Appendix for "On the Interaction Effects Between Prediction and Clustering"

A Proofs

A.1 Theorem 3.1

We begin by introducing a key lemma about the minimization of function mixtures.

Lemma A.1. *For functions $a, b : \Theta \to \mathbb{R}$ and $\alpha \in [0, 1]$,*

$$a(\arg\min_{\theta\in\Theta}(\alpha a(\theta) + (1-\alpha)b(\theta)))$$
$$b(\arg\min_{\theta\in\Theta}(\alpha a(\theta) + (1-\alpha)b(\theta)))$$

are monotonically decreasing and increasing, respectively, with respect to α.

Proof. Let $\Delta(\theta) = a(\theta) - b(\theta)$, $1 \geq j > i \geq 0$ and

$$\theta_i \in \arg\min_{\theta\in\Theta} b(\theta) + i\Delta(\theta)$$
$$\theta_j \in \arg\min_{\theta\in\Theta} b(\theta) + j\Delta(\theta)$$

Then a is monotonically decreasing with respect to α if and only if $a(\theta_i) \geq a(\theta_j)$.

Case 1: $\theta_i = \theta_j$. Then $a(\theta_i) = a(\theta_j)$, $b(\theta_i) = b(\theta_j)$ and the statements holds.

Case 2: $\theta_i \neq \theta_j$. Then both the following conditions must be true.

$$b(\theta_j) - b(\theta_i) + i\Delta(\theta_j) - i\Delta(\theta_i) > 0 \tag{8}$$
$$b(\theta_j) - b(\theta_i) + j\Delta(\theta_j) - j\Delta(\theta_i) < 0 \tag{9}$$

If Eq. (8) did not hold, then θ_j would have been optimal at $\alpha = i$, i.e. $\theta_j \in \arg\min_{\theta\in\Theta} b(\theta) + i\Delta(\theta)$. Likewise, if Eq. (9) did not hold, then $\theta_i \in \arg\min_{\theta\in\Theta} b(\theta) + j\Delta(\theta)$.

Together, they imply

$$\begin{aligned} b(\theta_j) - b(\theta_i) + i\Delta(\theta_j) - i\Delta(\theta_i) &> b(\theta_j) - b(\theta_i) + j\Delta(\theta_j) - j\Delta(\theta_i) \\ i\Delta(\theta_j) - i\Delta(\theta_i) &> j\Delta(\theta_j) - j\Delta(\theta_i) \\ (i-j)(\Delta(\theta_j) - \Delta(\theta_i)) &> 0 \\ \Delta(\theta_j) - \Delta(\theta_i) &< 0 \end{aligned}$$

since $i - j < 0$. Plugging this into Eq. (8),

$$\begin{aligned} b(\theta_j) - b(\theta_i) + i\Delta(\theta_j) - i\Delta(\theta_i) &> 0 \\ b(\theta_j) - b(\theta_i) &> i(\Delta(\theta_i) - \Delta(\theta_j)) \\ b(\theta_j) - b(\theta_i) &> 0 \end{aligned} \tag{10}$$

which proves the second statement. Finally, plugging Eq. (10) into Eq. (9) concludes the proof.

$$\begin{aligned} (1-j)(b(\theta_j) - b(\theta_i)) + j(a(\theta_j) - a(\theta_i)) &< 0 \\ a(\theta_j) - a(\theta_i) &< 0 \end{aligned}$$

□

The proof of Theorem 3.1 follows.

Proof. **Direction $\mathcal{V}$ to $\mathcal{T}$** We say f is optimal under its training distribution if

$$f(\cdot|\mathcal{T}) \in \arg\min_{f\in\mathcal{F}} \mathbb{E}_{x,y\sim P_\mathcal{T}} \ell(f(x), y).$$

Let $f_0, f_1, \ldots, f_n$ be models learned at each level of dependency leakage, such that each model is optimal under its training distribution, i.e.

$$f_i \in \arg\min_{f\in\mathcal{F}} \mathbb{E}_{x,y\sim M_{P_\mathcal{T},P_\mathcal{V}}(1-\frac{i}{n},\frac{i}{n})} \ell(f(x), y).$$

