Information-theoretic analysis of disfluencies in speech

Shiva Upadhye and Richard Futrell
{upadhyes,rfutrell}@uci.edu
Department of Language Science
University of California, Irvine

Abstract

This study proposes and examines an information-theoretic measure of planning in incremental speech production, and investigates the effects of planning, predictability, and interference-based measures on lexical substitution errors. We then present a rate-distortion theoretic model of speech production that explicates how these factors affect the production of lexical substitution errors.

1 Introduction

Spontaneous speech is punctuated with disfluencies that reflect delays or interruptions in language production mechanisms [1, 2, 3, 4]. One common kind of disfluency is a lexical substitution, where a speaker says one word (which we term a **distractor**) and then corrects it to another word (which we term a **target**) [5]. For example, see the following naturally-occurring utterance from the Switchboard corpus [6]:

(1) and I believe that mental acuity is easy to *sustain* **maintain** if you just simply continue to exercise your mind

Here the distractor is *sustain*, which the speaker corrects to *maintain*. More generally, several prominent models of speech production have attributed hesitations and disfluencies of various kinds to difficulties in selecting what word to say next [7, 8, 9].

Here we develop an information-theoretic analysis of lexical substitution errors in a dataset of naturally-occurring disfluencies in the Switchboard corpus, applying a new measure of planning. We leverage large language models to estimate the relevant information quantities, and perform a targeted analysis that predicts *which word* is likely to appear as a distractor in each context. Finally, we explicate our results within the framework of a rate–distortion theoretic model of speech production.

1.1 Background and related work

Computational characterizations of lexical selection have emphasized the effects of word frequency and predictability on the ease of retrieving the appropriate word from memory. Low frequency words were not only associated with delayed production in naming experiments [10, 11, 12], but were also more likely to be preceded by disfluencies in naturalistic speech [13, 14]. Similarly, words that were less predictable conditional on the past (low *forward* predictability) or within the surrounding context (low *contextual* predictability) were also more likely to be preceded by hesitations, repetitions, and corrections [1, 15, 13, 16, 17]. Intriguingly, disfluencies and slowdowns in speech were also associated with *backward* transitional probability: the probability of a word given its *following* context [2, 18]; this is usually considered to be related to a speaker's plans for future production. In addition to being a strong predictor of fillers and corrections [19, 20], backward predictability has

Submitted to Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

also been associated with delays caused by the reactivation of past material to aid retrieval of a target compatible with the planned future [21].

A complementary line of experimental research in word production has also demonstrated that delays or errors in selecting a target can arise due to the activation accrued by its semantic and/or phonological neighbors [22, 23, 24]. The semantic interference effect, in particular, has been a robust finding in the Picture Word Interference paradigm, where the degree of semantic relatedness between a pictorial target and an orthographic distractor predicts a delay in retrieving the target [25]. Similarly, the presentation of a phonologically-related distractor has been associated with an increase in activation of the segments that the distractor shares in common with the target [26]. Hence, higher proximity between the target and distractor along semantic and/or phonological dimensions correlates to a greater degree of co-activation during processing.

1.2 Planning measure

Integrating perspectives from both lines of inquiry, our analyses of lexical substitution errors incorporate the effects of frequency, incremental production, and planning along with distance-based metrics such as phonological and semantic distance. We propose an information-theoretic measure of planning based on Pointwise Mutual Information (PMI) [27]. In particular, we use the PMI of a word x with the future context c_f given the past context c_p: pmiFP $= \ln \frac{p(x|c_f, c_p)}{p(x|c_p)}$, where $p(x \mid c_f, c_p)$ is the probability of x given both the preceding and following contexts and $p(x|c_p)$ is its forward predictability. Unlike backward predictability, pmiFP does not assume that the future is planned independent of the past. Rather it serves as a measure of planning that (i) does not commit to a particular planning order and (ii) uses the information provided by the past context to estimate the association between the word and the future context. Therefore, pmiFP can accommodate both linear [28, 29] and hierarchical planning [30, 31]. Previous works investigating backward predictability have not considered the integration of past and future context represented in pmiFP. In Section 4, we give a more detailed theoretical justification for the use of this measure.

2 Methods

We develop a regression model that predicts whether a given word x, as opposed to any other word in the vocabulary, is the distractor that the speaker selects and eventually overrides after production. Specifically, we examine how proximity to the target, frequency, incremental or forward predictability, and planned production guide the selection of a particular distractor given an utterance context.

