Supermartingales and Termination Analysis of Statistical PCF

April 28, 2021

Abstract

⌜TODO⌟ ◁ L

Contents

Various theorems about probabilistic programs rely on the assumption that the program terminates almost surely. Proof rules based on relating the program state to supermartingales already exist for first-order imperative programs [**?**]. This paper's contribution is to extend this method to a higher-order setting.

1 Statistical PCF

1.1 Syntax of SPCF

The language SPCF is a simply-typed lambda calculus with sampling of real numbers from $[0, 1]$ and unbounded scoring, following **?**. Types and terms are

defined as follows, where r is a real number, x is a variable, $f : \mathbb{R}^n \to \mathbb{R}$ is any measurable function, and Γ is an environment:

types $A, B ::= \mathsf{R} \mid A \to B$

values $V ::= \lambda x.M \mid \underline{r}$

terms $M, N ::= V \mid x \mid M_1 \, M_2 \mid \underline{f}(M_1, \ldots, M_n) \mid \mathsf{Y}M \mid \mathsf{if}(M < 0, N_1, N_2) \mid \mathsf{sample} \mid \mathsf{score}(M)$

$$\frac{}{\Gamma; x : A \vdash x : A} \qquad \frac{\Gamma; x : A \vdash M : B}{\Gamma \vdash \lambda x.M : A \to B} \qquad \frac{}{\underline{r} : \mathsf{R}} \qquad \frac{\Gamma \vdash M : A \to B \quad \Gamma \vdash N : A}{\Gamma \vdash M \, N : B}$$

$$\frac{\Gamma \vdash M_1 : \mathsf{R} \ldots \Gamma \vdash M_n : \mathsf{R}}{\Gamma \vdash \underline{f}(M_1, \ldots, M_n) : \mathsf{R}} \, f : \mathbb{R}^n \to \mathbb{R} \qquad \frac{\Gamma \vdash M : (A \to B) \to (A \to B)}{\Gamma \vdash \mathsf{Y}M : (A \to B)}$$

$$\frac{\Gamma \vdash M : \mathsf{R} \quad \Gamma \vdash N_1 : A \quad \Gamma \vdash N_2 : A}{\Gamma \vdash \mathsf{if}(M < 0, N_1, N_2) : A} \qquad \frac{}{\Gamma \vdash \mathsf{sample} : \mathsf{R}} \qquad \frac{\Gamma \vdash s : \mathsf{R}}{\Gamma \vdash \mathsf{score}(s) : \mathsf{R}}$$

Terms are identified up to α-equivalence, as usual. The set of all terms is denoted Λ, and the set of closed terms is denoted Λ^0.

To define the reduction relation, let *evaluation contexts* be of the form:

$$E ::= [\,] \mid E \, M \mid V \, E \mid \underline{f}(\underline{r}_1, \ldots, \underline{r}_{k-1}, E, M_{k+1}, \ldots, M_n)$$
$$\mid \mathsf{Y}E \mid \mathsf{if}(E < 0, M_1, M_2) \mid \mathsf{score}(E)$$

then a term reduces if it is formed by substituting a redex in a context i.e.

$$E[(\lambda x.M)N] \to E[M[N/x]]$$
$$E[\underline{f}(\underline{r}_1, \ldots, \underline{r}_n)] \to E[\underline{f(r_1, \ldots, r_n)}]$$
$$E[\mathsf{Y}\lambda x.M] \to E[\lambda z.M[(\mathsf{Y}\lambda x.M)/x]z] \text{ where } z \text{ is not free in } s$$
$$E[\mathsf{if}(\underline{r} < 0, M_1, M_2)] \to E[M_1] \text{ where } r < 0$$
$$E[\mathsf{if}(\underline{r} < 0, M_1, M_2)] \to E[M_2] \text{ where } r \geq 0$$
$$E[\mathsf{sample}] \to E[\underline{r}] \text{ where } r \in [0, 1]$$
$$E[\mathsf{score}(\underline{r})] \to E[\underline{r}].$$

We write $\to^*$ for the reflexive, transitive closure of $\to$. $\lceil$The score rule should have a side condition: if $r > 0$.$\rfloor$ ⊜ L ◁ L

Every closed well-typed term either is a value or reduces to another term.

1.2 Sampling semantics

This version of the reduction relation allows sample to reduce to any number in $[0, 1]$. To more precisely specify the probabilities, an additional argument is needed to determine the outcome of random samples. Let $I = [0, 1] \subset \mathbb{R}$, and

let $S = I^{\mathbb{N}}$, with the Borel σ-algebra $\mathcal{F}$ and the probability measure, denoted μ, given by the limit of $1 \leftarrow I \leftarrow I^2 \leftarrow \cdots$, where the maps are the projections that ignore the last element. Equivalently, a basis of measurable sets is $\prod_{i=0}^{\infty} X_i$ where X_i are all Borel and all but finitely many are I, and $\mu(\prod_{i=0}^{\infty} X_i) = \prod_{i=0}^{\infty} \mathrm{Leb}(X_i)$. The maps $\pi_h : S \to I$, $\pi_t : S \to S$ popping the first element are then measurable. Following **?**, we call the probability space $(S, \mathcal{F}, \mu)$ the *entropy space*.

The σ-algebra and measure on Λ are defined by considering it as a disjoint union of equivalence classes under replacing all the real constants by a place-holder, where the measure on each class is that of $\mathbb{R}^n$, where n is the number of real constants. Precisely, following **?**, we view Λ as $\bigcup_{m \in \omega} (\mathsf{Sk}_m \times \mathbb{R}^m)$, where Sk_m is the set of SPCF terms with exactly m place-holders (a.k.a. *skeleton terms*) for numerals. Thus identified, we give Λ the countable disjoint union topology of the product topology of the discrete topology on Sk_m and the standard topology on $\mathbb{R}^m$. Note that the connected components of Λ have the form $\{M\} \times \mathbb{R}^m$, with M ranging over Sk_m, and m over ω. We fix the Borel algebra of this topology to be the σ-algebra on Λ.

The one-step reduction is given by the function $\mathrm{red} : \Lambda^0 \times S \to \Lambda^0 \times S$ where

$$
\mathrm{red}(M, s) = \begin{cases} (E[N], s) & \text{if } M = E[R], R \to N \text{ and } R \neq \mathsf{sample} \\ (E[\underline{\pi_h(s)}], \pi_t(s)) & \text{if } M = E[\mathsf{sample}] \\ (M, s) & \text{if } M \text{ is a value} \end{cases} \tag{1}
$$

⌈Both $\to$ and red are called reduction relation.⌋

The result after n steps is then simply $\mathrm{red}^n(M, s) = \overbrace{\mathrm{red}(\ldots \mathrm{red}(}^{n} M, s)\ldots)$, and the limit red^{∞} can then be defined as a partial function as $\lim_{n \to \infty} \mathrm{red}^n(M, s)$ whenever that sequence becomes constant by reaching a value. A term M terminates for a sample sequence s if the limit $\mathrm{red}^{\infty}(M, s)$ is defined.

The reduction function is measurable, and the set of values is measurable, therefore the set of s such that M terminates at s within n steps is measurable for any n, therefore $\{s \mid M \text{ terminates at } s\}$ is measurable. We say that a term M is *almost surely terminating* (AST) just if $\mu(\{s \mid M \text{ terminates at } s\}) = 1$. ⌈Alternatively, define the *runtime of M* to be the random variable

$$
T_M(s) := \begin{cases} \min\{n \mid \pi_0(\mathrm{red}^n(M, s)) \text{ is a value}\} & \text{if } \mathrm{red}^{\infty}(M, s) \text{ is defined} \\ \infty & \text{otherwise} \end{cases}
$$

Equivalently, we say that M is *almost surely terminating* (AST) if $T_M < \infty$ a.s.; and M is *positively almost surely terminating* (PAST) if $\mathbb{E}[T_M] < \infty$.⌋

For example, the term

$$
(\mathsf{Y} \lambda f, n : \mathsf{if}(\mathsf{sample} - 0.5 < 0, n, f(n+1))) \, \underline{0},
$$

which generates a geometric distribution, terminates on the set $S \setminus [0.5, 1]^{\mathbb{N}}$, which has measure 1, therefore it terminates almost surely, whereas

$$
\mathsf{if}(\mathsf{sample} - 0.5 < 0, \underline{0}, (\mathsf{Y}\lambda x.x)\underline{0})
$$

3

which terminates on the set $\pi_h^{-1}[[0, 0.5]]$, has probability 0.5 of failing to terminate.

Note that the requirement that M be closed is not important for the proofs of the theorems, but is simply required to ensure that M reaches a value, rather than a general normal form like $x\,y$.

This definition of almost sure termination is equivalent to that given in [**?**], although the program semantics is stated in a slightly different way. In particular, the argument to score is not relevant to termination (except for the possibility that its argument's evaluation wouldn't terminate).

2 Supermartingales

One approach to proving that a term terminates almost surely is to find some variant that is bounded below and, on average, decreases sufficiently quickly that it must eventually reach 0, similarly to the approach taken in **?** for imperative programs.

These variants are defined as functions from reachable terms (i.e. possible states of the program's execution) to real numbers. Specifically, let the set of reachable terms from a given closed starting term M be $Rch(M) := \{N \in \Lambda \mid M \to^* N\}$, with the σ-algebra induced as a subset of Λ.

Definition 1. A *ranking function on M* is a measurable function $f : Rch(M) \to \mathbb{R}$ such that $f(N) \geq 0$ for all N, and

1. $f(E[\mathsf{Y}\lambda x.N]) \geq 1 + f(E[\lambda z.N[(\mathsf{Y}\lambda x.N)/x]z])$ where z is not free in N

2. $f(E[\mathsf{sample}]) \geq \int_I f(E[\underline{x}]) \operatorname{Leb}(\mathrm{d}x)$

3. $f(E[R]) \geq f(E[R'])$ for any other redex R, where $R \to R'$.

We say that the ranking function f is *strict* if there exists $\epsilon > 0$ such that $f(N) = 0$ iff N is a value, and for all E and $R \to R'$, $f(E[R']) \leq f(E[R]) - \epsilon$.

Any closed term for which a ranking (respectively, strict ranking) function exists is called *rankable* (respectively, *strictly rankable*). For example, the term $(\mathsf{Y}\lambda x.x)\underline{0}$ is not rankable. It will be demonstrated later that for any rankable term M, if $(M_n)_{n\geq 0}$ is the reduction sequence starting from M, then $(f(M_n))_{n\geq 0}$ is a supermartingale, and M terminates almost surely, but first, some preliminaries about supermartingales. ✏ L

Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We have in mind $\Omega = S$, and $\mathbb{P} = \mu$.

Definition 2. 1. A sequence of random variables $(Y_n)_{n\geq 0}$ adapted to a filtration $(\mathcal{F}_n)_{n\geq 0}$ is a *supermartingale* if for all $n \in \omega$, Y_n is integrable (i.e. $\mathbb{E}[|Y_n|] < \infty$), and $\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq Y_n$ a.s. (i.e. $\forall A \in \mathcal{F}_n, \int_A \mathrm{d}\mu\, Y_{n+1} \leq \int_A \mathrm{d}\mu\, Y_n$). ✏ L

4

2. Let $\epsilon > 0$. A stopping time T and a supermartingale $(Y_n)_{n \geq 0}$, both adapted to filtration $(\mathcal{F}_n)_{n \geq 0}$, are a *$\epsilon$-ranking supermartingale* if for all n, $Y_n \geq 0$ and $\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq Y_n - \epsilon \cdot \mathbf{1}_{\{T>n\}}$.[1] A *ranking supermartingale* is a ϵ-ranking supermartingale for some $\epsilon > 0$.

Remark. The original definition of ranking supermartingale, due to [**??**], does not refer to a stopping time explicitly: Fioriti and Hermanns call a supermartingale $(Y_n)_{n>0}$ *ranking* if $Y_n \geq 0$, and there exists $\epsilon > 0$ such that $\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq Y_n - \epsilon \cdot \mathbf{1}_{\{Y_n > 0\}}$.

Intuitively Y_n gives the rank of the program after n steps of computation, and T is the time at which it reaches a value (which may be infinite if it fails to terminate). In a ϵ-ranking supermartingale, each computation step causes a strict decrease in rank of at least ϵ, provided the term in question is not a value.

Let T and T' be stopping times adapted to $(\mathcal{F}_n)_{n \geq 0}$. Recall the σ-algebra (consisting of measurable subsets "prior to T")

$$\mathcal{F}_T := \{A \in \mathcal{F} \mid \forall i \in \omega . A \cap \{T \leq i\} \in \mathcal{F}_i\}$$

and if $T \leq T'$, then $\mathcal{F}_T \subseteq \mathcal{F}_{T'}$.

The following is an iterated version of Doob's well-known Optional Sampling Theorem (see, e.g., [**?**, §6.7] and [**?**, Theorem 7.2]).

Theorem 1 (Optional Sampling). *Let $(X_n)_{n \geq 0}$ be a supermartingale, and $(T_n)_{n \geq 0}$ a sequence of increasing stopping times, both adapted to filtration $(\mathcal{F}_n)_{n \geq 0}$, then $(X_{T_n})_{n \geq 0}$ is a supermartingale adapted to $(\mathcal{F}_{T_n})_{n \geq 0}$ if one of the following conditions holds:*

1. *each T_n is bounded i.e. $T_n < c_n$ where c_n is a constant*

2. *$(X_n)_{n \geq 0}$ is uniformly integrable.*

Lemma 1. *Let $(Y_n)_{n \geq 0}$ be a ϵ-ranking supermartingale with respect to the stopping time T, then $T < \infty$ almost surely, and $\mathbb{E}[T] \leq \dfrac{\mathbb{E}[Y_0]}{\epsilon}$.*

Proof. We first prove by induction on n:

$$\forall n \in \omega . \mathbb{E}[Y_n] \leq \mathbb{E}[Y_0] - \epsilon \cdot \left(\sum_{i=0}^{n-1} \mathbb{P}[T > i] \right).$$

It follows that $\sum_{i=0}^{\infty} \mathbb{P}[T > i]$ converges; hence $\lim_{i \to \infty} \mathbb{P}[T > i] = 0$ and so $\mathbb{P}[T < \infty] = 1$. It then remains to observe: if $T < \infty$ a.s., then $\mathbb{E}[T] = \sum_{i=0}^{\infty} \mathbb{P}[T > i]$. □

Lemma 1 is a slight extension of [**?**, Lemma 5.5]. In Lemma 5.5 of op. cit., the stopping time T is assumed to be the random variable $s \mapsto \min\{n \mid Y_n(s) = 0\}$.

Our first result is the following theorem.

---

[1]For $A \in \mathcal{F}$, we write $\mathbf{1}_A$ for the *indicator function* of A, i.e., the random variable $s \mapsto \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{otherwise} \end{cases}$.

Theorem 2 (Rankability). *If a closed SPCF term is rankable (respectively, strictly rankable) by f, then $(f(M_n))_{n \geq 0}$ is a supermartingale (respectively, ranking supermartingale) adapted to the filtration $(\mathcal{F}_n)_{n \geq 0}$ where $\mathcal{F}_n = \sigma(M_1, \cdots, M_n)$.*

Notation. Given an ω-sequence (e.g. $s \in S$) and $m \in \omega$, we write $s_{\leq m} \in I^m$ to mean the prefix of s of length m.

Fix a closed SPCF term $M \in \Lambda^0$, and a ranking function f for it, satisfying $f(N) = 0$ iff N is a value. Let $n \in \omega$, define the following random variables on the probability space $(S, \mathcal{F}, \mu)$:

$$M_n(s) := \pi_0(\mathrm{red}^n(M, s))$$

$$\#\mathrm{draw}_n(s) := |\{k < n \mid \exists E : M_k(s) = E[\mathsf{sample}]\}|,$$

so that $\pi_1(\mathrm{red}^n(M, s)) = \overbrace{\pi_t(\cdots(\pi_t}^{\#\mathrm{draw}_n(s)}(s))$, and the filtration $\mathcal{F}_n := \sigma(M_1, \cdots, M_n)$. The $\mathcal{F}_n$-measurability of M_n (and hence of $\#\mathrm{draw}_n$) follows from [**?**]. Take $s \in A \in \mathcal{F}_n$ with $\#\mathrm{draw}_n(s) = l$. For any $s' \in S$, if $s_{\leq l} = s'_{\leq l}$ then $s' \in A$. It follows that $\{s_{\leq l}\} \cdot I^\omega \subseteq A$. [NOTE. To fix notation: $\sigma(M_n) = \sigma(\{M_n^{-1}(\alpha_j, U_{\alpha_j}) \mid$ ◁ L $\alpha_j \in \mathsf{Sk}_j, U_{\alpha_j} \in \mathcal{B}(\mathbb{R}^j), j \geq 0\})$]

We say that a given SPCF term is: *type-1* if it has the shape $E[\mathsf{Y}\lambda x.N]$, *type-2* if it has the shape $E[\mathsf{sample}]$; *type-3* if it has the shape $E[R]$ where R is any other redex, and *type-4* if it is a value. Henceforth fix an $n \geq 0$, and define $\mathbf{T}_i := \{s \mid M_n(s) \text{ is type-}i\}$. It is straightforward to see that each $\mathbf{T}_i \in \mathcal{F}_n$, and $\{\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3, \mathbf{T}_4\}$ is a partition of S. Hence it suffices to prove the following lemma.

Lemma 2. *For all $i \in \{1, 2, 3, 4\}$ and $A \in \mathcal{F}_n$*

$$\int_A \mu(\mathrm{d}s) \, f(M_{n+1})[s \in \mathbf{T}_i] \leq \int_A \mu(\mathrm{d}s) \, f(M_n)[s \in \mathbf{T}_i] \text{ }^2$$

Proof. We show the non-trivial case of $i = 2$. First we express

$$f(M_{n+1})[s \in \mathbf{T}_2] = \sum_{i \in \mathcal{I}} f(E_i[\underline{\sigma_i(s)}][\underline{\rho(s)}])[s \in U_i] \tag{2}$$

where

- $\mathcal{I}$ is a countable indexing set

- $E_i[\cdot][\mathsf{sample}] \in \mathsf{Sk}_{j_i}$, and $\sigma_i : S \to \mathbb{R}^{j_i}$, and $\rho(s) := \pi_h(\pi_1(\mathrm{red}^n(M, s))) = \pi_h(\pi_t^{l_i}(s)) \in \mathbb{R}$

- $\{U_i\}_{i \in \mathcal{I}}$ is a partition of $\mathbf{T}_2$, where each U_i is determined by a skeletal term $E_i[\cdot][\mathsf{sample}]$ and a number (of draws) $l_i \leq n$, satisfying

$$s \in U_i \iff (E_i[\underline{r}][\mathsf{sample}], \pi_t^{l_i}(s)) = \mathrm{red}^n(M, s), \text{ for some } r \in \mathbb{R}^{j_i}$$

---

^{2}where (the Iverson bracket) $[P] := 1$ if the statement P hold, and 0 otherwise.

(We use the (measurable) function $\sigma_i : S \to \mathbb{R}^{j_i}$ to skolemise the existentially quantified r.)

Equivalently $U_i := \#\mathrm{draw}_n^{-1}(l_i) \cap M_n^{-1}(E_i[\cdot][\mathsf{sample}], \mathbb{R}^{j_i}) \in \mathcal{F}_n$.

