Learning Covert URs via Disparity Minimization

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

Abstract

When considering the acquisition of underlying representations (URs), two common challenges are often levied against the inclusion of abstract URs in phonological theory: (1) permitting abstract URs causes the search space of potential URs to grow to a computationally intractable degree, and (2) learners have no recourse through which to prefer minimally abstract URs over increasingly abstract alternatives when both types of URs model the data with equal success. This paper directly addresses the second issue by implementing a MaxEnt learner equipped with a bias that penalizes disparities between UR inputs and their corresponding outputs. By favoring mappings with minimal divergence, the bias generates a preference for minimally abstract URs when competing candidates perform equally well in modeling the data. In addition, the paper proposes a conceptual framework for addressing the first issue, in which the space of potential URs is organized so that candidates are considered serially, beginning with those that exhibit the fewest disparities. This method offers a potential strategy for avoiding the added compute time introduced by permitting UR abstraction.

1 Introduction

A subject of significant debate since the advent of generative phonology concerns the level of abstraction that underlying representations (URs) are permitted to assume (Kenstowicz and Kisseberth, 1979). Classic generative phonology holds the rather strong position that a UR can be completely *covert* in relation to all of its allomorphs, never showing its true identity in surface forms. However, from a learning perspective, permitting this level of abstraction poses serious challenges. One of the most compelling objections is that covert URs render the learning problem intractable. Two key difficulties arise. First, the space of potential URs that a learner must consider becomes prohibitively large. When highly abstract URs are allowed, the search space expands dramatically, exceeding what can feasibly be explored in its entirety by a learner (Albright, 2002; Jarosz, 2015, 2019; Wang and Hayes, 2025).

Most models attempt to solve this issue by curtailing the level of abstraction URs can take, in essence shrinking the search space to a manageable size. For instance, Wang and Hayes (2025) constrain the search space by restricting the abstractness of candidate URs using a hierarchy of representational abstraction defined in Kenstowicz and Kisseberth (1977, ch.1). The model is impressive and successfully accounts for analyses at various levels of abstraction, but it fails to account for datasets requiring covert URs, like the Punjabi nasality pattern considered in this paper.

The second issue that arises when learning covert URs is that the learner has no means through which to prefer a less abstract UR over a highly abstract UR if both representations succeed in modeling the data. One particularly promising approach aimed at alleviating this computational burden is outlined in O'Hara (2017) with the use of a Maximum Entropy (MaxEnt) grammar called MaxLex. O'Hara provides compelling evidence from Klamath showing that a stem-final [i]-[ø] alternation in words like [ʔeːw-a] 'is deep' ∼ [ʔeːwitkʰ] 'deep' cannot be captured by either epenthesis or deletion but instead requires a covert UR, /e/, that deletes when not in the initial syllable, unless deletion would produce an illicit consonant cluster, in which case /e/ is raised to [i]. Importantly, /e/ is covert in

the stem-final position of stems like /ʔeːwe/ because it never surfaces in any allomorph. Moreover, O'Hara demonstrates that MaxLex has an *emergent* preference for minimally abstract URs, driven by an L2 Gaussian Prior that attempts to minimize increases in the weights of faithfulness constraints.

In this paper, I primarily address how the learner might come to prefer minimal UR abstraction. I first show that MaxLex *fails* to prefer minimally abstract URs over increasingly abstract alternatives for a set of non-alternating pre-nasal vowels in Pakistani Punjabi (Paramore, 2023). This failure arises because both the minimally abstract UR and more abstract alternatives provide equally accurate accounts of the data and require identical changes in faithfulness constraint weights to do so. As a solution, I propose an updated MaxLex learner equipped with a disparity bias that penalizes changes in UR→SR mappings. The effect of this bias is that, if two URs model a set of data equally well and do not differ in the minimization of the MaxLex L2 prior, the learner selects the UR that generates the minimum number of disparities. In addition to creating a preference for minimal UR abstraction, this disparity bias has potential to provide a mechanism through which the learner can efficiently search the space of potential URs without needing to stipulate its contents, as discussed in section 6.

2 MaxLex

The basic learning procedure taken by MaxLex is similar to other MaxEnt learning models (e.g. Hayes and Wilson, 2008; Pater et al., 2012; Wang and Hayes, 2025). Two general stages characterize the process. In the first stage, the learner is oblivious to morphological alternations and paradigmatic relations, and, as a consequence, the identity of underlying forms and mappings from those underlying forms to surface realizations is not considered. Instead, the learner has been confronted with a wealth of linguistic data and focuses on acquiring fluency in language-specific phonotactics, an aspect of the grammar that remains unchanged regardless of what the underlying forms turn out to be.

