Scalable Synthetic Coupling with Inference-Aware Optimization*

YoonHaeng Hur† and Tengyuan Liang‡

Abstract. Synthetic control has emerged as a powerful tool for estimating the effects of policy interventions in economics and social sciences. Despite many proposed extensions and improvements, the scalability of synthetic control remains a concern, particularly when the number of control units is large or multiple treated units are present. We address this issue by proposing and analyzing algorithms that efficiently solve the synthetic control problem with nearly dimension-free convergence rates. The key is entropic regularization, which enables us to explore the simplex of weights under a geometry based on the Kullback-Leibler divergence, calling the Sinkhorn algorithm as a subroutine. We delve into a statistical interpretation of entropic regularization in light of individualized inference, developing new insights into our formulation as an inference-aware optimization. The proposed algorithms enhance the scalability of synthetic control, thereby expanding its applicability to large-scale problems.

Key words. missing value imputation, optimal coupling, matching, entropic regularization, Sinkhorn algorithm

MSC codes. 90C25, 90C20, 62G20, 62P20

1. Introduction.

Synthetic control, initially proposed by [4] and popularized after [3], is now a mainstream tool for estimating the effects of policy interventions in economics and social sciences. The central idea is to estimate the counterfactual outcome of a treated unit by a certain weighted average of control units, where the weights are chosen to minimize the approximation error between the treated unit's covariate and the convex combination of the control units' covariates. Theoretical and algorithmic questions arising from synthetic control have attracted considerable attention from the econometrics, statistics, and computer science communities, while its scope of applications has expanded to various fields, including epidemiology and engineering, among others [1]. Quoting from [2]: "... a new literature has emerged in statistics, econometrics, and machine learning to extend and improve the synthetic control methodology, to study the properties of synthetic control estimators, and to adapt the method to better handle particular data configurations."

Synthetic control is essentially an optimization problem to find the optimal weights for the convex combination of control units that best approximates the treated unit in terms of covariates. In a nutshell, given a treated unit with covariate $x^t \in \mathbb{R}^d$ and a group of n control units with covariates $\{x_i^c\}_{i=1}^n \subset \mathbb{R}^d$, synthetic control solves the following convex quadratic optimization problem with a simplex constraint:

$$\min_{p=(p_1,\dots,p_n)\in\Delta_n} \|x^t - \textstyle\sum_{i=1}^n p_i x_i^c\|_2^2, \tag{1.1}$$

where $\Delta_n = \{a \in \mathbb{R}_+^n : \sum_{i=1}^n a_i = 1\}$ denotes the probability simplex. For a solution $p \in \Delta_n$ to (1.1), we expect the convex combination $\sum_{i=1}^n p_i x_i^c$ to approximate x^t well, and thus the

*Submitted to the editors DATE.

Funding: Liang acknowledges the generous support from the NSF Career Grant (DMS-2042473), and the support from the Wallman Society of Fellows.

†Department of Statistics, University of Chicago, Chicago, IL (yoonhaenghur@uchicago.edu).

‡Booth School of Business, University of Chicago, Chicago, IL (tengyuan.liang@chicagobooth.edu).

corresponding convex combination of the outcomes $\{Y_i^c\}_{i=1}^n$ observed for the control units, namely, $\sum_{i=1}^n p_i Y_i^c$, serves as an estimate for the counterfactual outcome of the treated unit. One may immediately notice similarities to the nearest neighbor matching—a simple instance of widely adopted matching techniques [60]—that also takes the weighted average $\sum_{i=1}^n p_i Y_i^c$ with $p_i = \frac{1}{k}$ for i's that are the k nearest neighbors of the treated unit in terms of the distance $\|x^t - x_i^c\|_2$, while $p_i = 0$ for the rest. If the treated unit is extremely far away from all control units, then the nearest neighbor matching is likely to commit a large estimation error, but the convex combination of synthetic control can simultaneously reduce the approximation error and the variance driven by the idiosyncratic errors under certain conditions. Figure 1 visualizes such a situation. Of course, this benefit comes at the cost of solving (1.1).

While (1.1) can be solved by generic convex solvers relying on the interior point method, their complexity is $O(n^3)$ or worse, which becomes prohibitive as n, the number of control units, grows. [39, 27] propose modifications to the synthetic control by dropping the simplex constraint and imposing some regularization, resulting in familiar regularized least squares problems. These modifications, however, no longer seek a convex combination of control units, and thus the interpretation and characteristics of synthetic control—detailed in Section 4 of [1]—are lost. The computational burden of synthetic control is much exacerbated when we consider multiple treated units, namely, solving (1.1) multiple times for different values of x^t. Instead of solving multiple synthetic control problems, [33, 44, 55, 17] aggregate the treated units and solve a single synthetic control problem for the pooled treated units. [7] tackles synthetic control for multiple treated units with certain penalization to promote uniqueness of the solution and reduce the bias. While [7] develops appealing theoretical properties, they do not address how to solve multiple synthetic control problems efficiently.

Our Contributions. The above discussion manifests a gap in the synthetic control literature regarding the computation, especially in the presence of many control units and multiple

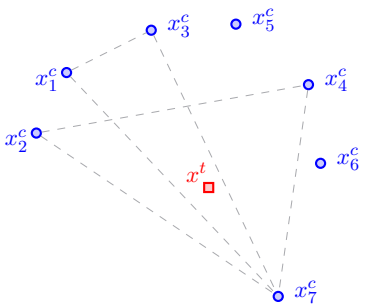


Figure 1. *Visualization of covariate/feature approximation. The blue circles are the control units, and the red square is the treated unit. The three nearest control units to the treated unit are x_4^c, x_6^c, and x_7^c. If all control units are far away from the treated unit, the average of these three is a poor approximation of x^t. Meanwhile, synthetic control finds the optimal convex combination of control units to minimize the approximation error. Any convex hull containing x^t gives a perfect approximation; for instance, the convex hull of (x_1^c, x_3^c, x_7^c) or (x_2^c, x_4^c, x_7^c)—shown as dashed triangles—is a valid choice. Under the linear relationship between the covariates and outcomes, synthetic control better estimates the counterfactual outcome than the nearest neighbor matching. The idea generalizes to features—certain transformations of the covariates—that are not necessarily linear, which can model nonlinear relationships between the covariates and outcomes, as we develop in Section 2.2.*

treated units. Such cases cannot be managed by generic convex solvers as the complexity becomes prohibitively large. In this paper, we fill this gap by proposing and analyzing concrete algorithms to solve the synthetic control even when the number of units—both treated and control—is large. The key is to consider entropic regularization to the optimization problem (1.1): for some $\lambda > 0$, we solve

$$\min_{p \in \Delta_n} \left(\|x^t - \textstyle\sum_{i=1}^n p_i x_i^c\|_2^2 + \lambda \displaystyle\sum_{i=1}^n p_i \log \tfrac{p_i}{e} \right). \tag{1.2}$$

We first clarify that the entropic regularization has already been proposed in the literature. For example, [33] introduces entropic regularization for covariate balancing. Indeed, entropic regularization promotes delocalization of the weights p_i's and thus can help reduce the variance of the estimation. Most importantly, it enables us to explore the simplex Δ_n under the geometry based on the Kullback-Leibler divergence, which allows for algorithms that solve multiple synthetic control problems simultaneously via parallelization, resulting in a complexity that is almost independent of the number of units. Furthermore, we analyze the entropic regularized synthetic control through the lens of individual uncertainty quantification, developing deeper insights into the regularization and its implications for the synthetic control.

Organization. The rest of the paper is organized as follows. After reviewing the related literature below, we formally state in Section 2 the entropic regularized synthetic control problem for multiple treated units simultaneously, where we also introduce nonlinear extensions and the coupling constraint devised to control the average treatment effect. In Section 3, we introduce and analyze two algorithms to solve the entropic regularized synthetic control problem, showing almost dimension-free convergence rates. Section 4 provides a statistical interpretation of the entropic regularization in light of individual uncertainty quantification. All technical proofs are collected in Appendix A. A real data application is deferred to Section SM1 of the supplementary materials due to space limit.

1.1. Related Literature. Matching [60, 57] is a widely adopted method for imputing counterfactual outcomes. For each subject, the counterfactual outcome is estimated by identifying units in the opposite treatment group similar to the subject in covariates and then averaging their outcomes. Due to its simplicity, nearest neighbor matching, which finds subjects closest under a suitable distance between the covariates, is favored in practice. However, it may produce a biased average treatment effect (ATE) estimate, which motivated certain bias correction methods [6, 5]. Exact or approximate matching with high-dimensional covariates can be difficult, and thus propensity score matching was proposed as a remedy [58]. There are more sophisticated matching methods to improve balance or forbid certain matches, often referred to as the optimal matching [57, 69]; they require solving combinatorial optimization problems by network optimization techniques or mixed integer programming, which can be computationally expensive. However, even analyzing the average treatment effect (ATE) estimates resulting from combinatorial optimization problems is highly nontrivial, and these estimates may differ from simple ATE estimates based on propensity score weighting [56, 36, 37] that enjoy nice statistical properties. The algorithms developed in this paper have much better complexity than the above optimal matching, while the bias in the average-level estimate will be controlled by the coupling constraint.

Regression imputation is another approach to estimating counterfactual outcomes. The idea is simple: under the ignorability or unconfoundedness assumption [58], the underlying functions that map the covariates to the responses under the treatment and the control are identifiable and estimable based on observational data. It translates a causal inference problem into a regression problem. Parametric or nonparametric regression techniques estimate the conditional response function under the treatment and the control, given the covariates, and in turn find the conditional average treatment effect (CATE), see [32, 34, 35, 47] for classical results and [12, 38, 43, 51, 45] for more recent nonparametric approaches. At the aggregate level, semiparametric efficient estimation of ATE leveraging regression imputation techniques is studied in [32]. Admittedly, if the quantity of interest is at the aggregate-level—a one-dimensional parameter such as ATE or ATE on the treated—the specific nonparametric regression technique matters less, as long as it estimates the function reasonably well. There has been a fruitful line of literature on the use of flexible machine learning methods in estimating CATE [15, 14, 23, 29, 30, 66].

Synthetic control was originally proposed for panel data [4, 6], where the outcomes of each unit are observed over time. See [1, 2] for an overview of the literature. The use of regularization to the synthetic control problem has been proposed in the literature, such as elastic net [27] and entropic regularization [33]. While any type of regularization can be added to the synthetic control problem, entropic regularization is particularly appealing as it allows for efficient algorithms to solve the problem, which is the main focus of this paper. Importantly, as we unveil in Section 4, entropic regularization plays a crucial role in controlling the uncertainty of the individual-level estimates, which is a unique insight of this paper. Combined with the existing techniques and insights from the causal panel data literature [11], a number of extensions of synthetic control have been proposed [9, 13, 16]. Statistical inference on synthetic control has also been an active area of research [46, 22, 10, 24], mainly studied under the factor model structure suited for causal panel data. As the main focus of this paper is on the optimization in the presence of multiple treated units, we do not specifically consider the panel data setting; instead, we simply treat any pretreatment outcomes as covariates.

As we will see in Section 3, the entropic regularized synthetic control problem with the coupling constraint minimizes a quadratic objective function over a coupling matrix, which can be seen as an extension of the optimal transport problem [65, 53]. In the standard optimal transport problem minimizing a linear objective function over a coupling matrix, entropic regularization and the resulting Sinkhorn algorithm [26] have shown to enjoy much better complexity than generic linear programming solvers [8, 21, 31]. As in the synthetic control problem, quadratic objective functions arise naturally in many applications. Gromov-Wasserstein (GW) distance [48] is an important example, which has been increasingly applied in machine learning and statistics [67, 20, 41, 49]. The GW distance involves a nonconvex quadratic objective function and thus is NP-hard to solve in general, which motivated entropic regularization [54]. Convex quadratic objectives arise naturally in many applications under different contexts [18, 28, 52]. However, the computational aspect of these problems has not received much attention. As mentioned earlier, generic convex solvers are computationally expensive for large-scale problems. The proposed algorithms provide solutions to the computational challenges of optimal transport with quadratic objectives, which can be applied to a wide range of applications.

Notation. We denote by $\|\cdot\|_2$ and $\langle\cdot,\cdot\rangle$ the Euclidean norm and inner product, respectively. For a square matrix A, let $\mathrm{tr}(A)$ denote its trace. Let $\langle A, B\rangle := \mathrm{tr}(A^\top B)$ denote the Frobenius inner product between matrices A and B of the same dimension. For a matrix $A \in \mathbb{R}^{n\times m}$, let $\|A\|_1 := \sum_{i=1}^n \sum_{j=1}^m |A_{ij}|$ and $\|A\|_\infty := \max_{1\le i\le n} \max_{1\le j\le m} |A_{ij}|$ denote its L^1 norm and L^∞ norm, respectively. For any integer $n \in \mathbb{N}$, let $1_n = (1,\dots,1) \in \mathbb{R}^n$ denote the vector whose entries are all 1, let $\Delta_n := \{a \in \mathbb{R}^n_+ : \sum_{i=1}^n a_i = 1\}$ denote the simplex of probability vectors, and let $\Delta_n^+ := \{a \in \Delta_n : a_i > 0 \ \ \forall i\}$ denote the interior of Δ_n.

2. Entropic Regularized Synthetic Control.

Following the standard potential outcome framework [50, 59] for a binary treatment, we consider n control units and m treated units. We denote the covariate vector of the i-th control unit by $x_i^c \in \mathbb{R}^d$ and that of the j-th treated unit by $x_j^t \in \mathbb{R}^d$. The observed outcome of the i-th control unit is denoted by $Y_i^c \in \mathbb{R}$ and that of the j-th treated unit by $Y_j^t \in \mathbb{R}$. For each treated unit j, the primary goal is to estimate its counterfactual outcome, denoted as $Y_j^t(0)$, which is the outcome of the treated unit had it not been treated.

The standard synthetic control method estimates $Y_j^t(0)$ by a convex combination of the outcomes of the control units, $\sum_{i=1}^n p_{ij} Y_i^c$, where $(p_{1j},\dots,p_{nj}) \in \Delta_n$ is a solution to (1.1) with x_j^t in place of x^t. The main challenge is to obtain the solution $(p_{1j},\dots,p_{nj})$ efficiently for all $j = 1,\dots,m$. We tackle this problem by imposing entropic regularization (1.2), which we can solve simultaneously for all j's using the Sinkhorn algorithm [62, 26].

2.1. Synthetic Coupling to Control the ATT.

Our goal is to solve m instances of (1.2) for $x^t \in \{x_1^t,\dots,x_m^t\}$ simultaneously. With an additional coupling constraint which we will explain below (hence called synthetic coupling), the task is equivalent to solving the following optimization problem:

$$\min_{\pi\in\mathbb{R}_+^{n\times m}} \quad \frac{1}{2m}\sum_{j=1}^m \|x_j^t - \textstyle\sum_{i=1}^n \frac{\pi_{ij}}{1/m} x_i^c\|_2^2 + \lambda \sum_{i=1}^n\sum_{j=1}^m \pi_{ij}\log\frac{\pi_{ij}}{e}, \tag{2.1}$$

$$\text{subject to} \quad \sum_{i=1}^n \pi_{ij} = \frac{1}{m} \quad \forall j = 1,\dots,m, \tag{2.2}$$

$$\sum_{j=1}^m \pi_{ij} = \frac{1}{n} \quad \forall i = 1,\dots,n. \tag{2.3}$$

First, we observe that solving (2.1) under (2.2) is exactly the same as solving m instances of (1.2) separately, each with a different x_j^t. Here, the objective function (2.1) is essentially the average of the objective functions of m instances of (1.2), which is separable in j, where $(\frac{\pi_{1j}}{1/m},\dots,\frac{\pi_{nj}}{1/m}) \in \Delta_n$ is the convex combination of the control units' covariates to approximate the covariate of the j-th treated unit. The variable $\pi \in \mathbb{R}_+^{n\times m}$, which we call a coupling matrix, collects the weights for all treated units so that its j-th column corresponds to the weights for the j-th treated unit.

