*Adv. Appl. Prob.* **50,** 102–130 (2018)
*doi:*10.1017/apr.2018.6


# FORWARD SENSITIVITY ANALYSIS FOR CONTRACTING STOCHASTIC SYSTEMS

THOMAS FLYNN,* *City University of New York*

**Abstract**

In this paper we investigate gradient estimation for a class of contracting stochastic systems on a continuous state space. We find conditions on the one-step transitions, namely differentiability and contraction in a Wasserstein distance, that guarantee differentiability of stationary costs. Then we show how to estimate the derivatives, deriving an estimator that can be seen as a generalization of the forward sensitivity analysis method used in deterministic systems. We apply the results to examples, including a neural network model.

*Keywords:* Markov chain; derivative estimation; contraction; iterated function system

2010 Mathematics Subject Classification: Primary 60J05
Secondary 90C31; 65P99

## 1. Introduction

Stationary gradient estimation starts with a Markov kernel $P$ that depends on a parameter $\theta$. Given a cost function $e$ defined on the states of the Markov chain, and assuming ergodicity of the process, the problem is to estimate the derivative of the average cost, at stationarity, with respect to the parameter $\theta$. That is, setting $\pi_\theta$ to the stationary measure of $P_\theta$, the problem is to estimate

$$\frac{\partial}{\partial\theta}\int_X e(x)\,\mathrm{d}\pi_\theta(x).$$

In this paper we investigate an approach to this problem based on forward sensitivity analysis, an algorithm used for estimating sensitivities in deterministic systems. We review this now to show the main idea.

Consider a continuous state space $X \subseteq \mathbb{R}^{n_X}$ and a parameter space $\Theta \subseteq \mathbb{R}^{n_\Theta}$. Let $f\colon X \times \Theta \to X$ be such that $f(\cdot,\theta)$ is a contraction mapping on $X$ for all values of $\theta$. Then $f$ has a unique fixed point $x^*(\theta)$ for each $\theta \in \Theta$. With further conditions on the differentiability of $f$, it holds that $x^*$ is differentiable in $\Theta$. The problem is to estimate

$$\frac{\partial}{\partial\theta}(e \circ x^*)(\theta). \tag{1.1}$$

Let $M = L(\mathbb{R}^{n_\Theta}, \mathbb{R}^{n_X})$, the space of linear maps from $\mathbb{R}^{n_\Theta}$ to $\mathbb{R}^{n_X}$. Define the map $T\colon X \times M \times \Theta \to X \times M$ by

$$T((x,m),\theta) = \biggl(f(x,\theta), \frac{\partial f}{\partial x}(x,\theta)m + \frac{\partial f}{\partial\theta}(x,\theta)\biggr).$$

Received 31 October 2016; revision received 15 November 2017.
* Current address: Computational Science Initiative, Brookhaven National Laboratory, P.O. Box 5000, Upton, NY 11973, USA. Email address: tflynn@bnl.gov

Using assumptions on the derivatives and contraction properties of $f$, one can show that $T(\cdot, \theta)$ is also a contraction, for a suitable metric on $X \times M$. Denoting by $(x^*, m^*)$ the fixed point of $T$ at $\theta$, it can be proven that the derivative of the fixed-point cost is

$$\frac{\partial}{\partial \theta}(e \circ x^*)(\theta) = \frac{\partial e}{\partial x}(x^*)m^*.$$

Based on this, to approximately compute (1.1) we can iterate $T$ to obtain a pair $(x, m)$ near $(x^*, m^*)$, and then prepare the gradient estimate by computing $(\partial e/\partial x)(x)m$. For more background on forward sensitivity analysis, we refer the reader to [6, Chapter 15].

In this paper we consider the method in the probabilistic setting. Let $P_\theta$ take the form

$$(P_\theta e)(x) = \int_\Xi e(f(x, \xi, \theta))\, \mathrm{d}\nu(\xi)$$

for a probability space $(\Xi, \Sigma, \nu)$ and a function $f\colon X \times \Xi \times \Theta \to X$. We find that if certain contraction and differentiability conditions are satisfied, then

$$\frac{\partial}{\partial \theta} \int_X e(x)\, \mathrm{d}\pi_\theta(x) = \int_{X \times M} \frac{\partial e}{\partial x}(x)m\, \mathrm{d}\gamma_\theta(x, m), \tag{1.2}$$

where $\gamma_\theta$ is the stationary measure on $X \times M$ of the recursion

$$x_{n+1} = f(x_n, \xi_{n+1}, \theta), \qquad m_{n+1} = \frac{\partial f}{\partial x}(x_n, \xi_{n+1}, \theta)m_n + \frac{\partial f}{\partial \theta}(x_n, \xi_{n+1}, \theta), \tag{1.3}$$

where the $\xi_n$ form an independent and identically distributed sequence of $\nu$-distributed random variables. There are several challenges associated with this. The first is to extend the contraction framework to include probabilistically interesting systems. The contraction framework should enable us to show convergence of the forward sensitivity process (1.3) as well as the underlying process. The second challenge is to show correctness of the procedure.

A simple case of our main result can be stated as follows. In the statement of this theorem and throughout the paper, a function is said to be $C^1$ if it is continuously differentiable, and the function is $C^2$ if it is twice continuously differentiable. For a function $h$ defined on a set $X$ and taking values in a normed space, $\|h\|_\infty = \sup_{x \in X} \|h(x)\|$.

**Theorem 1.1.** *Let the function $f$ and the probability space $(\Xi, \Sigma, \nu)$ be such that*

(i) $\int_\Xi \|f(x, \xi, \theta)\|^2\, \mathrm{d}\nu(\xi) < \infty$ *for all* $(x, \theta) \in X \times \Theta$;

(ii) $(x, \theta) \mapsto f(x, \xi, \theta)$ *is a $C^2$ function for each* $\xi \in \Xi$;

(iii) *for* $0 < i + j \leq 2$, *the functions* $L_{X^i, \Theta^j}(x, \theta) = \int_\Xi \|(\partial^{i+j} f/\partial x^i \partial \theta^j)(x, \xi, \theta)\|^2\, \mathrm{d}\nu(\xi)$ *are continuous and bounded on* $X \times \Theta$, *and, in particular,* $\sup_{(x,\theta)} L_X(x, \theta) < 1$.

*Then the forward sensitivity process (1.3) converges weakly to a stationary measure $\gamma_\theta$, and (1.2) holds for those $e\colon X \to \mathbb{R}$ that are $C^2$ with $\|\partial e/\partial x\|_\infty + \|\partial^2 e/\partial x^2\|_\infty < \infty$.*

The full version, stated below in Theorem 1.2, relaxes the assumptions. In the general version the various bounds are assumed to hold with respect to a Finsler structure.

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

### 1.1. Overview of the main results

First, the contraction framework is introduced. Second, criteria for differentiability of the stationary costs are presented. The third component is a set of conditions on the function $f$ that allows us apply the abstract result on stationary differentiability, establish convergence of the sensitivity process $(x_n, m_n)$, and allow us to show that (1.2) holds. Finally, we consider an application to neural networks.

1.1.1. *Contraction framework.* Given a matrix-valued function $A(x)$ and a norm $\|\cdot\|$ on $\mathbb{R}^{n_X}$, we consider the following ergodicity condition:

$$\sup_{x\in X}\left(\int_{\Xi}\left\|A(f(x,\xi))\frac{\partial f}{\partial x}(x,\xi)A(x)^{-1}\right\|^{p}\mathrm{d}\nu(\xi)\right)^{1/p}<1. \tag{1.4}$$

The object inside the norm is the composition of the linear maps $A(f(x,\xi))$, $(\partial f/\partial x)(x,\xi)$, and $A(x)^{-1}$. The norm in this inequality is that induced by $\|\cdot\|$ on the space of linear maps $L(\mathbb{R}^{n_X},\mathbb{R}^{n_X})$. Formally, the map $(x,u)\mapsto\|A(x)u\|$ defines a Finsler structure on the space $X$, which induces a metric $d_A$ on $X$. This is extended to a metric on probability measures using the Wasserstein distance $d_{p,A}$. Condition (1.4) implies that the Markov kernel $P$ is a contraction mapping for this distance. This is developed in Section 2. In Section 2.1 we consider interconnections of contracting systems, obtaining sufficient conditions for both feedback and hierarchical combinations of contracting systems to again be contracting. This is useful to analyze the forward sensitivity process, as it exhibits a hierarchical structure.

1.1.2. *Stationary differentiability.* In Section 3 we provide abstract conditions for stationary differentiability, using a variant of the proof technique of [9]. The equation

$$l = lP_\theta + \pi_\theta\frac{\partial}{\partial\theta}P_\theta \tag{1.5}$$

is shown to have a unique solution in the variable $l$, and this $l$ is shown to evaluate the stationary derivatives, meaning $l(e)=(\partial/\partial\theta)\int_X e(x)\,\mathrm{d}\pi_\theta(x)$. While similar formulas have been recovered by other authors (see [9]–[13]), we rederive this using assumptions that are relevant for the smooth systems we are interested in.

1.1.3. *Gradient estimation.* To study the forward sensitivity process, we define an appropriate metric on the space $X\times M$ and prove a pointwise contraction inequality for the joint system (1.3) in this distance. This is used together with a Lyapunov function for the joint system to establish ergodicity of the sensitivity process. This is carried out in Section 5. We then establish that the functional $e\mapsto\int_{X\times M}(\partial e/\partial x)(x)m\,\mathrm{d}\gamma_\theta(x,m)$ verifies (1.5). We conclude that (1.2) holds for the class of cost functions.

Before formally stating the assumptions and main results, we introduce some notation and conventions. For a function $f\colon X\to\mathbb{R}^n$, where $X\subseteq\mathbb{R}^m$, we denote by $(\partial f/\partial x)(x_0)$ the derivative of $f$ with respect to $x$ at the point $x_0$, and for a vector $u\in\mathbb{R}^m$, we denote by $(\partial f/\partial x)(x_0)u$ the $\mathbb{R}^n$-valued result of applying this linear map to the vector $u$. The second derivative of $f$ with respect to $x$ is $\partial^2 f/\partial x^2$, and $(\partial^2 f/\partial x^2)(x_0)[u,v]$ refers to the $\mathbb{R}^m$-valued result of applying this bilinear map to the arguments $u, v$. Given norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ on the space $\mathbb{R}^m$ and $\mathbb{R}^n$, recall that the norm of a linear map $E\colon\mathbb{R}^n\to\mathbb{R}^m$ is $\|E\|=\sup_{\|u\|_X=1}\|Eu\|_Y$. For a bilinear map $F$ defined on $\mathbb{R}^n\times\mathbb{R}^m$ and taking values in a third space with norm $\|\cdot\|_Z$, the norm is $\|F\|=\sup_{\|u\|_X=\|v\|_Y=1}\|F[u,v]\|_Z$. Given two linear maps $E$ and $F$, their direct sum is the linear map $(E\oplus F)(u,v)=(Eu,Fv)$. For reference, Appendix A contains a summary of notations and definitions of spaces used throughout the paper.

**Assumption 1.1.** *The set $X$ is a closed, convex subset of $\mathbb{R}^{n_X}$, and $\mathbb{R}^{n_X}$ carries a norm $\|\cdot\|_X$. The function $A\colon X \to L(\mathbb{R}^{n_X}, \mathbb{R}^{n_X})$ is continuous, such that each $A(x)$ is invertible, and $\sup_{x \in X} \|A(x)^{-1}\|_X < \infty$.*

We will require differentiability and integrability of $f$.

**Assumption 1.2.** *For an open set $\Theta \subseteq \mathbb{R}^{n_\Theta}$, the function $f\colon X \times \Xi \times \Theta \to X$ satisfies*

(i) $\xi \mapsto d_A(x, f(x, \xi, \theta))^2$ *is $\nu$-integrable for all $(x, \theta) \in X \times \Theta$;*

(ii) $(x, \theta) \mapsto f(x, \xi, \theta)$ *is twice continuously differentiable ($C^2$) for each $\xi \in \Xi$.*

We also require some bounds on $P$ as a function of $\theta$, formulated with the help of a function $B(x)$ taking values in the invertible $n_\Theta \times n_\Theta$ matrices.

**Assumption 1.3.** *Assume that $\mathbb{R}^{n_\Theta}$ has a norm $\|\cdot\|_\Theta$. The function $B\colon X \to L(\mathbb{R}^{n_\Theta}, \mathbb{R}^{n_\Theta})$ takes values in the invertible linear maps, and $x \mapsto \|B(x)\|_\Theta$ is a $d_A$-Lipschitz function.*

For an example when Assumption 1.3 is satisfied, consider the following. Let $g\colon X \to \mathbb{R}_{\geq 0}$ be a function that is Lipschitz continuous with respect to the underlying norm $\|\cdot\|_X$ on $X$. Then use $A(x) = \exp(g(x)) I_{n_X}$ and $B(x) = \exp(g(x)) I_{n_\Theta}$, where $I_n$ is the $n \times n$ identity matrix. Of course, the assumption always holds when $B(x) = I_{n_\Theta}$.

The next assumptions relate to the contraction property of $P$ and the differentiability properties of $P_\theta e$. Before continuing, we define several norms derived from $A$ and $B$. At each $x \in X$, the matrix $A(x)$ defines a norm $\|\cdot\|_{A(x)}$ on $\mathbb{R}^{n_X}$ by $\|u\|_{A(x)} = \|A(x)u\|$ and $B(x)$ defines a norm on $\mathbb{R}^{n_\Theta}$ by $\|v\|_{B(x)} = \|B(x)v\|$. These extend to norms on the various linear spaces. For example, if $l \in L(\mathbb{R}^{n_X}, \mathbb{R})$ then $\|l\|_{A(x)} = \|lA(x)^{-1}\|$. For a bilinear map $Q \in L(\mathbb{R}^{n_X}, \mathbb{R}^{n_X}; \mathbb{R})$, we can write $\|Q\|_{A(x),A(x)} = \|Q(A(x)^{-1} \oplus A(x)^{-1})\|$. Further extend this to functions from $X$ into the linear spaces by taking supremums, for example, if $h\colon X \to L(\mathbb{R}^{n_\Theta}, \mathbb{R})$ then $\|h\|_B = \sup_x \|h(x)\|_{B(x)}$. For the case of a real-valued $h\colon X \to \mathbb{R}$, let $\|h\|_A = \sup_x |h(x)|/(1 + d_A(x, x_0))$, where $x_0$ is an arbitrary basepoint in $X$.

We introduce the space of cost functions $\mathcal{E}^2$:

$$\mathcal{E}^2 = \left\{ h\colon X \to \mathbb{R} \mid h \text{ is } C^2 \text{ and } \|h\|_A + \left\| \frac{\partial h}{\partial x} \right\|_A + \left\| \frac{\partial^2 h}{\partial x^2} \right\|_{A,A} < \infty \right\}.$$

On $\mathcal{E}^2$ we put the norm

$$\|h\|_{\mathcal{E}^2} = \|h\|_A + \left\| \frac{\partial h}{\partial x} \right\|_A + \left\| \frac{\partial^2 h}{\partial x^2} \right\|_{A,A}.$$

We consider bounds on the derivatives of $f$ formulated using the following functions:

$$L_X(x, \theta) = \left( \int_\Xi \left\| A(f(x, \xi, \theta)) \frac{\partial f}{\partial x}(x, \xi, \theta) A(x)^{-1} \right\|^2 \mathrm{d}\nu(\xi) \right)^{1/2},$$

$$L_\Theta(x, \theta) = \left( \int_\Xi \left\| A(f(x, \xi, \theta)) \frac{\partial f}{\partial \theta}(x, \xi, \theta) B(x)^{-1} \right\|^2 \mathrm{d}\nu(\xi) \right)^{1/2},$$

$$L_{X^2}(x, \theta) = \int_\Xi \left\| A(f(x, \xi, \theta)) \frac{\partial^2 f}{\partial x^2}(x, \xi, \theta)(A(x)^{-1} \oplus A(x)^{-1}) \right\| \mathrm{d}\nu(\xi),$$

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

$$L_{\Theta^2}(x,\theta) = \int_\Xi \left\| A(f(x,\xi,\theta)) \frac{\partial^2 f}{\partial\theta^2}(x,\xi,\theta)(B(x)^{-1} \oplus B(x)^{-1}) \right\| \mathrm{d}\nu(\xi),$$

$$L_{X,\Theta}(x,\theta) = \int_\Xi \left\| A(f(x,\xi,\theta)) \frac{\partial^2 f}{\partial x \partial\theta}(x,\xi,\theta)(A(x)^{-1} \oplus B(x)^{-1}) \right\| \mathrm{d}\nu(\xi).$$

**Assumption 1.4.** *The functions $L_{X^i,\Theta^j}$ satisfy the following:*

(i) *they are continuous on $X \times \Theta$;*

(ii) *there is a $K_X \in [0,1)$ such that $\sup_{(x,\theta)\in X\times\Theta} L_X(x,\theta) \le K_X$;*

(iii) *for $0 < i + j \le 2$, there are $K_{X^i,\Theta^j}$ such that $\sup_{(x,\theta)\in X\times\Theta} L_{X^i,\Theta^j}(x,\theta) \le K_{X^i,\Theta^j}$.*

Using these assumptions and definitions, we can now state the main result.

