A. Proof for 2IWIL

A.1. Proof of Theorem 4.1

Theorem. *The classification risk (6) can be equivalently expressed as*

$$R_{\mathrm{SC},\ell}(g) = \mathbb{E}_{x,r\sim q}[r(\ell(g(x)) - \ell(-g(x))) + (1-\beta)\ell(-g(x))] + \mathbb{E}_{x\sim p}[\beta\ell(-g(x))],$$

where $\beta \in [0,1]$ is an arbitrary weight.

Proof. Similar to Eq. (4), we may express $p(x|y=-1)$ by the Bayes' rule as

$$p(x|y=-1) = \frac{(1-r(x))p(x)}{1-\alpha}. \tag{11}$$

Consequently, the statement can be confirmed as follows:

$$\begin{aligned}
R_{\mathrm{SC},\ell}(g) &= \int \alpha p(x|y=+1)\ell(g(x)) + (1-\alpha)p(x|y=-1)\ell(-g(x))dx \\
&= \int \alpha\frac{r(x)p(x)}{\alpha}\ell(g(x)) + (1-\alpha)\frac{(1-r(x))p(x)}{1-\alpha}\ell(-g(x))dx && (\because \text{Eqs. (4) and (11)}) \\
&= \int p(x)r(x)\ell(g(x)) + p(x)(1-r(x))\ell(-g(x))dx \\
&= \int \{r\ell(g(x)) + (1-r)\ell(-g(x))\}\, q(x,r)dxdr \\
&= \mathbb{E}_{x,r\sim q}[r\ell(g(x)) + (1-r)\ell(-g(x))] \\
&= \mathbb{E}_{x,r\sim q}\left[r\ell(g(x)) + (1-r)\ell(-g(x)) + \underbrace{\beta\ell(-g(x)) - \beta\ell(-g(x))}_{=0}\right] \\
&= \mathbb{E}_{x,r\sim q}[r(\ell(g(x)) - \ell(-g(x))) + (1-\beta)\ell(-g(x))] + \mathbb{E}_{x\sim p}[\beta\ell(-g(x))].
\end{aligned}$$

□

A.2. Proof of Proposition 4.2

Proposition. *Let σ_{cov} denote the covariance between $n_{\mathrm{c}}^{-1}\sum_{i=1}^{n_{\mathrm{c}}} r_i\{\ell(g(x_{\mathrm{c},i})) - \ell(-g(x_{\mathrm{c},i}))\}$ and $n_{\mathrm{c}}^{-1}\sum_{i=1}^{n_{\mathrm{c}}}\ell(-g(x_{\mathrm{c},i}))$. For a fixed g, the estimator $\widehat{R}_{\mathrm{SC},\ell}(g)$ of Eq. (7) has the minimum variance when $\beta = \frac{n_{\mathrm{u}}}{n_{\mathrm{c}}+n_{\mathrm{u}}} + \frac{\sigma_{\mathrm{cov}}}{\mathrm{Var}(\ell(-g(x)))}\frac{n_{\mathrm{c}}n_{\mathrm{u}}}{n_{\mathrm{c}}+n_{\mathrm{u}}}$ among estimators in the form of Eq. (7) for $\beta \in [0,1]$.*

Proof. Let

$$\begin{aligned}
\mu &\triangleq \mathbb{E}_{\mathcal{D}_{\mathrm{c}},\mathcal{D}_{\mathrm{u}}}[\widehat{R}_{\mathrm{SC},\ell}(g)], \\
\mu_1 &\triangleq \mathbb{E}_{\mathcal{D}_{\mathrm{c}}}\left[\frac{1}{n_{\mathrm{c}}}\sum_{i=1}^{n_{\mathrm{c}}}\ell(-g(x_{\mathrm{c},i}))\right] = \mathbb{E}_{\mathcal{D}_{\mathrm{u}}}\left[\frac{1}{n_{\mathrm{u}}}\sum_{i=1}^{n_{\mathrm{u}}}\ell(-g(x_{\mathrm{u},i}))\right] = \mathbb{E}_{x\sim p}[\ell(-g(x))], \\
w_1 &\triangleq \mathbb{E}_{\mathcal{D}_{\mathrm{c}}}\left[\frac{1}{n_{\mathrm{c}}}\sum_{i=1}^{n_{\mathrm{c}}} r(x_i)(\ell(g(x_{\mathrm{c},i})) - \ell(-g(x_{\mathrm{c},i})))\right], \\
w_2 &\triangleq \mathbb{E}_{\mathcal{D}_{\mathrm{c}}}\left[\left(\frac{1}{n_{\mathrm{c}}}\sum_{i=1}^{n_{\mathrm{c}}} r(x_i)(\ell(g(x_{\mathrm{c},i})) - \ell(-g(x_{\mathrm{c},i})))\right)^2\right], \\
\lambda &\triangleq \mathbb{E}_{\mathcal{D}_{\mathrm{c}}}\left[\left(\frac{1}{n_{\mathrm{c}}}\sum_{i=1}^{n_{\mathrm{c}}} r(x_i)(\ell(g(x_{\mathrm{c},i}) - \ell(-g(x_{\mathrm{c},i}))))\right)\left(\frac{1}{n_{\mathrm{c}}}\sum_{i=1}^{n_{\mathrm{c}}}\ell(-g(x_{\mathrm{c},i}))\right)\right], \\
\sigma_{\mathrm{cov}} &\triangleq \mathrm{Cov}\left(\frac{1}{n_{\mathrm{c}}}\sum_{i=1}^{n_{\mathrm{c}}} r_i(\ell(g(x_{\mathrm{c},i}) - \ell(-g(x_{\mathrm{c},i})))), \frac{1}{n_{\mathrm{c}}}\sum_{i=1}^{n_{\mathrm{c}}}\ell(-g(x_{\mathrm{c},i}))\right) = \lambda - w_1\mu_1
\end{aligned}$$

