6.036: Introduction to Machine Learning

Lecture start: Tuesdays 9:35am

Who's talking? Prof. Tamara Broderick

Questions? Ask on Piazza: "lecture (week) 4" folder

Materials: slides, video will all be available on Canvas

Live Zoom feed: https://mit.zoom.us/j/94238622313

Last Time(s)

I. Linear regression
   - data, hypothesis class, loss, regularizer

II. Gradient descent & SGD

Today's Plan

I. Linear classification

II. Linear logistic classification/logistic regression

Recall

Recall

Regression

Recall

Regression

- Datum i:

Recall

Regression

- Datum i: feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

Recall

Regression

- Datum i: feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

  - Label $y^{(i)} \in \mathbb{R}$

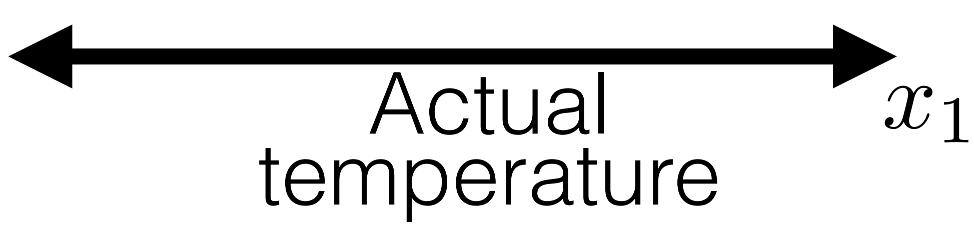
Recall

Regression

- Datum i: feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

  - Label $y^{(i)} \in \mathbb{R}$



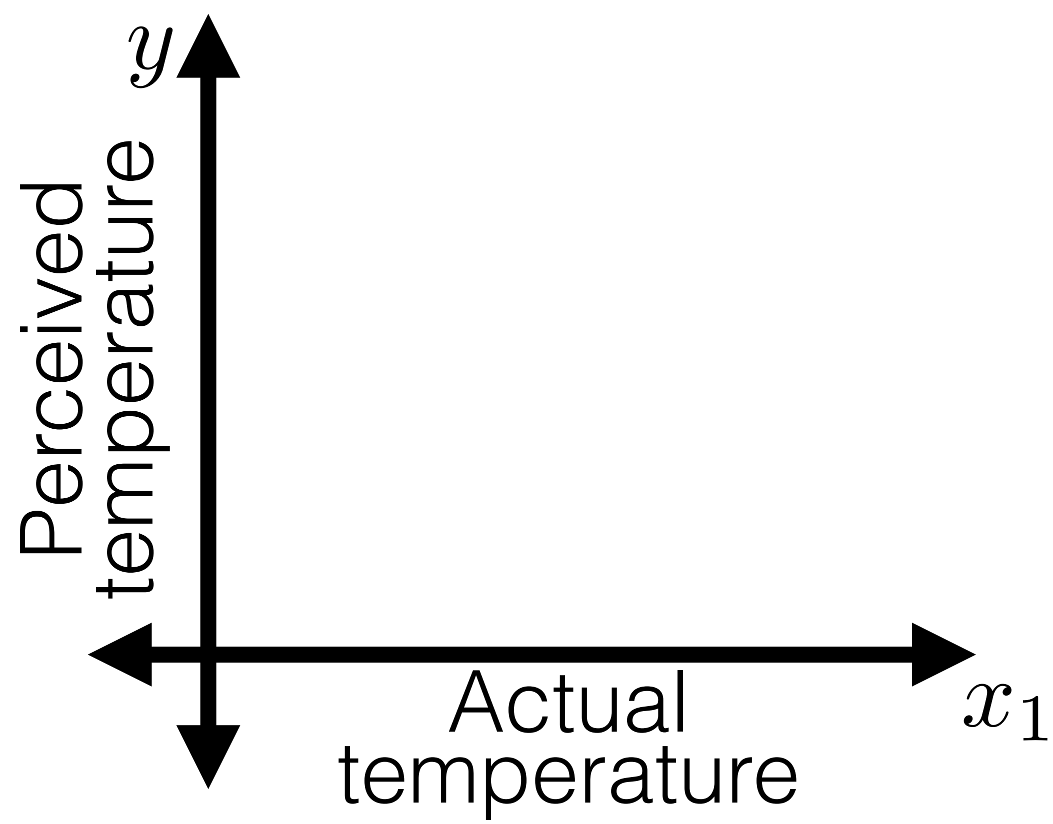
Recall

Regression

- Datum i: feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

  - Label $y^{(i)} \in \mathbb{R}$



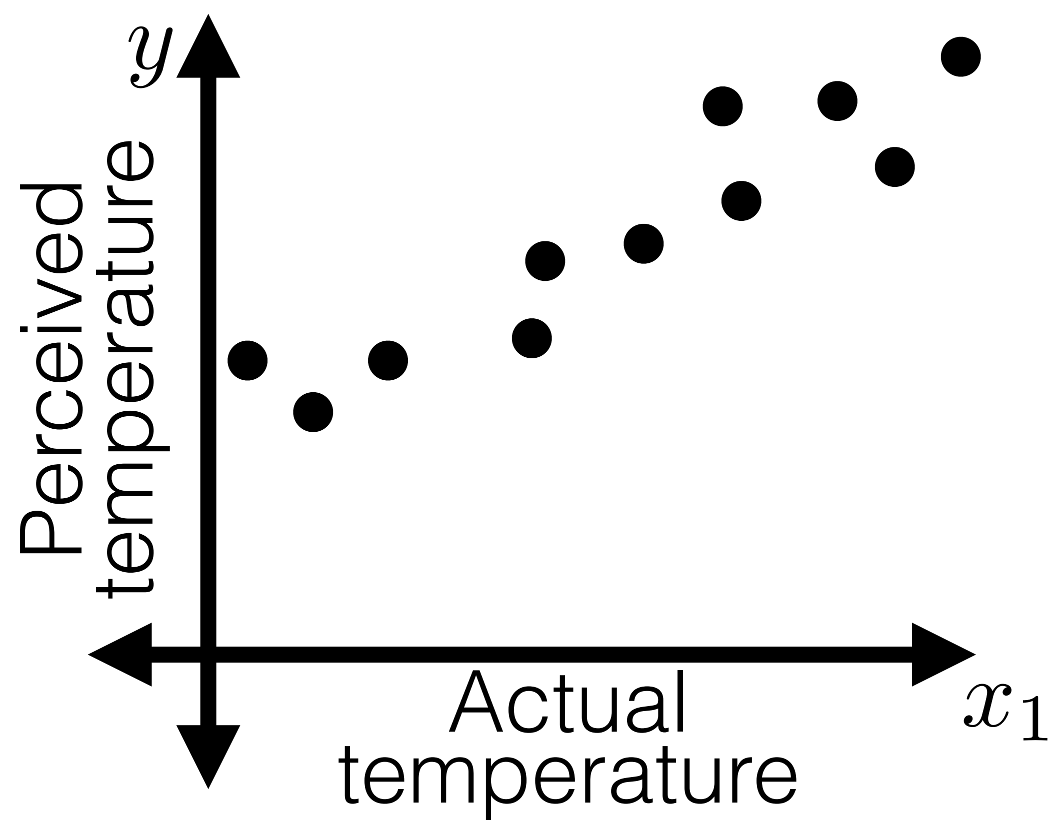
Recall

Regression

- Datum i: feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

  - Label $y^{(i)} \in \mathbb{R}$



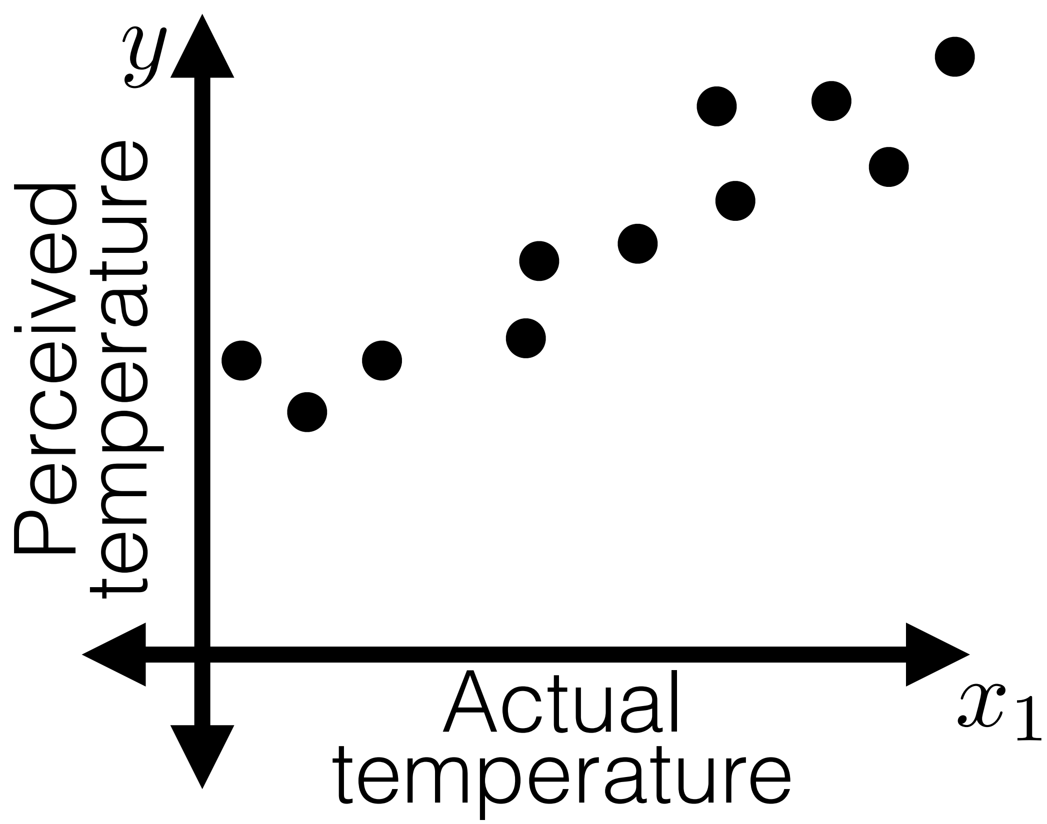
Recall

Regression

- Datum i: feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

  - Label $y^{(i)} \in \mathbb{R}$
- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$



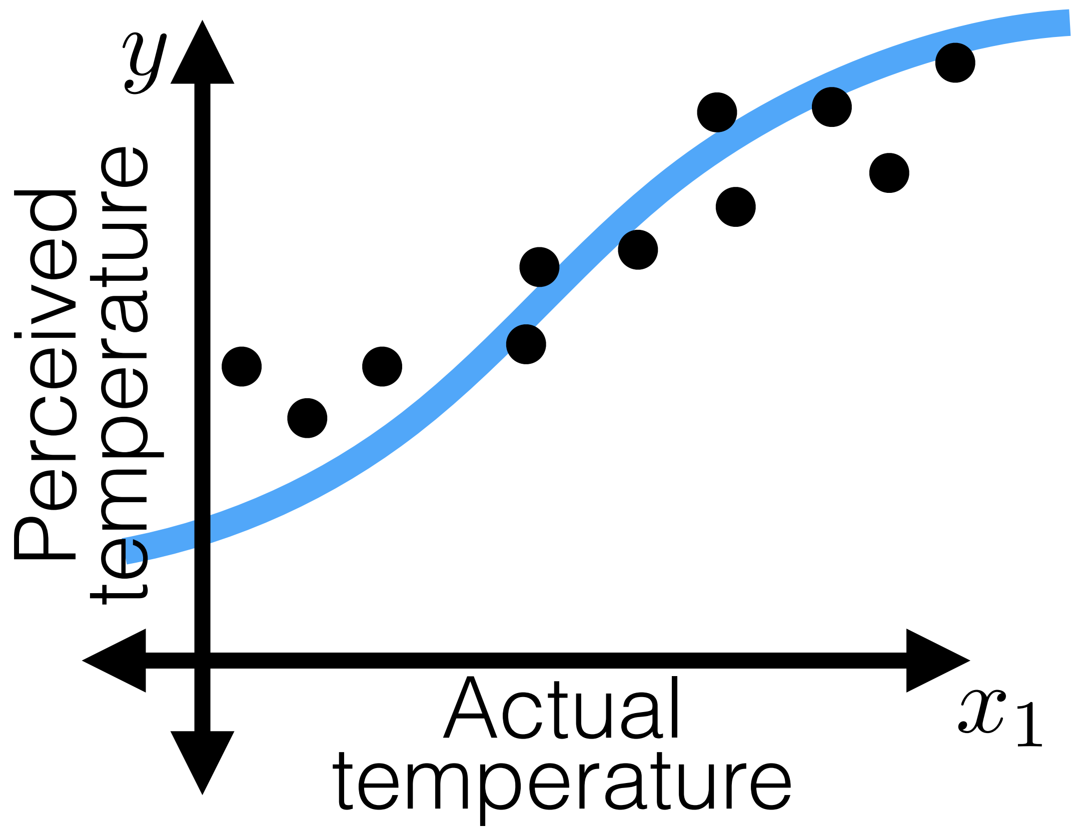
Recall

Regression

- Datum i: feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

  - Label $y^{(i)} \in \mathbb{R}$
- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$



Recall

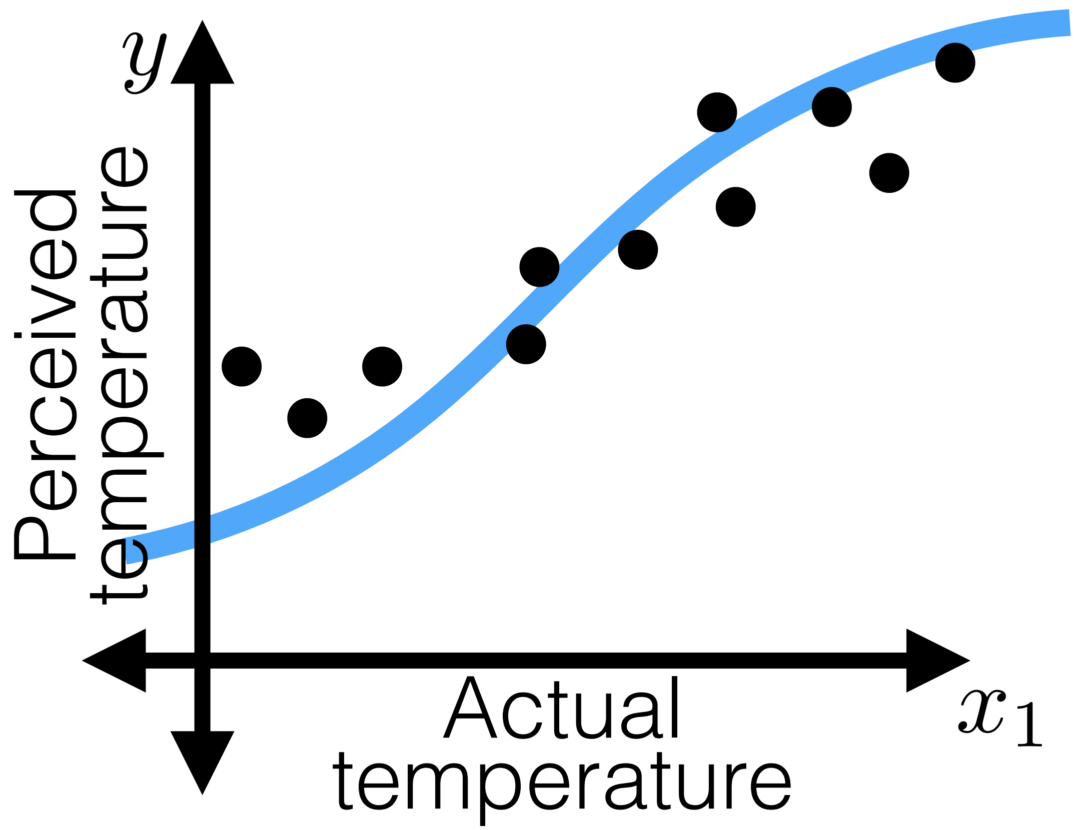
Compare

Regression

- Datum i: feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

  - Label $y^{(i)} \in \mathbb{R}$
- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$



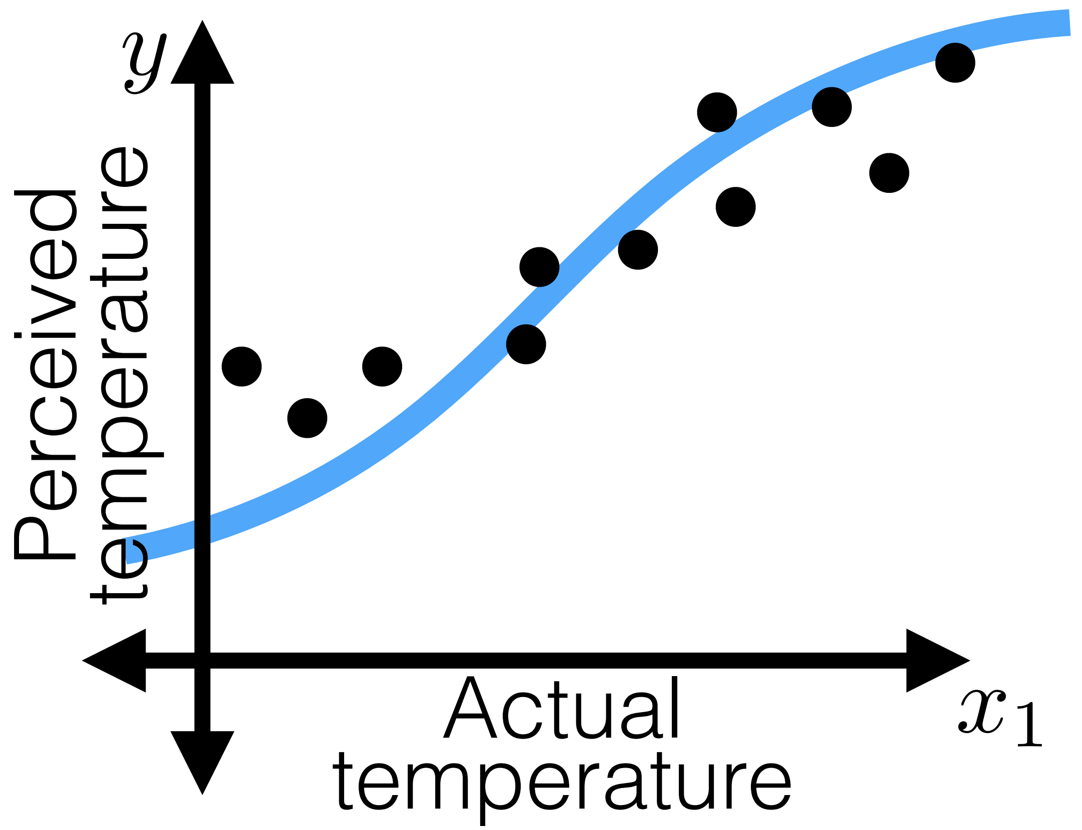
Recall

Regression

- Datum i: feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

  - Label $y^{(i)} \in \mathbb{R}$
- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$



