A Constrained Bayesian Approach to Out-of-Distribution Prediction
(Supplementary Material)

Ziyu Wang1  Binjie Yuan*1  Jiaxun Lu2  Bowen Ding3  Yunfeng Shao2  Qibin Wu3  Jun Zhu*1

1Dept. of Comp. Sci. & Tech., BNRist Center, Tsinghua-Huawei Joint Center for AI, THBI Lab, Tsinghua University
2Huawei Noah’s Ark Lab
3Huawei Technologies Co., Ltd.

Additional notations and conventions The following conventions are used throughout appendix: the denotation of constants (c1, c2, . . . , C1, C2, . . . ) may change from line to line. || · ||2 denotes the Euclidean norm for vectors, or the L2(Pm) norm for functions of x. ⟨ · , · ⟩2 denotes the respective inner product. || · ||F denotes the Frobenius norm. φx, Φx denote the PDF and CDF of the standard Gaussian distribution. Recall the inequality 1 − Φx(x) ≤ e−x2/2 for x > 0.

S1 PROOF FOR LEMMA [1] AND ADDITIONAL REMARKS

Proof for the lemma The first two claims in the lemma are implied by the following two lemmas:

Lemma S1. When α > 0, m ≪ max{1, σx2 d spu} the classifier parameterized by θ = (0, θs) = (0, αE tr bspu) satisfies

\[ R_{ε,01}(f_θ) ≤ c_1 \exp(-c_2σx2 d spu/m + o_p(1)) \rightarrow 0, \quad \forall \epsilon \in E tr. \]  \hspace{1cm} (S1)

Moreover, the same bound holds for the logistic loss, if in addition α ≤ O(poly(d spu/m)), ατ2 2 > 1/σx2 m.

Lemma S2. Denote by x1 ∈ Rd inv, xspu ∈ Rd spu the invariant and spurious components of the input x. For any e1, e2 ∈ E tr, let

\[ KL_{ij} := KL(p_{marg,e1} || p_{marg,e2}) = KL(p_{c1}(y, θs ⊤ x, x_i) || p_{c2}(y, θs ⊤ x, x_i)) \]

denote the KL divergence between the marginal distributions. When α > 0, m < d spu/16, with probability ≥ 1 − e−m/18 there exists an ℓ ∈ [m], determined by {βe spu : e ∈ E tr} s.t.

\[ KL\left( \bigotimes_{e ∈ E tr} p_{marg,e} \left\| p_{marg,e}^{⊗m} \right\| \right) = \sum_{j=1}^{m} KL_{ij}\ell ≤ \frac{256m}{σx2 d spu}, \]  \hspace{1cm} (S2)

The last claim in the text, Eq. [2], follows by a standard argument following [S2]: consider the scenario ne ≡ n for simplicity. Denote by D∗ := ((yκ, θs κ, xκ, xκ spu) : j ∈ [m] : k ∈ [n]) the transposed dataset which contains the same information as Dtr. Any test on Dtr for the null hypothesis “θs x is an invariant feature” always provides a two-sample test for

\[ H'_0 : D∗ ∼ \left( p_{marg,e}^{⊗m} \right)^{⊗n} =: P_{0,n}, \quad H'_1 : D∗ ∼ \left( \bigotimes_{e ∈ E tr} p_{marg,e} \right)^{⊗n} =: P_{1,n}, \]

with the same size and power in the two scenarios. However, any such test must have its combined error lower bounded by

\[ 1 − D_{TV}(P_{0,n}, P_{1,n}) ≥ 1 − \sqrt{2KL(P_{1,n} || P_{0,n})} ≥ 1 − n \cdot \frac{256m}{σx2 d spu}. \]

This completes the proof. □

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).
Remark 1 (IRM and GDRO). From lemma [S1] it is clear that \( \tilde{\theta} \) constitutes an approximate optima for ERM and GRDO, as well as the variance-penalty based approaches such as [Krueger et al. 2021]. It also shows that the predictor \( f = w \circ \Phi \), where \( \Phi = \text{id} \) and \( w(h) = \tilde{\theta}^\top h \), approximately satisfies the hard constraints in [IRM] and constitutes an approximate optima in this sense \([\text{IRMO}]\). 

Remark 2 (interpretations of the KL bound, additional error from feature learning). In addition to the testing-based interpretation as in Eq. [S2], (S2) can also be interpreted directly if we restrict to the family of tests based on a conditional KL divergence, since we have, for all \( j \in [m] \),

\[
\text{KL}(p_{e_j}(y \mid \tilde{\theta}^\top x_s) \parallel p_{e_j}(y \mid \hat{\theta}^\top x_s)) \leq \text{KL}(p_{e_j}(y \mid x_i, \tilde{\theta}^\top x_s) \parallel p_{e_j}(y \mid x_i, \hat{\theta}^\top x_s)) \leq \text{KL}_{d\ell} \leq \frac{256m}{\sigma^2_s d_{spu}}.
\]

Such tests can be viewed as restricting to the validation of the defition of invariant features, which concerns such conditionals; they should reject \( H_0 \) if the estimated KL divergence becomes larger than a threshold \( \delta_n = o_n(1) \), which at least needs to cover the estimation error for the KL divergence. It is thus clear that in the feature learning scenario we must use a threshold \( \delta'_{n} \gg \delta_n \), since even for the truly invariant part of the model, we can only learn approximately invariant features which will inevitably violate the KL bound by an extra margin. This makes for a larger threshold than (2): for example, if the feature learning process is such that \( \delta'_{n} \gtrsim d_{spu}/n \), we would have the indistinguishability result as long as \( d_{spu}/n \gg m/\sigma^2_s d_{spu} \), i.e.,

\[
\max_{e \in \mathcal{E}_r} n_{e} \ll \frac{\sigma^2_s d^2_{spu}}{m}.
\]

Another issue that exists for any possible test is that in the feature learning scenario, we need to apply to a collection of feature extractors; we thus needs more stringent requirements on the power of the test, rather than merely requiring them to be \( 1 - o(1) \). For \(|M|\) feature extractors with independent failure probabilities, we would require \( n \gtrsim |M| \sigma^2_s d_{spu}/m \) for reliable learning of invariant features. Note how these regimes allow for successful fitting of an in-distribution predictor.

