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

Ziyu Wang[*1] **Binjie Yuan**[*1] **Jiaxun Lu**[2] **Bowen Ding**[3] **Yunfeng Shao**[2] **Qibin Wu**[3] **Jun Zhu**[#1]

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

Additional notations and conventions The following conventions are used throughout appendix: the denotation of constants $(c_1, c_2, \ldots, C_1, C_2, \ldots)$ may change from line to line. $\|\cdot\|_2$ denotes the Euclidean norm for vectors, or the $L_3(P_x^*)$ norm for functions of x. $\langle\cdot,\cdot\rangle_2$ denotes the respective inner product. $\|\cdot\|_F$ denotes the Frobenius norm. ϕ_z, Φ_z denote the PDF and CDF of the standard Gaussian distribution. Recall the inequality $1 - \Phi_z(x) \leq e^{-x^2/2}$ 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 $\alpha > 0, m \ll \max\{1, \sigma_s^{-2}\tau_s^2\}d_{spu}$, the classifier parameterized by $\tilde{\theta} = (0, \tilde{\theta}_s) = (0, \alpha\sum_{e\in\mathcal{E}_{tr}}\bar{\beta}_{spu}^e)$ satisfies*

$$R_{e,01}(f_{\tilde{\theta}}) \leq c_1 \exp(-c_2\sigma_s^{-2}\tau_s^2 d_{spu}/m + o_p(1)) \to 0, \;\; \forall e \in \mathcal{E}_{tr}. \tag{S1}$$

Moreover, the same bound holds for the logistic loss, if in addition $\alpha \leq \mathcal{O}(\mathrm{poly}(d_{spu}/m)), \alpha\tau_s^2 \gtrsim 1/\sigma_s^2 m$.

Lemma S2. *Denote by $\mathbf{x}_i \in \mathbb{R}^{d_{inv}}, \mathbf{x}_s \in \mathbb{R}^{d_{spu}}$ the invariant and spurious components of the input $\mathbf{x}$. For any $e_1, e_2 \in \mathcal{E}_{tr}$, let*

$$KL_{ij} := \mathrm{KL}(p_{marg,e_1} \,\|\, p_{marg,e_2}) = \mathrm{KL}(p_{e_1}(\mathbf{y}, \tilde{\theta}_s^\top\mathbf{x}_s, \mathbf{x}_i) \,\|\, p_{e_2}(\mathbf{y}, \tilde{\theta}_s^\top\mathbf{x}_s, \mathbf{x}_i))$$

denote the KL divergence between the marginal distributions. When $\alpha > 0, m < d_{spu}/16$, with probability $\geq 1 - e^{-m/18}$ there exists an $\ell \in [m]$, determined by $\{\bar{\beta}_{spu}^e : e \in \mathcal{E}_{tr}\}$ s.t.

$$\mathrm{KL}\Big(\bigotimes_{e\in\mathcal{E}_{tr}} p_{marg,e} \,\Big\|\, p_{marg,e_\ell}^{\otimes m}\Big) = \sum_{j=1}^m \mathrm{KL}_{j\ell} \leq \frac{256m}{\sigma_s^2 d_{spu}}, \tag{S2}$$

The last claim in the text, Eq. 2, follows by a standard argument following (S2): consider the scenario $n_e \equiv n$ for simplicity. Denote by $\mathcal{D}_\top := (((y_k^{e_j}, \tilde{\theta}_s^\top x_{k,s}^{e_j}, x_{k,i}^{e_j}) : j \in [m]) : k \in [n])$ the transposed dataset which contains the same information as $\mathcal{D}_{tr}$. Any test on $\mathcal{D}_{tr}$ for the null hypothesis "$\tilde{\theta}_s^\top\mathbf{x}_s$ is an invariant feature" always provides a two-sample test for

$$H_0' : \mathcal{D}_\top \sim \big((p_{marg,e_\ell})^{\otimes m}\big)^{\otimes n} =: P_{0,n}, \;\; H_1' : \mathcal{D}_\top \sim \Big(\bigotimes_{e\in\mathcal{E}_{tr}} p_{marg,e}\Big)^{\otimes n} =: P_{1,n},$$

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

$$1 - D_{TV}(P_{0,n}, P_{1,n}) \geq 1 - \sqrt{2\mathrm{KL}(P_{1,n} \,\|\, P_{0,n})} \geq 1 - n \cdot \frac{256m}{\sigma_s^2 d_{spu}}.$$

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 $\Phi = \mathrm{id}$ and $w(h) = \tilde{\theta}^\top h$, approximately satisfies the hard constraints in (IRM) and constitutes an approximate optima in this sense.[1]

Remark 2 (interpretations of the KL bound, additional error from feature learning). In addition to the testing-based interpretation as in Eq. 2, (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 $j \in [m]$,

$$\mathrm{KL}(p_{e_j}(y \mid \tilde{\theta}^\top x_s) \,\|\, p_{e_\ell}(y \mid \tilde{\theta}^\top x_s)) \le \mathrm{KL}(p_{e_j}(y \mid x_i, \tilde{\theta}^\top x_s) \,\|\, p_{e_\ell}(y \mid x_i, \tilde{\theta}^\top x_s)) \le \mathrm{KL}_{j\ell} \le \frac{256m}{\sigma_s^2 d_{spu}}.$$

Such tests can be viewed as restricting to the validation of the definition 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 $d_{spu}/n \gg m/\sigma_s^2 d_{spu}$, i.e.,

$$\max_{e \in \mathcal{E}_{tr}} n_e \ll \frac{\sigma_s^2 d_{spu}^2}{m}.$$

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 $|\mathcal{M}|$ feature extractors with independent failure probabilities, we would require $n \gtrsim |\mathcal{M}|\sigma_s^2 d_{spu}/m$ 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: $\mathcal{E}_{tr} =: \{e_1, \ldots, e_m\}$, $\vec{\mathbf{1}} := \{1, \ldots, 1\} \in \mathbb{R}^m$, and define the $m \times m$ matrix $(\Sigma_{\mathcal{S}})_{ij} = (\bar{\beta}_{spu}^{e_i})^\top (\bar{\beta}_{spu}^{e_j})$, so that $\|\tilde{\theta}_s\|_2^2 = \alpha^2 \vec{\mathbf{1}}^\top \Sigma_{\mathcal{S}} \vec{\mathbf{1}}$. Define $\bar{\Sigma}_{\mathcal{S}} := \mathbb{E}_{\{\bar{\beta}_{spu}^e\}} \Sigma_{\mathcal{S}} = \tau_s^2 d_{spu} I$. Note that by covariance concentration [Wainwright, 2019, Ch. 6], we have, when $m \le d_{spu}$,

$$\mathbb{P}_{\{\bar{\beta}_{spu}^e\}}\left(\frac{1}{\tau_s\sqrt{d_{spu}}}\|\bar{\Sigma}_{\mathcal{S}} - \Sigma_{\mathcal{S}}\| \le 3\sqrt{\frac{m}{d_{spu}}} + \delta\right) \ge 1 - e^{-d_{spu}\delta^2/18}. \tag{S3}$$

Proof for lemma S1. We first derive the 0-1 loss for $\tilde{\theta}$. Note that in the setting of the example, we have, for any $\theta = (\theta_i, \theta_s)$,

$$R_{e,01}(f_\theta) := \mathbb{E}_{\mathbf{x}^e, \mathbf{y}^e} \mathbf{1}\{\mathrm{sgn}(\theta^\top \mathbf{x}^e) = \mathbf{y}^e\} = \Phi_z\left(-\frac{\theta^\top \mathbb{E}(\mathbf{x}^e \mid \mathbf{y}^e = 1)}{\sqrt{\theta^\top \mathrm{Cov}_e(\mathbf{x}^e (\mathbf{x}^e)^\top)\theta}}\right) = \Phi_z\left(-\frac{\theta_i^\top \bar{\beta}_{inv} + \theta_s^\top \bar{\beta}_{spu}^e}{\sqrt{\sigma_i^2\|\theta_i\|_2^2 + \sigma_s^2\|\theta_s\|_2^2}}\right).$$

