Counterfactual Learning of Stochastic Policies with Continuous Actions: from Models to Offline Evaluation

Houssam Zenati, Alberto Bietti, Matthieu Martin, Eustache Diemert, Julien Mairal

Counterfactual Learning of Stochastic Policies with Continuous Actions: from Models to Offline Evaluation

Houssam Zenati[1,2], Alberto Bietti[3], Matthieu Martin[1], Eustache Diemert[1], and Julien Mairal[2]

[1]Criteo AI Lab
[2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[3]Center for Data Science, New York University. New York, USA

Abstract

Counterfactual reasoning from logged data has become increasingly important for many applications such as web advertising or healthcare. In this paper, we address the problem of learning stochastic policies with continuous actions from the viewpoint of counterfactual risk minimization (CRM). While the CRM framework is appealing and well studied for discrete actions, the continuous action case raises new challenges about modelization, optimization, and offline model selection with real data which turns out to be particularly challenging. Our paper contributes to these three aspects of the CRM estimation pipeline. First, we introduce a modelling strategy based on a joint kernel embedding of contexts and actions, which overcomes the shortcomings of previous discretization approaches. Second, we empirically show that the optimization aspect of counterfactual learning is important, and we demonstrate the benefits of proximal point algorithms and differentiable estimators. Finally, we propose an evaluation protocol for offline policies in real-world logged systems, which is challenging since policies cannot be replayed on test data, and we release a new large-scale dataset along with multiple synthetic, yet realistic, evaluation setups.[1]

Keywords— counterfactual learning, offline contextual bandits, continuous actions

1 Introduction

Logged interaction data is widely available in many applications such as drug dosage prescription [16], recommender systems [20], or online auctions [6]. An important task is to leverage past data in order to find a good *policy* for selecting actions (*e.g.*, drug doses) from available features (or *contexts*), rather than relying on randomized trials or sequential exploration, which may be costly to obtain or subject to ethical concerns.

More precisely, we consider offline logged bandit feedback data, consisting of contexts and actions selected by a given *logging policy*, associated to observed rewards. This is known as *bandit feedback*, since the reward is only observed for the action chosen by the logging policy. The problem of finding a good policy thus requires a form of *counterfactual* reasoning to estimate what the rewards would have been, had we used a different policy. When the logging policy is stochastic, one may obtain unbiased reward estimates under a new policy through importance sampling with inverse propensity scoring [IPS, 12]. One may then use this estimator or its variants for optimizing new policies without the need for costly experiments [6, 9, 32, 33], an approach also known as counterfactual risk minimization (CRM). While this setting is not sequential, we assume that learning a *stochastic* policy is required so that one may gather new exploration data after deployment.

In this paper, we focus on stochastic policies with continuous actions, which, unlike the discrete setting, have received little attention in the context of counterfactual policy optimization [8, 16, 7]. As noted by Kallus and Zhou [16] and as our experiments confirm, addressing the continuous case with naive discretization strategies performs poorly. Our first contribution is about *data modeling*: we introduce a joint embedding of actions and contexts relying

[1]The code with open-source implementations for experimental reproducibility is available at `https://github.com/criteo-research/optimization-continuous-action-crm`.

on kernel methods, which takes into account the continuous nature of actions, leading to rich classes of estimators that prove to be effective in practice.

In the context of CRM, the problem of *estimation* is intrinsically related to the problem of *optimization* of a non-convex objective function. In our second contribution, we underline the role of optimization algorithms [6, 33]. We believe that this aspect was overlooked, as previous work has mostly studied the effectiveness of estimation methods regardless of the optimization procedure. In this paper, we show that appropriate tools can bring significant benefits. To that effect, we introduce differentiable estimators based on soft-clipping the importance weights, which are more amenable to gradient-based optimization than previous hard clipping procedures [6, 34]. We also find that proximal point algorithms [27] tend to dominate simpler off-the-shelf optimization approaches, while keeping a reasonable computation cost.

Finally, an open problem in counterfactual reasoning is the difficult question of reliable *evaluation* of new policies based on logged data only. Despite significant progress thanks to various IPS estimators, we believe that this issue is still acute, since we need to be able to estimate the quality of policies and possibly select among different candidate ones *before* being able to deploy them in practice. Our last contribution is a small step towards solving this challenge, and consists of a new offline evaluation benchmark along with a new large-scale dataset, which we call CoCoA, obtained from a real-world system. The key idea is to introduce importance sampling diagnostics [24] to discard unreliable solutions along with significance tests to assess improvements to a reference policy. We believe that this contribution will be useful for the research community; in particular, we are not aware of similar publicly available large-scale datasets for continuous actions.

1.1 Organization of the paper

We present below the different sections of this paper and summarize the contributions in each part:

- Modeling of continuous action policies: we review the CRM framework and introduce our counterfactual cost predictor (CCP) parametrization which uses a joint embedding of contexts and actions to parametrize stochastic policies.
- Optimization perspectives for CRM: we introduce a differentiable and smooth clipping strategy for importance sampling weights and explain how proximal point algorithms can be used to better optimize CRM objectives.
- Evaluation on real data: we release an open and real-world dataset for the logged bandit feedback problem with continuous actions. Moreover, we introduce our offline evaluation protocol that tackles the underlooked problem of offline evaluation in logged bandits problems.
- Experimental results: we provide extensive experiments to validate our proposed offline protocol and to discuss the relevance of continuous action strategies, the performance of our CCP model, as well as the optimization strategies.

We then discuss open questions for future research directions, such as the difficulty of developping a doubly robust estimator based on our CCP model, and on the links between optimization strategies and offline evaluation methods.

1.2 Related Work

A large effort has been devoted to designing CRM estimators that have less variance than the IPS method, through clipping importance weights [6, 34], variance regularization [32], or by leveraging reward estimators through doubly robust methods [9, 26]. In order to tackle an overfitting phenomenon termed "propensity overfitting", Swaminathan and Joachims [33] also consider self-normalized estimators [24]. Such estimation techniques also appear in the context of sequential learning in contextual bandits [2, 18], as well as for off-policy evaluation in reinforcement learning [13]. In contrast, the setting we consider is not sequential.

While most approaches for counterfactual policy optimization tend to focus on discrete actions, few works have tackled the continuous action case, again with a focus on estimation rather than optimization. In particular, propensity scores for continuous actions were considered by Hirano and Imbens [11]. More recently, evaluation and optimization of continuous action policies were studied in a non-parametric context by Kallus and Zhou [16], and by Demirer et al. [8] in a semi-parametric setting.

In contrast to these previous methods, (i) we focus on stochastic policies while they consider deterministic ones, even though the kernel smoothing approach of Kallus and Zhou [16] may be interpreted as learning a deterministic policy perturbed by Gaussian noise. (ii) The terminology of *kernels* used by Kallus and Zhou [16] refers to a different mathematical tool than the kernel embedding used in our work. We use positive definite kernels to define a nonlinear

representation of actions and contexts in order to model the reward function, whereas Kallus and Zhou [16] use *kernel density estimation* to obtain good importance sampling estimates. Chen et al. [7] also use a kernel embedding of contexts in their policy parametrization, while our method jointly models contexts and actions. Moreover, their method requires computing an $n \times n$ Gram matrix, which does not scale with large datasets; in principle, it should be however possible to modify their method to handle kernel approximations such as the Nyström method [35]. Besides, their learning formulation with a quadratic problem is not compatible with CRM regularizers introduced by [32, 33] which would change their optimization procedure. Eventually, we note that Krause and Ong [17] use similar kernels to ours for jointly modeling contexts and actions, but in the different setting of sequential decision making with upper confidence bound strategies. (iii) While Kallus and Zhou [16] and Demirer et al. [8] focus on policy *estimation*, our work introduces a new continuous-action data *representation* and encompasses *optimization*: in particular, we propose a new contextual policy parameterization, which leads to significant gains compared to baselines parametrized policies on the problems we consider, as well as further improvements related to the optimization strategy. We also note that, apart from [8] that uses an internal offline cross-validation for model selection, previous works did not perform offline model selection nor evaluation protocols, which are crucial for deploying methods on real data. We provide a brief summary in Table 1 to summarize the key differences with our work.

Method	Stochastic	Policy Parametrization	Kernels	CRM Regularizers	Offline evaluation protocol	Large-scale
Chen et al. [7]	✗	Linear	Embedding of contexts	✗	✗	✗ \ ✓
Kallus and Zhou [16]	✗ \ ✓	Linear	Kernel Density Estimation	✓	✗	✓
Demirer et al. [8]	✗	Any	Not used	✗	✓	✓
Ours	✓	CCP	Joint embedding of contexts/actions	✓	✓	✓

Table 1: Comparison to [7, 16, 8], CCP refers to our continuous action model, see section 2.2. For discussions on stochastic interpretation of [16] and the application of [7] to large-scale data, see main text.

Optimization methods for learning stochastic policies have been mainly studied in the context of reinforcement learning through the policy gradient theorem [3, 31, 36]. Such methods typically need to observe samples from the new policy at each optimization step, which is not possible in our setting. Other methods leverage a form of off-policy estimates during optimization [15, 29], but these approaches still require to deploy learned policies at each step, while we consider objective functions involving only a fixed dataset of collected data. In the context of CRM, Su et al. [30] introduce an estimator with a continuous clipping objective that achieves an improved bias-variance trade-off over the doubly-robust strategy. Nevertheless, this estimator is non-smooth, unlike our soft-clipping estimator.

2 Modeling of Continous Action Policies

We now review the CRM framework, and then present our modelling approach for policies with continuous actions.

2.1 Background

For a stochastic policy π over a set of actions $\mathcal{A}$, a contextual bandit environment generates i.i.d. context features $x \sim \mathcal{P}_{\mathcal{X}}$ in $\mathcal{X}$ for some probability distribution $P_{\mathcal{X}}$, actions $a \sim \pi(\cdot|x)$ and costs $y \sim \mathcal{P}_{\mathcal{Y}}(\cdot|x, a)$ for some conditional probability distribution $P_{\mathcal{Y}}$. We denote the resulting distribution over triplets (x, a, y) as $\mathcal{P}_\pi$.

Then, we consider a logged dataset $(x_i, a_i, y_i)_{i=1,\dots,n}$, where we assume $(x_i, a_i, y_i) \sim \mathcal{P}_{\pi_0}$ i.i.d. for a given stochastic logging policy π_0, and we assume the propensities $\pi_{0,i} := \pi_0(a_i|x_i)$ to be known. In the continuous action case, we assume that the probability distributions admit densities and thus propensities denote the density function of π_0 evaluated on the actions given a context. The expected loss or risk of a policy π is then given by

$$L(\pi) = \mathbb{E}_{(x,a,y)\sim\mathcal{P}_\pi}[y]\,. \tag{1}$$

For the logged bandit, the task is to determine a policy $\hat{\pi}$ in a set of *stochastic* policies Π with small risk. Note that this definition may also include deterministic policies by allowing Dirac measures, unless Π includes a specific constraint *e.g.*, minimum variance, which may be desirable in order to gather data for future offline experiments.

In our setting, the expectation in (1) cannot be computed directly for any π, as the available interaction data comes from a different distribution $\mathcal{P}_{\pi_0}$. Yet, multiple empirical estimators of the risk hereafter allow to derive an

empirical optimal policy that is found by solving

$$\hat{\pi} \in \arg\min_{\pi \in \Pi} \hat{L}(\pi) + \Omega(\pi), \tag{2}$$

where the objective consists of an empirical estimate $\hat{L}$ of the risk and of possible data-dependent regularizers, denoted by Ω. When using counterfactual estimators for $\hat{L}$, this method has been called *counterfactual risk minimization* [32].

