## A. A toy example

We would like to show that the optimal solution that minimize the worst-case risk across $E_2^{1\checkmark}$ and $E_2^{1\times}$ is to predict $Y$ only using $X_1$. Consider any classifier $f(Y \mid X_1, X_2)$ and its marginal

$$f(Y \mid X_1) \propto f(X_1, X_2 = 0, Y) + f(X_1, X_2 = 1, Y).$$

For any input $(x_1, x_2) \in E_2^{1\checkmark}$, based on our construction, the distribution $P_2^{1\checkmark}(X_1 = x_1, X_2 = x_2, Y)$ only has mass on one label value $y \in \{0, 1\}$. Thus $P_2^{1\checkmark}(Y = y \mid X_1 = x_1, X_2 = x_2) = 1$. We can then write the log risk of the classifier $f(Y \mid X_1, X_2)$ as

$$-\log \frac{f(x_1, x_2, y)}{f(x_1, x_2, y) + f(x_1, x_2, 1-y)}.$$

The log risk of the marginal classifier $f(Y \mid X_1)$ is defined as

$$\begin{aligned}-\log \Big( &(f(x_1, x_2, y) + f(x_1, 1-x_2, y)) \\ &/(f(x_1, x_2, y) + f(x_1, 1-x_2, y) \\ &\quad + f(x_1, x_2, 1-y) + f(x_1, 1-x_2, 1-y))\Big).\end{aligned}$$

Now suppose $f(Y \mid X_1, X_2)$ achieves a lower risk than $f(Y \mid X_1)$. This implies

$$\begin{aligned}&f(x_1, x_2, y)f(x_1, x_2, y) + f(x_1, x_2, 1-y)f(x_1, x_2, y) \\ &\quad + f(x_1, x_2, y)f(x_1, 1-x_2, y) \\ &\quad + f(x_1, x_2, 1-y)f(x_1, 1-x_2, y) \\ &< f(x_1, x_2, y)f(x_1, x_2, y) + f(x_1, x_2, y)f(x_1, x_2, 1-y) \\ &\quad + f(x_1, x_2, y)f(x_1, 1-x_2, y) \\ &\quad + f(x_1, x_2, y)f(x_1, 1-x_2, 1-y).\end{aligned}$$

Note that the first three terms on both side cancel out. We have

$$\begin{aligned}&f(x_1, x_2, 1-y)f(x_1, 1-x_2, y) \\ &\qquad < f(x_1, x_2, y)f(x_1, 1-x_2, 1-y).\end{aligned}$$

Now let's consider an input $(x_1, 1-x_2) \in E_2^{1\times}$. Based on our construction of the partitions, we have $P_2^{1\checkmark}(x_1, x_2, y) = P_2^{1\times}(x_1, 1-x_2, y)$. The log risk of the marginal classifier on $P_2^{1\times}$ is still the same, but the log risk of the classifier $f(Y \mid X_1, X_2)$ now becomes

$$-\log \frac{f(x_1, 1-x_2, y)}{f(x_1, 1-x_2, y) + f(x_1, 1-x_2, 1-y)}.$$

We claim that the log risk of $f(Y \mid X_1, X_2)$ is higher than $f(Y \mid X_1)$ on $P_2^{1\times}$. Suppose for contradiction that the log risk of $f(Y \mid X_1, X_2)$ is lower, then we have

$$\begin{aligned}&f(x_1, 1-x_2, y)f(x_1, x_2, y) + f(x_1, 1-x_2, 1-y)f(x_1, x_2, y) \\ &\quad + f(x_1, 1-x_2, y)f(x_1, 1-x_2, y) \\ &\quad + f(x_1, 1-x_2, 1-y)f(x_1, 1-x_2, y) \\ &< f(x_1, 1-x_2, y)f(x_1, x_2, y) + f(x_1, 1-x_2, y)f(x_1, 1-x_2, y) \\ &\quad + f(x_1, 1-x_2, y)f(x_1, x_2, 1-y) \\ &\quad + f(x_1, 1-x_2, y)f(x_1, 1-x_2, 1-y).\end{aligned}$$

Canceling out the terms, we obtain

$$\begin{aligned}&f(x_1, 1-x_2, 1-y)f(x_1, x_2, y) \\ &\qquad < f(x_1, 1-x_2, y)f(x_1, x_2, 1-y).\end{aligned}$$

Contradiction!

Thus the marginal $f(Y \mid X_1)$ will always reach a better worst-group risk compare to the original classifier $f(Y \mid X_1, X_2)$. As a result, the optimal classifier $f(Y \mid X_1, X_2)$ should satisfy $f(Y \mid X_1, X_2) = f(Y \mid X_1)$, i.e., it will only use $X_1$ to predict $Y$.

## B. Theoretical analysis

**Proposition 1.** *For a pair of environments $E_i$ and $E_j$, assuming that the classifier $f_i$ is able to learn the true conditional $P_i(Y \mid X_1, X_2)$, we can write the joint distribution $P_j$ of $E_j$ as the mixture of $P_j^{i\checkmark}$ and $P_j^{i\times}$:*

$$P_j(x_1, x_2, y) = \alpha_j^i P_j^{i\checkmark}(x_1, x_2, y) + (1-\alpha_j^i) P_j^{i\times}(x_1, x_2, y),$$

*where $\alpha_j^i = \sum_{x_1, x_2, y} P_j(x_1, x_2, y) \cdot P_i(y \mid x_1, x_2)$ and*

$$\begin{aligned}P_j^{i\checkmark}(x_1, x_2, y) &\propto P_j(x_1, x_2, y) \cdot P_i(y \mid x_1, x_2), \\ P_j^{i\times}(x_1, x_2, y) &\propto P_j(x_1, x_2, y) \cdot P_i(1-y \mid x_1, x_2).\end{aligned}$$

