Online Material

We provide the proofs for the propositions and theorem stated in the paper, as well as the additional numerical results.

A Proofs for the main results.

A.1 Proof for Proposition 1

Proof. The proofs for the lower bound often starts by converting the problem to a hypothesis testing task. Denote our parameter space by $\mathcal{B}(k) = \{\beta \in \mathbb{R}^d : \|\beta\|_0 \leq k\}$. The intuition is that suppose the data is generated by: (1). drawing β according to an uniform distribution on the parameter space; (2). conditioned on the particular β, the observed data is drawn. Then the problem is converted to determining according to the data if we can recover the underlying β as a canonical hypothesis testing problem.

For any δ-packing $\{\beta_1, \ldots, \beta_M\}$ of $\mathcal{B}(k)$, suppose B is sampled uniformly from the δ-packing, then following a standard argument of the Fano method [5], it holds that:

$$P\big(\min_{\hat{\beta}} \sup_{\|\beta^*\|_0 \leq k})\|\hat{\beta} - \beta^*\|_2 \geq \delta/2\big) \geq \min_{\tilde{\beta}} P\big(\tilde{\beta} \neq B\big), \tag{A.1}$$

where $\tilde{\beta}$ is a testing function that decides according to the data if the some estimated β equals to an element sampled from the δ-packing. The next step is to bound $\min_{\tilde{\beta}} P\big(\tilde{\beta} \neq B\big)$, whereas by the information-theoretical lower bound (Fano's Lemma), we have:

$$\min_{\tilde{\beta}} P\big(\tilde{\beta} \neq B\big) \geq 1 - \frac{I(y, B) + \log 2}{\log M}, \tag{A.2}$$

where $I(\cdot, \cdot)$ denotes the mutual information. Then we only need to bound the mutual information term. Let P_β be the distribution of $\mathbf{y}$ (which the vector consisting of the n samples) given $B = \beta$. Since $\mathbf{y}$ is distributed according to the mixture of: $\frac{1}{M}\sum_i P_{\beta_i}$, it holds:

$$I(y, B) = \frac{1}{M}\sum_i D_{KL}\big(P_{\beta_i} \| \frac{1}{M}\sum_j P_{\beta_j}\big) \leq \frac{1}{M^2}\sum_{i,j} D_{KL}\big(P_{\beta_i} \| P_{\beta_j}\big),$$

where D_{KL} is the Kullback-Leibler divergence. The next step is to determine M: the size of the $\delta-$packing, and the upper bound on $D_{KL}\big(P_{\beta_i} \| P_{\beta_j}\big)$ where P_{β_i}, P_{β_j} are elements of the $\delta-$packing.

For the first part, it has been shown that there exists a $1/2$-packing of $\mathcal{B}(k)$ in ℓ_2-norm with $\log M \geq \frac{k}{2} \log \frac{d-k}{k/2}$ [4]. As for the bound on the KL-divergence term, note that given β, P_β is a product distribution of the condition Gaussian: $y|\epsilon \sim N\big(\beta^\intercal \epsilon \frac{\sigma_z^2}{\sigma_\phi^2}, \beta^\intercal \beta(\sigma_z^2 - \sigma_z^4/\sigma_\phi^2)\big)$, where $\sigma_\phi^2 := \sigma_z^2 + \sigma_\epsilon^2$.

Henceforth, for any $\beta_1, \beta_2 \in \mathcal{B}(k)$, it is easy to compute that:

$$\begin{aligned}
&D_{KL}(P_{\beta_1} \| P_{\beta_2}) \\
&= \mathbb{E}_{P_{\beta_1}}\Big[\frac{n}{2}\log\Big(\frac{\beta_1^\intercal \beta_1(\sigma_z^2 - \sigma_z^4/\sigma_\phi^2)}{\beta_2^\intercal \beta_2(\sigma_z^2 - \sigma_z^4/\sigma_\phi^2)}\Big) + \frac{\big\|\mathbf{y} - \beta_2^\intercal \boldsymbol{\epsilon}\frac{\sigma_z^2}{\sigma_\phi^2}\big\|_2^2}{2\beta_2^\intercal \beta_2(\sigma_z^2 - \sigma_z^4/\sigma_\phi^2)} - \frac{\big\|\mathbf{y} - \beta_1^\intercal \boldsymbol{\epsilon}\frac{\sigma_z^2}{\sigma_\phi^2}\big\|_2^2}{2\beta_1^\intercal \beta_1(\sigma_z^2 - \sigma_z^4/\sigma_\phi^2)}\Big] \\
&= \frac{\sigma_z^2}{2\sigma_\epsilon^2}\|\boldsymbol{\epsilon}(\beta_1 - \beta_2)\|_2^2,
\end{aligned}$$

