DEEP FAIR LEARNING: TASK-AWARE FAIR REPRESENTATIONS VIA JOINT DISTANCE-COVARIANCE REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Ensuring fairness is essential as machine learning increasingly informs consequential decisions. However, many fairness-aware methods focus on the outputs of individual predictors, without directly controlling sensitive information retained in the underlying representations. We propose Deep Fair Learning (DFL), which combines distance covariance regularization with predictive loss to jointly learn representations and downstream predictors, promoting fairness at both levels while preserving task-relevant information. Its marginal and class-conditional formulations target independence and separation, respectively. Under suitable regularity conditions, we establish non-asymptotic joint excess-risk rates and convergence of the learned representation up to natural invariances. We further derive fairness-inheritance bounds linking representation-level dependence to downstream disparities over suitable predictor classes, extending fairness guarantees beyond the jointly trained predictor. Experiments on tabular, text, and image benchmarks show that DFL achieves lower fairness gaps than competing methods in many evaluated settings while maintaining competitive predictive accuracy, with fairness gains largely preserved after downstream retraining.

1 INTRODUCTION

Machine learning increasingly informs consequential decisions in lending, hiring, and healthcare, making fairness a central requirement alongside predictive accuracy (Hardt et al., 2016; Potash et al., 2015; Barocas et al., 2023). Many fairness-aware methods constrain the outputs of an individual predictor through penalties, adversaries, or post-processing (Wan et al., 2023; Caton & Haas, 2024). Such interventions can reduce disparities in the fitted model while leaving sensitive information in the features it uses. When those features are reused and the downstream predictor is retrained, that information remains available to influence predictions. This motivates a broader objective: learn representations that support accurate prediction while carrying fairness benefits beyond the jointly trained predictor.

Fair representation learning addresses this objective by controlling sensitive information in the feature space (Zemel et al., 2013; Madras et al., 2018). The central challenge is to reduce sensitive dependence without discarding task-relevant information. Optimizing these goals separately can shape features that are weakly associated with the sensitive attribute but poorly suited to the downstream task. Joint learning provides a direct way to balance them: the predictive loss determines which information the representation must preserve, while a representation-level penalty controls sensitive dependence.

We propose *Deep Fair Learning* (DFL), a framework that jointly learns a nonlinear representation map and a downstream predictor using predictive loss and distance covariance regularization. Marginal and class-conditional penalties target independence and separation, respectively. Distance covariance detects nonlinear dependence, accommodates discrete or continuous sensitive attributes, and admits a minibatch estimator without an adversarial optimization or a kernel-bandwidth parameter. In our fine-tuning implementation, a compact trainable encoder operates on frozen backbone features, making the resulting representation available for downstream retraining.

Existing fair representation methods provide important foundations: adversarial approaches connect representation alignment to downstream fairness (Madras et al., 2018; Zhao et al., 2020), while kernel

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison with related fairness methods. "Joint" denotes encoder–head training driven by the task loss; "Sep." denotes support for class-conditional fairness. Guarantees describe the published formulations. The experiments evaluate the subset listed in Section 4.1.

Method	Dependence measure	Level	Joint	Sep.	Guarantee
AdvDebias (Zhang et al., 2018)	adversarial (implicit)	output	—	✓	—
FairMixup (Chuang & Mroueh, 2021)	path-wise DP surrogate	output	—	✓	—
DRAlign (Li et al., 2023)	distributional alignment	output	—	✓	—
DiffMCDP (Jin et al., 2024)	smoothed MCDP surrogate	output	—	—	—
INLP (Ravfogel et al., 2020)	linear predictability	repr. (post-hoc)	—	—	—
RLACE (Ravfogel et al., 2022a)	linear adversarial game	repr. (post-hoc)	—	—	linear-erasure
LEACE (Belrose et al., 2023)	linear predictability	repr. (post-hoc)	—	—	exact linear erasure
SUP (Shi et al., 2024)	SDR subspace	repr. (post-hoc)	—	—	subspace
LAFTR (Madras et al., 2018)	adversarial alignment	repr.	✓	✓	downstream fairness
FairDisCo (Liu et al., 2022)	distance-based dependence	repr. (VAE)	—	—	dependence analysis
FARE (Jovanović et al., 2023)	group distributions	repr. (restricted)	—	—	fairness certificates
VFAE (Louizos et al., 2016)	MMD (Gaussian kernel)	repr.	✓	—	—
CFair (Zhao et al., 2020)	adversarial cond. alignment	repr.	✓	✓	trade-off lower bound
Grari et al. (2021)	HGR (adversarial)	repr.	✓	—	—
Quadrianto et al. (2019)	HSIC (Gaussian kernel)	data domain	✓	✓	—
Tan et al. (2020)	RKHS cross-covariance	repr. (RKHS)	—	✓	exact for in-subspace heads
Sadeghi et al. (2022)	kernel dependence	repr. (RKHS)	—	—	closed-form optimum
DFL (this paper)	(conditional) DC	repr.	✓	✓	joint risk; repr. convergence fairness inheritance

and distribution-matching methods characterize trade-offs and transfer guarantees (Tan et al., 2020; Sadeghi et al., 2022; Oneto et al., 2020). Distance-based approaches also learn fair representations through variational objectives (Liu et al., 2022). A central statistical question is how jointly learning a nonlinear encoder and its task predictor affects the estimation of both objects, and how the resulting representation-level control governs predictors fitted afterwards.

DFL connects training, representation recovery, and downstream reuse through three guarantees. A non-asymptotic bound controls the joint excess risk of the learned encoder and head at a fixed fairness–utility trade-off. Under identifiability and local growth, this bound yields convergence of the representation up to rigid transformations and calibration of predictive risk. Finally, representation-level distance covariance bounds the group disparity of every replacement head in a specified function class; the class-conditional version controls opportunity gaps. To our knowledge, this is the first analysis of jointly trained, distance-covariance-regularized neural representations to establish this combination of guarantees. Experiments across tabular, text, and image data evaluate the same framework and its persistence after head replacement. Table 1 summarizes the distinctions; detailed related work appears in Appendix A.

Contributions.

1. We develop a unified framework for independence and separation that couples representation-level dependence control with downstream predictive loss, preserving task-relevant information during fairness-aware training.
2. We establish non-asymptotic excess-risk rates for the jointly estimated encoder and head. Under identifiability and local growth, the rates imply representation recovery up to rigid transformations and convergence of predictive performance to the trade-off oracle.
3. We give uniform fairness-inheritance bounds for downstream heads with controlled complexity. Marginal and class-conditional versions convert representation-level dependence into quantitative score and opportunity guarantees after retraining.
4. Against strong prediction- and representation-level baselines, including RLACE, CFair-EO and LAFTR-EO-CE, DFL ranks first or second in 30 of 48 displayed accuracy and fairness comparisons. Downstream retraining largely preserves its gains, and controlled ablations examine the representation-level penalty.

2 FAIR REPRESENTATION LEARNING WITH DFL

Let $(\boldsymbol{X}, Y, \boldsymbol{Z})$ denote the input features, target, and sensitive attributes, with $\boldsymbol{X} \in \mathbb{R}^p$, $Y \in \{1, \ldots, K\}$, and $\boldsymbol{Z} \in \mathbb{R}^d$. The features may be supplied by a frozen backbone, and the sensitive attributes may be discrete, continuous, or multivariate. DFL learns an encoder $g_\theta : \mathbb{R}^p \to \mathbb{R}^p$ and a head $f_\phi : \mathbb{R}^p \to \mathbb{R}^K$. The representation is $\boldsymbol{R} = g_\theta(\boldsymbol{X})$; the head produces logits, with class probabilities $\mathrm{Softmax}(f_\phi(\boldsymbol{R}))$ and prediction $\widehat{Y} = \arg\max_k f_{\phi,k}(\boldsymbol{R})$. Write $\mathbf{e}_Y$ for the one-hot label vector.

2.1 FROM PREDICTION FAIRNESS TO REPRESENTATION FAIRNESS

Two widely used group-fairness criteria are *independence*, $\widehat{Y} \perp\!\!\!\perp \boldsymbol{Z}$, and *separation*, $\widehat{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid Y$ (Barocas et al., 2023). In binary classification with a binary sensitive attribute Z, demographic parity (DP) requires

$$\mathbb{P}(\widehat{Y} = 1 \mid Z = 0) = \mathbb{P}(\widehat{Y} = 1 \mid Z = 1), \tag{1}$$

while equalized odds (EO) requires, for each $y \in \{0, 1\}$,

$$\mathbb{P}(\widehat{Y} = 1 \mid Y = y, Z = 0) = \mathbb{P}(\widehat{Y} = 1 \mid Y = y, Z = 1). \tag{2}$$

DP equalizes positive-prediction probabilities across groups, whereas EO equalizes both true- and false-positive rates. These are population fairness criteria. Our experiments quantify disparities using the classwise TPR-RMS gap and the MCDP score-distribution metric, whose definitions and multiclass aggregation are given in Appendix B.

DFL applies these criteria to the representation itself:

$$\boldsymbol{R} \perp\!\!\!\perp \boldsymbol{Z} \quad \text{or} \quad \boldsymbol{R} \perp\!\!\!\perp \boldsymbol{Z} \mid Y. \tag{3}$$

This changes what the fairness intervention controls. If group-conditional representation distributions agree, any predictor using only those representations sees the same distribution across groups. Consequently, exact representation independence passes to every measurable downstream predictor; the conditional version gives the same conclusion within each target class. Reducing dependence therefore provides a route to fairness that can persist when the head is replaced. In finite samples dependence is generally nonzero, so Section 3.3 quantifies the remaining disparity for suitable predictor classes.

The two targets reflect different choices about which group differences to remove. Marginal independence aligns representations across groups without conditioning on the outcome; separation aligns them within each outcome class. When outcome frequencies differ between groups, these are different statistical requirements. DFL makes this choice through the penalty rather than through a change in architecture, so the same joint-learning framework can address either criterion.

2.2 MEASURING DEPENDENCE WITH DISTANCE COVARIANCE

Distance covariance detects general dependence between random vectors of possibly different dimensions (Székely et al., 2007; Székely & Rizzo, 2009). For vectors $\boldsymbol{U}, \boldsymbol{V}$ with finite second moments, we use DC to denote the *squared* distance covariance:

$$\begin{aligned}\mathrm{DC}(\boldsymbol{U}, \boldsymbol{V}) = \mathbb{E}\big[\|\boldsymbol{U} - \boldsymbol{U}'\| \, \|\boldsymbol{V} - \boldsymbol{V}'\|\big] - 2\mathbb{E}\big[\|\boldsymbol{U} - \boldsymbol{U}'\| \, \|\boldsymbol{V} - \boldsymbol{V}''\|\big] \\ + \mathbb{E}\|\boldsymbol{U} - \boldsymbol{U}'\| \, \mathbb{E}\|\boldsymbol{V} - \boldsymbol{V}'\|,\end{aligned} \tag{4}$$

where primes indicate independent copies of the pair and $\|\cdot\|$ is the Euclidean norm. It is nonnegative and equals zero precisely under independence. We write $\mathrm{dcov}(\boldsymbol{U}, \boldsymbol{V}) = \sqrt{\mathrm{DC}(\boldsymbol{U}, \boldsymbol{V})}$ when its square root is needed.

For separation, let $\rho_k = \mathbb{P}(Y = k) > 0$ be the population class probabilities and $\pi_k = \rho_k^4 / \sum_j \rho_j^4$ the corresponding population penalty weights. Define $\mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R} \mid Y) = \sum_k \pi_k \, \mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R} \mid Y = k)$. Because every π_k is positive, this quantity vanishes exactly when representation separation holds. Given a sample, $\widehat{\mathrm{DC}}_n$ is the unbiased degree-four U-statistic estimator of marginal DC. Its conditional counterpart uses empirical weights $w_{n,k}$ proportional to $\binom{n_k}{4}$, where n_k is the sample count for class k. The full formula and its treatment of classes with fewer than four observations are in Appendix C.

Theorem 1 (Consistency of the dependence estimators)**.** *For a fixed representation g, suppose $(\boldsymbol{Z}_i, \boldsymbol{R}_i, Y_i)$ are i.i.d., $\mathbb{E}(\|\boldsymbol{R}\|^2 + \|\boldsymbol{Z}\|^2) < \infty$, and each of finitely many classes has positive probability. Then, as $n \to \infty$,*

$$\widehat{\mathrm{DC}}_n(\boldsymbol{Z}, \boldsymbol{R}) \xrightarrow{\mathcal{P}} \mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R}), \qquad \widehat{\mathrm{DC}}_n(\boldsymbol{Z}, \boldsymbol{R} \mid Y) \xrightarrow{\mathcal{P}} \mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R} \mid Y).$$

The proof is in Appendix E.5. This result justifies the population targets of the empirical penalties; the data-dependent learned encoder is analyzed in Section 3.

Why distance covariance? DC detects nonlinear dependence and accommodates both discrete and continuous sensitive variables. Its estimator uses pairwise distances directly, requiring neither a density model nor an adversarial inner optimization or a kernel-bandwidth choice. These properties make it convenient for minibatch training. It is related to the Hilbert–Schmidt independence criterion (HSIC) through distance-induced kernels (Sejdinovic et al., 2013), and the associated energy-distance identity also supplies the link to downstream fairness.

2.3 Joint learning of representations and predictions

A fair representation must also retain the information needed for the target task: a constant map would remove sensitive dependence but lose predictive utility. DFL balances these goals by training the encoder and head together. With $\boldsymbol{R} = g_\theta(\boldsymbol{X})$ and cross-entropy $\ell(y, u) = -\log \mathrm{Softmax}(u)_y$, the marginal objective is

$$\mathcal{L}(\theta, \phi) = \mathbb{E}[\ell(Y, f_\phi(\boldsymbol{R}))] + \lambda \,\mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R}) - \mu \,\mathrm{DC}(\mathbf{e}_Y, \boldsymbol{R}), \tag{5}$$

where $\lambda > 0$ and $\mu \geq 0$. Replacing the sensitive-dependence term by its class-conditional version gives the separation objective. The predictive loss shapes features for the actual downstream task; the label-dependence term provides an additional informativeness incentive. The tuning parameters determine the fairness–utility trade-off.

Joint optimization makes the predictive task part of representation learning itself. Gradients from the head identify features useful for classification, while the dependence penalty discourages sensitive structure in those features. The predictive loss directly measures how effectively the head uses the retained label information, complementing the informativeness penalty. Once the encoder is frozen, a replacement head uses features shaped by both requirements, even though that new head need not include a fairness penalty in its own training objective.

The empirical objective $\mathcal{L}_n$ replaces the expectation by a sample average and DC by its sample estimator. We optimize both networks with minibatch stochastic gradients. In the fine-tuning implementation, a compact DenseNet-style encoder transforms frozen backbone features while preserving their dimension. After training, the encoder can be retained and a new head fitted on its outputs. Appendix H gives the architecture, normalization, and training protocol; Appendix C gives both empirical objectives.

The training and reuse stages have distinct roles. During training, every minibatch passes through the encoder and head, and both networks receive gradients from the joint objective. For separation, the fairness term compares sensitive dependence within target classes, while the predictive term uses all observations in the batch. After training, predictions are computed from $g_\theta(\boldsymbol{X})$ without requiring a separate sensitive-attribute input. Group annotations are used to construct the training penalty and to evaluate disparities. The trained encoder exports features in the original feature dimension, allowing a new classifier for the same target to be trained independently. This separation of stages motivates the head-replacement evaluation in Section 4.3.

3 Theoretical analysis

We study DFL at fixed fairness–utility weights and give a chain of guarantees for the fitted model: a non-asymptotic joint-risk rate, representation and predictive calibration, and an end-to-end fairness bound for subsequently fitted heads. The last guarantee includes the encoder's estimation error as well as dependence at the population trade-off oracle. The main text states the results in simplified form; Appendix E.1 gives the network classes, complete assumptions, and formal statements. The excess-risk analysis covers both objectives, with the conditional extension in Corollary E.15. Conditional fairness and inheritance are developed in Appendices E.8 and F.

3.1 JOINT ESTIMATION AND EXCESS RISK

Write $\mathcal{L}(g, f)$ for (5) as a function of the encoder and head. Let (g_0, f_0) minimize this objective over bounded measurable pairs, and let $(\widehat{g}_n, \widehat{f}_n)$ minimize the empirical objective over bounded ReLU network classes $\mathcal{G}_n \times \mathcal{F}_n$. The pair (g_0, f_0) is the *trade-off oracle*: the population optimum for the chosen balance of predictive utility and fairness. Define the joint excess risk

$$\Delta_n = \mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}(g_0, f_0).$$

Boundedness controls the scale of the dependence penalties; the formal class restrictions and their role are given in Appendices D and E.1.

Here n is the training sample size, p is the feature and representation dimension, and K is the number of target classes. The dimensions p and K are fixed as n grows. The classes $\mathcal{G}_n$ and $\mathcal{F}_n$ contain the candidate encoders and heads, respectively, with network size allowed to grow with n. A network class is specified by its depth D, maximum width W, parameter count S, and output bound B. We write B_g and B_f for the encoder and head bounds. Coordinatewise clipping at $[-10, 10]$ gives the experimental encoder the Euclidean bound $B_g = 10\sqrt{p}$, fixing a finite scale for dependence and inheritance.

Population losses average over a fresh observation from the data distribution. Because the empirical minimizer depends on the training sample, Δ_n is random; the expectation in the excess-risk bound also averages over that sample. The weights (λ, μ) stay fixed in this analysis, so the comparison is always with the same population trade-off.

Theorem 2 (Joint excess risk, simplified)**.** *Suppose the inputs and sensitive attributes are bounded and the oracle encoder and head are Hölder-smooth of orders β_g and β_f. For suitably sized ReLU networks and $s_f = \min\{\beta_f, 1\}$, up to logarithmic factors,*

$$\mathbb{E}[\Delta_n] \lesssim n^{-\beta_f/(2\beta_f+p)} + n^{-s_f\beta_g/(2\beta_g+p)} + n^{-\beta_g/(2\beta_g+p)}. \tag{6}$$

The full statement and complexity-dependent bound are Theorem E.1 and (E.2) *in Appendix E.1.*

The bound separates approximation of the oracle pair from sampling error in the predictive loss and dependence estimates. The encoder's approximation affects both the penalty and the composed predictor, which explains the two representation terms. When $\beta_f \geq 1$, these terms have the same order, leaving one rate for learning the encoder and one for the head.

The smoothness indices describe the regularity of the oracle functions: a Hölder order between zero and one controls how quickly function values vary, while a larger order additionally controls derivatives. The factor $s_f = \min\{\beta_f, 1\}$ enters because perturbing the representation changes the input to the head. Network size balances the ability to approximate these smooth functions against the statistical complexity of estimating them. The constants hidden by $\lesssim$ depend on the fixed smoothness, dimension, output bounds, and trade-off weights; the displayed powers describe how the error decreases with sample size.

More data improves estimation of the population optimum selected by the weights. The next result connects this objective-level convergence to recovery of the representation and its predictive performance.

3.2 REPRESENTATION CONVERGENCE AND DOWNSTREAM PERFORMANCE

Small excess risk concerns an objective value, whereas representation convergence concerns a function. The latter requires accounting for the objective's invariances: rotating, reflecting, or translating a representation leaves its pairwise distances unchanged, and an inverse transformation of the head preserves predictions. We therefore compare representations after the best rigid alignment. For the equivalence class $[g]$ of g under such transformations, define

$$d_\sim([g_1], [g_2]) = \inf_{O^\top O = I,\, c \in \mathbb{R}^p} \left(\mathbb{E}\|g_1(\boldsymbol{X}) - Og_2(\boldsymbol{X}) - c\|_2^2\right)^{1/2}.$$

The matrix $O \in \mathbb{R}^{p\times p}$ represents a rotation or reflection, and c is a translation. The expectation is with respect to the marginal input distribution P_X, so the distance measures alignment over the population rather than at a single input. The notation $[g]$ groups all rigidly transformed versions

of the same representation. Thus, $d_\sim = 0$ identifies representations with the same geometry after alignment, which is the geometry relevant to the DC objective. The identifiability condition requires the population optimum to have a unique equivalence class, separated in risk from classes a fixed distance away. For each representation, its profiled risk is the smallest joint risk obtainable by varying the head. A local growth condition requires the excess of this risk to grow at least as the κth power of the alignment distance, for some $\kappa > 0$.

Theorem 3 (Representation and predictive calibration, simplified)**.** *Under orbit identifiability, $\Delta_n \to 0$ in probability implies $d_\sim([\widehat{g}_n], [g_0]) \to 0$ in probability. Under local power growth, if $\Delta_n = O_\mathbb{P}(r_n)$ for a sequence $r_n \to 0$, then*

$$d_\sim([\widehat{g}_n], [g_0]) = O_\mathbb{P}(r_n^{1/\kappa}).$$

Under the boundedness conditions, the downstream cross-entropy converges to that of (g_0, f_0), with deviation $O_\mathbb{P}(r_n + r_n^{1/\kappa})$. Formal statements are Theorems E.2 and E.3 in Appendices E.2 and E.3.

Thus, the joint estimator recovers the oracle representation up to transformations that do not affect the learning objective, while its predictive performance approaches the oracle's. The same argument controls representation-level sensitive dependence (Theorem E.4). A Bayes classification bound separates estimation error from the bias associated with the fixed trade-off (Corollary E.27). These results explain what joint learning consistently estimates and connect the representation's statistical accuracy to its downstream utility.

The exponent κ describes the local shape of the profiled objective around the oracle orbit. For quadratic growth, $\kappa = 2$, an objective error of order r_n yields representation error of order $\sqrt{r_n}$. The notation $O_\mathbb{P}(r_n)$ means that the error divided by r_n is bounded in probability. Here r_n can be chosen as the excess-risk bound from Theorem 2, including its logarithmic factors. This conversion gives a direct interpretation of how improved joint estimation sharpens representation recovery.

The connection to predictive utility can also be read directly from the calibration inequality. Writing $\mathcal{L}_{\mathrm{ce}}(g, f) = \mathbb{E}[\ell(Y, f(g(\boldsymbol{X})))]$, the formal analysis gives

$$\left|\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_{\mathrm{ce}}(g_0, f_0)\right| \le \Delta_n + 16(\lambda B_Z + \mu B_Y) d_\sim([\widehat{g}_n], [g_0]), \tag{7}$$

where B_Z and B_Y bound the norms of the sensitive vector and one-hot label. In particular, $B_Y = 1$, and a sensitive vector with d binary coordinates admits $B_Z = \sqrt{d}$. These constants make the dependence on the attribute encoding explicit. The first term is joint estimation error; the second accounts for the difference between the learned and oracle representations. This explains why representation convergence is useful beyond identifying the encoder: it separates predictive performance from an objective that also contains fairness and informativeness penalties. Both errors are measured relative to the same trade-off oracle, making their contributions to predictive calibration explicit.

For classification, this distinction has a direct interpretation. Let $\mathcal{R}_{0\text{-}1}$ be the misclassification risk, $\mathcal{R}^\star_{0\text{-}1}$ its Bayes minimum, and $\mathcal{L}^\star_{\log}$ the minimum population log-loss. Define the oracle bias $b_{\lambda,\mu} = \mathcal{L}_{\mathrm{ce}}(g_0, f_0) - \mathcal{L}^\star_{\log}$ and the predictive estimation error $e_n = |\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_{\mathrm{ce}}(g_0, f_0)|$. Corollary E.27 then gives

$$\mathcal{R}_{0\text{-}1}(\widehat{g}_n, \widehat{f}_n) - \mathcal{R}^\star_{0\text{-}1} \le \sqrt{2b_{\lambda,\mu}} + \sqrt{2e_n}. \tag{8}$$

The first term quantifies the selected population trade-off, while the second is controlled by (7). Together they translate the joint-learning analysis into an interpretable classification guarantee.

3.3 Fairness inheritance after downstream retraining

Representation-level control matters because the trained head need not be the final consumer of the features. Exact independence, marginal or conditional, passes to any measurable head. For approximate independence, the head's complexity matters: a sufficiently sensitive decision rule can amplify a small remaining difference between group distributions. The following bound quantifies that dependence.

Theorem 4 (Fairness inheritance, simplified)**.** *Let $Z \in \{0, 1\}$ with $p_z = \mathbb{P}(Z = z) > 0$. For a bounded representation $\boldsymbol{R}$, let $\mathbb{H}_{\mathrm{dist}}$ be the reproducing-kernel Hilbert space of its distance-induced*

kernel (defined in Appendix F.1). Every score function h in this space with norm at most M satisfies

$$\left|\mathbb{E}[h(\boldsymbol{R}) \mid Z=1] - \mathbb{E}[h(\boldsymbol{R}) \mid Z=0]\right| \leq \frac{M}{2p_0p_1}\sqrt{\mathrm{DC}(Z, \boldsymbol{R})}.$$

The same result applies within each target class, with the corresponding conditional group probabilities. The formal marginal and conditional results are Theorems F.1 and F.2 in Appendix F.1.

This guarantee holds simultaneously for every head in the stated RKHS ball, irrespective of its training procedure. Combining it with representation calibration yields the following explicit bound for the learned encoder $\boldsymbol{R} = \widehat{g}_n(\boldsymbol{X})$ when $\lambda > 0$:

$$\Gamma(h) \leq \frac{M}{2p_0p_1}\left[\mathrm{DC}\left(Z, g_0(\boldsymbol{X})\right) + \frac{2\Delta_n}{\lambda} + \frac{16(\lambda B_Z + 2\mu B_Y)}{\lambda} d_{\sim}([\widehat{g}_n], [g_0])\right]^{1/2}. \tag{9}$$

Here $\Gamma(h)$ is the group mean-score gap on the learned representation. Thus, if the oracle has zero sensitive dependence and the joint excess risk is $O_{\mathbb{P}}(r_n)$, the gap for any fixed norm budget M is $O_{\mathbb{P}}((r_n + r_n^{1/\kappa})^{1/2})$. The class-conditional analogue bounds TPR and FPR disparities with class-prevalence and within-class group-probability factors. For deterministic decisions, an approximation term is needed. Appendix F gives these statements and the counterexample for unrestricted heads.

4 Experiments

The principal contribution is the joint-estimation and fairness-inheritance analysis. The experiments assess competitiveness with representative fairness methods across data modalities, persistence after replacing the trained head, and the effect of the objective's components. Detailed ablations, sensitivity studies, and multi-attribute results are deferred to Appendix I so the main text can focus on the central evidence.