In computational terms, at the outset of the phonotactic stage, MaxLex is fed a batch of data, a set of constraints with intermediate weights (e.g., 50), and the parameters for what constitutes a violation. Equipped with this information, the learner uses gradient descent optimization to minimize an objective function (in this case, the negative log-likelihood of the data) by adjusting the constraint weights appropriately until it arrives at the minimum possible value. A grammar with a 100% probability of producing the observed data will have an objective function value of zero, but a grammar with only a 50% probability of producing the observed data will result in a much higher objective function value.

In the second stage of learning, MaxLex becomes morphologically aware, understanding that words are constructed from morphemes, and those morphemes sometimes appear in phonologically distinct ways, depending on the context. For instance, during the phonotactic stage, the learner ignores the morphological relationship between the Punjabi words [sɑɑ] 'breath' and [sɑ̃ɑ̃ʊ̃ɑ̃ɑ̃] 'breaths', focusing only on phonotactic well-formedness. In the morphologically aware stage, however, the learner has discovered that the same morpheme for 'breath' occurs in both words and seeks to assign a single UR that can map to both of the observed forms. As such, the learner is confronted with a more complex learning problem in which it must work to determine what combination of constraint weights and underlying form probabilities maximizes the likelihood of observing the data to which it has been exposed (Jarosz, 2006a,b).

A crucial aspect of the morphologically aware learning stage that MaxLex capitalizes on is the way in which abstraction is mitigated in the choice of potential URs. Specifically, the objective function in MaxLex is constructed from the negative log-likelihood of the data plus the value of an L2 Gaussian Prior that prefers to use constraints active elsewhere in the grammar to account for abstract phonological patterns rather than altering the weight of novel constraints to accomplish the same task.[1] The negative log-

[1] Both Pater et al. (2012, p.66) and Wang and Hayes

likelihood (NLL) of a dataset, given in equation 1, is calculated by determining the combination of constraint weight (**w**) and UR probability ($\boldsymbol{\pi}$) values that maximize the likelihood (thereby minimizing the NLL) of observing a set of observed words (O_i - O_n).

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

To increase grammar restrictivity, the L2 Gaussian prior shown in equation 2 inherently favors markedness constraints with maximum weights of 100 and faithfulness constraints with minimum weights of zero. This bias is implemented by taking the squared difference of actual weight values (w_i) from their ideal weight (c_i).[2] If, however, the language data confronted by the learner indicates that different constraint weights would improve the success of the grammar in modeling the data (i.e., sufficiently lowering the NLL), these biases can be overcome. Thus, if a faithfulness constraint is given a non-zero weight to model some phonotactic pattern in the first stage of language learning, that same constraint will be preferred over a novel constraint with a zero weight to model another pattern concerning underlying forms, assuming both constraints can account for the observed data equally well. This preference to use the already-active faithfulness constraint falls out from the fact that the MaxLex prior seeks to minimize deviations in constraint weights from their optimal values. Because of this, O'Hara argues that a segment's UR is naturally restricted in its potential for abstraction by this bias.

$$\mathcal{O}_{Lex}(\mathbf{w}, \boldsymbol{\pi}) = \textit{NLL} + \underbrace{\sum_{w_i \in \mathbf{w}} \frac{(w_i - c_i)^2}{\sigma_i^2}}_{\text{L2 Gaussian Prior}} \quad (2)$$

The success of MaxLex in learning covert URs is demonstrated by examining a stem-final [i]∼[ø] alternation in a set of Klamath verbs, which, as O'Hara (2017) shows, capitalizes on a faithfulness constraint that is active in another area of the grammar to account for the alternation. As O'Hara delineates in detail in his computational proof, Maxlex takes advantage of these faithfulness constraint weight differences when deciding upon the optimal covert UR. However, that same learning process used to constrain UR abstraction in the Klamath [i]∼[ø] alternation is unavailable for the URs of non-alternating pre-N vowels in Punjabi.