*Proof.* For ease of notation, let $i = 1$, $j = 2$. For an input $(x_1, x_2)$, let's first consider the conditional probability $P_2^{1\times}(y \mid x_1, x_2)$ and $P_2^{1\checkmark}(y \mid x_1, x_2)$. Since the input is in $E_2$, the probability that it has label $y$ is given by $P_2(y \mid x_1, x_2)$. Since $f_1$ matches $P_1(y \mid x_1, x_2)$, the likelihood that the prediction is wrong is given by $P_1(1-y \mid x_1, x_2)$ and the likelihood that the prediction is correct is givn by $P_1(y \mid x_1, x_2)$. Thus, we have

$$\begin{aligned}P_2^{1\times}(y \mid x_1, x_2) &= \frac{P_1(1-y \mid x_1, x_2) P_2(y \mid x_1, x_2)}{\sum_{y'} P_1(1-y' \mid x_1, x_2) P_2(y' \mid x_1, x_2)}, \\ P_2^{1\checkmark}(y \mid x_1, x_2) &= \frac{P_1(y \mid x_1, x_2) P_2(y \mid x_1, x_2)}{\sum_{y'} P_1(y' \mid x_1, x_2) P_2(y' \mid x_1, x_2)}.\end{aligned}$$

Now let's think about the marginal of $(x_1, x_2)$ if it is in the set of mistakes $E_2^{1\times}$. Again, since the input is in $E_2$, the probability that it exists is given by the marginal in $E_2$:

$P_2(x_1, x_2)$. This input has two possibilities to be partitioned into $E_2^{1\times}$: 1) the label is $y$ and $f_1$ predicts it as $1-y$; 2) the label is $1-y$ and $f_1$ predicts it as $y$. Marginalizing over all $(x_1, x_2)$, we have

$$
\begin{aligned}
&P_2^{1\times}(x_1, x_2) \\
&= \frac{\frac{P_2(x_1,x_2)\sum_y P_1(1-y|x_1,x_2)P_2(y|x_1,x_2)}{\sum_y P_1(1-y|x_1,x_2)P_2(y|x_1,x_2)+P_1(y|x_1,x_2)P_2(y|x_1,x_2)}}{\sum_{x_1',x_2'}\frac{P_2(x_1',x_2')\sum_y P_1(1-y|x_1',x_2')P_2(y|x_1',x_2')}{\sum_y P_1(1-y|x_1',x_2')P_2(y|x_1',x_2')+P_1(y|x_1',x_2')P_2(y|x_1',x_2')}} \\
&= \frac{P_2(x_1, x_2)\sum_y P_1(1-y \mid x_1, x_2)P_2(y \mid x_1, x_2)}{\sum_{x_1',x_2'} P_2(x_1', x_2')\sum_y P_1(1-y \mid x_1', x_2')P_2(y \mid x_1', x_2')}
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&P_2^{1\checkmark}(x_1, x_2) \\
&= \frac{P_2(x_1, x_2)\sum_y P_1(y \mid x_1, x_2)P_2(y \mid x_1, x_2)}{\sum_{x_1',x_2'} P_2(x_1', x_2')\sum_y P_1(y \mid x_1', x_2')P_2(y \mid x_1', x_2')}
\end{aligned}
$$

Combining these all together using the Bayes' theorem, we have

$$
\begin{aligned}
&P_2^{1\times}(x_1, x_2, y) \\
&= \frac{P_1(1-y \mid x_1, x_2)P_2(y \mid x_1, x_2)P_2(x_1, x_2)}{\sum_{x_1',x_2'} P_2(x_1', x_2')\sum_{y'} P_1(1-y' \mid x_1', x_2')P_2(y' \mid x_1', x_2')}, \\
&= \frac{P_1(1-y \mid x_1, x_2)P_2(x_1, x_2, y)}{\sum_{x_1',x_2',y'} P_2(x_1', x_2', y')P_1(1-y' \mid x_1', x_2')}, \\
&\propto P_1(1-y \mid x_1, x_2)P_2(x_1, x_2, y), \\
&P_2^{1\checkmark}(x_1, x_2, y) \\
&= \frac{P_1(y \mid x_1, x_2)P_2(y \mid x_1, x_2)P_2(x_1, x_2)}{\sum_{x_1',x_2'} P_2(x_1', x_2')\sum_{y'} P_1(y' \mid x_1', x_2')P_2(y' \mid x_1', x_2')}, \\
&= \frac{P_1(y \mid x_1, x_2)P_2(x_1, x_2, y)}{\sum_{x_1',x_2',y'} P_2(x_1', x_2', y')P_1(y' \mid x_1', x_2')}, \\
&\propto P_1(y \mid x_1, x_2)P_2(x_1, x_2, y).
\end{aligned}
$$

Finally, it is straightforward to show that for $\alpha_2^1 = \sum_{x_1,x_2,y} P_2(x_1, x_2, y)P_1(y \mid x_1, x_2)$, we have

$$
\begin{aligned}
&\alpha_2^1 P_2^{1\checkmark}(x_1, x_2, y) + (1-\alpha_2^1)P_2^{1\times} \\
&= P_1(y \mid x_1, x_2)P_2(x_1, x_2, y) \\
&\qquad + P_1(1-y \mid x_1, x_2)P_2(x_1, x_2, y) \\
&= P_2(x_1, x_2, y).
\end{aligned}
$$

□

From now on, we assume that the marginal distribution of $Y$ is uniform in all joint distributions, i.e., $f_i$ performs equally well on different labels.