4.1 Datasets and evaluation protocol

We use two tabular datasets, Adult (income prediction, sensitive attribute gender) and Bank (deposit subscription, sensitive attribute age); two text datasets, BIOS (occupation prediction, gender) and MOJI (sentiment prediction, dialect group); and CelebA (facial-attribute prediction with gender, hair-color, or skin-color annotations). BIOS uses BERT features, MOJI uses DeepMoji features (Felbo et al., 2017), and CelebA uses extracted image features. Feature dimensions and method-specific architectures are detailed in Appendices H.1 and H.5.

The main benchmark tables use DFL's separation objective. Encoder outputs are clipped coordinate-wise to $[-10, 10]$ before entering both the dependence penalties and downstream head. Predictive and informativeness terms receive weight α; conditional fairness receives $1 - \alpha$. All methods use a common validation rule: minimize average TPR-RMS and MCDP subject to at least 95% of displayed STD accuracy, or select the most accurate candidate if none qualifies. DFL searches $\alpha \in \{0.1, \ldots, 0.9\}$; details are in Appendix H.

We compare with an unconstrained model (STD), four prediction-level methods (AdvDebias, Fair-Mixup, DRAlign, DiffMCDP), three representation-erasure methods (INLP, RLACE, SUP), and the conditional adversarial representation methods CFair-EO and LAFTR-EO-CE. Results report accuracy, classwise TPR-RMS (labelled TPR), and MCDP, all on a percentage scale. Zero TPR-RMS is equivalent to equalized odds for binary targets; its multiclass form summarizes diagonal class-conditional gaps. MCDP assesses marginal score disparity related to demographic parity. Parentheses give standard deviations; bold and underline mark the best and second-best fairness-aware means. Metric and baseline details are in Appendices B and H.5.

4.2 Fairness–utility comparison

Tables 2–4 compare fairness and predictive accuracy across the three data modalities; Appendix I completes the CelebA pairs. Across these 16 settings, DFL ranks first or second among fairness-aware methods in 30 of 48 metric comparisons: 11 accuracy, 10 TPR-gap, and 9 MCDP-gap comparisons. On Adult, DFL reduces MCDP from 12.46 to 7.31 and has accuracy 79.24 versus 77.67 for the

Table 2: Predictive accuracy and fairness on Adult and Bank. Parentheses give standard deviations. Among fairness-aware methods, best in **bold**, second best underlined.

Adult

Method	Accuracy ↑	TPR ↓	MCDP ↓
STD	77.67	3.55	12.46
AdvDebias	76.49(0.65)	3.27(0.28)	29.32(2.31)
FairMixup	74.48(0.43)	**3.14(0.23)**	24.91(1.58)
DRAlign	75.01(0.41)	3.66(0.17)	22.04(1.22)
DiffMCDP	73.93(0.31)	3.28(0.38)	11.50(1.09)
INLP	68.27(0.57)	3.96(0.44)	8.72(1.31)
RLACE	72.79(0.83)	3.82(0.16)	7.45(0.86)
SUP	70.53(0.36)	3.52(0.28)	8.16(1.25)
CFair-EO	77.69(0.98)	3.49(1.00)	**6.60(1.26)**
LAFTR-EO-CE	77.29(0.71)	3.36(1.62)	6.62(0.30)
DFL	**79.24(0.33)**	3.43(0.23)	7.31(0.43)

Bank

Method	Accuracy ↑	TPR ↓	MCDP ↓
STD	90.69	2.42	11.31
AdvDebias	87.73(0.84)	2.20(0.13)	17.55(1.39)
FairMixup	88.45(0.73)	2.21(0.24)	13.18(2.94)
DRAlign	86.63(0.66)	1.96(0.21)	13.61(1.23)
DiffMCDP	89.92(0.79)	2.18(0.13)	11.00(0.97)
INLP	87.85(0.63)	2.32(0.35)	10.25(0.68)
RLACE	88.32(0.67)	2.21(0.26)	8.83(0.35)
SUP	88.06(0.75)	2.46(0.21)	9.69(0.42)
CFair-EO	88.65(0.51)	**1.66(0.07)**	**8.22(1.32)**
LAFTR-EO-CE	88.82(0.44)	1.78(1.09)	10.23(2.34)
DFL	**90.55(0.67)**	2.25(0.16)	8.23(0.21)

Table 3: Predictive accuracy and fairness on BIOS and MOJI. Parentheses give standard deviations. Bold and underline mark the best and second-best fairness-aware means.

BIOS

Method	Accuracy ↑	TPR ↓	MCDP ↓
STD	79.79	15.59	6.67
AdvDebias	77.51(1.32)	13.51(1.24)	10.89(1.33)
FairMixup	75.01(0.83)	12.58(1.43)	11.72(1.29)
DRAlign	74.68(1.15)	11.36(1.31)	9.27(0.83)
DiffMCDP	75.18(1.30)	10.04(1.03)	6.42(0.76)
INLP	71.72(1.04)	9.68(0.44)	6.70(0.56)
RLACE	77.15(0.87)	13.15(0.94)	6.51(0.63)
SUP	76.36(0.59)	12.64(0.33)	6.63(0.41)
CFair-EO	78.39(1.59)	**8.87(1.13)**	**6.13(1.23)**
LAFTR-EO-CE	**78.82(0.73)**	12.44(0.15)	6.51(0.84)
DFL	77.94(0.34)	9.01(0.59)	6.24(0.51)

MOJI

Method	Accuracy ↑	TPR ↓	MCDP ↓
STD	71.19	38.95	39.86
AdvDebias	70.95(1.14)	30.41(1.46)	28.77(2.43)
FairMixup	69.71(1.21)	26.95(2.39)	23.85(2.69)
DRAlign	67.34(1.59)	21.41(1.75)	18.71(1.26)
DiffMCDP	68.55(0.83)	15.97(0.73)	14.14(0.91)
INLP	62.37(1.41)	15.28(0.64)	17.75(1.29)
RLACE	72.75(0.86)	15.21(0.47)	17.36(1.71)
SUP	69.25(0.78)	10.65(0.62)	18.35(1.43)
CFair-EO	**75.02(0.36)**	11.85(2.28)	9.99(2.90)
LAFTR-EO-CE	73.38(0.74)	12.82(4.49)	12.77(3.63)
DFL	74.59(0.34)	**10.47(0.53)**	**8.41(1.67)**

logistic-regression STD reference. On Bank, accuracy remains close to STD (90.55 versus 90.69), and MCDP falls from 11.31 to 8.23. DFL has the highest accuracy among fairness-aware methods on both tabular datasets; on Bank its MCDP gap is within 0.01 point of the smallest reported gap.

The text benchmarks show a stronger effect on both fairness metrics. On BIOS, DFL reduces the TPR gap from 15.59 to 9.01, with accuracy 77.94 compared with STD's 79.79. On MOJI, it reduces the TPR gap from 38.95 to 10.47 and MCDP from 39.86 to 8.41, while accuracy rises from 71.19 to 74.59. DFL attains the smallest TPR and MCDP gaps on MOJI, while CFair-EO has the highest accuracy there. On BIOS, CFair-EO has the smallest gaps on both metrics and DFL ranks second.

Table 4 reports the image tasks with gender as the sensitive attribute. DFL has the lowest MCDP gap on all four tasks and the lowest TPR gap for Attractive. For Wavy Hair, the TPR gap falls from 29.14 to 3.49 and MCDP from 48.13 to 18.04, with accuracy changing from 83.52 to 80.10. On Smiling, DiffMCDP has the lowest TPR gap, while DFL has the lowest MCDP gap. Additional sensitive attributes are evaluated in Appendix I.

4.3 Persistence under downstream retraining

We freeze the learned representation, discard the jointly trained head, and fit a fresh one-hidden-layer classifier using predictive loss alone. Table 5 shows that accuracy changes by at most 0.65 percentage points across the five reported datasets. Most large disparity reductions persist: the fresh head retains 98.4% of the TPR-gap reduction on MOJI, 92.2% on BIOS, and 92.2% on CelebA gender/Wavy Hair. MCDP reduction is also largely preserved on Adult, Bank, MOJI, and CelebA.

Across the tabular and text benchmarks, retraining changes the TPR gap by at most 0.51 points and MCDP by at most 0.71 points. The corresponding changes on CelebA are 2.00 and 1.50 points, relative to the substantial reductions from the original STD gaps. Table 5 reports each comparison.

Table 4: Predictive accuracy and fairness on CelebA with $\boldsymbol{Z}$ = gender. Parentheses give standard deviations. Best and second-best fairness-aware means are bold and underlined. Other sensitive attributes are in Appendix I.

$\boldsymbol{Y}$	Metric	STD	DFL	DiffMCDP	RLACE	CFair-EO	LAFTR-EO-CE
Young	Accuracy ↑	85.01	79.48(0.92)	79.15(0.74)	81.91(1.22)	**82.10(0.81)**	82.08(0.43)
	TPR Gap ↓	26.70	2.43(2.01)	12.29(1.14)	6.59(2.45)	2.70(1.43)	**2.22(1.74)**
	MCDP Gap ↓	49.74	**11.07(2.04)**	20.79(2.38)	26.47(2.78)	17.73(5.54)	11.92(3.36)
Attractive	Accuracy ↑	78.32	**75.53(1.87)**	74.79(0.85)	75.43(0.95)	74.59(0.48)	74.67(0.49)
	TPR Gap ↓	40.50	**2.89(1.31)**	18.66(1.45)	15.15(3.12)	7.84(2.34)	8.38(2.96)
	MCDP Gap ↓	57.20	**21.15(2.59)**	23.91(2.67)	30.42(3.15)	27.11(2.08)	28.39(2.44)
Smiling	Accuracy ↑	80.51	77.76(1.67)	77.81(0.68)	77.53(0.87)	**79.58(0.53)**	78.92(0.81)
	TPR Gap ↓	7.65	5.42(1.86)	**3.52(0.81)**	4.75(1.87)	5.07(0.64)	5.13(0.91)
	MCDP Gap ↓	16.63	**7.09(1.64)**	9.95(1.45)	8.84(1.96)	8.33(1.44)	8.92(2.77)
Wavy Hair	Accuracy ↑	83.52	80.10(1.08)	79.76(0.79)	80.45(1.15)	**81.27(0.67)**	80.15(0.50)
	TPR Gap ↓	29.14	3.49(2.38)	13.40(1.32)	10.08(2.83)	3.53(2.21)	**3.08(1.34)**
	MCDP Gap ↓	48.13	**18.04(2.79)**	21.12(2.42)	25.62(2.69)	19.80(4.48)	19.45(2.59)

Table 5: Persistence of fairness gains after downstream retraining. "DFL head" uses the jointly trained classifier $f_{\hat{\phi}}$; "fresh head" discards it and trains an unconstrained one-hidden-layer classifier on the frozen representation $g_{\hat{\theta}}$. δ is (fresh head − DFL head). The CelebA row uses $\boldsymbol{Z}$ = gender, $\boldsymbol{Y}$ = Wavy Hair; the remaining CelebA cells are in Appendix I. Means over 20 replications.

	Accuracy ↑			**TPR gap ↓**			**MCDP gap ↓**		
Dataset	STD	DFL head	fresh head (δ)	STD	DFL head	fresh head (δ)	STD	DFL head	fresh head (δ)
Adult	77.67	79.24	79.32 (+0.08)	3.55	3.43	3.47 (+0.04)	12.46	7.31	7.52 (+0.21)
Bank	90.69	90.55	90.13 (−0.42)	2.42	2.25	2.73 (+0.48)	11.31	8.23	8.66 (+0.43)
BIOS	79.79	77.94	77.83 (−0.11)	15.59	9.01	9.52 (+0.51)	6.67	6.24	6.83 (+0.59)
MOJI	71.19	74.59	74.16 (−0.43)	38.95	10.47	10.93 (+0.46)	39.86	8.41	9.12 (+0.71)
CelebA	83.52	80.10	79.45 (−0.65)	29.14	3.49	5.49 (+2.00)	48.13	18.04	19.54 (+1.50)

These results demonstrate empirical persistence after head replacement; they do not evaluate the numerical bound of Theorem 4, as explained in Appendix H.5.

4.4 What drives the fairness–utility trade-off?

On BIOS, representation- and output-level DC penalties share the predictive loss. At the same tuning, the representation penalty achieves 77.85% accuracy versus 64.78%, with MCDP gaps 6.52 versus 7.76; the corresponding TPR gaps are 8.97 and 5.12. Appendix I.4 reports the full ablation. Further sweeps vary the label-informativeness term, trade-off weight, encoder capacity, and sensitive attributes (Appendices I.5 and I.3). They show how DFL's tunable components shape the observed fairness–utility trade-off.

5 Conclusion

DFL jointly learns predictive representations and downstream classifiers while controlling sensitive dependence at the representation level. Its marginal and conditional formulations address independence and separation within one framework. We prove a non-asymptotic joint excess-risk rate; under identifiability and local growth, it yields representation recovery up to rigid transformations and predictive calibration to the trade-off oracle. A uniform inheritance bound controls the group score disparity of suitable replacement heads by representation dependence. For the learned encoder, the end-to-end guarantee adds oracle dependence and finite-sample estimation error, with a class-conditional counterpart for opportunity gaps. The bound also makes the roles of replacement-head complexity and group prevalence explicit. Across three data modalities, DFL ranks first or second in 30 of 48 reported fairness-aware accuracy and gap comparisons, and substantial gains persist after downstream retraining. Together, the guarantees and empirical results establish a foundation for learning fair, predictive representations and reusing them with downstream classifiers.

ETHICS STATEMENT

We study group fairness using public benchmarks with sensitive demographic annotations; no new human-subject data were collected. These annotations and group-level metrics simplify social concepts and do not establish fairness for every individual or subgroup. Deployment requires assessment of the application, population, and downstream predictor beyond the benchmark guarantees.

REPRODUCIBILITY STATEMENT

Appendix C specifies the estimators and objectives; Appendices D, E, and F give the proofs. Appendix H documents architectures, hyperparameters, and implementation details; Appendix I reports additional experimental results.

AI USE STATEMENT

Generative AI tools assisted with manuscript editing, notation and cross-reference checks. We have not used generative AI tools to generate synthetic datasets, implement the proposed methods, clean or reformat datasets, translate research material, or conduct qualitative or thematic data analysis; the remaining required disclosure tasks are not applicable to this work. The authors are responsible for the final text, mathematical claims, experimental results, and completeness of this disclosure.

REFERENCES

Kofi P Adragni and R Dennis Cook. Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405, 2009.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, 2023.

Richard A Berk, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Fair risk algorithms. *Annual Review of Statistics and Its Application*, 10(1):165–187, 2023.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1119–1130, 2016.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, 2024.

YinFeng Chen, YuLing Jiao, Rui Qiu, and Zhou Yu. Deep nonlinear sufficient dimension reduction. *The Annals of Statistics*, 52(3):1201–1226, 2024.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021.

R Dennis Cook and Bing Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474, 2002.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128, 2019.

Sepehr Dehdashtian, Bashir Sadeghi, and Vishnu Naresh Boddeti. Utility-fairness trade-offs and how to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12037–12046, 2024.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations*, 2016.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 11–21, 2018.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1615–1625, 2017. doi: 10.18653/v1/D17-1169.

Vincent Grari, Oualid El Hajouji, Sylvain Lamprier, and Marcin Detyniecki. Learning unbiased representations via rényi minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 749–764. Springer, 2021.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Umang Gupta, Aaron Ferber, Bistra Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7610–7619, 2021.

Xudong Han, Timothy Baldwin, and Trevor Cohn. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2760–2765, 2021.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Wassily Hoeffding. The strong law of large numbers for U-statistics. *Institute of Statistics Mimeograph Series*, 302, 1961.

Tailen Hsing and Haobo Ren. An rkhs formulation of the inverse regression dimension-reduction problem. *Ann. Statist.*, 37(1):726–755, 2009.

Jian Huang, Yuling Jiao, Xu Liao, Jin Liu, and Zhou Yu. Deep dimension reduction for supervised representation learning. *IEEE Transactions on Information Theory*, 2024.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691–716, 2023.

Jinqiu Jin, Haoxuan Li, and Fuli Feng. On the maximal local disparity of fairness-aware classifiers. In *Forty-first International Conference on Machine Learning*, 2024.

Nikola Jovanović, Mislav Balunovic, Dimitar Iliev Dimitrov, and Martin Vechev. FARE: Provably fair representation learning with practical certificates. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15401–15420, 2023.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pp. 35–50. Springer, 2012.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207, 1996.

Kuang Yao Lee, Bing Li, and Francesca Chiaromonte. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *Annals of Statistics*, 41(1):221–249, 2013.

Tianlin Li, Qing Guo, Aishan Liu, Mengnan Du, Zhiming Li, and Yang Liu. Fairer: fairness as decision rationale alignment. In *International Conference on Machine Learning*, pp. 19471–19489. PMLR, 2023.

Ji Liu, Zenan Li, Yuan Yao, Feng Xu, Xiaoxing Ma, Miao Xu, and Hanghang Tong. Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1088–1097, 2022. doi: 10.1145/3534678.3539302.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. In *International Conference on Learning Representations*, 2016.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393, 2018.

Daniel McNamara, Cheng Soon Ong, and Robert C. Williamson. Costs and benefits of fair representation learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 263–270, 2019.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Luca Oneto, Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer, and Massimiliano Pontil. Exploiting MMD and sinkhorn divergences for fair and transferable representation learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Eric Potash, Joe Brew, Alexander Loewi, Subhabrata Majumdar, Andrew Reece, Joe Walsh, Eric Rozier, Emile Jorgenson, Raed Mansour, and Rayid Ghani. Predictive modeling for public health: Preventing childhood lead poisoning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2039–2047, 2015.

Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8227–8236, 2019.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, 2020.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pp. 18400–18421. PMLR, 2022a.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6034–6055, 2022b.

Bashir Sadeghi, Sepehr Dehdashtian, and Vishnu Boddeti. On characterizing the trade-off in invariant representation learning. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=3gfpBR1ncr.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013.

Enze Shi, Lei Ding, Linglong Kong, and Bei Jiang. Debiasing with sufficient projection: A general theoretical framework for vector representations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5960–5975, 2024.

Enze Shi, Pankaj Bhagwat, Zhixian Yang, Linglong Kong, and Bei Jiang. Understanding fairness and prediction error through subspace decomposition and influence analysis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Gábor J. Székely and Maria L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, 2009.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Gábor J Székely and Maria L Rizzo. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, pp. 2382–2412, 2014.

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769, 2007.

Zilong Tan, Samuel Yeom, Matt Fredrikson, and Ameet Talwalkar. Learning fair representations for kernel models. In *International Conference on Artificial Intelligence and Statistics*, pp. 155–166. PMLR, 2020.

Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 17 (3):1–27, 2023.

Yi-Ren Yeh, Su-Yun Huang, and Yuh-Jye Lee. Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE transactions on Knowledge and Data Engineering*, 21(11):1590–1603, 2008.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20 (75):1–42, 2019.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Han Zhao and Geoffrey J. Gordon. Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research*, 23(57):1–26, 2022.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. Conditional learning of fair representations. In *International Conference on Learning Representations*, 2020.

APPENDIX

Appendix A reviews related work. Subsequent sections give the metrics, estimators, proofs, experimental details, and additional results.

A RELATED WORK

Output-level fairness. In-processing methods add fairness constraints or penalties to predictor training (Zafar et al., 2019; Berk et al., 2023), targeting demographic parity (Dwork et al., 2012; Kamishima et al., 2012), equality of opportunity (Hardt et al., 2016), or related criteria. Approaches include adversarial training (Zhang et al., 2018; Han et al., 2021), interpolation (Chuang & Mroueh, 2021), distributional alignment (Li et al., 2023), and differentiable disparity surrogates (Jin et al., 2024). These methods control the predictions of the fitted model. DFL instead regularizes the representation consumed by the predictor, enabling analysis of fairness after that predictor is replaced.

Fair representations and concept erasure. Fair representation learning includes early distribution-matching formulations (Zemel et al., 2013), adversarial censoring (Edwards & Storkey, 2016; Madras et al., 2018; Elazar & Goldberg, 2018), variational approaches with maximum mean discrepancy (Louizos et al., 2016), and information-theoretic control (Gupta et al., 2021). Post-hoc concept erasure modifies existing features: INLP iteratively removes linearly predictive directions (Ravfogel et al., 2020), RLACE uses a linear adversarial formulation (Ravfogel et al., 2022a), LEACE provides closed-form linear erasure (Belrose et al., 2023), and SUP removes a sufficient-dimension-reduction subspace (Shi et al., 2024). Kernel extensions address nonlinear structure (Ravfogel et al., 2022b). DFL learns a nonlinear encoder jointly with the task predictor, allowing the predictive loss to guide which features are retained.

Dependence-minimizing encoders. The closest approaches directly reduce dependence between representations and sensitive attributes. CFair uses class-conditional adversarial alignment and balanced error to target equalized odds and accuracy parity (Zhao et al., 2020); Grari et al. (2021) minimize adversarially estimated Hirschfeld–Gebelein–Rényi maximal correlation. Kernel approaches offer complementary constructions. Tan et al. (2020) identify an RKHS subspace whose cross-covariance with sensitive attributes vanishes, providing fairness for predictors within that subspace. Sadeghi et al. (2022) optimize an RKHS-linear encoder in closed form using dependence statistics for utility and invariance, while Dehdashtian et al. (2024) estimate achievable trade-offs to audit representations. Quadrianto et al. (2019) use HSIC to learn fair translations in the original data domain. Distance covariance is related to HSIC through distance-induced kernels (Sejdinovic et al., 2013). DFL uses a directly computed distance statistic together with predictive loss, and analyzes estimation of the nonlinear encoder and quantitative fairness inheritance for subsequently fitted heads. The feasibility of fair representations is also subject to fundamental accuracy–fairness limits (Zhao & Gordon, 2022; McNamara et al., 2019).

Statistical guarantees and downstream reuse. LAFTR combines theoretical downstream-fairness guarantees with experiments on new predictors and tasks (Madras et al., 2018). Multitask MMD/Sinkhorn methods derive transfer bounds (Oneto et al., 2020), and FARE gives finite-sample fairness certificates for restricted encoders (Jovanović et al., 2023). These results establish important precedents for theoretical and empirical analysis of fair representations. DFL's novelty concerns the combined analysis of a jointly estimated nonlinear encoder and task head, representation convergence modulo objective invariances, and quantitative inheritance under its distance-covariance penalty.

FairDisCo introduces a distance-based dependence criterion in a variational representation-learning framework (Liu et al., 2022). DFL adds a task-coupled joint estimator and follows it through finite-sample risk, representation recovery, and downstream fairness inheritance.

Prediction-level penalties and two-stage representation learning. The same dependence measure can regularize a fitted prediction, or train a representation before a separate predictor is fitted. In DFL, the representation and head solve one task-aware optimization problem. Within shared feasible classes, the global joint optimum has training objective no larger than that of any fixed two-stage candidate. The BIOS ablations in Section 4.4 isolate the effect of penalty location and the predictive-loss term.

B EVALUATION METRICS

Let $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$ be the one-hot true label and predicted class decision, respectively, and let $q_j \in [0,1]$ be the predicted probability of class j. The sensitive attribute is binary, $\boldsymbol{Z} \in \{0,1\}$. TPR-RMS uses class decisions; MCDP uses probability scores.

The TPR gap for class j is defined as:

$$\text{TPR}_{z,j} = P(\hat{\boldsymbol{Y}}_j = 1 \mid \boldsymbol{Z} = z, \boldsymbol{Y}_j = 1), \quad \text{TPR}_{gap,j} = \text{TPR}_{1,j} - \text{TPR}_{0,j}.$$

The TPR gap across K classes is then computed as:

$$\text{TPR}_{gap} = \sqrt{\frac{1}{K}\sum_{j=1}^{K} \text{TPR}_{gap,j}^2} \times 100\%.$$

The MCDP metric is defined as:

$$\text{MCDP}_j = \max_{y\in[0,1]} \big|F_{1,j}(y) - F_{0,j}(y)\big|, \quad \text{MCDP}_{gap} = \sqrt{\frac{1}{K-1}\sum_{j=1}^{K-1} \text{MCDP}_j^2} \times 100\%,$$

where $F_{z,j}(y) = P(q_j \le y \mid \boldsymbol{Z} = z)$ is the group-conditional score CDF. We follow the benchmark convention of averaging the first $K-1$ class-score gaps. For binary targets, either probability column gives the same exact CDF supremum. Equality of the two group score distributions implies demographic parity for every fixed score threshold. For multiclass argmax decisions, matching each marginal score distribution does not by itself match the joint score distribution or guarantee decision parity. The tabular STD reference and CFair-EO/LAFTR-EO-CE evaluations use a 100-bin approximation on $[0,1]$.

B.1 SEPARATION METRICS

The theory of Section 3.3 controls the true-positive-rate gap, the false-positive-rate gap and their maximum. For binary $\boldsymbol{Y}$ and binary $\boldsymbol{Z}$ these are

$$\Delta_{\text{TPR}} = \big|\mathbb{P}(\widehat{\boldsymbol{Y}}{=}1 \mid \boldsymbol{Y}{=}1, \boldsymbol{Z}{=}1) - \mathbb{P}(\widehat{\boldsymbol{Y}}{=}1 \mid \boldsymbol{Y}{=}1, \boldsymbol{Z}{=}0)\big|,$$
$$\Delta_{\text{FPR}} = \big|\mathbb{P}(\widehat{\boldsymbol{Y}}{=}1 \mid \boldsymbol{Y}{=}0, \boldsymbol{Z}{=}1) - \mathbb{P}(\widehat{\boldsymbol{Y}}{=}1 \mid \boldsymbol{Y}{=}0, \boldsymbol{Z}{=}0)\big|,$$
$$\Delta_{\text{EO}} = \max\{\Delta_{\text{TPR}}, \Delta_{\text{FPR}}\}, \qquad \Delta_{\text{EO}}^{\text{sum}} = \Delta_{\text{TPR}} + \Delta_{\text{FPR}}.$$

Separation holds if and only if $\Delta_{\text{EO}} = 0$. For K-class problems, replacing $\mathbb{P}(\widehat{\boldsymbol{Y}}_j{=}1 \mid \boldsymbol{Y}_j{=}1, \boldsymbol{Z}{=}z)$ by its class-j analogue and taking the RMS across j, as in the TPR-RMS definition above, gives the multiclass versions.

