gradient method. *Math. Program.*, 12(1):241–254, dec 1977. 2

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. 7, 25

Sachdeva, N., Su, Y., and Joachims, T. Off-policy bandits with deficient support. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '20, pp. 965–975, New York, NY, USA, 2020. Association for Computing Machinery. 2

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., and Abbeel, P. Trust region policy optimization. 2015. doi: 10.48550/ARXIV.1502.05477. 2, 7, 25

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. 2017. doi: 10.48550/ARXIV.1707.06347. 2, 7, 25

Simchi-Levi, D. and Xu, Y. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *SSRN Electronic Journal*, 2020. ISSN 1556-5068. doi: 10.2139/ssrn.3562765. 2

Sutton, R. S. and Barto, A. G. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981. 2, 6, 24

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning (ICML)*, 2015a. 1, 2, 3, 6, 12, 21, 22, 24

Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems (NIPS)*. 2015b. 1, 2, 3, 5, 7

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. Finite-Time Analysis of Kernelised Contextual Bandits. In *Uncertainty in Artificial Intelligence*, Bellevue, United States, July 2013. 2, 7, 25

van Erven, T. and Koolen, W. M. Metagrad: Multiple learning rates in online learning. In *Advances in Neural Information Processing Systems*, 2016. 6

van Erven, T., Grünwald, P. D., Mehta, N. A., Reid, M. D., and Williamson, R. C. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 2015. 6

Zenati, H., Bietti, A., Martin, M., Diemert, E., Gaillard, P., and Mairal, J. Counterfactual learning of stochastic policies with continuous actions: from models to offline evaluation. 2020. doi: 10.48550/ARXIV.2004.11722. 4, 5, 6, 22, 23, 24

Zenati, H., Bietti, A., Diemert, E., Mairal, J., Martin, M., and Gaillard, P. Efficient kernelized ucb for contextual bandits. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 5689–5720. PMLR, 28–30 Mar 2022. 2

Zhan, R., Ren, Z., Athey, S., and Zhou, Z. Policy learning with adaptively collected data. 2021. doi: 10.48550/ARXIV.2105.02344. 2

Łojasiewicz, S. *Une propriété topologique des sous ensembles analytiques réels*. Les équations aux dérivées partielles, 1963. 5

Łojasiewicz, S. Sur la géométrie semi et sous analytique. *Annales de l'institut Fourier*, 43(5):1575–1595, 1993. ISSN 2167-3888. 5

This appendix is organized as follows: in Appendix A, we provide additional explanations on counterfactual methods related to our approach. In Appendix B, we detail our analysis of our counterfactual estimator as well as the general SCRM procedure, as given in Alg. 1. Next, in Appendix D we present all the details of the empirical evaluation and eventually in Appendix E we provide all additional empirical results that were omitted from the main paper due to space limitation.

A. Additional details on counterfactual estimators

A.1. Unconfoundedness in sequential designs

In these explanations, we recall that the distributions of contexts as well as the distribution of losses are fixed. In other words, the latter do not vary from one batch to another. In the counterfactual risk minimization framework (CRM) (Swaminathan & Joachims, 2015a), the causal graph (using the conventions in (Peters et al., 2017)) can be represented as shown in Figure 5.

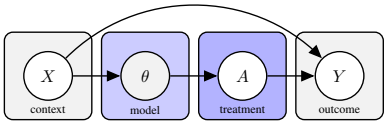

Figure 5: Causal Graph in a randomized data collection experiment. A denotes action (or treatment), X context, Y is the loss (or outcome). The causal influence of the contexts on actions is done through the model θ.

In the sequential counterfactual risk minimization (SCRM) framework, if we unfold the causal graph, the following representation can be given in Figure 6.

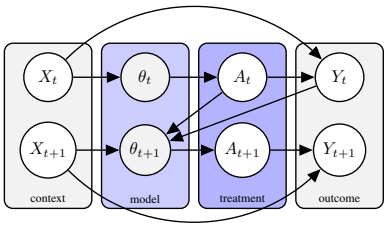

Figure 6: Causal Graph in a sequential randomized data collection experiment. A denotes action (or treatment), X context, Y is the loss (or outcome). The contextual treatments are taken through the models θ_t.

Therefore, it is clear that in general, $\theta_t \not\perp\!\!\!\perp \theta_{t+1}$. However, from d-separation and faithfullness (Peters et al., 2017), we have for $t' < t$:

$$\theta_t \perp\!\!\!\perp \theta_{t'} | \theta_{t-1}.$$

Therefore, given that all the dependencies are observed and that we can condition on the direct parents of a given model θ_t, sequential randomized data collection are possible. We eventually provide in Figure 7 an illustration of SCRM and CRM.

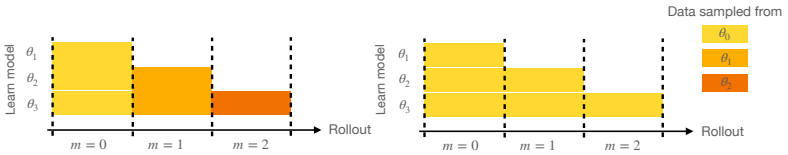

Figure 7: Graphical illustration of SCRM setup (left) and CRM (right), learned with same amount of data after each batch m. The training data are displayed with color block and the policy used to sample actions in these block are either adaptive (SCRM) or using the loggind model θ_0 (CRM).

A.2. Multiple Importance Sampling Estimators

Note that in order to avoid introducing dependencies in the excess risk bounds we analyzed, the theoretical algorithm we have studied uses geometric sample sizes to discard previous samples. However, using all past samples is effective in practice and developing guarantees for this case would be an interesting area for future research. We present in this section a estimators using aggregation of all previous information. In particular, we can use Multiple Importance Sampling (MIS) (Owen, 2013) over all previous samples. Consider in particular a partition of unity with $m > 1$ weight functions $\omega_t(a) > 0$ which satisfies $\sum_{t=0}^{m} \omega_{t,m}(a) = 1$ for all a and $m \in \{0, \dots M\}$. The MIS estimator writes:

$$\hat{L}_m^{\text{MIS}}(\theta) = \sum_{t=0}^{m} \frac{1}{n_t} \sum_{i=1}^{n_t} \omega_{t,m}(a_{t,i}) y_{t,i} w_{t,i}^{\theta}, \quad w_{t,i}^{\theta} = \frac{\pi_\theta(a_{t,i}|x_{t,i})}{\pi_{t,i}}. \tag{12}$$

In multiple importance sampling we usually assume that the behavior distributions are independent. In our case, when we optimize θ_t based on the models $\theta_{t-1}, \dots, \theta_0$, we break this assumption. However, as we will see, we can still have the unbiasedness property and derive an estimator for the variance of the estimator.

Proposition A.1 (Unbiasedness)**.** *The MIS estimator* (12) *is unbiased when the loss y is fixed (its distribution $\mathcal{P}_{\mathcal{Y}}(\cdot|x, a)$ does not depend on time rollout m).*

Proof. Let $m \in \{1, \dots M\}$. We recall that at all rounds $t < m$, models $\theta_t \in \Theta$ were deployed and sets s_t of n_t observations $s_t = (x_{t,i}, a_{t,i}, l_{t,i}, \pi_{t,i})_{i=1,\dots,n_t}$ were collected thereof, with propensities $\pi_{t,i} = \pi_{\theta_t}(a_{t,i}|x_{t,i})$ to learn the next model θ_{t+1}. To prove the unbiasedness we use the tower rule on the expectation and condition on previous observations $s_1, \dots s_{t-1}$:

$$\begin{aligned}
\mathbb{E}[\hat{L}_m^{\text{MIS}}(\theta)] &= \sum_{t=0}^{m} \frac{1}{n_t} \sum_{i=1}^{n_t} \mathbb{E}_{x,\theta_m,y} \left[\omega_t(a) y w_t^{\theta}\right] \\
&= \sum_{t=0}^{m} \mathbb{E}_{x,\theta_m,y} \left[\omega_t(a) y w_t^{\theta}\right] \\
&= \sum_{t=0}^{m} \mathbb{E}_{s_1 \dots s_{t-1}} \left[\mathbb{E}_{x,\theta_m,y} \left[\omega_t(a) y w_t^{\theta} \mid s_1 \dots s_{t-1}\right]\right] \\
&= \sum_{t=0}^{m} \mathbb{E}_{s_1 \dots s_{t-1}} \left[\mathbb{E}_{x,\theta,y} \left[\omega_t(a) y \mid s_1 \dots s_{t-1}\right]\right] \\
&= \sum_{t=0}^{m} \mathbb{E}_{x,\theta,y} \left[\omega_t(a) y\right] \\
&= \mathbb{E}_{x,\theta,y} \left[\left(\sum_{t=0}^{m} \omega_t(a)\right) y\right] \\
&= \mathbb{E}_{x,\theta,y} \left[y\right] \\
&= L(\theta),
\end{aligned}$$

where the second last line is true only when the distribution of y does not change over time roll-outs m. □

Among the proposals for functions $\omega_t(a)$, the most 'naive' and natural heuristic is to choose

$$\omega_t(a) = \frac{n_t}{\sum_{l=1}^{m} n_l}, \tag{13}$$

which gives the naive concatenation of all IPS estimators

$$\hat{L}_m^{\text{n-MIS}}(\theta) = \frac{1}{n} \sum_{t=0}^{m} \sum_{i=1}^{n_t} y_{t,i} \frac{\pi_\theta(a_{t,i}|x_{t,i})}{\pi_{\theta_t}(a_{t,i}|x_{t,i})}, \tag{14}$$

where $n = \sum_{t=0}^{m} n_t$.

