A Properties of PMD

We present lemmas relevant to the analysis of PMD. Key to the analysis is the Three-Point Descent Lemma, that relates the improvement of the proximal gradient update compared to an arbitrary point. It originally comes from [37] (Lemma 3.2) where a proof can be found, though we use a slightly modified version from [7] (Lemma 6).

Lemma A.1 (Three-Point Descent Lemma, Lemma 6 in [7])**.** *Suppose that $\mathcal{C} \subset \mathbb{R}^n$ is a closed convex set, $\phi : \mathcal{C} \to \mathbb{R}$ is a proper, closed convex function, $D_h(\cdot, \cdot)$ is the Bregman divergence generated by a function h of Legendre type and rint dom$h \cap \mathcal{C} \neq \emptyset$. For any $x \in$ rint dom$h$, let*

$$x^+ = \textit{argmin}_{u \in C}\{\phi(u) + D_h(u, x)\}. \tag{14}$$

Then $x^+ \in$ rint dom $h \cap C$ and $\forall u \in C$,

$$\phi(x^+) + D_h(x^+, x) \leq \phi(u) + D_h(u, x) - D_h(u, x^+) \tag{15}$$

The update (4) of PMD is an instance of the proximal minimisation (14) with $\mathcal{C} = \Delta(\mathcal{A})$, $x = \pi_s^k$ and $\phi(x) = -\eta_k \langle Q_s^k, x\rangle$. Plugging these into (15), Lemma A.1 relates the decrease in the proximal objective of π_s^{k+1} to any other policy, i.e. $\forall p \in \Delta(\mathcal{A})$,

$$-\eta_k \langle Q_s^k, \pi_s^{k+1}\rangle + D_h(\pi_s^{k+1}, \pi_s^k) \leq -\eta_k \langle Q_s^k, p\rangle + D_h(p, \pi_s^k) - D_h(p, \pi_s^{k+1}). \tag{16}$$

This equation is key to the analysis in Section 6. In particular, it allows us to prove the following lemma regarding the monotonic improvement in action-value of PMD iterates. This is an extension of Lemma 7 in [7].

Lemma A.2. *Consider the policies produced by the iterative updates of PMD in (4). Then for any $k \geq 0$,*

$$Q^{k+1}(s, a) \geq Q^k(s, a), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

A.1 Proof of Lemma A.2

We first present Lemma 7 from [7], from which Lemma A.2 almost immediately follows.

Lemma A.3 (Descent Property of PMD, Lemma 7 in [7])**.** *Consider the policies produced by the iterative updates of PMD in (4). Then for any $k \geq 0$*

$$\langle Q_s^k, \pi_s^{k+1} - \pi_s^k\rangle \geq 0, \quad \forall s \in \mathcal{S},$$
$$V^{k+1}(\rho) \geq V^k(\rho), \quad \forall \rho \in \Delta(\mathcal{S}).$$

Proof. From [7]. Recall that the Three-Point Descent Lemma states that $\forall p \in \Delta(\mathcal{A})$,

$$-\eta_k \langle Q_s^k, \pi_s^{k+1}\rangle + D_h(\pi_s^{k+1}, \pi_s^k) \leq -\eta_k \langle Q_s^k, p\rangle + D_h(p, \pi_s^k) - D_h(p, \pi_s^{k+1}).$$

Using this with $p = \pi_s^k$,

$$D_h(\pi_s^k, \pi_s^{k+1}) + D_h(\pi_s^{k+1}, \pi_s^k) \leq \eta_k \langle Q_s^k, \pi_s^{k+1} - \pi_s^k\rangle$$

and since the Bregman divergences are none-negative and $\eta_k > 0$,

$$0 \leq \langle Q_s^k, \pi_s^{k+1} - \pi_s^k\rangle$$

and the result follows by an application of the performance difference lemma (Appendix B)

$$\begin{aligned} V^{k+1}(\rho) - V^k(\rho) &= \frac{1}{1-\gamma}\mathbb{E}_{s \sim d_\rho^{k+1}}\left[\langle Q_s^k, \pi_s^{k+1} - \pi_s^k\rangle\right] \\ &\geq 0. \end{aligned}$$

Note that we use the performance difference lemma here because it gives a simple concise proof, but we do not actually need to. To maintain our claim that we avoid the use of the performance difference lemma, we can get the same result without it. We sketch how to do this as follows. From the first part of the lemma, we have

$$\langle Q_s^k, \pi_s^{k+1}\rangle \geq \langle Q_s^k, \pi_s^k\rangle = V^k(s),$$

in all states s. Now note that the left hand side above is

$$\begin{aligned}\langle Q_s^k, \pi_s^{k+1}\rangle &= \sum_a \pi^{k+1}(a|s) Q^k(s,a) \\ &= \sum_a \pi^{k+1}(a|s)\Big(r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^k(s')\Big)\end{aligned}$$

and we can then apply $\langle Q_{s'}^k, \pi_{s'}^{k+1}\rangle \geq V^k(s')$ at state s':

$$\begin{aligned}V^k(s) &\leq \sum_a \pi^{k+1}(a|s)\Big(r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^k(s')\Big) \\ &\leq \sum_a \pi^{k+1}(a|s)\Big(r(s,a) + \gamma \sum_{s'} p(s'|s,a) \sum_{a'} \pi^{k+1}(a'|s')\Big(r(s',a') + \gamma \sum_{s''} p(s''|s',a') V^k(s'')\Big)\Big)\end{aligned}$$

and as proceed iteratively in the limit you get exactly $V^{k+1}(s)$. ■

Since Lemma A.3 holds for any $\rho \in \Delta(\mathcal{S})$, it guarantees that the value in each state is non-decreasing for an update of PMD, i.e for all $s \in \mathcal{S}$,

$$V^{k+1}(s) - V^k(s) \geq 0.$$

Using this, we get

$$Q^{k+1}(s,a) - Q^k(s,a) = \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a)\Big(V^{k+1}(s') - V^k(s')\Big) \geq 0,$$

which concludes the proof. ■

A.2 Extension of Lemma A.2 to inexact setting:

As in the exact case, we first present Lemma 12 from [7] which is the extension of Lemma A.3 to the inexact case. We note that in the inexact case, we lose the monotonic increase of values due to the inaccuracy in our estimate $\widehat{Q}^k$ of Q_s^k.

