Appendix: Training Debiased Subnetworks with Contrastive Weight Pruning

Geon Yeong Park[1] Sangmin Lee[2] Sang Wan Lee[1*] Jong Chul Ye[1,2,3*]
[1]Bio and Brain Engineering, [2]Mathematical Sciences, [3]Kim Jaechul Graduate School of AI
Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea
{pky3436, leeleesang, sangwan, jong.ye}@kaist.ac.kr

The supplementary material is organized as follows. We first present the proof for Theorem 1 and 2. In section 2, we extend the presented theoretical example in the main paper to illustrate the risks of geometrical misalignment of embeddings arising from strong spurious correlations. Additional results are reported in section 3. Optimization setting, hyperparameter configuration, and other experimental details are provided in section 4.

1. Proofs

In this section, we present the detailed proofs for Theorems 1 and 2 explained in the main paper, followed by an illustration about the dynamics of weight ratio $\alpha_i(t) = \tilde{w}_{sp,i}(t)/\tilde{w}_{inv}(t)$.

1.1. Proof of Theorem 1

Theorem 1. *(Training and test bound) Assume that $p^e > 1/2$ in the biased training environment $e \in \mathcal{E}_{train}$. Define $\tilde{\boldsymbol{w}}(t)$ as weights pretrained for a finite time $t < T$. Then the upper bound of the error of training environment w.r.t. pruning parameters $\boldsymbol{\pi}$ is given as:*

$$\ell^e(\boldsymbol{\pi}) \leq 2\exp\Bigg(-\frac{2\big(\pi_{inv} + (2p^e-1)\sum_{i=1}^{D}\alpha_i(t)\pi_{sp,i}\big)^2}{4\sum_{i=1}^{D}\alpha_i(t)^2 + 1}\Bigg), \tag{1}$$

where the weight ratio $\alpha_i(t) = \tilde{w}_{sp,i}(t)/\tilde{w}_{inv}(t)$ is bounded below some positive constant. Given a test environment $e \in \mathcal{E}_{test}$ with $p^e = \frac{1}{2}$, the upper bound of the error of test environment w.r.t. $\boldsymbol{\pi}$ is given as:

$$\ell^e(\boldsymbol{\pi}) \leq 2\exp\Bigg(-\frac{2\pi_{inv}^2}{4\sum_{i=1}^{D}\alpha_i(t)^2 + 1}\Bigg), \tag{2}$$

which implies that there is an unavoidable gap between training bound and test bound.

Proof. We omit time t in $\tilde{\boldsymbol{w}}(t)$ and $\alpha_i(t)$ for notational simplicity throughout the proof of Theorem 1 and 2.

We recall the loss function defined in the main paper for convenience.

$$\begin{aligned}\ell^e(\boldsymbol{\pi}) &= \frac{1}{2}\mathbb{E}_{\boldsymbol{X}^e,Y^e,\boldsymbol{m}}[1 - Y^e\hat{Y}^e] \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{X}^e,Y^e,\boldsymbol{m}}\left[1 - Y^e \cdot \operatorname{sgn}\left(\tilde{\boldsymbol{w}}^T(\boldsymbol{X}^e \odot \boldsymbol{m})\right)\right],\end{aligned} \tag{3}$$

where $\hat{Y}^e$ is the prediction of binary classifier, $\tilde{\boldsymbol{w}}$ is the pretrained weight vector, $\operatorname{sgn}(\cdot)$ represents the sign function, and $\odot$ represents element-wise product.

The prediction from the classifier $\hat{Y}^e$ is defined as

$$\begin{aligned}\hat{Y}^e &= \operatorname{sgn}\left(\tilde{\boldsymbol{w}}^T(\boldsymbol{X}^e \odot \boldsymbol{m})\right) \\ &= \operatorname{sgn}\left(\mathcal{O}^e\right),\end{aligned} \tag{4}$$

where

$$\mathcal{O}^e := \tilde{w}_{inv} m_{inv} Z^e_{inv} + \sum_{i=1}^{D} \tilde{w}_{sp,i} m_{sp,i} Z^e_{sp,i}. \tag{5}$$

Assume that Y^e is uniformly distributed binary random variable. Then,

$$\mathbb{E}_{\boldsymbol{X}^e, Y^e, \boldsymbol{m}}[Y^e \hat{Y}^e] = \frac{1}{2}\mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\Big[\hat{Y}^e | Y^e = 1\Big] - \frac{1}{2}\mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\Big[\hat{Y}^e | Y^e = -1\Big], \tag{6}$$

where

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\Big[\hat{Y}^e | Y^e = 1\Big] &= \mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\Big[\operatorname{sgn}\Big(\mathcal{O}^e\Big) \,\Big|\, Y^e = 1\Big] \\
&= P\big(\mathcal{O}^e > 0 \,\big|\, Y^e = 1\big) - P\big(\mathcal{O}^e < 0 \,\big|\, Y^e = 1\big) \\
&= 1 - 2P\big(\mathcal{O}^e < 0 \,\big|\, Y^e = 1\big),
\end{aligned} \tag{7}$$

and

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\Big[\hat{Y}^e | Y^e = -1\Big] &= P\big(\mathcal{O}^e > 0 \,\big|\, Y^e = -1\big) - P\big(\mathcal{O}^e < 0 \,\big|\, Y^e = -1\big) \\
&= -\mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\Big[\hat{Y}^e | Y^e = 1\Big],
\end{aligned} \tag{8}$$

where we use $P\big(\mathcal{O}^e < 0 \,\big|\, Y^e = 1\big) = P\big(\mathcal{O}^e > 0 \,\big|\, Y^e = -1\big)$ and $P\big(\mathcal{O}^e > 0 \,\big|\, Y^e = 1\big) = P\big(\mathcal{O}^e < 0 \,\big|\, Y^e = -1\big)$ thanks to the symmetry. Therefore, we have

$$\begin{aligned}
\ell^e(\boldsymbol{\pi}) &= \frac{1}{2}\mathbb{E}_{\boldsymbol{X}^e, Y^e, \boldsymbol{m}}[1 - Y^e \hat{Y}^e] \\
&= \frac{1}{2} - \frac{1}{2}\mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\Big[\hat{Y}^e | Y^e = 1\Big] \\
&= P\big(\mathcal{O}^e < 0 \,\big|\, Y^e = 1\big).
\end{aligned} \tag{9}$$