**Theorem 1.2.** *Let Assumptions 1.1–1.4 be satisfied. Let $\theta$ be an arbitrary point of $\Theta$. Then the forward sensitivity process (1.3) possesses a unique stationary measure $\gamma_\theta$ and, for any $e \in \mathcal{E}^2$, (1.2) is valid. Furthermore, if the variables $(x_1, m_1)$ satisfy the integrability condition $\mathbb{E}[d_A(x_0, x_1) + \|A(x_1)m_1\|] < \infty$ for an arbitrary basepoint $x_0$, then $\mathbb{E}[(\partial e/\partial x)(x_n)m_n] \to (\partial/\partial\theta)\int_X e(x)\,\mathrm{d}\pi_\theta(x)$ as $n \to \infty$.*

1.1.4. *Neural network application.* In Section 6 two examples are considered. The first involves neural networks. In neural networks, a central problem is to compute derivatives of cost functionals with respect to network parameters (weights on the connections between nodes). We are concerned with long-term average cost problems, a type of problem that is relevant when a network has cycles. The back-propagation algorithm for calculating derivatives [16], originally formulated for a continuous state-space model with a finite-horizon objective, is also valid for calculating gradients in long-term average cost problems under contraction assumptions [15]. Our contribution addresses the long-term average cost problem for continuous stochastic networks.

The example system consists of a network with weights on connections between units. At each step every node updates its value based on the values of its neighbors, but only a random subset of possible connections are activated, leading to a stochastic process. We find contraction conditions based on a sparsity coefficient, and verify that stochastic forward sensitivity analysis can be used to calculate the derivative of stationary costs. We present a second example to illustrate using a nontrivial metric on the underlying system. We finish with a discussion in Section 7.

## 2. Contraction framework

We describe a class of metrics on Euclidean space that form the basis for the subsequent discussion of contraction. These metrics are defined by minimizing a length functional, and form a subclass of the Finsler metrics. Then we present ergodicity conditions which rely on pointwise contraction estimates involving such metrics.

Let $X$ be a closed convex subset of the Euclidean space $\mathbb{R}^n$ and let $[x \rightsquigarrow y]$ be the set of piecewise $C^1$ curves from $x$ to $y$. Given a norm $\|\cdot\|$ on $\mathbb{R}^n$ and a function $x \mapsto A(x)$ taking values in the invertible $n \times n$ matrices, we can define a metric on $X$ as follows.

**Proposition 2.1.** *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ and let $x \mapsto A(x)$ be a continuous function that assigns to each $x \in X$ an invertible linear map $A(x)$ on $\mathbb{R}^n$, in such a way that $\sup_{x\in X}\|A(x)^{-1}\| < \infty$. For a piecewise $C^1$ curve $\gamma\colon [\gamma_s, \gamma_e] \to X$, define $L(\gamma) = \int_{\gamma_s}^{\gamma_e} \|A(\gamma(t))\gamma'(t)\|\,\mathrm{d}t$. Then the*

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

*function* $d_A(x, y) = \inf_{\gamma\in[x\rightsquigarrow y]} L(\gamma)$ *defines a metric on* $X$ *compatible with the Euclidean topology, and* $(X, d_A)$ *is complete.*

*Proof.* See Appendix A. □

For instance, taking $A = I_n$ we recover the norm $d_A(x, y) = \|x-y\|$. Using $A(x) = V(x)I_n$ for a real-valued function $V$ means a cost $V(x)$ is assigned for going through each point $x$. Using a general matrix allows the cost for traveling through each point $x$ to also depend on the direction of the path at the point. For a function $e\colon X \to \mathbb{R}$, we let $\|e\|_{\mathrm{Lip}(A)}$ be the Lipschitz constant of a function $e\colon X \to \mathbb{R}$ with respect to the metric $d_A$. When the metric $d_A$ is clear we will just write $\|e\|_{\mathrm{Lip}}$.

The collection of Borel probability measures on $X$ is denoted $\mathcal{P}(X)$. We denote by $\mu(e)$ the expectation of $e$ under $\mu$. That is, $\mu(e) = \int_X e(x)\,\mathrm{d}\mu(x)$. For a number $k$, we let $\mathbb{R}_{\geq k}$ be the set $\{x \in \mathbb{R} \mid x \geq k\}$. For a probability measure $\mu$ and $p \geq 1$, we write $\|V\|_{L^p(\mu)} = (\int_X \|V(x)\|^p\,\mathrm{d}\mu(x))^{1/p}$. Given a function $V\colon X \to \mathbb{R}_{\geq 0}$, the space $\mathcal{P}_{p,V}(X)$ is defined to be all Borel measures $\mu$ on $X$ which can integrate $V^p$:

$$\mathcal{P}_{p,V}(X) = \left\{\mu \in \mathcal{P}(X) \;\middle|\; \int_X V(x)^p\,\mathrm{d}\mu(x) < \infty\right\}.$$

Given a Markov kernel $P$, we denote the image of measure $\mu$ under $P$ by $\mu P$. That is, $(\mu P)(A) = \int_X P(x, A)\,\mathrm{d}\mu(x)$. For $V\colon X \to \mathbb{R}_{\geq 1}$, let $\|e\|_V = \sup_{x\in X} |e(x)|/V(x)$. We say that $V\colon X \to \mathbb{R}_{\geq 1}$ is a $p$-Lyapunov function for $P$ if $V$ has compact sublevel sets and there exist numbers $\beta \in [0, 1)$, $K \geq 0$ so that $(PV^p(x))^{1/p} \leq \beta V(x) + K$ for all $x$. A measure $\mu \in \mathcal{P}(X \times X)$ is a coupling of $\mu_1$ and $\mu_2$ if $\mu(A \times X) = \mu_1(A)$ and $\mu(X \times A) = \mu_2(A)$ for each measurable set $A$. We define $\Gamma(\mu_1, \mu_2)$ to be the set of all couplings of $\mu_1$ and $\mu_2$.

Let the Markov kernel $P$ have an explicit representation as

$$(Pe)(x) = \int_\Xi e(f(x, \xi))\,\mathrm{d}\nu(\xi) \tag{2.1}$$

for a measurable function $f\colon X \times \Xi \to X$ and a probability space $(\Xi, \Sigma, \nu)$. In this section we present two separate conditions for the ergodicity of a Markov kernel given in the form (2.1). The first, Proposition 2.3, is weaker and is used to show convergence of the forward sensitivity system (consisting of the variables $x_n, m_n$). Proposition 2.4 relies on a stronger set of assumptions and is used to establish differentiability of the stationary costs. Both results utilize the following pointwise estimate of Proposition 2.2.

In this proposition, and throughout the paper, we consider a differentiable function defined on a closed subset $X$ of Euclidean space. In case $X$ is a strict subset of the space, we assume $f$ is the restriction of a function $\overline{f}$ that is defined and differentiable on an open set $U$ containing $X$. In this way, there is no ambiguity in defining the derivative of $f$ at each point of $X$.

**Proposition 2.2.** *Let $P$ be of the form (2.1), where*

(i) $x \mapsto f(x, \xi)$ *is* $C^1$ *for each* $\xi \in \Xi$*;*

(ii) $\sup_{x\in X} \sup_{u\in\mathbb{R}^n\colon \|u\|=1} (\int_\Xi \|A(f(x, \xi))(\partial f/\partial x)(x, \xi)A^{-1}(x)u\|^p\,\mathrm{d}\nu(\xi))^{1/p} \leq \alpha$ *for some* $\alpha \geq 0$.

*Then, for any* $x_1, x_2 \in X$*, we have*

$$\left(\int_\Xi d_A(f(x_1, \xi), f(x_2, \xi))^p\,\mathrm{d}\nu(\xi)\right)^{1/p} \leq \alpha d_A(x_1, x_2). \tag{2.2}$$

*Proof.* Let $x_1 \neq x_2$ be points of $X$, let $\varepsilon > 0$, and let $\gamma\colon [0, T] \to X$ be a piecewise $C^1$ path from $x_1$ to $x_2$ such that $L(\gamma) \le d_A(x_1, x_2) + \varepsilon$. We further assume that $\gamma$ is parameterized by arc length. For our definition of length, this means $\|A(\gamma(t))\gamma'(t)\| = 1$ for all $t$ and that $T = L(\gamma)$. Since $t \mapsto f(\gamma(t), \xi)$ defines a curve from $f(x_1, \xi)$ to $f(x_2, \xi)$, we have

$$\begin{aligned}\Big(\int_\Xi d_A(f(x_1,\xi), f(x_2,\xi))^p \,\mathrm{d}\nu(\xi)\Big)^{1/p} &\le \Big(\int_\Xi \Big(\int_0^T \Big\| A(f(\gamma(t),\xi))\frac{\partial f}{\partial x}(x,\xi)\gamma'(t)\Big\| \,\mathrm{d}t\Big)^p \,\mathrm{d}\nu(\xi)\Big)^{1/p} \\ &\le L(\gamma)^{(p-1)/p}\Big(\int_\Xi \int_0^T \Big\| A(f(\gamma(t),\xi))\frac{\partial f}{\partial x}(x,\xi)\gamma'(t)\Big\|^p \,\mathrm{d}t\,\mathrm{d}\nu(\xi)\Big)^{1/p}.\end{aligned}$$

In the first step the definition of length was applied. Then Jensen's inequality was used together with the fact that $L(\gamma) = T$. Next, note the integrand in the final expectation is of the form $(t, \xi) \mapsto g(t, \xi)$, where $g$ is nonnegative, continuous in $t$ for each $\xi$, and measurable in $\xi$ for each $t$. Then we may interchange the integrals, yielding

$$\begin{aligned}\Big(\int_\Xi d_A(f(x_1,\xi), f(x_2,\xi))^p \,\mathrm{d}\nu(\xi)\Big)^{1/p} &= L(\gamma)^{(p-1)/p}\Big(\int_0^T \int_\Xi \Big\| A(f(\gamma(t),\xi))\frac{\partial f}{\partial x}(x,\xi)\gamma'(t)\Big\|^p \,\mathrm{d}\nu(\xi)\,\mathrm{d}t\Big)^{1/p}.\end{aligned}$$

Using the identity $A(\gamma(t))^{-1}A(\gamma(t))\gamma'(t) = \gamma'(t)$, and the assumption on $\partial f/\partial x$, we obtain

$$\Big(\int_\Xi d_A(f(x_1,\xi), f(x_2,\xi))^p \,\mathrm{d}\nu(\xi)\Big)^{1/p} \le L(\gamma)^{(p-1)/p}\Big(\int_0^T \alpha^p \|A(\gamma(t))\gamma'(t)\|^p \,\mathrm{d}t\Big)^{1/p}.$$

Then since $\gamma$ is parameterized by arc length,

$$\Big(\int_\Xi d_A(f(x_1,\xi), f(x_2,\xi))^p \,\mathrm{d}\nu(\xi)\Big)^{1/p} = L(\gamma)^{(p-1)/p}\alpha L(\gamma)^{1/p} \le \alpha d_A(x_1, x_2) + \alpha\varepsilon.$$

As $\varepsilon > 0$ was arbitrary, the result follows. □

If a tuple $\{(\Xi, \Sigma, \nu), f, (\|\cdot\|, A)\}$ satisfies the conditions of Proposition 2.2 for some $\alpha < 1$, we say that a *pointwise p-contraction inequality* holds for the process.

Combining this with the assumption that the system carries a Lyapunov function yields the following ergodicity result.

**Proposition 2.3.** *Let the assumptions of Proposition 2.2 hold for $p \ge 1$ and $\alpha < 1$, and assume there is a $p$-Lyapunov function $V$ for $P$. Then $P$ has a unique invariant measure $\pi \in \mathcal{P}_{p,V}(X)$ and, for any $\mu \in \mathcal{P}_{p,V}$, $\sup_{\|e\|_{\mathrm{Lip}}+\|e\|_V \le 1} |\mu P^n(e) - \pi(e)| \to 0$ as $n \to \infty$. In particular, $\mu P^n$ converges weakly to $\pi$.*

*Proof.* The existence of a unique invariant measure $\pi$ is an immediate result of Corollary 4.23 and Theorem 4.25 of [7]. To show that $\pi \in \mathcal{P}_{p,V}$, we reason as follows. If $V$ is a $p$-Lyapunov function then $V^p$ is a 1-Lyapunov function (for possibly different values of the constants $\beta$ and $K$). Then apply Proposition 4.24 of [7].

We turn to convergence of the expectations $\mu P^n(e)$ as $n \to \infty$. Let $e$ have $\|e\|_{\mathrm{Lip}} + \|e\|_V < \infty$. Using (2.2), we see that $\|Pe\|_{\mathrm{Lip}} \le \alpha \|e\|_{\mathrm{Lip}}$ and by iterating the inequality, we have

$$|P^n e(x) - P^n e(y)| \le \alpha^n \|e\|_{\mathrm{Lip}} d_A(x, y). \tag{2.3}$$

By iterating the Lyapunov inequality, we obtain

$$|P^n e(x) - P^n e(y)| \le \|e\|_V \beta^n [V(x) + V(y)] + \|e\|_V K', \tag{2.4}$$

where $K' = 2K/(1-\beta)$. Combining (2.3) and (2.4), for any coupling $\gamma$ of $\mu$ and $\pi$,

$$|\mu P^n(e) - \pi(e)| \le (\|e\|_{\mathrm{Lip}} + \|e\|_V) \int_{X\times X} \min\{\alpha^n d_A(x, y), \beta^n [V(x) + V(y)] + K'\} \, \mathrm{d}\gamma(x, y).$$

It remains to show that the right-hand side of this inequality tends to 0 as $n \to \infty$. Letting $f_n(x, y) = \min\{\alpha^n d_A(x, y), \beta^n [V(x) + V(y)] + K'\}$, it is clear the pointwise convergence of $f_n$ to 0 holds. Since also $|f_n| \le V(x) + V(y) + K'$, the latter function being $\gamma$-integrable, the result follows by the dominated convergence theorem. □

Let $x_0$ be an arbitrary basepoint in $X$. The next result strengthens the conclusion in the $V(x) = 1 + d_A(x_0, x)$ case, and concerns contraction in the Wasserstein space $\mathcal{P}_{p,A}$. This is the set of all measures that can integrate $x \mapsto d_A(x_0, x)^p$, together with the metric

$$d_{p,A}(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \left( \int_{X\times X} d_A(x, y)^p \, \mathrm{d}\gamma(x, y) \right)^{1/p}.$$

The space $\mathcal{P}_{p,A}$ is complete if $(X, d_A)$ is complete. Furthermore, the Kantorovich duality formula holds for $p = 1$:

$$\sup_{\|e\|_{\mathrm{Lip}} \le 1} |\mu_1(e) - \mu_2(e)| = d_{1,A}(\mu_1, \mu_2). \tag{2.5}$$

See [22] for more background.

**Proposition 2.4.** *Let the assumptions of Proposition 2.2 hold for some $p \ge 1$ and $\alpha < 1$. Let $V(x) = 1 + d_A(x, x_0)$ be a $p$-Lyapunov function for the kernel $P$. Then $P$ determines a contraction mapping on the Wasserstein space $\mathcal{P}_{p,A}(X)$ and possesses a unique invariant measure $\pi \in \mathcal{P}_{p,A}$. Furthermore, if $\mu \in \mathcal{P}_{p,V}$,*

$$\sup_{\|e\|_{\mathrm{Lip}} \le 1} |\mu P^n(e) - \pi(e)| \le \alpha^n \sup_{\|e\|_{\mathrm{Lip}} \le 1} |\mu(e) - \pi(e)|. \tag{2.6}$$

*Proof.* Let $\gamma$ be any coupling in $\Gamma(\mu_1, \mu_2)$. For any points $x, y$ of $X$, we can form a coupling of $\delta_x P$ and $\delta_y P$ using common random numbers. Formally, this is the measure $C(x, y)$ which arises as the pushforward of $\nu$ under the map $\xi \mapsto (f(x, \xi), f(y, \xi))$. Then $C$ is a well-defined Markov kernel on $X \times X$, and according to Proposition 2.2,

$$\left( \int_{X\times X} d_A(x', y')^p \, \mathrm{d}(\delta_{(x,y)} C)(x', y') \right)^{1/p} \le \alpha d_A(x, y).$$

Then

$$\begin{aligned} d_{p,A}(\mu_1 P, \mu_2 P) &\le \left( \int_{X\times X} d_A(x, y)^p \, \mathrm{d}(\gamma C)(x, y) \right)^{1/p} \\ &\le \alpha \left( \int_{X\times X} d_A(x, y)^p \, \mathrm{d}\gamma(x, y) \right)^{1/p}. \end{aligned}$$

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

Since $\gamma$ was arbitrary, it follows that $P$ is a contraction. Since $\mathcal{P}_{p,A}$ is complete, $P$ has a unique stationary measure $\pi$ in $\mathcal{P}_{p,A}$. Inequality (2.6) results by combining the contraction property with the duality formula (2.5). □

Conditions similar to those used in Proposition 2.2 have been mentioned in other works. Steinsaltz [19] considered the case of a scalar potential $A(x) = V(x)I$. The metric viewpoint for the scalar potential can be found in [8] and [20]. The results of [1] may be helpful to find scalar weight functions. The contraction conditions were also motivated by work on contraction analysis for deterministic systems [5], [12].