### S1.1 PROOF FOR AUXILIARY LEMMAS

We first introduce the following notations: \( \mathcal{E}_{tr} := \{e_1, \ldots, e_m\} \), \( \mathbf{1} := \{1, \ldots, 1\} \in \mathbb{R}^m \), and define the \( m \times m \) matrix \((\Sigma_S)_{ij} = (\tilde{\beta}^e_{spu})^\top (\tilde{\beta}^e_{spu})\), so that \( \|\hat{\theta}\|^2 = \alpha^2 \mathbf{1}^\top \Sigma_S \mathbf{1} \). Define \( \Sigma_S := E(\tilde{\beta}^e_{spu})^2 = \sigma^2_s d_{spu} \mathbf{1} \). Note that by covariance concentration [Wainwright 2019 Ch. 6], we have, when \( m \leq d_{spu} \),

\[
\mathbb{P}(\{\tilde{\beta}_{spu}\left| 1 \right. : \Sigma_S - \Sigma_S \left. \geq 3 \frac{m}{d_{spu}} + \delta \right) \geq 1 - e^{-d_{spu}\delta^2/18}. \tag{S3}
\]

**Proof for lemma [S7]** We first derive the 0-1 loss for \( \tilde{\theta} \). Note that in the setting of the example, we have, for any \( \theta = (\theta_i, \theta_s) \),

\[
R_{e,01}(f_{\theta}) := \mathbb{E}_{x,c \sim \mathcal{D}} \mathbf{1}\{|\text{sgn}(\theta^\top x^e) = y^e\} = \Phi_{\mathbf{z}} \left( -\frac{\theta^\top \mathbb{E}(x^e \mid y^e = 1)}{\sqrt{\sigma^2_s \text{Cov}_{x^e}(x^e) \theta}^\top} \right) = \Phi_{\mathbf{z}} \left( -\frac{\theta^\top \tilde{\beta}_{spu} + \theta^\top \tilde{\beta}^e_{spu}}{\sqrt{\sigma^2_s \theta_i^2 + \sigma^2_s \theta_s^2}} \right).
\]

Thus when \( m < d_{spu} \), we have, by central limit theorem and (S3),

\[
\|\hat{\theta}\| = \alpha \sqrt{\mathbf{1}^\top \Sigma_S \mathbf{1}} \leq \alpha \sqrt{m(\|\Sigma_S\| + \|\Sigma_S - \Sigma_S\|)} = \alpha \mathbf{1} \sqrt{md_{spu}(1 + O_p((m/d_{spu})^{1/4}))}, \tag{S4}
\]

\[
\tilde{\beta}_{spu}^e = \alpha \left( \|\tilde{\beta}_{spu}\|^2 + \sum_{e' \in \mathcal{E}_{tr}, e' \neq e} (\tilde{\beta}_{spu}^e, \tilde{\beta}_{spu}^e)^2 \right) = \alpha \mathbf{1} \sqrt{d_{spu} + O_p(\sqrt{md_{spu}})} \tag{S5}
\]

Thus, when \( m \ll \min\{1, \sigma^2_s \mathbf{1}^\top \Sigma_S \mathbf{1}\} d_{spu} \), we have

\[
R_{e,01}(f_{\hat{\theta}}) = \Phi_{\mathbf{z}}(-\sigma^2_s^{-1} \mathbf{1} \sqrt{d_{spu}/m + o_p(1)}) \leq \exp(-2\sigma^2_s^{-1} \mathbf{1} \sqrt{d_{spu}/m + o_p(1)}) \overset{p}{\to} 1, \quad \forall e \in \mathcal{E}_{tr}.
\]

\(^1\)In practice, soft constraints based on gradient penalty are used for IRM. But it is easy to adapt our proof to show that the gradient norm \( \|\nabla_{w} R_{e,01}(w \circ \Phi)\|_2 = \|\nabla_{\hat{\theta}} R_{e,01}(f_{\hat{\theta}})\|_2 \) vanishes at a similar exponential rate. Thus, such a \((w, \Phi)\) pair is also an approximate optima for the soft constraint-based formulation.
Now we consider the logistic loss. Fix any $e \in \mathcal{E}_r$, and recall $\tilde{\theta}_r := \tilde{\beta}_r^e/2\sigma^2$ defines the Bayes classifier given the spurious features $x_{spu}^e \in \mathbb{R}^{d_{spu}}$. Introduce the random variables
\[
x \sim \mathcal{N}(\beta_{spu}^e, \sigma_s^2 I), \quad z_1 = \tilde{\theta}_r^T x, \quad z_2 = \tilde{\theta}_r^T x,
\]
so that
\[
R_{e, \text{log}}(\bar{\theta}) = \mathbb{E}_{x_{spu}} \sum_{y \in \{\pm 1\}} p(y^e = y | x_{spu}^e) \log(1 + e^{-y \bar{\theta}^T x_{spu}^e})
\]
\[
= \mathbb{E}_{z_1, z_2} \left( \mathbb{1}_{\{z_1 \geq 1\}} \log(1 + e^{-z_2}) + \mathbb{1}_{\{z_1 < 0\}} \log(1 + e^{z_2}) \right)
\]
\[
\leq \mathbb{E} \log(1 + e^{-z_2}) + \mathbb{E} \left( \mathbb{1}_{\{z_1 < 0\}} \max\{1, 2z_2\} \right)
\]
\[
\leq \mathbb{E} \log(1 + e^{-z_2}) + \mathbb{E} \mathbb{1}_{\{z_1 < 0\}} + \mathbb{E} \mathbb{1}_{\{z_1 < 0, z_2 > 1/2\}} 2z_2.
\]
We will bound the three terms in turn. Before that, we first derive the joint distribution of $(z_1, z_2)$. By \[\text{(S4)}\] - \[\text{(S5)}\], we have
\[
\mu_2 := \mathbb{E} z_2 = \alpha \tau_x^2 d_{spu}(1 + o_p(1)), \quad \sigma_2^2 := \text{Var}(z_2) = \sigma_s^2 \alpha^2 \tau_x^2 m d_{spu}(1 + o_p(1)), \quad \text{Cov}(z_1, z_2) = \frac{1 + o_p(1)}{2\sigma_2^2} \alpha \tau_x^2 d_{spu}.
\]
Moreover, we have
\[
\mu_1 := \mathbb{E} z_1 = \frac{\|\beta_{spu}^e\|^2_2}{2 \sigma_2^2} = \frac{\tau_x^2 (d_{spu} + O_p(\sqrt{d_{spu}}))}{2 \sigma_2^2}, \quad \sigma_1^2 := \text{Var}(z_1) = \frac{1}{2} \mu_1.
\]
We now return to the three terms for $R_{e, \text{log}}$. For (I), consider the decomposition
\[
(I) = \mathbb{E} \log(1 + e^{-z_2}) + \mathbb{E} \log(1 + e^{-z_2}) \mathbb{1}_{\{-1/2 < z_2 < A\}} + \mathbb{E} \log(1 + e^{-z_2}) \mathbb{1}_{\{z_2 > A\}}
\]
\[
\leq \mathbb{E} [-2z_2 \mid z_2 < -1/2] \mathbb{P}(z_2 < -1/2) + \mathbb{1}_{\{-1/2 < z_2 < A\}} \log(1 + e^{-A})
\]
\[
\leq \mathbb{E} [-2z_2 \mid z_2 < -1/2] \mathbb{P}(z_2 < -1/2) + \left(1 + \frac{\sigma_2^2}{\mu_2}\right) e^{-z_2^2/2\sigma_2^2} + e^{-(\mu_2 - A)^2/2\sigma_2^2} + \log(1 + e^{-A})
\]
\[
\leq \mathbb{E} [-2z_2 \mid z_2 < -1/2] \mathbb{P}(z_2 < -1/2) + \left(1 + \frac{\sigma_2^2}{\mu_2}\right) e^{-z_2^2/2\sigma_2^2} + e^{-(\mu_2 - A)^2/2\sigma_2^2} + e^{-A}
\]
\[
\leq \mathbb{E} [-2z_2 \mid z_2 < -1/2] \mathbb{P}(z_2 < -1/2) + \left(1 + \frac{\sigma_2^2}{\mu_2}\right) e^{-z_2^2/2\sigma_2^2} + e^{-(\mu_2 - A)^2/2\sigma_2^2} + e^{-A} \leq e^{-\mu_2^2/3\sigma_2^2}.
\]
In the above, (i) follows by the Gaussian CDF bound and lemma \[\text{(S3)}\] below, (ii) follows since $\mu_2^2 \gg \sigma_2^2 \gg 1$, $\alpha = O(\text{poly}(d_{spu}/m))$, and (iii) sets $A = \mu_2/10$. Now,
\[
(II) \leq e^{-\mu_2^2/2\sigma_2^2},
\]
\[
(III) = \int_{1/2}^\infty 2t \mathbb{P}(z_1 < 0 \mid z_2 = t) \mathbb{P}(z_2 = t) dt \leq \mathbb{P}(z_1 < 0) \mathbb{E}(2z_2 \mid z_2 > 1/2) \leq e^{-\mu_2^2/2\sigma_2^2} (\mu_2 + 2\sigma_2 e^{-\mu_2^2/2\sigma_2^2}),
\]
where the first inequality follows because $z_1$ and $z_2$ are positively correlated, and $\mu_2 > 0$, and the second follows by an application of lemma \[\text{(S3)}\]. Combining, we can see that in the regime $m \ll \sigma_s^2 \tau_x^2 d_{spu}$, there exists $c_1, c_2 > 0$ s.t.
\[
R_{e, \text{log}}(\bar{\theta}) \leq (I) + (II) + (III) \leq c_1 e^{-c_2 \tau_x^2 d_{spu}/m \sigma_s^2}.
\]
This proves \[\text{(S1)}\]. \qed

