Deterministic Policy Gradient Algorithms: Supplementary Material

A. Regularity Conditions

Within the text we have referred to regularity conditions on the MDP:

Regularity conditions A.1: $p(s'|s,a)$, $\nabla_a p(s'|s,a)$, $\mu_\theta(s)$, $\nabla_\theta \mu_\theta(s)$, $r(s,a)$, $\nabla_a r(s,a)$, $p_1(s)$ are continuous in all parameters and variables s, a, s' and x.

Regularity conditions A.2: there exists a b and L such that $\sup_s p_1(s) < b$, $\sup_{a,s,s'} p(s'|s,a) < b$, $\sup_{a,s} r(s,a) < b$, $\sup_{a,s,s'} ||\nabla_a p(s'|s,a)|| < L$, and $\sup_{a,s} ||\nabla_a r(s,a)|| < L$.

B. Proof of Theorem 1

proof of Theorem 1. The proof follows along the same lines of the standard stochastic policy gradient theorem in Sutton et al. (1999). Note that the regularity conditions A.1 imply that $V^{\mu_\theta}(s)$ and $\nabla_\theta V^{\mu_\theta}(s)$ are continuous functions of θ and s and the compactness of $\mathcal{S}$ further implies that for any θ, $||\nabla_\theta V^{\mu_\theta}(s)||$, $||\left.\nabla_a Q^{\mu_\theta}(s,a)\right|_{a=\mu_\theta(s)}||$ and $||\nabla_\theta \mu_\theta(s)||$ are bounded functions of s. These conditions will be necessary to exchange derivatives and integrals, and the order of integration whenever necessary in the following proof. We have,

$$
\begin{aligned}
\nabla_\theta V^{\mu_\theta}(s) &= \nabla_\theta Q^{\mu_\theta}(s, \mu_\theta(s)) \\
&= \nabla_\theta \left(r(s,\mu_\theta(s)) + \int_{\mathcal{S}} \gamma p(s'|s,\mu_\theta(s)) V^{\mu_\theta}(s') \mathrm{d}s' \right) \\
&= \nabla_\theta \mu_\theta(s) \left. \nabla_a r(s,a) \right|_{a=\mu_\theta(s)} + \nabla_\theta \int_{\mathcal{S}} \gamma p(s'|s,\mu_\theta(s)) V^{\mu_\theta}(s') \mathrm{d}s' \\
&= \nabla_\theta \mu_\theta(s) \left. \nabla_a r(s,a) \right|_{a=\mu_\theta(s)} \\
&\quad + \int_{\mathcal{S}} \gamma \left(p(s'|s,\mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') + \nabla_\theta \mu_\theta(s) \left. \nabla_a p(s'|s,a) \right|_{a=\mu_\theta(s)} V^{\mu_\theta}(s') \right) \mathrm{d}s' \qquad (1) \\
&= \nabla_\theta \mu_\theta(s) \nabla_a \left. \left(r(s,a) + \int_{\mathcal{S}} \gamma p(s'|s,a) V^{\mu_\theta}(s') \mathrm{d}s' \right) \right|_{a=\mu_\theta(s)} \\
&\quad + \int_{\mathcal{S}} \gamma p(s'|s,\mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') \mathrm{d}s' \\
&= \nabla_\theta \mu_\theta(s) \left. \nabla_a Q^{\mu_\theta}(s,a) \right|_{a=\mu_\theta(s)} + \int_{\mathcal{S}} \gamma p(s \to s', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s') \mathrm{d}s'.
\end{aligned}
$$

Where in (1) we used the Leibniz integral rule to exchange order of derivative and integration, requiring the regularity conditions, specifically continuity of $p(s'|s,a)$, $\mu_\theta(s)$, $V^{\mu_\theta}(s)$ and their derivatives w.r.t. θ. And now iterating this formula we have,

$$
\begin{aligned}
&= \nabla_\theta \mu_\theta(s)\ \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} \\
&\quad + \int_{\mathcal{S}} \gamma p(s \to s', 1, \mu_\theta) \nabla_\theta \mu_\theta(s')\ \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')}\ \mathrm{d}s' \\
&\quad + \int_{\mathcal{S}} \gamma p(s \to s', 1, \mu_\theta) \int_{\mathcal{S}} \gamma p(s' \to s'', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s'') \mathrm{d}s'' \mathrm{d}s' \\
&= \nabla_\theta \mu_\theta(s)\ \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} \\
&\quad + \int_{\mathcal{S}} \gamma p(s \to s', 1, \mu_\theta) \nabla_\theta \mu_\theta(s')\ \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')}\ \mathrm{d}s' \\
&\quad + \int_{\mathcal{S}} \gamma^2 p(s \to s', 2, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s') \mathrm{d}s' \qquad (2) \\
&\vdots \\
&= \int_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t p(s \to s', t, \mu_\theta) \nabla_\theta \mu_\theta(s')\ \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')}\ \mathrm{d}s'.
\end{aligned}
$$

