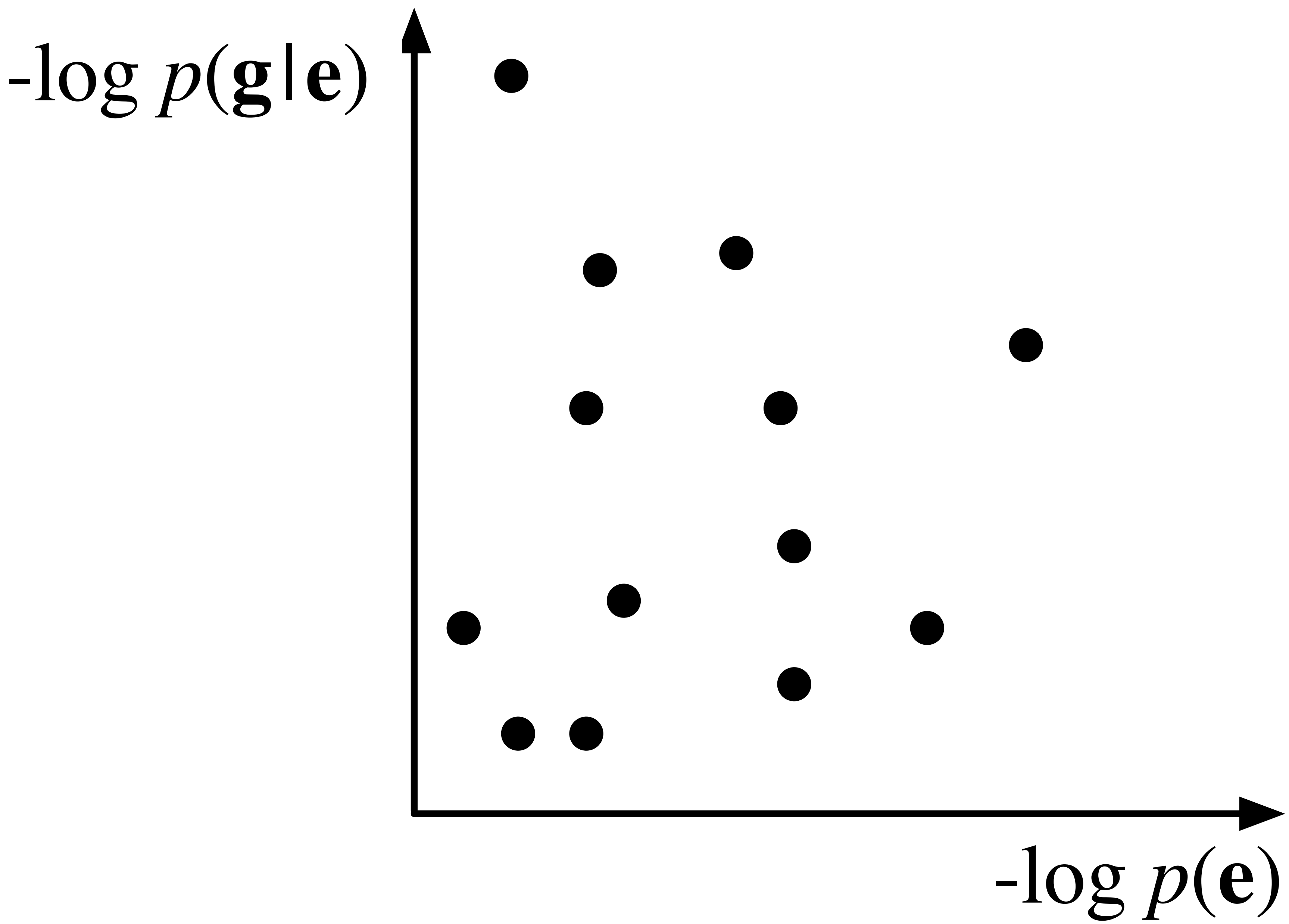
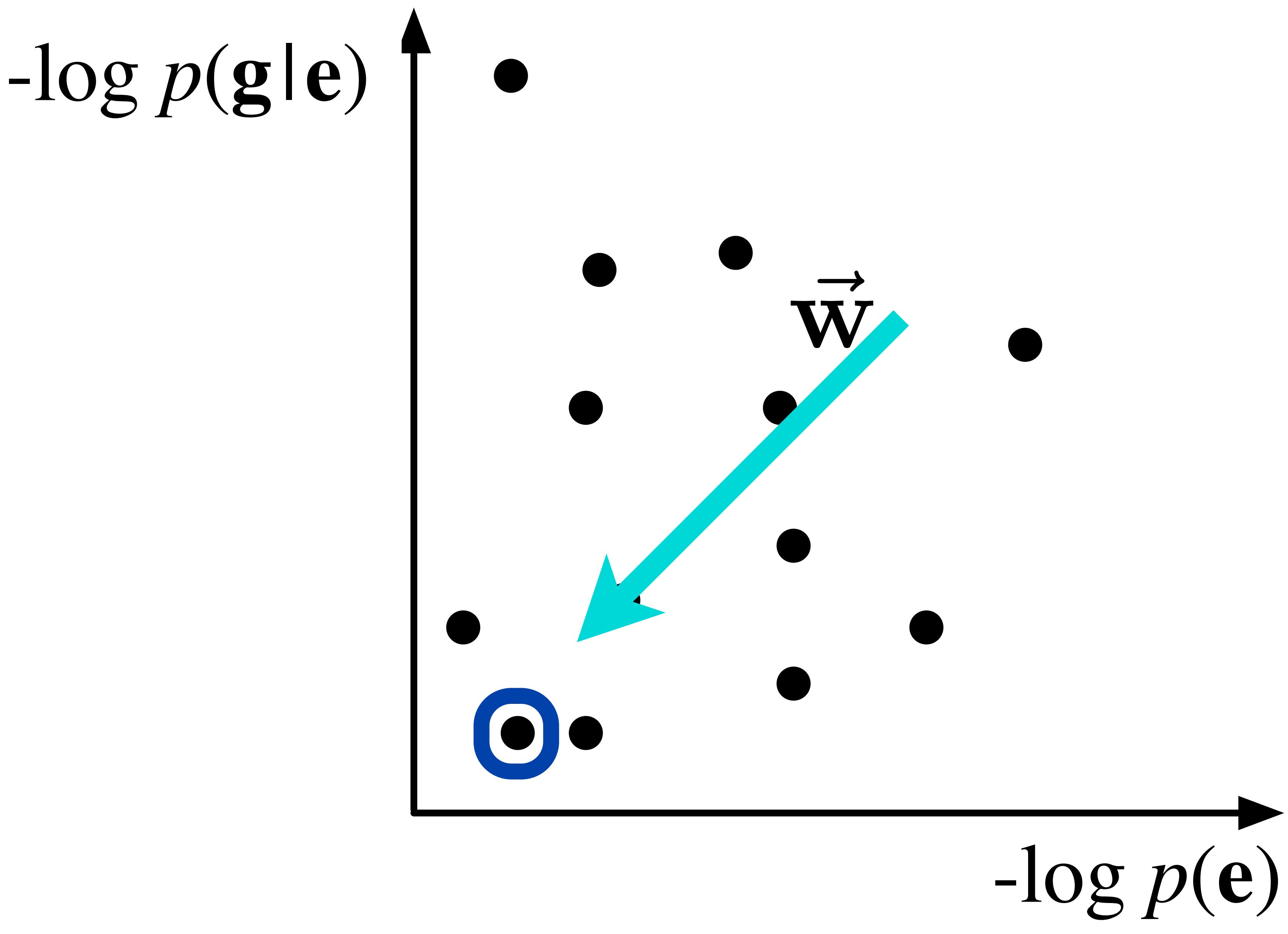
Discriminative Training II: MERT

