On the Relation between Policy Improvement and Off-Policy Minimum-Variance Policy Evaluation (Supplementary Material)

Alberto Maria Metelli[1] **Samuele Meta**[1] **Marcello Restelli**[1]

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy

1 PROOFS AND DERIVATIONS

In this appendix, we report the proofs and derivations, we have omitted in the main paper.

1.1 PROOFS OF SECTION 4

Proof of Proposition 4.1

Proof. Let us consider the following derivation:

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{I}_{h\circ f}[P]}[f(x)] - \mathbb{E}_{x\sim P}[f(x)] &= \int_{\mathcal{X}} \frac{p(x)h(f(x))}{\mathbb{E}_{x\sim P}[h(f(x))]} f(x)\mathrm{d}x - \mathbb{E}_{x\sim P}[f(x)] \\ &= \frac{\mathbb{E}_{x\sim P}[h(f(x))f(x)] - \mathbb{E}_{x\sim P}[f(x)]\mathbb{E}_{x\sim P}[h(f(x))]}{\mathbb{E}_{x\sim P}[h(f(x))]} \\ &= \frac{\mathrm{Cov}_{x\sim P}[h(f(x)), f(x)]}{\mathbb{E}_{x\sim P}[h(f(x))]},\end{aligned}$$

where we have exploited the definition of $\mathcal{I}_{h\circ f}$ and the definition of covariance. The result is obtained by recalling that h is increasing and the covariance between two increasing functions of the same random variable (i.e., h and the identity function) is non-negative [Cuadras, 2002]. □

Proof of Theorem 4.2

Proof. We are going to actually prove a more general statement in which we consider a non-negative monotonic increasing function h that is composed to function f, i.e., $h \circ f$. The theorem statement can be obtained by setting h to be the identity function.

We start with (i). First of all, we observe that since h is monotonically strictly-increasing it holds that $\mathrm{Var}_{x\sim P}[f(x)] = 0$ if and only if $\mathrm{Var}_{x\sim P}[h(f(x))] = 0$. P is a fixed point of $\mathcal{I}_{h\circ f}$, i.e., $P = \mathcal{I}_{h\circ f}[P]$ a.s. if and only if for all $x \in \mathcal{X}$ it holds a.s.:

$$p(x) = \frac{p(x)h(f(x))}{\mathbb{E}_{x\sim P}[h(f(x))]},$$

that occurs if and only if either $p(x) = 0$ ($x \notin \mathrm{supp}(P)$) or $h(f(x)) = \mathbb{E}_{x\sim P}[h(f(x))]$. (⇒) Whenever $p(x)$ is not zero, function $h(f(x))$ is a constant in $\mathrm{supp}(P)$ and, consequently, its variance under P is zero. (⇐) Suppose that $\mathrm{Var}_{x\sim P}[h(f(x))] = 0$, then $h(f(x)) = \mathbb{E}_{x\sim P}[h(f(x))]$ almost surely and, consequently $\frac{p(x)h(f(x))}{\mathbb{E}_{x\sim P}[h(f(x))]} = p(x)$ almost surely. Let us now consider (ii). First of all, we can easily observe that for every $k \in \mathbb{N}$:

$$\left(\mathcal{I}_{h\circ f}\right)^k [P](x) = \frac{p(x)f(x)^k}{\mathbb{E}_{x\sim P}[f(x)^k]}.$$

Let $f^* = \max_{x \in \mathrm{supp}(P)}\{f(x)\}$, consider the function $g_k(x) = p(x)\left(\frac{f(x)}{f^*}\right)^k$ and the limit:

$$\lim_{k\to\infty} g_k(x) = \lim_{k\to\infty} p(x)\left(\frac{f(x)}{f^*}\right)^k = \begin{cases} p(x) & \text{if } x \in \mathcal{X}^* \\ 0 & \text{otherwise} \end{cases}.$$

Thus, we have:

$$\begin{aligned} Q_\infty = \lim_{k\to\infty} \left(\mathcal{I}_{h\circ f}\right)^k[P](x) &= \lim_{k\to\infty} \frac{p(x)f(x)^k}{\int_{\mathcal{X}} p(x)f(x)^k \mathrm{d}x} \\ &= \lim_{k\to\infty} \frac{g_k(x)}{\int_{\mathcal{X}} g_k(x)\mathrm{d}x} = \begin{cases} \frac{p(x)}{\int_{\mathcal{X}^*} p(x)\mathrm{d}x} & \text{if } x \in \mathcal{X}^* \\ 0 & \text{otherwise} \end{cases}. \end{aligned}$$

Thus, the support of Q_∞ is given by $\mathcal{X}^*$. Consequently, the expectation of f under Q_∞ is given by:

$$\mathbb{E}_{x\sim Q_\infty}[f(x)] = \int_{\mathcal{X}} q_\infty(x)f(x)\mathrm{d}x = f^*.$$

□

Proof of Theorem 4.3

Proof. We are going to actually prove a more general statement in which we consider a non-negative monotonic increasing function h that is composed to function f, i.e., $h \circ f$. The theorem statement can be obtained by setting h to be the identity function.