Observe that if $s \in U_i$ then $\{s_{\leq l_i}\} \cdot I^\omega \subseteq U_i$; in fact $(U_i)_{\leq l_i} \cdot I^\omega = U_i$. This means that for any measurable $g : S \to \mathbb{R}_{\geq 0}$, if $g(s)$ only depends on the prefix of s of length $(l_i + 1)$, then, writing $\widehat{g} : I^{l_i+1} \to \mathbb{R}_{\geq 0}$ where $g(s) = \widehat{g}(s_{\leq l_i+1})$, we have: for any $A \in \mathcal{F}_n$

$$\int_{A \cap U_i} \mu(\mathrm{d}s)\, g(s) = \int_{(A \cap U_i)_{\leq l_i+1}} \mathrm{Leb}_{l_i+1}(\mathrm{d}t)\, \widehat{g}(t) \tag{3}$$

Take $s \in U_i$, and set $l = l_i$. Plainly $\sigma_i(s)$ depends on $s_{\leq l}$, and $\rho(s)$ depends on $s_{\leq l+1}$. Take $u \in (U_i)_{\leq l}$. It then follows from the definition of ranking function that

$$\int_I \mathrm{Leb}(\mathrm{d}r)\, f(E_i[\widehat{\sigma}_i(u)][\underline{r}]) \leq f(E_i[\widehat{\sigma}_i(u)][\mathsf{sample}])).$$

Take $A \in \mathcal{F}_n$, and integrating both sides, we get

$$\int_{(A \cap U_i)_{\leq l}} \mathrm{Leb}_l(\mathrm{d}u) \int_I \mathrm{Leb}(\mathrm{d}r)\, f(E_i[\widehat{\sigma}_i(u)][\underline{r}]))$$
$$\leq \int_{(A \cap U_i)_{\leq l}} \mathrm{Leb}_l(\mathrm{d}u)\, f(E_i[\widehat{\sigma}_i(u)][\mathsf{sample}])).$$

Since Leb_{l+1} is the (unique) product measure satisfying $\mathrm{Leb}_{l+1}(V \times B) = \mathrm{Leb}_l(V) \cdot \mathrm{Leb}(B)$, and $(U_i)_{\leq l_i} \cdot I^\omega = U_i$, we have

$$\int_{(A \cap U_i)_{\leq l+1}} \mathrm{Leb}_{l+1}(\mathrm{d}u')\, f(E_i[\widehat{\sigma}_i(u')][\underline{\widehat{\rho}(u')}]))$$
$$\leq \int_{(A \cap U_i)_{\leq l+1}} \mathrm{Leb}_{l+1}(\mathrm{d}u')\, f(E_i[\widehat{\sigma}_i(u')][\mathsf{sample}]))$$

and so, by (3)

$$\int_{A \cap U_i} \mu(\mathrm{d}s)\, f(E_i[\underline{\sigma_i(s)}][\underline{\rho(s)}])) \leq \int_{A \cap U_i} \mu(\mathrm{d}s)\, f(E_i[\underline{\sigma_i(s)}][\mathsf{sample}])). \tag{4}$$

7

Now, integrating both sides of (2), we have

$$\int_A \mu(\mathrm{d}s)\, f_{n+1}(M)[s \in \mathbf{T}_i] = \int_A \mu(\mathrm{d}s) \sum_{i \in \mathcal{I}} f(E_i[\underline{\sigma_i(s)}][\underline{\rho(s)}])[s \in U_i]$$

$$= \sum_{i \in \mathcal{I}} \int_{A \cap U_i} \mu(\mathrm{d}s)\, f(E_i[\underline{\sigma_i(s)}][\underline{\rho(s)}])$$

$$\leq \sum_{i \in \mathcal{I}} \int_{A \cap U_i} \mu(\mathrm{d}s)\, f(E_i[\underline{\sigma_i(s)}][\mathsf{sample}]) \quad \because (4)$$

$$= \int_A \mu(\mathrm{d}s) \sum_{i \in \mathcal{I}} f(E_i[\underline{\sigma_i(s)}][\mathsf{sample}])[s \in U_i]$$

$$= \int_A \mu(\mathrm{d}s) f_n(M)[s \in \mathbf{T}_2]$$

$$\qquad \square$$

As an immediate corollary of Lemma 2, each $f(M_n)$ is integrable:

$$\int_S \mu(\mathrm{d}s)\, |f(M_n)(s)| = \int_S \mu(\mathrm{d}s)\, f(M_n)(s) \leq \int_S \mu(\mathrm{d}s)\, f(M_0)(s) = f(M).$$

This concludes the proof of Theorem 2. $\qquad \square$

As before, fix a closed SPCF term $M \in \Lambda^0$. Define random variables on the probability space $(S, \mathcal{F}, \mu)$:

$$T_{-1}(s) := -1$$
$$T_{n+1}(s) := \min\{k \mid k > T_n(s), M_k(s) \text{ a value or of form } E[\mathsf{Y}\lambda x.N]\}$$
$$X_n(s) := f(M_n(s))$$
$$Y_n(s) := X_{T_n}(s)$$
$$T_M^{\mathsf{Y}}(s) := \min\{n \mid M_{T_n(s)}(s) \text{ is a value}\} \tag{5}$$

The random variable T_M^{Y}, which we call the Y-*runtime of M*, can equivalently be defined as the number of Y-reduction steps in the reduction sequence of M. Note that, as the type system ensures that the reduction relation excluding Y-reduction is strongly normalising, only finitely many reductions can occur in a row without one of them being a Y-reduction, therefore $T_M^{\mathsf{Y}} < \infty$ a.s. iff M is AST.

The r.v. T_M^{Y} is a stopping time adapted to the filtration $(\mathcal{F}_{T_n})_{n \geq 0}$.

Theorem 3 (Soundness). *1. If a closed SPCF term M is rankable, then M is AST (equivalently, $T_M^{\mathsf{Y}} < \infty$ almost surely), and M is Y-positively almost sure terminating (Y-PAST) i.e. $\mathbb{E}[T_M^{\mathsf{Y}}] < \infty$.*

2. If a closed SPCF term M is strictly rankable, then $\mathbb{E}[T_M] < \infty$ i.e. M is positively almost sure terminating (PAST).

8

Proof. Let f be a ranking function on M. For 1, observe that $T_0, T_1, T_2, \cdots$ are an increasing sequence of stopping times, each is bounded [Prove it.] and ◁ L adapted to $(\mathcal{F}_n)_{n \geq 0}$. Then, since $(f(M_n))_{n \geq 0}$ is a supermartingale also adapted to $(\mathcal{F}_n)_{n \geq 0}$ (Theorem 2), it follows from the Optional Sampling Theorem 1 that $(Y_n)_{n \geq 0}$ where $Y_n := f(M_{T_n})$ is a supermartingale adapted to the filtration $(\mathcal{F}_{T_n})_{n \geq 0}$. Since T_M^{Y} is a stopping time also adapted to $(\mathcal{F}_{T_n})_{n \geq 0}$, we have that $(Y_n)_{n \geq 0}$ is a 1-ranking supermartingale. Therefore, by Lemma 1, $T_M^{\mathsf{Y}} < \infty$ a.s. and $\mathbb{E}[T_M^{\mathsf{Y}}] < \infty$.

Statement 2 follows immediately from Lemma 1. □

Thus the method of (strict) ranking function is sound for proving (positive) a.s. termination of SPCF programs. It is in fact also complete in a sense: if $\mathbb{E}[T_N^{\mathsf{Y}}] < \infty$ for all $N \in Rch(M)$ then M is rankable (Theorem 4).

3   Constructing Ranking Functions

Although rankability implies almost sure termination, the converse does not hold in general. For example,

$$\text{if}(-\mathsf{sample} < 0, \underline{0}, (\mathsf{Y}\lambda x.x)\,\underline{0}) \tag{6}$$

terminates in 3 steps with probability 1, but isn't rankable because $(\mathsf{Y}\lambda x.x)\,\underline{0}$ is reachable, although that has probability 0. Not only is this counterexample AST, it's PAST.

A ranking function can be constructed under the stronger assumptions that, for every N reachable from M, the expected number of Y-reduction steps from N to a value is finite. In particular, the expected number of Y-reduction steps from each reachable term is a ranking function. Note that a finite number of expected Y-reduction steps does not necessarily imply a finite number of expected total reduction steps.

Example 1. The term

$$M = \big(\mathsf{Y}\lambda fn.\text{if}(\mathsf{sample} - 0.5 < 0, n\,\underline{0}, f(\lambda x.n(nx))))\big)(\lambda x.x + 1)$$

terminates with only 2 Y-reductions on average i.e. $\mathbb{E}[T_M^{\mathsf{Y}}] = 2$, but applies the increment function 2^n times with probability 2^{-1-n} for $n \geq 0$, which diverges i.e. $\mathbb{E}[T_M] = \infty$.

Theorem 4. *Given a closed term M, the function $f : Rch(M) \to \mathbb{R}$ given by $f(N) = \mathbb{E}[$the number of Y-reduction steps from N to a value$]$, if it exists, is the least of all possible ranking functions of M.*

Proof. Let f be the candidate least ranking function defined above, and suppose g is another ranking function such that $f(N) > g(N)$ for some $N \in Rch(M)$. The restrictions of f and g to $Rch(N)$ have the same properties assumed of f and g, so assume w.l.o.g. that $N = M$. The difference $g - f$ is then a

supermartingale (with the same setup as in the proof of Theorem 2); therefore $\mathbb{E}[g(M_n)] \leq \mathbb{E}[f(M_n)] + g(M) - f(M)$, for all n.

We have $\mathbb{E}[f(M_n)] = \sum_{k=n}^{\infty} \mathbb{P}[M_k = E[\mathsf{Y}N]$ for some $E, N] \to 0$ as $n \to \infty$; therefore as $g(M) - f(M) < 0$, eventually $\mathbb{E}[g(M_n)] < 0$, which is impossible. It follows that $g \geq f$ as required.

In order for f to be the least ranking function of M, it also has to actually be a ranking function itself. Each of the conditions on a ranking function is easily verified from the definition of f. $\qquad\square$

3.1 Partial ranking functions

Even in the case of reasonable simple terms, explicitly constructing a ranking function would be a lot of work, and Theorem 4 makes even stronger assumptions than almost sure termination, so it isn't useful for proving it.

Example 2 (Geometric distribution)**.** Consider the term

$$(\mathsf{Y}\lambda f, n.$$
$$\quad \mathsf{if}(\mathsf{sample} - 0.5 < 0, n, f(n+1))$$
$$)\,\underline{0}$$

which generates a geometric distribution (let $\Theta = \lambda f, n : \mathsf{if}(\mathsf{sample}\underline{-0.5} < 0, n, f(n\underline{+1}))$, then the term is $(\mathsf{Y}\Theta)\,\underline{0}$). Despite its simplicity, its *Rch* contains all the terms

$$\mathsf{Y}\,\Theta\,\underline{0}$$
$$\wr$$
$$(\lambda z.(\lambda n.\mathsf{if}(\mathsf{sample}\underline{-0.5} < 0, n, \mathsf{Y}\,\Theta\,(n\underline{+1})))z)\,\underline{i}$$
$$\wr$$
$$(\lambda n.\mathsf{if}(\mathsf{sample}\underline{-0.5} < 0, n, \mathsf{Y}\,\Theta\,(n\underline{+1})))\,\underline{i}$$
$$\wr$$
$$\mathsf{if}(\mathsf{sample}\underline{-0.5} < 0, \underline{i}, \mathsf{Y}\,\Theta\,(\underline{i}\,\underline{+1}))$$
$$\wr$$
$$\mathsf{if}(\underline{r}\,\underline{-0.5} < 0, \underline{i}, \mathsf{Y}\,\Theta\,(\underline{i}\,\underline{+1}))$$
$$\wr$$
$$\mathsf{if}(\underline{r} < 0, \underline{i}, \mathsf{Y}\,\Theta\,(\underline{i}\,\underline{+1}))$$
$$\searrow$$
$$\underline{i}$$
$$\mathsf{Y}\,\Theta\,(\underline{i}\,\underline{+1})$$
$$\wr$$
$$(\lambda z.(\lambda n.\mathsf{if}(\mathsf{sample}\underline{-0.5} < 0, n, \mathsf{Y}\,\Theta\,(n\underline{+1})))z)\,(\underline{i}\,\underline{+1})$$

Even in this simple case, defining a ranking function explicitly is awkward because of the number of cases, although in most cases, because the value need

only be greater than or equal to that of the next term in sequence, it suffices to take the ranking function as having the same value as the next term, so that overall it takes only 3 distinct values. (We will explain why later.)

The definition of rankability is also inconvenient for syntactic sugar. It could be useful, for example, to define $M \oplus_p N = \mathsf{if}(\mathsf{sample} - p < 0, M, N)$, where $M \oplus_p N$ reduces to M or N, depending on the first value of $s\lceil$Need a name ◁ L for $s.\rfloor$, with probability p resp. $(1 - p)$. Technically though, it reduces first to $\mathsf{if}(\underline{r} - p < 0, M, N)$ for all $r \in I$, so those terms all need values of the ranking function too.

In both of these cases, there are only some values of the ranking function that are semantically important.

Definition 3. Define a *partial ranking function* on a closed term M to be a partial function $f : Rch(M) \rightharpoonup \mathbb{R}$ such that

- $f(N) \geq 0$ for all N where f is defined.

- f is defined at M.

- For any N in the domain of definition of f, evaluation of N will eventually $\lceil$It would be nicer for this to be only almost certain, to make this more ◁ A neatly in-between rankability and ast, but that doesn't actually work because of terms reachable with 0 probability. Of course, astness doesn't imply rankability, so any definition of partial ranking function that includes all Y-past terms wouldn't always extend to a total ranking function.$\rfloor$reach some O which is either a value or in the domain of definition of f, and $f(N) \geq \mathbb{E}[f(O) + $ the number of Y-reduction steps from N to $O]$ (where $f(O)$ is taken to be 0 if O is a value outside of the domain of f).

A partial ranking function that is total is just a ranking function. Providing a partial ranking function is essentially part way between providing a ranking function and directly proving almost sure termination.

Theorem 5. *Every partial ranking function is a restriction of a ranking function.*

Proof. Take a closed term M and partial ranking function f on M. Define $f_1 : Rch(M) \rightharpoonup \mathbb{R}$ by

$$f_1(N) := f(N) \text{ whenever } f(N) \text{ is defined,}$$
$$f_1(V) := 0 \text{ for values } V \text{ not in the domain of } f.$$

Define $(\mathrm{next}(N, s), \_) = \mathrm{red}^n(N, s)$ for the least $n \geq 0$ such that it's in the domain of f_1, and $g(N, s) := |\{m < n \mid \mathrm{red}^m(N, s) \text{ is of the form } (E[\mathsf{Y}N'], s')\}|$. The function next is well-defined (i.e. n is finite) for all $N \in Rch(M)$ by induction on the path from M to N, by the third condition on partial ranking functions. Define $f_2(N) = \int_S f_1(\mathrm{next}(N, s)) + g(N, s) \, \mu(\mathrm{d}s)$. The (total) function f_2 agrees with f on f's domain, and it is a ranking function on M (in fact,

the least ranking function of which f is a restriction). $\lceil$TODO. The preceding ◁ L
sentence deserves a proof.$\rfloor$ □

As a corollary, any term which admits a partial ranking function terminates almost surely.

3.2 Examples

Let $M \oplus_p N = \mathsf{if}(\mathsf{sample} - p < 0, M, N)$, for $p \in (0, 1]$, then there are the pseudo-reduction relations

$$E[M \oplus_p N] \to^3 E[M]$$

$$E[M \oplus_p N] \to^3 E[N]$$

$$\mathrm{red}^3(E[M \oplus_p N], s) = \begin{cases} (E[M], \pi_t(s)) & \text{if } \pi_h(s) < p \\ (E[N], \pi_t(s)) & \text{if } \pi_h(s) \geq p. \end{cases}$$

A partial ranking function could be defined with respect to this shortcut reduction simply by replacing $\to$ and red in the definition of a partial ranking function by a version that goes straight from $N \oplus_p O$ to N or O. Such a pseudo-partial ranking function would then be a partial function from a subset of $Rch(M)$, so it could also be considered as a partial function from all of $Rch(M)$, and it would in fact also be an actual partial ranking function. It is therefore possible to prove rankability directly using the shortcutted reductions.

A similar procedure would work for other forms of syntactic sugar. If a closed term N eventually reduces to one of a set of other terms $\{N_i \mid i \in I\}$ with certain probabilities, a partial ranking function defined with respect to a reduction sequence that skips straight from N to N_i is also a valid partial ranking function for the original reduction function, and therefore its existence implies almost sure termination. There is a caveat, however, that Y-reduction steps skipped over in the shortcut still need to be counted for the expected number of Y-reduction steps.

With this abbreviation, the geometric distribution example from earlier can be written as $(\mathsf{Y}\lambda fn.\ n \oplus_{0.5} f(n+1))\ \underline{0}$. It is then easy to see that the following is a partial ranking function:

$$(\mathsf{Y}\lambda fn.\ n \oplus_{0.5} f(n+1))\ N \mapsto 2$$

$$\underline{i} \oplus_{0.5} (\mathsf{Y}\lambda fn.\ n \oplus_{0.5} f(n+1))\ (\underline{i}+1) \mapsto 1$$

$$\underline{i} \mapsto 0.$$

In fact, even the partial function $\mathsf{Y}\Theta N \mapsto 2$ alone is a partial ranking function for this term.

4 Antitone ranking functions

Example 3 (Random walk). 1. 1D biased (towards 0) random walk

$$M_1 = \big(\mathsf{Y}\,\lambda fn.\,\mathsf{if}(n = 0, 0, f\,(n - 1) \oplus_{3/2} f\,(n + 1))\big)\,10$$

2. 1D unbiased random walk

$$M_2 = \big(\mathsf{Y}\,\lambda fn.\,\mathsf{if}(n = 0, 0, f\,(n - 1) \oplus_{1/2} f\,(n + 1))\big)\,10$$

3. 1D biased (away from 0) random walk

$$M_3 = \big(\mathsf{Y}\,\lambda fn.\,\mathsf{if}(n = 0, 0, f\,(n - 1) \oplus_{1/3} f\,(n + 2))\big)\,10$$

The term M_1 is rankable, terminating in 31 Y-steps on average, and M_3 only has a $1/1024$ chance of terminating, but in between, M_2 is AST, but isn't Y-PAST, therefore it isn't rankable and Theorem 3 is insufficient to prove its termination. We therefore want to find a generalised notion of ranking function so that M_2 becomes rankable, and then prove it sound i.e. rankable in this generalised sense implies AST.

Idea. The definition of antitone-convex ranking function f (say) for $M \in \Lambda^0$ is the same as that of ranking function except that in the case of Y-redex, we require the existence of an antitone (meaning: $r < r'$ implies $\epsilon(r) \geq \epsilon(r')$) function $\epsilon : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ such that the ranking function $f : Rch(M) \to \mathbb{R}_{\geq 0}$ satisfies

$$f(E[R']) \leq f(E[R]) - \epsilon(f(E[R]))$$

where $R \to R'$ is the Y-redex rule.

Definition 4. We say that a supermartingale $(Y_n)_{n \geq 0}$ and a stopping time T adapted to filtration $(\mathcal{F}_n)_{n \geq 0}$ are *antitone strict* if $Y_n \geq 0$, and there exists an antitone function $\epsilon : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ satisfying

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq Y_n - \epsilon(Y_n) \cdot \mathbf{1}_{\{T > n\}}$$

Theorem 6. *Let $(Y_n)_{n \geq 0}$ be an antitone strict supermartingale via function $\epsilon : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$, and stopping time T. Then $T < \infty$ almost surely.*

Proof. First, as (Y_n) is a supermartingale, $\mathbb{E}[Y_n] \leq \mathbb{E}[Y_0]$. Therefore

$$\mathbb{E}[Y_n \mid T > n]$$
$$= \quad \{\text{ rearranging terms }\}$$
$$\frac{\mathbb{P}[T > n]\,\mathbb{E}[Y_n \mid T > n] + \mathbb{P}[T \leq n]\,\mathbb{E}[Y_n \mid T \leq n] - \mathbb{P}[T \leq n]\,\mathbb{E}[Y_n \mid T \leq n]}{\mathbb{P}[T > n]}$$
$$= \quad \{\text{ definition of conditional expectation }\}$$

$$\frac{\mathbb{E}[Y_n] - \mathbb{P}[T \leq n]\,\mathbb{E}[Y_n \mid T \leq n]}{\mathbb{P}[T > n]}$$

$\leq$ $\{\ Y_n \geq 0$ always $\}$

$$\frac{\mathbb{E}[Y_n]}{\mathbb{P}[T > n]}$$

$\leq$ $\{\ (Y_n)_n$ is a supermartingale $\}$

$$\frac{\mathbb{E}[Y_0]}{\mathbb{P}[T > n]}$$

Claim: For all $0 < x \leq 1$, $\mathbb{P}[T > B_x] \leq x$, where $B_x = \left\lceil \dfrac{\mathbb{E}[Y_0] + 1}{x\,\epsilon(\mathbb{E}[Y_0]\,x^{-1})} \right\rceil$.

As the convex hull of ϵ (the greatest convex function less than or equal to it) satisfies all the conditions assumed of ϵ, in addition to being convex, assume wlog that ϵ is convex.

Assume for a contradiction that $\mathbb{P}[T > B_x] > x$. Then, take $n \leq B_x$. We have

$\mathbb{E}[Y_n - Y_{n+1}]$

$=$ $\{\ \mathcal{F}_n \subseteq \mathcal{F}_{n+1}$, definition and linearity of conditional expectation $\}$

$\mathbb{E}[Y_n - \mathbb{E}[Y_{n+1} \mid \mathcal{F}_n]]$

$\geq$ $\{$ antitone strict assumption $\}$

$\mathbb{E}[\epsilon(Y_n) \cdot \mathbf{1}_{\{T > n\}}]$

$=$ $\{$ definition of expectation conditioning on an *event* $\}$

$\mathbb{P}[T > n]\,\mathbb{E}[\epsilon(Y_n) \mid T > n]$

$\geq$ $\{$ Jensen's inequality $\}$

$\mathbb{P}[T > n]\,\epsilon(\mathbb{E}[Y_n \mid T > n])$

$\geq$ $\{$ proved earlier $\}$

$\mathbb{P}[T > n]\,\epsilon\left(\dfrac{\mathbb{E}[Y_0]}{\mathbb{P}[T > n]}\right)$

$>$ $\{$ assumption, $\mathbb{P}[T > n] \geq \mathbb{P}[T > B_x] > x$ $\}$

$x\,\epsilon\left(\mathbb{E}[Y_0]\,x^{-1}\right)$.