where $\mathbf{y}$ and $\boldsymbol{\epsilon}$ are the vector and matrix consists of the n samples, i.e. $\mathbf{y} \in \mathbb{R}^n$ and $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times d}$, Since each row in the matrix $\boldsymbol{\epsilon}$ is drawn from $N(0, \sigma_\epsilon^2 I_{d\times d})$, standard concentration result shows that with high probability, $\|\boldsymbol{\epsilon}(\beta_1 - \beta_2)\|_2^2$ can be bounded by $C\|\beta_1 - \beta_2\|_2^2$ for some constant C. It gives us the final upper bound on the KL divergence term:

$$D_{KL}(P_{\beta_1} \| P_{\beta_2}) \lesssim \frac{n\sigma_z^2 \delta^2}{2\sigma_\epsilon^2}.$$

Substitute this result into (A.2) and (A.1), by choosing $\delta^2 = \frac{Ck\sigma_\epsilon^2}{\sigma_z^2 n} \log \frac{d-k}{k/2}$ and rearranging terms, we obtain the desired result that with probability at least $1/2$:

$$\inf_{\hat{\beta}} \sup_{\beta^*: \|\beta^*\|_0 \le k} \|\hat{\beta} - \beta^*\|_2 \gtrsim \frac{\sigma_\epsilon^2}{\sigma_z^2} \frac{d^* \log(d/d^*)}{n}.$$

□

A.2 Proof for Theorem 1

We first define the Rademacher and Gaussian complexity terms for the representation class Φ. We deliberately use the different complexity notions to differentiate the CL-based and same-structure pre-training. In particular, for CL-based pre-training with n triplets of (x_i, x_i^+, x_i^-), the empirical Rademacher complexity of Φ is given by:

$$\mathcal{R}_n(\Phi) = \mathbb{E}_{\vec{\sigma} \in \mathbb{R}^{3d}} \sup_{\phi \in \Phi} \sum_{i=1}^n \big\langle \vec{\sigma}, \big[\phi(x_i), \phi(x_i^+), \phi(x_i^-)\big] \big\rangle,$$

where $\vec{\sigma}$ is the vector of i.i.d Rademacher random variables. For the same-structure pre-training with n samples of (x_i, y_i), the empirical Gaussian complexity of Φ is given by:

$$\mathcal{G}_n(\Phi) = \mathbb{E}_{\vec{\gamma} \in \mathbb{R}^d} \sup_{\phi \in \Phi} \sum_{i=1}^n \big\langle \vec{\gamma}, \phi(x_i) \big\rangle,$$

where $\vec{\gamma}$ is the vector of i.i.d Gaussian random variables. Without loss of generality, we assume the loss functions for both CL-based and same-structure pre-training are bounded and L-Lipschitz. We first prove the result for the same-structure pre-training.

Proof. First recall from Section 4 that the downstream classifier is optimized on n sample drawn from P_τ by plugging in $\hat{\phi}$, which we denote by: $f_{\hat{\phi},\, P_{\tau,n}}$. Also, we have defined:

$$R^*_{\text{task}} = \min_{\phi \in \Phi} \mathbb{E}_{P_\tau \sim \mathcal{E}} \Big[\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim P_\tau} \ell\big(f \circ \phi(x), y\big) \Big],$$

with ϕ^* as the optimum, as well as:

$$R^*_{\text{task}}(\hat{\phi}) = \mathbb{E}_{P_\tau \sim \mathcal{E}} \mathbb{E}_{(x,y) \sim P_\tau} \ell\big(f_{\hat{\phi},\, P_{\tau,n}}(x), y\big). \tag{A.3}$$