Measures

Frequency: We use the frequencies from the SUBTLEXus corpus [32] to estimate unigram probability $p(x)$ of a given word x

Predictability-based measures: We estimate the forward, backward, and contextual probabilities of targets and distractors using XLNet [33], a large transformer-based model trained on both causal and masked modeling objectives. We select utterances where the length of both the preceding and following contexts exceed one word. To estimate $p(x \mid c_p)$ and $p(x \mid c_f)$ for a continuation x, we provide as input the past context c_p and the future context c_f (in reverse) respectively. To estimate $p(x \mid c_p, c_f)$, we provide the entire bidirectional utterance context (c_p, c_f) to the model with a *mask* applied to x in order to indicate the word to be predicted (see Appendix A for examples).

Distance measures: We estimate phonological distance between the target and distractor based on the Soundex algorithm [34] and semantic distance using cosine distance between pretrained GloVe embeddings [35].

Materials

For our analyses, we use Switchboard Annotations [36], a human-parsed subset of the Switchboard corpus of conversational speech [37] with annotations marking reparanda and repairs. In order to extract utterances with lexical substitutions, we identify two distinct signatures characteristic of these errors, namely (i) where the distractor or *reparandum* is immediately followed by the target or *repair* (1) and (ii) where a single word is substituted within a repeated phrase (Appendix A). We restrict our analyses to utterances with an equal number of reparanda and repairs in order to exclude cases of additions, deletions, or structural revisions. Utterances with multiple substitution errors were preprocessed into context frames with single substitution errors (see Appendix A for examples).

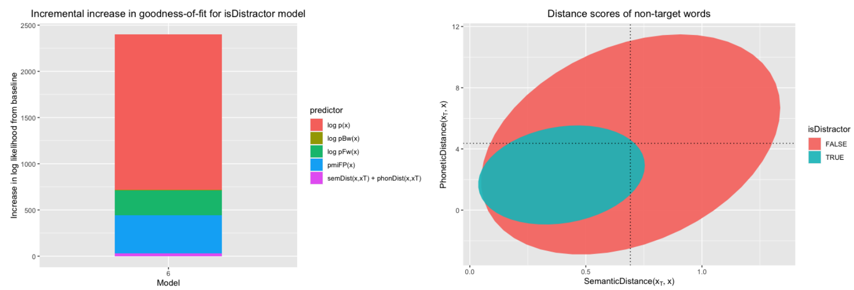

Figure 1: **Left:** Increase in log likelihood of isDistractor model with incrementally added predictors, in the order determined by goodness of fit in single-predictor models: $\log p$, distances, $\log pFw$, pmiFP, and $\log pBw$. The contribution of log backward probability is not visible. **Right:** Semantic and phonological distance to target for actual distractor words (blue) vs. all words in the vocabulary (red). Actual distractors are both semantically and phonologically close to the targets.

Model

We use a generalized linear mixed model (GLMM) [38] with a binary dependent variable *isDistractor* and the target identity as a random effect:

$$\text{isDistractor}(x) \sim \ln p(x) + \ln p(x \mid c_p) + \text{pmiFP}(x) + \text{phonDist}(x, x_T) + \text{semDist}(x, x_T) + (1 \mid x_T)$$

For each instance of a lexical substitution error (N = 1368), the model above is fit to predict which word out of a subset of nontarget words (N = 150) selected from the vocabulary with varying degrees of semantic and phonetic relatedness to the target is the chosen distractor.

We hypothesize that a nontarget x that is more frequent, and has higher forward predictability and *pmiFP* will have a higher likelihood of being the distractor that receives the activation required for selection. A word likely to be the distractor will also be co-activated with the target x_T by virtue of their common semantic and/or phonological features. Hence, we predict that the most probable candidate for the distractor will also have lower semantic and/or phonological distance from x_T.

3 Results

Both predictability and distance-based metrics emerged as significant predictors in the isDistractor model. As hypothesized, we observe a strong positive effect of frequency or unigram probability ($\beta = 0.240, p < 0.001$), forward predictability ($\beta = 0.269, p < 0.001$), and pmiFP ($\beta = 0.223, p < 0.001$). In contrast, we find a strong negative effect of semantic ($\beta = -0.695, p < 0.001$) and phonological distance ($\beta = -0.103, p < 0.001$).