**Theorem 1.** *Suppose $X_2$ is independent of $X_1$ given $Y$. For any environment pair $E_i$ and $E_j$, if $\sum_y P_i(x_2 \mid y) = \sum_y P_j(x_2 \mid y)$ for any $x_2$, then $\mathrm{Cov}(X_2, Y; P_i) > \mathrm{Cov}(X_2, Y; P_j)$ implies $\mathrm{Cov}(X_2, Y; P_j^{i\times}) < 0$ and $\mathrm{Cov}(X_2, Y; P_i^{j\times}) > 0$.*

*Proof.* By definition, we have

$$
\begin{aligned}
&\mathrm{Cov}(X_2, Y; P_j^{i\times}) \\
&= \mathbb{E}[X_2 Y; P_j^{i\times}] - \mathbb{E}[X_2; P_j^{i\times}]\,\mathbb{E}[Y; P_j^{i\times}] \\
&= \sum_{x_1,x_2} x_2 P_j^{i\times}(x_1, x_2, 1) \\
&\quad - \sum_{x_1,x_2,y} x_2 P_j^{i\times}(x_1, x_2, y) \sum_{x_1,x_2} P_j^{i\times}(x_1, x_2, 1) \\
&= \sum_{x_1,x_2,x_1',x_2',y'} x_2 P_j^{i\times}(x_1, x_2, 1) P_j^{i\times}(x_1', x_2', y') \\
&\quad - \sum_{x_1,x_2,y,x_1',x_2'} x_2 P_j^{i\times}(x_1, x_2, y) P_j^{i\times}(x_1', x_2', 1)
\end{aligned}
$$

Expanding the distributions of $P_j^{i\times}$, it suffices to show that

$$
\begin{aligned}
&\sum_{x_1,x_2,x_1',x_2',y'} \Big( x_2 P_j(x_1, x_2, 1) P_i(0 \mid x_1, x_2) \\
&\qquad\qquad P_j(x_1', x_2', y') P_i(1-y' \mid x_1', x_2') \Big) \\
&< \sum_{x_1,x_2,y,x_1',x_2'} \Big( x_2 P_j(x_1, x_2, y) P_i(1-y \mid x_1, x_2) \\
&\qquad\qquad P_j(x_1', x_2', 1) P_i(0 \mid x_1', x_2') \Big)
\end{aligned}
$$

Note that when $y = y' = 1$, two terms cancel out. Thus we need to show

$$
\begin{aligned}
&\sum_{x_1,x_2,x_1',x_2'} \Big( x_2 P_j(x_1, x_2, 1) P_i(0 \mid x_1, x_2) \\
&\qquad\qquad P_j(x_1', x_2', 0) P_i(1 \mid x_1', x_2') \Big) \\
&< \sum_{x_1,x_2,x_1',x_2'} \Big( x_2 P_j(x_1, x_2, 0) P_i(1 \mid x_1, x_2) \\
&\qquad\qquad P_j(x_1', x_2', 1) P_i(0 \mid x_1', x_2') \Big)
\end{aligned}
$$

Based on the assumption that the marginal distribution in $E_j^{i\times}$ is uniform, we have

$$
\begin{aligned}
&\sum_{x_1',x_2'} P_j(x_1', x_2', 0) P_i(1 \mid x_1', x_2') \\
&\doteq \sum_{x_1',x_2'} P_j(x_1', x_2', 1) P_i(0 \mid x_1', x_2').
\end{aligned}
$$

Thus we can simplify our goal as

$$
\begin{aligned}
&\sum_{x_1,x_2} x_2 P_j(x_1, x_2, 1) P_i(0 \mid x_1, x_2) \\
&< \sum_{x_1,x_2} x_2 P_j(x_1, x_2, 0) P_i(1 \mid x_1, x_2)
\end{aligned}
$$

Similarly, we can simplify the condition $\mathrm{Cov}(X_2, Y; P_i) > \mathrm{Cov}(X_2, Y; P_j)$ as

$$\sum_{x1,x_2} x_2(P_j(x_1, x_2, 1) - P_i(x_1, x_2, 1)) < \sum_{x1,x_2} x_2(P_j(x_1, x_2, 0) - P_i(x_1, x_2, 0))$$

Since $x_2$ is independent of $x_1$ given $y$, we have

$$\sum_{x1,x_2} x_2(P_j(x_1, y=1)P_j(x_2 \mid y=1) - P_i(x_1, y=1)P_i(x_2 \mid y=1)) < \sum_{x1,x_2} x_2(P_j(x_1, y=0)P_j(x_2 \mid y=0) - P_i(x_1, y=0)P_i(x_2 \mid y=0))$$

Since $x_1$ is the stable feature and the label marginal is the same across environments, we have $P_j(x_1, y=1) = P_i(x_1, y=1)$ and $P_j(x_1, y=0) = P_i(x_1, y=0)$. This implies

$$\sum_{x_1} P_j(x_1, y=1) \sum_{x_2} x_2(P_j(x_2 \mid y=1) - P_i(x_2 \mid y=1)) < \sum_{x_1} P_j(x_1, y=0) \sum_{x_2} x_2(P_j(x_2 \mid y=0) - P_i(x_2 \mid y=0))$$

Again, by uniform label marginals, we have

$$\sum_{x_2} x_2(P_j(x_2 \mid y=1) - P_i(x_2 \mid y=1)) < \sum_{x_2} x_2(P_j(x_2 \mid y=0) - P_i(x_2 \mid y=0).$$

