Table 1: Probability mass functions and cumulative distributions of integer-valued stochastic programs, where x is assumed in the domain of $X(p)$.

$X(p)$	Parameter	$\mathbb{P}(X(p) = x)$	$\mathbb{P}(X(p) \leq x)$	$\partial_p \mathbb{P}(X(p) \leq x)$
$\text{Ber}(p)$	Probability	$px + (1-p)(1-x)$	$(1-p) + px$	$x - 1$
$\text{Geo}(p)$	Probability	$p(1-p)^x$	$1 - (1-p)^x$	$x(1-p)^{x-1}$
$\text{Pois}(p)$	Rate	$p^x/(e^p x!)$	$\sum_{k=0}^{x} e^{-p} p^k / k!$	$-p^{x+1}/(e^p(x+1)!)$

In the supplement, we:

- provide a number of examples of stochastic derivatives and present hand-worked examples of how they compose (Appendix A),
- prove our formal statements regarding stochastic derivatives (Appendix B),
- provide an introduction to the particle filter methodology, and show how smoothed stochastic derivatives leads to unbiased differentiation of the resampling step (Appendix C),
- provide details about the hardware used for the experiments of the main text and the software dependencies of `StochasticAD.jl` (Appendix D).

A Examples of stochastic derivatives

In this section, we present examples of stochastic derivatives for a number of stochastic programs.

Example A.1 (Integer-valued stochastic programs)**.** The Bernoulli, Binomial, Geometric, and Poisson distributions are all discrete, and in fact (nonnegative) integer-valued. Suppose $X(p)$ follows one of these distributions, parameterized using the inversion method [31] over the interval $[0, 1]$. We present a worked derivation of the stochastic derivative of $X(p)$, following the construction used in Theorem 2.4.

We construct a stochastic derivative (δ, w, Y) at input p satisfying Definition 2.2. Since $X(p)$ is parameterized via the inversion method, $\mathrm{d}X(\varepsilon) \in \{-1, 0, 1\}$ for small enough ε. Thus, for small enough ε, the conditional distribution of $\mathrm{d}X(\varepsilon)$ given $X(p) = x$ is:

$$\mathbb{P}\left(\mathrm{d}X(\varepsilon) = 1 \mid X(p) = x\right) = \frac{\mathbb{P}(X(p+\varepsilon) = x + 1 \text{ and } X(p) = x)}{\mathbb{P}(X(p) = x)}, \tag{A.1}$$

$$\mathbb{P}\left(\mathrm{d}X(\varepsilon) = -1 \mid X(p) = x\right) = \frac{\mathbb{P}(X(p+\varepsilon) = x - 1 \text{ and } X(p) = x)}{\mathbb{P}(X(p) = x)}, \tag{A.2}$$

and $\mathrm{d}X(\varepsilon) = 0$ otherwise. According to the inversion method, $X(p)(\omega) = x$ for ω between $\mathbb{P}(X(p) \leq x - 1)$ and $\mathbb{P}(X(p) \leq x)$. Taking the derivative of the above quantities as $\varepsilon \to 0$ therefore yields:

$$w_+ := \lim_{\varepsilon \to 0} \frac{\mathbb{P}\left(\mathrm{d}X(\varepsilon) = 1 \mid X(p) = x\right)}{\varepsilon}, \tag{A.3}$$

$$= \frac{\frac{\mathrm{d}}{\mathrm{d}p}\mathbb{P}(X(p) \leq x)}{\mathbb{P}(X(p) = x)} \mathbf{1}\{\varepsilon \cdot \partial_p \mathbb{P}(X(p) \leq x) < 0\}, \tag{A.4}$$

and

$$w_- := \lim_{\varepsilon \to 0} \frac{\mathbb{P}\left(\mathrm{d}X(\varepsilon) = -1 \mid X(p) = x\right)}{\varepsilon}, \tag{A.5}$$

$$= \frac{\frac{\mathrm{d}}{\mathrm{d}p}\mathbb{P}(X(p) \leq x - 1)}{\mathbb{P}(X(p) = x)} \mathbf{1}\{\varepsilon \cdot \partial_p \mathbb{P}(X(p) \leq x - 1) > 0\}, \tag{A.6}$$

where the ε included on the right hand sides is an abuse of notation to express their dependence on the direction of the limit, and $\mathbf{1}\{C\} = \mathbf{1}_{\{C\}}$ for a condition C. By the differentiability of $\mathbb{P}(\mathrm{d}X(\varepsilon) \neq 0)$, $\mathrm{d}X(\varepsilon)/\varepsilon$ almost surely approaches 0, so $\delta = 0$. A simple choice for the bound B

Table 2: Stochastic derivatives of integer-valued stochastic programs

	Conditionally on $X(p) = x$	
$X(p)$	w_-	w_+
$\mathrm{Ber}(p)$	$-1/p \cdot \mathbf{1}\{x = 1 \text{ and } \varepsilon < 0\}$	$1/(1-p) \cdot \mathbf{1}\{x = 0 \text{ and } \varepsilon > 0\}$
$\mathrm{Bin}(n, p)$	$-x/p \cdot \mathbf{1}\{x > 0 \text{ and } \varepsilon < 0\}$	$(n-x)/(1-p) \cdot \mathbf{1}\{x < n \text{ and } \varepsilon > 0\}$
$\mathrm{Geo}(p)$	$x/\left(p(1-p)\right) \cdot \mathbf{1}\{x > 0 \text{ and } \varepsilon > 0\}$	$-(x+1)/p \cdot \mathbf{1}\{\varepsilon < 0\}$
$\mathrm{Pois}(p)$	$-x/p \cdot \mathbf{1}\{x > 0 \text{ and } \varepsilon < 0\}$	$\mathbf{1}\{\varepsilon > 0\}$

is 1: then $|\mathrm{d}X(\varepsilon)| > B\varepsilon = \varepsilon \iff \mathrm{d}X(\varepsilon) \in \{-1, 1\}$. We may now form w as the almost sure limit of $\mathbb{P}\left(A_B(\varepsilon) \mid X(p)\right)/\varepsilon = \mathbb{P}\left(\mathrm{d}X(\varepsilon) \neq 0 \mid X(p)\right)$ and Y as the limit of the conditional distribution of $X(p) + \mathrm{d}X(\varepsilon)$ given $X(p)$ and $\mathrm{d}X(\varepsilon) \neq 0$. By the above, w is given conditionally on $X(p) = x$ as

$$w = w_+ + w_-, \tag{A.7}$$

while the probability distribution of Y is given conditionally on $X(p) = x$ as

$$\mathbb{P}\left(Y = x + 1 \mid X(p) = x\right) = \frac{w_+}{w_+ + w_-}, \tag{A.8}$$

$$\mathbb{P}\left(Y = x - 1 \mid X(p) = x\right) = \frac{w_-}{w_+ + w_-}. \tag{A.9}$$

Table 2 lists the weights w_+ and w_- for the Bernoulli, Binomial, Poisson, and Geometric distributions, derived using the parameterizations given in Table 1. Using w_+ and w_-, we can easily express w and Y as above. Note that w_+ and w_- for a Binomial variable $X(p) \sim \mathrm{Bin}(n, p)$ may be derived via representation as the sum of Bernoulli variables, with the primal evaluation $X(p) = x$ corresponding to the sum of x Bernoulli successes and $n - x$ Bernoulli failures.

The stochastic derivative for a Binomial variable $\mathrm{Bin}(n, p)$ has weight w of order n and Y deterministically either 1 or -1 depending on the chosen direction of the derivative. Therefore, the derivative estimator $w(Y - X(p))$ has variance of order n. In contrast, the score method gives an estimator

$$\frac{X(p)(X(p) - np)}{p(1-p)} \tag{A.10}$$

which is also unbiased, but its variance

$$n^3 \frac{p}{1-p} + \mathcal{O}(n^2) \tag{A.11}$$

is growing with cubic rate in n.

Example A.2 (Differentiably parameterized continuous stochastic programs)**.** For continuous stochastic programs, our method reduces to the standard reparameterization trick. Specifically, for a differentiably parameterized continuous stochastic program $X(p)$, the stochastic derivative w.r.t a parameter p of the distribution has the form $(\delta, 0, 0)$. To satisfy Definition 2.2, a natural choice of bound is $B = |\delta| + 1$, so that $A_B^c(\varepsilon) = \{|\mathrm{d}X(\varepsilon)| \leq \varepsilon|\delta| + \varepsilon\}$ simply requires that the higher-order correction to the derivative linearization be smaller than ε when applied at a point at distance ε from p.