Lemma A.4. *Consider the policies produced by the iterative updates of IPMD in (9). For any $k \geq 0$, if $\|\widehat{Q}_s^k - Q_s^k\|_\infty \leq \tau$, then*

$$\begin{aligned}\langle \widehat{Q}_s^k, \pi_s^{k+1} - \pi_s^k\rangle \geq 0, \quad \forall s \in \mathcal{S}, \\ V^{k+1}(\rho) \geq V^k(\rho) - \frac{2\tau}{1-\gamma}, \quad \forall \rho \in \Delta(\mathcal{S}).\end{aligned}$$

Proof. From [7]. The Three-Point Descent Lemma applied to the IPMD update (9) gives $\forall p \in \Delta(\mathcal{A})$,

$$-\eta_k \langle \widehat{Q}_s^k, \pi_s^{k+1}\rangle + D_h(\pi_s^{k+1}, \pi_s^k) \leq -\eta_k \langle \widehat{Q}_s^k, p\rangle + D_h(p, \pi_s^k) - D_h(p, \pi_s^{k+1}).$$

Using this with $p = \pi_s^k$,

$$D_h(\pi_s^k, \pi_s^{k+1}) + D_h(\pi_s^{k+1}, \pi_s^k) \leq \eta_k \langle \widehat{Q}_s^k, \pi_s^{k+1} - \pi_s^k\rangle$$

and since the Bregman divergences are none-negative and $\eta_k > 0$,

$$0 \leq \langle \widehat{Q}_s^k, \pi_s^{k+1} - \pi_s^k\rangle,$$

which proves the first inequality. Now we cannot use the above inequality directly with the performance difference lemma since $\widehat{Q}_s^k$ is not the true action-value. Instead, we have

$$\begin{aligned}V^{k+1}(\rho) - V^k(\rho) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{k+1}}\Big[\langle Q_s^k, \pi_s^{k+1} - \pi_s^k\rangle\Big] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{k+1}}\Big[\langle Q_s^k - \widehat{Q}_s^k, \pi_s^{k+1} - \pi_s^k\rangle + \langle \widehat{Q}_s^k, \pi_s^{k+1} - \pi_s^k\rangle\Big] \\ &\geq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{k+1}}\Big[-\|Q_s^k - \widehat{Q}_s^k\|_\infty \|\pi_s^{k+1} - \pi_s^k\|_1\Big] \\ &\geq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{k+1}}\Big[-2\tau\Big] \\ &= -\frac{2\tau}{1-\gamma}\end{aligned}$$

which concludes the proof. ■

Using the above lemma, we can state and prove the extension of Lemma A.2 to the inexact setting.

Lemma A.5. *Consider the policies produced by the iterative updates of IPMD in (9). For any $k \geq 0$, if $\|\widehat{Q}_s^k - Q_s^k\|_\infty \leq \tau$, then*

$$\widehat{Q}^{k+1}(s,a) \geq \widehat{Q}^k(s,a) - \frac{2\tau\gamma}{1-\gamma}, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Proof. As in the exact case, since Lemma A.4 holds for any $\rho \in \Delta(\mathcal{S})$, it applies to each state, i.e for all $s \in \mathcal{S}$,

$$V^{k+1}(s) - V^k(s) \geq -\frac{2\tau}{1-\gamma}.$$

Using this, we immediately have

$$Q^{k+1}(s,a) - Q^k(s,a) = \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a)\Big(V^{k+1}(s') - V^k(s')\Big) \geq \frac{-2\tau\gamma}{1-\gamma},$$

which concludes the proof. ■

B Performance difference lemma

Lemma B.1 (Performance Difference Lemma)**.** *For any $\pi, \pi' \in \Pi$, we have*

$$V^{\pi}(\rho) - V^{\pi'}(\rho) = \frac{1}{1-\gamma}\mathbb{E}_{s \sim d_\rho^\pi}\Big[\langle Q_s^{\pi'}, \pi_s - \pi'_s \rangle\Big].$$

The performance difference lemma [14] is a property that relates the difference in values of policies to the policies themselves. The proof can be found in their paper under Lemma 6.1.

C Guarantees of Theorem 4.1 for various step-size choices

We give here two more choices of $\{c_k\}_{k \in \mathbb{Z}_{\geq 0}}$ for the step-size 5 of PMD and their corresponding guarantees from Theorem 4.1:

- $c_i = c_0$ for some $c_0 > 0$ yields a step-size with a constant component. The resulting bound is

$$\|V^\star - V^k\|_\infty \leq \gamma^k \|V^\star - V^0\|_\infty + \frac{c_0}{1-\gamma},$$

  which converges linearly up to some accuracy controlled by c_0.

- $c_i = \gamma^{i+1} c_0$ for some initial $c_0 > 0$ will yield a step-size with a component that is geometrically increasing as in [7], though at a slower rate than the one discussed in Section 4. The resulting bound is

$$\|V^\star - V^k\|_\infty \leq \gamma^k \Big(\|V^\star - V^0\|_\infty + k c_0\Big),$$

  which converges linearly with the sought-for γ-rate, though in early iterations the k factor may dominate.

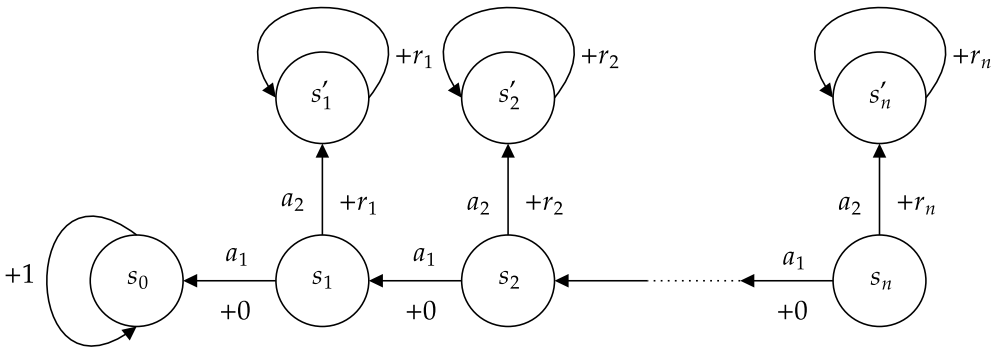

Figure 1: Example MDP used in the proof of Theorem 4.2

D Proof of Theorem 4.2

Fix $n > 0$ and $\delta \in (0, (1-\gamma)\gamma^n)$. Consider the MDP shown in Figure 1. The state space is $\mathcal{S} = \{s_0, s_1, s'_1, ..., s_n, s'_n\}$ and the action space is $\mathcal{A} = \{a_1, a_2\}$. There is a chain of states of length $n+1$ with the states indexed from 0 to n. The left-most state (s_0) is absorbing with reward +1. In the other states in the chain (s_i for $i = 1, ..., n$), the agent can take action a_1 and move left (to s_{i-1}) with reward of 0, or take action a_2 and move to an additional absorbing state unique to the state it is currently in (s'_i) with reward $r_i = \gamma^{i+1} + \delta$ (that the agent also receives in that state for all future time-steps). Summarising, we have for $1 \leq i \leq n$

