Global Optimiality Guarantees For Policy Gradient Methods: Technical Report with Supplementary Materials

Jalaj Bhandari and Daniel Russo

Columbia University

This technical report provide some further details related to our paper [Bhandari and Russo, 2019]. This helps separate our main contribution from either (i) slight adapatations of known results or (ii) technical details of specific examples that are somewhat orthogonal to the paper's main insights. The first part proves results related to the convergence of projected gradient descent in smooth (but possibly nonconvex) optimization. Similar results can be found in nonlinear optimization textbooks like Bertsekas [1997] or Beck [2017], but we include careful standalone proofs here. The second part provides some further technical results related to specific examples on which we instantiate our general results. Primarily, the effort here is on certifying that various functions are differentiable. We also extend our treatment of the optimal stopping example, providing stronger results.

The document should be read as a direct continuation of Bhandari and Russo [2019]. We refer to equations and citations from that paper throughout.

A Convergence proofs for first order methods.

To start, let us define some standard notions from first order optimization. For a convex set $\mathcal{X} \subset \mathbb{R}^d$, we say a function $f : \mathcal{X} \to \mathbb{R}$ is k-Lipshitz if $\|f(x) - f(y)\|_2 \leq k\|x - y\|_2$ for every $x, y \in \mathcal{X}$. We say a function is L-smooth if f is differentiable throughout $\mathcal{X}$ and ∇f is L-Lipschitz. A consequence of smoothness that will be useful throughout our proofs is often called the *descent lemma*. It implies a quadratic upper bound on function values. The proof follows by Taylor expansion and the mean-value theorem [Bertsekas, 1997].

Lemma 20 (Descent Lemma)**.** *If the function $f : \mathcal{D} \to \mathbb{R}$ is L-smooth over a set $\mathcal{X} \subseteq \mathcal{D}$, then for any $(x, y) \in \mathcal{X}$:*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2.$$

The following interpretation of projected gradient updates will be very useful for our proof. Recall the notation for orthogonal projection: $\text{Proj}_{\mathcal{X}}(x) = \arg\min_{y \in \mathcal{X}} \|y - x\|_2^2$. The projected gradient descent iteration can be equivalently written as

$$x_{t+1} = \text{Proj}_{\mathcal{X}}\left(x_t - \alpha_t \nabla f(x_t)\right) = \arg\min_{x \in \mathcal{X}} \left[f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\alpha_t}\|x - x_t\|_2^2 \right]. \tag{37}$$

giving a "proximal" interpretation of projection as minimizing a local quadratic approximation. See Beck [2017] for a simple proof.

A.1 Asymptotic convergence to stationary points: proof of Lemma 2

For convenience, we first restate the claim.

Lemma 2. *Consider the optimization problem* $\min_{x\in\mathcal{X}} f(x)$ *where* $\mathcal{X} \subset \mathbb{R}^d$ *is a closed convex set. Assume* f *is bounded below and its* β*–sublevel set* $\{x \in \mathcal{X} : f(x) \leq \beta\}$ *is bounded for each* $\beta \in \mathbb{R}$*. Consider the sequence* $x_{k+1} = \mathrm{Proj}_{\mathcal{X}}\left(x_k - \alpha\nabla f(x_k)\right)$ *for* $k \in \mathbb{N}$*.*

1. *[Beck, 2002, 2017] Assume* f *is differentiable on an open set containing* $\mathcal{X}$ *and* ∇f *is Lipschitz continuous on* $\mathcal{X}$ *with Lipschitz constant* L*. If* $\alpha \in (0, 1/L]$*, the sequence* $\{x_k\}$ *has at least one limit point and any limit point* x_∞ *is a stationary point of* $f(\cdot)$ *on* $\mathcal{X}$ *satisfying* $f(x_k) \downarrow f(x_\infty)$*.*

2. *Given a fixed initial iterate* x_0*, suppose* f *is continuously twice differentiable on an open set containing the sublevel set* $\{x \in \mathcal{X} : f(x) \leq f(x_0)\}$*. For a sufficiently small* $\alpha > 0$*, the sequence* $\{x_k\}$ *has at least one limit point and any limit point* x_∞ *is a stationary point of* $f(\cdot)$ *on* $\mathcal{X}$ *satisfying* $f(x_k) \downarrow f(x_\infty)$*.*