The counterfactual approach tackles the distribution mismatch between the logging policy $\pi_0(\cdot|x)$ and a policy π in Π via importance sampling. The IPS method [12] relies on correcting the distribution mismatch using the well-known relation

$$L(\pi) = \mathbb{E}_{(x,a,y)\sim\mathcal{P}_{\pi_0}} \left[y \frac{\pi(a|x)}{\pi_0(a|x)} \right], \tag{3}$$

assuming π_0 has non-zero mass on the support of π, which allows us to derive an unbiased empirical estimate

$$\hat{L}_{\mathrm{IPS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} y_i \frac{\pi(a_i|x_i)}{\pi_{0,i}}. \tag{4}$$

Clipped estimator. Since the empirical estimator $\hat{L}_{\mathrm{IPS}}(\pi)$ may suffer from large variance and is subject to various overfitting phenomena, regularization strategies have been proposed. In particular, this estimator may overfit negative feedback values y_i for samples that are unlikely under π_0 (see motivation for clipped estimators in Appendix 7), resulting in higher variances. Clipping the importance sampling weights in Eq. (5) as Bottou et al. [6] mitigates this problem, leading to a clipped (cIPS) estimator

$$\hat{L}_{\mathrm{cIPS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} y_i \min \left\{ \frac{\pi(a_i|x_i)}{\pi_{0,i}}, M \right\}. \tag{5}$$

Smaller values of M reduce the variance of $\hat{L}(\pi)$ but induce a larger bias. Swaminathan and Joachims [32] also propose adding an empirical variance penalty term controlled by a factor λ to the empirical risk $\hat{L}(\pi)$. Specifically, they write $\nu_i(\pi) = y_i \min\left(\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}, M\right)$ and consider the empirical variance for regularization:

$$\hat{V}(\pi) = \frac{1}{n-1} \sum_{i=1}^{n} (\nu_i(\pi) - \bar{\nu}(\pi))^2, \qquad \text{with} \quad \bar{\nu}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \bar{\nu}_i(\pi),$$

which is subsequently used to obtain a regularized objective $\mathcal{L}$ with hyperparameters M and λ for clipping and variance penalization, respectively:

$$\mathcal{L}(\pi) = \hat{L}_{\mathrm{cIPS}}(\pi) + \lambda \hat{V}(\pi). \tag{6}$$

The self-normalized estimator. Swaminathan and Joachims [33] also introduce a regularization mechanism for tackling the so-called *propensity overfitting* issue, occuring with rich policy classes, where the method would focus only on maximizing (resp. minimizing) the sum of ratios $\pi(a_i|x_i)/\pi_{0,i}$ for negative (resp. positive) costs. This effect is corrected through the following *self-normalized importance sampling* (SNIPS) estimator [24, see also], which is equivariant to additive shifts in cost values:

$$\hat{L}_{\mathrm{SNIPS}}(\pi) = \frac{\sum_{i=1}^{n} y_i w_i^{\pi}}{\sum_{i=1}^{n} w_i^{\pi}}, \quad \text{with} \quad w_i^{\pi} = \frac{\pi(a_i|x_i)}{\pi_{0,i}}. \tag{7}$$

The direct method DM and the optimal greedy policy. An important quantity is the expected cost given actions and context, denoted by $\eta^*(x,a) = \mathbb{E}[y|x,a]$. If this expected cost was known, an optimal (deterministic) greedy policy π^* would simply select actions that minimize the expected cost

$$\pi^*(x) = \arg\min_{a \in \mathcal{A}} \eta^*(x,a). \tag{8}$$

It is then tempting to use the available data to learn an estimator $\hat{\eta}(x,a)$ of the expected cost, for instance by using ridge regression to fit $y_i \approx \hat{\eta}(x_i,a_i)$ on the training data. Then, we may use the deterministic greedy policy $\hat{\pi}_{\mathrm{DM}}(x) = \arg\min_a \hat{\eta}(x,a)$. This approach, termed *direct method* (DM), has the benefit of avoiding the high-variance

problems of IPS-based methods, but may suffer from large bias since it ignores the potential mismatch between $\hat{\pi}_{\mathrm{DM}}$ and π_0. Specifically, the bias is problematic when the logging policy provides unbalanced data samples (*e.g.*, only samples actions in a specific part of the action space) leading to overfitting [6, 9, 33]. Conversely, counterfactual methods re-balance these generated data samples with importance weights and mitigate the distribution mismatch to better estimate reward function on less explored actions. Nevertheless, such cost estimators can be sometimes effective in practice and may be used to improve IPS estimators in the so-called doubly robust (DR) estimator [9] by applying IPS to the residuals $y_i - \hat{\eta}(x_i, a_i)$, thus using $\hat{\eta}$ as a control variate to decrease the variance of IPS.

While such greedy deterministic policies may be sufficient for exploitation, stochastic policies may be needed in some situations, for instance when one wants to still encourage some exploration in a future round of data logs. Using a stochastic policy also allows us to obtain more accurate off-policy estimates when performing cross-validation on logged data. Then, it may be possible to define a stochastic version of the direct method by adding Gaussian noise with variance σ^2:

$$\hat{\pi}_{\mathrm{SDM}}(\cdot|x) = \mathcal{N}(\hat{\pi}_{\mathrm{DM}}(x), \sigma^2), \tag{9}$$

In the context of offline evaluation on bandit data, such a smoothing procedure may also be seen as a form of kernel smoothing for better estimation [16].

2.2 Our Model for Continuous Actions Policies: the Counterfactual Cost Predictor (CCP)

When considering continuous action spaces, the choice of policies is more involved than in the discrete case. One may indeed naively discretize the action space into buckets and leverage discrete action strategies, but then local information within each bucket gets lost and it is non-trivial to choose an appropriate bucketization of the action space based on logged data, which contains non-discrete actions.

In this paper, we thus focus on stochastic policies belonging to certain classes of continuous distributions, such as normal or log-normal. Specifically, we consider a set of context-dependent policies $\Pi = \{\pi_\theta, \theta \in \Theta\}$ parameterized by $\theta = (\mu(x), \sigma)$, with context-dependent mean $\mu(x)$ and standard deviation σ. For example, for normal distributions, $\pi_\theta(\cdot|x) = \mathcal{N}(\mu(x), \sigma^2)$. This class of distributions require learning a parametrized model of the mean $\mu(x)$. Before introducing a more flexible model, we will consider the following baselines, given a context x in $\mathbb{R}^d$:

- *constant*: $\mu(x) = b$ (context-independent);
- *linear*: $\mu(x) = \langle x, \beta \rangle + b$ with β in $\mathbb{R}^d$;
- *poly*: $\mu(x) = x^\top B x + b = \langle xx^\top, B \rangle + b$ with B in $\mathbb{R}^{d \times d}$.

These baselines require learning the parameters b, β or B by using the CRM approach. Intuitively, the goal is to find a stochastic policy that is close to the optimal deterministic one from Eq. (8). While these approaches can be effective in simple problems, they may be limited in more difficult scenarios where the expected cost $\eta^*(x, a)$ has a complex behavior as a function of a. This motivates the need for classes of policies which can better capture such variability by considering a joint model $\eta(x, a)$ of the cost.

The counterfactual cost predictor (CCP). Assuming that we are given such a model $\eta(x, a)$, which we call *cost predictor*, we parametrize the mean of a stochastic policy by using a soft-argmin operator with temperature γ:

$$\mu_{\mathrm{CCP}}(x) = \sum_{i=1}^{m} a_i \frac{\exp(-\gamma \eta(x, a_i))}{\sum_{j=1}^{m} \exp(-\gamma \eta(x, a_j))}, \tag{10}$$

where $a_1, \ldots, a_m$ are anchor points (*e.g.*, a regular grid or quantiles of the action space), and μ_{CCP} may be viewed as a smooth approximation of a greedy policy $\mu_{\mathrm{greedy}}(x) = \arg\min_a \eta(x, a)$. The motivation for introducing a soft-argmin operator is to avoid the optimization over actions and to make the resulting CRM problem differentiable. This allows CCP policies to capture complex behavior of η^* as a function of a.

Cost predictors based on kernels. The above CCP model is parameterized by a model $\eta(x, a)$ that may be interpreted as a cost predictor. In a continuous action problem, a reasonable assumption is that costes vary smoothly as a function of actions. Then, a good choice is to take η in a space of smooth functions such as the reproducing kernel Hilbert space $\mathcal{H}$ (RKHS) defined by a positive definite kernel [28], so that one may control the smoothness of η through regularization with the RKHS norm. More precisely, we consider kernels of the form

$$K((x, a), (x', a')) = \langle \psi_{\mathcal{X}}(x), \psi_{\mathcal{X}}(x') \rangle e^{-\frac{\alpha}{2}\|a - a'\|^2}, \tag{11}$$

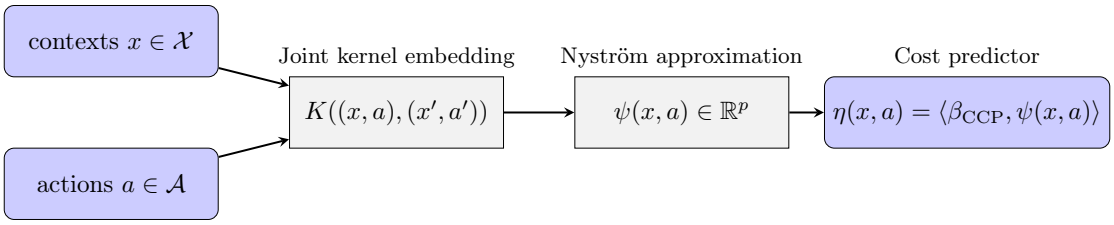

Figure 1: Illustration of the joint kernel embedding for the counterfactual cost predictor (CCP) and cost estimator.

where, for simplicity, $\psi_{\mathcal{X}}(x)$ is either a linear embedding $\psi_{\mathcal{X}}(x) = x$ or a quadratic one $\psi_{\mathcal{X}}(x) = (xx^T, x)$, while actions are compared via a Gaussian kernel, allowing to model complex interactions between contexts and actions.

Nyström method and explicit embedding. Since traditional kernel methods lack scalability, we rely on the classical Nyström approximation [35] of the Gaussian kernel, which provides us a finite-dimensional approximate embedding $\psi_{\mathcal{A}}(a)$ in $\mathbb{R}^m$ such that $e^{-\frac{\alpha}{2}\|a-a'\|^2} \approx \langle \psi_{\mathcal{A}}(a), \psi_{\mathcal{A}}(a') \rangle$ for all actions a, a'. This allows us to build a finite-dimensional embedding

$$\psi(x,a) = \psi_{\mathcal{X}}(x) \otimes \psi_{\mathcal{A}}(a),$$

where $\otimes$ denotes the tensorial product, such that

$$\begin{aligned} K((x,a),(x',a')) &\approx \langle \psi_{\mathcal{X}}(x), \psi_{\mathcal{X}}(x') \rangle \langle \psi_{\mathcal{A}}(a), \psi_{\mathcal{A}}(a') \rangle \\ &= \langle \psi(x,a), \psi(x',a') \rangle. \end{aligned}$$

More precisely, Nyström's approximation consists of projecting each point from the RKHS to a m-dimensional subspace defined as the span of m anchor points, representing here the mapping to the RKHS of m actions $\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_m$. In practice, we may choose $\bar{a}_i$ to be equal to the a_i in (10), since in both cases the goal is to choose a set of "representative" actions. For one-dimensional actions ($\mathcal{A} \subseteq \mathbb{R}$), it is reasonable to consider a uniform grid, or a non-uniform ones based on quantiles of the empirical distribution of actions in the dataset. In higher dimensions, one may simply use a K-means algorithms and assign anchor points to centroids.