Therefore, it holds that:

$$\begin{aligned} R_{\text{task}}(\hat{\phi}) - R^*_{\text{task}} = {} & R_{\text{task}}(\hat{\phi}) - \frac{1}{n}\sum_i \ell\big(f_{\hat{\phi}, P_{\tau,n}}(x_i), y_i\big) \\ & + \frac{1}{n}\sum_i \ell\big(f_{\hat{\phi}, P_{\tau,n}}(x_i), y_i\big) - \frac{1}{n}\sum_i \ell\big(f_{\phi^*, P_{\tau,n}}(x_i), y_i\big) \\ & + \frac{1}{n}\sum_i \ell\big(f_{\phi^*, P_{\tau,n}}(x_i), y_i\big) - \mathbb{E}_{(x_i,y_i) \sim P_{\tau,n}} \Big[\frac{1}{n}\sum_i \ell\big(f_{\phi^*, P_{\tau,n}}(x_i), y_i\big)\Big] \\ & + \mathbb{E}_{(x_i,y_i) \sim P_{\tau,n}} \Big[\frac{1}{n}\sum_i \ell\big(f_{\phi^*, P_{\tau,n}}(x_i), y_i\big)\Big] - \min_{f \in \mathcal{F}} \mathbb{E}_{(X,Y) \sim P_\tau} \ell\big(f \circ \phi(X), Y\big). \end{aligned} \tag{A.4}$$

We define: $f^* = \arg\min_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim P_\tau} \ell\big(f\circ\phi(x), y\big)$. Firstly, note that by the definition of $\hat{\phi}$, we have for the second line on RHS of (A.4) that:

$$\frac{1}{n}\sum_i \ell\big(f_{\hat{\phi},P_{\tau,n}}(x_i), y_i\big) - \frac{1}{n}\sum_i \ell\big(f_{\phi^*,P_{\tau,n}}(x_i), y_i\big) \leq 0.$$

In the next step, notice for the last line on RHS of (A.4) that:

$$\begin{aligned}
\mathbb{E}_{(x_i,y_i)\frac{1}{n}\sum_i\sim P_{\tau,n}}\Big[\frac{1}{n}\sum_i \ell\big(f_{\phi^*,P_{\tau,n}}(x_i), y_i\big)\Big] &= \mathbb{E}_{(x_i,y_i)\sim P_{\tau,n}} \min_{f\in\mathcal{F}} \frac{1}{n}\sum_i \ell\big(f\circ\phi^*(x_i), y_i\big) \\
&\leq \mathbb{E}_{(x_i,y_i)\frac{1}{n}\sum_i\sim P_{\tau,n}}\Big[\frac{1}{n}\sum_i \ell\big(f^*\circ\phi^*(x_i), y_i\big)\Big] \\
&\leq \min_{f\in\mathcal{F}} \mathbb{E}_{(X,Y)\sim P_\tau} \ell\big(f\circ h^*(X), Y\big).
\end{aligned} \tag{A.5}$$

Henceforth, the last line is also non-positive. As for the third line on RHS of (A.4), notice that is involves a bounded random variable $\frac{1}{n}\sum_i \ell\big(f_{\phi^*,P_{\tau,n}}(x_i), y_i\big)$ (since we assume the loss function is bounded) and its expectation. Using the regular Hoeffding bound, it holds with probability at least $1-\delta$ that:

$$\frac{1}{n}\sum_i \ell\big(f_{\phi^*,P_{\tau,n}}(x_i), y_i\big) - \mathbb{E}_{(x_i,y_i)\sim P_{\tau,n}}\Big[\frac{1}{n}\sum_i \ell\big(f_{\phi^*,P_{\tau,n}}(x_i), y_i\big)\Big] \lesssim \sqrt{\log(8/\delta)}.$$

Therefore, what remains is to bound the first line on RHS of (A.4), which can follows:

$$\begin{aligned}
&R_{\text{task}}(\hat{\phi}) - \frac{1}{n}\sum_i \ell\big(f_{\hat{\phi},P_{\tau,n}}(x_i), y_i\big) \leq \sup_{\phi\in\Phi}\Big\{R_{\text{task}}(\hat{\phi}) - \frac{1}{n}\sum_i \ell\big(f_{\phi,P_{\tau,n}}(x_i), y_i\big)\Big\} \\
&\leq \sup_{\phi} \mathbb{E}_{P_\tau\sim\mathcal{E}}\mathbb{E}_{(x_i,y_i)\sim P_{\tau,n}}\Big[\mathbb{E}_{(X,Y)\sim P_\tau}\ell\big(f\circ\phi(X), Y\big) - \frac{1}{n}\sum_i \ell\big(f_{\phi,P_{\tau,n}}(x_i), y_i\big)\Big] \\
&+ \sup_{\phi\in\Phi}\Big[\frac{1}{n}\sum_i \ell\big(f_{\phi,P_{\tau,n}}(x_i), y_i\big) - \mathbb{E}_{(x_i,y_i)\sim P_{\tau,n}}\frac{1}{n}\sum_i \ell\big(f_{\phi,P_{\tau,n}}(x_i), y_i\big)\Big].
\end{aligned} \tag{A.6}$$