Where in 2 we have used Fubini's theorem to exchange the order of integration, requiring the regularity conditions so that $||\nabla_\theta V^{\mu_\theta}(s)||$ is bounded. Now taking the expectation over S_1 we have,

$$
\begin{aligned}
\nabla_\theta J(\mu_\theta) &= \nabla_\theta \int_{\mathcal{S}} p_1(s) V^{\mu_\theta}(s) \mathrm{d}s \\
&= \int_{\mathcal{S}} p_1(s) \nabla_\theta V^{\mu_\theta}(s)\ \mathrm{d}s \qquad (3) \\
&= \int_{\mathcal{S}} \int_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t p_1(s) p(s \to s', t, \mu_\theta) \nabla_\theta \mu_\theta(s')\ \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')}\ \mathrm{d}s' \mathrm{d}s \\
&= \int_{\mathcal{S}} \rho^{\mu_\theta}(s) \nabla_\theta \mu_\theta(s)\ \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}\ \mathrm{d}s,
\end{aligned}
$$

where in (3) we used the Leibniz integral rule to exchange derivative and integral, requiring the regularity conditions, specifically so that $p_1(s)$ and $V^{\mu_\theta}(s)$ and derivatives w.r.t. θ are continuous. In the final line we again used Fubini's theorem to exchange the order of integration, requiring the boundedness of the integrand as implied by the regularity conditions. □

C. Proof of Theorem 2

We first restate Theorem 2 in detail, with discussion, and then prove the theorem. We first make a preliminary definition:

Conditions B1: Functions ν_σ parametrized by σ are said to be a *regular delta-approximation* on $\mathcal{R} \subseteq \mathcal{A}$ if they satisfy the following conditions:

1. The distributions ν_σ converge to a delta distribution: $\lim_{\sigma \downarrow 0} \int_{\mathcal{A}} \nu_\sigma(a', a) f(a) da = f(a')$ for $a' \in \mathcal{R}$ and suitably smooth f. Specifically we require that this convergence is uniform in a' and over any class $\mathcal{F}$ of L-Lipschitz and bounded functions, $||\nabla_a f(a)|| < L < \infty$, $\sup_a f(a) < b < \infty$, i.e.:

$$\lim_{\sigma \downarrow 0} \sup_{f \in \mathcal{F}, a' \in \mathcal{A}} \left| \int_{\mathcal{A}} \nu_\sigma(a', a) f(a) da - f(a') \right| = 0$$

2. For each $a' \in \mathcal{R}$, $\nu_\sigma(a', \cdot)$ is supported on some compact $\mathcal{C}_{a'} \subseteq \mathcal{A}$ with Lipschitz boundary $\mathrm{bd}(\mathcal{C}_{a'})$, vanishes on the boundary and is continuously differentiable on $\mathcal{C}_{a'}$.

3. For each $a' \in \mathcal{R}$, for each $a \in \mathcal{A}$, the gradient $\nabla_{a'} \nu_\sigma(a', a)$ exists.

4. Translation invariance: For all $a \in \mathcal{A}$, $a' \in \mathcal{R}$, and any $\delta \in \mathbb{R}^n$ such that $a + \delta \in \mathcal{A}$, $a' + \delta \in \mathcal{A}$, $\nu(a', a) = \nu(a' + \delta, a + \delta)$.

We restate the theorem:

Theorem. *Let $\mu_\theta : \mathcal{S} \to \mathcal{A}$. Denote the range of μ_θ by $\mathcal{R}_\theta := \mathrm{range}(\mu_\theta) \subseteq \mathcal{A}$, and $\mathcal{R} = \cup_\theta \mathcal{R}_\theta$. For each θ, Consider a stochastic policy $\pi_{\mu_\theta,\sigma}$ such that $\pi_{\mu_\theta,\sigma}(a|s) = \nu_\sigma(\mu_\theta(s), a)$, where ν_σ satisfy Conditions B1 on $\mathcal{R}$ above. Suppose further that the "regularity conditions" A.1 and A.2 (see Section A) on the MDP hold. Then,*

$$\lim_{\sigma \downarrow 0} \nabla_\theta J(\pi_{\mu_\theta,\sigma}) = \nabla_\theta J(\mu_\theta) \tag{4}$$

where on the l.h.s. the gradient is the standard stochastic policy gradient and on the r.h.s. the gradient is the deterministic policy gradient.

