
Counterfactual Learning for Machine Translation: Degeneracies and Solutions

Carolin Lawrence
 Computational Linguistics,
 Heidelberg University,
 Germany

lawrence@cl.uni-heidelberg.de

Pratik Gajane
 INRIA SequeL / Orange labs,
 France
 pratik.gajane@gmail.com

Stefan Riezler
 Computational Linguistics & IWR,
 Heidelberg University,
 Germany
 riezler@cl.uni-heidelberg.de

Abstract

Counterfactual learning is a natural scenario to improve web-based machine translation services by offline learning from feedback logged during user interactions. In order to avoid the risk of showing inferior translations to users, in such scenarios mostly exploration-free deterministic logging policies are in place. We analyze possible degeneracies of inverse and reweighted propensity scoring estimators, in stochastic and deterministic settings, and relate them to recently proposed techniques for counterfactual learning under deterministic logging.

1 Introduction

Machine translation has recently become a commodity service that is offered for free for online translation, e.g., by Google or Microsoft, or is integrated into e-commerce platforms (eBay) or social media (Facebook). Such commercial settings facilitate the collection of user feedback on the quality of machine translation output, either in form of an explicit user rating, or as indirect signal that can be inferred from the interaction of the user with the translated content. While user feedback in form of user clicks on displayed ads has been shown to be a valuable signal in response prediction for online advertising (Chapelle and Li, 2011; Bottou et al., 2013), the gold mine of free user feedback has not yet been exploited in the area of machine translation. Recent research has proposed bandit structured prediction (Sokolov et al., 2016; Kreutzer et al., 2017; Nguyen et al., 2017) for online learning of machine translation from weak user feedback to predicted translations, instead of from costly manually created reference translations. The scenario investigated in these works is still far removed from real world applications in commercial machine translation: Besides the fact that previous research has been confined to simulated feedback, online bandit learning is unrealistic in commercial settings due to the additional latency and the desire for offline testing of system updates before deployment. A natural solution would be to exploit counterfactual learning that reuses existing interaction data where the predictions have been made by a historic system different from the target system. However, both online learning and offline learning from logged data are plagued by the problem that exploration is prohibitive in commercial systems since it means to show inferior translations to users. This effectively results in deterministic logging policies that lack explicit exploration, making an application of off-policy methods theoretically questionable.

Lawrence et al. (2017) recently showed that bandit learning of machine translation is possible even under deterministic logging. They proposed an application of techniques such as doubly-robust policy evaluation and learning (Dudik et al., 2011) or weighted importance sampling (Jiang and Li, 2016; Thomas and Brunskill, 2016) to offline learning from deterministically logged data, and presented evidence from simulation experiments that confirmed their conjecture that these techniques effectively serve to smooth out deterministic components. The purpose of our paper is to give a formal account on the possible degeneracies of the standard inverse propensity scoring technique (Rosenbaum and Rubin, 1983) and its reweighted variant (Kong, 1992) under stochastic and deterministic logging, with the goal of a clearer understanding of the effectiveness of the techniques proposed in Lawrence et al. (2017).

2 Counterfactual Learning for Machine Translation

In the following, we give a short overview of the methods developed in Lawrence et al. (2017). They formalize the problem of counterfactual learning for bandit structured prediction as follows: \mathcal{X} denotes a structured input space, $\mathcal{Y}(x)$ denotes the set of possible structured output for input x , and $\delta : \mathcal{Y} \rightarrow [0, 1]$ denotes a reward function quantifying the quality of structured outputs. A data log is denoted as a set of tuples $\mathcal{D} = \{(x_t, y_t, \delta_t)\}_{t=1}^n$, where for inputs x_t , a logging policy μ produced an output y_t , which is logged with a corresponding reward $\delta_t \in [0, 1]$. In the case of stochastic logging, a propensity score $\mu(y_t|x_t) \in (0, 1]$ is logged in addition. Using the inverse propensity scoring approach (IPS), importance sampling achieves an unbiased estimate of the expected reward under the parametric target policy $\pi_w(y_t|x_t) \in [0, 1]$:

$$\begin{aligned}\hat{V}_{\text{IPS}}(\pi_w) &= \frac{1}{n} \sum_{t=1}^n \delta_t \frac{\pi_w(y_t|x_t)}{\mu(y_t|x_t)} \\ &\approx \mathbb{E}_{p(x)} \mathbb{E}_{\mu(y|x)} [\delta(y) \frac{\pi_w(y|x)}{\mu(y|x)}] \\ &= \mathbb{E}_{p(x)} \mathbb{E}_{\pi_w(y|x)} [\delta(y)].\end{aligned}\tag{1}$$