For instance, the stochastic derivative of $X(p) \sim \mathrm{Exp}(p)$ w.r.t. p, the scale of the Exponential distribution, is

$$\left(\frac{x}{p}, 0, 0\right) \tag{A.12}$$

conditionally on $X(p) = x$. Such rules are already implemented in the Julia AD ecosystem in `Distributions.jl` [1] and `DistributionsAD.jl` [2], and may be used without modification as the continuous case is a special case of our formalism.

Example A.3 (Categorical variable)**.** Consider a categorical variable $\mathrm{Categorical}_{\mathbf{a}}(p_1, p_2, p_3, \ldots, p_n)$ assuming fixed outputs $\mathbf{a} = a_1, a_2, \ldots, a_n$ with probabilities $p_1, p_2, p_3, \ldots, p_n > 0$. Assume that $p_1, p_2, p_3, \ldots, p_n$ implicitly depend on a single

parameter p, to fit our definition of stochastic derivative w.r.t. a single parameter, and let $X(p) \sim \text{Categorical}_{\mathbf{a}}(p_1, p_2, p_3, \ldots, p_n)$.

Suppose $X(p)$ is parameterized via the inversion method with outputs ordered as $a_1, a_2, \ldots a_n$. Similar logic to Example A.1 then yields that its stochastic derivative is of the form $(0, w, Y)$ with $Y \in \{a_{x-1}, a_{x+1}\}$, and

$$w \cdot \mathbb{P}(Y = a_{x+1}) = w_+ = \frac{|\sum_{i=1}^{x} \partial_p p_i|}{p_x} \cdot \mathbf{1}\left\{\varepsilon \cdot \sum_{i=1}^{x} \partial_p p_i < 0\right\}, \tag{A.13}$$

$$w \cdot \mathbb{P}(Y = a_{x-1}) = w_- = \frac{\left|\sum_{i=1}^{x-1} \partial_p p_i\right|}{p_x} \cdot \mathbf{1}\left\{\varepsilon \cdot \sum_{i=1}^{x-1} \partial_p p_i > 0\right\}. \tag{A.14}$$

In theory, $a_1, a_2, \ldots, a_n$ can be arbitrary discrete objects such as arrays or strings, as the categorical variable could represent an intermediate value of the program that is ultimately converted into a number. (Interpreted strictly, our formalism requires an embedding into Euclidean space; one can imagine a trivial embedding $a_i \mapsto i$ which is then provided to a learned embedding.) In handling such cases, one must be careful to speak in terms of *alternate* outputs rather than perturbations to the output. Our formalism of stochastic derivatives can already accommodate this idea, as Y takes on the alternate values rather than the values of the perturbations.

We now consider some general stochastic programs. Their stochastic derivatives arise automatically from composition and need not be hand-derived, but we do so to illustrate Theorem 2.6.

Example A.4 (Bernoulli plus Exponential)**.** Consider the program

$$X_1(p) \sim \text{Ber}(p), \tag{A.15}$$

$$X_2(p) \sim \text{Exp}(p), \tag{A.16}$$

$$X(p) = X_1(p) + X_2(p) \sim \text{Ber}(p) + \text{Exp}(p). \tag{A.17}$$

Then, $X(p)$ has right (left) stochastic derivative,

$$\left(\frac{x_2}{p}, w_1, Y_1\right) \tag{A.18}$$

conditionally on $X_1(p) = x_1$ and $X_2(p) = x_2$, where w_1 and Y_1 are given conditionally on the Bernoulli output $X_1(p) = x_1$ for $\varepsilon > 0$ ($\varepsilon < 0$) by Table 2.

Example A.5 (Cubing a geometric)**.** Consider the program

$$X_1(p) \sim \text{Geo}(p), \tag{A.19}$$

$$X(p) = X_1(p)^3. \tag{A.20}$$

We know from Example A.1 that $X_1(p)$ has right (left) stochastic derivative $(0, w_1, Y_1)$ with w_1 and Y_1 given by Table 2 for $\varepsilon > 0$ ($\varepsilon < 0$). In particular, $Y_1 \in \{x_1 - 1, x_1 + 1\}$ conditionally on $X_1(p) = x_1$. Thus, intuitively, when propagating through the cube function we only care about the discretely spaced alternate values $(x_1 - 1)^3$ and $(x_1 + 1)^3$; the conventional derivative $3x_1^2$ is irrelevant as the input to the cube function is integer-valued. Indeed, by Theorem 2.6 the stochastic derivative of $X(p)$ reads

$$(0, w_1, Y_1^3), \tag{A.21}$$

where $Y_1^3 \in \{(x_1 - 1)^3, (x_1 + 1)^3\}$.

Example A.6 (Parameter-scaled Bernoulli)**.** Consider the program,

$$X_1(p) \sim \text{Ber}(p), \tag{A.22}$$

$$X(p) = p \cdot X_1(p). \tag{A.23}$$

By Theorem 2.6, the stochastic derivative of $X(p)$ reads,

$$(X_1(p), w_1, pY_1), \tag{A.24}$$

where w_1 and Y_1 are given for $X_1(p)$ by Table 2.

Example A.7 (Two-step random walk)**.** Consider the program,

$$X_1(p) \sim \mathrm{Ber}(p), \tag{A.25}$$

$$X_2(x_1) \sim \begin{cases} \mathrm{Ber}(p) & \text{if } x_1 = 0 \\ \mathrm{Ber}(2p) & \text{if } x_1 = 1. \end{cases} \tag{A.26}$$

$$X(p) = X_1(p) + (X_2 \circ X_1)(p). \tag{A.27}$$

This represents a two-step random walk, where there is a transition $0 \to 1$ with probability p, a transition $1 \to 2$ with probability $2p$, and self loops otherwise. In this example, let us focus on a particular primal evaluation where $X_1(p) = 0$ and $X_2(X_1(p)) = 0$, so that $X(p) = 0$, and consider only the right stochastic derivative for simplicity. Conditionally on $X_1(p) = 0$, $X_1(p)$ has a right stochastic derivative (δ_1, w_1, Y_1) given by

$$\left(0, \frac{1}{1-p}, 1\right), \tag{A.28}$$

using Table 2. Similarly, conditionally on $X_2(X_1(p)) = X_2(0) = 0$, X_2 has right stochastic derivative (δ_2, w_2, Y_2) given by

$$\left(0, \frac{1}{1-p}, 1\right). \tag{A.29}$$

Let us now turn our attention to the stacked program $[X_1(p); (X_2 \circ X_1)(p)]$. By Theorem 2.6, the stacked program has right stochastic derivative $(0, w, Y_{12})$ where conditionally on $X_1(p) = 0$ and $X_2(0) = 0$,

$$Y_{12} \quad = \quad \begin{cases} [1; X_2(1)] & \text{with probability} \quad 1/2, \\ [0; 1] & \text{with probability} \quad 1/2, \end{cases} \tag{A.30}$$

and $w = \frac{2}{1-p}$. Note that conditionally on $X_2(0) = 0$, $X_2(1)$ has a chance $\frac{1-2p}{1-p}$ of also being 0, but a $\frac{p}{1-p}$ chance of flipping to 1, so that the first case expands into two events

$$Y_{12} \quad = \quad \begin{cases} [1; 1] & \text{with probability} \quad \frac{p}{2(1-p)}, \\ [1; 0] & \text{with probability} \quad \frac{1-2p}{2(1-p)}, \\ [0; 1] & \text{with probability} \quad 1/2. \end{cases} \tag{A.31}$$

Finally, the expression for $X(p)$ may be thought of as a unary sum function operating on this stacked program, with stochastic derivative $(0, w, Y)$ where conditionally on $X_1(p) = 0$ and $X_2(0) = 0$

$$Y \quad = \quad \begin{cases} 2 & \text{with probability} \quad \frac{p}{2(1-p)}, \\ 1 & \text{with probability} \quad \frac{1-2p}{2(1-p)} + 1/2. \end{cases} \tag{A.32}$$

Finally, we present an example using smoothed stochastic derivatives to rederive a popular gradient estimator.