Proof. Part 1 follows from the simple proofs in [Beck, 2002, 2017]. We show the claim in part 2. Throughout, let $\|x\|$ denotes the Euclidean norm of a vector x and $\|A\| = \max_{\|x\|\leq 2} \|Ax\|$ be induced operator norm of a matrix A. Note that the sub-level set $S := \{x \in \mathcal{X} : f(x) \leq f(x_0)\}$ is compact (continuity of $f(\cdot)$ implies its closed and we assume it to be bounded). Also, for a sufficiently small ϵ, $f(\cdot)$ is twice continuously differentiable over the compact set,

$$S_\epsilon := \{x + y : x \in S_1, \|y\| \leq \epsilon\}.$$

which follows by our assumption that $f(\cdot)$ is twice continuously differentiable on an open set containing S. We denote $G = \max_{x\in S} \|\nabla f(x)\|$ and $L = \max_{x\in S_\epsilon} \|\nabla^2 f(x)\|$. Note that G and L are finite since $\|\nabla f\|$ and $\|\nabla^2 f\|$ are continuous over the compact sets S and S_ϵ. Fix the step-size $\alpha = \min\{\epsilon/G, 1/L\}$. For any $x \in S_1$, define $x^+ = \mathrm{Proj}_{\mathcal{X}}\left(x - \alpha\nabla f(x)\right)$. For this choice of step-size, $x^+ \in S_\epsilon$ since

$$\|x^+ - x\|_2 = \|\mathrm{Proj}_{\mathcal{X}}\left(x - \alpha\nabla f(x)\right) - \mathrm{Proj}_{\mathcal{X}}(x)\| \leq \|\alpha\nabla f(x)\| \leq \alpha G \leq \epsilon,$$

which follows as projection operators are non-expansive. The optimality conditions for projection onto a convex set yield the standard property that $\hat{x} = \mathrm{Proj}_{\mathcal{X}}\left(x\right)$ if and only if $\langle \hat{x} - x, y - \hat{x}\rangle \geq 0$ for all $y \in \mathcal{X}$. Using this and some algebra, we get

$$\langle x - \alpha\nabla f(x) - x^+, x - x^+\rangle \leq 0 \implies \|x - x^+\|^2 - \alpha\langle \nabla f(x), x - x^+\rangle \leq 0.$$

As $x^+ \in S_\epsilon$,

$$\begin{aligned}
f(x^+) &\leq f(x) + \langle \nabla f(x)\,,\, x^+ - x\rangle + \frac{L}{2}\|x^+ - x\|^2 && \text{[smoothness of } f(\cdot) \text{ over } S_\epsilon] \\
&\leq f(x) + \left(\frac{L}{2} - \frac{1}{\alpha}\right)\|x^+ - x\|^2 \\
&\leq f(x). && [\alpha \leq 1/L]
\end{aligned}$$

Since the projected gradient update reduces cost, we know $x^+ \in S$. Repeating this argument inductively shows that $f(x_{k+1}) \leq f(x_k)$ and $x_k \in S$ for all k. Since $\{x_k\}$ is contained in a compact set S, it has a convergent sub-sequence, $\{x_{k_i}\}$ with some limit x_∞. We have ,

$$\lim_{k\to\infty} f(x_k) = \lim_{i\to\infty} f(x_{k_i}) = f(x_\infty),$$

where the first limit exists since $\{f(x_k)\}$ is monotone-decreasing and bounded below and the final inequality uses continuity of $f(\cdot)$. The proof to show that any limit point is a stationary point follows from [Beck, 2002, 2017]. See also [Bertsekas, 1997, Figure 3.3.2]. We omit this for brevity. □

A.2 Convergence rates under gradient dominance: Proof of Lemma 3.

We first restate the claim.

Lemma 3 (Convergence rates for gradient dominated smooth functions)**.** *Consider the problem,* $\min_{x\in\mathcal{X}} f(x)$ *where* $\mathcal{X} \subseteq \mathbb{R}^d$ *is nonempty. Assume* ∇f *is L–Lipschitz continuous on* $\mathcal{X}$*. Denote* $f^* = \inf_{x'\in\mathcal{X}} f(x')$*. Consider the sequence* $x_{t+1} = \mathrm{Proj}_{\mathcal{X}}\left(x_t - \alpha\nabla f(x_t)\right)$*.*

1. *Let* $\mathcal{X} \subset \mathbb{R}^d$ *be bounded. Set* $R = \sup_{x,x'\in\mathcal{X}} \|x-x'\|_2$ *and* $k = \sup_{x\in\mathcal{X}} \|\nabla f(x)\|_2$*. If* $\alpha \le \min\{\frac{1}{k}, \frac{1}{L}\}$ *and* f *is* $(c,0)$*–gradient-dominated, then,*

$$f(x_T) - f^* \le \sqrt{\frac{2R^2 c\left(f(x_0) - f^*\right)}{\alpha T}}.$$

2. *Assume* $\mathcal{X} = \mathbb{R}^d$ *and* $\alpha = 1/L$*. If* f *is* (c,μ)*–gradient-dominated for* $\mu > 0$*, then,*

$$f(x_T) - f^* \le \left(1 - \frac{\mu}{c^2 L}\right)^T \left(f(x_0) - f^*\right).$$

Proof of Lemma 3. Recall, by Definition 2 that a function f is defined to be (c,μ)*–gradient dominated* over $\mathcal{X}$ if there exists a constant $c > 0$ and $\mu \ge 0$ such that

$$f(x^*) \ge f(x) + \min_{y\in\mathcal{X}} \left[c\left\langle \nabla f(x), y - x\right\rangle + \frac{\mu}{2}\|y-x\|_2^2\right] \qquad \forall\, x \in \mathcal{X}.$$