Therefore, by a telescoping sum

$$\mathbb{E}[Y_{B_x}] = \mathbb{E}[Y_{B_x} - Y_0 + Y_0] \leq \mathbb{E}[Y_0] - B_x\,x\,\epsilon(\mathbb{E}[Y_0]\,x^{-1}) \leq -1 < 0$$

which is a contradiction, therefore the claim must be true, therefore $P[T > n]$ tends to 0 as n tends to infinity, therefore $T < \infty$ almost surely. $\qquad\square$

Note that it is essential in this proof that ϵ be defined for all $\mathbb{R}_{\geq 0}$, not just the values that $(Y_n)_{n \geq 0}$ actually takes (or at least if $(Y_n)_{n \geq 0}$ is uniformly bounded, ϵ must be positive at the supremum as well as the actual values).

Example 4. ⌈Consider a random variable X, uniformly distributed in the interval $[-1, 1]$, $T := \min\{n \mid 2^{-n} < X\}$,

$$Y_k := \begin{cases} 4 - 2^{-k} & \text{if } k < T \\ 0 & \text{otherwise} \end{cases}$$

⌋⌈The following reformulation defines T and Y_k directly as a measurable function on a probability space (rather than via a r.v. X).⌋

Take the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the closed interval of reals $[-1, 1]$, $\mathcal{F}$ the Borel σ-algebra, and $\mathbb{P}$ the corresponding Lebesgue probability measure. Let $\omega \in [-1, 1]$. Define random variables T and $(Y_k)_{k \in \omega}$:

$$T(\omega) := \begin{cases} \min\{n \in \mathbb{N} \mid \omega > 2^{-n}\} & \text{if } \omega \in (0, 1] \\ \infty & \text{otherwise} \end{cases}$$

$$Y_k(\omega) := \begin{cases} 4 - 2^{-k} & \text{if } \omega \in [-1, 2^{-k}] \text{ (equivalently } k < T(\omega)) \\ 0 & \text{otherwise} \end{cases}$$

⌈N.B. $\{T = n\} = (\frac{1}{2^n}, \frac{1}{2^{n-1}}]$, and so, $\{T > n\} = [-1, \frac{1}{2^n}]$.⌋Plainly, T is a stopping time, and $(Y_n)_{n \geq 0}$ is a supermartingale, adapted to the filtration $(\mathcal{F}_n)_{n \geq 0}$ where $\mathcal{F}_n = \mathcal{F}$ for all n. In this case, $(Y_n)_{n \geq 0}$ either tends to 4 as $n \to \infty$, or drops to 0 at some point, with an exponentially decreasing probability. With respect to the antitone function $\epsilon(x) = \frac{4-x}{4}$ and stopping time T, we have that $(Y_n)_{n \geq 0}$ is an antitone strict supermartingale, except that $\epsilon(4) = 0$, and $T = \infty$ with probability $\frac{1}{2}$, even though 4 is larger than any value $(Y_n)_{n \geq 0}$ can actually take, and $\epsilon(Y_n)$ never actually reaches 0.

Let f be a ranking function on a closed SPCF term M. We say that M is *antitone rankable* by f if the supermartingale $(f(M_n))_{n \geq 0}$ and the stopping time T_M^{Y}, both adapted to the fitration $(\mathcal{F}_n)_{n \geq 0}$ where $\mathcal{F}_n := \sigma(M_1, \ldots, M_n)$, are antitone strict.

Theorem 7 (Antitone ranking function soundness)**.** *If a closed SPCF term M is antitone rankable, then $T_M^{\mathsf{Y}} < \infty$ a.s. (equivalently, M is AST).*

Proof. As in Theorem 3, take the set S of traces ⌈TODO: We should find a better name for traces $s \in S$.⌋as a probability space, and define the same random variables $T_n, X_n, Y_n, T_M^{\mathsf{Y}}$. As before, $(Y_n)_{n \geq 0}$ is a supermartingale, and because M is now assumed to be antitone rankable, it is an antitone strict supermartingale with respect to the stopping time T_M^{Y}. Therefore, it follows from Theorem 6 that T_M^{Y} is almost surely finite. □

As before, constructing antitone ranking fuctions completely is not necessary, and there is a corresponding notion of an antitone partial ranking function.

Define an *antitone partial ranking function* on a closed term M to be a partial function $F : Rch(M) \rightharpoonup \mathbb{R}$ such that for some antitone function $\epsilon : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$

- $f(N) \geq 0$ for all N where f is defined.

- f is defined at M.

- For any N in the domain of definition of f, evaluation of N will eventually reach some O which is either a value or in the domain of definition of f, and $f(N) \geq \mathbb{E}[f(O) + \epsilon(f(O)) \times$ the number of Y-reduction steps from N to O] (where $f(O)$ is taken to be 0 if O is a value outside of the domain of f).

Theorem 8. *Every antitone partial ranking function is a restriction of an antitone ranking function.*

Proof. Take a closed term M and an antitone partial ranking function f on M, with a corresponding antitone function ϵ. Assume wlog that ϵ is convex (as if it isn't, we can just take its convex hull instead). As in Theorem 5, define $f_1 : Rch(M) \rightharpoonup \mathbb{R}$ by

$$f_1(N) = f(N) \text{ whenever } f(N) \text{ is defined,}$$
$$f_1(V) = 0 \text{ for values } V \text{ not in the domain of } f.$$

Define $(\text{next}(N, s), \_) = \text{red}^n(N, s)$ for the least $n \geq 0$ such that it's in the domain of f_1, and $g(N, s) = |\{m < n \mid \text{red}^m(N, s) \text{ is of the form } (E[\mathsf{Y}N'], s')\}|$. The function next is well-defined (i.e. n is finite) for all $N \in Rch(M)$ by induction on the path from M to N, by the third condition on antitone partial ranking functions. Define $f_2(N) = \int_S f_1(\text{next}(N, s)) + \epsilon(f_1(\text{next}(N, s)))g(N, s)\,\mu(\mathrm{d}s)$. For any term N where f is defined, $f_2(N) = f(N)$, and the value that f_2 would have at N if f were not defined at N is $\leq f(N)$, by the third condition on antitone partial ranking functions. In order to show that f_2 is an antitone partial ranking function, it therefore suffices to show that the value that f_2 would have had at each term if f were not defined at that term is at least the expectation of f_2 after one reduction step (plus $\epsilon(f_2(N))$ if the reduction step is a $t\mathsf{Y}$-reduction). For any term N which is not of the form $E[R]$ for some Y-redex R, this is trivial. If R is a Y-redex, then $\epsilon(f_2(N)) \leq \int_S \epsilon(f_1(\text{next}(N, s)))\,\mu(\mathrm{d}s)$ by the convexity of ϵ, because n is bounded. Therefore, the (total) function f_2, which agrees with f on f's domain, is an antitone ranking function on M (with the same function ϵ if it's convex). $\qquad\square$

As a corollary, any term which admits an antitone partial ranking function terminates almost surely.

Example 3 revisited ⌈All of these ranking functions are slightly incorrect, ◁ A because $\Xi\,\underline{n}$ reduces to something like $\Xi\,(\underline{n}+1)$ then $(\lambda z.\text{something}[\Xi/f])(\underline{n}+1)$, rather than reducing the argument first. This could be fixed by appealing to the confluent semantics version of the antitone ranking theorem, but it would be nice to have at least some examples here illustrating how antitone ranking functions work, rather than actually saving all these examples until the end.

Even though the difference is quite trivial, this was actually the original motivating example for the confluent trace semantics.⌋

Programs M_1 and M_2 are AST. For M_1, the partial ranking function

$$\left(\mathsf{Y}\,\lambda fn.\,\mathsf{if}(n=0,0,f\,(n-1)\oplus_{3/2}f\,(n+1))\right)\underline{x}\mapsto 3x+1$$

suffices to prove its termination (and could equivalently be considered an anti-tone partial ranking function with the constant antitone function $\epsilon_1(x)=1$).

For M_2, define the function $g_1:\mathbb{R}_{\geq 0}\to\mathbb{R}_{\geq 0}$ by $g_1(x)=\log(x+1)+1$. Using shorthand $\Theta_2=\mathsf{Y}\,\lambda fn.\,\mathsf{if}(n=0,0,f\,(n-1)\oplus_{1/2}f\,(n+1))$, we can define an antitone partial ranking function $f_2:Rch(M_2)\rightharpoonup\mathbb{R}_{\geq 0}$ defined as: for $n\in\mathbb{N}$

$$\Theta_2\,n\mapsto g_2(n)$$
$$0\mapsto 0.$$

For $n\geq 1$, $\Theta_2\,n$ reduces in several steps to either $\Theta_2\,(n-1)$ or $\Theta_2\,(n+1)$, each with probability $1/2$, with one Y-reduction. Now

$$
\begin{aligned}
g_2(n)-\frac{g_2(n-1)+g_2(n+1)}{2}&=\log\left(\frac{n+1}{\sqrt{n(n+2)}}\right)\\
&=\frac{1}{2}\log\left(\frac{(n+1)^2}{n(n+2)}\right)\\
&=\frac{1}{2}\log\left(1+\frac{1}{n(n+2)}\right)\\
&>\frac{1}{4n(n+2)}\\
&>\frac{1}{2}\frac{1}{5(n+1)^2}+\frac{1}{2}\frac{1}{5(n+3)^2}
\end{aligned}
$$

(The first inequality follows from the fact that for all $x\in(0,1)$, $\log(1+x)>x$.) ✍ L
Moreover $\Theta_2\,0$ reduces to 0 with one Y-reduction with a reduction in g_2 of 1. Therefore by setting $\epsilon_2(x)=\frac{1}{5(e^{x-1}+1)^2}$, the condition on how much g_2 must decrease is met. It is also defined at $M_2=\Theta_2\,10$, and is non-negative, therefore it is an antitone partial ranking function, and M_2 is AST.

Example 5 (Continuous random walk)**.** In SPCF we can construct a function whose argument changes by a random amount at each recursive call:

$$\underbrace{\left(\mathsf{Y}\,\lambda fx.\mathsf{if}(x\leq 0,0,f(x-\mathsf{sample}))\right)}_{\Theta}\,10$$

We can construct a partial ranking function f as follows:

$$
\begin{aligned}
\Theta\,\underline{l}&\mapsto l+2\\
\mathsf{if}(l\leq 0,0,\Theta\,(l-\mathsf{sample}))&\mapsto l+1\\
0&\mapsto 0.
\end{aligned}
$$

For a more complex (not Y-PAST) example, consider the following "continuous random walk"

$$\underbrace{\left(\mathsf{Y}\,f\,x.\mathsf{if}(x \le 0, 0, f(x - \mathsf{sample} + 1/2))\right)}_{\Xi}\,10$$

Let $g(x) := 2 + \log(x+1)$, and let ϵ be the function specified by

$$\epsilon(g(x - 1/2)) = g(x) - \int_{x-1/2}^{x+1/2} g(y)\,\mathrm{d}y.$$

The limit of this as $x \to \infty$ and $g(x + 1/2) \to \infty$ is 0, and $\frac{\mathrm{d}}{\mathrm{d}x}\epsilon(g(x + 1/2)) = \frac{1}{x+1} + \log(1 - \frac{1}{x+1/2}) < 0$, and g is monotonic increasing; therefore ϵ is antitone and bounded below by 0. We define a partial antitone ranking function as follows:

$$\Xi\,l \mapsto g(l)$$
$$0 \mapsto 0$$

The value of g after one Y-reduction step is at least $g(l - 1/2)$, therefore the expectation of ϵ after one Y-reduction step is at most $\epsilon(g(l - 1/2)) = g(x) - \int_{x-1/2}^{x+1/2} g(y)\,\mathrm{d}y$. Thus

- in case $l > 0$, g decreases by the required amount

- in case $l \le 0$, $g(\Xi\,l) \ge 2 + \log(1/2) > \log(\frac{3\sqrt{3}}{2}) - 1 = \epsilon(g(0 - 1/2))$ as well.

Hence this is a valid partial antitone ranking function and the term is AST.

Example 6 (Fair-in-the-limit random walk)**.** [**?**, §5.3]

$$\underbrace{\left(\mathsf{Y}\,f\,x.\mathsf{if}(x \le 0, 0, f(x - 1) \oplus_{\frac{x}{2x+1}} f(x + 1))\right)}_{\Xi}\,10$$

To construct an antitone ranking function, we solve the recurrence relation:

$$\begin{cases} z_0 &= 0 \\ n \ge 1, \quad z_n &> \dfrac{n}{2n+1}z_{n-1} + \dfrac{n+1}{2n+1}z_{n+1}. \end{cases}$$

For $n > 2$, $z_n = \log(n-1)$ works. The expected decrease is $\frac{1}{2}(\log(1 + \frac{1}{(n-1)^2 - 1}) - \frac{1}{2n+1}\log(1 + \frac{2}{n-2}))$, and using the fact that $\frac{x}{x+1} \le \log(1+x) \le x$ for $x > 0$, ⌈The preceding inequality is incorrect: for $x \in (0,1)$, $\log(1 + x) > x$.⌋ this is at least $\frac{1}{2}(\frac{1}{(n-1)^2} - \frac{1}{2n+1}\frac{2}{n-2}) = \frac{n^2}{2(n-1)^2(2n+1)(n-2)}$, which (again for $n > 2$) is positive and antitone. For $n = 0, 1, 2$, we take ϵ to be $9/40$ (the same as its value at 3), then set $z_2 = 2\log 2 - \log 3 - 1$, $z_1 = 3\log 2 - 2\log 3 - 3$, $z_0 = 4\log 2 - 3\log 3 - 6$. Some of those values are negative, but it's still bounded below, so by adding a constant offset it can be corrected, and the term is AST. ◁ L

Example 7 (Escaping spline). [**?**, §5.4]

$$\underbrace{\left(\mathsf{Y}\, f\, x.\, 0 \oplus_{\frac{1}{x+1}} f(x+1)\right)}_{\Xi}\, 10$$

In this case, the fact that the ranking function must decrease at each Y-step (in an antitone partial ranking function) by the expected value of ϵ not at the current term, but at the next term where the ranking function is defined, is a little harder to deal with, because the variable x can change all the way to 0 in one step, therefore simply adding a small offset doesn't suffice to compensate for this fact.

Consider the candidate ranking function $\Xi\, n \mapsto n+1$. For each n, $\Xi\, n$ reduces to either 0 (with probability $\frac{1}{n+1}$) or $\Xi\,(n+1)$, therefore the expected value of the ranking function is $\frac{n(n+2)}{n+1} = n + 1 - \frac{1}{(n+1)^2}$, and the required decrease is $\frac{\epsilon(0)+n\epsilon(n+2)}{n+1}$, therefore eventually, the expected decrease isn't enough, whatever the value of $\epsilon(0)$ is.

If instead, we take the ranking function to be defined after the Y-reduction but before the **sample**-reduction as well, this can be resolved. Letting $\Theta[n] = \underline{0} \oplus_{\frac{1}{n}} \Xi\, n$:

$$\Theta[n] \mapsto n$$
$$\Xi\, 10 \mapsto 12.$$

For each n, $\Theta[n]$ reduces to either 0 or $\Theta[n+1]$, with a Y-reduction only in the latter case, and the condition that this is an antitone partial ranking function is that $n \geq \frac{(n-1)(n+1)}{n} + \frac{n-1}{n}\epsilon(n+1)$ and $12 \geq 11 + \epsilon(11)$. These are satisfied by setting $\epsilon(x) = \min(1, 1/x)$, therefore this term is AST.

Completeness   It seems very likely that this method is complete in the case that every reachable term is AST, but we have been unable to actually prove this. It is certainly at least capable of proving the termination of terms which terminate arbitrarily slowly (in the sense that there is no similar limitation to Theorem 3, which can only prove termination of Y-PAST terms).

The following theorem does not prove completeness, but is suggestive in that direction:

Theorem 9. *For any stopping time T which is almost surely finite, if $(\mathcal{F}_n)_n$ is the coarsest filtration to which T is adapted, then there is a supermartingale $(Y_n)_n$ adapted to $(\mathcal{F}_n)_n$ and an antitone function ϵ such that $(Y_n)_n$ is an antitone ranking supermartingale with respect to T and ϵ.*

Proof. If T is bounded by b a.s. (for n constant), the statement is trivially true by taking ϵ to be constantly 1, and $Y_n = b-n$. Otherwise, let $(t_n)_{n\geq0}$ be defined recursively such that $t_0 = 0$, $t_{n+1} > t_n$ and $\mathbb{P}[T > t_{n+1}] \leq \frac{1}{4}\mathbb{P}[T > t_n]$. We will then define a supermartingale $(Y_n)_{n\geq0}$ and nonrandom sequence $(y_n)_{n\geq0}$ such that $Y_n = 0$ iff $T < n$, $Y_n = y_n$ iff $T \geq n$, and $y_n \geq 2^k$ for $n \geq t_k$.

The antitone function ϵ is defined piecewise and recursively in such a way as to force all the necessary constraints to hold:

$$\text{for } x \in [0,1),\ \epsilon(x) = 1$$

$$\text{for } x \in [2^k, 2^{k+1}),\ \epsilon(x) = \min\left(\epsilon(2^{k-1}), \frac{2^k}{t_{k+1} - t_k}\right).$$

This ensures that

- ϵ is weakly decreasing.

- ϵ is bounded below by 0.

- For $x \geq 2^k$, $\epsilon(x) \leq \frac{2^k}{t_{k+1} - t_k}$. ⌈This would imply $\frac{2^k}{t_{k+1} - t_k} \geq \epsilon(2^{k-1})$. ◁ L Why?⌉⌈Corrected: it should have been min, not max.⌉ ◁ A

The sequence $(y_n)_{n \geq 0}$ is then defined recursively by

$$y_0 = 2, \quad y_{n+1} = (y_n - \epsilon(y_n))\frac{\mathbb{P}[T > n]}{\mathbb{P}[T > n+1]}.$$

The fact $y_{t_k} \geq 2^{k+1}$ is proven by induction on k. For base case, $2 \geq 2$, as required. For the inductive case $k+1$, we first do another induction to prove that $y_n \geq 2^k$ for all $t_k \leq n \leq t_{k+1}$. The base case of this inner induction follows from the outer induction hypothesis a fortiori. Take the greatest k such that $n > t_k$. By the induction hypothesis, $y_{t_k} \geq 2^k$.

$$y_n = y_{t_k}\frac{\mathbb{P}[T > t_k]}{\mathbb{P}[T > n]} - \sum_{m=t_k}^{n-1} \epsilon(y_m)\frac{\mathbb{P}[T > m+1]}{\mathbb{P}[T > n]}$$

$$\geq \frac{\mathbb{P}[T > t_k]}{\mathbb{P}[T > n]}\left(y_{t_k} - \sum_{m=t_k}^{n-1} \epsilon(y_m)\right)$$

$$\geq \frac{\mathbb{P}[T > t_k]}{\mathbb{P}[T > n]}\left(y_{t_k} - \sum_{m=t_k}^{n-1} \frac{2^k}{t_{k+1} - t_k}\right)$$

$$= \frac{\mathbb{P}[T > t_k]}{\mathbb{P}[T > n]}\left(y_{t_k} - \frac{2^k(n - t_k)}{t_{k+1} - t_k}\right)$$

$$\geq \frac{\mathbb{P}[T > t_k]}{\mathbb{P}[T > n]}\left(y_{t_k} - 2^k\right)$$

$$\geq \frac{\mathbb{P}[T > t_k]}{\mathbb{P}[T > n]}\left(2^{k+1} - 2^k\right)$$

$$= \frac{\mathbb{P}[T > t_k]}{\mathbb{P}[T > n]}2^k$$

$$\geq 2^k$$

as required. Substituting $n = t_{k+1}$ and using the fact that $\frac{\mathbb{P}[T > t_k]}{\mathbb{P}[T > t_{k+1}]} > 4$, the same reasoning gives $y_{t_{k+1}} \geq 2^{k+2}$, as required.

Although the stronger induction hypothesis was necessary for the induction to work, the only reason this is needed is to prove that $y_n > 0$ for all n, which implies that $Y_n \geq 0$. The other condition for (Y_n) to be an antitone-strict supermartingale is

$$
\begin{aligned}
&\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \\
=&\mathbb{E}[y_{n+1} \mathbf{1}_{\{T \geq n+1\}} \mid \mathcal{F}_n] \\
=&y_{n+1} \mathbb{E}[\mathbf{1}_{\{T \geq n+1\}} \mid \mathcal{F}_n] \\
=&y_{n+1} \mathbf{1}_{\{T \geq n\}} \frac{\mathbb{P}[T > n+1]}{\mathbb{P}[T > n]} \\
=&(y_n - \epsilon(y_n)) \mathbf{1}_{\{T \geq n\}} \\
=&Y_n - \epsilon(y_n) \mathbf{1}_{\{T \geq n\}}
\end{aligned}
$$

therefore $(Y_n)_n$ is an antitone strict supermartingale with respect to (T, ϵ). $\square$

The assumption that $(\mathcal{F}_n)_{n \geq 0}$ is the coarsest filtration to which T is adapted is used in the equality between the third and fourth lines here. This condition is the main reason that this proof does not extend directly to a completeness result for antitone ranking functions.

