A Causal Inference with Observational Studies

In this section, we introduce necessary preliminaries about causal inference and treatment effect estimation, for readers that are unfamiliar with this area. We then present our theoretical insights based on these preliminaries.

A.1 Problem Formulation

This section formalizes the definitions, assumptions, and useful lemmas in causal inference from observational data. Following the notations in Section 2.1, an individual with covariates x has two potential outcomes, namely $Y_1(x)$ given it is treated and $Y_0(x)$ otherwise. The ground-truth individual treatment effect (CATE) is the difference in its potential outcomes.

Definition A.1. *The individual treatment effect (CATE) for a unit with covariates x is*

$$\tau(x) := \mathbb{E}\left[Y_1 - Y_0 \mid x\right], \tag{16}$$

where we abbreviate $Y_1(x)$ to Y_1 for brevity. The expectation is over the potential outcome space $\mathcal{Y}$.

Estimating CATE with observational data is a common practice in causal inference, which has long been confronted with two primary challenges:

- Missing counterfactuals: where only the factual outcome is observable. If a patient is treated, for instance, we can never observe what would have happened if the patient was untreated in the same situation.
- Treatment selection bias, where individuals have preferences for treatment selection. For example, doctors would adapt different treatment plans for patients with different health conditions. It would make the treated and untreated populations heterogeneous. CATE estimators naïvely trained to minimize the factual outcome error would overfit the respective group's properties and thus cannot generalize well to the entire population.

Pearl and Mackenzie [53] suggested a two-step methodology to overcome these two challenges. The first step is identification, which aims to construct an unbiased statistical estimand to identify the causal estimand (*e.g.*, $\tau(x)$) based on the adjustment formula. Note that not all causal estimands are identifiable, *e.g.*, CATE is identifiable only if Assumption A.1-A.4 hold.

Assumption A.1. *(Unconfoundedness). For all covariates x in the population of interest (i.e., x with $\mathbb{P}(X = x) > 0$), we have conditional independence $(Y_0, Y_1) \perp\!\!\!\perp T \mid X = x$. That is, potential outcomes are conditionally independent of treatment assignment.*

Assumption A.2. *(Consistency). For all covariates x in the population of interest, we have $Y = Y_t$. That is, the observed outcome is consistent with the potential outcome w.r.t. the assigned treatment.*

Assumption A.3. *(Positivity). For all covariates x in the population of interest, we have $0 < \mathbb{P}(T = 1 \mid X = x) < 1$. That is, all individuals have a chance to be assigned both treatments.*

Assumption A.4. *(SUTVA). The potential outcomes for any unit are not affected by the treatment assignments of other units, and there are no different forms or versions of each treatment level for each unit that can produce different potential outcomes [28].*

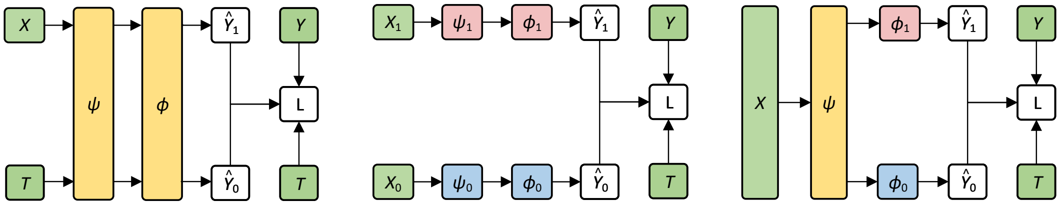
The second step is estimation, which aims to estimate the derived statistical estimand with observational data. Lemma A.1 illustrates how this two-step approach can be used for CATE estimation.

Lemma A.1. *The CATE estimand $\tau(x)$ can be identified as:*

$$\begin{aligned} \mathbb{E}\left[Y_1 - Y_0 \mid X = x\right] &= \mathbb{E}\left[Y_1 \mid X = x\right] - \mathbb{E}\left[Y_0 \mid X = x\right] \\ &\overset{(1)}{=} \mathbb{E}\left[Y_1 \mid X = x, T = 1\right] - \mathbb{E}\left[Y_0 \mid X = x, T = 0\right] \\ &\overset{(2)}{=} \mathbb{E}\left[Y \mid X = x, T = 1\right] - \mathbb{E}\left[Y \mid X = x, T = 0\right], \end{aligned} \tag{17}$$

where (1) stems from the unconfoundedness assumption A.1; (2) stems from the consistency assumption A.2. The derived estimand is fully composed of statistical estimands, which can only be estimated under the positivity assumption A.3. Otherwise, if the positivity assumption is violated, we have:

$$\begin{aligned} \mathbb{E}\left[Y \mid X = x, T = 1\right] &= \int y \cdot \mathbb{P}(Y = y \mid X = x, T = 1)\, dy \\ &= \int y \cdot \frac{\mathbb{P}(Y = y, X = x, T = 1)}{\mathbb{P}(T = 1 \mid X = x)\mathbb{P}(X = x)}\, dy, \end{aligned} \tag{18}$$



(a) S-learner (b) T-learner (c) TARNet

Figure 6: Architecture of Meta-learner based CATE estimators, consisting of inputs (green), outputs (white), shared mappings (yellow), and mappings for treated and untreated units (red and blue, respectively).

which is not computable as there exists $x \in \mathcal{X}$ which makes $\mathbb{P}(T = 1 \mid X = x) = 0$.

A.2 Meta-learners for CATE estimation with observational data

In an effort to solve missing counterfactuals, existing meta-learner based methods [36, 50] decompose the CATE estimation problem into several subproblems that can be solved with any supervised learning method. As depicted in Figure 6, S-learner regards the treatment indicator T as one of the covariates X, and utilizes the shared representation mapping ψ and outcome mapping ϕ to estimate the factual outcomes. However, because the network structure does not highlight the role of treatment indicator, it may be overlooked when treatment effects are minimal. T-learner models the factual outcomes for treated units X_1 and untreated units X_0 separately, which highlights the treatment indicator's effect; however, it reduces the data efficiency and is therefore inapplicable when the dataset is small. Künzel et al. [36] discuss the advantages and limitations of these two approaches in more detail.

Definition A.2. *Let $\psi : \mathcal{X} \to \mathcal{R}$ be a mapping from support $\mathcal{X}$ to $\mathcal{R}$. That is, $\forall x \in \mathcal{X}$, $\exists r = \psi(x) \in \mathcal{R}$. Let $\phi : \mathcal{R} \times \mathcal{T} \to \mathcal{Y}$ be a mapping from support $\mathcal{R} \times \mathcal{T}$ to $\mathcal{Y}$. That is, it maps the representations and treatment indicator to the corresponding factual outcome. For example, $Y_1 = \phi_1(R)$, $Y_0 = \phi_0(R)$, where we will always abbreviate $\phi(R, T = 1)$ and $\phi(R, T = 0)$ to $\phi_1(R)$ and $\phi_0(R)$, respectively.*

Assumption A.5. *$\phi : \mathcal{X} \to \mathcal{R}$ is differentiable and invertible, with its inverse ϕ^{-1} defined over $\mathcal{R}$.*

TARNet [67] in Figure 6 (c) obtains better results by absorbing the advantages of both T-learner and S-learner, which consists of a representation mapping ψ and an outcome mapping ϕ as defined in Definition A.2. For a unit with covariates X, TARNet estimates CATE as the difference in predicted outcomes when T is set to treated and untreated:

$$\hat{\tau}_{\psi,\phi}(X) := \hat{Y}_1 - \hat{Y}_0, \quad \text{where} \quad \hat{Y}_1 = \phi_1(\psi(X)), \quad \hat{Y}_0 = \phi_0(\psi(X)), \tag{19}$$

where ψ is trained over all units, ϕ_1 and ϕ_0 are trained over the treated and untreated units, respectively, to minimize the factual error $\epsilon_{\mathrm{F}}(\phi, \psi)$ in Definition A.3. Finally, the performance of the CATE estimator is mainly evaluated with PEHE:

$$\epsilon_{\mathrm{PEHE}}(\psi, \phi) = \int_{\mathcal{X}} \left(\hat{\tau}_{\psi,\phi}(x) - \tau(x)\right)^2 \mathbb{P}(x)\, dx. \tag{20}$$

