Bayesian Counterfactual Risk Minimization

Ben London [1] **Ted Sandler** [1]

Abstract

We present a Bayesian view of counterfactual risk minimization (CRM), or offline policy optimization from logged bandit feedback. Using PAC-Bayesian analysis, we derive a new generalization bound for the truncated IPS estimator. We apply the bound to a class of Bayesian policies, which motivates a novel, potentially data-dependent, regularization technique for CRM.

1. Introduction

In industrial applications of machine learning, model development is typically an iterative process, involving multiple trials of offline training and online experimentation. For example, a content streaming service might explore various recommendation strategies in a series of A/B tests. The data that is generated by this process—e.g., impression and interaction logs—can be used to augment training data and further refine a model. However, learning from logged interactions poses two fundamental challenges: (1) the feedback obtained from interaction is always incomplete, in that one only observes responses (usually referred to as *rewards*) for actions that were taken; (2) the distribution of observations is inherently biased by the *policy* that determines which action to take in each context.

This learning problem has been studied under various names by various authors [2, 4, 14, 15]. We adopt the moniker *counterfactual risk minimization* (CRM), introduced by Swaminathan & Joachims [15]. The goal of CRM is to learn a policy from data that was logged by a previous policy, such that the learned policy maximizes expected reward (alternatively, minimizes counterfactual risk) over draws of future contexts. Using an analysis based on Bennett's inequality, Swaminathan & Joachims derived an upper bound on the counterfactual risk of a stochastic policy, which motivates learning with variance-based regularization. In a similar vein, Strehl et al. [14] derived a lower bound on the expected reward of a deterministic policy.

[1]Amazon Music, Seattle, WA, USA. Correspondence to: Ben London <blondon@amazon.com>.

Proceedings of the FAIM'18 Workshop on CausalML, Stockholm, Sweden, 2018.

In this work, we study CRM from a Bayesian perspective, in which one's uncertainty over actions becomes uncertainty over models. That is, instead of learning a single stochastic policy from which actions are sampled, one learns a distribution over hypotheses, which induces a distribution over policies. This bridges the gap between CRM, which has until now been approached from the frequentist perspective, and Bayesian methods, which are often used to balance exploration and exploitation in contextual bandit problems [3].

Using a PAC-Bayesian analysis, we prove an upper bound on the counterfactual risk of a Bayesian policy. We then apply this bound to a class of Bayesian policies based on the mixed logit model. This analysis suggests a novel regularization strategy for CRM based on the L_2 distance from the logging policy's parameters. This regularizer is effectively similar to variance regularization, but simpler to implement. We also consider the scenario in which the logging policy is unknown; in this case, we propose to learn the logging policy, and provide a corresponding counterfactual risk bound based on data-dependent regularization.

Note *In the interest of brevity, all results are presented without proof. A full version of this work, with proofs, is available on arXiv [8].*

2. Preliminaries

Let $\mathcal{X}$ denote a space of *contexts*, and $\mathcal{A}$ denote a finite set of k discrete *actions*. We are interested in finding a *stochastic policy*, $\pi : \mathcal{X} \to \Delta^k$, which maps $\mathcal{X}$ to the k-dimensional probability simplex, Δ^k; in other words, π defines a conditional probability distribution over actions given contexts. For a given context, $x \in \mathcal{X}$, we denote the conditional distribution on $\mathcal{A}$ by $\pi(x)$, and the probability mass of a particular action, $a \in \mathcal{A}$, by $\pi(a \,|\, x)$.

Each action is associated with a stochastic, contextual *reward*, given by an unknown function, $\rho : \mathcal{X} \times \mathcal{A} \to [0, 1]$, which we assume is bounded. When an action is played in response to a context, we only observe the reward for said action. This type of incomplete feedback is commonly referred to as *bandit feedback*. We assume a stationary distribution, $\mathbb{D}$, over contexts and reward functions. Our goal will be to find a policy that maximizes the expected reward

over draws of $(x, \rho) \sim \mathbb{D}$ and $a \sim \pi(x)$; or, put differently, one that minimizes the *counterfactual risk*,

$$R(\pi) \triangleq 1 - \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}} \mathop{\mathbb{E}}_{a\sim\pi(x)} [\rho(x, a)] .$$

We assume that we have access to a dataset of logged observations (i.e., examples), $S \triangleq (x_i, a_i, r_i, p_i)_{i=1}^n$, where (x_i, ρ) were sampled from $\mathbb{D}$; action a_i was sampled with probability $p_i \triangleq \pi_0(a_i \,|\, x_i)$ from a stationary *logging policy*, π_0; and reward $r_i \triangleq \rho(x_i, a_i)$ was observed. The distribution of S, which we denote by $(\mathbb{D} \times \pi_0)^n$, is biased by the logging policy, in that we only observe rewards for actions that were sampled from its distribution. Nonetheless, we can obtain an unbiased estimate of $R(\pi)$ by scaling each reward by its *inverse propensity score* (IPS) [11], p_i^{-1}, which yields the *IPS empirical risk*,

$$\hat{R}(\pi, S) \triangleq 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(a_i \,|\, x_i)\, r_i}{p_i}.$$