$$
\begin{aligned}
&p(s_{i-1}|s_i, a_1) = 1, \quad r(s_i, a_1) = 0, \\
&p(s'_i|s_i, a_2) = 1, \quad r(s_i, a_2) = r_i = \gamma^{i+1} + \delta, \\
&p(s'_i|s'_i, a) = 1, \quad r(s'_i, a) = r_i = \gamma^{i+1} + \delta \quad \forall a \in \mathcal{A}.
\end{aligned}
$$

The value of δ is carefully restricted so that the optimal action in all the states of the chain is a_1. The proof will consist in showing that if the agent starts with an initial policy that places most probability mass on the sub-optimal action a_2, then it has to learn that a_1 is the optimal action in the state directly to the left before it can start switching from action a_2 to a_1 in the current state. And this can at best happen one iteration at a time starting starting from the left-most state. In particular, we consider π^0 s.t $\pi^0(a_1|s) = \alpha$, $\pi^0(a_2|s) = 1 - \alpha$ for all states and some α s.t $0 < \alpha \leq \delta(1-\gamma)$. We make the following claim from which the result will follow straightforwardly.

Claim: Fix $k < n$. The policies produced by PMD satisfy $\pi^k(a_1|s_i) \leq \alpha$ for $k < i \leq n$.

We prove this claim by induction.

Base Case: We want to show that $\pi^1(a_1|s_i) \leq \alpha$ for $i > 1$. We do this by showing that $Q^0(s_i, a_1) \leq Q^0(s_i, a_2)$ for $i > 1$ so that the probability of $\pi^1(a_1|s_i)$ cannot increase w.r.t $\pi^0(a_1|s_i)$, which is α

(this follows from $\langle Q_s^k, \pi_s^{k+1} - \pi_s^k \rangle \geq 0$ for all iterations of PMD). We have:

$$\begin{aligned}
Q^0(s_i, a_1) &= \gamma V^0(s_{i-1}) \\
&= \gamma\Big(\alpha Q^0(s_{i-1}, a_1) + (1-\alpha) Q^0(s_{i-1}, a_2)\Big) \\
&\leq \gamma\Big(\alpha \frac{\gamma^{i-1}}{1-\gamma} + \frac{r_{i-1}}{1-\gamma}\Big) \\
&\overset{(a)}{\leq} \gamma\Big(\delta(1-\gamma)\frac{\gamma^{i-1}}{1-\gamma} + \frac{\gamma^i + \delta}{1-\gamma}\Big) \\
&= \frac{\gamma^{i+1}}{1-\gamma} + \frac{\delta\gamma(1+\gamma^{i-1}-\gamma^i)}{1-\gamma} \\
&\overset{(b)}{\leq} \frac{\gamma^{i+1}}{1-\gamma} + \frac{\delta}{1-\gamma} \\
&= Q^0(s_i, a_2),
\end{aligned}$$

where we used $\alpha \leq \delta(1-\gamma)$ in (a) and $\gamma(1+\gamma^{i-1}-\gamma^i) < 1$ for $\gamma \in [0, 1)$ in (b). This concludes the base case.

Inductive Step: Now assume that the claim is true for k and we want to show that $\pi^{k+1}(a_1|s_i) \leq \alpha$ for $i > k+1$. We do this in the same way as the base case by showing that $Q^k(s_i, a_1) \leq Q^k(s_i, a_2)$ for $i > k+1$ so that the probability of $\pi^{k+1}(a_1|s_i)$ cannot increase w.r.t $\pi^k(a_1|s_i)$, which is less than or equal to α by the inductive hypothesis. We have:

$$\begin{aligned}
Q^k(s_i, a_1) &= \gamma V^k(s_{i-1}) \\
&= \gamma\Big(\pi^k(a_1|s_{i-1}) Q^k(s_{i-1}, a_1) + \pi^k(a_2|s_{i-1}) Q^k(s_{i-1}, a_2)\Big) \\
&\overset{(a)}{\leq} \gamma\Big(\alpha Q^k(s_{i-1}, a_1) + Q^k(s_{i-1}, a_2)\Big) \\
&\leq \gamma\Big(\alpha \frac{\gamma^{i-1}}{1-\gamma} + \frac{r_{i-1}}{1-\gamma}\Big) \\
&\overset{(b)}{\leq} \gamma\Big(\delta(1-\gamma)\frac{\gamma^{i-1}}{1-\gamma} + \frac{\gamma^i + \delta}{1-\gamma}\Big) \\
&= \frac{\gamma^{i+1}}{1-\gamma} + \frac{\delta\gamma(1+\gamma^{i-1}-\gamma^i)}{1-\gamma} \\
&\overset{(c)}{\leq} \frac{\gamma^{i+1}}{1-\gamma} + \frac{\delta}{1-\gamma} \\
&= Q^k(s_i, a_2),
\end{aligned}$$

where we used in (a) that $\pi^k(a_1|s_{i-1}) \leq \alpha$ for $i > k+1$, which is true by the inductive hypothesis since $i - 1 > k$, in (b) that $\alpha \leq \delta(1-\gamma)$ and in (c) that $\gamma(1+\gamma^{i-1}-\gamma^i) < 1$ for $\gamma \in [0, 1)$. This concludes the proof of the claim.