Let us consider the following derivation:

$$\begin{aligned} J := \int_{\mathcal{X}} \left(\left(I_{h\circ f}[P]\right)(x)\right)^\alpha p(x)^{1-\alpha}\mathrm{d}x &= \int_{\mathcal{X}} \left(\frac{p(x)h(f(x))}{\mathbb{E}_{x\sim P}[h(f(x))]}\right)^\alpha p(x)^{1-\alpha}\mathrm{d}x \\ &= \frac{\mathbb{E}_{x\sim P}[h(f(x))^\alpha]}{\mathbb{E}_{x\sim P}[h(f(x))]^\alpha}. \end{aligned}$$

By observing that $D_\alpha\left(I_{h\circ f}[P]\|P\right) = \frac{1}{\alpha-1}\log J$, we obtain the result. For $\alpha = 1$, we provide an independent derivation:

$$\begin{aligned} D_{\mathrm{KL}}(I_{h\circ f}[P]\|P) &= \int_{\mathcal{X}} \frac{p(x)h(f(x))}{\mathbb{E}_{x\sim P}[h(f(x))]} \log \frac{p(x)h(f(x))}{\mathbb{E}_{x\sim P}[h(f(x))]p(x)}\mathrm{d}x \\ &= \frac{\mathbb{E}_{x\sim P}[h(f(x))\log h(f(x))] - \mathbb{E}_{x\sim P}[h(f(x))]\mathbb{E}_{x\sim P}[\log h(f(x))]}{\mathbb{E}_{x\sim P}[h(f(x))]} \\ &= \frac{\mathrm{Cov}_{x\sim P}[h(f(x)), \log h(f(x))]}{\mathbb{E}_{x\sim P}[h(f(x))]}, \end{aligned}$$

where we exploited the definition of covariance in the last line. □

1.2 PROOFS OF SECTION 5

Proof of Proposition 5.1

Proof. We are going to actually prove a more general statement in which we consider a non-negative monotonic increasing function h that is composed to function f, i.e., $h \circ f$. The theorem statement can be obtained by setting h to be the identity function.

First of all, we observe that since $\mathbb{E}_{x\sim Q}\left[\frac{p(x)}{q(x)}h(f(x))\right] = \mathbb{E}_{x\sim P}[h(f(x))]$, for $\alpha \in [2, +\infty)$, the absolute central α-moment is smaller or equal than the (non-central) α-moment. Thus, for $\alpha \in [2, +\infty)$, we have:

$$\mathbb{E}_{x\sim Q}\left[\left|\frac{p(x)}{q(x)}h(f(x)) - \mathbb{E}_{x\sim P}[h(f(x))]\right|^\alpha\right] \leqslant \mathbb{E}_{x\sim Q}\left[\left(\frac{p(x)}{q(x)}h(f(x))\right)^\alpha\right]$$

$$
\begin{aligned}
&= \int_{\mathcal{X}} \left(\frac{p(x)h(f(x))}{\mathbb{E}_{x\sim P}[h(f(x))]} \right)^{\alpha} q(x)^{1-\alpha} \mathrm{d}x \mathbb{E}_{x\sim P}[h(f(x))]^{\alpha} \\
&= \int_{\mathcal{X}} \left((\mathcal{I}_{h\circ f}[P])(x) \right)^{\alpha} q(x)^{1-\alpha} \mathrm{d}x \mathbb{E}_{x\sim P}[h(f(x))]^{\alpha} \\
&= \exp\left\{ (\alpha-1) \frac{1}{\alpha-1} \log \int_{\mathcal{X}} \left((\mathcal{I}_{h\circ f}[P])(x) \right)^{\alpha} q(x)^{1-\alpha} \mathrm{d}x \right\} \mathbb{E}_{x\sim P}[h(f(x))]^{\alpha},
\end{aligned}
$$

where the first inequality follows from Lemma 1.1 with $y = \left(\frac{p(x)}{q(x)} h(f(x)) \right) / \mathbb{E}_{x\sim P}[h(f(x))]$. By applying the definition of Rényi divergences, we get the result. □

Proof of Theorem 5.2

Proof. Let us consider the following derivation:

$$
\begin{aligned}
\mathbb{E}_{x\sim Q}[h(f(x))^{\alpha}] &= \int_{\mathcal{X}} q(x) h(f(x))^{\alpha} \mathrm{d}x \\
&= \int_{\mathcal{X}} p(x) \frac{q(x)}{p(x)} h(f(x))^{\alpha} \mathrm{d}x \\
&= \int_{\mathcal{X}} p(x) h(f(x))^{\alpha} \mathrm{d}x + \int_{\mathcal{X}} p(x) \left(\frac{q(x)}{p(x)} - 1 \right) h(f(x))^{\alpha} \mathrm{d}x \\
&\geqslant \int_{\mathcal{X}} p(x) h(f(x))^{\alpha} \mathrm{d}x + \frac{1}{\alpha-1} \int_{\mathcal{X}} p(x) \left(1 - \left(\frac{p(x)}{q(x)} \right)^{\alpha-1} \right) h(f(x))^{\alpha} \mathrm{d}x \qquad (1) \\
&= \mathbb{E}_{x\sim P}[h(f(x))^{\alpha}] + \frac{1}{\alpha-1} \int_{\mathcal{X}} p(x) h(f(x))^{\alpha} \mathrm{d}x \\
&\quad - \frac{1}{\alpha-1} \int_{\mathcal{X}} p(x) \left(\frac{p(x)}{q(x)} \right)^{\alpha-1} h(f(x))^{\alpha} \mathrm{d}x \\
&= \mathbb{E}_{x\sim P}[h(f(x))^{\alpha}] + \mathbb{E}_{x\sim P}[h(f(x))]^{\alpha} \frac{1}{\alpha-1} \int_{\mathcal{X}} \left(\frac{p(x)h(f(x))}{\mathbb{E}_{x\sim P}[h(f(x))]} \right)^{\alpha} p(x)^{1-\alpha} \mathrm{d}x \\
&\quad - \mathbb{E}_{x\sim P}[h(f(x))]^{\alpha} \frac{1}{\alpha-1} \int_{\mathcal{X}} \left(\frac{p(x)h(f(x))}{\mathbb{E}_{x\sim P}[h(f(x))]} \right)^{\alpha} q(x)^{1-\alpha} \mathrm{d}x \\
&= \mathbb{E}_{x\sim P}[h(f(x))^{\alpha}] \\
&\quad + \mathbb{E}_{x\sim P}[h(f(x))]^{\alpha} \frac{1}{\alpha-1} \exp\left\{ (\alpha-1) \frac{1}{\alpha-1} \log \int_{\mathcal{X}} \left(\frac{p(x)h(f(x))}{\mathbb{E}_{x\sim P}[h(f(x))]} \right)^{\alpha} p(x)^{1-\alpha} \mathrm{d}x \right\} \\
&\quad - \mathbb{E}_{x\sim P}[h(f(x))]^{\alpha} \frac{1}{\alpha-1} \exp\left\{ (\alpha-1) \frac{1}{\alpha-1} \log \int_{\mathcal{X}} \left(\frac{p(x)h(f(x))}{\mathbb{E}_{x\sim P}[h(f(x))]} \right)^{\alpha} q(x)^{1-\alpha} \mathrm{d}x \right\} \\
&= \mathbb{E}_{x\sim P}[h(f(x))^{\alpha}] + \frac{\mathbb{E}_{x\sim P}[h(f(x))]^{\alpha}}{\alpha-1} \left(e^{(\alpha-1)D_{\alpha}(\mathcal{I}_{h\circ f}\|P)} - e^{(\alpha-1)D_{\alpha}(\mathcal{I}_{h\circ f}\|Q)} \right),
\end{aligned}
$$

