# Supplementary Material

## A The Nonparametric Bellman Equation

This section contains the proofs of Theorem 1 and Theorem 3.

**Proposition 1.** *In the limit of infinite samples the NPBE defined in Definition 2 with a data-set* $\lim_{n\to\infty} D_n$ *collected under distribution* $\beta$ *on the state-action space and MDP* $\mathcal{M}$ *converges to*

$$\begin{aligned}\hat{V}_\pi(\mathbf{s}) \quad &= \quad \int_{\mathcal{S}\times\mathcal{A}} \varepsilon_\pi(\mathbf{s},\mathbf{z},\mathbf{b})\Big(R_{\mathbf{z},\mathbf{b}} + \gamma \int_{\mathcal{S}} \hat{V}_\pi(\mathbf{s}')\phi(\mathbf{s}',\mathbf{z}'_{\mathbf{z},\mathbf{b}})\,\mathrm{d}\mathbf{s}'\Big)\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}, \\ &\quad \textit{with} \quad R_{\mathbf{z},\mathbf{b}} \sim R(\mathbf{z},\mathbf{b}) \quad \forall(\mathbf{z},\mathbf{b}) \in \mathcal{S}\times\mathcal{A}, \\ &\quad \textit{with} \quad \mathbf{z}'_{\mathbf{z},\mathbf{b}} \sim P(\cdot|\mathbf{z},\mathbf{b}) \quad \forall(\mathbf{z},\mathbf{b}) \in \mathcal{S}\times\mathcal{A}. \end{aligned} \tag{12}$$

*and*

$$\begin{cases} \varepsilon_\pi(\mathbf{s},\mathbf{z},\mathbf{b}) \coloneqq \int_{\mathcal{A}} \dfrac{\psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})}{\int_{\mathcal{S},\mathcal{A}} \psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}\pi(\mathbf{a}|\mathbf{s})\,\mathrm{d}\mathbf{a} & \textit{if } \pi \textit{ is stochastic,} \\ \varepsilon^\pi_i(\mathbf{s}) \coloneqq \dfrac{\psi(\mathbf{s},\mathbf{z})\varphi(\pi(\mathbf{s}),\mathbf{b})}{\int_{\mathcal{S},\mathcal{A}} \psi(\mathbf{s},\mathbf{z})\varphi(\pi(\mathbf{s}),\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}} & \textit{otherwise.} \end{cases}$$

*Proof.*

$$\begin{aligned}\hat{V}_\pi(\mathbf{s}) \quad &= \quad \lim_{n\to\infty} \int_{\mathcal{A}} \frac{\sum_{i=1}^n \psi_i(\mathbf{s})\varphi_i(\mathbf{a})\Big(r_i + \gamma \int_{\mathcal{S}} \phi_i(\mathbf{s}')\hat{V}_\pi(\mathbf{s}')\,\mathrm{d}\mathbf{s}\Big)}{\sum_{i=1}^n \psi_j(\mathbf{s})\varphi_j(\mathbf{a})}\pi(\mathbf{a}|\mathbf{s})\,\mathrm{d}\mathbf{a} \\ &= \quad \int_{\mathcal{A}} \frac{\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \psi_i(\mathbf{s})\varphi_i(\mathbf{a})\Big(r_i + \gamma \int_{\mathcal{S}} \phi_i(\mathbf{s}')\hat{V}_\pi(\mathbf{s}')\,\mathrm{d}\mathbf{s}\Big)}{\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \psi_j(\mathbf{s})\varphi_j(\mathbf{a})}\pi(\mathbf{a}|\mathbf{s})\,\mathrm{d}\mathbf{a} \\ &= \quad \int_{\mathcal{A}} \frac{\int_{\mathcal{S}\times\mathcal{A}} \psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\Big(R(\mathbf{z},\mathbf{b}) + \gamma \int_{\mathcal{S}} \phi(\mathbf{s}',\mathbf{z}')p(\mathbf{z}'|\mathbf{b},\mathbf{z})\hat{V}_\pi(\mathbf{s}')\,\mathrm{d}\mathbf{s}\Big)\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}{\int_{\mathcal{S}\times\mathcal{A}} \psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}\pi(\mathbf{a}|\mathbf{s})\,\mathrm{d}\mathbf{a}. \end{aligned}$$

Analogously we can derive the deterministic policy case. □

**Proposition 2.** *Both for finite samples and infinite samples, when* $R$ *is bounded by* $-R_{max}$ *and* $R_{max}$ *(where* $R_{max}$ *is non-negative defined), then the solution of the NPBE if exists it is bounded between* $\frac{-2R_{max}}{1-\gamma}$ *and* $\frac{2R_{max}}{1-\gamma}$.

*Proof.* Starting with the finite samples case. Suppose by absurd proposition that if the NPBE admits a solution $\hat{V}_\pi$ then $\sup_{\mathbf{s}} |\hat{V}_\pi(\mathbf{s})| = \frac{R_{\max}}{1-\gamma} + \epsilon$ with $\epsilon > 0$ strictly positive (and eventually $+\infty$). It immediately follows that $\sup_{\mathbf{s}_1,\mathbf{s}_2} |\hat{V}_\pi(\mathbf{s}_1) - \hat{V}_\pi(\mathbf{s}_2)| = \frac{2R_{\max}}{1-\gamma} + 2\epsilon$. Expanding this term

$$\begin{aligned}\sup_{\mathbf{s}_1,\mathbf{s}_2} |\hat{V}_\pi(\mathbf{s}_1) - \hat{V}_\pi(\mathbf{s}_2)| \quad &= \quad \sup_{\mathbf{s}_1,\mathbf{s}_2} \left|\Big(\boldsymbol{\varepsilon}^T_\pi(\mathbf{s}_1) - \boldsymbol{\varepsilon}^T_\pi(\mathbf{s}_2)\Big)\Big(\mathbf{r} + \gamma \int_{\mathcal{S}} \boldsymbol{\phi}(\mathbf{s}')\hat{V}_\pi(\mathbf{s}')\,\mathrm{d}\mathbf{s}'\Big)\right| \\ &\leq \quad \sup_{\mathbf{s}_1,\mathbf{s}_2} \left|\boldsymbol{\varepsilon}^T_\pi(\mathbf{s}_1) - \boldsymbol{\varepsilon}^T_\pi(\mathbf{s}_2)\right|\left|\mathbf{r} + \gamma \int_{\mathcal{S}} \boldsymbol{\phi}(\mathbf{s}')\hat{V}_\pi(\mathbf{s}')\,\mathrm{d}\mathbf{s}'\right| \\ &\leq \quad \sup_{\mathbf{s}_1,\mathbf{s}_2} \Big(\big|\boldsymbol{\varepsilon}^T_\pi(\mathbf{s}_1)\big| + \big|\boldsymbol{\varepsilon}^T_\pi(\mathbf{s}_2)\big|\Big)\left|\mathbf{r} + \gamma \int_{\mathcal{S}} \boldsymbol{\phi}(\mathbf{s}')\hat{V}_\pi(\mathbf{s}')\,\mathrm{d}\mathbf{s}'\right|. \end{aligned}$$