Aside from generality, there is a reason related to gradient estimation for considering matrix-valued functions $A$. Even if the underlying system has the unweighted average contraction property, meaning inequality (ii) of Proposition 2.2 holds with the function $A(x) = I$, this does not extend to the joint system (1.3). This is due to the factor $m$ in the auxiliary system of (1.3), which makes the Jacobian $\partial T/\partial z$ large at points $(x, m)$, where $\|m\|$ is large. One approach is to look beyond the scalar potentials to metrics that weigh the $x$ and $m$ directions differently. We will see in Section 5 that, for the case of unweighted contraction, a suitable metric involves a matrix $H(x, m)(u_x, u_m) = ((1 + h(x, m))u_x, u_m)$ for a scalar function $h(x, m)$.

### 2.1. Interconnections of contractions

In this section we provide conditions for the interconnection of two contracting systems to again be contracting. It is relevant to gradient estimation since the system (1.3) has a hierarchical form, the underlying system $x$ feeding into the system $m$. Interconnection theorems for contracting systems hold in other dynamical settings as well; results for deterministic continuous-time systems can be found in [17] and [18].

Let $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}^m$ be closed, convex sets, and let $Z = X \times Y$. For instance, when these results are applied later to the forward sensitivity process, the space $Y$ will be $L(\mathbb{R}^{n_X}, \mathbb{R}^{n_\Theta})$. Let $(\Xi, \Sigma, \nu)$ be a probability space and let $R$ be the Markov kernel that corresponds to following stochastic recursion on $Z$:

$$x_{n+1} = f(x_n, y_n, \xi_{n+1}), \qquad y_{n+1} = g(x_n, y_n, \xi_{n+1}),$$

where the $\xi_n$ are independent $\nu$-distributed random variables. For measurable $\phi\colon Z \to \mathbb{R}$, we have $(R\phi)(x, y) = \int_\Xi \phi(T(x, y, \xi))\, d\nu(\xi)$, where $T(x, y, \xi) = (f(x, y, \xi), g(x, y, \xi))$. We find conditions on $f$ and $g$ that guarantee the joint system is contracting.

**Assumption 2.1.** *Regarding the functions $f$, $g$ and the probability space $(\Xi, \Sigma, \nu)$,*

(i) *the maps $(x, y) \mapsto f(x, y, \xi)$ and $(x, y) \mapsto g(x, y, \xi)$ are $C^1$ for each $\xi \in \Xi$;*

(ii) *there are pairs $(\|\cdot\|_X, F)$, $(\|\cdot\|_Y, G)$, such that $\|\cdot\|_X$, $\|\cdot\|_Y$ are norms on $\mathbb{R}^n$, $\mathbb{R}^m$, respectively, $F\colon X \times Y \to \mathbb{R}^{n\times n}$ and $G\colon X \times Y \to \mathbb{R}^{m\times m}$ are continuous with values in the invertible matrices, and $\sup_{(x,y)\in X\times Y} \|F(x, y)^{-1}\|_X + \|G(x, y)^{-1}\|_Y < \infty$;*

(iii) *there are $\alpha_1$ and $\alpha_2$, both in $[0, 1)$, such that*

$$\sup_{z\in Z}\ \sup_{u\in\mathbb{R}^n:\ \|u\|_X=1} \left(\int_\Xi \left\| F(T(z, \xi))\frac{\partial f}{\partial x}(z, \xi)F^{-1}(z)u\right\|_X^p d\nu(\xi)\right)^{1/p} \le \alpha_1,$$

$$\sup_{z\in Z}\ \sup_{u\in\mathbb{R}^m:\ \|u\|_Y=1} \left(\int_\Xi \left\| G(T(z, \xi))\frac{\partial g}{\partial y}(z, \xi)G^{-1}(z)u\right\|_Y^p d\nu(\xi)\right)^{1/p} \le \alpha_2.$$

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

We are concerned with pointwise contraction as in Proposition 2.2. With further integrability assumptions, convergence to a unique stationary measure can be obtained with the results of the previous section.

**Proposition 2.5.** *Let Assumption 2.1 hold. Let* $K_1$, $K_2$, *and* $p \geq 1$ *be such that*

(i) $\sup_{z\in Z} \sup_{\|u_y\|_Y=1} (\int_\Xi \|F(T(z,\xi))(\partial f/\partial y)(z,\xi)G(z)^{-1}u_y\|_Y^p \, \mathrm{d}\nu(\xi))^{1/p} \leq K_1$;

(ii) $\sup_{z\in Z} \sup_{\|u_x\|_X=1} (\int_\Xi \|G(T(z,\xi))(\partial g/\partial x)(z,\xi)F(z)^{-1}u_x\|_X^p \, \mathrm{d}\nu(\xi))^{1/p} \leq K_2$;

(iii) $K_1K_2 < (1-\alpha_1)(1-\alpha_2)$.

*Choose* $\eta_1, \eta_2$ *so that* $\eta_2 K_2 < \eta_1(1-\alpha_1)$ *and* $\eta_1 K_1 < \eta_2(1-\alpha_2)$. *Then a pointwise p-contraction inequality holds for the system* $\{(\Xi, \Sigma, \nu), T, (\|\cdot\|_Z, H)\}$ *on Z, where*

$$H(z)(u_x, u_y) = (F(z)u_x,\ G(z)u_y), \qquad \|(u_x,u_y)\|_Z = \eta_1\|u_x\|_X + \eta_2\|u_y\|_Y. \tag{2.7}$$

*Proof.* We will apply Proposition 2.2. We need to find an $\alpha < 1$ so that

$$\sup_{z\in Z} \sup_{u\in\mathbb{R}^n\times\mathbb{R}^m:\ \|u\|_Z=1} \left(\int_\Xi \left\| H(T(z,\xi))\frac{\partial T}{\partial z}(z,\xi)H(z)^{-1}u\right\|_Z^p \mathrm{d}\nu(\xi)\right)^{1/p} \leq \alpha.$$

Let $z \in Z$ and let $u = (u_x, u_y)$ be any vector with $\eta_1\|u_x\|_X + \eta_2\|u_y\|_Y = 1$. Then

$$\begin{aligned}
&\left(\int_\Xi \left\| H(T(z,\xi))\frac{\partial T}{\partial z}(z,\xi)H(z)^{-1}u\right\|_Z^p \mathrm{d}\nu(\xi)\right)^{1/p} \\
&= \Bigg(\int_\Xi \Bigg[\eta_1 \left\| F(T(z,\xi))\frac{\partial f}{\partial x}(z,\xi)F(z)^{-1}u_x + F(T(z,\xi))\frac{\partial f}{\partial y}(z,\xi)G(z)^{-1}u_y\right\|_X \\
&\quad + \eta_2 \left\| G(T(z,\xi))\frac{\partial g}{\partial x}(x,\xi)F(z)^{-1}u_x + G(T(z,\xi))\frac{\partial g}{\partial y}(x,\xi)G(z)^{-1}u_y\right\|_Y\Bigg]^p \mathrm{d}\nu(\xi)\Bigg)^{1/p} \\
&\leq \eta_1\alpha_1\|u_x\|_X + \eta_1K_1\|u_y\|_Y + \eta_2K_2\|u_x\|_X + \eta_2\alpha_2\|u_y\|_Y \\
&\leq \max\left\{\alpha_1 + \frac{\eta_2}{\eta_1}K_2, \alpha_2 + \frac{\eta_1}{\eta_2}K_1\right\}.
\end{aligned}$$

Finally, note that satisfiability of the condition $\max\{\alpha_1 + (\eta_2/\eta_1)K_2, \alpha_2 + (\eta_1/\eta_2)K_2\} < 1$ is equivalent to the condition $K_1K_2 < (1-\alpha_1)(1-\alpha_2)$. □

The above can be specialized to hierarchical interconnections.

**Corollary 2.1.** *Let Assumption 2.1 hold. Say that* $f$ *does not depend on* $Y$ $(\partial f/\partial y = 0)$. *Let* $K$ *be such that*

$$\sup_{z\in Z} \sup_{\|u_x\|_X=1} \left(\int_\Xi \left\| G(T(z,\xi))\frac{\partial g}{\partial x}(z,\xi)F(z)^{-1}u_x\right\|_Y^p \mathrm{d}\nu(\xi)\right)^{1/p} \leq K. \tag{2.8}$$

*Choose* $\eta_1, \eta_2$ *so that* $\eta_2 K < \eta_1(1-\alpha_1)$. *Then a pointwise p-contraction property holds for the system* $\{(\Xi, \Sigma, \nu), T, (\|\cdot\|_Z, H)\}$ *on Z using the H and* $\|\cdot\|_Z$ *of (2.7).*

Condition (2.8) in Corollary 2.1 can be relaxed using a kind of Lyapunov function for the interconnection of the two systems, while requiring a stronger form of contraction on the input system.

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

**Proposition 2.6.** *Let Assumption 2.1 hold, with $p \geq 2q$ for some $q \geq 1$. Let $K$ and the continuous function $h\colon Z \to \mathbb{R}_{\geq 0}$ be such that, for all $z \in Z$,*

(i) $\sup_{\|u_x\|_X=1}(\int_\Xi \|G(T(z,\xi))(\partial g/\partial x)(z,\xi)F^{-1}(z)u_x\|_X^q \,\mathrm{d}\nu(\xi))^{1/q} \leq h(z)$;

(ii) $(\int_\Xi h(T(z,\xi))^p \,\mathrm{d}\nu(\xi))^{1/p} \leq h(z) + K$.

*Then there are some $\eta_1, \eta_2$ so that a pointwise $q$-contraction inequality holds for the system $\{(\Xi, \Sigma, \nu), T, (\|\cdot\|_Z, H)\}$ on $Z$, where*

$$H(z)(u_x, u_y) = ((1 + \eta_1 h(z))F(z)u_x,\ G(z)u_y), \qquad \|(u_x, u_y)\|_Z = \|u_x\|_X + \eta_2\|u_y\|_Y.$$

*Proof.* Let $\alpha_1, \alpha_2$ be contraction coefficients for $f, g$, respectively. Let $F_1(z) = [1 + \eta_3 h(z)]F(z)$, using an $\eta_3 \geq 0$ such that $\alpha_1(1 + \eta_3 K) < 1$. We aim to apply Corollary 2.1 to the pair of systems $f$ and $g$, using a metric defined by the pairs $(\|\cdot\|_X, F_1)$ and $(\|\cdot\|_Y, G)$, in order to find $q$-contraction of the joint system. Letting $\|u_x\|_X = 1$, then,

$$\begin{aligned}&\left(\int_\Xi \left\| F_1(T(z,\xi))\frac{\partial f}{\partial x}(x,\xi)F_1(z)^{-1}u_x\right\|_X^q \,\mathrm{d}\nu(\xi)\right)^{1/q}\\ &\quad = \left(\int_\Xi \left\| \frac{1 + \eta_3 h(T(z,\xi))}{1 + \eta_3 h(z)} F(T(z,\xi))\frac{\partial f}{\partial x}(x,\xi)F(z)^{-1}u_x\right\|_X^q \,\mathrm{d}\nu(\xi)\right)^{1/q}.\end{aligned}$$

Applying Hölder's inequality, and the assumption on $\partial f/\partial x$, yields

$$\begin{aligned}&\left(\int_\Xi \left\| F_1(T(z,\xi))\frac{\partial f}{\partial x}(x,\xi)F_1(z)^{-1}u_x\right\|_X^q \,\mathrm{d}\nu(\xi)\right)^{1/q}\\ &\quad \leq \frac{1}{1 + \eta_3 h(z)}\left(1 + \eta_3\left(\int_\Xi h(T(z,\xi))^{2q}\,\mathrm{d}\nu(\xi)\right)^{1/(2q)}\right)\alpha_1\\ &\quad \leq \frac{1 + \eta_3(h(z) + K)}{1 + \eta_3 h(z)}\alpha_1\\ &\quad \leq \alpha_1(1 + \eta_3)K.\end{aligned}$$

It remains to show that (2.8) holds. Let $\|u_x\|_X = 1$. Then

$$\begin{aligned}&\left(\int_\Xi \left\| G(T(z,\xi))\frac{\partial g}{\partial x}(z,\xi)F_1(z)^{-1}u_x\right\|_Y^q \,\mathrm{d}\nu(\xi)\right)^{1/q}\\ &\quad = \frac{1}{1 + \eta_3 h(z)}\left(\int_\Xi \left\| G(T(z,\xi))\frac{\partial g}{\partial x}(z,\xi_2)F(x)^{-1}u_x\right\|_Y^q \,\mathrm{d}\nu(\xi)\right)^{1/q}\\ &\quad \leq \frac{h(z)}{1 + \eta_3 h(z)}\\ &\quad \leq \frac{1}{\eta_3}.\end{aligned}$$

Let $\eta_1, \eta_2$ be chosen so that $\eta_2(1/\eta_3) < \eta_1(1 - \alpha_1(1 + \eta_3)K)$. Then, by Corollary 2.1, the tuple $\{(\Xi, \Sigma, \nu), T, (\|\cdot\|_Z, H)\}$ determines a $q$-contracting system, where $\|(u, v)\|_Z = \eta_1\|u\| + \eta_2\|v\|$ and $H(z)(u_x, u_y) = ((1 + \eta_3 h(z))F(z)u_x, G(z)u_y)$. One can take $\eta_1 = 1$ in these requirements, by choosing $\eta_2$ small enough that $\eta_2(1/\eta_3) < (1 - \alpha_1(1 + \eta_3)K)$. □

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

## 3. Stationary differentiability

Differentiability of stationary costs is established using properties of the Markov kernel $P$. In the next section, the assumptions are verified based on properties of the derivatives of the system.

Formally differentiating the equation $\pi_\theta = \pi_\theta P_\theta$ in $\theta$ suggests the stationary derivative $\pi'$ solves the equation $l = lP_\theta + \pi_\theta P'_\theta$ in the variable $l$. By defining $P'$ properly, as the linear map $e \mapsto (\partial/\partial\theta)P_\theta e$ on the space of cost functions, and considering this equation as being between functionals defined on the cost functions, we can show that it has a unique solution $l^*$, which is such that $l^*(e) = (\partial/\partial\theta)\int_X e(x)\,\mathrm{d}\pi_\theta(x)$. The line of argument used in this section is a variant of Theorem 2 of [9], adapted to the specific ergodicity and state space conditions that we work with. In that work, a class of functions with a norm $\|e\| = \sup_x |e(x)|/V(x)$ was considered, while the norm we will use also involves the derivatives of $e$. In Heidergott and Hordijk [9], an important role is played by the *deviation operator* $D_\theta$ (see Section 3 of that work) in that, in their setting, $D_\theta$ maps $\mathcal{E}^2$ back into itself. Dealing directly with the deviation operator in our case requires care since the space of functions will have more subtle topological properties due to the terms involving derivatives. We leave a possible unification of these two approaches to future work.

We introduce the assumptions on $P$ and the cost functions $\mathcal{E}$.

**Assumption 3.1.** *Denote by $X$ a Polish space, $\mathcal{E}$ a vector space of real-valued functions on $X$ with norm $\|\cdot\|_{\mathcal{E}}$, and $\mathcal{P}$ a space of probability measures on $X$. For any $\mu \in \mathcal{P}$, it is required that* $\sup_{\|e\|_{\mathcal{E}}\le 1} |\mu(e)| < \infty$.

Denote by $\Pi_\theta$ the Markov kernel $\Pi_\theta(x, A) = \pi_\theta(A)$. The parameter space is an open set $\Theta \subseteq \mathbb{R}^{n_\Theta}$ and we fix a $\theta_0 \in \Theta$. The space $\mathbb{R}^{n_\Theta}$ has a norm $\|\cdot\|_\Theta$. We show that the map sending a cost function $e$ to its stationary derivative at the fixed parameter $\theta_0$ is an element of the set $\mathcal{L}$ of linear maps from $\mathcal{E}$ to $L(\mathbb{R}^{n_\Theta}, \mathbb{R})$ that vanish on the constant functions and are bounded with respect to the norm $\|l\|_{\mathcal{L}} = \sup_{\|e\|_{\mathcal{E}}\le 1} \|l(e)\|_\Theta$:

$$\mathcal{L} = \{l \in L(\mathcal{E}, L(\mathbb{R}^{n_\Theta}, \mathbb{R})) \mid \|l\|_{\mathcal{L}} < \infty,\ l(\mathbf{1}) = 0\},$$

where $\mathbf{1}$ refers to the constant function $x \mapsto 1$. Note that $\mathcal{L}$ is a complete space.

