

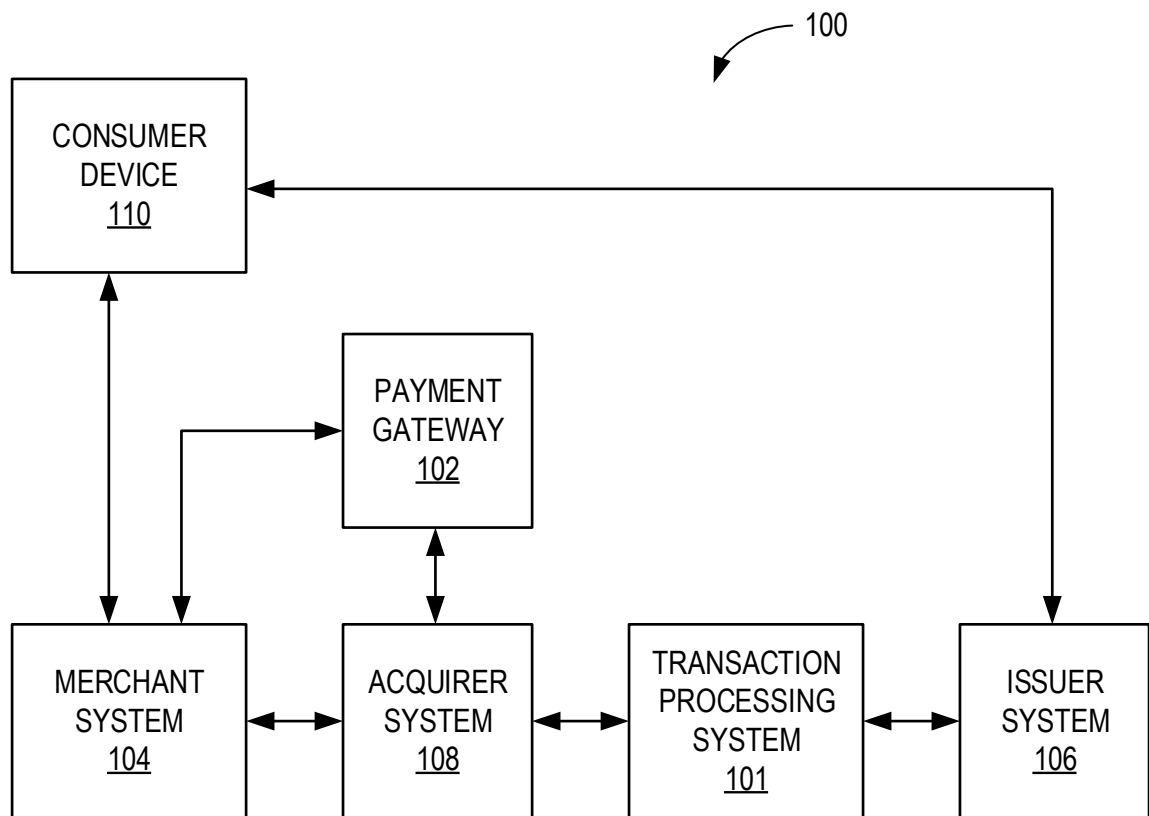


FIG. 1

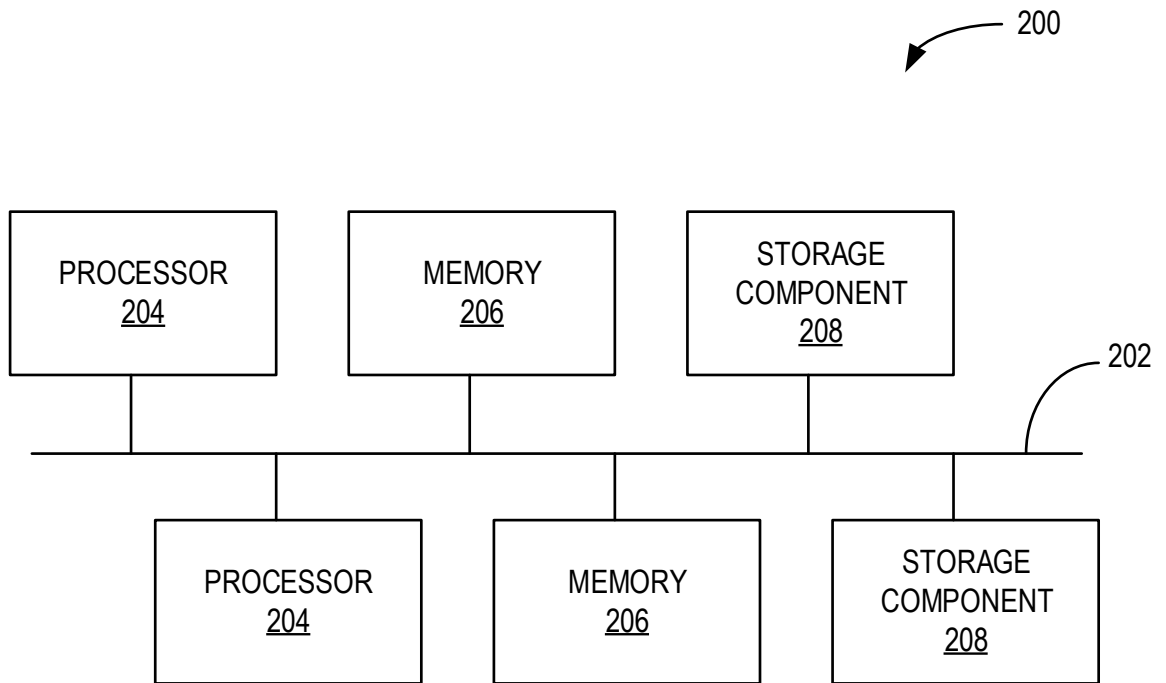