Definition A.3. *Let $\mathbb{L}$ be the loss function that measures the quality of outcome estimation, e.g., the squared loss. The expected loss for the units with covariates x and treatment indicator t is:*

$$l_{\psi,\phi}(x, t) := \int_{\mathcal{Y}} \mathbb{L}(Y_t, \phi(\psi(x), t)) \cdot \mathbb{P}(Y_t \mid x)\, dY_t. \tag{21}$$

where $\mathbb{L}$ is realized with the squared loss: $\mathbb{L}(Y_t, \psi(\phi(x), t)) = (Y_t - \psi(\phi(x), t))^2$ in our scenario. The expected factual outcome estimation error for treated, untreated and all units are:

$$\begin{aligned}
\epsilon_{\mathrm{F}}^{\mathrm{T=1}}(\psi, \phi) &:= \int_{\mathcal{X}} l_{\psi,\phi}(x, 1) \cdot \mathbb{P}^{\mathrm{T=1}}(x)\, dx, \\
\epsilon_{\mathrm{F}}^{\mathrm{T=0}}(\psi, \phi) &:= \int_{\mathcal{X}} l_{\psi,\phi}(x, 0) \cdot \mathbb{P}^{\mathrm{T=0}}(x)\, dx, \\
\epsilon_{\mathrm{F}}(\psi, \phi) &:= \int_{\mathcal{X} \times \mathcal{T}} l_{\psi,\phi}(x, t) \cdot \mathbb{P}(x, t)\, dx dt.
\end{aligned} \tag{22}$$

A.3 Representation-based Methods for Treatment Selection Bias

However, the treatment selection bias makes covariate distributions across groups shift. As such, ϕ_1 and ϕ_0 would overfit the respective group's properties and thus cannot generalize well to the entire population. For example, as shown in Figure 1(a), the potential outcome estimator ϕ_1 trained with treated units cannot generalize to the untreated units. Therefore, the resulting $\hat{\tau}$ would be biased.

Definition A.4. *Let* $\mathbb{P}^{\mathrm{T=1}}(x) := \mathbb{P}(x \mid T = 1)$ *and* $\mathbb{P}^{\mathrm{T=0}}(x) := \mathbb{P}(x \mid T = 0)$ *be the covariate distribution for treated and untreated groups, respectively. Let* $\mathbb{P}_{\psi}^{\mathrm{T=1}}(r)$ *and* $\mathbb{P}_{\psi}^{\mathrm{T=0}}(x)$ *be that of representations induced by the representation mapping* $r = \psi(x)$ *defined in Definition 2.2.*

To mitigate the effect of treatment selection bias, representation-based approaches [31, 67] minimize the distribution discrepancy of different groups in the representation space. In particular, the integral probability metric (IPM) in Definition A.4 is a widely used metric that measures the discrepancy of two distributions. Shalit et al. [67] propose to optimize the PEHE by minimizing the estimation error of factual outcomes ϵ_{F} and the IPM of learned representations between treated and untreated groups. They further provide theoretical results to back up their claim as per Theorem A.1.

Definition A.5. *Consider two distribution functions* $\mathbb{P}^{\mathrm{T=1}}(x)$ *and* $\mathbb{P}^{\mathrm{T=0}}(x)$ *supported over* $\mathcal{X}$*, let* $\mathcal{F}$ *be a sufficiently large function family, the integral probability metric induced by* $\mathcal{F}$ *is*

$$\mathrm{IPM}_{\mathcal{F}}\left(\mathbb{P}^{\mathrm{T=1}}, \mathbb{P}^{\mathrm{T=0}}\right) = \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f(x) \left(\mathbb{P}^{\mathrm{T=1}}(x) - \mathbb{P}^{\mathrm{T=0}}(x)\right) dx \right|, \tag{23}$$

Theorem A.1. *Let* ψ *and* ϕ *be the mappings in Definition 2.2,* $\mathcal{F}$ *be a predefined sufficiently large function family of* ϕ*,* $\mathrm{IPM}_{\mathcal{F}}$ *be the integral probability metric induced by* $\mathcal{F}$*. Assume there exists a constant* $B_\psi > 0$*, such that for* $t \in \{0, 1\}$*,* $\frac{1}{B_\psi} \cdot l_{\psi,\phi}(x, t) \in \mathcal{F}$ *holds. [67] demonstrate:*

$$\epsilon_{\mathrm{PEHE}}(\psi, \phi) \leq 2\left(\epsilon_{\mathrm{F}}^{\mathrm{T=0}}(\psi, \phi) + \epsilon_{\mathrm{F}}^{\mathrm{T=1}}(\psi, \phi) + B_\psi \mathrm{IPM}_{\mathcal{F}}\left(\mathbb{P}_{\psi}^{\mathrm{T=1}}, \mathbb{P}_{\psi}^{\mathrm{T=0}}\right) - 2\sigma_Y^2\right), \tag{24}$$

where $\epsilon_{\mathrm{F}}^{\mathrm{T=0}}$ *and* $\epsilon_{\mathrm{F}}^{\mathrm{T=1}}$ *follow Definition A.3,* $\mathbb{P}_{\psi}^{\mathrm{T=1}}(r)$ *and* $\mathbb{P}_{\psi}^{\mathrm{T=0}}(x)$ *follow Definition A.4.*

A.4 Theoretical Results and Extensions

Two problems with Theorem A.1 warrant further consideration. Firstly, the IPM metric, albeit with profound theoretical properties, is intractable. To counter this, note that the IPM holds for any sufficiently large function families, it is feasible to consider IPM in certain function families $\mathcal{F}$ to make it tractable. For example. in the 1-Lipschitz function family, the IPM is equivalent to the Wasserstein divergence as per Kantorovich-Rubinstein duality [67, 72]. As such, the IPM discrepancy can be casted to the Wasserstein discrepancy for computation as per Lemma A.2.

Lemma A.2. *Consider two distribution functions* $\mathbb{P}_1(x)$ *and* $\mathbb{P}_2(x)$ *supported over* $\mathcal{X}$*; let* $\mathcal{F}$ *be the family of* 1*-Lipschitz functions,* $\mathbb{W}$ *be the Wasserstein distance, Villani [72] demonstrate*

$$\mathrm{IPM}_{\mathcal{F}}\left(\mathbb{P}_1, \mathbb{P}_2\right) = \mathbb{W}\left(\mathbb{P}_1, \mathbb{P}_2\right) \tag{25}$$

Another issue that needs further consideration is sampling complexity. Specifically, Theorem A.1 holds if and only if the entire populations of treated and untreated groups are available. However, since the representation-based approaches update parameters with stochastic gradient methods, only a mini-batch of the population is accessible within each iteration. As such, it remains questionable how does Theorem A.1 perform at a mini-batch level in practice.

Lemma A.3. *Let* $\mathbb{P}(x)$ *be a probability measure supported over* $\mathcal{X} \in \mathbb{R}^d$ *satisfying* $T_1(\lambda)$ *inequality. Let* $\hat{\mathbb{P}}(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$ *be the corresponding empirical measure with* N *units. Bolley et al. [4] and Redko et al. [57] demonstrate that for any* $\mathrm{d}' > \mathrm{d}$ *and* $\lambda' < \lambda$*, there exists some constant* N_0*, such that for any* $\varepsilon > 0$ [4] *and* $N \geq N_0 \max(\varepsilon^{-(d+2)}, 1)$*, we have*

$$\mathbb{P}\left(\mathbb{W}\left(\mathbb{P}(x), \hat{\mathbb{P}}(x)\right) > \varepsilon\right) \leq \exp\left(-\frac{\lambda'}{2} N \varepsilon^2\right) \tag{26}$$

where d', λ' *can be calculated explicitly.*

---

[4] While there is a risk of symbol reuse, we use ε here to denote sampling error, and ϵ to control the strength of entropic regularization in optimal transport.

Hoeffding's inequality is a powerful statistical tool to quantify such sampling effects, which is proved to be applicable for $\mathbb{W}$ by [4]. Therefore, it is natural to expand $\mathbb{W}$ according to Lemma A.3 to extend Theorem A.1 to mini-batch situations, in order to quantify the sampling effects.