where line (1) derived from Lemma 1.2. The second inequality was provided in Proposition 6 of [Ghosh et al., 2020]. □

Proof of Theorem 5.3

Proof. We are going to actually prove a more general statement in which we consider a non-negative monotonic increasing function h that is composed to function f, i.e., $h \circ f$. The theorem statement can be obtained by setting h to be the identity function.

Let us consider the sequence of distributions $(Q_k)_{k\in\mathbb{N}}$, generated by the iterate in Equation (5), where possible ties are broken with an arbitrary (possibly with a tie-breaking rule T_k different for every k). From Theorem 5.2, we have for every $k \in \mathbb{N}$:

$$
\mathbb{E}_{x\sim Q_{k+1}}[h(f(x))^{\alpha}] - \mathbb{E}_{x\sim Q_k}[h(f(x))^{\alpha}]
$$

$$\geqslant \frac{\mathbb{E}_{x\sim Q_k}[h(f(x))]^\alpha}{\alpha-1}\left(e^{(\alpha-1)D_\alpha(\mathcal{I}_{h\circ f}[Q_k]\|Q_k)} - e^{(\alpha-1)D_\alpha(\mathcal{I}_{h\circ f}[Q_k]\|Q_{k+1})}\right) \geqslant 0,$$

where we simply exploited that $Q_k \in \arg\min_{Q\in\mathcal{Q}}\{D_\alpha(\mathcal{I}_{h\circ f}[Q_k]\|Q)\}$. Thus, $\mathbb{E}_{x\sim Q_k}\left[h(f(x))^\alpha\right]$ is a non-decreasing function of k. Since $h\circ f$ is bounded, it must be that $\lim_{k\to\infty}\mathbb{E}_{x\sim Q_k}\left[h(f(x))^\alpha\right] = \mu_\infty < \infty$, that proves convergence.[1]

Furthermore, being convergent, for $k\to\infty$ it must be that $\mathbb{E}_{x\sim Q_k}\left[h(f(x))^\alpha\right] = \mathbb{E}_{x\sim Q_{k+1}}\left[h(f(x))^\alpha\right]$ and consequently $D_\alpha(\mathcal{I}_{h\circ f}[Q_k]\|Q_k) = D_\alpha(\mathcal{I}_{h\circ f}[Q_k]\|Q_{k+1})$. Therefore, even if the tie-braking rule prescribes to select $Q_{k+1}\neq Q_k$ we could select Q_k instead, since it lead to the same divergence value. Consequently, being Q_k a solution, we can assert that it is a stationary point of the function $D_\alpha(\mathcal{I}_{h\circ f}[Q_k]\|\cdot)$ (as well as Q_{k+1}):

$$\begin{aligned}
0 &= \nabla_{q(\cdot)}D_\alpha(\mathcal{I}_{h\circ f}[Q_k]\|Q)|_{Q=Q_k} \\
&= \frac{1}{(\alpha-1)e^{(\alpha-1)D_\alpha(\mathcal{I}_{h\circ f}[Q_k]\|Q)}\mathbb{E}_{x\sim Q_k}[h(f(x))]}\nabla_{q(\cdot)}\int_{\mathcal{X}} h(f(x))^\alpha q_k(x)^\alpha q(x)^{1-\alpha}\mathrm{d}x|_{Q=Q_k} \\
&= -\frac{1}{e^{(\alpha-1)D_\alpha(\mathcal{I}_{h\circ f}[Q_k]\|Q)}\mathbb{E}_{x\sim Q_k}[h(f(x))]}\int_{\mathcal{X}} h(f(x))^\alpha q_k(x)^\alpha q(x)^{-\alpha}\mathrm{d}x|_{Q=Q_k} \\
&= -\frac{1}{e^{(\alpha-1)D_\alpha(\mathcal{I}_{h\circ f}[Q_k]\|Q)}\mathbb{E}_{x\sim Q_k}[h(f(x))]}\int_{\mathcal{X}} h(f(x))^\alpha\mathrm{d}x.
\end{aligned}$$

We observe that the latter expression is zero if and only if the gradient of $\mathbb{E}_{x\sim Q}[h(f(x))^\alpha]$ w.r.t. Q is zero. Indeed:

$$\nabla_{q(\cdot)}\mathbb{E}_{x\sim Q}[h(f(x))^\alpha] = \int_{\mathcal{X}} h(f(x))^\alpha\mathrm{d}x.$$