Notice that $\boldsymbol{\varepsilon}_\pi^T(\mathbf{s})$ is a stochastic vector (non-negative definite and sums up to 0),

$$\begin{aligned}
&\sup_{\mathbf{s}_1,\mathbf{s}_2}\Big(\big|\boldsymbol{\varepsilon}_\pi^T(\mathbf{s}_1)\big|+\big|\boldsymbol{\varepsilon}_\pi^T(\mathbf{s}_2)\big|\Big)\Big|\mathbf{r}+\gamma\int_{\mathcal{S}}\boldsymbol{\phi}(\mathbf{s}')\hat{V}_\pi(\mathbf{s}')\,\mathrm{d}\mathbf{s}'\Big| \\
\leq\ & 2R_{\max}+\gamma\sup_{\mathbf{s}_1,\mathbf{s}_2}\Big(\big|\boldsymbol{\varepsilon}_\pi^T(\mathbf{s}_1)\big|+\big|\boldsymbol{\varepsilon}_\pi^T(\mathbf{s}_2)\big|\Big)\Big|\int_{\mathcal{S}}\boldsymbol{\phi}(\mathbf{s}')\hat{V}_\pi(\mathbf{s}')\,\mathrm{d}\mathbf{s}'\Big| \\
\leq\ & 2R_{\max}+\gamma\Big(\frac{R_{\max}}{1-\gamma}+\epsilon\Big)\sup_{\mathbf{s}_1,\mathbf{s}_2}\Big(\big|\boldsymbol{\varepsilon}_\pi^T(\mathbf{s}_1)\big|+\big|\boldsymbol{\varepsilon}_\pi^T(\mathbf{s}_2)\big|\Big)\int_{\mathcal{S}}\boldsymbol{\phi}(\mathbf{s}')\,\mathrm{d}\mathbf{s}' \\
=\ & 2R_{\max}+\gamma\Big(\frac{R_{\max}}{1-\gamma}+\epsilon\Big)\sup_{\mathbf{s}_1,\mathbf{s}_2}\Big(\big|\boldsymbol{\varepsilon}_\pi^T(\mathbf{s}_1)\big|+\big|\boldsymbol{\varepsilon}_\pi^T(\mathbf{s}_2)\big|\Big)\mathbf{1} \\
\leq\ & 2R_{\max}+\gamma\Big(\frac{2R_{\max}}{1-\gamma}+2\epsilon\Big),
\end{aligned}$$

which implies that

$$\begin{aligned}
\sup_{\mathbf{s}_1,\mathbf{s}_2}|\hat{V}_\pi(\mathbf{s}_1)-\hat{V}_\pi(\mathbf{s}_2)| &\leq 2R_{\max}+\gamma\Big(\frac{2R_{\max}}{1-\gamma}+2\epsilon\Big) \\
\implies 2\frac{R_{\max}}{1-\gamma}+2\epsilon &\leq 2R_{\max}+\gamma\Big(\frac{2R_{\max}}{1-\gamma}+2\epsilon\Big) \\
\implies 0 &\leq \epsilon\big(\gamma(1-\gamma)-1\big). \qquad (13)
\end{aligned}$$

Since $\gamma(1-\gamma)-1$ is always negative (we defined $0\leq\gamma<1$), then there are no positive values for $\epsilon$ which satisfy the inequality, which is in clear contradiction with the absurd premise. For the infinite samples case we can do similar reasoning noting that $\phi,\beta,P$ are probability measures. □

**Proposition 3.** *If $R$ is bounded by $R_{max}$ and if $f^*:\mathcal{S}\to\mathbb{R}$ satisfies the NPBE, then there is no other function $f:\mathcal{S}\to\mathbb{R}$ for which $\exists\mathbf{z}\in\mathcal{S}$ and $|f^*(\mathbf{z})-f(\mathbf{z})|>0$.*

*Proof.* Suppose, by absurd assumption, that a function $g:\mathcal{S}\to\mathbb{R}$ exists such that $f(\mathbf{s})+g(\mathbf{s})$ satisfies Equation (12) for every $\mathbf{s}\in\mathcal{S}$ and a constant $G\in\mathbb{R}^+$ exists for which $|g(\mathbf{z})|>G$. Note that the existence of $f:\mathcal{S}\to\mathbb{R}$ as a solution for the NPBE implies the existence of

$$\int_{\mathcal{S}}\boldsymbol{\varepsilon}_\pi^T(\mathbf{s})\boldsymbol{\phi}(\mathbf{s}')f^*(\mathbf{s}')\,\mathrm{d}\mathbf{s}'\in\mathbb{R}, \qquad (14)$$

and similarly, the existence of $f(\mathbf{s})\in\mathbb{R}$ with $f(\mathbf{s})=f^*(\mathbf{s})+g(\mathbf{s})$ as a solution of the NPBE implies that

$$\int_{\mathcal{S}}\boldsymbol{\varepsilon}_\pi^T(\mathbf{s})\boldsymbol{\phi}(\mathbf{s}')f^*(\mathbf{s}')+g(\mathbf{s}')\,\mathrm{d}\mathbf{s}'\in\mathbb{R}. \qquad (15)$$

Note that the existence of the integral in Equations (14) and (15) implies

$$\int_{\mathcal{S}}\boldsymbol{\varepsilon}_\pi^T(\mathbf{s})\boldsymbol{\phi}(\mathbf{s}')g(\mathbf{s}')\,\mathrm{d}\mathbf{s}'\in\mathbb{R}. \qquad (16)$$

Note that

$$\begin{aligned}
|f^*(\mathbf{s})-f(\mathbf{s})| &= \Big|f^*(\mathbf{s})-\boldsymbol{\varepsilon}_\pi^T(\mathbf{s})\Big(\mathbf{r}+\gamma\int_{\mathcal{S}}\boldsymbol{\phi}(\mathbf{s}')\big(f(\mathbf{s}')+g(\mathbf{s}')\big)\,\mathrm{d}\mathbf{s}'\Big)\Big| \\
&= \Big|\boldsymbol{\varepsilon}_\pi^T(\mathbf{s})\Big(\mathbf{r}+\gamma\int_{\mathcal{S}}\boldsymbol{\phi}(\mathbf{s}')g(\mathbf{s}')\,\mathrm{d}\mathbf{s}'\Big)-\boldsymbol{\varepsilon}_\pi^T(\mathbf{s})\Big(\mathbf{r}+\gamma\int_{\mathcal{S}}\boldsymbol{\phi}(\mathbf{s}')\big(f^*(\mathbf{s}')+g(\mathbf{s}')\big)\,\mathrm{d}\mathbf{s}'\Big)\Big| \\
&= \gamma\Big|\boldsymbol{\varepsilon}_\pi^T(\mathbf{s})\int_{\mathcal{S}}\boldsymbol{\phi}(\mathbf{s}')g(\mathbf{s}')\,\mathrm{d}\mathbf{s}'\Big| \\
\implies |g(\mathbf{s})| &= \gamma\Big|\boldsymbol{\varepsilon}_\pi^T(\mathbf{s})\int_{\mathcal{S}}\boldsymbol{\phi}(\mathbf{s}')g(\mathbf{s}')\,\mathrm{d}\mathbf{s}'\Big|.
\end{aligned}$$

Using Jensen's inequality

$$|g(\mathbf{s})| \quad \leq \quad \gamma \boldsymbol{\varepsilon}_\pi^T(\mathbf{s}) \int_{\mathcal{S}} \boldsymbol{\phi}(\mathbf{s}')|g(\mathbf{s}')|\,\mathrm{d}\mathbf{s}'.$$