Coupling Constraint. The additional constraint (2.3) serves as an optional device to control the average treatment effect on the treated (ATT). Conceptually, (2.3) requires that all control units contribute equally by having the same amount of total weights across all treated units.

By algebra, this constrains the ATT to be equal to the difference-in-means of the treated and control units, which is a desirable average-level estimator in the randomized experiment setting. Let $\widehat{\pi}$ be a solution to (2.1) under (2.2) and (2.3). Then, we estimate the counterfactual outcome $Y_j^t(0)$ of treated unit j by the convex combination $\widehat{Y}_j^t(0) := \sum_{i=1}^n \frac{\widehat{\pi}_{ij}}{1/m} Y_i^c$, which in turn leads to the individual treatment effect estimate $\widehat{\tau}_j := Y_j^t - \widehat{Y}_j^t(0)$. By the constraint (2.3), the average of the estimated individual treatment effects coincides with the difference in the mean outcomes between the treatment and control groups, denoted as $\widehat{\tau}^{\mathrm{DiM}}$:

$$\frac{1}{m}\sum_{j=1}^m \widehat{\tau}_j = \frac{1}{m}\sum_{j=1}^m (Y_j^t - \textstyle\sum_{i=1}^n \frac{\widehat{\pi}_{ij}}{1/m} Y_i^c) = \frac{1}{m}\sum_{j=1}^m Y_j^t - \frac{1}{n}\sum_{i=1}^n Y_i^c =: \widehat{\tau}^{\mathrm{DiM}}. \tag{2.4}$$

Therefore, adding the extra constraint (2.3) to the minimization of (2.1) under (2.2) allows us to control the ATT to be $\widehat{\tau}^{\mathrm{DiM}}$, which allows practitioners to obtain granular individual-level treatment effect estimates while maintaining the desirable aggregate-level estimate.

Why Couple?. In the randomized experiment, $\widehat{\tau}^{\mathrm{DiM}}$ is clearly the most favorable aggregate-level estimator. If the experimenter's goal is solely to estimate the ATT, a single number $\widehat{\tau}^{\mathrm{DiM}}$ suffices. In many cases, however, individual-level estimates can be much more informative: the experimenter may analyze the treatment effect heterogeneity, while the participants of the experiment can acquire relevant treatment effect information personalized to them. To produce individual-level estimates, the experimenter may implement a method like the nearest neighbor matching or synthetic control for each treated unit separately, namely, solving (2.1) under (2.2) without (2.3). Then, however, the produced individual-level estimates will not be aggregated to the desired aggregate-level estimate $\widehat{\tau}^{\mathrm{DiM}}$, thereby leaving a gap between the individual-level analysis and the aggregate-level analysis. The coupling constraint (2.3) is a neat way to bridge this gap by ensuring that individual-level synthetic control estimates are consistent with the desired aggregate-level estimate $\widehat{\tau}^{\mathrm{DiM}}$. See Remark 2.1 for a further discussion on the coupling constraint.

Optimization. The resulting optimization is a smooth convex optimization problem. To see this, notice that the first term of the objective function (2.1) is the following quadratic function of π:

$$\frac{m}{2}\langle \pi, K_{cc}\pi\rangle - \langle \pi, K_{ct}\rangle + \frac{\mathrm{tr}(K_{tt})}{2m}, \tag{2.5}$$

where $K_{cc} = (\langle x_i^c, x_{i'}^c\rangle) \in \mathbb{R}^{n\times n}$, $K_{ct} = (\langle x_i^c, x_j^t\rangle) \in \mathbb{R}^{n\times m}$, $K_{tt} = (\langle x_j^t, x_{j'}^t\rangle) \in \mathbb{R}^{m\times m}$ are the Gram matrices whose entries are inner products of the covariates. This quadratic function is convex as the Gram matrix K_{cc} is positive semidefinite. The second term of (2.1), the entropy term, is convex. The constraint set resulting from (2.2) and (2.3) is a convex polytope consisting of couplings. In Section 3, we introduce and analyze efficient algorithms to solve this optimization problem based on a simple iterative matrix scaling procedure called the Sinkhorn algorithm [62]. These results easily translate to the standard uncoupled synthetic control problem to minimize (2.1) under (2.2) without (2.3) as well, which we explain in Section 3.4.

Lastly, we notice that the first term of (2.1), under the constraint (2.2), is upper bounded

as follows:

$$\frac{1}{2m}\sum_{j=1}^{m}\|x_j^t-\textstyle\sum_{i=1}^{n}\frac{\pi_{ij}}{1/m}x_i^c\|_2^2=\frac{1}{2m}\sum_{j=1}^{m}\|\textstyle\sum_{i=1}^{n}\frac{\pi_{ij}}{1/m}(x_j^t-x_i^c)\|_2^2\le\frac{1}{2m}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{\pi_{ij}}{1/m}\|x_j^t-x_i^c\|_2^2,$$

where the inequality uses Jensen's inequality. This upper bound analytically shows that the average approximation error based on convex combinations is smaller than that based on pairwise distances, which aligns with the motivation mentioned in Section 1. Replacing the first term of the objective function (2.1) by the above upper bound—while maintaining the constraints (2.2), (2.3)—leads to the optimal transport problem [65] with entropic regularization [26].

2.2. Extensions. We present two extensions of the formulation introduced in Section 2.1. The first is extending the main insights to cover the nonlinear case where the potential outcomes can be nonlinear functions of covariates, using a kernel trick [61, 63]. The second is incorporating notions of propensity scores [58] into the coupling constraint (2.3) to allow for different ATT estimates.

Nonlinear Extension via Kernelization. We can kernelize the formulation introduced in Section 2.1 to extend to nonlinear cases. The kernel trick is widely used in machine learning to modify methods that originated in the linear setting to accommodate nonlinearity, such as ridge regression, support vector machines, and principal component analysis. Recall that (2.1) relies on the Gram matrices as shown in (2.5). Therefore, to kernelize, we can replace the Gram matrices with the kernel matrices, namely, letting $K_{cc}=(k(x_i^c,x_{i'}^c))\in\mathbb{R}^{n\times n}$, $K_{ct}=(k(x_i^c,x_j^t))\in\mathbb{R}^{n\times m}$, and $K_{tt}=(k(x_j^t,x_{j'}^t))\times\mathbb{R}^{m\times m}$ for some kernel function $k\colon\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$. The kernelized formulation is essentially the same convex optimization problem. It can be tackled by the algorithms we introduce in Section 3.

Applying the above kernel trick is equivalent to replacing the first term of (2.1) with the approximation error in the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with the kernel k, namely, $\frac{1}{2m}\sum_{j=1}^{m}\|\phi_{x_j^t}-\sum_{i=1}^{n}\frac{\pi_{ij}}{1/m}\phi_{x_i^c}\|_{\mathcal{H}}^2$, where $\|\cdot\|_{\mathcal{H}}$ is the norm in $\mathcal{H}$ and $\phi_x=k(x,\cdot)\in\mathcal{H}$ denotes the canonical feature of $x\in\mathbb{R}^d$. In other words, we generalize the matching of covariates in the Euclidean space to the matching of embedded canonical features in $\mathcal{H}$, which leads to the following kernelized formulation:

$$\begin{aligned}\min_{\pi\in\mathbb{R}_+^{n\times m}}\quad&\frac{1}{2}\sum_{j=1}^{m}\|\phi_{x_j^t}-\sum_{i=1}^{n}\frac{\pi_{ij}}{1/m}\phi_{x_i^c}\|_{\mathcal{H}}^2+\lambda\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij}\log\frac{\pi_{ij}}{e},\\ \text{subject to}\quad&\sum_{i=1}^{n}\pi_{ij}=\frac{1}{m}\quad\forall j=1,\dots,m\quad\text{and}\quad\sum_{j=1}^{m}\pi_{ij}=w_i\quad\forall i=1,\dots,n,\end{aligned}\tag{2.6}$$

where $w=(w_1,\dots,w_n)\in\Delta_n^+$ is a strictly positive probability vector that generalizes the coupling constraint (2.3) to allow control units to have different weights. With a solution $\widehat{\pi}$, we estimate the counterfactual outcome $Y_j^t(0)$ of treated unit j by the convex combination $\widehat{Y}_j^t(0):=\sum_{i=1}^{n}\frac{\widehat{\pi}_{ij}}{1/m}Y_i^c$ as before, which gives the individual treatment effect estimate $\widehat{\tau}_j:=Y_j^t-\widehat{Y}_j^t(0)$.

General Weights. In (2.6), one can choose $w = (w_1, \dots, w_n) \in \Delta_n^+$ to allow for various average treatment effect estimators. There is a rich literature [69] on weighting methods for estimating average treatment effects, focusing on constructing w that balances the covariates between the treatment and control groups, such as overlap weights [25], stable weights [68], and distributinal balancing weights [42, 40], to name a few. These methods can be incorporated into the synthetic coupling framework by choosing w accordingly, resulting in the following aggregate-level estimate of the ATT:

$$\frac{1}{m}\sum_{j=1}^{m}\widehat{\tau}_j = \frac{1}{m}\sum_{j=1}^{m} Y_j^t - \sum_{i=1}^{n} w_i Y_i^c, \tag{2.7}$$

where the equality follows from the constraint. In the most ideal case, the propensity score is sufficient for achieving perfect balance in the covariates, and thus the optimal choice of w is the normalized propensity score. Concretely, let $\widehat{p}\colon \mathbb{R}^d \to (0,1)$ be the propensity score or its estimate. If the weights w are chosen based on the estimated propensity scores as

$$w_i = \frac{\widehat{p}(x_i^c)}{1-\widehat{p}(x_i^c)} \Big/ \sum_{i'=1}^{n} \frac{\widehat{p}(x_{i'}^c)}{1-\widehat{p}(x_{i'}^c)} \quad \forall i = 1, \dots, n, \tag{2.8}$$

the aggregate effects match the normalized Inverse Probability Weighting (IPW) estimator [36]:

$$\frac{1}{m}\sum_{j=1}^{m}\widehat{\tau}_j = \frac{\sum_{j=1}^{m} Y_j^t}{m} - \frac{\sum_{i=1}^{n} \frac{\widehat{p}(x_i^c)}{1-\widehat{p}(x_i^c)} Y_i^c}{\sum_{i'=1}^{n} \frac{\widehat{p}(x_{i'}^c)}{1-\widehat{p}(x_{i'}^c)}} =: \widehat{\tau}_{\mathrm{ATT}}^{\mathrm{IPW}}. \tag{2.9}$$

Hence, this version of synthetic coupling is useful when practitioners want to produce synthetic control estimates while controlling the ATT to be $\widehat{\tau}_{\mathrm{ATT}}^{\mathrm{IPW}}$, which can be useful in the observational study setting.

Remark 2.1 (Couple or Not?). As mentioned in Section 2.1, while it is informative to produce individual-level estimates, the results from synthetic control for each treated unit separately will not be aggregated to the desired aggregate-level estimate. The constraint (2.3) or $\sum_{j=1}^{m} \pi_{ij} = w_i$ in (2.6) is a device to bind the individual-level analysis and the aggregate-level analysis. Once this coupling constraint is imposed, the optimization problem is no longer separable in j. Still, we can efficiently solve this problem as we show in Section 3, which can also solve the standard uncoupled case much easily.

3. Optimization: Algorithms and Analysis. This section introduces and analyzes optimization algorithms to solve the entropic regularized synthetic control problem with multiple treated units. We will mainly focus on the coupled synthetic control problem (2.6). The standard case without the coupling constraint will later follow as a special case, which we describe in Section 3.4. Let us write (2.6) in a more general form: for $n, m \in \mathbb{N}$, $a \in \Delta_n^+$, $b \in \Delta_m^+$, $\lambda > 0$, and $g\colon \mathbb{R}^{n\times m} \to \mathbb{R}$, consider

$$\min_{\pi \in \Pi_{a,b}} \left(g(\pi) + \lambda h(\pi)\right), \tag{3.1}$$

where $\Pi_{a,b} = \{\pi \in \mathbb{R}_+^{n\times m} : \pi 1_m = a \text{ and } \pi^\top 1_n = b\}$ and $h\colon \mathbb{R}_+^{n\times m} \to \mathbb{R}$ is the entropy function defined by $h(\pi) = \sum_{i=1}^n \sum_{j=1}^m \pi_{ij} \log \frac{\pi_{ij}}{e}$. The coupled synthetic control problem (2.6) amounts to the case where g is a convex quadratic function. It can be solved using the interior point method in poly(nm) time, possibly slow in practice for coupling matrices with large nm.

In this section, we devise a faster algorithm that requires an almost *dimension-free* number of oracle calls to a subroutine—a simple iterative matrix scaling procedure called the Sinkhorn algorithm [62]. When λ is sufficiently large, the algorithm compares favorably to the interior point method, requiring only a $\log(1/\varepsilon)$ number of Sinkhorn subroutines. Later, we modify and extend the algorithm to cover all $\lambda \in \mathbb{R}_+$ by connecting to the steepest descent method under the Kullback-Leibler divergence. The convergence is admittedly slower for small λ: $\log(nm)/\varepsilon$ number of Sinkhorn subroutines is required, but it still compares favorably to the interior point method.

Additional Notation. Along with the notation defined earlier, we further introduce the following notation. Let h keep its definition as the entropy function with $\nabla h(A)$ denoting the gradient of h, which is $\log(A)$, the entrywise logarithm of A, provided all entries of A are positive. For any $a \in \Delta_n$ and $b \in \Delta_m$, let $\Pi_{a,b}^+ := \{A \in \Pi_{a,b} : A_{ij} > 0 \quad \forall i, j\}$, the set of couplings with strictly positive entries. Let $\Delta_{n,m} = \{A \in \mathbb{R}_+^{n\times m} : \sum_{i=1}^n \sum_{j=1}^m A_{ij} = 1\}$. For vectors $u, v \in \mathbb{R}^n$, let $\frac{u}{v}$ denote the entrywise division provided all the entries of v are nonzero.

3.1. Optimality Conditions and the Sinkhorn Algorithm. When g is convex and $\lambda > 0$, the strong convexity of h on $\Pi_{a,b}$ implies that (3.1) admits a unique minimizer. Moreover, as h prevents the minimizer from having a zero entry, (3.1) admits a unique minimizer on $\Pi_{a,b}^+$, which is the interior of $\Pi_{a,b}$. It turns out that this unique minimizer of (3.1) is the fixed point of an operator related to the entropic regularized optimal transport problem [26], as we shall show in Proposition 3.2. Later, in Theorem 3.3, we derive convergence to the minimizer via the iterative matrix scaling subroutine. This subroutine is referred to as the Sinkhorn algorithm [62], which we introduce below.

