Learning Latent Representations with Progressive Hypothesis Space Expansion

Jonathan Charles Paramore
University of California, Santa Cruz
jcparamo@ucsc.edu

Abstract

This paper introduces a learning model to address the computational challenges arising from including highly abstract underlying representations (URs) in morphophonemic learning. The proposed learner structures the UR hypothesis space by disparity distance and considers potential URs in batches, beginning with fully concrete URs, only expanding the UR candidate space if the current set of UR candidates fails to meet a predetermined likelihood threshold. Overall, the learner inherently restricts abstraction to cases where introducing it demonstrably improves likelihood, while avoiding issues associated with the exhaustive search of an unbounded hypothesis space. Applied to Pakistani Punjabi vowel nasality, the model is shown to successfully acquire abstract URs for phonological patterns that parallel learners fail to capture.

1 Introduction

Classic debates concerning the degree to which an underlying representation (UR) and its associated surface representations (SRs) may differ (e.g., Kenstowicz and Kisseberth 1977) have reemerged in recent years. This resurgent interest owes primarily to the difficulty that increased abstraction presents for phonological learnability. On the one hand, some phonological patterns are difficult to analyze without positing highly abstract URs (e.g., Klamath (O'Hara, 2017), Pakistani Punjabi (Paramore, 2025), and Sevillian Spanish (Gilbert, 2023)). In these cases, proposed URs are considered highly abstract in that they differ in one or more feature specifications from all of their corresponding surface exponents.

Nevertheless, despite the analytical motivations for assuming abstract URs, recent work on phonological learnability has raised serious concerns about the viability of such high degrees of UR abstraction. First, if multiple URs successfully model the observed data, what mechanism ensures that the learner converges on the minimally abstract UR? That is, how is UR abstraction permitted only to the extent necessary (Prince and Smolensky, 1993/2004)? Second, and perhaps more pressingly, if UR abstraction is unconstrained, the space of possible URs grows without bound, introducing a potentially intractable search space problem (Jarosz, 2019; Wang and Hayes, 2025).

While a number of learning algorithms have been proposed that minimize UR abstraction within Maximum Entropy frameworks (O'Hara, 2017; Paramore, 2025), the issue of search space size brought on by permitting increased UR abstraction remains largely unresolved. To date, most learning models have simply stipulated a finite set of potential URs for the learner to consider (O'Hara, 2017; Paramore, 2025). Wang and Hayes (2025), on the other hand, demonstrate that at certain levels of UR abstraction, the space of potential URs grows too large, preventing the learner from converging on a correct solution. Given these results, Wang and Hayes (2025) propose that the boundary on permissible UR abstraction should be defined by learnability itself. This conclusion places learnability-based limits on abstraction in direct tension with patterns in languages like those discussed above that appear to require highly abstract URs.

This paper introduces a novel learning algorithm—**the UR Progressive Hypothesis Space Expansion (UR PHaSE) learner**—which aims to reconcile these competing demands. The UR PHaSE learner solves both of the above issues posed by UR abstraction

without imposing an explicit upper bound on the degree of UR abstraction. The central insight builds on work by Tesar (2014, 2016), which structures the UR search space so that, rather than evaluating an infinite set of potential URs in parallel, the learner partitions the space into ordered subsets and considers candidates incrementally based on disparity count. Beginning with fully concrete URs, the learner only expands the space to include increasingly abstract URs when it fails to find an adequate learning solution at a lower level of abstraction. Moreover, a key novelty of the UR PHaSE learner is that expansion of the UR hypothesis space is *targeted*: the learner uses information from constraint weights and violations in observed forms to generate only those candidate URs that are potentially informative for learning. In this way, the hypothesis space remains manageable as the learner iteratively considers increasingly abstract UR candidates.

Using a vowel nasality pattern from Pakistani Punjabi (Paramore, 2025; Paramore and Bennett, 2025), I show that an appropriately structured hypothesis space of potential URs coupled with a serial search algorithm renders the required level of UR abstraction learnable without needing to provide external stipulations on the UR candidate set.

2 Pakistani Punjabi Nasality

Punjabi is an Indo-Aryan language spoken by approximately 110 million people. Paramore and Bennett (2025) describe a vowel nasality pattern in Pakistani Punjabi that requires vowels occurring before tautomorphemic nasal consonants (pre-N vowels) to be represented with abstract URs containing a nasality feature that never surfaces faithfully.

As shown by examples (i) and (iii) in Table 1, vowel nasality is contrastive in Punjabi. Importantly, two phonological processes shape the realization of nasality across words. First, all pre-N vowels are categorically nasalized, as shown in Table 1iv-vi. In addition, contrastive nasal vowels trigger regressive nasal harmony, such that nasality spreads to all glides and vowels, while other consonants block further propagation (Table 1i-ii). Crucially, however, while Paramore and Bennett (2025) show that pre-N vowels are phonetically indistinguishable from contrastive nasal vowels, they also demonstrate that pre-N vowels do not trigger nasal harmony, as shown by the examples in Table 1v-vi.

i.	[sɑ]	'breath'	ii.	[sɑ̃ʊ̃ɑ̃]	'breaths'
iii.	[sɑ̃]	'I was'	iv.	[sɑ̃ŋ]	'grindstone'
v.	[sijɑ̃ŋ]	'recognition'	vi.	[dʒəʊɑ̃ŋ]	'a youth'

Table 1: Forms illustrating contrastive vowel nasality and nasal harmony in Punjabi.