3 Pakistani Punjabi

Pakistani Punjabi is an Indo-Aryan language spoken by about 78 million people, primarily in the Punjab province of Pakistan (Bashir and Conners, 2019). Long vowels in Punjabi contrast in nasality, but this contrast is neutralized before nasal consonants (e.g., [tɑɑ] 'warmth' vs. [tãã] 'that' but [tããn] 'melody' vs. *[tɑɑn]). Additionally, Punjabi exhibits a process of nasal harmony, in which contrastive /ṼṼ/ vowels trigger the leftward spread of nasalization, with glides and vowels participating and other consonants acting as blockers, as shown in Table 1i. Pre-N vowels, on the other hand, surface as categorically nasalized and phonetically identical to contrastive /ṼṼ/ vowels, but they do not trigger nasal harmony (Table 1ii) (Paramore, 2023).

To account for the phonetic indistinguishability of /VVN/ and contrastive /ṼṼ/ vowels in terms of their nasality coupled with the fact that only contrastive /ṼṼ/ vowels trigger nasal harmony in Punjabi, /VVN/ vowels must be analyzed as underlyingly [-nas] without ever surfacing as such. In this view, the nasal harmony pattern in Punjabi serves as a straightforward example of counterfeeding opacity, in which underlyingly oral pre-N vowels undergo a predictable process of nasalization. Nevertheless, only underlying /ṼṼ/ vowels trigger nasal harmony. Harmony in Punjabi is thus sensitive to whether a vowel is underlyingly oral or nasal – even for vowels that are always *phonetically* nasal. This implies that /VVN/ vowels have abstract oral URs that are consistently distinct from their phonetic forms.

(2025, p.17, 34-35) incorporate similar biases favoring markedness constraints over faithfulness constraints.

[2] c_i is set to 100 for markedness constraints and zero for faithfulness constraints. O'Hara (2017) uses σ_i^2 as a plasticity constant (which he sets at 20 for markedness constraints and 25 for faithfulness constraints) to modulate how much deviations from ideal weights impact the value of the objective function.

i.	/saa-vãã/	→	[sããṽãã] 'breath-PL'
ii.	/taavaan/	→	[taavããn] 'penalty'

Table 1: Nasal Harmony in Punjabi.

i.	[saa]	'breath'	ii.	[sããṽãã]	'breaths'
iii.	[ʊʃaa]	'morning'	iv.	[ʊʃããṽãã]	'mornings'
v.	[gãã]	'cow'	vi.	[gããṽãã]	'cows'
vii.	[tʃʰãã]	'shade'	viii.	[tʃʰããṽãã]	'shades'
ix.	[taavããn]	'penalty'	x.	[prəvããn]	'accepted'

Table 2: Punjabi surface forms fed to MaxLex

4 MaxLex and Punjabi pre-N vowels

In attempting to learn the opaque nasalization patterns in Punjabi, MaxLex begins with an initial phonotactic learning stage. The observed data fed to the learner is given in Table 2. Forms 2i-iv show that underlyingly oral vowels are nasalized via nasal harmony when the appropriate suffix is attached (in this case, the plural marker). The forms in 2v-viii show the learner that a nasality contrast exists for vowels; otherwise, the learner may choose to analyze the vowels in 2i-ii as underlyingly nasal to explain the nasal harmony distinctions found between /VVN/ and contrastive /ṼṼ/ vowels. Finally, the forms in 2ix-x provide the learner with examples of the underapplication of nasal harmony for non-alternating /VVN/ vowels.

Individual Python scripts were developed for the phonotactic learning stage and morphologically aware learning stage to carry out the computational optimizations. The constraints used in the learner are provided in Table 3 with the initial weights set at 50, along with the weights acquired in the phonotactic learning stage in the rightmost column. Most of these constraints are straightforward, but a few merit further explanation.[3] First, as is well known, the standard parallel evaluation architecture of MaxEnt learners presents difficulty for the successful acquisition of opaque processes like nasal harmony in Punjabi (McCarthy, 2000, 2007). To handle this, I choose to analyze the nasality patterns using contextual faithfulness constraints (Hauser and Hughto, 2020), but other approaches capable of handling counterfeeding opacity in a parallel framework are equally viable. At its root, the contextual faithfulness constraint schema penalizes changes to a specified feature for a segment that occurs in a specified context in the input. The contextual faithfulness constraint relevant to the Punjabi nasalization data, Id[nas]/_V, penalizes changes in nasality to a segment occurring before a vowel that is oral in the input. When high-ranked, this constraint precludes underlying oral vowels – as /VVN/ vowels are proposed to be here – from continuing the transmission of nasal harmony to its immediately preceding segment.

