Statistica Sinica Preprint No: SS-2019-0115	
Title	Comment: Entropy Learning for Dynamic Treatment Regimes
Manuscript ID	SS-2019-0115
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0115
Complete List of Authors	Nathan Kallus
Corresponding Author	Nathan Kallus
E-mail	kallus@cornell.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0115

Statistica Sinica

Comment: Entropy Learning for Dynamic Treatment Regimes

Nathan Kallus

Cornell University

I would like to congratulate Profs. Binyan Jiang, Rui Song, Jialiang Li, and Donglin Zeng (JSLZ, henceforth) for an exciting development in conducting inferences on optimal dynamic treatment regimes (DTRs) learned via empirical risk minimization using the entropy loss as a surrogate. JSLZ's ingenuity was to carefully propagate the asymptotic distributions of M-estimators through a backward induction using a roll out of estimated individualized treatment regimes (ITRs) learned by weighted entropy loss minimization. This solved an open problem on how to conduct rigorous inference on DTRs (Laber et al., 2014).

JSLZ's approach leverages a rejection-and-importance-sampling estimate of the value of a given decision rule based on inverse probability weighting (IPW; see the first unnumbered display equation in JSLZ's Section 2.2) and its interpretation as a weighted (or cost-sensitive) classification, a celebrated reduction (Beygelzimer and Langford, 2009; Zhao et al.,

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0115

2012). Their use of smooth classification surrogates enables their careful approach to analyzing asymptotic distributions. However, even for evaluation purposes, the IPW estimate is problematic. The estimate is a weighted average of rewards, where, for a horizon of T steps, the weights are the product of T indicators of whether the decision rule's recommendations agree with the observed actions, divided by the product of T propensities for the observed actions. With even just two actions per step, the numerator is most often zero. At the same time, the denominator is invariably tiny, and minor differences in probabilities translate into large differences in their inverse products. The result is weights that discard most of the data and are extremely variable on whatever remains. This renders the estimator practically useless for any horizon T longer than 2–3 and any reasonably sized sample (see also Gottesman et al., 2019). So, while JSLZ's careful analysis enables us to conduct inferences on DTRs learned by optimizing this estimate (via a surrogate), one might question whether DTRs learned in this way are useful to begin with when $T \geq 3$ and n is realistic, given the unreliable evaluation.

In this comment, I discuss an optimization-based alternative to evaluating ITRs and DTRs, review several connections, and suggest directions forward. In Kallus (2018a), I proposed an approach for evaluating and learn-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0115

ing ITRs based on *optimal balance*. Optimal balance – a technique I have also developed for designing controlled experiments (Kallus, 2018c), designing observational studies (Kallus, 2017a,b, 2018b; Kallus et al., 2018), and estimating marginal structural models (Kallus and Santacatterina, 2018) – directly targets the error objective of interest by optimally choosing weights that minimize it, rather than relying on plug-in-and-pray approaches that fail for practically sized samples, such as IPW. I show how optimal balance extends to DTR evaluation and discuss why it holds promise.

Balanced Evaluation of ITRs

JSLZ motivate their approach by first considering ITRs; I will do the same. Indeed, using backward induction, evaluating and learning DTRs reduces to evaluating and learning ITRs. In their Eq. (2.1), JSLZ recall the central identity of importance sampling, as applied to ITR evaluation, which I repeat here using potential-outcome notation:

$$V(\mathcal{D} \mid X) \equiv \mathbb{E}\left[R(a) \textstyle\int_{a\in\mathcal{A}} d\mathcal{D}(a \mid X) \mid X\right] = \mathbb{E}\left[\frac{\mathcal{D}(A|X)}{\mathcal{L}(A|X)} R \mid X\right], \qquad (1.1)$$

where $R(a)$ is the potential reward of action a, for any possible action $a \in \mathcal{A}$ (I make no assumptions on $\mathcal{A}$; it can be discrete or continuous); $X \in \mathcal{X}$ are the prognostic covariates; $\mathcal{D}(a \mid X)$ is the probability (usually Dirac) of the decision rule choosing a when seeing X; A and R are the action and reward,

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0115

respectively, observed in the data; $\mathcal{L}(a \mid X)$ is the probability of A, given X, in the data; and we assume ignorable assignment: $R(a) \perp\!\!\!\perp A \mid X \; \forall a \in \mathcal{A}$.

