Bayesian inference for constraint-based phonology

Colin Wilson
Department of Cognitive Science
Johns Hopkins University

University of Massachusetts, Amherst
December 2, 2011

Outline

1. What is the *objective* of early learning of constraint-based phonology, and what prior assumptions are needed to (approximately) achieve that objective?
   - — restrictiveness and a rich prior (Tesar & Smolensky 1998, 2000; Prince & Tesar 2004, Hayes 2004, . . .)
   - \+ posterior sampling and an unbiased/uninformative prior

2. Gibbs algorithm for early phonological learning: a general Bayesian inference method applied to noisy Harmonic Grammar (e.g., Boersma & Pater 2008; Jesney & Tessier 2009)
   - observe positive data (surface forms) $\mathbf{y}$
   - iteratively sample *grammar* $\bar{\mathbf{w}}$, *weights* $\mathbf{W}$, and *inputs* $\mathbf{x}$
   - → guaranteed to eventually converge to the joint posterior distribution

3. Sample simulations and discussion of extensions
   - Pseudo-Korean, AZBA, and friends
   - hidden structure, paradigm learning, noisy OT, child production grammar

Outline

1. What is the *objective* of early learning of constraint-based phonology, and what prior assumptions are needed to (approximately) achieve that objective?

   − restrictiveness and a rich prior (Tesar & Smolensky 1998, 2000; Prince & Tesar 2004, Hayes 2004, . . .)

   \+ posterior sampling and an unbiased/uninformative prior

2. Gibbs algorithm for early phonological learning: a general Bayesian inference method applied to noisy Harmonic Grammar (e.g., Boersma & Pater 2008; Jesney & Tessier 2009)

   - observe positive data (surface forms) $\mathbf{y}$
   - iteratively sample *grammar* $\bar{\mathbf{w}}$, *weights* $\mathbf{W}$, and *inputs* $\mathbf{x}$

   → guaranteed to eventually converge to the joint posterior distribution

3. Sample simulations and discussion of extensions

   - Pseudo-Korean, AZBA, and friends
   - hidden structure, paradigm learning, noisy OT, child production grammar

Outline

1. What is the *objective* of early learning of constraint-based phonology, and what prior assumptions are needed to (approximately) achieve that objective?

   − restrictiveness and a rich prior (Tesar & Smolensky 1998, 2000; Prince & Tesar 2004, Hayes 2004, . . .)

   \+ posterior sampling and an unbiased/uninformative prior

2. Gibbs algorithm for early phonological learning: a general Bayesian inference method applied to noisy Harmonic Grammar (e.g., Boersma & Pater 2008; Jesney & Tessier 2009)

   - observe positive data (surface forms) $\mathbf{y}$
   - iteratively sample *grammar* $\bar{\mathbf{w}}$, *weights* $\mathbf{W}$, and *inputs* $\mathbf{x}$

   → guaranteed to eventually converge to the joint posterior distribution

3. Sample simulations and discussion of extensions

   - Pseudo-Korean, AZBA, and friends
   - hidden structure, paradigm learning, noisy OT, child production grammar

Early phonological learning

Infants 8-10m distinguish structures that are phonotactically legal (or common) in their native language from structures that are phonotactically illegal (or rare)

(e.g., Jusczyk et al. 1993ab, 1994; Friederici & Wessels 1993; Kajikawa et al. 2006; reviewed in Jusczyk 1997, Hayes 2004)

Ex. English syllable-/word-initial [sk], [st] vs. $^{??}$[vl], *[kn], *[zw]

Following Hayes (2004) and others, interpret this as evidence for acquisition of a constraint-based phonological perception or 'receptive' grammar

- from observation of legal examples (positive evidence)
- before development of the ability to reproduce these examples

Early phonological learning

Infants 8-10m distinguish structures that are phonotactically legal (or common) in their native language from structures that are phonotactically illegal (or rare)

(e.g., Jusczyk et al. 1993ab, 1994; Friederici & Wessels 1993; Kajikawa et al. 2006; reviewed in Jusczyk 1997, Hayes 2004)

Ex. English syllable-/word-initial [sk], [st] vs. $^{??}$[vl], *[kn], *[zw]

Following Hayes (2004) and others, interpret this as evidence for acquisition of a constraint-based phonological perception or 'receptive' grammar

- from observation of legal examples (positive evidence)
- before development of the ability to reproduce these examples

What is the objective of early constraint-based phonological learning, and what prior assumptions are needed to (approx.) achieve that objective?

The restrictiveness objective

Tesar & Smolensky (1998, 2000), Hayes (2004), Prince & Tesar (2004) and much subsequent work take the objective of learning to be **grammar restrictiveness**

(see also earlier work on the **subset problem**, e.g., Baker 1979; Angluin 1980; Pinker 1986, and the associated **subset principle**, e.g., Berwick 1982, 1986; Jacubowitz 1984, Wexler & Manzini 1987)

> *For the learning of phonotactic distributions . . . the goal is always to select the most restrictive grammar consistent with the data. The computational challenge is efficiently to determine, for any given set of data, which grammar is the most restrictive.* (Prince & Tesar 2004:249)

> *If the [learned] rankings are correct, the grammar will act as a filter: it will alter any illegal form to something similar which is legal, but it will allow legal forms to persist unaltered.* (Hayes 2004:168)

Running example: Pseudo-Korean (Hayes 2004)

- Laryngeal distributional pattern of stops
  - Voiceless aspirated vs. unaspirated contrast word-initially
    [t^ha], [ta] (*[da])
  - Voiceless aspirated vs. voiced contrast intervocalically
    [at^ha], [ada], [t^hat^ha], [t^hada], [tat^ha], [tada] (*[ata], *[t^hata], etc.)
  - Only voiceless unaspirated word-finally
    [at], [t^hat], [tat] (*[at^h], *[ad], *[t^had], etc.)

- Constraints

  M: *[+voice][-voice][+voice], *[+s.g.,+voice], *[-son,+voice], *[+s.g.], and 'crazy' *[-s.g.] (Hayes 2004:185)

  F: Ident[s.g.], Ident[s.g.]/_V, Ident[voice], Ident[voice]/_V

Difficulty of achieving restrictiveness

In principle:

> *Multiple grammars can be consistent with the same data, grammars which are empirically distinct in that they make different predictions about other forms not represented in the data. If learning is based upon only positive evidence, then the simple* consistency *of a grammatical hypothesis with all the observed data will not guarantee that this hypothesis is correct . . . .* (Prince & Tesar 2004:245)

In practice:

- Hayes (2004) ran the Recursive Constraint Demotion algorithm of Tesar (1995), Tesar & Smolensky (1998, 2000) on Pseduo-Korean and other cases
- The resulting grammars were quite *unrestrictive*: all IO-Faithfulness constraints in the top stratum, failing to 'filter' many types of illegal form

Toward restrictive learning: prior assumptions

Several *a priori* assumptions have been attributed to the language learner in work that takes restrictiveness as the objective

- **Input assumption**
  The input for phonological learning is identical to the adult surface form (or the 'observable' part of the adult surface form).

  early phonotactic learning: e.g., Tesar & Smolensky 1998, 2000; Hayes 2004, citing Daniel Albro; Prince & Tesar 2004

  production learning: e.g., Smith 1973; Demuth 1995, 1996; Pater & Paradis 1996; Tessier 2007, 2009

- **Low IO-Faithfulness assumption**
  The learner prefers grammars in which Markedness constraints are ranked/weighted more highly than input-output Faithfulness constraints.

  early phonotactic learning: e.g., Smolensky 1996; Tesar & Smolenksy 1998, 2000; Hayes 2004; Prince & Tesar 2004

  production learning: e.g., Demuth 1995; Pater & Paradis 1996; Itô & Mester 1999; Gnanadesikan 2004; Levelt & van de Vijver 2004

- **High OO-Faithfulness assumption** (McCarthy 1998; see also Jesney & Tessier 2009)

Toward restrictive learning: mechanisms

The Low IO-Faithfulness assumption has been enforced by an increasingly complex set of language-specific learning mechanisms

- Markedness $\gg$ IO-Faithfulness in the initial state

  (Smolensky 1996; Tesar & Smolensky 1998, 2000; originally based on a suggestion of Alan Prince)

- Biased Constraint Demotion (BCD) (Prince & Tesar 2004:269-270)
  ordered priorities: Faithfulness Delay, Avoid the Inactive; prefer Smallest Effective F Set; prefer Richest Markedness Cascade

- Low Faithfulness Constraint Demotion (Hayes 2004:177-182)
  ordered priorities: Favour Markedness; Favour Activeness; Favour Specificity; Favour Autonomy

- Elaborations of Favour Specificity that involve extra-grammatical tracking of phonological context co-occurrence (Hayes 2004:193 and fn.31; Tessier 2007:78)

  (see also Smith 2000 and especially Prince &Tesar 2004 on special/general relations both basic and derived)

Review of learning with the restrictiveness objective

- BCD, Low Faithfulness Constraint Demotion, and related algorithms are not guaranteed to maximize restrictiveness or its r-measure approximation

  (for discussion of the difficulties see Hayes 2004:186 and fn. 25; Prince & Tesar 2004:247,252)

  - see Hayes's website for a comparison of algorithms on Pseudo-Korean and other cases (www.linguistics.ucla.edu/people/hayes/acquisition/)
  - see Magri (2010) on the prospect of ever proving such a result for OT

- Existing algorithms are brittle, highly sensitive to the contents of the constraint set (see Hayes 2004:185, Prince & Tesar 2004:274-278)
- The r-measure approximation ($= \sum_F \sum_M I[M \gg F]$) does not accord with restrictiveness in all cases (Prince & Tesar 2004:252, 276)
- Special/general relations among Faithfulness constraints cannot be calculated independently of the phonological system being learned

  (Prince & Tesar 2004:271-278, Appendix 3; Tessier 2007, chapter 2)

Where to go from here?

- Switch from OT to HG can perhaps simplify the set of *a priori* learning assumptions (e.g., Jesney & Tessier 2009), but still no general guarantees about the restrictiveness of the learned grammars
- Jarosz (2006, 2007) proposes to change the learner's objective — this talk builds on Jarosz's work within a more general Bayesian framework, and with a very different type of inference algorithm

Where to go from here?

- *Embrace the learner's ignorance*

Where to go from here?

- *Embrace the learner's ignorance*
  - If the learner assumes that the input is identical to the output
    then high-ranking/weighted Faithfulness can explain all of the data,
    requiring additional assumptions and mechanisms to suppress Faithfulness.

Where to go from here?

- *Embrace the learner's ignorance*
  - If the learner assumes that the input is identical to the output then high-ranking/weighted Faithfulness can explain all of the data, requiring additional assumptions and mechanisms to suppress Faithfulness.
  - What if the Input assumption and the Low IO-Faithfulness assumption were both eliminated? (NB. not claiming that the learner fails to know the adult surface form)

Where to go from here?

- *Embrace the learner's ignorance*
  - If the learner assumes that the input is identical to the output then high-ranking/weighted Faithfulness can explain all of the data, requiring additional assumptions and mechanisms to suppress Faithfulness.
  - What if the Input assumption and the Low IO-Faithfulness assumption were both eliminated? (NB. not claiming that the learner fails to know the adult surface form)
  - What if no learning-specific assumptions were made and every unobserved variable had to be inferred from the positive data?

Where to go from here?