**Proof for lemma \[\text{(S2)}\]** For any $e_1, e_2 \in \mathcal{E}_r$ and fixed realizations of $\{\tilde{\beta}_r^e\}$ (i.e., the following display implicitly conditions on the two), we have
\[
\text{KL}_{ij} = \text{KL}(p_{e_1}(y, \tilde{\theta}_s^T x_s, x_i) \mid p_{e_2}(y, \tilde{\theta}_s^T x_s, x_i)) = \mathbb{E}_{p_{e_1}} \log \frac{p_{e_1}(y, \tilde{\theta}_s^T x_s, x_i)}{p_{e_2}(y, \tilde{\theta}_s^T x_s, x_i)} = \mathbb{E}_{p_{e_1}} \log \frac{p_{e_1}(\tilde{\theta}_s^T x_s \mid y, x_i)p_{e_2}(y, x_i)}{p_{e_2}(\tilde{\theta}_s^T x_s \mid y, x_i)p_{e_2}(y, x_i)}
\]
\[
= \mathbb{E}_{p_{e_1}} \log \frac{p_{e_1}(\tilde{\theta}_s^T x_s \mid y)}{p_{e_2}(\tilde{\theta}_s^T x_s \mid y)} = \text{KL}(\mathcal{N}(\mu_{e_1}, \sigma_2^2) \mid \mathcal{N}(\mu_{e_2}, \sigma_2^2)).
\]
where \( \mu_{2j} := \langle \beta_{2j}^e, \theta_j \rangle \forall j \in [m] \), and \( \sigma_2^2 := \sigma_2^2||\theta_i||^2 \). Plugging in the expression for KL divergence between Gaussian distributions, we find, for all \( i, j \in [m] \),

\[
\text{KL}_{ij} = \frac{(\mu_{2i} - \mu_{2j})^2}{2\sigma_2^2} = \frac{\alpha^2}{2\sigma_2^2} \left( \sum_{k=1}^{m} (\Sigma_S)_{ik} - (\Sigma_S)_{jk} \right)^2
\]

\[
\leq 4\alpha^2 \left( \sum_{k=1}^{m} (\Sigma_S)_{ik} - (\Sigma_S)_{jk} \right)^2 + 4\alpha^2 \sum_{k=1}^{m} (\Sigma_S)_{ik} - (\Sigma_S)_{jk} \right)^2
\]

\[
= 2\alpha^2 \left( \langle \mathbf{1}^\top (\Sigma_S - \Sigma_S)e_i \rangle^2 + \langle \mathbf{1}^\top (\Sigma_S - \Sigma_S)e_j \rangle^2 \right),
\]

where \( e_i = (\ldots, 0, 1, 0, \ldots) \) denotes the \( i \)-th Euclidean basis. We thus have, by symmetry and the trace formula, if \( m/d_{sp} \) we find, when \( d \geq 64m \), on that event we have

\[
\sum_{i=1}^{m} \sum_{j=1}^{m} \text{KL}_{ij} \leq 4\alpha^2 m^2 \mathbb{E} \left[ \frac{||\Sigma_S - \Sigma_S||^2}{\sigma_2^2} \right] \leq 4\alpha^2 m^2 \cdot 4\tau_s \sqrt{m} \leq \frac{64\alpha^2 \tau_s m^3}{\sigma_2^2 ||\theta||_2^2}
\]

\[
\leq \frac{64\alpha^2 \tau_s m^3}{\sigma_2^2 \alpha^2 \tau_s m d_{sp} / 4} = \frac{256m^2}{\sigma_2^2 d_{sp}},
\]

where we note that \( S_4 \) holds on the same event. Therefore, on that event there always exist some \( \ell \in [m] \) which is determined by \( \{\beta_{2j}^e\} \) s.t.

\[
\sum_{j=1}^{m} \text{KL}_{ij} \leq \frac{256m}{\sigma_2^2 d_{sp}}.
\]

This completes the proof.

