Bounded Off-Policy Evaluation with Missing Data for Course Recommendation and Curriculum Design

William Hoiles WHOILES@UCLA.EDU
University of California, Los Angeles

Mihaela van der Schaar MIHAELA@EE.UCLA.EDU
University of California, Los Angeles

Abstract

Successfully recommending personalized course schedules is a difficult problem given the diversity of students knowledge, learning behaviour, and goals. This paper presents personalized course recommendation and curriculum design algorithms that exploit logged student data. The algorithms are based on the regression estimator for contextual multi-armed bandits with a penalized variance term. Guarantees on the predictive performance of the algorithms are provided using empirical Bernstein bounds. We also provide guidelines for including expert domain knowledge into the recommendations. Using undergraduate engineering logged data from a post-secondary institution we illustrate the performance of these algorithms.

1. Introduction

The goal of this paper is to provide algorithms for personalized course recommendation and curriculum design. In an academic setting it is difficult to apply online reinforcement learning algorithms as this requires controlling the students actions. Standard methods for counterfactual estimation can not be applied for course recommendation as there is an extremely large number of counterfactuals resulting from differences in students knowledge backgrounds, learning behaviour, and learning goals. Missing data resulting from unobserved course schedules and associated grades also impedes learning which courses to recommend. These issues are also present for the optimal design of curricula for students. This paper presents two algorithms for personalized course recommendation and curriculum design that address these issues. The algorithms only require logged student data (i.e. historical data) mitigating the difficulties associated with online data collection. The logged student data depends on the courses taken, their difficulty, concepts taught and style in which they were taught as well as the students preferences, abilities, and knowledge, which are all evolving endogenously as they take classes. Thus, our recommender algorithms consider the evolving students characteristics when issuing recommendations. The algorithms do not just learn the students and courses characteristics and then match them; a richer structure is exploited: the algorithms consider the teaching style, course characteristics, as well as the students evolving characteristics as they take classes. The algorithms can also incorporate expert predictions for missing data resulting in refinement of course and curriculum recommendations. Guaranteed bounds are provided for the algorithms to ensure that reliable course recommendations and curricula are computed. Note that the algorithms can be used to identify curriculum gaps by detecting that certain students have difficulties and get low GPAs in certain classes. Such findings can be brought to the attention of administrators, who can add new classes or remedial materials to reinforce student learning.

Course recommendation and curriculum design can be naturally formulated using contextual multi-armed bandits (e.g. non-sequential, stateless Markov Decision Processes). In contextual multi-armed bandits the logged historical data consists of a collection of context, action, and reward tuples. An example of a context, action, reward tuple is the previous grade point average, the courses selected, and the resulting cumulative grade received. To perform off-policy evaluation we utilize the regression estimator introduced in (Li et al., 2015) in combination with a penalized variance term. The three main contributions of the paper are the finite bounds on the evaluations from the regression estimator, methods to incorporate missing data into the regression estimator, and course recommendation and curric-

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

ula design methods based on the regression estimator.

Evaluation Bounds: Using an extension of Bennet's inequality (Anthony & Bartlett, 2009) we construct an upper bound for the evaluation of a finite number of course scheduling policies. In the case of an infinite number of policies, typical for policy optimization, it is useful to introduce the generalization of the Vapnik-Chervonenkis dimension to real-valued functions known as the covering number (Anthony & Bartlett, 2009; Bartlett et al., 1997; Vovk et al., 2015; Bousquet et al., 2004). The covering number provides a measure of the "complexity" of a class of real-valued functions. As proven, the error in the off-policy evaluation from the regression estimator is bounded by the square-root of the finite-sample variance of the estimator and the covering number.

Missing Data: Logged data may not contain a context-action pair necessary to evaluate a policy. However, as we show, domain knowledge can be effectively introduced into the regression estimator to mitigate this issue with guaranteed bounds on the bias and variance of the estimator.

Recommendations: With the properties of the regression estimator known, we construct penalized variance optimization methods for course schedule recommendation and curricula design using logged data. The design of curricula involves the solutions from binary integer and convex nonlinear programs which can be solved numerically (Achterberg et al., 2008; Achterberg, 2009).

Related Work

There is a substantial amount of literature on recommending relevant courses to students based on their associated knowledge level, learning styles, and feedbacks (Lee & Brunskill, 2012; Mandel et al., 2014; Shishehchi et al., 2011; Chen et al., 2005; Klašnja-Milićević et al., 2011). However, several unique features of the course sequence recommendation problem make these approaches unsuitable. First, students vary tremendously in backgrounds, knowledge and goals. Therefore a personalized course recommendation system is vital to effectively provide course and curricula recommendations. Second, is the design of curricula with specific course constraints. Traditional recommendation systems deal with the problem of recommending a set of courses, however they do not consider providing a sequence of courses. Third, several online recommendation algorithms require active interaction with the student. However, in our setting only the logged data is available. Fourth, we provide methods to include missing data using the domain knowledge from the students or educators to evaluate a course schedule or curricula of interest.

To utilize logged data for course recommendations and curriculum design requires an off-policy estimator (also known as *counterfactual estimation*, *covariate shift estimation* in supervised learning, or *causal effect estimation* in statistics) that estimates how an unobserved policy performs given an observed policy. Recently the regression estimator (Li et al., 2015) has been introduced which satisfies minimax optimality. A limitation with using this estimator directly for the optimal design of a policy, is that no measure of variance is included for the policy evaluation. Recently several off-policy evaluation techniques, that include a variance reduction technique, have been proposed including: truncated importance sampling (Ionides, 2008), doubly robust estimators (Dudík et al., 2011), and the self-normalized estimator (Swaminathan & Joachims, 2015). In (Swaminathan & Joachims, 2015) the authors utilize the theory in (Anthony & Bartlett, 2009) to construct empirical Bernstein bound on the self-normalized estimator. However, a limitation with these estimators is that they contain are biased. In (Li et al., 2012) an unbiased estimator is provided with variance penalization which assumes that the reward can be represented by a generalized linear model (e.g. linear, logistic probit) dependent on the action. For specialized applications such as online news article recommendation systems these assumptions hold, however for student course recommendations we must consider a less restrictive assumption on the rewards. An additional limitation with these estimators is that no guidelines are provided for including missing data. We address all these issues to provide reliable course and curricula recommendations to students and educators.