April 3, 2014

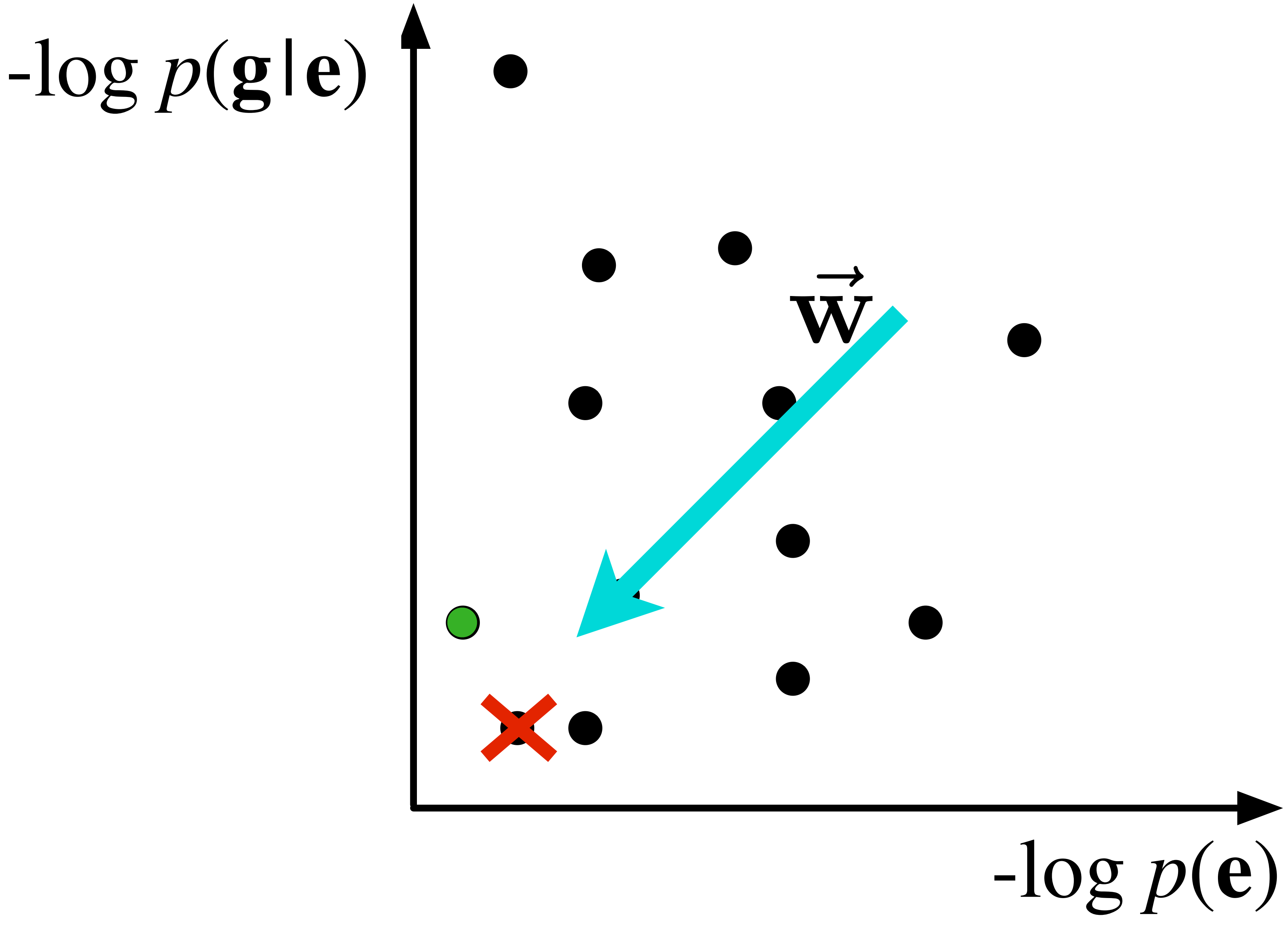
The Noisy Channel



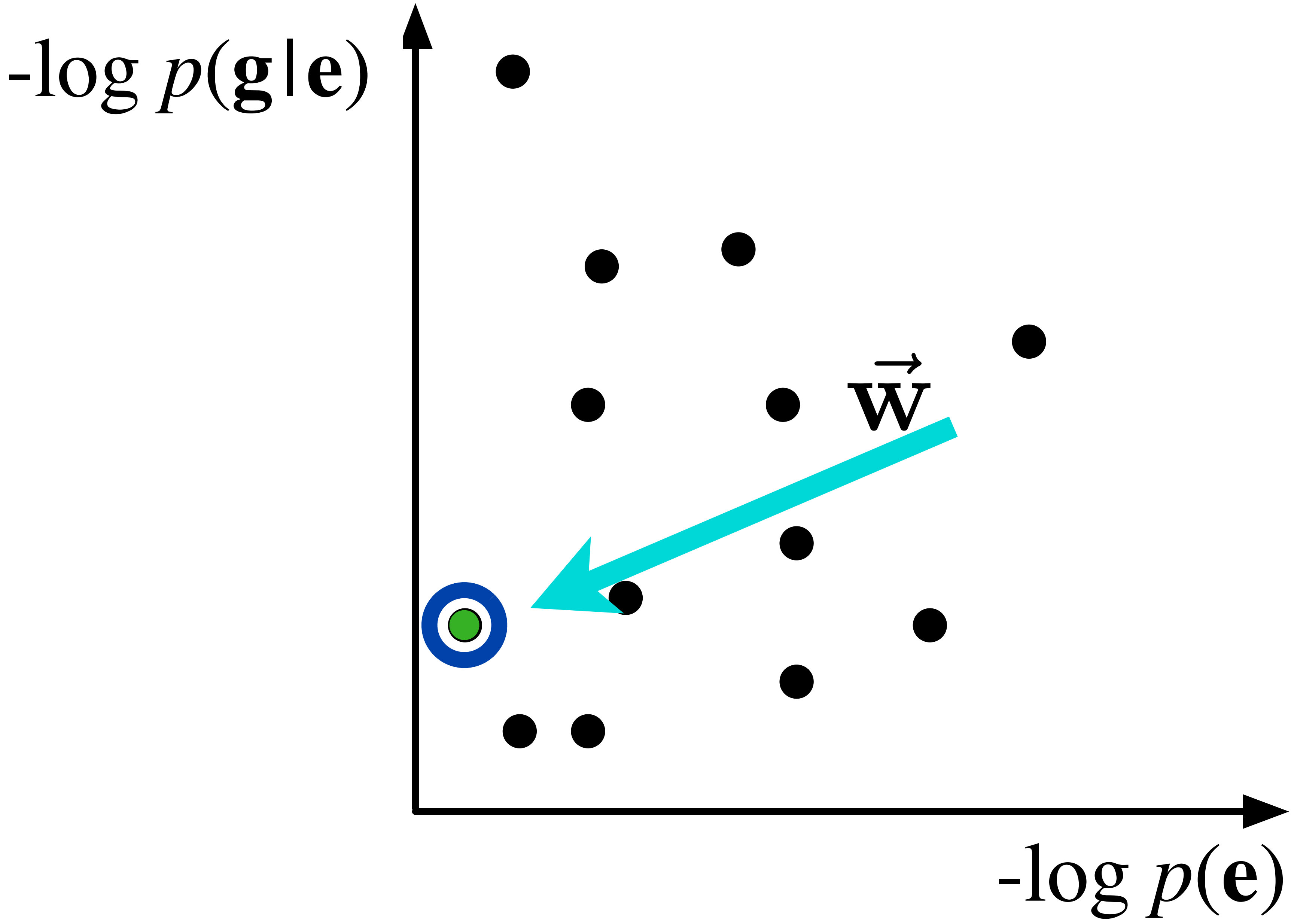
As a Linear Model



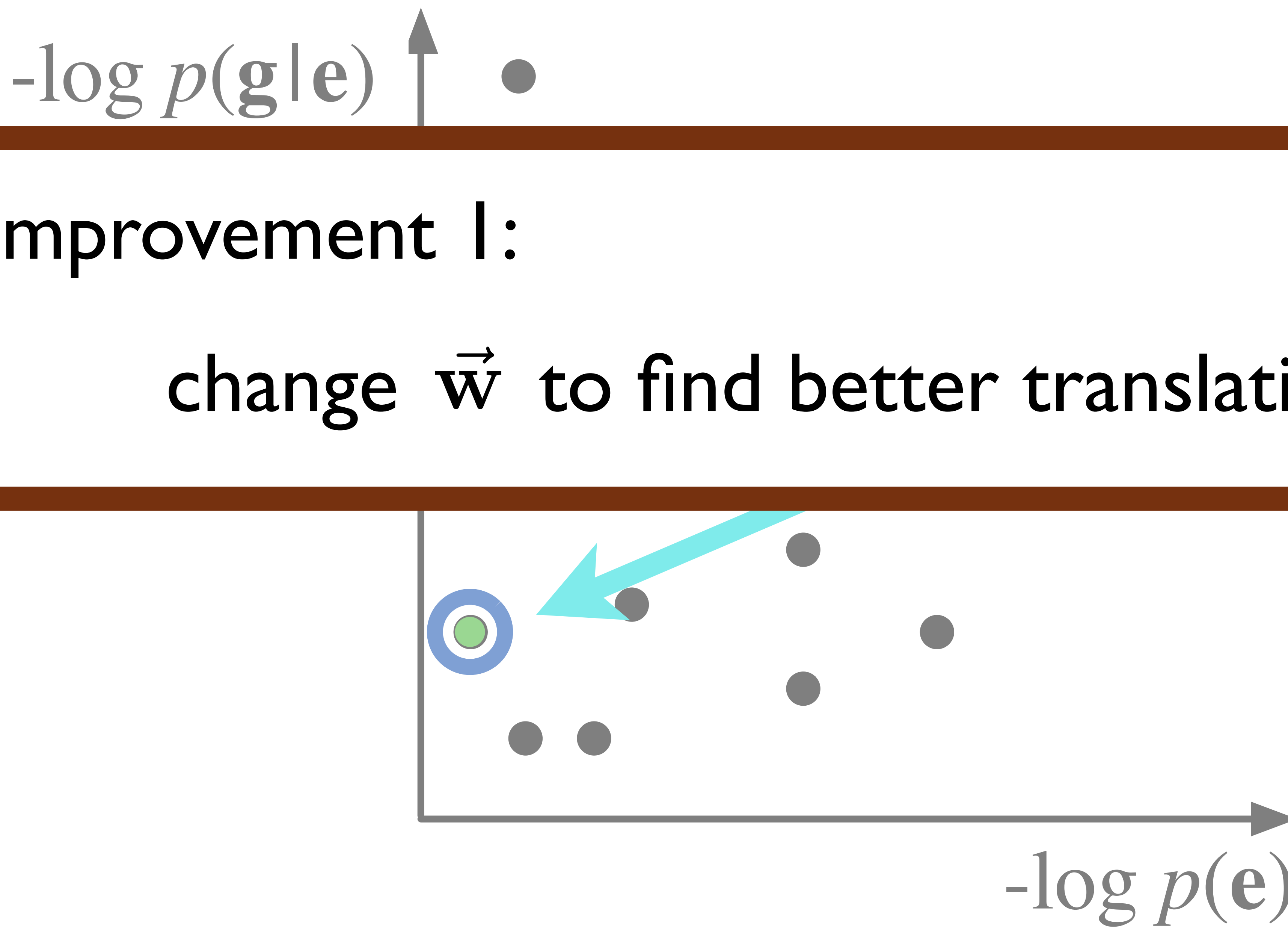
As a Linear Model



As a Linear Model



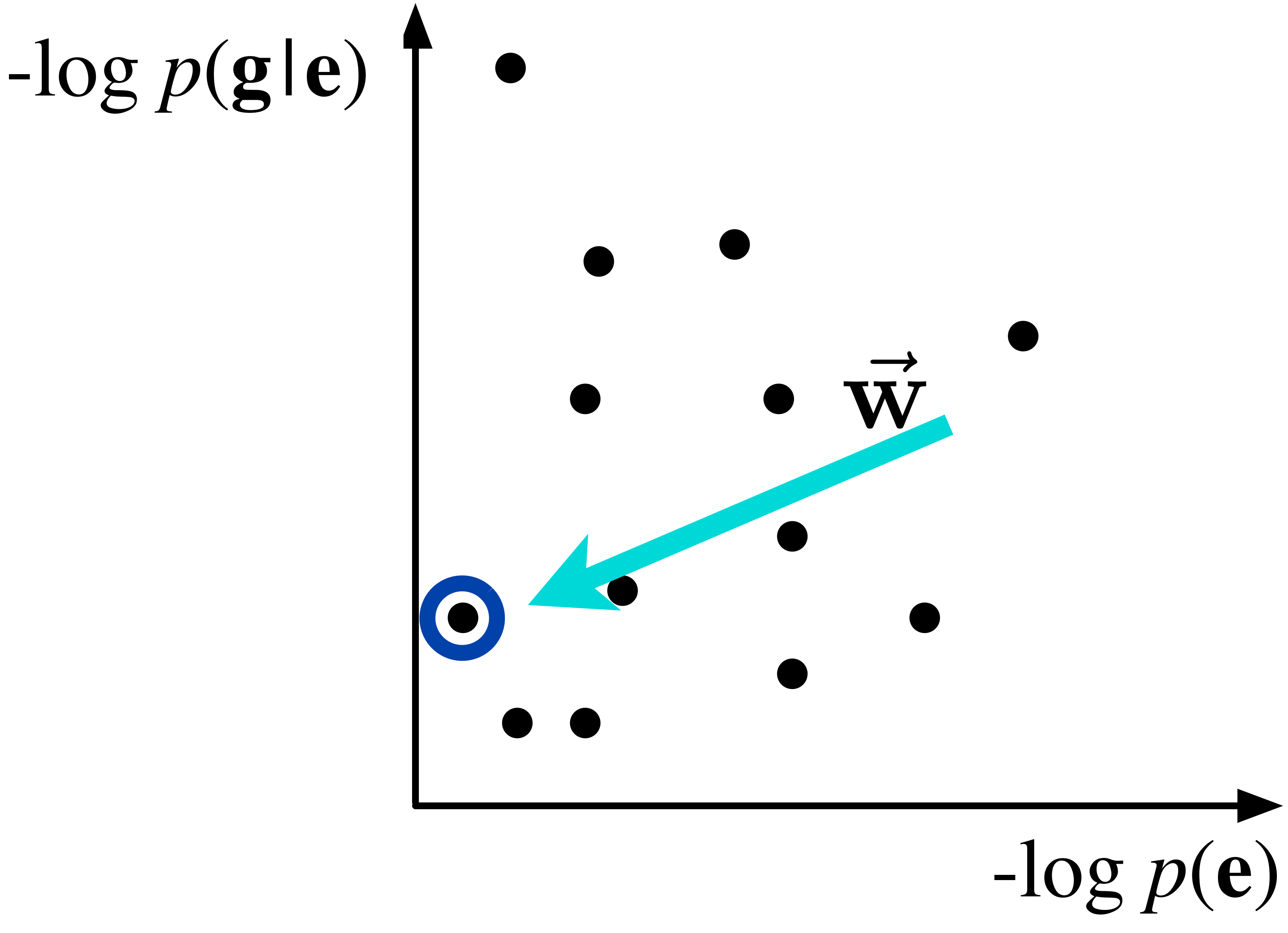
As a Linear Model



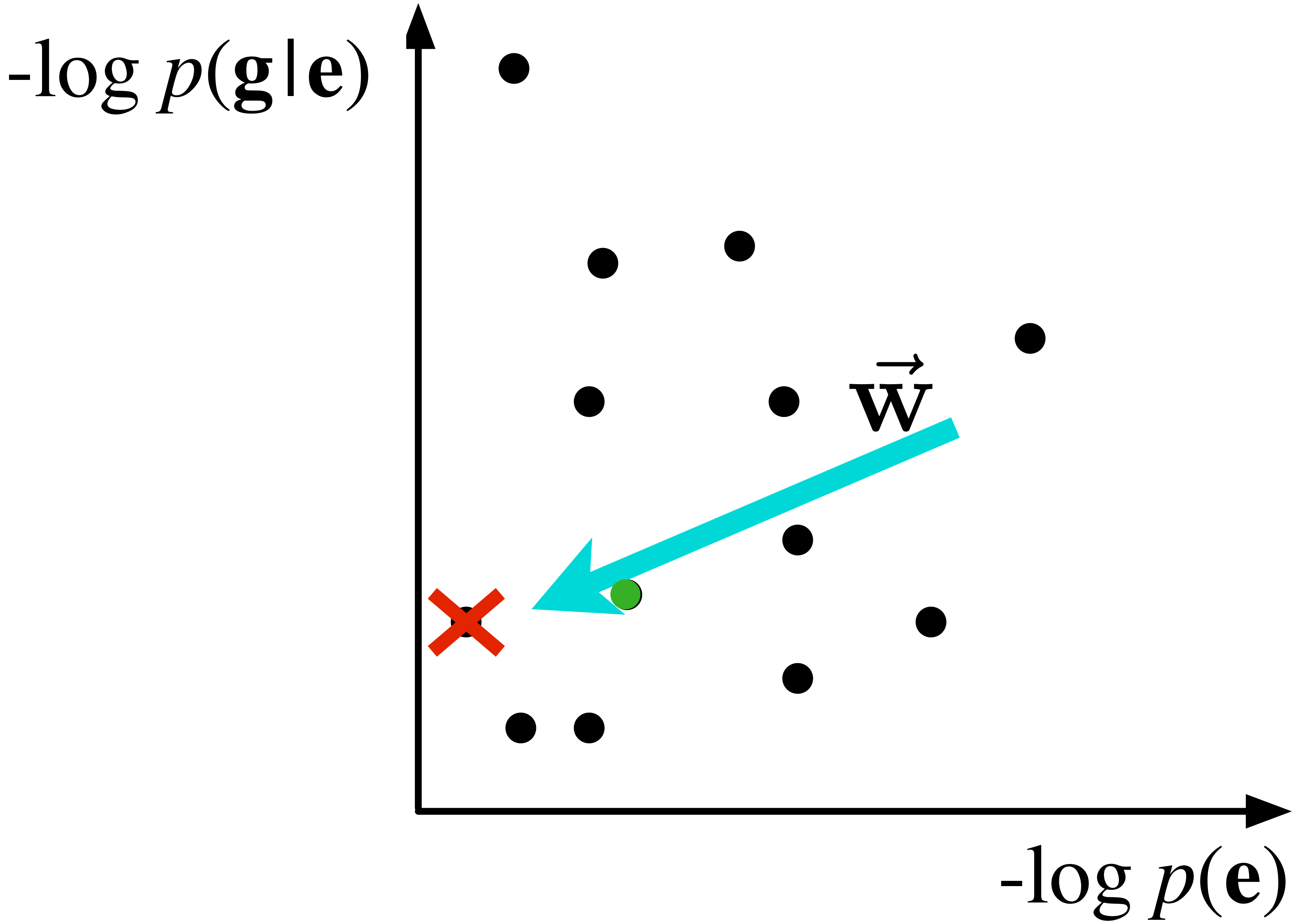
Improvement 1:

change $\vec{w}$ to find better translations

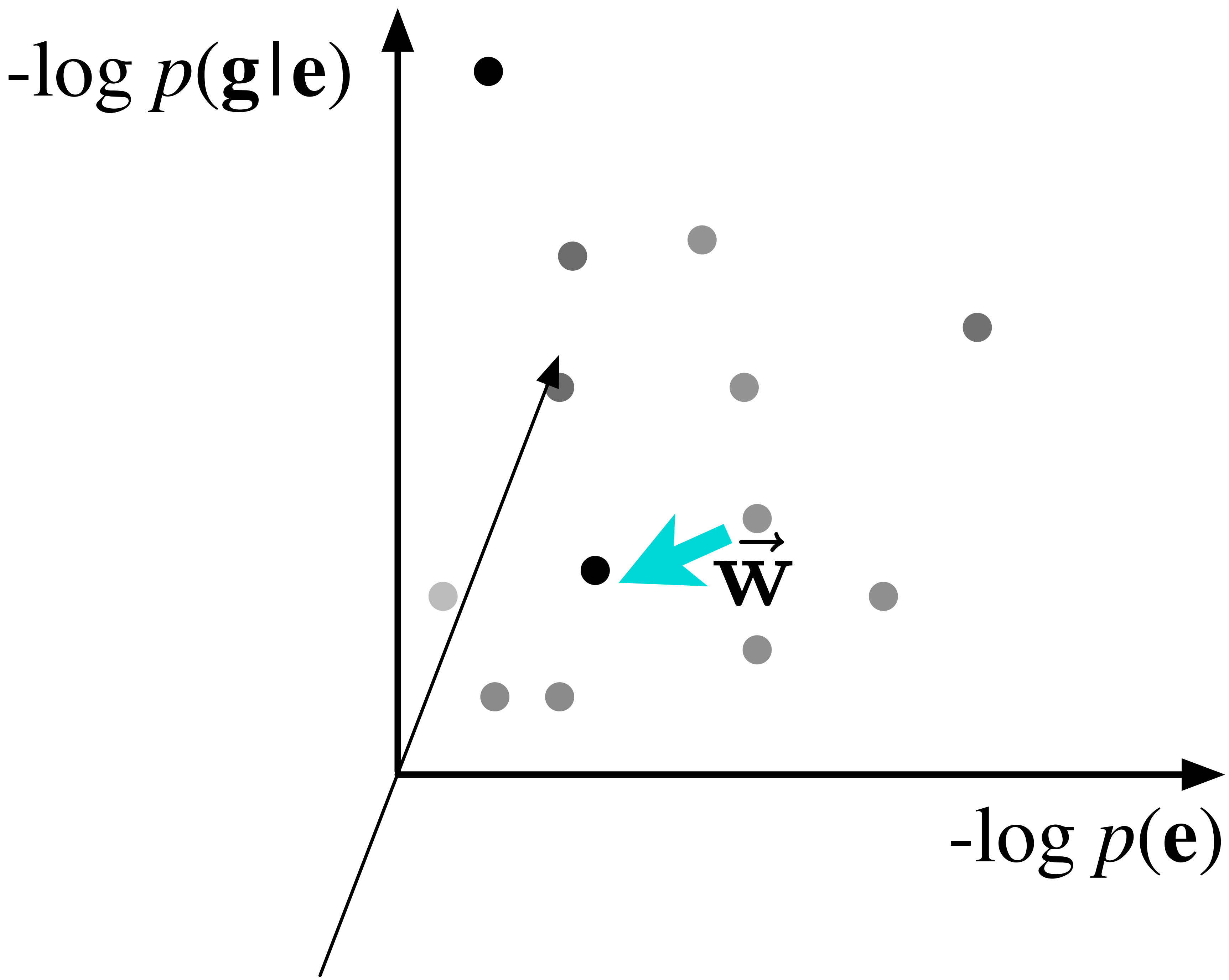
As a Linear Model



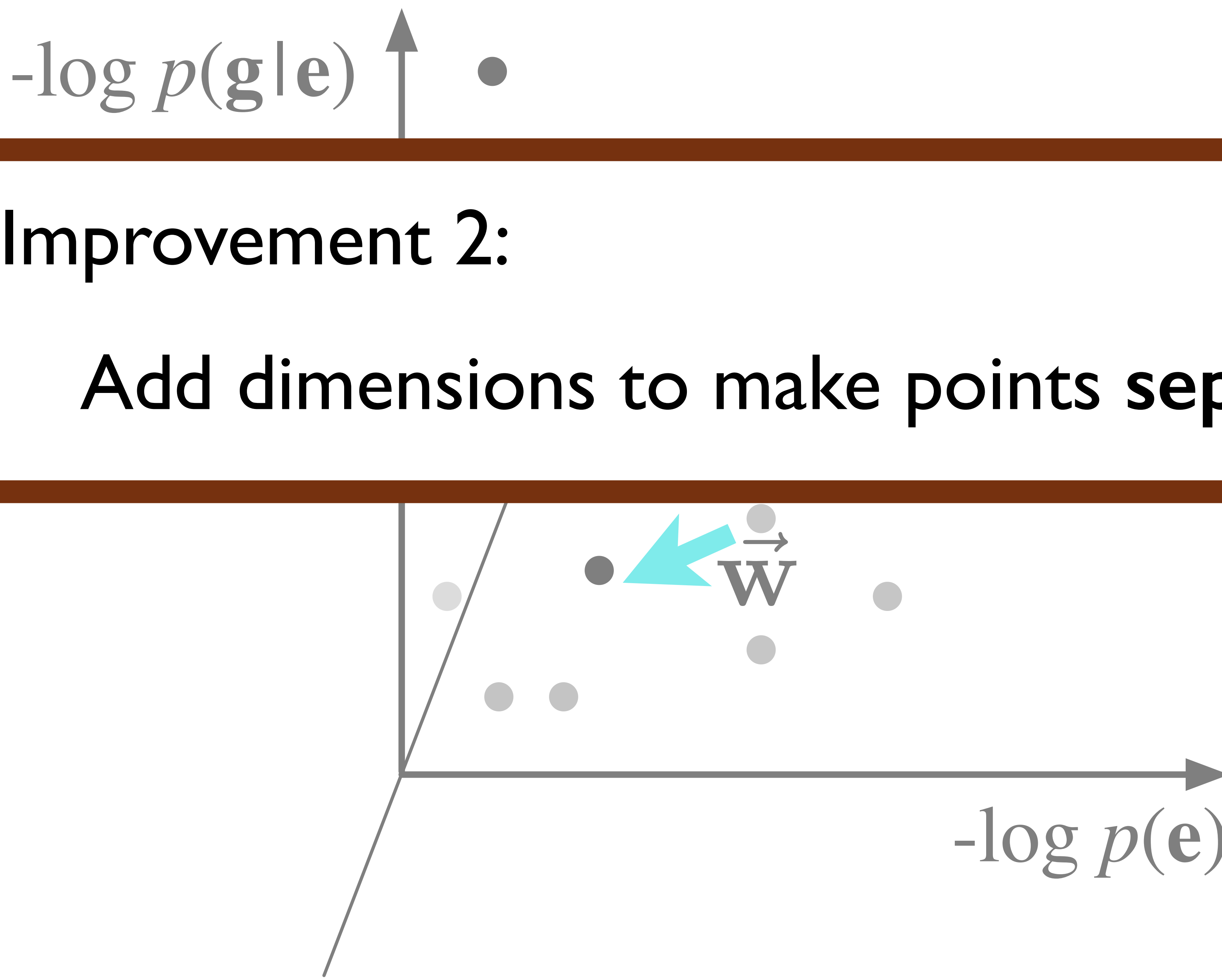
As a Linear Model



As a Linear Model



As a Linear Model

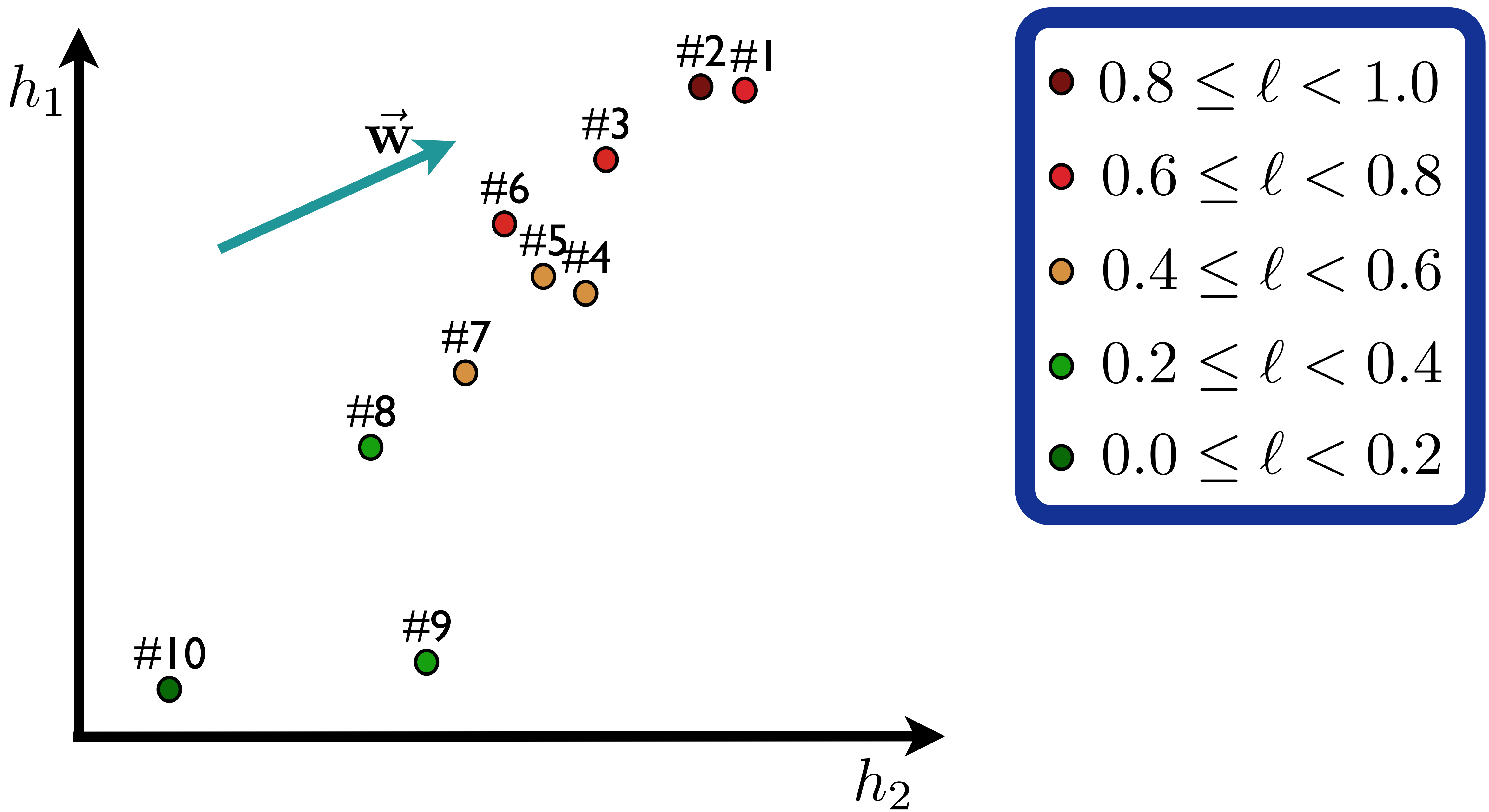


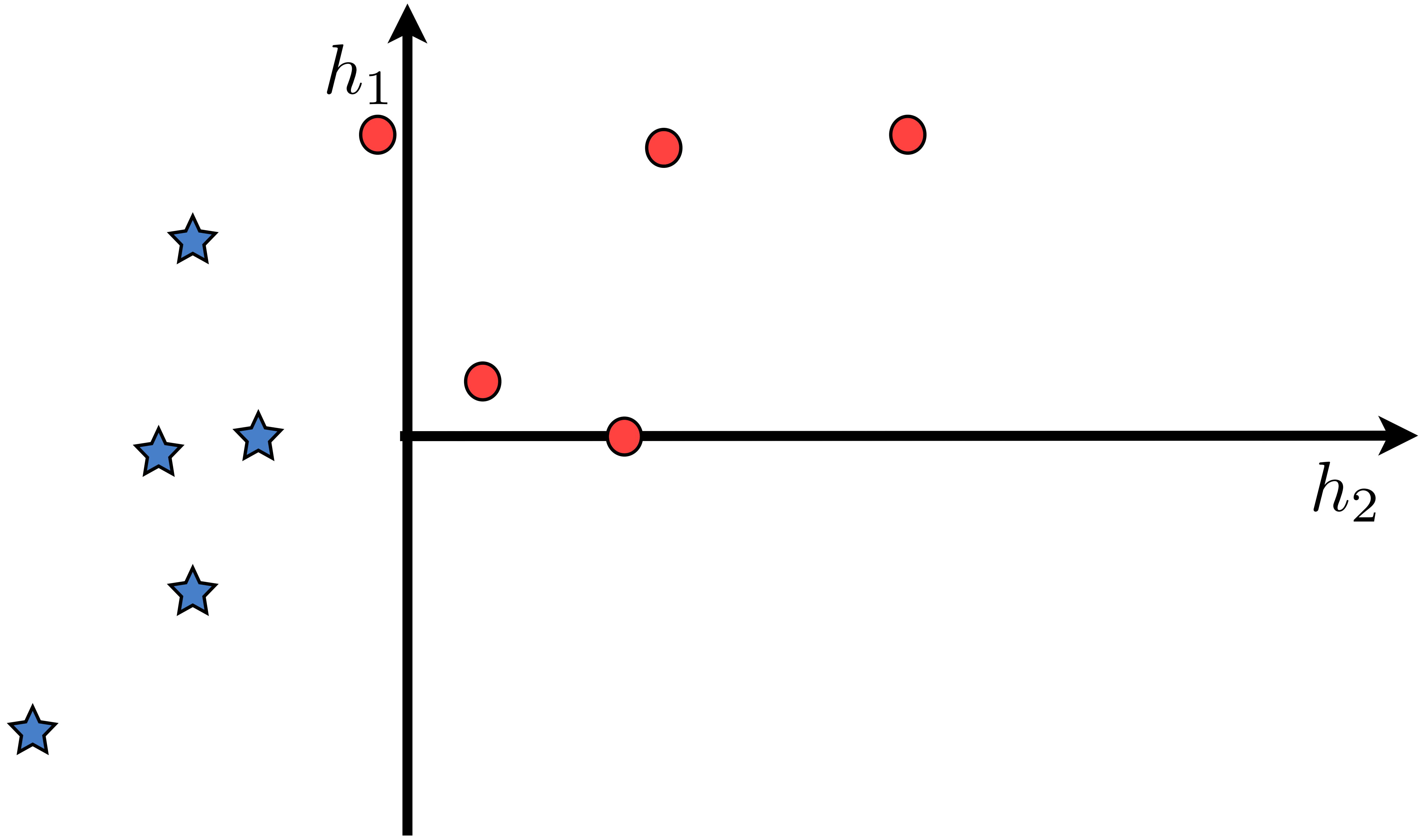
Improvement 2:

Add dimensions to make points **separable**

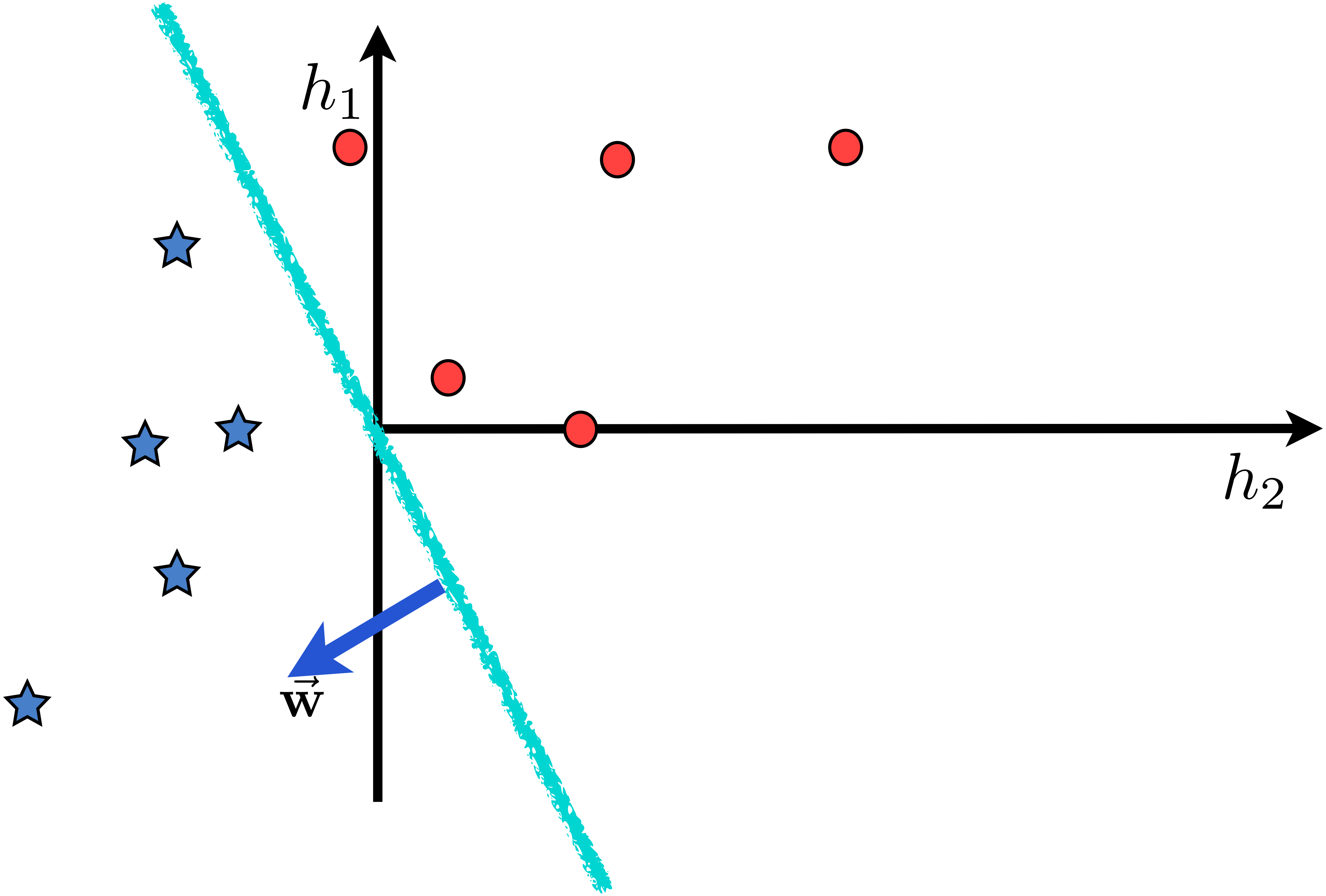
Parameter Learning: Review

K-Best List Example



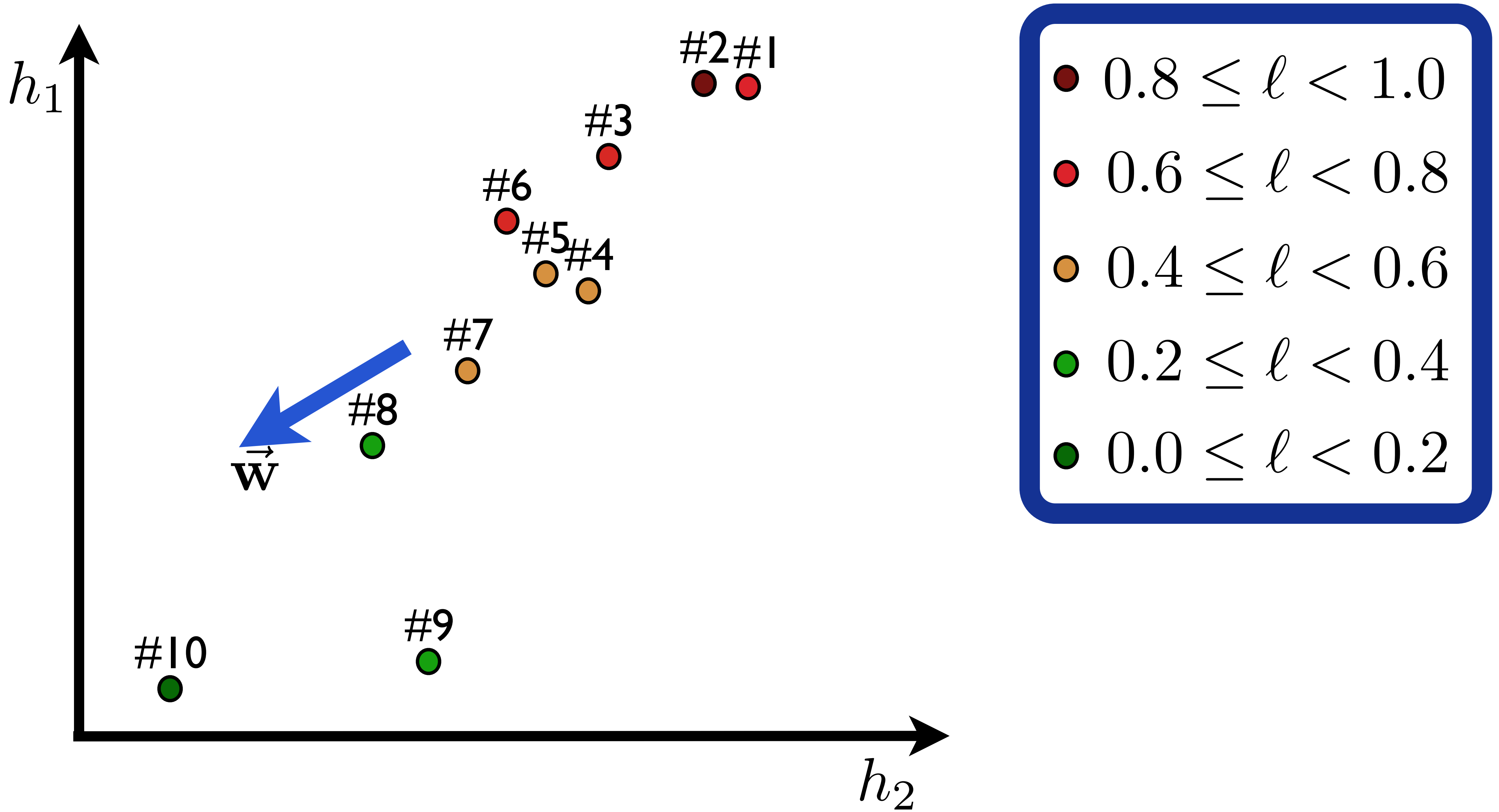


Fit a linear model



Fit a linear model

K-Best List Example



Limitations

- We can't optimize corpus-level metrics, like BLEU, on a test set
  - These don't decompose by sentence!
- We turn now to a kind of "direct cost minimization"

MERT

- **M**inimum **E**rror **R**ate **T**raining
- Directly target an automatic evaluation metric
  - BLEU is defined at the corpus level
  - MERT optimizes at the corpus level
- Downsides
  - Does not deal well with > ~20 features

MERT

Given weight vector $\mathbf{w}$, any hypothesis $\langle \mathbf{e}, \mathbf{a} \rangle$ will have a (scalar) score $m = \mathbf{w}^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a})$

Now pick a **search vector** $\mathbf{v}$, and consider how the score of this hypothesis will change:

$$\mathbf{w}_{\text{new}} = \mathbf{w} + \gamma \mathbf{v}$$

MERT

Given weight vector $\mathbf{w}$, any hypothesis $\langle \mathbf{e}, \mathbf{a} \rangle$ will have a (scalar) score $m = \mathbf{w}^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a})$

Now pick a **search vector** $\mathbf{v}$, and consider how the score of this hypothesis will change:

$$\mathbf{w}_{\text{new}} = \mathbf{w} + \gamma \mathbf{v}$$

$$m = (\mathbf{w} + \gamma \mathbf{v})^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a})$$

MERT

Given weight vector $\mathbf{w}$, any hypothesis $\langle \mathbf{e}, \mathbf{a} \rangle$ will have a (scalar) score $m = \mathbf{w}^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a})$

Now pick a **search vector** $\mathbf{v}$, and consider how the score of this hypothesis will change:

$$\begin{aligned} \mathbf{w}_{\text{new}} &= \mathbf{w} + \gamma \mathbf{v} \\ m &= (\mathbf{w} + \gamma \mathbf{v})^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a}) \\ &= \mathbf{w}^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a}) + \gamma\, \mathbf{v}^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a}) \end{aligned}$$