We may represent $\mathbb{E}_{\mathcal{D}_\mathrm{c}}\left[\left(\frac{1}{n_\mathrm{c}}\sum_{i=1}^{n_\mathrm{c}}\ell(-g(x_{\mathrm{c},i}))\right)^2\right]$ in terms of $\mathrm{Var}(\ell(-g(x)))$ and μ_1:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_\mathrm{c}}\left[\left(\frac{1}{n_\mathrm{c}}\sum_{i=1}^{n_\mathrm{c}}\ell(-g(x_{\mathrm{c},i}))\right)^2\right]=&\frac{1}{n_\mathrm{c}^2}\mathbb{E}_{\mathcal{D}_\mathrm{c}}\left[\sum_{i=1}^{n_\mathrm{c}}\ell(-g(x_{\mathrm{c},i}))^2+2\sum_{i=1}^{n_\mathrm{c}}a\sum_{j=1}^{i-1}\ell(-g(x_{\mathrm{c},i}))\ell(-g(x_{\mathrm{c},j}))\right]\\=&\frac{1}{n_\mathrm{c}^2}\left(n_\mathrm{c}\mathbb{E}_{x\sim p}\left[\ell(-g(x))^2\right]+n_\mathrm{c}(n_\mathrm{c}-1)\mathbb{E}_{x\sim p}\left[\ell(-g(x))\right]^2\right)\\=&\frac{1}{n_\mathrm{c}}\mathrm{Var}(\ell(-g(x)))+\mu_1^2.\end{aligned}$$

Similarly, we obtain $\mathbb{E}_{\mathcal{D}_\mathrm{u}}[(\frac{1}{n_\mathrm{u}}\sum_{i=1}^{n_\mathrm{u}}\ell(-g(x_{\mathrm{u},i})))^2]=n_\mathrm{u}^{-1}\mathrm{Var}(\ell(-g(x)))+\mu_1^2$. As a result,

$$\begin{aligned}&\mathrm{Var}(\widehat{R}_{\mathrm{SC},\ell}(g))\\&=\mathbb{E}_{\mathcal{D}_\mathrm{c},\mathcal{D}_\mathrm{u}}\left[\left(\widehat{R}_{\mathrm{SC},\ell}(g)\right)^2\right]-\mu^2\\&=\mathbb{E}_{\mathcal{D}_\mathrm{c},\mathcal{D}_\mathrm{u}}\left[\underbrace{\frac{1}{n_\mathrm{c}}\sum_{i=1}^{n_\mathrm{c}}r_i(\ell(g(x_{\mathrm{c},i}))-\ell(-g(x_{\mathrm{c},i})))}_{\text{(A)}}+(1-\beta)\underbrace{\frac{1}{n_\mathrm{c}}\sum_{i=1}^{n_\mathrm{c}}\ell(-g(x_{\mathrm{c},i}))}_{\text{(B)}}+\beta\underbrace{\frac{1}{n_\mathrm{u}}\sum_{i=1}^{n_\mathrm{u}}\ell(-g(x_{\mathrm{u},i}))}_{\text{(C)}}\right]-\mu^2\\&=\underbrace{w_2}_{\text{(A)}^2}+2(1-\beta)\underbrace{\lambda}_{\text{(A)(B)}}+2\beta\underbrace{w_1\mu_1}_{\text{(A)(C)}}+(1-\beta)^2\underbrace{\left(\frac{1}{n_\mathrm{c}}\mathrm{Var}(\ell(-g(x)))+\mu_1^2\right)}_{\text{(B)}^2}\\&\qquad+2(1-\beta)\beta\underbrace{\mu_1^2}_{\text{(B)(C)}}+\beta^2\underbrace{\left(\frac{1}{n_\mathrm{u}}\mathrm{Var}(\ell(-g(x)))+\mu_1^2\right)}_{\text{(C)}^2}-\mu^2\\&=\underbrace{\left(w_2+2\lambda-\mu^2+\frac{1}{n_\mathrm{c}}\mathrm{Var}(\ell(-g(x)))+\mu_1^2\right)}_{\text{const.w.r.t.}\beta}-2\left(\frac{\mathrm{Var}(\ell(-g(x)))}{n_\mathrm{c}}+\sigma_\mathrm{cov}\right)\beta+\mathrm{Var}(\ell(-g(x)))\left(\frac{n_\mathrm{c}+n_\mathrm{u}}{n_\mathrm{c}n_\mathrm{u}}\right)\beta^2\\&=\mathrm{Var}(\ell(-g(x)))\left(\frac{n_\mathrm{c}+n_\mathrm{u}}{n_\mathrm{c}n_\mathrm{u}}\right)\left(\beta-\left(\frac{n_\mathrm{u}}{n_\mathrm{c}+n_\mathrm{u}}+\frac{\sigma_\mathrm{cov}}{\mathrm{Var}(\ell(-g(x)))}\frac{n_\mathrm{c}n_\mathrm{u}}{n_\mathrm{c}+n_\mathrm{u}}\right)\right)^2+\mathrm{const.}\end{aligned}$$

Since $\mathrm{Var}(\ell(-g(x)))\left(\frac{n_\mathrm{c}+n_\mathrm{u}}{n_\mathrm{c}n_\mathrm{u}}\right)\geq 0$, and $\beta\in[0,1]$, $\mathrm{Var}(\widehat{R}_{\mathrm{SC},\ell}(g))$ is minimized when $\beta=\mathrm{clip}_{[0,1]}\left(\frac{n_\mathrm{u}}{n_\mathrm{c}+n_\mathrm{u}}+\frac{\sigma_\mathrm{cov}}{\mathrm{Var}(\ell(-g(x)))}\frac{n_\mathrm{c}n_\mathrm{u}}{n_\mathrm{c}+n_\mathrm{u}}\right)$. Note that $\mathrm{clip}_{[l,u]}(v)=\min\{\max\{v,l\},u\}$. □

A.3. Proof of Theorem 4.3

Theorem. *Let $\mathcal{G}$ be the hypothesis class we use. Assume that the loss function ℓ is ρ_ℓ-Lipschitz continuous, and that there exists a constant $C_\ell>0$ such that $\sup_{x\in\mathcal{X},y\in\{\pm1\}}|\ell(yg(x))|\leq C_\ell$ for any $g\in\mathcal{G}$. Let $\widehat{g}\triangleq\arg\min_{g\in\mathcal{G}}\widehat{R}_{\mathrm{SC},\ell}(g)$ and $g^*\triangleq\arg\min_{g\in\mathcal{G}}R_{\mathrm{SC},\ell}(g)$. For $\delta\in(0,1)$, with probability at least $1-\delta$ over repeated sampling of data for training $\widehat{g}$,*

$$R_{\mathrm{SC},\ell}(\widehat{g})-R_{\mathrm{SC},\ell}(g^*)\leq 16\rho_\ell((3-\beta)\mathfrak{R}_{n_\mathrm{c}}(\mathcal{G})+\beta\mathfrak{R}_{n_\mathrm{u}}(\mathcal{G}))+4C_\ell\sqrt{\frac{\log(8/\delta)}{2}}\left((3-\beta)n_\mathrm{c}^{-\frac{1}{2}}+\beta n_\mathrm{u}^{-\frac{1}{2}}\right).$$

