

Introduction to the Reading Group

On some aspects of language

13 October 2026

Motivation

1. **Structure** of language and language tasks

2. **Utility** of language and communication

Motivation

1. **Structure** of language and language tasks

- How/why does language acquisition give rise to emergent behavior?
- Extrapolation: ability to reason in completely new settings.
- Communication of novel ideas: the language can transmit and map to meaning that the receiver has not previously encountered before.

Motivation

2. **Utility** of language and communication

- How does/should communication arise and take shape?
 - Multi-agent systems benefit from/take advantage of communication.
 - We need reliable human–machine communication: how we specify the task to the machine, and how a machine explain its decisions to us.

Dynamics of Language

Natural language is a byproduct of a social process: its structure is shaped by how we **learn**, **use**, and **transmit** it.

Dynamics of Language

Natural language is a byproduct of a social process: its structure is shaped by how we **learn**, **use**, and **transmit** it.

Question: can language be made theoretically tractable through dynamics?

Dynamics of Language

Natural language is a byproduct of a social process: its structure is shaped by how we **learn**, **use**, and **transmit** it.

Question: can language be made theoretically tractable through dynamics?

- If language evolves to solve a very complicated multi-objective problem, what is that problem?

A Biased Selection of Readings

Readings

Theme: language learning and transmission

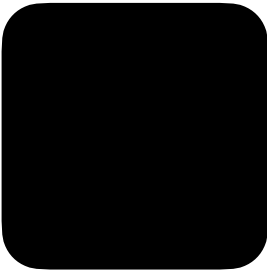
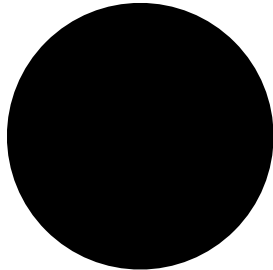
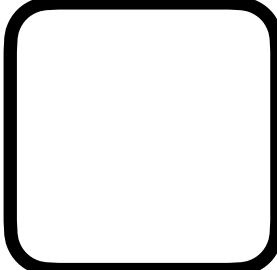
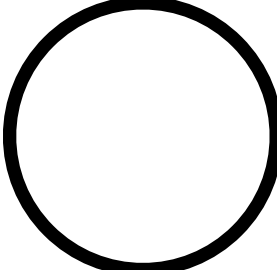
Readings

Theme: language learning and transmission

- Kirby (2017) “Culture and biology in the origins of linguistic structure” [survey]

A language must be learnable if it is to survive the “transmission bottleneck”. This forces the language to become more compositional, linguists argue.

Types of Languages

				
Holistic	A	B	C	D
Compositional	BS	BC	WS	WC
Degenerate	X	X	X	X

Compositionality allows us to communicate new meanings to each other, without requiring the recipient to have encountered that meaning before.

Readings

Theme: language learning and transmission

- Kirby (2017) “Culture and biology in the origins of linguistic structure” [survey]
- Gold (1967) “Language identification in the limit”
- Angluin (1980) “Inductive inference of formal languages from positive data”
- Kleinberg and Mullainathan (2024) “Language generation in the limit”

This series of works aim to characterize how hard it is to learn a language from positive examples.

Readings

The power distribution observed over natural, compositional data seems to be in some sense good for learnability.

Theme: language learning and transmission

- Kirby (2017) “Culture and biology in the origins of linguistic structure” [survey]
 - Gold (1967) “Language identification in the limit”
 - Angluin (1980) “Inductive inference of formal languages from positive data”
 - Kleinberg and Mullainathan (2024) “Language generation in the limit”
-
- Wang, Dang, Lee, Lyu (2026) “The power of power law: asymmetry enables compositional reasoning”

Readings

How do you teach a learner if you do not know its internal mechanism?

Theme: language learning and transmission

- Kirby (2017) “Culture and biology in the origins of linguistic structure” [survey]
 - Gold (1967) “Language identification in the limit”
 - Angluin (1980) “Inductive inference of formal languages from positive data”
 - Kleinberg and Mullainathan (2024) “Language generation in the limit”
-
- Wang, Dang, Lee, Lyu (2026) “The power of power law: asymmetry enables compositional reasoning”
 - Dasgupta, Hsu, Poulis, Zhu (2019) “Teaching a black-box learner”

Readings

Some formal models of how conventions and language evolve over time.

Theme: language learning and transmission

- Kirby (2017) “Culture and biology in the origins of linguistic structure” [survey]
- Gold (1967) “Language identification in the limit”
- Angluin (1980) “Inductive inference of formal languages from positive data”
- Kleinberg and Mullainathan (2024) “Language generation in the limit”
- Peyton Young (1993) “The evolution of conventions”
- Griffiths and Salish (2007) “Language evolution by iterated learning with Bayesian agents”
- Perfors and Navarro (2014) “Language evolution can be shaped by the structure of the world”
- Dmitriev, Franck, Heinzler, Sanyal (2025) “Learning in an echo chamber”
- Wang, Dang, Lee, Lyu (2026) “The power of power law: asymmetry enables compositional reasoning”
- Dasgupta, Hsu, Poulis, Zhu (2019) “Teaching a black-box learner”

Readings

Theme: utility of language and communication

Readings

Economics of information asymmetry.

Theme: utility of language and communication

- Crawford and Sobel (1982) “Strategic information transmission”
- Kamenica and Gentzkow (2011) “Bayesian persuasion”
- De Silva and Sigmund (2024) “Dynamics of signaling games” [survey]
- Brânzei and Peres (2021) “Multiplayer bandit learning: from competition to cooperation”
- Halpern and Moses (2000) “Knowledge and common knowledge in a distributed environment”

Readings

Language arising as a byproduct of cooperative multi-agent systems.

Theme: utility of language and communication

- Crawford and Sobel (1982) “Strategic information transmission”
- Kamenica and Gentzkow (2011) “Bayesian persuasion”
- De Silva and Sigmund (2024) “Dynamics of signaling games” [survey]
- Brânzei and Peres (2021) “Multiplayer bandit learning: from competition to cooperation”
- Halpern and Moses (2000) “Knowledge and common knowledge in a distributed environment”
- Wang, Liang, Manning (2016) “Learning language games through interaction”
- Kottur, Moura, Lee, Batra (2017) “Natural language does not emerge ‘naturally’ in multi-agent dialog”
- Lowe, Foerster, Boureau, Pineau, Dauphin (2019) “On the pitfalls of measuring emergent communication”
- [emergence.ai](#) (2026) “Emergence World: Adversarial stress-testing of long-horizon multi-agent systems”

Readings

Does learning language imply successful task generalization?

Theme: utility of language and communication

- Crawford and Sobel (1982) “Strategic information transmission”
- Kamenica and Gentzkow (2011) “Bayesian persuasion”
- De Silva and Sigmund (2024) “Dynamics of signaling games” [survey]
- Brânzei and Peres (2021) “Multiplayer bandit learning: from competition to cooperation”
- Halpern and Moses (2000) “Knowledge and common knowledge in a distributed environment”
- Wang, Liang, Manning (2016) “Learning language games through interaction”
- Kottur, Moura, Lee, Batra (2017) “Natural language does not emerge ‘naturally’ in multi-agent dialog”
- Lowe, Foerster, Boureau, Pineau, Dauphin (2019) “On the pitfalls of measuring emergent communication”
- [emergence.ai](#) (2026) “Emergence World: Adversarial stress-testing of long-horizon multi-agent systems”
- Bachmann and Nagarajan (2024) “The pitfalls of next-token prediction”

Schedule

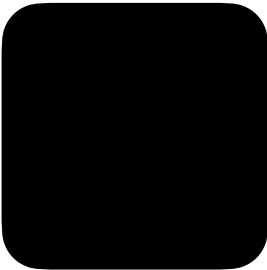
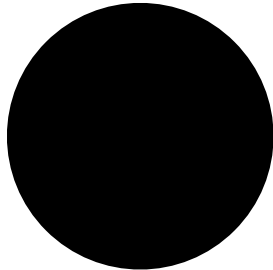
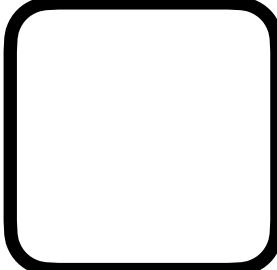
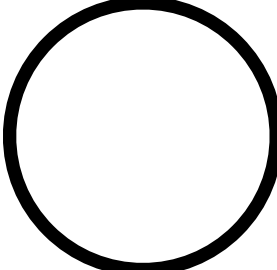
Date	Topic	Presenter
Tue 13.10	Structure and Dynamics of Language	Geelon
Tue 20.10	Emergence of Communication in Multi-Agent Systems	Giovanni
Tue 27.10	Learning to Identify/Generate Language	Tobi
Wed 11.11	Learning from Compositional Data	Julia
Wed 25.11	Language Generation with Replay—Theory of Model Collapse	Giorgio
Tue 1.12		

Learnability and Dynamics of Language

Linguistics Background

- Human languages seem to exhibit **linguistic universals** (patterns or commonalities shared by all natural languages).
 - e.g. compositionality/systematicity

Types of Languages

				
Holistic	A	B	C	D
Compositional	BS	BC	WS	WC
Degenerate	X	X	X	X

Compositionality allows us to communicate new meanings to each other, without requiring the recipient to have encountered that meaning before.

From the Worst-Case to the Best-Case World

For a language to persist...it must pass through a bottleneck of use. The bottleneck on cultural transmission leads inevitably to cumulative increase in the generalizability of the language being transmitted.

Kirby (2017)

From the Worst-Case to the Best-Case World

For a language to persist...it must pass through a bottleneck of use. The bottleneck on cultural transmission leads inevitably to cumulative increase in the generalizability of the language being transmitted.

Instead of asking what the *learner* needs to successfully acquire the target grammar, we ask what the *language* needs to be successfully acquired.

Kirby (2017)

From the Worst-Case to the Best-Case World

For a language to persist...it must pass through a bottleneck of use. The bottleneck on cultural transmission leads inevitably to cumulative increase in the generalizability of the language being transmitted.

Instead of asking what the *learner* needs to successfully acquire the target grammar, we ask what the *language* needs to be successfully acquired.

In other words, the learner can expect to be born into the best of all possible worlds for learning. This stands in stark contrast to some foundational results in [learning theory], which consider...the worst of all possible worlds for learning.

Kirby (2017)

Linguistics Question

Does compositionality/systematicity of natural language come from the need for language to be robustly transmissible?

Language Evolution by Iterated Learning with Bayesian Agents

Griffiths and Kalish (2007)

Language Evolution by Iterated Learning with Bayesian Agents

Griffiths and Kalish (2007)

Spoiler: they will not be able to explain compositionality, but they give an interesting formalization of language learning and also a nice example for an analysis of Markov chains.

Formal Setting

- Let $\mathcal{X}$ be a set of **meaning** (instances).

Formal Setting

- Let $\mathcal{X}$ be a set of **meaning** (instances).
- Let $\mathcal{Y}$ be a set of **utterances** (labels).

Formal Setting

- Let $\mathcal{X}$ be a set of **meaning** (instances).
- Let $\mathcal{Y}$ be a set of **utterances** (labels).
- Let $h : \mathcal{X} \rightarrow \mathcal{Y}$ denote a **language** (model).

Formal Setting

- Let $\mathcal{X}$ be a set of **meaning** (instances).
- Let $\mathcal{Y}$ be a set of **utterances** (labels).
- Let $h : \mathcal{X} \rightarrow \mathcal{Y}$ denote a **language** (model).
- Let $\mathcal{H}$ be a **set of languages** (model class).

Formal Setting

- Let $\mathcal{X}$ be a set of **meaning** (instances).
- Let $\mathcal{Y}$ be a set of **utterances** (labels).
- Let $h : \mathcal{X} \rightarrow \mathcal{Y}$ denote a **language** (model).
- Let $\mathcal{H}$ be a **set of languages** (model class).
- Let $(L_t)_t$ be a **sequence of language learners**.

Transmission Protocol

For $t = 1, 2, \dots$:

Transmission Protocol

For $t = 1, 2, \dots$:

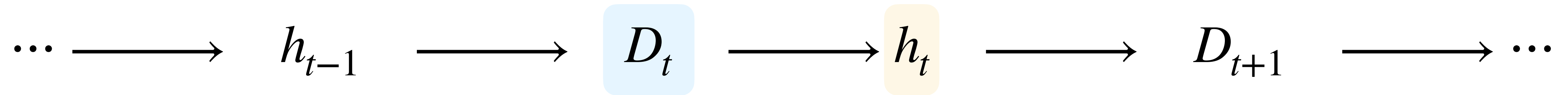
- Learner L_{t-1} generates a labeled dataset $D_t = \{(x, y), \dots\}$.

Transmission Protocol

For $t = 1, 2, \dots$:

- Learner L_{t-1} generates a labeled dataset $D_t = \{(x, y), \dots\}$.
- Learner L_t learns a model h_t using D_t .

Transmission Protocol



Data produced by generation $t - 1$ is what generation t learns from.