In order to derive a concentration inequality of $\ell^e(\boldsymbol{\pi})$, we compute a conditional expectation as follows:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\big[\mathcal{O}^e \,\big|\, Y^e = 1\big] &= \mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\Big[\tilde{w}_{inv} m_{inv} Z^e_{inv} + \sum_{i=1}^{D} \tilde{w}_{sp,i} m_{sp,i} Z^e_{sp,i} \,\Big|\, Y^e = 1\Big] \\
&= \mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\Big[\tilde{w}_{inv} m_{inv} + \sum_{i=1}^{D} \tilde{w}_{sp,i} m_{sp,i} Z^e_{sp,i} \,\Big|\, Y^e = 1\Big] \\
&= \tilde{w}_{inv}\pi_{inv} + \mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\Big[\sum_{i=1}^{D} \tilde{w}_{sp,i} m_{sp,i} Z^e_{sp,i} \,\Big|\, Y^e = 1\Big] \\
&= \tilde{w}_{inv}\pi_{inv} + \sum_{i=1}^{D} (2p^e - 1)\tilde{w}_{sp,i}\pi_{sp,i},
\end{aligned} \tag{10}$$

where the last equality follows from the independence of $Z_{sp,\cdot}$ and $m_{sp,\cdot}$ as assumed in the main paper. Then,

$$\begin{aligned}
P\big(\mathcal{O}^e < 0 \,\big|\, Y^e = 1\big) &= P\Big(\mathcal{O}^e - \mathbb{E}_{\boldsymbol{X}^e, \mathbf{m}}\big[\mathcal{O}^e\big] < -\mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\big[\mathcal{O}^e\big] \,\big|\, Y^e = 1\Big) \\
&\le P\Big(\,\Big|\, \mathcal{O}^e - \mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\big[\mathcal{O}^e\big] \,\Big|\, > \mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\big[\mathcal{O}^e\big] \,\big|\, Y^e = 1\Big) \\
&\le 2\exp\Big(-\frac{2\mathbb{E}_{\boldsymbol{X}^e, \boldsymbol{m}}\big[\mathcal{O}^e \,\big|\, Y^e = 1\big]^2}{\tilde{w}^2_{inv} + \sum_{i=1}^{D} 4\tilde{w}^2_{sp,i}}\Big) \\
&\le 2\exp\Big(-\frac{2\big(\tilde{w}_{inv}\pi_{inv} + \sum_{i=1}^{D}(2p^e - 1)\tilde{w}_{sp,i}\pi_{sp,i}\big)^2}{\tilde{w}^2_{inv} + \sum_{i=1}^{D} 4\tilde{w}^2_{sp,i}}\Big) \\
&\le 2\exp\Big(-\frac{2\big(\pi_{inv} + \sum_{i=1}^{D}(2p^e - 1)\alpha_i\pi_{sp,i}\big)^2}{1 + \sum_{i=1}^{D} 4\alpha^2_i}\Big),
\end{aligned} \tag{11}$$

where the second inequality is obtained using Hoeffding's inequality, third inequality is from (10), and last inequality is obtained by dividing both denominator and numerator with $\tilde{w}_{inv}^2$. We use the definition of weight ratio $\alpha_i = \tilde{w}_{sp,i}/\tilde{w}_{inv}$. For the second inequality, we use that $\tilde{w}_{inv} m_{inv} Z_{inv}^e \in \{0, \tilde{w}_{inv}\}$ and $\tilde{w}_{sp,i} m_{sp,i} Z_{sp,i}^e \in \{-\tilde{w}_{sp,i}, 0, \tilde{w}_{sp,i}\}\ \forall i$ in (5) to obtain the denominator.

Finally, the proof for the positivity of $\alpha_i(t)$ comes from Proposition 1 in section 1.3 in this appendix. This concludes the proof. □

1.2. Proof of Theorem 2

Theorem 2. *(Training bound with the mixture distribution) Assume that the defined mixture distribution P_{mix}^{η} is biased, i.e., for all $i \in \{1, \dots, D\}$,*

$$P_{mix}^{\eta}(Z_{sp,i}^{\eta} = -y \mid Y^e = y) \leq P_{mix}^{\eta}(Z_{sp,i}^{\eta} = y \mid Y^{\eta} = y). \tag{12}$$

Then, ϕ satisfies $0 \leq \phi \leq 1 - \frac{1}{2p^{\eta}}$. Then the upper bound of the error of training environment η w.r.t. the pruning parameters is given by

$$\ell^{\eta}(\boldsymbol{\pi}) \leq 2 \exp\left(- \frac{2(\pi_{inv} + (2p^{\eta}(1-\phi) - 1)\sum_{i=1}^{D} \alpha_i(t)\pi_{sp,i})^2}{4\sum_{i=1}^{D} \alpha_i(t)^2 + 1} \right). \tag{13}$$

Furthermore, when $\phi = 1 - \frac{1}{2p^{\eta}}$, the mixture distribution is perfectly debiased, and we have

$$\ell^{\eta}(\boldsymbol{\pi}) \leq 2 \exp\Big(- \frac{2\pi_{inv}^2}{4\sum_{i=1}^{D} \alpha_i(t)^2 + 1} \Big), \tag{14}$$

which is equivalent to the test bound in (2).

Proof. Recall that $Z_{sp,i}^{\eta}$ follows the mixture distribution P_{mix}^{η}:

$$P_{mix}^{\eta}(Z_{sp,i}^{\eta} \mid Y^{\eta} = y) = \phi P_{debias}^{\eta}(Z_{sp,i}^{\eta} \mid Y^{\eta} = y) + (1-\phi) P_{bias}^{\eta}(Z_{sp,i}^{\eta} \mid Y^{\eta} = y), \tag{15}$$

where

$$P_{debias}^{\eta}(Z_{sp,i}^{\eta} \mid Y^{\eta} = y) = \begin{cases} 1, & \text{if } Z_{sp,i}^{\eta} = -y \\ 0, & \text{if } Z_{sp,i}^{\eta} = y \end{cases} \tag{16}$$

is a debiasing distribution to weaken the correlation between Y^{η} and $Z_{sp,i}^{\eta}$ by setting the value of $Z_{sp,i}^{\eta}$ as $-Y^{\eta}$, and

$$P_{bias}^{\eta}(Z_{sp,i}^{\eta} \mid Y^{\eta} = y) = \begin{cases} p^{\eta}, & \text{if } Z_{sp,i}^{\eta} = y \\ 1 - p^{\eta}, & \text{if } Z_{sp,i}^{\eta} = -y. \end{cases} \tag{17}$$

Then, with definition in (16) and (17),

$$\begin{aligned} P_{mix}(Z_{sp,i}^{\eta} = -y | Y^{\eta} = y) &= \phi + (1-\phi)(1-p^{\eta}) \\ P_{mix}(Z_{sp,i}^{\eta} = y | Y^{\eta} = y) &= (1-\phi)p^{\eta}, \end{aligned} \tag{18}$$

for $y \in \{-1, 1\}$. Then, based on the assumption, $\phi + (1-\phi)(1-p^{\eta}) \leq (1-\phi)p^{\eta}$, which gives $\phi \leq 1 - \frac{1}{2p^{\eta}}$. Specifically, if $\phi = 1 - \frac{1}{2p^{\eta}}$, it turns out that $P_{mix}(Z_{sp,i}^{\eta} = -y | Y^{\eta} = y) = P_{mix}(Z_{sp,i}^{\eta} = y | Y^{\eta} = y) = \frac{1}{2}$, which implies that spurious features turns out to be random and the mixture distribution becomes perfectly debiased. If $\phi = 0$, the mixture distribution boils down into a biased distribution as similarly defined in the environment $e \in \mathcal{E}_{train}$.