To account for the distinct behavior of pre-N vowels and contrastive nasal vowels in their induction of nasal harmony, Paramore and Bennett (2025) analyze pre-N vowels as underlying oral, undergoing local categorical nasalization in the context of a following nasal consonant (I refer the interested reader to their paper for a more in-depth defense for analyzing pre-N vowels as abstract). Given this analysis, explaining the disparate behavior of pre-N and contrastive nasal vowels is straightforward: underlying nasal vowels trigger harmony, while derived nasal pre-N vowels do not.

Under this analysis, pre-N vowels are assigned abstract URs, oral underlyingly in order to block harmony triggering, but categorically nasal on the surface: /sijɑŋ/ → [sijɑ̃ŋ], *[sĩjɑ̃ŋ]. The level of abstraction posited for these pre-N vowels exceeds what Wang and Hayes (2025) find to be learnable using a learning approach that considers all UR candidates in parallel. Specifically, they demonstrate that permitting URs containing a feature that never surfaces in any allomorph as a viable UR candidate expands the hypothesis space beyond what can be effectively searched, resulting in learning failures (Wang and Hayes 2025, p.30-31).

3 The UR PHaSE Learner

The UR PHaSE Learner integrates Maximum Entropy (MaxEnt) learning (Hayes and Wilson, 2008) with expectation-maximization (EM) learning (Jarosz, 2006a, 2015; Wang and Hayes, 2025) to jointly identify optimal URs for all morphemes in a given dataset and the constraint weights used in UR→SR mappings. Learning proceeds in four steps: (1) phonotactic learning to establish a best initial

guess of constraint weights, (2) n-disparity UR expectation maximization learning, in which candidate URs with up to n disparities are evaluated; (3) Likelihood Threshold evaluation, which assesses whether the n-disparity UR candidate sets from step (2) sufficiently explain the observed data, and (4) generation of n+1 disparity UR candidates, which expands the UR candidate space when the current UR candidates fail the likelihood threshold evaluation in step (3). Steps (2)-(4) are repeated until the Likelihood Threshold is satisfied for all morphemes.

3.1 Step (1): Phonotactic Learning

The UR PHaSE learner first learns constraint weights to model phonotactic patterns before proceeding to consider URs for individual morphemes. Phonotactic learning here follows the standard error-driven learning approach to phonotactic learning undertaken by MaxEnt learning models, similar to Hayes and Wilson (2008). The learner takes as input a list of surface forms and pre-defined markedness constraints initialized at a neutral weight (e.g., set at 50 with constraint bounds between [0, 100]). Using gradient descent, the learner adjusts constraint weight values to minimize a loss function made up of two components: the negative log likelihood of the observed data and an L2 Gaussian prior (cf., O'Hara 2017) that favors high-weighted markedness constraints, thus increasing grammar restrictiveness. The output of the phonotactic stage is an array of optimized constraint weights that capture the surface phonotactic patterns in the data while maximally satisfying the prior's preference for a restrictive grammar.

3.2 Step (2): n-disparity UR EM Learning

After the phonotactic stage, learning progresses to a morphologically-aware stage in which the UR PHaSE learner jointly acquires a probability distribution over the set of possible URs and the optimal constraint weights that maximize the likelihood of the surface data. Expectation maximization (EM) is employed to acquire UR probability distributions and constraint weights (Dempster et al., 1977). Jarosz (2006a,b, 2009, 2010, 2015) demonstrates that EM learning within a probabilistic OT framework can successfully learn a broad range of hidden linguistic structure, including URs, and the EM learning principles developed in that work provide the inspiration for the approach taken here (see also Wang and Hayes 2025).

Morphologically-aware learning involves the simultaneous acquisition of URs for individual morphemes and the constraint weights that map those URs to their observed surface realizations (SRs) across contexts. This creates a classic chicken-and-egg problem. In order to correctly identify the UR for a morpheme that maximizes the likelihood of its SRs, the constraint weights governing the UR→SR mappings must already be known. At the same time, in order to arrive at the optimal constraint weights that maximize the likelihood of a UR→SR mapping, it is also crucial to know a morpheme's UR. The EM algorithm provides a solution to this issue by breaking the learning problem into two iterated steps.

3.2.1 Expectation Step (E-step)

The goal of the E-step is to update prior beliefs about a hypothesis given new evidence. For an individual morpheme, the initial hypothesis of the learner is that the probability of candidate URs is equivalent across the entire set for that morpheme. The initial new evidence is comprised of the combination of constraint weights learned during the phonotactic stage and the observed SRs of the morpheme.