To discuss stationary differentiability, we introduce the operator $(\partial/\partial\theta)P_{\theta_0}$. If $e \in \mathcal{E}$ then $(\partial/\partial\theta)P_{\theta_0}e$ is the function from $X$ into $L(\mathbb{R}^{n_\theta}, \mathbb{R})$ defined by

$$\left(\frac{\partial}{\partial\theta}P_{\theta_0}e\right)(x) = \frac{\partial}{\partial\theta}\left(P_{\theta_0}e(x)\right).$$

**Assumption 3.2.** *For any $\theta \in \Theta$ the following hold:*

(i) *if $\mu \in \mathcal{P}$ then $\mu P_\theta \in \mathcal{P}$ and $P_\theta$ has a stationary measure $\pi_\theta$ in $\mathcal{P}$;*

(ii) *if $e \in \mathcal{E}$ then $P_\theta e \in \mathcal{E}$, $\|P_\theta\|_{\mathcal{E}} < \infty$, and $\sum_{i=0}^{\infty} \|P^i_{\theta_0} - \Pi_{\theta_0}\|_{\mathcal{E}} \le K_{\theta_0}$ for some $K_{\theta_0} \ge 0$;*

(iii) *for $e \in \mathcal{E}$ and $x \in X$, the function $\theta \mapsto P_\theta e(x)$ is differentiable at $\|\pi_{\theta_0}(\partial/\partial\theta)P_{\theta_0}\|_{\mathcal{L}} < \infty$ and $\theta_0$;*

(iv) $(1/\|\Delta\theta\|_\Theta)\|\pi_{\theta_0}[P_{\theta_0+\Delta\theta} - P_{\theta_0} - (\partial/\partial\theta)P_{\theta_0}(\Delta\theta)]\|_{\mathcal{E}} \to 0$ *as* $\|\Delta\theta\|_\Theta \to 0$*;*

(v) $(1/\|\Delta\theta\|_\Theta)\|(\pi_{\theta_0+\Delta\theta} - \pi_{\theta_0})[P_{\theta_0+\Delta\Theta} - P_{\theta_0}]\|_{\mathcal{E}} \to 0$ *as* $\|\Delta\theta\|_\Theta \to 0$.

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

In part (iv), the functional $\pi_{\theta_0}[P_{\theta_0+\Delta\theta} - P_{\theta_0} - (\partial/\partial\theta)P_{\theta_0}(\Delta\theta)]$ maps a function $e \in \mathcal{E}$ to the number $\pi_{\theta_0}P_{\theta+\Delta\theta}(e) - \pi_{\theta_0}P_{\theta_0}(e) - \pi_{\theta_0}((\partial/\partial\theta)P_{\theta_0}e(\Delta\theta))$.

The main theorem on stationary differentiability now follows.

**Theorem 3.1.** *Under Assumptions 3.1 and 3.2, if $e \in \mathcal{E}^2$ then $\pi_\theta(e)$ is differentiable at $\theta_0$ and $(\partial/\partial\theta)\int_X e(x)\,\mathrm{d}\pi_{\theta_0}(x) = l^*(e)$, where $l^* \in \mathcal{L}$ satisfies $l^* = l^* P_{\theta_0} + \pi_{\theta_0}(\partial/\partial\theta)P_{\theta_0}$.*

*Proof.* First, define $T\colon \mathcal{L} \to \mathcal{L}$ as $T(l) := lP_{\theta_0} + \pi_{\theta_0}(\partial/\partial\theta)P_{\theta_0}$. The fact that $\pi_{\theta_0}(\partial/\partial\theta)P_{\theta_0}$ is in $\mathcal{L}$ was one of our assumptions along with $\|P_\theta\|_{\mathcal{E}} < \infty$, which implies that $T$ is well defined. Let $l^*$ be the functional $l^* = \sum_{i=0}^{\infty}(\pi_{\theta_0}(\partial/\partial\theta)P_{\theta_0})P_{\theta_0}^i$. This is in $\mathcal{L}$ since the space is Banach and, by Assumption 3.2(ii),

$$\sum_{i=0}^{\infty}\left\|\left(\pi_{\theta_0}\frac{\partial}{\partial\theta}P_{\theta_0}\right)P_{\theta_0}^i\right\|_{\mathcal{L}} = \sum_{i=0}^{\infty}\left\|\left(\pi_{\theta_0}\frac{\partial}{\partial\theta}P_{\theta_0}\right)(P_{\theta_0}^i - \Pi_{\theta_0})\right\|_{\mathcal{L}} \le \left\|\pi_{\theta_0}\frac{\partial}{\partial\theta}P_{\theta_0}\right\|_{\mathcal{L}} K.$$

To see that $l^*$ is a fixed point of $T$, note that

$$T(l^*) = \sum_{i=1}^{\infty}\left(\pi_{\theta_0}\frac{\partial}{\partial\theta}P_{\theta_0}\right)P_{\theta_0}^i + \pi_{\theta_0}\frac{\partial}{\partial\theta}P_{\theta_0} = l^*.$$

To show that $l^*$ is the unique fixed point, let $l$ be any other fixed point of $T$. Then

$$\|l - l^*\|_{\mathcal{L}} = \|T^n(l) - T^n(l^*)\|_{\mathcal{L}} = \|(l - l^*)(P_{\theta_0}^n - \Pi_{\theta_0})\|_{\mathcal{L}} \le \|l - l^*\|_{\mathcal{L}}\|P_{\theta_0}^n - \Pi_{\theta_0}\|_{\mathcal{E}}.$$

Using Assumption 3.2(ii) again, the right-hand side of this inequality goes to 0 as $n \to \infty$, hence, $T$ possesses a unique fixed point $l^*$ in $\mathcal{L}$.

Define $c(\Delta\theta)$ as the functional $c(\Delta\theta)(e) = \pi_{\theta_0+\Delta\theta}(e) - \pi_{\theta_0}(e) - l^*(e)(\Delta\theta)$. Assumption 3.1 and the definition of $\mathcal{L}$ guarantee that $c(\Delta\theta) \in L(\mathcal{E}, \mathbb{R})$. It suffices to show that $(1/\|\Delta\theta\|_\Theta)\|c(\Delta\theta)\|_{\mathcal{E}} \to 0$ as $\Delta\theta \to 0$. Using the fact that $T(l^*) = l^*$, we have

$$c(\Delta\theta) = \pi_{\theta_0}\left[P_{\theta_0+\Delta\theta} - P_{\theta_0} - \frac{\partial}{\partial\theta}P_{\theta_0}(\Delta\theta)\right] + (\pi_{\theta_0+\Delta\theta} - \pi_{\theta_0})[P_{\theta_0+\Delta\theta} - P_{\theta_0}] + c(\Delta\theta)P_{\theta_0}.$$

Iterating this, and noting that each summand is a functional vanishing on the constant functions, we obtain, for any $k > 0$,

$$\begin{aligned} c(\Delta\theta) = {}& \pi_{\theta_0}\left(P_{\theta_0+\Delta\theta} - P_{\theta_0} - \frac{\partial}{\partial\theta}P_{\theta_0}(\Delta\theta)\right)\sum_{i=0}^{k-1}(P_{\theta_0}^i - \Pi_{\theta_0}) \\ & + (\pi_{\theta_0+\Delta\theta} - \pi_{\theta_0})[P_{\theta_0+\Delta\theta} - P_{\theta_0}]\sum_{i=0}^{k-1}(P_{\theta_0}^i - \Pi_{\theta_0}) + c(\Delta\theta)(P_{\theta_0}^k - \Pi_{\theta_0}). \end{aligned}$$

Taking norms and letting $k \to \infty$, we see that

$$\begin{aligned} \|c(\Delta\theta)\|_{\mathcal{E}} \le {}& \left\|\pi_\theta\left(P_{\theta_0+\Delta\theta} - P_{\theta_0} - \frac{\partial}{\partial\theta}P_{\theta_0}(\Delta\theta)\right)\right\|_{\mathcal{E}} K_{\theta_0} \\ & + \|(\pi_{\theta_0+\Delta\theta} - \pi_{\theta_0})[P_{\theta_0+\Delta\theta} - P_{\theta_0}]\|_{\mathcal{E}} K_{\theta_0}. \end{aligned}$$

Finally, use parts (iv) and (v) of Assumption 3.2. □

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

## 4. State space conditions

Let $P_\theta$ be the transition kernel of the Markov chain

$$x_{n+1} = f(x_n, \xi_{n+1}, \theta) \tag{4.1}$$

with a $\nu$-distributed random input $\xi_n$. In this section we show how Assumptions 1.1–1.4 imply Assumptions 3.1 and 3.2, thereby establishing differentiability of the stationary costs for those cost functions $e \in \mathcal{E}^2$.

**Theorem 4.1.** *Let Assumptions 1.1–1.4 be satisfied. Then Assumptions 3.1 and 3.2 are verified for the space $\mathcal{P}_{2,A}(X)$ of probability measures and the space of cost functions $\mathcal{E}^2$, at any $\theta_0 \in \Theta$. Hence, $\pi_{\theta_0}(e)$ is differentiable for any $\theta_0 \in \Theta$ and $e \in \mathcal{E}^2$.*

To show this, several preliminary results will be used. The first is concerned with how $P_\theta$ varies with $\theta$. Recall that $x_0$ denotes an arbitrary basepoint.

**Proposition 4.1.** *Let $P_\theta$ be the transition kernel of the recursion (4.1), where*

- (i) *the map $\xi \mapsto d_A(x_0, f(x, \xi, \theta))^p$ is $\nu$-integrable for each $(x, \theta) \in X \times \Theta$;*
- (ii) *the function $(x, \theta) \mapsto f(x, \xi, \theta)$ is $C^1$ for each $\xi \in \Xi$;*
- (iii) *and*

$$\sup_{(x,\theta)\in X\times\Theta} \sup_{\|u_\theta\|=1} \bigg(\int_\Xi \bigg\| A(f(x,\xi,\theta))\frac{\partial f}{\partial\theta}(x,\xi,\theta)B(x)^{-1}u_\theta\bigg\|^p \,\mathrm{d}\nu(\xi)\bigg)^{1/p} \le K.$$

*Fix a $\theta_0 \in \Theta$. Then, for all $\Delta\theta$ sufficiently small and all $\mu \in \mathcal{P}_{p,A}(X)$, the inequality $d_{p,A}(\mu P_{\theta_0}, \mu P_{\theta_0+\Delta\theta}) \le K\|B\Delta\theta\|_{L^p(\mu)}$ holds.*

*Proof.* Let $\Delta\theta$ be so small that $\theta_0 + t\Delta\theta \in \Theta$ for $t \in [0, 1]$. If $(x, \xi)$ is distributed according to $\mu \times \nu$ then the law of $(f(x, \xi, \theta_0), f(x, \xi, \theta_0 + \Delta\theta))$ is a coupling of $\mu P_{\theta_0}$ and $\mu P_{\theta_0+\Delta\theta}$. Let $\gamma\colon [0, 1] \to \mathbb{R}^{n_\Theta}$ be $\gamma(t) = \theta_0 + t\Delta\theta$. Then $t \mapsto f(x, \xi, \gamma(t))$ determines a curve from $f(x, \xi, \theta_0)$ to $f(x, \xi, \theta_0 + \Delta\theta)$, and reasoning as in Proposition 2.2,

$$\begin{aligned}
&\bigg(\int_X\int_\Xi d_A(f(x,\xi,\theta_0), f(x,\xi,\theta_0+\Delta\theta))^p \,\mathrm{d}\nu(\xi)\,\mathrm{d}\mu(x)\bigg)^{1/p}\\
&\quad\le \bigg(\int_X\int_\Xi\bigg(\int_0^1 \bigg\|A(f(x,\xi,\gamma(t))\frac{\partial f}{\partial\theta}(x,\xi,\gamma(t))\Delta\theta\bigg\|\,\mathrm{d}t\bigg)^p \,\mathrm{d}\nu(\xi)\,\mathrm{d}\mu(x)\bigg)^{1/p}\\
&\quad\le \bigg(\int_0^1\int_X\int_\Xi \bigg\|A(f(x,\xi,\gamma(t)))\frac{\partial f}{\partial\theta}(x,\xi,\gamma(t))\Delta\theta\bigg\|^p \,\mathrm{d}\nu(\xi)\,\mathrm{d}\mu(x)\,\mathrm{d}t\bigg)^{1/p}\\
&\quad\le \bigg(\int_0^1\int_X K^p\|B(x)\Delta\theta\|^p \,\mathrm{d}\mu(x)\,\mathrm{d}t\bigg)^{1/p}\\
&\quad= K\|B\Delta\theta\|_{L^p(\mu)}.
\end{aligned}$$

□

The continuity assumptions on the $L_{X^i,\Theta^j}$ ensure that integration and differentiation can be exchanged. In the discussion of differentiability, it will be useful to introduce the following concept. A function $f\colon X \times \Xi \to \mathbb{R}^n$ is said to be $L^1(\nu)$-*continuous* when

- (i) $x \mapsto f(x, \xi)$ is continuous for each $\xi \in \Xi$;

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

(ii) $\xi \mapsto f(x, \xi)$ is measurable for each $x \in X$;

(iii) $x \mapsto \int_{\Xi} \|f(x, \xi)\| \, \mathrm{d}\nu(\xi)$ is continuous.

The following two properties are not difficult to show.

(i) If $f$, $g$ are $L^1(\nu)$-continuous functions then so are $\alpha f + \beta g$ for any numbers $\alpha, \beta$.

(ii) A monotonicity property holds: if $f$ is a function satisfying the first two requirements of $L^1(\nu)$-continuity and if $\|f(x, \xi)\| \leq \|g(x, \xi)\|$ for an $L^1(\nu)$-continuous function $g$, then $f$ is $L^1(\nu)$-continuous.

Using this notion we state a condition for interchanging derivatives and integrals, which is a generalized form of a result from [14], that considers a scalar parameter.

**Theorem 4.2.** (See [14, Theorem 3.13].) *Let $(\Xi, \Sigma, \nu)$ be a probability space and $W \subseteq \mathbb{R}^n$ be an open set. Let $h \colon W \times \Xi \to \mathbb{R}^m$ be a function such that*

(i) *$\xi \mapsto h(w, \xi)$ is integrable for each $w \in W$;*

(ii) *$w \mapsto h(w, \xi)$ is continuously differentiable for each $\xi \in \Xi$;*

(iii) *$\partial h / \partial w$ is $L^1(\nu)$-continuous.*

*Then $(\partial / \partial w) \int_{\Xi} h(w, \xi) \, \mathrm{d}\nu(\xi) = \int_{\Xi} (\partial h / \partial w)(w, \xi) \, \mathrm{d}\nu(\xi)$ for all $w \in W$.*

This criteria has the useful property that once it is established for $f$, it is easily extended to the function $e \circ f$. This is shown in the next proposition.

**Proposition 4.2.** *Let Assumptions 1.1–1.4 hold. If $e \in \mathcal{E}^2$ and $i + j \leq 2$ then, for any $(x, \theta) \in X \times \Theta$,*

$$\frac{\partial^{i+j}}{\partial x^i \partial \theta^j} \int_{\Xi} e(f(x, \xi, \theta)) \, \mathrm{d}\nu(\xi) = \int_{\Xi} \frac{\partial^{i+j}}{\partial x^i \partial \theta^j} e(f(x, \xi, \theta)) \, \mathrm{d}\nu(\xi).$$

*Proof.* Consider the derivative $\partial / \partial x$. To apply Theorem 4.2, we show that the map $x \mapsto \int_{\Xi} \|(\partial e / \partial x)(f(x, \xi, \theta))(\partial f / \partial x)(x, \xi, \theta)\| \, \mathrm{d}\nu(\xi)$ is continuous. Noting that

$$\left\| \frac{\partial e}{\partial x}(f(x, \xi, \theta)) \frac{\partial f}{\partial x}(x, \xi, \theta) \right\| \leq \left\| \frac{\partial e}{\partial x} \right\|_A \left\| A(f(x, \xi, \theta)) \frac{\partial f}{\partial x}(x, \xi, \theta) A(x)^{-1} \right\| \|A(x)\|,$$

the result follows by assumption on $\partial f / \partial x$ and the monotonicity property of $L^1(\nu)$-continuity. Next, consider $\partial^2 / \partial \theta^2$. We have

$$\begin{aligned} \left\| \frac{\partial^2}{\partial \theta^2} e(f(x, \xi, \theta)) \right\| \leq{}& \left\| \frac{\partial^2 e}{\partial x^2} \right\|_{A,A} \left\| A(f(x, \xi, \theta)) \frac{\partial f}{\partial \theta}(x, \xi, \theta) B(x)^{-1} \right\|^2 \|B(x)\|^2 \\ &+ \left\| \frac{\partial e}{\partial x} \right\|_A \left\| A(f(x, \xi, \theta)) \frac{\partial^2 f}{\partial \theta^2}(x, \xi, \theta)(B(x)^{-1} \oplus B(x)^{-1}) \right\| \|B(x)\|^2. \end{aligned}$$

The $L^1(\nu)$-continuity of the left-hand side follows by the $L^1(\nu)$-continuity of the right-hand side together with the monotonicity property. Similar reasoning yields the other cases. □

Using this result, we can obtain the contraction property of $P$ with respect to the class $\mathcal{E}^2$, and obtain some bounds on the second-order derivatives of $P_\theta e$.

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

**Proposition 4.3.** *Let Assumptions 1.2–1.4 be in effect. For $e \in \mathcal{E}^2$ and $\theta \in \Theta$,*

(i) $\|(\partial^2/\partial x^2)P_\theta e\|_{A,A} \le K_{X^2}\|(\partial e/\partial x)\|_A + K_X^2\|(\partial^2 e/\partial x^2)\|_{A,A}$;

(ii) $\|(\partial^2/\partial \theta^2)P_\theta e\|_{B,B} \le K_{\Theta^2}\|\partial e/\partial x\|_A + K_\Theta^2\|\partial^2 e/\partial x^2\|_{A,A}$;

(iii) $\|(\partial^2/\partial x\partial \theta)P_\theta e\|_{A,B} \le K_{X,\Theta}\|\partial e/\partial x\|_A + K_X K_\Theta\|\partial^2 e/\partial x^2\|_{A,A}$.