We analyze the effects of individual predictors by examining the goodness-of-fit of the isDistractor model with incrementally added features. We observe that a model with pmiFP as the planning measure provides a better fit than one with backward predictability ($\chi^2 = 800.5, p < 0.001$), the measure used in previous work. We also find that when both these predictors are included as planning measures, backward predictability has a negligible contribution to the increase in log likelihood (Figure 1).

4 Model sketch

To explicate our results, we present a sketch of a model of speech production within a rate–distortion framework, deriving an optimal probabilistic policy $p_g(x \mid c_p)$ for the selection of a word x given a communicative goal g and current state c_p consisting of a sequence of previous words, subject to a constraint on the policy's usage of information about the goal g [39, 40, 41, 42, 43, 44, 45, 46]. Following Todorov [39]'s KL-control framework, a policy is selected to maximize an average future-discounted **value-to-go**

$$v_g(x \mid c_p) = \ell_g(x \mid c_p) + \alpha \left\langle v_g(x \mid x, c_p) \right\rangle_{p^*_g(x \mid c_p)}, \tag{1}$$

where $\alpha \in [0, 1]$ is a future-discount parameter, $\langle \cdot \rangle_p$ indicates an average over distribution p, $p^*_g(x \mid c_p)$ is an optimal policy, and $\ell(x \mid c_p)$ is the **local value** of an action x given goal g and state c_p. The local value is given by the communicative value of a word x minus a **control cost** term which reflects the KL divergence between the policy p_g and an **automatic policy** p_0 which is not conditional on the goal g:

$$\ell_g(x \mid c_p) = \underbrace{R_g(x \mid c_p)}_{\text{Communicative value}} \; - \underbrace{\ln \frac{p_g(x \mid c_p)}{p_0(x \mid c_p)}}_{\text{Control cost}}. \tag{2}$$

Communicative value R_g is meant to signify how well the word x conveys the speaker's intended message to a listener. Viewed within a rate-distortion theoretic framework, maximizing the communicative value corresponds to minimizing the distortion subject to the rate or control cost, which quantifies the amount of information about the goal used to determine the optimal action. Given this setting, the policy that maximizes average value-to-go as derived by [39] is

$$p_g^*(x \mid c_p) \propto \exp\{\ln p_0(x \mid c_p) + R_g(x \mid c_p) + \alpha \langle v(x \mid c_p, x)\rangle\}. \tag{3}$$

We propose that the selection of the next word in speech is determined by Eq. 3. The policy predicts that what matters for the selection of a word x is (1) predictability given past context c_p, (2) the communicative value R_g of the word with respect to the current goal and state, and (3) the expected value of words following x, a kind of planning effect.

4.1 Application to lexical substitution errors

We consider lexical substitution errors to reflect cases where there are two words (the target and the distractor) that both receive high probability under Eq. 3. In that case, a word is likely to appear as a distractor whenever any of the three terms inside Eq. 3 are high. We will see that these correspond to forward predictability, semantic and phonetic distance, and pmiFP respectively.

To understand the effect of the communicative value of word x, we consider the difference in communicative value between the distractor x and target x_T, $\Delta R_g = R_g(x \mid c_p) - R_g(x_T \mid c_p)$. This value differential ΔR_g corresponds to a negative communicative cost for saying x instead of x_T. This cost should reflect both semantic distance, because semantically similar words will share many of their features relevant for communicative goals, and phonetic distance, because phonetically similar words may be indistinguishable to a listener.

In order to draw out predictions from Eq. 3, we make three simplifying assumptions. First, we assume production consists of a word x followed by a second word representing the entire future of the utterance, c_f. Second, we assume that the production of the future c_f is deterministic given the communicative goal, and third, that the communicative value of c_f is independent of the choice of x. Under these assumptions, the policy in Eq. 3 simplifies to

$$p_g^*(x \mid c_p) \propto \exp\{\ln p_0(x \mid c_p) + \Delta R_g + \alpha \ln p_0(c_f \mid c_p, x)\}. \tag{4}$$

Rewriting the probability of the future c_f using Bayes' Rule in terms of the probability of the current word x given the future c_f, $p_0(x \mid c_p, c_f)$, we get

$$p_g^*(x \mid c_p) \propto \exp\left\{ \underbrace{\ln p_0(x \mid c_p)}_{\text{Forward predictability}} + \underbrace{\Delta R_g}_{\text{Distance}} + \alpha \underbrace{\ln \frac{p_0(x \mid c_p, c_f)}{p_0(x \mid c_p)}}_{\text{pmiFP}} \right\}, \tag{5}$$

where the three terms correspond to factors that were shown to predict which words appear as distractors in our corpus studies, including pmiFP. See Appendix B for full derivations. In addition, the frequency effects that we observe may be accommodated by adding an additional control cost for use of information about the state c_p, but we leave this modeling question to future work.