Example A.8 (Recovering the straight-through gradient estimator of [6])**.** The straight-through gradient estimator formally assigns a derivative of 1 to the hard-thresholding function $\mathrm{HT} = \mathbf{1}_{[0,\infty)}$, such that a Bernoulli variable $\mathrm{Ber}(p) \sim \mathrm{HT}(p - U[0,1])$ is assigned a derivative of 1. Now, let $\widetilde{\delta}_L$ and $\widetilde{\delta}_R$ be the left and right smoothed stochastic derivatives of $\mathrm{Ber}(p)$. By Table 2 and Definition 2.7, for the right-sided ($\varepsilon > 0$) case,

$$\widetilde{\delta_R} = 1/(1-p) \cdot \mathbf{1}\{x = 0\}, \tag{A.33}$$

and for the left-sided ($\varepsilon < 0$) case,

$$\widetilde{\delta_L} = 1/p \cdot \mathbf{1}\{x = 1\}. \tag{A.34}$$

By linearity of expectation applied to Definition 2.7, any affine combination of $\widetilde{\delta_L}$ and $\widetilde{\delta_R}$ is also a valid smoothed stochastic derivative. In particular,

$$1 = (1-p) \cdot \widetilde{\delta}_L + p \cdot \widetilde{\delta}_R, \tag{A.35}$$

is a valid smoothed stochastic derivative of a Bernoulli variable, which explains why the straight-through estimator provides a low-bias estimate. Note that a constant-valued affine combination of the left and right smoothed stochastic derivatives is not possible in general, e.g. it is not possible for Geometric and Poisson random variables, in which cases smoothed stochastic derivatives generalize the straight-through estimator (mixing $\widetilde{\delta}_R$ and $\widetilde{\delta}_L$ may still be useful in these cases to reduce bias.)

B Proofs

B.1 Preliminaries

Recall from the main text,

Definition 2.1. A *stochastic program* $X(p)$ is a stochastic process with values in a Euclidean space E, whose index set I is either an open subset of a Euclidean space or a closed real interval.

Throughout the formalism, we let $X(p)$ denote such a stochastic program, where $I = [a, b] \subset \mathbb{R}$ is a closed interval. As noted in the main text, it is sufficient to consider this case because sensitivities of stochastic programs Z with more general index sets can be understood for an input $\mathbf{u}$ by studying at $p = 0$ the directional perturbation $X(p) = Z(\mathbf{u} + p\mathbf{v})$ in a direction $\mathbf{v}$, where $X(p)$ is then a stochastic program with index set a closed interval containing 0.

As in the main text, we use the shorthand

$$\mathrm{d}X(\varepsilon) = X(p + \varepsilon) - X(p). \tag{B.1}$$

A number of statements and propositions have identical forms for right and left stochastic derivatives, only differing in the direction of the limits. To accomodate this, we often use the notation $\varepsilon \to 0^{+/-}$ to indicate that statements and proofs for right and left stochastic derivatives can both be read off by choosing the appropriate side of the limit for all expressions, where objects such as w have different definitions depending on the chosen reading $\varepsilon \to 0^+$ or $\varepsilon \to 0^-$.

For clarity, we remark upon the (standard) notation,

$$W = \mathbb{E}\left[U \mid V\right], \tag{B.2}$$

where U and V are both random variables defined on a sample space Ω. In this setting, W is itself also a random variable defined on Ω. In particular, for a particular sample $\omega \in \Omega$, $W(\omega)$ is the conditional expectation $\mathbb{E}\left[U \mid V = V(\omega)\right]$. In the language of σ-algebras, $\mathbb{E}\left[U \mid V\right]$ is equivalent to $\mathbb{E}\left[U \mid \sigma(V)\right]$, where $\sigma(V)$ is the sub-σ-algebra generated by V.

B.2 Unbiasedness of stochastic derivatives

We first prove a general result regarding when a class of events $A(\varepsilon)$ allows for the pathwise gradient estimator δ to be applied when the probability space is restricted to $A^c(\varepsilon)$.

Proposition B.1. *Suppose $\mathrm{d}X(\varepsilon)/\varepsilon \to \delta$ almost surely, and $|\mathrm{d}X(\varepsilon)| \leq B|\varepsilon|$ holds given the event $A^c(\varepsilon)$, where $B > |\delta|$ is integrable. Then, for any function $f : E \to \mathbb{R}$ with bounded derivative,*

$$\lim_{\varepsilon \to 0^{+/-}} \mathbb{E}\left[\frac{f(X(p+\varepsilon)) - f(X(p))}{\varepsilon}\mathbf{1}_{A^c(\varepsilon)} \,\middle|\, X(p)\right] = \mathbb{E}\left[f'(X(p))\delta \mid X(p)\right]. \tag{B.3}$$

Proof. Since $\mathrm{d}X(\varepsilon)/\varepsilon \to \delta$ almost surely and $B > |\delta|$, $|\mathrm{d}X(\varepsilon)| \leq B|\varepsilon|$ holds almost surely as $\varepsilon \to 0$. So $\mathbf{1}_{A^c(\varepsilon)} \to 1$ almost surely as $\varepsilon \to 0$. By the chain rule,

$$\frac{f(X(p+\varepsilon)) - f(X(p))}{\varepsilon}\mathbf{1}_{A^c(\varepsilon)} \to f'(X(p))\delta. \tag{B.4}$$

almost surely as $\varepsilon \to 0^{+/-}$. Furthermore,

$$\left|\frac{f(X(p+\varepsilon)) - f(X(p))}{\varepsilon}\mathbf{1}_{A^c(\varepsilon)}\right| \leq \frac{1}{\varepsilon}\|f'\|_\infty |\mathrm{d}X(\varepsilon)|\mathbf{1}_{A^c(\varepsilon)} \tag{B.5}$$

$$\leq B\|f'\|_\infty =: G. \tag{B.6}$$

By assumption $\mathbb{E}\,G < \infty$. The proposition follows by the dominated convergence theorem for conditional expectations. □

As in the main text, define

$$A_B(\varepsilon) = \{|\mathrm{d}X(\varepsilon)| > B|\varepsilon|\} \tag{B.7}$$

and recall

Definition 2.2 (Stochastic derivative)**.** Suppose $X(p) \in E$ is a stochastic program with index set I a closed real interval. We say that the triple of random variables (δ, w, Y), with $w \in \mathbb{R}$ and $Y \in E$, is a right (left) *stochastic derivative* of X at the input $p \in I$ if $\mathrm{d}X(\varepsilon)/\varepsilon \to \delta$ almost surely as $\varepsilon \to 0$, and there is an integrable random variable $B > |\delta|$ such that for all bounded functions $f\colon E \to \mathbb{R}$ with bounded derivative it holds almost surely that

$$\mathbb{E}\left[w\left(f(Y) - f(X(p))\right) \mid X(p)\right] = \lim_{\varepsilon \to 0^{+/-}} \mathbb{E}\left[\frac{f(X(p+\varepsilon)) - f(X(p))}{\varepsilon} \mathbf{1}_{A_B(\varepsilon)} \;\middle|\; X(p)\right], \tag{2.4}$$

with limit taken from above (below), where $\mathbb{P}\left(A_B(\varepsilon) \mid X(p)\right)/\varepsilon$ is dominated by an integrable random variable for all $\varepsilon > 0$ ($\varepsilon < 0$).

Using Proposition B.1, we may show that outside the event $A_B(\varepsilon)$, the sensitivity is well-described by the pathwise gradient estimator, allowing us to prove Proposition 2.3.

Proposition 2.3 (Unbiasedness)**.** If (δ, w, Y) is a stochastic derivative of $X(p)$ at p, it holds that

$$\frac{\mathrm{d}\mathbb{E}\left[X(p)\right]}{\mathrm{d}p} = \mathbb{E}[\delta + w(Y - X(p))]. \tag{2.5}$$

Proof. Let B be the associated bound. Since $\mathrm{d}X(\varepsilon)/\varepsilon \to \delta$ almost surely as $\varepsilon \to 0$ and $A_B^{\mathrm{c}}(\varepsilon)$ implies $\mathrm{d}X(\varepsilon)/\varepsilon \leq B\varepsilon$, Proposition B.1 applied to $A_B(\varepsilon)$ with f as identity implies

$$\lim_{\varepsilon \to 0^{+/-}} \mathbb{E}\left[\frac{X(p+\varepsilon) - X(p)}{\varepsilon} \mathbf{1}_{A_B^{\mathrm{c}}(\varepsilon)} \;\middle|\; X(p)\right] = \mathbb{E}\left[\delta \mid X(p)\right], \tag{B.8}$$

while Eq. (2.4) of Definition 2.2 also applied with f as identity gives

$$\lim_{\varepsilon \to 0^{+/-}} \mathbb{E}\left[\frac{X(p+\varepsilon) - X(p)}{\varepsilon} \mathbf{1}_{A_B(\varepsilon)} \;\middle|\; X(p)\right] = \mathbb{E}\left[w(Y - X(p)) \mid X(p)\right]. \tag{B.9}$$

Summing Eq. (B.8) and Eq. (B.9), we have

$$\lim_{\varepsilon \to 0^{+/-}} \mathbb{E}\left[\frac{X(p+\varepsilon) - X(p)}{\varepsilon} \;\middle|\; X(p)\right] = \mathbb{E}\left[\delta + w(Y - X(p)) \mid X(p)\right], \tag{B.10}$$

and applying the tower property of conditional expectations, we obtain

$$\frac{\mathrm{d}\mathbb{E}[X(p)]}{\mathrm{d}p} = \lim_{\varepsilon \to 0^{+/-}} \mathbb{E}\left[\frac{X(p+\varepsilon) - X(p)}{\varepsilon}\right] = \mathbb{E}[\delta + w(Y - X(p))] = \mathbb{E}[\widetilde{X}(p)], \tag{B.11}$$

as desired. □

B.3 Construction of stochastic derivatives for elementary programs

We now give a technical condition under which a stochastic derivative can be constructed for an elementary stochastic program, in which we characterize w as the derivative of the conditional probability of $A_B(\varepsilon)$ given $X(p)$, and Y as realization of the weak limit of the conditional distribution of $X(p+\varepsilon)$ given $X(p)$ and $A_B(\varepsilon)$.