With the previous definition of the empirical mean estimator, we can now derive an empirical variance estimator, starting with the naive multi importance sampling estimator. We write the random variable $r^m = (\pi_\theta/\pi_{\theta_m})y$. We note that for inside a batch m each realization of $r_i^m = (\pi_\theta(a_{m,i}|x_{m,i})/\pi_{m,i})y_{m,i}$ and r_j^m are independent. But the realizations of the random variables r^m and $r^{m'}$ are dependent. Writing $n = \sum_{t=0}^m n_t$

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{n}\sum_{t=0}^{m}\sum_{i=1}^{n_m} r_i^m\right] &= \sum_{t=0}^{m}\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n_m} r_i^m\right] + 2\sum_{1\le p<q\le m}\mathrm{Cov}\left[\frac{1}{n}\sum_{i=1}^{n_p} r_i^p, \frac{1}{n}\sum_{j=1}^{n_q} r_j^q\right] \\
&= \frac{1}{n^2}\sum_{t=0}^{m}\mathrm{Var}\left[\sum_{i=1}^{n_m} r^m\right] + 2\frac{1}{n^2}\sum_{1\le p<q\le m}\sum_{i=1}^{n_p}\sum_{j=1}^{n_q}\mathrm{Cov}\left[r^p, r^q\right] \\
&= \frac{1}{n^2}\left[\sum_{t=0}^{m}\mathrm{Var}\left[\sum_{i=1}^{n_m} r^m\right] + 2\sum_{1\le p<q\le m} n_p n_q \mathrm{Cov}\left[r^p, r^q\right]\right],
\end{aligned}$$

where the second last equality is obtained with the bilinearity of the covariance. Given the latter expression of the variance, we propose the following estimator and with a linear sampling where all $n_p = n_q$ for $p, q \in \{1, \dots, M\}$:

$$\hat{V}_m^{\text{n-MIPS}}(\theta) = \frac{1}{n^2}\left[\sum_{t=0}^{m}\hat{V}(r^t) + 2\sum_{1\le p<q\le m} n_p n_q \left(\frac{1}{n_p}\sum_{k=1}^{n_p}\left(r_k^p - \bar{r}_p\right)\left(r_k^q - \bar{r}_q\right)\right)\right], \tag{15}$$

where $\hat{V}(r^m) = \frac{1}{n_m(n_m-1)}\sum_{i=1}^{n_m}\left(r_i^m - \bar{r}^m\right)^2$ and $\bar{r}^m = \frac{1}{n_m}\sum_{j=1}^{n_m} r_j^m$.

Note also that for other functions $\omega_t(a)$, the most studied one is the balance heuristic with $\omega_t \propto n_t \pi_{\theta_t}(a)$, that is:

$$\omega_t^{BH}(a) = \frac{n_t \pi_{\theta_t}(a)}{\sum_{l=1}^{m} n_l \pi_{\theta_l}(a)}. \tag{16}$$

The latter heuristic has been studied for its low variance (Owen, 2013) but these properties have been studied under an i.i.d assumption that is broken in our adaptive data collection strategy. Eventually, note that controlling the variance of this estimator with an implicit exploration estimator as we do in the i.i.d case would make a an interesting research direction.

B. Analysis details

In this section, we provide the details of our analysis by starting with essential definitions, then our proofs of variance dependent excess risk bounds and finally our regret analysis.

B.1. Definitions

$C_m(\Theta)$ is a complexity measure that will be upper-bounded by the metric entropy in sup-norm at level $\varepsilon = 1/n_m$ of the following function set,

$$\mathcal{F}_{m,\Theta} := \left\{ f_\theta : (x, a, y) \in \mathcal{X}\times\mathcal{A}\times\mathcal{Y} \mapsto \frac{1}{W} + \frac{1}{W} y\left(\frac{\pi_\theta(a|x)}{\pi_{\theta_m}(a|x) + \alpha\pi_\theta(a|x)} - 1\right) \quad \text{for } \theta\in\Theta \right\}. \tag{17}$$

The latter corresponds to clipped prediction errors of policies π_θ normalized into $[0, 1]$. More precisely, to define rigorously $C_m(\Theta)$, we denote for any $n_m \ge 1$ and $\varepsilon > 0$, the complexity of a class $\mathcal{F}$ by

$$\mathcal{H}_\infty(\varepsilon, \mathcal{F}, n) = \sup_{(x_i,a_i,y_i)\in(\mathcal{X}\times\mathcal{A}\times\mathcal{Y})^n} \mathcal{H}(\varepsilon, \mathcal{F}\big(\{x_i, a_i, y_i\}\big), \|\cdot\|_\infty), \tag{18}$$

where $\mathcal{F}\big(\{x_i, a_i, y_i\}\big) = \big\{\big(f(x_1, a_1, y_1), \dots, f(x_n, a_n, y_n)\big), f \in \mathcal{F}\big\} \subseteq \mathbb{R}^n$ and the number $\mathcal{H}(\varepsilon, A, \|\cdot\|_\infty)$ is the smallest cardinality $|A_0|$ of a set $A_0 \subseteq A$ such that A is contained in the finite union of ε-balls centered at points in A_0 in the metric induced by $\|\cdot\|_\infty$). Then, $C_m(\Theta)$ is defined by

$$C_m(\Theta) = \log \mathcal{H}_\infty(1/n_m, \mathcal{F}_{m,\Theta}, 2n_m). \tag{19}$$

B.2. Variance-dependent excess risk bounds

We will denote by $\mathbb{E}_m[\cdot] = \mathbb{E}[\cdot|s_0, \dots s_m]$ the conditional expectation given the set of observation samples $s_m = (x_{m,i}, a_{m,i}, y_{m,i}, \pi_{m,i})_{i=1,\dots,n_m}$ up to the rollout m. Here, we recall that $x_{m,i} \sim \mathcal{P}_{\mathcal{X}}$, $a_{m,i} \sim \pi_{\theta_m}(\cdot|x_{m,i})$, $y_{m,i} \sim \mathcal{P}_{\mathcal{Y}}(\cdot|x_{m,i}, a_{m,i})$, and $\pi_{m,i} = \pi_{\theta_m}(a_{m,i}|x_{m,i})$. Furthermore, throughout the document, $\mathbb{E}_{x,\theta_m,y}\left[\,\cdot\,\right]$ (resp. $\mathrm{Var}_{x,\theta_m,y}[\cdot]$) denotes the expectation (resp. variance) in (x, a, y) where $x \sim \mathcal{P}_{\mathcal{X}}$, $a \sim \pi_{\theta_m}(\cdot|x)$, and $y \sim \mathcal{P}_{\mathcal{Y}}(\cdot|x, a)$.

Proposition 4.1 (Generalization Error Bound)**.** *Let $\hat{L}_m^{\text{IPS-IX}}$ and $\hat{V}_m^{\text{IPS-IX}}$ be the empirical estimators defined respectively in Eq. (9) and Eq. (10). Let $\delta \in (0, 1)$, $\theta \in \Theta$, and $n_m \geq 2$ the number of samples associated to the logged dataset at round m. Then, with probability at least $1 - \delta$,*

$$L(\theta) \leq \hat{L}_m^{\text{IPS-IX}}(\theta) + \lambda\sqrt{\frac{\hat{V}_m^{\text{IPS-IX}}(\theta)}{n_m}} + \frac{2\lambda^2 W}{n_m} + \sqrt{\frac{\log(2/\delta)}{2n_m}}\,, \tag{20}$$

where $\lambda = \sqrt{18(C_m(\Theta) + \log(2/\delta))}$.

