On Mixing Rates for Bayesian CART

Jungeum Kim and Veronika Rockova

December 3 , 2025

ISBA NC STATE UNIVERSITY CHICAGO BOOTH
The University of Chicago Booth School of Business

Outline

Background : Tree and Mixing Rate

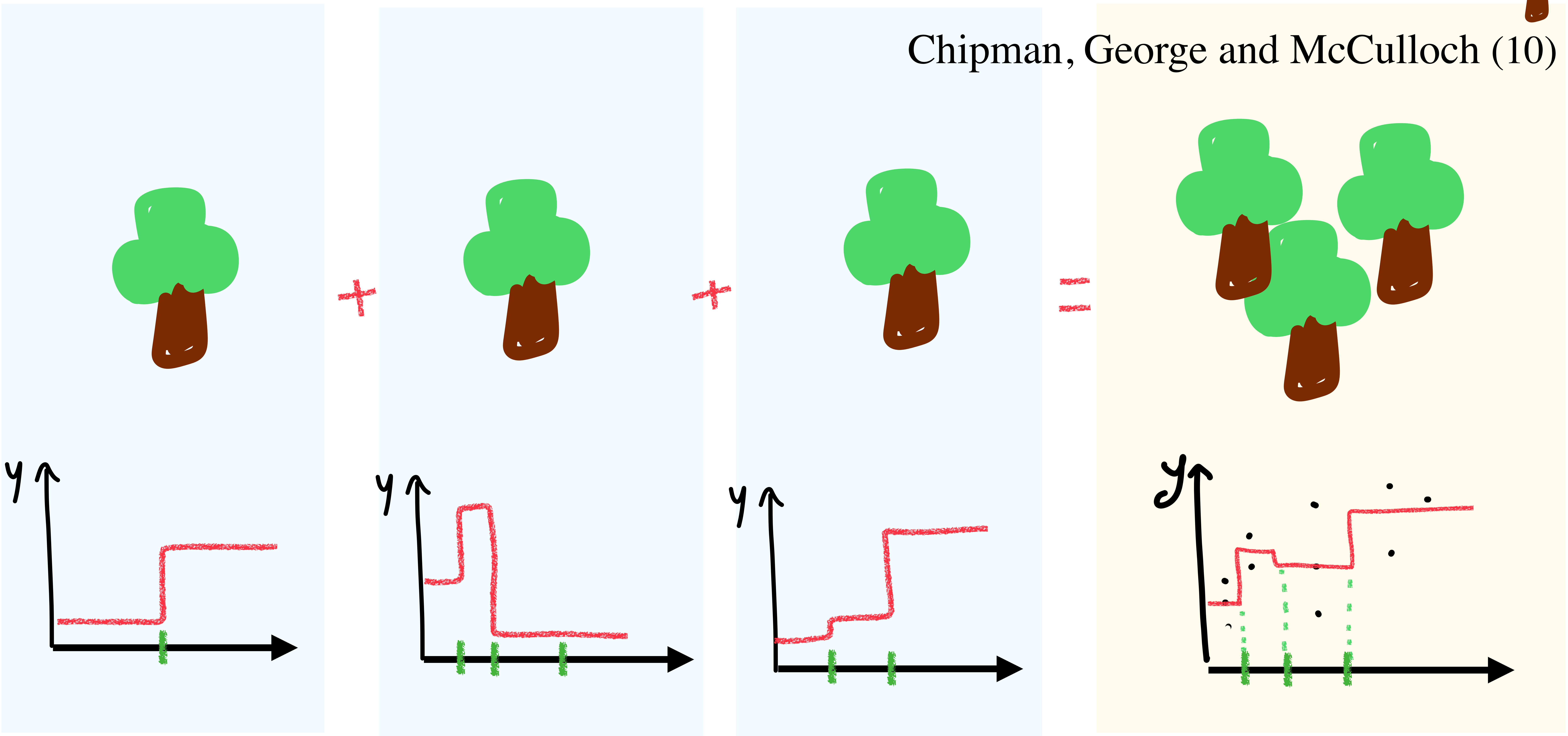
Setting ups and Mixing Time Framework

Theoretical and Numerical Results

Concluding Remarks

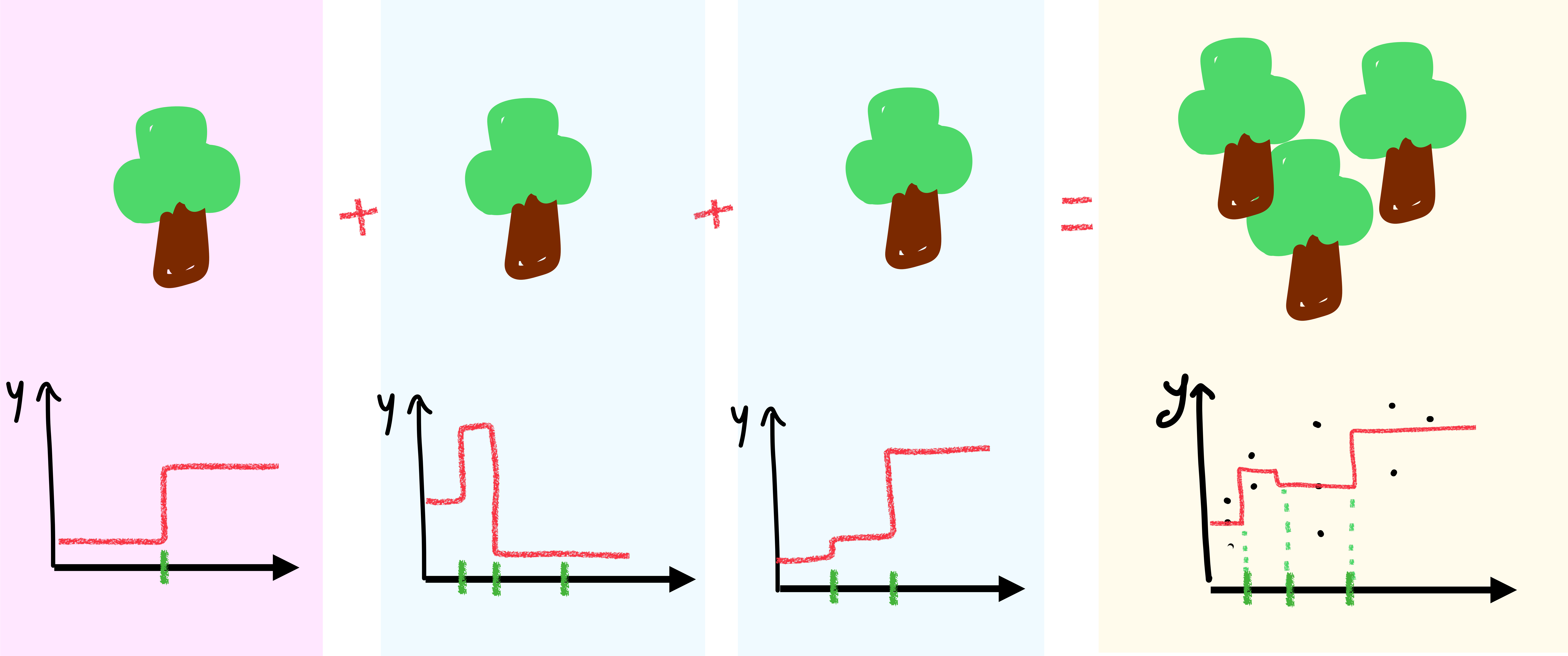
BART: Bayesian Additive Regression Tree

Chipman, George and McCulloch (10)



BART: Bayesian Additive Regression Tree

Chipman, George and McCulloch (10)

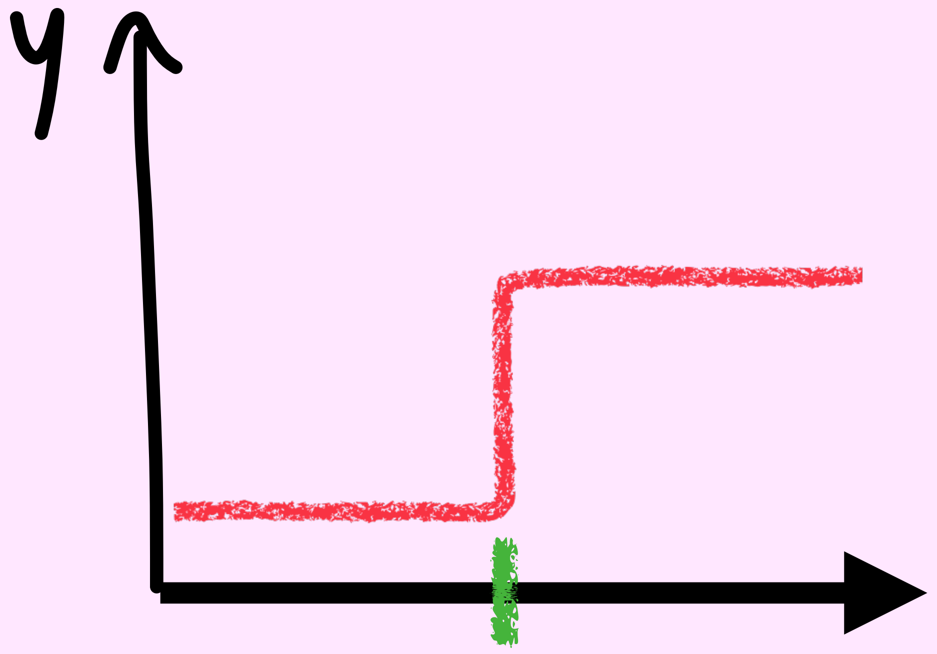


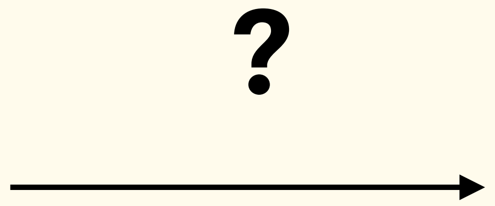
Bayesian Classification and Regression Tree

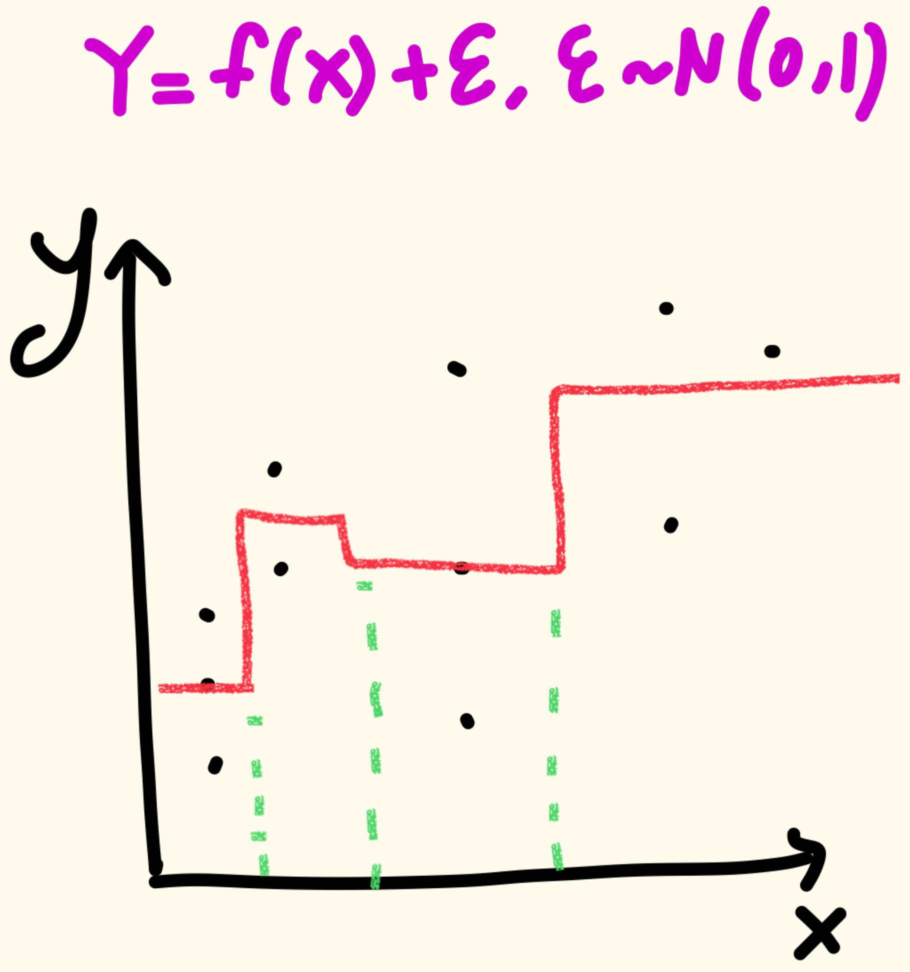
Chipman et al (98), Denison (98)

Bayesian Classification and Regression Tree

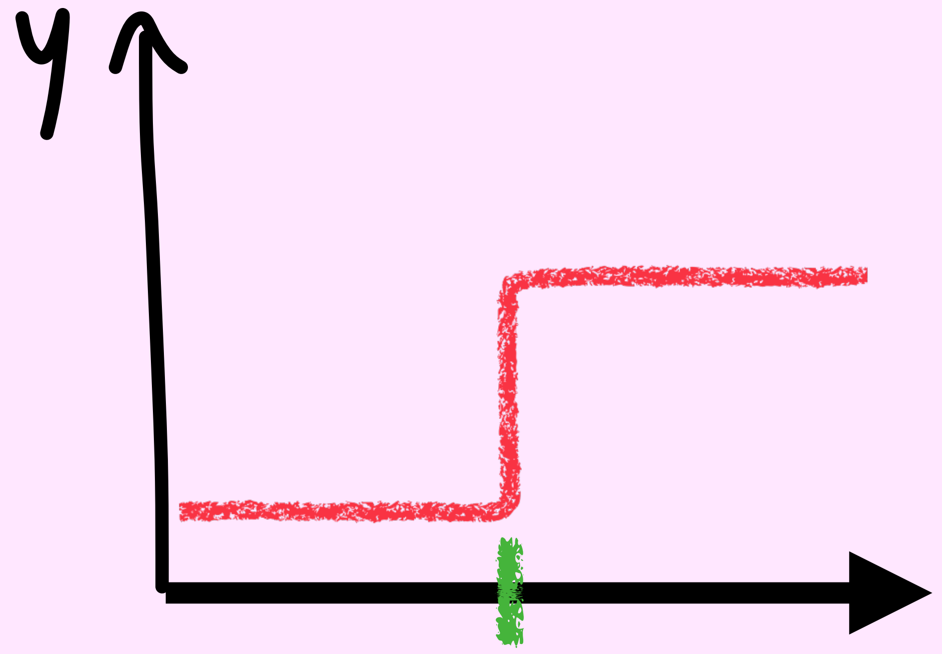
Chipman et al (98), Denison (98)





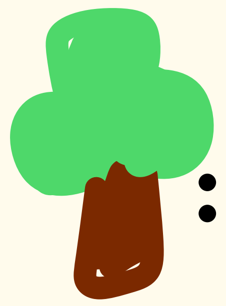


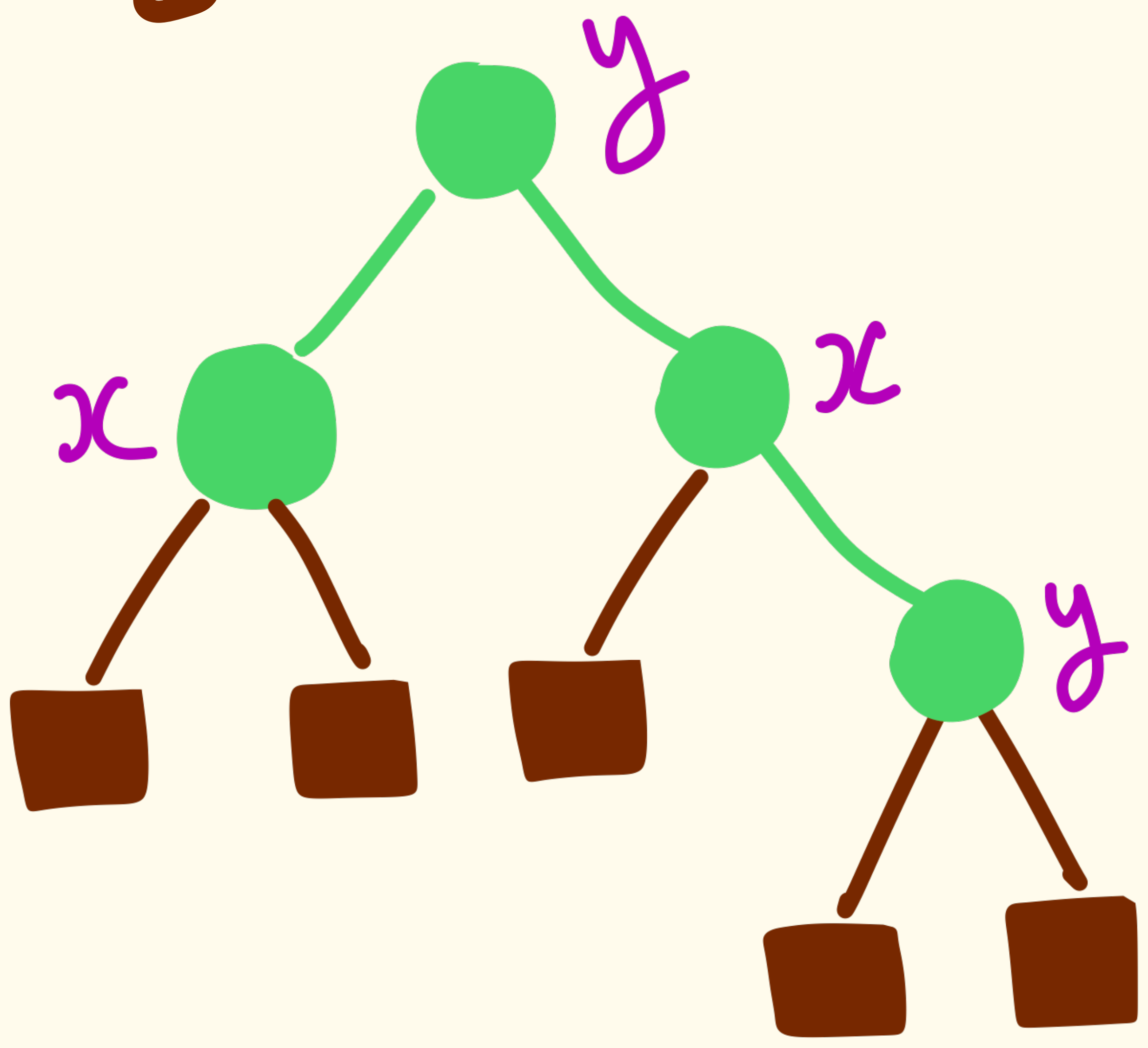
Bayesian Classification and Regression Tree



Chipman et al (98), Denison (98)

Tree ⟶ Partition

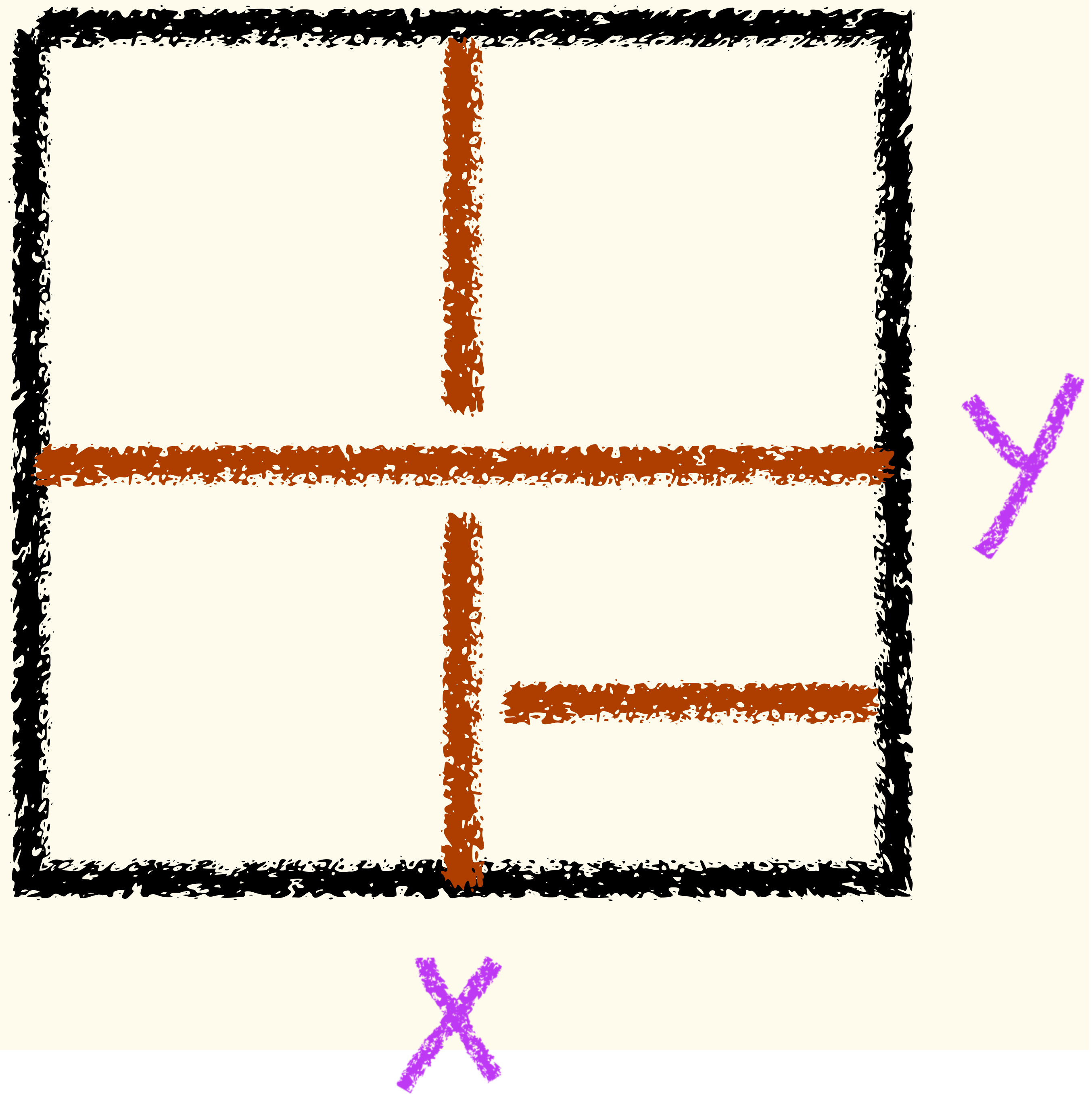

: Binary tree



Unique

Not unique

(Binary) Partition



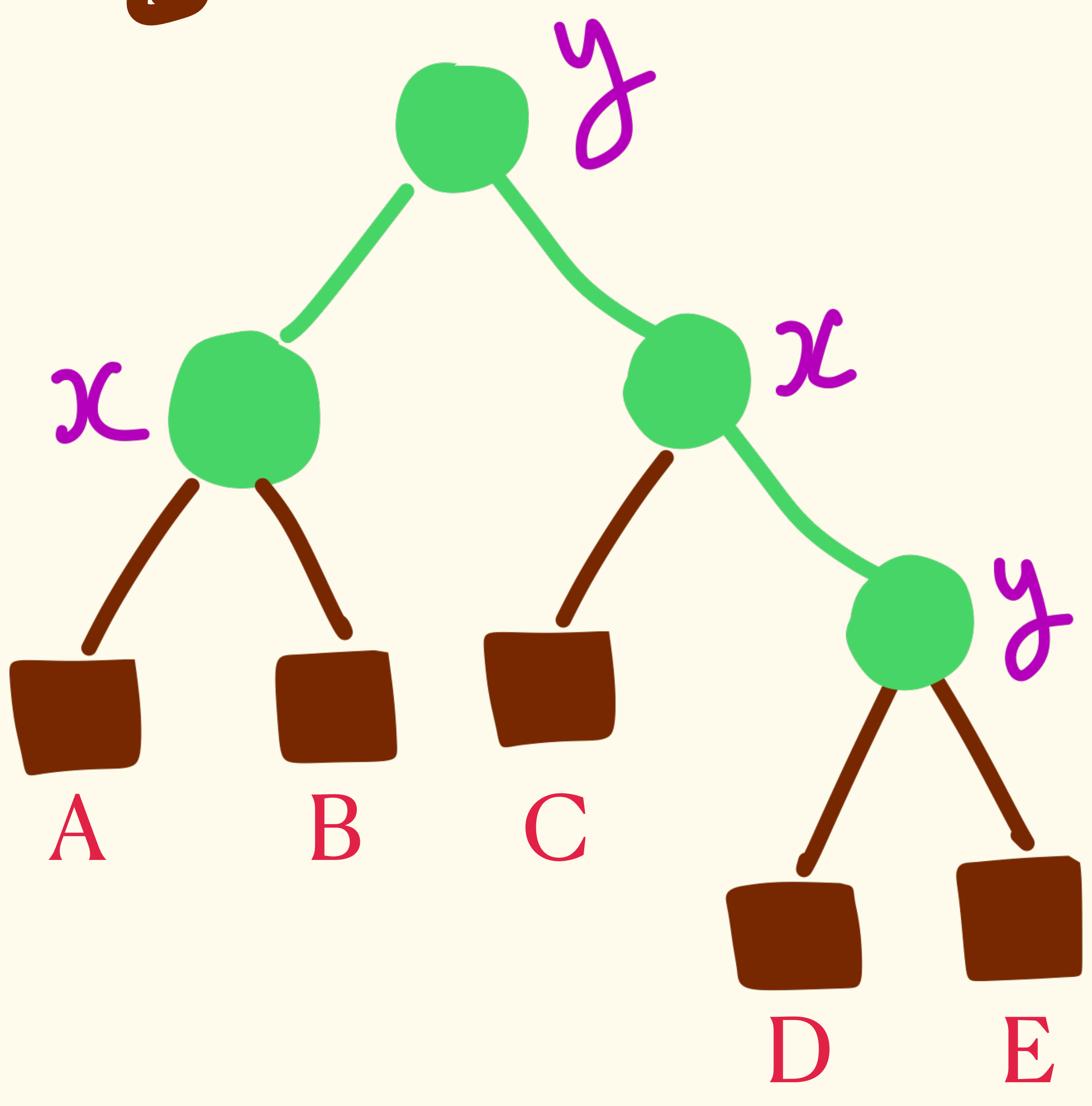
Given the box splits always at the center

Bayesian Classification and Regression Tree

Chipman et al (98), Denison (98)

Tree ⟶ Partition

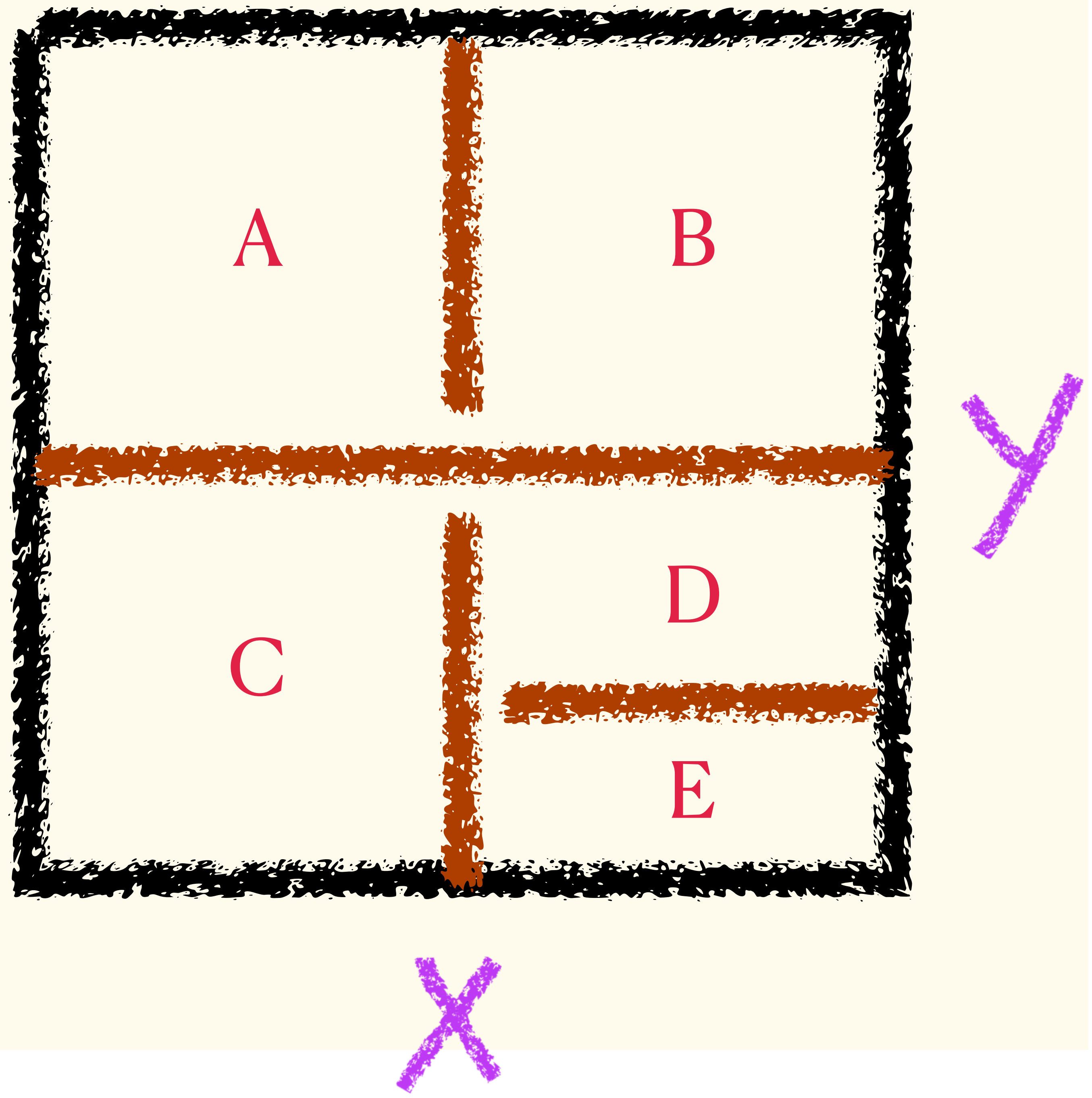
: **Binary tree**



Unique

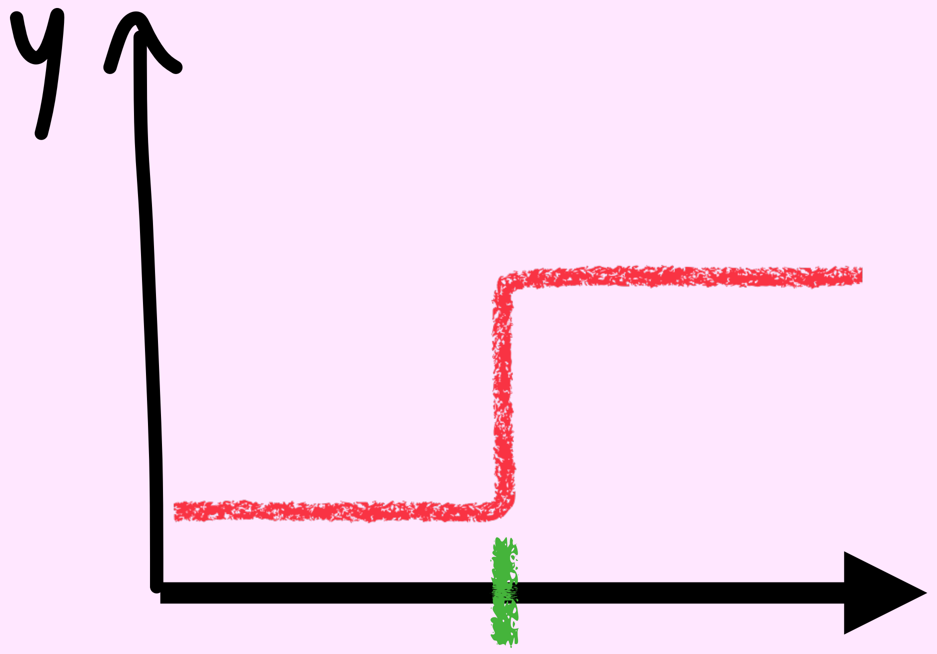
Not unique

(Binary) Partition



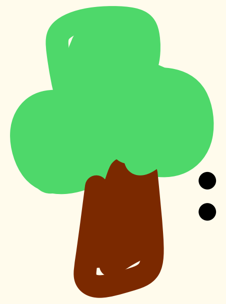
Given the box splits always at the center

Bayesian Classification and Regression Tree

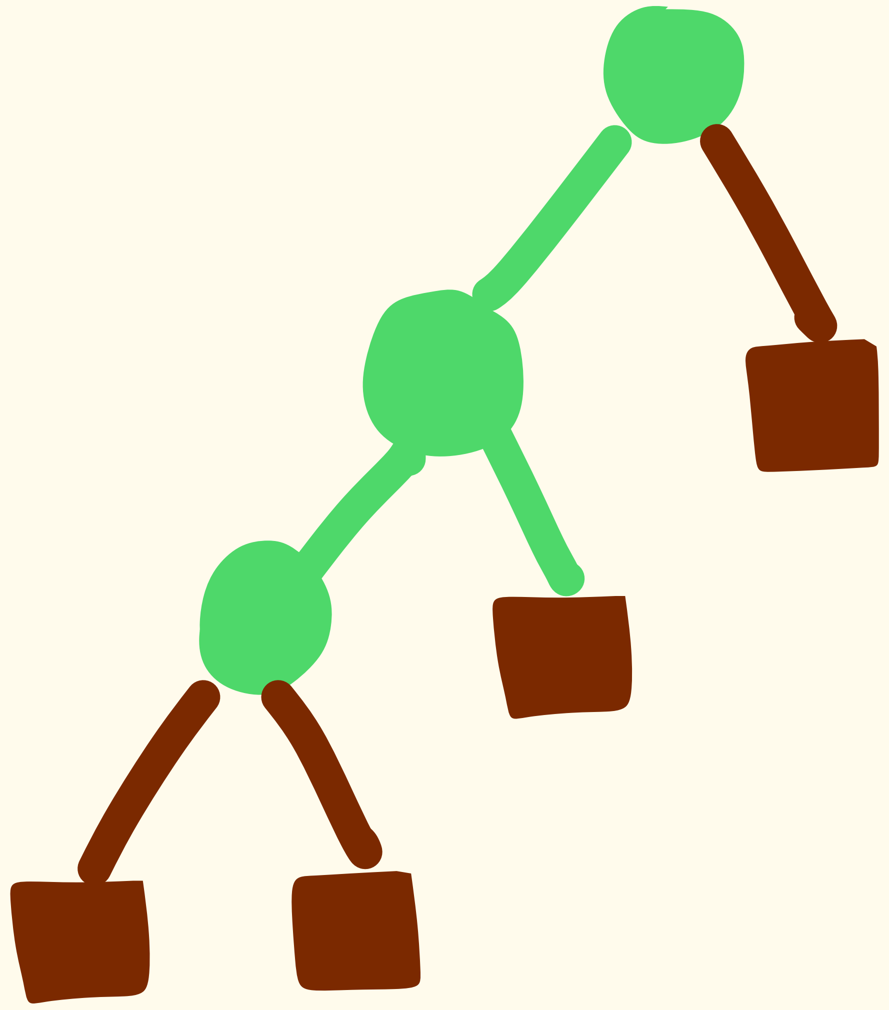


Chipman et al (98), Denison (98)

Tree → Partition

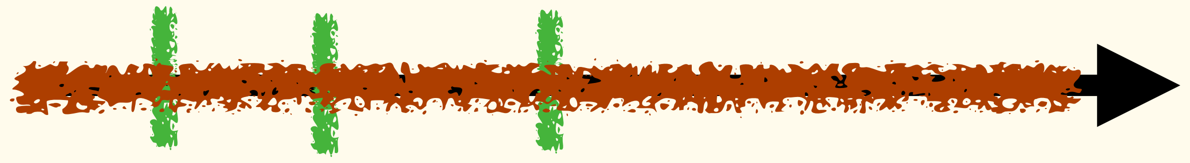
 **: Binary tree**

(Binary) Partition



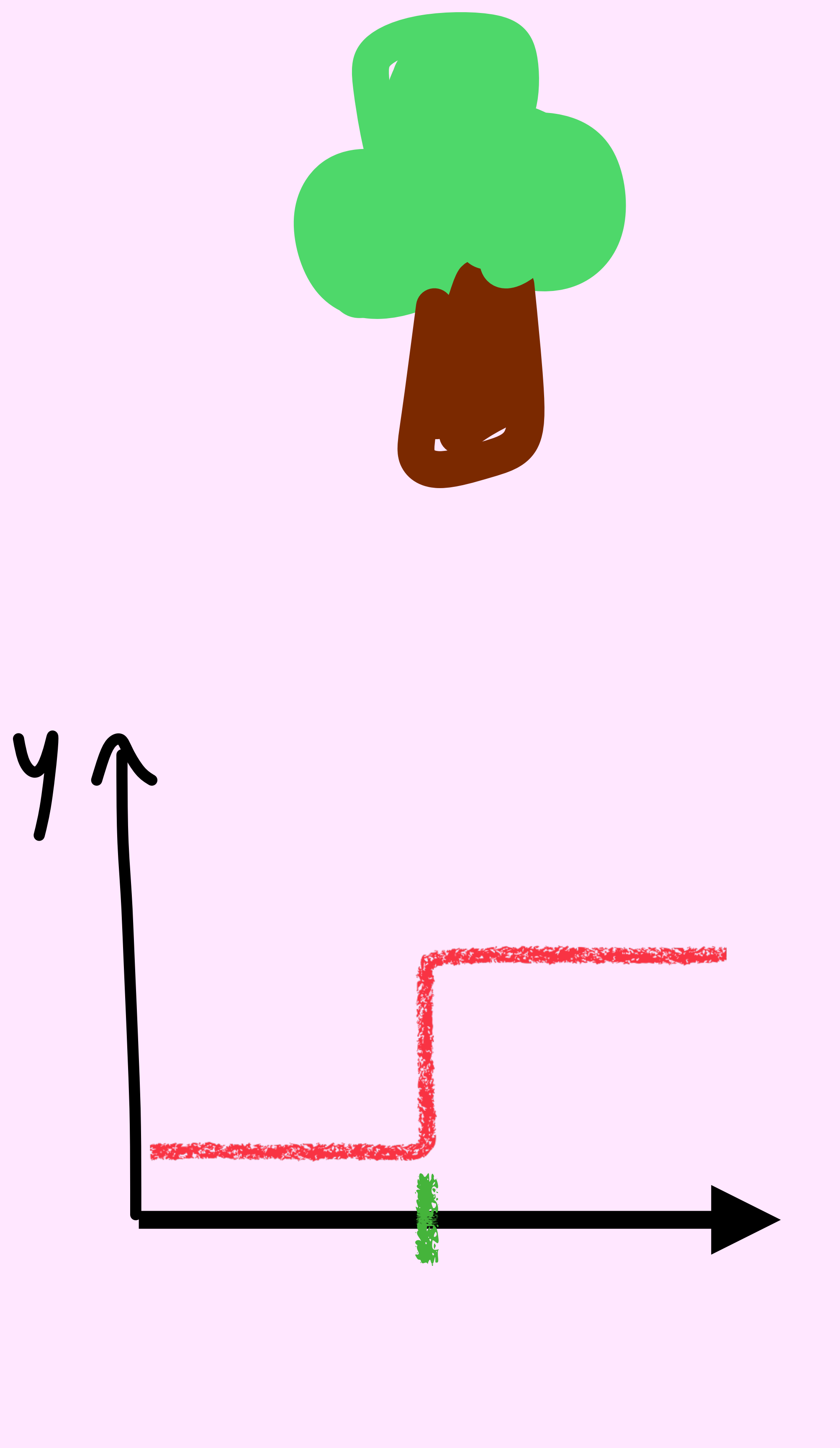
Unique

Unique



Given the box splits always at the center

Bayesian Classification and Regression Tree

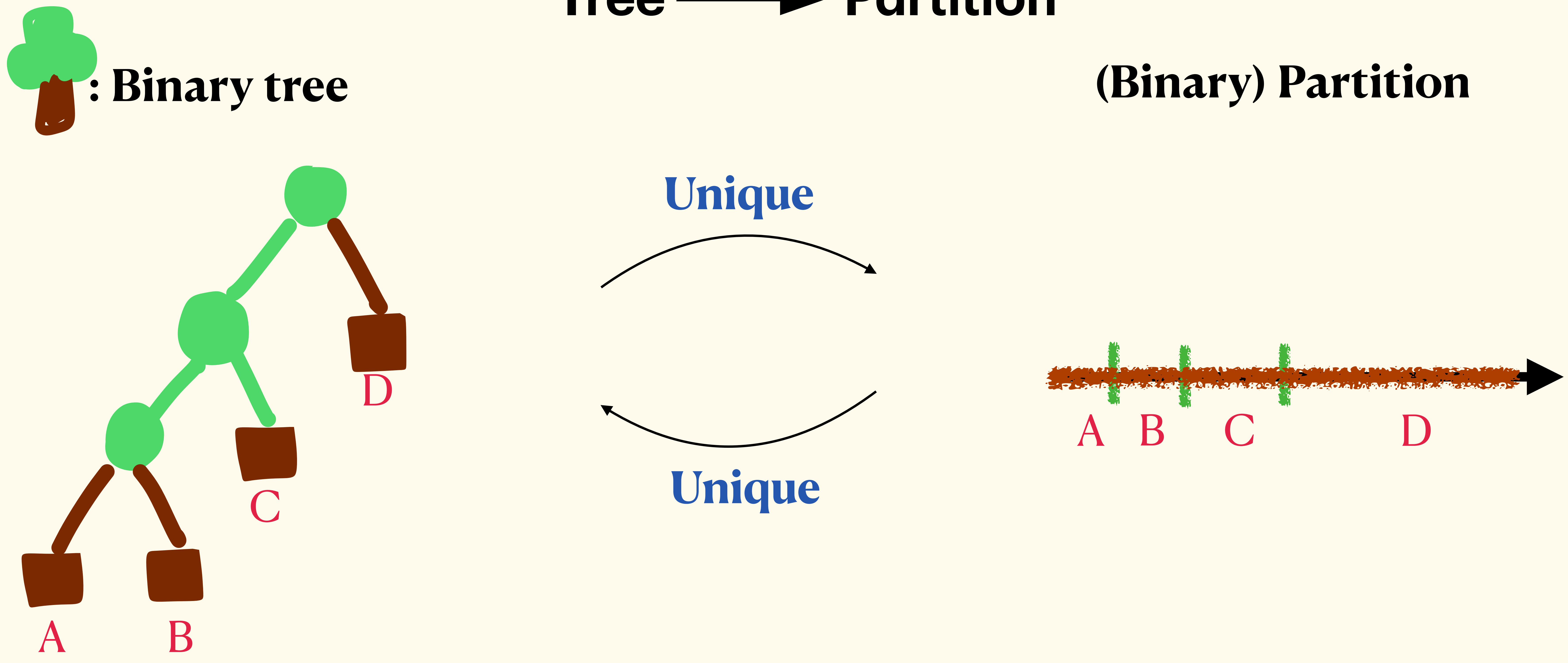


Chipman et al (98), Denison (98)

Tree ⟶ Partition

: **Binary tree**

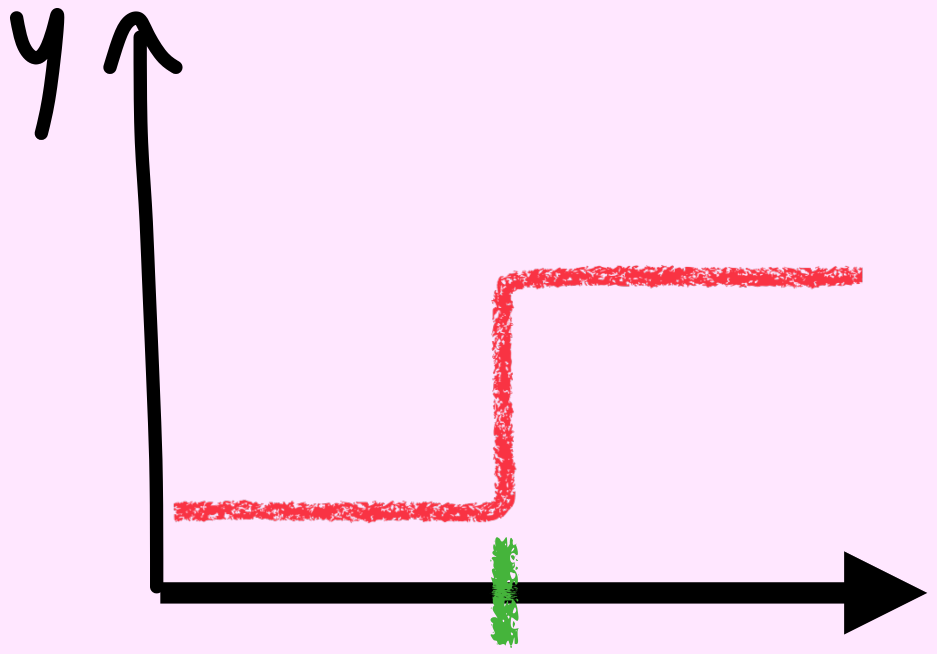
(Binary) Partition



Given the box splits always at the center

Bayesian Classification and Regression Tree

$\mathscr{T} \sim \pi(\mathscr{T})$



Chipman et al (98), Denison (98)

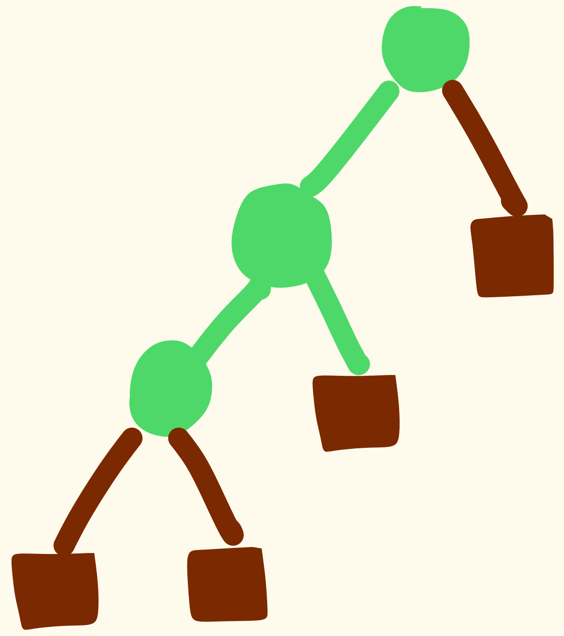
Step function:

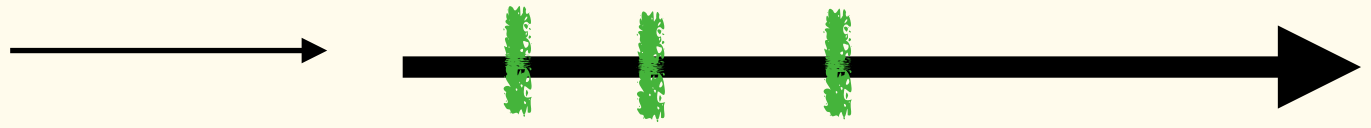
1) Partition: $\mathscr{T}$

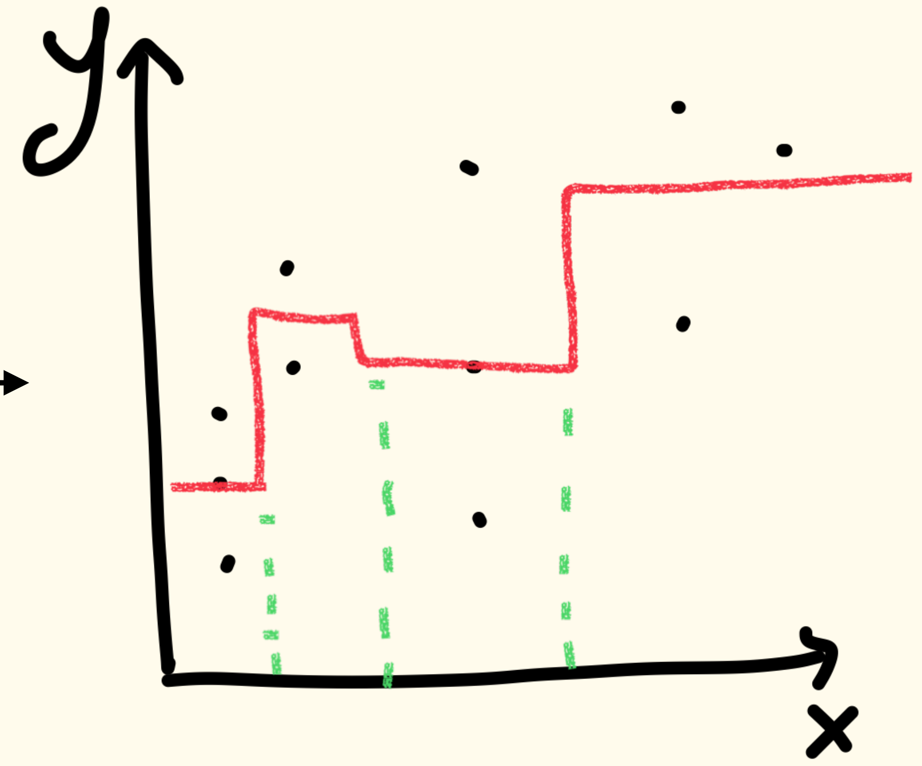
2) Jumps: regression function

(Very non-parametric modeling)

$Y = f(x) + \varepsilon,\ \varepsilon \sim N(0,1)$







$\mathscr{T} \sim \pi(\mathscr{T})$ $\qquad$ $L(Y|\mathscr{T})$

$$\pi(\mathscr{T} \mid Y) \propto \pi(\mathscr{T}) L(Y|\mathscr{T})$$

Rockova and Castillo (2021), Rockova and Rousseau (2023)

How to sample from $\pi(\mathscr{T} \mid Y)$? MCMC

The MH algorithm $\Rightarrow \mathscr{T}_t \sim P(\cdot \mid \mathscr{T}_{t-1})$

Given $\mathscr{T}_{t-1}$,

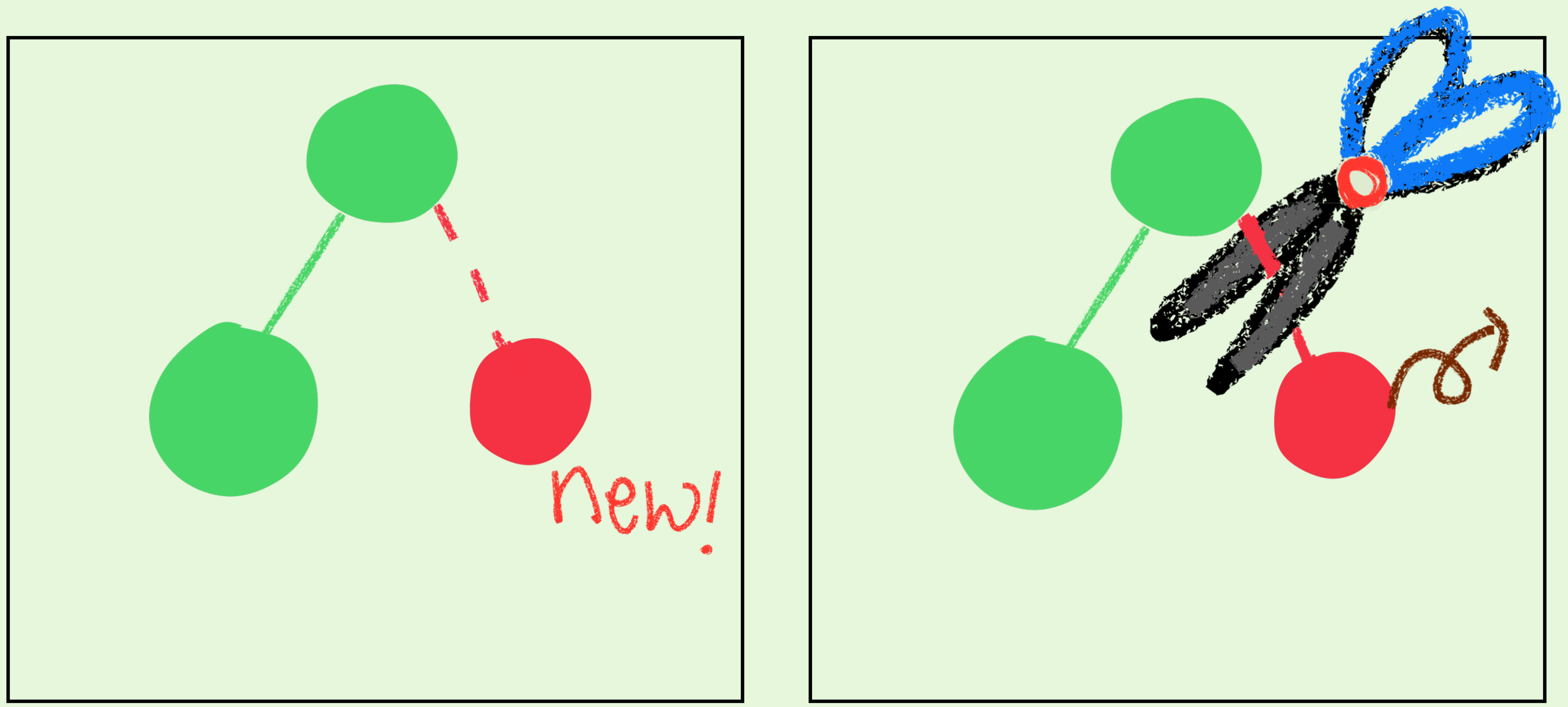
1. Generate $\tilde{\mathscr{T}} \sim S(\cdot \mid \mathscr{T}_{t-1})$
2. Accept $(\mathscr{T}_t = \tilde{\mathscr{T}})$ with probability α, where

$$\alpha = \min\left\{1, \frac{\pi(\tilde{\mathscr{T}} \mid Y)}{\pi(\mathscr{T}_{t-1} \mid Y)} \frac{S(\mathscr{T}_{t-1} \mid \tilde{\mathscr{T}})}{S(\tilde{\mathscr{T}} \mid \mathscr{T}_{t-1})}\right\}$$

Proposal of Bayesian CART

GROW

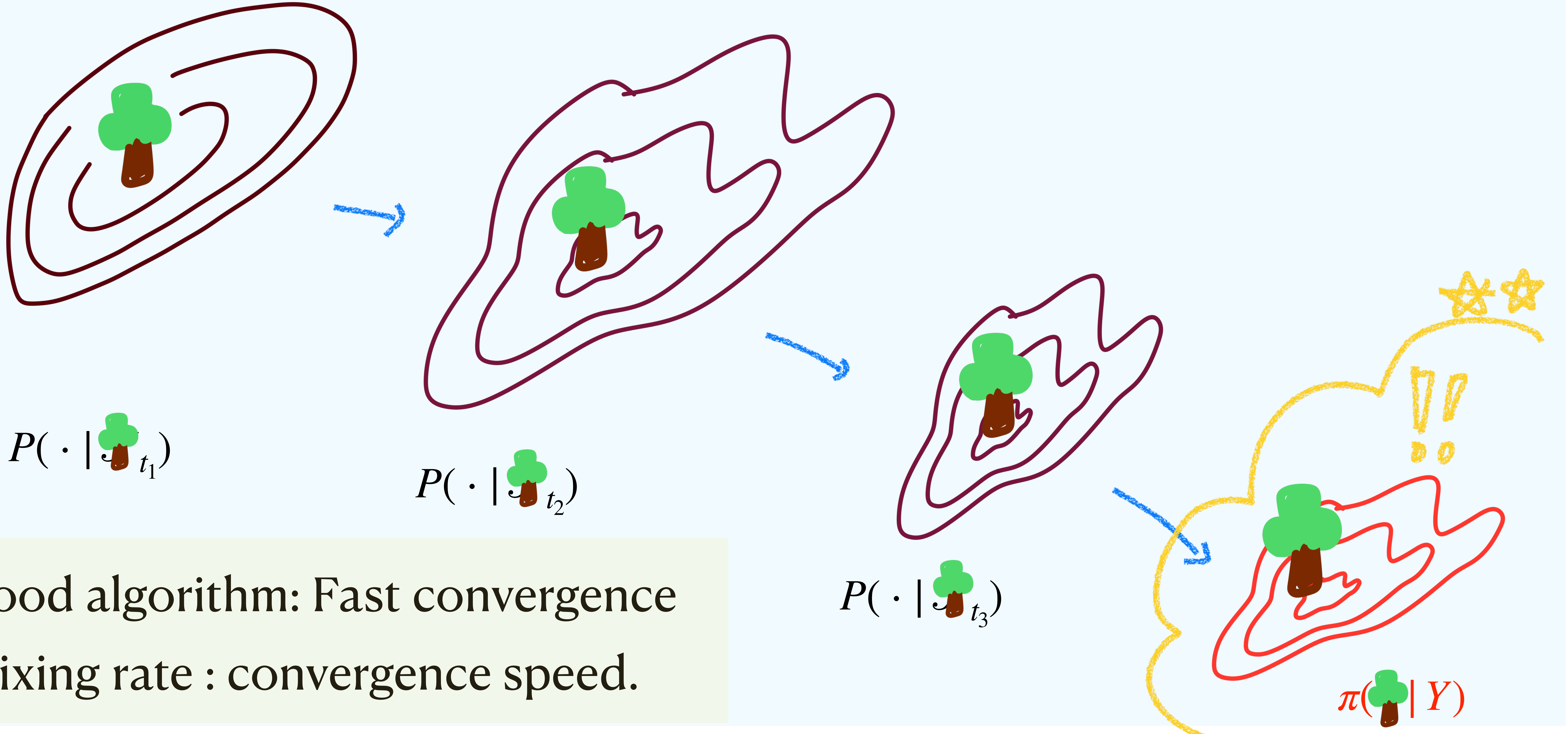
PRUNE



$$S(\cdot \mid \mathscr{T}_t) = \frac{1}{2} \sum_{m \in M} 1_m S_m(\cdot \mid \mathscr{T}_t), \text{ where } M = \{\text{grow, prune}\}$$

Mixing Rate: How fast?

$$P^t(\cdot \mid \mathcal{T}_0) \xrightarrow{t \to \infty} \pi(\cdot \mid Y)$$



Good algorithm: Fast convergence

Mixing rate : convergence speed.

Literature: Mixing rate can be slow

Bayesian CART mixes slow: Wu et al (2007), Chipman et al (2010).

New movements: Gramacy and Lee (2008), Lakshminarayanan et al. (2015), Pratola (2016)

New initializations: He and Hahn (2021)

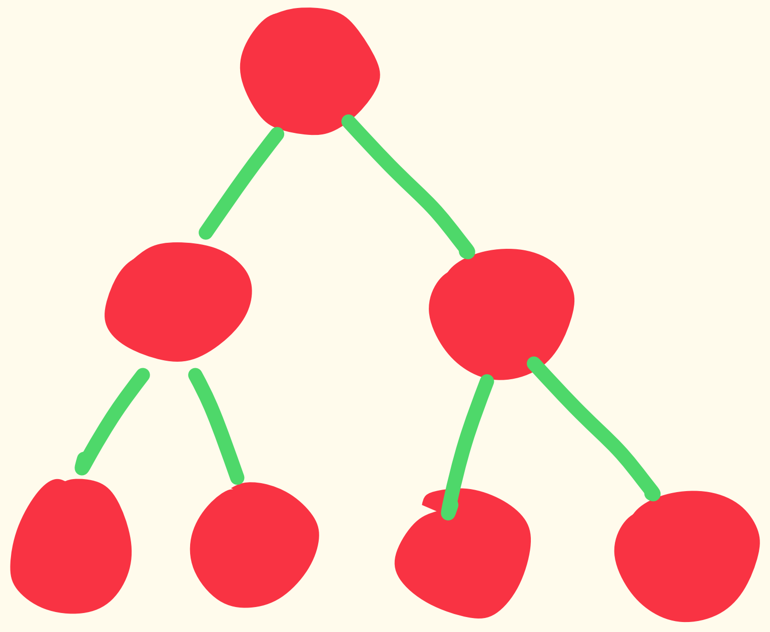
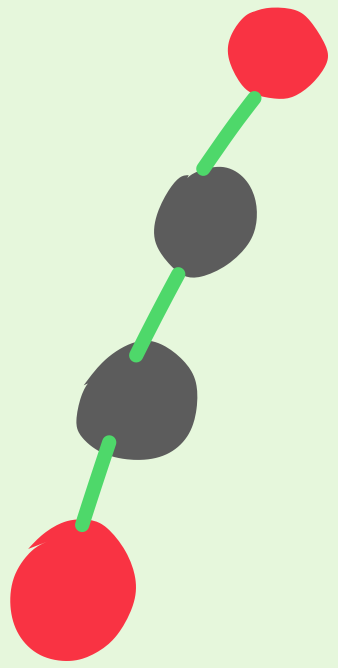
(Theoretical) Exponentially slow mixing : Ronen et al (2022)

In multi-dimensional setting

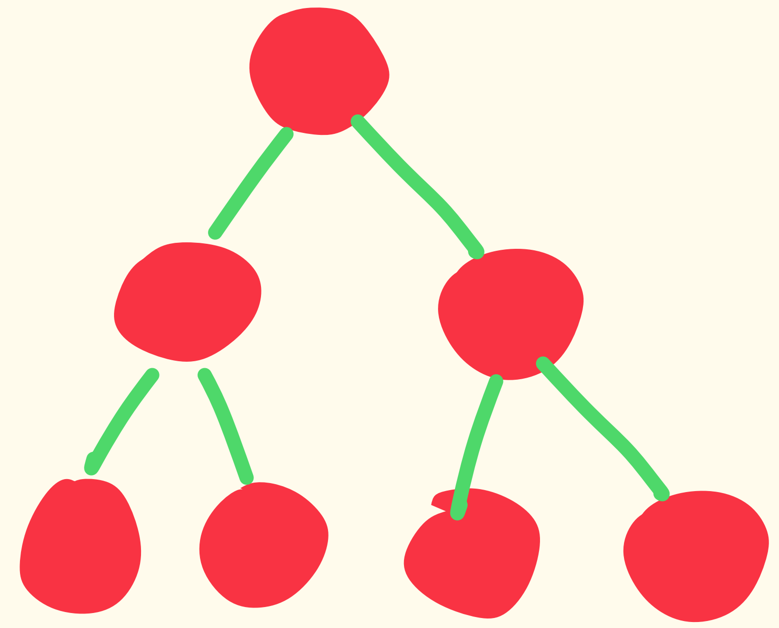
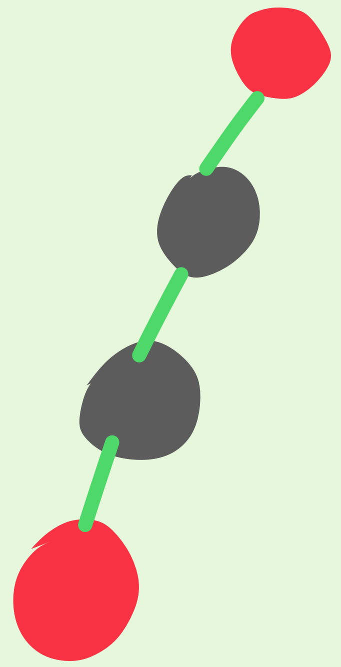
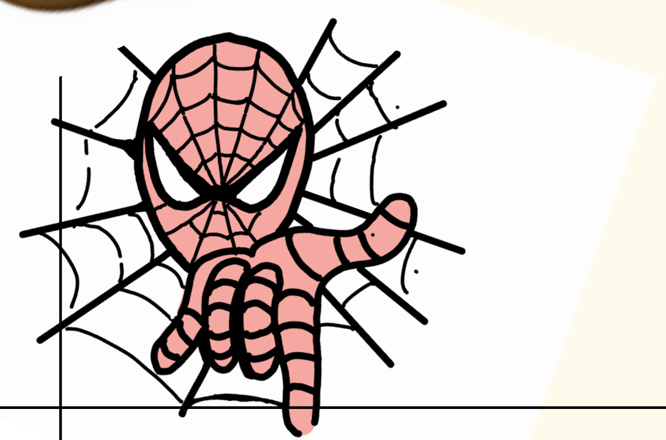
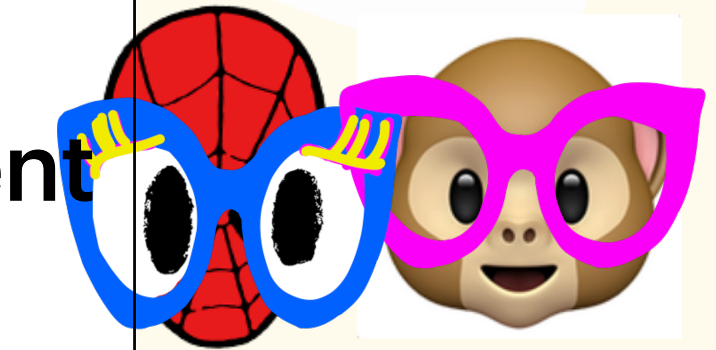
GROW and PRUNE

Our investigation : The mixing time highly depends on the data structure !

Our Contribution

	Type A 	Type A^c 
Original Movement	**Rapid**	**Slow**
New Movement	**Rapid**	**Rapid**
Informed Movement	**Improvement**	**Improvement?**

Our Contribution

	Type A 	Type A^c 
Original Moven 	Rapid	Slow
New Movement 	Rapid	Rapid
Informed Movement 	Improvement	Improvement?

Outline

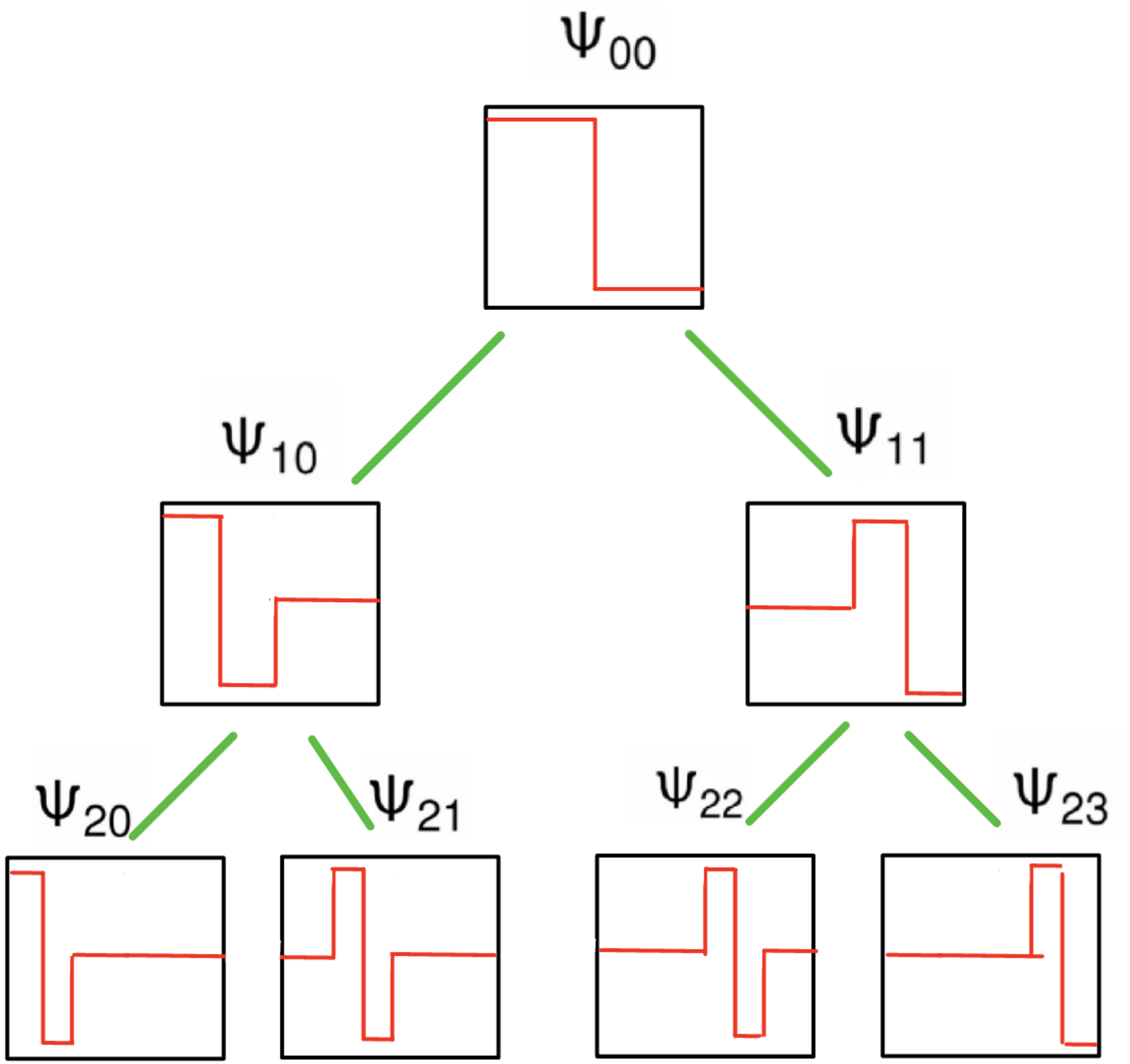
Background : Tree and Mixing Rate

Setting ups and Mixing Time Framework

Theoretical and Numerical Results

Concluding Remarks

Tree shaped wavelet reconstruction



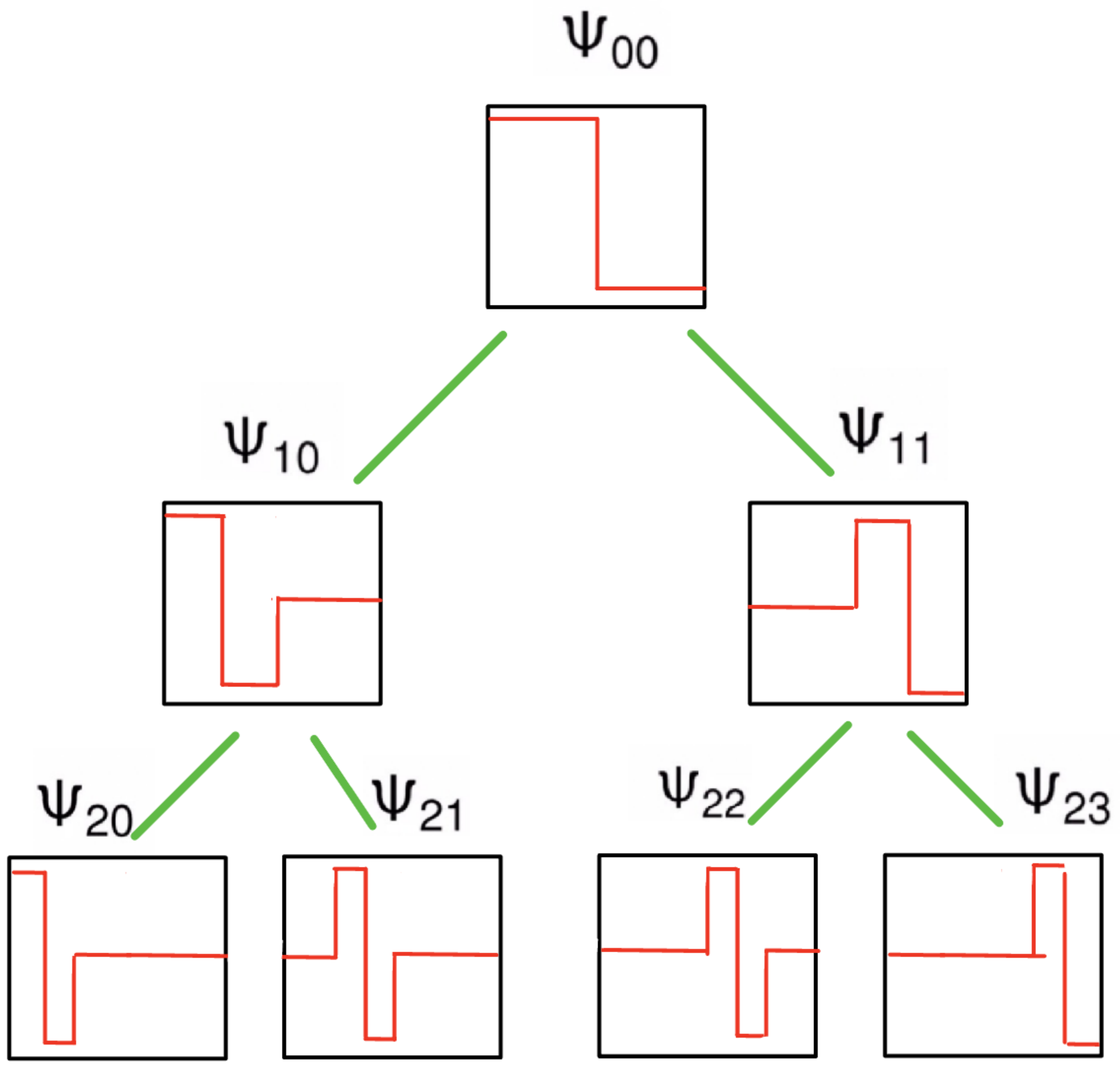
Haar wavelet basis

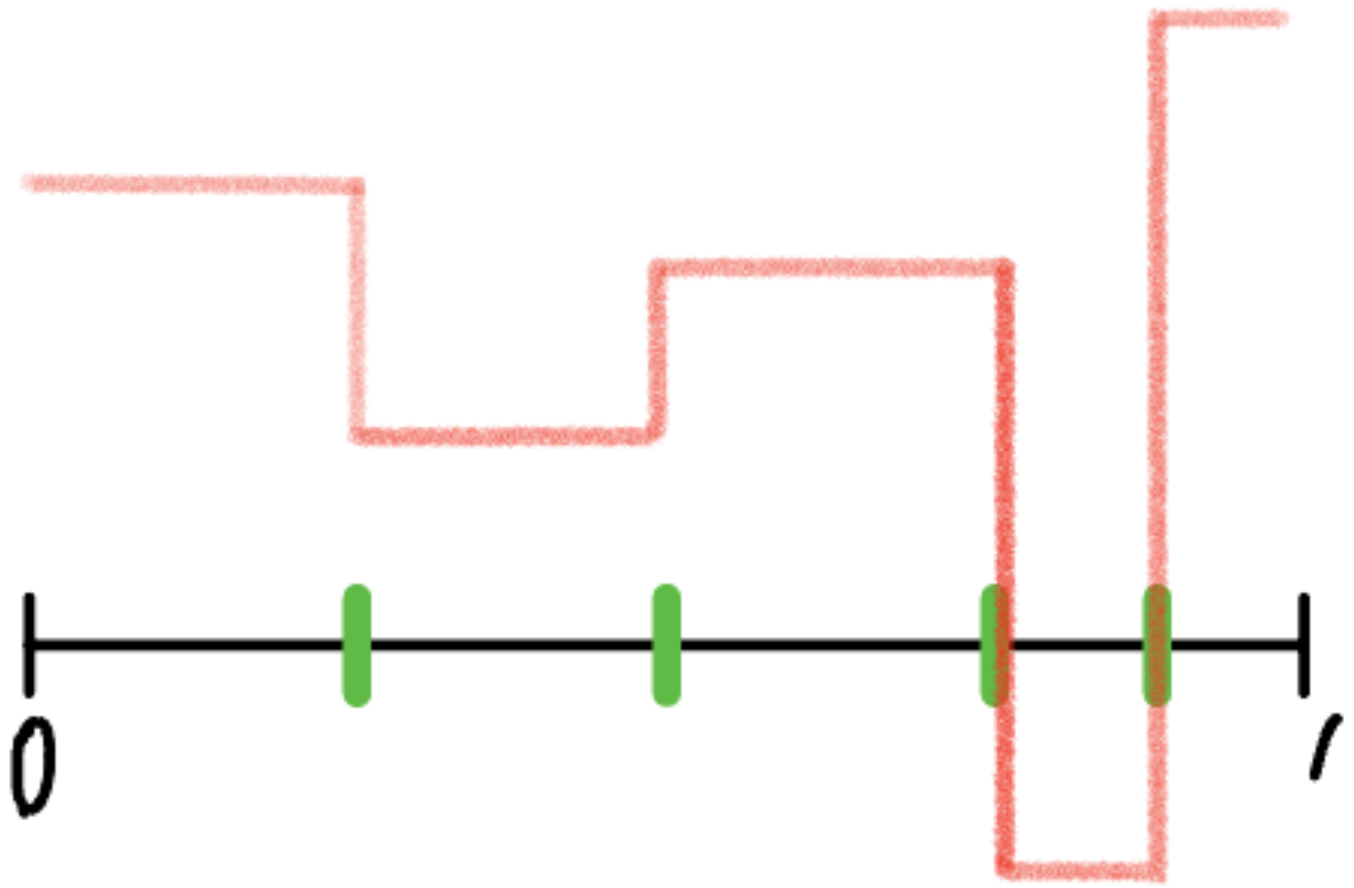
$$\psi_{-10}(x) = I_{[0,1]}(x)$$

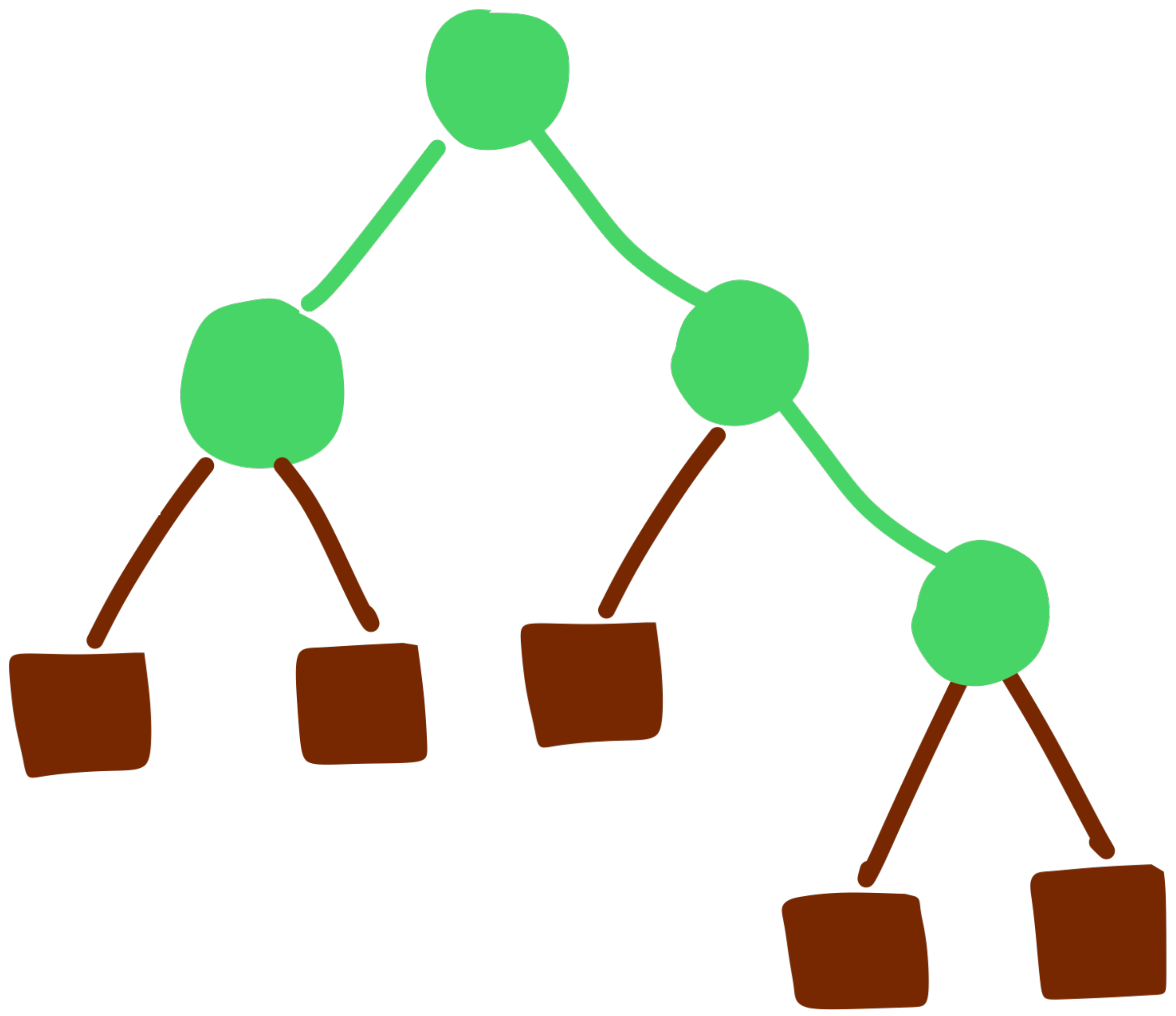
$$\psi_{lk}(x) = 2^{l/2}\psi(2^l x - k),$$

where $\psi = I_{(0,1/2]} - I_{(1/2,1]}$

Trees and Wavelets (Duality)

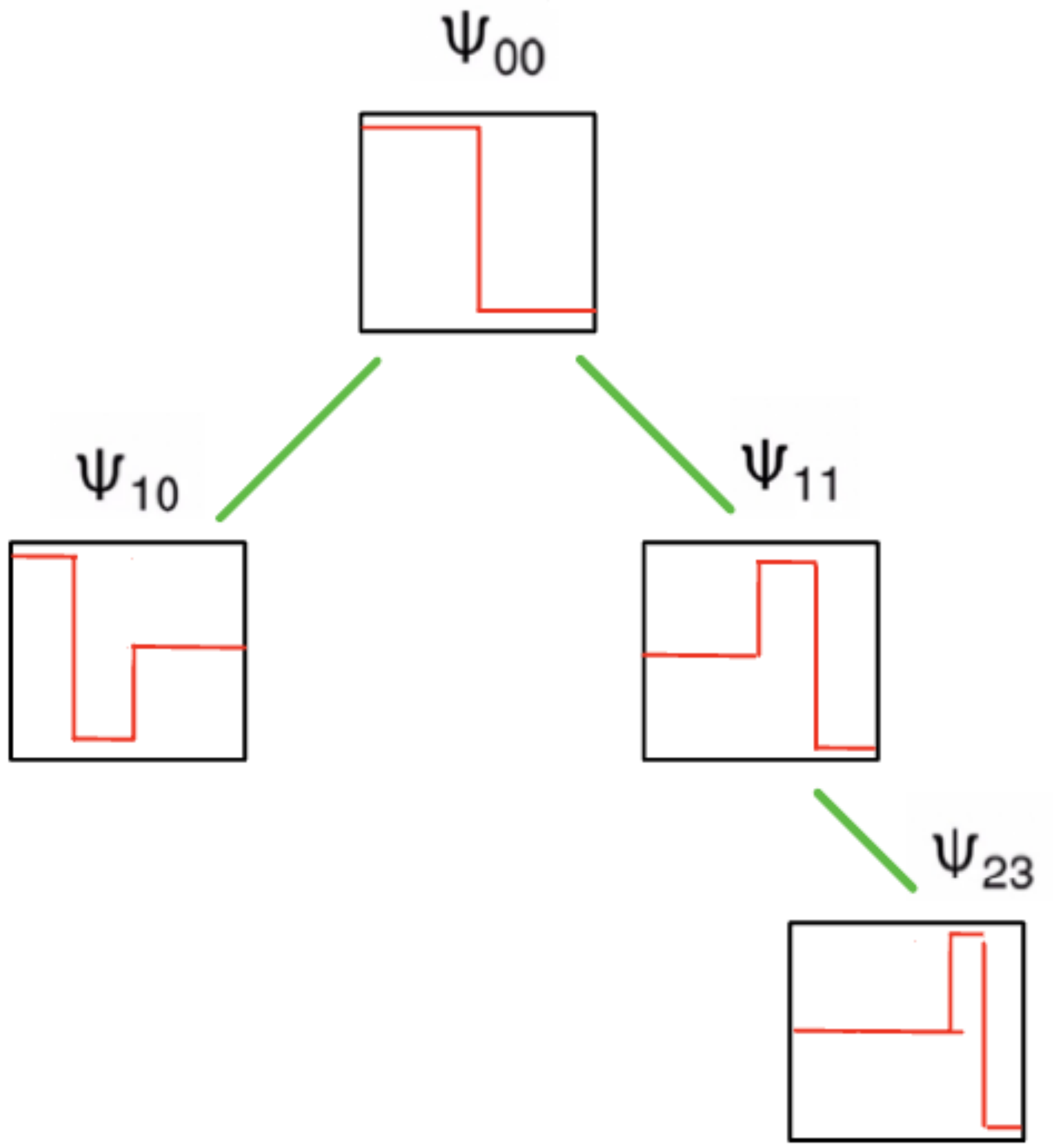


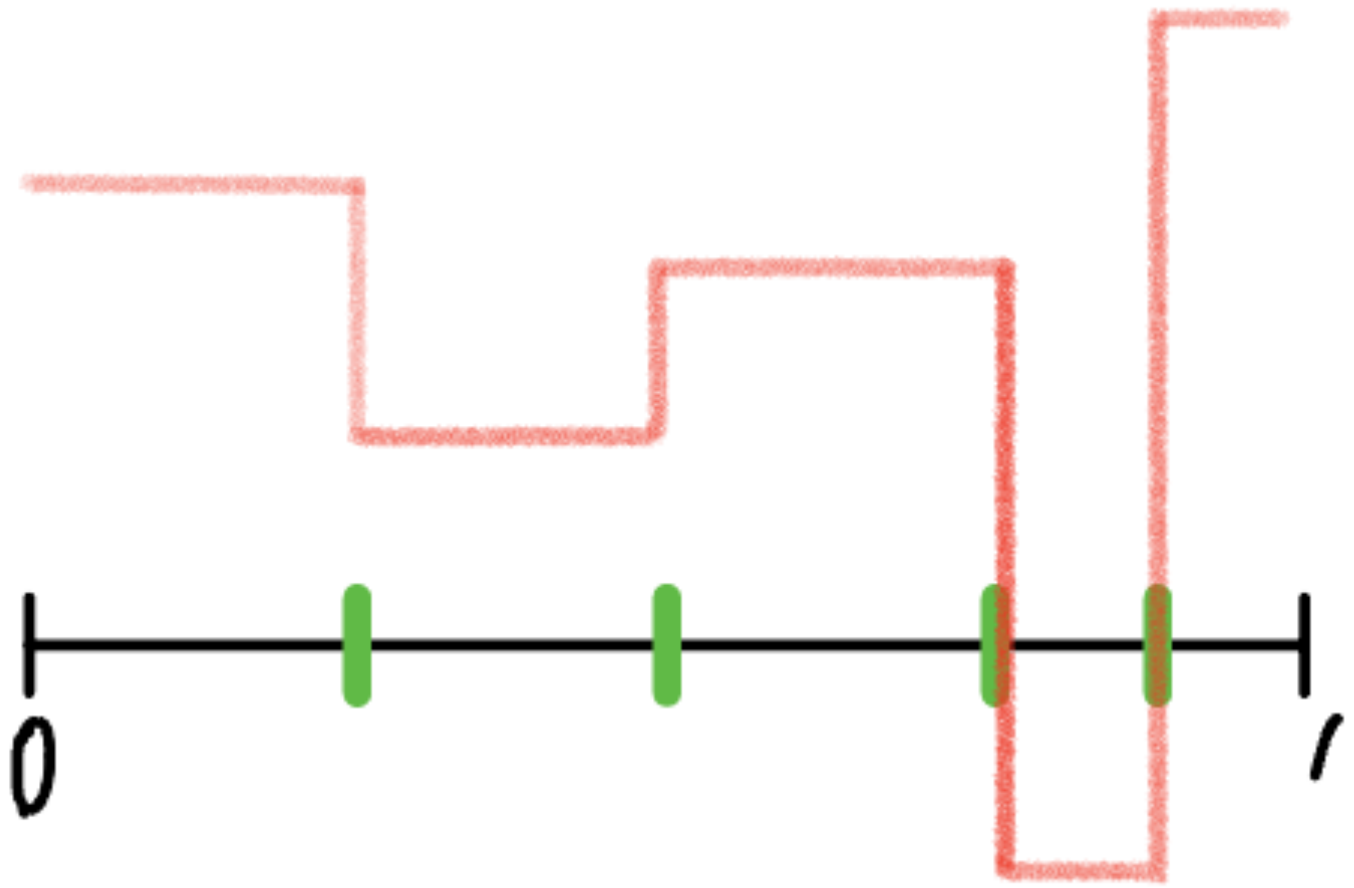


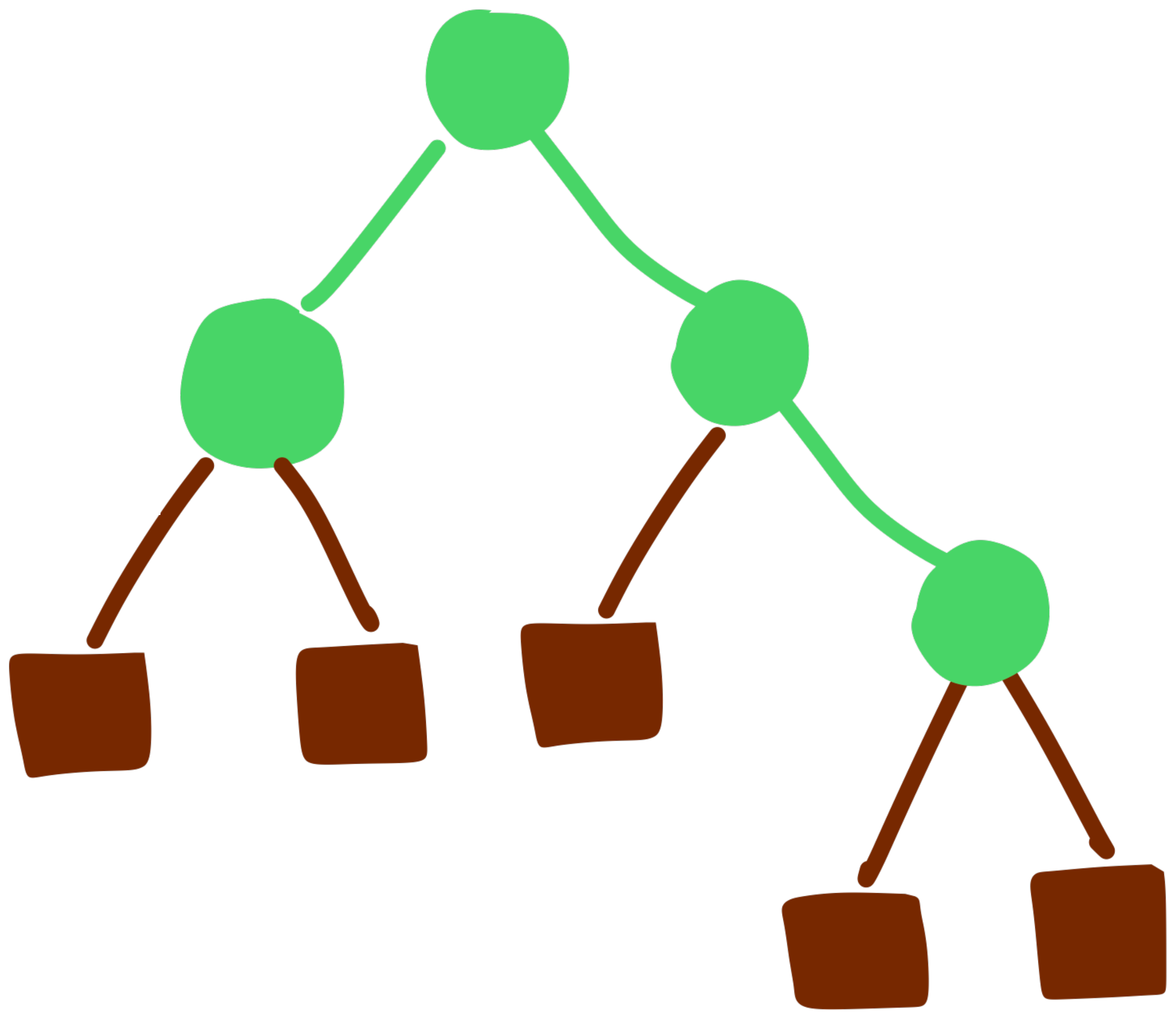


Donoho (1996)

Trees and Wavelets (Duality)

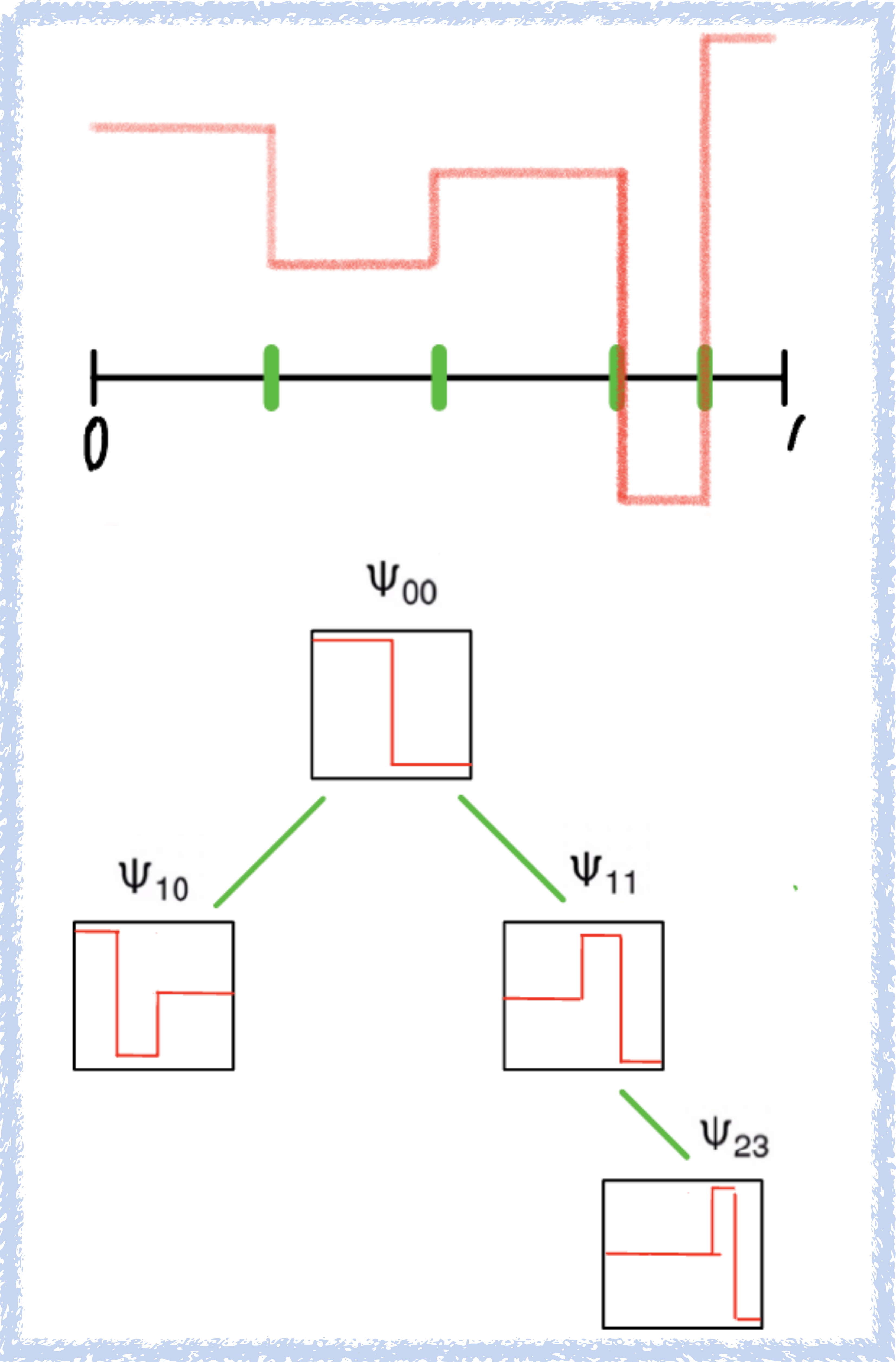






Donoho (1996)