Finally, the existing results of bounding empirical processes from Theorem 14 of [3] shows that with probability at least $1-\delta$, the third line above is bounded by:

$$\frac{\sqrt{2\pi}L\mathcal{G}_n(\Phi)}{\sqrt{n}} + \sqrt{9\log(2/\delta)},$$

and the second line is bounded by:

$$\frac{\sqrt{2\pi}}{n} Q \sup_{\phi\in\Phi} \mathbb{E}_{(X,Y)\sim P_\tau}\|\phi(X)\|_2^2,$$

where $Q := \tilde{\mathcal{G}}(\mathcal{F})$ is some complexity measure of the function class $\mathcal{F}$. By combining the above results, rearranging terms and simplifying the expressions, we obtain the desired result. ☐

In what follows, we provide the proof for the CL-based pre-training.

Proof. Recall that the risk of a downstream classifier f is given by: $R_\tau(f;\phi) := \mathbb{E}_{(X,Y)\sim P_\tau}\ell(f\circ\phi(x), y)$, where we let $\ell(\cdot)$ be the widely used logistic loss. When f is a linear model, it induces the loss as: $\ell\big(\theta_1^\intercal\phi(x) - \theta_2^\intercal\phi(x)\big)$, where θ_1, θ_2 corresponds to the two classes $y=0$ and $y=1$. We define a particular linear classifier whose class-specific parameters are given by: $\bar{\phi}^{(y)} := \mathbb{E}_{x\sim P_X^{(y)}}\phi(x)$, for $y\in\{0,1\}$. They correspond to using the average item embedding from the same class as the parameter vector. Therefore, we have:

$$R_\tau(\bar{\phi};\phi) := \mathbb{E}_{(X,Y)\sim P_\tau}\ell\big((\bar{\phi}^{(Y)})^\intercal\phi(x) - (\bar{\phi}^{(1-Y)})^\intercal\phi(x)\big).$$

The importance of studying this particular downstream classifier is because, as long as $\mathcal{F}$ includes linear model, it holds that: $\min_{f\in\mathcal{F}} R_\tau(f;\phi) \leq R_\tau(\bar{\phi};\phi)$. Further more, we will be able to derive meaningful results (upper bound) the risk associated with $\bar{\phi}$ with the CL-based pre-training risk. We first define the probability that two randomly drawn instances fall into the same class: $q := P_Y(y=1)^2 + P_Y(y=0)^2$. In particular, we observe that:

$$\begin{aligned}
R_{\text{CL}}(\phi) &= \mathbb{E}_{x,x^+\sim P_{\text{pos}},x^-\sim P_{\text{neg}}}\big[\ell\big(\phi(x)^\intercal(\phi(x^+)-\phi(x^-))\big)\big] \\
&= \mathbb{E}_{y^+,y^-\sim P_Y^2, x\sim P_X^{(y^+)}}\mathbb{E}_{x^+\sim P_X^{(y^+)},x^-\sim P_X^{(y^-)}}\big[\ell\big(\phi(x)^\intercal(\phi(x^+)-\phi(x^-))\big)\big] \\
&\geq \mathbb{E}_{y^+,y^-\sim P_Y^2, x\sim P_X^{(y^+)}}\Big[\ell\Big(\phi(x)\big(\bar{\phi}^{(y^+)}-\bar{\phi}^{(1-y^+)}\big)\Big)\Big] \text{ Jensen's inequality} \qquad (A.7)\\
&= (1-q)\mathbb{E}_{y^+,y^-\sim P_Y^2, x\sim P_X^{(y^+)}}\Big[\ell\Big(\phi(x)\big(\bar{\phi}^{(y^+)}-\bar{\phi}^{(1-y^+)}\big)\Big)\Big| y^+\neq y^-\Big] + q \\
&= (1-q)R_\tau(\bar{\phi};\phi) + q.
\end{aligned}$$