*Furthermore, for each $\theta$ there is an $L_\theta \ge 0$ such that $\|P_\theta e\|_{\mathcal{E}^2} \le L_\theta \|e\|_{\mathcal{E}^2}$ for all $e \in \mathcal{E}^2$.*

*Proof.* We prove (ii); (i) and (iii) are established similarly. We have

$$\frac{\partial^2}{\partial \theta^2} P_\theta e(x)(B^{-1}(x) \oplus B^{-1}(x)) = T_1 + T_2,$$

where $T_1$ and $T_2$ are defined as

$$T_1 = \int_\Xi \frac{\partial e}{\partial x}(f(x,\xi,\theta))\frac{\partial^2 f}{\partial \theta^2}(x,\xi,\theta)(B(x)^{-1} \oplus B(x)^{-1})\,\mathrm{d}\nu(\xi),$$

$$T_2 = \int_\Xi \frac{\partial^2 e}{\partial x^2}(f(x,\xi,\theta))\bigg(\frac{\partial f}{\partial \theta}(x,\xi,\theta)B^{-1}(x) \oplus \frac{\partial f}{\partial \theta}(x,\xi,\theta)B^{-1}(x)\bigg)\,\mathrm{d}\nu(\xi).$$

Using the identity $A(f(x,\xi,\theta))^{-1}A(f(x,\xi,\theta))(\partial^2 f/\partial\theta^2)(x,\xi,\theta) = (\partial^2 f/\partial\theta^2)(x,\xi,\theta)$, we obtain

$$\|T_1\| \le \left\|\frac{\partial e}{\partial x}\right\|_A K_{\Theta^2}, \tag{4.2}$$

while for $T_2$, we use the fact that

$$A(f(x,\xi,\theta))^{-1}A(f(x,\xi,\theta))\frac{\partial f}{\partial \theta}(x,\xi,\theta) = \frac{\partial f}{\partial \theta}(x,\xi,\theta)$$

to obtain

$$\|T_2\| \le \left\|\frac{\partial^2 e}{\partial x^2}\right\|_{A,A}\bigg(\int_\Xi \left\|A(f(x,\xi))\frac{\partial f}{\partial \theta}(x,\xi)B^{-1}(x)\right\|^2 \mathrm{d}\nu(\xi)\bigg) \le \left\|\frac{\partial^2 e}{\partial x^2}\right\|_{A,A} K_\Theta^2.$$

Combining this last inequality with (4.2) leads to

$$\left\|\frac{\partial^2}{\partial \theta^2}P_\theta e(x)\right\|_{B(x),\,B(x)} \le \left\|\frac{\partial e}{\partial x}\right\|_A K_{\Theta^2} + \left\|\frac{\partial^2 e}{\partial x^2}\right\|_{A,A} K_\Theta^2.$$

To show the boundedness with respect to $\|\cdot\|_{\mathcal{E}^2}$, note that, for any $e \in \mathcal{E}^2$,

$$\begin{aligned}|(P_\theta e)(x)| &\le |e(x_0)| + \left\|\frac{\partial e}{\partial x}\right\|_A \int_X d_A(x_0,y)\,\mathrm{d}(\delta_x P_\theta)(y)\\ &\le |e(x_0)| + \left\|\frac{\partial e}{\partial x}\right\|_A [C_\theta + K_X d_A(x,x_0)],\end{aligned}$$

where $C_\theta$ is the number $C_\theta = \int_X d_A(x_0,y)\,\mathrm{d}(\delta_{x_0}P_\theta)(y)$. This follows, since, for the Lipschitz function $h(x) = d(x_0,x)$, $|(Ph)(x)| \le |Ph(x_0)| + |(Ph)(x_0) - (Ph)(x)| \le C_\theta + K_X d_A(x_0,x)$. Also, for any $x \in X$, $|e(x_0)|/(1 + d_A(x_0,x)) \le |e(x_0)|/(1 + d_A(x_0,x_0)) \le \|e\|_A$. Therefore, $\|P_\theta e\|_A \le \|e\|_A + \max\{C_\theta, K_X\}\|\partial e/\partial x\|_A$. □

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

The following quadratic bound involving the metric $d_A$ will be used as well.

**Proposition 4.4.** *Let* $h\colon X \to \mathbb{R}^n$ *be differentiable, such that* $\|(\partial h/\partial x)(x)A(x)^{-1}\| \le B(x)$, *where* $B\colon X \to \mathbb{R}$ *is Lipschitz for the metric* $d_A$. *Then the following inequalities hold:*

(i) $\|h(x) - h(y)\| \le B(x)d_A(x, y) + \frac{1}{2}\|B\|_{\mathrm{Lip}} d_A(x, y)^2$;

(ii) *for any* $\mu_1, \mu_2 \in \mathcal{P}_{2,A}(X)$,

$$\left\| \int_X h(x)\,\mathrm{d}\mu_1(x) - \int_X h(y)\,\mathrm{d}\mu_2(y) \right\| \le \|B\|_{L^2(\mu)} d_{2,A}(\mu_1, \mu_2) + \tfrac{1}{2}\|B\|_{\mathrm{Lip}} d_{2,A}(\mu_1, \mu_2)^2.$$

*Proof.* See Appendix A. □

With these tools in hand we can proceed to the proof of Theorem 4.1.

*Proof of Theorem 4.1.* In order to apply Theorem 3.1, we establish the requirements of Assumptions 3.1 and 3.2. Assumption 3.1 requires that, for any $\mu$ in $\mathcal{P}_{2,A}(X)$, the bound $\sup_{\|e\|_{\mathcal{E}^2} \le 1} |\mu(e)| < \infty$ holds. Note that $|e(x_0)| = |e(x_0)|/(1 + d_A(x_0, x_0)) \le \|e\|_A$. Then

$$|\mu(e)| \le \int_X \left[ |e(x_0)| + \left\| \frac{\partial e}{\partial x} \right\|_A d_A(x_0, x) \right] \mathrm{d}\mu(x) \le \max\left\{1, \int_X d_A(x, x_0)\,\mathrm{d}\mu(x)\right\} \|e\|_{\mathcal{E}^2}.$$

The integrability part of Assumption 1.2 and the contraction part of Assumption 1.4 allow us to apply Proposition 2.4. Hence, $P_\theta$ is a contraction on the space $\mathcal{P}_{2,A}(X)$ with contraction coefficient $K_X$, and has a unique invariant measure $\pi_\theta$ for each $\theta \in \Theta$. Then Assumption 3.2(i) holds. Proposition 4.3 affirms that $P_\theta e \in \mathcal{E}^2$ if $e \in \mathcal{E}^2$, and $P_\theta$ is bounded for the norm $\|\cdot\|_{\mathcal{E}^2}$. We now establish $\|P_\theta^n - \Pi_\theta\|_{\mathcal{E}^2} \le \rho_\theta K_X^n$ for some constant $\rho_\theta$. We consider each of the terms in the norm $\|\cdot\|_{\mathcal{E}^2}$ . First, for $e \in \mathcal{E}^2$,

$$\|P_\theta^n(e) - \Pi_\theta(e)\|_A \le K_X^n \left\| \frac{\partial e}{\partial x} \right\|_A \max\{C_\theta, 1\}. \tag{4.3}$$

To see this, observe that

$$\begin{aligned} |(P_\theta^n(e) - \Pi_\theta(e))(x)| &= |P_\theta^n(e)(x) - P_\theta^n(e)(x_0) + P_\theta^n(e)(x_0) - \pi_\theta(e)| \\ &\le K_X^n \left\| \frac{\partial e}{\partial x} \right\|_A d_A(x, x_0) + K_X^n \left\| \frac{\partial e}{\partial x} \right\|_A C_\theta \\ &\le K_X^n \left\| \frac{\partial e}{\partial x} \right\|_A \max\{C_\theta, 1\}(1 + d_A(x, x_0)), \end{aligned}$$

where $C_\theta = \int_X d_A(x_0, y)\,\mathrm{d}\pi_\theta(y)$. Next,

$$\left\| \frac{\partial}{\partial x}(P_\theta^n(e) - \Pi_\theta(e)) \right\|_A \le K_X^n \left\| \frac{\partial e}{\partial x} \right\|_A. \tag{4.4}$$

This inequality follows from Proposition 4.2 and Assumption 1.4. Finally, by recursive application of Proposition 4.3(i),

$$\left\| \frac{\partial^2}{\partial x^2}(P_\theta^n(e) - \Pi_\theta(e)) \right\|_{A,A} \le K_{X^2} K_X^{n-1} \frac{1}{1 - K_X} \left\| \frac{\partial e}{\partial x} \right\|_A + K_X^{2n} \left\| \frac{\partial^2 e}{\partial x^2} \right\|_{A,A}. \tag{4.5}$$

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

Adding (4.3)–(4.5), we obtain

$$\begin{aligned}\|P_\theta^n(e) - \Pi_\theta(e)\|_{\mathcal{E}^2} &\le K_X^n\bigg(\max\{C_\theta, 1\} + 1 + K_{X^2}\frac{1}{K_X(1-K_X)}\bigg)\bigg\|\frac{\partial e}{\partial x}\bigg\|_A + K_X^{2n}\bigg\|\frac{\partial^2 e}{\partial x^2}\bigg\|_{A,A}\\ &\le K_X^n\bigg(\max\{C_\theta, 1\} + 1 + K_{X^2}\frac{1}{K_X(1-K_X)}\bigg)\bigg(\bigg\|\frac{\partial e}{\partial x}\bigg\|_A + \bigg\|\frac{\partial^2 e}{\partial x^2}\bigg\|_{A,A}\bigg)\\ &\le K_X^n\rho_\theta\|e\|_{\mathcal{E}^2},\end{aligned}$$

where $\rho_\theta = \max\{C_\theta, 1\} + 1 + K_{X^2}(1/K_X(1-K_X))$. In the second inequality, we have used the fact that $K_X < 1$. Thus, Assumption 3.2(ii) is satisfied.

Proposition 4.2 affirms that $\theta \mapsto P_\theta e(x)$ is differentiable for $e \in \mathcal{E}^2$ and $x \in X$. Proceeding as in the proof there, we see that $\|(\partial/\partial\theta)P_{\theta_0}e(x)\| \le \|\partial e/\partial x\|_A K_\Theta\|B(x)\|$. Therefore, $\|\pi_{\theta_0}(\partial/\partial\theta)P_{\theta_0}\|_{\mathcal{L}} \le K_\Theta\|B\|_{L^1(\pi_{\theta_0})}$, which confirms Assumption 3.2(ii).

Proposition 4.3(ii) means that, for any $e \in \mathcal{E}^2$ and $\theta \in \Theta$, $\|(\partial^2/\partial\theta^2)P_\theta e(x)\|_{B(x),\,B(x)} \le k_1\|e\|_{\mathcal{E}^2}$, where $k_1 = \max\{K_\Theta^2, K_{\Theta^2}\}$. Using the second-order version of Taylor's theorem, this implies that for all $\Delta\theta$ sufficiently small, for all $e \in \mathcal{E}^2$, and $x \in X$, we have

$$\bigg|P_{\theta_0+\Delta\theta}e(x) - P_{\theta_0}e(x) - \frac{\partial}{\partial\theta}P_{\theta_0}e(x)(\Delta\theta)\bigg| \le \frac{1}{2}k_1\|e\|_{\mathcal{E}^2}\|B(x)\Delta\theta\|^2. \tag{4.6}$$

Integrating (4.6) and dividing by $\|\Delta\theta\|$ leads to

$$\frac{1}{\|\Delta\theta\|}\bigg\|\pi_{\theta_0}\bigg[P_{\theta_0+\Delta\theta} - P_{\theta_0} - \frac{\partial}{\partial\theta}P_{\theta_0}(\Delta\theta)\bigg]\bigg\|_{\mathcal{E}^2} \le \frac{1}{2}k_1\|B\|^2_{L^2(\pi_{\theta_0})}\|\Delta\theta\|$$

and the right-hand side goes to 0 as $\|\Delta\theta\| \to 0$. Only Assumption 3.2(v) remains. By the fundamental theorem of calculus,

$$(P_{\theta_0+\Delta\theta} - P_{\theta_0})e(x) = \int_0^1\int_\Xi \frac{\partial e}{\partial x}(f(x,\xi,\theta+\lambda\Delta\theta))\frac{\partial f}{\partial\theta}(x,\xi,\theta+\lambda\Delta\theta)\Delta\theta\,\mathrm{d}\nu(\xi)\,\mathrm{d}t.$$

Differentiating the above with respect to $x$ and using Assumption 1.4(iii) yields

$$\bigg\|\frac{\partial}{\partial x}((P_{\theta_0+\Delta\theta} - P_{\theta_0})e(x))A(x)^{-1}\bigg\| \le \|e\|_{\mathcal{E}^2}k_2\|\Delta\theta\|\|B(x)\|,$$

where $k_2 = \max\{K_{X,\Theta}, K_X K_\Theta\}$. Applying Proposition 4.4, we have

$$\begin{aligned}&\|(\pi_{\theta_0+\Delta\theta} - \pi_\theta)(P_{\theta_0+\Delta\theta} - P_{\theta_0})e\|\\ &\quad\le k_2\|\Delta\theta\|\|e\|_{\mathcal{E}^2}[\|B\|_{L^2(\pi_{\theta_0})}d_{2,A}(\pi_{\theta_0+\Delta\theta}, \pi_{\theta_0}) + \tfrac{1}{2}\|B\|_{\mathrm{Lip}}d_{2,A}(\pi_{\theta_0+\Delta\theta}, \pi_{\theta_0})^2].\end{aligned}$$

For the terms $d_{2,A}$, first apply the contraction property of $P$ and Proposition 4.1:

$$\begin{aligned}d_{2,A}(\pi_{\theta+\Delta\theta}, \pi_\theta) &\le d_{2,A}(\pi_{\theta+\Delta\theta}P_{\theta+\Delta\theta}, \pi_\theta P_{\theta+\Delta\theta}) + d_{2,A}(\pi_\theta P_{\theta+\Delta\theta}, \pi_\theta P_\theta)\\ &\le K_X d_{2,A}(\pi_{\theta+\Delta\theta}, \pi_\theta) + K_\Theta\|B\Delta\theta\|_{L^2(\pi_\theta)}.\end{aligned}$$

Rearranging terms yields $d_{2,A}(\pi_{\theta+\Delta\theta}, \pi_\theta) \le (1/(1-K_X))K_\Theta\|B\Delta\theta\|_{L^2(\pi_\theta)}$. Hence,

$$\begin{aligned}&\|(\pi_{\theta_0+\Delta\theta} - \pi_\theta)(P_{\theta_0+\Delta\theta} - P_{\theta_0})\|_{\mathcal{L}}\\ &\quad\le k_2\|B\|^2_{L^2(\pi_{\theta_0})}\|\Delta\theta\|\bigg[\frac{1}{1-K_X}K_\Theta\|\Delta\theta\| + \frac{1}{2}\|B\|_{\mathrm{Lip}}\bigg(\frac{1}{1-K_X}K_\Theta\|\Delta\theta\|\bigg)^2\bigg],\end{aligned}$$

and Assumption 3.2(v) is verified. □

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

## 5. Gradient estimation

The goal of this section is to prove Theorem 1.2. The standing assumptions are Assumptions 1.1–1.4. We let $Z = X \times M$ and denote elements of this space by $z = (x, m)$. Denote by $R_\theta$ the Markov kernel corresponding to the recursion (1.3). In Proposition 5.1 and Corollary 5.1, we establish convergence of the forward sensitivity system in the sense of Proposition 2.3. It involves finding an appropriate Lyapunov function $V$ and metric $d_H$ on $X \times M$. In Proposition 5.2, we show that $(x, m) \mapsto (\partial e/\partial x)(x)m$ is an integrable function for $\gamma_\theta$, thereby establishing that the right-hand side of (1.2) is finite. Finally, we want to show that the functional $l$ defined by

$$l(e) = \int_{X\times M} \frac{\partial e}{\partial x}(x)m \,\mathrm{d}\gamma_\theta(x, m) \tag{5.1}$$

is bounded for the norm $\|\cdot\|_{\mathcal{L}}$ and satisfies the derivative equation of Theorem 3.1.

Define $g$ and $T$ to be the functions

$$g((x, m), \xi, \theta) = \frac{\partial f}{\partial x}(x, \xi, \theta)m + \frac{\partial f}{\partial \theta}(x, \xi, \theta),$$
$$T((x, m), \xi, \theta) = (f(x, \xi, \theta), g((x, m), \xi, \theta)).$$

As $\theta$ is fixed in this section, we simplify notation and denote the values of $g$ by $g(z, \xi)$. We use $u_x$, $u_\theta$, and $u_m$ to denote vectors in $\mathbb{R}^{n_X}$, $\mathbb{R}^{n_\Theta}$, and $L(\mathbb{R}^{n_\Theta}, \mathbb{R}^{n_X})$, respectively.