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

$$\|\tilde{\theta}_s\|_2 = \alpha\sqrt{\vec{\mathbf{1}}^\top \Sigma_{\mathcal{S}} \vec{\mathbf{1}}} \le \alpha\sqrt{m(\|\bar{\Sigma}_{\mathcal{S}}\| + \|\bar{\Sigma}_{\mathcal{S}} - \Sigma_{\mathcal{S}}\|)} = \alpha\tau_s\sqrt{m d_{spu}}(1 + \mathcal{O}_p((m/d_{spu})^{1/4})), \tag{S4}$$

$$\tilde{\theta}_s^\top \bar{\beta}_{spu}^e = \alpha\left(\|\bar{\beta}_{spu}^e\|_2^2 + \sum_{e' \in \mathcal{E}_{tr}, e' \neq e} \langle \bar{\beta}_{spu}^{e'}, \bar{\beta}_{spu}^e \rangle_2\right) = \alpha\tau_s^2(d_{spu} + \mathcal{O}_p(\sqrt{m d_{spu}})). \tag{S5}$$

Thus, when $m \ll \min\{1, \sigma_s^{-2}\tau_s^2\} d_{spu}$, we have

$$R_{e,01}(f_{\tilde{\theta}}) = \Phi_z(-\sigma_s^{-1}\tau_s(\sqrt{d_{spu}/m} + o_p(1))) \le \exp(-2\sigma_s^{-2}\tau_s^2 d_{spu}/m + o_p(1)) \xrightarrow{p} 1, \quad \forall e \in \mathcal{E}_{tr}.$$

[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 $\|\nabla_w R_{e,log}(w \circ \Phi)\|_2 = \|\nabla_{\tilde{\theta}} R_{e,log} f_{\tilde{\theta}}\|_2$ vanishes at a similar exponential rate. Thus, such a (w, Φ) pair is also an approximate optima for the soft constraint-based formulation.

Now we consider the logistic loss. Fix any $e \in \mathcal{E}_{tr}$, and recall $\bar{\theta}_e := \bar{\beta}^e_{spu}/2\sigma_s^2$ defines the Bayes classifier given the *spurious features* $\mathbf{x}^e_{spu} \in \mathbb{R}^{d_{spu}}$. Introduce the random variables

$$\mathbf{x} \sim \mathcal{N}(\bar{\beta}^e_{spu}, \sigma_s^2 I), \;\; \mathbf{z}_1 = \bar{\theta}_e^\top \mathbf{x}, \;\; \mathbf{z}_2 = \tilde{\theta}_s^\top \mathbf{x},$$

so that

$$\begin{aligned} R_{e,log}(\tilde{\theta}) &= \mathbb{E}_{\mathbf{x}^e_{spu}} \sum_{y\in\{\pm 1\}} p(\mathbf{y}^e = y \mid \mathbf{x}^e_{spu}) \log(1 + e^{-y\tilde{\theta}^\top \mathbf{x}^e_{spu}}) \\ &= \mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2} \mathbf{1}\{\mathbf{z}_1 \geq 0\} \log(1+e^{-\mathbf{z}_2}) + \mathbf{1}\{\mathbf{z}_1 < 0\} \log(1+e^{\mathbf{z}_2}) \\ &\leq \mathbb{E}\log(1+e^{-\mathbf{z}_2}) + \mathbb{E}(\mathbf{1}\{\mathbf{z}_1 < 0\} \max\{1, 2\mathbf{z}_2\}) \\ &\leq \underbrace{\mathbb{E}\log(1+e^{-\mathbf{z}_2})}_{(I)} + \underbrace{\mathbb{E}\mathbf{1}\{\mathbf{z}_1 < 0\}}_{(II)} + \underbrace{\mathbb{E}\mathbf{1}\{\mathbf{z}_1 < 0, \mathbf{z}_2 > 1/2\} 2\mathbf{z}_2}_{(III)}. \end{aligned}$$

We will bound the three terms in turn. Before that, we first derive the joint distribution of $(\mathbf{z}_1, \mathbf{z}_2)$. By (S4)-(S5), we have

$$\mu_2 := \mathbb{E}\mathbf{z}_2 = \alpha\tau_s^2 d_{spu}(1+o_p(1)), \;\sigma_2^2 := \mathrm{Var}(\mathbf{z}_2) = \sigma_s^2\alpha^2\tau_s^2 m d_{spu}(1+o_p(1)), \;\mathrm{Cov}(\mathbf{z}_1,\mathbf{z}_2) = \frac{1+o_p(1)}{2\sigma_s^2}\alpha\tau_s^2 d_{spu}.$$

Moreover, we have

$$\mu_1 := \mathbb{E}\mathbf{z}_1 = \frac{\|\bar{\beta}^e_{spu}\|_2^2}{2\sigma_s^2} = \frac{\tau_s^2(d_{spu} + \mathcal{O}_p(\sqrt{d_{spu}}))}{2\sigma_s^2}, \;\; \sigma_1^2 := \mathrm{Var}(\mathbf{z}_1) = \frac{1}{2}\mu_1.$$

We now return to the three terms for $R_{e,log}$. For (I), consider the decomposition

$$\begin{aligned} (I) &= \mathbb{E}\log(1+e^{-\mathbf{z}_2})\mathbf{1}\{\mathbf{z}_2 < -1/2\} + \mathbb{E}\log(1+e^{-\mathbf{z}_2})\mathbf{1}\{-1/2 < \mathbf{z}_2 < A\} + \mathbb{E}\log(1+e^{-\mathbf{z}_2})\mathbf{1}\{\mathbf{z}_2 > A\} \\ &\leq \mathbb{E}(-2\mathbf{z}_2 \mid \mathbf{z}_2 < -1/2)\mathbb{P}(\mathbf{z}_2 < -1/2) + \mathbf{1}\{-1/2 < \mathbf{z}_2 < A\} + \log(1+e^{-A}) \\ &\overset{(i)}{\lesssim} e^{-\mu_2^2/2\sigma_2^2} \cdot \left(1 + \frac{\sigma_2^2}{\mu_2}\right) + e^{-(\mu_2 - A)^2/2\sigma_2^2} + \log(1+e^{-A}) \\ &\overset{(ii)}{\lesssim} e^{-\mu_2^2/3\sigma_2^2} + e^{-(\mu_2-A)^2/2\sigma_2^2} + e^{-A} \overset{(iii)}{\lesssim} e^{-\mu_2^2/3\sigma_2^2}. \end{aligned}$$

In the above, (i) follows by the Gaussian CDF bound and lemma S3 below, (ii) follows since $\mu_2^2 \gg \sigma_2^2 \gg 1, \alpha = \mathcal{O}(\mathrm{poly}(d_{spu}/m))$, and (iii) sets $A = \mu_2/10$. Now,

$$\begin{aligned} (II) &\leq e^{-\mu_2^2/2\sigma_2^2}, \\ (III) &= \int_{1/2}^{\infty} 2t\mathbb{P}(\mathbf{z}_1 < 0 \mid \mathbf{z}_2 = t)\mathbb{P}_{\mathbf{z}_2}(dt) \leq \mathbb{P}(\mathbf{z}_1 < 0)\mathbb{E}(2\mathbf{z}_2 \mid \mathbf{z}_2 > 1/2) \lesssim e^{-\mu_1^2/2\sigma_1^2}(\mu_2 + 2\sigma_2 e^{-\mu_2^2/2\sigma_2^2}), \end{aligned}$$

where the first inequality follows because $\mathbf{z}_1$ and $\mathbf{z}_2$ are positively correlated, and $\mu_2 > 0$, and the second follows by an application of lemma S3. Combining, we can see that in the regime $m \ll \sigma_s^{-2}\tau_s^2 d_{spu}$, there exists $c_1, c_2 > 0$ s.t.