Now using the claim

$$\begin{aligned}
V^k(s_{k+1}) &= \pi^k(a_1|s_{k+1}) Q^k(s_{k+1}, a_1) + \pi^k(a_2|s_{k+1}) Q^k(s_{k+1}, a_2) \\
&\leq \alpha \frac{\gamma^{k+1}}{1-\gamma} + \frac{r_{k+1}}{1-\gamma} \\
&= \alpha \frac{\gamma^{k+1}}{1-\gamma} + \frac{\gamma^{k+2} + \delta}{1-\gamma},
\end{aligned}$$

so

$$\begin{aligned} V^\star(s_{k+1}) - V^k(s_{k+1}) &\geq \frac{\gamma^{k+1}}{1-\gamma} - \alpha\frac{\gamma^{k+1}}{1-\gamma} - \frac{\gamma^{k+2}+\delta}{1-\gamma} \\ &= \frac{\gamma^{k+1}(1-\gamma)}{1-\gamma} - \frac{\alpha\gamma^{k+1}+\delta}{1-\gamma} \\ &\geq \gamma^{k+1} - \frac{\alpha+\delta}{1-\gamma} \\ &\geq \gamma^{k+1} - \frac{2\delta}{1-\gamma}, \end{aligned} \tag{17}$$

where we used that $\alpha \leq \delta$. Now note that

$$\begin{aligned} V^0(s_1) &= \alpha Q^0(s_1, a_1) + (1-\alpha)Q^0(s_1, a_2) \\ &= \alpha\frac{\gamma}{1-\gamma} + (1-\alpha)\frac{\gamma^2+\delta}{1-\gamma}, \end{aligned}$$

so

$$\begin{aligned} V^\star(s_1) - V^0(s_1) &= \frac{\gamma}{1-\gamma} - \alpha\frac{\gamma}{1-\gamma} - (1-\alpha)\frac{\gamma^2+\delta}{1-\gamma} \\ &= (1-\alpha)\frac{\gamma}{1-\gamma} - (1-\alpha)\frac{\gamma^2+\delta}{1-\gamma} \\ &= \frac{1-\alpha}{1-\gamma}\Big(\gamma - \gamma^2 - \delta\Big) \\ &\leq \frac{1-\alpha}{1-\gamma}\Big(\gamma - \gamma^2\Big) \\ &= \gamma\frac{1-\alpha}{1-\gamma}\Big(1-\gamma\Big) \\ &= \gamma(1-\alpha) \\ &\leq \gamma \end{aligned}$$

and by induction we can show this is the case for all states (above is base case), the inductive step is as follows (assuming $V^\star(s_k) - V^0(s_k) \leq \gamma$),

$$\begin{aligned} V^\star(s_{k+1}) - V^0(s_{k+1}) &= \frac{\gamma^{k+1}}{1-\gamma} - (1-\alpha)\frac{\gamma^{k+2}+\delta}{1-\gamma} - \alpha\gamma V^0(s_k) \\ &= (1-\alpha)\Big[\frac{\gamma^{k+1} - \gamma^{k+2} - \delta}{1-\gamma}\Big] + \alpha\gamma\Big[V^\star(s_k) - V^0(s_k)\Big] \\ &\leq (1-\alpha)\gamma^{k+1} + \alpha\gamma^2 \\ &\leq \gamma \end{aligned}$$

and so

$$\|V^\star - V^0\|_\infty \leq \gamma,$$

which combining with (17) gives,

$$\begin{aligned} V^\star(s_{k+1}) - V^k(s_{k+1}) &\geq \gamma^k\|V^\star - V^0\|_\infty - \frac{2\delta}{1-\gamma} \\ \implies \|V^\star - V^k\|_\infty &\geq \gamma^k\|V^\star - V^0\|_\infty - \frac{2\delta}{1-\gamma}, \end{aligned}$$

which concludes the proof. ■

E Proof of Theorem 4.3

Consider the same MDP as in the proof of Theorem 4.2 in Appendix D (see Figure 1). Denote $c = \frac{1-\gamma}{8}$ and note that $c < \frac{\sqrt{\gamma}}{1+\sqrt{\gamma}}\frac{1-\gamma}{2}$ since $\frac{1}{4} < \frac{\sqrt{\gamma}}{1+\sqrt{\gamma}}$ for $\gamma > 0.2$.

Suppose you consider NPG updates with initial policy $\pi^0(a_1|s_i) = \alpha$. Recall that NPG is the instance of PMD with relative entropy as the mirror map. It can be shown that NPG has the closed form update

$$\pi^{k+1}(a|s) = \frac{\pi^k(a|s)e^{\eta_k Q^k(s,a)}}{\sum_{a'}\pi^k(a'|s)e^{\eta_k Q^k(s,a')}}.$$

We know from the proof of Theorem D that for any step-size regime, for $i > k+1$

$$Q^k(s_i, a_1) \leq Q^k(s_i, a_2).$$

Now, $\|V^\star - V^0\|_\infty = V^\star(s_1) - V^0(s_1) \leq \gamma - \frac{\delta}{1-\gamma}$ (see Section E.1 below). The idea of the proof is to show that satisfying the bound given in the statement of the theorem will imply that a certain condition on the step-size.

Fix a state s_k and let k_0 be the first iteration where $Q^{k_0}(s_k, a_1) > Q^{k_0}(s_k, a_2)$. By the above, we must have $k \leq k_0 + 1$, or $k_0 \geq k - 1$. By the proof of Theorem D, we also have $\pi^{k_0}(a_1|s_k) \leq \alpha$ (before iteration k_0, $Q(s_k, \cdot)$ favors a_2, so $\pi^{k_0}(a_1|s_k)$ has not increased compared to $\pi^0(a_1|s_k) = \alpha$).