Proof. Let $\delta \in (0, 1)$ and $\theta \in \Theta$. Since all functions in $\mathcal{F}_{m,\Theta}$ defined in Eq. (17) take values in $[0, 1]$, we can apply the concentration bound of Maurer & Pontil (2009, Theorem 6) to the set $\mathcal{F}_{m,\Theta}$. This yields, with probability at least $1 - \delta/2$,

$$\mathbb{E}_{x,\theta_m,y}[f_\theta(x, a, y)] - \frac{1}{n_m}\sum_{i=1}^{n_m} f_\theta(x_{m,i}, a_{m,i}, y_{m,i}) \leq \sqrt{\frac{18\hat{V}_{n_m}(f_\theta)(C_m(\Theta) + \log(2/\delta))}{n_m}} + \frac{15(C_m(\Theta) + \log(1/\delta))}{(n_m - 1)}\,, \tag{21}$$

where

$$\hat{V}_{n_m}(f_\theta) = \frac{1}{n_m - 1}\sum_{i=1}^{n_m}\Big(f_\theta(x_{m,i}, a_{m,i}, y_{m,i}) - \frac{1}{n_m}\sum_{j=1}^{n_m} f_\theta(x_{m,j}, a_{m,j}, y_{m,j})\Big)^2$$

is an estimation of the sample variance. Let $\alpha > 0$ and define the following biased estimator of the excess risk:

$$L_m^\alpha(\theta) = \mathbb{E}_{x,\theta_m,y}\left[y\left(\frac{\pi_\theta(a|x)}{\pi_{\theta_m}(a|x) + \alpha\pi_\theta(a|x)} - 1\right)\right] \qquad \forall\theta \in \Theta. \tag{22}$$

We recall that $\mathbb{E}_{x,\theta_m,y}\left[\,\cdot\,\right]$ denotes the expectation in (x, a, y) where $x \sim \mathcal{P}_{\mathcal{X}}$, $a \sim \pi_{\theta_m}(\cdot|x)$, and $y \sim \mathcal{P}_{\mathcal{Y}}(\cdot|x, a)$. By construction of f_θ (see Eq. (17)),

$$\begin{aligned}
\mathbb{E}_{x,\theta_m,y}[f_\theta(x, a, y)] &= \frac{1}{W} + \frac{1}{W}L_m^\alpha(\theta) \\
\frac{1}{n_m}\sum_{i=1}^{n_m} f_\theta(x_{m,i}, a_{m,i}, y_{m,i}) &= \frac{1}{W} + \frac{1}{W}\hat{L}_m^{\text{IPS-IX}}(\theta) - \frac{1}{Wn_m}\sum_{i=1}^{n_m} y_{m,i} \\
\hat{V}_{n_m}(f_\theta) &= \frac{1}{W^2}\hat{V}_m^{\text{IPS-IX}}(\theta)\,,
\end{aligned}$$

where $\hat{L}_m^{\text{IPS-IX}}$ and $\hat{V}_m^{\text{IPS-IX}}$ are defined respectively in Eq. (9) and Eq. (10). Thus, multiplying (21) by W, substituting the above terms, and using $\lambda = \sqrt{18(C_m(\Theta) + \log(2/\delta))}$, yields

$$L_m^\alpha(\theta) - \hat{L}_m^{\text{IPS-IX}}(\theta) + \frac{1}{n_m}\sum_{i=1}^{n_m} y_{m,i} \leq \lambda\sqrt{\frac{\hat{V}_m^{\text{IPS-IX}}(\theta)}{n_m}} + \frac{15\lambda^2 W}{18(n_m - 1)}, \tag{23}$$

with probability $1 - \delta/2$. Now, let us decompose

$$L_m^\alpha(\theta) = \mathbb{E}_{x,\theta_m,y}\left[y\left(\frac{\pi_\theta(a|x)}{\pi_{\theta_m}(a|x) + \alpha\pi_\theta(a|x)} - 1\right)\right] = \mathbb{E}_{x,\theta_m,y}\left[y\frac{\pi_\theta(a|x)}{\pi_{\theta_m}(a|x) + \alpha\pi_\theta(a|x)}\right] - L(\theta_m).$$

But, since the losses y are bounded in $[-1, 0]$ almost surely,

$$\mathbb{E}_{x,\theta_m,y}\left[y\frac{\pi_\theta(a|x)}{\pi_{\theta_m}(a|x) + \alpha\pi_\theta(a|x)}\right] \geq \mathbb{E}_{x,\theta_m,y}\left[y\frac{\pi_\theta(a|x)}{\pi_{\theta_m}(a|x)}\right] = L(\theta),$$

which, substituted into the previous equation, entails,

$$L_m^{\alpha}(\theta) \geq L(\theta) - L(\theta_m). \tag{24}$$

Lower-bounding the left-hand side of (26), we thus get w.p $1-\delta/2$,

$$L(\theta) - \hat{L}_m^{\text{IPS-IX}}(\theta) \leq \lambda\sqrt{\frac{\hat{V}_m^{\text{IPS-IX}}(\theta)}{n_m}} + \frac{15\lambda^2 W}{18(n_m-1)} + L(\theta_m) - \frac{1}{n_m}\sum_{i=1}^{n_m} y_{m,i}.$$

Using $\mathbb{E}_{m-1}[y_{m,i}] = L(\theta_m)$ and applying Hoeffding's inequality, this further yields w.p. $1-\delta$

$$L(\theta) \leq \hat{L}_m^{\text{IPS-IX}}(\theta) + \lambda\sqrt{\frac{\hat{V}_m^{\text{IPS-IX}}(\theta)}{n_m}} + \frac{15\lambda^2 W}{18(n_m-1)} + \sqrt{\frac{\log(2/\delta)}{2n_m}}. \tag{25}$$

Eventually, note that $(n_m-1)^{-1} \leq (2/n_m)$ since $n_m \geq 2$. Thus,

$$L(\theta) \leq \hat{L}_m^{\text{IPS-IX}}(\theta) + \lambda\sqrt{\frac{\hat{V}_m^{\text{IPS-IX}}(\theta)}{n_m}} + \frac{2\lambda^2 W}{n_m} + \sqrt{\frac{\log(2/\delta)}{2n_m}}, \tag{26}$$

which concludes the proof.

□

Proposition 4.2 (Conservative Excess Risk)**.** *Let $m \geq 0$ and $\theta_m \in \Theta$. Let $s_m = (x_{m,i}, a_{m,i}, y_{m,i}, \pi_{m,i})_{1\leq i\leq n_m}$ be a set of samples collected with $a_{m,i} \sim \pi_{\theta_m}(\cdot|x_{m,i})$. Then, under Assumptions 4.1 and 4.2, the solution θ_{m+1} of Problem (8) with the IPS-IX estimator in Eq. (11) on the samples s_m satisfies the excess risk upper-bound*

$$\Delta_{m+1} = L(\theta_{m+1}) - L(\theta^*) \lesssim \sqrt{\frac{d\log(n_m)+\log(1/\delta)}{n_m}\nu_m^2} + \frac{W^2 + W(d\log(n_m)+\log(1/\delta))}{n_m}, \tag{27}$$

where $\nu_m^2 = \mathrm{Var}_{x,\theta_m}\left(\frac{\pi_{\theta^*}(a|x)}{\pi_{\theta_m}(a|x)}\right)$.

Proof. We consider the notations of the proof of Proposition 4.1. Fix $\theta^* \in \Theta$. Applying, Theorem 15 of (Maurer & Pontil, 2009)[2] to the function set $\mathcal{F}_{m,\Theta}$ defined in (17), we get with probability $1-\delta$

$$\begin{aligned}&\mathbb{E}_{x,\theta_m,y}[f_{\theta_{m+1}}(x,a,y)] - \mathbb{E}_{x,\theta_m,y}[f_{\theta^*}(x,a,y)]\\ &\qquad\leq \sqrt{\frac{32\mathrm{Var}_{x,\theta_m,y}\big[f_{\theta^*}(x,a,y)\big]\big(C_m(\Theta)+\log\frac{30}{\delta}\big)}{n_m}} + \frac{22\big(C_m(\Theta)+\log\frac{30}{\delta}\big)}{n_m-1}.\end{aligned}$$

This can be written as:

$$\Delta_m^* \leq U_m^*, \tag{28}$$

with the following definitions:

$$\begin{aligned}\Delta_m^* &= \mathbb{E}_{x,\theta_m,y}[f_{\theta_{m+1}}(x,a,y)] - \mathbb{E}_{x,\theta_m,y}[f_{\theta^*}(x,a,y)]\\ U_m^* &= \sqrt{\frac{32\mathrm{Var}_{x,\theta_m,y}\big[f_{\theta^*}(x,a,y)\big]\big(C_m(\Theta)+\log\frac{30}{\delta}\big)}{n_m}} + \frac{22\big(C_m(\Theta)+\log\frac{30}{\delta}\big)}{n_m-1}.\end{aligned} \tag{29}$$

[2]Note that in their notation, $\log\mathcal{M}_n(\pi)$ equals $C_m(\Theta)+\log(10)$, $\mathbf{X}$ is the dataset $\{(x_i,a_i,y_i)\}_{1\leq i\leq n}$ where $(x_i,a_i,y_i) \overset{i.i.d.}{\sim} \mathcal{P}_{\mathcal{X}} \times \pi_{\theta_m}(\cdot|x) \times \mathcal{P}_{\mathcal{Y}}(\cdot|a,x)$, and $P(\cdot,\mu)$ is the expectation with respect to one test sample $\mathbb{E}_{x,\theta_m,y}[\,\cdot\,]$.