Note that since both $f^*$ and $f$ are bounded by $\frac{R_{\max}}{1-\gamma}$ then $|g(\mathbf{s})| \leq \frac{2R_{\max}}{1-\gamma}$, thus

$$\begin{aligned}
|g(\mathbf{s})| &\leq \gamma \boldsymbol{\varepsilon}_\pi^T(\mathbf{s}) \int_{\mathcal{S}} \boldsymbol{\phi}(\mathbf{s}')|g(\mathbf{s}')|\,\mathrm{d}\mathbf{s}' \qquad (17)\\
&\leq \gamma 2 \frac{R_{\max}}{1-\gamma} \boldsymbol{\varepsilon}_\pi^T(\mathbf{s}) \int_{\mathcal{S}} \boldsymbol{\phi}(\mathbf{s}')\,\mathrm{d}\mathbf{s}'\\
&= \gamma 2 \frac{R_{\max}}{1-\gamma}\\
\implies |g(\mathbf{s})| &\leq \gamma \frac{2R_{\max}}{1-\gamma}\\
\implies |g(\mathbf{s})| &\leq \gamma^2 \frac{2R_{\max}}{1-\gamma} \qquad \text{using (17)}\\
\implies |g(\mathbf{s})| &\leq \gamma^3 \frac{2R_{\max}}{1-\gamma} \qquad \text{using (17)}\\
&\dots\\
\implies |g(\mathbf{s})| &\leq 0,
\end{aligned}$$

which is in clear disagreement with the assumption made. Again here a similar procedure shows the same result for the infinite case. □

**Proof of Theorem 1**

*Proof.* Saying that $\hat{V}_\pi^*$ is a solution for Equation (12) is equivalent to saying

$$\hat{V}_\pi^*(\mathbf{s}) - \boldsymbol{\varepsilon}^\pi(\mathbf{s}) \left( \mathbf{r} + \gamma \int_{\mathcal{S}} \boldsymbol{\phi}(\mathbf{s}') \hat{V}_\pi^*(\mathbf{s}')\,\mathrm{d}\mathbf{s}' \right) = 0 \qquad \forall \mathbf{s} \in \mathcal{S}.$$

We can verify that by simple algebraic manipulation

$$\begin{aligned}
&\hat{V}_\pi^*(\mathbf{s}) - \boldsymbol{\varepsilon}_\pi^T(\mathbf{s}) \left( \mathbf{r} + \gamma \int_{\mathcal{S}} \boldsymbol{\phi}(\mathbf{s}') \hat{V}_\pi^*(\mathbf{s}')\,\mathrm{d}\mathbf{s}' \right)\\
= \;& \boldsymbol{\varepsilon}_\pi^T(\mathbf{s}) \boldsymbol{\Lambda}_\pi^{-1} \mathbf{r} - \boldsymbol{\varepsilon}^\pi(\mathbf{s}) \left( \mathbf{r} + \gamma \int_{\mathcal{S}} \boldsymbol{\phi}(\mathbf{s}') \boldsymbol{\varepsilon}_\pi^T(\mathbf{s}') \boldsymbol{\Lambda}_\pi^{-1} \mathbf{r}\,\mathrm{d}\mathbf{s}' \right)\\
= \;& \boldsymbol{\varepsilon}_\pi^T(\mathbf{s}) \left( \boldsymbol{\Lambda}_\pi^{-1} \mathbf{r} - \mathbf{r} - \gamma \int_{\mathcal{S}} \boldsymbol{\phi}(\mathbf{s}') \boldsymbol{\varepsilon}_\pi^T(\mathbf{s}') \boldsymbol{\Lambda}_\pi^{-1} \mathbf{r}\,\mathrm{d}\mathbf{s}' \right)\\
= \;& \boldsymbol{\varepsilon}_\pi^T(\mathbf{s}) \left( \left( I - \gamma \int_{\mathcal{S}} \boldsymbol{\phi}(\mathbf{s}') \boldsymbol{\varepsilon}_\pi^T(\mathbf{s}')\,\mathrm{d}\mathbf{s}' \right) \boldsymbol{\Lambda}_\pi^{-1} \mathbf{r} - \mathbf{r} \right)\\
= \;& \boldsymbol{\varepsilon}_\pi^T(\mathbf{s}) \left( \boldsymbol{\Lambda}_\pi \boldsymbol{\Lambda}_\pi^{-1} \mathbf{r} - \mathbf{r} \right)\\
= \;& 0. \qquad (18)
\end{aligned}$$

Since equation (12) has (at least) one solution, Proposition 3 guarantees that the solution $(\hat{V}_\pi^*)$ is unique. □

**Proof of Theorem 3.**

*Proof.* We perform the derivation for the stochastic policy, however the same derivation applies for the deterministic case almost identically. Expanding $\left| \mathbb{E}_D[\overline{V}_D(\mathbf{s})] - V^*(\mathbf{s}) \right|$ using the NPBE and the classic Bellman

equation,

$$\begin{aligned}\left|\mathop{\mathbb{E}}_{D}[\overline{V}_D(\mathbf{s})] - V^*(\mathbf{s})\right| &= \left|\mathop{\mathbb{E}}_{D}\left[\int_{\mathcal{S}\times\mathcal{A}} \varepsilon_\pi(\mathbf{s},\mathbf{z},\mathbf{b})\left(R_{\mathbf{z},\mathbf{b}} + \gamma\int_{\mathcal{S}} V_D(\mathbf{s}')\phi(\mathbf{s}',\mathbf{z}'_{\mathbf{z},\mathbf{b}})\,\mathrm{d}\mathbf{s}\right)\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}\right]\right. \\ &\quad \left. - \int_{\mathcal{A}}\left(\overline{R}(\mathbf{s},\mathbf{a}) + \gamma\int_{\mathcal{S}} V^*(\mathbf{s}')p(\mathbf{s}'|\mathbf{s},\mathbf{a})\,\mathrm{d}\mathbf{s}'\right)\pi(\mathbf{a}|\mathbf{s})\,\mathrm{d}\mathbf{a}\right|. \end{aligned} \tag{19}$$

As can be easily verified, $\varepsilon_\pi(\mathbf{s},\mathbf{z},\mathbf{b})\beta(\mathbf{z},\mathbf{b})$ is a density distribution over $\mathbf{z},\mathbf{b}$. Hence Equation (19) can be rewritten