Compare

(Two-class) Classification

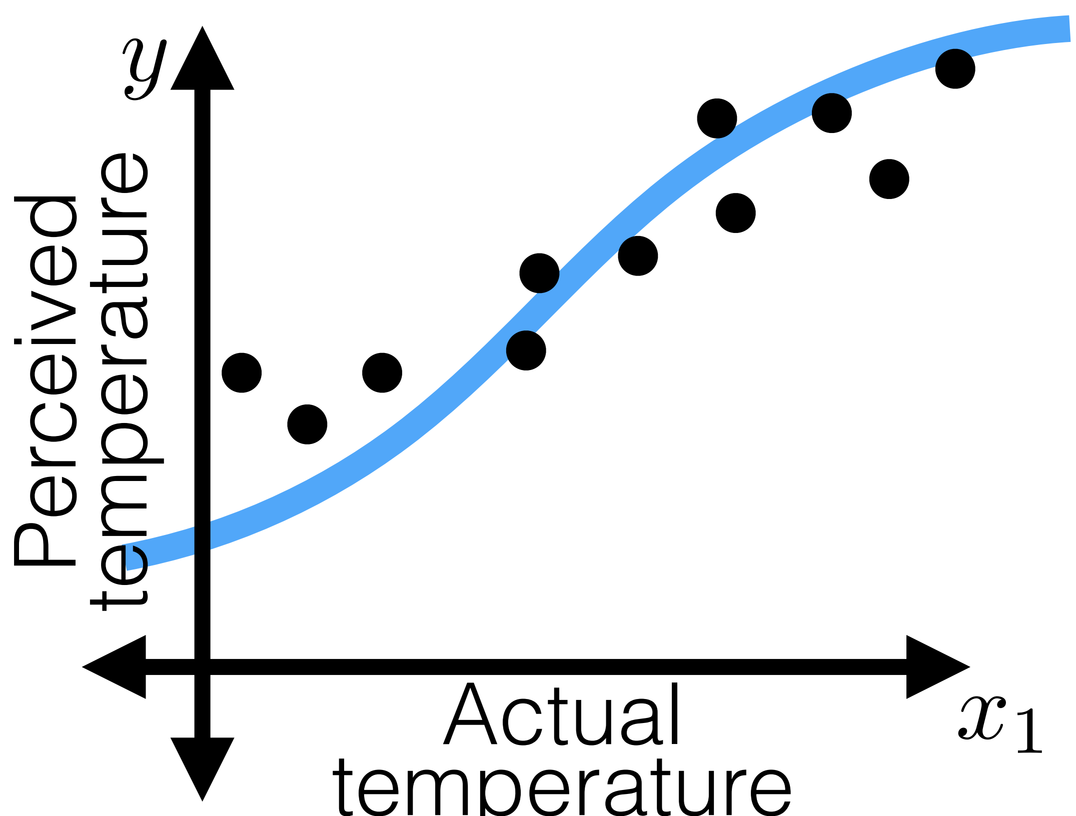
Recall

Regression

- Datum i: feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

  - Label $y^{(i)} \in \mathbb{R}$
- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$



Compare

(Two-class) Classification

- Datum i : feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

Recall

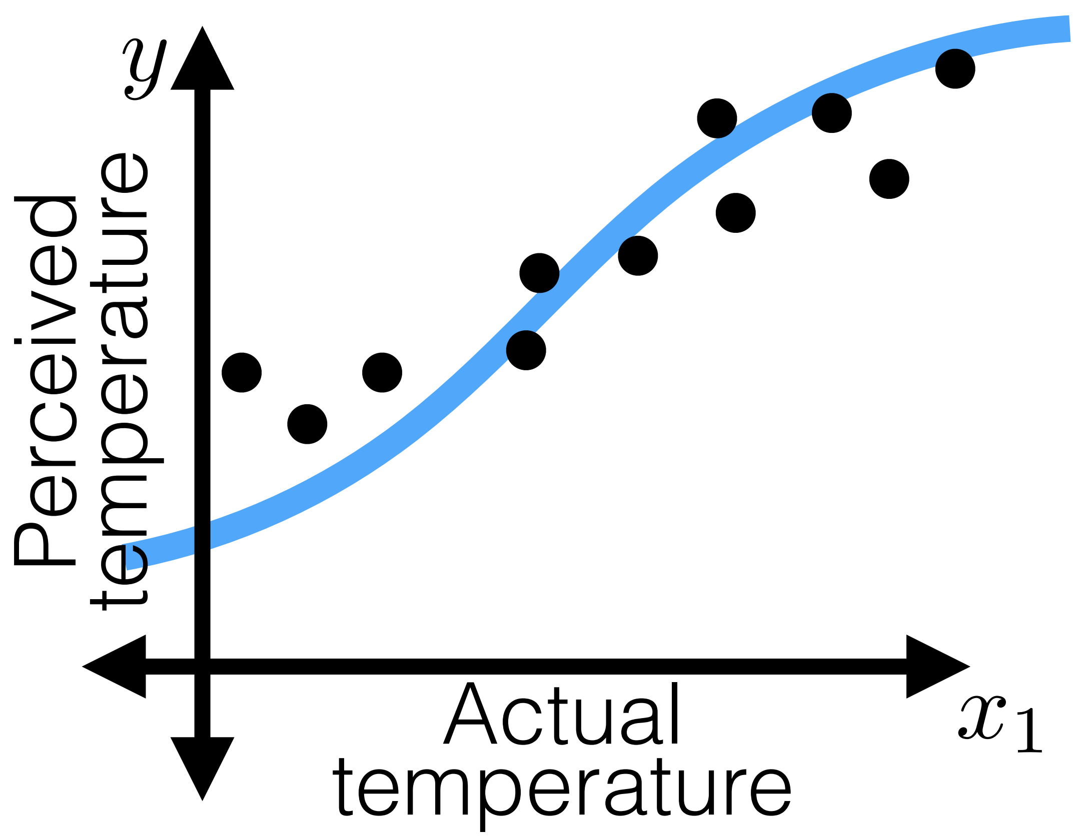
Regression

- Datum i: feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \mathbb{R}$

- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$



Compare

(Two-class) Classification

- Datum i : feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \{-1, +1\}$

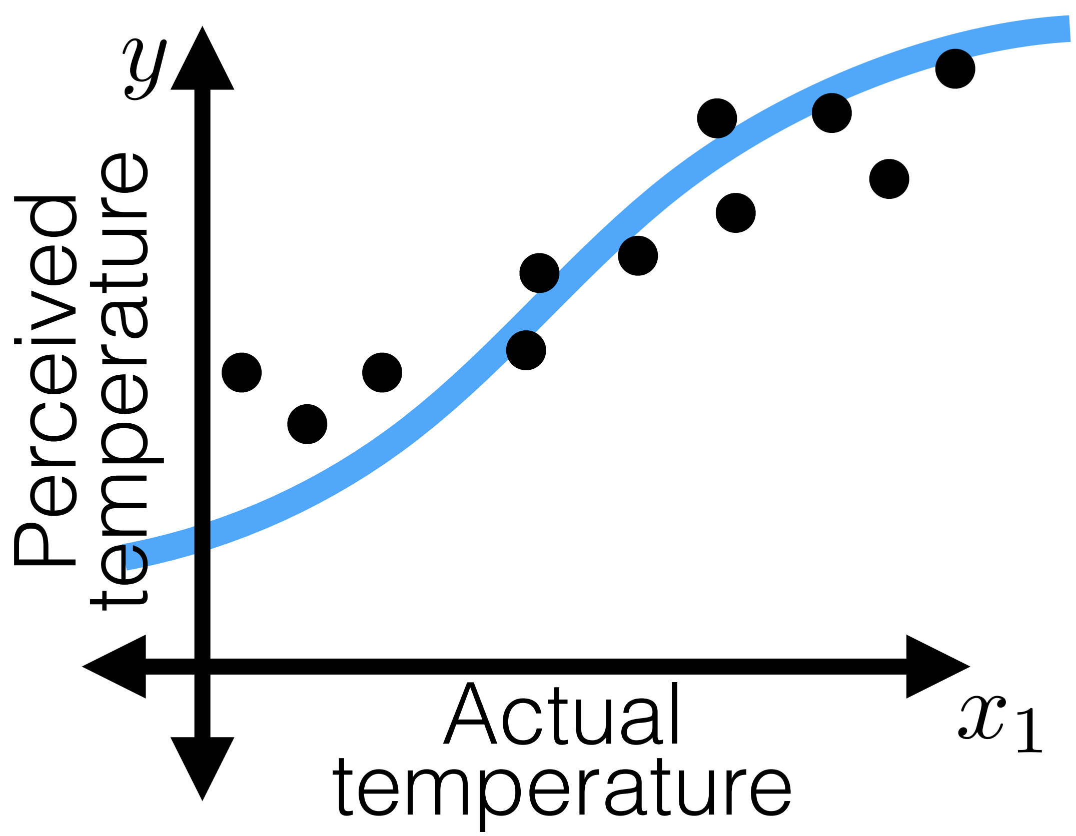
Recall

Regression

- Datum i: feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \mathbb{R}$
- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$