We want a γ-contraction at every iteration, i.e. we assume the following is satisfied:

$$V^\star(s_k) - V^{k_0+1}(s_k) \leq \gamma^{k_0+1}(\|V^\star - V^0\|_\infty + c) \leq \gamma^{k_0+1}(\gamma - \frac{\delta}{1-\gamma} + c).$$

Now, by direct computation,

$$\begin{aligned} V^\star(s_k) - V^{k_0+1}(s_k) &= \pi^{k_0+1}(a_1|s_k)\gamma(V^\star(s_{k-1}) - V^{k_0+1}(s_{k-1})) + \pi^{k_0+1}(a_2|s_2)\frac{\gamma^k - r_k}{1-\gamma} \\ &\geq \pi^{k_0+1}(a_2|s_2)\frac{\gamma^k - r_k}{1-\gamma} = \pi^{k_0+1}(a_2|s_2)(\gamma^k - \frac{\delta}{1-\gamma}). \end{aligned}$$

Putting this together with the above (this is an implication as this is about the necessity rather than sufficiency), we must have:

$$\begin{aligned} &\pi^{k_0+1}(a_2|s_2)(\gamma^k - \frac{\delta}{1-\gamma}) \leq \gamma^{k_0+1}(\gamma - \frac{\delta}{1-\gamma} + c) \\ \implies &\pi^{k_0+1}(a_2|s_2) \leq \frac{\gamma^{k_0+1}(\gamma - \frac{\delta}{1-\gamma} + c)}{(\gamma^k - \frac{\delta}{1-\gamma})} = \beta \end{aligned}$$

If we choose $\delta < \frac{1}{2}(1-\gamma)(1-\sqrt{\gamma})\gamma^k$ then $\beta < \sqrt{\gamma}$ and require

$$\pi^{k_0+1}(a_2|s_2) \leq \sqrt{\gamma}.$$

To see this, start from $\beta \leq \sqrt{\gamma}$, this is equivalent to

$$
\begin{aligned}
&\frac{\gamma^{k_0+1}(\gamma - \frac{\delta}{1-\gamma} + c)}{(\gamma^k - \frac{\delta}{1-\gamma})} \leq \sqrt{\gamma} \\
\Longleftarrow\ &\frac{\gamma^{k}(\gamma - \frac{\delta}{1-\gamma} + c)}{(\gamma^k - \frac{\delta}{1-\gamma})} \leq \sqrt{\gamma} \quad \text{since } k_0 + 1 \geq k \\
\Longleftrightarrow\ &\gamma^k(\gamma - \frac{\delta}{1-\gamma} + c) \leq \sqrt{\gamma}(\gamma^k - \frac{\delta}{1-\gamma}) \\
\Longleftarrow\ &\gamma^k(\gamma + c) \leq \sqrt{\gamma}(\gamma^k - \frac{\delta}{1-\gamma}) \\
\Longleftrightarrow\ &\gamma^{k-\frac{1}{2}}(\gamma + c) \leq \gamma^k - \frac{\delta}{1-\gamma} \\
\Longleftrightarrow\ &\frac{\delta}{1-\gamma} \leq \gamma^{k-\frac{1}{2}}(\sqrt{\gamma} - \gamma - c) \\
\Longleftarrow\ &\frac{\delta}{1-\gamma} \leq \gamma^{k-\frac{1}{2}}(\sqrt{\gamma} - \gamma - \frac{\sqrt{\gamma}}{1+\sqrt{\gamma}}\frac{1-\gamma}{2}) \quad \text{since } -c > -\frac{\sqrt{\gamma}}{1+\sqrt{\gamma}}\frac{1-\gamma}{2} \\
\Longleftrightarrow\ &\frac{\delta}{1-\gamma} \leq \gamma^{k-\frac{1}{2}}(\sqrt{\gamma} - \gamma - \frac{\sqrt{\gamma}}{2}(1-\sqrt{\gamma})) \\
\Longleftrightarrow\ &\frac{\delta}{1-\gamma} \leq \gamma^{k-\frac{1}{2}}(\sqrt{\gamma} - \gamma - \frac{\sqrt{\gamma}}{2} + \frac{\gamma}{2}) \\
\Longleftrightarrow\ &\frac{\delta}{1-\gamma} \leq \gamma^{k-\frac{1}{2}}(\frac{\sqrt{\gamma}}{2} - \frac{\gamma}{2}) \\
\Longleftrightarrow\ &\delta \leq \frac{1}{2}\gamma^k(1-\sqrt{\gamma})(1-\gamma),
\end{aligned}
$$

which is the condition for δ we imposed initially.

To achieve the above condition $\pi^{k_0+1}(a_2|s_2) \leq \sqrt{\gamma}$, recalling that $\pi^{k_0}(a_2|s_2) \geq 1-\alpha$, η_{k_0} has to satisfy

$$
\eta_{k_0} \geq \frac{1}{Q^{k_0}(s_k, a_1) - Q^{k_0}(s_k, a_2)}\Big[\log((1-\alpha)(1-\sqrt{\gamma})) + KL(\tilde{\pi}_{s_k}^{k_0+1}, \pi_{s_k}^{k_0})\Big]
$$

To see this, again start from $\pi^{k_0+1}(a_2|s_2) \leq \sqrt{\gamma}$, this is equivalent to (use $k_0 = m$ for simplicity of notation) using the closed-form update of NPG:

$$
\begin{aligned}
&\pi^m(a_2|s_2)\exp(\eta_m Q^m(s_k, a_2)) \leq \\
&\sqrt{\gamma}(\pi^m(a_2|s_2)\exp(\eta_m Q^m(s_k, a_2)) + \pi^m(a_1|s_2)\exp(\eta_m Q^m(s_k, a_1))) \\
\Longleftrightarrow\ &\frac{1}{\sqrt{\gamma}} \leq 1 + \frac{\pi^m(a_1|s_2)}{\pi^m(a_2|s_2)}\exp(\eta_m(Q^m(s_k, a_1) - Q^m(s_k, a_2))) \\
\Longleftrightarrow\ &\frac{1-\sqrt{\gamma}}{\sqrt{\gamma}}\frac{\pi^m(a_2|s_2)}{\pi^m(a_1|s_2)} \leq \exp(\eta_m(Q^m(s_k, a_1) - Q^m(s_k, a_2))) \\
\Longleftrightarrow\ &\eta_m(Q^m(s_k, a_1) - Q^m(s_k, a_2)) \geq \log\Big(\frac{1-\sqrt{\gamma}}{\sqrt{\gamma}}\frac{\pi^m(a_2|s_2)}{\pi^m(a_1|s_2)}\Big) \\
\Longleftrightarrow\ &\eta_m \geq \frac{1}{Q^m(s_k, a_1) - Q^m(s_k, a_2)}\Big[\log\Big(\frac{1-\sqrt{\gamma}}{\sqrt{\gamma}}\pi^m(a_2|s_2)\Big) + \log\Big(\frac{1}{\pi^m(a_1|s_2)}\Big)\Big] \\
\Longrightarrow\ &\eta_m \geq \frac{1}{Q^m(s_k, a_1) - Q^m(s_k, a_2)}\Big[\log\Big((1-\alpha)\frac{1-\sqrt{\gamma}}{\sqrt{\gamma}}\Big) + KL(\tilde{\pi}_{s_k}^{m+1}, \pi_{s_k}^{m})\Big] \\
\Longrightarrow\ &\eta_m \geq \frac{1}{Q^m(s_k, a_1) - Q^m(s_k, a_2)}\Big[\log\Big((1-\alpha)(1-\sqrt{\gamma})\Big) + KL(\tilde{\pi}_{s_k}^{m+1}, \pi_{s_k}^{m})\Big].
\end{aligned}
$$