Given a sample $\{(X_i, A_i, R_i) : i \leq n\}$, we can operationalize Eq. (1.1) by taking an empirical average of $\frac{\mathcal{D}(A_i|X_i)}{\mathcal{L}(A_i|X_i)} R_i$ (e.g., JSLZ's Eq (2.3)). However, this can prove problematic in practice, because the density ratio $\frac{\mathcal{D}(A_i|X_i)}{\mathcal{L}(A_i|X_i)}$ can vary wildly, giving some units much higher weight than others and leading to high-variance evaluation. Because of this fundamental problem, there have been many variations and iterations of this basic estimator, including weight normalization and clipping (Swaminathan and Joachims, 2015), "hybrid" clipping using estimates of $\mathbb{E}[R(a) \mid X]$ (Tsiatis and Davidian, 2007; Wang et al., 2017), using such estimates as control variates (Dudík et al., 2011), optimizing the choice of control variate (Cao et al., 2009; Farajtabar et al., 2018), among others. However, these and other estimators that do not rely completely on extrapolation via outcome modeling need to account for the covariate shift between $\mathcal{L}$ and $\mathcal{D}$ and to weight by the density ratio $\frac{\mathcal{D}(A|X)}{\mathcal{L}(A|X)}$, and ultimately suffer from its fundamental instability. This is particularly problematic when $\mathcal{D}(A \mid X)$ is Dirac, as is usually the case since optimal policies are deterministic, because it means that any data point that disagrees with $\mathcal{D}$'s recommendation is discarded, even if informative. Smoothing $\mathcal{D}(A \mid X)$ amounts to shrinking the esti-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0115

mate, by linearity. (When A is continuous, this means *all* data points are discarded; smoothing, as in Kallus and Zhou, 2018, becomes a necessity.)

I briefly explain my optimal balancing proposal for ITR evaluation from Kallus (2018a). Given *any* outcome-weighted estimator, $\hat{V} = \frac{1}{n}\sum_{i\leq n} W_i R_i$, with $W = W(X_{1:n}, A_{1:n})$, its conditional mean squared error, given the data upon which the weights depend, decomposes to:

$$\mathbb{E}\left[\left(\hat{V} - \tfrac{1}{n}\textstyle\sum_{i\leq n} V(\mathcal{D} \mid X_i)\right)^2 \mid X_{1:n}, A_{1:n}\right] = B^2(\mu; W) + \tfrac{1}{n^2}\textstyle\sum_{i\leq n} W_i \sigma_i^2,$$

where $\sigma_i^2 = \mathrm{Var}\left(R_i \mid X_i, A_i\right)$, $\mu(x,a) = \mathbb{E}[R_i \mid X_i = x, A_i = a]$, and

$$B(f; W) = \tfrac{1}{n}\textstyle\sum_{i\leq n} \int_{a\in\mathcal{A}} f(X_i, a) d(W_i \delta(a - A_i) - \mathcal{D}(a \mid X_i)),$$

which, for every W, is a linear operator on the space of functions $[\mathcal{A}\times\mathcal{X} \to \mathbb{R}]$. (A similar result holds if we augment the weighted estimator with an estimate $\hat{\mu}$, as in AIPW.) Because μ (or the difference $\mu - \hat{\mu}$) is unknown, this suggests seeking weights W that make $B(f; W)$ small for many functions $f \in \mathcal{F}$. Under appropriate conditions,

$$\sup_{f\in\mathcal{F}} B(f; W) = \sup_{\|f\|\leq 1} B(f; W) = \|B(\,\cdot\,; W)\|_*,$$

where $\|\cdot\|$ is the gauge of $\mathcal{F}$ and $\|\cdot\|_*$ its dual. Thus, we seek weights W that make the norm of the operator $B(\,\cdot\,; W)$ small, subject to some 2-norm regularization in order to control the variance. Because setting $W_i = \frac{\mathcal{D}(A_i|X_i)}{\mathcal{L}(A_i|X_i)}$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0115

