A. Gaussian Models: remarks and proofs

In this section we consider the case where the transition and reward models are Gaussian distributions.

A task $M_j = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}_j, \mathcal{R}_j, \gamma \rangle$ is jointly Gaussian when it has:

- *Gaussian Reward*: for any $(s,a) \in \mathcal{S} \times \mathcal{A}$

$$\mathcal{R}_j(\cdot|s,a) = \mathcal{N}(\mu_r^{(j)}(s,a), \sigma_j^2(s,a)). \tag{7}$$

- *Gaussian Transitions*: for any $(s,a) \in \mathcal{S} \times \mathcal{A}$

$$\mathcal{P}_j(\cdot|s,a) = \mathcal{N}(\boldsymbol{\mu}_p^{(j)}(s,a), \boldsymbol{\Sigma}_j(s,a)) \tag{8}$$

.

In this setting, by using a Gaussian process to estimate the transition and reward models, we can compute the expected importance weights, where the expectation is taken under the distribution induced by the GPs' predictions, in closed form. Note that, since reward and transition weights are independent, we can independently consider Assumptions 7 and 8 for the computation of w_r and w_p, respectively.

Theorem 4 (Reward Weights in Gaussian Models)**.** *Assume each task to have Gaussian reward distribution $\mathcal{R}_j(\cdot|s,a) = \mathcal{N}\left(\mu_r^{(j)}(s,a), \sigma_j^2(s,a)\right)$ with unknown mean. Given the available samples in $\widetilde{\mathcal{D}}$, we build an estimate of the reward distribution such that, for any MDP M_j, $\overline{r}^{(j)}(s,a) \sim \mathcal{N}(\mu_{GP_j}(s,a), \sigma^2_{GP_j}(s,a))$. Then, given a sample $\langle s,a,r \rangle$ from the j-th MDP, its importance weight $w = \frac{\mathcal{N}\left(r|\overline{r}^{(0)}(s,a), \sigma_0^2(s,a)\right)}{\mathcal{N}\left(r|\overline{r}^{(j)}(s,a), \sigma_j^2(s,a)\right)} \sim \mathcal{G}$, where $\mathcal{G}$ is the distribution induced by the GPs' predictions. Let $C = \frac{\sigma_j^2(s,a)}{\sigma_j^2(s,a) - \sigma^2_{GP_j}(s,a)}$ and suppose $\sigma^2_{GP_j}(s,a) < \sigma_j^2(s,a)$, then*

$$\mathbb{E}_{\mathcal{G}}[w] = C\, \frac{\mathcal{N}\left(r\middle|\mu_{GP_0}(s,a), \sigma_0^2(s,a) + \sigma^2_{GP_0}(s,a)\right)}{\mathcal{N}\left(r\middle|\mu_{GP_j}(s,a), \sigma_j^2(s,a) - \sigma^2_{GP_j}(s,a)\right)}. \tag{6}$$

Proof. In order to simplify our notation, we consider the dependence on (s,a,r) implicit. We have:

$$\begin{aligned}
\mathbb{E}[w] &= \int\int w\, \mathcal{N}\left(\overline{r}^{(0)}\middle|\mu_{GP_0}, \sigma^2_{GP_0}\right) \mathcal{N}\left(\overline{r}^{(j)}\middle|\mu_{GP_j}, \sigma^2_{GP_j}\right) \mathrm{d}\overline{r}^{(0)} \mathrm{d}\overline{r}^{(j)} \\
&= \int\int w \frac{\exp\left(-\frac{(\overline{r}^{(0)} - \mu_{GP_0})^2}{2\sigma^2_{GP_0}}\right)}{\sqrt{2\pi\sigma^2_{GP_0}}} \frac{\exp\left(-\frac{(\overline{r}^{(j)} - \mu_{GP_j})^2}{2\sigma^2_{GP_j}}\right)}{\sqrt{2\pi\sigma^2_{GP_j}}} \mathrm{d}\overline{r}^{(0)} \mathrm{d}\overline{r}^{(j)} \\
&= \int \frac{\exp\left(-\frac{(r - \overline{r}^{(0)})^2}{2\sigma_0^2}\right)}{\sqrt{2\pi\sigma_0^2}} \frac{\exp\left(-\frac{(\overline{r}^{(0)} - \mu_{GP_0})^2}{2\sigma^2_{GP_0}}\right)}{\sqrt{2\pi\sigma^2_{GP_0}}} \mathrm{d}\overline{r}^{(0)} \int \frac{\sqrt{2\pi\sigma_j^2}}{\exp\left(-\frac{(r - \overline{r}^{(j)})^2}{2\sigma_j^2}\right)} \frac{\exp\left(-\frac{(\overline{r}^{(j)} - \mu_{GP_j})^2}{2\sigma^2_{GP_j}}\right)}{\sqrt{2\pi\sigma^2_{GP_j}}} \mathrm{d}\overline{r}^{(j)}.
\end{aligned}$$

The first integral is over the product of two Gaussian densities, which is known to be (Bromiley, 2003):

$$\mathcal{N}\left(\overline{r}^{(0)}\middle|\bar{\mu}_0, \bar{\sigma}_0^2\right) \mathcal{N}\left(r\middle|\mu_{GP_0}, \sigma_0^2 + \sigma^2_{GP_0}\right), \tag{9}$$

where the values of the mean $\bar{\mu}_0$ and variance $\bar{\sigma}_0^2$ of the first density are not important to complete this proof since such density integrates out. By adopting the same procedure as the one described in (Bromiley, 2003), we can write the ratio of the two Gaussian densities in the second integral as:

$$\frac{\sigma_j^2}{\sigma_j^2 - \sigma^2_{GP_j}} \frac{\mathcal{N}\left(\overline{r}^{(j)}\middle|\bar{\mu}_j, \bar{\sigma}_j^2\right)}{\mathcal{N}\left(r\middle|\mu_{GP_j}, \sigma_j^2 - \sigma^2_{GP_j}\right)}, \tag{10}$$