From an implementation point of view, Nyström's approximation considers the embedding $\psi_{\mathcal{A}}(a) = K_{AA}^{-1/2} K_A(a)$, where $K_{AA} = [K_{\mathcal{A}}(\bar{a}_i, \bar{a}_j)]_{ij}$ and $K_A(a) = [K_{\mathcal{A}}(a, \bar{a}_i)]_i$ and $K_{\mathcal{A}}$ is the Gaussian kernel. Then, using the kernel K in (11) yields a model η parametrized by a vector β_{CCP} in

$$\eta(x,a) = \langle \beta_{\mathrm{CCP}}, \psi(x,a) \rangle,$$

and ℓ^2 regularization on β_{CCP} corresponds to controlling the RKHS norm and the smoothness of η. The anchor points that we use can be seen as the parameters of an *interpolation* strategy defining a smooth function, similar to knots in spline interpolation. Naive discretization strategies would prevent us from exploiting such a smoothness assumption on the cost with respect to actions and from exploiting the structure of the action space. Note that Section 5 provides a comparison with naive discretization strategies, showing important benefits of the kernel approach. Our goal was to design a stochastic, computationally tractable, differentiable approximation of the optimal (but unknown) greedy policy (8). The model is illustrated in Figure 1.

3 On Optimization Perspectives for CRM

Because our models yield non-convex CRM problems, we believe that it is crucial to study optimization aspects. Here, we introduce a differentiable clipping strategy for importance weights, and discuss optimization algorithms.

3.1 Soft Clipping IPS

The classical hard clipping estimator

$$\hat{L}_{\mathrm{cIPS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} y_i \min\left\{ \pi(a_i|x_i)/\pi_{0,i}, M \right\} \tag{12}$$

makes the objective function non-differentiable, and yields terms in the objective with clipped weights to have zero gradient. In other words, a trivial stationary point of the objective function is that of a stochastic policy that differs enough from the logging policy such that all importance weights are clipped. To alleviate this issue, we propose a differentiable logarithmic soft-clipping strategy. Given a threshold parameter $M \geq 0$ and an importance weight $w_i = \pi(a_i|x_i)/\pi_{0,i}$, we consider the soft-clipped weights:

$$\zeta(w_i, M) = \begin{cases} w_i & \text{if } w_i \leq M \\ \alpha(M) \log\left(w_i + \alpha(M) - M\right) & \text{otherwise,} \end{cases} \tag{13}$$

where $\alpha(M)$ is such that $\alpha(M)\log(\alpha(M)) = M$, which yields a differentiable operator. We illustrate the soft clipping expression in Figure 2 and give further explanations about the benefits of clipping strategies in Appendix 7.

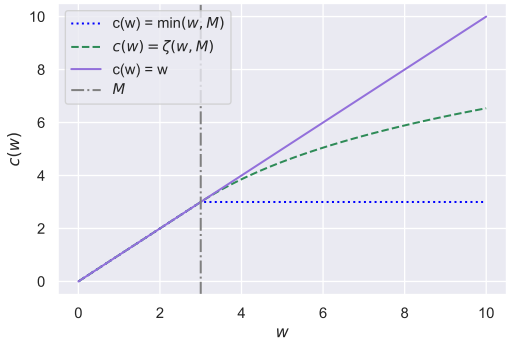

Figure 2: Different clipping strategies c on the importance weights w. Weights are clipped for $M = 3$, the hard clipping $c(w) = \min(w, M)$ provides no gradient for $w > M$, while the soft clipping $c(w) = \zeta(w, M)$ and the unclipped estimators $c(w) = w$ do.

Then, the IPS estimator with soft clipping becomes

$$\hat{L}_{\text{scIPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} y_i \zeta\left(\frac{\pi(a_i|x_i)}{\pi_{0,i}}, M\right). \tag{14}$$

We now provide a similar generalization bound to that of Swaminathan and Joachims [32] (for the hard-clipped version) for the variance-regularized objective of this soft-clipped estimator, justifying its use as a good optimization objective for minimizing the expected risk. Writing $\chi_i(\pi) = y_i \zeta\left(\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}, M\right)$, we recall the empirical variance with scIPS that is used for regularization:

$$\hat{V}(\pi) = \frac{1}{n-1}\sum_{i=1}^{n}(\chi_i(\pi) - \bar{\chi}(\pi))^2, \qquad \text{with} \quad \bar{\chi}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\bar{\chi}_i(\pi).$$

We assume that costs $y_i \in [-1, 0]$ almost surely, as in [32], and make the additional assumption that the importance weights $\pi(a_i|x_i)/\pi_0(a_i|x_i)$ are upper bounded by a constant W almost surely for all $\pi \in \Pi$. This is satisfied, for instance, if all policies have a given compact support (*e.g.*, actions are constrained to belong to a given interval) and π_0 puts mass everywhere in this support.

Proposition 3.1 (Generalization bound for $\hat{R}^M_{\text{scIPS}}$)**.** *Assume costs y_i in $[-1, 0]$ and importance weights bounded by W. With probability $1 - \delta$, for all π in Π,*

$$L(\pi) \leq \hat{L}_{scIPS}(\pi) + O\left(\sqrt{\frac{\hat{V}(\pi)C_n(\Pi, \delta)}{n}} + \frac{SC_n(\Pi, \delta)}{n}\right),$$

where $S = \zeta(W, M) = O(\log W)$, $\hat{V}$ denotes the empirical variance of the cost estimates, and $C_n(\Pi, \delta)$ is a complexity measure of the policy class.

We now prove the Proposition 3.1.

Proof. We begin by proving the result for a finite policy class and for which we use $C_n(\Pi, \delta) = \log(2|\Pi|/\delta)$. Let

$$f_i(\pi) = 1 + \frac{y_i}{S}\zeta\left(\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}, M\right).$$

We have $f_i(\pi) \in [0,1]$ and $\mathbb{E}_{(x_i,a_i,y_i)\sim\mathcal{D}_{\pi_0}}[f_i(\pi)] = 1 + R^M(\pi)/S$, with

$$L^M(\pi) = \mathbb{E}_{(x,a,y)\sim\mathcal{D}_{\pi_0}}\left[y\zeta\left(\frac{\pi(a|x)}{\pi_0(a|x)}, M\right)\right].$$

We can now apply the concentration bound of Maurer and Pontil [22, Corollary 5] to the f_i and rescale appropriately by S to obtain that with probability $1-\delta$, for all $\pi \in \Pi$,

$$L^M(\pi) \leq \hat{L}_{\text{scIPS}}(\pi) + \sqrt{\frac{2\hat{V}(\pi)\log(2|\Pi|/\delta)}{n}} + S\frac{7\log(2|\Pi|/\delta)}{3(n-1)}.$$

We then use that $L(\pi) \leq L^M(\pi)$, since $\ell(a) \leq 0$ and $\zeta(w, M) \leq w$ for all w.

We then conclude the proof by noting that the above result above can be extended to infinite policy classes by essentially replacing $|\Pi|$ above with an ℓ_∞ covering number of the set $\{(f_1(\pi), \ldots, f_n(\pi)) : \pi \in \Pi\}$, by leveraging Maurer and Pontil [22, Theorem 6] as in [32].

□

Note that while the bound requires importance weights bounded by a constant W, the bound only scales logarithmically with W when $W \gg M$, compared to a linear dependence for IPS. However we gain significant benefits in terms of optimization by having a smooth objective.

Remark 3.1. *If costs are in the range $[-c, 0]$, the constant S can be replaced by cS, making the bound homogeneous in the scale (indeed, the variance term is also scaled by c).*

Remark 3.2. *The scIPS estimator is less biased than the cIPS. We emphasize however that the main motivation for such a clipping strategy is to provide a differentiable estimator which is not the case for cIPS in areas where all point are clipped.*

3.2 Proximal Point Algorithms

Non-convex CRM objectives have been optimized with classical gradient-based methods [32, 33] such as L-BFGS [21], or the stochastic gradient descent approach [14]. Proximal point methods are classical approaches originally designed for convex optimization [27], which were then found to be useful for nonconvex functions [10, 25]. In order to minimize a function $\mathcal{L}$, the main idea is to approximately solve a sequence of subproblems that are better conditioned than $\mathcal{L}$, such that the sequence of iterates converges towards a better stationary point of $\mathcal{L}$. More precisely, the proximal point method consists of computing a sequence

$$\theta_k \approx \arg\min_\theta \left(\mathcal{L}(\theta) + \frac{\kappa}{2}\|\theta - \theta_{k-1}\|_2^2\right), \tag{15}$$

where $\mathcal{L}(\theta) = \hat{L}(\pi_\theta) + \Omega(\pi_\theta)$ and $\kappa > 0$ is a constant parameter. The regularization term Ω often penalizes the variance [33], see Appendix 9.2. The role of the quadratic function in (15) is to make subproblems "less nonconvex" and for many machine learning formulations, it is even possible to obtain convex sub-problems with large enough κ [see 25].

In this paper, we consider such a strategy (15) with a parameter κ, which we set to zero only for the last iteration. Note that the effect of the proximal point algorithm differs from the proximal policy optimization (PPO) strategy used in reinforcement learning [29], even though both approaches are related. PPO encourages a new stochastic policy to be close to a previous one in Kullback-Leibler distance. Whereas the term used in PPO modifies the objective function (and changes the set of stationary points), the proximal point algorithm optimizes and finds a stationary point of the original objective $\mathcal{L}$, even with fixed κ.

The proximal point algorithm (PPA) introduces an additional computational cost as it leads to solving multiple sub-problems instead of a single learning problem. In practice for 10 PPA iterations and with the L-BFGS solver,

the computational overhead was about 3× in comparison to L-BFGS without PPA. This overhead seems to be the price to pay to improve the test reward and obtain better local optima, as we show in the experimental section 5.2.2. Nevertheless, we would like to emphasize that computational time is often not critical for the applications we consider, since optimization is performed offline.

4 On Evaluation and Model Selection for Real World Data

The CRM framework helps finding solutions when online experiments are costly, dangerous or raising ethical concerns. As such it needs a reliable validation and evaluation procedure before rolling-out any solution in the real world. In the continuous action domain, previous work have mainly considered semi-simulated scenarios [5, 16], where contexts are taken from supervised datasets but rewards are synthetically generated. To foster research on practical continuous policy optimization, we release a new large-scale dataset called CoCoA, which to our knowledge is the first to provide logged exploration data from a real-world system with continuous actions. Additionally, we introduce a benchmark protocol for reliably evaluating policies using off-policy evaluation.

4.1 The CoCoADataset

The CoCoAdataset comes from the Criteo online advertising platform which ran an experiment involving a randomized, continuous policy for real-time bidding. Data has been properly anonymized so as to not disclose any private information. Each sample represents a bidding opportunity for which a multi-dimensional context x in $\mathbb{R}^d$ is observed and a continuous action a in $\mathbb{R}^+$ has been chosen according to a stochastic policy π_0 that is logged along with the reward $-y$ (meaning cost y) in $\mathbb{R}$. The reward represents an advertising objective such as sales or visits and is jointly caused by the action and context (a, x). Particular care has been taken to guarantee that each sample $(x_i, a_i, \pi_0(a_i|x_i), y_i)$ is independent. The goal is to learn a contextual, continuous, stochastic policy $\pi(a|x)$ that generates more reward in expectation than π_0, evaluated offline, while keeping some exploration (stochastic part). As seen in Table 2, a typical feature of this dataset is the high variance of the cost ($\mathbb{V}[Y]$), motivating the scale of the dataset N to obtain precise counterfactual estimates. The link to download the dataset is available in the code repository: `https://github.com/criteo-research/optimization-continuous-action-crm`.

Table 2: CoCoAdataset summary statistics.

N	d	$\mathbb{E}[-Y]$	$\mathbb{V}[Y]$	$\mathbb{V}[A]$	$\mathbb{P}(Y \neq 0)$
120.10^6	3	11.37	9455	.01	.07

4.2 Evaluation Protocol for Logged Data

In order to estimate the test performance of a policy on real-world systems, off-policy evaluation is needed, as we only have access to logged exploration data. Yet, this involves in practice a number of choices and difficulties, the most documented being i) potentially infinite variance of IPS estimators [6] and ii) propensity over-fitting [32, 33]. The former implies that it can be difficult to accurately assess the performance of new policies due to large confidence intervals, while the latter may lead to estimates that reflect large importance weights rather than rewards. A proper evaluation protocol should therefore guard against such outcomes.