Theorem A.2. *Let ψ and ϕ be the representation mapping and factual outcome mapping, respectively; $\hat{\mathbb{W}}_\psi$ be the discrepancy across groups at a mini-batch level. With the probability of at least $1-\delta$, we have:*

$$\epsilon_{\mathrm{PEHE}}(\psi,\phi) \le 2\left[\epsilon_{\mathrm{F}}^{\mathrm{T=1}}(\psi,\phi) + \epsilon_{\mathrm{F}}^{\mathrm{T=0}}(\psi,\phi) + B_\psi \hat{\mathbb{W}}_\psi - 2\sigma_Y^2 + \mathcal{O}(\frac{1}{\delta N})\right], \tag{27}$$

where $\epsilon_{\mathrm{F}}^{\mathrm{T=1}}$ and $\epsilon_{\mathrm{F}}^{\mathrm{T=0}}$ are the expected losses of factual outcome estimation over treated and untreated units, respectively. N is the batch size, σ_Y^2 is the variance of outcomes, B_ψ is some constant such that $\frac{1}{B_\psi}\cdot l_{\psi,\phi}(x,t)$ belongs to the family of 1-Lipschitz functions, $\mathcal{O}(\cdot)$ is the sampling complexity term.

Proof. According to Theorem A.1 we have:

$$\epsilon_{\mathrm{PEHE}}(\psi,\phi) \le 2\left(\epsilon_{\mathrm{F}}^{\mathrm{T=0}}(\psi,\phi) + \epsilon_{\mathrm{F}}^{\mathrm{T=1}}(\psi,\phi) + B_\psi \mathrm{IPM}_{\mathcal{F}}\left(\mathbb{P}_\psi^{\mathrm{T=1}}, \mathbb{P}_\psi^{\mathrm{T=0}}\right) - 2\sigma_Y^2\right). \tag{28}$$

Assuming that there exists a constant $B_\psi > 0$, such that for $t \in \{0,1\}$, $\frac{1}{B_\psi}\cdot l_{\psi,\phi}(x,t)$ belongs to the family of 1-Lipschitz functions. According to Lemma A.2, we have

$$\epsilon_{\mathrm{PEHE}}(\psi,\phi) \le 2\left(\epsilon_{\mathrm{F}}^{\mathrm{T=0}}(\psi,\phi) + \epsilon_{\mathrm{F}}^{\mathrm{T=1}}(\psi,\phi) + B_\psi \mathbb{W}\left(\mathbb{P}_\psi^{\mathrm{T=1}}, \mathbb{P}_\psi^{\mathrm{T=0}}\right) - 2\sigma_Y^2\right). \tag{29}$$

Following Definition 3.1, let $\hat{\mathbb{P}}_\psi^{\mathrm{T=1}}(r)$ and $\hat{\mathbb{P}}_\psi^{\mathrm{T=0}}(r)$ be the empirical distributions of representations at a mini-batch level, containing N_1 treated units and N_0 untreated units, respectively. Then we have:

$$\begin{aligned}\mathbb{W}\left(\mathbb{P}_\psi^{\mathrm{T=1}}, \mathbb{P}_\psi^{\mathrm{T=0}}\right) &\le \mathbb{W}\left(\mathbb{P}_\psi^{\mathrm{T=1}}, \hat{\mathbb{P}}_\psi^{\mathrm{T=1}}\right) + \mathbb{W}\left(\mathbb{P}_\psi^{\mathrm{T=0}}, \hat{\mathbb{P}}_\psi^{\mathrm{T=1}}\right)\\ &\le \mathbb{W}\left(\mathbb{P}_\psi^{\mathrm{T=1}}, \hat{\mathbb{P}}_\psi^{\mathrm{T=1}}\right) + \mathbb{W}\left(\mathbb{P}_\psi^{\mathrm{T=0}}, \hat{\mathbb{P}}_\psi^{\mathrm{T=0}}\right) + \mathbb{W}\left(\hat{\mathbb{P}}_\psi^{\mathrm{T=0}}, \hat{\mathbb{P}}_\psi^{\mathrm{T=1}}\right)\\ &:= \mathbb{W}\left(\mathbb{P}_\psi^{\mathrm{T=1}}, \hat{\mathbb{P}}_\psi^{\mathrm{T=1}}\right) + \mathbb{W}\left(\mathbb{P}_\psi^{\mathrm{T=0}}, \hat{\mathbb{P}}_\psi^{\mathrm{T=0}}\right) + \hat{\mathbb{W}}_\psi,\end{aligned} \tag{30}$$

because we have the triangular inequality for $\mathbb{W}$. The Hoeffding inequality in Lemma A.3 further gives the following inequality which holds with the probability at least $1-\delta$:

$$\begin{aligned}\mathbb{W}\left(\mathbb{P}_\psi^{\mathrm{T=1}}, \hat{\mathbb{P}}_\psi^{\mathrm{T=1}}\right) &\le \sqrt{2\log\left(\frac{1}{\delta}\right)/\lambda' N_1}\\ \mathbb{W}\left(\mathbb{P}_\psi^{\mathrm{T=0}}, \hat{\mathbb{P}}_\psi^{\mathrm{T=0}}\right) &\le \sqrt{2\log\left(\frac{1}{\delta}\right)/\lambda' N_0}.\end{aligned} \tag{31}$$

Denote $N := N_0 + N_1$ as the batch size, $\theta := N_1/N$ as the ratio of treated units in the current batch. Combining (30) and 31 we have

$$\begin{aligned}\mathbb{W}\left(\mathbb{P}_\psi^{\mathrm{T=1}}, \mathbb{P}_\psi^{\mathrm{T=0}}\right) &\le \hat{\mathbb{W}}_\psi + \sqrt{2\log\left(\frac{1}{\delta}\right)/\lambda' N_1} + \sqrt{2\log\left(\frac{1}{\delta}\right)/\lambda' N_0}\\ &= \hat{\mathbb{W}}_\psi + \sqrt{2\log\left(\frac{1}{\delta}\right)/\lambda' N}\left(\sqrt{\frac{1}{\theta}} + \sqrt{\frac{1}{1-\theta}}\right)\\ &:= \hat{\mathbb{W}}_\psi + \mathcal{O}(\frac{1}{\delta N}),\end{aligned} \tag{32}$$

that holds with the probability at least $(1-\delta)^2$. $\mathcal{O}(\cdot)$ satisfies

$$\sqrt{\log\left(\frac{1}{\delta}\right)/\lambda'}\left(1 + \sqrt{1/(N-1)}\right) \ge \mathcal{O}(\frac{1}{\delta N}) \ge 4\sqrt{\log\left(\frac{1}{\delta}\right)/\lambda' N}, \tag{33}$$

where $\mathcal{O}(\frac{1}{\delta N})$ reaches its maximum when $\theta = 1/N$ or $\theta = 1 - 1/N$, reaches its minimum when $\theta = 0.5$. This corollary can be derived by differentiating the function $f(x) = 1/\sqrt{x} + 1/\sqrt{1-x}$.

Combining (29) and (32), with the probability at least $(1-\delta)^2$, we have

$$\epsilon_{\mathrm{PEHE}}(\psi,\phi) \le 2\left[\epsilon_{\mathrm{F}}^{\mathrm{T=1}}(\psi,\phi) + \epsilon_{\mathrm{F}}^{\mathrm{T=0}}(\psi,\phi) + B_\psi \hat{\mathbb{W}}_\psi - 2\sigma_Y^2 + \mathcal{O}(\frac{1}{\delta N})\right], \tag{34}$$

where we denote $B_\psi \mathcal{O}(\frac{1}{\delta N})$ as $\mathcal{O}(\frac{1}{\delta N})$. Finally, it is straightforward to derive the probabilistic approximately correct format that holds with probability at least $(1-\delta')$ by setting $\delta = 1 - \sqrt{1-\delta'}$, and the proof is completed. ☐

Theorem A.2 extends Theorem A.1 and derives the upper bound of PEHE in the stochastic batch form, which demonstrates that the PEHE can be optimized by iteratively minimizing the factual outcome estimation error and the optimal transport discrepancy *at a mini-batch level*.

Corollary A.1. *The empirical variance of the PEHE estimates in* (27) *largely depends on the batch size and the proportion of treated and untreated units. Large batch size and balanced proportion correspond to low empirical variance, and vice versa.*

Proof. It can be drawn directly from (27) (batch size) and (33) (treatment proportion). ☐

Corollary A.2. *For discrete measures* $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{\mathbf{x}_i}$ *and* $\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{\mathbf{x}_j}$, *adding an outlier* $\delta_{\mathbf{x}'}$ *to* α *and denote the disturbed distribution as* α', *we have*

$$\mathbb{W}^{0,\kappa}(\alpha', \beta) - \mathbb{W}^{0,\kappa}(\alpha, \beta) \leq 2\kappa(1 - e^{-\sum_{j=1}^{m}(\mathbf{x}'-\mathbf{x}_j)^2/2\kappa})/(n+1), \tag{35}$$

which is upper bounded by $2\kappa/(n+1)$. $\mathbb{W}^{0,\kappa}$ *is the unbalanced discrepancy as per Definition* 3.2.