For binary $x_2 \in \{0, 1\}$, this implies $P_j(x_2 = 1 \mid y = 1) + P_i(x_2 = 1 \mid y = 0) < P_j(x_2 = 1 \mid y = 0) + P_i(x_2 = 1 \mid y = 1)$. Since $P_j(x_2 \mid y = 1) + P_j(x_2 \mid y = 0) = P_i(x_2 \mid y = 1) + P_i(x_2 \mid y = 0)$, we have

$$P_j(x_2 \mid y=1)P_j(x_2 \mid y=0) < P_j(x_2 \mid y=0)P_j(x_2 \mid y=1). \tag{1}$$

We can expand our goal in the same way:

$$\begin{aligned}
&\sum_{x_1,x_2} x_2 P_j(x_1, x_2, 1) P_i(0 \mid x_1, x_2) \\
&= \sum_{x_1,x_2} \Big( x_2 P_j(x_1, y=1) P_i(x_1, y=0) \\
&\qquad P_j(x_2 \mid y=1) P_i(x_2 \mid y=0) \Big) / P_i(x_1, x_2) \\
&= \sum_{x_1} P_j(x_1, y=1) P_i(x_1, y=0) \\
&\quad \cdot \sum_{x_2} \frac{x_2 P_j(x_2 \mid y=1) P_i(x_2 \mid y=0)}{P_i(x_1, x_2)} \\
&\sum_{x_1,x_2} x_2 P_j(x_1, x_2, 0) P_i(1 \mid x_1, x_2) \\
&= \sum_{x_1} P_j(x_1, y=0) P_i(x_1, y=1) \\
&\quad \cdot \sum_{x_2} \frac{x_2 P_j(x_2 \mid y=0) P_i(x_2 \mid y=1)}{P_i(x_1, x_2)},
\end{aligned}$$

Plug in Eq (1) and we complete the proof. The other inequality follows by symmetry. □

**Extension to multi-class classification:** In Theorem 1, we focus on binary classification for simplicity. For multi-class classification, we can convert it into a binary problem by defining $Y_c$ as a binary indicator of whether class $c$ is present or absent. Our strong empirical performance on MNIST (10-class classification) also confirms that our results generalize to the multi-class setting.

**Theorem 2.** *For any environment pair $E_i$ and $E_j$, $\mathrm{Cov}(X_2, Y; P_i) > \mathrm{Cov}(X_2, Y; P_j)$ implies*

$$\begin{aligned}
&\mathrm{Cov}(X_2, Y; P_j^{i\times}) \\
&< \frac{1-\alpha_j^i}{\alpha_i^i}\mathrm{Cov}(X_2, Y; P_i^{i\checkmark}) - \frac{1-\alpha_j^i}{\alpha_j^i}\mathrm{Cov}(X_2, Y; P_j^{i\checkmark}) \\
&\mathrm{Cov}(X_2, Y; P_i^{j\times}) \\
&> \frac{1-\alpha_i^j}{\alpha_j^j}\mathrm{Cov}(X_2, Y; P_j^{j\checkmark}) - \frac{1-\alpha_i^j}{\alpha_i^j}\mathrm{Cov}(X_2, Y; P_i^{j\checkmark})
\end{aligned}$$

*where $P_i^{i\checkmark}$ is the distribution of the correct predictions when applying $f_i$ on $E_i$.*

*Proof.* From the proof in Theorem 1, we can write the condition $\mathrm{Cov}(X_2, Y; P_i) > \mathrm{Cov}(X_2, Y; P_j)$ as

$$\sum_{x1,x_2} x_2(P_j(x_1, x_2, 1) - P_i(x_1, x_2, 1)) < \sum_{x1,x_2} x_2(P_j(x_1, x_2, 0) - P_i(x_1, x_2, 0))$$

Using $P_i(0 \mid x_1, x_2) + P_i(1 \mid x_1, x_2) = 1$,

$$\sum_{x1,x_2} x_2(P_j(x_1, x_2, 1) - P_i(x_1, x_2, 1)) \quad (P_i(0 \mid x_1, x_2) + P_i(1 \mid x_1, x_2)) < \sum_{x1,x_2} x_2(P_j(x_1, x_2, 0) - P_i(x_1, x_2, 0)) \quad (P_i(0 \mid x_1, x_2) + P_i(1 \mid x_1, x_2))$$

Since $P_i(x_1, x_2, 1)P_i(0 \mid x_1, x_2)$ and $P_i(x_1, x_2, 0)P_i(1 \mid$

$x_1, x_2)$ cancel out with each other. We have

$$\begin{aligned}
&\sum_{x1,x_2} x_2(P_j(x_1,x_2,1)P_i(0 \mid x_1,x_2) - \\
&\qquad P_j(x_1,x_2,0)P_i(1 \mid x_1,x_2)) \\
&< \sum_{x1,x_2} x_2(P_i(x_1,x_2,1)P_i(1 \mid x_1,x_2) - \\
&\qquad P_i(x_1,x_2,0)P_i(0 \mid x_1,x_2)) \\
&\quad - \sum_{x1,x_2} x_2(P_j(x_1,x_2,1)P_i(1 \mid x_1,x_2) - \\
&\qquad P_j(x_1,x_2,0)P_i(0 \mid x_1,x_2))
\end{aligned}$$