Unfortunately, without additional assumptions on the supports of π and π_0, this estimator has unbounded variance. This issue can be mitigated by *truncating* (or *clipping*) p_i to the interval $[\tau, 1]$ (as proposed in [14]), which yields

$$\hat{R}_\tau(\pi, S) \triangleq 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(a_i \,|\, x_i)\, r_i}{\max\{p_i, \tau\}}. \tag{1}$$

This estimator has finite variance, but at the cost of adding bias. However, since $\max\{p_i, \tau\} \geq p_i$, we have that $\hat{R}_\tau(\pi, S) \geq \hat{R}(\pi, S)$, which implies

$$\mathop{\mathbb{E}}_{S\sim(\mathbb{D}\times\pi_0)^n} \left[\hat{R}_\tau(\pi, S)\right] \geq \mathop{\mathbb{E}}_{S\sim(\mathbb{D}\times\pi_0)^n} \left[\hat{R}(\pi, S)\right] = R(\pi).$$

Thus, if $\hat{R}_\tau(\pi, S)$ concentrates around its mean, then by minimizing $\hat{R}_\tau(\,\cdot\,, S)$, we minimize a probabilistic upper bound on the counterfactual risk.

Remark 1. There are other estimators we can consider. For instance, we could instead truncate the ratio of the reward and the IPS, $\min\{r_i/p_i, \tau^{-1}\}$. However, this encourages a policy to favor low-reward actions when r_i and p_i are equally small relative to τ. Alternatively, we could truncate the ratio of the policy and the logging policy, $\min\{\pi(a_i \,|\, x_i)/p_i, \tau^{-1}\}$ (as proposed in [5, 15]). However, this form of truncation is incompatible with our subsequent analysis because the policy is inside the $\min$ operator. Avoiding truncation altogether, we could use the *self-normalizing* estimator [16], but this is also incompatible, since the estimator does not decompose as a sum of i.i.d. random variables. Finally, we note that our theory *does* apply, with small modifications, to the *doubly-robust* estimator [4].

2.1. Counterfactual Risk Minimization

Our work is heavily influenced by Swaminathan & Joachims [15], who coined the term *counterfactual risk minimization* (CRM) to refer to the problem of learning a policy from logged bandit feedback by minimizing an upper bound on the counterfactual risk. Their bound is a function of the truncated IPS estimator (with slightly different truncation), the sample variance of said estimator, $\hat{\mathrm{Var}}[\hat{R}_\tau(\pi, S)]$, and a measure of the complexity, $\mathcal{C}$, of the class of policies being considered, $\Pi \subseteq \{\pi : \mathcal{X} \to \Delta^k\}$. Ignoring constants, their bound is of the form

$$R(\pi) \leq \hat{R}_\tau(\pi, S) + \mathrm{O}\left(\sqrt{\frac{\hat{\mathrm{Var}}[\hat{R}_\tau(\pi, S)]\, \mathcal{C}(\Pi)}{n}} + \frac{\mathcal{C}(\Pi)}{n}\right). \tag{2}$$

This motivates a variance-regularized learning objective,

$$\mathop{\arg\min}_{\pi\in\Pi} \hat{R}_\tau(\pi, S) + \lambda \sqrt{\frac{\hat{\mathrm{Var}}[\hat{R}_\tau(\pi, S)]}{n}}, \quad \text{for } \lambda > 0,$$

which Swaminathan & Joachims minimize using a majorization-minimization algorithm.

3. PAC-Bayesian Analysis

In this work, we view CRM from a Bayesian perspective. We consider stochastic policies whose action distributions are induced by distributions over *hypotheses*. Instead of sampling directly from a distribution on the action space, we sample from a distribution on a *hypothesis space*, $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{A}\}$, in which each element is a deterministic mapping from contexts to actions.[1] As such, for a distribution, $\mathbb{Q}$, on $\mathcal{H}$, the probability of an action, $a \in \mathcal{A}$, given a context, $x \in \mathcal{X}$, is the marginal probability that a random hypothesis, $h \sim \mathbb{Q}$, maps a to x; that is,

$$\pi_{\mathbb{Q}}(a \,|\, x) \triangleq \mathop{\mathbb{E}}_{h\sim\mathbb{Q}} [\mathbb{1}\{h(x) = a\}] . \tag{3}$$

Usually, the hypothesis space consists of functions of a certain parametric form, so the distribution is actually over the parameter values. We analyze one such class in Section 4.

We will analyze Bayesian policies using the *PAC-Bayesian* framework (also known as simply *PAC-Bayes*). The PAC-Bayesian learning paradigm proceeds as follows:

1. We fix a hypothesis space, $\mathcal{H}$, and a *prior* distribution, $\mathbb{P}$, on $\mathcal{H}$.
2. We receive some data, S.
3. Using S, we learn a *posterior* distribution, $\mathbb{Q}$, on $\mathcal{H}$.

In our PAC-Bayesian formulation of CRM, the learned posterior becomes our stochastic policy (Equation 3). Given a

[1] This view of stochastic policies was also used by Seldin et al. [12] to analyze contextual bandits in the PAC-Bayes framework.

context, $x \in \mathcal{X}$, we sample an action by sampling $h \sim \mathbb{Q}$ (independent of x) and returning $h(x)$. (In PAC-Bayesian terminology, this is often called the *Gibbs classifier*.)