- *Embrace the learner's ignorance*
  - If the learner assumes that the input is identical to the output then high-ranking/weighted Faithfulness can explain all of the data, requiring additional assumptions and mechanisms to suppress Faithfulness.
  - What if the Input assumption and the Low IO-Faithfulness assumption were both eliminated? (NB. not claiming that the learner fails to know the adult surface form)
  - What if no learning-specific assumptions were made and every unobserved variable had to be inferred from the positive data?
  - Need a learning framework in which multiple interdependent hidden variables can be jointly inferred from the available evidence . . .

Bayesian inference for constraint-based phonology with Gibbs sampling

Bayesian grammatical inference

$$\underbrace{p(grammar|data)}_{posterior} \propto \underbrace{p(data|grammar)}_{likelihood} \times \underbrace{p(grammar)}_{prior}$$

- Likelihood term favors grammars that assign higher probability to the positive data (observed surface forms) — this is related to consistency vs. restrictiveness
- Prior term favors grammars that accord with *a priori* assumptions
  - – could include Low Faithfulness, Favour Specificity, r-measure value, etc.
  - \+ instead assume that the prior is maximally unbiased (uninformative)
- Goal of inference is not to find a single grammar with maximal posterior probability, but to *sample grammars according to their posterior values*

Specialization to noisy Harmonic Grammar

- HG defines harmony as the **weighted sum** of constraint violations, and optimality as **most harmonic** among the competitors

(Legendre, Miyata & Smolensky 1990ab, Smolensky & Legendre 2006, Keller 2006, Pater 2009, Potts et al. 2010)

/t^{h}ata/	*[+voice][-voice][+voice] 12	*[-son,+voice] 10	Ident[voice] 1	H
[t^{h}ata]	-1	0	0	-12
☞ [t^{h}ada]	0	-1	-1	-11

- Given a constraint set, a particular HG grammar is defined by a weight vector $\bar{\mathbf{w}} = (\bar{w}_1, \ldots, \bar{w}_m)$ with one component per constraint (each $\bar{w}_k \geq 0$)

Specialization to noisy Harmonic Grammar

- A noisy HG system is defined by a weight vector $\bar{\mathbf{w}}$ and a (Gaussian) noise distribution $\mathcal{N}(0, \sigma^2 I)$. Each evaluation involves drawing a weight vector

$$\mathbf{w} \sim \bar{\mathbf{w}} + \mathcal{N}(0, \sigma^2 I)$$

and calculating harmony and optimality with the sampled weights

(Boersma & Pater 2008, Jesney & Tessier 2009, with precedent in Boersma 1998, Boersma & Hayes 2001; cf. Goldrick & Daland 2009, who add noise to constraint violations instead of weights)

/t^{h}ata/	*[+voice][-voice][+voice] 12 + 0.476	*[-son,+voice] 10 - 1.25	Ident[voice] 1 - 0.329	H
[t^{h}ata]	-1	0	0	-12.476
☞ [t^{h}ada]	0	-1	-1	-9.421

/t^{h}ata/	*[+voice][-voice][+voice] 12 - 0.711	*[-son,+voice] 10 + 1.096	Ident[voice] 1 + 1.154	H
☞ [t^{h}ata]	-1	0	0	-11.289
[t^{h}ada]	0	-1	-1	-13.25

Specialization to noisy Harmonic Grammar

Independence and domain of weight sampling

- The sampling distribution

$$\mathbf{w} \sim \bar{\mathbf{w}} + \mathcal{N}(0, \sigma^2 I)$$

  is equivalent to sampling each weight $w_k \sim \bar{w}_k + \mathcal{N}(0, \sigma^2)$ independently of the other weights, giving the probability of the vector $\mathbf{w}$ a simple form:

$$p(\mathbf{w}|\bar{\mathbf{w}}, \sigma^2) = \prod_{k=1}^{m} \mathcal{N}(w_k; \bar{w}_k, \sigma^2)$$

- Technically weight sampling is from truncated Gaussians restricted to $[0, \infty)$ — no negative weights are allowed — but this does not affect the inference.

Notation

grammar	$\bar{\mathbf{w}}$	
weights	$\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n\}$	
inputs	$\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$	
outputs	$\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$	(the observed, positive data)
harmony	$h_{\mathbf{w}}(x, y)$	
optimality	$\mathsf{HG}(\mathbf{w}, x) = \{\ y \mid \forall y' \in \mathsf{Gen}(x) :\ h_{\mathbf{w}}(x, y) \geq h_{\mathbf{w}}(x, y')\ \}$	

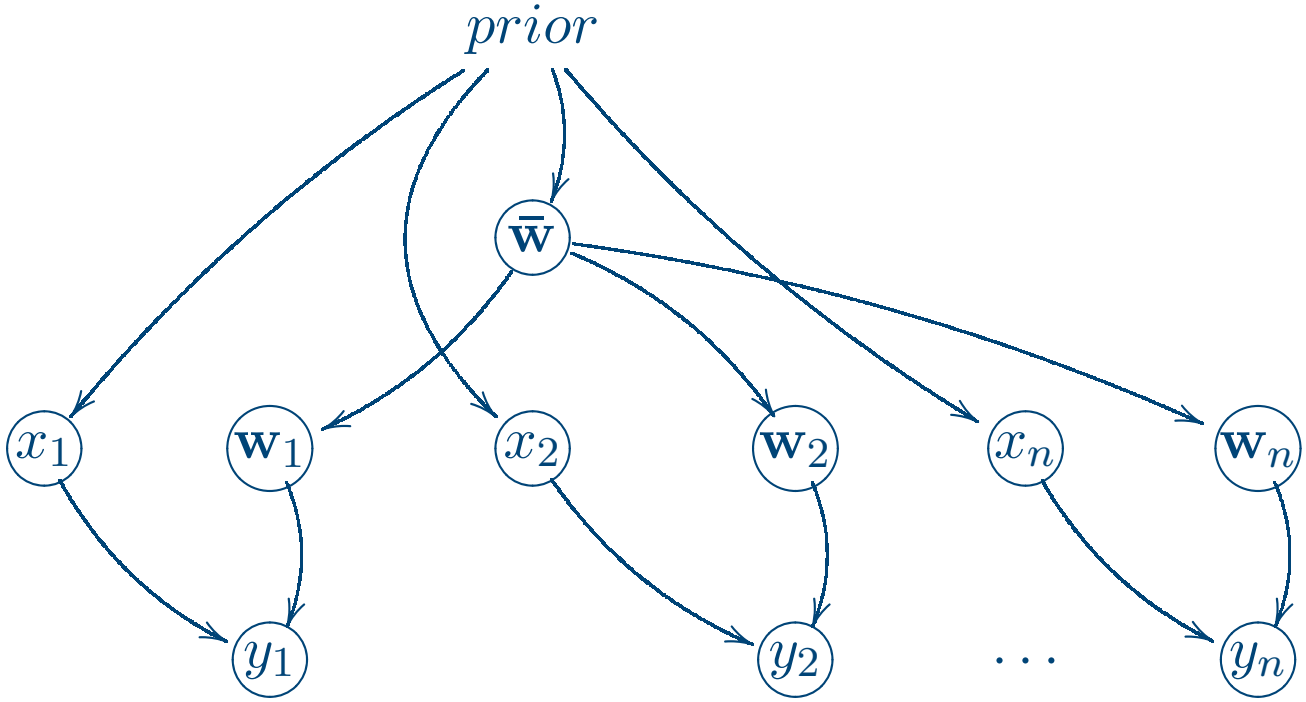
Proposed learner maintains the invariant that $y_i \in \mathsf{HG}(\mathbf{w}_i, x_i)$ (for all $i = 1, \ldots, n$), where typically $\{y_i\} = \mathsf{HG}(\mathbf{w}_i, x_i)$ [i.e., harmony ties are rare]

Graphical model of the data

grammar, noise

inputs, weights

observed outputs



(variance parameter σ^2 is assumed to be fixed here, but could also be inferred)

The inference problem

Joint posterior distribution of the grammar, weight samples, and inputs given the observed data:

$$\underbrace{p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}|\mathbf{y})}_{posterior} \propto \underbrace{p(\mathbf{y}|\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x})}_{likelihood} \times \underbrace{p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x})}_{prior}$$

The inference problem

Joint posterior distribution of the grammar, weight samples, and inputs given the observed data:

$$\underbrace{p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}|\mathbf{y})}_{posterior} \propto \underbrace{p(\mathbf{y}|\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x})}_{likelihood} \times \underbrace{p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x})}_{prior}$$

Assuming a completely *unbiased (uninformative) prior* the joint posterior distribution becomes:

$$\underbrace{p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}|\mathbf{y})}_{posterior} \propto \underbrace{p(\mathbf{y}|\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x})}_{likelihood} \times 1$$

The inference problem

Joint posterior distribution of the grammar, weight samples, and inputs given the observed data:

$$\underbrace{p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}|\mathbf{y})}_{posterior} \propto \underbrace{p(\mathbf{y}|\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x})}_{likelihood} \times \underbrace{p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x})}_{prior}$$

Assuming a completely *unbiased (uninformative) prior* the joint posterior distribution becomes:

$$\underbrace{p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}|\mathbf{y})}_{posterior} \propto \underbrace{p(\mathbf{y}|\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x})}_{likelihood} \times 1$$

⋆ **Learning objective**: draw samples from this posterior distribution, and in particular find grammars ($\bar{\mathbf{w}}$) that have high posterior probability given the data

Analyzing the inference problem

The joint posterior distribution factorizes as follows (recall graphical model):

$$p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}) \times 1$$

$$\propto p(\mathbf{y}|\mathbf{W}, \mathbf{x}) \times p(\mathbf{W}|\bar{\mathbf{w}}) \times 1$$

Analyzing the inference problem

The joint posterior distribution factorizes as follows (recall graphical model):

$$p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}) \times 1$$

$$\propto p(\mathbf{y}|\mathbf{W}, \mathbf{x}) \times p(\mathbf{W}|\bar{\mathbf{w}}) \times 1$$

where

- $p(\mathbf{y}|\mathbf{W}, \mathbf{x}) = 1$ iff, for all $i = 1, \ldots, n$, the observed output y_i is optimal given input x_i and weight sample $\mathbf{w}_i$: that is, $y_i \in \mathsf{HG}(\mathbf{w}_i, x_i)$
  - otherwise the observed data is not generated and $p(\mathbf{y}|\mathbf{W}, \mathbf{x}) = 0$
  - learner proposed below maintains the invariant that $p(\mathbf{y}|\mathbf{W}, \mathbf{x}) = 1$

Analyzing the inference problem

The joint posterior distribution factorizes as follows (recall graphical model):

$$p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}) \times 1$$

$$\propto p(\mathbf{y}|\mathbf{W}, \mathbf{x}) \times p(\mathbf{W}|\bar{\mathbf{w}}) \times 1$$

where

- $p(\mathbf{y}|\mathbf{W}, \mathbf{x}) = 1$ iff, for all $i = 1, \ldots, n$, the observed output y_i is optimal given input x_i and weight sample $\mathbf{w}_i$: that is, $y_i \in \mathsf{HG}(\mathbf{w}_i, x_i)$
  - otherwise the observed data is not generated and $p(\mathbf{y}|\mathbf{W}, \mathbf{x}) = 0$
  - learner proposed below maintains the invariant that $p(\mathbf{y}|\mathbf{W}, \mathbf{x}) = 1$
- $p(\mathbf{W}|\bar{\mathbf{w}}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{w}_i|\bar{\mathbf{w}}, \sigma^2 I) = \prod_{k=1}^{m} \left[\prod_{i=1}^{n} \mathcal{N}(w_{ik}|\bar{w}_k, \sigma^2)\right]$