$$R_{e,\log}(\tilde{\theta}) \leq (I) + (II) + (III) \leq c_1 e^{-c_2\tau_s^2 d_{spu}/m\sigma_s^2}.$$

This proves (S1). □

Proof for lemma S2. For any $e_1, e_2 \in \mathcal{E}_{tr}$ and fixed realizations of $\{\bar{\beta}^{e_1}_{spu}, \bar{\beta}^{e_2}_{spu}\}$ (i.e., the following display implicitly conditions on the two), we have

$$\begin{aligned} \mathrm{KL}_{ij} = \mathrm{KL}(p_{e_1}(\mathbf{y}, \tilde{\theta}_s^\top \mathbf{x}_s, \mathbf{x}_i) \,\|\, p_{e_2}(\mathbf{y}, \tilde{\theta}_s^\top \mathbf{x}_s, \mathbf{x}_i)) &= \mathbb{E}_{p_{e_1}} \log \frac{p_{e_1}(\mathbf{y}, \tilde{\theta}_s^\top \mathbf{x}_s, \mathbf{x}_i)}{p_{e_2}(\mathbf{y}, \tilde{\theta}_s^\top \mathbf{x}_s, \mathbf{x}_i)} = \mathbb{E}_{p_{e_1}} \log \frac{p_{e_1}(\tilde{\theta}_s^\top \mathbf{x}_s \mid \mathbf{y}, \cancel{\mathbf{x}_i}) p_{e_1}(\mathbf{y}, \mathbf{x}_i)}{p_{e_2}(\tilde{\theta}_s^\top \mathbf{x}_s \mid \mathbf{y}, \cancel{\mathbf{x}_i}) p_{e_2}(\mathbf{y}, \mathbf{x}_i)} \\ &= \mathbb{E}_{p_{e_1}} \log \frac{p_{e_1}(\tilde{\theta}_s^\top \mathbf{x}_s \mid \mathbf{y})}{p_{e_2}(\tilde{\theta}_s^\top \mathbf{x}_s \mid \mathbf{y})} = \mathrm{KL}(\mathcal{N}(\mu_{21}, \sigma_2^2) \,\|\, \mathcal{N}(\mu_{22}, \sigma_2^2)) \end{aligned}$$

where $\mu_{2j} := \langle \bar{\beta}^{e_j}_{spu}, \tilde{\theta}_s \rangle_2 \ \forall j \in [m]$, and $\sigma_2^2 := \sigma_s^2 \|\tilde{\theta}_s\|_2^2$. Plugging in the expression for KL divergence between Gaussian distributions, we find, *for all* $i, j \in [m]$,

$$\begin{aligned}
\mathrm{KL}_{ij} &= \frac{(\mu_{2i} - \mu_{2j})^2}{2\sigma_2^2} = \frac{\alpha^2}{2\sigma_2^2} \Big(\sum_{k=1}^m (\Sigma_{\mathcal{S}})_{ik} - (\Sigma_{\mathcal{S}})_{jk} \Big)^2 \\
&= \frac{\alpha^2}{2\sigma_2^2} \Big(\sum_{k=1}^m (\Sigma_{\mathcal{S}})_{ik} - (\bar{\Sigma}_{\mathcal{S}})_{ik} + \sum_{k=1}^m (\bar{\Sigma}_{\mathcal{S}})_{jk} - (\Sigma_{\mathcal{S}})_{jk} + \cancel{\sum_{k=1}^m (\bar{\Sigma}_{\mathcal{S}})_{ik} - (\bar{\Sigma}_{\mathcal{S}})_{jk}} \Big)^2 \\
&\leq \frac{4\alpha^2}{2\sigma_2^2} \bigg(\Big(\sum_{k=1}^m (\Sigma_{\mathcal{S}})_{ik} - (\bar{\Sigma}_{\mathcal{S}})_{ik} \Big)^2 + \Big(\sum_{k=1}^m (\Sigma_{\mathcal{S}})_{jk} - (\bar{\Sigma}_{\mathcal{S}})_{jk} \Big)^2 \bigg) \\
&= \frac{2\alpha^2}{\sigma_2^2} \big((\vec{\mathbf{1}}^\top (\Sigma_{\mathcal{S}} - \bar{\Sigma}_{\mathcal{S}}) \mathbf{e}_i)^2 + (\vec{\mathbf{1}}^\top (\Sigma_{\mathcal{S}} - \bar{\Sigma}_{\mathcal{S}}) \mathbf{e}_j)^2 \big),
\end{aligned} \tag{S6}$$

where $\mathbf{e}_i = (\underbrace{\ldots, 0}_{(i-1) \text{ zeros}}, 1, 0, \ldots)$ denotes the i-th Euclidean basis. We thus have, by symmetry and the trace formula,

$$\sum_{i=1}^m \sum_{j=1}^m \mathrm{KL}_{ij} \leq 2m \cdot \sum_{l=1}^m \frac{2\alpha^2}{\sigma_2^2} (\vec{\mathbf{1}}^\top (\Sigma_{\mathcal{S}} - \bar{\Sigma}_{\mathcal{S}}) \mathbf{e}_l)^2 = 2m \cdot \sum_{l=1}^m \frac{2\alpha^2}{\sigma_2^2} \vec{\mathbf{1}}^\top (\Sigma_{\mathcal{S}} - \bar{\Sigma}_{\mathcal{S}})^2 \vec{\mathbf{1}}.$$

Plugging in (S3) with $\delta \leftarrow \sqrt{m/d_{spu}}$ we find, when $d \geq 64m$, on that event we have

$$\begin{aligned}
\sum_{i=1}^m \sum_{j=1}^m \mathrm{KL}_{ij} &\leq \frac{4\alpha^2 m \|\vec{\mathbf{1}}\|_2^2 \|\Sigma_{\mathcal{S}} - \bar{\Sigma}_{\mathcal{S}}\|^2}{\sigma_2^2} \overset{\text{(S3)}}{\leq} \frac{4\alpha^2 m^2 \cdot (4\tau_s \sqrt{m})^2}{\sigma_2^2} = \frac{64\alpha^2 \tau_s^2 m^3}{\sigma_s^2 \|\tilde{\theta}_s\|_2^2} \\
&\overset{\text{(S4)}}{\leq} \frac{64\alpha^2 \tau_s^2 m^3}{\sigma_s^2 \alpha^2 \tau_s^2 m d_{spu}/4} = \frac{256 m^2}{\sigma_s^2 d_{spu}},
\end{aligned}$$

where we note that (S4) holds on the same event. Therefore, on that event there always exist some $\ell \in [m]$ which is determined by $\{\bar{\beta}^e_{spu}\}$ s.t.

$$\sum_{j=1}^m \mathrm{KL}_{\ell j} \leq \frac{256m}{\sigma_s^2 d_{spu}}.$$

This completes the proof. □

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

$$\mathbb{E}(\mathbf{z} \mid \mathbf{z} < b) \overset{(i)}{=} \mu - \sigma \frac{\phi_z((b-\mu)/\sigma)}{\Phi_z((b-\mu)/\sigma)} \overset{(ii)}{\geq} b - \frac{\sigma^2}{\mu - b}, \ \ \mathbb{E}(\mathbf{z} \mid \mathbf{z} < b) \overset{(i)}{=} \mu + \sigma \frac{\phi_z((b-\mu)/\sigma)}{1 - \Phi_z((b-\mu)/\sigma)} \leq \mu + \sigma \phi_z((b-\mu)/\sigma).$$