Definition 3.1 ([62, 26]). *Fix $\lambda > 0$. The following operator $\Phi_\lambda\colon \mathbb{R}^{n\times m} \to \Pi_{a,b}^+$ is well-defined:*

$$\Phi_\lambda(C) := \operatorname*{arg\,min}_{\pi\in\Pi_{a,b}} \left(\langle C, \pi\rangle + \lambda h(\pi)\right) \quad \forall C \in \mathbb{R}^{n\times m},$$

where we often call the right-hand side the entropic regularized optimal transport problem, given a cost matrix C. Moreover, for any $C \in \mathbb{R}^{n\times m}$, one can obtain $\Phi_\lambda(C)$ by iteratively scaling the rows and columns of the matrix $\exp(-C/\lambda)$ using the Sinkhorn algorithm summarized in Algorithm 3.1, namely, $\Phi_\lambda(C)$ is the limit of the sequence produced from $\mathrm{Sinkhorn}(e^{-C/\lambda}, a, b)$.

For simplicity, Algorithm 3.1 states the Sinkhorn algorithm such that the column sum of $P^{(k)}$ matches with b, namely, $(P^{(k)})^\top 1_n = b$ for any $k \geq 0$. Then, to assess the accuracy of the scaled matrix $P^{(k)}$, we may only need to check a suitable discrepancy between a and the row sum of $P^{(k)}$, that is, $P^{(k)} 1_m$.

The following proposition shows that when g is convex, the unique minimizer of (3.1) is the fixed point of an operator given by the composition of Φ_λ and ∇g.

Algorithm 3.1 Sinkhorn(P, a, b)

Require: A matrix $P \in \mathbb{R}^{n\times m}$ with positive entries, $a \in \Delta_n^+$, and $b \in \Delta_m^+$.

Initialize $k \leftarrow 0$, $x^{(k)} \leftarrow 1_n$, and $y^{(k)} \leftarrow \frac{b}{P^\top x^{(k)}}$.

repeat

$$x^{(k+1)} \leftarrow \frac{a}{Py^{(k)}} \quad \text{and} \quad y^{(k+1)} \leftarrow \frac{b}{P^\top x^{(k+1)}}.$$

$P^{(k+1)} = \mathrm{diag}(x^{(k+1)})P\mathrm{diag}(y^{(k+1)})$.

$k \leftarrow k + 1$.

until Discrepancy between $P^{(k)}1_m$ and a reaches a desired level.

return $P^{(k)}$.

Proposition 3.2. *Fix $\lambda > 0$. For a function $g\colon \mathbb{R}^{n\times m} \to \mathbb{R}$ such that ∇g exists, define an operator $T_\lambda\colon \mathbb{R}^{n\times m} \to \Pi_{a,b}^+$ by $T_\lambda = \Phi_\lambda \circ \nabla g$, that is, for any $A \in \mathbb{R}^{n\times m}$,*

$$T_\lambda(A) = \operatorname*{arg\,min}_{\pi\in\Pi_{a,b}}(\langle \nabla g(A), \pi\rangle + \lambda h(\pi)). \tag{3.2}$$

If g is convex, (3.1) admits a unique minimizer contained in $\Pi_{a,b}^+$, say, $\pi^\star \in \Pi_{a,b}^+$, and $\pi^\star$ is the unique fixed point of T_λ, namely, $\pi^\star = T_\lambda(\pi^\star)$.

Based on the established connection between the optimality condition of (3.1) and the Sinkhorn algorithm, we propose two algorithms to solve (3.1) in the following sections.

3.2. Fixed-Point Algorithm. By Proposition 3.2, finding the fixed point of T_λ is equivalent to solving (3.1). Therefore, we propose solving (3.1) by the fixed-point iterations of the operator $T_\lambda\colon \mathbb{R}^{n\times m} \to \Pi_{a,b}^+$, which is simply choosing an initial point $\pi^{(0)} \in \Pi_{a,b}$ and iterating $\pi^{(k+1)} \leftarrow T_\lambda(\pi^{(k)})$ for $k \geq 0$, where $T_\lambda(\pi^{(k)}) = \Phi_\lambda(\nabla g(\pi^{(k)}))$ is approximated by Sinkhorn$(e^{-\nabla g(\pi^{(k)})/\lambda}, a, b)$ as explained in Definition 3.1. Algorithm 3.2 summarizes this procedure.

Algorithm 3.2 FixedPoint(g, λ, a, b)

Require: A map $g\colon \mathbb{R}^{n\times m} \to \mathbb{R}$, a parameter $\lambda > 0$, $a \in \Delta_n^+$, and $b \in \Delta_m^+$.

Pick any $\pi^{(0)} \in \Pi_{a,b}$ and set $k \leftarrow 0$.

repeat

$\pi^{(k+1)} \leftarrow$ Sinkhorn$(e^{-\nabla g(\pi^{(k)})/\lambda}, a, b)$.

$k \leftarrow k + 1$.

until Discrepancy between $\pi^{(k)}$ and $\pi^{(k-1)}$ reaches a desired level.

return $\pi^{(k)}$.

We show that T_λ is a contraction if the gradient of g is Lipschitz and λ is larger than the Lipschitz constant. In such a case, Algorithm 3.2 converges to the fixed point quickly. When g is a quadratic function, we can write the Lipschitz constant in terms of the L^∞ norm of the

Hessian matrix, which is independent of the input dimension nm. We do not require g to be convex for this result; T_λ admits a unique fixed point due to the contraction argument, which is independent of Proposition 3.2. The role of Proposition 3.2 is to translate this convergence result into the convergence to the minimizer of (3.1) for convex g by equating the fixed point to the minimizer.

Theorem 3.3. *Let $g\colon \mathbb{R}^{n\times m} \to \mathbb{R}$ be a quadratic function given by $g(\pi) = \frac{1}{2}\langle \pi, H\pi\rangle + \langle C, \pi\rangle$ for some $H \in \mathbb{R}^{n\times n}$ that is symmetric and $C \in \mathbb{R}^{n\times m}$. Assume $\lambda > \|H\|_\infty$. Then, the operator T_λ defined in (3.2) is a contraction under the distance defined by $\|\cdot\|_1$ and has a unique fixed-point, say, $\pi^\star \in \Pi^+_{a,b}$. Moreover, Algorithm 3.2, assuming the inner loop* Sinkhorn *is always exact, outputs a sequence $(\pi^{(k)})_{k\geq 0}$ such that for any $T \in \mathbb{N}$,*

$$\|\pi^{(T)} - \pi^\star\|_1 \leq \frac{(\|H\|_\infty/\lambda)^T}{1-(\|H\|_\infty/\lambda)}\|\pi^{(1)} - \pi^{(0)}\|_1. \tag{3.3}$$

Remark 3.4. Theorem 3.3 shows that to obtain an ε-approximate fixed point $\|\pi^{(T)} - \pi^\star\|_1^2 \leq \varepsilon$ when $\lambda > \|H\|_\infty$, we need $\log(1/\varepsilon)/2\log(\lambda/\|H\|_\infty)$ number of Sinkhorn routines, which is a dimension-free quantity.

3.3. Local Algorithm with Kullback-Leibler Geometry. The fixed-point algorithm may not converge for sufficiently small λ. To fill in the gap when $\lambda \in [0, \|H\|_\infty)$, which is left unanswered by Theorem 3.3, we employ the steepest descent method under Bregman divergence to solve (3.1) as follows: let $f = g + \lambda h$,

$$\pi^{(k+1)} = \operatorname*{arg\,min}_{\pi\in\Pi_{a,b}} \left(\langle \nabla f(\pi^{(k)}), \pi\rangle + \frac{D_h(\pi, \pi^{(k)})}{\tau_k} \right), \tag{3.4}$$

where $D_h(\pi, \pi^{(k)}) = h(\pi) - h(\pi^{(k)}) - \langle \nabla h(\pi^{(k)}), \pi - \pi^{(k)}\rangle$ is the Bregman divergence, which equals the Kullback-Leibler divergence on $\Pi_{a,b}$, and $\tau_k > 0$ is a suitable step size. With the entropic regularization h, we implement the steepest descent over the polytope $\Pi_{a,b}$ under the Kullback-Leibler divergence. We emphasize that this can be solved using the Sinkhorn algorithm as well since the right-hand side of (3.4) is equivalent to minimizing $\langle \nabla g(\pi^{(k)}) + (\lambda - \tau_k^{-1})\nabla h(\pi^{(k)}), \pi\rangle + \tau_k^{-1} h(\pi)$, which leads to

$$\pi^{(k+1)} = \Phi_{\tau_k^{-1}}\left(\nabla g(\pi^{(k)}) + (\lambda - \tau_k^{-1})\nabla h(\pi^{(k)}) \right).$$

This can be approximated by applying Algorithm 3.1 to $e^{-(\nabla g(\pi^{(k)}) + (\lambda-\tau_k^{-1})\nabla h(\pi^{(k)}))/\tau_k^{-1}}$. Algorithm 3.3 summarizes this procedure. Note that letting $\tau_k = \lambda^{-1}$ for all $k \geq 0$ in Algorithm 3.3 recovers Algorithm 3.2.

We show that Algorithm 3.3 outputs a sequence that converges to the minimum of (3.1) provided the step size is below a certain threshold.

Theorem 3.5. *Let $g\colon \mathbb{R}^{n\times m} \to \mathbb{R}$ be a convex quadratic function given by $g(\pi) = \frac{1}{2}\langle \pi, H\pi\rangle + \langle C, \pi\rangle$ for some $H \in \mathbb{R}^{n\times n}$ that is symmetric and positive semidefinite and $C \in \mathbb{R}^{n\times m}$. Then, for any $\lambda \geq 0$, if we run Algorithm 3.3, assuming the inner loop* Sinkhorn *is always exact,*

with constant step size $\tau_k = \tau$ *for all* $k \geq 0$*, where* $\tau^{-1} \geq \|H\|_\infty + \lambda$*, it outputs a sequence* $(\pi^{(k)})_{k\geq 0}$ *such that for any* $T \in \mathbb{N}$*,*

$$g(\pi^{(T)}) + \lambda h(\pi^{(T)}) - \min_{\pi \in \Pi_{a,b}} (g(\pi) + \lambda h(\pi)) \leq \frac{1}{T} \frac{\log(nm)}{\tau}. \tag{3.5}$$

When $\lambda > 0$*, we have* $\pi^\star \in \Pi^+_{a,b}$*, the unique minimizer of* (3.1)*, and the following holds:*

$$\|\pi^{(T)} - \pi^\star\|_1^2 \leq \frac{1}{T} \frac{2\log(nm)}{\lambda\tau}. \tag{3.6}$$

Remark 3.6. The $\lambda \in [\|H\|_\infty, \infty)$ case has been studied in Theorem 3.3. Theorem 3.5 shows that for any $\lambda \in [0, \|H\|_\infty)$, with the choice $\tau^{-1} = 2\|H\|_\infty$, we need $\frac{1}{\varepsilon} \cdot 2\|H\|_\infty \log(nm)$ number of Sinkhorn subroutines to get ε-close to the objective value. Further more, when λ is strictly positive, we can assure that after $\frac{1}{\varepsilon} \cdot 4\|H\|_\infty \log(nm) \cdot \frac{1}{\lambda}$ number of Sinkhorn subroutines, $\|\pi^{(T)} - \pi^\star\|_1^2 \leq \varepsilon$, with a logarithmic dependence on the problem dimension.

Algorithm 3.3 SteepestDescentKL$(g, \lambda, \{\tau_k\}_{k\geq 0}, a, b)$

Require: A map $g\colon \mathbb{R}^{n\times m} \to \mathbb{R}$, a parameter $\lambda > 0$, $a \in \Delta^+_n$, and $b \in \Delta^+_m$.
Require: Step sizes $\{\tau_k\}_{k\geq 0}$.
1: Pick any $\pi^{(0)} \in \Pi_{a,b}$ and set $k \leftarrow 0$.
2: **repeat**
3: $\quad \pi^{(k+1)} \leftarrow \text{Sinkhorn}(e^{-(\nabla g(\pi^{(k)}) + (\lambda - \tau_k^{-1})\nabla h(\pi^{(k)}))/\tau_k^{-1}}, a, b)$.
4: $\quad k \leftarrow k + 1$.
5: **until** Discrepancy between $\pi^{(k)}$ and $\pi^{(k-1)}$ reaches a desired level.
6: **return** $\pi^{(k)}$.

3.4. Analysis of the Standard Uncoupled Synthetic Control. So far, we have mainly focused on the coupled synthetic control problem. The standard case without the coupling constraint amounts to lifting the row sum constraint, which reads

$$\min_{\pi \in \Pi_{,b}} (g(\pi) + \lambda h(\pi)), \tag{3.7}$$

where $\Pi_{,b} = \{\pi \in \mathbb{R}_+^{n\times m} : \pi^\top 1_n = b\}$. Though this problem is now separable with respect to the columns of π, we can solve a single optimization problem over the coupling matrix π instead of m independent problems. The key is that all the results we have established for the coupled case translate to this uncoupled case by replacing $\Pi_{a,b}$ with $\Pi_{,b}$ and thus running column normalization in place of the Sinkhorn algorithm. Accordingly, we modify Algorithms 3.2 and 3.3 to solve (3.7), which are summarized as Algorithms 3.4 and 3.5, respectively. The only difference is that we replace the Sinkhorn subroutine with the exact column normalization. The optimality result—Proposition 3.2—and the convergence results—Theorems 3.3 and 3.5—carry over to the uncoupled case with suitable modifications as we summarize in the following corollaries.

Corollary 3.7. *Fix $\lambda > 0$. For a function $g\colon \mathbb{R}^{n\times m} \to \mathbb{R}$ such that ∇g exists, define an operator $T_\lambda\colon \mathbb{R}^{n\times m} \to \Pi_{,b}^{+} = \{A \in \Pi_{,b} : A_{ij} > 0 \ \ \forall i, j\}$ by*

$$T_\lambda(A) = e^{-\nabla g(A)/\lambda}\mathrm{diag}\left(\frac{b}{1_n^\top e^{-\nabla g(A)/\lambda}}\right). \tag{3.8}$$

which normalizes the columns of $e^{-\nabla g(A)/\lambda}$ so that the column sum becomes b. If g is convex, (3.7) admits a unique minimizer contained in $\Pi_{,b}^{+}$, say, $\pi^\star \in \Pi_{,b}^{+}$, and $\pi^\star$ is the unique fixed point of T_λ, namely, $\pi^\star = T_\lambda(\pi^\star)$.

Corollary 3.8. *Let $g\colon \mathbb{R}^{n\times m} \to \mathbb{R}$ be a quadratic function given by $g(\pi) = \frac{1}{2}\langle \pi, H\pi\rangle + \langle C, \pi\rangle$ for some $H \in \mathbb{R}^{n\times n}$ that is symmetric and $C \in \mathbb{R}^{n\times m}$. Assume $\lambda > \|H\|_\infty$. Then, the operator T_λ defined in (3.8) is a contraction under the distance defined by $\|\cdot\|_1$ and has a unique fixed-point, say, $\pi^\star \in \Pi_{,b}^{+}$. Moreover, Algorithm 3.4 outputs a sequence $(\pi^{(k)})_{k\geq 0}$ such that for any $T \in \mathbb{N}$,*

$$\|\pi^{(T)} - \pi^\star\|_1 \leq \frac{(\|H\|_\infty/\lambda)^T}{1 - (\|H\|_\infty/\lambda)}\|\pi^{(1)} - \pi^{(0)}\|_1.$$

Corollary 3.9. *Let $g\colon \mathbb{R}^{n\times m} \to \mathbb{R}$ be a convex quadratic function given by $g(\pi) = \frac{1}{2}\langle \pi, H\pi\rangle + \langle C, \pi\rangle$ for some $H \in \mathbb{R}^{n\times n}$ that is symmetric and positive semidefinite and $C \in \mathbb{R}^{n\times m}$. Then, for any $\lambda > 0$, if we run Algorithm 3.5 with constant step size $\tau_k = \tau$ for all $k \geq 0$, where $\tau^{-1} > \|H\|_\infty + \lambda$, it outputs a sequence $(\pi^{(k)})_{k\geq 0}$ such that for any $T \in \mathbb{N}$,*

$$g(\pi^{(T)}) + \lambda h(\pi^{(T)}) - \min_{\pi\in\Pi_{,b}}(g(\pi) + \lambda h(\pi)) \leq \frac{1}{T}\frac{\log(nm)}{\tau}.$$

When $\lambda > 0$, we have $\pi^\star \in \Pi_{,b}^{+}$, the unique minimizer of (3.7), and the following holds:

$$\|\pi^{(T)} - \pi^\star\|_1^2 \leq \frac{1}{T}\frac{2\log(nm)}{\lambda}.$$

Algorithm 3.4 FixedPoint(g, λ, b)

Require: A map $g\colon \mathbb{R}^{n\times m} \to \mathbb{R}$, a parameter $\lambda > 0$, and $b \in \Delta_m^{+}$.