Remark 2. We can alternatively view the posterior as a distribution over policies, $\{\pi : \mathcal{X} \to \Delta^k\}$, and the Bayesian policy as the expected policy, $\bar{\pi}_{\mathbb{Q}}(a \,|\, x) \triangleq \mathbb{E}_{\pi \sim \mathbb{Q}}[\pi(a \,|\, x)]$. However, it is more traditional in PAC-Bayes to think in terms of the Gibbs classifier.

It is important to note that the prior cannot depend on the training data; however, *the prior can generate the data*. Indeed, we can generate S by sampling $(x_i, \rho) \sim \mathbb{D}$, $h \sim \mathbb{P}$ and logging $(x_i, h(x_i), \rho(x_i, h(x_i)), \pi_0(h(x_i) \,|\, x_i))$, for $i = 1, \ldots, n$. Thus, in the PAC-Bayesian formulation of CRM, *the prior can be the logging policy*. We elaborate on this idea in Section 4.

3.1. PAC-Bayesian Counterfactual Risk Bounds

The heart of our analysis is an application of the PAC-Bayesian theorem—a generalization bound for Bayesian learning—to upper-bound the counterfactual risk. The particular PAC-Bayesian bound we use is by McAllester [9].

Lemma 1. *Let $\mathbb{D}$ denote a fixed distribution on an instance space, $\mathcal{Z}$. Let $L : \mathcal{H} \times \mathcal{Z} \to [0, 1]$ denote a* loss function*. For a distribution, $\mathbb{Q}$, on the hypothesis space, $\mathcal{H}$, and a dataset, $S \triangleq (z_1, \ldots, z_n) \in \mathcal{Z}^n$, let $R(\mathbb{Q}) \triangleq \mathbb{E}_{h \sim \mathbb{Q}} \mathbb{E}_{z \sim \mathbb{D}}[L(h, z)]$ and $\hat{R}(\mathbb{Q}, S) \triangleq \mathbb{E}_{h \sim \mathbb{Q}} \left[\frac{1}{n} \sum_{i=1}^{n} L(h, z_i)\right]$ denote the risk and empirical risk, respectively. For any $n \geq 1$, $\delta \in (0, 1)$, and fixed prior, $\mathbb{P}$, on $\mathcal{H}$, with probability at least $1 - \delta$ over draws of $S \sim \mathbb{D}^n$, the following holds simultaneously for all posteriors, $\mathbb{Q}$, on $\mathcal{H}$:*

$$R(\mathbb{Q}) \leq \hat{R}(\mathbb{Q}, S) + \sqrt{\frac{2\hat{R}(\mathbb{Q}, S)\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \frac{n}{\delta}\right)}{n-1}} + \frac{2\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \frac{n}{\delta}\right)}{n-1}.$$

The hallmark of a PAC-Bayesian bound is the KL divergence from the fixed prior to a learned posterior. This quantity can be interpreted as a complexity measure, similar to the VC dimension, covering number or Rademacher complexity [10]. The divergence penalizes posteriors that stray from the prior, effectively penalizing overfitting.

One attractive property of this particular bound is that, if the empirical risk is sufficiently small, then the generalization error, $R(\mathbb{Q}) - \hat{R}(\mathbb{Q}, S)$, can be of order $\mathrm{O}(n^{-1})$. Thus, the bound captures both realizable and non-realizable learning problems.

By applying Lemma 1 to an appropriate loss function (the details of which we will not go into here), we obtain a PAC-Bayesian counterfactual risk bound.

Theorem 1. *Let $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{A}\}$ denote a hypothesis space mapping contexts to actions. For any $n \geq 1$, $\delta \in (0, 1)$, $\tau \in (0, 1)$ and fixed prior, $\mathbb{P}$, on $\mathcal{H}$, with probability at least $1 - \delta$ over draws of $S \sim (\mathbb{D} \times \pi_0)^n$, the following holds simultaneously for all posteriors, $\mathbb{Q}$, on $\mathcal{H}$:*

$$R(\pi_{\mathbb{Q}}) \leq \hat{R}_\tau(\pi_{\mathbb{Q}}, S) + \sqrt{\frac{2\left(\frac{1}{\tau} - 1 + \hat{R}_\tau(\pi_{\mathbb{Q}}, S)\right)\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \frac{n}{\delta}\right)}{\tau\,(n-1)}} + \frac{2\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \frac{n}{\delta}\right)}{\tau\,(n-1)}.$$

It is important to note that the truncated IPS empirical risk, $\hat{R}_\tau$, can be negative, achieving its minimum at $1 - \tau^{-1}$. This means that when $\hat{R}_\tau$ is minimized, the middle $\mathrm{O}(n^{-1/2})$ term disappears and the $\mathrm{O}(n^{-1})$ term dominates the bound, yielding the "fast" learning rate. That said, our bound may not be as tight as Swaminathan & Joachims' (Equation 2), since the sample variance can sometimes be smaller than the average. To achieve a similar rate, we could perhaps use Seldin et al.'s PAC-Bayesian Bernstein bound [13].