Question

How does language evolve in this model?

Goal in linguistics: can learning dynamics explain linguistic universals?

Parts of the Dynamics

- **Learning mechanism:** Which language is acquired from data?

$$\mathbf{M}_{\text{learn}} : D_t \mapsto h_t$$

Parts of the Dynamics

- **Learning mechanism:** Which language is acquired from data?

$$\mathbf{M}_{\text{learn}} : D_t \mapsto h_t$$

- **Generation mechanism:** Which utterances are transmitted?

$$\mathbf{M}_{\text{gen}} : h_t \mapsto D_{t+1}$$

Bayesian Learning Mechanism

- Each learner begins with a prior $\pi(h_{t-1})$ over the previous generation's language.

Bayesian Learning Mechanism

- Each learner begins with a prior $\pi(h_{t-1})$ over the previous generation's language.
- It also has a model $\pi(D_t | h_{t-1})$ for how utterances are generated.

Bayesian Learning Mechanism

- Each learner begins with a prior $\pi(h_{t-1})$ over the previous generation's language.
- It also has a model $\pi(D_t | h_{t-1})$ for how utterances are generated.
- Upon observing a dataset D_t , a learner computes the posterior $\mu_t \equiv \pi(h | D_t)$

$$\mu_t(h) = \frac{\pi(h, D_t)}{\pi(D_t)} = \frac{\pi(h) \cdot \pi(D_t | h)}{\sum_{h'} \pi(h') \cdot \pi(D_t | h')} .$$

Bayesian Learning Mechanism

- Each learner begins with a prior $\pi(h_{t-1})$ over the previous generation's language.
- It also has a model $\pi(D_t | h_{t-1})$ for how utterances are generated.
- Upon observing a dataset D_t , a learner computes the posterior $\mu_t \equiv \pi(h | D_t)$

$$\mu_t(h) = \frac{\pi(h, D_t)}{\pi(D_t)} = \frac{\pi(h) \cdot \pi(D_t | h)}{\sum_{h'} \pi(h') \cdot \pi(D_t | h')} .$$

- The learner can choose h_t by
 - solving for the language maximizing μ_t .

Bayesian Learning Mechanism

- Each learner begins with a prior $\pi(h_{t-1})$ over the previous generation's language.
- It also has a model $\pi(D_t | h_{t-1})$ for how utterances are generated.
- Upon observing a dataset D_t , a learner computes the posterior $\mu_t \equiv \pi(h | D_t)$

$$\mu_t(h) = \frac{\pi(h, D_t)}{\pi(D_t)} = \frac{\pi(h) \cdot \pi(D_t | h)}{\sum_{h'} \pi(h') \cdot \pi(D_t | h')} .$$

- The learner can choose h_t by
 - solving for the language maximizing μ_t .
 - sampling from the posterior μ_t .

Bayesian Learning Mechanism

- Each learner begins with a prior $\pi(h_{t-1})$ over the previous generation's language.
- It also has a model $\pi(D_t | h_{t-1})$ for how utterances are generated.
- Upon observing a dataset D_t , a learner computes the posterior $\mu_t \equiv \pi(h | D_t)$

$$\mu_t(h) = \frac{\pi(h, D_t)}{\pi(D_t)} = \frac{\pi(h) \cdot \pi(D_t | h)}{\sum_{h'} \pi(h') \cdot \pi(D_t | h')}.$$

- The learner can choose h_t by
 - solving for the language maximizing μ_t .
 - sampling from the posterior μ_t .

Goal: understand the evolution of $(\mu_t)_t$.

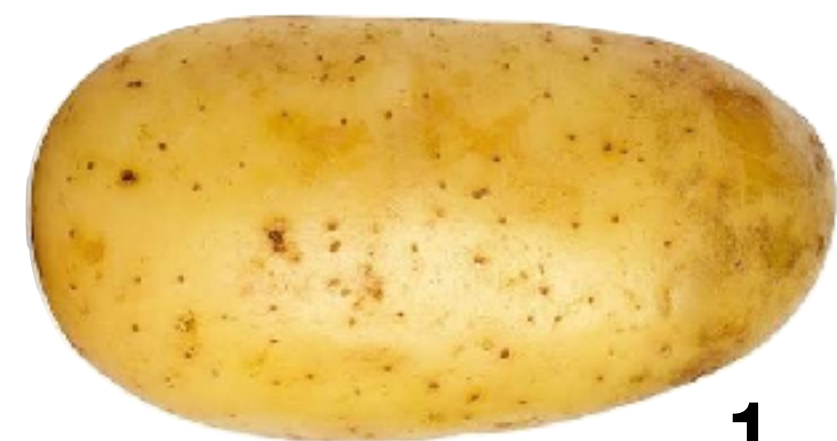
Warm Up

Warm Up

- Let $\mathcal{H} = \{h_A, h_B\}$ consist of **two binary-valued languages** (Quine 1960).

$$h(x) \in \{ \text{Gavagai}, \text{Not Gavagai} \}$$

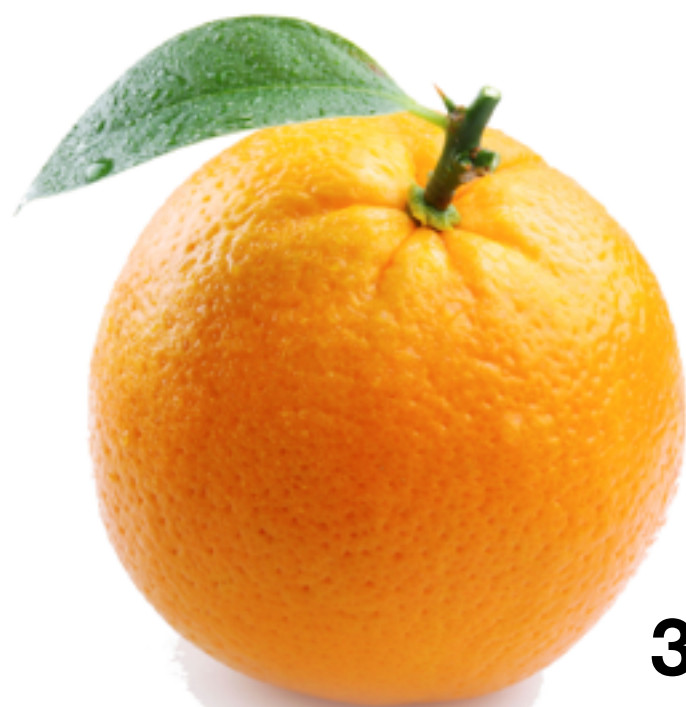
Warm Up



1



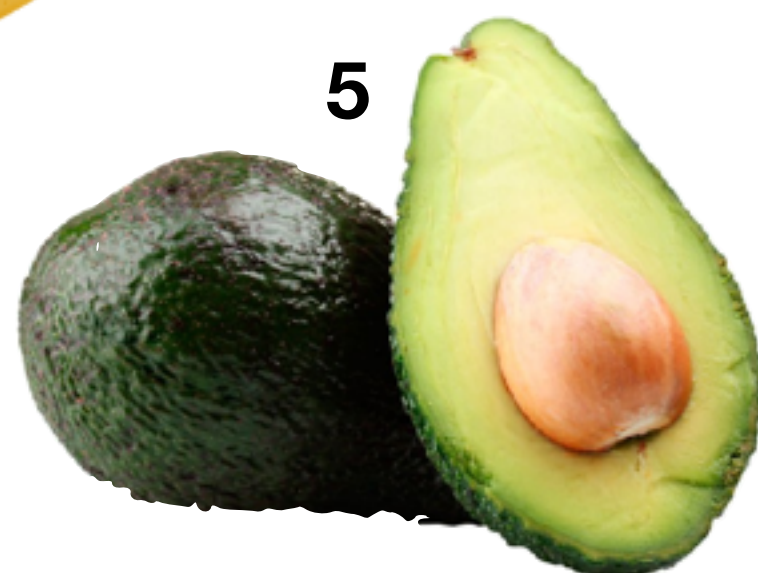
2



3



4



5



6



7



8



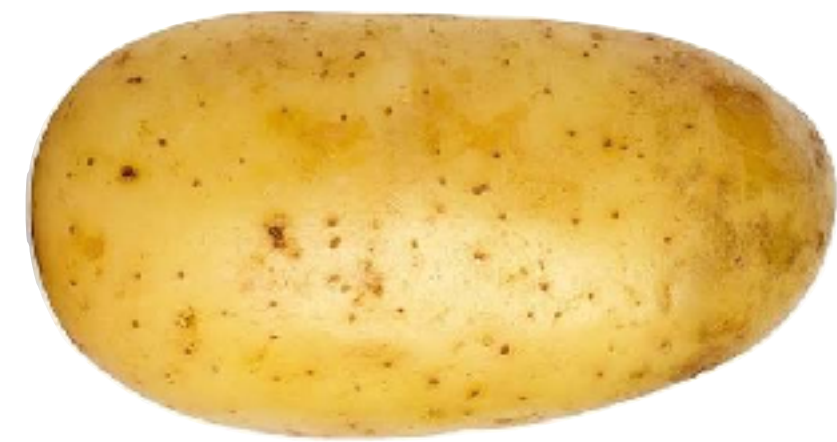
9



10

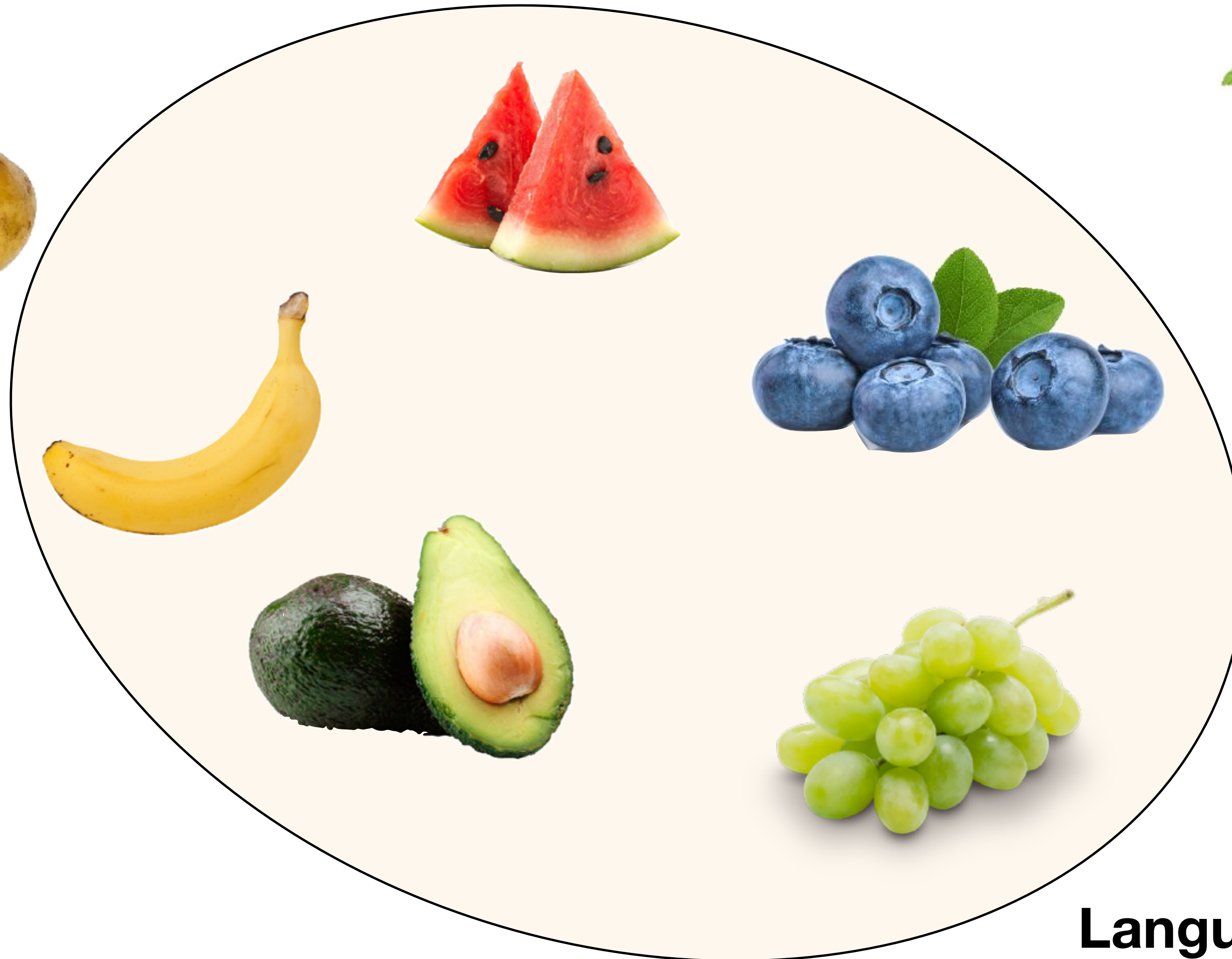
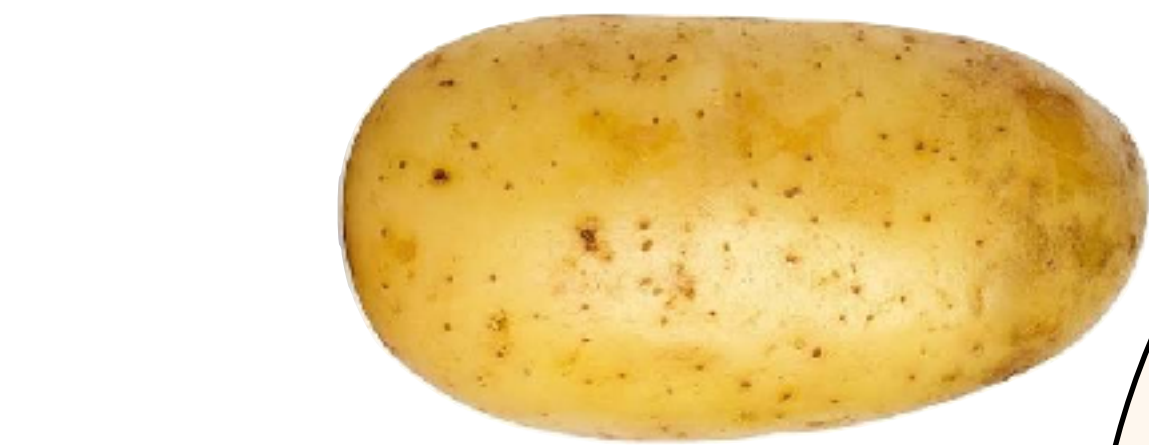
Instance space $\mathcal{X}$