Table 1: ITR evaluation performance in Kallus (2018a, Example 1)

Weights	Outcome Weighting			Augmented OW (DR)			$\|W\|_0$
	RMSE	Bias	SD	RMSE	Bias	SD	
IPW	2.209	−0.005	2.209	4.196	0.435	4.174	13.6 ± 2.9
NIPW	0.519	−0.181	0.487	0.754	0.408	0.634	13.6 ± 2.9
Balanced	**0.280**	0.227	0.163	**0.251**	−0.006	0.251	90.7 ± 3.2

makes $B(f; W)$ a sum of independent mean-zero terms, a straightforward empirical process argument (see, e.g., Pollard, 1990) shows that, under appropriate conditions on $\mathcal{F}$, these weights also make $\|B(\,\cdot\,; W)\|_* \to 0$. However, in practice, these plug-in weights still have all the problems of extreme values and being mostly zeros. Instead, my proposal for optimally balanced evaluation of ITRs is to choose weights that directly optimize the error objective of interest:

$$W^* \in \operatorname{argmin}_{W \geq 0\,:\,\frac{1}{n}\sum_{i\leq n} W_i = 1} \ \|B(\,\cdot\,; W)\|_*^2 + \tfrac{\lambda}{n^2}\|W\|_2^2, \qquad (1.2)$$

which is a linearly constrained convex optimization problem.

To illustrate how this works, I include an excerpt from Kallus (2018a) in Table 1, where I apply this to an example with $|\mathcal{A}| = 5$, $n = 100$, and low overlap between $\mathcal{L}$ and $\mathcal{D}$. For simplicity, I let $\mathcal{F}$ be the unit ball of the RKHS with kernel $\mathcal{K}((x,a),(x',a')) = \delta(a-a')e^{-\|x-x'\|_2^2}$ and $\lambda = 1$. I include augmented (DR) estimators, using $\hat{\mu}$ fitted by XGBoost, as well as normalized (Hájek) IPW. IPW discards about 86% of the data; the balanced approach only 9%, and correspondingly performs much better.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0115

Balanced Evaluation of DTRs

When considering sequential decisions, the fragility of IPW only becomes worse: the weights become even sparser and more extreme, because they are now the ratio of the product of T indicators and the product of T probabilities. Fortunately, the approach to balanced evaluation extends to the case of DTRs, which holds promise for salvaging DTR value estimators that rely on density ratio weighting in any way.

In the sequential setting, we are interested in evaluating the DTR value:

$$V(\mathcal{D}_{1:T}) \equiv \sum_{t \leq T} \left\{ V_t(\mathcal{D}_{1:t}) \equiv \mathbb{E} \int_{a_{1:t} \in \mathcal{A}_{1:t}} R_t(a_{1:t}) d\mathcal{D}_{1:t}(a_{1:t} \mid X_{1:t}(a_{1:t-1}), a_{1:t-1}) \right\},$$

where $\mathcal{D}_{1:t}(a_{1:t} \mid X_{1:t}(a_{1:t-1}), a_{1:t-1}) = \prod_{s \leq t} \mathcal{D}_s(a_s \mid X_{1:s}(a_{1:s-1}), a_{1:s-1})$ and, for each t and sequence of actions $a_{1:t} \in \mathcal{A}_{1:t} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_t$, we now have potential outcomes for both the reward at time t and the time-dependent covariates at time $t+1$. Our data consist of observations of trajectories $X_{1:T}, A_{1:T}, R_{1:T}$, assuming sequentially ignorable assignment:

$$R_{t:T}(a_{1:T}), X_{t+1:T}(a_{1:T-1}) \perp\!\!\!\perp A_t(a_{1:t-1}) \mid X_{1:t}(a_{1:t-1}), A_{1:t-1}(a_{1:t-2}).$$