**Lemma S3.** Let \( z \sim \mathcal{N}(\mu, \sigma^2) \), \( b < \mu \). Then we have

\[
\mathbb{E}(z \mid z < b) \overset{(i)}{=} \mu - \sigma \frac{\Phi_z((b - \mu)/\sigma)}{\Phi_z((b - \mu)/\sigma)} \geq b - \frac{\sigma}{\mu - b}, \quad \mathbb{E}(z \mid z < b) \overset{(ii)}{=} \mu + \sigma \frac{\phi_z((b - \mu)/\sigma)}{1 - \Phi_z((b - \mu)/\sigma)} \leq \mu + \sigma \phi_z((b - \mu)/\sigma).
\]

**Proof.** (i) is a known property of truncated normal distribution [Greene, 2003]. (ii) follows by the bound \( \Phi_z(-x) \geq \frac{x}{x^2 + 1} \phi_z(-x) \) [Abramowitz et al., 1988].

---

**S2 PROOF FOR PROPOSITION 2**

Throughout the proof we will work with the transformed data and parameters: \( x^c \leftarrow M^{-1} x^c, \theta \leftarrow M^\top \theta \). This allows us to ignore the presence of \( M \), as long as we replace \( U \) with \( U^c := \|M\|U \). We also introduce the following notations:

\[
d := d_{\text{inv}} + d_{sp}, \quad X := \{x_i, \ldots, x_{n_i} \} \in \mathbb{R}^{n_i \times d}, \quad Y := \{y_i, \ldots, y_{n_i} \} \in \mathbb{R}^{n_i \times 1},
\]

\[
\tilde{\theta}^e_{sp} := (\tilde{b}_{inv}, \tilde{\theta}^e_{sp}), \quad S_{sp} := \langle \tilde{b}^e_{sp} : e \in \mathcal{E}_{sp} \rangle.
\]

Note that given the above transformation, \( \tilde{\theta}^e_{sp} \) now parameterizes the Bayes predictor for environment \( e \). And across all environment \( e \), we will also have

\[
R^c(\theta) = \mathbb{E}_{x^c, y^c} (\langle \theta \Sigma x^c - y^c \rangle)^2 = \|\theta - \tilde{\theta}^e_{sp}\|_2^2 + \sigma^2,
\]
and the constraint set, for which we have fixed \( \rho + \varepsilon_n = 0 \), reduces to
\[
C_{tr} = \{ \theta \in \mathbb{R}^d : \| \theta - \bar{\theta}_{spu} \|_2^2 \leq \| \theta_{inv} - \bar{\theta}_{spu} \|_2^2 \ \forall e \in \mathcal{E}_{tr} \}
\]
\[
= \{ \theta = (\beta, \beta_s) \in \mathbb{R}^d : \| \beta - \beta_{inv} \|_2^2 + \| \beta_s - \beta_{spu} \|_2^2 \leq \| \beta_{spu} \|_2^2 \ \forall e \in \mathcal{E}_{tr} \}
\]
\[
\subset \{ \theta = (\beta, \beta_s) \in \mathbb{R}^d : \langle \beta, \beta_{spu} \rangle_2 \geq \frac{1}{2} \| \beta_s \|_2^2 \ \forall e \in \mathcal{E}_{tr} \}.
\]
The constrained parameter space is
\[
\mathcal{F} = \{ f_0 : \| \theta \|_2^2 \leq U', \theta \in C_{tr} \}.
\]
Note that by construction, \( \bar{\theta}_{inv} \in \mathcal{F} \) always holds.
Our main task is to establish improved metric entropy bounds for the following “localized” space:
\[
\partial \mathcal{F}_\delta := \left\{ f_0 - f_\theta : f_\theta \in \mathcal{F}, \| f_0 - f_\theta \|_{n,2}^2 := \frac{1}{n_s} \sum_{i=1}^{n_s} (f_0(x^*_i) - f_\theta(x^*_i))^2 \leq \delta \right\}.
\]
We first note that

**Lemma S4.** When \( n_s \geq 5d \), there exists \( c > 0 \) s.t.
\[
\mathbb{P}_X \left( \forall \theta, \theta' \in \mathbb{R}^d, \frac{1}{2} \| \theta - \theta' \|_2^2 \leq \| f_0 - f_\theta' \|_{n,2}^2 \leq 2 \| \theta - \theta' \|_2^2 \right) \geq 1 - e^{-cn_s}.
\]

(S7)

**Proof.** By the fact that \( \| f_\theta - f_\theta' \|_{n,2} = \| n_s^{-1/2} X(\theta - \theta') \|_2 \), and the concentration of Gaussian covariance matrices [Wainwright 2019, Theorem 6.1]. \( \square \)

**Proposition S5** (entropy bound). On the event (S7) we have, for any \( \zeta, \delta, t > 0 \), with \( S_{tr} \)-probability \( \geq 1 - \zeta \),
\[
\log N(\partial \mathcal{F}_\delta, \| \cdot \|_{n,2}, t) \leq \min \left\{ d_{spu} \log \left( 1 + \frac{c' \delta}{e c n_s t^2} \right), d_{spu} \log \left( 1 + \frac{c' \delta}{2m d_{spu} t} \right) \right\} + d_{inv} \log \left( 1 + \frac{c' \delta}{t} \right) + \log \zeta^{-1}.
\]

(S8)

