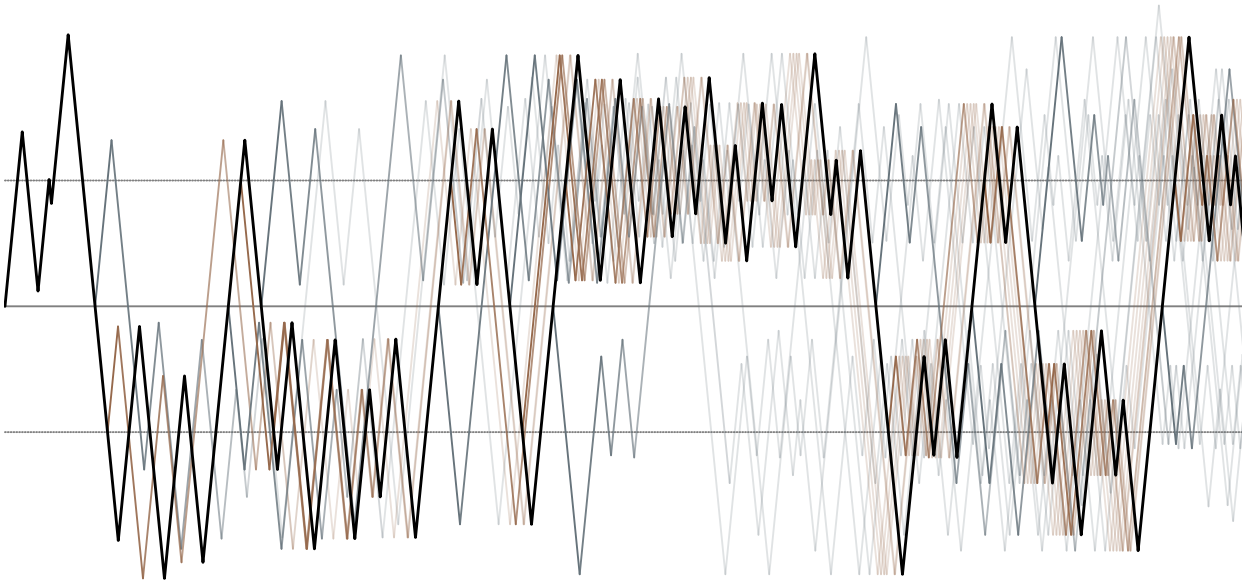


Differentiable Monte Carlo Samplers with Piecewise Deterministic Markov Processes

Master's thesis in Engineering Mathematics and Computational Science

RUBEN SEYER

DEPARTMENT OF MATHEMATICAL SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023
www.chalmers.se

Master's thesis 2023

Differentiable Monte Carlo Samplers with Piecewise Deterministic Markov Processes

Ruben Seyer

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2023

Differentiable Monte Carlo Samplers with
Piecewise Deterministic Markov Processes
Ruben Seyer

Supervisor: Moritz Schauer
Examiner: Axel Ringh

Master's Thesis 2023
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Given a one-dimensional Zig-Zag trajectory (definition 2.21) targeting a bimodal Gaussian (example 3.14), every alternative trajectory constructed by teleportation (definition 3.11) is displayed. The computational cost of handling alternatives would be severe were it not for the recoupling that occurs thanks to the construction (theorem 3.12).

Typeset in Erewhon with XeLaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2023

Differentiable Monte Carlo Samplers with
Piecewise Deterministic Markov Processes
Ruben Seyer
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

ABSTRACT

Gradient estimation by Monte Carlo methods, to e.g. find optimization directions, is an important component of many problems in statistics and machine learning. In one approach, related to the reparameterization trick, the sampling method itself is differentiated pathwise to obtain a sampler for the gradient. Unfortunately, the Hamiltonian Monte Carlo and other common methods contain a non-differentiable rejection step, for which pathwise derivatives do not provide unbiased estimates and corrections are computationally expensive. Here we use recently developed rejection-free methods based on piecewise deterministic Markov processes (PDMPs) to construct differentiable Monte Carlo methods. These handle unnormalized target densities as well as unbiased estimates of the target density. We find couplings (reparameterizations) for two PDMP methods, the Bouncy Particle sampler and the Zig-Zag sampler, which make them differentiable. The former is pathwise differentiable while the latter requires correction for large sample path perturbations, made efficient by our coupling. We investigate the theoretical properties of the resulting estimators, which only require a single sampler run. This opens up a promising new approach to stochastic gradient estimation problems.

Keywords: Piecewise deterministic Markov processes, Monte Carlo, gradient estimation, pathwise derivatives, reparameterization trick, probabilistic programming.

ACKNOWLEDGEMENTS

Without a doubt, half a year of work cannot be done in a vacuum. A part of this thesis belongs also to the people who supported me, whose patience was tested, and whose help made my work possible:

I begin with expressing my deepest gratitude to Moritz Schauer, my supervisor, whose efforts cannot be done justice with words. Not only would this project not exist without him, its completion would not have been possible without his patience and invaluable suggestions. I can only hope his commitment gave him half as much as it gave me. I am also extremely grateful to Axel Ringh, my examiner, who has gone above and beyond with both suggestions and advice. Were it not for this duo together, my life and career would probably look very different. It is with great excitement I join you as a colleague at the department!

I wish to extend special thanks to Gaurav Arya, whose attentive reading improved the final work at a critical point (in both the English as well as the thesis' sense of critical); I look forward to our fruitful collaboration in the future! I would also like to gratefully acknowledge the effort of my opponent Joakim Blomqvist in improving this work.

Finally, thank you to my friends for their support and office companionship. And last but not least, thank you to my patient Lotta, who has suffered enough musings about stochastic derivatives for a lifetime.

Ruben Seyer, Mölndal, 2023-06-15

CONTENTS

1 INTRODUCTION

The impact of Monte Carlo methods on the field of statistics in the last few decades cannot be overstated. Today, one would struggle to find a problem in Bayesian inference where no such technique is used to sample from a posterior, especially as datasets have grown simultaneously with our computational capabilities [1]. The standard tool for sampling is *Markov chain Monte Carlo* (MCMC), in particular the Metropolis-Hastings algorithm and its descendants, one of which is the popular Hamiltonian Monte Carlo algorithm [2]. Recently, a new type of samplers using *piecewise deterministic Markov processes* (PDMPs) have been developed [3]. In contrast to the common MCMC methods, which are discrete-time, PDMPs are continuous-time processes that follow deterministic dynamics but switch states at random event times. PDMP samplers are especially suited to large scale Bayesian applications, allowing subsampling techniques and unnormalized densities (cf. section 2.3).

An important application for Monte Carlo methods, particularly in machine learning, is stochastic gradient estimation, i.e. estimating the gradient of an expectation with respect to parameters of the underlying distribution. Such problems are ubiquitous in machine learning and statistics, where the estimators are used for optimization and for sensitivity analysis, with applications in e.g. approximate inference, reinforcement learning, financial mathematics, and experimental design [4] (cf. section 1.2). One class of gradient estimator which has enjoyed renewed interest is the *pathwise derivative* or the *reparameterization trick*, whereby the sampling is done indirectly through applying a transformation to independent randomness. The transformation can then be differentiated with respect to the parameters to obtain a gradient estimator, which in turn can be sampled using Monte Carlo methods [5, 6].

Gradient estimation has become even more important and accessible with the advent of *automatic differentiation* (AD), referring to methods which efficiently and automatically compute the derivatives of input functions [7]. AD has enabled differentiable *probabilistic programming languages*, where one can fluently describe generative models that incorporate samples of random variables, thus expressing complicated distributions. The models are then compiled into programs which sample from the resulting distribution using Monte Carlo methods, and if this method is differentiable the program essentially internalizes the machine learning concept of backpropagation without manual derivations [8, 9].

One strategy for general gradient estimation is therefore to directly differentiate the sampler program itself, viewing it as a complicated transformation of random numbers. However, Hamiltonian Monte Carlo uses a Metropolis-Hastings type acceptance-rejection step, which is not directly differentiable. Many applications omit this step to obtain a differentiable sampler [9], but the resulting estimator is then not guaranteed to be unbiased without introducing further corrections, such as resorting to simulating all alternative paths [10, 11]. This drawback is not shared by the rejection-free PDMPs. Hence, this thesis combines the two fields of PDMPs and stochastic gradient estimation, with the ultimate goal of producing a novel differentiable Monte Carlo method for unbiased gradient estimation.

1.1 OBJECTIVE

More precisely, we introduce the objective of this thesis as follows: Let μ_θ be a target probability distribution on $[\mathbb{R}, \mathscr{B}_\mathbb{R}]$, the measurable space where $\mathbb{R}$ denotes the real numbers and $\mathscr{B}_\mathbb{R}$ denotes the corresponding Borel σ-algebra, dependent on the parameter θ. In practice, this target may be given only by an unnormalized density. By using a suitable Monte Carlo sampler $\{Z_t^\theta\}_{t\geq 0}$ targeting this distribution, in our case a PDMP sampler, we can obtain expectations for this distribution according to the ergodic theorem

$$\mathbb{E}_{X^\theta\sim\mu_\theta}\left[f(X^\theta)\right] = \int f\,d\mu_\theta = \lim_{T\to\infty}\frac{1}{T}\int_0^T f(Z_t^\theta)\,dt \tag{1.1}$$

for integrable functions f (cf. theorem 2.18). In practice, approximate estimates of the expectation are obtained by finite-horizon trajectories. Other functionals of the sample path than expectations are possible, but we will focus on expectations. Now, we want to estimate the gradient in a similar way

$$\frac{\partial}{\partial\theta}\mathbb{E}_{X^\theta\sim\mu_\theta}\left[f(X^\theta)\right] \overset{?}{=} \lim_{T\to\infty}\frac{\partial}{\partial\theta}\frac{1}{T}\int_0^T f(Z_t^\theta)\,dt \tag{1.2}$$

applying stochastic derivative methodology to the Monte Carlo sampler. The estimator we construct should ideally be unbiased when starting according to the target distribution, consistent, require only a single trajectory of the sampler to compute, and exhibit low variance. In pursuit of this goal, we must identify assumptions on f and μ_θ under which our constructed estimator is valid.

1.2 APPLICATIONS

The applications of such a method are at the intersection of machine learning and Bayesian inference. Chandra, Li, Tenenbaum & Ragan-Kelley [9] have

catalogued several examples from the literature where differentiable Monte Carlo methods are used for gradient estimation:

- *Variational inference* [12]: The goal in variational inference is to fit a parameterized distribution, often represented by complex models such as neural networks, as an approximation to the true posterior. This transforms Bayesian inference into optimization. Fitting this distribution can be done using gradient estimates for the *evidence lower bound* and is one of the earliest applications of differentiable Monte Carlo in the machine learning literature.
- *Probabilistic programming* [9]: As previously mentioned, gradient estimates can be used to optimize over the expected output of probabilistic programs. In [9] a probabilistic model for human perception is used in optimization to generate optical illusions.
- *Bayesian learning* [13]: More direct inference tasks, such as fitting a Bayesian linear regression model with the sampler targeting the posterior, have also been done using gradient estimates.
- *Hyperparameter tuning* [14]: Monte Carlo methods themselves often involve several hyperparameters, and tuning them systematically to speed up convergence could be done by optimizing on gradient estimates for the expected final potential.

The cited literature for the examples all share a common approach of using Hamiltonian Monte Carlo to sample from the posterior, and thus run into the aforementioned issue of having to omit the acceptance-rejection step to obtain a differentiable sampler, in doing so obtaining a bias in the estimates. This illustrates a gap which can be filled by our objective of deriving a single-run unbiased gradient estimator.

1.3 OUTLINE

Following on this introduction, chapter 2 presents the necessary theory of stochastic derivatives, PDMPs, and more required to derive our estimators. In particular, an explicit theory of filtered stochastic derivatives is introduced. Chapter 3 then builds up to a general result using the Zig-Zag sampler for one-dimensional target distributions, with a proof of an unbiased, consistent gradient estimator. Next, chapter 4 uses many of the same ideas to construct an estimator using the Bouncy Particle sampler for multi-dimensional target distributions. Finally, chapter 5 discusses the big picture of the previous chapters and suggests future avenues of exploration.

2 BACKGROUND

2.1 STOCHASTIC DERIVATIVES

The general objective of recovering an unbiased estimator of the gradient makes it natural to consider methods to differentiate random variables. The immediate relation one might have to this topic is the subject of *stochastic calculus*, in which a theory of integration and stochastic differential equations is developed. However, in this thesis *time* derivatives of the stochastic processes involved are not the focus—in fact, differentiating in time would be easy for PDMPs instead of a major difficulty—but rather derivatives with respect to some *parameter* of the process, so that we consider a whole parametric family of processes.

The simplest possible example is to consider a single random variable $X^\theta \sim \mu_\theta$ whose distribution depends on some parameter θ. Our goal is to understand how X^θ behaves when the parameter is perturbed. The immediate naïve approach might be to recall first-year calculus and consider difference quotients of the type

$$\frac{1}{\varepsilon}\left(X^{\theta+\epsilon} - X^\theta\right) \tag{2.1}$$

for independent realizations $X^\theta, X^{\theta+\epsilon}$ and a small perturbation ε. Unfortunately, such differences are not necessarily well-behaved; the result can exhibit arbitrarily large fluctuations simply due to the inherent randomness, in particular if $\varepsilon \to 0$ [4].

To make this limit well-defined and reduce the variance, we devise a *coupling* between the two variables X^θ and $X^{\theta+\epsilon}$: a manufactured joint distribution such that each variable has the correct marginal distribution, but taken together there is some (positive) correlation or dependence. A small but important portion of this thesis is dedicated to finding couplings that make our analysis of stochastic derivatives easier, or even possible in the first place.

We present a few different approaches to this problem relevant for this thesis, but there are many variations and extensions; one suggestion for further reading is the survey by Mohamed, Rosca, Figurnov & Mnih [4], which contains several examples of applications.

2.1.1 Pathwise derivatives

The *pathwise derivative* [5, 15], also known as *infinitesimal perturbation analysis* and the *reparameterization trick* [6] is a method which directly targets the structure by which the process is (or could be) simulated. Our presentation here follows Glasserman [5, ch. 1].

Let $\{Z_t^\theta\}_{t\geq 0}$ be a continuous parametric family of processes defined on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Define a *performance functional* $F(Z^\theta)$, a map from a sample path $t \mapsto Z_t^\theta$ to a real value. The performance is itself random and dependent on θ due to the random path. Hence, we can view F as a function of the random outcome $\omega \in \Omega$ and the parameter θ, giving us a random variable $F(\cdot;\theta)$. Then consider the following (purely analytical) computation:

Definition 2.1. The *pathwise derivative* of $F(Z_t^\theta)$ wrt θ is the random variable

$$\frac{\partial F}{\partial \theta}(\omega;\theta) = \lim_{\epsilon \to 0} \frac{F(\omega;\theta+\varepsilon) - F(\omega;\theta)}{\varepsilon} \tag{2.2}$$

if this limit exists a.s. (that is, for a.e. $\omega \in \Omega$).

The name comes from the fact that it is obtained for each possible path corresponding to some ω. Under sufficient conditions on F we can then use it as an estimator:

Theorem 2.2 (Unbiasedness of pathwise derivative, [16, theorems 1.1–1.2])**.** *Let $F(\omega;\theta)$ be a performance functional as above. If $F(\cdot;\theta)$ is continuous and piecewise differentiable wrt θ a.s., and $\mathbb{E}\left[\sup_\theta \left|\frac{\partial F}{\partial \theta}(\cdot;\theta)\right|\right] < \infty$ taken over θ for which $F(\cdot;\theta)$ is differentiable, then*

$$\frac{\partial}{\partial \theta}\mathbb{E}[F(\cdot;\theta)] = \lim_{\epsilon\to 0}\mathbb{E}\left[\frac{F(\cdot;\theta+\varepsilon)-F(\cdot;\theta)}{\varepsilon}\right] = \mathbb{E}\left[\frac{\partial F}{\partial \theta}(\cdot;\theta)\right], \tag{2.3}$$

that is, the pathwise derivative is an unbiased estimator for the derivative of the expected performance.

In practice we do not observe an infinite path of the process, but rather some finite $\{Z_t^\theta\}_{t=0}^T$. If we make longer and longer observations on the trajectory we desire some kind of convergence of the derivative, assuming of course that the performance functional itself has this property.

Theorem 2.3 (Consistency of pathwise derivative, [16, theorem 1.3])**.** *Let the set of parameters Θ be compact, and let $F(\omega;\theta)$ be a performance functional defined as above. Furthermore, let $\{F_n(\omega;\theta)\}$ be a family of performance functionals such that $\lim_{n\to\infty} F_n(\cdot;\theta) = \mathbb{E}[F(\cdot;\theta)]$ a.s. (i.e. a strong 'law of large numbers' holds, with a.s. convergence to the mean) and the corresponding pathwise derivatives exist a.s.. Suppose that the following is satisfied for all $\theta \in \Theta$:*

(i) $\mathbb{E}\left[\frac{\partial F_n}{\partial \theta}(\cdot;\theta)\right] = \frac{\partial}{\partial\theta}\mathbb{E}[F_n(\cdot;\theta)] + h_n(\theta)$ *where $h_n(\theta) \to 0$ as $n \to \infty$,*

(ii) $\lim_{n\to\infty} \mathbb{E}[F_n(\cdot;\theta)] = \mathbb{E}[F(\cdot;\theta)]$,

(iii) $\lim_{n\to\infty} \frac{\partial F_n}{\partial\theta}(\cdot;\theta) = \lim_{n\to\infty} \mathbb{E}\left[\frac{\partial F_n}{\partial\theta}(\cdot;\theta)\right]$ *a.s. (with the limits existing a.s.),*

(iv) *There exists a bound* $g(\theta)$ *integrable on* Θ *such that for all* $\theta \in \Theta$ *it holds that* $\left|\mathbb{E}\left[\frac{\partial F_n}{\partial\theta}(\cdot;\theta)\right] - h_n(\theta)\right| < g(\theta)$.

Then $\mathbb{E}[F(\cdot;\theta)]$ *is a.e. differentiable, and* $\frac{\partial F_n}{\partial\theta}(\cdot;\theta)$ *is a strongly consistent estimator for its derivative, that is* $\frac{\partial F_n}{\partial\theta}(\cdot;\theta) \to \frac{\partial}{\partial\theta}\mathbb{E}[F(\cdot;\theta)]$ *a.s. as* $n \to \infty$.

Remark. In our setting, several of these assumptions hold indirectly. Clearly (ii) is dependent on the setting rather than the derivative scheme. If the finite horizon estimators $\frac{\partial F_n}{\partial\theta}(\cdot;\theta)$ are unbiased, then (i) holds with $h_n \equiv 0$. If we have Lipschitz continuity of $F(\cdot;\theta)$, then both the boundedness desired for unbiasedness as well as (iv) in the unbiased case (through Jensen's inequality) follow.

The abstract nature of this formulation hides the main challenge with it, which is to determine a mechanism by which we can connect 'independent randomness' in $\omega \in \Omega$ through a sample path $t \mapsto Z_t^\theta(\omega)$ to obtain the value of $F(\omega;\theta)$. It is precisely this connection which we differentiate, and so we must have a concrete relationship in mind to apply the theorems. The only constraint on such a connection is that $F(\cdot;\theta)$ has the correct marginal distribution for every θ, but we are free to choose the joint dependency—a *coupling*—to improve the properties or conditions for existence of the estimator. Even better if we can write $\frac{\partial F}{\partial\theta}(Z_t^\theta;\theta)$ so that the derivative can be recovered from the trajectory itself, thus requiring no extra simulation; then the connection can be purely theoretical and we are free to simulate the process any way we want.

2.1.2 *Stratified derivatives*

Nevertheless, the (simplified) condition that the performance should be a.s. Lipschitz in θ is relatively strong. Even in cases where the limit interchange in the pathwise derivative fails, the expectation can still have a well-defined derivative (as seen later in e.g. example 3.7), and we could conceivably construct some estimator. One way to correct the pathwise derivative is *smoothing* [15], also known as *smoothed perturbation analysis* [16]. Our presentation here is inspired by Heidergott & Vázquez-Abad [15], Fu & Hu [16, chs. 3–5] and Arya, Schauer, Schäfer & Rackauckas [17]; the latter call their construction *stochastic derivatives*.

On a high level, this smoothing consists of taking the conditional expectation of the performance with respect to some σ-algebra $\mathcal{G}$ such that

$$\frac{\partial}{\partial\theta}\mathbb{E}[F(\cdot;\theta)] = \mathbb{E}\left[\frac{\partial}{\partial\theta}\mathbb{E}[F(\cdot;\theta) \mid \mathcal{G}]\right]. \tag{2.4}$$

where now $\mathbb{E}[F(\cdot;\theta) \mid \mathcal{G}]$ is a.s. Lipschitz in θ. The difficulty lies in selecting $\mathcal{G}$ to be small enough that one 'integrates out' the problematic points where the

derivative fails, but large enough that no significant extra computation over the pathwise derivative is required. In general, there is no method to select $\mathscr{G}$ that works for all processes and performance functions, and any results obtained are only relevant to the specific setting.

Our approach will be to handle the discontinuities separately, hence the name *stratified* derivatives. The pathwise derivative contains infinitesimal perturbations which occur almost surely, while the discontinuities are finite perturbations that occur with infinitesimal probability. In the words of Heidergott & Vázquez-Abad [15], we partially integrate the discontinuities by instead differentiating the infinitesimal probabilities.

Definition 2.4. A *critical event* $A(\varepsilon,\theta)$ for the performance $F(\cdot;\theta)$ where $\varepsilon \neq 0$ is an event

$$A(\varepsilon,\theta) \subseteq \left\{\omega \in \Omega : \left|F(\omega;\theta+\tfrac{1}{2}\varepsilon) - F(\omega;\theta-\tfrac{1}{2}\varepsilon)\right| > B(\omega)|\varepsilon|\right\} \tag{2.5}$$

i.e. the local Lipschitz condition fails for some integrable random bound $B > 0$, such that

(i) there exists a limiting complement event $A^*(\theta)$ which has probability 1, where we have for all $\omega \in A^*(\theta)$ that $\lim_{\epsilon\downarrow 0} \mathbf{1}_{A(\epsilon,\theta)^\complement}(\omega) = 1$,

(ii) the *critical rate* $p'_\theta := \lim_{\epsilon\downarrow 0} \frac{1}{\epsilon}\mathbb{P}(A(\varepsilon,\theta)) > 0$ exists and is finite, and

(iii) the conditional distributions given $A(\varepsilon,\theta)$ of the performance discontinuities $F(\cdot;\theta+\frac{1}{2}\varepsilon) - F(\cdot;\theta-\frac{1}{2}\varepsilon)$ are uniformly integrable, and converge in distribution as $\varepsilon \downarrow 0$ to an integrable random variable Δ_F, called the *jump*.

Remark. The definition of $A(\varepsilon,\theta)$ is symmetric in the sign of ε. One can replace the right limits with a left limits, which only results in flipping the signs of the critical rate and jump. Note that a single critical event does not need to describe *all* possible discontinuities that may occur, i.e. with equality in (2.5). Furthermore, if the critical rate were zero or the limiting event were empty, the impact of the perturbation will vanish in the limit (so that the event no longer merits being called critical).

