Interpretable Almost-Exact Matching for Causal Inference
Supplementary Material

Awa Dieng, Yameng Liu, Sudeepa Roy, Cynthia Rudin, Alexander Volfovsky

Appendix

A Naïve AME solutions

In this section we present the complete outline of the two straightforward (but inefficient) solution to the AME problem (described in Section 3) for all units.

AME Solution 1 (quadratic in n, linear in p): For all treatment units t, we (i) iterate over all control units c, (ii) find the vector $\boldsymbol{\theta}_{tc} \in \{0,1\}^p$ where $\boldsymbol{\theta}_{tcj} = 1$ if t and c match on covariate j and 0 otherwise, (iii) find the control unit(s) with the highest value of $\boldsymbol{\theta}_{tc}^T\mathbf{w}$, and (iv) return them as the main matched group for the treatment unit t. Repeat the same procedure for each control unit c. Note that the CATE for each unit is computed based on its main matched group which means that the outcome of each unit can contribute to the computation of CATEs for multiple units. This algorithm is polynomial in both n and p, however, the quadratic time complexity in n also makes this approach impractical for large datasets (for instance, when we have more than a million units with half being treatment units).

AME Solution 2 (order $n \log n$, exponential in p:) This approach solves the AME problem simultaneously for all treatment and control units for a fixed weight vector $\mathbf{w}$. First, (i) enumerate every $\boldsymbol{\theta} \in \{0,1\}^p$ (which serves as an indicator for a subset of covariates), (ii) order the $\boldsymbol{\theta}$'s according to $\boldsymbol{\theta}^T\mathbf{w}$, (iii) call `GroupedMR` for every $\boldsymbol{\theta}$ in the predetermined order, (iv) the first time each unit is matched during a `GroupedMR` procedure, mark that unit with a 'done' flag, and record its corresponding main matched group and compute the CATE for each treatment and control unit using its main matched group. Each unit's outcome will be used to estimate CATEs for every auxiliary group that it is a member of, as before. Although this approach can use an efficient 'group by' function (e.g., an implementation using *bit-vectors* or *database/SQL queries* as discussed by Wang et al. (2017)), which can be implemented in $O(n \log n)$ time by sorting the units, iterating over all possible vectors $\boldsymbol{\theta} \in \{0,1\}^p$ makes this approach unsuitable for practical purposes (exponential in p).

B Proof of Proposition 4.1

Proposition 4.1 *If for a superset r of a newly processed set s where $|s| = k$ and $|r| = k+1$, all subsets s' of r of size k have been processed (i.e. r is eligible to be active after s is processed), then r is included in the set Z returned by* `GenerateNewActiveSets`*.*

Proof. Suppose all subsets of r of size k are already processed and belong to Δ^k. Let f be the covariate in $r \setminus s$. Clearly, f would appear in Δ^k, since at least one subset $s' \neq s$ of r of size k would contain f, and $s' \in \Delta^k$. Further all covariates in r, including f and those in s will have support at least k in Δ^k. To see this, note that there are $k+1$ subsets of r of size k, and each covariate in r appears in exactly k of them. Hence $f \in \Omega$, which the set of high support covariates. Further, the 'if' condition to check minimum support for all covariates in s is also satisfied. In addition, the final 'if' condition to eliminate false positives is satisfied too by assumption (that all subsets of r are already processed). Therefore r will be included in Z returned by the procedure. □

C Proof of Theorem 4.2

Theorem 4.2 ***(Correctness)*** *The* `DAME` *algorithm solves the AME problem.*

Proof. Consider any treatment unit t. Let s be the set of covariates in its main matched group returned in `DAME` (the while loop in `DAME` runs as long as there is a treated unit and the stopping criteria have not been met, and the `GroupedMR` returns the main matched group for every unit when it is matched for the first time). Let $\boldsymbol{\theta}_s$ be the indicator vector of s (see Eq. 1). Since the `GroupedMR` procedure returns a main matched group only if it is a *valid* matched group containing at least one treated and one control unit (see Algorithm 2), and since all units in the matched group on s have the same value of covariates in $\mathcal{J} \setminus s$, there exists a unit ℓ with $T_\ell = 0$ and $\mathbf{x}_\ell \circ \boldsymbol{\theta}_s = \mathbf{x}_t \circ \boldsymbol{\theta}_s$.