The sequence $e_0, e_1, \ldots, e_n$ is monotonically decreasing when

$$e_i - e_{i+1} \geq 0 \quad \forall i \in \{0, \ldots, n-1\}.$$

Starting from the definition of e and using the notational shorthand $\ell_P(f) = \mathbb{E}_{x,y\sim P}\ell(f(x), y)$,

$$\begin{aligned}
e_i &= \mathbb{E}_{x,y\sim P_\mathcal{V}} \ell(f_i(x), y) \\
&= \ell_{P_\mathcal{V}}(f_i) \\
&= \ell_{P_\mathcal{V}}(\arg\min_{f\in\mathcal{F}} \mathbb{E}_{x,y\sim M_{P_\mathcal{T},P_\mathcal{V}}\left(1-\frac{i}{n},\frac{i}{n}\right)} \ell(f(x), y)) \\
&= \ell_{P_\mathcal{V}}\left(\arg\min_{f\in\mathcal{F}} \frac{i}{n}\ell_{P_\mathcal{V}}(f) + \left(1-\frac{i}{n}\right)\ell_{P_\mathcal{T}}(f)\right)
\end{aligned} \tag{11}$$

By Lemma A.1, e is monotonically decreasing with respect to $\frac{i}{n}$, and thus also with respect to i since n is a fixed constant.

Direction $\mathcal{T}$ to $\mathcal{V}$. In this direction, e will further be linear:

$$\begin{aligned}
e_0 &= \mathbb{E}_{x,y\sim P_\mathcal{V}, \mathcal{T}\overset{n}{\sim}P_\mathcal{T}} \ell(f(x|\mathcal{T}), y) \\
e_n &= \mathbb{E}_{x,y\sim P_\mathcal{T}, \mathcal{T}\overset{n}{\sim}P_\mathcal{T}} \ell(f(x|\mathcal{T}), y) \\
e_i &= \mathbb{E}_{x,y\sim M_{P_\mathcal{T},P_\mathcal{V}}\left(\frac{i}{n},1-\frac{i}{n}\right), \mathcal{T}\overset{n}{\sim}P_\mathcal{T}} \ell(f(x|\mathcal{T}), y) \\
&= \left(\frac{i}{n}\right) \mathbb{E}_{x,y\sim P_\mathcal{T}, \mathcal{T}\overset{n}{\sim}P_\mathcal{T}} \ell(f(x|\mathcal{T}), y) + \left(1-\frac{i}{n}\right) \mathbb{E}_{x,y\sim P_\mathcal{V}, \mathcal{T}\overset{n}{\sim}P_\mathcal{T}} \ell(f(x|\mathcal{T}), y) \\
&= \left(\frac{i}{n}\right) e_n + \left(1-\frac{i}{n}\right) e_0
\end{aligned}$$

and $e_n \leq e_0$ by the assumption that f is optimal under its training distribution. □

A.2 Theorem 3.2

We begin by introducing a lemma on the minimization of mixtures of convex functions.

Lemma A.2. *For $\alpha \in [0, 1]$, let $a, b : \Theta \to \mathbb{R}$ be strictly convex and differentiable (where $\dot{a}$ denotes $\frac{\partial a}{\partial \theta}$) over*

$$\begin{aligned}
\Theta^* &= \{\theta \in \arg\min_{\theta\in\Theta}(\alpha b(\theta) + \alpha\Delta(\theta))\} \quad \forall \alpha \in [0, 1] \\
&= \{\theta \in g(\alpha)\} \quad \forall \alpha \in [0, 1] \subseteq \Theta.
\end{aligned}$$

If $\frac{\dot{a}}{\dot{b}}$ is convex, decreasing over Θ^*, then

$$a(\arg\min_{\theta\in\Theta}(\alpha a(\theta) + (1-\alpha)b(\theta)))$$

is convex over α.

Proof. If $\frac{\dot{a}}{\dot{b}}$ is convex, decreasing then $\frac{-\dot{b}}{\dot{\Delta}}$ is also convex decreasing.

$$\frac{\dot{a}}{\dot{b}}\text{convex, decreasing} \Leftrightarrow \frac{-\dot{\Delta}}{\dot{b}}\text{concave, increasing} \tag{12}$$

because $\frac{-\dot{\Delta}}{\dot{b}} = \frac{\dot{b}-\dot{a}}{\dot{b}} = 1 - \frac{\dot{a}}{\dot{b}}$.

Further, we know $\frac{-\dot{\Delta}}{\dot{b}} \geq 0$ because $\dot{a} \leq 0$ and $\dot{b} \geq 0$ by Lemma A.1. Then $\frac{-\dot{b}}{\dot{\Delta}}$ is convex decreasing by the composition of the convex, decreasing function $\frac{1}{x}$ and the concave increasing $\frac{-\dot{\Delta}}{\dot{b}}$. Note in the case where $\dot{\Delta} = 0$, $g(\alpha)$ is constant and the lemma holds.