Warm Up



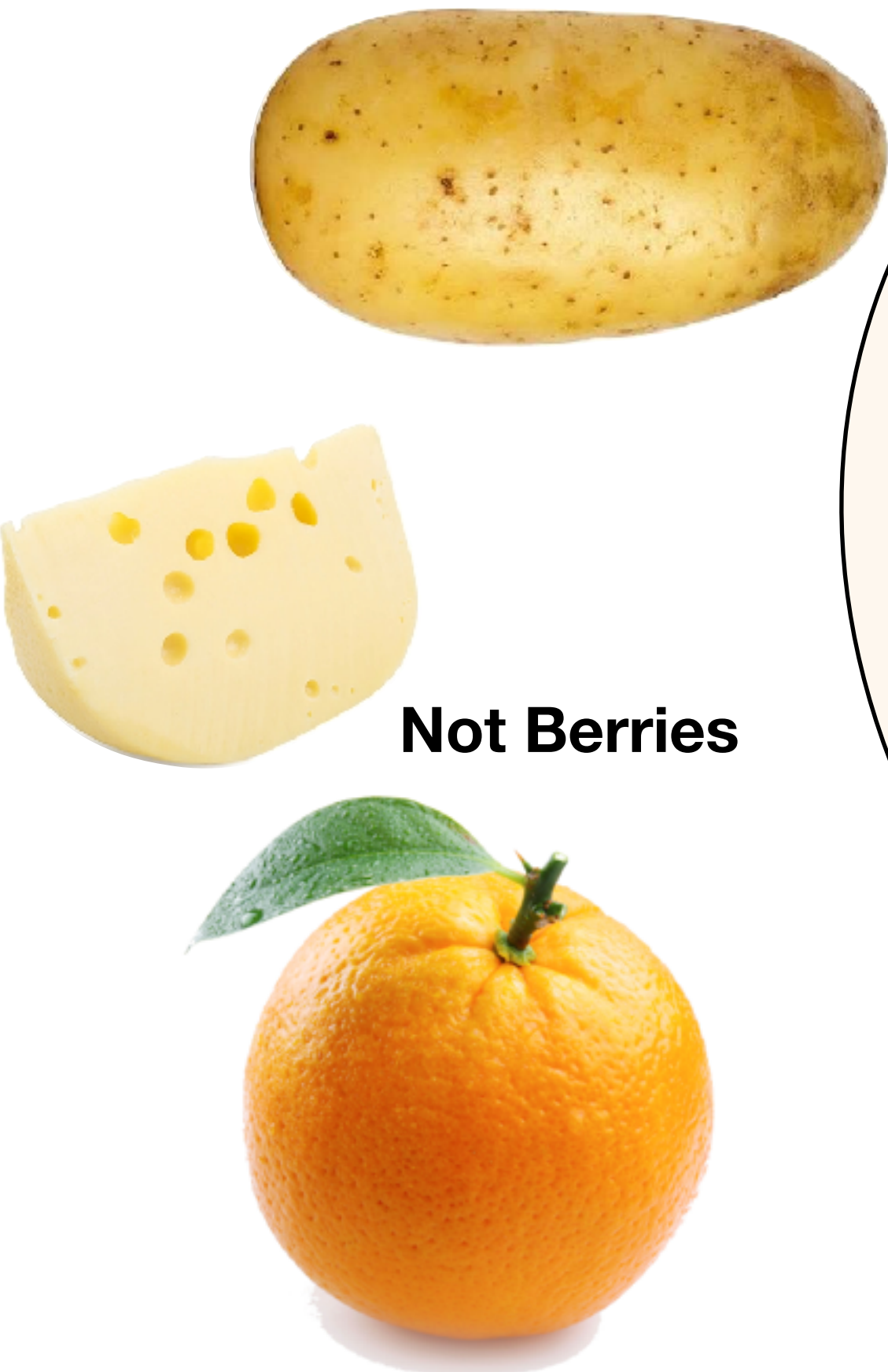
Language A: “Gavagai!”

Warm Up

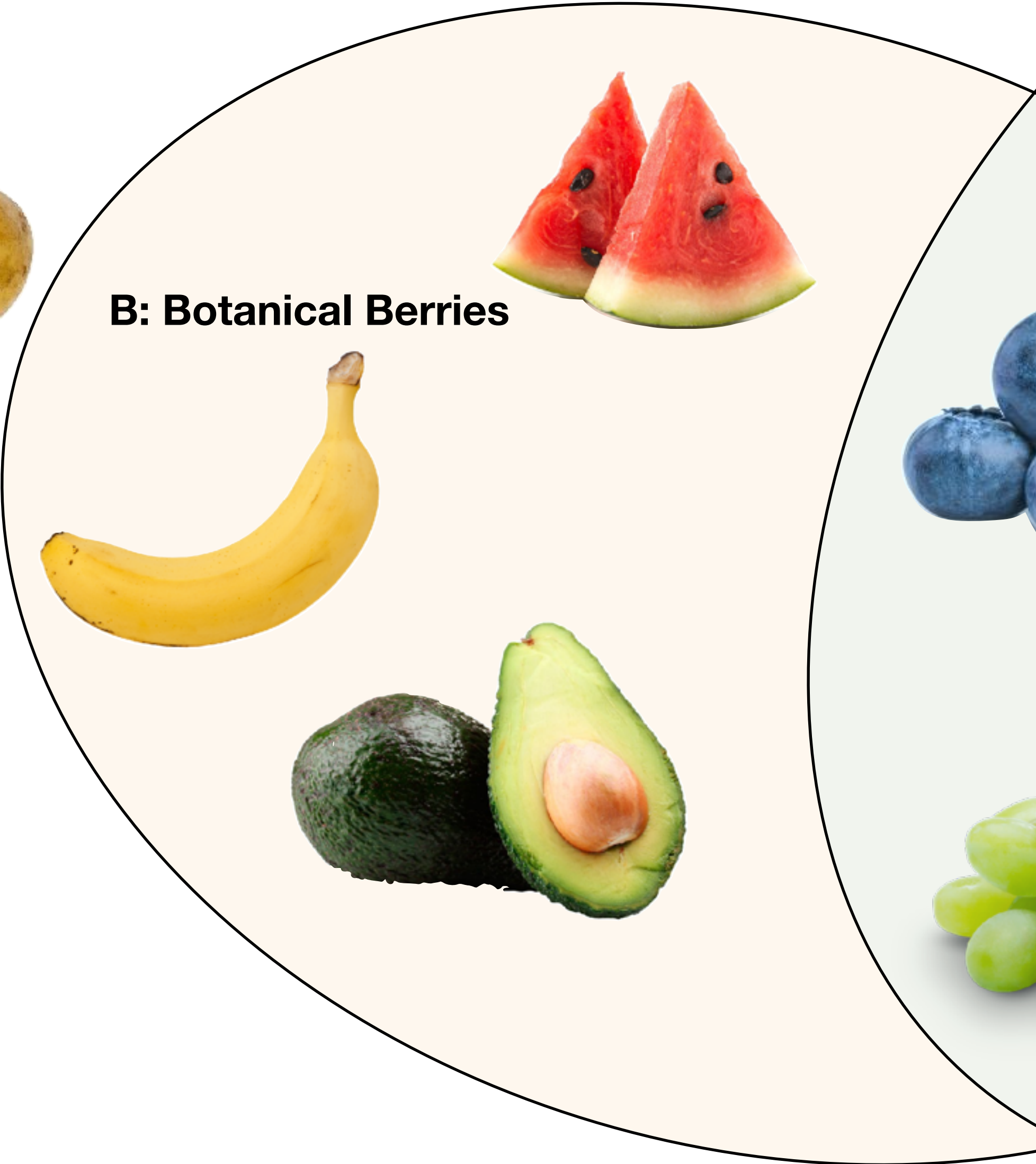


Language B: “Gavagai!”

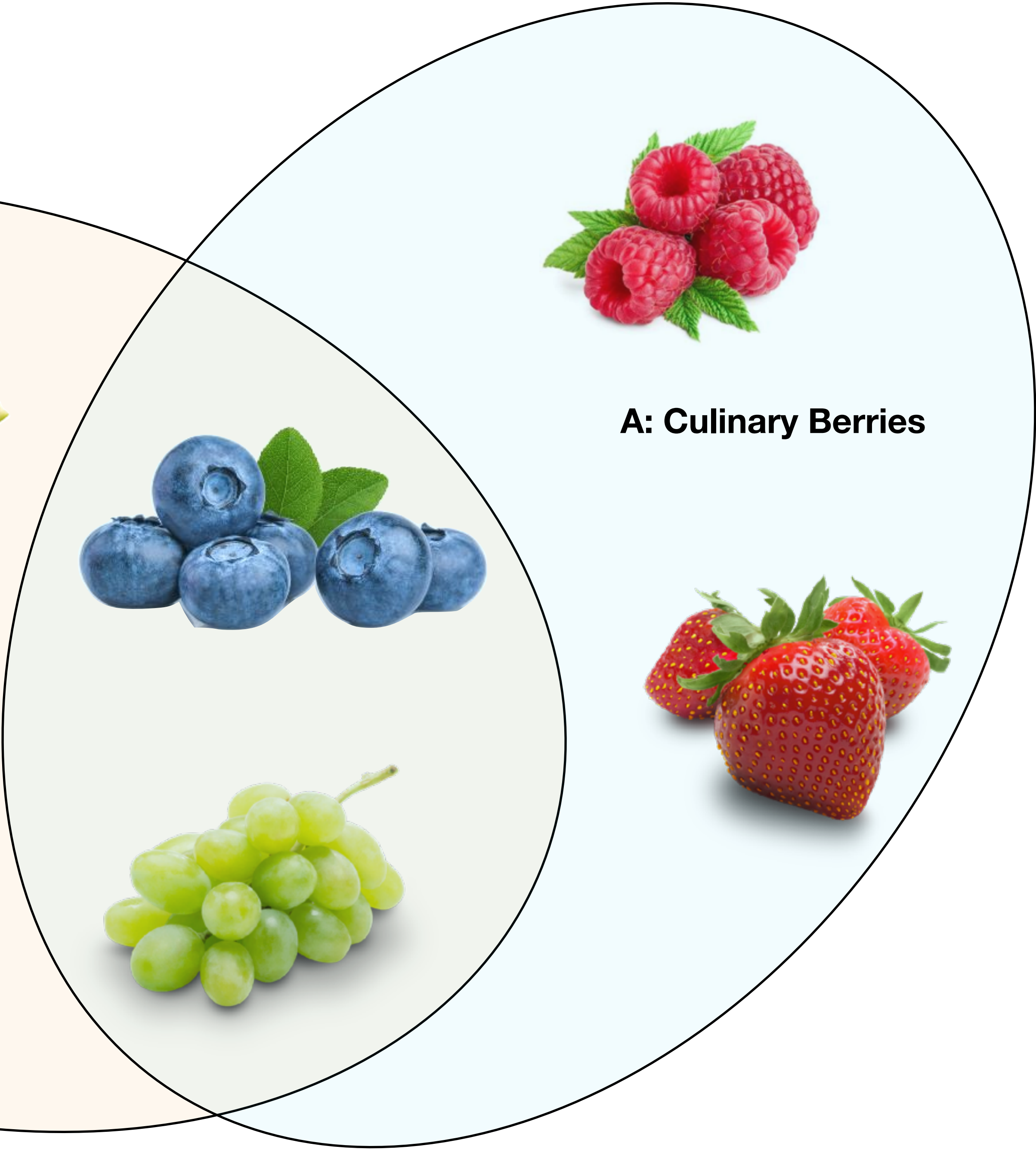
Warm Up



Not Berries



B: Botanical Berries



A: Culinary Berries

Warm Up: 1-example transmission

- Each generation **transmits a single data point** (x_t, y_t) :
 - Nature selects a random instance $x_t \sim \text{Unif}(\mathcal{X})$
 - The previous generation labels it $D_t = (x_t, h_{t-1}(x_t))$

Warm Up: 1-example transmission

- Each generation **transmits a single data point** (x_t, y_t) :
 - Nature selects a random instance $x_t \sim \text{Unif}(\mathcal{X})$
 - The previous generation labels it $D_t = (x_t, h_{t-1}(x_t))$
- The next generation learns the **maximum *a posteriori* (MAP) estimator**:

$$h_t = \arg \max_h \mu_t(h).$$

Warm Up: 1-example transmission

Lemma. Let $p > 0$ be the fraction of instances on which both languages agree.
Let $h^\star = \arg \max_h \pi(h)$.