Hence it remains to show that the covariate set s in the main matched group for t corresponds to the maximum weight $\boldsymbol{\theta}^T\mathbf{w}$ over all $\boldsymbol{\theta}$ for which there is a valid matched group. Assume that there exists another covariate-set r such that $\boldsymbol{\theta}_r^T \mathbf{w} > \boldsymbol{\theta}_s^T \mathbf{w}$, there exists a unit ℓ' with $T_{\ell'} = 0$ and $\mathbf{x}_{\ell'} \circ \boldsymbol{\theta}_r = \mathbf{x}_t \circ \boldsymbol{\theta}_r$, and gives the maximum weight $\boldsymbol{\theta}_r^T \mathbf{w}$ over all such r. Then,

(i) r cannot be a (strict) subset of s, since `DAME` ensures that all subsets are processed before a superset is processed to satisfy the downward closure property in Proposition 3.1.

(ii) r cannot be a (strict) superset of s. Recall that $\theta_{s,j}$ is 1 for covariates j that are not in s (analogously for r). If r is a strict superset of s, then we would have $\boldsymbol{\theta}_r^T \mathbf{w} \leq \boldsymbol{\theta}_s^T \mathbf{w}$, which violates the assumption that $\boldsymbol{\theta}_r^T \mathbf{w} > \boldsymbol{\theta}_s^T \mathbf{w}$ for non-negative weights.

Given (i) and (ii), r and s must be incomparable (there exist covariates in both $r \setminus s$ and $s \setminus r$). Suppose the active set s was chosen in iteration h. If r was processed in an earlier iteration $h' < h$, since r forms a valid matched group for t, it would give the main matched group for t, violating the assumption that s was chosen by `DAME` to form the main matched group for t, rather than r.

Next, we argue that r must be active at the start of iteration h, and will be chosen as the best covariate set in iteration h, leading to a contradiction.

Note that we start with all singleton sets as active sets in $\Lambda_{(0)} = \{\{1\}, \cdots, \{p\}\}$ in the `DAME` algorithm. Consider any singleton subset $r_0 \subseteq r$ (comprising a single covariate in r). Due to the downward closure property in Proposition 3.1, $\boldsymbol{\theta}_{r_0}^T \mathbf{w} \geq \boldsymbol{\theta}_r^T \mathbf{w} > \boldsymbol{\theta}_s^T \mathbf{w}$. Hence all of the singleton subsets of r will be processed in earlier iterations $h' < h$, and will belong to the set of processed covariate sets $\Delta_{(h-1)}$.

Repeating the above argument, consider any subset $r' \subseteq r$. It holds that $\boldsymbol{\theta}_{r'}^T \mathbf{w} \geq \boldsymbol{\theta}_r^T \mathbf{w} > \boldsymbol{\theta}_s^T \mathbf{w}$. All subsets r' of r will be processed in earlier iterations $h' < h$ starting with the singleton subsets of r. In particular, all subsets of size $|r| - 1$ will belong to $\Delta_{(h-1)}$. As soon as the last of those subsets is processed, the procedure `GenerateNewActiveSets` will include r in the set of active sets in a previous iteration $h' < h$. Hence if r is not processed in an earlier iteration, it must be active at the start of iteration h, leading to a contradiction.

Hence for all treatment units t, the covariate-set r giving the maximum value of $\boldsymbol{\theta}_r^T \mathbf{w}$ will be used to form the main matched group of t, showing the correctness of the `DAME` algorithm. □

D Details of Breaking the Cycle of Drugs and Crime Study

D.1 Details About Survey

A survey was conducted in Alabama, Florida, and Washington regarding the program's effectiveness, with high quality data for over 380 individuals. These data (and this type of data generally) can be a powerful tool in the war against opioids, and our ability to draw interpretable, trustworthy conclusions from it depends on our ability to construct high-quality matches. For the survey, participants were chosen to receive screening shortly after arrest and participate in a drug intervention under supervision. Similar defendants before the start of the BTC program were selected as the control group. Features are listed in Table 2.