Theorem 2 holds for a very wide class of policies when $\mathcal{A} = \mathbb{R}^n$: any continuously differentiable, compactly supported $\xi : \mathbb{R}^n \to \mathbb{R}$ with total integral 1, can be used to construct $\nu_\sigma(a, a') = 1/\sigma^n \xi((a' - a)/\sigma)$ which satisfies our conditions, and the space of such functions is large: given any compact support such a function can be constructed. It is easy to check that any $\nu_\sigma(a, a')$ constructed on compact support with Lipschitz boundary in this way will satisfy Conditions B1. A simple example is any "bump function" such as, in 1 dimension, $\xi(a) = \begin{cases} e^{-\frac{1}{1-|a|^2}} & |a| < 1 \\ 0 & |a| \geq 1 \end{cases}$, or multidimensional versions.

We now prove the theorem. Throughout the proof we denote the time t marginal density at state s following policy π by $p_t^\pi(s)$. We begin with preliminary lemmas:

Lemma 1. *Let $\mathcal{U} \times \mathcal{V} \subseteq \mathbb{R}^n \times \mathbb{R}^n$. Let $\nu : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ be differentiable on $\mathcal{U} \times \mathcal{V}$. Then $(A) \Leftrightarrow (B) \Rightarrow (C)$ where,*

(A) Translation invariance: For all $u \in \mathcal{U}$, $v \in \mathcal{V}$, and any $\delta \in \mathbb{R}^n$ such that $u+\delta \in \mathcal{U}$, $v+\delta \in \mathcal{V}$, $\nu(u, v) = \nu(u+\delta, v+\delta)$.

(B) There exists some function $\chi : \mathbb{R}^n \to \mathbb{R}$ such that $\nu(u, v) = \chi(u - v)$.

(C) $\nabla_u \nu(u, v) = -\nabla_v \nu(u, v)$, wherever the gradients exist.

If furthermore $\mathcal{U} \times \mathcal{V}$ is convex then $C \Rightarrow A$, i.e. all properties are equivalent.

proof of Lemma 1. A $\Rightarrow$ B: For any $c \in \mathcal{U} - \mathcal{V}$ define $\chi : \mathbb{R}^n \to \mathbb{R}$ by $\chi : c \mapsto \nu(w, w - c)$ for any $w \in \mathcal{U}$ such that $c = w - v$ for some $v \in \mathcal{V}$. Observe that this defines χ uniquely on all of $\mathcal{U} - \mathcal{V}$. Thus given any $u \in \mathcal{U}$, $v \in \mathcal{V}$ we can choose $w = u$ and we have,

$$\begin{aligned} \chi(u - v) &= \nu(u, u - (u - v)) \\ &= \nu(u, v) \end{aligned}$$

B $\Rightarrow$ A: Trivial

B $\Rightarrow$ C: Let $h(u, v) = u - v$ then by the chain rule $\nabla_u \nu(u, v) = \nabla_h \chi(h)|_{h(u,v)} \nabla_u h(u, v) = \nabla_h \chi(h)|_{h(u,v)} = -\nabla_h \chi(h)|_{h(u,v)} \nabla_v h(u, v) = -\nabla_v \nu(u, v)$

(C and Convexity) $\Rightarrow$ A: Suppose $\mathcal{U} \times \mathcal{V}$ is convex. Consider any $(u, v) \in \mathcal{U} \times \mathcal{V}$, and any $\delta \in \mathbb{R}^n$, we have

$$\begin{aligned} \langle \nabla_{(u,v)} \nu(u, v), (\delta, \delta) \rangle &= \langle \nabla_u \nu(u, v), \delta \rangle + \langle \nabla_v \nu(u, v), \delta \rangle \\ &= \langle \nabla_u \nu(u, v), \delta \rangle - \langle \nabla_u \nu(u, v), \delta \rangle \\ &= 0 \end{aligned}$$

hence ν is constant in the direction (δ, δ). Since (u, v) and δ were arbitrary, ν is constant in the direction (δ, δ) for all $\delta \in \mathbb{R}^n$. Now since $\mathcal{U} \times \mathcal{V}$ is convex, for any $A = (u, v) \in \mathcal{U} \times \mathcal{V}$ and $B = (u + \delta, v + \delta) \in \mathcal{U} \times \mathcal{V}$ we have that the straight line connecting A and B is entirely contained $\mathcal{U} \times \mathcal{V}$. Thus, since ν is constant along the path $\nu(A) = \nu(B)$. □