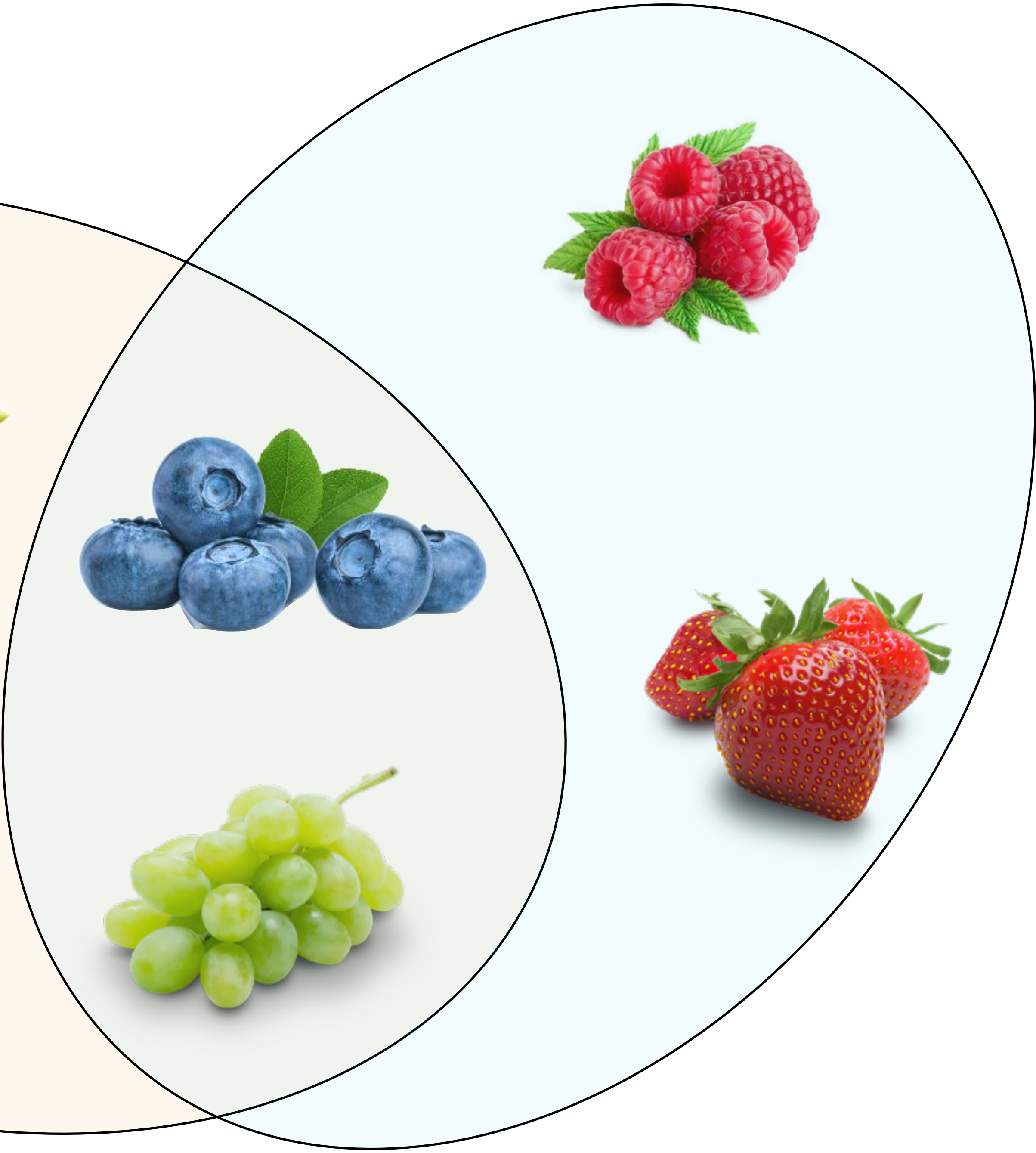
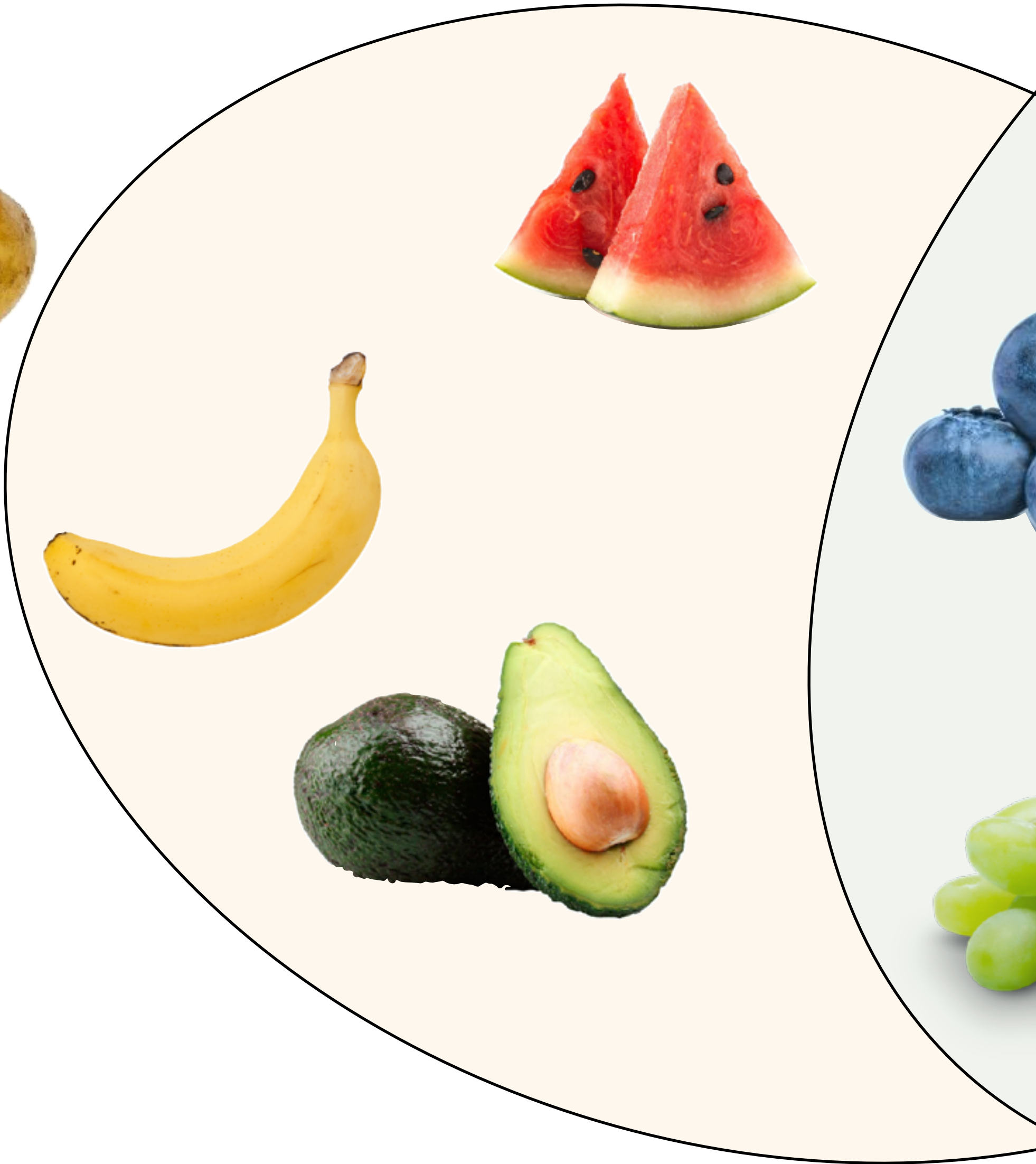
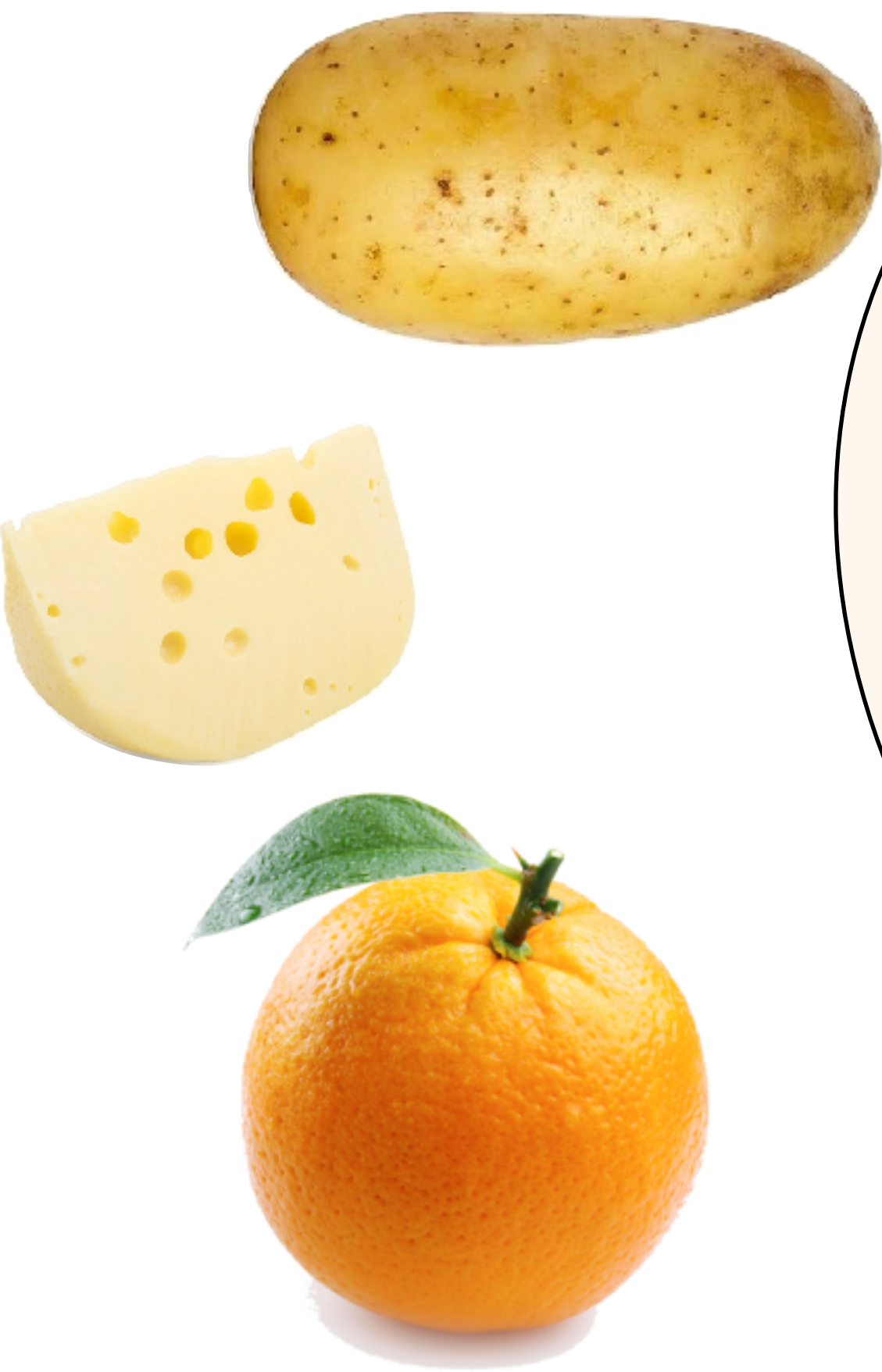
Warm Up: 1-example transmission

Lemma. Let $p > 0$ be the fraction of instances on which both languages agree.
Let $h^\star = \arg \max_h \pi(h)$.