The prediction from the classifier $\mathcal{O}^{\eta}$ is defined as similar to $\mathcal{O}^e$ in (5). Then in order to derive a concentration inequality of $\ell^{\eta}(\boldsymbol{\pi})$, we derive a conditional expectation of $\mathcal{O}^{\eta}$ as done in (10):

$$\begin{aligned} \mathbb{E}_{\boldsymbol{X}^{\eta}, \boldsymbol{m}}\left[\mathcal{O}^{\eta} \,\middle|\, Y^{\eta} = 1\right] &= \mathbb{E}_{\boldsymbol{X}^{\eta}, \boldsymbol{m}}\left[\tilde{w}_{inv} m_{inv} Z_{inv}^{\eta} + \sum_{i=1}^{D} \tilde{w}_{sp,i} m_{sp,i} Z_{sp,i}^{\eta} \,\middle|\, Y^{\eta} = 1\right] \\ &= \mathbb{E}_{\boldsymbol{X}^{\eta}, \boldsymbol{m}}\left[\tilde{w}_{inv} m_{inv} + \sum_{i=1}^{D} \tilde{w}_{sp,i} m_{sp,i} Z_{sp,i}^{\eta} \,\middle|\, Y^{\eta} = 1\right]. \end{aligned} \tag{19}$$

Then, with the definition in (15), the second term in the above conditional expectation of (19) is defined as follows:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}^\eta,\boldsymbol{m}}\Big[\sum_{i=1}^{D}\tilde{w}_{sp,i}m_{sp,i}Z^\eta_{sp,i}\mid Y^\eta=1\Big]\\
&\quad=\sum_{i=1}^{D}\tilde{w}_{sp,i}\pi_{sp,i}\Big(\phi\mathbb{E}_{debias}[Z^\eta_{sp,i}\mid Y^\eta=1]+(1-\phi)\mathbb{E}_{bias}[Z^\eta_{sp,i}\mid Y^\eta=1]\Big)\\
&\quad=\sum_{i=1}^{D}\tilde{w}_{sp,i}\pi_{sp,i}\Big(\phi\cdot(-1)+(1-\phi)(2p^\eta-1)\Big)\\
&\quad=\sum_{i=1}^{D}\tilde{w}_{sp,i}\pi_{sp,i}\big(2p^\eta(1-\phi)-1\big),
\end{aligned}\tag{20}$$

where $\mathbb{E}_{debias}$ and $\mathbb{E}_{bias}$ in the first equality denote the conditional expectation with respect to distribution P^η_{debias} and P^η_{bias} in (16) and (17), respectively. Plugging (20) into (19), we get

$$\mathbb{E}_{\boldsymbol{X}^\eta,\mathbf{m}}\big[\mathcal{O}^\eta\;\big|\;Y^\eta=1\big]=\tilde{w}_{inv}\pi_{inv}+\sum_{i=1}^{D}\big(2p^\eta(1-\phi)-1\big)\tilde{w}_{sp,i}\pi_{sp,i}.\tag{21}$$

Then we can derive the upper bound of $\ell^\eta(\boldsymbol{\pi})=P(\mathcal{O}^\eta<0\mid Y^\eta=1)$ similarly to (11):

$$\begin{aligned}
P\big(\mathcal{O}^\eta<0\,\big|\,Y^\eta=1\big)&\le P\Big(\Big|\,\mathcal{O}^\eta-\mathbb{E}_{\boldsymbol{X}^\eta,\boldsymbol{m}}\big[\mathcal{O}^\eta\big]\,\Big|>\mathbb{E}_{\boldsymbol{X}^\eta,\boldsymbol{m}}\big[\mathcal{O}^\eta\big]\mid Y^\eta=1\Big)\\
&\le 2\exp\left(-\frac{2\mathbb{E}_{\boldsymbol{X}^\eta,\boldsymbol{m}}\big[\mathcal{O}^\eta\mid Y^\eta=1\big]^2}{\tilde{w}^2_{inv}+4\sum_{i=1}^{D}\tilde{w}^2_{sp,i}}\right)\\
&\le 2\exp\left(-\frac{2\big(\tilde{w}_{inv}\pi_{inv}+\sum_{i=1}^{D}\big(2p^\eta(1-\phi)-1\big)\tilde{w}_{sp,i}\pi_{sp,i}\big)^2}{\tilde{w}^2_{inv}+4\sum_{i=1}^{D}\tilde{w}^2_{sp,i}}\right)\\
&\le 2\exp\left(-\frac{2\big(\pi_{inv}+\sum_{i=1}^{D}(2p^\eta(1-\phi)-1)\alpha_i\pi_{sp,i}\big)^2}{1+\sum_{i=1}^{D}4\alpha_i^2}\right),
\end{aligned}\tag{22}$$

where the first inequality is obtained by Hoeffding's inequality, and second inequality is from (21). The denominator is obtained as same as in (11), since $\tilde{w}_{inv}m_{inv}Z^\eta_{inv}\in\{0,\tilde{w}_{inv}\}$ and $\tilde{w}_{sp,i}m_{sp,i}Z^\eta_{sp,i}\in\{-\tilde{w}_{sp,i},0,\tilde{w}_{sp,i}\}\;\forall i$ as-is. If we plug-in the upper bound value of $\phi=1-\frac{1}{2p^\eta}$ obtained from (18) into (22), it boils down into the test bound in (2). □

1.3. Dynamics of the weight ratio

We omit an index of environment e in the proposition below for notational simplicity.

Proposition 1. *Consider a binary classification problem of linear classifier $f_{\boldsymbol{w}}$ under exponential loss. Let $(\boldsymbol{X},Y)\sim P$, where each input random variable $\boldsymbol{X}$ and the corresponding label Y is generated by*

$$\boldsymbol{X}=\begin{pmatrix}Z_{inv}\\ \boldsymbol{Z}_{sp}\end{pmatrix},Y=Z_{inv},$$

where $\boldsymbol{Z}_{sp}=(2\boldsymbol{z}-1)Z_{inv}$ for a random variable $\boldsymbol{z}\in\{0,1\}^D$ which is chosen from multivariate Bernoulli distribution ($z_i\sim Bern(p)$) with $p>\frac{1}{2}$, i.e., p denotes p^e in the main paper. Let $\boldsymbol{w}=\begin{pmatrix}w_{inv}\\ \boldsymbol{w}_{sp}\end{pmatrix}\in\mathbb{R}^{D+1}$ be the weight of the linear classifier $f_{\boldsymbol{w}}(\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{x}$. Assume that $0<w_{inv}(0)$, i.e., w_{inv} is initialized with a positive value, and $0<w_{sp,i}(0)<\frac{1}{2}\log\frac{p}{1-p}$. Then, after sufficient time of training, w_{inv} diverges to $+\infty$ and $w_{sp,i}$ converges to $\frac{1}{2}\log\frac{p}{1-p}$, which means $\alpha_i:=\frac{w_{sp,i}}{w_{inv}}$ converges to 0 for all $i\in\{1,2,\cdots,D\}$. More precisely,

$$\log\Big(e^{w_{inv}(0)}+[4p(1-p)]^{\frac{D}{2}}t\Big)\le w_{inv}(t)\le\log\left(e^{w_{inv}(0)}+t\prod_{i=1}^{D}\Big(pe^{-w_{sp,i}(0)}+\sqrt{p(1-p)}\Big)\right).$$

However, for a fixed $t<T$, each α_i is positive and its lower bound converges to some positive value.