Proof. This is a direct extension to the Lemma 1 by Fatras et al. [23], under the assumption that all the units including the outlier $\delta_{\mathbf{x}'}$ share the same mass (*i.e.*, uniform mass distribution in each group). Specifically, when adding an outlier to α and obtaining a disturbed measure α', the mass of each unit in α' is $1/(n+1)$ (the OT problem would normalize the mass of units, *i.e.*, the total mass of the measure equals to 1). Based on this assumption, we set the ζ in the Lemma 1 by Fatras et al. [23] to $n/(n+1)$ and derived the Equation (35) with the denominator being $(n+1)$. ☐

B Discrete Optimal Transport

This section proposes the definitions and algorithms to calculate optimal transport between discrete measures. We have omitted the case of general measures [49] since it is beyond the scope of this work. Instead, we provide an equivalent interpretation under discrete measures. Readers interested in this topic should refer to [19, 55] for details.

B.1 Problem Formulation

Consider n warehouses and m factories, where the i-th warehouse contains $\mathbf{a}_i$ units of materials; the j-th factory needs $\mathbf{b}_j$ units of materials [55]. Now we construct a *mapping* from warehouses to factories, satisfying: (1) all materials of warehouses are transported; (2) all requirements of factories are satisfied; (3) materials from one warehouse are transported to *no more than one* factory (mapping constraint). Every feasible mapping is associated with a *global* cost, calculated by aggregating the *local* cost of moving a unit of material from the i-th warehouse to the j-th factory. Our objective, to find a feasible mapping that minimizes the transport cost, is formulated in Definition B.1.

Definition B.1. *For discrete measures* $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{\mathbf{x}_i}$ *and* $\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{\mathbf{x}_j}$, *the Monge problem seeks for a mapping* $\mathbb{T} : \{\mathbf{x}_i\}_{i=1}^{n} \rightarrow \{\mathbf{x}_j\}_{j=1}^{m}$ *that associates to each point* $\mathbf{x}_i$ *a single point* $\mathbf{x}_j$ *and pushes the mass of* α *to* β. *That is,* $\forall j \in \{1, \ldots, m\}$ *we have* $\mathbf{b}_j = \sum_{i:\mathbb{T}(\mathbf{x}_i)=\mathbf{x}_j} \mathbf{a}_i$. *This mass-preserving constraint is abbreviated as* $\mathbb{T}_\sharp \alpha = \beta$. *The mapping should also minimize the transportation cost denoted as* $c(x, y)$. *To this end, Monge problem for discrete measures is formulated as:*

$$\min_{\mathbb{T}:\mathbb{T}_\sharp \alpha = \beta} \left\{ \sum_i c(\mathbf{x}_i, \mathbb{T}(\mathbf{x}_i)) \right\}. \tag{36}$$

This problem was further utilized to compare two probability measures where $\sum_i \mathbf{a}_i = \sum_j \mathbf{b}_j = 1$. However, Monge's formulation cannot guarantee the existence and uniqueness of solutions [55]. [34] relaxed the mapping constraint by allowing the transport from one warehouse to many factories and reformulated the Monge problem as a linear programming problem in Definition B.2.

Algorithm 1 Sinkhorn Algorithm

Input: discrete measures $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{\mathbf{x}_i}$ and $\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{\mathbf{x}_j}$, distance matrix $\mathbf{D}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$.
Parameter: ϵ: strength of entropic regularization; $\ell_{\max}$: maximum iterations.
Output: $\boldsymbol{\pi}^{\epsilon}$: the entropic regularized optimal transport matrix.

```
$\mathbf{K} \leftarrow \exp(-\mathbf{D}/\epsilon)$
$\mathbf{u} \leftarrow \mathbf{1}_n, \mathbf{v} \leftarrow \mathbf{1}_m, \ell \leftarrow 1$
while $\ell < \ell_{\max}$ do
    $\mathbf{u} \leftarrow \mathbf{a}/(\mathbf{K}\mathbf{v})$
    $\mathbf{v} \leftarrow \mathbf{b}/(\mathbf{K}^{\mathrm{T}}\mathbf{u})$
    $\ell \leftarrow \ell + 1$
$\boldsymbol{\pi}^{\epsilon} \leftarrow \operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v})$
```

Definition B.2. *For discrete measures $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{\mathbf{x}_i}$ and $\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{\mathbf{x}_j}$, the Kantorovich problem aims to find a feasible plan $\pi \in \mathbb{R}_{+}^{n \times m}$ which transports α to β at minimum cost:*

$$\mathbb{W}(\alpha, \beta) := \min_{\boldsymbol{\pi} \in \Pi(\alpha, \beta)} \langle \mathbf{D}, \boldsymbol{\pi} \rangle, \quad \Pi(\alpha, \beta) := \left\{ \boldsymbol{\pi} \in \mathbb{R}_{+}^{n \times m} : \boldsymbol{\pi} \mathbf{1}_m = \mathbf{a}, \boldsymbol{\pi}^{\mathrm{T}} \mathbf{1}_n = \mathbf{b} \right\}, \tag{37}$$

where $\mathbb{W}(\alpha, \beta) \in \mathbb{R}$ is the Wasserstein discrepancy between α and β; $\mathbf{D} \in \mathbb{R}_{+}^{n \times m}$ is the unit-wise distance[5] *between α and β; $\mathbf{a}$ and $\mathbf{b}$ indicate the mass of units in α and β, and Π is the feasible transportation plan set which ensures the mass-preserving constraint holds.*

B.2 Sinkhorn Discrepancy and Algorithm

Exact solutions to the Kantorovich problem suffer from great computational costs. The interior-point and network-simplex methods, for example, have a complexity of $\mathcal{O}(n^3 \log n)$ [54]. A shortcut is to add an entropic regularizer as

$$\mathbb{W}^{\epsilon}(\alpha, \beta) := \langle \mathbf{D}, \boldsymbol{\pi}^{\epsilon} \rangle, \quad \boldsymbol{\pi}^{\epsilon} := \underset{\boldsymbol{\pi} \in \Pi(\alpha, \beta)}{\arg\min} \langle \mathbf{D}, \boldsymbol{\pi} \rangle - \epsilon \mathrm{H}(\boldsymbol{\pi}), \quad \mathrm{H}(\boldsymbol{\pi}) := -\sum_{i,j} \boldsymbol{\pi}_{ij} \left(\log(\boldsymbol{\pi}_{ij}) - 1 \right), \tag{38}$$

which makes the problem ϵ-convex and solvable with the Sinkhorn algorithm [19], with a lower complexity of $\mathcal{O}(n^2/\epsilon^2)$. Besides, the Sinkhorn algorithm consists of matrix-vector products only, which makes it suited to be accelerated with GPUs. Specifically, let $\mathbf{f} \in \mathbb{R}^n$ and $\mathbf{g} \in \mathbb{R}^m$ be the lagrangian multipliers, the Lagrangian of (38) is:

$$\Phi(\boldsymbol{\pi}, \mathbf{f}, \mathbf{g}) = \langle \mathbf{D}, \boldsymbol{\pi} \rangle - \epsilon \mathrm{H}(\boldsymbol{\pi}) - \langle \mathbf{f}, \boldsymbol{\pi} \mathbf{1}_n - \mathbf{a} \rangle - \left\langle \mathbf{g}, \boldsymbol{\pi}^{\mathrm{T}} \mathbf{1}_m - \mathbf{b} \right\rangle \tag{39}$$

According to the first-order condition of constraint optimization problem, we have:

$$\frac{\partial \Phi(\boldsymbol{\pi}, \mathbf{f}, \mathbf{g})}{\partial \boldsymbol{\pi}_{ij}} = \mathbf{D}_{ij} + \varepsilon \log(\boldsymbol{\pi}_{ij}) - \mathbf{f}_i - \mathbf{g}_j = 0, \tag{40}$$

or equivalently, the best transport matrix $\boldsymbol{\pi}^{\epsilon}$ should satisfy:

$$\boldsymbol{\pi}_{ij}^{\epsilon} = \exp\left(\frac{\mathbf{f}_i}{\epsilon}\right) * \exp\left(-\frac{\mathbf{D}_{ij}}{\epsilon}\right) * \exp\left(\frac{\mathbf{g}_j}{\epsilon}\right). \tag{41}$$