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: $\mathbf{x}^e \leftarrow M^{-1} \mathbf{x}^e, \theta \leftarrow M^\top \theta$. This allows us to ignore the presence of M, as long as we replace U with $U' := \|M\| U$. We also introduce the following notations:

$$\begin{gathered}
d := d_{inv} + d_{spu}, \ X := \{x_1^*, \ldots, x_{n_*}^*\} \in \mathbb{R}^{n_* \times d}, \ Y := \{y_1^*, \ldots, y_{n_*}^*\} \in \mathbb{R}^{n_* \times 1}, \\
\bar{\theta}^e_{spu} := (\bar{\beta}_{inv}, \bar{\beta}^e_{spu}), \ \mathcal{S}_{tr} := \{\bar{\beta}^e_{spu} : e \in \mathcal{E}_{tr}\}.
\end{gathered}$$

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

$$R^e(\theta) = \mathbb{E}_{\mathbf{x}^e, \mathbf{y}^e} (\theta^\top \mathbf{x}^e - \mathbf{y}^e)^2 = \|\theta - \bar{\theta}^e_{spu}\|_2^2 + \sigma^2,$$

and the constraint set, for which we have fixed $\rho + \varepsilon_n = 0$, reduces to

$$\begin{aligned}\mathcal{C}_{tr} &= \{\theta \in \mathbb{R}^d : \|\theta - \bar\theta^e_{spu}\|_2^2 \le \|\bar\theta_{inv} - \bar\theta^e_{spu}\|_2^2 \ \ \forall e \in \mathcal{E}_{tr}\} \\ &= \{\theta = (\beta_i, \beta_s) \in \mathbb{R}^d : \|\beta_i - \bar\beta_{inv}\|_2^2 + \|\beta_s - \bar\beta^e_{spu}\|_2^2 \le \|\bar\beta^e_{spu}\|_2^2 \ \ \forall e \in \mathcal{E}_{tr}\} \\ &\subset \left\{\theta = (\beta_i, \beta_s) \in \mathbb{R}^d : \langle \beta_s, \bar\beta^e_{spu}\rangle_2 \ge \frac{1}{2}\|\beta_s\|_2^2 \ \ \forall e \in \mathcal{E}_{tr}\right\}.\end{aligned}$$

The *constrained* parameter space is

$$\mathcal{F} = \{f_\theta : \|\theta\|_2^2 \le U', \theta \in \mathcal{C}_{tr}\}.$$

Note that by construction, $\bar\theta_{inv} \in \mathcal{F}$ always holds.

Our main taks is to establish improved metric entropy bounds for the following "localized" space:

$$\partial\mathcal{F}_\delta := \left\{f_\theta - f_{\theta'} : f_\theta, f_{\theta'} \in \mathcal{F}, \|f_\theta - f_{\theta'}\|_{n,2}^2 := \frac{1}{n_*}\sum_{i=1}^{n_*}(f_\theta(x_i^*) - f_{\theta'}(x_i^*))^2 \le \delta\right\}.$$

We first note that

Lemma S4. *When $n_* \ge 5d$, there exists $c > 0$ s.t.*

$$\mathbb{P}_X\left(\forall \theta, \theta' \in \mathbb{R}^d,\ \frac{1}{2}\|\theta - \theta'\|_2^2 \le \|f_\theta - f_{\theta'}\|_{n,2}^2 \le 2\|\theta - \theta'\|_2^2\right) \ge 1 - e^{-cn_*}. \tag{S7}$$

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

Proposition S5 (entropy bound). *On the event* (S7) *we have, for any $\zeta, \delta, t, Z > 0$, with $\mathcal{S}_{tr}$-probability $\ge 1 - \zeta$,*

$$\begin{aligned}\log N(\partial\mathcal{F}_\delta, \|\cdot\|_{n,2}, t) \le \min\Bigg\{&d_{spu}\log\left(1 + \frac{c'\delta}{e^{mZ^2\delta^2/2}t}\right) + d_{spu}\log\left(1 + \frac{c'Z\delta}{t}\right), \\ &d_{spu}\log\left(1 + \frac{c'\delta}{2^{-m/d_{spu}}t}\right)\Bigg\} + d_{inv}\log\left(1 + \frac{c'\delta}{t}\right) + \log\zeta^{-1}.\end{aligned}$$

Proof. We condition on the event (S7) throughout the proof. Then, any $\theta = (\beta_i, \beta_s)$ s.t. $f_\theta \in \partial\mathcal{F}_\delta$ must satisfy

$$\|\beta_i - \bar\beta_{inv}\|_2^2 + \|\beta_s\|_2^2 = \|\theta - \bar\theta_{inv}\|_2^2 \overset{\text{(S7)}}{\le} 2\|f_\theta - f_{\bar\theta_{inv}}\|_{n,2}^2 \le 2\delta^2 \ \ \Rightarrow \ \ \max\{\|\beta_i - \bar\beta_{inv}\|_2, \|\beta_s\|_2\} \le \sqrt{2}\delta,$$

and

$$\beta_s \in \mathcal{C}_s := \{\beta_s : \langle\beta_s, \bar\beta^e_{spu}\rangle_2 \ge \frac{1}{2}\|\beta_s\|_2^2 \ \ \forall e \in \mathcal{E}_{tr}\}.$$

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 $\mathbb{B}_{i,\delta} := \{\beta_i \in \mathbb{R}^{d_{inv}} : \|\beta_i - \bar\beta_{inv}\|_2 \le \sqrt{2}\delta\}$, and a $(\|\cdot\|_2, t/3)$-covering for $\mathcal{C}_s \cap \mathbb{B}_{s,\delta}$, where $\mathbb{B}_{s,\delta} := \{\beta_s \in \mathbb{R}^{d_{spu}} : \|\beta_s\|_2 \le \sqrt{2}\delta\}$. Thus, we have, for some $c, c' > 0$,

$$\begin{aligned}\log N(\partial\mathcal{F}_\delta, \|\cdot\|_{n,2}, t) &\le \log N(\mathbb{B}_{i,\delta}, \|\cdot\|_2, ct) + \log N(\mathbb{B}_{s,\delta} \cap \mathcal{C}_s, \|\cdot\|_2, ct) \\ &\le d_{inv}\log(1 + c'\delta/t) + \log N(\mathbb{B}_{s,\delta} \cap \mathcal{C}_s, \|\cdot\|_2, ct),\end{aligned} \tag{S8}$$

where the second inequality can be found as Wainwright [2019, Example 5.8]. It remains to bound $\log N(\mathbb{B}_{s,\delta} \cap \mathcal{C}_s, \|\cdot\|_2, ct)$.