1: Pick any $\pi^{(0)} \in \Pi_{,b}$ and set $k \leftarrow 0$.
2: **repeat**
3: $\quad \pi^{(k+1)} \leftarrow e^{-\nabla g(\pi^{(k)})/\lambda}\mathrm{diag}\left(\frac{b}{1_n^\top e^{-\nabla g(\pi^{(k)})/\lambda}}\right)$.
4: $\quad k \leftarrow k + 1$.
5: **until** Discrepancy between $\pi^{(k)}$ and $\pi^{(k-1)}$ reaches a desired level.
6: **return** $\pi^{(k)}$.

In summary, we solve the standard uncoupled synthetic control problem by iteratively applying the column normalization subroutine, where each column corresponds to a treated unit. This allows for parallel computation of the synthetic control for multiple units, which is particularly useful in large-scale applications.

Algorithm 3.5 SteepestDescentKL$(g, \lambda, \{\tau_k\}_{k\geq 0}, b)$

Require: A map $g\colon \mathbb{R}^{n\times m} \to \mathbb{R}$, a parameter $\lambda > 0$, and $b \in \Delta_m^+$.
Require: Step sizes $\{\tau_k\}_{k\geq 0}$.
1: Pick any $\pi^{(0)} \in \Pi_{,b}$ and set $k \leftarrow 0$.
2: **repeat**
3: $\quad \pi^{(k+1)} \leftarrow e^{-(\nabla g(\pi^{(k)})+(\lambda-\tau_k^{-1})\nabla h(\pi^{(k)}))/\tau_k^{-1}} \operatorname{diag}\left(\frac{b}{1_n^\top e^{-(\nabla g(\pi^{(k)})+(\lambda-\tau_k^{-1})\nabla h(\pi^{(k)}))/\tau_k^{-1}}}\right)$.
4: $\quad k \leftarrow k+1$.
5: **until** Discrepancy between $\pi^{(k)}$ and $\pi^{(k-1)}$ reaches a desired level.
6: **return** $\pi^{(k)}$.

4. Individual Uncertainty Quantification. This section analyzes the entropic regularized synthetic control through the lens of individual uncertainty quantification, developing deeper insights into the regularization and its role as inference-aware optimization for synthetic control. We discuss how the entropic regularization parameter λ drives the bias and variance tradeoff. We show that optimizing the entropic regularized synthetic control problem (2.6) with the desired choice of λ is closely connected to optimizing the average of the squared lengths of the individual confidence intervals.

Notation. To facilitate the analysis, we switch the notation from the previous sections and consider N units indexed by $1, \ldots, N$. For unit i whose covariate vector is $X_i \in \mathbb{R}^d$, let $Y_i(1), Y_i(0) \in \mathbb{R}$ denote the potential outcome variables under treatment, control, respectively, and let $Z_i \in \{0, 1\}$ denote the treatment assignment. Y_i is the observed outcome of unit i, which we can write as $Y_i = Z_iY_i(1) + (1 - Z_i)Y_i(0)$. Let $\mathcal{T} = \{i \in [N] : Z_i = 1\}$ and $\mathcal{C} = \{i \in [N] : Z_i = 0\}$ denote the sets of indices of the treated units and control units, respectively, so that $\mathcal{T} \cup \mathcal{C} = \{1, \ldots, N\} =: [N]$. Also, let $N_t := |\mathcal{T}|$ and $N_c := |\mathcal{C}|$ denote the numbers of the treated and control units.

4.1. Fixed Design Inference Framework for Imputation. Thanks to the proposed algorithms, we can now efficiently deploy synthetic control to obtain the imputed counterfactual outcomes $\{\widehat{Y}_j(0)\}_{j\in\mathcal{T}}$ even when there are many units. In this section, we construct an interval around each imputed counterfactual outcome $\widehat{Y}_j(0)$ for $j \in \mathcal{T}$, which provides a range of potential counterfactual outcomes for each treated unit.

We conduct inference by conditioning on the design matrix $\mathbf{X} = [X_1, \ldots, X_N] \in \mathbb{R}^{d\times N}$ and the treatment assignment vector $\mathbf{Z} = (Z_1, \ldots, Z_N) \in \mathbb{R}^N$. In other words, we restrict the inference to the given covariates and treatment assignment. The motivation stems from an individual's perspective: for each unit of the study, such as a participant of a randomized experiment, the primary interest lies in knowing the missing counterfactual outcome, rather than how the results generalize to new samples with different covariates. Quantifying the uncertainty around the imputed counterfactual outcomes is a transparent approach that directly addresses this primary interest, providing a range of values that each individual cares about most.

Adopting this perspective, we postulate the following assumptions.

Assumption 4.1. Let $\mathbf{X} = [X_1, \ldots, X_N] \in \mathbb{R}^{d\times N}$ and $\mathbf{Z} = (Z_1, \ldots, Z_N) \in \mathbb{R}^N$ denote

the design matrix and the treatment assignment vector, respectively, and let $\varepsilon_{0,i} := Y_i(0) - \mathbb{E}[Y_i(0) \mid X_i]$ for each $i \in [N]$.

(i) There is a function $f_0 \colon \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}[Y_i(0) \mid X_i = x] = f_0(x)$ for all $i \in [N]$ and $x \in \mathbb{R}^d$.
(ii) $\mathbb{E}[Y_i(0) \mid \mathbf{X}] = \mathbb{E}[Y_i(0) \mid X_i]$ and $\mathrm{Var}(Y_i(0) \mid \mathbf{X}) = \mathrm{Var}(Y_i(0) \mid X_i)$ for all $i \in [N]$.
(iii) $\{\varepsilon_{0,i}\}_{i=1}^N$ are independent random variables given $\mathbf{X}$.
(iv) $\mathbf{Z} \perp\!\!\!\perp \{\varepsilon_{0,i}\}_{i=1}^N \mid \mathbf{X}$.

Remark 4.2. Typical assumptions in causal inference consist of the sampling and unconfoundedness assumptions: (a) $\{Y_i(0), Y_i(1), X_i, Z_i\}_{i=1}^N$ are i.i.d. from some joint population distribution defined on $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d \times \{0,1\}$, and (b) for each $i \in [N]$, the treatment assignment Z_i is independent of $(Y_i(0), Y_i(1))$ conditional on X_i. As we condition on $\mathbf{X}, \mathbf{Z}$ for inference, requiring units to be i.i.d. is not necessary. Hence, we impose assumptions weaker than (a) and (b). Indeed, (i)-(iii) of Assumption 4.1 are implied by (a), while (iv) is implied by (a) and (b). Particularly, as Assumption 4.1 does not require $\mathbf{X}$ or $\mathbf{Z}$ to be i.i.d., Z_i's are allowed to have arbitrary dependence across units or even to be adversarially chosen conditioned on the fixed design $\mathbf{X}$.

Procedure. We analyze the formulation (2.6) given some kernel function $k \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Let $\mathcal{H}$ be the reproducing kernel Hilbert space (RKHS) associated k, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the RKHS inner product, $\| \cdot \|_{\mathcal{H}}$ denotes the RKHS norm, and $\phi_x := k(x, \cdot) \in \mathcal{H}$ is the corresponding reproducing kernel function. Given some suitable positive weights $(w_i)_{i \in \mathcal{C}}$ such that $\sum_{i \in \mathcal{C}} w_i = 1$, we can rewrite (2.6) based on the new notation as follows:

$$\begin{aligned} \min_{\pi \in \mathbb{R}_+^{N_c \times N_t}} \quad & \frac{1}{2} \sum_{j \in \mathcal{T}} \Big\| \phi_{X_j} - \sum_{i \in \mathcal{C}} \tfrac{\pi_{ij}}{1/N_t} \phi_{X_i} \Big\|_{\mathcal{H}}^2 + \lambda \sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{T}} \pi_{ij} \log \tfrac{\pi_{ij}}{e}, \\ \text{subject to} \quad & \sum_{i \in \mathcal{C}} \pi_{ij} = \tfrac{1}{N_t} \quad \forall j \in \mathcal{T} \quad \text{and} \quad \sum_{j \in \mathcal{T}} \pi_{ij} = w_i \quad \forall i \in \mathcal{C}, \end{aligned} \tag{4.1}$$

where $\pi \in \mathbb{R}_+^{N_c \times N_t}$ is a matrix whose entries are indexed by (i, j) for $i \in \mathcal{C}$ and $j \in \mathcal{T}$, instead of the usual indexing by $i \in \{1, \ldots, N_c\}$ and $j \in \{1, \ldots, N_t\}$, to denote that π_{ij} is the weight between control unit i and treated unit j. Equivalently, as noted in (2.5), we may rewrite (4.1) as follows:

$$\begin{aligned} \min_{\pi \in \mathbb{R}_+^{N_c \times N_t}} \quad & \frac{m}{2} \langle \pi, K_{cc} \pi \rangle - \langle \pi, K_{ct} \rangle + \frac{\mathrm{tr}(K_{tt})}{2m} + \lambda \sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{T}} \pi_{ij} \log \tfrac{\pi_{ij}}{e}, \\ \text{subject to} \quad & \sum_{i \in \mathcal{C}} \pi_{ij} = \tfrac{1}{N_t} \quad \forall j \in \mathcal{T} \quad \text{and} \quad \sum_{j \in \mathcal{T}} \pi_{ij} = w_i \quad \forall i \in \mathcal{C}, \end{aligned}$$

where $K_{cc} = (k(X_i, X_{i'}))_{i,i' \in \mathcal{C}}$, $K_{ct} = (k(X_i, X_j))_{(i,j) \in \mathcal{C} \times \mathcal{T}}$, and $K_{tt} = (k(X_j, X_{j'}))_{j,j' \in \mathcal{T}}$ are the kernel matrices.

4.2. Coverage. Under the above setup, for each treated unit $j \in \mathcal{T}$, we quantify the uncertainty around the imputed counterfactual outcome $\widehat{Y}_j(0)$ by constructing a confidence interval containing the conditional expectation $f_0(X_j) = \mathbb{E}[Y_j(0) \mid X_j]$. More specifically, we

aim to construct an interval $[\widehat{L}_j, \widehat{U}_j]$ such that the following holds: given $\alpha \in (0,1)$,

$$\mathbb{P}\left(f_0(X_j) \in [\widehat{L}_j, \widehat{U}_j] \mid \mathbf{X}, \mathbf{Z}\right) \geq 1-\alpha, \tag{4.2}$$

namely, the probability that the interval $[\widehat{L}_j, \widehat{U}_j]$ contains the conditional expectation $f_0(X_j)$ is at least $1-\alpha$, where the probability is conditional on the design matrix $\mathbf{X} = [X_1, \ldots, X_N]$ and the treatment assignment vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$.

Let $\widehat{\pi}$ be the solution of (4.1) and define a probability transition matrix $\widehat{P} = (\widehat{p}_{ij})_{(i,j)\in\mathcal{C}\times\mathcal{T}}$, where $\widehat{p}_{ij} = N_t\widehat{\pi}_{ij}$ so that $\sum_{i\in\mathcal{C}} \widehat{p}_{ij} = 1$ for any $j \in \mathcal{T}$. We remark that the optimal coupling $\widehat{\pi}$—and thus $\widehat{P} = N_t\widehat{\pi}$ —is calculated based solely on $\mathbf{X}$ and $\mathbf{Z}$, not the outcomes Y_i's. Recall that we impute the counterfactual outcome of treated unit j as

$$\widehat{Y}_j(0) = \sum_{i\in\mathcal{C}} \widehat{p}_{ij} Y_i(0). \tag{4.3}$$

Now, recall the following bias-variance decomposition of the point estimate $\widehat{Y}_j(0)$:

$$\widehat{Y}_j(0) = \mathbb{E}[Y_j(0) \mid \mathbf{X}, \mathbf{Z}] + \underbrace{\mathbb{E}[\widehat{Y}_j(0) \mid \mathbf{X}, \mathbf{Z}] - \mathbb{E}[Y_j(0) \mid \mathbf{X}, \mathbf{Z}]}_{:=\mathcal{B}_j} + \underbrace{\widehat{Y}_j(0) - \mathbb{E}[\widehat{Y}_j(0) \mid \mathbf{X}, \mathbf{Z}]}_{:=\mathcal{V}_j}. \tag{4.4}$$

Assuming the correct model specification, namely, $f_0 \in \mathcal{H}$, we analyze the bias and variance terms $\mathcal{B}_j$ and $\mathcal{V}_j$ as follows.

Bias-Awareness. The point estimate could be biased as

$$\mathcal{B}_j = \mathbb{E}[\widehat{Y}_j(0) \mid \mathbf{X}, \mathbf{Z}] - \mathbb{E}[Y_j(0) \mid \mathbf{X}, \mathbf{Z}] = \sum_{i\in\mathcal{C}} \widehat{p}_{ij} f_0(X_i) - f_0(X_j) = \langle f_0, \sum_{i\in\mathcal{C}} \widehat{p}_{ij}\phi_{X_i} - \phi_{X_j}\rangle_{\mathcal{H}},$$

where the second equality uses $\mathbb{E}[Y_j(0) \mid \mathbf{X}, \mathbf{Z}] = f_0(X_j)$, which is true under Assumption 4.1, and the last equality follows from the reproducing property of the RKHS and $f_0 \in \mathcal{H}$. Since f_0 is unknown and estimating such a function uniformly on the support of the covariate vector can be difficult, we rely on a bias-awareness correction by invoking the Cauchy-Schwarz inequality:

$$|\mathcal{B}_j| \leq \|f_0\|_{\mathcal{H}} \cdot \|\phi_{X_j} - \sum_{i\in\mathcal{C}} \widehat{p}_{ij}\phi_{X_i}\|_{\mathcal{H}} = \|f_0\|_{\mathcal{H}} \cdot (K_{tt} + \widehat{P}^\top K_{cc}\widehat{P} - 2K_{ct}^\top \widehat{P})_{jj}^{1/2}. \tag{4.5}$$

If we knew the bias $\mathcal{B}_j$, we could directly correct the point estimate $\widehat{Y}_j(0)$ by subtracting it, yielding a bias-corrected point estimate $\widehat{Y}_j(0) - \mathcal{B}_j$. However, we do not know the bias $\mathcal{B}_j$ in practice. Instead, based on (4.5), we subtract and add the term $\|f_0\|_{\mathcal{H}} \cdot \|\phi_{X_j} - \sum_{i\in\mathcal{C}} \widehat{p}_{ij}\phi_{X_i}\|_{\mathcal{H}}$ to the point estimate $\widehat{Y}_j(0)$, which leads to a bias-aware interval. In practice, the model parameter $\|f_0\|_{\mathcal{H}}$ is unknown and thus should be estimated. As this is a scalar parameter, one can estimate this easily using the kernel ridge regression as we explain at the end of this section.