5   Y stopping times are finite

As before, fix a $M \in \Lambda^0$. Recall the random variables on $(S, \mathcal{F}, \mu)$:

$$T_0(s) := 0$$
$$T_{n+1}(s) := \min\{k \mid k > T_n(s), M_k(s) \text{ a value or of form } E[\mathsf{Y}\lambda x.N]\}$$

Lemma 3. $T_0, T_1, T_2, \cdots$ *are an increasing sequence of stopping times adapted to* $(\mathcal{F}_n)_{n \geq 0}$*, and each* T_i *is bounded.*

We first show that T_1 is bounded i.e. $T_1 \leq n$, for some $n \in \omega$.

A first proof idea is to construct a reduction argument to the strong normalisation of the simply-typed lambda-calculus. There is a classical transform of conditionals into the pure lambda-calculus. However, each evaluation of sample (almost surely) returns a different number, which complicates such a transform.

Given a closed SPCF term M, we transform it to a term $\ulcorner M \urcorner$ of the nondeterministic simply-typed lambda-calculus as follows. (We may assume that the nondeterministic calculus is generated from a base type ι with a ι-type constant

symbol r, and a function symbol $\bot : A$ for each function type A.)

$$\ulcorner\mathsf{sample}\urcorner := r$$
$$\ulcorner y\urcorner := y$$
$$\ulcorner \lambda y.N\urcorner := \lambda y.\ulcorner N\urcorner$$
$$\ulcorner M_1\,M_2\urcorner := \ulcorner M_1\urcorner\ulcorner M_2\urcorner$$
$$n \geq 0,\ \ulcorner f\,M_1\cdots M_n\urcorner := (\lambda z_1\cdots z_n.r)\ulcorner M_1\urcorner\cdots\ulcorner M_n\urcorner$$
$$\ulcorner\mathsf{if}(B,M_1,M_2)\urcorner := \big(\lambda z.(\ulcorner M_1\urcorner + \ulcorner M_2\urcorner)\big)\ulcorner B\urcorner$$
$$\ulcorner\mathsf{Y}\,f\,y.N\urcorner := \lambda y.\ulcorner N\urcorner[\bot/f]$$
$$\ulcorner\mathsf{score}(N)\urcorner := \ulcorner N\urcorner$$

In the above, variables $z, z_1, \cdots, z_n$ are assumed to be fresh; f ranges over primitive functions and numerals. In transforming $\mathsf{score}(N)$, we set aside the requirement that score is undefined on negative numbers.

The idea is that $\ulcorner M\urcorner$ captures the initial, non-recursive operational behaviour of M, so that $\ulcorner M\urcorner$ simulates the reduction of M until the latter reaches a value, or a term of the form $E[\mathsf{Y}\lambda x.N]$.

It is straightforward to see that for every trace $s \in S$, there is a reduction sequence of $\ulcorner M\urcorner$ that simulates (an initial subsequene of) the reduction of M under s; and the simulation can be made lockstep.

Finally thanks to the following

Theorem 10 (de Groote)**.** *The simply-typed nondeterministic lambda-calculus is strongly normalising.* $\qquad\square$

we have:

1. Every $\ulcorner M\urcorner$-reduction terminates on reaching r, or a term of the shape $E[\bot N_1\cdots N_n]$.

2. Further there is a finite bound (say l) on the length of such $\ulcorner M\urcorner$-reduction sequences; and l bounds the stopping time T_1.

⌜Note: **?** considered a call-by-name version of a simply-typed nondeterministic   ◁ L
lambda-calculus, and proved it strongly normalising, using a reducibility argument à la Tait. One should check that it extends to the call-by-value calculus.
⌟

6 Confluent Trace Semantics

When proving almost sure termination in this way, it is necessary to consider which terms a given term may reduce to. Sometimes however, the reduction that the programmer has in mind may not be strictly the call-by-value order defined so far, or considering an alternative reduction order may be simpler or more intuitive.

Non-probabilistic lambda calculi generally have the Church-Rosser property, that if a term A reduces to both B_1 and B_2, there is some C to which both B_1 and B_2 reduce, so the reduction order mostly doesn't matter. In the probabilistic case, this may not be true, because β-reduction can duplicate samples, so the outputs of the copies of the sample may be identical or independent, depending on whether the sample is taken before or after β-reduction. There are, however, some restricted variations on the reduction order that do not have this problem.

Even with this restriction, a sampling semantics in the style of the one already defined would not be entirely Church-Rosser, as, for example, $\mathrm{red}^3(\mathsf{sample} - \mathsf{sample}, (1, 0, \dots))$ would be either 1 or -1 depending on the order of evaluation of the samples, as that determines which sample from the pre-selected sequence is used for each one. To fix this, rather than pre-selecting samples according to the order they'll be drawn in, select them according to the position in the term where they'll be used instead.

A *position* is a finite sequence of steps into a term, defined inductively as

$$\alpha ::= \cdot \mid \lambda; \alpha \mid @_1; \alpha \mid @_2; \alpha \mid \underline{f}_i; \alpha \mid \mathsf{Y}; \alpha \mid \mathsf{if}_1; \alpha \mid \mathsf{if}_2; \alpha \mid \mathsf{if}_3; \alpha \mid \mathsf{score}; \alpha.$$

The *subterm of M at a position α*, denoted $M \mid \alpha$, is defined as

$$M \mid \cdot = M$$
$$\lambda x.M \mid \lambda; \alpha = M \mid \alpha$$
$$M_1 M_2 \mid @_i; \alpha = M_i \mid \alpha \quad \text{for } i = 1, 2$$
$$\underline{f}(M_1, \dots, M_n) \mid \underline{f}_i; \alpha = M_i \mid \alpha \quad \text{for } i \leq n$$
$$\mathsf{Y} M \mid \mathsf{Y}; \alpha = M \mid \alpha$$
$$\mathsf{if}(M_1 < 0, M_2, M_3) \mid \mathsf{if}_i; \alpha = M_i \mid \alpha \quad \text{for } i = 1, 2, 3$$
$$\mathsf{score}(M) \mid \mathsf{score}; \alpha = M \mid \alpha$$

so that every subterm is located at a unique position, but not every position corresponds to a subterm (e.g. $x\,y \mid \lambda$ is undefined). A position such that $M \mid \alpha$ does exist is said to *occur* in M. *Substitution* of N at position α in M, written $M[N/\alpha]$, is defined similarly. For example, let

$$M = \lambda x\,y.y\,(\mathsf{if}(x < 0, y\,(\underline{f}(x)), \underline{3})) \qquad \alpha = \lambda; \lambda; @_2; \mathsf{if}_2; @_2$$

then $M[\mathsf{sample}/\alpha] = \lambda x\,y.y\,(\mathsf{if}(x < 0, y\,\mathsf{sample}, \underline{3}))$.

Two subterms N_1 and N_2 of a term M, corresponding to positions α_1 and α_2, can overlap in a few different ways. If α_1 is an initial segment (i.e. prefix) of α_2 (written as $\alpha_1 \leq \alpha_2$), then N_2 is also a subterm of N_1. If neither $\alpha_1 \leq \alpha_2$ nor $\alpha_1 \geq \alpha_2$, the positions are said to be *disjoint*. The notion of disjointness is mostly relevant in that if α_1 and α_2 are disjoint, performing a substitution at α_1 will leave the subterm at α_2 inaffected. ⌈The preceding sentence is unclear. ◁ L Typo?⌋⌈It wasn't a typo, but I've made it a little more explicit now. Is this ◁ A clearer?⌋

With this notation, a more general reduction relation $\rightarrow$ can be defined. ⌈It ◁ A would almost work to have one of the neater versions of Y-reduction here.⌋

Definition 5. The binary relation $\to$ is defined by the following rules, each is conditional on a redex occurring at position α in the term M:

$$\text{if } M \mid \alpha = (\lambda x.N)V, \; M \to M[N[V/x]/\alpha]$$

$$\text{if } M \mid \alpha = \underline{f}(\underline{r}_1, \ldots, \underline{r}_n), \; M \to M[\underline{f(r_1, \ldots, r_n)}/\alpha]$$

$$\text{if } M \mid \alpha = \mathsf{Y}\lambda x.N, \; M \to M[\lambda z.N[(\mathsf{Y}\lambda x.N)/x]z/\alpha] \text{ where } z \text{ is not free in } N$$

$$\text{if } M \mid \alpha = \mathsf{if}(\underline{r} < 0, N_1, N_2), \; M \to M[N_1/\alpha] \text{ where } r < 0$$

$$\text{if } M \mid \alpha = \mathsf{if}(\underline{r} < 0, N_1, N_2), \; M \to M[N_2/\alpha] \text{ where } r \geq 0$$

$$\text{if } M \mid \alpha = \mathsf{sample} \text{ and } \lambda \text{ does not occur after } @_2 \text{ or } \mathsf{Y} \text{ in } \alpha,$$

$$M \to M[\underline{r}/\alpha] \text{ where } r \in [0, 1]$$

$$\text{if } M \mid \alpha = \mathsf{score}(\underline{r}), \; M \to M[\underline{r}/\alpha].$$

In each of these cases, $M \mid \alpha$ is the *redex*, and reduction takes place at α.

Labelling the pre-chosen samples by the positions in the term would also not work because in some cases, a sample will be duplicated before being reduced, for example, in $(\lambda x.x\,\underline{0} + x\,\underline{0})(\lambda y.\mathsf{sample})$, both of the sample redexes that eventually occur originate at $@_2; \lambda$. It is therefore necessary to consider possible positions that may occur in other terms reachable from the original term. Even this is itself inadequate because some of the positions in different reachable terms need to be considered the same, and the number of reachable terms is in general uncountable, which leads to measure-theoretic issues.

Define a *skeleton* to be a term but, instead of having real constants $\underline{r}$, it has a placeholder X, so that each term M has a skeleton $Sk(M)$, and each skeleton S can be converted to a term $S[r]$ given a vector r of n real numbers to substitute in, where n is the number of occurrences of X in S. ⌈Do we actually use the ◁ L notation $S[r]$ in what follows?⌋⌈Yes, in Lemma 11 for example.⌋⌈Skeleton is ◁ A defined in Sec. 1.2.⌋The positions in a skeleton and the reduction relation on ◁ L skeletons can be extended from the definitions on terms in the obvious way, with $\mathsf{if}(\mathsf{X} < 0, A, B)$ reducing nondeterministically to both A and B, and sample reducing to X. For example, we have $(\lambda x.\mathsf{if}(x < 0, x, \mathsf{X}))\,\mathsf{sample} \to (\lambda x.\mathsf{if}(x < 0, x, \mathsf{X}))\,\mathsf{X} \to \mathsf{if}(\mathsf{X} < 0, \mathsf{X}, \mathsf{X}) \to \mathsf{X}$

Given a closed term M, let $L_0(M)$ be the set of pairs, the first element of which is a $\to$-reduction sequence of skeletons starting at $Sk(M)$, and the second of which is a position in the final skeleton of the reduction sequence. As with the traces from $I^{\mathbb{N}}$ used to pre-select samples to use in the usual linear sampling semantics, there will be (with one more caveat introduced later) a sample from I pre-selected for each element of $L_0(M)$, which will then be used if a sample reduction is ever performed at that position.

Reduction sequences are used rather than reachable skeletons because if the same skeleton is reached twice, different samples may be needed. For example, consider the term $M = \mathsf{Y}(\lambda fx. \; \mathsf{if}(\mathsf{sample} - \underline{0.5} < 0, f\,x, x))\,\underline{0}$, which reduces after a few steps to $N = \mathsf{if}(\mathsf{sample} - \underline{0.5} < 0, M, \underline{0})$. If the pre-selected sample for $(N, \mathsf{if}_1; \underline{\;}_1)$ is less than 0.5, N reduces back to M, then N again, then the

same sample is used the next time, therefore it's an infinite loop, whereas if samples are labelled by reduction sequences, the samples for $M \to \cdots \to N$ are independent from the samples for $M \to \cdots \to N \to M \to \cdots \to N$, and so on.

The reduction sequences of skeletons will often be discussed as though they were just skeletons, identifying them with their final skeletons. With this abuse of notation, a reduction sequence N (actually $N_1 \to \cdots \to N_n = N$) may be said to reduce to a reduction sequence O, where the reduction sequence implicitly associated with the final skeleton O is $N_1 \to \cdots \to N_n \to O$.

This is still not quite sufficient because sometimes the same samples must be used at corresponding positions in different reduction sequences. For example, the term $M = \mathsf{sample} + \mathsf{sample}$ has the reachable skeletons $N_1 = \mathsf{X} + \mathsf{sample}$, $N_2 = \mathsf{sample} + \mathsf{X}$, $O = \mathsf{X} + \mathsf{X}$ and X, with reductions $M \to N_1 \to O \to \mathsf{X}$ and $M \to N_2 \to O \to \mathsf{X}$. In the reduction $M \to N_1$, the sample labelled $(M, \pm_2)$ is used, and in the reduction $N_2 \to O$, the sample labelled $(M \to N_2, \pm_2)$ is used. Each of these samples becomes the value of the second numeral in O in their respective reduction sequences, therefore in order for Church-Rosserness to be attained, they must be the same. Which elements of $L_0(M)$ must match can be described by the relation $\sim^*$:

Definition 6. The relation $\sim$ is defined as the union of the minimal symmetric relations $\sim_p$ ("p" for parent-child) and $\sim_c$ ("c" for cousin) satisfying

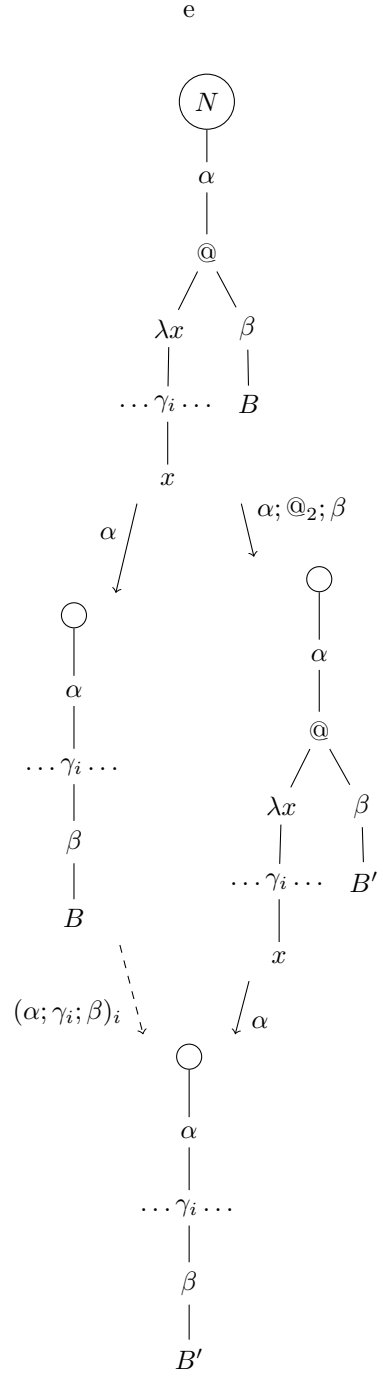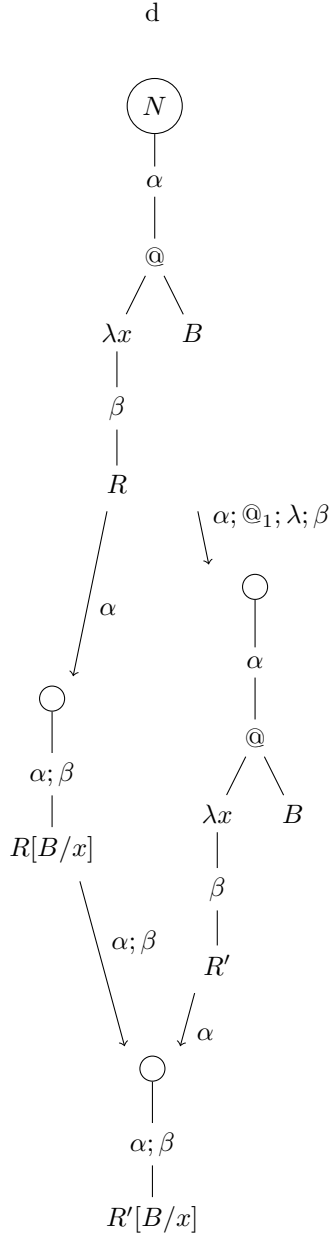1. If N reduces to O with the redex at position α, and β is a position in N disjoint from α, then $(N, \beta) \sim_p (O, \beta)$.

2. If N β-reduces to O at position α, β is a position in $N \mid \alpha; @_1; \lambda$ and $N \mid \alpha; @_1; \lambda; \beta$ is not the variable involved in the reduction, $(N, \alpha; @_1; \lambda; \beta) \sim_p (O, \alpha; \beta)$

3. If N if-reduces to O at position α, with the first resp. second branch being taken, and $\alpha; \mathsf{if}_i; \beta$ occurs in N (where $i = 2$ resp. 3), $(N, \alpha; \mathsf{if}_i; \beta) \sim_p (O, \alpha; \beta)$

4. If N, O_1 and O_2 match any of the following cases:

   (a) N contains redexes at disjoint positions α_1 and α_2, O_1 is N reduced first at α_1 then α_2 and O_2 is N reduced first at α_2 then at α_1.

   (b) $N \mid \alpha = \mathsf{if}(\underline{r} < 0, N_1, N_2)$, where $r < 0$ (or, respectively, $r \geq 0$), $(N_2$ resp. $N_1) \mid \beta$ is a redex, and O_1 is N reduced at α and O_2 is N reduced first at $\alpha; (\mathsf{if}_3$ resp. $\mathsf{if}_2); \beta$ then at α

   (c) $N \mid \alpha = \mathsf{if}(\underline{r} < 0, N_1, N_2)$, where $r < 0$ (or, respectively, $r \geq 0$), $(N_1$ resp. $N_2) \mid \beta$ is a redex, and O_1 is N reduced first at α then at $\alpha; \beta$ and O_2 is N reduced first at $\alpha; (\mathsf{if}_2$ resp. $\mathsf{if}_3); \beta$ then at α

   (d) $N \mid \alpha = (\lambda x.A)B$, there is a redex in A at position β, O_1 is N reduced first at α then at $\alpha; \beta$, and O_2 is N reduced first at $\alpha; @_1; \lambda; \beta$ then at α

(e) $N \mid \alpha = (\lambda x.A)B$, $B \mid \beta$ is a redex, $(\gamma_i)_i$ is a list of all the positions in A where $A \mid \gamma = x$, ordered from left to right, O_1 is N reduced first at $\alpha; @_2; \beta$ then at α, and O_2 is N reduced first at α then at $\alpha; \gamma_i; \beta$ for each i in order.