Compare

(Two-class) Classification

- Datum i : feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \{-1, +1\}$

Recall

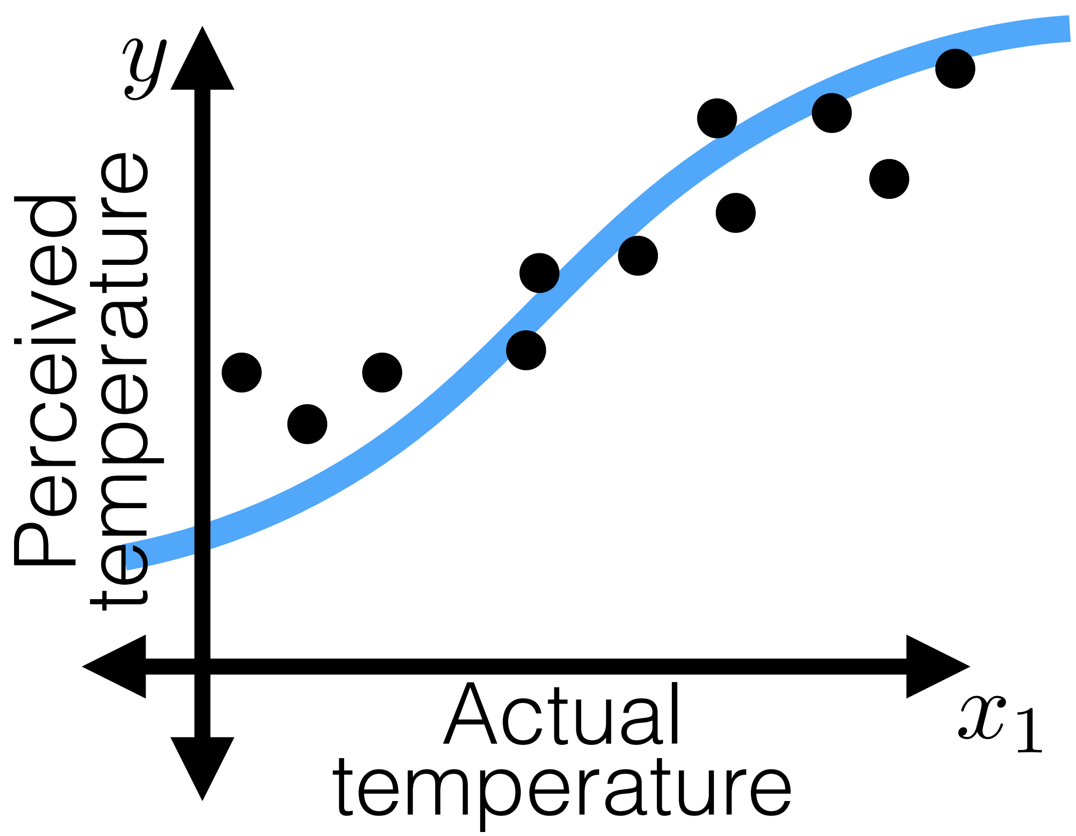
Regression

- Datum i: feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \mathbb{R}$

- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$



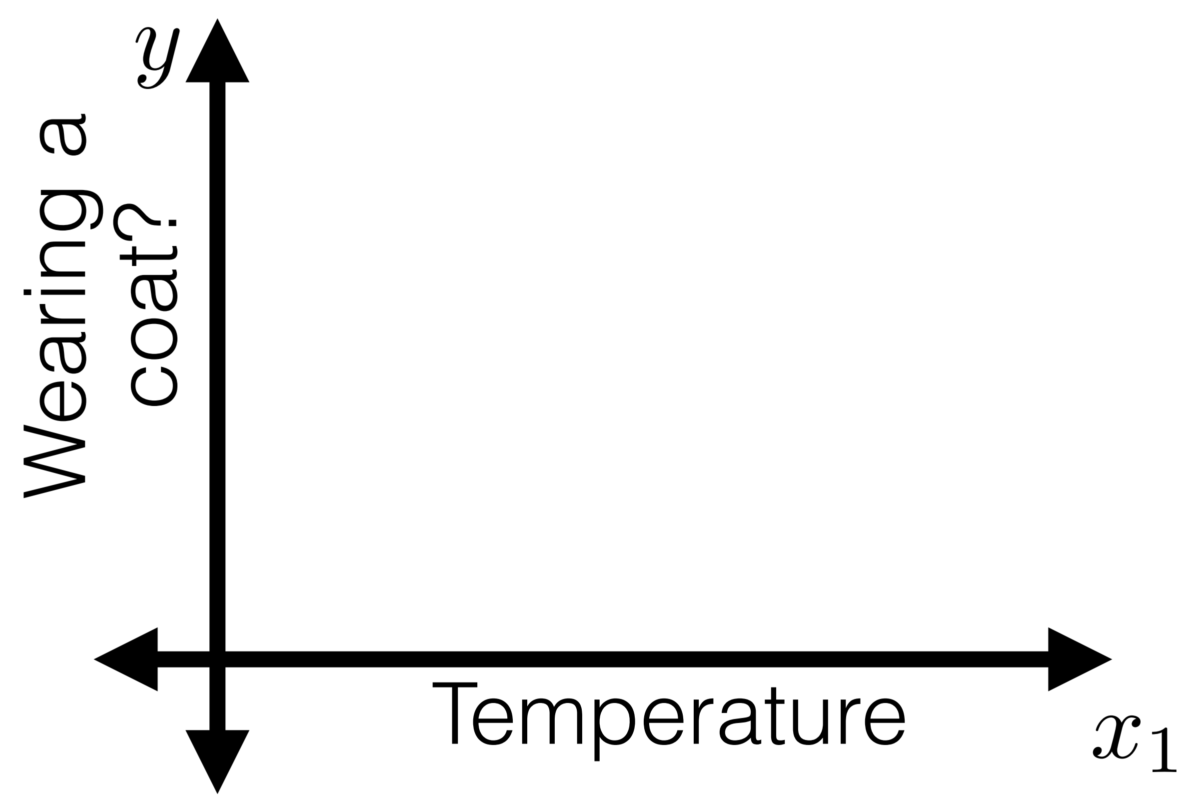
Compare

(Two-class) Classification

- Datum i : feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \{-1, +1\}$



Recall

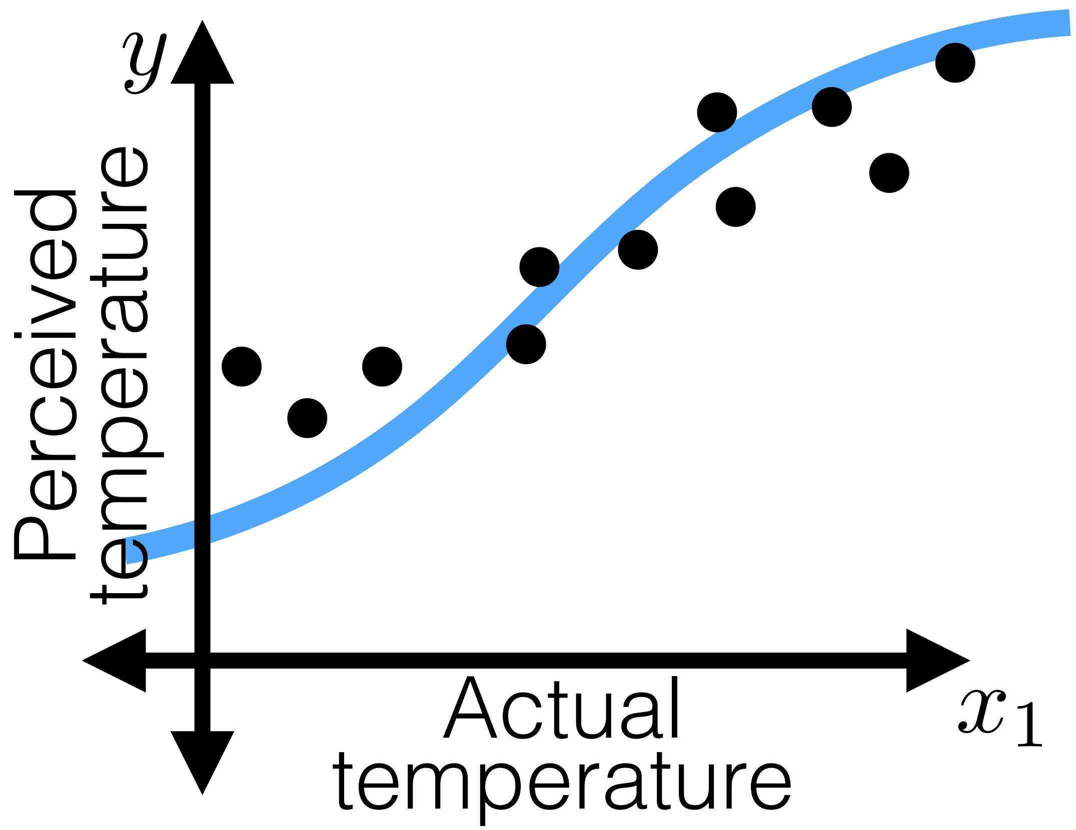
Regression

- Datum i: feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \mathbb{R}$

- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$



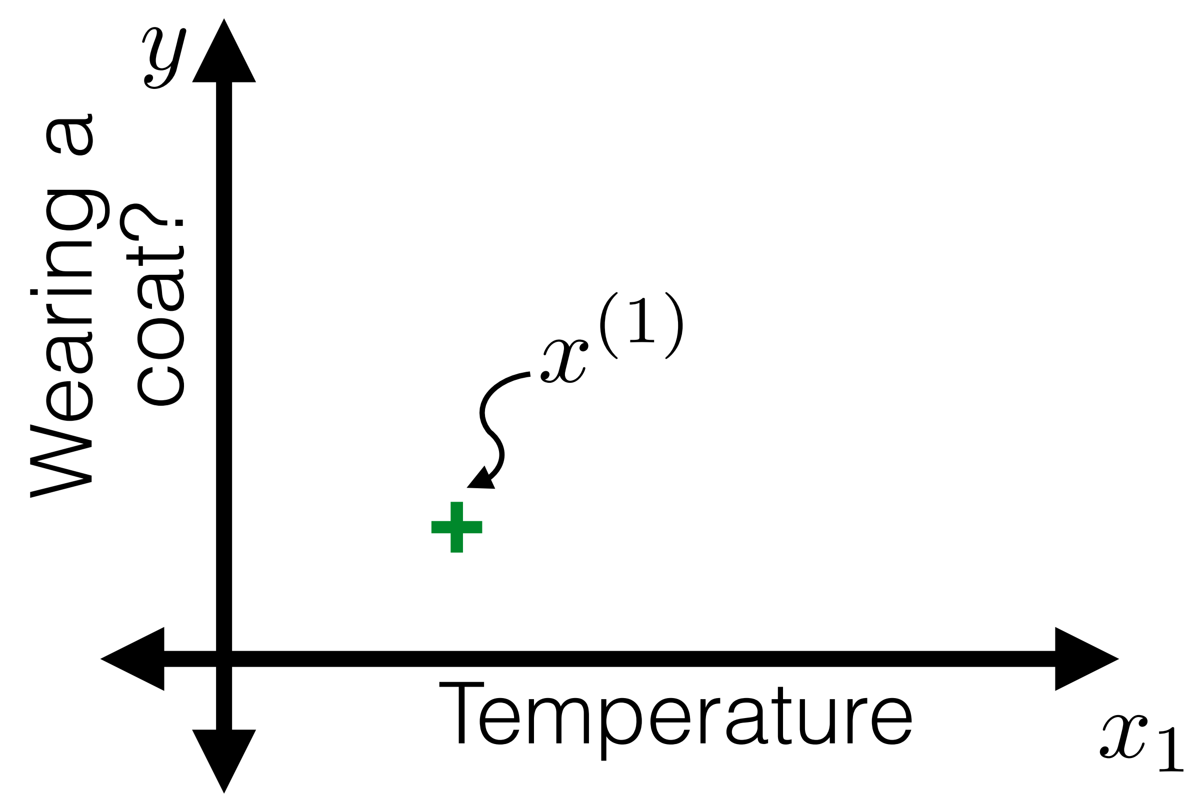
Compare

(Two-class) Classification

- Datum i : feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \{-1, +1\}$



Recall

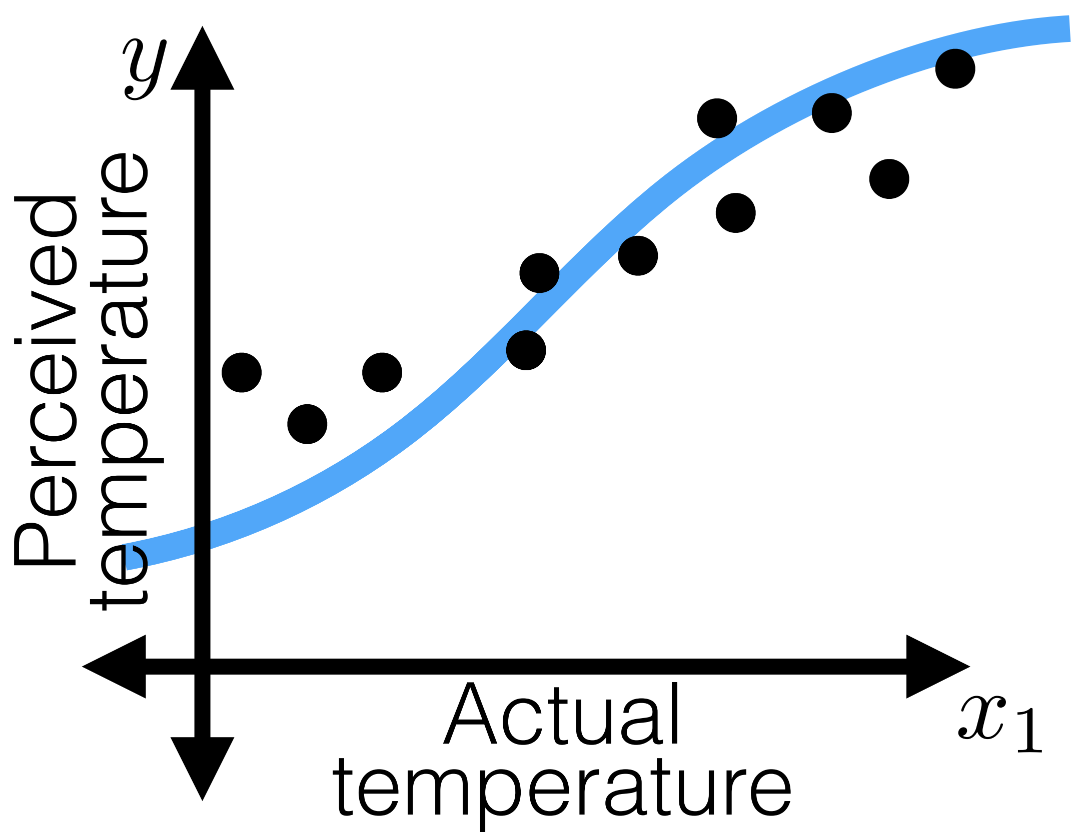
Regression

- Datum i: feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \mathbb{R}$

- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$



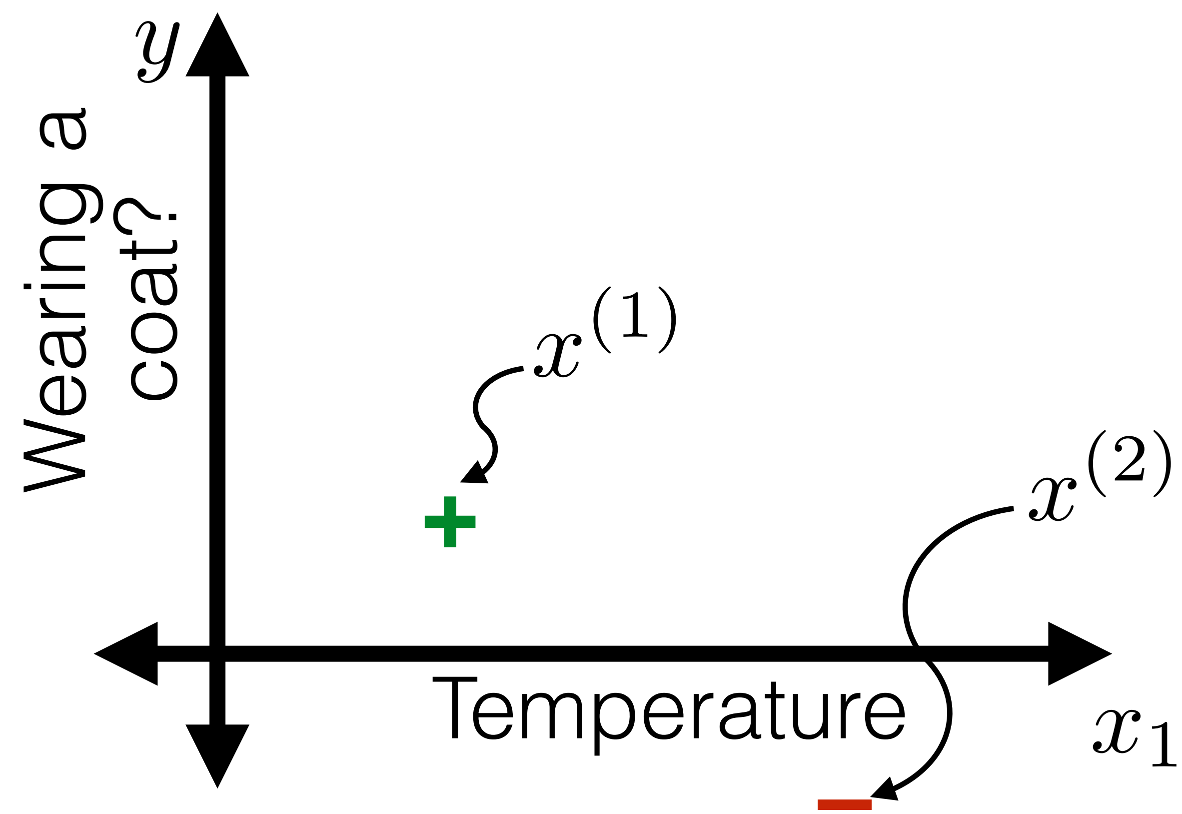
Compare

(Two-class) Classification

- Datum i: feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \{-1, +1\}$



Recall

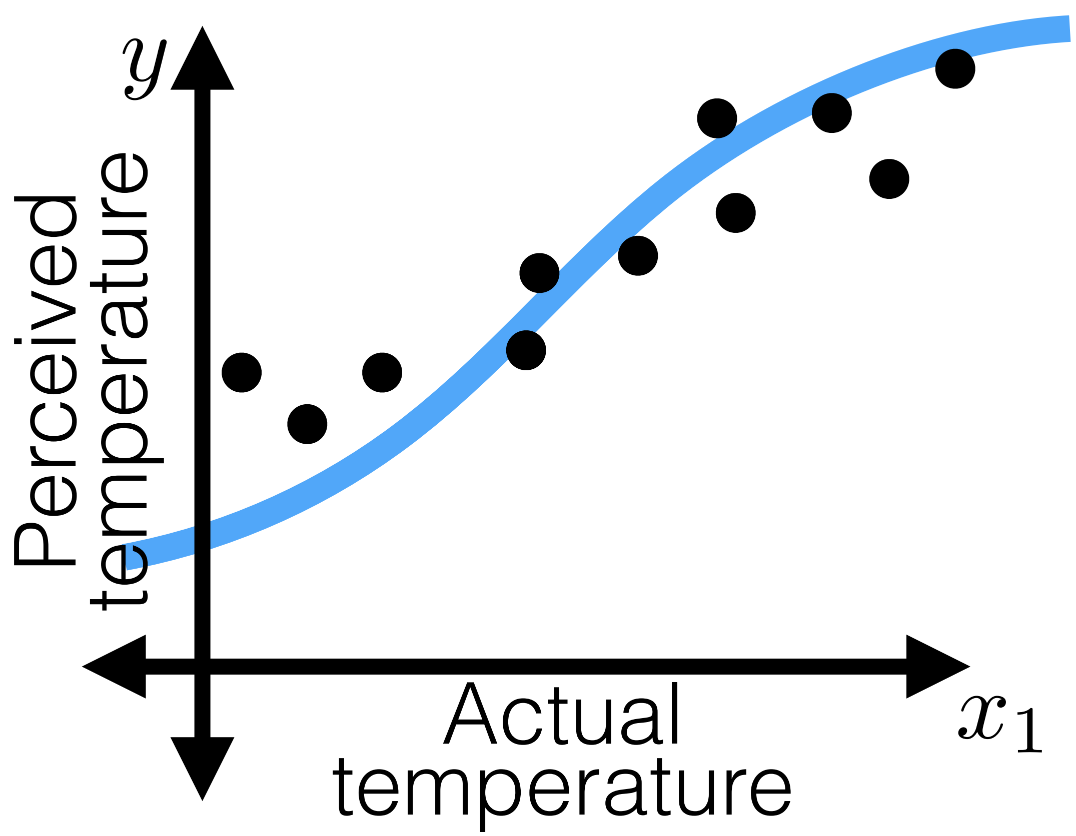
Regression

- Datum i: feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \mathbb{R}$

- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$



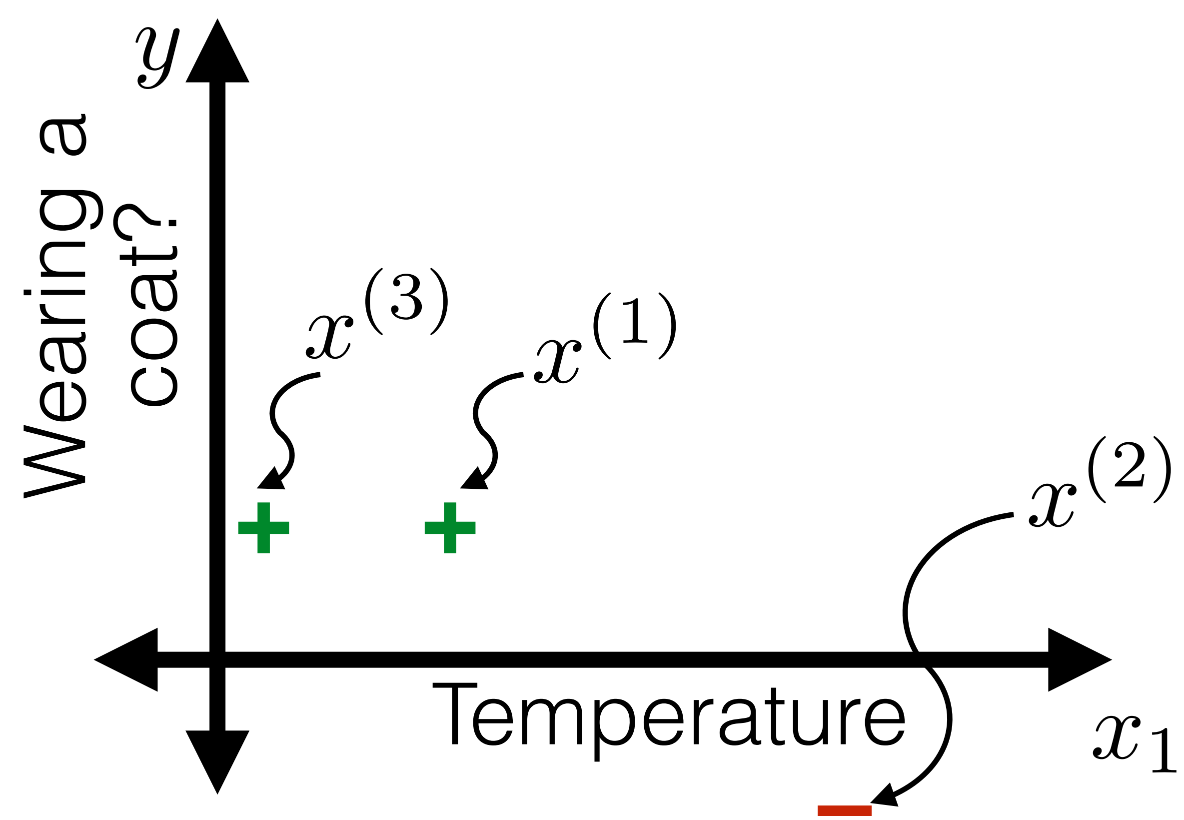
Compare

(Two-class) Classification

- Datum i : feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \{-1, +1\}$



Recall

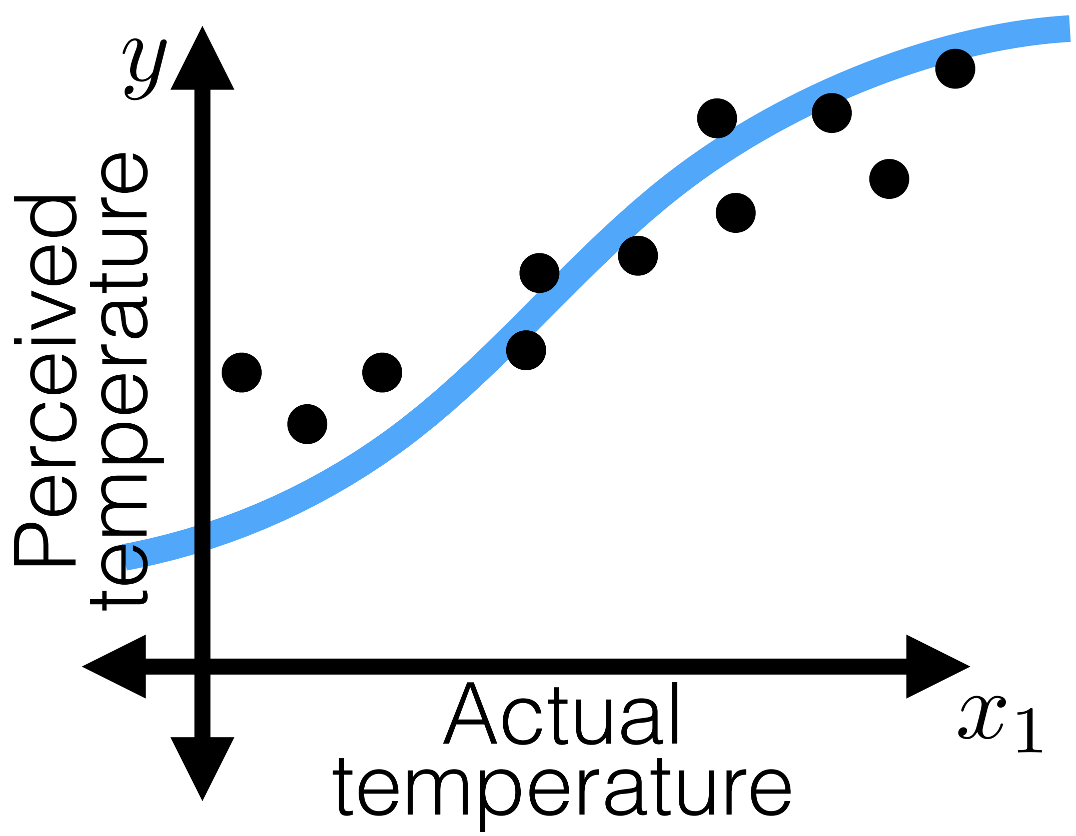
Regression

- Datum i: feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \mathbb{R}$

- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$



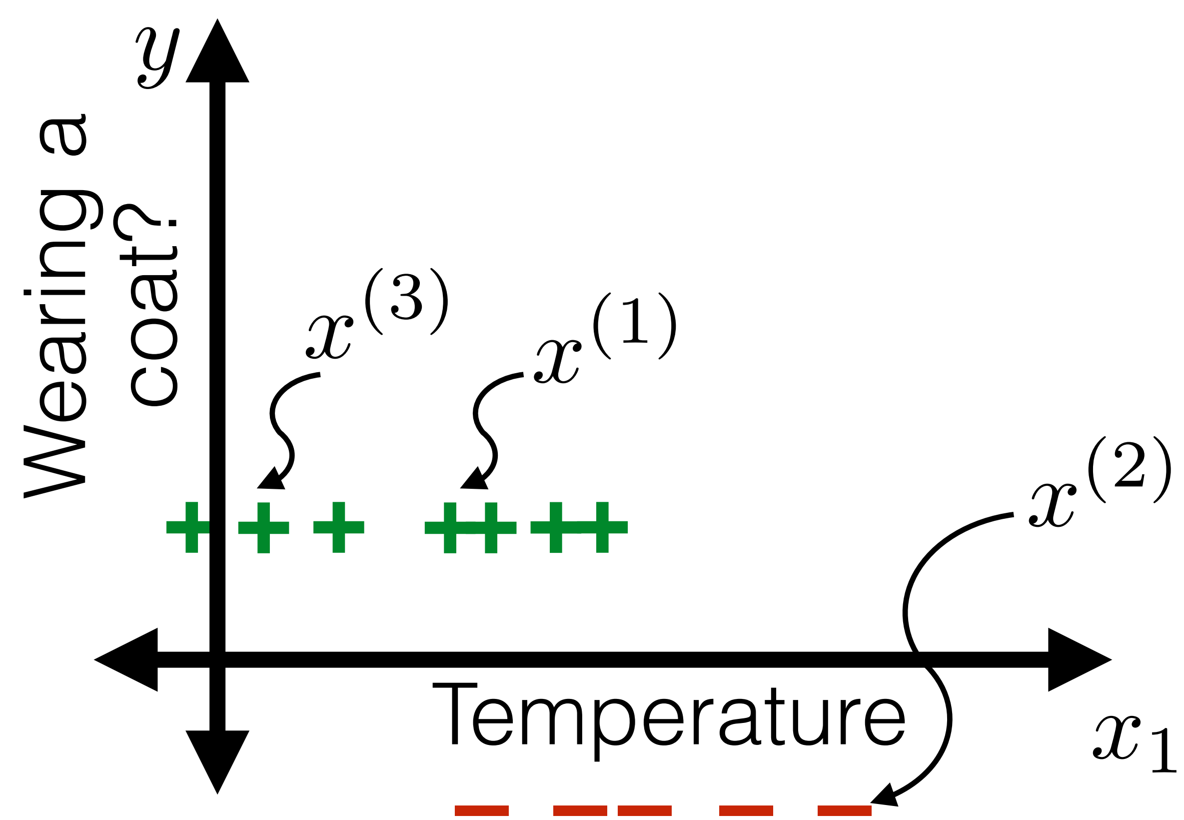
Compare

(Two-class) Classification

- Datum i : feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \{-1, +1\}$



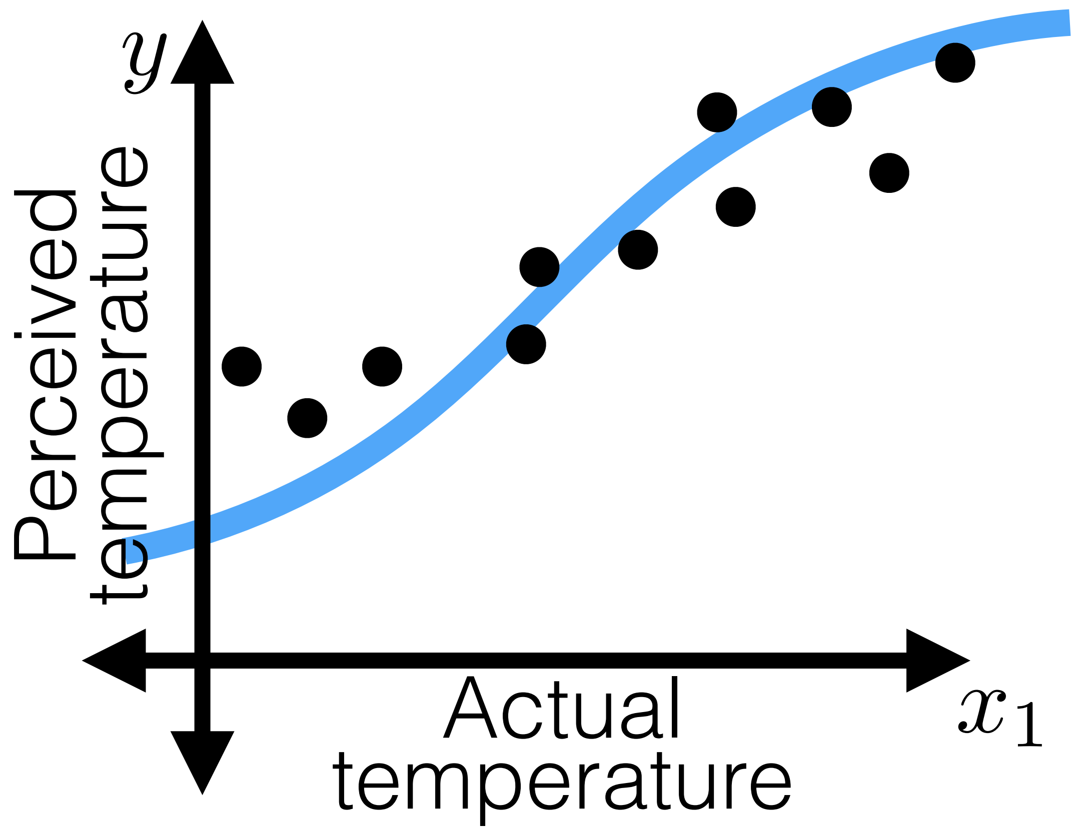
Recall

Regression

- Datum i: feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \mathbb{R}$
- Hypothesis $h : \mathbb{R}^d \rightarrow \mathbb{R}$