As in the case of ITRs, consider estimating $V_t(\mathcal{D}_{1:t})$ by a weighted average of outcomes. To streamline the already cumbersome notation, I discuss this in terms of population averages. Thus, I consider the weighted average of observables $\hat{V}_t = \mathbb{E}[W_{1:t} R_t]$, for some weights $W_{1:t} = \prod_{s \leq t} W_s$ where $W_s =$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0115

$W_s(X_{1:s}, A_{1:s})$. Then, iteratively applying sequential ignorability yields a decomposition similar to the ITR case:

$$\hat{V}_t - V_t(\mathcal{D}_{1:t}) = \textstyle\sum_{s \le t} B_s(\mu_{t,s}; W_s), \tag{1.3}$$

$$B_s(f; W_s) \equiv \mathbb{E} \textstyle\int_{a_s \in \mathcal{A}_s} f\left(X_{1:s}, A_{1:s-1}, a_s\right) d\left(W_s \delta\left(a_s - A_s\right) - \mathcal{D}_s(a_s \mid X_{1:s}, A_{1:s-1})\right),$$

$$\mu_{t,s}(x_{1:s}, a_{1:s}) \equiv W_{1:s-1}(x_{1:s-1}, a_{1:s-1}) \mathbb{E}\left[R^{\mathcal{D}}_{t,s}(a_{1:s}) \mid X_{1:s} = x_{1:s}, A_{1:s-1} = a_{1:s-1}\right],$$

$$R^{\mathcal{D}}_{t,s}(a_{1:s}) \equiv \textstyle\int_{a_{s+1:t} \in \mathcal{A}_{s+1:t}} R_t(a_{1:t}) d\mathcal{D}_{s+1:t}(a_{s+1:t} \mid X_{1:t}(a_{1:t-1}), a_{1:t-1}).$$

This looks rather complicated, but has a simple message: the error is a sum over $s = 1, \ldots, t$ of a particular moment mismatch (B_s) in variables $X_{1:s}, A_{1:s}$ between the weighted data distribution and the distribution induced by deviating and following $\mathcal{D}_s$ at step s. Therefore, to obtain a good estimate, we require weights that make this mismatch small for many functions $f : \mathcal{X}_{1:s} \times \mathcal{A}_{1:s} \to \mathbb{R}$. As before, setting $W_s = \frac{\mathcal{D}_s(A_s|X_{1:s}, A_{1:s-1})}{\mathcal{L}_s(A_s|X_{1:s}, A_{1:s-1})}$ achieves this at the population level or for very large samples, but can fail horribly in realistically sized samples. (JSLZ actually use weights $\prod_{s=1}^{T} \frac{\mathcal{D}_s(A_s|X_{1:s}, A_{1:s-1})}{\mathcal{L}_s(A_s|X_{1:s}, A_{1:s-1})}$ on $\sum_{t \le T} R_t$, which is also unbiased, but even more unstable; when estimating the average reward at time t, multiplying by density ratios for times after t is superfluous and just increases the variance.) However, given any sample and some function class $\mathcal{F}_s$, we can seek weights that minimize the (empirical) worst-case mismatches $\|B_s(\,\cdot\,; W_s)\|_{s*}$, subject to some 2-norm