Proof. In this proof, $w_{inv}(t)$ denotes the invariant weight at time t, while we often omit the time t and interchangeably use w_{inv} for notational simplicity, and likewise for $w_{sp,i}(t)$.

Note that the network output is given by

$$\begin{aligned} f_{\boldsymbol{w}}(\boldsymbol{x}) &= \boldsymbol{w}^T \boldsymbol{x} \\ &= Z_{inv} w_{inv} + \boldsymbol{Z}_{sp}^T \boldsymbol{w}_{sp} \\ &= Z_{inv} w_{inv} + \sum_{i=1}^{D} Z_{sp,i} w_{sp,i}. \end{aligned}$$

The exponential loss is defined by

$$\begin{aligned} L(\boldsymbol{w}) &= \mathbb{E}_{(\boldsymbol{X},Y)}[e^{-f_{\boldsymbol{w}}(\boldsymbol{X})Y}] \\ &= \mathbb{E}_{\boldsymbol{z}}\big[\exp\big(-(Z_{inv} w_{inv} + \sum_{i=1}^{D} Z_{sp,i} w_{sp,i}) Z_{inv}\big)\big] \\ &= \mathbb{E}_{\boldsymbol{z}}\big[\exp(-w_{inv} - (2z_1 - 1) w_{sp,1} - \cdots - (2z_D - 1) w_{sp,D})\big] \\ &= e^{-w_{inv}} \prod_{i=1}^{D} \mathbb{E}_{\boldsymbol{z}}[e^{-(2z_i - 1) w_{sp,i}}] \\ &= e^{-w_{inv}} \prod_{i=1}^{D} (p e^{-w_{sp,i}} + (1-p) e^{w_{sp,i}}). \end{aligned}$$

Then, thanks to symmetry of $\boldsymbol{w}_{sp}$, it is enough to consider $\alpha := \frac{w_{sp,1}}{w_{inv}}$. We first compute the gradient:

$$\begin{aligned} \frac{\partial L}{\partial w_{inv}} &= -e^{-w_{inv}} \prod_{i=1}^{D} (p e^{-w_{sp,i}} + (1-p) e^{w_{sp,i}}) \\ \frac{\partial L}{\partial w_{sp,1}} &= -e^{-w_{inv}} (p e^{-w_{sp,1}} - (1-p) e^{w_{sp,1}}) \prod_{i=2}^{D} (p e^{-w_{sp,i}} + (1-p) e^{w_{sp,i}}). \end{aligned}$$

Since $\frac{d}{dt} w_{inv} = -\frac{\partial L}{\partial w_{inv}}$, the dynamics is given by the following differnetial equations.

$$\begin{aligned} \frac{d}{dt} w_{inv} &= e^{-w_{inv}} \prod_{i=1}^{D} (p e^{-w_{sp,i}} + (1-p) e^{w_{sp,i}}) \\ \frac{d}{dt} w_{sp,1} &= e^{-w_{inv}} (p e^{-w_{sp,1}} - (1-p) e^{w_{sp,1}}) \prod_{i=2}^{D} (p e^{-w_{sp,i}} + (1-p) e^{w_{sp,i}}). \end{aligned}$$

First we show that $w_{inv}(t)$ diverges to $+\infty$ as t goes ∞. We show this by computing its lower bound.

$$\begin{aligned} \frac{d}{dt} w_{inv} &= e^{-w_{inv}} \prod_{i=1}^{D} (p e^{-w_{sp,i}} + (1-p) e^{w_{sp,i}}) \\ &\geq e^{-w_{inv}} \prod_{i=1}^{D} (2\sqrt{p(1-p)}) \\ &= e^{-w_{inv}} [4p(1-p)]^{\frac{D}{2}}, \end{aligned}$$

where the inequality is obtained by AM-GM inequality. This implies $e^{w_{inv}} dw_{inv} \geq [4p(1-p)]^{\frac{D}{2}} dt$. Integrating both sides from 0 to t, we get

$$e^{w_{inv}(t)} - e^{w_{inv}(0)} \geq [4p(1-p)]^{\frac{D}{2}} t$$

or

$$w_{inv}(t) \geq \log\left(e^{w_{inv}(0)} + [4p(1-p)]^{\frac{D}{2}}t\right), \tag{23}$$

which shows that $w_{inv}(t)$ diverges to $+\infty$ as $t \to \infty$. Note also that w_{inv} strictly increases since $\frac{d}{dt}w_{inv} > 0$.

For $w_{sp,i}$, $\frac{d}{dt}w_{sp,i} = 0$ implies $w_{sp,i}$ converges to $w^*_{sp,i}$ such that

$$pe^{-w^*_{sp,i}} - (1-p)e^{w^*_{sp,i}} = 0,$$

namely, $w^*_{sp,i} = \frac{1}{2}\log\frac{p}{1-p}$.

As similar to w_{inv}, $w_{sp,1}$ strictly increases if and only if $w_{sp,1} < \frac{1}{2}\log\frac{p}{1-p}$. Based on the assumptions that $0 < w_{sp,i}(0) < \frac{1}{2}\log\frac{p}{1-p}$, we conclude that $w_{sp,1}$ monotonically converges to $\frac{1}{2}\log\frac{p}{1-p}$. As p goes to 1, $\frac{1}{2}\log\frac{p}{1-p}$ is sufficiently large and we can assume $w_{sp,i}(0) < \frac{1}{2}\log\frac{p}{1-p}$.

Now, we fix $0 < t < T$ for given T and compute an upper bound of w_{inv}. Using $w_{sp,i}(t) < \frac{1}{2}\log\frac{p}{1-p}$, we get

$$\begin{aligned}
\frac{d}{dt}w_{inv} &= e^{-w_{inv}}\prod_{i=1}^{D}(pe^{-w_{sp,i}} + (1-p)e^{w_{sp,i}}) \\
&< e^{-w_{inv}}\prod_{i=1}^{D}\left(pe^{-w_{sp,i}(0)} + (1-p)\sqrt{\frac{p}{1-p}}\right) \\
&= e^{-w_{inv}}\prod_{i=1}^{D}\left(pe^{-w_{sp,i}(0)} + \sqrt{p(1-p)}\right)
\end{aligned}$$

which implies

$$e^{w_{inv}}dw_{inv} < \prod_{i=1}^{D}\left(pe^{-w_{sp,i}(0)} + \sqrt{p(1-p)}\right)dt.$$

Integrating both sides from 0 to t, we get

$$w_{inv}(t) < \log\left(e^{w_{inv}(0)} + \prod_{i=1}^{D}\left(pe^{-w_{sp,i}(0)} + \sqrt{p(1-p)}\right)t\right). \tag{24}$$

Similarly, we compute a lower bound of $w_{sp,1}$ on $0 < t < T$. Before we start, note that $w_{inv}(t) < w_{inv}(T) =: M$ from monotonicity.