Thus, the process converges to a stationary point of $\mathbb{E}_{x\sim Q_k}[h(f(x))^\alpha]$. □

Proof of Theorem 5.4

Proof. The proof is a simple application of Lemma 1.3, by taking $Q\leftarrow P$, $Q^*\leftarrow Q^\dagger$, and $P\leftarrow\mathcal{I}_f[P]$. □

1.3 PROOFS OF SECTION 6

Proof of Theorem 6.1

Proof. We start observing that each addendum of $\widehat{d}_\alpha\left(\mathcal{I}_f[Q_{\boldsymbol{\xi}_i}]\|Q_{\boldsymbol{\xi}};\Phi_{i,j}\right)$ is non negative. Since all terms are i.i.d., we can apply unilateral Bernstein's inequality that allows achieving an exponential concentration. Thus, for every $\delta\in[0,1]$, with probability at least $1-\delta$ it holds that:

$$\begin{aligned}
\mathbb{E}_{x\sim\boldsymbol{\xi}}\left[\left(\frac{q_{\boldsymbol{\xi}_i}(x)}{q_{\boldsymbol{\xi}}(x)}f(x)\right)^\alpha\right] &\leqslant \widehat{d}_\alpha\left(\mathcal{I}_f[Q_{\boldsymbol{\xi}_i}]\|Q_{\boldsymbol{\xi}};\Phi_{i,j}\right) \\
&\quad + \sqrt{2\mathrm{Var}_{x_i\sim\Phi_{i,j}}\left[\widehat{d}_\alpha\left(\mathcal{I}_f[Q_{\boldsymbol{\xi}_i}]\|Q_{\boldsymbol{\xi}};\Phi_{i,j}\right)\right]\log\frac{1}{\delta}}.
\end{aligned}$$

Thus, it remains to provide a bound on the variance term. We exploit the fact that $h(f(x))\leqslant\overline{m}$ and that each addendum represents an i.i.d. random variable:

$$\begin{aligned}
&\mathrm{Var}_{x_i\sim\Phi_{i,j}}\left[\widehat{d}_\alpha\left(\mathcal{I}_f[Q_{\boldsymbol{\xi}_i}]\|Q_{\boldsymbol{\xi}};\Phi_{i,j}\right)\right] \\
&\qquad\leqslant \frac{1}{(nj)^2}\sum_{k\in[j]}\sum_{l\in[n]}\mathbb{E}_{x_{k,l}\sim\Phi_{i,j}}\left[\left(\frac{q_{\boldsymbol{\xi}_i}(x_{k,l})^\alpha}{\Phi_{i,j}(x_{k,l})q_{\boldsymbol{\xi}}(x_{k,l})^{\alpha-1}}f(x)^\alpha\right)^2\right] \\
&\qquad\leqslant \frac{\overline{m}^{2\alpha}}{(nj)^2}\sum_{k\in[j]}\sum_{l\in[n]}\mathbb{E}_{x_{k,l}\sim\Phi_{i,j}}\left[\left(\frac{q_{\boldsymbol{\xi}_i}(x_{k,l})^\alpha}{\Phi_{i,j}(x_{k,l})q_{\boldsymbol{\xi}}(x_{k,l})^{\alpha-1}}\right)^2\right]
\end{aligned}$$

[1] Notice that the improvement holds also for $\alpha<1$. Indeed, while it is true that $\frac{\mathbb{E}_{x\sim Q_k}[h(f(x))]^\alpha}{\alpha-1}<0$, but in such a case function $e^{(\alpha-1)(\cdot)}$ is decreasing in its argument.

$$= \frac{\overline{m}^{2\alpha}}{nj} \mathbb{E}_{x \sim \Phi_{i,j}} \left[\left(\frac{q_{\boldsymbol{\xi}_i}(x)^\alpha}{\Phi_{i,j}(x) q_{\boldsymbol{\xi}}(x)^{\alpha-1}} \right)^2 \right].$$

□

1.4 TECHNICAL LEMMAS

Lemma 1.1. *Let $\alpha \in [2, +\infty)$ and let y be a non-negative random variable with expectation 1. Then, it holds that $\mathbb{E}[|y-1|^\alpha]^{1/\alpha} \leqslant \mathbb{E}[y^\alpha]^{1/\alpha}$.*

Proof. When $y \geqslant 0$ and $\alpha \in [2, +\infty)$, it holds that $y^\alpha - |y-1|^\alpha \geqslant y - 1$. Consequently, we have:

$$\mathbb{E}[|y-1|^\alpha] \leqslant \mathbb{E}[y^\alpha - y + 1] \leqslant \mathbb{E}[y^\alpha].$$

□

Lemma 1.2. *For every $x \geqslant 0$ and $\alpha \in (0,1) \cup (1, \infty)$, it holds that:*

$$x - 1 \geqslant \frac{1}{\alpha - 1} \left(1 - \frac{1}{x^{\alpha-1}} \right).$$

Furthermore, for $\alpha = 1$, it holds that:

$$x - 1 \geqslant \log x.$$

Proof. Consider the auxiliary function $g_\alpha(x) = x - 1 - \frac{1}{\alpha-1}\left(1 - \frac{1}{x^{\alpha-1}}\right)$. We are going to prove that the minimum of $g_\alpha(x)$ is zero. Suppose $\alpha > 1$, then $g_\alpha(0) = \infty$ and $g_a(\infty) = \infty$. Thus, the minimum must lie in between and since function g_α is differentiable, we have:

$$\frac{\partial}{\partial x} g_\alpha(x) = 1 - x^{-\alpha} = 0 \qquad \Longrightarrow \qquad x = 1.$$