Let $\mathbf{u}_i := \exp(\mathbf{f}_i/\epsilon)$, $\mathbf{v}_j := \exp(\mathbf{g}_j/\epsilon)$, $\mathbf{K}_{ij} := \exp(-\mathbf{D}_{ij}/\epsilon)$, then we have $\boldsymbol{\pi}^{\epsilon} = \operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v})$. The transport matrix should also satisfy the mass-preserving constraint, such that:

$$\operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v})\mathbf{1}_m = \mathbf{a}, \qquad \operatorname{diag}(\mathbf{v})\mathbf{K}^{\top}\operatorname{diag}(\mathbf{u})\mathbf{1}_n = \mathbf{b}, \tag{42}$$

or equivalently, let $\odot$ be the entry-wise multiplication of vectors, we have:

$$\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a} \quad \text{and} \quad \mathbf{v} \odot \left(\mathbf{K}^{\mathrm{T}}\mathbf{u}\right) = \mathbf{b}. \tag{43}$$

(43) is known as the matrix scaling problem. An intuitive approach is to solve them iteratively:

$$\mathbf{u}^{(\ell+1)} = \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(\ell)}} \quad \text{and} \quad \mathbf{v}^{(\ell+1)} = \frac{\mathbf{b}}{\mathbf{K}^{\mathrm{T}}\mathbf{u}^{(\ell+1)}} \tag{44}$$

which is the critical step of Sinkhorn algorithm in Algorithm 1. The optimal transport matrix $\boldsymbol{\pi}^{\epsilon}$ acting as a constant matrix further induces the *Sinkhorn discrepancy* $\mathbb{W}^{\epsilon}$ following (38). As $\mathbf{D}$ is differentiable to α and β, it is feasible to minimize $\mathbb{W}^{\epsilon}$ by adjusting the generation process of α and β, *i.e.*, the representation mapping in Definition A.2 with gradient-based optimizers.

---

[5] In this work, we calculate the unit-wise distance with the squared Euclidean metric following [17].

Algorithm 2 Generalized Sinkhorn Algorithm for Unbalanced Optimal Transport

Input: discrete measures $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{\mathbf{x}_i}$ and $\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{\mathbf{x}_j}$, distance matrix $\mathbf{D}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$.
Parameter: ϵ: strength of entropic regularizer; κ: strength of mass preserving; $\ell_{\max}$: max iterations.
Output: $\boldsymbol{\pi}^{\epsilon,\kappa}$: the entropic regularized unbalanced optimal transport matrix.

```
$\mathbf{K} \leftarrow \exp(-\mathbf{D}/\epsilon)$.
$\mathbf{f} \leftarrow \mathbf{0}_n$, $\mathbf{g} \leftarrow \mathbf{0}_m$, $\ell \leftarrow 1$.
while $\ell < \ell_{\max}$ do
    $\mathbf{u} \leftarrow \exp(\mathbf{f}_i/\epsilon)$, $\mathbf{v} \leftarrow \exp(\mathbf{g}_j/\epsilon)$
    $\boldsymbol{\pi} \leftarrow \mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v})$.
    $\mathbf{a}' \leftarrow \boldsymbol{\pi}\mathbf{1}_n$, $\mathbf{b}' \leftarrow \boldsymbol{\pi}^{\mathrm{T}}\mathbf{1}_m$.
    if $\ell//2 = 0$ then
        $\mathbf{f} \leftarrow \left[\frac{\mathbf{f}}{\epsilon} + \log(\mathbf{a}) - \log(\mathbf{a}')\right]\frac{\epsilon\kappa}{\epsilon+\kappa}$
    else
        $\mathbf{g} \leftarrow \left[\frac{\mathbf{g}}{\epsilon} + \log(\mathbf{b}) - \log(\mathbf{b}')\right]\frac{\epsilon\kappa}{\epsilon+\kappa}$
    $\ell \leftarrow \ell + 1$.
$\boldsymbol{\pi}^{\epsilon,\kappa} \leftarrow \mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v})$.
```

B.3 Unbalanced optimal transport and generalized sinkhorn

We have reported the mini-batch sampling effect (MSE) issue of $\mathbb{W}^{\epsilon}$ in Section 3.2, and attributed it to the mass-preserving constraint in (38). An intuitive approach to mitigate MSE is to relax the marginal constraint and allow for the creation and destruction of mass. To this end, RMPR is proposed in Definition B.3, which replaces the hard marginal constraint with a soft penalty.

Definition B.3. *For empirical distributions α and β with n and m units, respectively, unbalanced optimal transport seeks a transport plan at minimum cost:*

$$\mathbb{W}^{\epsilon,\kappa}(\alpha,\beta) := \min_{\boldsymbol{\pi}} \langle \mathbf{D}, \boldsymbol{\pi} \rangle, \boldsymbol{\pi} := \arg\min_{\boldsymbol{\pi}} \langle \mathbf{D}, \boldsymbol{\pi} \rangle + \epsilon \mathrm{H}(\boldsymbol{\pi}) + \kappa(\mathbf{KL}(\boldsymbol{\pi}\mathbf{1}_n, \mathbf{a}) + \mathbf{KL}(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{1}_m, \mathbf{b})), \tag{45}$$

where $\mathbf{D} \in \mathbb{R}_+^{n\times m}$ is the unit-wise distance, and $\mathbf{a}$ and $\mathbf{b}$ indicate the mass of units in α and β.

The unbalanced optimal transport problem in Definition B.3 has a similar structure with (38) and thus can be solved with a generalized Sinkhorn algorithm [14]. The derivation starts from the Fenchel-Legendre dual form of (45):

$$\begin{aligned}
\max_{\mathbf{f}\in\mathbb{R}^n, \mathbf{g}\in\mathbb{R}^m} &- F^*(-\mathbf{f}) - G^*(-\mathbf{g}) - \epsilon \sum_{i,j} \exp\left(\frac{\mathbf{f}_i + \mathbf{g}_j - \mathbf{D}_{ij}}{\epsilon}\right), \\
F^*(\mathbf{f}) &= \max_{\mathbf{z}\in\mathbb{R}^n} \mathbf{z}^{\top}\mathbf{f} - \kappa\mathbf{KL}(\mathbf{z}\|\mathbf{a}) = \kappa\left\langle e^{\mathbf{f}/\kappa}, \mathbf{a}\right\rangle - \mathbf{a}^{\top}\mathbf{1}_n, \\
G^*(\mathbf{g}) &= \max_{\mathbf{z}\in\mathbb{R}^m} \mathbf{z}^{\top}\mathbf{g} - \kappa\mathbf{KL}(\mathbf{z}\|\mathbf{b}) = \kappa\left\langle e^{\mathbf{g}/\kappa}, \mathbf{b}\right\rangle - \mathbf{b}^{\top}\mathbf{1}_m,
\end{aligned} \tag{46}$$

where the functions $F^*(\cdot)$ and $G^*(\cdot)$ are the Legendre transformation of KL divergence. Ignoring the constant terms, we can obtain the equivalent optimization problem:

$$\min_{\mathbf{f}\in\mathbb{R}^n, \mathbf{g}\in\mathbb{R}^m} \epsilon \sum_{i,j=1}^{n} \exp\left(\frac{\mathbf{f}_i + \mathbf{g}_j - \mathbf{D}_{ij}}{\epsilon}\right) + \kappa\left\langle e^{-\mathbf{f}/\kappa}, \mathbf{a}\right\rangle + \kappa\left\langle e^{-\mathbf{g}/\kappa}, \mathbf{b}\right\rangle. \tag{47}$$

According to the first-order condition, the minimizer's gradient of (47) should be zero. As such, fixing $\mathbf{g}^{\ell}$, the updated $\mathbf{f}^{\ell+1}$ ought to satisfy:

$$\exp\left(\frac{\mathbf{f}_i^{\ell+1}}{\epsilon}\right) \sum_{j=1}^{n} \exp\left(\frac{\mathbf{g}_j^{\ell} - \mathbf{D}_{ij}}{\epsilon}\right) = \exp\left(-\frac{\mathbf{f}_i^{\ell+1}}{\kappa}\right)\mathbf{a}_i, \tag{48}$$