MERT

Given weight vector $\mathbf{w}$, any hypothesis $\langle \mathbf{e}, \mathbf{a} \rangle$ will have a (scalar) score $m = \mathbf{w}^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a})$

Now pick a **search vector** $\mathbf{v}$, and consider how the score of this hypothesis will change:

$$\mathbf{w}_{\text{new}} = \mathbf{w} + \gamma \mathbf{v}$$

$$\begin{aligned} m &= (\mathbf{w} + \gamma \mathbf{v})^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a}) \\ &= \underbrace{\mathbf{w}^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a})}_{b} + \gamma \underbrace{\mathbf{v}^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a})}_{a} \end{aligned}$$

$$m = a\gamma + b$$

MERT

Given weight vector $\mathbf{w}$, any hypothesis $\langle \mathbf{e}, \mathbf{a} \rangle$ will have a (scalar) score $m = \mathbf{w}^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a})$

Now pick a **search vector** $\mathbf{v}$, and consider how the score of this hypothesis will change:

$$\mathbf{w}_{\text{new}} = \mathbf{w} + \gamma \mathbf{v}$$

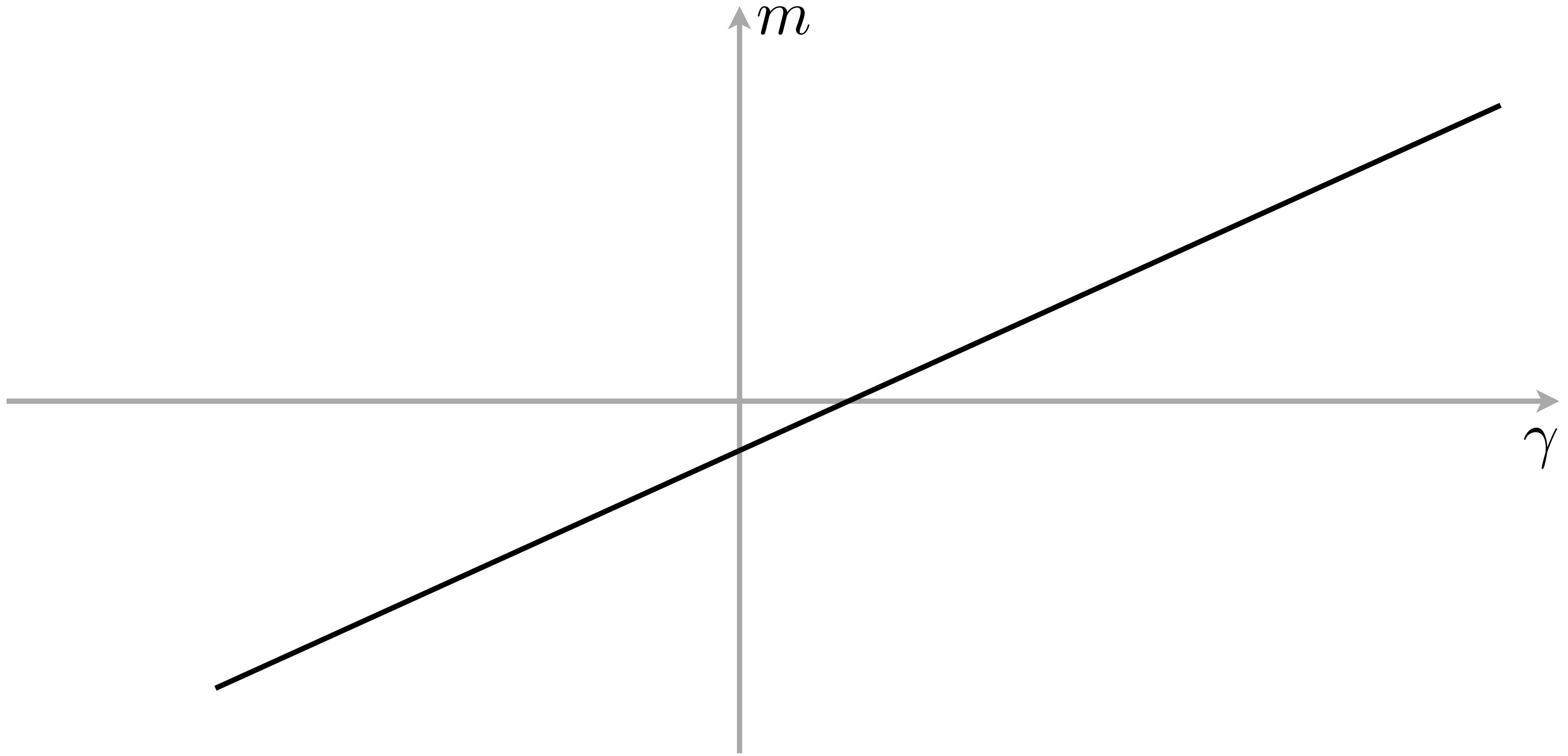
$$m = (\mathbf{w} + \gamma \mathbf{v})^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a})$$

$$= \underbrace{\mathbf{w}^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a})}_{b} + \gamma \underbrace{\mathbf{v}^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a})}_{a}$$

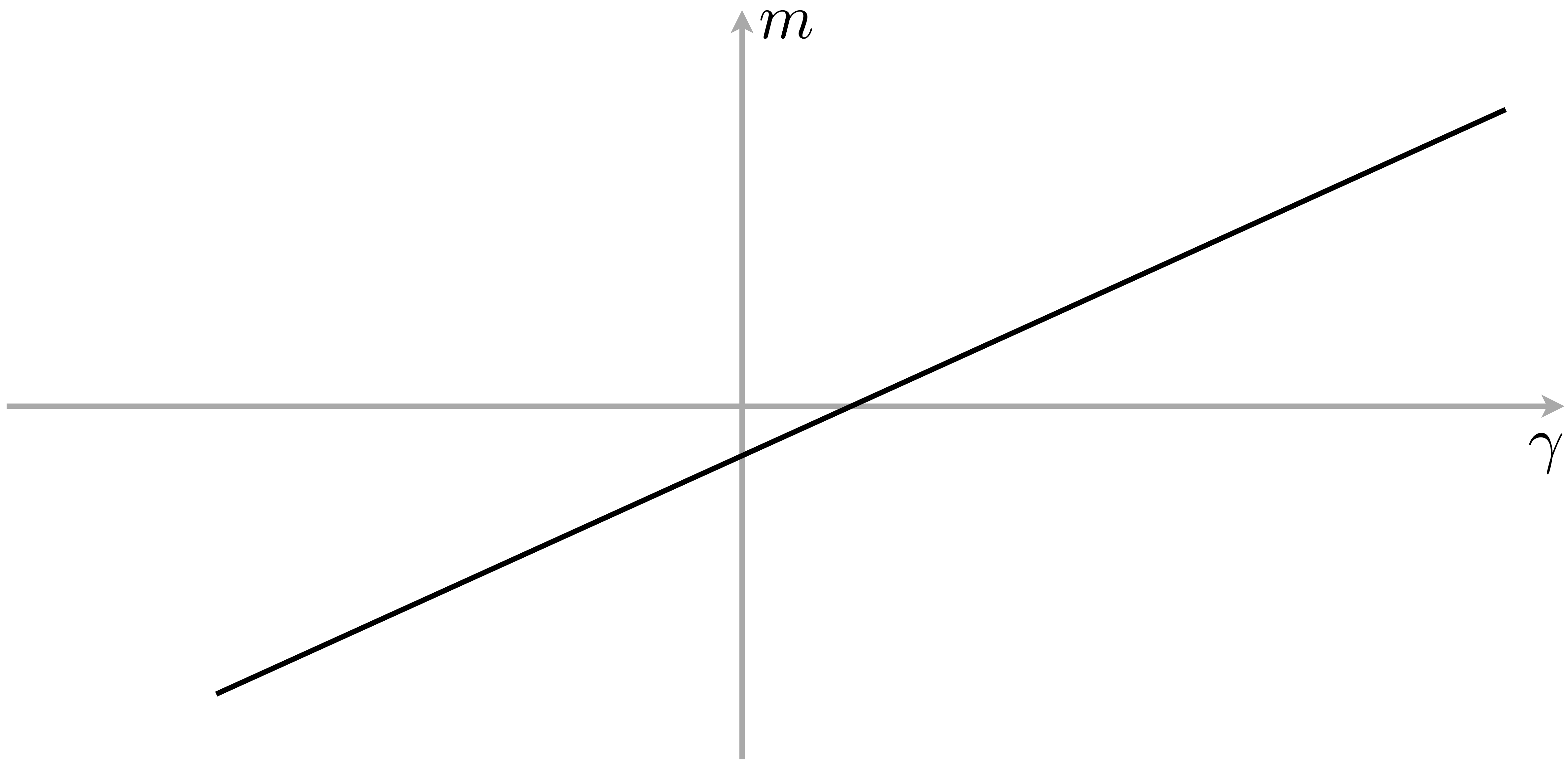
$$\boxed{m = a\gamma + b}$$

Linear function in 2D!