Step: Lower bounding Δ_m^* Using the definition of $f_\theta(x,a,y)$ in (17) and that of L_m^α in Eq. (22), we have

$$\mathbb{E}_{x,\theta_m,y}[f_{\theta_{m+1}}(x,a,y)] = \frac{1}{W} + \frac{1}{W} L_m^\alpha(\theta_{m+1}).$$

Thus, Δ_m^* can be re-written as

$$\Delta_m^* = \frac{1}{W}\left(L_m^\alpha(\theta_{m+1}) - L_m^\alpha(\theta^*)\right),$$

which we now lower-bound. To do so, we begin by upper-bounding $L_m^\alpha(\theta^*)$. It can be expressed as

$$L_m^\alpha(\theta^*) = \mathbb{E}_{x,\theta_m,y}\left[y\frac{\pi_{\theta^*}(a|x)}{\pi_{\theta_m}(a|x) + \alpha\pi_{\theta^*}(a|x)}\right] - L(\theta_m). \tag{30}$$

To shorten notation, from now on and throughout this proof, we write π_θ instead of $\pi_\theta(a|x)$, omitting the dependence on a and x. Using the inequality $(1+x)^{-1} \geq 1 - x$ for $x \geq 0$, we have

$$\begin{aligned}
\mathbb{E}_{x,\theta_m,y}\left[y\frac{\pi_{\theta^*}}{\pi_{\theta_m} + \alpha\pi_{\theta^*}}\right] &= \mathbb{E}_{x,\theta_m,y}\left[y\frac{\pi_{\theta^*}}{\pi_{\theta_m}}\frac{1}{1+\alpha\frac{\pi_{\theta^*}}{\pi_{\theta_m}}}\right] && (31)\\
&\leq \mathbb{E}_{x,\theta_m,y}\left[y\frac{\pi_{\theta^*}}{\pi_{\theta_m}}\right] - \alpha\mathbb{E}_{x,\theta_m,y}\left[y\left(\frac{\pi_{\theta^*}}{\pi_{\theta_m}}\right)^2\right]\\
&= L(\theta^*) - \alpha\mathbb{E}_{x,\theta_m,y}\left[y\left(\frac{\pi_{\theta^*}}{\pi_{\theta_m}}\right)^2\right]\\
&\leq L(\theta^*) + \alpha W^2, && (32)
\end{aligned}$$

where the last inequality is by Assumption 4.1 and because $y \in [-1, 0]$. Together with (30), we get

$$L_m^\alpha(\theta^*) \leq L(\theta^*) + \alpha W^2 - L(\theta_m).$$

We recall that $L(\theta_{m+1}) - L(\theta_m) \leq L_m^\alpha(\theta_{m+1})$ by Eq.(24). Therefore,

$$\frac{1}{W}(L(\theta_{m+1}) - L(\theta^*) - \alpha W^2) \leq \frac{1}{W}\left(L_m^\alpha(\theta_{m+1}) - L_m^\alpha(\theta^*)\right),$$

which finally gives

$$\frac{1}{W}(L(\theta_{m+1}) - L(\theta^*) - \alpha W^2) \leq \Delta_m^*. \tag{33}$$

Step: Upper bound U_m^* By definition of $f_\theta(x,a,y)$ in (17), we have

$$\begin{aligned}
\mathrm{Var}_{x,\theta_m,y}\left[f_{\theta^*}(x,a,y)\right] &= \frac{1}{W^2}\mathrm{Var}_{x,\theta_m,y}\left[y\left(\frac{\pi_{\theta^*}}{\pi_{\theta_m} + \alpha\pi_{\theta^*}} - 1\right)\right]\\
&\leq \frac{1}{W^2}\mathbb{E}_{x,\theta_m,y}\left[y^2\left(\frac{\pi_{\theta^*}}{\pi_{\theta_m} + \alpha\pi_{\theta^*}} - 1\right)^2\right] \leq \frac{1}{W^2}\mathbb{E}_{x,\theta_m}\left[\left(\frac{\pi_{\theta^*}}{\pi_{\theta_m} + \alpha\pi_{\theta^*}} - 1\right)^2\right].
\end{aligned}$$

Then, using the inequality $(x+y)^2 \leq 2x^2 + 2y^2$, for $x, y \in \mathbb{R}$, this may be upper-bounded as

$$\begin{aligned}
&\mathrm{Var}_{x,\theta_m,y}\left[f_{\theta^*}(x,a,y)\right]\\
&\leq \frac{2}{W^2}\mathbb{E}_{x,\theta_m}\left[\left(\frac{\pi_{\theta^*}}{\pi_{\theta_m} + \alpha\pi_{\theta^*}} - \mathbb{E}_{x,\theta_m}\left[\frac{\pi_{\theta^*}}{\pi_{\theta_m} + \alpha\pi_{\theta^*}}\right]\right)^2\right] + \frac{2}{W^2}\left(\mathbb{E}_{x,\theta_m}\left[\frac{\pi_{\theta^*}}{\pi_{\theta_m} + \alpha\pi_{\theta^*}}\right] - 1\right)^2. \quad (34)
\end{aligned}$$

On the one hand, the first term of the right-hand side may be upper-bounded as

$$\mathbb{E}_{x,\theta_m}\left[\left(\frac{\pi_{\theta^*}}{\pi_{\theta_m} + \alpha\pi_{\theta^*}} - \mathbb{E}_{x,\theta_m}\left[\frac{\pi_{\theta^*}}{\pi_{\theta_m} + \alpha\pi_{\theta^*}}\right]\right)^2\right] = \mathrm{Var}_{x,\theta_m}\left[\frac{\pi_{\theta^*}}{\pi_{\theta_m} + \alpha\pi_{\theta^*}}\right] \leq \nu_m^2,$$

where $\nu_m^2 = \mathrm{Var}_{x,\theta_m}\left[\frac{\pi_{\theta^*}}{\pi_{\theta_m}}\right]$. On the other hand, for the second term, we use the same factorization as in Eq. (31) to get

$$-\alpha\, \mathbb{E}_{x,\theta_m}\left[\left(\frac{\pi_{\theta^*}}{\pi_{\theta_m}}\right)^2\right] \leq \mathbb{E}_{x,\theta_m}\left[\frac{\pi_{\theta^*}}{\pi_{\theta_m} + \alpha\pi_{\theta^*}}\right] - 1 \leq 0\,,$$

which yields the upper-bound

$$\left(\mathbb{E}_{x,\theta_m}\left[\frac{\pi_{\theta^*}}{\pi_{\theta_m} + \alpha\pi_{\theta^*}}\right] - 1\right)^2 \leq \alpha^2\, \mathbb{E}_{x,\theta_m}\left[\left(\frac{\pi_{\theta^*}}{\pi_{\theta_m}}\right)^2\right] \leq \alpha^2 W^2.$$

Therefore, substituting the last two upper-bounds into (34) entails

$$\mathrm{Var}_{x,\theta_m,y}\big[f_{\theta^*}(x,a,y)\big] \leq \frac{2}{W^2}\big(\nu_m^2 + \alpha^2 W^2\big)\,.$$

Then, replacing this upper-bound into the definition of U_m^* in (29) and using Assumption 4.2 to upper bound the terms in $C_m(\Theta) \leq d\log(n_m)$, we obtain the following upper-bound

$$\begin{aligned} U_m^* &\leq \frac{1}{W}\sqrt{\frac{64(\nu_m^2 + \alpha^2 W)\big(d\log(n_m) + \log\frac{30}{\delta}\big)}{n_m}} + \frac{22\big(d\log(n_m) + \log\frac{30}{\delta}\big)}{n_m - 1} \\ &\leq \frac{1}{W}\sqrt{\frac{64(\nu_m^2 + \alpha^2 W)\big(d\log(n_m) + \log\frac{30}{\delta}\big)}{n_m}} + \frac{44\big(d\log(n_m) + \log\frac{30}{\delta}\big)}{n_m}\,, \end{aligned} \tag{35}$$

where the last inequality is because $n_m \geq 2$.