FIG. 2

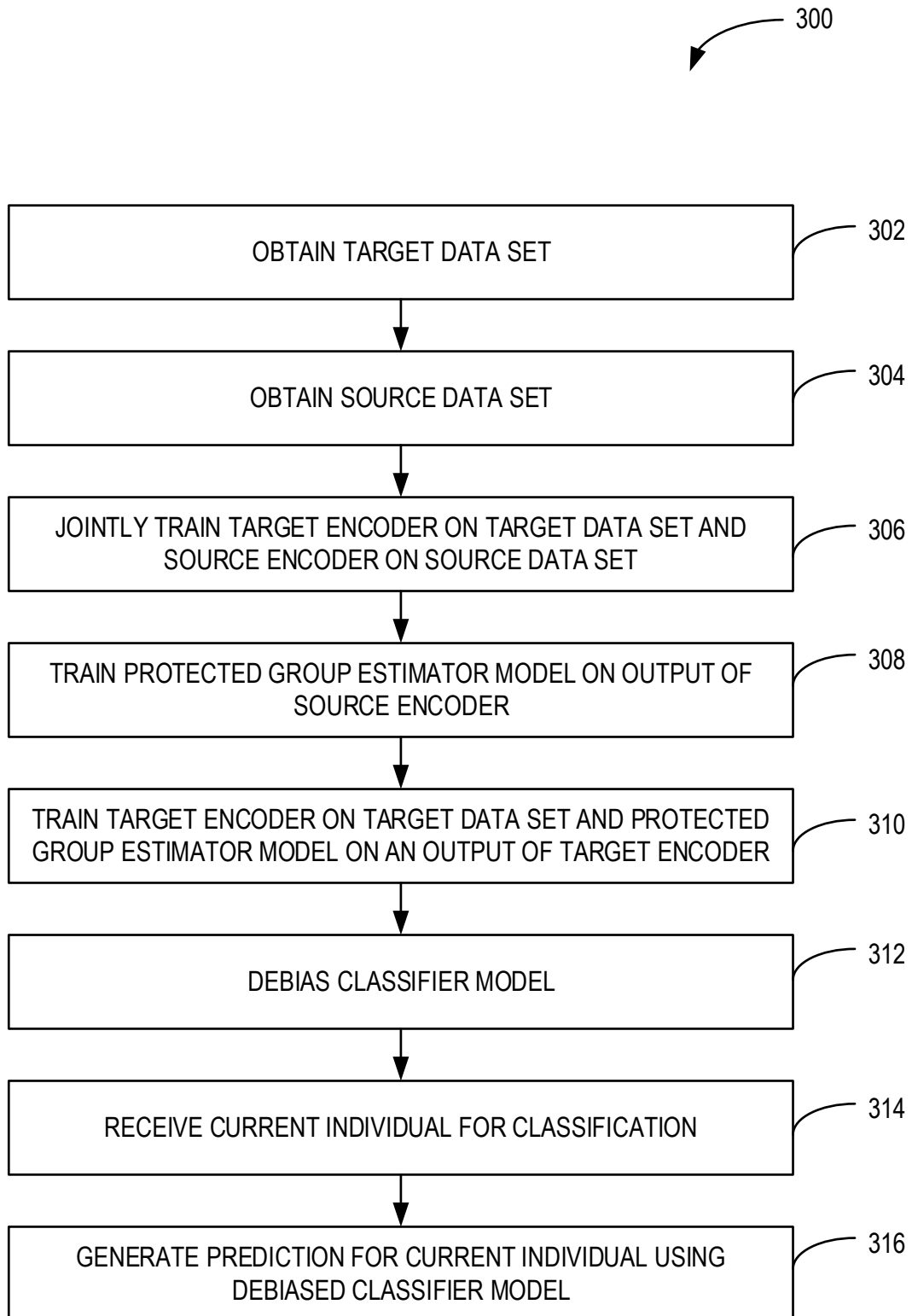
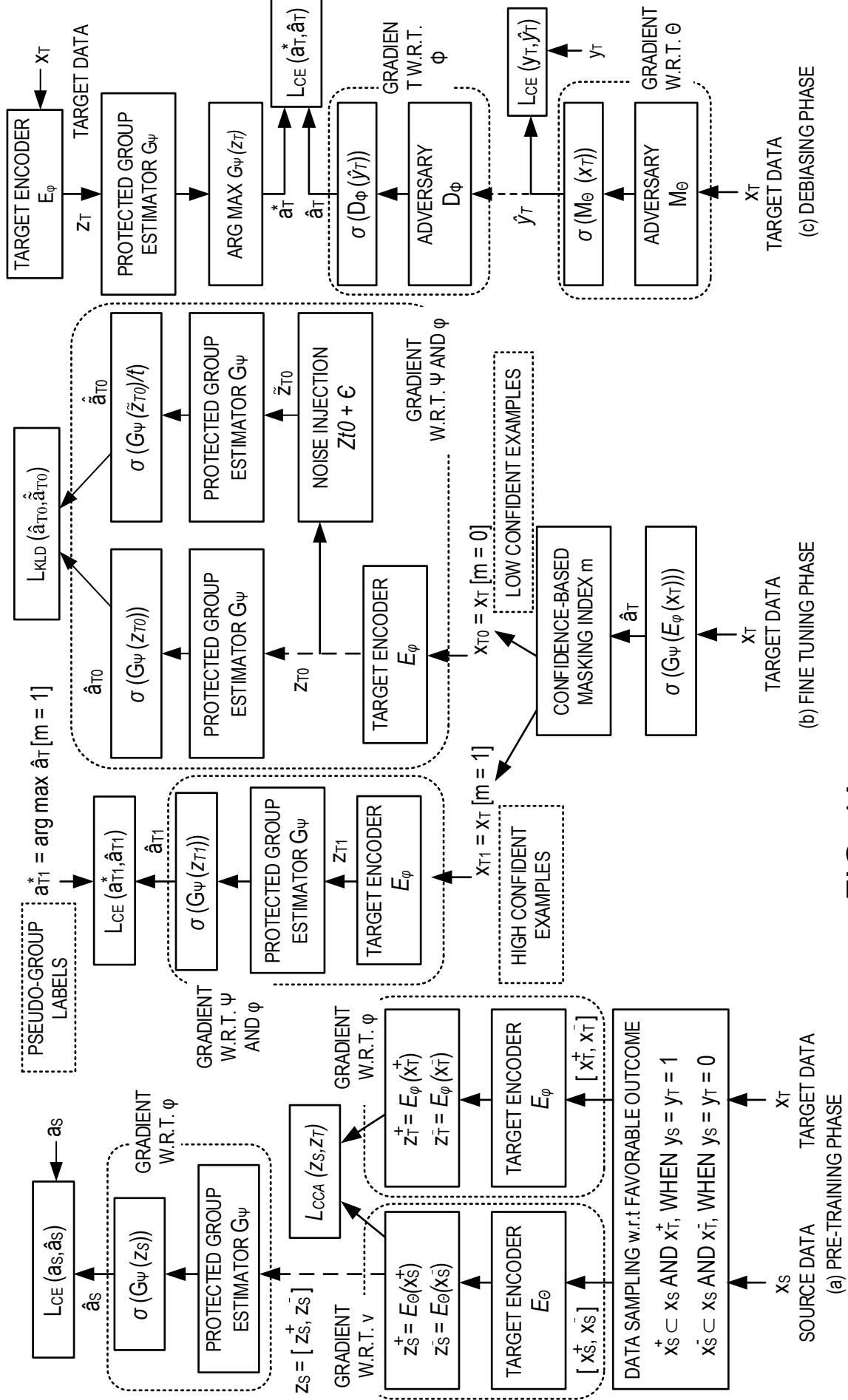