Another important note is the inclusion of Id[rd] and *LowRd in the constraint set. For reasons that will become clearer when discussing the updated learning algorithm in section 5, I provide the learner with two potential covert URs to choose between. The restrictedly abstract and intuitively most appealing covert UR for a /VVN/ vowel like [ãã] in [taavããn] is /aa/. /aa/ possesses an identical feature set to [ãã] except for one disparity: nasality. Because nasality is the key underlying feature that results in distinct harmony patterns for /VVN/ and contrastive /ṼṼ/ vowels, it makes sense for nasality to be the only feature that changes between the UR and SR of /VVN/ vowels. With that said, MaxLex does not contain an inherent mechanism to act upon this sensible conclusion. Instead, the learner is free to choose any covert UR that models the data and minimizes changes in constraint weights from their biases, regardless of whether there are one or fifty feature disparities in the UR→SR mapping.

To focus on the learner's preference for minimally abstract URs, I provide MaxLex with one additional potential covert UR, /ɒɒ/. Just like its unrounded counterpart /aa/, the low round back vowel /ɒɒ/ is quite similar to its corresponding SR, [ãã], except it contains *two* disparities rather than one: nasality and roundedness. Importantly, any increasingly abstract UR (e.g., diacritics) would suffice in the following discussion, but /ɒɒ/ is an especially good candidate because it is *more* abstract than /aa/ (/ɒɒ/ never surfaces in Punjabi and has more disparities in the input-output mapping) but only minimally so. Thus, /ɒɒ/ serves as a stand-in for any overly abstract covert UR that needs to be ruled out,

[3] See 5 in the appendix for a full set of constraint definitions.

Constraints	Type	initial w	final w
ID[nas]	faith.	50.00	51.37
IDFIN[nas]	faith.	50.00	44.83
SPRD-L[nas]	mark.	50.00	92.83
*NASOBS	mark.	50.00	100.00
*NASG	mark.	50.00	99.48
ID[nas]/_V	contfaith.	50.00	100.00
*VVVN	mark.	50.00	100.00
ID[rd]	faith.	50.00	0.00
*LOWRD	mark.	50.00	100.00

Table 3: Constraint weights after phonotactic learning with MaxLex.

and if /ɒɒ/ is ruled out, potential URs with greater disparities will also be ruled out.[4]

The weights acquired in the phonotactic learning stage of MaxLex demonstrate three phonotactic restrictions in Punjabi that must hold regardless of the particular UR chosen for /VVN/ vowels. First, low round vowels never surface in Punjabi, so *LOWRD is undominated and ID[rd] is inactive and set to zero. As shown in (1), this weighting relationship appropriately unrounds all inputs containing /ɒɒ/ with a probability of 1.0.

(1) Low Round vowels never surface

/sɒɒ/	*LOWRD 100.00	ID[rd] 0.00	$\mathcal{H}$	$\tilde{\mathcal{P}}$
a. ☞ sɑɑ		-1	0	1.0
b. sɒɒ	-1		-100	$4e^{-44}$

Another phonotactic restriction MaxLex acquires is the absolute ban on nasal obstruents in Punjabi. To accomplish this, *NASOBS must outweigh SPRD-L, as in (2).

(2) Obstruents never nasalized

/sɑɑʋãã/	*NASOBS 100.00	SPRD-L 92.83	$\mathcal{H}$	$\tilde{\mathcal{P}}$
a. ☞ sããʋ̃ãã		-1	-92.83	0.999
b. s̃ããʋ̃ãã	-1		-100	$8e^{-4}$

Finally, in order for /VVN/ vowels to surface consistently as nasal vowels, either *VVN or SPRD-L must outweigh ID[nas]. In fact, both constraints end up outweighing ID[nas],

[4]Note that a covert UR like the *nasalized low back round vowel /õõ/* only has a single disparity in its mapping to [ãã] (roundedness), so it would tie /ɑɑ/ in its performance on the disparity component of the objective function. However, just like the concrete UR /ãã/ fails to model the lack of harmony triggered by /VVN/ vowels in Punjabi, any other nasal vowel would run into the same issue.

Constraints	Type	initial w	final w
ID[nas]	faith.	51.37	3.36
IDFIN[nas]	faith.	44.83	99.96
SPRD-L	mark.	92.83	5.65
*NASOBS	mark.	100.00	100.00
*NASG	mark.	99.48	0.19
ID[nas]/_V	contfaith.	100.00	100.00
*VVN	mark.	100.00	100.00
ID[rd]	faith.	0.00	0.00
*LOWRD	mark.	100.00	100.00

UR	$\mathcal{P}$
/tɑɑʋããn/	1.0

Table 4: Constraint weights and UR probabilities with concrete URs only

resulting in /VVN/ vowels always surfacing as nasal, as in (3).