$$\begin{aligned}&\left|\mathop{\mathbb{E}}_{D}\left[\int_{\mathcal{S}\times\mathcal{A}} \varepsilon_\pi(\mathbf{s},\mathbf{z},\mathbf{b})\left(R_{\mathbf{z},\mathbf{b}} + \gamma\int_{\mathcal{S}} V_D(\mathbf{s}')\phi(\mathbf{s}',\mathbf{z}'_{\mathbf{z},\mathbf{b}})\,\mathrm{d}\mathbf{s}'\right)\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}\right]\right. \\ &\quad \left. - \int_{\mathcal{A}}\left(\overline{R}(\mathbf{s},\mathbf{a}) + \gamma\int_{\mathcal{S}} V^*(\mathbf{s}')p(\mathbf{s}'|\mathbf{s},\mathbf{a})\,\mathrm{d}\mathbf{s}'\right)\pi(\mathbf{a}|\mathbf{s})\,\mathrm{d}\mathbf{a}\right| \\ =\ &\left|\mathop{\mathbb{E}}_{D}\left[\int_{\mathcal{A}} \frac{\int_{\mathcal{S}\times\mathcal{A}}\psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\big(R_{\mathbf{z},\mathbf{b}} - \overline{R}(\mathbf{s},\mathbf{a})\big)\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}{\int_{\mathcal{S},\mathcal{A}}\psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}\pi(\mathbf{a}|\mathbf{s})\,\mathrm{d}\mathbf{a}\right]\right. \\ +\ &\left.\gamma\int_{\mathcal{A}}\mathop{\mathbb{E}}_{D}\left[\frac{\int_{\mathcal{S}\times\mathcal{A}}\psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\big(\int_{\mathcal{S}} V_D(\mathbf{s}')\phi(\mathbf{s}',\mathbf{z}'_{\mathbf{z},\mathbf{b}})\,\mathrm{d}\mathbf{s}' - \int_{\mathcal{S}} V^*(\mathbf{s}')p(\mathbf{s}'|\mathbf{s},\mathbf{a})\,\mathrm{d}\mathbf{s}'\big)\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}{\int_{\mathcal{S},\mathcal{A}}\psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}\right]\pi(\mathbf{a}|\mathbf{s})\,\mathrm{d}\mathbf{a}\right| \\ \leq\ &\left|\mathop{\mathbb{E}}_{D}\left[\int_{\mathcal{A}}\underbrace{\frac{\int_{\mathcal{S}\times\mathcal{A}}\psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\big(R_{\mathbf{z},\mathbf{b}} - \overline{R}(\mathbf{s},\mathbf{a})\big)\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}{\int_{\mathcal{S},\mathcal{A}}\psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}}_{\mathrm{A}}\pi(\mathbf{a}|\mathbf{s})\,\mathrm{d}\mathbf{a}\right]\right| \\ +\ &\gamma\left|\int_{\mathcal{A}}\underbrace{\mathop{\mathbb{E}}_{D}\left[\frac{\int_{\mathcal{S}\times\mathcal{A}}\psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\big(\int_{\mathcal{S}} V_D(\mathbf{s}')\phi(\mathbf{s}',\mathbf{z}'_{\mathbf{z},\mathbf{b}})\,\mathrm{d}\mathbf{s}' - \int_{\mathcal{S}} V^*(\mathbf{s}')p(\mathbf{s}'|\mathbf{s},\mathbf{a})\,\mathrm{d}\mathbf{s}'\big)\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}{\int_{\mathcal{S},\mathcal{A}}\psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}\right]}_{\mathrm{B}}\pi(\mathbf{a}|\mathbf{s})\,\mathrm{d}\mathbf{a}\right| \\ \leq\ &\mathrm{A}_{\mathrm{Bias}} + \gamma\mathrm{B}_{\mathrm{Bias}}. \end{aligned} \tag{20}$$

It is evident that the term $A$ is the Nadaraya-Watson kernel regression, as it is possible to observe in the beginnin of the proof at page twelve of Tosatto et al. (2020), therefore Theorem 2 applies

$$\mathrm{A}_{\mathrm{Bias}} = \frac{L_R\sum_{k=1}^d \mathbf{h}_k\left(\prod_{i\neq k}^d e^{\frac{L_\beta^2\mathbf{h}_i^2}{2}}\left(1+\mathrm{erf}\left(\frac{\mathbf{h}_i L_\beta}{\sqrt{2}}\right)\right)\right)\left(\frac{1}{\sqrt{2\pi}} + L_\beta\mathbf{h}_k\frac{e^{\frac{L_\beta^2\mathbf{h}_k^2}{2}}}{2}\left(1+\mathrm{erf}\left(\frac{\mathbf{h}_k L_\beta}{\sqrt{2}}\right)\right)\right)}{\prod_{i=1}^d e^{\frac{L_\beta^2 h_i^2}{2}}\left(1-\mathrm{erf}\left(\frac{\mathbf{h}_i L_\beta}{\sqrt{2}}\right)\right)},$$

where $\mathbf{h} = [\mathbf{h}_\psi, \mathbf{h}_\varphi]$ and $d = d_s + d_a$.
Returning to the estimate of $\mathrm{B}_{\mathrm{Bias}}$

$$\begin{aligned}&\left|\int_{\mathcal{A}}\mathop{\mathbb{E}}_{D}\left[\frac{\int_{\mathcal{S}\times\mathcal{A}}\psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\big(\int_{\mathcal{S}} V_D(\mathbf{s}')\phi(\mathbf{s}',\mathbf{z}'_{\mathbf{z},\mathbf{b}})\,\mathrm{d}\mathbf{s}' - \int_{\mathcal{S}} V^*(\mathbf{s}')p(\mathbf{s}'|\mathbf{s},\mathbf{a})\,\mathrm{d}\mathbf{s}'\big)\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}{\int_{\mathcal{S},\mathcal{A}}\psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}\right]\pi(\mathbf{a}|\mathbf{s})\,\mathrm{d}\mathbf{a}\right| \\ =\ &\left|\int_{\mathcal{A}}\frac{\int_{\mathcal{S}\times\mathcal{A}}\psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\big(\int_{\mathcal{S}}\mathbb{E}\left[V_D(\mathbf{s}')\phi(\mathbf{s}',\mathbf{z}'_{\mathbf{z},\mathbf{b}})\right]\mathrm{d}\mathbf{s}' - \int_{\mathcal{S}} V^*(\mathbf{s}')p(\mathbf{s}'|\mathbf{s},\mathbf{a})\,\mathrm{d}\mathbf{s}'\big)\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}{\int_{\mathcal{S},\mathcal{A}}\psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}\pi(\mathbf{a}|\mathbf{s})\,\mathrm{d}\mathbf{a}\right| \end{aligned}$$

One my ask whether the terms in $\mathbb{E}[V_D(\mathbf{s}')\phi(\mathbf{s}',\mathbf{z}'_{\mathbf{z},\mathbf{b}})]$ are uncorrelated. The answer it is affirmative, since, even if $V_D$ depends by $\mathbf{z}_{\mathbf{z},\mathbf{b}}$ (integral in Equation (12)), this corresponds only in the variation of a single point in the integral, and therefore the overall estimate does not change. This argument, however, does not immediately hold for the case of an infinitesimal bandwidth, and therefore we provide the results for that case separately.