We now note that the regularity conditions and properties of ν imply the following lemmas which we will need to prove Theorem 2.

Lemma 2. *1. For any stochastic policy π and any t, $\sup_s p_t^\pi(s) < b$ and similarly for deterministic policies.*

2. *For any stochastic policy π, $\sup_s \rho^\pi(s) < b/(1-\gamma)$ and similarly for deterministic policies.*

3. *for any stochastic policy π, $\sup_{a,s} \{||\nabla_a Q^\pi(a,s)||\} < c < \infty$ and similarly for deterministic policies.*

Proof. 1. The claim is true for $t = 1$ by the regularity conditions A.2, then for $t \geq 1$,

$$\sup_{s'} p^\pi_{t+1}(s') = \sup_{s'} \int p^\pi_t(s) \int \pi(a|s) p(s'|s,a) da ds \\ \leq \sup_{s',a,s} p(s'|s,a) < b$$

2. $\sup_s \rho^\pi(s) \leq \sum_{t=1}^\infty \gamma^{t-1} \sup_s p^\pi_t(s) \leq b/(1-\gamma)$

3. We have that,

$$\sup_{s,a} ||\nabla_a Q^\pi(a,s)|| \leq \sup_{s,a} ||\nabla_a r(s,a)|| + \gamma \sup_{s,a} \int ||\nabla_a p(s'|s,a)|| |V^\pi(s')| ds' \\ \leq L + \gamma \int Lb/(1-\gamma) ds' \\ < \infty$$

where the final line follows since $\mathcal{S}$ is compact and the integral over $\mathcal{S}$ is finite.

□

Lemma 3. $\lim_{\sigma \downarrow 0} \rho^{\pi_{\mu_\theta,\sigma}}(s) = \rho^{\pi_{\mu_\theta,0}}(s)$ *and the convergence is uniform w.r.t. s, i.e.*

$$\lim_{\sigma \downarrow 0} \sup_s |\rho^{\pi_{\mu_\theta,\sigma}}(s) - \rho^{\pi_{\mu_\theta,0}}(s)| = 0 \tag{5}$$

Proof. We have that $\rho^\pi(s) = \sum_{t=1}^\infty \gamma^{t-1} p^\pi_t(s)$. Clearly $p_1^{\pi_{\mu_\theta,\sigma}}(s) = p_1(s) = p_1^{\pi_{\mu_\theta,0}}(s)$. Note that by the definition of ν_σ, given any $\epsilon_1 > 0$ we can choose σ^* such that for all $\sigma < \sigma^*$,

$$\sup_s |\int \pi_{\mu_\theta,\sigma}(a|s) p(s'|s,a) da - \int \pi_{\mu_\theta,0}(a|s) p(s'|s,a) da| \leq \epsilon_1.$$

Now suppose (for induction) that for some $t \geq 1$ we have that

$$\sup_s |p_t^{\pi_{\mu_\theta,\sigma}}(s) - p_t^{\pi_{\mu_\theta,0}}(s)| \leq \epsilon_2(t),$$

then,

$$\sup_{s'} \left| p_{t+1}^{\pi_{\mu_\theta,\sigma}}(s') - p_{t+1}^{\pi_{\mu_\theta,0}}(s') \right| \leq \sup_{s'} \int |p_t^{\pi_{\mu_\theta,\sigma}}(s) - p_t^{\pi_{\mu_\theta,0}}(s)| \int \pi_{\mu_\theta,\sigma}(a|s) p(s'|s,a) da ds \\ + \sup_{s'} \int p_t^{\pi_{\mu_\theta,0}}(s) \left| \int \pi_{\mu_\theta,\sigma}(a|s) p(s'|s,a) da - \int \pi_{\mu_\theta,0}(a|s) p(s'|s,a) da \right| ds \\ \leq \epsilon_2(t) \int b ds + \epsilon_1 \\ = \epsilon_2(t) b\zeta + \epsilon_1,$$

where $\zeta = \int 1 ds < \infty$. Since $\epsilon_2(1) = 0$ we therefore have that

$$\sup_s |p_t^{\pi_{\mu_\theta,\sigma}}(s) - p_t^{\pi_{\mu_\theta,0}}(s)| \leq \epsilon_1 (b\zeta + 1)^{t-1},$$