(3) /VVN/ vowels always nasalized

/siin/	*VVN 100.00	ID[nas] 51.37	SPRD-L 92.83	$\mathcal{H}$	$\tilde{\mathcal{P}}$
a. ☞ sĩĩn		-1	-1	-144.2	1.0
b. siin	-1		-2	-285.66	$3e^{-62}$

Once the morphologically aware learning stage begins, MaxLex recognizes that surface alternations such as [sɑɑ] and [sãã] belong to the same underlying morpheme. We will first consider the use of concrete URs to model the data. For our purposes, the important morphemes are those containing non-alternating pre-N vowels like [tɑɑʋããn]. Because [tɑɑʋããn] only exhibits a single surface form, only one concrete UR is available to MaxLex, and using it prevents MaxLex from accurately modeling the data. The results for constraint weights and UR probabilities with only concrete URs are given in Table 4. Again, because [tɑɑʋããn] does not exhibit morphological alternations, there is only one potential UR, and it receives all of the probability as the correct UR for modeling the data.

However, using only concrete URs results in the model's inability to successfully learn the appropriate constraint weights and an almost zero probability of learning the correct nasalization pattern of forms with /VVN/ vowels. This is exemplified by the tableau in (4). Because the URs for both /VVN/ and contrastive /ṼṼ/ vowels are identical, MaxLex cannot correctly learn the pattern. When presented with /tɑɑʋããn/, the learner incorrectly as-

signs almost all the probability to the candidate that exhibits nasal harmony.

(4) Failure of Concrete URs to model Punjabi nasalization

/taavɑ̃ɑ̃n/		*VVN 100.00	SPRD-L 5.65	*NASG 0.19	ID[nas] 3.36	$\mathcal{H}$	$\tilde{\mathcal{P}}$
a.	taavɑ̃ɑ̃n		-3			-16.95	0.012
b.	taavɑan	-1	-4		-1	-125.96	$6e^{-50}$
c.	tɑ̃ɑ̃vɑ̃ɑ̃n		-1	-1	-2	-12.56	0.988

Up to this point, the learning process has followed the same general pattern as the Klamath [i]-[ø] alternation discussed in O'Hara (2017). The phonotactic patterns were learned, and using a concrete UR for /VVN/ vowels resulted in a failure to accurately predict the observed data. Now, just as for Klamath, MaxLex is provided two covert URs to consider when modeling the data. The results of the morphologically aware learning stage with /ɑ̃ɑ̃/, /ɑɑ/, and /ɒɒ/ included as potential URs are provided in Table 5. Here, the final constraint weights are quite similar to the weights when concrete URs were the only potential option, but the inclusion of the covert representations as potential URs for forms with /VVN/ vowels allows MaxLex to accurately model the data, with a .98 total probability of observing the correct surface forms for all words fed to the learner. However, while MaxLex is successful in modeling the data with the inclusion of these two covert URs, it is unsuccessful in discriminating between them, instead assigning an equal 0.5 probability to both covert URs. In other words, the MaxLex prior cannot distinguish between a restrictedly abstract UR like /ɑɑ/ and an unnecessarily abstract UR like /ɒɒ/. The reason for this is that changes in constraint weights from the phonotactic to the morphologically-aware learning stage are identical regardless of which covert UR is used. To permit the nasal harmony pattern in forms with contrastive /ṼṼ/ vowels, ID[nas] and *NASG need to lower so that their combined sum is less than SPRD-L. This change holds regardless of whether the UR for the /VVN/ vowel in [taavɑ̃ɑ̃n] is /ɑɑ/ or /ɒɒ/. Additionally, ID[rd] – the faithfulness constraint associated with the increasingly abstract UR, /ɒɒ/ – remains at zero without any pressure to increase. This is because no alternation exists for /VVN/ vowels, so faithfulness constraints are not driving their surface realization. In cases like Punjabi, then, when an alternation does not exist but a covert UR is still needed, the MaxLex prior fails to restrict abstraction because minimally abstract URs like /ɑɑ/ and increasingly abstract URs like /ɒɒ/ do not rely on distinct constraint weights to accurately model the data.