Step: excess risk upper bound Setting $\alpha = \frac{1}{n_m}$ and using the two previous bounds (33) and (35) respectively on Δ_m^* and on U_m^* into (28), we get

$$L(\theta_{m+1}) - L(\theta^*) \leq \sqrt{\frac{64\big(d\log(n_m) + \log\frac{30}{\delta}\big)}{n_m}\big(\nu_m^2 + \frac{1}{n_m^2}W^2\big)} + W\frac{44\big(d\log(n_m) + \log\frac{30}{\delta}\big)}{n_m} + \frac{1}{n_m}W^2. \tag{36}$$

Using that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we have that

$$\sqrt{\frac{64\big(d\log(n_m) + \log\frac{30}{\delta}\big)}{n_m}\big(\nu_m^2 + \frac{1}{n_m^2}W^2\big)} \leq \sqrt{\frac{64\big(d\log(n_m) + \log\frac{30}{\delta}\big)}{n_m}\nu_m^2} + \frac{W}{n_m}\sqrt{\frac{64\big(d\log(n_m) + \log\frac{30}{\delta}\big)}{n_m}}.$$

Then, since $n_m \geq 2$ and $\delta < 1$, we have $d\log(n_m) + \log(30/\delta) \geq \log(2) + \log(30) \geq 4$, which yields

$$\frac{1}{n_m}\sqrt{\frac{64\big(d\log(n_m) + \log\frac{30}{\delta}\big)}{n_m}} \leq \frac{\sqrt{32\big(d\log(n_m) + \log\frac{30}{\delta}\big)}}{n_m} \leq \frac{\sqrt{8}\big(d\log(n_m) + \log\frac{30}{\delta}\big)}{n_m}.$$

Substituting the last two inequalities into (36) finally entails

$$L(\theta_{m+1}) - L(\theta^*) \leq 8\sqrt{\frac{d\log(n_m) + \log\frac{30}{\delta}}{n_m}\nu_m^2} + 47W\frac{d\log(n_m) + \log\frac{30}{\delta}}{n_m} + \frac{W^2}{n_m}, \tag{37}$$

which concludes the proof. □

B.3. Regret analysis

Proposition 5.1 (Regret upper-bound). *Let $n_0, n \geq 2$ and $\theta^* \in \arg\min_\theta L(\theta)$. Let $n_m = n_0 2^m$ for $m = 0, \ldots, M = \lfloor \log_2(1 + \frac{n}{n_0})\rfloor$. Then, under Assumptions 4.1, 4.2 and 5.1, the SCRM procedure (Alg. 1) satisfies the excess risk upper-bound*

$$L(\theta_M) - L(\theta^*) \leq O\Big(n^{-\frac{1}{2-\beta}}\log n\Big)\,.$$

Moreover, the expected regret is upper-bounded as follows:

$$R_n = \mathbb{E}\left[\sum_{m=0}^{M} n_{m+1}\big(L(\theta_m) - L(\theta^*)\big)\right] \leq O\Big(n^{\frac{1-\beta}{2-\beta}}\log(n)^2\Big).$$

Proof. First, note that for $n_m = n_0 2^m$ and $M = \lfloor \log_2(1+\frac{n}{n_0})\rfloor$, we have $\sum_{m=0}^{M-1} n_m = n_0(2^M - 1) \leq n$. Hence, Alg. 1 has collected at most n samples to design the estimator θ_M. For $m \geq 0$, we recall $\Delta_m = L(\theta_m) - L(\theta^*)$ and use Eq. (37) to write

$$\begin{aligned}
\Delta_{m+1} &\leq 8\sqrt{\frac{d\log(n_m) + \log\frac{30}{\delta}}{n_m}\nu_m^2} + 47W\frac{d\log(n_m) + \log\frac{30}{\delta}}{n_m} + \frac{W^2}{n_m} \\
&\leq 8\sqrt{\frac{d\log(n) + \log\frac{30}{\delta}}{n_m}\nu_m^2} + 47W\frac{d\log(n) + \log\frac{30}{\delta}}{n_m} + \frac{W^2}{n_m} \\
&= C\sqrt{\frac{\nu_m^2}{n_m}} + \frac{B}{n_m}\,, \qquad (38)
\end{aligned}$$

where $C = 8\sqrt{d\log(n) + \log\frac{30}{\delta}}$ and $B = W^2 + 47W(d\log(n) + \log\frac{30}{\delta})$ are independent of m.

Step: Obtaining a recurrence relation for Δ_{m+1} By Assumption 5.1, there exist $\gamma > 0$ and $\beta \in [0,1]$ such that

$$\nu_m^2 = \mathrm{Var}_{x,\theta_m}\left(\frac{\pi_{\theta^*}}{\pi_{\theta_m}}\right) \leq \frac{1}{\gamma}\big(L(\theta_m) - L(\theta^*)\big)^\beta = \frac{\Delta_m^\beta}{\gamma}\,.$$

Replacing ν_m^2 in Eq. (38) thus entails

$$\begin{aligned}
\Delta_{m+1} &\leq C\sqrt{\frac{1}{\gamma}\frac{\Delta_m^\beta}{n_m}} + \frac{B}{n_m} \\
&\leq C2^{-\frac{m}{2}}\sqrt{\frac{n_0}{\gamma}\Delta_m^\beta} + B2^{-m}n_0 \qquad \leftarrow n_m = n_0 2^m \\
&= C\sqrt{\frac{n_0}{\gamma}}2^{-\frac{m}{2}}\Delta_m^{\beta/2} + B2^{-m}n_0\,. \qquad (39)
\end{aligned}$$

Step: Solving the recurrence relation for Δ_m We then insure by induction that Δ_m satisfies

$$\Delta_m \leq c_0 2^{\frac{-m}{2-\beta}}\,, \qquad (40)$$

for some $c_0 > 0$ that will be specified by the analysis.

Base step Since losses take values in $[-1, 0]$, $\Delta_0 = L(\theta_0) - L(\theta^*) \leq 1$. Equation (40) is thus satisfied for $m = 0$ as soon as $c_0 \geq 1$.

Induction step Let $m \geq 0$. We assume that $\Delta_m \leq c_0 2^{\frac{-m}{2-\beta}}$ and prove Equation (40) for Δ_{m+1}. Using Eq. (39), we have

$$\begin{aligned}
\Delta_{m+1} &\leq C\sqrt{\frac{n_0}{\gamma}}2^{-\frac{m}{2}}\Delta_m^{\beta/2} + B2^{-m}n_0 \\
&\leq C\sqrt{\frac{n_0}{\gamma}}2^{-\frac{m}{2}}{c_0}^{\beta/2}2^{-\frac{m\beta}{\beta(2-\beta)}} + B2^{-m}n_0 \qquad \leftarrow \quad \text{by induction} \\
&\leq \max\left\{2C\sqrt{\frac{n_0}{\gamma}}{c_0}^{\frac{\beta}{2}}2^{-\frac{m}{2}-\frac{m}{2-\beta}}, 2B2^{-m}n_0\right\}. \qquad (41)
\end{aligned}$$

Now, we show that both terms inside the maximum can be upper-bounded by $c_0 2^{-(m+1)/(2-\beta)}$ as soon as c_0 is large enough. On the one hand, if $c_0 \geq 4Bn_0$, we have

$$2B2^{-m}n_0 \leq c_0 2^{-(m+1)} \leq c_0 2^{-\frac{m+1}{2-\beta}} .$$

On the other hand, if $c_0 \geq (4C^2 n_0/\gamma)^{1/(2-\beta)}$, we also have

$$2C\sqrt{\frac{n_0}{\gamma}} c_0^{\frac{\beta}{2}} 2^{-\frac{m}{2}-\frac{m}{2-\beta}} \leq 2C\sqrt{\frac{n_0}{\gamma}} c_0^{\frac{\beta}{2}} 2^{-\frac{m+1}{2-\beta}} \leq c_0 2^{-\frac{m+1}{2-\beta}} .$$

Combining the above two upper-bounds with (41) concludes the induction step under the condition

$$c_0 \geq \max\left\{1, \left(\frac{4C^2 n_0}{\gamma}\right)^{\frac{1}{2-\beta}}, 4Bn_0\right\} .$$

Step: conclusion Finally, setting the above value for c_0 we proved that for all $m \geq 0$, we have

$$\begin{aligned}
\Delta_m &\leq \max\left\{1, \left(\frac{4C^2 n_0}{\gamma}\right)^{\frac{1}{2-\beta}}, 4Bn_0\right\} 2^{-\frac{m}{2-\beta}} \\
&\leq \left(1 + \left(\frac{4C^2 n_0}{\gamma}\right)^{\frac{1}{2-\beta}} + 4Bn_0\right) 2^{-\frac{m}{2-\beta}} \\
&= \left(1 + \left(\frac{256(d\log n + \log\frac{30}{\delta})n_0}{\gamma}\right)^{\frac{1}{2-\beta}} + W^2 n_0 + 47Wn_0\Big(d\log n + \log\frac{30}{\delta}\Big)\right) 2^{-\frac{m}{2-\beta}} ,
\end{aligned} \tag{42}$$

where the last equality is by substituting the values of B and C from (38). For the final step $M = \lfloor \log_2(\frac{n}{n_0}+1)\rfloor$, this yields

$$\begin{aligned}
\Delta_M &\leq \left(1 + \left(\frac{256(d\log n + \log\frac{30}{\delta})n_0}{\gamma}\right)^{\frac{1}{2-\beta}} + W^2 n_0 + 47Wn_0\Big(d\log n + \log\frac{30}{\delta}\Big)\right) 2^{-\frac{M}{2-\beta}} \\
&\leq 2\left(1 + \left(\frac{256(d\log n + \log\frac{30}{\delta})n_0}{\gamma}\right)^{\frac{1}{2-\beta}} + W^2 n_0 + 47Wn_0\Big(d\log n + \log\frac{30}{\delta}\Big)\right) \times \left(\frac{n_0}{n}\right)^{\frac{1}{2-\beta}} \\
&= O\Big(n^{-\frac{1}{2-\beta}} \log n\Big) .
\end{aligned}$$

This concludes the first part of the proof.