Thus, we have $g_\alpha(1) = 0$. Suppose now that $\alpha < 1$, we have $g_\alpha(0) = \frac{\alpha}{1-\alpha} > 0$ and $g_\alpha(\infty) = \infty$. Thus, again, the minimum must lie in between and with the same calculations as before, we conclude $g_\alpha(1) = 0$. The case $\alpha = 1$ is trivial. □

Lemma 1.3. *Let $P \in \mathscr{P}(\mathcal{X})$ and let $\alpha \in [0,1)$. Let $\mathcal{Q} \subseteq \mathscr{P}(\mathcal{X})$ be an $(1-\alpha)$-convex [van Erven and Harremoës, 2014, Definition 4] subset of distributions. Let $Q^* \in \mathcal{Q}$ be the α-moment projection:*

$$Q^* = \operatorname*{arg\,min}_{Q \in \mathcal{Q}} \left\{ D_\alpha(P \| Q) \right\}.$$

If Q^* exists, then for every $Q \in \mathcal{Q}$ if holds that:

$$D_\alpha(P\|Q) \geqslant D_\alpha(P\|Q^*) + D_\alpha(Q^*\|Q).$$

Proof. The proof of the result is inspired to [van Erven and Harremoës, 2014, Theorem 14]. Let $\lambda \in [0,1]$ and let us define Q_λ as the $(1-\alpha, (1-\lambda, \lambda))$-mixture of Q^* and Q:

$$q_\lambda(x) = Z_\lambda^{-1} \left((1-\lambda) q^*(x)^{1-\alpha} + \lambda q(x)^{1-\alpha} \right)^{\frac{1}{1-\alpha}},$$
$$Z_\lambda = \int_{\mathcal{X}} \left((1-\lambda) q^*(x)^{1-\alpha} + \lambda q(x)^{1-\alpha} \right)^{\frac{1}{1-\alpha}} \mathrm{d}x.$$

Let us first observe that for $\lambda = 0$, we have $Q_0 = Q^*$ and $Z_0 = \int_{\mathcal{X}} q^*(x) \mathrm{d}x = 1$. Since $\mathcal{Q}$ is $(1-\alpha)$-convex and Q^* is the minimizer over $\mathcal{Q}$, it holds that $\frac{\partial}{\partial \lambda} D_\alpha(P\|Q_\lambda)|_{\lambda=0} \geqslant 0$. First of all, we compute:

$$\int_{\mathcal{X}} p(x)^\alpha q_\lambda(x)^{1-\alpha} \mathrm{d}x = Z_\lambda^{\alpha-1} \int_{\mathcal{X}} \left[(1-\lambda) p(x)^\alpha q^*(x)^{1-\alpha} + \lambda p(x)^\alpha q(x)^{1-\alpha} \right] \mathrm{d}x$$

$$\frac{\partial}{\partial \lambda} Z_\lambda = \frac{1}{1-\alpha} \int_{\mathcal{X}} \left((1-\lambda) q^*(x)^{1-\alpha} + \lambda q(x)^{1-\alpha} \right)^{\frac{\alpha}{1-\alpha}} \left(q(x)^{1-\alpha} - q^*(x)^{1-\alpha} \right) \mathrm{d}x.$$

The latter, for $\lambda = 0$, becomes: $\frac{\partial}{\partial \lambda} Z_\lambda \Big|_{\lambda=0} = \frac{1}{1-\alpha} \left[\int_{\mathcal{X}} q^*(x)^\alpha q(x)^{1-\alpha} - 1 \right]$. For calculation easiness, instead of directly operating on $D_\alpha(P \| Q_\lambda)$, we consider:

$$\begin{aligned} \frac{\partial}{\partial \lambda} \int_{\mathcal{X}} p(x)^\alpha q_\lambda(x)^{1-\alpha} \mathrm{d}x = Z_\lambda^{\alpha-1} & \int_{\mathcal{X}} \left[-p(x)^\alpha q^*(x)^{1-\alpha} + p(x)^\alpha q(x)^{1-\alpha} \right] \mathrm{d}x, \\ & + (\alpha - 1) Z_\lambda^{\alpha-2} \frac{\partial}{\partial \lambda} Z_\lambda \int_{\mathcal{X}} \left[(1-\lambda) p(x)^\alpha q^*(x)^{1-\alpha} + \lambda p(x)^\alpha q(x)^{1-\alpha} \right] \mathrm{d}x. \end{aligned}$$

We now evaluate it at $\lambda = 0$:

$$\begin{aligned} \frac{\partial}{\partial \lambda} \int_{\mathcal{X}} p(x)^\alpha q_\lambda(x)^{1-\alpha} \mathrm{d}x \Big|_{\lambda=0} = & - \int_{\mathcal{X}} p(x)^\alpha q^*(x)^{1-\alpha} \mathrm{d}x + \int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha} \mathrm{d}x \\ & - \int_{\mathcal{X}} p(x)^\alpha q^*(x)^{1-\alpha} \mathrm{d}x \left[\int_{\mathcal{X}} q^*(x)^\alpha q(x)^{1-\alpha} \mathrm{d}x - 1 \right]. \end{aligned}$$

For $\alpha \geqslant 1$, we require $\frac{\partial}{\partial \lambda} \int_{\mathcal{X}} p(x)^\alpha q_\lambda(x)^{1-\alpha} \mathrm{d}x \Big|_{\lambda=0} \geqslant 0$, to obtain:

$$\int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha} \mathrm{d}x \geqslant \int_{\mathcal{X}} p(x)^\alpha q^*(x)^{1-\alpha} \mathrm{d}x \int_{\mathcal{X}} q^*(x)^\alpha q(x)^{1-\alpha} \mathrm{d}x.$$