In case of deterministic logging, outputs are logged with propensity $\mu_t = 1, t = 1, \dots, n$ of the historical system. This results in empirical risk minimization, or empirical reward maximization, without correction of the sampling bias of the logging policy:

$$\hat{V}_{\text{DPM}}(\pi_w) = \frac{1}{n} \sum_{t=1}^n \delta_t \pi_w(y_t|x_t).\tag{2}$$

Lawrence et al. (2017) call equation (2) the *deterministic propensity matching (DPM)* objective, and propose a first modification by the use of weighted importance sampling (Precup et al., 2000; Jiang and Li, 2016; Thomas and Brunskill, 2016). The new objective is the *reweighted deterministic propensity matching (DPM+R)* objective:

$$\begin{aligned}\hat{V}_{\text{DPM+R}}(\pi_w) &= \frac{1}{n} \sum_{t=1}^n \delta_t \bar{\rho}_w(y_t|x_t) \\ &= \frac{\frac{1}{n} \sum_{t=1}^n \delta_t \rho_w(y_t|x_t)}{\frac{1}{n} \sum_{t=1}^n \rho_w(y_t|x_t)},\end{aligned}\tag{3}$$

with $\rho_w(y_t|x_t) = \pi_w(y_t|x_t)$. Setting $\rho_w(y_t|x_t) = \frac{\pi_w(y_t|x_t)}{\mu(y_t|x_t)}$ recovers IPS with reweighting, $\hat{V}_{\text{IPS+R}}(\pi_w)$ (Swaminathan and Joachims, 2015).

Lawrence et al. (2017) present further modifications of Equation (3) by the incorporation of a direct reward estimation method into IPS as proposed in the doubly-robust (DR) estimator (Dudik et al., 2011; Jiang and Li, 2016; Thomas and Brunskill, 2016). Let $\hat{\delta}(x_t, y_t)$ be a regression-based reward model trained on the logged data, and let \hat{c} be a scalar that allows to optimize the estimator for minimal variance (Ross, 2013). They define a *doubly controlled* empirical risk minimization objective

Table 1: Gradients of counterfactual objectives.

$\nabla \hat{V}_{\text{IPS/DPM}} = \frac{1}{n} \sum_{t=1}^n \delta_t \rho_w(y_t x_t) \nabla \log \pi_w(y_t x_t).$
$\nabla \hat{V}_{\text{IPS+R/DPM+R}} = \frac{1}{n} \sum_{t=1}^n [\delta_t \bar{\rho}_w(y_t x_t) (\nabla \log \pi_w(y_t x_t) - \sum_{u=1}^n \bar{\rho}_w(y_u x_u) \nabla \log \pi_w(y_u x_u))].$
$\nabla \hat{V}_{\hat{c} \text{ DC}/\hat{c} \text{ DR}} = \frac{1}{n} \sum_{t=1}^n [(\delta_t - \hat{c} \hat{\delta}) \bar{\rho}_w(y_t x_t) (\nabla \log \pi_w(y_t x_t) - \sum_{u=1}^n \bar{\rho}_w(y_u x_u) \nabla \log \pi_w(y_u x_u)) + \hat{c} \sum_{y \in \mathcal{Y}(x_t)} \hat{\delta}(x_t, y) \pi_w(y x_t) \nabla \log \pi_w(y x_t)].$