Proof of Part (a): We assume $\mu = 0$ in which case for any $x \in \mathcal{X}$, we have

$$\min_{y\in\mathcal{X}} \left[c\left\langle \nabla f(x), y - x\right\rangle\right] \le f(x^*) - f(x) \tag{38}$$

Therefore, for any $x \neq x^*$, we have $\min_{y\in\mathcal{D}} \left\langle \nabla f(x_t), y - x\right\rangle < 0$. Let $\{x_t\}$ be the iterates produced by projected gradient descent. At iterate x_t, let $\bar{y} = \arg\min_{y\in\mathcal{X}} \left\langle \nabla f(x_t), y - x_t\right\rangle$ and denote $\delta_t = \min_{y\in\mathcal{X}} \langle \nabla f(x_t), y - x_t\rangle$. Note that $\delta_t \le 0$ and $|\delta_t| \le \|\nabla f(x_t)\|\|y - x_t\| \le kR$ as f is assumed to be k-Lipschitz. We take a constant stepsize, $\alpha_t = \alpha \le \min\{\frac{1}{k}, \frac{1}{L}\}$. Then,

$$\begin{aligned}
f(x_{t+1}) - f(x_t) &\overset{(a)}{\le} \min_{y\in\mathcal{D}} \left[\left\langle \nabla f(x_t), y - x_t\right\rangle + \frac{1}{2\alpha}\|y - x_t\|_2^2\right] \\
&\overset{(b)}{=} \min_{\beta\in[0,1]} \left[\left\langle \nabla f(x_t), x_t + \beta(\bar{y} - x_t) - x_t\right\rangle + \frac{1}{2\alpha}\|x_t + \beta(\bar{y} - x_t) - x_t\|_2^2\right] \\
&= \min_{\beta\in[0,1]} \left[\beta\langle \nabla f(x_t), (\bar{y} - x_t)\rangle + \frac{\beta^2}{2\alpha}\|\bar{y} - x_t\|_2^2\right] \\
&\le \min_{\beta\in[0,1]} \left[\beta\delta_t + \frac{\beta^2 R^2}{2\alpha}\right] = \frac{-\alpha\delta_t^2}{2R^2}
\end{aligned} \tag{39}$$

where the minimizer $\beta^* = -\delta_t\alpha/R^2 \le k\alpha/R \le 1$ as $\alpha \le \min\{\frac{1}{k}, \frac{1}{L}\}$ (we assume $R > 1$ without loss of generality as we can take any upper bound while minimizing in (39)). Here (a) follows by using the equivalence shown in (37) and the quadratic upper bound on the function values implied by the descent lemma. Equality (b) uses the fact that right hand side of (a) can be optimized by searching over the steepest descent direction $x_t \to y$. Using (38), we get

$$f(x_{t+1}) - f(x_t) \le \frac{-\alpha}{2R^2c^2}\left(f(x^*) - f(x_t)\right)^2$$

Rearranging, we get our desired result

$$\begin{aligned}\min_{t\le T}\left(f(x_t)-f(x^*)\right)^2 \le \frac{1}{T}\sum_{t=0}^{T-1}\left(f(x_t)-f(x^*)\right)^2 &\le \frac{2R^2c^2}{\alpha T}\sum_{t=0}^{T-1} f(x_t)-f(x_{t+1})\\ &\le \frac{2R^2c^2}{\alpha T}\left(f(x_0)-f(x_T)\right)\\ &\le \frac{2R^2c^2}{\alpha T}\left(f(x_0)-f(x^*)\right)\end{aligned}$$

Since also $f(x_T)\le f(x_{T-1})\le\cdots f(x_1)$, we have

$$f(x_T)-f(x^*)\le \min_{t\le T}\left\{f(x_t)-f(x^*)\right\}\le\sqrt{\frac{2R^2c^2\left(f(x_0)-f(x^*)\right)}{\alpha T}}.$$

Proof of Part (b): We refer readers to the proof in Karimi et al. [2016], which can be dated back to Polyak [1963]. □

B Example Details

B.1 Stability in discounted LQ control: Proof of Lemma 15

We first recall the claim.

Lemma 15. *In the LQ control problem formulated in Example 2, $\ell(\theta)<\infty$ if and only if $\theta\in\Theta_S$.*

Proof. The average cost can be written as

$$\ell(\theta)=(1-\gamma)\mathbb{E}_\rho^{\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\left(s_t^\top\left[\theta R\theta+C\right]s_t\right)\right]$$

where the expectation is over $s_0\sim\rho$ and it is implicit that the state evolves according to the linear dynamical system $s_{t+1}=(A+B\theta)s_t$ for $t=0,1\cdots$. We can bound $\ell(\theta)$ above and below as,

$$\lambda_{\min}(\theta R\theta+C)\mathbb{E}_\rho^{\pi_\theta}\left[\sum_{t=0}^{\infty}\|\sqrt{\gamma}s_t\|_2^2\right]\le(1-\gamma)^{-1}\ell(\theta)\le\lambda_{\max}(\theta R\theta+C)\mathbb{E}_\rho^{\pi_\theta}\left[\sum_{t=0}^{\infty}\|\sqrt{\gamma}s_t\|_2^2\right].$$