where, again, the values of $\bar{\mu}_j$ and $\bar{\sigma}_j^2$ are not relevant to our proof. Finally, we can write:

$$\begin{aligned}\mathbb{E}\left[w(r)\right] =& \int \mathcal{N}\left(\overline{r}^{(0)}\middle|\bar{\mu}_0, \bar{\sigma}_0^2\right) \mathcal{N}\left(r\middle|\mu_{GP_0}, \sigma_0^2 + \sigma_{GP_0}^2\right) \mathrm{d}\overline{r}^{(0)} \int \frac{\sigma_j^2}{\sigma_j^2 - \sigma_{GP_j}^2} \frac{\mathcal{N}\left(\overline{r}^{(j)}\middle|\bar{\mu}_j, \bar{\sigma}_j^2\right)}{\mathcal{N}\left(r\middle|\mu_{GP_j}, \sigma_j^2 - \sigma_{GP_j}^2\right)} \mathrm{d}\overline{r}^{(j)} \\ =& \frac{\sigma_j^2}{\sigma_j^2 - \sigma_{GP_j}^2} \frac{\mathcal{N}\left(r\middle|\mu_{GP_0}, \sigma_0^2 + \sigma_{GP_0}^2\right)}{\mathcal{N}\left(r\middle|\mu_{GP_j}, \sigma_j^2 - \sigma_{GP_j}^2\right)} \int \mathcal{N}\left(\overline{r}^{(0)}\middle|\bar{\mu}_0, \bar{\sigma}_0^2\right) \mathrm{d}\overline{r}^{(0)} \int \mathcal{N}\left(\overline{r}^{(j)}\middle|\bar{\mu}_j, \bar{\sigma}_j^2\right) \mathrm{d}\overline{r}^{(j)} \\ =& \frac{\sigma_j^2}{\sigma_j^2 - \sigma_{GP_j}^2} \frac{\mathcal{N}(r\middle|\mu_{GP_0}, \sigma_0^2 + \sigma_{GP_0}^2)}{\mathcal{N}(r\middle|\mu_{GP_j}, \sigma_j^2 - \sigma_{GP_j}^2)}.\end{aligned}$$

□

We can derive a similar result for the transition model by considering Assumption 8.

Theorem 5 (Transition Weights in Gaussian Models)**.** *Assume each task to have Gaussian transition distribution $\mathcal{P}_j(\cdot|s,a) = \mathcal{N}(\boldsymbol{\mu}_p^{(j)}(s,a), \boldsymbol{\Sigma}_j(s,a))$ with unknown mean. Furthermore, suppose that $\boldsymbol{\Sigma}_j(s,a) = diag\left(\delta_{j,1}^2(s,a), \delta_{j,2}^2(s,a), \ldots, \delta_{j,D}^2(s,a)\right)$. Given the available samples in $\widetilde{\mathcal{D}}$, we build an estimate of the transition distribution such that, for any MDP M_j and state component d, $\overline{p}_d^{(j)}(s,a) \sim \mathcal{N}(\mu_{GP_{j,d}}(s,a), \sigma_{GP_{j,d}}^2(s,a))$. Then, given a sample $\langle s, a, s'\rangle$ from the j-th MDP, the importance weight for the transition model given by $w = \prod_{d=1}^{D} \frac{\mathcal{N}\left(s'_d\middle|\overline{p}_d^{(0)}(s,a), \delta_{0,d}^2(s,a)\right)}{\mathcal{N}\left(s'_d\middle|\overline{p}_d^{(j)}(s,a), \delta_{j,d}^2(s,a)\right)}$ is distributed according to some distribution $\mathcal{G}$ induced by the GPs' predictions. Let $C_d = \frac{\delta_{j,d}^2(s,a)}{\delta_{j,d}^2(s,a) - \sigma_{GP_{j,d}}^2(s,a)}$ and suppose $\sigma_{GP_{j,d}}^2(s,a) < \delta_{j,d}^2(s,a)$ for all d, then*

$$\mathbb{E}_{\mathcal{G}}\left[w\right] = \prod_{d=1}^{D} C_d \frac{\mathcal{N}\left(s'_d\middle|\mu_{GP_{0,d}}(s,a), \delta_{0,d}^2(s,a) + \sigma_{GP_{0,d}}^2(s,a)\right)}{\mathcal{N}\left(s'_d\middle|\mu_{GP_{j,d}}(s,a), \delta_{j,d}^2(s,a) - \sigma_{GP_{j,d}}^2(s,a)\right)}. \tag{11}$$

Proof. Notice that the density $\mathcal{P}_j(\cdot|s,a)$ decomposes into:

$$\mathcal{P}_j(\cdot|s,a) = \prod_{d=1}^{D} \mathcal{N}\left(\cdot\middle|\mu_{p,d}^{(j)}(s,a), \delta_{j,d}^2(s,a)\right). \tag{12}$$

By noticing that the d-th GP for task j provides an independent estimate of the transition mean's d-th component, $\overline{p}_d^{(j)}$, the proof of Theorem 4 can be straightforwardly applied to each component of s', thus proving the theorem. □

