not just surprisal

towards a theory of incremental processing cost

Jacob Louis Hoover
MCQLL 2023-04-04

slides at: jahoo.github.io/2023-04

not just surprisal

towards a theory of incremental processing cost

1 **sampling algorithms** ——— superlinearity in surprisal theory

2 **divergence theory** ———————— beyond surprisal theory

slides at: jahoo.github.io/2023-04

sentence processing

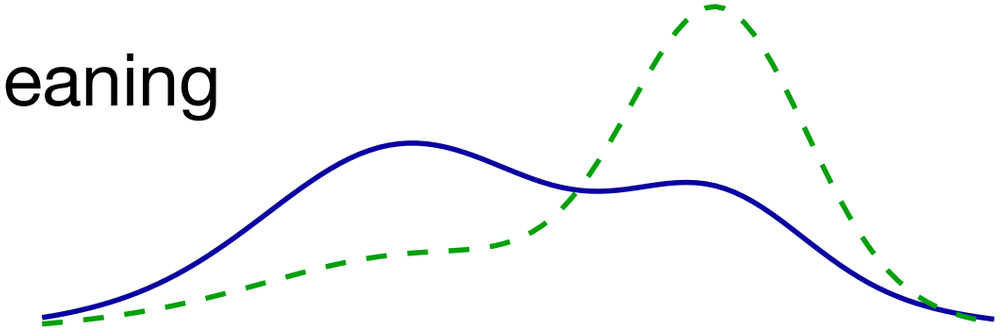
iterative inference problem

$z =$ 

$\mathbf{u} =$ Near that bank there is an otter.

What is going on when you read a sentence?

- **incrementally** observe each word
- build some **representation** of the meaning
  - $p($ Z $|$ $u_1, u_2, \ldots$ $)$
    meaning utterance so far
  - $p(Z \mid u_1, \ldots, u_{i-1}) \overset{u_i}{\rightsquigarrow} p(Z \mid u_1, \ldots, u_{i-1}, u_i)$



How? …with what processing algorithm?

- one clue: for humans, processing difficulty is not constant
- generally **surprising words take longer**.

How? …with what processing algorithm?

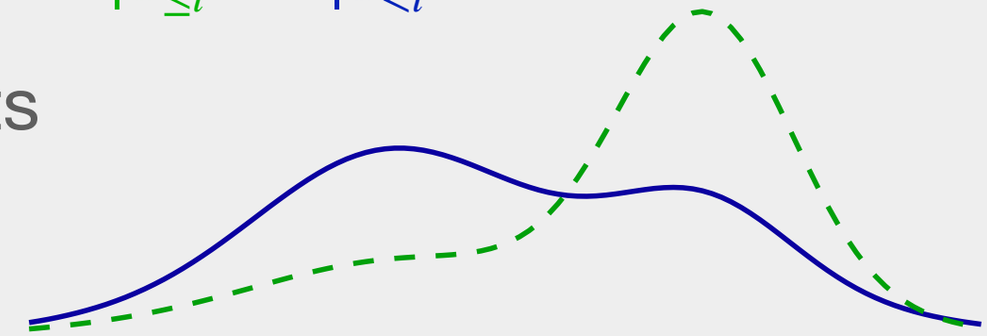
- one clue: for humans, processing difficulty is not constant
- generally **surprising words take longer**.

standard *surprisal theory* (Hale 2001, Levy 2005, 2008):

$$\text{cost}(u_i) \propto \overbrace{-\log p(u_i \mid \mathbf{u}_{<i})}^{:=\,\text{surp}(u_i)} = \mathrm{D}_{\mathrm{KL}}(p_{Z|\mathbf{u}_{\leq i}} \| p_{Z|\mathbf{u}_{<i}})$$

this equality holds *only with assumption* of deterministic yield. *(to be revisited!)*

= size of belief update



based on **computational level** arguments

(optimal / rational analysis)

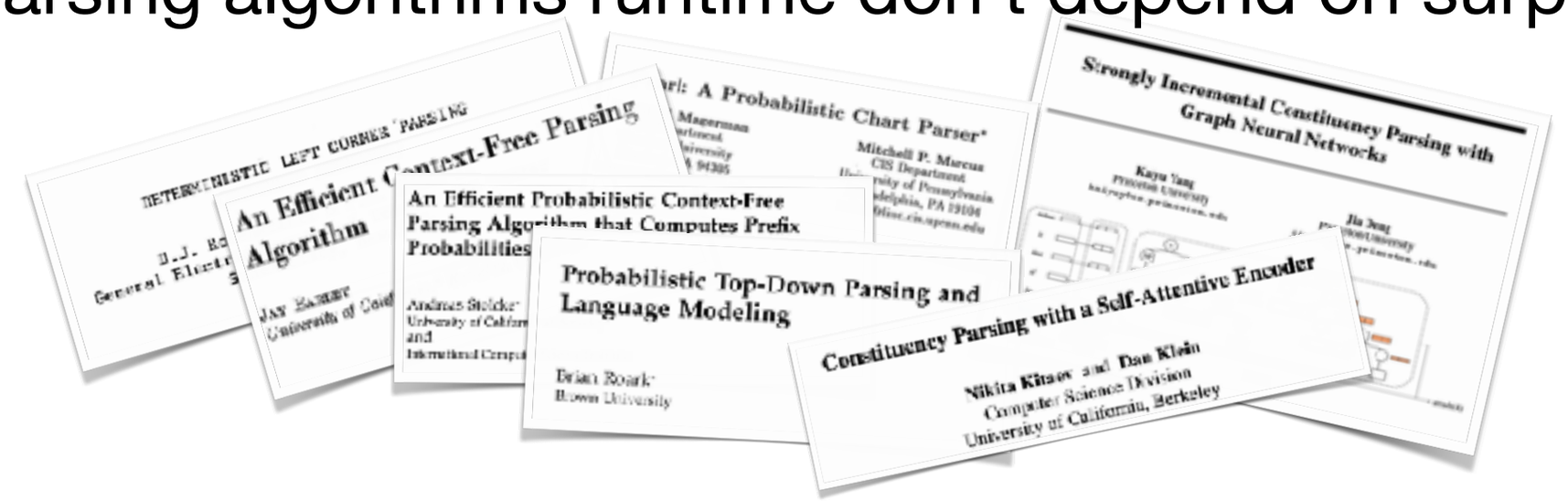
no algorithmic level theory

- *does an algorithm exist which would behave like this?*

How? …with what processing algorithm?

- one clue: for humans, processing difficulty is not constant
- generally **surprising words take longer**.

but, most parsing algorithms runtime don't depend on surprisal



what general kinds of algorithm *does* have this property?

- ones that prioritize higher-probability meanings: **sampling algorithms**
- sampling algorithms' runtime scales in divergence:
  - importance sampling: to approx. p with q, #samples $\approx e^{\mathrm{D}_{\mathrm{KL}}(p\|q)}$ (Chatterjee & Diaconis 2018)

The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

Jacob Louis Hoover[1,2], Morgan Sonderegger[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

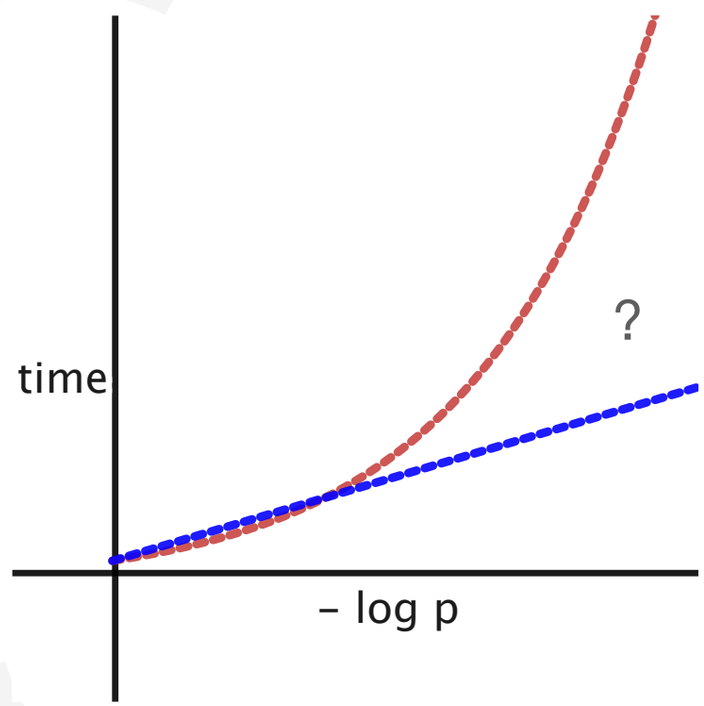
Jacob Louis Hoover[1,2], Morgan Sonderegger[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

Theoretical prediction about function linking surprisal to cost

- *sampling algorithms* predict
  ⇒ **superlinear** in surprisal
  ⇒ **increasing** variance
- contrasts with standard *surprisal theory* (Levy 2008, etc.)
  ⇒ **linear** in surprisal
  ⇒ **constant** variance (or no prediction)
  ⇒ (…. without a proposed algorithm)

Empirical question:

- what shape is the linking function?
  - fit scale-location GAMs



The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

Jacob Louis Hoover[1,2], Morgan Sonderegger[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

linking function: empirical study

is the mean superlinear? does variance increase?

predictor of interest: surprisal

- **Transformer-based LLMs**

response: processing time

- self-paced reading time

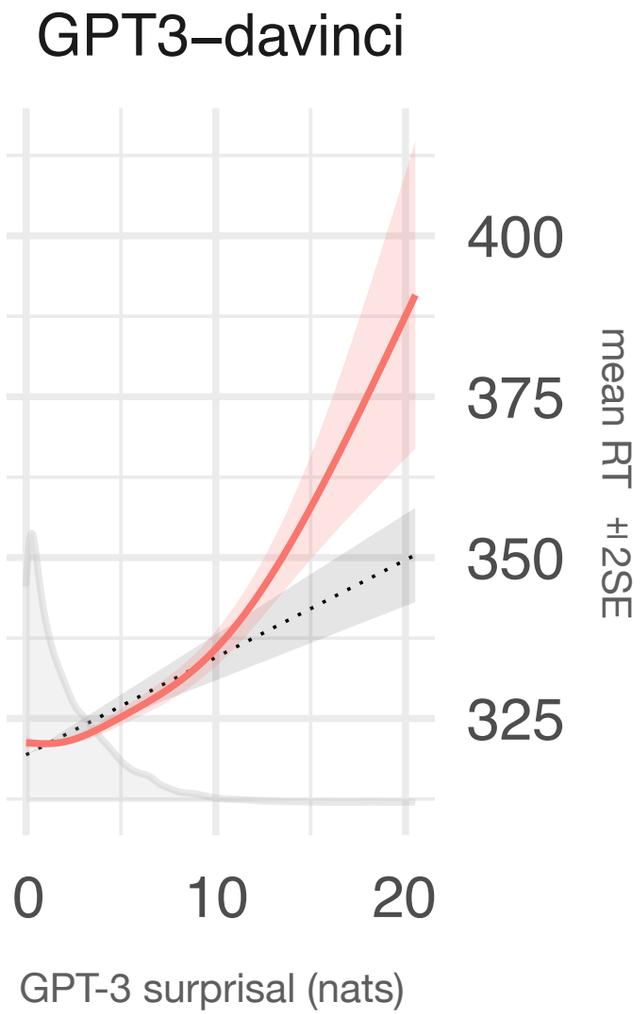
The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

Jacob Louis Hoover[1,2], Morgan Sonderegger[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

linking function: empirical study

is the mean superlinear? does variance increase?

Yes



The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

Jacob Louis Hoover[1,2], Morgan Sonderegger[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

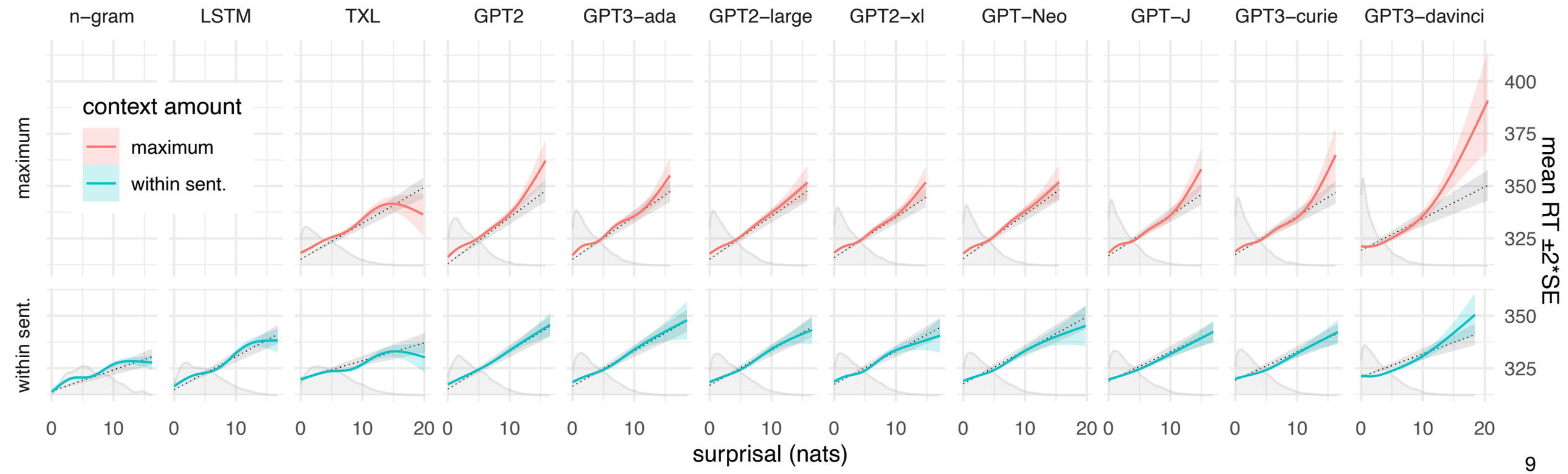
linking function: empirical study

is the mean superlinear? does variance increase?