Analyzing the inference problem

The factorized form of the joint posterior

$$p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}|\mathbf{y}) \;\propto\; p(\mathbf{y}|\mathbf{W}, \mathbf{x}) \times p(\mathbf{W}|\bar{\mathbf{w}}) \times 1$$

suggests an iterative inference (sampling) strategy, looping the following steps:

Analyzing the inference problem

The factorized form of the joint posterior

$$p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}|\mathbf{y}) \;\propto\; p(\mathbf{y}|\mathbf{W}, \mathbf{x}) \times p(\mathbf{W}|\bar{\mathbf{w}}) \times 1$$

suggests an iterative inference (sampling) strategy, looping the following steps:

1. Sample grammar $\bar{\mathbf{w}}$ holding fixed $\mathbf{W}$ and $\mathbf{x}$
   If the weight samples were $\mathbf{W}$*, what would be a likely grammar* $\bar{\mathbf{w}}$*?*

Analyzing the inference problem

The factorized form of the joint posterior

$$p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}|\mathbf{y}) \;\propto\; p(\mathbf{y}|\mathbf{W}, \mathbf{x}) \times p(\mathbf{W}|\bar{\mathbf{w}}) \times 1$$

suggests an iterative inference (sampling) strategy, looping the following steps:

1. Sample grammar $\bar{\mathbf{w}}$ holding fixed $\mathbf{W}$ and $\mathbf{x}$
   If the weight samples were $\mathbf{W}$, what would be a likely grammar $\bar{\mathbf{w}}$?

2. Sample weights $\mathbf{W}$ holding fixed $\bar{\mathbf{w}}$ and $\mathbf{x}$
   If the grammar were $\bar{\mathbf{w}}$ and the inputs were $\mathbf{x}$, what are likely weight samples $\mathbf{W}$ could map the inputs to the observed outputs?

Analyzing the inference problem

The factorized form of the joint posterior

$$p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}|\mathbf{y}) \;\propto\; p(\mathbf{y}|\mathbf{W}, \mathbf{x}) \times p(\mathbf{W}|\bar{\mathbf{w}}) \times 1$$

suggests an iterative inference (sampling) strategy, looping the following steps:

1. Sample grammar $\bar{\mathbf{w}}$ holding fixed $\mathbf{W}$ and $\mathbf{x}$
   If the weight samples were $\mathbf{W}$, what would be a likely grammar $\bar{\mathbf{w}}$?

2. Sample weights $\mathbf{W}$ holding fixed $\bar{\mathbf{w}}$ and $\mathbf{x}$
   If the grammar were $\bar{\mathbf{w}}$ and the inputs were $\mathbf{x}$, what are likely weight samples $\mathbf{W}$ could map the inputs to the observed outputs?

3. Sample inputs $\mathbf{x}$ holding fixed $\bar{\mathbf{w}}$ and $\mathbf{W}$
   If the weight samples were $\mathbf{W}$, what possible inputs $\mathbf{x}$ could be mapped to the observed outputs $\mathbf{y}$?

Interim summary

Up to this point:

- We have eliminated the Input assumption and Low IO-Faithfulness assumption of previous work on early phonological learning. In fact, we are entertaining a completely unbiased prior: $p(input) \propto 1$, $p(grammar) \propto 1$

Interim summary

Up to this point:

- We have eliminated the Input assumption and Low IO-Faithfulness assumption of previous work on early phonological learning. In fact, we are entertaining a completely unbiased prior: $p(input) \propto 1$, $p(grammar) \propto 1$
- This makes the phonological learning problem harder: now must infer inputs and weight samples in addition to the grammar

Interim summary

Up to this point:

- We have eliminated the Input assumption and Low IO-Faithfulness assumption of previous work on early phonological learning. In fact, we are entertaining a completely unbiased prior: $p(input) \propto 1$, $p(grammar) \propto 1$
- This makes the phonological learning problem harder: now must infer inputs and weight samples in addition to the grammar
- The joint posterior distribution over $\langle$grammar, weight samples, inputs$\rangle$ can be written down and factored, but not obvious how to do anything with it

Interim summary

Up to this point:

- We have eliminated the Input assumption and Low IO-Faithfulness assumption of previous work on early phonological learning. In fact, we are entertaining a completely unbiased prior: $p(input) \propto 1$, $p(grammar) \propto 1$
- This makes the phonological learning problem harder: now must infer inputs and weight samples in addition to the grammar
- The joint posterior distribution over $\langle$grammar, weight samples, inputs$\rangle$ can be written down and factored, but not obvious how to do anything with it
- Suggested an iterative strategy of inferring (sampling) each hidden variable while holding the others constant — is this legitimate? how to execute it?

Gibbs sampling

- Gibbs sampling is a general Markov Chain Monte Carlo (MCMC) technique for generating samples from complicated joint distributions

  (Geman & Geman 1993; MacKay 2003; Bishop 2006)

- Given hidden variables $\theta_1, \theta_2, \ldots, \theta_n$ and observed variables, the technique involves a loop in which each θ_i is sampled holding all others constant
  - Requirement: must know and be able to sample from the conditional distribution $p(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$ for each hidden variable θ_i
  - Derivation of conditional distributions involves factorization of the joint and liberal application of Bayes' Theorem, even when the prior is uniform

- Given sufficient time, the Gibbs sampler is guaranteed to draw samples from the joint distribution: **the objective of posterior sampling is achievable**

Gibbs sampling algorithm for noisy HG

Notation: $\alpha^{(t)}$ is the value of variable α at discrete 'time' or step t, where initialization of all variables occurs at time $t = 0$

Initialization

- For each observed surface form y_i, select an initial input $x_i^{(0)}$ and a weight sample $\mathbf{w}_i^{(0)}$ such that $\mathbf{w}_i^{(0)}$ maps x_i to y_i: that is, $y_i \in \mathsf{HG}(\mathbf{w}_i^{(0)}, x_i^{(0)})$

  - can initially set $x_i^{(0)} := y_i$ (recall the Input assumption)

  - given $(x_i^{(0)}, y_i)$, can find $\mathbf{w}_i^{(0)}$ using the **simplex algorithm**

    (Dantzig 1981/1982, Chvátal 1983, Cormen et al. 2001; see Potts et al. 2010 on HG specifcially)

Gibbs sampling algorithm for noisy HG

Repeatedly until convergence

1. Sample grammar $\bar{\mathbf{w}}^{(t+1)}$ given the currently-sampled weight vectors $\mathbf{W}^{(t)} = \mathbf{w}_1^{(t)}, \mathbf{w}_2^{(t)}, \ldots, \mathbf{w}_n^{(t)}$ and fixed noise variance σ^2
2. Sample weight $\mathbf{w}_i^{(t+1)}$ for each output y_i given grammar $\bar{\mathbf{w}}^{(t+1)}$ and $x_i^{(t)}$
3. Sample input $x_i^{(t+1)}$ for each output y_i given $\mathbf{w}_i^{(t+1)}$

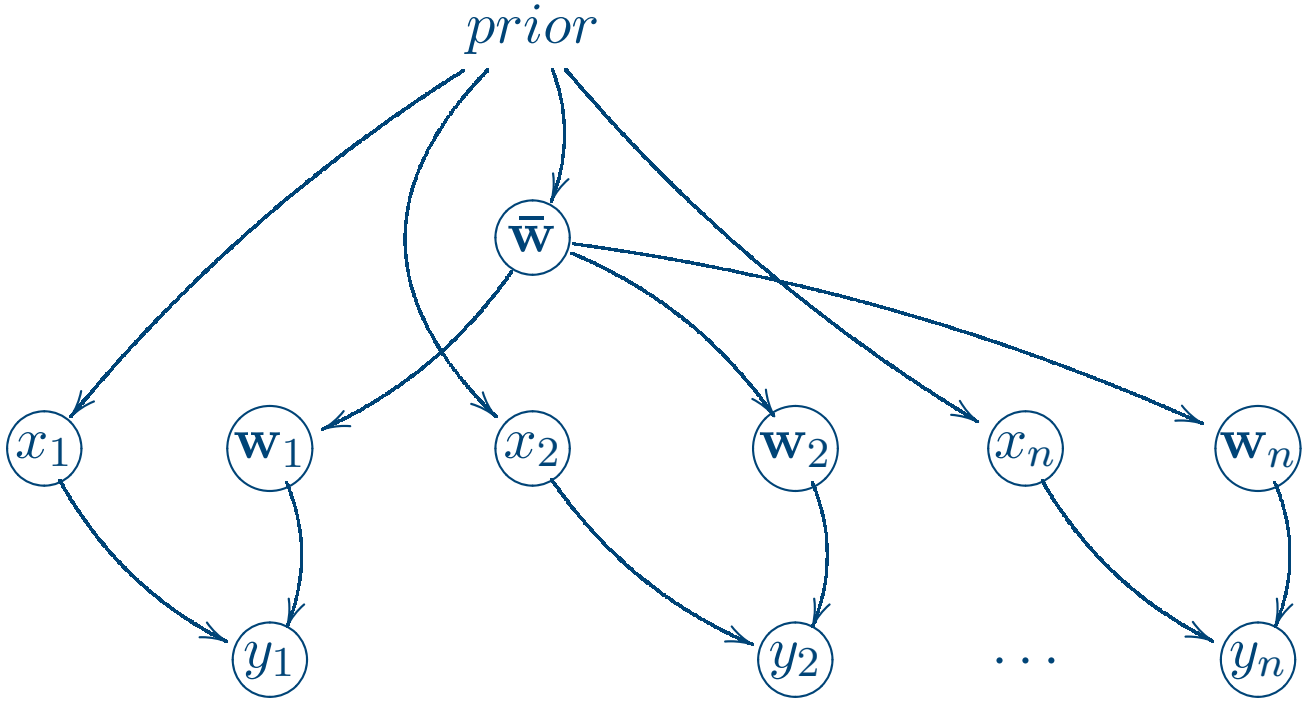
Steps (2) and (3) can be done in parallel, asynchronously with all updates accumulating into the grammar update (1)

Graphical model of the data (repeated)

grammar, noise

inputs, weights

observed outputs



Recall factored posterior: $p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{W}, \mathbf{x}) \times p(\mathbf{W}|\bar{\mathbf{w}})$

1. Sampling a grammar

The grammar $\bar{\mathbf{w}}$ is *conditionally independent* of the observed outputs and sampled inputs given weight samples $\mathbf{W} = \mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_n \sim \mathcal{N}(\bar{\mathbf{w}}, \sigma^2 I)$ i.i.d.

Therefore, grammar sampling reduces to the well-known problem of *inferring the posterior of a multivariate normal mean given known variance and i.i.d. samples*

$$p(\bar{\mathbf{w}}|\mathbf{W}, \sigma^2) \propto \mathcal{N}(\bar{\mathbf{w}}|\tilde{\mathbf{w}}, \frac{\sigma^2}{n} I) \quad \times \quad 1$$

where $\tilde{\mathbf{w}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_i$ is the mean of the current weight samples $\mathbf{W}$

(e.g., Gelman et al. 2004, Bishop 2006)

Presumably could add an across-the-board 'weight decay' prior (e.g., Goldwater & Johnson 2003; Hayes & Wilson 2008) or other weight regularizer without significantly affecting the results. Recall also that samples are actually drawn from a truncated Gaussian restricted to return positive weights.