Model



Data: $Y_i = f_0(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \quad i = 1,\ldots,n$

$$f_0(x) = \sum_{(l,k)\in\mathcal{B}} \psi_{lk}(x)\beta^*_{lk}, \text{ for } \mathcal{B} \subset \{(l,k) : l < L\}$$

Model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$

X grid: $\mathcal{X} = \{x_i = i/n : 1 \le i \le n\}$

Design matrix: $\mathbf{X} = (x_{ij})_{(n\times p)}$, where $p = 2^L$

$$x_{ij} = \begin{cases} \psi_{-10}(x_i) = 1 & \text{for} \quad j = 1 \\ \psi_{lk}(x_i) & \text{for} \quad j = 2^l + k + 1. \end{cases}$$

Bayesian CART prior

Tree -shaped wavelet shrinkage prior

$$\mathcal{T} \sim \Pi(\mathcal{T})$$

$$\{\beta_{lk}\}_{l \le L_{\max},k} \,|\, \mathcal{T} \sim \Pi(\boldsymbol{\beta}_{\mathcal{T}}) \otimes \bigotimes_{(l,k)\notin\mathcal{T}_{int}} \delta_0(\beta_{lk})\,.$$

Galton Watson Process

$$p_{lk} = \mathbb{P}[(l,k) \in \mathcal{T}_{int}] = n^{-c}$$

$$\Pi(\mathcal{T}) \propto \prod_{(l,k)\in\mathcal{T}_{int}} p_{lk} \prod_{(l,k)\in\mathcal{T}_{ext}} (1-p_{lk})$$

G-prior ($g_n = n$)

$$\pi(\boldsymbol{\beta}_{\mathcal{T}}) = N(0,I)$$

Posterior $\quad \pi(\mathcal{T} \,|\, Y) \propto \pi(\mathcal{T})L(\mathcal{T} \,|\, Y)$

Bayesian CART prior

Tree -shaped wavelet shrinkage prior

$$\mathcal{T} \sim \Pi(\mathcal{T})$$

$$\{\beta_{lk}\}_{l\le L_{\max},k} \mid \mathcal{T} \sim \Pi(\boldsymbol{\beta}_{\mathcal{T}}) \otimes \bigotimes_{(l,k)\notin\mathcal{T}_{int}} \delta_0(\beta_{lk})\,.$$

$$L(Y\mid\mathcal{T}) = \frac{\exp\left\{-\frac{1}{2}Y'[I - X_{\mathcal{T}}\Sigma_{\mathcal{T}}X_{\mathcal{T}}]'Y\right\}}{(2\pi)^{n/2}(1+n)^{|\mathcal{T}_{ext}|/2}}$$

$\Sigma_{\mathcal{T}} = c_n(X'_{\mathcal{T}}X_{\mathcal{T}})^{-1}$ with $c_n = n/(n+1)$

Posterior $\pi(\mathcal{T}\mid Y) \propto \pi(\mathcal{T})L(\mathcal{T}\mid Y)$

Galton Watson Process

$$p_{lk} = \mathbb{P}[(l,k)\in\mathcal{T}_{int}] = n^{-c}$$

$$\Pi(\mathcal{T}) \propto \prod_{(l,k)\in\mathcal{T}_{int}} p_{lk} \prod_{(l,k)\in\mathcal{T}_{ext}} (1-p_{lk})$$

G-prior ($g_n = n$)

$$\pi(\boldsymbol{\beta}_{\mathcal{T}}) = N(0,I)$$

Assumptions

Data: $Y_i = f_0(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \quad i = 1,\ldots,n$

$$f_0(x) = \sum_{(l,k)\in\mathscr{B}} \psi_{lk}(x)\beta^*_{lk}, \text{ for } \mathscr{B} \subset \{(l,k) : l < L\}$$

Signals: $A \log n/\sqrt{n} < |\beta^*_{lk}| < C_{f_0}$ for all $(l,k) \in \mathscr{B}$ for some $A > 0,\ C_{f_0} > 0.$

$\mathscr{T}^* \in \mathbb{T}_L$: the smallest tree that contains all signals $\mathscr{B}$

Assumptions

Data: $Y_i = f_0(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \quad i = 1,\ldots,n$

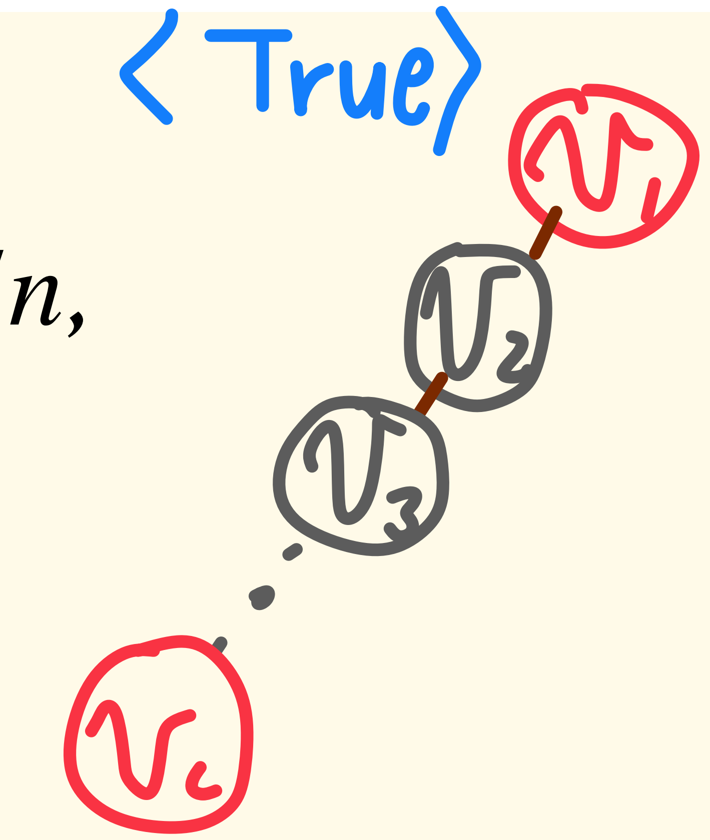
$$f_0(x) = \sum_{(l,k)\in\mathscr{B}} \psi_{lk}(x)\beta^*_{lk}, \text{ for } \mathscr{B} \subset \{(l,k) : l < L\}$$

Signals: $A \log n/\sqrt{n} < |\beta^*_{lk}| < C_{f_0}$ for all $(l,k) \in \mathscr{B}$ for some $A > 0,\ C_{f_0} > 0.$

$\mathscr{T}^* \in \mathbb{T}_L$: the smallest tree that contains all signals $\mathscr{B}$

Consistency: With probability $\geq 1 - 4/n$,

$$\pi(\mathscr{T}^* \mid Y) \geq 1 - o(1)$$



Assumptions

Data: $Y_i = f_0(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \quad i = 1,\ldots,n$

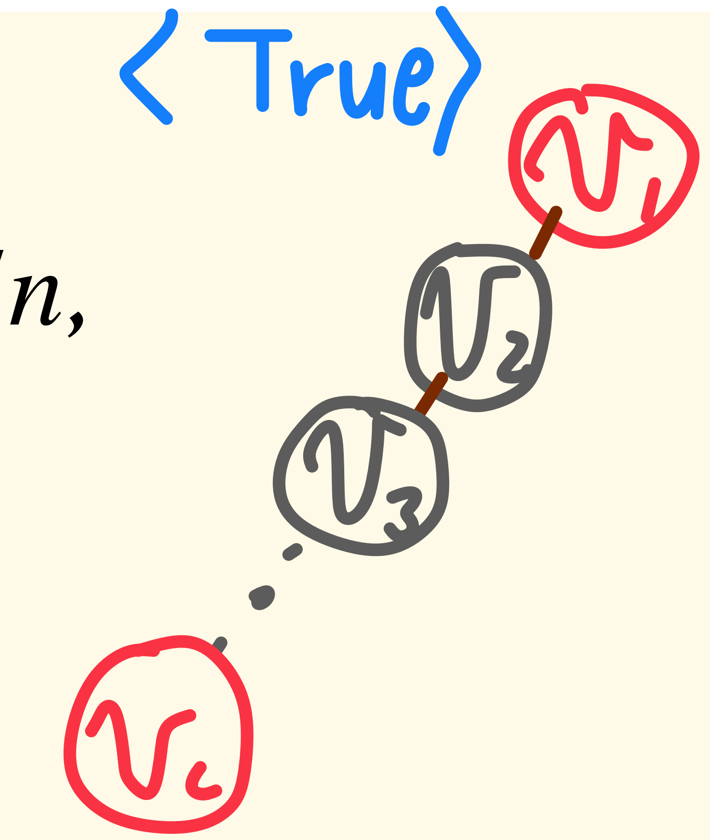
$$f_0(x) = \sum_{(l,k)\in\mathscr{B}} \psi_{lk}(x)\beta^*_{lk}, \text{ for } \mathscr{B} \subset \{(l,k) : l < L\}$$

Signals: $A \log n/\sqrt{n} < |\beta^*_{lk}| < C_{f_0}$ for all $(l,k) \in \mathscr{B}$ for some $A > 0,\ C_{f_0} > 0.$

$\mathscr{T}^* \in \mathbb{T}_L$: the smallest tree that contains all signals $\mathscr{B}$

Consistency: With probability $\geq 1 - 4/n$,

$$\pi(\mathscr{T}^* \mid Y) \geq 1 - o(1)$$



Variable selection:

⟨True⟩

v_1, v_L

$$A\sqrt{\log n/n} < |\beta^*_{lk}| < C_{f_0}$$

Yang et al (2016)

Assumptions

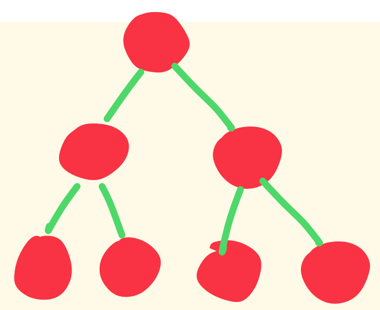
Data: $Y_i = f_0(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \quad i = 1,\ldots,n$

$$f_0(x) = \sum_{(l,k)\in\mathcal{B}} \psi_{lk}(x)\beta^*_{lk}, \text{ for } \mathcal{B} \subset \{(l,k) : l < L\}$$

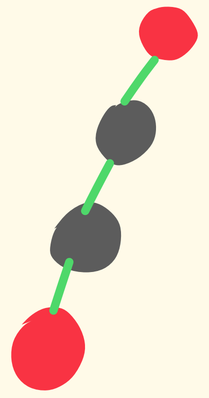
Signals: $A \log n/\sqrt{n} < |\beta^*_{lk}| < C_{f_0}$ for all $(l,k) \in \mathcal{B}$ for some $A > 0,\ C_{f_0} > 0.$

$\mathcal{T}^* \in \mathbb{T}_L$: the smallest tree that contains all signals $\mathcal{B}$

(Connected signals) Assume that $\mathcal{B} \in \mathbb{T}_L$, i.e. $\mathcal{T}^* = \mathcal{B}$

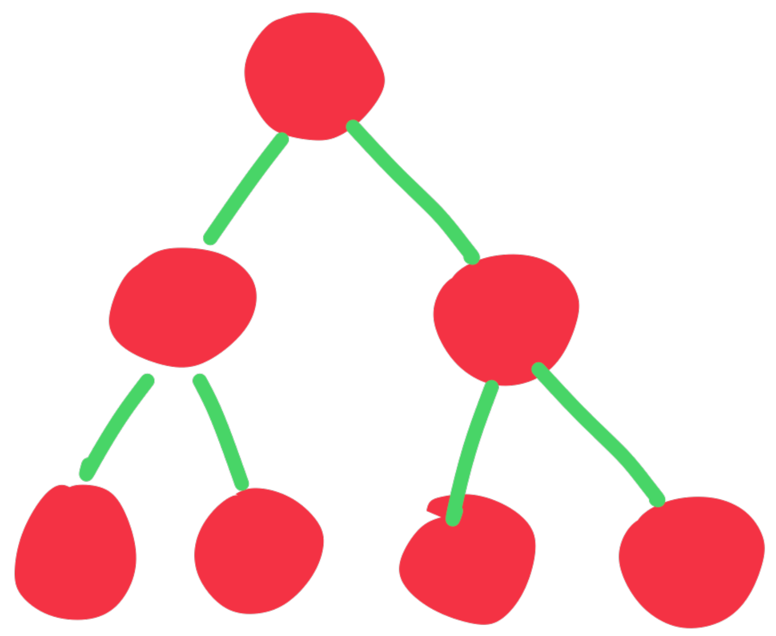


(Disconnected signals) Assume that $\mathcal{B} \notin \mathbb{T}_L$ (at least all preterminal nodes are signals)

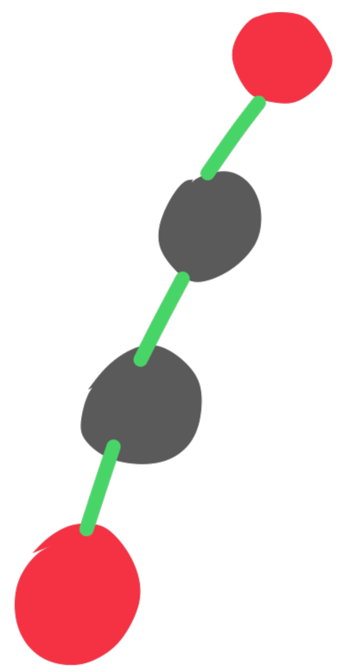


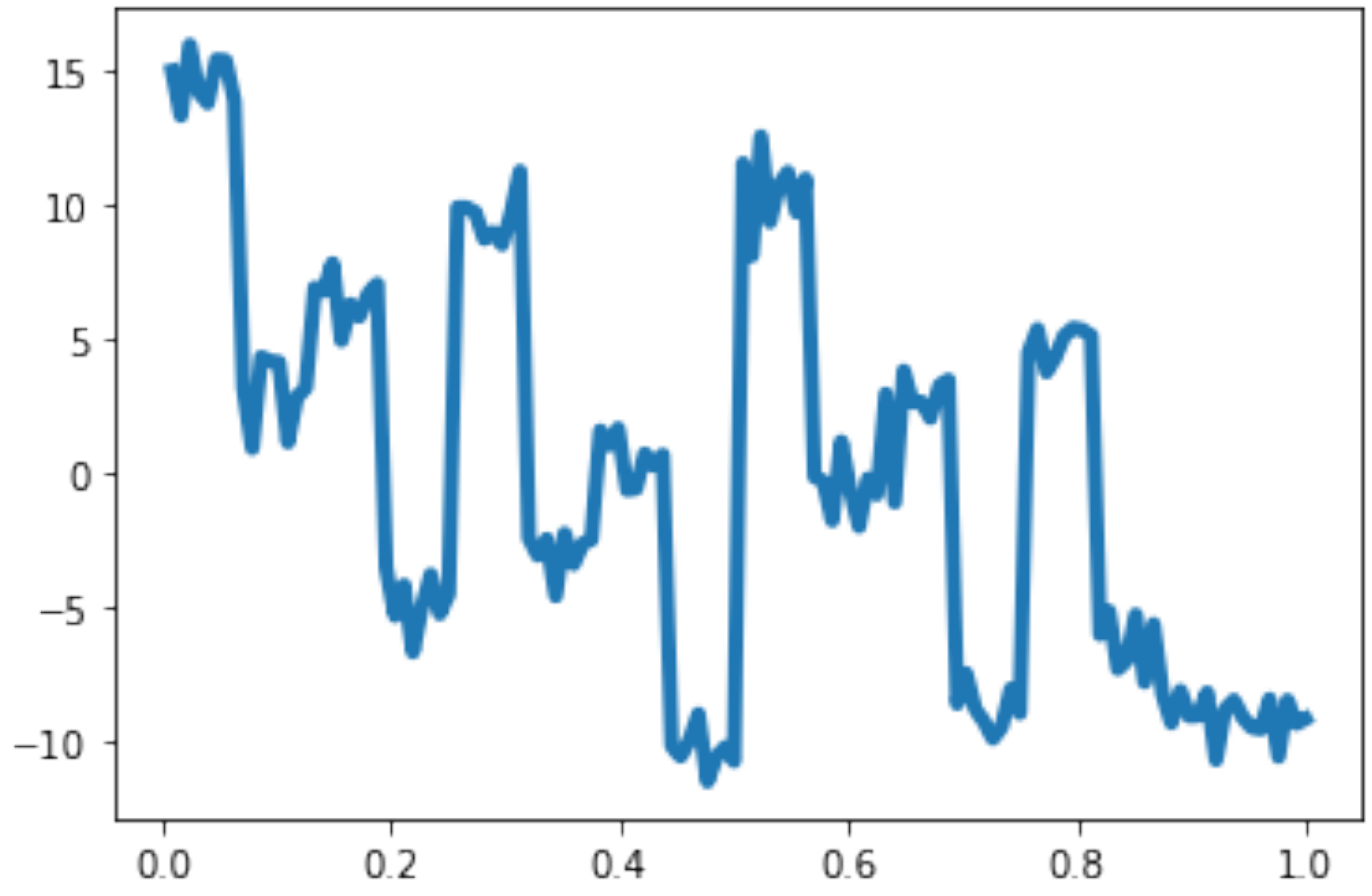
Data looks like:

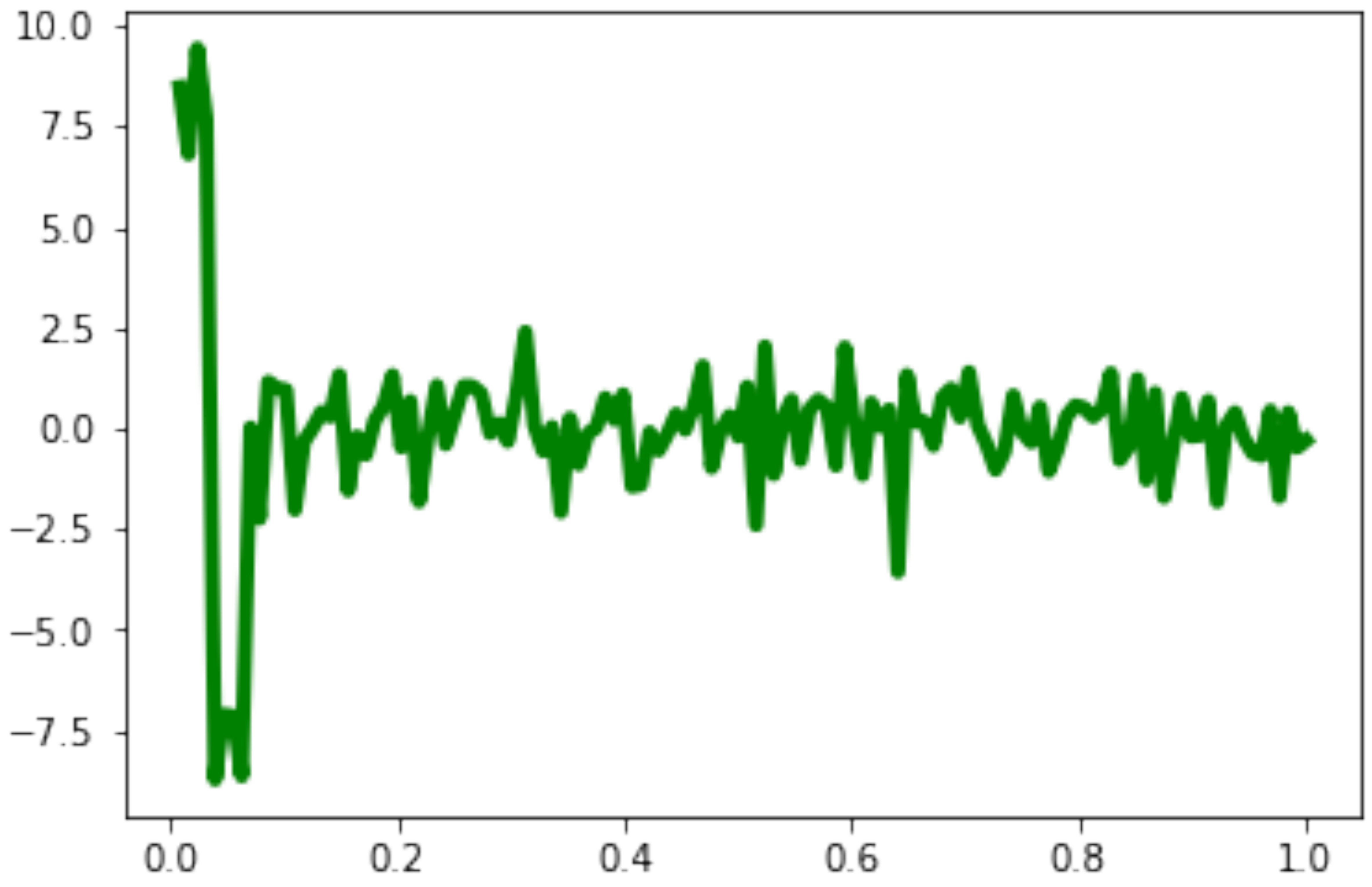
(Connected signals) $\mathscr{T}^* = \mathscr{B}$



(Disconnected signals) $\mathscr{B} \notin \mathbb{T}_L$,







Mixing time

$P^t(\mathscr{T}, \cdot\,)$: the distribution of the state at time t with an initial condition $\mathscr{T}$

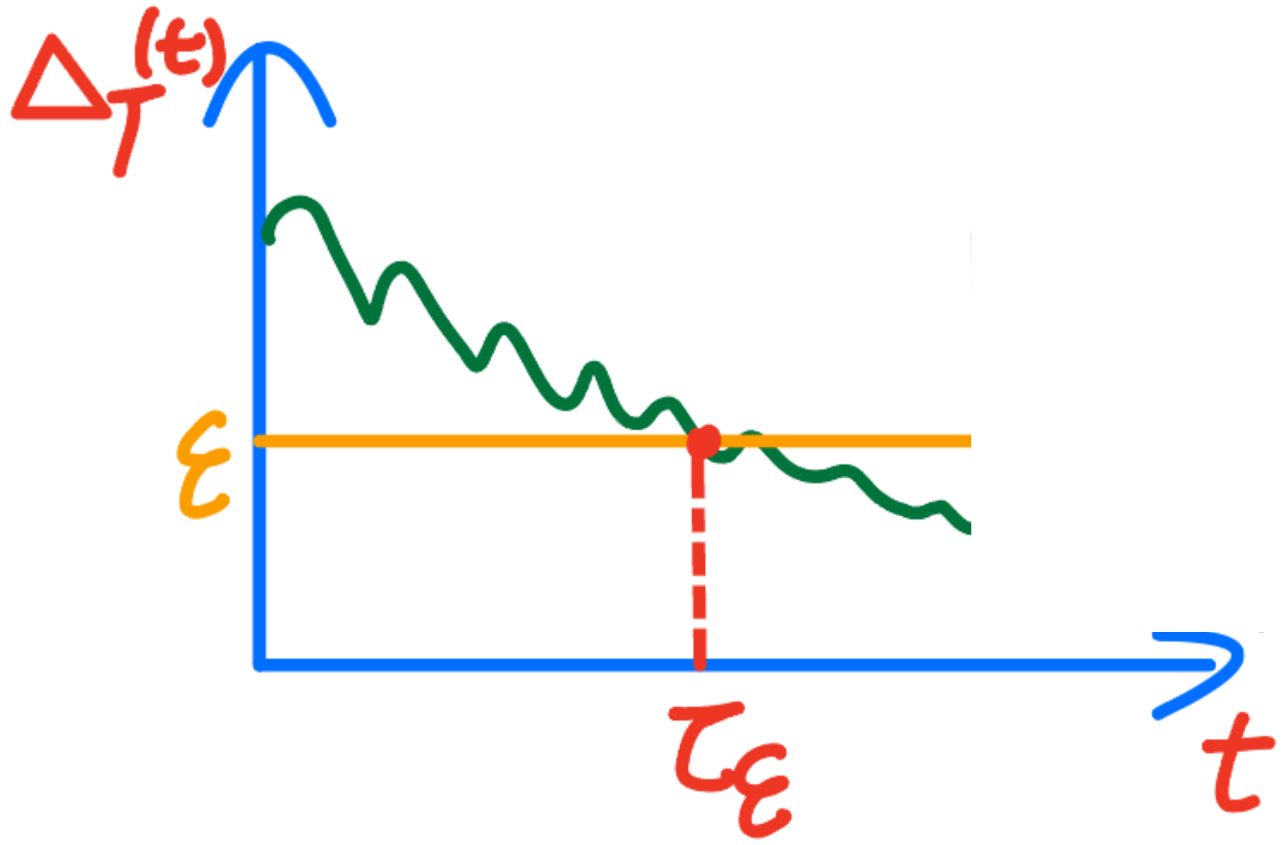
The total variation distance to the stationary distribution after t iterations

$$\Delta_{\mathscr{T}}(t) = \| P^t(\mathscr{T}, \cdot\,) - \Pi[\,\cdot \mid Y] \|_{TV} \equiv \max_{S \subset \mathbb{T}_L} | P^t(\mathscr{T}, S) - \Pi[S \mid Y] |$$

Definition

The ϵ-mixing time of the Markov chain is defined as

$$\tau_\epsilon \equiv \max_{\mathscr{T} \in \mathbb{T}_L} \min\{t \in \mathscr{N} : \Delta_{\mathscr{T}}(t') \leq \epsilon \quad \text{for all} \quad t' \geq t\}$$



Outline

Background : Tree and Mixing Rate

Setting ups and Mixing Time Framework

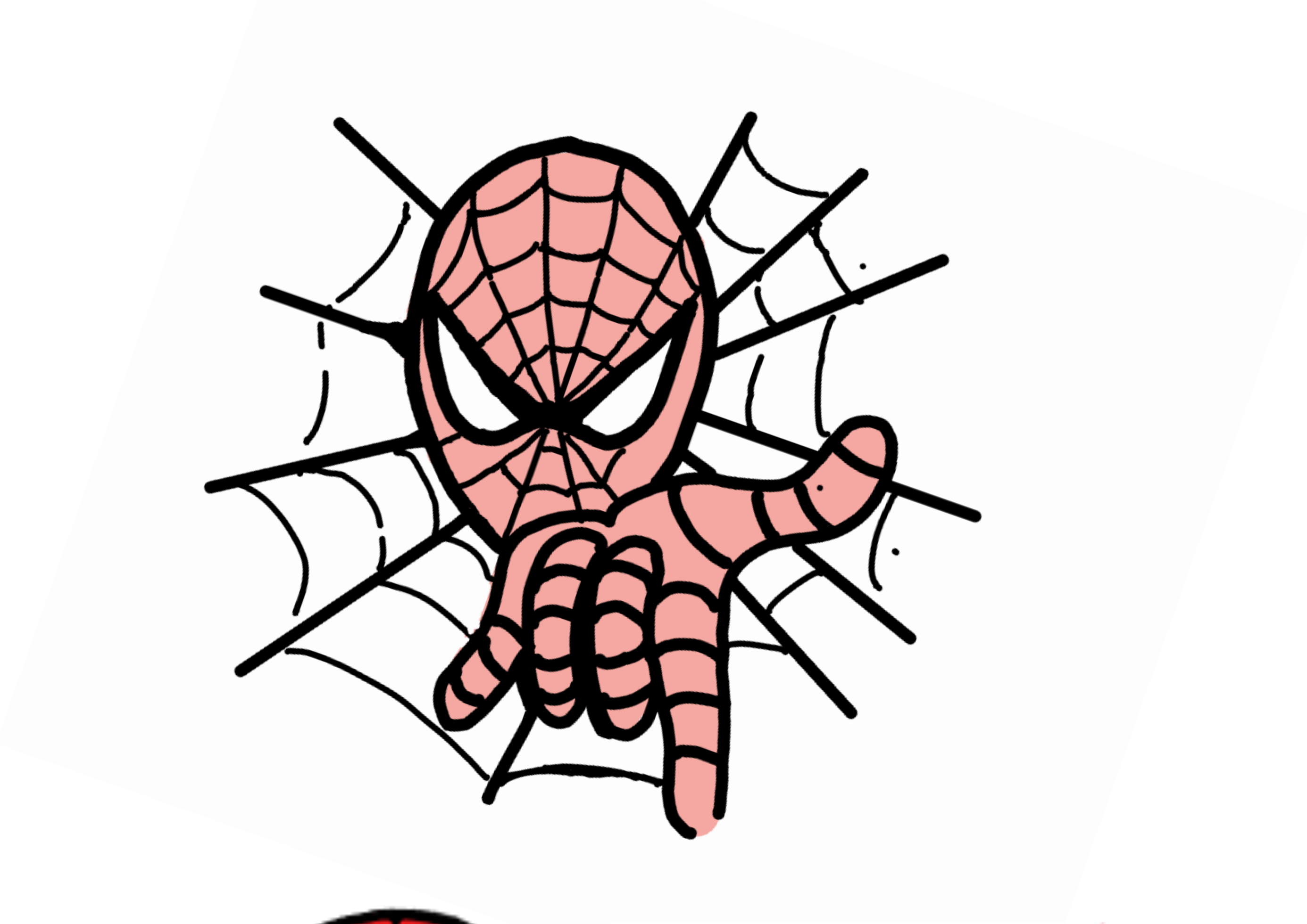
Theoretical and Numerical Results

Concluding Remarks

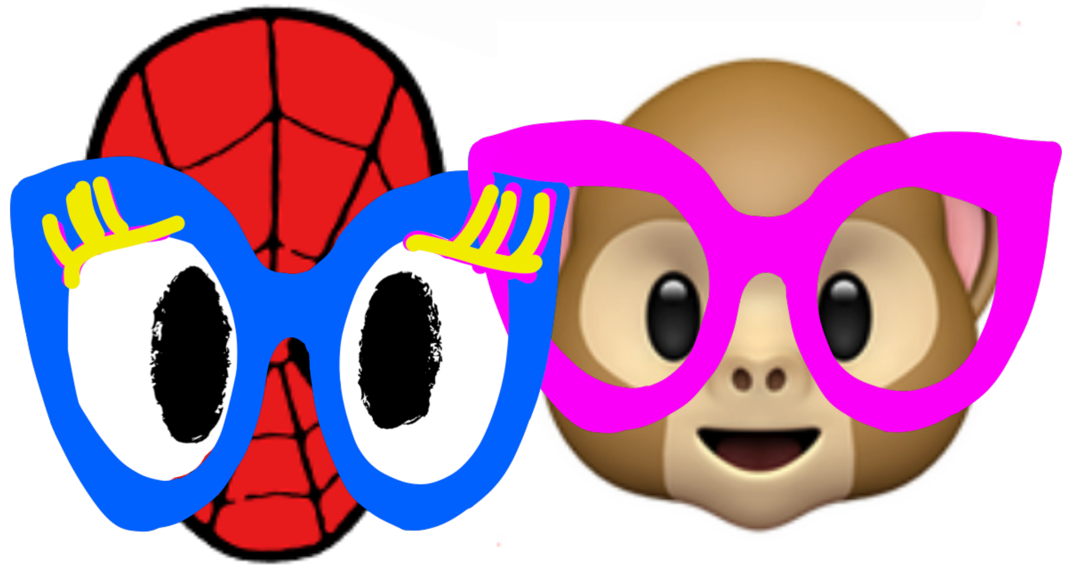
Theoretical and Numerical Results

1. Bayesian CART

2. Twiggy CART

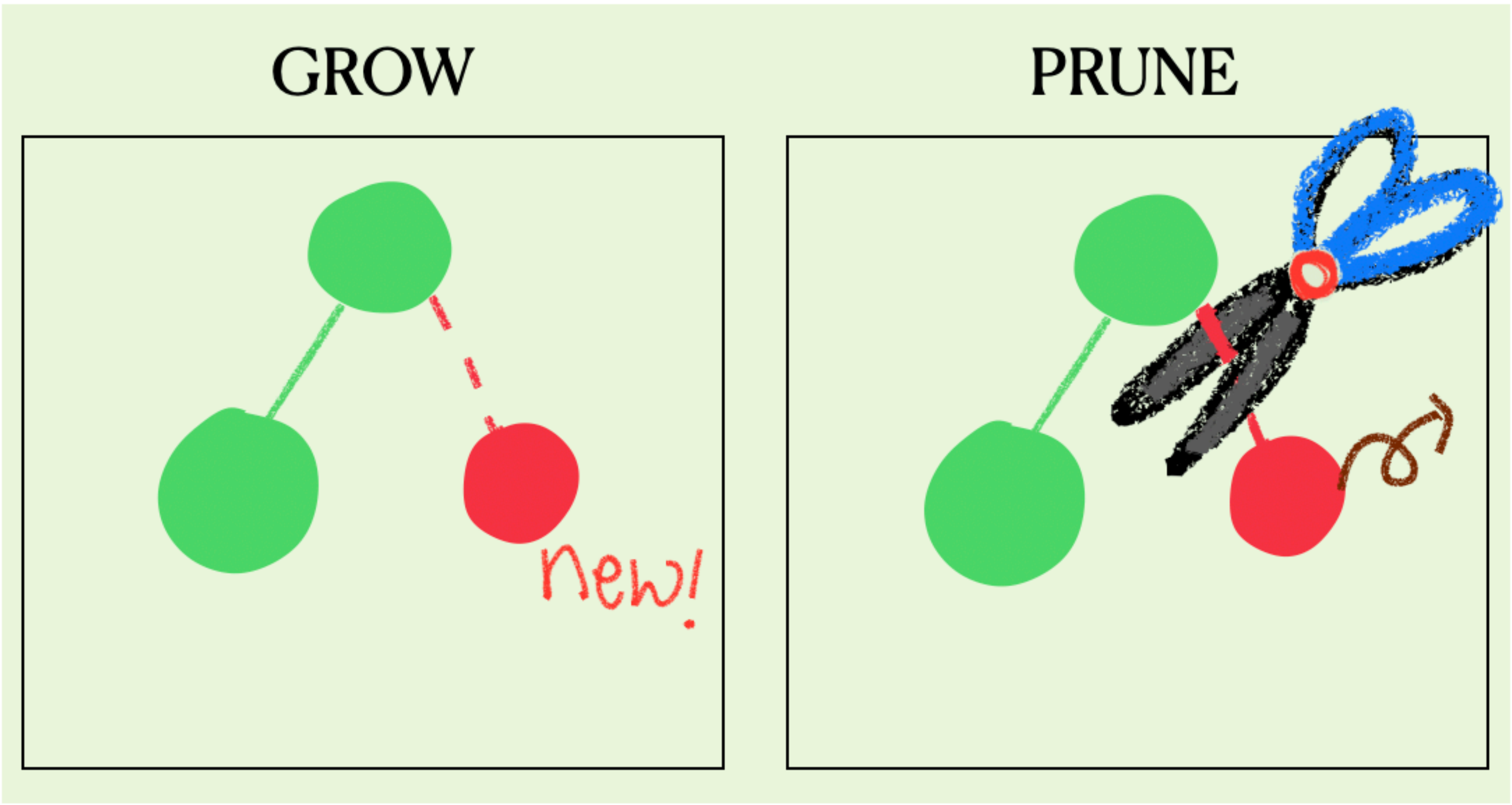


3. Informed MCMC

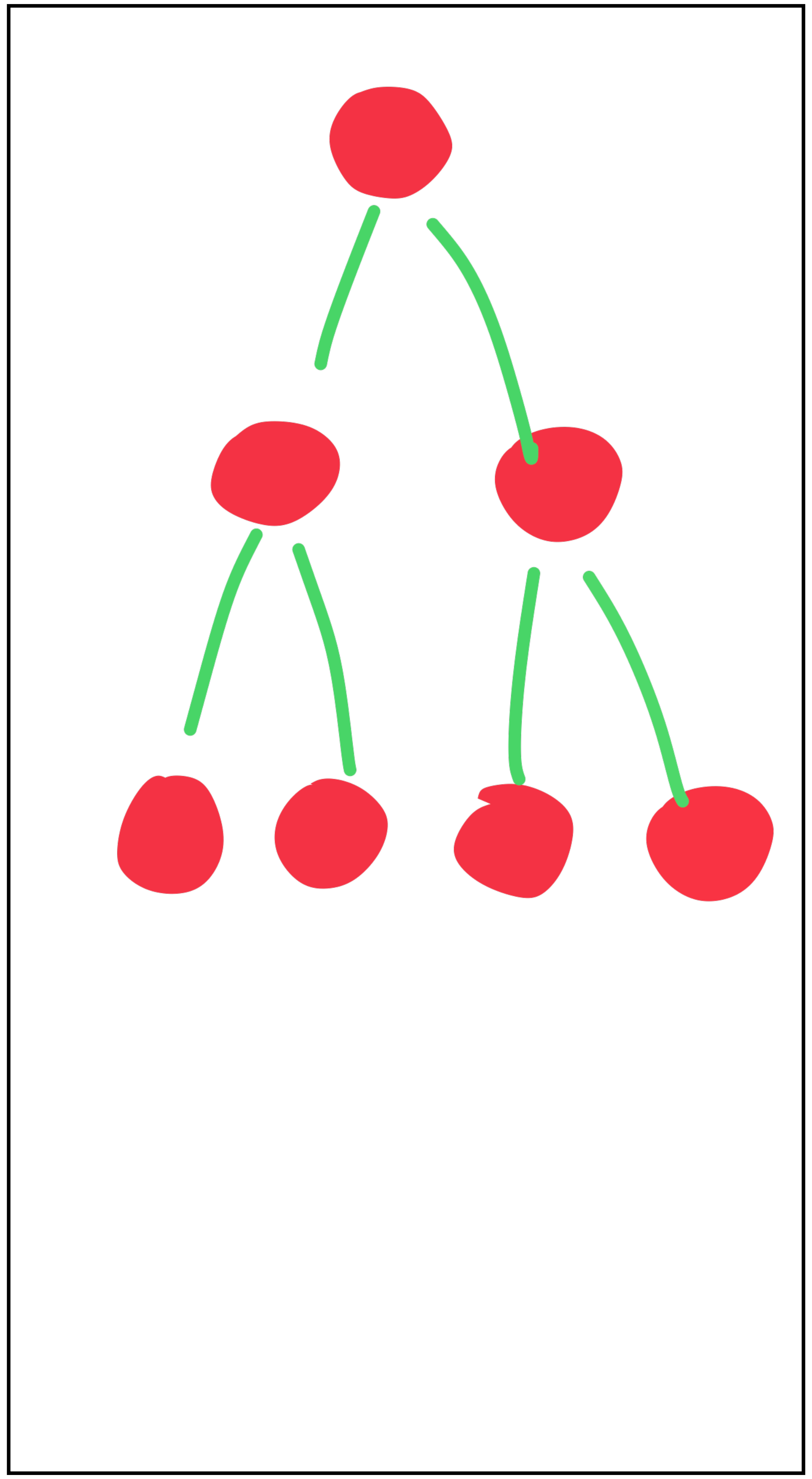


1. Bayesian CART

Proposal of Bayesian CART



Connected Signals: Rapid Mixing



Theorem 1 (Bayesian CART algorithm)

Assume the connected signals.

$$\tau_\epsilon \leq 2^{2L+3} \left\{ n \left[\left(c + \frac{1}{2} \right) \log(1+n) + |\mathcal{T}^*_{int}| C^2_{f_0} + 1 \right] + 4 |\mathcal{T}^*_{int}| \log n + \log \left(\frac{2}{\epsilon} \right) \right\}$$

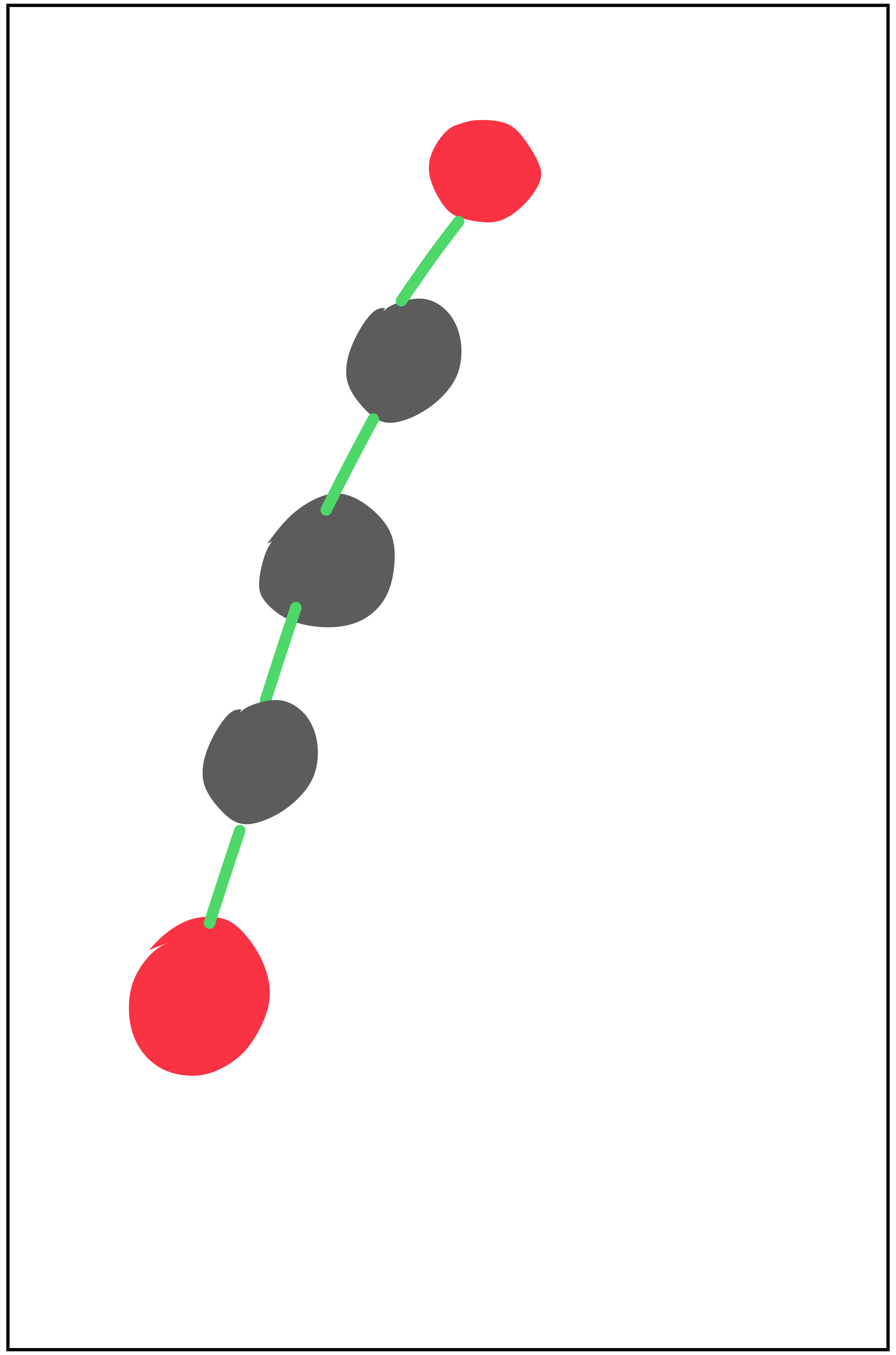
with probability at least $1 - 4/n$.

$p = 2^L$: the number of problem parameters $(2^L \leq n/2)$

$|\mathcal{T}^*_{int}|$: the number of internal nodes of true tree $\mathcal{T}^*$

C_{f_0}: the bound for the maximum signal strength

Disconnected Signals: Slow Mixing



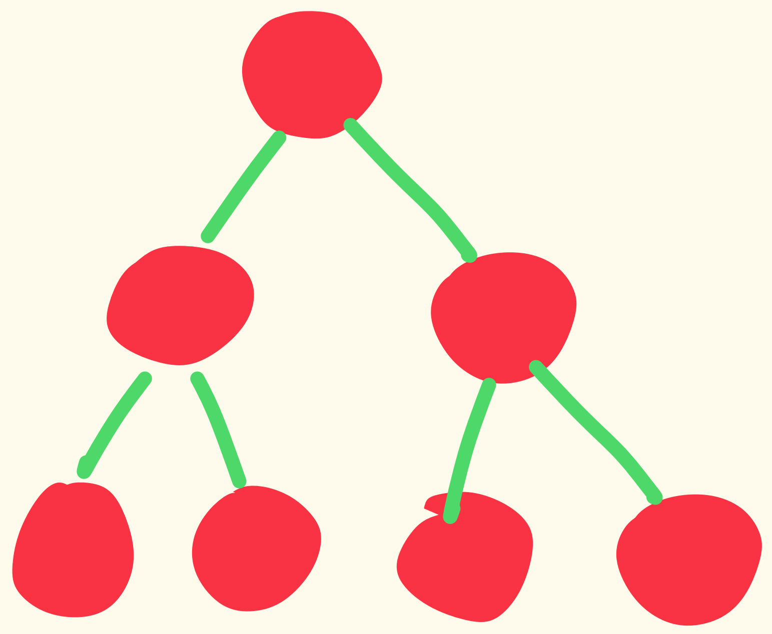
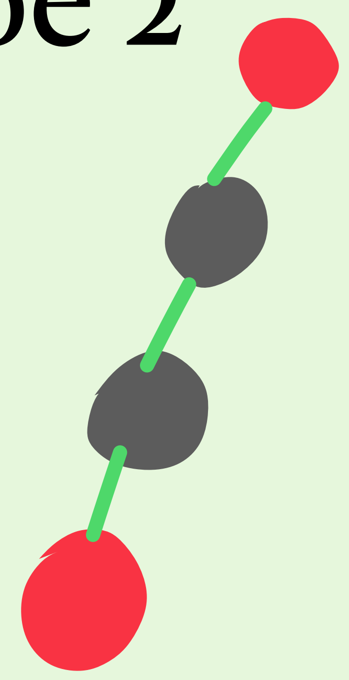
Theorem 2 (Bayesian CART algorithm)

Assume the disconnected signals. There exists f_0 such that

$$\tau_\epsilon > \log\left(\frac{1}{2\epsilon}\right)\frac{1}{4}\left[\left(\frac{n^{(c-3/2)}/4-1}{C}\right)^{L-2}-3\right]$$

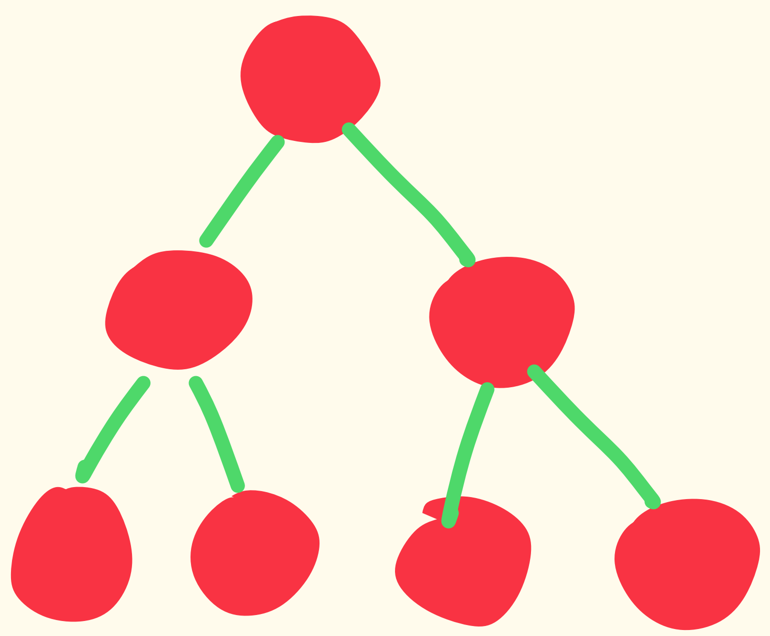
with probability at least $1-4/n$ for some $C>1$.