Regret upper-bound To upper bound the cumulative regret, using $n_{m+1} = n_0 2^{m+1}$, we write

$$R_n = \sum_{m=0}^{M} \Delta_m n_{m+1} \overset{(42)}{\leq} D \sum_{m=0}^{M} 2^{-\frac{m}{2-\beta}} n_{m+1} = 2Dn_0 \sum_{m=0}^{M} 2^{\left(\frac{1-\beta}{2-\beta}\right)m} ,$$

where

$$D = 1 + \left(\frac{256(d\log n + \log\frac{30}{\delta})n_0}{\gamma}\right)^{\frac{1}{2-\beta}} + W^2 n_0 + 47Wn_0\Big(d\log n + \log\frac{30}{\delta}\Big) .$$

Then, computing the sum for $M = \lfloor \log_2(\frac{n}{n_0}+1)\rfloor$, we have

$$R_n \leq 2Dn_0 \sum_{m=0}^{M} 2^{\left(\frac{1-\beta}{2-\beta}\right)m} \leq 2Dn_0(M+1)2^{\left(\frac{1-\beta}{2-\beta}\right)M} \leq 2Dn_0\Big(1 + \log_2\Big(\frac{n}{n_0}+1\Big)\Big) \times \Big(1 + \frac{n}{n_0}\Big)^{\frac{1-\beta}{2-\beta}} .$$

Using that $D = O(\log n)$, we finally obtain

$$R_n \leq O\Big(n^{\frac{1-\beta}{2-\beta}} \log(n)^2\Big) .$$

□

C. Additional discussions on the Hölderian Bound Assumption 5.1

In this appendix, we discuss Assumption 5.1 on different particular examples.

C.1. Verification of the assumption on a toy example with Gaussian families

We consider the setting of Example 3.1. In the latter, the policies are Gaussian of the form $\pi_\theta = \mathcal{N}(\theta, \sigma^2)$ and the loss is defined by $l_t(a) = (a - y_t)^2 - 1$ where $y_t \sim \mathcal{N}(\theta^*, \sigma^2)$. There is no loss in generality in assuming $\sigma^2 = 1$. Then, we can compute

$$L(\theta) - L(\theta^*) = (\theta - \theta^*)^2 \qquad \text{and} \qquad \mathrm{Var}_\theta\left[\frac{\pi_{\theta^*}(a)}{\pi_\theta(a)}\right] = \exp\left((\theta^* - \theta)^2\right) - 1\,.$$

We recall that we are interested in verifying the existence of $\gamma > 0$ and $\beta > 0$ for which Assumption 5.1 holds, that is in this case for any $\theta \in \Theta$:

$$\gamma \mathrm{Var}_\theta\left[\frac{\pi_{\theta^*}(a)}{\pi_\theta(a)}\right] \leq \left(L(\theta) - L(\theta^*)\right)^\beta, \tag{43}$$

which may be re-written here as

$$\gamma\left(\exp\left((\theta^* - \theta)^2\right) - 1\right) \leq (\theta - \theta^*)^{2\beta}\,.$$

The latter is satisfied for any $\beta \leq 1$ as soon as Θ is a bounded interval. Note that the constant γ may decrease exponentially fast as the diameter of Θ increases. To illustrate, the existence of such couples (β, γ), we plot in Fig. 8 different values of the following ratio

$$R(\theta, \beta) = \frac{\mathrm{Var}_\theta\left[\frac{\pi_{\theta^*}(a)}{\pi_\theta(a)}\right]}{\left(L(\theta) - L(\theta^*)\right)^\beta} = \frac{\exp\left((\theta^* - \theta)^2\right) - 1}{\left(\|\theta - \theta^*\|^2\right)^\beta}. \tag{44}$$

The value of γ can be found for different values of β in Fig. 8 by taking $\frac{1}{\gamma} = \max_\theta R(\theta, \beta)$. Higher values of β induce

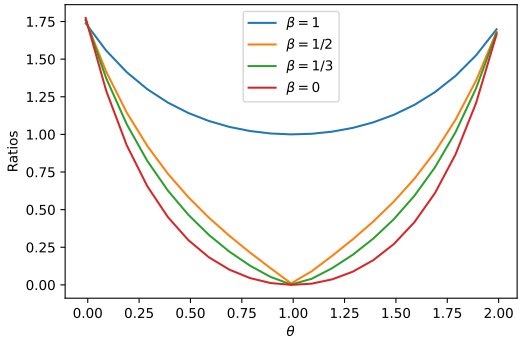

Figure 8: Ratio R defined in (44) with different values of β.

faster rates and lower values of γ induce worst constant terms in the excess risk and regret bounds. Eventually, note that SCRM does not need those parameters to run and those two parameters γ, β are automatically calibrated by SCRM to find the best trade-off.

C.2. Discussion of Assumption 5.1 for Exponential Families

In this section, we consider a more realistic example in which policies belong to an exponential family. That is, we assume that the policies are parameterized by a parameter $\eta \in \mathbb{R}^q$ and can be written in the form:

$$\forall a \in \mathcal{A}, \qquad \pi_\eta(a) = e^{\eta \cdot t(a) - A(\eta)} h(a),$$

for some known function $h : \mathcal{A} \to \mathbb{R}_+$ and sufficient statistic $t : \mathcal{A} \to \mathbb{R}^q$. Here, $A(\eta)$ is a normalization constant, so that $e^{A(\eta)} = \int_a e^{\eta \cdot t(a)} h(a)\,\mathrm{d}a$. We provide in Example C.1 a concrete example considered by (Swaminathan & Joachims,

2015a; Faury et al., 2020). To ease the notation, we removed here the dependency on contexts, but the generalization to contextual policies can be made similarly. The importance weight ratio may be written as,

$$\frac{\pi_\eta(a)}{\pi_{\eta_m}(a)} = e^{(\eta-\eta_m)t(a)-(A(\eta)-A(\eta_m))}. \tag{45}$$

To verify Assumption 5.1, we need to upper bound their variance, which we shall write as,

$$\mathrm{Var}_{a\sim\pi_{\eta_m}}\left[\frac{\pi_\eta(a)}{\pi_{\eta_m}(a)}\right] = e^{2(A(\eta_m)-A(\eta))}\mathrm{Var}_{a\sim\pi_{\eta_m}}\left[e^{(\eta-\eta_m)t(a)}\right].$$

Now, computing the moment generating function (MGF) of the statistic $t(a) \in \mathbb{R}^q$

$$M_t(s) = \mathbb{E}\left[e^{s\cdot t(a)}\right] = \int_a e^{s\cdot t(a)}e^{\eta_m\cdot t(a)-A(\eta_m)}h(a)\,\mathrm{d}a = e^{-A(\eta_m)}\int_a e^{(\eta_m+s)\cdot t(a)}e^{\eta_m\cdot t(a)}h(a)\,\mathrm{d}a = e^{A(\eta_m+s)-A(\eta_m)},$$

the variance term may be written as

$$\mathrm{Var}_{a\sim\pi_{\eta_m}}\left[e^{(\eta-\eta_m)t(a)}\right] = M_t(2(\eta-\eta_m)) - M_t^2(\eta-\eta_m) = e^{A(2\eta-\eta_m)-A(\eta_m)} - e^{2(A(\eta)-A(\eta_m))}.$$

This eventually leads us to

$$\mathrm{Var}_{a\sim\pi_{\eta_m}}\left[\frac{\pi_\eta(a)}{\pi_{\eta_m}(a)}\right] = e^{A(2\eta-\eta_m)+A(\eta_m)-2A(\eta)} - 1. \tag{46}$$

We now discuss two cases that are used for discrete actions (Swaminathan & Joachims, 2015a) and continuous actions (Kallus & Zhou, 2018; Zenati et al., 2020).

Bounded sufficient statistic Supposing that there exists an upper bound A such that $\|t(a)\| \leq A$, Cauchy-Schwartz inequality states that $|(\eta-\eta_m)\cdot t(a)| \leq \|\eta-\eta_m\|A$, which entails

$$\begin{aligned}
\mathrm{Var}_{a\sim\pi_{\eta_m}}\left[\frac{\pi_\eta(a)}{\pi_{\eta_m}(a)}\right] &= e^{A(2\eta-\eta_m)+A(\eta_m)-2A(\eta)} - 1 \\
&= \frac{\int_a e^{(2\eta-\eta_m)\cdot t(a)}h(a)\,\mathrm{d}a \int_a e^{\eta_m\cdot t(a)}h(a)\,\mathrm{d}a}{\left(\int_a e^{\eta\cdot t(a)}h(a)\,\mathrm{d}a\right)^2} - 1 \\
&= \frac{\int_a e^{(\eta-\eta_m)\cdot t(a)}e^{\eta\cdot t(a)}h(a)\,\mathrm{d}a \int_a e^{(\eta_m-\eta)\cdot t(a)}e^{\eta\cdot t(a)}h(a)\,\mathrm{d}a}{\left(\int_a e^{\eta\cdot t(a)}h(a)\,\mathrm{d}a\right)^2} - 1 \\
&\leq e^{\|\eta-\eta_m\|A} - 1\,.
\end{aligned}$$

Assuming that the parameter space is compact, i.e, $\max_{\eta,\eta'}\|\eta-\eta'\| \leq D$, there exists a constant C that depends on A and D such that, this may be further upper-bounded as

$$\mathrm{Var}_{a\sim\pi_{\eta_m}}\left[\frac{\pi_\eta(a)}{\pi_{\eta_m}(a)}\right] \leq C\|\eta-\eta_m\|\,.$$

Therefore, Assumption 5.1 is implied by

$$\gamma C\|\eta-\eta_m\|^2 \leq (L(\theta)-L(\theta^*))^{2\beta}\,.$$

The latter is implied by a local version of strong convexity for $\beta = 1/2$ (d'Aspremont et al., 2021), and holds with $\gamma = C^{-1}D^{-2}$ for $\beta = 0$.