Therefore, we conclude the relation between the $\bar{\phi}$-induced classifier and the CL-based pre-training risk:

$$R_\tau(\bar{\phi};\phi) \leq \frac{1}{1-q}\big(R_{\text{CL}}(\phi) - q\big), \text{for any } \phi\in\Phi.$$

The next step is to study the generalization bound regarding $R_{\text{CL}}(\phi), \forall\phi\in\Phi$. Suppose the loss function $\ell(\cdot)$ is bounded by B, and is L-Lipschitz. Both assumptions holds for the logistic loss that we study. We define the CL-specific loss function class on top of $\phi\in\Phi$:

$$\mathcal{H}_\Phi := \Big\{\frac{1}{B}\ell\big(\phi(x)^\intercal\big(\phi(x^+)-\phi(x^-)\big)\big)\big|\phi\in\Phi\Big\},$$

such that for $h_\phi\in\mathcal{H}_\Phi$ we have: $h_\phi(x,x^+,x^-) = \frac{1}{B}\ell\circ\tilde{\phi}(x,x^+,x^-)$, where $\tilde{\phi}$ is the mapping of: $\phi(x),\phi(x^+),\phi(x^-)\mapsto\phi(x)^\intercal\big(\phi(x^+)-\phi(x^-)$. The classical generalization result [2] shows that with probability at least $1-\delta$:

$$\mathbb{E}h_\phi \leq \frac{1}{n}\sum_{i=1}^n h_\phi(x_i,x_i^+,x_i^-) + \frac{2\mathcal{R}_n(\mathcal{H}_\Phi)}{n} + 3\sqrt{\frac{\log(4/\delta)}{n}}. \qquad (A.8)$$

In what follows, we connect the complexity of $\mathcal{R}_n(\mathcal{H}_\Phi)$ to the desired $\mathcal{R}_n(\Phi)$. Note that the Jacobian associated with the mapping of $\tilde{\phi}$ is given by:

$$J := \big[\phi(x^+)-\phi(x^-), \phi(x), -\phi(x)\big],$$

so it holds that $\|J\|_2 \leq \|J\|_F \leq 3\sqrt{2}R$, where R is the uniform bound on $\phi\in\Phi$. Hence, $\ell\circ\phi$ is $(3\sqrt{2}LR/B)$-Lipschitz on the domain of $\big(\phi(x),\phi(x^+),\phi(x^-)\big)$. In what follows, using the Telegrand contraction inequality for Rademacher complexity, we reach: $\mathcal{R}_n(\mathcal{H}_\Phi)\leq 3\sqrt{2}LR/B\mathcal{R}_n(\Phi)$. Combining the above results, we see that for any $\phi\in\Phi$, it holds with probability at least $1-\delta$ that:

$$R_{\text{CL}}(\phi) \leq \frac{1}{n}\sum_{i=1}^n \ell\big(\phi(x_i)^\intercal\big(\phi(x_i^+)-\phi(x_i^-)\big)\big) + \mathcal{O}\Big(R\mathcal{R}(\Phi) + \sqrt{\frac{\log(4/\delta)}{n}}\Big).$$

Finally, since we have the decomposition: $R_{\text{CL}}(\phi) = R_{\text{CL}}^G(\phi) + R_{\text{CL}}^B(\phi)$, it remains to bound: $R_{\text{CL}}^B(\phi) = \mathbb{E}_y\mathbb{E}_{x,x+,x^-\sim P_X^y}\Big[\ell\big(\phi(x)^\intercal\big(\phi(x^+)-\phi(x^-)\big)\Big]$.

Let $z_i := \phi(x_i)^\intercal\big(\phi(x_i^+)-\phi(x_i^-)\big)$ and $z = \max z_i$. It is straightforward to show for logistic loss that: $R_{\text{CL}}^B(\phi)\leq\mathbb{E}|z|$. Further more, we have:

$$\begin{aligned}
E|z| \leq \mathbb{E}\big[\max_i |z_i|\big] &\leq n\mathbb{E}[|z_1|] \\
&\leq n\mathbb{E}_x\Big[\|\phi(x)\|\sqrt{\mathbb{E}_{x^+,x^-}\Big(\phi(x)/\|\phi(x)\|\big(\phi(x^+)-\phi(x^-)\big)\Big)^2}\Big] \qquad (A.9)\\
&\lesssim R\mathbb{E}_y\big\|\text{cov}_{P_X^{(y)}}\phi\big\|_2.
\end{aligned}$$