D.2 Order of Dropping Covariates

For both `DAME` and FLAME we used ridge regression as the machine learning method for the Full-AME problem, calculating variable importance as the difference in mean squared error before and after dropping the variable. The order in which `DAME` and FLAME process covariates could be different. Table 3 shows the order in which the dynamic versions of the two algorithms process the covariates. The first covariate that the two algorithms process is identical: "Have problem getting along with father in life" but the two diverge afterwards. At the second round, `DAME` processes the covariate "Have an automobile." On the other hand, at that same second round, FLAME processes "Have serious depression or anxiety in past 30 days", which now is dropped along with "Have problem getting along with father in life." What is important is that `DAME` is able to construct matched groups by only dropping subsets of what FLAME drops as early as the second and third iteration of the algorithm.

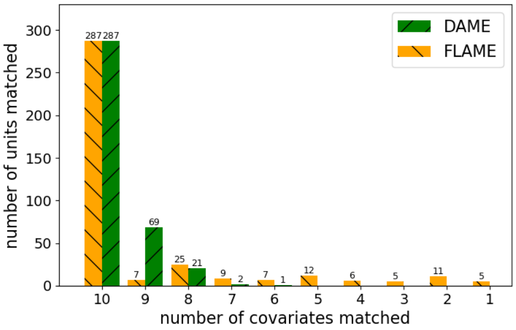

Figure 4: Number Matched: Number of units matched per covariates for the BTC data

Table 2: Features for BTC data.

Feature
1. Live with anyone with an alcohol problem
2. Have trouble understanding in life
3. Live with anyone using non prescription drugs
4. Have problem getting along with father in life
5. Have an automobile
6. Have drivers license
7. Have serious depression or anxiety in past 30 days
8. Have serious anxiety in life
9. SSI benefit last 6 months
10. Have serious depression in life

Table 3: Order in which features were processed for DAME and FLAME. The feature numbers correspond to the feature numbers in Table 2. The number in the parenthesis corresponds to the number of units matched for the first time at that round. Before any covariates are dropped, 287 individuals are matched on all features, which is 75% of the data.

	DAME	FLAME
1st	4: problem with father (15 new units matched)	4 (7 units)
2nd	5: have an automobile (9 units)	4,7 (25 units)
3rd	7: have serious depression (24 units)	4,7,9 (9 units)
4th	4,7 (3 units)	4,7,9,1 (7 units)
5th	5,7 (1 unit)	4,7,9,1,8 (12 units)
6th	4,5 (7 units)	4,7,9,1,8,10 (6 units)
7th	4,5,7 (0 units)	4,7,9,1,8,10,6 (5 units)
8th	9 (8 units)	4,7,9,1,8,10,6,5 (11 units)
9th	4,9 (0 units)	4,7,9,1,8,10,6,5,2 (5 units)
⋮		
196th	1,2,4,5 (1 unit)	

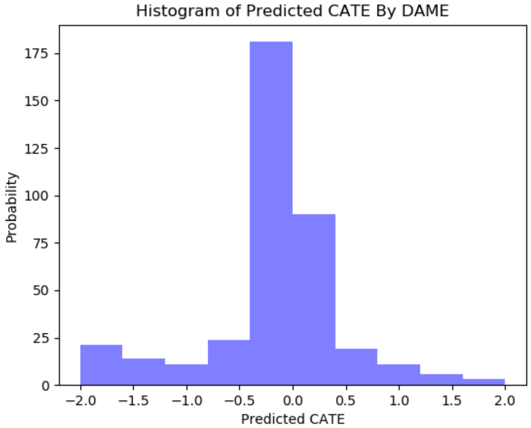

Figure 5: Histogram of estimated CATE by DAME. For individuals where the CATE is negative, it means that BTC was estimated to reduce crime.