By applying both sides the $\log$ function and dividing by $\frac{1}{\alpha-1} > 0$ we get the result. Symmetrically, for $\alpha < 1$, we require the converse $\frac{\partial}{\partial \lambda} \int_{\mathcal{X}} p(x)^\alpha q_\lambda(x)^{1-\alpha} \mathrm{d}x \Big|_{\lambda=0} \leqslant 0$. Recalling that $\frac{1}{\alpha-1} < 0$, we obtain the desired result. □

2 CLOSED FORM OF THE INTEGRAL FOR GAUSSIANS

In this appendix, we derive a closed form for the integral involved in the computation of the bound of Theorem 6.1 in the case that all involved distributions are Gaussians and for $\alpha = 2$. Let us introduce the notation:

$$\mu = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}), \qquad \phi = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\phi}}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}}), \qquad \nu = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\nu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}}). \tag{2}$$

We have to compute the following integral:

$$\int_{\mathcal{X}} \frac{\mu^4(\mathbf{x})}{\phi(\mathbf{x}) \nu(\mathbf{x})^2} \mathrm{d}\mathbf{x}.$$

Let us start elaborating on the integrand function, denoting for properly sized vector $\mathbf{x}$ and matrix $\mathbf{S}$, $\|\mathbf{m}\|_{\mathbf{S}} = \mathbf{x}^T \mathbf{S} \mathbf{x}$ and $|\mathbf{S}|$ the determinant of $\mathbf{S}$:

$$\begin{aligned} \frac{\mu^4(\mathbf{x})}{\phi(\mathbf{x}) \nu(\mathbf{x})^2} & = \frac{(2\pi)^{-2k} |\boldsymbol{\Sigma}_{\boldsymbol{\mu}}|^{-2} \exp\left(-2 \|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\mu}}\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}} \right)}{(2\pi)^{-k/2} |\boldsymbol{\Sigma}_{\boldsymbol{\phi}}|^{-1/2} \exp\left(-1/2 \|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\phi}}\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1}} \right) (2\pi)^{-k} |\boldsymbol{\Sigma}_{\boldsymbol{\nu}}|^{-1} \exp\left(- \|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\nu}}\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1}} \right)} \\ & = \frac{(2\pi)^{-k/2} |\boldsymbol{\Sigma}_{\boldsymbol{\mu}}|^{-2}}{|\boldsymbol{\Sigma}_{\boldsymbol{\phi}}|^{-1/2} |\boldsymbol{\Sigma}_{\boldsymbol{\nu}}|^{-1}} \exp\left(-2 \|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\mu}}\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}} + 1/2 \|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\phi}}\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1}} + \|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\nu}}\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1}} \right). \end{aligned}$$

Now, we have to deal with the argument of the exponential:

$$\begin{aligned} & -2 \|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\mu}}\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}} + 1/2 \|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\phi}}\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1}} + \|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\nu}}\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1}} \\ & \quad = -\frac{1}{2} \mathbf{x}^T \underbrace{\left(4 {\boldsymbol{\Sigma}_{\boldsymbol{\mu}}}^{-1} - {\boldsymbol{\Sigma}_{\boldsymbol{\phi}}}^{-1} - 2 {\boldsymbol{\Sigma}_{\boldsymbol{\nu}}}^{-1} \right)}_{\mathbf{M}} \mathbf{x} + \underbrace{\left(4 {\boldsymbol{\Sigma}_{\boldsymbol{\mu}}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\mu}} - {\boldsymbol{\Sigma}_{\boldsymbol{\phi}}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\phi}} - 2 {\boldsymbol{\Sigma}_{\boldsymbol{\nu}}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\nu}} \right)^T}_{\mathbf{b}^T} \mathbf{x} \\ & \qquad - \frac{1}{2} \underbrace{\left(4 {\boldsymbol{\mu}_{\boldsymbol{\mu}}}^T {\boldsymbol{\Sigma}_{\boldsymbol{\mu}}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\mu}} - {\boldsymbol{\mu}_{\boldsymbol{\phi}}}^T {\boldsymbol{\Sigma}_{\boldsymbol{\phi}}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\phi}} - 2 {\boldsymbol{\mu}_{\boldsymbol{\nu}}}^T {\boldsymbol{\Sigma}_{\boldsymbol{\nu}}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\nu}} \right)}_{\mathbf{c}}. \end{aligned}$$

We now proceed completing the square:

$$\mathbf{x}^T\mathbf{M}\mathbf{x} - 2\mathbf{b}^T\mathbf{x} = (\mathbf{x} - \mathbf{M}^{-1}\mathbf{b})^T\mathbf{M}(\mathbf{x} - \mathbf{M}^{-1}\mathbf{b}) - \mathbf{b}^T\mathbf{M}^{-1}\mathbf{b}.$$

Thus, we have:

$$-\frac{1}{2}\left(\mathbf{x}^T\mathbf{M}\mathbf{x} - 2\mathbf{b}^T\mathbf{x} + \mathbf{c}\right) = -\frac{1}{2}(\mathbf{x} - \mathbf{M}^{-1}\mathbf{b})^T\mathbf{M}(\mathbf{x} - \mathbf{M}^{-1}\mathbf{b}) + \frac{1}{2}\mathbf{b}^T\mathbf{M}^{-1}\mathbf{b} - \frac{1}{2}\mathbf{c}.$$

Moreover, we observe that the following expression is the density of a k-variate normal distribution with mean $M^{-1}b$ and covariance matrix M^{-1}:

$$(2\pi)^{-k/2}|\mathbf{M}^{-1}|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{M}^{-1}\mathbf{x})^T\mathbf{M}(\mathbf{x} - \mathbf{M}^{-1}\mathbf{b})\right).$$