Theorem 2.5 (variant of [17, theorems B.1, 2.3] and [15, section 2.1])**.** *Let $F(\omega;\theta)$ be a performance functional defined as above, and suppose it is a.s. piecewise continuous and differentiable. Let*

$$A(\varepsilon,\theta) = \left\{\omega \in \Omega : \left|F(\omega;\theta+\tfrac{1}{2}\varepsilon) - F(\omega;\theta-\tfrac{1}{2}\varepsilon)\right| > B(\omega)|\varepsilon|\right\}, \tag{2.6}$$

(with equality) for some given random bound $B > 0$, and suppose it fulfils the conditions of a critical event. Then

$$\frac{\partial}{\partial\theta}\mathbb{E}\left[F(\cdot;\theta)\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta}F(\cdot;\theta)\right] + \mathbb{E}\left[\Delta_F\right]p'_\theta. \tag{2.7}$$

Proof. On the complements $A(\varepsilon,\theta)^{\complement}$ the conditions are fulfilled for the existence of the pathwise derivative using the bound of $|F(\omega;\theta+\frac{1}{2}\varepsilon)-F(\omega;\theta-\frac{1}{2}\varepsilon)|\leq B(\omega)|\varepsilon|$, which is a Lipschitz condition. Hence we have a.s.

$$\lim_{\epsilon\to 0}\frac{F(\omega;\theta+\frac{1}{2}\varepsilon)-F(\omega;\theta-\frac{1}{2}\varepsilon)}{\varepsilon}\mathbf{1}_{A(\epsilon,\theta)^{\complement}}=\frac{\partial}{\partial\theta}F(\cdot;\theta) \tag{2.8}$$

and the bound implies by dominated convergence that

$$\lim_{\epsilon\to 0}\mathbb{E}\left[\frac{F(\cdot;\theta+\frac{1}{2}\varepsilon)-F(\cdot;\theta-\frac{1}{2}\varepsilon)}{\varepsilon}\mathbf{1}_{A(\epsilon,\theta)^{\complement}}\right]=\mathbb{E}\left[\frac{\partial}{\partial\theta}F(\cdot;\theta)\right]. \tag{2.9}$$

Next, the assumptions on the critical event together imply

$$\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\frac{F(\cdot;\theta+\frac{1}{2}\varepsilon)-F(\cdot;\theta-\frac{1}{2}\varepsilon)}{\varepsilon}\mathbf{1}_{A(\epsilon,\theta)}\right] \tag{2.10}$$

$$=\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\left(F(\cdot;\theta+\tfrac{1}{2}\varepsilon)-F(\cdot;\theta-\tfrac{1}{2}\varepsilon)\right)\frac{\mathbf{1}_{A(\epsilon,\theta)}}{\varepsilon}\right] \tag{2.11}$$

$$=\lim_{\epsilon\downarrow 0}\mathbb{E}\left[F(\cdot;\theta+\tfrac{1}{2}\varepsilon)-F(\cdot;\theta-\tfrac{1}{2}\varepsilon)\;\middle|\;A(\varepsilon,\theta)\right]\frac{\mathbb{P}(A(\varepsilon,\theta))}{\varepsilon}=\mathbb{E}\left[\Delta_F\right]p'_\theta \tag{2.12}$$

where uniform integrability was required for convergence of the means. Taking the limit from the left yields instead as the final line $\mathbb{E}\left[-\Delta_F\right](-p'_\theta)$ where the two signs cancel out, so that the limits are equal and the limit as $\varepsilon\to 0$ exists. Partitioning on the two sets and taking the limit yields the result. □

The preceding theorem introduces in a very abstract manner the notion of the random jump Δ_F. One may instead understand this through the concept of the *alternatives* $F(\omega;\theta^+):=\lim_{\epsilon\downarrow 0}F(\omega;\theta+\varepsilon)$ and $F(\cdot;\theta^-):=\lim_{\epsilon\downarrow 0}F(\omega;\theta-\varepsilon)$. If we let $\mathbb{Q}$ be the resulting probability measure from the weak convergence to Δ_F, which informally describes the distribution between different discontinuities conditional on a jump occurring, we can take the limits pointwise for each $\omega\in\operatorname{supp}\mathbb{Q}$, and note they exist using the integrability of jumps and piecewise continuity of F. Then

$$\mathbb{E}[\Delta_F]=\int\left(F(\omega;\theta^+)-F(\omega;\theta^-)\right)d\mathbb{Q}(d\omega). \tag{2.13}$$

At the occurrence of a critical event, two very different performances are possible simply by infinitesimally perturbing θ. Since the performance depends on θ through the sample path, these two different performances really come from two alternative limiting sample paths. With this perspective we can in many cases obtain a much more illuminating description of the alternative than the abstract definition. Note that the corrected estimator requires us to sample the alternative trajectory, so a better understanding of it will hopefully allow us to do so efficiently.

In practice, characterizing a 'complete' critical event as in theorem 2.5 can be difficult. The possibility of discontinuities can depend on the sample path

(and thus $\mathbb{Q}$ might be too complicated). We extend this theorem to a version where we have a collection of critical events, and want to apply a filtered approach. The introduction of a filtration will later allow us to exploit Markov structure in the underlying process in our characterization.

Definition 2.6. Let $\{\mathscr{F}_i\}_{i=0}^N$ be a filtration. A *filtered collection of critical events* (FCCE) for the performance $F(\cdot;\theta)$ is an event $A^*(\varepsilon,\theta)$ and a collection of events $\{A_i(\varepsilon,\theta)\}_{i=1}^N$, all pairwise disjoint, such that $A_i(\varepsilon,\theta) \in \mathscr{F}_i$ for $\varepsilon \neq 0$ where

$$A^*(\varepsilon,\theta) = \left\{\omega \in \Omega : \left|F(\omega;\theta+\tfrac{1}{2}\varepsilon) - F(\omega;\theta-\tfrac{1}{2}\varepsilon)\right| \leq B(\omega)|\varepsilon|\right\} \tag{2.14}$$

for some integrable random bound $B > 0$, and it holds

(i) there exists a limiting complement event $A^*(\theta)$ which has probability 1, where we have for all $\omega \in A^*(\theta)$ that $\lim_{\epsilon\downarrow 0} \mathbf{1}_{A^*(\epsilon,\theta)}(\omega) = 1$,

(ii) $\mathbb{P}\left(A^*(\varepsilon,\theta) \cup \bigcup_{i=1}^N A_i(\varepsilon,\theta)\right) = 1 - o(\varepsilon)$,

(iii) for all $i = 1,\dots,N$

a) the *conditional critical rate* $w_i^\theta := \lim_{\epsilon\downarrow 0} \frac{1}{\epsilon}\mathbb{E}[\mathbf{1}_{A_i(\epsilon,\theta)} \mid \mathscr{F}_{i-1}]$ a.s. exists and is uniformly bounded, and

b) the conditional distributions given $A_i(\varepsilon,\theta)$ of the conditional performance discontinuities $\mathbb{E}[F(\cdot;\theta+\frac{1}{2}\varepsilon) - F(\cdot;\theta-\frac{1}{2}\varepsilon) \mid \mathscr{F}_{i-1}]$ are uniformly integrable, and converge in distribution as $\varepsilon \downarrow 0$ to an integrable random variable $\mathbb{E}[\Delta_F^{(i)} \mid \mathscr{F}_{i-1}]$, called the *conditional jump*.

Theorem 2.7. *Let $F(\omega;\theta)$ be a performance functional defined as above, and suppose it is a.s. piecewise continuous and differentiable. Let $\{\mathscr{F}_i\}_{i=0}^N$ be a filtration, and let $A^*(\varepsilon,\theta)$, $\{A_i(\varepsilon,\theta)\}_{i=1}^N$ be an FCCE for F. Then*

$$\frac{\partial}{\partial\theta}\mathbb{E}\left[F(\cdot;\theta)\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta}F(\cdot;\theta)\right] + \mathbb{E}\left[\sum_{i=1}^N \mathbb{E}\left[\Delta_F^{(i)} \mid \mathscr{F}_{i-1}\right] w_i^\theta\right]. \tag{2.15}$$

Proof. The hypothesis (i) on the FCCE ensures that, following exactly the same argument as in theorem 2.5 but taking the limit on $A^*(\varepsilon,\theta)$ instead, the pathwise derivative is applicable on $A^*(\theta)$.

Furthermore, by hypothesis (ii) the FCCE is 'sufficiently close' to a partition to capture the perturbations; the remaining complement will have probability $o(\varepsilon)$ and thus vanish through a 'partial integration' as in theorem 2.5.

The final step is to partially integrate each term belonging to a critical event $A_i(\varepsilon,\theta)$, similar to the argument in theorem 2.5, although this time conditional on $\mathscr{F}_{i-1}$. The product converging to the critical jump and critical rate

are by FCCE hypotheses (ii) and (iii) uniformly integrable and still converge in distribution. Hence

$$\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\frac{F(\cdot;\theta+\frac{1}{2}\varepsilon)-F(\cdot;\theta-\frac{1}{2}\varepsilon)}{\varepsilon}\mathbf{1}_{A_i(\epsilon,\theta)}\right] \tag{2.16}$$

$$=\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\mathbb{E}\left[\left(F(\cdot;\theta+\tfrac{1}{2}\varepsilon)-F(\cdot;\theta-\tfrac{1}{2}\varepsilon)\right)\frac{\mathbf{1}_{A_i(\epsilon,\theta)}}{\varepsilon}\middle|\mathscr{F}_{i-1}\right]\right] \tag{2.17}$$

$$=\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\mathbb{E}\left[F(\cdot;\theta+\tfrac{1}{2}\varepsilon)-F(\cdot;\theta-\tfrac{1}{2}\varepsilon)\mid A_i(\varepsilon,\theta),\mathscr{F}_{i-1}\right]\frac{\mathbb{P}(A_i(\varepsilon,\theta)\mid\mathscr{F}_{i-1})}{\varepsilon}\right] \tag{2.18}$$

$$=\mathbb{E}\left[\mathbb{E}\left[\Delta_F^{(i)}\mid\mathscr{F}_{i-1}\right]w_i^\theta\right]. \tag{2.19}$$

Again, by a similar argument as in theorem 2.5 the left limit leads to two signs that cancel each other, so the limit as $\varepsilon\to 0$ exists. Partitioning on the FCCE and taking the limit yields the result. □

We close this subsection with a few examples to illustrate the practical use of our stratified derivative definitions, which hopefully yields some intuitive understanding for the abstractions.

Example 2.8 (Bernoulli parameter)**.** Let $(\Omega,\mathscr{F},\mathbb{P})=([0,1],\mathscr{B}_{[0,1]},\mathsf{Leb})$. Consider now $F(\omega;\theta)=\mathbf{1}_{[0,\theta]}(\omega)$, so that $F(\cdot;\theta)\sim\mathsf{Ber}(\theta)$. Obviously $\mathbb{E}[F(\cdot;\theta)]=\theta$ and hence $\frac{\partial}{\partial\theta}\mathbb{E}[F(\cdot;\theta)]=1$. But F is also piecewise constant, so the pathwise derivative is a.s. zero. We instead apply theorem 2.5.

The finite difference in performance occurs if the same ω ends up in different branches. We have a clear characterization of the critical event $A(\varepsilon,\theta)=(\theta-\frac{1}{2}\varepsilon,\theta+\frac{1}{2}\varepsilon]$ (with say $B=1$ a.s.). Then we compute the critical rate

$$\frac{\mathbb{P}((\theta-\frac{1}{2}\varepsilon,\theta+\frac{1}{2}\varepsilon])}{\varepsilon}=1\xrightarrow[\epsilon\downarrow 0]{}1=p'_\theta \tag{2.20}$$

and finally, since the performance discontinuity is 1 for all $\omega\in A(\varepsilon,\theta)$, the jump sequence is a fortiori uniformly integrable with $\Delta_F=1$ $\mathbb{Q}$-a.s. and $\mathbb{Q}=\delta_\theta$ the Dirac measure on θ, as this is the only jump point remaining in the limit. In this very simple example we could thus characterize the conditional distribution fully. By theorem 2.5 indeed $\frac{\partial}{\partial\theta}\mathbb{E}[F(\cdot;\theta)]=\mathbb{E}[\Delta]p'_\theta=1$.

Example 2.9 (Geometric parameter)**.** Consider now $X_\theta\sim\mathsf{Geom}(\theta)$ defined as $\mathbb{P}(X_\theta=k)=(1-\theta)^{k-1}\theta$ for $k=1,2,\ldots$ i.e. the number-of-trials definition. Then $\mathbb{E}[X_\theta]=1/\theta$ and hence $\frac{\partial}{\partial\theta}\mathbb{E}[X_\theta]=-1/\theta^2$. If we were to use a direct inversion of the cumulative distribution function and simulate a uniform latent, then $F(\omega;\theta)=\lceil\log(\omega)/\log(1-\theta)\rceil$. We then have countably many discontinuities in F that get arbitrarily close to each other. Hence, it seems difficult to produce a valid characterization of a complete critical event for a fixed $\varepsilon>0$.

However, we have considerable freedom on the underlying probability space. Let $(\Omega,\mathscr{F},\mathbb{P})=((0,\infty),\mathscr{B}_{(0,\infty)},\mathbb{P}_{\mathsf{Exp}})$ with the cumulative distribution function

$\mathbb{P}_{\mathsf{Exp}}(0,x] = 1-e^{-x}$, i.e. an $\mathsf{Exp}(1)$ probability measure. One can imagine this as using exponential latents in the simulation procedure. Then let $F(\omega;\theta) = \lceil -\omega/\log(1-\theta)\rceil$. We now have equidistant jumps of size -1 (because increasing θ will lower the number of trials) concentrating on points $-k\log(1-\theta)$, $k = 1,2,\ldots$. The critical event will now be the union of intervals about these points $(-k\log(1-\theta+\varepsilon/2), -k\log(1-\theta-\varepsilon/2)]$, $k = 1,2,\ldots$, and for sufficiently small ε we may assume these intervals are disjoint, so that

$$\frac{\mathbb{P}(A(\varepsilon,\theta))}{\varepsilon} = \frac{1}{\varepsilon}\sum_{k=1}^{\infty}\left(e^{k\log(1-\theta+\epsilon/2)} - e^{k\log(1-\theta-\epsilon/2)}\right) \tag{2.21}$$

$$= \frac{1}{\varepsilon}\cdot\frac{-4\varepsilon}{\varepsilon^2-4\theta^2} \xrightarrow[\epsilon\downarrow 0]{} \frac{1}{\theta^2} \tag{2.22}$$

and we once again recover the derivative as $-1/\theta^2$ by theorem 2.5. Therefore, there is considerable value in selecting the right coupling for our derivatives.

2.1.3 Score method

A different approach to stochastic derivatives is to differentiate the *measure* corresponding to X^θ rather than the simulation path itself, and we very briefly cover this as the main alternative to the estimators of the thesis. One general estimator is the *score method* [4, 18], also known as *likelihood ratio method* [19] and *REINFORCE*. It relies on the *score*, defined for a distribution with density $\pi(x;\theta)$ as

$$\frac{\partial}{\partial\theta}\log\pi(x;\theta) = \frac{\frac{\partial}{\partial\theta}\pi(x;\theta)}{\pi(x;\theta)} \tag{2.23}$$

by a classic calculus identity. Under suitable conditions this allows us to obtain an unbiased estimator; let $X^\theta \sim \pi(x;\theta)$, then

$$\frac{\partial}{\partial\theta}\mathbb{E}_{X^\theta\sim\pi(x;\theta)}[f(X^\theta)] = \frac{\partial}{\partial\theta}\int f(x)\pi(x;\theta)\,dx = \int f(x)\frac{\partial}{\partial\theta}\pi(x;\theta)\,dx \tag{2.24}$$

$$= \int f(x)\left[\frac{\partial}{\partial\theta}\log\pi(x;\theta)\right]\pi(x;\theta)\,dx \tag{2.25}$$

$$= \mathbb{E}_{X^\theta\sim\pi(x;\theta)}\left[f(X^\theta)\frac{\partial}{\partial\theta}\log\pi(X^\theta;\theta)\right] \tag{2.26}$$

by exchanging the derivative and integral. Hence the derivative can be estimated with the same sampler by changing the performance function appropriately. The conditions are fairly weak for distributions supported on $\mathbb{R}$:

Theorem 2.10 ([18, section 2])**.** *Let X^θ be distributed according to the density $\pi(x;\theta)$ parameterized by $\theta\in\Theta$ where $\Theta\subseteq\mathbb{R}$ is an open interval. Suppose that*

(i) for all $x\in\mathbb{R}$, $\pi(x;\theta)$ is continuously differentiable in θ, and

(ii) there exists an integrable function $h:\mathbb{R}\to\mathbb{R}$ such that, for all $\theta\in\Theta$, $\left|f(x)\frac{\partial}{\partial\theta}\pi(x;\theta)\right| \le h(x)$.

Then

$$\frac{\partial}{\partial\theta}\mathbb{E}_{X^\theta\sim\pi(x;\theta)}[f(X^\theta)] = \mathbb{E}_{X^\theta\sim\pi(x;\theta)}\left[f(X^\theta)\frac{\partial}{\partial\theta}\log\pi(X^\theta;\theta)\right] \tag{2.27}$$

that is, the score method yields an unbiased estimator for the derivative of the expectation.

Remark. As a special case, by taking $f \equiv 1$ constant so the left-hand side is the derivative of a constant, we see that the expected score is zero.

The score method is appealingly simple and completely general with respect to the distribution, and unlike the pathwise approach does not require a coupling. The estimator is unaffected by large magnitudes of performance derivatives, but the variance of the score method estimator grows with the dimensionality of the parameter space while the variance of the pathwise estimator is bounded by the squared Lipschitz constant of f. Therefore, no approach is uniformly better; for further reading, [4] contains some simple simulation studies that compare the variance of estimates.

2.2 POISSON PROCESSES

Poisson processes, ubiquitous in probability theory, appear in this thesis as important tools. We summarize the necessary definitions and results, adapting them principally from Çınlar & Sollenberger [20, chapter 4].

Definition 2.11 ([20, 4.7.1])**.** An *inhomogeneous Poisson process* (IPP) is the process $\{N_t\}_{t\geq 0}$ valued in $\mathbb{N}_0 = \{0,1,2,\dots\}$ defined by

(i) $N_0 = 0$,

(ii) $t \mapsto N_t$ is (weakly) increasing and càdlàg,

(iii) $t \mapsto N_t$ a.s. only increases by jumps of size one, and

(iv) for all $t, h \geq 0$, the increment $N_{t+h} - N_t$ is independent of the past $\{N_s\}_{0\leq s\leq t}$.

We define the corresponding *cumulative intensity function* (CIF) of the IPP as $\Lambda(t) = \mathbb{E}[N_t]$, $t \geq 0$. We call $t_i = \inf\{t \geq 0 : N_t = i\}$, $i \in \mathbb{N}$, the ith *arrival time* and $\tau_i = t_i - t_{i-1}$, $i \in \mathbb{N}$, the ith *interarrival time* (with $t_0 = 0$).

Note that the CIF is (weakly) increasing and right-continuous by monotone convergence ensuring it inherits these properties from the IPP [20, 5.7.2].

Definition 2.12 ([20, 4.7.11])**.** Let $\{N_t\}_{t\geq 0}$ be an IPP. If there exists $\lambda : [0,\infty) \to [0,\infty)$ such that for the corresponding CIF $\Lambda(t)$ it holds that

$$\Lambda(t) = \int_0^t \lambda(r)\,dr \tag{2.28}$$

we say that the IPP has *intensity* or *rate* $\lambda(t)$. If the rate is constant $\lambda(t) \equiv \lambda$ we say that the IPP is *homogeneous*.

Intuitively, $\lambda(t)\,dt$ represents the infinitesimal probability of an event occurring in the infinitesimal time interval $[t, t+dt]$.

Definition 2.13 ([21, 2.1])**.** Let $\Lambda : [0,\infty) \to [0,\infty)$ be (weakly) increasing and right-continuous. Then it has a *pseudoinverse* $\Lambda^{\leftarrow} : [0,\infty) \to [0,\infty)$ defined by

$$\Lambda^{\leftarrow}(\omega) = \inf\{t \geq 0 : \Lambda(t) \geq \omega\}, \tag{2.29}$$

i.e. where there are multiple candidates for t such that $\Lambda(t) = \omega$, we pick the one making $\Lambda^{\leftarrow}$ left-continuous.

Note that where Λ is locally constant we have a jump discontinuity in $\Lambda^{\leftarrow}$ and vice versa, and $\Lambda^{\leftarrow}$ is itself (weakly) increasing by construction. Furthermore, if Λ is strictly increasing, then it is invertible and the inverse Λ^{-1} agrees with $\Lambda^{\leftarrow}$. The pseudoinverse has an important theoretical use in describing the distribution of arrival times:

Proposition 2.14 ([20, 4.7.7–8])**.** *Let $\{N_t\}_{t\geq 0}$ be an IPP with continuous CIF. Then $\{N_{\Lambda^{\leftarrow}(t)}\}_{t\geq 0}$ is a homogeneous Poisson process with unit rate. In particular, for two consecutive arrival times of N it holds that $\Lambda(t_{n+1}) - \Lambda(t_n) \sim \mathsf{Exp}(1)$, and each such increment is independent of the others.*

Corollary 2.15 (variant of [20, 4.7.10])**.** *Let $\{N_t\}_{t\geq 0}$ be an IPP with continuous CIF. Then*

$$\mathbb{P}(\tau_{n+1} > t \mid t_1, \dots, t_n) = e^{-[\Lambda(t_n+t)-\Lambda(t_n)]} \tag{2.30}$$

and if the rate $\lambda(t)$ exists then we further have

$$\mathbb{P}(\tau_{n+1} > t \mid t_1, \dots, t_n) = e^{-\int_{t_n}^{t_n+t} \lambda(r)\,dr} \tag{2.31}$$

$$f_{\tau_{n+1}|t_1,\dots,t_n}(t \mid t_1, \dots, t_n) = \lambda(t_n + t)e^{-\Lambda(t_n+t)},\ t \geq 0. \tag{2.32}$$

Note that the interarrival times need not be independent of each other.

Hence, an IPP can be simulated through sampling independent *latents* $\eta_i \sim \mathsf{Exp}(1)$, $i \in \mathbb{N}$ and computing arrival times $t_i = \Lambda^{\leftarrow}(\Lambda(t_{i-1}) + \eta_i)$, $i \in \mathbb{N}$. Inversion is not the most numerically efficient way to simulate an IPP in general, as computing $\Lambda^{\leftarrow}$ may be intractable, and in practice some variant of *thinning* is often used (see e.g. [22]). We are not concerned with the details of simulation in this thesis, preferring the theoretical connection provided by inversion.

Our final set of results concern superposition and thinning, where IPPs possess a remarkable closure property.

Theorem 2.16 ([20, variant of 4.4.2 using 4.7.9])**.** *Let $\{N_t^{(1)}\}_{t\geq 0}$ be an IPP with CIF Λ_1 and let $\{N_t^{(2)}\}_{t\geq 0}$ be an IPP with CIF Λ_2, both independent of each other. Then the* superposition *$\{N_t^{(1)} + N_t^{(2)}\}_{t\geq 0}$ is an IPP with CIF $\Lambda_1 + \Lambda_2$.*

Theorem 2.17 ([22, theorem 1])**.** *Let $\{N_t\}_{t\geq 0}$ be an IPP with rate λ, and let $p(t) : [0,\infty) \to [0,1]$, interpreting $p(t)$ as the time-dependent* thinning *probability. Construct $\{N_t^p\}_{t\geq 0}$ by counting the ith arrival of N independently with probability $p(t_i)$, and construct $\{N_t^{p'}\}_{t\geq 0}$ as $N_t^{p'} = N_t - N_t^p$. Then $\{N_t^p\}_{t\geq 0}$ and $\{N_t^{p'}\}_{t\geq 0}$ are independent IPPs with rate $p(t)\lambda(t)$ and $(1-p(t))\lambda(t)$ respectively.*

2.3 PIECEWISE DETERMINISTIC MONTE CARLO

The class of Monte Carlo methods that we study in this thesis are based on piecewise deterministic Markov processes (PDMPs); thus we term them Piecewise Deterministic Monte Carlo methods to contrast with the more widely known *Markov chain Monte Carlo* (MCMC). PDMP-based samplers are a fairly recent development, with most literature appearing in the last decade.

There are several motivations behind their development. Firstly, these continuous-time non-reversible methods are more efficient, in the sense of mixing time to the target distribution, than the classic discrete-time reversible MCMC counterparts such as Metropolis-Hastings. The need for fast mixing has also motivated further development of new MCMC methods, such as the common Hamiltonian Monte Carlo. But unlike Hamiltonian Monte Carlo, PDMP-based samplers do not rely on approximate integration for the dynamics, instead truly yielding a continuous sample path. [3, 23]

Secondly, they require only the gradient of the potential of the target distribution, i.e. $\nabla\psi(x) = \nabla\left[-\log\pi(x)\right]$ where $\pi(x)$ is the target density. The normalization constant, often intractable, then disappears from the expression. In the particular case of a posterior formed by a product over data point likelihoods, $\nabla\psi(x)$ becomes a sum. This implies a computational efficiency particularly suitable for modern Big Data workloads, and there are even schemes which use only an unbiased estimate of $\nabla\psi$ from subsampling the data but still sample from the target distribution. [3, 23]

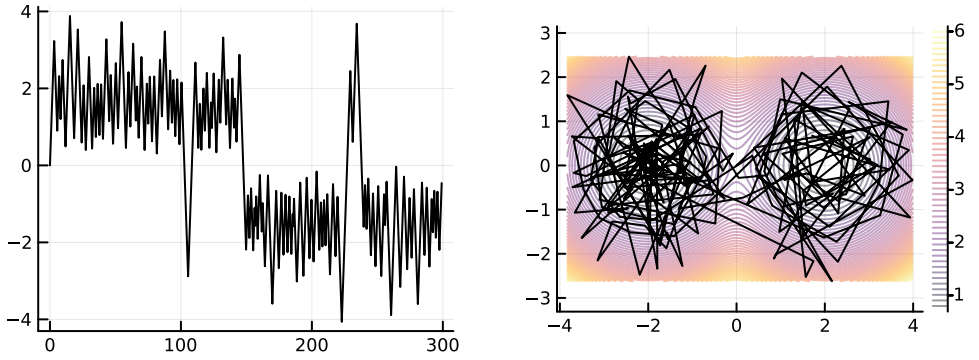
Ultimately, estimating expectations with any Monte Carlo method relies on the same fundamental result, the ergodic theorem. It allows us to pass from the invariant or stationary distribution, i.e. a distribution preserved by transitions over any finite time interval, to long term behaviour:

Theorem 2.18 (special case of [24, theorem 5.1], see also [25])**.** *Suppose $\{Z_t^\theta\}_{t\geq 0}$ has an invariant distribution μ_θ on its state space E and is ergodic, let $f : E \to \mathbb{R}$ be μ_θ-integrable, and suppose $t \mapsto f(Z_t^\theta)$ is a.s. locally integrable. Then a.s.*

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T f(Z_t^\theta)\,dt = \int_E f\,d\mu_\theta = \mathbb{E}_{X^\theta\sim\mu_\theta}[f(X^\theta)]. \tag{2.33}$$

The precise definition of *ergodic* for a Markov process beyond the intuitive 'mixing to a unique invariant distribution no matter the starting state' is out of scope here, and in general proving that a PDMP is ergodic is a non-trivial undertaking. However, if we have ergodicity we have a consistent estimator for the expectation wrt the target distribution μ_θ.