As we take $\alpha \to 0$, the KL term will dominate. In particular, note $\alpha < 1-\gamma$ so $1-\alpha > \gamma$ and

$$
(1-\alpha)(1-\sqrt{\gamma}) > \gamma(1-\sqrt{\gamma})
$$

and if we further impose the condition $\alpha < \gamma^2(1-\sqrt{\gamma})^2$ then

$$(1-\alpha)(1-\sqrt{\gamma}) > \sqrt{\alpha} > \sqrt{\pi^{k_0}(a_1|s_2)}$$

and the step-size needs to satisfy the following condition:

$$\begin{aligned}
\eta_{k_0} &\geq \frac{1}{Q^{k_0}(s_k,a_1) - Q^{k_0}(s_k,a_2)}\Big[\log(\sqrt{\pi^{k_0}(a_1|s_2)}) + KL(\tilde{\pi}_{s_k}^{k_0+1}, \pi_{s_k}^{k_0})\Big] \\
&= \frac{1}{Q^{k_0}(s_k,a_1) - Q^{k_0}(s_k,a_2)}\Big[-\frac{1}{2}KL(\tilde{\pi}_{s_k}^{k_0+1}, \pi_{s_k}^{k_0}) + KL(\tilde{\pi}_{s_k}^{k_0+1}, \pi_{s_k}^{k_0})\Big] \\
&= \frac{1}{2(Q^{k_0}(s_k,a_1) - Q^{k_0}(s_k,a_2))}KL(\tilde{\pi}_{s_k}^{k_0+1}, \pi_{s_k}^{k_0}) \qquad (18)
\end{aligned}$$

Distinct Iterations: Note that the iteration $k_0(s_k)$ where $Q(\cdot, s_k)$ starts becoming bigger at a_1 that a_2 is distinct for each s_k. Fix any s_k and $k_0 = k_0(s_k)$. We have

$$\begin{aligned}
Q^{k_0}(s_k,a_1) &< Q^{k_0}(s_k,a_2) \\
Q^{k_0+1}(s_k,a_2) &\leq Q^{k_0+1}(s_k,a_1)
\end{aligned}$$

then $\pi^{k_0+1}(a_1|s_k) \leq \pi^{k_0}(a_1|s_k) \leq \alpha$ (since Q^t points towards a_2 in s_k for all $t \leq k$). Then applying exactly the same steps as in the proof of Theorem 4.2, we have

$$Q^{k_0+1}(s_{k+1},a_1) < Q^{k_0+1}(s_{k+1},a_2),$$

meaning that $k_0(s_k)$ is disctinct to $k_0(s_{k+1})$.

Upper Bounding Q-value difference: We want to upper-bound the Q-value difference appearing in the step-size condition above. We have,

$$\begin{aligned}
Q^{k_0}(s_k,a_2) &= \frac{r_k}{1-\gamma} = \frac{\gamma^{k+1}+\delta}{1-\gamma} \\
Q^{k_0}(s_k,a_1) &= \gamma V^{k_0}(s_{k-1}) \leq \frac{\gamma^k}{1-\gamma}.
\end{aligned}$$

So,

$$\begin{aligned}
Q^{k_0}(s_k,a_1) - Q^{k_0}(s_k,a_2) &\leq \frac{\gamma^k}{1-\gamma} - \frac{\gamma^{k+1}+\delta}{1-\gamma} \\
&= \gamma^k - \frac{\delta}{1-\gamma} \\
&\leq \gamma^k.
\end{aligned}$$

Plugging this into the above bound (18), if the iterates of NPG are to satisfy the bound with the γ-rate in the statement of the theorem, the step-size must at least satisfy the following condition:

$$\eta_{k_0} \geq \frac{1}{2\gamma^k}KL(\tilde{\pi}_{s_k}^{k_0+1}, \pi_{s_k}^{k_0}),$$

which concludes the proof. ■

E.1 Largest sub-optimality gap at iteration 0

In this section, we prove the claim that

$$\|V^\star - V^0\|_\infty = V^\star(s_1) - V^0(s_1) \leq \gamma - \frac{\delta}{1-\gamma}$$

Proof: First of all, $V^\star(s_1) - V^0(s_1) = \pi^0(a_2|s_1)\frac{\gamma - r_1}{1-\gamma} = (1-\alpha)(\gamma - \frac{\delta}{1-\gamma}) \leq \gamma - \frac{\delta}{1-\gamma}$. For the first part, we proceed by induction. We will use throughout that

$$\frac{\gamma^k - r_k}{1-\gamma} = \gamma^k - \frac{\delta}{1-\gamma} \leq V^\star(s_1) - V^0(s_1) = (1-\alpha)(\gamma - \frac{\delta}{1-\gamma}).$$

This is true if (when LHS is the largest)

$$\gamma^2 - \frac{\delta}{1-\gamma} \leq (1-\alpha)(\gamma - \frac{\delta}{1-\gamma})$$

which holds when

$$\begin{aligned} &\alpha \leq \frac{\gamma(1-\gamma)^2}{\gamma(1-\gamma)-\delta} \\ \Longleftarrow\ &\alpha \leq 1-\gamma \end{aligned}$$

Base Case:

$$\begin{aligned} V^\star(s_2) - V^0(s_2) &= \alpha\gamma(V^\star(s_1) - V^0(s_1)) + (1-\alpha)\frac{\gamma^2 - r_2}{1-\gamma} \\ &\leq \alpha\gamma(V^\star(s_1) - V^0(s_1)) + (1-\alpha)(V^\star(s_1) - V^0(s_1)) \\ &\leq V^\star(s_1) - V^0(s_1) \end{aligned}$$