(f) $N \mid \alpha = \mathsf{Y}(\lambda x.A)$, A reduced at β is A', $(\gamma_i)_i$ is a list of all the positions where $A' \mid \gamma = x$, ordered from left to right, O_1 is N reduced first at $\alpha; \mathsf{Y}; \lambda; \beta$ then at α, and O_2 is N reduced first at α then at $\alpha; \lambda; @_1; \gamma_i; Y; \lambda; \beta$ for each i in order where γ_i is left of β then at $\alpha; \lambda; @_1; \beta$ then at $\alpha; \lambda; @_1; \gamma_i; Y; \lambda; \beta$ for the remaining values of i.
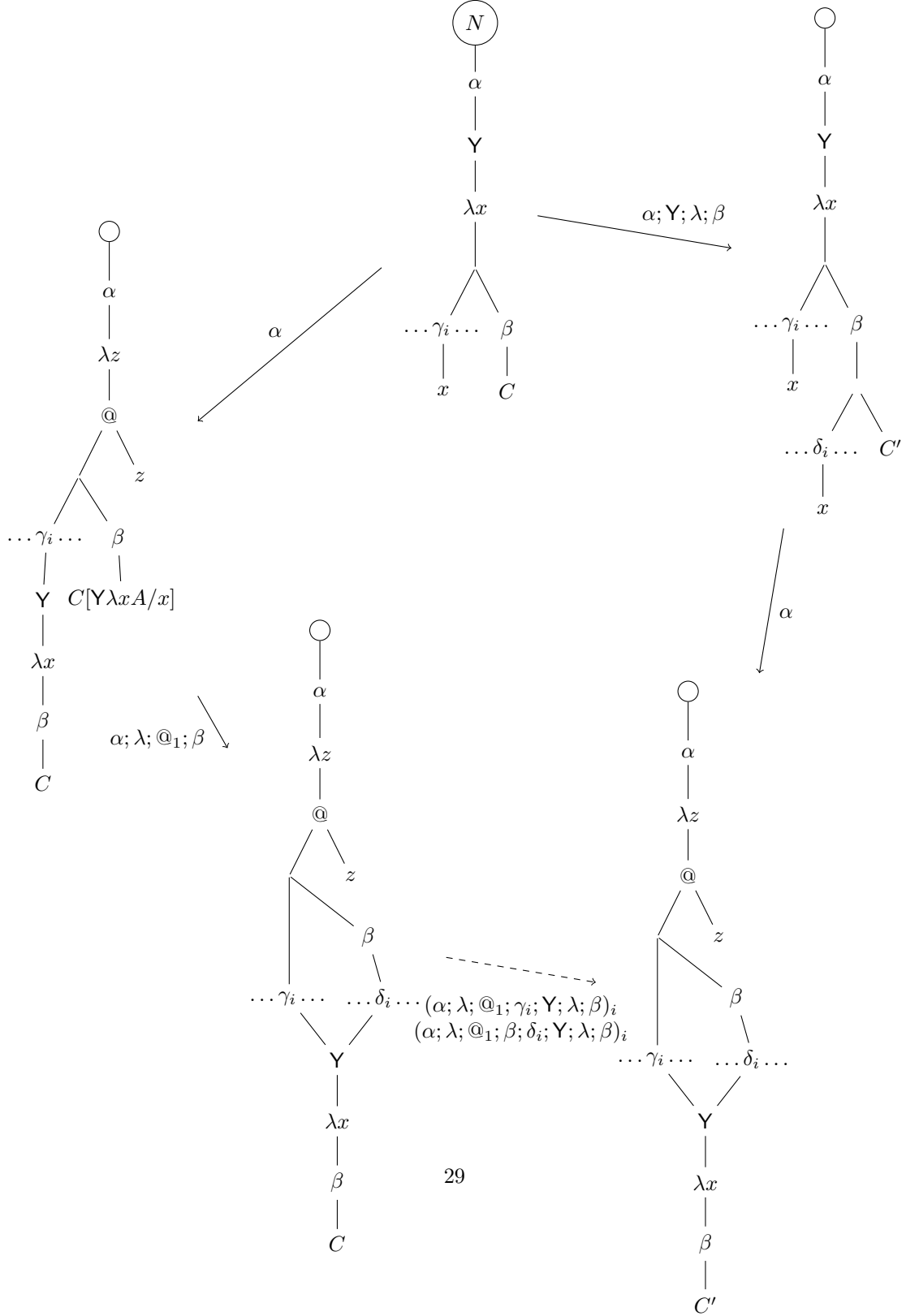
(in which case O_1 and O_2 are equal as skeletons, but with different reduction sequences), O_1' and O_2' are the results of applying some reduction sequence to each of O_1 and O_2 (the same reductions in each case, which is always possible because they're equal terms), and δ is a position in O_1' (or equivalently O_2'), then $(O_1', \delta) \sim_c (O_2', \delta)$.

The $\sim_c$-rules are illustrated in the following diagrams.

a

N

α_1 α_2

A_1 A_2

α_1 α_2

α_1 α_2 α_1 α_2

A'_1 A_2 A_1 A'_2

α_2 α_1

α_1 α_2

A'_1 A'_2

b

N

α

if

$\underline{r}$ A β

B

$\alpha; \mathsf{if}_3; \beta$

α

if

$\underline{r}$ A β

B'

α

α

α

A

c

N

α

if

$\underline{r}$ A β

B

$\alpha; \mathsf{if}_3; \beta$

α

α

$\alpha; \beta$

if

B

$\underline{r}$ A β

B'

$\alpha; \beta$

α

$\alpha; \beta$

B'

27

d

N

α

@

λx B

β

R

α

$\alpha; @_1; \lambda; \beta$

α

@

λx B

$\alpha; \beta$

$R[B/x]$

$\alpha; \beta$

β

R'

$\alpha; \beta$

α

$\alpha; \beta$

$R'[B/x]$

e

N

α

@

λx β

$\dots \gamma_i \dots$ B

x

α

$\alpha; @_2; \beta$

α

α

$\dots \gamma_i \dots$

α

@

λx β

β

$\dots \gamma_i \dots$ B'

B

x

$(\alpha; \gamma_i; \beta)_i$

α

α

$\dots \gamma_i \dots$

β

B'

28

f

$\alpha; \mathsf{Y}; \lambda; \beta$

α

$\alpha; \lambda; @_1; \beta$

$C[\mathsf{Y}\lambda x A/x]$

$(\alpha; \lambda; @_1; \gamma_i; \mathsf{Y}; \lambda; \beta)_i$
$(\alpha; \lambda; @_1; \beta; \delta_i; \mathsf{Y}; \lambda; \beta)_i$

29

The reflexive transitive closure $\sim^*$ of this relation is used to define the set of *potential positions* $L(M) = L_0(M)/\sim^*$, and each equivalence class can be considered as the same position as it may occur across multiple reachable skeletons. If $(N, \alpha) \sim^* (O, \beta)$, then $N \mid \alpha$ and $O \mid \beta$ both have the same shape (i.e. they're either both the placeholder X, both variables, both applications, both samples etc.), therefore it's well-defined to talk of the set of potential positions where there is a sample, $L_s(M)$. Formally $L_s(M) := \{[(X, \alpha)]_{\sim_M^*} : X \mid \alpha = \text{sample}\}$. The new sample space is then defined as $I^{L_s(M)}$, with the Borel σ-algebra and product measure. Since $I^{L_s(M)}$ is a countable product, the measure space is $\circledcirc$ L well-defined [?, Cor. 2.7.3].

Although the six cases in the definition of $\sim_c$ are defined separately, there are some commonalities worth noting. Each case (except a, which is symmetrical) has a right branch and a left branch (the branches of case b are unfortunately displayed the wrong way around in the diagram). The right branch has reduction steps at two positions, β then α, where $\beta > \alpha$ (i.e. β is inside α), and in the left branch, the order of these reductions are swapped, so the reductions are at α then $\beta_1, \ldots, \beta_n$. The position α of one of the reductions is unchanged, and the other reduction may have multiple (in cases e and f) or 0 (in cases b and e) images in the left branch. The images (β_i) are all still inside (greater than or equal to) α, and although they are not generally equal to β, the subskeletons at these positions have a similar shape to the subskeleton (initially) at position β, so that the reductions at these positions are the same type of reduction (e.g. both β-reduction or both if-reduction). Case a is also similar, except that the relevant positions are all disjoint rather than contained in one another, and either branch could be considered the right branch.

Before defining the new version of the reduction relation red, a few lemmas about the properties of $\sim$ are necessary for it to be well-defined.

Lemma 4. *If $(N_1, \alpha_1) \sim^* (N_2, \alpha_2)$, there is some descendants N_1', N_2' of N_1 resp. N_2, and a position α' in both of them such that $(N_1, \alpha_1) \sim_p^* (N_1', \alpha') \sim_c^* (N_2', \alpha') \sim_p^* (N_2, \alpha_2)$.*

Proof. The $\sim_p$ relation can be split into $\sim_\downarrow \cup \sim_\uparrow$, where $(A, \alpha) \sim_\downarrow (B, \beta)$ if $A \to B$ and $(A, \alpha) \sim_p (B, \beta)$, and similarly $(A, \alpha) \sim_\uparrow (B, \beta)$ if $B \to A$ and $(A, \alpha) \sim_p (B, \beta)$. At each stage of this proof, it will be assumed that the $\sim_c$ steps and the $\sim_p$ steps in the sequence from (N_1, α_1) to (N_2, α_2) are rearranged such that there is never a $\sim_c$ immediately before a $\sim_\downarrow$, and there is never a $\sim_c$ immediately after a $\sim_\uparrow$. This rearrangement is always possible, because if there is some subsequence $(A, \alpha) \sim_c (B, \alpha) \sim_\downarrow (C, \beta)$, then there is an alternate path $(A, \alpha) \sim_\downarrow (B', \beta) \sim_c (C, \beta)$, where the reduction and the $\sim_\downarrow$ from A to B' are the same as those from B to C, and the $\sim_c$ step is the same, but with the reduction sequences $O_1 \to^* O_1'$ and $O_2 \to^* O_2'$ in the definition of $\sim_c$ extended by the reduction $B \to C$. With these rearrangements assumed, it is sufficient

to prove that the $\sim_p$ steps can be rearranged so that all of the $\sim_\downarrow$ steps come before all of the $\sim_\uparrow$ steps (possibly introducing some $\sim_c$ steps in the process).

The rearrangement to put all of the $\sim_\downarrow$s before all of the $\sim_\uparrow$s proceeds by induction on the number of $\sim_\downarrow$s that occur after the first occurrence of $\sim_\uparrow$. If it's 0, all of the $\sim_p$s are already in the correct order, and we're done. Otherwise, take the subsequence from the first $\sim_\uparrow$ to the first $\sim_\downarrow$ after that. This must be rearranged to some $\sim_\downarrow$s followed by some $\sim_\uparrow$s, then once that's done, the number of $\sim_\downarrow$s after the first $\sim_\uparrow$ in the overall sequence will have decreased by 1.

Let this subsequence be $A \sim_\uparrow^n B \sim_\downarrow C$. Suppose for induction that $A \sim_\uparrow^k (D,\delta) \sim_\downarrow^* (E,\epsilon) \sim_\uparrow^* C$ for some D,δ,E,ϵ, and $0 \le k \le n$, where in the reduction sequence $D \to^* E$, for any pair of reductions at positions which aren't disjoint, the reduction at the innermost (greater) position occurs first. Reduction sequences with this property will be called "parallel", as it is analagous to parallel reduction as in ⌈TODO: find this reference, for the proof of the Church-Rosser theorem by Tait and Martin-Löf⌋. This induction is in reverse, with k decreasing from n to 0. If $k = n$, simply set $(D,\delta) = B$ and $(E,\epsilon) = C$. If $k = 0$, then A is related to B by a sequence of $\sim_\downarrow$s then $\sim_\uparrow$s, as desired. In the intermediate steps, k must be decreased by one, so a subsequence of one $\sim_\uparrow$ followed by a parallel sequence of $\sim_\downarrow$s must be replaced by a parallel sequence of $\sim_\downarrow$s followed by some $\sim_\uparrow$s.

For any individual $\sim_\uparrow$ then $\sim_\downarrow$ pair, let $(A,\alpha) \sim_\uparrow (B,\beta) \sim_\downarrow (C,\gamma)$. The skeleton B reduces to both A and C. If these are the same reduction, then $(A,\alpha) = (C,\gamma)$ and this subsequence may be removed. If they are different, then the way they overlap corresponds to one of the cases in the definition of $\sim_c$, either case a if the reduction positions are disjoint, or one of the cases b–f if one of the reduction positions is inside the other. In any case, there is a reduction sequence from each of A and C to some common skeleton D (O_1 and O_2 in the definition of $\sim_c$). For the same reason that $(B,\beta) \sim_\downarrow (C,\gamma)$, there is some δ such that $(A,\alpha) \sim_\downarrow^* (D,\delta)$, and similarly for the same reason that $(B,\beta) \sim_\downarrow (A,\alpha)$, for the same δ, $(C,\gamma) \sim_\downarrow^* (D,\delta)$. There are a lot of cases to check for this statement, but all of them are rather simple, and similar to each other. There is therefore an alternative $\sim^*$ sequence from (A,α) to (C,γ), of the form $(A,\alpha) \sim_\downarrow^* (D_1,\delta) \sim_c (D_2,\delta) \sim_\uparrow^* (C,\gamma)$, where D_1 and D_2 are the alternative reduction sequences leading to the same skeleton D. There are only multiple $\sim_\downarrow$ steps in the result if the position of the reduction $B \to C$ is inside the position of the reduction $B \to A$, and similarly, there are only multiple $\sim_\uparrow$ steps in the result if the position of the reduction $B \to A$ is inside the position of the reduction $B \to C$. If there are multiple $\sim_\downarrow$ steps, they are not necessarily parallel, but in the only case where they aren't (case f of sim_c), the subsequence of $\sim_\downarrow$ steps with the reductions at $\alpha;\lambda;@_1;\beta$ then $\alpha;\lambda;@_1;\gamma_i;\mathsf{Y};\lambda;\beta$ for each i where $\gamma_i \ge \beta$ (in the notation of the definition of $\sim_c$) can be replaced by $\sim_\downarrow$s with the reductions at $\alpha;\lambda;@_1;\gamma_i';\mathsf{Y};\lambda;\beta$ then $\alpha;\lambda;@_1;\beta$, then another $\sim_c$, where (γ_i') is the list of positions in A (*before* the reduction at β to A', unlike (γ_i)) where $A|\gamma_i' = x$ and $\gamma_i' > \beta$.

The effect of this rearrangement is (ignoring the $\sim_c$s) to swap the order of

the $\sim_\uparrow$ and the $\sim_\downarrow$, except that if the positions of one of these reductions is inside the other, that inner $\sim_p$ may be duplicated. If the positions are disjoint, both of them are unchanged (corresponding to case a of $\sim_c$), and if one is inside the other, then the outer position is unchanged and the other resultant positions, although they aren't equal to the original inner position, are still inside the outer position. Taking the sub-sequence of one $\sim_\uparrow$ followed by a parallel sequence of $\sim_\downarrow$s, this rearrangement can be repeatedly applied to move the $\sim_\uparrow$ further along the sequence. If at some point a $\sim_\uparrow$ and $\sim_\downarrow$ match (by relating the same (skeleton, position) pairs in opposite directions), they may be removed and the process may stop early. The $\sim_\uparrow$ may be duplicated if it passes a $\sim_\downarrow$ whose position is outside of the $\sim_\uparrow$'s position, but by the assumption that the sequence of $\sim_\downarrow$s is initially parallel, no position inside of this occurs later in the sequence, therefore all of the resultant $\sim_\uparrow$s pass the remaining $\sim_\downarrow$s without changing or duplicating them. The $\sim_\uparrow$s may be further duplicated, but the number of steps left in this process is bounded by the product, for all of the remaining $\sim_\downarrow$s, of 2+the number of occurrences of the variable relevant to the reduction, because for each $\sim_\downarrow$, that is more than the number of duplicates it could produce for any $\sim_\uparrow$ that passes it. It is also possible (earlier) for some of the $\sim_\downarrow$s to be duplicated, but the only case where this process would not terminate, that both the $\sim_\downarrow$s and the $\sim_\uparrow$s continue being duplicated forever so that the number of switches left to do never decreases, is prevented by the parallelness condition, as all of the duplications of $\sim_\downarrow$s occur before all of the duplications of $\sim_\uparrow$s.

It still remains to be shown that the sequence of $\sim_\downarrow$s left at the end of this can be made parallel. The changes that may have occurred since the previous version of this sequence (which was known to be parrallel already), are that some of the $\sim_\downarrow$s may have been duplicated by passing the $\sim_\uparrow$, and if that $\sim_\uparrow$ matches one of the $\sim_\downarrow$s, that $\sim_\downarrow$ is removed. The position of the $\sim_\uparrow$'s reduction is the same for all of these duplications, because it is not changed until later in the sequence, when the $\sim_\downarrow$s with positions outside of that may occur. The resultant positions after a duplication are in the same position relative to each other, and therefore the order of the remaining $\sim_\downarrow$s is automatically parallel, except that (letting the position of the $\sim_\uparrow$'s reduction be α) if there are $\sim_\downarrow$s with reductions at positions $\alpha; @_1; \lambda; \beta$ and $\alpha; @_2; \gamma$, or $\alpha; \mathsf{Y}; \lambda; \beta$ and $\alpha; \mathsf{Y}; \lambda; \gamma$, some of the resultant $\sim_\downarrow$s' positions may be inside each other where they were originally not. Similarly to the case where a Y reduction and a reduction at a position inside it are swapped though, the order can be fixed by swapping some of the $\sim_\downarrow$ steps with each other. In the first case, the reduction at position $\alpha; @_1; \lambda; \beta$ ends up at $\alpha; \beta$, and the reduction at position $\alpha; @_2; \gamma$ ends up at the positions $\alpha; \delta_i; \gamma$, where (δ_i) is the positions of the variable relevant to the $\sim_\uparrow$'s reduction inside its lambda. The positions of the variable in the lambda may be different before the reduction at β, but the positions of the redexes corresponding to the original redex at $\alpha; @_2; \gamma$ are changed similarly. Rather than reducing at $\alpha; \beta$ then at $\alpha; \delta_i; \gamma$, it is therefore possible to reach the same point by reducing at $\alpha; \delta_i'; \gamma$ then at $\alpha; \beta$, where (δ_i') is the positions of the relevant variable within the lambda before the reduction at $\alpha; @_1; \lambda; \beta$ (and again, a $\sim_c$ of some sort must also be introduced). The other case, that the $\sim_\uparrow$'s reduction is a Y-reduction,

is similar. Each reduction which is duplicated has one image at $\alpha; \lambda; @_1$;its original relative position, and one for each occurrence of the variable. Those corresponding to variable occurrences may need to be moved to before those at $\lambda; @_1$.

In summary, to reach the desired order of $\sim_p$ steps, each $\sim_\downarrow$ may have to be moved past some $\sim_\uparrow$ steps, and it may get duplicated in the process, but the resultant sequence of $\sim_\downarrow$s can all be put in parallel order, and it has been shown that moving a $\sim_\uparrow$ past a sequence of $\sim_\downarrow$s is always possible if they're in parallel order, therefore the overall process terminates and reaches a point where all of the $\sim_\downarrow$s precede all of the $\sim_\uparrow$s, and all of the $\sim_c$s are in between them. All of these rearrangements leave the end points of the sequence unchanged, therefore this is an alternative sequence of $\sim$ steps between (N_1, α_1) and (N_2, α_2) of the form $(N_1, \alpha_1) \sim_\downarrow^* (N_1', \alpha') \sim_c^* (N_2', \alpha') \sim_\uparrow^* (N_2, \alpha_2)$. □

If $(X, \alpha) \sim_c (Y, \alpha)$ then the final skeletons in X and Y are equal, and $(X, \beta) \sim_c (Y, \beta)$ for any β that occurs in this skeleton, therefore $\sim_c$ can be considered to apply to reduction sequences even without positions, $X \sim_c Y$ iff $(X, \cdot) \sim_c (Y, \cdot)$. The structure of $\sim_c$ can be seen more clearly by choosing a canonical example from each equivalence class. For this, the member of the class that's in call-by-value order is used. A reduction sequence is in call-by-value order if for every pair of reductions in the sequence at positions α and β, where the reduction at α happens first, either $\alpha \leq \beta$ or α and β are disjoint, with α to the left of β (i.e. the first elements of the sequences α, β where they differ are $@_1$ and $@_2$, $\underline{f}_i$ and $\underline{f}_j$ for $i < j$, or if_i and if_j for $i < j$, in that order). Note that although the reductions that occur in the sequence are in call-by-value order, it may not be the actual call-by-value reduction sequence starting from that point because some redexes may remain un-reduced even when other reductions that should happen afterwards occur. It is only required that the reductions that do occur occur in the correct order.

Lemma 5. *For any reduction sequence X, there is a unique reduction sequence X_{cbv} that is related to X by $\sim_c$ and is in call-by-value order, and if $X \sim_c^* Y$, then $X_{cbv} = Y_{cbv}$.*

⌈As defined, $\sim_c$ is not symmetric.⌋   ◁ L

Proof. Define X_n recursively so that $X_0 = X$ and if X_n is not in CBV order, take the last pair of reductions in X_n which aren't in CBV order. They are adjacent in the sequence, as the CBV order is a total order. Let the positions of these reductions be α and β respectively. If they are disjoint, they can be considered as one of the branches of $\sim_c$ case a. Otherwise, $\beta < \alpha$, and the sub-skeleton at position β at the appropriate point in the reduction sequence is a redex with non-trivial subterms, therefore it's of one of the forms $\text{if}(X > 0, A, B)$, $(\lambda x.A)B$ or $Y\lambda x.A$, therefore the reductions α and β form the O_2 branch of one of the cases b, c, d, e or f of $\sim_c$. In either case, define X_{n+1} as equal to X_n except taking the other (left) branch, so that the order of the reductions at α and β

33

are switched (although the equivalent(s) of the reduction at α may not actually be at that position).

This sequence of swaps rearranges the reductions in X rather like insertion-sort (but in some cases duplicating or deleting the new element), and eventually reaches some X_n that is in CBV order. This is defined to be X_{cbv}.