$\hat{V}_{\hat{c} \text{ DC}}$ as follows:

$$\hat{V}_{\hat{c} \text{ DC}}(\pi_w) = \frac{1}{n} \sum_{t=1}^n \left[(\delta_t - \hat{c} \hat{\delta}_t) \bar{\rho}_w(y_t|x_t) + \hat{c} \sum_{y \in \mathcal{Y}(x_t)} \hat{\delta}(x_t, y) \rho_w(y|x_t) \right], \quad (4)$$

with $\rho_w(y_t|x_t) = \pi_w(y_t|x_t)$. Setting $\hat{c} = 1$ yields an objective called \hat{V}_{DC} . Setting $\rho_w(y_t|x_t) = \frac{\pi_w(y_t|x_t)}{\mu(y_t|x_t)}$ recovers the standard stochastic doubly-robust estimator $\hat{V}_{\hat{c} \text{ DR}}$. The optimal scalar parameter \hat{c} can be derived easily by taking the derivative of the variance term, leading to $\hat{c} = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)}$.

The learning algorithms in Lawrence et al. (2017) are defined by applying a stochastic gradient ascent update rule $w_{t+1} = w_t + \eta \nabla \hat{V}(\pi_w)_t$ to the objective functions defined above. The gradients are shown in Table 1. In the experiments reported in Lawrence et al. (2017), the policy distribution is assumed to be a Gibbs model $\pi_w(y_t|x_t) = \frac{e^{\alpha(w^\top \phi(x_t, y_t))}}{\sum_{y \in \mathcal{Y}(x_t)} e^{\alpha(w^\top \phi(x_t, y))}}$, based on a feature representation $\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$, a weight vector $w \in \mathbb{R}^d$, and a smoothing parameter $\alpha \in \mathbb{R}^+$, yielding the following simple derivative $\nabla \log \pi_w(y_t|x_t) = \alpha(\phi(x_t, y_t) - \sum_{y \in \mathcal{Y}(x_t)} \phi(x_t, y) \pi_w(y|x_t))$.

3 Degenerate Behaviour in Counterfactual Learning

Both the IPS and the DPM estimators can exhibit a degenerate behavior in that they can be maximized by simply setting all logged outputs to probability 1, i.e., if $\pi_w(y_t|x_t) = 1$ for $\forall (y_t, x_t, \delta_t) \in \mathcal{D} = \{(x_t, y_t, \delta_t)\}_{t=1}^n$. This is the case irrespective of whether data are logged stochastically (IPS) or deterministically (DPM). Obviously, this is undesired as the probability for low reward outputs should not be raised. For abbreviation, we set $\pi_t = \pi_w(y_t|x_t)$ and $\mu_t = \mu(y_t|x_t)$.

Theorem 1. $\max_{\pi} \hat{V}_{\text{IPS}} \wedge \max_{\pi} \hat{V}_{\text{DPM}}$ if $\forall (y_t, x_t, \delta_t) \in \mathcal{D} : \pi(y_t|x_t) = 1 \wedge \delta_t > 0$.

Proof. We start by showing that the value of \hat{V}_{IPS} where $\forall (y_t, x_t, \delta_t) \in \mathcal{D} : \pi_t = 1$ is greater than the value of \hat{V}_{IPS} where $\exists (x_t, y_t, \delta_t) : \pi_t \in [0, 1)$. W.l.o.g. assume that (x_n, y_n, δ_n) is the tuple with $\pi_t \in [0, 1)$. Then

$$\begin{aligned} \sum_{t=1}^n \frac{\delta_t}{\mu_t} &> \sum_{t=1}^{n-1} \frac{\delta_t}{\mu_t} + \frac{\delta_n \pi_n}{\mu_n}, \\ \frac{\delta_n}{\mu_n} &> \frac{\delta_n \pi_n}{\mu_n}, \\ 1 &> \pi_n, \end{aligned} \quad (5)$$

where the last line is true by assumption $\pi_n \in [0, 1)$. Because DPM is a special case of IPS with $\mu_t = 1$ for $\forall (y_t, x_t) \in \mathcal{D} = \{(x_t, y_t, \delta_t)\}_{t=1}^n$, the proof also holds for DPM. \square

The degenerate behavior of IPS and DPM described in Theorem 1 can be fixed by using reweighting, which results in defining a probability distribution over the log \mathcal{D} . Under reweighting, increasing the probability of a low reward output takes away probability mass from the higher reward output. This decreases the value of the estimator, and will thus be avoided in learning.