Yes

- better LM ⇒ more superlinear

GAM fits of the effect of surprisal on reading time



The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

Jacob Louis Hoover[1,2], Morgan Sonderegger[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

linking function: empirical study

is the mean superlinear?

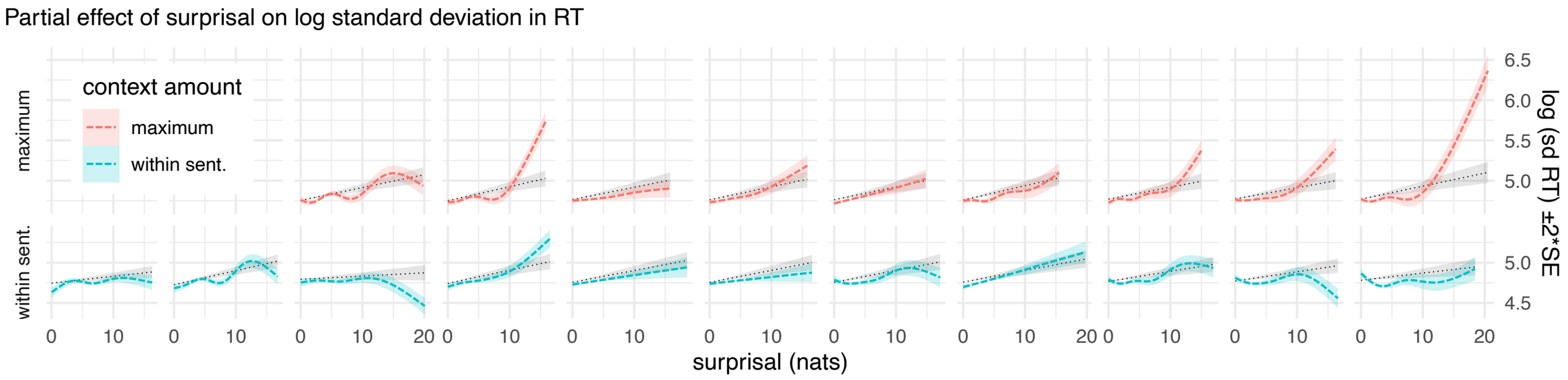
Yes

- more so for lower-perplexity language models

does variance increase?

Yes

- in general



The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

Jacob Louis Hoover[1,2], Morgan Sonderegger[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

linking function: empirical study

is the mean superlinear?

Yes

- more so for lower-perplexity language models

does variance increase?

Yes

- in general

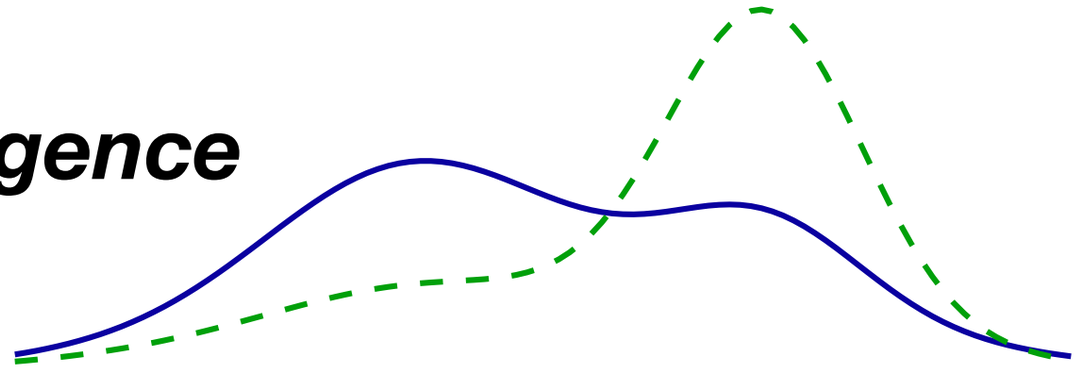
consistent with sampling algorithms' predictions

$\Longrightarrow$ sampling algorithms for processing

2 divergence theory

divergence theory

processing cost scales in divergence
(not just surprisal)



recall: sampling algorithms' runtime scales in divergence:

- e.g., importance sampling: to approx. p with q, #samples $\approx e^{\mathrm{D}_{\mathrm{KL}}(p\|q)}$

relationship with surprisal theory: assume surprisal = KL.

but in general, KL ≠ surprisal …there's an additional term:

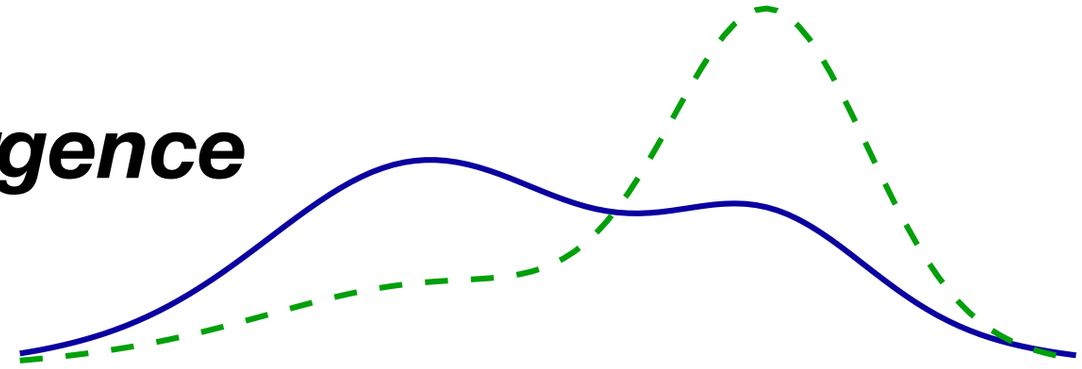
- $$\mathrm{D}_{\mathrm{KL}}(p_{Z|\mathbf{u}_{\leq i}}\|p_{Z|\mathbf{u}_{<i}}) = \underbrace{-\log p(u_i \mid \mathbf{u}_{<i})}_{\mathrm{surp}(u_i)} \underbrace{-\mathbb{E}_{z|\mathbf{u}_{\leq i}}[-\log p(u_i \mid z)]}_{R(u_i)}$$

- call $R(u_i)$ the "reconstruction uncertainty"
  - R measures uncertainty about u_i that remains *even under posterior*
  - if observed yield is part of Z, R is zero. *(R=0 iff $p(u_i \mid z) = 1$ $p_{z|\mathbf{u}_{\leq i}}$-a.e.)*

surprisal theory implicitly assumes this is zero

divergence theory

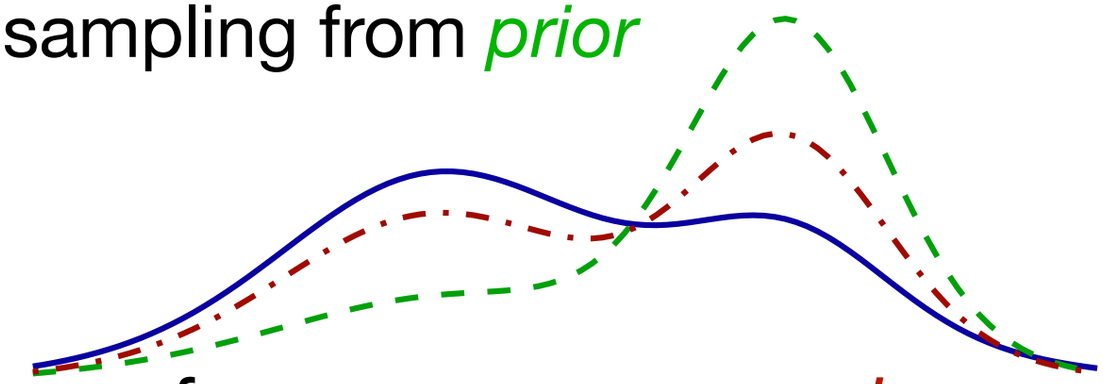
processing cost scales in divergence



surprisal theory: $\text{surp}(u_i)$

divergence theory: $\mathrm{D}_{\mathrm{KL}}(p_{Z|\mathbf{u}_{\leq i}} \| p_{Z|\mathbf{u}_{<i}}) = \text{surp}(u_i) - R(u_i)$

- algorithmic motivation: runtime of sampling from *prior*



or, more generally, if samples are drawn from some *proposal*, q (rather than the prior), **add one (last!) term**:

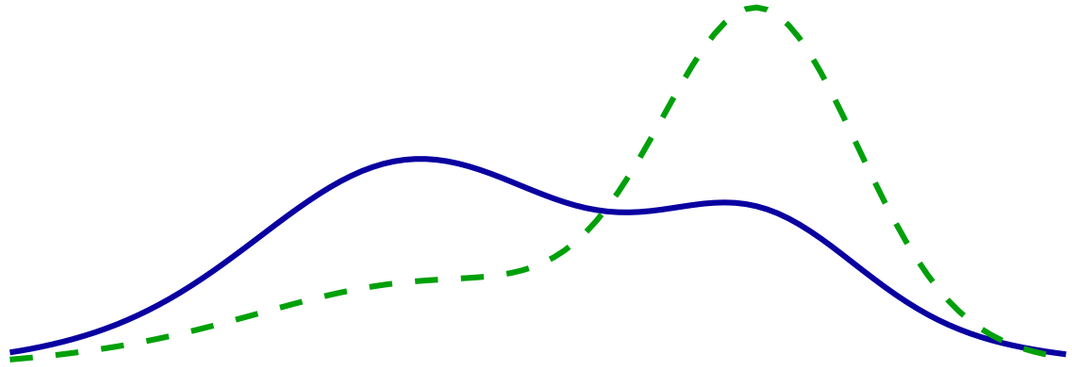
$$\mathrm{D}_{\mathrm{KL}}(p_{Z|\mathbf{u}_{\leq i}} \| q) = \text{surp}(u_i) - R(u_i) - \underbrace{\mathop{\mathbb{E}}_{z|\mathbf{u}_{\leq i}} \log \frac{q}{p_{Z|\mathbf{u}_{<i}}}}_{D(u_i)}$$

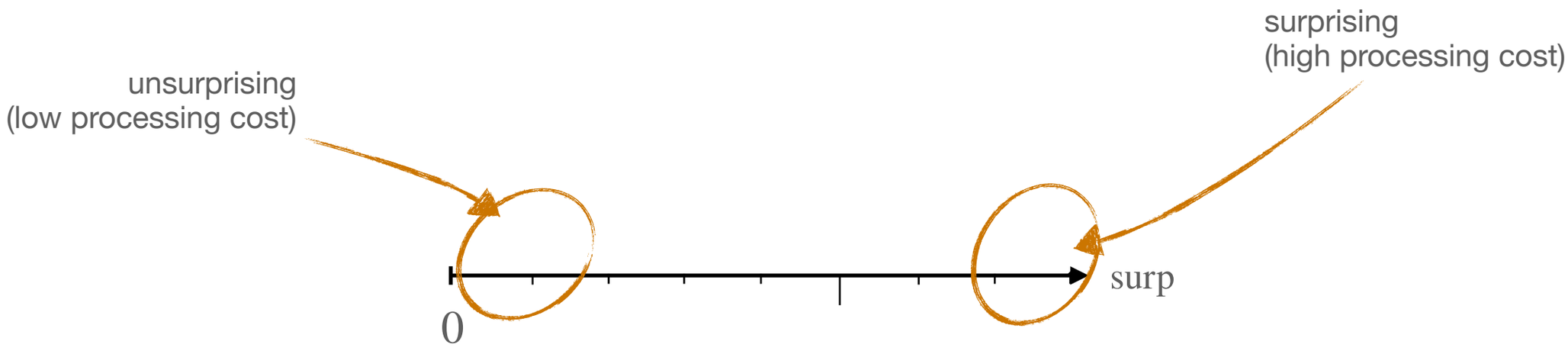
measures how much better q is than $p_{Z|\mathbf{u}_{<i}}$.

expected reduction in belief update afforded by using samples from proposal rather than the prior

divergence theory

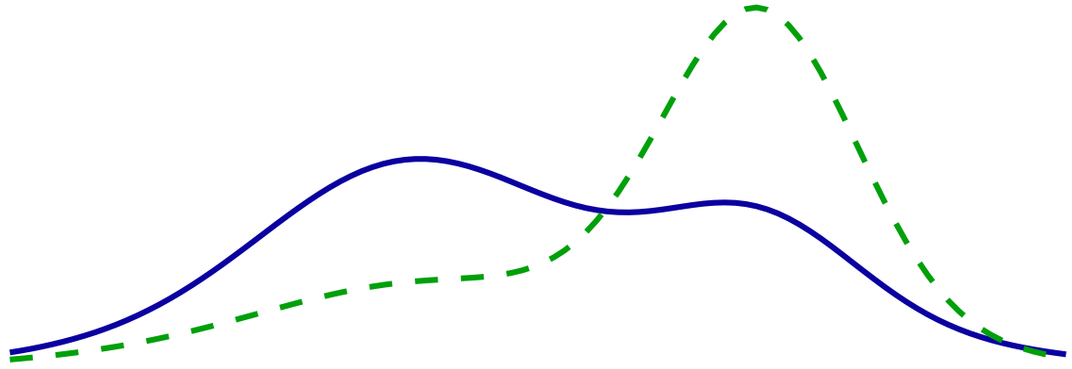
surprisal theory: $\text{surp}(u_i)$



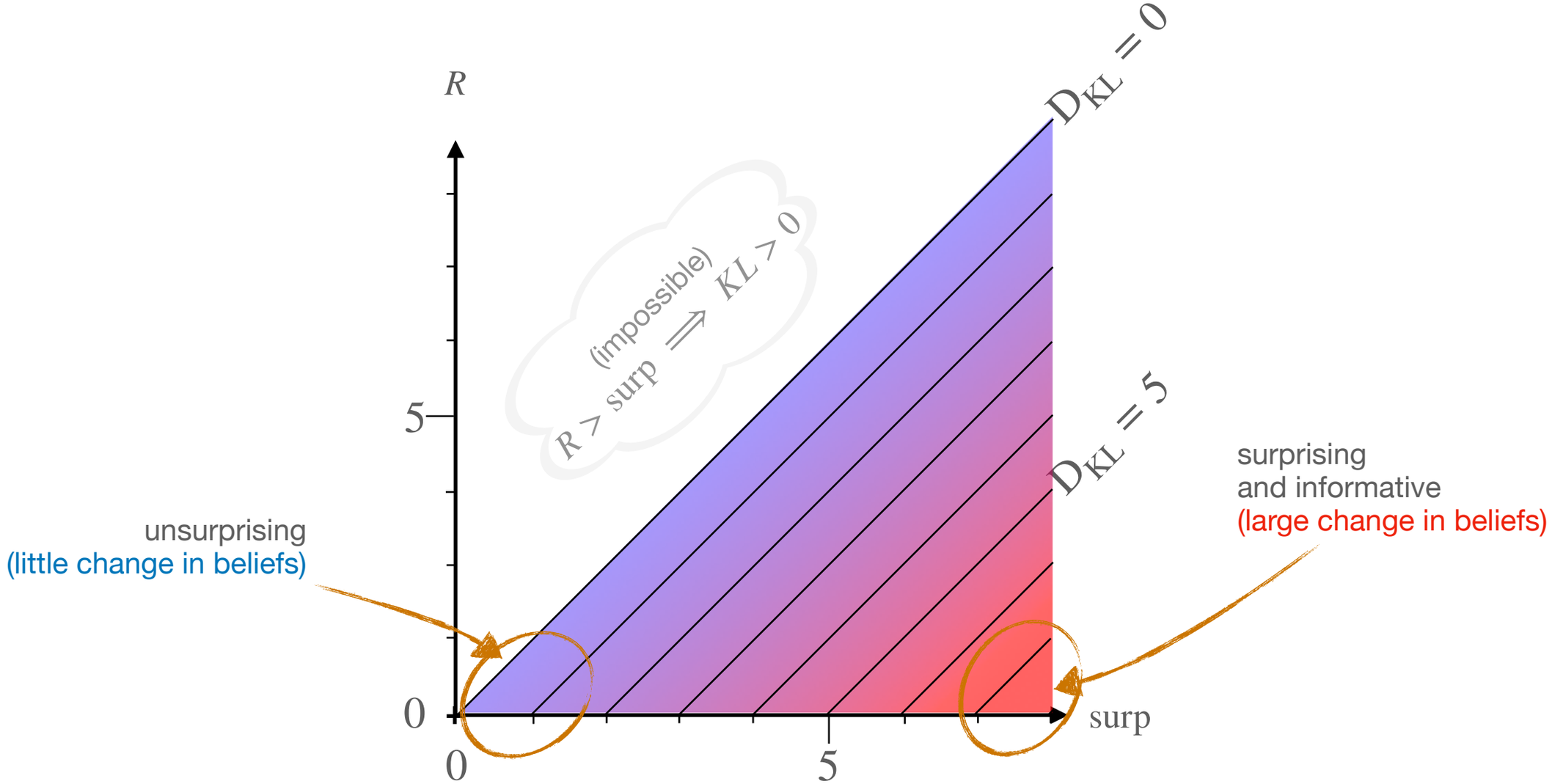


divergence theory

surprisal theory: $\text{surp}(u_i)$

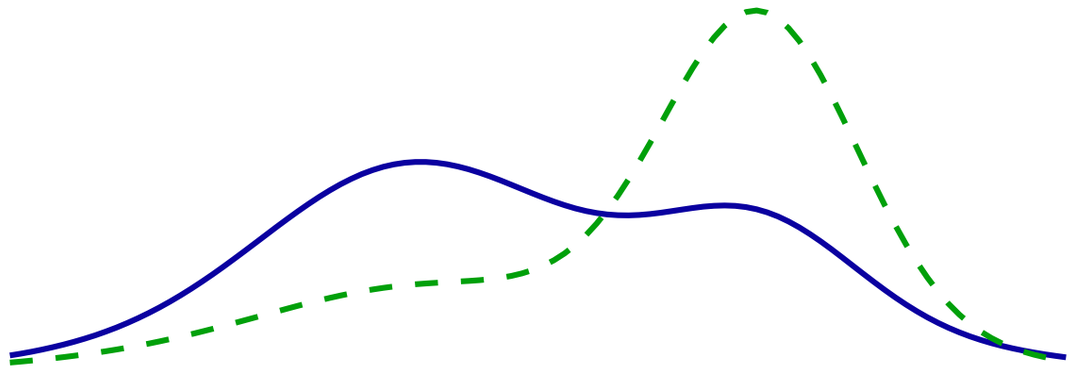


divergence theory: $\text{surp}(u_i) - R(u_i)$

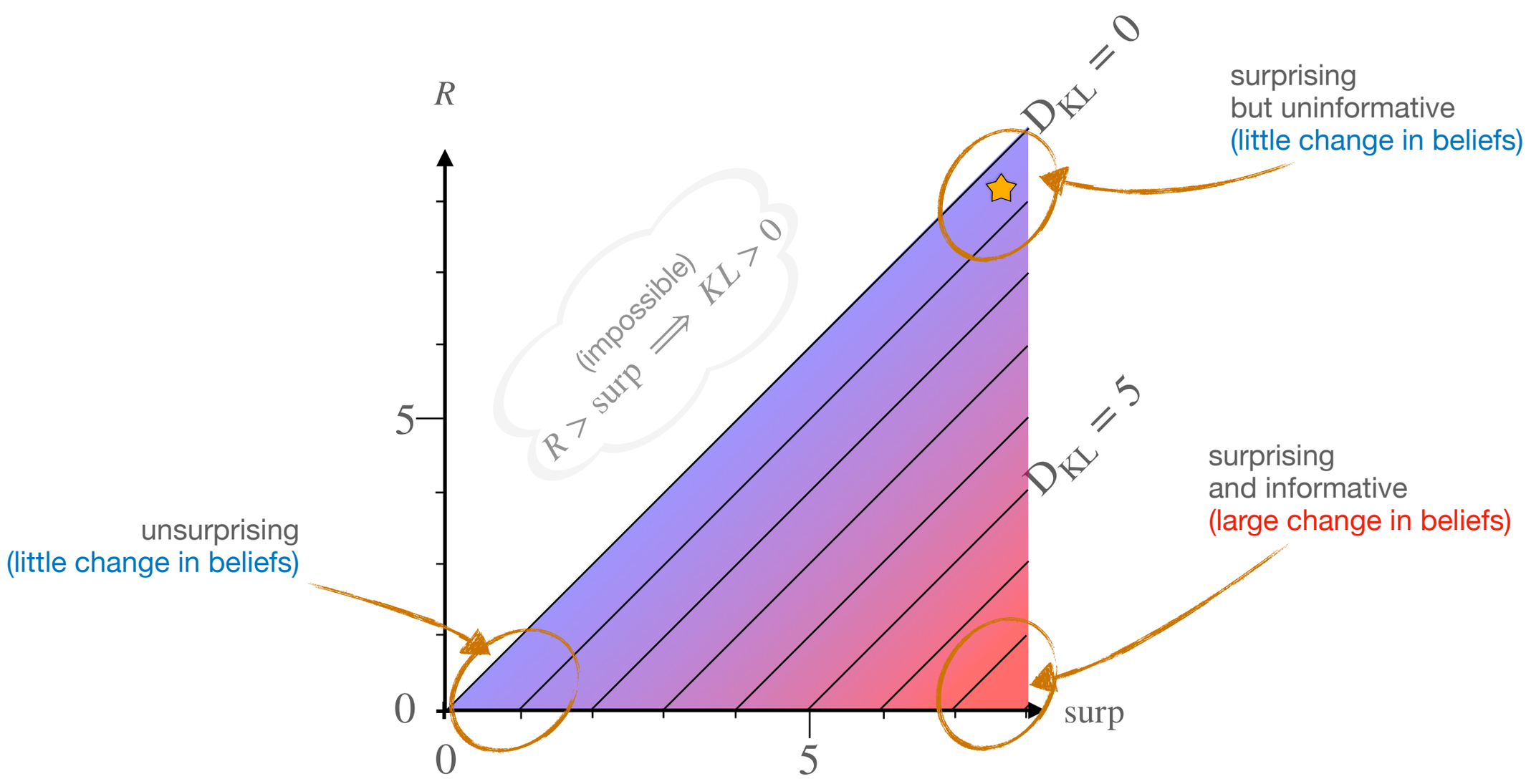


divergence theory

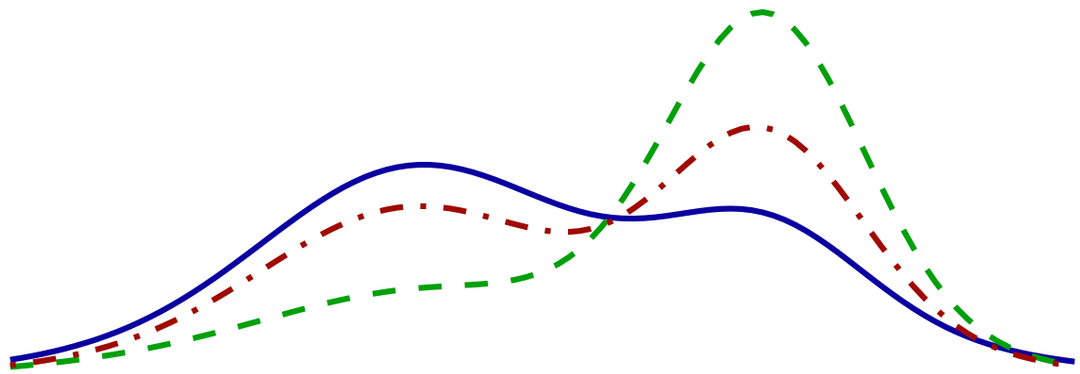
surprisal theory: $\text{surp}(u_i)$



divergence theory: $\text{surp}(u_i) - R(u_i)$

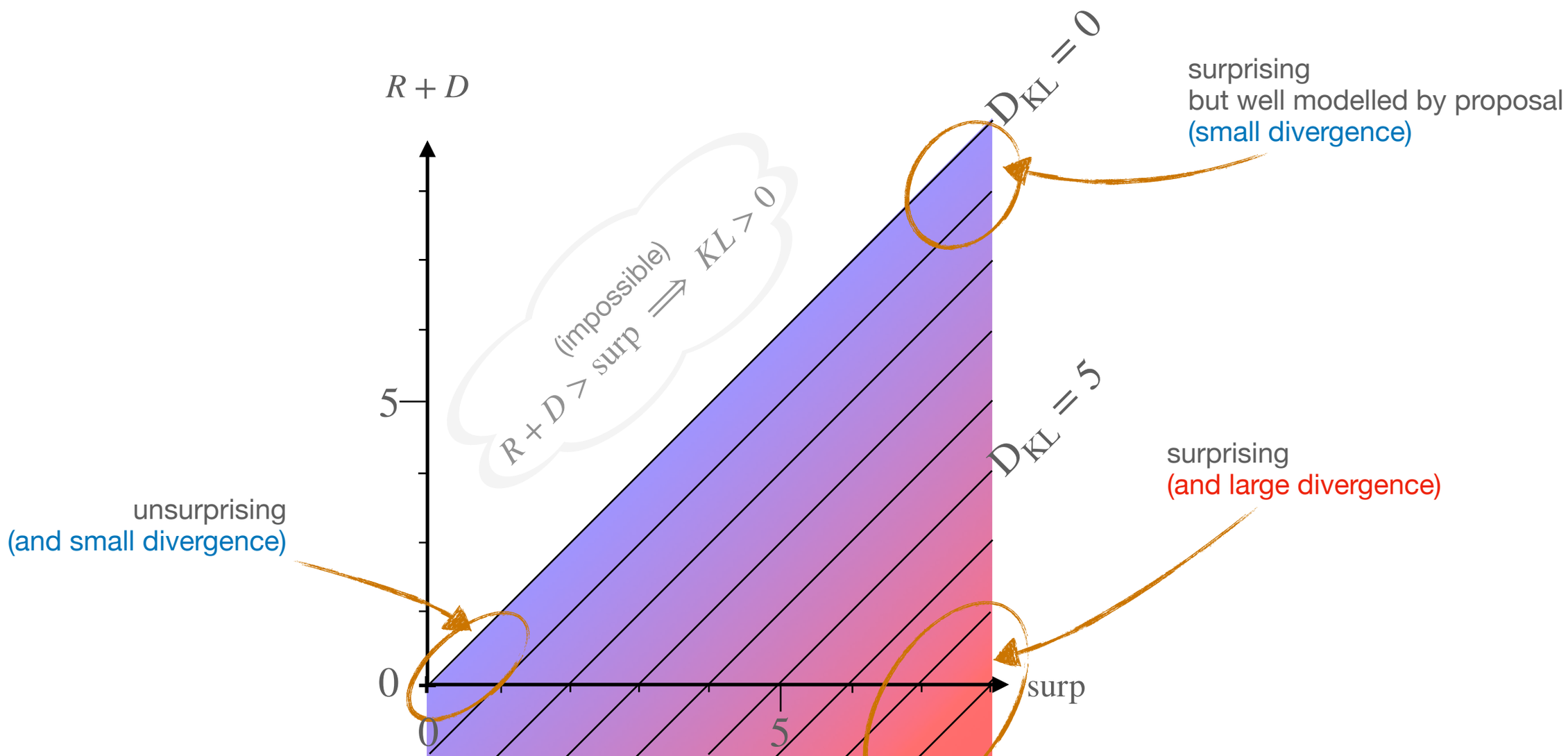


divergence theory



surprisal theory: $\text{surp}(u_i)$

divergence theory: $\text{surp}(u_i) - [R(u_i) + D(u_i)]$



divergence theory

informativeness of u_i

$$\text{cost}(u_i) \approx \exp \quad \mathrm{D}_{\mathrm{KL}}(p_{Z|\mathbf{u}_{\leq i}} \| q)$$