Assumption B.2. We assume that $X(p)$ is almost surely differentiable, so that $\mathrm{d}X(\varepsilon)/\varepsilon \to \delta$ almost surely as $\varepsilon \to 0$ for some δ, and that we may find an integrable random bound $B > |\delta|$ such that

- for the quantity
$$w(\varepsilon) = \mathbb{P}\left(A_B(\varepsilon) \mid X(p)\right), \tag{B.12}$$
$w(\varepsilon)/\varepsilon$ is dominated in ε by an integrable random variable and converges almost surely to a random variable w as $\varepsilon \to 0^{+/-}$.

- the conditional distribution of $X(p+\varepsilon)$ given $X(p)$ and $A_B(\varepsilon)$ converges in distribution to the distribution of a random variable Y as $\varepsilon \to 0^{+/-}$. Specifically, Y must satisfy

$$\mathbb{E}\left[w \cdot f(Y) \mid X(p)\right] = \lim_{\varepsilon \to 0^{+/-}} \mathbb{E}\left[w \cdot f(X(p+\varepsilon)) \mid X(p), A_B(\varepsilon)\right] \tag{B.13}$$

  for all bounded continuous functions $f\colon E \to \mathbb{R}$.

Given Assumption B.2, we may construct a stochastic derivative for $X(p)$.

Theorem 2.4. Given a stochastic program $X(p)$ satisfying Assumption B.2, there exists a right stochastic derivative (δ, w_R, Y_R) with $w_R \geq 0$ at any $p \in [a, b)$ and a left stochastic derivative (δ, w_L, Y_L) with $w_L \leq 0$ at any $p \in (a, b]$.

Proof. Construct (δ, w, Y) via Assumption B.2 (taking the limit from above for the case of right stochastic derivatives, and the limit from below for the case of left stochastic derivatives). For bounded f with bounded derivative, write

$$\mathbb{E}\left[\frac{f(X(p+\varepsilon)) - f(X(p))}{\varepsilon}\mathbf{1}_{A_B(\varepsilon)} \;\middle|\; X(p)\right] \tag{B.14}$$

$$= \mathbb{E}\left[w\left(f(X(p+\varepsilon)) - f(X(p))\right)\frac{\mathbf{1}_{A_B(\varepsilon)}}{w(\varepsilon)} \;\middle|\; X(p)\right]$$

$$+ \mathbb{E}\left[\left(\frac{w(\varepsilon)}{\varepsilon} - w\right)\left(f(X(p+\varepsilon)) - f(X(p))\right)\frac{\mathbf{1}_{A_B(\varepsilon)}}{w(\varepsilon)} \;\middle|\; X(p)\right]. \tag{B.15}$$

Note that $\mathbb{E}[\mathbf{1}_{A_B(\varepsilon)} \mid X(p)] = \mathbb{P}(A_B(\varepsilon) \mid X(p)) = w(\varepsilon)$. Thus, as $\varepsilon \to 0^{+/-}$ the first term approaches

$$\mathbb{E}\left[w\left(f(X(p+\varepsilon)) - f(X(p))\right)\frac{\mathbf{1}_{A_B(\varepsilon)}}{w(\varepsilon)} \;\middle|\; X(p)\right] \tag{B.16}$$

$$= \mathbb{E}\left[w\left(f\left(X(p+\varepsilon)\right) - f(X(p))\right) \mid X(p), A_B(\varepsilon)\right] \tag{B.17}$$

$$\to \mathbb{E}\left[w\left(f(Y) - f(X(p))\right) \mid X(p)\right] \tag{B.18}$$

by Assumption B.2. Now, using the $X(p)$-measurability of $w(\varepsilon)$ and w, we may bound the magnitude of the second term as

$$\left|\mathbb{E}\left[\left(\frac{w(\varepsilon)}{\varepsilon} - w\right)\left(f(X(p+\varepsilon)) - f(X(p))\right)\frac{\mathbf{1}_{A_B(\varepsilon)}}{w(\varepsilon)} \;\middle|\; X(p)\right]\right| \tag{B.19}$$

$$\leq 2\|f\|_\infty \cdot \left|\frac{w(\varepsilon)}{\varepsilon} - w\right| \cdot \mathbb{E}\left[\frac{\mathbf{1}_{A_B(\varepsilon)}}{w(\varepsilon)} \;\middle|\; X(p)\right] \tag{B.20}$$

$$= 2\|f\|_\infty \cdot \left|\frac{w(\varepsilon)}{\varepsilon} - w\right|, \tag{B.21}$$

which approaches 0 almost surely as $\varepsilon \to 0^{+/-}$ by the almost sure convergence of $w(\varepsilon)/\varepsilon$ to w. Thus,

$$\mathbb{E}\left[w\left(f(Y) - f(X(p)\right) \mid X(p)\right] = \lim_{\varepsilon \to 0^{+/-}} \mathbb{E}\left[\frac{f(X(p+\varepsilon)) - f(X(p))}{\varepsilon}\mathbf{1}_{A_B(\varepsilon)} \;\middle|\; X(p)\right]. \tag{B.22}$$

Additionally, $\mathbb{P}\left(A_B(\varepsilon) \mid X(p)\right)/\varepsilon$ is dominated in ε by an integrable random variable and $\mathrm{d}X(\varepsilon)/\varepsilon \to \delta$ almost surely by assumption. So (δ, w, Y) is a stochastic derivative of X at p. Finally, note that $w(\varepsilon)/\varepsilon \geq 0$ for $\varepsilon > 0$ while $w(\varepsilon)/\varepsilon \leq 0$ for $\varepsilon < 0$, by the non-negativity of $w(\varepsilon)$. So applying Assumption B.2 with limits from above for $p \in [a, b)$ indeed produces a right stochastic derivative, while applying Assumption B.2 with limits from below for $p \in (a, b]$ produces a left stochastic derivative. □

B.4 Composition of stochastic derivatives

We now provide a proof of the chain rule for stochastic derivatives.

Theorem 2.6 (Chain rule)**.** Consider independent stochastic programs X_1 and X_2 and their composition $X_2 \circ X_1$. Suppose that X_1 has a right (left) stochastic derivative at $p \in \mathbb{R}$ given by (δ_1, w_1, Y_1) with bound B_1, and X_2 has a right stochastic derivative (δ_2, w_2, Y_2) in the direction[1] $\widehat{\delta}_1 = \delta_1/|\delta_1|$ with bound B_2 given conditionally on its input $X_1(p)$, where with $\mathrm{d}X_2(\mathbf{v}) = X_2(X_1(p) + \mathbf{v}) - X_2(X_1(p))$ the event

$$\widetilde{A}_2(\varepsilon) = \{|\mathrm{d}X_2(\mathbf{v})| > B_2|\mathbf{v}| \text{ for some } \mathbf{v} \text{ satisfying } |\mathbf{v}| \leq \varepsilon\} \tag{B.23}$$

has $\mathbb{P}\left(\widetilde{A}_2(\varepsilon) \;\middle|\; X_1(p), X_2(p)\right)/\varepsilon$ dominated in ε by an integrable random variable W, and $\mathrm{d}X_2(\mathbf{v})$ is almost surely differentiable with respect to $\mathbf{v}$. Then, if $B_1 B_2$ and $|\delta_1|W$ are integrable, the stacked program $[X_1; X_2 \circ X_1]$ has a right (left) stochastic derivative at p given by (δ, w, Y) where $\delta = [\delta_1; |\delta_1|\delta_2]$,

$$Y \quad = \quad \begin{cases} [Y_1; X_2(Y_1)] & \text{with probability} \quad \dfrac{w_1}{w_1 + |\delta_1| w_2}, \\ [X_1(p); Y_2] & \text{with probability} \quad \dfrac{|\delta_1| w_2}{w_1 + |\delta_1| w_2}, \end{cases} \tag{2.7}$$

and $w = w_1 + |\delta_1| w_2$, with associated bound $B = B_1 + B_1 B_2$.