FIG. 3



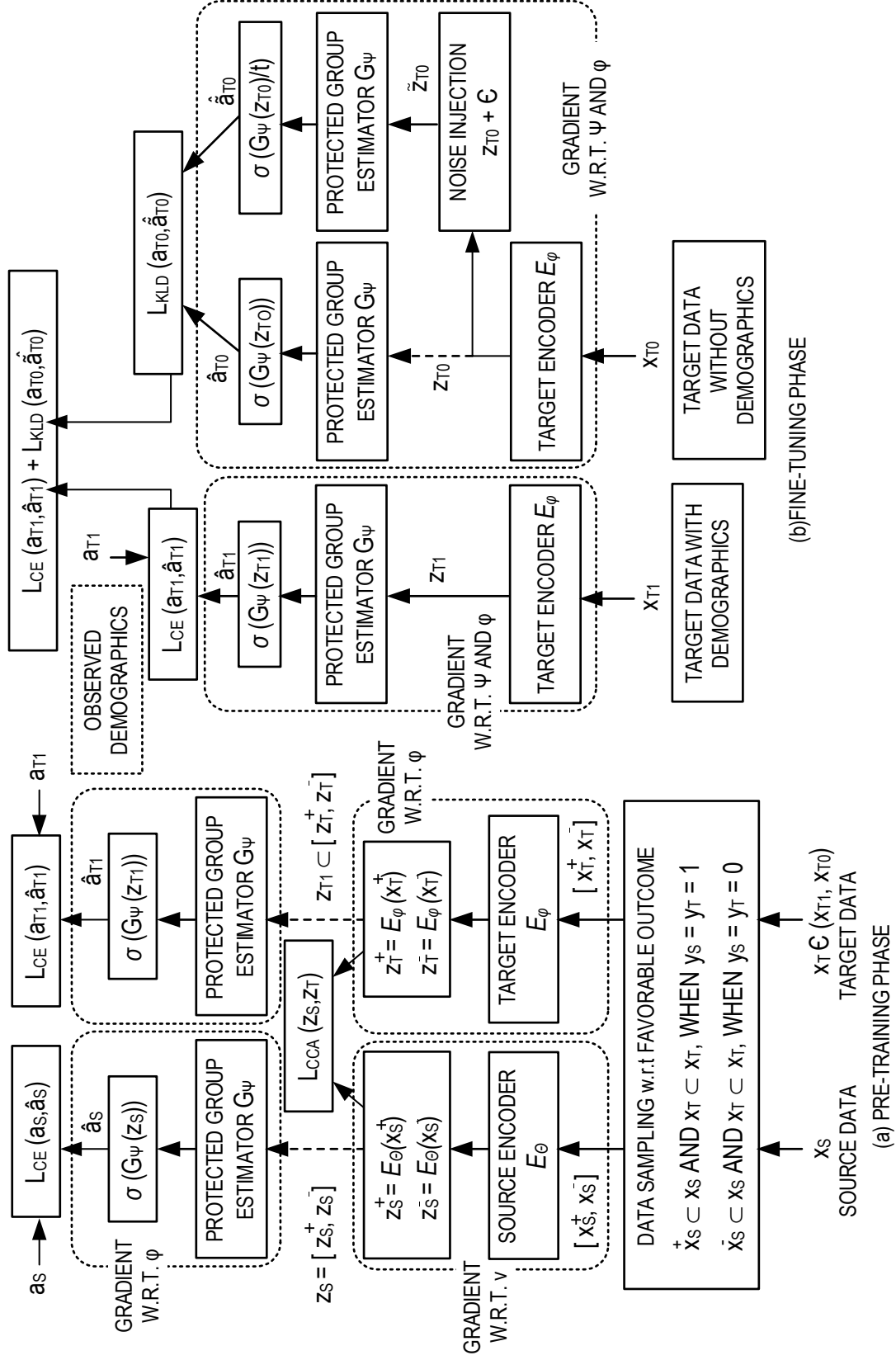


FIG. 4B

DATASET	METHOD	BAL. ACC. ↑ OVERALL	BAL. ACC. ↑ MIN	AUC ↑ OVERALL	AUC ↑ MIN	DPR ↑	DPD ↓
ADULT	BASELINE	0.773±0.009	0.764±0.007	0.910±0.001	0.910±0.001	0.888±0.009	0.164±0.012
	DRO	0.735±0.022	0.716±0.020	0.903±0.002	0.903±0.002	0.879±0.002	0.146±0.027
	ARL	0.778±0.004	0.769±0.003	0.912±0.000	0.912±0.000	0.890±0.000	0.170±0.004
	SLSD	0.700±0.028	0.695±0.023	0.828±0.035	0.828±0.035	0.822±0.028	0.078±0.033
	R-SLSD	0.763±0.012	0.750±0.012	0.900±0.003	0.900±0.003	0.885±0.001	0.103±0.019
DEFAULT	BASELINE	0.781±0.007	0.737±0.005	0.885±0.001	0.885±0.001	0.841±0.001	0.114±0.004
	DRO	0.753±0.017	0.726±0.013	0.874±0.002	0.874±0.002	0.830±0.002	0.114±0.007
	ARL	0.778±0.005	0.744±0.003	0.880±0.001	0.880±0.001	0.839±0.002	0.175±0.006
	SLSD	0.752±0.092	0.714±0.079	0.835±0.157	0.835±0.157	0.789±0.146	0.084±0.029
	R-SLSD	0.775±0.010	0.737±0.006	0.882±0.003	0.882±0.003	0.841±0.002	0.080±0.013
ACSINCOME	BASELINE	0.798±0.003	0.794±0.003	0.894±0.000	0.888±0.000	0.698±0.007	0.123±0.004
	DRO	0.771±0.018	0.766±0.020	0.875±0.002	0.870±0.002	0.724±0.032	0.106±0.009
	ARL	0.801±0.003	0.797±0.003	0.896±0.000	0.890±0.000	0.700±0.007	0.124±0.003
	SLSD	0.723±0.048	0.715±0.051	0.817±0.056	0.804±0.061	0.727±0.036	0.096±0.020
	R-SLSD	0.797±0.003	0.792±0.003	0.892±0.001	0.889±0.000	0.786±0.013	0.083±0.006