For this purpose, first note that, for any fixed $\beta \ne 0$, we have

$$\begin{aligned}\mathbb{P}_{\mathcal{S}_{tr}}(\beta \in \mathcal{C}_s) &= \prod_{e \in \mathcal{E}_{tr}} \mathbb{P}_{\bar\beta^e_{spu} \sim \mathcal{N}(0, d_{spu}^{-1}I)}\left(\langle\beta, \bar\beta^e_{spu}\rangle_2 \ge \frac{1}{2}\|\beta\|_2^2\right) \\ &= \prod_{j=1}^m \mathbb{P}\left(\mathcal{N}(0, d_{spu}^{-1}\|\beta\|_2^2) \ge \frac{1}{2}\|\beta\|_2^2\right) \le \min\{e^{-md_{spu}\|\beta\|_2^2/4},\ 2^{-m}\}.\end{aligned} \tag{S9}$$

Introduce

$$B_{s1} := \{\beta_s \in \mathbb{R}^{d_{spu}} : \|\beta_s\|_2 \le \sqrt{2}Z\delta\}, \ \ B_{s2} := \mathbb{B}_{s,\delta} \backslash B_{s1},$$

and $\mathrm{C}_{s1}, \mathrm{C}_{s2}$ be $(\|\cdot\|, ct)$-coverings for B_{s1}, B_{s2} with the optimal cardinality. Then

$$\begin{aligned}\mathbb{E}_{\mathcal{E}_{tr}} N(\mathbb{B}_{s,\delta} \cap \mathcal{C}_s, \|\cdot\|_2, ct) &\le \sum_{\beta\in \mathrm{C}_{s1}} \mathbb{P}(\beta \in \mathcal{C}_s) + \sum_{\beta\in \mathrm{C}_{s2}} \mathbb{P}(\beta \in \mathcal{C}_s) \\ &\overset{\text{(S9)}}{\le} N(B_{s1}, \|\cdot\|, ct) + N(\mathbb{B}_{s,\delta}, \|\cdot\|, ct) e^{-m d_{spu} Z^2\delta^2/2} \\ &\le (1 + cZ\delta/t)^{d_{spu}} + (1 + c'\delta/t)^{d_{spu}} e^{-M d_{spu} Z^2\delta^2/2}.\end{aligned}$$

By Markov's inequality and the monotonicity of $\log(\cdot)$, we find that for all $\zeta \in (0, 1)$, with $\mathcal{S}_{tr}$-probability $\ge 1 - \zeta$,

$$\begin{aligned}\log N(\mathbb{B}_{s,\delta} \cap \mathcal{C}_s, \|\cdot\|_2, ct) &\le d_{spu} \log(1 + c'Z\delta/t) + d_{spu} \log\left(\frac{1 + c'\delta/t}{e^{MZ^2\delta^2/2}}\right) + \log \zeta^{-1} \\ &< d_{spu} \log(1 + c'Z\delta/t) + d_{spu} \log\left(1 + \frac{c'\delta}{e^{MZ^2\delta^2/2} \cdot t}\right) + \log \zeta^{-1}.\end{aligned}$$

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}). $\square$

This allows us to prove our main result:

Proof for Proposition 2. It suffices to bound the critical radius $\hat\delta_n^2$ of the Gaussian complexity of $\partial\mathcal{F}_\delta$ [Wainwright, 2019, Eq. (13.42a)]; given a high-probability bound $\hat\delta_n^2 \le \delta_n^2$ that holds w. p. $1 - \zeta'$, we will have, with probability $\ge 1 - \zeta' - c_1 e^{-c_2 n\delta_n^2}$,

$$\|f_{\hat\theta} - f_{\bar\theta_{inv}}\|_2^2 \overset{\text{(S7)}}{\le} 2\|f_{\hat\theta} - f_{\bar\theta_{inv}}\|_2^2 \le c_0 \|f_{\bar\theta_{inv}} - f_{\bar\theta^e_{spu}}\|_2^2 + c_1\delta_n^2, \tag{S10}$$

where the last inequality is Wainwright [2019, Theorem 13.13].

By Wainwright [2019, Corollary 13.7], any solution δ to the following inequality will bound $\hat\delta_n^2$:

$$\frac{16}{\sqrt{n}} \int_{\delta^2}^{\delta} \sqrt{\log N(\partial\mathcal{F}_\delta, \|\cdot\|_{n,2}, t)}\, dt \le \frac{\delta^2}{4\sigma}. \tag{S11}$$

Let us restrict to $\delta \ge n^{-1/2}$ and bound the LHS. Define $t_j = 2^{-j\delta}$ for $j \le J := \lceil \log n^{1/2}\delta \rceil$. Then

$$\int_{\delta^2}^{\delta} \sqrt{\log N(\partial\mathcal{F}_\delta, \|\cdot\|_{n,2}, t)}\, dt \le \sum_{j=0}^{J} \sqrt{\log N(\partial\mathcal{F}_\delta, \|\cdot\|_{n,2}, t_j)}(t_j - t_{j+1}).$$

For any fixed $Z > 0, \delta \ge n^{-1/2}$, a union bound over J applications of proposition S5 to $(\zeta \leftarrow n^{-10}, \delta, t \leftarrow t_j, Z)$ shows that, with $\mathcal{S}_{tr}$-probability $\ge 1 - n^{-10}\log(n^{1/2}\delta)$,

$$\begin{aligned}\sum_{j=0}^{J} \sqrt{\log N(\partial\mathcal{F}_\delta, \|\cdot\|_{n,2}, t_j)}(t_j - t_{j+1}) &< \sqrt{10\log n}\delta + \sum_{j=0}^{J}(t_j - t_{j+1}) \cdot \Bigg(d_{inv}\log\left(1 + \frac{c'\delta}{t}\right) + \\ &\qquad d_{spu}\log\left(1 + \frac{c'\delta}{e^{mZ^2\delta^2/2}\cdot t}\right) + d_{spu}\log\left(1 + \frac{c'Z\delta}{t}\right)\Bigg)^{1/2} \\ &\le \sqrt{10\log n}\delta + \sum_{j=0}^{J}(t_j - t_{j+1}) \cdot \Bigg(\sqrt{d_{inv}\log\left(1 + \frac{c'\delta}{t}\right)} + \\ &\qquad \sqrt{d_{spu}\log\left(1 + \frac{c'\delta}{e^{MZ^2\delta^2/2}\cdot t}\right)} + \sqrt{d_{spu}\log\left(1 + \frac{c'Z\delta}{t}\right)}\Bigg),\end{aligned} \tag{S12}$$

and that

$$\sum_{j=0}^{J}\sqrt{\log N(\partial\mathcal{F}_\delta,\|\cdot\|_{n,2},t_j)}(t_j-t_{j+1})<\sqrt{10\log n}\delta+\sum_{j=0}^{J}(t_j-t_{j+1})\Bigg(\sqrt{d_{inv}\log\left(1+\frac{c'\delta}{t}\right)}+\sqrt{d_{spu}\log\left(1+\frac{c'\delta}{2^{m/d_{spu}}\cdot t}\right)}\Bigg). \quad \text{(S13)}$$

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

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

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

$$\delta\ge\max\{\delta_{0,n},n^{-1/2}\},\ \sqrt{10\log n}\,\delta+c''(\sqrt{d_{inv}}+\min\{Z+e^{-mZ^2\delta_{0,n}^2/2},2^{-m/d_{spu}}\}\sqrt{d_{spu}})\delta\le\frac{\sqrt{n}\delta^2}{64\sigma^2}$$

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

$$\delta_n^2\le c'''\left(\frac{\log n+d_{inv}}{n}+\sqrt{\frac{d_{spu}\log m}{nm}}\right),$$

while considering the second argument of $\min$ yields

$$\delta_n^2\le c'''\frac{\log n+d_{inv}+2^{-2m/d_{spu}}d_{spu}}{n}.$$

Both bounds hold with the aforementioned $\mathcal{S}_{tr}$-probability. Plugging back to (S10) completes the proof. □