Since $s_t=\left[A+B\theta\right]^t s_0$, applying Lemma 14 at each timestep with a choice of $M=(\sqrt{\lambda}\left[A+B\theta\right])^t$ yields

$$\lambda_{\min}\left(\Sigma_\rho\right)\sum_{t=0}^{\infty}\|(\sqrt{\gamma}\left[A+B\theta\right])^t\|_F^2\le\mathbb{E}_\rho^{\pi_\theta}\left[\sum_{t=0}^{\infty}\|\sqrt{\gamma}s_t\|_2^2\right]\le\lambda_{\max}\left(\Sigma_\rho\right)\sum_{t=0}^{\infty}\|(\sqrt{\gamma}\left[A+B\theta\right])^t\|_F^2.$$

We recall Gelfand's formula, which says that for any square matrix M with spectral radius $\rho(M)$ and any matrix norm $\|\cdot\|$, we have $\|M^t\|^{1/t}\to\rho(M)$ as $t\to\infty$. This implies $\sum_{t=0}^{\infty}\|(\sqrt{\gamma}\left[A+B\theta\right])^t\|_F^2$ is finite if and only if $\sqrt{\gamma}\left[A+B\theta\right]$ has spectral radius strictly less than one. □

B.2 Further results on optimal stopping

This section continues the analysis from the appendix of the main paper. It is intended t be read side-by-side with the notation and results there.

Condition 2.B: Gradient dominance. We now establish a gradient dominance result that strengthens the previous analysis. We first recall the claim.

Lemma 17 (Gradient dominance for optimal stopping)**.** *Consider the optimal stopping problem formulated in Example 5. For any $\pi \in \Pi_\Theta$, the function $\theta \mapsto \mathcal{B}(\theta|\eta_\pi, J_\pi)$ is $(\beta, 0)$–gradient-dominated where $\beta = \max_{x\in\mathcal{X},y\in\mathcal{Y}} q_x(y)/\min_{x\in\mathcal{X},y\in\mathcal{Y}} q_x(y)$.*

Proof. Fix any $\pi \in \Pi_\Theta$. As we formulate the optimal stopping as a reward maximization problem, following our notion of gradient dominance in Definition 2, we want to show that

$$\max_{\theta'\in\Theta} \langle \nabla_\theta \mathcal{B}(\theta|\eta_\pi, J_\pi), \theta' - \theta\rangle \geq \frac{1}{\beta}\left(\mathcal{B}(\theta^+|\eta_\pi, J_\pi) - \mathcal{B}(\theta|\eta_\pi, J_\pi)\right) \tag{40}$$

where θ^+ is the parameter of a policy iteration update to π, i.e. $\theta^+ = \arg\max_{\theta\in\Theta} \mathcal{B}(\theta|\eta_\pi, J_\pi)$. As shown in (34), $\theta_x^+ = \max\{y_{\min}, c_\pi(x)\}$. We first lower bound the left hand side of (40) as,

$$\begin{aligned}
\max_{\theta'\in\Theta} \langle \nabla_\theta \mathcal{B}(\theta|\eta_\pi, J_\pi), \theta' - \theta\rangle &= \sum_{x\in\mathcal{X}} \max_{\theta'_x\in\mathcal{Y}} \frac{\partial \mathcal{B}(\theta|\eta_\pi, J_\pi)}{\partial \theta_x} \cdot (\theta'_x - \theta_x) \\
&= \sum_{x\in\mathcal{X}} \max_{\theta'_x\in\mathcal{Y}} (c_\pi(x) - \theta_x) \cdot \eta'_\pi(x) q_x(\theta_x) \cdot (\theta'_x - \theta_x) \\
&\geq \left(\min_{x'\in\mathcal{X},y'\in\mathcal{Y}} q_{x'}(y')\right) \sum_{x\in\mathcal{X}} \eta'_\pi(x) \left\{\max_{\theta'_x\in\mathcal{Y}} \left[(c_\pi(x) - \theta_x) \cdot (\theta'_x - \theta_x)\right]\right\},
\end{aligned} \tag{41}$$

where second equality uses the derivative calculation in (35). We now upper bound the right hand side of (40):

$$\begin{aligned}
&\mathcal{B}(\theta^+|\eta_\pi, J_\pi) - \mathcal{B}(\theta|\eta_\pi, J_\pi) \\
&= \sum_{x\in\mathcal{X}} \eta'_\pi(x) \int_{y_{\min}}^{y_{\max}} \left[Q_\pi((x,y)\,,\, \mathbb{1}(y > \theta_x^+)) - Q_\pi((x,y)\,,\, \mathbb{1}(y > \theta_x))\right] q_x(y) dy \\
&\leq \left(\max_{x'\in\mathcal{X},y'\in\mathcal{Y}} q_{x'}(y')\right) \sum_{x\in\mathcal{X}} \eta'_\pi(x) \underbrace{\int_{y_{\min}}^{y_{\max}} \left[Q_\pi((x,y)\,,\, \mathbb{1}(y > \theta_x^+)) - Q_\pi((x,y)\,,\, \mathbb{1}(y > \theta_x))\right] dy}_{:= G_x(\theta_x^+, \theta_x)}.
\end{aligned}$$