2. Sampling a weight vector

Each weight vector $\mathbf{w}_i$ is conditionally independent of all other variables and the prior given $\bar{\mathbf{w}}$, σ^2, and the input-output pair (x_i, y_i)

Unnormalized conditional distribution of each $\mathbf{w}_i$ is given by restricting the domain of $\mathcal{N}(\bar{\mathbf{w}}, \sigma^2 I)$ to weight vectors $\mathbf{w}$ that map input x_i to output y_i

$$p(\mathbf{w}_i|\bar{\mathbf{w}}, \sigma^2, x_i, y_i) \propto \begin{cases} \mathcal{N}(\mathbf{w}_i|\bar{\mathbf{w}}, \sigma^2 I) & \text{if } y_i \in \mathsf{HG}(\mathbf{w}, x_i) \\ 0 & \text{otherwise} \end{cases}$$

$\star$ How can we draw samples from this restricted multivariate normal distribution?

2. Sampling a weight vector

Embedded *rejection sampler*?

- Repeatedly draw $\mathbf{w}$ from $\mathcal{N}(\mathbf{w}|\bar{\mathbf{w}}, \sigma^2 I)$ until the condition $y_i \in \mathsf{HG}(\mathbf{w}, x_i)$ is satisfied, then set $\mathbf{w}_i := \mathbf{w}$
- This is guaranteed to 'work' — generate samples from the desired restricted distribution — if we are willing to wait for a very, very long time!

2. Sampling a weight vector

Embedded *Gibbs sampler*

- The condition $y_i \in \mathrm{HG}(\mathbf{w}, x_i)$ can be rewritten as a system of linear inequalities on harmony values (Keller 2006, Potts et al. 2010)

$$\begin{aligned} h_{\mathbf{w}}(x_i, y_i) &\geq h_{\mathbf{w}}(x_i, y') \\ h_{\mathbf{w}}(x_i, y_i) &\geq h_{\mathbf{w}}(x_i, y'') \\ &\vdots \end{aligned}$$

(one harmony inequality per rival candidate in $\mathrm{Gen}(x_i)$)

- The statistics literature provides Gibbs samplers for multivariate normal distributions subject to sets of linear inequalities (e.g., Rodriguez-Yam et al. 2004)

(see example on next slide)

2. Sampling a weight vector

Ex. Suppose that in the process of learning Pseudo-Korean the Gibbs sampler encounters the following partially-described situation:

hypothesized input: /t^{h}ata/	*[-son,+voice] 10	Ident[voice] 5	*[+v][-v][+v] 1
observed output: [t^{h}ada]	-1	-1	0
[t^{h}ata]	0	0	-1

It is highly improbable for a draw from $\mathcal{N}(\bar{\mathbf{w}}, \sigma^2 I)$ to yield a weight sample that maps /t^{h}ata/ to [t^{h}ada], since $\bar{w}$*[-son,+voice] + $\bar{w}$Ident[voice] $>>$ $\bar{w}$*[+v][-v][+v].

Rejection sampling would be very wasteful (billions, trillions of tries required before a single success)

2. Sampling a weight vector

Ex. Suppose that in the process of learning Pseudo-Korean the Gibbs sampler encounters the following partially-described situation:

hypothesized input: /t^{h}ata/	*[-son,+voice] 10	Ident[voice] 5	*[+v][-v][+v] 1
observed output: [t^{h}ada]	-1	-1	0
[t^{h}ata]	0	0	-1

Instead sample $\mathbf{w}$ from the restricted distribution

$\mathcal{N}(\bar{\mathbf{w}}, \sigma^2 I)$ subject to:

$$\underbrace{(-1)(w\text{*[-son,+voice]}) + (-1)(w\text{Ident[voice]})}_{\text{harmony of /t}^h\text{ata/} \to \text{[t}^h\text{ada]}} \geq \underbrace{(-1)(w\text{*[+v][-v][+v]})}_{\text{harmony of /t}^h\text{ata/} \to \text{[t}^h\text{ata]}}$$

General solution: restrict $\mathcal{N}(\bar{\mathbf{w}}, \sigma^2 I)$ by as many linear inequalities as are needed to guarantee $x_i \to y_i$; pass the problem to embedded Rodriguez-Yam sampler

3. Sampling an input

Each input x_i is conditionally independent of all other random variables given the observed output y_i, the weight sample $\mathbf{w}_i$, and the prior

Conditional distribution of x_i is given by restricting the space of all inputs to those that are mapped to y_i under the weighting $\mathbf{w}_i$ (and normalizing)

$$p(x_i|\mathbf{w}_i, y_i) \propto \begin{cases} prior(x_i) & \text{if } y_i \in \mathsf{HG}(\mathbf{w}_i, x_i) \\ 0 & \text{otherwise} \end{cases}$$

What is the prior distribution over inputs? Assume uninformative, $prior(x) \propto 1$, stochastic counterpart of Richness of the Base (see also Jarosz 2006, 2007)

$\star$ How can we draw samples from this restricted distribution over inputs?

3. Sampling an input

Embedded *rejection sampler*?

- Repeatedly draw x from the rich base until the condition $y_i \in \mathsf{HG}(\mathbf{w}_i, x)$ is satisfied, then set $x_i := x$
- As before, this is guaranteed to eventually produce a sample from the desired restricted distribution eventually but can be extremely inefficient

3. Sampling an input

Embedded *Metropolis-Hastings sampler*

- The Input assumption of previous work asserted that the learner's input is *always identical* to the adult surface form
- A more flexible version of this assumption states that valid inputs are likely to lie 'close' to the observed output
- Provisional implementation
  - Consider an observed output y_i as the 'center' of a discrete **string-edit distribution** over potential inputs $q(x|y_i)$ (see for example Cortes et al. 2004, 2008 on the class of **rational kernels**, Smith & Eisner 2005 on **contrastive estimation**)
  - Potential inputs are *proposed* by $q(x|y_i)$ and accepted according to the standard Metropolis-Hastings rule (Metropolis et al. 1953; Hastings 1970; MacKay 2003)

3. Sampling an input

Embedded *Metropolis-Hastings sampler*

Sampling an input x_i given output y_i and weight sample $\mathbf{w}$

- initialize $x_{(0)}$ (e.g., set $x_{(0)} := x_i^{(t-1)}$)
- for $r = 1, \ldots, R$
  - draw a new proposal $x_{(r)} \sim q(x|y_i)$
  - if $y_i \notin \mathsf{HG}(\mathbf{w}_i, x_{(r)})$ reject the proposal
  - else accept the new proposal with probability

$$\min\left\{\frac{1 \times q(x_{(r-1)}|y_i)}{1 \times q(x_{(r)}|y_i)}, 1\right\} \qquad \text{(Metropolis-Hastings rule)}$$

- set $x_i := x_R$ and return

3. Sampling an input

Comments on the string-edit proposal distribution

- A simple way of parameterizing $q(x|y_i)$ is to assess a penalty of $-\log\delta$, $0 \leq \delta \leq 1$, for each change (string edit) made in transforming y into x

  (To calculate the proposal probability $q(x|y_i)$, sum over all possible edits that transform y into x, using standard finite-state techniques; see for example Mohri 2009)

- The Input assumption corresponds to the limit of infinite change penalty, $\delta \to 0$ ($\Rightarrow -\log\delta \to \infty$)

- The present implementation makes a much weaker assumption
  - δ is large enough that proposed inputs are often different from the output
  - sampling is from $p(x|\mathbf{w}, y)$, with $q(x|y)$ just an intermediary tool

3. Sampling an input

Alternative input samplers?

- Could penalize edits in proportion to featural or perceptual distance

  (related to proposals based on the **P-map**, Steriade 1999, 2001, and **harmonic serialism**, McCarthy 2007, et seq., 2011; see also experimental data on perceptual similarity in infants, e.g., Stager & Werker 1997; Pater et al. 2004; Swingley & Aslin 2007)

> *This morning we were talking about freezing water, and I said she would learn about Science in school someday. I said "You probably don't know what Science means yet" and she said "Yes I do Grammy, it means sshhhhhhhh!"* (Carol Wilson, p.c., on EW 3y)

- Calculation of proposal probabilities (and normalizing constant) involve dynamic programming — faster alternative?

Your suggestions are very welcome!

Summary of Gibbs sampler

Repeatedly for $t = 1, \ldots$ after arbitrary valid initialization

1. Sample grammar $\bar{\mathbf{w}}^{(t+1)}$ conditioned on weight samples $\mathbf{W}^{(t)} = \{\mathbf{w}_i^{(t)}\}_{i=0}^{N}$, fixed noise σ^2
   - standard multivariate normal posterior with fixed variance

2. Sample each weight vector $\mathbf{w}_i^{(t+1)}$ conditioned on input sample $x_i^{(t)}$, observed output y_i, grammar $\bar{\mathbf{w}}^{(t+1)}$, and fixed noise σ^2
   - embedded Gibbs sampler for multivariate normal distribution constrained by a system of linear inequalities

3. Sample each input $x_i^{(t+1)}$ conditioned on observed output y_i, $\mathbf{w}_i^{(t+1)}$, and prior
   - embedded Metropolis-Hastings sampler with proposals generated from a string-edit distribution centered on the observed form y_i

Products of the learner

	grammar	weights	inputs
initialization	$\bar{\mathbf{w}}^{(0)}$	$\mathbf{w}_1^{(0)}, \mathbf{w}_2^{(0)}, \ldots, \mathbf{w}_n^{(0)}$	$x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}$
burn-in phase	$\vdots$	$\vdots$	$\vdots$
samples	$\bar{\mathbf{w}}^{(t)}$	$\mathbf{w}_1^{(t)}, \mathbf{w}_2^{(t)}, \ldots, \mathbf{w}_n^{(t)}$	$x_1^{(t)}, x_2^{(t)}, \ldots, x_n^{(t)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
stop	$\bar{\mathbf{w}}^{(T)}$	$\mathbf{w}_1^{(T)}, \mathbf{w}_2^{(T)}, \ldots, \mathbf{w}_n^{(T)}$	$x_1^{(T)}, x_2^{(T)}, \ldots, x_n^{(T)}$

Achieving the learning objective: for t, T large enough, guaranteed that $\bar{\mathbf{w}}^{(t)}, \ldots, \bar{\mathbf{w}}^{(T)}$ is a sample from the posterior distribution over grammars

Examples, analysis, and extensions

Pseudo-Korean simulation

- Exhaustive set of inputs constructed from {t^h, t, d, d^h, a} with skeleta CV, VC, CVC, CVCV

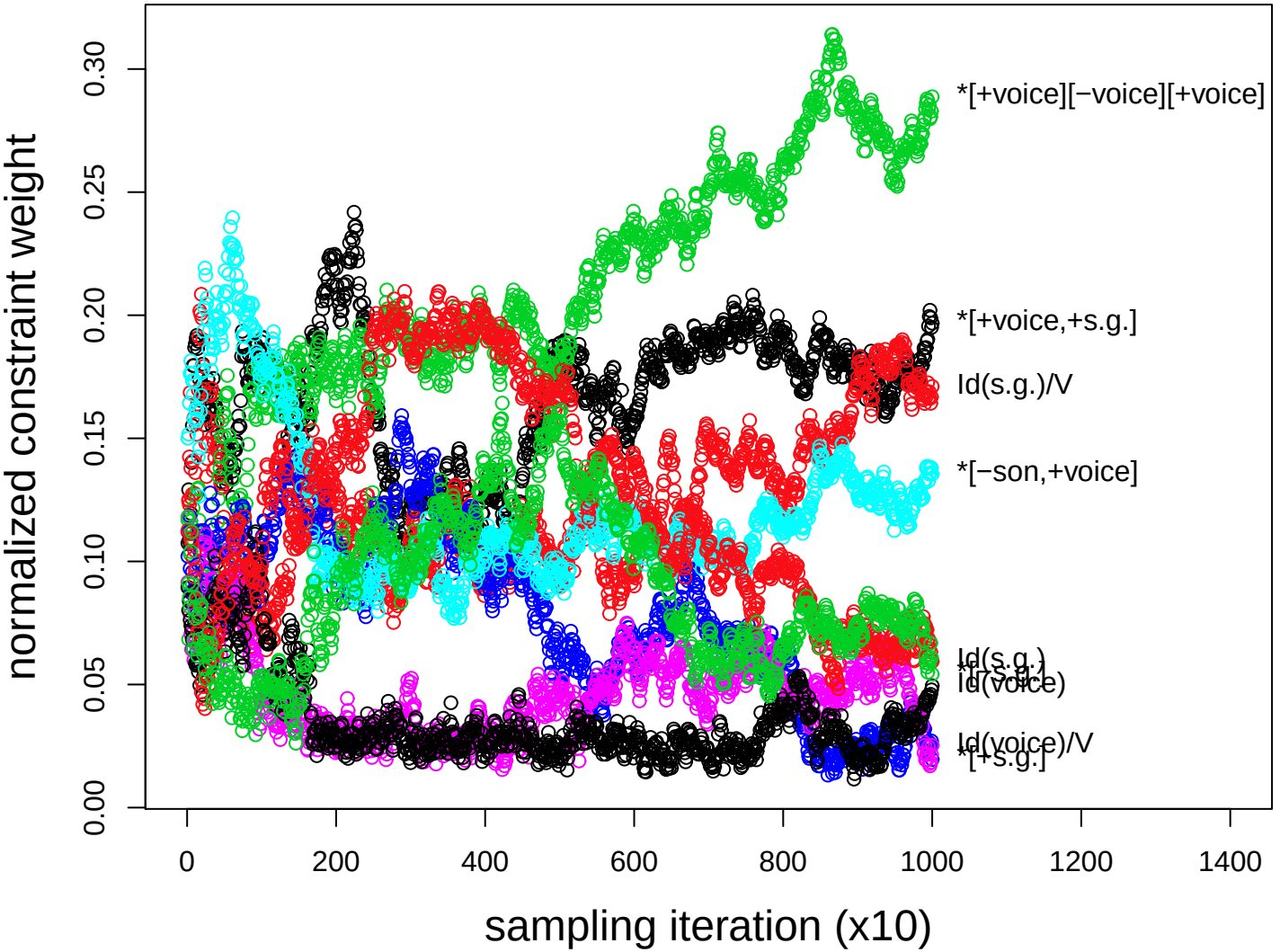
- Adult grammar

M:*[+sg., +voice]	100
F:Ident[s.g.]/_V	100
M:*[+voice][-voice][+voice]	75
M:*[+s.g.]	75
M:*[-son,+voice]	50
F:Ident[voice]/_V	1
F:Ident[voice]	1
F:Ident[s.g.]	1

- Adult language effectively deterministic with $\sigma^2 = 1$

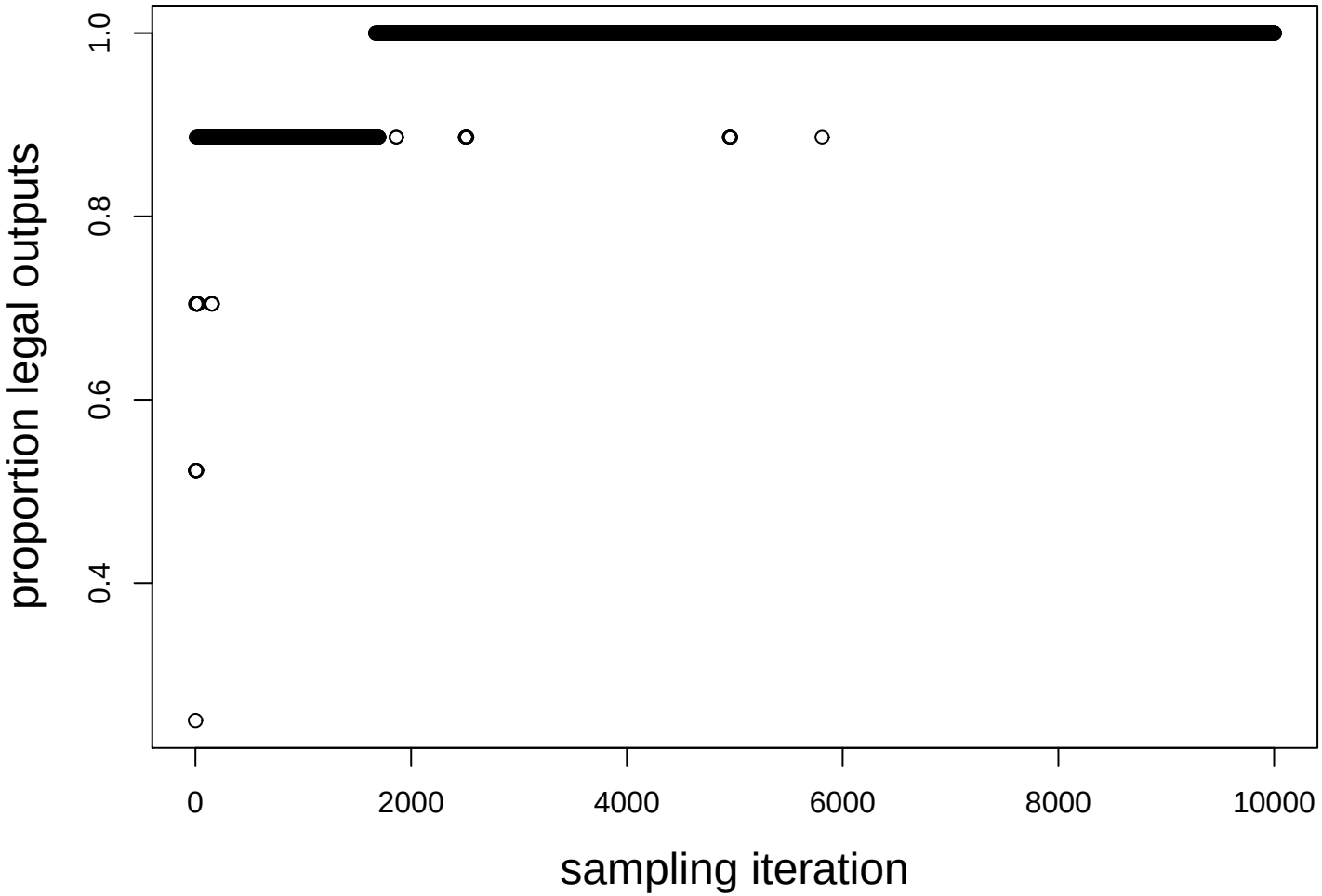
Pseudo-Korean simulation

Weight samples (normalized to sum to unity)



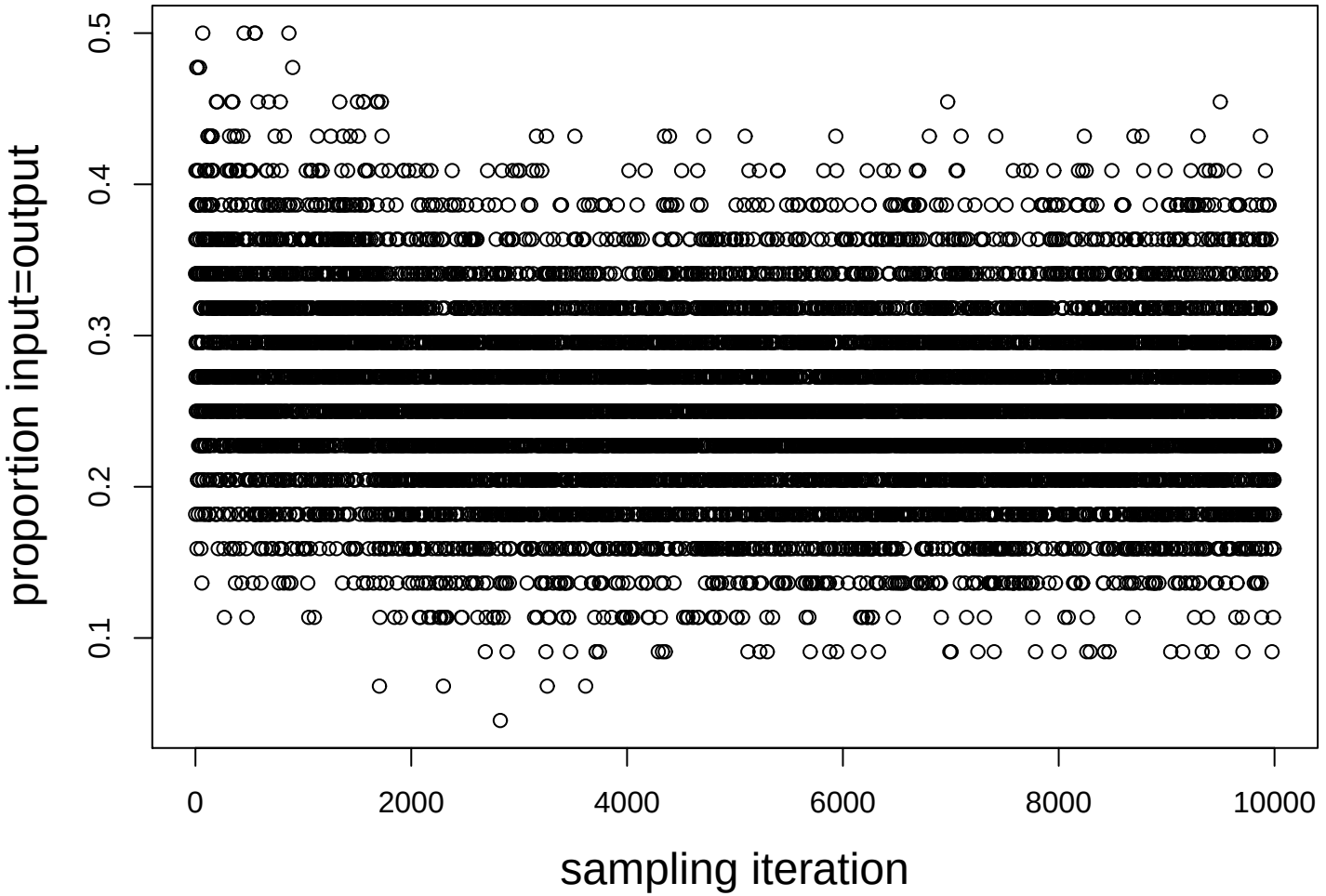
Pseudo-Korean simulation

Restrictiveness test: all inputs in the rich base 'filtered' by grammar samples



Pseudo-Korean simulation

Sampled inputs often differ from observed outputs



Pseudo-Korean simulation

- Late grammar samples generate **all and only** the set of legal outputs, with massive neutralization of inputs from the stochastic rich base

$$\begin{array}{rcl} /t^ha/,\ /d^ha/ & \rightarrow & [t^ha] \\ /ta/,\ /da/ & \rightarrow & [da] \\ /tat^h/,\ /tat/,\ /tad^h/,\ /dat/,\ /dat^h/,\ /dad/,\ /dad^h/ & \rightarrow & [tat] \\ \vdots & & \vdots \end{array}$$

- Sampled grammars richly exploit 'ganging up' property of HG

  Ex. /tat^{h}a/ $\rightarrow$ [tat^{h}a] (not [tada], *[tad^{h}a]) because several constraints *[+voice,+s.g.], Id[s.g.]/_V, *[-son,+voice], Id[voice]/_V, Id[s.g.], Id[voice] gang up to defeat highest-weighted *[+voice][-voice][+voice]

AZBA simulation

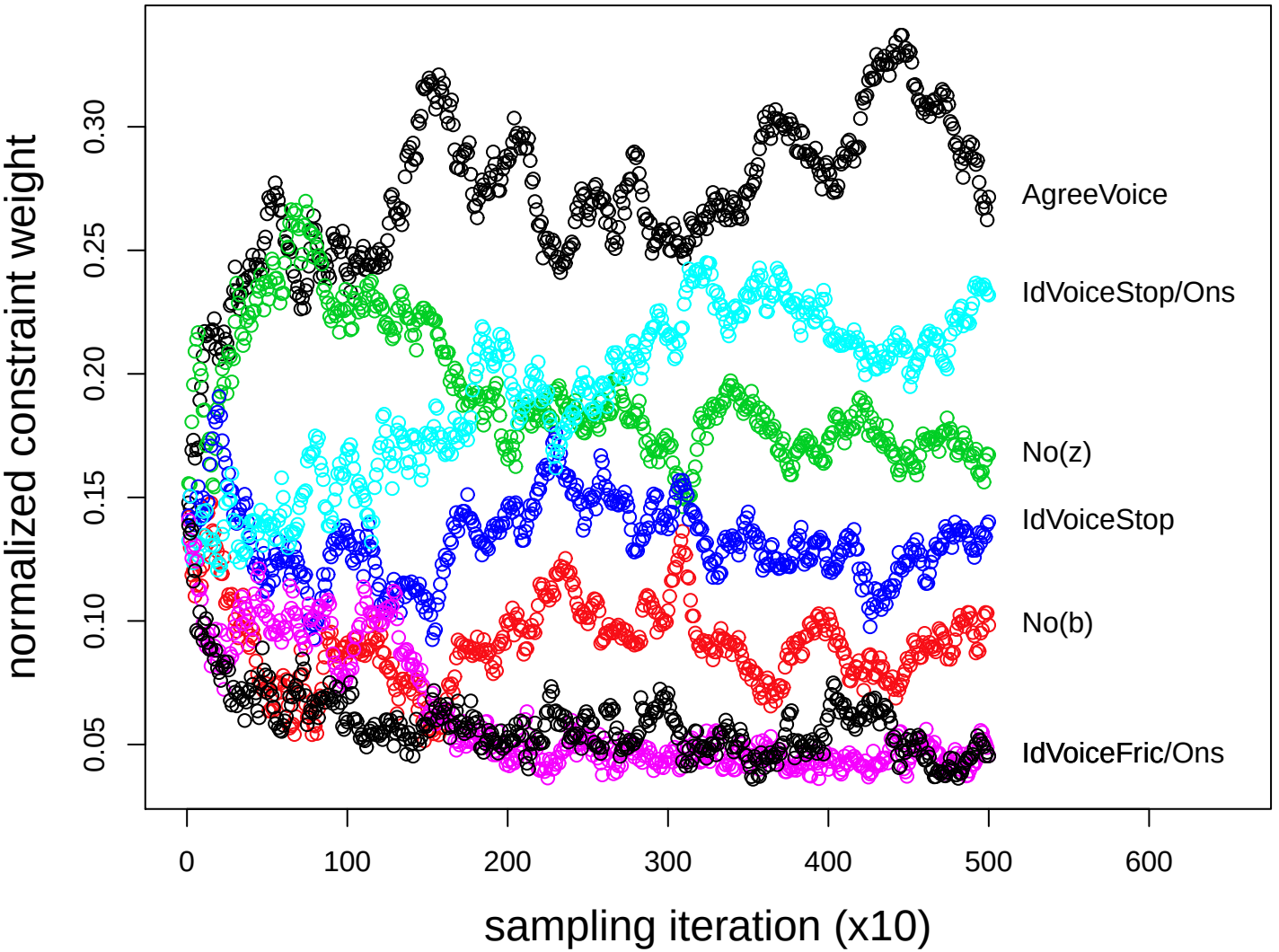
AZBA language type (Prince & Tesar 2004, Hayes 2004)

- Laryngeal phonotactics of stops and fricatives
  - Contrast between voiced and voiceless stops word-initial and word-finally
    [pa, ba, ap, ab]
  - No contrast between voiced and voiceless fricatives
    [sa], [as] (*[za], *[az])
  - Voiced fricatives occur allophonically under regressive voicing assimilation; voicing contrast in stops is neutralized under the same conditions
    [aspa], [azba], [apsa] (*[azpa], *[asba], *[absa], *[abza])

- Constraints
  M: AgreeVoice, NoVcdStop, NoVcdFric
  F: IdentVoiceStop, IdentVoiceStop/Onset
  IdentVoiceFric, IdentVoiceFric/Onset

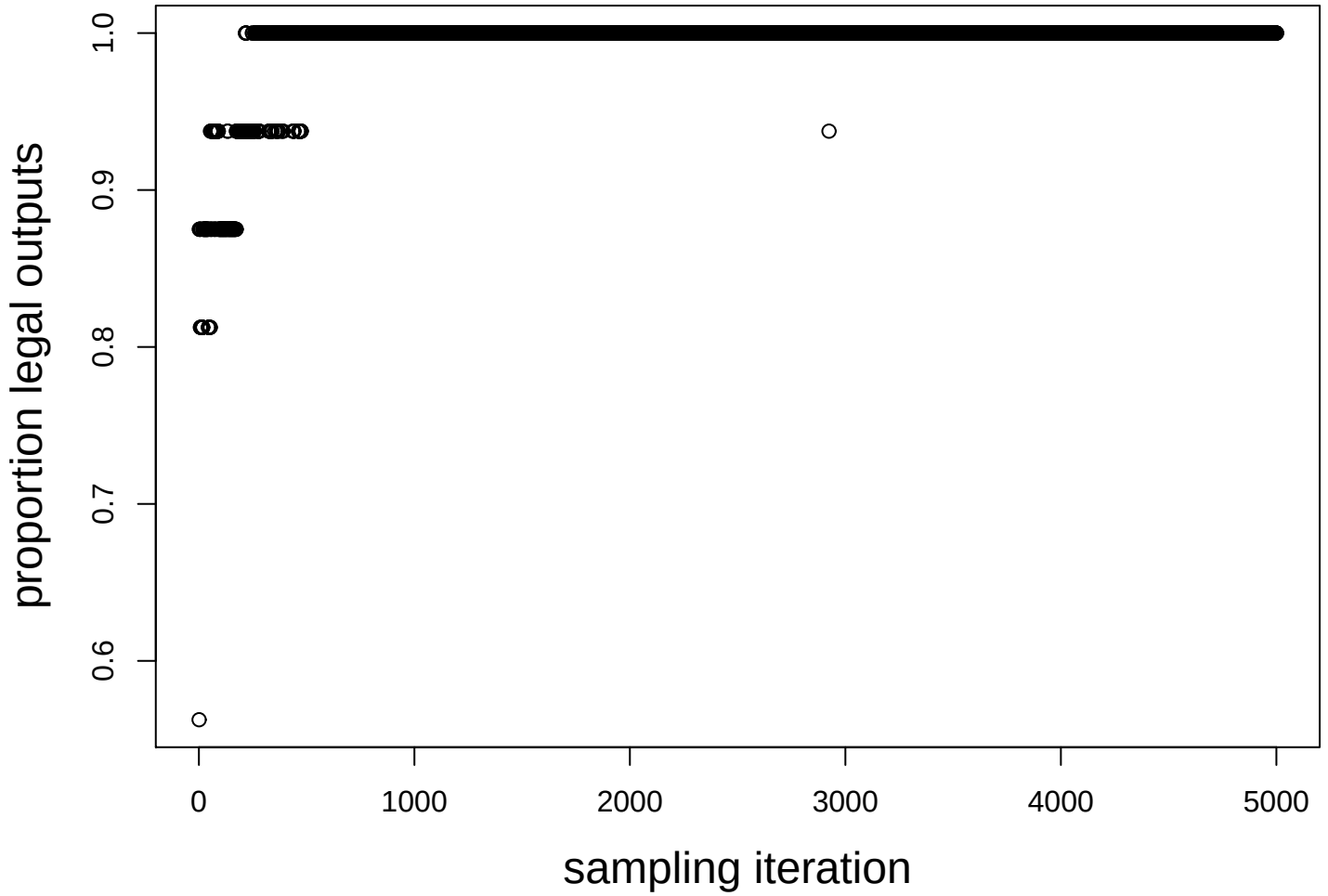
AZBA simulation

Weight samples (normalized to sum to unity)



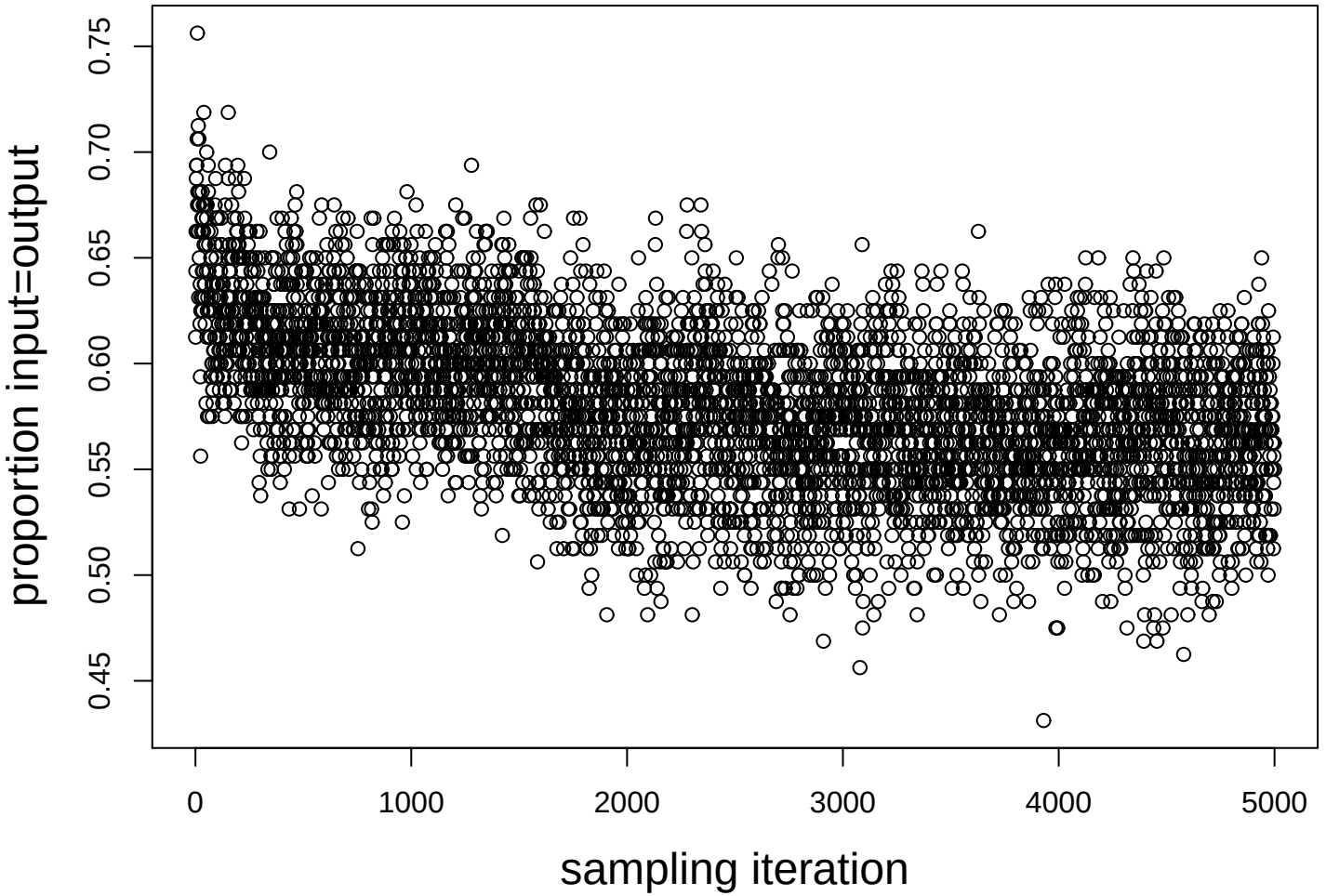
AZBA simulation

Restrictiveness test: all inputs in the rich base 'filtered' by grammar samples



AZBA simulation

Sampled inputs often differ from observed outputs



PAKA simulation

PAKA language type (Prince & Tesar 2004, Hayes 2004, Tessier 2007)

- Only initial-syllable vowels can be stressed (but not all are, perhaps because 'stress' is actually pitch accent)

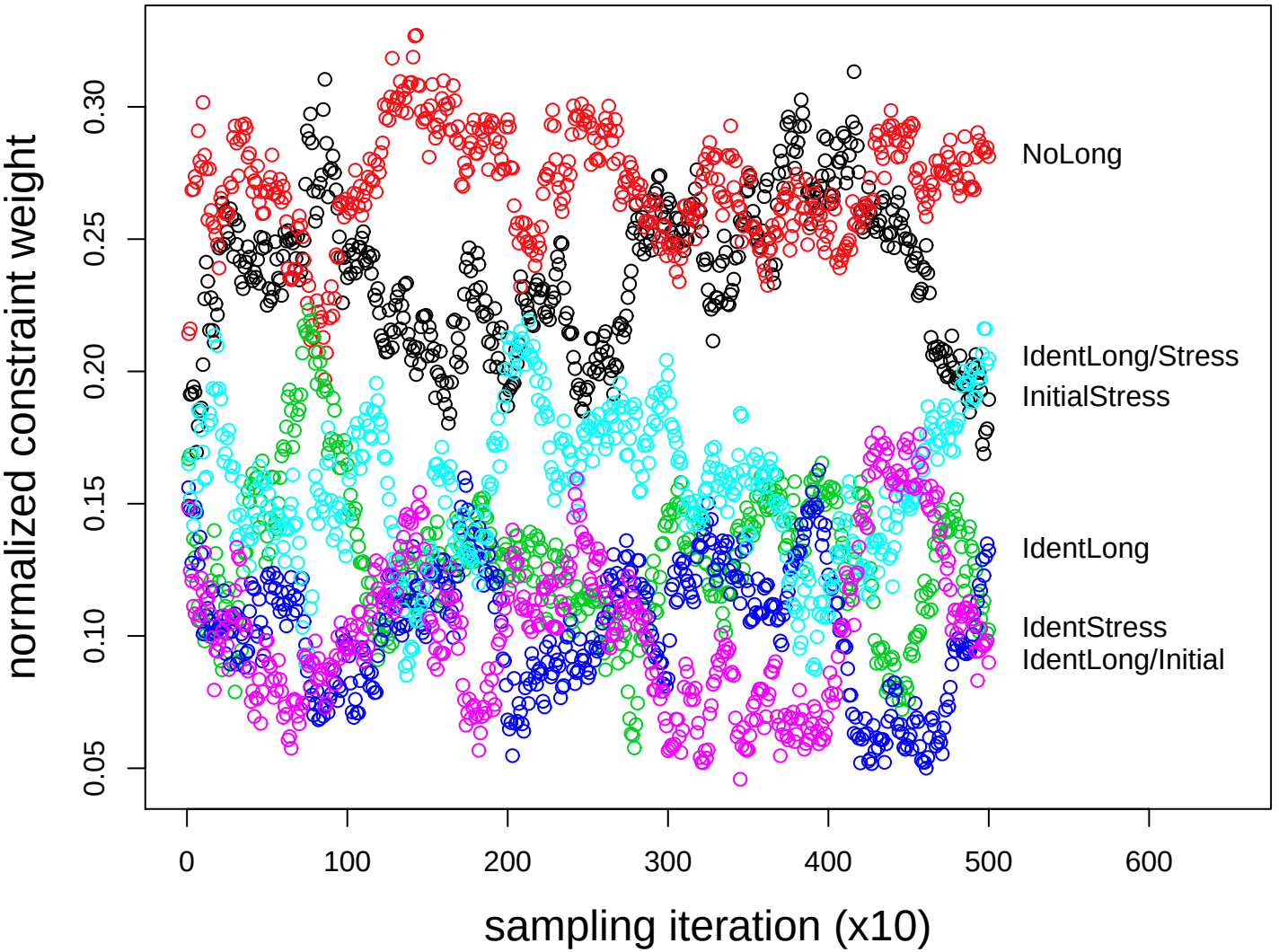
- Only stressed vowels contrast for length

  [páka], [pá:ka], [paka] (*[paká], *[paká:], *[pa:ka], etc.)

- Constraints
  M: InitialStress, NoLong
  F: IdentStress, IdentLong, IdentLong/Stressed, IdentLong/Initial

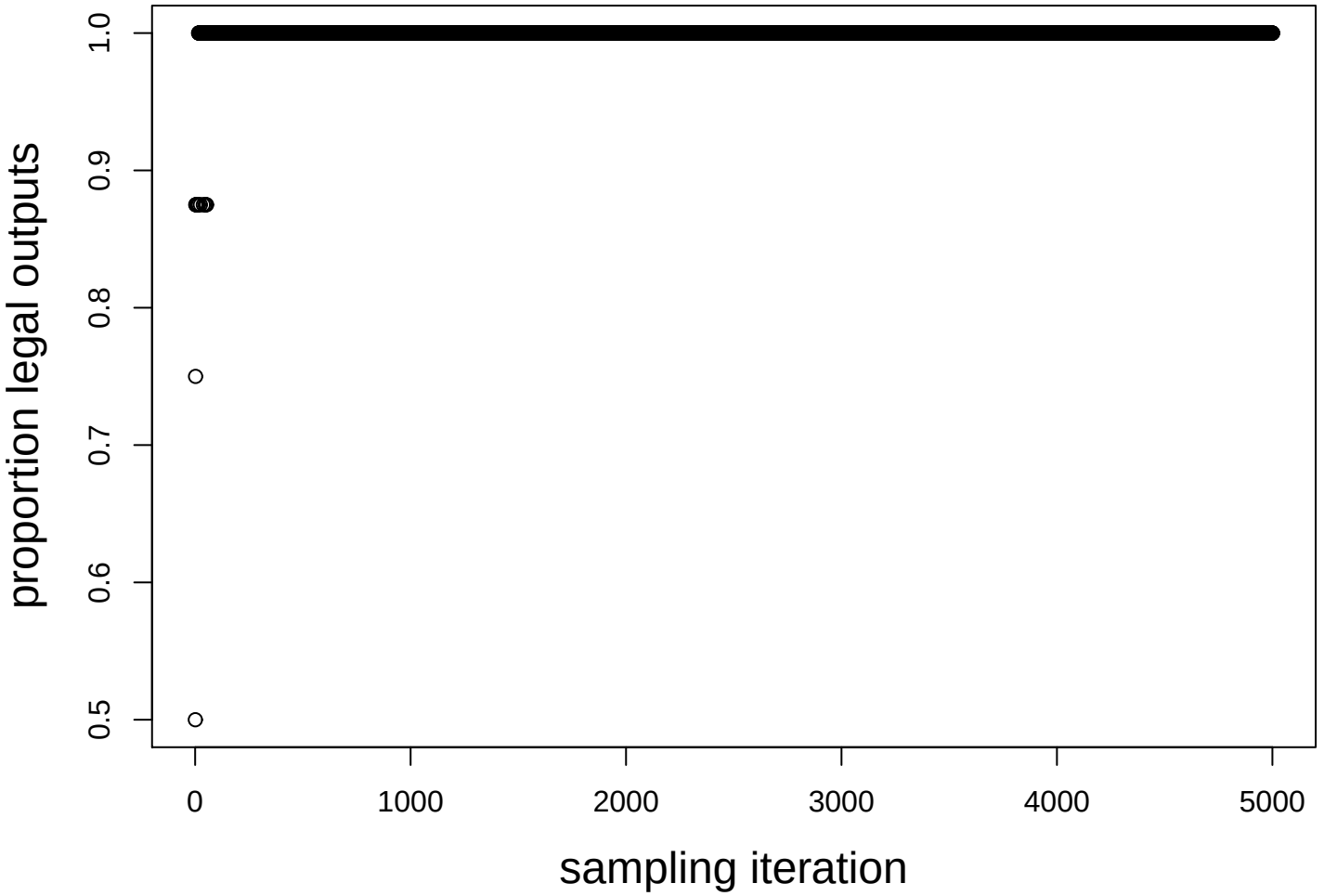
PAKA simulation

Weight samples (normalized to sum to unity)