Variance. The variance term $\mathcal{V}_j$ satisfies

$$\mathcal{V}_j = \widehat{Y}_j(0) - \sum_{i\in\mathcal{C}} \widehat{p}_{ij} f_0(X_i) = \sum_{i\in\mathcal{C}} \widehat{p}_{ij}\varepsilon_{0,i}, \tag{4.6}$$

where the residuals $(\varepsilon_{0,i})_{i\in\mathcal{C}}$ given $\mathbf{X}, \mathbf{Z}$ are independent mean zero random variables with variance $\mathrm{Var}(\varepsilon_{0,i} \,|\, \mathbf{X}, \mathbf{Z}) = \mathrm{Var}(Y_i(0) \,|\, X_i) =: \sigma_{0,i}^2$ under Assumption 4.1. Assuming homoskedasticity, namely, $\sigma_{0,i}^2 = \sigma_0^2$ for all $i \in \mathcal{C}$, we have

$$\mathrm{Var}(\mathcal{V}_j \,|\, \mathbf{X}, \mathbf{Z}) = \sum_{i\in\mathcal{C}} \widehat{p}_{ij}^2 \cdot \mathrm{Var}(\varepsilon_{0,i} \,|\, \mathbf{X}, \mathbf{Z}) = \sigma_0^2 \sum_{i\in\mathcal{C}} \widehat{p}_{ij}^2. \tag{4.7}$$

When we the sample size is large, the convex combination $\mathcal{V}_j = \sum_{i\in\mathcal{C}} \widehat{p}_{ij}\varepsilon_{0,i}$ is asymptotically normal with mean zero and variance (4.7) if the following Lindeberg condition holds: for any $\varepsilon > 0$, we have

$$\lim_{N_c\to\infty} \frac{1}{\sigma_0^2 \sum_{i\in\mathcal{C}} \widehat{p}_{ij}^2} \sum_{i\in\mathcal{C}} \mathbb{E}\, \widehat{p}_{ij}^2 \varepsilon_{0,i}^2 1\left\{ |\widehat{p}_{ij}\varepsilon_{0,i}| > \varepsilon \sqrt{\sigma_0^2 \sum_{i\in\mathcal{C}} \widehat{p}_{ij}^2} \right\} = 0. \tag{4.8}$$

Using the asymptotic normality, we can construct confidence intervals for the imputed counterfactual outcomes $\widehat{Y}_j(0)$ satisfying (4.2) in the large-sample limit.

Proposition 4.3. *Under Assumption 4.1, suppose* $\mathrm{Var}(Y_i(0) \,|\, X_i) = \sigma_0^2$ *for all* $i \in \mathcal{C}$. *Suppose* f_0 *belongs to the RKHS* $\mathcal{H}$ *associated with the kernel* k. *Let* $\widehat{\pi}$ *be the solution of* (4.1) *and let* $\widehat{P} = N_t\widehat{\pi}$. *Then, for each treated unit* $j \in \mathcal{T}$, *we construct the interval around the imputed counterfactual outcome* $\widehat{Y}_j(0)$ *defined in* (4.3) *as follows: for* $\alpha \in (0,1)$,

$$\widehat{L}_j := \widehat{Y}_j(0) - \|f_0\|_{\mathcal{H}} \cdot \|\phi_{X_j} - \sum_{i\in\mathcal{C}} \widehat{p}_{ij}\phi_{X_i}\|_{\mathcal{H}} - z_{1-\frac{\alpha}{2}} \cdot \sigma_0 \sqrt{\sum_{i\in\mathcal{C}} \widehat{p}_{ij}^2}, \tag{4.9}$$

$$\widehat{U}_j := \widehat{Y}_j(0) + \|f_0\|_{\mathcal{H}} \cdot \|\phi_{X_j} - \sum_{i\in\mathcal{C}} \widehat{p}_{ij}\phi_{X_i}\|_{\mathcal{H}} + z_{1-\frac{\alpha}{2}} \cdot \sigma_0 \sqrt{\sum_{i\in\mathcal{C}} \widehat{p}_{ij}^2}, \tag{4.10}$$

where $z_{1-\frac{\alpha}{2}}$ *is the* $(1-\frac{\alpha}{2})$*-th quantile of the standard normal distribution* $N(0,1)$. *If the condition* (4.8) *holds for any* $\varepsilon > 0$, *then the interval* $[\widehat{L}_j, \widehat{U}_j]$ *satisfies the following asymptotic coverage property:*

$$\lim_{N_c\to\infty} \mathbb{P}\left(f_0(X_j) \in [\widehat{L}_j, \widehat{U}_j] \mid \mathbf{X}, \mathbf{Z} \right) \geq 1 - \alpha. \tag{4.11}$$

Remark 4.4. One can deduce that the condition (4.8) is satisfied if the entropic regularization is sufficient enough to ensure that the weights $\widehat{p}_{ij}$ are delocalized in the sense $\max_{i\in\mathcal{C}} \widehat{p}_{ij} \leq O(1/N_c)$ for large N_c. In this case, (4.8) follows. To see this, consider the residuals with bounded third moments, namely, there is a constant $R > 0$ such that $\mathbb{E}\,|\varepsilon_{0,i}|^3 \leq R$ for all $i \in \mathcal{C}$. Then, the summand on the left hand-side of (4.8) is bounded by

$$\mathbb{E}\, \widehat{p}_{ij}^2 \varepsilon_{0,i}^2 1\left\{ |\widehat{p}_{ij}\varepsilon_0| > \varepsilon\sigma_0 \sqrt{\sum_{i\in\mathcal{C}} \widehat{p}_{ij}^2} \right\} \leq \frac{\mathbb{E}\, \widehat{p}_{ij}^3 |\varepsilon_{0,i}|^3}{\varepsilon\sigma_0 \sqrt{\sum_{i\in\mathcal{C}} \widehat{p}_{ij}^2}} \leq \frac{R \cdot \widehat{p}_{ij}^3}{\varepsilon\sigma_0 \sqrt{\sum_{i\in\mathcal{C}} \widehat{p}_{ij}^2}}$$

Hence, the left-hand side of (4.8) is bounded by

$$\frac{R}{\varepsilon\sigma_0^3}\cdot\frac{\sum_{i\in\mathcal{C}}\widehat{p}_{ij}^3}{(\sum_{i\in\mathcal{C}}\widehat{p}_{ij}^2)^{3/2}}\le\frac{R}{\varepsilon\sigma_0^3}\cdot\frac{\max_{i\in\mathcal{C}}\widehat{p}_{ij}}{\sqrt{\sum_{i\in\mathcal{C}}\widehat{p}_{ij}^2}}\le\frac{R}{\varepsilon\sigma_0^3}\cdot\sqrt{N_c}\cdot\max_{i\in\mathcal{C}}\widehat{p}_{ij}\le O\left(\frac{1}{\sqrt{N_c}}\right),$$

where the second inequality is due to the Cauchy-Schwarz inequality with $\sum_{i\in\mathcal{C}}\widehat{p}_{ij}=1$.

4.3. Efficiency and Inference-Aware Optimization. We now explain how the width of the confidence interval is related to the objective function of (2.6). We have already noticed in Remark 4.4 that the entropic regularization parameter λ essentially controls the delocalization of the weights $\widehat{p}_{ij}$, which in turn affects the variance term $\mathcal{V}_j=\sum_{i\in\mathcal{C}}\widehat{p}_{ij}\varepsilon_{0,i}$ in (4.6). We take a closer look at this idea by analyzing the overall efficiency of the confidence intervals. We show that the entropic regularization parameter λ serves as a tuning parameter to trade off bias and variance for overall interval length efficiency.

Proposition 4.5. *Consider the interval $[\widehat{L}_j,\widehat{U}_j]$ defined in (4.9) and (4.10). For each $j\in\mathcal{T}$, suppose that $(\widehat{p}_{ij})_{i\in\mathcal{C}}\in\Delta_{N_c}$, the j-th column of $\widehat{P}$, is delocalized in the sense $\max_{i\in\mathcal{C}}\widehat{p}_{ij}\le(M/2-1)/N_c$ for some constant $M>4$. Then, the interval length $\mathrm{Len}_j:=\widehat{U}_j-\widehat{L}_j$ satisfies the following bounds:*

$$\mathrm{Len}_j^2\le 8\left\{\|f_0\|_{\mathcal{H}}^2\cdot\|\phi_{X_j}-\sum_{i\in\mathcal{C}}\frac{\widehat{\pi}_{ij}}{1/N_t}\phi_{X_i}\|_{\mathcal{H}}^2+\frac{2Mz_{1-\frac{\alpha}{2}}^2\cdot\sigma_0^2}{N_c/N_t}\left[\sum_{i\in\mathcal{C}}\widehat{\pi}_{ij}\log\frac{\widehat{\pi}_{ij}}{e}+\frac{\frac{1}{M}+1+\log(N_cN_t)}{N_t}\right]\right\},$$

$$\mathrm{Len}_j^2\ge 4\left\{\|f_0\|_{\mathcal{H}}^2\cdot\|\phi_{X_j}-\sum_{i\in\mathcal{C}}\frac{\widehat{\pi}_{ij}}{1/N_t}\phi_{X_i}\|_{\mathcal{H}}^2+\frac{2z_{1-\frac{\alpha}{2}}^2\cdot\sigma_0^2}{N_c/N_t}\left[\sum_{i\in\mathcal{C}}\widehat{\pi}_{ij}\log\frac{\widehat{\pi}_{ij}}{e}+\frac{2+\log(N_cN_t)}{N_t}\right]\right\}.$$

Averaging over $j\in\mathcal{T}$ and barring the constant M in Proposition 4.5, the upper and lower bounds on the right-hand sides remind us of the objective function in (2.6), with the regularization parameter

$$\lambda\asymp\frac{z_{1-\frac{\alpha}{2}}}{N_c}\frac{\sigma_0^2}{\|f_0\|_{\mathcal{H}}^2}.$$

Later in Section 4.4, we confirm the following tradeoff through numerical simulations: for a high signal-to-noise problem, to balance bias and variance, we need a smaller λ for better approximation error; for a low signal-to-noise problem, a larger λ is preferred to reduce variance. See Figure 2.

Proposition 4.5 provides a statistical interpretation of entropic regularized synthetic control as inference-aware optimization. We now see that searching for a coupling minimizing the objective of (2.6) is closely related to minimizing the interval length, provided the regularization parameter λ is appropriately chosen. Therefore, optimizing the entropic regularized problem (2.6) with the desired λ chosen above is closely connected to optimizing the average of the squared lengths of the individual confidence intervals $\frac{1}{N_t}\sum_{j\in\mathcal{T}}\mathrm{Len}_j^2$.

Estimated Intervals. In practice, the model parameters $\|f_0\|_{\mathcal{H}},\sigma_0$ are unknown, which we replace with suitable estimates $\widehat{\theta},\widehat{\sigma}_0$, respectively. By applying kernel ridge regression to the control group data, we estimate $f_0\in\mathcal{H}$ by $\widehat{f}_0(x)=\sum_{i\in\mathcal{C}}\widehat{\beta}_ik(X_i,x)$ with $\widehat{\beta}=(K_{cc}+$

$\rho I_{N_c})^{-1} Y_c \in \mathbb{R}^{N_c}$, where $\rho > 0$ is a ridge regularization parameter and $Y_c = (Y_i)_{i \in \mathcal{C}} \in \mathbb{R}^{N_c}$ is a vector of the control outcomes. Then, we estimate $\|f_0\|_{\mathcal{H}}$ and the variance σ_0^2 by $\widehat{\theta} := \|\widehat{f_0}\|_{\mathcal{H}} = \sqrt{\langle \widehat{\beta}, K_{cc} \widehat{\beta} \rangle}$ and $\widehat{\sigma}_0^2 := \frac{1}{N_c} \sum_{i \in \mathcal{C}} (Y_i - \widehat{f_0}(X_i))^2 = \frac{1}{N_c} \langle Y_c - K_{cc} \widehat{\beta}, Y_c - K_{cc} \widehat{\beta} \rangle$. By plugging in the estimates $\widehat{\theta}$ and $\widehat{\sigma}_0$ into (4.9) and (4.10), we obtain the following confidence intervals for each treated unit $j \in \mathcal{T}$:

$$\widehat{L}_j := \widehat{Y}_j(0) - \widehat{\theta} \sqrt{\left(K_{tt} + \widehat{P}^\top K_{cc} \widehat{P} - 2 K_{ct}^\top \widehat{P} \right)_{jj}} - z_{1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_0 \sqrt{\left(\widehat{P}^\top \widehat{P} \right)_{jj}}, \tag{4.12}$$

$$\widehat{U}_j := \widehat{Y}_j(0) + \widehat{\theta} \sqrt{\left(K_{tt} + \widehat{P}^\top K_{cc} \widehat{P} - 2 K_{ct}^\top \widehat{P} \right)_{jj}} + z_{1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_0 \sqrt{\left(\widehat{P}^\top \widehat{P} \right)_{jj}}. \tag{4.13}$$

4.4. Simulation: Visualization of the Individual Intervals. We visualize the individual intervals constructed in Section 4.2 using a simulation study. We consider fixed one-dimensional covariates $X_1, \ldots, X_N \in [0, 1]$ with $N = 500$ and randomly choose $N_t = 200$ treatment units. Let $\mathcal{H}$ be the RKHS associated with the Gaussian kernel $k(x, x') = e^{-\gamma \cdot |x - x'|^2}$ with $\gamma = 2.5$, and let $f_0(x) = k(0.5, x)$, which yields $\|f_0\|_{\mathcal{H}} = (k(0.5, 0.5))^{1/2} = 1$. Then, we generate the potential outcomes by $Y_i(0) = f_0(X_i) + \varepsilon_{0,i}$, where $\varepsilon_{0,1}, \ldots, \varepsilon_{0,N}$ are independently drawn from $N(0, \sigma_0^2)$. As $X_1, \ldots, X_N$ and $Z_1, \ldots, Z_N$ are fixed, the randomness is solely from the residuals $\varepsilon_{0,i}$'s, which we sample 1,000 times.