We first define piecewise deterministic Markov processes (PDMPs). Then we consider two particular PDMP samplers: the *Zig-Zag sampler* and the *Bouncy Particle sampler*. For further reading, a more in-depth but still accessible survey is that of Fearnhead, Bierkens, Pollock & Roberts [3]. We focus on the case of target distributions that are absolutely continuous wrt the Lebesgue measure and supported on $\mathbb{R}^d$, but extensions are possible, see [26, 27].



(a) 1D Zig-Zag sampler targeting a bimodal distribution (see example 3.15). Position (vertical) over time (horizontal).

(b) 2D Bouncy Particle sampler targeting a Gaussian mixture (see example 4.11). Phase portrait in the x_1x_2-plane.

Figure 2.1: Example trajectories of the two PDMP samplers we consider.

2.3.1 *Piecewise Deterministic Markov Processes (PDMPs)*

PDMPs were introduced by Davis [28, 29] to provide a common theoretical framework for, in his words, 'non-diffusion models of applied probability'. Many common types of stochastic processes, such as Markov chains (both discrete- and continuous-time), IPPs, renewal processes, and queues can be formulated as PDMPs, connecting results from these fields together. We will here consider a slightly simplified definition suitable for our purposes.

Definition 2.19 ([29, section 24])**.** A *piecewise deterministic Markov process* (PDMP) is a continuous-time càdlàg stochastic process $\{Z_t\}_{t\geq 0}$ on a state space $E \subseteq \mathbb{R}^n$ described by its *local characteristics*:

- A *drift* $\xi : E \to E$, a locally Lipschitz vector field such that we have a *flow* $\varphi(t, z_0) = z(t)$ with $z(t)$ given by the solution to the ODE

$$\frac{dz}{dt} = \xi(z),\ z(0) = z_0 \in E. \tag{2.34}$$

  This flow φ will describe the deterministic behaviour of the process between random events. We assume that the flow does not explode at a finite time unless the process would cross a boundary of E; for a more rigorous treatment of this technicality see [29, pp. 57–58].

- A *rate* $\lambda : E \to [0, \infty)$ such that for all $z_0 \in E$, there exist $\varepsilon > 0$ such that $\lambda(\varphi(t, z_0))$ is integrable on $[0, \varepsilon)$.

  Then $t \mapsto \lambda(\varphi(t, z_0))$ is a suitable rate of an IPP, from which we draw the time to the next event along a segment starting at z_0. We assume the time to the next event is a.s. finite.

- A *transition kernel* $Q(z, dz')$ such that for each $z \in E$, $Q(z, \cdot)$ is a probability measure on $[E, \mathscr{B}]$ with $\mathscr{B}$ the Borel σ-algebra on E.

The kernel determines the distribution of random transitions from state z on an event. (One often assumes that there are a.s. no self-transitions, that is for all $z \in E$ we have $Q(z, \{z\}) = 0$, to ensure λ truly describes the rate of jumps, but this is not strictly necessary.)

Now, construct $\{Z_t\}_{t\geq 0}$ recursively as follows. Given the current state Z_{t_k} (with $t_0 = 0$ and $Z_0 \in E$ fixed):

1. Sample the next interarrival time τ_{k+1} as the first arrival of the IPP with rate $\lambda(\varphi(t, Z_{t_k}))$, and let $t_{k+1} = t_k + \tau_{k+1}$.
2. For $s \in [t_k, t_{k+1})$ we let $Z_s = \varphi(s - t_k, Z_{t_k})$.
3. Sample the transition $Z_{t_{k+1}} \sim Q(\varphi(\tau_{k+1}, Z_{t_k}), \cdot)$.

We assume that the expected number of events on $[0, t]$ for fixed t is finite.

Hence, a sample trajectory of a PDMP is entirely described by the corresponding sequence of events with times and new states, since the deterministic dynamics are followed in between. This description in terms of *skeleton points* will frequently be used throughout the thesis.

The construction in the definition directly suggests a scheme for simulating a PDMP, where the main difficulty is tractable simulation of the interarrival times. Explicitly connecting the simulation to IPPs allows us to make use of the techniques discussed in the previous section. Furthermore, the construction hints that a PDMP indeed is Markov:

Theorem 2.20 ([29, theorem 25.5])**.** *A PDMP $\{Z_t\}_{t\geq 0}$ satisfies the* strong Markov property*: Let $\mathscr{F}_t = \sigma(Z_r, 0 \leq r \leq t)$ be the natural filtration, let τ be a $\mathscr{F}_t$-stopping time, and let $f : E \to \mathbb{R}$ be bounded and measurable. Then for $s \geq 0$*

$$\mathbb{E}[f(Z_{T+s})\mathbf{1}_{\{T<\infty\}} \mid \mathscr{F}_T] = \mathbb{E}[f(Z_{T+s})\mathbf{1}_{\{T<\infty\}} \mid Z_T]. \tag{2.35}$$

In particular, the event arrival times are stopping times, so the evolution of the process on each segment only depends on the state at segment start, and is independent of the trajectory up to that point.

2.3.2 *Zig-Zag sampler*

Definition 2.21 (adapting [30, section 2.4.2])**.** Let μ_θ be a probability distribution on $\mathbb{R}$ parameterized by $\theta \in \Theta$, such that μ_θ has a density π_θ with respect to the Lebesgue measure. Write $\psi(x;\theta) = -\log \pi_\theta(x)$ for the corresponding potential. The *Zig-Zag sampler* targeting μ_θ is a PDMP with state space $E = \mathbb{R} \times \{+1, -1\}$ characterized by

- drift corresponding to uniform motion (i.e. no acceleration)

$$\xi(x, v) = (v, 0), \tag{2.36}$$

- event rate

$$\lambda(x, v; \theta) = \max\left\{0, v\frac{\partial}{\partial x}\psi(x;\theta)\right\}, \tag{2.37}$$

- and transition kernel

$$Q(x, v, dx', dv') = \delta_x(dx')\delta_{-v}(dv') \tag{2.38}$$

  corresponding to deterministically flipping the sign of v on an event.

Theorem 2.22 ([23, theorem 2.2 and proposition 2.5])**.** *Let μ_θ be a probability distribution on $\mathbb{R}$ with potential $\psi \in \mathscr{C}^1$, such that $\lim_{x\to\pm\infty} \frac{\partial}{\partial x}\psi(x) = \pm\infty$. Then the one-dimensional Zig-Zag sampler targeting μ_θ has invariant distribution $\mu_\theta \otimes \mathsf{Unif}\{-1, +1\}$ and is ergodic.*

For brevity, we will write Z_t for the positional component of the sampler only.

The Zig-Zag sampler in one dimension was first proposed by Bierkens & Roberts [31]. This was then extended in [23] to multiple dimensions, best understood as parallel event 'clocks' that flip each coordinate of the velocity.

2.3.3 *Bouncy Particle sampler*

Definition 2.23 (adapting [30, section 2.4.1])**.** Let μ_θ be a probability distribution on $\mathbb{R}^d$ parameterized by $\theta \in \Theta$, such that μ_θ has a density π_θ with respect to the Lebesgue measure. Write $\psi(\boldsymbol{x};\theta) = -\log\pi_\theta(\boldsymbol{x})$ for the corresponding potential. Define the reflection update rule

$$R(\boldsymbol{x}, \boldsymbol{v}) = \boldsymbol{v} - 2\frac{\boldsymbol{v}^\mathsf{T}\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta)}{\|\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta)\|^2}\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta) \tag{2.39}$$

corresponding to an elastic reflection tangential to the gradient of the potential, whence we derive the adjective *bouncy*.

Then, the *Bouncy Particle sampler* is a PDMP with state space $E = \mathbb{R}^d \times \mathbb{R}^d$ characterized by

- drift corresponding to uniform motion (i.e. no acceleration)

$$\xi(\boldsymbol{x}, \boldsymbol{v}) = \begin{pmatrix} \boldsymbol{v} \\ 0 \end{pmatrix}, \tag{2.40}$$

- event rate

$$\lambda(\boldsymbol{x}, \boldsymbol{v};\theta) = \max\{0, \boldsymbol{v}^\mathsf{T}\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta)\} + \lambda_{\text{ref}} \tag{2.41}$$

  for some constant *refreshment rate* $\lambda_{\text{ref}} \geq 0$, and

- transition kernel

$$Q(\boldsymbol{x}, \boldsymbol{v}, d\boldsymbol{x}', d\boldsymbol{v}') = \frac{\delta_{\boldsymbol{x}}(d\boldsymbol{x}')}{\lambda(\boldsymbol{x}, \boldsymbol{v};\theta)}\left(\lambda_{\text{ref}}\rho(d\boldsymbol{v}') + \max\{0, \boldsymbol{v}^\mathsf{T}\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta)\}\delta_{R(\boldsymbol{x},\boldsymbol{v})}(d\boldsymbol{v}')\right) \tag{2.42}$$

  for the *velocity distribution* $\rho = \mathsf{N}(0, I^d)$.

A possibly more intuitive interpretation of the event rate and transition kernel is as a race between two times, each with rate corresponding to one of the terms. The transition kernel corresponds to either reflecting or refreshing depending on the winning time with the position unchanged. Such a formulation is equivalent by superposition/thinning results for IPPs. [30, section 2.4.1]

Theorem 2.24 ([32, proposition 1 and theorem 1])**.** *Let μ_θ be a probability distribution on $\mathbb{R}^d$ with potential $\psi \in \mathscr{C}^1$. Then the Bouncy Particle sampler targeting μ_θ has invariant distribution $\mu_\theta \otimes \rho$. If additionally $\lambda_{\text{ref}} > 0$, the sampler is ergodic.*

Without refreshments one can find counterexamples where the sampler gets stuck, unable to explore the whole space (see [32, section 4.1]). For brevity, as in the previous section, we will write $\boldsymbol{Z}_t$ for the positional component of the sampler only.

It will be convenient for us to augment the sequence of skeleton points with information about whether a reflection or a refreshment occurred on the segment. This can be done formally by moving to the state space $E' = E \times \{\text{reflect}, \text{refresh}\}$, letting the drift be zero on the new component, and adjusting the transition kernel to transition into the appropriate state. The main reason this is not part of the definition is that it adds unnecessary complications in formulating the invariant distribution.

The Bouncy Particle sampler was first proposed by Bouchard-Côté, Vollmer & Doucet [32]. It may be interesting to note that it in one dimension is equivalent to the Zig-Zag sampler (without refreshments).

3 DIFFERENTIATION OF THE ZIG-ZAG SAMPLER IN ONE DIMENSION

The first concrete problem in this thesis concerns the Zig-Zag sampler in one dimension. Given a target distribution μ_θ on $\mathbb{R}$ parameterized by θ, and a function $f : \mathbb{R} \to \mathbb{R}$, our goal is to use a trajectory of the sampler to estimate $\frac{\partial}{\partial\theta}\mathbb{E}_{X^\theta\sim\mu_\theta}[f(X^\theta)]$. The extension to gradients of parameter vectors is naturally done through all the directional derivatives along coordinate axes.

Our standard setting in this chapter is the following:

Assumption 3.1. Let $\{Z_t^\theta\}_{t=0}^T$ be a Zig-Zag sampler targeting a distribution μ_θ parameterized by $\theta \in \Theta \subseteq \mathbb{R}$ (in particular, μ_θ satisfies the assumptions in theorem 2.22). Let $f : \mathbb{R} \to \mathbb{R}$ be such that the expectation $\mathbb{E}_{X^\theta\sim\mu_\theta}[f(X^\theta)]$ exists and is differentiable wrt θ. Furthermore, the potential $\psi(x;\theta)$ corresponding to μ_θ is $\mathscr{C}^2$. We identify trajectories of the sampler with the sequence of skeleton points $\{(t_i, x_i, v_i)\}_{i=0}^N$.

The outline of this chapter is as follows: To obtain such an estimator, we begin by constructing a coupling for derivative analysis. We then split the problem into differentiating the integral that estimates the expectation from the sampler, and differentiating the actual trajectory segment-by-segment through the pathwise derivatives of events. This suffices for unimodal μ_θ, where we in fact can prove consistency. In the general case, we must in addition handle jumps in the perturbed trajectory where the difference in expectation is not Lipschitz by a stratified derivative. This requires us to develop understanding of when such jumps occur, and how to sample them. The end result is an unbiased estimator of the derivative of the expected performance. We close this chapter with some examples and simulations.

3.1 SHADOWING COUPLING

As put forward in the discussion of stochastic derivatives, we require a *coupling* between the two trajectories to produce single-run estimators for the gradient wrt θ of the expectation. The ideal result in the case of a finite perturbation ε, in the sense of having stable variance properties, should resemble a copy of the trajectory *almost* following the other one, hence our name *shadowing coupling*.

The 'randomness' of the Zig-Zag sampler comes entirely from the underlying IPP that drives the event times, since the dynamics are deterministic and all

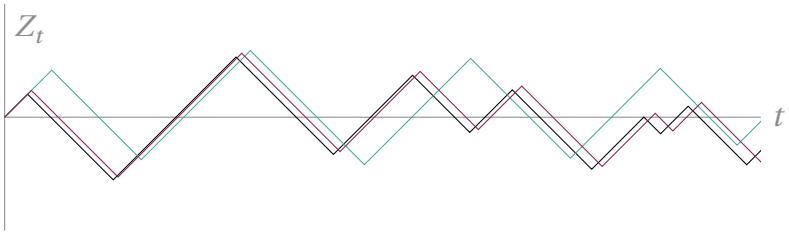


Figure 3.1: A shadow trajectory (purple) obtained through coupling with a primal trajectory (black), for a perturbation of a small but finite ε. Compare with the completely independent new trajectory (teal).

transitions are reflections. If these are driven by the same *latent* independent randomness we hope to obtain the necessary coupling between the primal and shadow trajectories (fig. 3.1). The simplest such connection is provided by the inversion method. Our shadowing coupling is then achieved by inverting the inversion to provide an (a.s.) differentiable connection:

Definition 3.2 (Shadowing coupling, Zig-Zag)**.** Let $\{Z_t^\theta\}_{t=0}^{T_\theta}$ denote the primal trajectory, which is perfectly described by the sequence of interarrival times $\{\tau_i\}_{i=1}^N$ and the starting state. Corresponding to perturbing θ by ε, form the *shadow trajectory* $\{Z_t^{\theta+\epsilon}\}_{t=0}^{T_{\theta+\epsilon}}$ with the new sequence of interarrival times $\{\tau_i'\}_{i=1}^N$ defined by

$$\tau_i' = \Lambda_i^{\leftarrow}(\Lambda_i(\tau_i;\theta);\theta+\varepsilon). \tag{3.1}$$

where Λ_i is the CIF of the corresponding IPP for the ith event. (Note that the dependency of the CIF on θ is nontrivial, as it also depends on the state at the start of each segment which itself may depend on θ.) The one-to-one relationship between events implies that the shadow trajectory has the same number of events N as the primal trajectory (which is a priori random), and hence may have a perturbed total duration $T_{\theta+\epsilon}$.

3.2 SMOOTH CASE

We can now use the shadowing coupling to analyse the sensitivity of the estimates, ostensibly taking the limit between the trajectories as $\varepsilon \to 0$. This is done in two stages: first pathwise differentiating the *expectation functional* (the integral or performance) with respect to the trajectory, and then pathwise differentiating the trajectory with respect to the parameter. Hence, this section works under an important assumption:

Assumption 3.3. The CIF of the IPP on a segments in the sampler is invertible in time on $(0,\infty)$, and the inverse is differentiable in θ. Intuitively speaking, infinitesimal perturbations of the parameter translate to infinitesimal perturbations of the trajectory.

This is a fairly strong assumption which does not hold in many common cases, and to avoid it requires extensions deferred to the next section, but it holds in e.g. the case of a unimodal target distribution.

To build intuition, consider simply stretching the process trajectory infinitesimally at each reflection as in fig. 3.2. Since we must keep unit velocity, this corresponds to an infinitesimal delay ε at each reflection, incurring the same delay when returning 'home' for the next reflection. The need to return is due

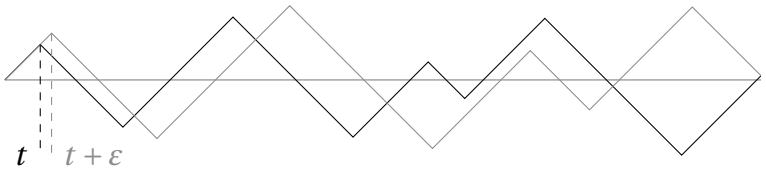


Figure 3.2: Illustration of uniformly delaying a trajectory by ε.

to the inability to reflect when moving towards the mode, which the process does immediately after reflecting. The result is that we accumulate the delay in arrival times over time; a local perturbation has global effects. However, the relative position of the reflection points only change by $\pm\varepsilon$, and this motivates focusing on the sensitivities of these points, as is done in the analysis that follows.

3.2.1 *Derivative of expectation functional*

First, given the sensitivities of the reflection points, how does the perturbation affect the expectation functional?

Lemma 3.4. *Under assumption 3.1, suppose the sensitivities of the reflection positions $\frac{\partial x_i}{\partial \theta}$ (which are in fact pathwise derivatives of random variables) are well-defined and exist (a.s.) for all i.*

Then the (a.s.) derivative wrt θ of the functional estimating the expectation

$$F(Z_t^\theta) = \frac{1}{T_\theta} \int_0^{T_\theta} f(Z_t^\theta)\, dt \tag{3.2}$$

is given by

$$\frac{\partial}{\partial \theta} F(Z_t^\theta) = \frac{1}{T_\theta} \sum_{i=1}^{N} (2 - \mathbf{1}_{\{N\}}(i)) \left(f(x_i) - F(Z_t^\theta) \right) \left(-\frac{1}{v_i} \frac{\partial x_i}{\partial \theta} \right). \tag{3.3}$$

Proof. Use the shadowing coupling (definition 3.2) in order to consider the difference in the functional between the primal trajectory $\{Z_t^\theta\}_{t=0}^{T_\theta}$ and the

shadow trajectory $\{Z_t^{\theta+\epsilon}\}_{t=0}^{T_{\theta+\epsilon}}$ for some ε. Note that $T = T_\theta$ since the coupling fixes the number of events and hence the total time may depend on θ.

$$F(Z_t^{\theta+\epsilon}) - F(Z_t^\theta) = \frac{1}{T_{\theta+\epsilon}} \int_0^{T_{\theta+\epsilon}} f(Z_t^{\theta+\epsilon})\, dt - \frac{1}{T_\theta} \int_0^{T_\theta} f(Z_t^\theta)\, dt \tag{3.4}$$

$$= \frac{1}{T_{\theta+\epsilon}} \left(\int_0^{T_{\theta+\epsilon}} f(Z_t^{\theta+\epsilon})\, dt - \int_0^{T_\theta} f(Z_t^\theta)\, dt \right) \tag{3.5}$$

$$+ \left(\frac{1}{T_{\theta+\epsilon}} - \frac{1}{T_\theta} \right) \int_0^{T_\theta} f(Z_t^\theta)\, dt \tag{3.6}$$

The difference in each reflection position is $\frac{\partial x_i}{\partial \theta}\varepsilon + o(\varepsilon)$. This contributes to a difference in the integral on each adjacent segment, with special cases of the starting point being fixed and the ending point contributing half, by the fundamental theorem of calculus and using that $v_i = -v_{i-1}$:

$$\frac{\partial}{\partial \theta} \int_{t_{i-1}}^{t_i} f(x_{i-1} + v_{i-1} r)\, dr = \frac{\partial}{\partial \theta} \left[\frac{1}{v_{i-1}} \int_{x_{i-1}}^{x_i} f(u)\, du \right] \tag{3.7}$$

$$= -\frac{1}{v_i} \frac{\partial x_i}{\partial \theta} f(x_i) - \frac{1}{v_{i-1}} \frac{\partial x_{i-1}}{\partial \theta} f(x_{i-1}). \tag{3.8}$$

Outside of the reflections, the coupling ensures that the shadow trajectory follows the primal trajectory perfectly albeit with a time shift, thus cancelling each other's contributions. Hence the first term in eq. (3.5) is

$$\frac{1}{T_{\theta+\epsilon}} \sum_{i=1}^{N} (2 - \mathbf{1}_{\{N\}}(i)) f(x_i) \left(-\frac{1}{v_i} \frac{\partial x_i}{\partial \theta} \varepsilon + o(\varepsilon) \right). \tag{3.9}$$

Similarly, the unit speed of the Zig-Zag implies that the shadow trajectory reflection times are perturbed by the same amount, so that

$$T^{\theta+\epsilon} = T^\theta + \sum_{i=1}^{N} (2 - \mathbf{1}_{\{N\}}(i)) \left(-\frac{1}{v_i} \frac{\partial x_i}{\partial \theta} \varepsilon + o(\varepsilon) \right) \tag{3.10}$$

which may be inserted into eq. (3.6), and by the chain rule

$$\frac{1}{T_{\theta+\epsilon}} - \frac{1}{T_\theta} = -\frac{1}{(T_\theta)^2} \sum_{i=1}^{N} (2 - \mathbf{1}_{\{N\}}(i)) \left(-\frac{1}{v_i} \frac{\partial x_i}{\partial \theta} \varepsilon + o(\varepsilon) \right). \tag{3.11}$$

Now divide by ε and take $\varepsilon \to 0$ in both expressions. □

Remark. Note that the result holds even if f is not differentiable, as in our common example of estimating the probability of some event A, where we have $f \equiv \mathbf{1}_A$. Furthermore, the derivation does not depend on the form of the rate, abstracting it away behind the point sensitivities.

3.2.2 *Pathwise derivative of segments*

We continue by determining the reflection point sensitivities.

Lemma 3.5. *Under assumption 3.1, then (a.s.)*

$$-\frac{1}{v_i}\frac{\partial x_i}{\partial\theta} = \left(-\frac{1}{v_{i-1}}\frac{\partial x_{i-1}}{\partial\theta}\right)\left(-\frac{\lambda(x_{i-1},v_{i-1};\theta)}{\lambda(x_i,v_{i-1};\theta)}\right) - \frac{\frac{1}{v_{i-1}}\int_{x_{i-1}}^{x_i}\frac{\partial\lambda}{\partial\theta}(y,v_{i-1};\theta)\,dy}{\lambda(x_i,v_{i-1};\theta)} \tag{3.12}$$

for $i = 1,2,\dots,N$.

Proof. We consider the effect of a perturbation segment by segment; by appeal to the strong Markov property, the starting point of the segment is sufficient to reason about the evolution on the segment. A perturbation in the end point of a given segment can therefore come from perturbations in the starting point of the segment and/or perturbations in the actual dynamics along the segment.

For the ith segment, denote by

$$L_i(t;\theta) = \int_0^t \lambda(x_{i-1} + v_{i-1}r, v_{i-1};\theta)\,dr \tag{3.13}$$

the CIF of the corresponding IPP with the starting point fixed. Let L_i^{-1} be the corresponding inverse in time. Assumption 3.3 implies that L_i is strictly increasing and differentiable, hence truly invertible with differentiable inverse. (Without this assumption, the preceding only holds a.e., but we can still proceed in those cases, albeit obtaining a weaker result of a.s. existence.)

Suppose we have an arbitrary perturbation ε of θ such that the previous reflection is delayed by δ_ϵ, instead occurring at $t_{i-1}+\delta_\epsilon$. A direct consequence is the perturbation of the starting point of the segment to $x_{i-1} - v_{i-1}\delta_\epsilon$. (The sign comes from the fact that v_{i-1} is the velocity after the reflection, so that $-v_{i-1} = v_{i-2}$, the actual incoming velocity. Working with the perturbation in time leads to more concise notation in what follows.) Hence

$$\Lambda_i(t;\theta+\varepsilon) = \int_0^t \lambda(x_{i-1} + v_{i-1}(r-\delta_\epsilon), v_{i-1};\theta+\varepsilon)\,dr \tag{3.14}$$

is the CIF of the resulting perturbed IPP.