We further multiply both sides by $\exp(\mathbf{f}_i^{\ell}/\epsilon)$:

$$\exp\left(\frac{\mathbf{f}_i^{\ell+1}}{\epsilon}\right)\mathbf{a}_i' = \exp\left(\frac{\mathbf{f}_i^{\ell}}{\epsilon}\right)\exp\left(-\frac{\mathbf{f}_i^{\ell+1}}{\kappa}\right)\mathbf{a}_i \tag{49}$$

Algorithm 3 ESCFR Algorithm

```
Input: covariates of treated units $\{\mathbf{x}_i\}_{i=1}^{n}$ and untreated units $\{\mathbf{x}_j\}_{j=1}^{m}$; factual outcomes $\{y_i\}_{i=1}^{n}$ and $\{y_j\}_{j=1}^{m}$; representation mapping $\psi$; outcome mapping $\phi$.
Parameter: $\lambda$: strength of optimal transport; $\epsilon$: strength of entropic regularizer; $\kappa$: strength of RMPR; $\gamma$: strength of PFOR; $\ell_{\max}$: max iterations
Output: $\mathcal{L}_{\mathrm{ESCFR}}^{\epsilon,\kappa,\gamma,\lambda}$: the learning objective of ESCFR.
1: $\{\mathbf{r}_i\}_{i=1}^{n} \leftarrow \{\psi(\mathbf{x}_i)\}_{i=1}^{n}, \quad \{\mathbf{r}_j\}_{j=1}^{m} \leftarrow \{\psi(\mathbf{x}_j)\}_{j=1}^{m}$.
2: $\{\hat{y}_i\}_{i=1}^{n} \leftarrow \{\phi(\mathbf{r}_i, 1)\}_{i=1}^{n}, \quad \{\hat{y}_j\}_{j=1}^{m} \leftarrow \{\phi(\mathbf{r}_j, 0)\}_{j=1}^{m}$.
3: $\{\tilde{y}_i\}_{i=1}^{n} \leftarrow \{\phi(\mathbf{r}_i, 0)\}_{i=1}^{n}, \quad \{\tilde{y}_j\}_{j=1}^{m} \leftarrow \{\phi(\mathbf{r}_j, 1)\}_{j=1}^{m}$.
4: $\mathbf{D}_{ij}^{\gamma} \leftarrow \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 + \gamma \cdot \|y_i - \tilde{y}_j\|_2^2 + \gamma \cdot \|y_j - \tilde{y}_i\|_2^2$.
5: $\mathbf{D}_{\mathrm{stop}}^{\gamma} \leftarrow \mathrm{stopgradient}(\mathbf{D}^{\gamma})$.
6: $\boldsymbol{\pi}^{\epsilon,\kappa,\gamma} \leftarrow \mathrm{Algorithm2}\left(\alpha = \{\mathbf{r}_i\}_{i=1}^{n}, \beta = \{\mathbf{r}_j\}_{j=1}^{m}, \mathbf{D} = \mathbf{D}_{\mathrm{stop}}^{\gamma}\right)$.
7: $\mathcal{L}_{\mathrm{F}}(\psi, \phi) \leftarrow \frac{1}{n}\sum_{i=1}^{n} \|\hat{y}_i - y_i\|_2^2 + \frac{1}{m}\sum_{j=1}^{m} \|\hat{y}_j - y_j\|_2^2$.
8: $\mathcal{L}_{\mathrm{D}}^{\epsilon,\kappa,\gamma}(\psi) \leftarrow \langle \mathbf{D}^{\gamma}, \boldsymbol{\pi}^{\epsilon,\kappa,\gamma} \rangle$.
9: $\mathcal{L}_{\mathrm{ESCFR}}^{\epsilon,\kappa,\gamma,\lambda} \leftarrow \mathcal{L}_{\mathrm{F}}(\psi, \phi) + \lambda \cdot \mathcal{L}_{\mathrm{D}}^{\epsilon,\kappa,\gamma}(\psi)$.
```

Table 3: Running time (mean+std) in seconds of Algorithm 1-2 with 100 runs.

Parameter	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$
Algorithm1	0.0266±0.0102	0.0241±0.0075	0.0326±0.0088	0.0499±0.0099	0.0725±0.0128	0.1430±0.0259
Algorithm2	0.0050±0.0004	0.0051±0.0001	0.0065±0.0002	0.0104±0.0005	0.0138±0.0008	0.0256±0.0007
Parameter	$\epsilon = 0.1$	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 5.0$	$\epsilon = 10.0$	$\epsilon = 100.0$
Algorithm1	0.1683±0.0038	0.1207±0.0102	0.0699±0.0095	0.0153±0.0013	0.0097±0.0009	0.0072±0.0009
Algorithm2	0.0166±0.0019	0.0068±0.0010	0.0052±0.0011	0.0047±0.0010	0.0045±0.0011	0.0043±0.0009
Parameter	$\kappa = 0.1$	$\kappa = 0.5$	$\kappa = 1.0$	$\kappa = 5.0$	$\kappa = 10.0$	$\kappa = 100.0$
Algorithm2	0.0050±0.0011	0.0059±0.0008	0.0060±0.0011	0.0112±0.0014	0.0162±0.0016	0.1039±0.0033

where $\mathbf{a}' := \boldsymbol{\pi}\mathbf{1}_n$ with $\boldsymbol{\pi}_{ij} := \exp(\mathbf{f}_i^{\ell} + \mathbf{g}_j^{\ell} - \mathbf{D}_{ij})$. Similarly, fixing $\mathbf{f}$ we have $\mathbf{g}^{\ell+1}$ as:

$$\exp\left(\frac{\mathbf{g}_j^{\ell+1}}{\epsilon}\right)\mathbf{b}_j' = \exp\left(\frac{\mathbf{g}_j^{\ell}}{\epsilon}\right)\exp\left(-\frac{\mathbf{g}_j^{\ell+1}}{\kappa}\right)\mathbf{b}_j \tag{50}$$

where $\mathbf{b}' := \boldsymbol{\pi}^{\mathrm{T}}\mathbf{1}_m$. (49) and (50) construct the critical iteration steps of the generalized Sinkhorn algorithm [14], which we formulate in Algorithm 2. The transport matrix $\boldsymbol{\pi}^{\epsilon,\kappa}$ further induces the *generalized Sinkhorn discrepancy* $\mathbb{W}^{\epsilon,\kappa}$ in Definition B.3. As $\mathbf{D}$ is differentiable with respect to α and β, it is feasible to minimize $\mathbb{W}^{\epsilon,\kappa}$ by adjusting the generation process of α and β, *i.e.*, the representation mapping in Definition A.2, with gradient-based optimizers.

B.4 Optimization of Entire Space Counterfactual Regression

Algorithm 3 shows how to calculate the learning objective at a mini-batch level. Specifically, we first calculate the factual outcome estimates (step 2), counterfactual outcome estimates (step 3), and the unit-wise distance matrix with PFOR (step 4). Afterwards, fix the gradient of the distance matrix (step 5) and calculate the transport matrix with Algorithm 2 (step 6). Finally, calculate the factual outcome estimation error (step 7) and distribution discrepancy (step 8), and aggregate them to acquire the learning objective of ESCFR (step 9). According to Section B.3, the learning objective is differentiable to ψ and ϕ and thus can be optimized end-to-end with stochastic gradient methods.

B.5 Complexity Analysis

One primary concern would be the overall complexity of solving discrete optimal transport problems. Exact algorithms, *e.g.*, the interior-point method and network-simplex method, suffer from a high computational cost of $\mathcal{O}(n^3 \log n)$ [54]. An entropic regularizer is thus introduced in (5), making the problem solvable by the Sinkhorn algorithm [19] in Algorithm 1. The complexity was shown to be $\mathcal{O}(n^2/\epsilon^3)$ by [1] in terms of the absolute error of the mass-preservation constraints. [22]

improved it to $\mathcal{O}(n^2/\epsilon^2)$, which can be further accelerated with greedy algorithm by [47]. Several recent explorations [3, 29] have also attempted to further reduce the complexity to $\mathcal{O}(n^2/\epsilon)$.