Under these conditions, the learner computes posterior probabilities over the space of candidate URs for each morpheme, updating the probability that each UR is the correct UR for a given SR, using Bayes' Theorem in equation 1.

$$\underbrace{\mathbb{P}(UR_i \mid SR_j)}_{Posterior} = \frac{\overbrace{\mathbb{P}(SR_j \mid UR_i)}^{Likelihood}\ \overbrace{\mathbb{P}(UR_i)}^{Prior}}{\underbrace{\sum_{UR_k \in S} \mathbb{P}(SR_j \mid UR_k)\,\mathbb{P}(UR_k)}_{Marginal}} \quad (1)$$

Bayes' theorem provides a principled mechanism for updating the learner's uncertainty over which UR is the correct representation for a given morpheme. With a set of observed SRs and fixed constraint weights, the posterior probability of a candidate UR can be computed by multiplying the joint likelihood of

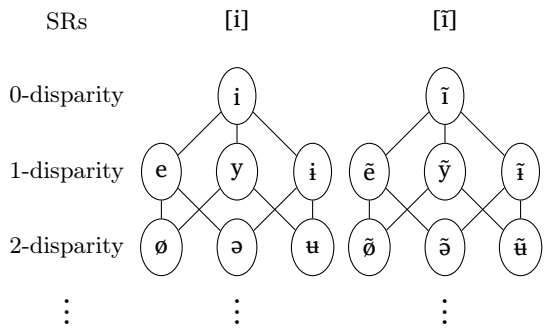

Figure 1: Lattice for a structured search space of candidate URs for a single morpheme realized contextually as [i] and [ĩ]. Visualized disparities only include the features [HIGH], [RD], and [FRONT].

observing all SRs of a morpheme given that UR by the prior probability assigned to the UR, and then normalizing by the marginal probability of the surface forms across all candidate URs for the morpheme. The UR's prior probability is initially uniform across candidate URs before any learning has taken place and is re-estimated in the M-step as the average posterior responsibility assigned to that UR across the SRs of the morpheme.

Crucially, the size of a morpheme's UR candidate space is what typically results in failures during UR learning (Wang and Hayes, 2025, p.33-35). As increasingly abstract URs are permitted, the hypothesis space size explodes, making traversal across the candidate set costly and leading to learning failures (Jarosz, 2019, p.78-79).

One of the main insights of the UR PHaSE learner is that the UR hypothesis space is structured based on the number of disparities in a UR→SR mapping for each morpheme (cf., Tesar 2014, 2016). This structure is illustrated by the lattice for a hypothetical morpheme with surface forms [i] and [ĩ] and possible disparities for the features [FRONT], [HIGH], and [RD] in Figure 1. As Tesar (2014, p.21) defines it, a featural disparity in a UR→SR mapping occurs when corresponding input-output segments differ in their value for a feature.[1]

The position of a UR in a morpheme's UR hypothesis space is determined by the disparity distance between that UR and the closest SR. URs that are realized faithfully in at least one allomorph count as 0-disparity URs and reside at the top of the UR space. 1-disparity URs, on the other hand, can be reached by flipping exactly one feature value in one segment for at least one of the SRs. 2-disparity URs are those that can be reached by flipping exactly two feature values—either in one segment or across two segments—for at least one of the SRs. URs with more than two disparities (not depicted in Figure 1) are positioned on the appropriate disparity level.

Importantly, while the feature flips resulting in the level 2 candidate URs in Figure 1 are implemented on a single segment, this does not have to be the case. If a morpheme is constructed of multiple segments, there must be exactly two combined feature flips across all segments for a UR to be considered a 2-disparity UR.

Underlying the motivation to structure the UR candidate space based on disparities is the assumption that learners inherently disprefer differences in UR→SR mappings (Prince and Smolensky, 1993/2004). That is, the learner assumes morphemes will be represented concretely unless evidence suggests otherwise. The framework of the UR PHaSE learner incorporates this assumption into its algorithm by considering potential UR candidates in batches, beginning with all 0-disparity candidates and only considering additional disparity candidates if the likelihood of the data remains low using URs with the current number of permitted disparities.

Note that nothing constrains the number of possible disparities a candidate UR can take, which means the entire hypothesis space will be quite large for most morphemes. Nevertheless, the UR PHaSE learner's algorithm takes advantage of the structure of the space to avoid searching it exhaustively.

3.2.2 Maximization Step (M-step)

After defining the UR candidate space for each morpheme and calculating the posterior probability distributions of each UR candidate space, the learner moves to the M-step, seeking to maximize the likelihood of the observed data (Jarosz, 2006a,b). To do so, the learner updates the constraint weights (**w**), us-

[1] I set aside disparities caused by insertion and deletion in this paper due to space constraints, but future work will incorporate them into the learner.

ing gradient descent to minimize a loss function made up of the negative log likelihood of all n observed SRs (2a) and an L2 Gaussian prior (2b) that maximizes grammar restrictiveness by preferring high-weighted markedness constraints and low-weighted faithfulness constraints (O'Hara, 2017; Paramore, 2025).[2,3]

$$\text{NLL} = -\ln\left[\prod_{i=1}^{n}\left(\mathbb{P}[SR_i \mid (\mathbf{w}, \boldsymbol{\pi}_{\mathbf{post}})]\right)\right] \quad (2a)$$

$$\text{L2 Prior} = \sum_{w_i \in \mathbf{w}} \frac{(w_i - c_i)^2}{\sigma_i^2} \quad (2b)$$

Once the optimal constraint weights are discovered in the M-step, the learner returns to the E-step, updating these posterior probability distributions based on the newly optimized constraint weights. Iteration between these steps continues until the learner converges on stable constraint weight values and UR probability distributions.