**For Finite Bandwidth:**

$$
\begin{aligned}
&\left| \int_{\mathcal{A}} \frac{\int_{\mathcal{S}\times\mathcal{A}} \psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\big(\int_{\mathcal{S}} \mathbb{E}\left[V_D(\mathbf{s}')\phi(\mathbf{s}',\mathbf{z}'_{\mathbf{z},\mathbf{b}})\right] \mathrm{d}\mathbf{s}' - \int_{\mathcal{S}} V^*(\mathbf{s}')p(\mathbf{s}'|\mathbf{s},\mathbf{a})\,\mathrm{d}\mathbf{s}'\big)\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}{\int_{\mathcal{S},\mathcal{A}} \psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}} \pi(\mathbf{a}|\mathbf{s})\,\mathrm{d}\mathbf{a}\right| \\
\leq\ & \max_{\mathbf{s},\mathbf{a}} \left| \frac{\int_{\mathcal{S}\times\mathcal{A}} \psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\big(\int_{\mathcal{S}\times\mathcal{S}} \overline{V}(\mathbf{z}')\phi(\mathbf{z}',\mathbf{s}')p(\mathbf{s}'|\mathbf{s},\mathbf{a})\,\mathrm{d}\mathbf{s}'\,\mathrm{d}\mathbf{z}' - \int_{\mathcal{S}} V^*(\mathbf{s}')p(\mathbf{s}'|\mathbf{s},\mathbf{a})\,\mathrm{d}\mathbf{s}'\big)\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}{\int_{\mathcal{S},\mathcal{A}} \psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}} \right| \\
=\ & \max_{\mathbf{s},\mathbf{a}} \left| \frac{\int_{\mathcal{S}\times\mathcal{A}} \psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\big(\int_{\mathcal{S}}\int_{\mathcal{S}} \big(\overline{V}(\mathbf{z}')\phi(\mathbf{z}',\mathbf{s}') - V^*(\mathbf{s}')\big)p(\mathbf{s}'|\mathbf{s},\mathbf{a})\,\mathrm{d}\mathbf{s}'\,\mathrm{d}\mathbf{z}'\big)\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}{\int_{\mathcal{S},\mathcal{A}} \psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}} \right| \\
\leq\ & \max_{\mathbf{s},\mathbf{a},\mathbf{s}'} \left| \frac{\int_{\mathcal{S}\times\mathcal{A}} \psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\big(\int_{\mathcal{S}} \overline{V}(\mathbf{z}')\phi(\mathbf{z}',\mathbf{s}') - V^*(\mathbf{s}')\,\mathrm{d}\mathbf{z}'\big)\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}{\int_{\mathcal{S},\mathcal{A}} \psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}} \right| \\
=\ & \max_{\mathbf{s},\mathbf{a},\mathbf{s}'} \left| \frac{\int_{\mathcal{S}\times\mathcal{A}} \psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}}{\int_{\mathcal{S},\mathcal{A}} \psi(\mathbf{s},\mathbf{z})\varphi(\mathbf{a},\mathbf{b})\beta(\mathbf{z},\mathbf{b})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{b}} \left( \int_{\mathcal{S}} \overline{V}(\mathbf{z}')\phi(\mathbf{z}',\mathbf{s}') - V^*(\mathbf{s}')\,\mathrm{d}\mathbf{z}' \right) \right| \\
=\ & \max_{\mathbf{s},\mathbf{a},\mathbf{s}'} \left| \int_{\mathcal{S}} \overline{V}(\mathbf{z}')\phi(\mathbf{z}',\mathbf{s}') - V^*(\mathbf{s}')\,\mathrm{d}\mathbf{z}' \right| \\
=\ & \max_{\mathbf{s},\mathbf{a},\mathbf{s}'} \left| \int_{\mathcal{S}} \overline{V}(\mathbf{s}'+\mathbf{l})\phi(\mathbf{s}+\mathbf{l},\mathbf{s}') - V^*(\mathbf{s}')\,\mathrm{d}\mathbf{l} \right|. \qquad (21)
\end{aligned}
$$

Note that

$$
\phi(\mathbf{s}'+\mathbf{l},\mathbf{s}') = \prod_{i=1}^{d_s} \frac{e^{-\frac{l_i^2}{2h_{\phi,i}^2}}}{\sqrt{2\pi h_{\phi,i}^2}},
$$

thus

$$
\begin{aligned}
&\max_{\mathbf{s},\mathbf{a},\mathbf{s}'} \left| \int_{\mathcal{S}} \overline{V}(\mathbf{s}'+\mathbf{l})\phi(\mathbf{s}+\mathbf{l},\mathbf{s}') - V^*(\mathbf{s}')\,\mathrm{d}\mathbf{l} \right| \\
\leq\ & \max_{\mathbf{s},\mathbf{a},\mathbf{s}'} \left| \overline{V}(\mathbf{s}') - V^*(\mathbf{s}') \right| + \int_{\mathcal{S}} L_V \Big( \sum_{i=1}^{d_s} |l_i| \Big) \prod_{i=1}^{d_s} \frac{e^{-\frac{l_i^2}{2h_{\phi,i}^2}}}{\sqrt{2\pi h_{\phi,i}^2}}\,\mathrm{d}\mathbf{l}
\end{aligned}
$$

Using Proposition **??**

$$
\begin{aligned}
&\max_{\mathbf{s},\mathbf{a},\mathbf{s}'} \left| \overline{V}(\mathbf{s}') - V^*(\mathbf{s}') \right| + L_V \int_{\mathcal{S}} \left( \sum_{i=1}^{d_s} |l_i| \right) \prod_{i=1}^{d_s} \frac{e^{-\frac{l_i^2}{2h_{\phi,i}^2}}}{\sqrt{2\pi h_{\phi,i}^2}}\,\mathrm{d}\mathbf{l} \\
=\ & \max_{\mathbf{s},\mathbf{a},\mathbf{s}'} \left| \overline{V}(\mathbf{s}') - V^*(\mathbf{s}') \right| + L_V \sum_{k=1}^{d_s} \left( \prod_{i\neq k}^{d_s} \int_{-\infty}^{+\infty} \frac{e^{-\frac{l_i^2}{2h_{\phi,i}^2}}}{\sqrt{2\pi h_{\phi,i}^2}}\,\mathrm{d}l_i \right) \int_{-\infty}^{+\infty} |l_k| \frac{e^{-\frac{l_k^2}{2h_{\phi,k}^2}}}{\sqrt{2\pi h_{\phi,k}^2}}\,\mathrm{d}l_k \\
=\ & \max_{\mathbf{s},\mathbf{a},\mathbf{s}'} \left| \overline{V}(\mathbf{s}') - V^*(\mathbf{s}') \right| + L_V 2 \sum_{k=1}^{d_s} \int_{0}^{+\infty} l_k \frac{e^{-\frac{l_k^2}{2h_{\phi,k}^2}}}{\sqrt{2\pi h_{\phi,k}^2}}\,\mathrm{d}l_k \\
=\ & \max_{\mathbf{s},\mathbf{a},\mathbf{s}'} \left| \overline{V}(\mathbf{s}') - V^*(\mathbf{s}') \right| + L_V \sum_{k=1}^{d_s} \frac{h_{\phi,k}}{\sqrt{2\pi}}
\end{aligned} \qquad (22)
$$

which means that when $\mathbf{h}$ not infinitesimal

$$
\left| \overline{V}(\mathbf{s}) - V^*(\mathbf{s}) \right| \leq \mathrm{A}_{\mathrm{Bias}} + \gamma \left( \max_{\mathbf{s},\mathbf{a},\mathbf{s}'} \left| \overline{V}(\mathbf{s}') - V^*(\mathbf{s}') \right| + L_V \sum_{k=1}^{d_s} \frac{h_{\phi,k}}{\sqrt{2\pi}} \right).
$$

It is however known that $\left|\overline{V}(\mathbf{s}) - V^*(\mathbf{s})\right| \leq 2\frac{R_{\max}}{1-\gamma}$, thus

$$\left|\overline{V}(\mathbf{s}) - V^*(\mathbf{s})\right| \leq \mathrm{A}_{\mathrm{Bias}} + \gamma\left(\max_{\mathbf{s},\mathbf{a},\mathbf{s}'}\left|\overline{V}(\mathbf{s}') - V^*(\mathbf{s}')\right| + L_V \sum_{k=1}^{d_s} \frac{h_{\phi,k}}{\sqrt{2\pi}}\right) \tag{23}$$

$$\left|\overline{V}(\mathbf{s}) - V^*(\mathbf{s})\right| \leq \mathrm{A}_{\mathrm{Bias}} + \gamma\left(2\frac{R_{\max}}{1-\gamma} + L_V \sum_{k=1}^{d_s} \frac{h_{\phi,k}}{\sqrt{2\pi}}\right)$$

$$\implies \quad \left|\overline{V}(\mathbf{s}) - V^*(\mathbf{s})\right| \leq \mathrm{A}_{\mathrm{Bias}} + \gamma\left(\mathrm{A}_{\mathrm{Bias}} + \gamma\left(2\frac{R_{\max}}{1-\gamma} + L_V \sum_{k=1}^{d_s} \frac{h_{\phi,k}}{\sqrt{2\pi}}\right) + L_V \sum_{k=1}^{d_s} \frac{h_{\phi,k}}{\sqrt{2\pi}}\right) \quad \text{using Equation (23)}$$

$$\implies \quad \left|\overline{V}(\mathbf{s}) - V^*(\mathbf{s})\right| \leq \sum_{t=0}^{\infty} \gamma^t \left(\mathrm{A}_{\mathrm{Bias}} + \gamma L_V \sum_{k=1}^{d_s} \frac{h_{\phi,k}}{\sqrt{2\pi}}\right) \quad \text{using Equation (23)}$$

$$\implies \quad \left|\overline{V}(\mathbf{s}) - V^*(\mathbf{s})\right| \leq \frac{1}{1-\gamma}\left(\mathrm{A}_{\mathrm{Bias}} + \gamma L_V \sum_{k=1}^{d_s} \frac{h_{\phi,k}}{\sqrt{2\pi}}\right).$$