Entropic regularization trick is still applicable to speed up the solution of the unbalanced optimal transport problem in RMPR, represented by the Sinkhorn-like algorithm in Algorithm 2. [56] further proved that the complexity of Algorithm 2 is $\tilde{\mathcal{O}}(n^2/\epsilon)$.

Table 3 reports the practical running time at the commonly-used batch settings. In general, the computational cost of optimal transmission is not a concern at the mini-batch level. Notice that enlarging ϵ speeds up the computation while making the resulting transfer matrix biased, hindering the transportation performance, as per Figure 5. In addition, a large relaxation parameter κ makes the computed results closer to those by Sinkhorn algorithm yet significantly contributes to more iterations, which is discussed and mitigated by [66].

C Reproduction Details

C.1 Datasets

We conduct experiments on two semi-synthetic benchmarks to validate our models. For the IHDP[6] benchmark, we report the results over 10 simulation realizations following [85]. However, the limited size (747 observations and 25 covariates) makes the results highly volatile. As such, we mainly validate the models on the ACIC benchmark, which was released by the ACIC-2016 competition[7].

All datasets are randomly shuffled and partitioned in a 0.7:0.15:0.15 ratio for training, validation, and test, where we maintain the same ratio of treated units in all three splits to avoid numerical unreliability in the validation and test phases. We find that these datasets are overly easy to fit by the model because they are semi-synthetic. To increase the distinguishability of the results, we omit preprocessing strategies, such as min-max scaling, to increase the difficulty of the learning task.

C.2 Baselines

The collection of baselines involves statistical estimators [36, 67], matching estimators [18, 60, 73] and representation-based estimators [31, 67]. We implement these baselines based on Pytorch for neural network models, Sklearn for statistical models, and EconML for tree and forest models.

D Additional Discussions

D.1 Additional Discussion for Stochastic Optimal Transport

According to Theorem 3.1, one critical hyperparameter for CFR-WASS and ESCFR is the batch size, which directly affects the variance of stochastic optimal transport in Section 3.1 and thus the performance of both methods. As such, it is necessary to verify whether ESCFR outperforms CFR for different batch sizes. We conduct extensive experiments and summarize the results in Table 4. Interesting observations are noted:

- Increasing batch size in a wide range improves the performance of CFR-WASS and ESCFR. For example, The PEHE of CFR-WASS decreases from 3.114 at b = 32 to 2.932 at b = 128, and the PEHE of ESCFR exhibits a similar pattern. The performance gain is attributed to the decreased variance in (6), which backs up Theorem 3.1.
- By finetuning batch size, we can easily exceed the performance we report in Table 1. However, we did not finetune it as the PEHE is invisible during our hyper-parameter tuning process[8].
- The performance drop given overly large batch sizes comes from the sub-optimal backbone (TARNet) performance. Due to the limited training samples, *e.g.*, 4.8k * 70% units for ACIC and 0.7k * 70% units for IHDP, a large batch size might impede the optimizer from escaping saddle points [30] and sharp minima [82], thus deteriorating the quality of factual outcome estimation.

---

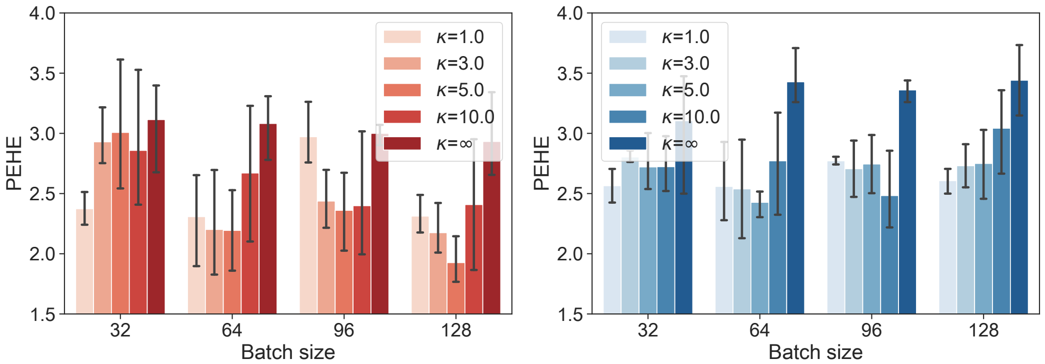
[6] It can be downloaded from https://www.fredjo.com/

[7] It can be downloaded from https://jenniferhill7.wixsite.com/acic-2016/competition

[8] Most of the experiments in Table 1 were performed with a fixed batch size b = 32, which is selected by the factual estimation performance of TARNet.

Table 4: Out-of-sample PEHE of ESCFR and important baselines with different batch sizes b.

Model	b = 32	b = 64	b = 96	b = 128	b = 196	b = 256
TARNet	3.3293±0.1853	3.2054±0.2676	3.0869±0.2812	2.9262±0.3160	3.4619±0.6652	3.6309±0.2026
CFR-WASS	3.1143±0.4578	3.0819±0.3407	2.9998±0.1017	2.9326±0.4142	4.0740±1.4127	3.4675±0.1552
ESCFR	2.3736±0.1621	2.3082±0.4334	2.9719±0.2889	2.3125±0.1836	2.0373±0.1538	2.2777±0.4230



(a) Out-of-sample Performance (b) In-sample Performance

Figure 7: PEHE of ESCFR and CFR-WASS ($\kappa = \infty$) under different batch size.

D.2 Additional Discussion for Relaxed Mass-Preserving Regularizer

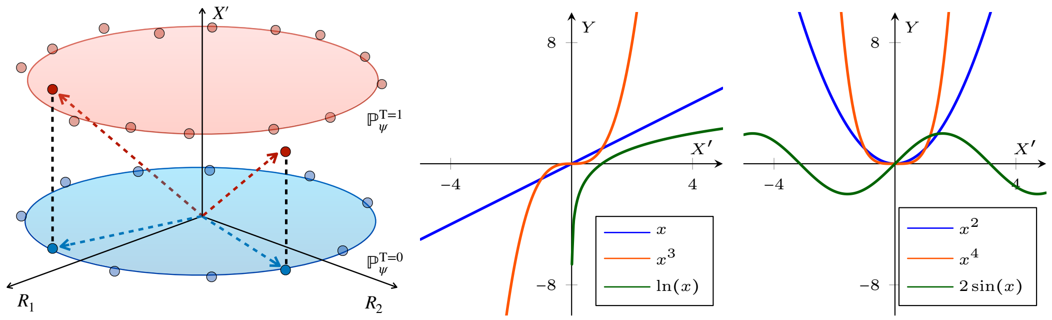
Existing methods [31, 67, 85] suffer from the mini-batch sampling effect (MSE) issue, as indicated by the two bad cases in Figure 2. RMPR mitigates the MSE issue by relaxing the mass-preserving constraint, the performance of which is affected by two critical hyperparameters, *i.e.*, the batch size b and the strength of mass-preserving constraint κ. On top of the ablation studies, it is necessary to explore the performance of ESCFR at different settings of b and κ, to investigate 1) how RMPR works; 2) the limitation and bottleneck of RMPR; 3) the robustness of RMPR to hyperparameter setting. The results are presented in Figure 7, and the observations are summarized as follows.

- The optimal value of κ increases with the increase of batch size. For example, the optimal κ is 1.0 at $b = 32$, and 5.0 at $b = 128$. This observation partially verifies how RMPR works as described in Section 3.2. Specifically, at small batch sizes where sampling outliers dominate the sampled batches, a small κ effectively relaxes the mass-preserving constraint and avoids the damage of mini-batch outliers, thus improving the performance effectively and robustly. At large batch sizes, the noise of sampling outliers is reduced, and it is reasonable to increase κ to match more units and obtain more accurate wasserstein distance estimates.
- Even with large batch sizes, oversized κ, *e.g.*, $\kappa \geq 10$ does not perform well. Although the effect of sampling outliers is reduced, some patterns such as outcome imbalance are present for all batch sizes, resulting in false matching given large mass-preserving constraint strength κ, which might be a primary bottleneck of RMPR.
- Hyper-parameter tuning is not necessarily the reason why ESCFR works well, since all ESCFR implementations outperform the strongest baseline CFR-WASS ($\kappa = \infty$) on all batch sizes, most of which are statistically significant. This can be further supported by our extensive ablation study in Section 4.3 and parameter study in Section 4.5.

In summary, it is necessary to relax the mass-preserving constraint under all settings of batch size, which strongly verifies the effectiveness of RMPR in Section 3.2.