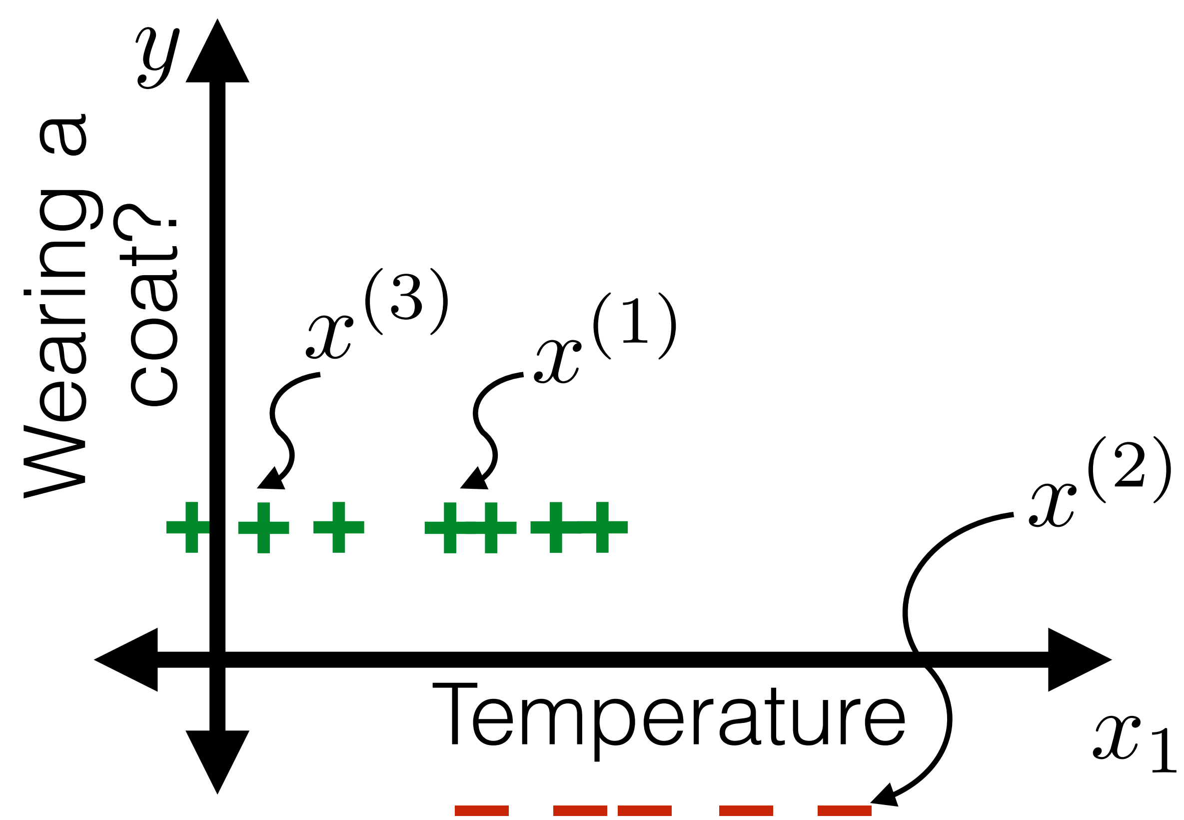
Compare

(Two-class) Classification

- Datum i : feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \{-1, +1\}$
- Hypothesis $h : \mathbb{R}^d \rightarrow \{-1, +1\}$



Recall

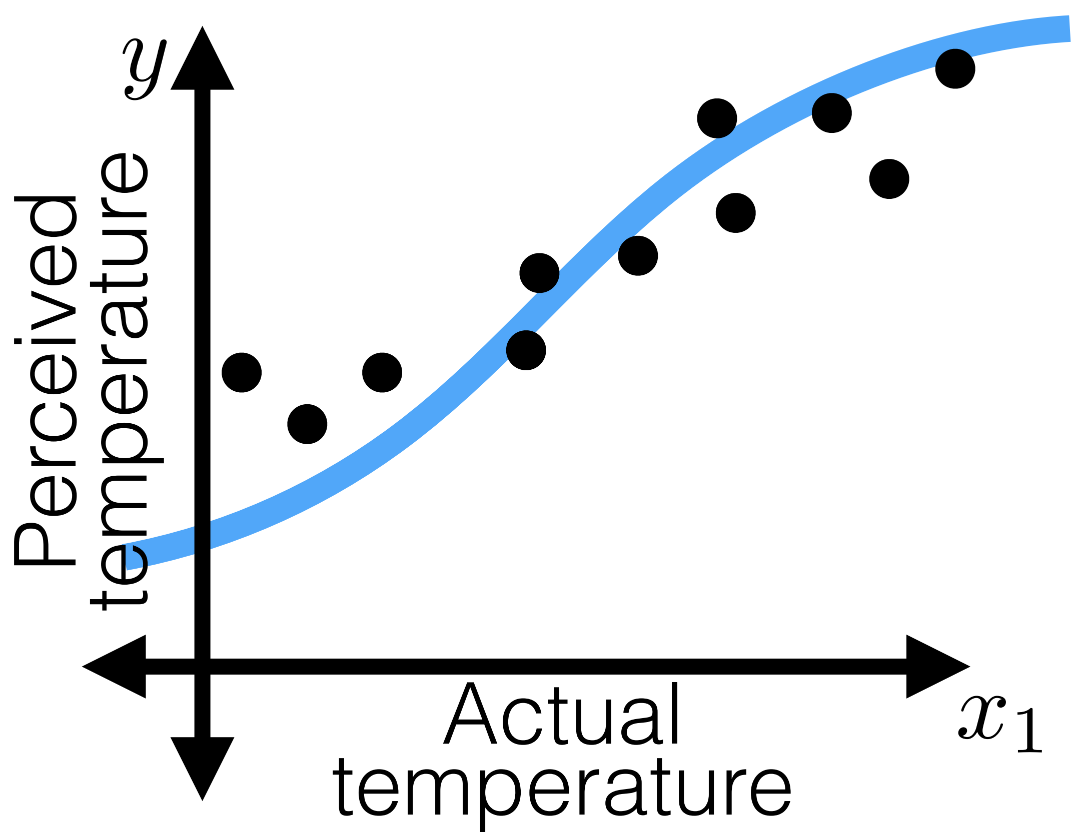
Regression

- Datum i: feature vector

$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \mathbb{R}$

- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$



Compare

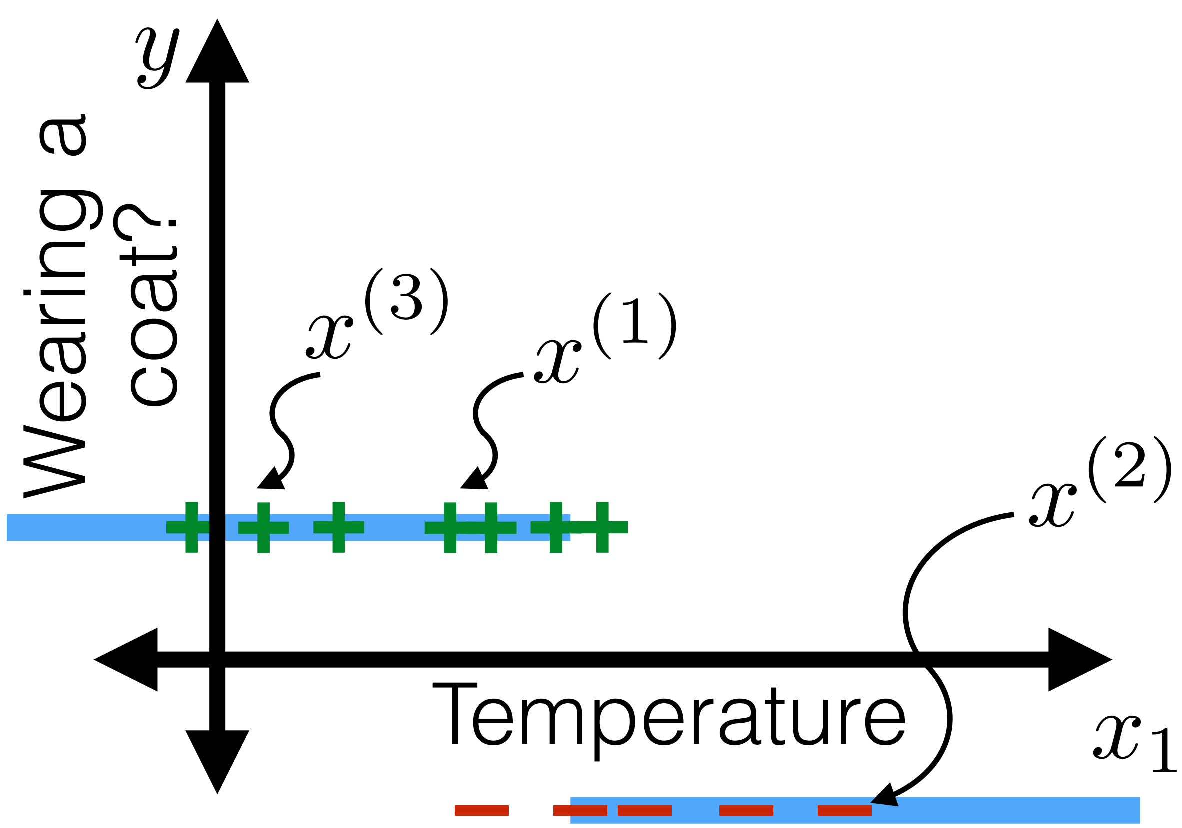
(Two-class) Classification

- Datum i : feature vector

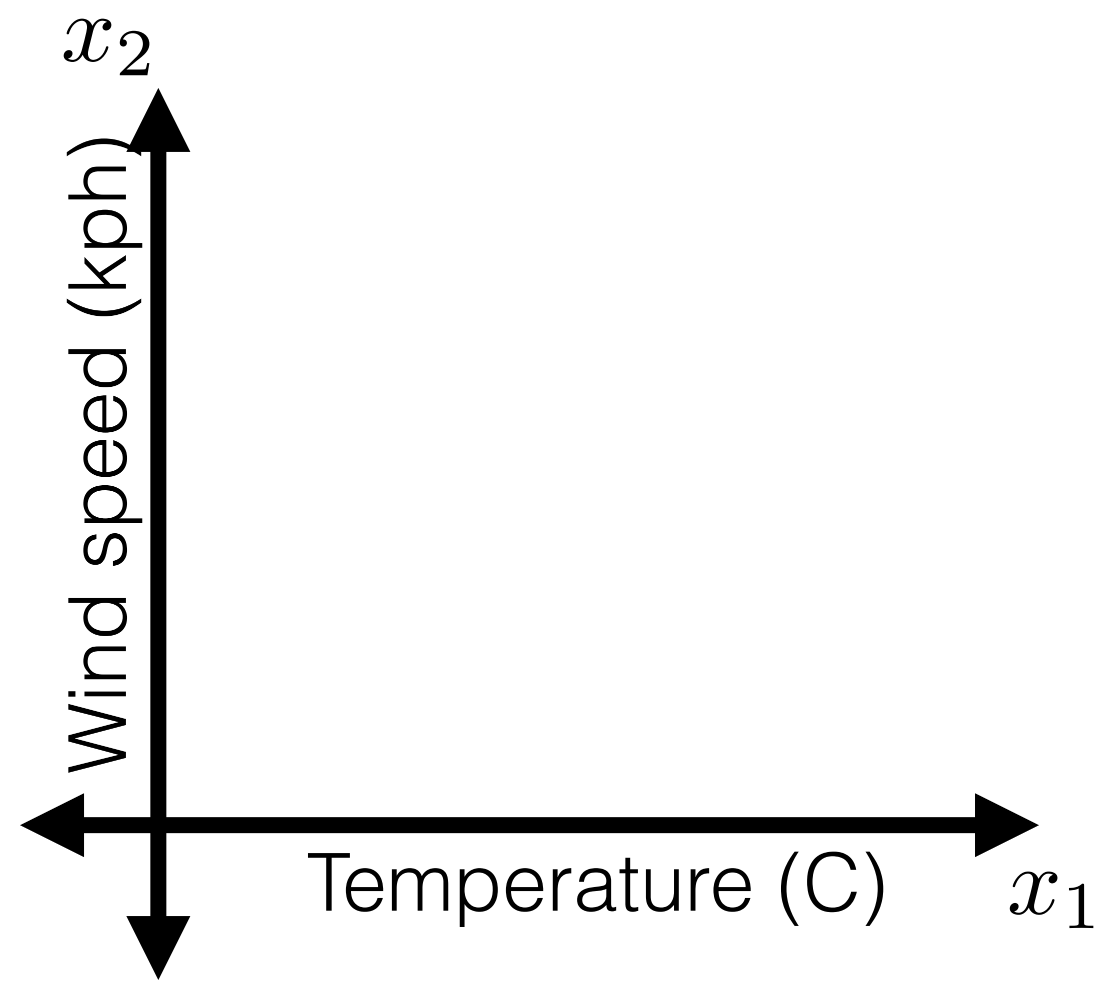
$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$