Henceforth, $R_{\text{CL}}(\phi) \lesssim R^G_{\text{CL}}(\phi) + R\mathbb{E}_y \big\|\text{cov}_{P_X^{(y)}}\phi\big\|_2$. Recall that $R^*_{\text{task}} = \min_\phi R^G_{CL}(\phi)$ and for all $\phi \in \Phi$, we have $R_{\text{task}}(\phi) \leq \min_{f\in\mathcal{F}} R_\tau(f;\phi) \leq R_\tau(\hat{\phi};\phi)$. Hence, by rearranging terms and discarding constant factors, we reach the final result:

$$R_{\text{task}}(\hat{\phi}) - R^*_{\text{task}} \lesssim \frac{\mathcal{G}_n(\Phi)}{\sqrt{n}} + \frac{R\tilde{\mathcal{G}}(\mathcal{F})}{n} + \sqrt{\log(8/\delta)},$$

$\square$

A.3 Proof for Proposition 2

Proof. Recall that the kernel-based classifier is given by:

$$f_\phi(x) = \frac{E_{x'}\big[y'k_\phi(x,x')\big]}{\sqrt{\mathbb{E}[k_\phi^2]}},$$

where $y \in \{-1,+1\}$ and R^{OOD} is the out-of-distribution risk associated with a $0-1$ classification risk. We first define for $x \in \mathcal{X}$:

$$\gamma_\phi(x) := \sqrt{\frac{\mathbb{E}_{x'}\big[K_\phi(x,x')\big]}{\mathbb{E}_{x,x'}\big[K_\phi(x,x')\big]}},$$

where the expectation is taken wrt. the underlying distribution. Using the Markov inequality, we immediately have: $|\gamma(x)| \leq \frac{1}{\sqrt{\delta}}$ with probability at least $1-\delta$. It then holds that:

$$\begin{aligned} 1 - R^{\text{OOD}}(f_\phi) &= P\big(yf_\phi(x) \geq 0\big) \\ &\geq \mathbb{E}\Big[\frac{yf_\phi(x)}{\gamma(x)} \cdot 1[yf_\phi(x)\geq 0]\Big] \\ &\geq \mathbb{E}\Big[\frac{yf_\phi(x)}{\gamma(x)}\Big] \geq \frac{\mathbb{E}\big[K_Y(y,y')K_\phi(x,x')\big]}{\sqrt{\mathbb{E}K_\phi^2}}\sqrt{\delta} \text{ ,with probability } 1-\delta, \end{aligned}$$

where $K_Y(y,y') = 1[y=y']$. It concludes the proof. $\square$

A.4 Proof for Proposition 3

Proof. Recall that the sequential interaction model is given by:

$$p\big(x_{k+1}\,|\,s\big) = \lambda p_0(x_{k+1}) + (1-\lambda)\frac{\exp\big(\langle\phi(x_{k+1}),\varphi(s)\rangle\big)}{Z_s},\ \lambda \in (0,1), \tag{A.10}$$

so the likelihood of the sequence $\{x_1,\ldots,x_{k+1}\}$ is given by:

$$\prod_{i=1}^{k+1}\Big(\lambda p_0(x_i) + (1-\lambda)\frac{\exp\big(\langle\phi(x_i),\varphi(s)\rangle\big)}{Z_s}\Big).$$

As a result, the log-likelihood of the sequence embedding $\varphi(s)$, for a particular x_i is given by:

$$l_i\big(\varphi(s)\big) = \log\Big(\lambda p_0(x_i) + (1-\lambda)\frac{\exp\big(\langle\phi(x_i),\varphi(s)\rangle\big)}{Z_s}\Big),$$

and by Taylor approximation, we immediately have:

$$f_i(\varphi(s)) = \frac{1-\lambda}{\lambda Z_s p_0(x_i) + (1-\lambda)}\big\langle\phi(x_i),\varphi(s)\big\rangle + f_i(\mathbf{0}) + \text{residual}. \tag{A.11}$$