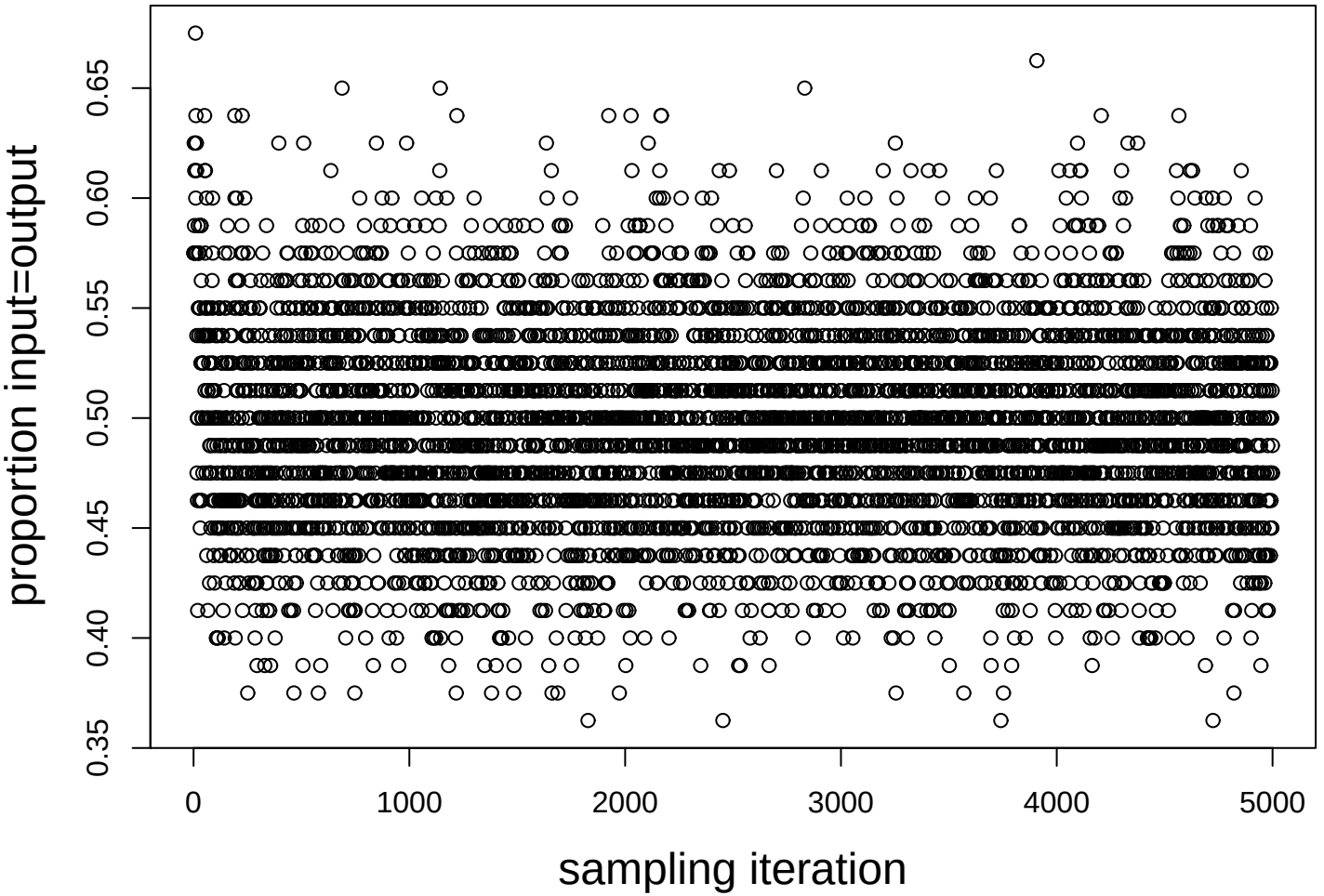
PAKA simulation

Restrictiveness test: all inputs in the rich base 'filtered' by grammar samples



PAKA simulation

Sampled inputs often differ from observed outputs



Restrictive grammars from an uninformative prior

Why do the preceding simulations succeed in finding restrictive grammars?

The restrictive grammars maximize the prob. of the data — and hence their own likelihood — by mapping more inputs from the rich base to the observed forms

- AZBA

$\bar{w}$NoVcdFric $>$ $\bar{w}$IdentVoiceFric + $\bar{w}$IdentVoiceFric/Onset	/sa/	→ [sa]
	/za/	→ [sa]
$\bar{w}$IdentVoiceFric + $\bar{w}$IdentVoiceFric/Onset $>$ $\bar{w}$NoVcdFric	/sa/	→ [sa]
	/za/	→ [za]*

- PAKA

$\bar{w}$NoLong $>$ $\bar{w}$IdentStress/Initial	/paka/	→ [paka]
	/paːka/	→ [paka]
$\bar{w}$IdentStress/Initial $>$ $\bar{w}$NoLong	/paka/	→ [paka]
	/paːka/	→ [paːka]*

Restrictive grammars from an uninformative prior

The *posterior distribution over grammars* is the joint distribution marginalized over all possible weight samples and inputs:

$$p(\bar{\mathbf{w}}|\mathbf{y}) \propto \int_{\mathbf{W}} \sum_{\mathbf{x}} \alpha(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}, \mathbf{y})$$

where

$$\alpha(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{W}, \mathbf{X}) \ \times \ p(\mathbf{W}|\bar{\mathbf{w}})$$

which can be thought of as a positive, fractional 'point' that grammar $\bar{\mathbf{w}}$ receives if the inputs in $\mathbf{x}$ are mapped to the observed outputs by the weights $\mathbf{W}$

Grammars with larger point totals by def. have higher posterior probability

Restrictive grammars from an uninformative prior

Posterior/restrictiveness connection

More restrictive grammars have higher posterior probability because they receive larger probability contributions from more combinations of weights and inputs

(similar ideas suggested by Paul Smolensky, p.c., pursued in Riggle 2006, Jarosz 2006, 2007)

- Only possible if Input assumption is relaxed
- If correct, renders Low IO-Faithfulness assumption and associated learning-specific mechanisms unnecessary

(see example on next slide)

Restrictive grammars from an uninformative prior

Ex. Small fragment of Pseudo-Korean: observe only [ta] not *[da]

/ta/	*[-son,+voice] 10	Id(voice) 5
☞[ta]	0	0
[da]	-1	-1

/da/	*[-son,+voice] 10	Id(voice) 5
☞[ta]	0	-1
[da]	-1	0

/ta/	Id(voice) 10	*[-son,+voice] 5
☞[ta]	0	0
[da]	-1	-1

/da/	Id(voice) 10	*[-son,+voice] 5
[ta]	0	-1
☞[da]	-1	0

$\bar{w}$*[-son,+v] > $\bar{w}$Id(voice)
(more restrictive grammar)

$\bar{w}$Id(voice) > $\bar{w}$*[-son,+v]
(less restrictive grammar)

Restrictive grammars from an uninformative prior

Summary: two-part strategy for guaranteeing grammatical restrictiveness in early phonological learning

- Learning objective: sample from the posterior distribution over grammars
  - Gibbs sampler guaranteed to do this given sufficient time
  - Quite possibly other sampling methods would be even more efficient
- Posterior/restrictiveness connection
  - Prove that grammars with higher posterior probability are more restrictive
  - Trivial given the uninformative prior if more restrictive $\Leftrightarrow$ higher likelihood

(see Jarosz 2006, 2007 for relevant discussion and proof of a simple case)

Extensions

More hidden variables

- Learner receives *partially-observed* surface forms $\mathbf{z}$ (e.g, Tesar & Smolensky 2000; Tesar 2004; Naradowsky et al. 2010; Pater et al. 2010)
  - learning objective: sample from joint posterior $p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}, \mathbf{y}|\mathbf{z})$
  - addition of hidden $\mathbf{y}$ does not qualitatively change the inference problem or the Gibbs sampling strategy
  - harder if elements of $\mathbf{y}$ are derivations of unknown length (e.g., Pater & Staubs 2010)

Extensions

More hidden variables

- Learner receives *partially-observed* surface forms $\mathbf{z}$ (e.g, Tesar & Smolensky 2000; Tesar 2004; Naradowsky et al. 2010; Pater et al. 2010)
  - learning objective: sample from joint posterior $p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}, \mathbf{y}|\mathbf{z})$
  - addition of hidden $\mathbf{y}$ does not qualitatively change the inference problem or the Gibbs sampling strategy
  - harder if elements of $\mathbf{y}$ are derivations of unknown length (e.g., Pater & Staubs 2010)
- Learning morphological relations (e.g., Tesar & Smolensky 2000; Tesar NELS 36, Jarosz 2006, 2007; Jesney et al. 2009; see also Dreyer & Eisner 2008, 2009; Dreyer 2011)
  - (softly) tie inputs across surface realizations of a morpheme

Extensions

More hidden variables

- Learner receives *partially-observed* surface forms $\mathbf{z}$ (e.g, Tesar & Smolensky 2000; Tesar 2004; Naradowsky et al. 2010; Pater et al. 2010)
  - learning objective: sample from joint posterior $p(\bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}, \mathbf{y}|\mathbf{z})$
  - addition of hidden $\mathbf{y}$ does not qualitatively change the inference problem or the Gibbs sampling strategy
  - harder if elements of $\mathbf{y}$ are derivations of unknown length (e.g., Pater & Staubs 2010)
- Learning morphological relations (e.g., Tesar & Smolensky 2000; Tesar NELS 36, Jarosz 2006, 2007; Jesney et al. 2009; see also Dreyer & Eisner 2008, 2009; Dreyer 2011)
  - (softly) tie inputs across surface realizations of a morpheme
- Learning constraints (e.g., Pater & Staubs 2010; Hayes & Wilson 2008; Pater, to appear)
  - learning objective: sample from joint posterior $p(\textsc{Con}, \bar{\mathbf{w}}, \mathbf{W}, \mathbf{x}, |\mathbf{z})$?!

Extensions

Different constraint-based formalisms

- Noisy OT
  - learning objective and inference strategy remain largely the same
  - feasibility depends on parameterization of ranking probabilities

  (e.g., Boersma 1997, Boersma & Hayes 2001 vs. Jarosz 2006, 2007)

- Conditional maximum entropy
  - easier inference problem: no need for weight samples $\mathbf{W}$, because grammars are inherently stochastic

In the near future: comparison of multiple constraint-based grammatical frameworks on learning-theoretic grounds (see already Jesney 2009)

Extensions

Later stages of phonological acquisition

- Many proposals assume (at least partly) distinct grammars for perception and production (e.g., Smith 1973; Demuth 1995, 1996; Pater & Paradis 1996; Boersma 1998; Pater 2004; Hayes 2004; Jarosz 2006, 2007; cf. Smolensky 1996)
  - broadly supported by neurophysiological and neuropsychological evidence for distinct input and output lexicons and processing in adults (e.g., Hickok 2000, Poeppel & Hickok 2004, 2007; Guenther 1995; Guenther et al. 1998, 2006; Shallice 2000)
  - production grammar maps adult-like surface representation to child form (e.g., Smith 1973; Demuth 1995; Gnanadesikan 1995; Pater & Paradis 1996; Levelt and Van de Vijver 1998; Hayes 2004)

- Tentative proposal: train initial production grammar from perception grammar by mapping all adult surface forms to silence (null output)

$$p(\bar{\mathbf{w}}_{prod}) \propto sim(\bar{\mathbf{w}}_{prod}, \bar{\mathbf{w}}_{perc}) \times \delta(null\ outputs|\bar{\mathbf{w}}_{prod})$$

Summary

- Bayesian inference of (sampling from) the posterior distribution on noisy Harmonic Grammars can be achieved with modern MCMC techniques

- More restrictive grammars in test cases have higher posterior probability, despite the unbiased prior, because they receive support from more combinations of the hidden variables (= weight samples, inputs)
  - supported by replicable simulations
  - difficulty of general proof hinges on def. of 'restrictive'

- Pursuing simplification in learning theory
  - avoid learning-specific assumptions
  - embrace learner's ignorance of hidden variables
  - pursue consequences of inference with an unbiased prior

Thank you!