D.3 Match Quality for FLAME and DAME

We compare the quality of matches in the BTC data between FLAME and DAME in terms of the number of covariates used to match within the groups. Many of the units matched exactly on all covariates and thus were matched by both algorithms at the first round. In fact 75% of the data are matched on all covariates. This is important, because exact matching alone yields the highest quality CATE estimates for most of the data; if we had used a classical propensity score matching technique, we may not have noticed this important aspect of the data.

For the remaining units that do not have exact matches on all covariates, DAME matches on more covariates than FLAME. In Figure 4 we see that DAME matched many more units on 9 out of the 10 variables than FLAME; FLAME cannot match the same data on so many variables.

D.4 CATEs from BTC analysis

We plot a histogram of the estimated CATEs for BTC in Figure 5. The program does not seem to provide uniform protection from future arrests, but does seem to protect some individuals. The majority of people

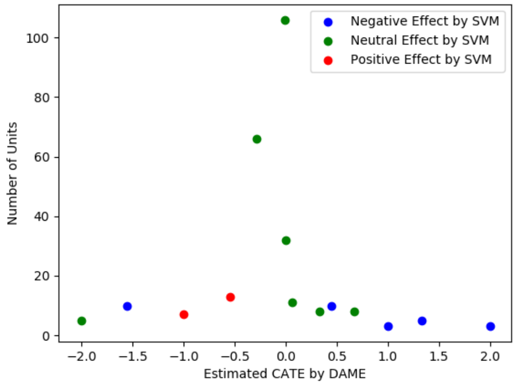

Figure 6: Comparison between DAME and SVM-based method

are estimated to experience little to no effect from the program.

D.5 A Comparison of DAME with SVM-Based Method Minimax Surrogate Loss

We can use DAME as a tool to check the performance of a black box machine learning approach. We chose a recent method that predicts whether treatment effects are positive, negative, or neutral, using a support vector machine formulation (Goh and Rudin, 2018). We ran DAME on the BTC dataset and saved the CATE for each treatment and control unit that were matched. Units with a positive CATE (outcome on treatment unit minus outcome on control unit) are considered to have a negative treatment effect, meaning that the program increased the probability of crime. Units with a negative CATE analogously had a positive predicted treatment effect. We also implemented the SVM approach and recorded a prediction of positive, negative, or neutral treatment effect for each unit. Figure 6 plots the CATEs for all the units that were matched exactly by DAME and colors them according to the output of the SVM. Since the distribution of some covariates is unbalanced, the number of matched groups is small with most units belonging to large groups.

Figure 6 shows that DAME and the SVM approach agree on the direction of the treatment effect for most of the matched units: Most positive CATEs corresponded to negative treatment effects from the SVM. Only two points have a mismatch between DAME and SVM: the left-most green (neutral) labeled and blue (negative) labeled points.

The easiest way to explain the discrepancy between the two methods is that DAME is a matching method, not a statistical model and so does not smooth CATEs. CATEs are sometimes computed using a very small number of units, so it is possible that the SVM simply smoothed out the treatment effect estimates so that there was a different predicted treatment effect on some of the units. To evaluate this hypothesis, we computed the Hamming distance between the special group's units (this is the group where DAME and the SVM disagree) with units in other groups to investigate.

In Figure 6, the units within the leftmost blue (negative) labeled matched group were much closer to other blue (negative) labeled matched groups than to green (neutral) or red (positive) labeled groups, suggesting that smoothing the estimates after running DAME would likely make them consistent with the SVM results. The units within the leftmost green (neutral) labeled matched group are not closer to other green (neutral) labeled matched groups than other colors, suggesting that neither SVM nor DAME have information to properly identify the causal effect for this group. We similarly investigated the blue (negative) labeled group for which CATE= 0.5 and again, the covariate values of its units were closer in Hamming distance to other blue (negative) labeled groups than to other points. Thus, additional smoothing of the CATEs from the matched groups could likely yield estimated positive and negative treatment effects similar to those of the SVMs.