**Proof.** We condition on the event (S7) throughout the proof. Then, any \( \theta = (\beta, \beta_s) \) s.t. \( f_\theta \in \partial \mathcal{F}_\delta \) must satisfy
\[
\| \beta_i - \bar{\beta}_{inv} \|_2^2 + \| \beta_s \|_2^2 = \| \theta - \bar{\theta}_{inv} \|_2^2 \leq 2 \| f_0 - f_{\bar{\theta}_{inv}} \|_{n,2}^2 \leq 2\delta^2 \Rightarrow \max\{\| \beta_i - \bar{\beta}_{inv} \|_2, \| \beta_s \|_2 \} \leq \sqrt{2}\delta,
\]
and
\[
\beta_s \in C_s := \{ \beta_s : \langle \beta_s, \beta_{spu} \rangle_2 \geq \frac{1}{2} \| \beta_s \|_2^2 \ \forall e \in \mathcal{E}_{tr} \}.
\]
By (S7), we also know that a \( (\| \cdot \|_{n,2}, t) \)-covering for \( \partial \mathcal{F}_\delta \) can be constructed as the Cartesian product of a \( (\| \cdot \|_2, t/3) \)-covering for the ball \( \mathbb{B}_{\mathbb{L},\delta} := \{ \beta_i \in \mathbb{R}^d : \| \beta_i - \bar{\beta}_{inv} \|_2 \leq \sqrt{2}\delta \} \), and a \( (\| \cdot \|_2, t/3) \)-covering for \( C_s \cap \mathbb{B}_{\mathbb{L},\delta} \), where \( \mathbb{B}_{\mathbb{L},\delta} := \{ \beta_s \in \mathbb{R}^d : \| \beta_s \|_2 \leq \sqrt{2}\delta \} \). Thus, we have, for some \( c, c' > 0 \),
\[
\log N(\partial \mathcal{F}_\delta, \| \cdot \|_{n,2}, t) \leq \log N(\mathbb{B}_{\mathbb{L},\delta}, \| \cdot \|_2, ct) + \log N(C_s \cap \mathbb{B}_{\mathbb{L},\delta}, \| \cdot \|_2, ct)
\]
\[
\leq d_{inv} \log (1 + c' \delta t) + \log N(\mathbb{B}_{\mathbb{L},\delta} \cap C_s, \| \cdot \|_2, ct),
\]
(S8)

where the second inequality can be found as [Wainwright 2019, Example 5.8]. It remains to bound \( \log N(\mathbb{B}_{\mathbb{L},\delta} \cap C_s, \| \cdot \|_2, ct) \).
For this purpose, first note that, for any fixed \( \beta \neq 0 \), we have
\[
\mathbb{P}_{S_{tr}}(\beta \in C_s) = \prod_{e \in \mathcal{E}_{tr}} \mathbb{P}_{\beta_{spu} \sim \mathcal{N}(0, d_{spu}^{-1} t)} \left( \| \beta \|_2^2 \right) \geq \frac{1}{2} \| \beta \|_2^2 \leq \min\{e^{-md_{spu} \| \beta \|_2^2/4}, 2^{-m}\}.
\]
(S9)
Introduce
\[ B_{s_1} := \{ \beta \in \mathbb{R}^{d_{spu}} : \| \beta \|_2 \leq \sqrt{2Z}\delta \}, \quad B_{s_2} := B_{s, \delta} \setminus B_{s_1}, \]
and \( C_{s_1}, C_{s_2} \) be \((\| \cdot \|, ct)\)-coverings for \( B_{s_1}, B_{s_2} \) with the optimal cardinality. Then
\[
\mathbb{E}_{\mathcal{E}_r} N(B_{s, \delta} \cap C_s, \| \cdot \|_2, ct) \leq \sum_{\beta \in C_s} \mathbb{P}(\beta \in C_s) + \sum_{\beta \in C_{s_2}} \mathbb{P}(\beta \in C_s) \leq N(B_{s_1}, \| \cdot \|, ct) + N(B_{s, \delta}, \| \cdot \|, ct)e^{-md_{spu}Z^2\delta^2/2} \leq (1 + cZ\delta/t)^d_{spu} + (1 + c'd\delta/t)^d_{spu} e^{-Md_{spu}Z^2\delta^2/2}.
\]
By Markov’s inequality and the monotonicity of \( \log(\cdot) \), we find that for all \( \zeta \in (0, 1) \), with \( S_{tr} \)-probability \( \geq 1 - \zeta \),
\[
\log N(B_{s, \delta} \cap C_s, \| \cdot \|_2, ct) \leq d_{spu} \log(1 + c'Z\delta/t) + d_{spu} \log \left( \frac{1 + c'd\delta/t}{e^{MZ^2\delta^2/2}} \right) + \log \zeta^{-1}.
\]
Plugging back to (S8) proves the first claim. The second claim (involving \( 2^{-m/d_{spu}} \)) can be proved similarly, by using the second case in (S9) (involving \( 2^{-m} \)).

This allows us to prove our main result:

**Proof for Proposition** \(^2\) It suffices to bound the critical radius \( \delta_n^2 \) of the Gaussian complexity of \( \partial F_3 \) [Wainwright 2019, Eq. (13.42a)]; given a high-probability bound \( \delta_n^2 \leq \delta_n^2 \) that holds w. p. \( 1 - \zeta' \), we will have, with probability \( \geq 1 - \zeta' - c_1 e^{-c_2 n\delta_n^2} \),
\[
\| f_\theta - f_{\theta_{inv}} \|_2^2 \leq 2\| f_\theta - f_{\theta_{inv}} \|_2^2 \leq c_0 \| f_{\theta_{inv}} - f_{\theta_{spu}} \|_2^2 + c_1 \delta_n^2,
\]
where the last inequality is [Wainwright 2019, Theorem 13.13].

By [Wainwright 2019, Corollary 13.7], any solution \( \delta \) to the following inequality will bound \( \delta_n^2 \):
\[
\frac{16}{\sqrt{\pi}} \int_{\delta^2}^{\delta} \sqrt{\log N(\partial F_3, \| \cdot \|_{n, 2}, t)} \, dt \leq \delta^2 / 4\delta.
\]
Let us restrict to \( \delta \geq n^{-1/2} \) and bound the LHS. Define \( t_j = 2^{-j\delta} \) for \( j \leq J := \lfloor \log n^{1/2}\delta \rfloor \).
Then
\[
\int_{\delta^2}^{\delta} \sqrt{\log N(\partial F_3, \| \cdot \|_{n, 2}, t)} \, dt \leq \sum_{j=0}^{J} \sqrt{\log N(\partial F_3, \| \cdot \|_{n, 2}, t)} (t_j - t_{j+1}).
\]
For any fixed \( Z > 0, \delta \geq n^{-1/2} \), a union bound over \( J \) applications of proposition[S5 to \((\zeta \leftarrow n^{-10}, \delta, t \leftarrow t_j, Z)\) shows that, with \( S_{tr} \)-probability \( \geq 1 - n^{-10} \log(n^{1/2}\delta) \),
\[
\sum_{j=0}^{J} \sqrt{\log N(\partial F_3, \| \cdot \|_{n, 2}, t)} (t_j - t_{j+1}) < \sqrt{10 \log n\delta} + \sum_{j=0}^{J} (t_j - t_{j+1}) \cdot \left( d_{inv} \log \left( 1 + \frac{c'd}{t} \right) + d_{spu} \log \left( 1 + \frac{c'Z\delta}{t} \right) \right)^{1/2}
\]
\[
\leq \sqrt{10 \log n\delta} + \sum_{j=0}^{J} (t_j - t_{j+1}) \cdot \left( \sqrt{d_{inv} \log \left( 1 + \frac{c'd}{t} \right)} + \sqrt{d_{spu} \log \left( 1 + \frac{c'Z\delta}{t} \right)} \right),
\]
(S12)
and that