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0115

Table 2: DTR evaluation performance

Weights	$T = 3$			$T = 5$			$T = 7$		
	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD
IPW_T	$5e2$	0.96	$5e2$	$4e4$	-42.94	$4e4$	$2e2$	28.61	$2e2$
IPW	$2e2$	0.41	$2e2$	$1e4$	-11.52	$1e4$	$1e4$	-2.08	$1e4$
NIPW_T	11.82	8.39	8.32	38.07	38.01	2.03	63.10	63.09	0.64
NIPW	6.90	4.64	5.10	26.94	26.27	5.96	51.57	51.22	5.98
Bal. $\mathcal{K}_G$	**6.28**	-0.57	6.26	**11.73**	9.69	6.61	**18.65**	17.44	6.61
Bal. $\mathcal{K}_M$	6.87	-0.26	6.87	12.71	10.06	7.78	19.43	17.80	7.78

regularization to control the variance. Doing so amounts to nothing more than solving Eq. (1.2), for each of $t = 1, \ldots, T$, to obtain W_t, each time considering $X_{1:t}, A_{1:t-1}$ as the "prognostic covariates" being balanced and a_t as the "action." (We could have also placed the $W_{1:s-1}$ term in B_s, rather than in $\mu_{t,s}$, which would have amounted to a simple reweighting of the moment conditions being balanced; however, I focus on the simplest reduction to repeatedly solving problems of the form of Eq. (1.2). We can also apply Eq. (1.3) to the residuals and use an augmented DR-style estimator.)

A DTR Evaluation Example

To demonstrate how this works, I include a simple example. Let T vary and, for $t \leq T$, let $\mathcal{A}_t = \{-1, +1\}$, $\mathcal{X}_t = \mathbb{R}^2$, $R_t(a_{1:t}) = 5a_t + X_{t,1}(a_{t-1}) + \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0,1)$, $X_{1,j} \sim \mathcal{N}(0,1)$, $X_{t+1,j}(a_{1:t}) = a_t + X_{t,j}(a_{t-1}) + \xi_{t,j}$, $\xi_{t,j} \sim \mathcal{N}(0,1)$, $\mathcal{L}(+1 \mid x_{1:t}, a_{1:t-1}) = \text{expit}(2(X_{t,1} + X_{t,2})A_{t-1})$, and $\mathcal{D}(+1 \mid x_{1:t}, a_{1:t-1}) = \mathbb{I}\left[(X_{t,1} + X_{t,2})A_{t-1} < 0\right]$. I consider 2,000 replications of $n = 800$ for each

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0115

$T \in \{3, 5, 7\}$. To apply balanced evaluation, I let $\mathcal{F}_t$ be the unit ball of the RKHS with kernel $\mathcal{K}((x_{1:t}, a_{1:t}), (x'_{1:t}, a'_{1:t})) = \delta(a_{t-1:t} - a'_{t-1:t})\mathcal{K}_x(x_t, x'_t)$, where $\mathcal{K}_x$ is either the Gaussian ($\mathcal{K}_G$) or Matérn ($\mathcal{K}_M$, $\nu = 5/2$) kernel. I compare this with IPW and normalized IPW. I also include the variation in JSLZ in which we multiply $\sum_{t \le T} R_t$ by density ratios up to T, referred to as IPW$_T$.

The results appear in Table 2. The large variance of IPW renders it unusable even with a reasonably sized data set. The variance is so large that it throws off the bias estimated by 2,000 replications (zero in theory). NIPW mitigates this variance, but is actually equal to the uniform weights 37%, 99%, or 100% of the time, for $T = 3, 5, 7$, respectively, and has correspondingly large bias. Balancing has both low bias (indistinguishable from that estimated for IPW) and low variance (comparable to NIPW).

Estimating DTR value when horizons are long is a fundamentally difficult task. Whereas IPW discards most of the data, estimating reward and transition models requires strong modeling assumptions and precarious extrapolations. Balancing could provide a fruitful middle ground: rather than throwing away imperfectly matching trajectories, we imbue the problem with some structure to allow these to be used, while ensuring that our weights achieve the same consistency guarantees afforded by IPW asymp-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0115

totically (see, e.g., Kallus, 2017b, 2018a).

Beyond Evaluation: Learning and Inference

I have argued the merits of using optimal balance to evaluate DTRs. An immediate question is how to use this to learn DTRs. As before, we can optimize the value estimate. Although computationally challenging, this is the approach I took in Kallus (2018a) for ITRs. To apply this to DTRs requires just an application of backward induction with roll out.