Theorem 1 assumes that the truncation parameter, τ, is fixed *a priori*. However, using a covering technique, we can derive a counterfactual risk bound that holds for all τ simultaneously—meaning, τ can be data-dependent, such as the 10^{th} percentile of the logged propensities.

Theorem 2. *Let $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{A}\}$ denote a hypothesis space mapping contexts to actions. For any $n \geq 1$, $\delta \in (0, 1)$ and fixed prior, $\mathbb{P}$, on $\mathcal{H}$, with probability at least $1 - \delta$ over draws of $S \sim (\mathbb{D} \times \pi_0)^n$, the following holds simultaneously for all posteriors, $\mathbb{Q}$, on $\mathcal{H}$, and all $\tau \in (0, 1)$:*

$$R(\pi_{\mathbb{Q}}) \leq \hat{R}_\tau(\pi_{\mathbb{Q}}, S) + \sqrt{\frac{4\left(\frac{2}{\tau} - 1 + \hat{R}_\tau(\pi_{\mathbb{Q}}, S)\right)\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \frac{2n}{\delta\tau}\right)}{\tau\,(n-1)}} + \frac{4\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \frac{2n}{\delta\tau}\right)}{\tau\,(n-1)}.$$

Theorems 1 and 2 hold for any fixed prior, but they have an intriguing interpretation when the prior is defined as the logging policy. In this case, one can minimize an upper bound on the counterfactual risk by minimizing the empirical risk while keeping the learned policy close to the logging policy. We explore this idea, and its relationship to variance regularization, in the next section.

4. Mixed Logit Models

We will apply our PAC-Bayesian analysis to the following class of stochastic policies. We first define a hypothesis class,

$$\mathcal{H} \triangleq \{h_{\mathbf{w},\boldsymbol{\gamma}} : \mathbf{w} \in \mathbb{R}^d, \boldsymbol{\gamma} \in \mathbb{R}^k\}, \tag{4}$$

of functions of the form

$$h_{\mathbf{w},\boldsymbol{\gamma}}(x) \triangleq \arg\max_{a \in \mathcal{A}} \mathbf{w} \cdot \phi(x,a) + \gamma_a, \tag{5}$$

where $\phi(x,a) \in \mathbb{R}^d$ outputs features of the context and action, whose norm we assume is uniformly bounded, $\sup_{x \in \mathcal{X}, a \in \mathcal{A}} \|\phi(x,a)\| \leq B$. If each γ_a is sampled from a *standard Gumbel* distribution, $\mathrm{Gumbel}(0,1)$ (location 0, scale 1), then $h_{\mathbf{w},\boldsymbol{\gamma}}(x)$ produces a sample from a *multinomial logit* model,

$$\begin{aligned}\pi_{\mathbf{w}}(a \,|\, x) &\triangleq \frac{\exp(\mathbf{w} \cdot \phi(x,a))}{\sum_{a' \in \mathcal{A}} \exp(\mathbf{w} \cdot \phi(x,a'))} \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\gamma} \sim \mathrm{Gumbel}(0,1)^k} \left[\mathbb{1}\{h_{\mathbf{w},\boldsymbol{\gamma}}(x) = a\}\right].\end{aligned} \tag{6}$$

Further, if $\mathbf{w}$ is normally distributed, then $h_{\mathbf{w},\boldsymbol{\gamma}}(x)$ has a *logistic-normal* distribution [1].

Given some *learned* logit parameters, $\hat{\mathbf{w}} \in \mathbb{R}^d$, we define the posterior, $\mathbb{Q}$, as a $\hat{\mathbf{w}}$-mean, isotropic Gaussian over logit parameters, $\mathcal{N}(\hat{\mathbf{w}}, \sigma^2 \mathbf{I})$, for $\sigma^2 \in (0, \infty)$, with i.i.d. standard Gumbel-distributed perturbations, $\mathrm{Gumbel}(0,1)^k$. As such, we have that

$$\begin{aligned}\pi_{\mathbb{Q}}(a \,|\, x) &= \mathop{\mathbb{E}}_{(\mathbf{w},\boldsymbol{\gamma}) \sim \mathbb{Q}} \left[\mathbb{1}\{h_{\mathbf{w},\boldsymbol{\gamma}}(x) = a\}\right] \\ &= \mathop{\mathbb{E}}_{\mathbf{w} \sim \mathcal{N}(\hat{\mathbf{w}}, \sigma^2 \mathbf{I})} \left[\pi_{\mathbf{w}}(a \,|\, x)\right].\end{aligned} \tag{7}$$

This model is alternately referred to as a *mixed logit* or *random parameter logit*.