**Proposition 5.1.** *Define* $h\colon Z \to \mathbb{R}_{\geq 0}$ *as* $h(z) = \eta_1\|A(x)m\| + \eta_2\|B(x)\| + \eta_3 d_A(x_0, x)$. *Then there are* $\eta_1, \eta_2, \eta_3, \eta_4, \eta_5$ *so that* $\{(\Xi, \Sigma, \nu), T, (\|\cdot\|_Z, H)\}$ *satisfies a* 1*-contraction inequality, where*

$$H(z)(u_x, u_m) = ((1 + \eta_4 h(z))A(x)u_x, A(x)u_m), \qquad \|(u_x, u_m)\|_Z = \|u_x\| + \eta_5\|u_m\|.$$

*Proof.* We will apply Proposition 2.6 to the map $T(z, \xi) = (f(x, \xi, \theta), g((x, m), \xi))$ in order to find contraction in the metric $d_H$. The norm $\|\cdot\|_M$ is the usual norm on $M$ induced by $\|\cdot\|_X$ and $\|\cdot\|_\Theta$. For Assumption 2.1(iii), we have

$$\begin{aligned}
&\sup_{\|u_m\|=1} \int_\Xi \left\| A(f(x, \xi, \theta))\frac{\partial g}{\partial m}(z, \xi)A(x)^{-1}u_m \right\| \mathrm{d}\nu(\xi) \\
&\quad = \sup_{\|u_m\|=1} \int_\Xi \sup_{\|u_x\|=1} \left\| A(f(x, \xi, \theta))\frac{\partial f}{\partial x}(x, \xi, \theta)A(x)^{-1}u_m u_x \right\| \mathrm{d}\nu(\xi) \\
&\quad \leq K_X
\end{aligned}$$

and, directly by assumption,

$$\sup_{\|u_x\|=1} \left( \int_\Xi \left\| A(f(x, \xi, \theta))\frac{\partial f}{\partial x}(x, \xi, \theta)A(x)^{-1}u_x \right\|^2 \mathrm{d}\nu(\xi) \right)^{1/2} \leq K_X.$$

We now establish Proposition 2.6(i). The function $(\partial g/\partial x)(z, \xi)$ is a linear map from $\mathbb{R}^{n_X}$ to $L(\mathbb{R}^{n_\Theta}, \mathbb{R}^{n_X})$, and we identify this with a bilinear map from $\mathbb{R}^{n_X} \times \mathbb{R}^{n_\Theta}$ to $\mathbb{R}^{n_X}$. Specifically,

$$\frac{\partial g}{\partial x}(z, \xi)[u_x, u_\theta] = \frac{\partial^2 f}{\partial x^2}(x, \xi, \theta)[u_x, m\, u_\theta] + \frac{\partial^2 f}{\partial x \partial \theta}(x, \xi, \theta)[u_x, u_\theta],$$

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

and $A(f(x,\xi,\theta))(\partial g/\partial x)(z,\xi)A(x)^{-1}$ is the linear map from $\mathbb{R}^{n_X}$ to $L(\mathbb{R}^{n_\Theta},\mathbb{R}^{n_X})$, where

$$A(f(x,\xi,\theta))\frac{\partial g}{\partial x}(z,\xi)A(x)^{-1}[u_x,u_\theta] = A(f(x,\xi,\theta))\frac{\partial^2 f}{\partial x^2}(x,\xi,\theta)[A(x)^{-1}u_x, m\,u_\theta] + A(f(x,\xi,\theta))\frac{\partial^2 f}{\partial x\partial\theta}(x,\xi,\theta)[A(x)^{-1}u_x, u_\theta].$$

For the first term, we have, using the assumption on $\partial^2 f/\partial x^2$ from Assumption 1.4 and the identity $m\,u_\theta = A(x)^{-1}A(x)m\,u_\theta$,

$$\sup_{\|u_x\|=1}\int_\Xi \sup_{\|u_\theta\|=1}\left\|A(f(x,\xi,\theta))\frac{\partial^2 f}{\partial x^2}(x,\xi,\theta)[A(x)^{-1}u_x, m\,u_\theta]\right\|\mathrm{d}\nu(\xi) \le K_{X^2}\|A(x)m\|.$$

For the second, use the identity $u_\theta = B(x)^{-1}B(x)u_\theta$ and our assumption on $\partial^2 f/\partial x\partial\theta$,

$$\sup_{\|u_x\|=1}\int_\Xi \sup_{\|u_\theta\|=1}\left\|A(f(x,\xi,\theta))\frac{\partial^2 f}{\partial x\partial\theta}(x,\xi,\theta)[A(x)^{-1}u_x, u_\theta]\right\|\mathrm{d}\nu(\xi) \le K_{X,\Theta}\|B(x)\|.$$

Combining these two inequalities, while assuming $K_{X^2}\le\eta_1$ and $K_{X,\Theta}\le\eta_2$,

$$\sup_{\|u_x\|=1}\int_\Xi\left\|A(f(x,\xi,\theta))\frac{\partial g}{\partial x}(z,\xi)A(x)^{-1}u_x\right\|\mathrm{d}\nu(\xi)\le K_{X^2}\|A(x)m\| + K_{X,\Theta}\|B(x)\|\le h(z).$$

Next, we confirm Proposition 2.6(ii) by showing the Lyapunov property of the function $h$. We consider the three terms of the function, starting with $\|A(x)m\|$:

$$\begin{aligned}\left(\int_\Xi\|A(f(x,\xi,\theta))g(z,\xi)\|^2\,\mathrm{d}\nu(\xi)\right)^{1/2} &\le \left(\int_\Xi\left\|A(f(x,\xi,\theta))\frac{\partial f}{\partial x}(x,\xi,\theta)m\right\|^2\mathrm{d}\nu(\xi)\right)^{1/2}\\ &\quad+\left(\int_\Xi\left\|A(f(x,\xi,\theta))\frac{\partial f}{\partial\theta}(x,\xi,\theta)\right\|^2\mathrm{d}\nu(\xi)\right)^{1/2}\\ &\le K_X\|A(x)m\| + K_\Theta\|B(x)\|.\end{aligned}$$

Next, we consider $\|B(x)\|$. Fix a basepoint $x_0$ and set $B_0 = (\int_\Xi\|B(f(x_0,\xi,\theta))\|^2\,\mathrm{d}\nu(\xi))^{1/2}$. Then

$$\begin{aligned}\left(\int_\Xi\|B(f(x,\xi,\theta))\|^2\,\mathrm{d}\nu(\xi)\right)^{1/2} &\le B_0 + \|B\|_{\mathrm{Lip}}\left(\int_\Xi d_A(f(x_0,\xi,\theta), f(x,\xi,\theta))^2\,\mathrm{d}\nu(\xi)\right)^{1/2}\\ &\le B_0 + \|B\|_{\mathrm{Lip}}\,K_X d_A(x_0,x).\end{aligned}$$

The first inequality uses Assumption 1.3 and the second uses the pointwise contraction property of $f$ which comes from Proposition 2.2. For the term $d_A(x_0,x)$, we have, setting $D_0 = (\int_\Xi d_A(x_0, f(x_0,\xi,\theta))^2\,\mathrm{d}\nu(\xi))^{1/2}$,

$$\begin{aligned}\left(\int_\Xi d_A(x_0, f(x,\xi,\theta))^2\,\mathrm{d}\nu(\xi)\right)^{1/2} &\le D_0 + \left(\int_\Xi d_A(f(x_0,\xi,\theta), f(x,\xi,\theta))^2\,\mathrm{d}\nu(\xi)\right)^{1/2}\\ &\le D_0 + K_X d_A(x_0,x).\end{aligned}$$

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

Combining these, we obtain

$$\left(\int_\Xi h(T(z,\xi))^2\,\mathrm{d}\nu(\xi)\right)^{1/2}$$
$$\leq \eta_1 K_X\|A(x)m\| + \eta_1 K_\theta\|B(x)\| + (\eta_2\|B\|_{\mathrm{Lip}}K_X + \eta_3 K_X)d_A(x_0,x) + K_4,$$

where $K_4 = \eta_2 B_0 + \eta_3 D_0$. Based on this inequality, it is evident that $\eta_1, \eta_2, \eta_3$ can be chosen so that the Lyapunov condition on $h$ is satisfied. Specifically, we take $K_{X^2} \leq \eta_1$, $\max\{K_{X,\Theta}, \eta_1 K_\Theta\} < \eta_2$, and $\eta_2\|B\|_{\mathrm{Lip}}K_X < \eta_3(1-K_X)$. □

We can use $h$ to obtain a Lyapunov function, yielding ergodicity of the sensitivity process.

**Corollary 5.1.** *Let the $\eta_1, \eta_2, \eta_3$ of Proposition 5.1 be chosen so that they are all positive. Let $V$ be the function $V(z) = \eta_1\|A(x)m\| + \eta_2\|B(x)\| + \eta_3 d_A(x_0,x) + 1$. Then the kernel $R_\theta$ has a unique invariant measure $\gamma_\theta \in \mathcal{P}_{1,V}(Z)$, and, for $\mu \in \mathcal{P}_{1,V}(Z)$, $\sup_{\|g\|_{\mathrm{Lip}(H)}+\|g\|_V\leq 1}|\mu R_\theta^n(g) - \gamma_\theta(g)| \to 0$ as $n \to \infty$.*

*Proof.* We apply Proposition 2.3, using the metric $d_H$ defined in Proposition 5.1. In Proposition 5.1 we established the pointwise contraction inequality needed for Proposition 2.3. For some $\beta \in [0,1)$, the inequality $\int_\Xi V(T(z,\xi,\theta))\,\mathrm{d}\nu(\xi) \leq \beta V(z) + (K_4+1)$ holds at $z \in Z$, as we have already shown in the proof of Proposition 5.1. It remains to show that $V$ has compact sublevel sets. Note that if $V(x,m) \leq r$ then $d_A(x_0,x) \leq r/\eta_3$ and $\|m\| \leq rK/\eta_1$, where $K$ is such that $\sup_{x\in X}\|A(x)^{-1}\| \leq K$. Thus, $V^{-1}[0,r]$ is contained in the compact set $\{(x,m) \in Z \mid d_A(x_0,x) \leq r/\eta_3 \text{ and } \|m\| \leq rK/\eta_1\}$. □

To ensure that the function $(x,m) \mapsto (\partial e/\partial x)(x)m$ is integrable for the measure $\gamma_\theta$, it suffices that it is Lipschitz for the metric $d_H$, and bounded for Lyapunov function $V$.

**Proposition 5.2.** *For any $e \in \mathcal{E}^2$, the map $(x,m) \mapsto (\partial e/\partial x)(x)m$ is a Lipschitz function in the metric $d_H$ of Proposition 5.1, and is also bounded for the norm $\|\cdot\|_V$.*

*Proof.* Let the $\eta_i$ be as in Proposition 5.1. Let $g(x,m) = (\partial e/\partial x)(x)m$. We have

$$\|g(x,m)\| \leq \left\|\frac{\partial e}{\partial x}\right\|_A \|A(x)m\| \leq \|e\|_{\mathcal{E}^2}\|A(x)m\| \leq \frac{1}{\eta_1}\|e\|_{\mathcal{E}^2}V(x,m);$$

hence, $\|g\|_V \leq (1/\eta_1)\|e\|_{\mathcal{E}^2}$. Next, we show that $\|g\|_{\mathrm{Lip}} < \infty$ for the metric $d_H$. This is equivalent to showing $\|\partial g/\partial x\|_H < \infty$. Let $(u_x, u_m)$ be a vector in $\mathbb{R}^{nx} \times L(\mathbb{R}^{n_\Theta}, \mathbb{R}^{nx})$. Then $H(z)^{-1}(u_x,u_m)$ is $H(z)^{-1}(u_x,u_m) = ((1/(1+\eta_4 h(z)))A^{-1}(x)u_x, A(x)^{-1}u_m)$ and $(\partial g/\partial z)(z)$ is the linear map from $\mathbb{R}^{nx} \times L(\mathbb{R}^{n_\Theta}, \mathbb{R}^{nx})$ to $L(\mathbb{R}^{n_\Theta}, \mathbb{R})$, where

$$\frac{\partial g}{\partial z}(z)[u_x,u_m][u_\theta] = \frac{\partial^2 e}{\partial x^2}(x)[u_x, mu_\theta] + \frac{\partial e}{\partial x}(x)[u_m u_\theta].$$

Fix $(u_x,u_m)$ with $\|u_x\| + \eta_5\|u_m\| = 1$. Then

$$\begin{aligned}
\left\|\frac{\partial g}{\partial z}(z)H(z)^{-1}(u_x,u_m)\right\| &= \sup_{\|u_\theta\|=1}\left|\frac{(\partial^2 e/\partial x^2)(x)[A^{-1}(x)u_x, mu_\theta]}{1+\eta_4 h(z)} + \frac{\partial e}{\partial x}(x)A^{-1}(x)u_m u_\theta\right| \\
&\leq \sup_{\|u_\theta\|=1}\frac{\|\partial^2 e/\partial x^2\|_{A,A}\|u_x\|\|A(x)m\|\|u_\theta\|}{1+\eta_4 h(z)} + \left\|\frac{\partial e}{\partial x}\right\|_A\|u_m\|\|u_\theta\| \\
&\leq \frac{\|\partial^2 e/\partial x^2\|_{A,A}\|u_x\|\|A(x)m\|}{1+\eta_4 h(z)} + \left\|\frac{\partial e}{\partial x}\right\|_A\|u_m\|.
\end{aligned}$$

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

To continue, note by definition of $h$ that $\|A(x)m\|/(1+\eta_4 h(z)) \leq 1/\eta_1\eta_4$. Then

$$\begin{aligned}\left\|\frac{\partial g}{\partial z}(z)H(z)^{-1}(u_x,u_m)\right\| &\leq \left\|\frac{\partial^2 e}{\partial x^2}\right\|_{A,A}\|u_x\|\frac{1}{\eta_1\eta_4}+\frac{\eta_5}{\eta_5}\left\|\frac{\partial e}{\partial x}\right\|_A\|u_m\| \\ &\leq \max\left\{\left\|\frac{\partial^2 e}{\partial x^2}\right\|_{A,A}\frac{1}{\eta_1\eta_4}, \frac{1}{\eta_5}\left\|\frac{\partial e}{\partial x}\right\|_A\right\} \\ &\leq \|e\|_{\mathcal{E}^2}\max\left\{\frac{1}{\eta_1\eta_4},\frac{1}{\eta_5}\right\}.\end{aligned}$$

Therefore, a Lipschitz constant for the function $g$ is $\|e\|_{\mathcal{E}^2}\max\{1/\eta_1\eta_4, 1/\eta_5\}$. □

We now continue to the proof of Theorem 1.2.

*Proof of Theorem 1.2.* By Corollary 5.1, the forward sensitivity process converges to a unique stationary measure $\gamma_\theta$ in $\mathcal{P}_{1,V}(Z)$. Let $g$ be the function $g(x,m) = (\partial e/\partial x)(x)m$. By Proposition 5.2, we see that $\|g\|_{\mathrm{Lip}} + \|g\|_V < \infty$, which means, in particular, that the integral on the right-hand side of (5.1) is well defined.

We show that the functional $l$ of (5.1) is bounded for the norm $\|\cdot\|_{\mathcal{L}}$. We have $\|l(e)\| \leq \|e\|_{\mathcal{E}^2}\int_Z \|A(x)m\|\,\mathrm{d}\gamma_\theta(z)$, with the latter integral being finite since $\gamma_\theta \in \mathcal{P}_{1,V}(Z)$. Then $\|l\|_{\mathcal{L}} < \infty$. It remains to show that $T(l) = l$. By the identity $\gamma_\theta = \gamma_\theta R_\theta$,

$$\begin{aligned}l(e) &= \int_{X\times M}\frac{\partial e}{\partial x}(x)m\,\mathrm{d}\gamma_\theta(x,m) \\ &= \int_{X\times M}\left(\int_\Xi \frac{\partial e}{\partial x}(f(x,\xi,\theta))\left(\frac{\partial f}{\partial x}(x,\xi,\theta)m+\frac{\partial f}{\partial\theta}(x,\xi,\theta)\right)\mathrm{d}\nu(\xi)\right)\mathrm{d}\gamma_\theta(x,m). \qquad (5.2)\end{aligned}$$

Recall that the definition of $T$ is $T(l)e = lP_\theta e + \pi_\theta(\partial/\partial\theta)P_\theta e$. With our definition of $l$, and applying Proposition 4.2, these two terms are

$$\begin{aligned}lP_\theta(e) &= \int_{X\times M}\frac{\partial}{\partial x}(P_\theta e)(x)m\,\mathrm{d}\gamma_\theta(x,m) \\ &= \int_{X\times M}\left(\int_\Xi\frac{\partial e}{\partial x}(f(x,\xi,\theta))\frac{\partial f}{\partial x}(x,\xi,\theta)\,\mathrm{d}\nu(\xi)\right)m\,\mathrm{d}\gamma_\theta(x,m) \qquad (5.3)\end{aligned}$$

and

$$\pi_\theta\frac{\partial}{\partial\theta}P_\theta e = \int_X\left(\int_\Xi\frac{\partial e}{\partial x}(f(x,\xi,\theta))\frac{\partial f}{\partial\theta}(x,\xi,\theta)\,\mathrm{d}\nu(\xi)\right)\mathrm{d}\pi_\theta(x). \qquad (5.4)$$

Add (5.3) to (5.4) and compare with (5.2) to see $T(l) = l$. □

To finish this section, we discuss how this estimator can be implemented. One option is to iterate the joint recursion (1.3) for a large number of steps, to obtain a sample $(x_n, m_n)$, and then prepare the estimate by forming the product $\Delta_n = (\partial e/\partial x)(x_n)m_n$. This requires the ability to compute the derivatives of $e$ and $f$. According to Theorem 1.2, the estimate $\Delta_n$ has the property that $\mathbb{E}[\Delta_n] \to (\partial/\partial\theta)\int_X e(x)\,\mathrm{d}\pi_\theta(x)$ as $M \to \infty$. To control the variance of the estimate, one can form the running averages $A_n = (1/n)\sum_{i=1}^n \Delta_i$. The results of [11] can be used in certain cases to quantify how the variance of the $A_n$ decreases with time.