$$\begin{aligned}
\frac{d}{dt}w_{sp,1} &= e^{-w_{inv}}(pe^{-w_{sp,1}} - (1-p)e^{w_{sp,1}})\prod_{i=2}^{D}(pe^{-w_{sp,i}} + (1-p)e^{w_{sp,i}}) \\
&> e^{-M}(pe^{-w_{sp,1}} - (1-p)e^{w_{sp,1}})\prod_{i=2}^{D}(2\sqrt{p(1-p)}) \\
&= e^{-M}[4p(1-p)]^{\frac{D-1}{2}}(pe^{-w_{sp,1}} - (1-p)e^{w_{sp,1}})
\end{aligned}$$

induces

$$\frac{1}{pe^{-w_{sp,1}} - (1-p)e^{w_{sp,1}}}dw_{sp,1} > e^{-M}[4p(1-p)]^{\frac{D-1}{2}}dt.$$

Integrating both sides from 0 to $t < T$, we get

$$\left[\frac{1}{\sqrt{p(1-p)}}\tanh^{-1}\left(\sqrt{\frac{1-p}{p}}e^{w_{sp,1}}\right)\right]_0^t > e^{-M}[4p(1-p)]^{\frac{D-1}{2}}t$$

or

$$w_{sp,1}(t) > \frac{1}{2}\log\frac{p}{1-p} + \log\ \tanh\left(\tanh^{-1}(\sqrt{\frac{1-p}{p}}e^{w_{sp,1}(0)}) + e^{-M}2^{D-1}[p(1-p)]^{\frac{D}{2}}t\right). \tag{25}$$

Combining (24) and (25), we conclude that

$$\alpha_p(t) = \frac{w_{sp,1}(t)}{w_{inv}(t)} \tag{26}$$

$$> \frac{\frac{1}{2}\log\frac{p}{1-p} + \log\ \tanh\left(\tanh^{-1}(\sqrt{\frac{1-p}{p}}e^{w_{sp,1}(0)}) + e^{-M}2^{D-1}[p(1-p)]^{\frac{D}{2}}t\right)}{\log\left(e^{w_{inv}(0)} + t\prod_{i=1}^{D}\left(pe^{-w_{sp,i}(0)} + \sqrt{p(1-p)}\right)\right)} \tag{27}$$

for $0 < t < T$. Note that $\alpha_p(t)$ is positive in $0 < t < T$, since both $w_{sp,1}(t)$ and $w_{inv}(t)$ is monotonically increasing in $0 < t < T$, and $0 < w_{sp,1}(0), w_{inv}(0)$ by assumptions.

The numerator becomes

$$\begin{aligned}
&\frac{1}{2}\log\frac{p}{1-p} + \log\ \tanh\left(\tanh^{-1}(\sqrt{\frac{1-p}{p}}e^{w_{sp,1}(0)}) + e^{-M}2^{D-1}[p(1-p)]^{\frac{D}{2}}t\right)\\
&= \log\left[\sqrt{\frac{p}{1-p}}\tanh\left(\tanh^{-1}(\sqrt{\frac{1-p}{p}}e^{w_{sp,1}(0)}) + e^{-M}2^{D-1}[p(1-p)]^{\frac{D}{2}}t\right)\right]\\
&= \log\left[\sqrt{\frac{p}{1-p}}\left(\sqrt{\frac{1-p}{p}}e^{w_{sp,1}(0)} + e^{-M}2^{D-1}[p(1-p)]^{\frac{D}{2}}t\ \operatorname{sech}^2 c\right)\right]
\end{aligned}$$

for some c such that

$$\tanh^{-1}(\sqrt{\frac{1-p}{p}}e^{w_{sp,1}(0)}) < c < \tanh^{-1}(\sqrt{\frac{1-p}{p}}e^{w_{sp,1}(0)}) + e^{-M}2^{D-1}[p(1-p)]^{\frac{D}{2}}t.$$

We use $f(x+y) = f(x) + yf'(c)$ by the Mean Value Theorem (MVT) at the last line.

Notably, if we take a limit $p \to 1$, the numerator becomes

$$\lim_{p\to 1}\log\left[e^{w_{sp,1}(0)} + e^{-M}2^{D-1}p^{\frac{D+1}{2}}(1-p)^{\frac{D-1}{2}}t\ \operatorname{sech}^2 c\right] = w_{sp,1}(0).$$

Similarly, the denominator becomes

$$\begin{aligned}
&\lim_{p\to 1}\log\left(e^{w_{inv}(0)} + t\prod_{i=1}^{D}\left(pe^{-w_{sp,i}(0)} + \sqrt{p(1-p)}\right)\right)\\
&= \log\left(e^{w_{inv}(0)} + t\prod_{i=1}^{D}e^{-w_{sp,i}(0)}\right)\\
&= \log\left(e^{w_{inv}(0)} + t\exp\left(-\sum_{i=1}^{D}w_{sp,i}(0)\right)\right)
\end{aligned}$$

Therefore, for a fixed $0 < t < T$, we conclude that

$$\begin{aligned}
\lim_{p\to 1}\alpha_p(t) &= \lim_{p\to 1}\frac{w_{sp,1}(t)}{w_{inv}(t)}\\
&\geq \frac{w_{sp,1}(0)}{\log\left(e^{w_{inv}(0)} + t\exp\left(-\sum_{i=1}^{D}w_{sp,i}(0)\right)\right)}\\
&> \frac{w_{sp,1}(0)}{\log(e^{w_{inv}(0)} + T\exp\left(-\sum_{i=1}^{D}w_{sp,i}(0)\right))}\\
&\geq \frac{w_{sp,1}(0)}{\log T + \frac{1}{T}\exp\left(w_{inv}(0) + \sum_{i=1}^{D}w_{sp,i}(0)\right) - \sum_{i=1}^{D}w_{sp,i}(0)}
\end{aligned} \tag{28}$$

where we use the inequality $\log(x+y) \leq \log x + \frac{y}{x}$ in the last line. □

The key insights from Proposition 1 can be summarized as follows:
(1) Weight ratio $\alpha_i(t)$ converges to 0 as $t \to \infty$.
(2) However, for a fixed $t < T$, $\alpha_i(t) > 0$.
(3) When $t < T$ and $p \to 1$, i.e., the environment is almost perfectly biased, the convergence rate of (1) is remarkably slow as in (28). In other words, there exists $c > 0$ such that $\frac{c}{\log t} < \alpha_p(t)$ over $0 < t < T$ if p is sufficiently close to 1.

This results afford us intriguing perspective on the fundamental factors behind the biased classifiers. If we situate the presented theoretical example in an ideal scenario in which infinitely many data and sufficient training time is provided, our result (1) shows that the pretrained classifier becomes fully invariant to the spurious correlations. However, in practical setting with finite training time and number of samples, our result (2) shows that the pretrained model inevitably rely on the spuriously correlated features.

Beyond theoretical results, we empirically observe that the weight ratio α_i of pretrained classifiers indeed increases as $p^e \to 1$. We simulate the example presented in section 3.2 of the main paper, where the dimensionality D is set to 15, and probability p^e varies from 0.6 (weakly biased) to 0.99 (severely biased). We train a linear classifier for 500 epochs with batch size of 1024, and measure the unbiased accuracy on test samples generated from environment $e \in \mathcal{E}_{test}$. We also measure weight ratio $\text{mean}(\tilde{w}_{sp})/\tilde{w}_{inv}$, where $\text{mean}(\tilde{w}_{sp})$ denotes the average of pretrained spurious weights $\{w_{sp,i}\}_{i=1}^{D}$. To enable the end-to-end training, we use binary cross entropy loss instead of exponential loss, with setting $\mathcal{Y} = \{0, 1\}$ instead of $\mathcal{Y} = \{-1, 1\}$. We do not consider pruning process in this implementation. Figure 1 shows that the weight ratio increases to 1 in average as $p^e \to 1$. It implies that the spurious features $\boldsymbol{Z}_{sp}^e$ participate almost equally to the invariant feature Z_{inv}^e in the presence of strong spurious correlations. In this worst case, it is frustratingly difficult to discriminate weights necessary for OOD generalization in biased environment, resulting in the failure of learning optimal pruning parameters. Simulation results are averaged on 15 different random seeds.