However, IPS+R and DPM+R still can behave in a degenerate manner, as we will show in the following. We define the set \mathcal{D}^{\max} that contains all tuples that receive the highest reward δ_{\max} observed in the log, and we assume $\delta_{\max} > 0$, leading to a cardinality of \mathcal{D}^{\max} of at least one.

Definition 1. Let $\mathcal{D}^{max} = \max_{\delta} \mathcal{D}$, then $\mathcal{D} = \mathcal{D}^{max} \cup \mathcal{D} \setminus \mathcal{D}^{max}$.

We will show that the estimators can be maximized by simply setting the probability of at least one tuple in \mathcal{D}^{max} to a value higher than 0, while leaving all other tuples in \mathcal{D}^{max} at their probabilities $[0, 1]$, and setting the probability of tuples in the set $\mathcal{D} \setminus \mathcal{D}^{max}$ to 0. Clearly, this is undesired as outputs with a reward close to δ_{max} should not receive a probability of 0. Furthermore, this learning goal is easy to achieve since a degenerate estimator only needs to be concerned about lowering the probability of tuples in $\mathcal{D} \setminus \mathcal{D}^{max}$ as long as there is one tuple of \mathcal{D}^{max} with a probability above 0. We want to prove the following theorem:

Theorem 2. $\max_{\pi} \hat{V}_{IPS+R} \wedge \max_{\pi} \hat{V}_{DPM+R}$ if $\exists (x_t, y_t, \delta_{max}) \in \mathcal{D}^{max} : \pi_t \in (0, 1] \wedge \forall (y_t, x_t, \delta_t) \in \mathcal{D} \setminus \mathcal{D}^{max} : \pi_t = 0$.

We introduce a definition of data indices belonging to the sets \mathcal{D}^{max} and its complement in \mathcal{D} :

Definition 2. Let

$$\hat{V}_{IPS+R}(\pi_w) = \frac{\sum_{t=1}^n \delta_t \frac{\pi_t}{\mu_t}}{\sum_{t=1}^n \frac{\pi_t}{\mu_t}} = \frac{\delta_{max} \sum_{t=1}^{s-1} \frac{\pi_t}{\mu_t} + \sum_{t=s}^n \delta_t \frac{\pi_t}{\mu_t}}{\sum_{t=1}^n \frac{\pi_t}{\mu_t}},$$

where w.l.o.g. indices $(1 \dots (s-1))$ refer to tuples in \mathcal{D}^{max} and indices $(s \dots n)$ refer to indices in $\mathcal{D} \setminus \mathcal{D}^{max}$. Thus, $\mathcal{D}^{max} = \{(x_t, y_t, \delta_{max})\}_{t=1}^{s-1}$ and $\mathcal{D} \setminus \mathcal{D}^{max} = \{(x_t, y_t, \delta_t)\}_{t=s}^n$.

Proof. We need to show that the value of \hat{V}_{IPS+R} where $\pi_t = 0$ for $\forall (y_t, x_t, \delta_t) \in \mathcal{D}^{max}$ is lower than the value of \hat{V}_{IPS+R} where $\exists (x_t, y_t, \delta_{max}) \in \mathcal{D}^{max}$ with $\pi_t \in (0, 1]$. Then

$$\begin{aligned} \frac{\sum_{t=s}^n \delta_t \frac{\pi_t}{\mu_t}}{\sum_{t=s}^n \frac{\pi_t}{\mu_t}} &< \frac{\delta_{max} \sum_{t=1}^{s-1} \frac{\pi_t}{\mu_t} + \sum_{t=s}^n \delta_t \frac{\pi_t}{\mu_t}}{\sum_{t=1}^n \frac{\pi_t}{\mu_t}} \\ 0 &< \frac{\delta_{max} \sum_{t=1}^{s-1} \frac{\pi_t}{\mu_t}}{\sum_{t=1}^n \frac{\pi_t}{\mu_t}}, \end{aligned} \tag{6}$$

where the last line is true for $\delta_{max} > 0$ as long as $\exists (x_t, y_t, \delta_{max}) \in \mathcal{D}^{max}$ with $\pi_t > 0$ as $\mu_t \in (0, 1]$ by definition.