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

## 6. Examples

**Example 6.1.** We consider a stochastic neural network where at each time only a subset of the edges in the network are activated. There are $N$ nodes so that the state space $X$ is $[0, 1]^N$. The random input is a binary vector in $\Xi = \{0, 1\}^{N\times N}$. Let $\sigma$ be the sigmoid function $\sigma(x) = (1 + \exp(-x))^{-1}$. The function $f\colon X \times \Xi \times \Theta \to X$ is

$$f_i(x, \xi, \theta) = \sigma(u_i(x, \xi, \theta)),$$

where $u_i(x, \xi, \theta) = \sum_{k=1}^{n} \xi_{i,k}\theta_{i,k}x_k$. The $b_i$ are biases and considered fixed. A vector $\xi \in \Xi$ indicates which edges are active at each time step; the edge $(i, j)$ from $j$ to $i$ is only used if $\xi_{i,j} = 1$. The probability measure on $\Xi$ is defined by $\nu(\xi) := \prod_{(i,j)\in E} \rho^{1-\xi_{i,j}}(1-\rho)^{\xi_{i,j}}$. Under this law, in the extreme $\rho = 1$, we have $\xi_{i,j} = 0$ for all $i, j$ with probability 1. The parameter space $\Theta$ is the $N \times N$ matrix $\mathbb{R}^{N\times N}$, which are the weights $\theta_{i,j}$ between each unit. We set $A(x) = I$ and $\|\cdot\|_X = \|\cdot\|_\infty$; hence, $d_A(x, y) = \|x - y\|_\infty$. We set $B(x) = I$. We need to find conditions so that Assumptions 1.1–1.4 hold. After setting $\Theta$ to be an arbitrary open ball, the only nontrivial part is the contraction criteria, Assumption 1.4(ii). Observe that $(\partial f_i/\partial x_j)(x, \xi, \theta) = \sigma'(u_i(x, \xi, \theta))\xi_{i,j}\theta_{i,j}$. With the norm $\|\cdot\|_\infty$ on $X$ and as $|\sigma'(u)| \le \frac{1}{4}$,

$$\left\| \frac{\partial f}{\partial x}(x, \xi, \theta)\right\|_\infty \le \frac{1}{4}\|\theta\|_\infty \sup_{i,j} \xi_{i,j}.$$

Note that $(\int_\Xi (\sup_{i,j} \xi_{i,j})^2 \,d\nu(\xi))^{1/2} = (1 - \nu(\xi = 0))^{1/2} = (1 - \rho^{|E|})^{1/2}$, so a sufficient condition for contraction in $d_2$ is $\|w\|_\infty(1 - \rho^{|E|})^{1/2} < 4$. The matrix norm induced by $\|\cdot\|_\infty$ is the maximum absolute row sum; then the condition is that the sum of magnitudes of incoming weights at each node must be bounded in this way.

The requirements for applying forward sensitivity analysis are met. For completeness we derive the exact form of the sensitivity system. The space $M$ consists of the linear maps from $\mathbb{R}^{N\times N}$ to $\mathbb{R}^N$ and $(\partial f_i/\partial\theta_{(j,k)})(x, \xi, \theta) = \delta_{i,j}\sigma'(u_i(x, \xi, \theta))\xi_{i,k}x_k$. We use subscripts to denote time, and $v(k)$ means the $k$th component of vector $v$. Then

$$x_{n+1}(i) = \sigma(u(x_n, \xi_{n+1}, \theta)(i)),$$

$$m_{n+1}(i, (j, k)) = \sigma'(u(x_n, \xi_{n+1}, \theta)(i))\left[\delta_{i,j}\xi_{n+1}(i, k)x_n(k) + \sum_{q=1}^{n} \xi_{n+1}(i, q)\theta(i, q)m_n(q, (j, k))\right].$$

At time $n + 1$, node $i$ has to pull from each node $q$ that connects to it the data $m_n(q, (j, k))$ and the state variable $x_n(q)$.

**Example 6.2.** Let $\Xi = \mathbb{R}^2$ and let $\nu$ be the law of two independent random variables $\xi_1, \xi_2$, such that $\mathbb{E}[\exp(6|\xi_1|) + |\xi_2|^2] < \infty$. Let $f\colon \mathbb{R}^2 \times \Xi \times \Theta \to \mathbb{R}^2$ be the function

$$f(x, \xi, \theta) = (f_1(x_1, \xi, \theta), f_2(x_1, x_2, \xi, \theta)), \tag{6.1}$$

where $f_1(x_1, \xi, \theta) = \frac{1}{2}x_1 + \theta + \varepsilon\xi_1$ and $f_2(x_1, x_2, \xi, \theta) = \frac{1}{2}x_1x_2 + \varepsilon\xi_2$. Let $g_1, g_2$ be the real-valued functions $g_1(x) = \exp(2|x_1|)(1+|x_2|)$ and $g_2(x) = \exp(2|x_1|)$. The metric $d_A$ will be defined using the pair $(\|\cdot\|, A)$, where $\|(u, v)\| = p_1|u| + p_2|v|$ and $A(x) = g_1(x) \oplus g_2(x)$, with $p_1, p_2$ determined below. The parameter $\theta$ is a number and $B$ is $B(x) = g_1(x)$. We seek conditions on $\varepsilon$ and $\theta$ that guarantee contraction and the applicability of stochastic forward sensitivity analysis.

**Proposition 6.1.** *Let the following hold:*

(i) *the parameter space is* $\Theta = (-\frac{1}{4}\log 2, \frac{1}{4}\log 2)$*;*

(ii) $\varepsilon < 1$ *and* $(1 + \varepsilon(\int_\Xi |\xi_2|^2 \, \mathrm{d}\nu(\xi))^{1/2})(\int_\Xi \exp(2\varepsilon|\xi_1|)^2 \, \mathrm{d}\nu(\xi))^{1/2} < 2^{1/4}$*;*

(iii) *the coefficients* $p_1, p_2$ *are any positive numbers such that* $1 + p_2/p_1 < 2^{1/4}$*.*

*For* $\theta \in \Theta$*, the stochastic forward sensitivity method is applicable for the system (6.1).*

*Proof.* See Appendix A for a sketch of the calculations involved. □

Based on the definition of $\mathcal{E}^2$, the cost functions are those $e\colon \mathbb{R}^2 \to \mathbb{R}$ satisfying

$$\sup_x \left|\frac{\partial e}{\partial x_i}(x)\right| g_i(x)^{-1} < \infty \quad \text{and} \quad \sup_x \left|\frac{\partial^2 e}{\partial x_i \partial x_j}(x)\right| g_i^{-1}(x) g_j^{-1}(x) < \infty \quad \text{for } 1 \le i,\ j \le 2.$$

Note that since $g_i \ge 1$, the functions in $\mathcal{E}$ include those with $\sup_x \|(\partial e/\partial x)(x)\| < \infty$ and $\sup_x \|(\partial^2 e/\partial x^2)(x)\| < \infty$. The joint process takes the following form. We denote the $k$th component of a vector $v$ by $v(k)$, and use a subscript to denote time. Thus,

$$x_{n+1}(1) = \tfrac{1}{2}x_n(1) + \theta + \varepsilon\xi_{n+1}(1), \qquad x_{n+1}(2) = \tfrac{1}{2}x_n(1)x_n(2) + \varepsilon\xi_{n+1}(2),$$
$$m_{n+1}(1) = \tfrac{1}{2}m_n(1) + 1, \qquad m_{n+1}(2) = \tfrac{1}{2}x_n(2)m_n(1) + \tfrac{1}{2}x_n(1)m_n(2).$$

## 7. Discussion

Our approach to establishing differentiability can be compared with works on measure-valued differentiation, such as [9] and [10]. The ergodicity framework in those works is based on normed ergodicity [1], while ours is also based on a norm but involves the derivatives of the cost functions as well. The approach to establishing differentiability is based on setting up a certain equation between linear functionals, showing that any solution to that equation must evaluate the stationary derivative, and showing that the equation indeed has a solution. In this sense, it is similar to [21], which works with the class of bounded measurable cost functions, and in a different ergodicity framework. Pflug [13] also used contraction in the Wasserstein distance in an ergodicity framework for stationary gradient estimation. This work was motivated by derivative estimation and optimization in neural networks. The back-propagation procedure is based on *adjoint sensitivity analysis*, as opposed to the forward sensitivity analysis studied here. Adjoint sensitivity analysis is often preferred as the auxiliary system in this case evolves in a space which has dimension $n_X$ as opposed to $n_\theta \times n_X$. In [3] and [4], the author analyzed joint gradient estimation/optimization schemes based on adjoint sensitivity analysis. It may be that the methods of this paper can be extended to adjoint sensitivity analysis. A counter example to this possibility would also be very interesting.

Another interesting extension may be to apply recursively the construction to obtain estimators for higher derivatives. Calculating $(\partial^2/\partial\theta^2)\mathbb{E}_{\pi_\theta}[e(x)]$ should be equivalent to computing $(\partial/\partial\theta)\mathbb{E}_{\gamma_\theta}[g(x)]$ for the 'cost function' $g(x) = (\partial e/\partial x)(x)m$.

## Appendix A.

We adopt the following notation. Denote $\Theta$ the space of parameters, $n_X$ the dimensionality of state space for the underlying system, $n_\Theta$ the dimensionality of parameter space, $L(\mathbb{R}^n, \mathbb{R}^m)$ the space of linear maps from $\mathbb{R}^n$ to $\mathbb{R}^m$, $M$ the space $L(\mathbb{R}^{n_\Theta}, \mathbb{R}^{n_X})$, and let $\|V\|_{L^p(\mu)}$ be shorthand for $(\int_X \|V(x)\|^p \, \mathrm{d}\mu)^{1/p}$. Further denote $\mathcal{P}(X)$ as the Borel probability measures

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

on $X$, $\mathscr{P}_{p,V}(X)$ the measures in $\mathscr{P}(X)$ that such that $\|V\|_{L^p(\mu)} < \infty$, $d_A$ the metric induced by a Finsler structure, $\mathscr{P}_{p,A}(X)$ the measures such that $\int_X d_A(x,x_0)^p \,\mathrm{d}\mu(x) < \infty$, $d_{p,A}$ the Wasserstein distance on the space $\mathscr{P}_{p,A}$, $\|\cdot\|_{\mathrm{Lip}}$ the Lipschitz constant for a function between metric spaces, and $(E \oplus F)(u,v) = (Eu, Fv)$ the direct sum of linear maps; $(E \oplus F)(u,v) = (Eu, Fv)$. Finally, denote $\|\cdot\|_{A,A}$ as the norm for a bilinear map: $\|m\|_{A,A} = \sup_{\|u\|=\|v\|=1} \|A[u,v]\|$, $\|\cdot\|_{\mathscr{E}^2}$ the norm $\|e\|_{\mathscr{E}^2} = \|e\|_A + \|\partial e/\partial x\|_A + \|\partial^2 e/\partial x^2\|_{A,A}$, and $I_n$ the identity matrix on $\mathbb{R}^n$.

*Proof of Proposition 2.1.* The metric axioms follow the approach of [2, Chapter 2]. We show the completeness. The condition on $A(x)^{-1}$ means that for some $k$, the inequality

$$\|x - y\| \le k d_A(x,y) \tag{A.1}$$

holds for all $x, y \in X$. The continuity of $A$ means that $\|A\|$ is bounded on compact subsets of $X$. Combining this with (A.1), it follows that $d_A$ and the metric determined on $\|\cdot\|$ are strongly equivalent on compact subsets of $X$. Using (A.1), one can show that any $d_A$-Cauchy sequence is contained in a compact subset of $X$. By the strong equivalence, $d_A$ is complete on this subset. □

*Proof of Proposition 4.4.* We will make use of the following: whenever $\gamma : [0,T] \to X$ is a curve from $x$ to $y$ that is

- parameterized by arc length, and
- such that $L(\gamma) \le d_A(x,y) + \varepsilon$, then

$$\int_0^T d_A(\gamma(t), x)\,\mathrm{d}t \le \frac{(d_A(x,y)+\varepsilon)^2}{2}. \tag{A.2}$$

To see this, note that for any curve parameterized by arc length, $d_A(\gamma(t), x) \le t$. Integrating both sides of this inequality and using the first assumption yields the result.

We now proceed to the proof of (i). Let $h : X \to \mathbb{R}^n$ be a function satisfying the assumptions of the proposition. Given $\varepsilon > 0$, let $\gamma : [0,T] \to X$ be a piecewise $C^1$ curve from $x$ to $y$ with $L(\gamma) \le d_A(x,y) + \varepsilon$. Assume that $\gamma$ is parameterized by arc length. By the identity $\gamma'(t) = A(\gamma(t))^{-1} A(\gamma(t))\gamma'(t)$, and the assumption on $h$,

$$\begin{aligned}\|h(x) - h(y)\| &\le \int_0^T \left\| \frac{\partial h}{\partial x}(\gamma(t))\gamma'(t) \right\| \mathrm{d}t \\ &= \int_0^T \left\| \frac{\partial h}{\partial x}(\gamma(t)) A(\gamma(t))^{-1} A(\gamma(t))\gamma'(t) \right\| \mathrm{d}t \\ &\le \int_0^T B(\gamma(t))\,\mathrm{d}t.\end{aligned}$$

Noting that $B$ is Lipschitz, and invoking (A.2),

$$\begin{aligned}\|h(x) - h(y)\| &\le \int_0^T (B(x) + \|B\|_{\mathrm{Lip}} d_A(\gamma(t), x))\,\mathrm{d}t \\ &\le B(x) \int_0^T 1\,\mathrm{d}t + \|B\|_{\mathrm{Lip}} \int_0^T d_A(\gamma(t), x)\,\mathrm{d}t \\ &\le B(x)[d_A(x,y) + \varepsilon] + \|B\|_{\mathrm{Lip}} \left[ \frac{d_A(x,y)^2}{2} + d_A(x,y)\varepsilon + \frac{\varepsilon^2}{2} \right].\end{aligned}$$

Since $\varepsilon$ was arbitrary, we have $\|h(x) - h(y)\| \le B(x) d_A(x,y) + \frac{1}{2}\|B\|_{\mathrm{Lip}} d_A(x,y)^2$.

For (ii), let $\gamma$ be any coupling of $\mu_1$ with $\mu_2$ such that $(\int_{X\times X} d_A(x,y)^2\,\mathrm{d}\gamma(x,y))^{1/2} \le d_{2,A}(\mu_1,\mu_2)+\varepsilon$. Then

$$\begin{aligned}
&\left\| \int_X h(x)\,\mathrm{d}\mu_1(x) - \int_X h(y)\,\mathrm{d}\mu_2(y) \right\| \\
&\quad \le \int_{X\times X} \|h(x)-h(y)\|\,\mathrm{d}\gamma(x,y) \\
&\quad \le \int_{X\times X} B(x) d_A(x,y)\,\mathrm{d}\gamma(x,y) + \frac{1}{2}\|B\|_{\mathrm{Lip}} \int_{X\times X} d_A(x,y)^2\,\mathrm{d}\gamma(x,y) \\
&\quad \le \|B\|_{L^2(\mu_1)}(d_{2,A}(\mu_1,\mu_2)+\varepsilon) + \tfrac{1}{2}\|B\|_{\mathrm{Lip}}(d_{2,A}(\mu_1,\mu_2)+\varepsilon)^2.
\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, the proof is complete. □

*Proof of Proposition 6.1.* We verify Assumptions 1.1–1.4. For Assumption 1.1, the continuity is obvious. As $A$ has a diagonal structure, $\|A(x)^{-1}\| = \max\{g_1(x)^{-1},\ g_2(x)^{-1}\}$, so it is clear that $\|A(x)^{-1}\| \le 1$ for all $x$.

