

## Supplementary Material for Oblivator

This supplementary material provides additional details to support the main paper. It includes mathematical proofs, implementation specifics, ablation studies, and additional results to further illustrate and validate the effectiveness of **Oblivator**. The contents are organized as follows:

- Mathematical Proofs. (§A)
  - Cross-Covariance Operator (§A.1)
  - Hilbert-Schmidt Operators (§A.2)
  - Empirical Estimation of the Objective (§A.3)
  - Finite-Dimensional Feature Map (§A.4)
  - RKHS Encoders for Feature-Space Alignment (§A.5)
- Implementation Details. (§B)
  - **Oblivator**’s Training Procedure (§B.1)
  - Datasets (§B.2)
  - Experimental Setup (§B.3)
- Probing Networks. (§C)
  - Our Choice of Probing Networks (§C.1)
  - Ablation Studies with Different Probing Networks. (§C.2)
  - Statistical Significance of the Obtained Trade-off (§C.3)
- Additional Results. (§D)
  - Single Step Erasure Vs Multi-Step Erasure (§D.1)
  - Hyperparameter Sensitivity (§D.2)
  - Visualization (§D.3)
- Societal Impact (§E)

## A Mathematical Proofs

### A.1 Cross-Covariance Operator

Let  $\mathcal{F} : \mathbf{x} \mapsto \phi(\mathbf{x})$  and  $\mathcal{G} : \mathbf{y} \mapsto \psi(\mathbf{y})$  denote reproducing kernel Hilbert spaces (RKHSs) with feature maps  $\phi$  and  $\psi$ , respectively. We aim to find functions  $f \in \mathcal{F}$  and  $g \in \mathcal{G}$  that maximize the following objective:

$$\sup_{g \in \mathcal{G}} \sup_{f \in \mathcal{F}} \mathbb{E}[(f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})])(g(\mathbf{y}) - \mathbb{E}[g(\mathbf{y})])] \quad s.t. \quad \|f\|_{\mathcal{F}} = \|g\|_{\mathcal{G}} = 1 \quad (14)$$

Using the reproducing property of RKHSs, the centered evaluation of  $f$  can be expressed as:

$$\begin{aligned} \bar{f}(\mathbf{x}) &= \langle \phi(\mathbf{x}), f \rangle_{\mathcal{F}} - \mathbb{E}[\langle \phi(\mathbf{x}), f \rangle_{\mathcal{F}}] \\ &= \langle \phi(\mathbf{x}), f \rangle_{\mathcal{F}} - \langle \mathbb{E}[\phi(\mathbf{x})], f \rangle_{\mathcal{F}} \\ &= \langle \phi(\mathbf{x}) - \mathbb{E}[\phi(\mathbf{x})], f \rangle_{\mathcal{F}} \\ &= \langle \bar{\phi}(\mathbf{x}), f \rangle_{\mathcal{F}} \end{aligned} \quad (15)$$

where  $\bar{\phi}(\mathbf{x}) := \phi(\mathbf{x}) - \mathbb{E}[\phi(\mathbf{x})]$ . A similar expression holds for  $g(\mathbf{y})$ . Substituting into the objective, we obtain:

$$\sup_{g \in \mathcal{G}} \sup_{f \in \mathcal{F}} \mathbb{E}[\bar{f}(\mathbf{x})\bar{g}(\mathbf{y})] = \mathbb{E}[\langle \bar{\phi}(\mathbf{x}), f \rangle_{\mathcal{F}} \langle \bar{\psi}(\mathbf{y}), g \rangle_{\mathcal{G}}] \quad (16)$$

Recall that the tensor product  $g \otimes f : \mathcal{F} \rightarrow \mathcal{G}$  is a rank-one operator defined by:

$$(g \otimes f)h := g\langle f, h \rangle_{\mathcal{F}}, \quad \text{for all } h \in \mathcal{F}. \quad (17)$$

357 Using this, the objective becomes:

$$\begin{aligned}
\mathbb{E} [\langle \bar{\phi}(\mathbf{x}), f \rangle_{\mathcal{F}} \langle \bar{\psi}(\mathbf{y}), g \rangle_{\mathcal{G}}] &= \mathbb{E} [\langle \bar{\psi}(\mathbf{y}) \langle \bar{\phi}(\mathbf{x}), f \rangle_{\mathcal{F}}, g \rangle_{\mathcal{G}}] \\
&= \mathbb{E} [\langle \bar{\psi}(\mathbf{y}) \otimes \bar{\phi}(\mathbf{x}) f, g \rangle_{\mathcal{G}}] \\
&= \langle \mathbb{E}[\bar{\psi}(\mathbf{y}) \otimes \bar{\phi}(\mathbf{x})] f, g \rangle_{\mathcal{G}} \\
&= \langle \text{Cov}_{yx} f, g \rangle_{\mathcal{G}} = \langle g, \text{Cov}_{yx} f \rangle_{\mathcal{G}}
\end{aligned} \tag{18}$$

358 where  $\text{Cov}_{yx} := \mathbb{E}[\bar{\psi}(\mathbf{y}) \otimes \bar{\phi}(\mathbf{x})]$  denotes the cross-covariance operator from  $\mathcal{F}$  to  $\mathcal{G}$ .

359 **Remark.** Strictly speaking, interchanging the expectation with the inner product requires justifying  
360 that  $\bar{\psi}(\mathbf{y}) \otimes \bar{\phi}(\mathbf{x})$  defines a Hilbert–Schmidt (HS) operator and that its expectation exists in the Hilbert  
361 space of HS operators. We provide a minimal justification in the following section. The derivation  
362 above is retained for its close resemblance to the finite-dimensional cross-covariance formulation.

## 363 A.2 Hilbert–Schmidt Operators

364 Let  $\mathcal{F}$  and  $\mathcal{G}$  be separable Hilbert spaces and  $\{q_i\}_{i=1}^{\infty}$  an orthonormal basis of  $\mathcal{F}$ . For a bounded  
365 operator  $\mathcal{L} : \mathcal{F} \rightarrow \mathcal{G}$  the *Hilbert–Schmidt (HS)* norm is defined as

$$\|\mathcal{L}\|_{\text{HS}}^2 = \sum_{i=1}^{\infty} \|\mathcal{L} q_i\|_{\mathcal{G}}^2 \tag{19}$$

366 If this series converges,  $\mathcal{L}$  is called *Hilbert–Schmidt*. The set of Hilbert–Schmidt operators mapping  
367 from  $\mathcal{F}$  to  $\mathcal{G}$  is a Hilbert space denoted by  $\text{HS}(\mathcal{F}, \mathcal{G})$  with the inner product

$$\langle \mathcal{L}, \mathcal{T} \rangle_{\text{HS}} = \sum_{i=1}^{\infty} \langle \mathcal{L} q_i, \mathcal{T} q_i \rangle_{\mathcal{G}} \tag{20}$$

368 **Rank one tensor product is HS.** For  $f \in \mathcal{F}$  and  $g \in \mathcal{G}$ :

$$\begin{aligned}
\|g \otimes f\|_{\text{HS}}^2 &= \sum_{i=1}^{\infty} \|g \langle f, q_i \rangle_{\mathcal{F}}\|_{\mathcal{G}}^2 \\
&= \|g\|_{\mathcal{G}}^2 \sum_{i=1}^{\infty} |\langle f, q_i \rangle_{\mathcal{F}}|^2 \\
&= \|g\|_{\mathcal{G}}^2 \|f\|_{\mathcal{F}}^2 < \infty
\end{aligned} \tag{21}$$

369 so  $g \otimes f \in \text{HS}(\mathcal{F}, \mathcal{G})$ .