MERT

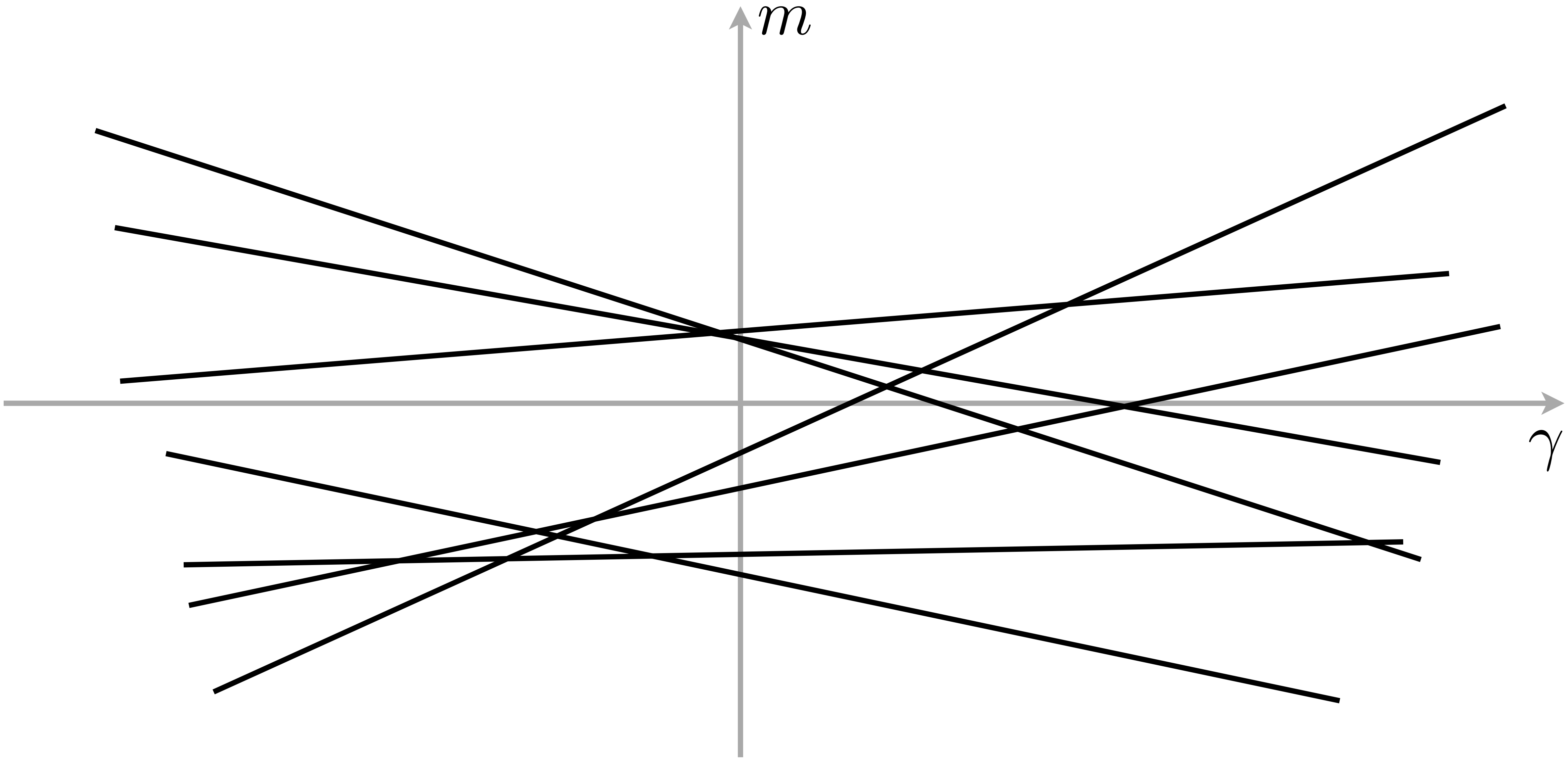


MERT



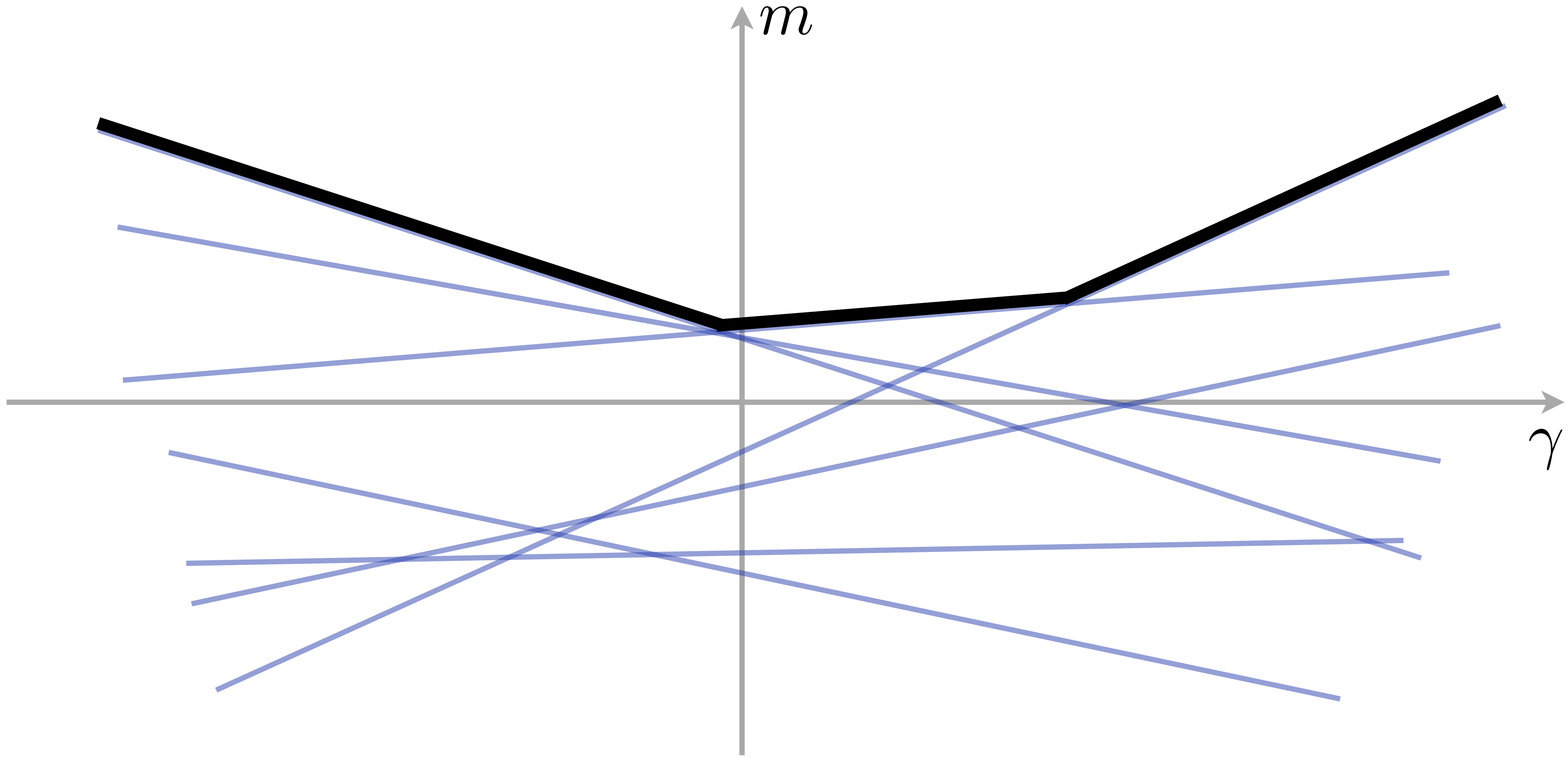
Recall our k-best set $\{\langle \mathbf{e}_i^*, \mathbf{a}_i^* \rangle\}_{i=1}^{K}$

MERT

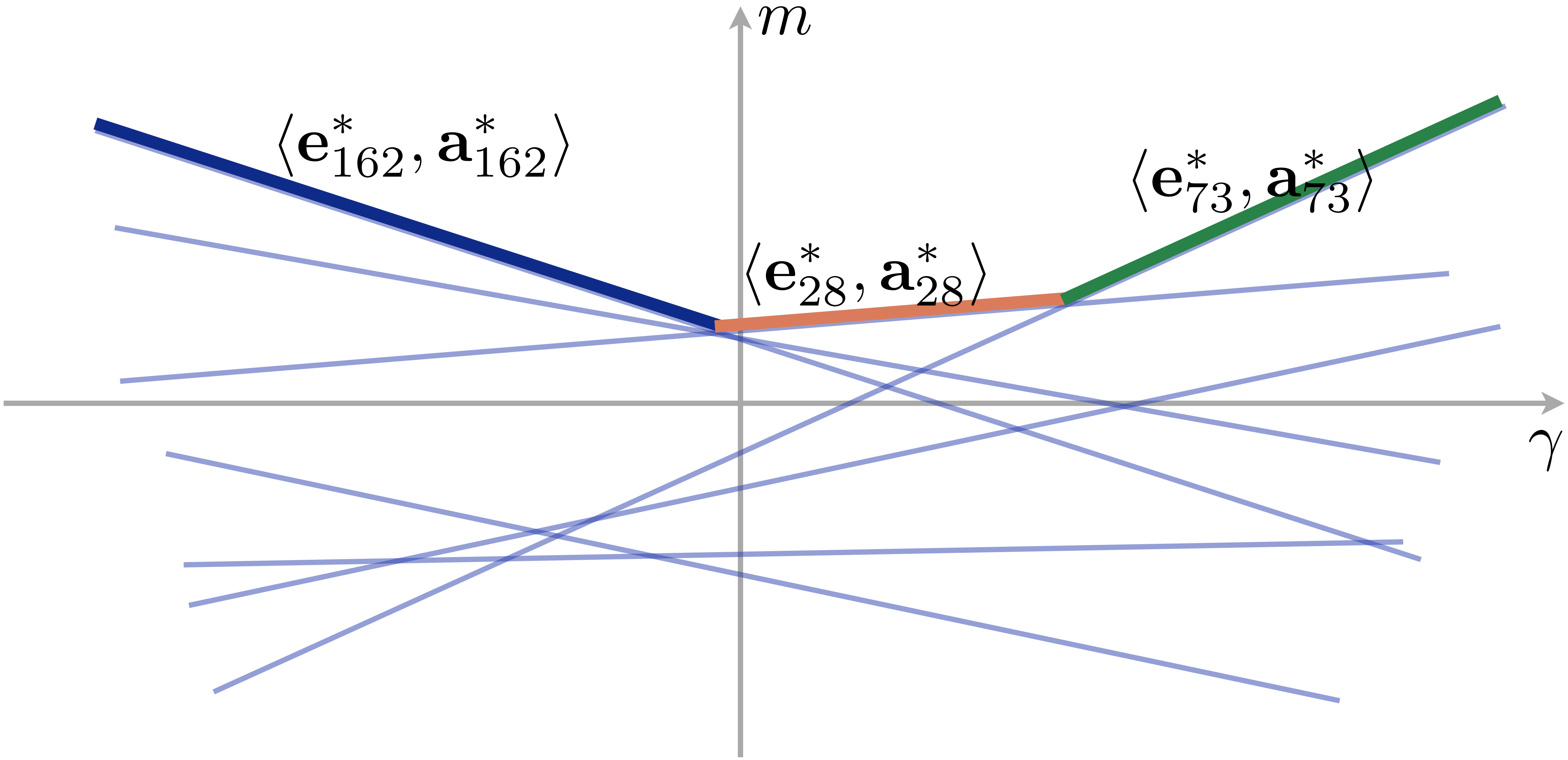


Recall our k-best set $\{\langle \mathbf{e}_i^*, \mathbf{a}_i^* \rangle\}_{i=1}^{K}$

MERT

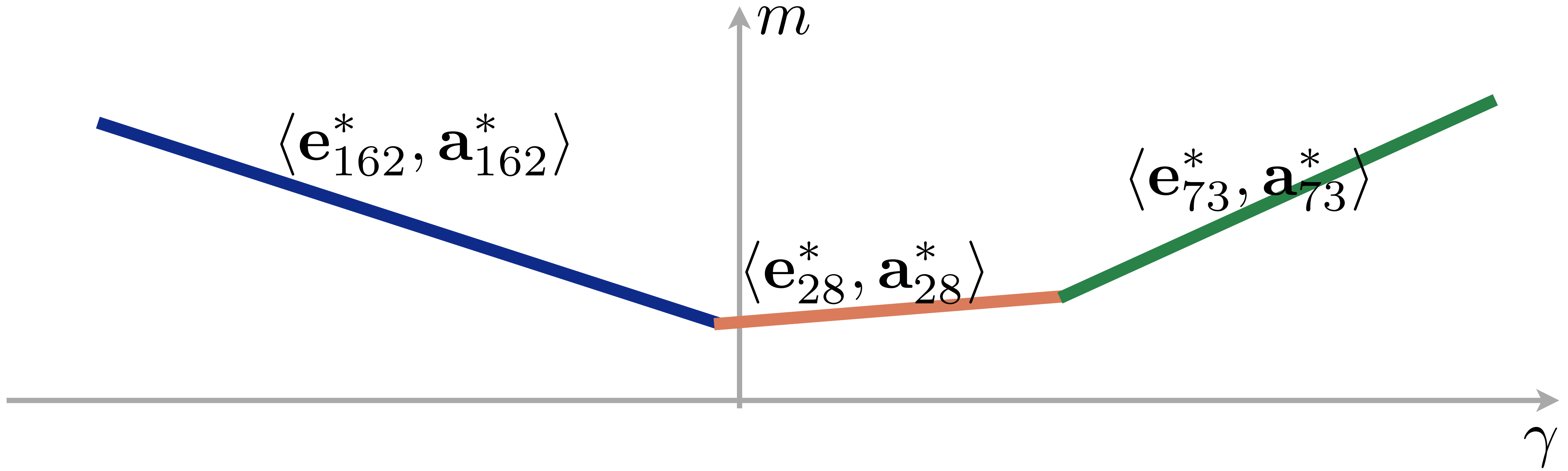


MERT



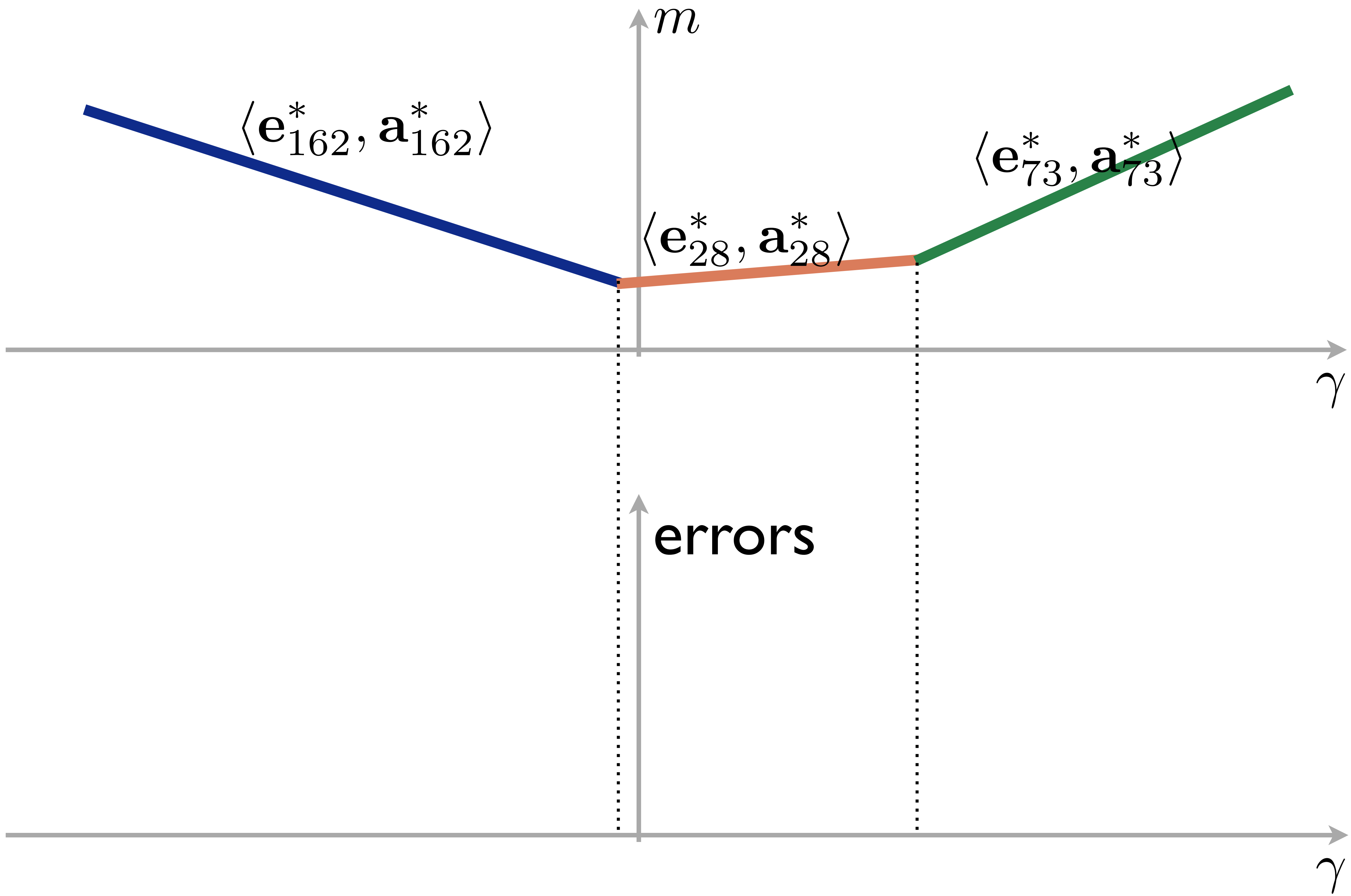
m
$\langle \mathbf{e}^*_{162}, \mathbf{a}^*_{162} \rangle$
$\langle \mathbf{e}^*_{73}, \mathbf{a}^*_{73} \rangle$
$\langle \mathbf{e}^*_{28}, \mathbf{a}^*_{28} \rangle$
γ

MERT



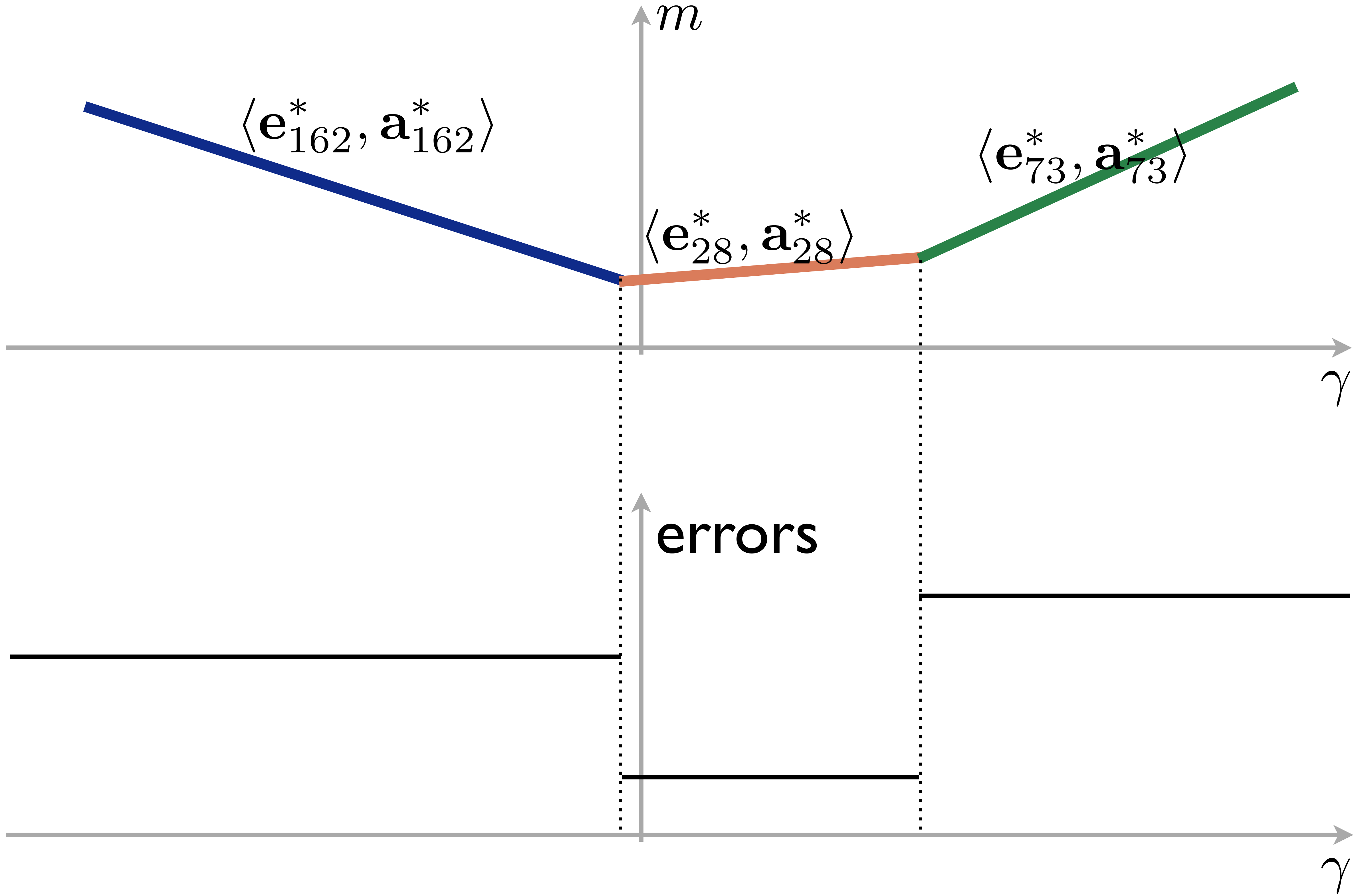
m
$\langle \mathbf{e}^*_{162}, \mathbf{a}^*_{162} \rangle$
$\langle \mathbf{e}^*_{73}, \mathbf{a}^*_{73} \rangle$
$\langle \mathbf{e}^*_{28}, \mathbf{a}^*_{28} \rangle$
γ

MERT



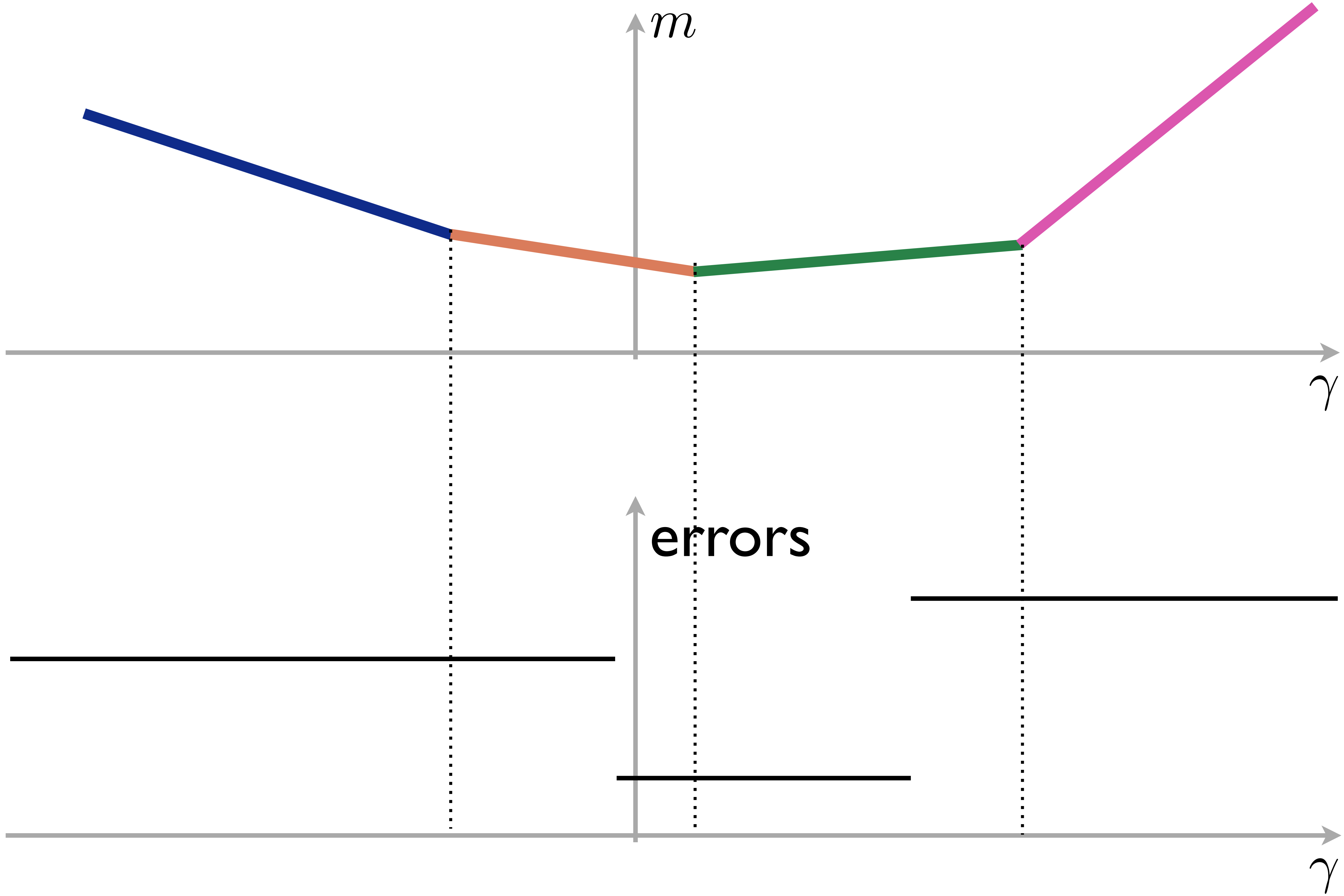
m
$\langle \mathbf{e}^*_{162}, \mathbf{a}^*_{162} \rangle$
$\langle \mathbf{e}^*_{73}, \mathbf{a}^*_{73} \rangle$
$\langle \mathbf{e}^*_{28}, \mathbf{a}^*_{28} \rangle$
γ
errors
γ

MERT

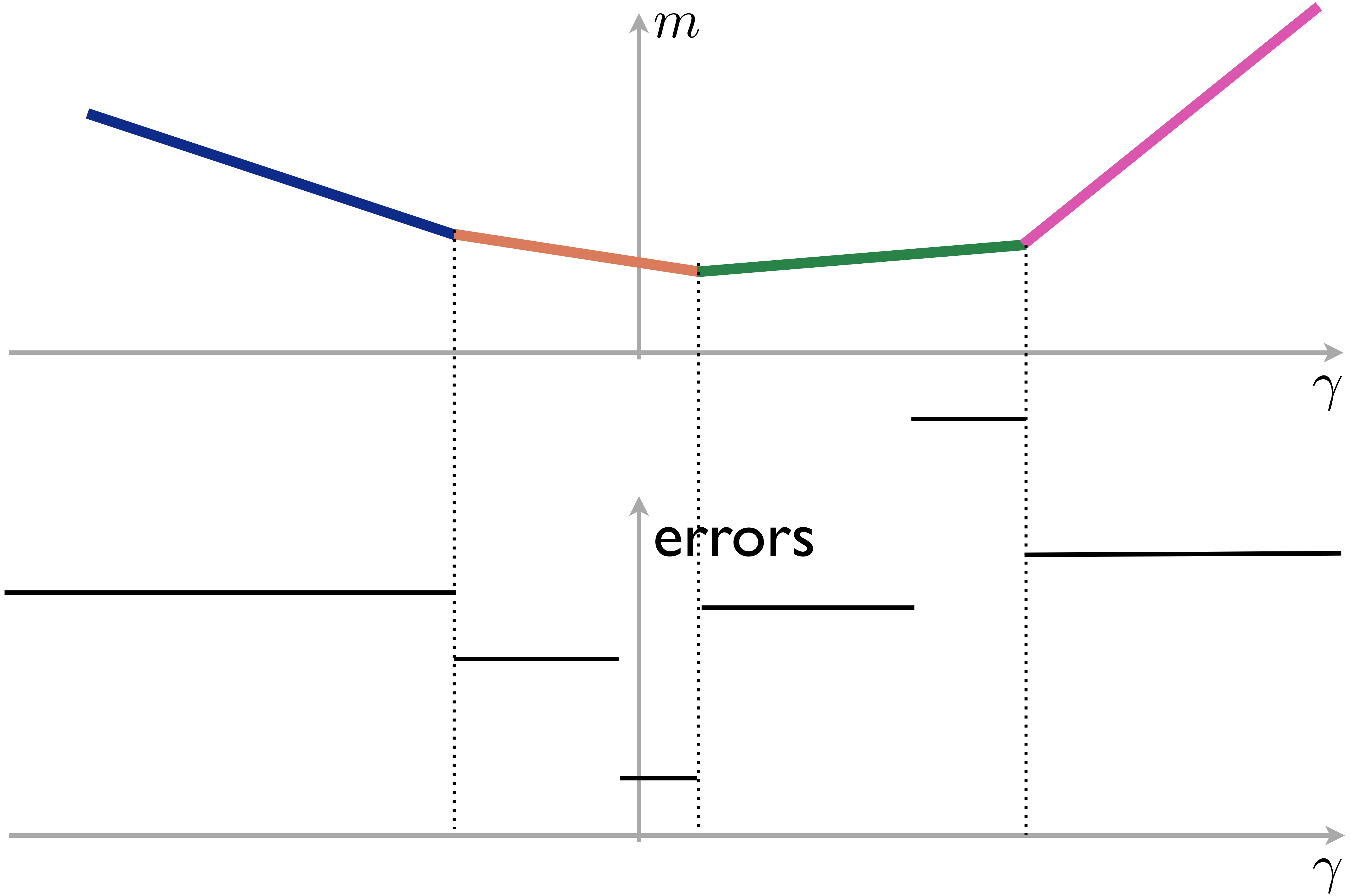


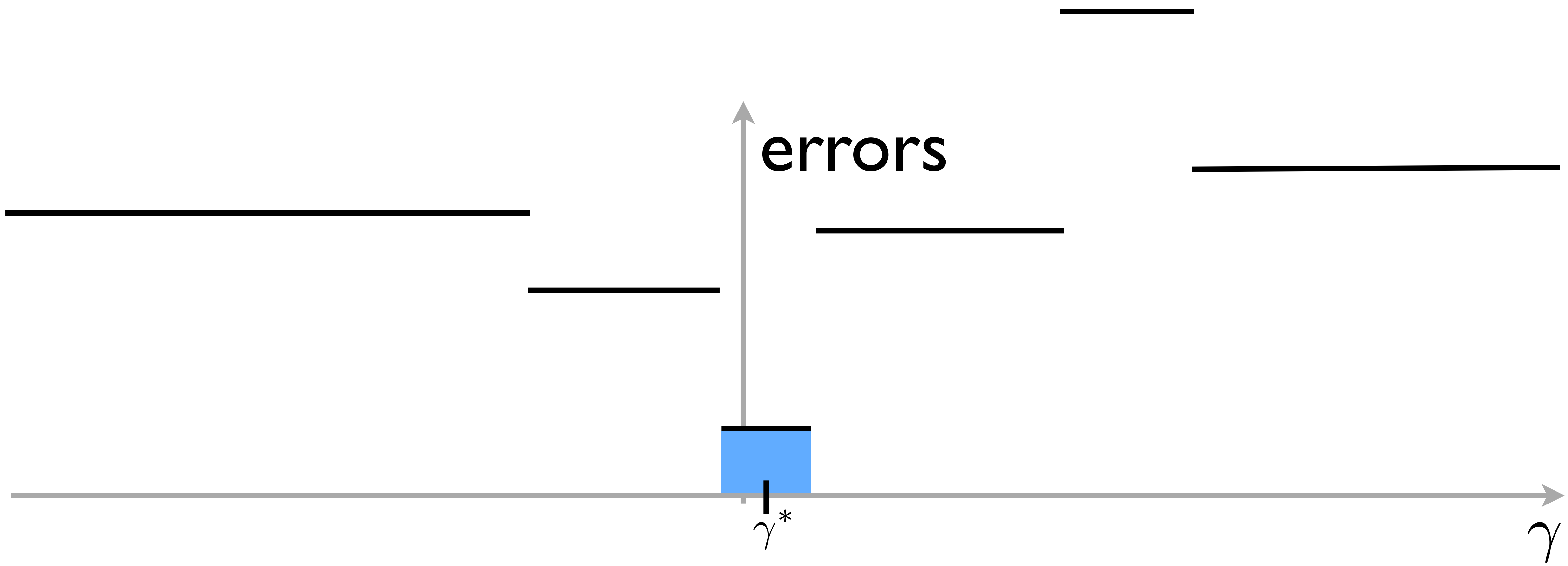
m
$\langle \mathbf{e}_{162}^*, \mathbf{a}_{162}^* \rangle$
$\langle \mathbf{e}_{73}^*, \mathbf{a}_{73}^* \rangle$
$\langle \mathbf{e}_{28}^*, \mathbf{a}_{28}^* \rangle$
γ
errors
γ

MERT



MERT





Let $\mathbf{w}_{\text{new}} = \gamma^* \mathbf{v} + \mathbf{w}$

MERT

- In practice “errors” are sufficient statistics for evaluation metrics (e.g., BLEU)
  - Can maximize or minimize
- **How do you pick the search direction?**

Dynamic Programming MERT

Other Algorithms

- Given a hypergraph translation space
- In the Viterbi (Inside) algorithm, there are two operations
  - **Multiplication** (extend path)
  - **Maximization** (choose between paths)
- Semirings generalize these to compute other quantities

Semirings

semiring	$\mathbb{K}$	$\oplus$	$\otimes$	$\bar{0}$	$\bar{1}$	notes
Boolean	$\{0,1\}$	$\vee$	$\wedge$	0	1	idempotent
count	$\mathbb{N}_0 \cup \{\infty\}$	$+$	$\times$	0	1	
probability	$\mathbb{R}_+ \cup \{\infty\}$	$+$	$\times$	0	1	
tropical	$\mathbb{R} \cup \{-\infty, \infty\}$	max	$+$	$-\infty$	0	idempotent
log	$\mathbb{R} \cup \{-\infty, \infty\}$	$\oplus_{\log}$	$+$	$-\infty$	0	

Inside Algorithm

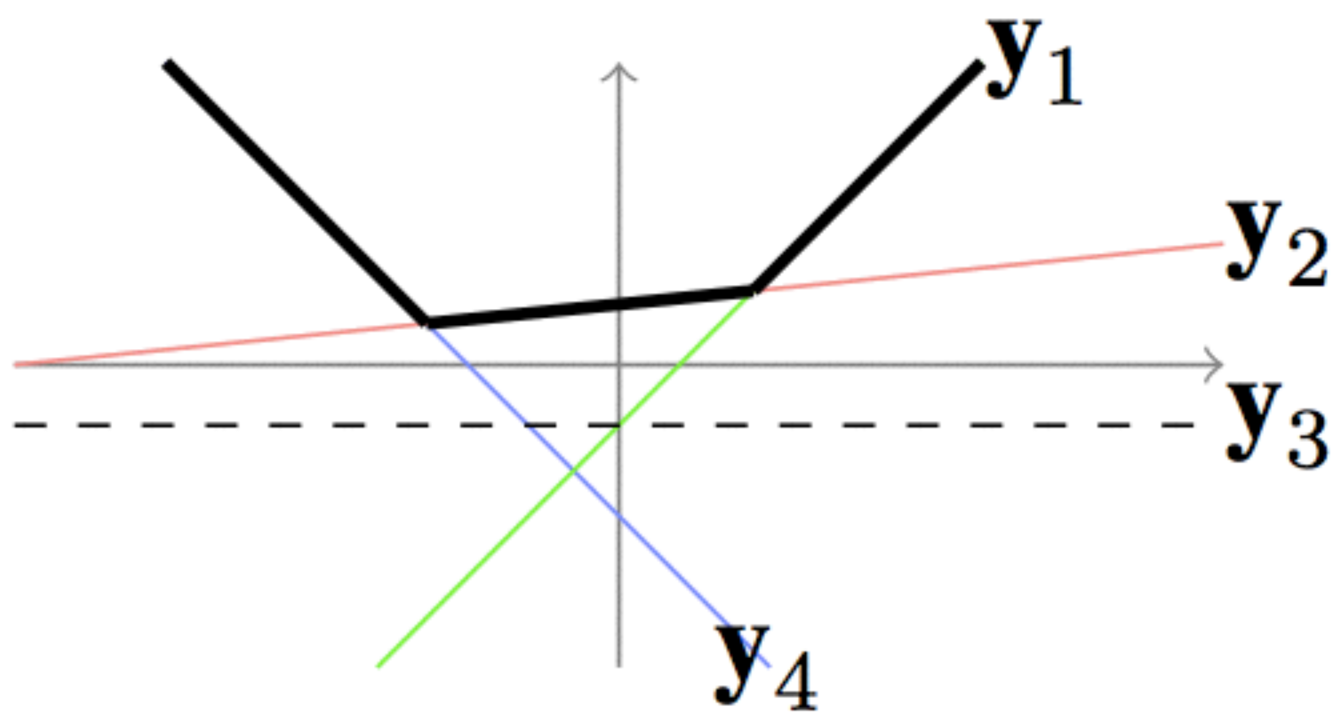
$$\alpha(q_{goal}) = \bigoplus_{\mathbf{d} \in \mathcal{G}} \bigotimes_{e \in \mathbf{d}} w(e)$$

function INSIDE($\mathcal{G}, K$) $\triangleright$ $\mathcal{G}$ is an acyclic hypergraph and K is a semiring
  for q in topological order in $\mathcal{G}$ **do**
    if $B(q) = \emptyset$ **then**
      $\alpha(q) \leftarrow \overline{1}$ $\triangleright$ assume states with no in-edges are axioms
    else
      $\alpha(q) \leftarrow \overline{0}$
      for all $e \in B(q)$ **do** $\triangleright$ all in-coming edges to node q
        $k \leftarrow w(e)$
        for all $r \in \mathbf{t}(e)$ **do** $\triangleright$ all tail (previous) nodes of edge e
          $k \leftarrow k \otimes \alpha(r)$
        $\alpha(q) \leftarrow \alpha(q) \oplus k$
  return α

Point-Line Duality

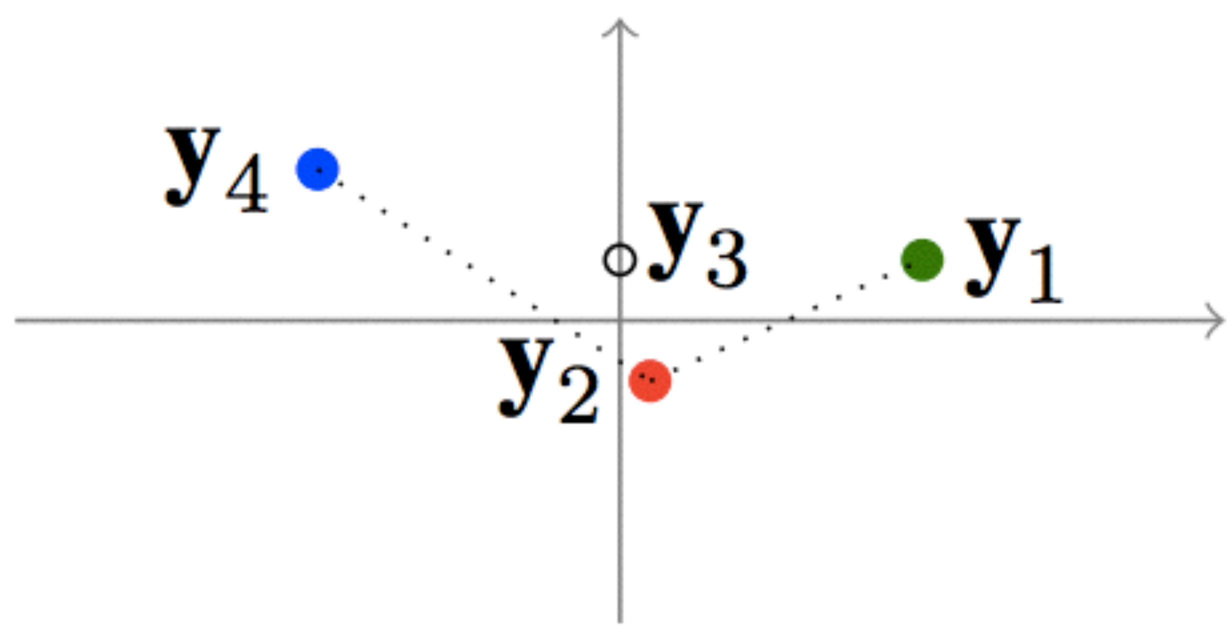
- Represent a set of lines as a set of points (and vice-versa)
  - $y = mx + b$ => $(m, -b)$
- The slope between dual points is the intersection x-axis of the pair of lines
- An upper envelope is dual to a **lower convex hull**

Primal



$\mathbf{y}_1$
$\mathbf{y}_2$
$\mathbf{y}_3$
$\mathbf{y}_4$

Dual



$\mathbf{y}_4$
$\mathbf{y}_3$
$\mathbf{y}_1$
$\mathbf{y}_2$

Convex Hull Semiring

Definition 2. ***The Convex Hull Semiring.*** *Let* $(\mathbb{K}, \oplus, \otimes, \overline{0}, \overline{1})$ *be defined as follows:*

$\mathbb{K}$	*A set of points in the plane that are the extreme points of a convex hull.*
$A \oplus B$	$\text{conv}\,[A \cup B]$
$A \otimes B$	*convex hull of the Minkowski sum, i.e.,* $\text{conv}\{(a_1 + b_1, a_2 + b_2) \mid (a_1, a_2) \in A \ \wedge \ (b_1, b_2) \in B\}$
$\overline{0}$	$\emptyset$
$\overline{1}$	$\{(0,0)\}$

Theorem 1. *The Convex Hull Semiring fulfills the semiring axioms and is commutative and idempotent.*

Theorem 2

- The Inside algorithm with the computes the convex hull dual to the MERT upper envelope generated from the ∞-best list of derivations

Summary

- Evaluation metrics
  - Figure out how well we're doing
  - Figure out if a feature helps
  - **Train your system**
- What's a great way to improve translation?
  - **Improve evaluation!**