\[
\sum_{j=0}^{J} \sqrt{\log N(\partial \mathcal{F}_\delta, \| \cdot \|_{n, 2}, t_j)(t_j - t_{j+1})} < \sqrt{10 \log n \delta} + \sum_{j=0}^{J} (t_j - t_{j+1}) \left( \sqrt{d_{\text{inv}} \log \left( 1 + \frac{c\delta}{t} \right)} + \sqrt{d_{\text{spu}} \log \left( 1 + \frac{c\delta}{2m/d_{\text{spu}} \cdot t} \right)} \right).
\]  

(S13)

By basic calculus and a scaling argument as in [Wainwright, 2019, p. 427], we find the summation is bounded by

\[
\sqrt{10 \log n \delta} + c'(\sqrt{d_{\text{inv}} + \min \{ Z + e^{-mZ^2\delta^2/2}, 2^{-m/d_{\text{spu}}} \}} \sqrt{d_{\text{spu}}}) \delta.
\]

Therefore, for any \( Z, \delta_{0,n} > 0 \), any solution to

\[
\delta \geq \max \{ \delta_{0,n}, n^{-1/2} \}, \quad \sqrt{10 \log n \delta} + c''(\sqrt{d_{\text{inv}} + \min \{ Z + e^{-mZ^2\delta^2/2}, 2^{-m/d_{\text{spu}}} \}} \sqrt{d_{\text{spu}}}) \delta \leq \sqrt{\frac{7n\delta^2}{64\sigma^2}}
\]

always solves (S11) with the claimed probability. Choosing \( \delta_{0,n} = \sqrt{\frac{Z^2d_{\text{spu}}}{n}}, Z^2 = \left( \frac{2n\log m}{md_{\text{spu}}} \right)^{1/4} \) yields

\[
\delta_n^2 \leq c'' \left( \log n + d_{\text{inv}} + \frac{d_{\text{spu}} \log m}{nm} \right),
\]

while considering the second argument of \( \min \) yields

\[
\delta_n^2 \leq c'' \frac{\log n + d_{\text{inv}} + 2^{-2m/d_{\text{spu}}}d_{\text{spu}}}{n}.
\]

Both bounds hold with the aforementioned \( \mathcal{S}_{e,\tau} \)-probability. Plugging back to (S10) completes the proof.

\[ \square \]

S3 EXPERIMENT SETUP AND FULL RESULTS

S3.1 FULL RESULTS FOR SECTION 6.1

Full results for all methods in the setting of Figure 1 are reported in Figure S1, where we add the results for BLR-LC for \( N := 0.3n_s \), BLR-Prior for \( \alpha \in \{1, 3\} \), and CBLR for \( \rho = 0.2 \). The results are consistent with the discussion in the text. We have also conducted experiments in a larger-scale setting, with \( n_c \leftarrow 12000 \) for classification and \( n_c \leftarrow 18000 \), \( d_{\text{inv}} \leftarrow 80 \), \( d_{\text{spu}} \leftarrow 160 \), \( m \leftarrow 5 \) for regression. As shown in Figure S2, the results are qualitatively similar, with our methods performing slightly better for the regression task.

We further experimented with generalized variants of both the regression and classification experiments, where we replace the definition of \( \tilde{\beta}_{\text{spu}}^* \) with

\[
\tilde{\beta}_{\text{spu}}^* \sim \mathcal{N}\left( \frac{2\alpha}{m} \sum_{c \in \mathcal{C}r} \tilde{\beta}_{\text{spu},c} (1 - \alpha^2) \tau_c^2 I \right),
\]

where \( \tau_c = 1 \) for classification and \( d_{\text{spu}}^{-1/2} \) for regression. For regression, we also introduce environment-specific correlations between the invariant and spurious features, by replacing the generating process of \( x_{\text{spu},i}^c \) with

\[
x_{\text{spu},i}^c \sim \mathcal{N}(\beta A^c x_{\text{inv},i}^c (1 - \beta^2) I),
\]

where \( A^c \in \mathbb{R}^{d_{\text{spu}} \times d_{\text{inv}}} \) is a random matrix with i.i.d. \( \mathcal{N}(0, d_{\text{inv}}^{-1}) \) components. The data generating process in the text thus corresponds to \( \alpha = 1, \beta = 0 \). Results for other choices of \( (\alpha, \beta) \) are reported in Figure S3 and Figure S4, where we plot the distribution of test losses across 32 independently samples for \( \{ \tilde{\beta}_{\text{inv}}, \tilde{\beta}_{\text{spu}}^* \} \). As we can see, our method remains competitive across all settings.
S3.2 SETUP AND RESULTS FOR SECTION 6.2

Full results for all methods in the setting of Table 1 and Table 2 are reported in Table S1 and Table S2, respectively, where we also report the results for our method with \( \rho \in \{0.05, 0.2\} \). As we can see, our method achieves robust performance across all settings. In order to maintain a high acceptance rate for the M-H test of the Langevin Monte Carlo steps, we set the step-size upper bound \( \bar{\eta}_k = 0.001 \) for ColoredMNIST and \( \bar{\eta}_k = 0.0025 \) for PACS using binary search. To guarantee convergence, we run \( 2 \times 10^4 \) steps with 50 parallel chains for each method. In accordance with the ERM baseline, the batch size for the BLR-LC method is set to 32 for each domain. For the BLR-LC-N_train method, because of the relatively large number of training examples \( n_e \) on Colored MNIST, we set \( N := 0.02n_e \) for Colored MNIST, while \( N := n_e \) for PACS. The training-domain validation set selection approach from [Gulrajani and Lopez-Paz, 2020] is used to search the ERM baseline through 20 hyperparameter configurations \( \times 3 \) trials, which is important for ERM to establish itself as a strong baseline.

References


Figure S2: Synthetic experiment: results for all methods at a larger scale.