Figure 2 shows the constructed confidence intervals $[\widehat{L}_j, \widehat{U}_j]$—based on (4.13), (4.12)—for each treated unit j for different values of σ_0, with $\alpha = 0.05$. For each $\sigma_0 \in \{0.1, 1, 3\}$, we solve (4.1) for $\lambda \in \{0.1, 0.01, 0.001\}$ with uniform weights w_i's, run the kernel ridge regression to obtain $\widehat{\theta}, \widehat{\sigma}_0$, and produce $[\widehat{L}_j, \widehat{U}_j]$ which is shown as blue shaded regions in the plots. We compare this interval with the "oracle" interval $[L_j^*, U_j^*]$ defined as follows:

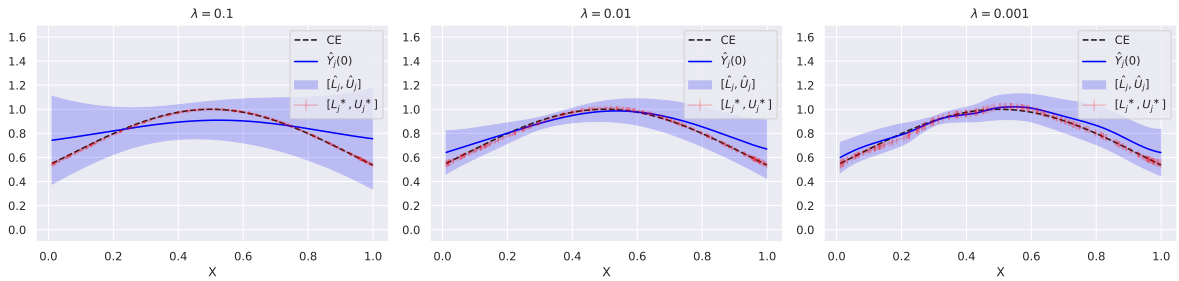
$$L_j^* := \widehat{Y}_j(0) - \mathcal{B}_j - z_{1-\frac{\alpha}{2}} \cdot \sigma_0 \sqrt{\sum_{i \in \mathcal{C}} \widehat{p}_{ij}^2}, \tag{4.14}$$

$$U_j^* := \widehat{Y}_j(0) - \mathcal{B}_j + z_{1-\frac{\alpha}{2}} \cdot \sigma_0 \sqrt{\sum_{i \in \mathcal{C}} \widehat{p}_{ij}^2}, \tag{4.15}$$

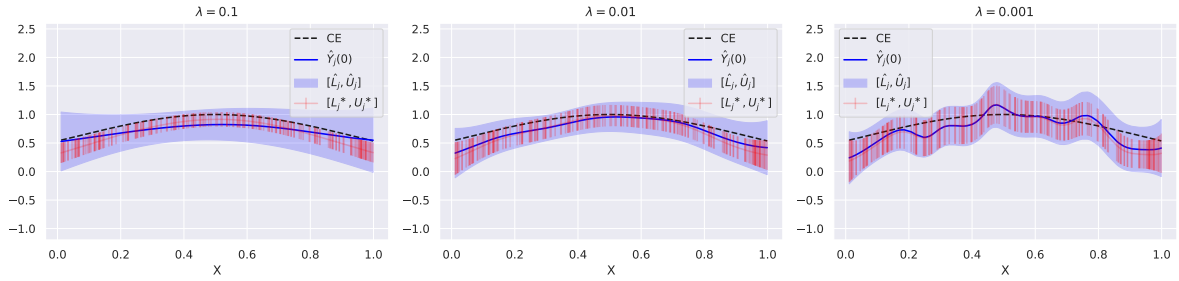
The oracle interval $[L_j^*, U_j^*]$ is based on the unknown quantities $\mathcal{B}_j, \sigma_0$ (hence named "oracle") and thus is guaranteed to have the exact $1 - \alpha$ coverage under this Gaussian setting, namely, we have

$$\mathbb{P}\left(f_0(X_j) \in [L_j^*, U_j^*] \mid \mathbf{X}, \mathbf{Z} \right) = 1 - \alpha. \tag{4.16}$$

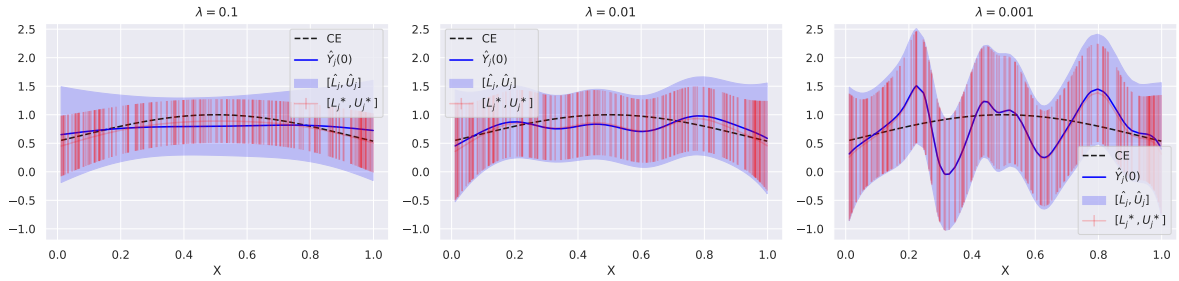
By construction, if the estimates $\widehat{\theta}, \widehat{\sigma}_0$ are accurate enough, namely, $\widehat{\theta} \approx \|f_0\|_{\mathcal{H}}$ and $\widehat{\sigma}_0 \approx \sigma_0$, the interval $[\widehat{L}_j, \widehat{U}_j]$ must contain the oracle interval $[L_j^*, U_j^*]$, which can be verified in Figure 2. For each σ_0, decreasing λ results in smaller approximation errors, namely, the bias correction reflected in $\left(K_{tt} + \widehat{P}^\top K_{cc} \widehat{P} - 2 K_{ct}^\top \widehat{P} \right)_{jj}^{1/2}$, and the estimated counterfactual outcomes (solid blue curves) that are more wiggly, as the obtained weights $\widehat{P}$ have more



(a) $\sigma_0 = 0.1$



(b) $\sigma_0 = 1$



(c) $\sigma_0 = 3$

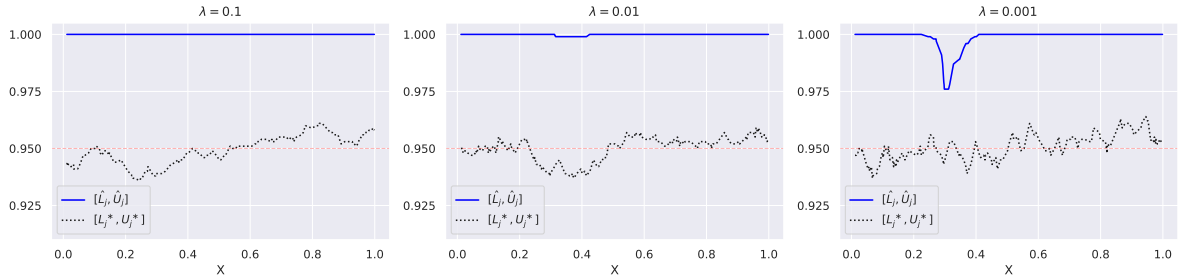
Figure 2. *Confidence intervals of potential outcomes generated by* $Y_i(0) = f_0(X_i) + \varepsilon_{0,i}$, *where* $\varepsilon_{0,1}, \ldots, \varepsilon_{0,N}$ *are i.i.d. from* $N(0, \sigma_0^2)$, *with* $\alpha = 0.05$. *The black dashed curve is the ground truth conditional expectation function* $f_0(x) = e^{-2.5 \times |x - 0.5|^2}$. *The blue solid curve is the estimated counterfactual outcome* $\widehat{Y}_j(0) = \sum_{i \in \mathcal{C}} \widehat{p}_{ij} Y_i$ *for each treated unit* j. *The blue shaded region shows* $[\widehat{L}_j, \widehat{U}_j]$ *defined in* (4.12), (4.13), *where* $\widehat{\theta}, \widehat{\sigma}_0$ *are estimated by the kernel ridge regression explained earlier. The oracle interval* $[L_j^*, U_j^*]$ *defined in* (4.14), (4.15) *is shown as red error bars.*

variations. Assuming $\widehat{\sigma}_0 \approx \sigma_0$, we can see that $[\widehat{L}_j, \widehat{U}_j]$ is wider than $[L_j^*, U_j^*]$ by the twice of the bias correction, namely,

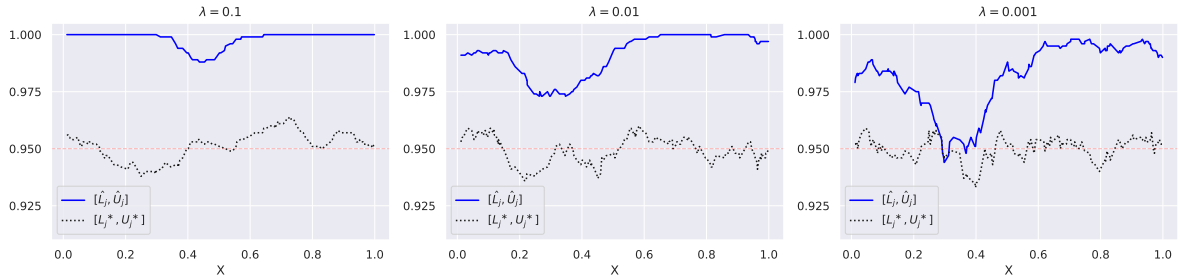
$$(\widehat{U}_j - \widehat{L}_j) - (U_j^* - L_j^*) \approx 2 \cdot \widehat{\theta} \cdot \left(K_{tt} + \widehat{P}^\top K_{cc} \widehat{P} - 2 K_{ct}^\top \widehat{P} \right)_{jj}^{1/2}.$$

We can verify from Figure 2 that this gap between the two intervals gets smaller as λ decreases, consistent with the above observation that the bias decreases.

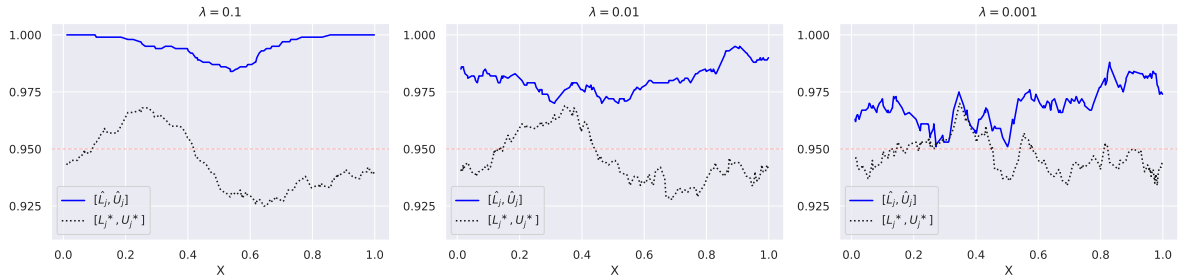
Figure 3 plots the coverage of $[\widehat{L}_j, \widehat{U}_j]$ and $[L_j^*, U_j^*]$, namely, the conditional probability that they contain $f_0(X_j)$, estimated using the 1,000 samples. When $\lambda = 0.1$, we can see that



(a) $\sigma_0 = 0.1$



(b) $\sigma_0 = 1$



(c) $\sigma_0 = 3$

Figure 3. *Coverage of the confidence intervals* $[\widehat{L}_j, \widehat{U}_j]$ *and* $[L_j^*, U_j^*]$ *estimated using the 1,000 samples. The red dashed line shows the nominal coverage* $1 - \alpha = 0.95$.

the coverage of $[\widehat{L}_j, \widehat{U}_j]$ is almost 1, meaning that the interval is too conservative, which results from the fact that the bias term is not small enough compared to the variance of the residuals. As λ decreases, we can see that the coverage of $[\widehat{L}_j, \widehat{U}_j]$ gets closer to $1 - \alpha = 0.95$, which is consistent with the fact that the bias decreases. Particularly, when $\sigma_0 = 3$ and $\lambda = 0.001$, the coverage of $[\widehat{L}_j, \widehat{U}_j]$ is closest to the coverage of the oracle interval $[L_j^*, U_j^*]$, suggesting that the bias is dominated by the variance. Meanwhile, for $(\sigma_0, \lambda) = (1, 0.01)$ and $(\sigma_0, \lambda) = (3, 0.1)$, the coverage of $[\widehat{L}_j, \widehat{U}_j]$ is larger than the coverage of the oracle interval $[L_j^*, U_j^*]$ by approximately a constant, which suggests that the bias term and the variance term are nearly balanced. These observations are consistent with the bias-variance tradeoff discussed in Section 4.3.

5. Conclusion. We proposed and analyzed algorithms that efficiently solve the synthetic control problem with nearly dimension-free convergence rates. By the coupling constraint to control the aggregate-level estimate, we provide a neat way to close the gap between the

individual-level analysis and the aggregate-level analysis. After placing the coupling constraint, we solve the entropic regularized synthetic control problem by calling the Sinkhorn algorithm as a subroutine. We provided a statistical interpretation of entropic regularization in light of individualized inference, developing new insights into our formulation as an inference-aware optimization. Entropic regularization plays a crucial role in both efficiency in inference and computation: first, the entropic regularized objective function is gauged for minimizing the width of the individual confidence intervals, trading off bias and variance; second, the entropic regularization enables us to design fast algorithms. The simulation confirms our theoretical insight on the level of entropic regularization needed.

Appendix A. Proofs.

A.1. Proof of Proposition 3.2.

Proof of Proposition 3.2. We have already discussed that (3.1) must admit a unique minimizer $\pi^\star \in \Pi^+_{a,b}$. We show that $\pi^\star$ is a unique fixed point of T_λ by using the Karush-Kuhn-Tucker (KKT) conditions. To see this, let us first rewrite (3.1) as

$$\begin{aligned} \min_{\pi \in \mathbb{R}^{n\times m}_+} \quad & (g(\pi) + \lambda h(\pi)) \\ \text{subject to} \quad & \pi 1_m = a \quad \text{and} \quad \pi^\top 1_n = b. \end{aligned}$$

Then, define the Lagrangian $L\colon \mathbb{R}^{n\times m}_+ \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ by

$$L(\pi, \mu, \nu) = g(\pi) + \lambda h(\pi) + \langle \mu, \pi 1_m - a\rangle + \langle \nu, \pi^\top 1_n - b\rangle.$$

The KKT conditions for $(\pi, \mu, \nu) \in \mathbb{R}^{n\times m}_+ \times \mathbb{R}^n \times \mathbb{R}^m$ are $\pi \in \Pi_{a,b}$ and π is a minimizer of $L(\cdot, \mu, \nu)$ over $\mathbb{R}^{n\times m}_+$. Due to h, we can see that $L(\cdot, \mu, \nu)$ must admit a unique minimizer at the interior of $\mathbb{R}^{n\times m}_+$, namely, all the entries of the unique minimizer are strictly positive. Accordingly, the minimizer must satisfy the first-order condition:

$$\nabla_\pi L(\pi, \mu, \nu) = \nabla g(\pi) + \lambda \nabla h(\pi) + \mu 1_m^\top + 1_n \nu^\top = 0,$$

which leads to

$$\pi = \exp\left(-\frac{\mu 1_m^\top + 1_n \nu^\top + \nabla g(\pi)}{\lambda}\right). \tag{A.1}$$

Hence, $(\pi, \mu, \nu) \in \mathbb{R}^{n\times m}_+ \times \mathbb{R}^n \times \mathbb{R}^m$ satisfies the KKT conditions if $\pi \in \Pi_{a,b}$ and (A.1) holds. Meanwhile, for any $(\pi, \mu, \nu) \in \mathbb{R}^{n\times m}_+ \times \mathbb{R}^n \times \mathbb{R}^m$, if the right-hand side of (A.1) belongs to $\Pi_{a,b}$, it must be $\Phi_\lambda(\nabla g(\pi)) = T_\lambda(\pi)$ which can be deduced from Theorem 1 of [62].