If $Y \sim_c X \sim_c^* X_{cbv}$, and the reductions involved in $Y \sim_c X$ are not the last pair of reductions in Y that aren't in CBV order, the order of the $\sim_c$s can be rearranged so that the sequence $Y \sim_c^* X_{cbv}$ is the canonical sequence given above, therefore $Y_{cbv} = X_{cbv}$. To be more precise about this rearrangement, consider the sequence of $\sim_c$ steps from Y to X_{cbv}. The first step takes some subsequence of the reduction sequence Y and swaps the order of those reductions. If $X = X_{cbv}$, then $X = Y_1 = Y_{cbv}$ already. Otherwise, consider the cases according to whether the image of this subsequence in X occurs earlier than the next pair of reductions to be swapped (i.e. the last pair of reductions in X that occur not in CBV order). If not, then either the step $Y \sim_c X$ changes this subsequence into CBV order, or out of it. If it is into, then it was the last out-of-order pair in Y therefore the whole sequence $Y \sim_c^* X_{cbv}$ is already in the canonical order and $Y_{cbv} = X_{cbv}$. If it switches the pair to the wrong order, then the result of that switch is the last out-of-order pair in X therefore $Y = X_1$, and the sequence $X_1 \sim_c^* X_{cbv}$ is just a suffix of the sequence $X \sim_c^* X_{cbv}$ therefore $(X_1)_{cbv} = X_{cbv}$ and again $Y_{cbv} = X_{cbv}$.

The remaining cases are when the subsequence of reduction steps involved in $Y \sim_c X$ occur before the last out-of-order pair in X, but possibly overlapping. If there is no overlap, then the $\sim_c$ steps do not interfere with each other at all, and can simply be performed in the other order. This gives a different sequence of $\sim_c$ steps $Y \sim_c Y_1 \sim_c X_1 \sim_c^* X_{cbv}$. As $Y_{cbv} = (Y_1)_{cbv}$, we can proceed by induction in this case and use the fact that $(Y_1)_{cbv} = (X_1)_{cbv} = X_{cbv}$ to acheive the result.

In the other case, there is some overlap between the regions involved in $Y \sim_c X$ and $X \sim_c X_1$, but the last reduction in $X \sim_c X_1$ is later than the last reduction in $Y \sim_c X$. There are only 2 reductions in X involved in $X \sim_c X_1$, therefore the region involved in $Y \sim_c X$ ends on the first of the reductions involved in $X \sim_c X_1$. Now consider the specific sequence of positions of the reductions in X involved in either of these steps. Let the last of these positions be α, the first γ, and the second-last, third-last and so on $\beta_1, \beta_2, \ldots$ respectively, and let the positions of the redexes in Y involved in $Y \sim_c X$ be β then (γ_i) in order.. Then $X \sim_c X_1$ swaps β_1 and α, and α comes before β_1 in CBV order (i.e. $\alpha < \beta_1$ or α is left of β_1), and the step $Y \sim_c X$ either proceeds forwards towards CBV order, swapping β below $\gamma_1 = \gamma$ to form γ then (β_i), or proceeds backwards, taking Y even farther from CBV order, swapping the (γ_i) to above β, forming γ then $\beta_1 = \beta$. In each of the cases below, it is required to show that the canonical sequence $(X_i)_i$ eventually reaches some term which is the same as one in $(Y_j)_j$, and from that point on, they match, therefore they reach the same end result: $X_{cbv} = Y_{cbv}$.

- $Y \sim_c X$ in the backwards direction (so that Y is closer to CBV order than X is): By swapping the roles of X and Y, and β and γ, this is equivalent to the $Y \sim_c X$ forwards case. The assumption that α is before β_1 in CBV order (so that β_1, α is the pair to be swapped in $X \sim_c X_1$) maps to the assumption that α is before γ_1 (so that the result doesn't just follow trivially by Y_1 being equal to X) and vice-versa. Then one of the cases below establishes that $X_i = Y_j$ for some i and j, therefore after swapping back, $Y_i = X_j$.

- $Y \sim_c X$ forwards and α comes after γ in CBV order: In this case, the γ and α reduction steps are already in the correct order in Y, therefore β and γ are already the last out-of-order pair of reductions in Y, $Y_1 = X$, and $Y \sim_c X \sim_c^* X_{cbv}$ is already the canonical sequence from Y therefore $Y_{cbv} = X_{cbv}$.

- $Y \sim_c X$ forwards and α is disjoint from γ and β: In this case, the sequence $(X_i)_i$ starts by swapping α and β_1 by case a, then because γ is disjoint from α and comes after is in CBV order, γ is right of α, therefore either all the remaining β_is, are $\geq \gamma$ (if β and γ aren't disjoint) or there are no remaining β_is (if β and γ are disjoint, in which case they swap by case a and there is only β_1). The sequence $(X_i)_i$ therefore proceeds to swap all the remaining β_is then γ, in order, below α, resulting in X_i for some i having as its relevant sub-reduction-sequence $\alpha, \gamma, \beta_k, \ldots, \beta_1$. The other sequence $(Y_j)_j$ starts by swapping γ then β below α, then swapping γ and β. Because α is disjoint from γ, the reduction there does not affect the sub-skeleton at position γ, therefore this last $\sim_c$ step proceeds identically to how it did in $Y \sim_c X$, resulting in $\gamma, \beta_k, \ldots, \beta_i$. Overall, this results in Y_3 having the relevant sub-reduction-sequence $\alpha, \gamma, \beta_k, \ldots, \beta_1$, therefore $Y_3 = X_i$ for the aforementioned value of i.

- $Y \sim_c X$ forwards, $\alpha \leq \gamma$ and α (and therefore also γ) is disjoint from β: In this case, $(X_i)_i$ proceeds by swapping $\beta_1 = \beta$ below α by case a, then swapping γ below α by one of the other cases, and $(Y_j)_j$ proceeds by swapping γ and α, then swapping β below α then all of the images of γ (which are $\geq \alpha$ and therefore left of β). In both cases, the subterm at α when α and γ are swapped is unaffected by the reduction at β, therefore it produces the same result in both cases, therefore the overall sequence in both cases is α then the images of γ then β.

- Both γ and β are $> \alpha$ but they're disjoint from each other, and either the skeleton at position α is of the form $\mathsf{if}(\mathsf{X} > 0, A, B)$ or it is of the form AB and either both γ and β are $> \alpha; @_1$ or they are both $> \alpha; @_2$: In this case, $(X_i)_i$ proceeds by swapping β then γ below α, in each case possibly forming 0 or multiple images. The only way the number of images of each of these positions can be different is if the subskeleton at α is an if, and there are 0 of one of them and 1 of the other, in which case they're trivially in the correct order already. Otherwise, all of the images of β and

35

γ are disjoint from one another, therefore none of the positions, or their relative order, change when they are swapped, therefore the next part of the sequence $(X_i)_i$ is just insertion-sort running on the images of β and γ with $\sim_c$ case a swaps until they're in the correct order. The sequence $(Y_j)_j$ starts by swapping γ then β below α, then as before, sorting the images into the correct order by case a swaps. The images originally formed of γ and β are the same in both cases, therefore the final order is the same, so $X_i = Y_j$ for some i and j.

- $Y \sim_c X$ forwards, $\alpha; @_1 < \gamma$, and $\alpha; @_2 < \beta$: Let x be the variable involved in the β-reduction at α, so that the sub-skeleton at α is $(\lambda x.A)B$ for some A, B, where $\gamma = \alpha; @_1; \lambda; \gamma'$ and $A \to A'$ at γ', and $\beta = \alpha; @_2; \beta'$, and $B \to B'$ at β'. In this case, $(X_i)_i$ starts by swapping β below α by case e, producing one image of β for each x in A', then swapping γ below α by case d, producing a single image $\alpha; \gamma'$. For each of the instances of x in A left of γ', the reduction at $\alpha; \gamma'$ is then swapped below the corresponding image of β. The sequence $(Y_j)_j$ starts by swapping γ below α then β below α, but in this case the images of β produced may be different. In this case the swap happens earlier in the reduction sequence than γ, so that the body of the lambda is still A rather than A', and the reduction at γ may rearrange or change the number of instances of x. Next, each of the images of β to the right of $\alpha; \gamma'$ are swapped with it by case a. At this point, the next few reductions before $\alpha; \gamma'$ will in general be images of β at positions inside $\alpha; \gamma'$. The subskeleton at position $\alpha; \gamma'$ before these reductions is $(A|\gamma')[B/x]$, then by the images of β this reduces to $(A|\gamma')[B'/x]$, then it reduces at its root position ($\alpha; \gamma'$ in the overall skeleton) to $(A'|\gamma')[B'/x]$. For each position of an x in $(A|\gamma')$, there are 0 or more corresponding positions of x in $(A'|\gamma')$, depending on what type of reduction $A \to A'$ is. Because B and B' cannot contain any instances of the variable involved in the reduction $A \to A'$ (if it's a β-reduction or Y-reduction), each instance of B in $(A|\gamma')[B/x]$ has the same set of images in $(A'|\gamma')[B/x]$ as the corresponding x in $(A|\gamma')$. When these images of β that overlap with the image of γ are swapped below it, each of them has its own images, one at each position of B in $(A'|\gamma')[B/x]$ corresponding to its original copy of B in $(A|\gamma')[B/x]$. After all of them are swapped below $\alpha; \gamma'$ (and rearranged among themselved by case a), there is therefore one reduction at each copy of B in $(A'|\gamma')[B/x]$, i.e. one at position β' relative to each x in $(A'|\gamma')$. They are therefore the same as the images of β in X_1 that overlap with $\alpha; \gamma'$, because those are also one reduction at position β' relative to each x in $(A'|\gamma)$. Both $(X_i)_i$ and $(Y_j)_j$ therefore eventually reach a point where the relevant sub-reduction-sequence is α, the images of β to the left of $\alpha; \gamma'$, $\alpha; \gamma'$ itself, the aforementioned images of β that overlap with $\alpha; \gamma'$, then all the images of β to the right of $\alpha; \gamma'$.

- $Y \sim_c X$ forwards, γ and β are disjoint and both $> \alpha$, and the reduction at α is a Y-reduction: This is similar to the previous case, but more complicated, so it will need some definitions to explain properly what's

going on. Let the term at α before any of the reductions be $\mathsf{Y}\lambda x.A$, let $\gamma = \alpha; \mathsf{Y}; \lambda; \gamma'$, let $\beta = \alpha; \mathsf{Y}; \lambda; \beta'$, let $A|\beta' = B \to B'$ with the reduction at $\cdot$, let $A|\gamma' = C \to C'$ with the reduction at $\cdot$, let the positions in A where x occurs that are left of γ', between γ' and β' (but still disjoint from both), and right of β' respectively be (δ_i^l), (δ_i^m) and (δ_i^r), let the positions in B, C, B' and C' respectively where x occurs be (δ_i^B), (δ_i^C), $(\delta_i^{B'})$ and $(\delta_i^{C'})$ (so that all the positions in A where x occurs, in left-to-right order, are $(\delta_i^l)_i, (\delta_i^C)_i, (\delta_i^m)_i, (\delta_i^B)_i, (\delta_i^r)_i$). The sequence $(X_i)_i$ starts by swapping β then γ below α. With the shorthand that $\gamma'(\epsilon) = \alpha; \lambda; @_1; \epsilon; \mathsf{Y}; \lambda; \gamma'$, $\gamma_0' = \alpha; \lambda; @_1; \gamma'$, and similarly for β', the relevant portion of X_2 is then α, $(\gamma'(\delta_i^l))_i, \gamma_0', (\gamma'(\delta_i^{C'}))_i, (\gamma'(\delta_i^m))_i, (\gamma'(\delta_i^B))_i, (\gamma'(\delta_i^r))_i, (\beta'(\delta_i^l))_i, (\beta'(\delta_i^{C'}))_i$, $(\beta'(\delta_i^m))_i, \beta_0', (\beta'(\delta_i^{B'}))_i, (\beta'(\delta_i^r))_i$. Next in $(X_k)_k$, for each i, $\gamma'(\delta_i^r)$ are swapped past all the images of β until it is immediately before $\beta'(\delta_i^r)$ by case a, then for each i, $\gamma'(\delta_i^B)$ is swapped past all the images of β until β_0' by case a, then swapped with β_0' by one of the other cases depending on what type of reduction $B \to B'$ is and the relative positions. Even if it's case d or f, all of the images of $\gamma'(\delta_i^B)$ are disjoint from each other (and from the positions of the other reductions after β_0') because the redex, C, doesn't contain any instances of the variable involved in the reduction $B \to B'$. Expanding the definitions, this is swapping $\alpha; \lambda; @_1; \delta_i^B; \mathsf{Y}; \lambda; \gamma'$ with $\alpha; \lambda; @_1; \beta'$ at a point in the reduction sequence where the subskeleton at $\alpha; \lambda; @_1; \beta'$ is B with a mixture of $\mathsf{Y}\lambda x.A$ and $\mathsf{Y}\lambda x.A[C'/\gamma']$ substituted for its occurrences of x, and one of these occurrences of x is at δ_i^B. For each δ_j^B, there are some corresponding $\delta_k^{B'}$, where the images of the x at δ_j^B end up after the reduction $B \to B'$. These cover all the $\delta_k^{B'}$, and uniquely, and for each δ_i^B, the images of $\zeta; \delta_i^B; \theta$ after swapping it with $\zeta; \beta'$ are precisely $(\zeta; \delta_k^{B'}; \theta)$ for those same values of k, therefore after swapping all of the $(\gamma'(\delta_i^B))_i$ down past β_0', their images are $(\gamma'(\delta_i^{B'}))_i$. After swapping with β_0', each of these images than swaps by case a to take its place among the $(\beta'(\delta_i^{B'}))_i$. Next up in $(X_i)_i$, the $(\gamma'(\delta_i^m))_i$s and the $(\gamma'(\delta_i^{C'}))_i$s swap down past some images of β they're disjoint from to take their place before their matching image of β, then γ_0' swaps past the $(\beta'(\delta_i^l))_i$, stopping immediately before $\beta'(\delta_1^{C'})$, then the $(\gamma'(\delta_i^l))_i$ and the $(\beta'(\delta_i^l))_i$ mix. The end result of the rearrangement of (this subsequence of) the reduction sequence is therefore α, $(\gamma'(\delta_i^l), \beta'(\delta_i^l))_i$, γ_0', $(\gamma'(\delta_i^{C'}), \beta'(\delta_i^{C'}))_i$, $(\gamma'(\delta_i^m), \beta'(\delta_i^m))_i$, β_0', $(\gamma'(\delta_i^{B'}), \beta'(\delta_i^{B'}))_i$, $(\gamma'(\delta_i^r), \beta'(\delta_i^r))_i$.

The other sequence, $(Y_j)_j$, proceeds similarly, but with the images of γ starting out after the images of β. It swaps all the images of β that are right of γ_0 to the appropriate places by case a, then swaps all the $(\beta'(\delta_i^C)_i$ first to then past γ_0', resulting in $(\beta'(\delta_i^{C'}))_i$ for the same reason as with swapping those images of γ that overlapped with β_0' past it in the $(X_i)_i$ case above, then finally swapping the $(\beta'(\delta_i^l))_i$s and the $(\gamma'(\delta_i^l))_i$s into the correct order. As required, this produces the same result as in $(X_i)_i$.

- $Y \to X$ forwards, and $\alpha < \gamma < \beta$: Let γ' be such that $\gamma = \alpha; \mathsf{if}_i; \gamma'$,

$\alpha; @_1; \lambda; \gamma'$, $\alpha; @_2; \gamma'$ or $\alpha; \mathsf{Y}; \lambda; \gamma'$, so that γ' is the freely varying later part of γ as in the definition of $\sim_c$, whichever case applies to swapping γ and α. Similarly, let β' be the freely varying part of β relative to γ. Let the images of γ after swapping it with α (where this swap occurs later in the reduction sequence than β) then be $(\alpha; \delta_i; \gamma')_i$, and the images of β after swapping it with γ be $(\gamma; \epsilon_i; \beta')_i$. Depending on the skeleton at α, and γ's position within it, $(\delta_i)_i$ may be empty (case b), a singleton containing $\cdot$ (cases c and d), all the positions of the relevant variable in the lambda in the skeleton at α (case e), or $\lambda; @_1$ and $\lambda; @_1; \zeta; \mathsf{Y}; \lambda$ for each position ζ of the relevant variable (case f), but in all cases (except a, which is excluded because none of α, γ and β are disjoint) the set of images has this general structure. Furthermore, the positions $(\delta_i)_i$ are determined only by the general position of γ within α (the part excluded from γ'), and the positions of the relevant variable within the skeleton at α. They do not depend on α or γ, so that if both initial positions were moved somewhere else the relative positions of the images of γ would be unaffected, and similarly if some position within γ were swapped with α, the relative positions of its images would be the same (except to the extent that, in case f, the positions of the relevant variable are taken after the inner reduction takes place, so that some of those may still change). Because there are so many of them and they don't actually affect the multiset of positions where reductions take place, I will be ignoring swaps by case a in the proof of this case, and just assuming that all the final positions end up in the correct order.

The final set of reduction positions that I will be proving both $(X_i)_i$ and $(Y_j)_j$ reach is $\alpha, (\alpha; \delta_i; \gamma', (\alpha; \delta_i; \gamma'; \epsilon_j; \beta')_j)_i$, where as usual the sequences are expanded out to the full list. For the sequence $(X_i)_i$, first the images of β in X_0 are $(\gamma, \epsilon_i, \beta')$. For each of these in turn, it is swapped with α. By the fact mentioned above that sub-positions are mapped to the same set of images except in case f, the images of $\gamma, \epsilon_j, \beta'$ after this swap are $(\alpha; \delta_i^j; \gamma', \epsilon_j, \beta')_i$, where $(\delta_i^j)_i$ is the same as $(\delta_i)_i$ except that in the case where this is not the first image of β swapped below α and the reduction at α is a Y-reduction (and therefore swaps with it proceed by case f), the positions of the relevant variable are taken after the reductions at γ and $\gamma; \epsilon_k; \beta'$ for $k \leq j$ (this implies that the first of these, $\delta_i^n = \delta_i$). All the images $\alpha; \delta_i^j; \gamma'; \epsilon_j; \beta'$ where $\delta_i^j > \lambda; @_1; \gamma'; \epsilon_{j+1}; \beta'$ are then swapped with $\alpha; \lambda; @_1; \gamma'; \epsilon_{j+1}; \beta'$ (which is $\alpha; \delta_i^{j+1}; \gamma'; \epsilon_{j+1}; \beta'$ for some i). As in the case where $\alpha; \mathsf{Y} < \gamma, \beta$ and γ and β are disjoint, the effects of these swaps may duplicate or delete the reductions in such a way that all the remaining images of $\gamma; \epsilon_j; \beta'$ are $(\alpha; \delta_i^{j+1}; \gamma'; \epsilon_j; \beta'$. This is repeated with δ_i^{j+2} and so on until these images are all in the correct order, at which point they are $(\alpha; \delta_i; \gamma'; \epsilon_j; \beta')_i$. After all of these are done, γ is swapped with α, yielding $(\alpha; \delta_i^{-1}; \gamma')_i$. As with the images of β, the subset of these which are inside $\alpha; \lambda; @_1; \gamma'; \epsilon_j; \beta'$ for each j in turn are swapped with it, until they are all in the correct order and the images left of γ are $(\alpha; \delta_i; \gamma')_i$. At this point,

X_i has the desired value.

The sequence $(Y_j)_j$ proceeds similarly. First γ swaps with α, yielding α and $(\alpha; \delta_i; \gamma')_i$, then β swaps with α, yielding one image of β for each δ_i, except that, in case f, those images at positions inside $\alpha; \lambda; @_1; \gamma'$ may differ because they are earlier in the reduction sequence than any of the images of γ. Each of the images of β in turn it moved to the corresponding image of γ, except that those overlapping with $\alpha; \lambda; @_1; \gamma'$ may be duplicated or deleted on the way. The images of β after just this process (which does not actually correspond to any term in $(Y_j)_j$, but it's sufficiently independent that the order doesn't matter that much) are $(\alpha; \delta_i; \gamma'; \beta'')$ (where $\gamma; \beta'' = \beta$). The images of these for each i, after swapping with $\alpha; \delta_i; \gamma'$, is $(\alpha; \delta_i; \gamma'; \epsilon_j; \beta')_j$, because the reductions produced by a swap are unaffected by its overall location, and the skeleton at $\alpha; \delta_i; \gamma'$ at the relevant point in the sequence is equal to the skeleton at γ initially, therefore this is equivalent to immediately swapping β and γ as in $Y \sim_c X$ (except that in the case where the reduction at α is a Y-reduction, the skeleton at $\alpha; \lambda; @_1; \gamma'$ is not actually equal to the skeleton initially at γ because something was substituted in for the variable bound at $\alpha; \mathsf{Y}$, but this doesn't affect the variable involved in the reduction at γ, so this doesn't actually matter). Therefore also in this case, eventually the set of reductions in the relevant portion of Y_j is $\alpha, (\alpha; \delta_i; \gamma', (\alpha; \delta_i; \gamma'; \epsilon_j; \beta')_j)_i$, therefore it is equal to some X_i.

In summary, if the swap in $Y \sim_c X$ doesn't overlap with the swap $X \sim_c X_0$, the sequence of swaps can be rearranged until it does, and in any other case, there is some Y_j that's equal to some X_i, therefore these sequences reach the same end point and if $Y \sim_c X$, then $Y_{cbv} = X_{cbv}$, and by chaining these together, if $Y \sim_c^* X$, $Y_{cbv} = X_{cbv}$.