- If each learner computes a MAP estimate, then with probability at least $1 - \delta$, the language converges to $h_t = h^\star$ for all $t > T$ where

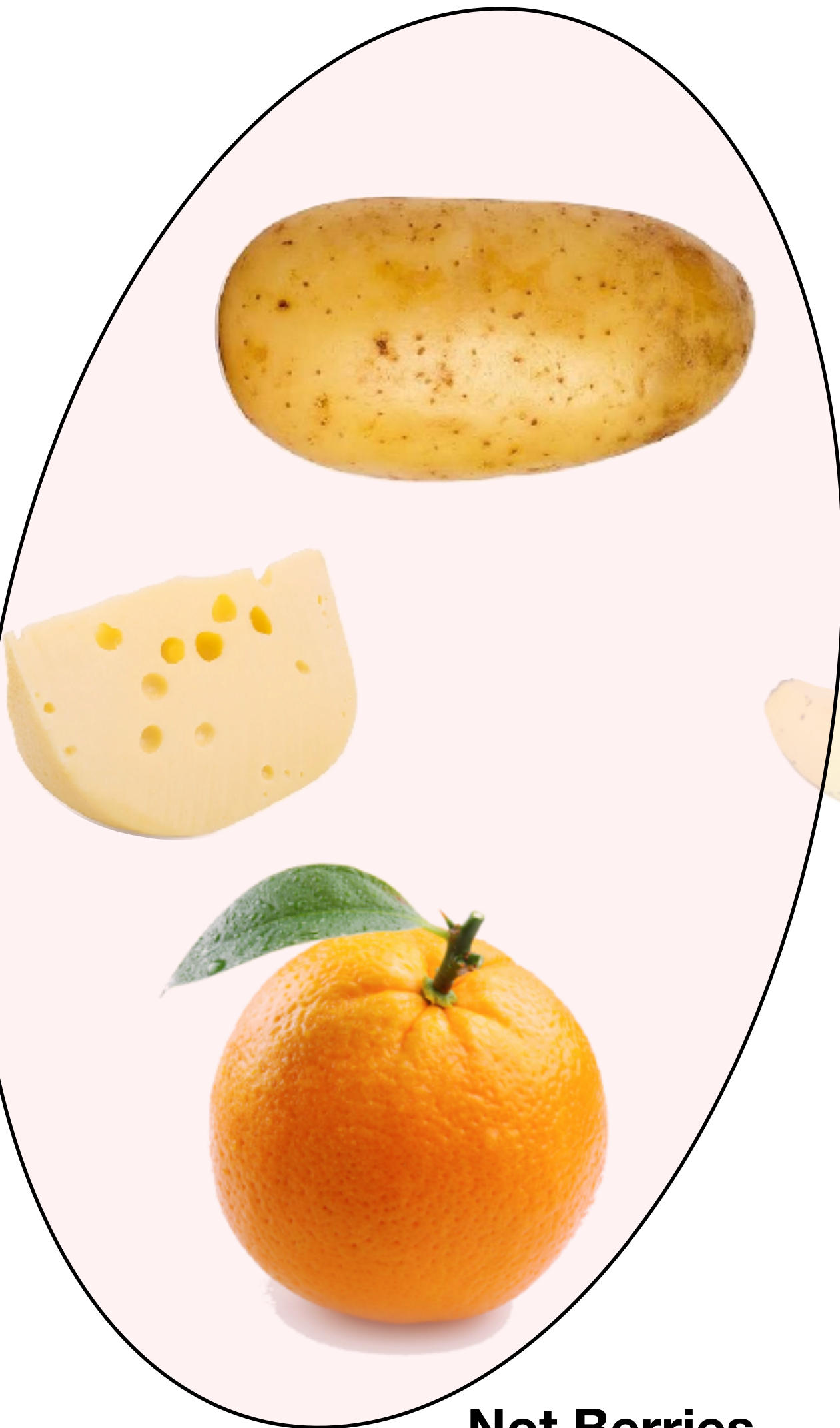
$$T > \frac{1}{p} \log \frac{1}{\delta}.$$

Intuition

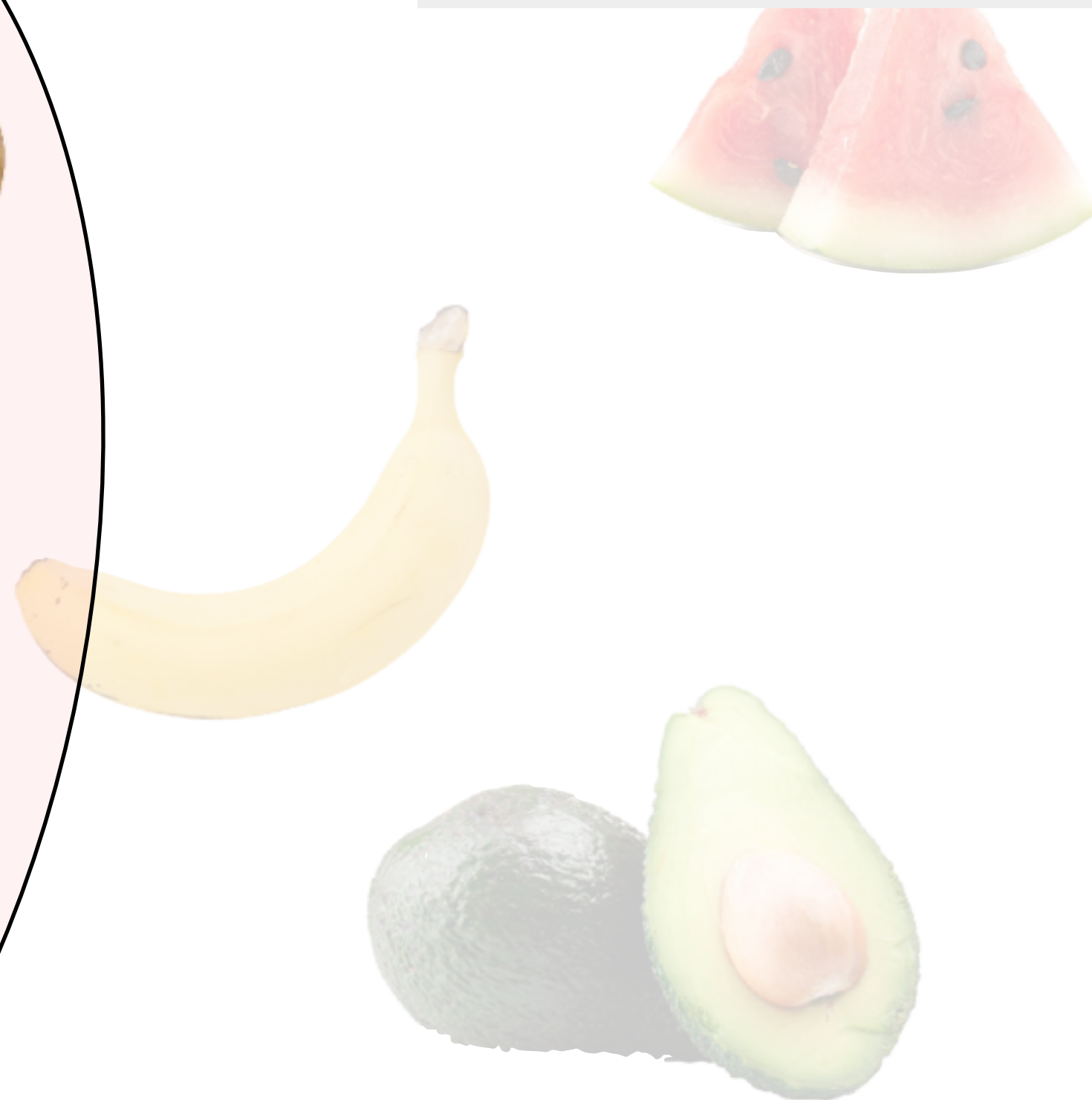


Intuition

If both languages agree on x_t , then the transmitted data fails to provide any information about the prior generation's language.



Not Berries



Berries



Warm Up: 1-example transmission

Proof Sketch.

- The language h_0 is preserved while x_t comes from the disagreement region.
- Language transmission fails with probability p each generation.

Warm Up: 1-example transmission

Proof Sketch.

- The language h_0 is preserved while x_t comes from the disagreement region.
- Language transmission fails with probability p each generation.
- The past's influence decays exponentially quickly

$$\Pr[\text{successful transmission over } t \text{ rounds}] = (1 - p)^t \leq e^{-pt}.$$

Warm Up: 1-example transmission

Proof Sketch.

- The language h_0 is preserved while x_t comes from the disagreement region.
- Language transmission fails with probability p each generation.
- The past's influence decays exponentially quickly

$$\Pr[\text{successful transmission over } t \text{ rounds}] = (1 - p)^t \leq e^{-pt}.$$

When $t > \frac{1}{p} \log \frac{1}{\delta}$, transmission beyond t rounds occurs w.p. $\leq \delta$.

Warm Up: 1-example transmission

Takeaway, according to Griffiths and Kalish (2007):

- The prior governs the long-term dynamics of the language.

Question. How should we interpret the prior?

Warm Up: **noisy** transmission model

- Let $\mathcal{H} = \{h_A, h_B\}$ be constant maps $h_A \equiv A$ and $h_B \equiv B$.

Warm Up: **noisy** transmission model

- Let $\mathcal{H} = \{h_A, h_B\}$ be constant maps $h_A \equiv A$ and $h_B \equiv B$.
- Each generation **transmits a set of data** $(x_1, y_1), \dots$

Warm Up: **noisy** transmission model

- Let $\mathcal{H} = \{h_A, h_B\}$ be constant maps $h_A \equiv A$ and $h_B \equiv B$.
- Each generation **transmits a set of data** $(x_1, y_1), \dots$

Learner's Subjective Model of the Generation Mechanism

1. Sample an instance $x \sim \text{Unif}(\mathcal{X})$
2. Label it $y = \begin{cases} h_{t-1}(x) & \text{w.p. } 1 - \varepsilon \\ \neg h_{t-1}(x) & \text{w.p. } \varepsilon \end{cases}$

Warm Up: **noisy** transmission model

- Let $\mathcal{H} = \{h_A, h_B\}$ be constant maps $h_A \equiv A$ and $h_B \equiv B$.
- Each generation **transmits a set of data** $(x_1, y_1), \dots$

Learner's Subjective Model of the Generation Mechanism

1. Sample an instance $x \sim \text{Unif}(\mathcal{X})$

2. Label it $y = \begin{cases} h_{t-1}(x) & \text{w.p. } 1 - \varepsilon \\ \neg h_{t-1}(x) & \text{w.p. } \varepsilon \end{cases}$

ε -error rate in transmitted data.

Warm Up: **noisy** transmission model

- Let $\mathcal{H} = \{h_A, h_B\}$ be constant maps $h_A \equiv A$ and $h_B \equiv B$.
- Each generation **transmits a set of data** $(x_1, y_1), \dots$

Does not need to reflect how data is objectively generated.

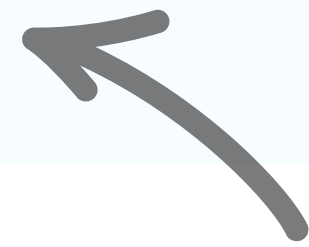


Learner's **Subjective** Model of the Generation Mechanism

1. Sample an instance $x \sim \text{Unif}(\mathcal{X})$

2. Label it $y = \begin{cases} h_{t-1}(x) & \text{w.p. } 1 - \varepsilon \\ \neg h_{t-1}(x) & \text{w.p. } \varepsilon \end{cases}$

ε -error rate in transmitted data.



Warm Up: **noisy** transmission model

- Suppose the **prior** π satisfies $\pi(h_A) \gg \pi(h_B)$.

Warm Up: **noisy** transmission model

- Suppose the **prior** π satisfies $\pi(h_A) \gg \pi(h_B)$.
- Let the learner receive n_A and n_B examples of A and B respectively:

$$D = \{(x_1, A), (x_2, A), (x_3, B), \dots\}.$$

Warm Up: **noisy** transmission model

- Suppose the **prior** π satisfies $\pi(h_A) \gg \pi(h_B)$.
- Let the learner receive n_A and n_B examples of A and B respectively:

$$D = \{(x_1, A), (x_2, A), (x_3, B), \dots\}.$$

- The learner decides language B is more likely only if $\pi(h_A | D) < \pi(h_B | D)$:

$$\frac{\pi(h_A)(1 - \varepsilon)^{n_A}\varepsilon^{n_B}}{\pi(D)} < \frac{\pi(h_B)\varepsilon^{n_A}(1 - \varepsilon)^{n_B}}{\pi(D)}.$$

Warm Up: **noisy** transmission model

- Suppose the **prior** π satisfies $\pi(h_A) \gg \pi(h_B)$.
- Let the learner receive n_A and n_B examples of A and B respectively:

$$D = \{(x_1, A), (x_2, A), (x_3, B), \dots\}.$$

- The learner decides language B is more likely only if $\pi(h_A | D) < \pi(h_B | D)$:

$$n_B > n_A + \log_{1/\varepsilon} \frac{\pi(h_A)}{\pi(h_B)}.$$

Warm Up: **noisy** transmission model

- Suppose the **prior** π satisfies $\pi(h_A) \gg \pi(h_B)$.
- Let the learner receive n_A and n_B examples of A and B respectively:

$$D = \{(x_1, A), (x_2, A), (x_3, B), \dots\}.$$

- The learner decides language B is more likely only if $\pi(h_A | D) < \pi(h_B | D)$:

$$n_B > n_A + \log_{1/\varepsilon} \frac{\pi(h_A)}{\pi(h_B)}.$$

How much evidence for B over A is required for the learner to choose that model.