The result follows if we can conclude,

$$\max_{\theta'_x\in\mathcal{Y}} \left[(c_\pi(x) - \theta_x) \cdot (\theta'_x - \theta_x)\right] \geq G_x(\theta_x^+, \theta_x), \tag{42}$$

which we do below by separately considering two cases. For this purpose, it is helpful to use a more explicit formula for $G_x(\theta_x^+, \theta_x)$, which follows directly from the formula for the Q function given in this section's preliminaries:

$$G_x(\theta_x^+, \theta_x) = \begin{cases} \int_{\theta_x}^{\theta_x^+} (c_\pi(x) - y) dy, & \text{for } \theta_x \leq \theta^+ \\ \int_{\theta_x^+}^{\theta_x} (y - c_\pi(x)) dy, & \text{for } \theta_x > \theta^+ \end{cases}$$

Case (1): Assume that $c_\pi(x) \in (y_{\min}, y_{\max})$ in which case $\theta_x^+ = c_\pi(x)$. Then,

$$G_x(\theta_x^+, \theta_x) = \frac{1}{2}(c_\pi(x) - \theta_x)^2 \leq (c_\pi(x) - \theta_x)^2 \leq \max_{\theta'_x\in\mathcal{Y}} (c_\pi(x) - \theta_x) \cdot (\theta'_x - \theta_x)$$

where the formula for $G_x(\theta_x^+, \theta_x)$ follows by calculating the integral and the final inequality uses that a choice of $\theta'_x = c_\pi(x)$ is feasible. The establishes (42).

Case (2): Now assume that $c_\pi(x) \notin (y_{\min}, y_{\max})$ in which case we know $c_\pi(x) \leq y_{\min}$ (since $c_\pi(x) \leq \gamma y_{\max}$) and therefore $\theta_x^+ = y_{\min}$. One can check that

$$G_x(\theta_x^+, \theta_x) = \int_{y_{\min}}^{\theta_x} (y - c_\pi(x))dy \leq (\theta_x - c_\pi(x))(\theta_x - y_{\min}) \leq \max_{\theta'_x \in \mathcal{Y}} (c_\pi(x) - \theta_x) \cdot (\theta'_x - \theta_x)$$

where again the final inequality is immediate since $\theta'_x = y_{\min}$ is a feasible choice. This establishes (42). □

Smoothness results: Condition 0 and Lemma 18 *Additional notation.* We simplify notation to write $\eta_\theta := \eta_{\pi_\theta}$, $\eta'_\theta := \eta'_{\pi_\theta}$, $T_\theta := T_{\pi_\theta}$, and $J_\theta := J_{\pi_\theta}$. We define $J'_\theta(x) := \int_{\mathcal{Y}} J_\theta(x, y) q_x(y) dy$ to be the expected cost-to-go function from context x. Similarly, denote $g'_\theta(x) := \int_{\mathcal{Y}} \mathbb{1}(y \geq \theta_x)\, y q_x(y) dy$ to be the expected reward earned from context x. Take $J'_\theta(\tau) = 0$ and $g'_\theta(\tau) = 0$. Recall that $\nu(x)$ is the probability the initial state (x_0, y_0) has $x_0 = x$. We extend ν to be a probability distribution over $\mathcal{X} \cup \{\tau\}$ by defining $\nu(\tau) = 0$.

We let $P'_\theta \in \mathbb{R}^{(|\mathcal{X}|+1)\times(|\mathcal{X}|+1)}$ denote the transition matrix over $\mathcal{X} \cup \{\tau\}$ under π_θ, defined as

$$P'_\theta(x'|x) = p(x'|x) \int_{\mathcal{Y}} \mathbb{1}(y < \theta_x) q_x(y) dy, \quad P'_\theta(\tau|x) = \int_{\mathcal{Y}} \mathbb{1}(y \geq \theta_x) q_x(y) dy, \quad P'_\theta(\tau|\tau) = 1, \tag{43}$$

for all $x', x \in \mathcal{X}$. One can write key quantities in "vector form" as

$$J'_\theta = (I - \gamma P'_\theta)^{-1} g'_\theta \qquad \& \qquad \eta'_\theta = (1-\gamma)\nu(I - \gamma P'_\theta)^{-1} \tag{44}$$

where $\nu \in \mathbb{R}^{|\mathcal{X}|+1}$ is a row vector $g'_\theta \in \mathbb{R}^{|\mathcal{X}|+1}$ is a column vector. The matrix $(I - \gamma P'_{\overline{\theta}})$ is invertible as $P'_{\overline{\theta}}$ is a stochastic matrix.