  - Label $y^{(i)} \in \{-1, +1\}$

- Hypothesis $h : \mathbb{R}^d \to \{-1, +1\}$

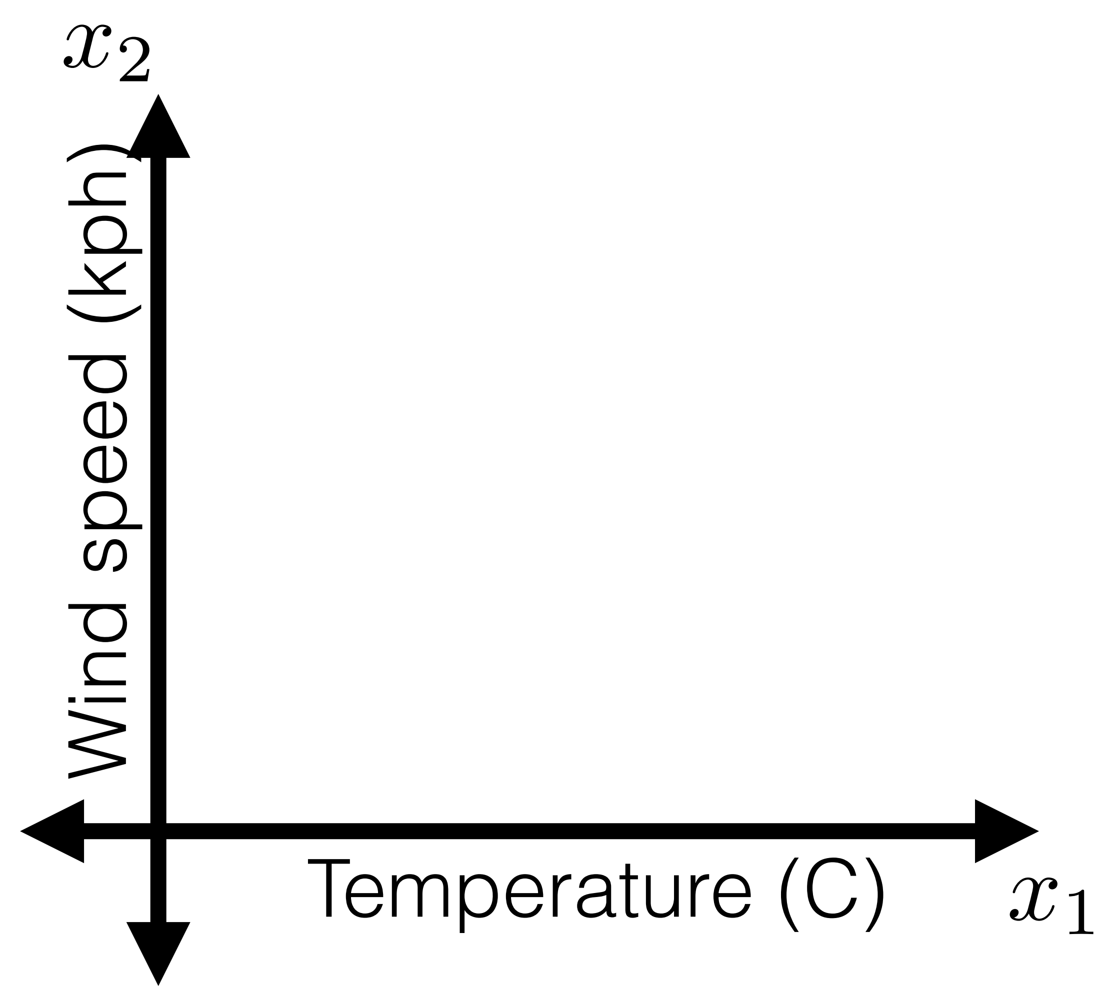


Linear classifiers

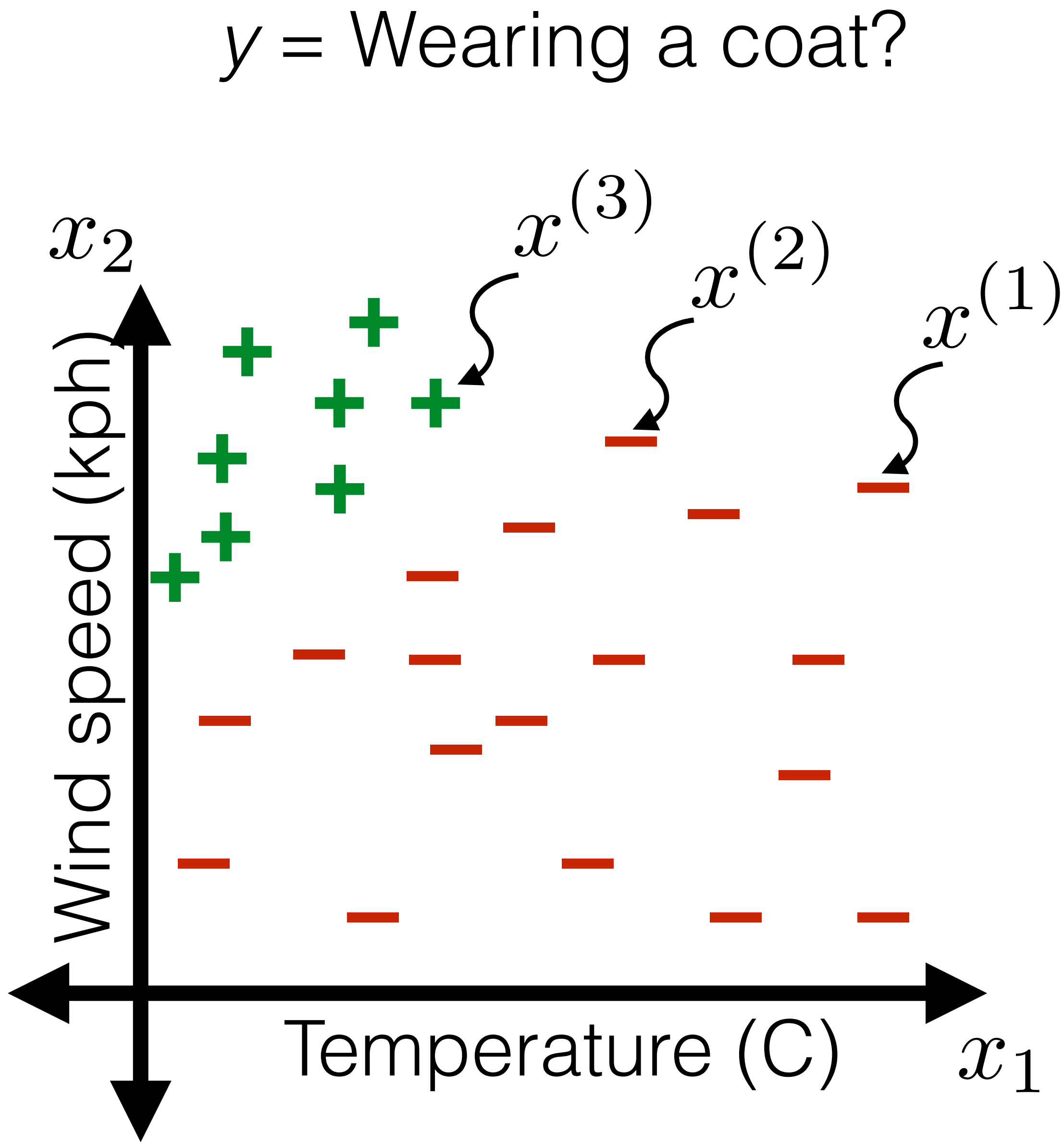


Linear classifiers

y = Wearing a coat?



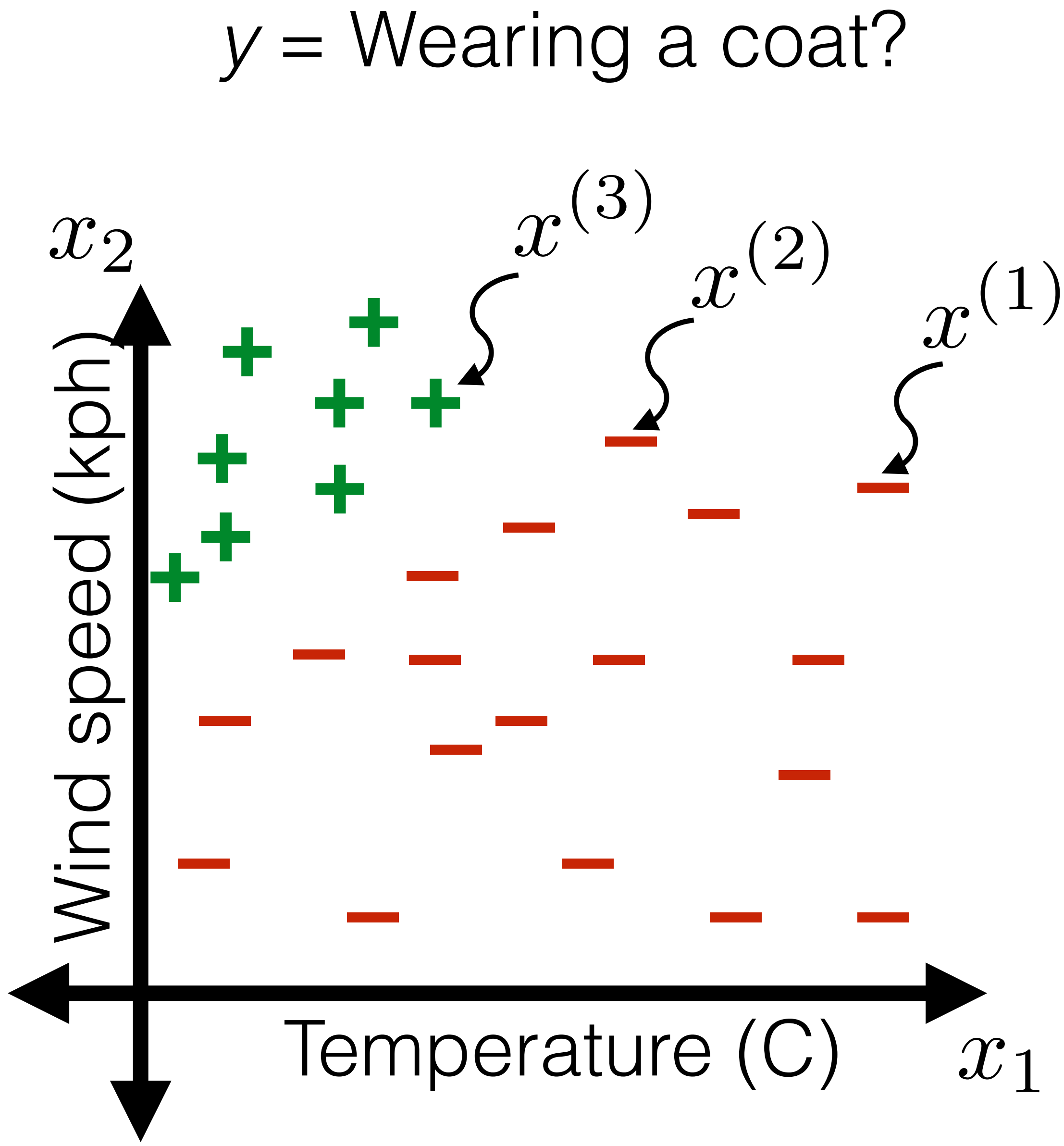
Linear classifiers



Linear classifiers

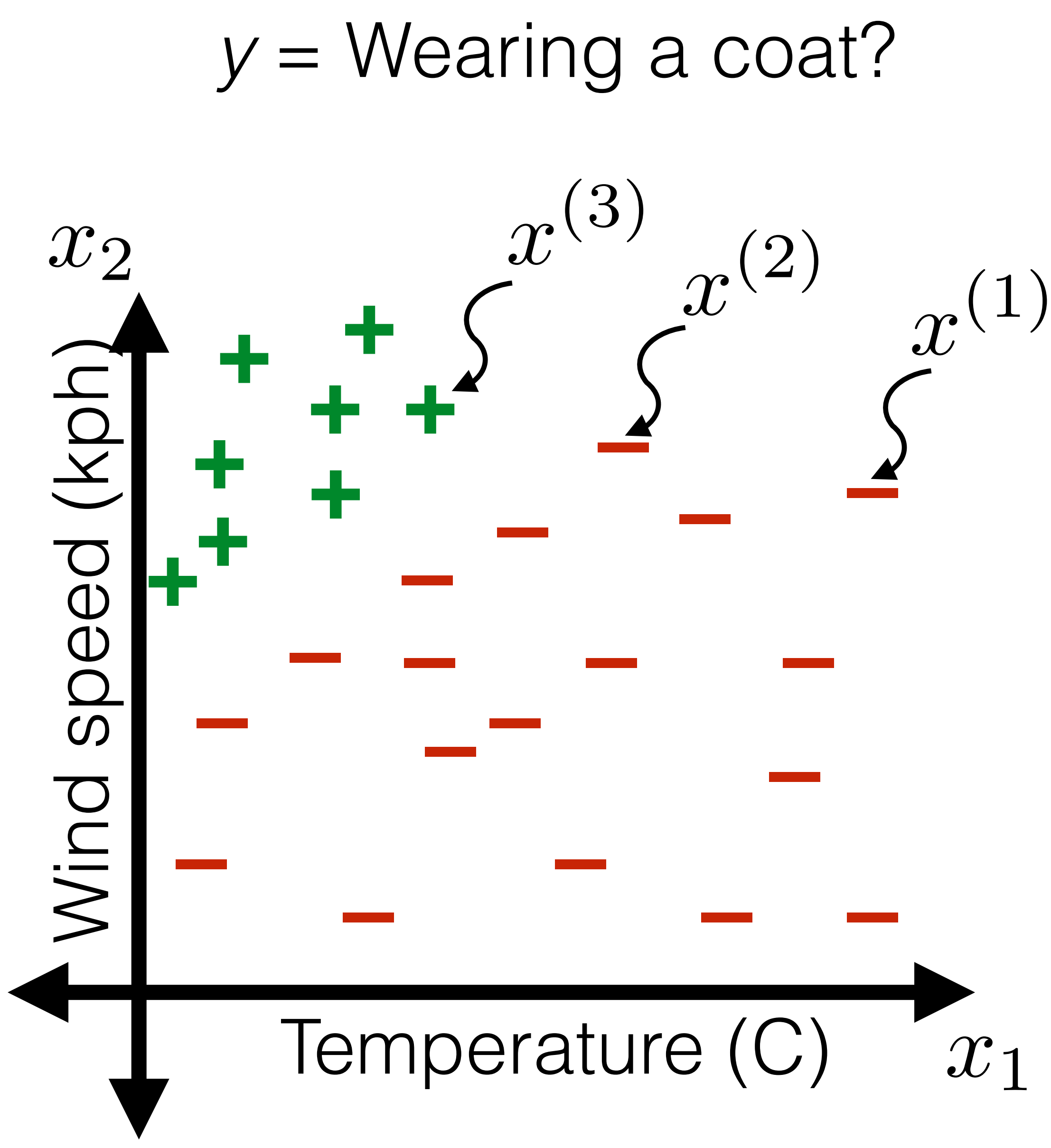
- Classification hypothesis:

$$h : \mathbb{R}^d \to \{-1, +1\}$$



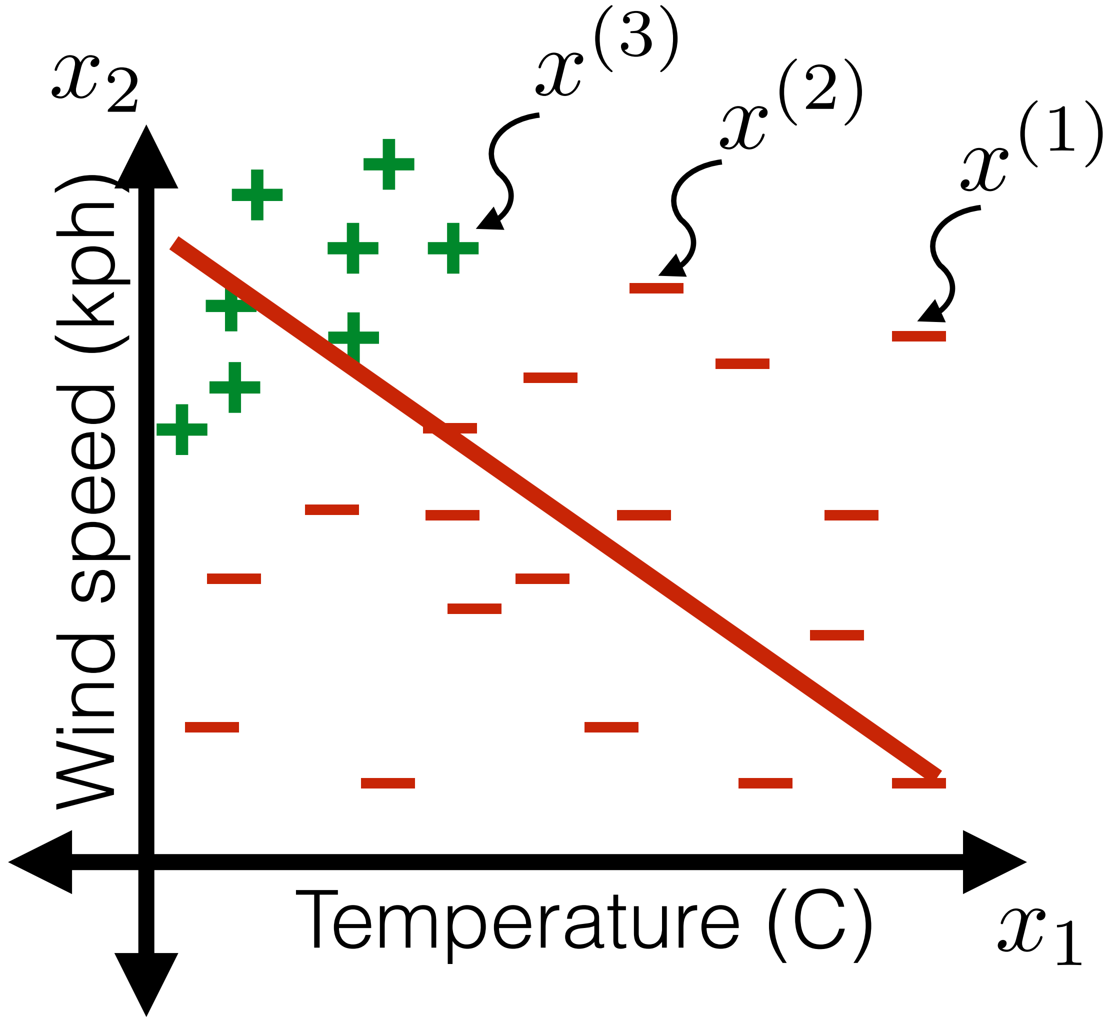
Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



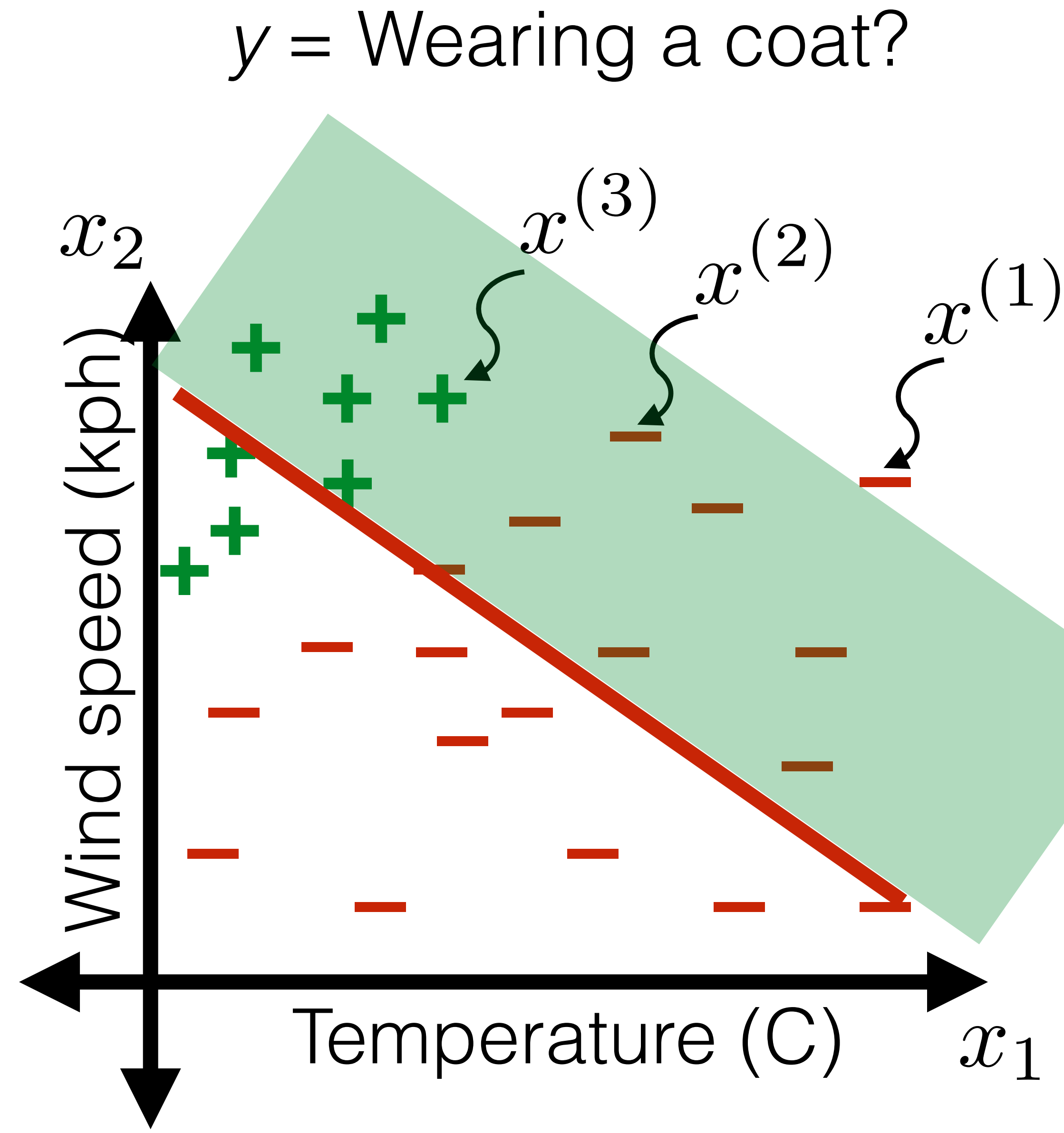
Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