3.3 Step (3): Likelihood Threshold Evaluation

Convergence on a set of optimal UR distributions and constraint weights in the EM learning step is not equivalent to solving the learning problem. Rather, it means that the learner has arrived at a locally optimal solution with the current UR candidates under consideration for each morpheme, but it does not mean this solution is adequate. To determine adequacy, step (3) evaluates the likelihood of each individual SR in the observed data using the UR probabilities and constraint weights learned in step (2). If all SRs exceed a .95 likelihood threshold, the learner—satisfied that the current URs generate the observed data with sufficiently high probability—stops its search for a more optimal model. If, however, any single SR exhibits less than .95 likelihood, the learner has failed to meet the Likelihood Threshold and continues iterating through the learning pathway (cf., Belth 2026).

Establishing a Likelihood Threshold at which the learner is satisfied with its learned representations is a defining feature of the UR PHaSE learner, and is similar to how humans approach decision making. Rather than examine all possible candidates in a hypothesis space exhaustively, humans almost always settle on a solution once it meets some threshold that satisfies their expectations (Simon, 1955, 1956). Of course, it is certainly true that acceptability thresholds vary from individual to individual and decision to decision. In this sense, the .95 likelihood threshold offered here serves as a stand-in for individual variation of what a satisfactory solution may be.

3.4 Step (4): UR Candidate Expansion

If any surface form falls below the .95 threshold in step (3) using n-disparity URs, the UR PHaSE learner will expand the UR hypothesis space for all morphemes by introducing n + 1-disparity UR candidates. Crucially, even at low-disparity levels, the unrestricted generation of UR candidates can cause the hypothesis space to grow too large to support effective learning.

To avoid this combinatorial explosion, the UR PHaSE learner does not generate all logically possible n + 1-disparity URs. Instead, it exploits information already present in the grammar: only features associated with markedness constraints that (i) are violated by at least one surface representation and (ii) currently exhibit low (i.e., non-maximal) weights are eligible to be flipped to generate new UR candidates.

This targeted strategy equips the learner with a principled means of expanding the UR search space. Rather than indiscriminately introducing all higher-disparity candidates, the learner proposes only those URs that introduce additional UR→SR mapping candidates that make violations of these markedness constraints informative for learning. In this way, the *selective* introduction of UR candidates at a single level of increased abstraction provides the learner with precisely the evidence needed to revise its grammatical commitments, while keeping the hypothesis space

[2] See Hayes and Wilson (2008) for a detailed explanation of gradient ascent, the maximization counterpart to the gradient descent process used here.

[3] To maximize restrictiveness, c_i is set to 100 for markedness constraints and 0 for faithfulness constraints. σ_i^2 is a plasticity constant used to modulate how much deviations from ideal weights impact the value of the loss function.

i.	[CV]	stem1	ii. [CṼ-G̃Ṽ]	stem1-suff1
iii.	[CṼ]	stem2	iv. [CṼ-G̃Ṽ]	stem2-suff1
v.	[CVGṼN]	stem3		

Table 2: Punjabi surface form categories

tractable.

3.5 Iteration Until Convergence

After the UR candidate space is expanded in step (4), the UR PHaSE learner repeats the learning process in steps (2)-(4) until all morphemes exceed the Likelihood Threshold evaluation in step (3).

4 A Test Case: Punjabi Pre-N Vowels

This section demonstrates that the level of UR abstraction required for pre-N vowels in Punjabi is learnable under the architecture of the UR PHaSE learner. The model was implemented in Python. The code and data used in this section are available at github.com/joncparamore/UR_PHaSE_Learner.

Surface forms exemplified by the five word categories in Table 2 served as the input data for the learning model. The entire dataset is provided in the appendix. Six words were included from each of the five categories, resulting in a total of thirty forms. Along with the observed surface form of each word, the learner was also provided with morpheme boundary information.[4] Categories in Table 2.i-ii demonstrate the nasal harmony pattern triggered by contrastive nasal vowels, and the patterns in Table 2.iii-iv illustrate that vowel nasality is contrastive. Finally, the category in Table 2.v demonstrates that pre-N vowels are nasalized but do not trigger harmony.