Constraints	Type	initial w	final w
ID[nas]	faith.	51.37	0.07
IDFIN[nas]	faith.	44.83	100.00
SPRD-L	mark.	92.83	5.42
*NASOBS	mark.	100.00	100.00
*NASG	mark.	99.48	0.02
ID[nas]/_V	contfaith.	100.00	100.00
*VVN	mark.	100.00	100.00
ID[rd]	faith.	0.00	0.00
*LOWRD	mark.	100.00	100.00

UR	$\mathcal{P}$
/taavɑan/	0.5
/taavɒɒn/	0.5
/taavɑ̃ɑ̃n/	0.0

Table 5: Constraint weights and UR probabilities with abstract URs included

5 Learning via Disparity Minimization

In this section, I propose an update to the MaxLex learner that generates a preference for minimally abstract URs over increasingly abstract alternatives, even when the minimally abstract UR does not outperform the increasingly abstract UR in either its accuracy in modeling the data or its deviation from a prior on constraint weights. Specifically, if the disparity component in equation (3) is added to the objective function, assigning probability to URs that introduce disparities increases the loss. Consequently, abstraction will only be preferred if doing so sufficiently increases the likelihood of observing the data.

$$\mathrm{D}(\mathrm{IO}_j) = \sum_{i=1}^{k_j} \left[\mathbf{1}_{\{s^I_{ij} \oplus s^O_{ij} = \varnothing\}} + \sum_{f \in \mathrm{F}} \mathbf{1}_{\{s^I_{fij} \neq s^O_{fij}\}} \right]^2 \tag{3}$$

As shown in the equation, the disparity value for the jth input-output mapping is computed by summing squared segment-level disparity terms across all k_j aligned segments.

Each term within the summation compares the ith input segment (s^I_{ij}) with the corresponding output segment (s^O_{ij}). Two indicator functions contribute to segment-level disparities: the first returns 1 if exactly one of the two segments is null (i.e., an insertion or deletion has occurred); the second iterates over all features f in the feature set F, returning 1 whenever the corresponding input-output segments differ on that feature. When either s^I_{ij} or s^O_{ij} are null, the second term contributes 0 vacuously, since the null segment has no features over which to compare. In effect, incentivizing the minimization of the disparity bias encourages the learner to acquire input-output mappings with as few differences as possible between corresponding segments. Squaring segment-level disparities before aggregating them results in a quadratic increase of the disparity bias as the number of disparities for a given segment increases, thereby enacting harsher penalties for underlying segments that are increasingly divorced from their realization.

The inclusion of a disparity bias in the learner is motivated by both theoretical assumptions and empirical observations about how underlying representations are selected. From a modeling perspective, the updated learner satisfies Occam's Razor: among competing hypotheses that account equally well for the data, the disparity bias favors the simplest one. In the context of UR selection, increasingly abstract URs introduce additional complexity by requiring more transformations between the underlying and surface forms. In the absence of independent motivation, positing such abstract forms results in unnecessary representational complexity.

Indeed, linguists often assume that URs reflect SRs faithfully unless motivated otherwise (Kiparsky, 1982; Baković et al., 2022). This assumption is formalized in Tesar (2014, p.1) through the principle of surface-orientedness, whereby "disparities between input and output are introduced only to the extent necessary" to satisfy independent grammatical restrictions. Similarly, Prince and Smolensky (1993/2004, p.225–226) propose the Lexicon Optimization Principle, which holds that learners should select URs that result in the most harmonic output, minimizing violations unless a more abstract UR yields a demonstrable advantage. Finally, empirical evidence supports the notion that language learners disprefer abstract URs. As shown by Kiparsky (1973), covert URs are often reanalyzed over time as surface-true by successive generations of learners, suggesting a robust bias in favor of minimizing disparities.

What follows demonstrates the computational success of incorporating the disparity bias into the MaxLex learner. The procedure begins in the same way as MaxLex, with an initial stage of phonotactic learning followed by a morphologically-aware learning stage. Here, as in the previous section, the algorithm is provided with two potential covert URs to consider, /ɑɑ/ and /ɒɒ/. Importantly, these are the only two URs that need to be considered under the present analysis to demonstrate that the model prefers minimal abstraction. That is, if /ɒɒ/ can be ruled out by the disparity bias, any other covert UR with a superset of the disparities of /ɑɑ/ can also be ruled out. In this case, the UR of /VVN/ vowels must be oral to appropriately model the data, and /ɑɑ/ only differs from the surface form [ɑ̃ɑ̃] in its nasality value. As such, any other potential UR that could effectively model the observed Punjabi forms with a sufficiently high likelihood necessarily possesses a superset of the disparities of /ɑɑ/ and will, therefore, be dispreferred by the disparity bias.