B. Proofs of Theorems

Theorem 2. *Let $\mathcal{H} \subset \mathcal{B}(X, F_{\max})$ be a functional space. Suppose we have a dataset of N i.i.d. samples $\mathcal{D} = \{(x_i, y_i)\}$ distributed according to $Q(X,Y) = q(Y|X)\mu(X)$, while $P(X,Y) = p(Y|X)\mu(X)$ is the target distribution. Assume $|Y| \leq F_{\max}$ almost surely. Let $w(x,y) = \frac{p(y|x)}{q(y|x)}$, $\widetilde{w}(x,y)$ be any positive function, $\widehat{h}(x) = \arg\min_{f\in\mathcal{H}} \widehat{\mathbb{E}}_{\mathcal{D}}\left[\widetilde{w}(X,Y)|f(X) - Y|^2\right]$, $h^*(x) = \mathbb{E}_p[Y|x]$, $g(x) = \mathbb{E}_q[\widetilde{w}(x,Y)|x] - 1$, and $M(\widetilde{w}) = \sqrt{\mathbb{E}_Q[\widetilde{w}(X,Y)^2]} + \sqrt{\widehat{\mathbb{E}}_{\mathcal{D}}[\widetilde{w}(X,Y)^2]}$, where $\widehat{\mathbb{E}}_{\mathcal{D}}$ denotes the empirical expectation on $\mathcal{D}$. Furthermore, assume $d = Pdim(\{|f(x) - y|^2 : f \in \mathcal{H}\}) < \infty$ and $\mathbb{E}_Q[\widetilde{w}(X,Y)^2] < \infty$. Then, for any $\delta > 0$, the following holds with probability at least $1 - 2\delta$:*

$$\begin{aligned}\|\widehat{h} - h^*\|_\mu \leq& \inf_{f\in\mathcal{H}} \|f - h^*\|_\mu + F_{\max}\sqrt{\|g\|_{1,\mu}} \\ &+ 2^{13/8} F_{\max} \sqrt{M(\widetilde{w})} \left(\frac{d \log \frac{2Ne}{d} + \log \frac{4}{\delta}}{N}\right)^{\frac{3}{16}} \\ &+ 2F_{\max}\|\widetilde{w} - w\|_Q\end{aligned}$$

Proof. Applying Hölder's inequality, for all $f \in \mathcal{H}$:

$$\mathbb{E}_Q\left[\left(\widetilde{w}(X,Y)|f(X)-Y|^2\right)^2\right] \leq 16F_{\max}^4\mathbb{E}_Q[\widetilde{w}(X,Y)^2] < \infty. \tag{13}$$

Thus, by applying Corollary 1 of (Cortes et al., 2010) to $L_h(x,y) = \widetilde{w}(x,y)|h(x)-y|^2$, we can write:

$$\mathbb{E}_Q[\widetilde{w}(X,Y)|\widehat{h}(X)-Y|^2] \leq \frac{1}{N}\sum_{i=1}^N \widetilde{w}(x_i,y_i)|\widehat{h}(x_i)-y_i|^2 + 2^{13/4}F_{\max}^2\sqrt{\mathbb{E}_Q\left[\widetilde{w}(X,Y)^2\right]}\sqrt[8]{\frac{d\log\frac{2Ne}{d}+\log\frac{4}{\delta}}{N}}. \tag{14}$$

Let us now expand the left-hand side of (14):

$$\begin{aligned}
\mathbb{E}_Q\left[\widetilde{w}(X,Y)|\widehat{h}(X)-Y|^2\right] &= \mathbb{E}_\mu\left[\mathbb{E}_q\left[\widetilde{w}(X,Y)\left(\widehat{h}^2(X)+Y^2-2\widehat{h}(X)Y\right) \mid X\right]\right] \\
&= \mathbb{E}_\mu\Big[\widehat{h}^2(X)\mathbb{E}_q[\widetilde{w}(X,Y)\mid X] + \mathbb{E}_q[\widetilde{w}(X,Y)Y^2\mid X] - 2\widehat{h}(X)\mathbb{E}_q[\widetilde{w}(X,Y)Y\mid X] \\
&\quad \pm\mathbb{E}_q^2[\widetilde{w}(X,Y)Y\mid X] \pm \widehat{h}^2(X)\Big] \\
&= \mathbb{E}_\mu\Big[\left(\widehat{h}(X)-\mathbb{E}_q[\widetilde{w}(X,Y)Y\mid X]\right)^2 + \widehat{h}^2(X)\mathbb{E}_q[\widetilde{w}(X,Y)\mid X] + \mathbb{E}_q[\widetilde{w}(X,Y)Y^2\mid X] \\
&\quad -\mathbb{E}_q^2[\widetilde{w}(X,Y)Y\mid X] - \widehat{h}^2(X)\Big] \\
&= \|\widehat{h}-\widetilde{h}\|_\mu^2 + \mathbb{E}_\mu[\widehat{h}^2(X)\left(\mathbb{E}_q[\widetilde{w}(X,Y)-1\mid X]\right)] + K,
\end{aligned} \tag{15}$$

where $K = \mathbb{E}_\mu[\mathbb{E}_q[\widetilde{w}(X,Y)Y^2 \mid X] - \mathbb{E}_q^2[\widetilde{w}(X,Y)Y \mid X]]$ is a constant term (independent of $\widehat{h}$) and $\widetilde{h}(x) = \mathbb{E}_q[\widetilde{w}(x,Y)Y \mid x]$ is the regression function weighted by $\widetilde{w}$. Plugging this into (14) we get:

$$\begin{aligned}
&\|\widehat{h}-\widetilde{h}\|_\mu^2 + \mathbb{E}_\mu[\widehat{h}^2(X)\left(\mathbb{E}_q[\widetilde{w}(X,Y)-1\mid X]\right)] + K \leq \\
&\frac{1}{N}\sum_{i=1}^N \widetilde{w}(x_i,y_i)|\widehat{h}(x_i)-y_i|^2 + 2^{13/4}F_{max}^2\sqrt{\mathbb{E}_Q[\widetilde{w}(X,Y)]}\sqrt[8]{\frac{d\log\frac{2Ne}{d}+\log\frac{4}{\delta}}{N}}.
\end{aligned} \tag{16}$$

Consider now the hypothesis $h_0 \in \mathcal{H}$ such that $h_0 = \underset{f\in\mathcal{H}}{argmin}\ \|f-\widetilde{h}\|_\mu^2$. Since h_0 is in $\mathcal{H}$ and $\widehat{h}$ was defined as the hypothesis minimizing the empirical weighted loss, we have:

$$\frac{1}{N}\sum_{i=1}^N \widetilde{w}(x_i,y_i)|\widehat{h}(x_i)-y_i|^2 \leq \frac{1}{N}\sum_{i=1}^N \widetilde{w}(x_i,y_i)|h_0(x_i)-y_i|^2. \tag{17}$$