**For Infinitesimal Bandwidth:** In the case of an infinitesimal bandwidth note that, even if $V_D$ and $\phi$ are correlated the overall integral reduces only on a single point, and the same argument made in the case of finite bandwidth applies,

$$\int_{\mathcal{S}} \mathbb{E}\left[V_D(\mathbf{s}')\phi(\mathbf{s}', \mathbf{z}'_{\mathbf{z},\mathbf{b}})\right] \mathrm{d}\mathbf{s}' = \mathbb{E}\left[\int_{\mathcal{S}} V_D(\mathbf{s}')\phi(\mathbf{s}', \mathbf{z}'_{\mathbf{z},\mathbf{b}}\,\mathrm{d}\mathbf{s}')\,\mathrm{d}\mathbf{s}'\right] = \mathbb{E}\left[V_D(\mathbf{z}'_{\mathbf{z},\mathbf{b}})\right] = \int_{\mathcal{S}} \overline{V}_D(\mathbf{s}')p(\mathbf{s}'|\mathbf{s},\mathbf{a})\,\mathrm{d}\mathbf{s}'.$$

It follows that, proceeding similarly to Equation (21), we obtain

$$\left|\mathop{\mathbb{E}}_{D}[\overline{V}_D(\mathbf{s})] - V^*(\mathbf{s})\right| \leq \max_{\mathbf{s},\mathbf{a},\mathbf{s}'}\left|\overline{V}(\mathbf{s}') - V^*(\mathbf{s}')\right|, \tag{24}$$

which yields

$$\left|\overline{V}(\mathbf{s}) - V^*(\mathbf{s})\right| \leq \frac{1}{1-\gamma}\mathrm{A}_{\mathrm{Bias}}. \tag{25}$$

□

# B Empirical Evaluation Detail

## B.1 Linear Quadratic Regulator Experiment

Here we detail the experiment presented in Figure 1. We use a discrete infinite-horizon discounted Linear Quadratic Regulator system of the form

$$\max_{\vec{x}_t, \vec{u}_t} J = \frac{1}{2}\sum_{t=0}^{\infty} \gamma^t \left(\vec{x}_t^\top \mathbf{Q}\vec{x}_t + \vec{u}_t^\top \mathbf{R}\vec{u}_t\right)$$
$$\vec{x}_{t+1} = \mathbf{A}\vec{x}_t + \mathbf{B}\vec{u}_t \quad \forall t,$$

where $\vec{x}_t \in \mathbb{R}^{d_x}$, $\vec{u}_t \in \mathbb{R}^{d_u}$, $\mathbf{Q} \in \mathbb{R}^{d_x \times d_x}$, $\mathbf{R} \in \mathbb{R}^{d_u \times d_u}$, $\mathbf{A} \in \mathbb{R}^{d_x \times d_x}$, $\mathbf{B} \in \mathbb{R}^{d_x \times d_u}$, $\gamma \in [0, 1)$ and $\vec{x}_0$ given.

In this example we use consider a 2-dimensional problem with the following quantities

$$\mathbf{A} = \begin{bmatrix} 1.2 & 0 \\ 0 & 1.1 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.2 \end{bmatrix}$$
$$\mathbf{Q} = \begin{bmatrix} -0.5 & 0 \\ 0 & -0.25 \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}$$
$$\vec{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \gamma = 0.9.$$

For this LQR problem we impose a linear controller as a diagonal matrix

$$\mathbf{K} = \begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix}. \tag{26}$$

#### B.1.1 Deterministic Experiment

For each dataset we run 100 trajectories of 30 steps. Each trajectory is generated by following the dynamics of the described LQR and using at each time step a fixed policy initialized as

$$\mathbf{K} = \begin{bmatrix} k_1 + \varepsilon & \varepsilon \\ \varepsilon & k_2 + \varepsilon \end{bmatrix},\ \varepsilon \sim \mathcal{N}(0, 1),$$

where $k_1 = 0.7$ and $k_2 = -0.7$.

NOPG-D optimized for each dataset a policy encoded as in (26) with: learning rate 0.5 with ADAM optimizer; bandwidths (on average) for the state space $\vec{h}_\psi = [0.03, 0.05]$ and for the action space $\vec{h}_\varphi = [0.33, 0.27]$; discount factor $\gamma = 0.9$; and keeping 5 elements per row after sparsification of the $\mathbf{P}$ matrix.

DPG optimized for each dataset a policy encoded as in (26) with: learning rate 0.5 with ADAM optimizer; $Q$-function encoded as $Q(\vec{x}, \vec{u}) = \vec{x}^\top \mathbf{Q} \vec{x} + \vec{u}^\top \mathbf{R} \vec{u}$ (with $\mathbf{Q}$ and $\mathbf{R}$ to be learned); discount factor $\gamma = 0.9$; two target networks are kept to stabilize learning and soft-updated using $\tau = 0.01$ (similar to DDPG).

#### B.1.2 Stochastic Experiment

For each dataset we run 100 trajectories of 30 steps. Each trajectory is generated by following the dynamics of the described LQR, and using at each time step a stochastic policy as

$$\vec{u}_t = \mathbf{K}\vec{x}_t + \vec{\varepsilon},\ \vec{\varepsilon} \sim \mathcal{N}\left(\vec{\mu} = \vec{0}, \mathbf{\Sigma} = \mathrm{diag}(0.01, 0.01)\right), \tag{27}$$

where $\mathbf{K} = \mathrm{diag}(0.35, -0.35)$.

NOPG-S optimized for each dataset a policy encoded as in (27) with: learning rate 0.25 with ADAM optimizer; bandwidths (on average) for the state space $\vec{h}_\psi = [0.008, 0.003]$ and for the action space $\vec{h}_\varphi = [0.02, 0.02]$; discount factor $\gamma = 0.9$; and keeping 10 elements per row after sparsification of the $\mathbf{P}$ matrix.

PWIS optimized for each dataset a policy encoded as in (27) with: learning rate $2.5 \times 10^{-4}$ with ADAM optimizer; and discount factor $\gamma = 0.9$.

### B.2 Other Experiments Configurations

We use a policy encoded as neural network with parameters $\vec{\theta}$. A deterministic policy is encoded with a neural network $\mathbf{a} = f_{\vec{\theta}}(\mathbf{s})$. The stochastic policy is encoded as a Gaussian distribution with parameters determined by a neural network with two outputs, the mean and covariance. In this case we represent by $f_{\vec{\theta}}(\mathbf{s})$ the slice of the output corresponding to the mean and by $g_{\vec{\theta}}(\mathbf{s})$ the part of the output corresponding to the covariance.