amount *surprisal* exceeds *informativeness*

$$\text{cost}(u_i) \approx \exp\Big[\quad \text{surp}(u_i) - [R(u_i) + D(u_i)] \quad\Big]$$

reconstruction uncertainty
(remaining under the posterior)

reduction in belief update afforded
by using proposal rather than prior

divergence theory

testing predictions

reconstruction uncertainty
(remaining under the posterior)

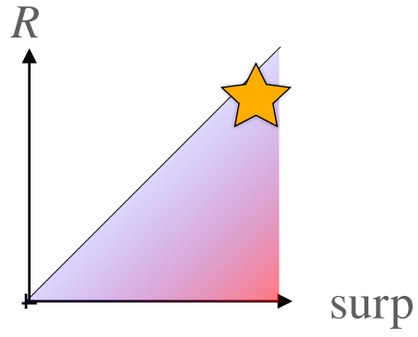
$$\text{cost}(u_i) \approx \exp\left[\ \text{surp}(u_i) - [R(u_i) + D(u_i)]\ \right]$$

reduction in belief update afforded
by using proposal rather than prior

study ideas

study R:

- Phenomena to look at: where processing is easier than expected under surprisal theory
- agreement attraction; grammatical illusions
  - *The key to all the cabinets <u>are</u> on the table*.
  - *The bills that no senator voted for will <u>ever</u> become law*.
    E.g. from data with self-paced reading time (Wagers et al 2009)
- or typos/speech errors



divergence theory

testing predictions

reconstruction uncertainty
(remaining under the posterior)

$$\text{cost}(u_i) \approx \exp\left[\ \text{surp}(u_i) - [R(u_i) + D(u_i)]\ \right]$$

reduction in belief update afforded
by using proposal rather than prior

study R+D together:

- $$\overbrace{\mathop{\mathbb{E}}_{z|\mathbf{u}_{\leq i}} \log \frac{1}{p(u_i \mid z)}}^{R(u_i)} + \overbrace{\mathop{\mathbb{E}}_{z|\mathbf{u}_{\leq i}} \log \frac{q(z; \mathbf{u}_{\leq i})}{p(z \mid \mathbf{u}_{<i})}}^{D(u_i)} = \mathop{\mathbb{E}}_{z|\mathbf{u}_{\leq i}} \log \frac{q(z; \mathbf{u}_{\leq i})}{p(z, u_i \mid \mathbf{u}_{<i})}$$

- Idea to estimate R+D with pretrained LMs:
  - use **continuation** $\mathbf{u}_{>i}$ as proxy for Z
  - use pretrained LM as estimator of prior, and of proposal
  - $$\widehat{[R + D]}_{LM}(u_i) = \mathop{\mathbb{E}}_{\mathbf{u}_{>i}|\mathbf{u}_{\leq i}} \log \frac{\text{LM}(\mathbf{u}_{>i} \mid \mathbf{u}_{\leq i})}{\text{LM}(\mathbf{u}_{>i}, u_i \mid \mathbf{u}_{<i})}$$
- Estimate surprisal with same LM.
- Use estimates to model human processing time
  - E.g., fit self-paced reading on Natural Stories, and compare goodness-of-fit of divergence-theory against surprisal-theory.

divergence theory

testing predictions

reconstruction uncertainty
(remaining under the posterior)

$$\text{cost}(u_i) \approx \exp\left[\ \text{surp}(u_i) - [R(u_i) + D(u_i)]\ \right]$$

reduction in belief update afforded
by using proposal rather than prior

study D alone: (very similar, but a little more complicated)

- $D(u_i) = \underset{z|\mathbf{u}_{\leq i}}{\mathbb{E}} \log \frac{q(z; \mathbf{u}_{\leq i})}{p(z \mid \mathbf{u}_{<i})}$
- Idea to estimate R+D with pretrained LMs:
  - use **continuation** $\mathbf{u}_{>i}$ as proxy for Z
  - estimate prior (without u_i): $p(\mathbf{u}_{>i} \mid \mathbf{u}_{<i}) \approx \mathbb{E}_{u_i' | \mathbf{u}_{<i}} \text{LM}(\mathbf{u}_{>i} \mid \mathbf{u}_{<i}, u_i')$
  - estimate proposal (with u_i): $\text{q}(\mathbf{u}_{>i}; \mathbf{u}_{\leq i}) \approx \text{LM}(\mathbf{u}_{>i}; \mathbf{u}_{\leq i})$
- so, $\hat{D}_{LM}(u_i) = \underset{\mathbf{u}_{>i}|\mathbf{u}_{\leq i}}{\mathbb{E}} \log \frac{\text{LM}(\mathbf{u}_{>i}; \mathbf{u}_{\leq i})}{\mathbb{E}_{u_i' | \mathbf{u}_{<i}} \text{LM}(\mathbf{u}_{>i} \mid \mathbf{u}_{<i}, u_i')}$

divergence theory

testing predictions

reconstruction uncertainty
(remaining under the posterior)

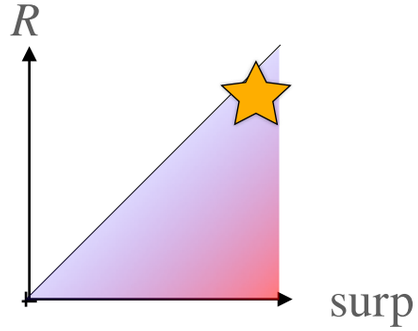
$$\text{cost}(u_i) \approx \exp\left[\ \text{surp}(u_i) - [R(u_i) + D(u_i)]\ \right]$$

reduction in belief update afforded
by using proposal rather than prior

study ideas

study R: look at specific phenomena

- where processing is easier than expected under surprisal theory
  - e.g., agreement attraction; grammatical illusions
    - *The key to all the cabinets <u>are</u> on the table*.
    - *The bills that no senator voted for will <u>ever</u> become law*.
- (what dataset?)



study R+D together: model processing time on corpus

- estimate with LM: $\widehat{[R + D]}_{LM}(u_i) = \mathop{\mathbb{E}}_{\mathbf{u}_{>i}|\mathbf{u}_{\leq i}} \log \frac{\text{LM}(\mathbf{u}_{>i} \mid \mathbf{u}_{\leq i})}{\text{LM}(\mathbf{u}_{>i}, u_i \mid \mathbf{u}_{<i})}$
- fit models of human processing time

thank you!

references

Chatterjee and Diaconis (2018). The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2).