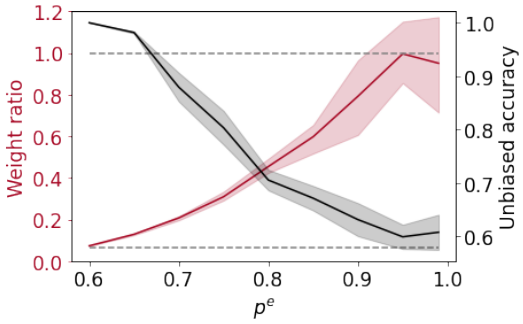

Figure 1. Implemented results of presented example.

2. Example of geometrical misalignment

In this section, we present a simple example illustrating the potential adverse effect of spurious correlations on latent representations. Consider independent arbitrary samples within the same class $\boldsymbol{X}_i^b, \boldsymbol{X}_j^b \sim P^b_{\boldsymbol{X}^b|Y^b=y}$ and $\boldsymbol{X}^d \sim P^d_{\boldsymbol{X}^d|Y^d=y}$ for a common $y \in \{-1, 1\}$ and environments b, d where $b \in \mathcal{E}_{train}$ and $d \in \mathcal{E}_{test}$. Let $\boldsymbol{W} \in \mathbb{R}^{Q\times(D+1)}$ be a weight matrix representation of a linear mapping $T : \{-1, 1\}^{D+1} \to \mathbb{R}^Q$ which encodes the embedding vector of a given sample. We denote such embedding as $\boldsymbol{h}^e = \boldsymbol{W}\boldsymbol{X}^e$ for some $e \in \mathcal{E}$. We assume that $\boldsymbol{W}$ is initialized as to be semi-orthogonal [4, 12] for simplicity. Then the following lemma reveals the geometrical misalignment of embeddings in the presence of strong spurious correlations:

Lemma 1. *Given $y \in \{-1, 1\}$, let $\boldsymbol{h}_i^b, \boldsymbol{h}_j^b, \boldsymbol{h}^d$ be embeddings of $\boldsymbol{X}_i^b, \boldsymbol{X}_j^b, \boldsymbol{X}^d$ respectively. Then, the expected cosine similarity between $\boldsymbol{h}_i^b$ and $\boldsymbol{h}^d$ is derived as:*

$$\mathbb{E}\left[\frac{\langle \boldsymbol{h}_i^b, \boldsymbol{h}^d\rangle}{\|\boldsymbol{h}_i^b\| \cdot \|\boldsymbol{h}^d\|} \,\middle|\, Y^b = y, Y^d = y\right] = \frac{1}{D+1}, \tag{29}$$

while the expected cosine similarity between $\boldsymbol{h}_i^b$ and $\boldsymbol{h}_j^b$ is derived as:

$$\mathbb{E}\left[\frac{\langle \boldsymbol{h}_i^b, \boldsymbol{h}_j^b\rangle}{\|\boldsymbol{h}_i^b\| \cdot \|\boldsymbol{h}_j^b\|} \,\middle|\, Y^b = y\right] = \frac{1 + D(2p^b - 1)^2}{D+1}, \tag{30}$$

where p^b is a probability parameter of Bernoulli distribution of i.i.d variable $Z_{sp,i}^b$, similar to p^e in the main paper.

Proof. Let $\boldsymbol{X}^e = \boldsymbol{V}^e_{inv} + \boldsymbol{V}^e_{sp}$ for the sample from an arbitrary environment e in general, where $\boldsymbol{v}^e_{inv}, \boldsymbol{v}^e_{sp} \in \{-1, 1\}^{D+1}$ are invariant and spurious component vector, respectively:

$$V^e_{inv,j} = \begin{cases} Z^e_{inv}, & \text{if } j = 1 \\ 0, & \text{otherwise} , \end{cases} \tag{31}$$

$$V^e_{sp,j} = \begin{cases} Z^e_{sp,j}, & \text{if } j = 2, \dots, D+1 \\ 0, & \text{otherwise} . \end{cases} \tag{32}$$

Thus, $\boldsymbol{V}^e_{inv}$ and $\boldsymbol{V}^e_{sp}$ are orthogonal. Given $Y^b = y$ and $Y^d = y$ for some $y \in \{-1, 1\}$, the cosine similarity between $\boldsymbol{h}^b_i$ and $\boldsymbol{h}^d$ is expressed as follows:

$$\begin{aligned} \mathbb{E}\left[\frac{\langle \boldsymbol{h}^b_i, \boldsymbol{h}^d \rangle}{\|\boldsymbol{h}^b_i\| \|\boldsymbol{h}^d\|} \,\middle|\, Y^b = y, Y^d = y \right] &= \mathbb{E}\left[\frac{\langle \boldsymbol{X}^b_i, \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{X}^d \rangle}{\|\boldsymbol{h}^b_i\| \|\boldsymbol{h}^d\|} \,\middle|\, Y^b = y, Y^d = y \right] \\ &= \mathbb{E}\left[\frac{\langle \boldsymbol{X}^b_i, \boldsymbol{X}^d \rangle}{D+1} \,\middle|\, Y^b = y, Y^d = y \right] \\ &= \mathbb{E}\left[\frac{\langle \boldsymbol{V}^b_{i,inv} + \boldsymbol{V}^b_{i,sp}, \boldsymbol{V}^d_{inv} + \boldsymbol{V}^d_{sp} \rangle}{D+1} \,\middle|\, Y^b = y, Y^d = y \right] \\ &= \frac{1}{D+1}, \end{aligned} \tag{33}$$

where $\boldsymbol{V}^b_{i,inv}$ and $\boldsymbol{V}^b_{i,sp}$ represent the invariant and spurious component vector of $\boldsymbol{X}^b_i$, respectively, and the second equality comes from the semi-orthogonality of $\boldsymbol{W}$. The last equality comes from the orthogonality of spurious component vector from different environment $b \in \mathcal{E}_{train}$ and $d \in \mathcal{E}_{test}$.