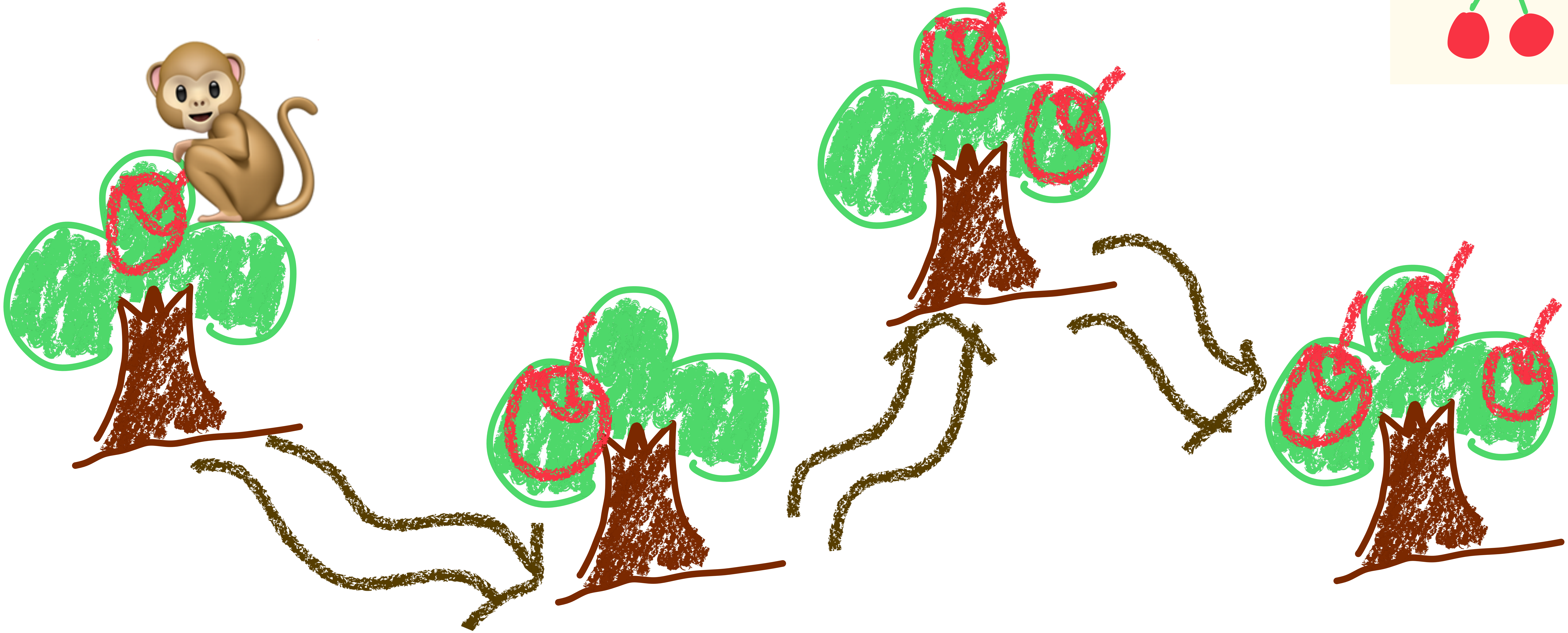
Deeper the signal is, worse the mixing rate becomes.

	Type 1 	Type 2 
B-CART	Rapid	Slow

Connected Signals

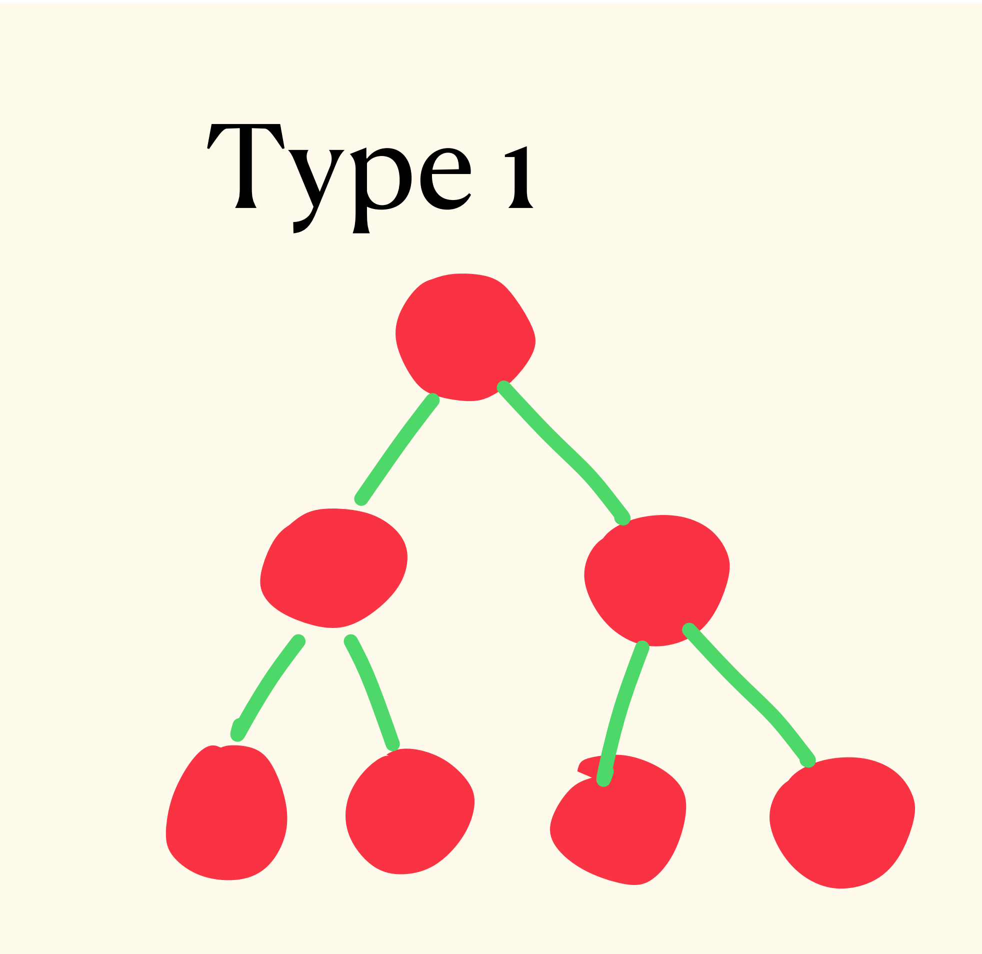
Type 1

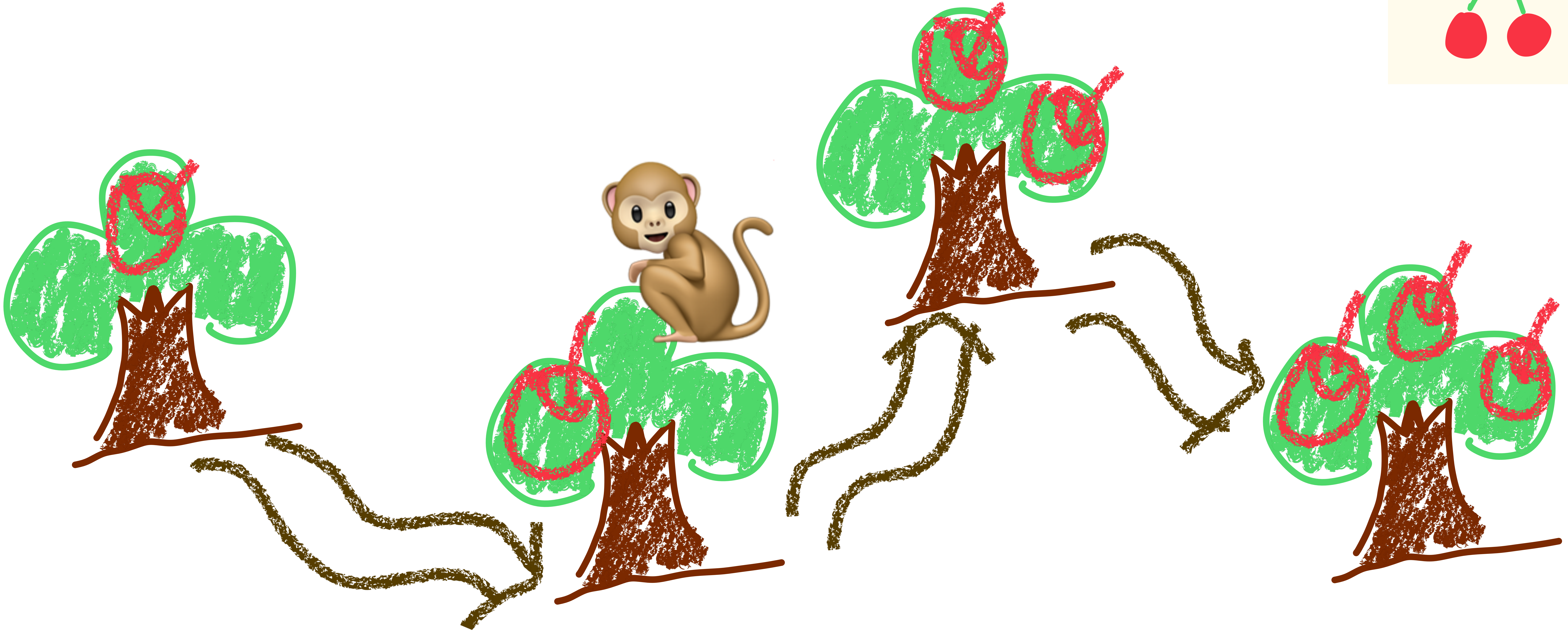




There were plenty of food on the way to the apple tree $\mathscr{T}^*$

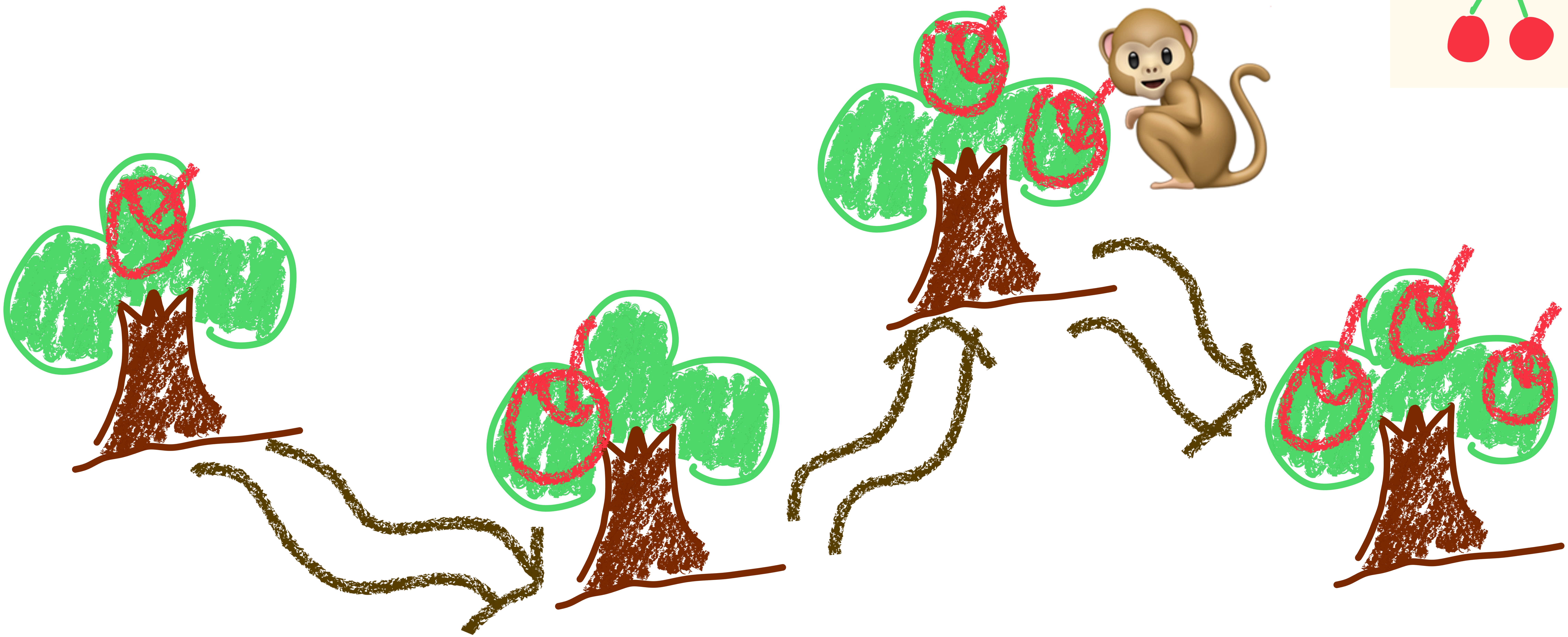
Connected Signals

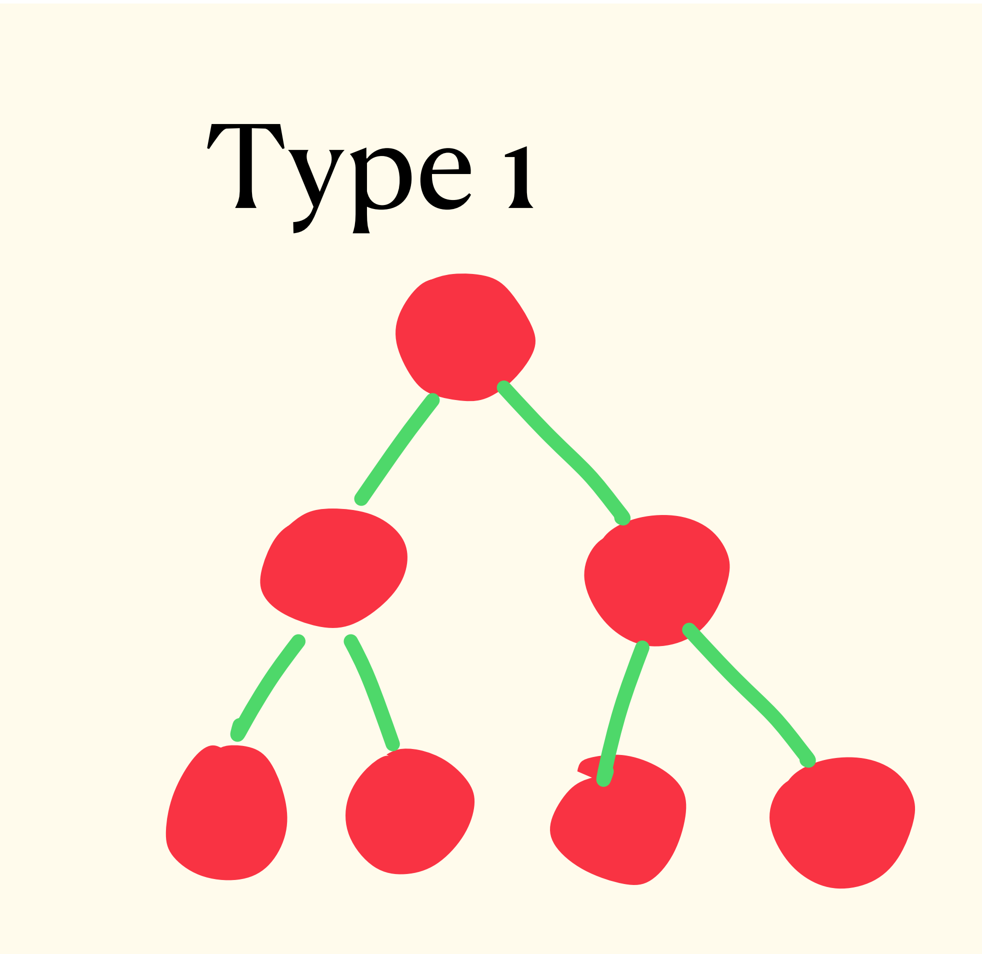




There were plenty of food on the way to the apple tree $\mathscr{T}^*$

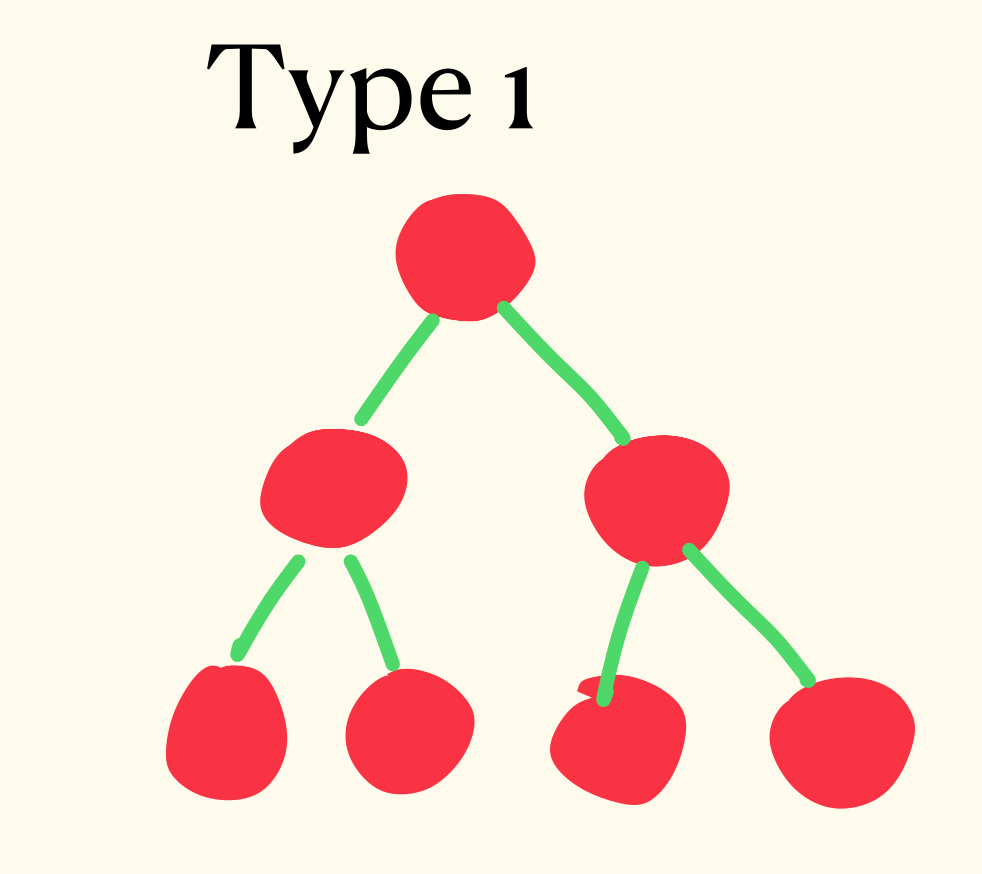
Connected Signals

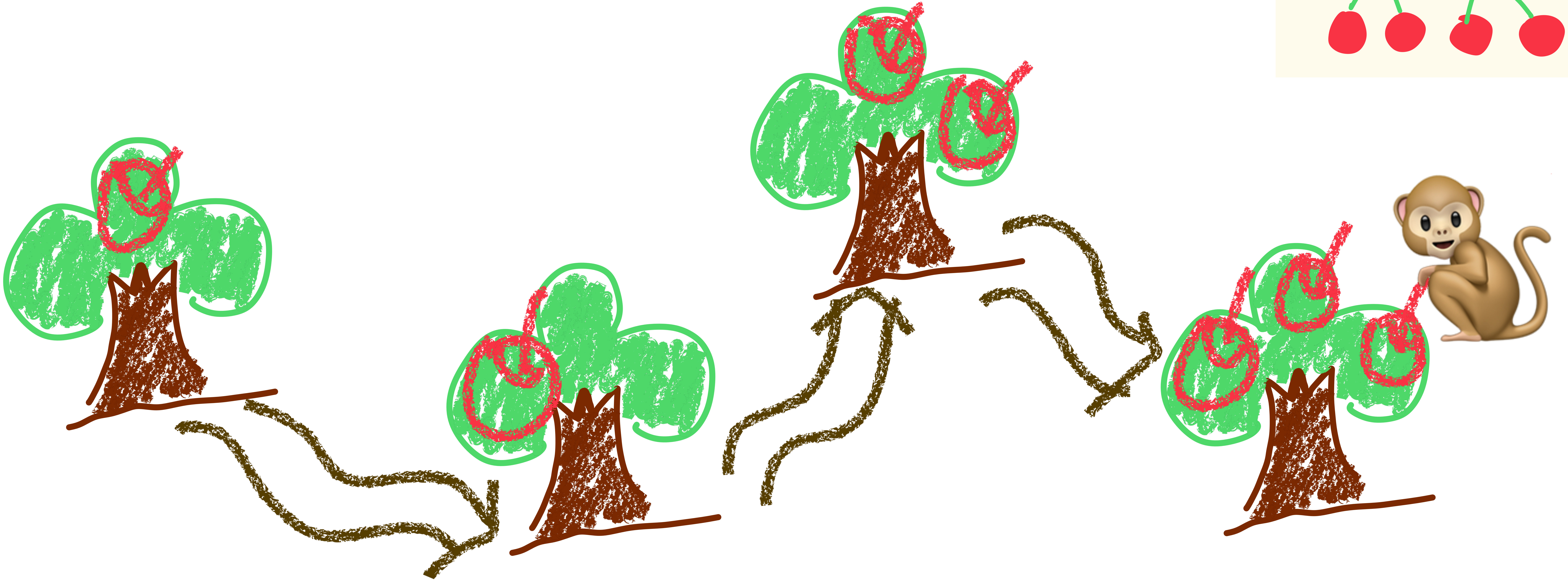




There were plenty of food on the way to the apple tree $\mathscr{T}^*$

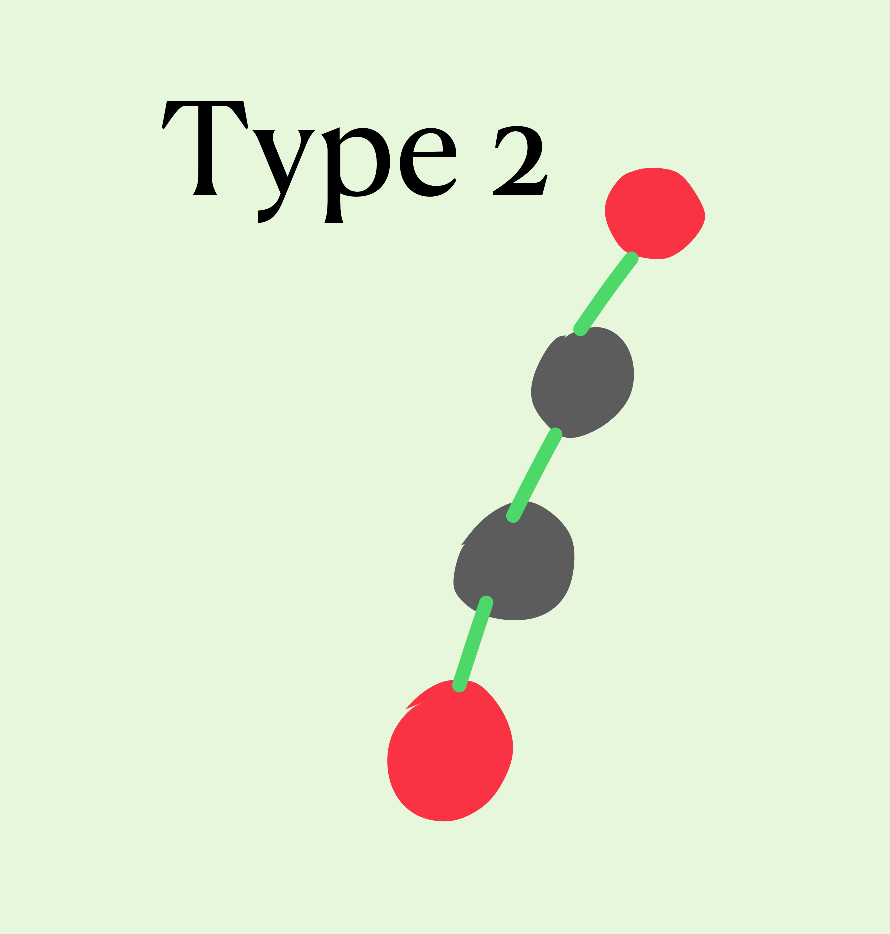
Connected Signals

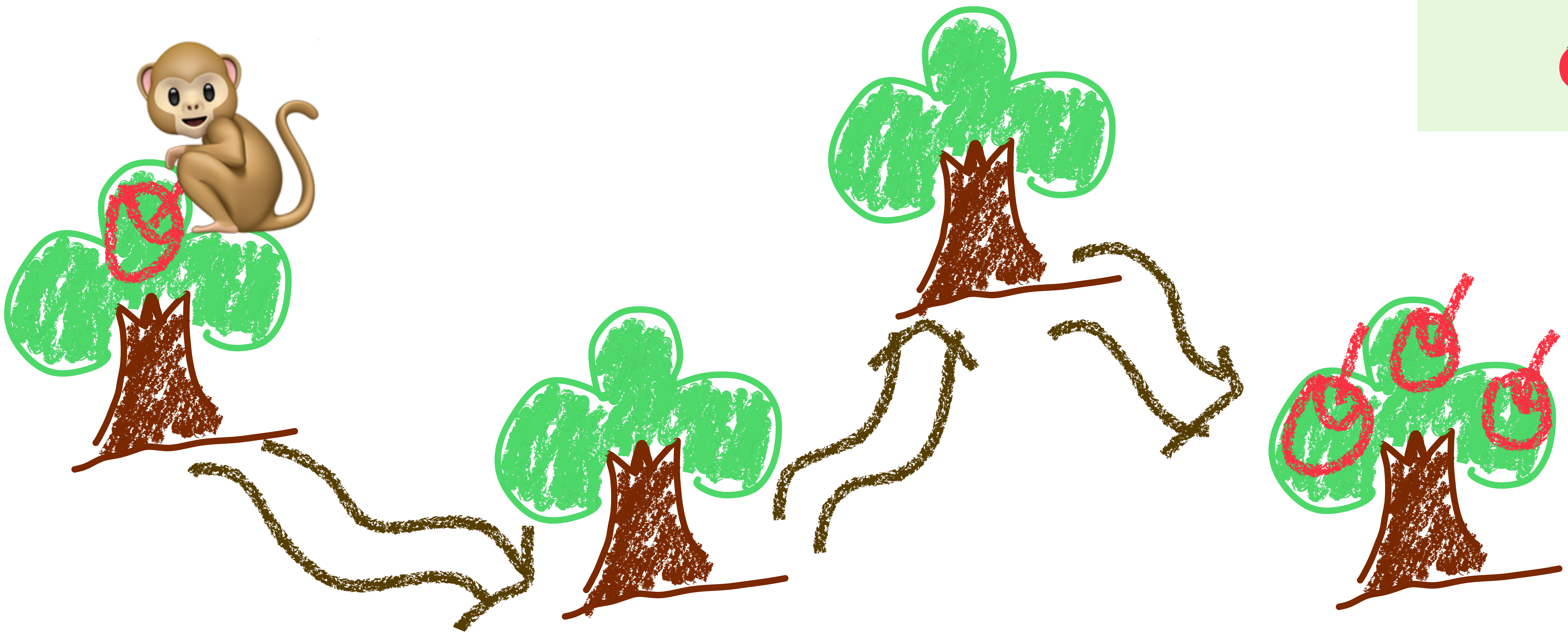




There were plenty of food on the way to the apple tree $\mathcal{T}^*$

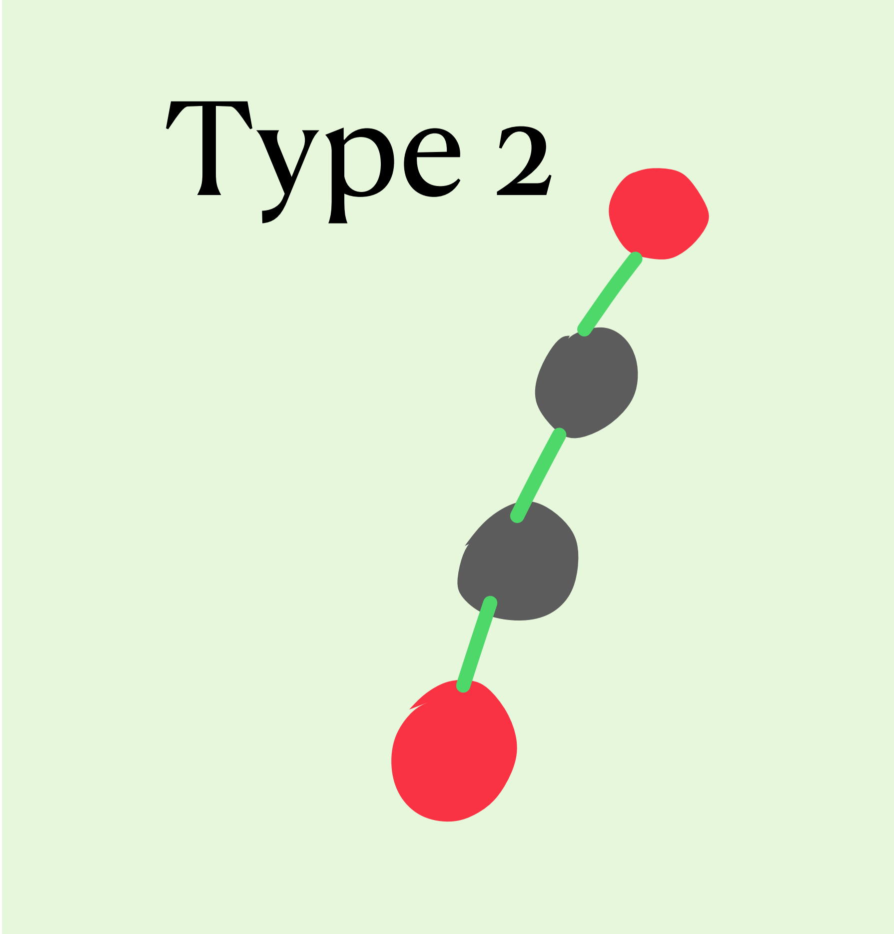
Disconnected Signals





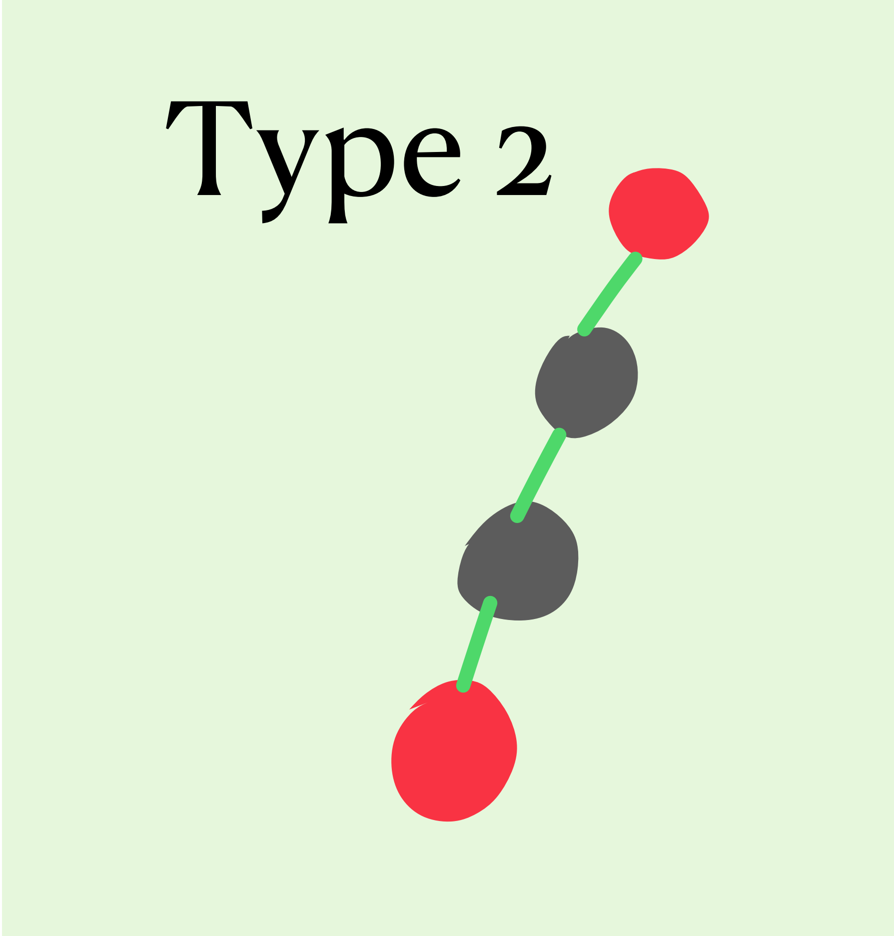
The monkey became hungry on the way to the apple tree $\mathscr{T}^*$, so went back home.

Disconnected Signals



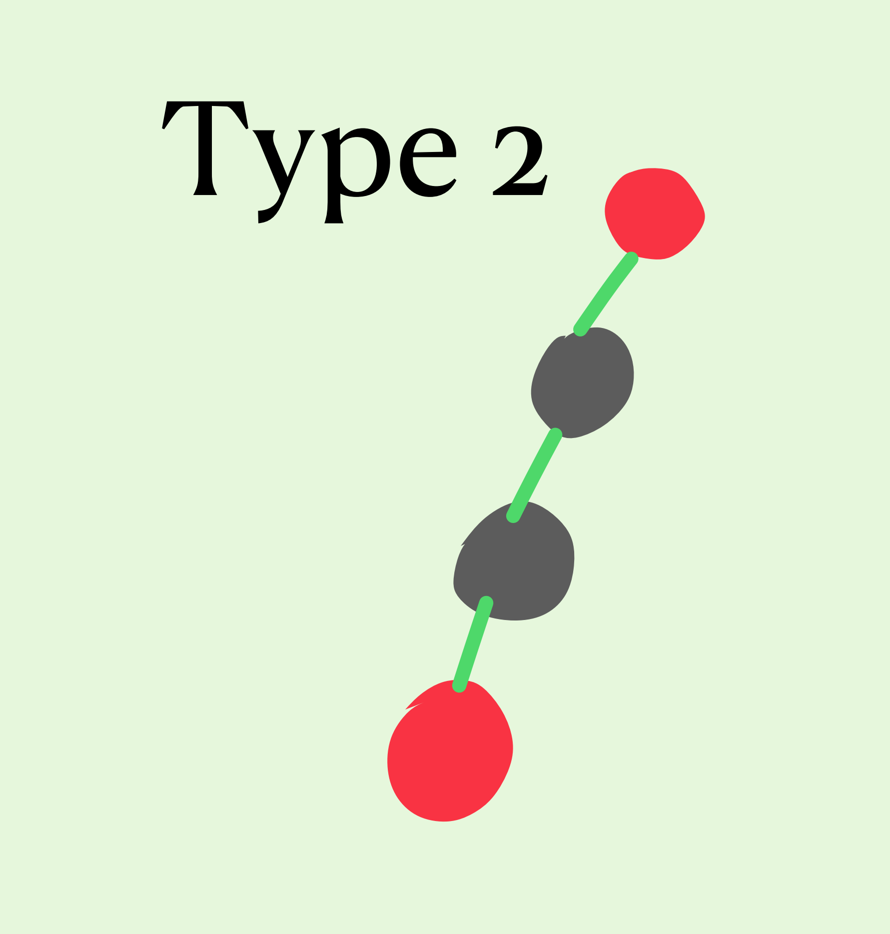
The monkey became hungry on the way to the apple tree $\mathscr{T}^*$, so went back home.

Disconnected Signals



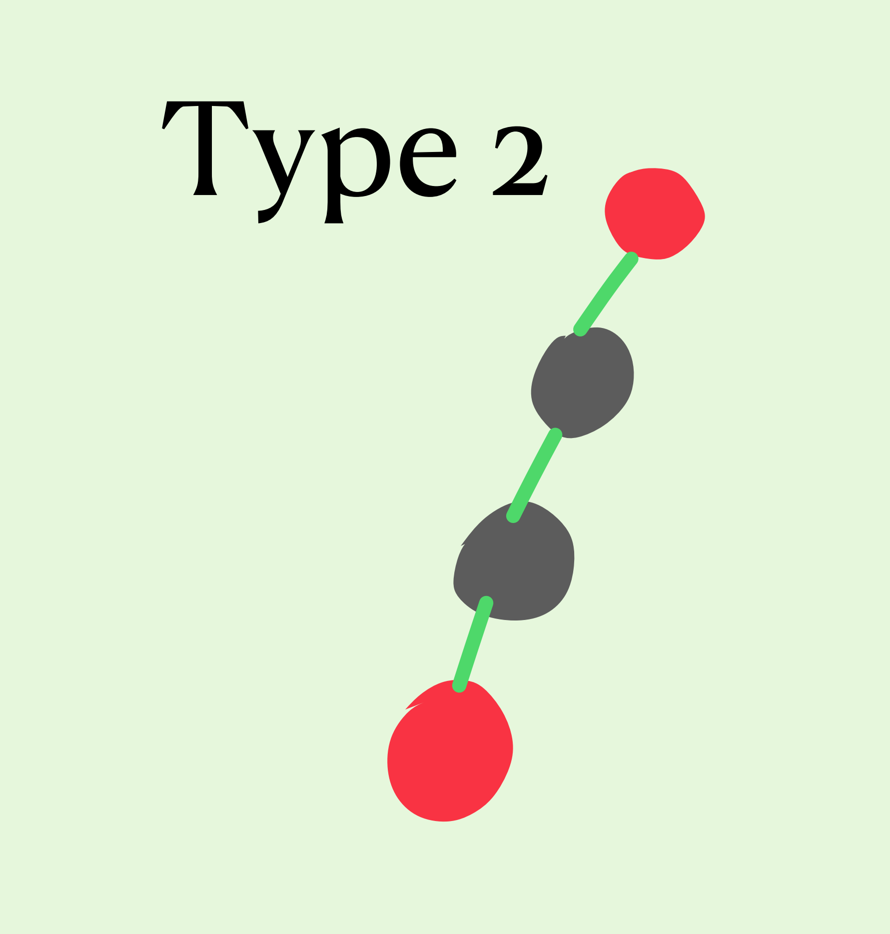
The monkey became hungry on the way to the apple tree $\mathscr{T}^*$, so went back home.

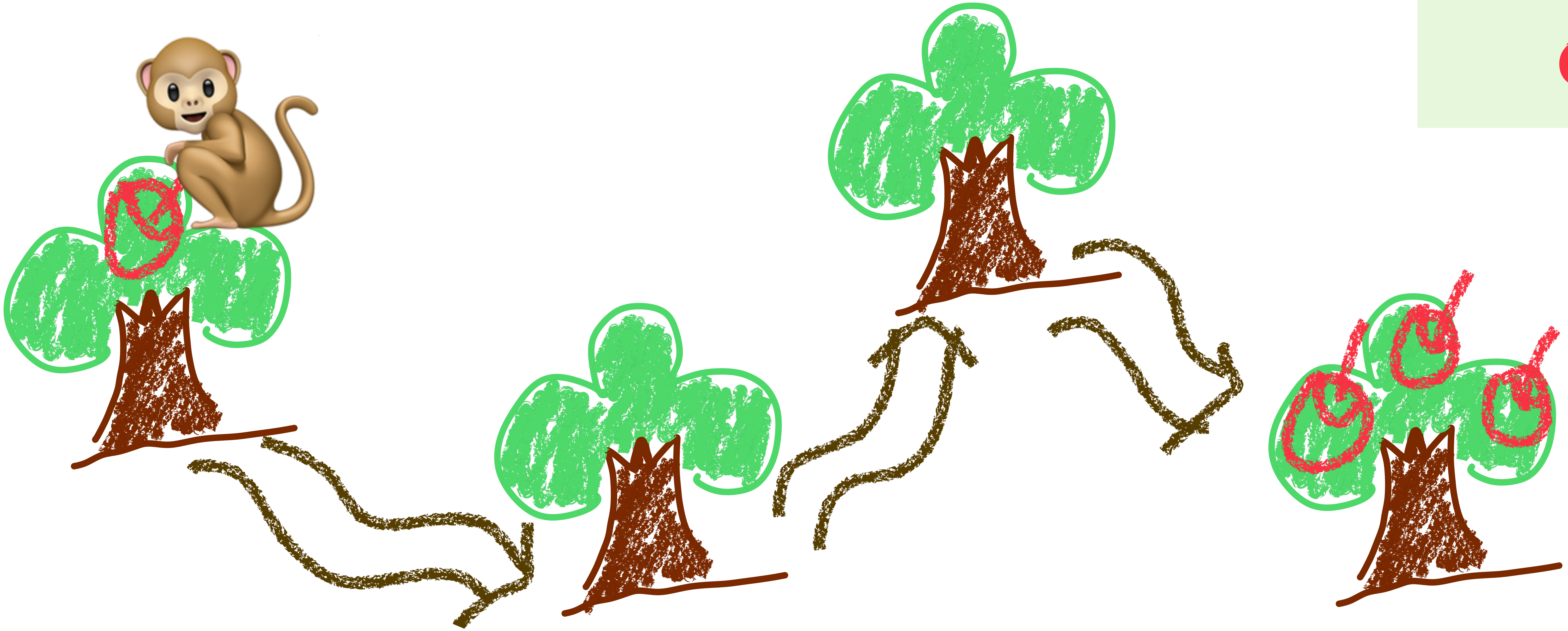
Disconnected Signals



The monkey became hungry on the way to the apple tree $\mathscr{T}^*$, so went back home.

Disconnected Signals





The monkey became hungry on the way to the apple tree $\mathscr{T}^*$, so went back home.