4.1 Step (1): Phonotactic Learning

The ten constraints listed in Table 3, along with their definitions (provided in the appendix), were also supplied to the learner. All constraint weights were initialized at the midpoint value of 50 at the outset of step (1). ID[NAS]/_V is a unique constraint in that it is a contextual faithfulness constraint (Hauser and Hughto, 2020) that penalizes changes to a segment's [NAS] feature value when directly preceding a vowel specified as [-NAS] in the *input*. As is well known, parallel versions of Optimality Theory are unable to capture many opaque phonological processes (McCarthy, 2000, 2007), including the opaque counterfeeding pattern in which pre-N vowels are nasalized on the surface but do not trigger harmony in Punjabi. To address this, contextual faithfulness constraints reference a segment's input context to explain why vowels and glides preceding pre-N vowels do not undergo nasalization.

Constraint	Type	initial w	final w
Sprd-L[nas]	mark.	50.00	0.13
*NasObs	mark.	50.00	100.00
*NasG	mark.	50.00	0.47
*NasV	mark.	50.00	0.00
*VN	mark.	50.00	100.00
ID[NAS]	faith.	50.00	0.00
IDFin[NAS]	faith.	50.00	0.00
ID[NAS]/_V	contfaith.	50.00	0.00
ID[RD]	faith.	50.00	0.00
*LowRD	mark.	50.00	100.00

Table 3: Constraint weights with their initial and optimized weights in the phonotactic stage.

The weights learned in the phonotactic stage are provided in the final column of Table 3 and serve as the input to the morphologically-aware learning stage that follows. From a cursory examination, it is clear that the phonotactic grammar correctly bans nasal obstruents, oral vowels before nasal consonants, and low round vowels—none of which occur in Punjabi—by maximizing the associated markedness constraints. Also, because the grammar does not consider URs during the phonotactic stage, all faithfulness constraints were reduced to zero to allow for a maximally restrictive grammar. *NasG must have a low weight during the phonotactic stage because nasal glides are phonotactically licit in Punjabi (e.g., [sãṽã] 'breaths'). Likewise, because pre-N forms do not exhibit nasal harmony (e.g., [tɑʋə̃n] 'penalty'), Sprd-L[nas] must also be assigned a low weight in the phonotactic stage. Otherwise, the grammar would impose nasal harmony in all applicable environments, including in pre-N forms.

[4] See Wang and Hayes (2025) for a method of learning morpheme boundaries during morphophonemic learning.

4.2 Step (2): 0-disparity UR EM Learning

After phonotactic learning, step (2) uses the Expectation–Maximization (EM) algorithm to search for the combination of URs and constraint weights that maximize the likelihood of the observed surface forms.

Although the learner optimizes UR probability distributions for all morphemes across the dataset, I report updates for only three representative word forms here: [sã] 'breath', [sãṽã] 'breaths', and [tɑvãɳ] 'penalty'. Together, these forms are sufficient to diagnose whether the learner has acquired (i) nasal harmony triggered by contrastive nasal vowels and (ii) the absence of nasal harmony when vowel nasality is derived in pre-N contexts.

In this round of EM learning, the UR PHaSE learner constructs the UR candidate space for each morpheme based solely on the surface alternations it has encountered. The 0-disparity UR candidates for the morphemes 'breath' and 'penalty' (Table 4) are all identical to at least one SR of their respective morpheme.

The results of EM learning using only 0-disparity URs are given in Table 4. Note that the final posterior UR probabilities the learner converged on remain unchanged from their initial uniform distributions. Additionally, the likelihood of observing [tɑvãɳ] 'penalty' is 1, indicating that nasal harmony is correctly not expected in pre-N forms. For the two realizations of 'breath' [sɑ]∼[sãṽã], however, the learner is unable to correctly model nasal harmony given the current UR options, so instead optimizes the likelihood as much as possible by distributing UR probabilities evenly across the two candidate URs. This results in approximately 50% likelihood of observing the correct nasality patterns of 'breath'.

The reason for the learning failure is straightforward: given only concrete forms as UR candidates, the learner cannot find constraint weights that induce nasal harmony in 'breath' while simultaneously suppressing it in the pre-N form 'penalty'.

4.3 Step (3): Likelihood Threshold Eval. for 0-disparity URs

As is already evident from the sampling of SRs in Table 4 above, the six singular surface forms with an oral vowel in the stem (like [sɑ] 'breath') and the six surface forms with those same stems in the plural (like [sãṽã] 'breaths') all fail to meet the Likelihood Threshold of .95 when using only 0-disparity URs.

Constraint	Type	initial w	final w
Sprd-L[nas]	mark.	0.13	62.92
*NasObs	mark.	100.00	100.00
*NasG	mark.	0.47	60.24
*NasV	mark.	0.00	23.16
*VN	mark.	100.00	100.00
ID[NAS]	faith.	0.00	83.44
IDFin[NAS]	faith.	0.00	89.46
ID[NAS]/_V	contfaith.	0.00	100.00
ID[RD]	faith.	0.00	6.29
*LowRD	mark.	100.00	100.00

Morpheme	UR candidate	Prior	Posterior
'breath'	/sɑ/	0.5	0.5
	/sã/	0.5	0.5
'penalty'	/tɑvãɳ/	1.0	1.0

SR	Likelihood
[sɑ]	0.50
[sãṽã]	0.48
[tɑvãɳ]	1.00

Table 4: Constraint weights, UR probabilities, and SR likelihoods after 0-disparity EM learning.