370 **Lemma.** For any  $\mathcal{L} \in \text{HS}(\mathcal{F}, \mathcal{G})$ ,  $f \in \mathcal{F}$ , and  $g \in \mathcal{G}$ ,

$$\langle \mathcal{L}, g \otimes f \rangle_{\text{HS}} = \langle g, \mathcal{L} f \rangle_{\mathcal{G}} \tag{22}$$

371 *Proof.* Choose an orthonormal basis of  $\mathcal{F}$  that begins with the normalised vector  $\tilde{f} := f/\|f\|_{\mathcal{F}}$ , i.e.  
372  $\{\tilde{f}\} \cup \mathcal{B}_{\perp}$  with  $\mathcal{B}_{\perp} := \{\tilde{f}_i^{\perp}\}_{i=1}^{\infty}$ :

$$\begin{aligned}
\langle \mathcal{L}, g \otimes f \rangle_{\text{HS}} &= \langle \mathcal{L} \tilde{f}, g \otimes f \tilde{f} \rangle_{\mathcal{G}} + \overbrace{\sum_{q_i \in \mathcal{B}_{\perp}} \langle \mathcal{L} q_i, g \otimes f q_i \rangle_{\mathcal{G}}}^{=0} \\
&= \frac{1}{\|f\|_{\mathcal{F}}} \langle \mathcal{L} f, g \langle f, \tilde{f} \rangle_{\mathcal{F}} \rangle_{\mathcal{G}} \\
&= \langle g, \mathcal{L} f \rangle_{\mathcal{G}}
\end{aligned} \tag{23}$$

373 The second term vanishes because  $f$  is orthogonal to every  $q_i \in \mathcal{B}_{\perp}$ . □

374 Using this lemma we have :

$$\langle \bar{\psi}(\mathbf{y}) \otimes \bar{\phi}(\mathbf{x}), g \otimes f \rangle_{\text{HS}} = \langle \bar{\phi}(\mathbf{x}), f \rangle_{\mathcal{F}} \langle \bar{\psi}(\mathbf{y}), g \rangle_{\mathcal{G}} \quad (24)$$

375 and

$$\mathbb{E}[\bar{f}(\mathbf{x})\bar{g}(\mathbf{y})] = \mathbb{E}[\langle \bar{\psi}(\mathbf{y}) \otimes \bar{\phi}(\mathbf{x}), g \otimes f \rangle_{\text{HS}}] \quad (25)$$

376 **Boundedness of the expectation functional.** Consider the following linear functional:

$$\mathcal{F} : \text{HS}(\mathcal{F}, \mathcal{G}) \rightarrow \mathbb{R}, \quad \mathcal{F}(\mathcal{L}) = \mathbb{E}[\langle \bar{\psi}(\mathbf{y}) \otimes \bar{\phi}(\mathbf{x}), \mathcal{L} \rangle_{\text{HS}}] \quad (26)$$

377 Applying Jensen and then Cauchy–Schwarz inequalities:

$$\begin{aligned} |\mathcal{F}(\mathcal{L})| &\leq \mathbb{E}[|\langle \bar{\psi}(\mathbf{y}) \otimes \bar{\phi}(\mathbf{x}), \mathcal{L} \rangle_{\text{HS}}|] \\ &\leq \mathbb{E}[\|\bar{\psi}(\mathbf{y}) \otimes \bar{\phi}(\mathbf{x})\|_{\text{HS}} \|\mathcal{L}\|_{\text{HS}}] \\ &\leq \|\mathcal{L}\|_{\text{HS}} \mathbb{E}[\|\bar{\phi}(\mathbf{x})\|_{\mathcal{F}} \|\bar{\psi}(\mathbf{y})\|_{\mathcal{G}}] \\ &\leq \|\mathcal{L}\|_{\text{HS}} \mathbb{E}[\sqrt{k(\mathbf{x}, \mathbf{x}) l(\mathbf{y}, \mathbf{y})}] < \infty \end{aligned} \quad (27)$$

378 so  $\mathcal{F}$  is a bounded linear functional on  $\text{HS}(\mathcal{F}, \mathcal{G})$ .<sup>4</sup>

379 **Riesz representation and the covariance operator.** Recall that on any Hilbert space, the Riesz  
380 representation theorem states that every bounded linear functional can be expressed as an inner  
381 product with a unique element of that space. Applying this to the bounded functional  $\mathcal{F}$  defined  
382 above, we obtain a unique operator  $\mathbb{C}\text{ov}_{yx} \in \text{HS}(\mathcal{F}, \mathcal{G})$  satisfying:

$$\langle \mathbb{C}\text{ov}_{yx}, \mathcal{L} \rangle_{\text{HS}} = \mathcal{F}(\mathcal{L}) \quad \forall \mathcal{L} \in \text{HS}(\mathcal{F}, \mathcal{G}) \quad (28)$$

383 Taking  $\mathcal{L} = g \otimes f$  recovers the cross-covariance identity used in the main text. We now turn to the  
384 empirical estimation of cross-covariance operator.

### 385 A.3 Empirical Estimation of the Objective

386 Given a dataset  $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ , the empirical estimate of the cross-covariance operator is given  
387 by:

$$\widehat{\mathbb{C}\text{ov}}_{yx} = \frac{1}{n} \sum_{i=1}^n \bar{\psi}(\mathbf{y}_i) \otimes \bar{\phi}(\mathbf{x}_i). \quad (29)$$

388 Next, by examining the action of the tensor product operator in Equation (17), together with the  
389 unit-norm constraints on  $f$  and  $g$ , we conclude<sup>5</sup>:

$$f = \sum_{i=1}^n \alpha_i \bar{\phi}(\mathbf{x}_i), \quad \text{and} \quad g = \sum_{i=1}^n \beta_i \bar{\psi}(\mathbf{y}_i) \quad (30)$$

390 Substituting this representation into the objective in Equation (18), we obtain:

$$\mathbb{E}[\bar{f}(\mathbf{x})\bar{g}(\mathbf{x})] \approx \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \alpha_j \beta_k \langle \bar{\psi}(\mathbf{y}_i), \bar{\psi}(\mathbf{y}_k) \rangle_{\mathcal{G}} \cdot \langle \bar{\phi}(\mathbf{x}_i), \bar{\phi}(\mathbf{x}_j) \rangle_{\mathcal{F}} \quad (31)$$

391 Letting  $k$  and  $l$  denote the kernels associated with  $\mathcal{F}$  and  $\mathcal{G}$ , respectively, the expression simplifies to:

$$\mathbb{E}[\bar{f}(\mathbf{x})\bar{g}(\mathbf{x})] \approx \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \beta_k \bar{l}_{ik} \bar{k}_{ij} \alpha_j = \frac{1}{n} \boldsymbol{\beta}^\top \bar{\mathbf{L}} \bar{\mathbf{K}} \boldsymbol{\alpha} \quad (32)$$

392 where  $\bar{\mathbf{K}}$  and  $\bar{\mathbf{L}}$  are the centered kernel matrices corresponding to  $k$  and  $l$ , respectively. Following  
393 the representation of  $f$  and  $g$  from Equation (30), the unit-norm constraint on  $f$  can be expressed as:

$$\begin{aligned} \|f\|_{\mathcal{F}}^2 &= \left\langle \sum_{i=1}^n \alpha_i \bar{\phi}(\mathbf{x}_i), \sum_{j=1}^n \alpha_j \bar{\phi}(\mathbf{x}_j) \right\rangle_{\mathcal{F}} \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \langle \bar{\phi}(\mathbf{x}_i), \bar{\phi}(\mathbf{x}_j) \rangle_{\mathcal{F}} \\ &= \boldsymbol{\alpha}^\top \bar{\mathbf{K}} \boldsymbol{\alpha} \end{aligned} \quad (33)$$

<sup>4</sup>We assume  $k(\mathbf{x}, \mathbf{x})$  and  $l(\mathbf{y}, \mathbf{y})$  have finite first moments.