Inductive Step: Assume true for k. Then,

$$\begin{aligned} V^\star(s_{k+1}) - V^0(s_{k+1}) &= \alpha\gamma(V^\star(s_k) - V^0(s_k)) + (1-\alpha)\frac{\gamma^{k+1} - r_{k+1}}{1-\gamma} \\ &\leq \alpha\gamma(V^\star(s_1) - V^0(s_1)) + (1-\alpha)(V^\star(s_1) - V^0(s_1)) \\ &\leq V^\star(s_1) - V^0(s_1), \end{aligned}$$

which concludes the proof. ■

F Inexact policy mirror descent and the generative model

The following Lemma from [7] controls the accuracy of the estimator $\widehat{Q}_s^k$ specified in (10) of Section 5 with respect to H and M_k:

Lemma F.1 (Lemma 15 in [7])**.** *Consider using (10) to estimate Q_s^k for all state-action pairs for K iterations of IPMD. Then for $\delta \in (0,1)$, if for all $k \leq K$,*

$$M_k \geq \frac{\gamma^{-2H}}{2} log\big(\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta}\big).$$

Then with probability at least $1-\delta$, we have for all $k \leq K$,

$$\|\widehat{Q}_s^k - Q_s^k\|_\infty \leq \frac{2\gamma^H}{1-\gamma}.$$

The proof of this result can be found in Lemma 15 of [7].

F.1 Proof of Theorem 5.1

This proof is similar to that of [7] (Theorem 14). It is also similar in structure to the proof of Theorem 4.1 in Section 6.

Fix a state $s \in \mathcal{S}$ and an integer $k \geq 0$. For now let's assume that our Q-estimates are τ-accurate ($\tau > 0$), i.e.

$$\|Q^k - \widehat{Q}^k\|_\infty \leq \tau$$

for all $k \geq 0$. With this assumption, we have from Lemma A.5 in Appendix A.1,

$$Q^{k+1}(s,a) \geq Q^k(s,a) - \frac{2\gamma\tau}{1-\gamma}, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Now proceeding in a similar way to Section 6,

$$\begin{aligned}
\langle \widehat{Q}_s^k, \pi_s^\star - \pi_s^{k+1} \rangle &= \langle Q_s^k, \pi_s^\star - \pi_s^{k+1} \rangle + \langle \widehat{Q}_s^k - Q_s^k, \pi_s^\star - \pi_s^{k+1} \rangle \\
&\geq \langle Q_s^k, \pi_s^\star \rangle - \langle Q_s^k, \pi_s^{k+1} \rangle - \|\widehat{Q}_s^k - Q_s^k\|_\infty \|\pi_s^\star - \pi_s^{k+1}\|_1 \\
&\geq \langle Q_s^k, \pi_s^\star \rangle - \langle Q_s^{k+1}, \pi_s^{k+1} \rangle - \frac{2\gamma\tau}{1-\gamma} - 2\tau \\
&\geq \langle Q_s^k, \pi_s^\star \rangle - V^{k+1}(s) - \frac{4\gamma\tau}{1-\gamma} \\
&= \langle Q_s^k - Q_s^\star, \pi_s^\star \rangle + V^\star(s) - V^{k+1}(s) - \frac{4\gamma\tau}{1-\gamma} \\
&\geq -\|Q_s^\star - Q_s^k\|_\infty + V^\star(s) - V^{k+1}(s) - \frac{4\gamma\tau}{1-\gamma} \\
&\geq -\gamma\|V^\star - V^k\|_\infty + V^\star(s) - V^{k+1}(s) - \frac{4\gamma\tau}{1-\gamma}.
\end{aligned}$$

Now again proceeding exactly as in Section 6 with this extra τ-term using the step-size condition ($c_k = \gamma^{2k+1}$), we end up with

$$\|V^\star - V^{k+1}\|_\infty \leq \gamma\|V^\star - V^k\|_\infty + \gamma^{2k+1} + \frac{4\gamma\tau}{1-\gamma}.$$

Unravelling this recursion yields

$$\begin{aligned}
\|V^\star - V^k\|_\infty &\leq \gamma^k \Big(\|V^\star - V^0\|_\infty + \sum_{i=1}^{k} \gamma^{-i} \gamma^{2(i-1)+1} \Big) + \frac{4\gamma\tau}{1-\gamma} \sum_{i=0}^{k-1} \gamma^i \\
&\leq \gamma^k \Big(\|V^\star - V^0\|_\infty + \frac{1}{1-\gamma} \Big) + \frac{4\gamma\tau}{(1-\gamma)^2}.
\end{aligned}$$

Now using the properties of the estimator (10) in Lemma F.1, we have with probability $1-\delta$ for all $0 \le k \le K$,

$$\tau = \frac{2\gamma^H}{1-\gamma},$$

giving

$$\begin{aligned}\|V^\star - V^k\|_\infty &\le \gamma^k\Big(\|V^\star - V^0\|_\infty + \frac{1}{1-\gamma}\Big) + \frac{8\gamma^H}{(1-\gamma)^3} \\ &\le \frac{2}{1-\gamma}\gamma^k + \frac{8\gamma^H}{(1-\gamma)^3}.\end{aligned}$$

This establishes the first bound. Now

$$\begin{aligned}K > \frac{1}{1-\gamma}\log\frac{4}{(1-\gamma)\varepsilon} &\implies \frac{2}{1-\gamma}\gamma^k \le \varepsilon/2, \\ H \ge \frac{1}{1-\gamma}\log\frac{16}{(1-\gamma)^3\varepsilon} &\implies \frac{8\gamma^H}{(1-\gamma)^3} \le \varepsilon/2\end{aligned}$$

giving

$$\|V^\star - V^k\|_\infty \le \varepsilon/2 + \varepsilon/2 = \varepsilon$$

as required. In terms of M, we have

$$\begin{aligned}M &\ge \frac{\gamma^{-2H}}{2}\log\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta} \\ &\ge \frac{1}{2}\Big(\frac{16}{(1-\gamma)^3\varepsilon}\Big)^2\log\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta} \\ &= \frac{16^2}{2(1-\gamma)^6\varepsilon^2}\log\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta}\end{aligned}$$

and the corresponding number of calls to the sampling model, i.e. the sample complexity is (what we have shown above is actually a lower bound but can choose K, H, M so that it is of the following order),

$$|\mathcal{S}| \cdot |\mathcal{A}| \cdot K \cdot H \cdot M = \tilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^8\varepsilon^2}\Big),$$

where the notation $\tilde{O}()$ hides poly-logarithmic factors. This completes the proof. ■

G MDP examples

G.1 MDP on which distribution-mismatch coefficient scales with size of state space

We construct an MDP on which

$$\theta_\rho = \frac{1}{1-\gamma}\left\|\frac{d_\rho^\star}{\rho}\right\|_\infty,$$

scales with $|\mathcal{S}|$, and hence so does the iteration complexity of the bound of [7] for exact PMD.