4.4 Step (4): UR Candidate Expansion

Unsatisfied with this learning result, the UR PHaSE learner proceeds to consider more abstract 1-disparity URs as potential additions to the UR candidate space. Importantly, 1-disparity URs are generated for all morphemes the learner encounters, not only those that fail to meet the likelihood threshold. Expanding the hypothesis spaces of all morphemes prevents the learner from cementing the 0-disparity analysis into part of the lexicon while allowing abstraction elsewhere, which can distort subsequent EM updates.

For instance, at the end of EM learning with concrete 0-disparity URs in section 4.2, the learner is unable to simultaneously model the presence of nasal harmony in alternations like [sɑ∼sãṽã] 'breath'∼'breaths' and the lack of harmony in pre-N forms like [tɑvãn] 'penalty'. As a compromise, the learner lowers the weights of the constraints enforcing harmony and, in doing so, predicts the absence of harmony in pre-N forms with perfect probability, as shown in Table 4. If only the candi-

date spaces of morphemes with low likelihood were expanded, the learner would preserve the incorrect URs assigned to pre-N forms during the 0-disparity stage, thereby preventing it from ever arriving at the correct abstract, underlyingly oral URs for pre-N forms.

In generating 1-disparity URs, the learner first examines the violation profiles of all markedness constraints with respect to the observed surface forms at the conclusion of 0-disparity learning (Table 4). If a markedness constraint exhibits a non-maximal weight and is violated at least once by an observed SR, the feature associated with that markedness constraint is eligible to be flipped to construct new 1-disparity URs. A violated markedness constraint that fails to reach its upper bound indicates that weight re-optimization alone cannot explain the relevant patterns across SRs under current UR assumptions. The learner, therefore, treats such intermediate weights as evidence that the UR may be misspecified along the associated feature and permits UR abstraction in order to test whether an alternative UR yields higher likelihood.

In fact, only SPRD-L[NAS], *NASG, and *NASV satisfy both criteria, rendering [NAS] as the only feature eligible to be flipped to generate new 1-disparity candidate URs. Crucially, *LOWRD both reaches the upper weight bound of 100 and is not violated by any observed SRs. Consequently, the learner determines that introducing URs with a [RD] disparity is unlikely to improve likelihood. Thus, the expansion of the UR candidate space with the addition of 1-disparity URs will be relatively modest, only introducing URs that flip the [NAS] feature on eligible segments.

Table 5 shows the 1-disparity URs added to the hypothesis spaces of the morphemes 'breath' and 'penalty'. The two morphemes differ sharply in how their UR candidate spaces expand. For 'breath', the candidate space remains unchanged because flipping the [NAS] feature on the vowel yields URs that already belong to the 0-disparity level. Furthermore, nasalized [s̃] is unattested in Punjabi, leaving the learner without a phonological category that could support representations such as /s̃ɑ̃/ or /s̃ɑ/. In contrast, the UR candidate space for 'penalty' expands from the single SR to four potential URs. By flipping the [NAS] feature on all segments for which a corresponding segmental category exists, the learner introduces a small but analytically consequential set of 1-disparity URs.

Morpheme	UR Candidate Space
'breath'	/sɑ/ /sɑ̃/
'penalty'	/tɑʊɑ̃ɳ/ /tɑʊ̃ɑ̃ɳ/ /tɑ̃ʊɑ̃ɳ/ /tɑʊɑɳ/

Table 5: UR candidates for 'breath' and 'penalty', expanded to include all possible 1-disparity URs generated by flipping the [NAS] feature.

4.5 Step (2): 1-disparity UR EM Learning

Equipped with the newly expanded hypothesis spaces, the learner returns to step (2) to reinstantiate EM learning, the results of which are shown in Table 6. As is evident, the combination of learned weights and UR probability distributions enables the learner to predict the observed SRs with high degrees of likelihood. This is achieved by assigning all of the UR probability for 'breath' to oral /sɑ/ and deriving nasality on the stem by way of nasal harmony triggered by the vowel in the plural suffix. Conversely, the learner also correctly assigned all of the probability to the abstract UR /tɑʊɑɳ/ for 'penalty', thereby enabling the pre-N vowel to surface as nasal (due to high weight *VN) while not predicting nasal harmony on these forms (due to high weight on the contextual faithfulness constraint, ID[NAS]/_V).

4.6 Step (3): Likelihood Threshold Eval. for 1-disparity URs

With the addition of 1-disparity URs, the learner converged on a combination of constraint weights and UR probabilities that yielded a likelihood greater than .95 for each SR, thus successfully acquiring a grammar and lexicon that accounts for the data.