<sup>5</sup>This follows directly from the Representer Theorem.

and likewise for  $g$ :

$$\|g\|_g^2 = \beta^\top \bar{L} \beta \quad (34)$$

Therefore, the empirical estimation of the objective reduces to the following constrained optimization problem:

$$\sup_{\alpha} \sup_{\beta} \frac{1}{n} \beta^\top \bar{L} \bar{K} \alpha \quad \text{s.t.} \quad \alpha^\top \bar{K} \alpha = \beta^\top \bar{L} \beta = 1 \quad (35)$$

Next, let  $\bar{K} = \mathbf{J} \mathbf{J}^\top$  and define  $\mathbf{u} = \mathbf{J}^\top \alpha$ , and similarly let  $\bar{L} = \mathbf{D} \mathbf{D}^\top$  with  $\mathbf{v} = \mathbf{D}^\top \beta$ . Then, the objective can be rewritten as:

$$\sup_{\mathbf{u}} \sup_{\mathbf{v}} \frac{1}{n} \mathbf{v}^\top \mathbf{D}^\top \mathbf{J} \mathbf{u} \quad \text{s.t.} \quad \mathbf{v}^\top \mathbf{v} = \mathbf{u}^\top \mathbf{u} = 1 \quad (36)$$

This is maximized by the largest singular value of  $\mathbf{D}^\top \mathbf{J}$ , i.e., it is upper bounded by  $\sigma_{\max}(\mathbf{D}^\top \mathbf{J})$ . Next, by considering the sum of squared singular values, we obtain:

$$\sum_{i=1}^n \sigma_i^2 = \frac{1}{n^2} \text{tr}(\mathbf{J}^\top \mathbf{D} \mathbf{D}^\top \mathbf{J}) = \frac{1}{n^2} \text{tr}(\mathbf{J} \mathbf{J}^\top \mathbf{D} \mathbf{D}^\top) = \frac{1}{n^2} \text{tr}(\bar{K} \bar{L}) = \frac{1}{n^2} \text{tr}(\mathbf{K} \mathbf{H} \mathbf{L} \mathbf{H}) \quad (37)$$

where we used the cyclic property of the trace and the centering matrix  $\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top$ .

Finally, the squared *Hilbert–Schmidt* norm of the empirical cross-covariance operator reproduces the same scalar quantity:

$$\begin{aligned} \|\widehat{\text{Cov}}_{yx}\|_{\text{HS}}^2 &= \left\langle \frac{1}{n} \sum_{i=1}^n \bar{\psi}(\mathbf{y}_i) \otimes \bar{\phi}(\mathbf{x}_i), \frac{1}{n} \sum_{j=1}^n \bar{\psi}(\mathbf{y}_j) \otimes \bar{\phi}(\mathbf{x}_j) \right\rangle_{\text{HS}} \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \langle \bar{\psi}(\mathbf{y}_j), (\bar{\psi}(\mathbf{y}_i) \otimes \bar{\phi}(\mathbf{x}_i)) \bar{\phi}(\mathbf{x}_j) \rangle_{\mathcal{G}} \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \langle \bar{\phi}(\mathbf{x}_i), \bar{\phi}(\mathbf{x}_j) \rangle_{\mathcal{F}} \langle \bar{\psi}(\mathbf{y}_j), \bar{\psi}(\mathbf{y}_i) \rangle_{\mathcal{G}} \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \bar{k}_{ij} \bar{l}_{ji} = \frac{1}{n^2} \text{tr}(\bar{K} \bar{L}) = \frac{1}{n^2} \text{tr}(\mathbf{K} \mathbf{H} \mathbf{L} \mathbf{H}) \end{aligned} \quad (38)$$

In the next section we show that, in the empirical setting, one can work with a *finite-dimensional* feature map obtained from a factorization of the kernel matrix. This representation makes subsequent covariance expressions far more transparent.

#### A.4 Finite-Dimensional Feature Map

Let  $\mathcal{F}: \mathbf{x} \mapsto \phi(\mathbf{x})$  be an RKHS with reproducing kernel  $k(\cdot, \mathbf{x})$ . By the Representer Theorem (cf. Eq. (30)), any solution subject to norm constraints admits the form:

$$f = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i) \quad (39)$$

so  $f$  lies in the finite subspace  $\mathcal{X} := \text{span}\{\phi(\mathbf{x}_i)\}_{i=1}^n$ . For  $f, g \in \mathcal{X}$  we have:

$$\langle f, g \rangle_{\mathcal{F}} = \left\langle \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i), \sum_{j=1}^n \beta_j \phi(\mathbf{x}_j) \right\rangle_{\mathcal{F}} = \alpha^\top \mathbf{K} \beta, \quad (40)$$

where  $\mathbf{K}$  is the kernel matrix  $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ . Because  $\mathbf{K}$  is symmetric positive semi-definite it admits a factorization:

$$\mathbf{K} = \mathbf{J} \mathbf{J}^\top \quad (41)$$

with  $\mathbf{J} \in \mathbb{R}^{n \times r}$  full rank. Row  $i$  of  $\mathbf{J}$  (denoted by  $\mathbf{J}_i^\top$ ) therefore provides an *implicit finite-dimensional feature vector* for  $\phi(\mathbf{x}_i)$ : indeed  $\mathbf{J}_i^\top \mathbf{J} = \mathbf{K}_{i\cdot} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{F}}$ . Moreover:

$$\langle f, g \rangle_{\mathcal{F}} = \alpha^\top \mathbf{J} \mathbf{J}^\top \beta = (\mathbf{J}^\top \alpha)^\top (\mathbf{J}^\top \beta) = \mathbf{u}^\top \mathbf{v} \quad (42)$$

where we have set  $\mathbf{u} = \mathbf{J}^\top \boldsymbol{\alpha}$  and  $\mathbf{v} = \mathbf{J}^\top \boldsymbol{\beta}$ . Hence  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^r$  are equivalence of  $f$  and  $g$  in this finite-dimensional feature map. Using the reproducing property, the centred evaluation vector  $\bar{\mathbf{f}} = [f(\mathbf{x}_1) - \mu_f, \dots, f(\mathbf{x}_n) - \mu_f]^\top$ , with  $\mu_f = \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i)$ , satisfies :

$$\bar{\mathbf{f}} = \mathbf{J}\mathbf{u} - \mathbf{1}\left(\frac{1}{n}\mathbf{1}^\top \mathbf{J}\mathbf{u}\right) = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\mathbf{J}\mathbf{u} = \mathbf{H}\mathbf{J}\mathbf{u} \quad (43)$$