Note that: $\arg\max_{v:\|v\|_2=1}\langle v,\phi(x_i)\rangle = \phi(x_i)/\|\phi(x_i)\|_2$, so putting aside the residual terms, the approximate optimal achieved is given by:

$$\arg\max_{\varphi(s)}\sum_{i=1}^{k+1}\Big(\frac{1-\lambda}{\lambda Z_s p_0(x_i) + (1-\lambda)}\big\langle\phi(x_i),\varphi(s)\big\rangle\Big) \propto \sum_{i=1}^{k}\frac{\alpha}{p_0(x_i)+\alpha}\phi(x_i),$$

where $\alpha = (1-\lambda)/(\lambda Z_s)$. This concludes the proof. $\square$

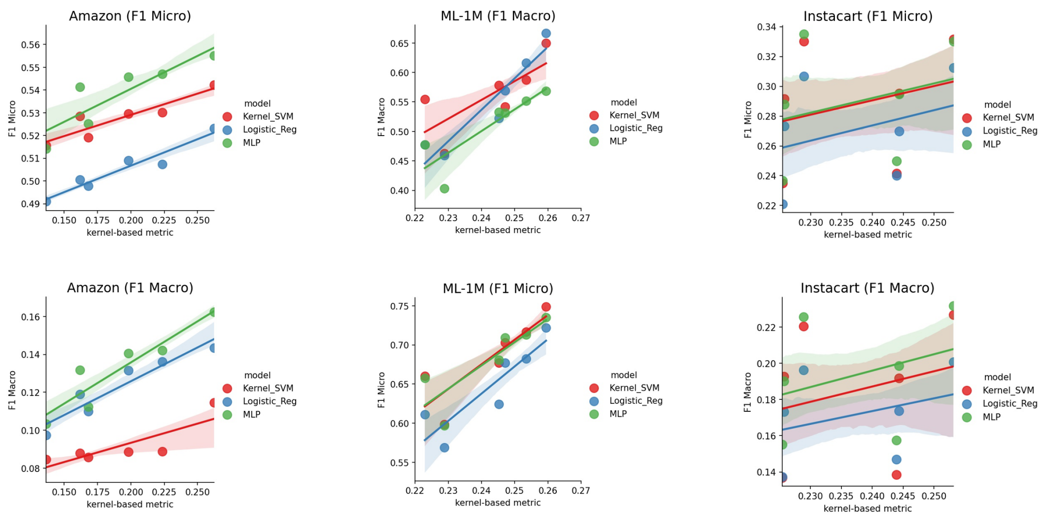

Figure A.1: Classification results for the relation between the kernel-based evaluation score and the classification metrics for Kernel SVM, logistic regression and two-layer MLP.

B Addition numerical results

We provide in the part of the appendix the additional numerical results for examining the effectiveness of the kernel-based evaluation for entity tasks. As we mentioned in the paper, we set the window size and #negative samples (for the Item2vec-type pre-training model [1]) to different values, to obtain item embeddings with different qualities. We fix the embedding dimensions as $d = 32$.

The kernel-based evaluation metric is computed via:

$$\frac{\sum_{(x',y')} \big[K_Y(y,y')K_\phi(x,x')\big]}{\sqrt{\sum_{x,x'} K_\phi(x,x')^2}},$$

where the target kernel is simply given by: $K_Y(y,y') = 1[y = y']$. Since we deal with multi-class classification, we use *micro-F1* score as well as *macro-F1* score to evaluate the models' predictions on the testing data. Since our purpose is to examine the performance of the kernel-based metric, we do not particularly tune the classification models. Instead, we directly train them for 80% of the items, and evaluate them on the remaining 20%. The outcomes we plot in Figure A.1, where we observe a significant positive correlation between the kernel-based evaluation metric and the classification outcomes, both in terms of micro-F1 and macro-F1 scores. They suggest the effectiveness of the kernel-based evaluation metric for the OOD performance on entity tasks. We do notice a very high variance on the Instacart dataset. We suspect that it is because there is no ordering information (only shopping baskets), so the Item2vec algorithm is unable to leverage the inductive bias that items show up closer in the sequence tend to be contextually similar. Even so, our appraoch is still able to produce positive signal for the quality of the pre-trained embeddings.

References

[1] O. Barkan and N. Koenigstein. Item2vec: neural item embedding for collaborative filtering. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2016.

[2] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[3] A. Maurer, M. Pontil, and B. Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

[4] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ell-q balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

[5] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.