FIG. 5

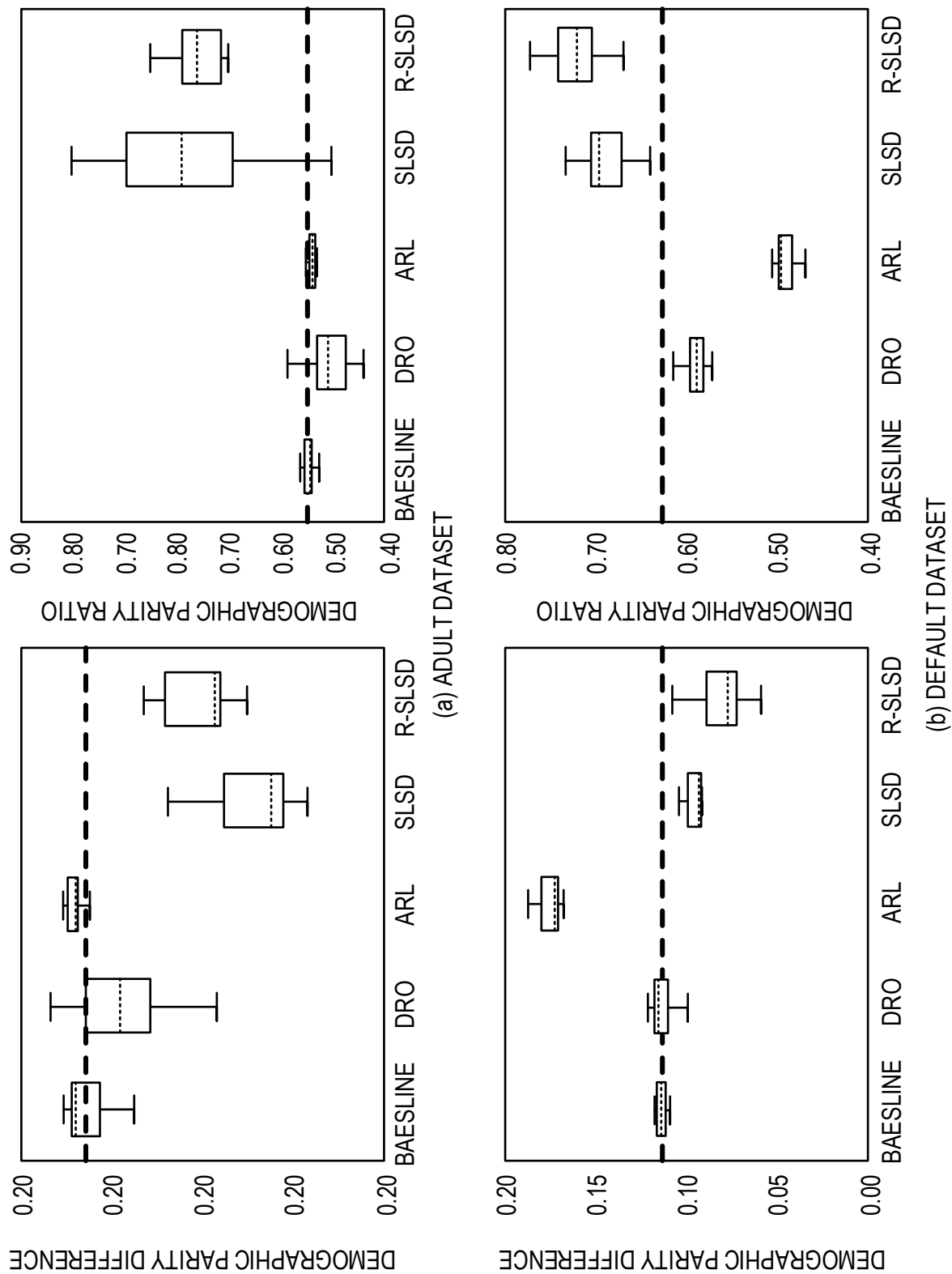


FIG. 6

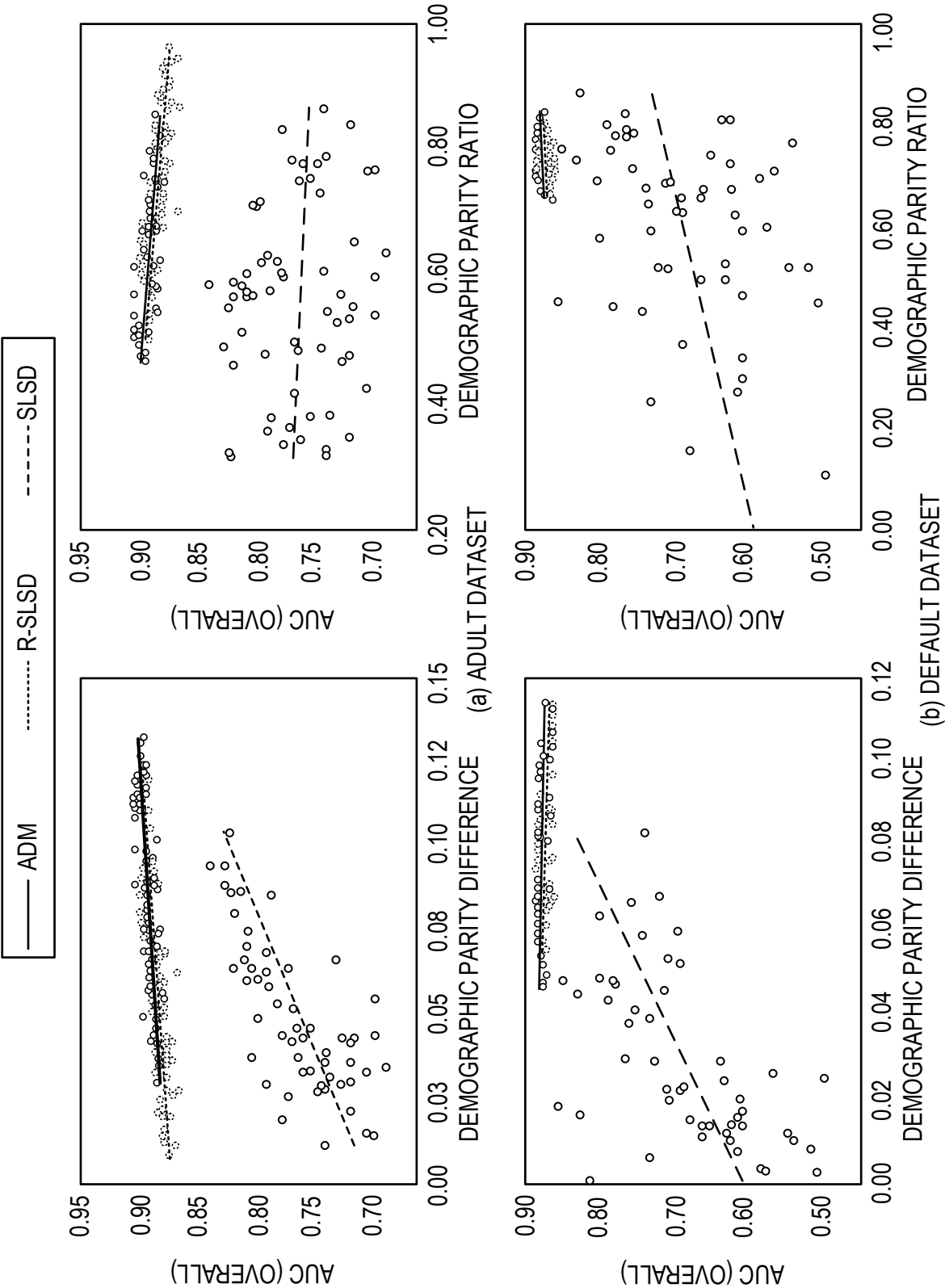


FIG. 7

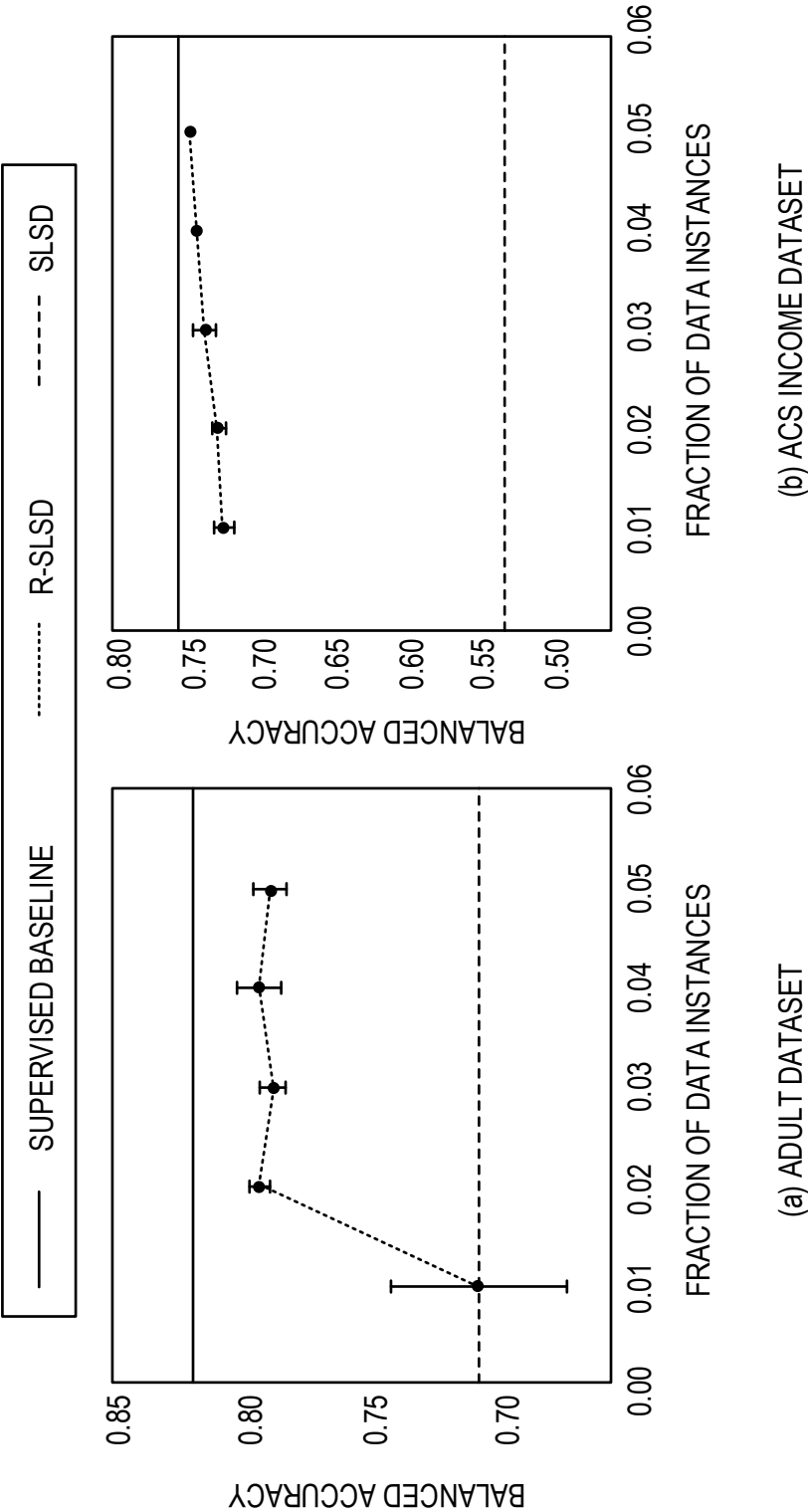


FIG. 8

METHOD	BAL. ACC. ↑ OVERALL	AUC ↑ OVERALL	DPR ↑	DPD ↓