5 Conclusion

We have presented an analysis of lexical substitution errors in speech that predicts which words surface as distractors, and shown how the results can be accommodated in a rate–distortion control framework. The work opens the way for information-theoretic models of speech production that are tightly linked with rate–distortion models in other fields such as neuroscience and psychophysics.

References

[1] F. Goldman-Eisler, Speech production and language statistics, Nature 180 (1957) 1497–1497.

[2] F. Goldman-Eisler, Psycholinguistics: Experiments in spontaneous speech (1968).

[3] J. E. Fox Tree, The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech, Journal of Memory and Language 34 (1995) 709–738.

[4] E. Shriberg, Disfluencies in Switchboard, in: Proceedings of International Conference on Spoken Language Processing, volume 96, Citeseer, 1996, pp. 11–14.

[5] W. J. Levelt, Monitoring and self-repair in speech, Cognition 14 (1983) 41–104.

[6] J. J. Godfrey, E. C. Holliman, J. McDaniel, SWITCHBOARD: Telephone speech corpus for research and development, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92), volume 1, IEEE, 1992, pp. 517–520.

[7] V. A. Fromkin, Slips of the tongue, Scientific American 229 (1973) 110–117.

[8] G. Kempen, E. Hoenkamp, An incremental procedural grammar for sentence formulation, Cognitive Science 11 (1987) 201–258. URL: `https://www.sciencedirect.com/science/article/pii/S036402138780006X`. doi:`https://doi.org/10.1016/S0364-0213(87)80006-X`.

[9] K. Bock, W. J. M. Levelt, Language production : Grammatical encoding, 1994.

[10] R. C. Oldfield, A. Wingfield, Response latencies in naming objects, Quarterly Journal of Experimental Psychology 17 (1965) 273–281.

[11] Z. M. Griffin, K. Bock, Constraint, word frequency, and the relationship between lexical processing levels in spoken word production, Journal of Memory and Language 38 (1998) 313–338. URL: `https://www.sciencedirect.com/science/article/pii/S0749596X9792547X`. doi:`https://doi.org/10.1006/jmla.1997.2547`.

[12] J. Almeida, M. Knobel, M. Finkbeiner, A. Caramazza, The locus of the frequency effect in picture naming: When recognizing is not enough, Psychonomic bulletin & review 14 (2007) 1177–1182.

[13] G. Beattie, B. Butterworth, Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech, Language and Speech 22 (1979) 201 – 211.

[14] V. Kapatsinski, Frequency of use leads to automaticity of production: Evidence from repair in conversation, Language and Speech 53 (2010) 105 – 71.

[15] P. H. Tannenbaum, F. A. Williams, C. S. Hillier, Word predictability in the environments of hesitations, Journal of Verbal Learning and Verbal Behavior 4 (1965) 134–140.

[16] E. Shriberg, A. Stolcke, Word predictability after hesitations: a corpus-based study, Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96 3 (1996) 1868–1871 vol.3.

[17] Z. Harmon, V. Kapatsinski, Studying the dynamics of lexical access using disfluencies, Papers presented at DiSS (2015) 41–44.

[18] A. Bell, J. M. Brenier, M. L. Gregory, C. Girand, D. Jurafsky, Predictability effects on durations of content and function words in conversational english, Journal of Memory and Language 60 (2009) 92–111.

[19] S. Dammalapati, R. Rajkumar, S. Agarwal, Expectation and locality effects in the prediction of disfluent fillers and repairs in English speech, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 103–109. URL: `https://aclanthology.org/N19-3015`. doi:`10.18653/v1/N19-3015`.

[20] S. Dammalapati, R. Rajkumar, S. Agarwal, Effects of duration, locality, and surprisal in speech disfluency prediction in English spontaneous speech, in: Proceedings of the Society for Computation in Linguistics 2021, Association for Computational Linguistics, Online, 2021, pp. 91–101. URL: `https://aclanthology.org/2021.scil-1.9`.