NOPG can be described with the following hyper-parameters

| **NOPG Parameters** | Meaning |
|---|---|
| dataset sizes | number of samples contained in the dataset used for training |
| discount factor $\gamma$ | usual discount factor in infinite horizon MDP |
| state $\vec{h}_{\text{factor}}$ | constant used to decide the bandwidths for the state-space |
| action $\vec{h}_{\text{factor}}$ | constant used to decide the bandwidths for the action-space |
| policy | parametrization of the policy |
| policy output | how is the output of the policy encoded |
| learning rate | the learning rate and the gradient ascent algorithm used |
| $N_\pi^{\text{MC}}$ (NOPG-S) | number of samples drawn to compute the integral $\boldsymbol{\varepsilon}_\pi(\mathbf{s})$ with MonteCarlo sampling |
| $N_\phi^{\text{MC}}$ | number of samples drawn to compute the integral over the next state $\int \phi(\mathbf{s}')\, \mathrm{d}\mathbf{s}'$ |
| $N_{\mu_0}^{\text{MC}}$ | number of samples drawn to compute the integral over the initial distribution $\int \hat{V}_\pi(\mathbf{s})\mu_0(\mathbf{s})\, \mathrm{d}\mathbf{s}$ |
| policy updates | number of policy updates before returning the optimized policy |

A few considerations about NOPG parameters. If $N_{\phi}^{\text{MC}} = 1$ we use the mean of the kernel $\phi$ as a sample to approximate the integral over the next state. When optimizing a stochastic policy represented by a Gaussian distribution, we set and linearly decay the variance over the policy optimization procedure. The kernel bandwidths are computed in two steps: first we find the best bandwidth for each dimension of the state and action spaces using cross validation; second we multiply each bandwidth by an empirical constant factor ($\vec{h}_{\text{factor}}$). This second step is important to guarantee that the state and action spaces do not have a zero density. For instance, in a continuous action environment, when sampling actions from a uniform grid we have to guarantee that the space between the grid points have some density. The problem of estimating the bandwidth in kernel density estimation is well studied, but needs to be adapted to the problem at hand, specially with a low number of samples. We found this approach to work well for our experiments but it still can be improved.

### B.2.1 Pendulum with Uniform Dataset

Tables 3 and 4 describe the hyper-parameters used to run the experiment shown in the first plot of Figure 2.

**Dataset Generation** The dataset have been generated using a grid over the state-action space $\theta, \dot{\theta}, u$, where $\theta$ and $\dot{\theta}$ are respectively angle and angular velocity of the pendulum, and $u$ is the torque applied. In Table 3 are enumerated the different dataset used.

| $\#\theta$ | $\#\dot{\theta}$ | $\#u$ | Sample size |
|---|---|---|---|
| 10 | 10 | 2 | 200 |
| 15 | 15 | 2 | 450 |
| 20 | 20 | 2 | 800 |
| 25 | 25 | 2 | 1250 |
| 30 | 30 | 2 | 1800 |
| 40 | 40 | 2 | 3200 |

Table 3: **Pendulum uniform grid dataset configurations** This table shows the level of discretization for each dimension of the state space ($\#\theta$ and $\#\dot{\theta}$) and the action space ($\#u$). Each line corresponds to a uniformly sampled dataset, where $\theta \in [-\pi, \pi]$, $\dot{\theta} \in [-8, 8]$ and $u \in [-2, 2]$. The entries under the states' dimensions and action dimension correspond to how many linearly spaced states or actions are to be queried from the corresponding intervals. The Cartesian product of states and actions dimensions is taken in order to generate the state-action pairs to query the environment transitions. The rightmost column indicates the total number of corresponding samples.

**Algorithm details.** The configuration used for NOPG-D and NOPG-S are listed in Table 4.

| **NOPG** | |
|---|---|
| discount factor $\gamma$ | 0.97 |
| state $\vec{h}_{\text{factor}}$ | 1.0 1.0 1.0 |
| action $\vec{h}_{\text{factor}}$ | 50.0 |
| policy | neural network parameterized by $\vec{\theta}$ |
| | 1 hidden layer, 50 units, ReLU activations |
| policy output | $2\tanh(f_{\vec{\theta}}(\mathbf{s}))$ (NOGP-D) |
| | $\mu = 2\tanh(f_{\vec{\theta}}(\mathbf{s}))$, $\sigma = \text{sigmoid}(g_{\vec{\theta}}(\mathbf{s}))$ (NOGP-S) |
| learning rate | $10^{-2}$ with ADAM optimizer |
| $N_{\pi}^{\text{MC}}$ (NOPG-S) | 15 |
| $N_{\phi}^{\text{MC}}$ | 1 |
| $N_{\mu_0}^{\text{MC}}$ | (non applicable) fixed initial state |
| policy updates | $1.5 \cdot 10^3$ |

Table 4: **NOPG configurations for the Pendulum uniform grid experiment**

### B.2.2 Pendulum with Random Agent

The following table shows the hyper-parameters used for generating the second plot starting from the left in Figure 2

| **NOPG** | |
|---|---|
| dataset sizes | $10^2$, $5 \cdot 10^2$, $10^3$, $1.5 \cdot 10^3$, $2 \cdot 10^3$, $3 \cdot 10^3$, $5 \cdot 10^3$, $7 \cdot 10^3$, $9 \cdot 10^3$, $10^4$ |
| discount factor $\gamma$ | 0.97 |
| state $\vec{h}_{\text{factor}}$ | 1.0 1.0 1.0 |
| action $\vec{h}_{\text{factor}}$ | 25.0 |
| policy | neural network parameterized by $\vec{\theta}$<br>1 hidden layer, 50 units, ReLU activations |
| policy output | $2\tanh(f_{\vec{\theta}}(\mathbf{s}))$ (NOGP-D)<br>$\mu = 2\tanh(f_{\vec{\theta}}(\mathbf{s}))$, $\sigma = \text{sigmoid}(g_{\vec{\theta}}(\mathbf{s}))$ (NOGP-S) |
| learning rate | $10^{-2}$ with ADAM optimizer |
| $N^{\text{MC}}_{\pi_0}$ (NOPG-S) | 10 |
| $N^{\text{MC}}_{\phi}$ | 1 |
| $N^{\text{MC}}_{\mu_0}$ | (non applicable) fixed initial state |
| policy updates | $2 \cdot 10^3$ |

| **DDPG** | |
|---|---|
| discount factor $\gamma$ | 0.97 |
| rollout steps | 1000 |
| actor | neural network parameterized by $\vec{\theta}_{\text{actor}}$<br>1 hidden layer, 50 units, ReLU activations |
| actor output | $2\tanh(f_{\vec{\theta}_{\text{actor}}}(\mathbf{s}))$ |
| actor learning rate | $10^{-3}$ with ADAM optimizer |
| critic | neural network parameterized by $\vec{\theta}_{\text{critic}}$<br>1 hidden layer, 50 units, ReLU activations |
| critic output | $f_{\vec{\theta}_{\text{critic}}}(\mathbf{s}, \mathbf{a})$ |
| critic learning rate | $10^{-2}$ with ADAM optimizer |
| soft update | $\tau = 10^{-3}$ |
| policy updates | $3 \cdot 10^5$ |