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 $N:=0.3n_*$, `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 $n_e\leftarrow 12000$ for classification and $n_e\leftarrow 18000, d_{inv}\leftarrow 80, d_{spu}\leftarrow 160, m\leftarrow 5$ 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 $\bar\beta^*_{spu}$ with

$$\bar\beta^*_{spu}\sim\mathcal{N}\Big(\frac{2\alpha}{m}\sum_{e\in\mathcal{E}_{tr}}\bar\beta^e_{spu},(1-\alpha^2)\tau_*^2 I\Big),$$

where $\tau_*=1$ for classification and $d_{spu}^{-1/2}$ for regression. For regression, we also introduce environment-specific correlations between the invariant and spurious features, by replacing the generating process of $x^e_{spu,i}$ with

$$x^e_{spu,i}\sim\mathcal{N}(\beta A^e x^e_{inv,i},(1-\beta^2)I),$$

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

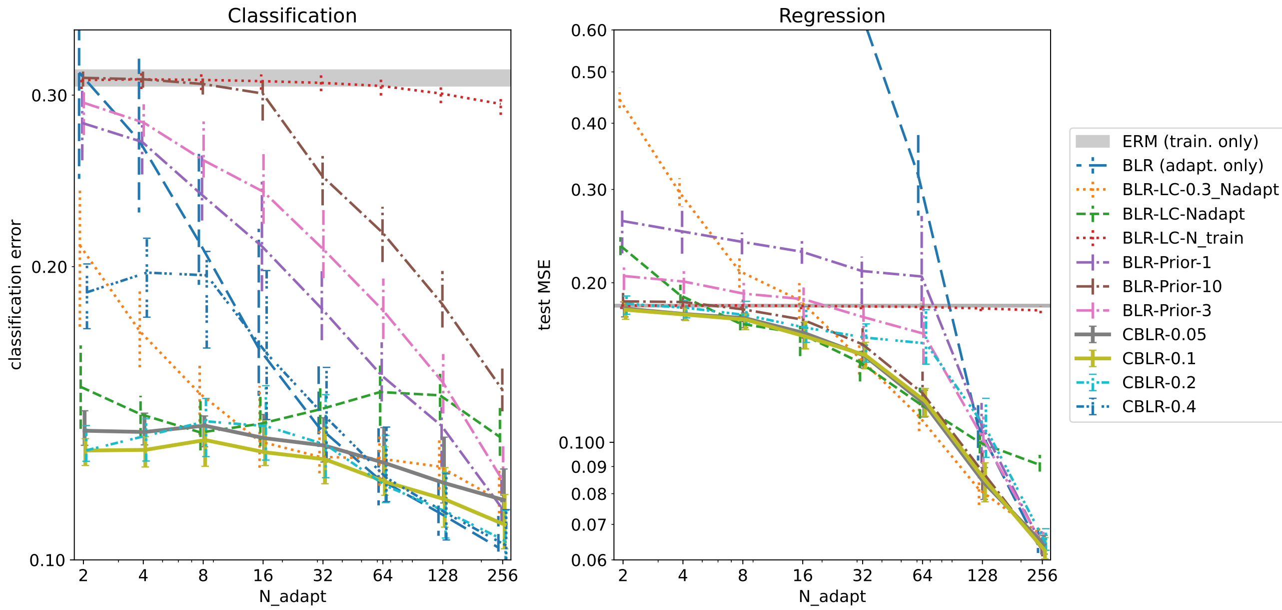

Figure S1: Synthetic experiment: results for all methods in the setting of Figure 1.

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 2×10^4 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

Milton Abramowitz, Irene A Stegun, and Robert H Romer. Handbook of mathematical functions with formulas, graphs, and mathematical tables, 1988.

William H Greene. *Econometric analysis*. Pearson Education India, 2003.

Ishaan Gulrajani and David Lopez-Paz. In Search of Lost Domain Generalization. *arXiv:2007.01434 [cs, stat]*, July 2020. URL `http://arxiv.org/abs/2007.01434`. arXiv: 2007.01434.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

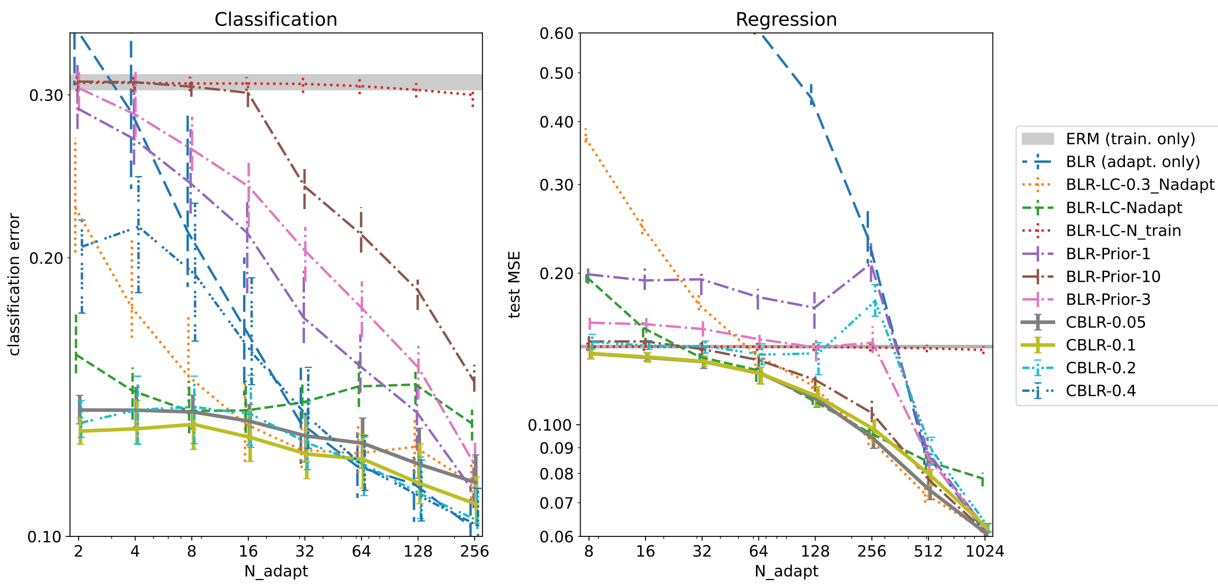

Figure S2: Synthetic experiment: results for all methods at a larger scale.

.