For binary targets, the class-0 TPR gap is the negative of the class-1 FPR gap, so $\text{TPR-RMS} = \sqrt{(\Delta_{\text{TPR}}^2 + \Delta_{\text{FPR}}^2)/2} \times 100\%$. It is zero if and only if equalized odds holds, although it is a different aggregation from the maximum Δ_{EO}. For multiclass targets, the reported TPR-RMS summarizes diagonal entries of the class-conditional confusion matrices and does not assess every off-diagonal disparity. MCDP is a marginal score-distribution metric for demographic parity, not a class-conditional separation metric.

C OBJECTIVES, ESTIMATORS AND NOTATION

C.1 DISTANCE-COVARIANCE ESTIMATORS

We use DC for squared distance covariance and

$$\text{dcov}(\boldsymbol{U}, \boldsymbol{V}) := \sqrt{\text{DC}(\boldsymbol{U}, \boldsymbol{V})}. \tag{C.1}$$

Definition (4) is the moment form of the squared distance covariance. It is equivalent to the characteristic-function form that makes the characterization of independence transparent: with

$\psi_U, \psi_V, \psi_{U,V}$ the characteristic functions of $\boldsymbol{U} \in \mathbb{R}^d$, $\boldsymbol{V} \in \mathbb{R}^q$ and the pair,

$$\mathrm{DC}(\boldsymbol{U}, \boldsymbol{V}) = \int_{\mathbb{R}^{d+q}} \frac{\left|\psi_{U,V}(\boldsymbol{t}, \boldsymbol{s}) - \psi_U(\boldsymbol{t})\psi_V(\boldsymbol{s})\right|^2}{c_d c_q \|\boldsymbol{t}\|^{d+1} \|\boldsymbol{s}\|^{q+1}} \, d\boldsymbol{t} \, d\boldsymbol{s}, \qquad c_k = \frac{\pi^{(k+1)/2}}{\Gamma((k+1)/2)}, \tag{C.2}$$

from which $\mathrm{DC}(\boldsymbol{U}, \boldsymbol{V}) \geq 0$ with equality if and only if $\psi_{U,V} = \psi_U \psi_V$ almost everywhere, i.e. if and only if $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{V}$ (Székely et al., 2007).

Given n i.i.d. copies $\{(Z_i, R_i)\}_{i=1}^n$ of $(\boldsymbol{Z}, \boldsymbol{R})$, the unbiased U-statistic estimator of the squared distance covariance (4) is

$$\widehat{\mathrm{DC}}_n(\boldsymbol{Z}, \boldsymbol{R}) = \binom{n}{4}^{-1} \sum_{1 \leq i_1 < i_2 < i_3 < i_4 \leq n} h\big((Z_{i_1}, R_{i_1}), \ldots, (Z_{i_4}, R_{i_4})\big), \tag{C.3}$$

with the symmetric kernel

$$\begin{aligned} h\big((Z_1, R_1), \ldots, (Z_4, R_4)\big) = \frac{1}{4} \sum_{\substack{1 \leq i,j \leq 4 \\ i \neq j}} \|Z_i - Z_j\| \, \|R_i - R_j\| \\ + \frac{1}{24} \Big(\sum_{\substack{1 \leq i,j \leq 4 \\ i \neq j}} \|Z_i - Z_j\| \Big) \Big(\sum_{\substack{1 \leq k,l \leq 4 \\ k \neq l}} \|R_k - R_l\| \Big) \\ - \frac{1}{4} \sum_{i=1}^{4} \sum_{\substack{1 \leq j \leq 4 \\ j \neq i}} \|Z_i - Z_j\| \sum_{\substack{1 \leq j \leq 4 \\ j \neq i}} \|R_i - R_j\|, \end{aligned}$$

where $\|\cdot\|$ is the Euclidean norm. The class-conditional estimator aggregates within-class estimates,

$$\widehat{\mathrm{DC}}_n(\boldsymbol{Z}, \boldsymbol{R} \mid \boldsymbol{Y}) = \sum_{k : n_k \geq 4} w_{n,k} \, \widehat{\mathrm{DC}}_{n_k}(\boldsymbol{Z}, \boldsymbol{R}), \qquad w_{n,k} = \binom{n_k}{4} \Big/ \sum_{j=1}^{K} \binom{n_j}{4}, \tag{C.4}$$

computed on $S_k = \{i : Y_i = k\}$ with $n_k = |S_k|$, provided $\sum_j \binom{n_j}{4} > 0$. Classes with $n_k < 4$ have zero empirical weight and do not contribute to this finite-sample statistic. These are the U-statistic weights used in our implementation. A balanced alternative would give equal weight to all classes, but would require at least four observations in every included class; we did not use it.

The consistency of both estimators is stated in Theorem 1.

C.2 Empirical objectives

The empirical objective under the independence criterion is

$$\mathcal{L}_n(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(Y_i, f_\phi(g_\theta(X_i))\big) + \lambda \, \widehat{\mathrm{DC}}_n\big(\boldsymbol{Z}, g_\theta(\boldsymbol{X})\big) - \mu \, \widehat{\mathrm{DC}}_n\big(\mathbf{e}_{\boldsymbol{Y}}, g_\theta(\boldsymbol{X})\big), \tag{C.5}$$

and under the separation criterion

$$\mathcal{L}_n(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(Y_i, f_\phi(g_\theta(X_i))\big) + \lambda \, \widehat{\mathrm{DC}}_n\big(\boldsymbol{Z}, g_\theta(\boldsymbol{X}) \mid \boldsymbol{Y}\big) - \mu \, \widehat{\mathrm{DC}}_n\big(\mathbf{e}_{\boldsymbol{Y}}, g_\theta(\boldsymbol{X})\big). \tag{C.6}$$

Both are minimized jointly over (θ, ϕ) by stochastic gradient descent. The reported experiments use the separation form, in the single-parameter α form of (H.1).

C.3 Function-space notation

For fixed (λ, μ) we write the population objective directly over the function pair (g, f),

$$\mathcal{L}(g, f) := \mathbb{E}\big[\ell\big(\boldsymbol{Y}, f(g(\boldsymbol{X}))\big)\big] + \lambda \, \mathrm{DC}\left(\boldsymbol{Z}, g(\boldsymbol{X})\right) - \mu \, \mathrm{DC}\left(\mathbf{e}_Y, g(\boldsymbol{X})\right), \tag{C.7}$$

with empirical version $\mathcal{L}_n(g, f)$. The bounded network classes and joint empirical minimizer are specified in Appendix E.1, equation (E.1). We write $(g_0, f_0) = (g_{\lambda,\mu}, f_{\lambda,\mu})$ for the trade-off oracle over bounded measurable pairs.

Definition C.1 (Hölder class)**.** Let $\beta = k + a$ with $k \in \mathbb{N}_0$ and $a \in (0, 1]$. The Hölder class $H^\beta(\mathcal{D}, B)$ consists of functions $h : \mathcal{D} \to \mathbb{R}$ whose partial derivatives $\partial^\gamma h$ exist for all $\|\gamma\|_1 \le k$ and satisfy

$$\max_{\|\gamma\|_1 \le k} \|\partial^\gamma h\|_\infty \le B, \qquad \max_{\|\gamma\|_1 = k} \sup_{x \ne x'} \frac{|\partial^\gamma h(x) - \partial^\gamma h(x')|}{\|x - x'\|_2^a} \le B.$$

We write L_{f_0} for the Hölder constant implied by $f_0 \in H^{\beta_f}(\cdot, B_{f0})$.

Standard asymptotic notation is used throughout: $a \wedge b = \min\{a, b\}$, $a \vee b = \max\{a, b\}$; $a_n = \mathcal{O}(b_n)$, $a_n \lesssim b_n$ and $b_n \gtrsim a_n$ all mean $\limsup_n |a_n/b_n| < \infty$; and $a_n \asymp b_n$ means both directions hold. For a multi-index $\alpha \in \mathbb{N}_0^d$, $\partial^\alpha h$ denotes the corresponding partial derivative, and $\|f\|_\infty$ the sup-norm.

D Proofs for Section 2

Throughout the analysis, $\|g\|_\infty = \operatorname{ess\,sup}_x \|g(x)\|_2$. The coordinatewise clipped experimental representations satisfy $B_g = 10\sqrt{p}$, as described in Appendix H. The theoretical bounded classes allow a general fixed B_g.

Proposition D.1 (Unbounded representations make the objective ill-posed)**.** *Let* $\mathcal{L}(g, f) = \mathbb{E}[\ell(\boldsymbol{Y}, f(g(\boldsymbol{X})))] + \lambda \operatorname{DC}(\boldsymbol{Z}, g(\boldsymbol{X})) - \mu \operatorname{DC}(\mathbf{e}_Y, g(\boldsymbol{X}))$ *with* $\lambda, \mu \ge 0$*. Suppose there exists a measurable* g *with* $\lambda \operatorname{DC}(\boldsymbol{Z}, g(\boldsymbol{X})) - \mu \operatorname{DC}(\mathbf{e}_Y, g(\boldsymbol{X})) < 0$*. Then* $\inf_{g,f} \mathcal{L}(g, f) = -\infty$*, where the infimum is over measurable pairs.*

D.1 Proof of Proposition D.1

Proof of Proposition D.1. Inspect (4): every term is a product of one factor $\|\boldsymbol{U} - \boldsymbol{U}'\|$ and one factor $\|\boldsymbol{V} - \boldsymbol{V}'\|$ (or $\|\boldsymbol{V} - \boldsymbol{V}''\|$), so replacing $\boldsymbol{V}$ by $t\boldsymbol{V}$ for $t > 0$ multiplies each term by t and hence $\operatorname{DC}(\boldsymbol{U}, t\boldsymbol{V}) = t \operatorname{DC}(\boldsymbol{U}, \boldsymbol{V})$.

Let g be measurable with $c := \mu \operatorname{DC}(\mathbf{e}_Y, g(\boldsymbol{X})) - \lambda \operatorname{DC}(\boldsymbol{Z}, g(\boldsymbol{X})) > 0$, and let f be any measurable head. For $t > 0$ set $g_t := t\, g$ and $f_t := f(\cdot/t)$. Then $f_t \circ g_t = f \circ g$ pointwise, so the cross-entropy term of $\mathcal{L}$ is unchanged, while by 1-homogeneity

$$\lambda \operatorname{DC}(\boldsymbol{Z}, g_t(\boldsymbol{X})) - \mu \operatorname{DC}(\mathbf{e}_Y, g_t(\boldsymbol{X})) = t\big(\lambda \operatorname{DC}(\boldsymbol{Z}, g(\boldsymbol{X})) - \mu \operatorname{DC}(\mathbf{e}_Y, g(\boldsymbol{X}))\big) = -tc.$$

Hence $\mathcal{L}(g_t, f_t) = \mathbb{E}[\ell(\boldsymbol{Y}, f(g(\boldsymbol{X})))] - tc \to -\infty$ as $t \to \infty$. □

The pathology is not exotic. Any g that is more informative about $\boldsymbol{Y}$ than about $\boldsymbol{Z}$, relative to the ratio λ/μ, triggers it, and gradient descent moves toward it by inflating the representation scale — a direction along which the fairness penalty grows but the objective still falls.

Proposition D.2 (The bounded objective is bounded below)**.** *Suppose* $\|g\|_\infty := \operatorname{ess\,sup}_x \|g(x)\|_2 \le B_g$*. Then*

$$\operatorname{DC}\left(\mathbf{e}_{\boldsymbol{Y}}, g(\boldsymbol{X})\right) \;\le\; 2\sqrt{2}\, B_g, \qquad \textit{hence} \qquad \mathcal{L}(g, f) \;\ge\; -2\sqrt{2}\, \mu B_g$$

for every f*, uniformly over* $\lambda \ge 0$*.*

Proof. Write $\boldsymbol{V} = g(\boldsymbol{X})$ and let $\boldsymbol{V}'$ be an independent copy. Conditioning on $\boldsymbol{V}$ and applying Jensen's inequality, $\mathbb{E}\big[\|\boldsymbol{V} - \boldsymbol{V}'\|\, \|\boldsymbol{V} - \boldsymbol{V}''\|\big] = \mathbb{E}\big[(\mathbb{E}[\|\boldsymbol{V} - \boldsymbol{V}'\| \mid \boldsymbol{V}])^2\big] \ge (\mathbb{E}\|\boldsymbol{V} - \boldsymbol{V}'\|)^2$, so from (4) with $\boldsymbol{U} = \boldsymbol{V}$,

$$\operatorname{DC}(\boldsymbol{V}, \boldsymbol{V}) \le \mathbb{E}\|\boldsymbol{V} - \boldsymbol{V}'\|^2 - (\mathbb{E}\|\boldsymbol{V} - \boldsymbol{V}'\|)^2 \le \mathbb{E}\|\boldsymbol{V} - \boldsymbol{V}'\|^2 \le (2B_g)^2.$$

The same computation with $\boldsymbol{V} = \mathbf{e}_{\boldsymbol{Y}}$, using $\|\mathbf{e}_y - \mathbf{e}_{y'}\| \le \sqrt{2}$, gives $\operatorname{DC}(\mathbf{e}_{\boldsymbol{Y}}, \mathbf{e}_{\boldsymbol{Y}}) \le 2$. The Cauchy–Schwarz inequality for distance covariance (Székely et al., 2007, Theorem 3), $\operatorname{DC}(\boldsymbol{U}, \boldsymbol{V}) \le \sqrt{\operatorname{DC}(\boldsymbol{U}, \boldsymbol{U})}\sqrt{\operatorname{DC}(\boldsymbol{V}, \boldsymbol{V})}$, then yields $\operatorname{DC}(\mathbf{e}_{\boldsymbol{Y}}, g(\boldsymbol{X})) \le \sqrt{2} \cdot 2B_g$. Since the cross-entropy term and $\lambda \operatorname{DC}(\boldsymbol{Z}, g(\boldsymbol{X}))$ are nonnegative, $\mathcal{L}(g, f) \ge -2\sqrt{2}\, \mu B_g$. □

D.2 A SCALE-FREE VARIANT

An alternative repair replaces every distance covariance by the squared distance correlation

$$\varrho^2(\boldsymbol{U},\boldsymbol{V}) := \frac{\mathrm{DC}(\boldsymbol{U},\boldsymbol{V})}{\sqrt{\mathrm{DC}(\boldsymbol{U},\boldsymbol{U})\ \mathrm{DC}(\boldsymbol{V},\boldsymbol{V})}} \in [0,1], \qquad \mathrm{DC}(\boldsymbol{U},\boldsymbol{U})\,\mathrm{DC}(\boldsymbol{V},\boldsymbol{V}) > 0, \tag{D.1}$$

giving $\mathcal{L}^{\varrho}(g,f) = \mathbb{E}[\ell(\boldsymbol{Y}, f(g(\boldsymbol{X})))] + \lambda\varrho^2(\boldsymbol{Z}, g(\boldsymbol{X})) - \mu\varrho^2(\mathbf{e}_Y, g(\boldsymbol{X}))$.

Proposition D.3. *ϱ^2 is invariant under $g \mapsto cOg + b$ for any $c > 0$, $O \in \mathbb{O}(p)$, $b \in \mathbb{R}^p$; $\varrho^2(\boldsymbol{U},\boldsymbol{V}) = 0$ if and only if $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{V}$; and $\mathcal{L}^{\varrho}(g,f) \geq -\mu$ for all (g,f), with no boundedness assumption on g.*

Proof. Invariance follows from the 1-homogeneity established above, which scales numerator and denominator identically, together with the rigid-motion invariance of DC (Lemma E.18). The characterization of independence is that of DC. The lower bound follows from $\varrho^2 \in [0,1]$, which is the Cauchy–Schwarz inequality used in Proposition D.2, and the nonnegativity of the cross-entropy term. □

The normalizer must also be estimated from each minibatch, so this alternative introduces a ratio estimator. The reported experiments use the DC objective with coordinatewise clipped encoder outputs. We include the scale-free variant as an alternative way of controlling scale; its denominator depends on the representation distribution and changes the optimization problem.

E PROOFS FOR SECTION 3

This section states and proves Theorems E.1, E.4 and E.3, together with Theorem 1. Appendix E.6 bounds the joint excess risk; Appendix E.7 develops the quotient-space results stated in Section 3.2; Appendices E.8 and E.9 then split the excess-risk bound into the population fairness and utility statements.

Throughout, $\mathcal{L}_{\mathrm{ce}}(g,f) := \mathbb{E}[\ell(\boldsymbol{Y}, f(g(\boldsymbol{X})))]$ and $\mathcal{L}_{\mathrm{DC}}(g) := \lambda\ \mathrm{DC}(\boldsymbol{Z}, g(\boldsymbol{X})) - \mu\ \mathrm{DC}(\mathbf{e}_Y, g(\boldsymbol{X}))$, so that $\mathcal{L} = \mathcal{L}_{\mathrm{ce}} + \mathcal{L}_{\mathrm{DC}}$, and $\Delta_n := \mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}(g_0, f_0)$.

E.1 FORMAL SETUP AND JOINT EXCESS RISK

Let $\sigma(t) = \max\{t, 0\}$ and let $\mathcal{NN}^{(r\to s)}_{D,W,S,B}$ denote ReLU feedforward networks $\mathbb{R}^r \to \mathbb{R}^s$ of depth D, width $\leq W$, at most S parameters, and sup-norm at most B. Set $\mathcal{G}_n = \mathcal{NN}^{(p\to p)}_{D_g,W_g,S_g,B_g}$ and $\mathcal{F}_n = \mathcal{NN}^{(p\to K)}_{D_f,W_f,S_f,B_f}$, where the output norm is Euclidean and $\|g\|_\infty = \operatorname{ess\,sup}_x \|g(x)\|_2$. Write $\mathcal{L}(g,f)$ for the population objective (5) viewed as a function of the pair, and $\mathcal{L}_n$ for its empirical version. The joint empirical minimizer is:

$$(\widehat{g}_n, \widehat{f}_n) \in \arg\min_{(g,f)\in\mathcal{G}_n\times\mathcal{F}_n} \mathcal{L}_n(g,f). \tag{E.1}$$

Let $(g_0, f_0) = (g_{\lambda,\mu}, f_{\lambda,\mu})$ minimize $\mathcal{L}$ over bounded measurable pairs; we call it the *trade-off oracle* at (λ,μ), and $\Delta_n := \mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}(g_0, f_0) \geq 0$ the joint excess risk.

Assumption E.1 (Bounded data)**.** $\mathrm{supp}(\boldsymbol{X}) \subseteq [0,1]^p$, $\|\boldsymbol{Z}\| \leq B_Z$ and $\|\mathbf{e}_{\boldsymbol{Y}}\| \leq B_Y$ almost surely.

Assumption E.2 (Oracle smoothness)**.** Each coordinate of g_0 lies in the Hölder class $H^{\beta_g}([0,1]^p, B_{g0})$ of Definition C.1 (Appendix C) and each coordinate of f_0 lies in $H^{\beta_f}([-B_g, B_g]^p, B_{f0})$, for some $\beta_g, \beta_f > 0$.

Theorem E.1 (Non-asymptotic joint excess risk)**.** *For positive integers M_g, N_g, M_f, N_f, take $D_g \asymp M_g \log(8M_g)$, $W_g \asymp N_g \log(8N_g)$, $D_f \asymp M_f \log(8M_f)$, and $W_f \asymp N_f \log(8N_f)$. Under*

Assumptions E.1–E.2, for sufficiently large n,

$$\mathbb{E}[\Delta_n] \le C_1\Big\{(M_f N_f)^{-2\beta_f/p} + (M_g N_g)^{-2s_f\beta_g/p} + (M_g N_g)^{-2\beta_g/p}\Big\} + C_2 B_\ell \sqrt{\frac{(M_g+M_f)(N_g^2 M_g + N_f^2 M_f)\log(2+N_g^2 M_g + N_f^2 M_f)\log^5 n}{n}} + C_2 B_g(\lambda B_Z + \mu B_Y)\sqrt{\frac{M_g^2 N_g^2 \log(2+N_g^2 M_g)\log^5 n}{n}}. \tag{E.2}$$

where $s_f = \min\{\beta_f, 1\}$, $B_\ell = \log K + 2B_f$, and C_1, C_2 depend only on $(\beta_g, \beta_f, p, K, B_g, B_f, B_Z, B_Y, \lambda, \mu, L_{f_0}, B_{g0}, B_{f0})$. Taking M_g, M_f fixed and choosing $N_g \asymp n^{p/(4\beta_g+2p)}$ and $N_f \asymp n^{p/(4\beta_f+2p)}$ balances the terms as detailed in Appendix E.6 and yields (E.3) *up to polylogarithmic factors.*

$$\mathbb{E}[\Delta_n] \lesssim n^{-\beta_f/(2\beta_f+p)} + n^{-s_f\beta_g/(2\beta_g+p)} + n^{-\beta_g/(2\beta_g+p)}, \tag{E.3}$$

where the rate is understood up to polylogarithmic factors. In subsequent rate conversions, r_n denotes the upper bound including those factors from (E.2)*.*

For an oracle head with at least Lipschitz regularity ($\beta_f \ge 1$), the representation terms have the same rate. The result describes the statistical performance of an empirical minimizer; it does not certify that stochastic gradient descent reaches that minimizer.

E.2 Quotient-space calibration of the learned representation

Rigid motions of the representation preserve pairwise distances and can be absorbed by the head. The appropriate estimation target is therefore the oracle representation up to rotation, reflection, and translation. Let P_X denote the marginal distribution of $\boldsymbol{X}$, and let $\mathbb{H} = L^2(P_X; \mathbb{R}^p)$, with $\|g\|_{\mathbb{H}} = (\mathbb{E}\|g(\boldsymbol{X})\|_2^2)^{1/2}$, and let $\mathbb{G} = \mathbb{O}(p) \ltimes \mathbb{R}^p$ be the rigid-motion group, acting by $((O,c)\cdot g)(x) = Og(x) + c$.

Definition E.1 (Representation equivalence and distance)**.** Write $g_1 \sim g_2$ when $g_2 = Og_1 + c$ almost surely for some $(O,c) \in \mathbb{G}$, and denote the equivalence class by $[g]$. Its quotient distance is

$$d_\sim([g_1],[g_2]) = \inf_{(O,c)\in\mathbb{G}} \|g_1 - (O,c)\cdot g_2\|_{\mathbb{H}}. \tag{E.4}$$

Profiling out the head gives

$$\mathcal{L}^\star(g) := \inf_{\|f\|_\infty \le B_f} \mathcal{L}(g,f), \qquad \mathcal{L}^\star_{\min} := \inf_{\|g\|_\infty \le B_g} \mathcal{L}^\star(g). \tag{E.5}$$

The profiled objective is constant on rigid-motion orbits. We impose identifiability of its minimizing orbit and a local growth condition to convert excess risk into representation error.

Assumption E.3 (Tradeoff-oracle uniqueness and orbit identifiability)**.** Given (λ,μ), there exists a joint minimizer (g_0, f_0) of $\mathcal{L}$ over bounded measurable functions $\|g\|_\infty \le B_g$ and $\|f\|_\infty \le B_f$, such that for any other joint minimizer (g,f) under the same bounds, we have $g \sim g_0$. Moreover, for every $\varepsilon > 0$,

$$\delta(\varepsilon) := \inf\Big\{\mathcal{L}^\star(g) - \mathcal{L}^\star_{\min} :\ \|g\|_\infty \le B_g,\ d_\sim([g],[g_0]) \ge \varepsilon\Big\} \ >\ 0. \tag{E.6}$$

Assumption E.4 (Local power growth on the quotient)**.** There exist constants $c_{\mathrm{LG}} > 0$, $\kappa > 0$, and $r_{\mathrm{LG}} > 0$ such that for all measurable g with $\|g\|_\infty \le B_g$,

$$d_\sim\big([g],[g_0]\big) \le r_{\mathrm{LG}} \quad\Longrightarrow\quad \mathcal{L}^\star(g) - \mathcal{L}^\star_{\min} \ \ge\ c_{\mathrm{LG}}\, d_\sim\big([g],[g_0]\big)^\kappa. \tag{E.7}$$

Theorem E.2 (Quotient-space calibration: consistency and rate conversion)**.** *Under Assumption E.3, let $(\widehat{g}_n, \widehat{f}_n)$ be any estimator with $\|\widehat{f}_n\|_\infty \le B_f$ and $\|\widehat{g}_n\|_\infty \le B_g$, and define*

$$\Delta_n := \mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}(g_0, f_0) = \mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}^\star_{\min}.$$

*(i) (*Consistency*) If $\Delta_n \to 0$ in probability, then*

$$d_\sim\big([\widehat{g}_n],[g_0]\big) \to 0 \qquad \textit{in probability.}$$

In particular, if $\mathbb{E}[\Delta_n] \to 0$, then $d_\sim([\widehat{g}_n],[g_0]) \to 0$ in probability.

*(ii) (*Rates*) If additionally Assumption E.4 holds and $\Delta_n = O_\mathbb{P}(r_n)$ for some $r_n \downarrow 0$, then*

$$d_\sim\big([\widehat{g}_n],[g_0]\big) = O_\mathbb{P}\big(r_n^{1/\kappa}\big).$$

By Markov's inequality, (E.3) and the local growth condition give $d_\sim([\widehat{g}_n],[g_0]) = O_\mathbb{P}((r_n)^{1/\kappa})$ whenever $r_n \to 0$.

E.3 Downstream classification performance

Let $\mathcal{L}_{\mathrm{ce}}(g,f) = \mathbb{E}[\ell(Y, f(g(\boldsymbol{X})))]$. Representation convergence separates predictive utility from the joint objective.

Theorem E.3 (Downstream cross-entropy calibration to the tradeoff oracle)**.** *Work under Assumption E.3 and Assumption E.1. Let Δ_n be as in Theorem E.2, then*

$$\big|\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n,\widehat{f}_n) - \mathcal{L}_{\mathrm{ce}}(g_0,f_0)\big| \;\le\; \Delta_n \;+\; 16\big(\lambda B_Z + \mu B_Y\big)\, d_\sim\big([\widehat{g}_n],[g_0]\big). \tag{E.8}$$

Consequently, if $\Delta_n \to 0$ in probability, then $\big|\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n,\widehat{f}_n) - \mathcal{L}_{\mathrm{ce}}(g_0,f_0)\big| \to 0$ in probability. If moreover Assumption E.4 holds and $\Delta_n = O_\mathbb{P}(r_n)$ for some $r_n \downarrow 0$, then

$$\big|\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n,\widehat{f}_n) - \mathcal{L}_{\mathrm{ce}}(g_0,f_0)\big| = O_\mathbb{P}(r_n) + O_\mathbb{P}(r_n^{1/\kappa}).$$

The comparison is with the optimum at the chosen fairness–utility trade-off. A standard log-loss calibration also gives a Bayes classification bound with separate trade-off bias and estimation terms (Corollary E.27, Appendix E.9).

E.4 Fairness of the learned representation

Theorem E.4 (Fairness of the learned representation)**.** *Under Assumptions E.1 and E.3, for any $(\widehat{g}_n,\widehat{f}_n)$ with $\|\widehat{g}_n\|_\infty \le B_g$, $\|\widehat{f}_n\|_\infty \le B_f$,*

$$\mathrm{DC}\left(\boldsymbol{Z},\widehat{g}_n(\boldsymbol{X})\right) \;\le\; \mathrm{DC}\left(\boldsymbol{Z},g_0(\boldsymbol{X})\right) \;+\; \frac{2\Delta_n}{\lambda} \;+\; \frac{16\big(\lambda B_Z + 2\mu B_Y\big)}{\lambda}\, d_\sim\big([\widehat{g}_n],[g_0]\big). \tag{E.9}$$

Under Assumptions E.3–E.4 and (E.3)*, the last two terms are $O_\mathbb{P}\big((r_n)/\lambda\big) + O_\mathbb{P}\big((r_n)^{1/\kappa}\big)$. The separation criterion obeys the same bound with $\mathrm{DC}(\boldsymbol{Z},\cdot)$ replaced by $\mathrm{DC}(\boldsymbol{Z},\cdot \mid \boldsymbol{Y})$ and a class-weighting constant in place of B_Z; see Proposition E.25 and the remark following it.*

At fixed λ, the bound approaches the oracle's dependence level; increasing sample size controls estimation error without requiring that level to be zero. Section 3.3 translates this representation guarantee into fairness bounds for replacement heads.

E.5 Proof of Theorem 1

Proof of Theorem 1. Throughout this proof we write $\boldsymbol{X}$ for the second argument of the estimator; the statement for a representation $\boldsymbol{R} = g(\boldsymbol{X})$ follows by applying the argument to the transformed variable, which again has finite second moment under the stated conditions.

We first consider the empirical distance covariance defined in (C.3), which is equivalent to Equation (3.2) in Proposition 1 of Székely & Rizzo (2014). It is shown therein that this expression is an unbiased estimator of the population distance covariance $\mathrm{DC}(\boldsymbol{Z},\boldsymbol{X})$.

Moreover, (C.3) defines a U-statistic with a symmetric kernel h of degree four, and the strong law of large numbers for U-statistics (Hoeffding, 1961) gives $\widehat{\mathrm{DC}}_n(\boldsymbol{Z},\boldsymbol{X}) \to \mathbb{E}[h] = \mathrm{DC}(\boldsymbol{Z},\boldsymbol{X})$ almost surely, hence in probability, as soon as $\mathbb{E}|h| < \infty$. The kernel is a finite linear combination of products

$\|Z_i - Z_j\|\,\|X_k - X_l\|$; when the index pairs are disjoint the expectation of such a product factorizes into first moments, and when they share an index it is bounded by $\mathbb{E}[\|Z_i - Z_j\|\,\|X_k - X_l\|] \le \big(\mathbb{E}\|Z_i - Z_j\|^2\,\mathbb{E}\|X_k - X_l\|^2\big)^{1/2} \le 4\big(\mathbb{E}\|\boldsymbol{Z}\|^2\,\mathbb{E}\|\boldsymbol{X}\|^2\big)^{1/2}$ by the Cauchy–Schwarz inequality. Thus $\mathbb{E}|h| < \infty$ under $\mathbb{E}[\|\boldsymbol{X}\|^2 + \|\boldsymbol{Z}\|^2] < \infty$, and

$$\widehat{\mathrm{DC}}_n(\boldsymbol{Z}, \boldsymbol{X}) \xrightarrow{\mathcal{P}} \mathrm{DC}(\boldsymbol{Z}, \boldsymbol{X}).$$

(A Chebyshev argument through Hoeffding's variance decomposition would require the stronger condition $\mathbb{E}[h^2] < \infty$, i.e. a finite fourth-order cross moment, which the stated assumption does not supply; it holds trivially when $\boldsymbol{X}$ is replaced by a bounded representation $g(\boldsymbol{X})$ as in Section 3.)

For the conditional estimator, let $\rho_k = \mathbb{P}(Y = k) > 0$. The strong law gives $n_k/n \to \rho_k$ almost surely, hence $w_{n,k} \to \rho_k^4 / \sum_j \rho_j^4 = \pi_k$. Within each class, the observations form an i.i.d. sample from the conditional law, with $n_k \to \infty$. Finite unconditional second moments imply finite conditional second moments for every positive-probability class. Applying the same U-statistic law within each class and summing the finitely many terms yields $\widehat{\mathrm{DC}}_n(\boldsymbol{Z}, \boldsymbol{R} \mid Y) \to \sum_k \pi_k\,\mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R} \mid Y = k)$ almost surely, hence in probability.

□

E.6 Proof of Theorem E.1

Throughout this appendix we work under Assumptions E.1–E.2, and the multiplicative constant C varies without further remarks. Recall the joint population and empirical objectives:

$$\mathcal{L}(g, f) = \mathbb{E}\big[\ell\big(Y, f(g(\boldsymbol{X}))\big)\big] + \lambda\,\mathrm{DC}\big(\boldsymbol{Z}, g(\boldsymbol{X})\big) - \mu\,\mathrm{DC}\big(\mathbf{e}_Y, g(\boldsymbol{X})\big),$$

and

$$\mathcal{L}_n(g, f) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(Y_i, f(g(\boldsymbol{X}_i))\big) + \lambda\,\widehat{\mathrm{DC}}_n\big(\boldsymbol{Z}, g(\boldsymbol{X})\big) - \mu\,\widehat{\mathrm{DC}}_n\big(\mathbf{e}_Y, g(\boldsymbol{X})\big),$$

where $\widehat{\mathrm{DC}}_n$ is the unbiased U-statistic estimator defined in (C.3). The joint estimator $(\widehat{g}_n, \widehat{f}_n)$ is any measurable minimizer of (E.1).

We prove Theorem E.1 by the usual decomposition into an approximation term and a stochastic term. The proof is organized as follows: (i) a basic inequality for empirical minimizers; (ii) approximation bounds via Hölder approximation by ReLU networks; (iii) stochastic bounds for the cross-entropy term and the $\widehat{\mathrm{DC}}_n$ U-statistics; (iv) covering-number/pseudo-dimension bounds for ReLU network classes; and (v) assembling all pieces.

E.6.1 Step 1: A basic inequality for the ERM

Lemma E.5 (Basic inequality / excess-risk decomposition)**.** *Let $(\widehat{g}_n, \widehat{f}_n)$ be any minimizer of $\mathcal{L}_n$ over $\mathcal{G}_n \times \mathcal{F}_n$. Fix any oracle pair (g_0, f_0) (measurable) and define the oracle value $\mathcal{L}(g_0, f_0)$. Then*

$$\mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}(g_0, f_0) \;\le\; 2 \sup_{(f,g)\in\mathcal{F}_n\times\mathcal{G}_n} \big|\mathcal{L}(g, f) - \mathcal{L}_n(g, f)\big| + \inf_{(f,g)\in\mathcal{F}_n\times\mathcal{G}_n} \Big\{\mathcal{L}(g, f) - \mathcal{L}(g_0, f_0)\Big\}. \tag{E.10}$$

Proof. Fix any $(g, f) \in \mathcal{G}_n \times \mathcal{F}_n$. Add and subtract $\mathcal{L}_n(\widehat{g}_n, \widehat{f}_n)$ and $\mathcal{L}_n(g, f)$:

$$\begin{aligned}\mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}(g_0, f_0) = \big(\mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_n(\widehat{g}_n, \widehat{f}_n)\big) + \big(\mathcal{L}_n(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_n(g, f)\big) \\ + \big(\mathcal{L}_n(g, f) - \mathcal{L}(g, f)\big) + \big(\mathcal{L}(g, f) - \mathcal{L}(g_0, f_0)\big).\end{aligned}$$

By definition of $(\widehat{g}_n, \widehat{f}_n)$ as an empirical minimizer, $\mathcal{L}_n(\widehat{g}_n, \widehat{f}_n) \le \mathcal{L}_n(g, f)$, hence the middle difference is non-positive. Therefore,

$$\mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}(g_0, f_0) \le \big|\mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_n(\widehat{g}_n, \widehat{f}_n)\big| + \big|\mathcal{L}_n(g, f) - \mathcal{L}(g, f)\big| + \big(\mathcal{L}(g, f) - \mathcal{L}(g_0, f_0)\big).$$

Taking the supremum over $(g, f) \in \mathcal{G}_n \times \mathcal{F}_n$ for the first two terms and then the infimum over $(g, f) \in \mathcal{G}_n \times \mathcal{F}_n$ for the last term yields (E.10). □

E.6.2 Step 2: Approximation error

We first recall a deep ReLU approximation result for Hölder functions.

Lemma E.6 (Deep ReLU approximation of Hölder functions, Corollary 3.1 in Jiao et al. (2023)). *Let $r \in H^\beta([0,1]^q, B)$ for some $\beta > 0$ and integer $q \geq 1$. Then for any integers $M, N \geq 1$, there exists a ReLU network ϕ with depth $D \asymp M \log(8M)$ and width $W \asymp N \log(8N)$ such that*

$$\sup_{x \in [0,1]^q} |r(x) - \phi(x)| \;\leq\; C_{\beta,q}\, B\, (MN)^{-2\beta/q}, \tag{E.11}$$

where $C_{\beta,q} > 0$ depends only on (β, q).

Next we will show two facts of Lipschitz or boundedness conditions that will be used repeatedly.

Lemma E.7 (Cross-entropy: Lipschitz and bounded). *Define the (multiclass) cross-entropy loss by*

$$\ell(y, z) := -z_y + \log\Big(\sum_{k=1}^{K} e^{z_k}\Big), \qquad y \in \{1, \ldots, K\},\ z \in \mathbb{R}^K.$$

Then for any y and any $z, z' \in \mathbb{R}^K$,

$$|\ell(y, z) - \ell(y, z')| \;\leq\; 2\|z - z'\|_\infty. \tag{E.12}$$

Moreover, if $\|z\|_\infty \leq B_f$, then

$$0 \leq \ell(y, z) \leq \log K + 2B_f. \tag{E.13}$$

Proof. Fix y and define $L(z) := \ell(y, z)$. Its gradient is $\nabla L(z) = \pi(z) - \mathbf{e}_y$, where $\pi(z)$ is the softmax vector. Since $\|\pi(z)\|_1 = 1$ and $\|\mathbf{e}_y\|_1 = 1$, we have $\|\nabla L(z)\|_1 \leq 2$. By the mean value theorem,

$$|L(z) - L(z')| \leq \sup_{\tilde{z} \in [z, z']} \|\nabla L(\tilde{z})\|_1 \cdot \|z - z'\|_\infty \leq 2\|z - z'\|_\infty,$$

giving (E.12). For (E.13), note that

$$\ell(y, z) = -\log \pi_y(z) \geq 0,$$
$$\ell(y, z) \leq \|z\|_\infty + \log\big(K e^{\|z\|_\infty}\big) = \log K + 2\|z\|_\infty \leq \log K + 2B_f.$$

□

Lemma E.8 (Distance covariance: Lipschitz in the representation argument). *Let (U, U', V) be random vectors defined on the same probability space such that $\mathbb{E}\|U\|_2 + \mathbb{E}\|U'\|_2 < \infty$ and $\|V\|_2 \leq B_V$ almost surely. Then*

$$\big|\mathrm{DC}(U, V) - \mathrm{DC}(U', V)\big| \;\leq\; 16\, B_V\, \mathbb{E}\|U - U'\|_2. \tag{E.14}$$

In particular, if $U = g(\boldsymbol{X})$ and $U' = g_0(\boldsymbol{X})$ for a common $\boldsymbol{X}$ and $\|g - g_0\|_\infty \leq \delta_g$, then

$$\big|\mathrm{DC}(V, g(\boldsymbol{X})) - \mathrm{DC}(V, g_0(\boldsymbol{X}))\big| \;\leq\; 16\, B_V\, \delta_g.$$

Proof. Let (U_1, U_1', V_1) and (U_2, U_2', V_2) be i.i.d. copies of (U, U', V), independent of each other and of (U, U', V). Recall the population identity (cf. (4)):

$$\begin{aligned}
\mathrm{DC}(U, V) &= \mathbb{E}\big[\|U - U_1\|_2\, \|V - V_1\|_2\big] - 2\,\mathbb{E}\big[\|U - U_1\|_2\, \|V - V_2\|_2\big] \\
&\quad + \mathbb{E}\|U - U_1\|_2 \cdot \mathbb{E}\|V - V_1\|_2, \\
\mathrm{DC}(U', V) &= \mathbb{E}\big[\|U' - U_1'\|_2\, \|V - V_1\|_2\big] - 2\,\mathbb{E}\big[\|U' - U_1'\|_2\, \|V - V_2\|_2\big] \\
&\quad + \mathbb{E}\|U' - U_1'\|_2 \cdot \mathbb{E}\|V - V_1\|_2.
\end{aligned}$$

Set

$$A := \|U - U_1\|_2, \qquad A' := \|U' - U_1'\|_2, \qquad B_1 := \|V - V_1\|_2, \qquad B_2 := \|V - V_2\|_2.$$

Then

$$\mathrm{DC}(U, V) - \mathrm{DC}(U', V) = \underbrace{\mathbb{E}\big[(A - A')B_1\big]}_{\text{(I)}} - 2\underbrace{\mathbb{E}\big[(A - A')B_2\big]}_{\text{(II)}} + \underbrace{\big(\mathbb{E}A - \mathbb{E}A'\big)\,\mathbb{E}B_1}_{\text{(III)}}.$$

We bound the three terms separately.

Step A: control $\mathbb{E}|A - A'|$. By the reverse triangle inequality and then the triangle inequality,

$$|A - A'| = \Big| \|U - U_1\|_2 - \|U' - U_1'\|_2 \Big| \le \big\| (U - U_1) - (U' - U_1') \big\|_2 \\ \le \|U - U'\|_2 + \|U_1 - U_1'\|_2.$$

Taking expectations and using that $(U_1, U_1') \overset{d}{=} (U, U')$ gives

$$\mathbb{E}|A - A'| \;\le\; \mathbb{E}\|U - U'\|_2 + \mathbb{E}\|U_1 - U_1'\|_2 \;=\; 2\,\mathbb{E}\|U - U'\|_2. \tag{E.15}$$

Step B: bound (I) and (II). Since $\|V\|_2 \le B_V$ a.s., we have $B_1 \le 2B_V$ and $B_2 \le 2B_V$ almost surely. Hence

$$|(\mathrm{I})| \le \mathbb{E}\big[|A - A'|\, B_1\big] \le 2B_V\, \mathbb{E}|A - A'| \overset{(E.15)}{\le} 4B_V\, \mathbb{E}\|U - U'\|_2,$$

and similarly

$$|(\mathrm{II})| \le \mathbb{E}\big[|A - A'|\, B_2\big] \le 2B_V\, \mathbb{E}|A - A'| \overset{(E.15)}{\le} 4B_V\, \mathbb{E}\|U - U'\|_2.$$

Step C: bound (III). We have $|\mathbb{E}A - \mathbb{E}A'| \le \mathbb{E}|A - A'|$ and $\mathbb{E}B_1 \le 2B_V$, hence

$$|(\mathrm{III})| \le \mathbb{E}|A - A'| \cdot 2B_V \overset{(E.15)}{\le} 4B_V\, \mathbb{E}\|U - U'\|_2.$$

Step D: assemble. Combining the bounds yields

$$\big|\mathrm{DC}(U, V) - \mathrm{DC}(U', V)\big| \le |(\mathrm{I})| + 2|(\mathrm{II})| + |(\mathrm{III})| \\ \le (4 + 2 \cdot 4 + 4) B_V \mathbb{E}\|U - U'\|_2 = 16 B_V \mathbb{E}\|U - U'\|_2.$$

□

We can now translate uniform approximation of (f_0, g_0) into approximation of the objective $\mathcal{L}$.

Lemma E.9 (Approximation error for the joint objective)**.** *Assume Assumptions E.1–E.2. Let (f, g) be any pair of measurable functions such that $\|g - g_0\|_\infty \le \delta_g$ on $[0,1]^p$ and $\|f - f_0\|_\infty \le \delta_f$ on the domain of f_0. Let $s_f := \beta_f \wedge 1 \in (0, 1]$ and let L_{f_0} denote the Hölder constant implied by $f_0 \in H^{\beta_f}(\cdot, B_{f0})$, namely*

$$\|f_0(u) - f_0(u')\|_\infty \le L_{f_0}\, \|u - u'\|_2^{s_f}. \tag{E.16}$$

Then

$$\mathcal{L}(g, f) - \mathcal{L}(g_0, f_0) \;\le\; 2\delta_f \;+\; 2L_{f_0}\,(\sqrt{p}\,\delta_g)^{s_f} \;+\; 16\big(\lambda B_Z + \mu B_Y\big)\,\sqrt{p}\,\delta_g, \tag{E.17}$$

where B_Y is the almost-sure bound on $\|\mathbf{e}_Y\|_2$ (in particular $B_Y = 1$ under one-hot encoding).

Proof. By definition of $\mathcal{L}$,

$$\mathcal{L}(g, f) - \mathcal{L}(g_0, f_0) = \Big(\mathbb{E}\ell\big(Y, f(g(\boldsymbol{X}))\big) - \mathbb{E}\ell\big(Y, f_0(g_0(\boldsymbol{X}))\big)\Big) \\ + \lambda\Big(\mathrm{DC}\big(\boldsymbol{Z}, g(\boldsymbol{X})\big) - \mathrm{DC}\big(\boldsymbol{Z}, g_0(\boldsymbol{X})\big)\Big) - \mu\Big(\mathrm{DC}\big(\mathbf{e}_Y, g(\boldsymbol{X})\big) - \mathrm{DC}\big(\mathbf{e}_Y, g_0(\boldsymbol{X})\big)\Big).$$

Cross-entropy term. By Lemma E.7,

$$\mathbb{E}\ell\big(Y, f(g(\boldsymbol{X}))\big) - \mathbb{E}\ell\big(Y, f_0(g_0(\boldsymbol{X}))\big) \le 2\,\mathbb{E}\big\|f(g(\boldsymbol{X})) - f_0(g_0(\boldsymbol{X}))\big\|_\infty.$$

Insert and subtract $f_0(g(\boldsymbol{X}))$ and apply the triangle inequality:

$$\big\|f(g(\boldsymbol{X})) - f_0(g_0(\boldsymbol{X}))\big\|_\infty \le \big\|f(g(\boldsymbol{X})) - f_0(g(\boldsymbol{X}))\big\|_\infty + \big\|f_0(g(\boldsymbol{X})) - f_0(g_0(\boldsymbol{X}))\big\|_\infty.$$

The first term is bounded by $\|f - f_0\|_\infty \le \delta_f$. For the second term, use (E.16) and $\|g(\boldsymbol{X}) - g_0(\boldsymbol{X})\|_2 \le \sqrt{p}\,\|g - g_0\|_\infty \le \sqrt{p}\,\delta_g$:

$$\big\|f_0(g(\boldsymbol{X})) - f_0(g_0(\boldsymbol{X}))\big\|_\infty \le L_{f_0}\,\|g(\boldsymbol{X}) - g_0(\boldsymbol{X})\|_2^{s_f} \le L_{f_0}\,(\sqrt{p}\,\delta_g)^{s_f}.$$

Combining gives

$$\mathbb{E}\ell\big(Y, f(g(\boldsymbol{X}))\big) - \mathbb{E}\ell\big(Y, f_0(g_0(\boldsymbol{X}))\big) \le 2\delta_f + 2L_{f_0}(\sqrt{p}\,\delta_g)^{s_f}.$$

Distance covariance terms. Apply Lemma E.8 with $V = \boldsymbol{Z}$ and then with $V = \mathbf{e}_Y$:

$$\big|\mathrm{DC}(\boldsymbol{Z}, g(\boldsymbol{X})) - \mathrm{DC}(\boldsymbol{Z}, g_0(\boldsymbol{X}))\big| \le 16B_Z\,\mathbb{E}\|g(\boldsymbol{X}) - g_0(\boldsymbol{X})\|_2 \le 16B_Z\,\sqrt{p}\,\delta_g,$$

and similarly

$$\big|\mathrm{DC}(\mathbf{e}_Y, g(\boldsymbol{X})) - \mathrm{DC}(\mathbf{e}_Y, g_0(\boldsymbol{X}))\big| \le 16B_Y\,\sqrt{p}\,\delta_g.$$

Multiplying by λ and μ and inserting into the decomposition yields (E.17). □

Apply Lemma E.6 coordinate-wise to g_0 and f_0. Under Assumption E.2 applied to g_0, there exists $g \in \mathcal{G}_n$ such that

$$\|g - g_0\|_\infty \;\le\; C_g(M_g N_g)^{-2\beta_g/p}, \tag{E.18}$$

where the depth and width satisfy the scaling in Theorem E.1. Similarly, under Assumption E.2 applied to f_0, there exists $f \in \mathcal{F}_n$ such that

$$\|f - f_0\|_\infty \;\le\; C_f(M_f N_f)^{-2\beta_f/p}, \tag{E.19}$$

where p is the representation dimension and chosen to be the same as the data dimension. Plugging (E.18)–(E.19) into Lemma E.9 yields the approximation terms appearing in Theorem E.1.

E.6.3 Step 3: Stochastic error bounds

We use uniform empirical covers to control both the predictive loss and the dependence penalties. For a class $\mathcal{A}$ of scalar- or vector-valued functions, define

$$d_{\infty,s}(a, b; x_{1:s}) := \max_{1\le i\le s} \|a(x_i) - b(x_i)\|_2, \qquad \mathcal{N}_\infty(\delta, \mathcal{A}, s) := \sup_{x_{1:s}} \mathcal{N}(\delta, \mathcal{A}, d_{\infty,s}(\cdot,\cdot; x_{1:s})).$$

For scalar functions the norm is absolute value. Covering centers may be taken in the class, at the cost of a universal change in radius. These covering numbers are deterministic upper bounds over all evaluation samples. Suprema below are assumed measurable; equivalently, the bounds can be read with outer expectations.

Lemma E.10 (Maximal inequality for finite Rademacher families)**.** *Let $\epsilon_1, \dots, \epsilon_s$ be independent Rademacher signs and let $\xi_j = \sum_{i=1}^s a_{j,i}\epsilon_i$, $j = 1, \dots, M$. Then*

$$\mathbb{E}_\epsilon \max_{j\le M} |\xi_j| \le \sqrt{2\log(2M)} \max_{j\le M} \Big(\sum_{i=1}^s a_{j,i}^2\Big)^{1/2}.$$

Proof. For every j, $\mathbb{E}_\epsilon e^{t\xi_j} \le \exp(t^2 \sum_i a_{j,i}^2/2)$. Sum this bound for both signs of all M variables, apply Jensen's inequality to the logarithm of the sum, and minimize over $t > 0$. □

Lemma E.11 (One-scale bound using uniform empirical covers)**.** *Let $W_1, \dots, W_s$ be i.i.d. with law P, and suppose $|h| \le B$ for all $h \in \mathcal{H}$. For every $\delta > 0$,*

$$\mathbb{E}\sup_{h\in\mathcal{H}} |(P_s - P)h| \le 2\delta + 2B\sqrt{\frac{2\log\big(2\mathcal{N}_\infty(\delta, \mathcal{H}, s)\big)}{s}}. \tag{E.20}$$

Proof. Symmetrization gives

$$\mathbb{E}\sup_{h\in\mathcal{H}} |(P_s - P)h| \le 2\mathbb{E}\mathbb{E}_\epsilon \sup_{h\in\mathcal{H}} \left|\frac{1}{s}\sum_{i=1}^s \epsilon_i h(W_i)\right|.$$

Conditional on $W_{1:s}$, choose a δ-net in the empirical sup metric. Approximating each h by its net element changes the signed empirical average by at most δ. Each net element has coefficient squared sum at most B^2/s. Lemma E.10 therefore bounds the conditional Rademacher supremum by

$$\delta + B\sqrt{\frac{2\log(2M)}{s}}, \qquad M \le \mathcal{N}_\infty(\delta, \mathcal{H}, s).$$

The uniform covering bound is deterministic, so taking expectation over $W_{1:s}$ proves (E.20). A cover for a single population $L_2(P)$ metric cannot replace the empirical cover in this step, because it does not uniformly control the sample approximation error. □

Cross-entropy loss. Let $\mathcal{Q}_n = \{f \circ g : g \in \mathcal{G}_n,\ f \in \mathcal{F}_n\}$ be the vector-valued logit class, and define

$$\mathcal{H}_n = \{(x, y) \mapsto \ell(y, q(x)) : q \in \mathcal{Q}_n\}, \qquad B_\ell = \log K + 2B_f.$$

Lemma E.12 (Uniform deviation bound for the cross-entropy term)**.** *For bounded logits $\|f\|_\infty \le B_f$,*

$$\mathbb{E} \sup_{h \in \mathcal{H}_n} |(P_n - P)h| \le \frac{2B_\ell}{n} + 2B_\ell\sqrt{\frac{2\log\left(2\mathcal{N}_\infty(B_\ell/n, \mathcal{H}_n, n)\right)}{n}}. \tag{E.21}$$

Moreover, $\mathcal{N}_\infty(\delta, \mathcal{H}_n, n) \le \mathcal{N}_\infty(\delta/2, \mathcal{Q}_n, n)$.

Proof. Lemma E.7 gives $0 \le \ell \le B_\ell$ and $|\ell(y, u) - \ell(y, v)| \le 2\|u - v\|_\infty \le 2\|u - v\|_2$. Apply Lemma E.11 with $\delta = B_\ell/n$. The same Lipschitz inequality sends a logit cover to a loss cover, uniformly over the input-label sample. We bound the entropy of $\mathcal{Q}_n$ below; we do not treat the softmax/log-loss composition as a piecewise-linear network. □

Distance-covariance kernels. Write $W = (X, V)$ and $h_g(W_1, \dots, W_4) = h((g(X_1), V_1), \dots, (g(X_4), V_4))$, where h is the symmetric kernel in (C.3). Let $\mathcal{K}_n = \{h_g : g \in \mathcal{G}_n\}$. Then $\mathbb{E}h_g = \mathrm{DC}(V, g(X))$.

Lemma E.13 (Kernel envelope and samplewise Lipschitz bound)**.** *If $\|g\|_\infty \le B_g$ and $\|V\|_2 \le B_V$, then*

$$|h_g| \le 72 B_g B_V. \tag{E.22}$$

For any two admissible encoders,

$$|h_g(W_1, \dots, W_4) - h_{\widetilde{g}}(W_1, \dots, W_4)| \le 72 B_V \sum_{j=1}^{4} \|g(X_j) - \widetilde{g}(X_j)\|_2. \tag{E.23}$$

Proof. Set $a_{ij} = \|g(X_i) - g(X_j)\|_2$ and $b_{ij} = \|V_i - V_j\|_2$. Then $a_{ij} \le 2B_g$ and $b_{ij} \le 2B_V$. The three kernel terms in (C.3) are bounded in absolute value by $12B_gB_V$, $24B_gB_V$, and $36B_gB_V$, respectively, which proves (E.22). For $\Delta_j = \|g(X_j) - \widetilde{g}(X_j)\|_2$, $|a_{ij} - \widetilde{a}_{ij}| \le \Delta_i + \Delta_j$. Applying this inequality to the three terms gives bounds $3B_V \sum_j \Delta_j$, $6B_V \sum_j \Delta_j$, and $9B_V \sum_j \Delta_j$. Their sum is bounded by the stated, convenient constant $72B_V \sum_j \Delta_j$. □

Lemma E.14 (Uniform deviation bound for the distance-covariance U-statistic)**.** *For $n \ge 4$, $\|g\|_\infty \le B_g$, and $\|V\|_2 \le B_V$, there is a universal constant C such that, for every $\delta > 0$,*

$$\mathbb{E} \sup_{g \in \mathcal{G}_n} |\widehat{\mathrm{DC}}_n(V, g(X)) - \mathrm{DC}(V, g(X))| \le C B_V \delta + C B_g B_V \sqrt{\frac{\log\left(2\mathcal{N}_\infty(\delta, \mathcal{G}_n, n)\right)}{n}}. \tag{E.24}$$

Proof. Let $m = \lfloor n/4 \rfloor$ and, for a permutation σ of $\{1, \dots, n\}$, set

$$T_\sigma(g) = \frac{1}{m} \sum_{j=1}^{m} h_g(W_{\sigma(4j-3)}, W_{\sigma(4j-2)}, W_{\sigma(4j-1)}, W_{\sigma(4j)}).$$

The symmetry of h_g gives the exact permutation identity $\widehat{\mathrm{DC}}_n(V, g(X)) = (n!)^{-1} \sum_\sigma T_\sigma(g)$, including when n is not divisible by four. The supremum of absolute deviations is convex, hence

$$\mathbb{E} \sup_g |\widehat{\mathrm{DC}}_n(V, g(X)) - \mathbb{E}h_g| \le \frac{1}{n!} \sum_\sigma \mathbb{E} \sup_g |T_\sigma(g) - \mathbb{E}h_g|$$
$$= \mathbb{E} \sup_{h \in \mathcal{K}_n} \left| \frac{1}{m} \sum_{j=1}^{m} h(B_j) - \mathbb{E}h(B_1) \right|, \tag{E.25}$$

where $B_j = (W_{4j-3}, \dots, W_{4j})$ are independent blocks with law $P_W^{\otimes 4}$.

A δ-cover of $\mathcal{G}_n$ on the $4m$ input points induces, by (E.23), a $288 B_V \delta$-cover of $\mathcal{K}_n$ on the m blocks. Thus,

$$\mathcal{N}_\infty(288 B_V \delta, \mathcal{K}_n, m) \le \mathcal{N}_\infty(\delta, \mathcal{G}_n, 4m) \le \mathcal{N}_\infty(\delta, \mathcal{G}_n, n).$$

Apply Lemma E.11 to the independent blocks, using the envelope $72 B_g B_V$ and radius $288 B_V \delta$. Since $m \ge n/8$ for $n \ge 4$, the resulting inequality is (E.24) after absorbing numerical constants. If $B_V = 0$ or $B_g = 0$, the statistic and its population target are zero. □

Extension to the empirical separation penalty. The data-dependent class weights can be handled by the same block argument. Let $W_i = (X_i, Z_i, Y_i)$, with h_g acting on the first two components, and let $\rho_k = \mathbb{P}(Y = k)$ and $b = \sum_{k=1}^K \rho_k^4 > 0$. For a four-tuple I, write $J_I = \mathbf{1}\{Y_i \text{ is identical for all } i \in I\}$ and define

$$A_n(g) = \binom{n}{4}^{-1} \sum_{|I|=4} J_I h_g(W_I), \qquad B_n = \binom{n}{4}^{-1} \sum_{|I|=4} J_I.$$

For $n > 3K$, at least one class contains four observations, so $B_n > 0$. The conditional estimator in (C.4) is exactly $A_n(g)/B_n$. Its population counterpart is $A(g)/b$, where $A(g) = \sum_k \rho_k^4 \, \mathrm{DC}(Z, g(X) \mid Y = k)$.

Corollary E.15 (Uniform deviation for the class-conditional estimator)**.** *For fixed finite K, bounded $V = Z$, and $n > 3K$, the conditional estimator obeys*

$$\mathbb{E} \sup_{g \in \mathcal{G}_n} |\widehat{\mathrm{DC}}_n(Z, g(X) \mid Y) - \mathrm{DC}(Z, g(X) \mid Y)|$$
$$\le \frac{C}{b} \left\{ B_Z \delta + B_g B_Z \sqrt{\frac{\log(2\mathcal{N}_\infty(\delta, \mathcal{G}_n, n))}{n}} \right\} + C B_g B_Z e^{-cnb^2} \tag{E.26}$$

for universal $c, C > 0$ and every $\delta > 0$.

Proof. The kernel $J_I h_g$ has the same envelope and Lipschitz bound as h_g, so the proof of Lemma E.14 controls $\mathbb{E} \sup_g |A_n(g) - A(g)|$ by its right-hand side. Replacing one observation changes B_n by at most $4/n$. Bounded differences therefore give $\mathbb{P}(B_n < b/2) \le 2\exp(-nb^2/32)$ and $\mathbb{E}|B_n - b| \le C/\sqrt{n}$. Put $H = 72 B_g B_Z$. Since $|A(g)| \le Hb$, on $\{B_n \ge b/2\}$,

$$\sup_g \left| \frac{A_n(g)}{B_n} - \frac{A(g)}{b} \right| \le \frac{2}{b} \sup_g |A_n(g) - A(g)| + \frac{2H}{b} |B_n - b|.$$

On the complementary event the error is at most $2H$, because both ratios are averages of kernels bounded by H. Combine the bounds and absorb $n^{-1/2}$ using $\log(2\mathcal{N}_\infty) \ge \log 2$. □

Consequently, Theorem E.1 has the same powers of n for the separation objective, with constants also depending on b, when the boundedness and smoothness assumptions are imposed on the oracle of that objective. The approximation bound is unchanged: the population conditional penalty is a convex combination of within-class DC functionals. The exponential remainder is absorbed into the stochastic term for sufficiently large n.

E.6.4 Step 4: Covering numbers and pseudo-dimension for ReLU networks

Lemma E.16 (Uniform empirical covers via pseudo-dimension)**.** *Let $\mathcal{A}$ be a scalar function class bounded by $B > 0$, with pseudo-dimension $v < \infty$. For $s \geq 1$ and $0 < \delta \leq B$,*

$$\log\big(2\mathcal{N}_\infty(\delta, \mathcal{A}, s)\big) \leq C(v+1)\log\left(\frac{CBs}{\delta}\right). \tag{E.27}$$

For scalar-output ReLU networks of depth D and parameter count at most S,

$$\mathrm{Pdim}\big(\mathcal{NN}^{(r\to 1)}_{D,W,S,B}\big) \leq CDS\log(S+1). \tag{E.28}$$

Proof. On any fixed sample of s points, discretize function values in $[-B, B]$ at resolution proportional to δ. The resulting value patterns are determined by subgraph classifications on at most CsB/δ point-threshold pairs. The subgraph class has VC dimension v; the growth-function bound yields (E.27), with $v+1$ covering the constant-class case. This bound is uniform in the sample. Standard pseudo-dimension entropy estimates are also given in Anthony & Bartlett (2009); the ReLU bound follows from Bartlett et al. (2019). □

For a vector-valued class with q coordinates, construct coordinate covers at radius $\delta/(2\sqrt{q})$ and choose an admissible representative from each occupied product cell. This gives a vector cover at radius δ, with log cardinality bounded by the sum of the coordinate log cardinalities. Because p and K are fixed, these factors are absorbed into the constants. Apply this construction to $\mathcal{G}_n$ and the composed logit class $\mathcal{Q}_n$. The loss-class bound follows from the Lipschitz pullback in Lemma E.12.

E.6.5 Step 5: Proof of Theorem E.1

Proof of Theorem E.1. Lemma E.5 gives

$$\begin{aligned}\mathbb{E}[\Delta_n] \leq 2\mathbb{E}\sup_{g,f}|\mathcal{L}(g,f) - \mathcal{L}_n(g,f)| \\ + \inf_{g\in\mathcal{G}_n,\ f\in\mathcal{F}_n}\{\mathcal{L}(g,f) - \mathcal{L}(g_0,f_0)\}.\end{aligned} \tag{E.29}$$

The approximation argument is unchanged: Lemma E.9 and (E.18)–(E.19) bound the second term by

$$C\{(M_fN_f)^{-2\beta_f/p} + (M_gN_g)^{-2s_f\beta_g/p} + (M_gN_g)^{-2\beta_g/p}\}, \qquad s_f = \beta_f \wedge 1.$$

For the stochastic term, apply Lemma E.12 to the predictive loss and Lemma E.14 to $V = Z$ and $V = \mathbf{e}_Y$. With $\delta = 1/n$,

$$\begin{aligned}\mathbb{E}\sup_{g,f}|\mathcal{L}(g,f) - \mathcal{L}_n(g,f)| \leq C\Bigg\{&\frac{B_\ell + \lambda B_Z + \mu B_Y}{n} \\ &+ B_\ell\sqrt{\frac{\log(2\mathcal{N}_\infty(B_\ell/(2n), \mathcal{Q}_n, n))}{n}} \\ &+ B_g(\lambda B_Z + \mu B_Y)\sqrt{\frac{\log(2\mathcal{N}_\infty(1/n, \mathcal{G}_n, n))}{n}}\Bigg\}.\end{aligned}$$

Both entropy terms are uniform empirical entropies; no fixed-population $L_2(P_X)$ cover is used to approximate a realized sample.

Under the stated network shapes, parameter counts satisfy $S_g \lesssim D_gW_g^2$ and $S_f \lesssim D_fW_f^2$. A composition $f \circ g$ is a ReLU network with depth at most $D_g + D_f$ and parameter count at most $S_g + S_f$. The coordinatewise pseudo-dimensions are therefore bounded by

$$CD_gS_g\log(S_g+1), \qquad C(D_g+D_f)(S_g+S_f)\log(S_g+S_f+1),$$

respectively. Substituting these bounds into (E.27), then into the stochastic inequality, gives the two complexity terms in (E.2). The n^{-1} remainder is absorbed into those terms for sufficiently large n and fixed positive class bounds; the displayed $\log^5 n$ factor is a common upper bound on the resulting logarithmic factors.

Finally, choose fixed M_g, M_f and $N_g \asymp n^{p/(4\beta_g+2p)}$, $N_f \asymp n^{p/(4\beta_f+2p)}$. The encoder and head approximation terms are of orders $n^{-\beta_g/(2\beta_g+p)}$ and $n^{-\beta_f/(2\beta_f+p)}$. Their statistical terms have the same respective powers, and the composition approximation contributes $n^{-s_f\beta_g/(2\beta_g+p)}$. This proves (E.3), up to the stated logarithmic factors. □

E.7 Proofs for Section 3.2

Throughout, P_X denotes the marginal law of $\boldsymbol{X}$ and $\mathbb{H} = L^2(P_X; \mathbb{R}^p)$ with $\|g\|_{\mathbb{H}} = (\mathbb{E}\|g(\boldsymbol{X})\|_2^2)^{1/2}$. Let $\mathbb{G} := \mathbb{O}(p) \ltimes \mathbb{R}^p$ and recall the actions $((O,c)\cdot g)(x) = Og(x) + c$ and $((O,c)\cdot f)(u) = f(O^\top(u-c))$.

Translations can be eliminated explicitly by centering. Let $\bar{g} := \mathbb{E}[g(\boldsymbol{X})] \in \mathbb{R}^p$. Then

$$d_\sim\big([g_1],[g_2]\big) = \inf_{O\in\mathbb{O}(p)} \Big\| \big(g_1 - \bar{g}_1\big) - O\big(g_2 - \bar{g}_2\big) \Big\|_{\mathbb{H}}, \tag{E.30}$$

and the infimum is attained (Lemma E.20). In particular, if $\|g_1\|_\infty, \|g_2\|_\infty \le B_g$, then the optimizing translation has norm at most $2B_g$, so incorporating translations into the orbit is compatible with bounded representations.

E.7.1 Rigid-motion invariance and quotient geometry

Lemma E.17 (Rigid motions can be absorbed by the classifier)**.** *For any $(O,c) \in \mathbb{G}$ and any measurable pair (g,f),*

$$\big((O,c)\cdot f\big)\Big(\big((O,c)\cdot g\big)(x)\Big) = f(g(x)), \qquad \forall x \in \mathbb{R}^p.$$

Consequently, $\mathbb{E}[\ell(Y, ((O,c)\cdot f)(((O,c)\cdot g)(\boldsymbol{X})))] = \mathbb{E}[\ell(Y, f(g(\boldsymbol{X})))]$.

Proof. By definition, $((O,c)\cdot g)(x) = Og(x)+c$ and $((O,c)\cdot f)(u) = f(O^\top(u-c))$. Hence

$$((O,c)\cdot f)(((O,c)\cdot g)(x)) = f\big(O^\top((Og(x)+c)-c)\big) = f(O^\top Og(x)) = f(g(x)).$$

Apply $\ell(Y,\cdot)$ and take expectations. □

Lemma E.18 (Distance covariance is invariant under rigid motions)**.** *Assume $\mathbb{E}\|U\|_2^2 < \infty$ and $\mathbb{E}\|V\|_2^2 < \infty$. Then for any $(O,c) \in \mathbb{G}$,*

$$\mathrm{DC}\big((O,c)\cdot U,\, V\big) = \mathrm{DC}(U,V), \qquad \mathrm{DC}\big(V,\, (O,c)\cdot U\big) = \mathrm{DC}(V,U).$$

Proof. Write $(O,c)\cdot U = OU + c$. In the population identity (4), U enters only through Euclidean distances $\|U - U'\|_2$ for i.i.d. copies U'. Since O is orthogonal and c cancels,

$$\|(OU+c) - (OU'+c)\|_2 = \|O(U-U')\|_2 = \|U-U'\|_2.$$

Thus every term in (4) is unchanged. □

Lemma E.19 (Invariance of the joint objective)**.** *For any $(O,c) \in \mathbb{G}$ and any measurable pair (g,f),*

$$\mathcal{L}\big((O,c)\cdot g,\, (O,c)\cdot f\big) = \mathcal{L}(g,f).$$

Proof. The cross-entropy term is invariant by Lemma E.17. Both distance-covariance terms are invariant by Lemma E.18 with $U = g(\boldsymbol{X})$ and $V \in \{\boldsymbol{Z}, \mathbf{e}_Y\}$. Combine these facts. □

Lemma E.20 (Explicit elimination of translations)**.** *Let $g_1, g_2 \in \mathbb{H}$ and define $\bar{g}_j := \mathbb{E}[g_j(\boldsymbol{X})] \in \mathbb{R}^p$. Then*

$$d_\sim\big([g_1],[g_2]\big) = \inf_{O\in\mathbb{O}(p)} \Big\| \big(g_1 - \bar{g}_1\big) - O\big(g_2 - \bar{g}_2\big) \Big\|_{\mathbb{H}},$$

and the infimum is attained.

Proof. Fix $O \in \mathbb{O}(p)$. For any $c \in \mathbb{R}^p$,

$$\|g_1 - (O,c)\cdot g_2\|_{\mathbb{H}}^2 = \mathbb{E}\|g_1(\boldsymbol{X}) - Og_2(\boldsymbol{X}) - c\|_2^2.$$

As a function of c, the right-hand side is a strictly convex quadratic uniquely minimized at $c^\star(O) = \mathbb{E}[g_1(\boldsymbol{X}) - Og_2(\boldsymbol{X})] = \bar{g}_1 - O\bar{g}_2$. Substituting $c = c^\star(O)$ yields

$$\inf_{c\in\mathbb{R}^p} \|g_1 - (O,c)\cdot g_2\|_{\mathbb{H}}^2 = \Big\|\big(g_1 - \bar{g}_1\big) - O\big(g_2 - \bar{g}_2\big)\Big\|_{\mathbb{H}}^2.$$

Taking the infimum over O gives the claim. Since $\mathbb{O}(p)$ is compact and the objective is continuous in O, the infimum is attained. □

E.7.2 Profiled objective and separation on compact quotient sets

Lemma E.21 (Attainment of the inner minimization defining $\mathcal{L}^\star$)**.** *Fix a measurable g with $\|g\|_\infty \le B_g$. Then the infimum $\mathcal{L}^\star(g) = \inf_{f:\|f\|_\infty\le B_f} \mathcal{L}(g,f)$ is attained by some measurable $f_g^\star$ with $\|f_g^\star\|_\infty \le B_f$, i.e. $\mathcal{L}^\star(g) = \mathcal{L}(g, f_g^\star)$.*

Proof. Write $U := g(\boldsymbol{X})$. Since the distance-covariance terms do not depend on f, it suffices to prove attainment for the cross-entropy component $\mathcal{L}_{\mathrm{ce}}(g,f) = \mathbb{E}[\ell(Y, f(U))]$ under $\|f\|_\infty \le B_f$. Let $\eta^g(u) := (\mathbb{P}(Y=1 \mid U=u), \dots, \mathbb{P}(Y=K \mid U=u))$. For each u, define the conditional risk

$$\varphi(u,z) := \sum_{k=1}^K \eta_k^g(u)\,\ell(k,z), \qquad z \in [-B_f, B_f]^K.$$

For fixed u, $\varphi(u,\cdot)$ is continuous on the compact cube $[-B_f,B_f]^K$, hence it attains the minimum. Moreover, $\varphi(\cdot,z)$ is measurable for each fixed z because $\eta^g(\cdot)$ is measurable. Therefore the argmin correspondence

$$\Gamma(u) := \arg\min_{z\in[-B_f,B_f]^K} \varphi(u,z)$$

is nonempty and compact-valued and has a measurable graph. By a standard measurable selection result (e.g. Berge's maximum theorem), there exists a measurable selection $z^\star(u) \in \Gamma(u)$. Define $f_g^\star(u) := z^\star(u)$. Then $\|f_g^\star\|_\infty \le B_f$ by construction and

$$\mathcal{L}_{\mathrm{ce}}(g, f_g^\star) = \mathbb{E}\big[\min_{z\in[-B_f,B_f]^K} \varphi(U,z)\big] = \inf_{f:\|f\|_\infty\le B_f} \mathcal{L}_{\mathrm{ce}}(g,f).$$

Adding back the f-independent distance-covariance terms yields $\mathcal{L}^\star(g) = \mathcal{L}(g, f_g^\star)$. □

Lemma E.22 (Basic properties of $\mathcal{L}^\star$)**.** *Let $\mathcal{L}^\star$ be defined in* (E.5)*. Then: (i) $\mathcal{L}^\star(g) \le \mathcal{L}(g,f)$ for all admissible (g,f); (ii) $\mathcal{L}^\star((O,c)\cdot g) = \mathcal{L}^\star(g)$ for all $(O,c) \in \mathbb{G}$; (iii) if (g_0,f_0) is any joint minimizer of $\mathcal{L}$ under the bounds, then $\mathcal{L}_{\min}^\star = \mathcal{L}(g_0,f_0)$.*

Proof. (i) is immediate from the definition as an infimum over f. (ii) follows from Lemma E.19 and the bijection $f \leftrightarrow (O,c)^{-1}\cdot f$ on the bounded class. (iii) follows from $\inf_g \inf_f \mathcal{L}(g,f) = \inf_{g,f} \mathcal{L}(g,f)$ and optimality of (g_0,f_0). □

E.7.3 Proof of Theorem E.2

Proof of Theorem E.2. Let $\Delta_n = \mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_{\min}^\star$.

Part (i). Fix $\varepsilon > 0$ and let $\delta(\varepsilon) > 0$ be as in (E.6). On the event $\{d_\sim([\widehat{g}_n],[g_0]) \ge \varepsilon\}$, we have $\|\widehat{g}_n\|_\infty \le B_g$ by assumption, hence

$$\mathcal{L}^\star(\widehat{g}_n) - \mathcal{L}_{\min}^\star \;\ge\; \delta(\varepsilon).$$

Since $\mathcal{L}^\star(\widehat{g}_n) \le \mathcal{L}(\widehat{g}_n, \widehat{f}_n)$, it follows that

$$\delta(\varepsilon) \le \mathcal{L}^\star(\widehat{g}_n) - \mathcal{L}_{\min}^\star \le \mathcal{L}(\widehat{g}_n,\widehat{f}_n) - \mathcal{L}_{\min}^\star = \Delta_n.$$

Therefore,

$$\{d_\sim([\widehat{g}_n],[g_0]) \ge \varepsilon\} \subseteq \{\Delta_n \ge \delta(\varepsilon)\},$$

and hence $\mathbb{P}\big(d_\sim([\widehat{g}_n],[g_0]) \geq \varepsilon\big) \leq \mathbb{P}\big(\Delta_n \geq \delta(\varepsilon)\big)$. If $\Delta_n \to 0$ in probability, the right-hand side converges to 0 for every fixed $\varepsilon > 0$, proving $d_\sim([\widehat{g}_n],[g_0]) \to 0$ in probability. If instead $\mathbb{E}[\Delta_n] \to 0$, Markov's inequality gives $\mathbb{P}(\Delta_n \geq \delta(\varepsilon)) \leq \mathbb{E}[\Delta_n]/\delta(\varepsilon) \to 0$, yielding the same conclusion.

Part (ii). Assume also Assumption E.4 and $\Delta_n = O_\mathbb{P}(r_n)$ with $r_n \downarrow 0$. Define $d_n := d_\sim([\widehat{g}_n],[g_0])$. Fix $M > 0$ and set $t_n := M r_n^{1/\kappa}$. For n large enough, $t_n < r_{\mathrm{LG}}$. Then

$$\mathbb{P}(d_n > t_n) \leq \mathbb{P}(d_n > r_{\mathrm{LG}}) + \mathbb{P}(t_n < d_n \leq r_{\mathrm{LG}}).$$

By Part (i), $\Delta_n = O_\mathbb{P}(r_n)$ implies $\Delta_n \to 0$ in probability, hence $d_n \to 0$ in probability and therefore $\mathbb{P}(d_n > r_{\mathrm{LG}}) \to 0$.

On the event $\{t_n < d_n \leq r_{\mathrm{LG}}\}$, Assumption E.4 yields

$$c_{\mathrm{LG}}\, d_n^\kappa \leq \mathcal{L}^\star(\widehat{g}_n) - \mathcal{L}^\star_{\min} \leq \mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}^\star_{\min} = \Delta_n,$$

so in particular $\Delta_n > c_{\mathrm{LG}} t_n^\kappa = c_{\mathrm{LG}} M^\kappa r_n$. Thus

$$\mathbb{P}(t_n < d_n \leq r_{\mathrm{LG}}) \leq \mathbb{P}\big(\Delta_n > c_{\mathrm{LG}} M^\kappa r_n\big).$$

Since $\Delta_n = O_\mathbb{P}(r_n)$, choosing M large makes the right-hand side arbitrarily small for all large n. Therefore $d_\sim([\widehat{g}_n],[g_0]) = O_\mathbb{P}(r_n^{1/\kappa})$. □

E.7.4 Cross-entropy calibration

Recall that $\mathcal{L}^\star_{\mathrm{ce}}(g) := \inf_{f:\|f\|_\infty \leq B_f} \mathcal{L}_{\mathrm{ce}}(g,f)$. Let g be fixed and write $\mathcal{L}(g,f) = \mathcal{L}_{\mathrm{ce}}(g,f) + \mathcal{L}_{\mathrm{DC}}(g)$, where $\mathcal{L}_{\mathrm{DC}}(g) := \lambda\, \mathrm{DC}(\boldsymbol{Z}, g(\boldsymbol{X})) - \mu\, \mathrm{DC}(\mathbf{e}_Y, g(\boldsymbol{X}))$. Then any minimizer of $f \mapsto \mathcal{L}(g,f)$ over $\|f\|_\infty \leq B_f$ is also a minimizer of $f \mapsto \mathcal{L}_{\mathrm{ce}}(g,f)$ over $\|f\|_\infty \leq B_f$. In particular, if (g_0, f_0) is a joint minimizer, then $\mathcal{L}_{\mathrm{ce}}(g_0,f_0) = \mathcal{L}^\star_{\mathrm{ce}}(g_0)$. For fixed g, $\mathcal{L}_{\mathrm{DC}}(g)$ does not depend on f. Therefore, minimizing $\mathcal{L}(g,f)$ over f is equivalent to minimizing $\mathcal{L}_{\mathrm{ce}}(g,f)$ over f.

Lemma E.23 (Lipschitz control of the DC regularizer in quotient distance)**.** *Assume Assumption E.1. Then for any $g \in \mathbb{H}$,*

$$\big|\mathcal{L}_{\mathrm{DC}}(g) - \mathcal{L}_{\mathrm{DC}}(g_0)\big| \;\leq\; 16\big(\lambda B_Z + \mu B_Y\big)\, d_\sim\big([g],[g_0]\big). \tag{E.31}$$

Proof. We bound the two distance-covariance components separately.