2. Contextual Multi-Armed Bandits and Logged Data

In this section we formulate the reinforcement learning process of students who select between a set of courses to take given their prior skill level in each course. The goal of the student is to maximize their expected cumulative grade for each course schedule selected. This type of repeated actions-response can be formulated using contextual multi-armed bandits (Langford & Zhang, 2008).

Consider a set of actions $\mathcal{A} = \{1, 2, \ldots, A\}$ representing possible course selections. In each term t, the student selects an action and receives an associated reward via the following process:

1. The student is presented with a context $x_t \in \mathcal{X}$ with $\mathcal{X} = \{1, \ldots, X\}$ a finite set of states which account for the previous grade and skill level of the student. x_t can be computed from their entry grade-point-average (GPA) or scholastic assessment test (SAT) scores.

2. The student, given x_t, then selects an action $a_t \in \mathcal{A}$ of courses via the policy $\pi(a|x)$. That is $a_t \sim \pi(a|x)$ where $\pi(a|x)$ is a discrete probability distribution function over the finite context-action space $\mathcal{X} \times \mathcal{A}$.

3. Given the context x_t and action a_t the student receives a reward $r_t \sim \Phi(r|a_t, x_t)$ where the reward is

generated from some unknown family of distributions $\{\Phi(r|a,x)\}_{a\in\mathcal{A},x\in\mathcal{X}}$. For students the reward is the cumulative grade from completing the courses a_t.

At the completion of each term t the following tuple is generated (x_t, a_t, r_t). Note that each tuple (x_t, a_t, r_t) is generated by an i.i.d. process. After the completion of T terms, the logged dataset $\mathcal{D} = \{(x_t, a_t, r_t)\}_{t=1}^T$ is generated using the policy $\pi_\mathcal{D} \in \Pi(\mathcal{X}, \mathcal{A})$ and reward distributions Φ.

Given the logged dataset $\mathcal{D}$, there are three main problems. First, given $\mathcal{D}$ from a policy $\pi_\mathcal{D}$, how to perform off-policy evaluation. That is, using an unobserved policy $\pi \in \Pi(\mathcal{X}, \mathcal{A})$, it is desirable to have an estimate of the expected reward of this policy:

$$v_\Phi^\pi = \mathbb{E}_{a\sim\pi(\cdot|x),r\sim\Phi(\cdot|a,x)}[r]. \tag{1}$$

The estimate of v_Φ^π given $\mathcal{D}$ is denoted by $\hat{v}_\Phi^\pi$. Important properties of an estimator of v_Φ^π (1) include knowing if the estimator is biased, the variance of the estimated value, and a measure of the number of samples T in $\mathcal{D}$ necessary for $v_\Phi^\pi \to \hat{v}_\Phi^\pi$. The second main problem is how to mitigate the issue of missing data in $\mathcal{D}$. If the expected reward is known for a context-action pair (x, a) then we illustrate how inclusion of this information into $\mathcal{D}$ effects the bias and finite-sample variance of the off-policy estimator for $\hat{v}_\Phi^\pi$. The third main problem is how to utilize results from the off-policy estimator to provide course recommendations to students, and design curricula for educators.

3. Off-Policy Evaluation with Missing Data

To reliably estimate a cumulative grade associated with a new course scheduling policy π using the dataset $\mathcal{D}$, we utilize a combination of counterfactual estimation and missing data inclusion. The algorithms are based on the off-policy regression estimator in (Li et al., 2015) for computing $\hat{v}_\Phi^\pi$ for an unobserved policy π. We prove finite sample convergence bounds for this estimator to ensure the results are reliable. Additionally, methods are presented for including missing data $(x, a) \notin \mathcal{D}$ into the regression estimator for unobserved course schedules.

3.1. Regression Estimator

The regression estimator presented in (Li et al., 2015) is given by:

$$\hat{v}_\Phi^\pi = \sum_{x,a} \mu(x)\pi(a|x)\hat{r}(x,a) \tag{2}$$

$$\hat{r}(x,a) = \frac{\sum_{t=1}^T \mathbf{1}\{x_t = x, a_t = a\} r_t}{\sum_{t=1}^T \mathbf{1}\{x_t = x, a_t = a\}}$$

where $\mu(\cdot)$ is the context distribution, $\pi(\cdot)$ is the policy, $\hat{r}(\cdot)$ is the expected reward, and $\mathbf{1}\{\cdot\}$ denotes the indicator function. The parameter $\mu(\cdot)$ is the probability that a student has the historical performance x. $\pi(a|x)$ is the students course scheduling policy given the performance x. $\hat{r}(x,a)$ is the estimated cumulative grade for the historical performance x and selected course schedule a. Given a logged dataset $\mathcal{D}$ with observed course selection policy $\hat{\pi}_\mathcal{D}$, the regression estimator (2) can be used to estimate the value of a different course selection policy π. The off-policy estimator (2) can be computed in $O(T)$ time complexity, is unbiased, and satisfies minmax optimality. The minimax optimality condition states that $\mathbb{E}[(v_\Phi^\pi - \hat{v}_\Phi^\pi)^2]$ is bounded by some constant. Remarkably the estimator (2) is likely to have a much smaller $\mathbb{E}[(v_\Phi^\pi - \hat{v}_\Phi^\pi)^2]$ compared with popular off-policy estimators such as the propensity score estimator (Bottou et al., 2013).

There are three limitations with using the regression estimator for course recommendations and curriculum design. First, the regression estimator (2) can not be used to evaluate a policy $\pi(x,a)$ that is dependent on a context-action pair $(x,a) \notin \mathcal{D}$. For example, if $\mathcal{D}$ only contains information for students that always select math courses, then the regression estimator can not be used to estimate the performance for a course selection policy that selects chemistry courses. Second, no estimate of the variance of the estimator is provided for a particular policy of interest π. Third, no estimate of how including missing data into the regression estimator impact the estimator's bias or variance. In the following sections we address these issues allowing the regression estimator to be used for course recommendations and curriculum design using the dataset $\mathcal{D}$.