At the minimum of $b(\theta) + \alpha\Delta(\theta)$,

$$0 = \dot{b} + \alpha\dot{\Delta}$$
$$\alpha = \frac{-\dot{b}}{\dot{\Delta}}$$

Thus, $g^{-1}(\theta) = \frac{-\dot{b}}{\dot{\Delta}}$ is convex, decreasing and $g(\alpha)$ is concave, increasing. Finally $a(g(\alpha))$ is convex, decreasing by the composition of a convex, non-increasing and concave function. □

The proof for Theorem 3.2 follows.

Proof. **Direction $\mathcal{T}$ to $\mathcal{V}$** Holds by Theorem 3.1, as linearity implies convexity.

Direction $\mathcal{V}$ to $\mathcal{T}$ Starting from the definition of e and using the notational shorthand $\ell_P(f) = \mathbb{E}_{x,y\sim P}\ell(f(x), y)$,

$$\begin{aligned} e_i &= \mathbb{E}_{x,y\sim P_{\mathcal{V}}}\ell(f_i(x), y) \\ &= \ell_{P_{\mathcal{V}}}(f_i) \\ &= \ell_{P_{\mathcal{V}}}(\arg\min_{f\in\mathcal{F}} \mathbb{E}_{x,y\sim M_{P_{\mathcal{T}},P_{\mathcal{V}}}(1-\frac{i}{n},\frac{i}{n})}\ell(f(x), y)) \\ &= \ell_{P_{\mathcal{V}}}\left(\arg\min_{f\in\mathcal{F}} \frac{i}{n}\ell_{P_{\mathcal{V}}}(f) + \left(1 - \frac{i}{n}\right)\ell_{P_{\mathcal{T}}}(f)\right) \end{aligned} \tag{13}$$

By Lemma A.2, e is convex with respect to $\frac{i}{n}$, and thus also with respect to i since n is a fixed constant. □

A.3 Proof of Theorem 5.1

Proof.

$$(A\Psi)_{ij} = \mathop{\mathbb{E}}_{k_n\sim\text{Binomial}(n,p_i)}\psi_j\left(\frac{k_n}{n}\right)$$

By the weak law of large numbers, $\frac{k_n}{n} \xrightarrow{p} p_i$. Further, $\psi_j\left(\frac{k_n}{n}\right) \xrightarrow{p} \psi_j(p_i)$ by the continuous mapping theorem. Finally, $\mathbb{E}\psi_j\left(\frac{k_n}{n}\right) \rightsquigarrow \mathbb{E}\psi_j(p_i) = \psi_j(p_i)$ by the Portmanteau lemma. The matrix formed by $\psi_j(p_i)$ is invertible by the Unisolvence theorem when $p_0, \ldots, p_s$ are unique. □

A.4 Proof of Theorem 5.2

Proof. Let e' be the solution to the sketched system $Se' = b$. Then

$$\begin{aligned} Se' = Ae &= b \\ e' &= S^{-1}Ae \\ e'_0 &= (S^{-1}A)_{00}e_0 + \sum_{i=1}^{n}(S^{-1}A)_{0i}e_i \\ e'_0 - e_0 &= \sum_{i=1}^{n}(S^{-1}A)_{0i}e_i. \end{aligned}$$

Let s' be the first row of S^{-1}. By the Cauchy-Schwarz inequality, for all $i \geq 1$

$$\begin{aligned}|(S^{-1}A)_{0i}| &= |s' \cdot A_i| \\ &= |s' \cdot A_i - s' \cdot S_{r(i)}| \\ &\leq \|s'\| \|A_i - S_{r(i)})\| \\ &= \epsilon \|s'\|.\end{aligned}$$

Finally, by Theorem 3.1

$$\begin{aligned}|e'_o - e_0| &\leq \sum_{i=1}^{n} |(S^{-1}A)_{0i}| e_i \\ &\leq \sum_{i=1}^{n} \epsilon \|s'\| e_i \\ &\leq \epsilon n \|s'\| e_0.\end{aligned}$$

□

B Connections to the Bernstein basis and Bézier curves

B.1 Bernstein basis

Recall that a Bernstein basis of degree n is defined as

$$b_{j,n}(x) = \binom{n}{j} x^j (1-x)^{n-j} \quad j = 0, \ldots, n \tag{14}$$

and that this forms a basis for polynomials at most degree n. Then the Bernstein polynomial is defined as

$$B_n(x) = \sum_{j=0}^{n} \beta_j b_{j,n}(x) \tag{15}$$

where B_j are the Bernstein coefficients. The B3 estimator $b_i = \sum_{j=0}^{n} e_j A_{ij}$ is equivalent to solving for the Bernstein coefficients $e_j = \beta_j$, where the Berstein basis is $A_{ij} = b_{j,n}(p_i)$.