Similarly to what we did for $\widehat{h}$, we can bound the empirical error of h_0. According to Corollary 1 of (Cortes et al., 2010), we have that for any $\delta > 0$, with probability at least $1-\delta$:

$$\frac{1}{N}\sum_{i=1}^N \widetilde{w}(x_i,y_i)|h_0(x_i)-y_i|^2 \leq \mathbb{E}_Q[\widetilde{w}(X,Y)|h_0(X)-Y|^2] + 2^{13/4}F_{max}^2\sqrt{\widehat{\mathbb{E}}_{\mathcal{D}}[\widetilde{w}(X,Y)]}\sqrt[8]{\frac{d\log\frac{2Ne}{d}+\log\frac{4}{\delta}}{N}}. \tag{18}$$

By adopting (15) to expand the expected error of h_0, we obtain:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N \widetilde{w}(x_i,y_i)|h_0(x_i)-y_i|^2 \leq \\
&\inf_{f\in\mathcal{H}}\|f-\widetilde{h}\|_\mu^2 + \mathbb{E}_\mu[h_0^2(X)\left(\mathbb{E}_q[\widetilde{w}(X,Y)-1\mid X]\right)] + K + 2^{13/4}F_{max}^2\sqrt{\widehat{\mathbb{E}}_{\mathcal{D}}[\widetilde{w}(X,Y)]}\sqrt[8]{\frac{d\log\frac{2Ne}{d}+\log\frac{4}{\delta}}{N}}.
\end{aligned} \tag{19}$$

If we now put (16) and (19) together by means of (17), we get that, with probability at least $1-2\delta$:

$$\begin{aligned}
\|\widehat{h}-\widetilde{h}\|_\mu^2 &\leq \inf_{f\in\mathcal{H}}\|f-\widetilde{h}\|_\mu^2 + \mathbb{E}_\mu[\left(h_0^2(X)-\widehat{h}^2(X)\right)(\mathbb{E}_q[\widetilde{w}(X,Y)-1\mid X])] \\
&\quad + 2^{13/4}F_{max}^2\left(\sqrt{\mathbb{E}_Q[\widetilde{w}(X,Y)^2]}+\sqrt{\widehat{\mathbb{E}}_{\mathcal{D}}[\widetilde{w}(X,Y)^2]}\right)\sqrt[\frac{3}{8}]{\frac{d\log\frac{2Ne}{d}+\log\frac{4}{\delta}}{N}} \\
&\leq \inf_{f\in\mathcal{H}}\|f-\widetilde{h}\|_\mu^2 + F_{max}^2\mathbb{E}_\mu\left[|\mathbb{E}_q[\widetilde{w}(X,Y)\mid X]-1|\right] \\
&\quad + 2^{13/4}F_{max}^2\left(\sqrt{\mathbb{E}_Q[\widetilde{w}(X,Y)^2]}+\sqrt{\widehat{\mathbb{E}}_{\mathcal{D}}[\widetilde{w}(X,Y)^2]}\right)\sqrt[\frac{3}{8}]{\frac{d\log\frac{2Ne}{d}+\log\frac{4}{\delta}}{N}} \\
&= \inf_{f\in\mathcal{H}}\|f-\widetilde{h}\|_\mu^2 + F_{max}^2\|g\|_{1,\mu} + 2^{13/4}F_{max}^2 M(\widetilde{w})\left(\frac{d\log\frac{2Ne}{d}+\log\frac{4}{\delta}}{N}\right)^{\frac{3}{8}}.
\end{aligned} \tag{20}$$

By taking the square root of both sides of (20) and using $\sqrt{\sum_i a_i}\leq\sum_i\sqrt{a_i}$ for $a_i\geq 0$, we obtain:

$$\|\widehat{h}-\widetilde{h}\|_\mu \leq \inf_{f\in\mathcal{H}}\|f-\widetilde{h}\|_\mu + F_{max}\sqrt{\|g\|_{1,\mu}} + 2^{13/8}F_{max}\sqrt{M(\widetilde{w})}\left(\frac{d\log\frac{2Ne}{d}+\log\frac{4}{\delta}}{N}\right)^{\frac{3}{16}}. \tag{21}$$

Furthermore:

$$\inf_{f\in\mathcal{H}}\|f-\widetilde{h}\|_\mu \leq \inf_{f\in\mathcal{H}}\|f-h^\star\|_\mu + \|h^\star-\widetilde{h}\|_\mu. \tag{22}$$

We can now bound the expected error of $\widehat{h}$ with respect to $h^\star$ by:

$$\|\widehat{h}-h^\star\|_\mu \leq \|\widehat{h}-\widetilde{h}\|_\mu + \|\widetilde{h}-h^\star\|_\mu. \tag{23}$$

We already provided a bound on the first term, so let us analyze the second one. We have:

$$\begin{aligned}
\|\widetilde{h}-h^\star\|_\mu^2 &= \mathbb{E}_\mu[|\widetilde{h}(X)-h^\star(X)|^2] \\
&= \mathbb{E}_\mu[|\mathbb{E}_q[\widetilde{w}(X,Y)Y\mid X]-\mathbb{E}_p[Y\mid X]|^2] \\
&= \mathbb{E}_\mu[|\mathbb{E}_q[\widetilde{w}(X,Y)Y\mid X]-\mathbb{E}_q[w(X,Y)Y\mid X]|^2] \\
&= \mathbb{E}_\mu[|\mathbb{E}_q[Y\left(\widetilde{w}(X,Y)-w(X,Y)\right)\mid X]|^2] \\
&\leq \mathbb{E}_\mu[\mathbb{E}_q[|Y|^2|\widetilde{w}(X,Y)-w(X,Y)|^2\mid X]] \\
&\leq F_{max}^2\mathbb{E}_Q\left[|\widetilde{w}(X,Y)-w(X,Y)|^2\right] \\
&= F_{max}^2\|\widetilde{w}-w\|_Q^2.
\end{aligned} \tag{24}$$