For Assumption 1.2, the differentiability is evident. For the integrability, using the basepoint $(0,0)$, it suffices that $(\int_\Xi d_A(0, f(x,\xi,\theta))^2\,\mathrm{d}\nu(\xi))^{1/2} < \infty$ for any $(x,\theta)\in X\times\Theta$. Consider the curve $t \mapsto t f(x,\xi,\theta)$ for $t\in[0,1]$ from 0 to $f(x,\xi,\theta)$. Then $d_A(0, f(x,\xi,\theta)) \le \int_0^1 \|A(t f(x,\xi,\theta)) f(x,\xi,\theta)\|\,\mathrm{d}t$. Next, by the definition of $\|\cdot\|$,

$$\begin{aligned}
&\|A(t f(x,\xi,\theta)) f(x,\xi,\theta)\| \\
&\quad = p_1\,|g_1(t f(x,\xi,\theta)) f_1(x,\xi,\theta)| + p_2\,|g_2(t f(x,\xi,\theta)) f_2(x,\xi,\theta)|.
\end{aligned}$$

For the first term on the right-hand side of this equation, we have

$$\begin{aligned}
&|g_1(t f(x,\xi,\theta)) f_1(x,\xi,\theta)| \\
&\quad = \exp\!\big(2\big|t\tfrac{1}{2}x_1 + t\theta + t\varepsilon\xi_1\big|\big)\big(1+\big|t\tfrac{1}{2}x_1x_2 + t\varepsilon\xi_2\big|\big)\big|\tfrac{1}{2}x_1+\theta+\varepsilon\xi_1\big| \\
&\quad \le \exp(|x_1|)\exp(2|\theta|)\exp(2\varepsilon|\xi_1|)\big(1+\tfrac{1}{2}|x_1||x_2|+\varepsilon|\xi_2|\big)\big(\tfrac{1}{2}|x_1|+|\theta|+\varepsilon|\xi_1|\big) \\
&\quad \le \exp(2|x_1|+|x_1||x_2|)\exp(2|\theta|)\exp(3\varepsilon|\xi_1|)(1+\varepsilon|\xi_2|).
\end{aligned}$$

In the last inequality, we used the fact that $\theta < \frac{1}{2}$. Likewise, for the second term,

$$\begin{aligned}
|g_2(tf(x,\xi,\theta)) f_2(x,\xi,\theta)| &= \exp\!\big(2\big|t\tfrac{1}{2}x_1+t\theta+t\varepsilon\xi_1\big|\big)\big|\tfrac{1}{2}x_1x_2+\varepsilon\xi_2\big| \\
&\le \exp(|x_1|+|x_1x_2|)\exp(2|\theta|)\exp(2\varepsilon|\xi_1|)\varepsilon|\xi_2|.
\end{aligned}$$

Combining these, we obtain a bound for $d_A(0, f(x,\xi,\theta))$:

$$\begin{aligned}
d_A(0, f(x,\xi,\theta)) &\le p_1\exp(2|x_1|+|x_1x_2|)\exp(2|\theta|)\exp(3\varepsilon|\xi_1|)(1+\varepsilon|\xi_2|) \\
&\quad + p_2\exp(|x_1|+|x_1x_2|)\exp(2|\theta|)\exp(2\varepsilon|\xi_1|)\varepsilon|\xi_2| \\
&\le (p_1+p_2)\exp(2|x_1|+|x_1x_2|)\exp(2|\theta|)\exp(3\varepsilon|\xi_1|)(1+\varepsilon|\xi_2|). \qquad \text{(A.3)}
\end{aligned}$$

Let $Q = (\int_\Xi |\xi_2|^2\,\mathrm{d}\nu(\xi))^{1/2}$ and set $R = (\int_\Xi \exp(2\varepsilon|\xi_1|)^2\,\mathrm{d}\nu(\xi))^{1/2}$. Squaring and integrating (A.3) yields

$$\begin{aligned}
&\left(\int_\Xi d_A(0, f(x,\xi,\theta))^2\,\mathrm{d}\nu(\xi)\right)^{1/2} \\
&\quad \le (p_1+p_2)\exp(2|x_1|+|x_1x_2|)\exp(2|\theta|)\left(\int_\Xi \exp(3\varepsilon|\xi_1|)^2\,\mathrm{d}\nu(\xi)\right)^{1/2}(1+\varepsilon Q),
\end{aligned}$$

which is finite by the assumption that $\exp(6|\xi_1|)$ is integrable and that $\varepsilon < 1$.

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

For Assumption 1.3, the invertibility of $B(x)$ follows since $g_1 > 1$. Next, we show that $\|B(x)\|$ is Lipschitz for $d_A$. Since $\|e\|_{\text{Lip}} = \|\partial e/\partial x\|_A$ when $e$ is differentiable, the Lipschitz continuity of $g_1$ can be shown as follows. Let $x = (x_1, x_2)$ be a point of differentiability for $(|x_1|, |x_2|)$, and let $p_1|u| + p_2|v| = 1$. Then

$$\begin{aligned}\left|\frac{\partial g_1}{\partial x}(x)A(x)^{-1}(u,v)\right| &= \left|\frac{\partial g_1}{\partial x}(x)(g_1(x)^{-1}u, g_2(x)^{-1}v)\right| \\ &= \left|\frac{\partial g_1}{\partial x_1}(x)g_1(x)^{-1}u + \frac{\partial g_1}{\partial x_2}(x)g_2(x)^{-1}v\right| \\ &\leq \max\left\{\frac{1}{p_1}\left|\frac{\partial g_1}{\partial x_1}(x)g_1(x)^{-1}\right|, \frac{1}{p_2}\left|\frac{\partial g_1}{\partial x_2}(x)g_2(x)^{-1}\right|\right\},\end{aligned}$$

where $|(\partial g_1/\partial x_1)(x)g_1(x)^{-1}| \leq 2$ and $|(\partial g_1/\partial x_2)(x)g_2(x)^{-1}| \leq 1$. By an argument using a mollification of $|\cdot|$, this is extended to all points of $X$. Therefore, $\|g\|_{\text{Lip}} \leq \max\{2/p_1, 1/p_2\}$.

We turn to the functions $L_{X^i,\Theta^j}$, starting with $L_X$. Note the inequalities

$$g_1(f(x,\xi,\theta))\left|\frac{\partial f_1}{\partial x_1}(x,\xi,\theta)\right|g_1(x)^{-1} \leq \tfrac{1}{2}\exp(2|\theta|)\exp(2\varepsilon|\xi_1|)\exp(|x_1|)\big(1+\tfrac{1}{2}|x_1|+\varepsilon|\xi_2|\big)\exp(-2|x_1|), \tag{A.4}$$

$$g_2(f(x,\xi,\theta))\left|\frac{\partial f_2}{\partial x_1}(x,\xi,\theta)\right|g_1(x)^{-1} \leq \tfrac{1}{2}\exp(2|\theta|)\exp(2\varepsilon|\xi_1|)\exp(|x_1|)\exp(-2|x_1|), \tag{A.5}$$

$$g_2(f(x,\xi,\theta))\left|\frac{\partial f_2}{\partial x_2}(x,\xi,\theta)\right|g_2(x)^{-1} \leq \tfrac{1}{2}\exp(2|\theta|)\exp(2\varepsilon|x_1|)\exp(|x_1|)|x_1|\exp(-2|x_1|). \tag{A.6}$$

Next, note that

$$\begin{aligned}&\left\|A(f(x,\xi,\theta))\frac{\partial f}{\partial x}(x,\xi,\theta)A(x)^{-1}\right\| \\ &\quad\leq \max\left\{g_1(f(x,\xi,\theta))\left|\frac{\partial f_1}{\partial x_1}(x,\xi,\theta)\right|g_1(x)^{-1}\right. \\ &\qquad\left. + \frac{p_2}{p_1}g_2(f(x,\xi,\theta))\left|\frac{\partial f_2}{\partial x_1}(x,\xi,\theta)\right|g_1(x)^{-1}, g_2(f(x,\xi,\theta))\left|\frac{\partial f_2}{\partial x_2}(x,\xi,\theta)\right|g_2(x)^{-1}\right\}.\end{aligned}$$

Combining this with the three inequalities (A.4)–(A.6), we obtain

$$\begin{aligned}&\left\|A(f(x,\xi,\theta))\frac{\partial f}{\partial x}(x,\xi,\theta)A(x)^{-1}\right\| \\ &\quad\leq \frac{1}{2}\exp(2|\theta|)\exp(2\varepsilon|\xi_1|)\exp(|x_1|)\max\left\{1+\frac{1}{2}|x_1|+\varepsilon|\xi_2|+\frac{p_2}{p_1}, |x_1|\right\}\exp(-2|x_1|) \\ &\quad\leq \frac{1}{2}\exp(2|\theta|)\exp(2\varepsilon|\xi_1|)\exp(|x_1|)\left[1+|x_1|+\varepsilon|\xi_2|+\frac{p_2}{p_1}\right]\exp(-2|x_1|).\end{aligned}$$

Squaring and integrating the right-hand side of the last inequality, and using the independence of the $\xi_1$ and $\xi_2$ variables yields

$$L_X(x,\theta) \leq \frac{1}{2}\exp(2|\theta|)R\exp(|x_1|)\left(1+\varepsilon Q+\frac{p_2}{p_1}+|x_1|\right)\exp(-2|x_1|).$$

https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press

This is a continuous function of $(x, \theta)$, so the continuity of $L_X$ holds. We now show the contraction property. Using the inequality $a + x \leq a\exp(x/a)$, we obtain

$$\leq \left(1 + \varepsilon Q + \frac{p_2}{p_1}\right)\frac{1}{2}\exp(2|\theta|)R\exp\left(\left[1 + \left(1 + \varepsilon Q + \frac{p_2}{p_1}\right)^{-1}\right]|x_1|\right)\exp(-2|x_1|).$$

Based on this, the contraction property holds if $\varepsilon, \theta, p_1$, and $p_2$ are such that $(1 + \varepsilon Q + p_2/p_1)\exp(2|\theta|)R < 2$ and one can verify that assumptions (i)–(iii) mean that this indeed is the case. Now consider $L_\Theta$. Let $\|\cdot\|_\Theta = |\cdot|$. Then $\|A(f(x,\xi,\theta))(\partial f/\partial\theta)(x,\xi,\theta)B(x)^{-1}\| = g_1(f(x,\xi,\theta))g_1(x)^{-1}$. Using a similar analysis as above,

$$g_1(f(x,\xi,\theta))g_1(x)^{-1} \leq \exp(2|\theta|)\exp(2\varepsilon|\xi_1|)\exp(|x|)\big(1 + \tfrac{1}{2}|x_1| + \varepsilon|\xi_2|\big)\exp(-2|x_1|).$$

Squaring and integrating the right-hand side of this equation yields

$$\begin{aligned} L_\Theta(x,\theta) &\leq \exp(2|\theta|)R\exp(|x_1|)\big(1 + \tfrac{1}{2}|x_1| + \varepsilon Q\big)\exp(-2|x_1|) \\ &\leq (1+\varepsilon Q)\exp(2|\theta|)R\exp\left(\left[1 + \frac{1}{2(1+\varepsilon Q)} - 2\right]|x_1|\right) \leq (1+\varepsilon Q)\exp(2|\theta|)R. \end{aligned}$$

From the first inequality, we can see that $L_\Theta$ is continuous. From the last, we can see that $L_\Theta$ is bounded on the set $X \times \Theta$. It remains to verify conditions on the higher derivatives. The higher derivatives vanish except for $\partial^2 f/\partial x^2$. This is defined as follows:

$$\frac{\partial^2 f_k}{\partial x_i \partial x_j}(x,\xi,\theta) = \begin{cases} \frac{1}{2} & \text{if } k = 2 \text{ and } i \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

For $i = 1, 2$, we have $A(x)^{-1}e_i = g_i^{-1}(x)e_i$, and by the basic properties of bilinear maps

$$A(f(x,\xi,\theta))\frac{\partial^2 f}{\partial x^2}(z)(A(x)^{-1}e_i, A(x)^{-1}e_j) = A(f(x,\xi,\theta))g_i^{-1}(x)g_j^{-1}(x)\frac{\partial^2 f}{\partial x_i\partial x_j}(x,\xi,\theta).$$

Note that $(\partial^2 f/\partial x_i\partial x_j)(x,\xi,\theta) = 0$ if $i = j$. When $i \neq j$, we have $(\partial^2 f/\partial x_i\partial x_j)(x,\xi,\theta) = (0, \frac{1}{2})$ and $A(f(x,\xi))g_1^{-1}(x)g_2^{-1}(x)(0,\frac{1}{2}) = (0, g_2(f(x,\xi))g_1^{-1}g_2^{-1}(x))$. Then, for any $i$, $j$,

$$\left\|A(f(x,\xi,\theta))\frac{\partial^2 f}{\partial x^2}(x,\xi,\theta)(A(x)^{-1}e_i, A(x)^{-1}e_j)\right\| \leq p_2 g_2(f(x,\xi,\theta))g_1(x)^{-1}g_2(x)^{-1}.$$

Note that $|g_1^{-1}(x)| \leq 1$, and the norms $\|\cdot\|_1$ and $\|\cdot\|_X$ satisfy $\|\cdot\|_1 \leq \max\{1/p_1, 1/p_2\}\|\cdot\|_X$. With this, we obtain

$$\begin{aligned} &\left\|A(f(x,\xi,\theta))\frac{\partial^2 f}{\partial x^2}(x,\xi,\theta)(A(x)^{-1} \oplus A(x)^{-1})\right\| \\ &\quad\leq \left(\max\left\{\frac{1}{p_1}, \frac{1}{p_2}\right\}\right)^2 p_2 g_2(f(x,\xi,\theta))g_2(x)^{-1} \\ &\quad= \max\left\{\frac{p_2}{p_1^2}, \frac{1}{p_2}\right\} g_2(f(x,\xi,\theta))g_2(x)^{-1} \\ &\quad= \max\left\{\frac{p_2}{p_1^2}, \frac{1}{p_2}\right\}\exp(2|\theta| + \varepsilon|\xi_1|). \end{aligned}$$

Integrating yields $L_{X^2}(x,\theta) \leq \max\{p_2/p_1^2, 1/p_2\}\exp(2|\theta|)\int_\Xi \exp(2\varepsilon|\xi_1|)\,\mathrm{d}\nu(\xi)$, which is bounded and continuous on $X \times \Theta$. □

## References


[1] Borovkov, A. A. and Hordijk, A. (2004). Characterization and sufficient conditions for normed ergodicity of Markov chains. *Adv. Appl. Prob.* **36,** 227–242.

[2] Burago, D., Burago, Y. and Ivanov, S. (2001). *A Course in Metric Geometry* (Graduate Stud. Math. **33**). American Mathematical Society, Providence, RI.

[3] Flynn, T. (2015). Timescale separation in recurrent neural networks. *Neural Comput.* **27,** 1321–1344.

[4] Flynn, T. (2016). Convergence of one-step adjoint methods. In *Proceedings of the 22nd International Symposium on Mathematical Theory of Networks and Systems*.

[5] Forni, F. and Sepulchre, R. (2014). A differential Lyapunov framework for contraction analysis. *IEEE Trans. Automatic Control* **59,** 614–628.

[6] Griewank, A. and Walther, A. (2008). *Evaluating Derivatives*, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia, PA.

[7] Hairer, M. (2006). Ergodic properties of Markov processes. Lecture given at the University of Warwick. Available at http://www.hairer.org/notes/Markov.pdf.

[8] Hairer, M. and Mattingly, J. C. (2008). Spectral gaps in Wasserstein distances and the 2D stochastic Navier–Stokes equations. *Ann. Prob.* **36,** 2050–2091.

[9] Heidergott, B. and Hordijk, A. (2003). Taylor series expansions for stationary Markov chains. *Adv. Appl. Prob.* **35,** 1046–1070. (Correction: **36** (2004), 1300.)

[10] Heidergott, B., Hordijk, A. and Weisshaupt, H. (2006). Measure-valued differentiation for stationary Markov chains. *Math. Operat. Res.* **31,** 154–172.

[11] Joulin, A. and Ollivier, Y. (2010). Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Prob.* **38,** 2418–2442.

[12] Lohmiller, W. and Slotine, J.-J. E. (1998). On contraction analysis for non-linear systems. *Automatica J.* **34,** 683–696.

[13] Pflug, G. C. (1992). Gradient estimates for the performance of Markov chains and discrete event processes. *Ann. Operat. Res.* **39,** 173–194.

[14] Pflug, G. C. (1996). *Optimization of Stochastic Models: The Interface Between Simulation and Optimization*. Kluwer, Boston, MA.

[15] Pineda, F. J. (1988). Dynamics and architecture for neural computation. *J. Complexity* **4,** 216–245.

[16] Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* **323,** 533–536.

[17] Russo, G., di Bernardo, M. and Sontag, E. D. (2010). Global entrainment of transcriptional systems to periodic inputs. *PLoS Comput. Biol.* **6,** e1000739.

[18] Simpson-Porco, J. W. and Bullo, F. (2014). Contraction theory on Riemannian manifolds. *Systems Control Lett.* **65,** 74–80.

[19] Steinsaltz, D. (1999). Locally contractive iterated function systems. *Ann. Prob.* **27**, 1952–1979.

[20] Stenflo, Ö. (2012). A survey of average contractive iterated function systems. *J. Difference Equat. Appl.* **18,** 1355–1380.

[21] Vázquez-Abad, F. J. and Kushner, H. J. (1992). Estimation of the derivative of a stationary measure with respect to a control parameter. *J. Appl. Prob.* **29,** 343–352.

[22] Villani, C. (2009). *Optimal Transport: Old and New*. Springer, Berlin.


https://doi.org/10.1017/apr.2018.6 Published online by Cambridge University Press