Proof. Note that $\widehat{g}$ and g^* are the minimizers of $\widehat{R}_{\mathrm{SC},\ell}(g)$ and $R_{\mathrm{SC},\ell}(g)$, respectively. Then,

$$\begin{aligned}
R_{\mathrm{SC},\ell}(\widehat{g}) - R_{\mathrm{SC},\ell}(g^*) &= R_{\mathrm{SC},\ell}(\widehat{g}) - \widehat{R}_{\mathrm{SC},\ell}(\widehat{g}) + \widehat{R}_{\mathrm{SC},\ell}(\widehat{g}) - \widehat{R}_{\mathrm{SC},\ell}(g^*) + \widehat{R}_{\mathrm{SC},\ell}(g^*) - R_{\mathrm{SC},\ell}(g^*) \\
&\leq \sup_{g\in\mathcal{G}} \left(R_{\mathrm{SC},\ell}(g) - \widehat{R}_{\mathrm{SC},\ell}(g) \right) + 0 + \sup_{g\in\mathcal{G}} \left(\widehat{R}_{\mathrm{SC},\ell}(g) - R_{\mathrm{SC},\ell}(g) \right) \\
&\leq 2 \sup_{g\in\mathcal{G}} \left| \widehat{R}_{\mathrm{SC},\ell}(g) - R_{\mathrm{SC},\ell}(g) \right|.
\end{aligned}$$

From now on, our goal is to bound the uniform deviation $\sup_{g\in\mathcal{G}} \left| \widehat{R}_{\mathrm{SC},\ell}(g) - R_{\mathrm{SC},\ell}(g) \right|$. Since

$$\begin{aligned}
&\sup_{g\in\mathcal{G}} \left| \widehat{R}_{\mathrm{SC},\ell}(g) - R_{\mathrm{SC},\ell}(g) \right| \\
&\quad\leq \sup_{g\in\mathcal{G}} \left| \frac{1}{n_{\mathrm{c}}} \sum_{i=1}^{n_{\mathrm{c}}} \{ r_i(\ell(g(x_{\mathrm{c},i})) - \ell(-g(x_{\mathrm{c},i}))) + (1-\beta)\ell(-g(x_{\mathrm{c},i})) \} \right. \\
&\qquad\qquad \left. - \mathbb{E}_{x,r\sim q} \left[r(\ell(g(x)) - \ell(-g(x))) + (1-\beta)\ell(-g(x)) \right] \right| \\
&\qquad + \beta \sup_{g\in\mathcal{G}} \left| \frac{1}{n_{\mathrm{u}}} \sum_{i=1}^{n_{\mathrm{u}}} \ell(-g(x_{\mathrm{u},i})) - \mathbb{E}_{x\sim p} \left[\ell(-g(x)) \right] \right| \\
&\quad\leq \sup_{g\in\mathcal{G}} \left| \frac{1}{n_{\mathrm{c}}} \sum_{i=1}^{n_{\mathrm{c}}} r_i \ell(g(x_{\mathrm{c},i})) - \mathbb{E}_{x,r\sim q}[r\ell(g(x))] \right| + \sup_{g\in\mathcal{G}} \left| \frac{1}{n_{\mathrm{c}}} \sum_{i=1}^{n_{\mathrm{c}}} r_i \ell(-g(x_{\mathrm{c},i})) - \mathbb{E}_{x,r\sim q}[r\ell(-g(x))] \right| \\
&\qquad + (1-\beta) \sup_{g\in\mathcal{G}} \left| \frac{1}{n_{\mathrm{c}}} \sum_{i=1}^{n_{\mathrm{c}}} \ell(-g(x_{\mathrm{c},i})) - \mathbb{E}_{x,r\sim q}[\ell(-g(x))] \right| + \beta \sup_{g\in\mathcal{G}} \left| \frac{1}{n_{\mathrm{u}}} \sum_{i=1}^{n_{\mathrm{u}}} \ell(-g(x_{\mathrm{u},i})) - \mathbb{E}_{x\sim p}[\ell(-g(x))] \right|,
\end{aligned} \tag{12}$$

all we need to do is to bound four terms appearing in the RHS independently, which can be done by McDiarmid's inequality (McDiarmid, 1989). For the first term, since $\sum_{i=1}^{n_{\mathrm{c}}} r_i \ell(g(x_{\mathrm{c},i})) - \mathbb{E}_{x,r\sim q}[r\ell(g(x))]$ is the bounded difference with a constant C_L/n_{c} for every replacement of $x_{\mathrm{c},i}$, McDiarmid's inequality state that

$$\begin{aligned}
\Pr \left[\sup_{g\in\mathcal{G}} \left(\frac{1}{n_{\mathrm{c}}} \sum_{i=1}^{n_{\mathrm{c}}} r_i \ell(g(x_{\mathrm{c},i})) - \mathbb{E}_{x,r\sim q}[r\ell(g(x))] \right) - \mathbb{E} \left[\sup_{g\in\mathcal{G}} \left(\frac{1}{n_{\mathrm{c}}} \sum_{i=1}^{n_{\mathrm{c}}} r_i \ell(g(x_{\mathrm{c},i})) - \mathbb{E}_{x,r\sim q}[r\ell(g(x))] \right) \right] \geq \varepsilon \right] \\
\leq \exp \left(-\frac{2\varepsilon^2}{C_L^2/n_{\mathrm{c}}} \right),
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
&\sup_{g\in\mathcal{G}} \left(\frac{1}{n_{\mathrm{c}}} \sum_{i=1}^{n_{\mathrm{c}}} r_i \ell(g(x_{\mathrm{c},i})) - \mathbb{E}_{x,r\sim q}[r\ell(g(x))] \right) \\
&\qquad \leq \mathbb{E} \left[\sup_{g\in\mathcal{G}} \left(\frac{1}{n_{\mathrm{c}}} \sum_{i=1}^{n_{\mathrm{c}}} r_i \ell(g(x_{\mathrm{c},i})) - \mathbb{E}_{x,r\sim q}[r\ell(g(x))] \right) \right] + C_L \sqrt{\frac{\log(8/\delta)}{2n_{\mathrm{c}}}},
\end{aligned}$$