Verifying Condition 0. We establish the main condition needed to invoke the policy gradient formula in Lemma 6. That $\frac{\partial}{\partial \overline{\theta}_x} \mathcal{B}(\overline{\theta}|\eta_\theta, J_\theta)$ exists and is continuous in $\overline{\theta}$ is an immediate consequence of (35) together with the continuity of $q_x(\cdot)$. We also need to verify that $\mathcal{B}(\theta|\eta_{\overline{\theta}}, J_\theta)$ is continuously differentiable as a function of $\overline{\theta}$. Since $J'_\theta(\tau) = 0$,

$$\mathcal{B}(\theta|\eta_{\overline{\theta}}, J_\theta) = \sum_{x \in \mathcal{X}} \eta'_{\overline{\theta}}(x) \int_{\mathcal{Y}} Q_{\pi_\theta}((x, y), \pi_\theta(x, y)) q_x(y) dy = \sum_{x \in \mathcal{X}} \eta'_{\overline{\theta}}(x) J'_\theta(x).$$

Observe that $P'_{\overline{\theta}}$ is continuously differentiable in $\overline{\theta}$, as $q_x(\cdot)$ is assumed to be continuous. Therefore, $(I - \gamma P_{\overline{\theta}})^{-1}$ is also continuously differentiable and $\eta'_{\overline{\theta}}(\cdot)$ is as well.

Verifying Lemma 18. We now establish the following smoothness result for the policy gradient loss $\ell(\cdot)$, which is useful for invoking convergence rate results for first order methods.

Lemma 18. *For the optimal stopping problem in Example 5,* $\max_{\theta \in \Theta} \|\nabla^2 \ell(\theta)\| < \infty$.

Proof. Using the policy gradient theorem as shown in Lemma 6 and the derivative calculations in (35), we have

$$\frac{\partial}{\partial \theta_x} \ell(\theta) = \frac{\partial}{\partial \overline{\theta}_x} \mathcal{B}(\overline{\theta}|\eta_\theta, J_\theta)\Big|_{\overline{\theta}=\theta} = (c_{\pi_\theta}(x) - \theta_x)\, \eta'_\theta(x) q_x(\theta_x).$$

We argued above when verifying Condition 0 that $\eta'_\theta(x)$ is continuously differentiable. By assumption, $q_x(\theta_x)$ is continuously differentiable in θ_x. Therefore, $\nabla \ell(\theta)$ has a continuous derivative if $c_{\pi_\theta}(x)$ is continuously differentiable in θ. To show this, recall that by definition,

$$c_{\pi_\theta}(x) = \gamma \sum_{x' \in \mathcal{X}} p(x'|x) \int_{\mathcal{Y}} J_\theta(x', y') q_{x'}(y') dy = \gamma \sum_{x' \in \mathcal{X}} p(x'|x) J'_\theta(x') \qquad \forall x \in \mathcal{X}. \tag{45}$$

While verifying Condition 0 above, we argued that $(I - \gamma P'_\theta)^{-1}$ is is continuously differentiable. It is also easily shown that g'_θ is continuously differentiable. Due to (44), this implies J'_θ is continuously differentiable. Hence using (45), we find that c_{π_θ} is continuously differentiable in θ.

We have shown $\nabla^2 \ell(\theta)$ exists and is continuous. Since Θ is compact, the Extreme Value theorem implies $\max_{\theta \in \Theta} \|\nabla^2 \ell(\theta)\|_2$ exists and is finite. □

B.3 Finite horizon inventory control

We consider the inventory control problem as described in Example 6. To review notation, recall that in this setting, the inventory level evolves as: $x_{t+1} = x_t + a_t - w_t$ for non-negative orders a_t and i.i.d demands $w_t \in [0, w_{\max}]$. We let $s_t = (x_t, h_t)$ denote the state at time t and consider the class of base-stock policies which orders inventory, $\pi_\theta(s_t) = \max\{0, \theta_{h_t} - x_t\}$ in state s_t. We consider the policy class, $\Pi_\Theta = \{\pi_\theta : \theta \in \Theta\}$ with bounded parameter space $\Theta = [0, 2w_{\max}]^H$ which implies the inventory levels, x_t, at every period are bounded in $\mathcal{I} = [-w_{\max}, 2w_{\max}]$. We assume the initial distribution factorizes as, $\rho(dx, h) = \nu(h) q_h(x) dx_0$, where $\nu(h) > 0$ for every $h \in \{1, \ldots, H\}$ and $q_h(\cdot)$ is a twice differentiable PDF supported over $\mathcal{I}$ for each h. Clearly, this choice of ρ satisfies the regularity condition in Assumption 3. The problem setup also clearly follows the factorization structure assumed in Condition 3. Here, we verify the differentiability condition along with Condition 4, both of which are needed for Theorem 3.