Consider an MDP with state-space $\mathcal{S} = \{s_1, s_2, ..., s_n\}$ of size n and arbitrary action space $\mathcal{A}$. s_1 is an absorbing state giving out rewards of 1 at each time-step, regardless of the action taken, i.e

$$p(s_1|s_1, a) = 1, \quad r(s_1, a) = 1 \quad \forall a \in \mathcal{A}.$$

All others states have an action, say a_1, that gives out a reward of 1 and with probability $1-\delta$ brings the agent to state s_1 for some $\delta > 0$ and spreads the remaining δ probability arbitrarily amongst the other states. The other actions have arbitrary rewards strictly less than 1 associated to them, and arbitrary transition probabilities that place 0 mass on state s_1, i.e

$$\begin{aligned} &p(s_1|s, a_1) = 1-\delta, \quad r(s, a_1) = 1 \quad \forall s \neq s_1, \\ &p(s_1|s, a) = 0, \quad r(s, a) < 1 \quad \forall s \neq s_1, \forall a \neq a_1. \end{aligned}$$

Denote $r_{\max} = \max_{s\neq s_1, a\neq a_1} r(s, a) < 1$. The following condition ensures that a_1 is the optimal action in all states,

$$\delta \leq \frac{1-\gamma}{\gamma}(1 - r_{\max})$$

so that $\pi^\star(s) = a_1$ for all states s. To see this, consider $s_i \neq s_1$, $a_m \neq a_1$ and an arbitrary policy π,

$$\begin{aligned} Q^\pi(s_i, a_1) &= 1 + \gamma\Big(\frac{1-\delta}{1-\gamma} + \sum_{j=2}^n p(s_j|s_i, a_1)V^\pi(s_j)\Big) \\ &\geq 1 + \gamma\frac{1-\delta}{1-\gamma} \\ Q^\pi(s_i, a_m) &= r(s_i, a_m) + \gamma\sum_{j=2}^n p(s_j|s_i, a_1)V^\pi(s_j) \\ &\leq r_{\max} + \gamma\frac{1}{1-\gamma} \end{aligned}$$

and solving

$$r_{\max} + \gamma\frac{1}{1-\gamma} \leq 1 + \gamma\frac{1-\delta}{1-\gamma}$$

will yield the condition above.

Then for $t \geq 1$ (abusing notation, s_t denotes the state at time t),

$$\begin{aligned} \mathbb{P}^{\pi^\star}(s_t = s_1|s_0 = s) &= \sum_{s'} \mathbb{P}^{\pi^\star}(s_t = s_1, s_{t-1} = s'|s_0 = s) \\ &= \sum_{s'} p(s_1|s', a_1)\mathbb{P}^{\pi^\star}(s_{t-1} = s'|s_0 = s) \\ &\geq \sum_{s'} (1-\delta)\mathbb{P}^{\pi^\star}(s_{t-1} = s'|s_0 = s) \\ &= 1 - \delta \end{aligned}$$

and

$$\begin{aligned} d_\rho^\star(s_1) &= (1-\gamma)\sum_s \rho(s)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}^{\pi^\star}(s_t = s_1 | s_0 = s) \\ &\geq (1-\gamma)\sum_s \rho(s)\sum_{t=1}^{\infty}\gamma^t (1-\delta) \\ &\geq (1-\gamma)\sum_s \rho(s)\frac{\gamma}{1-\gamma}(1-\delta) \\ &= \gamma(1-\delta). \end{aligned}$$

Now

$$\left\| \frac{d_\rho^\star}{\rho} \right\|_\infty \geq \frac{d_\rho^\star(s_1)}{\rho(s_1)} \geq \frac{\gamma(1-\delta)}{\rho(s_1)}$$

and depending on what ρ you consider, θ_ρ can be arbitrarily large. In particular, the natural choice of the uniform starting-state distribution $\rho(s) = 1/n$ leads to

$$\theta_\rho \geq n\frac{\gamma(1-\delta)}{(1-\gamma)}$$

which gives an iteration complexity under the result of [7] for an ε-optimal policy that is

$$n\frac{\gamma(1-\delta)}{(1-\gamma)}\log\frac{2}{(1-\gamma)\varepsilon}.$$

Recall that $n = |\mathcal{S}|$, so this iteration complexity scales linearly with the size of the state space.

G.2 Family of MDPs on which sub-optimality gaps can be made arbitrarily small

We present how to construct a family of MDPs on which $\Delta^k(s)$ defined in Section 4 can be made arbitrarily small.

Consider an arbitrary MDP $\mathcal{M}$ with state space $\mathcal{S}$ and action space $\mathcal{A}$. For each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, create a duplicate action a' s.t the transitions from that action in that state are the same as for the original pair, i.e

$$p(s'|s,a) = p(s'|s,a') \quad \forall s' \in \mathcal{S}$$

and the reward is shifted down by $\delta > 0$ from the original reward, i.e

$$r(s,a') = r(s,a) - \delta.$$

This results in a new MDP $\mathcal{M}'$ with an augmented action space $\mathcal{A}'$, that is twice the size of the action space of the original MDP $\mathcal{M}$. In terms of action-value of an arbitrary policy π, this results in

$$Q^\pi_{\mathcal{M}'}(s,a) - Q^\pi_{\mathcal{M}'}(s,a') = \delta,$$

where the notation $Q^\pi_{\mathcal{M}'}$ refers to action-values in the MDP $\mathcal{M}'$. In terms of sub-optimality gaps, this gives

$$\Delta^\pi(s) \leq \delta.$$

Choosing δ small enough, we can make the step-size of [9] arbitrarily large, at least in early iterations. The step-size condition (5) of Theorem 4.1 will be less affected by this issue as it does not depend directly on $\Delta^k(s)$, and not at all in the first iteration. Beyond its generality to PMD, this illustrates the benefit of our result restricted to NPG over the result of [9].