Now

$$\begin{aligned}
\Lambda_i(t;\theta+\varepsilon) &= \int_0^t \lambda(x_{i-1} + v_{i-1}(r-\delta_\epsilon), v_{i-1};\theta+\varepsilon)\,dr && (3.15)\\
&= \int_{-\delta_\epsilon}^{t-\delta_\epsilon} \lambda(x_{i-1} + v_{i-1}r, v_{i-1};\theta+\varepsilon)\,dr && (3.16)\\
&= L_i(t-\delta_\epsilon;\theta+\varepsilon) + \int_{-\delta_\epsilon}^{0} \lambda(x_{i-1} + v_{i-1}r, v_{i-1};\theta+\varepsilon)\,dr && (3.17)
\end{aligned}$$

where we extended the domain of L_i appropriately if $\delta_\epsilon > 0$, and so inverting $\omega = \Lambda_i(t;\theta+\varepsilon)$ in time yields

$$\Lambda_i^{-1}(\omega;\theta+\varepsilon) = L_i^{-1}\left(\omega - \int_{-\delta_\epsilon}^{0} \lambda(x_{i-1}+v_{i-1}r, v_{i-1};\theta+\varepsilon)\,dr\,;\theta+\varepsilon\right) + \delta_\epsilon. \quad (3.18)$$

We must now expand this in terms of the perturbation ε in two steps. First, nothing that δ_ϵ depends on ε, let

$$I(\varepsilon) = -\int_{-\delta_\epsilon}^{0} \lambda(x_{i-1}+v_{i-1}r, v_{i-1};\theta+\varepsilon)\,dr \quad (3.19)$$

so that

$$I(\varepsilon) = \int_{0}^{-\delta_\epsilon} \lambda(x_{i-1}+v_{i-1}r, v_{i-1};\theta+\varepsilon)\,dr = -\lambda(x_{i-1}, v_{i-1};\theta)\delta_\epsilon + o(\varepsilon) \quad (3.20)$$

by application of the Leibniz rule (theorem A.2, using that the inner θ-derivative vanishes as $\varepsilon \to 0$). Hence, by the chain rule

$$\Lambda_i^{-1}(\omega;\theta+\varepsilon) = L_i^{-1}(\omega;\theta) + \frac{\partial L_i^{-1}}{\partial\omega}(\omega;\theta)\cdot I(\varepsilon) + \frac{\partial L_i^{-1}}{\partial\theta}(\omega;\theta)\varepsilon + \delta_\epsilon + o(\varepsilon) \quad (3.21)$$

$$= L_i^{-1}(\omega;\theta) + \left(1 - \frac{\partial L_i^{-1}}{\partial\omega}(\omega;\theta)\lambda(x_{i-1}, v_{i-1};\theta)\right)\delta_\epsilon + \frac{\partial L_i^{-1}}{\partial\theta}(\omega;\theta)\varepsilon + o(\varepsilon) \quad (3.22)$$

By application of the inverse function theorem (theorem A.1) to $t = L_i^{-1}(\omega;\theta)$

$$\frac{\partial L_i^{-1}}{\partial\omega}(\omega;\theta) = \frac{1}{\frac{\partial L_i}{\partial t}(L_i^{-1}(\omega;\theta);\theta)} = \frac{1}{\lambda(x_{i-1}+v_{i-1}t, v_{i-1};\theta)} \quad (3.23)$$

$$\frac{\partial L_i^{-1}}{\partial\theta}(\omega;\theta) = \frac{-\frac{\partial L_i}{\partial\theta}(L_i^{-1}(\omega;\theta);\theta)}{\frac{\partial L_i}{\partial t}(L_i^{-1}(\omega;\theta);\theta)} = \frac{-\frac{1}{v_{i-1}}\int_{x_{i-1}}^{x_{i-1}+v_{i-1}t}\frac{\partial\lambda}{\partial\theta}(y, v_{i-1};\theta)\,dy}{\lambda(x_{i-1}+v_{i-1}t, v_{i-1};\theta)} \quad (3.24)$$

where the last equality performed an interchange of the integral and derivative, valid even though $\frac{\partial\lambda}{\partial\theta}$ may only exist almost everywhere by the fact that λ is continuously zero at those points where $\frac{\partial\lambda}{\partial\theta}$ may fail to exist.

The final detail is to determine δ_ϵ. One may think it is equal to the previous interarrival time perturbation, but that is too simplistic; the previous segment can itself be perturbed by both sources of perturbation. By definition of the reflection point sensitivity, as the point must move in space and time simultaneously, the delay is precisely

$$\delta_\epsilon = -\frac{1}{v_{i-1}}\frac{\partial x_{i-1}}{\partial\theta}\varepsilon + o(\varepsilon). \quad (3.25)$$

and now eqs. (3.22) to (3.25) combine to the desired derivative $\frac{\partial\Lambda_i^{-1}}{\partial\theta}(s;\theta)$ which can be used with the shadowing coupling (definition 3.2) to determine the interarrival time sensitivities accounting for both sources of perturbations as

$$\frac{\partial\tau_i}{\partial\theta} = \left(-\frac{1}{v_{i-1}}\frac{\partial x_{i-1}}{\partial\theta}\right)\left(1 - \frac{\lambda(x_{i-1}, v_{i-1};\theta)}{\lambda(x_i, v_{i-1};\theta)}\right) - \frac{\frac{1}{v_{i-1}}\int_{x_{i-1}}^{x_i}\frac{\partial\lambda}{\partial\theta}(y, v_{i-1};\theta)\,dy}{\lambda(x_i, v_{i-1};\theta)}. \quad (3.26)$$

Finally, the definition of the segments $x_i = x_{i-1} + v_{i-1}\tau_i$ yields a recursive expression for the sensitivity

$$-\frac{1}{v_i}\frac{\partial x_i}{\partial \theta} = \frac{1}{v_{i-1}}\frac{\partial x_i}{\partial \theta} = \frac{1}{v_{i-1}}\frac{\partial x_{i-1}}{\partial \theta} + \frac{\partial \tau_i}{\partial \theta} \tag{3.27}$$

with the initial position x_0 constant and having zero sensitivity. Inserting (3.26) recovers the result. □

Remark. Only a single step here used explicitly that we are targeting the Zig-Zag rate, and with some work it could be extended to a more general result; additional conditions might be required for the Leibniz rule in eq. (3.24) to apply, without which one might be forced to leave the derivative outside the integral.

This completes the work required to have an estimator under assumption 3.3, as we can sequentially compute point sensitivities with lemma 3.5 that are then used in lemma 3.4 to provide an estimate for the derivative. The next step would be to prove unbiasedness of the pathwise derivative, by noting that $\Lambda^{\leftarrow}$ is Lipschitz in θ by assumption 3.3 and applying theorem 2.2 (assuming e.g. that the θ of interest are a compact set, which poses no practical restriction). We omit the details in favour of a later result.

Nevertheless, such pathwise unbiasedness results are subtle, implying only that the finite-horizon pathwise derivative is an unbiased estimator of the derivative of the finite-horizon expectation. This is desirable but not enough for our final goal, since clearly a very short trajectory starting from a fixed point need not mix and produce a good estimate for the target expectation. We also want a strong consistency result, i.e. that the pathwise derivative estimator a.s. converges to the true derivative of the expectation; the error is then only a matter of how fast it converges.

Theorem 3.6. *Under assumptions 3.1 and 3.3, the estimate of the derivative of the expectation functional derived from lemmata 3.4 and 3.5 is strongly consistent, that is*

$$\frac{\partial}{\partial \theta}\mathbb{E}_{X^\theta \sim \mu_\theta}[f(X^\theta)] = \lim_{T\to\infty}\frac{\partial}{\partial \theta}F(Z_T^\theta) \quad a.s. \tag{3.28}$$

where $F(Z_T^\theta)$ is the finite-horizon estimator up to T.

Proof. The key idea is to cut the trajectory into a sequence of *excursions* as illustrated in fig. 3.3, hence establishing a regenerative structure, which allows us to apply the strong law of large numbers to 'interchange' the derivative and limit in the ergodic theorem. This would be sufficient to apply theorem 2.3, but we will for illustrative purposes prove it directly.

Let $m_\theta = \max\{x \in \mathbb{R} : \frac{\partial \psi}{\partial x}(x;\theta) = 0\}$; for unimodal target distributions this corresponds to the mode. By hypothesis, m_θ exists and the sampler will return to m_θ infinitely often. Denote the times of upcrossings of m_θ by $\{c_{i,\theta}\}_{i=1}^{\infty}$, and note that they are (a.s. finite) stopping times for the sampler. The cut points

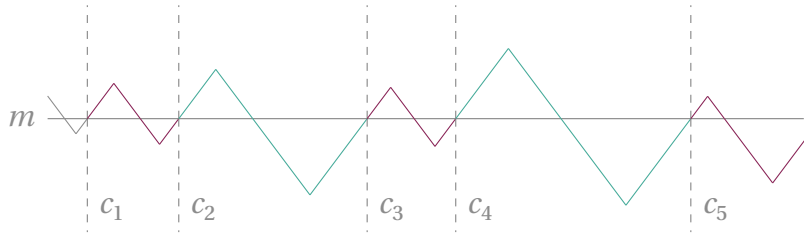


Figure 3.3: Cutting a sample trajectory into i.i.d. excursions after burn-in.

for the excursions correspond to the times $c_{i,\theta}$. Each such excursion corresponds to a short trajectory of the Zig-Zag sampler with starting state $(m_\theta, +1)$ that is run until stopped at the next upcrossing of m_θ. By the strong Markov property, each excursion is independent and identically distributed. Let I_i be the integral along the ith excursion, i.e.

$$I_i = \int_{c_{i,\theta}}^{c_{i+1,\theta}} f(Z_t^\theta)\, dt \tag{3.29}$$

and let $T_i = c_{i+1,\theta} - c_{i,\theta}$ be the elapsed time. If the starting state is not $(m_\theta, +1)$, a 'burn-in' period until the first upcrossing is necessary, and we denote its integral by $B = \int_0^{c_{1,\theta}} f(Z_t^\theta)\, dt$.

We are now ready to connect the regenerative structure to the ergodic theorem. Consider a sequence of N excursions with possible burn-in. By theorem 2.22, $\mathbb{E}_{X^\theta \sim \mu_\theta}[f(X^\theta)] = \lim_{N\to\infty} F(Z_{c_{N+1,\theta}})$ a.s., while at the same time

$$F(Z_{c_{N+1,\theta}}) = \frac{1}{c_{N+1,\theta}} \int_0^{c_{N+1,\theta}} f(Z_t^\theta)\, dt = \frac{B + \sum_{i=1}^N I_i}{c_{1,\theta} + \sum_{i=1}^N T_i} \tag{3.30}$$

$$= \frac{B + \sum_{i=1}^N I_i}{N} \cdot \frac{N}{c_{1,\theta} + \sum_{i=1}^N T_i} \xrightarrow[N\to\infty]{\text{a.s.}} \frac{\mathbb{E} I_1}{\mathbb{E} T_1} \tag{3.31}$$

by the strong law of large numbers. Using the regenerative structure, we have thus related the true expectation to expectations of special finite-horizon estimators, allowing us to proceed with proving the consistency through the use of finite-horizon unbiasedness.

We now differentiate pathwise wrt θ. Clearly $\frac{\partial T_i}{\partial \theta}$ are i.i.d., since they are based on the separate excursions. It is less obvious that $\frac{\partial I_i}{\partial \theta}$ are i.i.d., but note from the form of the pathwise estimators in lemmata 3.4 and 3.5 that only parts moving away from m_θ contribute to the derivative, each such excursion containing two terms dependent on each reflection point. Furthermore, as the performance is Lipschitz and the pathwise derivative is then unbiased by theorem 2.2, we remark that by extension these finite horizon pathwise derivatives are unbiased estimates of the corresponding derivatives as well.

Hence, considering a sequence of N excursions with possible burn-in as before, we have the pathwise derivatives

$$\frac{\partial}{\partial \theta} \int_0^{c_{N+1,\theta}} f(Z_t^\theta)\, dt = \frac{\partial B}{\partial \theta} + \sum_{i=1}^N \frac{\partial I_i}{\partial \theta}, \tag{3.32}$$

$$\frac{\partial}{\partial\theta}c_{N+1,\theta} = \frac{\partial}{\partial\theta}c_{1,\theta} + \sum_{i=1}^{N}\frac{\partial T_i}{\partial\theta}, \tag{3.33}$$

so that

$$\frac{\partial}{\partial\theta}F(Z^\theta_{c_{N+1,\theta}}) = \frac{1}{c_{N+1,\theta}}\int_0^{c_{N+1,\theta}} f(Z_t^\theta)\,dt \tag{3.34}$$

$$= \frac{1}{c_{N+1,\theta}}\frac{\partial}{\partial\theta}\int_0^{c_{N+1,\theta}} f(Z_t^\theta)\,dt - \frac{\frac{\partial}{\partial\theta}c_{N+1,\theta}}{(c_{N+1,\theta})^2}\int_0^{c_{N+1,\theta}} f(Z_t^\theta)\,dt \tag{3.35}$$

$$= \frac{\frac{\partial B}{\partial\theta} + \sum_{i=1}^N \frac{\partial I_i}{\partial\theta}}{c_{1,\theta} + \sum_{i=1}^N T_i} - \frac{\left(\frac{\partial}{\partial\theta}c_{1,\theta} + \sum_{i=1}^N \frac{\partial T_i}{\partial\theta}\right)\left(B + \sum_{i=1}^N I_i\right)}{\left(c_{1,\theta} + \sum_{i=1}^N T_i\right)^2} \tag{3.36}$$

$$\xrightarrow[N\to\infty]{\text{a.s.}} \frac{\mathbb{E}\frac{\partial I_1}{\partial\theta}}{\mathbb{E}\,T_1} - \frac{\mathbb{E}\frac{\partial T_1}{\partial\theta}\mathbb{E}\,I_1}{(\mathbb{E}\,T_1)^2} = \frac{\frac{\partial}{\partial\theta}\mathbb{E}\,I_1}{\mathbb{E}\,T_1} - \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}\,T_1\right)\cdot\mathbb{E}\,I_1}{(\mathbb{E}\,T_1)^2} = \frac{\partial}{\partial\theta}\frac{\mathbb{E}\,I_1}{\mathbb{E}\,T_1} \tag{3.37}$$

by the strong law of large numbers applied similarly as before, and the chain rule in reverse using finite-horizon unbiasedness. However, we showed above that $\frac{\mathbb{E}I_1}{\mathbb{E}T_1} = \mathbb{E}_{X^\theta\sim\mu_\theta}[f(X^\theta)]$ and so we have established the consistency along the discrete sequence. Consistency for the continuous limit $T\to\infty$ follows by separately sandwiching numerator and denominator between full sets of excursions for any given T, and thus we are done. □

3.3 GENERAL CASE

Not every reasonable IPP that arises for the Zig-Zag sampler satisfies assumption 3.3 of invertible CIF Λ with the inverse smooth in θ. It suffices that there exist regions where the rate λ becomes zero again, e.g. if the target is multimodal. Then both parts of the assumption are violated: although Λ is not uniquely invertible, we may form the pseudoinverse $\Lambda^{\leftarrow}$ and the inversion method remains valid, but this results in jump discontinuities in $\Lambda^{\leftarrow}$ and we now only have a.s. differentiability.

Then, lemmata 3.4 and 3.5 only hold a.s., and this is not sufficient in general to construct an unbiased estimator: an infinitesimal perturbation of θ can then lead to a finite change in the expectation by 'pushing' reflections across these regions where the process cannot reflect, as illustrated in fig. 3.4. We call these regions where reflections cannot occur *tunnels*. Even though this in the limit happens with probability zero the derivative can then fail to exist. The main difficulty is that we never observe this probability zero event in practice when sampling our primal trajectory. Therefore, we require more tools to capture this perturbation, which we now develop in this section, beginning with a concrete example and culminating in a stratified stochastic derivative estimator.

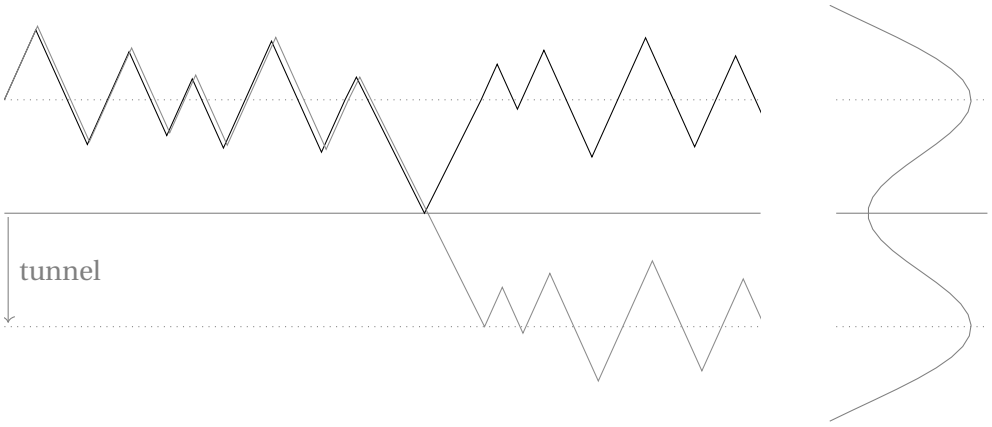


Figure 3.4: 'Losing the coupling' due to a tunnel. A prototypical target density is sketched to the right, showing how two modes can create a tunnel.

Example 3.7 (Truncated IPP)**.** Consider the piecewise constant rate $\lambda(t;\theta) = \mathbf{1}_{[0,\theta]\cup(2\theta,\infty)}(t)$, $\theta > 0$. We obtain the CIF and pseudoinverse:

$$\Lambda(t;\theta) = \int_0^t \lambda(t';\theta)\,dt' = \begin{cases} t, & 0 \le t \le \theta \\ \theta, & \theta < t \le 2\theta \\ t-\theta, & 2\theta < t \end{cases} \tag{3.38}$$

$$\Lambda^{\leftarrow}(\omega;\theta) = \begin{cases} \omega, & 0 \le \omega \le \theta \\ \omega+\theta, & \theta < \omega \end{cases} \tag{3.39}$$

shown in fig. 3.5. Note that the pseudoinverse has a jump discontinuity at θ with jump size θ, exemplifying the two ways θ can affect such discontinuities.

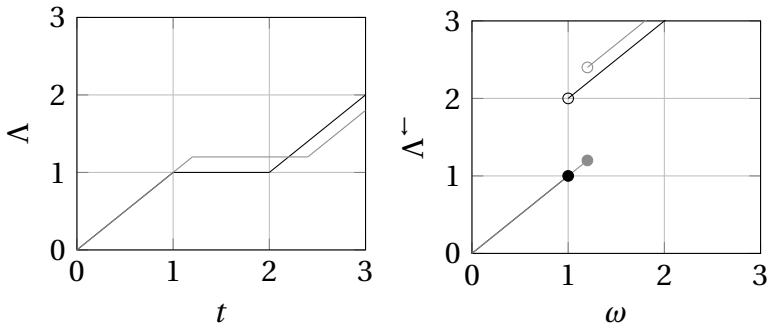


Figure 3.5: CIF and pseudoinverse in example 3.7, for two different θ.

We can in the standard manner compute the expectation of the corresponding first arrival time and then differentiate:

$$\mathbb{E}\tau = \int_0^\theta e^{-t}\,dt + \int_\theta^{2\theta} e^{-\theta}\,dt + \int_{2\theta}^\infty e^{-t+\theta}\,dt = 1 + \theta e^{-\theta} \tag{3.40}$$

$$\frac{\partial}{\partial\theta}\mathbb{E}\tau = (1-\theta)e^{-\theta}. \tag{3.41}$$

However, attempting to directly use the pathwise derivative obtained through inversion (with the derivative that exists a.s.) fails:

$$\frac{\partial \tau}{\partial \theta} = \frac{\partial \Lambda^{\leftarrow}}{\partial \theta}(\Lambda(\tau;\theta);\theta) = \mathbf{1}_{(\theta,\infty)}(\Lambda(\tau;\theta)) = \mathbf{1}_{(2\theta,\infty)}(\tau) \tag{3.42}$$

$$\mathbb{E}\left[\frac{\partial \tau}{\partial \theta}\right] = \int_{2\theta}^{\infty} e^{-t+\theta}\, dt = e^{-\theta} \neq \frac{\partial}{\partial \theta}\mathbb{E}\,\tau. \tag{3.43}$$

That is, even though we have a.s. differentiability, this is not sufficient to interchange expectation and derivative. The small perturbation becomes a large difference if we 'switch sides' of the jump discontinuity. However, simply from fig. 3.5 one can see that the stratified derivative will provide the missing piece of the puzzle, because the jump is of size $-\theta$ with critical rate $e^{-\theta}$.

3.3.1 *Tunnels and teleportation*

We present an exact definition of tunnel and study the behaviour of trajectories interacting with them for finite perturbations ε.

Definition 3.8 (Tunnel)**.** Given an IPP with CIF $\Lambda(t)$, a *tunnel* $\mathfrak{T}$ is a compact interval $[a,b] \subset (0,\infty)$ in time such that

- $\Lambda(b) - \Lambda(a) = 0$,
- $\forall \varepsilon > 0 \;\; \Lambda(b+\varepsilon) - \Lambda(b) > 0$,
- $\forall \varepsilon \in (0,a] \;\; \Lambda(a) - \Lambda(a-\varepsilon) > 0$.

This precise description makes clear that a tunnel is not only an interval of zero rate, but also has *openings* at both ends a, b so that reflections can potentially occur both before and after. Hence there is no tunnel containing time zero, even though the rate commonly is zero immediately after an reflection, because no reflections can be pushed earlier than the start of the segment. Of course, for the parameterized rates $\lambda(x,v;\theta)$ that we consider where the time enters through the sampler state, a tunnel $\mathfrak{T}(\theta)$ is also sensitive to the parameter, and one can identify tunnel openings with two zeros of $\frac{\partial \psi}{\partial x}(x;\theta)$. The sign of v determines which zero corresponds to $\inf \mathfrak{T}(\theta) = a$.

For a given CIF $\Lambda(t)$ we can identify each tunnel in the corresponding IPP with one *jump discontinuity* of $\Lambda^{\leftarrow}(\omega)$. Indeed, if we have a tunnel $\mathfrak{T}$ and let $\omega_{\mathfrak{T}} = \Lambda(\inf \mathfrak{T})$, then it holds that $\Lambda^{\leftarrow}(\omega_{\mathfrak{T}}) = \inf \mathfrak{T}$ and $\lim_{\omega \downarrow \omega_{\mathfrak{T}}} \Lambda^{\leftarrow}(\omega) = \sup \mathfrak{T}$ by definition of pseudoinverse. This connection can be used to succinctly describe the critical set when the pathwise derivative fails.

Lemma 3.9. *For a given* CIF $\Lambda(t;\theta)$ *corresponding to one segment of a Zig-Zag process with a single tunnel* $\mathfrak{T}(\theta)$*, the set*

$$A(\varepsilon,\theta) = \left\{\omega \in (0,\infty) : \left|\Lambda^{\leftarrow}(\omega;\theta + \tfrac{1}{2}\varepsilon) - \Lambda^{\leftarrow}(\omega;\theta - \tfrac{1}{2}\varepsilon)\right| > \alpha|\varepsilon|\right\} \tag{3.44}$$

for $\varepsilon \neq 0$ *and some bound* $\alpha > 0$*, is either empty or equals whichever is nonempty of* $(\omega_{\mathfrak{T}(\theta-\epsilon/2)}, \omega_{\mathfrak{T}(\theta+\epsilon/2)}]$ *or* $(\omega_{\mathfrak{T}(\theta+\epsilon/2)}, \omega_{\mathfrak{T}(\theta-\epsilon/2)}]$.

Proof. The argument is essentially illustrated in fig. 3.5; if the ω considered is between $\omega_{\mathfrak{T}(\theta-\epsilon/2)}$ and $\omega_{\mathfrak{T}(\theta+\epsilon/2)}$, then the resulting times land on different sides of the tunnel and the jump contributes to the difference, ensuring the difference does not become arbitrarily small. However, if ω is such that the times both land on the same side of the tunnel, the local θ-differentiability of $\Lambda^{\leftarrow}$ ensures that it in particular is Lipschitz in the argument for this ω and hence fails to satisfy the condition. Finally, α is determined by the maximal Lipschitz constant obtained on each of the piecewise derivatives of $\Lambda^{\leftarrow}$. □

Intuitively, if $\omega_{\mathfrak{T}(\theta-\epsilon/2)} < \omega \leq \omega_{\mathfrak{T}(\theta+\epsilon/2)}$, the perturbation causes the reflection to be pushed before the tunnel, while if $\omega_{\mathfrak{T}(\theta+\epsilon/2)} < \omega \leq \omega_{\mathfrak{T}(\theta-\epsilon/2)}$ the perturbation causes the reflection to be pushed after the tunnel. In the special case $\omega_{\mathfrak{T}(\theta-\epsilon/2)} = \omega_{\mathfrak{T}(\theta+\epsilon/2)}$ the perturbation causes no jumps at all.

Corollary 3.10. *In the limit, perturbing reflections exactly at tunnel openings cause jumps. That is,* $\lim_{\epsilon\to 0} A(\epsilon,\theta) = \{\omega_{\mathfrak{T}(\theta)}\} = \{\tau = \inf\mathfrak{T}(\theta)\}$ *if nonempty.*

The impact of a perturbation pushing a reflection across a tunnel on level of a whole trajectory is thus quantified by the difference between two limiting trajectories at the openings of the tunnel, corresponding to reflections 'infinitesimally before' or 'infinitesimally after' the tunnel. The strategy we will use to efficiently generate such trajectories from a sample primal trajectory is by *teleporting* the particle between crossings of the openings of the tunnel. This generates new simulated events which correspond to the desired reflections.

The benefit of this approach is that these constructed trajectories eventually end back up on the same trajectory (translated in time) as the primal trajectory they were constructed from. Hence one can very easily compute the change in a linear performance functional (e.g. expectation) generated by this perturbation, as by linearity it suffices to compute this performance for the 'skipped' segment which is much shorter than the whole trajectory. At the same time, the construction ensures that the trajectories deviate minimally, thus computing the impact with relatively little variance. Nevertheless, such reflections can only be forced on segments where the primal trajectory crossed the tunnel in the first place, and thus there is an additional conditioning on this occurring when using such trajectories.

Definition 3.11. Suppose assumption 3.1 holds. Consider the ith segment of the trajectory, and suppose there is a tunnel $\mathfrak{T}(\theta)$ on this segment. Suppose we hit the tunnel, i.e. $\tau_i > \inf\mathfrak{T}(\theta)$. Then we can construct the limiting trajectory $\{Z_t^{\theta^+(i)}\}_{t=0}^{T^+(i)}$ (and $\{Z_t^{\theta^-(i)}\}_{t=0}^{T^-(i)}$ respectively, replacing inf by sup) from the primal, illustrated in fig. 3.6, as follows:

1. Replace τ_i with $\inf\mathfrak{T}(\theta)$ ($\sup\mathfrak{T}(\theta)$) and from it recompute x_i, v_i, so the reflection occurs at the beginning (end) of the tunnel.
2. Jump ahead to the next time Z_t^θ crosses $\inf\mathfrak{T}(\theta)$ ($\sup\mathfrak{T}(\theta)$), or if no such point exists, end the trajectory.