where  $\mathbf{H}$  is the centring matrix. Now let  $\mathcal{G}: \mathbf{y} \mapsto \psi(\mathbf{y})$  with Gram matrix  $\mathbf{L} = \mathbf{D}\mathbf{D}^\top$ . Repeating the same construction yields  $\bar{\mathbf{g}} = \mathbf{H}\mathbf{D}\mathbf{v}$  for  $g(\mathbf{y}) = \sum_{j=1}^n \beta_j \psi(\mathbf{y}_j)$ . Thus the empirical objective becomes:

$$\begin{aligned} \sup_{g \in \mathcal{G}} \sup_{f \in \mathcal{F}} \mathbb{E}[\bar{f}(\mathbf{x}) \bar{g}(\mathbf{y})] &\approx \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - \mu_f) (g(\mathbf{y}_i) - \mu_g) \\ &= \frac{1}{n} (\mathbf{H}\mathbf{D}\mathbf{v})^\top (\mathbf{H}\mathbf{J}\mathbf{u}) \\ &= \frac{1}{n} \mathbf{v}^\top \mathbf{D}^\top \mathbf{H}\mathbf{J}\mathbf{u} = \mathbf{v}^\top \hat{\mathbf{C}}_{yx} \mathbf{u} \end{aligned} \quad (44)$$

where  $\hat{\mathbf{C}}_{yx} = \frac{1}{n} \mathbf{D}^\top \mathbf{H}\mathbf{J}$  is the (finite-sample) cross-covariance matrix corresponding to the empirical cross-covariance operator  $\widehat{\mathbb{C}\text{ov}}_{yx}$ . In the following section we derive the Equation (11).

## A.5 RKHS Encoders for Feature-Space Alignment

Recall from Section 4.2 the following random variables:

- $X^i$ : encoder input at iteration  $i$
- $Z^i$ : encoder output at iteration  $i$
- $X$ : initial representation
- $S$ : undesired concept labels
- $Y$ : target task labels

Let  $\mathcal{F}$  be an RKHS in which we seek encoders  $f$  that optimize the following objective:

$$\begin{aligned} \sup_{\{g_{\mathcal{I}}\}} \sup_f \mathbb{E}[\bar{g}_{x^i}(X^i) \bar{f}(Z^i)] + \tau_x \mathbb{E}[\bar{g}_x(X) \bar{f}(Z^i)] + \tau_y \mathbb{E}[\bar{g}_y(Y) \bar{f}(Z^i)] \\ \text{s.t. } \sup_{g_s} \mathbb{E}[\bar{g}_s(S) \bar{f}(Z^i)] = 0, \quad \|g_{\mathcal{I}}\|_{\mathcal{G}_{\mathcal{I}}} = \|f\|_{\mathcal{F}} = \|g_s\|_{\mathcal{G}_s} = 1 \end{aligned} \quad (45)$$

Here,  $\mathcal{G}_{\mathcal{I}}$  (with  $\mathcal{I} = \{x^i, x, y\}$ ) and  $\mathcal{G}_s$  refer to corresponding RKHSs for  $g_{\mathcal{I}}$  and  $g_s$ . Let the kernel matrices for  $Z^i, X^i, X, Y$ , and  $S$  be factorized as

$$\mathbf{K}_{z^i} = \mathbf{L}_{z^i} \mathbf{L}_{z^i}^\top, \quad \mathbf{K}_{x^i} = \mathbf{J}_{x^i} \mathbf{J}_{x^i}^\top, \quad \mathbf{K}_x = \mathbf{J}_x \mathbf{J}_x^\top, \quad \mathbf{K}_y = \mathbf{L}_y \mathbf{L}_y^\top, \quad \mathbf{K}_s = \mathbf{L}_s \mathbf{L}_s^\top$$

Using the finite-dimensional feature representation discussed earlier, the empirical estimate for each term in the objective, for instance  $\mathbb{E}[\bar{g}_y(Y) \bar{f}(Z^i)]$ , can be written as:

$$\mathbb{E}[\bar{g}_y(Y) \bar{f}(Z^i)] \approx \frac{1}{n} \sum_{j=1}^n \bar{g}_y(\mathbf{y}_j) \bar{f}(\mathbf{z}_j^i) = \frac{1}{n} \mathbf{u}_y^\top \mathbf{L}_y^\top \mathbf{H} \mathbf{L}_{z^i} \mathbf{w} = \mathbf{u}_y^\top \hat{\mathbf{C}}_{yz^i} \mathbf{w} \quad (46)$$

where  $\mathbf{u}_y$  and  $\mathbf{w}$  are the equivalent vectors to  $g_y$  and  $f$  in the corresponding finite dimensional feature map, and  $\hat{\mathbf{C}}_{yz^i}$  denotes the empirical cross-covariance matrix. Note that the following two optimization problems over  $\mathbf{w}$  are equivalent:

$$\sup_{\mathbf{u}_y} \sup_{\mathbf{w}} \mathbf{u}_y^\top \hat{\mathbf{C}}_{yz^i} \mathbf{w} \quad \equiv \quad \sup_{\mathbf{w}} \mathbf{w}^\top \hat{\mathbf{C}}_{yz^i}^\top \hat{\mathbf{C}}_{yz^i} \mathbf{w} \quad (47)$$

This equivalence follows since the optimal  $\mathbf{u}_y$  is aligned with  $\hat{\mathbf{C}}_{yz^i} \mathbf{w}$  and satisfies:

$$\mathbf{u}_y = \frac{\hat{\mathbf{C}}_{yz^i} \mathbf{w}}{\|\hat{\mathbf{C}}_{yz^i} \mathbf{w}\|_2} \quad (48)$$

437 Hence, the full empirical objective becomes:

$$\sup_{\mathbf{w}} \mathbf{w}^\top (\hat{\mathbf{C}}_{x^i z^i}^\top \hat{\mathbf{C}}_{x^i z^i} + \tau_x \hat{\mathbf{C}}_{xz^i}^\top \hat{\mathbf{C}}_{xz^i} + \tau_y \hat{\mathbf{C}}_{yz^i}^\top \hat{\mathbf{C}}_{yz^i}) \mathbf{w} \quad (49)$$

438 Now consider the constraint in Eq. (45), which ensures that  $f$  does not increase alignment with the  
439 undesired attribute  $S$ . Its empirical estimate becomes:

$$\sup_{g_s} \mathbb{E}[\bar{g}_s(S) \bar{f}(Z^i)] \approx \sup_{\mathbf{u}_s} \mathbf{u}_s^\top \hat{\mathbf{C}}_{sz^i} \mathbf{w} = 0 \quad (50)$$

To satisfy this constraint, we require

$$\mathbf{w} \in \text{Null}(\hat{\mathbf{C}}_{sz^i})$$

440 Let  $\mathbf{Q}$  be an orthonormal basis for this null space. Then we can write:

$$\mathbf{w} = \mathbf{Q} \mathbf{v} \quad (51)$$

441 Since orthonormal transformations preserve norm, the constraint remains unchanged. Substituting  
442 into the objective yields:

$$\begin{aligned} \sup_{\mathbf{v}} \mathbf{v}^\top \mathbf{Q}^\top (\hat{\mathbf{C}}_{x^i z^i}^\top \hat{\mathbf{C}}_{x^i z^i} + \tau_x \hat{\mathbf{C}}_{xz^i}^\top \hat{\mathbf{C}}_{xz^i} + \tau_y \hat{\mathbf{C}}_{yz^i}^\top \hat{\mathbf{C}}_{yz^i}) \mathbf{Q} \mathbf{v} \\ \text{s.t. } \|\mathbf{v}\|_2 = 1 \end{aligned} \quad (52)$$