3.2. Finite Sample Convergence

In this section we provide a method to estimate the number of samples T necessary to perform a reliable off-policy evaluation using (2). This is challenging as the regression estimator utilizes the logged data $\mathcal{D}$ generated from a policy $\pi_\mathcal{D} \in \Pi$ that may be very different from the policy $\pi \in \Pi$ we are attempting to evaluate. As an example, if the policy $\pi_\mathcal{D}$ always recommends taking math courses, then it is difficult to evaluate a policy π that always recommends chemistry courses. Two conditions are required to provide a reliable evaluation of π using the information in $\mathcal{D}$. First, the course selection policy π can not deviate significantly from the observed course selection policy $\pi_\mathcal{D}$. Second, there must be sufficient samples in $\mathcal{D}$ to reliably estimate the value of π–that is, the variance associated with the evaluation of π can not be too large.

To construct the off-policy bound we first define the function class $\mathcal{F}_\Pi = \{f_\pi : \mathcal{X} \times \mathcal{A} \mapsto [0,1]\}$ where each f_π represents a policy $\pi \in \Pi$ and is given by:

$$f_\pi(x,a) = M\hat{r}(x,a)\frac{\mu(x)\pi(a|x)}{\hat{\mu}(x)\hat{\pi}_\mathcal{D}(a|x)} = Mu_\mathcal{D}^\pi(a,x). \tag{3}$$

In (3) $\hat{r}(x,a) \in [0,1]$ is the normalized expected re-

ward, $M = \min\{\hat{\mu}(x)\hat{\pi}(a|x)\}$, and $\hat{\mu}(x)\hat{\pi}_{\mathcal{D}}(a|x) = \sum_{t=1}^{T} \mathbf{1}\{a_t = a, x_t = x\}/T$. Using f_π (3) and the theory developed in (Maurer & Pontil, 2009), we bound the variance of (2). The main idea in (Maurer & Pontil, 2009) is to extend Bennet's inequality (Anthony & Bartlett, 2009) to upper bound the value of a random variable by an empirically computed variance. The function f_π (3) satisfies:

$$\begin{aligned} &\mathbb{E}_{x\sim\mu(\cdot)}\mathbb{E}_{a\sim\pi_{\mathcal{D}}(\cdot|x)}[f_\pi(x,a)] = M v_\Phi^\pi, \\ &\mathbb{V}[f(u_{\mathcal{D}}^\pi)] = M^2\mathbb{V}[u_{\mathcal{D}}^\pi] \end{aligned} \tag{4}$$

with $\mathbb{E}[\cdot]$ the expected value and $\mathbb{V}[\cdot]$ the variance. Notice that (4) provides the key relations between the functions f_π and the output of the regression estimator $\hat{v}_\Phi^\pi$ (2) for off-policy evaluation using $\mathcal{D}$. Therefore if we bound the expected value of f_π, then we can bound the expected value of $\hat{v}_\Phi^\pi$ using (4) to make reliable course schedule and curricula recommendations.

Consider a student faced with the following dilemma. They would like to select between two policies, the first π_1 favours course schedules with a significant math component, the second π_2 favours courses with a significant chemistry component. How can the dataset $\mathcal{D}$ be utilized to decide between these two policies. In this two policy case, the finite function class is given by $\mathcal{F}_\Pi = \{f_{\pi_1}, f_{\pi_2} : \mathcal{X}\times\mathcal{A} \mapsto [0,1]\}$. For $\mathcal{F}_\Pi$, $T \geq 2$ and $\gamma > 0$ we have the following probabilistic bound on the estimated value of policies π_1 and π_2:

$$\begin{aligned} &\mathbb{P}\Big(v_\Phi^\pi \leq \hat{v}_\Phi^\pi + \sqrt{\frac{2\hat{\mathbb{V}}[u_{\mathcal{D}}^\pi]\ln(\frac{|\mathcal{F}_\Pi|}{\gamma})}{T}} + \frac{7\ln(\frac{|\mathcal{F}_\Pi|}{\gamma})}{M(T-1)}\Big) \geq 1-\gamma \\ &\hat{\mathbb{V}}[u_{\mathcal{D}}^\pi] = \frac{1}{T-1}\sum_{t=1}^{T}\Big(u_{\mathcal{D}}^\pi(a_t,x_t) - \frac{1}{T}\sum_{\tau=1}^{T} u_{\mathcal{D}}^\pi(a_\tau,x_\tau)\Big)^2. \end{aligned} \tag{5}$$

The probabilistic relation (5) follows directly from (Theorem 4 in (Maurer & Pontil, 2009)) using the function definition (3), statistical relations (4), and the *union bound* from probability. Eq. 5 is a particularly useful result as it provides an empirical bound on the difference between the expected and actual reward $(v_\Phi^\pi - \hat{v}_\Phi^\pi)$ based on the number of samples T in $\mathcal{D}$, and the empirical variance $\hat{\mathbb{V}}[\cdot]$ which is dependent on the policy π. $\hat{\mathbb{V}}[\cdot]$ in (5) encodes the variation in the data–that is, if the student's CGPA (i.e. rewards) deviate significantly for different student historical performances x and course schedules a, or for the ratio $\pi(a|x)/\hat{\pi}_{\mathcal{D}}(a|x)$ of student course scheduling policies, this will result in a high variance. As such, to obtain a reasonable estimate of a course scheduling policy π for such cases requires numerous samples T. Given (5) and $\mathcal{D}$ the student can reliably select the best policy π_1 or π_2 using the results from the regression estimator (2).

Of importance to students and educators is to not only select between a finite number of policies, but to select optimal policies from a continuum. That is, in optimal policy selection problems it is desirable to find a policy $\pi \in \Pi$ where the policy set Π is composed of an infinite number of policies. Given that each π corresponds to a function f_π (3), then the function class $\mathcal{F}_\Pi$ will contain an infinite number of functions. Since the cardinality of the function class $\mathcal{F}_\Pi$ is uncountable we can not utilize (5) to bound the estimated value of $\hat{v}_\Phi^\pi$. To mitigate this issue a different measure of the capacity or complexity of $\mathcal{F}_\Pi$ is required. A commonly used measure of complexity for function classes $\mathcal{F}_\Pi$ is the *uniform covering number* $\mathcal{N}_\infty(\varepsilon, \mathcal{F}_\Pi, T)$ with ε the size of the ε-cover, and T the number of samples (Anthony & Bartlett, 2009; Bartlett et al., 1997; Vovk et al., 2015; Bousquet et al., 2004).