Furthermore, we need to show that the value of \hat{V}_{IPS+R} where $\exists (y_t, x_t, \delta_t) \in \mathcal{D} \setminus \mathcal{D}^{max}$ with $\pi_t \in (0, 1]$ is lower than the value of \hat{V}_{IPS+R} with $\pi_t = 0$ for $\forall (y_t, x_t, \delta_t) \in \mathcal{D} \setminus \mathcal{D}^{max}$.

From the above, it is clear that $\exists (x_t, y_t, \delta_{max}) \in \mathcal{D}^{max}$ with $\pi_t \in (0, 1]$, thus $\frac{\delta_{max} \sum_{t=1}^{s-1} \frac{\pi_t}{\mu_t}}{\sum_{t=1}^{s-1} \frac{\pi_t}{\mu_t}}$ is defined. W.l.o.g. assume that $(y_s, x_s, \delta_s) \in \mathcal{D} \setminus \mathcal{D}^{max}$ is the tuple with $\pi_s \in (0, 1]$. Then

$$\begin{aligned} \frac{\delta_{max} \sum_{t=1}^{s-1} \frac{\pi_t}{\mu_t} + \delta_s \frac{\pi_s}{\mu_s} + \sum_{t=s+1}^n \delta_t \frac{0}{\mu_t}}{\sum_{t=1}^s \frac{\pi_t}{\mu_t} + \sum_{t=s+1}^n \frac{0}{\mu_t}} &< \frac{\delta_{max} \sum_{t=1}^{s-1} \frac{\pi_t}{\mu_t} + \sum_{t=s}^n \delta_t \frac{0}{\mu_t}}{\sum_{t=1}^{s-1} \frac{\pi_t}{\mu_t} + \sum_{t=s}^n \frac{0}{\mu_t}} = \delta_{max}, \tag{7} \\ \delta_{max} \sum_{t=1}^{s-1} \frac{\pi_t}{\mu_t} + \delta_s \frac{\pi_s}{\mu_s} &< \delta_{max} \sum_{t=1}^s \frac{\pi_t}{\mu_t}, \\ \delta_{max} \sum_{t=1}^{s-1} \frac{\pi_t}{\mu_t} + \delta_s \frac{\pi_s}{\mu_s} &< \delta_{max} \sum_{t=1}^{s-1} \frac{\pi_t}{\mu_t} + \delta_{max} \frac{\pi_s}{\mu_s}, \\ \delta_s \frac{\pi_s}{\mu_s} &< \delta_{max} \frac{\pi_s}{\mu_s}. \end{aligned} \tag{8}$$

Equation 8 is true as $\delta_s < \delta_{max}$ by Definition 1. As DPM+R is a special case of IPS+R with $\mu_t = 1$ for $\forall (y_t, x_t) \in \mathcal{D} = \{(x_t, y_t, \delta_t)\}_{t=1}^n$, the proof also holds for DPM+R. \square

While employing stochastic gradient ascent, IPS+R and DPM+R can be prevented from reaching their degenerate state by performing early stopping on a validation set. However, one cannot control

what happens to the probability mass that is freed when lowering the probability of a logged output. The freed probability mass could be allocated to outputs that receive a lower reward than the logged output which would create a system that is worse than the logging system.

The estimators $(\hat{c})DR$ and $(\hat{c})DC$ successfully solve this problem. The direct reward predictor takes the whole output space into account and thus assigns rewards to any structured output. The objective will now be increased if the probability of outputs with high estimated reward is increased, and decreased for outputs with low estimated reward. For this to happen, high reward outputs other than the ones with maximal reward will be considered, even if the outputs have not been seen in the training log. This will shift probability mass to unseen data with high estimated reward, which is a desired property in learning.