From Root to $\mathscr{T}^*$

Important: $\pi(\mathscr{T} \mid Y) \propto \pi(\mathscr{T})L(Y \mid \mathscr{T})$

$$L(Y \mid \mathscr{T}) = \frac{\exp\left\{-\frac{1}{2}Y'[I - X_{\mathscr{T}}\Sigma_{\mathscr{T}}X_{\mathscr{T}}]'Y\right\}}{(2\pi)^{n/2}(1+n)^{|\mathscr{T}_{ext}|/2}}$$

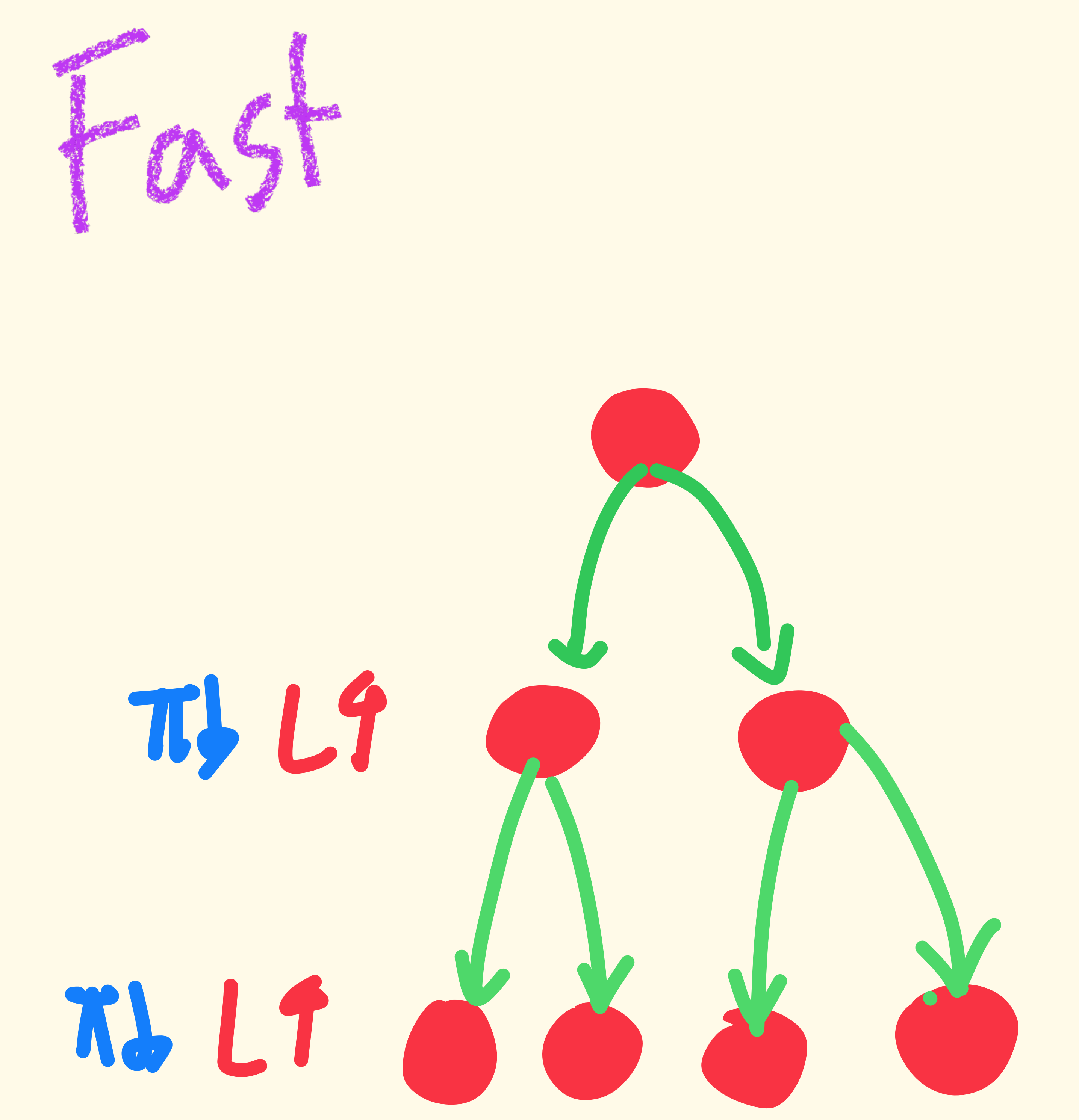
Initialization: $\mathscr{T}_0 = (0,0)$

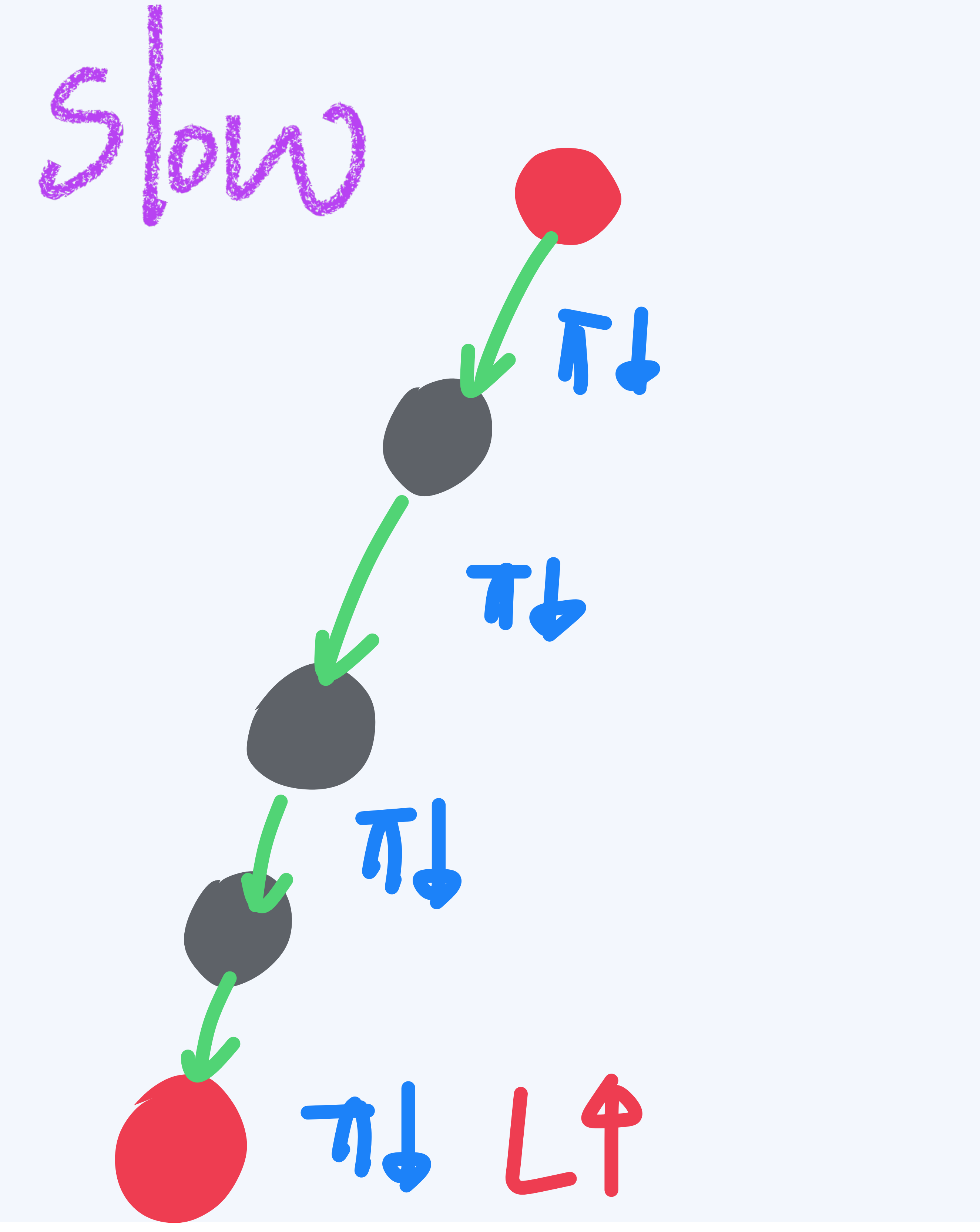
Metropolis-Hastings algorithm

1. Generate $\tilde{\mathscr{T}} \sim S(\cdot \mid \mathscr{T}_{t-1})$

2. Accept $(\mathscr{T}_t = \tilde{\mathscr{T}})$ with

$$\alpha = \min\left\{1, \frac{\pi(\tilde{\mathscr{T}} \mid Y)}{\pi(\mathscr{T}_{t-1} \mid Y)} \frac{S(\mathscr{T}_{t-1} \mid \tilde{\mathscr{T}})}{S(\tilde{\mathscr{T}} \mid \mathscr{T}_{t-1})}\right\}$$





Proof

Spectral Gap

Spectral gap: $Gap(P) = 1 - \lambda_{max}$, where $\lambda_{max} = \max\{\lambda_2, |\lambda_{|\mathbb{T}_L|}|\}$

Sinclair (1992) If the chain $\mathcal{T}_0, \ldots, \mathcal{T}_t$ from P is ergodic,

$$\frac{1}{2}\left(\frac{1}{Gap(P)} - 1\right)\log\left[\frac{1}{2\epsilon}\right] \leq \tau_\epsilon \leq \frac{\log[1/\min_{\mathcal{T}\in\mathbb{T}_L}\Pi(\mathcal{T}\mid Y)] + \log 1/\epsilon}{Gap(P)}.$$

By considering $\tilde{P} = \dfrac{P+I}{2}$, all eigen values made non-negative

1) The chain: ergodic

2) $Gap(\tilde{P}) = 1 - \lambda_2$

$$Gap(\tilde{P}) = 1 - \lambda_2$$

$$\tilde{P} = V\Lambda V^T$$

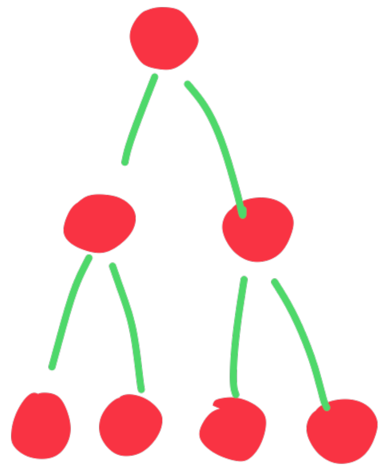
$$\tilde{P}^k = V\Lambda^k V^T \xrightarrow{k\to\infty} \begin{bmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{bmatrix}$$

$$\lambda_1 = 1 > \lambda_2 \geq \lambda_3 \geq \ldots \geq 0$$

$$\lim_{k\to\infty} \Lambda^k = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$\begin{bmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{bmatrix} = vv^T$, where v is the first coloumn of V

Proof of rapid mixing



Sinclair, 1992 (Canonical Path)

For any reversible Markov chain and any choice of $\mathscr{E}$

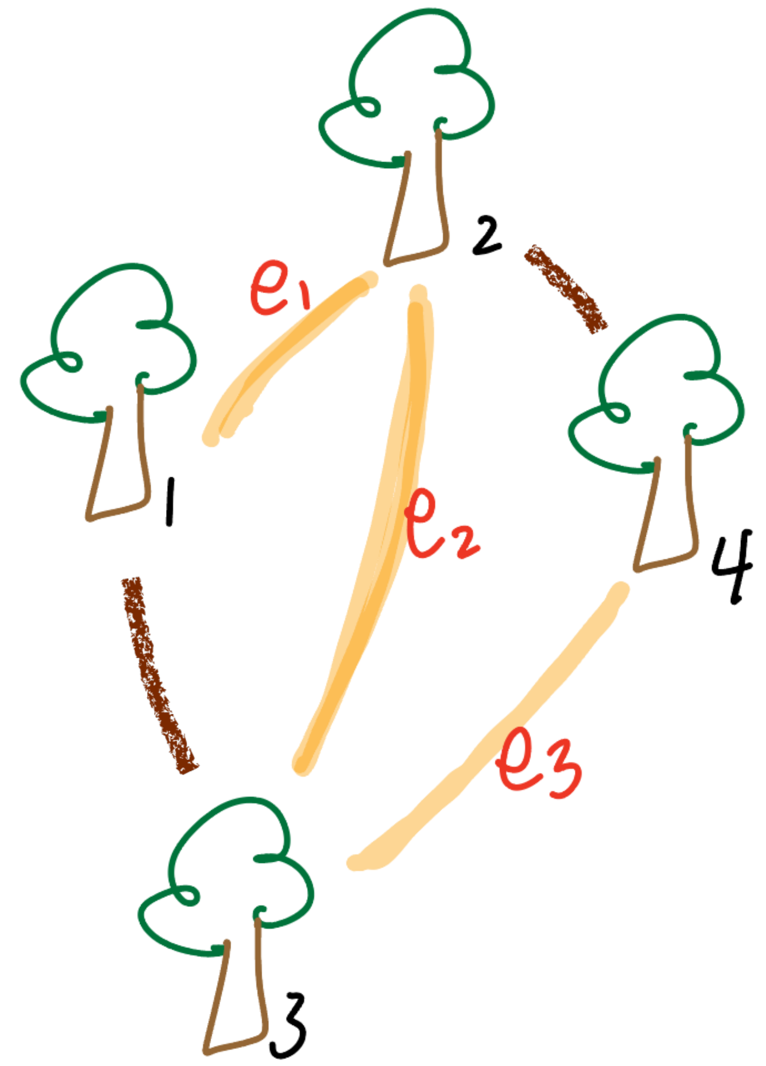
$$Gap(P)^{-1} \leq l(\mathscr{E})\rho(\mathscr{E})$$

$T_{\mathscr{T},\mathscr{T}'}$: Simple path between $(\mathscr{T},\mathscr{T}') \in \mathbb{T} \times \mathbb{T}$

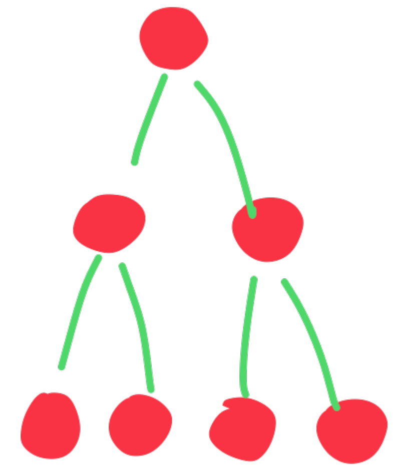
$\mathscr{E}$: Canonical path ensemble:

$$\mathscr{E} = \{T_{\mathscr{T},\mathscr{T}'} : (\mathscr{T},\mathscr{T}') \in \mathbb{T} \times \mathbb{T}\}$$

<Graph (Adjacency: P)>



Proof of rapid mixing



Sinclair, 1992 (Canonical Path)

For any reversible Markov chain and any choice of $\mathscr{E}$

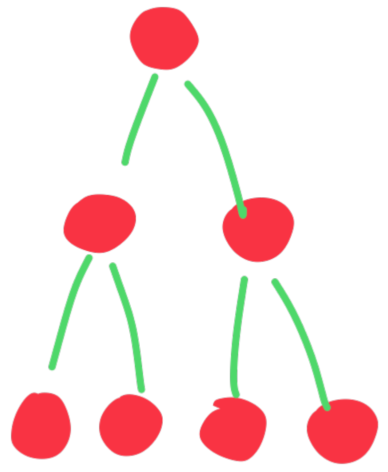
$$Gap(P)^{-1} \le l(\mathscr{E})\rho(\mathscr{E})$$

$l(\mathscr{E})$: the length of the longest path in $\mathscr{E}$

$\rho(\mathscr{E})$: the path congestion parameter of $\mathscr{E}$

< Canonical Path >

Proof of rapid mixing



Sinclair, 1992 (Canonical Path)

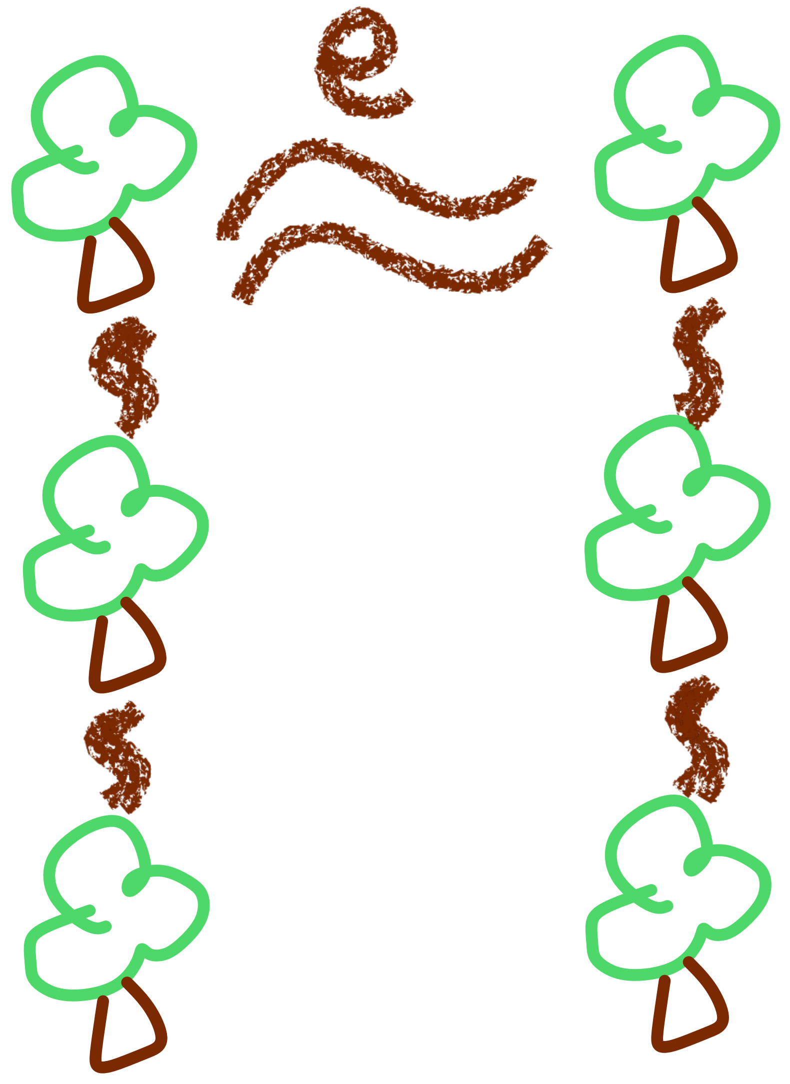
For any reversible Markov chain and any choice of $\mathscr{E}$

$$Gap(P)^{-1} \leq l(\mathscr{E})\rho(\mathscr{E})$$

$l(\mathscr{E})$: the length of the longest path in $\mathscr{E}$

$\rho(\mathscr{E})$: the path congestion parameter of $\mathscr{E}$

< Canonical Path >



Yang et al, 2016 (Variable selection)

$\mathscr{E}: \mathscr{T} \to \mathscr{T}^* \to \mathscr{T}'$

Congestion parameter

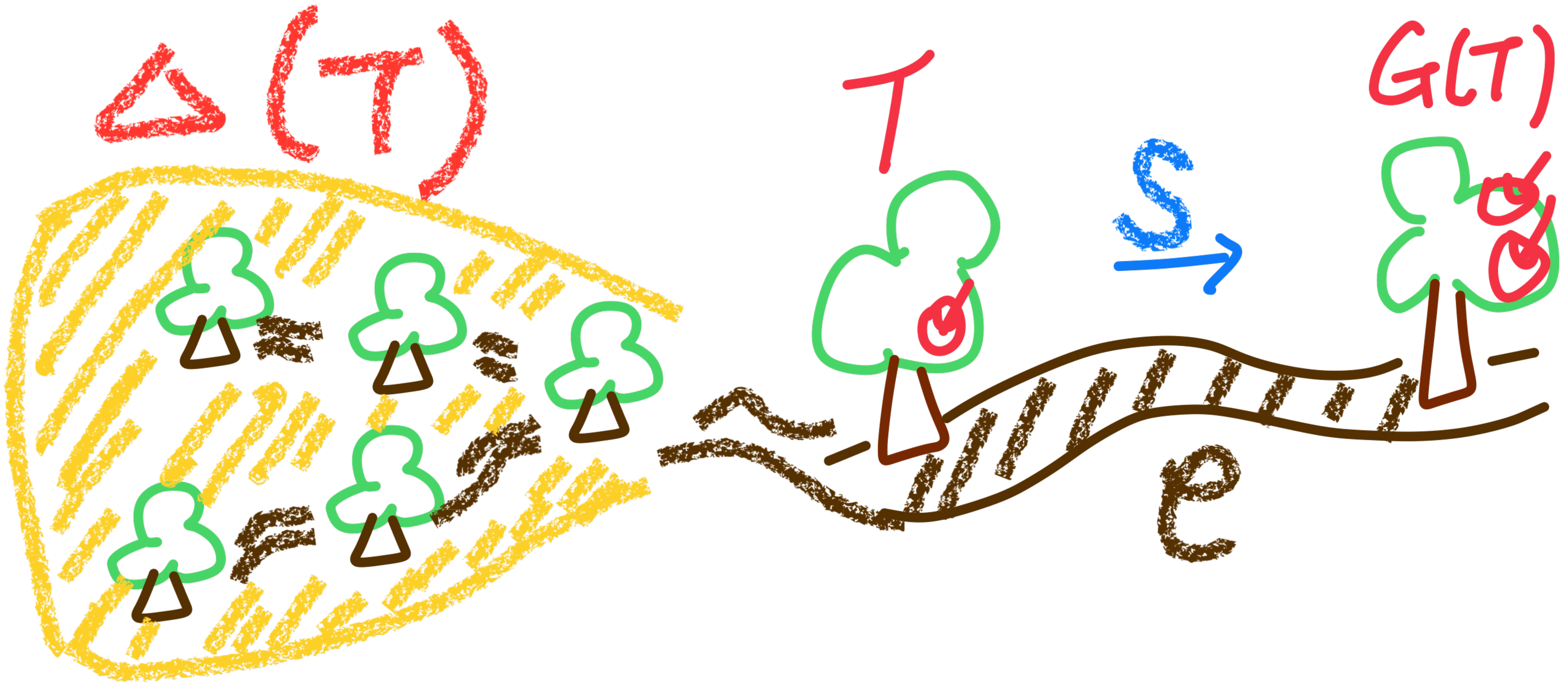
$$\rho(\mathscr{E}) = \max_{e \in \mathscr{E}} \frac{1}{Q(e)} \sum_{(\mathscr{T},\mathscr{T}'):e \in T_{\mathscr{T},\mathscr{T}'}} \Pi(\mathscr{T} \mid Y)\Pi(\mathscr{T}' \mid Y)$$

$Q(e)$: the natural capacity of e

$Q(e) \equiv Q(\mathscr{T},\mathscr{T}') = \Pi(\mathscr{T} \mid Y)P(\mathscr{T},\mathscr{T}')$

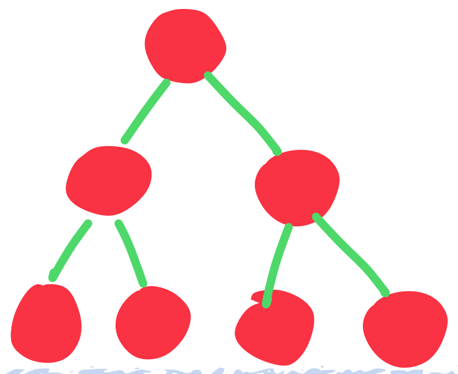
$$\leq 2 \max_{\mathscr{T}} \frac{\pi(\Delta(\mathscr{T}) \mid Y)}{\pi(\mathscr{T} \mid Y)} \frac{1}{S(\mathscr{T} \to G(\mathscr{T}))} \max(1, \alpha^{-1})$$

$G(\mathscr{T})$: the next tree of $\mathscr{T}$ in $\mathscr{E}$



$\longrightarrow \mathscr{T}^*$

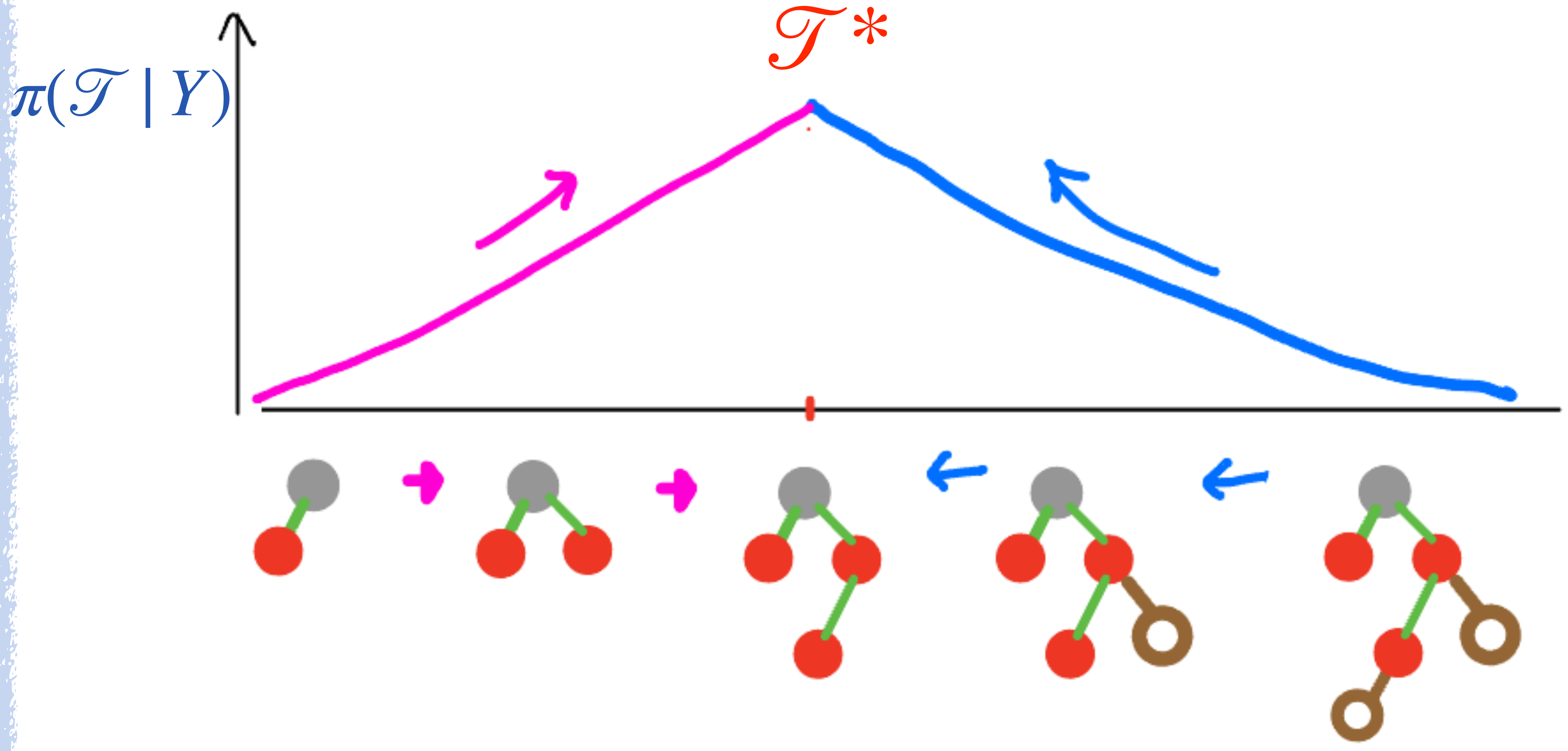
Canonical Path (Sinclair, 1992)



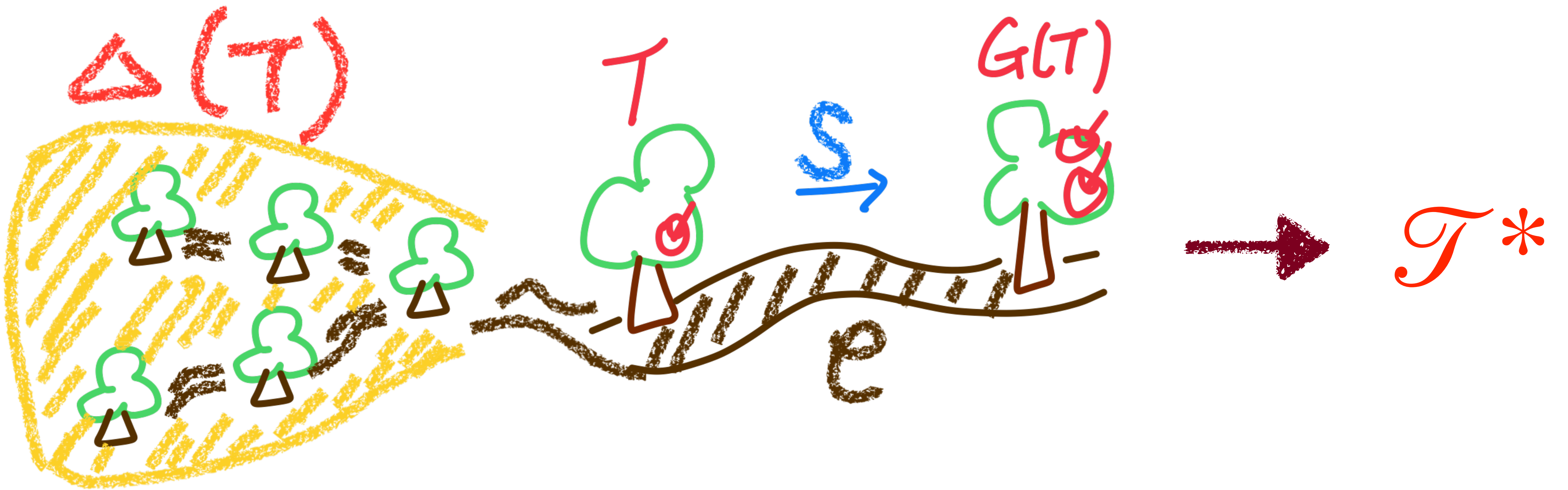
Our Canonical Path

$$\text{Partial Overfit} \xrightarrow{\text{PRUNE}} \text{Underfit} \xrightarrow{\text{GROW}} \mathscr{T}^* \xleftarrow{\text{PRUNE}} \text{Overfit}$$

Posterior Shaping

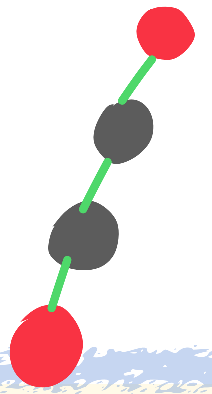


Congestion parameter



$$\rho(\mathscr{E}) \le 2 \max_{\mathscr{T}} \frac{\pi(\Delta(\mathscr{T}) \mid Y)}{\pi(\mathscr{T} \mid Y)} \frac{1}{S(\mathscr{T} \to G(\mathscr{T}))} \max(1, \alpha^{-1})$$

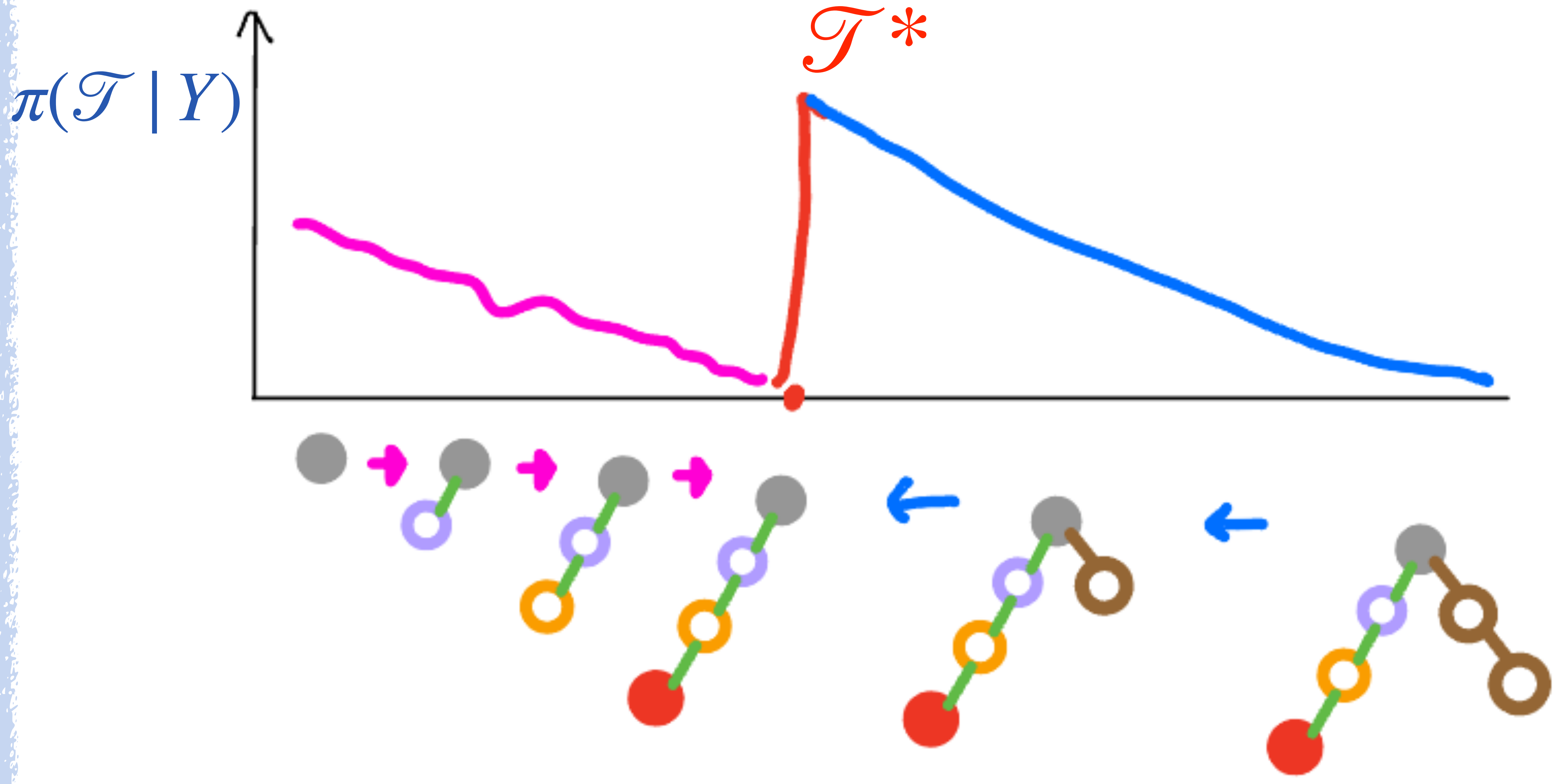
Canonical Path (Sinclair, 1992)



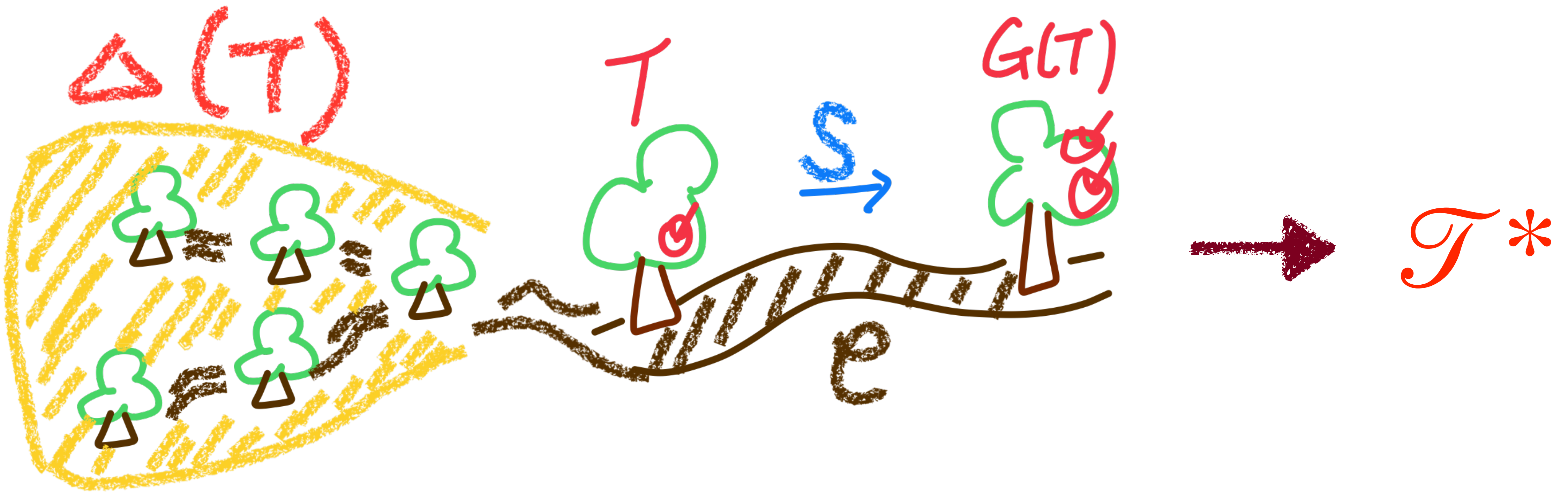
Our Canonical Path

Partial Overfit $\xrightarrow{\text{PRUNE}}$ **Underfit** $\xrightarrow{\text{GROW}}$ $\mathscr{T}^*$ $\xleftarrow{\text{PRUNE}}$ **Overfit**

Posterior Shaping



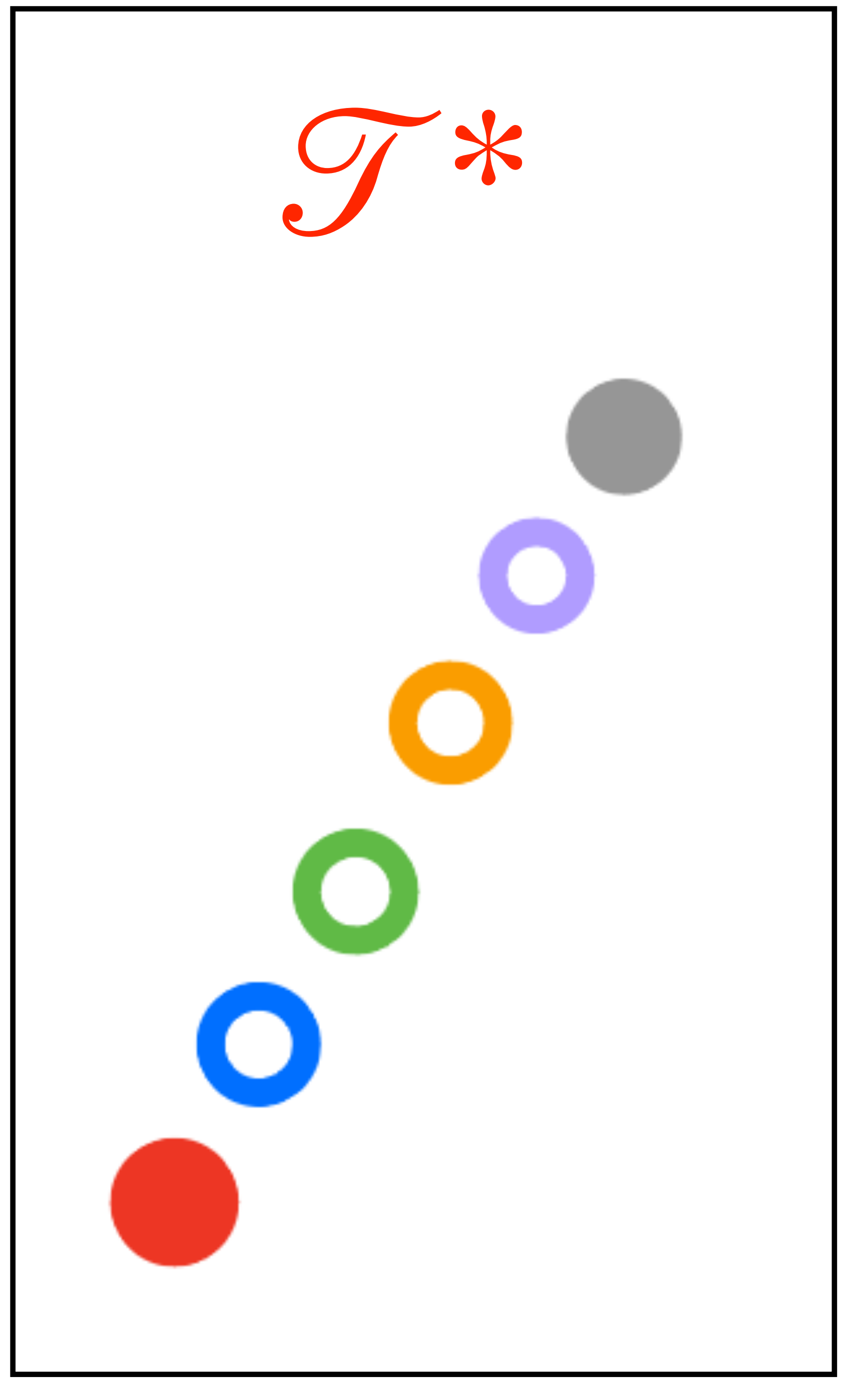
Congestion parameter

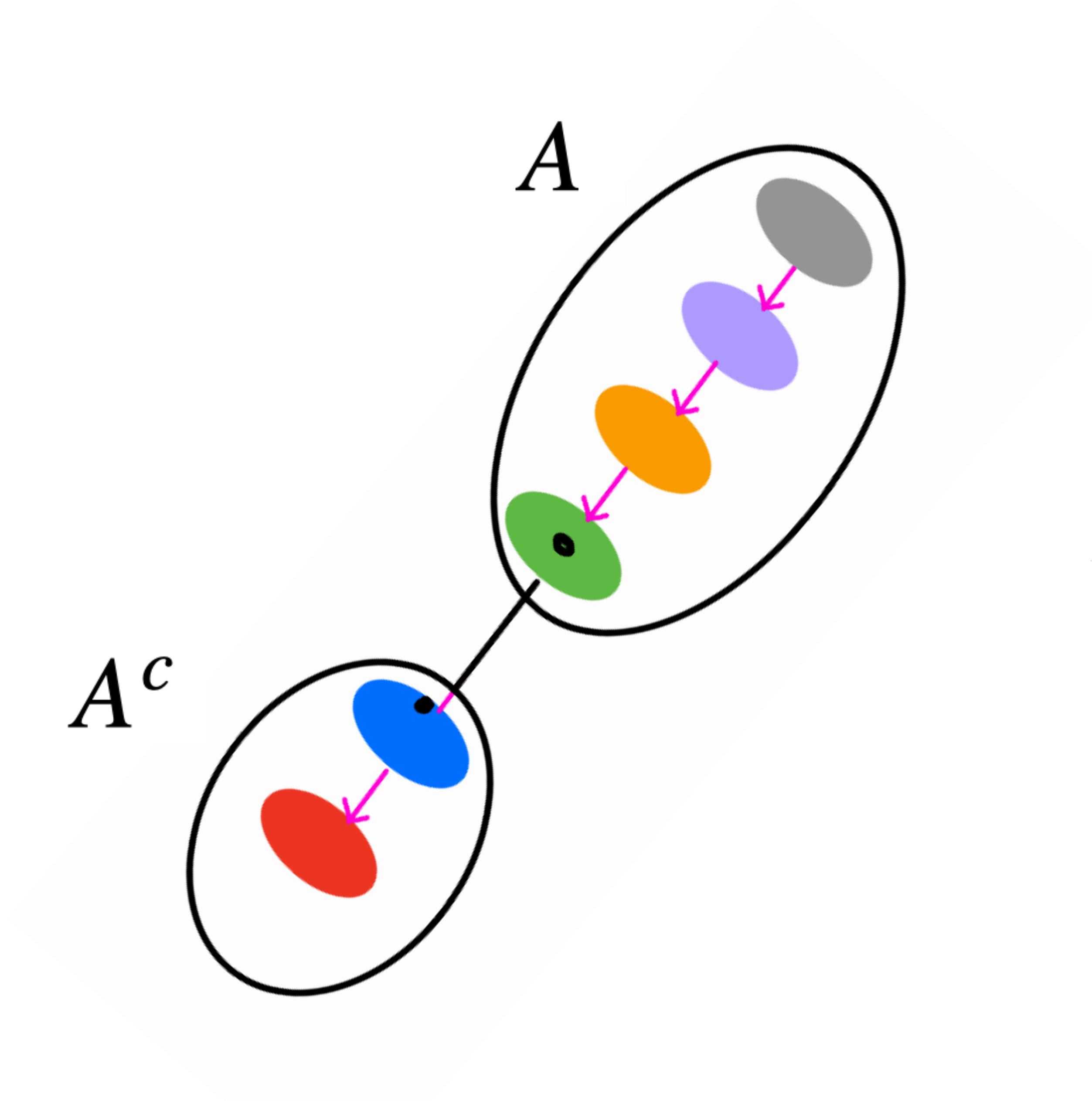


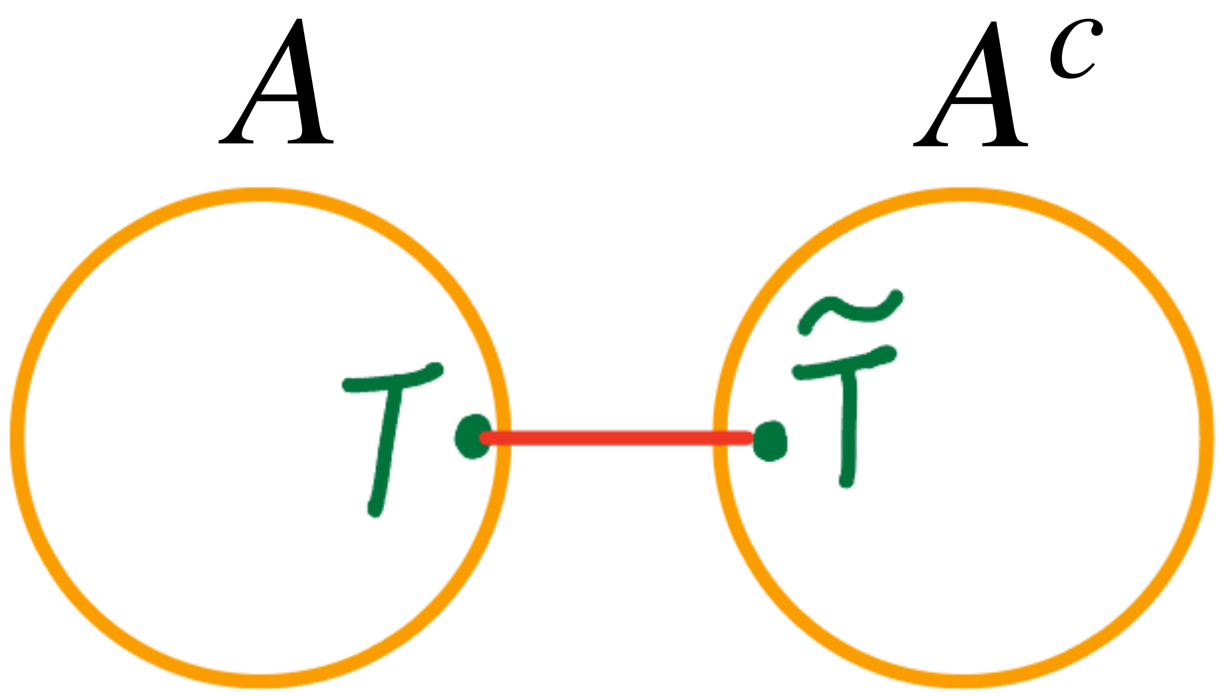
$$\rho(\mathscr{E}) \leq 2 \max_{\mathscr{T}} \frac{\pi(\Delta(\mathscr{T}) \mid Y)}{\pi(\mathscr{T} \mid Y)} \frac{1}{S(\mathscr{T} \rightarrow G(\mathscr{T}))} \max(1, \alpha^{-1})$$

Bottleneck ratio bound (Sinclair 1992)

Lower Bound





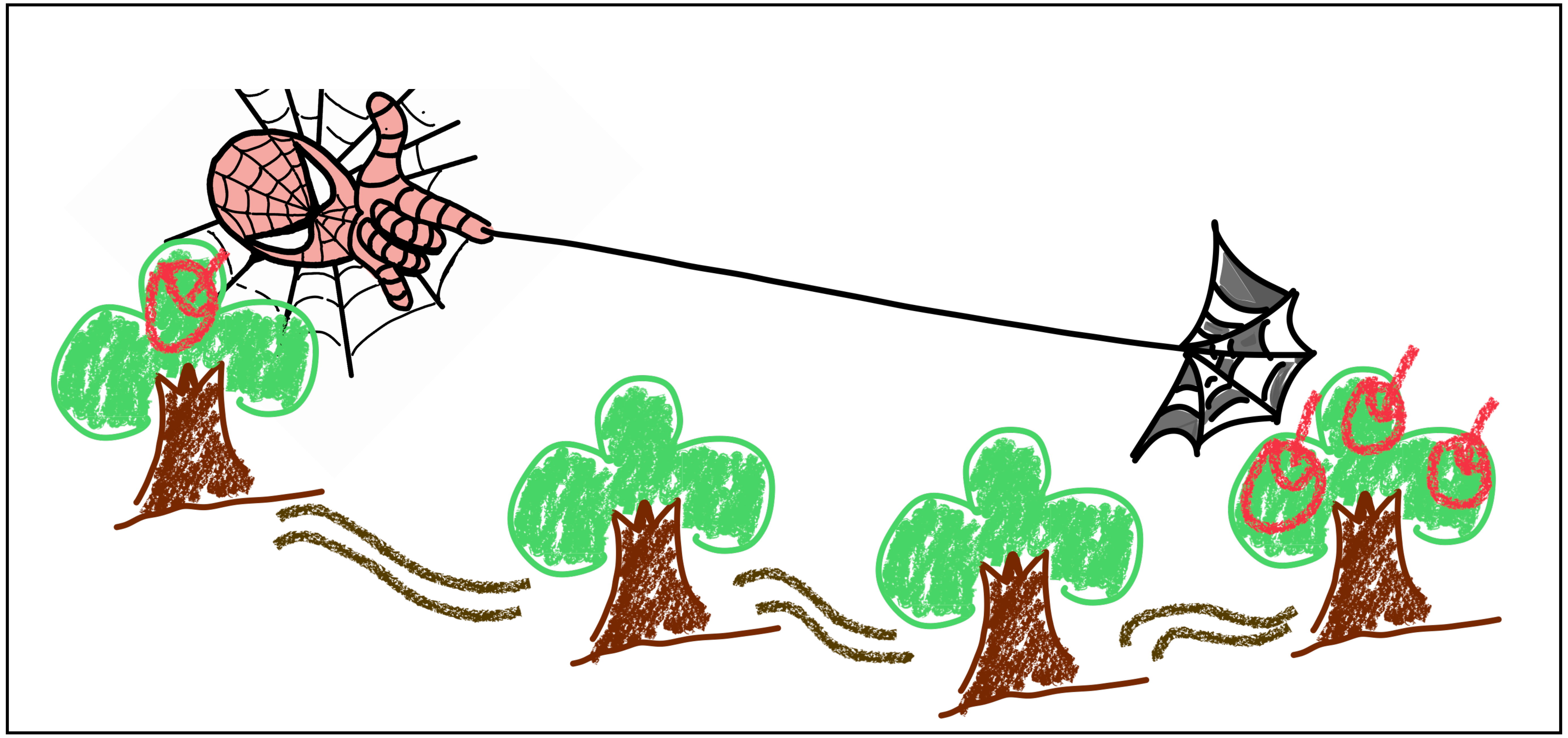


$$\Phi^2/2 \leq Gap(P) \leq 2\Phi$$

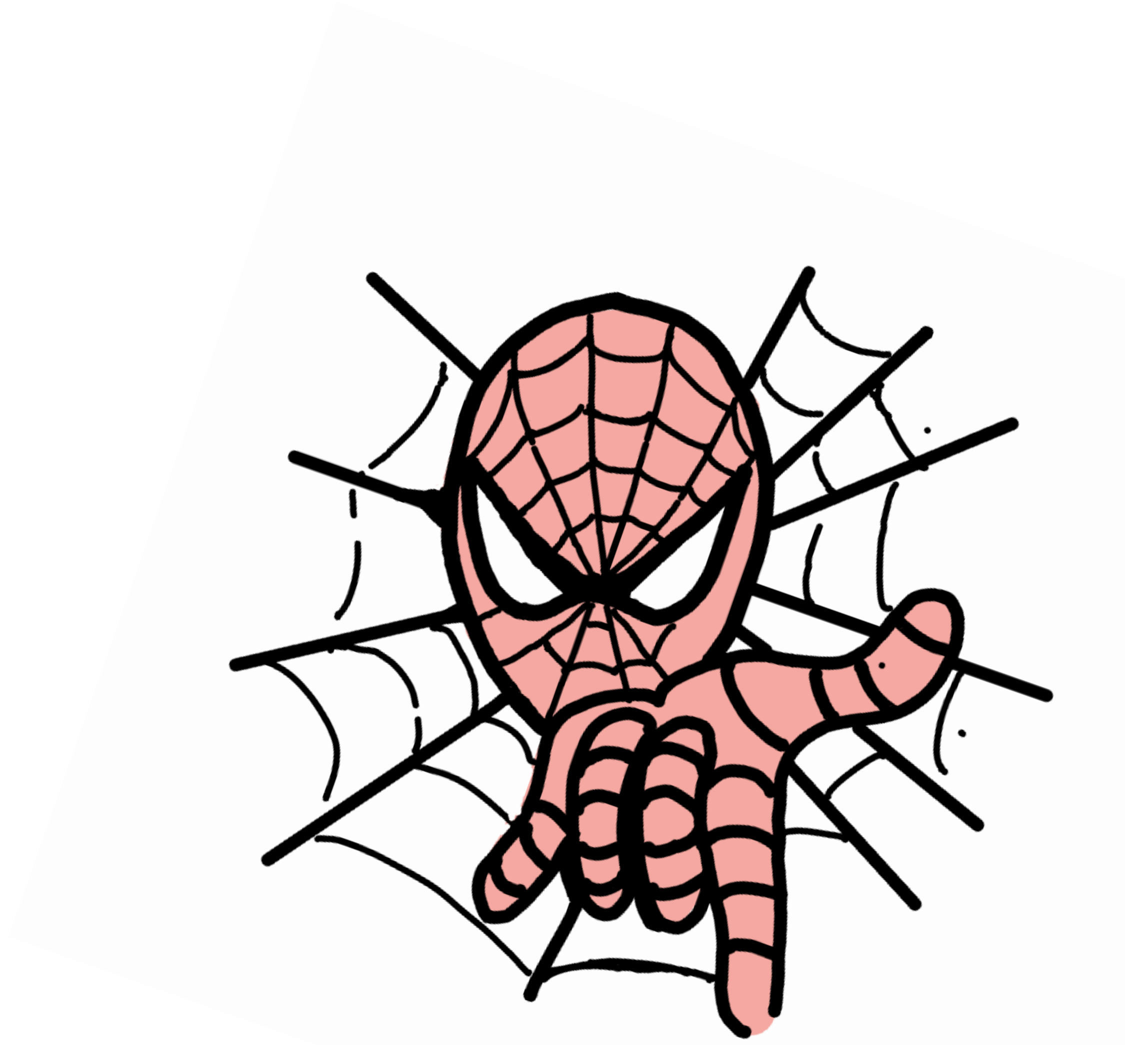
Ability to escape from any small region of the state space

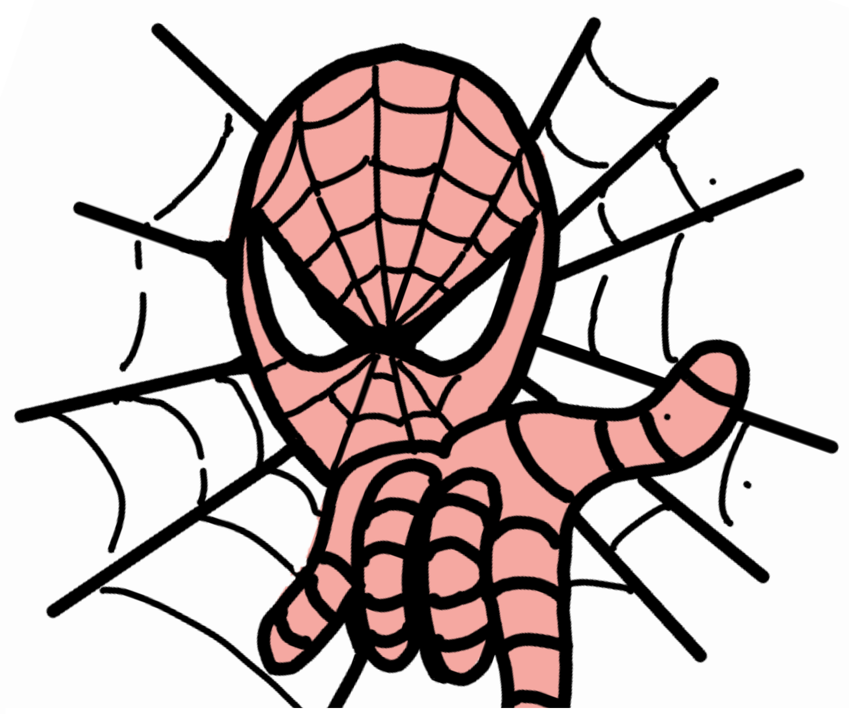
$$\Phi = \min_{0<\Pi[A|Y]\leq 1/2} \frac{\sum_{\mathscr{T}\in A, \mathscr{T}'\in \mathbb{T}\setminus A} \Pi(\mathscr{T} \mid Y) P(\mathscr{T}, \mathscr{T}')}{\Pi(A \mid Y)}$$

Twiggy CART

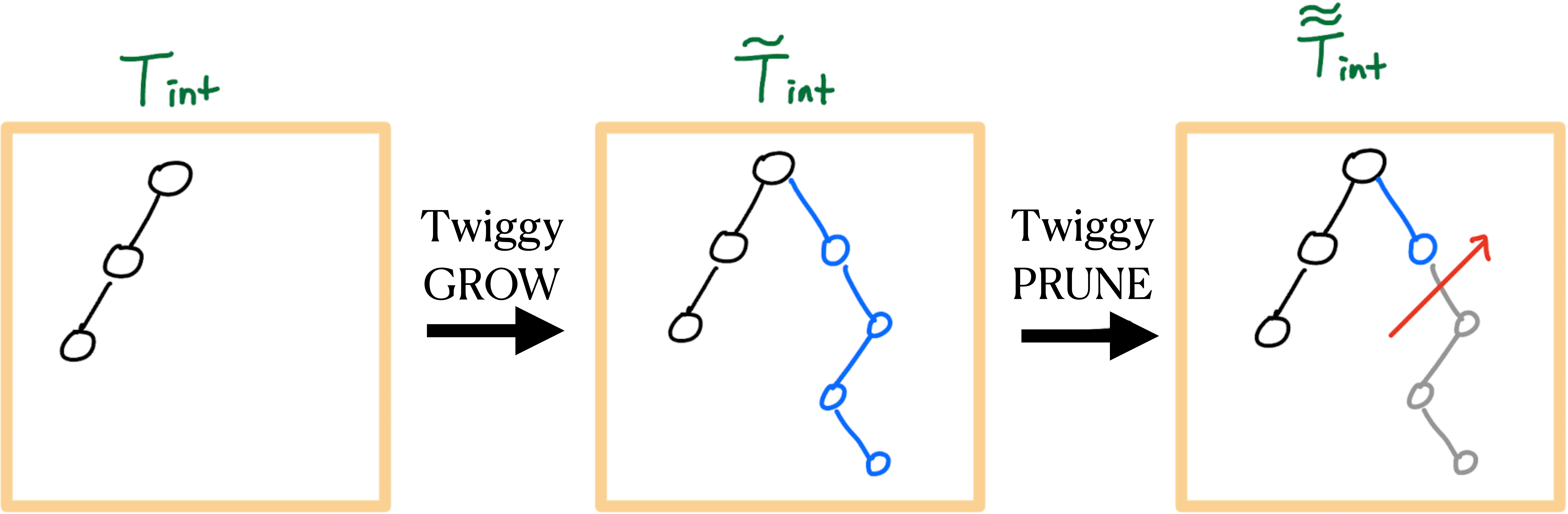


The Spiderman directly flew to the apple tree $\mathscr{T}^*$,
and ate instant rewarding apples.



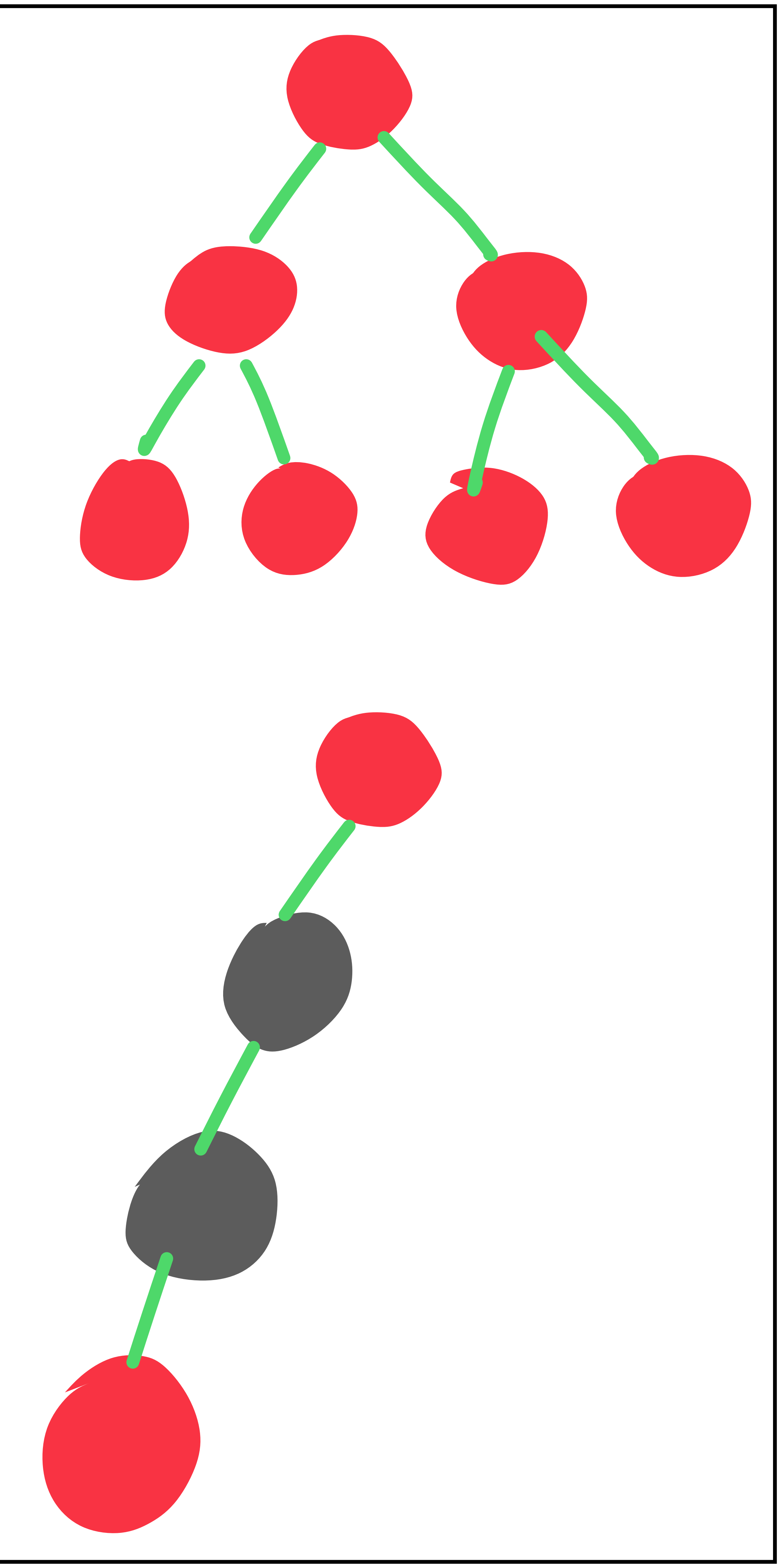


New movement: Twiggy GROW and PRUNE



$$S(\cdot \mid \mathscr{T}_t) = \frac{1}{2} \sum_{m \in M} 1_m S_m(\cdot \mid \mathscr{T}_t), \text{ where } M = \{\text{twiggy grow, twiggy prune}\}$$

Both type Signals: Rapid Mixing



Theorem 3 (Twiggy Bayesian CART algorithm)

Assume the connected signals or disconnected signals.

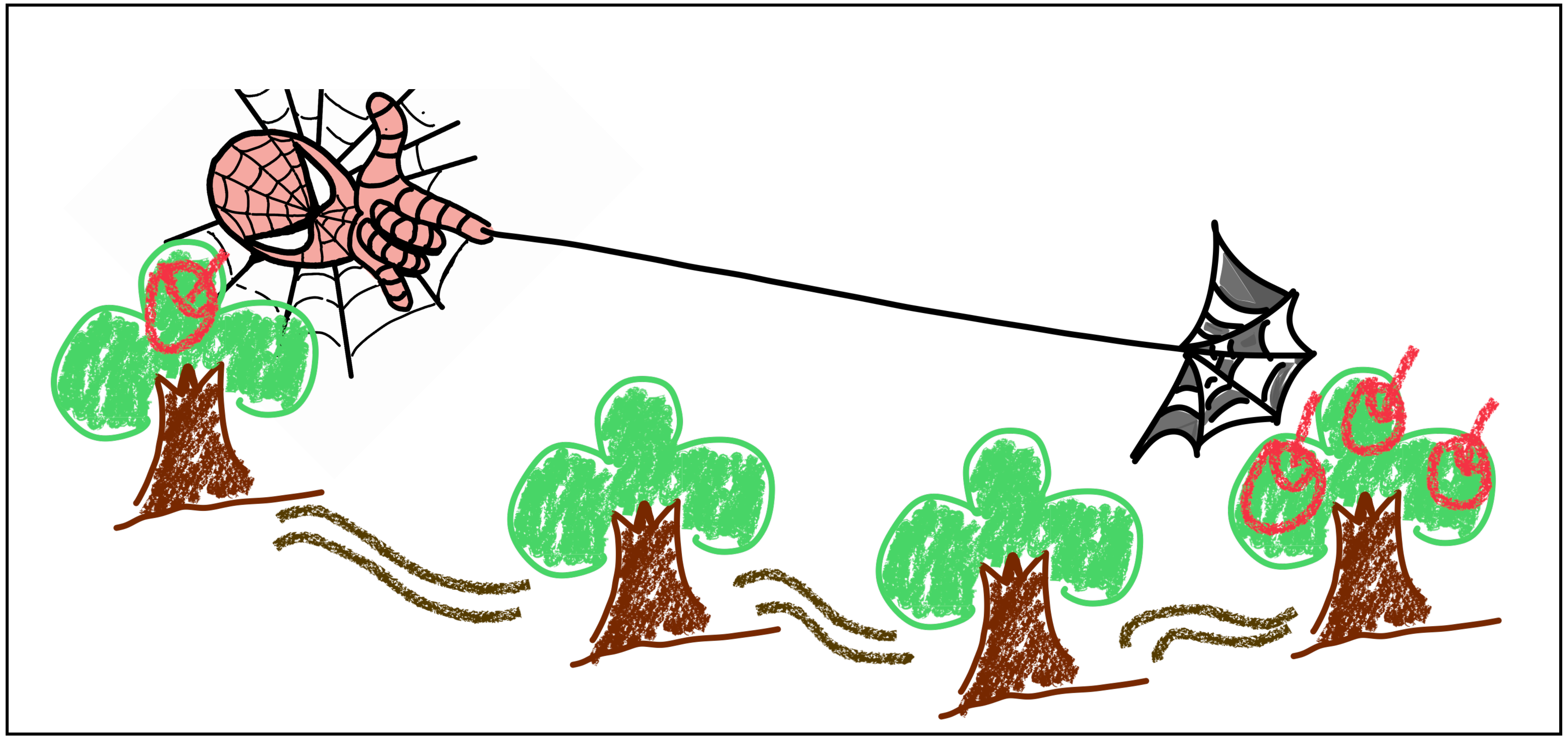
With $|\mathscr{T}^*_{int}| \lesssim \log^2 n$,

$$\tau_\epsilon \le 2^{2L+3}\left\{n\left[\left(c+\frac{1}{2}\right)\log(1+n)+|\mathscr{T}^*_{int}|C_{f_0}^2+1\right] + 4|\mathscr{T}^*_{int}|\log n+\log\left(\frac{2}{\epsilon}\right)\right\}$$

with probability at least $1-4/n$.

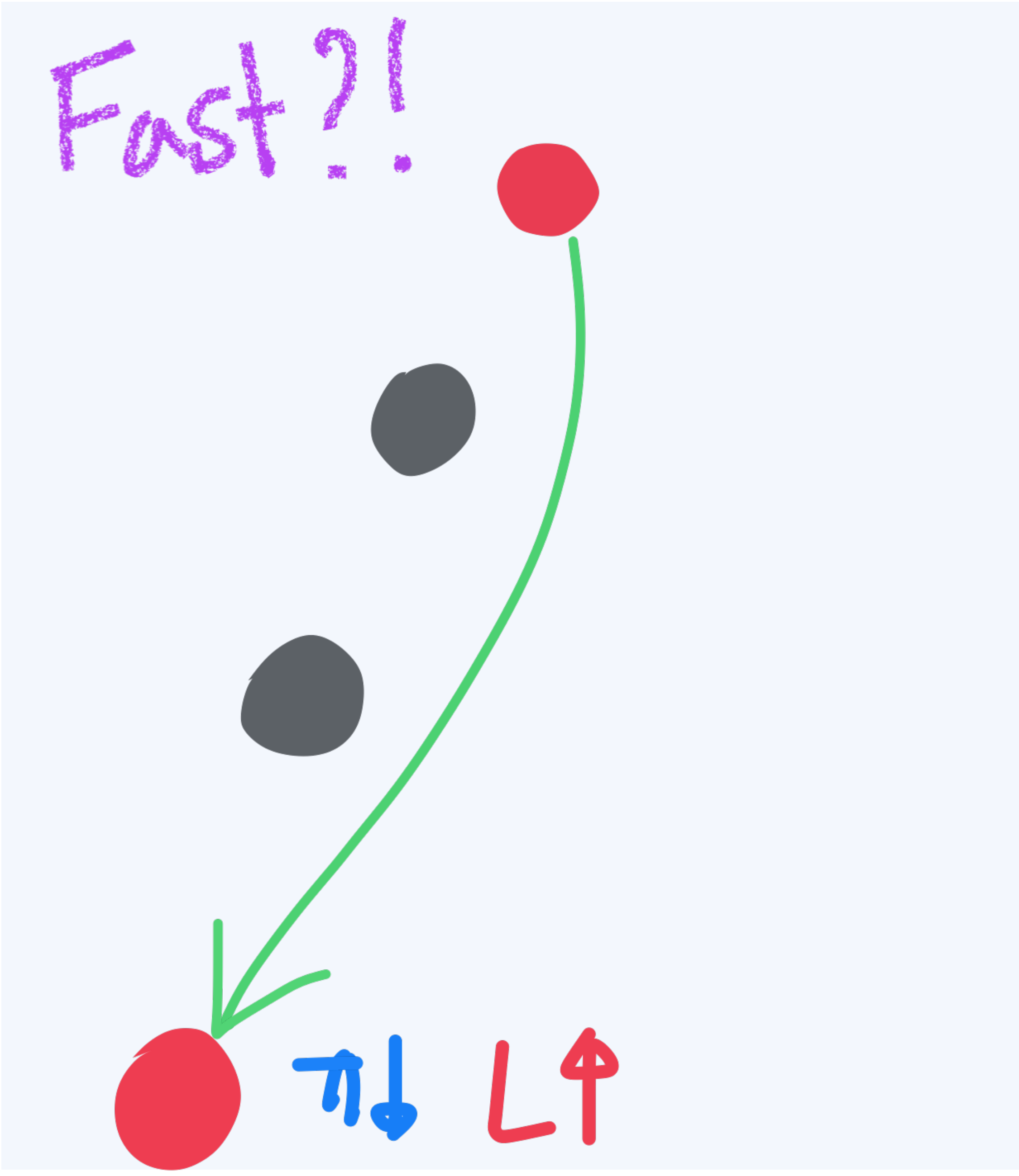
$p = 2^L$: the number of problem parameters $(2^L \le n/2)$

C_{f_0}: the bound for the maximum signal strength



The Spiderman directly flew to the apple tree $\mathscr{T}^*$,
and ate instant rewarding apples.

Bottleneck Broken

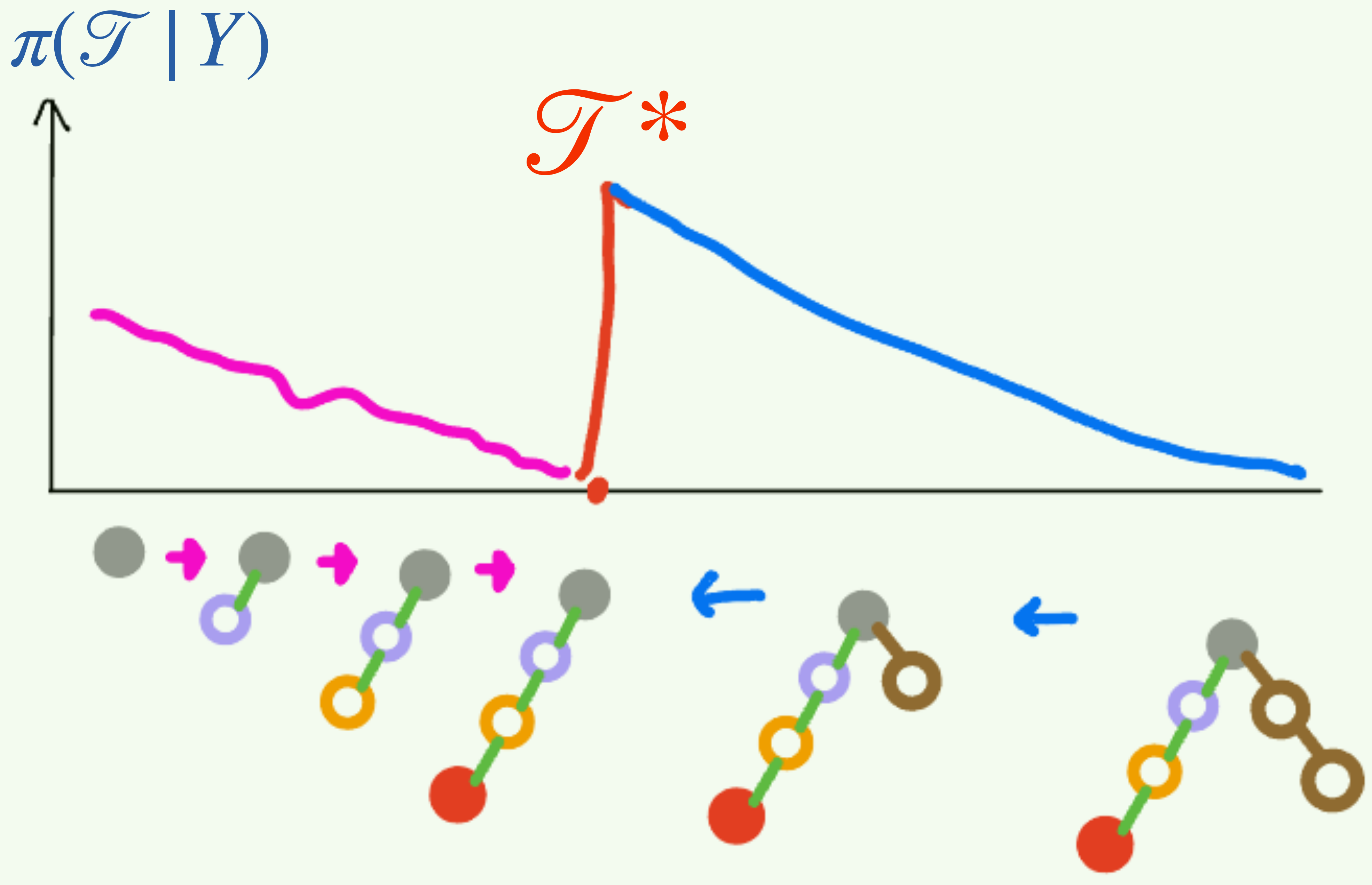


$$\pi(\mathscr{T} \mid Y) \propto \pi(\mathscr{T})L(\mathscr{T} \mid Y)$$

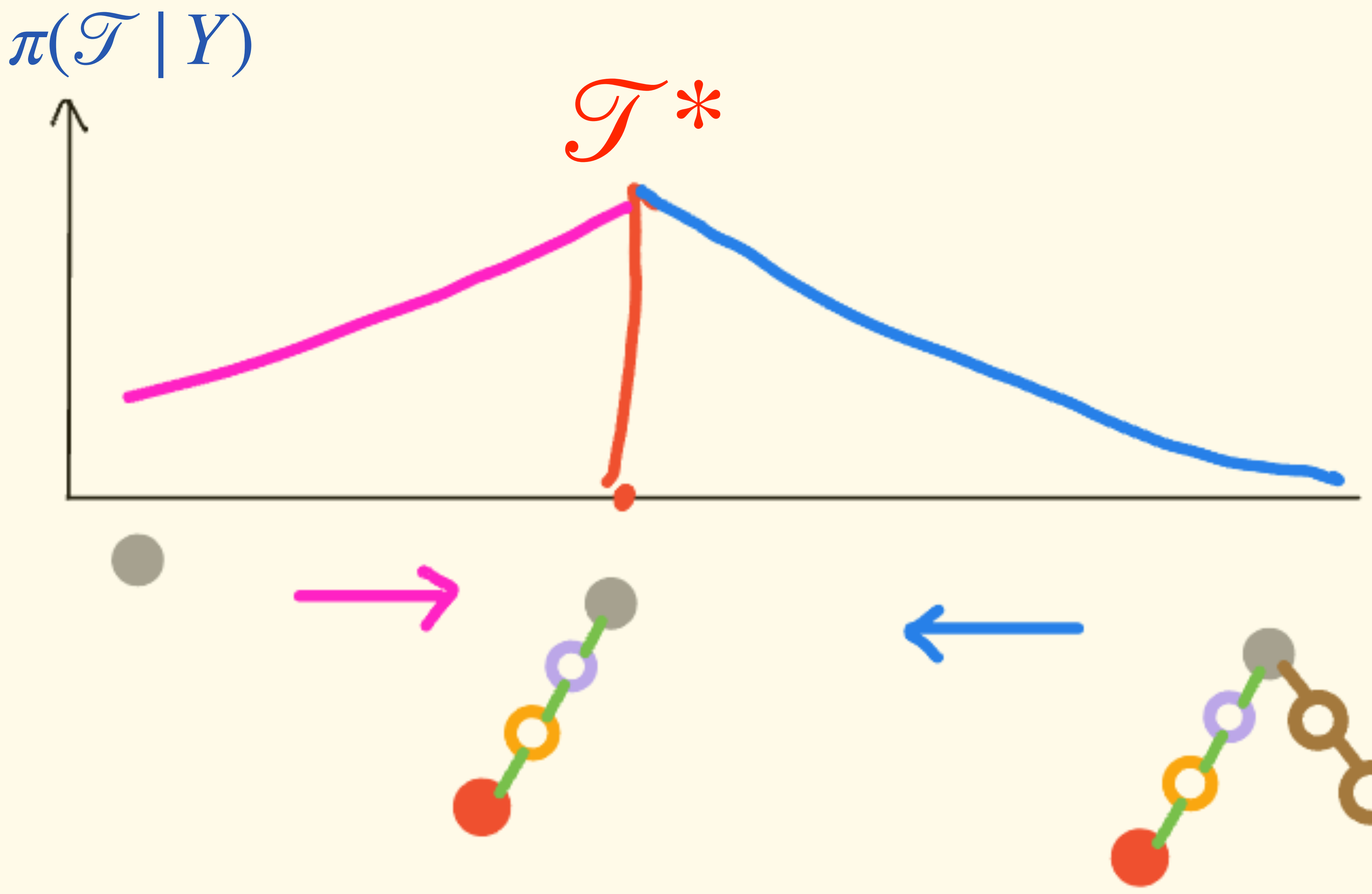
$$\alpha = \min \left\{ 1, \frac{\pi(\tilde{\mathscr{T}} \mid Y)}{\pi(\mathscr{T}_{t-1} \mid Y)} \frac{S(\mathscr{T}_{t-1} \mid \tilde{\mathscr{T}})}{S(\tilde{\mathscr{T}} \mid \mathscr{T}_{t-1})} \right\}$$

Twiggy effect

Canonical path
broken by B-CART moves

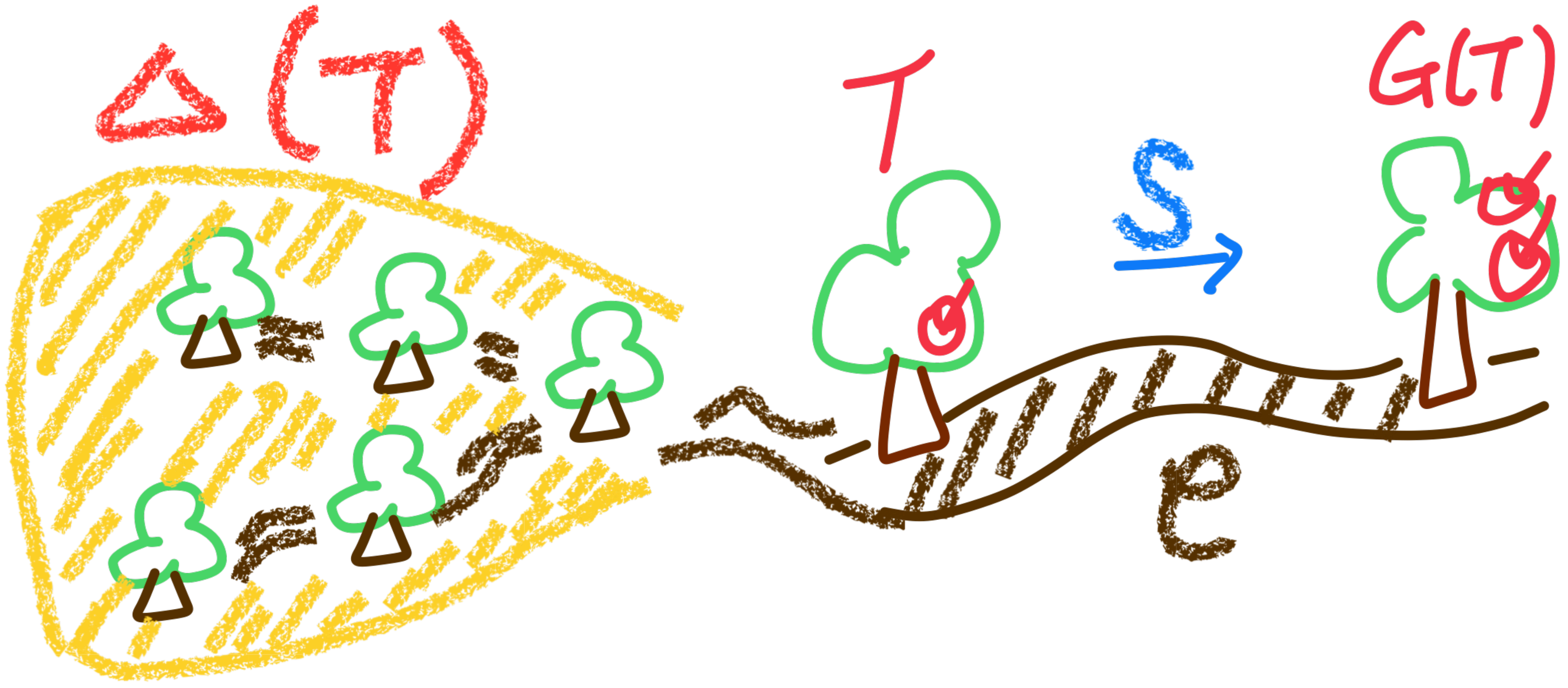


Canonical path
recovered by Twiggy moves



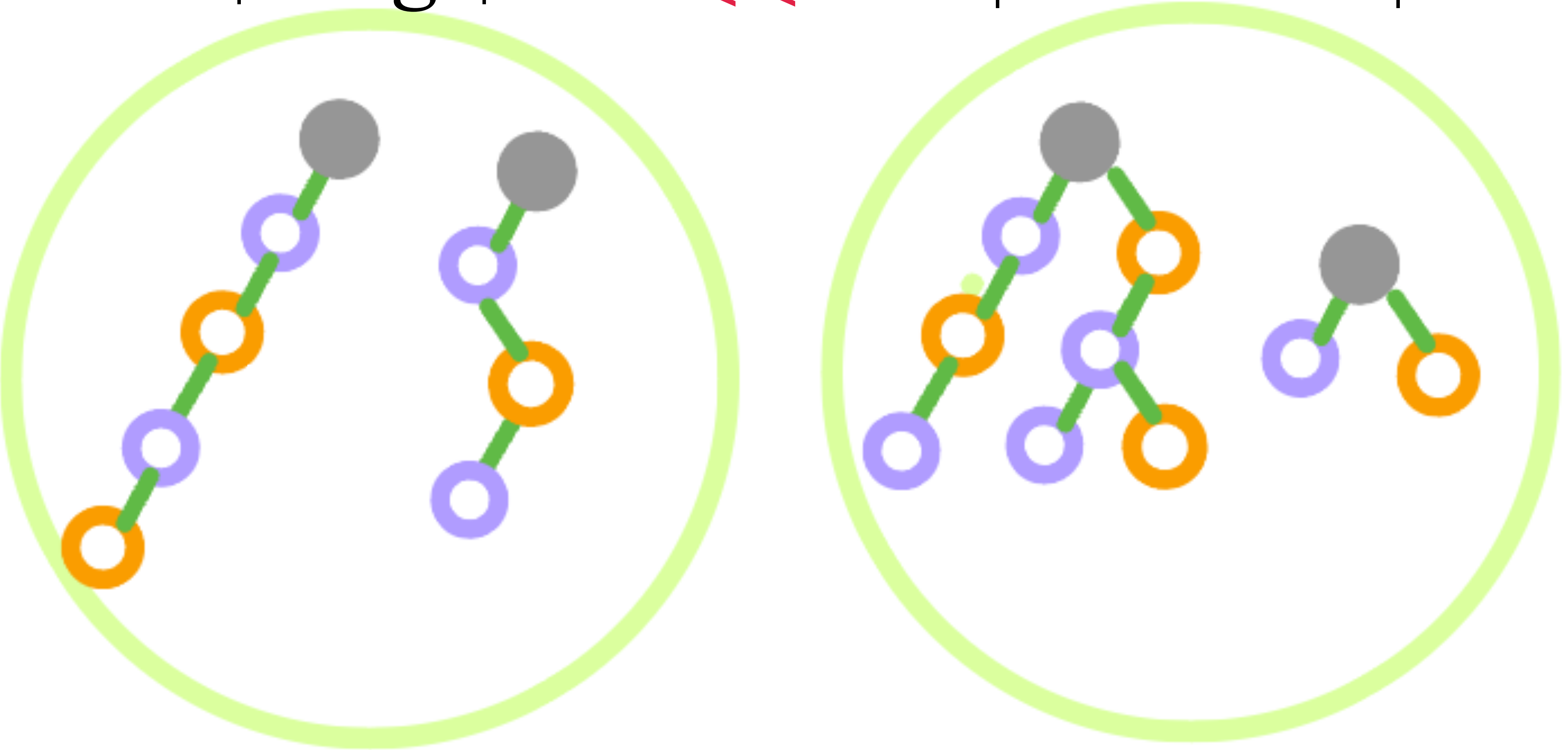
Why not small mini-trees?

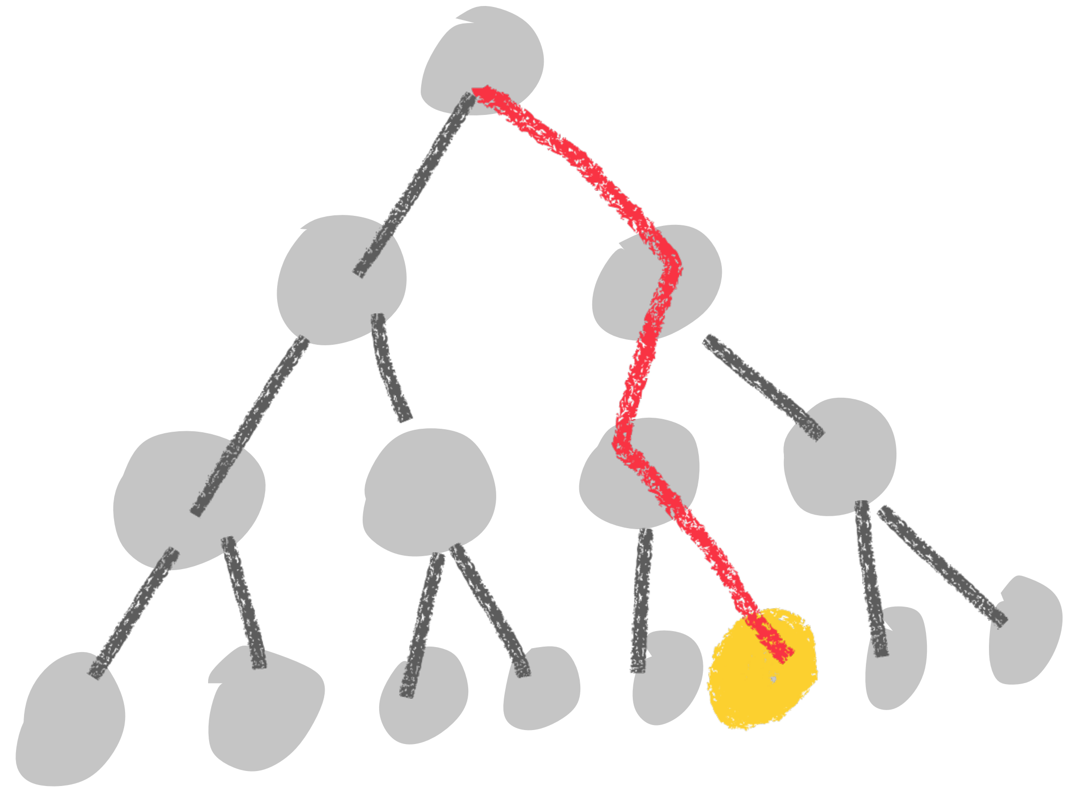
Congestion parameter



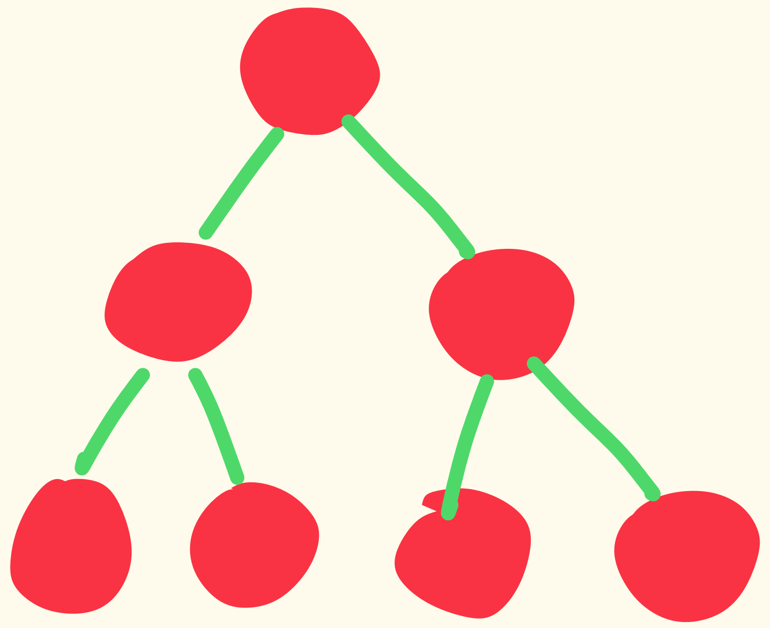
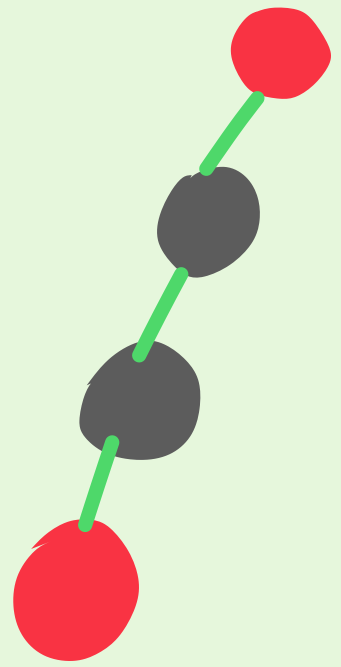
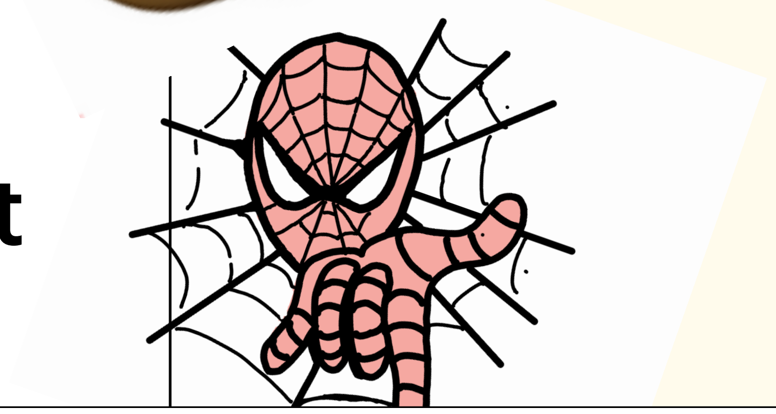
$$\rho(\mathscr{E}) \leq 2 \max_{\mathscr{T}} \frac{\pi(\Delta(\mathscr{T}) \mid Y)}{\pi(\mathscr{T} \mid Y)} \boxed{\frac{1}{S(\mathscr{T} \to G(\mathscr{T}))}} \max(1, \alpha^{-1})$$

|Twigs| << |Mini-trees|

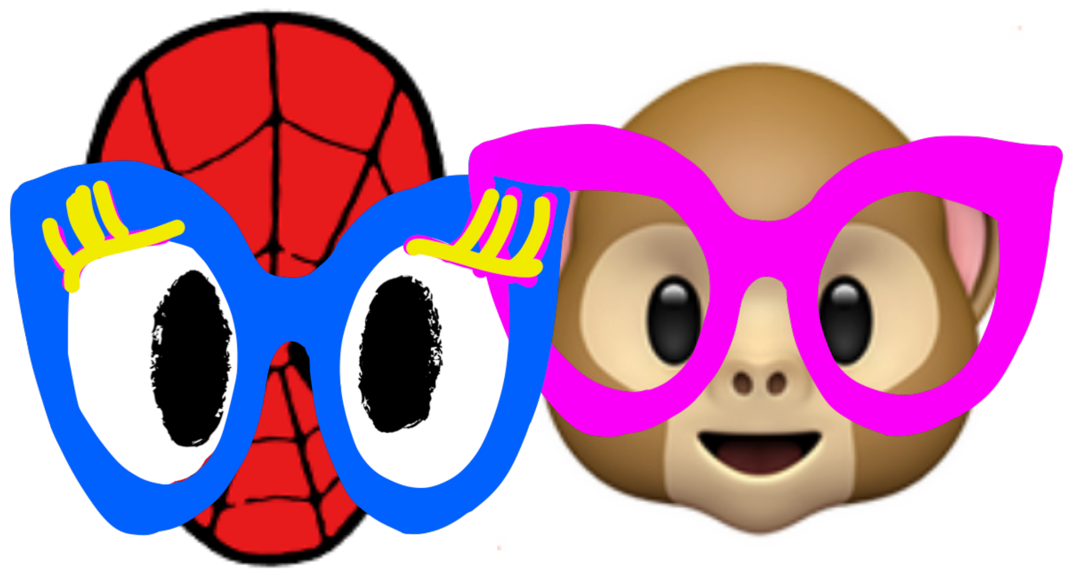




Our Contribution

	Type A 	Type A^c 
Original Moven 	Rapid	Slow
New Movement 	Rapid	Rapid

Informed MCMC



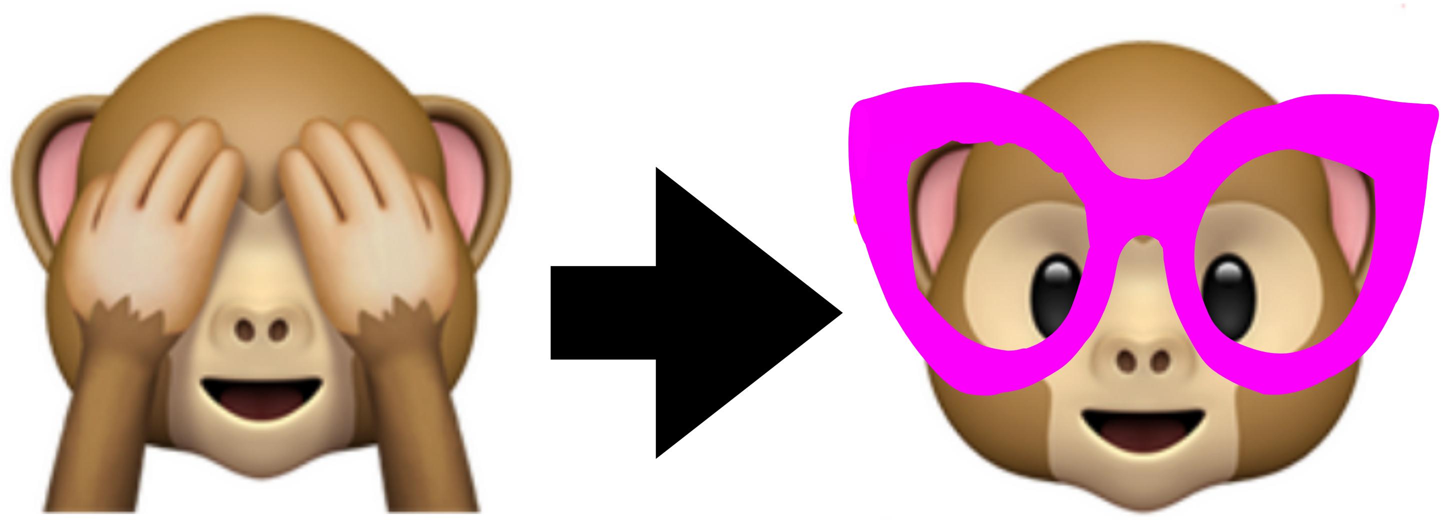
Informed Proposal

Zanella et al (2020), Zhou et al (2022)

$$w_G(\tilde{\mathscr{T}} \mid \mathscr{T}) = \frac{\pi(\tilde{\mathscr{T}} \mid Y)}{\pi(\mathscr{T} \mid Y)} \wedge \delta_G, \qquad Z_G(\mathscr{T}) = \sum_{\mathscr{T} \in \mathscr{N}_G(\mathscr{T})} w_G(\tilde{\mathscr{T}} \mid \mathscr{T})$$

$$w_P(\tilde{\mathscr{T}} \mid \mathscr{T}) = 1 \vee \frac{\pi(\tilde{\mathscr{T}} \mid Y)}{\pi(\mathscr{T} \mid Y)} \wedge \delta_P, \qquad Z_P(\mathscr{T}) = \sum_{\mathscr{T} \in \mathscr{N}_P(\mathscr{T})} w_P(\tilde{\mathscr{T}} \mid \mathscr{T})$$

$$S(\mathscr{T} \rightarrow \cdot\,) = \frac{1}{2} \sum_{m \in M} 1_{N_M(\mathscr{T})} \frac{w_m(\,\cdot \mid \mathscr{T})}{Z_m(\mathscr{T})}, \text{ where } M = \{\mathrm{G}, \mathrm{P}\}$$