3. Suppose the next time this crossing occurs is on the $(j+1)$th segment, after the jth reflection: replace τ_{i+1} with $(x_{j+1} - x_i)/v_i$, teleporting to the crossing, and recompute x_{i+1}, rejoining the primal.
4. Then replay segments $j+1, j+2, \ldots$ until the end of the trajectory.

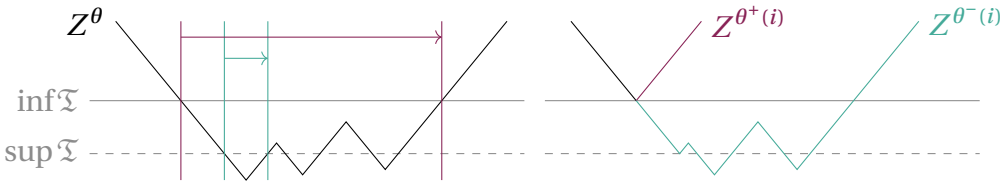


Figure 3.6: Regaining the coupling through forcing reflections by teleporting the sampler between tunnel crossings: Left, we see the teleportations on the primal trajectory (black). Right, creating the limiting trajectories from these: before (purple) and after (teal) the tunnel.

3.3.2 *Stratified derivative of trajectories*

With a characterization of discontinuity events and the constructed limiting trajectories we finally have all the necessary elements to construct a stratified derivative, where we smooth out the jumps caused by tunnels at the trajectory level to obtain a general unbiased estimator.

Theorem 3.12. *Suppose assumption 3.1 holds. Suppose further that Z^θ has a single tunnel, which we identify with two zeros $c_\theta(v)$ of $\frac{\partial \psi}{\partial x}(x;\theta)$ (selecting the correct zero and tunnel opening depending on the sign of v). On each segment i, let $\mathfrak{T}_i(\theta)$ be this tunnel (in terms of time) for the corresponding IPP.*

Define the functional estimating the expectation

$$F(Z_t^\theta) = \frac{1}{T_\theta} \int_0^{T_\theta} f(Z_t^\theta)\, dt. \tag{3.45}$$

Then an unbiased estimator of $\frac{\partial}{\partial \theta} \mathbb{E}\left[F(Z_t^\theta)\right]$ is given by

$$\frac{\partial}{\partial \theta} F(Z_t^\theta) + \sum_{i=1}^{N} \mathbf{1}_{\{\tau_i > \inf \mathfrak{T}_i(\theta)\}} J_{\theta(i)} \left(\frac{1}{v_i} \int_{x_i}^{c_\theta(v_i)} \frac{\partial \lambda}{\partial \theta}(y, v_i; \theta)\, dy \right) \tag{3.46}$$

where the first term is given by lemmata 3.4 and 3.5, and the performance difference $J_{\theta(i)} = F(Z_t^{\theta^+(i)}) - F(Z_t^{\theta^-(i)})$ using the construction in definition 3.11.

Proof. Let $\{\mathscr{F}_i\}_{i=0}^N$ be the filtration obtained by the information at successive reflections, i.e. $\mathscr{F}_i = \sigma(Z_s^\theta, 0 \le s \le t_i)$. Use once again the shadowing coupling (definition 3.2) in order to consider the primal trajectory $\{Z_t^\theta\}_{t=0}^{T_\theta}$ and the shadow trajectory $\{Z_t^{\theta+\delta}\}_{t=0}^{T_{\theta+\delta}}$ for some small δ. Hence, we can let $\eta_i = \Lambda_i(\tau_i;\theta)$ following the shadowing coupling and obtain a vector of latents $(\eta_i)_{i=1}^N$, independent and $\mathsf{Exp}(1)$ distributed, so that $\mathscr{F}_i = \sigma(\eta_1, \ldots, \eta_i)$. Here $Z^{\theta+\delta}$ is not

adapted to $\{\mathscr{F}_i\}$ in the usual sense, since the arrival times of the reflections may be perturbed, but the reflections are 'unveiled' successively.

Suppose $\varepsilon > 0$. We now define a FCCE for the expectation functional using the shadowing coupling, by letting $A^*(\varepsilon,\theta)$ be the event where no jumps occur, that is for the CIFs Λ_i on each segment i we have $\Lambda_i^{\leftarrow}$ locally θ-differentiable for all i, and letting $A_i(\varepsilon,\theta)$ be the event where a jump occurs on the ith segment *only*, that is the Lipschitz condition being violated by $\Lambda_i^{\leftarrow}$. Taken as a fixed event, the definition is difficult to write out fully, since it depends on the exact sequence of latents. We proceed conditional on $\mathscr{F}_{i-1}$, which allows us to use the characterization of lemma 3.9. Then the limiting event exists and has measure zero, corresponding to an empty event (if no tunnel exists or the tunnel is not parameter sensitive) or the latent hitting a single value.

Suppose that the tunnel $\mathfrak{T}_i(\theta)$ in question has $\omega_{\mathfrak{T}_i(\theta-\epsilon/2)} < \omega_{\mathfrak{T}_i(\theta+\epsilon/2)}$. Then

$$\mathbb{P}(A_i(\varepsilon,\theta) \mid \mathscr{F}_{i-1}) = \int_{\omega_{\mathfrak{T}_i(\theta-\epsilon/2)}}^{\omega_{\mathfrak{T}_i(\theta+\epsilon/2)}} e^{-s}\,ds \tag{3.47}$$

$$= -\left(e^{-\Lambda_i(\inf\mathfrak{T}_i(\theta+\epsilon/2);\theta+\epsilon/2)} - e^{-\Lambda_i(\inf\mathfrak{T}_i(\theta-\epsilon/2);\theta-\epsilon/2)}\right) \tag{3.48}$$

and hence

$$w_i^\theta = \lim_{\epsilon\downarrow 0}\frac{1}{\varepsilon}\mathbb{P}(A_i(\varepsilon,\theta) \mid \mathscr{F}_{i-1}) = \frac{\partial\Lambda_i}{\partial\theta}(\inf\mathfrak{T}_i(\theta);\theta)\,e^{-\Lambda_i(\inf\mathfrak{T}_i(\theta);\theta)} \tag{3.49}$$

showing the ($\mathscr{F}_{i-1}$-measurable) critical rate exists (with a Lipschitz-type bound for sufficiently small ε), where the dependency of $\inf\mathfrak{T}_i(\theta)$ on θ does not contribute as the rate λ is zero at the opening. If $\omega_{\mathfrak{T}_i(\theta-\epsilon/2)} > \omega_{\mathfrak{T}_i(\theta+\epsilon/2)}$, then w_i^θ is instead the negative of (3.49). The last case of $\omega_{\mathfrak{T}_i(\theta-\epsilon/2)} = \omega_{\mathfrak{T}_i(\theta+\epsilon/2)}$ implies that the event is not critical, but w_i^θ in any case is zero.

For $A^*(\varepsilon,\theta)$, a sequence of such conditional characterizations, where the CIFs and random bounds depend on the previous trajectory, show that the limiting event also exists. There the performance is then (a.s.) Lipschitz so that the pathwise derivatives and lemmata 3.4 and 3.5 apply. The collection of events is also sufficiently fine: if we were to account for multiple jumps along the trajectory, the critical rate would vanish, since the probability of two or more jumps is $o(\varepsilon)$ by the above calculation on the separate segments by the Markov property. It follows that $\lim_{\epsilon\downarrow 0}\mathbb{P}(A^*(\varepsilon,\theta)) = 1$.

Finally, the difference between trajectories reflecting at either end of the tunnel, all else equal, is certainly a.s. bounded in absolute value since our segment is a.s. finite, and given $\mathscr{F}_i$ we can obviously determine whether a jump occurred in the last segment, so the necessary hypotheses on the FCCE are satisfied.

Hence the conditions for theorem 2.7 are fulfilled, and

$$\frac{\partial}{\partial\theta}\mathbb{E}\left[F(Z_t^\theta)\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta}F(Z_t^\theta)\right] + \mathbb{E}\left[\sum_{i=1}^{N}\mathbb{E}\left[\Delta_F^{\theta(i)} \mid \mathscr{F}_{i-1}\right] w_i^\theta\right] \tag{3.50}$$

where $\mathbb{E}\left[\Delta_F^{\theta(i)} \mid \mathscr{F}_{i-1}\right]$ is composed of (a priori random) limiting trajectories that differ only in which end of the tunnel reflection i happens. We select $\Delta_F^{\theta(i)}$ as the difference before and after the tunnel rather than precisely right and left limits; this represents reflections being pulled earlier, corresponding to $\frac{\partial \Lambda_i}{\partial \theta} > 0$. However, if $\frac{\partial \Lambda_i}{\partial \theta} < 0$ at the opening, the expression for w_i^θ contributes a negative sign so that we have the correct difference representing reflections being pushed later, and if $w_i^\theta = 0$ the difference should also be zero. This slight abuse of notation where we let the right limit always mean 'before the tunnel' and vice versa allows us to have a single expression no matter the sign of the derivative.

Although this provides an estimator, it is not practical to sample these limiting trajectories if the primal trajectory did not cross the tunnel, since it will require simulating new trajectories after the reflection. We would like to use the construction in definition 3.11 (yielding $J_{\theta(i)}$), but it is only valid when the primal crosses the tunnel on that segment. However

$$\mathbb{P}(\tau_i > \inf \mathfrak{T}_i(\theta) \mid \mathscr{F}_{i-1}) = e^{-\Lambda_i(\inf \mathfrak{T}_i(\theta);\theta)} \tag{3.51}$$

and certainly conditional on $\mathscr{F}_{i-1}$ the indicator $\mathbf{1}_{\{\tau_i > \inf \mathfrak{T}_i(\theta)\}}$ of hitting the tunnel is independent of the limiting trajectories, since they do not depend on the primal τ_i. It follows that the sampling bias can be corrected:

$$\mathbb{E}\left[\sum_{i=1}^N \mathbb{E}\left[\Delta_F^{\theta(i)} \mid \mathscr{F}_{i-1}\right] w_i^\theta\right] = \mathbb{E}\left[\sum_{i=1}^N \mathbf{1}_{\{\tau_i > \inf \mathfrak{T}_i(\theta)\}} \mathbb{E}\left[\Delta_F^{\theta(i)} \mid \mathscr{F}_{i-1}\right] e^{\Lambda_i(\inf \mathfrak{T}_i(\theta);\theta)} w_i^\theta\right]. \tag{3.52}$$

In this setting $J_{\theta(i)} \overset{d}{=} \mathbb{E}\left[\Delta_F^{\theta(i)} \mid \mathscr{F}_{i-1}\right]$. It remains only to simplify the last factor according to our known expression for $\frac{\partial \Lambda_i}{\partial \theta}$

$$e^{\Lambda_i(\inf \mathfrak{T}_i(\theta);\theta)} w_i^\theta = \frac{1}{v_i} \int_{x_i}^{c_\theta(v_i)} \frac{\partial \lambda}{\partial \theta}(y, v_i; \theta)\, dy \tag{3.53}$$

and we obtain the result. □

Remark. We emphasize that we only need to observe a single trajectory to produce an estimate of the derivative according to the theorem. The extension to multiple tunnels on an interval is straightforward by further dividing the critical events on each segment, generating several possible jump terms, and the conditional difference $\mathbb{E}\left[\Delta_F^{\theta(i)} \mid \mathscr{F}_{i-1}\right]$ will be a mixture between the jumps.

Remark. We conjecture that we have consistency of the one-dimensional estimator even in the presence of tunnels. It suffices to establish a regenerative structure like in theorem 3.6, where the main change in the proof will be that the form of the excursions is more complicated and finite-horizon unbiasedness is established by the preceding theorem.

3.4 EXAMPLES

As a proof-of-concept, we implement the estimator we have derived in theorem 3.12 in Julia [33], using `ZigZagBoomerang.jl` [34] to run the Zig-Zag sampler. (More details on the implementation are available in appendix B.) Using the implementation, we test a few example problems. All sampler runs are started in the mean with a random initial direction.

Example 3.13 (Gaussian scale)**.** Consider $X^\theta \sim \mathsf{N}(0,\theta^2)$, and $f \equiv \mathbf{1}_{[a,b]}$ for some interval. This example is artificial, since there is a direct reparameterization $X^\theta = \theta X^1$ moving the parameter dependency out to the the performance and avoiding differentiating the sampler entirely, but of course such shortcuts are not generally available.

The two required derivatives of the potential are

$$\frac{\partial \psi}{\partial x}(x;\theta) = \frac{x}{\theta^2}, \qquad \frac{\partial^2 \psi}{\partial \theta \partial x}(x;\theta) = -\frac{2x}{\theta^3} \tag{3.54}$$

and the unimodal distribution implies the smoothness assumptions holds, so that there are no tunnels in this example.

Here we have exact expressions for the true values, with the derivative even composed of elementary functions:

$$\mathbb{E}[\mathbf{1}_{[a,b]}(X^\theta)] = \frac{1}{2}\left(\operatorname{erf}\left(\frac{b}{\sqrt{2}\theta}\right) - \operatorname{erf}\left(\frac{a}{\sqrt{2}\theta}\right)\right) \tag{3.55}$$

$$\frac{\partial}{\partial\theta}\mathbb{E}[\mathbf{1}_{[a,b]}(X^\theta)] = -\frac{1}{\sqrt{2\pi}\theta^2}\left(b\exp\left(-\frac{b^2}{2\theta^2}\right) - a\exp\left(-\frac{a^2}{2\theta^2}\right)\right). \tag{3.56}$$

Comparing with the results from the sampler in fig. 3.7, we see that we have good convergence to the true result with very little variance. This situation is more or less the ideal case, with a unimodal target density and the sampler efficiently exploring regions that contribute to the derivative.

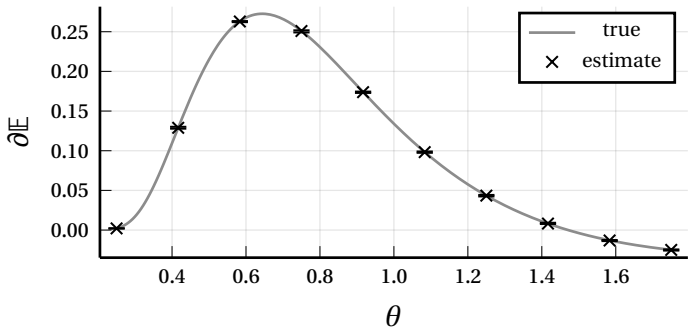


Figure 3.7: Derivative estimates for Zig-Zag targeting a Gaussian parameterized by scale, with $f \equiv \mathbf{1}_{[1,2]}$. At each θ we sample 10 trajectories for $T = 10^6$ and show means (×), minima and maxima (error bars).

Example 3.14 (Gaussian mixture)**.** The natural first bimodal example is a mixture of two Gaussians. We will see that the jump term is required for an unbiased estimate in this situation, where a tunnel is present. Consider (with abuse of notation) $X^\theta \sim \frac{1}{2}\mathsf{N}(2+\theta,1) + \frac{1}{2}\mathsf{N}(-2-\theta,1)$ with $\theta \geq 0$, and $f \equiv \mathbf{1}_{[a,b]}$ for some interval.

The two required derivatives of the potential are

$$\frac{\partial \psi}{\partial x}(x;\theta) = x - (2+\theta)\tanh((2+\theta)x) \tag{3.57}$$

$$\frac{\partial^2 \psi}{\partial \theta \partial x}(x;\theta) = -\frac{x(2+\theta)}{\cosh^2((2+\theta)x)} - \tanh((2+\theta)x) \tag{3.58}$$

and we omit the true derivative of the expectation, only noting that it can be obtained in elementary closed form.

The results from the sampler in fig. 3.8 also show the contribution from the pathwise derivative and the smoothed jumps separately. Indeed both terms are required to obtain an unbiased estimator. The difficulty of crossing between modes contributes to the variance, so that longer trajectories become necessary, but the jump term estimator is very stable thanks to the teleportation construction.

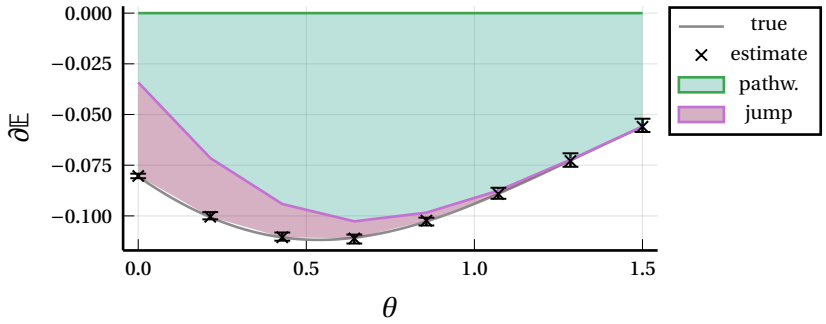


Figure 3.8: Derivative estimates for Zig-Zag targeting a mixture of two Gaussians, with $f \equiv \mathbf{1}_{[1,2]}$. At each θ we sample 15 trajectories for $T = 10^6$ and show means (×), minima and maxima (error bars). We also show the contribution of each estimate term.

It may seem unintuitive that the jump term vanishes as the modes move further apart. One can explain this by connecting the jump term to the distortion of tunnels; the shape of the distribution changes significantly more when the modes are closer together. (And recall that parameter-insensitive tunnels do not cause jumps.)

Example 3.15 (Alpha-skew-Normal)**.** The *α-skew-Normal* distribution [35] has a shape parameter $\alpha \in \mathbb{R}$, and transitions from unimodal to bimodal at $\alpha \approx \pm 1.34$. We say a random variable is $\mathsf{ASN}(\alpha)$-distributed if it has density

$$\pi(x) = \frac{(1-\alpha x)^2 + 1}{2+\alpha^2}\varphi(x), x \in \mathbb{R} \tag{3.59}$$

where $\varphi(x)$ is the standard Normal density. Note that ASN(0) is equal in distribution to a standard Normal. To our knowledge there is no closed form inverse nor a continuous reparameterization for general α.

We target ASN(θ), and consider $f \equiv \mathbf{1}_{[a,b]}$ for some interval. The two required derivatives of the potential are

$$\frac{\partial \psi}{\partial x}(x;\theta) = x + \frac{2\theta(1-\theta x)}{1+(1-\theta x)^2} \tag{3.60}$$

$$\frac{\partial^2 \psi}{\partial \theta \partial x}(x;\theta) = \frac{2(2-4\theta x+\theta^2 x^2)}{(2-2\theta x+\theta^2 x^2)^2} \tag{3.61}$$

and yet again we omit the true derivative of the expectation and note it can be obtained in elementary closed form.

Like the previous example, fig. 3.9 shows the contribution from the pathwise derivative and the smoothed jumps separately. We see the transition where the tunnel appears and the jump contribution becomes non-zero.

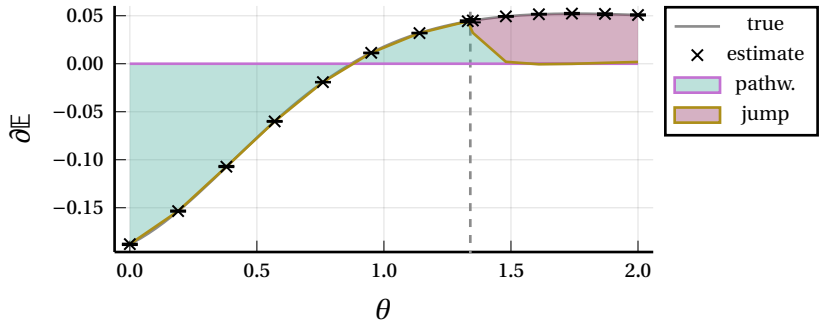


Figure 3.9: Derivative estimates for Zig-Zag targeting ASN(θ), with $f \equiv \mathbf{1}_{[1,2]}$. At each θ we sample 15 trajectories for $T = 10^6$ and show means (×), minima and maxima (error bars). We also show the contribution of each estimate term, and indicate the theoretical shift between uni- and bimodality.

Example 3.16 (An inverse inverse problem)**.** The following problem illustrating an application of differentiable Monte Carlo is taken from Chandra, Li, Tenenbaum & Ragan-Kelley [9].

Suppose the prior model for a temperature $T \sim \mathsf{N}(70, 5^2)$ degrees Fahrenheit. Our thermometer measures $M = T + E$, where the error $E \sim \mathsf{N}(0, 2^2)$.

(i) If we observe $M = 100$ degrees Fahrenheit, what is our estimate for T?

(ii) What measurement M would we have to observe to infer that $T = 100$ degrees Fahrenheit?

By conjugacy, this Bayesian inference problem can be solved to obtain the exact posterior $T \mid M = m \sim \mathsf{N}((25m + 280)/29, 100/29)$. Hence, the answer to the first question is $\mathbb{E}[T \mid M = 100] = \frac{2780}{29} \approx 95.86$. Numerically, the resulting posterior for $T \mid M = 100$ could be sampled with the Zig-Zag sampler to compute the expectation.

However, the second question is an 'inverse inverse problem' in the terminology of [9]. In this simple case, an analytical solution is available: $\mathbb{E}[T \mid M = m] = 100$ has solution $m^* = \frac{524}{5} = 104.8$. But it is also a simple example of a problem which could be solved numerically by differentiable Monte Carlo. By considering m as a parameter of the posterior, we can minimize the objective $\ell(m) = \mathbb{E}[(T - 100)^2 \mid M = m]$ to find the solution m^*. (The choice of quadratic loss is somewhat arbitrary; all we require is the correct global minimum, and the exact choice of loss function in the expectation only affects numerical properties of the minimization.)

Since ℓ has the form of the expectation of some function of the posterior, it is an instance of the prototype problem in this chapter. Thus an estimator for $\ell'(m)$ can be obtained from a Zig-Zag process sampling the posterior together with our derivative construction, and we can by gradient descent find m^* numerically. The results of a gradient descent-type optimization starting in $m_0 = 100$ is illustrated in fig. 3.10, where we see that with some tuning we converge quickly to the true m^*.

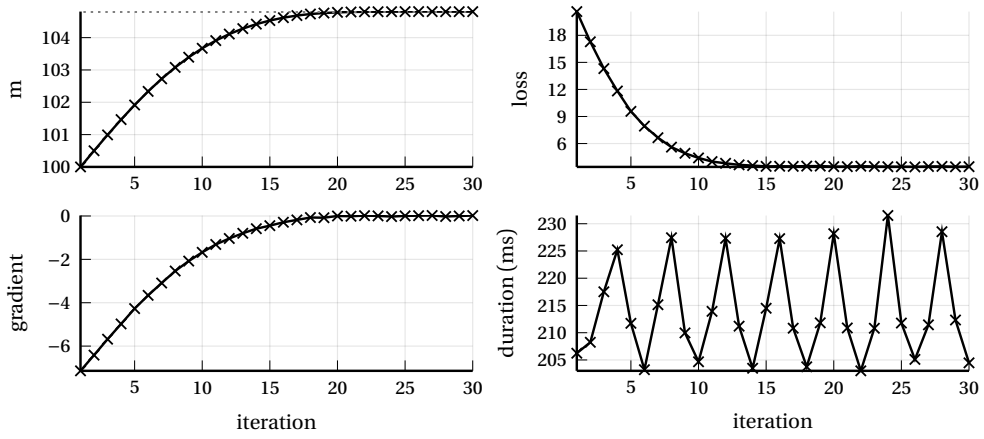


Figure 3.10: Optimization diagnostics for example 3.16. The Adam optimizer [36] (learning rate $\alpha = 0.5$, decays $\beta_1 = 0.5, \beta_2 = 0.99$) is run using our Zig-Zag-based estimates of $\ell'(m)$ by targeting the posterior ($T = 10^5$).

4 DIFFERENTIATION OF THE BOUNCY PARTICLE SAMPLER IN MULTIPLE DIMENSIONS

We now extend the problem in the previous chapter by considering multi-dimensional distributions. Given a target distribution μ_θ on $\mathbb{R}^d$, $d \in \mathbb{N}$, parameterized by θ, and a function $f : \mathbb{R}^d \to \mathbb{R}$, our goal is to use a trajectory of the sampler to estimate $\frac{\partial}{\partial\theta}\mathbb{E}_{\boldsymbol{X}^\theta\sim\mu_\theta}[f(\boldsymbol{X}^\theta)]$. Again, the extension to gradients of parameter vectors is naturally done through directional derivatives.

Rather than use the multi-dimensional Zig-Zag sampler, we will instead use the Bouncy Particle sampler. The reason is that the discrete set of directions allowed in the Zig-Zag sampler creates direction change perturbations which are not infinitesimal, making the pathwise derivative inapplicable. Methods exist to work with the sensitivities of discrete randomness (see e.g. [17]) but are likely to require simulating alternative trajectories. Instead, the Bouncy Particle sampler has a continuous and a.e. differentiable reflection rule.

However, the Bouncy Particle sampler requires the introduction of *refreshment* events to ensure ergodicity. This leads to a new difficulty we did not have in one dimension: handling a perturbation causing the interchange of reflections and refreshments. We will see that an appropriate choice of coupling, using an equivalent formulation of the sampler, will allow us to control the size of such perturbations.