Figure S3: Synthetic classification experiment: violin plot of classification errors, across independently sampled environments, in the setting of Figure 1. From left to right: results for $\alpha \in \{0, 0.5, 1\}$. From top to bottom: results for $n_* \in \{4, 32, 256\}$. Best viewed when zoomed.
Figure S4: Synthetic regression experiment: violin plot of test MSE across independently sampled environments in the larger-scale setting ($n_e = 18000, m = 5, d = 240$). From left to right: results for $(\alpha, \beta) \in \{(0,0), (1,0), (1,0.5)\}$. 
Table S1: Colored MNIST: test accuracy for all methods on all domains. We report the 20th percentile / mean / 80th percentile across 20 independent runs.

<table>
<thead>
<tr>
<th>n_x</th>
<th>Method / ε_x</th>
<th>0.1</th>
<th>0.2</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>ERM</td>
<td>88.5</td>
<td>87.2</td>
<td>71.5</td>
</tr>
<tr>
<td>4</td>
<td>CBLR</td>
<td>87.8 / 88.2 / 88.5</td>
<td>86.7 / 87.0 / 87.3</td>
<td>81.5 / 81.9 / 85.6</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.05</td>
<td>87.8 / 87.9 / 88.0</td>
<td>87.1 / 87.3 / 87.7</td>
<td>79.8 / 81.9 / 85.9</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.20</td>
<td>85.0 / 85.7 / 87.2</td>
<td>86.0 / 86.6 / 87.6</td>
<td>79.3 / 76.4 / 86.9</td>
</tr>
<tr>
<td></td>
<td>BLR</td>
<td>75.2 / 80.6 / 88.1</td>
<td>77.5 / 79.3 / 85.8</td>
<td>88.4 / 87.5 / 90.0</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_adapt</td>
<td>88.1 / 88.0 / 88.7</td>
<td>86.6 / 87.0 / 87.4</td>
<td>80.5 / 81.8 / 85.6</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_train</td>
<td>88.3 / 88.4 / 88.5</td>
<td>87.2 / 87.3 / 87.4</td>
<td>67.9 / 69.3 / 71.1</td>
</tr>
<tr>
<td></td>
<td>DivDis</td>
<td>88.4 / 88.4 / 88.5</td>
<td>87.8 / 87.8 / 87.9</td>
<td>70.7 / 70.7 / 70.9</td>
</tr>
<tr>
<td>8</td>
<td>CBLR</td>
<td>87.8 / 88.2 / 88.5</td>
<td>86.5 / 87.0 / 87.4</td>
<td>85.0 / 85.7 / 87.1</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.05</td>
<td>87.8 / 88.1 / 88.6</td>
<td>87.0 / 87.2 / 87.5</td>
<td>83.9 / 85.1 / 86.8</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.20</td>
<td>85.4 / 86.1 / 88.0</td>
<td>85.1 / 86.0 / 87.1</td>
<td>85.3 / 86.7 / 88.9</td>
</tr>
<tr>
<td></td>
<td>BLR</td>
<td>83.0 / 86.1 / 88.4</td>
<td>80.9 / 81.7 / 86.5</td>
<td>89.2 / 89.3 / 90.0</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_adapt</td>
<td>87.9 / 88.3 / 88.7</td>
<td>87.0 / 87.1 / 87.5</td>
<td>82.9 / 83.5 / 86.0</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_train</td>
<td>88.3 / 88.4 / 88.5</td>
<td>87.2 / 87.3 / 87.4</td>
<td>68.3 / 69.9 / 71.8</td>
</tr>
<tr>
<td></td>
<td>DivDis</td>
<td>88.4 / 88.4 / 88.5</td>
<td>87.8 / 87.8 / 87.8</td>
<td>70.6 / 70.7 / 70.8</td>
</tr>
<tr>
<td>16</td>
<td>CBLR</td>
<td>87.6 / 88.2 / 88.7</td>
<td>86.8 / 87.0 / 87.4</td>
<td>87.1 / 87.7 / 88.5</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.05</td>
<td>87.9 / 88.1 / 88.6</td>
<td>86.9 / 87.2 / 87.5</td>
<td>86.4 / 87.0 / 87.9</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.20</td>
<td>85.7 / 86.3 / 88.5</td>
<td>86.0 / 86.5 / 87.3</td>
<td>87.9 / 88.6 / 89.4</td>
</tr>
<tr>
<td></td>
<td>BLR</td>
<td>86.0 / 87.4 / 88.5</td>
<td>85.4 / 85.5 / 87.0</td>
<td>89.1 / 89.4 / 90.0</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_adapt</td>
<td>88.3 / 88.5 / 88.7</td>
<td>87.1 / 87.3 / 87.5</td>
<td>83.9 / 84.9 / 86.9</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_train</td>
<td>88.2 / 88.4 / 88.6</td>
<td>87.2 / 87.3 / 87.4</td>
<td>70.1 / 71.4 / 73.4</td>
</tr>
<tr>
<td></td>
<td>DivDis</td>
<td>88.4 / 88.4 / 88.5</td>
<td>87.8 / 87.8 / 87.9</td>
<td>70.7 / 70.7 / 70.8</td>
</tr>
<tr>
<td>32</td>
<td>CBLR</td>
<td>88.2 / 88.5 / 88.8</td>
<td>86.7 / 87.0 / 87.4</td>
<td>88.6 / 88.8 / 89.1</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.05</td>
<td>88.1 / 88.4 / 88.7</td>
<td>87.0 / 87.2 / 87.5</td>
<td>87.5 / 88.0 / 88.4</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.20</td>
<td>87.2 / 87.7 / 88.5</td>
<td>86.2 / 86.8 / 87.4</td>
<td>89.1 / 89.4 / 89.7</td>
</tr>
<tr>
<td></td>
<td>BLR</td>
<td>87.8 / 88.2 / 88.8</td>
<td>86.3 / 86.8 / 87.3</td>
<td>89.6 / 89.8 / 90.0</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_adapt</td>
<td>88.5 / 88.6 / 88.7</td>
<td>87.2 / 87.3 / 87.5</td>
<td>84.8 / 85.6 / 86.4</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_train</td>
<td>88.2 / 88.4 / 88.6</td>
<td>87.2 / 87.3 / 87.4</td>
<td>72.7 / 73.6 / 75.4</td>
</tr>
<tr>
<td></td>
<td>DivDis</td>
<td>88.4 / 88.4 / 88.5</td>
<td>87.8 / 87.8 / 87.9</td>
<td>70.7 / 70.7 / 70.8</td>
</tr>
</tbody>
</table>
Table S2: PACS: test accuracy for all methods on all domains. We report the 20th percentile / mean / 80th percentile across 20 independent runs.