[21] Z. Harmon, V. Kapatsinski, A theory of repetition and retrieval in language production., Psychological review (2021).

[22] A. Roelofs, A spreading-activation theory of lemma retrieval in speaking, Cognition 42 (1992) 107–142.

[23] G. M. Oppenheim, G. S. Dell, M. F. Schwartz, The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production, Cognition 114 (2010) 227–252.

[24] M. A. Goldrick, Limited interaction in speech production: Chronometric, speech error, and neuropsychological evidence, Language and Cognitive Processes 21 (2006) 817 – 855.

[25] R. R. Rosinski, Picture-word interference is semantically based., Child Development 48 (1977) 643–647.

[26] A. S. Meyer, H. Schriefers, Phonological facilitation in picture-word interference experiments: Effects of stimulus onset asynchrony and types of interfering stimuli., Journal of Experimental Psychology: Learning, Memory and Cognition 17 (1991) 1146–1160.

[27] K. W. Church, P. Hanks, Word association norms, mutual information, and lexicography, in: 27th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Vancouver, British Columbia, Canada, 1989, pp. 76–83. URL: `https://aclanthology.org/P89-1010`. doi:`10.3115/981623.981633`.

[28] A. S. Meyer, Lexical access in phrase and sentence production: Results from picture–word interference experiments, Journal of Memory and Language 35 (1996) 477–496.

[29] S. Brown-Schmidt, M. K. Tanenhaus, Watching the eyes when talking about size: An investigation of message formulation and utterance planning, Journal of Memory and Language 54 (2006) 592–609.

[30] K. Christianson, F. Ferreira, Conceptual accessibility and sentence production in a free word order language (odawa), Cognition 98 (2005) 105–135.

[31] S. Momma, V. S. Ferreira, Beyond linear order: The role of argument structure in speaking, Cognitive Psychology 114 (2019).

[32] M. Brysbaert, B. New, Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english, Behavior research methods 41 (2009) 977–990.

[33] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).

[34] R. Russell, M. Odell, The soundex indexing system, National Archives and Records Administration (1918).

[35] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[36] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, D. I. Beaver, The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue, Language Resources and Evaluation 44 (2010) 387–419.

[37] J. J. Godfrey, E. Holliman, J. McDaniel, Switchboard: telephone speech corpus for research and development, [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing 1 (1992) 517–520 vol.1.

[38] D. G. Clayton, Generalized linear mixed models, Markov chain Monte Carlo in practice 1 (1996) 275–302.

[39] E. Todorov, Efficient computation of optimal actions, Proceedings of the National Academy of Sciences 106 (2009) 11478–11483.

[40] N. Tishby, D. Polani, Information theory of decisions and actions, in: Perception-action cycle, Springer, 2011, pp. 601–636.

[41] S. G. Van Dijk, D. Polani, C. L. Nehaniv, Hierarchical behaviours: getting the most bang for your bit, in: European Conference on Artificial Life, Springer, 2009, pp. 342–349.

[42] P. A. Ortega, D. A. Braun, Thermodynamics as a theory of decision-making with information-processing costs, Proceedings of the Royal Society A: Mathematical, Physical, and Engineering Sciences 469 (2013) 20120683.

[43] T. Genewein, F. Leibfried, J. Grau-Moya, D. A. Braun, Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle, Frontiers in Robotics and AI 2 (2015) 27.

[44] S. J. Gershman, Origin of perseveration in the trade-off between reward and complexity, bioRxiv (2020).

[45] N. Zaslavsky, J. Hu, R. P. Levy, A Rate–Distortion view of human pragmatic reasoning, arXiv preprint arXiv:2005.06641 (2020).

[46] R. Futrell, An information-theoretic account of semantic interference in word production, Frontiers in Psychology 12 (2021) 672408.