Our standard setting in this chapter is the following:

Assumption 4.1. Let $\{\boldsymbol{Z}_t^\theta\}_{t=0}^T$ be a Bouncy Particle sampler targeting a distribution μ_θ parameterized by $\theta \in \Theta \subseteq \mathbb{R}$ (in particular, μ_θ satisfies the assumptions in theorem 2.24). Let $f : \mathbb{R}^d \to \mathbb{R}$ be such that the expectation $\mathbb{E}_{\boldsymbol{X}^\theta\sim\mu_\theta}[f(\boldsymbol{X}^\theta)]$ exists and is differentiable wrt θ. Furthermore, the potential $\psi(\boldsymbol{x};\theta)$ corresponding to μ_θ is $\mathscr{C}^2$. We identify trajectories of the sampler with the sequence of skeleton points $\{(t_i, \boldsymbol{x}_i, \boldsymbol{v}_i, r_i)\}_{i=0}^N$, where $r_i \in \{\text{reflect}, \text{refresh}\}$ indicates the event type.

Notably, the assumptions on the potential are the same as in the previous chapter, although we will now make use of all second order derivatives.

The outline of this chapter parallels the previous one: We begin by constructing a coupling, which requires exploring how the sampler behaves when a perturbation reorders reflections and refreshments. Next, we once again differentiate in two steps under a smoothness assumption, first the expectation integral wrt the sampler, then the trajectory wrt the parameter. Surprisingly, unlike in one dimension, tunnels do not pose an issue for the Bouncy Particle

sampler as there is no difference in the trajectory at the endpoints, and hence we will in fact derive a general estimator, subject to some boundedness conditions. We finish with some examples and simulations.

4.1 SHADOWING COUPLING

We proceed with defining a coupling for use in deriving our derivative estimators. The approach chosen here is hierarchical: we simulate first refreshment events, then reflection events in-between. This avoids unnecessary deviations between the primal and shadow trajectories due to refreshments, exploiting that they do not depend on the parameter θ. The connection between reflection interarrival times in the two trajectories is still provided by the inversion method, like in the previous chapter.

Definition 4.2 (Shadowing coupling, BPS)**.** We refer to the characteristics in definition 2.23. Let $\{\eta_{i,j}\}_{i\geq 1, j\geq 1}$ be a matrix of iid $\mathsf{Exp}(1)$ latents, let $\{\tau_i^{\text{refr}}\}_{i\geq 1}$ be a sequence of iid $\mathsf{Exp}(\lambda_{\text{refr}})$ refreshment interarrival times, and let $\{\boldsymbol{u}_i\}_{i\geq 1}$ be a sequence of iid ρ refreshment directions. These form the randomness which couple the trajectories.

Denote by $\{Z_t^\theta\}_{t\geq 0}$ the primal trajectory, and by $\{Z_t^{\theta+\epsilon}\}_{t\geq 0}$ the shadow trajectory for a perturbation ε. Construct each of them separately as follows:

Initialize a trajectory Z^θ at Z_0
for $i = 1, 2, \ldots$ **do**
  $j \leftarrow 1$, $\delta \leftarrow \tau_i^{\text{refr}}$
  $\tau_j \leftarrow$ *Obtain* the next reflection time from Z^θ using $\eta_{i,j}$
  while $\tau_j < \delta$ **do**
    Move Z^θ deterministically for τ_j
    Reflect Z^θ ▷ Emits an event.
    $j \leftarrow j+1$, $\delta \leftarrow \delta - \tau_j$
    $\tau_j \leftarrow$ *Obtain* the next reflection time from Z^θ using $\eta_{i,j}$
  end while
  Move Z^θ deterministically for δ
  Refresh Z^θ with new direction $\boldsymbol{u}_i$ ▷ Emits an event.
end for

where we sample reflection times from the Bouncy Particle sampler IPP without refreshments ($\lambda_{\text{refr}} = 0$) starting from the current state by inversion, that is $\tau_j^\theta = \Lambda_j^{\leftarrow}(\eta_{i,j}; \theta)$. Therefore, between refreshments we use latents from row i to obtain reflection times in both trajectories. The movement, reflections and refreshments are done according to the Bouncy Particle sampler flow and transitions.

Importantly, the primal and shadow trajectories have the same refreshment arrival times, but may have a different number of events between refreshments due to the perturbation. The coupling ensures this does not decouple

the common randomness following the refreshment. Finally, we may truncate the processes at an appropriate stopping time to obtain a finite trajectory; the ideal choice is at the arrival of a refreshment, as this will make the final time T independent of θ.

Proposition 4.3. *Using definition 4.2, we have that the generated* $\{Z_t^\theta\}_{t\geq 0}$ *and* $\{Z_t^{\theta+\epsilon}\}_{t\geq 0}$ *marginally are Bouncy Particle samplers.*

Proof. It suffices to argue for the primal trajectory $\{Z_t^\theta\}_{t\geq 0}$, since the two trajectories are simulated the same way, just with a different parameter θ. Following a remark on definition 2.23, we note that the IPP on each segment can be viewed as the superposition of two IPPs, one for reflections and one for refreshments, each with separate transition kernels. Thus it suffices to simulate interarrival times from the separate IPPs and select the event type using whichever occurs first; we are free to specify that the refreshment 'wins' if they are equal, since that almost never occurs.

The reflection times are correctly distributed by inversion, but the refreshments are simulated in a different way than the general PDMP construction. There, we draw a new first arrival from the IPP for each segment. However, the refreshments come from a homogeneous IPP, and thus by memorylessness of the exponential distribution the interarrival time after waiting for a reflection will still be exponential. Thus this scheme also simulates the refreshments correctly.

Finally, we note that the reflection kernel is deterministic, and the refreshment kernel is deterministic conditional on the new direction, which is given by the correctly distributed $\{\boldsymbol{u}_i\}_{i\geq 1}$; hence the times are sufficient to determine the evolution of the sampler, and we are done. □

4.1.1 *The reordering problem*

Note that simply by observing the primal trajectory we do not a priori possess enough information to generate the shadow trajectory for an arbitrary finite perturbation ε; the shadow trajectory *could* have additional reflections on a section between two refreshments that did not occur in the primal trajectory, and information on those latents is not available. Moreover, such a *reordering* may cause the shadow trajectory to deviate further from the primal trajectory, which is not desirable for controlling the variance. However, as

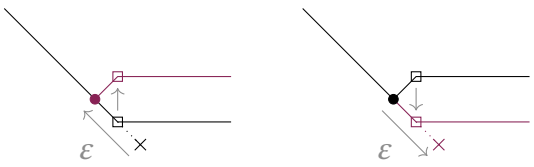


Figure 4.1: Illustration of a reordering adding or removing a reflection from a segment. Note that the two cases are equal up to a sign change.

$\varepsilon \to 0$, the probability of this occurring vanishes; in the limit a reordering only occurs if a reflection and refreshment arrives simultaneously. If we can verify that the coupling behaves nicely under a reordering, so that the perturbations a reordering causes are sufficiently small, we may thus neglect it for the purposes of applying the pathwise derivative.

Proposition 4.4. *For a sufficiently small but finite perturbation ε, the change between $\{Z_t^\theta\}_{t\geq 0}$ and $\{Z_t^{\theta+\epsilon}\}_{t\geq 0}$ (defined by the shadowing coupling) in refreshment positions caused by a reordering is surely Lipschitz. Furthermore, the probability of a reordering is $O(\varepsilon)$. Hence neglecting reorderings does not introduce a bias in the pathwise derivative.*

Proof. Consider first the prototypical case in fig. 4.1 of a single segment, from a starting point at time zero, to a refreshment at time τ_R. (In general, the starting point may be of any event type and thus sensitive to the perturbation ε.) Depending on how ε affects the time to reflection τ_L, an additional reflection might be introduced before the refreshment if $\tau_R^\theta \leq \tau_L^\theta$ but $\tau_L^{\theta+\epsilon} < \tau_R^{\theta+\epsilon}$. The interarrival time until the refreshment on the extra segment then becomes

$$\tau_R'(\varepsilon) = \max\{\tau_R^{\theta+\epsilon} - \tau_L^{\theta+\epsilon}, 0\} \tag{4.1}$$

and the refreshment moves to

$$\begin{aligned} \boldsymbol{x}_R'(\varepsilon) &= \boldsymbol{x}_R^{\theta+\epsilon} + (\boldsymbol{v}'(\varepsilon) - \boldsymbol{v}^{\theta+\epsilon})\tau_R'(\varepsilon) && (4.2)\\ &= \boldsymbol{x}_R^{\theta+\epsilon} + (R(\boldsymbol{x}_L^{\theta+\epsilon}, \boldsymbol{v}^{\theta+\epsilon}) - \boldsymbol{v}^{\theta+\epsilon}) \max\{\tau_R^{\theta+\epsilon} - \tau_L^{\theta+\epsilon}, 0\} && (4.3)\end{aligned}$$

where we with the explicit parameter dependencies account for sensitivities in the starting point (which affects the refreshment interarrival time) and reflection rate. Hence, the difference caused by a reordering is Lipschitz as a function of the event interarrival times and positions. The opposite case where the perturbation ε excludes a final reflection before the refreshment is analogous, with a sign change in the difference.

Whether a reflection is added or removed depends on the sign of $\frac{\partial \tau_L}{\partial \theta}$, but also to some extent on the starting point sensitivity. In fact, the conditional probability of an additional reflection is

$$\mathbb{P}(\tau_R^\theta \leq \tau_L^\theta, \tau_L^{\theta+\epsilon} < \tau_R^{\theta+\epsilon} \mid \tau_R^\theta) = \int_{\Lambda(\tau_R^\theta;\theta)}^{\Lambda(\tau_R^{\theta+\epsilon};\theta+\epsilon)} e^{-s}\, ds \tag{4.4}$$

which is $O(\varepsilon)$ by continuity of the integral and CIF, with the opposite sign for a removed reflection. (The extra sign in that case cancels with the extra sign of the derivative.)

Now, let $\{\mathscr{F}_i\}_{i\geq 0}$ be the filtration obtained by the information up to the ith *refreshment*, that is $\mathscr{F}_i = \sigma(\{\eta_{k,j}\}_{k=1,\dots,i;\, j\geq 1}, \{\tau_k^{\text{refr}}\}_{k=1}^i, \{\boldsymbol{u}_k\}_{k=1}^i)$. By construction, for both the primal and shadow trajectory the state at the ith refreshment is $\mathscr{F}_i$-measurable. Given the next refreshment time, reflections are simulated according to the coupling, and for sufficiently small ε there is only a possible

difference of a single reflection between the two. But this means the refreshment position changes as in the prototypical case, because we can include all the accumulated sensitivity up to the last common reflection. Furthermore, this holds for an arbitrary refreshment time, so by total probability the reordering probability is still $O(\varepsilon)$ and the impact is surely Lipschitz. Hence, reorderings do not violate the conditions for unbiasedness in theorem 2.2 (although whether unbiasedness holds depends on how the perturbation affects the aforementioned skeleton points). □

This is a remarkable property of the coupling, necessary for our approach to work at all. Without it, every refreshment would give rise to some alternative trajectory caused by a possible reordering, but this lemma shows that we may without loss of generality assume a one-to-one correspondence between the events in the primal and shadow trajectories when taking pathwise derivatives, as is done in the previous chapter.

4.2 SMOOTH CASE

The shadowing coupling will be used similarly to the previous chapter to perform a pathwise differentiation of the trajectory. The proofs will follow the same ideas, just with more involved computations. We thus introduce a similar smoothness assumption for this section:

Assumption 4.5. The CIF of the IPP in the sampler is invertible in time on $(0,\infty)$, and the inverse is differentiable in θ.

A sufficient condition is that ψ is quasiconvex (or equivalently that the distribution is quasiconcave). This ensures no tunnels appear that cause the problematic jumps we previously studied. In one dimension this would imply unimodality of the distribution, just like in the corresponding smooth case of the previous chapter.

Definition 4.6. A function $\psi : \mathbb{R}^d \to \mathbb{R}$ is *quasiconvex* if for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$ and $\delta \in [0,1]$ it holds that

$$\psi(\delta \boldsymbol{x}_1 + (1-\delta)\boldsymbol{x}_2) \leq \max\{\psi(\boldsymbol{x}_1), \psi(\boldsymbol{x}_2)\}. \tag{4.5}$$

For convenience, we will use matrix calculus notation to succinctly represent the derivatives that appear. Hence for a vector $\boldsymbol{y}$ (or more precisely a vector-valued function) we write for scalar θ and vector $\boldsymbol{x}$

$$\frac{\partial \boldsymbol{y}}{\partial \theta} = \begin{bmatrix} \frac{\partial y_1}{\partial \theta} & \cdots & \frac{\partial y_n}{\partial \theta} \end{bmatrix}^\mathsf{T} \tag{4.6}$$

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix} \tag{4.7}$$

where $y_1, \dots, y_n$ and $x_1, \dots, x_n$ are the components of $\boldsymbol{y}$ and $\boldsymbol{x}$ respectively. Computations are made with the help of `matrixcalculus.org` [37]. To efficiently express the chain rule, we also write $\frac{d}{d\theta}$ for the *total derivative* which simply amounts to applying the chain rule viewing all arguments as functions of θ.

To differentiate the Bouncy Particle reflection rule $R(\boldsymbol{x}, \boldsymbol{v})$ in (2.39), define the shorthand *projection operator* for projecting $\boldsymbol{u}$ onto $\boldsymbol{y}$

$$P_{\boldsymbol{y}}\boldsymbol{u} = \frac{\boldsymbol{u}^\mathsf{T}\boldsymbol{y}}{\|\boldsymbol{y}\|^2}\boldsymbol{y} \tag{4.8}$$

so that $R(\boldsymbol{x}, \boldsymbol{v}) = \boldsymbol{v} - 2P_{\nabla_x \psi(\boldsymbol{x};\theta)}\boldsymbol{v}$, and note that if both $\boldsymbol{u}$ and $\boldsymbol{y}$ are θ-differentiable

$$\frac{dP_{\boldsymbol{y}}\boldsymbol{u}}{d\theta} = \left[\frac{\partial P_{\boldsymbol{y}}\boldsymbol{u}}{\partial \boldsymbol{u}}\right]^\mathsf{T}\frac{d\boldsymbol{u}}{d\theta} + \left[\frac{\partial P_{\boldsymbol{y}}\boldsymbol{u}}{\partial \boldsymbol{y}}\right]^\mathsf{T}\frac{d\boldsymbol{y}}{d\theta} \tag{4.9}$$

$$= \frac{\boldsymbol{y}\boldsymbol{y}^\mathsf{T}}{\|\boldsymbol{y}\|^2}\frac{d\boldsymbol{u}}{d\theta} + \left(\frac{\boldsymbol{y}\boldsymbol{u}^\mathsf{T}}{\|\boldsymbol{y}\|^2} - 2\frac{(\boldsymbol{u}^\mathsf{T}\boldsymbol{y})\boldsymbol{y}\boldsymbol{y}^\mathsf{T}}{\|\boldsymbol{y}\|^4} + \frac{(\boldsymbol{u}^\mathsf{T}\boldsymbol{y})I}{\|\boldsymbol{y}\|^2}\right)\frac{d\boldsymbol{y}}{d\theta}. \tag{4.10}$$

4.2.1 Derivative of expectation functional

We now prove the multi-dimensional analogue to lemma 3.4. Here, the cancellation that provided a geometric argument in one dimension does not necessarily occur due to perturbations of the reflection angle, and we have to analyse the impact segment by segment.

Lemma 4.7. *Under assumption 4.1, suppose the sensitivities of the event interarrival times, positions and velocities $\frac{\partial \tau_i}{\partial \theta}, \frac{\partial \boldsymbol{x}_i}{\partial \theta}, \frac{\partial \boldsymbol{v}_i}{\partial \theta}$ are well-defined and exist (a.s.) for all i. Suppose further that f is differentiable.*

Then the (a.s.) derivative wrt θ of the functional estimating the expectation

$$F(\boldsymbol{Z}_t^\theta) = \frac{1}{T_\theta}\int_0^{T_\theta} f(\boldsymbol{Z}_t^\theta)\, dt \tag{4.11}$$

is given by

$$\frac{\partial}{\partial\theta}F(Z_t^\theta) = \frac{1}{T_\theta}\sum_{i=1}^{N}\left(\left[f(\boldsymbol{x}_i) - F(\boldsymbol{Z}_t^\theta)\right]\frac{\partial \tau_i}{\partial \theta}\right. \tag{4.12}$$

$$\left. + \int_0^{\tau_1} \nabla f(\boldsymbol{x}_{i-1} + \boldsymbol{v}_{i-1}r)^\mathsf{T}\left[\frac{\partial \boldsymbol{x}_{i-1}}{\partial\theta} + \frac{\partial \boldsymbol{v}_{i-1}}{\partial\theta}r\right] dr\right) \tag{4.13}$$

Proof. We use a direct computational approach, with the shadowing coupling (definition 4.2) as the implicit tool to obtain the sensitivities in the hypothesis. Differentiating yields

$$\frac{\partial}{\partial\theta}F(Z_t^\theta) = \frac{1}{T_\theta}\left(\frac{\partial}{\partial\theta}\int_0^{T_\theta} f(\boldsymbol{Z}_t^\theta)\, dt\right) + \left(\frac{\partial}{\partial\theta}\frac{1}{T_\theta}\right)\int_0^{T_\theta} f(\boldsymbol{Z}_t^\theta)\, dt. \tag{4.14}$$

For the first term, consider the impact segment by segment. Fix $i \in \{1, \dots, N\}$. Then apply the Leibniz rule for the integral along the ith segment

$$\frac{\partial}{\partial\theta}\int_0^{\tau_i} f(\boldsymbol{x}_{i-1}+\boldsymbol{v}_{i-1}r)\,dr = f(\boldsymbol{x}_i)\frac{\partial\tau_i}{\partial\theta}+\int_0^{\tau_i}\nabla f(\boldsymbol{x}_{i-1}+\boldsymbol{v}_{i-1}r)^\intercal\left[\frac{\partial\boldsymbol{x}_{i-1}}{\partial\theta}+\frac{\partial\boldsymbol{v}_{i-1}}{\partial\theta}r\right]dr \tag{4.15}$$

and sum over all segments.

For the second term, we use $T_\theta = \sum_{i=1}^N \tau_i$ and differentiate directly to obtain

$$\left(\frac{\partial}{\partial\theta}\frac{1}{T_\theta}\right)\int_0^{T_\theta} f(\boldsymbol{Z}_t^\theta)\,dt = \left(-\frac{1}{(T_\theta)^2}\sum_{i=1}^N\frac{\partial\tau_i}{\partial\theta}\right)\int_0^{T_\theta} f(\boldsymbol{Z}_t^\theta)\,dt = \frac{1}{T_\theta}\sum_{i=1}^N\left(-F(\boldsymbol{Z}_t^\theta)\frac{\partial\tau_i}{\partial\theta}\right) \tag{4.16}$$

and combining the two yields the desired result. □

Remark. It is possible to weaken the assumption that f is differentiable with distributional derivatives or other ways to assign meaning to $\int\nabla f$, making the results applicable to e.g. $f \equiv \mathbf{1}_A$, but it would needlessly complicate the presentation here.

Note that lemma 4.7 simplifies to the same equation as lemma 3.4 in one dimension, as then the Bouncy Particle sampler (without refreshment) and the Zig-Zag sampler are equivalent, and we have deterministic velocities after reflection. However, the proof in the special case completely avoids any assumption of differentiability of f.

4.2.2 *Pathwise derivative of segments*

Next, we prove the multi-dimensional analogue to lemma 3.5. We require sensitivities not only for the position, but also the interarrival time and the velocity, making this generalization more complicated.

Lemma 4.8. *Under assumption 4.1, then (a.s.)*

$$\frac{\partial t_i}{\partial\theta} = \begin{cases} \dfrac{\partial t_{i-1}}{\partial\theta} - \dfrac{\sum\limits_{(a_j,b_j)\subset P_i}\nabla_x\psi(\boldsymbol{x}_{i-1}+\boldsymbol{v}_{i-1}r;\theta)^\intercal\left(\frac{\partial\boldsymbol{x}_{i-1}}{\partial\theta}+r\frac{\partial\boldsymbol{v}_{i-1}}{\partial\theta}\right)\Big|_{r=a_j}^{r=b_j}}{\lambda(\boldsymbol{x}_i,\boldsymbol{v}_{i-1};\theta)} & r_i = \text{reflect,} \\ \qquad - \dfrac{\int_0^{\tau_i}\frac{\partial\lambda}{\partial\theta}(\boldsymbol{x}_{i-1}+\boldsymbol{v}_{i-1}r,\boldsymbol{v}_{i-1};\theta)\,dr}{\lambda(\boldsymbol{x}_i,\boldsymbol{v}_{i-1};\theta)} & \\ 0 & r_i = \text{refresh} \end{cases} \tag{4.17}$$

$$\frac{\partial\tau_i}{\partial\theta} = \frac{\partial t_i}{\partial\theta} - \frac{\partial t_{i-1}}{\partial\theta} \tag{4.18}$$

$$\frac{\partial\boldsymbol{x}_i}{\partial\theta} = \left[\frac{\partial\boldsymbol{x}_{i-1}}{\partial\theta}+\frac{\partial\boldsymbol{v}_{i-1}}{\partial\theta}\tau_i\right]+\boldsymbol{v}_{i-1}\frac{\partial\tau_i}{\partial\theta} \tag{4.19}$$

$$\frac{\partial \boldsymbol{v}_i}{\partial \theta} = \begin{cases} \left(I - 2\dfrac{\boldsymbol{p}_i \boldsymbol{p}_i^\top}{\|\boldsymbol{p}_i\|^2}\right)\dfrac{\partial \boldsymbol{v}_{i-1}}{\partial \theta} \\ \quad -2\left(\dfrac{\boldsymbol{p}_i \boldsymbol{v}_{i-1}^\top}{\|\boldsymbol{p}_i\|^2} - 2\dfrac{(\boldsymbol{v}_{i-1}^\top \boldsymbol{p}_i)\boldsymbol{p}_i \boldsymbol{p}_i^\top}{\|\boldsymbol{p}_i\|^4} + \dfrac{(\boldsymbol{v}_{i-1}^\top \boldsymbol{p}_i) I}{\|\boldsymbol{p}_i\|^2}\right) & r_i = \text{reflect}, \\ \qquad \left(\nabla_{\boldsymbol{x}}^2 \psi(\boldsymbol{x}_i;\theta)\dfrac{\partial \boldsymbol{x}_i}{\partial \theta} + \dfrac{\partial \nabla_{\boldsymbol{x}} \psi}{\partial \theta}(\boldsymbol{x}_i;\theta)\right) \\ 0 & r_i = \text{refresh} \end{cases} \tag{4.20}$$

for $i = 1, 2, \ldots, N$, *where* $P_i = \{r \in [0, \tau_i] : \lambda(\boldsymbol{x}_{i-1} + \boldsymbol{v}_{i-1} r, \boldsymbol{v}_{i-1}; \theta) > 0\}$ *with the sum over the set of disjoint intervals* (a_j, b_j) *whose union is* P_i, *and* $\boldsymbol{p}_i = \nabla_{\boldsymbol{x}} \psi(\boldsymbol{x}_i; \theta)$.

Proof. We consider the effect of a perturbation segment by segment. A perturbation has two qualitatively different sources, just like in one dimension: either it is the result of accumulated perturbations of earlier segments, or the result of perturbing the dynamics along the current segment.

First, eqs. (4.17) and (4.18) make use of the shadowing coupling (definition 4.2). Each segment may end in either a reflection or a refreshment. In the case of a refreshment, we work with *arrival* time, which itself is independent of θ and by proposition 4.4 we may neglect any possibility of reordering (which would alter the number of events). Hence the derivative is zero.

In the case of a reflection, we instead work with *interarrival* times. Differentiate through the inversion coupling and apply the relevant inverse function theorem (theorem A.1)

$$\frac{\partial \tau_i}{\partial \theta} = \frac{\partial \Lambda_i^{-1}}{\partial \theta}(\Lambda_i(\tau_i;\theta);\theta) = -\frac{\frac{\partial \Lambda_i}{\partial \theta}(\tau_i;\theta)}{\frac{\partial \Lambda_i}{\partial t}(\tau_i;\theta)} \tag{4.21}$$

where

$$\frac{\partial \Lambda_i}{\partial t}(\tau_i;\theta) = \lambda(\boldsymbol{x}_i, \boldsymbol{v}_{i-1};\theta) \tag{4.22}$$

$$\frac{\partial \Lambda_i}{\partial \theta}(\tau_i;\theta) = \int_0^{\tau_i} \frac{d\lambda}{d\theta}(\boldsymbol{x}_{i-1} + \boldsymbol{v}_{i-1} r, \boldsymbol{v}_{i-1};\theta)\, dr. \tag{4.23}$$

Note the interchange of integral and derivative in the second equation, which uses the same justification as lemma 3.5: $\frac{d\lambda}{d\theta}$ may only exist a.e., but for t where $\frac{d\lambda}{d\theta}$ may fail to exist we have that λ is continuously zero. The total derivative appears due to the dependency of both λ as well as $\boldsymbol{x}_{i-1}, \boldsymbol{v}_{i-1}$ on θ.