Reminder on the Leibniz rule. We appeal to the Leibniz rule, which states sufficient conditions for differentiating the integral of a function. For a bounded set $\Psi \subset \mathbb{R}$, consider the function $F : \Psi \to \mathbb{R}$ given by

$$F(\psi) = \mathbb{E}[f(\psi, s)] = \int_{\mathcal{S}} f(\psi, s) P(ds)$$

where $f : \Psi \times \mathcal{S} \to \mathbb{R}$ is a real valued function, P is a probability measure supported over $\mathcal{S}$ and for each $\psi \in \Psi$, the function $f(\psi, \cdot)$ is P-integrable, i.e. $\mathbb{E}[|f(\psi, s)|] < \infty$. Let $\mathcal{D}_s(\psi)$ be the set of points $s \in \mathcal{S}$ such that $f(\cdot, s)$ is non-differentiable at ψ. By the Leibniz rule, F is differentiable at ψ if (i) $\mathcal{D}_s(\psi)$ has zero measure under P, i.e. $P[\mathcal{D}_s(\psi)] = 0$ and (ii) $f'(\psi, \cdot)$ is P-integrable, in which case $F'(\psi) = \int_{\mathcal{S}} f'(\psi, s) P(ds)$. By applying the same steps again, we can calculate the second derivative of $F(\cdot)$ as well.

The Leibniz rule is useful as a threshold policy, $\pi_\theta(z) = \max\{0, \theta - z\}$, is differentiable everywhere except at one point, $z = \theta$.

Condition 0: continuous differentiability. We make the abbreviation $\eta_\theta = \pi_\theta$, $J_\theta \equiv J_{\pi_\theta}$ and $Q_\theta \equiv Q_{\pi_\theta}$. To establish Condition 0, we need to show that the function $\mathcal{B}(\theta^+ | \eta_{\overline{\theta}}, J_\theta) = \int Q_\theta(s, \pi_{\theta^+}(s)) \eta_{\overline{\theta}}(ds)$, is (i) continuously differentiable as a function of θ^+ when $\overline{\theta} = \theta$ and (ii) continuously differentiable in $\overline{\theta}$ when $\theta^+ = \theta$. The approach is to (somewhat tediously) rewrite $\mathcal{B}(\cdot)$ in terms of a single multi-dimensional integral that appears in (47) below. From there, the result follows by applying the Leibniz rule to differentiate under the integral.

We write:

$$\begin{aligned}
\mathcal{B}(\theta^+ | \eta_{\overline{\theta}}, J_\theta) &= (1-\gamma) \mathbb{E}_\rho^{\pi_{\overline{\theta}}} \left[\sum_{t=0}^{\infty} \gamma^t Q_\theta(s_t, \pi_{\theta^+}(s_t)) \right] \\
&= (1-\gamma) \mathbb{E}_\rho^{\pi_{\overline{\theta}}} \left[\sum_{t=0}^{H-h_0} \gamma^t Q_\theta(s_t, \pi_{\theta^+}(s_t)) \right] \\
&= (1-\gamma) \mathbb{E}_\rho^{\pi_{\overline{\theta}}} \left[\sum_{t=0}^{H-h_0} \underbrace{\mathbb{E}_\rho^{\pi_{\overline{\theta}}} \left[\gamma^t Q_\theta(s_t, \pi_{\theta^+}(s_t)) \mid h_0 \right]}_{:= \mathcal{B}_{h_0, t}(\theta^+ | \eta_{\overline{\theta}}, J_\theta)} \right].
\end{aligned}$$

The first equality is the definition of the state occupancy measure. The second equality uses that the process transitions to a terminal state after stage $h_t = H + 1$ is reached. The final equality is uses the tower property of conditional expectation — there the outer expectation is over $h_0 \in \{1, \cdots, H\}$ drawn from the initial distribution and the inner expecation is over the state $s_t = (x_t, h_0 + t)$, and in particular the inventory level x_t, that is reached t periods later under the policy $\pi_{\overline{\theta}}$.

We show that each individual term, $\mathcal{B}_{h_0,t}(\theta^+|\eta_{\overline{\theta}},\, J_\theta)$, is continuously differentiable in $\overline{\theta}$ or in θ^+. This completes the result, since $\mathcal{B}(\theta^+|\eta_{\overline{\theta}}, J_\theta)$ a finite mixture of such terms. Notice that this expression involves three different policies that are applied at different timesteps:

1. $\pi_{\overline{\theta}}$ is applied prior to timestep t and it is these dynamics that determine s_t.
2. π_{θ^+} is applied during timestep t only. (That is the meaning of the action in the Q function.)
3. π_θ is applied from timestep $t+1$ until the final stage is reached. (That is the meaning of the subscript of the Q function.)