443 This is a Rayleigh quotient maximization problem. The optimal solution  $\mathbf{v}$  corresponds to the  
444 eigenvector of the matrix with the largest eigenvalue. Define:

$$\mathbf{A} = \mathbf{Q}^\top (\hat{\mathbf{C}}_{x^i z^i}^\top \hat{\mathbf{C}}_{x^i z^i} + \tau_x \hat{\mathbf{C}}_{xz^i}^\top \hat{\mathbf{C}}_{xz^i} + \tau_y \hat{\mathbf{C}}_{yz^i}^\top \hat{\mathbf{C}}_{yz^i}) \mathbf{Q} \quad (53)$$

445 The matrix  $\mathbf{A}$  is a sum of symmetric positive semi-definite matrices, and is itself symmetric and PSD.  
446 Using its eigen-decomposition:

$$\begin{aligned} \mathbf{A} &= \mathbf{D} \mathbf{\Lambda} \mathbf{D}^\top \\ \mathbf{v}^\top \mathbf{A} \mathbf{v} &= \mathbf{v}^\top \mathbf{D} \mathbf{\Lambda} \mathbf{D}^\top \mathbf{v} = \mathbf{a}^\top \mathbf{\Lambda} \mathbf{a} = \sum a_i^2 \lambda_i \leq \lambda_{\max} \end{aligned} \quad (54)$$

447 where  $\mathbf{a} = \mathbf{D}^\top \mathbf{v}$  and  $\|\mathbf{a}\|_2^2 = 1$ , since the eigenvectors of symmetric matrices form an orthonormal  
448 basis. Subsequent encoders, constrained to be orthogonal to the previously selected ones, can be  
449 obtained by iteratively maximizing the same Rayleigh quotient over orthogonal complements. By  
450 the *Courant–Fischer* min–max principle, these correspond to the eigenvectors associated with the  
451 next largest eigenvalues of  $\mathbf{A}$ . One may then select the top  $d$  eigenvectors—ranked by eigenvalue  
452 magnitude or a normalized criterion—as encoder directions for the next iteration.

## 453 B Implementation Details

### 454 B.1 Oblivator’s Training Procedure

455 The practical implementation of **Oblivator** is outlined in Algorithm 1. Depending on the availability  
456 of target task labels, the erasure setting is configured as either supervised (when target labels are  
457 available) or unsupervised. In the supervised case, an additional term involving  $\tau_y \mathbf{K} y$  is included  
458 in the encoder loss (13), and correspondingly in the RKHS-based eigenvalue problem (11) as  
459  $\tau_y \hat{\mathbf{C}}_{yz^i}^\top \hat{\mathbf{C}}_{yz^i}$ . This term is omitted under the unsupervised setting.

460 The encoder is then trained for a fixed number of epochs, which must be sufficient to enable effective  
461 transfer of information from the original representation. If the encoder is trained for too few epochs,  
462 its capacity to preserve relevant features may be limited. In such cases, one practical strategy is to  
463 pre-train the encoder for a few epochs without applying the undesired-concept removal term, and  
464 then perform erasure for a smaller number of epochs. However, excessive pretraining can also be  
465 detrimental: the encoder may learn overly complex representations that are no longer smooth enough  
466 to be effectively constrained by the smooth witness functions. This, in turn, may necessitate stronger  
467 (and potentially less smooth) witnesses, which can hinder erasure quality. Nonetheless, since initial  
468 representations from language models are typically expressive, even random projections can retain  
469 most task-relevant information. Thus, the encoder can often preserve key features with relatively

---

**Algorithm 1** Obliviator Training Procedure

---

```
1: Input: data  $\{x_i\}$ , unwanted labels  $\{s_i\}$ , optional target labels  $\{y_i\}$ 
2: for  $j = 1$  to  $M$  do
3:   if  $\{y_i\}$  is available then
4:     RVs  $\leftarrow (x^j, x, y)$  ▷ Supervised
5:   else
6:     RVs  $\leftarrow (x^j, x)$  ▷ Unsupervised
7:   end if
8:   Train encoder  $\varepsilon^j$  using loss (13)  $\leftarrow$  RVs ▷ Adversarial Training
9:    $z^j \leftarrow \varepsilon^j(x^j)$ 
10:  Solve EVP (11), obtain encoder ▷ RKHS Refinement
11:   $x^{j+1} \leftarrow$  encode  $z^j$  via (12)
12: end for
```

---

few training iterations.<sup>6</sup> In our experiments, we found that 10–15 full-batch iterations were typically sufficient.

After training the encoder, its output is passed to the eigenvalue problem defined in (11). Since the feature space induced by the kernel can be high-dimensional—even for a moderate number of training samples—explicit factorization of the kernel matrix is generally intractable. To address this, we employ approximation methods such as the Nyström method or Random Fourier Features (RFF) [22] to obtain a finite-dimensional approximation of the feature map. Next, the resulting eigenvalues are normalized by the largest eigenvalue, and the top  $d$  eigenvectors are selected based on a predefined threshold. These eigenvectors correspond to functions in the RKHS, which serve as new encoder. We apply these functions to the encoder’s output to generate a transformed representation. This transformed representation then becomes the input to the encoder in the subsequent iteration.

## B.2 Datasets

We evaluate on three benchmark datasets commonly used for concept erasure: BIAS IN BIOS, DIAL-SENTIMENT and DIAL-MENTION.

- **BIAS IN BIOS** [8]: This dataset consists of biographical texts, each annotated with a profession (the primary task) and a gender label (the sensitive/protected attribute). It includes 28 distinct professions and two gender categories, with 53.7% of the data associated with male subjects and 46.3% with female subjects. The most common profession in the dataset is "professor," accounting for 30% of the total samples. To ensure a fair comparison with FaRM, we used the same dataset split as FaRM [6] for the fine-tuned BERT representations. For the frozen representations, we followed the dataset split used by [24].
- **DIAL** [4]: This dataset consists of two subsets: **DIAL-SENTIMENT** and **DIAL-MENTION**.
  - **DIAL-SENTIMENT** is labeled for sentiment analysis, with sentiment as the primary target variable (*happy* 54.57%, *sad* 45.43%). It also includes *race* labels (*African-American English* 36.93%, *Standard American English* 63.07%).
  - **DIAL-MENTION** is a binary classification dataset for detecting whether a tweet mentions another user (50% conversational, 50% non-conversational). The *race* labels in this subset are equally distributed (50%-50%).

## B.3 Experimental Setup

We use a multilayer perceptron (MLP) as our encoder, consisting of a single hidden layer with 256 units and the SiLU activation function. Optimization is performed using the AdamW optimizer with default hyperparameters. We set the learning rate to  $5 \times 10^{-4}$  and apply a weight decay of 0.001. For the BIAS IN BIOS dataset, the encoder is trained for 50 iterations in the first step and 30 iterations in subsequent steps. Due to the infeasibility of computing the full kernel matrix on the entire dataset, exact kernel-based training is only possible via stochastic gradient descent with a reasonable batch size. However, we observe that with RFF and training on the full-batch consistently

---

<sup>6</sup>This is consistent with the Johnson–Lindenstrauss lemma.