A first, structuring choice is the IPS estimator. While variants of IPS exist to reduce variance, such as clipped IPS, we found Self-Normalized IPS [SNIPS, 33, 19, 24, 23] to be more effective in practice. Indeed, it avoids the choice of a clipping threshold, generally reduces variance and is equivariant with respect to translation of the reward.

A second component is the use of importance sampling diagnostics to prevent propensity over-fitting. Lefortier et al. [19] propose to check if the empirical average of importance weights deviates from 1. However, there is no precise guideline based on this quantity to reject estimates. Instead, we recommend to use a diagnostic on the *effective sample size* $n_{\text{eff}} = (\sum_{i=1}^n w_i)^2 / \sum_{i=1}^n w_i^2$, which measures how many samples are actually usable to perform estimation of the counterfactual estimate; we follow Owen [24], who recommends to reject the estimate when the relative effective sample size n_{eff}/n is less than 1%. A third choice is a statistical decision procedure to check if $L(\pi) < L(\pi_0)$. In theory, any statistical test against a null hypothesis H_0: $L(\pi) \geq L(\pi_0)$ with confidence level $1 - \delta$ can be used.

Finally, we present our protocol in Algorithm 1. Since we cannot evaluate such a protocol on purely offline data, we performed an empirical evaluation on synthetic setups where we could analytically design true positive ($L(\pi) < L(\pi_0)$)

Algorithm 1: Evaluation Protocol

Input: $1-\delta$: confidence of statistical test (def: 0.95); ν: a max deviance ratio for effective sample size (def: 0.01);

Output: counter-factual estimation of $L(\pi)$ and decision to reject the null hypothesis H_0: $L(\pi) \geq L(\pi_0)$.

1. Split dataset $\mathcal{D} \mapsto \mathcal{D}^{\text{train}}, \mathcal{D}^{\text{valid}}, \mathcal{D}^{\text{test}}$
2. Train new policy π on $\mathcal{D}^{\text{train}}$ and tune hyper-parameters on $\mathcal{D}^{\text{valid}}$ (can be replaced by internal cross-validation)
3. Estimate effective sample size $n_{\text{eff}}(\pi)$ on $\mathcal{D}^{\text{valid}}$

if $\frac{n_{eff}}{n} > \nu$ **then**

  Estimate $\hat{L}_{\text{SNIPS}}(\pi)$ on $\mathcal{D}^{\text{test}}$ and test $\hat{L}_{\text{SNIPS}}(\pi) < \hat{L}(\pi_0)$ on $\mathcal{D}^{\text{test}}$ with confidence $1-\delta$.

else

  Keep H_0, consider the estimate to be invalid.

end

and true negative policies. We discuss in Section 5 the concrete parameters of Alg. 1 and their influence on false (non-)discovery rates in practice.

Model Selection with the Offline Protocol In order to make realistic evaluations, hyperparameter selection is always conducted by estimating the loss of a new policy π in a counterfactual manner. This requires using a validation set (or cross-validation) with propensities obtained from the logging policy π_0 of the training set. Such estimates are less accurate than online ones, which would require to gather new data obtained from π, which we assume is not feasible in real-world scenarios.

To solve this issue, we have chosen to discard unreliable estimates that do not pass the effective sample size test from Alg. 1. When doing cross-validation, it implies discarding folds that do not pass the test, and averaging estimates computed on the remaining folds. Although this induces a bias in the cross-validation procedure, we have found it to significantly reduce the variance and dramatically improve the quality of model selection when the number of samples is small, especially for the Warfarin dataset in Sec. 5.

5 Experimental Setup and Evaluation

We first provide an empirical evaluation of our offline protocol and proceed with a presentation of our empirical findings on synthetic datasets that allow evaluation for any test policy, before presenting our results including the CoCoAdataset.

5.1 Experimental Validation of the Protocol

In this section, we study the ability of Alg. 1 to accurately decide if a candidate policy π is better than a reference logging policy π_0 (condition $L(\pi) \leq L(\pi_0)$) on synthetic data. Here we simulate logging policy π_0 being a lognormal distribution of known mean and variance, and an optimal policy π^* being a Gaussian distribution. We generate a logged dataset by sampling actions $a \sim \pi_0$ and trying to evaluate policies $\hat{\pi}$ with costs observed under the logging policy. We compare the costs predicted using IPS and SNIPS offline metrics to the online metric as the setup is synthetic, it is then easy to check that indeed they are better or worse than π_0. We compare the IPS and SNIPS estimates along with their level of confidences and the influence of the effective sample size diagnostic. Offline evaluations of policies $\hat{\pi}$ illustrated in Figure 3 are estimated from logged data $(x_i, a_i, y_i, \pi_0)_{i=1\ldots n}$ where $a_i \sim \pi_0(\cdot|x_i)$ and where the policy risk would be optimal under the oracle policy π^*.

While the goal of counterfactual learning is to find a policy $\hat{\pi}$ which is as close a possible to the optimal policy π^*, based on samples drawn from a logging policy π_0, it is in practice hard to assess the statistical significance of a policy that is too "far" from the logging policy. Offline importance sampling estimates are indeed limited when the distribution mismatch between the evaluated policy and the logging policy (in terms of KL divergence $D_{KL}(\pi_0||\hat{\pi})$) is large. Therefore we create a setup where we evaluate the quality of offline estimates for policies (i) "close" to the logging policy (meaning the KL divergence $D_{KL}(\pi_0||\hat{\pi})$ is low) and (ii) "close" to the oracle optimal policy (meaning

Figure 3: Illustration of policies: logging policy π_0, optimal π^* and example policy $\hat{\pi}$.

the KL divergence $D_{KL}(\pi^*||\hat{\pi})$ is low). In this experiment, we focus on evaluating the ability of the offline protocol to correctly assess whether $L(\hat{\pi}) \leq L(\pi_0)$ or not by comparing to online truth estimates. Specifically, for both setups (i) and (ii), we compare the number of False Positives (FP) and False Negatives (FN) of the two offline protocols for $N = 2000$ initializations, by adding Gaussian noise to the parameters of the closed form policies. False negatives are generated when the offline protocol keeps $H_0 : L(\pi) \geq L(\pi_0)$ while the online evaluation reveals that $L(\pi) \leq L(\pi_0)$, while false positives are generated in the opposite case when the protocol rejects H_0 while it is true. We also show histograms of the differences between online and offline boundary decisions for $(L(\hat{\pi}) < L(\pi_0))$, using bootstrapped distribution of SNIPS estimates to build confidence intervals.

Validation of the use of SNIPS estimates for the offline protocol. To assess the performance of our evaluation protocol, we first compare the use of IPS and SNIPS estimates for the offline evaluation protocol and discard solutions with low importance sampling diagnostics $\frac{n_{\text{eff}}}{n} < \nu$ with the recommended value $\nu = 0.01$ from [24]. In Table 3, we provide an analysis of false positives and false negatives in both setups. We first observe that for setup (i) the SNIPS estimates has both fewer false positives and false negatives. Note that is setup is probably more realistic for real-world applications where we want to ensure incremental gains over the logging policy. In setup (ii) where importance sampling is more likely to fail when the evaluated policy is too "far" from the logging policy, we observe that the SNIPS estimate has a drastically lower number of false negatives than the IPS estimate, though it slightly has more false positives, thus illustrating how conservative this estimator is.

Table 3: Comparison of false positives and false negatives: Perturbation to the logging policy π_0 (setup (i)) and perturbation to the optimal policy (setup (ii)). The SNIPS estimator yields less FN and FP on setup (i), while being more effective on setup (ii) as well by inducing a drastically lower FP rate than IPS and a low FN rate. The effective sample size threshold is fixed at $\nu = 0.01$

Offline Protocol		Setup (i)				Setup (ii)			
		IPS		SNIPS		IPS		SNIPS	
		$\hat{\pi} \succeq \pi_0$	Keep H_0	$\hat{\pi} \succeq \pi_0$	Keep H_0	$\hat{\pi} \succeq \pi_0$	Keep H_0	$\hat{\pi} \succeq \pi_0$	Keep H_0
"Truth"	$\hat{\pi} \succeq \pi_0$	1282	24	1296	**10**	1565	67	1631	**1**
	Keep H_0	19	675	**0**	694	**0**	368	6	362

We then provide in Fig. 4 histograms of the differences of the upper boundary decisions between online estimates and bootstrapped offline estimates over all samples for both setups (i, left) and (ii, right). Both histograms illustrate how the IPS estimate underestimates the value of the reward with regard to the online estimate, unlike the SNIPS estimates. In the setup (ii) in particular, the IPS estimate underestimates severely the reward, which may explain why IPS has lower number of false positives when the evaluated policy is far from the logging policy. However in both setups, IPS has a higher number of false negatives. We also observed that our SNIPS estimates were highly correlated to the true (online) reward (average correlation $\rho = .968$, 30% higher than IPS, see plots in Appendix 8.1) for the synthetic setups presented in section 5.2.1, which therefore confirms our findings.

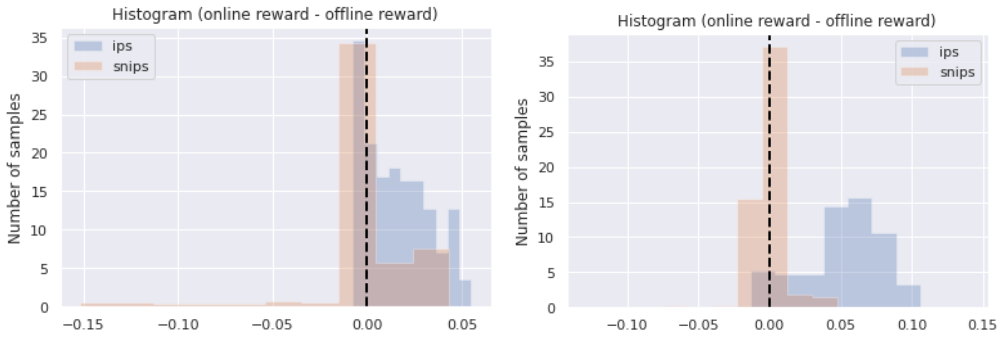

Figure 4: **Histogram of differences between online reward and offline lower confidence bound.** Perturbation to the logging policy π_0 (left), perturbation to the optimal policy π^* (right). Effective sample size threshold $\nu = 0.01$

Influence of the effective sample size criteria in the evaluation protocol In this setup we vary the effective sample size (ESS) threshold and show in Fig. 5 how it influences the performance of the offline evaluation protocol for the two previously discussed setups where we consider evaluations of (i) perturbations of the logging policy (left) and (ii) perturbations of the optimal policy (right) in our synthetic setup. We compute precision, recall and F1 scores for each threshold values between 0 and 1. One can see that for low threshold values where no policies are filtered, precision, recall and F1 scores remain unchanged. Once the ESS raises above a certain threshold, undesirable policies start being filtered but more false negatives are created when the ESS is too high. Overall, ESS criterion is relevant for both setups. However, we observe that on simple synthetic setups the effective sample size criterion $\nu = n_{\text{eff}}/n$ is seldom necessary for policies close to the logging policy ($\pi \approx \pi_0$). Conversely, for policies which are not close to the logging policy the standard statistical significance testing at $1 - \delta$ level was by itself not enough to guarantee a low false discovery rate (FDR) which justified the use of ν. Adjusting the effective sample size can therefore influence the performance of the protocol (see Appendix 8.2 for further illustrations of importance sampling diagnostics in what-if simulations).