From the derivations in Theorem 1, we know that

$$\begin{aligned}
&\frac{1}{2(1-\alpha_j^i)}\mathrm{Cov}(X_2, Y; P_j^{i\times}) \\
&= \sum_{x1,x_2} x_2\Big(P_j(x_1,x_2,1)P_i(0 \mid x_1,x_2) - \\
&\qquad P_j(x_1,x_2,0)P_i(1 \mid x_1,x_2)\Big) \\
&\frac{1}{2\alpha_j^i}\mathrm{Cov}(X_2, Y; P_j^{i\checkmark}) \\
&= \sum_{x1,x_2} x_2\Big(P_j(x_1,x_2,1)P_i(1 \mid x_1,x_2) \\
&\qquad P_j(x_1,x_2,0)P_i(0 \mid x_1,x_2)\Big) \\
&\frac{1}{2\alpha_i^i}\mathrm{Cov}(X_2, Y; P_i^{i\checkmark}) \\
&= \sum_{x1,x_2} x_2\Big(P_i(x_1,x_2,1)P_i(1 \mid x_1,x_2) - \\
&\qquad P_i(x_1,x_2,0)P_i(0 \mid x_1,x_2)\Big).
\end{aligned}$$

Combining these, we have

$$\begin{aligned}
&\mathrm{Cov}(X_2, Y; P_j^{i\times}) \\
&< \frac{1-\alpha_j^i}{\alpha_i^i}\mathrm{Cov}(X_2, Y; P_i^{i\checkmark}) - \frac{1-\alpha_j^i}{\alpha_j^i}\mathrm{Cov}(X_2, Y; P_j^{i\checkmark})
\end{aligned}$$

Similarly, by using $P_j(0 \mid x_1, x_2) + P_j(1 \mid x_1, x_2) = 1$, we can get

$$\begin{aligned}
&\mathrm{Cov}(X_2, Y; P_i^{j\times}) \\
&> \frac{1-\alpha_i^j}{\alpha_j^j}\mathrm{Cov}(X_2, Y; P_j^{j\checkmark}) - \frac{1-\alpha_i^j}{\alpha_i^j}\mathrm{Cov}(X_2, Y; P_i^{j\checkmark})
\end{aligned}$$

□

# C. Experimental Setup

## C.1. Datasets and Models

### C.1.1. MNIST

**Data** We use the official train-test split of MNIST. Training environments are constructed from training split, with 14995 examples per environment. Validation data and testing data is constructed based on the testing split, with 2497 examples each. Following Arjovsky et al. (2019), We convert each grey scale image into a $10 \times 28 \times 28$ tensor, where the first dimension corresponds to the spurious color feature.

**Model:** The input image is passed to a CNN with 2 convolution layers and 2 fully connected layers. We use the architecture from PyTorch's MNIST example[6].

### C.1.2. Beer Review

**Data** We use the data processed by Lei et al. (2016). Reviews shorter than 10 tokens or longer than 300 tokens are filtered out. For each aspect, we sample training/validation/testing data randomly from the dataset and maintain the marginal distribution of the label to be uniform. Each training environment contains 4998 examples. The validation data contains 4998 examples and the testing data contains 5000 examples. The vocabulary sizes for the three aspects (look, aroma, palate) are: 10218, 10154 and 10086. The processed data will be publicly available.

**Model** We use a standard CNN text classifier (Kim, 2014). Each input is first encoded by pre-trained FastText embeddings (Mikolov et al., 2018). Then it is passed into a 1D convolution layer followed by max pooling and ReLU activation. The convolution layer uses filter size $3, 4, 5$. Finally we attach a linear layer with Softmax to predict the label.

### C.1.3. CelebA

**Data** We use the official train/val/test split of CelebA (Liu et al., 2015b). The training environment {`female`} contains 94509 examples and the training environment {`male`} contains 68261 examples. The validation set has 19867 examples and the test set has 19962 examples.

**Model** We use the Pytorch torchvision implementation of the ResNet50 model, starting from pretrained weights. We re-initalize the final layer to predict the target attribute `hair color`.

### C.1.4. ASK2ME

**Data** Since the original data doesn't have a standard train/val/test split, we randomly split the data and use 50% for training, 20% for validation, 30%for testing. There are 2227 examples in the training environment {`breast_cancer=0`}, 1394 examples in the training environment {`breast_cancer=1`}. The validation set contains 1448 examples and the test set contains 2173 examples. The vocabulary size is 16310. The processed data will be publicly available.

[6] https://github.com/pytorch/examples/blob/master/mnist/main.py

**Model** The model architecture is the same as the one for Beer review.

### C.2. Implementation details

**For all methods:** We use batch size 50 and evaluate the validation performance every 100 batch. We apply early stopping once the validation performance hasn't improved in the past 20 evaluations. We use Adam (Kingma & Ba, 2014) to optimize the parameters and tune the learning rate $\in \{10^{-3}, 10^{-4}, 10^{-5}\}$. For simplicity, we train all methods without data augmentation. Following Sagawa et al. (2019), we apply strong regularizations to avoid over-fitting. Specifically, we tune the dropout rate $\in \{0.1, 0.3, 0.5\}$ for text classification datasets (Beer review and ASK2ME) and tune the weight decay parameters $\in \{10^{-0}, 10^{-1}, 10^{-2}, 10^{-3}\}$ for image datasets (MNIST and CelebA).

**DRO** and **Ours** We directly optimize the $\min - \max$ objective. Specifically, at each step, we sample a batch of example from each group, and minimize the worst-group loss. We found the training process to be pretty stable when using the Adam optimizer. On CelebA, we are able to match the performance reported by Sagawa et al. (2019).

**IRM** We implement the gradient penalty based on the official implementation of IRM[7]. The gradient penalty is applied to the last hidden layer of the network. We tune the weight of the penalty term $\in \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$ and the annealing iterations $\in \{10, 10^2, 10^3\}$.

**RGM** For the per-environment classifier in RGM, we use a MLP with one hidden layer. This MLP takes the last layer of the model as input and predicts the label. Similar to IRM, we tune the weight of the regret $\in \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$ and the annealing iterations $\in \{10, 10^2, 10^3\}$.