We can define the prior in any way that seems reasonable—without access to training data, of course. A common choice of prior is the standard (zero mean, unit variance) multivariate normal distribution. This prior corresponds to L_2 regularization. However, the fact that the data is generated according to a logging policy motivates a different kind of prior (hence, regularizer). A natural assumption is that the logging policy has a multinomial logit form (Equation 6), with parameters $\mathbf{w}_0 \in \mathbb{R}^d$. We can therefore define the prior, $\mathbb{P}$, as a Gaussian centered at the logging policy's logit parameters, $\mathcal{N}(\mathbf{w}_0, \sigma_0^2 \mathbf{I})$, for $\sigma_0^2 \in (0, \infty)$, with i.i.d. standard Gumbel-distributed perturbations, $\mathrm{Gumbel}(0,1)^k$. This does not break the requirement that the prior be independent of the training data, since the prior is fixed before generating the training data. The policy induced by this prior may not correspond to the logging policy, depending on the prior's variance; regardless, we can define the prior any way want, and certain choices for σ_0^2 have nice analytic properties, which we discuss later.

4.1. Bounding the KL Divergence

The KL divergence for the above prior and posterior constructions motivates an interesting regularizer for counterfactual risk minimization. We first derive an upper bound on the KL divergence as a function of the model parameters.

Lemma 2. *For* $\mathbb{P} \triangleq \mathcal{N}(\mathbf{w}_0, \sigma_0^2 \mathbf{I}) \times \mathrm{Gumbel}(0,1)^k$*, and* $\mathbb{Q} \triangleq \mathcal{N}(\hat{\mathbf{w}}, \sigma^2 \mathbf{I}) \times \mathrm{Gumbel}(0,1)^k$*, where* $\mathbf{w}_0, \hat{\mathbf{w}} \in \mathbb{R}^d$ *and* $0 < \sigma^2 \leq \sigma_0^2 < \infty$*,*

$$D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) \leq \frac{\|\hat{\mathbf{w}} - \mathbf{w}_0\|^2}{2\sigma_0^2} + \frac{d}{2} \ln \frac{\sigma_0^2}{\sigma^2}.$$

One implication of Lemma 2, captured by the term $\|\hat{\mathbf{w}} - \mathbf{w}_0\|^2$, is that, to generalize, the learned policy's parameters should stay close to the logging policy's parameters.[2] This intuition concurs with Swaminathan & Joachims's variance regularization [15], which implicitly penalizes diverging from the logging policy; hence, one way to reduce variance is to not stray too far from the logging policy. Implementing this guideline in practice requires a simple modification to the usual L_2 regularization: instead of $\lambda \|\hat{\mathbf{w}}\|^2$ (where $\lambda > 0$ controls the strength of the regularization), use $\lambda \|\hat{\mathbf{w}} - \mathbf{w}_0\|^2$. Of course, this assumes that the logging policy's parameters, $\mathbf{w}_0$, are known; we address the scenario in which the logging policy is unknown in Section 5.

Another implication of Lemma 2 is that the variance parameters of the prior and posterior—σ_0^2 and σ^2, respectively—affect the KL divergence, which can be thought of as the variance of the risk estimator. As we show in Section 4.2, σ^2 also affects the bias of the risk estimator. Thus, selecting these parameters controls the bias-variance trade-off. We discuss this trade-off in Section 4.3.

Remark 3. We used an isotropic Gaussian construction for the posterior to simplify our analysis and presentation. That said, it is possible to define a variance parameter for each dimension of $\mathbf{w}$, which could capture uncertainty in the model estimate for each feature of the context and action.

4.2. Approximating the Action Probabilities

In practice, computing the posterior action probabilities (Equation 7) of a mixed logit model is difficult, since there is no analytical expression for the mean of the logistic-normal distribution [1]. It is therefore difficult to log action probabilities, or to compute the empirical risk (Equation 1), which is a function of the learned and logged action probabilities. However, we can bound the action probabilities by an easily computable function of the mean parameters, $\hat{\mathbf{w}}$.

[2] Interestingly, a similar bound holds for the KL divergence between action distributions, $D_{\mathrm{KL}}(\pi_{\hat{\mathbf{w}}}(x) \| \pi_{\mathbf{w}_0}(x)) \leq \mathrm{O}(\|\hat{\mathbf{w}} - \mathbf{w}_0\|)$, due to Fenchel duality and Cauchy-Schwarz.

Lemma 3. *If* $\sup_{x\in\mathcal{X},a\in\mathcal{A}} \|\phi(x,a)\| \leq B$, *then*

$$\pi_{\hat{\mathbf{w}}}(a\,|\,x)\,e^{-\frac{\sigma^2B^2}{2}} \;\leq\; \pi_{\mathbb{Q}}(a\,|\,x) \;\leq\; \pi_{\hat{\mathbf{w}}}(a\,|\,x)\,e^{\sigma^2B^2}.$$

By Lemma 3, the action probabilities induced by the mean parameters provide lower and upper bounds on the action probabilities of the mixed logit model. The bounds tighten as the variance, σ^2, becomes smaller. For instance, if $\sigma^2 = \mathrm{O}(n^{-1})$, then $\pi_{\mathbb{Q}}(a\,|\,x) \to \pi_{\hat{\mathbf{w}}}(a\,|\,x)$ as $n \to \infty$. During learning, we can use the lower bound of the learned action probabilities to upper-bound the empirical risk. Likewise, when the learned posterior is deployed, we can log the upper bound of the action probabilities, so that future training with the logged data has an upper bound on the IPS empirical risk.