Step 1: the $\mathrm{DC}(\boldsymbol{Z},\cdot)$ term. Using symmetry of distance covariance, $\mathrm{DC}(\boldsymbol{Z}, g(\boldsymbol{X})) = \mathrm{DC}(g(\boldsymbol{X}), \boldsymbol{Z})$ and similarly for g_0. Fix any $(O,c) \in \mathbb{G}$. By rigid-motion invariance (Lemma E.18),

$$\mathrm{DC}(g_0(\boldsymbol{X}), \boldsymbol{Z}) = \mathrm{DC}\left(((O,c)\cdot g_0)(\boldsymbol{X}), \boldsymbol{Z}\right).$$

Therefore,

$$\big|\,\mathrm{DC}(\boldsymbol{Z}, g(\boldsymbol{X})) - \mathrm{DC}(\boldsymbol{Z}, g_0(\boldsymbol{X}))\big| = \big|\,\mathrm{DC}(g(\boldsymbol{X}), \boldsymbol{Z}) - \mathrm{DC}(((O,c)\cdot g_0)(\boldsymbol{X}), \boldsymbol{Z})\big|.$$

Apply the distance-covariance Lipschitz bound (Lemma E.8 with $V = \boldsymbol{Z}$) to obtain

$$\begin{aligned}\big|\,\mathrm{DC}(g(\boldsymbol{X}), \boldsymbol{Z}) - \mathrm{DC}(((O,c)\cdot g_0)(\boldsymbol{X}), \boldsymbol{Z})\big| &\leq 16 B_Z\, \mathbb{E}\big\|g(\boldsymbol{X}) - ((O,c)\cdot g_0)(\boldsymbol{X})\big\|_2 \\ &\leq 16 B_Z\, \|g - (O,c)\cdot g_0\|_{\mathbb{H}}.\end{aligned}$$

Taking the infimum over $(O,c) \in \mathbb{G}$ yields

$$\big|\,\mathrm{DC}(\boldsymbol{Z}, g(\boldsymbol{X})) - \mathrm{DC}(\boldsymbol{Z}, g_0(\boldsymbol{X}))\big| \leq 16 B_Z\, d_\sim\big([g],[g_0]\big).$$

Step 2: the $\mathrm{DC}(\mathbf{e}_Y,\cdot)$ term. The same argument with $V = \mathbf{e}_Y$ (so $\|\mathbf{e}_Y\|_2 \leq B_Y$ a.s.) yields

$$\big|\,\mathrm{DC}(\mathbf{e}_Y, g(\boldsymbol{X})) - \mathrm{DC}(\mathbf{e}_Y, g_0(\boldsymbol{X}))\big| \leq 16 B_Y\, d_\sim\big([g],[g_0]\big).$$

Step 3: combine. By definition of $\mathcal{L}_{\mathrm{DC}}$ and the triangle inequality,

$$\begin{aligned}&\big|\mathcal{L}_{\mathrm{DC}}(g) - \mathcal{L}_{\mathrm{DC}}(g_0)\big| \\ =&\big|\lambda\big(\mathrm{DC}(\boldsymbol{Z}, g(\boldsymbol{X})) - \mathrm{DC}(\boldsymbol{Z}, g_0(\boldsymbol{X}))\big) - \mu\big(\mathrm{DC}(\mathbf{e}_Y, g(\boldsymbol{X})) - \mathrm{DC}(\mathbf{e}_Y, g_0(\boldsymbol{X}))\big)\big| \\ \leq&\lambda\,\big|\,\mathrm{DC}(\boldsymbol{Z}, g(\boldsymbol{X})) - \mathrm{DC}(\boldsymbol{Z}, g_0(\boldsymbol{X}))\big| + \mu\,\big|\,\mathrm{DC}(\mathbf{e}_Y, g(\boldsymbol{X})) - \mathrm{DC}(\mathbf{e}_Y, g_0(\boldsymbol{X}))\big| \\ \leq&16(\lambda B_Z + \mu B_Y)\, d_\sim\big([g],[g_0]\big),\end{aligned}$$

which is (E.31). □

Proof of Theorem E.3. Recall $\Delta_n = \mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}(g_0, f_0)$ and $\mathcal{L}(g, f) = \mathcal{L}_{\mathrm{ce}}(g, f) + \mathcal{L}_{\mathrm{DC}}(g)$. Therefore,

$$\begin{aligned}\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_{\mathrm{ce}}(g_0, f_0) &= \big(\mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_{\mathrm{DC}}(\widehat{g}_n)\big) - \big(\mathcal{L}(g_0, f_0) - \mathcal{L}_{\mathrm{DC}}(g_0)\big) \\ &= \Delta_n - \big(\mathcal{L}_{\mathrm{DC}}(\widehat{g}_n) - \mathcal{L}_{\mathrm{DC}}(g_0)\big).\end{aligned}$$

Taking absolute values and using $\Delta_n \geq 0$ yields

$$\big|\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_{\mathrm{ce}}(g_0, f_0)\big| \leq \Delta_n + \big|\mathcal{L}_{\mathrm{DC}}(\widehat{g}_n) - \mathcal{L}_{\mathrm{DC}}(g_0)\big|.$$

Applying Lemma E.23 with $g = \widehat{g}_n$ gives (E.8).

If $\Delta_n \to 0$ in probability, then Theorem E.2(i) implies $d_\sim([\widehat{g}_n], [g_0]) \to 0$ in probability, hence the right-hand side of (E.8) converges to 0 in probability. If moreover Assumption E.4 holds and $\Delta_n = O_{\mathbb{P}}(r_n)$, then Theorem E.2(ii) yields $d_\sim([\widehat{g}_n], [g_0]) = O_{\mathbb{P}}(r_n^{1/\kappa})$, and the stated rate conclusion follows by substitution into (E.8). □

E.7.5 Bayes 0–1 calibration by log-loss

Lemma E.24 (0–1 calibration by log-loss, with KL identity)**.** *Let $\eta(x) := (\mathbb{P}(Y = 1 \mid \boldsymbol{X} = x), \ldots, \mathbb{P}(Y = K \mid \boldsymbol{X} = x)) \in \Delta^K$ and let $h^\star(x) := \arg\max_k \eta_k(x)$ be a Bayes classifier (measurable tie-breaking if needed). For any measurable predictor $p : \mathcal{X} \to \Delta^K$ and induced classifier $h_p(x) := \arg\max_k p_k(x)$, define*

$$\mathcal{L}_{\log}(p) := \mathbb{E}[-\log p_Y(\boldsymbol{X})], \qquad \mathcal{L}^\star_{\log} := \mathbb{E}[-\log \eta_Y(\boldsymbol{X})].$$

Then

$$\mathbb{P}(h_p(\boldsymbol{X}) \neq Y) - \mathbb{P}(h^\star(\boldsymbol{X}) \neq Y) \leq \sqrt{2\big(\mathcal{L}_{\log}(p) - \mathcal{L}^\star_{\log}\big)}.$$

Proof. Condition on $\boldsymbol{X} = x$ and write $\eta = \eta(x)$ and $p = p(x)$. Let $k^\star \in \arg\max_k \eta_k$ and $\hat{k} \in \arg\max_k p_k$. The conditional 0–1 risk of h_p at x is $1 - \eta_{\hat{k}}$, while that of $h^\star$ is $1 - \eta_{k^\star}$, hence the conditional excess risk equals

$$\big(1 - \eta_{\hat{k}}\big) - \big(1 - \eta_{k^\star}\big) = \eta_{k^\star} - \eta_{\hat{k}}.$$

Since $p_{\hat{k}} \geq p_{k^\star}$,

$$\eta_{k^\star} - \eta_{\hat{k}} \leq |\eta_{k^\star} - p_{k^\star}| + |p_{\hat{k}} - \eta_{\hat{k}}| \leq \sum_{k=1}^K |\eta_k - p_k| = \|\eta - p\|_1.$$

Pinsker's inequality (in the form $\|\eta - p\|_1^2 \leq 2\,\mathrm{KL}(\eta\|p)$) yields

$$\eta_{k^\star} - \eta_{\hat{k}} \leq \sqrt{2\,\mathrm{KL}(\eta\|p)}.$$

We now relate KL to the conditional log-loss gap. By definition,

$$\mathbb{E}[-\log p_Y \mid \boldsymbol{X} = x] = \sum_{k=1}^K \mathbb{P}(Y = k \mid \boldsymbol{X} = x)\,(-\log p_k) = \sum_{k=1}^K \eta_k\,(-\log p_k),$$

and similarly

$$\mathbb{E}[-\log \eta_Y \mid \boldsymbol{X} = x] = \sum_{k=1}^K \eta_k\,(-\log \eta_k).$$

Therefore,

$$\begin{aligned}\mathbb{E}[-\log p_Y \mid \boldsymbol{X} = x] - \mathbb{E}[-\log \eta_Y \mid \boldsymbol{X} = x] &= \sum_{k=1}^K \eta_k\big(-\log p_k + \log \eta_k\big) = \sum_{k=1}^K \eta_k \log\Big(\frac{\eta_k}{p_k}\Big) \\ &= \mathrm{KL}(\eta\|p).\end{aligned}$$

Combining the two displays gives, conditionally on $\boldsymbol{X} = x$,

$$\mathbb{E}\big[\mathbf{1}\{h_p(\boldsymbol{X}) \neq Y\} - \mathbf{1}\{h^\star(\boldsymbol{X}) \neq Y\} \mid \boldsymbol{X} = x\big] \leq \sqrt{2\Big(\mathbb{E}[-\log p_Y \mid \boldsymbol{X} = x] - \mathbb{E}[-\log \eta_Y \mid \boldsymbol{X} = x]\Big)}.$$

Taking expectation over $\boldsymbol{X}$ and applying Jensen's inequality to the concave function $\sqrt{\cdot}$ yields

$$\mathbb{P}(h_p(\boldsymbol{X}) \neq Y) - \mathbb{P}(h^\star(\boldsymbol{X}) \neq Y) \leq \sqrt{2\Big(\mathcal{L}_{\log}(p) - \mathcal{L}^\star_{\log}\Big)}.$$

□

E.8 Proof of Theorem E.4

Proof of Theorem E.4. Write $\Delta_n = \mathcal{L}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}(g_0, f_0)$ and decompose $\mathcal{L} = \mathcal{L}_{\mathrm{ce}} + \mathcal{L}_{\mathrm{DC}}$:

$$\mathcal{L}_{\mathrm{DC}}(\widehat{g}_n) - \mathcal{L}_{\mathrm{DC}}(g_0) = \Delta_n - \big(\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_{\mathrm{ce}}(g_0, f_0)\big). \tag{E.32}$$

By Theorem E.3, $\big|\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_{\mathrm{ce}}(g_0, f_0)\big| \leq \Delta_n + 16(\lambda B_Z + \mu B_Y)\, d_\sim([\widehat{g}_n], [g_0])$, so (E.32) gives

$$\mathcal{L}_{\mathrm{DC}}(\widehat{g}_n) - \mathcal{L}_{\mathrm{DC}}(g_0) \;\leq\; 2\Delta_n + 16(\lambda B_Z + \mu B_Y)\, d_\sim\big([\widehat{g}_n], [g_0]\big). \tag{E.33}$$

Next, by definition of $\mathcal{L}_{\mathrm{DC}}$,

$$\lambda\Big(\mathrm{DC}(\boldsymbol{Z}, \widehat{g}_n(\boldsymbol{X})) - \mathrm{DC}(\boldsymbol{Z}, g_0(\boldsymbol{X}))\Big) = \Big(\mathcal{L}_{\mathrm{DC}}(\widehat{g}_n) - \mathcal{L}_{\mathrm{DC}}(g_0)\Big) + \mu\Big(\mathrm{DC}(\mathbf{e}_Y, \widehat{g}_n(\boldsymbol{X})) - \mathrm{DC}(\mathbf{e}_Y, g_0(\boldsymbol{X}))\Big).$$

Step 2 of the proof of Lemma E.23 bounds the second bracket: combining the rigid-motion invariance of DC (Lemma E.18) with the Lipschitz bound of Lemma E.8 applied with $V = \mathbf{e}_Y$, and optimizing over $(O, c) \in \mathbb{G}$,

$$\big|\,\mathrm{DC}(\mathbf{e}_Y, \widehat{g}_n(\boldsymbol{X})) - \mathrm{DC}(\mathbf{e}_Y, g_0(\boldsymbol{X}))\big| \leq 16 B_Y\, d_\sim\big([\widehat{g}_n], [g_0]\big).$$

Substituting this and (E.33) and dividing by $\lambda > 0$ yields (E.9). The rate statement follows from Theorem E.1 together with Theorem E.2(ii) and Markov's inequality. □

Proposition E.25 (Separation version)**.** *Let Assumptions E.1 and F.1 hold and let $\mathcal{L}_{\mathrm{DC}}$ be formed with the conditional penalty $\mathrm{DC}(\boldsymbol{Z}, g(\boldsymbol{X}) \mid \boldsymbol{Y}) = \sum_k \pi_k\, \mathrm{DC}(\boldsymbol{Z}, g(\boldsymbol{X}) \mid \boldsymbol{Y} = k)$. Then*

$$\mathrm{DC}\left(\boldsymbol{Z}, \widehat{g}_n(\boldsymbol{X}) \mid \boldsymbol{Y}\right) \leq \mathrm{DC}\left(\boldsymbol{Z}, g_0(\boldsymbol{X}) \mid \boldsymbol{Y}\right) + \frac{2\Delta_n}{\lambda} + \frac{16\big(\lambda C_w B_Z + 2\mu B_Y\big)}{\lambda}\, d_\sim\big([\widehat{g}_n], [g_0]\big),$$

where $C_w := \sum_{k=1}^K \pi_k / \mathbb{P}(\boldsymbol{Y} = k)$.

Proof. The argument is that of Theorem E.4; only the Lipschitz step changes. Fix $(O, c) \in \mathbb{G}$ and write $\Delta g := \widehat{g}_n - (O, c) \cdot g_0$. Conditioning on $\boldsymbol{Y} = k$ and applying Lemma E.8 within that class, $\big|\,\mathrm{DC}(\boldsymbol{Z}, \widehat{g}_n \mid \boldsymbol{Y}{=}k) - \mathrm{DC}(\boldsymbol{Z}, g_0 \mid \boldsymbol{Y}{=}k)\big| \leq 16 B_Z\, \mathbb{E}[\|\Delta g(\boldsymbol{X})\| \mid \boldsymbol{Y} = k]$, where we used that a rigid motion of g_0 leaves each within-class distance covariance unchanged (Lemma E.18 applied to the conditional law). Since $\mathbb{E}[\|\Delta g(\boldsymbol{X})\| \mid \boldsymbol{Y} = k] \leq \mathbb{E}\|\Delta g(\boldsymbol{X})\|/\mathbb{P}(\boldsymbol{Y} = k) \leq \|\Delta g\|_{\mathbb{H}}/\mathbb{P}(\boldsymbol{Y} = k)$ by Jensen's inequality, summing against the weights π_k and taking the infimum over (O, c) gives the constant $16 C_w B_Z$. Assumption F.1 guarantees $C_w < \infty$. □

Remark E.26 (Reading the constant C_w). For the class-proportional weights $\pi_k \propto \mathbb{P}(\boldsymbol{Y} = k)^4$ of (C.4), $C_w = \sum_k \mathbb{P}(\boldsymbol{Y} = k)^3 / \sum_j \mathbb{P}(\boldsymbol{Y} = j)^4$; for balanced weights $\pi_k = 1/K$, $C_w = K^{-1} \sum_k \mathbb{P}(\boldsymbol{Y} = k)^{-1}$. Both equal K when the classes are balanced, and $C_w \geq 1$ always. They differ in how they respond to skew: the balanced weighting blows up when any class is rare (the $\mathbb{P}(\boldsymbol{Y} = k)^{-1}$ term), whereas the class-proportional weighting shrinks toward 1 under extreme skew because it puts almost all of the penalty mass on the dominant class — which is precisely why it gives a weak guarantee for rare classes, as discussed after Theorem F.2. Neither weighting is uniformly preferable; the choice should follow which classes the separation guarantee is meant to protect.

E.9 Proof of Corollary E.27

To convert cross-entropy control into a Bayes classification bound, define

$$p_n(x) := \mathrm{softmax}\big(\widehat{f}_n(\widehat{g}_n(x))\big) \in \Delta^K, \qquad \mathcal{L}_{\log}(p) := \mathbb{E}\big[-\log p_Y(\boldsymbol{X})\big],$$

so that $\mathcal{L}_{\log}(p_n) = \mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n)$. Let $\eta(x) := (\mathbb{P}(Y = 1 \mid \boldsymbol{X} = x), \dots, \mathbb{P}(Y = K \mid \boldsymbol{X} = x))$ and $\mathcal{L}^\star_{\log} := \mathbb{E}[-\log \eta_Y(\boldsymbol{X})]$, and let $\mathcal{R}^\star_{0\text{-}1}$ be the Bayes 0–1 risk.

Corollary E.27 (Bayes 0–1 excess risk and irreducible tradeoff bias)**.** *Let* $\widehat{h}_n(x) := \arg\max_k p_{n,k}(x)$ *and* $\mathcal{R}_{0\text{-}1}(\widehat{g}_n, \widehat{f}_n) := \mathbb{P}(\widehat{h}_n(\boldsymbol{X}) \neq Y)$*. Then*

$$\begin{aligned}\mathcal{R}_{0\text{-}1}(\widehat{g}_n, \widehat{f}_n) - \mathcal{R}^\star_{0\text{-}1} \;&\le\; \sqrt{2\Big(\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}^\star_{\log}\Big)} \\ &\le\; \sqrt{2\big(\mathcal{L}_{\mathrm{ce}}(g_0, f_0) - \mathcal{L}^\star_{\log}\big)} + \sqrt{2|\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_{\mathrm{ce}}(g_0, f_0)|}. \qquad \text{(E.34)}\end{aligned}$$

Remark E.28 (Interpretation of the bias term)*.* The first term in (E.34) is an irreducible bias induced by the fixed fairness–utility tradeoff (λ, μ): even the tradeoff oracle (g_0, f_0) need not achieve the Bayes log-loss $\mathcal{L}^\star_{\log}$. The second term is an estimation term, controlled by Theorem E.3, and it vanishes as the joint excess risk vanishes. This decomposition makes explicit how fairness constraints can create unavoidable degradation in predictive accuracy relative to Bayes, while still allowing consistent learning of the best predictor under the imposed tradeoff.

Proof of Corollary E.27. Apply Lemma E.24 with $p = p_n(x) := \mathrm{softmax}(\widehat{f}_n(\widehat{g}_n(x)))$. By definition, $\mathcal{L}_{\log}(p_n) = \mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n)$. This gives the first inequality in (E.34). For the second inequality, note that

$$\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}^\star_{\log} = \big(\mathcal{L}_{\mathrm{ce}}(g_0, f_0) - \mathcal{L}^\star_{\log}\big) + \big(\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_{\mathrm{ce}}(g_0, f_0)\big),$$

and hence

$$|\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}^\star_{\log}| \le |\mathcal{L}_{\mathrm{ce}}(g_0, f_0) - \mathcal{L}^\star_{\log}| + |\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_{\mathrm{ce}}(g_0, f_0)|.$$

Applying $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ with $a = |\mathcal{L}_{\mathrm{ce}}(g_0, f_0) - \mathcal{L}^\star_{\log}|$ and $b = |\mathcal{L}_{\mathrm{ce}}(\widehat{g}_n, \widehat{f}_n) - \mathcal{L}_{\mathrm{ce}}(g_0, f_0)|$ yields the second inequality in (E.34). □

Proof of Corollary F.3. Apply Theorem F.1 with the representation $\boldsymbol{R} = \widehat{g}_n(\boldsymbol{X})$, which satisfies $\|\boldsymbol{R}\| \le B_g$ because $\widehat{g}_n \in \mathcal{G}_n$, and then bound $\mathrm{DC}(\boldsymbol{Z}, \widehat{g}_n(\boldsymbol{X}))$ by the right-hand side of (E.9), using that $t \mapsto \sqrt{t}$ is increasing. The separation statement follows in the same way from Theorem F.2 and Proposition E.25. □

F Proofs for Section 3.3

F.1 Formal inheritance theorems

Let $\|\boldsymbol{R}\| \le B_g$ almost surely and let P_z denote the conditional distribution of $\boldsymbol{R}$ given $\boldsymbol{Z} = z$. For binary $\boldsymbol{Z}$, write $p = \mathbb{P}(\boldsymbol{Z} = 1)$ and $q = 1 - p$; in this section p denotes group probability. The key identity is $\mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R}) = 2p^2q^2\mathcal{E}(P_1, P_0)$, where $\mathcal{E}$ is energy distance (Lemma F.6). For the distance kernel $\kappa(u, v) = \frac{1}{2}(\|u\| + \|v\| - \|u - v\|)$ on $\mathcal{B}_g = \{r : \|r\| \le B_g\}$, energy distance equals twice the squared maximum mean discrepancy. Its reproducing-kernel Hilbert space (RKHS), $\mathbb{H}_\kappa$, therefore provides a predictor class with explicit disparity control. This is the space denoted by $\mathbb{H}_{\mathrm{dist}}$ in Theorem 4.

Theorem F.1 (Fairness inheritance under head replacement: independence)**.** *Let* $\boldsymbol{Z} \in \{0, 1\}$ *with* $p = \mathbb{P}(\boldsymbol{Z} = 1) \in (0, 1)$*,* $q = 1 - p$*. For a head* $h : \mathcal{B}_g \to \mathbb{R}$ *write* $\Gamma(h) = \big|\mathbb{E}[h(\boldsymbol{R}) \mid \boldsymbol{Z} = 1] - \mathbb{E}[h(\boldsymbol{R}) \mid \boldsymbol{Z} = 0]\big|$*. Then for every* $h \in \mathbb{H}_\kappa$ *with* $\|h\|_{\mathbb{H}_\kappa} \le M$*,*

$$\Gamma(h) \;\le\; \frac{M}{2pq}\sqrt{\mathrm{DC}\left(\boldsymbol{Z}, \boldsymbol{R}\right)} \;=\; \frac{M}{2pq}\,\mathrm{dcov}\left(\boldsymbol{Z}, \boldsymbol{R}\right). \qquad \text{(F.1)}$$

This guarantee applies simultaneously to all heads in the RKHS ball, including heads fitted after the representation is frozen. Predictor complexity and group imbalance enter through $M/(2pq)$. Hard decisions require an additional approximation term, stated in Corollary F.7.

F.2 Class-conditional fairness

Assumption F.1 (Nonvanishing cells)**.** $\boldsymbol{Y}$ takes finitely many values with $\rho_k = \mathbb{P}(\boldsymbol{Y} = k) > 0$ for all k, and $p_k := \mathbb{P}(\boldsymbol{Z} = 1 \mid \boldsymbol{Y} = k) \in (0, 1)$ for all k; write $q_k = 1 - p_k$ and $c_\star = \min_k p_k q_k > 0$.

Theorem F.2 (Fairness inheritance: separation)**.** *Under Assumption F.1, let* $\mathrm{DC}_k := \mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R} \mid \boldsymbol{Y} = k)$ *and let* $\mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R} \mid \boldsymbol{Y}) = \sum_k \pi_k \mathrm{DC}_k$ *be the weighted conditional distance covariance of Section 2.2 with weights* $\pi_k > 0$*. Then for every* $h \in \mathbb{H}_\kappa$ *with* $\|h\|_{\mathbb{H}_\kappa} \le M$ *and every class* k*,*

$$\big|\mathbb{E}[h(\boldsymbol{R}) \mid \boldsymbol{Z} = 1, \boldsymbol{Y} = k] - \mathbb{E}[h(\boldsymbol{R}) \mid \boldsymbol{Z} = 0, \boldsymbol{Y} = k]\big| \le \frac{M}{2p_k q_k}\sqrt{\mathrm{DC}_k} \le \frac{M}{2c_\star}\sqrt{\frac{\mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R} \mid \boldsymbol{Y})}{\pi_k}}.$$

For binary $\boldsymbol{Y}$ *and a head* h *with values in* $[0, 1]$*, read as the probability of the positive decision given* $\boldsymbol{R}$*, the left-hand side at* $k = 1$ *is* $\Delta_{\mathrm{TPR}}(h)$ *and at* $k = 0$ *is* $\Delta_{\mathrm{FPR}}(h)$*, and hence* $\Delta_{\mathrm{EO}}(h) \le \frac{M}{2c_\star}\sqrt{\mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R} \mid \boldsymbol{Y})/\min_k \pi_k}$*. For a hard decision rule* D *the argument of Corollary F.7 applied under the conditional law given* $\boldsymbol{Y} = k$ *adds* $2A_k(M, D)$ *to the bound for class* k*, where* $A_k(M, D) := \inf_{\|h\|_{\mathbb{H}_\kappa} \le M} \max_z \mathbb{E}[|D(\boldsymbol{R}) - h(\boldsymbol{R})| \mid \boldsymbol{Z} = z, \boldsymbol{Y} = k]$.

The factor $\pi_k^{-1/2}$ makes the cost of underweighting rare classes explicit. One-hot categorical attributes admit a groupwise extension (Remark F.8); a general vector-valued penalty does not directly yield the same binary-group constants.

F.3 From estimation to downstream fairness

Combining the representation bound from Section 3.2 with marginal inheritance gives a guarantee for the estimated representation and any replacement head in the specified class.

Corollary F.3 (End-to-end guarantee for retrained heads)**.** *Under the assumptions of Theorem E.4, let* $\boldsymbol{Z} \in \{0, 1\}$ *with* $p = \mathbb{P}(\boldsymbol{Z} = 1) \in (0, 1)$*,* $q = 1 - p$*. For every* $h \in \mathbb{H}_\kappa$ *with* $\|h\|_{\mathbb{H}_\kappa} \le M$*, fitted on* $\widehat{g}_n(\boldsymbol{X})$ *by any procedure,*

$$\Gamma(h) \;\le\; \frac{M}{2pq}\left\{\mathrm{DC}\left(\boldsymbol{Z}, g_0(\boldsymbol{X})\right) + \frac{2\Delta_n}{\lambda} + \frac{16(\lambda B_Z + 2\mu B_Y)}{\lambda}\, d_\sim\big([\widehat{g}_n], [g_0]\big)\right\}^{1/2}.$$

For the separation objective, Proposition E.25 and Theorem F.2 give the within-class analogue. Probability-valued heads yield randomized-decision TPR and FPR bounds; deterministic decisions additionally require the approximation terms in Corollary F.7.

F.4 Supporting inheritance results

Proposition F.4 (Exact inheritance)**.** *If* $\boldsymbol{R} \perp\!\!\!\perp \boldsymbol{Z}$ *then* $h(\boldsymbol{R}) \perp\!\!\!\perp \boldsymbol{Z}$ *for every measurable* h*, so every downstream head built on* $\boldsymbol{R}$ *satisfies independence and* $\Delta_{\mathrm{DP}} = 0$*. If* $\boldsymbol{R} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{Y}$ *then* $h(\boldsymbol{R}) \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{Y}$ *for every measurable* h*, so every downstream head satisfies separation and* $\Delta_{\mathrm{TPR}} = \Delta_{\mathrm{FPR}} = \Delta_{\mathrm{EO}} = 0$.

Example F.5 (No uniform bound over measurable heads). Let $\boldsymbol{Z} \sim \mathrm{Bernoulli}(1/2)$ and $\boldsymbol{R} = \varepsilon \boldsymbol{Z} \in \mathbb{R}$ for $\varepsilon > 0$. By Lemma F.6 below, $\mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R}) = 2 \cdot (1/2)^2 (1/2)^2 \cdot 2\varepsilon = \varepsilon/4$, which tends to 0. Yet the head $h(r) = \mathbf{1}\{r > \varepsilon/2\}$ reconstructs $\boldsymbol{Z}$ exactly, so its demographic-parity gap equals 1 for every $\varepsilon > 0$. Hence no bound of the form $\Delta_{\mathrm{DP}}(h) \le \psi(\mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R}))$ with $\psi(0^+) = 0$ can hold uniformly over measurable h, and in particular $\mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R}) \ge \mathrm{DC}(\boldsymbol{Z}, h(\boldsymbol{R}))$ is false in general.

For distributions P, Q on $\mathbb{R}^p$, let $\mathcal{E}(P, Q) = 2\mathbb{E}\|U - V\| - \mathbb{E}\|U - U'\| - \mathbb{E}\|V - V'\|$, where $U, U' \sim P$ and $V, V' \sim Q$ are independent (Székely & Rizzo, 2013). Write $P_z = \mathcal{L}(\boldsymbol{R} \mid \boldsymbol{Z} = z)$.

Lemma F.6 (Distance covariance against a binary attribute is an energy distance)**.** *Let* $\boldsymbol{Z} \in \{0, 1\}$ *with* $p = \mathbb{P}(\boldsymbol{Z} = 1) \in (0, 1)$ *and* $q = 1 - p$*. Then*

$$\mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R}) \;=\; 2\,p^2 q^2\; \mathcal{E}(P_1, P_0). \tag{F.2}$$

Corollary F.7 (Hard decisions)**.** *For any measurable decision rule* $D : \mathbb{R}^p \to \{0, 1\}$*,*

$$\Delta_{\mathrm{DP}}(D) \;\le\; \inf_{M > 0}\Big\{\frac{M}{2pq}\,\mathrm{dcov}(\boldsymbol{Z}, \boldsymbol{R}) \;+\; 2A(M, D)\Big\},$$

where

$$A(M, D) := \inf_{\|h\|_{\mathbb{H}_\kappa} \le M} \max_{z \in \{0,1\}} \mathbb{E}\big[|D(\boldsymbol{R}) - h(\boldsymbol{R})| \,\big|\, \boldsymbol{Z} = z\big].$$

The approximation modulus $A(M, D)$ measures how closely a norm-bounded RKHS head matches the decision rule under both group distributions. It is not estimated in the experiments.

Remark F.8 (Multi-group and multivariate $\boldsymbol{Z}$). Applying Theorem F.1 to the indicator of each group against its complement gives, for a categorical $\boldsymbol{Z}$ with G levels of probability p_g, the one-vs-rest bound $\Gamma_g(h) \le M \operatorname{dcov}(\mathbf{1}\{\boldsymbol{Z} = g\}, \boldsymbol{R})/(2p_g(1-p_g))$. When $\boldsymbol{Z}$ enters the penalty one-hot encoded, the aggregate penalty controls every such term: because DC is linear in the distance kernel of its first argument and $\|\mathbf{e}_z - \mathbf{e}_{z'}\| = \sqrt{2}\,\mathbf{1}\{z \ne z'\} = \frac{1}{\sqrt{2}} \sum_g \big|\mathbf{1}\{z = g\} - \mathbf{1}\{z' = g\}\big|$,

$$\sum_{g=1}^{G} \mathrm{DC}\left(\mathbf{1}\{\boldsymbol{Z} = g\}, \boldsymbol{R}\right) = \sqrt{2}\ \mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R}),$$

$$\Gamma_g(h) \le \frac{2^{1/4} M}{2p_g(1-p_g)}\ \operatorname{dcov}(\boldsymbol{Z}, \boldsymbol{R}) \quad \text{for every } g.$$

(Remark F.9). For continuous or multivariate $\boldsymbol{Z}$ — including a vector of several binary attributes, as in Section 4.4 — exact inheritance (Proposition F.4) still holds, but the energy identity (F.2) is specific to a binary indicator, the Euclidean distance between attribute vectors is not a sum of per-attribute indicators, and the quantitative statement must be made per induced grouping. We regard that case as open.

F.5 Exact inheritance

Proof of Proposition F.4. If $\boldsymbol{R} \perp\!\!\!\perp \boldsymbol{Z}$ then for measurable h the σ-algebra $\sigma(h(\boldsymbol{R})) \subseteq \sigma(\boldsymbol{R})$ is independent of $\sigma(\boldsymbol{Z})$, so $h(\boldsymbol{R}) \perp\!\!\!\perp \boldsymbol{Z}$; taking h to be the composition of a head with the $\arg\max$ map gives $\widehat{\boldsymbol{Y}} \perp\!\!\!\perp \boldsymbol{Z}$ and hence $\Delta_{\mathrm{DP}} = 0$. The conditional statement is identical after replacing every probability by its conditional version given $\boldsymbol{Y} = k$ and noting that conditional independence is likewise preserved by measurable functions of the conditioned variable. □

F.6 Proof of Lemma F.6

Proof of Lemma F.6. Write $p = \mathbb{P}(\boldsymbol{Z} = 1)$, $q = 1 - p$, and let $R_1 \sim P_1$, $R_0 \sim P_0$ be independent draws from the group-conditional laws, with $a_{11} = \mathbb{E}\|R_1 - R_1'\|$, $a_{00} = \mathbb{E}\|R_0 - R_0'\|$ and $a_{10} = \mathbb{E}\|R_1 - R_0\|$. Since $\boldsymbol{Z}$ is binary, $\|\boldsymbol{Z} - \boldsymbol{Z}'\| = \mathbf{1}\{\boldsymbol{Z} \ne \boldsymbol{Z}'\}$ and $\mathbb{E}\|\boldsymbol{Z} - \boldsymbol{Z}'\| = 2pq$. We evaluate the three terms of (4) with $\boldsymbol{U} = \boldsymbol{Z}$, $\boldsymbol{V} = \boldsymbol{R}$.

First, $\mathbb{E}\big[\mathbf{1}\{\boldsymbol{Z} \ne \boldsymbol{Z}'\}\|\boldsymbol{R} - \boldsymbol{R}'\|\big] = 2pq\, a_{10}$, because the event $\{\boldsymbol{Z} \ne \boldsymbol{Z}'\}$ has probability $2pq$ and on it $(\boldsymbol{R}, \boldsymbol{R}')$ is a pair of independent draws from P_1 and P_0.

Second, with $(\boldsymbol{Z}'', \boldsymbol{R}'')$ an independent copy and $\boldsymbol{R}''$ drawn from the marginal $pP_1 + qP_0$,

$$\mathbb{E}\big[\mathbf{1}\{\boldsymbol{Z} \ne \boldsymbol{Z}'\}\|\boldsymbol{R} - \boldsymbol{R}''\|\big] = pq\big(\mathbb{E}\|R_1 - \boldsymbol{R}''\| + \mathbb{E}\|R_0 - \boldsymbol{R}''\|\big) = pq\big(p\,a_{11} + q\,a_{00} + a_{10}\big),$$

using $\mathbb{E}\|R_1 - \boldsymbol{R}''\| = p\,a_{11} + q\,a_{10}$ and $\mathbb{E}\|R_0 - \boldsymbol{R}''\| = p\,a_{10} + q\,a_{00}$.

Third, $\mathbb{E}\|\boldsymbol{R} - \boldsymbol{R}'\| = p^2 a_{11} + q^2 a_{00} + 2pq\,a_{10}$, so the product term equals $2pq\,(p^2 a_{11} + q^2 a_{00} + 2pq\,a_{10})$.

Collecting the three and using $p^2 - p = -pq$ and $q^2 - q = -pq$,

$$\mathrm{DC}(\boldsymbol{Z}, \boldsymbol{R}) = 2pq\Big[a_{10} - p\,a_{11} - q\,a_{00} - a_{10} + p^2 a_{11} + q^2 a_{00} + 2pq\,a_{10}\Big] = 2p^2q^2\big(2a_{10} - a_{11} - a_{00}\big),$$

and $2a_{10} - a_{11} - a_{00} = \mathcal{E}(P_1, P_0)$ by definition. □

Two immediate consequences are worth recording. First, $\mathcal{E} \ge 0$ with equality if and only if $P_1 = P_0$ (Székely & Rizzo, 2013), which re-derives "DC $= 0$ iff independent" in this special case. Second, the factor $2p^2q^2$ makes explicit that for a rare group the same distributional discrepancy produces a much smaller value of the penalty: at $p = 0.01$ the factor is 1.96×10^{-4} against 0.125 at $p = 1/2$,

so a fixed penalty budget tolerates roughly 6.4×10^2 times more group separation for the rare group, and by Theorem F.1 the resulting downstream guarantee is about 25 times weaker. Practitioners who care about small groups should reweight the penalty accordingly.

F.7 Quantitative inheritance

Proof of Theorem F.1. Since $\|\boldsymbol{R}\| \le B_g$, both P_1 and P_0 are supported on the ball $\mathcal{B}_g$, on which κ is continuous and bounded ($0 \le \kappa(u,u) = \|u - r_0\| \le 2B_g$ for $r_0 \in \mathcal{B}_g$), so the kernel mean embeddings $\varphi_\kappa(P) = \int \kappa(u,\cdot)\, dP(u)$ exist in $\mathbb{H}_\kappa$ and every $h \in \mathbb{H}_\kappa$ is bounded on $\mathcal{B}_g$, with $|h(u)| \le \|h\|_{\mathbb{H}_\kappa} \sqrt{\kappa(u,u)}$. The Euclidean metric on $\mathbb{R}^p$ is a semimetric of negative type, and κ is a kernel that generates it in the sense $\kappa(u,u) + \kappa(v,v) - 2\kappa(u,v) = \|u - v\|$; for such pairs Sejdinovic et al. (2013) show that the energy distance is twice the squared MMD,

$$\mathcal{E}(P,Q) = 2\big\|\varphi_\kappa(P) - \varphi_\kappa(Q)\big\|^2_{\mathbb{H}_\kappa} = 2\,\mathrm{MMD}^2_\kappa(P,Q),$$

for all P, Q with finite first moments. The identity is a two-line expansion: writing $a(u) = \|u - r_0\|$, $\mathrm{MMD}^2_\kappa(P,Q) = \mathbb{E}\kappa(U,U') + \mathbb{E}\kappa(V,V') - 2\mathbb{E}\kappa(U,V)$ with $U, U' \sim P$ and $V, V' \sim Q$ independent, and substituting $\kappa = \frac{1}{2}(a(u) + a(v) - \|u - v\|)$ the a-terms cancel, leaving $\mathbb{E}\|U - V\| - \frac{1}{2}\mathbb{E}\|U - U'\| - \frac{1}{2}\mathbb{E}\|V - V'\| = \frac{1}{2}\mathcal{E}(P,Q)$. The left-hand side does not involve r_0, so the value of $\mathrm{MMD}_\kappa(P,Q)$ is the same for every base point; the RKHS $\mathbb{H}_\kappa$ and hence the head class $\{\|h\|_{\mathbb{H}_\kappa} \le M\}$ do depend on r_0, which is why the theorem fixes it.

By the reproducing property, $\mathbb{E}_P h = \langle h, \varphi_\kappa(P)\rangle_{\mathbb{H}_\kappa}$, so the Cauchy–Schwarz inequality gives the variational form $\mathrm{MMD}_\kappa(P_1,P_0) = \sup_{\|h\|_{\mathbb{H}_\kappa} \le 1} \big|\mathbb{E}_{P_1} h - \mathbb{E}_{P_0} h\big|$ (Gretton et al., 2012), and for any h with $\|h\|_{\mathbb{H}_\kappa} \le M$,

$$\Gamma(h) = \big|\mathbb{E}_{P_1} h - \mathbb{E}_{P_0} h\big| \le M\,\mathrm{MMD}_\kappa(P_1,P_0) = M\sqrt{\tfrac{1}{2}\mathcal{E}(P_1,P_0)}.$$

Substituting $\mathcal{E}(P_1,P_0) = \mathrm{DC}(\boldsymbol{Z},\boldsymbol{R})/(2p^2q^2)$ from Lemma F.6 gives $\Gamma(h) \le M\sqrt{\mathrm{DC}(\boldsymbol{Z},\boldsymbol{R})/(4p^2q^2)} = M\,\mathrm{dcov}(\boldsymbol{Z},\boldsymbol{R})/(2pq)$. The bound involves h only through M, so it holds simultaneously for the whole ball. □

Proof of Corollary F.7. Fix $M > 0$ and h with $\|h\|_{\mathbb{H}_\kappa} \le M$. Then

$$\begin{aligned}
\Delta_{\mathrm{DP}}(D) &= \big|\mathbb{E}[D(\boldsymbol{R}) \mid \boldsymbol{Z}{=}1] - \mathbb{E}[D(\boldsymbol{R}) \mid \boldsymbol{Z}{=}0]\big| \\
&\le \big|\mathbb{E}[h(\boldsymbol{R}) \mid \boldsymbol{Z}{=}1] - \mathbb{E}[h(\boldsymbol{R}) \mid \boldsymbol{Z}{=}0]\big| + \sum_{z\in\{0,1\}} \mathbb{E}\big[|D(\boldsymbol{R}) - h(\boldsymbol{R})| \;\big|\; \boldsymbol{Z} = z\big] \\
&\le \Gamma(h) + 2\max_z \mathbb{E}\big[|D(\boldsymbol{R}) - h(\boldsymbol{R})| \;\big|\; \boldsymbol{Z} = z\big].
\end{aligned}$$

Apply Theorem F.1 to the first term, take the infimum over admissible h to obtain $A(M,D)$, and then the infimum over $M > 0$. □

Proof of Theorem F.2. Fix k. Under Assumption F.1, $\mathbb{P}(\boldsymbol{Y} = k) > 0$ and $p_k \in (0,1)$, so the conditional law of $(\boldsymbol{Z},\boldsymbol{R})$ given $\boldsymbol{Y} = k$ is well defined and $\boldsymbol{Z}$ remains a nondegenerate binary variable under it. Applying Lemma F.6 and Theorem F.1 under this conditional law gives

$$\big|\mathbb{E}[h(\boldsymbol{R}) \mid \boldsymbol{Z} = 1, \boldsymbol{Y} = k] - \mathbb{E}[h(\boldsymbol{R}) \mid \boldsymbol{Z} = 0, \boldsymbol{Y} = k]\big| \le \frac{M}{2p_kq_k}\sqrt{\mathrm{DC}_k}.$$

Since $\mathrm{DC}(\boldsymbol{Z},\boldsymbol{R} \mid \boldsymbol{Y}) = \sum_j \pi_j\,\mathrm{DC}_j$ with all $\pi_j > 0$ and all $\mathrm{DC}_j \ge 0$, we have $\pi_k\,\mathrm{DC}_k \le \mathrm{DC}(\boldsymbol{Z},\boldsymbol{R} \mid \boldsymbol{Y})$, i.e. $\mathrm{DC}_k \le \mathrm{DC}(\boldsymbol{Z},\boldsymbol{R} \mid \boldsymbol{Y})/\pi_k$; combined with $p_kq_k \ge c_\star$ this gives the second inequality. For binary $\boldsymbol{Y}$ and h with values in $[0,1]$ read as $\mathbb{P}(\widehat{\boldsymbol{Y}} = 1 \mid \boldsymbol{R})$, one has $\mathbb{P}(\widehat{\boldsymbol{Y}} = 1 \mid \boldsymbol{Y} = k, \boldsymbol{Z} = z) = \mathbb{E}[h(\boldsymbol{R}) \mid \boldsymbol{Y} = k, \boldsymbol{Z} = z]$, so $k = 1$ is the true-positive case and $k = 0$ the false-positive case, and Δ_{EO} is the maximum of the two. For a hard rule D, the decomposition in the proof of Corollary F.7, written under the conditional law given $\boldsymbol{Y} = k$, gives the additional term $2A_k(M,D)$. □

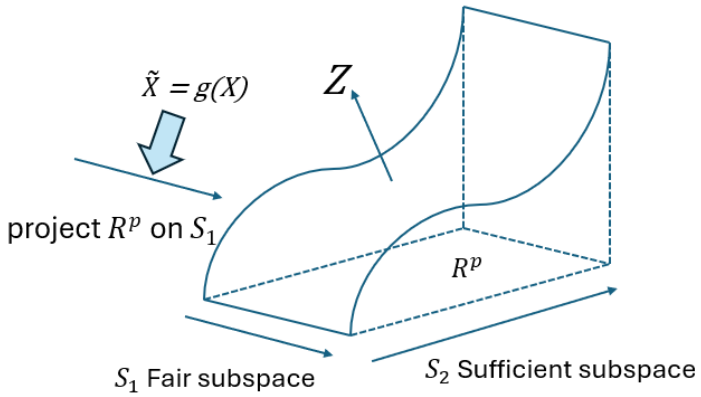


Figure G.1: Fair and sufficient subspaces under the component-independence conditions in the text. The sufficient component carries the dependence on $\boldsymbol{Z}$, while the retained projection is independent of it.

Remark F.9 (Penalizing a categorical attribute group-wise). Theorem F.1 is a statement about a binary indicator. For a categorical $\boldsymbol{Z}$ with levels $1, \dots, G$, one-hot encoded as $\mathbf{e}_{\boldsymbol{Z}}$, the aggregate penalty nevertheless controls each one-vs-rest indicator. Write $a(u, u')$ for the distance kernel of the first argument in (4), so that DC is linear in a for fixed second argument (each of the three terms is linear in a). The one-hot kernel is $a_{\mathrm{oh}}(z, z') = \|\mathbf{e}_z - \mathbf{e}_{z'}\| = \sqrt{2}\,\mathbf{1}\{z \neq z'\}$, and the indicator kernel for level g is $a_g(z, z') = |\mathbf{1}\{z = g\} - \mathbf{1}\{z' = g\}|$, which equals one exactly when one of z, z' is g and the other is not. Since $\sum_g a_g(z, z') = 2 \cdot \mathbf{1}\{z \neq z'\} = \sqrt{2}\, a_{\mathrm{oh}}(z, z')$, linearity gives $\sum_{g=1}^{G} \mathrm{DC}(\mathbf{1}\{\boldsymbol{Z} = g\}, \boldsymbol{R}) = \sqrt{2}\ \mathrm{DC}(\mathbf{e}_{\boldsymbol{Z}}, \boldsymbol{R})$. Every summand is nonnegative, so $\mathrm{DC}(\mathbf{1}\{\boldsymbol{Z} = g\}, \boldsymbol{R}) \leq \sqrt{2}\ \mathrm{DC}(\mathbf{e}_{\boldsymbol{Z}}, \boldsymbol{R})$ for each g, and Theorem F.1 applied to the indicator yields the per-group bound displayed in Remark F.8. The factor $2^{1/4}$ is the price of penalizing the aggregate rather than the G indicators separately; the two penalties differ only by the constant $\sqrt{2}$, so nothing is gained by penalizing the indicators explicitly. Note that this uses the one-hot encoding: for an integer encoding $\boldsymbol{Z} \in \{1, \dots, G\}$ with kernel $|z - z'|$ the identity fails, because $|z - z'|$ is not a fixed multiple of $\mathbf{1}\{z \neq z'\}$.

G GEOMETRIC BACKGROUND

This section gives geometric intuition for representation-level debiasing.

Suppose the representation space decomposes as $\mathbb{R}^p = \mathcal{S}_1 \oplus \mathcal{S}_2$, where $\mathcal{S}_2$ collects sensitive directions and $\mathcal{S}_1$ contains the retained coordinates (Figure G.1). Linear sufficient dimension reduction seeks $B \in \mathbb{R}^{p \times m}$ with $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{X} \mid B^\top \boldsymbol{X}$ (Cook & Li, 2002; Adragni & Cook, 2009). With $B^\top B = I_m$, set $\mathcal{S}_2 = \mathrm{Span}(B)$, $\mathcal{S}_1 = \mathcal{S}_2^{\perp}$, and $P = I - BB^\top$. If the retained and sufficient components also satisfy $P\boldsymbol{X} \perp\!\!\!\perp B^\top \boldsymbol{X}$, then $\boldsymbol{Z} \perp\!\!\!\perp P\boldsymbol{X}$: conditioning on $B^\top \boldsymbol{X}$ and using the two independence relations factorizes their joint law. These conditions give a precise setting in which projection onto $\mathcal{S}_1$ removes sensitive information. They motivate the subspace view of linear erasure methods such as SUP (Shi et al., 2024); the interaction between the fair subspace and the target-relevant subspace is the geometric form of the fairness–utility trade-off (Shi et al., 2025).

Nonlinear SDR replaces $B^\top \boldsymbol{X}$ by a vector-valued map $\boldsymbol{h}$ with $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{h}(\boldsymbol{X})$ (Lee et al., 2013), estimated by kernel (Yeh et al., 2008; Hsing & Ren, 2009) or deep (Chen et al., 2024; Huang et al., 2024) methods. A nonlinear sufficient map does not in general specify an orthogonal complement. DFL instead couples predictive loss to dependence regularization directly in the learned feature space, allowing task-aware transformations beyond linear projection. Its statistical and inheritance analyses are formulated for the learned representation itself.

H EXPERIMENTAL SETTINGS

In this section, we present the detailed experimental setup, including the selection of representation networks $g_\theta(\cdot)$ and downstream classifiers $f_\phi(\cdot)$ for different datasets, as well as the chosen hyperparameters.

H.1 DATASETS AND FROZEN INPUT FEATURES

The five benchmarks cover tabular classification, text classification, and image attribute prediction. The target Y specifies the task, while $\boldsymbol{Z}$ specifies the groups used in the fairness penalty and evaluation. Backbone features remain fixed during DFL training; the additional representation network learns a task-aware transformation of those features. Table H.1 gives the feature dimensions and encoder settings used in the reported experiments.

Adult. Adult is a census-income benchmark (Kohavi, 1996). The binary target indicates whether annual income exceeds \$50,000, and the sensitive attribute is the dataset's binary sex annotation, referred to as gender in the result tables. The tabular feature representation has dimension 101. This task tests whether fairness regularization can retain income-predictive information while reducing disparities between the annotated groups.

Bank. Bank contains records from a Portuguese bank's marketing campaigns (Moro et al., 2014). The task is to predict whether a client subscribes to a term deposit. The sensitive variable is the binary age group $Z = \mathbf{1}\{\text{age} \geq 25\}$, and the input feature dimension is 62. The threshold defines the groups used by both the conditional penalty and the reported fairness metrics.

BIOS. Bias in Bios contains biographies annotated with gender and one of 28 occupations (De-Arteaga et al., 2019). We follow the dataset split of De-Arteaga et al. (2019), predict the occupation, and use gender as the sensitive attribute. Frozen BERT representations have 768 coordinates. The multiclass setting requires dependence control within occupation classes; accordingly, the reported TPR metric aggregates class-specific group gaps using the RMS definition in Appendix B.

MOJI. MOJI is a tweet-sentiment task with happy/sad labels from emoji annotations and African-American English (AAE) versus Standard American English (SAE) group annotations (Blodgett et al., 2016). These dialect labels are a demographic proxy, rather than self-reported race. The training mixture is AAE–happy 40%, SAE–happy 10%, AAE–sad 10%, and SAE–sad 40% (Han et al., 2021), creating an association between sentiment and dialect group. The reported experiments use frozen, $2{,}304$-dimensional DeepMoji representations (Felbo et al., 2017). Thus, BIOS and MOJI use different text encoders.

CelebA. CelebA contains $202{,}599$ celebrity face images with 40 binary attribute annotations (Liu et al., 2015). We consider four separate binary prediction tasks: Young, Attractive, Smiling, and Wavy Hair. Each task is evaluated with Male, Black Hair, or Pale Skin as the sensitive annotation; these names refer to the dataset's attributes. A frozen ResNet-18 supplies 512-dimensional features. In the multiple-attribute setting, the three sensitive annotations form a vector and the penalty acts on their joint conditional dependence with the representation. Reported gaps remain per-attribute summaries, as specified in Appendix B.

H.2 ENCODER OUTPUTS AND OBJECTIVE

Let $\widetilde{g}_\theta(x)$ be the encoder output before clipping. The representation used by both the head and the dependence penalties is

$$g_{\theta,j}(x) = \min\{10, \max\{-10, \widetilde{g}_{\theta,j}(x)\}\}, \qquad j = 1, \ldots, p.$$

Hence $\|g_\theta(x)\|_2 \leq 10\sqrt{p}$. This supplies a Euclidean representation bound $B_g = 10\sqrt{p}$ for the boundedness conditions in the analysis; 10 is the coordinate bound. The theoretical restrictions on the oracle and head class are stated separately in Appendix E.1.

The experimental objective is

$$\begin{aligned}\mathcal{L}_n^{\alpha}(\theta,\phi) = \alpha\Big[\tfrac{1}{n}\textstyle\sum_i \ell\big(Y_i, f_\phi(g_\theta(X_i))\big) - \widehat{\mathrm{DC}}_n\big(\mathbf{e}_{\boldsymbol{Y}}, g_\theta(\boldsymbol{X})\big)\Big] \\ + (1-\alpha)\,\widehat{\mathrm{DC}}_n\big(\boldsymbol{Z}, g_\theta(\boldsymbol{X}) \mid \boldsymbol{Y}\big)\end{aligned} \tag{H.1}$$

for $\alpha \in (0,1)$. Dividing by α yields the separation objective with $\lambda = (1-\alpha)/\alpha$ and $\mu = 1$. All ablated objectives below use the same clipped representation; only the included objective terms change.