where the first inequality is due to Jensen's inequality. Thus, $\|\widetilde{h}-h^\star\|_\mu \leq F_{max}\|\widetilde{w}-w\|_Q$. By combining (21), (22), (23), and (24), we have:

$$\begin{aligned}
\|\widehat{h}-h^\star\|_\mu &\leq \|\widehat{h}-\widetilde{h}\|_\mu + \|\widetilde{h}-h^\star\|_\mu \\
&\leq \inf_{f\in\mathcal{H}}\|f-\widetilde{h}\|_\mu + F_{max}\sqrt{\|g\|_{1,\mu}} + 2^{13/8}F_{max}\sqrt{M(\widetilde{w})}\left(\frac{d\log\frac{2Ne}{d}+\log\frac{4}{\delta}}{N}\right)^{\frac{3}{16}} + \|\widetilde{h}-h^\star\|_\mu \\
&\leq \inf_{f\in\mathcal{H}}\|f-h^\star\|_\mu + F_{max}\sqrt{\|g\|_{1,\mu}} + 2^{13/8}F_{max}\sqrt{M(\widetilde{w})}\left(\frac{d\log\frac{2Ne}{d}+\log\frac{4}{\delta}}{N}\right)^{\frac{3}{16}} + 2\|\widetilde{h}-h^\star\|_\mu \\
&\leq \inf_{f\in\mathcal{H}}\|f-h^\star\|_\mu + F_{max}\sqrt{\|g\|_{1,\mu}} + 2^{13/8}F_{max}\sqrt{M(\widetilde{w})}\left(\frac{d\log\frac{2Ne}{d}+\log\frac{4}{\delta}}{N}\right)^{\frac{3}{16}} + 2F_{max}\|\widetilde{w}-w\|_Q
\end{aligned}$$

which concludes the proof. □

Lemma 1. *Let $\mathcal{H} \subset B(\mathcal{S} \times \mathcal{A}, Q_{max})$ be a functional space. Call $g_r(s,a) = \mathbb{E}_{\mathcal{R}_S}[\widetilde{w}_r(r|s,a)] - 1$ and $M(\widetilde{w}_r) = \sqrt{\mathbb{E}_{\phi_S^R}[\widetilde{w}_r(r|s,a)^2]} + \sqrt{\widehat{\mathbb{E}}_{\mathcal{D}}[\widetilde{w}_r(r|s,a)^2]}$, where $\phi_S^R(r|s,a) = \mu(s,a)\mathcal{R}_s(r|s,a)$ and $\mathcal{D}$ is a dataset of N i.i.d. samples. Assume $d = Pdim(\{|f(s,a) - r|^2 \ : f \in \mathcal{H}\}) < \infty$ and $\mathbb{E}_{\phi_S^R}[\widetilde{w}_r(r|s,a)^2] < \infty$. Let $\widehat{R}$ be as defined in (2). Then, for any $\delta > 0$, with probability at least $1 - 2\delta$:*

$$
\begin{aligned}
||R - \widehat{R}||_\mu \leq \inf_{f \in \mathcal{H}} ||f - R||_\mu + r_{max}\sqrt{||g_r||_{1,\mu}} \\
+ 2^{13/8} Q_{max} \sqrt{M(\widetilde{w}_r)} \left(\frac{d \log \frac{2Ne}{d} + \log \frac{4}{\delta}}{N} \right)^{\frac{3}{16}} \\
+ 2 r_{max} ||\widetilde{w}_r - w_r||_{\phi_S^R}.
\end{aligned}
\tag{25}
$$

Proof. The result follows straightforwardly by applying Theorem 2. □

Lemma 2. *Let $\mathcal{H} \subset B(\mathcal{S} \times \mathcal{A}, Q_{max})$ be a functional space. Call $g_p(s,a) = \mathbb{E}_{\mathcal{P}_S}[\widetilde{w}_p(s'|s,a)] - 1$ and $M(\widetilde{w}_p) = \sqrt{\mathbb{E}_{\phi_S^P}[\widetilde{w}_p(s'|s,a)^2]} + \sqrt{\widehat{\mathbb{E}}_{\mathcal{D}}[\widetilde{w}_p(s'|s,a)^2]}$, where $\phi_S^P(r|s,a) = \mu(s,a)\mathcal{P}_s(r|s,a)$ and $\mathcal{D}$ is a dataset of N i.i.d. samples. Assume $d = Pdim(\{|f(s,a) - r|^2 \ : f \in \mathcal{H}\}) < \infty$ and $\mathbb{E}_{\phi_S^P}[\widetilde{w}_p(s'|s,a)^2] < \infty$. Let Q_{k+1} be as defined in (4) and denote $\widetilde{L}^* Q(s,a) \coloneqq \widehat{R}(s,a) + \int_{\mathcal{S}} \mathcal{P}_T(\mathrm{d}s'|s,a) \max_{a'} Q(s',a)$. Then, for any $\delta > 0$, with probability at least $1 - 2\delta$:*

$$
\begin{aligned}
||\widetilde{L}^* Q_k - Q_{k+1}||_\mu \leq \inf_{f \in \mathcal{H}} ||f - \widetilde{L}^* Q_k||_\mu + Q_{max}\sqrt{||g_p||_{1,\mu}} \\
+ 2^{13/8} Q_{max} \sqrt{M(\widetilde{w}_p)} \left(\frac{d \log \frac{2Ne}{d} + \log \frac{4}{\delta}}{N} \right)^{\frac{3}{16}} \\
+ 2 Q_{max} ||\widetilde{w}_p - w_p||_{\phi_S^P}.
\end{aligned}
\tag{26}
$$