4.3. Bayesian Counterfactual Risk Minimization for Mixed Logit Models

We now state a counterfactual risk bound for the Bayesian policy, $\pi_{\mathbb{Q}}$, in terms of the non-Bayesian policy, $\pi_{\hat{\mathbf{w}}}$, given by the mean parameters, $\hat{\mathbf{w}}$. This bound motivates a new regularized learning objective for Bayesian CRM.

In light of Lemma 3, we overload our previous notation to define a new estimator,

$$\hat{R}_{\tau,\sigma^2}(\mathbf{w},S) \triangleq 1 - \frac{1}{n\,\exp(\frac{\sigma^2B^2}{2})}\sum_{i=1}^{n}\frac{\pi_{\mathbf{w}}(a_i\,|\,x_i)\,r_i}{\max\{p_i,\tau\}}.$$

This estimator is biased, but the bias decreases with σ^2. Importantly, $\hat{R}_{\tau,\sigma^2}(\mathbf{w},S)$ is easy to compute, since it avoids the logistic-normal integral.

The following bound is based on Theorem 1, for fixed τ, though one can easily derive an analogous bound for data-dependent τ using Theorem 2.

Theorem 3. *Let* $\mathcal{H}$ *denote the space of hypotheses defined in Equations 4 and 5, and let* $\pi_{\mathbb{Q}}$ *denote the mixed logit policy defined in Equation 7. For any* $n \geq 1$, $\delta \in (0,1)$, $\tau \in (0,1)$, $\mathbf{w}_0 \in \mathbb{R}^d$ *and* $\sigma_0^2 \in (0,\infty)$, *with probability at least* $1-\delta$ *over draws of* $S \sim (\mathbb{D}\times\pi_0)^n$, *the following holds simultaneously for all* $\hat{\mathbf{w}} \in \mathbb{R}^d$ *and* $\sigma^2 \in (0,\sigma_0^2]$*:*

$$\begin{aligned} R(\pi_{\mathbb{Q}}) &\leq \hat{R}_{\tau,\sigma^2}(\hat{\mathbf{w}},S) \\ &\quad + \sqrt{\frac{\left(\frac{1}{\tau}-1+\hat{R}_{\tau,\sigma^2}(\hat{\mathbf{w}},S)\right)\left(\Gamma(\mathbf{w}_0,\sigma_0^2,\hat{\mathbf{w}},\sigma^2)+2\ln\frac{n}{\delta}\right)}{\tau\,(n-1)}} \\ &\quad + \frac{\Gamma(\mathbf{w}_0,\sigma_0^2,\hat{\mathbf{w}},\sigma^2)+2\ln\frac{n}{\delta}}{\tau\,(n-1)}, \end{aligned} \tag{8}$$

$$\textit{where}\quad \Gamma(\mathbf{w}_0,\sigma_0^2,\hat{\mathbf{w}},\sigma^2) \triangleq \frac{\|\hat{\mathbf{w}}-\mathbf{w}_0\|^2}{\sigma_0^2} + d\ln\frac{\sigma_0^2}{\sigma^2}.$$

Theorem 3 provides an upper bound on the counterfactual risk that can be easily computed with training data. Moreover, the bound is differentiable and smooth, but not convex. We can derive a simpler, convex upper bound; then, absorbing constants into a regularization parameter, λ, we obtain a Bayesian CRM objective.

Proposition 1. *The following convex optimization, with* $\lambda \triangleq \left(\sigma_0^2\,\tau\,(n-1)\right)^{-1}$, *minimizes an upper bound on Equation 8:*

$$\operatorname*{arg\,min}_{\hat{\mathbf{w}}\in\mathbb{R}^d}\;\frac{1}{n}\sum_{i=1}^{n}-\frac{r_i\ln\pi_{\hat{\mathbf{w}}}(a_i\,|\,x_i)}{\max\{p_i,\tau\}} \;+\; \lambda\,\|\hat{\mathbf{w}}-\mathbf{w}_0\|^2. \tag{9}$$

Equation 9 is equivalent to a weighted multinomial logistic regression (with a modified L_2 regularizer). This optimization can be solved by standard methods, with guaranteed convergence to a global optimum.

In practice, one usually tunes λ to optimize the empirical risk on a held-out validation dataset. By Proposition 1, this is equivalent to tuning the variance of the prior, σ_0^2. Though $\mathbf{w}_0$ could in theory be any fixed vector, the case when it is the parameters of the logging policy corresponds to an interesting regularizer. This regularizer instructs the learning algorithm to keep the learned policy close to the logging policy, which effectively reduces the estimator's variance.