Bayesian Framework

The prior determines “how much evidence that a learner would need to see in order to adopt a particular language”

Griffiths and Kalish (2007)

Bayesian Framework

The prior determines “how much evidence that a learner would need to see in order to adopt a particular language”

“The Bayesian framework is not supposed to be interpreted as a mechanistic process by which language acquisition takes place...”

Griffiths and Kalish (2007)

Bayesian Framework

The prior determines “how much evidence that a learner would need to see in order to adopt a particular language”

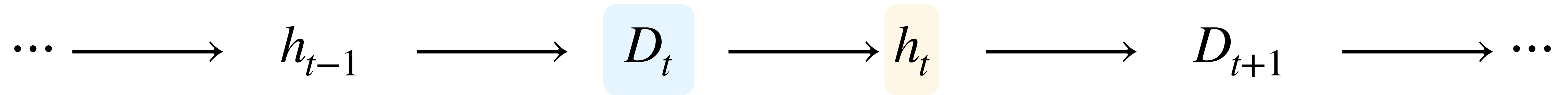
“The Bayesian framework is not supposed to be interpreted as a mechanistic process by which language acquisition takes place...it is a *computational level* analysis, as is generally emphasized in rational models of cognition.”

Griffiths and Kalish (2007)

[*italics* are their emphasis]

Bayesian Iterated Learning

Transmission Protocol



Data produced by generation $t - 1$ is what generation t learns from.

Statistical Production Mechanism

- Let Nature fix a distribution P_X over $\mathcal{X}$.
 - Each generation, a fresh set of instances are drawn from P_X .

Statistical Production Mechanism

- Let Nature fix a distribution P_X over $\mathcal{X}$.
 - Each generation, a fresh set of instances are drawn from P_X .
- The learner speaking language h annotate these instances.

Statistical Production Mechanism

- Let Nature fix a distribution P_X over $\mathcal{X}$.
 - Each generation, a fresh set of instances are drawn from P_X .
- The learner speaking language h annotate these instances.
- Each learner has the same prior π over a finite set of languages $\mathcal{H}$.

Iterated Learning (IL) Dynamics

For $t = 1, 2, \dots$:

Iterated Learning (IL) Dynamics

For $t = 1, 2, \dots$:

- Nature selects a set of contexts $x_1, \dots, x_n \sim P_X$.

Iterated Learning (IL) Dynamics

For $t = 1, 2, \dots$:

- Nature selects a set of contexts $x_1, \dots, x_n \sim P_X$.
- Learner L_{t-1} produces dataset $D_t = \{(x_i, h_{t-1}(x_i)) : i \in [n]\}$.

Iterated Learning (IL) Dynamics

For $t = 1, 2, \dots$:

- Nature selects a set of contexts $x_1, \dots, x_n \sim P_X$.
- Learner L_{t-1} produces dataset $D_t = \{(x_i, h_{t-1}(x_i)) : i \in [n]\}$.
- Learner L_t samples a model $h_t \sim \mu_t$ from posterior

Iterated Learning (IL) Dynamics

For $t = 1, 2, \dots$:

- Nature selects a set of contexts $x_1, \dots, x_n \sim P_X$.
- Learner L_{t-1} produces dataset $D_t = \{(x_i, h_{t-1}(x_i)) : i \in [n]\}$.
- Learner L_t samples a model $h_t \sim \mu_t$ from posterior

$$\mu_t(h) = \frac{\pi(h)\pi(D_t | h)}{\sum_{h'} \pi(h')\pi(D_t | h')}.$$

Gibbs Sampling Detour

The Gibbs Sampler: Motivation

Problem. Let ν be a distribution over $(Z_1, \dots, Z_N)$. We want to sample from ν .

The Gibbs Sampler: Motivation

Problem. Let ν be a distribution over $(Z_1, \dots, Z_N)$. We want to sample from ν .

- It can be hard to sample from ν .

The Gibbs Sampler: Motivation

Problem. Let ν be a distribution over $(Z_1, \dots, Z_N)$. We want to sample from ν .

- It can be hard to sample from ν .
- For some probabilistic models, sampling from conditional distributions can be easy, where $\nu(z_i | Z_{-i})$ is the conditional probability of

$$Z_i | Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_N.$$

The Gibbs Sampler: Motivation

Problem. Let ν be a distribution over $(Z_1, \dots, Z_N)$. We want to sample from ν .

- It can be hard to sample from ν .
- For some probabilistic models, sampling from conditional distributions can be easy, where $\nu(z_i | Z_{-i})$ is the conditional probability of

$$Z_i | Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_N.$$

The Gibbs sampler is an **approximate sampling** algorithm.

The Gibbs Sampler

Algorithm. Initialize $(Z_1, \dots, Z_N)$.

For $t = 1, 2, \dots$:

The Gibbs Sampler

Algorithm. Initialize $(Z_1, \dots, Z_N)$.

For $t = 1, 2, \dots$:

- Let $i \leftarrow t \bmod N$.

The Gibbs Sampler

Algorithm. Initialize $(Z_1, \dots, Z_N)$.

For $t = 1, 2, \dots$:

- Let $i \leftarrow t \bmod N$.
- Resample Z'_i from the conditional distribution $\nu(z_i | Z_{-i})$.

The Gibbs Sampler

Algorithm. Initialize $(Z_1, \dots, Z_N)$.

For $t = 1, 2, \dots$:

- Let $i \leftarrow t \bmod N$.
- Resample Z'_i from the conditional distribution $\nu(z_i | Z_{-i})$.
- Replace the i th component Z_i with Z'_i .

The Gibbs Sampler

Algorithm. Initialize $(Z_1, \dots, Z_N)$.

For $t = 1, 2, \dots$:

- Let $i \leftarrow t \bmod N$.
- Resample Z'_i from the conditional distribution $\nu(z_i | Z_{-i})$.
- Replace the i th component Z_i with Z'_i .

Return $(Z_1, \dots, Z_N)$.

The Gibbs Sampler

Observation. The stationary distribution of the Gibbs dynamics is ν .

The Gibbs Sampler

Observation. The stationary distribution of the Gibbs dynamics is ν .

- Let $(Z_1, \dots, Z_N) \sim \nu$.

The Gibbs Sampler

Observation. The stationary distribution of the Gibbs dynamics is ν .

- Let $(Z_1, \dots, Z_N) \sim \nu$.
- Resample the i th component $Z'_i \sim \nu(z_i | Z_{-i})$.

The Gibbs Sampler

Observation. The stationary distribution of the Gibbs dynamics is ν .

- Let $(Z_1, \dots, Z_N) \sim \nu$.
- Resample the i th component $Z'_i \sim \nu(z_i | Z_{-i})$.
- Both Z_i and Z'_i have conditional distribution $\nu(z_i | Z_{-i})$, thus

$$(Z_1, \dots, Z_{i-1}, Z_i, Z_{i+1}, \dots, Z_N) \stackrel{d}{=} (Z_1, \dots, Z_{i-1}, Z'_i, Z_{i+1}, \dots, Z_N).$$

Analysis of the Iterated Learning Dynamics

Iterated Learning is a Gibbs Sampler

Iterated Learning is a Gibbs Sampler

Define the distribution ν over (h, D) language-dataset pairs where

$h \sim \pi$ and D is a random set of data from $P_X^{\otimes n}$ labeled by h .

Iterated Learning is a Gibbs Sampler

Define the distribution ν over (h, D) language-dataset pairs where

$h \sim \pi$ and D is a random set of data from $P_X^{\otimes n}$ labeled by h .

Iterated Learning Dynamics. For $t = 1, 2, \dots$:

- Nature selects a set of contexts $x_1, \dots, x_n \sim P_X$.
- Learner L_{t-1} produces dataset $D_t = \{(x_1, h_{t-1}(x_1)), \dots\} \sim \nu(D \mid h_{t-1})$.
- Learner L_t samples a model $h_t \sim \mu_t = \nu(h \mid D_t)$ from the posterior.

Analysis of Iterated Learning (IL) Dynamics

Theorem. Let $\mathcal{H}$ be a finite set of languages and $\pi(h) > 0$ for all $h \in \mathcal{H}$.

Analysis of Iterated Learning (IL) Dynamics

Theorem. Let $\mathcal{H}$ be a finite set of languages and $\pi(h) > 0$ for all $h \in \mathcal{H}$. Let h_0 be the initial language.

Analysis of Iterated Learning (IL) Dynamics

Theorem. Let $\mathcal{H}$ be a finite set of languages and $\pi(h) > 0$ for all $h \in \mathcal{H}$. Let h_0 be the initial language. Then, π is the stationary distribution of (IL)

The steady-state distribution over languages is π

Analysis of Iterated Learning (IL) Dynamics

Theorem. Let $\mathcal{H}$ be a finite set of languages and $\pi(h) > 0$ for all $h \in \mathcal{H}$. Let h_0 be the initial language. Then, π is the stationary distribution of (IL), and

$$\|\mu_t - \pi\|_{\text{TV}} \leq C \exp(-\gamma t),$$

exponential convergence rate to π

Analysis of Iterated Learning (IL) Dynamics

Theorem. Let $\mathcal{H}$ be a finite set of languages and $\pi(h) > 0$ for all $h \in \mathcal{H}$. Let h_0 be the initial language. Then, π is the stationary distribution of (IL), and

$$\|\mu_t - \pi\|_{\text{TV}} \leq C \exp(-\gamma t),$$

where $(h, D) \sim \nu$ and γ is given by

$$\gamma = \min_{f: \text{Var}_{\pi}(f) > 0} \frac{\mathbb{E}_D[\text{Var}(f(h) | D)]}{\text{Var}_{\pi}(f)} \quad \text{where } f: \mathcal{H} \rightarrow \mathbb{R}.$$

the rate is controlled by a notion of correlation $1 - \gamma$ between a language h and sample utterances from the language D

Transition Kernel of (IL)

Definition. Let $K(h_t, h_{t+1})$ be the *transition kernel* on $\mathcal{H}$ defined

$$K(h, h') = \sum_D \nu(D | h) \nu(h' | D).$$

Transition Kernel of (IL)

Definition. Let $K(h_t, h_{t+1})$ be the *transition kernel* on $\mathcal{H}$ defined

$$K(h, h') = \sum_D \nu(D | h) \nu(h' | D).$$

- The value of $K(h, h')$ is the probability of transitioning from h to h' .

Transition Kernel of (IL)

Definition. Let $K(h_t, h_{t+1})$ be the *transition kernel* on $\mathcal{H}$ defined

$$K(h, h') = \sum_D \nu(D | h) \nu(h' | D).$$

- The value of $K(h, h')$ is the probability of transitioning from h to h' .
- By definition, we have $\mu_t K = \mu_{t+1}$, where

Transition Kernel of (IL)

Definition. Let $K(h_t, h_{t+1})$ be the *transition kernel* on $\mathcal{H}$ defined

$$K(h, h') = \sum_D \nu(D | h) \nu(h' | D).$$

- The value of $K(h, h')$ is the probability of transitioning from h to h' .
- By definition, we have $\mu_t K = \mu_{t+1}$, where

$$(\mu_t K)(h') = \sum_h \sum_D \mu_t(h) \nu(D | h) \nu(h' | D) = \mu_{t+1}(h').$$

Reversibility of the Transition Kernel of (IL)

Definition. Let $K(h_t, h_{t+1})$ be the *transition kernel* on $\mathcal{H}$ defined

$$K(h, h') = \sum_D \nu(D | h) \nu(h' | D).$$

Lemma. The kernel K is *reversible* with respect to π :

$$\pi(h)K(h, h') = \pi(h')K(h', h).$$

Reversibility of the Transition Kernel of (IL)

Definition. Let $K(h_t, h_{t+1})$ be the *transition kernel* on $\mathcal{H}$ defined

$$K(h, h') = \sum_D \nu(D | h) \nu(h' | D).$$

Lemma. The kernel K is *reversible* with respect to π :

$$\pi(h)K(h, h') = \pi(h')K(h', h).$$

- This implies that π is a *stationary* distribution: the amount of mass flowing from h to h' is equal to the amount of mass flowing from h' to h .