with probability at least $1 - \delta/8$. Following the symmetrization device (Lemma 6.3 in Ledoux & Talagrand (1991)) and Ledoux-Talagrand's contraction inequality (Theorem 4.12 in Ledoux & Talagrand (1991)), we obtain

$$\begin{aligned}
\mathbb{E} \left[\sup_{g\in\mathcal{G}} \left(\frac{1}{n_{\mathrm{c}}} \sum_{i=1}^{n_{\mathrm{c}}} r_i \ell(g(x_{\mathrm{c},i})) - \mathbb{E}_{x,r\sim q}[r\ell(g(x))] \right) \right] &\leq 2\mathfrak{R}_{n_{\mathrm{c}}}(\ell \circ \mathcal{G}) && \text{(symmetrization)} \\
&\leq 4\rho_L \mathfrak{R}_{n_{\mathrm{c}}}(\mathcal{G}) && \text{(contraction)}.
\end{aligned}$$

Note that $0 \leq r_i \leq 1$ for $i = 1, \ldots, n_{\mathrm{c}}$. Thus, one-sided uniform deviation bound is obtained: with probability at least $1 - \delta/8$,

$$\sup_{g\in\mathcal{G}} \left(\frac{1}{n_{\mathrm{c}}} \sum_{i=1}^{n_{\mathrm{c}}} r_i \ell(g(x_{\mathrm{c},i})) - \mathbb{E}_{x,r\sim q}[r\ell(g(x))] \right) \leq 4\rho_L \mathfrak{R}_{n_{\mathrm{c}}}(\mathcal{G}) + C_L \sqrt{\frac{\log(8/\delta)}{2n_{\mathrm{c}}}}.$$

Applying it twice, the two-sided uniform deviation bound is obtained: with probability at least $1-\delta/4$,

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{n_{\mathrm{c}}}\sum_{i=1}^{n_{\mathrm{c}}} r_i\ell(g(x_{\mathrm{c},i}))-\mathbb{E}_{x,r\sim q}[r\ell(g(x))]\right|\leq 8\rho_L\mathfrak{R}_{n_{\mathrm{c}}}(\mathcal{G})+2C_L\sqrt{\frac{\log(8/\delta)}{2n_{\mathrm{c}}}}.$$

Similarly, the remaining three terms in the RHS of Eq. (12) can be bounded. Since the second, third, and fourth terms are the bounded differences with constants C_L/n_{c}, C_L/n_{c}, and C_L/n_{u}, respectively, the following inequalities hold with probability at least $1-\delta/4$:

$$\begin{aligned}
\sup_{g\in\mathcal{G}}\left|\frac{1}{n_{\mathrm{c}}}\sum_{i=1}^{n_{\mathrm{c}}} r_i\ell(-g(x_{\mathrm{c},i}))-\mathbb{E}_{x,r\sim q}[r\ell(-g(x))]\right|&\leq 8\rho_L\mathfrak{R}_{n_{\mathrm{c}}}(\mathcal{G})+2C_L\sqrt{\frac{\log(8/\delta)}{2n_{\mathrm{c}}}},\\
\sup_{g\in\mathcal{G}}\left|\frac{1}{n_{\mathrm{c}}}\sum_{i=1}^{n_{\mathrm{c}}} \ell(-g(x_{\mathrm{c},i}))-\mathbb{E}_{x,r\sim q}[\ell(-g(x))]\right|&\leq 8\rho_L\mathfrak{R}_{n_{\mathrm{c}}}(\mathcal{G})+2C_L\sqrt{\frac{\log(8/\delta)}{2n_{\mathrm{c}}}},\\
\sup_{g\in\mathcal{G}}\left|\frac{1}{n_{\mathrm{u}}}\sum_{i=1}^{n_{\mathrm{u}}} \ell(-g(x_{\mathrm{u},i}))-\mathbb{E}_{x\sim p}[\ell(-g(x))]\right|&\leq 8\rho_L\mathfrak{R}_{n_{\mathrm{u}}}(\mathcal{G})+2C_L\sqrt{\frac{\log(8/\delta)}{2n_{\mathrm{u}}}}.
\end{aligned}$$

After all, we can bound the original estimation error: with probability at least $1-\delta$,

$$R_{\mathrm{SC},\ell}(\widehat{g})-R_{\mathrm{SC},\ell}(g^*)\leq 16\rho_L((3-\beta)\mathfrak{R}_{n_{\mathrm{c}}}(\mathcal{G})+\beta\mathfrak{R}_{n_{\mathrm{u}}}(\mathcal{G}))+4C_L\sqrt{\frac{\log(8/\delta)}{2}}\left((3-\beta)n_{\mathrm{c}}^{-\frac{1}{2}}+\beta n_{\mathrm{u}}^{-\frac{1}{2}}\right).$$

□

B. Proof for IC-GAIL

B.1. Proof of Theorem 4.4

Theorem. *Denote that*

$$V(\pi_\theta,D_w)=\mathbb{E}_{x\sim p}[\log(1-D_w(x))]+\mathbb{E}_{x\sim p'}[\log D_w(x)],$$

and that $C(\pi_\theta)=\max_w V(\pi_\theta,D_w)$. *Then,* $V(\pi_\theta,D_w)$ *is maximized when* $D_w=\frac{p'}{p+p'}(\triangleq D_w^*)$, *and its maximum value is* $C(\pi_\theta)=-\log 4+2\mathrm{JSD}(p\|p')$. *Thus,* $C(\pi_\theta)$ *is minimized if and only if* $p_\theta=p_{\mathrm{opt}}$ *almost everywhere.*