| **DDPG Offline** | |
|---|---|
| dataset sizes | $10^2$, $5 \cdot 10^2$, $10^3$, $2 \cdot 10^3$, $5 \cdot 10^3$, $7.5 \cdot 10^3$, $10^4$, $1.2 \cdot 10^4$, $1.5 \cdot 10^4$, $2 \cdot 10^4$, $2.5 \cdot 10^4$ |
| discount factor $\gamma$ | 0.97 |
| actor | neural network parameterized by $\vec{\theta}_{\text{actor}}$<br>1 hidden layer, 50 units, ReLU activations |
| actor output | $2\tanh(f_{\vec{\theta}_{\text{actor}}}(\mathbf{s}))$ |
| actor learning rate | $10^{-2}$ with ADAM optimizer |
| critic | neural network parameterized by $\vec{\theta}_{\text{critic}}$<br>1 hidden layer, 50 units, ReLU activations |
| critic output | $f_{\vec{\theta}_{\text{critic}}}(\mathbf{s}, \mathbf{a})$ |
| critic learning rate | $10^{-2}$ with ADAM optimizer |
| soft update | $\tau = 10^{-3}$ |
| policy updates | $2 \cdot 10^3$ |

| **PWIS** | |
|---|---|

| | |
|---|---|
| dataset sizes | $10^2$, $5 \cdot 10^2$, $10^3$, $2 \cdot 10^3$, $5 \cdot 10^3$, $7.5 \cdot 10^3$, $10^4$, $1.2 \cdot 10^4$, $1.5 \cdot 10^4$, $2 \cdot 10^4$, $2.5 \cdot 10^4$ |
| discount factor $\gamma$ | 0.97 |
| policy | neural network parameterized by $\vec{\theta}$<br>1 hidden layer, 50 units, ReLU activations |
| policy output | $\mu = 2\tanh(f_{\vec{\theta}}(\mathbf{s}))$, $\sigma = \text{sigmoid}(g_{\vec{\theta}}(\mathbf{s}))$ |
| learning rate | $10^{-2}$ with ADAM optimizer |
| policy updates | $2 \cdot 10^3$ |

Table 5: **Algorithms configurations for the Pendulum random data experiment**

### B.2.3 Cart-pole with Random Agent

The following tables show the hyper-parameters used to generate the third plot in Figure 2.

| **NOPG** | |
|---|---|
| dataset sizes | $10^2$, $2.5 \cdot 10^2$, $5 \cdot 10^2$, $10^3$, $1.5 \cdot 10^3$, $2.5 \cdot 10^3$, $3 \cdot 10^3$, $5 \cdot 10^3$, $6 \cdot 10^3$, $8 \cdot 10^3$, $10^4$ |
| discount factor $\gamma$ | 0.99 |
| state $\vec{h}_{\text{factor}}$ | 1.0 1.0 1.0 |
| action $\vec{h}_{\text{factor}}$ | 20.0 |
| policy | neural network parameterized by $\vec{\theta}$<br>1 hidden layer, 50 units, ReLU activations |
| policy output | $5\tanh(f_{\vec{\theta}}(\mathbf{s}))$ (NOGP-D)<br>$\mu = 5\tanh(f_{\vec{\theta}}(\mathbf{s}))$, $\sigma = \text{sigmoid}(g_{\vec{\theta}}(\mathbf{s}))$ (NOGP-S) |
| learning rate | $\cdot 10^{-2}$ with ADAM optimizer |
| $N_{\pi}^{\text{MC}}$ (NOPG-S) | 10 |
| $N_{\phi}^{\text{MC}}$ | 1 |
| $N_{\mu_0}^{\text{MC}}$ | 15 |
| policy updates | $2 \cdot 10^3$ |

| **DDPG** | |
|---|---|
| discount factor $\gamma$ | 0.99 |
| rollout steps | 1000 |
| actor | neural network parameterized by $\vec{\theta}_{\text{actor}}$<br>1 hidden layer, 50 units, ReLU activations |
| actor output | $5\tanh(f_{\vec{\theta}_{\text{actor}}}(\mathbf{s}))$ |
| actor learning rate | $10^{-3}$ with ADAM optimizer |
| critic | neural network parameterized by $\vec{\theta}_{\text{critic}}$<br>1 hidden layer, 50 units, ReLU activations |
| critic output | $f_{\vec{\theta}_{\text{critic}}}(\mathbf{s}, \mathbf{a})$ |
| critic learning rate | $10^{-2}$ with ADAM optimizer |
| soft update | $\tau = 10^{-3}$ |
| policy updates | $2 \cdot 10^5$ |

| **DDPG Offline** | |
|---|---|
| dataset sizes | $10^2$, $5 \cdot 10^2$, $10^3$, $2 \cdot 10^3$, $3.5 \cdot 10^3$, $5 \cdot 10^3$, $8 \cdot 10^3$, $10^4$, $1.5 \cdot 10^4$, $2 \cdot 10^4$, $2.5 \cdot 10^4$ |
| discount factor $\gamma$ | 0.99 |

| | |
|---|---|
| actor | neural network parameterized by $\vec{\theta}_{\text{actor}}$<br>1 hidden layer, 50 units, ReLU activations |
| actor output | $5 \tanh(f_{\vec{\theta}_{\text{actor}}}(\mathbf{s}))$ |
| actor learning rate | $10^{-2}$ with ADAM optimizer |
| critic | neural network parameterized by $\vec{\theta}_{\text{critic}}$<br>1 hidden layer, 50 units, ReLU activations |
| critic output | $f_{\vec{\theta}_{\text{critic}}}(\mathbf{s}, \mathbf{a})$ |
| critic learning rate | $10^{-2}$ with ADAM optimizer |
| soft update | $\tau = 10^{-3}$ |
| policy updates | $2 \cdot 10^3$ |

| **PWIS** | |
|---|---|
| dataset sizes | $10^2$, $5 \cdot 10^2$, $10^3$, $2 \cdot 10^3$, $3.5 \cdot 10^3$, $5 \cdot 10^3$,<br>$8 \cdot 10^3$, $10^4$, $1.5 \cdot 10^4$, $2 \cdot 10^4$, $2.5 \cdot 10^4$ |
| discount factor $\gamma$ | 0.99 |
| policy | neural network parameterized by $\vec{\theta}$<br>1 hidden layer, 50 units, ReLU activations |
| policy output | $\mu = 5 \tanh(f_{\vec{\theta}}(\mathbf{s}))$, $\sigma = \text{sigmoid}(g_{\vec{\theta}}(\mathbf{s}))$ |
| learning rate | $10^{-3}$ with ADAM optimizer |
| policy updates | $2 \cdot 10^3$ |

Table 6: **Algorithms configurations for the CartPole random data experiment**.

### B.2.4 Mountain Car with Human Demonstrator

Here the detail of the experiment shown in Figure 4. The dataset in this experiment (10 trajectories) has been generated by a human demonstrator. The dataset used is available in the source code provided.

| **NOPG** | |
|---|---|
| discount factor $\gamma$ | 0.99 |
| state $\vec{h}_{\text{factor}}$ | 1.0 1.0 |
| action $\vec{h}_{\text{factor}}$ | 50.0 |
| policy | neural network parameterized by $\vec{\theta}$<br>1 hidden layer, 50 units, ReLU activations |
| policy output | $1 \tanh(f_{\vec{\theta}}(\mathbf{s}))$ (NOGP-D)<br>$\mu = 1 \tanh(f_{\vec{\theta}}(\mathbf{s}))$, $\sigma = \text{sigmoid}(g_{\vec{\theta}}(\mathbf{s}))$ (NOGP-S) |
| learning rate | $10^{-2}$ with ADAM optimizer |
| $N_{\pi}^{\text{MC}}$ (NOPG-S) | 15 |
| $N_{\phi}^{\text{MC}}$ | 1 |
| $N_{\mu_0}^{\text{MC}}$ | 15 |
| policy updates | $1.5 \cdot 10^3$ |

Table 7: **NOPG configurations for the MountainCar experiment.**