Thus, its integral is 1. Therefore, coming to the initial expression:

$$\begin{aligned}\int_{\mathcal{X}} \frac{\mu^4(\mathbf{x})}{\phi(\mathbf{x})\nu(\mathbf{x})^2}\mathrm{d}\mathbf{x} &= \frac{(2\pi)^{-k/2}|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}|^{-2}}{|\boldsymbol{\Sigma}_{\boldsymbol{\phi}}|^{-1/2}|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}|^{-1}}\left((2\pi)^{-k/2}|\mathbf{M}^{-1}|^{-1/2}\right)^{-1}\exp\left(\frac{1}{2}\mathbf{b}^T\mathbf{M}^{-1}\mathbf{b} - \frac{1}{2}\mathbf{c}\right) \\ &= \frac{|\boldsymbol{\Sigma}_{\boldsymbol{\phi}}|^{1/2}|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}|}{|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}|^2|\mathbf{M}|^{1/2}}\exp\left(\frac{1}{2}\left(\mathbf{b}^T\mathbf{M}^{-1}\mathbf{b} - \mathbf{c}\right)\right)\end{aligned}$$

3 GRADIENT OF THE OBJECTIVE FUNCTION OF THEOREM 6.1

In this appendix, we report the expression of the gradient of the right hand side of Theorem 6.1:

$$\begin{aligned}&(1-\alpha)\frac{1}{nj}\sum_{k\in[j]}\sum_{l\in[n]}\frac{q_{\boldsymbol{\xi}_i}(x_{k,l})^\alpha}{\Phi_{i,j}(x_{k,l})q_{\boldsymbol{\xi}}(x_{k,l})^{\alpha-1}}\left(\nabla_{\boldsymbol{\xi}}\log q_{\boldsymbol{\xi}}(x_{k,l})\right)f(x_{k,l})^\alpha \\ &\qquad - 2(\alpha-1)\overline{m}^\alpha\sqrt{\frac{\log(1/\delta)}{2nj\int_{\mathcal{X}}\frac{q_{\boldsymbol{\xi}_i}(x)^{2\alpha}}{\Phi_{i,j}(x)q_{\boldsymbol{\xi}}(x)^{2(\alpha-1)}}\mathrm{d}x}}\int_{\mathcal{X}}\frac{q_{\boldsymbol{\xi}_i}(x)^{2\alpha}}{\Phi_{i,j}(x)q_{\boldsymbol{\xi}}(x)^{2(\alpha-1)}}\left(\nabla_{\boldsymbol{\xi}}\log q_{\boldsymbol{\xi}}(x)\right)\mathrm{d}x\end{aligned}$$

The integral present in the second addendum can be either evaluated from samples (i.e., replacing the expectation with the sample mean) or computed exactly for common classes of distributions, e.g. Gaussian distributions, as we show in Appendix 2.

4 EXPERIMENTAL DETAILS

In this appendix, we report the experimental details and additional experimental results.

Infrastructure The experiments have been run on two machines:

- 2 x CPUs Intel(R) Xeon(R) CPU E7-8880 v4 @ 2.20GHz (22 cores, 44 thread, 55 MB cache) and 128 GB RAM;
- 4 x Intel(R) Xeon(R) CPU E5-4610 v2 @ 2.30GHz (8 cores, 16 thread, 16 MB cache) and 256 GB RAM.

Environments The environments are the rllab implementations [Duan et al., 2016], MIT license, `https://github.com/rll/rllab`. The Swimmer environment belongs to the Mujoco suite [Todorov et al., 2012], MuJoCo Personal License, `http://www.mujoco.org/`.

Algorithms The TRPO implementation is taken from baselines, MIT licence, `https://github.com/openai/baselines`. For POIS we use the original implementation [Metelli et al., 2018], MIT license, `https://github.com/T3p/baselines`.

Hyperparameters In order to properly compare the algorithms, a set of 20 seeds has been chosen. A subset of 5 seeds, underlined, was used to test the performances during the tuning phase. Once the optimal hyperparameters were found, the experiments were extended to the other 15 seeds. In the following, we report the hyperparameter values for MBPExPI.

The *shift return* refers to the need for making the return non-negative in order to perform the optimization of the α-moment in MBPExPI. This procedure is carried out independently at each algorithm iteration by subtracting the minimum return among the ones observed. The *variance init* hyperparameter refers to the logarithm of the standard deviation. All experiments have been carried out with Gaussian policies linear with mean linear in the state variables and constant variance uniform over the state space.

Cartpole

- seeds: 0, 3, 11, 16, 19, 42, 66, 72, 84, 87, 90, 123, 222, 343, 404, 452, 542, 875, 943, 999
- max iters: 500
- policy: linear
- policy init: zeros
- capacity: 1
- inner: 1
- variance init: -1
- step size: 1 / gradient norm
- penalization: True
- delta: 0.75
- max offline iters: 10

Mountain Car

- seeds: 0, 3, 11, 16, 19, 42, 66, 72, 84, 87, 90, 123, 222, 343, 404, 452, 542, 875, 943, 999
- max iters: 500
- policy: linear
- policy init: zeros
- capacity: 1
- inner: 1
- variance init: -1
- step size: 2 / gradient norm
- penalization: True
- delta: 0.9
- max offline iters: 10
- shift return: True

Inverted Double Pendulum

- seeds: 0, 3, 11, 16, 19, 42, 66, 72, 84, 87, 90, 123, 222, 343, 404, 452, 542, 875, 943, 999
- max iters: 500
- policy: linear
- policy init: zeros
- capacity: 1
- inner: 1
- variance init: -1

- step size: 2 / gradient norm
- penalization: True
- delta: 0.99
- max offline iters: 10

Swimmer

- seeds: 0, 3, 11, 16, 19, 42, 66, 72, 84, 87, 90, 123, 222, 343, 404, 452, 542, 875, 943, 999
- max iters: 500
- policy: linear
- policy init: zeros
- capacity: 1
- inner: 1
- log-std init: -0.6
- step size: 1 / gradient norm
- penalization: True
- delta: 0.99
- max offline iters: 10
- shift return: True

For POIS (both AB and PB) and TRPO, the same hyperparameter value have been used, except for the algorithm-specific ones that have been tuned with the same protocol discussed above ($\delta_{\mathrm{KL}} \in \{0.001, 0.01, 0.1, 1\}$). In particular, for POIS, we employ the line search procedure presented in the original paper for setting the step-size. The following table summarizes the algorithm-specific hyperparameter values for the different algorithms and environments.