In summary, $(\pi, \mu, \nu) \in \mathbb{R}^{n\times m}_+ \times \mathbb{R}^n \times \mathbb{R}^m$ satisfies the KKT conditions if

$$\pi = \exp\left(-\frac{\mu 1_m^\top + 1_n \nu^\top + \nabla g(\pi)}{\lambda}\right) = T_\lambda(\pi).$$

Meanwhile, for (3.1), we can check that the Slater condition—see for instance, Section 5.2.3 of [19]—is satisfied, and thus we have an optimizer $(\mu^\star, \nu^\star) \in \mathbb{R}^n \times \mathbb{R}^m$ of the dual problem with

strong duality. Accordingly, the KKT conditions are satisfied by $(\pi^\star, \mu^\star, \nu^\star)$. This implies that $\pi^\star = T_\lambda(\pi^\star)$. Hence, T_λ has a fixed point $\pi^\star$. For uniqueness, suppose $\pi = T_\lambda(\pi)$ holds for some $\pi \in \Pi^+_{a,b}$. Again, by Theorem 1 of [62], the matrix $T_\lambda(\pi) = \Phi_\lambda(\nabla g(\pi))$ must take the form of the right-hand side of (A.1) for some $(\mu, \nu) \in \mathbb{R}^n \times \mathbb{R}^m$, meaning that the KKT conditions are satisfied by (π, μ, ν) and thus π is a minimizer of (3.1). As we have already shown that (3.1) admits a unique minimizer $\pi^\star$, we conclude $\pi = \pi^\star$. Hence, $\pi^\star$ is the unique fixed point of T_λ. ∎

A.2. Proof of Theorem 3.3. Theorem 3.3 is based on the following lemma that analyzes the Lipschitz property of the operator Φ_λ—defined in Proposition 3.1—w.r.t. the L^∞ metric.

Lemma A.1. *Fix $\lambda > 0$. Define a map $\Phi_\lambda \colon \mathbb{R}^{n\times m} \to \Pi^+_{a,b}$ by*

$$\Phi_\lambda(C) = \operatorname*{arg\,min}_{\pi \in \Pi_{a,b}} \left(\langle C, \pi \rangle + \lambda h(\pi) \right).$$

Then, for any $C_1, C_2 \in \mathbb{R}^{n\times m}$, we have

$$\|\Phi_\lambda(C_1) - \Phi_\lambda(C_2)\|_1 \le \frac{\|C_1 - C_2\|_\infty}{\lambda}. \tag{A.2}$$

Proof of Lemma A.1. We first claim that for any $\pi \in \Pi_{a,b}$, we have

$$\langle C_1 + \lambda \nabla h(\Phi_\lambda(C_1)), \pi - \Phi_\lambda(C_1) \rangle \ge 0. \tag{A.3}$$

Suppose not, namely, there exists $\pi \in \Pi_{a,b}$ such that

$$\langle C_1 + \lambda \nabla h(\Phi_\lambda(C_1)), \pi - \Phi_\lambda(C_1) \rangle < 0.$$

In other words, letting $q(\pi) := \langle C_1, \pi \rangle + \lambda h(\pi)$, we have $\pi \in \Pi_{a,b}$ such that

$$\langle \nabla q(\Phi_\lambda(C_1)), \pi - \Phi_\lambda(C_1) \rangle < 0.$$

As q is differentiable at $\Phi_\lambda(C_1)$ and q is convex on $\Pi_{a,b}$, this means that we can find a point π' on the line segment between $\Phi_\lambda(C_1)$ and π such that $q(\pi') < q(\Phi_\lambda(C_1))$, which contradicts that $\Phi_\lambda(C_1)$ is a minimizer of q over $\Pi_{a,b}$. Hence, (A.3) must hold for any $\pi \in \Pi_{a,b}$. Letting $\pi = \Phi_\lambda(C_2)$ in (A.3), we have

$$\langle C_1 + \lambda \nabla h(\Phi_\lambda(C_1)), \Phi_\lambda(C_2) - \Phi_\lambda(C_1) \rangle \ge 0.$$

Changing the role of C_1 and C_2,

$$\langle C_2 + \lambda \nabla h(\Phi_\lambda(C_2)), \Phi_\lambda(C_1) - \Phi_\lambda(C_2) \rangle \ge 0.$$

Combining the above two inequalities, we have

$$\langle \nabla h(\Phi_\lambda(C_1)) - \nabla h(\Phi_\lambda(C_2)), \Phi_\lambda(C_1) - \Phi_\lambda(C_2) \rangle \le -\frac{1}{\lambda} \langle C_1 - C_2, \Phi_\lambda(C_1) - \Phi_\lambda(C_2) \rangle. \tag{A.4}$$

By applying Hölder's inequality to the right-hand side of (A.4), we have

$$-\frac{1}{\lambda}\langle C_1 - C_2, \Phi_\lambda(C_1) - \Phi_\lambda(C_2)\rangle \le \frac{\|C_1 - C_2\|_\infty \cdot \|\Phi_\lambda(C_1) - \Phi_\lambda(C_2)\|_1}{\lambda}. \tag{A.5}$$

Also, as h is 1-strongly convex with respect to $\|\cdot\|_1$ on $\Delta_{n,m}$, we have for any $P_1, P_2 \in \Delta_{n,m}$,

$$\langle \nabla h(P_1) - \nabla h(P_2), P_1 - P_2\rangle \ge \|P_1 - P_2\|_1^2,$$

which allows us to lower bound the left-hand side of (A.4) as follows:

$$\langle \nabla h(\Phi_\lambda(C_1)) - \nabla h(\Phi_\lambda(C_2)), \Phi_\lambda(C_1) - \Phi_\lambda(C_2)\rangle \ge \|\Phi_\lambda(C_1) - \Phi_\lambda(C_2)\|_1^2. \tag{A.6}$$

Combining (A.4), (A.5), and (A.6), we have (A.2). ∎

Proof of Theorem 3.3. By (A.2) of Lemma A.1, we have for any $A_1, A_2 \in \mathbb{R}^{n\times m}$,

$$\begin{aligned}\|T_\lambda(A_1) - T_\lambda(A_2)\|_1 &= \|\Phi_\lambda(\nabla g(A_1)) - \Phi_\lambda(\nabla g(A_2))\|_1 \\ &\le \frac{\|\nabla g(A_1) - \nabla g(A_2)\|_\infty}{\lambda} \\ &\le \frac{\|H\|_\infty}{\lambda}\|A_1 - A_2\|_1,\end{aligned}$$

where the last inequality follows from $\|\nabla g(A_1) - \nabla g(A_2)\|_\infty = \|H(A_1 - A_2)\|_\infty \le \|H\|_\infty \|A_1 - A_2\|_1$. Therefore, T_λ is a contraction on $\Pi_{a,b}$ equipped with $\|\cdot\|_1$ if $\lambda > \|H\|_\infty$. By the Banach-Caccioppoli theorem, T_λ has a unique fixed point, and we have (3.3). ∎

A.3. Proof of Theorem 3.5.

Proof of Theorem 3.5. We first show the following from (3.4): for any $k \ge 0$ and $\pi \in \Pi_{a,b}$,

$$\langle \nabla f(\pi^{(k)}), \pi^{(k+1)} - \pi\rangle \le \frac{\langle \nabla h(\pi^{(k)}) - \nabla h(\pi^{(k+1)}), \pi^{(k+1)} - \pi\rangle}{\tau_k}. \tag{A.7}$$

To see this, pick $\pi \in \Pi_{a,b}$ and define $q(\varepsilon) = f((1-\varepsilon)\pi^{(k+1)} + \varepsilon\pi) + \tau_k^{-1} D_h((1-\varepsilon)\pi^{(k+1)} + \varepsilon\pi, \pi^{(k)})$ for $\varepsilon \in [0,1]$. As $q(\varepsilon) \ge q(0)$ for $\varepsilon \in [0,1]$, the right derivative of q at $\varepsilon = 0$ must be nonnegative, which leads to (A.7). Next, letting D_f be the Bregman divergence with respect to f, we have

$$\begin{aligned}f(\pi^{(k+1)}) - f(\pi) &= f(\pi^{(k+1)}) - f(\pi^{(k)}) + f(\pi^{(k)}) - f(\pi) \\ &= D_f(\pi^{(k+1)}, \pi^{(k)}) + \langle \nabla f(\pi^{(k)}), \pi^{(k+1)} - \pi^{(k)}\rangle + f(\pi^{(k)}) - f(\pi) \\ &\le D_f(\pi^{(k+1)}, \pi^{(k)}) + \langle \nabla f(\pi^{(k)}), \pi^{(k+1)} - \pi^{(k)}\rangle + \langle \nabla f(\pi^{(k)}), \pi^{(k)} - \pi\rangle \\ &\le D_f(\pi^{(k+1)}, \pi^{(k)}) + \frac{\langle \nabla h(\pi^{(k)}) - \nabla h(\pi^{(k+1)}), \pi^{(k+1)} - \pi\rangle}{\tau_k},\end{aligned}$$

where the first inequality uses the convexity of f and the second inequality follows from (A.7). For the last term on the right-hand side, we have

$$\begin{aligned}\langle \nabla h(\pi^{(k)}) - \nabla h(\pi^{(k+1)}), \pi^{(k+1)} - \pi \rangle &= \langle \log \tfrac{\pi^{(k)}}{\pi^{(k+1)}}, \pi^{(k+1)} - \pi \rangle \\ &= \langle \log \tfrac{\pi}{\pi^{(k)}} - \log \tfrac{\pi}{\pi^{(k+1)}}, \pi \rangle - \langle \log \tfrac{\pi^{(k+1)}}{\pi^{(k)}}, \pi^{(k+1)} \rangle \\ &= D_h(\pi, \pi^{(k)}) - D_h(\pi, \pi^{(k+1)}) - D_h(\pi^{(k+1)}, \pi^{(k)}).\end{aligned}$$

Hence,

$$f(\pi^{(k+1)}) - f(\pi) \le D_f(\pi^{(k+1)}, \pi^{(k)}) + \tfrac{D_h(\pi,\pi^{(k)}) - D_h(\pi,\pi^{(k+1)}) - D_h(\pi^{(k+1)},\pi^{(k)})}{\tau_k}. \tag{A.8}$$

Meanwhile,

$$\begin{aligned}D_f(\pi^{(k+1)}, \pi^{(k)}) &= \tfrac{\langle \pi^{(k+1)} - \pi^{(k)}, H(\pi^{(k+1)} - \pi^{(k)}) \rangle}{2} + \lambda D_h(\pi^{(k+1)}, \pi^{(k)}) \\ &\le \frac{\|H\|_\infty}{2} \|\pi^{(k+1)} - \pi^{(k)}\|_1^2 + \lambda D_h(\pi^{(k+1)}, \pi^{(k)}) \\ &\le (\|H\|_\infty + \lambda) D_h(\pi^{(k+1)}, \pi^{(k)}),\end{aligned}$$

where the first inequality follows from Hölder's inequality and the second inequality uses Pinsker's inequality $\|\pi^{(k+1)} - \pi^{(k)}\|_1^2 \le 2D_h(\pi^{(k+1)}, \pi^{(k)})$; see, Lemma 2.5 of [64]. Now, as $\tau_k^{-1} = \tau^{-1} \ge \|H\|_\infty + \lambda$, we have $D_f(\pi^{(k+1)}, \pi^{(k)}) \le \frac{D_h(\pi^{(k+1)}, \pi^{(k)})}{\tau}$, which, together with (A.8), leads to

$$f(\pi^{(k+1)}) - f(\pi) \le \frac{D_h(\pi, \pi^{(k)}) - D_h(\pi, \pi^{(k+1)})}{\tau}$$

for any $\pi \in \Pi_{a,b}$. By letting $\pi = \pi^{(k)}$, we have $f(\pi^{(k+1)}) \le f(\pi^{(k)})$ for all $k \ge 0$. Hence,

$$f(\pi^{(T)}) - f(\pi) \le \frac{1}{T} \sum_{k=0}^{T-1} (f(\pi^{(k+1)}) - f(\pi)) \le \frac{D_h(\pi, \pi^{(0)}) - D_h(\pi, \pi^{(T)})}{T\tau} \le \frac{D_h(\pi, \pi^{(0)})}{T\tau}.$$

As f always admits a minimizer on $\Pi_{a,b}$ (even when $\lambda = 0$), plugging a minimizer to π and using the fact that $D_h(\pi, \pi^{(0)}) \le \log(nm)$, we obtain (3.5). Lastly,

$$\begin{aligned}f(\pi^{(T)}) - f(\pi^\star) &= \langle \nabla f(\pi^\star), \pi^{(T)} - \pi^\star \rangle + D_f(\pi^{(T)}, \pi^\star) \\ &= \langle \nabla f(\pi^\star), \pi^{(T)} - \pi^\star \rangle + \tfrac{\langle \pi^{(T)} - \pi^{(k)}, H(\pi^{(T)} - \pi^{(k)}) \rangle}{2} + \lambda D_h(\pi^{(T)}, \pi^\star) \\ &\ge \frac{\lambda}{2} \|\pi^{(T)} - \pi^\star\|_1^2,\end{aligned}$$

where the last inequality uses Pinsker's inequality and $\langle \nabla f(\pi^\star), \pi^{(T)} - \pi^\star \rangle \ge 0$, which holds as $\pi^\star$ is a minimizer of f. Hence, we have (3.6). ∎

A.4. Proof of Proposition 4.3.

Proof of Proposition 4.3. Under Assumption 4.1, note that $\{\varepsilon_{0,i}\}_{i\in\mathcal{C}}$ given $\mathbf{X},\mathbf{Z}$ are independent mean zero random variables. If (4.8) holds, the Lindeberg central limit theorem implies that the variance term $\mathcal{V}_j$ given $\mathbf{X},\mathbf{Z}$ is asymptotically normal with mean zero and variance $\sigma_0^2\sum_{i\in\mathcal{C}}\widehat{p}_{ij}^2$, namely,

$$\frac{\mathcal{V}_j}{\sqrt{\sigma_0^2\sum_{i\in\mathcal{C}}\widehat{p}_{ij}^2}} = \frac{\sum_{i\in\mathcal{C}}\widehat{p}_{ij}\varepsilon_{0,i}}{\sigma_0\sqrt{\sum_{i\in\mathcal{C}}\widehat{p}_{ij}^2}} \mid \mathbf{X},\mathbf{Z} \rightsquigarrow N(0,1) \quad \text{as } N_c\to\infty. \tag{A.9}$$

This implies

$$\lim_{N_c\to\infty}\mathbb{P}\left(|\mathcal{V}_j| \le z_{1-\frac{\alpha}{2}}\cdot\sigma_0\sqrt{\sum_{i\in\mathcal{C}}\widehat{p}_{ij}^2} \mid \mathbf{X},\mathbf{Z}\right) = 1-\alpha,$$

By (4.4) and (4.5), we have

$$|\widehat{Y}_j(0) - f_0(X_j)| \le |\mathcal{B}_j| + |\mathcal{V}_j| \le \|f_0\|_{\mathcal{H}}\cdot\|\phi_{X_j} - \sum_{i\in\mathcal{C}}\widehat{p}_{ij}\phi_{X_i}\|_{\mathcal{H}} + |\mathcal{V}_j|.$$

Therefore, we have

$$\begin{aligned}
&\mathbb{P}\left(f_0(X_j)\in[\widehat{L}_j,\widehat{U}_j] \mid \mathbf{X},\mathbf{Z}\right)\\
&= \mathbb{P}\left(|\widehat{Y}_j(0) - f_0(X_j)| \le \|f_0\|_{\mathcal{H}}\cdot\|\phi_{X_j} - \sum_{i\in\mathcal{C}}\widehat{p}_{ij}\phi_{X_i}\|_{\mathcal{H}} + z_{1-\frac{\alpha}{2}}\cdot\sigma_0\sqrt{\sum_{i\in\mathcal{C}}\widehat{p}_{ij}^2} \mid \mathbf{X},\mathbf{Z}\right)\\
&\ge \mathbb{P}\left(|\mathcal{V}_j| \le z_{1-\frac{\alpha}{2}}\cdot\sigma_0\sqrt{\sum_{i\in\mathcal{C}}\widehat{p}_{ij}^2} \mid \mathbf{X},\mathbf{Z}\right).
\end{aligned}$$

Hence, we have (4.11). ∎

A.5. Proof of Proposition 4.5. First, we need the following simple facts on the Kullback-Leibler divergence, Hellinger distance, and χ^2 distance, which will help us relate the entropy term and the standard Euclidean norm of the weights.