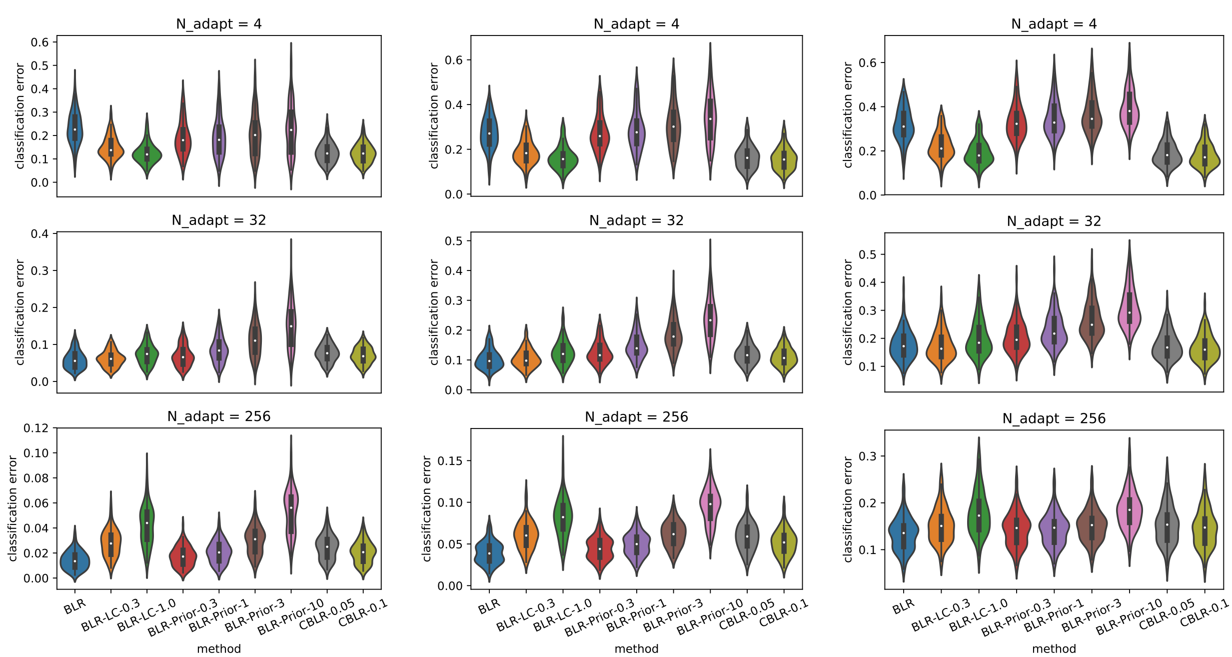

Figure S3: Synthetic classification experiment: violin plot of classification errors, across independently sampled environments, in the setting of Figure 1. From left to right: results for $\alpha \in \{0, 0.5, 1\}$. From top to bottom: results for $n_* \in \{4, 32, 256\}$. Best viewed when zoomed.

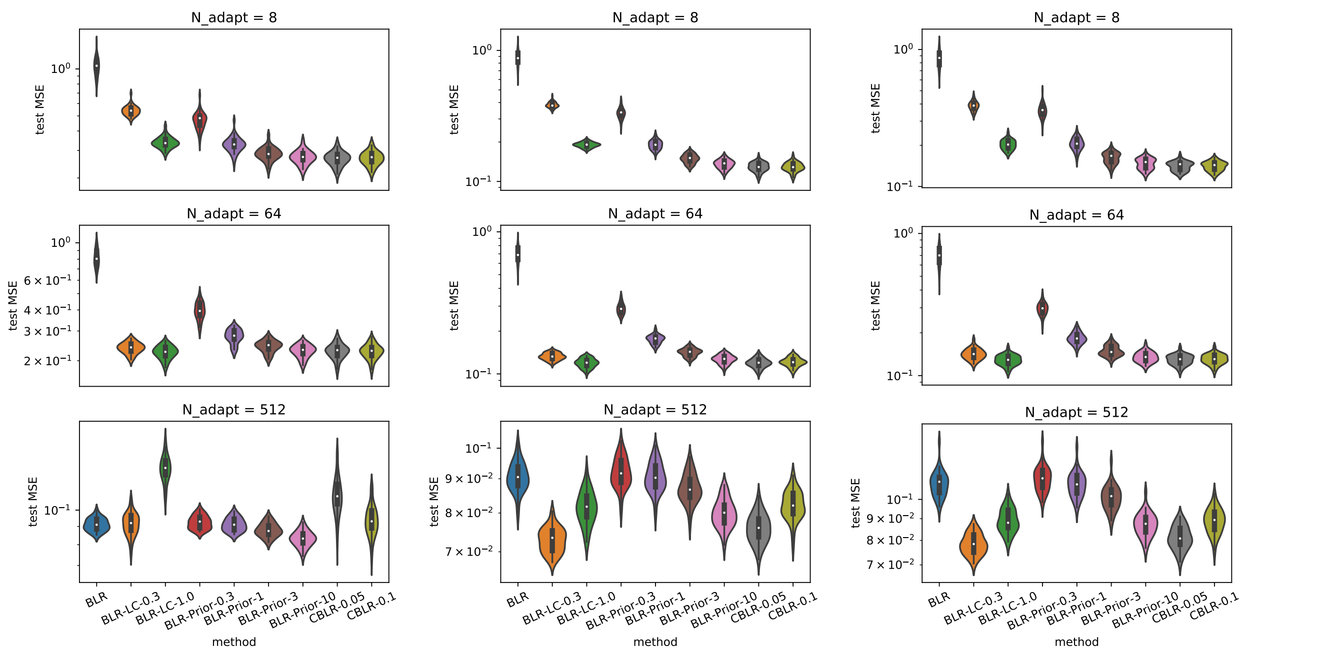

Figure S4: Synthetic regression experiment: violin plot of test MSE across independently sampled environments in the larger-scale setting ($n_e = 18000, m = 5, d = 240$). From left to right: results for $(\alpha, \beta) \in \{(0, 0), (1, 0), (1, 0.5)\}$.

.

Table S1: Colored MNIST: test accuracy for all methods on all domains. We report the 20th percentile / mean / 80th percentile across 20 independent runs.

n_*	Method / e_*	0.1	0.2	0.9
0	ERM	88.5	87.2	71.5
4	CBLR	87.8 / 88.2 / 88.5	86.7 / 87.0 / 87.3	81.5 / 81.9 / 85.6
	CBLR_0.05	87.8 / 87.9 / 88.0	87.1 / 87.3 / 87.7	79.8 / 81.9 / 85.9
	CBLR_0.20	85.0 / 85.7 / 87.2	86.0 / 86.6 / 87.6	79.3 / 76.4 / 86.9
	BLR	75.2 / 80.6 / 88.1	77.5 / 79.3 / 85.8	88.4 / 87.5 / 90.0
	BLR-LC-N_adapt	88.1 / 88.0 / 88.7	86.6 / 87.0 / 87.4	80.5 / 81.8 / 85.6
	BLR-LC-N_train	88.3 / 88.4 / 88.5	87.2 / 87.3 / 87.4	67.9 / 69.3 / 71.1
	DivDis	88.4 / 88.4 / 88.5	87.8 / 87.8 / 87.9	70.7 / 70.7 / 70.9
8	CBLR	87.8 / 88.2 / 88.5	86.5 / 87.0 / 87.4	85.0 / 85.7 / 87.1
	CBLR_0.05	87.8 / 88.1 / 88.6	87.0 / 87.2 / 87.5	83.9 / 85.1 / 86.8
	CBLR_0.20	85.4 / 86.1 / 88.0	85.1 / 86.0 / 87.1	85.3 / 86.7 / 88.9
	BLR	83.0 / 86.1 / 88.4	80.9 / 81.7 / 86.5	89.2 / 89.3 / 90.0
	BLR-LC-N_adapt	87.9 / 88.3 / 88.7	87.0 / 87.1 / 87.5	82.9 / 83.5 / 86.0
	BLR-LC-N_train	88.3 / 88.4 / 88.5	87.2 / 87.3 / 87.4	68.3 / 69.9 / 71.8
	DivDis	88.4 / 88.4 / 88.5	87.8 / 87.8 / 87.8	70.6 / 70.7 / 70.8
16	CBLR	87.6 / 88.2 / 88.7	86.8 / 87.0 / 87.4	87.1 / 87.7 / 88.5
	CBLR_0.05	87.9 / 88.1 / 88.6	86.9 / 87.2 / 87.5	86.4 / 87.0 / 87.9
	CBLR_0.20	85.7 / 86.3 / 88.5	86.0 / 86.5 / 87.3	87.9 / 88.6 / 89.4
	BLR	86.0 / 87.4 / 88.5	85.4 / 85.5 / 87.0	89.1 / 89.4 / 90.0
	BLR-LC-N_adapt	88.3 / 88.5 / 88.7	87.1 / 87.3 / 87.5	83.9 / 84.9 / 86.9
	BLR-LC-N_train	88.2 / 88.4 / 88.6	87.2 / 87.3 / 87.4	70.1 / 71.4 / 73.4
	DivDis	88.4 / 88.4 / 88.5	87.8 / 87.8 / 87.9	70.7 / 70.7 / 70.8
32	CBLR	88.2 / 88.5 / 88.8	86.7 / 87.0 / 87.4	88.6 / 88.8 / 89.1
	CBLR_0.05	88.1 / 88.4 / 88.7	87.0 / 87.2 / 87.5	87.5 / 88.0 / 88.4
	CBLR_0.20	87.2 / 87.7 / 88.5	86.2 / 86.8 / 87.4	89.1 / 89.4 / 89.7
	BLR	87.8 / 88.2 / 88.8	86.3 / 86.8 / 87.3	89.6 / 89.8 / 90.0
	BLR-LC-N_adapt	88.5 / 88.6 / 88.7	87.2 / 87.3 / 87.5	84.8 / 85.6 / 86.4
	BLR-LC-N_train	88.2 / 88.4 / 88.6	87.2 / 87.3 / 87.4	72.7 / 73.6 / 75.4
	DivDis	88.4 / 88.4 / 88.5	87.8 / 87.8 / 87.9	70.7 / 70.7 / 70.8