If Y is in CBV order and $Y \sim_c^* X$, then $Y_{cbv} = X_{cbv}$ but also $Y_{cbv} = Y$ because the sequence $(Y_i)_i$ terminates immediately therefore X_{cbv} is the unique reduction sequence in call-by-value order that's related to X by $\sim_c^*$. □

Lemma 6. *The relation $\sim$ is defined on $L_0(M)$ with reference to a particular starting term M, so different versions, $\sim_M$ and $\sim_N$, can be defined starting at different terms. If $M \to N$, then $\sim_N^*$ is equal to the restriction of $\sim_M^*$ to $L_0(N)$.*

Proof. $\sim_N^*$ is trivially a subset of $\sim_M^*$ because $\sim_N$ is a subset of $\sim_M$.

In the other direction, suppose $(N_1, \alpha_1) \sim_M^* (N_2, \alpha_2)$ where both N_1 and N_2 are descendants of N. By Lemma 4, take $(N_1, \alpha_1) \sim_{p,M}^* (N_1', \alpha') \sim_{c,M}^* (N_2', \alpha') \sim_{p,M}^* (N_2, \alpha_2)$. The $\sim_p^*$ steps remain within $Rch(N)$, and $\sim_p$ does not depend on the history of the reduction sequences, therefore $(N_1, \alpha_1) \sim_{p,N}^* (N_1', \alpha') \sim_{c,M}^* (N_2', \alpha') \sim_{p,N}^* (N_2, \alpha_2)$.

Let X be the call-by-value reduction sequence related to both N_1' and N_2' by $\sim_{c,M}$ given by Lemma 5. As $M \to N$ is the first reduction in the sequence N_1', it is the last to be affected by the $\sim_{c,M}$ sequence $N_1' \sim_{c,M}^* X$ given by Lemma

[5](#) therefore it can be split into $N_1' \sim_{c,M}^* Y_1 \sim_{c,M}^* X$ where Y_1 is in CBV order except possibly for its first reduction, which is still $M \to N$. Let the position of the reduction $M \to N$ be β. ⌈TODO: fill in the details that are missing here. It shouldn't be too hard, it's just a bit sketchy at present.⌋The rearrangement of the reduction sequence $Y_1 \sim_{c,M} X$ consists of moving the reduction at β down past the other reductions in Y_1, possibly duplicating or deleting it in the process, but not affecting the positions or the correct order for any of the other reductions. The reductions derived from $M \to N$ can be identified as follows:

◁ A

A position in some (reduction sequence of) term(s) is *derived from* another if it is related by the reflexive transitive closure of $\rightsquigarrow$, where $(V, \gamma) \rightsquigarrow (W, \delta)$ just if $V \to W$ and one of the following cases holds:

- $(V, \gamma) \sim_p (W, \delta)$

- $\gamma = \delta$, $V \to W$ at ϵ, and $\epsilon > \gamma$

- $V \mid \epsilon = (\lambda x.U)Z, U|\zeta = x, \gamma = \epsilon; @_2; \theta$ and $\delta = \epsilon; \zeta; \theta$

- $V \mid \epsilon = \mathsf{Y}(\lambda x.U), \gamma = \epsilon; \mathsf{Y}; \lambda; \zeta$ and $\delta = \epsilon; \lambda; @_1; \zeta$

- $V \mid \epsilon = \mathsf{Y}(\lambda x.U), U|\zeta = x, \gamma = \epsilon; \theta$ and $\delta = \epsilon; \lambda; @_1; \zeta; \theta$

Crucially, the reductions in X can be partitioned into 2 sets: those at positions derived from (M, β), and those at positions equal to the reductions in Y_1 (in the same order as they occur in Y_1). Using the same construction for N_2' shows that the positions of the reductions in Y_2 are also the positions of the reductions in X other than those derived from (M, β), therefore $Y_1 = Y_2$ therefore $N_1' \sim_{c,M} Y_1 \sim_{c,M} N_2'$ and every reduction sequence in this sequence starts with $M \to N$ therefore $N_1' \sim_{c,N} N_2'$.

Combining this with the $\sim_{p,N}^*$s at the beginning and end then yields the desired result that $(N_1, \alpha_1) \sim_N^* (N_2, \alpha_2)$, therefore the restriction of $\sim_M^*$ to $\sim_N$'s domain $(Rch(N))$ is a subset of $\sim_N^*$ therefore the two versions of $\sim$ match on this domain, as desired. □

At each reduction step $M \to N$, the sample space must be restricted from $I^{L_s(M)}$ to $I^{L_s(N)}$. The injection $L_0(N) \to L_0(M)$ is trivial to define by appending $Sk(M) \to Sk(N)$ to each path, and using Lemma [6](#), this induces a corresponding injection on the quotient, $L(N) \to L(M)$. The corresponding map $L_s(N) \to L_s(M)$ is then denoted $i(M \to N)$.

Unlike in the purely call-by-value case, the version of the reduction relation that takes into account samples is still a general relation rather than a function, so it is denoted "$\Rightarrow$" instead of "red", and it relates $\biguplus_{M \in \Lambda_0} I^{L_s(M)}$ to itself. We write an element of $\biguplus_{M \in \Lambda_0} I^{L_s(M)}$ as (M', s) where the *term $M' \in \Lambda^0$* and $s \in I^{L_s(M')}$.

◷ L

$(M, s) \Rightarrow (N, s \circ i(M \to N))$ if either

$M \to N$ and the redex is not sample, or

$M \mid \alpha = \mathsf{sample}$, $N = M[\underline{s(Sk(M), \alpha)}/\alpha]$ and λ does not occur after $@_2$ or Y in α

Lemma 7. *The relation $\Rightarrow$ is Church-Rosser.*

Proof. Suppose that $(M, s) \Rightarrow^* (M_1, s_1)$ and also $(M, s) \Rightarrow^* (M_2, s_2)$, then it is required to prove that there is some (M', s') such that both $(M_1, s_1) \Rightarrow^* (M', s')$ and $(M_2, s_2) \Rightarrow^* (M', s')$. First consider the special case where $(M, s) \Rightarrow (M_1, s_1)$ and $(M, s) \Rightarrow (M_2, s_2)$, with only a single step in each case. Let the positions of the redexes in $M \to M_1$ and $M \to M_2$ be α_1 and α_2 respectively.

First consider the case that α_1 and α_2 are disjoint. Let $M_1 = M[X_1/\alpha_1]$ and similarly for X_2, then let $M' = M[X_1/\alpha_1][X_2/\alpha_2]$. As the positions are disjoint, the substitutions commute and both M_1 and M_2 reduce (with $\to$) to M'. Let $s' = s \circ i(M \to M_1) \circ i(M_1 \to M')$. The injection $i(M \to M_1) \circ i(M_1 \to M')$ consists of prepending $M \to M_1 \to M'$ to each reduction sequence, but by case a of $\sim_c$, this is equivalent to prepending $M \to M_2 \to M'$, which is $i(M \to M_2) \circ i(M_2 \to M')$. In the case that the reduction $M \to M_2$ isn't a sample-reduction, this is enough to establish that $(M_1, s_1) \Rightarrow (M', s')$ (and similarly if the redex of $M \to M_1$ isn't sample, $(M_2, s_2) \Rightarrow (M', s')$). If it is sample though, in order for it to be the case that $(M_1, s_1) \Rightarrow (M', s')$, it is additionally necessary that X_2, the result of the reduction at α_2, be $s_1(M_1, \alpha_2)$, which follows from the fact that $(M, \alpha_2) \sim (M_1, \alpha_2)$ by case 1 of $\sim_p$. The case that the reduction $M \to M_1$ is a sample-reduction is similar.

The case that $\alpha_1 = \alpha_2$ is trivial, because there is at most one possible $\Rightarrow$ reduction at any given position, therefore $(M_1, s_1) = (M_2, s_2)$ already.

The remaining case is that $\alpha_1 < \alpha_2$ or $\alpha_1 > \alpha_2$. Assume without loss of generality that $\alpha_1 < \alpha_2$. For each possible case of what type of redex $M|\alpha_1$ is, and α_2's position within it, there is a corresponding case of $\sim_c$, and similarly to the case where α_1 and α_2 are disjoint, the term $O_1 = O_2$ from the definition of $\sim_c$ is a suitable value of M'. The term $M \mid \alpha_1$ can't be sample, because it has strict subterms, but the case that $M \mid \alpha_2 = $ sample is still somewhat more complicated. α_2 can't be within $\alpha_1; @_2$ or $\alpha_1; Y$ (cases e and f of $\sim_c$) because those are not valid positions to reduce a sample, and the case that α_2 is within the branch of an if statement that is deleted by the reduction at α_1 (case b) doesn't actually present a problem because there is no corresponding reduction to $M \to M_2$ in the other branch, which leaves cases c and d, that α_2 is in the if$(\cdot, \cdot, \cdot)$ branch that isn't deleted, and that $\alpha_2 < \alpha_1; @_1; \lambda$. The values that the samples take in these cases match because $(M, \alpha) \sim_p (M_1, \alpha; \beta)$ by cases 3 and 2 of $\sim_p$ respectively.

In the case that $(M, s) \Rightarrow^* (M_1, s_1)$ or $(M, s) \Rightarrow^* (M_2, s_2)$ by multiple steps, it is possible to repeatedly replace a pair of the form $A \Leftarrow B \Rightarrow C$ by $A \Rightarrow^* D \Leftarrow^* C$ by the construction above, but it is not immediate that this process terminates, because each $\Rightarrow$ or $\Leftarrow$ may be replaced by multiple, so the sequence of $\Rightarrow$s and $\Leftarrow$s from (M_1, s_1) to (M_2, s_2) may get longer at some steps. However, termination can be proved by noting that the structure of this process is identical to the process of swapping $\sim_\uparrow$ and $\sim_\downarrow$ steps in Lemma 4. To be more precise, in

this case there is a sequence of $\Leftarrow$ and $\Rightarrow$ steps, each of which has an associated reduction position and initial and final terms, and if a $\Leftarrow$ immediately precedes a $\Rightarrow$, they may be swapped to produce some number of $\Rightarrow$s, followed by some number of $\Leftarrow$s. In the case of Lemma 4, there is a sequence of $\sim_\uparrow$ and $\sim_\downarrow$ steps, each of which has an associated reduction position and initial and final skeletons, and if a $\sim_\uparrow$ immediately precedes a $\sim_\downarrow$, they may be swapped to produce some number of $\sim_\downarrow$s followed by some number of $\sim_\uparrow$s. In both cases, the number and reduction positions of the resultant steps are determined by the case of $\sim_c$ that matches the way the initial reduction positions overlap, and the initial skeleton (or the skeleton of the initial term). The same argument that the process in Lemma 4 terminated is therefore applicable here too. At every stage, the sequence of reductions that results from one of the initial $\Rightarrow$s is a parallel sequence, and swapping a parallel sequence of $\Rightarrow$s with a $\Leftarrow$ always terminates, therefore each $\Rightarrow$ in turn can be moved past all of the $\Leftarrow$s, and the process as a whole will terminate in a state where all of the $\Rightarrow$s precede all of the $\Leftarrow$s, i.e. a pair of reduction sequences $(M_1, s_1) \Rightarrow^* (M', s') \Leftarrow^* (M_2, s_2)$. $\square$

Lemma 8. *If A is some descendant of M and $(A, \gamma) \sim^* (M, \delta)$, then $(A, \gamma) \sim_p^* (M, \delta)$, with the length of the reduction sequences decreasing by one each step from A to M.*

Proof. This is a simple induction on $\sim^*$. In the base case, $(A, \gamma) = (M, \delta)$ therefore $(A, \gamma) \sim_p^* (M, \delta)$ trivially. Otherwise, suppose that $(M, \delta) \sim_p^* (B, \epsilon) \sim (A, \gamma)$. Either $(B, \epsilon) \sim_p (A, \gamma)$ or $(B, \epsilon) \sim_c (A, \gamma)$. In the first case, either $B \to A$, in which case the result follows directly, or $A \to B$, in which case the fact that each (reduction sequence of) skeleton(s) has only one parent implies that a $(M, \delta) \sim_p^* (A, \gamma)$ directly as a subsequence of the path to (B, ϵ).

In the $\sim_c$ case, consider the definition of $\sim_c$. Either $O_1' = A$ and $O_2' = B$ or vice-versa. As $M \to^* N \to^* O_1 \to^* B$ and $(M, \delta) \sim_p^* (B, \epsilon)$, and $\sim_p$ only relates positions in a term and its parent, there are some positions ζ, θ such that $(M, \delta) \sim_p^* (N, \zeta) \sim_p^* (O_1, \theta) \sim_p^* (B, \epsilon)$. It follows that $(O_2, \theta) \sim_p^* (A, \gamma)$ by following the same path, therefore it suffices to provide the only missing portion of the path from M to A, i.e. to prove that $(N, \zeta) \sim_p^* (O_2, \theta)$ given $(N, \zeta) \sim_p^* (O_1, \theta)$ (or vice-versa).

If ζ is disjoint from all the positions of reduction from N to O_1 and O_2 (and consequently $\zeta = \theta$), this follows from case 1 of $\sim_p$. Otherwise, this can be proved by taking cases from the definition of $\sim_c$. This is rather long, but all of the cases are similar. The general idea is that the reductions from N to O_1 correspond to the reductions from N to O_2, so that if a position is related by $\sim_p$ across that reduction, it is related in the other branch for the same reason. Case d, where $B = O_1'$ rather than O_2', is given here in more detail as an illustrative example:

Let I be N reduced at α, and J be N reduced at $\alpha; @_1; \lambda; \beta$, so that $N \to I \to O_1$ and $N \to J \to O_2$. All of the reduction positions are $\geq \alpha$, and ζ is not disjoint from all of them, therefore ζ is not disjoint from α. Let ι be the position such that $(N, \zeta) \sim_p (I, \iota) \sim_p (O_1, \theta)$. The fact that

$(N, \zeta) \sim_p (I, \iota)$ implies that $\zeta > \alpha; @_1; \lambda$. Let $\zeta = \alpha; @_1; \lambda; \kappa$ and $\theta = \alpha; \kappa'$. If κ is disjoint from β, then $(N, \zeta) \sim_p (J, \zeta) \sim_p (O_2, \alpha; \kappa') = (O_2; \theta)$ by cases 1 and 2 of $\sim_p$. In the other case, that κ isn't disjoint from β, $\kappa > \beta$ because none of the positions $<= \alpha; @_1; \lambda; \beta$ in N are related to any position in I by $\sim_p$. As $(I, \alpha; \kappa) \sim_p (O_1, \alpha; \kappa')$ (with the redex at $\alpha; \beta$), for exactly the same reason $(N, \alpha; @_1; \lambda; \kappa) \sim_p (J, \alpha; @_1; \lambda; \kappa')$ (with the redex at $\alpha; @_1; \lambda; \beta$). Because $(N, \alpha; @_1; \lambda; \kappa) \sim_p (J, \alpha; @_1; \lambda; \kappa')$, $N|\alpha; @_1; \lambda; \kappa = J|\alpha; @_1; \lambda; \kappa'$, and $N|\alpha; @_1; \lambda; \kappa \neq$ the variable of $N|\alpha; @_1$, therefore $J|\alpha; @_1; \lambda; \kappa'$ is also not the variable therefore $(J, \alpha; @_1; \lambda; \kappa') \sim_p (O_2, \alpha; \kappa') = (O_2, \theta)$ by case 2 of $\sim_p$. Combining these results, $(N, \zeta) \sim_p^* (O_2; \theta)$ as desired. $\qquad \square$

Lemma 9. *If $M \to N$, with the redex at position α, then no position in any term reachable from N is related by $\sim^*$ to (M, α).*

Proof. Suppose on the contrary that $(M, \alpha) \sim^* (A, \beta)$, where $N \to^* A$, then by Lemma 8, $(M, \alpha) \sim_p^* (A, \beta)$. $M \neq A$ therefore $(M, \alpha) \sim_p$ some position in N, but in all the cases of the definition of $\sim_p$, no position in the child term is related to the position of the redex. $\qquad \square$

The reduction relation $\Rightarrow$ is nondeterministic, so it admits multiple possible reduction strategies. A *reduction strategy* starting from a closed term M is a partial function f from $Rch(M)$ to positions, such that for any reachable term N where f is defined, $f(N)$ is a position of a redex in N, and if $f(N)$ is not defined, N is a value. Using a reduction strategy f, a subset of $\Rightarrow$ that isn't nondeterministic, $\Rightarrow_f$, can be defined by $(N, s) \Rightarrow_f (N', s')$ just if $(N, s) \Rightarrow (N', s')$ and N reduces to N' with the redex at $f(N)$.

Essentially what Lemma 9 demonstrates is that the samples taken during any reduction sequence are independent of each other. This is made more precise in the following lemmas.

Lemma 10. *For any skeletons $M \to N$, with the redex at position α, and measurable set of samples $S \subset I^{L_s(N)}$, $\mu(S) = \mu(\{s \in I^{L_s(M)} \mid s \circ i(M \to N) \in S\})$, and furthermore, if $M|\alpha = \mathsf{sample}$, for any $S \subset I \times I^{L_s(N)}$, $\mu(S) = \mu(\{s \in I^{L_s(M)} \mid (s(M, \alpha), s \circ i(M \to N)) \in S\})$.*

Proof. If the result holds for all sets S of the form $\{s \in I^{L_s(N)} \mid \forall j : s(j) \in x_i\}$, where $(x_j)_{j \in J}$ is a family of measurable subsets of I indexed by some finite set $J \subset L_s(N)$ of positions, it also holds for all other S by taking limits and disjoint unions.

Take such a $(x_j)_{j \in J}$, where $x_j = x_k$ and define $K = i(M \to N)[J]$. Because $i(M \to N)$ is injective, it defines a bijection between J and K. We can then calculate the measure $\mu(\{s \in I^{L_s(M)} \mid s \circ i(M \to N) \in S\}) = \mu(\{s \in I^{L_s(M)} \mid \forall k : s(k) \in x_{i(M \to N)^{-1}(k)}\}) = \prod_{k \in K} \mu_I(x_{i(M \to N)^{-1}(k)}) = \prod_{j \in J} \mu_I(x_j) = \mu(S)$.

In the case that $M|\alpha = \mathsf{sample}$, we can similarly consider only those sets S of the form $x_{\mathsf{sample}} \times \{s \in I^{L_s(N)} \mid \forall j : s(j) \in x_i\}$. The position α in M is not related to any position in $L_s(N)$, therefore $(M, \alpha) \notin K$. Again, we can calculate the measure $\mu(\{s \in I^{L_s(M)} \mid (s(M, \alpha), s \circ i(M \to N)) \in$

$S\}) = \mu(\{s \in I^{L_s(M)} \mid s(M,\alpha) \in x_{\mathsf{sample}}, \forall k : s(k) \in x_{i(M\to N)^{-1}(k)}\}) = \mu_I(x_{\mathsf{sample}}) \prod_{k\in K} \mu_I(x_{i(M\to N)^{-1}(k)}) = \mu_I(x_{\mathsf{sample}}) \prod_{j\in J} \mu_I(x_j) = \mu(S)$. $\qquad\square$

Lemma 11. *For any initial term M, reduction strategy f on M, natural number n, skeleton N with k holes, measurable set $T \subset \mathbb{R}^k$ and measurable set S of samples in $I^{L_s(N)}$, $\mu(\{s \in I^{L_s(M)} \mid \exists r \in T, s' \in S : (M,s) \Rightarrow_f^n (N[r], s')\}) = \mu(\{s \in I^{L_s(M)} \mid \exists r \in T, s' \in I^{L_s(N)} : (M,s) \Rightarrow_f^n (N[r], s')\})\mu(S)$*

Proof. Suppose, to begin with, that $n = 0$. Either $\exists r \in T : N[r] = M$, in which case both sides of the equation are $\mu(S)$, or there is no such r, in which case both sides of the equation are 0.