And now given any $\epsilon > 0$ if we choose T sufficiently large such that, $\sum_{t=T+1}^{\infty} \gamma^{t-1} b < \epsilon/2$ and then we choose ϵ_1 and the corresponding σ^* sufficiently small so that, $\sum_{t=1}^{T} \gamma^{t-1} \epsilon_1 (b\zeta + 1)^{t-1} < \epsilon/2$, then we ensure that for any $\sigma < \sigma^*$,

$$\begin{aligned}
\sup_s |\rho^{\pi_{\mu_\theta,\sigma}}(s) - \rho^{\pi_{\mu_\theta,0}}(s)| &= \sup_s |\sum_{t=1}^{\infty} \gamma^{t-1} p_t^{\pi_{\mu_\theta,\sigma}}(s) - \sum_{t=1}^{\infty} \gamma^{t-1} p_t^{\pi_{\mu_\theta,0}}(s)| \\
&\leq \sum_{t=1}^{T} \gamma^{t-1} \sup_s |p_t^{\pi_{\mu_\theta,\sigma}}(s) - p_t^{\pi_{\mu_\theta,0}}(s)| \\
&+ \sum_{t=T+1}^{\infty} \gamma^{t-1} \sup_s |p_t^{\pi_{\mu_\theta,\sigma}}(s) - p_t^{\pi_{\mu_\theta,0}}(s)| \\
&\leq \sum_{t=1}^{T} \gamma^{t-1} \epsilon_1 (b\zeta + 1)^{t-1} + \sum_{t=1}^{\infty} \gamma^{t-1} b \\
&\leq \epsilon
\end{aligned}$$

as required. □

Lemma 4. *For all $s \in \mathcal{S}$, θ, the convergence $\nabla_a Q^{\pi_{\mu_\theta,\sigma}}(a, s) \to \nabla_a Q^{\pi_{\mu_\theta,0}}(a, s)$, as $\sigma \to 0$, is uniform in (s, a), i.e.*

$$\lim_{\sigma \downarrow 0} \sup_{(s,a)} ||\nabla_a Q^{\pi_{\mu_\theta,\sigma}}(a, s) - \nabla_a Q^{\pi_{\mu_\theta,0}}(a, s)|| = 0$$

Proof. $\nabla_a Q^{\pi}(a, s) = \nabla_a \left(r(s, a) + \gamma \int p(s'|s, a) V^{\pi}(s') ds' \right)$, so

$$\begin{aligned}
\sup_{(s,a)} ||\nabla_a Q^{\pi_{\mu_\theta,\sigma}}(a, s) - \nabla_a Q^{\pi_{\mu_\theta,0}}(a, s)|| &\leq \gamma \int \sup_{(s',s,a)} ||\nabla_a p(s'|s, a)|| |V^{\pi_{\mu_\theta,\sigma}}(s') - V^{\pi_{\mu_\theta,0}}(s')| ds' \\
&\leq \gamma \zeta L \sup_{s'} |V^{\pi_{\mu_\theta,\sigma}}(s') - V^{\pi_{\mu_\theta,0}}(s')|
\end{aligned}$$

where $\zeta = \int 1 ds < \infty$. Now, given any ϵ_1, ϵ_2 there exists σ^* such that for all $\sigma < \sigma^*$ we have that,

$$\sup_s | \int r(s, a) \left(\pi_{\mu_\theta,\sigma}(a|s) - \pi_{\mu_\theta,0}(a|s) \right) da| < \epsilon_1$$

and

$$\sup_{s,\hat{s}} |\rho_{\hat{s}}^{\pi_{\mu_\theta,\sigma}}(s) - \rho_{\hat{s}}^{\pi_{\mu_\theta,0}}(s)| < \epsilon_2 \tag{6}$$

where $\rho_{\hat{s}}^{\pi}(s)$ is analogous to $\rho^{\pi}(s)$, but conditioned on starting in distribution $\int p(s|a, \hat{s}) \pi(a|\hat{s}) da$ at $t = 1$ rather than in distribution p_1 (the result (6) result can be proved in an identical fashion to Lemma 3 noting that the result does not depend upon p_1 other than through its boundedness). Then,