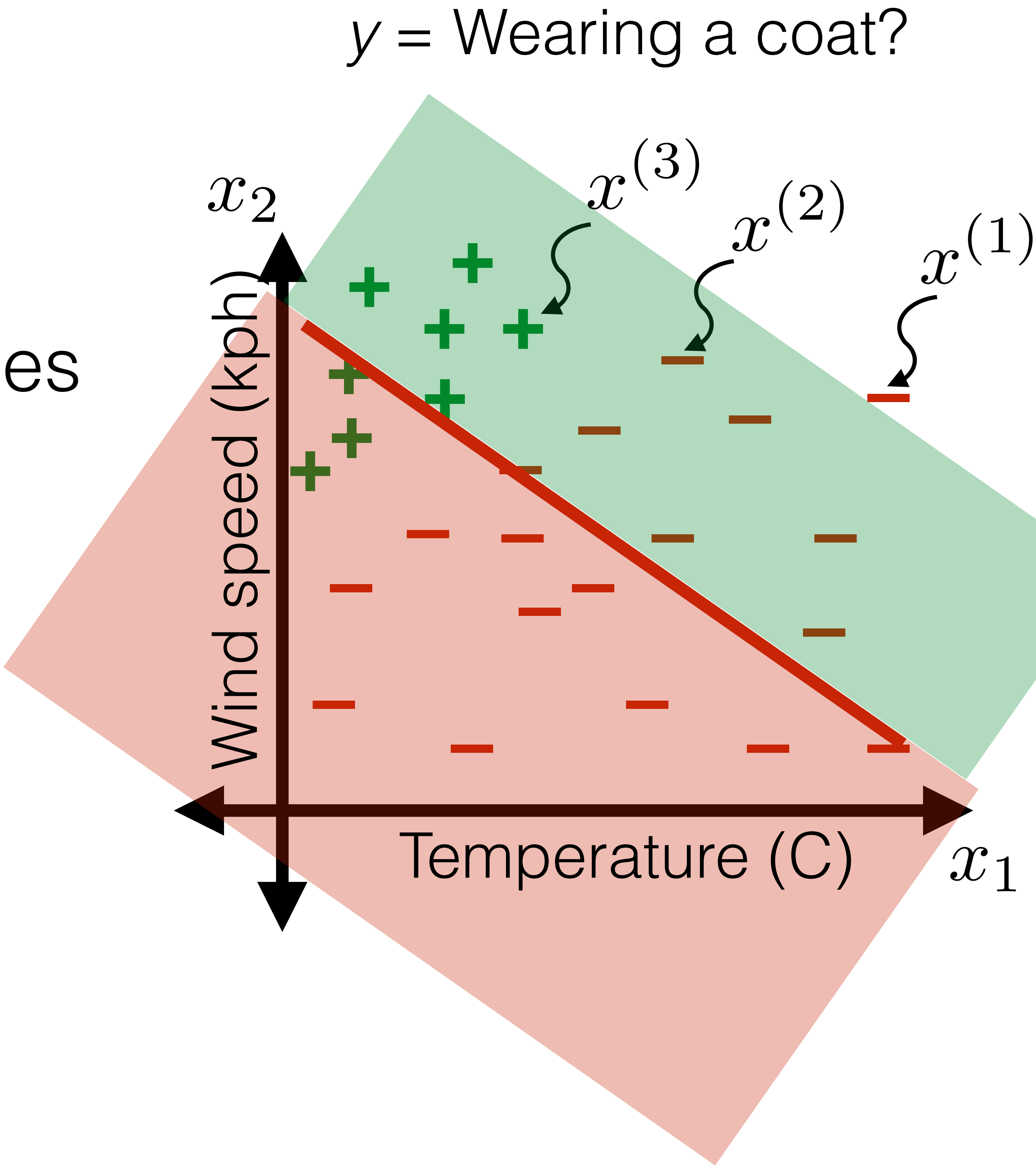
Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



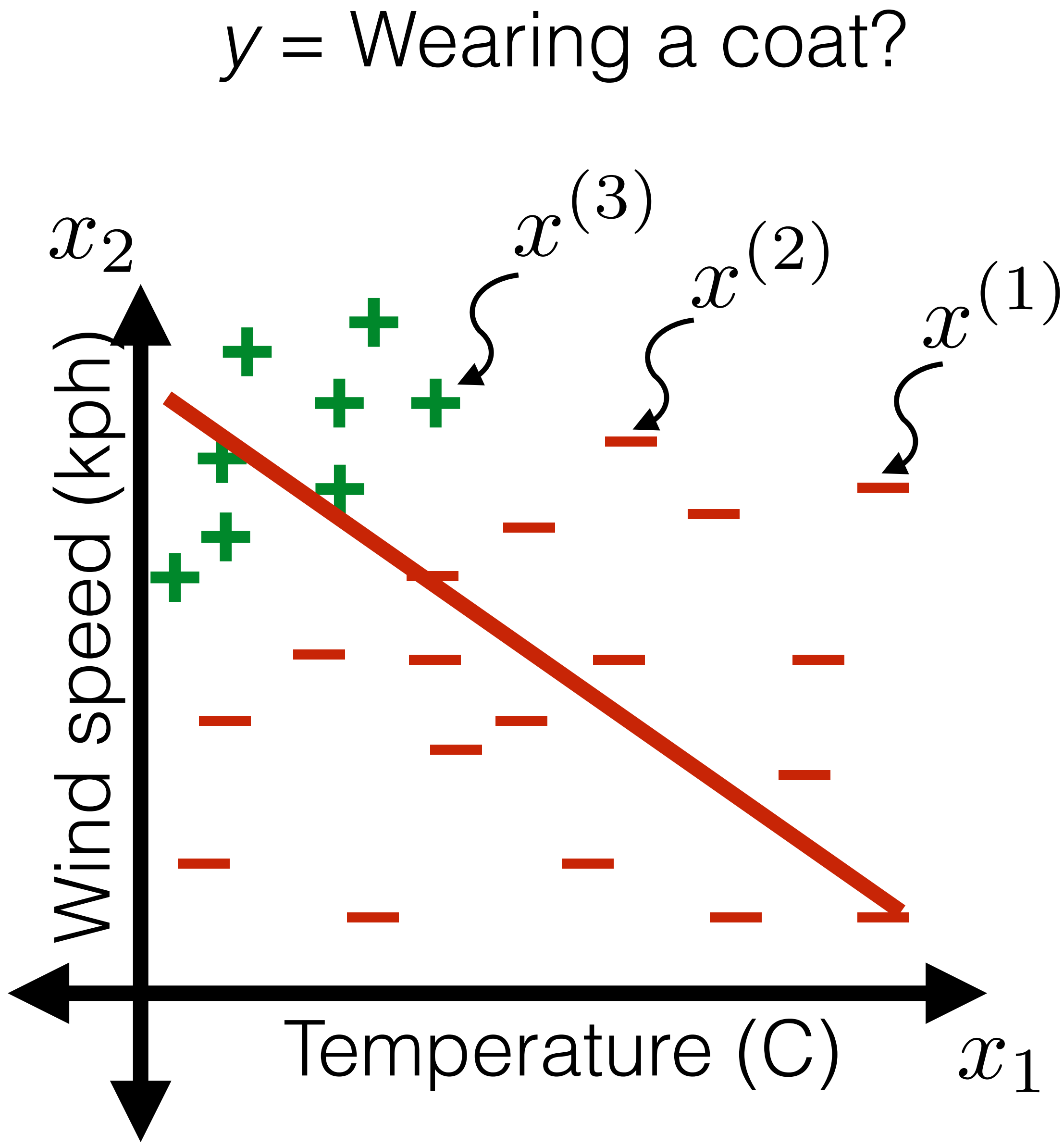
Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



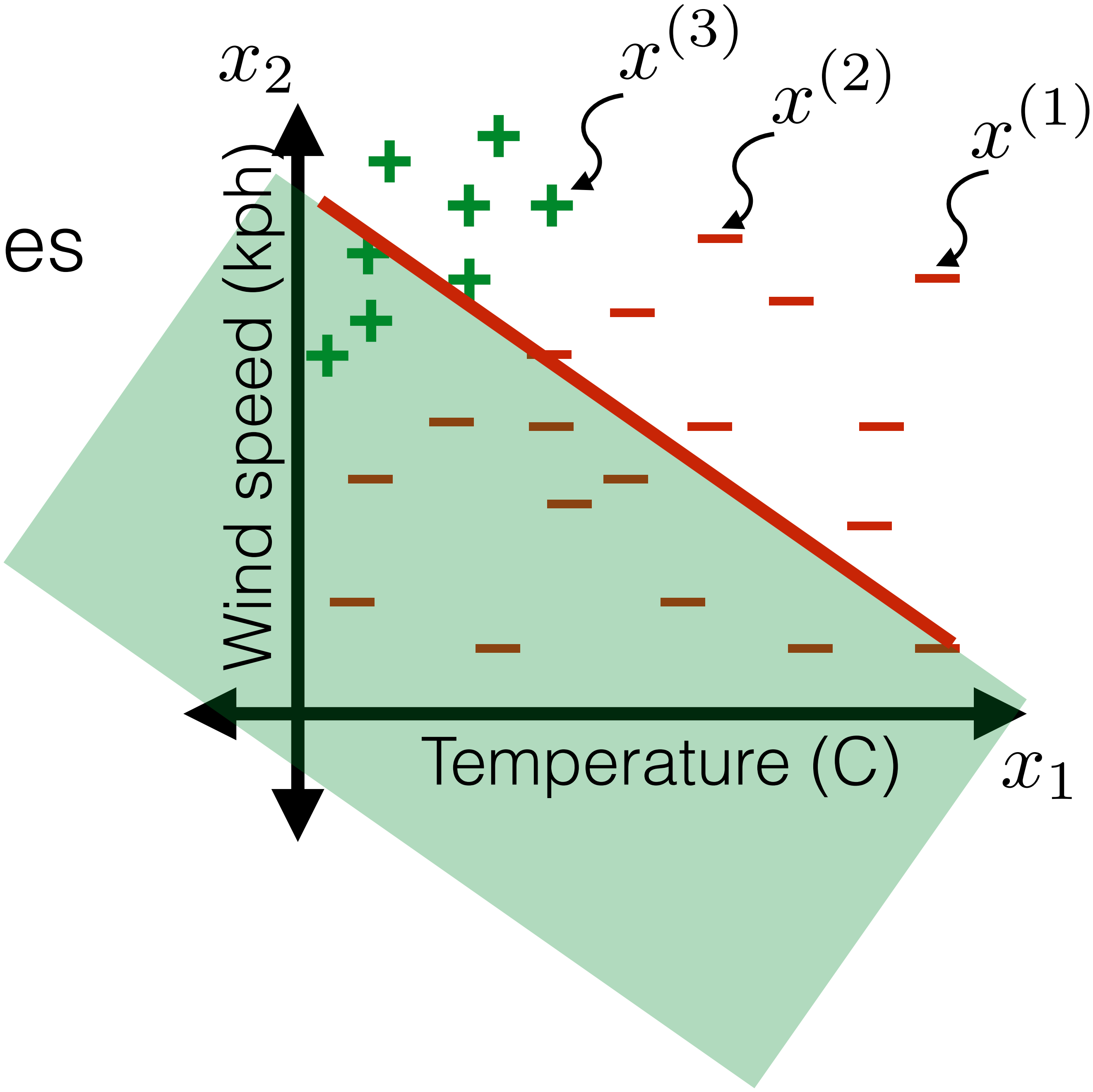
Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



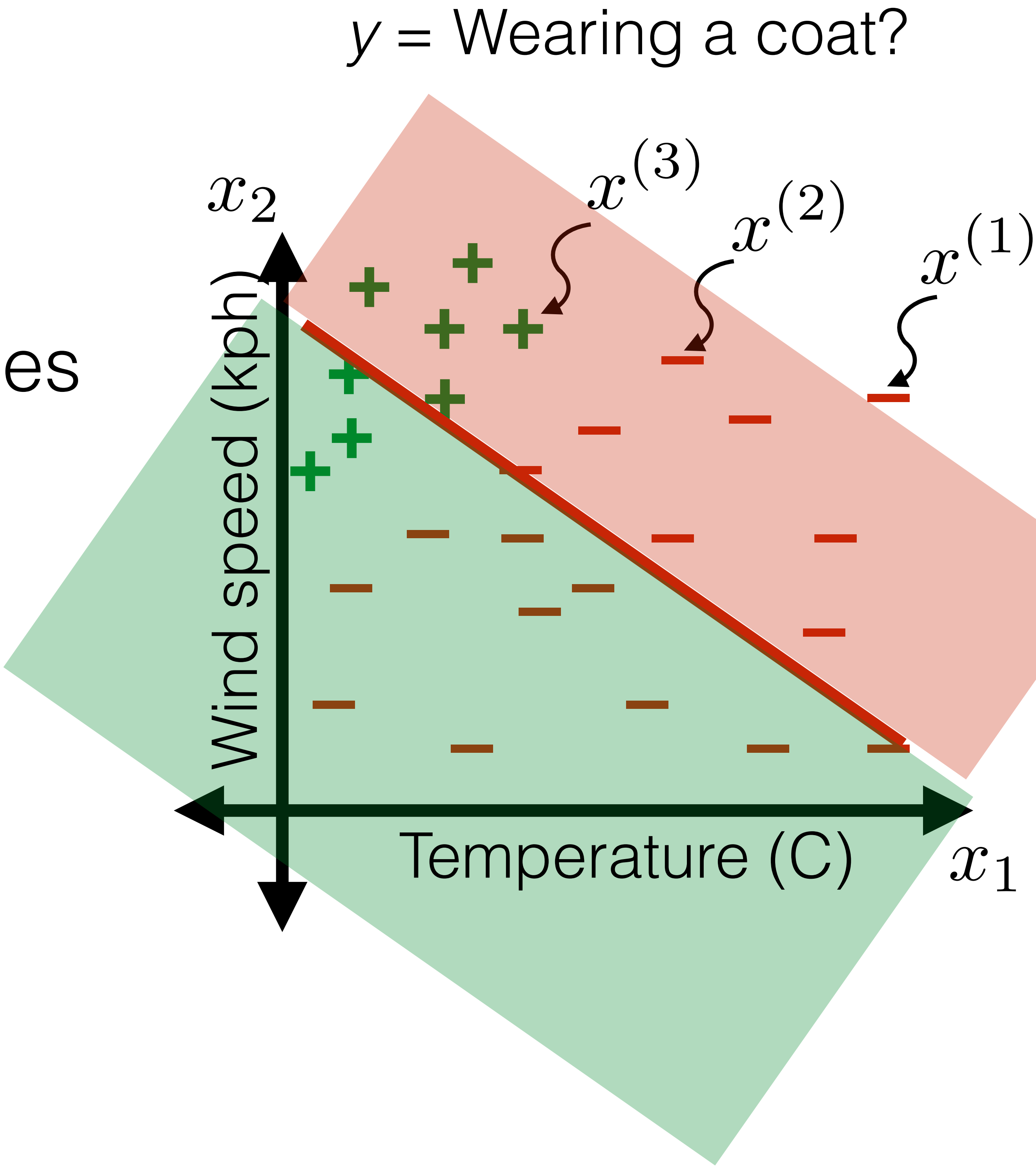
Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



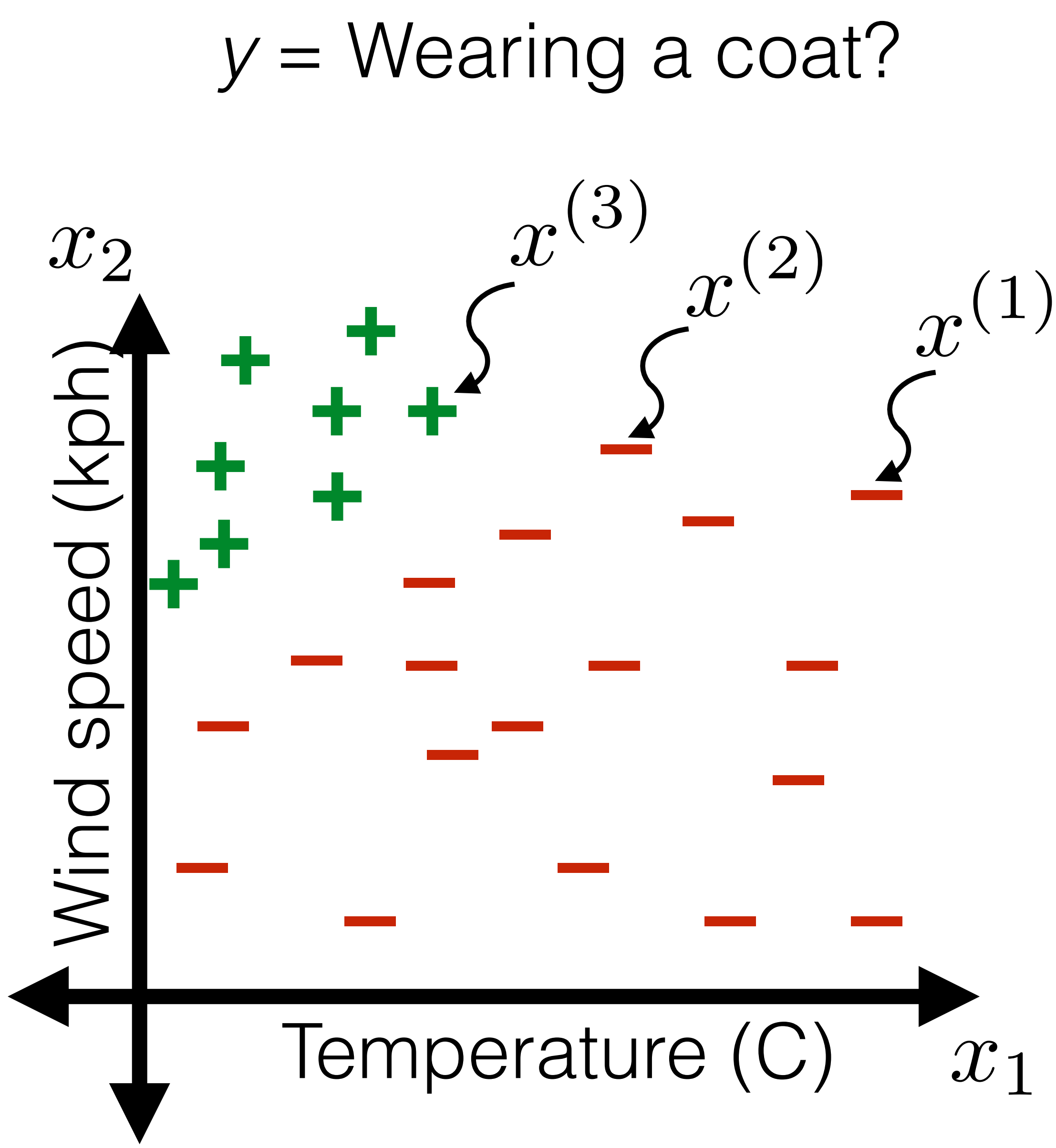
Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