To obtain $\frac{d\lambda}{d\theta}$, recall that $\lambda(\boldsymbol{x}, \boldsymbol{v};\theta) = \max\{\boldsymbol{v}^\top \nabla_{\boldsymbol{x}} \psi(\boldsymbol{x};\theta), 0\}$, and differentiate the two pieces separately (yielding an a.e. derivative). In the zero case, the derivative is clearly zero, so we focus on the positive case and compute the total derivative. Let (r) be shorthand for evaluating in $(\boldsymbol{x}_{i-1} + \boldsymbol{v}_{i-1} r;\theta)$. Then

$$\frac{d}{d\theta} \boldsymbol{v}_{i-1}^\top \nabla_{\boldsymbol{x}} \psi(r) = \nabla_{\boldsymbol{x}} \psi(r)^\top \frac{\partial \boldsymbol{v}_{i-1}}{\partial \theta} + \boldsymbol{v}_{i-1}^\top \frac{d \nabla_{\boldsymbol{x}} \psi}{d\theta}(r). \tag{4.24}$$

The total derivative of the the gradient of the potential is by the chain rule

$$\frac{d\nabla_{\boldsymbol{x}}\psi}{d\theta} = \frac{\partial\nabla_{\boldsymbol{x}}\psi}{\partial\boldsymbol{x}}\frac{\partial\boldsymbol{x}}{\partial\theta} + \frac{\partial\nabla_{\boldsymbol{x}}\psi}{\partial\theta} = \nabla^2_{\boldsymbol{x}}\psi\frac{\partial\boldsymbol{x}}{\partial\theta} + \frac{\partial\nabla_{\boldsymbol{x}}\psi}{\partial\theta} \tag{4.25}$$

where we recovered the Hessian $\nabla^2_{\boldsymbol{x}}\psi$. And since $\frac{d}{d\theta}(\boldsymbol{x}+\boldsymbol{v}r) = \frac{\partial\boldsymbol{x}}{\partial\theta} + \frac{\partial\boldsymbol{v}}{\partial\theta}r$ we finally write (4.24) as

$$\frac{d}{d\theta}\boldsymbol{v}_{i-1}^\intercal\nabla_{\boldsymbol{x}}\psi(r) = \nabla_{\boldsymbol{x}}\psi(r)^\intercal\frac{\partial\boldsymbol{v}_{i-1}}{\partial\theta} + \boldsymbol{v}_{i-1}^\intercal\left(\nabla^2_{\boldsymbol{x}}\psi(r)\left[\frac{\partial\boldsymbol{x}_{i-1}}{\partial\theta} + r\frac{\partial\boldsymbol{v}_{i-1}}{\partial\theta}\right] + \frac{\partial\nabla_{\boldsymbol{x}}\psi}{\partial\theta}(r)\right). \tag{4.26}$$

Although we have the ability to compute this expression, we are not yet done as (4.23) requires the integral of the total derivative. We can achieve some simplification through integration by parts: for the first term

$$\int\nabla_{\boldsymbol{x}}\psi(r)^\intercal\frac{\partial\boldsymbol{v}_{i-1}}{\partial\theta}\,dr = r\nabla_{\boldsymbol{x}}\psi(r)^\intercal\frac{\partial\boldsymbol{v}_{i-1}}{\partial\theta} - \int r\,\boldsymbol{v}_{i-1}^\intercal\nabla^2_{\boldsymbol{x}}\psi(r)\frac{\partial\boldsymbol{v}_{i-1}}{\partial\theta}\,dr \tag{4.27}$$

cancelling a part of the second term, and noting

$$\int\boldsymbol{v}_{i-1}^\intercal\nabla^2_{\boldsymbol{x}}\psi(r)\frac{\partial\boldsymbol{x}_{i-1}}{\partial\theta}\,dr = \int\frac{d}{dr}\left(\nabla_{\boldsymbol{x}}\psi(r)^\intercal\frac{\partial\boldsymbol{x}_{i-1}}{\partial\theta}\right)dr = \nabla_{\boldsymbol{x}}\psi(r)^\intercal\frac{\partial\boldsymbol{x}_{i-1}}{\partial\theta} \tag{4.28}$$

we can substitute both to obtain a succinct antiderivative of (4.26) (i.e. still in the positive case)

$$\int\frac{d}{d\theta}\boldsymbol{v}_{i-1}^\intercal\nabla_{\boldsymbol{x}}\psi(r)\,dr = \nabla_{\boldsymbol{x}}\psi(r)^\intercal\left(\frac{\partial\boldsymbol{x}_{i-1}}{\partial\theta} + r\frac{\partial\boldsymbol{v}_{i-1}}{\partial\theta}\right) + \int\boldsymbol{v}_{i-1}^\intercal\frac{\partial\nabla_{\boldsymbol{x}}\psi}{\partial\theta}(r)\,dr. \tag{4.29}$$

Combining everything, and ensuring we evaluate the antiderivative accordingly to the positivity restriction (which can be done by partitioning the time interval on the zeros of the potential), we arrive at (4.17).

Next, (4.19) is immediately recovered by differentiating the segment relation $\boldsymbol{x}_i = \boldsymbol{x}_{i-1} + \boldsymbol{v}_{i-1}\tau_i$ and hence the same for both reflections and refreshments.

Finally, (4.20) depends on the event type. In the case of a refreshment, it suffices to note that the new velocity is independent of the parameter. In the case of a reflection, using the expression differentiating the reflection rule

$$\frac{\partial\boldsymbol{v}_i}{\partial\theta} = \frac{\partial\boldsymbol{v}_{i-1}}{\partial\theta} - 2\frac{dP_{\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x}_i;\theta)}\boldsymbol{v}_{i-1}}{d\theta}, \tag{4.30}$$

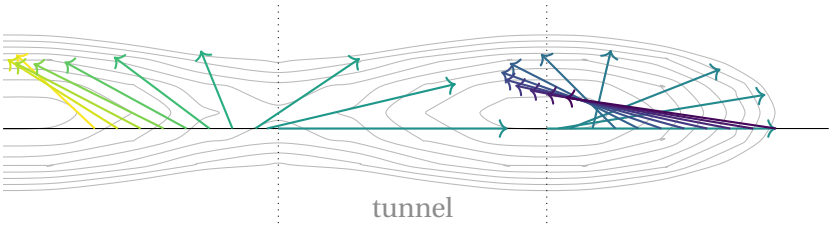
expanding the projection derivative by (4.10) and obtaining the total derivative of the gradient of the potential from (4.25) evaluated in $(\boldsymbol{x}_i;\theta)$, we arrive at the result. This completes all expressions. □

Similarly to the previous chapter, lemmata 4.7 and 4.8 allow us to obtain an estimator under the smoothness assumption for this section. The computations are more involved than in one dimension, but can be done sequentially top to bottom according to lemma 4.8 for each segment.

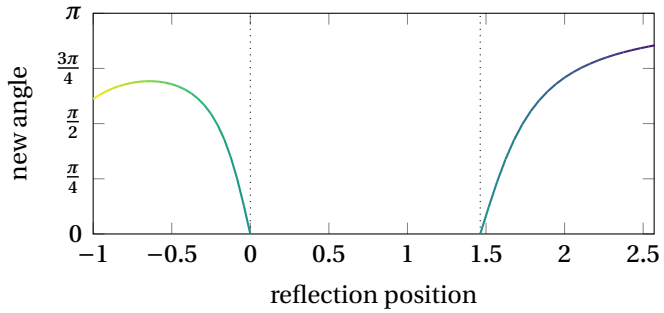
4.3 GENERAL CASE

The assumption of quasiconvexity excludes many popular models such as Gaussian mixtures, so without lifting it the applicability of the estimator would be limited. Up until this point the arguments have essentially followed the previous chapter, only with a more complicated coupling. Hence, one could reasonably believe that we also will extend the multi-dimensional case by separately handling jumps across tunnels with a stratified derivative.

However, the Bouncy Particle sampler reflections behave differently close to tunnels compared to the Zig-Zag sampler, since the reflection angle depends on the potential. At the opening of a tunnel, $\boldsymbol{v}^\mathsf{T}\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta) = 0$, and assuming $\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta) \neq 0$ we will therefore have $R(\boldsymbol{x},\boldsymbol{v}) = \boldsymbol{v}$ causing a 'reflection' leaving the velocity unchanged. The condition that $\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta) \neq 0$ is important, because otherwise the reflection rule is undefined, and the limit approaching the point may be non-zero.



(a) Sample reflections at different times, showing the new direction.



(b) The new trajectory angle as a function of reflection position. Recall that reflections cannot occur inside a tunnel, so the reflection position 'jumps' between 0 and 1.5.

Figure 4.2: Illustrations of the continuity wrt θ of the trajectory across tunnels, in terms of dependency on reflection position.

Thus, even though the reflection times jump when crossing the tunnel, the trajectory itself is identical for reflections at either end. In particular, there will be no performance difference in this limiting case. This suggests the rather surprising conclusion that tunnels do not matter at all for the validity of the derivative, because the trajectory changes continuously as we perturb the parameter. Handling the tunnels separately as done in one dimension would, the fact that the events cannot be critical notwithstanding, simply lead to a zero term and provide no contribution to the resulting estimator.

There is no special result in this section, as the conditions for unbiasedness in theorem 2.2 impose no requirements on the underlying representation of the performance derivative; although the skeleton point sensitivities often are discontinuous across the tunnel (as in the example of fig. 4.2), as long as the whole trajectory is sufficiently well-behaved the estimator will be unbiased. The observant reader might note that the continuity of the trajectory is not sufficient, as we require Lipschitz continuity or at least an integrable bound on the performance derivative. Conditions need to be imposed on f and ψ for this to hold, and the more complicated interlinked expressions of lemma 4.8 makes boundedness harder to show than in one dimension, where it is incorporated into the FCCE construction. Furthermore, distributions where $\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta) \neq 0$ in more than isolated points could theoretically cause discontinuities at tunnels.

Ultimately, in practice the estimator obtained from lemmata 4.7 and 4.8 appears to work numerically, and one can then investigate the theoretical unbiasedness on a case-by-case basis.

4.4 EXAMPLES

Like the previous chapter, we implement the estimator we have derived in Julia [33], using `ZigZagBoomerang.jl` [34] to run the Bouncy Particle sampler. (More details on the implementation are available in appendix B.) We once again run the sampler starting in the mean with a random initial direction.

Compared to the previous chapter, this estimator has significantly increased variance. Depending on the properties of $\nabla_{\boldsymbol{x}}\psi$, relatively large perturbations may occur in the velocity with significant impact on the trajectory. Reflections close to tunnel openings (see once again fig. 4.2, where the angle varies over the entire half circle) lead to particularly large perturbations. Hence, for numerical purposes we must increase the refreshment rate λ_{refr} to combat large velocity perturbations, noting that refreshments reset the accumulated velocity sensitivity. It is important to find a balance, since increasing λ_{refr} implies increasing the computational effort required simply due to more events occurring per time unit.

Example 4.9 (Gaussian correlation)**.** Consider $\boldsymbol{X}^\theta \sim \mathsf{N}(0, \Sigma(\theta))$ where

$$\Sigma(\theta) = \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix} \implies \Sigma^{-1}(\theta) = \frac{1}{1-\theta^2}\begin{pmatrix} 1 & -\theta \\ -\theta & 1 \end{pmatrix} \tag{4.31}$$

with the correlation $\theta \in (-1, 1)$. This target distribution is quasiconcave and thus satisfies assumption 4.5. It is an artificial example, since we can continuously reparameterize as

$$\boldsymbol{X}^\theta = \begin{pmatrix} 1 & 0 \\ \theta & \sqrt{1-\theta^2} \end{pmatrix} \boldsymbol{X}^0 \tag{4.32}$$

thus avoiding any differentiation of the sampler, but it still illustrates the easiest possible case for our method. (Similarly, we could apply the inverse transform for preconditioning or aligning the refreshment directions.)

We let $f(\boldsymbol{x}) = x_1 x_2$, which in this zero mean, unit variance case will lead to us estimating the correlation of the two components. Hence the true expectation and its derivative are by construction

$$\mathbb{E}[f(\boldsymbol{X}^\theta)] = \theta, \quad \frac{\partial}{\partial\theta}\mathbb{E}[f(\boldsymbol{X}^\theta)] = 1. \tag{4.33}$$

The required derivatives of the potential are

$$\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta) = \Sigma^{-1}(\theta)\boldsymbol{x}, \quad \nabla^2_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta) = \Sigma^{-1}(\theta), \tag{4.34}$$

$$\frac{\partial}{\partial\theta}\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta) = \frac{1}{(1-\theta^2)^2}\begin{pmatrix} 2\theta & -1-\theta^2 \\ -1-\theta^2 & 2\theta \end{pmatrix}\boldsymbol{x}. \tag{4.35}$$

The final step required to run the sampler and obtain derivative estimates is tuning the refreshment rate. We find by experimentation that $\lambda_{\text{ref}} = 8$ yields sufficiently stable single estimates (for larger θ we probably could adjust it even higher). A relatively high refreshment rate means the process becomes quite diffusive (see fig. 4.3), with about 93% of the events being refreshments.

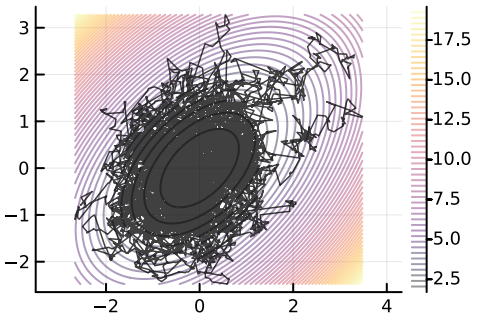


Figure 4.3: Sample trajectory for Bouncy Particle sampler targeting a 2D Gaussian parameterized by correlation, with $T = 10^3$, $\theta = 0.5$, and $\lambda_{\text{ref}} = 8$.

Results for several different parameter values are shown in fig. 4.4. The variability is indeed much greater than in one dimension, and it visibly grows with increasing correlation θ as the potential becomes steeper. Nevertheless, we still achieve reasonably good single estimates for longer runs, and taking the mean of 10–20 runs essentially yields the true value with two decimals of precision. Again, this represents the ideal scenario for the estimator, as we have no tunnels that generally imply derivatives that are much greater in magnitude.

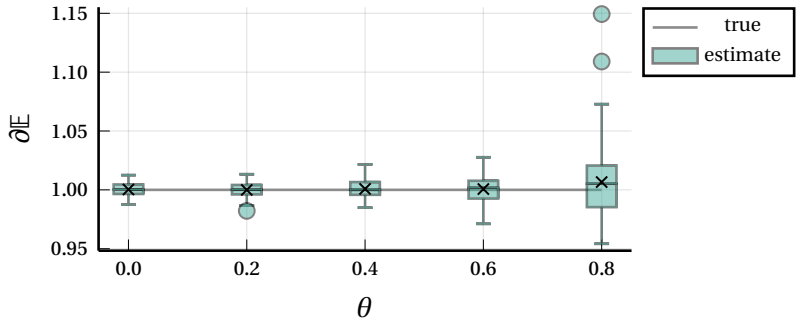


Figure 4.4: Derivative estimates for Bouncy Particle sampler targeting a 2D Gaussian parameterized by correlation, with $f(\boldsymbol{x}) = x_1 x_2$. At each θ we sample 100 trajectories for $T = 10^6$ and show means ($\times$) and boxplots.

Example 4.10 (Rosenbrock-type 'banana' density)**.** Consider $\boldsymbol{X}^\theta$ parameterized by $\theta > 0$ and distributed according to the density

$$\pi([x_1\ x_2];\theta) = \varphi\left(\sqrt{\gamma_x}x_1\right)\varphi\left(\sqrt{\gamma_y}(x_2 - \theta x_1^2)\right), \quad [x_1\ x_2] \in \mathbb{R}^2 \tag{4.36}$$

where $\varphi(t)$ is the standard one-dimensional Normal density and $\gamma_x, \gamma_y > 0$ are precision constants. We will fix $\gamma_x = \frac{1}{4}$ and $\gamma_y = \frac{4}{9}$. This density is a Gaussian smoothing of a Rosenbrock-type function, which are common test cases in optimization [38]. It leads to a curved ridge which must be explored by the sampler, and the distribution is non-quasiconvex with tunnels occurring across the curve (see fig. 4.5). However, the sampler does not need to cross tunnels to explore the whole ridge, and thus this is a fairly weak form of non-quasiconvexity in the problem, with little impact for small θ.

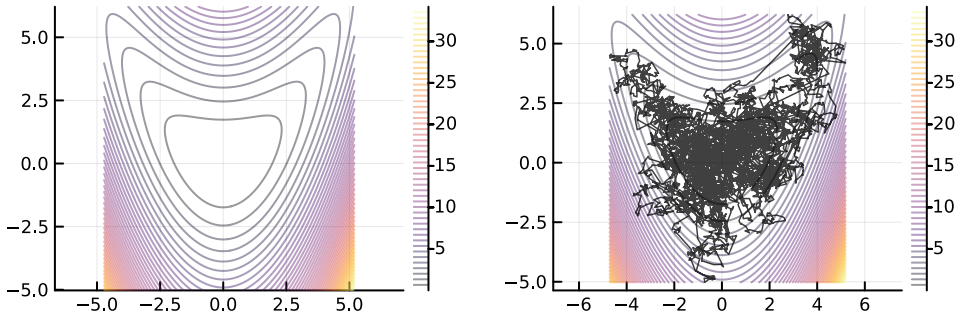


Figure 4.5: Contour and sample trajectory for Bouncy Particle sampler targeting the 'banana' density, with $T = 10^3$, $\theta = 0.25$, and $\lambda_{\text{ref}} = 8$.

We let $f \equiv \mathbf{1}_{[0,\infty)^2}$, i.e. the indicator of the first quadrant. This choice is not differentiable and thus strictly speaking violates our assumptions. However, following a previous remark we can still make sense of $\int \nabla f$ using distributional derivatives, yielding contributions from crossings of the boundary; in

fact, for numerical purposes and assuming unbiasedness holds, we do not need to worry about these details and could replace lemma 4.7 by AD on a function computing the integral, using lemma 4.8 as derivatives of the skeleton points.

The required derivatives of the potential are

$$\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta) = \begin{pmatrix} x_1(\gamma_x - 2\theta(x_2 - \theta x_1^2)\gamma_y) \\ (x_2 - \theta x_1^2)\gamma_y \end{pmatrix}, \quad \frac{\partial}{\partial\theta}\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta) = \begin{pmatrix} (4\theta x_1^3 - 2x_1x_2)\gamma_y \\ -x_1^2\gamma_y \end{pmatrix}, \tag{4.37}$$

$$\nabla_{\boldsymbol{x}}^2\psi(\boldsymbol{x};\theta) = \begin{pmatrix} \gamma_x - 2\theta(x_2 - 3\theta x_1^2)\gamma_y & -2\theta x_1\gamma_y \\ -2\theta x_1\gamma_y & \gamma_y \end{pmatrix} \tag{4.38}$$

and here there is no closed form for the true expectation and its derivative; we obtain various hypergeometric series, which can be computed numerically.

We set $\lambda_{\text{ref}} = 8$, and run the estimator for a few different parameter values to obtain the results in fig. 4.6. With a high refreshment rate we can still obtain acceptable estimates in a non-quasiconvex setting, although the scaling is worse and we likely need to increase the refreshment rate for larger θ.

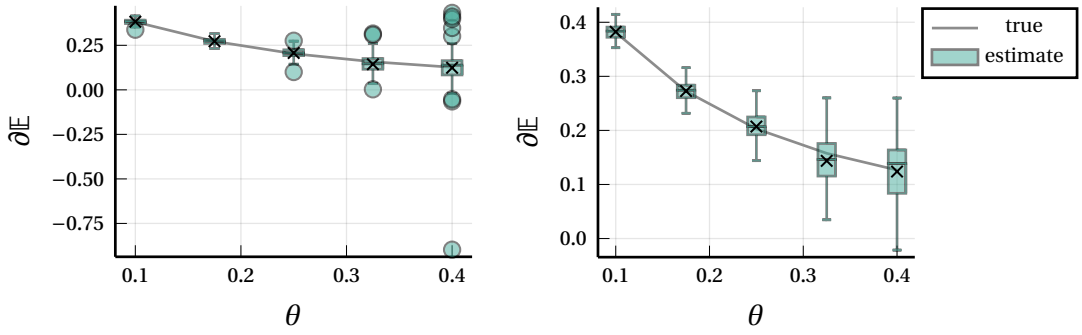


Figure 4.6: Derivative estimates for Bouncy Particle sampler targeting the 'banana' density, with $f \equiv \mathbf{1}_{[0,\infty)^2}$. At each θ we sample 100 trajectories for $T = 10^6$ and show means (×) and boxplots. To the right, enlarged version without outliers.

Example 4.11 (Gaussian mixture)**.** We return to the standard example of a Gaussian mixture. Consider (with abuse of notation) the bimodal distribution $\boldsymbol{X}^\theta \sim \frac{1}{2}\mathsf{N}(r[\cos\theta, \sin\theta], I) + \frac{1}{2}\mathsf{N}(-r[\cos\theta, \sin\theta], I)$ with $\theta \in \mathbb{R}$, $r > 1$, and $f \equiv \mathbf{1}_{[0,\infty)^2}$. This corresponds to the modes of the mixture rotating about the origin as θ changes, yielding a non-trivial shape dependency while keeping tunnel length constant. The potential is non-quasiconvex, and furthermore it is necessary to cross tunnels to sample properly from the whole distribution. Hence, there is no avoiding large perturbations from reflections close to tunnels on longer trajectories, further increasing the variance of the estimator and forcing even higher refreshment rates.

The required derivatives of the potential are

$$\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta) = \begin{pmatrix} x_1 - r\cos(\theta)\tanh(rx_1\cos(\theta) + rx_2\sin(\theta)) \\ x_2 - r\sin(\theta)\tanh(rx_1\cos(\theta) + rx_2\sin(\theta)) \end{pmatrix}, \tag{4.39}$$

$$\frac{\partial}{\partial\theta}\nabla_{\boldsymbol{x}}\psi(\boldsymbol{x};\theta) = \begin{pmatrix} \frac{r^2\cos(\theta)(-x_2\cos(\theta)+x_1\sin(\theta))}{\cosh^2(rx_1\cos(\theta)+rx_2\sin(\theta))} + r\sin(\theta)\tanh(rx_1\cos(\theta) + rx_2\sin(\theta)) \\ \frac{r^2\sin(\theta)(-x_2\cos(\theta)+x_1\sin(\theta))}{\cosh^2(rx_1\cos(\theta)+rx_2\sin(\theta))} - r\cos(\theta)\tanh(rx_1\cos(\theta) + rx_2\sin(\theta)) \end{pmatrix}, \tag{4.40}$$

$$\nabla_{\boldsymbol{x}}^2\psi(\boldsymbol{x};\theta) = \begin{pmatrix} 1 - \frac{r^2\cos^2(\theta)}{\cosh^2(rx_1\cos(\theta)+rx_2\sin(\theta))} & -\frac{r^2\sin(\theta)\cos(\theta)}{\cosh^2(rx_1\cos(\theta)+rx_2\sin(\theta))} \\ -\frac{r^2\sin(\theta)\cos(\theta)}{\cosh^2(rx_1\cos(\theta)+rx_2\sin(\theta))} & 1 - \frac{r^2\sin^2(\theta)}{\cosh^2(rx_1\cos(\theta)+rx_2\sin(\theta))} \end{pmatrix} \tag{4.41}$$

and we omit the true expectation and its derivative, which do not exist in closed form but are amenable to numerical computation.

We let $r = \frac{3}{2}$. Experimentation indicates we now have to increase to $\lambda_{\text{ref}} = 12$ to see some convergence of the mean, in doing so compensating for the increased velocity variance. At such a high refreshment rate the process becomes extremely diffusive; compare fig. 4.7 with the more standard fig. 2.1b. Such a configuration would definitely be less than ideal for generating samples from the distribution, but without it the derivative estimates become unstable.

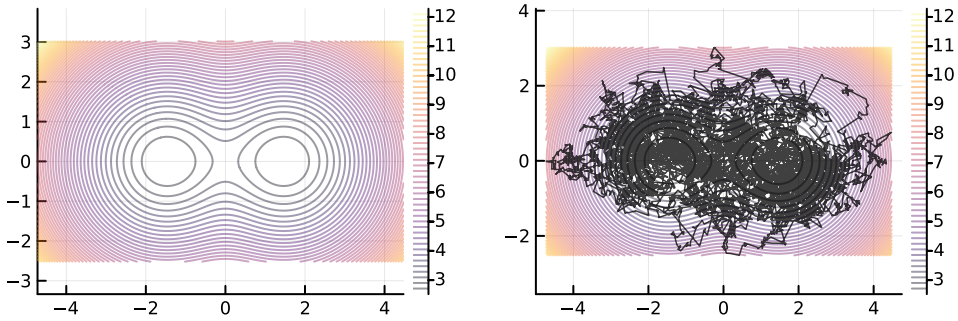


Figure 4.7: Contour and sample trajectory for Bouncy Particle sampler targeting a Gaussian mixture, with $T = 10^3$, $\theta = 0.0$, and $\lambda_{\text{ref}} = 12$.

We run the estimator for a few different parameter values to obtain the results in fig. 4.8. Since we require more samples to obtain a good estimate of the mean, we reduce the trajectory length significantly to avoid excessively long computations. Shorter trajectories also contribute to the observed greater variability, with several outliers far from the true value. Unlike our first example, here we do not yet achieve a sufficiently stable single-run estimate of the derivative, but the estimator does still appear to be unbiased.