Table 2: Hyperparameters used for the training of **Oblivator** across different datasets.

| Dataset        | Erasure      | Encoder Parameters Eq.(13) |          |          | EVP Parameters Eq.(11) |          |                    |  |
|----------------|--------------|----------------------------|----------|----------|------------------------|----------|--------------------|--|
|                |              | $\tau_{x^i}$               | $\tau_x$ | $\tau_y$ | $\tau_x$               | $\tau_y$ | Threshold          | $\gamma_x$ $\gamma_{x^i}$ $\gamma_{z^i}$ |
| BIAS IN BIOS   | Supervised   | 0.05                       | 0.05     | 3        | 0.5                    | 2        | $10^{-4}$          | 0.05 0.5 0.5                             |
|                | Unsupervised | 0.1                        | 0.1      | —        | 0.7                    | —        | $10^{-4}$          | 0.05 0.5 0.5                             |
| DIAL-SENTIMENT | Supervised   | 0.05                       | 0.05     | 5        | 0.5                    | 3        | $5 \times 10^{-5}$ | 0.05 0.2 0.2                             |
|                | Unsupervised | 0.1                        | 0.1      | —        | 0.7                    | —        | $5 \times 10^{-5}$ | 0.05 0.2 0.2                             |
| DIAL-MENTION   | Supervised   | 0.05                       | 0.05     | 5        | 0.5                    | 3        | $5 \times 10^{-5}$ | 0.05 0.2 0.2                             |
|                | Unsupervised | 0.1                        | 0.1      | —        | 0.7                    | —        | $5 \times 10^{-5}$ | 0.05 0.2 0.2                             |

yields better performance. Consequently, we approximate the kernel using RFF throughout our experiments. The RFF dimensionality is set to 2500 in the first iteration and reduced to 1500 in later steps, as the encoder’s input dimension is reduced after the first transformation. To ensure that HSIC is not artificially increased through isotropic scaling, we evaluated two normalization strategies: (1) feature-wise normalization to unit variance and (2) sample-wise normalization. The latter proves more effective in our setup. For the DIAL-MENTION and DIAL-SENTIMENT datasets all settings are the same except the initial training is set to 30 iterations. All reported trade-off curves are over three independent runs where we reported minimum performance (lowest trade-off curve). For the RKHS encoder, the Kernel matrix is approximated using RFF with a dimension of 1500. The hyper-parameters used for training can be found in Table 2. Training is conducted on a single NVIDIA RTX A6000 GPU. For the training of DeepSeek and LLaMa on BIAS IN BIOS dataset all the parameters are similar to Table 2 except we set the threshold to  $10^{-5}$ .

## C Probing Networks

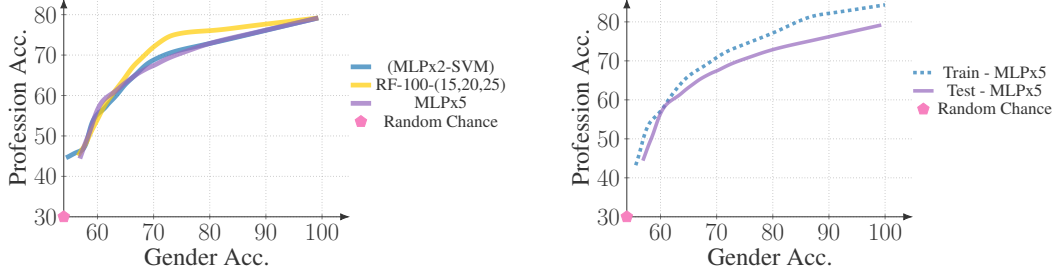
### C.1 Our Choice of Probing Networks

Unlike prior work [1, 6], we do not restrict the nonlinear adversary to the default MLP classifier provided by scikit-learn. Instead, we evaluate all methods using two types of adversarial classifiers: an SVM with an RBF kernel (implemented via cuML [23]) and an MLP with two hidden layers of 128 neurons (implemented in PyTorch [19]). The MLP classifier is trained three times with different random seeds, while the SVM is evaluated under seven hyperparameter settings  $(\gamma, C) \in \{(10, 5), (10, 10), (5, 10), (5, 5), (1, 5), (1, 1), (0.5, 1)\}$ . We report the maximum accuracy obtained across all ten runs (three MLP + seven SVM) for each method. For the target task classifier, we employ the same architecture as the adversarial MLP, a two-layer Multi-Layer Perceptron with 128 neurons per hidden layer. An SVM with a large kernel scale ( $\gamma = 10$ ) and hard margin ( $C = 10$ ) serves as a coarse decision boundary, useful for testing whether erasure occurs by overlapping distributions across unwanted attributes. We must note that, the erasure process is carried by a deterministic function and only shuffling data samples across unwanted attributes is not sufficient for erasure as this shuffling can be still invertible. Therefore, true erasure happens upon distribution matching.

In our experiments, we observed that applying dimensionality reduction often improved adversarial accuracy. For instance, using a strong encoder—an MLP with four hidden layers(see §D.1)—without any iterative refinement led our MLP adversaries to perform at random chance level while SVM with  $\gamma = 10, C = 10$  reported around 80% accuracy. However, when we applied RKHS refinement via (11) as a dimensionality reduction step and re-evaluated the MLP adversaries again, their accuracy was similar to that of SVM with  $\gamma = 10, C = 10$ . This indicates that sensitive information was still present in the representation, and the prior adversary failures were primarily due to training difficulty rather than successful erasure.

### C.2 Ablation Studies with Different Probing Networks

To evaluate the robustness of our probing setup, we assess the trade-off curves using multiple types of classifiers. Specifically, we consider a Multi-Layer Perceptron (MLP) with five hidden layers of size 128, as well as Random Forest classifiers with 100 estimators and maximum depths of 15, 20, and 25. For the MLP, we report the highest accuracy across three independent runs. For the Random Forest, we report the best accuracy obtained across the three depth settings. The resulting trade-off



(a) Comparison of obtained trade-off with deeper MLP (5 hidden layers) and Random Forest with our choice of classifier (MLPx2-SVM).

(b) Train and test accuracy for MLP with 5 hidden layer at different step of the erasure.

Figure 7: Ablation studies with different probing networks. Dataset is BIAS IN BIOS and language model is BERT.

curves are shown in Figure 7a, alongside the curve obtained using our default probing classifier (as described earlier). As illustrated in the figure, our choice of classifier yields a consistently lower trade-off profile, demonstrating its robustness for evaluating erasure performance.

In contrast to prior work, we do not rely on visualizations produced by t-SNE or UMAP to assess overlap with respect to the sensitive attribute (e.g., gender). While such techniques can be visually appealing, they are unreliable indicators of structure in high-dimensional spaces, often distorting distances and cluster separability. Instead, we adopt an evaluation based on the training accuracy of an expressive classifier. In particular, if a sufficiently flexible probing model is unable to overfit the training data, this provides stronger evidence that the representation has been overlapped across the unwanted attribute, suggesting successful erasure. However, it is important to note that we are working with finitely many samples in a high dimensional space. It is unlikely that the data points become so close that no probing classifier can overfit to training samples. For this reason, the flexibility of the probing model must be chosen reasonably so that this evaluation becomes meaningful. Figure 10 presents the train and test accuracy of the MLP with five hidden layers over the course of erasure. As the process progresses, we observe a drop in both training and test accuracy, with the training accuracy falling to near random chance. This indicates that even a high-capacity classifier cannot overfit to the training samples, suggesting that the representation has been successfully aligned across the sensitive groups via distributional overlap.