BASELINE	0.684	0.745	0.642	0.235
DRO	0.665	0.740	0.671	0.236
ARL	0.685	0.745	0.642	0.234
SLSD	0.676	0.736	0.660	0.221
R-SLSD	0.653	0.739	0.718	0.224

FIG. 9

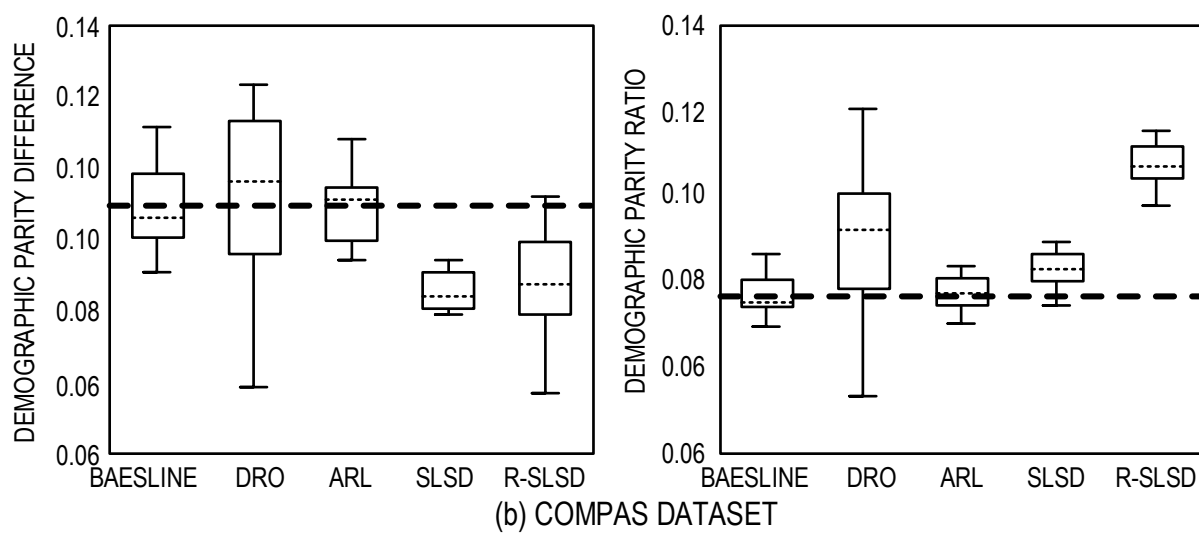
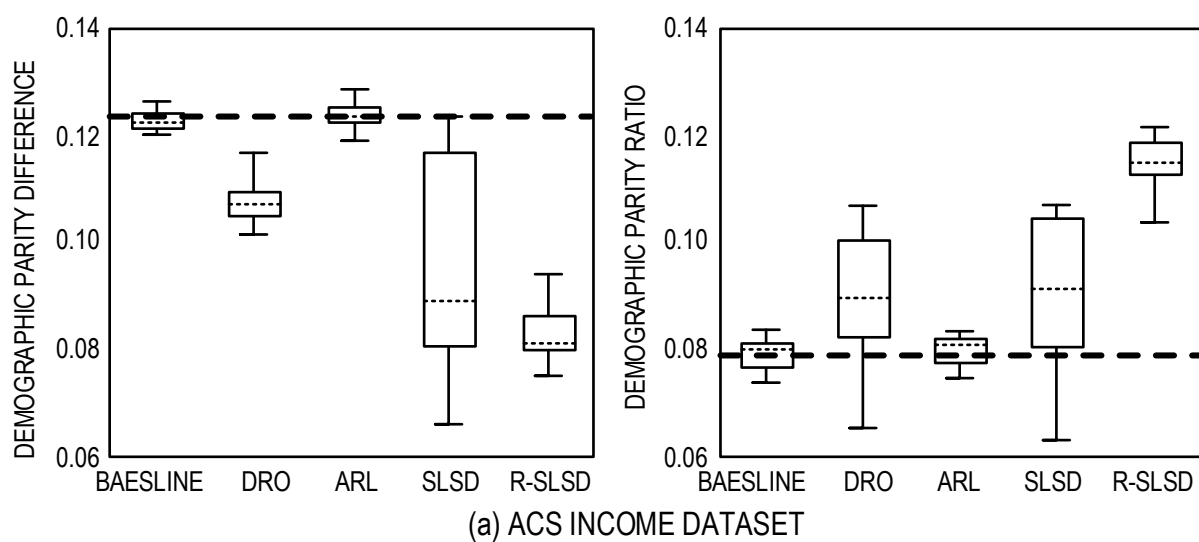