5 Discussion & Future Research

This paper introduced the UR Progressive Hypothesis Space Expansion (UR PHaSE) learner, a novel learning architecture designed to reconcile the empirical need for abstract URs with the learnability concerns raised by permitting increased UR abstraction. Rather than evaluating an effectively unbounded

Constraint	Type	initial w	final w
Sprd-L[nas]	mark.	0.13	48.45
*NasObs	mark.	100.00	100.00
*NasG	mark.	0.47	48.45
*NasV	mark.	0.00	0.00
*VN	mark.	100.00	100.00
ID[NAS]	faith.	0.00	0.00
IDFin[NAS]	faith.	0.00	100.00
ID[NAS]/_V	contfaith.	0.00	100.00
ID[RD]	faith.	0.00	6.28
*LowRD	mark.	100.00	100.00

Morpheme	UR candidate	Prior	Posterior
'breath'	/sɑ/	0.5	1.00
	/sɑ̃/	0.5	0.00
'penalty'	/tɑvɑ̃ɳ/	0.25	0.00
	/tɑvɑɳ/	0.25	1.00
	/tɑ̃vɑ̃ɳ/	0.25	0.00
	/tɑ̃vɑ̃ɳ/	0.25	0.00

SR	Likelihood
[sɑ]	1.00
[sɑ̃ṽɑ̃]	0.96
[tɑvɑ̃ɳ]	1.00

Table 6: Constraint weights, UR probabilities, and SR likelihoods after 1-disparity EM learning.

UR hypothesis space in parallel, the learner structures that space by disparity count and searches it serially, beginning with fully concrete (0-disparity) URs and expanding to include more abstract candidates only when the current level fails to meet a Likelihood Threshold.

The UR PHaSE learner was applied to a Pakistani Punjabi vowel nasality pattern that requires abstract URs to account for pre-N vowels, showing that 0-disparity EM learning predictably fails to acquire the Punjabi nasality patterns because concrete URs cannot simultaneously induce harmony in contrastive-nasal contexts while suppressing it in pre-N contexts. The learner then selectively expands the UR candidate spaces using information already present in the learned grammar to generate only those higher-disparity URs that are plausibly informative, keeping hypothesis space growth tractable. More specifically, when learning fails at lower levels of abstraction, the learner uses the constraints that highlight uncertainty/inconsistency in the grammar (i.e., constraints with non-maximal weights) to collect a set of features that could be changed in the representations to resolve this uncertainty. With this targeted expansion, the learner successfully converges on the minimally abstract URs required by the Punjabi pattern, thereby demonstrating that restrictedly abstract URs are learnable when hypothesis space expansion is progressive, structured, and threshold-governed.

The UR PHaSE learner developed in this paper shows that constraints on UR abstraction need not be imposed as an external stipulation. Building on the conception of structuring UR hypothesis spaces by disparity count proposed by Tesar (2014, 2016), it avoids evaluating an unbounded set of candidates in parallel by generating and considering URs in batches, based on their disparity distance. This architecture naturally restricts abstraction because minimally abstract URs are always preferred if they are sufficient. At the same time, the learner addresses search space concerns presented by UR abstraction by expanding the hypothesis space in a targeted way. As the Punjabi example shows, this combination allows the learner to arrive at the highly abstract URs needed for pre-N vowels.

Several directions for future work follow from these results. First, the present paper demonstrates success on a single pattern that motivates UR abstraction, Pakistani Punjabi vowel nasality. The learner needs to be tested on additional case studies that have been argued to require abstract URs (e.g., O'Hara (2017) and Wang and Hayes (2025)). Second, because the UR PHaSE learner does not impose an explicit upper bound on abstraction, it predicts that if a pattern truly requires 2-disparity or 3-disparity URs, the learner will eventually incorporate those URs into the candidate space. At the same time, after the learner has climbed deep enough into the disparity ladder, the UR candidate space may still grow too large, ultimately yielding prohibitive compute time or convergence failures. If so, then we should not expect to find stable phonological systems in which very high degrees of UR→SR mapping disparities are analytically necessary. Exploring where this boundary lies could clarify how far abstraction can plausibly extend under realistic learning constraints and why, over time, learners tend to reanalyze highly abstract URs as concrete (Kiparsky, 1973; Kuo, 2024a,b).

References

Caleb Belth. 2026. A learning-based account of phonological tiers. *Linguistic Inquiry*, 57(2):229--265.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. In *Royal Statistical Society*.

Madeline Gilbert. 2023. Testing for underlying representations: Segments and clusters in Sevillian Spanish. *Natural Language & Linguistic Theory*, 42:493--531.

Ivy Hauser and Coral Hughto. 2020. Analyzing opacity with contextual faithfulness constraints. *Glossa: a journal of general linguistics*, 5(1):1--33.

Bruce Hayes and Colin Wilson. 2008. A maxiumum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3):379--440.

Gaja Jarosz. 2006a. *Rich lexicons and restrictive grammars - Maximum likelihood learning in Optimality Theory*. Ph.D. thesis, Johns Hopkins University.

Gaja Jarosz. 2006b. Richness of the base and probabilistic unsupervised learning in Optimality Theory. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology*, pages 50--59.

Gaja Jarosz. 2009. Restrictiveness and phonological grammar and lexicon learning. In *Proceedings of the 43rd Annual Meeting of the Chicago Linguistics Society*, volume 43, pages 125--134.