Intuitively from the result in (5), of use for selecting between a finite number of policies, we expect that a similar bound will result using the covering number $\mathcal{N}_\infty(\cdot)$ when selecting between an infinite number of policies. Such a bound is provided by Theorem 1.

Theorem 1 *Let $u_{\mathcal{D}}^\pi(a,x)$ be a random variable with T i.i.d samples in $\mathcal{D}$. Then with probability $1-\gamma$ the random vector $(a_t,x_t) \sim \pi_{\mathcal{D}}$, for a stochastic hypothesis class $\pi \in \Pi$ with covering number $\mathcal{C}(\Pi) = \mathcal{N}_\infty(1/T, \mathcal{F}_\Pi, 2T)$ and $T \geq 16$, satisfies*

$$v_\Phi^\pi \leq \hat{v}_\Phi^\pi + \sqrt{\frac{18\hat{\mathbb{V}}[u_{\mathcal{D}}^\pi]\ln(10\mathcal{C}(\Pi)/\gamma)}{T}} + \frac{15\ln(10\mathcal{C}(\Pi)/\gamma)}{M(T-1)}. \tag{6}$$

□

Note that as a result of the combinatorial lemma by (Vapnik & Chervonenkis, 2015; Sauer, 1972), the covering number for the function class $\mathcal{F}_\Pi$ increases at most polynomially with T. Several interesting insights are provided by Theorem 1 for course and curricula recommendations. First, if very few instances of a particular students historical performance and course schedules (i.e. context-action pairs) are present, then the maximum value of $\pi \in \arg\max\{\hat{v}_\Phi^\pi : \pi \in \Pi\}$ will significantly overestimate v_Φ^π as $\hat{v}_\Phi^\pi$ is upper bounded by the empirical variance in (6). Formally if $M \ll 1$ in (6) then $T \gg 1$ for the regression estimator (2) to reliably evaluate a course scheduling policy π. Second, the difference $v_\Phi^\pi - \hat{v}_\Phi^\pi$ is dependent on the finite sample variance $\mathbb{V}[\cdot]$ and the logarithm of the covering number, also known as the entropy of the class $\mathcal{F}_\Pi$. Qualitatively this is expected as the covering number encodes the complexity of the function class. If $\hat{\pi}_{\mathcal{D}}$ from $\mathcal{D}$ is significantly different from the course scheduling policy π being evaluated, this will introduce a large error in the estimated value of π from (2). To restrict this possibility the constraint $\pi(a,x)/\hat{\pi}_{\mathcal{D}}(a,x) < \beta$ can be applied for $\beta \in \mathbb{R}_+$.

Note that the entropy of the function class $\mathcal{F}_\Pi$ is dependent on all $\pi \in \Pi$, however only the empirical variance $\mathbb{V}[\cdot]$ and $\hat{v}^\pi_\Phi$ in (6) are dependent on the specific policy π being evaluated. That is, as long as the course scheduling policy π is similar to the observed course scheduling policy $\hat{\pi}_\mathcal{D}$, the accuracy of the off-policy evaluation $\hat{v}^\pi_\Phi$ from the regression estimator (2) is dependent on the sample size T and the empirical variance $\mathbb{V}[\cdot]$. The key insight from Theorem 1 is that $v^\pi_\Phi \leq \hat{v}^\pi_\Phi + \lambda\sqrt{\mathbb{V}[u^\pi_\mathcal{D}]/T}$ for some constant $\lambda \in \mathbb{R}_+$. For selecting optimal course scheduling policies $\pi \in \Pi$, if $\sqrt{\mathbb{V}[u^\pi_\mathcal{D}]/T}$ is large then the regression estimator likely overestimate the value of v^π_Φ, therefore it is vital to ensure this variance parameter is not too large.

3.3. Off-Policy Evaluation with Missing Data

In several real-world applications the dataset $\mathcal{D}$ will not contain every context-action pair of interest. For example if there are no examples in $\mathcal{D}$ of a particular course schedule the student is interested in selecting. In such cases the regression estimator (2) can not be used to evaluate $\pi(a|x)$ for $(x, a) \notin \mathcal{D}$ as there is no observation of the reward $\hat{r}(a, x)$. In many cases however it is possible to estimate $\hat{r}(a, x)$, written $\bar{r}(a, x)$, either from expert domain-knowledge or can be inferred from a predictor. In this section we construct mean and variance bounds on the estimated reward from (2) when including $\bar{r}(a, x)$ into the estimator.

Given the predicted CGPA $\bar{r}(a, x)$ for an unobserved course schedule and historical performance pair, Theorem 2 characterizes the bias of $\hat{v}^\pi_\Phi$ from (2) when including this missing information into the estimator.

Theorem 2 *Write $\bar{r}(a, x)$ as the expected reward for the unobserved context-action pair $(a, x) \notin \mathcal{D}$. Then for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$ the estimation of $\hat{v}^\pi_\Phi$ using (2) with the dataset $\mathcal{D} \cup (a, x)$ satisfies:*

$$\mathbb{E}[\hat{v}^\pi_\Phi] - v^\pi_\Phi = \sum_{x,a} \mu(x)\pi(a|x)(1 - p(a, x))^T \Delta r(a, x)$$

with $p(a, x)$ the probability of observing the context-action pair (x, a) and $\Delta r(a, x) = (\bar{r}(a, x) - \hat{r}(a, x))$. □

As expected, Theorem 2 illustrates that if the predicted CGPA $\bar{r}(a, x)$ is equal to the actual CGPA $r(a, x)$ then the estimate $\hat{v}^\pi_\Phi$ from (2) is not biased. If $\Delta r(a, x) \neq 0$ then the incurred bias of (2) can be reduced by selecting a policy that has a small probability of selecting the context-action pair associated with the predicted CGPA $\bar{r}(a, x)$. That is, if the error $\Delta r(a, x)$ is large, then select $\pi(a|x) \ll 1$ to reduced the bias of estimating $\hat{v}^\pi_\Phi$.