Proposition A.2. *For any probability vector $p\in\Delta_n$,*

$$\frac{1}{2(n\|p\|_\infty+1)} \le \frac{\sum_{i=1}^n p_i\log(p_i)+\log(n)}{n\sum_{i=1}^n p_i^2-1} \le 1.$$

Proof of Proposition A.2. For any $p=(p_1,\dots,p_n)\in\Delta_n$, we have

$$\sum_{i=1}^n p_i\log(p_i)+\log(n) \le n\sum_{i=1}^n p_i^2 - 1,$$

where the left-hand side is the Kullback-Leibler (KL) divergence of p from $\frac{1}{n}1_n$ which is known to be bounded by the χ^2 distance of p from $\frac{1}{n}1_n$ on the right-hand side; see Lemma 2.7 of

[64]. Meanwhile, the KL divergence is bounded below by the Hellinger distance; see Lemma 2.4 of [64]. Therefore, we have

$$\sum_{i=1}^{n} p_i \log(p_i) + \log(n) \geq \sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{1/n})^2,$$

where the right-hand side is the Hellinger distance between p and $\frac{1}{n}1_n$. The right-hand side can be further lower bounded,

$$\sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{1/n})^2 = \sum_{i=1}^{n} \frac{(p_i - 1/n)^2}{(\sqrt{p_i} + \sqrt{1/n})^2} \geq \frac{n \sum_{i=1}^{n} p_i^2 - 1}{2(n\|p\|_\infty + 1)}.$$

Combining these inequalities, we have

$$\frac{1}{2(n\|p\|_\infty + 1)} \leq \frac{\sum_{i=1}^{n} p_i \log(p_i) + \log(n)}{n \sum_{i=1}^{n} p_i^2 - 1} \leq 1.$$

■

Proof of Proposition 4.5. The interval length is as follows:

$$\mathrm{Len}_j := 2\|f_0\|_{\mathcal{H}} \cdot \|\phi_{X_j} - \sum_{i \in \mathcal{C}} \widehat{p}_{ij} \phi_{X_i}\|_{\mathcal{H}} + 2z_{1-\frac{\alpha}{2}} \cdot \sigma_0 \sqrt{\sum_{i \in \mathcal{C}} \widehat{p}_{ij}^2}.$$

Hence,

$$\mathrm{Len}_j^2 \leq 8\|f_0\|_{\mathcal{H}}^2 \cdot \|\phi_{X_j} - \sum_{i \in \mathcal{C}} \widehat{p}_{ij} \phi_{X_i}\|_{\mathcal{H}}^2 + 8z_{1-\frac{\alpha}{2}}^2 \cdot \sigma_0^2 \sum_{i \in \mathcal{C}} \widehat{p}_{ij}^2. \tag{A.10}$$

By Proposition A.2, we have

$$\begin{aligned} \sum_{i \in \mathcal{C}} \widehat{p}_{ij}^2 &\leq \frac{1 + 2(N_c \cdot \max_{i \in \mathcal{C}} \widehat{p}_{ij} + 1)(\sum_{i \in \mathcal{C}} \widehat{p}_{ij} \log \widehat{p}_{ij} + \log(N_c))}{N_c} \\ &\leq \frac{1 + M(N_t \cdot \sum_{i \in \mathcal{C}} \widehat{\pi}_{ij} \log \frac{\widehat{\pi}_{ij}}{e} + 1 + \log(N_t) + \log(N_c))}{N_c}, \end{aligned}$$

which, combined with (A.10), leads to the desired upper bound on Len_j^2. Meanwhile, we have

$$\mathrm{Len}_j^2 \geq 4\|f_0\|_{\mathcal{H}}^2 \cdot \|\phi_{X_j} - \sum_{i \in \mathcal{C}} \widehat{p}_{ij} \phi_{X_i}\|_{\mathcal{H}}^2 + 4z_{1-\frac{\alpha}{2}}^2 \cdot \sigma_0^2 \sum_{i \in \mathcal{C}} \widehat{p}_{ij}^2. \tag{A.11}$$

By Proposition A.2, we have

$$\begin{aligned} \sum_{i \in \mathcal{C}} \widehat{p}_{ij}^2 &\geq \frac{1 + \sum_{i \in \mathcal{C}} \widehat{p}_{ij} \log \widehat{p}_{ij} + \log(N_c)}{N_c} \\ &\geq \frac{1 + N_t \cdot \sum_{i \in \mathcal{C}} \widehat{\pi}_{ij} \log \frac{\widehat{\pi}_{ij}}{e} + 1 + \log(N_t) + \log(N_c)}{N_c}, \end{aligned}$$

which, combined with (A.11), leads to the desired lower bound on Len_j^2. ■

Acknowledgments. The authors thank Max Farrell, Avi Feller, Chris Hansen, and Guido Imbens for suggestions.

REFERENCES

[1] A. Abadie, *Using Synthetic Controls: Feasibility, Data requirements, and Methodological Aspects*, Journal of Economic Literature, 59 (2021), pp. 391–425.
[2] A. Abadie and M. D. Cattaneo, *Introduction to the Special Section on Synthetic Control Methods*, Journal of the American Statistical Association, 116 (2021), pp. 1713–1715.
[3] A. Abadie, A. Diamond, and J. Hainmueller, *Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program*, Journal of the American Statistical Association, 105 (2010), pp. 493–505.
[4] A. Abadie and J. Gardeazabal, *The Economic Costs of Conflict: A Case Study of the Basque Country*, American Economic Review, 93 (2003), pp. 113–132.
[5] A. Abadie and G. W. Imbens, *Large Sample Properties of Matching Estimators for Average Treatment Effects*, Econometrica, 74 (2006), pp. 235–267.
[6] A. Abadie and G. W. Imbens, *Bias-Corrected Matching Estimators for Average Treatment Effects*, Journal of Business & Economic Statistics, 29 (2011), pp. 1–11.
[7] A. Abadie and J. L'hour, *A Penalized Synthetic Control Estimator for Disaggregated Data*, Journal of the American Statistical Association, 116 (2021), pp. 1817–1834.
[8] J. Altschuler, J. Niles-Weed, and P. Rigollet, *Near-linear Time Approximation Algorithms for Optimal Transport via Sinkhorn Iteration*, in Advances in Neural Information Processing Systems, 2017.
[9] D. Arkhangelsky, S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager, *Synthetic Difference-in-Differences*, American Economic Review, 111 (2021), pp. 4088–4118.
[10] D. Arkhangelsky and D. Hirshberg, *Large-Sample Properties of the Synthetic Control Method under Selection on Unobservables*, arXiv preprint arXiv:2311.13575, (2023).
[11] D. Arkhangelsky and G. Imbens, *Causal Models for Longitudinal and Panel Data: A Survey*, The Econometrics Journal, 27 (2024), pp. C1–C61.
[12] T. B. Armstrong and M. Kolesár, *Optimal Inference in a Class of Regression Models*, Econometrica, 86 (2018), pp. 655–683.
[13] S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi, *Matrix Completion Methods for Causal Panel Data Models*, Journal of the American Statistical Association, 116 (2021), pp. 1716–1730.
[14] S. Athey and G. Imbens, *Recursive Partitioning for Heterogeneous Causal Effects*, Proceedings of the National Academy of Sciences, 113 (2016), pp. 7353–7360.
[15] A. Belloni, V. Chernozhukov, and C. Hansen, *Inference on Treatment Effects after Selection among High-Dimensional Controls*, Review of Economic Studies, 81 (2014), pp. 608–650.
[16] E. Ben-Michael, A. Feller, and J. Rothstein, *The Augmented Synthetic Control Method*, Journal of the American Statistical Association, 116 (2021), pp. 1789–1803.
[17] E. Ben-Michael, A. Feller, and J. Rothstein, *Synthetic Controls with Staggered Adoption*, Journal of the Royal Statistical Society Series B: Statistical Methodology, 84 (2022), pp. 351–381.
[18] M. Blondel, V. Seguy, and A. Rolet, *Smooth and Sparse Optimal Transport*, in International Conference on Artificial Intelligence and Statistics, 2018.
[19] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
[20] C. Bunne, D. Alvarez-Melis, A. Krause, and S. Jegelka, *Learning Generative Models across Incomparable Spaces*, in International Conference on Machine Learning, 2019.
[21] G. Carlier, *On the Linear Convergence of the Multimarginal Sinkhorn Algorithm*, SIAM Journal on Optimization, 32 (2022), pp. 786–794.
[22] M. D. Cattaneo, Y. Feng, and R. Titiunik, *Prediction Intervals for Synthetic Control Methods*, Journal of the American Statistical Association, 116 (2021), pp. 1865–1880.
[23] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey, *Double/Debiased/Neyman Machine Learning of Treatment Effects*, American Economic Review, 107

(2017), pp. 261–265.
[24] V. Chernozhukov, K. Wüthrich, and Y. Zhu, *An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls*, Journal of the American Statistical Association, 116 (2021), pp. 1849–1864.
[25] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik, *Dealing with Limited Overlap in Estimation of Average Treatment Effects*, Biometrika, 96 (2009), pp. 187–199.
[26] M. Cuturi, *Sinkhorn Distances: Lightspeed Computation of Optimal Transport*, Advances in Neural Information Processing Systems, 26 (2013).
[27] N. Doudchenko and G. W. Imbens, *Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis*, tech. report, National Bureau of Economic Research, 2016.
[28] M. Essid and J. Solomon, *Quadratically Regularized Optimal Transport on Graphs*, SIAM Journal on Scientific Computing, 40 (2018), pp. A1961–A1986.
[29] M. H. Farrell, *Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations*, Journal of Econometrics, 189 (2015), pp. 1–23.
[30] M. H. Farrell, T. Liang, and S. Misra, *Deep Neural Networks for Estimation and Inference*, Econometrica, 89 (2021), pp. 181–213.
[31] P. Ghosal and M. Nutz, *On the Convergence Rate of Sinkhorn's Algorithm*, Mathematics of Operations Research, 51 (2026), pp. 1080–1096.
[32] J. Hahn, *On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects*, Econometrica, (1998), pp. 315–331.
[33] J. Hainmueller, *Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies*, Political Analysis, 20 (2012), pp. 25–46.
[34] J. J. Heckman, H. Ichimura, and P. E. Todd, *Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme*, Review of Economic Studies, 64 (1997), pp. 605–654.
[35] J. J. Heckman and E. Vytlacil, *Structural Equations, Treatment Effects, and Econometric Policy Evaluation*, Econometrica, 73 (2005), pp. 669–738.
[36] K. Hirano and G. W. Imbens, *Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization*, Health Services and Outcomes Research Methodology, 2 (2001), pp. 259–278.
[37] K. Hirano, G. W. Imbens, and G. Ridder, *Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score*, Econometrica, 71 (2003), pp. 1161–1189.
[38] D. A. Hirshberg and S. Wager, *Augmented Minimax Linear Estimation*, The Annals of Statistics, 49 (2021), pp. 3206–3227.
[39] C. Hsiao, H. Steve Ching, and S. Ki Wan, *A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Economic Integration of Hong Kong with Mainland China*, Journal of Applied Econometrics, 27 (2012), pp. 705–740.
[40] J. D. Huling and S. Mak, *Energy Balancing of Covariate Distributions*, Journal of Causal Inference, 12 (2024), p. 20220029.
[41] Y. Hur, W. Guo, and T. Liang, *Reversible Gromov–Monge Sampler for Simulation-Based Inference*, SIAM Journal on Mathematics of Data Science, 6 (2024), pp. 283–310.
[42] N. Kallus, *Generalized Optimal Matching Methods for Causal Inference*, Journal of Machine Learning Research, 21 (2020), pp. 1–54.
[43] E. H. Kennedy, *Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects*, Electronic Journal of Statistics, 17 (2023), pp. 3008–3049.
[44] N. Kreif, R. Grieve, D. Hangartner, A. J. Turner, S. Nikolova, and M. Sutton, *Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units*, Health Economics, 25 (2016), pp. 1514–1528.
[45] L. Lei and E. J. Candès, *Conformal Inference of Counterfactuals and Individual Treatment Effects*, Journal of the Royal Statistical Society Series B: Statistical Methodology, 83 (2021), pp. 911–938.
[46] K. T. Li, *Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods*, Journal of the American Statistical Association, 115 (2020), pp. 2068–2083.
[47] T. MaCurdy, X. Chen, and H. Hong, *Flexible Estimation of Treatment Effect Parameters*, American Economic Review, 101 (2011), pp. 544–551.

[48] F. Mémoli, *Gromov-Wasserstein Distances and the Metric Approach to Object Matching*, Foundations of Computational Mathematics, 11 (2011), pp. 417–487.
[49] F. Mémoli and T. Needham, *Comparison results for Gromov–Wasserstein and Gromov–Monge distances*, ESAIM: Control, Optimisation and Calculus of Variations, 30 (2024), p. 78.
[50] J. Neyman, *On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.* Translated from Polish to English with discussion in *Statistical Science*, 5(4):465–472, 1990, 1923.
[51] X. Nie and S. Wager, *Quasi-Oracle Estimation of Heterogeneous Treatment Effects*, Biometrika, 108 (2021), pp. 299–319.
[52] M. Nutz, *Quadratically Regularized Optimal Transport: Existence and Multiplicity of Potentials*, arXiv preprint arXiv:2404.06847, (2024).
[53] G. Peyré and M. Cuturi, *Computational Optimal Transport: With Applications to Data Science*, Foundations and Trends® in Machine Learning, 11 (2019), pp. 355–607.
[54] G. Peyré, M. Cuturi, and J. Solomon, *Gromov-Wasserstein Averaging of Kernel and Distance Matrices*, in International Conference on Machine Learning, 2016.
[55] M. W. Robbins, J. Saunders, and B. Kilmer, *A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention*, Journal of the American Statistical Association, 112 (2017), pp. 109–126.
[56] P. R. Rosenbaum, *Model-Based Direct Adjustment*, Journal of the American Statistical Association, 82 (1987), pp. 387–394.
[57] P. R. Rosenbaum, *Modern Algorithms for Matching in Observational Studies*, Annual Review of Statistics and Its Application, 7 (2020), pp. 143–176.
[58] P. R. Rosenbaum and D. B. Rubin, *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, Biometrika, 70 (1983), pp. 41–55.
[59] D. B. Rubin, *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies*, Journal of Educational Psychology, 66 (1974), p. 688.
[60] D. B. Rubin, *Matched Sampling for Causal Effects*, Cambridge University Press, 2006.
[61] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.
[62] R. Sinkhorn, *Diagonal Equivalence to Matrices with Prescribed Row and Column Sums*, The American Mathematical Monthly, 74 (1967), pp. 402–405.
[63] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer, 2008.
[64] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, Springer, 2009.
[65] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, 2003.
[66] S. Wager and S. Athey, *Estimation and Inference of Heterogeneous Treatment Effects using Random Forests*, Journal of the American Statistical Association, 113 (2018), pp. 1228–1242.
[67] H. Xu, D. Luo, and L. Carin, *Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching*, in Advances in Neural Information Processing Systems, 2019.
[68] J. R. Zubizarreta, *Stable Weights that Balance Covariates for Estimation with Incomplete Outcome Data*, Journal of the American Statistical Association, 110 (2015), pp. 910–922.
[69] J. R. Zubizarreta, E. A. Stuart, D. S. Small, and P. R. Rosenbaum, *Handbook of Matching and Weighting Adjustments for Causal Inference*, CRC Press, 2023.