On the other hand, the expected cosine similarity between two arbitrary embeddings $\boldsymbol{h}^b_i$ and $\boldsymbol{h}^b_j$ from the biased environment b is expressed as follows:

$$\begin{aligned} \mathbb{E}\left[\frac{\langle \boldsymbol{h}^b_i, \boldsymbol{h}^b_j \rangle}{\|\boldsymbol{h}^b_i\| \|\boldsymbol{h}^b_j\|} \,\middle|\, Y^b = y \right] &= \mathbb{E}\left[\frac{\langle \boldsymbol{V}^b_{i,inv} + \boldsymbol{V}^b_{i,sp}, \boldsymbol{V}^b_{j,inv} + \boldsymbol{V}^b_{j,sp} \rangle}{D+1} \,\middle|\, Y^e = y \right] \\ &= \frac{1 + D(2p^b - 1)^2}{D+1}, \end{aligned} \tag{34}$$

where the last equality comes from the expectation of product of independent Bernoulli variables. □

The gap between (29) and (30) unveils the imbalance of distance between same-class embeddings from different environments on the unit hypersphere; embeddings from the training environment are more closely aligned to other embeddings from the same environment than embeddings from test environment at initial even when all samples are generated within the same class. While the Lemma 1 is only applicable to the initialized $\boldsymbol{W}$ before training, such imbalance may be worsened if $\boldsymbol{W}$ learns to project the samples on the high-dimensional subspace where most of its basis are independent to the invariant features. This sparks interests in designing weight pruning masks to aggregate the representations from same-class samples all together. Indeed, in this simple example, we can address this misalignment by masking out every weight in $\boldsymbol{W}$ except the first column, which is associated with the invariant feature.

From this point of view, we revisit the proposed alignment loss in main paper:

$$\ell_{align}\left(\{x_i, y_i\}_{i=1}^{|S|}; \tilde{\boldsymbol{W}}, \Theta \right) = \mathbb{E}_{\boldsymbol{m} \sim G(\Theta)} \Big[\ell_{con}(S_{bc}, S; \boldsymbol{m} \odot \tilde{\boldsymbol{W}}) + \ell_{con}(S_{ba}, S_{bc}; \boldsymbol{m} \odot \tilde{\boldsymbol{W}}) \Big], \tag{35}$$

where the first term reduces the gap between bias-conflicting samples and others, while the second term prevents bias-aligned samples from being aligned too close each other. In other words, the first term is aimed at increasing the cosine similarity between representations of same-class samples with different spurious attributes, as $\boldsymbol{h}^b_i$ and $\boldsymbol{h}^d$ in this example. The second term serves as a regularizer that pulls apart same-class bias-aligned representations, as $\boldsymbol{h}^b_i$ and $\boldsymbol{h}^b_j$ in this example. Thus we can leverage abundant bias-aligned samples as negatives regardless of their class in second term, while [16] limits the negatives to samples with different target label but same bias label, which are often highly scarce in a biased dataset.

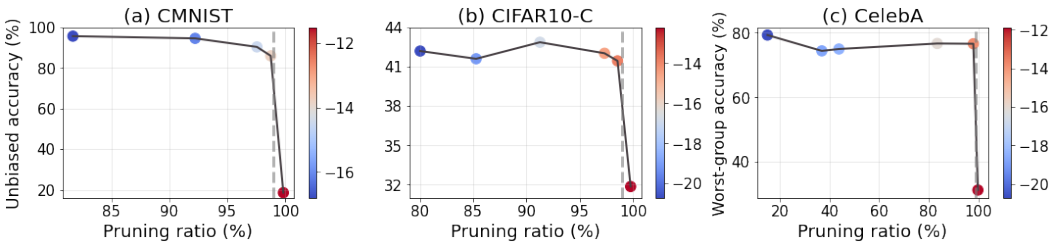

Figure 2. Analysis on the sparsity level. Bias ratio=1% for (a), (b). Color bar: log-scaled λ_{ℓ_1}. Dotted line: ratio = 99%.

3. Additional results

Comparisons to the pruning baselines. Pruning (debiasing) appears to suffer from the generalization-efficiency tradeoff; improving computational efficiency (OOD generalization) does not always guarantee improvement in OOD generalization (efficiency). Unlike this, our framework reliably improves both generalization and efficiency as shown in Table 1. Note that the standard pruning algorithms [14] fail to improve the unbiased accuracy in CIFAR10-C.

Table 1. Test (unbiased) accuracy (%) on standard and corrupted CIFAR10 (Bias ratio=5%). Pruning ratio=90.0% for GraSP and (92.4%, 90.4%) for DCWP on (CIFAR10, CIFAR10-C).

Dataset	Full-size	GraSP [14]	DCWP
CIFAR10	86.76	85.64	86.32
CIFAR10-C	45.66	44.21	60.24

Analysis of sparsity level. One may concern that a trade-off between performance and sparsity may exist. For example, networks with mild sparsity may still be over-parameterized and thus not fully debiased, whereas networks with high sparsity do not have enough capacity to preserve the averaged accuracy. In order to investigate the trade-off, we measure the unbiased accuracy by explicitly controlling the pruning ratio with varying λ_{ℓ_1}. Figure 2 shows that (1) the trade-off between performance and sparsity does exist, while (2) the proposed framework is reasonably tolerant to high sparsity in terms of generalization. We conjecture that such tolerance is owing to the *prioritized* elimination of spurious weights; the networks can be compressed to a significant extent without hurting the generalization after pruning out the spurious weights.

4. Experimental setup

4.1. Datasets

We mainly follow [7, 9] to evaluate our framework on Color-MNIST (CMNIST), Corrupted CIFAR-10 (CIFAR10-C) and Biased FFHQ (BFFHQ) as presented in Figure 3.

CMNIST. We first consider the prediction task of digit class which is spuriously correlated to the pre-assigned color, following the existing works [1, 7, 9, 13]. Each digit is colored with certain type of color, following [7, 9]. The ratio of bias-conflicting samples, i.e., bias ratio, is varied in range of $\{0.5\%, 1.0\%, 2.0\%, 5.0\%\}$, where the exact number of (bias-aligned, bias-conflicting) samples is set to: (54,751, 249)-0.5%, (54,509, 491)-1%, (54,014, 986)-2%, and (52,551, 2,449)-5%.

CIFAR10-C. Each sample in this dataset is generated by corrupting original samples in CIFAR-10 with certain types of corruption. Among 15 different corruptions introduced in the original paper [2], we select 10 types which are `Brightness`, `Contrast`, `Gaussian Noise`, `Frost`, `Elastic Transform`, `Gaussian Blur`, `Defocus Blur`, `Impulse Noise`, `Saturate`, and `Pixelate`, following [7]. Each of these corruption is spuriously correlated to the object classes of CIFAR-10, which are `Plane`, `Car`, `Bird`, `Cat`, `Deer`, `Dog`, `Frog`, `Horse`, `Ship`, and `Truck`. We use the samples corrupted in most severe level among five different severity, following [7]. The exact number of (bias-aligned, bias-conflicting) samples is set to: (44,832, 228)-0.5%, (44,527, 442)-1%, (44,145, 887)-2%, and (42,820, 2,242)-5%.

BFFHQ. Each sample in this biased dataset are selected from Flickr-Faces-HQ (FFHQ) Dataset [5], where we conduct binary classifications with considering (`Age`, `Gender`) as target and spuriously correlated attribute pair following [6, 7]. Specifically, majority of training images correspond to either young women (i.e., aged 10-29) or old men (i.e., aged 40-59).