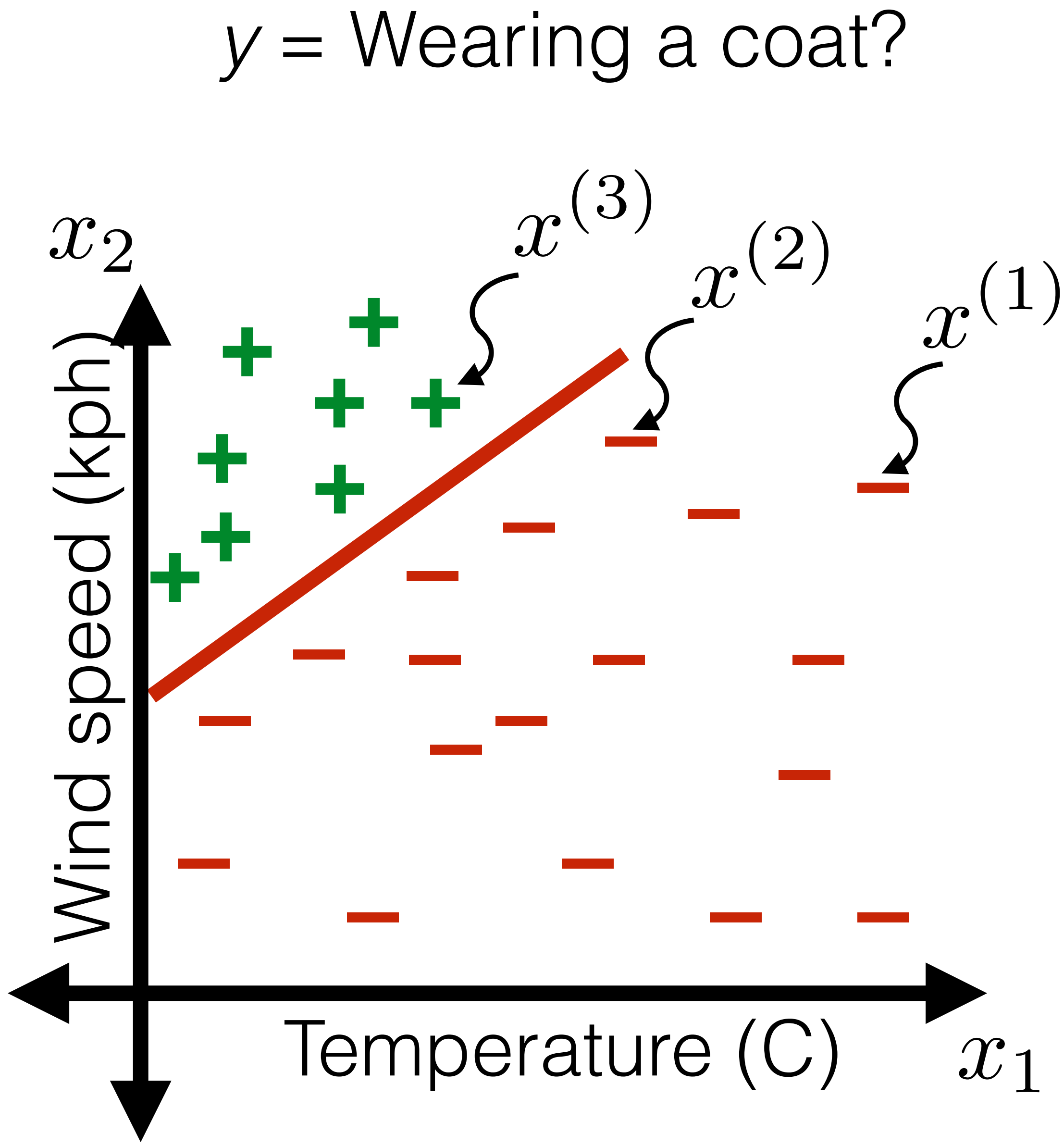


Linear classifiers

- Classification hypothesis:

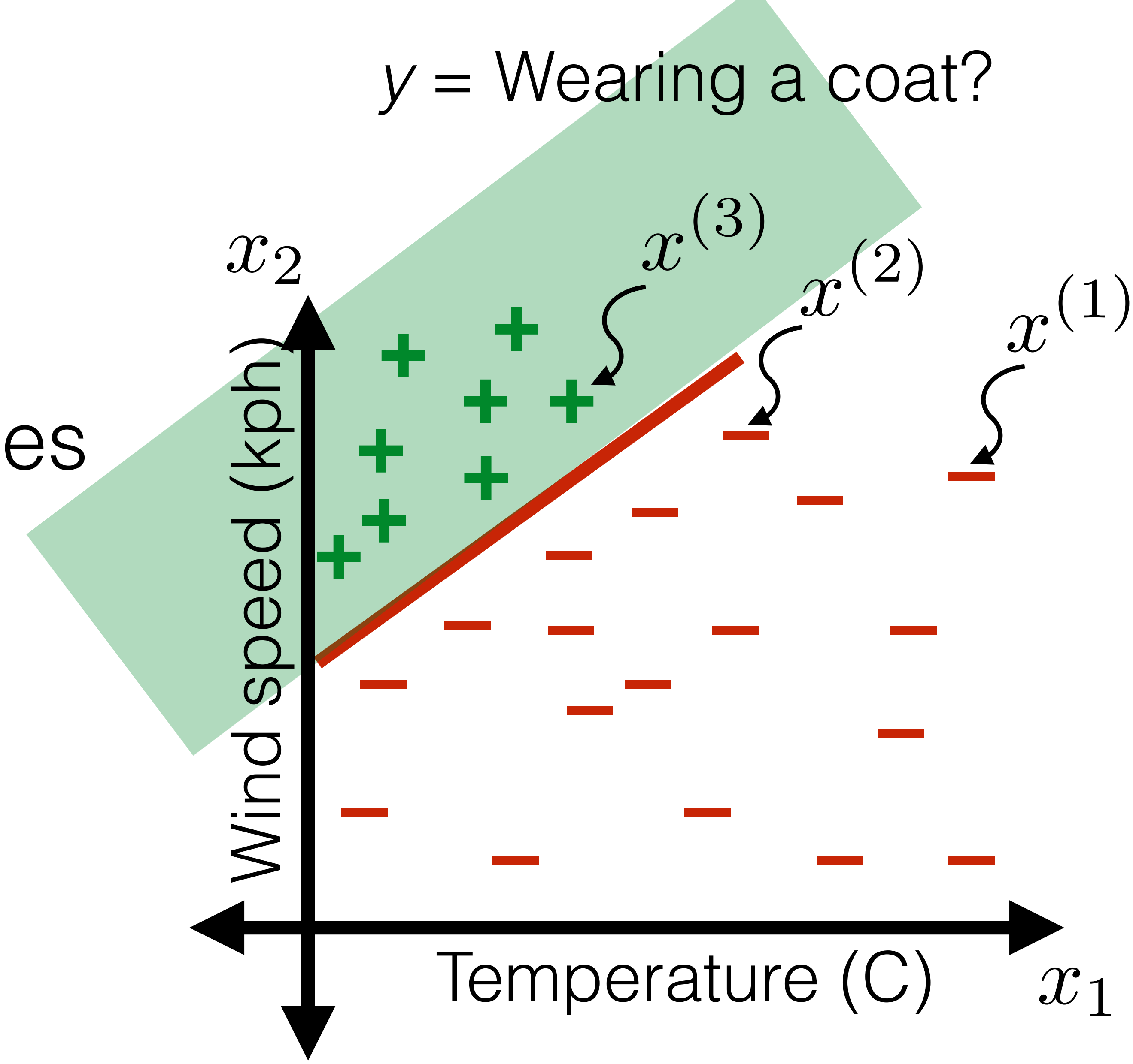
$$h : \mathbb{R}^d \to \{-1, +1\}$$

- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



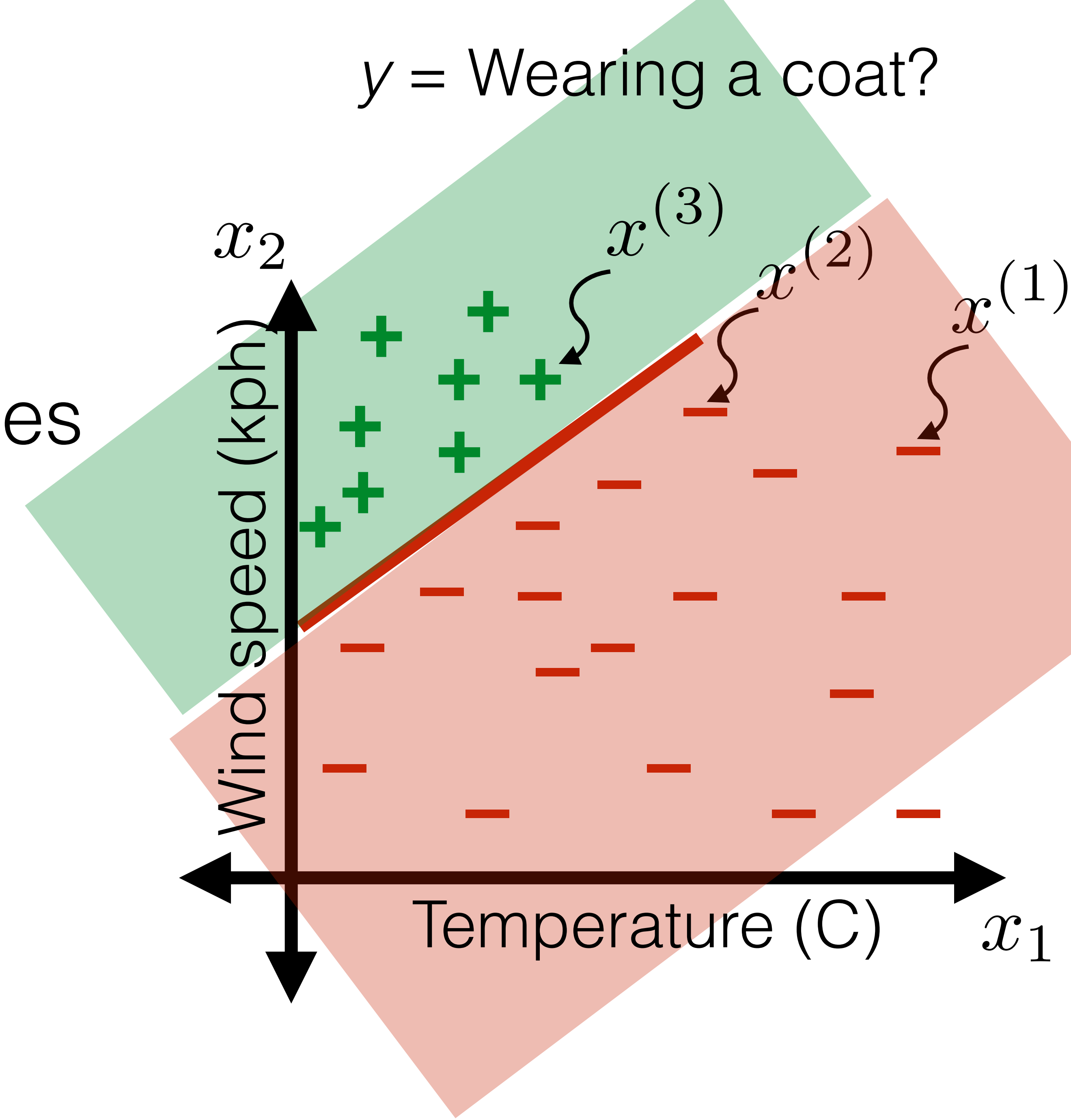
Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



Linear classifiers

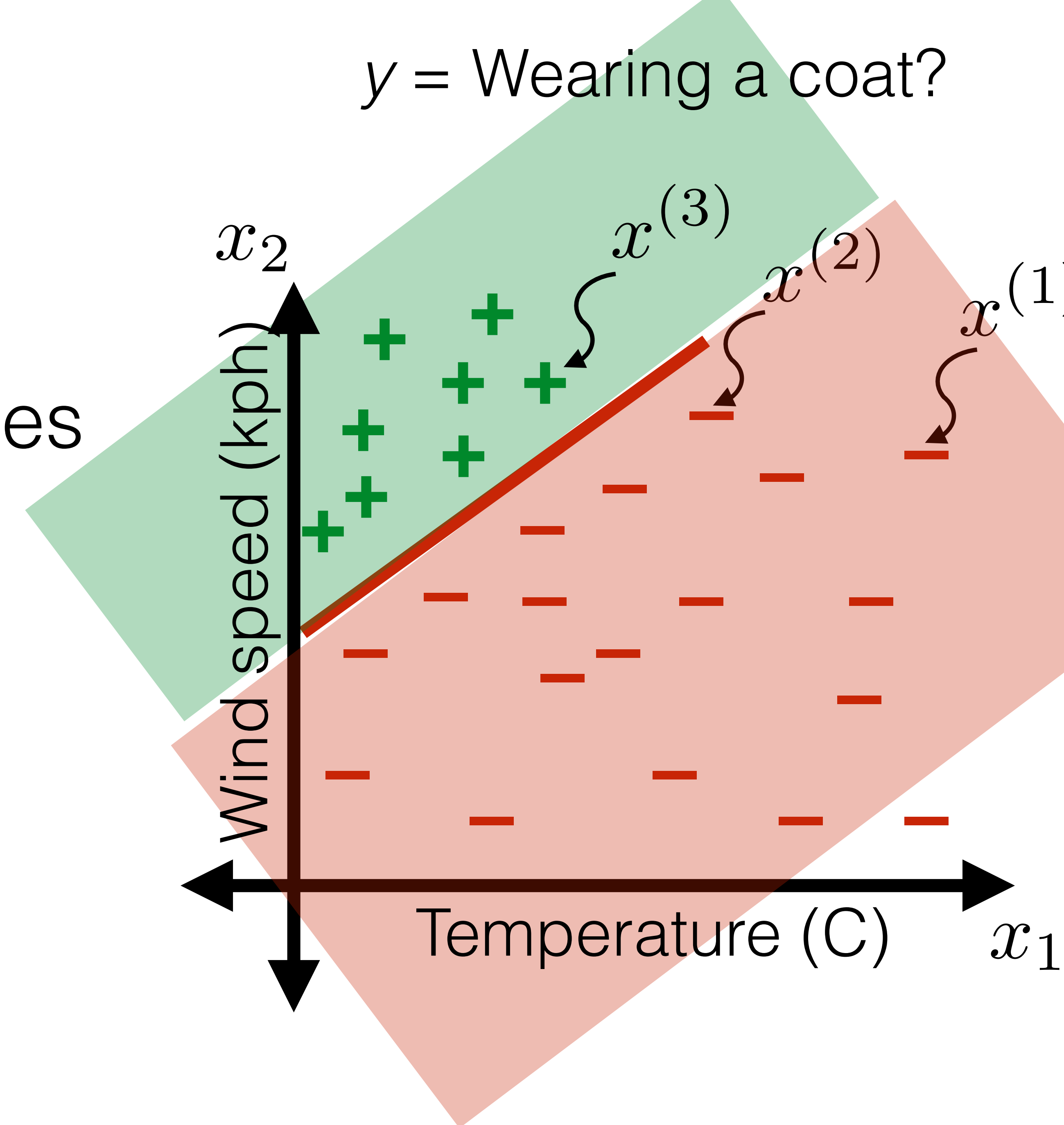
- Classification hypothesis:
$$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

Math facts!



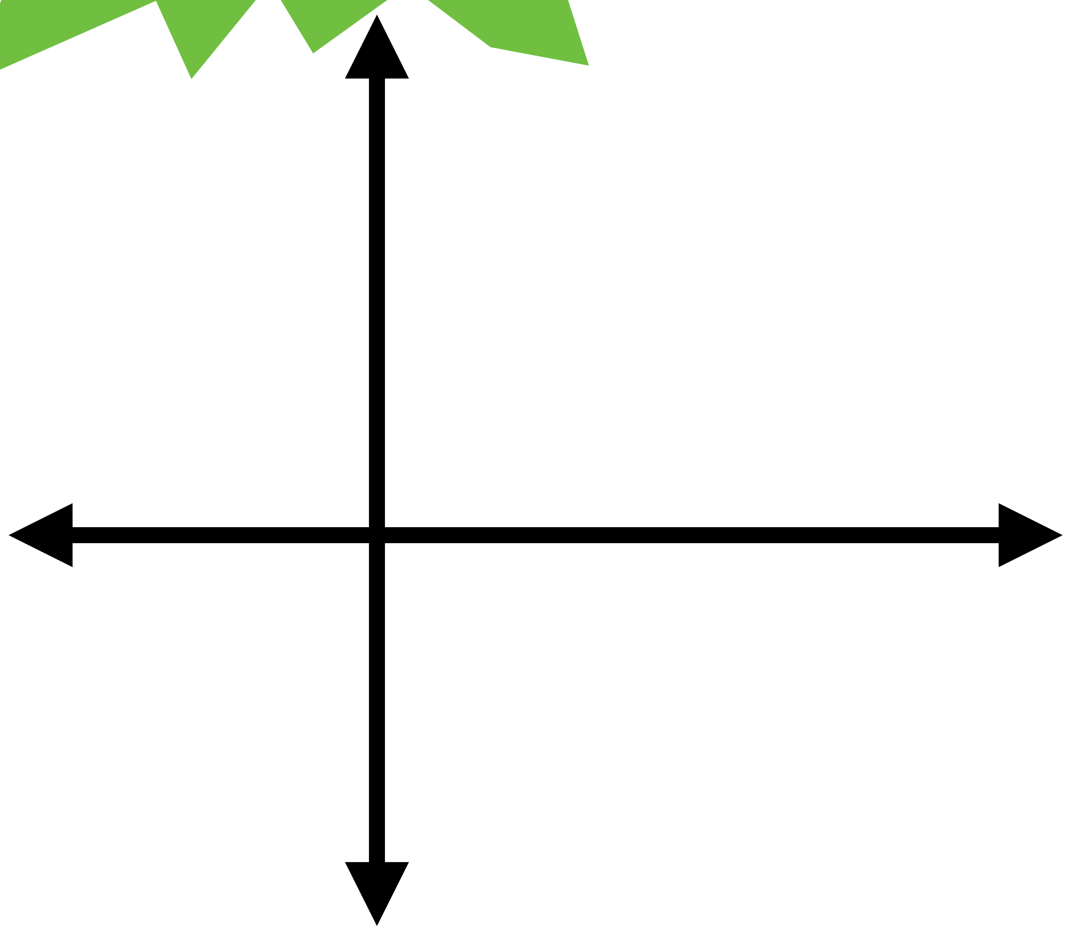
Linear classifiers

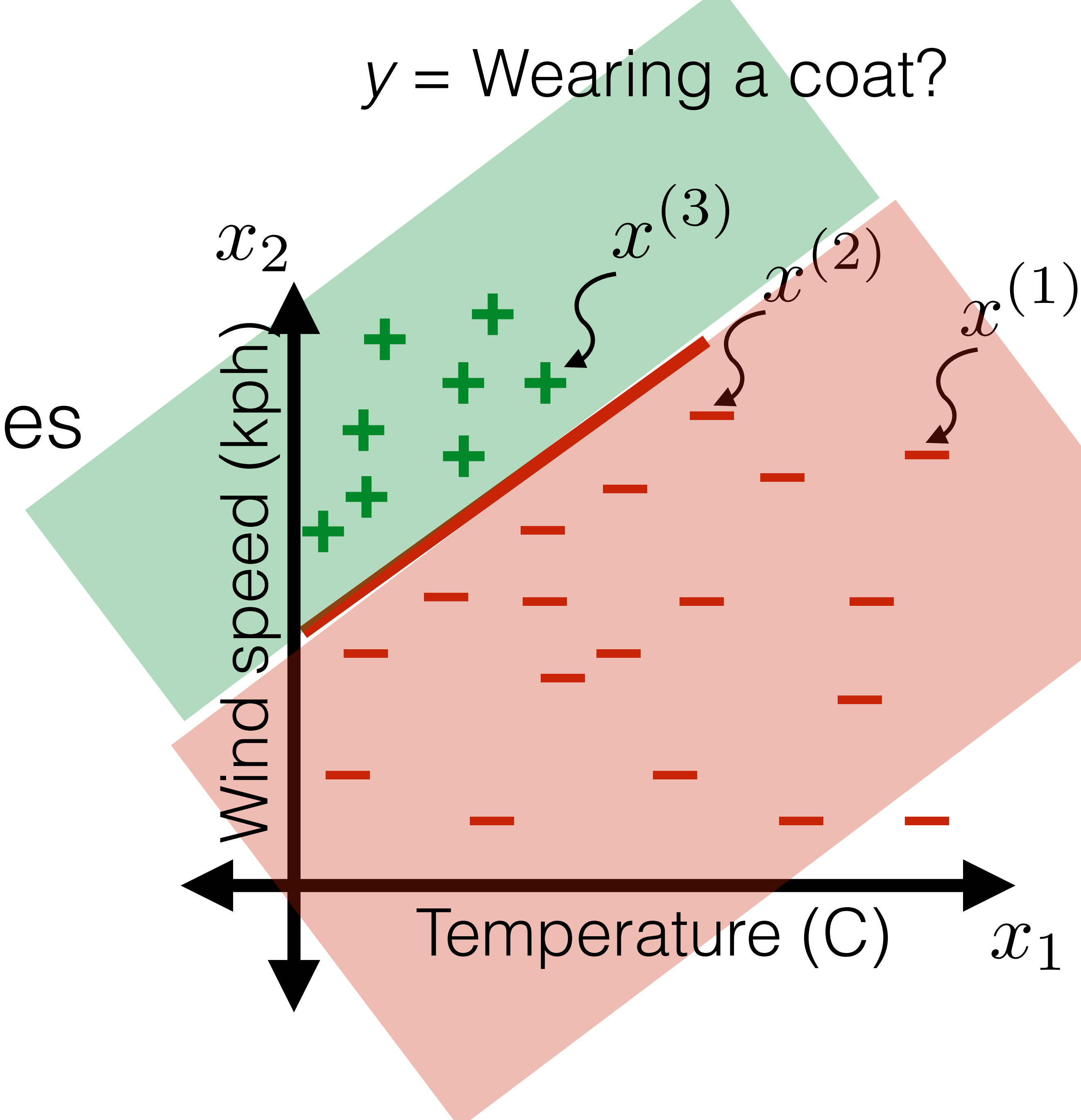
- Classification hypothesis:

$$h : \mathbb{R}^d \to \{-1, +1\}$$

- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

Math facts!