Example C.1. *For discrete actions $\mathcal{A} = \{a_1,\dots,a_K\}$, we consider, as in (Swaminathan & Joachims, 2015a) and (Faury et al., 2020), policies where given a context x, probabilities $p_i(x)$ of sampling an action a_i are given by*

$$p_i(x) = \frac{\exp(\theta^\top\phi(x,a_i))}{\sum_{j=1}^K \exp(\theta^\top\phi(x,a_j))}. \tag{47}$$

The function ϕ is typically a feature map associated to a kernel in a RKHS. In this case, the natural parameter η and the sufficient statistic $t(a)$ may be written as

$$\eta = \begin{bmatrix} \log(\frac{p_1}{p_K}) \\ \vdots \\ \log(\frac{p_{K-1}}{p_K}) \\ 0 \end{bmatrix} \quad t(a) = \begin{bmatrix} \mathbb{1}\{a = a_1\} \\ \vdots \\ \mathbb{1}\{a = a_K\} \end{bmatrix}. \tag{48}$$

Lognormal and Normal distributions For normal $\mathcal{N}(\mu, \sigma^2)$ and lognormal Lognormal(μ, σ^2) distributions with fixed variance σ^2 as considered by (Kallus & Zhou, 2018; Zenati et al., 2020), the normalizing constant writes $A(\eta) = \frac{\eta^2}{2}$, and we then obtain that:

$$A(2\eta - \eta_m) + A(\eta_m) - 2A(\eta) = (\eta - \eta_m)^2 \,,$$

which gives:

$$\mathrm{Var}_{a\sim\pi_{\eta_m}}\left[\frac{\pi_\eta(a)}{\pi_{\eta_m}(a)}\right] = e^{\|\eta-\eta_m\|^2} - 1\,.$$

In that case, it is again possible for a bounded parameter space to linearize $e^{\|\eta-\eta_m\|^2} - 1 \lesssim \|\eta - \eta_m\|^2$, consider losses that verify: for all η, there exists an optimal η^* such that

$$\gamma\|\eta_m - \eta^*\|^2 \leq \big(L(\eta_m) - L(\eta^*)\big)^\beta. \tag{49}$$

Again, this holds generally for $\beta = 0$ and for locally strongly convex losses for $\beta = 1$.

D. Experiment details

D.1. Code

All the code to reproduce figures and tables is available in the following repository: https://github.com/criteo-research/sequential-conterfactual-risk-minimization.

D.2. Empirical settings details

Pricing The pricing application in (Demirer et al., 2019) considers a "personalized pricing" setting where given contexts x, prices p (which are the actions) need to be predicted to maximize the revenue:

$$r(x, p) = p(a(x) - b(x)p + \varepsilon)$$

where $\varepsilon \sim \mathcal{N}(0, 1)$ and $d = a(x) + b(x)p + \varepsilon$ is akin to an unknown context-specifidemand function. The data generating process uses contexts $x \in [1, 2]^k$ for $k > 1$ a positive integer. Only $l < k$ dimensions however affect the demand, that is if we write $\bar{x} = \frac{1}{l}(z_1, \ldots, z_l)$. The price p is generated from a Gaussian logging policy $p \sim \mathcal{N}(\bar{x}, 1)$ centered in $\bar{x}$. We consider in our example the quadratic functionnal $a(x) = 2x^2$ and $b(x) = 0.6x$ as in the original paper.

Advertising The advertising simulation in (Zenati et al., 2020) consists in predicting the potential $p \in]0, +\infty[$ of a user that may be compared to their a priori responsiveness to a treatment. The potential is caused by an unobserved random group variable g in G (groups of "high" or "low" potential users in their responsiveness) that influences context x of users. The goal is then to find a policy $\pi(a|x)$ that maximizes reward by adapting to an unobserved potential. The potentials are normally distributed conditionally on the group index, $p|g \sim \mathcal{N}(\mu_g, \sigma_g^2)$ where $\sigma_g = 0.5$ and $\mu_g = 1$ or 3 for two groups. The observed reward $-y$ is then a function of the action a and the context x through the associated potential p_x of the user x. The reward function mimics reward over the offline continuous bidding dataset in (Zenati et al., 2020) with the form:

$$r_l(p_x, a) = \begin{cases} \frac{a}{p_x} & \text{if } a < p_x \\ \frac{1}{2}(p_x - a) + 1 & \text{else} \end{cases}$$
$$r(p_x, a) = \max(r_l(p_x, a), -0.1)$$

The logging policy is a lognormal distribution as it is common in advertising applications (Bottou et al., 2013). In particular, as in (Zenati et al., 2020), $\pi_{\theta_0} = \text{Lognormal}(\mu, \sigma^2)$ where the mean $\exp(\mu + \sigma^2/2) = 2$ and the variance $(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2) = 1$.

Yeast*, *Scene*, *TMC2007 We follow (Swaminathan & Joachims, 2015a). We now recall briefly the setup. The problem is a binary multilabel classification with $|\mathcal{A}| = 2^K$ potential labels. All models are parametrized by $\pi_\theta(a|x) \propto \exp(\theta^\top(x \bigotimes a))$. The baseline (resp. skyline) is a supervised, full information model with identical parameter space than CRM methods trained on 5% (resp. 100%) of the training data. Our main modification it to consider the class of probabilistic policies that satisfy Assumption 5.1 by predicting actions in an Epsilon Greedy fashion (Sutton & Barto, 1998)): $\pi_\theta^\varepsilon(a, x) = (1 - \varepsilon)\pi_\theta(a, x) + \varepsilon/|\mathcal{A}|$ where $\varepsilon = .1$. The loss is the Hamming loss (number of incorrectly assigned labels - both false positives and false negatives in the action vector):

$$L(\theta) = \frac{1}{nK} \sum_{i=1}^{n} \sum_{j=1}^{K} \mathbb{1}_{[y_i^j = a_i^j]} \tag{50}$$

where y_i^j (resp. a_i^j) is the j-th component of the label vector (resp. action vector) of line i. A uniform policy will thus evaluate at a loss of .5.

D.3. Implementation details

Counterfactual methods In this paragraph we start by detailing the non adaptive counterfactual risk minimization that we compare to in this work.

Algorithm 2 Counterfactual Risk Minimization

Input: Logged observations $(x_{0,i}, a_{0,i}, y_{0,i}, \pi_{0,i})_{i=1,\dots,n_0}$, parameter $\lambda > 0$
for $m = 1$ *to* M **do**
  Build $\mathcal{L}_m$ from observations s_m using Eq. (11)
  Learn θ using Eq. (8)
  Re-deploy the logging model θ_0 and collect observations $s_{m+1} = (x_{m+1,i}, a_{m+1,i}, l_{m+1,i}, \pi_{m+1,i})_{i=1,\dots,n_{m+1}}$
end

We also provide the grid of hyperparameters for the λ evaluated in CRM and SCRM methods $\lambda \in [1e-5, 1e-4, 1e-3, 1e-2, 1e-1]$.

Batch Bandits Let $k : (\mathcal{X} \times \mathcal{A}) \times (\mathcal{X} \times \mathcal{A}) \to \mathbb{R}$ be a bounded positive definite Kernel associated to a RKHS $\mathcal{H}$, $\phi : \mathcal{X} \times \mathcal{A} \to \mathcal{H}$ is the feature map such that $k(s, s') = \langle \phi(s), \phi(s') \rangle$ for any $s, s' \in \mathcal{X} \times \mathcal{A}$. Context-actions pairs are written as $s_{m,i} := (x_{m,i}, a_{m,i}) \in \mathcal{X} \times \mathcal{A}$ and $\mathcal{S}_m := \{s_{1,0}, \dots, s_{n_m,m}\}$ denoting the history of all context-actions pairs seen up until the end of batch m. K_m is the kernel matrix of all context-actions seen until the end of the batch $m \geq 1$. Eventually, $K_{\mathcal{S}}(s')$ is the kernel column vector $[k(s_1, s'), \dots, k(s_l, s')]^\top$ of size $|\mathcal{S}| = l$. $Y_m = [-y_{0,1}, \cdots - y_{0,n_0}, \cdots - y_{m,1}, \cdots - y_{m,n_m}]$ denotes the vector of concatenated rewards observed up until the end of the batch m.