Given $\bar{r}(a, x)$, the variance of $\hat{v}^\pi_\Phi$ introduced by including $\bar{r}$ into (2) is given by Theorem 3.

Theorem 3 *Write $\bar{r}(a, x)$ as the expected reward for the unobserved context-action pair $(a, x) \notin \mathcal{D}$. Then for all $x, y \in \mathcal{X}$ and $a, b \in \mathcal{A}$ the estimation of $\hat{v}^\pi_\Phi$ using (2) with the dataset $\mathcal{D} \cup (a, x)$ satisfies:*

$$\mathbb{V}[\hat{v}^\pi_\Phi] = \sum_{x,a} (\mu(x)\pi(a|x))^2 \sigma^2(r) \mathbb{E}\Big[\frac{\mathbf{1}\{n(a, x) > 0\}}{n(a, x)}\Big] + \mathbb{V}\Big[\sum_{x,a} \mu(x)\pi(a|x)\mathbf{1}\{n(a, x) > 0\}\Delta r(a, x) + v^\pi_\Phi\Big],$$

where $\sigma^2(r) = \mathbb{V}[r|a, x]$ is the variance of the reward conditional on the context-action pair. □

The parameter $\mathbb{V}[r|a, x]$ in Theorem 3 can be interpreted as the uncertainty associated with the CGPA for the historical performance x and course schedule a. As expected, if the course scheduling policy $\pi(a|x)$ places more weight on course schedules with uncertain rewards then the variance of the regression estimator (2) increases. Additionally, since $\mathbb{E}[\cdot] \propto (1 - (1 - p(a, x))^T)$ in the first term, if the probability of observing the historical performance and course schedule is low ($p(a, x) \ll 1$), then this will also increase the variance if the policy $\pi(a|x)$ place more weight on (a, x). The second term in Theorem 3 illustrates the variance introduced by using the estimated CGPA instead of the observed CGPA. If every (a, x) is observed then the variance of the second term is $\mathbb{V}[v^\pi_\Phi]$. However, if (a, x) is unobserved then the variance increase by a value proportional to the error $\Delta r(a, x)$ between the estimated and actual CGPA, and the weight from the scheduling policy $\pi(a, x)$. Notice that, unlike the first term in Theorem 3 if $p(a, x) \ll 1$ this will decrease the variance of $\hat{v}^\pi_\Phi$.

Theorem 2 and Theorem 3, provide insight into how including missing data into the estimator (2) increase the bias and variance of $\hat{v}^\pi_\Phi$. To reduce the bias and variance it is vital to ensure that $\Delta r(a, x)$ is minimized.

4. Student Recommendation and Curriculum Design with Variance Penalization

This section presents methods for student course scheduling and curricula design recommendations based on the regression estimator (2).

4.1. Student Recommendation with Sample Variance Mitigation

The objective of the student is to select courses in each term that are likely to give the highest cumulative grade. That is, the student should select the policy π to maximize v^π_Φ (1). Using the logged dataset $\mathcal{D}$ and the regression estimator (2) the student can evaluate an unobserved policy π, yet if we were to just maximize $\hat{v}^\pi_\Phi$ as the objective, the expected reward of π is likely to have a large variance (Theorem 1).

To select a policy $\pi \in \Pi$ that maximizes v_Φ^π with a low variance, we utilize the convex optimization problem:

$$\pi^* \in \arg\max_{\pi\in\Pi} \left\{ \hat{v}_\Phi^\pi - \lambda \sqrt{\frac{\hat{\mathbb{V}}[u_\mathcal{D}^\pi]}{T}} \right\} \tag{7}$$
$$\text{s.t.} \quad \sum_{i=1}^{A} \pi(a_i|x) = 1 \quad \forall x \in \mathcal{X}, \quad \forall i \in \{1,\dots,A\}$$
$$\pi(a_i|x) \geq 0, \quad \pi(a_i|x) = 0 \text{ if } \hat{\pi}_\mathcal{D}(a_i|x) = 0.$$

Including the regularization term $\hat{\mathbb{V}}[\cdot]$ is motivated by the results of Theorem 1. The regularization term balances the maximization of the expected policy v_Φ^π with the accuracy of the estimated policy using the dataset $\mathcal{D}$–formally it accounts for the finite-sample variance associated with estimating a policy π. Notice that for $\hat{v}_\Phi^\pi \geq v_\Phi^\pi$, the empirical variance $\hat{\mathbb{V}}[u_\mathcal{D}^\pi]$ monotonically increases for decreasing λ. Therefore λ can be selected by solving $\min\{\lambda : \hat{\mathbb{V}}[u_\mathcal{D}^\pi] - \hat{\mathbb{V}}[u_\mathcal{D}^{\hat{\pi}_\mathcal{D}}] \leq \delta\}$ with δ a design parameter.

4.2. Curriculum Design with Sample Variance Mitigation

The objective of a curriculum designer is to ensure that students gain sufficient knowledge in a short period of time. Consider a curriculum designer that has a defined number of courses in each course category $k \in \{1,2,\dots,K\}$ that must be taken to graduate. Algorithm 1 provides a method that simultaneously ensures the educators course constraints are met and that students achieve the highest grade possible. The main idea is to construct personalized curricula based on the logged dataset $\mathcal{D}$. To graduate students must complete a defined number courses in each course category–$\boldsymbol{c} = [c(1),\dots,c(K)]$. Given the dataset $\mathcal{D}$, Algorithm 1 constructs the optimal context distribution $\mu(x)$ for a series of possible course graduation dates. Given $\mu(x)$ for a new set of students, it is then possible to provide the curriculum that ensures students graduate in T terms.