Proof. Given a fixed agent policy π_θ, the discriminator maximize the quantity $V(\pi_\theta,D_w)$, which can be rewritten in the same way we did in Eq. (13), such as

$$\begin{aligned}
V(\pi_\theta,D_w)&=\mathbb{E}_{x\sim p}[\log(1-D_w(x))]+\mathbb{E}_{x\sim p'}[\log D_w(x)]\\
&=\int p'(x)\log D_w(x)+p(x)\log(1-D_w(x))dx.
\end{aligned}$$

This maximum is achieved when $D_w(x)=D_{w^*}(x)=\frac{p'(x)}{p'(x)+p(x)}$, with the same discussion as Proposition 1 in Goodfellow et al. (2014). As a result, we may derive $\max_w V(\pi_\theta,D_w)$ with $D_w^*(x)$,

$$C(\pi_\theta)=V(\pi_\theta,D_w^*)=\mathbb{E}_{x\sim p}\left[\log\frac{p}{p'+p}\right]+\mathbb{E}_{x\sim p'}\left[\log\frac{p'}{p'+p}\right],$$

where $p'=\alpha p_\theta+(1-\alpha)p_{\mathrm{non}}$. Note that $C(\pi_\theta)=\mathbb{E}_{x\sim p}[\log\frac{1}{2}]+\mathbb{E}_{x\sim p'}[\log\frac{1}{2}]=-\log 4$ when $p'=p$. We may rewrite $C(\pi_\theta)$ as follows:

$$\begin{aligned}
C(\pi_\theta)=&\mathbb{E}_{x\sim p}\left[\log\frac{p}{p'+p}\right]+\mathbb{E}_{x\sim p'}\left[\log\frac{p'}{p'+p}\right]\\
=&-\log 4+\mathbb{E}_{x\sim p}\left[\log\frac{p'}{(p'+p)/2}\right]+\mathbb{E}_{x\sim p'}\left[\log\frac{p}{(p'+p)/2}\right]\\
=&-\log 4+2\mathrm{JSD}(p\|p'),
\end{aligned}$$

where $\mathrm{JSD}(p_1 \| p_2) \triangleq \frac{1}{2}\mathbb{E}_{p_1}[\log \frac{p_1}{(p_1+p_2)/2}] + \frac{1}{2}\mathbb{E}_{p_2}[\log \frac{p_2}{(p_1+p_2)/2}]$ is Jensen-Shannon divergence. Since Jensen-Shannon divergence is greater or equal to zero and it is minimized and only if $p' = p$, we obtain that $C(\pi_\theta)$ is minimized if and only if

$$\begin{aligned} p' = p &\Rightarrow \alpha p_\theta + (1-\alpha)p_{\mathrm{non}} = \alpha p_{\mathrm{opt}} + (1-\alpha)p_{\mathrm{non}} \text{ almost everywhere} \\ &\Rightarrow p_\theta = p_{\mathrm{opt}} \text{ almost everywhere.} \end{aligned}$$

□

B.2. Proof of Theorem 4.5

Theorem. *$V(\pi_\theta, D_w)$ can be transformed to $\widetilde{V}(\pi_\theta, D_w)$, which is defined as follows:*

$$\widetilde{V}(\pi_\theta, D_w) = \mathbb{E}_{x\sim p}[\log(1 - D_w(x))] + \alpha\mathbb{E}_{x\sim p_\theta}[\log D_w(x)] + \mathbb{E}_{x,r\sim q}[(1-r)\log D_w(x)].$$

Proof. The statement can be confirmed as follows:

$$\begin{aligned} &\mathbb{E}_{x\sim p}[\log(1 - D_w(x))] + \mathbb{E}_{x\sim p'}[\log D_w(x)] \\ &\quad= \mathbb{E}_{x\sim p}[\log(1 - D_w(x))] + \alpha\mathbb{E}_{x\sim p_\theta}[\log D_w(x)] + (1-\alpha)\mathbb{E}_{x\sim p_{\mathrm{non}}}[\log D_w(x)] \\ &\quad= \mathbb{E}_{x\sim p}[\log(1 - D_w(x))] + \alpha\mathbb{E}_{x\sim p_\theta}[\log D_w(x)] + (1-\alpha)\mathbb{E}_{x,r\sim q}\left[\frac{1-r}{1-\alpha}\log D_w(x)\right] \\ &\quad= \mathbb{E}_{x\sim p}[\log(1 - D_w(x))] + \alpha\mathbb{E}_{x\sim p_\theta}[\log D_w(x)] + \mathbb{E}_{x,r\sim q}[(1-r)\log D_w(x)], \end{aligned} \tag{13}$$

where the first identity comes from the definition $p' = \alpha p_\theta + (1-\alpha)p_{\mathrm{non}}$, and the second identity holds since

$$\begin{aligned} \mathbb{E}_{x\sim p_{\mathrm{non}}}[\log D_w(x)] &= \int \log D_w(x) p_{\mathrm{non}}(x) dx \\ &= \int \log D_w(x) \frac{1-r(x)}{1-\alpha} p(x) dx \qquad \text{(note } p_{\mathrm{non}}(x) = p(x|y=-1)\text{)} \\ &= \int \log D_w(x) \frac{1-r}{1-\alpha} q(x,r) dx dr \\ &= \mathbb{E}_{x,r\sim q}\left[\frac{1-r}{1-\alpha}\log D_w(x)\right]. \end{aligned}$$

□

B.3. Proof of Theorem 4.6

Theorem. *Let $\mathcal{W}$ be a parameter space for training the discriminator and $D_{\mathcal{W}} \triangleq \{D_w \mid w \in \mathcal{W}\}$ be its hypothesis space. Assume that $\max\{\sup_{x\in\mathcal{X}, w\in\mathcal{W}} |\log D_w(x)|, \sup_{x\in\mathcal{X}, w\in\mathcal{W}} |\log(1 - D_w(x))|\} \le C_L$, and that $\max\{\sup_{w\in\mathcal{W}} |\log D_w(x) - \log D_w(x')|, \sup_{w\in\mathcal{W}} |\log(1-D_w(x)) - \log(1-D_w(x'))|\} \le \rho_L |x - x'|$ for any $x, x' \in \mathcal{X}$. For a fixed agent policy π_θ, let $D_{\widehat{w}} \triangleq \arg\max_{w\in\mathcal{W}} \widehat{V}(\pi_\theta, D_w)$ and $D_{w^*} \triangleq \arg\max_{w\in\mathcal{W}} V(\pi_\theta, D_w)$. For $\delta \in (0,1)$, with probability at least $1-\delta$ over repeated sampling of data for training $D_{\widehat{w}}$,*