Analysis of Iterated Learning (IL) Dynamics

Theorem. Let $\mathcal{H}$ be a finite set of languages and $\pi(h) > 0$ for all $h \in \mathcal{H}$. Let h_0 be the initial language. Then, π is the stationary distribution of (IL), and

$$\|\mu_t - \pi\|_{\text{TV}} \leq C \exp(-\gamma t),$$

where $(h, D) \sim \nu$ and γ is given by

$$\gamma = \min_{f: \text{Var}_{\pi}(f) > 0} \frac{\mathbb{E}_D[\text{Var}(f(h) | D)]}{\text{Var}_{\pi}(f)} \quad \text{where } f: \mathcal{H} \rightarrow \mathbb{R}.$$

Reversibility of the Transition Kernel of (IL)

Definition. Let $K(h_t, h_{t+1})$ be the *transition kernel* on $\mathcal{H}$ defined

$$K(h, h') = \sum_D \nu(D | h) \nu(h' | D).$$

Lemma. The kernel K is *reversible* with respect to π :

$$\pi(h)K(h, h') = \pi(h')K(h', h).$$

- This implies that π is a *stationary* distribution: the amount of mass flowing from h to h' is equal to the amount of mass flowing from h' to h .
- This condition is called *detailed balance*.

Reversibility of the Transition Kernel of (IL)

Proof of Reversibility. We verify by a direct computation:

$$\pi(h)K(h, h') = \sum_D \pi(h)\nu(D | h)\nu(D | h')$$

Reversibility of the Transition Kernel of (IL)

Proof of Reversibility. We verify by a direct computation:

$$\begin{aligned}\pi(h)K(h, h') &= \sum_D \pi(h)\nu(D | h)\nu(D | h') \\ &= \sum_D \pi(h)\nu(D | h) \cdot \frac{\pi(h')\nu(D | h')}{\sum_{h''} \pi(h'')\nu(D | h'')}\end{aligned}$$

Reversibility of the Transition Kernel of (IL)

Proof of Reversibility. We verify by a direct computation:

$$\begin{aligned}\pi(h)K(h, h') &= \sum_D \pi(h)\nu(D | h)\nu(D | h') \\ &= \sum_D \pi(h)\nu(D | h) \cdot \frac{\pi(h')\nu(D | h')}{\sum_{h''} \pi(h'')\nu(D | h'')}\end{aligned}$$

Reversibility of the Transition Kernel of (IL)

Proof of Reversibility. We verify by a direct computation:

$$\begin{aligned}\pi(h)K(h, h') &= \sum_D \pi(h)\nu(D | h)\nu(D | h') \\ &= \sum_D \pi(h)\nu(D | h) \cdot \frac{\pi(h')\nu(D | h')}{\sum_{h''} \pi(h'')\nu(D | h'')} \\ &= \sum_D \frac{\pi(h)\nu(D | h)}{\sum_{h''} \pi(h'')\nu(D | h'')} \cdot \pi(h')\nu(D | h')\end{aligned}$$

Reversibility of the Transition Kernel of (IL)

Proof of Reversibility. We verify by a direct computation:

$$\begin{aligned}\pi(h)K(h, h') &= \sum_D \pi(h)\nu(D | h)\nu(D | h') \\ &= \sum_D \pi(h)\nu(D | h) \cdot \frac{\pi(h')\nu(D | h')}{\sum_{h''} \pi(h'')\nu(D | h'')} \\ &= \sum_D \frac{\pi(h)\nu(D | h)}{\sum_{h''} \pi(h'')\nu(D | h'')} \cdot \pi(h')\nu(D | h') = \pi(h')K(h', h) .\end{aligned}$$

Eigenvectors of the Transition Kernel of (IL)

Eigenvector of K .

- From reversibility, it follows that π is a left-eigenvector of K with eigenvalue 1

$$\pi K = \pi.$$

Eigenvectors of the Transition Kernel of (IL)

Eigenvector of K .

- From reversibility, it follows that π is a left-eigenvector of K with eigenvalue 1

$$\pi K = \pi.$$

Goal in remaining analysis.

Eigenvectors of the Transition Kernel of (IL)

Eigenvector of K .

- From reversibility, it follows that π is a left-eigenvector of K with eigenvalue 1

$$\pi K = \pi.$$

Goal in remaining analysis.

Show that the eigenvector π corresponds to the largest eigenvalue of K .

Eigenvectors of the Transition Kernel of (IL)

Eigenvector of K .

- From reversibility, it follows that π is a left-eigenvector of K with eigenvalue 1

$$\pi K = \pi.$$

Goal in remaining analysis.

Show that the eigenvector π corresponds to the largest eigenvalue of K .

- Compute the spectral gap between the first and second largest eigenvalue.

Eigenvectors of the Transition Kernel of (IL)

Eigenvector of K .

- From reversibility, it follows that π is a left-eigenvector of K with eigenvalue 1

$$\pi K = \pi.$$

Goal in remaining analysis.

Show that the eigenvector π corresponds to the largest eigenvalue of K .

- Compute the spectral gap between the first and second largest eigenvalue.
- Convergence rate to π is $\exp(-\gamma t)$ where γ is the spectral gap.

Convergence Rate from Spectral Gap

Suppose that K has a set of eigenbasis $v_1, v_2, \dots, v_M$

i.e. assume that K is diagonalizable

Convergence Rate from Spectral Gap

Suppose that K has a set of eigenbasis $v_1, v_2, \dots, v_M$ where

$$\lambda_1 > \lambda_2 \geq \dots \geq \lambda_M$$

with a spectral gap $\gamma = \lambda_1 - \lambda_2$.

Convergence Rate from Spectral Gap

Suppose that K has a set of eigenbasis $v_1, v_2, \dots, v_M$ where

$$\lambda_1 > \lambda_2 \geq \dots \geq \lambda_M$$

with a spectral gap $\gamma = \lambda_1 - \lambda_2$.

Let $\mu_0 = \sum_m \alpha_m v_m$ be the initial language distribution.

Convergence Rate from Spectral Gap

Suppose that K has a set of eigenbasis $v_1, v_2, \dots, v_M$ where

$$\lambda_1 > \lambda_2 \geq \dots \geq \lambda_M$$

with a spectral gap $\gamma = \lambda_1 - \lambda_2$.

Let $\mu_0 = \sum_m \alpha_m v_m$ be the initial language distribution. Then

$$\mu_t = \mu_0 K^t = \sum_m \alpha_m \lambda_m^t v_m$$

Convergence Rate from Spectral Gap

Suppose that K has a set of eigenbasis $v_1, v_2, \dots, v_M$ where

$$\lambda_1 > \lambda_2 \geq \dots \geq \lambda_M$$

with a spectral gap $\gamma = \lambda_1 - \lambda_2$.

Let $\mu_0 = \sum_m \alpha_m v_m$ be the initial language distribution. Then

$$\mu_t = \mu_0 K^t = \sum_m \alpha_m \lambda_m^t v_m \rightarrow \pi$$

Convergence Rate from Spectral Gap

Suppose that K has a set of eigenbasis $v_1, v_2, \dots, v_M$ where

$$\lambda_1 > \lambda_2 \geq \dots \geq \lambda_M$$

with a spectral gap $\gamma = \lambda_1 - \lambda_2$.

Let $\mu_0 = \sum_m \alpha_m v_m$ be the initial language distribution. Then

$$\mu_t = \mu_0 K^t = \sum_m \alpha_m \lambda_m^t v_m \rightarrow \pi$$

since $(\lambda_1, v_1) = (1, \pi)$ and $\lambda_m \leq 1 - \gamma$ for all $m > 1$.

All eigendirections but π decay exponentially fast.

Analysis of Iterated Learning (IL) Dynamics

Theorem. Let $\mathcal{H}$ be a finite set of languages and $\pi(h) > 0$ for all $h \in \mathcal{H}$. Let h_0 be the initial language. Then, π is the stationary distribution of (IL), and

$$\|\mu_t - \pi\|_{\text{TV}} \leq C \exp(-\gamma t),$$

where $(h, D) \sim \nu$ and γ is given by

$$\gamma = \min_{f: \text{Var}_{\pi}(f) > 0} \frac{\mathbb{E}_D[\text{Var}(f(h) | D)]}{\text{Var}_{\pi}(f)}$$

where $f: \mathcal{H} \rightarrow \mathbb{R}$.

Eigenvectors of the Transition Kernel of (IL)

Eigenvector of K .

- From reversibility, it follows that π is a left-eigenvector of K with eigenvalue 1

$$\pi K = \pi.$$

- Convergence rate to π is $\exp(-\gamma t)$ where γ is the spectral gap.

Goal in remaining analysis.

Show that the eigenvector π corresponds to the largest eigenvalue of K .

- Compute the spectral gap between the first and second largest eigenvalue.

Kernel as an Operator

Let $f : \mathcal{H} \rightarrow \mathbb{R}$ be functions on the set of languages.

Kernel as an Operator

Let $f : \mathcal{H} \rightarrow \mathbb{R}$ be functions on the set of languages. The kernel is an operator:

$$(Kf)(h) = \sum_{h'} K(h, h')f(h').$$

Kernel as an Operator

Let $f : \mathcal{H} \rightarrow \mathbb{R}$ be functions on the set of languages. The kernel is an operator:

$$(Kf)(h) = \sum_{h'} K(h, h') f(h').$$

- Let $f(h_t)$ gives a measurement of the language h_t .

Kernel as an Operator

Let $f : \mathcal{H} \rightarrow \mathbb{R}$ be functions on the set of languages. The kernel is an operator:

$$(Kf)(h) = \sum_{h'} K(h, h') f(h').$$

- Let $f(h_t)$ gives a measurement of the language h_t .
- The function $Kf : \mathcal{H} \rightarrow \mathbb{R}$ gives the conditional expectation

$$(Kf)(h_t) = \mathbb{E} [f(h_{t+1}) | h_t].$$

Kernel as an Operator

Let $f : \mathcal{H} \rightarrow \mathbb{R}$ be functions on the set of languages. The kernel is an operator:

$$(Kf)(h) = \sum_{h'} K(h, h') f(h').$$

- Let $f(h_t)$ gives a measurement of the language h_t .
- The function $Kf : \mathcal{H} \rightarrow \mathbb{R}$ gives the conditional expectation

$$(Kf)(h_t) = \mathbb{E} [f(h_{t+1}) | h_t].$$

If f tells you about a language today, Kf tells you about what it will like in expectation tomorrow.

Symmetry of the Kernel

Let $L^2(\pi)$ be the function space consisting of functions $f, g : \mathcal{H} \rightarrow \mathbb{R}$ with

$$\langle f, g \rangle_\pi :=$$

Symmetry of the Kernel

Let $L^2(\pi)$ be the function space consisting of functions $f, g : \mathcal{H} \rightarrow \mathbb{R}$ with

$$\langle f, g \rangle_\pi := \sum_{\mathcal{H}} \pi(h) f(h) g(h)$$

Symmetry of the Kernel

Let $L^2(\pi)$ be the function space consisting of functions $f, g : \mathcal{H} \rightarrow \mathbb{R}$ with

$$\langle f, g \rangle_\pi := \sum_{\mathcal{H}} \pi(h) f(h) g(h) = \mathbb{E}_{h \sim \pi} [f(h) g(h)] .$$

Symmetry of the Kernel

Let $L^2(\pi)$ be the function space consisting of functions $f, g : \mathcal{H} \rightarrow \mathbb{R}$ with

$$\langle f, g \rangle_\pi := \sum_{\mathcal{H}} \pi(h) f(h) g(h) = \mathbb{E}_{h \sim \pi} [f(h) g(h)] .$$

Lemma. The kernel K is symmetric in $L^2(\pi)$. That is,

$$\langle Kf, g \rangle_\pi = \langle f, Kg \rangle_\pi .$$

Symmetry of the Kernel

Let $L^2(\pi)$ be the function space consisting of functions $f, g : \mathcal{H} \rightarrow \mathbb{R}$ with

$$\langle f, g \rangle_\pi := \sum_{\mathcal{H}} \pi(h) f(h) g(h) = \mathbb{E}_{h \sim \pi} [f(h) g(h)] .$$

Lemma. The kernel K is symmetric in $L^2(\pi)$. That is,

$$\langle Kf, g \rangle_\pi = \langle f, Kg \rangle_\pi .$$

This is useful because symmetric operators are diagonalizable.

Symmetry of the Kernel

Proof of symmetry. This follows from reversibility:

$$\langle Kf, g \rangle_\pi = \sum_h \pi(h)(Kf)(h)g(h)$$

Symmetry of the Kernel

Proof of symmetry. This follows from reversibility:

$$\begin{aligned}\langle Kf, g \rangle_\pi &= \sum_h \pi(h)(Kf)(h)g(h) \\ &= \sum_h \sum_{h'} \pi(h)K(h, h')f(h')g(h)\end{aligned}$$

Symmetry of the Kernel

Proof of symmetry. This follows from reversibility:

$$\begin{aligned}\langle Kf, g \rangle_\pi &= \sum_h \pi(h)(Kf)(h)g(h) \\ &= \sum_h \sum_{h'} \pi(h)K(h, h')f(h')g(h)\end{aligned}$$

Symmetry of the Kernel

Proof of symmetry. This follows from reversibility:

$$\begin{aligned}\langle Kf, g \rangle_\pi &= \sum_h \pi(h)(Kf)(h)g(h) \\ &= \sum_h \sum_{h'} \pi(h)K(h, h')f(h')g(h) \\ &= \sum_h \sum_{h'} \pi(h')K(h', h)f(h')g(h)\end{aligned}$$

Symmetry of the Kernel

Proof of symmetry. This follows from reversibility:

$$\begin{aligned}\langle Kf, g \rangle_\pi &= \sum_h \pi(h)(Kf)(h)g(h) \\ &= \sum_h \sum_{h'} \pi(h)K(h, h')f(h')g(h) \\ &= \sum_h \sum_{h'} \pi(h')K(h', h)f(h')g(h) \\ &= \sum_{h'} \pi(h')f(h')(Kg)(h') = \langle f, Kg \rangle_\pi.\end{aligned}$$

Symmetry of the Kernel

Alternate proof of symmetry. Consider the evolution $(h_t)_t$ where $h_0 \sim \pi$.