4 Experimental Evidence

For completeness, we report the experimental evidence that Lawrence et al. (2017) provide to show the effectiveness of their proposed techniques. They report an application of counterfactual learning in a domain-adaptation setup for machine translation. A model is trained using out-of-domain data using the hierarchical phrase-based machine translation framework that is based on a linear learner. The model is given in-domain data to translate, and outputs are logged together with their per-sentence BLEU score to the true reference, which simulates the reward signal. Experiments are conducted on two language pairs. The first is German-to-English and its baseline system is trained on the concatenation of the Europarl corpus, the Common Crawl corpus and the News corpus. The target domain is represented by a corpus containing transcribed TED talks. The second language pair is French-to-English. Its out-of-domain system is trained on the Europarl corpus and the target domain is the News corpus.

Table 2: BLEU increase over the out-of-domain baseline on validation and test set for deterministically and stochastically created logs.

			BLEU	BLEU difference			BLEU
			out-of-domain	DPM+R	DC	$\hat{c}DC$	in-domain
determin.	TED	validation	22.39	+0.59	+1.50	+1.89	25.43
		test	22.76	+0.67	+1.41	+2.02	25.58
	News	validation	24.64	+0.62	+0.99	+1.02	27.62
		test	25.27	+0.94	+1.05	+1.13	28.08
stochastic	TED	out-of-domain		IPS+R	DR	$\hat{c}DR$	in-domain
		validation	22.39	+0.57	+1.92	+1.95	25.43
		test	22.76	+0.58	+2.04	+2.09	25.58
	News	validation	24.64	+0.71	+1.00	+0.71	27.62
		test	25.27	+0.81	+1.18	+0.95	28.08

As shown in Table 2, under deterministic logging, the best results are obtained by the combining reweighting and double control in the $\hat{c}DC$ method. The relations between the algorithms and even the absolute improvements are quite similar under stochastic logging. For an extended discussion see Lawrence et al. (2017).

5 Discussion

We presented an analysis of possible degenerate behavior in counterfactual learning scenarios. We analyzed the degeneracies of the standard inverse propensity scoring method and its weighted variant, both under stochastic and deterministic logging. Our analysis facilitates a clearer understanding why doubly robust learning techniques serve to avoid such degeneracies, and why such techniques even allow to perform counterfactual learning under deterministic logging.

Lawrence et al. (2017) also discuss a possible implicit exploration effect by the stochastic selection of inputs. This phenomenon has recently been given a formal account by Bastani et al. (2017) and has not been analyzed formally in this paper.

An open question is the application of the techniques proposed by Lawrence et al. (2017) to machine translation with neural networks. For example, the necessity to normalize probabilities over the full set of logged data creates a memory bottleneck which makes it difficult to transfer the reweighting approach to neural networks.

Acknowledgments

The research reported in this paper was supported in part by the German research foundation (DFG).

References

Bastani, H., Bayati, M., and Khosravi, K. (2017). Exploiting the natural exploration in contextual bandits. *ArXiv e-prints*, 1704.09011.

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

Chapelle, O. and Li, L. (2011). An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems (NIPS)*, Granada, Spain.

Dudik, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, Bellevue, WA.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, New York, NY.

Kong, A. (1992). A note on importance sampling using standardized weights. Technical Report 348, Department of Statistics, University of Chicago, Illinois.

Kreutzer, J., Sokolov, A., and Riezler, S. (2017). Bandit structured prediction for neural sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver, Canada.

Lawrence, C., Sokolov, A., and Riezler, S. (2017). Counterfactual learning from bandit feedback under deterministic logging: A case study in statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark.

Nguyen, K., Daumé, H., and Boyd-Graber, J. (2017). Reinforcement learning for bandit neural machine translation with simulated feedback. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark.

Precup, D., Sutton, R. S., and Singh, S. P. (2000). Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, San Francisco, CA.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Ross, S. M. (2013). *Simulation*. Elsevier, fifth edition.

Sokolov, A., Kreutzer, J., Lo, C., and Riezler, S. (2016). Stochastic structured prediction under bandit feedback. In *Advances in Neural Information Processing Systems (NIPS)*, Barcelona, Spain.

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

Thomas, P. S. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, New York, NY.