With regard to inference (JSLZ's primary concern), this remains open for the balanced approach, but there may be promising directions. Asymptotically, under appropriate conditions on $\mathcal{F}$ and the class of rules being considered, optimal sample weights will uniformly concentrate, so we may consider the distribution when we use the optimal population weights. However, it remains unclear how the estimated rules are distributed (even ITRs). A possible hybrid approach is to use JSLZ's Eq. (2.8), but to replace $\prod_{s\geq t+1} \frac{\mathcal{D}_s(A_s|X_{1:s},A_{1:s-1})}{\mathcal{L}_s(A_s|X_{1:s},A_{1:s-1})}$ with the optimal balancing weights $W^*_{t+1:T}$, while keeping $\frac{\mathcal{D}_t(A_t|X_{1:t},A_{1:t-1})}{\mathcal{L}_t(A_t|X_{1:t},A_{1:t-1})}$ and replacing its numerator with a smooth surrogate. This will at least alleviate issues with longer horizons by limiting IPW to one step, while still being an M-estimator.

While JSLZ's advance is a breakthrough, further advances are neces-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0115

sary. Currently, using IPW and its derivatives to evaluate and learn DTRs when T is moderate and n is realistic is woefully impractical.

References

Beygelzimer, A. and J. Langford (2009). The offset tree for learning with partial labels. In *KDD*, pp. 129–138.

Cao, W., A. A. Tsiatis, and M. Davidian (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika 96*(3), 723–734.

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

Farajtabar, M., Y. Chow, and M. Ghavamzadeh (2018). More robust doubly robust off-policy evaluation. In *ICML*, pp. 1446–1455.

Gottesman, O., F. Johansson, M. Komorowski, A. Faisal, D. Sontag, F. Doshi-Velez, and L. Celi (2019). Guidelines for reinforcement learning in healthcare. *Nat Med 25*(1), 16–18.

Kallus, N. (2017a). A framework for optimal matching for causal inference. In *AISTATS*, pp. 372–381.

Kallus, N. (2017b). Generalized optimal matching methods for causal inference.

Kallus, N. (2018a). Balanced policy evaluation and learning. In *NeurIPS*, pp. 8909–8920.

Kallus, N. (2018b). DeepMatch: Balancing deep covariate representations for causal inference

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0115

using adversarial training.

Kallus, N. (2018c). Optimal a priori balance in the design of controlled experiments. *J Roy Stat Soc B 80*(1), 85–112.

Kallus, N., B. Pennicooke, and M. Santacatterina (2018). More robust estimation of sample average treatment effects using kernel optimal matching in an observational study of spine surgical interventions.

Kallus, N. and M. Santacatterina (2018). Optimal balancing of time-dependent confounders for marginal structural models.

Kallus, N. and A. Zhou (2018). Policy evaluation and optimization with continuous treatments. In *AISTATS*, pp. 1243–1251.

Laber, E. B., D. J. Lizotte, M. Qian, W. E. Pelham, and S. A. Murphy (2014). Dynamic treatment regimes: Technical challenges and applications. *Electron J Stat 8*(1), 1225.

Pollard, D. (1990). Empirical processes: theory and applications.

Swaminathan, A. and T. Joachims (2015). The self-normalized estimator for counterfactual learning. In *NeurIPS*, pp. 3231–3239.

Tsiatis, A. A. and M. Davidian (2007). Comment: Demystifying double robustness. *Stat Sci 22*(4), 569.

Wang, Y.-X., A. Agarwal, and M. Dudik (2017). Optimal and adaptive off-policy evaluation in contextual bandits. In *ICML*, pp. 3589–3597.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0115

Zhao, Y., D. Zeng, A. J. Rush, and M. R. Kosorok (2012). Estimating individualized treatment rules using outcome weighted learning. *J Am Stat Assoc 107*(499), 1106–1118.

School of Operations Research and Information Engineering and Cornell Tech, Cornell University, New York, NY 10044, USA.

kallus@cornell.edu

Statistica Sinica