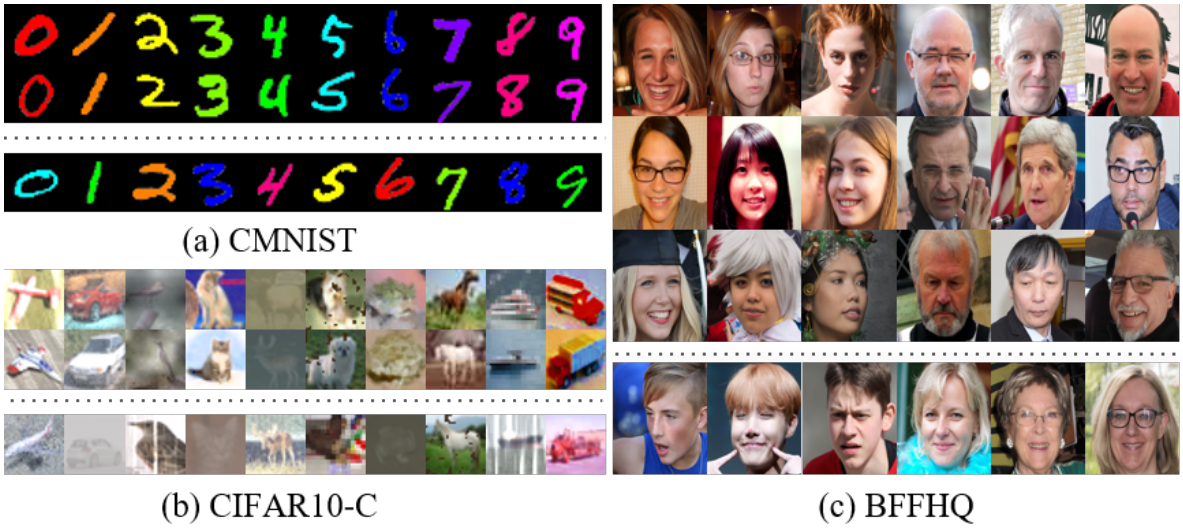

Figure 3. Example images of datasets. The images above the dotted line denote the bias-aligned samples, while the ones below the dotted line are the bias-conflicting samples. For CMNIST and CIFAR10-C, each column indicates each class. For BFFHQ, the group of three columns indicates each class.

This dataset consists of 19,104 number of such bias-aligned samples and 96 number of bias-conflicting samples, i.e., old women and young men.

CelebA. For CelebA, we consider (`Blonde Hair`, `Male`) as (target, spurious) attribute pair, following [3, 9, 10]. Pixel resolutions and batch size are 256×256 and 128, respectively. The exact number of samples for the prediction task follows that from [3].

4.2. Simulation settings

Architecture details. We use a simple convolutional network with three convolution layers for CMNIST, with feature map dimensions of 64, 128 and 256, each followed by a ReLU activation and a batch normalization layer following [15]. For CIFAR10-C and BFFHQ, we use ResNet-18 with pretrained weights provided in PyTorch `torchvision` implementations. Each convolutional network and ResNet-18 includes 1.3×10^6 and 2.2×10^7 number of parameters, respectively. We assign a pruning parameter for each weight parameter except bias in deep networks. Each of pruning parameter is initialized with value 1.5 so that the initial probability of preserving the corresponding weight is set to $\sigma(1.5) \approx 0.8$ in default.

Training details. We first train bias-capturing networks using GCE loss (q=0.7) for (CMNIST, BFFHQ, CelebA), with (2000, 10000, 10000) iterations, respectively. For CIFAR10-C, we use epoch-ensemble-based mining algorithms presented in [17], which select samples cooperated with an ensemble of predictions at each epoch to prevent overfitting. We use b-c score threshold $\tau = 0.8$ and the confidence threshold $\eta = 0.05$ as suggested in the original paper.

Then, main networks are pretrained for 10000 iterations using an Adam optimizer with learning rates 0.01, 0.001, 0.001 and 0.0001 for CMNIST, CIFAR10-C, BFFHQ, and CelebA, respectively.

We train pruning parameters for 2000 iterations using a learning rate 0.01, upweighting hyperparameter $\lambda_{up} = 80$ and a balancing hyperparameter $\lambda_{align} = 0.05$ for each dataset. We use a Lagrangian multiplier $\lambda_{\ell_1} = 10^{-8}$ for CMNIST, and $\lambda_{\ell_1} = 10^{-9}$ for CIFAR10-C, BFFHQ and CelebA. Specifically, we set λ_{ℓ_1} by considering the size of deep networks, where we found that the value within range $\mathcal{O}(0.1 * n^{-1})$ serves as a good starting point where n is the number of parameters.

After pruning, we finetune the networks with decaying learning rate to 0.001 for CMNIST and 0.0005 for others. We use $\lambda_{align} = 0.05$ consistently. Then, we use $\lambda_{up} = 80$ for BFFHQ, $\lambda_{up} = 20$ for CelebA, and $\lambda_{up} = \{10, 30, 50, 80\}$ for CMNIST and CIFAR10-C with $\{0.5\%, 1.0\%, 2.0\%, 5.0\%\}$ of bias ratio, respectively.

Considering the pruning as a strong regularization, we did not use additional capacity control techniques such as early stopping or strong ℓ_2 regularization presented in [8, 11].

Data augmentations. We did not use any kinds of data augmentations which may implicitly enforce networks to encode invariances. For the BFFHQ and CelebA dataset, we only apply random horizontal flip. For the CIFAR10-C dataset, we take 32×32 random crops from image padded by 4 pixels followed by random horizontal flip, following [9]. We do not use any

kinds of augmentations in CMNIST.

Baselines. We use the official implementations of Rebias, LfF, DisEnt and JTT released by authors, and reproduce EnD and MRM by ourselves. For DisEnt, we use the official hyperparameter configurations provided in the original paper. We use $q = 0.7$ for LfF as suggested by authors on every experiment. For Rebias, we use the official hyperparameter configurations for CMNIST, and train for 200 epochs using Adam optimizer with learning rate 0.001 and RBF kernel radius of 1 for other datasets. For MRM, we use λ_{ℓ_1} of 10^{-8} for CMNIST following the original paper, and 10^{-9} for the others. For EnD, we set the multipliers α for disentangling and β for entangling to 1.

References

[1] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020. 10

[2] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 10

[3] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34:26449–26461, 2021. 11

[4] Wei Hu, Lechao Xiao, and Jeffrey Pennington. Provable benefit of orthogonal initialization in optimizing deep linear networks. *arXiv preprint arXiv:2001.05992*, 2020. 8

[5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 10

[6] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14992–15001, 2021. 10

[7] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34:25123–25133, 2021. 10

[8] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 11

[9] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 10, 11

[10] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 11

[11] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020. 11

[12] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013. 8

[13] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. End: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13508–13517, 2021. 10

[14] Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. *arXiv preprint arXiv:2002.07376*, 2020. 10

[15] Dinghuai Zhang, Kartik Ahuja, Yilun Xu, Yisen Wang, and Aaron Courville. Can subnetwork structure be the key to out-of-distribution generalization? In *International Conference on Machine Learning*, pages 12356–12367. PMLR, 2021. 11

[16] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022. 9

[17] Bowen Zhao, Chen Chen, Qi Ju, and Shutao Xia. Learning debiased models with dynamic gradient alignment and bias-conflicting sample mining. *arXiv preprint arXiv:2111.13108*, 2021. 11