Gaja Jarosz. 2010. Implicational markedness and frequency in constraint-based computational models of phonological learning. *Journal of Child Language. Special Issue on Computational Models of Child Language Learning*, 37(3):565--606.

Gaja Jarosz. 2015. Expectation driven learning of phonology. University of Massachusetts manuscript.

Gaja Jarosz. 2019. Computational modeling of phonological learning. *Annual Review of Linguistics*, 5:67--90.

Michael Kenstowicz and Charles Kisseberth. 1977. *Topics in Phonological Theory*. Academic Press.

Paul Kiparsky. 1973. *Abstractness, opacity, and global rules*, pages 57--86. Tokyo: TEC.

Jennifer Kuo. 2024a. Phonetic naturalness in the reanalysis of Samoan thematic consonant alternations. *Journal of Phonetics*, 107:1--18.

Jennifer Kuo. 2024b. Phonological reanalysis is guided by markedness: the case of malagasy weak stems. *Phonology*, 41(3):1--35.

John J. McCarthy. 2000. Harmonic serialism and parallelism. In *Proceedings of the 30th meeting of the North East Linguistic Society*, pages 501--524.

John J. McCarthy. 2007. *Hidden Generalizations: Phonological Opacity in Optimality Theory*. Sheffield: Equinox.

Charlie O'Hara. 2017. How abstract is more abstract? learning abstract underlying representations. *Phonology*, 34:325--345.

Jonathan Charles Paramore. 2025. Learning covert URs via disparity minimization. In *Eighth Annual Meeting of the Society for Computation in Linguistics (SCiL)*, volume 8.

Jonathan Charles Paramore and Ryan T. Bennett. 2025. Covert URs: evidence from nasalization in Pakistani Punjabi. Manuscript, UC Santa Cruz.

Alan Prince and Paul Smolensky. 1993/2004. *Optimality Theory: constraint interaction in Generative grammar*. Malden, Mass: Blackwell Publishers.

Herbert A. Simon. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99--118.

Herbert A. Simon. 1956. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129--138.

Bruce Tesar. 2014. *Output-Driven Phonology*. Cambridge: Cambridge University Press.

Bruce Tesar. 2016. Phonological learning with output-driven maps. *Language Acquisition*, 24(2):148--167.

Rachel Walker. 2003. *Reinterpreting transparency in nasal harmony*, pages 37--72. Amsterdam: John Benjamins.

Yang Wang and Bruce Hayes. 2025. Learning phonological underlying representations: the role of abstractness. *Linguistic Inquiry*.

A Appendix

(1) Surface forms provided to the learner

i. **CV**: [sɑ], [do], [to], [pe], [ge], [si]

ii. **CṼ-G̃Ṽ**: [sɑ̃ṽɑ̃], [dõṽɑ̃], [tõṽɑ̃], [pẽṽɑ̃], [gẽṽɑ̃], [sĩṽɑ̃]

iii. **CṼ**: [gɑ̃], [tʃʰɑ̃], [kõ], [ʒɑ̃], [kɑ̃], [põ]

iv. **CṼ-G̃Ṽ**: [gãṽã], [tʃʰãṽã], [kõṽã], [ʒãṽã], [kãṽã], [põṽã]

v. **CVGṼN**: [tɑvãɳ], [prəvãɳ], [tejãɳ], [ʒəvãɳ], [ɑvãm], [sijãɳ]

(2) Constraint Definitions

i. SPRD-L[NAS] (cf. Walker, 2003, p.47)
For every occurrence of a [+NAS] feature in a prosodic word, if that [+NAS] feature is dominated by some segment, assign a violation for every segment to the left of that segment in the prosodic word that does not dominate the [+NAS] feature.

ii. *NASOBS (Walker, 2003, p.51)
Assign a violation for every obstruent that dominates a [+NAS] feature.

iii. *NASG (Walker, 2003, p.51)
Assign a violation for every glide that dominates a [+NAS] feature.

iv. *NASV (Walker, 2003, p.51)
Assign a violation for every vowel that dominates a [+NAS] feature.

v. *VN
Assign a violation for every vowel that dominates a [-NAS] feature when directly preceding a nasal consonant.

vi. ID[NAS]
For every segment, *A*, assign a violation if the output value for the [NAS] feature dominated by *A* does not match the input value for the [NAS] feature dominated by *A*.

vii. IDFIN[NAS]
For every segment, *A*, assign a violation if the output value for the [NAS] feature dominated by *A* does not match the input value for the [NAS] feature dominated by *A* in the final syllable of a prosodic word.

viii. ID[NAS]/_V
Let *A* be a segment that occurs before an oral vowel, _V, in the input. Assign one violation if the output correspondent of *A* does not have the same specifications for [NAS] as *A*.

ix. ID[RD]
For every segment, *A*, assign a violation if the output value for the [RD] feature dominated by *A* does not match the input value for the [RD] feature dominated by *A*.

x. *LOWRD
Assign a violation for every vowel that dominates a [rd] feature and a [LOW] feature simultaneously.