FIG. 10

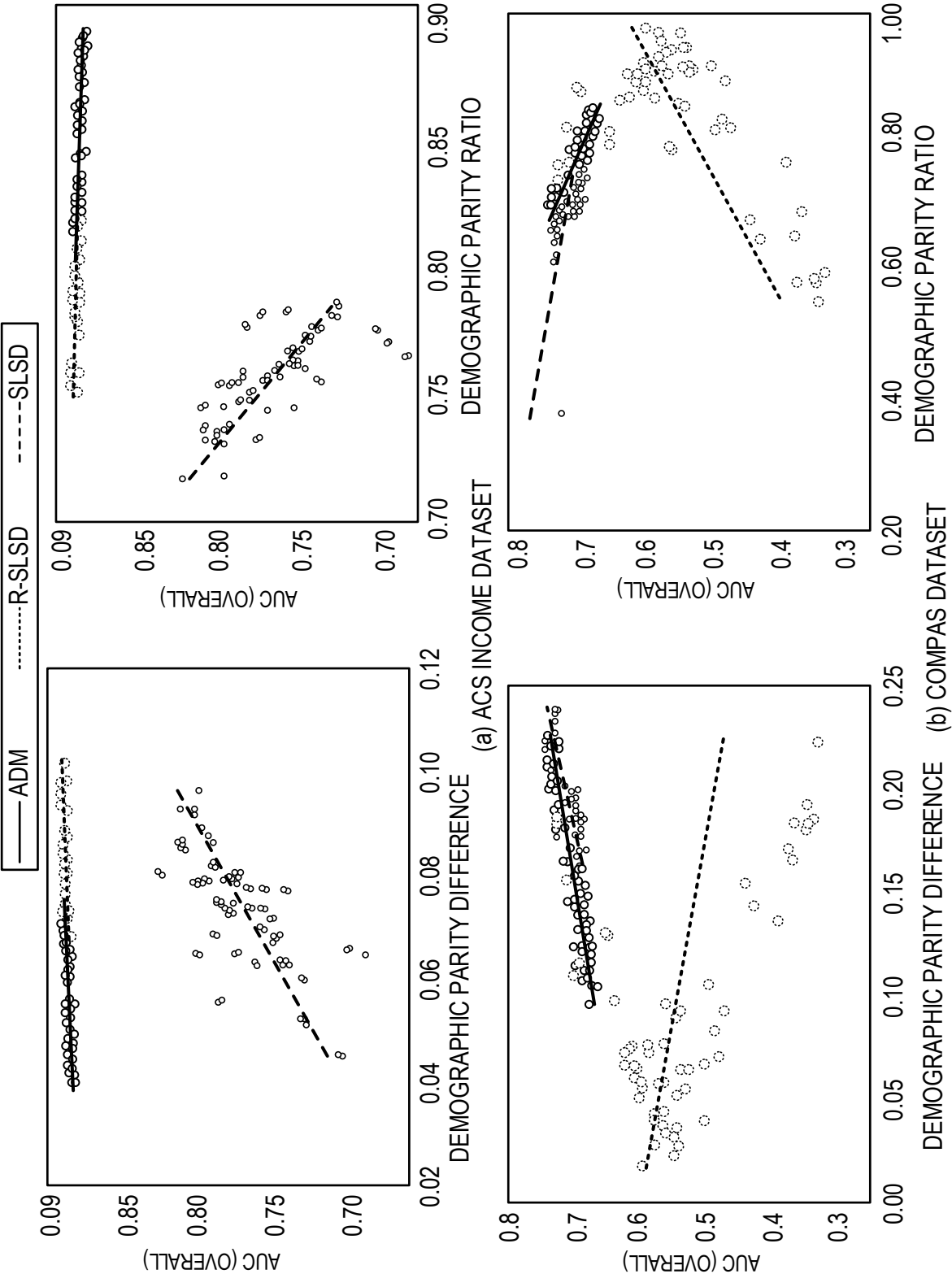


FIG. 11

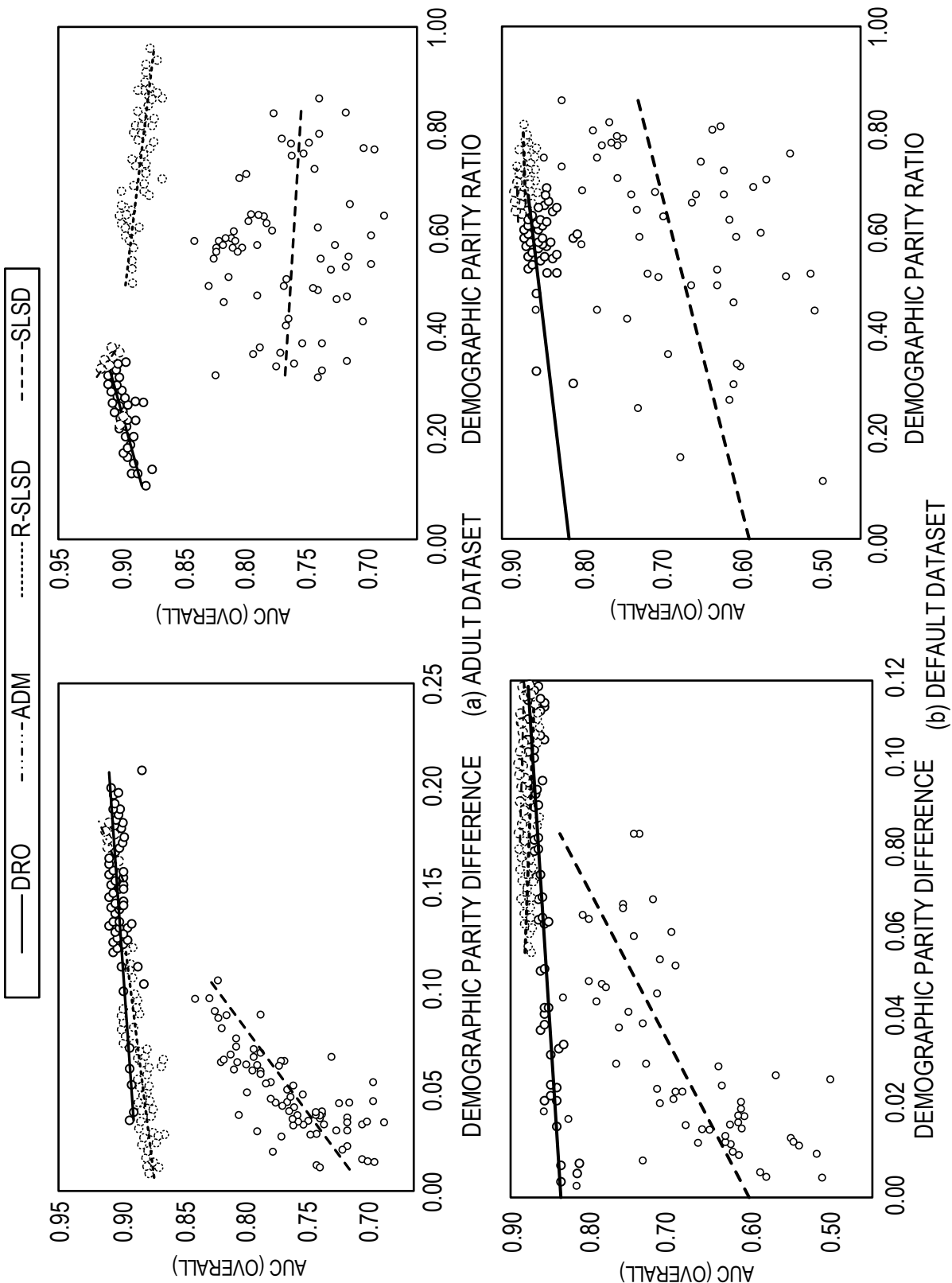


FIG. 12A

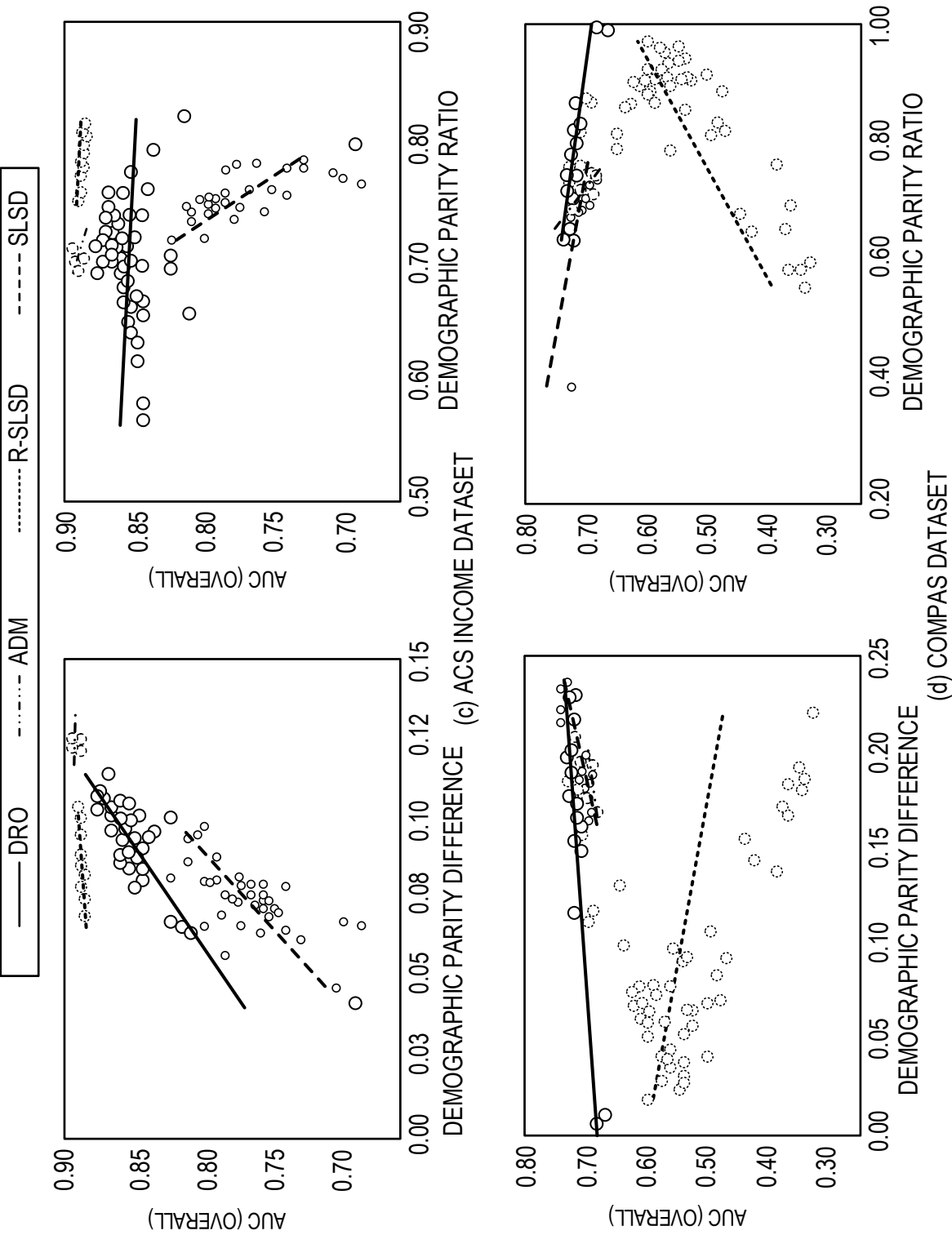


FIG. 12B

SOURCE DATASET					TARGET DATASET				
DATASET	SIZE	NO. OF FEATURES	PROTECTED	PREDICTION TASK	DATASET	SIZE	NO. OF FEATURES	PROTECTED	PREDICTION TASK
ACS INCOME	1664500	11	SEX	INCOME ≥50k?	ADULT	48842	15	SEX	INCOME ≥50k?
ADULT	48842	15	SEX	INCOME ≥50k?	DEFAULT	30000	24	SEX	DEFAULT PAYMENT?
ADULT	48842	15	RACE	INCOME ≥50k?	COMPAS	7214	11	RACE	RECIDIVATE IN 2 YEARS?
ADULT	48842	15	SEX	INCOME ≥50k?	ACS INCOME	16645016	11	SEX	INCOME ≥50k?

FIG. 13

ALGORITHM 1: PROTECTED GROUP ESTIMATION IN SLSD

INPUT:

SOURCE DATASET $\{(x^{(i)}_S, y^{(i)}_S, a^{(i)}_S)\}_{i=1}^{n_S}$;

TARGET DATASET $\{(x^{(i)}_T, y^{(i)}_T)\}_{i=1}^{n_T}$;