D.3 Additional Discussion for Proximal Factual Outcome Regularizer

Existing representation-based methods block the backdoor path $X \rightarrow T$ by balancing the distribution of the observed covariates in a latent space. Given the unconfoundedness assumption A.1, this approach effectively handles the treatment selection bias. However, Assumption A.1 is usually



(a) Toy example (b) Applicable cases (c) Bad cases

Figure 8: A diagram showing how PFOR works and its limitations. (a) A toy example of PFOR, where R and X' indicate the balanced representations and an unobserved confounder, respectively; scatters indicate the empirical distribution of units in the treated and control groups; for solid scatters with balanced R, the colored dashed line indicates the ground truth outcome $Y = \sqrt{R_1^2 + R_2^2 + X'^2}$ in each group, the black dashed line measures the difference of unobserved X'. (b) Cases that satisfy Assumption D.1, where the the outcome Y is monotone with unobserved X' given observed confounders in R. (c) Cases that violate Assumption D.1, where the Y is non-monotone with X'.

violated in practice, which invalidates this approach as the backdoor path from the unobserved confounder X' to T is not blocked.

According to the designed causal graph in Figure 3(b), all factors associated with outcome Y include the observed confounders X, treatment T, and unobserved confounders X'. Therefore, it is reasonable to derive that given balanced X and identical T, the only variable reflecting the variation of X' is the outcome Y. As such, inspired by the joint distribution transport technique [see 16], PFOR calibrates the unit-wise distance $\mathbf{D}$ with the potential outcomes in (12). The underlying regularization is: units with similar (observed and unobserved) confounders should have similar potential outcomes. Equivalently, for a pair of units with similar observed covariates, *i.e.*, $\|r_i - r_j\|^2 \approx 0$, if their potential outcomes under the same treatment $t = \{0, 1\}$ differ significantly, *i.e.*, $\|y_i^t - y_j^t\| >> 0$, their unobserved confounders should also differ significantly. As such, it is reasonable to utilize the difference of outcomes to calibrate the unobserved confounding effect.

Assumption D.1. *(Monotonicity). For all observed covariates $X = x$ in the population of interest, let $T = t$ and $X' = x'$ be the treatment assignment and unobserved confounders, respectively, we have $\mathbb{E}[Y \mid X = x, X' = x', T = t]$ is monotonically increasing or decreasing with respect to x'.*

Advantages. The advantages of PFOR can be further interpreted as follows.

- From a statistical perspective, PFOR encourages units with similar outcomes to share similar representations. It is a valid prior that inspires many learning algorithms, *e.g.*, K-nearest neighbors and gaussian process [see 78]. As an effective statistical regularizer, PFOR also works in the absence of unobserved confounders, especially on small data sets.
- From a domain adaptation perspective, vanilla Sinkhorn aligns the distributions $\mathbb{P}_{\psi}^{\mathrm{T}=1}(r)$ and $\mathbb{P}_{\psi}^{\mathrm{T}=0}(r)$, where r is the learned representations in Definition A.4. PFOR further aligns the transition probabilities $\mathbb{P}^{\mathrm{T}=1}(Y(T = t) \mid r)$ and $\mathbb{P}^{\mathrm{T}=0}(Y(T = t) \mid r)$ for $t = 0, 1$. The discrepancy between transition probabilities can be attributed to unobserved confounders that can be viewed as parameters of the transition probabilities [16]. As such, it is feasible to align the unobserved confounders by aligning the transition probabilities.

Toy example. Let the ground truth $Y := \sqrt{R_1^2 + R_2^2 + X'^2}$ where T is omitted as we only consider one group, R_1 and R_2 are the representations of observed confounders that have been aligned with Sinkhorn algorithm. Let the unobserved $X' = 0$ for controlled units and $X' = 1$ for treated units, which makes X' an unobserved confounder as it is related to Y and different between groups. As shown in Figure 8(a), given balanced R_1 and R_2, the variation of Y reveals that of X'. As such, it is reasonable to employ Y to calibrate the unit-wise distance $\mathbf{D}$ that ignores X'.

Synthetic labels. PFOR remains effective for semi-synthetic data, where the outcomes are synthetic from the covariates and treatment assignments. One source of hidden confounders in such data is information loss from the raw data space to the representation space, where not all valuable information (*e.g.*, confounders) is extracted and preserved, in particular when the representation mapping ψ is not invertible. Besides, this improvement could also come from the statistical regularization, encouraging units with similar outcomes to share similar representations, which is an effective prior according to the K-nearest neighboring methods and warrants further investigation in the context of treatment effect estimation.

Limitations. PFOR fails to handle confounders that add constant effects to all units. Specifically, for unobserved confounder X' and treatment assignment $t = 0, 1$, if $\mathbb{E}[Y \mid X, X' = x_1, T = t] = \mathbb{E}[Y \mid X, X' = x_2, T = t]$, PFOR fails to eliminate the confounding effect of X'. Examples can be found in Figure 8 (c). However, in real scenarios, it is rare that different values of X' only add a constant effect to the outcome [see 52, 70, 91], making PFOR still effective in a wide range of application scenarios.

This limitation is formalized as the Assumption D.1, where the outcome should increase or decrease monotonically with unobserved confounders given observed confounders and treatment assignment, as shown in Figure 8 (b). Notably, it is a commonly used assumption in confounder analysis [70, 91]. Besides, this assumption is often plausible, at least approximately, conditional on $T = t$ [91] . For example, it naturally holds for binary confounders; and generally holds in applications such as epidemiology [52]. Finally, this assumption is only imposed on the hidden confounder X' following [91], which further weakens Assumption D.1 significantly.

Further discussion. PFOR is mainly built upon the assumption of the causal graph in Figure 3(b), where the roles of the adjustment variables and the noise variables are excluded. Actually, it is a standard setting in many existing work of causal inference, such as Figure 1.1 in [62] and Figure 5 in [6]. Nevertheless, we would like to further discuss the applicability of PROR when there are noise variables that are parents only of Y. We provide the following analysis.

- If these variables are both observable and predictive of Y, they are known as adjustment variables. Aligning these variables does not introduce additional bias and can lead to a reduction in the variance of estimated treatment effects [25, 90]. Therefore, many studies [13, 27] do not differentiate them from confounders and align them together with confounders. Therefore, this is not an issue with ESCFR as these variables can be adjusted along with confounders.
- If these variables are non-observable and predictive of Y, we can rely on the monotonicity assumption to adjust for them using PFOR alongside unobserved confounders. This method does not introduce any additional bias and can still reduce the variance in the estimated treatment effects [25, 90].
- If these variables are pure noise, *i.e.*, non-predictive of Y, we believe they will interference the calculation of PFOR. Nevertheless, we mildly argue that the effect of this noise is not catastrophic, since such independent noise is also present in X and does not impede the success of canonical OT in fundamental domains such as computer vision and neural language processing.
- Finally, we find it would be interesting to discuss the robustness of the OT discrepancy to the volume of noise (in both X and Y), with the aim at devising a more robust OT discrepancy. There have been many attempts in this topic, including but not limited to UOT, Relaxed OT, semi-UOT, etc. These robust approaches could further mitigate the negative impact of noise and handle this piece of weakness. In future work, we will allocate some effort to exploring this interesting topic.

An important approach with unobserved confounders is the partial identification. Specifically, an estimand denoted as θ is partially identified if the observed data distribuion is compatible with multiple values of θ. In causal inference, challenges like unobserved confoundings might prevent precise causal effect pinpointing. A weaker alternative is to obtain a range of possible causal effects, known as a "identified set", and reduce the size of the set using proper assumptions. For example, given certain assumptions, *e.g.*, monotone treatment selection, we can narrow the bound of treatment effect estimate.

It is interesting to investigate the connection between PFOR from the partial identification view. We note that the transport strategy derived by the canonical Kantorovich problem in (4) is non-identifiable given the existence of (an) unobserved confounder X'. That is, assuming that X' has multiple candidate values, there should be multiple corresponding transport strategies, which makes the ground-truth transport strategy non-identifiable. Ideally, we can only identify a huge strategy set (by enumerating possible values of X'). Nevertheless, under the monotonic assumption between X' and outcomes, we can calibrate the unit-wise distance with the outcome differences, to reduce the size of strategy set and achieve more accurate estimation among possible transport strategies, which largely share the intuition of partial identification methodology. The partial identification method in causality from an OT view warrants further investigation.