Algorithm 1 selects a course curriculum that reliably maximizes a students grades while ensuring the required number of courses are taken. Step 1 selects course schedules that meet the required curriculum constraints by computing all feasible solutions of a multidimensional Knapsack problem for each term of length T. An exhaustive search to solve this problem has time complexity $O(A^{T+1}(K+1))$ with A the number of course schedules, and K the number of course constraints. If we approximate the number of feasible solutions to within $1\pm\epsilon$, (Dyer et al., 1993) provide an algorithm with time complexity $T2^{O((K+1)\sqrt{A}(log(A))^{5/2})}\epsilon^{-2}$ that is subexponential in A. Additionally, Algorithm 1 can be redesigned to allow for courses with general pre-requisites. If $\mathcal{A}_j$ denotes the set of pre-requisite courses necessary for a_j, then the following constraint can be added to (8): $\lfloor \frac{1}{|\mathcal{A}_j|} \sum_{a\in\mathcal{A}_j} b_c(a;\tau) \rfloor = b_c(a_j;t)\ \forall \tau < t$. Given the possible course curricula, the objective is to compute the curriculum that maximizes the students CGPA. Notice that the key feature of Algorithm 1 is that the entrance requirements, or historical grades $\mu(x)$ are optimized for each curriculum $T_c \in \{T_1,\dots,T_C\}$. Thus the course curricula $T_c \in \{T_1,\dots,T_C\}$ are personalized for a set of student with historical performance $\mu(x)$. To ensure that the results from the optimization of $\mu(x)$ (9) are reasonable, comparison is made between the logged historical performance distribution and the optimally computed distribution, as well as ensuring the empirical vari-

Algorithm 1 Curriculum Design Variance Penalization

Step 0: Given the dataset $\mathcal{D}$, select the desired number of courses in each category $\boldsymbol{c} \in \mathbb{Z}_+^K$, and the number of terms $\{T_1, T_2, \dots, T_C\}$ for analysis.

Step 1: Compute all feasible course schedules $a \in \mathcal{D}$ that satisfy the course requirements $\boldsymbol{c}$ by term $T_c \in \{T_1, T_2, \dots, T_C\}$. Write $b_c \in \{0,1\}^{A\times T_c}$ as the binary variable indicating the feasible course schedule for term T_c. The total set of feasible course schedules is given by:

$$M = \left\{ b_c : \sum_{t=1}^{T} b_c(a_i;t)\boldsymbol{n}_i = \boldsymbol{c} \quad \forall i \in \{1,\dots,A\}, \right.$$
$$\left. b \in \{0,1\}^{A\times T_c} \right\}_{c=1}^{C} \tag{8}$$

with $\boldsymbol{n}_i = [n_i(1),\dots,n_i(K)]$ the number of courses in each category associated with taking action a_i.

Step 2: For each feasible solution $b_c \in M$ (8), compute the optimal entrance distribution μ^* by solving the following convex nonlinear program:

$$\mu_c^* \in \arg\max_{\mu} \left\{ \hat{v}_\Phi^\mu - \lambda_c \sqrt{\frac{\hat{\mathbb{V}}[u_\mathcal{D}^\mu]}{T_c}} \right\}$$
$$\text{s.t.} \quad \sum_{i=1}^{A} \mu(x_i) = 1 \quad \forall i \in \{1,\dots,X\} \tag{9}$$
$$\mu(x_i) \geq 0, \quad \mu(x_i) = 0 \text{ if } \hat{\mu}(x_i) = 0.$$

using the dataset $\mathcal{D}_c \subseteq \mathcal{D}$ which contains course schedules a_i if $b_c(a_i;t) \neq 0 \quad \forall t \in \{1,\dots,T_c\}$.

Step 3: Compute the value $\hat{v}_\Phi^\mu$ and variance $\hat{\mathbb{V}}[u_\mathcal{D}^\mu]$ for the logging contextual distribution $\hat{\mu}$.

Step 4: Select the course curricula $\{T_1, T_2, \dots, T_C\}$ with the highest policy value $\hat{v}_\Phi^\mu$ and that have a sufficiently low variance $\hat{\mathbb{V}}[u_\mathcal{D}^\mu]$ that is comparable with the logged policy variance from Step 3.

ance $\mathbb{V}[u^{\mu}_{\mathcal{D}}]$ is sufficiently low. If this is not satisfied then a different λ_c is required for curriculum T_c which can be selected using the method for selecting λ in (7).

5. Real-World Application of Student Recommendation and Curriculum Design

In this section a real-world dataset for undergraduate engineering students, from an accredited post-secondary institution, is used to study the performance of the student recommender (7) and curriculum design (9) methods.

The logged dataset consists of course grades of 920 anonymized students that graduated between the year 2013 and 2015. The logged dataset also includes contextual information of the students including their Scholastic Assessment Test (SAT) scores for *mathematics*, *verbal*, *computers*, *writing* and their high school grade-point-average (GPA). We consider each SAT/GPA score to be either above average or below average which results in a possible context set of $x \in \mathcal{X} = \{1, \dots, 2^5\}$. We group all the possible courses into the following $K = 5$ categories of *math, chemistry, physics, engineering, electives*. If n is the maximum number of courses in each category a student can take per term, this results in $n^K - 1$ possible actions $a \in \mathcal{A}$. Note that in real-world applications the cardinality of $\mathcal{A}$ will be significantly less than $n^K - 1$ as a result of course schedule restrictions. The dataset $\mathcal{D} = \{(x_t, a_t, r_t)\}_{t=1}^{T}$ is generated by defining the rewards r_t as the CGPA achieved as a result of taking the context-action pair (a_t, x_t). The dataset $\mathcal{D}$ contains $A = 112$ unique actions a, $X = 28$ unique contexts x, 965 unique context-action (x, a) pairs, and a total of $T = 10,488$ samples. Using the dataset $\mathcal{D}$ we provide an optimal course selection strategy for students using (7), and design curricula using Algorithm 1.

5.1. Course Schedule Evaluation with Missing Data

In this section we illustrate the impact of including missing data into the regression estimator (2). From Fig. 1, the bias of $\hat{v}^{\pi}_{\Phi}$ increases as the probability of observing the context-action pair (a, x) increases. This is expected from Theorem 2 as the bias of the regression estimator (2) increases as the probability of observing $(a, x) \notin \mathcal{D}$ increase is if the estimated CGPA is not equal to the actual CGPA. Additionally, the empirical variance of $\hat{v}^{\pi}_{\Phi}$ also increases as the $p(a, x)$ increases. This is in agreement with the results in Theorem 3 that illustrate that if the probability of observing the missing pair (a, x) increases then the associated variance of $\hat{v}^{\pi}_{\Phi}$ also increases. Note that for $p(a, x) < 10^{-3}$, the empirical variance of $\hat{v}^{\pi}_{\Phi}$ is small even though $\mathbb{V}[r|a, x]$ is large. From Theorem 3 this results as the course scheduling policy $\pi(a|x)$ places a small weight on (a, x) therefore the variance of $\hat{v}^{\pi}_{\Phi}$ is negligibly effected.

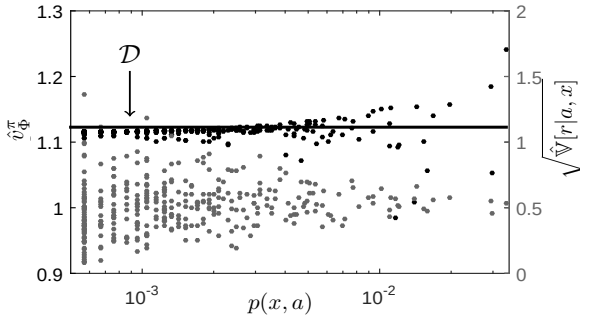

Figure 1. Computed off-policy value $\hat{v}^{\pi}_{\Phi}$ and reward standard deviation $\sqrt{\mathbb{V}[r|a, x]}$ vs. the probability of observing the context-action pair (x, a) (i.e. $p(x, a)$). The dataset is composed of $(x, a) \notin \mathcal{D}$ with all entries (x, a, r) replaced with $(x, a, \bar{r})$, where $\bar{r}$ is the mean value of the reward r. The solid line indicates the policy value given all of $\mathcal{D}$.

5.2. Personalized Student Recommendations

In this section we apply (7) to construct the optimal course scheduling policy for students. The first step is to compute the parameter λ in (7) which balances the maximization of grades with the reliability of the predicted grades. To gain insight into a selection procedure for λ, Fig. 2 plots the optimal off-policy evaluation $\hat{v}^{\pi}_{\Phi}$ and empirical variance $\mathbb{V}[u^{\pi}_{\mathcal{D}}]$ (5) for different values of $0 \leq \lambda \leq 10$. As seen, for small values of λ the computed policy is significantly higher than the logged policy, however the reliability of this off-policy estimate is low. For large values of λ the computed policy value is less than the logged policy–not a useful off-policy evaluation. From the results in Fig. 2, the value of λ that balances the value of the regression estimator $\hat{v}^{\pi}_{\Phi}$ (7) while ensuring a sufficiently low variance $\hat{\mathbb{V}}[u^{\pi}_{\mathcal{D}}]$ (5) is given by a $\lambda = 1$.

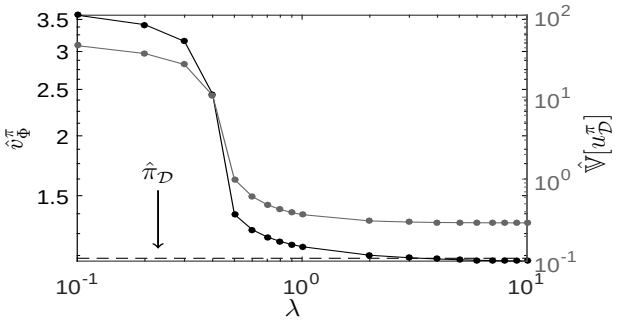

Figure 2. Computed $\hat{v}^{\pi}_{\Phi}$ using (7) for different values of $\lambda \in [0, 10]$. The black line indicates the value of $\hat{v}^{\pi}_{\Phi}$, the gray line is the finite sample variance $\hat{\mathbb{V}}[u^{\pi}_{\mathcal{D}}]$ (5), and the dotted line indicates the value of $\hat{v}^{\pi}_{\Phi}$ for the policy $\hat{\pi}_{\mathcal{D}}$. All computations are performed using the dataset $\mathcal{D}$ defined in Sec. 5.

From Theorem 1 we know that the empirical variance $\hat{\mathbb{V}}[u^{\pi}_{\mathcal{D}}]$ must be sufficiently low to obtain an accurate off-policy estimate, however the ratio of the policies $\pi/\hat{\pi}$ must also not deviate significantly for an accurate course recom-

mendation to be given by $\hat{v}_\Phi^\pi$ (7). Fig. 3 plots the computed policy π^* for $\lambda = 1$, π for $\lambda = 0$, and $\hat{\pi}_\mathcal{D}$. As seen from Fig. 3, the optimization (7) with $\lambda = 1$ ensures that the estimated policy π^* does not deviate significantly from the data generating policy $\hat{\pi}_\mathcal{D}$. In Fig. 3 we see that for $\lambda = 0$ (π) the computed policy has a high $\hat{v}_\Phi^\pi$, however it significantly deviates from $\hat{\pi}_\mathcal{D}$ such that the estimate has a high empirical variance. Additionally, with $\lambda = 1$ the ratio $\pi/\hat{\pi}$ does not deviate significantly between the unique context-action pairs. Therefore the selection $\lambda = 1$ ensures that the optimization (7) can be utilized to recommend course scheduling policies to students.

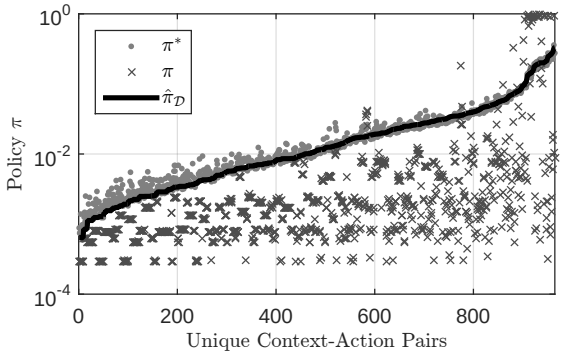

Figure 3. Computed policies π from the solution (7) for $\lambda = 1$ (π^*), $\lambda = 0$ (π), and the data generating policy ($\hat{\pi}_\mathcal{D}$). All computations are performed using the dataset $\mathcal{D}$ defined in Sec. 5.