Remark 4. In Theorem 3, we can see how the parameters σ_0^2 and σ^2 affect the bias-variance trade-off. Recall that, by Lemma 3, higher values of σ^2 increase the bias of the estimator, $\hat{R}_{\tau,\sigma^2}$. To reduce this bias, we want σ^2 to be small; e.g., setting $\sigma^2 \triangleq \mathrm{O}(n^{-1})$ would result in a negligible bias. However, if $\sigma^2 \ll \sigma_0^2$, then $\Gamma(\mathbf{w}_0,\sigma_0^2,\hat{\mathbf{w}},\sigma^2)$—which can be interpreted as the variance of the estimator—has a term that is $\mathrm{O}(d\ln n)$, which depends on the number of features, d. When d is large, this term can make the risk bound vacuous. We can eliminate this dependence on d by setting $\sigma_0^2 \triangleq \sigma^2$; but if $\sigma^2 = \mathrm{O}(n^{-1})$, then $\Gamma(\mathbf{w}_0,\sigma_0^2,\hat{\mathbf{w}},\sigma^2) = \mathrm{O}(\|\hat{\mathbf{w}}-\mathbf{w}_0\|^2\,n)$, which definitely makes the risk bound vacuous. Thus, a good rule of thumb is: if d is small, set $\sigma_0^2 \triangleq \mathrm{O}(1)$ and $\sigma^2 \triangleq \mathrm{O}(n^{-1})$, which yields $\Gamma(\mathbf{w}_0,\sigma_0^2,\hat{\mathbf{w}},\sigma^2) = \mathrm{O}(\|\hat{\mathbf{w}}-\mathbf{w}_0\|^2 + d\ln n)$ and vanishing bias; and if d is large, set $\sigma_0^2 \triangleq \sigma^2 \triangleq \mathrm{O}(1/\ln n)$, which yields $\Gamma(\mathbf{w}_0,\sigma_0^2,\hat{\mathbf{w}},\sigma^2) = \mathrm{O}(\|\hat{\mathbf{w}}-\mathbf{w}_0\|^2\ln n)$ and slightly larger bias. Of course, one can also learn σ^2 from the training data, but σ_0^2 must be fixed independently.

5. When the Logging Policy Is Unknown

In Section 4, we assumed that the logging policy was known and used it to construct a prior. However, there may be settings in which the logging policy is unknown. We can nonetheless construct a prior that approximates the logging policy by *learning* from its logged actions.

At first, this idea may sound counterintuitive. After all, the prior is supposed to be a fixed distribution that is independent of the training data. However, the expected value, $\mu \triangleq \mathbb{E}_S[f(S)]$, of a function, f, is constant with respect to

S; hence, μ is independent of any realization of the training data. Based on this fact, the expected estimator of the logging policy is independent of the training data, and can thus serve as a valid prior. Further, if the estimator concentrates around its mean, then we can probabilistically bound the distance between the prior and the learned logging policy, which yields a data-dependent regularizer.

Overloading our previous notation, let $L : \mathbb{R}^d \times \mathcal{X} \times \mathcal{A} \to \mathbb{R}_+$ denote a loss function that measures how well a multinomial logit model parameterized by $\mathbf{w}$ can predict action a given context x. We will assume that L is both convex and β-Lipschitz with respect to $\mathbf{w}$. For a dataset, $S \in (\mathcal{X} \times \mathcal{A})^n$, let

$$F(\mathbf{w}, S) \triangleq \frac{1}{n} \sum_{i=1}^{n} L(\mathbf{w}, x_i, a_i) + \lambda \left\|\mathbf{w}\right\|^2 \qquad (10)$$

denote the *regularized empirical risk*. Let

$$\hat{\mathbf{w}}_0(S) \triangleq \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^d} F(\mathbf{w}, S) \qquad (11)$$

denote the minimizer of the regularized empirical risk, or RERM, and let $\bar{\mathbf{w}}_0 \triangleq \mathbb{E}_{S \sim (\mathbb{D} \times \pi_0)^n}[\hat{\mathbf{w}}_0(S)]$ denote the expected RERM. Since $\bar{\mathbf{w}}_0$ is a constant, it is independent of any realization of the training data. We can therefore construct a Gaussian prior around $\bar{\mathbf{w}}_0$, which corresponds to a regularizer proportional to $\left\|\hat{\mathbf{w}} - \bar{\mathbf{w}}_0\right\|^2$.

Due to the strong convexity of F, the RERM exhibits *uniform algorithmic stability*; meaning, it is robust to perturbations of the training data. Because of this property, the random variable $\hat{\mathbf{w}}_0(S)$ concentrates around its mean, $\bar{\mathbf{w}}_0$. We therefore have that, with high probability, the distance from $\hat{\mathbf{w}}_0(S)$ to $\bar{\mathbf{w}}_0$ is small. Thus, by the triangle inequality, $\|\hat{\mathbf{w}} - \bar{\mathbf{w}}_0\|$ is approximately $\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_0(S)\|$, with high probability.

We use this reasoning to prove the following. Our analysis is similar to Lever et al.'s analysis of distribution-dependent PAC-Bayesian priors [6], and uses a concentration bound from Liu et al. [7].