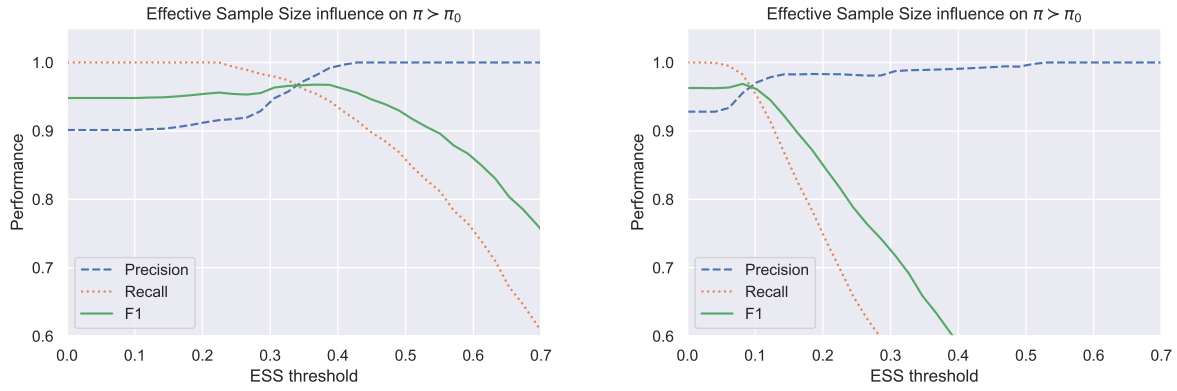

Figure 5: **Precision, recall and F1 score varying with the ESS threshold on synthetic setups (i) and (ii)**. Setup (i) perturbation of the logging policy (left) and setup (ii) perturbation to the optimal policy (right). The ESS threshold can maximize the F1 score.

5.2 Experimental Evaluation of the Continuous Modelling and the Optimization Perspectives

In this section we introduce our empirical settings for evaluation and present our proposed CCP policy parametrization, and the influence of optimization in counterfactural risk minimization problems.

5.2.1 Experimental Setup

We present the synthetic potential prediction setup, a semi-synthetic setup as well as our real-world setup.

Synthetic potential prediction. We introduce simple synthetic environments with the following generative process: an unobserved random group index g in $\mathcal{G}$ is drawn, which influences the drawing of a context x and of an unobserved "potential" p in $\mathbb{R}$, according to a joint conditional distribution $\mathcal{P}_{\mathcal{X},P|\mathcal{G}}$. Intuitively, the potential p may be compared to users a priori responsiveness to a treatment. The observed reward $-y$ is then a function of the context x, action a, and potential p. The causal graph corresponding to this process is given in Figure 6.

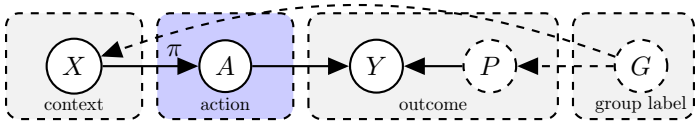

Figure 6: Causal Graph of the synthetic setting. A denotes action, X context, G unobserved group label, Y outcome and P unobserved potentials. Unobserved elements are dotted.

Then, we generate three datasets ("noisymoons, noisycircles, and anisotropic", abbreviated respect. "moons, circles, and GMM" in Table 4 and illustrated in Figure 7, with two-dimensional contexts on 2 or 3 groups and different choices of $\mathcal{P}_{\mathcal{X},P|\mathcal{G}}$.

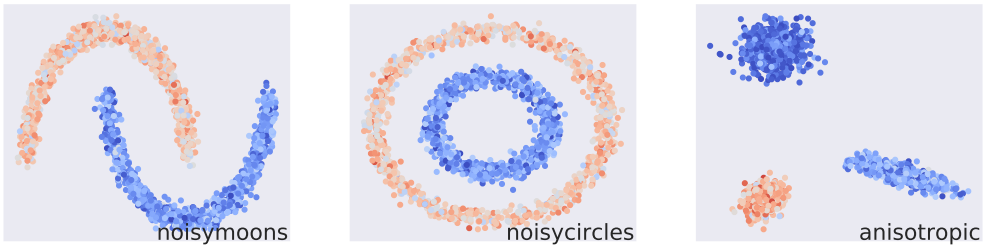

Figure 7: Contexts (points in $\mathbb{R}^2$), and potentials represented by a color map for the synthetic datasets. Learned policies should vary with the context to adapt to the underlying potentials.

The goal is then to find a policy $\pi(a|x)$ that maximizes reward by adapting to an unobserved potential. For our experiments, potentials are normally distributed conditionally on the group index, $p|g \sim \mathcal{N}(\mu_g, \sigma^2)$. As many real-world applications feature a reward function that increases first with the action up to a peak and finally drops, we have chosen a piecewise linear function peaked at $a = p$ (see Appendix 9.1, Figure 16), that mimics reward over the CoCoAdataset presented in Section 4. In bidding applications, a potential may represent an unknown true value for an advertisement, and the reward is then maximized when the bid (action) matches this value. In medicine, increasing drug dosage may increase treatment effectiveness but if dosage exceeds a threshold, secondary effects may appear and eclipse benefits [4].

Semi-synthetic setting with medical data. We follow the setup of Kallus and Zhou [16] using a dataset on dosage of the Warfarin blood thinner drug [1]. The dataset consists of covariates about patients along with a dosage treatment prescription by a medical expert, which is a scalar value and thus makes the setting useful for continuous action modelling. While the dataset is supervised, we simulate a contextual bandit environment by using

a hand-crafted reward function that is maximal for actions a that are within 10% of the expert's therapeutic drug dosage, following Kallus and Zhou [16].

Specifically, the semi-synthetic cost inputs prescriptions from medical experts to obtain $y(a, x) = \max(|a - t^*| - 0.1t^*, 0)$, so as to mimic the expert prediction. The logging policy π_0 samples actions $a \sim \pi_0$ contextually to a patient's body mass index (BMI) score $Z_{BMI} = \frac{x_{\text{BMI}} - \mu_{\text{BMI}}}{\sigma_{\text{BMI}}}$ and can be analytically written with i.i.d noise $e \sim \mathcal{N}(0, 1)$, moments of the therapeutic dose distribution μ_T^*, σ_T^* such that $a = \mu_T^* + \sigma_T^* \sqrt{\theta} Z_{BMI} + \sigma_T^* \sqrt{1-\theta}\epsilon$ ($\theta = 0.5$ in the setup of [16]). The logging probability density function thus is a continuous density of a standard normal distribution over the quantity $\frac{a - \mu_T^* + \sigma_T^* \sqrt{\theta} Z_{BMI}}{\sigma_T^* \sqrt{1-\theta}}$.

Evaluation methodology For synthetic datasets, we generate training, validation, and test sets of size 10 000 each. For the CoCoA dataset, we consider a 50%-25%-25% training-validation-test sets. We then run each method with 5 different random intializations such that the initial policy is close to the logging policy. Hyperparameters are selected on a validation set with logged bandit feedback as explained in Algorithm 1. We use an offline SNIPS estimate of the obtained policies, while discarding solutions deemed unsafe with the importance sampling diagnostic. On the semi-synthetic Warfarin dataset we used a cross-validation procedure to improve model selection due to the low dataset size. For estimating the final test performance and confidence intervals on synthetic and on semi-synthetic datasets, we use an online estimate by leveraging the known reward function and taking a Monte Carlo average with 100 action samples per context: this accounts for the randomness of the policy itself over given fixed samples. For offline estimates we leverage the randomness across samples to build confidence intervals: we use a 100-fold bootstrap and take percentiles of the distribution of rewards. For the CoCoA dataset, we report SNIPS estimates for the test metrics.

5.2.2 Empirical Evaluation

We now evaluate our proposed CCP policy parametrization and the influence of optimization in counterfactural risk minimization problems.

Continuous action space requires more than naive discretization. In Figure 8, we compare our continuous parametrization to discretization strategies that bucketize the action space and consider stochastic discrete-action policies on the resulting buckets, using IPS, DM or DR estimators. On all synthetic datasets, continuous modelings such as CCP perform significantly better than discrete approaches (see also Appendix 10.1), across all choices considered for the number of anchor points/buckets. To achieve a reasonable performance, naive discretization strategies require a much finer grid, and are thus also more computationally costly. The plots also show that our (stochastic) direct method strategy, where we add a minimal amount of noise to the deterministic DM in order to pass the $n_{\text{eff}}/n > \nu$ validation criterion, has similar benefits in terms of robustness to anchor points, thanks to our proposed Nyström parameterization. Nevertheless, it is overall outperformed by CCP, highlighting a benefit of using counterfactual methods compared to a direct fit to observed rewards.

Counterfactual cost predictor (CCP) provides a competitive parameterization for continuous-action policy learning. We compare our CCP modelling approach to other parameterized modelings (constant, linear and non-linear described in Section 2.2) on our synthetic and semi-synthetic setups described in Section 5.2.1 as well as the CoCoAdataset presented in Section 4.1. We also provide a baseline comparison to Chen et al. [7] using their surrogate loss formulation for continous actions and Kallus and Zhou [16], who propose a counterfactual method using kernel density estimation. Their approach is based on an automatic kernel bandwidth selection procedure which did not perform well on our datasets except Warfarin; instead, we select the best bandwidth on a grid through cross-validation and selecting it through our offline protocol. We experimented using the generic doubly robust method from Demirer et al. [8] but could not reach satisfactory results using the parameters and feature maps that were used in their empirical section and with the specific closed form estimators for their applications. Nevertheless, by adapting their method with more elaborated models and feature maps, we managed to obtain performances beating the logging policy; these modifications would make promising directions for future research venues.

In Table 4, we show a comparison of test rewards for several estimators. We report here the performances of scIPS and SNIPS strategies. For the Warfarin dataset, following Kallus and Zhou [16], we only consider the linear context parametrization baseline, since the dataset has categorical features and higher-dimensional contexts. All results are obtained by using the offline model selection procedure for optimizing hyper-parameters (see Section 4.2).

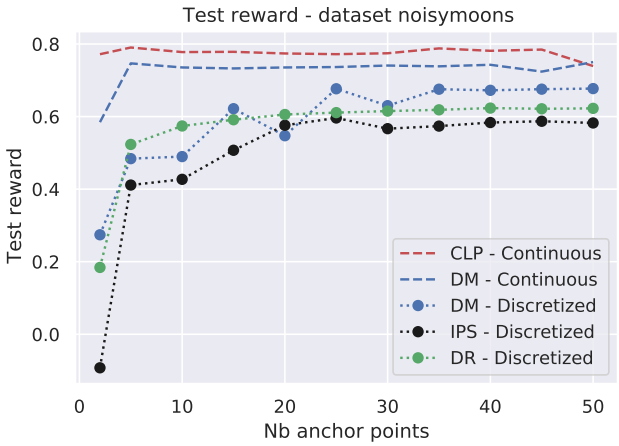

Figure 8: Continuous vs discretization strategies. Test rewards on NoisyMoons dataset with varying numbers of anchor points for our continuous parametrization for CCP and DM, versus naive discretization. Note that few anchor points are sufficient to achieve good results on this dataset; this is not the case for more complicated ones (*e.g.*, Warfarin requires at least 15 anchor points).

Overall, we find our CCP parameterization to improve over all baselines, which highlights the effectiveness of the reward predictor at exploiting the continuous action structure. Note that unlike synthetic setups, it is harder to obtain policies that beat the logging policy with large statistical significance on the CoCoAdataset where the logging policy already makes a satisfactory baseline for real-world deployment. Only CCP and Kallus and Zhou [16] are passing the significance test.

Table 4: Test rewards (higher the better) for several estimators (see main text for details). * refers to the original result from Kallus and Zhou [16], denoted by KZ.