In fact, we have throughout these examples systematically compared the true value with the finite-horizon estimators, thus suggesting a stronger claim

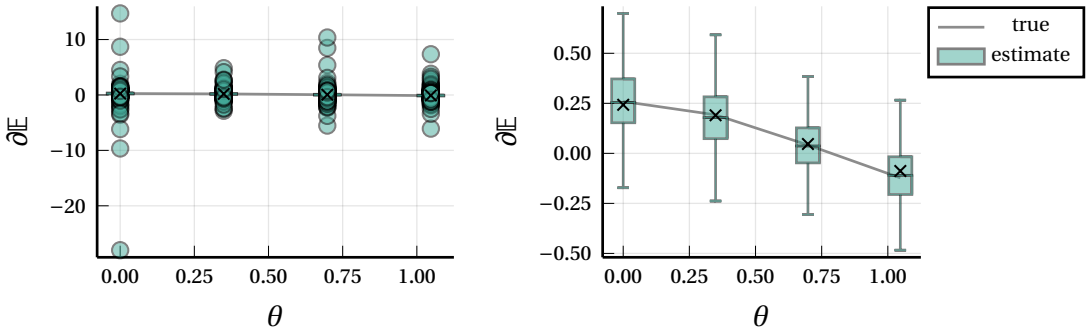


Figure 4.8: Derivative estimates for Bouncy Particle sampler targeting a Gaussian mixture, with $f \equiv \mathbf{1}_{[0,\infty)^2}$. At each θ we sample 1000 trajectories for $T = 10^3$ and show means ($\times$) and boxplots. To the right, enlarged version without outliers.

than just that of having an unbiased estimate of the derivative of the finite-horizon expectation functional, which may not estimate the true derivative of the expectation. That our numerical estimates appear to be unbiased for the true derivative indicates that we also have vanishing bias from the starting point as the sampler is mixing or at least symmetric truncation error, which is important to achieve a single-run derivative estimator.

5 CONCLUSION

We have in this thesis presented a blueprint for the construction of a general differentiable Monte Carlo method based on PDMPs, by two exemplifying such constructions that are strongly related, each with its own difficulties. For both samplers that we consider, we have applied the same technique of a carefully constructed coupling of trajectories to derive the pathwise derivatives of the sampler trajectories, noting that the actual simulation of the trajectories need not be done by the coupling for the estimator to be valid.

In one dimension we developed an unbiased, low-variance, and single-run derivative estimator, and we also proved it consistent in a special case, conjecturing its consistency in general. Note that consistency is the key step that allows us to go from an unbiased estimator for the derivative of the finite-horizon performances to an estimator for the desired derivative of the expectation, much like we would use our sampler for the expectation. The general unbiased estimator required the correction of the pathwise derivatives by stratifying on discontinuities in the sampler, using our formulation of a filtered theory for stochastic derivatives that partially integrate these discontinuities. In doing so, we recoupled the primal and shadow trajectories across discontinuities, while estimating the impact of alternative trajectories in a way which requires no extra simulation.

We find particularly appealing this notion of stratification, where we have both infinitesimal perturbations that occur almost surely and finite perturbations that occur almost never, only solving our problem by handling these cases separately. It challenges our intuition about the behaviour of the derivatives that something, which is never observed in practice, can have a disproportionate impact on the expectation. The strategy to recover the coupling, where we create alternative trajectories that eventually recouple with the primal trajectory to obtain low-variance estimates of the impact, is also an example of a promising technique which we would like to highlight, which can be applied in other settings than PDMPs (see e.g. [39] where both ideas are applied to Metropolis-Hastings).

In multiple dimensions we used the same ideas to develop an estimator, by the power of the chosen coupling showing that no corrections were required at all in many cases, since the trajectory difference became continuous with respect to the perturbations. With a.s. valid skeleton point sensitivities we open up many applications where arbitrary stochastic programs use the trajectories as input, as our estimates then enable AD through sampling routines.

Here, the strength of the technique of coupling shines. Ultimately, the shadowing couplings show that the effect of perturbations are local and propagate, which makes our estimation possible. In one dimension this can be understood by the sensitivity essentially stretching time proportional to the derivative. Furthermore, to obtain the desired *perfect* shadowing in multiple dimensions, we could not resort to the naïve choice of common randomness (i.e. common seeds for the random number generator), but had to adjust the simulation procedure in order to achieve our goal.

5.1 PRACTICAL CONSIDERATIONS

A natural consideration in discussing the results is how far they are from use in real-world applications. In the one-dimensional case, we have as previously mentioned essentially developed the theory necessary to show that the estimator fulfils our objective. The only missing piece is a more mature implementation. However, there is a limit to the possible intractability of practical one-dimensional problems, and few if any new major applications are made possible by this estimator.

The multi-dimensional case, on the other hand, does suggest a scheme which solves a current gap in applications of an unbiased single-run estimator. In particular, it will have the remarkable ability to handle unknown but parameter-dependent normalization constants. Nevertheless, we do not yet have general conditions on its applicability, and we must consider the assumptions carefully when using the estimator to ensure it is unbiased. Sufficient conditions for unbiasedness which are easier to check together with a mature, performant implementation are likely required to adopt the estimator in practice. A proof of consistency, which is also desirable for practical purposes, will require additional theory beyond what is presented in this thesis.

Furthermore, the multi-dimensional estimator suffers from relatively high variance compared to the one-dimensional estimator (although it should be noted that we do not compare either to other approaches), and in harder settings one cannot yet rely on a single run for a good approximation. The study of methods for variance reduction is thus of high importance to further enable application of the estimator, and in particular reducing the refreshment rate to improve performance while still maintaining convergence.

One idea is to pool the results of shorter trajectories that already have reached stationarity rather than running long trajectories which accumulate sensitivity contributions over time. To alleviate bias from incomplete exploration of the target density this requires varying the starting state and representing the estimator as conditional on the starting state:

$$\mathbb{E}_{\boldsymbol{Z}_0^\theta \sim \mu_\theta}\left[F(\boldsymbol{Z}^\theta)\right] = \mathbb{E}_{\boldsymbol{Z}_0^\theta \sim \mu_\theta}\left[\mathbb{E}\left[F(\boldsymbol{Z}^\theta)\middle|\boldsymbol{Z}_0^\theta\right]\right] = \int \mathbb{E}\left[F(\boldsymbol{Z}^\theta)\middle|\boldsymbol{z}\right]\pi_\theta(\boldsymbol{z})\,d\boldsymbol{z}. \tag{5.1}$$

In practice, we use the sampler itself to obtain starting states $\boldsymbol{Z}_0^\theta$ from a 'mother' trajectory tuned for that purpose. Hence, the starting state itself is sensitive to the parameter, and this sensitivity cannot a priori be neglected when differentiating. We handle this sensitivity by a hybrid approach, using the score method to compensate:

$$\frac{\partial}{\partial\theta}\mathbb{E}_{\boldsymbol{Z}_0^\theta\sim\mu_\theta}\left[F(\boldsymbol{Z}^\theta)\right] = \int\left(\frac{\partial}{\partial\theta}\mathbb{E}\left[F(\boldsymbol{Z}^\theta)\middle|\boldsymbol{z}\right]\pi_\theta(\boldsymbol{z}) + \mathbb{E}\left[F(\boldsymbol{Z}^\theta)\middle|\boldsymbol{z}\right]\frac{\partial}{\partial\theta}\pi_\theta(\boldsymbol{z})\right)d\boldsymbol{z} \tag{5.2}$$

$$= \mathbb{E}_{\boldsymbol{Z}_0^\theta\sim\mu_\theta}\left[\frac{\partial}{\partial\theta}\mathbb{E}\left[F(\boldsymbol{Z}^\theta)\middle|\boldsymbol{Z}_0^\theta\right]\right] + \mathbb{E}_{\boldsymbol{Z}_0^\theta\sim\mu_\theta}\left[\mathbb{E}\left[F(\boldsymbol{Z}^\theta)\middle|\boldsymbol{Z}_0^\theta\right]\frac{\partial}{\partial\theta}\log\pi_\theta(\boldsymbol{Z}_0^\theta)\right] \tag{5.3}$$

where the first term is the pooling of our 'daughter' trajectory derivative estimators and the second term is the score method compensation for sensitivity in the starting state. Note that for sufficiently long 'daughter' trajectories, the second term tends to zero as the 'memory' of the starting state is lost, the expectation estimate becomes effectively independent, and the expected score is zero. Of course, such pooling will not fulfil our goal of a single-run estimator. However, this scheme suggests the validity of restarting the trajectory at points where the score is zero, and so the sensitivities of a long trajectory could be reset at such points, thus reducing the accumulation of sensitivities without introducing bias or requiring pooling.

5.2 FUTURE WORK

There are several directions in which one can continue this work. Firstly, to prepare for practical applications a major simulation study of the variance with comparisons to other methods is required. Although we note that each gradient estimation method fills different niches, and the assumptions and knowledge about the target required to apply, say, the score method, differs from applying our PDMP-based pathwise derivative estimators, it is still of interest to have a reference comparison. The existing literature on pathwise derivatives already contains comparisons with other gradient estimation methodology that one can build on. This would also certainly entail investigating possibilities for variance reduction, such as the one discussed in the previous section.

Another extension would be to other performance functionals than the expectation functional. The theory of pathwise derivatives and stratified derivatives are not restricted to any specific functional. Furthermore, we made a major point of separating our derivations into two lemmata, first obtaining the derivative of the functional with respect to the trajectory, then obtaining the derivative of the trajectory with respect to the parameter. Much of the theory showcased can therefore be adapted to other functionals, thus enabling even more applications than what has been considered in this thesis. In par-

ticular, we are not restricted to only working with the marginal distributions and e.g. gradient estimates of the autocovariance

$$C(Z,h) = \frac{1}{T}\int_0^{T-h} (Z_t - \mathbb{E}[X])(Z_{t+h} - \mathbb{E}[X])\,dt \tag{5.4}$$

with respect to hyperparameters could be used for tuning, in some sense literally measuring the 'performance' of the sampler. Compared to the tuning application covered in the introduction, we have a much wider class of performance metrics available in this setting as we are not dependent on an explicit reparameterization in terms of hyperparameters.

It is also interesting to further investigate other PDMP samplers and related extensions. An aspect of this is extending to samplers on other domains than $\mathbb{R}^d$ and with target densities with respect to other measures than Lebesgue, where we hope that our general methodology is applicable as well. We opted early to not use the Zig-Zag sampler in multiple dimensions, because of the possibility of unavoidable discrete perturbations in the velocities after reflection. However, our intuition of difficulties with the Bouncy Particle sampler due to reflections and even tunnels was disproven by further scrutiny. There is a hope that the restricted set of velocities of the Zig-Zag sampler may alleviate much of the variance problems due to velocity perturbations that we see in our multidimensional estimator. Perhaps the right choice of coupling, such that we could obtain a similar low-variance solution to recoupling the alternative trajectory as that of tunnels in one dimension, will make the Zig-Zag sampler in multiple dimensions a viable low-variance approach.

Finally, the Bouncy Particle sampler examples in the previous chapter where we increased λ_{ref} to achieve sufficient stability led to much more diffusive trajectories. It suggests the importance not of the sampler position itself, as the high refreshment rate leads to less effective sampling of the target, but rather of how the shape of the distribution is perturbed, with the increased exploration of 'bad' directions by the trajectories. There is the possibility of the existence of some diffusion limit, perhaps in which the sampler approximates a Hamiltonian SDE. Stochastic calculus, which we emphasized the differences of in the opening statements about stochastic derivatives, might very well return as a tool to analyse this emergent phenomenon. We also saw that a portion of the variability comes from the inability to explore the distribution without tunnels, and there is the possibility that the diffusivity could be restricted by better ways to move between modes or a symmetric use of the tunnels by the introduction of teleportation events as in [27].

In conclusion, this thesis only scratches the surface of the possibilities that PDMP-based differentiable Monte Carlo methods provide. It is our hope that in presenting these first examples and theoretical steps we have motivated further interest in this methodology, which in time may pose a solution to many of the stochastic gradient estimation problems introduced at the beginning of the thesis.

BIBLIOGRAPHY

1. Robert, C. & Casella, G. A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Statistical Science* **26(1)** (1st Feb. 2011).
2. Neal, R. M. MCMC using Hamiltonian dynamics. in *Handbook of Markov Chain Monte Carlo* (Chapman & Hall/CRC, 2011). arXiv: 1206.1901v1.
3. Fearnhead, P., Bierkens, J., Pollock, M. & Roberts, G. O. Piecewise Deterministic Markov Processes for Continuous-Time Monte Carlo. *Statistical Science* **33(3),** 386–412 (1st Aug. 2018).
4. Mohamed, S., Rosca, M., Figurnov, M. & Mnih, A. Monte Carlo Gradient Estimation in Machine Learning. *Journal of Machine Learning Research* **21(132),** 1–62 (2020).
5. Glasserman, P. *Gradient estimation via perturbation analysis* (Kluwer Academic Publishers, Boston, 1991).
6. Kingma, D. P. & Welling, M. *Auto-Encoding Variational Bayes* in *Proceedings of the 2nd International Conference on Learning Representations* International Conference on Learning Representations (ICLR, Ithaca, 2014). arXiv: 1312.6114v11.
7. Baydin, A. G., Pearlmutter, B. A., Radul, A. A. & Siskind, J. M. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research* **18(153),** 1–43 (2018).
8. Van de Meent, J.-W., Paige, B., Yang, H. & Wood, F. *An Introduction to Probabilistic Programming* 19th Oct. 2021. arXiv: 1809.10756v2.
9. Chandra, K., Li, T.-M., Tenenbaum, J. & Ragan-Kelley, J. *Designing Perceptual Puzzles by Differentiating Probabilistic Programs* in *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings* SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference (ACM, Vancouver BC Canada, 7th Aug. 2022), 1–9.
10. Zoltowski, D. M., Cai, D. & Adams, R. P. *Slice Sampling Reparameterization Gradients* in *Advances in Neural Information Processing Systems* NeurIPS 2021. **34** (2021), 23532–23544.
11. Naesseth, C. A., Ruiz, F. J. R., Linderman, S. W. & Blei, D. M. *Reparameterization Gradients through Acceptance-Rejection Sampling Algorithms* 12th Feb. 2020. arXiv: 1610.05683v3.
12. Salimans, T., Kingma, D. P. & Welling, M. *Markov Chain Monte Carlo and Variational Inference: Bridging the Gap* in *Proceedings of the 32nd International Conference on Machine Learning* International Conference on Machine Learning. **37** (PMLR, 2015), 1218–1226.

13. Zhang, G., Hsu, K., Li, J., Finn, C. & Grosse, R. *Differentiable Annealed Importance Sampling and the Perils of Gradient Noise* in *Advances in Neural Information Processing Systems* NeurIPS 2021. **34** (2021), 19398–19410.
14. Campbell, A., Chen, W., Stimper, V., Hernandez-Lobato, J. M. & Zhang, Y. *A Gradient Based Strategy for Hamiltonian Monte Carlo Hyperparameter Optimization* in *Proceedings of the 38th International Conference on Machine Learning* International Conference on Machine Learning. **139** (PMLR, 2021), 1238–1248.
15. Heidergott, B. & Vázquez-Abad, F. J. Measure-Valued Differentiation for Markov Chains. *Journal of Optimization Theory and Applications* **136(2),** 187–209 (Feb. 2008).
16. Fu, M. & Hu, J.-Q. *Conditional Monte Carlo* (Springer US, Boston, MA, 1997).
17. Arya, G., Schauer, M., Schäfer, F. & Rackauckas, C. *Automatic Differentiation of Programs with Discrete Randomness* in *Advances in Neural Information Processing Systems* NeurIPS 2022. **36** (2022).
18. Kleijnen, J. P. & Rubinstein, R. Y. Optimization and sensitivity analysis of computer simulation models by the score function method. *European Journal of Operational Research* **88(3),** 413–427 (Feb. 1996).
19. Glynn, P. W. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM* **33(10),** 75–84 (Oct. 1990).
20. Çınlar, E. & Sollenberger, N. J. *Introduction to stochastic processes* Dover edition (Dover Publications, Inc, Mineola, New York, 2013).
21. Embrechts, P. & Hofert, M. A note on generalized inverses. *Mathematical Methods of Operations Research* **77(3),** 423–432 (June 2013).
22. Lewis, P. A. W. & Shedler, G. S. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly* **26(3),** 403–413 (Sept. 1979).
23. Bierkens, J., Fearnhead, P. & Roberts, G. The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data. *The Annals of Statistics* **47(3)** (1st June 2019).
24. Maruyama, G. & Tanaka, H. Ergodic Property of N-Dimensional Recurrent Markov Processes. *Memoirs of the Faculty of Science, Kyushi University. Series A* **13(2),** 157–172 (1959).
25. Bierkens, J., Roberts, G. O. & Zitt, P.-A. Ergodicity of the zigzag process. *The Annals of Applied Probability* **29(4)** (1st Aug. 2019).
26. Bierkens, J., Grazzi, S., Meulen, F. v. d. & Schauer, M. Sticky PDMP samplers for sparse and local inference problems. *Statistics and Computing* **33(1),** 8 (Feb. 2023).
27. Bierkens, J., Grazzi, S., Roberts, G. & Schauer, M. *Methods and applications of PDMP samplers with boundary conditions* 14th Mar. 2023. arXiv: 2303.08023v1.

28. Davis, M. H. A. Piecewise-Deterministic Markov Processes: A General Class of Non-Diffusion Stochastic Models. *Journal of the Royal Statistical Society. B* **46(3),** 353–388 (1984).
29. Davis, M. H. A. *Markov models and optimization* 1st ed. (Chapman & Hall, London ; New York, 1993).
30. Vanetti, P. *Piecewise-deterministic Markov chain Monte Carlo* PhD thesis (University of Oxford, 2019).
31. Bierkens, J. & Roberts, G. A piecewise deterministic scaling limit of lifted Metropolis–Hastings in the Curie–Weiss model. *The Annals of Applied Probability* **27(2)** (1st Apr. 2017).
32. Bouchard-Côté, A., Vollmer, S. J. & Doucet, A. The Bouncy Particle Sampler: A Nonreversible Rejection-Free Markov Chain Monte Carlo Method. *Journal of the American Statistical Association* **113(522),** 855–867 (3rd Apr. 2018).
33. Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. Julia: A Fresh Approach to Numerical Computing. *SIAM Review* **59(1),** 65–98 (Jan. 2017).
34. Schauer, M., Grazzi, S. & Scherrer, C. *mschauer/ZigZagBoomerang.jl* version v0.13.1. 31st Aug. 2022. doi:10.5281/ZENODO.3931118.
35. Elal-Olivero, D. Alpha-skew-normal distribution. *Proyecciones (Antofagasta)* **29(3)** (Dec. 2010).
36. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization* in *Proceedings of the 3rd International Conference for Learning Representations* International Conference on Learning Representations (ICLR, San Diego, 2015). arXiv: 1412.6980v9.
37. Laue, S., Mitterreiter, M. & Giesen, J. *Computing Higher Order Derivatives of Matrix and Tensor Expressions* in *Advances in Neural Information Processing Systems* NeurIPS 2018. **31** (2018).
38. Pagani, F., Wiegand, M. & Nadarajah, S. An n-dimensional Rosenbrock distribution for Markov chain Monte Carlo testing. *Scandinavian Journal of Statistics* **49(2),** 657–680 (June 2022).
39. Arya, G., Seyer, R., Schäfer, F., Lew, A., Huot, M., Mansinghka, V. K., Rackauckas, C., Chandra, K. & Schauer, M. *Differentiating Metropolis-Hastings to Optimize Intractable Densities* 13th June 2023. arXiv: 2306.07961v1.
40. Oswaldo De Oliveira. The Implicit and the Inverse Function Theorems: Easy Proofs. *Real Analysis Exchange* **39(1),** 207 (2014).
41. Klenke, A. *Probability Theory: A Comprehensive Course* 2nd ed. (Springer, London, 2014).

A SUPPLEMENTARY BACKGROUND

Here we collect a few extra preliminaries, required to fill in some technical details of the proofs but which may not be known to the reader in this form.

Theorem A.1 (corollary of [40, theorem 4]). *Let $F : D \times \Theta \to \Omega$, where D, Θ, Ω are open intervals in $\mathbb{R}$, be continuously differentiable. Consider $(t, \theta) \in D \times \Theta$, let $\omega = F(t, \theta)$, and suppose $\frac{\partial F}{\partial t}(t, \theta) \neq 0$. Then it holds that F is t-invertible in a neighbourhood of (t, θ), i.e. there exists a F^{-1} so that $F^{-1}(\omega; \theta) = t$, and*

$$\frac{\partial F^{-1}}{\partial \omega}(\omega; \theta) = \frac{1}{\frac{\partial F}{\partial t}(F^{-1}(\omega; \theta); \theta)}, \tag{A.1}$$

$$\frac{\partial F^{-1}}{\partial \theta}(\omega; \theta) = \frac{-\frac{\partial F}{\partial \theta}(F^{-1}(\omega; \theta); \theta)}{\frac{\partial F}{\partial t}(F^{-1}(\omega; \theta); \theta)}. \tag{A.2}$$

This is a special case of the inverse or implicit function theorems from multivariate calculus, applied to the explicit form rather than the usual implicit $G(t, \omega; \theta) = F(t; \theta) - \omega$. In fact, specialized forms for sampling random variables with the inversion method appear in the literature, see [5, theorem 1.3].

Theorem A.2 (Leibniz rule, [41, theorem 6.28]). *Let $(D, \mathscr{B}, \mu)$ be a measure space, let $\Theta \subset \mathbb{R}$ be a nontrivial open interval, and let $f : D \times \Theta \to \mathbb{R}$ be a function such that*

(i) for all $\theta \in \Theta$, $x \mapsto f(x; \theta)$ is μ-integrable,

(ii) for a.e. $x \in D$, $\theta \mapsto f(x; \theta)$ is differentiable with derivative $\theta \mapsto \frac{\partial f}{\partial \theta}(x; \theta)$,

(iii) there exists $h : D \to [0, \infty)$ that is μ-integrable such that $|\frac{\partial f}{\partial \theta}(\cdot; \theta)| \leq h$ μ-a.e. for all $\theta \in \Theta$.

Then, for all $\theta \in \Theta$, we have $x \mapsto \frac{\partial f}{\partial \theta}(x; \theta)$ is μ-integrable and

$$\frac{\partial}{\partial \theta} \int f(x; \theta)\, \mu(dx) = \int \frac{\partial f}{\partial \theta}(x; \theta)\, \mu(dx). \tag{A.3}$$

If the integral endpoints depend on θ, we handle this through application of the usual chain rule, which allows us to work around cases when we only have a.e. θ-differentiability.

B IMPLEMENTATION NOTES

The numerical results exhibited in the examples are obtained with the aforementioned proof-of-concept implementation of the results available at

> Seyer, R. Implementation archive for Differentiable Monte Carlo Samplers with Piecewise Deterministic Markov Processes. 2023. doi:10.5281/ZENODO.8028965.

Although the actual implementation is not the focus of this thesis, and the code is not a complete end product, we do have some remarks for the reader who wants to use the theoretical results for numerical computation.

In one dimension, most of the numerical difficulty lies in integration of $\frac{\partial\lambda}{\partial\theta}$. Imprecise computations of this integral may lead to considerable error in the final estimate. It is helpful to compute the zeros of $\frac{\partial}{\partial\theta}\psi(x;\theta)$, which not only appear explicitly in theorem 3.12 as tunnel openings, but also help numerically in the smooth case.

- The integrand $\frac{\partial\lambda}{\partial\theta}$ is piecewise continuous (and in pratice often piecewise smooth), and in general numerical methods will have smaller error if one handles the pieces separately. In terms of x these cut points correspond to the aforementioned zeros.
- It is important both for numerical, performance and variance reduction purposes to exploit the linearity of the integral in computing the change in performance due to a jump across a tunnel. The performance only needs to be computed for the removed segment to obtain an unbiased estimate for the performance based on the primal trajectory.

In higher dimensions, with the variable angle of the trajectory we no longer have the theoretical cancellation that simplified calculations considerably in the one-dimensional case. A larger computational effort is required, which means implementation choices play a larger role, and limiting the error in the steps improves the final estimates.

- As mentioned in the example section it is very important to have sufficiently large λ_{ref} for long trajectories. Refreshments allow the reset of the accumulated sensitivity in the velocity of the sampler, which otherwise over time leads to large deviations and catastrophic cancellation. However, a larger refreshment rate slows the convergence as the trajectory becomes more diffusive, and increases the computational effort since more events occur per time unit. It may be useful to use

out-of-core storage or online computations for the trajectory to avoid memory restrictions impacting the maximum possible λ_{ref}.

- The problem of finding zeros of $\nabla\psi$ becomes in higher dimensions the more general one of describing isopotentials. Although one can use standard root-finding methods (using gradients, since the potential is $\mathscr{C}^2$), if we have knowledge of these isopotentials or they have special structure (e.g. are polynomial curves) we can use faster and more precise methods. Zeros along the trajectory appear as cut points both for integration of $\frac{\partial\lambda}{\partial\theta}$ as well as points a_j, b_j in the partial integration of lemma 4.8.
- The many dot and matrix products that appear can be taken in an order that avoids allocation of large matrices, and in particular it is unnecessary to build the large 'outer product' matrices $\boldsymbol{p}_i\boldsymbol{p}_i^\mathsf{T}$ of the projection operator derivative.

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023
www.chalmers.se

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY