Tropical Phylogenetics

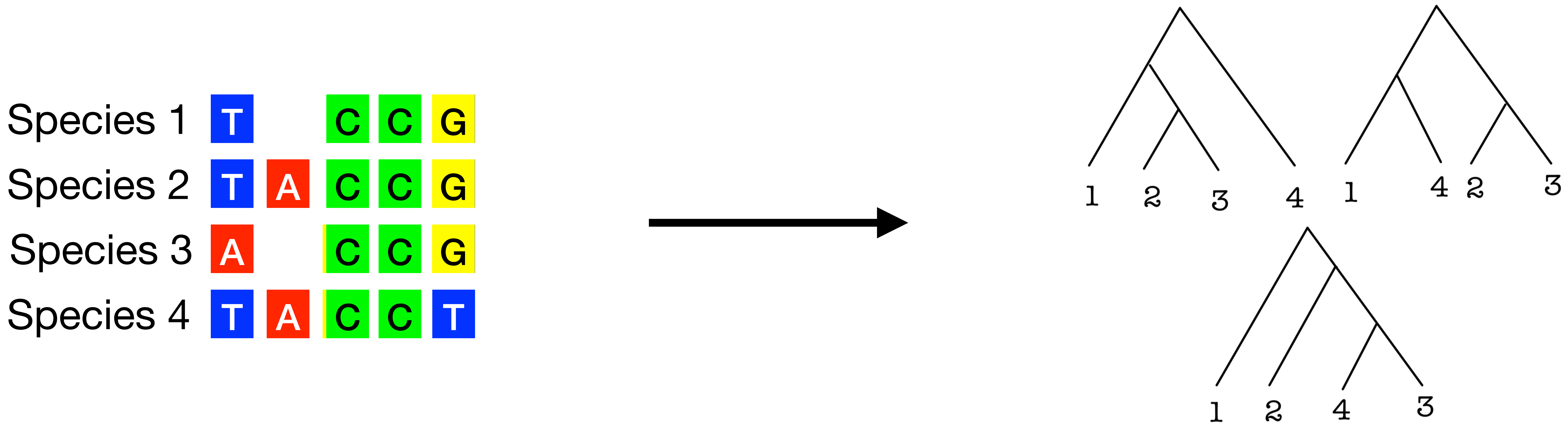
Shelby Cox

IMSI New Directions in Algebraic Statistics 2025

Tree Data

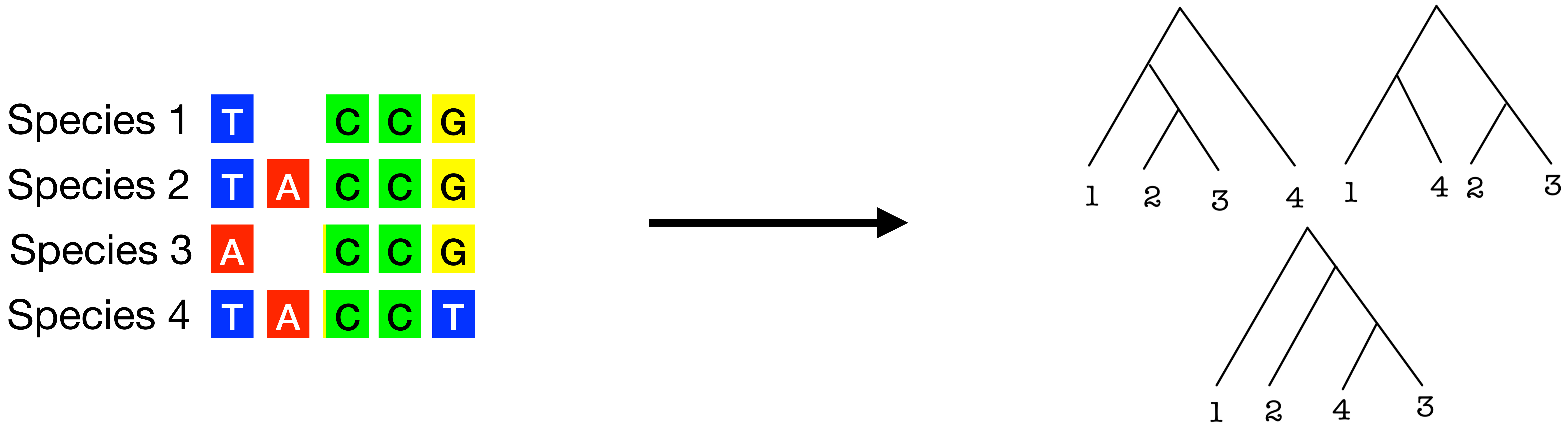
Phylogenetics

Phylogenetic reconstruction methods convert data from present-day taxa into trees that describe the evolutionary relationships between the taxa.



Phylogenetics

Phylogenetic reconstruction methods convert data from present-day taxa into trees that describe the evolutionary relationships between the taxa.



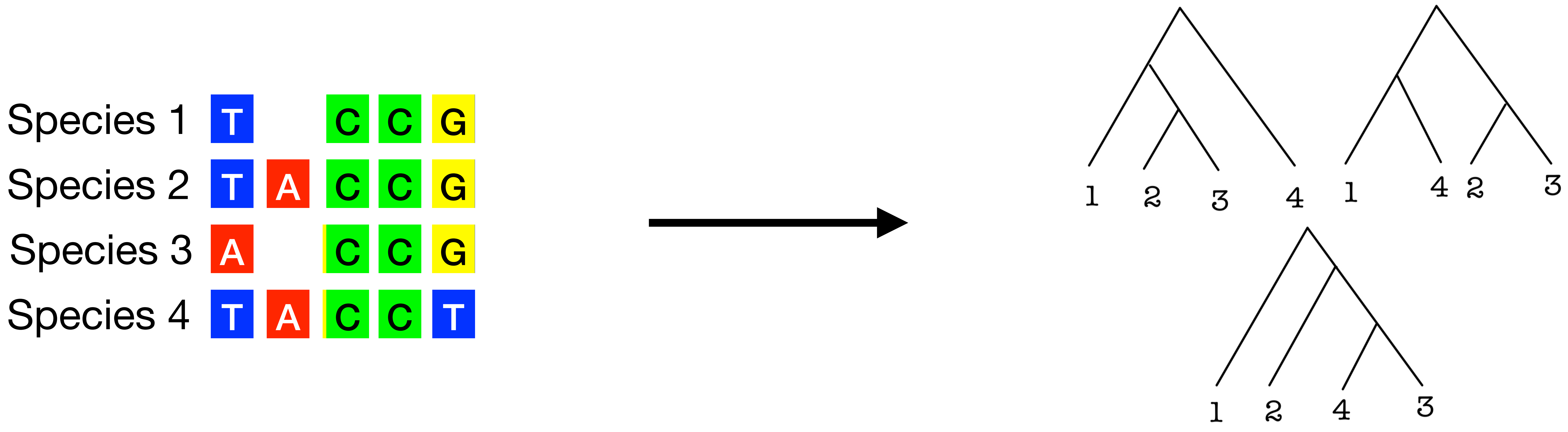
Problems:

Phylogenetic reconstruction methods often *disagree*, even with the same input data.

Gene trees often *disagree* with each other and with species trees.

Phylogenetics

Phylogenetic reconstruction methods convert data from present-day taxa into trees that describe the evolutionary relationships between the taxa.

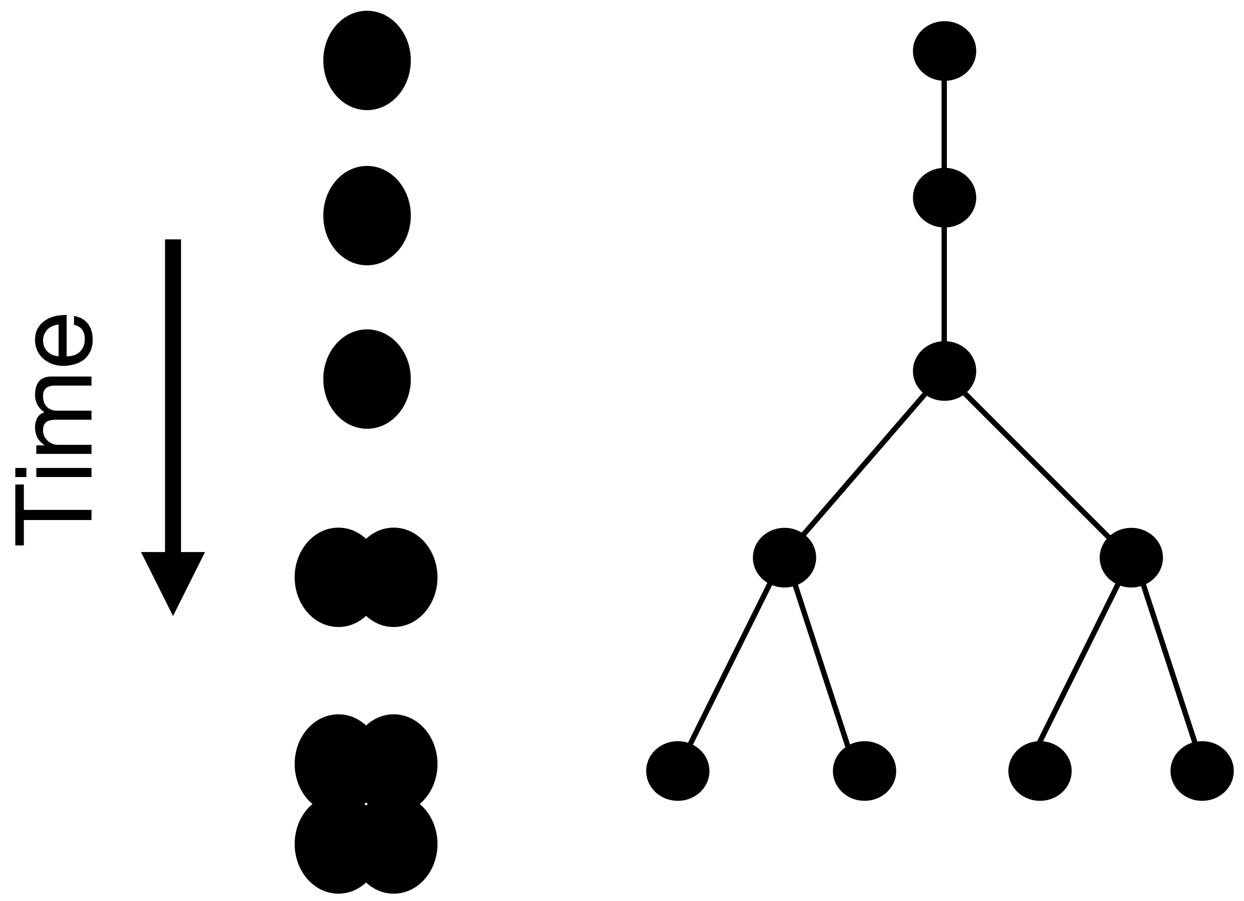


Problems: Phylogenetic reconstruction methods often *disagree*, even with the same input data.

Gene trees often *disagree* with each other and with species trees.

Question: How do we reconcile different reconstructed trees?

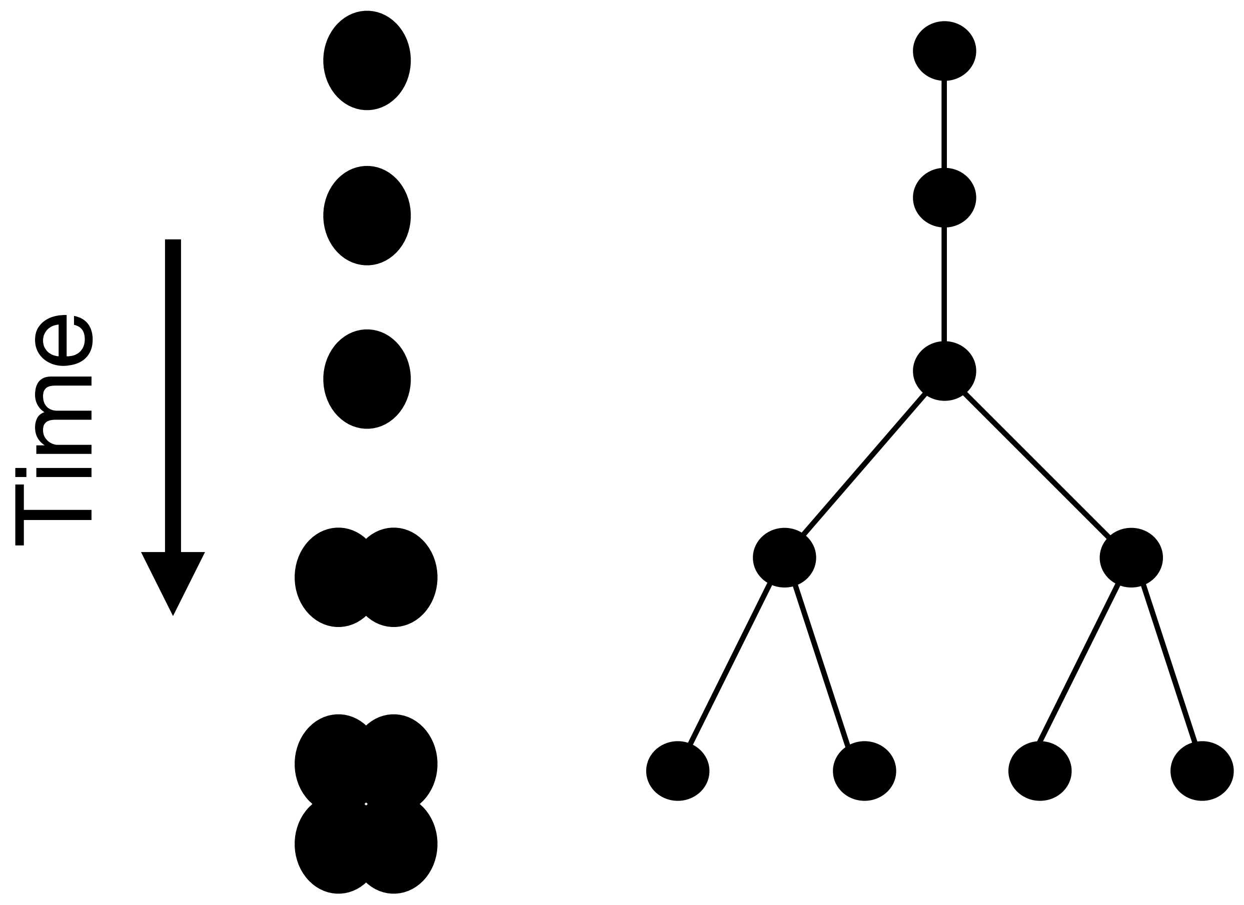
Developmental Biology



Cell lineages trace the development of an organism starting from a single cell.

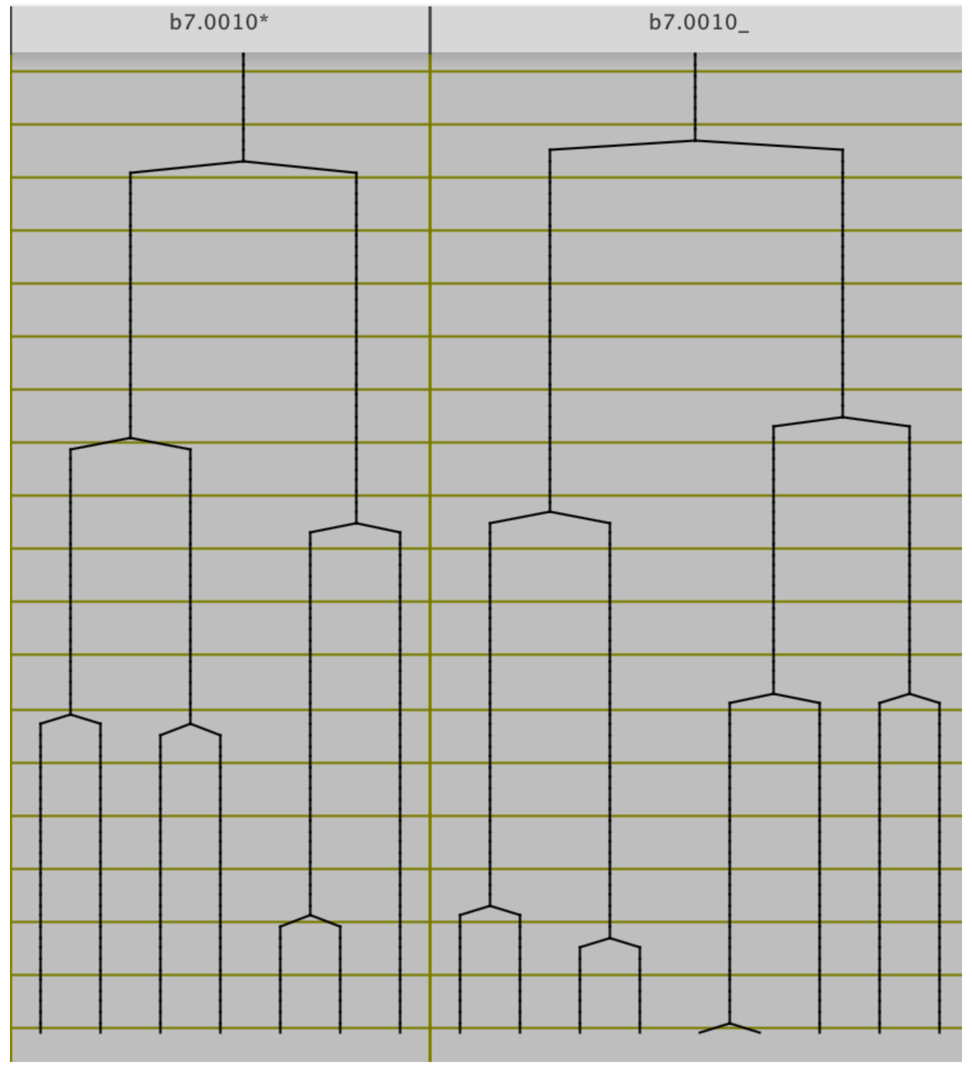
Each node in the tree represents a cell of the organism at a specific time.

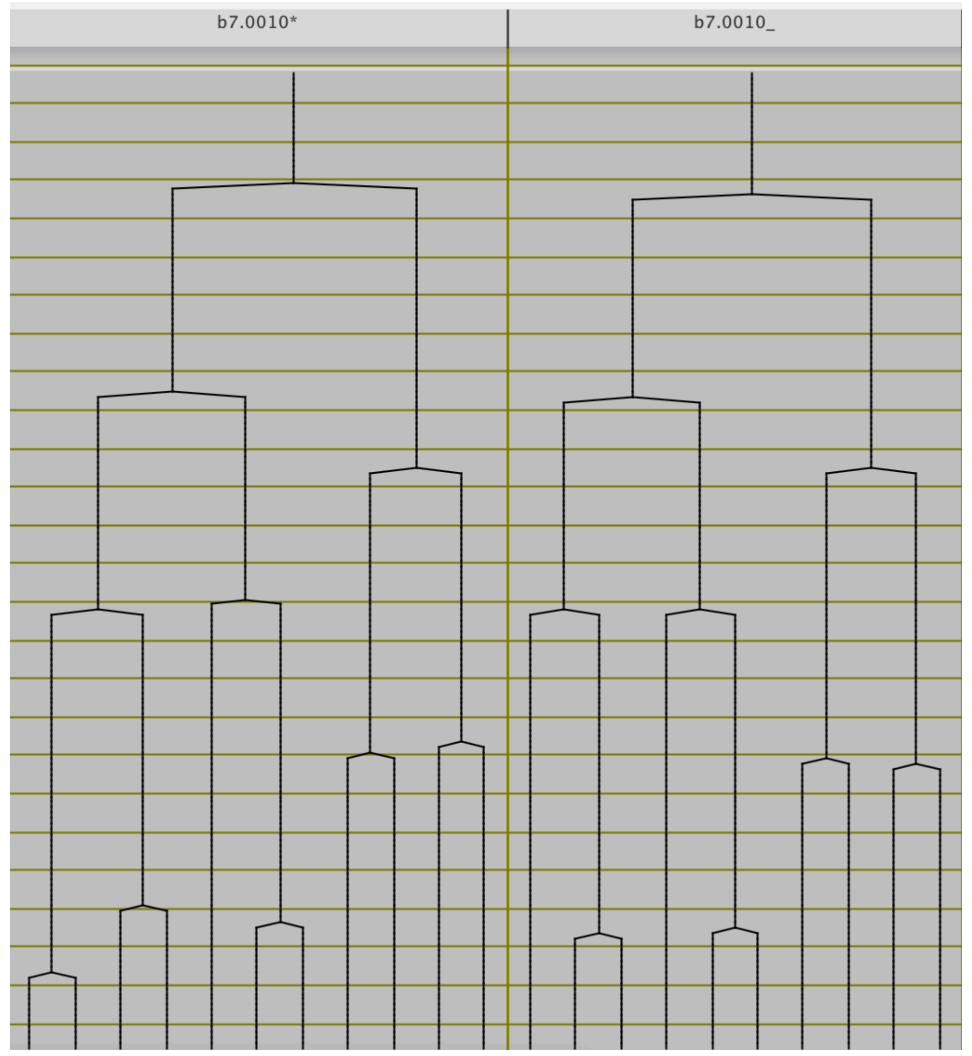
Developmental Biology



Cell lineages trace the development of an organism starting from a single cell.

Each node in the tree represents a cell of the organism at a specific time.



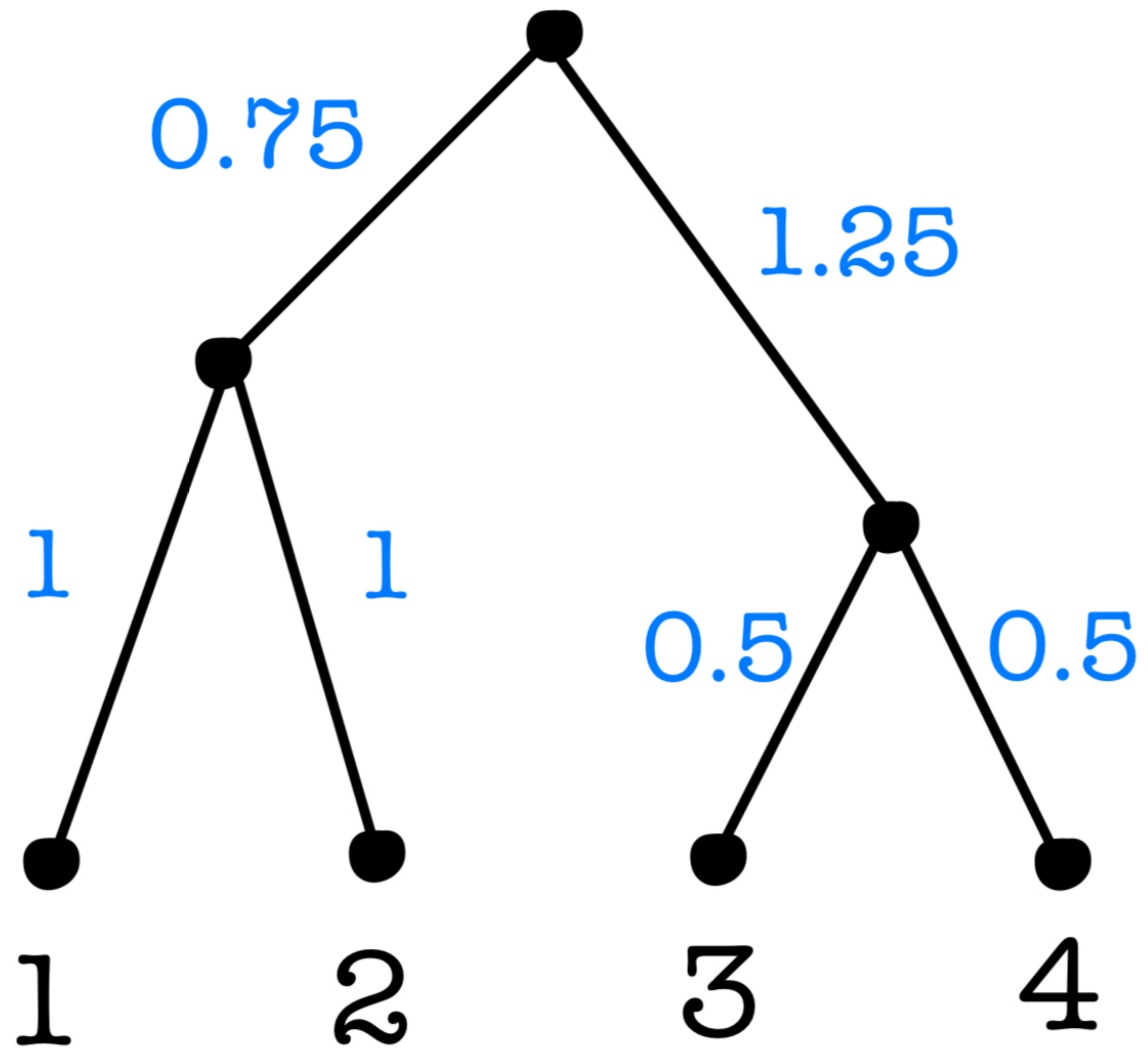


Phallusia mammillata developmental lineage trees [L], viewed in Mastodon.

Photo by Diego Delso [D].

“Phylogenetic” Trees

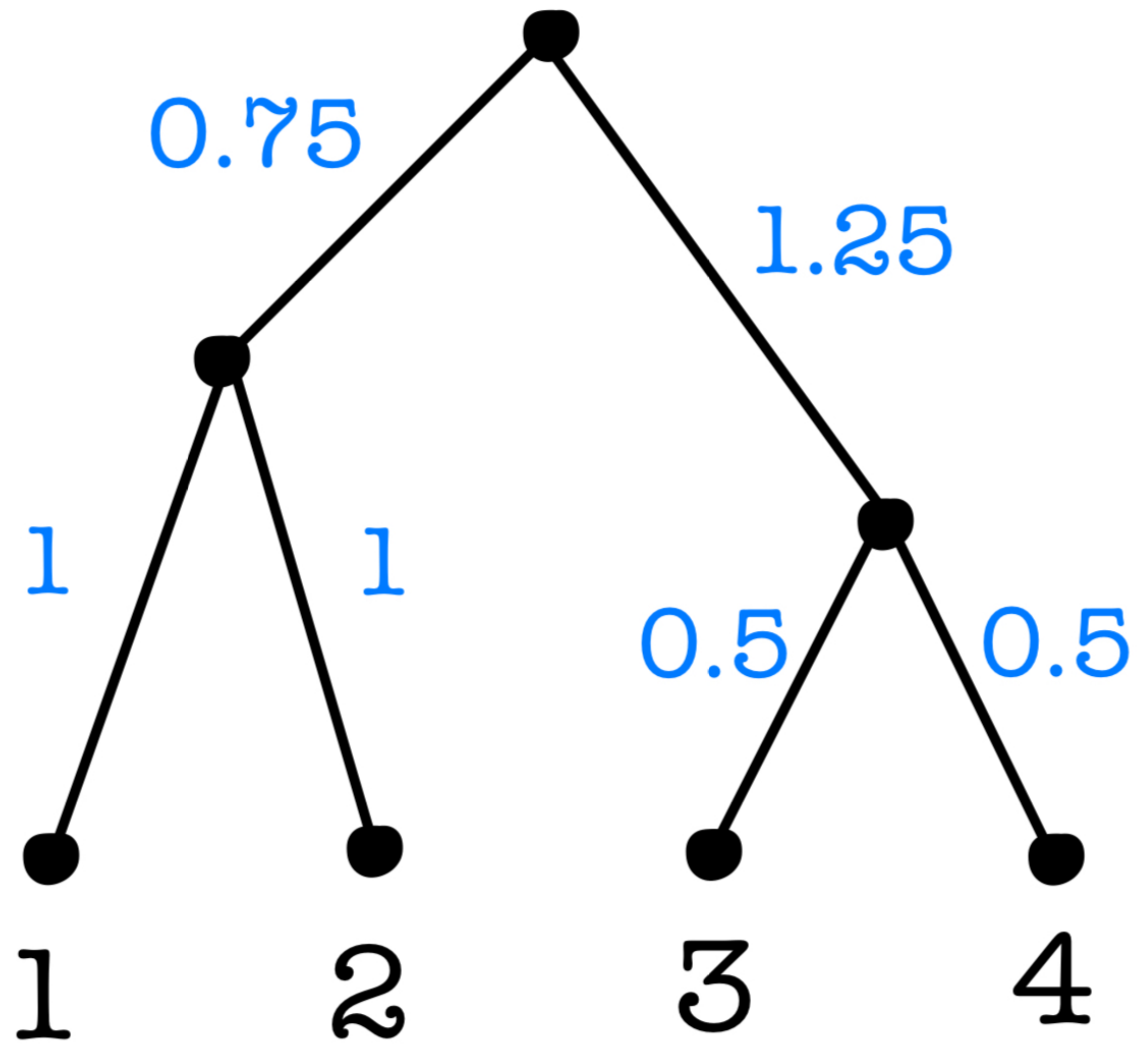
A **phylogenetic tree** is a rooted, leaf-labeled, $\mathbb{R}$-weighted tree, with no degree 2 nodes other than the root.



“Phylogenetic” Trees

A **phylogenetic tree** is a rooted, leaf-labeled, $\mathbb{R}$-weighted tree, with no degree 2 nodes other than the root.

The **topology** of a phylogenetic tree is the underlying unweighted leaf-labeled graph.



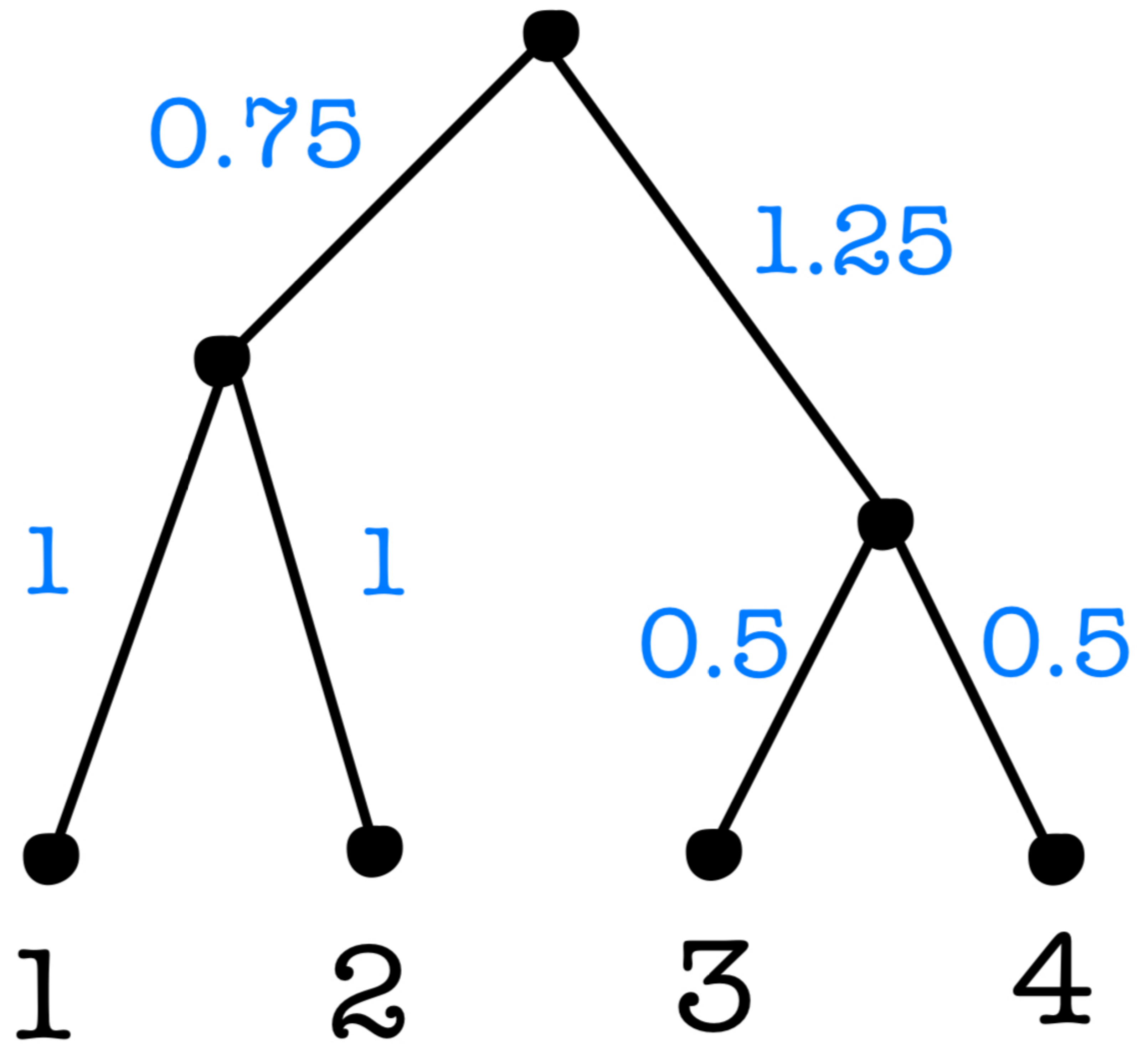
“Phylogenetic” Trees

A **phylogenetic tree** is a rooted, leaf-labeled, $\mathbb{R}$-weighted tree, with no degree 2 nodes other than the root.

The **topology** of a phylogenetic tree is the underlying unweighted leaf-labeled graph.

Key assumption: We assume our trees are **equidistant**, meaning the distance from the root to each leaf is the same.

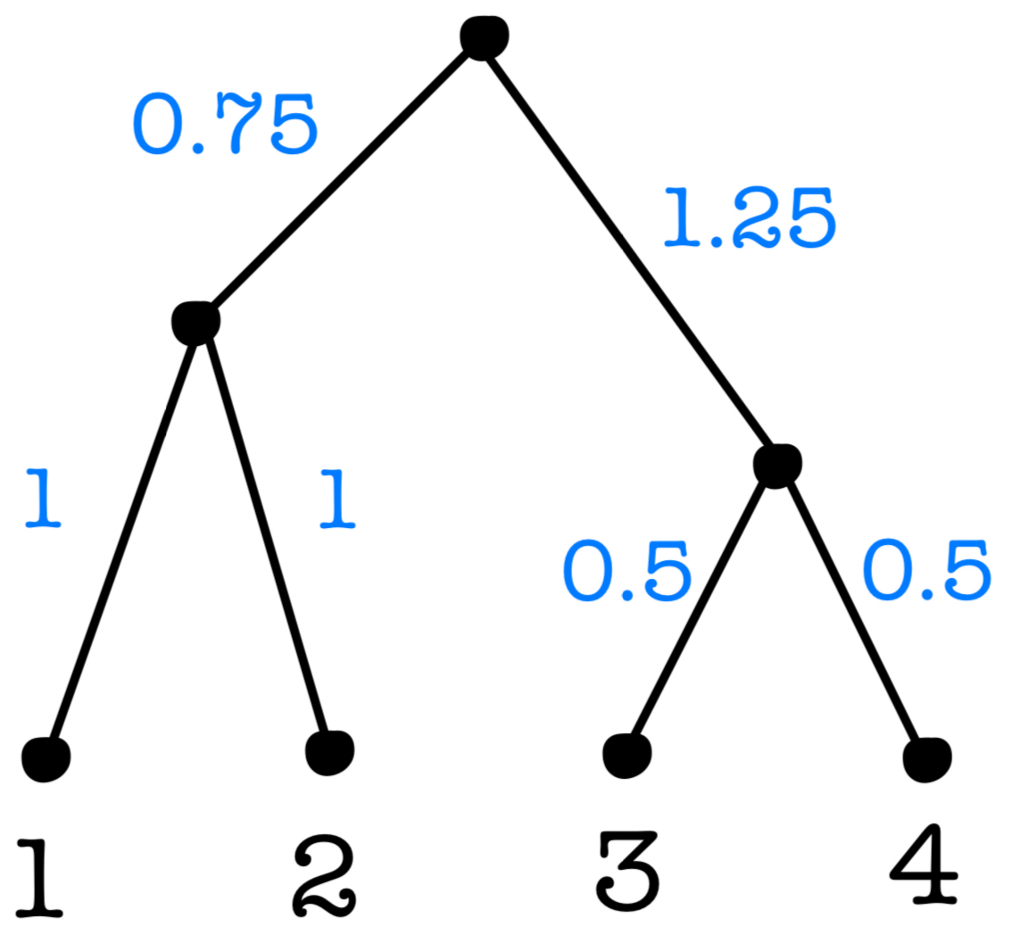
In phylogenetics, equidistant trees appear under the *molecular clock hypothesis,* which assumes that each lineage evolves at the same rate.



Tree Metric Embedding

T defines a **tree metric**: $\forall i, j \in [n]$

$d_T(i,j) =$ distance from i to j in T.



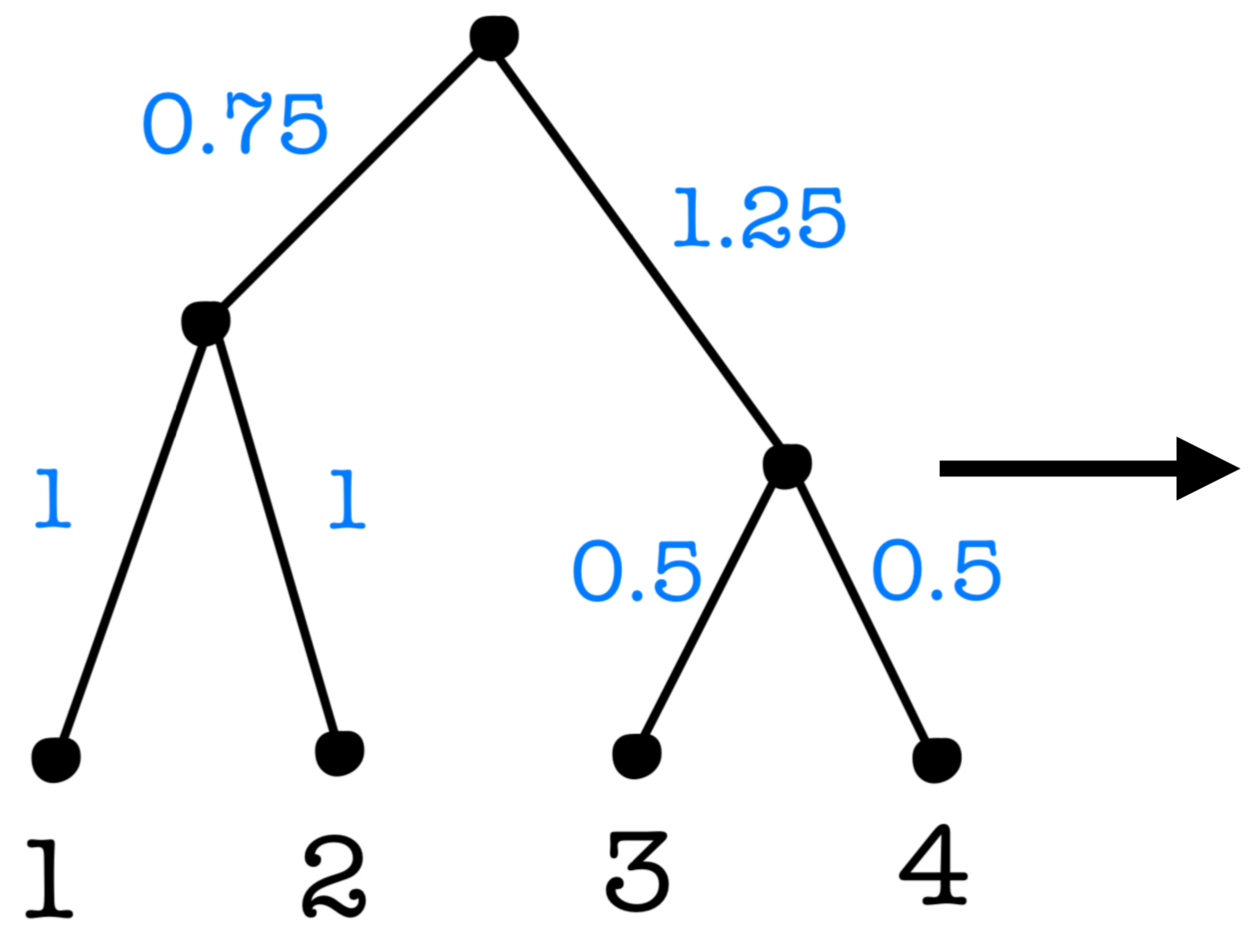
12 13 14 23 24 34

$d_T = (2,\ 3.5,\ 3.5,\ 3.5,\ 3.5,\ 1)$

Tree Metric Embedding

T defines a **tree metric**: $\forall i, j \in [n]$

$d_T(i,j)$ = distance from i to j in T.



12 13 14 23 24 34

$d_T = (2,\ 3.5,\ 3.5,\ 3.5,\ 3.5,\ 1)$

Identify T with

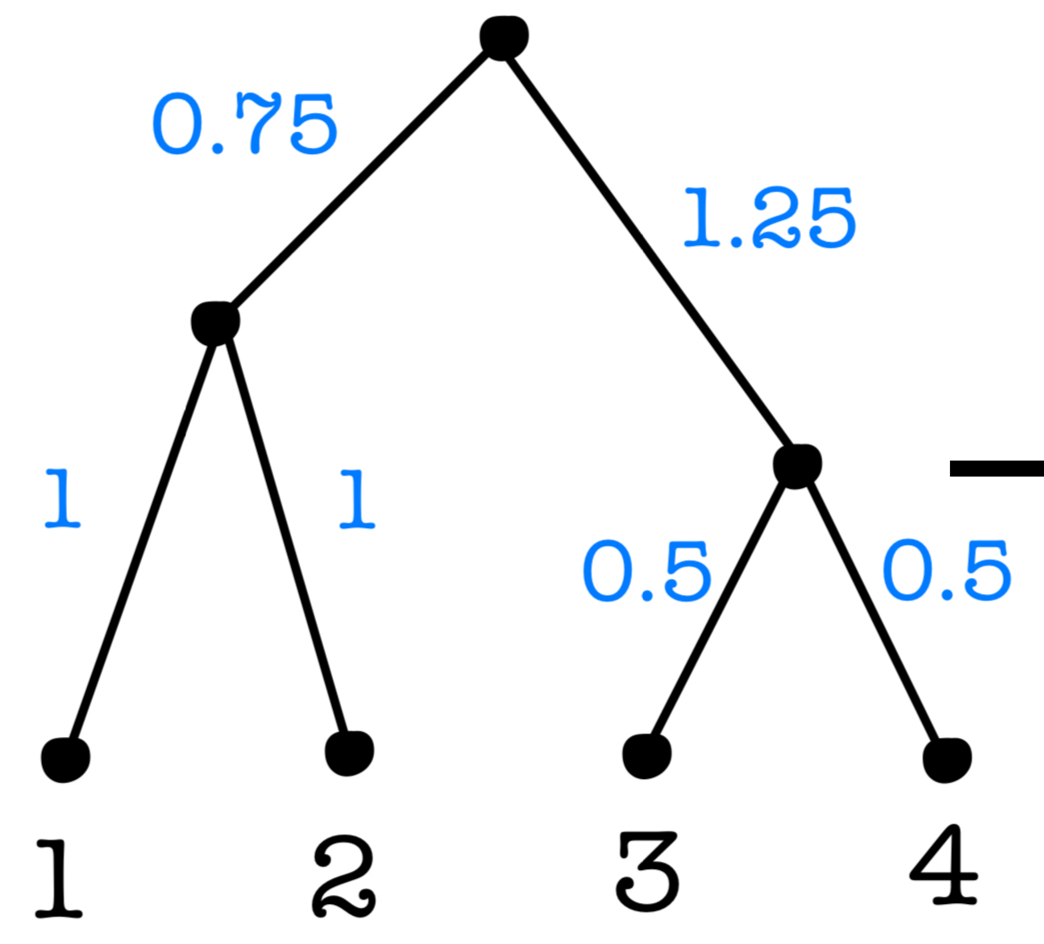
$(d_T(i,j))_{1 \le i < j \le n} \in \mathbb{R}^{\binom{n}{2}}$

Tree Metric Embedding

T defines a **tree metric**: $\forall i,j \in [n]$

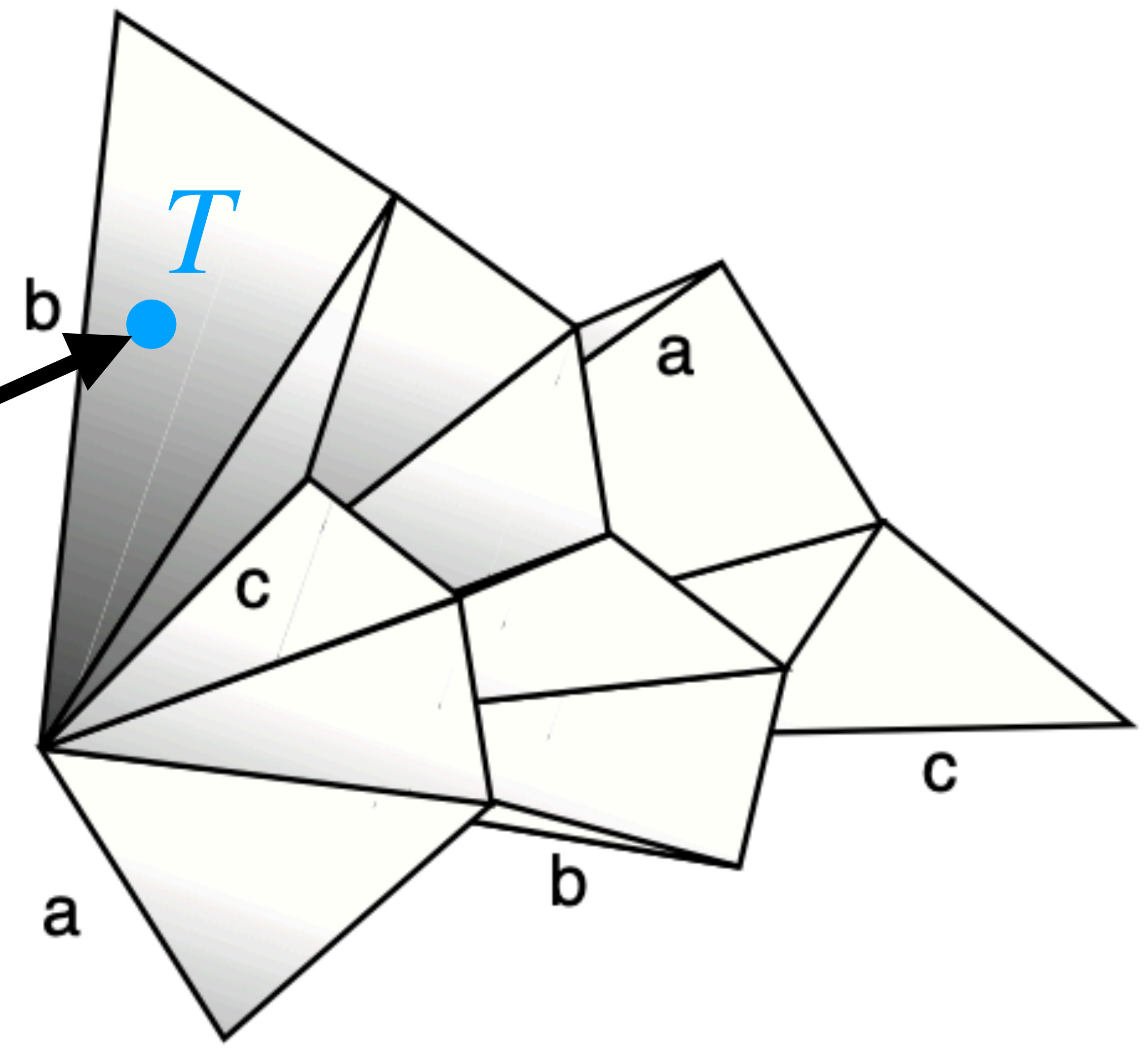
$d_T(i,j)$ = distance from i to j in T.

The set of all tree metrics on n leaves forms a polyhedral fan.



12 13 14 23 24 34

$d_T = (2,\ 3.5,\ 3.5,\ 3.5,\ 3.5,\ 1)$



BHV space for (unrooted) trees on 5 leaves [BHV].

Identify T with

$(d_T(i,j))_{1\leq i<j\leq n} \in \mathbb{R}^{\binom{n}{2}}$

Tropical Geometry

Tropical Tree Spaces

12	13	14	23	24	34
(2,	3.5,	3.5,	3.5,	3.5,	1)

What dependencies do the entries of d_T have?

Tropical Tree Spaces

$\oplus = \max,$
$\odot = +.$

$$\begin{matrix} 12 & 13 & 14 & 23 & 24 & 34 \\ (2, & 3.5, & 3.5, & 3.5, & 3.5, & 1) \end{matrix}$$

Three-point condition: A point $d \in \mathbb{R}^{\binom{n}{2}}$ is a tree metric if and only if

$$\max\{d(i,j), d(i,k), d(j,k)\}$$

is achieved at least twice for all $i, j, k \in [n]$.

These are tropical vanishing equations!

Tropical Tree Spaces

$\oplus = \max,$
$\odot = +.$

$$\begin{matrix} 12 & 13 & 14 & 23 & 24 & 34 \\ (2, & 3.5, & 3.5, & 3.5, & 3.5, & 1) \end{matrix}$$

Three-point condition: A point $d \in \mathbb{R}^{\binom{n}{2}}$ is a tree metric if and only if

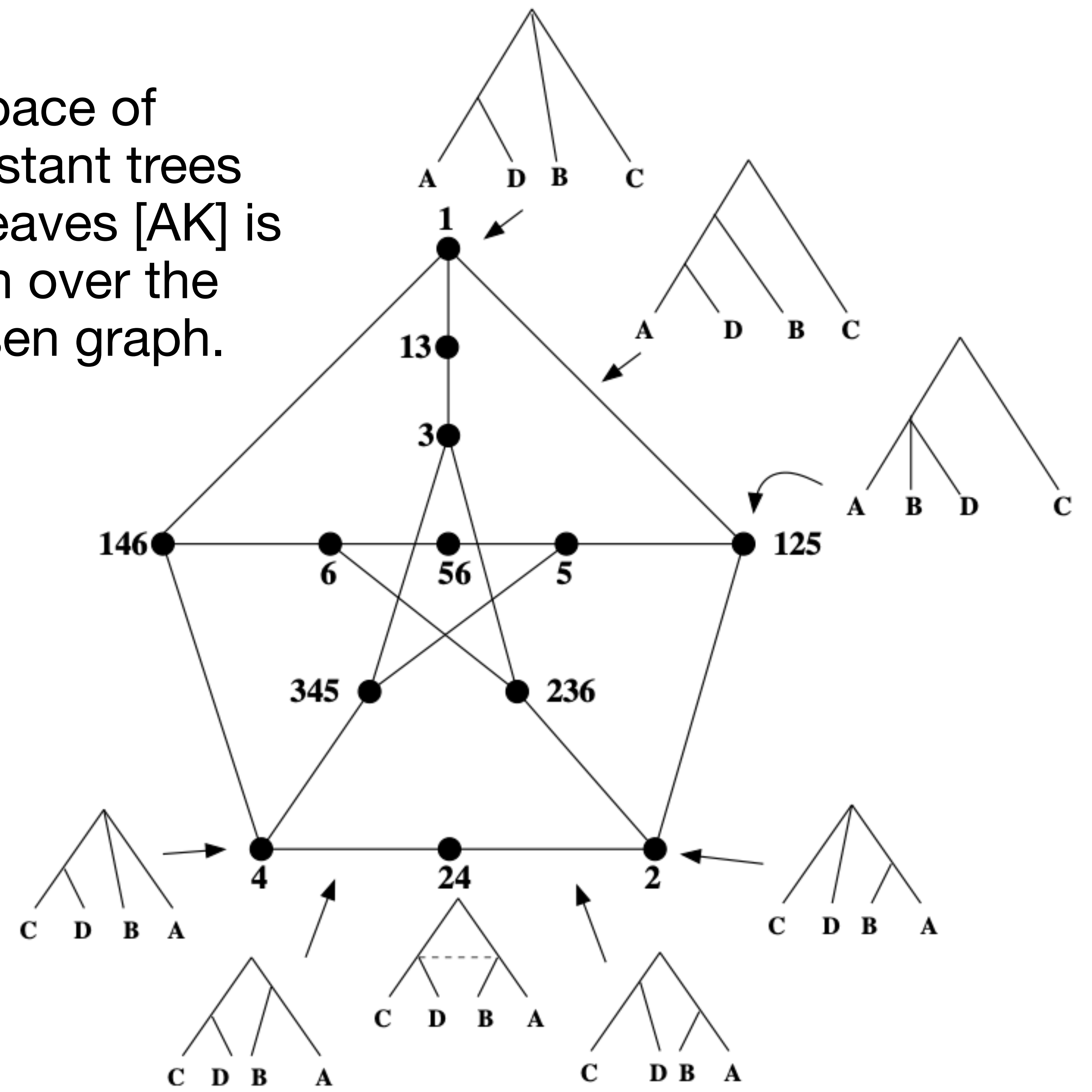
$$\max\{d(i,j), d(i,k), d(j,k)\}$$

is achieved at least twice for all $i, j, k \in [n]$.

These are tropical vanishing equations!

Theorem (Ardila-Klivans '06) The space of equidistant trees on n leaves is the *tropical linear space* $\text{Trop } M(K_n)$.

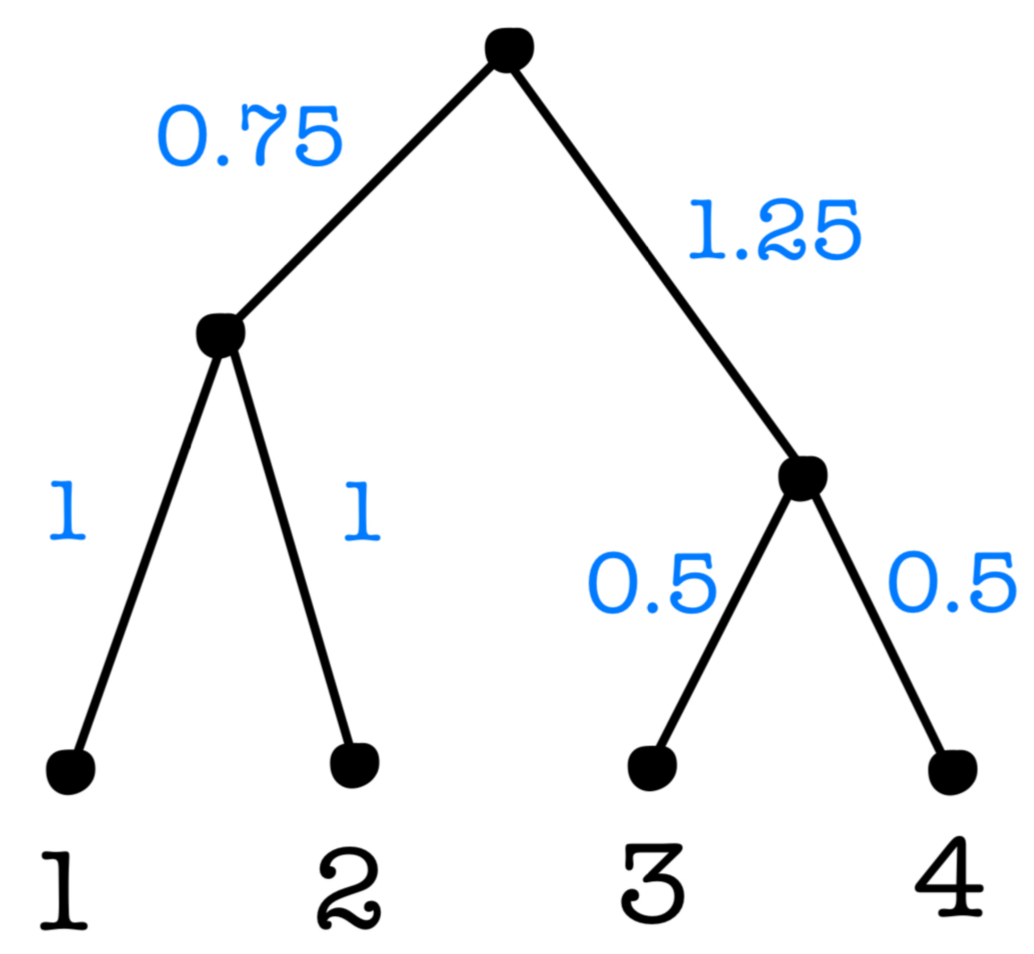
The space of equidistant trees on 4 leaves [AK] is the fan over the Petersen graph.



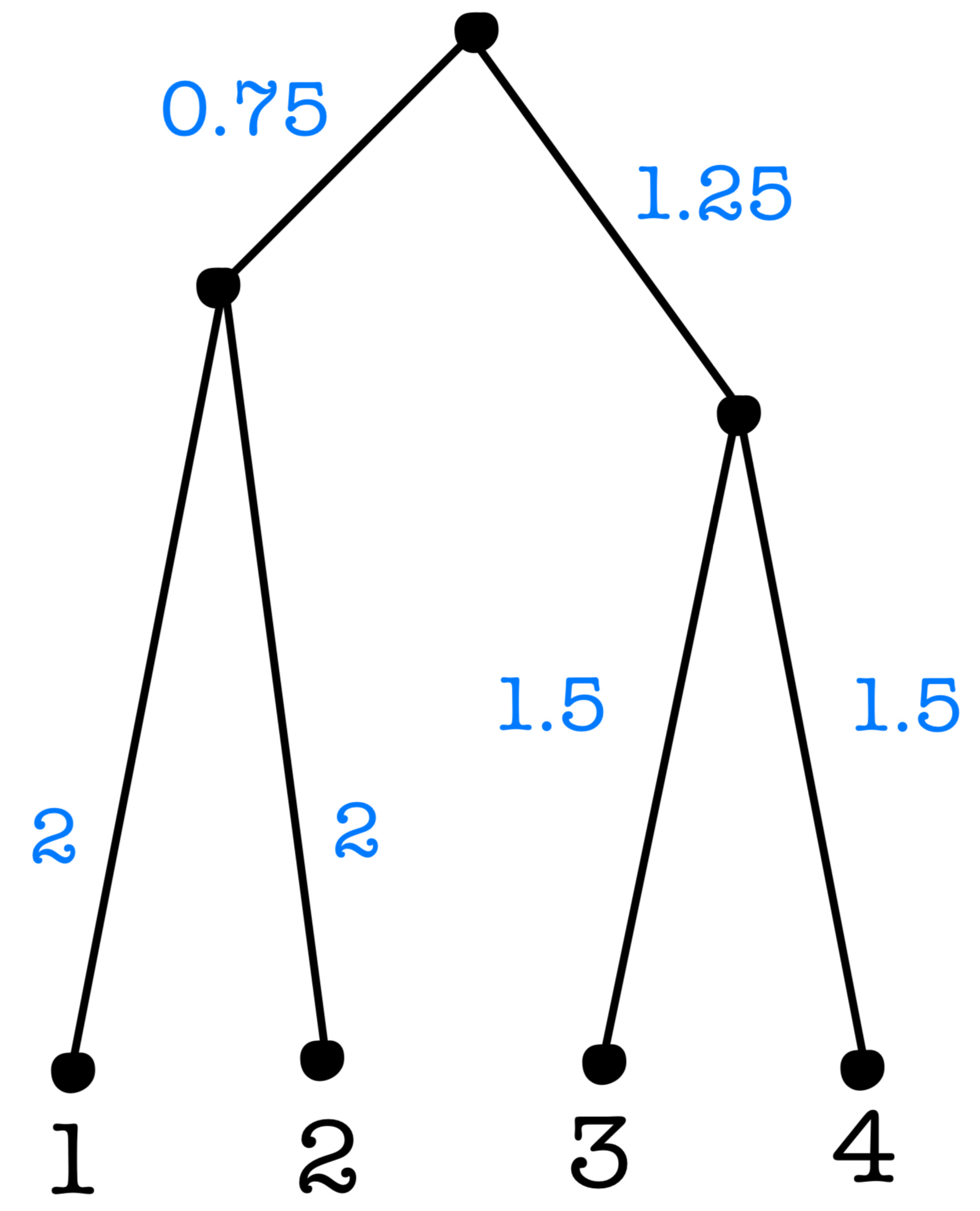
The Tropical Projective Torus $\mathbb{R}^n/\mathbb{R}\mathbf{1}$

$$x \sim y \iff x = \lambda \odot y,\ \exists \lambda \in \mathbb{R} \iff x = (\lambda + y_1, \ldots, \lambda + y_n),\ \exists \lambda \in \mathbb{R}$$

$\mathbb{R}^n/\mathbb{R}\mathbf{1} \cong \mathbb{R}^{n-1}$



$\sim$



12	13	14	23	24	34

$d_T = (2, 3.5, 3.5, 3.5, 3.5, 1) \sim d_T = (4, 5.5, 5.5, 5.5, 5.5, 3)$

12	13	14	23	24	34

Tropical Convexity

Tropical linear spaces are **tropically convex**, meaning they are closed under taking tropical linear combinations.

Exercise: the tropical sum (coordinate-wise max) of two equidistant tree metrics is an equidistant tree metric.

Tropical Convexity

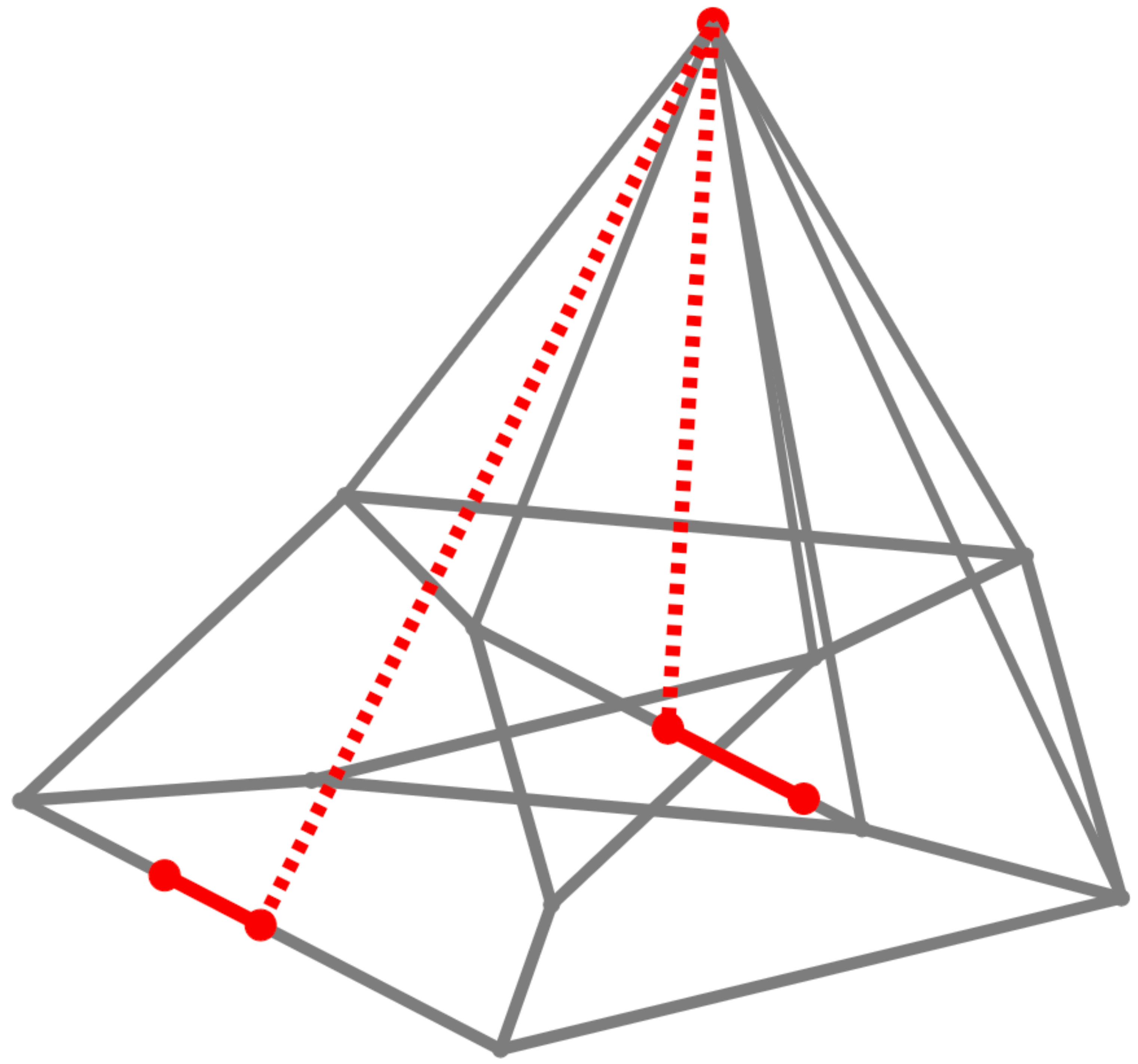
Tropical linear spaces are **tropically convex**, meaning they are closed under taking tropical linear combinations.

Exercise: the tropical sum (coordinate-wise max) of two equidistant tree metrics is an equidistant tree metric.

A **tropical line segment** is a *concatenation of Euclidean line segments.*

Theorem (Page-Yoshida-Zhang '20)

In the relative interior of each segment, all trees have the same topology.

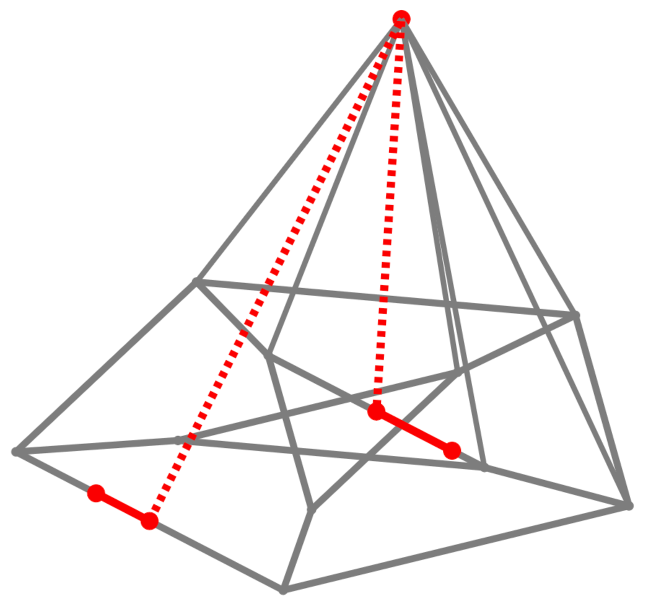


A tropical line segment in $\text{Trop } M(K_4)$, the space of equidistant trees on 4 leaves.

Tropical Line Segments

…in the space of equidistant trees.

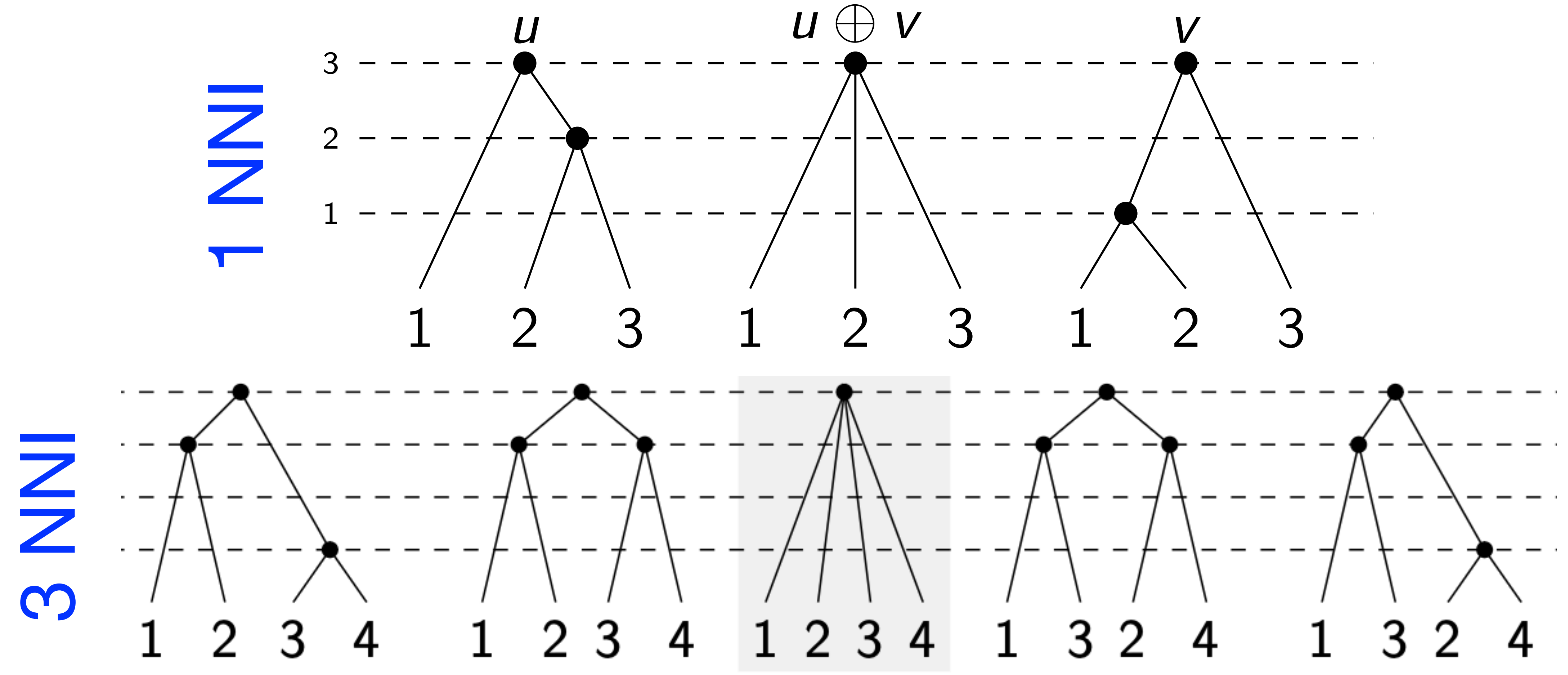
A tropical line segment is a *concatenation of Euclidean line segments.*



Theorem (C. ’22)

For sufficiently general edge weights, there are two possible topology changes along the tropical line segment:

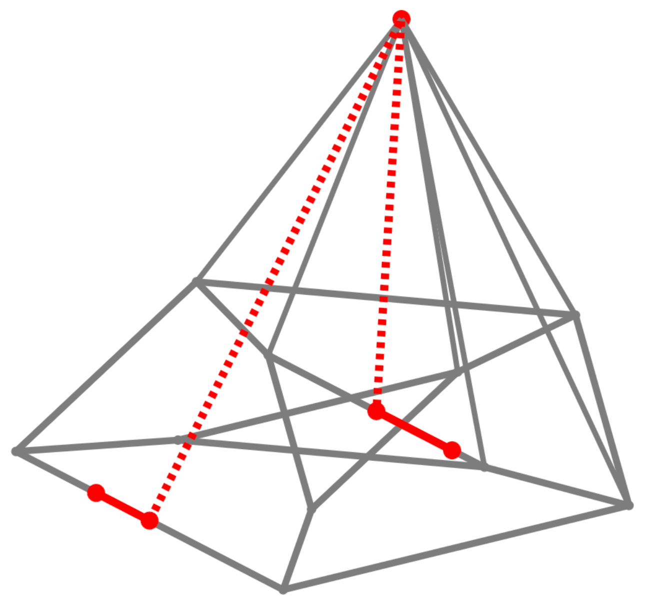
1. A single NNI (three branches come together)
2. A 4-clade rearrangement (four branches come together)



Tropical Line Segments

…in the space of equidistant trees.

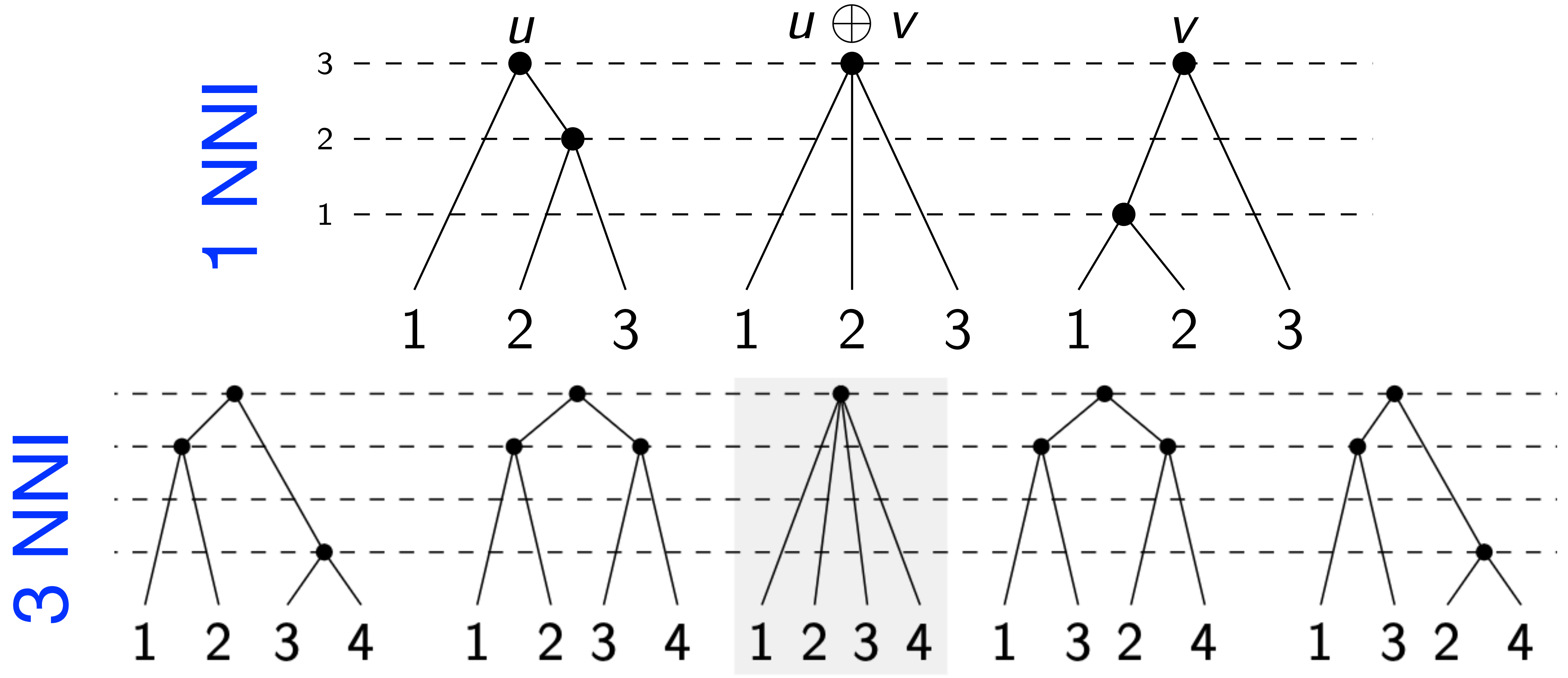
A tropical line segment is a *concatenation of Euclidean line segments.*



Theorem (C. ’22)

For sufficiently general edge weights, there are two possible topology changes along the tropical line segment:

1. A single NNI (three branches come together)
2. A 4-clade rearrangement (four branches come together)



How does this compare to other tree spaces?

BHV

- geodesics in polynomial time
  same as tropical line segments
- geodesics may pass through cones of arbitrary codimension
  codim $\leq$ 2 for tropical line segments

NNI (graph)

- shortest path is NP hard to compute
- average distance is $O(n \log n)$
  $O(n(\log n)^4)$ *NNI moves on a tropical line segment* **(C. ’22)**

A long tropical line segment

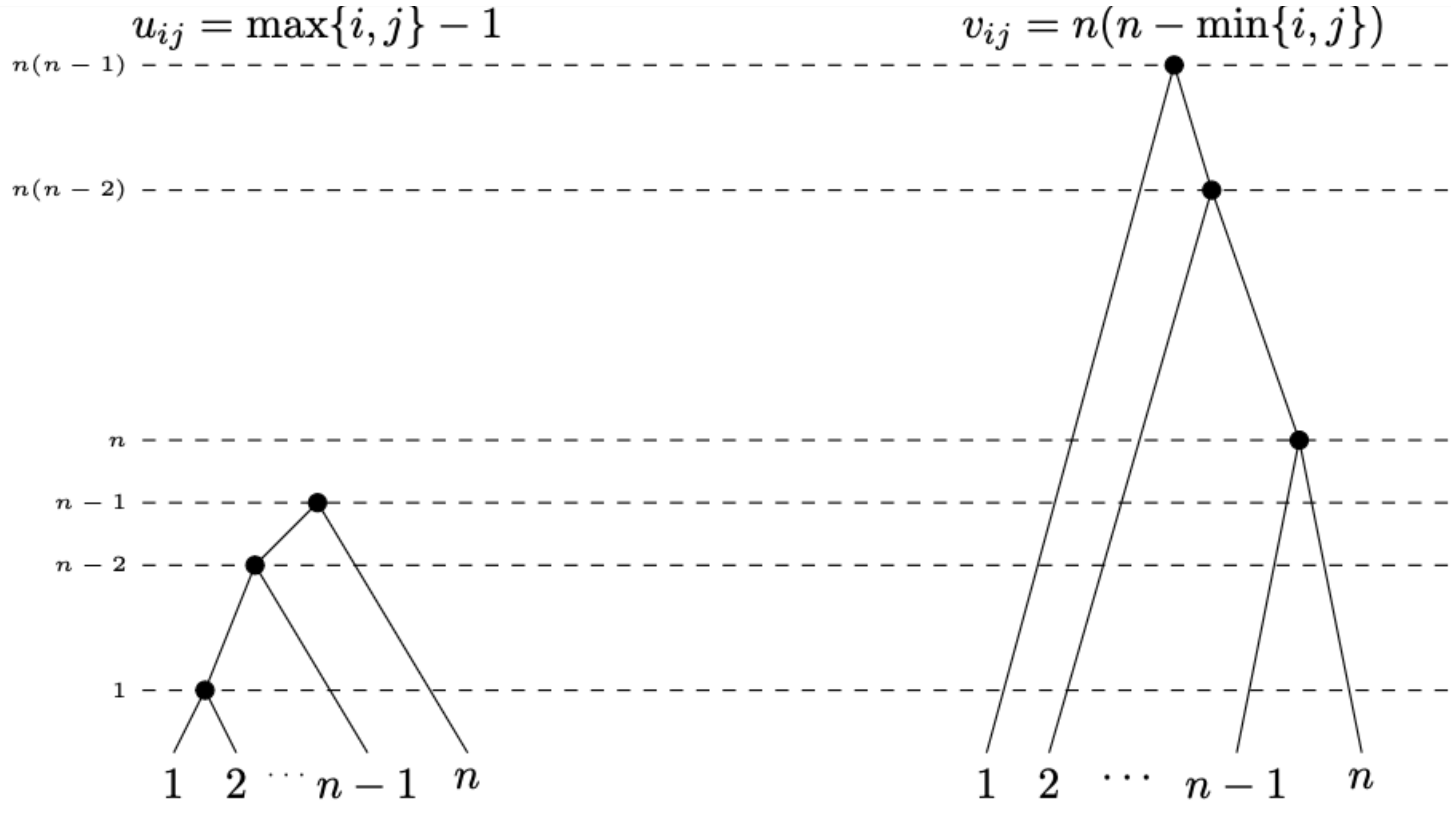
Not all tropical line segments in tree space are well-behaved…

The *tropical line segment* takes

$\binom{n-1}{2}$ NNI moves **(C. ’22)**.

But in the NNI graph, the trees are

$n-2$ NNI moves apart.

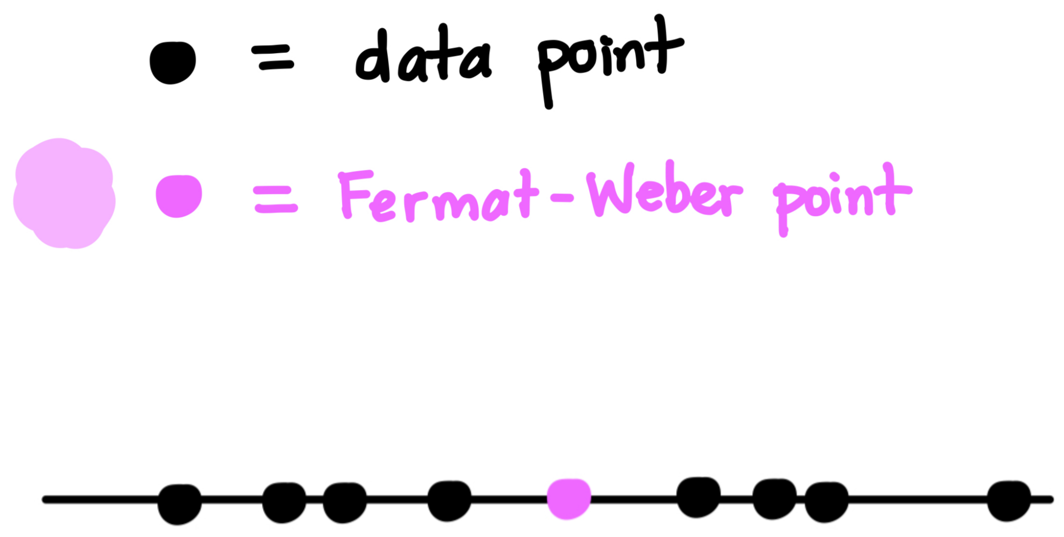


Tropical Fermat-Weber Points

Fermat-Weber Points

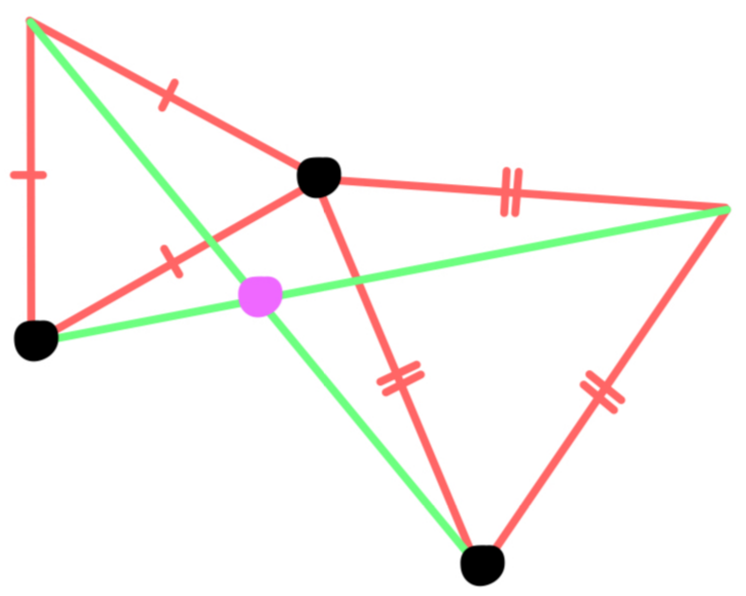
Given data $v_1, \ldots, v_m$ in a metric space (X, d), a **Fermat-Weber point** or **geometric median** is a point x^* minimizing the average distance to the data.

$$x^* \in \underset{x \in X}{\operatorname{argmin}} \sum_{i=1}^{m} d(v_i, x)$$



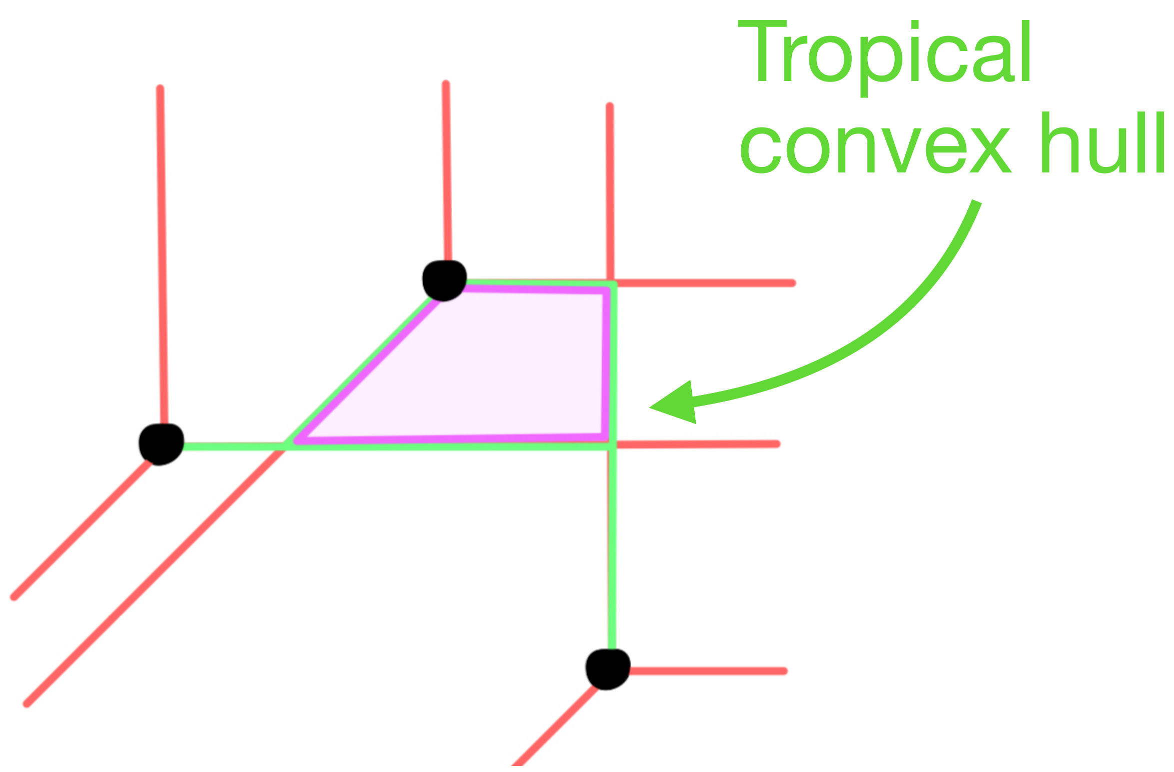
$X = \mathbb{R}, d(x,y) = |x - y|$

Median



$X = \mathbb{R}^2, d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$

Fermat-Torricelli point



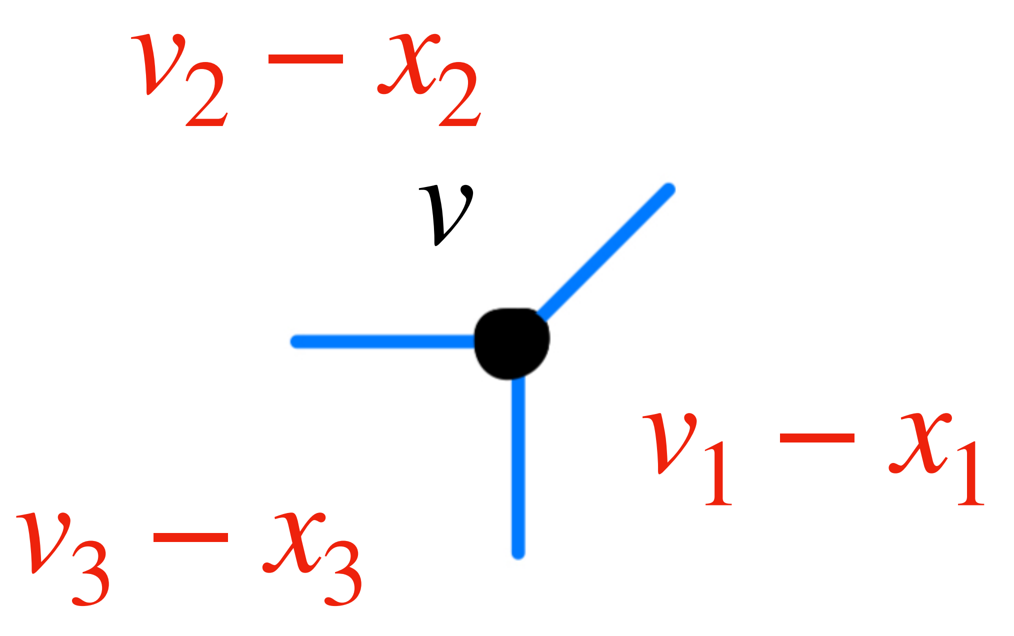
$X = \mathbb{R}^3/\mathbb{R}\mathbf{1}, d(x,y) = d_\Delta(x,y)$

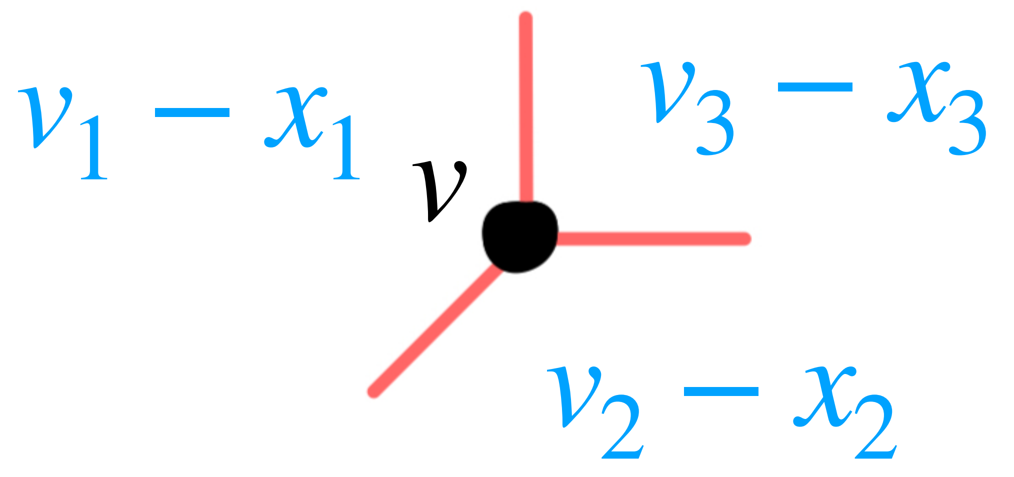
Asymmetric tropical FW set

Symmetric Tropical Fermat-Weber Sets

(Symmetric) tropical distance $v_i, x \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$

$$d_\Delta(v_i, x) = \max_{j\in[n]}(v_{ij} - x_j) - \min_{j\in[n]}(v_{ij} - x_j)$$

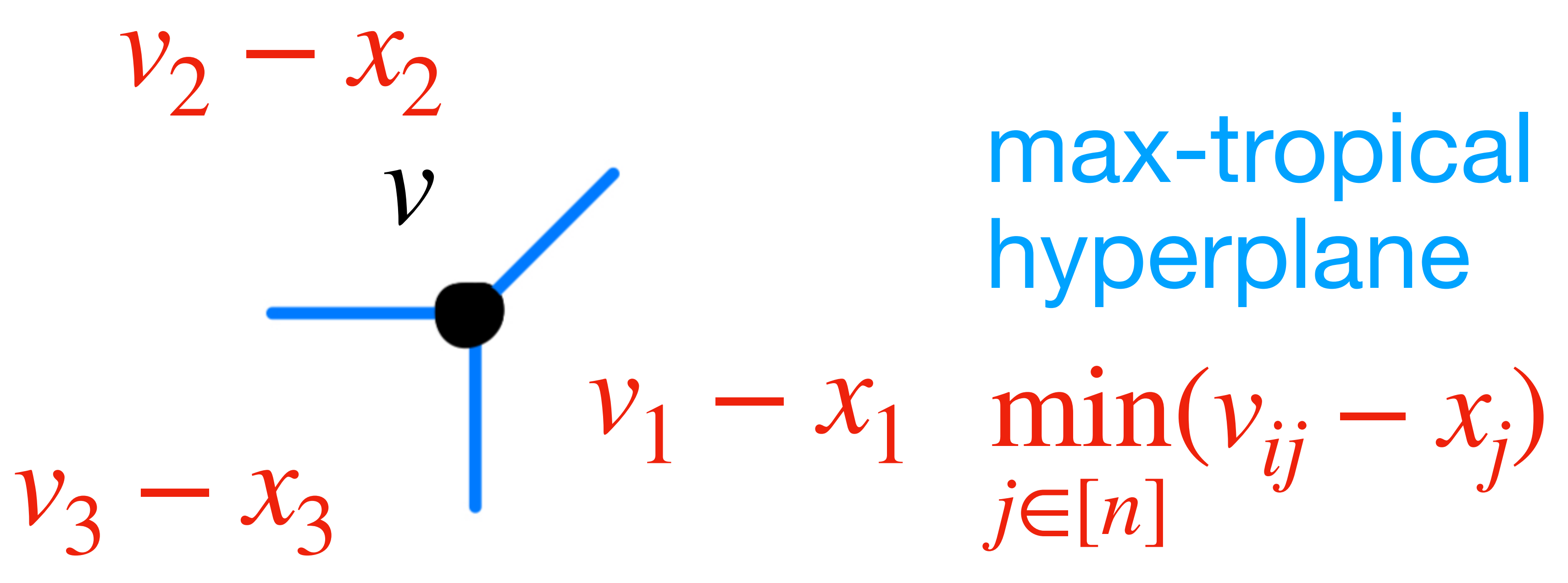




Symmetric Tropical Fermat-Weber Sets

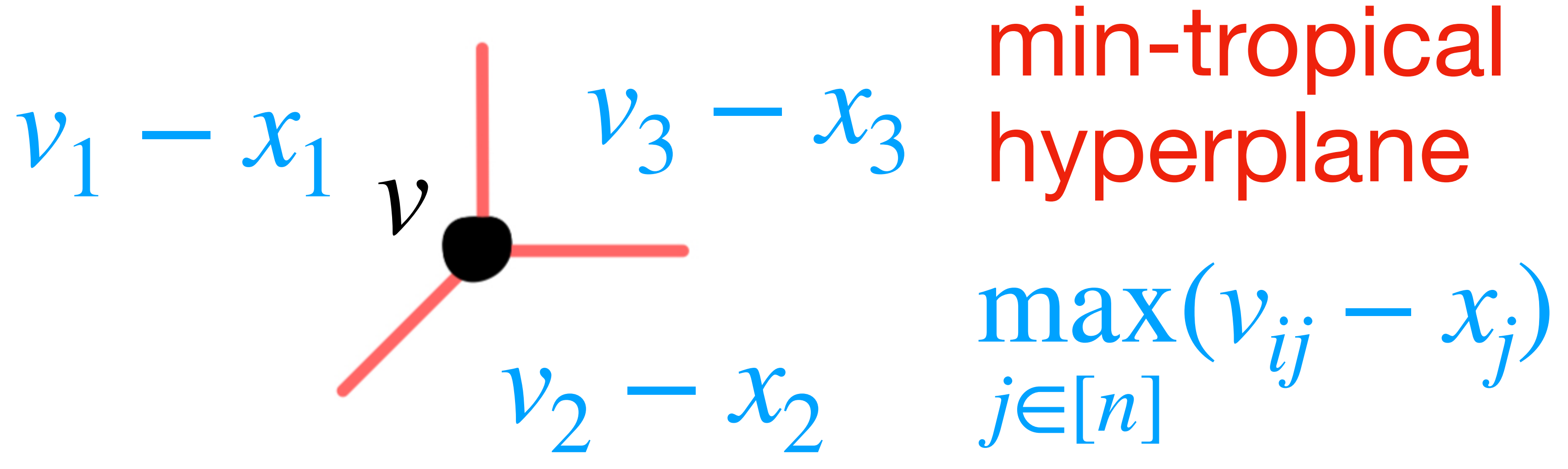
(Symmetric) tropical distance $v_i, x \in \mathbb{R}^n / \mathbb{R}\mathbf{1}$

$$d_\Delta(v_i, x) = \max_{j \in [n]}(v_{ij} - x_j) - \min_{j \in [n]}(v_{ij} - x_j)$$



- **Theorem (Lin et al. ’17)** In general, $\mathrm{FW}(v_1, \ldots, v_m) \nsubseteq \mathrm{tconv}(v_1, \ldots, v_m)$.

- **Theorem (Sabol et al. ’25+)** $\mathrm{FW}(v_1, \ldots, v_m)$ is a region of the **min- and max**-tropical hyperplane arrangement centered at $v_1, \ldots, v_m$.



- **Theorem (C.-Talbut-Sabol-Yoshida ’25+)** $\mathrm{FW}(v_1, \ldots, v_m) \cap \mathrm{tconv}(v_1, \ldots, v_m) \neq \varnothing$.

Symmetric Tropical Fermat-Weber Sets

(Symmetric) tropical distance $v_i, x \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$

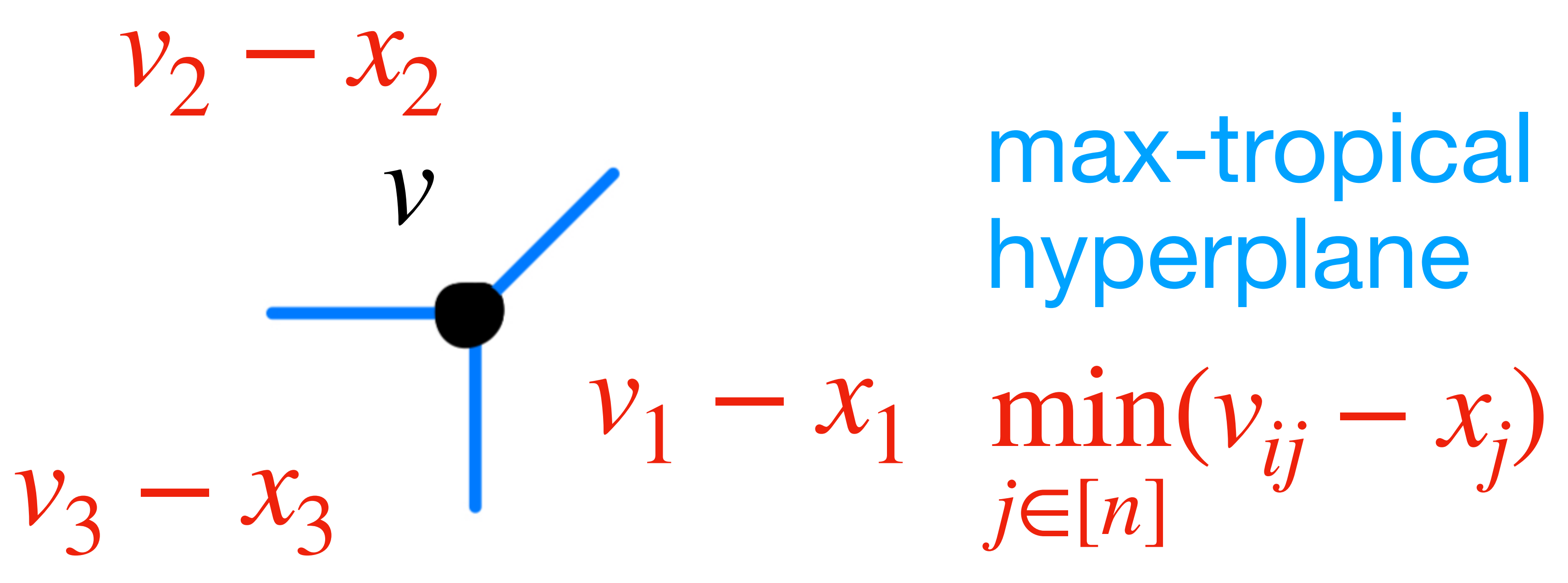
$$d_\Delta(v_i, x) = \max_{j\in[n]}(v_{ij} - x_j) - \min_{j\in[n]}(v_{ij} - x_j)$$

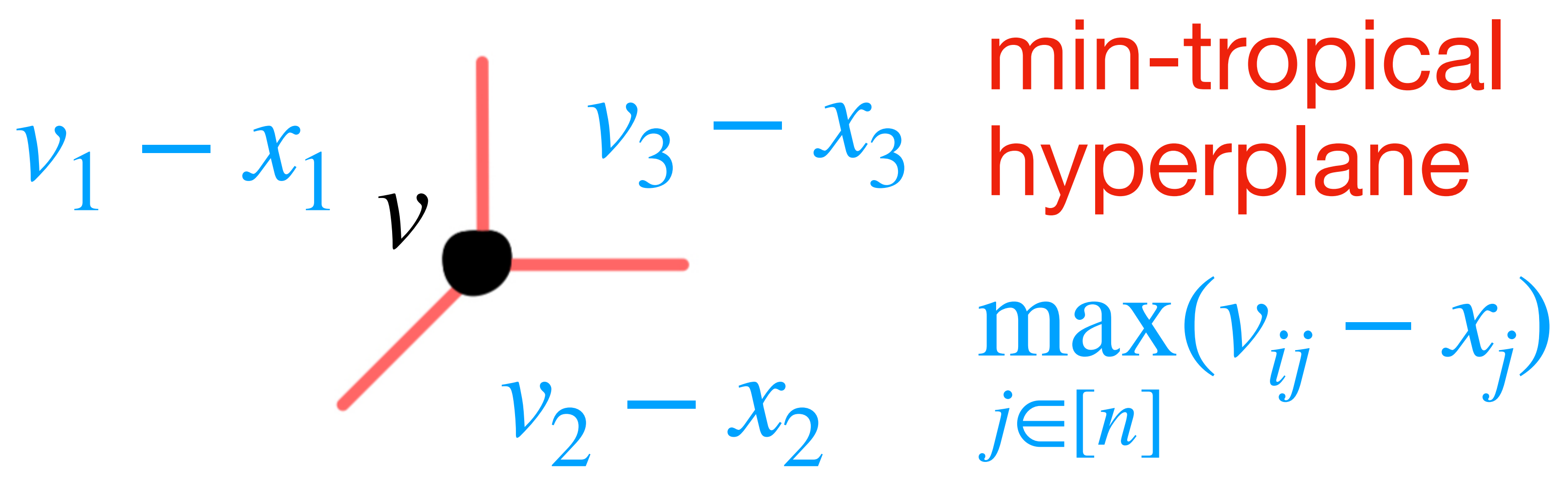
- **Theorem (Lin et al. '17)** In general, $\mathrm{FW}(v_1, \ldots, v_m) \nsubseteq \mathrm{tconv}(v_1, \ldots, v_m)$.

Median of tree metrics is not always a tree metric

- **Theorem (Sabol et al. '25+)** $\mathrm{FW}(v_1, \ldots, v_m)$ is a region of the **min- and max-**tropical hyperplane arrangement centered at $v_1, \ldots, v_m$.

- **Theorem (C.-Talbut-Sabol-Yoshida '25+)** $\mathrm{FW}(v_1, \ldots, v_m) \cap \mathrm{tconv}(v_1, \ldots, v_m) \neq \emptyset$.





Symmetric Tropical Fermat-Weber Sets

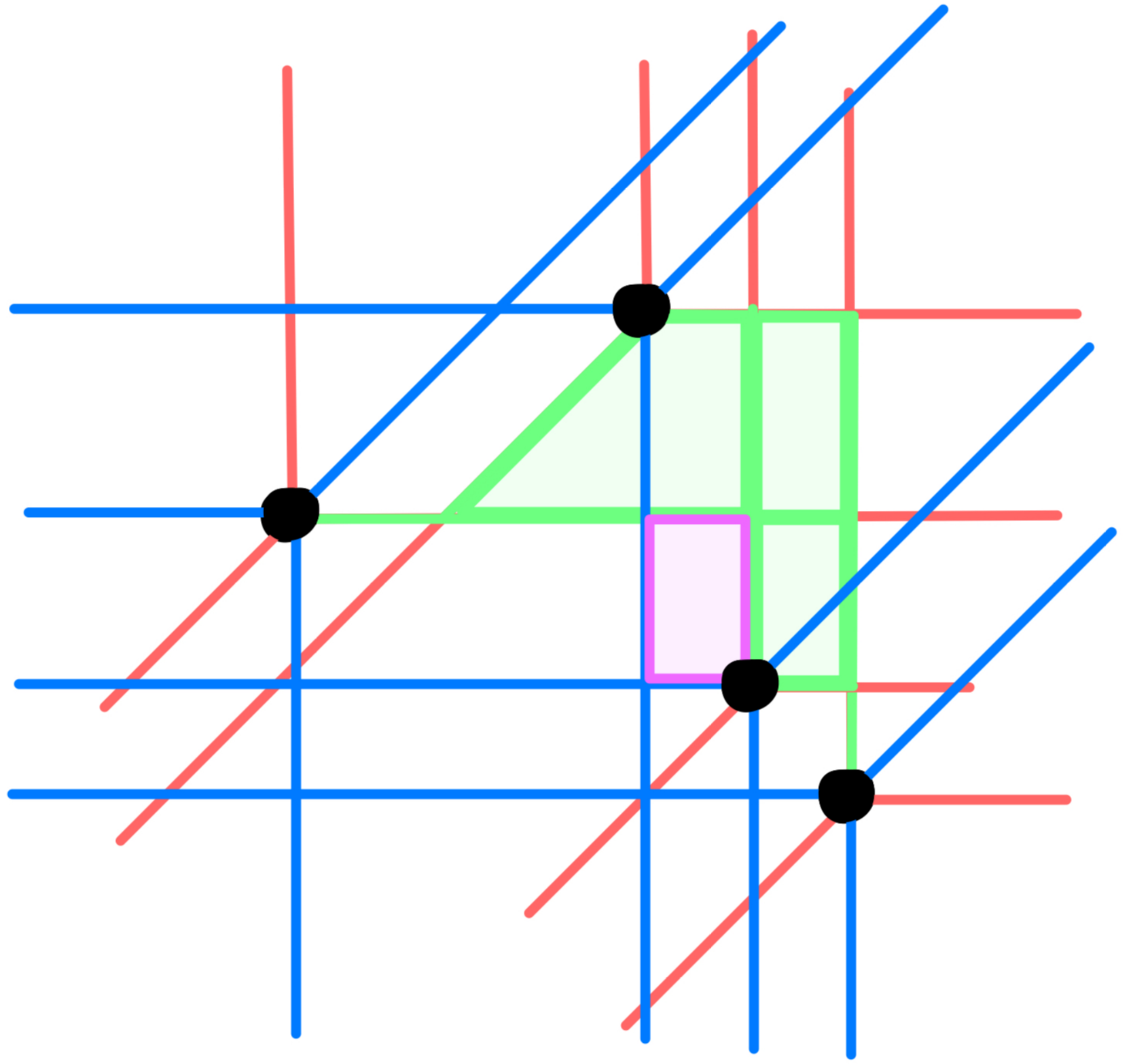
(Symmetric) tropical distance $v_i, x \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$

$$d_\Delta(v_i, x) = \max_{j \in [n]}(v_{ij} - x_j) - \min_{j \in [n]}(v_{ij} - x_j)$$

- **Theorem (Lin et al. '17)** In general, $\mathrm{FW}(v_1, \ldots, v_m) \nsubseteq \mathrm{tconv}(v_1, \ldots, v_m)$.

Median of tree metrics is not always a tree metric

- **Theorem (Sabol et al. '25+)** $\mathrm{FW}(v_1, \ldots, v_m)$ is a region of the **min- and max-**tropical hyperplane arrangement centered at $v_1, \ldots, v_m$.
- **Theorem (C.-Talbut-Sabol-Yoshida '25+)** $\mathrm{FW}(v_1, \ldots, v_m) \cap \mathrm{tconv}(v_1, \ldots, v_m) \neq \varnothing$.



Symmetric Tropical Fermat-Weber Sets

(Symmetric) tropical distance $v_i, x \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$

$$d_\Delta(v_i, x) = \max_{j\in[n]}(v_{ij} - x_j) - \min_{j\in[n]}(v_{ij} - x_j)$$

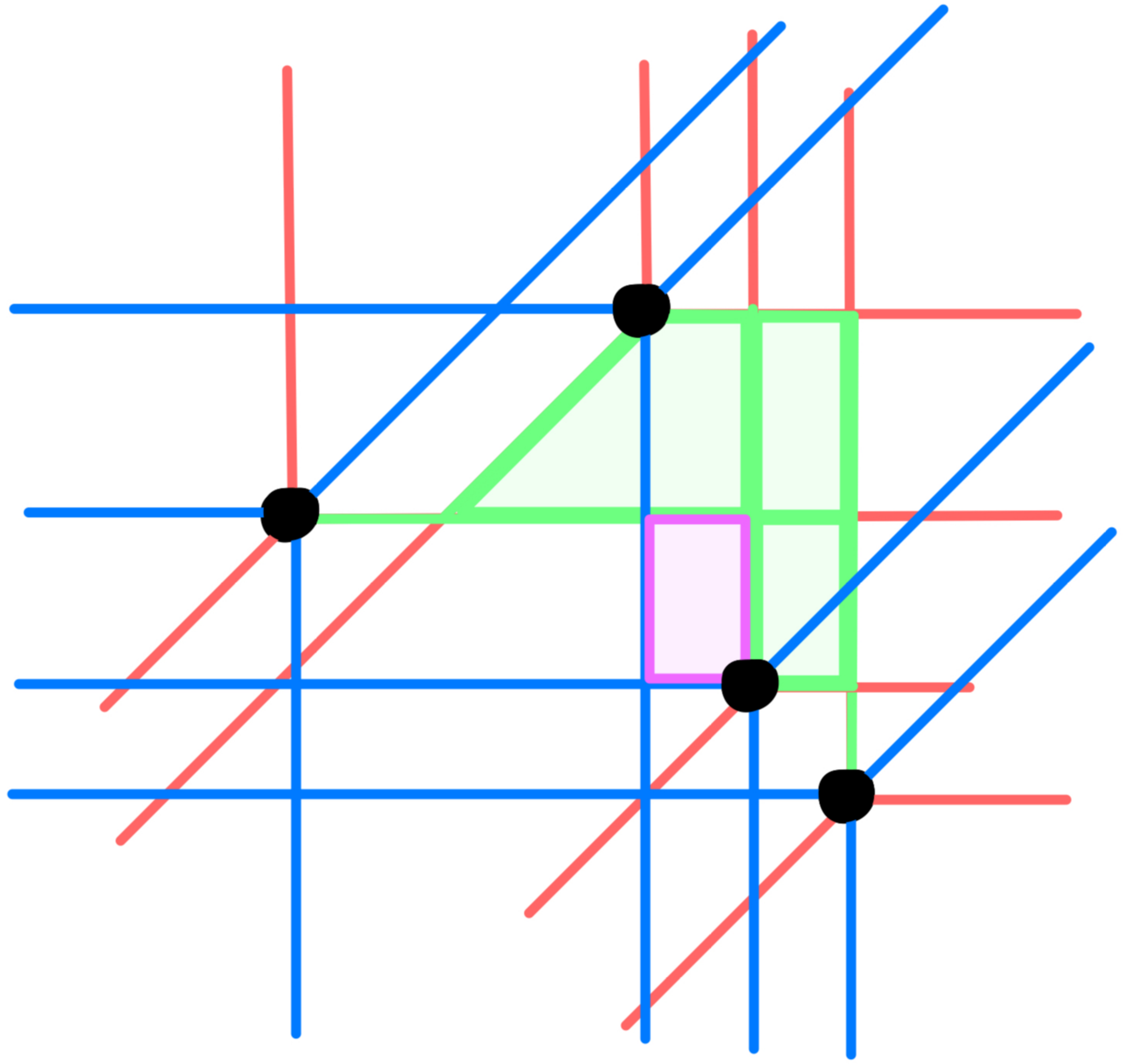
- **Theorem (Lin et al. '17)** In general, $\mathrm{FW}(v_1, \ldots, v_m) \not\subseteq \mathrm{tconv}(v_1, \ldots, v_m)$.

Median of tree metrics is not always a tree metric

- **Theorem (Sabol et al. '25+)** $\mathrm{FW}(v_1, \ldots, v_m)$ is a region of the **min- and max-**tropical hyperplane arrangement centered at $v_1, \ldots, v_m$.

- **Theorem (C.-Talbut-Sabol-Yoshida '25+)** $\mathrm{FW}(v_1, \ldots, v_m) \cap \mathrm{tconv}(v_1, \ldots, v_m) \neq \emptyset$.

But some median is a tree metric



Asymmetric Tropical Fermat-Weber Sets

Asymmetric tropical distance (Comaneci-Joswig 2024) $v_i, x \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$

$$d_\Delta(v_i, x) = n \max_{j\in[n]}(v_{ij} - x_j) + \sum_{j\in[n]} x_j - v_{ij}\,.$$

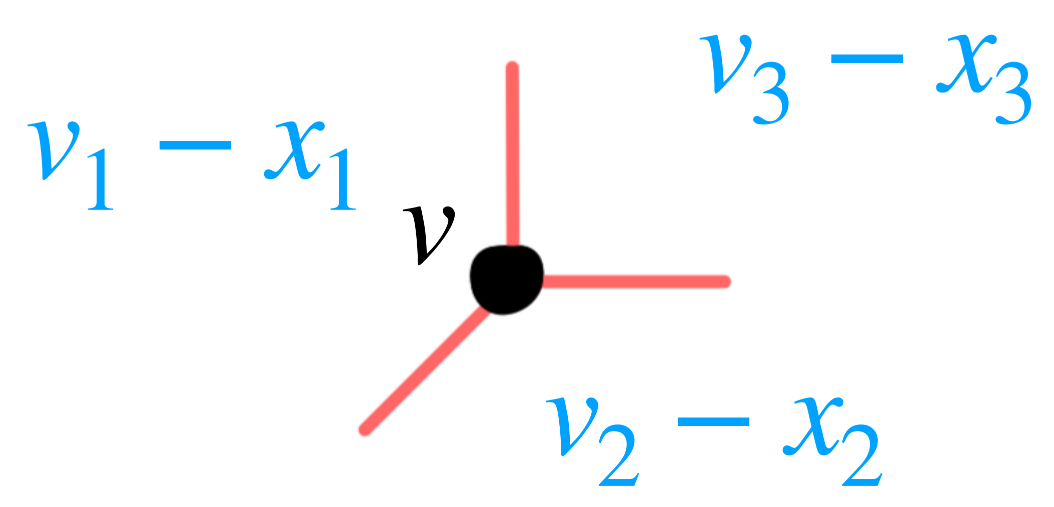
Asymmetric Tropical Fermat-Weber Sets

Asymmetric tropical distance (Comaneci-Joswig 2024) $v_i, x \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$

$$d_\Delta(v_i, x) = n \max_{j\in[n]}(v_{ij} - x_j) + \sum_{j\in[n]} x_j - v_{ij}\,. \quad = n \max_{i\in[n]}(v_{ij} - x_i), \text{ for}$$

$$x, v_i \in \{z \in \mathbb{R}^n : z_1 + \cdots + z_n = 0\} \cong \mathbb{R}^n/\mathbb{R}\mathbf{1}\,.$$

$d_\Delta(v, x)$ is linear on the regions of the min-tropical hyperplane centred at v.



Asymmetric Tropical Fermat-Weber Sets

Asymmetric tropical distance (Comaneci-Joswig 2024) $v_i, x \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$

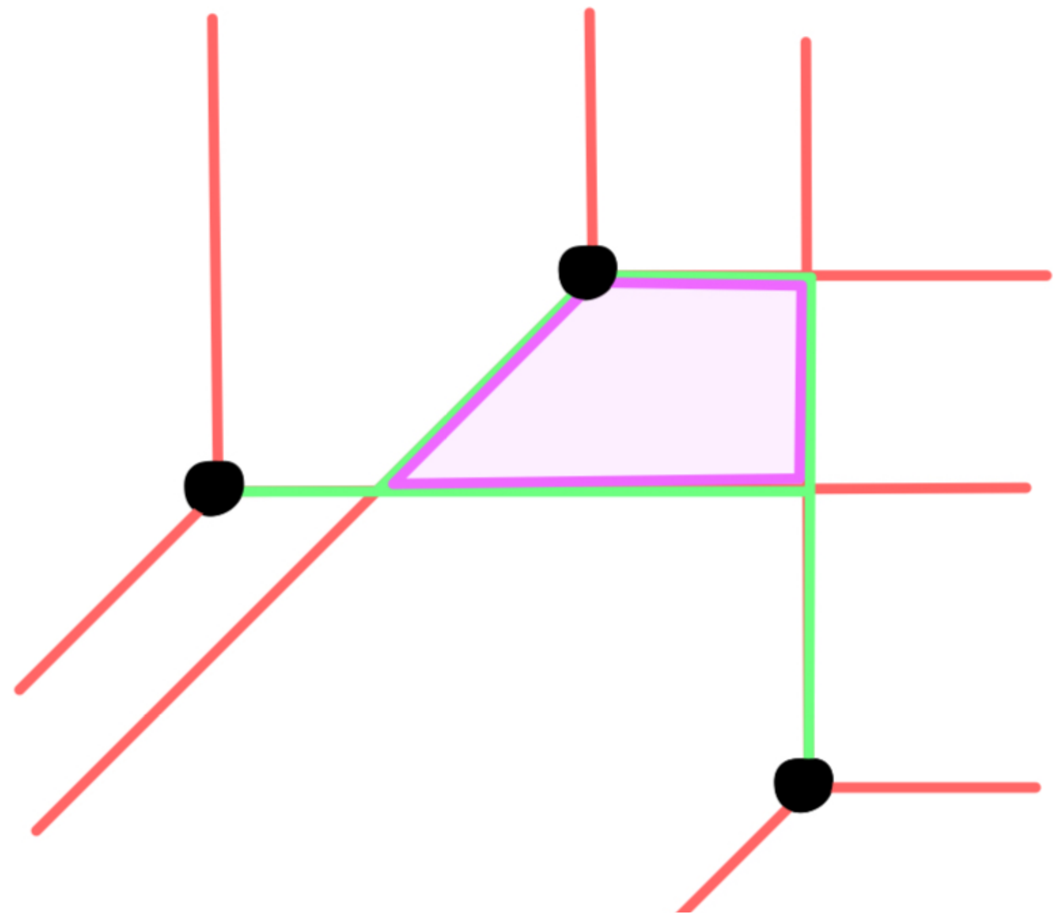
$$d_\Delta(v_i, x) = n \max_{j\in[n]}(v_{ij} - x_j) + \sum_{j\in[n]} x_j - v_{ij}\,. \quad = n \max_{i\in[n]}(v_{ij} - x_i), \text{ for}$$

$$x, v_i \in \{z \in \mathbb{R}^n : z_1 + \cdots + z_n = 0\} \cong \mathbb{R}^n/\mathbb{R}\mathbf{1}\,.$$

Theorem (Comaneci-Joswig '24)

- $\mathrm{FW}(v_1, \ldots, v_m)$ is the *central cell* of the min-tropical hyperplane arrangement centered at $v_1, \ldots, v_m$.
- $\mathrm{FW}(v_1, \ldots, v_m) \subseteq \mathrm{tconv}^{\max}(v_1, \ldots, v_m)\,.$
- If $v_1, \ldots, v_m$ represent equidistant trees $T_1, \ldots, T_m$, then every FW point is also an equidistant tree.

$d_\Delta(v, x)$ is linear on the regions of the min-tropical hyperplane centred at v.



(weighted)

Asymmetric Tropical Fermat-Weber Sets

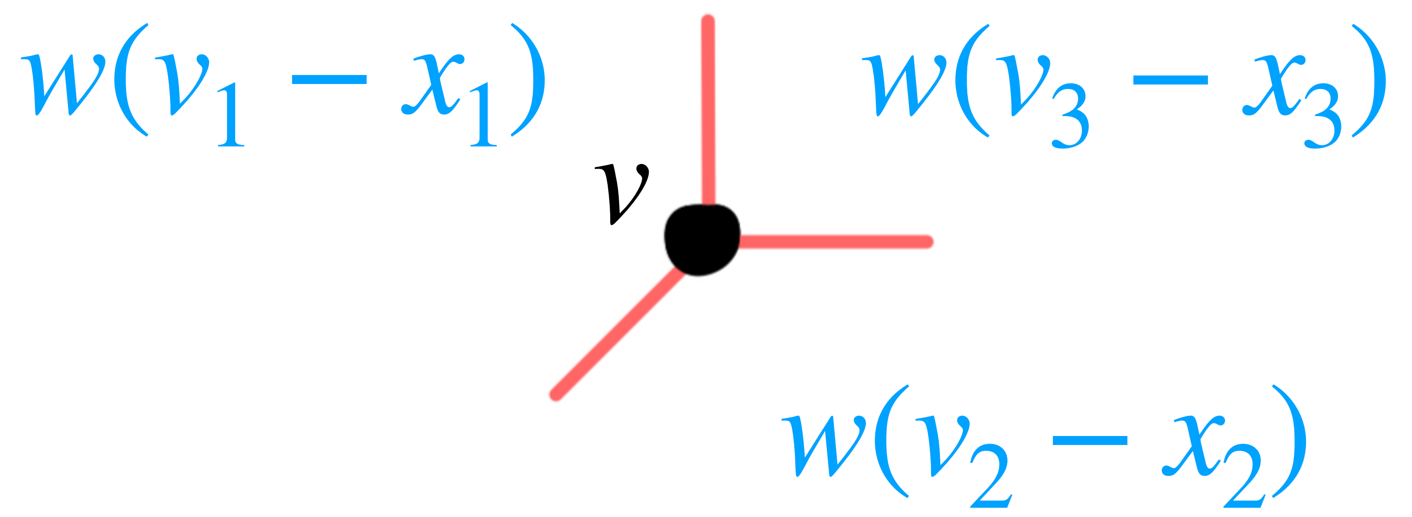
Weighted Fermat-Weber Points

Given weights $w_1, \ldots, w_m > 0$, $\quad x^* \in \operatorname{argmin}_{x \in X} \sum_{i=1}^{m} w_i \, d_\Delta(v_i, x)$.

Theorem (C.-Curiel 2025+)

- $\mathrm{FW}(v_1, \ldots, v_m, \mathbf{w})$ is a bounded cell* of the min-tropical hyperplane arrangement centered at $v_1, \ldots, v_m$.
  *generically a point
- $\mathrm{FW}(v_1, \ldots, v_m, \mathbf{w}) \subseteq \mathrm{tconv}^{\max}(v_1, \ldots, v_m)$.
- If $v_1, \ldots, v_m$ represent equidistant trees $T_1, \ldots, T_m$, then every weighted FW point is also an equidistant tree.

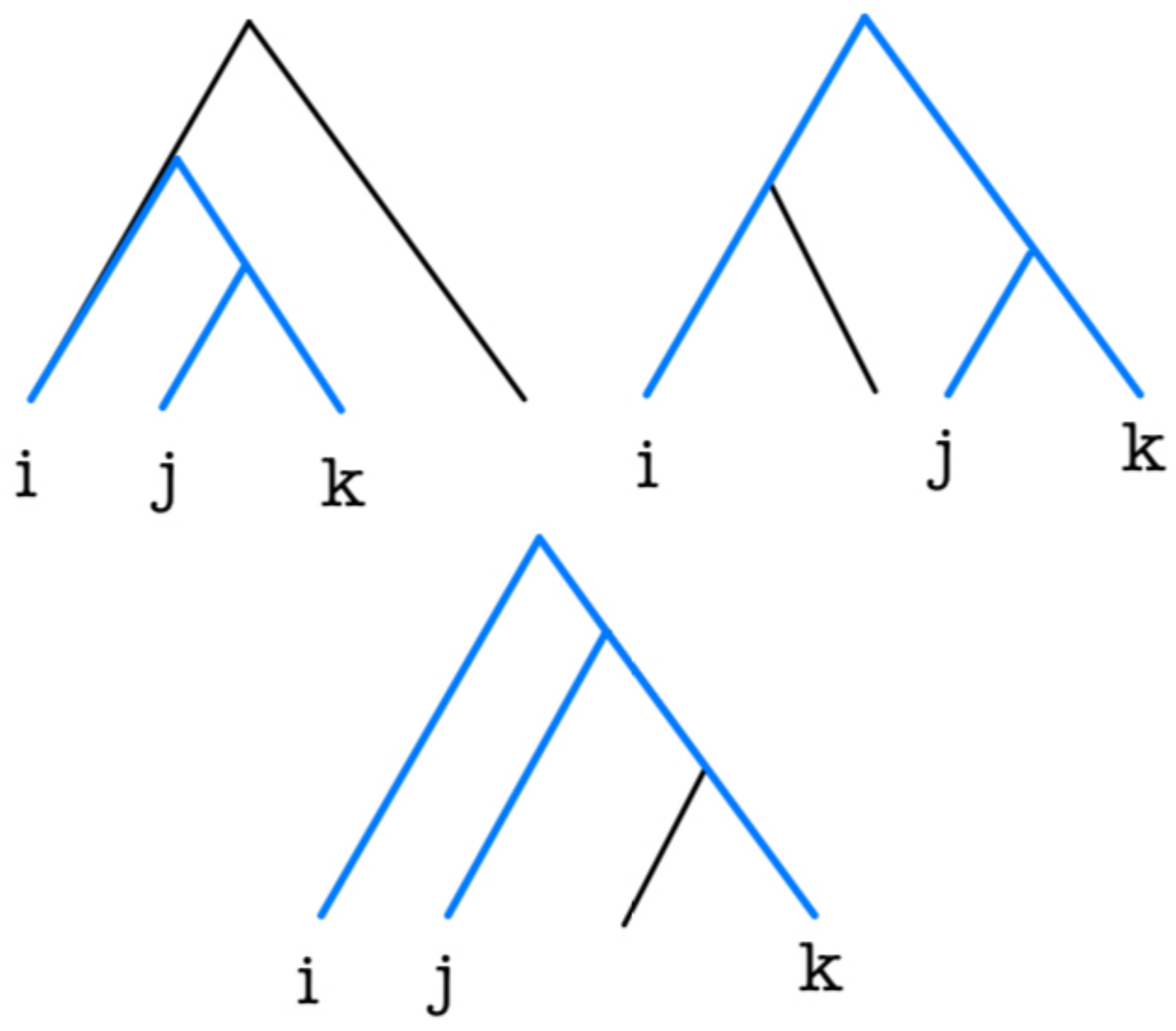
$w \cdot d_\Delta(v, x)$ is linear on the regions of the tropical hyperplane centred at v.



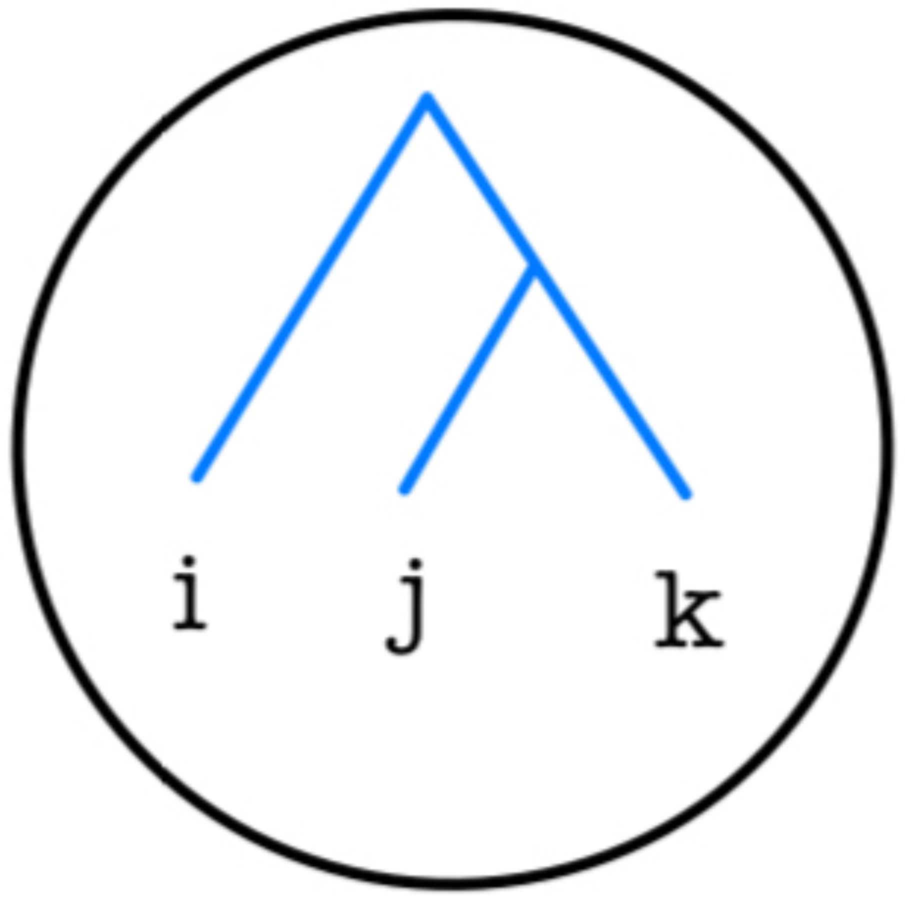
Good trees, quickly

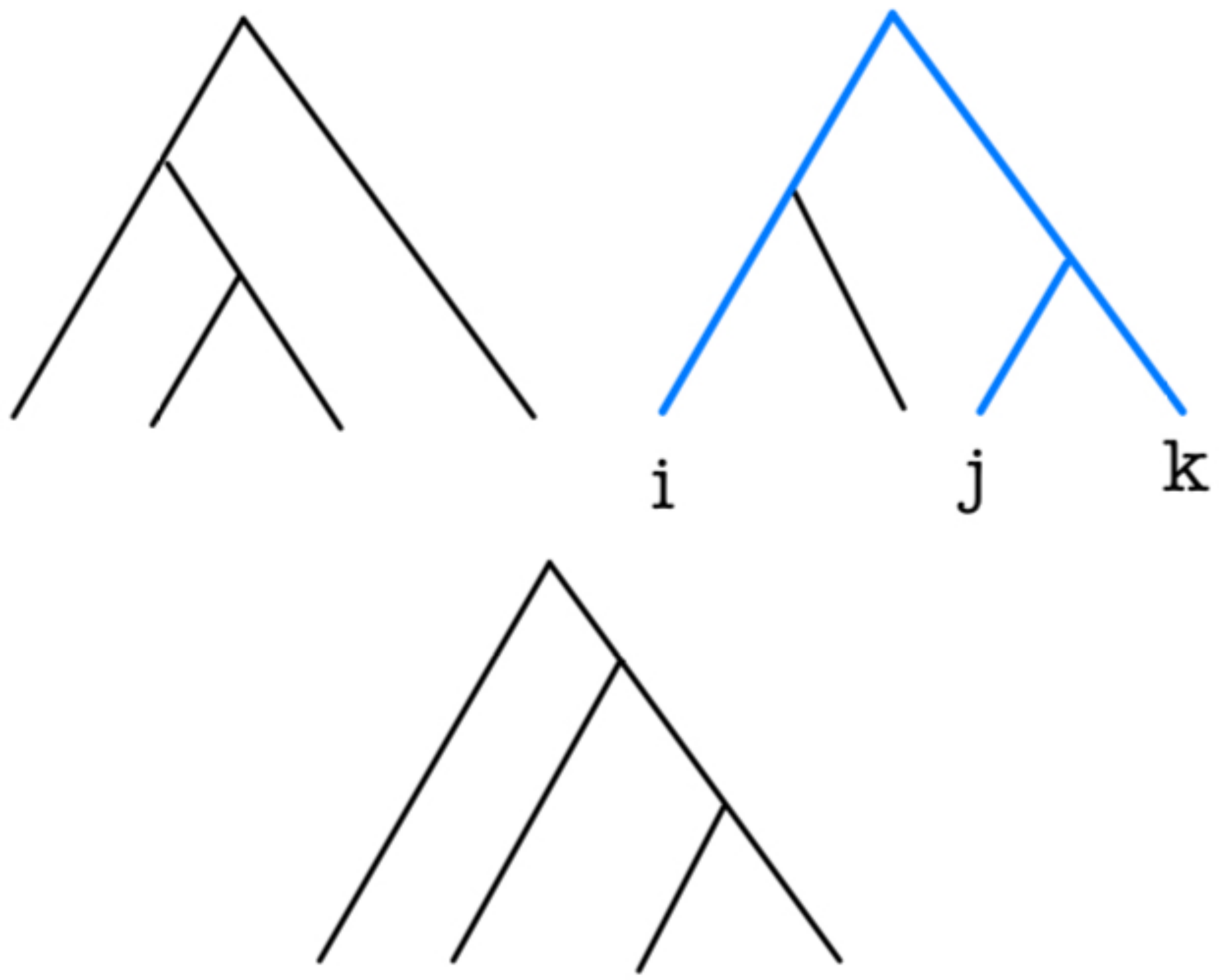
Theorem (Comaneci-Joswig '24) Every point in $\mathrm{tconv}^{\max}(T_1, \ldots, T_m)$ is Pareto and co-Pareto on rooted triples. The tropical vertices of this set can be computed in strongly polynomial time.

→ Every point in $\mathrm{FW}(v_1, \ldots, v_m, \mathbf{w})$ is Pareto and co-Pareto on rooted triples of $v_1, \ldots, v_m$.



median tree





Every sample tree contains the rooted triple $i \mid jk$.

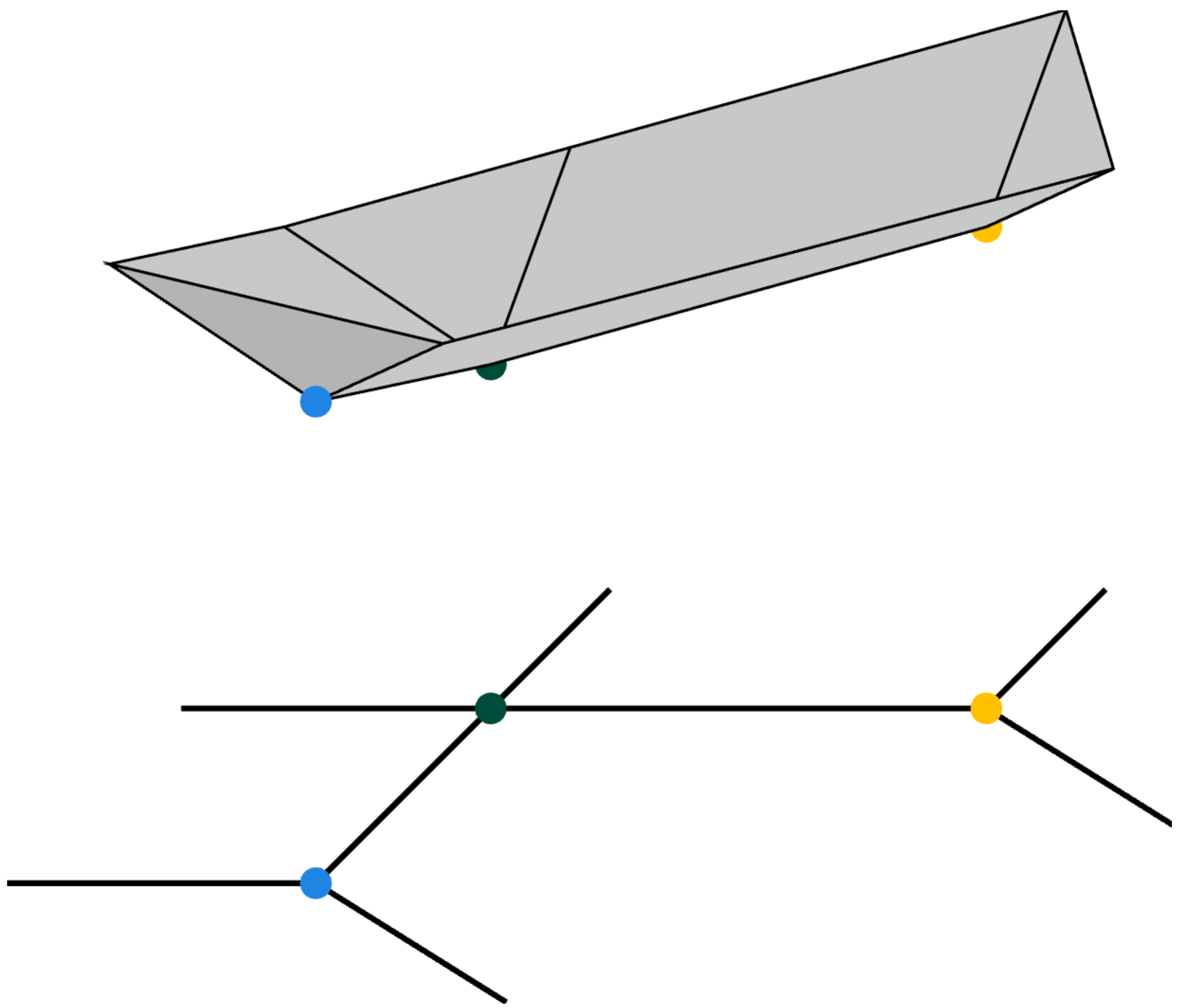
Pareto →

The median tree contains the rooted triple $i \mid jk$.

co-Pareto →

At least one sample tree contains the rooted triple $i \mid jk$.

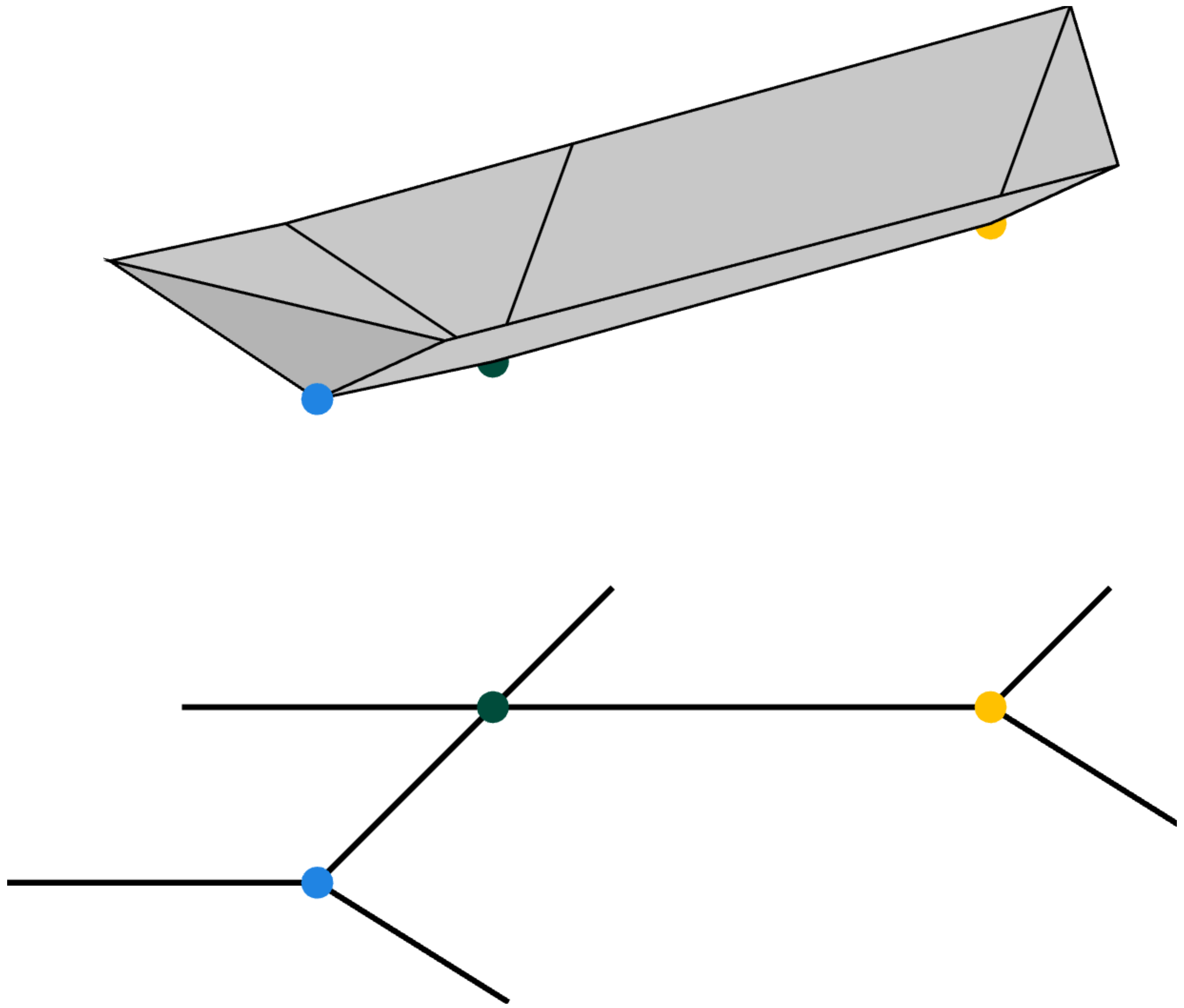
Tropical Fermat-Weber Points



The graph of

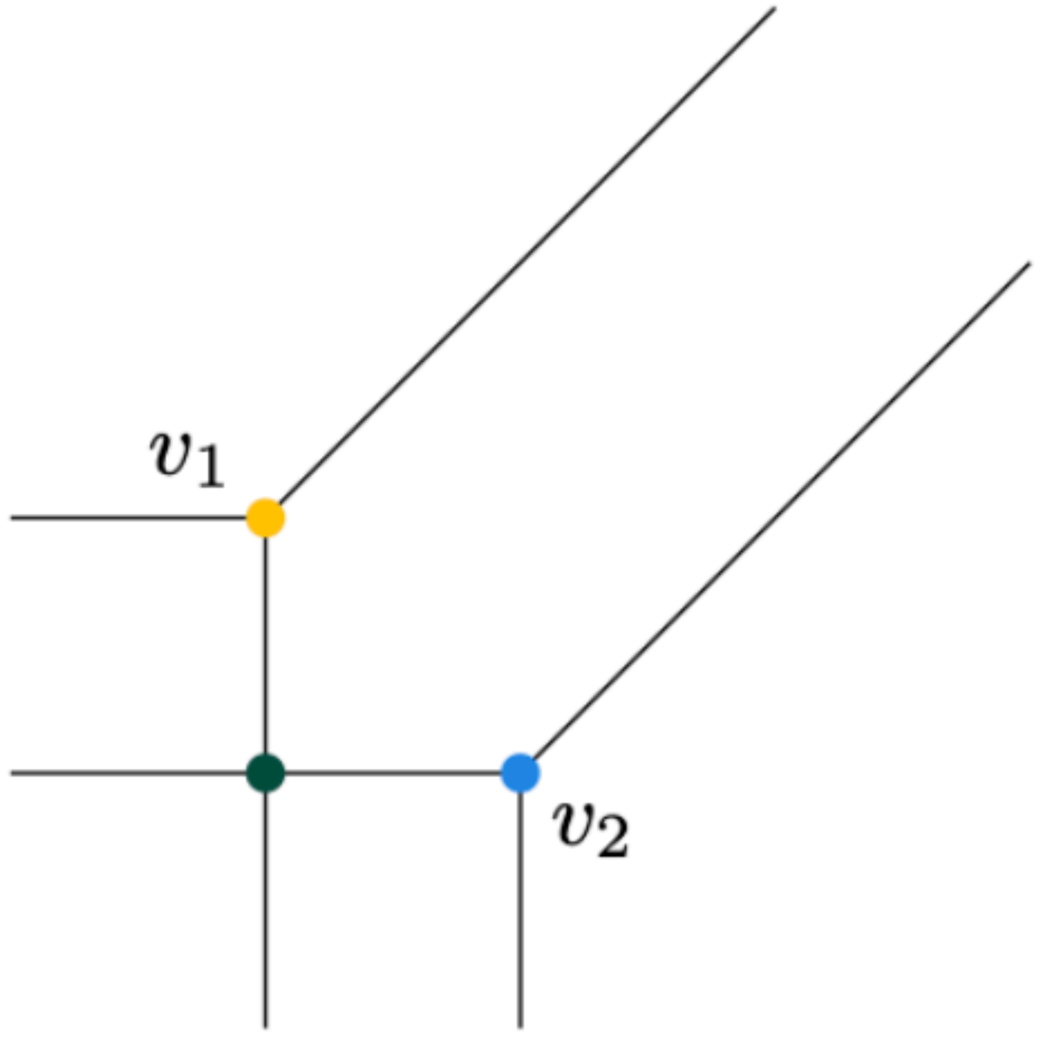
$$d_{\Delta}(x, v_1) + d_{\Delta}(x, v_2).$$

Tropical Fermat-Weber Points



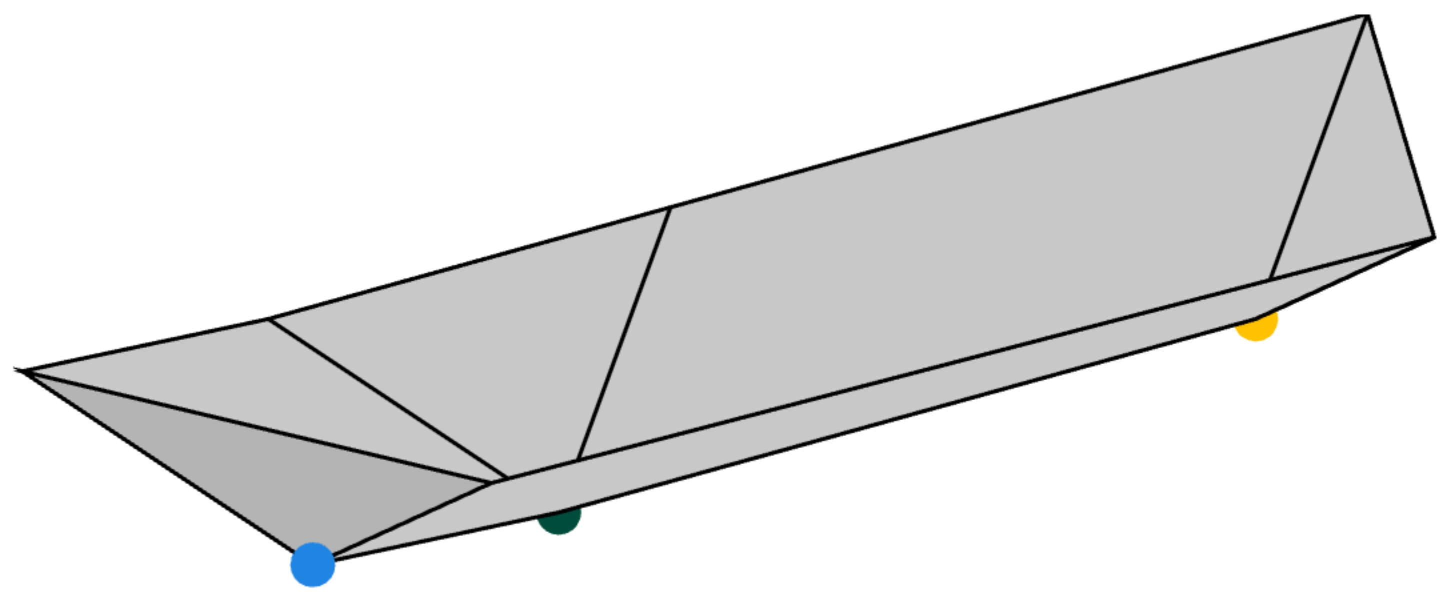
The graph of

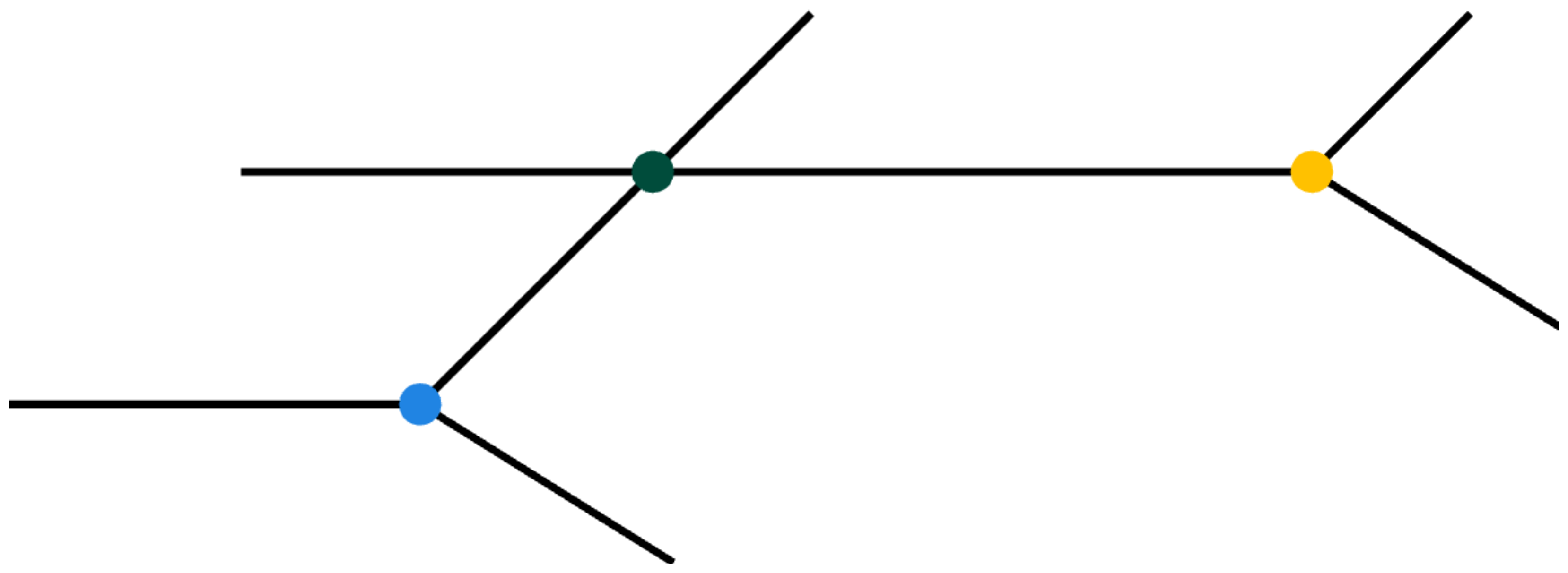
$$d_\Delta(x, v_1) + d_\Delta(x, v_2).$$



The tropical hyperplane arrangement.

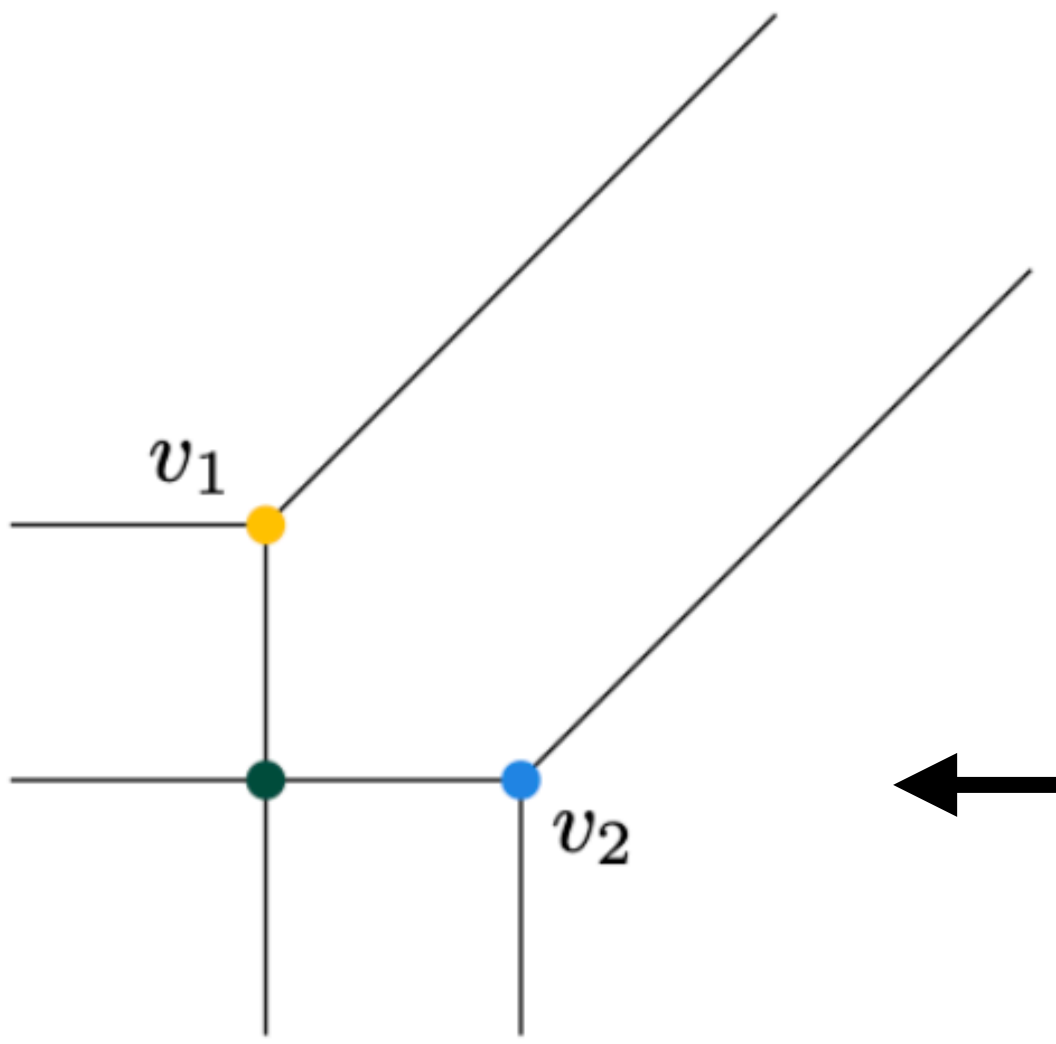
Tropical Fermat-Weber Points



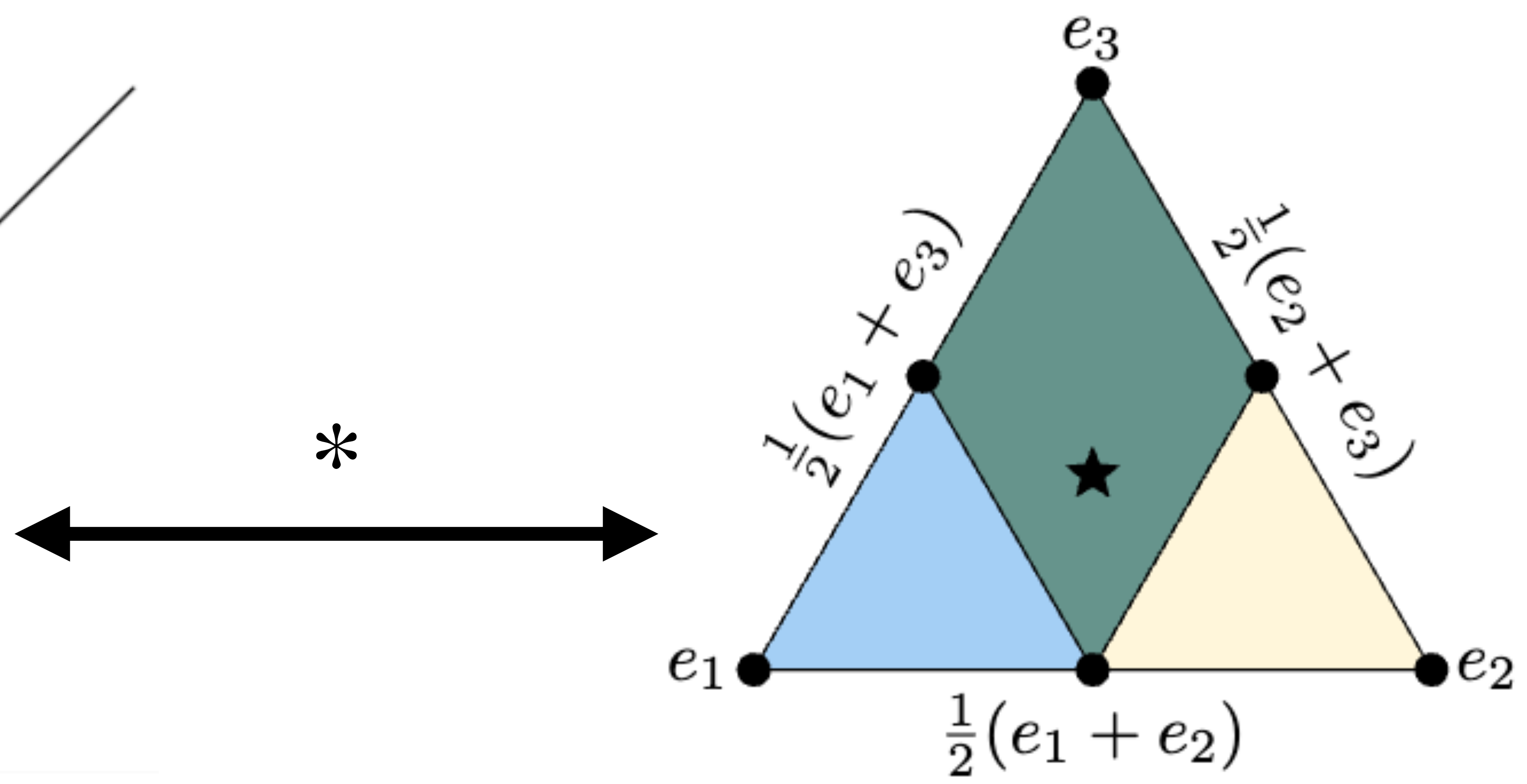


The graph of

$d_\Delta(x, v_1) + d_\Delta(x, v_2)$.

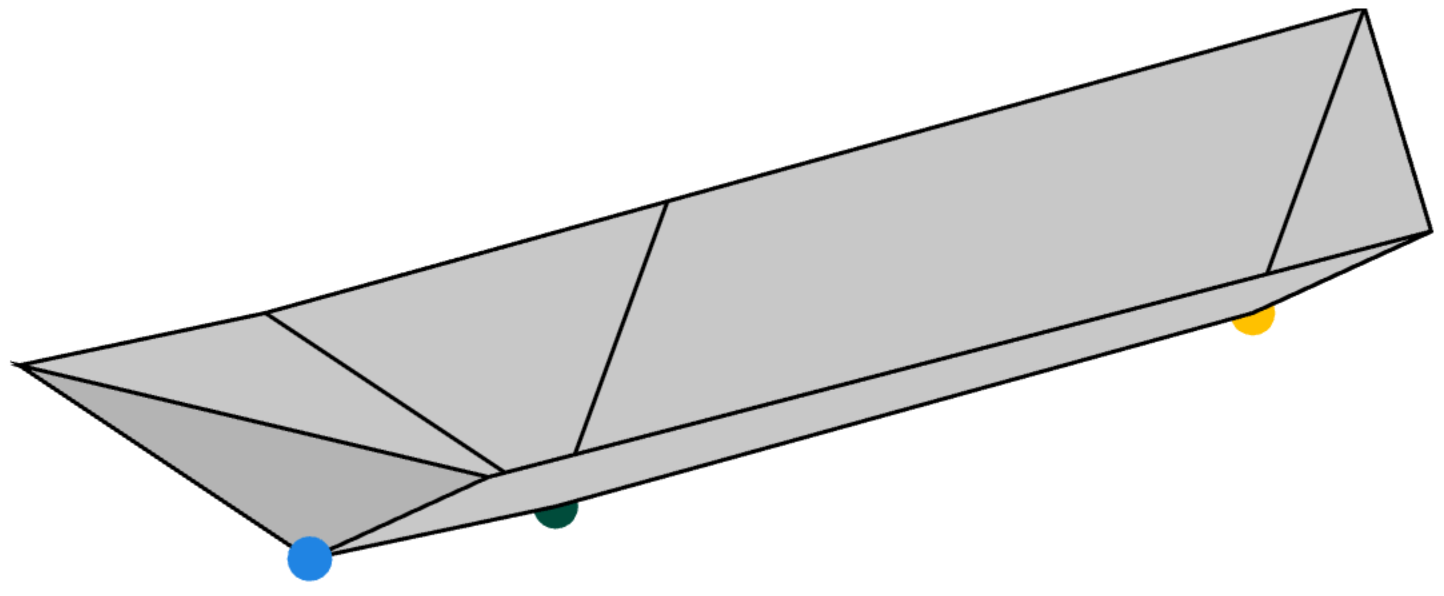


The tropical hyperplane arrangement.



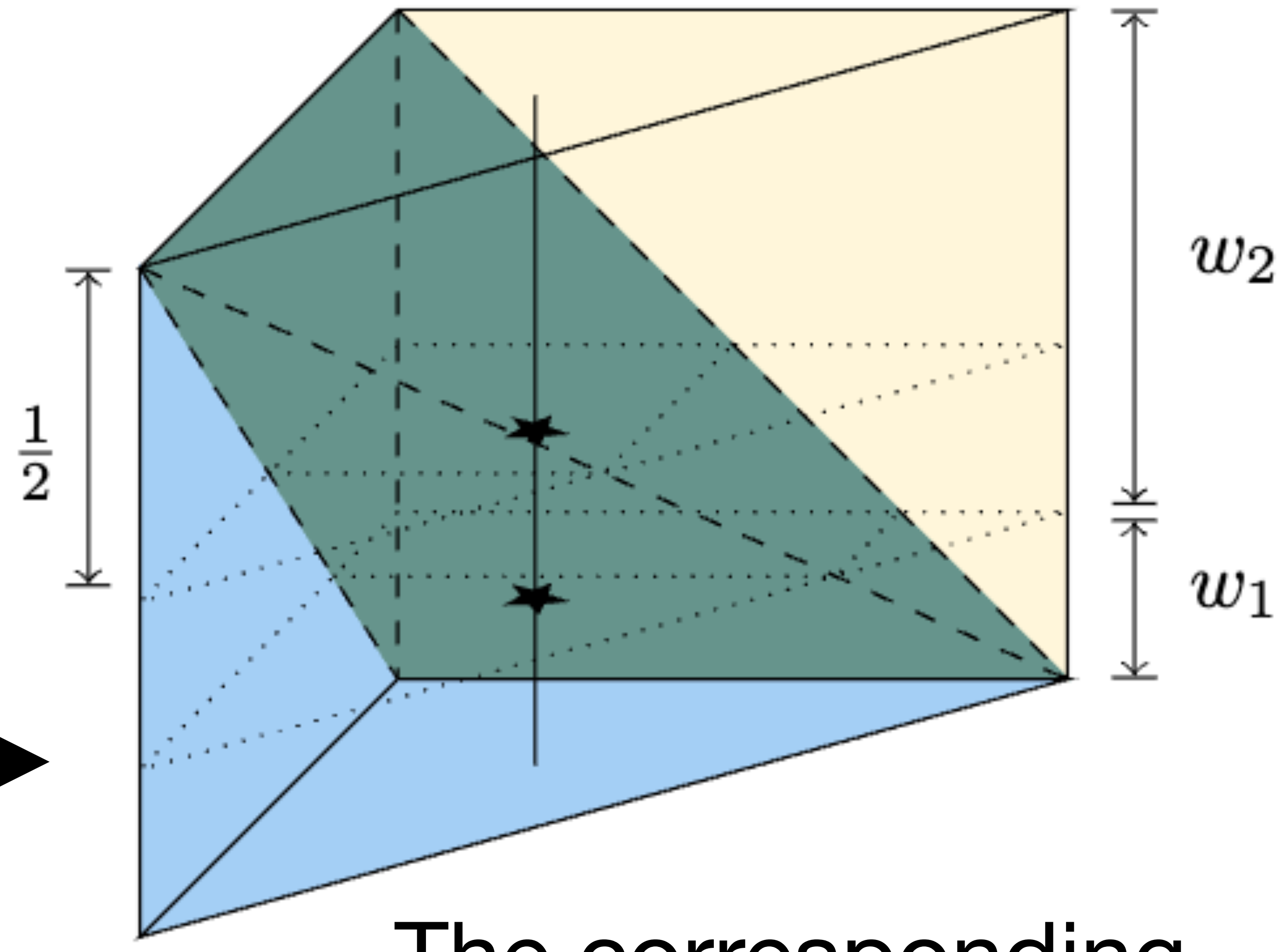
The Newton polytope of $d_\Delta(x, v_1) + d_\Delta(x, v_2)$.

Tropical Fermat-Weber Points



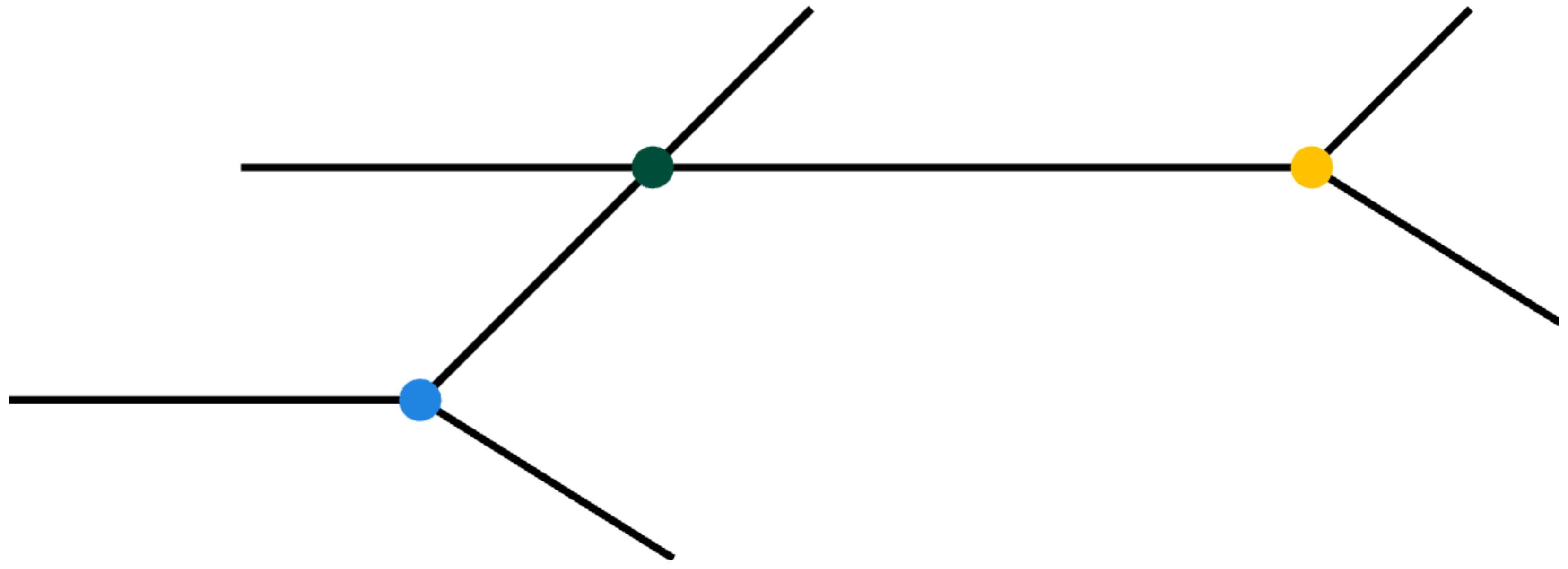
The envelope

$\mathscr{E}(v_1, \ldots, v_m) \subset \mathbb{R}^m \times \mathbb{R}^n$ $\overset{*}{\longleftrightarrow}$



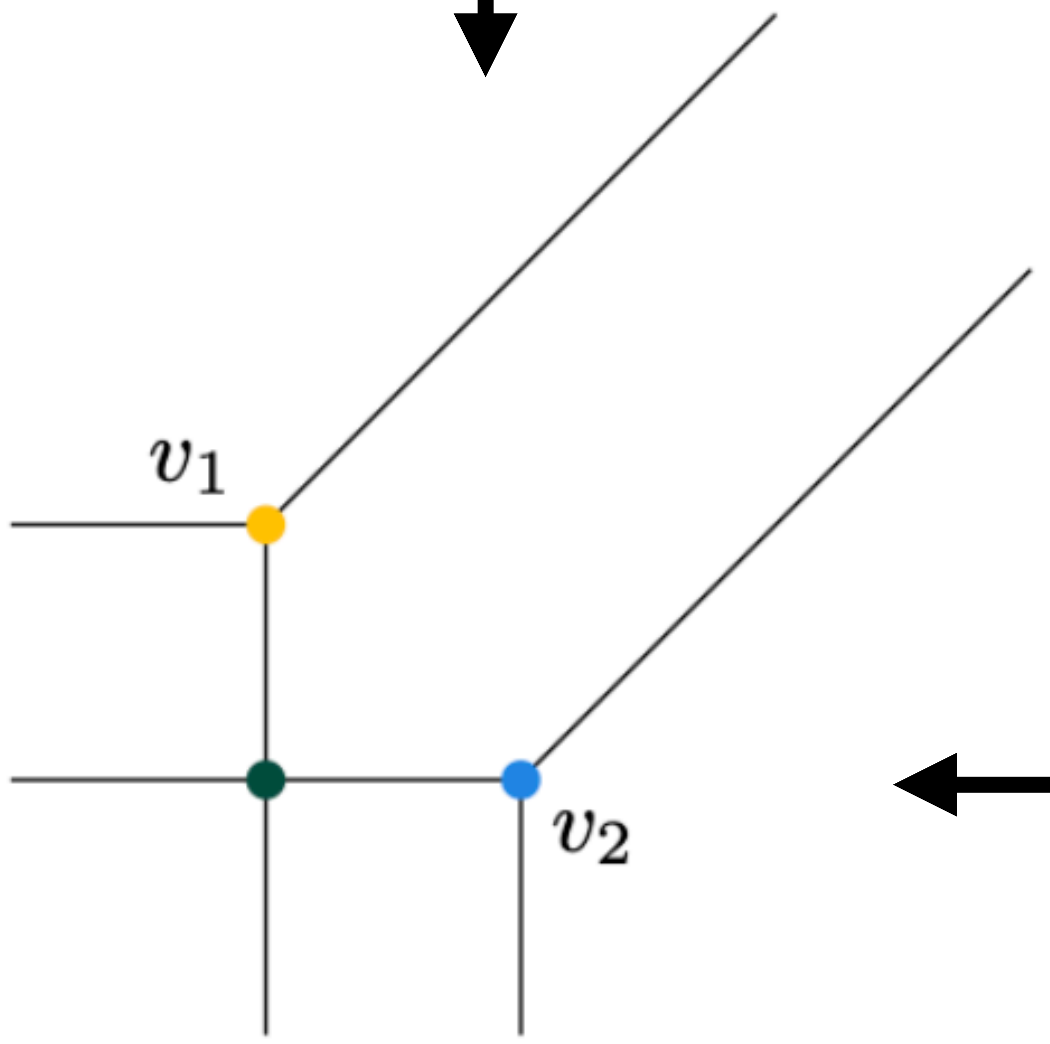
The corresponding Cayley polytope subdivision

$\downarrow (t, x) \mapsto x$



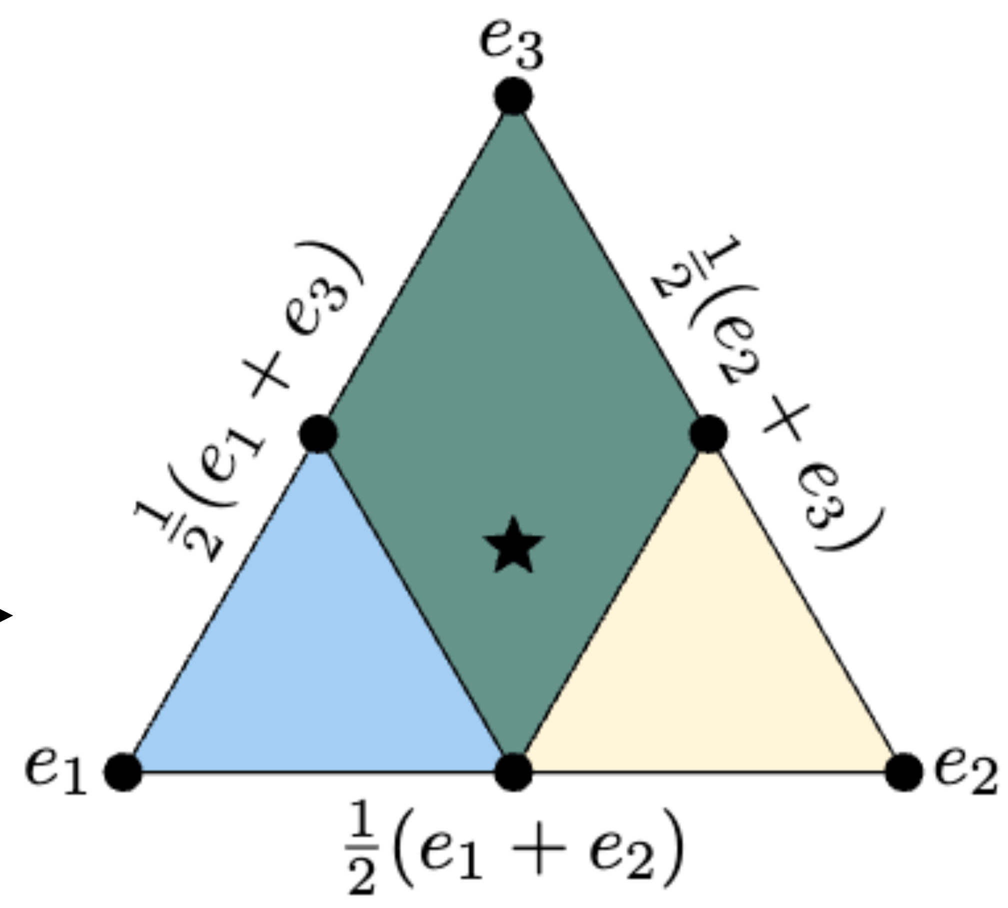
The graph of

$d_\Delta(x, v_1) + d_\Delta(x, v_2)$.



The tropical hyperplane arrangement.

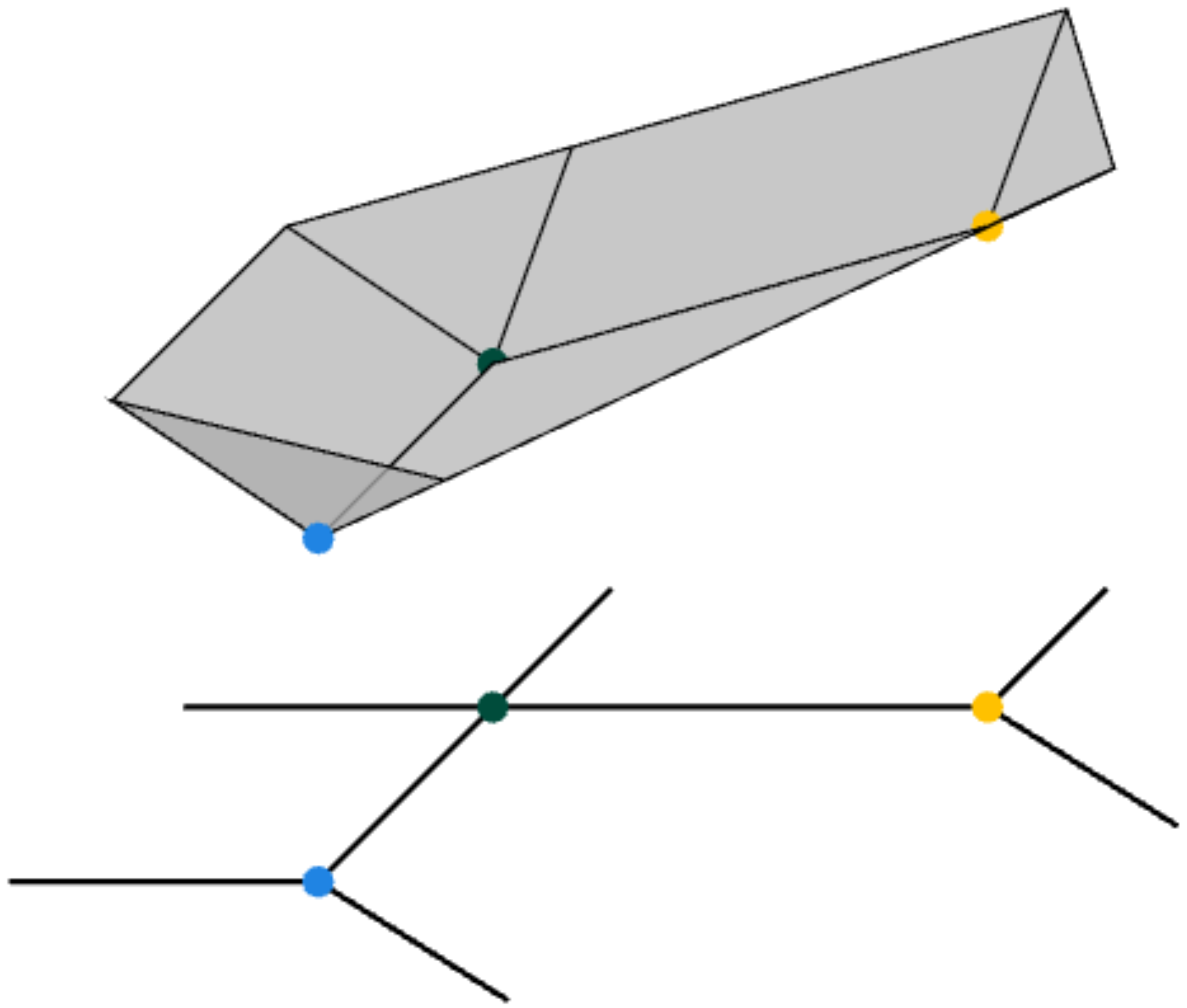
$\overset{*}{\longleftrightarrow}$



The Newton polytope of $d_\Delta(x, v_1) + d_\Delta(x, v_2)$.

(weighted) $w_1 = 1/3,\ w_2 = 2/3$

Tropical Fermat-Weber Points



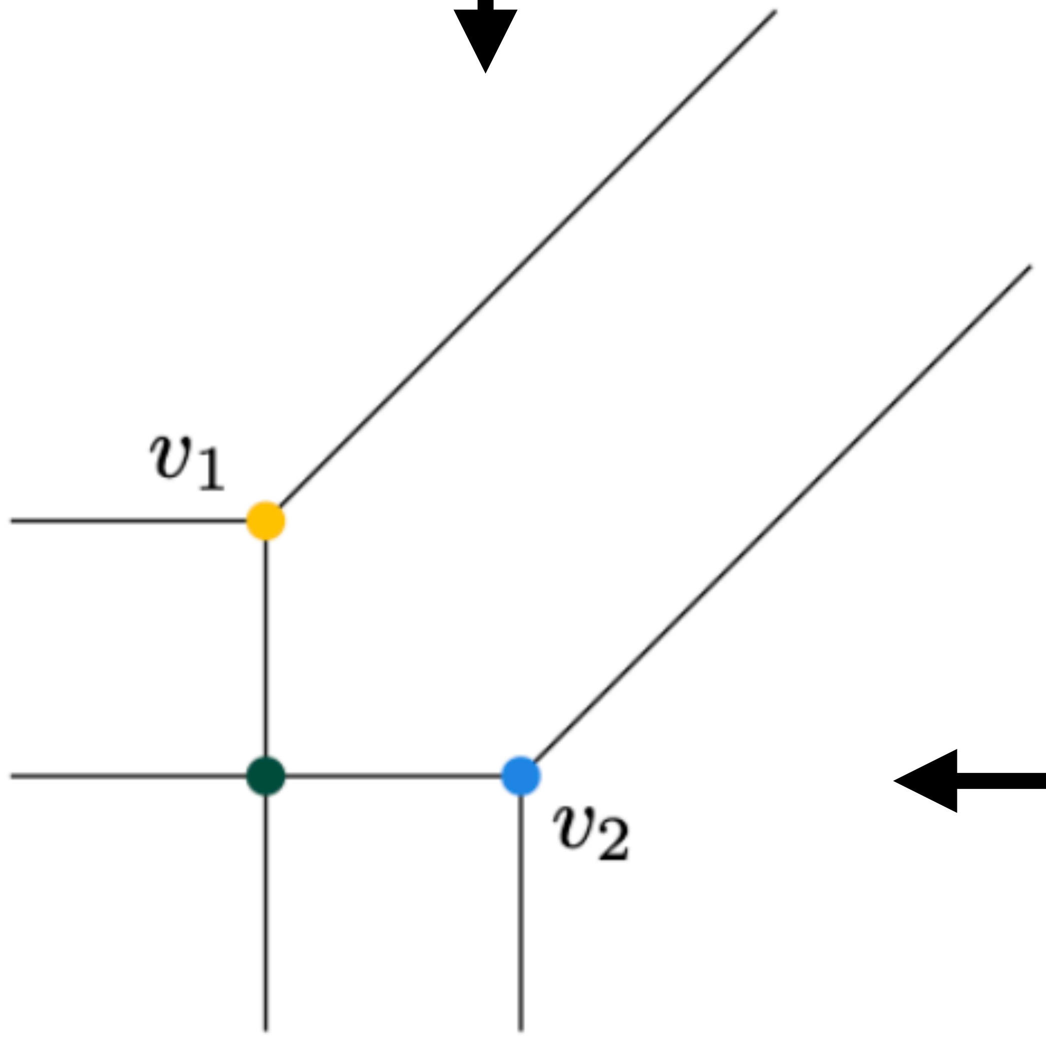
The graph of

$$\frac{1}{3} d_\Delta(x, v_1) + \frac{2}{3} d_\Delta(x, v_2).$$

The envelope

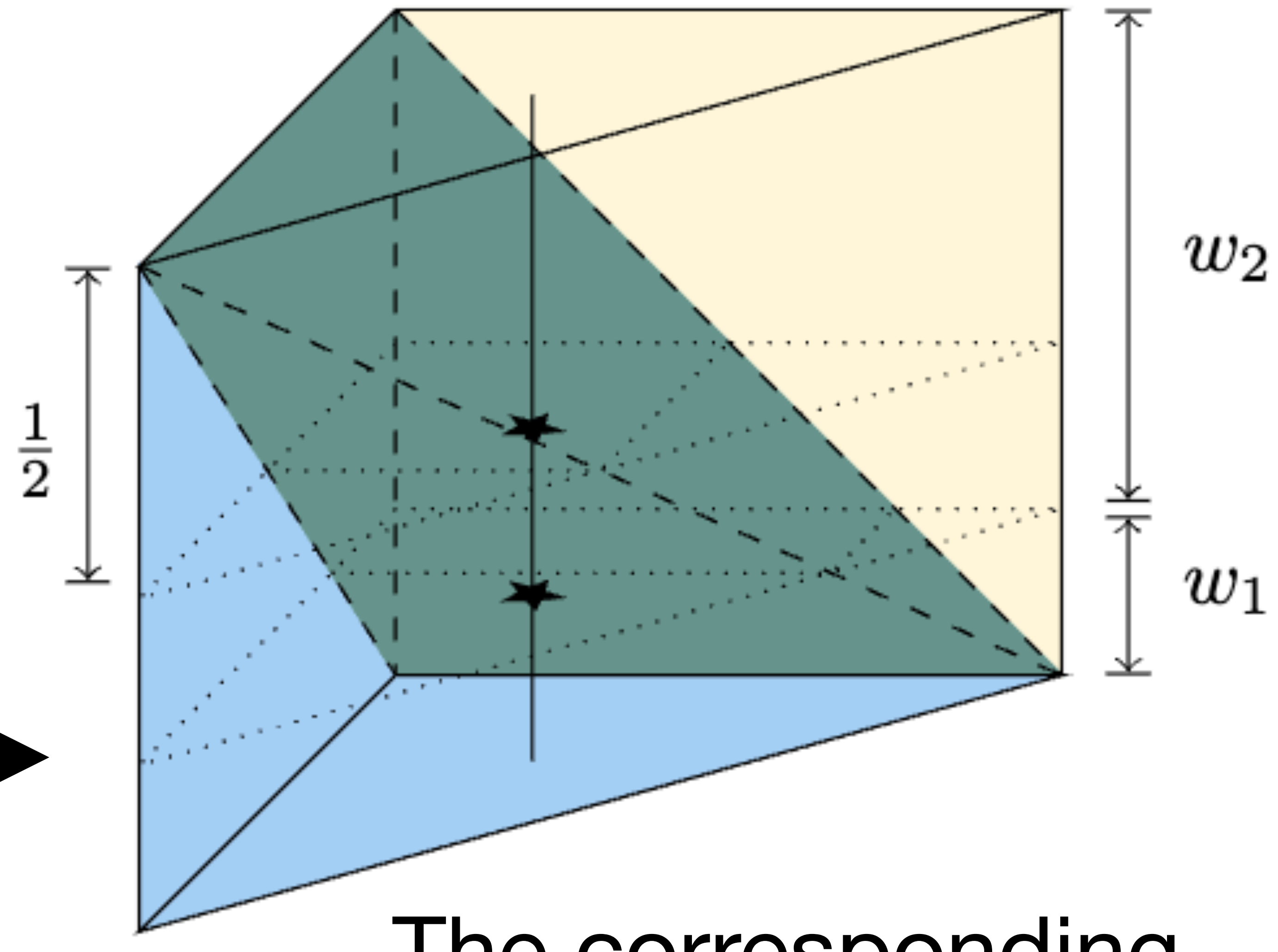
$$\mathscr{E}(v_1, \ldots, v_m, \mathbf{w}) \subset \mathbb{R}^m \times \mathbb{R}^n \quad \overset{*}{\longleftrightarrow}$$

$$\downarrow (t, x) \mapsto x$$

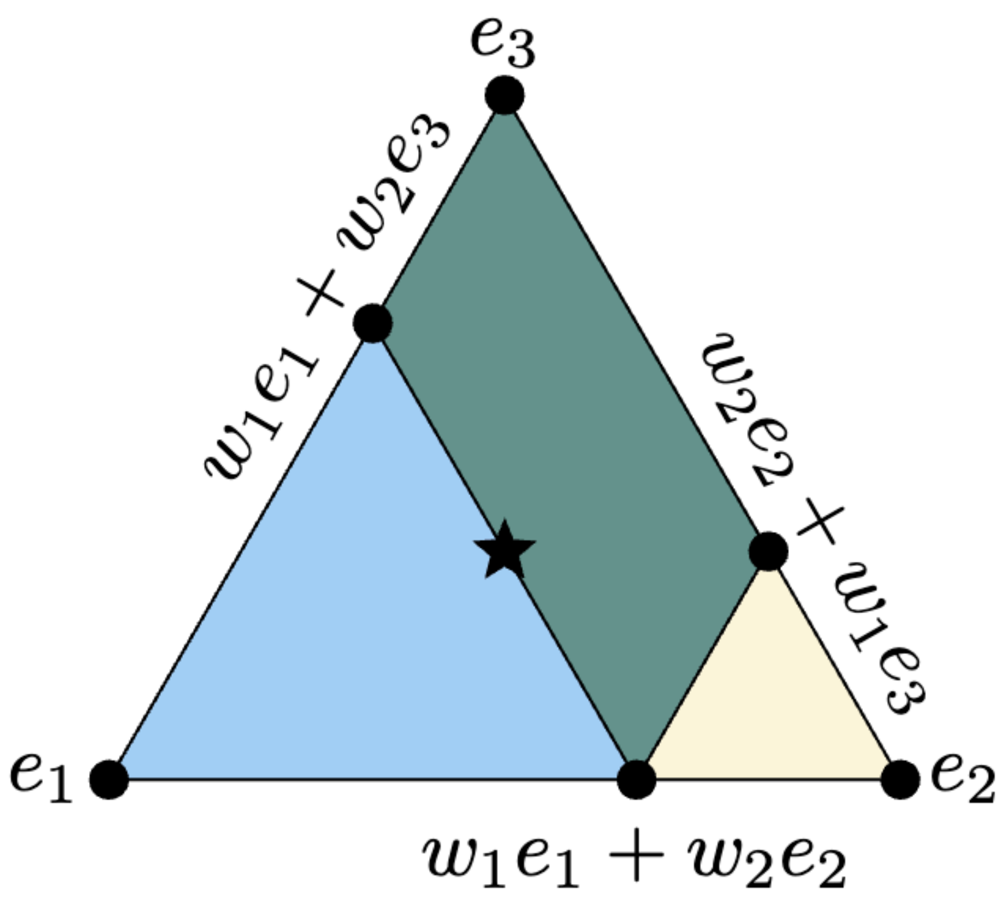


The tropical hyperplane arrangement.

$\overset{*}{\longleftrightarrow}$



The corresponding Cayley polytope subdivision



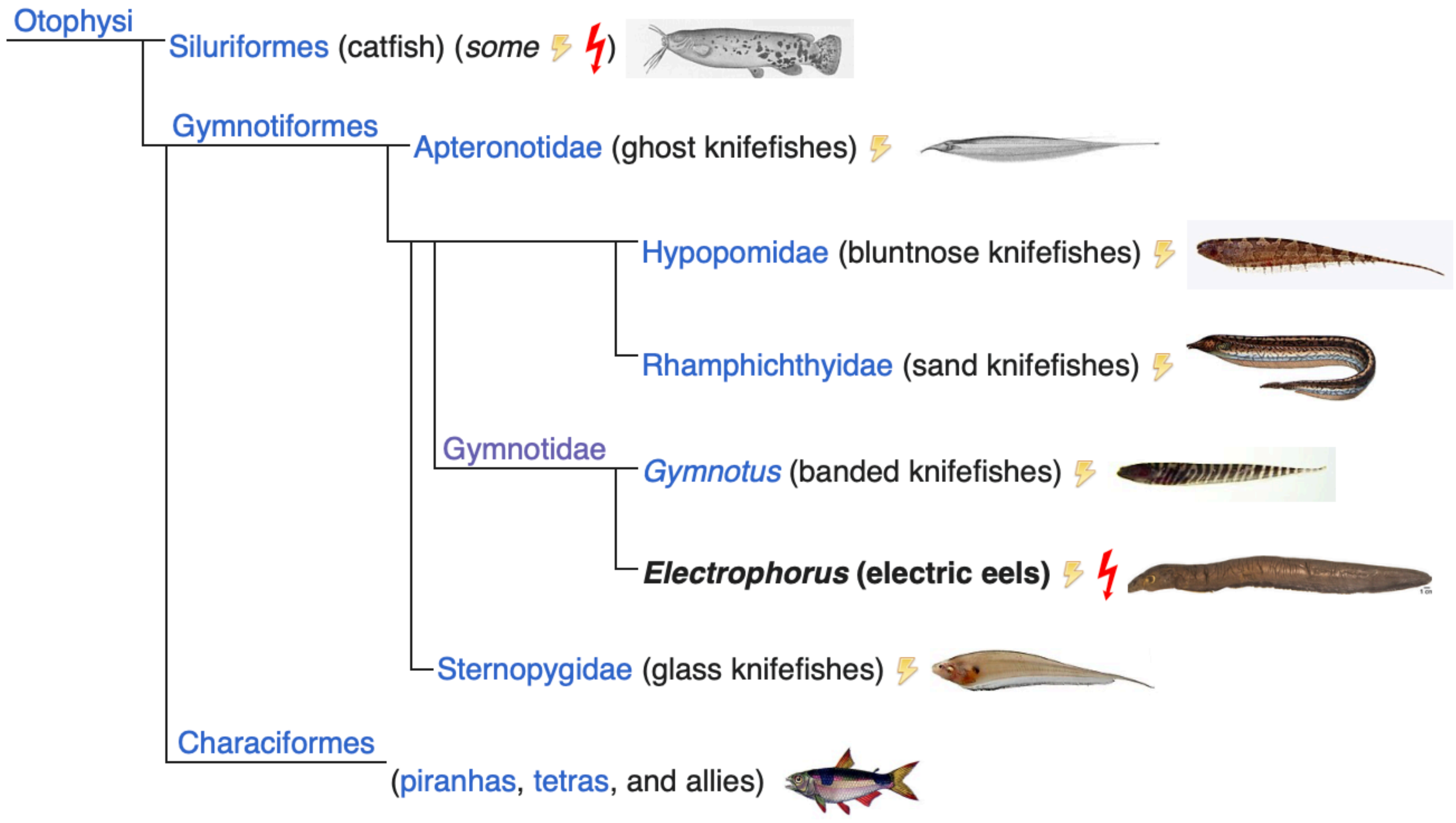
The Newton polytope of

$$\frac{1}{3} d_\Delta(x, v_1) + \frac{2}{3} d_\Delta(x, v_2).$$

Back to the Data

Weighted data

- Data collected may not align with actual population proportions.
- Some data is higher quality than other data.
    - In phylogenetics, reconstructed trees can be weighted by a function of bootstrap scores on edges [MBX].



From the Wikipedia page on electric eels.

What's Next?

Normal distributions on tropical tree space

Problems:

- If d_T is a tree metric, then $-d_T$ is usually **not** a tree metric.
- If we write down the usual normal distribution formula on a tree space, the normalizing constant depends on the center of the distribution.

What other statistics problems should be explored in tree space?

Thank you!

References

- [AK] Ardila, F., & Klivans, C. J. (2006). The Bergman complex of a matroid and phylogenetic trees. *Journal of Combinatorial Theory, Series B*, *96*(1), 38-49.
- [BHV] Billera, L. J., Holmes, S. P., & Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, *27*(4), 733-767.
- [BSYM] Barnhill, D., Sabol, J., Yoshida, R., & Miura, K. (2024). Tropical Fermat-Weber Polytopes. *arXiv preprint arXiv:2402.14287*.
- [C] **Cox, S.** (2022). Classifying Tree Topology Changes along Tropical Line Segments. *Algebraic Statistics*.
- [CJ] Comăneci, A., & Joswig, M. (2024). Tropical medians by transportation. *Mathematical Programming*, *205*(1), 813-839.
- [CC] **Cox, S.**, & Curiel, M. (2023). The tropical polytope is the set of all weighted tropical Fermat-Weber points. *arXiv preprint arXiv:2310.07732*.
- [DS] Develin, M., & Sturmfels, B. (2004). Tropical convexity. *Documenta Mathematica*, *9*, 1-27.
- [D] Diego Delso, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons.
- [L] Lemaire, Patrick (2020). Astec-Pm13 (Wild type Phallusia mammillata embryo, live SPIM imaging. figshare. Dataset.
- [MBX] Makarenkov, V., Boc, A., Xie, J. *et al.* (2010). Weighted bootstrapping: a correction method for assessing the robustness of phylogenetic trees. *BMC Evol Biol* **10**, 250.
- [PYZ] Page, R., Yoshida, R., & Zhang, L. (2020). Tropical principal component analysis on the space of phylogenetic trees. *Bioinformatics*, *36*(17), 4590-4598.