5.3. Curriculum Design with Personalized Student Recommendations

In this section Algorithm 1 is used to design a course curriculum composed of 3 math, 2 chemistry, 2 physics, 3 engineering, and 1 elective course (i.e. $\boldsymbol{c} = [3, 2, 2, 3, 1]$) using the dataset $\mathcal{D}$. In $\mathcal{D}$ there are a possible $A = 112$ unique course schedules, however several of these schedules will not be feasible as entries in n_i, the total number of each course associated with a_i, will be larger than the entries in $\boldsymbol{c}$. Therefore these can be removed from the possible course schedules leaving a possible $A = 61$ to construct the curriculum. Solving for all feasible course schedules using Step 1 of Algorithm 1, the possible term end dates are $\{3, 4, 5, 6, 7\}$ to complete the $\boldsymbol{c}$ courses. To compute the optimal entrance requirements and curriculum end date, we solve (9) for all possible course schedules and term end dates from Step 1 of Algorithm 1. The variance penalization minimization in (9) ensures that we do not estimate an optimal historical performance distribution (i.e. entrance requirement) that has a high variance. From the solution of (9), the optimal term end date is $T = 6$ which has an expected CGPA of $v_\Phi^\mu = 3.71$. If no optimization is performed on the entrance requirements then the expected CGPA is $v_\Phi^{\hat{\mu}} = 3.35$. If $T = 7$ is selected then the expected CGPA is $v_\Phi^{\hat{\mu}} = 3.30$, therefore it is not always optimal to select the curriculum with the longest duration to maximize the CGPA of students. Insight into how the entrance requirements change for $\hat{\mu}_\mathcal{D}$ compare with μ^* is provided by Fig. 4. Since the ratio $\mu^*/\hat{\mu}_\mathcal{D}$ is sufficiently small, Theorem 1 guarantees that we have maintained a sufficiently low variance on our computed entrance distribution given the observed entrance distribution $\hat{\mu}_\mathcal{D}$ to reliably estimate the CGPA. From Fig. 4, the main difference between μ^* and $\hat{\mu}_\mathcal{D}$ is related to the students that are below average in mathematics and above average in all other courses. The predominant characteristic of the entrance requirement $\hat{\mu}_\mathcal{D}$ and μ^* is that students with above average mathematics skills will perform better in the $T = 6$ curriculum compared with students that are below average in mathematics.

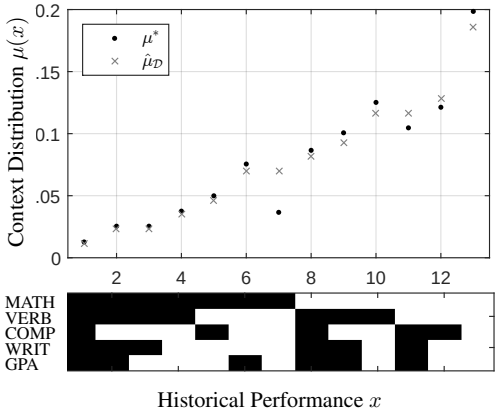

Figure 4. Comparison of the contextual distributions $\hat{\mu}_\mathcal{D}$ from $\mathcal{D}$ and μ^* computed using Algorithm 1 with the parameters defined in Sec. 5. The black squares indicate below average, and the white squares above average for the 13 contexts for the $T = 6$ course curriculum.

6. Conclusion and Future Work

In this paper methods for computing personalized course recommendations and curricula based on the logged data of students are presented. Personalization is a key feature of these methods as students vary tremendously in backgrounds, knowledge, and goals. The methods do not require active interaction with the students. We provide data-driven bounds to ensure that recommendations from the methods are reasonable. Additionally, methods are provided to include missing data into the estimators of use when domain knowledge is available. We illustrate the performance of these methods using logged data from undergraduate engineering students from an accredited post-secondary institution. Extension of the current work includes extending the regression estimator to allow for sequential off-policy evaluations, applicable to several real-world applications. To perform accurate evaluations, confidence bounds can be constructed based on the extension of covering numbers to sequential covering numbers.

Acknowledgments

Funding for the research presented in this paper was provided by the National Science Foundation: Division of Electrical, Communications and Cyber Systems (NSF ECCS 1407712).

References

Achterberg, T. SCIP: solving constraint integer programs. *Mathematical Programming Computation*, 1(1):1–41, 2009.

Achterberg, T., Heinz, S., and Koch, T. Counting solutions of integer programs using unrestricted subtree detection. In *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*, pp. 278–282. Springer, 2008.

Anthony, M. and Bartlett, P. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

Bartlett, P., Kulkarni, S., and Posner, S. Covering numbers for real-valued function classes. *IEEE Transactions on Information Theory*, 43(5):1721–1724, 1997.

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

Bousquet, O., Luxburg, U., and Rätsch, G. Advanced lectures on machine learning. In *ML Summer Schools 2003*, 2004.

Chen, C., Lee, H., and Chen, Y. Personalized e-learning system using item response theory. *Computers & Education*, 44(3): 237–255, 2005.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. *ICML*, pp. 1097–1104, 2011.

Dyer, M., Frieze, A., Kannan, R., Kapoor, A., Perkovic, L., and Vazirani, U. A mildly exponential time algorithm for approximating the number of solutions to a multidimensional knapsack problem. *Combinatorics, Probability and Computing*, 2 (03):271–284, 1993.

Ionides, E. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

Klašnja-Milićević, A., Vesin, B., Ivanović, M., and Budimac, Z. E-learning personalization based on hybrid recommendation strategy and learning style identification. *Computers & Education*, 56(3):885–899, 2011.

Langford, J. and Zhang, T. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pp. 817–824, 2008.

Lee, J. In and Brunskill, E. The impact on individualizing student models on necessary practice opportunities. *International Educational Data Mining Society*, 2012.

Li, L., Chu, W., Langford, Jo., Moon, T., and Wang, X. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. *JMLR*, 26:19–36, 2012.

Li, L., Munos, R., and Szepesvári, C. Toward minimax off-policy value estimation. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 608–616, 2015.

Mandel, T., Liu, Y., Levine, S., Brunskill, E., and Popovic, Z. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 1077–1084. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *COLT*, 2009.

Sauer, N. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

Shishehchi, S., Banihashem, S., Zin, M., and Noah, S. Review of personalized recommendation techniques for learners in e-learning systems. In *Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on*, pp. 277–281. IEEE, 2011.

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

Vapnik, V. and Chervonenkis, Y. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of Complexity*, pp. 11–30. Springer, 2015.

Vovk, V., Papadopoulos, H., and Gammerman, A. Measures of complexity. 2015.