Informed Proposal

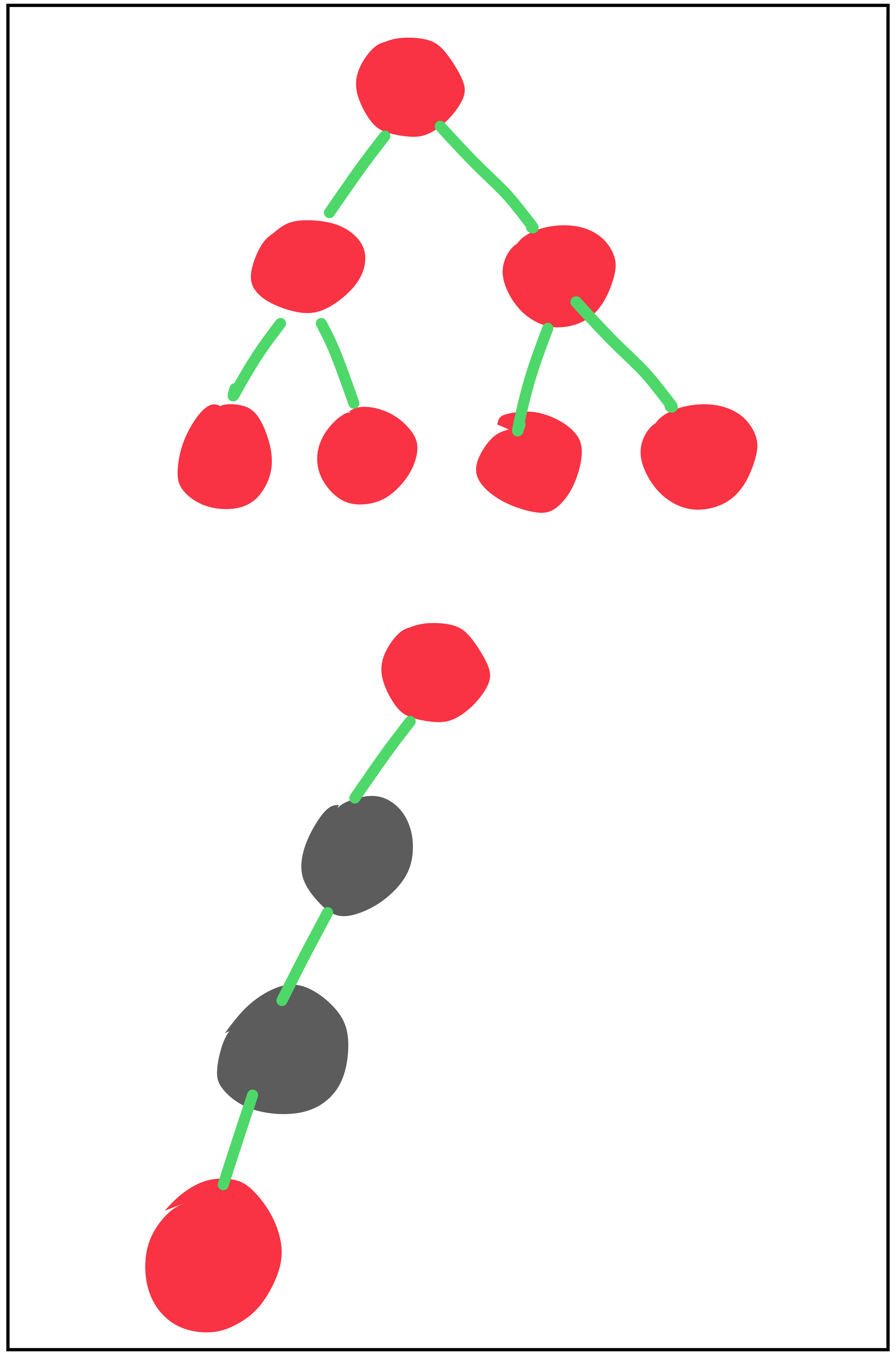
Zanella et al (2020), Zhou et al (2022)

$$w_G(\tilde{\mathcal{T}} \mid \mathcal{T}) = \frac{\pi(\tilde{\mathcal{T}} \mid Y)}{\pi(\mathcal{T} \mid Y)} \wedge \delta_G, \qquad Z_G(\mathcal{T}) = \sum_{\mathcal{T} \in \mathcal{N}_G(\mathcal{T})} w_G(\tilde{\mathcal{T}} \mid \mathcal{T})$$

$$w_P(\tilde{\mathcal{T}} \mid \mathcal{T}) = 1 \vee \frac{\pi(\tilde{\mathcal{T}} \mid Y)}{\pi(\mathcal{T} \mid Y)} \wedge \delta_P, \qquad Z_P(\mathcal{T}) = \sum_{\mathcal{T} \in \mathcal{N}_P(\mathcal{T})} w_P(\tilde{\mathcal{T}} \mid \mathcal{T})$$

$$P(\tilde{\mathcal{T}} \mid \mathcal{T}) = \min\left\{ S(\mathcal{T} \to \tilde{\mathcal{T}}), \frac{\pi(\tilde{\mathcal{T}} \mid Y)}{\pi(\mathcal{T} \mid Y)} S(\tilde{\mathcal{T}} \to \mathcal{T}) \right\}$$

Improved Mixing rate



Theorem 3 (Informed Twiggy Bayesian CART)

Assume the dis/connected signals.

$$\tau_\epsilon \lesssim \log(6/\epsilon) \max\left(\frac{9\, C_{f_0}^2}{A^2}\frac{2^L n}{\log^2 n}, 2^{L+5}\right),$$

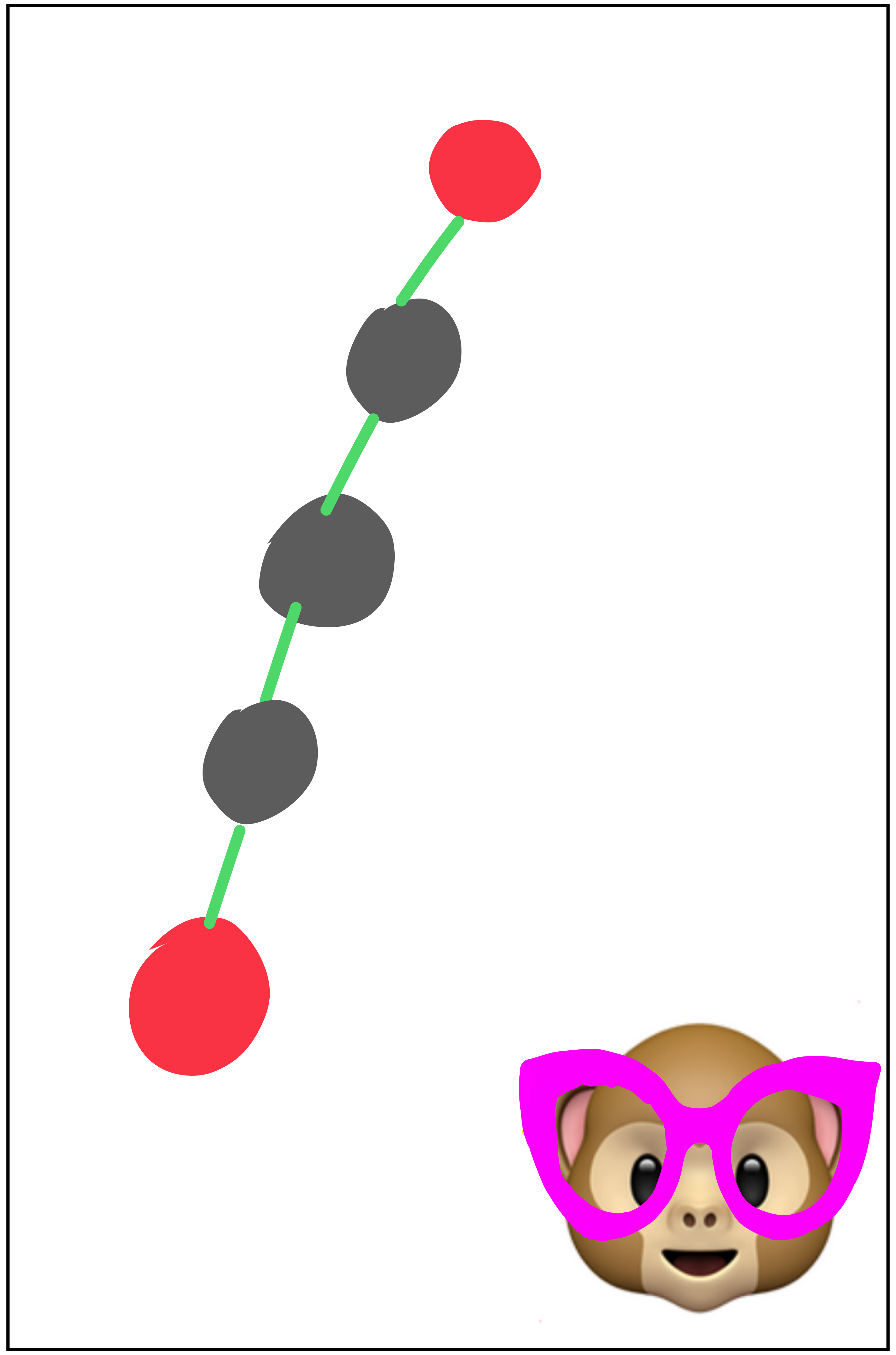
with probability at least $1 - 4/n$.

$p = 2^L$: the number of problem parameters $(2^L \leq n/2)$

Similar result on the informed Bayesian CART,

on connected signals.

Disconnected Signals: Slow Mixing



(Informed Bayesian CART algorithm)

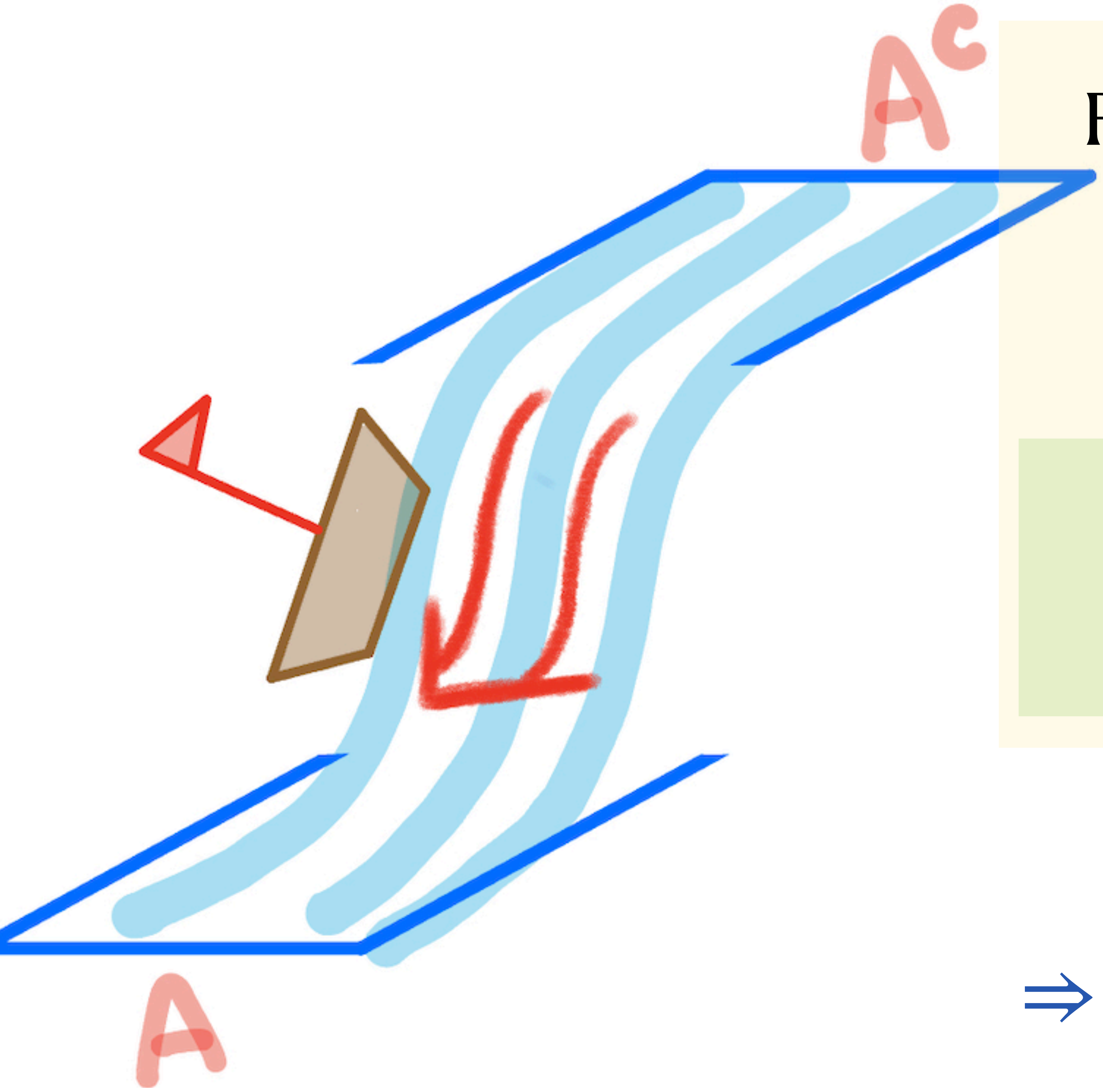
Assume the disconnected signals. There exists f_0 such that

$$\tau_\epsilon > \log\left(\frac{1}{2\epsilon}\right)\frac{1}{4}\left[\left(\frac{n^{(c-3/2)}/4 - 1}{C}\right)^{L-2} - 3\right]$$

with probability at least $1 - 4/n$ for some $C > 1$.

Deeper the signal is, worse the mixing rate becomes.

Drift Condition



For $A \subset \mathbb{T}_L$, define $\tau_A = \min\{t \geq 0 : \mathcal{T}_t \in A\}$

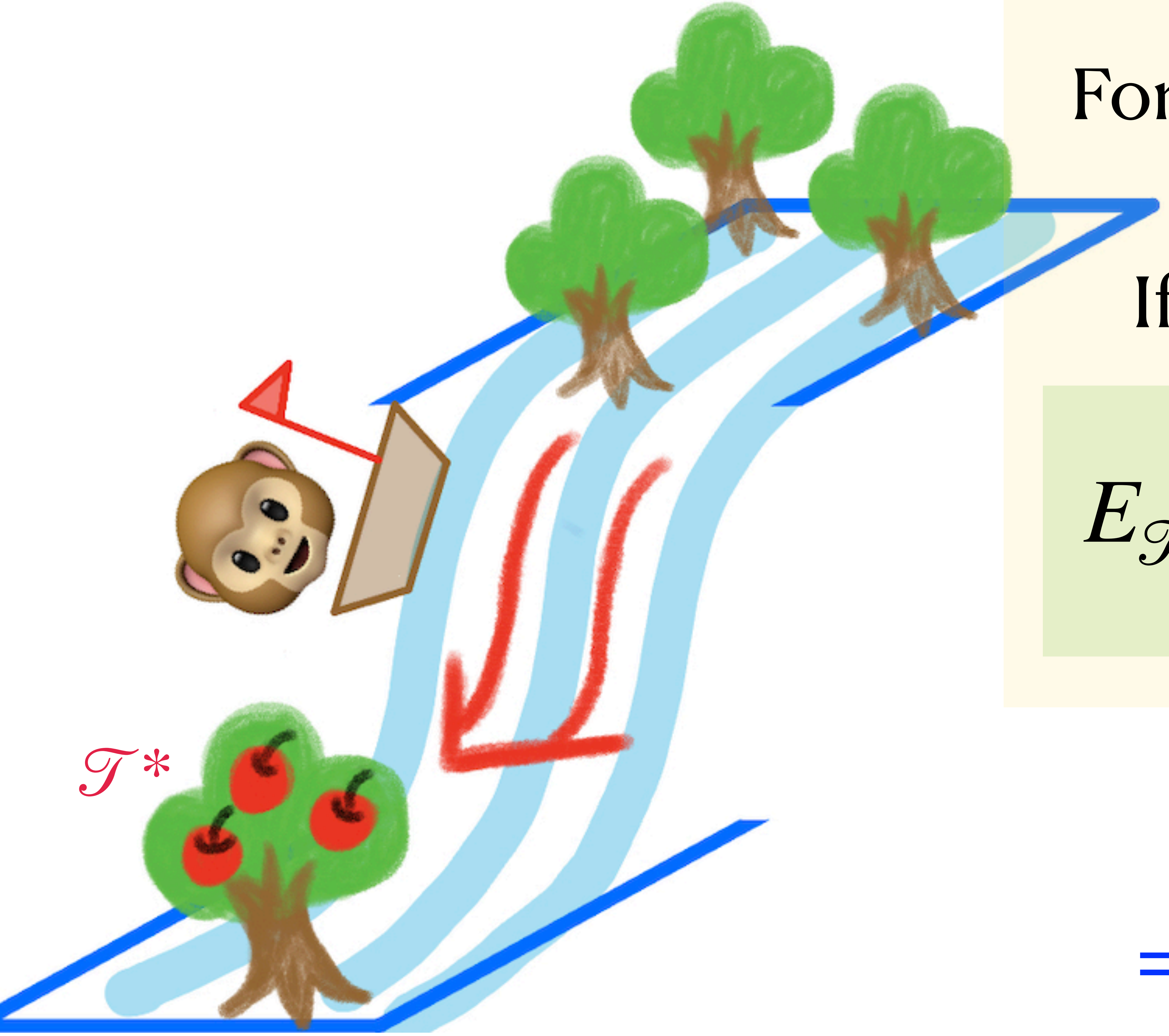
If $\exists V : \mathbb{T}_L \to [1,\infty)$ and $\exists \lambda \in (0,1)$ such that

$$E_{\mathcal{T}_0 = \mathcal{T}}\left[V(\mathcal{T}_1)\right] \leq \lambda V(\mathcal{T}) \quad \text{for all} \quad \mathcal{T} \in A^c,$$

$\Rightarrow$ for any $\mathcal{T} \in \mathbb{T}_L$, $P_{\mathcal{T}}(\tau_A \geq t) \leq \lambda^t V(\mathcal{T})$ for all $t \in \mathbb{N}$

Drift condition = Entering condition: Entering to A would not be too late!

Drift Condition



For $\mathscr{T}^* \in \mathbb{T}_L$, define $\tau^* = \min\{t \geq 0 : \mathscr{T}_t = \mathscr{T}^*\}$

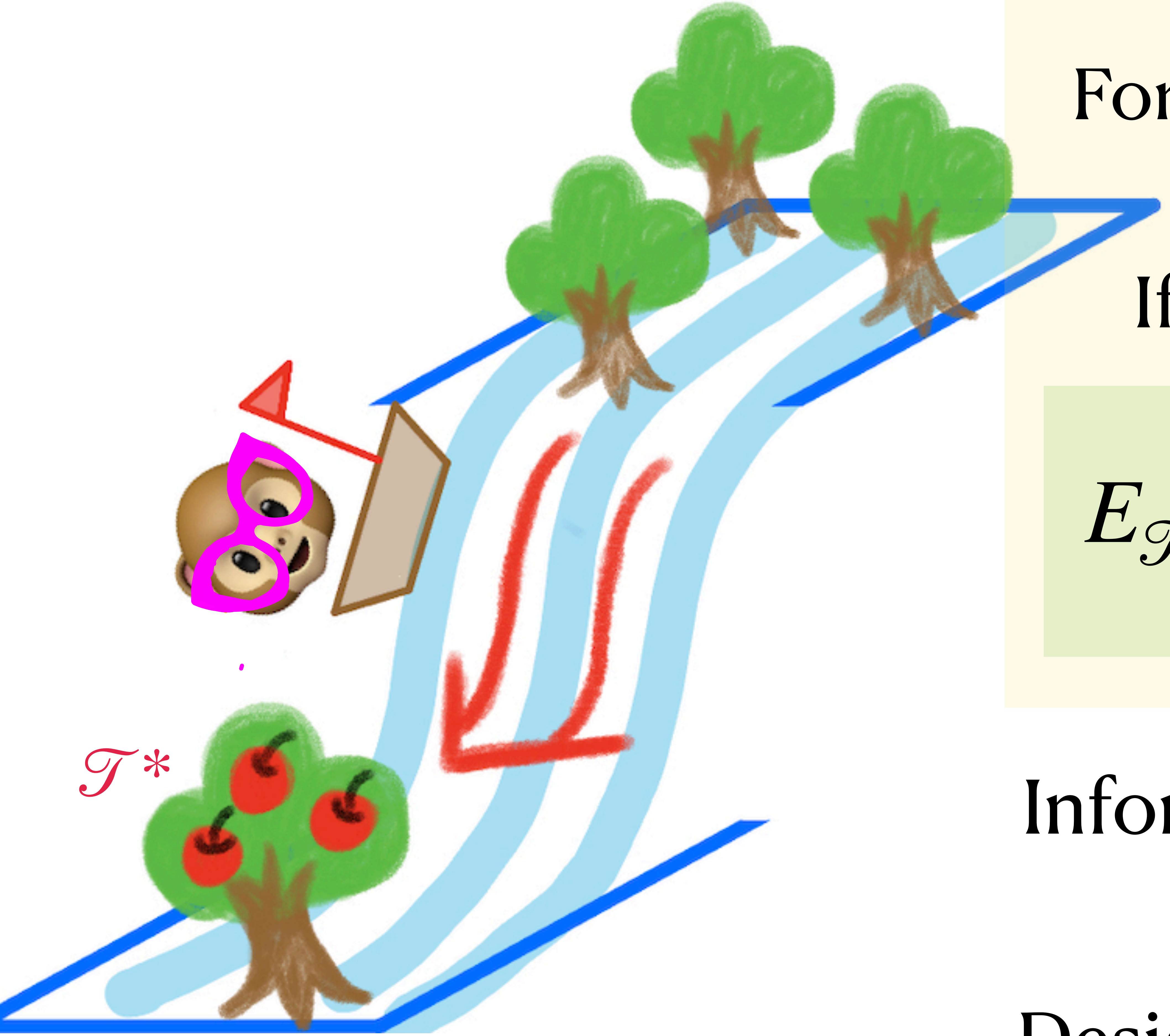
If $\exists V : \mathbb{T}_L \to [1,\infty)$ and $\exists \lambda \in (0,1)$ such that

$$E_{\mathscr{T}_0=\mathscr{T}}\left[V(\mathscr{T}_1)\right] \leq \lambda V(\mathscr{T}) \quad \text{for all} \quad \mathscr{T} \neq \mathscr{T}^*,$$

$\Rightarrow \|P^t(\mathscr{T}, \cdot\,) - \Pi[\,\cdot \mid Y]\|_{TV} \leq 2V(\mathscr{T})\lambda^{t+1}$ (Jerison, 2016)

Drift condition = Hitting condition: Hitting $\mathscr{T}^*$ would not be too late!

Drift Condition



For $\mathscr{T}^* \in \mathbb{T}_L$, define $\tau^* = \min\{t \geq 0 : \mathscr{T}_t = \mathscr{T}^*\}$

If $\exists V : \mathbb{T}_L \to [1,\infty)$ and $\exists \lambda \in (0,1)$ such that

$$E_{\mathscr{T}_0=\mathscr{T}}\left[V(\mathscr{T}_1)\right] \leq \lambda V(\mathscr{T}) \quad \text{for all} \quad \mathscr{T} \neq \mathscr{T}^*,$$

Informed Algorithm:

$$\pi(\mathscr{T} \mid Y) < < \pi(\tilde{\mathscr{T}} \mid Y)$$

Design intension for V :

If $\tilde{\mathscr{T}}$ is a desired movement, $V(\mathscr{T}) > > V(\tilde{\mathscr{T}})$

Our Drift Functions

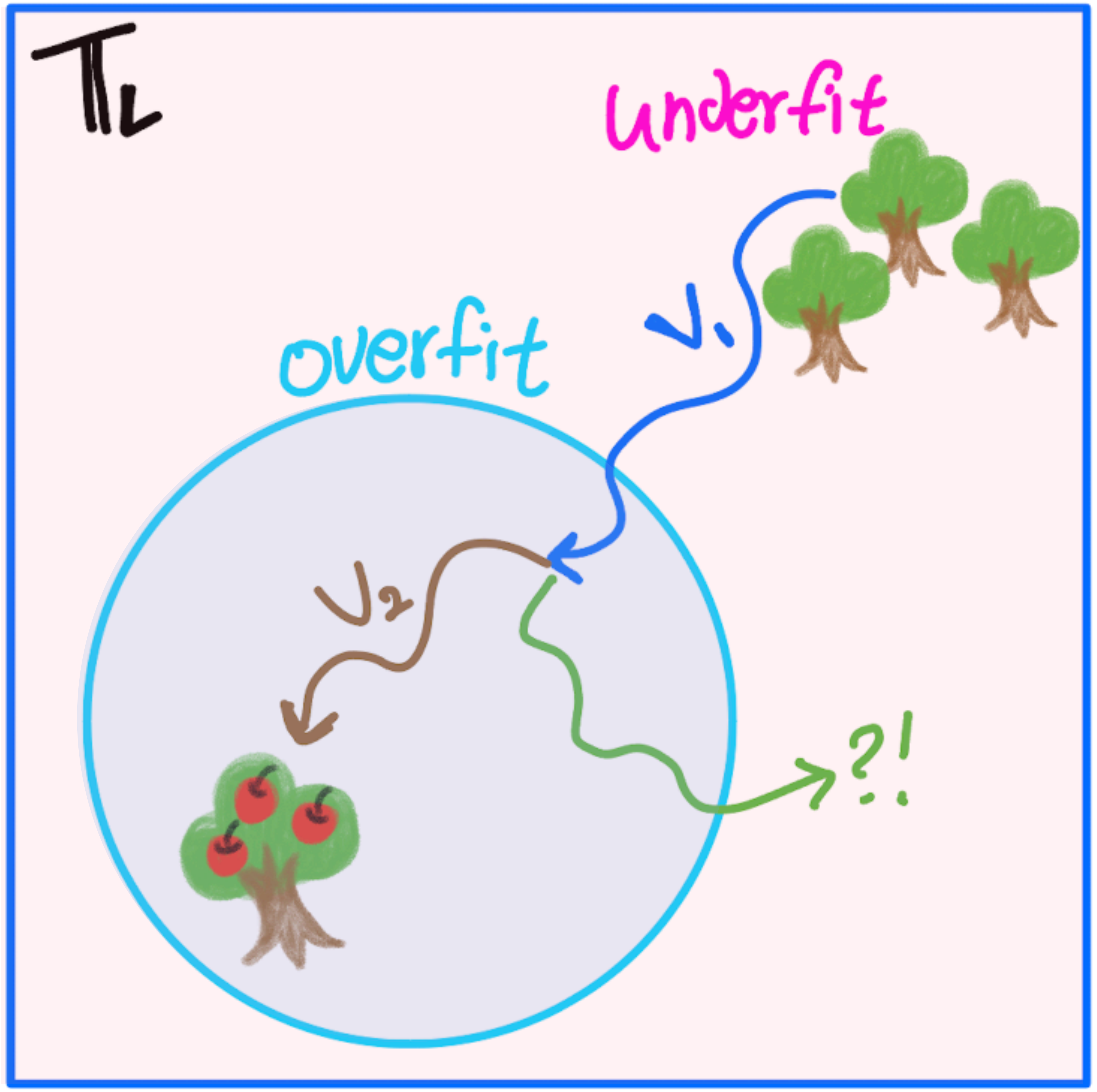
$$V_1(\mathscr{T}) = \exp\left\{ \frac{1}{C_{f_0}^2\, n^2/2} \left(Y^T(I - P_{\mathscr{T}}/n)Y \right) \right\}$$

=> Escape underfit!

$$V_2(\mathscr{T}) = \begin{cases} \exp\left\{ |\mathscr{T}_{int} \backslash \mathscr{T}^*_{int}| / 2^L \right\} & \text{if } \mathscr{T} \text{ overfits} \\ \exp\left\{ 1 - |T^*_{int}| / 2^L \right\} & \text{otherwise} \end{cases}$$

=> Escape overfit!

Two drift condition (Zhou et al., 2021)



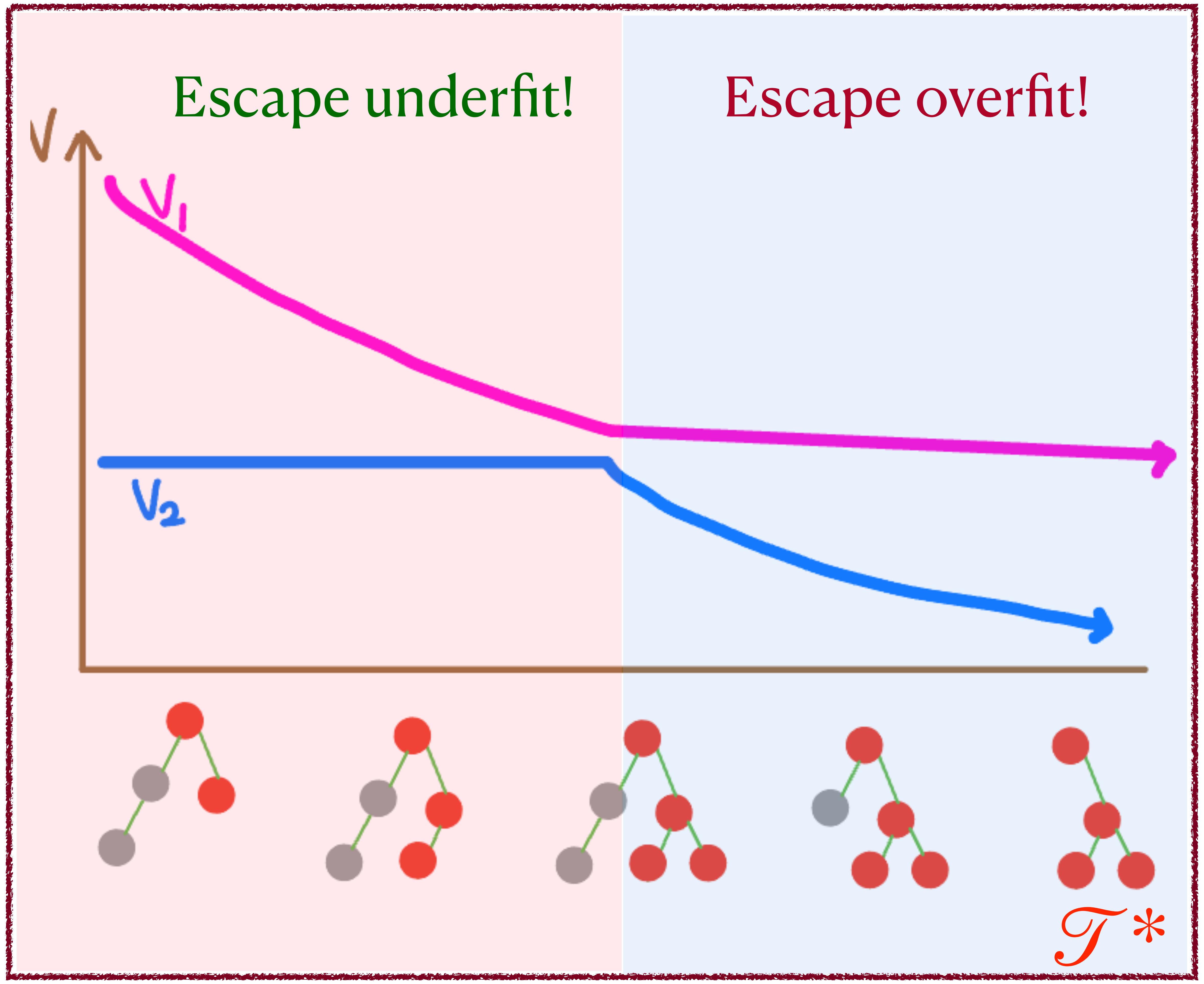
Our Drift Functions

$$V_1(\mathscr{T}) = \exp\left\{\frac{1}{C_{f_0}^2\, n^2/2}\left(Y^T(I - P_{\mathscr{T}}/n)Y\right)\right\}$$

=> Escape underfit!

$$V_2(\mathscr{T}) = \begin{cases} \exp\left\{|\mathscr{T}_{int}\backslash\mathscr{T}^*_{int}|/2^L\right\} & \text{if } \mathscr{T} \text{ overfits} \\ \exp\left\{1 - |T^*_{int}|/2^L\right\} & \text{otherwise} \end{cases}$$

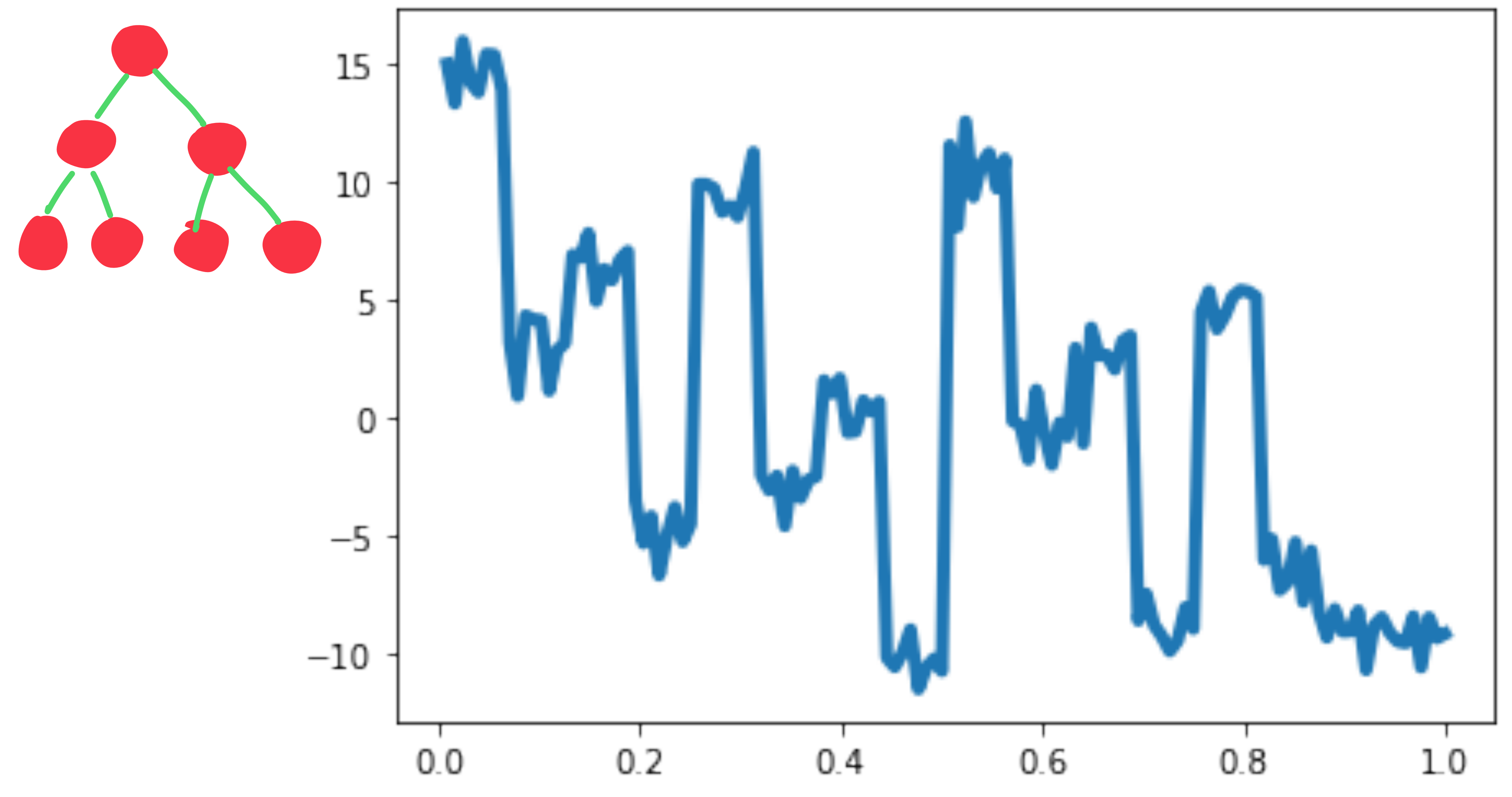
=> Escape overfit!

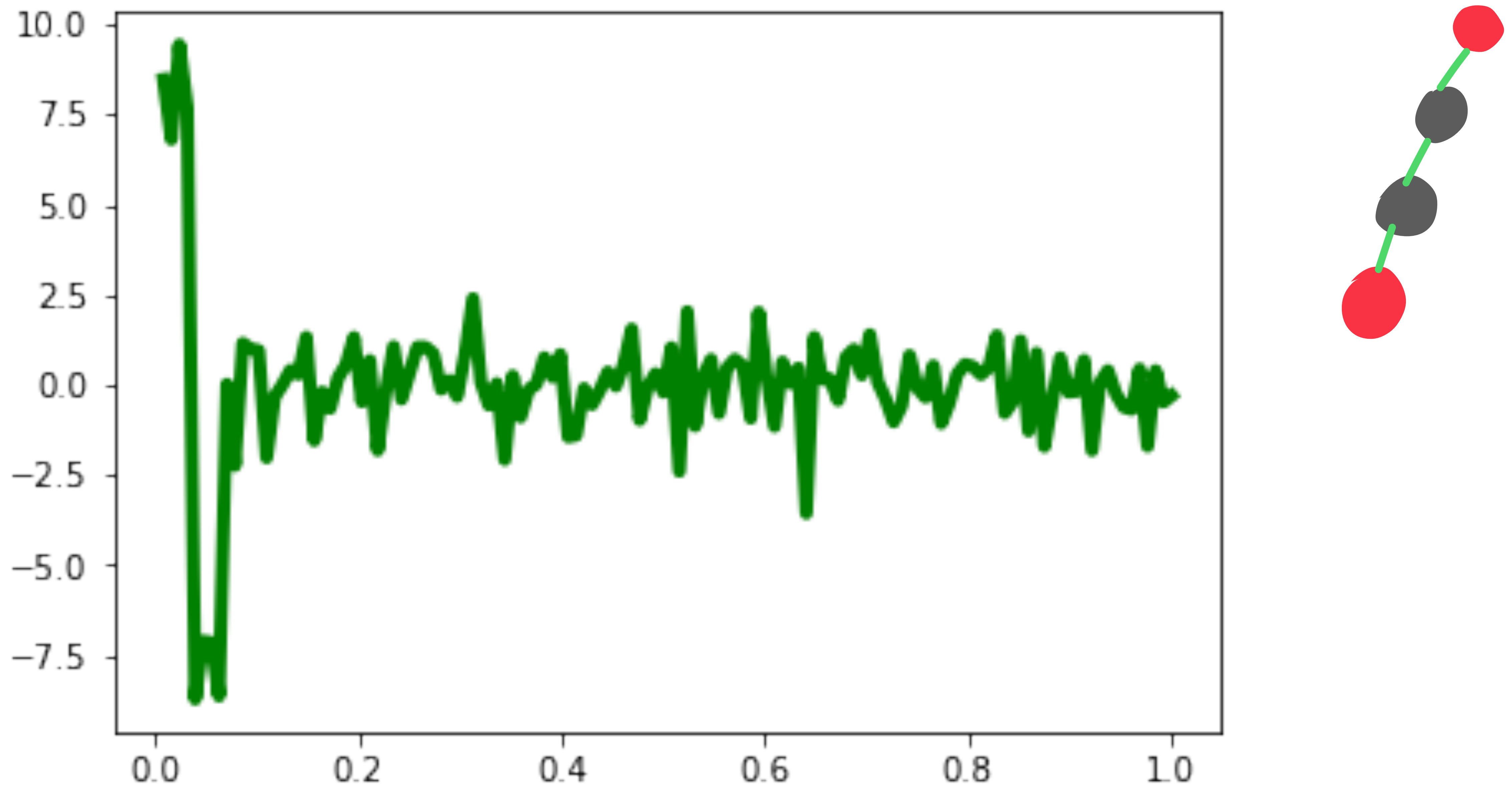


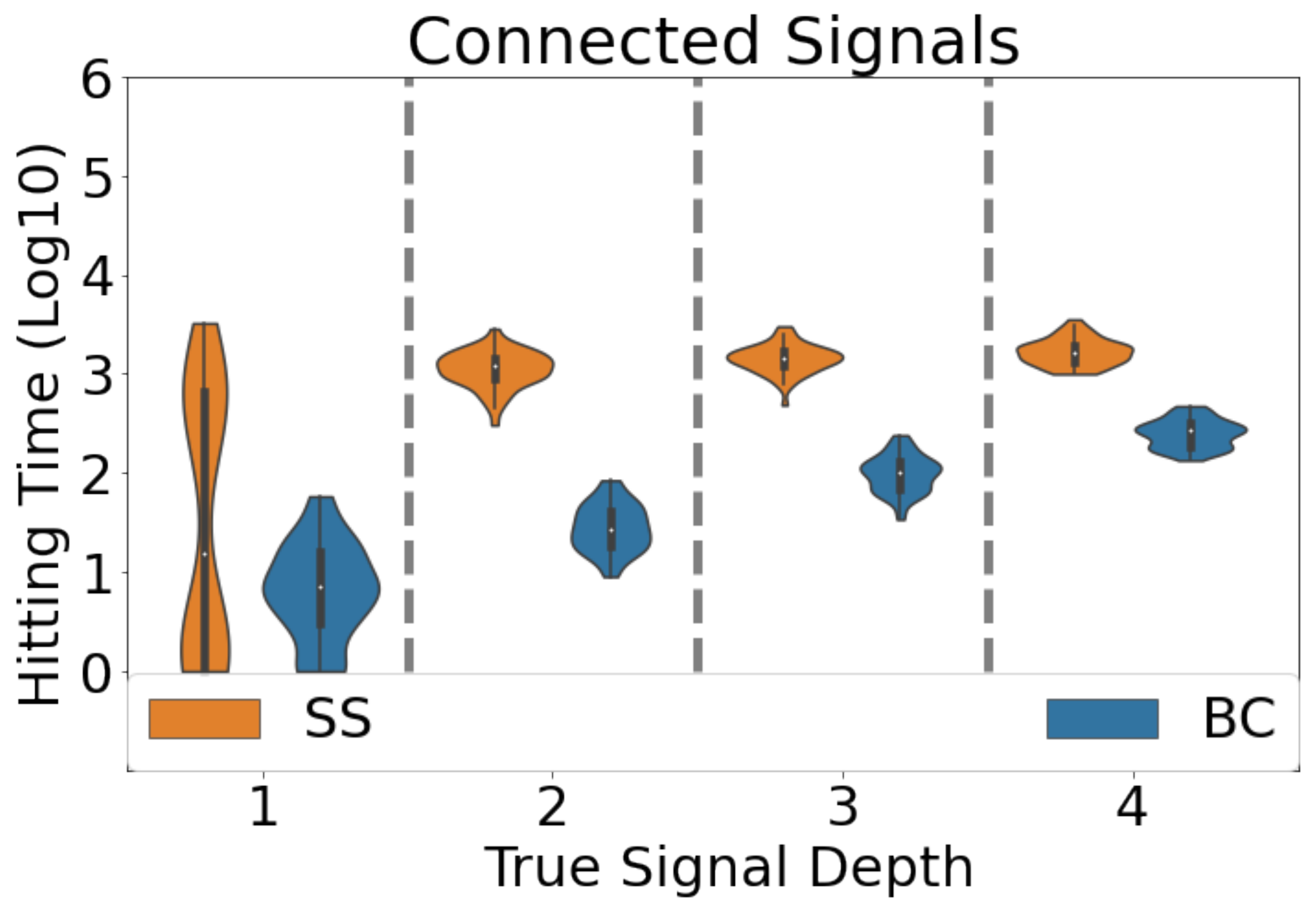
We apply two drift condition of Zhou (2021)

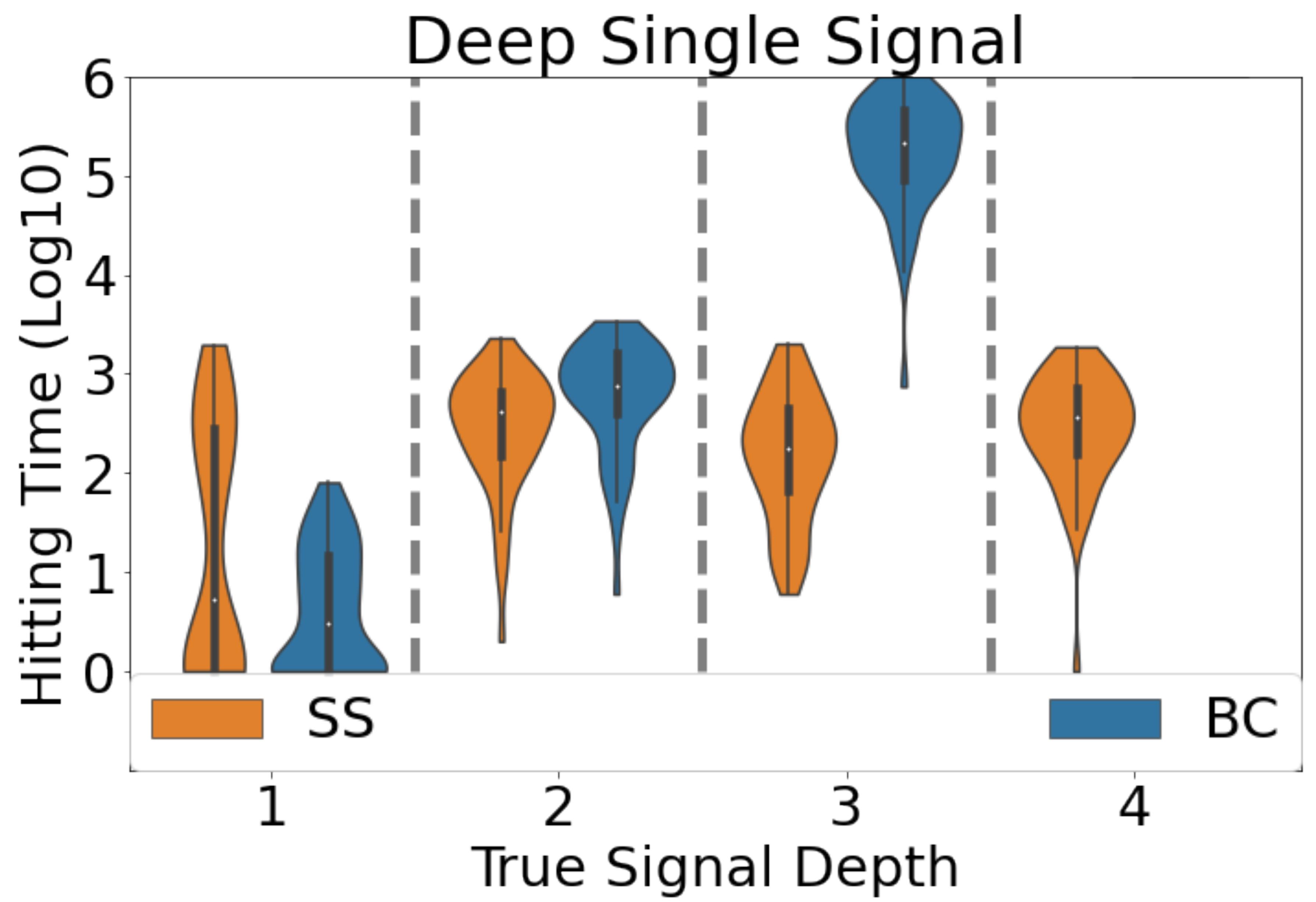
Experiment

Performance of Original Proposal

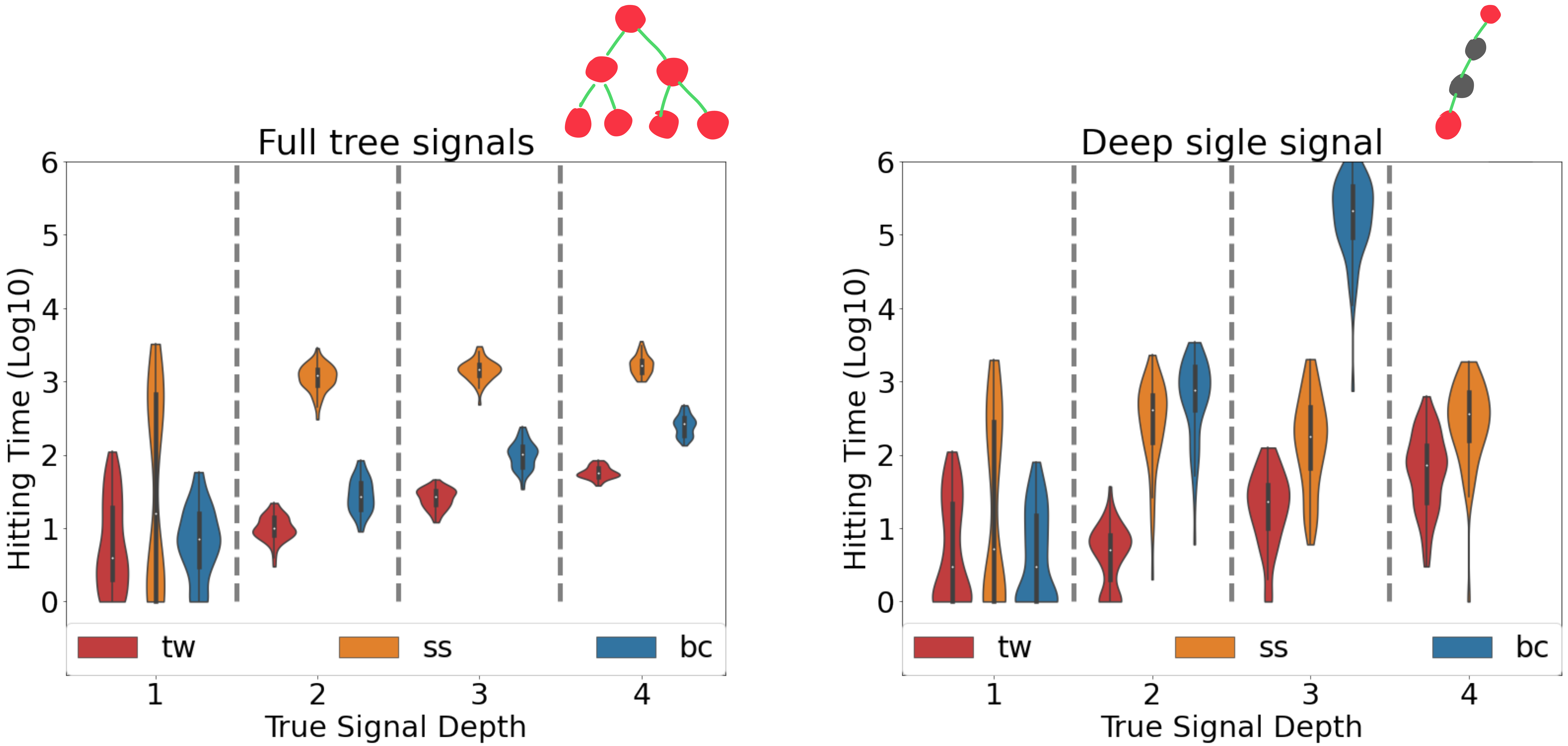






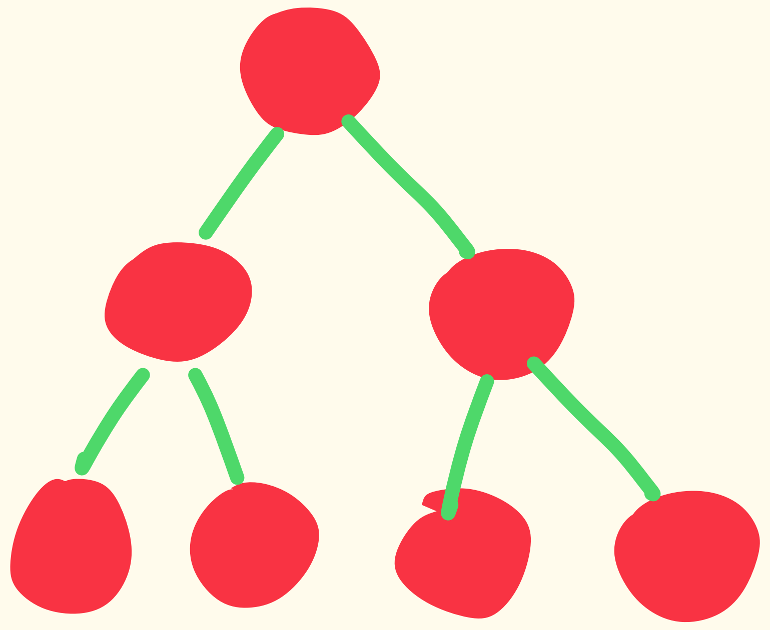
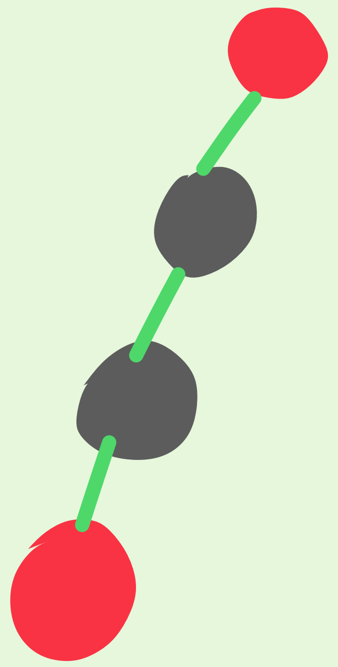
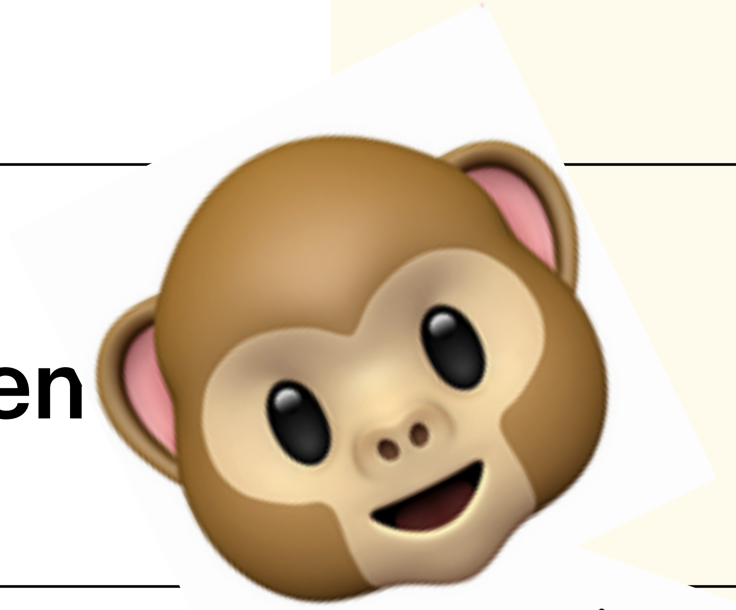
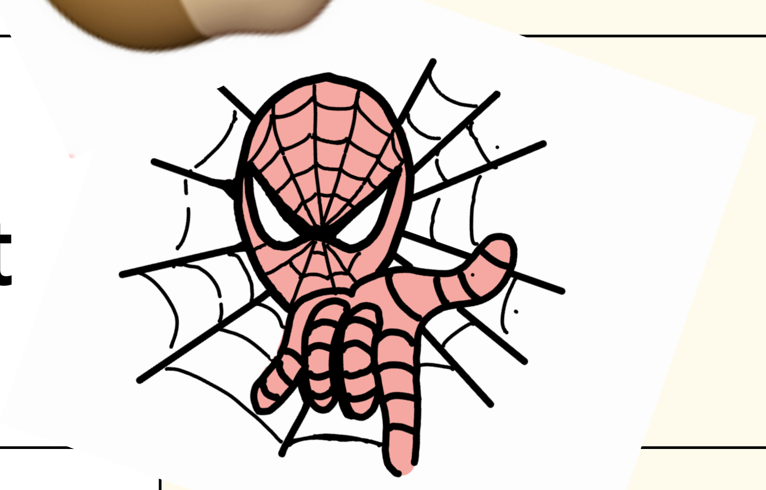
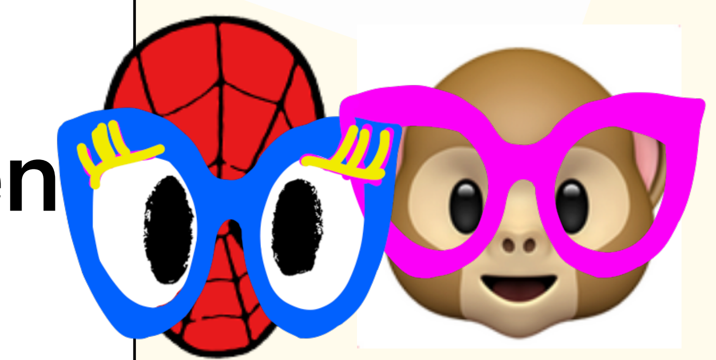


Improvement of Twiggy Proposal



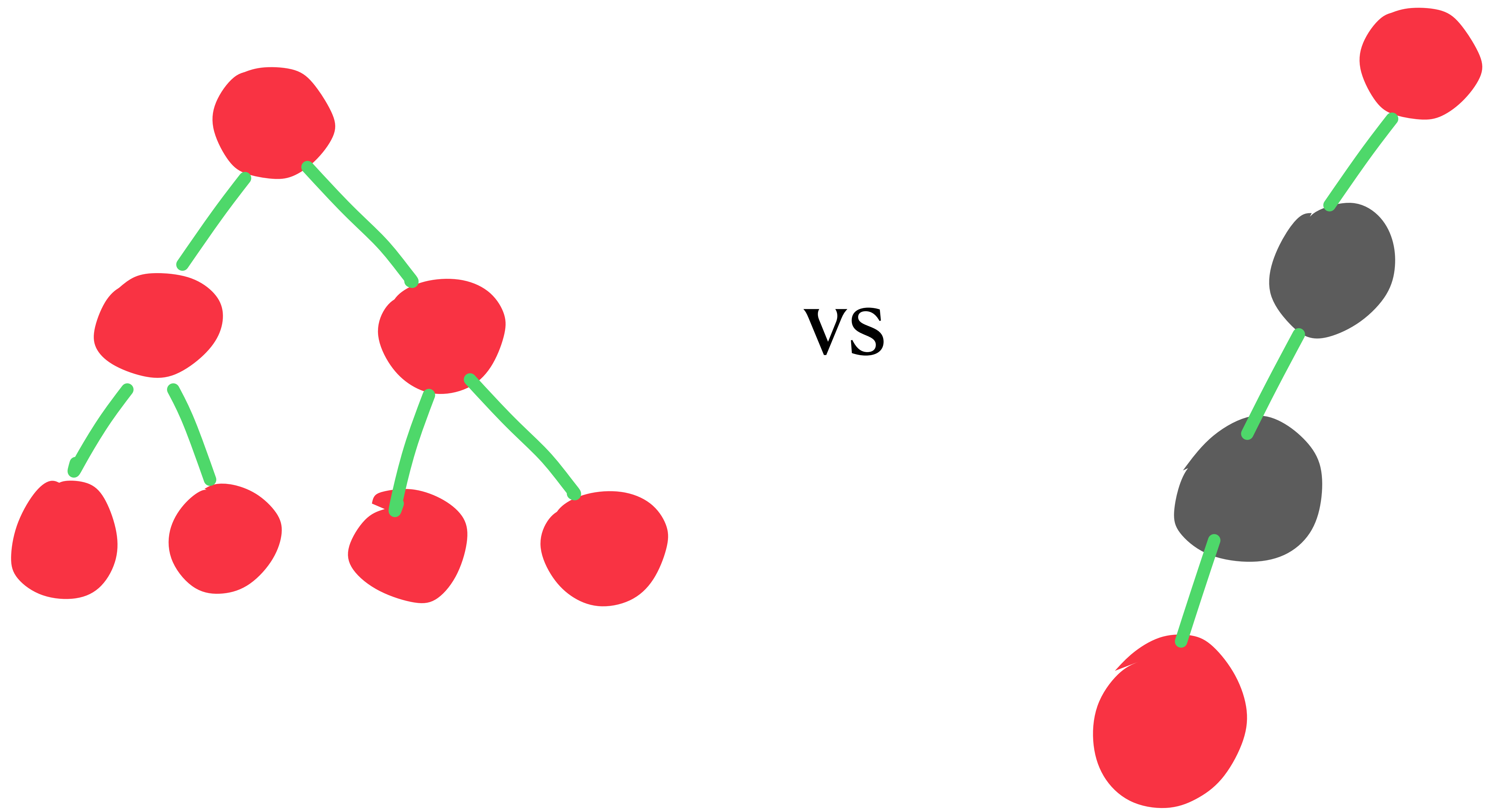
Improvement of Informed Proposal

Our Contribution

	Type A 	Type A^c 
Original Moven 	Rapid	Slow
New Movement 	Rapid	Rapid
Informed Movemen 	Improvement	Improvement?

Today's key message

1. GROW, PRUNE: myopic →The mixing depends on the data structure

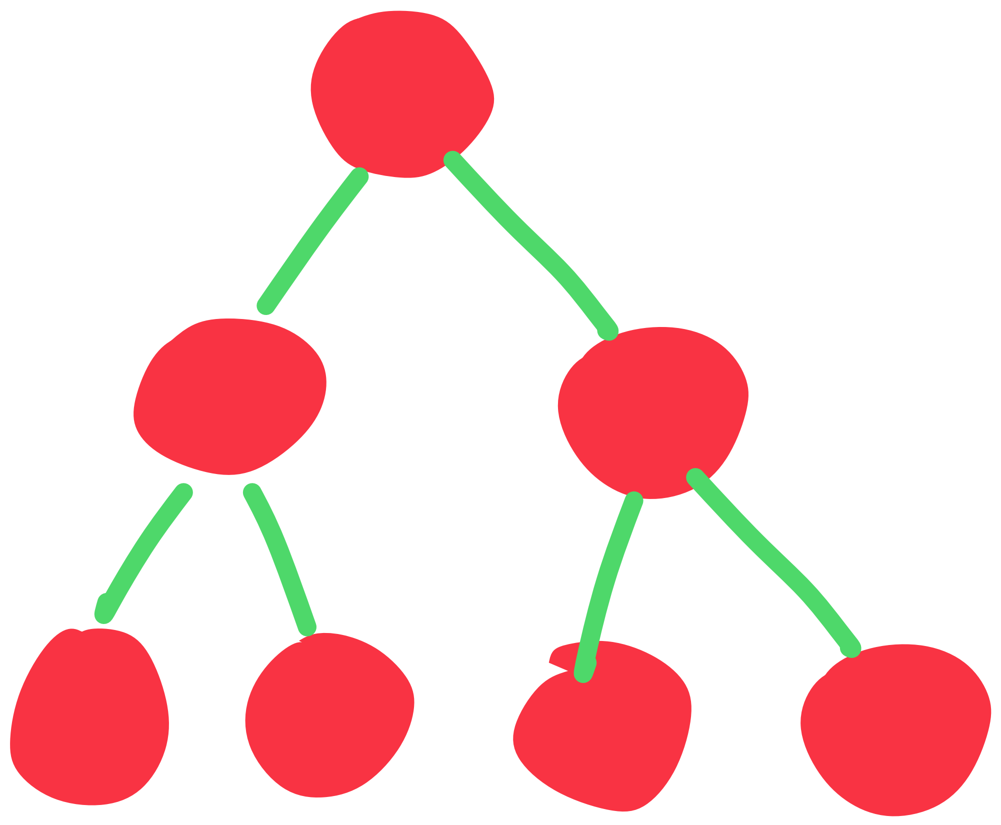


2. We may need more global movements

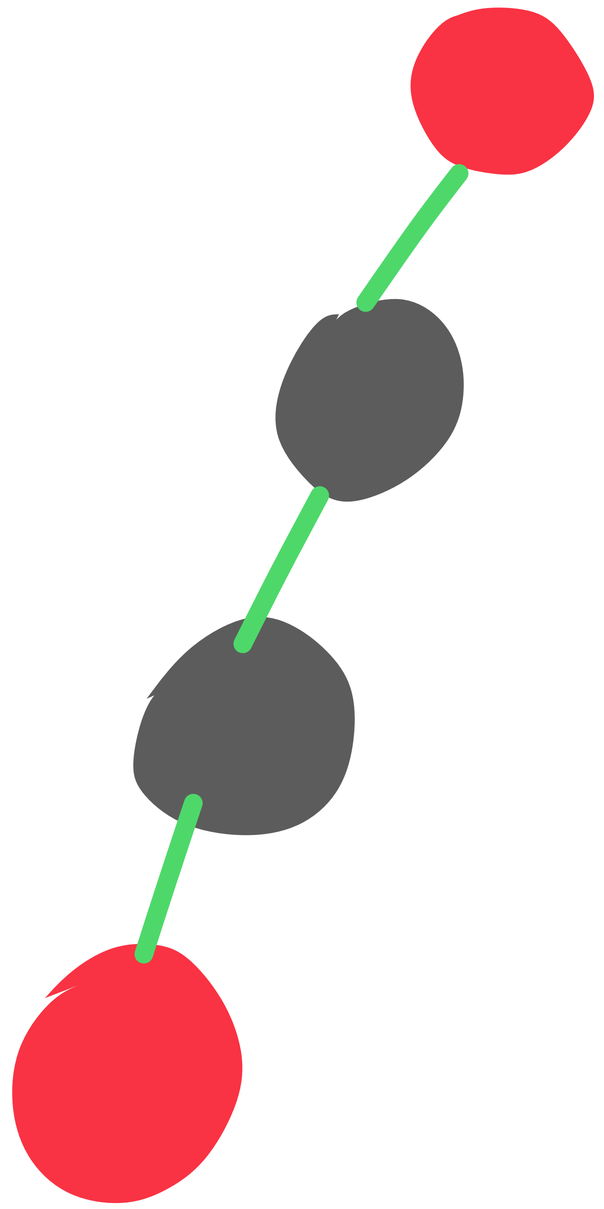
3. We may also need to open the eyes when moving (Informed MCMC)

Today's key message

1. GROW, PRUNE: myopic →The mixing depends on the data structure



vs

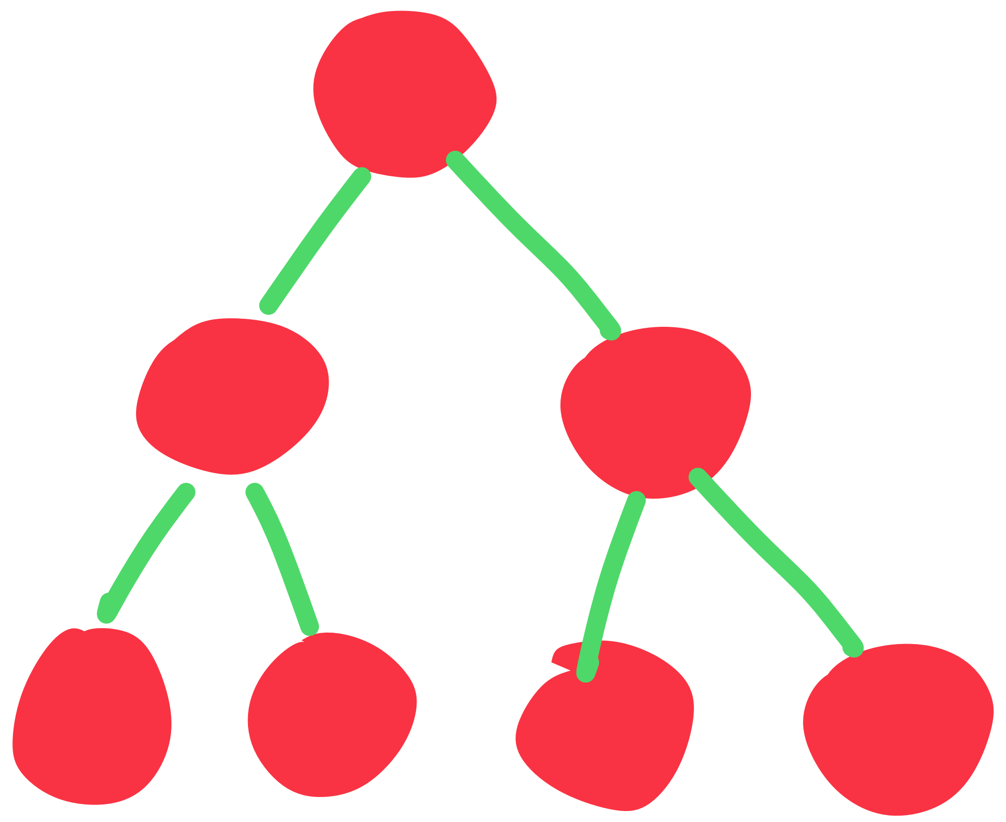


2. We may need more global movements

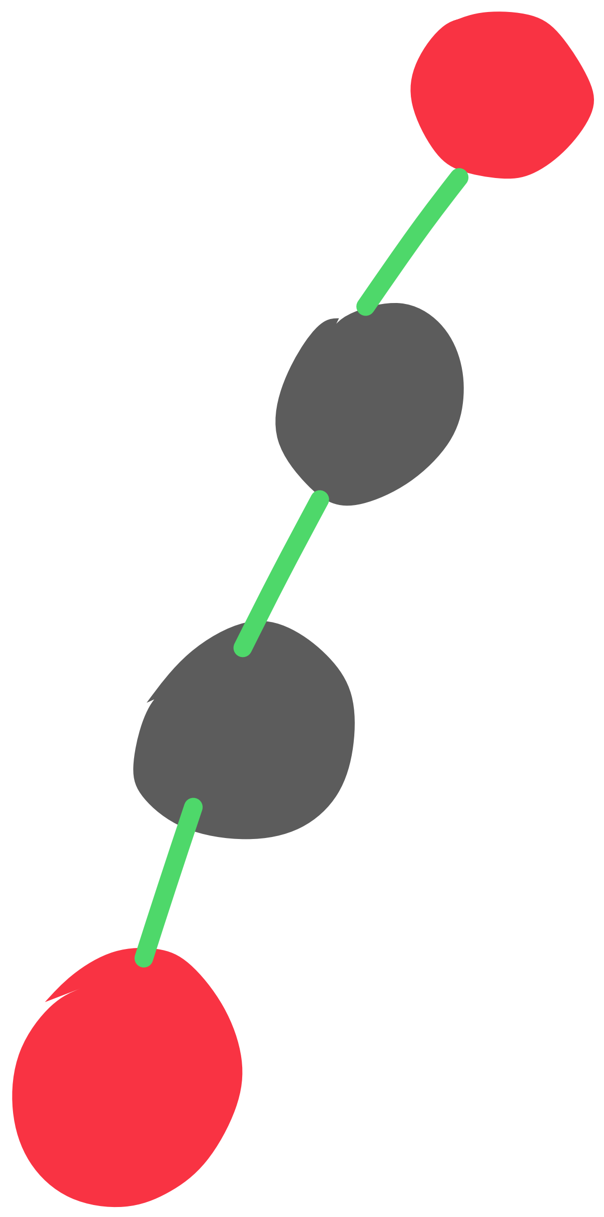
3. We may also need to open the eyes when moving (Informed MCMC)

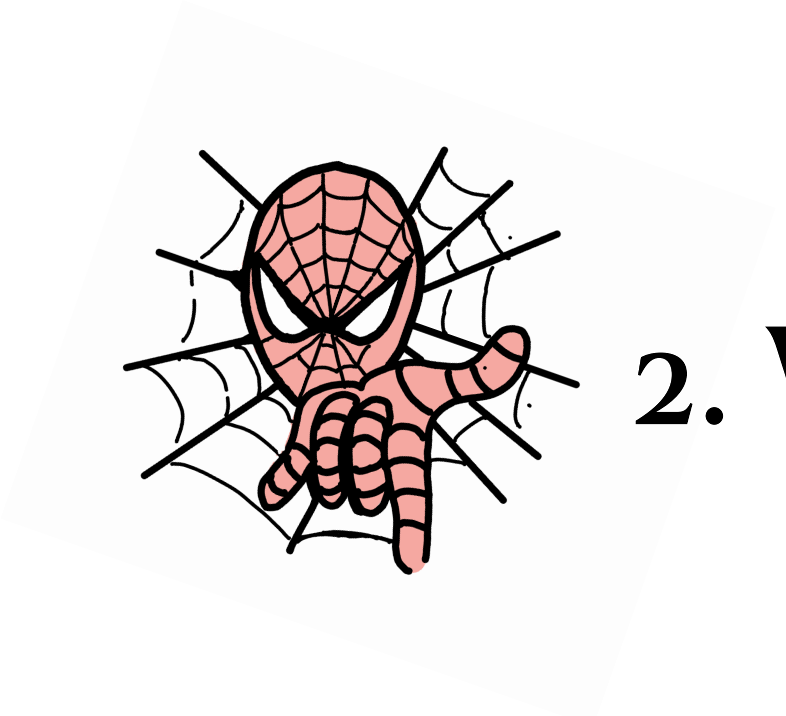
Today's key message

1. GROW, PRUNE: myopic →The mixing depends on the data structure



vs



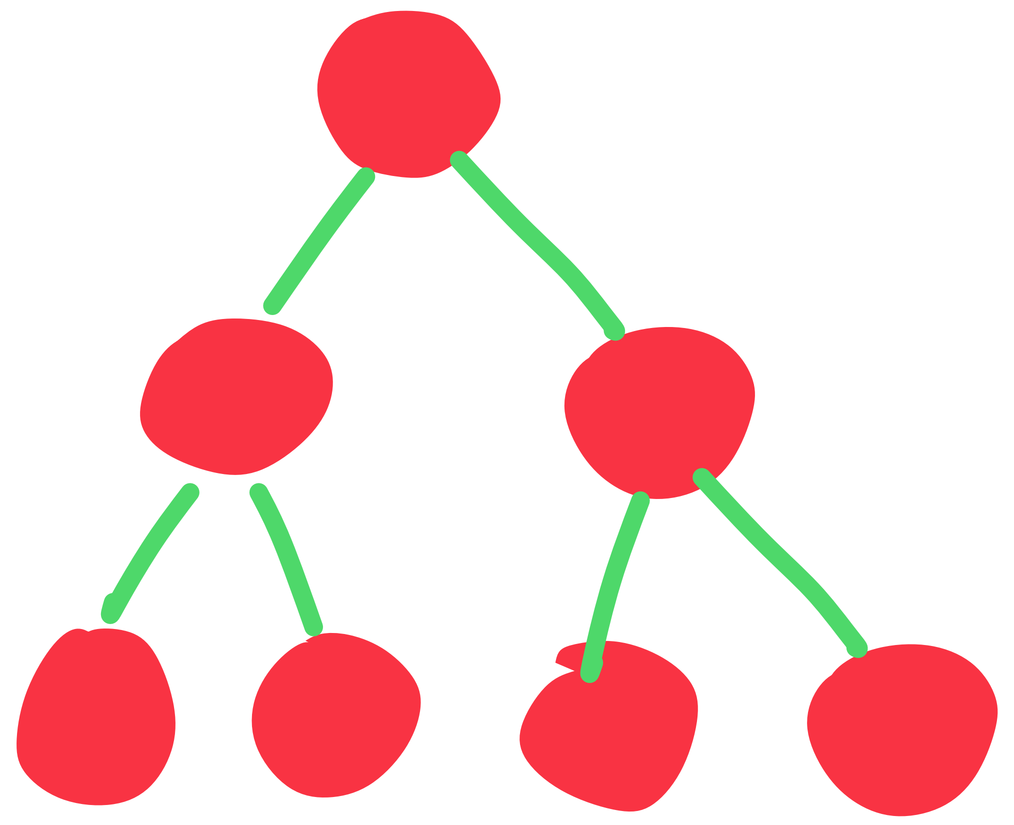


2. We may need more global movements

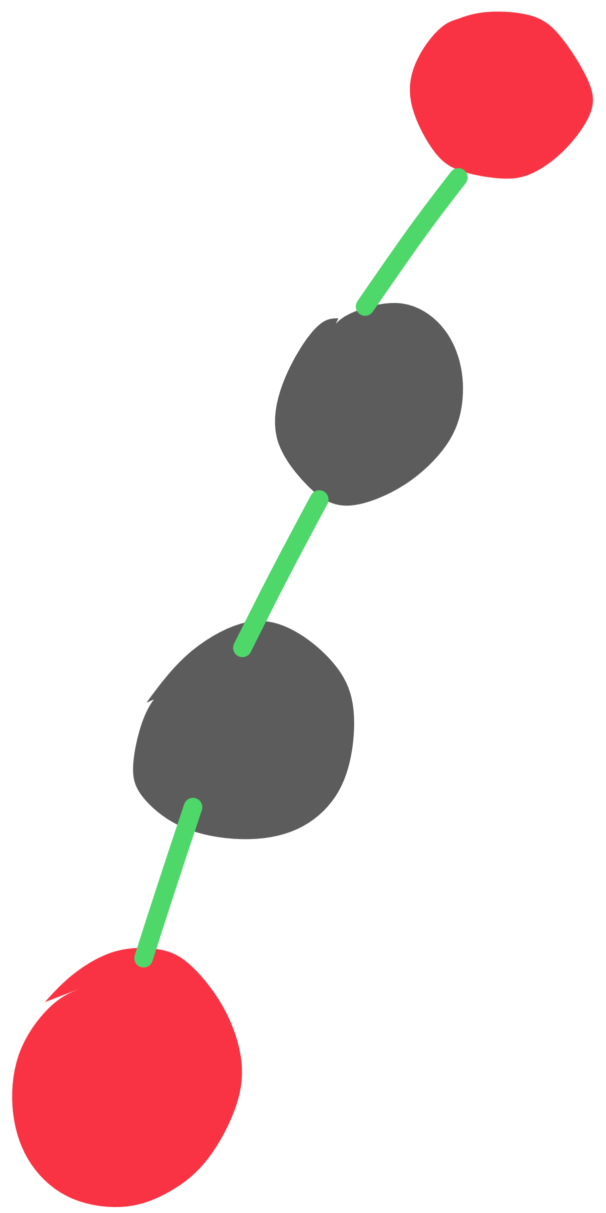
3. We may also need to open the eyes when moving (Informed MCMC)

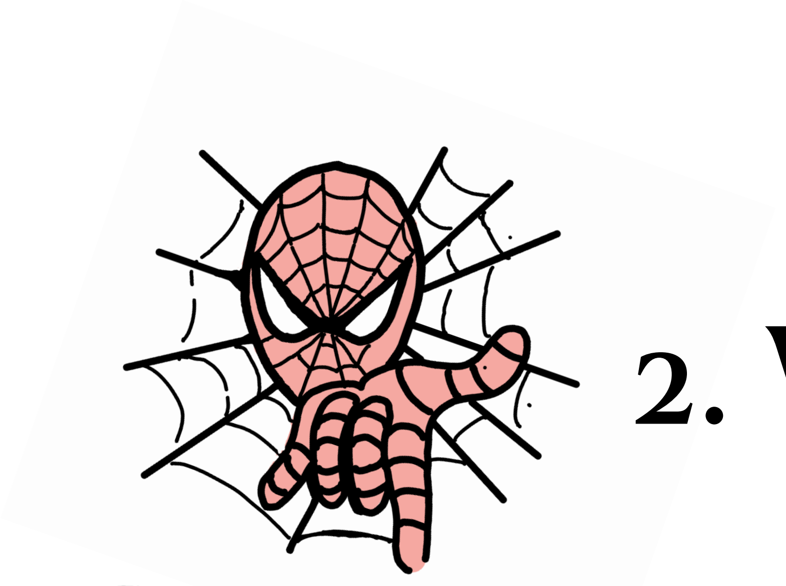
Today's key message

1. GROW, PRUNE: myopic → The mixing depends on the data structure

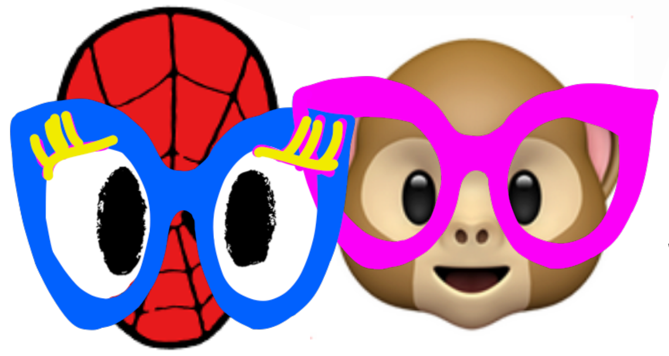


vs





2. We may need more global movements



3. We may also need to open the eyes when moving (Informed MCMC)

Acknowledgement:

Joint work with Dr. Veronika Rockova

Thank you! 

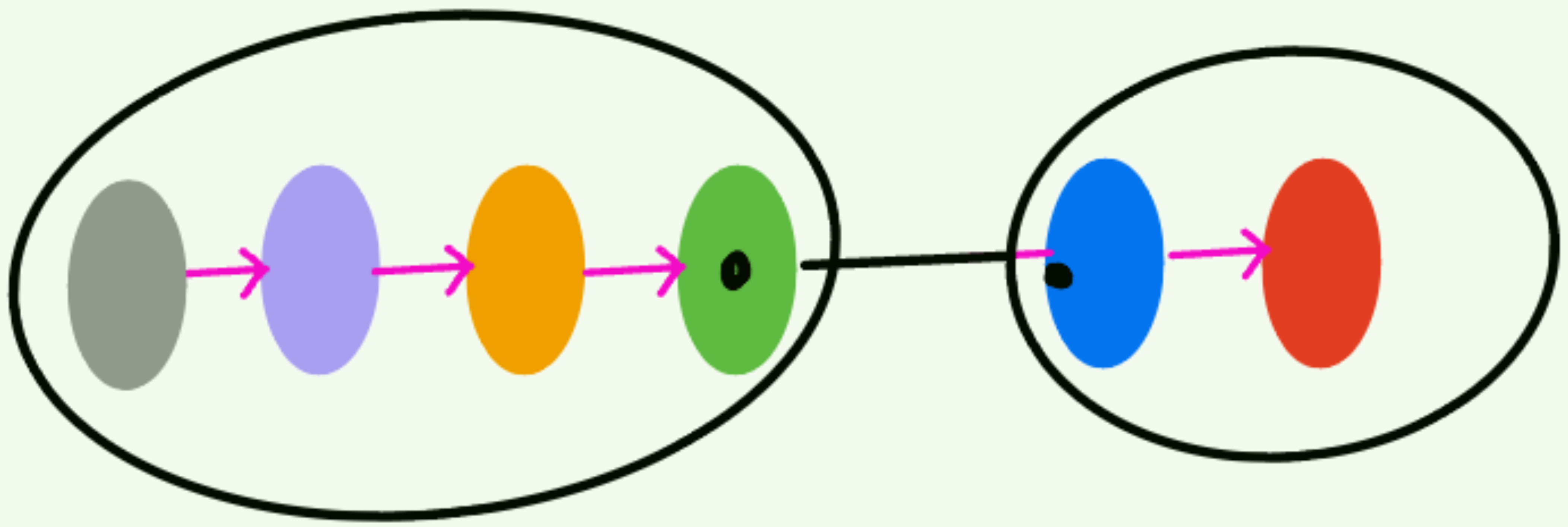
Why Not Drift Conditions for Original B-CART?

$$E_{\mathcal{T}_0=\mathcal{T}}\left[\frac{V(\mathcal{T}_1)}{V(\mathcal{T})}\right] = 1 + \sum_{m\in[Grow,Prune]}\ \sum_{\tilde{\mathcal{T}}\in N_m(\mathcal{T})}\left(\frac{V(\tilde{\mathcal{T}}\mid Y)}{V(\mathcal{T}\mid Y)} - 1\right)P(\tilde{\mathcal{T}}\mid\mathcal{T})$$

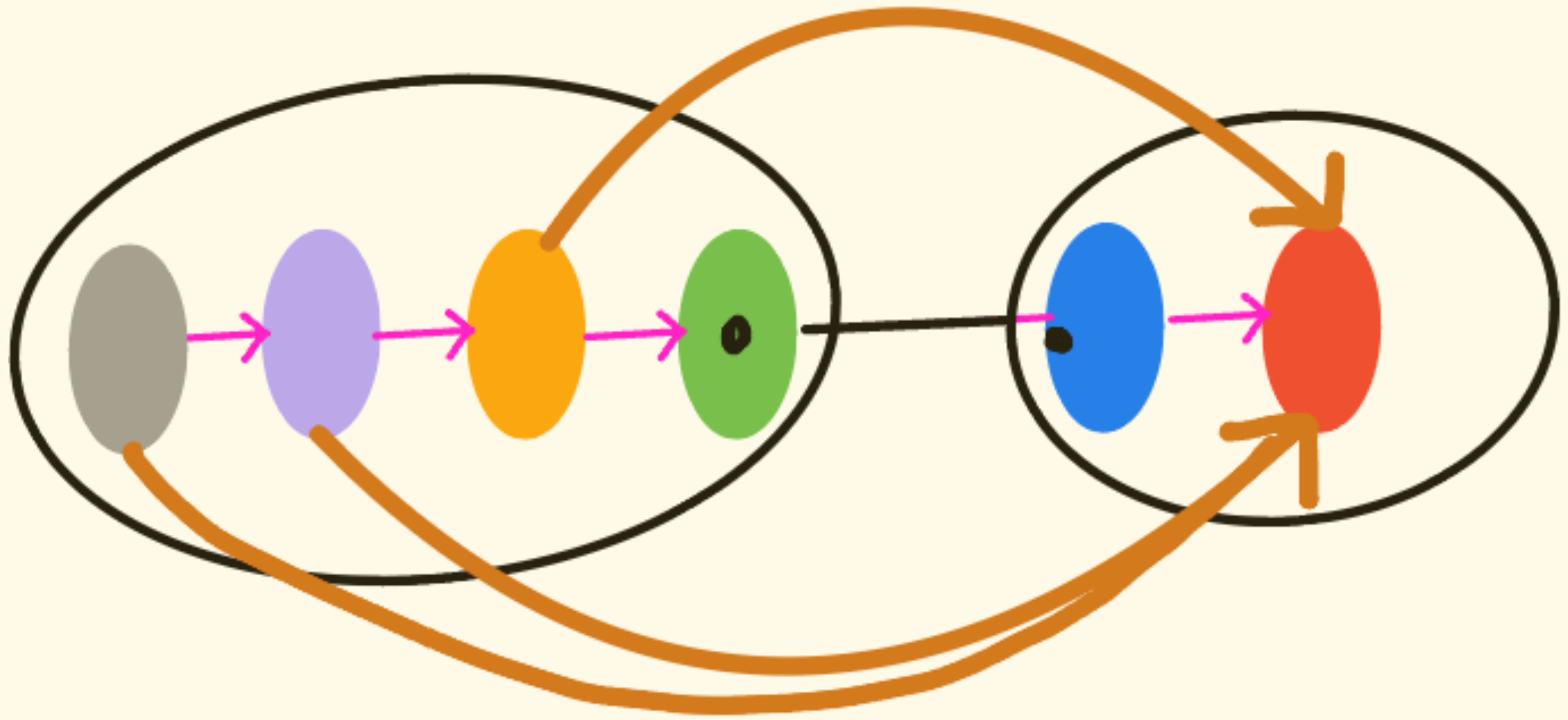
$$\geq 1 + \sum_{V(\tilde{\mathcal{T}})<V(\mathcal{T})}\left(\frac{V(\tilde{\mathcal{T}}\mid Y)}{V(\mathcal{T}\mid Y)} - 1\right)S(\mathcal{T}\to\tilde{\mathcal{T}})$$

Twiggy effect

Bottleneck made by B-CART moves



Bottleneck broken by Twiggy moves



Two Drift Conditions

Zhou et al (2022)

For $A \subset \mathcal{T}_L$ and $\mathcal{T}^* \in A$,

if $\exists V_1, V_2 : \mathcal{T}_L \to [1,\infty)$ and $\exists \lambda_1, \lambda_2 \in (0,1)$ such that

$$(PV_1)(\mathcal{T}) \le \lambda_1 V_1(\mathcal{T}) \quad \text{for all} \quad \mathcal{T} \in A^c,$$

$$(PV_2)(\mathcal{T}) \le \lambda_2 V_2(\mathcal{T}) \quad \text{for all} \quad \mathcal{T} \in A \backslash \mathcal{T}^*.$$

With some additional regularity conditions on V_1 and V_2,

$$\|P^t(\mathcal{T}, \cdot\,) - \pi\|_{TV} \le 4\alpha(\lambda_1, \lambda_2)^{t+1}\left(1 + \frac{V_1(\mathcal{T})}{M}\right),$$

where $\alpha(\lambda_1, \lambda_2)$ is an increasing function of λ_1 and λ_2

Two Drift Conditions

Zhou et al (2022)

For $A \subset \mathscr{T}_L$ and $\mathscr{T}^* \in A$,

if $\exists V_1, V_2 : \mathscr{T}_L \to [1,\infty)$ and $\exists \lambda_1, \lambda_2 \in (0,1)$ such that

$$E_{\mathscr{T}_0=\mathscr{T}}[V_1(\mathscr{T}_1)] \leq \lambda_1 V_1(\mathscr{T}) \quad \text{for all} \quad \mathscr{T} \in A^c,$$

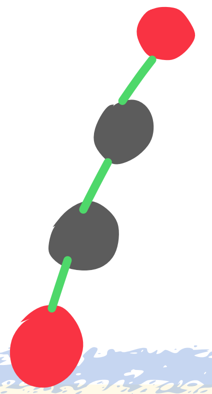
$$E_{\mathscr{T}_0=\mathscr{T}}[V_2(\mathscr{T}_1)] \leq \lambda_2 V_2(\mathscr{T}) \quad \text{for all} \quad \mathscr{T} \in A \backslash \mathscr{T}^* .$$

With some additional regularity conditions on V_1 and V_2,

$$\|P^t(\mathscr{T}, \cdot\,) - \pi\|_{TV} \leq 4\alpha(\lambda_1, \lambda_2)^{t+1}\left(1 + \frac{V_1(\mathscr{T})}{M}\right),$$

where $\alpha(\lambda_1, \lambda_2)$ is an increasing function of λ_1 and λ_2

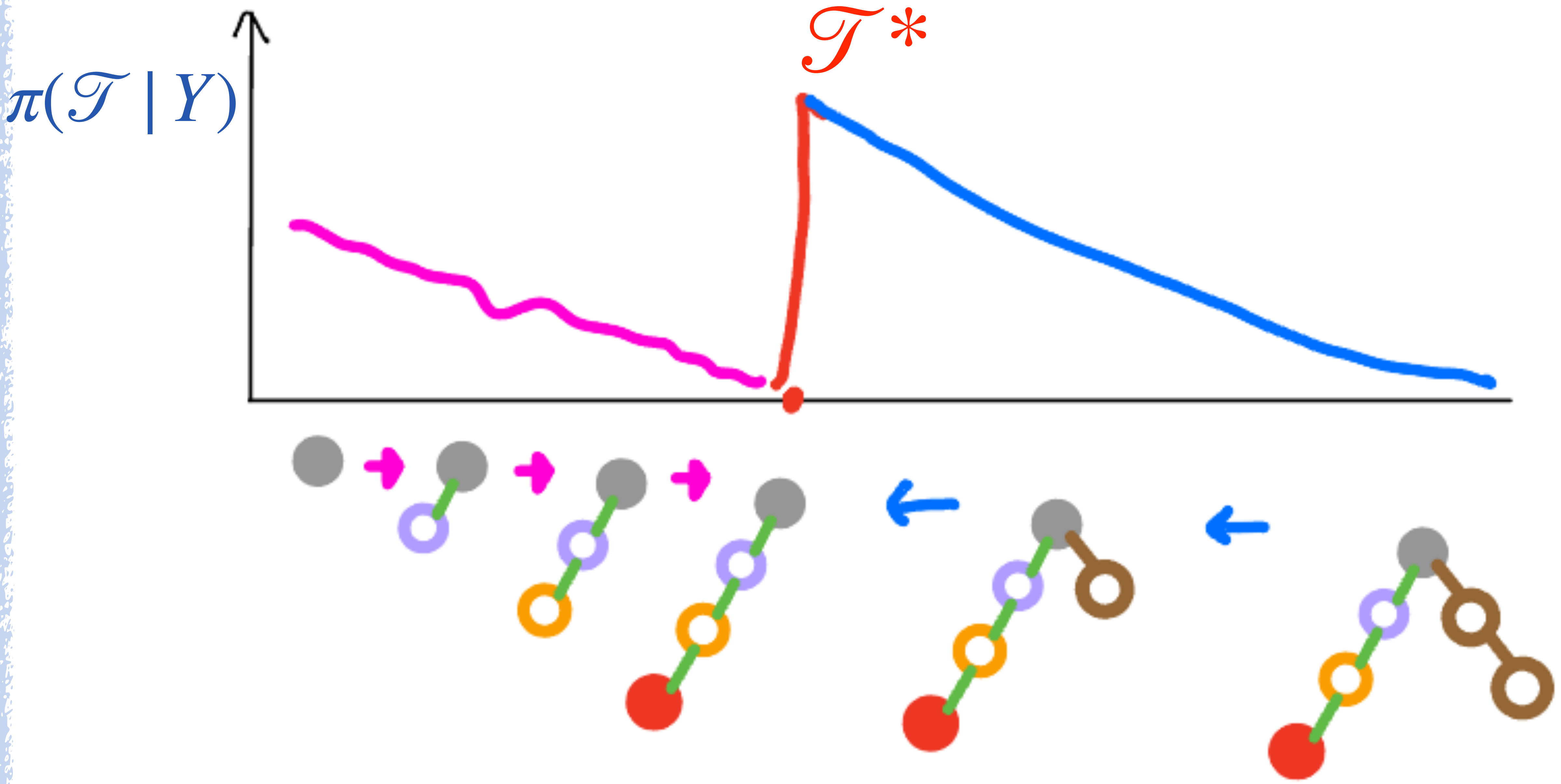
Canonical Path (Sinclair, 1992)



Our Canonical Path

Partial Overfit $\xrightarrow{\text{PRUNE}}$ Underfit $\xrightarrow{\text{GROW}}$ $\mathscr{T}^*$ $\xleftarrow{\text{PRUNE}}$ Overfit

Posterior Shaping



Exponential Lower Bound