		Noisycircles	NoisyMoons	Anisotropic	Warfarin	CoCoA
Logging policy π_0		0.5301	0.5301	0.4533	-13.377	11.34
scIPS	Constant	0.6115 ± 0.0000	0.6116 ± 0.0000	0.6026 ± 0.0000	-8.964 ± 0.001	11.36 ± 0.13
	Linear	0.6113 ± 0.0001	0.7326 ± 0.0001	0.7638 ± 0.0005	-12.857 ± 0.002	11.35 ± 0.02
	Poly	0.6959 ± 0.0001	0.7281 ± 0.0001	0.7448 ± 0.0008	-	10.36 ± 0.11
	CCP	$\mathbf{0.7674} \pm 0.0008$	$\mathbf{0.7805} \pm 0.0004$	$\mathbf{0.7703} \pm 0.0002$	$\mathbf{-8.720} \pm 0.001$	$\mathbf{11.44}^* \pm 0.10$
SNIPS	Constant	0.6115 ± 0.0001	0.6115 ± 0.0001	0.5930 ± 0.0001	-9.511 ± 0.001	11.32 ± 0.13
	Linear	0.6115 ± 0.0001	0.7360 ± 0.0001	0.7103 ± 0.0003	-10.583 ± 0.005	10.34 ± 0.12
	Poly	0.6969 ± 0.0001	0.7370 ± 0.0001	0.5801 ± 0.0002	-	11.13 ± 0.08
	CCP	$\mathbf{0.6972} \pm 0.0001$	$\mathbf{0.74091} \pm 0.0004$	$\mathbf{0.7899} \pm 0.0002$	$\mathbf{-9.161} \pm 0.001$	$\mathbf{11.48}^* \pm 0.14$
Chen et al. [7]		0.608 ± 0.0002	0.645 ± 0.0003	0.754 ± 0.0002	-9.407 ± 0.004	-
Kallus and Zhou [16]		0.612 ± 0.0001	0.734 ± 0.0001	0.785 ± 0.0002	-10.19^*	11.38 ± 0.07

Soft-clipping improves performance of the counterfactual policy learning. Figure 9 shows the improvements in test reward of our optimization-driven strategies for the soft-clipping estimator for the synthetic datasets (see also Appendix 10.2). The points correspond to different choices of the clipping parameter M, models and initialization, with the rest of the hyper-parameters optimized on the validation set using the offline evaluation protocol. This plot also shows that soft clipping provides benefits over hard clipping, perhaps thanks to a more favorable optimization landscape. Overall, these figures confirm that the optimization perspective is important when considering CRM problems.

Proximal point algorithm (PPA) influences optimization of non-convex CRM objective functions and policy learning performance. We illustrate in Figure 10 the improvements in test reward and in training objective of our optimization-driven strategies with the use of the proximal point algorithm (see also

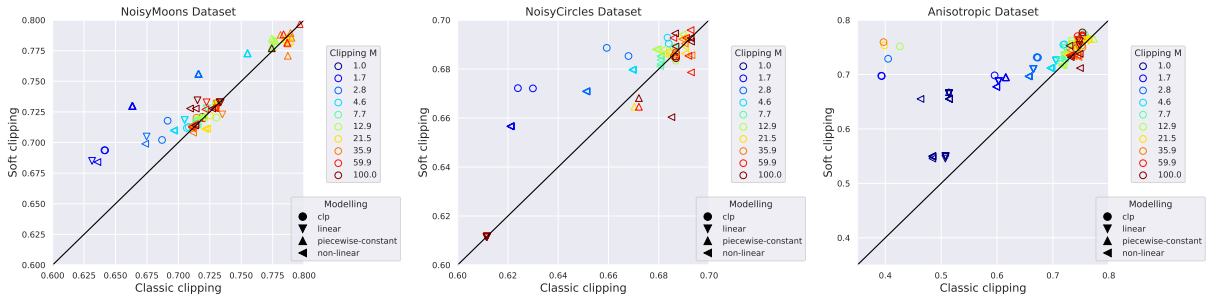

Figure 9: Influence of soft-clipping. Relative improvements in the test performance for soft- vs hard-clipping on synthetic datasets. The points correspond to different choices of the clipping parameter, models and initialization.

Appendix 10.2). Here, each point compares the test metric for fixed models as well as initialization seeds, while optimizing the remaining hyperparameters on the validation set with the offline evaluation protocol. Figure 10 (left) illustrates the benefits of the proximal point method when optimizing the (non-convex) CRM objective in a wide range of hyperparameter configurations, while Figure 10 (center) shows that in many cases this improves the test reward as well. In our experiments, we have chosen L-BFGS because it was performing best among the solvers we tried (nonlinear conjugate gradient (CG) and Newton) and used 10 PPA iterations. For further information, Figure 10 (right) presents a comparison between CG and L-BFGS for different parameters κ and number of iterations. As for computational time, for 10 PPA iterations, the computational overhead was about $3\times$ in comparison to L-BFGS without PPA. This overhead seems to be the price to pay to improve the test reward and obtain better local optima. Overall, these figures confirm that the proximal point algorithm improves performance in CRM optimization problems.

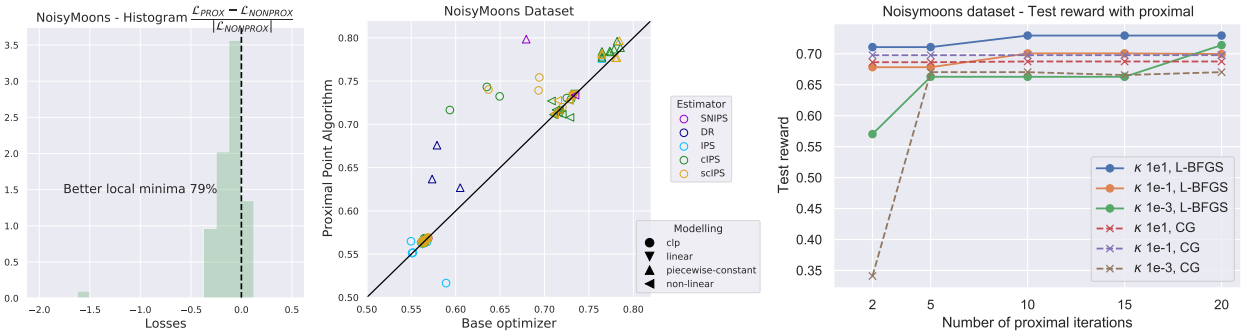

Figure 10: Influence of proximal point optimization. Relative improvements in the training objective w and w/o using the proximal point method (left), relative improvements in the test performance w and w/o using the proximal point method (center) comparison of test performance with different number of proximal steps and (right).

6 Conclusion and Discussion

In this paper, we address the problem of counterfactual learning of stochastic policies on real data with continuous actions. This raises several challenges about different steps of the CRM pipeline such as (i) modelization, (ii) optimization, and (iii) evaluation. First, we propose a new parametrization based on a joint kernel embedding of contexts and actions, showing competitive performance. Second, we underline the importance of optimization in CRM formulations with soft-clipping and proximal point methods. Third, we propose an offline evaluation protocol and a new large-scale dataset, which, to the best of our knowledge, is the first with real-world logged propensities and continuous actions. For future research directions, we would like to discuss the doubly robust estimator (which achieves

the best results in the discrete action case) with the CCP parametrization of stochastic policies with continuous actions, as well as further optimization perspectives and the offline model selection.

Doubly-robust estimators for continuous action policies While Demirer et al. [8] provide a doubly robust (DR) estimator on continuous action using a semi-parametric model of the policy value function, we did not propose a doubly-robust estimator along with our CCP modelling. Indeed, their policy learning is performed in two stages (i) estimate a doubly robust parameter $\theta^{DR}(x, a, r)$ in the semi-parametric model of the value function $\mathbb{E}[y|a, x] = V(a, x) = \langle \theta_*(x), \phi(a, x) \rangle$ and (ii) learn a policy in the empirical Monte Carlo estimate of the policy value by solving

$$\min_{\pi \in \Pi} \left\{ \hat{V}^{DR}(\pi) := \frac{1}{n} \sum_{i=1}^{n} \langle \hat{\theta}^{DR}(x_i, a_i, r_i), \phi(\pi(x_i), x_i) \rangle \right\}.$$

The doubly robust estimation is performed with respect to the first parameter learned in (i) for the value function, while we follow the CRM setting [32] and directly derive estimators of the policy value (risk) itself, which would correspond to the phase (ii). To derive an estimate a DR estimator of such policy values, we tried extending the standard DR approach for discrete actions from Dudik et al. [9] to continuous actions by using our anchors points, but these worked poorly in practice. Actually, a proper DR method for estimating the expectation of a policy risk likely requires new techniques for dealing with integration over the training policy in the direct method term, which is non-trivial and goes beyond the scope of this work. We hope to be able to do this in the future.

On further optimization perspectives and offline model selection As mentioned in Section 3, our use of the proximal point algorithm differs from approaches that enhance policies to stay close to the logging policies which modify the objective function as in [29] in reinforcement learning. Another venue for future work would be to investigate distributionnally robust methods that do such modifications of the objective function or add constraints on the distribution being optimized. The policy thereof obtained would thus be closer to the logging policy in the CRM context. Moreover, as we showed in Section 4.2 with importance sampling estimates and diagnostics, the offline decision becomes less statistically significant as the evaluated policy is far from the logging policy. Investigating how the distributionally robust optimization would yield better CRM solutions with regards to the offline evaluation protocol would make an interesting future direction of research.

Acknowledgments

The authors thank Christophe Renaudin (Criteo AI Lab) for his help in the data collection and engineering which was necessary for this project. JM was supported by the ERC grant number 714381 (SOLARIS project) and by ANR 3IA MIAI@Grenoble Alpes, (ANR-19-P3IA-0003). AB acknowledges support from the ERC grant number 724063 (SEQUOIA project).

References

[1] Estimation of the Warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360 (8):753–764, 2009.

[2] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning (ICML)*, 2014.

[3] Z. Ahmed, N. Le Roux, M. Norouzi, and D. Schuurmans. Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning (ICML)*, 2019.

[4] C. D. Barnes and L. G. Eltherington. *Drug dosage in laboratory animals: a handbook*. Univ of California Press, 1966.

[5] D. Bertsimas and C. McCord. Optimization over continuous and multi-dimensional decisions with observational data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

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

[7] G. Chen, D. Zeng, and M. R. Kosorok. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, 111(516):1509–1521, 2016.

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

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

[10] M. Fukushima and H. Mine. A generalized proximal point algorithm for certain non-convex minimization problems. *International Journal of Systems Science*, 12(8):989–1000, 1981.

[11] K. Hirano and G. W. Imbens. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84, 2004.

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

[13] N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2016.

[14] T. Joachims, A. Swaminathan, and M. de Rijke. Deep learning with logged bandit feedback. In *International Conference on Learning Representations (ICLR)*, 2018.

[15] S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2002.

[16] N. Kallus and A. Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[17] A. Krause and C. S. Ong. Contextual gaussian process bandit optimization. In *Advances in neural information processing systems (NIPS)*, 2011.

[18] J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[19] D. Lefortier, A. Swaminathan, X. Gu, T. Joachims, and M. de Rijke. Large-scale validation of counterfactual learning methods: A test-bed. *arXiv preprint arXiv:1612.00367*, 2016.

[20] L. Li, W. Chu, J. Langford, T. Moon, and X. Wang. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, 2012.

[21] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[22] A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. In *Conference on Learning Theory (COLT)*, 2009.

[23] T. Nedelec, N. L. Roux, and V. Perchet. A comparative study of counterfactual estimators. *arXiv preprint arXiv:1704.00773*, 2017.

[24] A. Owen. *Monte Carlo theory, methods and examples.* 2013.

[25] C. Paquette, H. Lin, D. Drusvyatskiy, J. Mairal, and Z. Harchaoui. Catalyst for gradient-based nonconvex optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[26] J. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[27] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[28] B. Schölkopf and A. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[29] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[30] Y. Su, L. Wang, M. Santacatterina, and T. Joachims. Cab: Continuous adaptive blending for policy evaluation and learning. In *International Conference on Machine Learning (ICML)*, 2019.

[31] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems (NIPS)*, 2000.

[32] A. Swaminathan and T. Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning (ICML)*, 2015.

[33] A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems (NIPS)*. 2015.

[34] Y.-X. Wang, A. Agarwal, and M. Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning (ICML)*, 2017.

[35] C. K. Williams and M. Seeger. Using the nyström method to speed up kernel machines. *Adv. Neural Information Processing Systems (NIPS)*, 2001.