Proof. We prove the case of right stochastic derivatives, as the case of left stochastic derivatives is analogous. Let $X = X_2 \circ X_1$ as a shorthand. Now, for $\varepsilon > 0$ consider the events

$$A_1(\varepsilon) = \{|\mathrm{d}X_1(\varepsilon)| > B_1 \varepsilon\}, \tag{B.24}$$

$$A_2(\varepsilon) = \{|\mathrm{d}X_2(\varepsilon\delta_1)| > |\delta_1| B_2 \varepsilon\}, \tag{B.25}$$

$$A(\varepsilon) = \{|\mathrm{d}X(\varepsilon)| > B_1 B_2 \varepsilon\}. \tag{B.26}$$

Intuitively, the events $A_1(\varepsilon)$, $A_2(\varepsilon)$, and $A(\varepsilon)$ represent large jumps in X_1, X_2, and X respectively. Throughout the proof, we let f denote a bounded function with bounded derivative.

Step 1: The composed program is almost surely differentiable.

Since $X_1(p)$ and $\mathrm{d}X_2(\mathbf{v})$ are almost surely differentiable, the chain rule implies that $X(p) = X_2(X_1(p))$ is almost surely differentiable, with derivative $|\delta_1|\delta_2$, i.e. $\mathrm{d}X(\varepsilon)/\varepsilon \to |\delta_1|\delta_2$ almost surely as $\varepsilon \to 0$. Additionally, the directional perturbation $\varepsilon \mapsto X_2(X_1(p) + \varepsilon\delta_1)$ is also almost surely differentiable by the almost-sure differentiability of $X_1(p)$ and the definition of δ_2, with the same derivative $|\delta_1|\delta_2$.

Step 2: No large jump in X_1 or X_2 implies no large jump in X.

Given $A_1^c(\varepsilon)$ and $\widetilde{A}_2^c(\varepsilon|\delta_1|)$, we have

$$|\mathrm{d}X(\varepsilon)| = |\mathrm{d}X_2(\mathrm{d}X_1(\varepsilon))| \tag{B.27}$$

$$\leq B_1 B_2 \varepsilon, \tag{B.28}$$

which implies $A^c(\varepsilon)$. We now remark that $\widetilde{A}_2(\varepsilon|\delta_1|)$ and $A_2(\varepsilon)$ may be interchanged when considering the sensitivity of the directional perturbation $\varepsilon \mapsto X_2(X_1(p) + \varepsilon\delta_1)$. To see this, note that

$$\widetilde{A}_2^c(\varepsilon|\delta_1|) \subseteq A_2^c(\varepsilon) = \{|X_2(X_1(p) + \varepsilon\delta_1) - X(p)| \leq |\delta_1| B_2 \varepsilon\}. \tag{B.29}$$

Thus, since $X_2(X_1(p)+\varepsilon\delta_1)$ is almost surely differentiable, we may apply the dominated convergence argument of Proposition B.1 to both $A_2(\varepsilon)$ and $\widetilde{A}_2(\varepsilon|\delta_1|)$, obtaining

$$\lim_{\varepsilon \to 0^+} \mathbb{E}\left[\frac{f(X_2(X_1(p) + \varepsilon\delta_1)) - f(X(p))}{\varepsilon} \mathbf{1}_{A_2^c(\varepsilon)} \;\middle|\; X_1(p), X(p)\right] \tag{B.30}$$

$$= \lim_{\varepsilon \to 0^+} \mathbb{E}\left[f'(X_2(X_1(p)))\delta_2|\delta_1| \mid X_1(p), X(p)\right] \tag{B.31}$$

$$= \lim_{\varepsilon \to 0^+} \mathbb{E}\left[\frac{f(X_2(X_1(p) + \varepsilon\delta_1))) - f(X(p))}{\varepsilon} \mathbf{1}_{\widetilde{A}_2^c(\varepsilon|\delta_1|)} \;\middle|\; X_1(p), X(p)\right]. \tag{B.32}$$

[1] when $\delta_1 = 0$, we may choose $\widehat{\delta}_1$ arbitrarily, as δ_2, w_2 and Y_2 ultimately have no contribution to the stochastic derivative in this case.

By a similar argument, it holds that $A_1(\varepsilon) \cup \widetilde{A}_2(\varepsilon|\delta_1|)$ may be interchanged with $A(\varepsilon)$ when considering the sensitivity of $X_2 \circ X_1$. Specifically, since $X_2(X_1(p))$ is almost surely differentiable and both $A_1^c(\varepsilon) \cap \widetilde{A}_2^c(\varepsilon|\delta_1|)$ and $A_2^c(\varepsilon)$ bound $|\mathrm{d}X(\varepsilon)|$ by $B_1 B_2 \varepsilon$, we may apply Proposition B.1 for $X = X_2 \circ X_1$ to both event classes, obtaining

$$\lim_{\varepsilon \to 0^+} \mathbb{E}\left[\frac{f(X(p+\varepsilon)) - f(X(p))}{\varepsilon} \mathbf{1}_{A_1^c(\varepsilon) \cap \widetilde{A}_2^c(\varepsilon|\delta_1|)} \;\middle|\; X_1(p), X(p) \right] \tag{B.33}$$

$$= \lim_{\varepsilon \to 0^+} \mathbb{E}\left[f'(X_2(X_1(p)))\delta_2 |\delta_1| \mid X_1(p), X(p) \right] \tag{B.34}$$

$$= \lim_{\varepsilon \to 0^+} \mathbb{E}\left[\frac{f(X(p+\varepsilon)) - f(X(p))}{\varepsilon} \mathbf{1}_{A_2^c(\varepsilon)} \;\middle|\; X_1(p), X(p) \right]. \tag{B.35}$$

Step 3: A large jump in both X_1 and X_2 has negligible probability.

Note that $A_1(\varepsilon)$ and $\widetilde{A}_2(\varepsilon|\delta_1|)$ are independent conditional on $X_1(p)$. Therefore, in the limit, we have

$$\frac{\mathbb{P}\left(A_1(\varepsilon) \cap \widetilde{A}_2(\varepsilon|\delta_1|) \;\middle|\; X_1(p), X(p) \right)}{\varepsilon} \tag{B.36}$$

$$= \frac{\mathbb{P}\left(A_1(\varepsilon) \mid X_1(p) \right)}{\varepsilon} \mathbb{P}\left(\widetilde{A}_2(\varepsilon|\delta_1|) \;\middle|\; X_1(p), X(p) \right) \tag{B.37}$$

$$\to 0 \tag{B.38}$$

almost surely as $\varepsilon \to 0^+$ by the domination of $\mathbb{P}\left(A_1(\varepsilon) \mid X_1(p) \right) / \varepsilon$ by an integrable random variable and that $\mathbb{P}\left(\widetilde{A}_2(\varepsilon|\delta_1|) \;\middle|\; X_1(p), X(p) \right) \to 0$. By boundedness of f this implies

$$\lim_{\varepsilon \to 0^+} \mathbb{E}\left[\frac{f(X(p+\varepsilon)) - f(X(p))}{\varepsilon} \mathbf{1}_{A_1(\varepsilon) \cap \widetilde{A}_2(\varepsilon|\delta_1|)} \;\middle|\; X_1(p), X(p) \right] = 0, \tag{B.39}$$

almost surely, so the sensitivity of X is negligible under $A_1(\varepsilon) \cap \widetilde{A}_2(\varepsilon|\delta_1|)$.

Step 4: Sensitivity of $X(p)$ given a large jump in X_1 is described by w_1, Y_1.

The random variable

$$Z(y) = \mathbb{E}[f(X_2(y)) \mid X_1(p), X(p)] \tag{B.40}$$

is $X_1(p)$, $X(p)$ measurable and almost surely continuous, and the function f in the characterization Eq. (2.4) can be taken as bounded and almost surely continuous in $X_1(p)$ by an extension of the Portmanteau theorem [47]. It follows, as $X(p)$ and $X_1(p+\varepsilon)$ are independent given $X_1(p)$, that

$$\lim_{\varepsilon \to 0^+} \mathbb{E}\left[\frac{Z(X_1(p+\varepsilon)) - Z(X_1(p))}{\varepsilon} \mathbf{1}_{A_1(\varepsilon)} \;\middle|\; X_1(p), X(p) \right] \tag{B.41}$$

$$= \mathbb{E}\left[w_1 \left(Z(Y_1) - Z(X_1(p)) \right) \mid X_1(p), X_2(p) \right], \tag{B.42}$$

so that

$$\lim_{\varepsilon \to 0^+} \mathbb{E}\left[\frac{f(X(p+\varepsilon)) - f(X(p))}{\varepsilon} \mathbf{1}_{A_1(\varepsilon)} \;\middle|\; X_1(p), X(p) \right] \tag{B.43}$$

$$= \mathbb{E}\left[w_1 \left(f(X_2(Y_1)) - f(X(p)) \right) \mid X_1(p), X(p) \right]. \tag{B.44}$$

Step 5: Sensitivity of $X(p)$ given a large jump in X_2 is described by w_2, Y_2.