Proof. The result follows straightforwardly by applying Theorem 2. □

Theorem 3. *Let $\mathcal{H} \subset B(\mathcal{S} \times \mathcal{A}, Q_{\max})$ be a hypothesis space, μ a distribution over $\mathcal{S} \times \mathcal{A}$, $(Q_i)_{i=0}^{k+1}$ a sequence of Q-functions as defined in Equation (4), and L^* the optimal Bellman operator of the target task. Suppose to have a dataset of N i.i.d. samples $\mathcal{D}$ drawn from the source task M_S according to a joint distribution ϕ_S. Let w_p, w_r denote the ideal importance weights defined in (5) and (3), and $\widetilde{w}_r(r|s,a)$, $\widetilde{w}_p(s'|s,a)$ denote arbitrary positive functions with bounded second moments. Define $g_r(s,a) = \mathbb{E}_{\mathcal{R}_S}[\widetilde{w}_r(r|s,a)|s,a] - 1$, $M(\widetilde{w}_r) = \sqrt{\mathbb{E}_{\phi_S^R}[\widetilde{w}_r(r|s,a)^2]} + \sqrt{\widehat{\mathbb{E}}_{\mathcal{D}}[\widetilde{w}_r(r|s,a)^2]}$, where $\phi_S^R(r|s,a) = \mu(s,a)\mathcal{R}_S(r|s,a)$. Similarly, define g_p, $M(\widetilde{w}_p)$, and $\phi_S^P(s'|s,a)$ for the transition model. Then, for any $\delta > 0$, with probability at least $1 - 4\delta$:*

$$
\begin{aligned}
&\|L^* Q_k - Q_{k+1}\|_\mu \leq Q_{\max}\sqrt{\|g_p\|_{1,\mu}} + 2 r_{\max}\sqrt{\|g_r\|_{1,\mu}} \\
&+ 2 Q_{\max} \|\widetilde{w}_p - w_p\|_{\phi_S^P} + 4 r_{\max} \|\widetilde{w}_r - w_r\|_{\phi_S^R} \\
&+ \inf_{f \in \mathcal{H}} \|f - (L^*)^{k+1} Q_0\|_\mu + 2 \inf_{f \in \mathcal{H}} \|f - R\|_\mu \\
&+ \frac{Q_{\max}}{2^{-\frac{13}{8}}} \left(\sqrt{M(\widetilde{w}_p)} + 2\sqrt{M(\widetilde{w}_r)} \right) \left(\frac{d \log \frac{2Ne}{d} + \log \frac{4}{\delta}}{N} \right)^{\frac{3}{16}} \\
&+ \sum_{i=0}^{k-1} (\gamma C_{\mathrm{AE}}(\mu))^{i+1} \|\epsilon_{k-i-1}\|_\mu,
\end{aligned}
$$

where C_{AE} is the concentrability coefficient of one-step transitions as defined in (Farahmand, 2011, Definition 5.2).

Proof. We can decompose the error at iteration k into:

$$
\begin{aligned}
\|\epsilon_k\|_\mu &= \|L^* Q_k - Q_{k+1}\|_\mu \\
&\leq \|L^* Q_k - \widetilde{L}^* Q_k\|_\mu + \|\widetilde{L}^* Q_k - Q_{k+1}\|_\mu \\
&= \|R - \widehat{R}\|_\mu + \|\widetilde{L}^* Q_k - Q_{k+1}\|_\mu,
\end{aligned}
\tag{27}
$$

where, for any pair (s,a), $\widetilde{L}^*Q(s,a) \coloneqq \widehat{R}(s,a) + \int_S \mathcal{P}_T(\mathrm{d}s'|s,a)\max_{a'} Q(s',a)$ is the optimal Bellman operator of the target task using the approximated reward function defined in (2). The two terms in (27) can be bounded straightforwardly by applying Lemma 1 and Lemma 2, respectively. The application of Lemma 2 to the second term gives rise to $\inf_{f\in\mathcal{H}} \|f - \widetilde{L}^*Q_k\|_\mu$, which can be further bounded by noticing that:

$$\inf_{f\in\mathcal{H}} \|f - \widetilde{L}^*Q_k\|_\mu \leq \inf_{f\in\mathcal{H}} \|f - L^*Q_k\|_\mu + \|L^*Q_k - \widetilde{L}^*Q_k\|_\mu. \tag{28}$$

The second term in (28) is again $\|R - \widehat{R}\|_\mu$, while the first term has already been bounded in Theorem 5.3 of (Farahmand, 2011):

$$\inf_{f\in\mathcal{H}} \|f - L^*Q_k\|_\mu \leq \inf_{f\in\mathcal{H}} \|f - (L^*)^{k+1}Q_0\|_\mu + \sum_{i=0}^{k-1} (\gamma C_{AE}(\mu))^{i+1} \|\epsilon_{k-i-1}\|_\mu. \tag{29}$$

Then, by combining the bounds from Lemma 1 and Lemma 2 with (28) and (29), we can write:

$$\begin{aligned}
\|\epsilon_k\|_\mu &\leq \|R - \widehat{R}\|_\mu + \|\widetilde{L}^*Q_k - Q_{k+1}\|_\mu \\
&\leq \|R - \widehat{R}\|_\mu + \inf_{f\in\mathcal{H}} ||f - \widetilde{L}^*Q_k||_\mu + Q_{max}\sqrt{||g_p||_{1,\mu}} \\
&\quad + 2^{13/8} Q_{max} \sqrt{M(\widetilde{w}_p)} \left(\frac{d \log \frac{2Ne}{d} + \log \frac{4}{\delta}}{N} \right)^{\frac{3}{16}} + 2Q_{max} ||\widetilde{w}_p - w_p||_{\phi_S^P} \\
&\leq 2\|R - \widehat{R}\|_\mu + \inf_{f\in\mathcal{H}} \|f - (L^*)^{k+1}Q_0\|_\mu + \sum_{i=0}^{k-1} (\gamma C_{AE}(\mu))^{i+1} \|\epsilon_{k-i-1}\|_\mu + Q_{max}\sqrt{||g_p||_{1,\mu}} \\
&\quad + 2^{13/8} Q_{max} \sqrt{M(\widetilde{w}_p)} \left(\frac{d \log \frac{2Ne}{d} + \log \frac{4}{\delta}}{N} \right)^{\frac{3}{16}} + 2Q_{max} ||\widetilde{w}_p - w_p||_{\phi_S^P} \\
&\leq Q_{\max}\sqrt{\|g_p\|_{1,\mu}} + 2r_{\max}\sqrt{\|g_r\|_{1,\mu}} + 2Q_{\max}\|\widetilde{w}_p - w_p\|_{\phi_S^P} + 4r_{\max}\|\widetilde{w}_r - w_r\|_{\phi_S^R} \\
&\quad + \inf_{f\in\mathcal{H}} \|f - (L^*)^{k+1}Q_0\|_\mu + 2 \inf_{f\in\mathcal{H}} \|f - R\|_\mu \\
&\quad + 2^{\frac{13}{8}} Q_{\max} \left(\sqrt{M(\widetilde{w}_p)} + 2\sqrt{M(\widetilde{w}_r)} \right) \left(\frac{d \log \frac{2Ne}{d} + \log \frac{4}{\delta}}{N} \right)^{\frac{3}{16}} + \sum_{i=0}^{k-1} (\gamma C_{\mathrm{AE}}(\mu))^{i+1} \|\epsilon_{k-i-1}\|_\mu,
\end{aligned}$$

where we recall that $g_p(s,a) = \mathbb{E}_{\mathcal{P}_S}[\widetilde{w}_p(s'|s,a)|s,a] - 1$, $M(\widetilde{w}_p) = \sqrt{\mathbb{E}_{\phi_S^P}[\widetilde{w}_p(s'|s,a)^2]} + \sqrt{\widehat{\mathbb{E}}_{\mathcal{D}}[\widetilde{w}_p(s'|s,a)^2]}$, and $\phi_S^P(s'|s,a) = \mu(s,a)\mathcal{P}_S(s'|s,a)$. This concludes the proof. □

C. Additional Details on the Experiments

C.1. Puddle World

Our first experimental domain is a modified version of the puddle world environment presented in (Sutton, 1996). Puddle world is a two-dimensional continuous grid with a goal area and some elliptical "puddles". The goal is to drive the agent from a starting position to the goal area while avoiding the puddles. The state-space is $[0, 10]^2$, while the action-space is discrete and allows the agent to move in the four cardinal directions. At each time-step, the agent receives a reward of -1 plus a penalization proportional to the distance from all puddles: $R(s,a) = -1 - 100\sum_{u\in\mathcal{U}} W_u(s)$, where $\mathcal{U}$ is the set of puddles and $W_u(s)$ is the weight of puddle u for state s. In the goal the reward is zero. In our experiments, we modeled $W_u(s)$ as a bivariate Gaussian. Each action moves the agent by α in the corresponding direction. In particular, we consider two versions of the environment: I) *shared dynamics* where $\alpha = 1$ and II) *puddle-based dynamics* where puddles also slow-down the agent by: $\alpha = (1 + 5\sum_{u\in\mathcal{U}} W_u(s'))^{-1}$. Finally, a white Gaussian noise of $\sigma_r^2 = 0.01$ and $\sigma_p^2 = 0.04$ is added to the reward and the transition model, respectively. In our experiments we set $\gamma = 0.99$ and a maximum horizon of 50 time-steps.

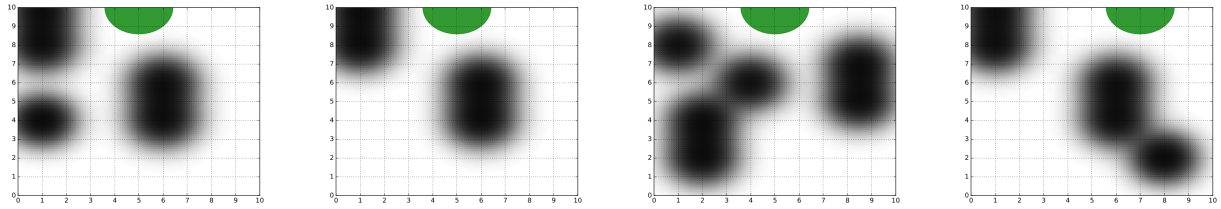

Figure 4. From left to right: the target task and the three source tasks. The agent always starts in the bottom-left corner and must reach the goal area (shown in green). Puddles are shown in black.

We provide additional details on the puddle world experiments. The target task and the three source tasks can be seen in Figure 4. Notice that the optimal paths to solve each task have at least a small overlapping, thus allowing some knowledge transfer. However, the optimal policy for one task is likely to cross a puddle if carelessly used in another domain. This makes the transfer problem more challenging since the algorithm has to figure out which samples should be retained and which should be discarded.

In both experiments, 20 episodes were generated beforehand from each source task. For IWFQI, a Gaussian process was fitted on each of the three source datasets using the squared exponential kernel. The noises of the reward and transition models were estimated as 10 times their true value. In each algorithm, FQI was run for 50 iterations using Extra-Trees with 50 estimators and a minimum of 2 samples to split a node. Results were averaged over 20 independent runs.

C.2. Acrobot

We provide a precise description of the two tasks used in the Acrobot experiment. For both tasks, the state-space is composed of the two link angles (θ_1, θ_2) and their velocities $(\dot{\theta}_1, \dot{\theta}_2)$. The transition dynamics are the ones described in (Sutton & Barto, 1998). The agent can only apply a torque of $+2$ or -2 to the joint between the two links. The initial state is $(\theta_1, 0, 0, 0)$, where $\theta_1 \sim \mathcal{U}(-2, 2)$. Performance is evaluated starting from multiple states $(\theta_1, 0, 0, 0)$, with θ_1 evenly spaced in $[-2, 2]$. The swing-up task has reward:

$$R_{sw}(\theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2) = -cos(\theta_1) - cos(\theta_1 + \theta_2) - 2, \tag{30}$$

and terminates whenever $-cos(\theta_1) - cos(\theta_1 + \theta_2) > 1$ or 100 time-steps are reached. The constant-spin task has reward:

$$R_{cs}(\theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2) = -|\dot{\theta}_1 - \pi|, \tag{31}$$

and terminates whenever 100 time-steps are reached.