[36] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Appendix

This appendix is organized as follows: in Appendix 7, we present discussions and toy experiments to motivate the need for clipping strategies on real datasets. Next, we provide motivations in Appendix 8 for the offline evaluation protocol with experiments justifying the need for appropriate diagnostics and statistical testing for importance sampling. Then, Appendix 9 is devoted to experimental details that were omitted from the main paper for space limitation reasons, and which are important for reproducing our results (see also the code provided with the submission). In Appendix 10, we present additional experimental results to those in the main paper.

7 Motivation for Clipped Estimators

In this section we provide a motivation example for clipping strategies in counterfactual systems in a toy example.

In Figure 11 we provide an example of unbounded variance and loss overfitting problem.

We recall the data generation: a hidden group label g in $\mathcal{G}$ is drawn, and influences the associated context distribution x and of an unobserved potential p in $\mathbb{R}$, according to a joint conditional distribution $P_{X,P|G}$ The observed reward r is then a function of the context x, action a, and potential p. Here, we design one outlier (big red dark dot on Figure 2 left). This point has a noisy reward r, higher than neighbors, and a potential p high as its neighbors have a low potential. We artificially added a noise in the reward function f that can be written as:

$$r = f(a, x, p) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 1)$$

As explained in Section 5.2.1, the reward function is a linear function, with its maximum localized at the point $x = p(x)$, i.e. at the potential sampled. The observability of the potential p is only through this reward function f. Hereafter, we compare the optimal policy computed, using different type of estimators.

The task is to predict the high potentials (red circles) and low potentials (blue circles) in the ground truth data (left). Unfortunately, a rare event sample with high potential is put in the low potential cluster (big dark red dot). The action taken by the logging policy is low while the reward is high: this sample is an outlier because it has a high reward while being a high potential that has been predicted with a low action. The resulting unclipped estimator is biased and overfits this high reward/low propensity sample. The rewards of the points around this outlier are low as the diameter of the points in the middle figure show. Inversely, clipped estimator with soft-clipping succeeds to learn the potential distributions, does not overfit the outlier, and has larger rewards than the clipping policies as the diameter of the points show.

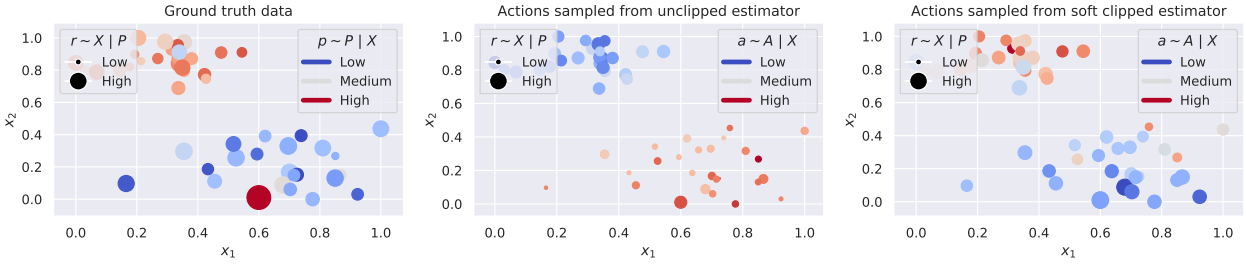

Figure 11: Unbounded variance and loss overfitting. Unlikely ($\pi_{0,i} \approx 0$) sample $(x_1, x_2) = (0.6, 0.)$ with high reward r (left) results in larger variance and loss overfitting for the unclipped estimator (middle) unlike clipped estimator (right).

8 Motivation for Offline Evaluation Protocol

In this part we demonstrate the offline/online correlation of the estimator we use for real-world systems and for validation of our methods even in synthetic and semi-synthetic setups. We provide further explanations of the necessity of importance sampling diagnostics and we perform experiments to empirically assess the rate of false discoveries of our protocol.

8.1 Correlation of Self-Normalized Importance Sampling with Online Rewards

We show in Figures 12,14,13 comparisons of IPS and SNIPS against an on-policy estimate of the reward for policies obtained from our experiments for linear and non-linear contextual modellings on the synthetic datasets, where policies can be directly evaluated online. Each point represents an experiment for a model and a hyperparameter combination. We measure the R^2 score to assess the quality of the estimation, and find that the SNIPS estimator is indeed more robust and gives a better fit to the on-policy estimate. Note also that overall the IPS estimates illustrate severe variance compared to their SNIPS estimate. While SNIPS indeed reduces the variance of the estimate, the bias it introduces does not deteriorate too much its (positive) correlation with the online evaluation.

These figures further justify the choice of the self-normalized estimator SNIPS [33] for offline evaluation and validation to estimate the reward on held-out logged bandit data. The SNIPS estimator is indeed more robust to the reward distribution thanks to its equivariance property to additive shifts and does not require hyperparameter tuning.

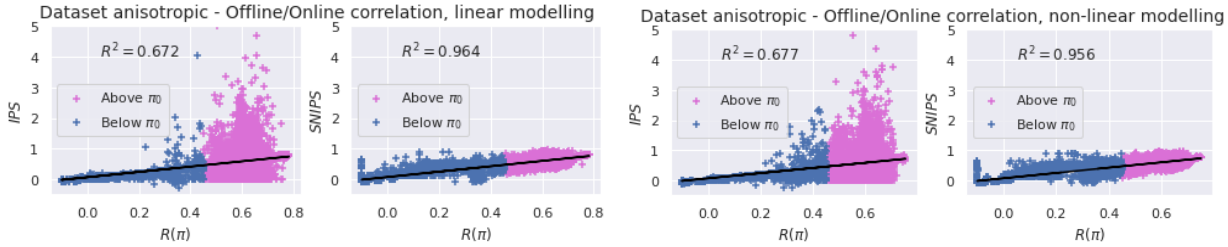

Figure 12: Correlation between offline and online estimates on Anisotropic synthetic data. Linear (left) and non-linear (right) contextual modellings. Ideal fit would be $y = x$.

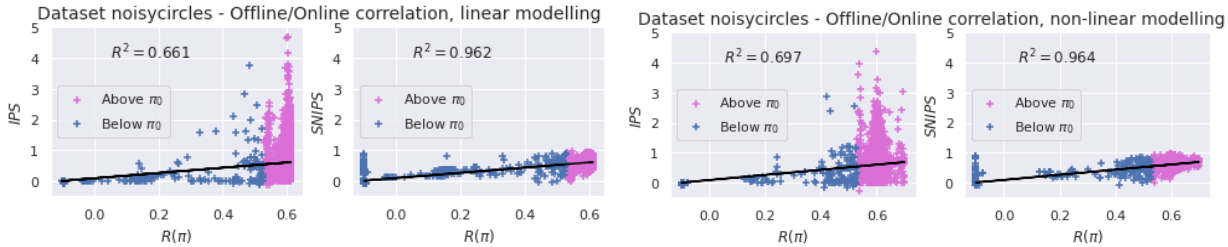

Figure 13: Correlation between offline and online estimates on NoisyCircles synthetic data. Linear (left) and non-linear (right) contextual modellings. Ideal fit would be $y = x$.

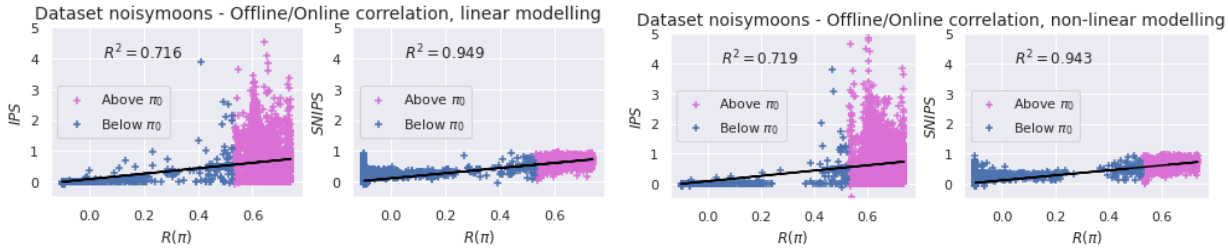

Figure 14: Correlation between offline and online estimates on NoisyMoons synthetic data. Linear (left) and non-linear (right) contextual modellings. Ideal fit would be $y = x$.

8.2 Importance Sampling Diagnostics in What-If simulations

Importance sampling estimators rely on weighted observations to address the distribution mismatch for offline evaluation, which may cause large variance of the estimator. Notably, when the evaluated policy differs too much from the logging policy, many importance weights are large and the estimator is inaccurate. We provide here a motivating example to illustrate the effect of importance sampling diagnostics in a simple scenario.

When evaluating with SNIPS, we consider an "effective sample size" quantity given in terms of the importance weights $w_i = \pi(a_i|x_i)/\pi_0(a_i|x_i)$ by $n_e = (\sum_{i=1}^n w_i)^2 / \sum_{i=1}^n w_i^2$. When this quantity is much smaller than the sample size n, this indicates that only few of the examples contribute to the estimate, so that the obtained value is likely a poor estimate. Apart from that, we note also that IPS weights have an expectation of 1 when summed over the logging policy distribution (that is $\mathbb{E}_{(x,a)\sim\pi_0}[\pi(a|x)/\pi_0(a|x)] = 1.$). Therefore, another sanity check, which is valid for any estimator, is to look for the empirical mean $1/n \sum_{i=1}^n \pi(a_i|x_i)/\pi_{0,i}$ and compare its deviation to 1. In the example below, we illustrate three diagnostics: (i) the one based on effective sample size described in Section 4; (ii) confidence intervals, and (iii) empirical mean of IPS weights. The three of them coincide and allow us to remove test estimates when the diagnostics fail.

Example 8.1. *What-if simulation: For x in $\mathbb{R}^d$, let $\max(x) = \max_{1\leq j\leq d} x_j$; we wish to estimate $\mathbb{E}(\max(X))$ for X i.i.d $\sim \pi_\mu = \mathcal{N}(\mu, \sigma)$ where samples are drawn from a logging policy $\pi_0 = \log\mathcal{N}(\lambda_0, \sigma_0)$ $(d = 3, (\lambda_0, \sigma_0) = (1, 1/2))$ and analyze parameters μ around the mode of the logging policy μ_{π_0} with fixed variance $\sigma = 1/2$. In this parametrized policy example, we see in Fig. 15 that $n_e/n \ll 1$, confidence interval range increases and $\sum_{i=1}^n \frac{\pi(a_i|x_i)}{\pi_{0,i}} \neq 1$ when the parameter μ of the policy being evaluated is far away from the logging policy mode μ_{λ_0}.*

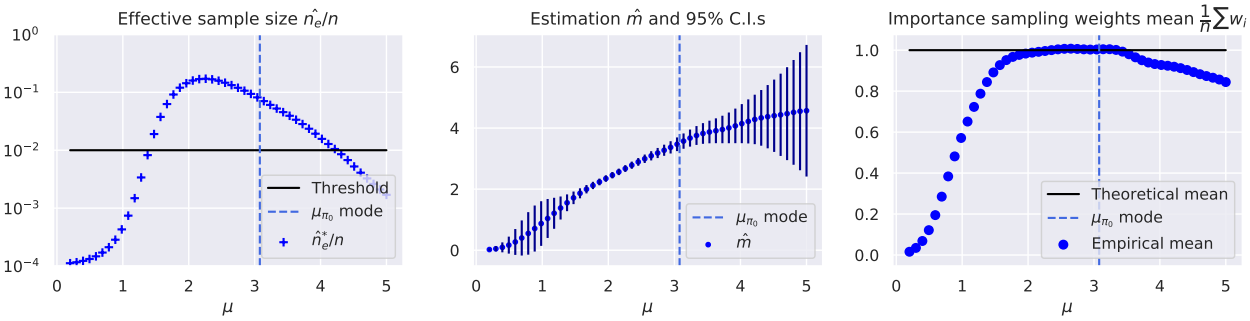

Figure 15: Importance sampling diagnostics. Ideal importance sampling: i) effective sample n_e/n close to 1, ii) low confidence intervals (C.I.s) for $\hat{m}$, iii) empirical mean $\frac{1}{n}\sum_i w_i$ close to 1. Note that when μ differs too much from μ_{π_0}, importance sampling fails.