Using Eq. (B.28), given $\widetilde{A}_2^c(\varepsilon|\delta_1|) \cap A_1^c(\varepsilon)$ it holds that

$$\left| \frac{f\left(X_2(X_1(p) + \mathrm{d}X_1(\varepsilon)) \right) - f(X_2(p))}{\varepsilon} \right| \leq \|f'\|_\infty \frac{|\mathrm{d}X(\varepsilon)|}{\varepsilon} \leq \|f'\|_\infty B_1 B_2 \tag{B.45}$$

is dominated by an integrable random variable. Therefore, since $\mathrm{d}X_1(\varepsilon)/\varepsilon \to \delta_1$ almost surely,

$$\lim_{\varepsilon\to 0^+} \mathbb{E}\left[\frac{f(X_2(X_1(p)+\mathrm{d}X(\varepsilon))) - f(X(p))}{\varepsilon}\mathbf{1}_{\widetilde{A}_2^c(\varepsilon|\delta_1|)\cap A_1^c(\varepsilon)} \;\middle|\; X_1(p), X(p)\right] \tag{B.46}$$

$$= \lim_{\varepsilon\to 0^+} \mathbb{E}\left[\frac{f(X_2(X_1(p)+\varepsilon\delta_1)) - f(X(p))}{\varepsilon}\mathbf{1}_{\widetilde{A}_2^c(\varepsilon|\delta_1|)\cap A_1^c(\varepsilon)} \;\middle|\; X_1(p), X(p)\right] \tag{B.47}$$

by dominated convergence. Putting this together with Eq. (B.32), by definition of w_2 and Y_2,

$$\mathbb{E}\left[|\delta_1| w_2\left(f(Y_2) - f\big(X(p)\big)\right) \mid X_1(p), X(p)\right] \tag{B.48}$$

$$= \lim_{\varepsilon\to 0^+} \mathbb{E}\left[\frac{f(X_2(X_1(p)+\varepsilon\delta_1)) - f(X(p))}{\varepsilon}\mathbf{1}_{A_2(\varepsilon)} \;\middle|\; X_1(p), X(p)\right] \tag{B.49}$$

$$= \lim_{\varepsilon\to 0^+} \mathbb{E}\left[\frac{f(X_2(X_1(p)+\varepsilon\delta_1))) - f(X(p))}{\varepsilon}\mathbf{1}_{\widetilde{A}_2(\varepsilon|\delta_1|)} \;\middle|\; X_1(p), X(p)\right] \tag{B.50}$$

$$= \lim_{\varepsilon\to 0^+} \mathbb{E}\left[\frac{f(X_2(X_1(p)+\varepsilon\delta_1))) - f(X(p))}{\varepsilon}\mathbf{1}_{\widetilde{A}_2(\varepsilon|\delta_1|)\cap A_1^c(\varepsilon)} \;\middle|\; X_1(p), X(p)\right] \tag{B.51}$$

$$= \lim_{\varepsilon\to 0^+} \mathbb{E}\left[\frac{f(X_2(X_1(p)+\mathrm{d}X_1(\varepsilon))) - f(X(p))}{\varepsilon}\mathbf{1}_{\widetilde{A}_2(\varepsilon|\delta_1|)\cap A_1^c(\varepsilon)} \;\middle|\; X_1(p), X(p)\right] \tag{B.52}$$

$$= \lim_{\varepsilon\to 0^+} \mathbb{E}\left[\frac{f(X(p+\varepsilon)) - f(X(p))}{\varepsilon}\mathbf{1}_{\widetilde{A}_2(\varepsilon|\delta_1|)} \;\middle|\; X_1(p), X(p)\right], \tag{B.53}$$

where we freely neglect (or add back in) $\mathbf{1}_{A_1(\varepsilon)\cap\widetilde{A}_2(\varepsilon|\delta_1|)}$ in the conditional expectation by Eq. (B.39).

Step 6: (δ, w, Y) is a stochastic derivative of the stacked program $[X_1(p); X(p)]$.

Since δ_1 is an almost-sure derivative of $X_1(p)$ and $|\delta_1|\delta_2$ is an almost-sure derivative of $X(p)$, $\delta = [\delta_1; |\delta_1|\delta_2]$ is an almost sure derivative of $[X_1(p); X(p)]$.

Given $A(\varepsilon)$, $|\mathrm{d}X(\varepsilon)| > B_1B_2\varepsilon$, and given $A_1(\varepsilon)$, $|\mathrm{d}X_1(\varepsilon)| > B_1\varepsilon$. Thus we may choose the bound $B = B_1B_2 + B_1$ for the stacked program, so that $A_B(\varepsilon) \subseteq A_1(\varepsilon) \cup A(\varepsilon)$. Now, since $A(\varepsilon) \subseteq A_1(\varepsilon) \cup \widetilde{A}_2(\varepsilon|\delta_1|)$,

$$\frac{\mathbb{P}\left(A(\varepsilon) \mid X_1(p), X(p)\right)}{\varepsilon} \leq \frac{\mathbb{P}\left(A_1(\varepsilon) \mid X_1(p), X(p)\right)}{\varepsilon} + \frac{\mathbb{P}\left(\widetilde{A}_2(\varepsilon|\delta_1|) \;\middle|\; X_1(p), X(p)\right)}{\varepsilon} \tag{B.54}$$

$$\leq \frac{\mathbb{P}\left(A_1(\varepsilon) \mid X_1(p), X(p)\right)}{\varepsilon} + |\delta_1|W, \tag{B.55}$$

where $\mathbb{P}\left(A_1(\varepsilon) \mid X_1(p), X(p)\right)/\varepsilon$ and $|\delta_1|W$ are integrable by assumption. Therefore, $\mathbb{P}\left(A_B(\varepsilon) \mid X_1(p), X(p)\right)$ is also dominated by an integrable random variable, as desired.

Now, let f operate on the stacked program, and for convenience write $f(x_1; x_2) = f([x_1; x_2])$. As a shorthand, let $\overline{X}(p) = [X_1(p); X(p)]$. With w and Y as given in the statement,

$$\mathbb{E}\left[w\big(f(Y) - f(X_1(p); X(p))\big) \mid X_1(p), X(p)\right] \tag{B.56}$$

$$= \mathbb{E}\left[w_1\big(f\left(X_1(p); X_2(Y_1)\right) - f(X_1(p); X(p))\big) \mid X_1(p), X(p)\right]$$
$$+ \mathbb{E}\left[|\delta_1| w_2\big(f(Y_1; X_2(Y_1)) - f(X_1(p); X(p))\big) \mid X_1(p), X(p)\right] \tag{B.57}$$

$$= \lim_{\varepsilon\to 0^+} \mathbb{E}\left[\frac{f(\overline{X}(p+\varepsilon)) - f(\overline{X}(p))}{\varepsilon}\left(\mathbf{1}_{A_1(\varepsilon)} + \mathbf{1}_{\widetilde{A}_2(\varepsilon|\delta_1|)}\right) \;\middle|\; X_1(p), X(p)\right] \tag{B.58}$$

$$= \lim_{\varepsilon\to 0^+} \mathbb{E}\left[\frac{f(\overline{X}(p+\varepsilon)) - f(\overline{X}(p))}{\varepsilon}\mathbf{1}_{A_1(\varepsilon)\cup\widetilde{A}_2(\varepsilon|\delta_1|)} \;\middle|\; X_1(p), X(p)\right] \tag{B.59}$$

$$= \lim_{\varepsilon\to 0^+} \mathbb{E}\left[\frac{f(\overline{X}(p+\varepsilon)) - f(\overline{X}(p))}{\varepsilon}\mathbf{1}_{A_1(\varepsilon)\cup A(\varepsilon)} \;\middle|\; X_1(p), X(p)\right] \tag{B.60}$$

$$= \lim_{\varepsilon\to 0^+} \mathbb{E}\left[\frac{f(\overline{X}(p+\varepsilon)) - f(\overline{X}(p))}{\varepsilon}\mathbf{1}_{A_B(\varepsilon)} \;\middle|\; X_1(p), X(p)\right]. \tag{B.61}$$

where we use Eq. (B.35) in the penultimate equality, and the last equality follows from applying Proposition B.1 to both $A_B(\varepsilon)$ and $A_1(\varepsilon) \cup A(\varepsilon)$ as in Eq. (B.35), noting that $A_B^c(\varepsilon)$ bounds $|\mathrm{d}\overline{X}(\varepsilon)|$ by $B\varepsilon$ and that $A_1^c(\varepsilon) \cap A_2^c(\varepsilon) \subseteq A_B^c(\varepsilon)$. We conclude that (δ, w, Y) is a valid stochastic derivative of $\overline{X}(p) = [X_1(p); X(p)]$.

□

B.5 Unbiasedness of pruning strategy

In the main text, we note that we can employ a pruning strategy so that we only ever track one alternative path (i.e. one sample from the stochastic derivative component Y) and yet still obtain an unbiased estimate. The following shows that the pruning method, whereby one chooses between two samples of Y by picking one with probability proportional to its weight, is indeed unbiased.

We prove by induction that the currently tracked alternative path is an unbiased choice amongst all possible alternative paths seen so far. The base case, where the first alternative path observed is followed, is trivial as it is the only choice. Now, suppose that n alternative branches have been observed so far, where $w = w_1 + w_2 + \cdots + w_n$ is the summed weight so far. Suppose we observe an $(n+1)$th branch, with weight w_{n+1}. By our pruning strategy it is chosen with probability

$$\frac{w_{n+1}}{w_{n+1} + w} = \frac{w_{n+1}}{w_1 + w_2 + \cdots + w_{n+1}}, \tag{B.62}$$

while the jth path for some $j \leq n$ will have been chosen after this step with probability

$$\frac{w_j}{w_1 + w_2 + \cdots + w_n} \cdot \frac{w_1 + w_2 + \cdots + w_n}{w_1 + w_2 + \cdots + w_{n+1}} = \frac{w_j}{w_1 + w_2 + \cdots + w_{n+1}}, \tag{B.63}$$

as desired for unbiasedness. Note that this enables us to prune *online*, i.e. without knowing the full structure of the computation a priori, which we exploit in `StochasticAD.jl` for $\mathcal{O}(1)$ memory overhead.

B.6 Smoothing

Recall from the main text,

Definition 2.7 (Smoothed stochastic derivative)**.** For a stochastic program $X(p)$ with a right (left) stochastic derivative (δ, w, Y) at input p, a right (left) smoothed stochastic derivative $\widetilde{\delta}$ of X at input p is given as

$$\widetilde{\delta} = \mathbb{E}\left[\delta + w(Y - X(p)) \mid X(p)\right]. \tag{2.8}$$

Using Definition 2.2, we can write the above in an alternative form that does not rely on the definition of the stochastic derivative,

$$\widetilde{\delta} = \lim_{\varepsilon \to 0^{+/-}} \frac{\mathbb{E}\left[\mathrm{d}X(\varepsilon) \mid X(p)\right]}{\varepsilon}, \tag{B.64}$$

as given in [26, 27]. Smoothed stochastic derivatives propagate through functions that are *locally* linear over the range of Y conditionally on $X(p)$, as we now formalize. Note that this is a much weaker requirement than global linearity over the full range of $X(p)$, as the range of Y conditional on $X(p)$ is generally more restricted (e.g. $Y \in \{X(p) - 1, X(p) + 1\}$ for a binomial variable).

Proposition B.3. *Suppose that f is linear over the range of Y conditionally on $X(p) = x$, for all x. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}p}\mathbb{E}f(X(p)) = \mathbb{E}\left[f'(X(p))\widetilde{\delta}\right]. \tag{B.65}$$

Proof. Note that $(f'(X(p)) \cdot \delta, w, f(Y))$ is a stochastic derivative of $f \circ X$ by Theorem 2.6. Now, by Proposition 2.3 and the local linearity of f, we have the simplification,

$$\frac{\mathrm{d}}{\mathrm{d}p}\mathbb{E}\left[f(X(p))\right] = \mathbb{E}\left[f'(X(p)) \cdot \delta + w\left(f(Y) - f(X(p))\right)\right] \tag{B.66}$$

$$= \mathbb{E}\left[f'(X(p))\left(\delta + w(Y - X(p))\right)\right] \tag{B.67}$$

$$= \mathbb{E}\left[f'(X(p))\widetilde{\delta}\right]. \tag{B.68}$$

□

In most cases, local linearity will only hold approximately, leading to bias in the estimate produced by propagating smoothed stochastic derivatives. However, in the case of a particle filter resampling step an exact estimate is produced, as we show in the following.

C Formalism of the particle filter

C.1 Hidden Markov model

Let us consider a hidden Markov model with random states $X_1, \ldots, X_n$ as specified by a stochastic program $X_1(\theta)$ giving the starting value depending on parameters θ and consecutive states given by pointwise differentiable stochastic programs $X_i(x_{i-1}, \theta)$ depending on the previous state $x_{i-1} \sim X_{i-1}$ (Markov property) and θ. In general, we allow continuous probability densities $p(x_1 \mid \theta)$ of X_1 and continuous transition probability densities $p(x_i \mid x_{i-1}, \theta)$ to depend arbitrarily on θ. This latent process $X_1, \ldots, X_n$ is indirectly observed as the process $y_1 \sim Y_1, \ldots, y_n \sim Y_n$ with n observations $Y_i(x_i)$ depending on $x_i \sim X_i$ which are specified to have smooth conditional probability densities $p(y_i \mid x_i, \theta)$ depending only on x_i and θ. As a concrete, special case, we consider the following linear Gaussian state-space model with a d-dimensional latent process,

$$X_i = \Phi X_{i-1} + \mathrm{Normal}(0, Q), \tag{C.1}$$

$$Y_i = X_i + \mathrm{Normal}(0, R), \tag{C.2}$$

where $Q = 0.02 \cdot \mathbb{1}_{d\times d}$, $R = 0.01 \cdot \mathbb{1}_{d\times d}$, $x_1 \sim \mathrm{Normal}(\mu, 0.001 \cdot \mathbb{1}_{d\times d})$, $\mu \sim \mathrm{Normal}(0, \mathbb{1}_{d\times d})$ is a random initial position, and Φ is a d-dimensional rotation matrix. Here, the parameters θ are defined by the entries of Φ, i.e. $\theta = \mathrm{vec}(\Phi)$. We use a particle filter to compute an estimate of the likelihood $\mathcal{L} = p(y_1, \ldots, y_n \mid \theta)$.

C.2 Differentiating a particle filter with a resampling step

Given n observations $y_1 \sim Y_1, \ldots, y_n \sim Y_n$ of the hidden Markov model defined via Eqs. (C.1) and (C.2), a bootstrap particle filter allows us to approximate the posterior distributions of the states and the likelihood $\mathcal{L}$ of these observations by propagating a cloud of K *weighted* particles. We denote the kth particle by $x_i^{(k)}$. Each particle evolves independently according to the stochastic program [Eqs. (C.1) and (C.2)] and carries a weight $\varpi_i^{(k)}$ measuring how well the trajectory of the particle so far is matching the observations. This importance weight is updated by $\varpi_i^{(k)} = p(y_i \mid x_i^{(k)}, \theta)\varpi_{i-1}^{(k)}$, such that the empirical measure $\sum_k \varpi_i^{(k)} \delta_{x_i^{(k)}} / \sum_k \varpi_i^{(k)}$ of the weighted particles approximates the filtering distribution $p(x_i \mid y_1, \ldots, y_i, \theta)$.

So far, the particles and weight trajectories are differentiable with respect to the parameter [3]. However, weight degeneracy, the collapse of all but a few weights, is a common problem. Our goal is to discard unlikely particles, so that numerical resources are not wasted on particles with vanishing weight. The strategy to accomplish this goal is to include resampling steps, where we pick the particles that best match the observations. However, such resampling steps present discrete randomness, where the particle population is resampled according to the particle weights $\varpi_i^{(k)}$ to form a new population $(x')_i^{(k)}$ with equal weight $\varpi_i' = 1/K \cdot \sum_{k=1}^K \varpi_i^{(k)}$. To differentiate the resampling

step, we need to describe how perturbing the weight distribution $\varpi_i^{(k)}$ provided to the resampling procedure affects the resampled particles and weights. Importantly, the marginal likelihood of a parameter θ can be read off using the weights at the last step

$$p(y_1, \ldots, y_n \mid \theta) \approx \sum_{k=1}^{K} \varpi_n^{(k)}. \tag{C.3}$$

The strategies for resampling vary, but what they have in common is that the *marginal* distribution of each resampled particle $x_i'^{(k)}$ is a multinomial distribution over the original particles with weights given by the normalized weight vector $(\varpi_i^{(1)}/\varpi_i, \ldots, \varpi_i^{(K)}/\varpi_i)$ with $\varpi_i = \sum_{k=1}^{K} \varpi_i^{(k)}$. Thus, a weighted sample of the marginal distribution of each resampled particle can be obtained by repeating the following procedure until obtaining a weight of ϖ_i:

1. Sample an integer k uniformly from 1 to K.
2. Return the particle $x_i^{(k)}$ with assigned weight $\varpi_i \operatorname{Ber}(\varpi_i^{(k)}/\varpi_i)$.