$$V(\pi_\theta, D_{w^*}) - V(\pi_\theta, D_{\widehat{w}}) \le 16\rho_L(\mathfrak{R}_{n_{\mathrm{u}}}(D_{\mathcal{W}}) + \alpha\mathfrak{R}_{n_{\mathrm{a}}}(D_{\mathcal{W}}) + \mathfrak{R}_{n_{\mathrm{c}}}(D_{\mathcal{W}})) + 4C_L\sqrt{\frac{\log(6/\delta)}{2}}\left(n_{\mathrm{u}}^{-\frac{1}{2}} + \alpha n_{\mathrm{a}}^{-\frac{1}{2}} + n_{\mathrm{c}}^{-\frac{1}{2}}\right).$$

Proof. Denote $\mathcal{V}(w) \triangleq V(\pi_\theta, D_w)$ and $\widehat{\mathcal{V}}(w) \triangleq \widehat{V}(\pi_\theta, D_w)$. Note that $\widehat{w}$ and w^* are the minimizers of $\mathcal{V}(w)$ and $\widehat{\mathcal{V}}(w)$, respectively. Then,

$$\begin{aligned} \mathcal{V}(w^*) - \mathcal{V}(\widehat{w}) &= \mathcal{V}(w^*) - \widehat{\mathcal{V}}(w^*) + \widehat{\mathcal{V}}(w^*) - \widehat{\mathcal{V}}(\widehat{w}) + \widehat{\mathcal{V}}(\widehat{w}) - \mathcal{V}(\widehat{w}) \\ &\le \sup_{w\in\mathcal{W}}\left(\mathcal{V}(w) - \widehat{\mathcal{V}}(w)\right) + 0 + \sup_{w\in\mathcal{W}}\left(\widehat{\mathcal{V}}(w) - \mathcal{V}(w)\right) \\ &\le 2\sup_{w\in\mathcal{W}}\left|\widehat{\mathcal{V}}(w) - \mathcal{V}(w)\right|. \end{aligned}$$

From now on, our goal is to bound the uniform deviation $\sup_{w\in\mathcal{W}}\left|\widehat{\mathcal{V}}(w)-\mathcal{V}(w)\right|$. Since

$$\begin{aligned}\sup_{w\in\mathcal{W}}\left|\widehat{\mathcal{V}}(w)-\mathcal{V}(w)\right| \leq & \sup_{w\in\mathcal{W}}\left|\frac{1}{n_{\mathrm{u}}}\sum_{i=1}^{n_{\mathrm{u}}}\log(1-D_w(x_{\mathrm{u},i}))-\mathbb{E}_{x\sim p}\left[\log(1-D_w(x))\right]\right| \\ & +\alpha\sup_{w\in\mathcal{W}}\left|\frac{1}{n_{\mathrm{a}}}\sum_{i=1}^{n_{\mathrm{a}}}\log D_w(x_{\mathrm{a},i})-\mathbb{E}_{x\sim p_\theta}\left[\log D_w(x)\right]\right| \\ & +\sup_{w\in\mathcal{W}}\left|\frac{1}{n_{\mathrm{c}}}\sum_{i=1}^{n_{\mathrm{c}}}(1-r_i)\log D_w(x_{\mathrm{c},i})-\mathbb{E}_{x,r\sim q}\left[(1-r)\log D_w(x)\right]\right|, \end{aligned} \tag{14}$$

three terms appearing in the RHS must be bounded independently, utilizing McDiarmid's inequality (McDiarmid, 1989). For the first term, since $\sum_{i=1}^{n_{\mathrm{u}}}\log(1-D_w(x_{\mathrm{u},i}))-\mathbb{E}_{x\sim p}[\log(1-D_w(x))]$ has the bounded difference property with a constant C_L/n_{u} for every replacement of $x_{\mathrm{u},i}$, we can conclude by McDiarmid's inequality that

$$\begin{aligned}\Pr\Bigg[\sup_{w\in\mathcal{W}}\Bigg(\sum_{i=1}^{n_{\mathrm{u}}}\log(1-D_w(x_{\mathrm{u},i}))-\mathbb{E}_{x\sim p}[\log(1-D_w(x))]\Bigg) \\ -\mathbb{E}\left[\sup_{w\in\mathcal{W}}\sum_{i=1}^{n_{\mathrm{u}}}\log(1-D_w(x_{\mathrm{u},i}))-\mathbb{E}_{x\sim p}[\log(1-D_w(x))]\right]\geq\varepsilon\Bigg]\leq\exp\left(-\frac{2\varepsilon^2}{C_L^2/n_{\mathrm{u}}}\right),\end{aligned}$$

which is equivalent to

$$\begin{aligned}&\sup_{w\in\mathcal{W}}\left(\sum_{i=1}^{n_{\mathrm{u}}}\log(1-D_w(x_{\mathrm{u},i}))-\mathbb{E}_{x\sim p}[\log(1-D_w(x))]\right) \\ &\quad\leq\mathbb{E}\left[\sup_{w\in\mathcal{W}}\sum_{i=1}^{n_{\mathrm{u}}}\log(1-D_w(x_{\mathrm{u},i}))-\mathbb{E}_{x\sim p}[\log(1-D_w(x))]\right]+C_L\sqrt{\frac{\log(6/\delta)}{2n_{\mathrm{u}}}},\end{aligned}$$

with probability at least $1-\delta/6$. Following symmetrization device (Lemma 6.3 in Ledoux & Talagrand (1991)) and Ledoux-Talagrand's contraction inequality (Theorem 4.12 in Ledoux & Talagrand (1991)), we obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{w\in\mathcal{W}}\sum_{i=1}^{n_{\mathrm{u}}}\log(1-D_w(x_{\mathrm{u},i}))-\mathbb{E}_{x\sim p}[\log(1-D_w(x))]\right] &\leq 2\mathfrak{R}_{n_{\mathrm{u}}}(\log\circ D_{\mathcal{W}}) && \text{(symmetrization)} \\ &\leq 4\rho_L\mathfrak{R}_{n_{\mathrm{u}}}(D_{\mathcal{W}}). && \text{(contraction inequality)}\end{aligned}$$