Taken together, these probing results indicate that **ObliViator** consistently achieves full concept erasure. This is consistent with the theoretical motivation: minimizing  $\text{HSIC}(S, X)$  enforces statistical independence between the representation  $X$  and the sensitive attribute  $S$ .

### C.3 Statistical Significance of the Obtained Trade-off

In this section, we evaluate the statistical significance of the obtained trade-off curve. As described in Section Appendix B.3, each trade-off profile is computed using the evaluation protocol outlined in Section Appendix C.1, and reflects the lowest performance across three independent runs. We repeat this entire process five times and report the mean trade-off profile along with the 95% confidence interval, computed using a Student’s  $t$ -test as  $\pm \sigma / \sqrt{5} \cdot t_{4,0.975}$ . Figure 8 presents results under the *frozen + unsupervised* erasure setting on the BIAS IN BIOS dataset, with BERT as the language model. The dashed curve corresponds to the trade-off profile reported in the main paper, while the solid curve shows the mean across five repeated evaluations. The shaded region represents the confidence interval. The low variance observed across repetitions reflects two complementary factors: 1) The rigor of our evaluation protocol—based on multiple adversaries and conservative selection of worst-case performance, naturally reduces variability; 2) It also highlights the inherent stability of our multi-step framework, where erasure remains consistent across independent runs.

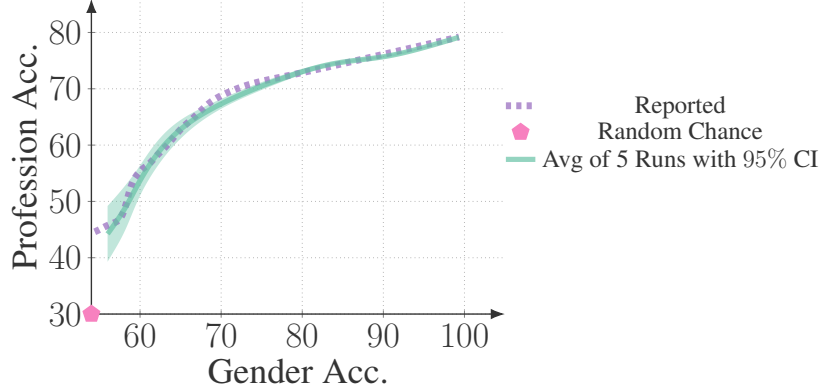


Figure 8: Statistical Significance of the obtained trade-off curve. Dataset is BIAS IN BIOS and language model is BERT.

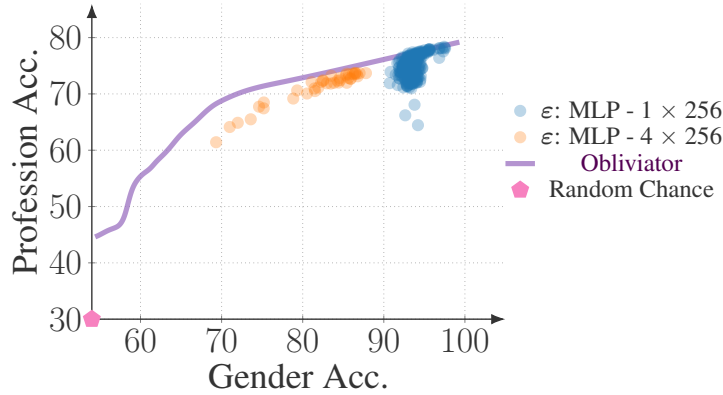


Figure 9: Multi-Step Erasure vs. Single-Step Erasure. Comparison of erasure performance using a single-step encoder with either 1 or 4 hidden layers (MLP), versus the proposed multi-step erasure framework (**Oblivator**). Results are shown on the BIAS IN BIOS dataset using BERT representations. The multi-step approach consistently demonstrates more effective erasure.

## 582 D Additional Results

### 583 D.1 Single Step Erasure Vs Multi-Step Erasure

584 To highlight the importance of each component in **Oblivator**, we evaluate a simplified baseline using  
 585 a single MLP encoder with either one or four hidden layers (each containing 256 neurons). The goal  
 586 is to perform erasure using only the encoder trained with the loss in (13), without employing the  
 587 multi-step framework or RKHS-based refinement. This setup allows us to isolate the contribution of  
 588 these components and assess their effect on erasure and adversarial optimization. During training, we  
 589 evaluate both the target task and undesired attribute accuracy every 30 steps. As shown in Figure 9,  
 590 neither of the single-step encoder configurations achieves full concept erasure. Although both gender  
 591 and profession accuracies initially decline, they eventually plateau and begin to oscillate, failing to  
 592 reach full erasure. While a more expressive encoder might hypothetically improve performance, a  
 593 direct comparison reveals that the empirical trade-off achieved by **Oblivator** is consistently better.  
 594 This demonstrates that the multi-step procedure and RKHS refinement, results in a more robust and  
 595 effective erasure process.

### 596 D.2 Hyperparameter Sensitivity

597 We conduct two additional studies to evaluate the impact of hyperparameters on the performance  
 598 of **Oblivator**. First, we examine the effect of the RBF kernel bandwidth parameters  $\gamma_{z^i}$  and  $\gamma_{x^i}$ ,  
 599 which are defined as the inverse of the kernel width (i.e.,  $\gamma = 1/2\sigma^2$ ). These parameters control the

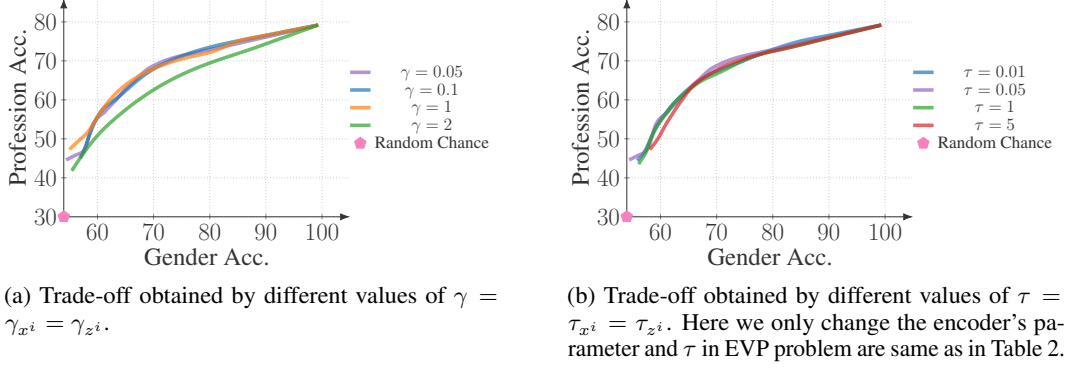


Figure 10: Analysis of the effect of hyperparameters on **Oblivator**'s performance. Dataset is BIAS IN BIOS and language model is BERT.