A Appendix: Lexical substitution examples and data processing

Example of a lexical substitution within repeated material:

1 so until i see the entire quote old guard of the soviet *military* of the soviet **government** completely roll over and disappear preferably buried i still consider them a threat

Example of utterance with multiple substitution errors preprocessed into contexts with single substitution errors:

A it depends on whether *you* whether **we** figure that we have a defense oriented military or an *aggressive* **aggression** oriented military

- a it depends on whether [*you*/**we**] figure that we have a defense oriented military or an aggression oriented military
- b it depends on whether we figure that we have a defense oriented military or an [*aggression*/**aggressive**] oriented military

Examples of preprocessed XLnet inputs for estimating forward, backward, and masked probabilities:

A.3 it depends on whether <mask>

A.4 military oriented aggression an or military oriented defense a have we that figure <mask>

A.5 it depends on whether <mask> figure that we have a defense oriented military or an aggression oriented military

Code for preprocessing, calculating metrics, and analysis can be found at: InfoTheoreticDisfModel

B Appendix: Derivation of pmiFP from the speech production model

Starting with policy of Eq. 3, we can get to Eq. 5 by assuming (1) production consists of a word x followed by a second word representing the entire future of the utterance, c_f, (2) production of the future c_f is deterministic given the communicative goal, and (3) that the communicative value of c_f is independent of the choice of x. Under these assumptions, the policy in Eq. 3. The first assumption means that we only need to consider a finite time horizon with one future action. The second assumption means that the policy $p_g^*(c \mid c_p, x) = \delta_{cc_f}$. The third assumption means that $R_g(c_f \mid x, c_p)$ is the same for all x. These assumptions represent a scenario where the speaker has high certainty about what they will say next, and where the next part of an utterance is relatively independent of the current part, for example at the end of a phrase or clause.

We start with the policy probability (setting $\alpha = 1$ to save writing):

$$p_g^*(x \mid c_p) \propto \exp\{\ln p_0(x \mid c_p) + R_g(x \mid c_p) + \langle v(x' \mid c_p, x)\rangle\}. \tag{6}$$

First, we rewrite $R_g(x \mid c_p) = R_g(x_T \mid c_p) + \Delta R_g$. Because $R_g(x_T \mid c_p)$ is not a function of the action under consideration x, it can be absorbed into the normalizing constant of Eq. 3, giving

$$p_g^*(x \mid c_p) \propto \exp\{\ln p_0(x \mid c_p) + \Delta R_g + \langle v(x' \mid c_p, x)\rangle\}. \tag{7}$$

Now using assumption (1), we can rewrite the policy in Eq. 3 as:

$$p_g^*(x \mid c_p) \propto \exp\left\{\ln p_0(x \mid c_p) + \Delta R_g - \left\langle R_g(c \mid c_p, x) + \ln \frac{p_g^*(c \mid c_p, x)}{p_0(c \mid c_p, x)}\right\rangle_{p_g^*(c \mid c_p, x)}\right\}. \tag{8}$$

Using $p_g^*(c \mid c_p, x) = \delta_{cc_f}$ for the expectation over future actions (assumption 2), we get

$$p_g^*(x \mid c_p) \propto \exp\left\{\ln p_0(x \mid c_p) + \Delta R_g + R_g(c_f \mid c_p, x) - \ln \frac{1}{p_0(c_f \mid c_p, x)}\right\}. \tag{9}$$

Because $R_g(c_f \mid c_p, x)$ is invariant to x (assumption 3), it can be absorbed into the implicit normalizing constant of Eq. 9, giving

$$p_g^*(x \mid c_p) \propto \exp\{\ln p_0(x \mid c_p) + \Delta R_g + \ln p_0(c_f \mid c_p, x)\}. \tag{10}$$

Now applying Bayes' rule to the term $\ln p_0(c_f \mid c_p, x)$, and then applying logarithm rules, we get

$$\begin{aligned} p_g^*(x \mid c_p) &\propto \exp\left\{\ln p_0(x \mid c_p) + \Delta R_g + \ln \frac{p_0(x \mid c_p, c_f) p_0(c_f \mid c_p)}{p_0(x \mid c_p)}\right\} && (11)\\ &= \exp\left\{\ln p_0(x \mid c_p) + \Delta R_g + \ln \frac{p_0(x \mid c_p, c_f)}{p_0(x \mid c_p)} + \ln p_0(c_f \mid c_p)\right\}. && (12)\end{aligned}$$

Again, the last term is invariant to x, so it can be absorbed into the normalizing constant, leaving us with the policy considered in the main text

$$p_g^*(x \mid c_p) \propto \exp\left\{\ln p_0(x \mid c_p) + \Delta R_g + \ln \frac{p_0(x \mid c_p, c_f)}{p_0(x \mid c_p)}\right\}. \tag{13}$$