At a batch m, a context $x_{m,i}$ is sampled for $i \in \{1, n_m\}$, and then to sample an action a, the following decision rule is applied:

$$a \in \arg\max_{a \in \mathcal{A}} \hat{q}_{m,i,a}. \tag{51}$$

In batch Kernel UCB, $\hat{q}_{m,i,a}$ is defined as

$$\hat{q}_{m,i,a} = \hat{m}_{m,i,a} + \beta_m \hat{\sigma}_{m,i,a}, \tag{52}$$

where

$$\hat{\mu}_{m,i,a} = K_{\mathcal{S}_{t-1}}\big((x_{m,i}, a)\big)^\top K_{m-1}^{-1} Y_{m-1}$$
$$\hat{\sigma}^2_{m,i,a} = \frac{1}{\lambda} k\big((x_{m,i}, a), (x_{m,i}, a)\big) - \frac{1}{\lambda} K_{\mathcal{S}_{m-1}}\big((x_{m,i}, a)\big)^\top K_{m-1}^{-1} K_{\mathcal{S}_{m-1}}\big((x_{m,i}, a)\big),$$

and β_m is a theoretical parameter that is set to $\beta_m = \frac{1}{\sqrt{m}}$ in practical heuristics (Lattimore & Szepesvari, 2019). In SBPE (Han et al.), $\hat{q}_{m,i,a}$ is defined directly as

$$\hat{q}_{m,i,a} = K_{\mathcal{S}_{t-1}}\big((x_{m,i}, a)\big)^\top K_{m-1}^{-1} Y_{m-1}. \tag{53}$$

Algorithm 3 Batch bandit - SBPE (Han et al.) and Kernel UCB (Valko et al., 2013)

Input: Logged observations $(x_{0,i}, a_{0,i}, y_{0,i}, \pi_{0,i})_{i=1,\dots,n_0}$, λ regularization and exploration parameters, k the kernel function initialization

$K_\lambda = [k(s_{0,i}, s_{0,j})]_{1\le i,j\le n_0} + \lambda I, Y_0 = [-y_{0,i}]_{1\le i\le n_0}$

for $m = 1$ *to* M **do**

  for $i = 1$ *to* n_m **do**

    Observe context $x_{i,m}$

    Choose $a_{i,m} \leftarrow \arg\max_{a\in\mathcal{A}} \hat{q}_{m,i,a}$ using Eq. (53) or (52)

  end

  Observe losses $y_{i,m}$ for all i in past batch $\{1, \dots, n_m\}$

  Update $Y_m \leftarrow [-y_{0,1}, \cdots - y_{0,n_0}, \cdots - y_{m,1}, \cdots - y_{m,n_m}]$

  Update the translated gram matrix $K_\lambda \leftarrow [k(s_{i,p}, s_{j,p})]_{1\le i,j\le n_p, 1\le p\le m} + \lambda I$

end

SBPE (Han et al.) uses a linear modelling, therefore we used a linear kernel. For the Kernel UCB (Valko et al., 2013) method, we used Gaussian and Polynomial kernels in our experiments. Note also that no regularization parameter λ is used in SBPE so we set $\lambda = 0$ in our experiments, and for K-UCB we chose λ in the grid $[1e0, 1e1, 1e2]$.

Note in particular that we adapted the batch bandit baselines to the CRM setting by benefiting the initialization with the logged dataset to set the gram matrix K_λ as well as the reward vector Y_0 with information from the logging data. This modification changes the original methods which take random actions at initializations.

Eventually, the baselines were carefully optimized using the Jax library (`https://github.com/google/jax`) to allow for just in time compilations of algebraic blocks in both methods and to maximize their scaling capacity.

RL baselines In order to compare our method to the two known off-policy online RL algorithm `PPO` (Schulman et al., 2017) and `TRPO` (Schulman et al., 2015), we do the following:

1. we use the `stable_baselines3`(Raffin et al., 2021) library for the implementation. When necessary we call multiple times the model `PPO` or `TRPO`, to have buffer size of geometrical increase.

2. we initialize the `ActorCriticPolicy` with a simpler MLP model having only one layer with output dimension of 1, (with argument `net_arch= [1]`, that is mathematically the same modelling as in CRM and SCRM baselines).

3. At the initial step only and to enable a fair comparison with counterfactual methods using a logging dataset, we pretrain the RL policies to imitate the actions sampled from the logging policy: we process by multiple step of the Adam optimizer, minimizing a loss being the sum of 2 terms:

   - a MSE term between the sampled action of the `ActorCriticPolicy` for the contexts in the n_0 instances, and the actions sampled by the logging policy.
   - the ENTROPY term guaranteeing to keep a minimum of exploration in order to initialize the RL algorithm $(-\sum p_i \log(p_i))$

4. we combine the 2 last terms with a linear combinaison with hyperparameters being tuned a posteriori, i.e. LOSS $=$ MSE $+ \lambda$ ENTROPY with the hyperparam $\lambda \in \{.5, 1, 2, 5, 10\}$

E. Additional empirical results

E.1. SCRM compared to CRM

We provide here the additional plot in the *Pricing* setting.

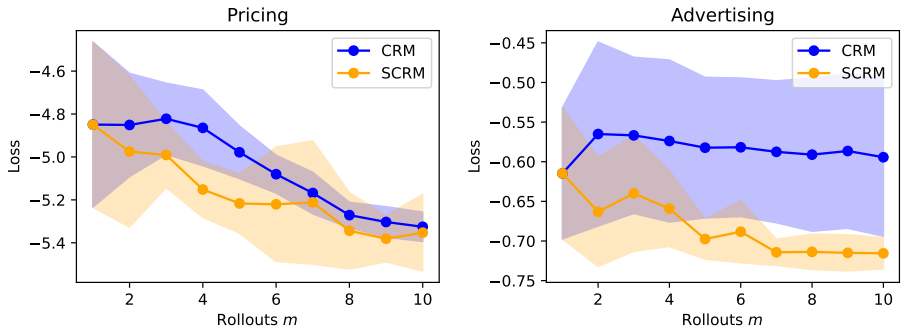

Figure 9: Test loss as a function of sample size on *Pricing*, *Advertising* (from left to right).

E.2. Evaluation of IPS-IX

We provide here the plots for the whole setting considered in policy evaluation with IPS-IX.

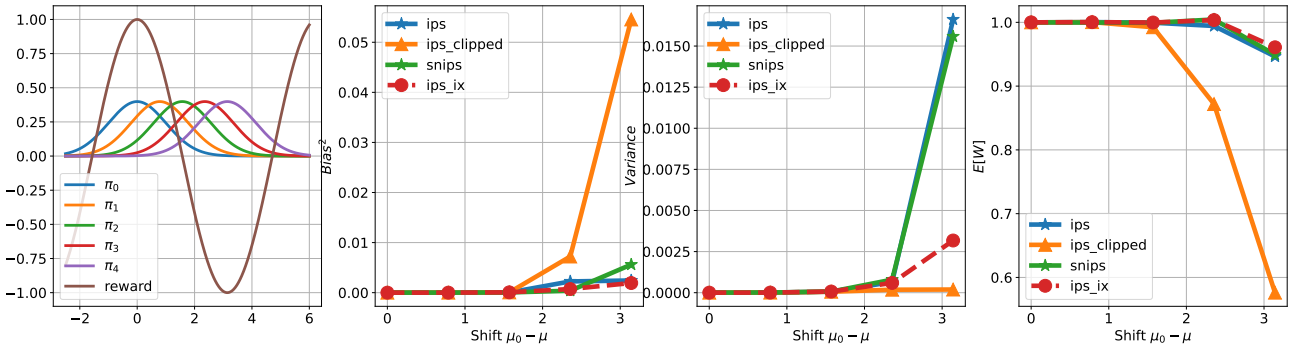

Figure 10: Comparison of IPS estimators on a Cosine reward and series of shifted Gaussian policies. Setup (left), Bias (middle left), Variance (middle right), Average IPS weight (right). IPS-IX shows a low bias and compares favorably to IPS and SNIPS in terms of variance.

E.3. Exploration/Exploitation tradeoff

In this part we give the details used for the experiment described in Section 6.3. We consider again Example 3.1 with the Gaussian parametrized policies $\pi_\theta = \mathcal{N}(\theta, \sigma^2)$ and a loss $l_t(a) = (a - y_t)^2 - 1$ where $y_t \sim \mathcal{N}(\theta^*, \sigma^{*2})$ with $\sigma^* = 0.3$. Recall that $\pi_{\theta_0} = \mathcal{N}(\theta_0, \sigma)$. We consider a grid of $\sigma \in [0.1, 0.3, 1, 3]$ and consider $\theta^* = 1$. Our experiment aims at illustrating the influence of sequential exploration that is an important detail of the SCRM and CRM principles.