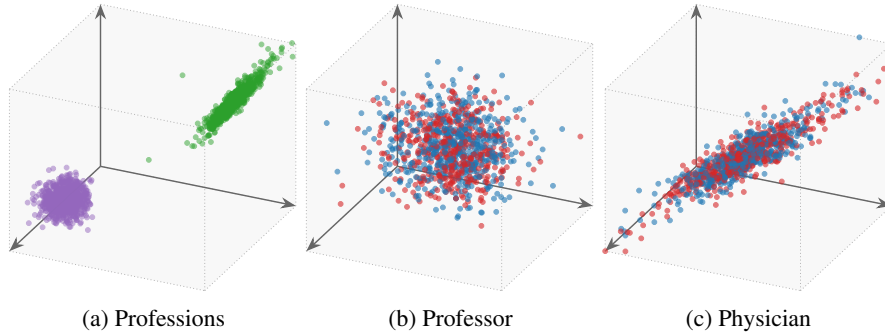


Figure 11: Representations Learned by **Oblivator** on Finetuned BIAS IN BIOS Representations (BERT). The professions Professor and Physician are shown separately to better visualize the distribution of gender within each class. While the two professions are clearly separated (green and purple), gender labels (blue and red) are indistinguishable within each profession—indicating that **Oblivator** effectively erases gender information while preserving task-relevant structure.

smoothness and expressivity of the witness functions used in dependency estimation. Intuitively, higher  $\gamma$  values (corresponding to narrower kernels) yield more complex witness functions, which are more sensitive to fine-grained correlations. This can result in more aggressive erasure by capturing spurious dependencies. This behavior is consistent with the patterns observed in Figure 10a. Furthermore, we can see that **Oblivator** performs consistently across a broad, reasonable range of values for  $\gamma$ , demonstrating robustness to this choice. Note that, in this experiment, we considered  $\gamma = \gamma_{z^i} = \gamma_{x^i}$ .

Next, we study the effect of the weighting coefficients  $\tau_{x^i}$  and  $\tau_x$ , which appear in both the adversarial loss (13) and the eigenvalue problem in (11). In this experiment, we vary only the encoder loss parameters while keeping the EVP parameters fixed, as specified in Table 2. As before, we set  $\tau_{x^i} = \tau_x$ . The results, shown in Figure 10b, indicate that very large values of  $\tau$  lead to a slight degradation in performance. Moreover, increasing  $\tau$  slows convergence, requiring more iterations to reach the same level of erasure.

### D.3 Visualization

Here, we visualize the representations of two professions—Professor and Physician—from the BIAS IN BIOS dataset using fine-tuned representations at an intermediate stage of erasure. This representation corresponds to a point on the trade-off curve where gender classification accuracy is approximately 60%, while profession classification accuracy remains around 85% (see Figure 4c). The visualization in Figure 11 also corresponds to this stage of erasure.

619 At this point, the number of RKHS encoders obtained from (11) is three, meaning we visualize the  
620 actual transformed representation used by the model. Notably, while the two professions remain  
621 well-separated, the gender labels within each profession are indistinguishable—highlighting that  
622 **Oblivator** successfully removes gender information while preserving task-relevant structure.

## 623 **E Societal Impact**

624 **Oblivator** enables targeted concept erasure from learned representations, which can be beneficial—for  
625 instance, by removing sensitive demographic attributes to reduce reliance on them in decision-making,  
626 or by erasing personal identifiers to protect privacy. However, if task-critical attributes are treated as  
627 “unwanted”, such as age in healthcare or income level in finance, model performance may degrade, or  
628 the model may rely on spurious correlations with remaining features like gender or race. Transparent  
629 documentation, domain-informed definitions, and independent oversight are essential, especially in  
630 high-stakes applications. Without careful review, erasure may discard socially meaningful information  
631 or fail to generalize across contexts.

## References

- [1] Somnath Basu Roy Chowdhury, Sayan Ghosh, Yiyuan Li, Junier Oliva, Shashank Srivastava, and Snigdha Chaturvedi. Adversarial scrubbing of demographic information for text classification. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 550–562, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.43. URL <https://aclanthology.org/2021.emnlp-main.43/>.
- [2] Somnath Basu Roy Chowdhury, Nicholas Monath, Kumar Avinava Dubey, Amr Ahmed, and Snigdha Chaturvedi. Robust concept erasure via kernelized rate-distortion maximization. *Advances in Neural Information Processing Systems*, 36:43284–43306, 2023.
- [3] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36, 2024.
- [4] 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*, pages 1119–1130, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1120. URL <https://aclanthology.org/D16-1120>.
- [5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.
- [6] Somnath Basu Roy Chowdhury and Snigdha Chaturvedi. Learning fair representations via rate-distortion maximization. *Transactions of the Association for Computational Linguistics*, 10:1159–1174, 2022.
- [7] Somnath Basu Roy Chowdhury, Nicholas Monath, Kumar Avinava Dubey, Amr Ahmed, and Snigdha Chaturvedi. Robust concept erasure via kernelized rate-distortion maximization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [8] 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*, pages 120–128, 2019.
- [9] DeepSeek-AI. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024. URL <https://github.com/deepseek-ai/DeepSeek-LLM>.
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423/>.
- [11] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations (ICLR2016)*, pages 1–14, May 2016. URL <https://iclr.cc/archive/www/doku.php%3Fid=iclr2016:main.html>. 4th International Conference on Learning Representations, ICLR 2016 ; Conference date: 02-05-2016 Through 04-05-2016.
- [12] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1002. URL <https://aclanthology.org/D18-1002/>.

- [13] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In Jill Burstein, Christy Doran, and Tamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1061. URL <https://aclanthology.org/N19-1061/>.
- [14] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [15] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.
- [16] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, Bernhard Schölkopf, and Aapo Hyvärinen. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(12), 2005.
- [17] Yazhe Li, Roman Pogodin, Danica J Sutherland, and Arthur Gretton. Self-supervised learning with kernel dependence maximization. *Advances in Neural Information Processing Systems*, 34:15543–15556, 2021.
- [18] Yitong Li, Timothy Baldwin, and Trevor Cohn. Towards robust and privacy-preserving text representations. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2005. URL <https://aclanthology.org/P18-2005/>.
- [19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf).
- [20] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL <http://www.aclweb.org/anthology/D14-1162>.
- [21] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- [22] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.
- [23] Sebastian Raschka, Joshua Patterson, and Corey Nolet. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *arXiv preprint arXiv:2002.04803*, 2020.
- [24] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL <https://aclanthology.org/2020.acl-main.647/>.
- [25] Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421. PMLR, 2022.

- 735 [26] Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Adversarial concept erasure in kernel space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6034–6055, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.405. URL <https://aclanthology.org/2022.emnlp-main.405/>.  
736  
737  
738  
739  
740
- 741 [27] Bashir Sadeghi, Sepehr Dehdashtian, and Vishnu Boddeti. On characterizing the trade-off in invariant representation learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL <https://openreview.net/forum?id=3gfpBR1ncr>. Featured Certification.  
742  
743  
744
- 745 [28] Shun Shao, Yftah Ziser, and Shay B. Cohen. Gold doesn’t always glitter: Spectral removal of linear and nonlinear guarded attribute information. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1611–1622, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.118. URL <https://aclanthology.org/2023.eacl-main.118/>.  
746  
747  
748  
749  
750
- 751 [29] Shun Shao, Yftah Ziser, and Shay B Cohen. Erasure of unaligned attributes from neural representations. *Transactions of the Association for Computational Linguistics*, 11:488–510, 2023.  
752  
753
- 754 [30] Ben Verhoeven, Walter Daelemans, and Barbara Plank. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth international conference on language resources and evaluation (LREC’16)*, pages 1632–1637, 2016.  
755  
756
- 757 [31] Liwen Wang, Yuanmeng Yan, Keping He, Yanan Wu, and Weiran Xu. Dynamically disentangling social bias from task-oriented representations with adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3740–3750, 2021.  
758  
759  
760
- 761 [32] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.  
762