### C.3. Computing Infrastructure and Running Time Analysis

We have used the following graphics cards for our experiments: Tesla V100-32GB, GeForce RTX 2080 Ti and A100-40G.

We conducted our running time analysis on MNIST and ASK2ME using GeForce RTX 2080 Ti. Table 5 and 6 shows the results. We observe that due to the direct optimization of the $\min\max$ objective, the running time of DRO, PI and Oracle is roughly 4 times comparing to other methods (proportional to the number of groups). Also, while our model needs to train additional environment-specific classifiers (comparing to DRO), its running time is very similar to DRO across the two datasets. We believe by using the online learning algorithm proposed by Sagawa et al. (2019), we can further reduce the running time of our algorithm.

[7] https://github.com/facebookresearch/InvariantRiskMinimization

| | TIME | Train | Val | Test |
|---|---|---|---|---|
| ERM | 2 MIN 58 SEC | 83.61 | 81.21 | 15.65 |
| IRM | 3 MIN 37 SEC | 83.42 | 80.41 | 12.89 |
| RGM | 3 MIN 7 SEC | 82.60 | 81.41 | 13.97 |
| DRO | 17 MIN 19 SEC | 79.44 | 80.65 | 16.05 |
| OURS | 11 MIN 58 SEC | 65.04 | 71.16 | 71.56 |
| ORACLE | 14 MIN 31 SEC | 68.96 | 72.28 | 70.04 |

*Table 5.* Running time and model performance on MNIST. Here the validation data is sampled from the training environments. Our algorithm requires training additional environment-specific classifiers. However, it converges faster than DRO in the third stage (50 epochs vs. 72 epochs) and generalizes much better.

| | TIME | Train | Val | Test |
|---|---|---|---|---|
| ERM | 3 MIN 35 SEC | 99.44 | 66.01 | 59.04 |
| IRM | 3 MIN 21 SEC | 98.70 | 63.10 | 57.85 |
| RGM | 5 MIN 36 SEC | 99.78 | 64.07 | 59.99 |
| DRO | 16 MIN 40 SEC | 86.77 | 77.66 | 67.34 |
| PI (Ours) | 18 MIN | 97.09 | 78.64 | 74.14 |

*Table 6.* Running time and model performance on ASK2ME. Here the validation accuracy is computed based on the `breast_cancer` attribute. The test accuracy is the average worst-group accuracy across all 17 attributes. Our algorithm's running time is similar to DRO.

## D. Additional results

***What features does* PI *look at?*** To understand what features different methods rely on, we plot the word importance on Beer Look in Figure 5. For the given input example, we evaluate the prediction change as we mask out each input token. We observe that only PI and Oracle ignore the spurious feature and predict the label correctly. Comparing to ERM, IRM and RGM focus more on the causal feature such as 'tiny'. However, they still heavily rely on the spurious feature.

| | ERM | | DRO | | IRM | | RGM | | Ours | |
|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy | Worst | Avg | Worst | Avg | Worst | Avg | Worst | Avg | Worst | Avg |
| Adenocarcinoma | 33.33 | 72.91 | 77.29 | 79.23 | 55.56 | 78.40 | 55.56 | 78.12 | 80.24 | 84.74 |
| Polyp syndrom | 44.44 | 74.63 | 77.29 | 78.79 | 55.56 | 76.31 | 66.67 | 78.73 | 69.23 | 81.28 |
| Brain cancer | 55.56 | 78.51 | 77.14 | 78.09 | 55.56 | 78.59 | 67.55 | 82.33 | 79.94 | 87.95 |
| Breast cancer | 66.49 | 80.47 | 75.00 | 78.84 | 66.87 | 80.56 | 64.38 | 79.81 | 80.32 | 83.12 |
| Colorectal cancer | 66.54 | 80.50 | 69.31 | 77.94 | 64.96 | 81.28 | 66.93 | 80.33 | 76.24 | 81.71 |
| Endometrial cancer | 66.98 | 80.60 | 76.19 | 80.21 | 66.03 | 82.60 | 66.98 | 81.77 | 80.32 | 83.26 |
| Gastric cancer | 62.96 | 79.94 | 76.95 | 81.65 | 62.96 | 80.03 | 59.26 | 78.87 | 79.44 | 85.92 |
| Hepatobiliary cancer | 44.44 | 73.01 | 60.00 | 73.89 | 55.56 | 77.19 | 55.56 | 76.22 | 60.00 | 78.94 |
| Kidney cancer | 16.67 | 66.66 | 50.00 | 68.70 | 33.33 | 73.07 | 33.33 | 71.31 | 50.00 | 74.76 |
| Lung cancer | 44.44 | 74.76 | 62.50 | 74.58 | 38.89 | 74.28 | 50.00 | 74.75 | 70.31 | 78.85 |
| Melanoma | 66.67 | 80.55 | 66.67 | 78.87 | 66.67 | 83.32 | 66.67 | 79.69 | 80.06 | 86.67 |
| Neoplasia | 50.00 | 75.98 | 33.33 | 69.10 | 33.33 | 71.97 | 50.00 | 75.18 | 70.00 | 80.06 |
| Ovarian cancer | 65.31 | 80.16 | 77.20 | 79.30 | 66.80 | 80.64 | 66.33 | 79.53 | 73.47 | 82.76 |
| Pancreatic cancer | 67.18 | 80.93 | 75.82 | 78.74 | 63.64 | 79.69 | 63.64 | 79.67 | 80.06 | 84.31 |
| Prostate cancer | 63.96 | 85.77 | 51.04 | 77.48 | 64.29 | 85.21 | 65.58 | 83.92 | 78.90 | 86.75 |
| Rectal cancer | 66.67 | 78.78 | 64.10 | 80.37 | 66.67 | 78.86 | 67.54 | 80.80 | 71.79 | 84.59 |
| Thyroid cancer | 50.00 | 77.18 | 75.00 | 83.06 | 66.86 | 84.05 | 67.73 | 82.56 | 80.23 | 87.85 |
| Average | 54.80 | 77.73 | 67.34 | 77.58 | 57.86 | 79.18 | 60.81 | 79.03 | 74.15 | 83.15 |