Environment / Algorithm	MBPExPI (delta)	AB-POIS (delta)	TRPO (max kl)
Cartpole	0.75	0.4	0.01
Mountain Car	0.9	0.9	0.01
Inverted Double Pendulum	0.99	0.1	0.001
Swimmer	0.99	0.8	0.01

Environment / Algorithm	PB-POIS (delta)	PB-MBPExPI (delta)
Cartpole	0.4	0.6
Mountain Car	1	0.00001
Inverted Double Pendulum	0.1	0.999999
Swimmer	0.4	0.4

4.1 NOISE ROBUSTNESS

As we have already observed, using the trajectory return $\mathcal{R}(\tau)$ as function f does no longer allow to provide performance improvement guarantees. Nevertheless, we conjecture that the loss of this property is compensated by the variance reduction implicit in our approach. In the direction of empirically showing this aspect, we tested the parameter-based version of MBPExPI in the Inverted Double Pendulum environment, with forced stochasticity in the environment. Specifically, whenever an action is prescribed by the policy the actual action to be executed is obtained by adding while Gaussian noise with standard deviation σ. The results are shown in Figure 1. We observe that our algorithm is overall competitive with PB-POIS and, in the case of $\sigma = 1$, significantly outperforms PB-POIS.

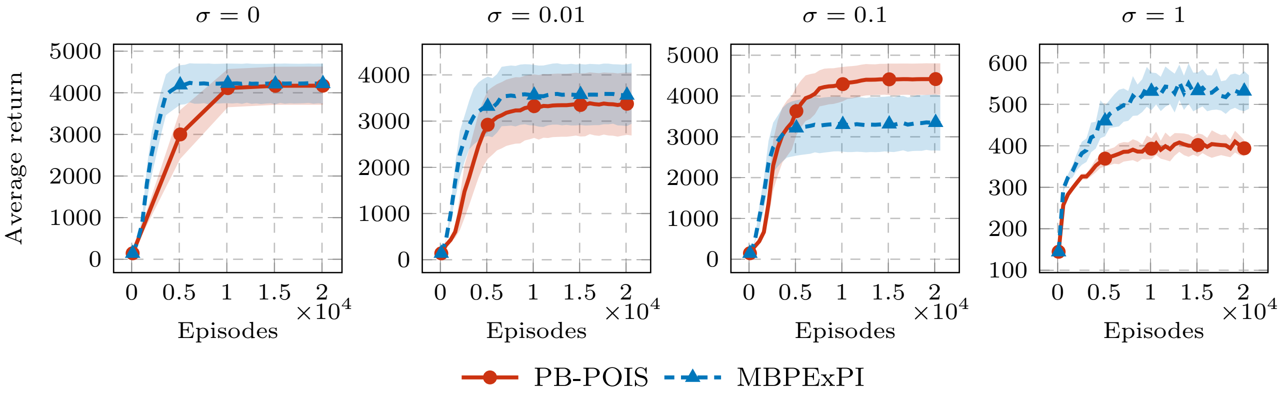

Figure 1: Learning curves comparing PB-POIS and PB-MBPExPI with increasing magnitude of the noise (20 runs, 95% c.i.).

4.2 ABOUT RETURN TRANSLATION

Our approach can be employed for non-negative functions f. Since in the PO experimental evaluation we employ $f = \mathcal{R}(\tau)$. Under the assumption that the immediate reward is bounded $R(s,a) \in [R_{\min}, R_{\max}]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we can make the return function with a simple translation and preserving the optimality of policies:

$$\overline{R}(\tau) = \mathcal{R}(\tau) - \underbrace{R_{\min}\frac{1-\gamma^H}{1-\gamma}}_{-c_{\min}},$$

where $R_{\min}\frac{1-\gamma^H}{1-\gamma}$ is the minimum achievable return. Of course, we can perform the translation even by using a constant $c \geqslant c_{\min} = -R_{\min}\frac{1-\gamma^H}{1-\gamma}$ and still obtain a translated return that remains positive. It is worth noting, from Theorem 4.3 that the size of the trust region is larger as the constant approaches the its minimum possible value.

For instance, we consider $\alpha = 2$, $f \geqslant 0$, and we apply a further translation with $c \geqslant 0$. From Theorem 4.3, we have:

$$D_2(I_{+c\circ f}[P] \| P) = \log \frac{\mathbb{E}_{x\sim P}[(f(x)+c)^2]}{\mathbb{E}_{x\sim P}[f(x)+c]^2} = \log \frac{\mathbb{E}_{x\sim P}[f(x)^2] + c^2 + 2c\mathbb{E}_{x\sim P}[f(x)]}{\mathbb{E}_{x\sim P}[f(x)]^2 + c^2 + 2c\mathbb{E}_{x\sim P}[f(x)]}.$$

Since $\mathbb{E}_{x\sim P}[f(x)^2] \geqslant \mathbb{E}_{x\sim P}[f(x)]^2$, we have that this expression is maximized with the smallest value of c, i.e., $c = 0$.