Thus, one-sided uniform deviation bound is obtained: with probability at least $1-\delta/6$,

$$\sup_{w\in\mathcal{W}}\left(\sum_{i=1}^{n_{\mathrm{u}}}\log(1-D_w(x_{\mathrm{u},i}))-\mathbb{E}_{x\sim p}[\log(1-D_w(x))]\right)\leq 4\rho_L\mathfrak{R}_{n_{\mathrm{u}}}(D_{\mathcal{W}})+C_L\sqrt{\frac{\log(6/\delta)}{2n_{\mathrm{u}}}}.$$

Applying it twice, the two-sided uniform deviation bound is obtained: with probability at least $1-\delta/3$,

$$\sup_{w\in\mathcal{W}}\left|\sum_{i=1}^{n_{\mathrm{u}}}\log(1-D_w(x_{\mathrm{u},i}))-\mathbb{E}_{x\sim p}[\log(1-D_w(x))]\right|\leq 8\rho_L\mathfrak{R}_{n_{\mathrm{u}}}(D_{\mathcal{W}})+2C_L\sqrt{\frac{\log(6/\delta)}{2n_{\mathrm{u}}}}.$$

Similarly, the second and third terms on the RHS of Eq. (14) can be bounded. Since they have the bounded difference property with constants C_L/n_{a} and C_L/n_{c}, respectively (note that $|1-r(x)|\leq 1$ for any x), both of the following inequalities hold independently with probability at least $1-\delta/3$:

$$\begin{aligned}\sup_{w\in\mathcal{W}}\left|\sum_{i=1}^{n_{\mathrm{a}}}\log D_w(x_{\mathrm{a},i})-\mathbb{E}_{x\sim p_\theta}[\log D_w(x)]\right|&\leq 8\rho_L\mathfrak{R}_{n_{\mathrm{a}}}(D_{\mathcal{W}})+2C_L\sqrt{\frac{\log(6/\delta)}{2n_{\mathrm{a}}}}, \\ \sup_{w\in\mathcal{W}}\left|\sum_{i=1}^{n_{\mathrm{c}}}(1-r_i)\log D_w(x_{\mathrm{c},i})-\mathbb{E}_{x,r\sim q}[(1-r)\log D_w(x)]\right|&\leq 8\rho_L\mathfrak{R}_{n_{\mathrm{c}}}(D_{\mathcal{W}})+2C_L\sqrt{\frac{\log(6/\delta)}{2n_{\mathrm{c}}}}.\end{aligned}$$

Combining the above all, we can bound the original estimation error: the following bound holds with probability at least $1-\delta$,

$$\mathcal{V}(w^*) - \mathcal{V}(\widehat{w}) \leq 16\rho_L(\mathfrak{R}_{n_\mathrm{u}}(D_\mathcal{W}) + \alpha\mathfrak{R}_{n_\mathrm{a}}(D_\mathcal{W}) + \mathfrak{R}_{n_\mathrm{c}}(D_\mathcal{W})) + 4C_L\sqrt{\frac{\log(6/\delta)}{2}}\left(n_\mathrm{u}^{-\frac{1}{2}} + \alpha n_\mathrm{a}^{-\frac{1}{2}} + n_\mathrm{c}^{-\frac{1}{2}}\right).$$

□

C. Implementation and Experimental Details

We use the same neural net architecture and hyper-parameters for all tasks. For the architectures of all neural networks, we use two hidden layers with size 100 and Tanh as activation functions. Please refer to Table 2 for more details. Specification of each tasks is shown in Table 3, where we show the average return of the optimal and the uniformly random policies. The average return is used to normalize the performance in Sec. 5 so that 1.0 indicates the optimal policy and 0.0 the random policy.

Table 2. Hyper-parameters used for all tasks.

HYPER-PARAMETERS	VALUE
γ	0.995
τ (GENERALIZED ADVANTAGE ESTIMATION)	0.97
BATCH SIZE	$5,000$
LEARNING RATE (VALUE NETWORK)	3×10^{-4}
LEARNING RATE (DISCRIMINATOR)	1×10^{-3}
OPTIMIZER	ADAM
LOSS FUNCTION (2IWIL)	LOGISTIC LOSS

Table 3. Specification of each tasks. Optimal policy and random policy columns indicate the average return.

TASKS	$\mathcal{S}$	$\mathcal{A}$	n_u	n_c	OPTIMAL POLICY	RANDOM POLICY
HALFCHEETAH-V2	$\mathbb{R}^{17}$	$\mathbb{R}^6$	2000	500	3467.32	-288.44
WALKER-V2	$\mathbb{R}^{17}$	$\mathbb{R}^6$	1600	400	3694.13	1.91
ANT-V2	$\mathbb{R}^{111}$	$\mathbb{R}^8$	480	120	4143.10	-72.30
SWIMMER-V2	$\mathbb{R}^8$	$\mathbb{R}^2$	20	5	348.99	2.31
HOPPER-V2	$\mathbb{R}^{11}$	$\mathbb{R}^3$	16	4	3250.67	18.04

C.1. Performance comparison

The numeric performance of the proposed methods and other baselines are shown in Table 4.

Table 4. Comparison of the proposed methods with other baselines. We report the average normalized return over 5 trials. We show the best and equivalent methods based on the 5% t-test in bold.