Conditional on the assigned weight being ϖ_i, the returned particle obeys the multinomial distribution. The key insight is that the alternate possibility of a resampled particle $x_i^{(k)}$ not being chosen can be written instead as the alternate possibility of its weight $\varpi_i^{(k)}$ changing to 0.

Recall from Example 2.5 that there is an asymmetry between the left and right stochastic derivative of the Bernoulli distribution $\operatorname{Ber}(\varpi_i^{(k)}/\varpi)$ conditioned on the output 0 or 1 being chosen. In this case, it is convenient to take the *left* stochastic derivative $(0, w_i^{(k)}, Y_i^{(k)})$ with $Y_i^{(k)} = 0$ and $w_i^{(k)} = -\varpi_i^k/\varpi_i$ of the particle's weight ϖ_i because it is 0 when the assigned weight is 0, meaning that this case has no influence on the derivative estimate. Therefore, particles that are not resampled [and thus do not contribute to the primal value, cf. Eq. (C.3)] do (also) not contribute to the derivative computation. The fact that only the weights have a stochastic derivative, but not the particles imposes the following setting:

Let $X(p)$ be a stochastic program approximated by the program $\widetilde{X}(p)$. Assume we can sample from $\widetilde{X}(p)$. Assuming absolute continuity, we may write expectations as $\mathbb{E}\varpi(p)f(\widetilde{X}(p)) = \mathbb{E}f(X(p))$ where $\varpi(p)$ is the Radon-Nikodym derivative [4] of the law of $\widetilde{X}(p)$ with respect to the law of $X(p)$, evaluated in $\widetilde{X}(p)$. We thus consider the program which returns the pair of weight and value

$$(\varpi(p), \widetilde{X}(p)). \tag{C.4}$$

Proposition C.1. *If $\widetilde{X}(p)$ is a continuous program, reparameterized such that it is differentiable pointwise, and $\varpi(p)$ has a smoothed stochastic derivative $\widetilde{\delta}$,*

$$\frac{\mathrm{d}}{\mathrm{d}p}\mathbb{E}f(X(p)) = \mathbb{E}\left[\varpi(p)f'(\widetilde{X}(p))\widetilde{X}'(p) + \widetilde{\delta}f(\widetilde{X}(p))\right]. \tag{C.5}$$

Proof. By assumption,

$$\mathbb{E}f(X(p)) = \mathbb{E}\varpi(p)f(\widetilde{X}(p)). \tag{C.6}$$

As the function $(\varpi, x) \mapsto \varpi f(x)$ is linear in ϖ, the statement follows from Proposition B.3. □

Therefore, we can replace the stochastic derivative of $\varpi \operatorname{Ber}(\varpi_i^{(k)}/\varpi)$ with its smoothed stochastic derivative $w_i^{(k)}(1 - Y_i^{(k)}) = \varpi_i^k/\varpi_i$ obtained using Table 2. This is convenient, as smoothed stochastic derivatives permit forward- and reverse-mode whereas reverse-mode AD becomes superior than forward-mode AD for functions f from $\mathbb{R}^n$ to $\mathbb{R}^m$ with $m \gg n$ [2]. Since the weights are used in a purely linear fashion, Proposition B.3 and Proposition C.1 guarantee that the derivative estimator is unbiased.

Numerically, we accomplish this in our code with a formally differentiable weight function `new_weight(p)` whose primal value is always 1 but whose derivative is the left smoothed stochastic derivative of $\mathrm{Ber}(p)$ for primal output 1,

$$\widetilde{\delta_L} = \frac{1}{p}, \tag{C.7}$$

so that a particle $x_i^{(k)}$ has weight given as

$$\varpi_i \cdot \texttt{new\_weight}(\varpi_i^{(k)}/\varpi_i). \tag{C.8}$$

In [7], an equivalent expression is derived by different means, using the stop-gradient operator $\perp$:

$$\varpi_i \cdot \frac{(\varpi_i^{(k)}/\varpi_i)}{\perp(\varpi_i^{(k)}/\varpi_i)}, \tag{C.9}$$

where $\perp$ is formally assigned a derivative of 0.

D Implementation details

D.1 Experiment details

All computation times in Fig. 5 were measured on an Intel Xeon Platinum 8260 CPU and Julia version 1.6. Garbage collection times are included in the total run time. We provide code and instructions to run the examples in the tutorials folder of `StochasticAD.jl`.

D.2 Software dependencies

`StochasticAD.jl` is implemented in the Julia Language [4] and uses internally `Distributions.jl` and `DistributionsAD.jl` [1, 2], `ChainRulesCore.jl` [5], and `ForwardDiff.jl` [6].

The presented examples are using the additional packages: `BenchmarkTools.jl` [7], `Zygote.jl` [8], `GaussianDistributions.jl` [9], and `Plots.jl` [10].

We refer the reader to the documentation of `StochasticAD.jl` for more detailed information on the package implementation.

E Funding details

This material is based upon work supported by the National Science Foundation under grant no. OAC-1835443, grant no. SII-2029670, grant no. ECCS-2029670, grant no. OAC-2103804, and grant no. PHY-2021825. The information, data, or work presented herein was funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0001211 and DE-AR0001222. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No HR00112290091. We also gratefully acknowledge the U.S. Agency for International Development through Penn State for grant no. S002283-USAID. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. This material was supported by The Research Council of Norway and Equinor ASA through Research Council project "308817 - Digital wells for optimal production and drainage". Research was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. We also acknowledge financial support from the NCCR QSIT funded by the Swiss National Science Foundation (Grant No. 51NF40-185902), as well as from the Swiss National Science Foundation individual grant (Grant No. 200020_200481).

Supplemental References

[SI1] Dahua Lin, John Myles White, Simon Byrne, Douglas Bates, Andreas Noack, John Pearson, Alex Arslan, Kevin Squire, David Anthoff, Theodore Papamarkou, Mathieu Besançon, Jan Drugowitsch, Moritz Schauer, and contributors. JuliaStats/Distributions.jl: a Julia package for probability distributions and associated functions, July 2019. URL https://doi.org/10.5281/zenodo.2647458.

[SI2] Mathieu Besançon, Theodore Papamarkou, David Anthoff, Alex Arslan, Simon Byrne, Dahua Lin, and John Pearson. Distributions.jl: Definition and modeling of probability distributions in the juliastats ecosystem. *Journal of Statistical Software*, 98(16):1–30, 2021.

[SI3] Rico Jonschkowski, Divyam Rastogi, and Oliver Brock. Differentiable particle filters: End-to-end learning with algorithmic priors. *arXiv:1805.11122*, 2018.

[SI4] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

[SI5] Frames Catherine White, Michael Abbott, Miha Zgubic, Jarrett Revels, Alex Arslan, Seth Axen, Simeon Schaub, Nick Robinson, Yingbo Ma, Gaurav Dhingra, Will Tebbutt, Niklas Heim, David Widmann, Andrew David Werner Rosemberg, Niklas Schmitz, Christopher Rackauckas, Rainer Heintzmann, Frank Schäfer, Carlo Lucibello, Keno Fischer, Alex Robson, Jerry Ling, Matt Brzezinski, Andrei Zhabinski, Daniel Wennberg, Mathieu Besançon, Pietro Vertechi, Shashi Gowda, and Andrew Fitzgibbon. JuliaDiff/ChainRules.jl: v1.35.0, May 2022. URL https://doi.org/10.5281/zenodo.6574605.

[SI6] Jarrett Revels, Miles Lubin, and Theodore Papamarkou. Forward-mode automatic differentiation in Julia. *arXiv:1607.07892 [cs.MS]*, 2016.

[SI7] Jiahao Chen and Jarrett Revels. Robust benchmarking in noisy environments. *arXiv:1608.04295*, 2016.

[SI8] Michael Innes. Don't unroll adjoint: Differentiating SSA-form programs. *arXiv:1810.07951*, 2018.

[SI9] Moritz Schauer and contributors. GaussianDistributions.jl, 2018. URL https://github.com/mschauer/GaussianDistributions.jl.

[SI10] Simon Christ, Daniel Schwabeneder, and Christopher Rackauckas. Plots.jl–a user extendable plotting API for the julia programming language. *arXiv:2204.08775*, 2022.