H.3 Network architecture

We describe the structure of the DFL framework, which consists of a representation network $g_\theta(\cdot)$ followed by a downstream classifier $f_\phi(\cdot)$. The entire framework is built upon a modified DenseNet architecture. The key input parameters of the DFL framework are:

- **Input Dimension** (p): The dimensionality of the input representation $\boldsymbol{X}$.
- **Output Dimension** (K): The number of classes for the target variable $\boldsymbol{Y}$.
- **Growth Rate**: Controls the number of features added by each layer in the dense blocks.
- **Depth**: Specifies the total number of layers in the network.
- **Reduction Factor**: Determines the compression applied in transition layers.

The network consists of three key components: input processing, dense blocks with transition layers, and output generation. The input is a p-dimensional vector, processed through a fully connected layer that maps it to an initial feature space of size $2 \times$ growthRate.

The network includes three dense blocks, each composed of multiple layers with dense connectivity. Each layer compresses features into an intermediate space of width $4 \times$ growthRate, and its output is concatenated with the inputs of all preceding layers. After each dense block, a transition layer reduces the feature dimensions using batch normalization and a fully connected layer, controlled by the reduction factor.

Following the final dense block, a batch normalization layer and a fully connected layer produce the latent representation of dimensionality p. Each coordinate is then clipped to $[-10, 10]$ as specified above, and the clipped representation is passed through another fully connected layer to generate logits for K-class classification. Softmax gives class probabilities, and log-softmax supplies the log-probabilities used in the loss.

Table H.1: DFL benchmark configurations.

Dataset	p	K	Growth Rate	Depth	Reduction Factor	α
Adult	101	2	20	10	0.2	0.3
Bank	62	2	20	10	0.2	0.1
BIOS	768	28	64	10	0.2	0.7
MOJI	2304	2	64	10	0.2	0.2
CelebA	512	2	64	10	0.2	0.5

H.4 Hyperparameter selection

For the reported DFL runs, training is set to 100 epochs with learning rate 10^{-3}. Figures I.1 and I.2 show the observed trajectories. Batch size is 128 except on BIOS, where it is 256 to include more of its 28 classes in each batch for the conditional DC terms. Other methods use the implementations described below.

The one-parameter objective is (H.1). We select α on validation data from $\{0.1, 0.2, \ldots, 0.9\}$, minimizing the average TPR-RMS and MCDP gaps subject to accuracy at least 95% of the displayed STD validation accuracy. The same rule applies to each fairness-aware method's candidate set; when no candidate meets the floor, the highest-validation-accuracy candidate is selected. The five-point sweep in Table I.7 is a separate sensitivity study, rather than the nine-point selection grid. Network

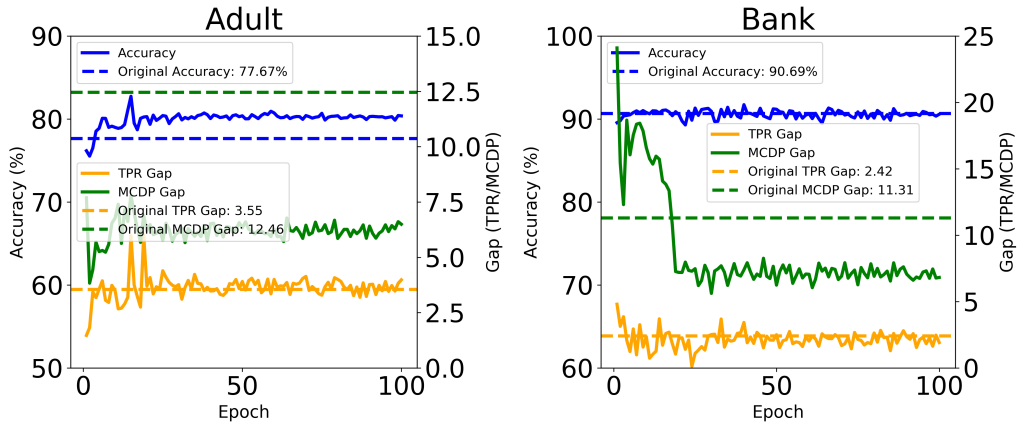


Figure I.1: DFL fine-tuning performance on Adult and Bank

weights use Kaiming initialization and biases are initialized to zero. Table H.1 lists the dataset-specific architecture and selected trade-off weight.

H.5 Baseline implementations and comparison protocol

The Adult and Bank STD rows use the standard multinomial logistic-regression configuration. Their reported MCDP uses 100 histogram bins on $[0, 1]$. The CFair-EO and LAFTR-EO-CE rows train the same vector DenseNet encoder with class-conditional balanced adversaries. They use a training-only three-fold weight search and validation-selected checkpoints for Adult, Bank, BIOS, and MOJI. The CelebA rows use separate $(\boldsymbol{Y}, \boldsymbol{Z})$ pairs, one fixed adversarial weight, and a held-out validation checkpoint.

For baseline methods, we use the official code released by the respective authors. Specifically, for AdvDebias, FairMixup, DRAlign, and DiffMCDP, we use the implementation provided in Jin et al. (2024). For INLP, we set the number of iterations to 100. For RLACE, we follow the training hyperparameters used in the original paper. For SUP, we use $(p + 1)$ slices and select the dimension to remove by 10-fold cross-validation under the common validation rule above.

CFair-EO and LAFTR-EO-CE share a vector DenseNet encoder; the other baselines use their method-specific settings. INLP, RLACE, and SUP are post-hoc transformations of extracted features. Adult and Bank STD use logistic regression.

The head-replacement experiment trains a one-hidden-layer classifier on the frozen DFL representation. Its measured TPR-RMS and MCDP gaps show empirical persistence, while Theorem 4 concerns a distance-kernel RKHS class.

I Additional experimental results

I.1 Trajectories of the performance metrics during fine-tuning

We display the fine-tuning trajectories of the performance metrics (accuracy, TPR gap, and MCDP gap) for the tabular datasets (Adult and Bank) and the contextual datasets (BIOS and MOJI) in Figures I.1 and I.2, respectively.

I.2 Additional results for CelebA

In this section, we present additional numeric results for the CelebA datasets. Tables I.1 and I.2 show the results for $\boldsymbol{Z}$ = Black Hair and $\boldsymbol{Z}$ = Pale Skin across different target attributes. Tables I.3 and I.4 present the results for single and multiple sensitive attributes after discarding the DFL head and training a new unconstrained classifier on the frozen representation.

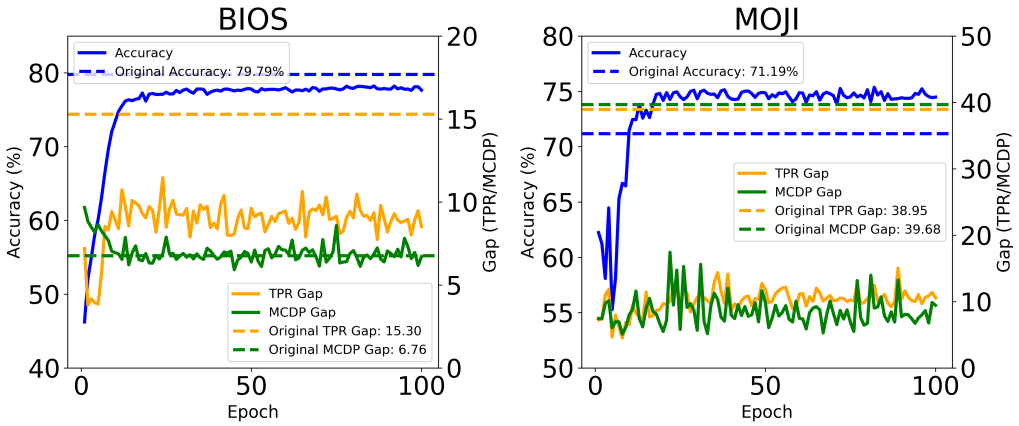


Figure I.2: DFL fine-tuning performance on BIOS and MOJI.

Table I.1: CelebA results with Black Hair as the sensitive attribute. Parentheses give standard deviations. Best and second-best fairness-aware means are bold and underlined.

Target Attribute Y	Metric	Standard	DFL	DiffMCDP	RLACE	CFair-EO	LAFTR-EO-CE
Young	Accuracy ↑	85.01	81.68(1.41)	81.38(0.78)	82.33(1.11)	**82.55(0.23)**	82.38(0.45)
	TPR Gap ↓	19.50	**8.39(2.47)**	8.97(0.95)	12.10(1.67)	8.72(1.37)	8.55(1.12)
	MCDP Gap ↓	11.40	6.18(0.83)	8.72(0.87)	9.07(1.31)	5.46(1.41)	**4.73(0.98)**
Attractive	Accuracy ↑	78.32	77.01(1.09)	76.85(0.88)	76.50(1.25)	76.24(0.71)	**77.15(0.63)**
	TPR Gap ↓	5.83	5.78(0.43)	5.68(0.74)	**4.62(1.44)**	5.36(0.17)	5.58(0.50)
	MCDP Gap ↓	5.63	**3.13(0.66)**	3.35(0.65)	4.99(1.04)	3.34(0.47)	3.19(0.46)
Smiling	Accuracy ↑	80.51	**79.30(2.10)**	76.91(0.74)	78.61(1.32)	78.82(1.13)	78.87(1.10)
	TPR Gap ↓	0.57	1.06(0.71)	**0.56(0.11)**	0.65(0.98)	0.70(0.46)	0.81(0.31)
	MCDP Gap ↓	1.12	1.98(0.49)	**1.47(0.21)**	1.60(0.92)	1.65(0.37)	1.61(0.77)
Wavy Hair	Accuracy ↑	83.52	82.36(1.84)	80.90(0.85)	81.55(1.06)	**82.50(0.45)**	81.90(1.35)
	TPR Gap ↓	9.75	1.78(0.82)	4.49(0.87)	3.05(1.12)	3.31(0.85)	**0.82(0.23)**
	MCDP Gap ↓	15.67	7.31(0.85)	8.55(1.13)	9.33(1.76)	8.53(0.68)	**6.72(0.59)**

Table I.2: CelebA results with Pale Skin as the sensitive attribute. Parentheses give standard deviations. Best and second-best fairness-aware means are bold and underlined.

Target Attribute Y	Metric	Standard	DFL	DiffMCDP	RLACE	CFair-EO	LAFTR-EO-CE
Young	Accuracy ↑	85.01	**84.53(0.44)**	83.20(0.75)	81.88(1.10)	84.23(0.50)	84.15(0.57)
	TPR Gap ↓	8.07	6.62(1.32)	7.72(0.68)	7.01(1.23)	**6.20(0.86)**	6.80(0.69)
	MCDP Gap ↓	18.58	9.94(0.91)	9.77(1.04)	9.89(1.88)	8.74(1.81)	**7.34(0.91)**
Attractive	Accuracy ↑	78.32	77.51(0.53)	**77.96(0.81)**	77.48(1.14)	75.81(0.99)	75.99(0.94)
	TPR Gap ↓	13.90	9.13(1.65)	10.40(0.92)	10.63(1.90)	**8.05(1.40)**	8.49(2.24)
	MCDP Gap ↓	23.10	18.93(0.76)	**16.65(1.39)**	18.29(2.05)	17.18(1.11)	19.08(1.56)
Smiling	Accuracy ↑	80.51	**79.92(1.11)**	77.04(0.83)	79.50(0.91)	78.81(0.69)	78.62(0.78)
	TPR Gap ↓	10.03	4.58(1.38)	4.61(0.91)	6.23(1.34)	**4.37(1.10)**	4.78(1.07)
	MCDP Gap ↓	20.64	15.68(1.09)	15.63(1.29)	16.98(2.13)	15.93(1.06)	**14.89(1.24)**
Wavy Hair	Accuracy ↑	83.52	**83.20(0.50)**	82.89(0.77)	82.49(1.03)	82.30(0.69)	82.84(0.77)
	TPR Gap ↓	1.91	1.58(0.89)	1.89(0.27)	**1.48(0.93)**	1.92(0.80)	2.51(0.95)
	MCDP Gap ↓	8.21	8.08(1.58)	**7.43(0.91)**	8.37(1.43)	8.31(0.70)	8.05(1.13)

Table I.3: Results for the CelebA dataset with a single sensitive attribute, using the frozen DFL encoder $g_{\hat{\theta}}$ with a freshly trained unconstrained head (the head-replacement setting of Section 4.3).

Metric	$\boldsymbol{Z}$ \ $\boldsymbol{Y}$	Young STD	Young DFL	Attractive STD	Attractive DFL	Smiling STD	Smiling DFL	Wavy Hair STD	Wavy Hair DFL
Accuracy ↑	Male	85.01	79.12(1.08)	78.32	74.85(1.65)	80.51	77.06(1.34)	83.52	79.45(0.85)
	Black Hair	85.01	81.43(0.99)	78.32	76.54(1.03)	80.51	78.81(1.85)	83.52	81.69(1.56)
	Pale Skin	85.01	83.98(0.67)	78.32	77.15(0.89)	80.51	79.35(1.42)	83.52	82.91(0.68)
TPR Gap ↓	Male	26.70	4.85(2.45)	40.50	5.91(1.74)	7.65	6.14(2.03)	29.14	5.49(2.78)
	Black Hair	19.50	9.81(2.93)	5.83	7.02(0.87)	0.57	1.06(1.04)	9.75	4.15(1.06)
	Pale Skin	8.07	7.98(1.56)	13.90	10.23(1.98)	10.03	6.73(1.52)	1.91	2.15(1.25)
MCDP Gap ↓	Male	49.74	12.98(2.85)	57.20	23.54(3.12)	16.63	8.94(2.04)	48.13	19.54(3.02)
	Black Hair	11.40	7.15(1.25)	5.63	4.34(0.76)	1.12	2.02(0.87)	15.67	8.03(1.25)
	Pale Skin	18.58	10.24(1.03)	23.10	20.43(1.24)	20.64	16.92(1.94)	8.21	9.24(1.63)

Table I.4: Results for the CelebA dataset with multiple sensitive attributes, using the frozen DFL encoder $g_{\hat{\theta}}$ with a freshly trained unconstrained head.

Metric	$\boldsymbol{Z}$ \ $\boldsymbol{Y}$	Young STD	Young DFL	Attractive STD	Attractive DFL	Smiling STD	Smiling DFL	Wavy Hair STD	Wavy Hair DFL
Accuracy ↑	Male	85.01	81.36(1.12)	78.32	75.14(1.45)	80.51	79.54(1.02)	83.52	82.11(1.06)
	Black Hair	85.01	81.36(1.12)	78.32	75.14(1.45)	80.51	79.54(1.02)	83.52	82.11(1.06)
	Pale Skin	85.01	81.36(1.12)	78.32	75.14(1.45)	80.51	79.54(1.02)	83.52	82.11(1.06)
TPR Gap ↓	Male	26.73	5.94(3.74)	40.50	15.12(3.11)	7.64	7.35(2.72)	29.13	13.24(2.85)
	Black Hair	19.52	15.02(1.87)	5.84	5.94(1.64)	0.57	1.15(1.43)	9.75	5.31(1.84)
	Pale Skin	8.01	3.27(1.82)	13.90	11.35(1.88)	10.02	9.32(0.83)	1.90	2.81(0.85)
MCDP Gap ↓	Male	49.73	24.14(5.21)	57.20	34.62(4.83)	16.64	14.01(5.83)	48.12	30.47(4.95)
	Black Hair	11.40	11.56(2.45)	5.63	6.14(1.88)	1.14	2.51(1.41)	15.60	10.52(1.79)
	Pale Skin	18.52	14.11(1.84)	23.10	20.34(1.95)	20.64	19.81(2.62)	8.21	6.41(1.92)

I.3 Multiple sensitive attributes on CelebA

Table I.5 reports the multi-attribute setting of Section 4.4, in which $\boldsymbol{Z}$ concatenates Male, Black Hair and Pale Skin and the conditional DC penalty is applied to the three-dimensional vector. Gaps are computed per coordinate, with the other coordinates marginalized out. As stated in the paper, these are marginal and not intersectional quantities: they do not measure disparity on the eight subgroups defined by the attribute combinations, and the reductions below should not be read as evidence about those subgroups. On most coordinates the reduction is smaller than in the corresponding single-attribute run, which is expected when one penalty must serve three attributes; two Pale Skin cells go the other way (Young TPR falls $8.01 \to 1.97$ jointly against $8.07 \to 6.62$ singly, and Wavy Hair MCDP falls $8.21 \to 5.41$ against $8.21 \to 8.08$), and on a few already-small cells the gap increases slightly.

I.4 Ablation studies

In this section, we conduct ablation studies on the BIOS dataset to investigate the contribution of different components in the DFL objective. Specifically, we define the following terms:

- $T_1 = \mathbb{E}[\ell(\boldsymbol{Y}, f_\phi(g_\theta(\boldsymbol{X})))]$,
- $T_2 = \mathrm{DC}(\mathbf{e}_{\boldsymbol{Y}}, g_\theta(\boldsymbol{X}))$,
- $T_3 = \mathrm{DC}(\boldsymbol{Z}, g_\theta(\boldsymbol{X}) \mid \boldsymbol{Y})$,
- $T_4 = \mathrm{DC}(\boldsymbol{Z}, f_\phi(g_\theta(\boldsymbol{X})) \mid \boldsymbol{Y})$.

Table I.5: Results for the CelebA dataset with multiple sensitive attributes using DFL. Mean (standard deviation) over 20 replications. Marginal per-coordinate gaps; see Appendix **??** for why these are not intersectional.

	$\boldsymbol{Y}$	Young		Attractive		Smiling		Wavy Hair	
Metric	$\boldsymbol{Z}$	STD	DFL	STD	DFL	STD	DFL	STD	DFL
Accuracy ↑	Male	85.01	81.79(0.83)	78.32	75.87(0.90)	80.51	80.31(0.66)	83.52	82.70(0.46)
	Black Hair	85.01	81.79(0.83)	78.32	75.87(0.90)	80.51	80.31(0.66)	83.52	82.70(0.46)
	Pale Skin	85.01	81.79(0.83)	78.32	75.87(0.90)	80.51	80.31(0.66)	83.52	82.70(0.46)
TPR Gap ↓	Male	26.73	3.94(3.41)	40.50	13.45(4.79)	7.64	6.93(2.12)	29.13	12.08(4.39)
	Black Hair	19.52	12.89(1.67)	5.84	6.03(1.09)	0.57	1.65(1.19)	9.75	4.68(1.63)
	Pale Skin	8.01	1.97(1.54)	13.90	10.11(1.29)	10.02	8.42(2.58)	1.90	1.92(0.68)
MCDP Gap ↓	Male	49.73	23.31(3.69)	57.20	32.58(3.58)	16.64	13.36(2.63)	48.12	29.36(3.29)
	Black Hair	11.40	10.03(1.92)	5.63	5.24(1.29)	1.14	2.65(1.19)	15.60	9.12(1.58)
	Pale Skin	18.52	13.04(1.51)	23.10	19.02(1.47)	20.64	19.38(2.21)	8.21	5.41(1.68)

Table I.6: Effect of the objective components on BIOS. Mean (standard deviation) over 5 replications, all other settings fixed. L_2 versus L_5 isolates representation-level against prediction-level dependence control; L_1 versus L_2 isolates the informativeness term T_2.

Objective	**Composition**	**Accuracy ↑**	**TPR gap ↓**	**MCDP gap ↓**
L_1	$\alpha(T_1 - T_2) + (1-\alpha)T_3$ (full objective)	78.09 (0.42)	9.08 (0.65)	6.21 (0.81)
L_2	$\alpha T_1 + (1-\alpha)T_3$ (no informativeness term)	77.85 (0.38)	8.97 (0.74)	6.52 (0.71)
L_3	$\alpha(T_1 - T_2) + (1-\alpha)(T_3 + T_4)$	61.46 (0.33)	7.35 (0.69)	6.99 (0.66)
L_4	$\alpha T_1 + (1-\alpha)(T_3 + T_4)$	60.52 (0.49)	4.28 (0.75)	8.52 (0.77)
L_5	$\alpha T_1 + (1-\alpha)T_4$ (prediction-level DC only)	64.78 (0.36)	5.12 (0.70)	7.76 (0.62)
L_6	$-\alpha T_2 + (1-\alpha)T_3$ (no task loss)	74.79 (0.51)	11.01 (0.62)	7.83 (0.74)

We evaluate six loss functions:

$$
\begin{aligned}
L_1 &= \alpha(T_1 - T_2) + (1-\alpha)T_3 \quad \text{(Original DFL)} \\
L_2 &= \alpha T_1 + (1-\alpha)T_3 \quad \text{(Without } T_2\text{)} \\
L_3 &= \alpha(T_1 - T_2) + (1-\alpha)(T_3 + T_4) \quad \text{(Adding } T_4\text{)} \\
L_4 &= \alpha T_1 + (1-\alpha)(T_3 + T_4) \quad \text{(Without } T_2\text{, with } T_4\text{)} \\
L_5 &= \alpha T_1 + (1-\alpha)T_4 \quad \text{(Without } T_2, T_3\text{, only } T_4\text{)} \\
L_6 &= -\alpha T_2 + (1-\alpha)T_3 \quad \text{(Without downstream loss)}
\end{aligned}
$$

The results, summarized in Table I.6, report accuracy, TPR gap, and MCDP gap averaged over five replications.

The primary comparisons are penalty location (L_2 versus L_5) and the informativeness term (L_1 versus L_2), discussed in Sections 4.4 and 4.4. These are comparisons at one operating point rather than complete fairness–utility frontiers. The L_6 objective contains no head parameters, and the head-fitting procedure for its reported accuracy is undocumented. That row cannot establish the performance of a controlled two-stage procedure. Removing T_2 changes accuracy from 78.09% to 77.85% at this setting, slightly reduces the TPR-RMS gap, and raises MCDP. This mixed effect does not establish that T_2 is essential at other operating points.

I.5 Hyperparameter sensitivity

Remark I.1 (Cross-table variation in separate sweeps). Two differences affect cross-table comparisons. First, the configuration $\alpha = 0.7$, depth 10, growth rate 64 appears both in Table I.7 ($77.82/9.44/6.42$) and in Table I.8 below ($77.67/9.52/6.04$). These were separate sweeps, and the differences lie within the reported replication standard deviations; the entries are independent summaries rather than repeated measurements from an identical run. Second, the STD baseline for a given $(\boldsymbol{Y}, \boldsymbol{Z})$ cell differs by up to 0.07 between tables (for example, Male/Young TPR is 26.70 in Table 4 and 26.73 in Table I.5). These tables retain their separately evaluated STD references; small cross-table differences should not be interpreted as treatment effects.

Table I.7: Sensitivity to the fairness–utility weight on BIOS. Mean (standard deviation) over 5 replications; depth 10, growth rate 64 throughout.

α	Accuracy ↑	TPR gap ↓	MCDP gap ↓
0.1	48.53 (0.39)	4.77 (0.51)	5.38 (0.63)
0.3	62.33 (0.48)	8.76 (0.66)	6.31 (0.61)
0.5	71.67 (0.37)	8.52 (0.72)	6.04 (0.68)
0.7	77.82 (0.52)	9.44 (0.62)	6.42 (0.69)
0.9	78.76 (0.47)	12.02 (0.73)	6.83 (0.71)

Table I.8: Performance of DFL on BIOS dataset under different DenseNet architecture (growth rates and depths). Values are reported as mean (std) for 5 replications.

Alpha	Depth	Growth Rate	Parameters	Acc ↑	TPR Gap ↓	MCDP Gap ↓
0.7	10	32	132139	77.45 (0.44)	9.93 (0.64)	6.51 (0.76)
0.7	10	64	295030	77.67 (0.37)	9.52 (0.72)	6.04 (0.68)
0.7	10	96	510973	77.82 (0.33)	9.49 (0.71)	5.53 (0.70)
0.7	20	32	206187	78.42 (0.36)	9.41 (0.67)	5.84 (0.63)
0.7	20	64	529775	78.61 (0.41)	10.21 (0.75)	6.11 (0.53)
0.7	20	96	991071	78.86 (0.39)	10.03 (0.69)	5.63 (0.66)

To assess the robustness of DFL, we perform a sensitivity analysis with respect to two key aspects: the loss trade-off parameter α and the architecture of the DenseNet backbone.

Effect of Trade-off Parameter α

The parameter $\alpha \in [0, 1]$ controls the balance between the utility term (cross-entropy loss) and the fairness constraint in the DFL objective. Table I.7 reports the performance of DFL on the BIOS dataset across various values of α, with all other settings held fixed.

As α increases, more weight is placed on the prediction loss. This leads to improved accuracy, from 48.53% at $\alpha = 0.1$ to 78.76% at $\alpha = 0.9$, but comes at the cost of fairness. Specifically, both the TPR gap and the MCDP gap increase overall, though not monotonically — both dip at $\alpha = 0.5$ — reflecting reduced fairness enforcement. This illustrates the expected fairness–utility trade-off and highlights $\alpha = 0.7$ as a balanced choice, offering competitive accuracy (77.82%) while keeping fairness metrics at acceptable levels.

Effect of Network Architecture

We further analyze the impact of the DenseNet encoder's architecture by varying its depth and growth rate, summarized in Table I.8. All experiments are conducted with $\alpha = 0.7$, the reference setting used in this sweep.

Increasing the growth rate (i.e., the width of the network) improves overall performance, with the best fairness observed at a growth rate of 96. For example, increasing the growth rate from 32 to 96 (at depth 10) reduces the MCDP gap from 6.51 to 5.53 and marginally improves accuracy. Increasing depth helps accuracy but not fairness uniformly: at growth rate 64 the TPR gap rises from 9.52 at depth 10 to 10.21 at depth 20. The combination of depth 20 and growth rate 96 attains the highest accuracy (78.86%) with an MCDP gap of 5.63.

These results indicate that DFL's behaviour is not an artifact of one network size: over a $7.5\times$ range of parameter counts accuracy moves by 1.4 points and the MCDP gap by 1.0. Larger encoders help slightly, but not uniformly — at growth rate 64, going from depth 10 to depth 20 raises the TPR gap from 9.52 to 10.21.