We collect 100 episodes from the first source task (corresponding to 3400 samples) and 50 episodes from the second source task (corresponding to 5000 samples). For all algorithms, FQI uses extra-trees with 50 estimators and a minimum of 20 samples to split a node. Data is collected in batches of 10 episodes using an ϵ-greedy policy ($\epsilon = 0.1$). For IWFQI, GPs use the squared exponential kernel with parameters estimated by maximum likelihood on the data.

To further demonstrate the advantages of our approach, we show what happens when only the constant-spin source task is available. Clearly, most of the reward samples should be discarded, and conversely for the transition samples. As we can see from Figure 5a and 5b, RBT now performs significantly worse than FQI. This is due to the fact that, by transferring samples jointly, it cannot avoid introducing bias. Our approach, on the other hand, is able to discard the reward samples, thus being robust to negative transfer. Furthermore, it achieves a little improvement over FQI due to the few samples transferred.

C.3. Water Reservoir Control

All tasks used in this experiment are modeled according to the dynamics described in the paper. For the sake of simplicity, each water reservoir is supposed to have capacity of 500 Mm^3, minimum storage of 50 Mm^3, flooding threshold of 300 Mm^3, and per-day demand of 10 Mm^3. Due to the different geographic locations, each task has different inflow function $i_j(t) = \bar{i}_j(t) + \mathcal{N}(0, \sigma_p^2)$, where $\sigma_p^2 = 2.0$ is the fixed noise variance. The different mean-inflow functions are shown in Figure 6a. Furthermore, each water reservoir weighs the flooding and demand objectives differently. This is

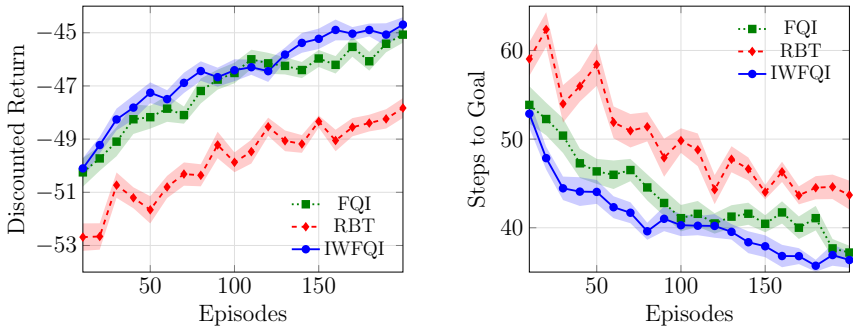

Figure 5. Transfer of samples from the constant-spin task to the swing-up task. (a) discounted expected reward and (b) number of steps before reaching a goal state.

modeled by changing the respective weights α and β. The values for all tasks are reported in Table 1. Notice that there is no source task that is globally similar to the target: either some reward structure is shared or some transition structure is, never both. This makes transfer very challenging since samples have to be accurately selected to prevent detrimental consequences.

In this experiment, we run FQI using extra-trees for 80 iterations with 100 estimators and a minimum of 10 samples to split a node. GPs use the anisotropic squared exponential kernel. For each of the 6 source reservoirs, we gather 30 years of historical data where controls are applied by a human operator's policy. We learn the target task by collecting batches of 1 year, each using an ϵ-greedy policy ($\epsilon = 0.3$) on the previously learned Q-function. Evaluation is performed by averaging 10 trajectories of 1 year, each starting from January 1^{st} and with an initial storage of 200 Mm^3 of water.

To better demonstrate the difficulty of this control problem, we run FQI for 500 episodes (equivalent to 500 years of interaction). Furthermore, to make the problem simpler, we allow the agent to sample the state-action space arbitrarily, so as to have a better exploration. The result is shown in Figure 6b. Although we significantly simplified the problem and we allowed FQI to gather an enormous amount of data, the algorithm still needs almost 500 years to achieve optimal performance. This demonstrates that solving this task by directly interacting with the real environment is clearly impractical. Thus, transfer of previous knowledge is, in this case, mandatory to achieve good performance.

Table 1. Reward parameters for the different water reservoirs.

Parameter	Target	Source 1	Source 2	Source 3	Source 4	Source 5	Source 6
α	0.3	0.8	0.35	0.7	0.4	0.6	0.45
β	0.7	0.2	0.65	0.3	0.6	0.4	0.55

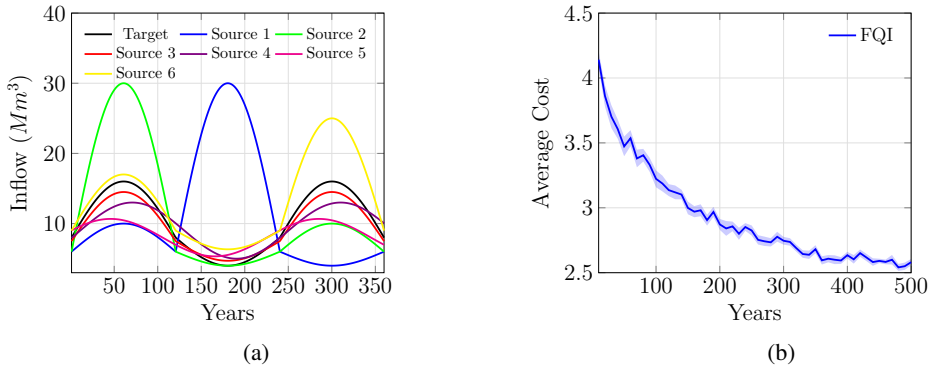

Figure 6. Water reservoir control. (a) Inflow profiles for all tasks. (b) Learning without transfer for 500 years.