Note that in this example, the parameterized distribution that is learned (multivariate Gaussian) is not the same as the parameterized distribution of the logging policy (multivariate Lognormal) which skewness may explain the asymmetry of the plots. This points out another practical problem: even though different parametrization of policies is theoretically possible, the probability density masses overlap is in practice what is most important to ensure successful importance sampling. This observation is of utmost interest for real-life applications where the initialization of a policy to be learned needs to be "close" to the logging policy; otherwise importance sampling may fail from the very first iteration of an optimization in learning problems.

9 Details on the Experiment Setup and Reproducibility

In this section we give additional details on synthetic and semi-synthetic datasets, we provide details on the evaluation methodology and information for experiment reproducibility.

9.1 Synthetic and Semi-Synthetic setups

Synthetic setups As many real-world applications feature a reward function that increases first with the action, then plateaus and finally drops, we have chosen a piecewise linear function as shown in Fig. 16 that mimics reward buckets over the CoCoA dataset presented in Section 4.

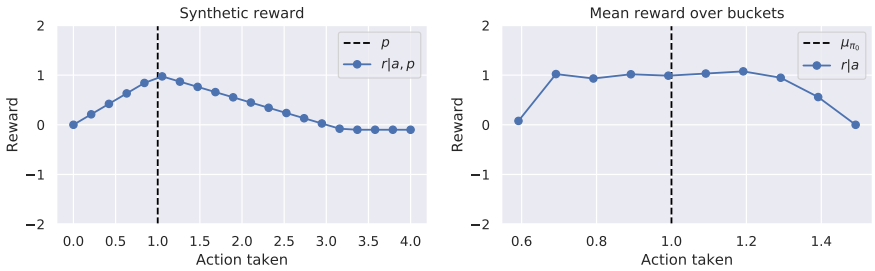

Figure 16: Synthetic reward engineering. The synthetic reward (left) is inspired from real-dataset reward buckets (right).

Semi-synthetic medical setup

9.2 Reproducibility

We provide code for reproducibility and all experiments were run on a cpu cluster, each node consisting on 24 CPU cores (2 x Intel(R) Xeon(R) Gold 6146 CPU@ 3.20GHz), with 500Go of RAM.

Policy parametrization. In our experiments, we consider two forms of parametrizations: (i) a lognormal distribution with $\theta = (\theta_\mu, \sigma), \pi_{(\mu,\sigma)} = \log \mathcal{N}(m, s)$ with $s = \sqrt{\log\left(\sigma^2/\mu^2 + 1\right)}; m = \log(\mu) - s^2/2$, so that $\mathbb{E}_{a\sim\pi_{(\mu,\sigma)}}[a] = \mu$ and $\mathrm{Var}_{a\sim\pi_{(\mu,\sigma)}}[a] = \sigma^2$; (ii) a normal distribution $\pi_{(\mu,\sigma)} = \mathcal{N}(\mu, \sigma)$. In both cases, the mean μ may depend on the context (see Section 3), while the standard deviation σ is a learned constant. We add a positivity constraint for σ and add an entropy regularization term to the objective in order to encourage exploratory policies and avoid degenerate solutions.

Models. For parametrized distributions, we experimented both with normal and lognormal distributions on all datasets, and different baseline parameterizations including constant, linear and quadratic feature maps. We also performed some of our experiments on low-dimensional datasets with a stratified piece-wise contextual parameterization, which partitions the space by bucketizing each feature by taking K (for e.g $K = 4$) quantiles, and taking the cross product of these partitions for each feature. However this baseline is not scalable for higher dimensional datasets such as the Warfarin dataset.

Hyperparameters. In Table 5 we show the hyperparameters considered to run the experiments to reproduce all the results. Note that the grid of hyperparameters is larger for synthetic data. For our experiments involving anchor points, we validated the number of anchor points and kernel bandwidths similarly to other hyperparameters.

Table 5: Table of hyperparameters for the Synthetic and CoCoA datasets

	Synthetic	Warfarin	CoCoA
Variance reg. λ	$\{0., 0.001, 0.01, 0.1, 1, 10, 100\}$	$\{0.00010.0010.010.1\}$	$\{0., 0.001, 0.1\}$
Clipping M	$\{1, 1.7, 2.8, 4.6, 7.7, 12.9, 21.5, 35.9, 59.9, 100.0\}$	$\{1, 2.1, 4.5, 9.5, 20\}$	$\{1, 2.1, 4.5, 9.5, 10, 20, 100\}$
Prox. κ	$\{0.001, 0.01, 0.1, 1\}$	$\{0.001, 0.01, 0.1\}$	$\{0.001, 0.01, 0.1\}$
Reg. param. C	$\{0.00001, 0.0001, 0.001, 0.01, 0.1\}$	$\{0.00001, 0.0001, 0.001, 0.01, 0.1\}$	$\{0.00001, 0.0001, 0.001, 0.01, 0.1\}$
Number of anchor points	$\{2, 3, 5, 7, 10\}$	$\{5, 7, 10, 12, 15, 20\}$	$\{2, 3, 5\}$
Softmax γ	$\{1, 10, 100\}$	$\{1, 5, 10\}$	$\{0.1, 0.5, 1, 5\}$

10 Additional Results and Additional Evaluation Metrics

In this section we provided additional results on both contextual modeling and optimization driven approaches of CRM.

10.1 Continuous vs Discrete strategies in continuous-action space

We provide in Figure 17 additional plots for the continuous vs discrete strategies for the synthetic setups described in Section 5.2.1.

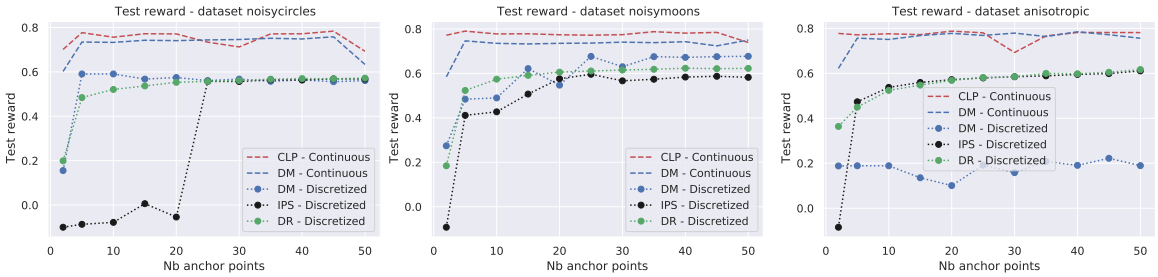

Figure 17: Continuous vs discrete. Test rewards for CLP and (stochastic) direct method (DM) with Nyström parameterization, versus a discrete approach, with varying numbers of anchor points. We add a minimal amount of noise to the deterministic DM in order to pass the n_{eff} validation criterion.

10.2 Optimization Driven Approaches of CRM

In this part we provide additional results on optimization driven approaches of CRM for the Noisycircles, Anisotropic, Warfarin and CoCoAdatasets.

Both Noisycircles and Anisotropic datasets in Figure 18 show the improvements in test reward and in training objective of our optimization-driven strategies, namely the soft-clipping estimator and the use of the proximal point algorithm. Overall we see that for most configurations, the proximal point method better optimizes the objective function and provides better test performances, while the soft-clipping estimator performs better than its hard-clipping variant, which may be attributed to the better optimization properties. For semi-synthetic Warfarin and real-world CoCoA datasets in Figure 18 we also show the improvements in test reward and in training objective of our optimization-driven strategies. More particularly we demonstrate the effectiveness of proximal point methods on the Warfarin dataset where most proximal configurations perform better than the base algorithm. Moreover, soft-clipping strategies perform better than its hard-clipping variant on real-world dataset with outliers and noises, which demonstrate the effectiveness of this smooth estimator for real-world setups.

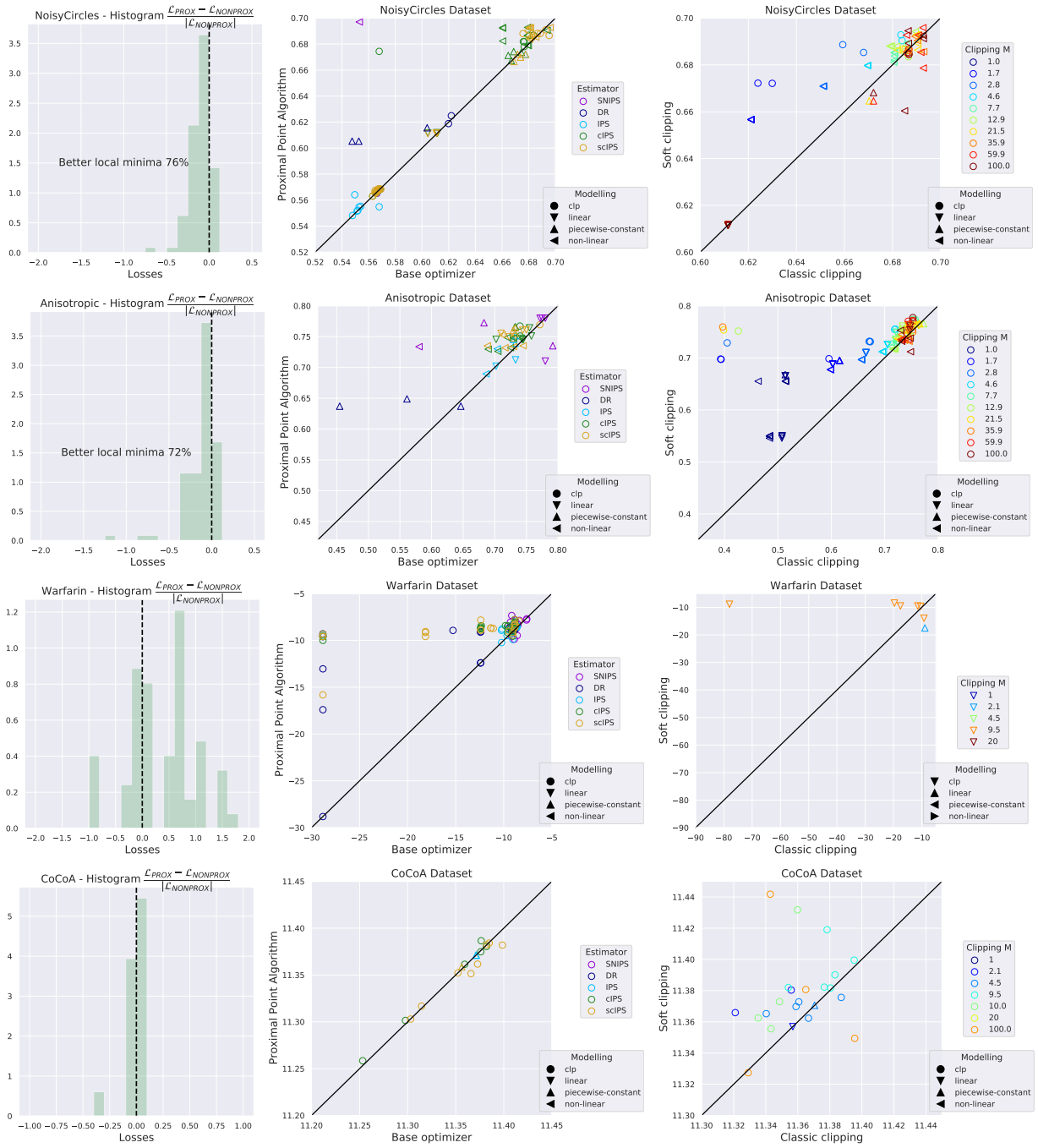

Figure 18: Optimization-driven approaches (NoisyCircles, Anisotropic, Warfarin and CoCoA datasets). Relative improvements in the training objective from using the proximal point method (left), comparison of test rewards for proximal point vs the simpler gradient-based method (center), and for soft- vs hard-clipping (right). See also Figures 9, 10 for the NoisyMoons dataset.