Hale (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second meeting of the north American chapter of the association for computational linguistics*.

Hoover, Sonderegger, Piantadosi, and O'Donnell. (2022). The plausibility of sampling as an algorithmic theory of sentence processing. Preprint.

Levy (2005). *Probabilistic models of word order and syntactic discontinuity*. PhD thesis, Stanford University.

Levy (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Wagers, Lau, and Phillips (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2), 206–237.

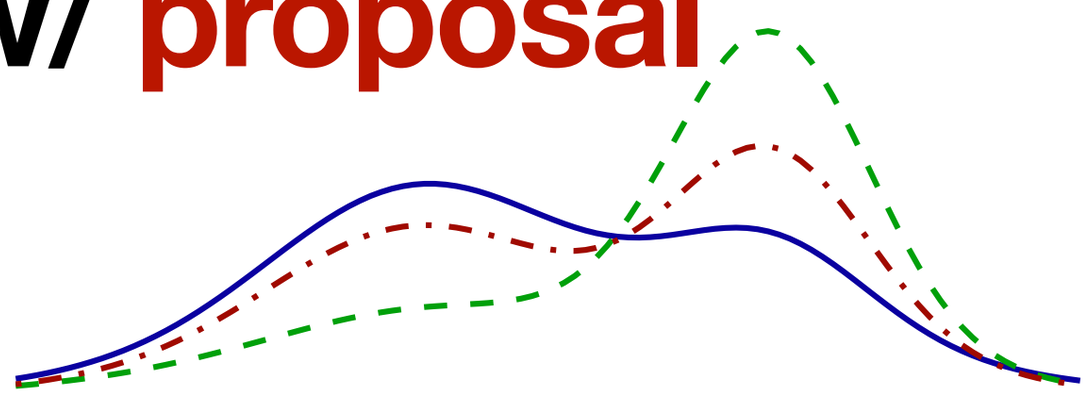
divergence theory

R study: typo/error example

$$\mathrm{D}_{\mathrm{KL}}(p_{Z|\mathbf{u}_{\leq i}} \| p_{Z|\mathbf{u}_{<i}}) = \mathrm{surp}(u_i) - R(u_i)$$

context	word	S	R	KL = S - R	
find my keys on the	table	LOW	LOW	LOW	
''	double	HIGH	LOW	HIGH	
''	hmmrph	HIGH	HIGH	LOW	
''	fable	MID	MID	LOW	

divergence theory w/ proposal



if you use some $q(\,\cdot\,; \mathbf{u}_{\leq i})$, rather than sampling from the prior $p_{Z|\mathbf{u}_{<i}}$, introduce one final term:

$$\mathrm{D}_{\mathrm{KL}}(p_{Z|\mathbf{u}_{\leq i}} \| q) = \overbrace{\log \frac{1}{p(u_i \mid \mathbf{u}_{<i})}}^{\mathrm{surp}(u_i)} - \Bigg(\overbrace{\mathop{\mathbb{E}}_{z|\mathbf{u}_{\leq i}} \log \frac{1}{p(u_i \mid z)}}^{R(u_i)} + \overbrace{\mathop{\mathbb{E}}_{z|\mathbf{u}_{\leq i}} \log \frac{q}{p_{Z|\mathbf{u}_{<i}}}}^{D(u_i)} \Bigg)$$

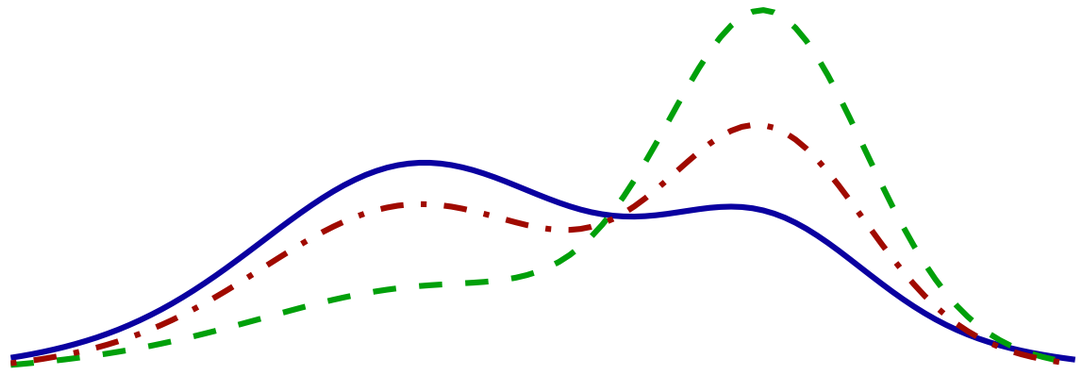
- $D > 0$ if proposal q is on average better than prior $p_{Z|\mathbf{u}_{\leq i}}$
- $D < 0$ if proposal q is on average worse than prior $p_{Z|\mathbf{u}_{\leq i}}$
- $D = 0$ if $q = p_{Z|\mathbf{u}_{\leq i}}$

D is difference in KLs, (the *reduction* in excess surprise from using proposal q instead of the prior $p_{Z|\mathbf{u}_{\leq i}}$)
or equivalently, difference in cross-entropies:

$$\begin{aligned} \mathop{\mathbb{E}}_{z\sim p} \log \frac{q}{p_{Z|\mathbf{u}_{<i}}} &= \mathrm{D}_{\mathrm{KL}}(p_{Z|\mathbf{u}_{\leq i}} \| p_{Z|\mathbf{u}_{<i}}) - \mathrm{D}_{\mathrm{KL}}(p_{Z|\mathbf{u}_{\leq i}} \| q) \\ &= \mathrm{H}[p_{Z|\mathbf{u}_{\leq i}}, p_{Z|\mathbf{u}_{<i}}] - \mathrm{H}[p_{Z|\mathbf{u}_{\leq i}}, q] \end{aligned}$$

divergence theory

sampling from a proposal



$$\mathrm{D}_{\mathrm{KL}}(p_{Z|\mathbf{u}_{\leq i}} \| q) = \overbrace{\log \frac{1}{p(u_i \mid \mathbf{u}_{<i})}}^{\mathrm{surp}(u_i)} - \Bigg(\overbrace{\mathop{\mathbb{E}}_{z\sim p} \log \frac{1}{p(u_i \mid z)}}^{R(u_i)} + \overbrace{\mathop{\mathbb{E}}_{z\sim p} \log \frac{q}{p_{Z|\mathbf{u}_{<i}}}}^{D(u_i)} \Bigg)$$