Based on this, we write $\tilde{\theta} = (\overline{\theta}_1, \cdots, \overline{\theta}_{h_t-1}, \theta^+_{h_t}, \theta_{h_t+1}, \cdots, \theta_H)$. Since we condition on h_0 and treat a fixed t throughout the remainder of the proof, we omit dependence on these quantities in notation. Then, one can write

$$\mathcal{B}_{h_0,t}(\theta^+|\eta_{\overline{\theta}},\, J_\theta) = \mathbb{E}_\rho^{\pi_{\tilde{\theta}}}\left[\sum_{\ell=t}^{H-h_0} \gamma^\ell g(s_t, a_t) \,\middle|\, h_0\right],$$

where we used the observations 1-3 above together with the definition of a Q-function as an expected disctounted sum of single-period costs $g(\cdot)$. In this inventory control problem,

$$g(s_t, a_t) = c \cdot a_t + \mathbb{E}[r(x_t + a_t - w_t) \mid x_t, a_t] = c \cdot a_t + \mathbb{E}[r(x_{t+1}) \mid x_t, a_t],$$

where $r : \mathbb{R} \to \mathbb{R}$ is defined as $r(z) = b\max(0, z) + p\max(0, -z)$ and denotes the holding/backlogging costs. The conditional expectation in the middle equation integrates over the i.i.d draw of the demand w_t and the second equation uses form of inventory dynamics $x_{t+1} = x_t + a_t - w_t$. By the tower property of conditional expectation, we can rewrite

$$\mathcal{B}_{h_0,t}(\theta^+|\eta_{\overline{\theta}},\, J_\theta) = \mathbb{E}_\rho^{\pi_{\tilde{\theta}}}\left[\sum_{\ell=t}^{H-h_0} \gamma^\ell \left(c \cdot a_t + r(x_{t+1})\right) \,\middle|\, h_0\right]. \tag{46}$$

Finally, we are ready to put this in the form of a single multi-dimensional integral. For any sequence of of inventory levels write $\vec{x} = (x_0, \cdots, x_{H-h_0+1})$. Set $G(\vec{x}, \tilde{\theta}) = \sum_{\ell=t}^{H-h_0} \gamma^\ell \left(c \cdot [\tilde{\theta}_{h_t} - x_t]^+ + r(x_{t+1})\right)$, representing the cost incurred for an arbitrary sequence of inventory levels $\vec{x}$ assuming the ordering costs are based on the values $[\tilde{\theta}_{h_t} - x_t]^+ = \max\{0, \tilde{\theta}_{h_t} - x_t\}$ prescribed by the base-stock levels $\tilde{\theta}$. Now, to assign probabilities over the sequence $\vec{x}$, recall that conditioned on h_0, the initial inventory level x_0 is drawn from a PDF $q_x(\cdot)$. Let $f(\cdot)$ denote the PDF of the demand distribution. Both PDFs are assumed to be twice differentiable. Then,

$$\mathcal{B}_{h_0,t}(\theta^+|\eta_{\overline{\theta}},\, J_\theta) = \int G(\vec{x}, \tilde{\theta}) p(\vec{x} \mid \tilde{\theta}) d\vec{x} \tag{47}$$

where

$$p(\vec{x} \mid \tilde{\theta}) = q_{h_0}(x_0) \prod_{t=0}^{H-h_0} f(x_{t+1} - x_t - [\tilde{\theta}_{h_t} - x_t]^+)$$

is the probability density function of $\vec{x}$. Here we have used that $x_{t+1} = x_t + a_t - w_t$ to relate the transitions of the inventory level to demand realizations.

Now, we can justify that $G(\vec{x}, \tilde{\theta})p(\vec{x}, \tilde{\theta})$ is almost surely twice differentiable with bounded first and second derivatives, so the Leibniz rule implies (47) is twice differentiable (and hence continuously differentiable).

Notice that $G(\vec{x}, \tilde{\theta})$ and $p(\vec{x}, \tilde{\theta})$ may fail to be differentiable at values of $\vec{x}$ where $x_t = \tilde{\theta}_{h_t}$ for some t, but the set of such sequences of inventory levels has zero Lebesgue measure. Boundedness of first and second derivatives of $G(\vec{x}, \tilde{\theta})$ is immediate since it involves piece-wise linear functions of $\tilde{\theta}$. Boundedness of first and second derivatives of $p(\vec{x} \mid \tilde{\theta})$ can be shown from the assumption that $f(\cdot)$ and $q_{h_0}(\cdot)$ are twice continuously differentiable and both $\vec{x}$ and $\tilde{\theta}$ take values in bounded sets. (See Example 6 where it is noted that $x_t \in [-w_{\max}, 2w_{\max}]$ almost surely.).

Remark 2. *It is also possible to complete the argument by claiming that the discounted sum inside* (46) *is differentiable* pathwise*, i.e for each realization of the random initial inventory level x_0 and the demand realizations (except on a set of measure zero). A benefit of this approach is that it does not require differentiability of $f(\cdot)$, whereas directly differentiating $p(\vec{x}|\tilde{\theta})$ in* (47) *does. See [Bertsekas, 1997, Section 2.6] for an introduction on how to calculate pathwise derivatives using the chain rule. Glasserman and Tayur [1995] applies this technique rigorously to an inventory control problem.*