B.2 Bézier curves

Bézier curves are closely related to Bernstein polynomials, using slightly different notation

$$B(t) = \sum_{j=0}^{n} \binom{n}{j} t^j (1-t)^{n-j} \mathbf{P}_j \tag{16}$$

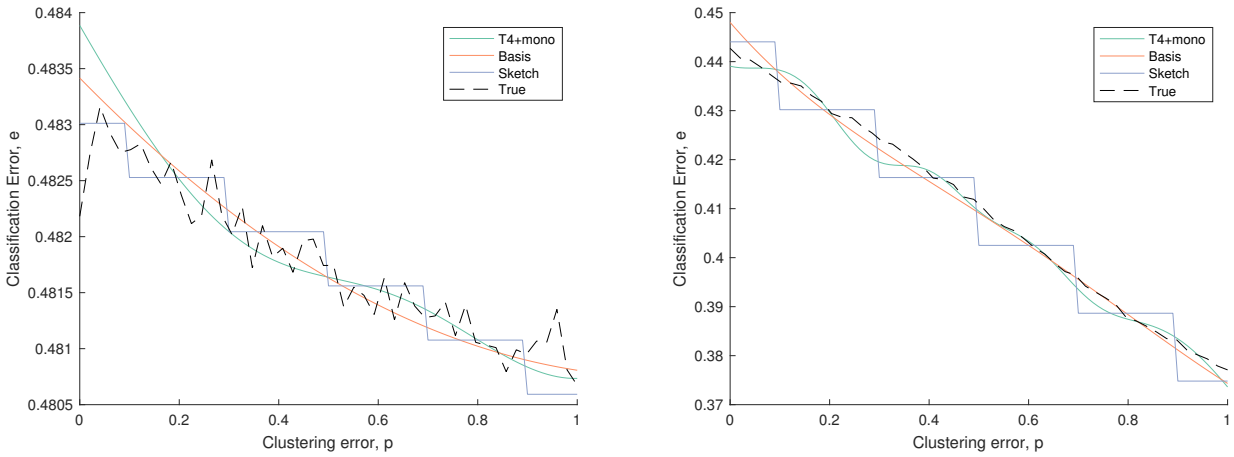
$$= \sum_{j=0}^{n} b_{j,n}(t) \mathbf{P}_j \tag{17}$$

where $\mathbf{P}_j$ are the Bézier control points. Once again, A_{ij} from the B3 estimator is equivalent to the Bernstein basis function $b_{j,n}(p_i)$, and we solve for the Bézier control points $\mathbf{P}_0, \ldots, \mathbf{P}_n$

C Additional Experiments

C.1 Additions to Fig. 2

Due to space limitations, the Dota 2 and Parkinson's experiments were excluded from Fig. 2. We include them here in Fig. 4 to demonstrate the theoretical properties in Section 3 held across all datasets. The Dota 2 experiments had higher variance than others, though the monotonic and convex trend appears to hold true.



(a) Dota 2 (b) Parkinson's

Figure 4: Additions to Fig. 2. Empirical results show the loss is indeed convex and monotonically decreasing, validating our theoretical results in Section 3. Note our methods are able to recover the full loss in addition to the true OOC loss e_0.

C.2 Synthetic Experiments

As an initial exploratory experiment, we generated synthetic data according to a partition model with $k = 2$ parts, $m = 2n$ rows in A (i.e. levels of dependency leakage) and initial dependency leakage probability $p_0 = 0.1$. Our learner is a linear regression model with mean squared error loss. Our exploratory synthetic results, presented in the Figure 5 box plot, demonstrate that the basis function and matrix sketching estimators perform comparably or better than baseline methods.