The results of the simulation with the updated learner are provided in Table 6. The weights the learner arrives at are almost identical to the weights learned by the original MaxLex learner. The key difference here is the probability given to the three potential URs considered for [tɑɑʋɑ̃ɑ̃n]. Whereas MaxLex assigned equal probability to both covert URs because they model the grammar equally well and minimize the prior to the same degree, the updated learner assigns essentially all of the probability to the minimally abstract covert UR, /tɑɑʋɑɑn/.

In sum, O'Hara (2017) demonstrated that MaxLex effectively constrains UR abstraction in cases where surface alternations are present and potential covert URs do not dif-

Constraints	Type	initial w	final w
ID[nas]	faith.	51.37	0.00
IDFIN[nas]	faith.	44.83	100.00
SPRD-L	mark.	92.83	4.61
*NASOBS	mark.	100.00	100.00
*NASG	mark.	99.48	0.00
ID[nas]/_V	contfaith.	100.00	100.00
*VVN	mark.	100.00	100.00
ID[rd]	faith.	0.00	0.00
*LOWRD	mark.	100.00	100.00

UR	$\mathcal{P}$
/taavaan/	1.00
/taavɒɒn/	$9e^{-15}$
/taavããn/	$2e^{-15}$

Table 6: Constraint weights and UR probabilities with abstract URs and the DISPARITY bias.

fer in their disparity count (as in Klamath). Incorporating an explicit disparity bias into MaxLex extends its utility by enabling it to constrain unnecessary abstraction in forms that lack alternations but still require a covert UR for an adequate analysis.

6 Traversing the Search Space

The proposed disparity bias in equation (3) is intimately connected to output-driven maps defined in Tesar (2014, 2016). Tesar's framework shows how disparities between underlying and surface forms can be used to organize the space of potential URs in a way that allows the learner to search efficiently and avoid unnecessary computations.

Output-driven phonology imposes entailment relationships on UR-SR mappings based on their disparity profiles. If a UR maps to a given surface form with n disparities, then any UR that maps to that same surface form with a proper subset of those n disparities must also be grammatical. For instance, if the mapping is grammatical, then /to/ → [tu] must also be grammatical because /to/ → [tu] possesses a proper subset of /tɑ/ → [tu]'s disparities. However, this relationship does not hold between URs that have non-nested disparity sets; for example, /ti/ differs from [tu] in two features (e.g., [front], [round]), but /to/ differs in only one ([high]). Because the disparities in /ti/ → [tu] are not a superset of those in /to/ → [tu], no entailment of grammaticality follows between these mappings.

These entailment relationships allow the learner to organize the space of potential URs for a given surface form into a structured lattice (Figure 1), with the fully faithful UR at the top and increasingly abstract URs further down. Each node represents a potential UR, and edges lead to forms lower down in the lattice that differ by one additional disparity. If a UR at some level of the lattice fails to generate the observed SR, then all URs that include a superset of that UR's disparities (i.e., nodes further down the lattice) can be immediately ruled out. This structure allows the learner to efficiently eliminate broad swaths of the search space.

Importantly, the use of output-driven phonology by Tesar (2014, 2016) to structure the space of potential URs is primarily *negative*: it is designed to rule out more abstract URs based on the failure of a less abstract UR – one higher in the lattice – to map successfully to the surface form. It does not address how a learner might efficiently traverse the remaining space of *successful* URs that can generate the correct SR but differ in the number of disparities they require. Consider again the example lattice in Figure 1. If a learner considers /to/ as a potential UR for [tu] and finds that it is successful in modeling the data, no mechanism exists to prevent it from also needing to consider /tɑ/, /tɒ/, /tõ/, or any other potential UR that contains a proper superset of disparities in its /UR/→[SR] mapping to [tu].

I propose extending output-driven phonology in precisely this direction. A learner equipped with the disparity bias outlined in the previous section and a likelihood threshold at which success in modeling the data is 'good enough' can use the lattice structure not only to eliminate chains of incompatible URs, but also to stop searching the space once this likelihood threshold has been reached and further levels of abstraction only trivially improve the likelihood of observing the data.

More precisely, the search for the optimal UR could be conducted serially rather than initializing UR optimization with the full set of potential URs in contention simultaneously. A learner would begin by considering URs with 0 disparities and then move on to generate and consider URs with successively more disparities as needed. As a result, the size of

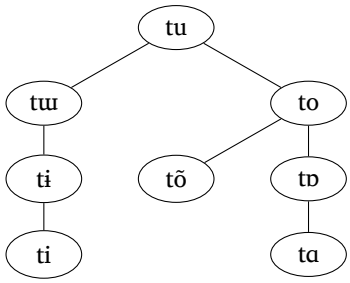

Figure 1: Example lattice for the output form [tu] (c.f. Tesar, 2016)

the search space would be irrelevant because the learner does not need to cover the entire space (or even most of it) to decide on the optimal UR.