<table>
<thead>
<tr>
<th>n_e</th>
<th>Method / c_e</th>
<th>A</th>
<th>C</th>
<th>P</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>ERM</td>
<td>87.8</td>
<td>72.6</td>
<td>96.1</td>
<td>76.3</td>
</tr>
<tr>
<td></td>
<td>CBLR</td>
<td>87.8/88.5/89.5</td>
<td>80.8/81.7/82.7</td>
<td>96.7/97.1/97.6</td>
<td>77.6/78.3/79.5</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.05</td>
<td>86.6/87.4/88.8</td>
<td>79.5/80.2/81.6</td>
<td>96.1/96.7/97.3</td>
<td>76.1/76.9/78.0</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.20</td>
<td>83.9/85.4/87.0</td>
<td>78.2/79.3/81.0</td>
<td>95.8/96.3/96.7</td>
<td>73.8/74.9/76.4</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_adapt</td>
<td>86.8/87.7/89.2</td>
<td>76.1/78.8/82.9</td>
<td>95.5/96.1/97.0</td>
<td>70.1/72.5/75.5</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_train</td>
<td>86.1/86.8/87.5</td>
<td>79.7/80.0/80.6</td>
<td>96.7/96.9/97.0</td>
<td>76.1/76.4/77.6</td>
</tr>
<tr>
<td></td>
<td>DivDis</td>
<td>85.1/85.6/86.3</td>
<td>79.3/79.3/79.7</td>
<td>96.4/96.8/97.6</td>
<td>77.7/78.0/79.1</td>
</tr>
<tr>
<td>16</td>
<td>CBLR</td>
<td>88.8/89.8/90.7</td>
<td>82.7/83.3/84.2</td>
<td>96.7/97.1/97.3</td>
<td>78.6/79.2/80.1</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.05</td>
<td>87.5/88.5/89.5</td>
<td>80.3/81.6/82.7</td>
<td>96.1/96.6/97.0</td>
<td>77.2/78.3/79.5</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.20</td>
<td>86.1/87.0/88.5</td>
<td>79.5/80.8/82.1</td>
<td>96.1/96.4/97.0</td>
<td>75.4/76.8/78.2</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_adapt</td>
<td>88.8/89.6/90.7</td>
<td>82.1/82.9/84.4</td>
<td>96.4/96.9/97.6</td>
<td>76.4/77.6/79.2</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_train</td>
<td>89.0/89.5/90.2</td>
<td>81.6/82.5/83.8</td>
<td>97.0/97.3/97.6</td>
<td>76.4/77.0/79.9</td>
</tr>
<tr>
<td></td>
<td>DivDis</td>
<td>85.1/85.1/85.6</td>
<td>79.1/79.6/81.2</td>
<td>96.4/96.8/97.3</td>
<td>77.1/77.3/78.5</td>
</tr>
<tr>
<td>32</td>
<td>CBLR</td>
<td>89.7/90.8/91.7</td>
<td>84.2/85.0/85.7</td>
<td>97.0/97.4/97.9</td>
<td>79.9/80.5/81.4</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.05</td>
<td>89.0/89.4/90.2</td>
<td>82.3/83.3/84.2</td>
<td>96.4/96.9/97.3</td>
<td>78.5/79.4/80.4</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.20</td>
<td>88.0/88.8/90.0</td>
<td>82.7/83.1/83.8</td>
<td>96.4/96.8/97.3</td>
<td>76.7/78.3/79.6</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_adapt</td>
<td>90.0/90.5/91.4</td>
<td>84.2/85.3/86.1</td>
<td>97.0/97.3/97.9</td>
<td>79.4/80.9/81.3</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_train</td>
<td>89.2/90.0/90.7</td>
<td>82.3/82.7/84.4</td>
<td>97.0/97.2/97.6</td>
<td>77.6/79.0/80.5</td>
</tr>
<tr>
<td></td>
<td>DivDis</td>
<td>85.6/85.9/86.6</td>
<td>78.6/79.1/79.9</td>
<td>96.4/96.8/97.3</td>
<td>77.2/77.6/78.6</td>
</tr>
<tr>
<td>64</td>
<td>CBLR</td>
<td>90.7/91.5/92.2</td>
<td>86.3/86.9/87.8</td>
<td>97.3/97.6/98.2</td>
<td>81.5/82.4/83.7</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.05</td>
<td>90.0/90.4/91.4</td>
<td>84.2/85.5/87.6</td>
<td>97.3/97.3/97.6</td>
<td>80.8/81.4/83.4</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.20</td>
<td>89.5/90.6/92.4</td>
<td>84.0/84.8/87.4</td>
<td>96.7/96.9/97.3</td>
<td>79.2/79.7/81.3</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_adapt</td>
<td>90.5/91.1/91.7</td>
<td>86.1/86.6/87.4</td>
<td>96.7/97.4/97.9</td>
<td>80.1/81.2/82.5</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_train</td>
<td>89.2/89.9/90.7</td>
<td>82.7/83.2/84.2</td>
<td>97.0/97.3/97.9</td>
<td>79.1/79.8/80.6</td>
</tr>
<tr>
<td></td>
<td>DivDis</td>
<td>84.8/85.4/86.1</td>
<td>78.8/79.7/80.6</td>
<td>96.4/96.9/97.3</td>
<td>76.7/77.2/78.1</td>
</tr>
<tr>
<td>128</td>
<td>CBLR</td>
<td>91.9/92.5/92.9</td>
<td>86.3/86.7/88.5</td>
<td>97.6/97.9/98.2</td>
<td>83.7/84.2/85.1</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.05</td>
<td>91.9/92.3/92.7</td>
<td>78.4/82.9/87.6</td>
<td>97.3/97.5/97.9</td>
<td>82.9/83.5/84.2</td>
</tr>
<tr>
<td></td>
<td>CBLR_0.20</td>
<td>91.4/91.9/92.4</td>
<td>85.7/86.9/88.7</td>
<td>97.3/97.5/97.9</td>
<td>81.8/82.9/83.8</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_adapt</td>
<td>91.4/91.8/92.4</td>
<td>85.9/86.5/87.4</td>
<td>97.3/97.6/98.2</td>
<td>80.6/81.7/82.8</td>
</tr>
<tr>
<td></td>
<td>BLR-LC-N_train</td>
<td>89.7/90.0/90.5</td>
<td>80.8/82.5/84.2</td>
<td>97.0/97.4/97.9</td>
<td>77.2/78.3/80.1</td>
</tr>
<tr>
<td></td>
<td>DivDis</td>
<td>85.3/85.8/87.0</td>
<td>80.6/80.6/80.8</td>
<td>96.7/96.9/97.3</td>
<td>76.3/77.0/78.1</td>
</tr>
</tbody>
</table>