Theorem 4. *Let $\mathcal{H}$ denote the space of hypotheses defined in Equations 4 and 5, and let $\pi_{\mathbb{Q}}$ denote the mixed logit policy defined in Equation 7. Let $\hat{\mathbf{w}}_0(S)$ denote the RERM defined by Equation 11, for a convex, β-Lipschitz loss function. For any $n \geq 1$, $\delta \in (0, 1)$, $\tau \in (0, 1)$ and $\sigma_0^2 \in (0, \infty)$, with probability at least $1 - \delta$ over draws of $S \sim (\mathbb{D} \times \pi_0)^n$, the following holds simultaneously for all $\hat{\mathbf{w}} \in \mathbb{R}^d$ and $\sigma^2 \in (0, \sigma_0^2]$:*

$$R(\pi_{\mathbb{Q}}) \leq \hat{R}_{\tau,\sigma^2}(\hat{\mathbf{w}}, S) + \sqrt{\frac{\left(\frac{1}{\tau} - 1 + \hat{R}_{\tau,\sigma^2}(\hat{\mathbf{w}}, S)\right)\left(\hat{\Gamma}(\hat{\mathbf{w}}_0(S), \sigma_0^2, \hat{\mathbf{w}}, \sigma^2) + 2\ln\frac{2n}{\delta}\right)}{\tau\,(n-1)}} + \frac{\hat{\Gamma}(\hat{\mathbf{w}}_0(S), \sigma_0^2, \hat{\mathbf{w}}, \sigma^2) + 2\ln\frac{2n}{\delta}}{\tau\,(n-1)},$$

where

$$\hat{\Gamma}(\hat{\mathbf{w}}_0(S), \sigma_0^2, \hat{\mathbf{w}}, \sigma^2) \triangleq \frac{\left(\left\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_0(S)\right\| + \frac{\beta}{\lambda}\sqrt{\frac{2\ln\frac{4}{\delta}}{n}}\right)^2}{\sigma_0^2} + d\ln\frac{\sigma_0^2}{\sigma^2}.$$

It is straightforward to show that Proposition 1 holds for Theorem 4 with $\mathbf{w}_0 \triangleq \hat{\mathbf{w}}_0(S)$. Thus, Theorem 4 motivates the following 2-step learning procedure for Bayesian CRM:

1. Using logged data, S, but ignoring the rewards and propensities, train a multinomial logit model (Equation 6) by minimizing the regularized empirical risk (Equation 10), which outputs parameters $\hat{\mathbf{w}}_0(S)$.
2. Using S again, including the rewards and propensities, train a mixed logit model by minimizing the Bayesian CRM objective (Equation 9), using $\mathbf{w}_0 \triangleq \hat{\mathbf{w}}_0(S)$, and setting σ_0^2 and σ^2 however seems appropriate (see Remark 4).

Remark 5. Throughout, we have assumed that the log data includes the action probabilities (propensities), which enables IPS weighting. Given that we can learn to approximate the logging policy, it seems natural to use the learned propensities in the absence of the true propensities. In practice, this may work, though we cannot provide any formal guarantees for this approximation without further assumptions on the true logging policy and analysis of the RERM estimator. We leave this as a task for future work.

6. Conclusion

We have presented a PAC-Bayesian analysis of counterfactual risk minimization, for learning Bayesian policies from logged bandit feedback. Like Swaminathan & Joachims's risk bound (Equation 2), ours achieves a "fast" learning rate under certain conditions—though theirs suggests variance regularization, while ours suggests regularizing by the posterior's divergence from the prior. We applied our risk bound to a class of mixed logit policies, which led to these key insights: (1) to minimize the counterfactual risk, the learned policy should minimize the IPS empirical risk while staying as close as possible to the logging policy, which can be implemented as L_2 regularization; (2) when the logging policy is unknown, one can learn the logging policy from logged data, thus motivating a two-step learning procedure for data-dependent regularization. Though our contributions are theoretical, they suggest practical, actionable advice for practitioners, which we will evaluate empirically in future work.

References

[1] Aitchison, J. and Shen, S. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2): 261–272, 1980.

[2] Bottou, L., Peters, J., nonero Candela, J. Qui Charles, D., Chickering, D., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14: 3207–3260, 2013.

[3] Chapelle, O. and Li, L. An empirical evaluation of Thompson sampling. In *Neural Information Processing Systems*, 2011.

[4] Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, 2011.

[5] Ionides, Edward L. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

[6] Lever, G., Laviolette, F., and Shawe-Taylor, J. Distribution-dependent PAC-Bayes priors. In *Algorithmic Learning Theory*, 2010.

[7] Liu, T., Lugosi, G., Neu, G., and Tao, D. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, 2017.

[8] London, B. and Sandler, T. Bayesian counterfactual risk minimization. *CoRR*, abs/1806.11500, 2018.

[9] McAllester, D. Simplified PAC-Bayesian margin bounds. In *COLT*, pp. 203–215, 2003.

[10] Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, 2012. ISBN 978-0-262-01825-8.

[11] Rosenbaum, P. and Rubin, D. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[12] Seldin, Y., Auer, P., Laviolette, F., Shawe-Taylor, J., and Ortner, R. PAC-Bayesian analysis of contextual bandits. In *Neural Information Processing Systems*, 2011.

[13] Seldin, Y., Laviolette, F., Cesa-Bianchi, N., Shawe-Taylor, J., and Auer, Peter. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

[14] Strehl, A., Langford, J., Li, L., and Kakade, S. Learning from logged implicit exploration data. In *Neural Information Processing Systems*, 2010.

[15] Swaminathan, A. and Joachims, T. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16:1731–1755, 2015.

[16] Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. In *Neural Information Processing Systems*, 2015.