Table S2: PACS: test accuracy for all methods on all domains. We report the 20th percentile / mean / 80th percentile across 20 independent runs.

n_*	Method / e_*	A	C	P	S
0	ERM	87.8	72.6	96.1	76.3
16	CBLR	87.8 / 88.5 / 89.5	80.8 / 81.7 / 82.7	96.7 / 97.1 / 97.6	77.6 / 78.3 / 79.5
	CBLR_0.05	86.6 / 87.4 / 88.8	79.5 / 80.2 / 81.6	96.1 / 96.7 / 97.3	76.1 / 76.9 / 78.0
	CBLR_0.20	83.9 / 85.4 / 87.0	78.2 / 79.3 / 81.0	95.8 / 96.3 / 96.7	73.8 / 74.9 / 76.4
	BLR	86.8 / 87.7 / 89.2	76.1 / 78.8 / 82.9	95.5 / 96.1 / 97.0	70.1 / 72.5 / 75.5
	BLR-LC-N_adapt	88.3 / 88.7 / 89.5	81.8 / 82.6 / 83.5	97.0 / 97.3 / 97.6	77.8 / 78.7 / 79.9
	BLR-LC-N_train	86.1 / 86.8 / 87.5	79.7 / 80.0 / 80.6	96.7 / 96.9 / 97.0	76.1 / 76.4 / 77.6
	DivDis	85.1 / 85.6 / 86.3	79.3 / 79.3 / 79.7	96.4 / 96.8 / 97.6	77.7 / 78.0 / 79.1
32	CBLR	88.8 / 89.8 / 90.7	82.7 / 83.3 / 84.2	96.7 / 97.1 / 97.3	78.6 / 79.2 / 80.1
	CBLR_0.05	87.5 / 88.5 / 89.5	80.3 / 81.6 / 82.7	96.1 / 96.6 / 97.0	77.2 / 78.3 / 79.5
	CBLR_0.20	86.1 / 87.0 / 88.5	79.5 / 80.8 / 82.1	96.1 / 96.4 / 97.0	75.4 / 76.8 / 78.2
	BLR	88.8 / 89.6 / 90.7	82.1 / 82.9 / 84.4	96.4 / 96.9 / 97.6	76.4 / 77.6 / 79.2
	BLR-LC-N_adapt	89.0 / 89.5 / 90.2	81.6 / 82.5 / 83.8	97.0 / 97.3 / 97.6	76.4 / 77.0 / 79.9
	BLR-LC-N_train	86.1 / 86.9 / 87.5	79.7 / 80.3 / 81.0	96.7 / 97.0 / 97.3	75.9 / 76.4 / 77.5
	DivDis	85.1 / 85.1 / 85.6	79.1 / 79.6 / 81.2	96.4 / 96.8 / 97.3	77.1 / 77.3 / 78.5
64	CBLR	89.7 / 90.8 / 91.7	84.2 / 85.0 / 85.7	97.0 / 97.4 / 97.9	79.9 / 80.5 / 81.4
	CBLR_0.05	89.0 / 89.4 / 90.2	82.3 / 83.3 / 84.2	96.4 / 96.9 / 97.3	78.5 / 79.4 / 80.4
	CBLR_0.20	88.0 / 88.8 / 90.0	82.7 / 83.1 / 83.8	96.4 / 96.8 / 97.3	76.7 / 78.3 / 79.6
	BLR	90.0 / 90.5 / 91.4	84.2 / 85.3 / 86.1	97.0 / 97.3 / 97.9	79.4 / 80.0 / 81.3
	BLR-LC-N_adapt	89.2 / 90.0 / 90.7	82.3 / 82.7 / 84.4	97.0 / 97.2 / 97.6	77.6 / 79.0 / 80.5
	BLR-LC-N_train	87.0 / 87.4 / 88.0	79.3 / 79.9 / 80.6	97.0 / 97.0 / 97.3	75.5 / 76.3 / 77.2
	DivDis	85.6 / 85.9 / 86.8	78.6 / 79.1 / 79.9	96.4 / 96.8 / 97.3	77.2 / 77.6 / 78.6
128	CBLR	90.7 / 91.5 / 92.2	86.3 / 86.9 / 87.8	97.3 / 97.6 / 98.2	81.5 / 82.4 / 83.7
	CBLR_0.05	90.0 / 90.4 / 91.4	84.2 / 85.5 / 87.6	97.3 / 97.3 / 97.6	80.8 / 81.4 / 83.4
	CBLR_0.20	89.5 / 90.6 / 92.4	84.0 / 84.8 / 87.4	96.7 / 96.9 / 97.3	79.2 / 79.7 / 81.3
	BLR	90.5 / 91.1 / 91.7	86.1 / 86.6 / 87.4	96.7 / 97.4 / 97.9	80.1 / 81.2 / 82.5
	BLR-LC-N_adapt	89.2 / 89.9 / 90.7	82.7 / 83.2 / 84.2	97.0 / 97.3 / 97.9	79.1 / 79.8 / 80.6
	BLR-LC-N_train	86.6 / 87.2 / 88.0	79.5 / 80.4 / 81.2	96.7 / 97.0 / 97.3	75.7 / 76.4 / 77.6
	DivDis	84.8 / 85.4 / 86.1	78.8 / 79.7 / 80.6	96.4 / 96.9 / 97.3	76.7 / 77.2 / 78.1
256	CBLR	91.9 / 92.5 / 92.9	86.3 / 86.7 / 88.5	97.6 / 97.9 / 98.2	83.7 / 84.2 / 85.1
	CBLR_0.05	91.9 / 92.3 / 92.7	78.4 / 82.9 / 87.6	97.3 / 97.5 / 97.9	82.9 / 83.5 / 84.2
	CBLR_0.20	91.4 / 91.9 / 92.4	85.7 / 86.9 / 88.7	97.3 / 97.5 / 97.9	81.8 / 82.9 / 83.8
	BLR	91.4 / 91.8 / 92.4	85.9 / 86.5 / 87.4	97.3 / 97.6 / 98.2	80.6 / 81.7 / 82.8
	BLR-LC-N_adapt	89.7 / 90.0 / 90.5	80.8 / 82.5 / 84.2	97.0 / 97.4 / 97.9	77.2 / 78.3 / 80.1
	BLR-LC-N_train	87.5 / 87.9 / 88.5	79.9 / 80.4 / 81.0	96.7 / 96.8 / 97.0	76.9 / 77.2 / 78.5
	DivDis	85.3 / 85.8 / 87.0	80.6 / 80.6 / 80.8	96.7 / 96.9 / 97.3	76.3 / 77.0 / 78.1