$$\begin{aligned}
\sup_{s'} |V^{\pi_{\mu_\theta,\sigma}}(s') - V^{\pi_{\mu_\theta,0}}(s')| &\leq \sup_{s'} \left| \int r(s', a) (\pi_{\mu_\theta,\sigma}(a|s') - \pi_{\mu_\theta,0}(a|s')) da \right| \\
&+ \gamma \sup_{s'} \left| \int \int \rho_{s'}^{\pi_{\mu_\theta,\sigma}}(s) \pi_{\mu_\theta,\sigma}(a|s) r(s, a) da ds - \int \int \rho_{s'}^{\pi_{\mu_\theta,0}}(s) \pi_{\mu_\theta,0}(a|s) r(s, a) da ds \right| \\
&\leq \epsilon_1 + \sup_{s'} \int \int |\rho_{s'}^{\pi_{\mu_\theta,\sigma}}(s) - \rho_{s'}^{\pi_{\mu_\theta,\sigma}}(s)| |r(s, a)| \pi_{\mu_\theta,0}(a|s) da ds \\
&+ | \sup_{s'} \int \rho_{s'}^{\pi_{\mu_\theta,0}}(s) \int r(s, a) \left(\pi_{\mu_\theta,\sigma}(a|s) - \pi_{\mu_\theta,0}(a|s) \right) da ds| \\
&\leq \epsilon_1 + \epsilon_2 \zeta b + \epsilon_1 / (1 - \gamma)
\end{aligned}$$

which can thus be made arbitrarily small by choosing σ sufficiently small. □

proof of Theorem 2. Translation invariance, and Lemma 1 implies that $\nabla_{a'}\nu_\sigma(a',a)|_{a'=\mu_\theta(s)} = -\nabla_a\nu_\sigma(\mu_\theta(s),a)$. Then integration by parts implies that,

$$\begin{aligned}\int_\mathcal{A} Q^{\pi_{\mu_\theta,\sigma}}(s,a)\ \nabla_{a'}\nu_\sigma(a',a)|_{a'=\mu_\theta(s)}\,\mathrm{d}a &= -\int_\mathcal{A} Q^{\pi_{\mu_\theta,\sigma}}(s,a)\nabla_a\nu_\sigma(\mu_\theta(s),a)\mathrm{d}a \\ &= \int_{\mathcal{C}_{\mu_\theta(s)}} \nabla_a Q^{\pi_{\mu_\theta,\sigma}}(s,a)\nu_\sigma(\mu_\theta(s),a)\mathrm{d}a + \text{boundary terms} \\ &= \int_{\mathcal{C}_{\mu_\theta(s)}} \nabla_a Q^{\pi_{\mu_\theta,\sigma}}(s,a)\nu_\sigma(\mu_\theta(s),a)\mathrm{d}a\end{aligned}$$

Where the boundary terms are zero since ν_σ vanishes on the boundary. We have, from the stochastic policy gradient theorem,

$$\begin{aligned}\lim_{\sigma\downarrow 0}\nabla_\theta J(\pi_{\mu_\theta,\sigma}) &= \lim_{\sigma\downarrow 0}\int_\mathcal{S}\rho^{\pi_{\mu_\theta,\sigma}}(s)\int_\mathcal{A} Q^{\pi_{\mu_\theta,\sigma}}(s,a)\nabla_\theta\pi_{\mu_\theta,\sigma}(a|s)\ \mathrm{d}a\mathrm{d}s \\ &= \lim_{\sigma\downarrow 0}\int_\mathcal{S}\rho^{\pi_{\mu_\theta,\sigma}}(s)\int_\mathcal{A} Q^{\pi_{\mu_\theta,\sigma}}(s,a)\nabla_\theta\mu_\theta(s)\nabla_{a'}\ \nu_\sigma(a',a)|_{a'=\mu_\theta(s)}\,\mathrm{d}a\mathrm{d}s \\ &= \lim_{\sigma\downarrow 0}\int_\mathcal{S}\rho^{\pi_{\mu_\theta,\sigma}}(s)\nabla_\theta\mu_\theta(s)\int_{\mathcal{C}_{\mu_\theta(s)}}\nabla_a Q^{\pi_{\mu_\theta,\sigma}}(s,a)\nu_\sigma(\mu_\theta(s),a)\mathrm{d}a\mathrm{d}s \\ &= \int_\mathcal{S}\lim_{\sigma\downarrow 0}\rho^{\pi_{\mu_\theta,\sigma}}(s)\nabla_\theta\mu_\theta(s)\int_{\mathcal{C}_{\mu_\theta(s)}}\nabla_a Q^{\pi_{\mu_\theta,\sigma}}(s,a)\nu_\sigma(\mu_\theta(s),a)\mathrm{d}a\mathrm{d}s,\end{aligned} \tag{7}$$