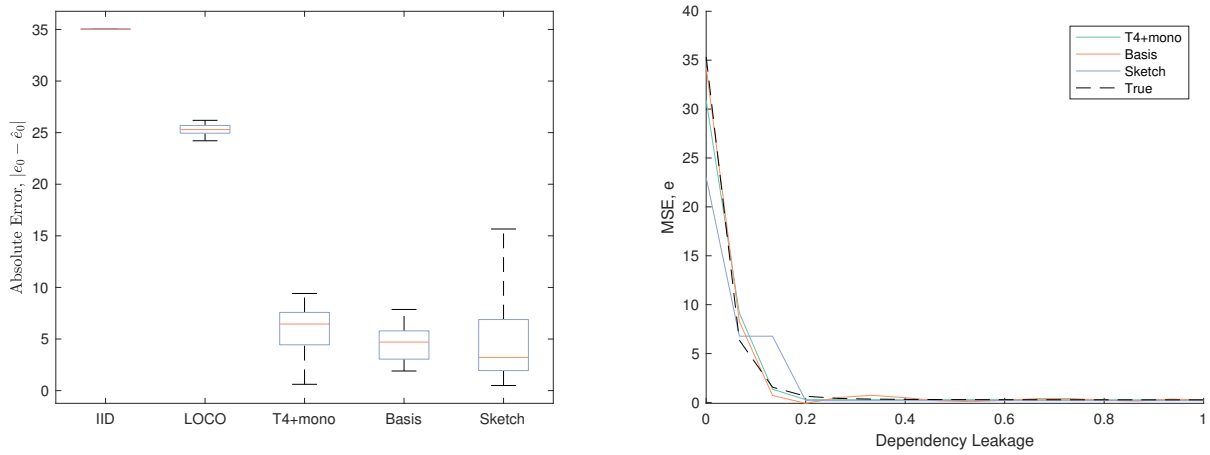


Figure 5: Synthetic regression results.

D Experimental Details

In the sketching approximation, we formed k nearly equally sized groups of adjacent columns from A when forming the sketched matrix S. Even after sketching, we found it beneficial to add some regularization comparable to T4+mono, referred to as λ_s (the regularization used in T4+mono is referred to as λ_{T4}). We found that other approaches, including using k-medoids to group the columns of A, did not provide any benefits and were more complicated. In all experiments we set $k = 7$.

In the basis function approximation, we found that using simple, low-order polynomials was sufficient. Higher order polynomials tended to be unstable. After observing b, we chose to use either a 2nd or 7th order polynomial, depending on the curvature of b.

The whisker plots in Fig. 3 are generated over 10 independent trials, where the whiskers correspond to most extreme values over those trials (i.e. no outliers removed).

The complete set of experimental parameters are shown in Table 2. We made an effort to limit fitting to a specific dataset, and kept most parameters the same across all experiments. In the Dota 2 experiments, the availability of sufficient training data allowed us to increase $|\mathcal{T}|$ to 1000. Further, after completing the Heart and Census experiments, we reduced the number of rows m in A by an order of magnitude to speed up experimentation, and correspondingly increased the regularization λ.

Table 2: Parameters used in all experiments. n is the number of samples in the training set, $|\mathcal{V}|$ is the number of samples in the validation set, t is the number of resamples in Algorithm 1, λ's are the regularization strengths in the T4+mono and sketching method, m is the number of corruption levels (i.e. the number of rows in A), k is the number of sketching groups and d is the number of features in the dataset.

	Parameter													
Dataset	n	$\lvert\mathcal{T}\rvert$	$\lvert\mathcal{V}\rvert$	d	t	λ_{T4}	λ_s	s	m	k	Latent cluster	Training clusters	Validation clusters	Features
Synthetic	∞	15	1000	2	1000	0.1	0.01	7	30	10	-	-	-	-
Heart[a]	100	100	100	12	1000	10	0.1	7	200	7	Location	Cleveland, VA, Switzerland	Hungary	age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, thal
1994 US Census[b]	100	100	100	5	10000	10	0.1	7	200	7	Native country	United States, El Salvador, Germany, Mexico, Philippines, Puerto Rico	India, Canada	age, education_num, hours_per_week, race, occupation
Parkinson[c]	100	100	100	26	10000	1000	0.1	2	20	7	Subject	$2, 3, 4, 6, 7, 8, \ldots$	$1, 5, 9, \ldots$	jitter_local, jitter_abs, jitter_rap, jitter_ppq5, jitter_ddp, shimmer_local, shimmer_db, shimmer_apq3, shimmer_apq5, shimmer_apq11, shimmer_dda, ac, nth, htn, median_pitch, mean_pitch, std_dev, min_pitch, max_pitch, pulses, periods, mean_period, std_dev_period, unvoiced, breaks, deg_breaks
Dota 2[d]	100	100	100	114	1000	1000	0.1	2	20	7	Type	1, 2	3	hero0, hero1, . . . , hero112

[a] https://archive.ics.uci.edu/ml/datasets/heart+Disease
[b] https://archive.ics.uci.edu/ml/datasets/adult
[c] https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with++Multiple+Types+of+Sound+Recordings
[d] https://archive.ics.uci.edu/ml/datasets/Dota2+Games+Results