*Table 7.* Worst-group and average-group accuracy across 17 attributes on ASK2ME.

| method | input example |
|---|---|
| ERM | <art_positive> gold color with almost a surprisingly tiny head . |
| DRO | <art_positive> gold color with almost a surprisingly tiny head . |
| IRM | <art_positive> gold color with almost a surprisingly tiny head . |
| RGM | <art_positive> gold color with almost a surprisingly tiny head . |
| PI | <art_positive> gold color with almost a surprisingly tiny head . |
| Oracle | <art_positive> gold color with almost a surprisingly tiny head . |

*Figure 5.* Visualizing word importance on Beer Look. Only `PI` and `Oracle` ignore the artificial token and correctly predict the input as negative. We will add more examples in the update.

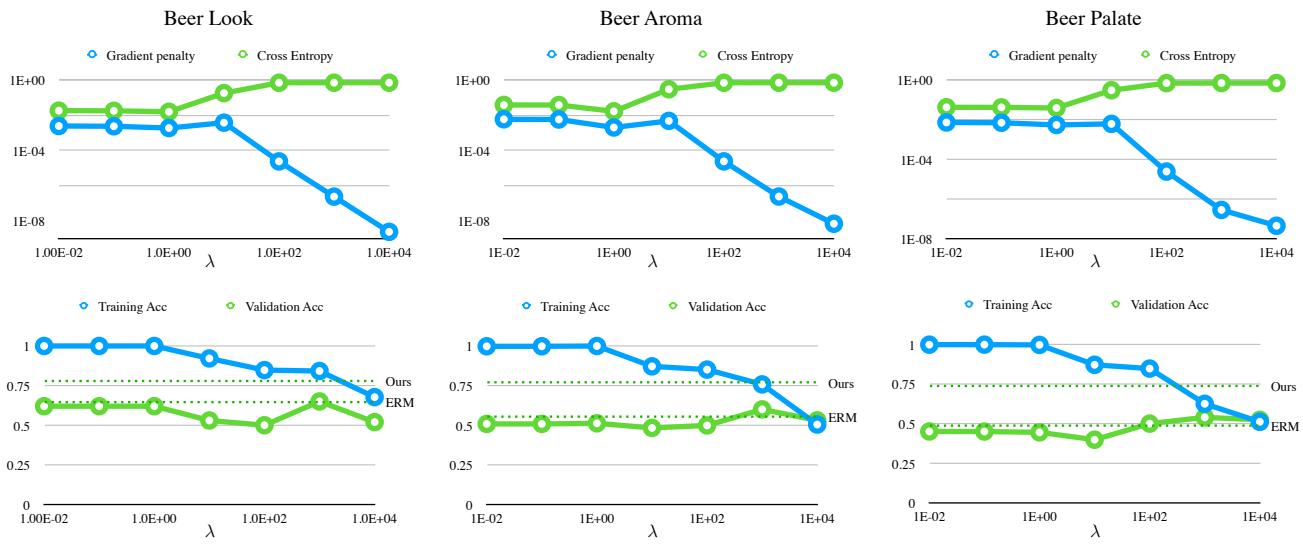


*Figure 6.* Performance of IRM as we adjust the weight of the gradient penalty. We observe that while the gradient penalty term is always orders of magnitude smaller than the cross entropy loss, the model is still able to overfit the unstable correlations in the training environments. As we further increase the penalty, the training & validation performance quickly drop to that of ERM.