where exchange of limit and integral in (7) follows by dominated convergence (in Banach spaces) where we can take the dominating function (which is bounded by Lemma 2),

$$\begin{aligned}g_\theta(s) &= \sup_\sigma\{\rho^{\pi_{\mu_\theta,\sigma}}(s)\}\sup_{a\in\mathcal{C}_{\mu_\theta(s)},\sigma}\{||\nabla_a Q^{\pi_{\mu_\theta,\sigma}}(a,s)||\}\,||\nabla_\theta\mu_\theta(s)||_{\mathrm{op}} \\ &\geq ||\rho^{\pi_{\mu_\theta,\sigma}}(s)\int_{\mathcal{C}_{\mu_\theta(s)}}\nabla_a Q^{\pi_{\mu_\theta,\sigma}}(s,a)\nu_\sigma(\mu_\theta(s),a)\mathrm{d}a\nabla_\theta\mu_\theta(s)||.\end{aligned} \tag{8}$$

Where $||\cdot||_{\mathrm{op}}$ denotes the operator norm, or largest singular value. Now note that by uniform convergence of $\nabla_a Q^{\pi_{\mu_\theta,\sigma}}(s,a)$, Lemma 4, given any ϵ_1, ϵ_2 there exists σ^* such that for all $\sigma<\sigma^*$ we have

$$||\nabla_a Q^{\pi_{\mu_\theta,\sigma}}(s,a)-\nabla_a Q^{\pi_{\mu_\theta,0}}(s,a)||<\epsilon_1$$

so that

$$||\int_{\mathcal{C}_{\mu_\theta(s)}}\nabla_a Q^{\pi_{\mu_\theta,\sigma}}(s,a)\nu_\sigma(\mu_\theta(s),a)\mathrm{d}a-\int_{\mathcal{C}_{\mu_\theta(s)}}\nabla_a Q^{\pi_{\mu_\theta,0}}(s,a)\nu_\sigma(\mu_\theta(s),a)\mathrm{d}a||<\epsilon_1,$$

and also that,

$$||\int_{\mathcal{C}_{\mu_\theta(s)}}\nabla_a Q^{\pi_{\mu_\theta,0}}(s,a)\nu_\sigma(\mu_\theta(s),a)\mathrm{d}a-\ \nabla_a Q^{\pi_{\mu_\theta,0}}(s,a)|_{a=\mu_\theta(s)}\,||<\epsilon_2.$$

Hence,

$$||\int_{\mathcal{C}_{\mu_\theta(s)}}\nabla_a Q^{\pi_{\mu_\theta,\sigma}}(s,a)\nu_\sigma(\mu_\theta(s),a)\mathrm{d}a-\ \nabla_a Q^{\pi_{\mu_\theta,0}}(s,a)|_{a=\mu_\theta(s)}\,||<\epsilon_1+\epsilon_2$$

and from this and Lemma 3 we have,

$$\begin{aligned}(7) &= \int_\mathcal{S}\rho^{\pi_{\mu_\theta,0}}(s)\nabla_\theta\mu_\theta(s)\lim_{\sigma\downarrow 0}\int_{\mathcal{C}_{\mu_\theta(s)}}\nabla_a Q^{\pi_{\mu_\theta,\sigma}}(s,a)\nu_\sigma(\mu_\theta(s),a)\mathrm{d}a\mathrm{d}s \\ &= \int_\mathcal{S}\rho^{\pi_{\mu_\theta,0}}(s)\nabla_\theta\mu_\theta(s)\ \nabla_a Q^{\pi_{\mu_\theta,0}}(s,a)|_{a=\mu_\theta(s)}\,\mathrm{d}s \\ &= \int_\mathcal{S}\rho^{\mu_\theta}(s)\nabla_\theta\mu_\theta(s)\ \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}\,\mathrm{d}s\end{aligned}$$

□