METHODS	HALFCHEETAH-V2	ANT-V2	HOPPER-V2	SWIMMER-V2	WALKER2D-V2
OURS (2IWIL)	0.798 ± 0.019	0.687 ± 0.073	$\mathbf{0.769 \pm 0.219}$	$\mathbf{0.973 \pm 0.027}$	$\mathbf{0.675 \pm 0.098}$
OURS (IC-GAIL)	$\mathbf{0.902 \pm 0.037}$	$\mathbf{0.850 \pm 0.077}$	$\mathbf{0.974 \pm 0.039}$	$\mathbf{0.952 \pm 0.023}$	$\mathbf{0.695 \pm 0.039}$
GAIL (U+C)	0.636 ± 0.139	0.058 ± 0.200	0.561 ± 0.287	0.415 ± 0.310	0.528 ± 0.031
GAIL (REWEIGHT)	0.659 ± 0.077	0.379 ± 0.196	0.389 ± 0.183	0.510 ± 0.298	0.353 ± 0.122
GAIL (C)	0.342 ± 0.116	0.178 ± 0.132	0.314 ± 0.082	0.445 ± 0.369	0.234 ± 0.070

C.2. Non-negative risk estimator

By observing the risk estimator of Eq. (7), it is possible that the empirical estimation is negative and this may lead to overfitting (Kiryo et al., 2017). Since we know that the expected risk is nonnegative, we can borrow the idea from Kiryo et al. (2017) to mitigate this problem by simply adding the max operator to prevent the empirical risk from becoming negative by first rewriting the empirical risk as

$$\widehat{R}_{\mathrm{SC},\ell}(g) = \widehat{R}_{C}^{+}(g) + \widehat{R}_{C,U}^{-}(g), \tag{15}$$

where

$$\widehat{R}_{C}^{+}(g) = \frac{1}{n_{\mathrm{c}}} \sum_{i=1}^{n_{\mathrm{c}}} r(x_{\mathrm{c},i}) \ell(g(x_{\mathrm{c},i})),$$

and

$$\widehat{R}_{C,U}^{-}(g) = \frac{1}{n_{\mathrm{c}}} \sum_{i=1}^{n_{\mathrm{c}}} (1 - \beta - r(x_i)) \ell(-g(x_{\mathrm{c},i})) + \frac{1}{n_{\mathrm{u}}} \sum_{i=1}^{n_{\mathrm{u}}} \beta \ell(-g(x_{\mathrm{u},i})).$$

Note that $R_{C,U}^{-} \geq 0$ holds for all g. However, it is not the case for $\widehat{R}_{C,U}^{-}(g)$, which is a potential reason to overfit. Based on Eq. (15), we achieve the *non-negative risk estimator* that gives the non-negative empirical risk as follows.

$$\widehat{R}_{\mathrm{SC},\ell}(g) = \widehat{R}_{C}^{+}(g) + \max\left\{0, \widehat{R}_{C,U}^{-}(g)\right\}. \tag{16}$$

C.3. Ant-v2 Figures

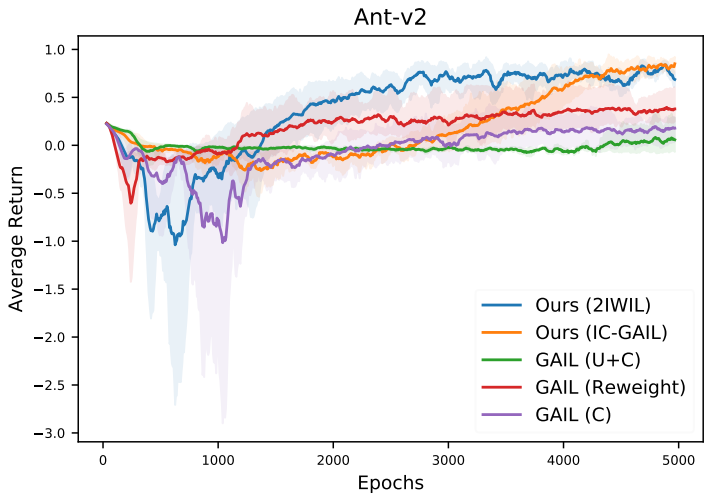

Figure 4. Learning curves of our 2IWIL and IC-GAIL versus baselines.

We empirically found that when using GAIL-based approaches in Ant-v2 environment, the performance degrades quickly in early training stages. The uncropped figures are Figs. 4, 5 and 6.

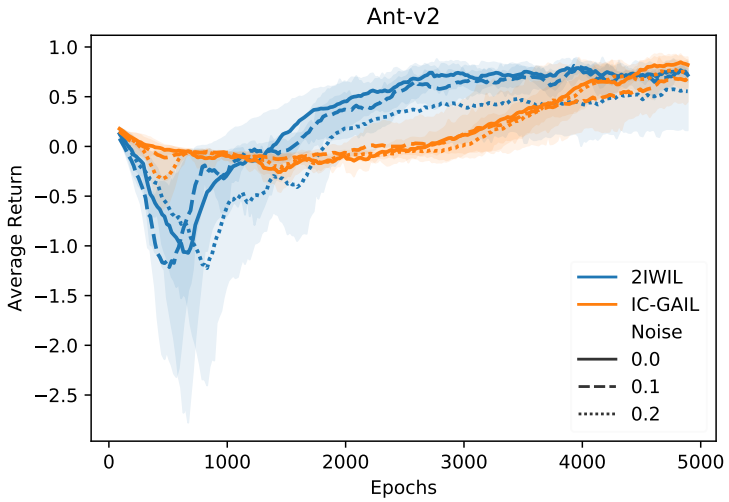

Figure 5. Learning curves of proposed methods with different standard deviations of Gaussian noise added to confidence. The numbers in the legend indicate the standard deviation of the Gaussian noise.

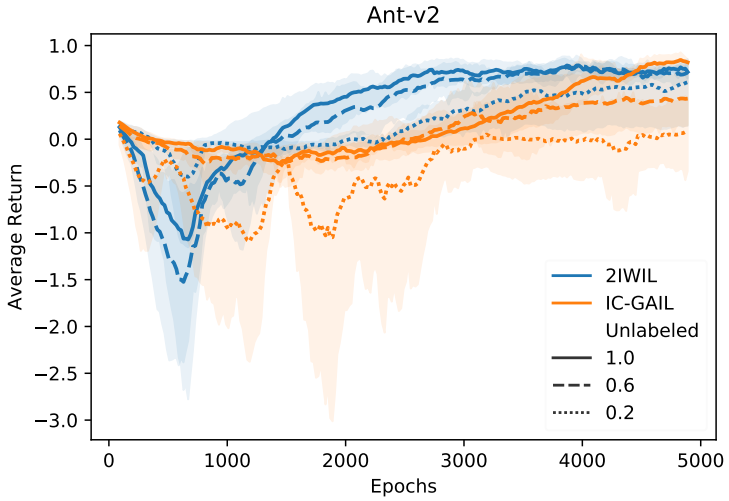

Figure 6. Learning curves of the proposed methods with different number of unlabeled data. The numbers in the legend suggest the proportion of unlabeled data used as demonstrations.