In sum, the disparity bias does more than minimize abstraction: it also provides a principled way to structure and efficiently search an otherwise infinite space of potential URs. By combining the lattice structure from output-driven phonology with a disparity bias and principled likelihood threshold of acceptability, the framework not only curtails unnecessary abstraction but also offers a computationally efficient method for identifying the optimal UR.

7 Conclusion

This paper introduced a disparity bias as an addition to the MaxLex learner from O'Hara (2017) to improve its preference for minimally abstract underlying representations when multiple URs generate the same surface data with similar likelihood. By penalizing input-output disparities, the model favors URs that more closely resemble their surface realizations, thus curtailing unnecessary abstraction.

In addition to implementing this disparity bias, the paper outlined a blueprint for addressing a second major challenge posed by abstract URs. Specifically, permitting abstraction causes the space of potential URs to grow beyond a size that is computationally feasible to search. Drawing on insights from output-driven phonology, I proposed organizing the UR space into a lattice structured by disparity count and conducting a serial search through this space. By incorporating a likelihood threshold that defines when a UR adequately models the data, the learner can stop the search once candidates with additional disparities fail to meaningfully improve the likelihood of observing data.

While the paper provided a computational implementation of the disparity bias, the proposed method for structuring and traversing the UR space remains conceptual. Future work is required to develop this proposal computationally. This is a non-trivial task. Although concrete URs can be easily identified, generating the set of potential URs for the learner to consider at each increasing disparity level poses a combinatorial challenge. That is, as the number of disparities grows, the number of combined ways in which a segment could be altered to achieve that number of disparities explodes. The matter only worsens when considering multiple segments in a UR. Thus, additional work is needed to determine principled ways to constrain the set of potential URs at each disparity level considered by the learner.

A second open question concerns the likelihood threshold. Although I suggested a threshold as a stopping point, future research must investigate how this value can be grounded empirically. It may be that no single threshold is appropriate across a population of learners, and that the stopping criterion must be calibrated on a speaker-specific basis.

In addition, future work should explore how the disparity bias interacts with the MaxLex prior introduced in O'Hara (2017). This paper has shown that the MaxLex prior alone is insufficient for limiting abstraction in the case of Punjabi pre-N vowels. However, the prior remains crucial in cases like Klamath, where multiple URs generate the same surface form with equivalent disparity counts. Thus, it should be examined whether the disparity component and the MaxLex prior ever conflict, and if so, how such conflicts would be resolved in the learning process.

Finally, the disparity bias was implemented on data from Punjabi, but its application to phonological patterns from other languages that require varying degrees of abstraction is necessary. The cases discussed in Wang and Hayes (2025) would be an interesting set of case studies to begin with in this regard.

References

Adam C. Albright. 2002. *The identification of bases in morphological paradigms*. Ph.D. thesis, UCLA.

Eric Baković, Jeffrey Heinz, and Jonathan Rawski. 2022. *Phonological abstraction in the mental lexicon*. Oxford Academic.

Elena Bashir and Thomas J. Conners. 2019. *A descriptive grammar of Hindko, Panjabi, and Saraiki*. Mouton-CASL Grammar Series. De Gruyter Mouton.

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. 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.

Michael Kenstowicz and Charles Kisseberth. 1979. *Generative phonology: description and theory*. New York: Academic Press.

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

Paul Kiparsky. 1982. *How abstract is phonology?*, chapter 6. Foris Publications.

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. 2023. Covert URs: evidence from Pakistani Punjabi (talk). In *Formal Approaches to South Asian Languages (FASAL) 14*.

Joe Pater, Robert Staubs, Karen Jesney, and Brian Smith. 2012. Learning probabilities over underlying representations. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 62--71.

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

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

(5) Constraints used in Modeling Punjabi

i. SPRD-L[nas] (cf. Walker, 2003, 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, 51)
Assign a violation for every obstruent that dominates a [+nas] feature.

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

iv. 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*.

v. 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.

vi. *VVN
Assign a violation for every vowel that dominates a [-nas] feature when directly preceding a nasal consonant.

vii. 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*.

viii. 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*.

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