For $n > 0$, suppose for induction that the lemma is true for $n - 1$, for all N, T and S. If $\exists r \in T, s' \in S : (M,s) \Rightarrow_f^n (N[r], s')$, the r and s' are necessarily unique, therefore we may assume that T is the product of k measurable setsets of $\mathbb{R}$, $(T_j)_{0\le j<k}$, as all of the other cases follow by taking unions and limits of these rectangles. Let

$$R_s = \{\text{the skeleton of } O \mid O \in Rch(M), r \in \mathbb{R}^k, O|f(O) = \mathsf{sample}, O \to N[r]\}$$

$$R_d = \{\text{the skeleton of } O \mid O \in Rch(M), r \in \mathbb{R}^k, O|f(O) \neq \mathsf{sample}, O \to N[r]\}.$$

Together, R_s and R_d contain all of the skeletons of terms that can reduce to N, and they're disjoint. For each of these skeletons, there is a map g_O from the reals in O to the reals in N, i.e. $g_O(r) = r'$ if $O[r] \to N[r']$ for $O \in R_d$, and $g_O(s,r) = r'$ if $O[r] \to N[r']$ with s as the value the sample takes in the reduction. All of these functions are measurable, as they are just rearrangements of vector components, or in the case that $O|f(O) = \underline{x}(\mathsf{X}, \ldots, \mathsf{X})$, a combination fo the measurable function x with a rearrangement of vector components. In the case that $O \in R_s$, the function g_O simply inserts the sample into the vector of reals at a certain index, call it h_O.

$$\mu(\{s \in I^{L_s(M)} \mid \exists r \in T, s' \in S : (M,s) \Rightarrow_f^n (N[r], s')\})$$
$$=\mu(\{s \in I^{L_s(M)} \mid \exists O \in R_s, r \in \mathbb{R}^{k-1}, s' \in I^{L_s(O)}, r' \in T, s'' \in S :$$
$$(M,s) \Rightarrow_f^{n-1} (O[r], s') \Rightarrow_f (N[r'], s'')\})$$
$$+\mu(\{s \in I^{L_s(M)} \mid \exists O \in R_d, r \in \bigcup_l \mathbb{R}^l, s' \in I^{L_s(O)}, r' \in T, s'' \in S :$$
$$(M,s) \Rightarrow_f^{n-1} (O[r], s') \Rightarrow_f (N[r'], s'')\})$$

$$= \sum_{O \in R_s} \mu(\{s \in I^{L_s(M)} \mid \exists r \in \mathbb{R}^{k-1}, s' \in I^{L_s(O)} :$$

$$g_O(r, s'(O, f(O))) \in T, s' \circ i(O[r] \to N[g_O(r, s'(O, f(O)))]) \in S, (M, s) \Rightarrow_f^{n-1} (O[r], s')\})$$

$$+ \sum_{O \in R_d} \mu(\{s \in I^{L_s(M)} \mid \exists r \in \bigcup_l \mathbb{R}^l, s' \in I^{L_s(O)} :$$

$$g_O(r) \in T, s' \circ i(O[r] \to N[g_O(r)]) \in S, (M, s) \Rightarrow_f^{n-1} (O[r], s')\})$$

$$= \sum_{O \in R_s} \mu(\{s \in I^{L_s(M)} \mid \exists r \in \prod_{j \neq h(O)} T_j, s' \in I^{L_s(O)} :$$

$$s'(O, f(O)) \in T_{h(O)}, s' \circ i(O[r] \to N[g_O(r, s'(O, f(O)))]) \in S, (M, s) \Rightarrow_f^{n-1} (O[r], s')\})$$

$$+ \sum_{O \in R_d} \mu(\{s \in I^{L_s(M)} \mid \exists r \in \bigcup_l \mathbb{R}^l, s' \in I^{L_s(O)} :$$

$$g_O(r) \in T, s' \circ i(O[r] \to N[g_O(r)]) \in S, (M, s) \Rightarrow_f^{n-1} (O[r], s')\})$$

$$= \sum_{O \in R_s} \mu(\{s \in I^{L_s(M)} \mid \exists r \in \prod_{j \neq h(O)} T_j, s' \in I^{L_s(O)} : (M, s) \Rightarrow_f^{n-1} (O[r], s')\})$$

$$\mu(\{s' \in I^{L_s(O)} \mid s'(O, f(O)) \in T_{h(O)}, s' \circ i(O[r] \to N[g_O(r, s'(O, f(O)))]) \in S\})$$

$$+ \sum_{O \in R_d} \mu(\{s \in I^{L_s(M)} \mid \exists r \in \bigcup_l \mathbb{R}^l, s' \in I^{L_s(O)} : g_O(r) \in T, (M, s) \Rightarrow_f^{n-1} (O[r], s')\})$$

$$\mu(s' \in I^{L_s(O)} \mid s' \circ i(O[r] \to N[g_O(r)]) \in S)$$

$$= \sum_{O \in R_s} \mu(\{s \in I^{L_s(M)} \mid \exists r \in \prod_{j \neq h(O)} T_j, s' \in I^{L_s(O)} : (M, s) \Rightarrow_f^{n-1} (O[r], s')\})\mu(T_{h(O)})\mu(S)$$

$$+ \sum_{O \in R_d} \mu(\{s \in I^{L_s(M)} \mid \exists r \in \bigcup_l \mathbb{R}^l, s' \in I^{L_s(O)} : g_O(r) \in T, (M, s) \Rightarrow_f^{n-1} (O[r], s')\})\mu(S)$$

$$= \sum_{O \in R_s} \mu(\{s \in I^{L_s(M)} \mid \exists r \in \prod_{j \neq h(O)} T_j, s' \in I^{L_s(O)} : s'(O, f(O)) \in T_{h(O)}, (M, s) \Rightarrow_f^{n-1} (O[r], s')\})\mu(S)$$

$$+ \sum_{O \in R_d} \mu(\{s \in I^{L_s(M)} \mid \exists r \in \bigcup_l \mathbb{R}^l, s' \in I^{L_s(O)} : g_O(r) \in T, (M, s) \Rightarrow_f^{n-1} (O[r], s')\})\mu(S)$$

$$= \sum_{O \in R_s} \mu(\{s \in I^{L_s(M)} \mid \exists r \in \mathbb{R}^{k-1}, s' \in I^{L_s(O)} : g_O(r, s'(O, f(O))) \in T, (M, s) \Rightarrow_f^{n-1} (O[r], s')\})\mu(S)$$

$$+ \sum_{O \in R_d} \mu(\{s \in I^{L_s(M)} \mid \exists r \in \bigcup_l \mathbb{R}^l, s' \in I^{L_s(O)} : g_O(r) \in T, (M, s) \Rightarrow_f^{n-1} (O[r], s')\})\mu(S)$$

$$= \mu(\{s \in I^{L_s(M)} \mid \exists O \in R_s, r \in \mathbb{R}^{k-1}, s' \in I^{L_s(O)}, r' \in T, s'' \in I^{L_s(N)} :$$

$$(M, s) \Rightarrow_f^{n-1} (O[r], s') \Rightarrow_f (N[r'], s'')\})\mu(S)$$

$$+ \mu(\{s \in I^{L_s(M)} \mid \exists O \in R_d, r \in \bigcup_l \mathbb{R}^l, s' \in I^{L_s(O)}, r' \in T, s'' \in I^{L_s(N)} :$$

$$(M, s) \Rightarrow_f^{n-1} (O[r], s') \Rightarrow_f (N[r'], s'')\})\mu(S)$$

$$= \mu(\{s \in I^{L_s(M)} \mid \exists r \in T, s' \in I^{L_s(N)} : (M, s) \Rightarrow_f^n (N[r], s')\})\mu(S)$$

$\square$

The usual call-by-value semantics can be implemented as one of these reduction strategies, given by (with V a value and T a term that isn't a value and M a general term)

$$\mathrm{cbv}(TM) = @_1; \mathrm{cbv}(T)$$
$$\mathrm{cbv}(VT) = @_2; \mathrm{cbv}(T)$$
$$\mathrm{cbv}(\underline{f}(V_1, \ldots, V_{k-1}, T, M_{k+1}, \ldots, M_n)) = \underline{f}_k; \mathrm{cbv}(T)$$
$$\mathrm{cbv}(\mathsf{Y}T) = \mathsf{Y}; \mathrm{cbv}(T)$$
$$\mathrm{cbv}(\mathsf{if}(T < 0, M_1, M_2)) = \mathsf{if}_1; \mathrm{cbv}(T)$$
$$\mathrm{cbv}(\mathsf{score}(T)) = \mathsf{score}; \mathrm{cbv}(T)$$
$$\mathrm{cbv}(V) \text{ is undefined}$$
$$\mathrm{cbv}(T) = \cdot \text{ otherwise}$$

(this last case covers redexes at the root position).

A closed term M terminates with a given reduction strategy f and samples s if there is some natural number n such that $(M, s) \Rightarrow_f^n (N, s')$ where f gives no reduction at N. The term terminates almost surely with respect to f if it terminates with f for almost all s.

Theorem 11. *A closed term M is AST with respect to* cbv *iff it is AST.*

Proof. ⌈I'm not sure whether this proof is sufficiently detailed. It would be pretty easy to fill it in more, just very verbose.⌋ ◁ A

For both red and $\Rightarrow_{\mathrm{cbv}}$, the probability measure on the sample space can be used to define a measure on the reduction sequences of each finite length. The distributions are equal for each length, by induction, as follows.

The base case, length 0, is trivial, because in both cases the distribution has probability 1 on the sequence containing only M.

For the inductive case, suppose that the distributions of reduction sequences of length n are equal. ⌈TODO: mention the case that it has already terminated by step n, or is about to. The two versions of the semantics handle it a little differently.⌋ ◁ A

For any term N which is not a value, there is a unique environment E and redex R such that $N = E[R]$. The position of the hole in E is equal to $\mathrm{cbv}(N)$, so that $N \mid \mathrm{cbv}(N) = R$, as the cases in the definition of cbv match the cases in the definition of environments. If $R \neq \mathsf{sample}$, let R' be the result of reducing R (which is equal in both versions of the semantics), then $E[R'] = N[R' / \mathrm{cbv}(N)]$ therefore $\mathrm{red}(N, s) = (E[R'], s)$ and $(N, s) \Rightarrow_{\mathrm{cbv}} (E[R'], s)$. The distribution of next terms, conditional on the previous term, in the case that that previous term reduces deterministically, is therefore equal for red and $\Rightarrow_{\mathrm{cbv}}$.

In the other case that $R = \mathsf{sample}$, the first (term) part of $\mathrm{red}(N, s)$ is $E[\underline{\pi_h(s)}]$. If k is the number of samples already taken in this reduction sequence, this is equivalent to $E[\underline{s_0(k)}]$, where s_0 is the initial sample. For any sufficiently

small neighbourhood of N (i.e. containing only reduction sequences with the same skeletons), the probability of reaching this neighbourhood is independent of $s(k)$, because it depends only on the samples taken so far, and the samples are independent. The distribution of $s(k)$ conditional on reaching N is therefore the uniform distribution on I. Similarly for the other semantics, if the distribution of remaining samples after n steps is independent of the term by Lemma 11, therefore the distribution of $s(Sk(N), \mathrm{cbv}(N))$ conditional on N is the uniform distribution on I. In either case, the distribution of the next term conditional on the previous term, in the case that it reduces randomly, is equal to the image under $r \mapsto E[\underline{r}]$ of the uniform distribution on I.

In any case, the distribution of reduction sequences after $n + 1$ steps, conditional on the reduction sequence after n steps, as defined by $\Rightarrow_{\mathrm{cbv}}$ and by red, is equal, any by the inductive hypothesis the distributions after n steps are equal, therefore by integrating over the reduction sequence up to step n, the distributions on reduction sequences up to step $n + 1$ are equal.

The distributions on reduction sequences of any finite length as defined by $\Rightarrow_{\mathrm{cbv}}$ and red are therefore equal, therefore so is the probability of having reached a value after n steps, therefore so is its limit, the probability of termination, therefore the probability of termination is 1 iff the probability of termination with respect to cbv is 1. $\qquad\square$

Theorem 12. *If M terminates with some reduction strategy f and samples s, it terminates with* cbv *and s.*

Proof. Suppose M terminates with the reduction strategy f and samples s. Let $(M, s) = (M_0, s_0) \Rightarrow_f \cdots \Rightarrow_f (M_n, s_n)$, where M_n is a value. By Lemma 5, the reduction sequence $(Sk(M_i))_i$ is related by $\sim_c^*$ to some skeletal reduction sequence X which is in call-by-value order. For each skeletal reduction sequence X_j in the sequence $(Sk(M_i))_i \sim_c \cdots \sim_c X$, it is possible to define a reduction sequence of terms such that $(M, s) = (N_{j,0}, s_{j,0}) \Rightarrow \cdots \Rightarrow (N_{j,n_j}, s_{j,n_j})$ and $Sk(N_{j,k}) = X_{j,k}$, by at each step applying the reduction at the same place as in X_j, essentially just filling in the reals in the skeletal reduction sequence to make it a reduction sequence of terms (it will be shown shortly that the correct branch is taken in each if-reduction). Assume for induction on j that $(N_{j,n_j}, s_{j,n_j}) = (M_n, s_n)$. To prove that this holds true for $j + 1$ as well, it suffices to show that the corresponding term and samples immediately after the reductions involved in this $\sim_c$ step match, as from that point on, the $\Rightarrow$-sequences will match, as they start from the same point and have all the same reduction positions. Letting the term and remaining samples immediately before the reductions involved in this $\sim_c$ step be (N, t) (which is the same in both N_j and N_{j+1} for the same reason again, they start at the same point and the sequences of reduction positions up to this point are the same), and the terms and samples after these reductions be (O_1, t_1) and (O_2, t_2), so that these match the N, O_1 and O_2 in the definition of $\sim_c$, $t_1 = t_2$ because both of them are, by Lemma 6, equal to the restriction of t to the subset of $Rch(N)$ that starts with the reduction sequences to O_1 and O_2 respectively, and the potential positions

in those subsets of $Rch(N)$ are identified with each other by $\sim$ because of the same case of $\sim_c$ that means N_j and N_{j+1} are related. If none of the relevant reductions are sample-reductions, $O_1 = O_2$ trivially. If there are samples taken (which must be the arbitrary reduction at β), it is also required that the same samples are taken in both branches. Similarly to Lemma 7, the positions of these sample-reductions in each branch are related by $\sim_p^*$, therefore they correspond to the same sample in t, therefore $O_1 = O_2$. The fact that the terms are equal also implies that the same branch of an if statements will be taken.

There is therefore a reduction sequence $(M, s) = (N_0, s_0') \Rightarrow \cdots \Rightarrow (N_m, s_m')$ which ends in a value ($N_m = M_n$) and is in CBV order. It doesn't immediately follow that this is the $\Rightarrow_{\mathrm{cbv}}$ reduction sequence starting from (M, s) because the fact that $(N_i)_i$ is in CBV order only means that the reductions that do take place in this reduction sequence are in the correct order, not that the reduction sequence finishes once it reaches a value, or that there are no other reductions that would come earlier in CBV order that don't occur at all in $(N_i)_i$. Suppose that at some point j in the reduction sequence, the reduction that occurs is not the next reduction in CBV order, but something else, and that the reduction that should be next is at position $\alpha = \mathrm{cbv}(N_i)$, and the position of the actual reduction $N_j \to N_{j+1}$ is β. Suppose further that N_j is not a value. Either $\alpha = \cdot$, the root position, or the reduction at the root position is dependent in some way on the reduction at α (e.g. $\alpha = \mathrm{if}_1$, or $\alpha = \underline{+}; @_1$). In any case, all the reductions in the rest of the reduction sequence are after α in CBV order, therefore none of them are $\leq \alpha$, and there continues to be an unreduced redex at α, therefore the reduction at the root position never occurs, therefore N_m is not a value, which is a contradiction, therefore at the point (if any) where $(N_i)_i$ departs from the $\Rightarrow_{\mathrm{cbv}}$ reduction sequence starting at (M, s), it has already reached a value, therefore $(M, s) \Rightarrow_{\mathrm{cbv}} \ldots$ does reach a value eventually, i.e. M terminates with the reduction strategy cbv and samples s. $\qquad\square$

Corollary 1. *If M is AST with respect to any reduction strategy, it is AST.*

Proof. Suppose M is AST with respect to f. Let the set of samples with which it terminates with this reduction strategy be X. By Theorem 12, M also terminates with cbv and every element of X, and X has measure 1, by assumption, therefore M is AST with respect to cbv therefore by Theorem 11 it is AST. $\quad\square$

All of the theorems on the termination of rankable terms therefore extend to other reduction strategies too. The proofs of Theorems 3, 5, 7 and 8 are all sufficiently generic with respect to what the reduction relation actually is that they can be directly applied to other reduction strategies. They only require that the number of reductions that can occur without any of them being a Y-reduction is bounded for any starting term (which is true, because Theorem 10 applies equally to any reduction strategy). For a reduction strategy r and term M, just substitute $N[r(N)]$ being a Y-redex for N being of the form $E[Y\lambda x.O]$, and $r(N)$ being undefined for N being a value.

The domain of definition of the ranking functions also needs to be changed from the reachable terms $Rch(M)$ to the *reachable terms with respect to the reduction strategy r*,

$$Rch_r(M) = \{N \mid \exists n, N_0, \ldots, N_n : N_0 = M, N_n = N, N_i \to N_{i+1} \text{ at } r(N_i)\}.$$

To be more explicit, the modified forms of the theorems are:

Definition 7. A *ranking function on M with respect to a reduction strategy r* is a measurable function $f : Rch_r(M) \to \mathbb{R}$ such that $f(N) \geq 0$ for all N, and

- $f(N) \geq 1 + f(N')$ if $N[r(N)]$ is a Y-redex and N reduces to N' at $r(N)$

- $f(N) \geq \int_I f(N[\underline{x}/r(N)]) \operatorname{Leb}(\mathrm{d}x)$ if $N[r(N)] = \mathsf{sample}$

- $f(N) \geq f(N')$ if $r(N)$ is any other redex, where $N \to N'$ at $r(N)$.

Any closed term for which a ranking function with respect to a reduction strategy r exists is called *rankable* with respect to r.

Definition 8. A *partial ranking function on M with respect to a reduction strategy r* is a partial function $f : Rch_r(M) \rightharpoonup \mathbb{R}$ such that

- $f(N) \geq 0$ for all N where f is defined.

- f is defined at M.

- For any N in the domain of definition of f, evaluation of N at the positions specified by r will eventually reach some O such that either $r(O)$ isn't defined or $f(O)$ is, and $f(N) \geq \mathbb{E}[f(O) + \text{ the number of Y-reduction steps from } N \text{ to } O]$ (where $f(O)$ is taken to be 0 if O is a value outside of the domain of f).

Definition 9. An *antitone ranking function on M with respect to a reduction strategy r* is a measurable function $f : Rch_r(M) \to \mathbb{R}$ such that $f(N) \geq 0$ for all N, and there exists an antitone function $\epsilon : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ such that

- $f(N) \geq \epsilon(f(N)) + f(N')$ if $N[r(N)]$ is a Y-redex and N reduces to N' at $r(N)$

- $f(N) \geq \int_I f(N[\underline{x}/r(N)]) \operatorname{Leb}(\mathrm{d}x)$ if $N[r(N)] = \mathsf{sample}$

- $f(N) \geq f(N')$ if $r(N)$ is any other redex, where $N \to N'$ at $r(N)$.

Any closed term for which an antitone ranking function with respect to a reduction strategy r exists is called *antitone rankable* with respect to r.

Definition 10. An *antitone partial ranking function on M with respect to a reduction strategy r* is a partial function $f : Rch_r(M) \rightharpoonup \mathbb{R}$ such that there exists some antitone function $\epsilon : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ such that

- $f(N) \geq 0$ for all N where f is defined.

- f is defined at M.

- For any N in the domain of definition of f, evaluation of N at the positions specified by r will eventually reach some O such that either $r(O)$ isn't defined or $f(O)$ is, and $f(N) \geq \mathbb{E}[f(O) + \epsilon(f(O)) \times$ the number of Y-reduction steps from N to $O]$ (where $f(O)$ is taken to be 0 if O is a value outside of the domain of f).

Theorem 13. *If a closed SPCF term M has a ranking function, partial ranking function, antitone ranking function or antitone partial ranking function with respect to a reduction strategy r, then M is AST with respect to r, and AST.*

⌈Give an example or two of a reduction strategy (for $\Rightarrow$) that is not cbv.⌋ ◁ L
⌈NOTE. For the standard CBV semantics, this example is indeed not a problem. ◁ A
In the alternative semantics, defining

$$A[x] = \mathsf{if}(\mathsf{if}(x > 0, I, I)\,\mathsf{sample} - 0.5 > 0, 0, \Omega)$$
$$B = \mathsf{if}(\mathsf{sample} - 0.5 > 0, 0, \Omega)$$

The set of samples where it terminates is

$$\bigcup\nolimits_{r \in [0,1]} \{ s \in I^{L_s(A[\mathsf{sample}])} \mid s([A[\mathsf{sample}]], \mathsf{if}_1; \underline{}_1; @_1; \mathsf{if}_1) = r,$$
$$s([A[\mathsf{sample}], A[r], \dots, B, \mathsf{if}_1) > 0.5 \}.$$

This is a rather unwieldy expression, but the crucial part is that r occurs twice in the conditions on s: once as the value a sample must take, and once in the location of a sample. It is this set that's unmeasurable.⌋

References

⌈22 Jan: ◁ L
Further directions   An obvious next step is to extend the result to the score construct.

The following are highly topical, and could form the basis of an interesting and novel DPhil thesis.

1. Devise methods for proving (positive) a.s. termination.

   - For example, develop a type system satisfying the property: if a term is typable then it is (positively) a.s. terminating. See [**??**].

   - Another approach is to develop algorithms that synthesise ranking supermartingales, following, for example, [**?**].

2. Develop principles (e.g. in the form of "proof rules") for reasoning about (positively) a.s. termination, in the style of [**?**].

3. Design algorithms that synthesise probabilistic invariants (*qua* martingales), à la **?**; see also [**?**].

⌋