| | ERM | | DRO | | IRM | | RGM | | Ours | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Worst | Avg | Worst | Avg | Worst | Avg | Worst | Avg | Worst | Avg |
| 5_o_Clock_Shadow | 53.33 | 81.51 | 90.00 | 92.05 | 80.00 | 85.93 | 66.67 | 87.22 | 83.33 | 89.65 |
| Arched_Eyebrows | 72.14 | 87.16 | 90.42 | 92.68 | 84.43 | 87.85 | 88.95 | 92.60 | 90.56 | 92.31 |
| Attractive | 67.21 | 85.84 | 90.77 | 92.22 | 82.57 | 87.28 | 86.64 | 91.94 | 89.98 | 91.86 |
| Bags_Under_Eyes | 72.46 | 86.16 | 90.59 | 92.04 | 81.34 | 86.84 | 88.52 | 92.13 | 89.12 | 92.12 |
| Bald | 75.98 | 91.23 | 91.73 | 93.04 | 71.39 | 82.21 | 91.50 | 94.81 | 91.68 | 93.42 |
| Bangs | 73.85 | 87.80 | 90.84 | 92.91 | 81.70 | 87.38 | 88.05 | 92.21 | 90.24 | 92.33 |
| Big_Lips | 73.46 | 87.14 | 90.59 | 92.55 | 84.16 | 87.87 | 89.54 | 92.65 | 90.52 | 92.17 |
| Big_Nose | 71.43 | 86.00 | 91.58 | 92.97 | 84.99 | 88.43 | 91.22 | 93.78 | 91.36 | 92.87 |
| Black_Hair | 75.98 | 90.91 | 89.62 | 93.77 | 78.63 | 89.16 | 90.66 | 94.02 | 88.10 | 93.30 |
| Blurry | 51.23 | 81.06 | 86.56 | 90.14 | 79.36 | 85.73 | 79.01 | 89.71 | 85.61 | 89.39 |
| Brown_Hair | 43.68 | 79.17 | 64.37 | 85.74 | 78.16 | 83.39 | 72.41 | 87.30 | 59.77 | 83.83 |
| Bushy_Eyebrows | 72.73 | 86.52 | 72.73 | 88.83 | 81.82 | 87.51 | 81.82 | 91.26 | 81.82 | 90.77 |
| Chubby | 9.52 | 70.69 | 61.90 | 84.63 | 76.19 | 82.91 | 47.62 | 82.32 | 71.43 | 86.59 |
| Double_Chin | 50.00 | 80.73 | 90.66 | 91.76 | 78.52 | 86.35 | 91.50 | 92.74 | 90.21 | 92.45 |
| Eyeglasses | 58.06 | 82.83 | 90.32 | 92.02 | 80.44 | 85.71 | 77.42 | 89.34 | 88.71 | 91.17 |
| Goatee | 0.00 | 68.26 | 0.00 | 70.08 | 84.80 | 90.83 | 91.50 | 95.66 | 91.63 | 94.59 |
| Gray_Hair | 60.71 | 82.53 | 69.08 | 87.73 | 42.60 | 76.20 | 85.71 | 89.56 | 68.26 | 88.18 |
| Heavy_Makeup | 66.06 | 85.68 | 89.69 | 92.22 | 84.18 | 87.20 | 84.43 | 91.49 | 90.01 | 91.86 |
| High_Cheekbones | 73.33 | 86.62 | 90.78 | 92.21 | 84.42 | 87.13 | 89.02 | 92.27 | 90.39 | 91.72 |
| Gender. | 46.67 | 80.14 | 85.56 | 90.87 | 74.44 | 83.93 | 70.00 | 87.73 | 90.56 | 91.52 |
| Mouth_Slightly_Open | 74.22 | 87.01 | 91.27 | 92.33 | 84.51 | 87.42 | 91.01 | 92.56 | 91.74 | 91.85 |
| Mustache | 50.00 | 80.89 | 91.72 | 95.38 | 50.00 | 78.58 | 91.50 | 95.97 | 91.60 | 94.93 |
| Narrow_Eyes | 69.23 | 85.54 | 90.05 | 91.85 | 82.94 | 87.00 | 88.46 | 91.84 | 91.69 | 91.90 |
| No_Beard | 39.39 | 78.10 | 84.85 | 90.97 | 72.73 | 83.80 | 57.58 | 85.00 | 84.85 | 90.43 |
| Oval_Face | 75.16 | 87.20 | 90.71 | 92.40 | 84.22 | 87.70 | 91.24 | 92.76 | 90.31 | 91.90 |
| Pale_Skin | 75.44 | 87.99 | 90.30 | 91.54 | 81.67 | 85.97 | 91.37 | 92.46 | 89.55 | 92.02 |
| Pointy_Nose | 73.34 | 87.18 | 91.19 | 92.42 | 84.87 | 87.69 | 89.29 | 92.55 | 91.07 | 92.00 |
| Receding_Hairline | 66.67 | 84.75 | 90.98 | 91.86 | 80.56 | 84.34 | 83.33 | 91.11 | 87.96 | 90.97 |
| Rosy_Cheeks | 74.90 | 88.17 | 91.40 | 93.32 | 84.88 | 88.59 | 90.55 | 93.00 | 91.49 | 92.71 |
| Sideburns | 38.46 | 77.84 | 84.62 | 90.69 | 76.92 | 84.14 | 76.92 | 89.72 | 91.35 | 93.75 |
| Smiling | 75.91 | 86.87 | 91.59 | 92.31 | 84.14 | 87.29 | 91.10 | 92.49 | 91.53 | 91.88 |
| Straight_Hair | 74.00 | 86.60 | 90.27 | 92.05 | 84.37 | 87.41 | 88.36 | 92.37 | 91.55 | 91.92 |
| Wavy_Hair | 74.22 | 86.88 | 91.41 | 92.40 | 84.15 | 87.41 | 88.81 | 92.21 | 91.64 | 91.88 |
| Wearing_Earrings | 75.36 | 86.77 | 91.67 | 92.63 | 84.78 | 87.70 | 90.88 | 92.55 | 91.51 | 92.19 |
| Wearing_Hat | 7.69 | 70.39 | 46.15 | 82.28 | 46.15 | 77.34 | 61.54 | 86.26 | 53.85 | 84.56 |
| Wearing_Lipstick | 59.37 | 83.61 | 89.47 | 91.88 | 82.53 | 86.01 | 79.37 | 90.10 | 90.32 | 91.57 |
| Wearing_Necklace | 74.57 | 87.26 | 91.09 | 92.47 | 82.26 | 87.42 | 89.53 | 92.17 | 90.73 | 92.21 |
| Wearing_Necktie | 25.00 | 74.52 | 90.00 | 91.56 | 80.00 | 84.31 | 35.00 | 79.32 | 91.44 | 92.53 |
| Young | 71.60 | 86.07 | 89.13 | 91.64 | 76.21 | 85.96 | 90.19 | 92.03 | 87.23 | 91.51 |
| Average | 60.06 | 83.63 | 84.25 | 90.83 | 78.51 | 85.79 | 82.52 | 90.95 | 87.04 | 91.41 |

*Table 8.* Worst-group and average-group accuracy for hair color prediction on CelebA.