SOFTMAX TEMPERATURE $t = 0.4$.

PRE-TRAINING STEP:

- RANDOMLY INITIALIZE SOURCE ENCODER E_θ , TARGET ENCODER E_ϕ ,
- AND PROTECTED GROUP ESTIMATOR (PGE) G_ψ .
- DRAW n_S^+ INDIVIDUALS $x_S^+ \subset x_S$ WHEN $y_S = 1$ AND n_S^-
- INDIVIDUALS $x_S^- \subset x_S$ WHEN $y_S = 0$.
- SIMILARLY DRAW $x_T^+ \subset x_T$ AND $x_T^- \subset x_T$ VIA OVERSAMPLING OR DOWNSAMPLING TO ENSURE $n_T^+ = n_S^+$ AND $n_T^- = n_S^-$ INDIVIDUALS.
- FOR EACH ITERATION:
 - ESTIMATE $[z_S^+, z_S^-] = E_\psi([x_S^+, x_S^-])$.
 - ESTIMATE $[z_T^+, z_T^-] = E_\phi([x_T^+, x_T^-])$.
 - CONCATENATE $z_S = [z_S^+, z_S^-]$ AND $z_T = [z_T^+, z_T^-]$.
 - ESTIMATE $\hat{a}_S = \sigma(G_\psi(z_S))$
 - UPDATE USING ADAM OPTIMIZATION VIA BACKPROPAGATION:

$$\min_{\psi, \phi, \psi} L_{CCA}(z_S, z_T) + L_{CE}(\hat{a}_S, a_S).$$

FINE-TUNING STEP:

- ESTIMATE $\hat{a}_T = \sigma(G_\psi(z_T))$, WHERE $z_T = E_\phi(x_T)$.
- SET MASKING INDEX $m = 1$ FOR THE TOP $n_T/2$ INDIVIDUAL W.R.T.
- PROBABILITIES $\hat{a}_T$ AND $m = 0$ FOR THE REMAINING INDIVIDUALS.
- EXTRACT THE HIGH-CONFIDENT SAMPLES $x_{T1} = x_T[m = 1]$, AND THE LOW-CONFIDENT SAMPLES $x_{T0} = x_T[m = 0]$.
- GENERATE PSEUDO-GROUP LABELS $a_{T1} = \arg \max \sigma(G_\psi(z_{T1}))$.
- FOR EACH ITERATION:
 - ESTIMATE $z_{T1} = E_\phi(x_{T1})$ AND $z_{T0} = E_\phi(x_{T0})$.
 - ESTIMATE $\hat{a}_{T1} = \sigma(G_\psi(z_{T1}))$.
 - ESTIMATE $\hat{a}_{T0} = \sigma(G_\psi(z_{T0}))$.
 - ESTIMATE $\hat{\hat{a}}_{T0} = \sigma(G_\psi(z_{T0} + \epsilon)/t)$,
WHERE $\epsilon \sim \text{CAUCHY}(0, 200)$.
 - UPDATE USING ADAM OPTIMIZATION VIA BACKPROPAGATION:

$$\min_{\phi, \psi} L_{CE}(\hat{a}_{T1}, a_{T1}) + \text{LKLD}(\hat{\hat{a}}_{T0}, \hat{a}_{T0}).$$

OUTPUT:

FINE-TUNED TARGET ENCODER E_ϕ AND FINE-TUNED PGE G_ψ

FIG. 14

ALGORITHM 2: PROTECTED GROUP ESTIMATION IN R-SLSD

INPUT:

SOURCE DATASET $\{(x^{(i)}_S, y^{(i)}_S, a^{(i)}_S)\}_{i=1}^{n_S}$;

TARGET DATA SUBSET $\{(x^{(i)}_{T1}, y^{(i)}_{T1}, a^{(i)}_{T1})\}_{i=1}^{n_{T1}}$; AND
 $\{(x^{(i)}_{T0}, y^{(i)}_{T0})\}_{i=1}^{n_{T0}}$, WHERE $n_{T0} \gg n_{T1}$;

SOFTMAX TEMPERATURE $t = 0.4$.

PRE-TRAINING STEP:

- RANDOMLY INITIALIZE SOURCE ENCODER E_θ , TARGET ENCODER E_ϕ , AND PROTECTED GROUP ESTIMATOR (PGE) G_ψ .
- DRAW n_S^+ INDIVIDUALS $x_S^+ \subset x_S$ WHEN $y_S = 1$ AND n_S^- INDIVIDUALS $x_S^- \subset x_S$ WHEN $y_S = 0$.
- SIMILARLY DRAW $x_T^+ \subset \{x_{T0}, x_{T1}\}$ AND $x_T^- \subset \{x_{T0}, x_{T1}\}$ VIA OVERSAMPLING OR DOWNSAMPLING TO ENSURE $n_T^+ = n_S^+$ AND $n_T^- = n_S^-$ INDIVIDUALS.
- FOR EACH ITERATION:
 - ESTIMATE $[z_S^+, z_S^-] = E_\psi([x_S^+, x_S^-])$.
 - ESTIMATE $[z_T^+, z_T^-] = E_\phi([x_T^+, x_T^-])$.
 - CONCATENATE $z_S = [z_S^+, z_S^-]$ AND $z_T = [z_T^+, z_T^-]$.
 - ESTIMATE $\hat{a}_S = \sigma(G_\psi(z_S))$.
 - ESTIMATE $\hat{a}_{T1} = \sigma(G_\psi(E_\phi(z_{T1})))$.
 - UPDATE USING ADAM OPTIMIZATION VIA BACKPROPAGATION:

$$\min_{\psi, \phi, \Psi} \text{LCCA}(z_S, z_T) + \text{LCE}(\hat{a}_S, a_S) + \text{LCE}(\hat{a}_{T1}, a_{T1}).$$

FINE-TUNING STEP:

- FOR EACH ITERATION
 - ESTIMATE $z_{T1} = E_\phi(x_{T1})$ AND $z_{T0} = E_\phi(x_{T0})$.
 - ESTIMATE $\hat{a}_{T1} = \sigma(G_\psi(z_{T1}))$.
 - ESTIMATE $\hat{a}_{T0} = \sigma(G_\psi(z_{T0}))$.
 - ESTIMATE $\hat{\hat{a}}_{T0} = \sigma(G_\psi(z_{T0} + \epsilon)/t)$,
 WHERE $\epsilon \sim \text{CAUCHY}(0, 200)$.
 - UPDATE USING ADAM OPTIMIZATION VIA BACKPROPAGATION:

$$\min_{\phi, \Psi} \text{LCE}(\hat{a}_{T1}, a_{T1}) + \text{LKLD}(\hat{\hat{a}}_{T0}, \hat{a}_{T0}).$$

OUTPUT:

FINE-TUNED TARGET ENCODER E_ϕ AND FINE-TUNED PGE G_ψ

FIG. 15

ALGORITHM 3: DEBASING PROCEDURE FOR SLSD AND R-SLSD

INPUT:

TARGET DATA $\{(x^{(i)}_T, y^{(i)}_T, a^{(i)}_T)\}_{i=1}^{n_T}$, FAIRNESS TUNING PARAMETER λ ,
FINE-TUNED TARGET ENCODER E_ϕ AND FINE-TUNED PGE G_ψ .

- RANDOMLY INITIALIZE CLASSIFIER M_θ AND ADVERSARY D_ϕ .

PRE-TRAINING CLASSIFIER:

- FOR EACH ITERATION:
 - ESTIMATE $\hat{y}_T = \sigma(M_\theta(x_T))$.
 - UPDATE USING ADAM OPTIMIZATION VIA BACKPROPAGATION:

$$\min_{\theta} L_{CE}(\hat{y}_T, y_T).$$

PRE-TRAINING ADVERSARY:

- FOR EACH ITERATION:
 - ESTIMATE GROUP $a_T^* = \arg \max G_\psi(E_\phi(x_T))$.
 - ESTIMATE $\hat{y}_T = \sigma(M_\theta(x_T))$ AND $\hat{a}_T = \sigma(D_\phi(\hat{y}_T))$.
 - UPDATE USING ADAM OPTIMIZATION VIA BACKPROPAGATION:

$$\min_{\phi} \lambda L_{CE}(\hat{a}_T, a_T^*).$$

SIMULTANEOUS TRAINING:

- FOR EACH ITERATION:
 - ESTIMATE GROUP $a_T^* = \arg \max G_\psi(E_\phi(x_T))$.
 - ESTIMATE $\hat{y}_T = \sigma(M_\theta(x_T))$ AND $\hat{a}_T = \sigma(D_\phi(\hat{y}_T))$.
 - UPDATE USING ADAM OPTIMIZATION VIA BACKPROPAGATION:

$$\min_{\theta} \max_{\phi} \lambda L_{CE}(\hat{y}_T, y_T) - \lambda L_{CE}(\hat{a}_T, a_T^*).$$

OUTPUT:

DEBIASED CLASSIFIER M_θ .

FIG. 16