Symmetry of the Kernel

Alternate proof of symmetry. Consider the evolution $(h_t)_t$ where $h_0 \sim \pi$.

$$\langle Kf, g \rangle_\pi = \mathbb{E} [f(h_{t+1})g(h_t)]$$

Symmetry of the Kernel

Alternate proof of symmetry. Consider the evolution $(h_t)_t$ where $h_0 \sim \pi$.

$$\begin{aligned}\langle Kf, g \rangle_\pi &= \mathbb{E} [f(h_{t+1})g(h_t)] \\ &= \mathbb{E}_{D_t} [\mathbb{E}[f(h_{t+1}) \mid D_t] \cdot \mathbb{E}[g(h_t) \mid D_t]]\end{aligned}$$

Symmetry of the Kernel

Alternate proof of symmetry. Consider the evolution $(h_t)_t$ where $h_0 \sim \pi$.

$$\begin{aligned}\langle Kf, g \rangle_\pi &= \mathbb{E} [f(h_{t+1})g(h_t)] \\ &= \mathbb{E}_{D_t} [\mathbb{E}[f(h_{t+1}) | D_t] \cdot \mathbb{E}[g(h_t) | D_t]] \\ &= \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t] \cdot \mathbb{E}[g(h_{t+1}) | D_t]]\end{aligned}$$

Symmetry of the Kernel

Alternate proof of symmetry. Consider the evolution $(h_t)_t$ where $h_0 \sim \pi$.

$$\begin{aligned}\langle Kf, g \rangle_\pi &= \mathbb{E} [f(h_{t+1})g(h_t)] \\ &= \mathbb{E}_{D_t} [\mathbb{E}[f(h_{t+1}) | D_t] \cdot \mathbb{E}[g(h_t) | D_t]] \\ &= \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t] \cdot \mathbb{E}[g(h_{t+1}) | D_t]] \\ &= \mathbb{E} [f(h_t)g(h_{t+1})] = \langle f, Kg \rangle_\pi.\end{aligned}$$

Variational Characterization of Eigenvalues

Spectral Theorem for Symmetric Operators. Let K be a symmetric operator with respect to $\langle \cdot, \cdot \rangle$. It has an orthonormal basis $(v_1, \dots, v_M)$ with eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M,$$

where they are computed by solving the optimization problem

$$\lambda_m = \max_{\substack{v \perp v_1, \dots, v_{m-1} \\ v \neq 0}} \frac{\langle v, Kv \rangle}{\langle v, v \rangle}$$

and v_m is the argmax of this Rayleigh quotient.

Bounding the Eigenvalues of the Kernel

Lemma. Consider the evolution $(h_t)_t$ where $h_0 \sim \pi$. For every $f : \mathcal{H} \rightarrow \mathbb{R}$,

$$\langle f, Kf \rangle_\pi = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t]^2]$$

$$\langle f, (I - K)f \rangle_\pi = \mathbb{E}_{D_t} [\text{Var}(f(h_t) | D_t)] .$$

Bounding the Eigenvalues of the Kernel

Lemma. Consider the evolution $(h_t)_t$ where $h_0 \sim \pi$. For every $f : \mathcal{H} \rightarrow \mathbb{R}$,

$$\langle f, Kf \rangle_\pi = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t]^2]$$

$$\langle f, (I - K)f \rangle_\pi = \mathbb{E}_{D_t} [\text{Var}(f(h_t) | D_t)] .$$

Proof.

$$\langle f, Kf \rangle_\pi = \mathbb{E}[f(h_t)f(h_{t+1})]$$

Bounding the Eigenvalues of the Kernel

Lemma. Consider the evolution $(h_t)_t$ where $h_0 \sim \pi$. For every $f : \mathcal{H} \rightarrow \mathbb{R}$,

$$\langle f, Kf \rangle_\pi = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t]^2]$$

$$\langle f, (I - K)f \rangle_\pi = \mathbb{E}_{D_t} [\text{Var}(f(h_t) | D_t)] .$$

Proof.

$$\langle f, Kf \rangle_\pi = \mathbb{E}[f(h_t)f(h_{t+1})] = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t] \cdot \mathbb{E}[f(h_{t+1}) | D_t]]$$

Bounding the Eigenvalues of the Kernel

Lemma. Consider the evolution $(h_t)_t$ where $h_0 \sim \pi$. For every $f : \mathcal{H} \rightarrow \mathbb{R}$,

$$\langle f, Kf \rangle_\pi = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t]^2]$$

$$\langle f, (I - K)f \rangle_\pi = \mathbb{E}_{D_t} [\text{Var}(f(h_t) | D_t)] .$$

Proof.

$$\langle f, Kf \rangle_\pi = \mathbb{E}[f(h_t)f(h_{t+1})] = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t] \cdot \mathbb{E}[f(h_{t+1}) | D_t]] = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t]^2] .$$

Bounding the Eigenvalues of the Kernel

Lemma. Consider the evolution $(h_t)_t$ where $h_0 \sim \pi$. For every $f : \mathcal{H} \rightarrow \mathbb{R}$,

$$\langle f, Kf \rangle_\pi = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t]^2]$$

$$\langle f, (I - K)f \rangle_\pi = \mathbb{E}_{D_t} [\text{Var}(f(h_t) | D_t)] .$$

Proof.

$$\langle f, Kf \rangle_\pi = \mathbb{E}[f(h_t)f(h_{t+1})] = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t] \cdot \mathbb{E}[f(h_{t+1}) | D_t]] = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t]^2] .$$

- The second equation follows by definition of variance

Bounding the Eigenvalues of the Kernel

Lemma. Consider the evolution $(h_t)_t$ where $h_0 \sim \pi$. For every $f : \mathcal{H} \rightarrow \mathbb{R}$,

$$\langle f, Kf \rangle_\pi = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t]^2]$$

$$\langle f, (I - K)f \rangle_\pi = \mathbb{E}_{D_t} [\text{Var}(f(h_t) | D_t)] .$$

Proof.

$$\langle f, Kf \rangle_\pi = \mathbb{E}[f(h_t)f(h_{t+1})] = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t] \cdot \mathbb{E}[f(h_{t+1}) | D_t]] = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t]^2] .$$

- The second equation follows by definition of variance, since

$$\langle f, f \rangle_\pi = \mathbb{E}[f(h_t)^2]$$

Bounding the Eigenvalues of the Kernel

Lemma. Consider the evolution $(h_t)_t$ where $h_0 \sim \pi$. For every $f : \mathcal{H} \rightarrow \mathbb{R}$,

$$\langle f, Kf \rangle_\pi = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t]^2]$$

$$\langle f, (I - K)f \rangle_\pi = \mathbb{E}_{D_t} [\text{Var}(f(h_t) | D_t)] .$$

Proof.

$$\langle f, Kf \rangle_\pi = \mathbb{E}[f(h_t)f(h_{t+1})] = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t] \cdot \mathbb{E}[f(h_{t+1}) | D_t]] = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t) | D_t]^2] .$$

- The second equation follows by definition of variance, since

$$\langle f, f \rangle_\pi = \mathbb{E}[f(h_t)^2] = \mathbb{E}_{D_t} [\mathbb{E}[f(h_t)^2 | D_t]] .$$

Bounding the Eigenvalues of the Kernel

Proposition. The operator K is PSD and $0 \preceq K \preceq I$.

- In particular, all eigenvalues are contained in $[0,1]$.
- The top eigenvalue is 1 corresponding to eigenvector π .

Bounding the Eigenvalues of the Kernel

Proposition. The operator K is PSD and $0 \preceq K \preceq I$.

- In particular, all eigenvalues are contained in $[0,1]$.
- The top eigenvalue is 1 corresponding to eigenvector π .

Proof.

- PSD follows from $\langle f, Kf \rangle_\pi = \mathbb{E}[\mathbb{E}[f(h) | D]^2] \geq 0$.

Bounding the Eigenvalues of the Kernel

Proposition. The operator K is PSD and $0 \preceq K \preceq I$.

- In particular, all eigenvalues are contained in $[0,1]$.
- The top eigenvalue is 1 corresponding to eigenvector π .

Proof.

- PSD follows from $\langle f, Kf \rangle_\pi = \mathbb{E}[\mathbb{E}[f(h) | D]^2] \geq 0$.
- The second follows from $\langle f, (I - K)f \rangle_\pi = \mathbb{E}[\text{Var}(f(h) | D)] \geq 0$.

Characterizing the Spectral Gap

Corollary. Let $(h, D) \sim \nu$. The spectral gap of K is given by

$$\gamma = \min_{f: \text{Var}_{\pi}(f) > 0} \frac{\mathbb{E}[\text{Var}(f(h) \mid D)]}{\text{Var}_{\pi}(f)}.$$

Proof.

- Since the largest eigenvalue of K is 1, the spectral gap $1 - \lambda_2$ corresponds to the smallest eigenvalue of $I - K$, given by the above Rayleigh quotient.

Iterated Learning (IL) Dynamics

Theorem. Let $\mathcal{H}$ be a finite set of languages and $\pi(h) > 0$ for all $h \in \mathcal{H}$. Let h_0 be the initial language. Then, π is the stationary distribution of (IL), and

$$\|\mu_t - \pi\|_{\text{TV}} \leq C \exp(-\gamma t),$$

where $(h, D) \sim \nu$ and γ is given by

$$\gamma = \min_{f: \text{Var}_{\pi}(f) > 0} \frac{\mathbb{E}_D[\text{Var}(f(h) | D)]}{\text{Var}_{\pi}(f)} \quad \text{where } f: \mathcal{H} \rightarrow \mathbb{R}.$$

Interpreting the Spectral Gap

The spectral gap of K is given by

$$\gamma = \min_{\text{Var}_{\pi}(f)=1} \mathbb{E} [\text{Var}(f(h) \mid D)].$$

Interpreting the Spectral Gap

The spectral gap of K is given by

$$\gamma = \min_{\text{Var}_{\pi}(f)=1} \mathbb{E} [\text{Var}(f(h) \mid D)].$$

Interpretation.

- We can think of $\text{Var}(f(h) \mid D)$ as a “reconstruction error” of learning the language h from sampled data D ; $1 - \gamma$ is a correlation between h and D .

Interpreting the Spectral Gap

The spectral gap of K is given by

$$\gamma = \min_{\text{Var}_{\pi}(f)=1} \mathbb{E} [\text{Var}(f(h) \mid D)].$$

Interpretation.

- We can think of $\text{Var}(f(h) \mid D)$ as a “reconstruction error” of learning the language h from sampled data D ; $1 - \gamma$ is a correlation between h and D .
- A small gap $\gamma \approx 0$ means that the language can be transmitted well.

Interpreting the Spectral Gap

The spectral gap of K is given by

$$\gamma = \min_{\text{Var}_{\pi}(f)=1} \mathbb{E} [\text{Var}(f(h) \mid D)].$$

Interpretation.

- We can think of $\text{Var}(f(h) \mid D)$ as a “reconstruction error” of learning the language h from sampled data D ; $1 - \gamma$ is a correlation between h and D .
- A small gap $\gamma \approx 0$ means that the language can be transmitted well.
- A large gap $\gamma \approx 1$ means that the Markov chain mixes rapidly.

Conclusion of the Paper

- In the long-run, the original language will be lost, and what remains is a reflection of the prior of the family of learner.
 - This conclusion is also related to more recent work on ‘mode collapse’.
 - Prior work tried to empirically demonstrate that compositionality arose out of the transmission bottleneck. This work suggests that the computational model matters a lot.

Follow up questions

- It may be more interesting to take a more fine-grained view of the dynamics: when are specific features of the language lost?
- What if there is covariate shift, as what we want to say changes over time?
- How do we incorporate a utility aspect to the language? Currently, the iterated learning dynamics are analyzed in a vacuum.
- Bayesian model of learners is also an interesting formalization of “how much evidence” is required for a learner to choose a particular hypothesis.
- Model of Bayesian learning also interesting from the perspective of transmission of information across agents with bounded memory.
- How do I communicate to a learner if I do not know its prior?

References

Griffiths and Kalish (2007) “Language Evolution by Iterated Learning with Bayesian Agents”.

Kirby (2017) “Culture and biology in the origins of linguistic structure”.

Roberts and Rosenthal (2004) “General state space Markov chains and MCMC algorithms”.