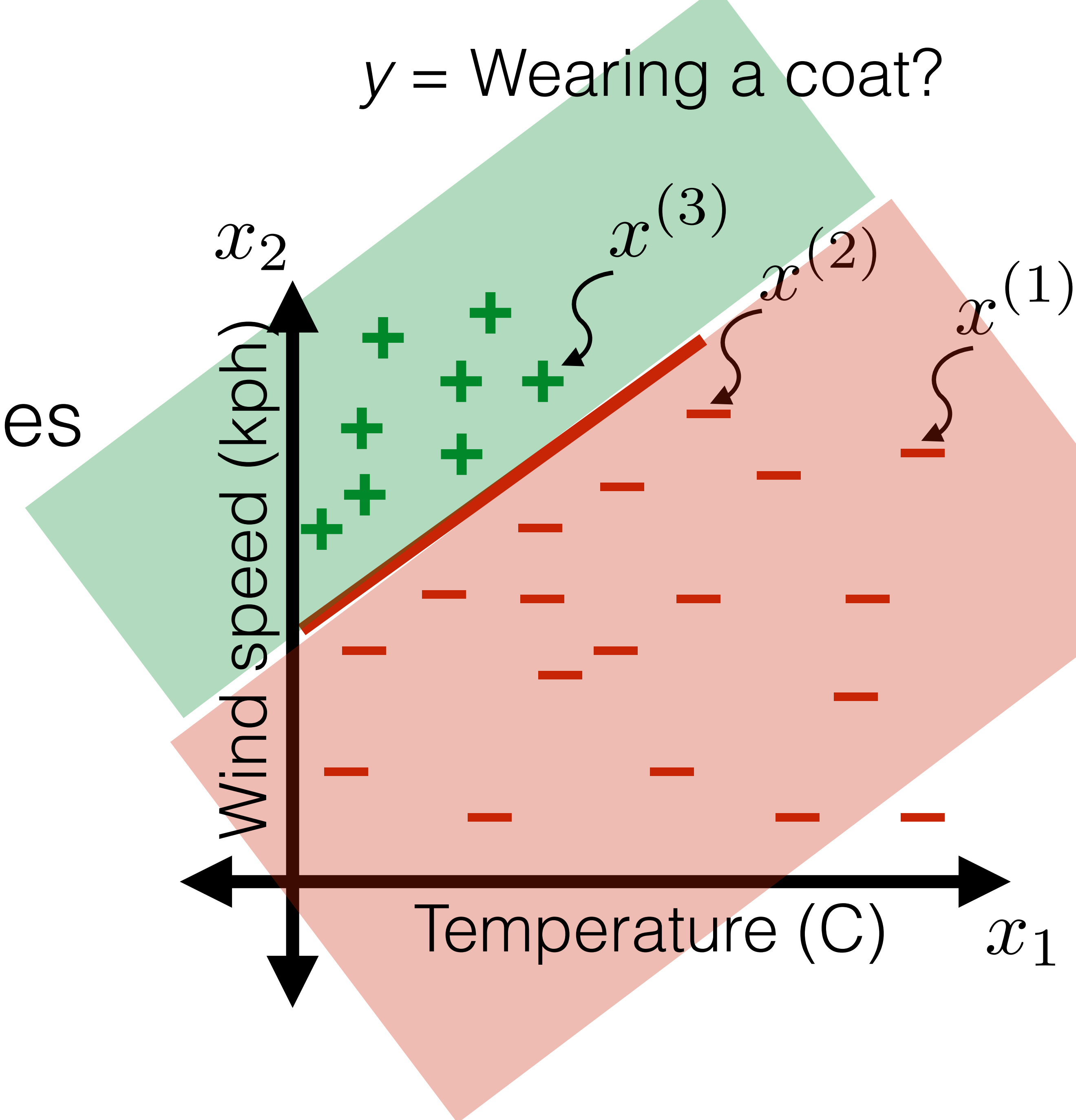


Linear classifiers

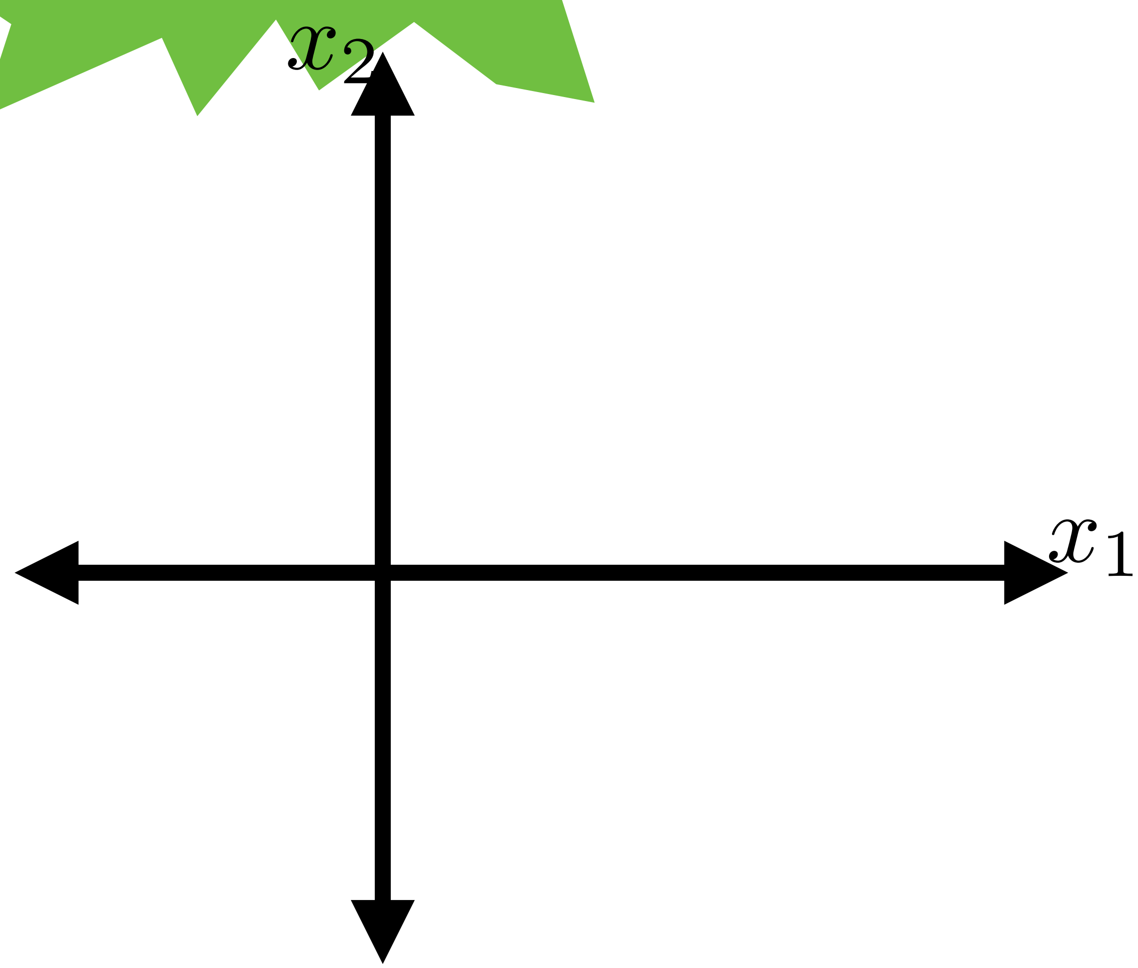
- Classification hypothesis:

$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$

- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

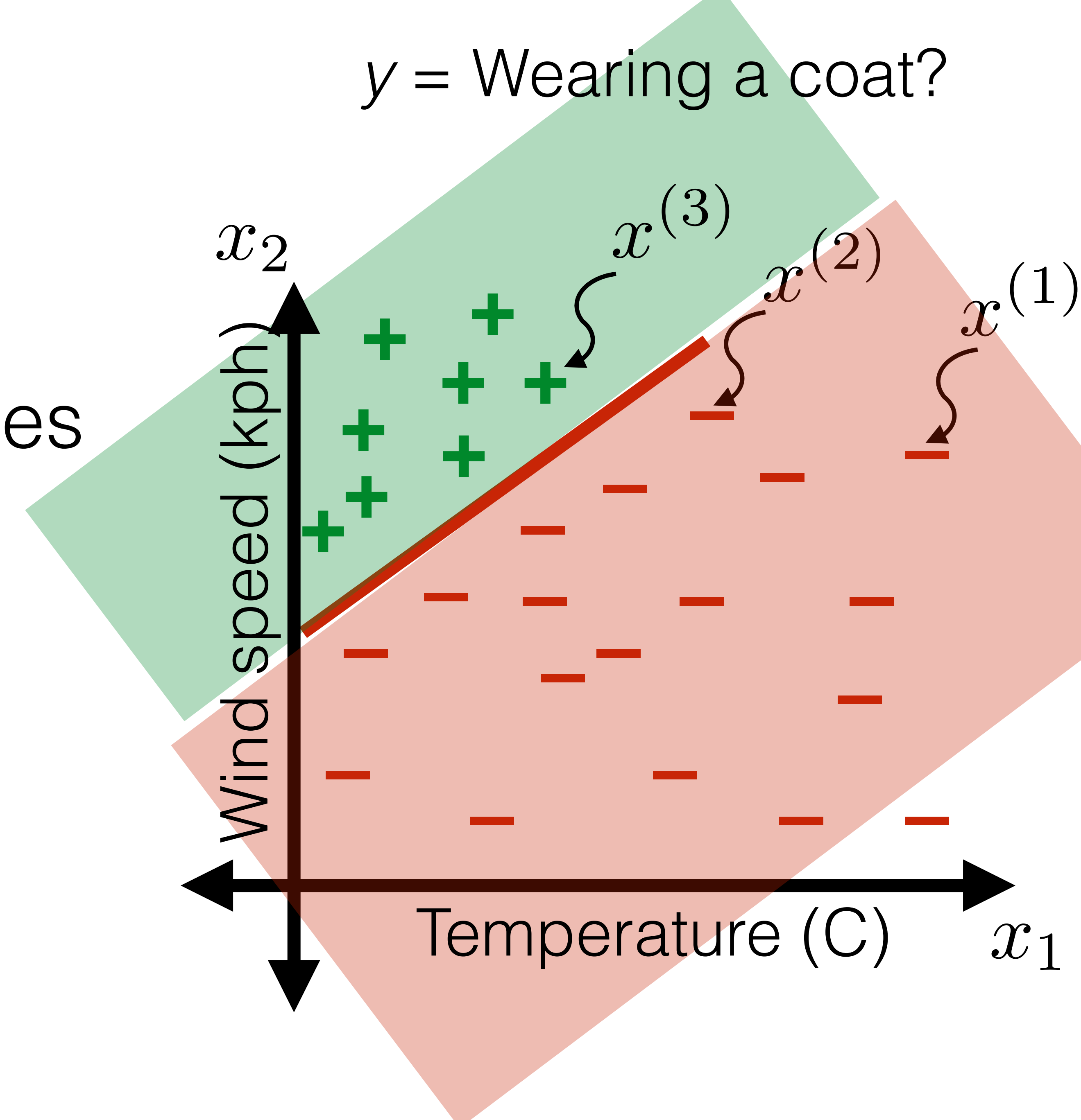


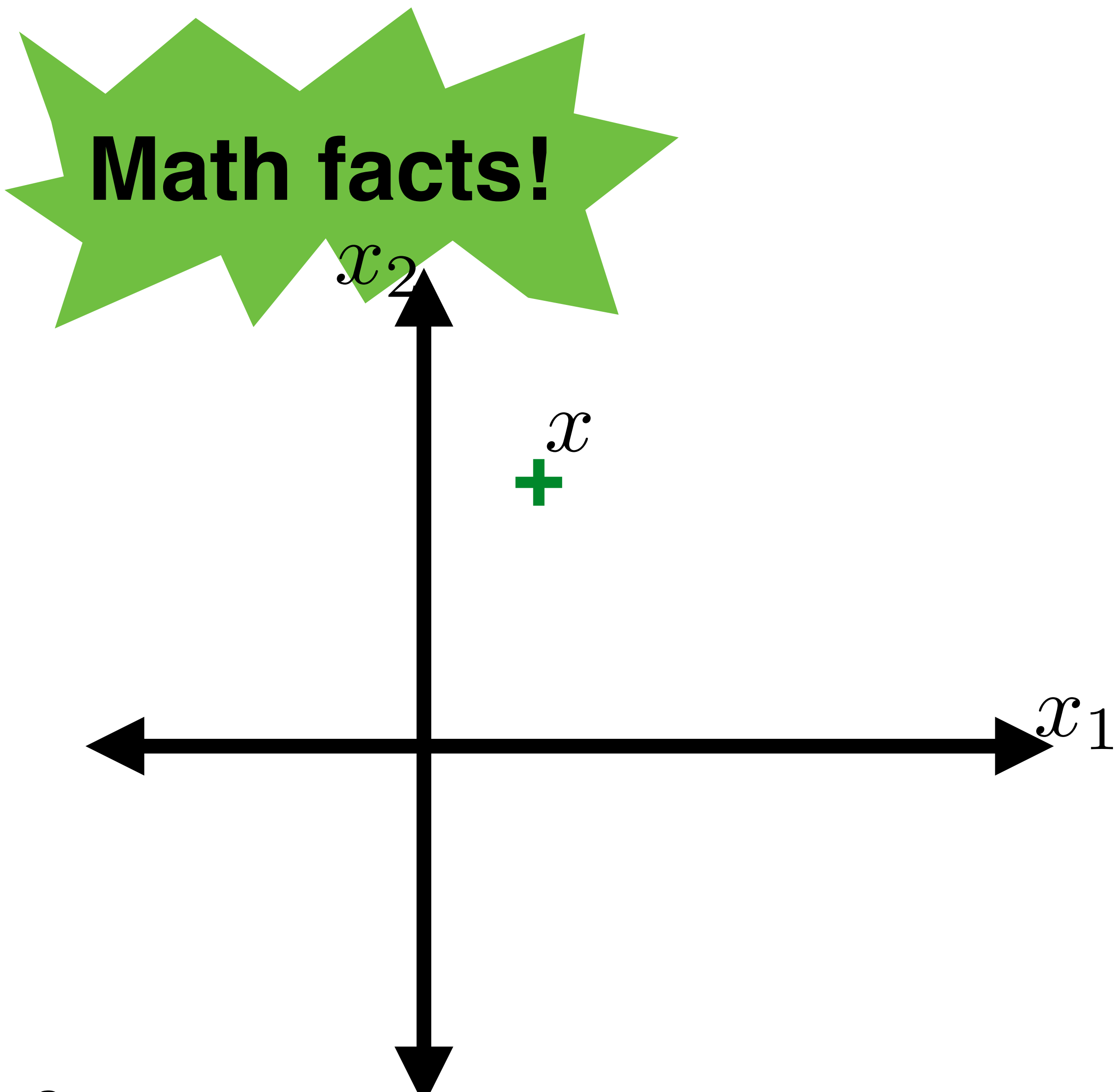
Math facts!



Linear classifiers

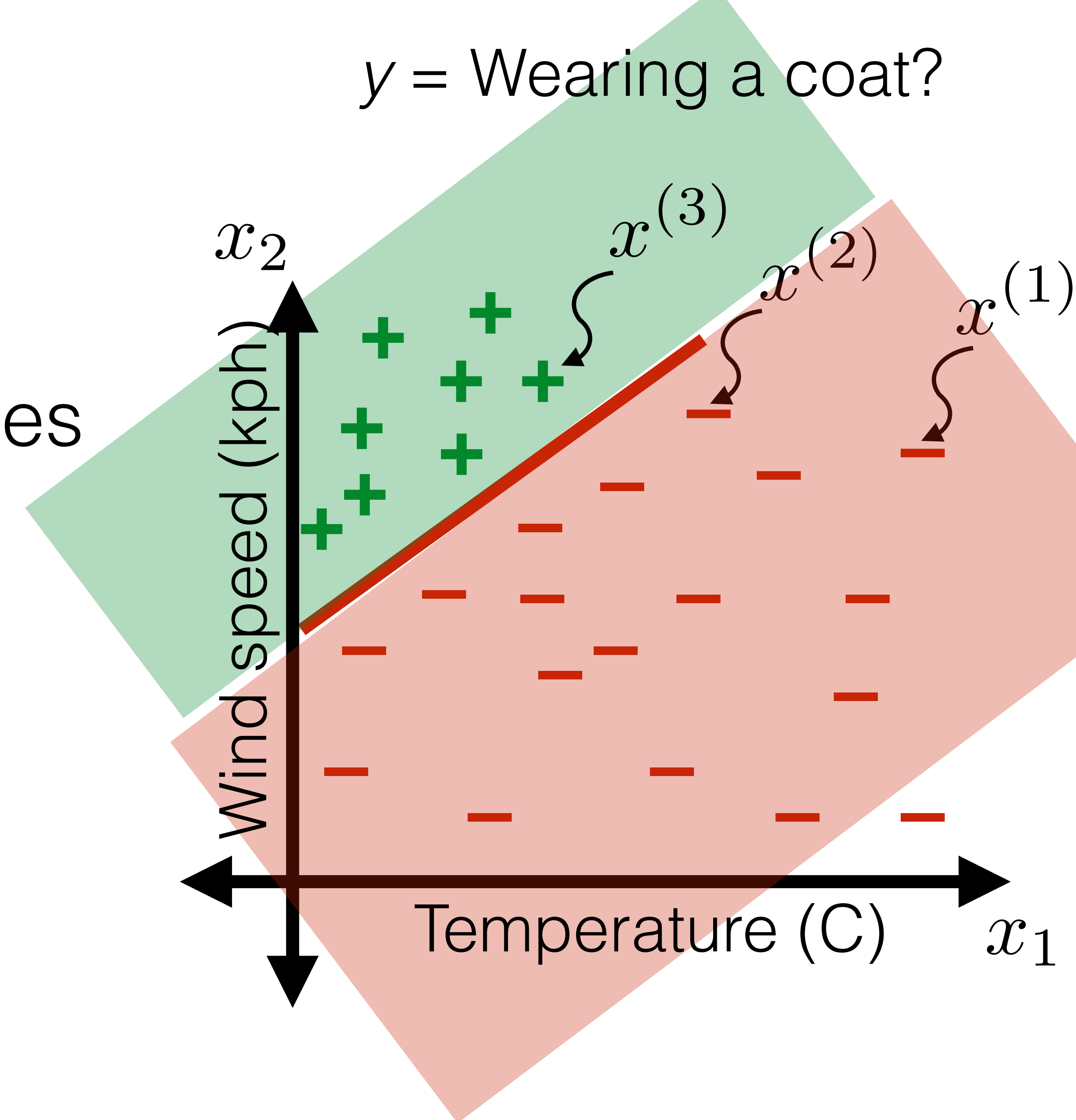
- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

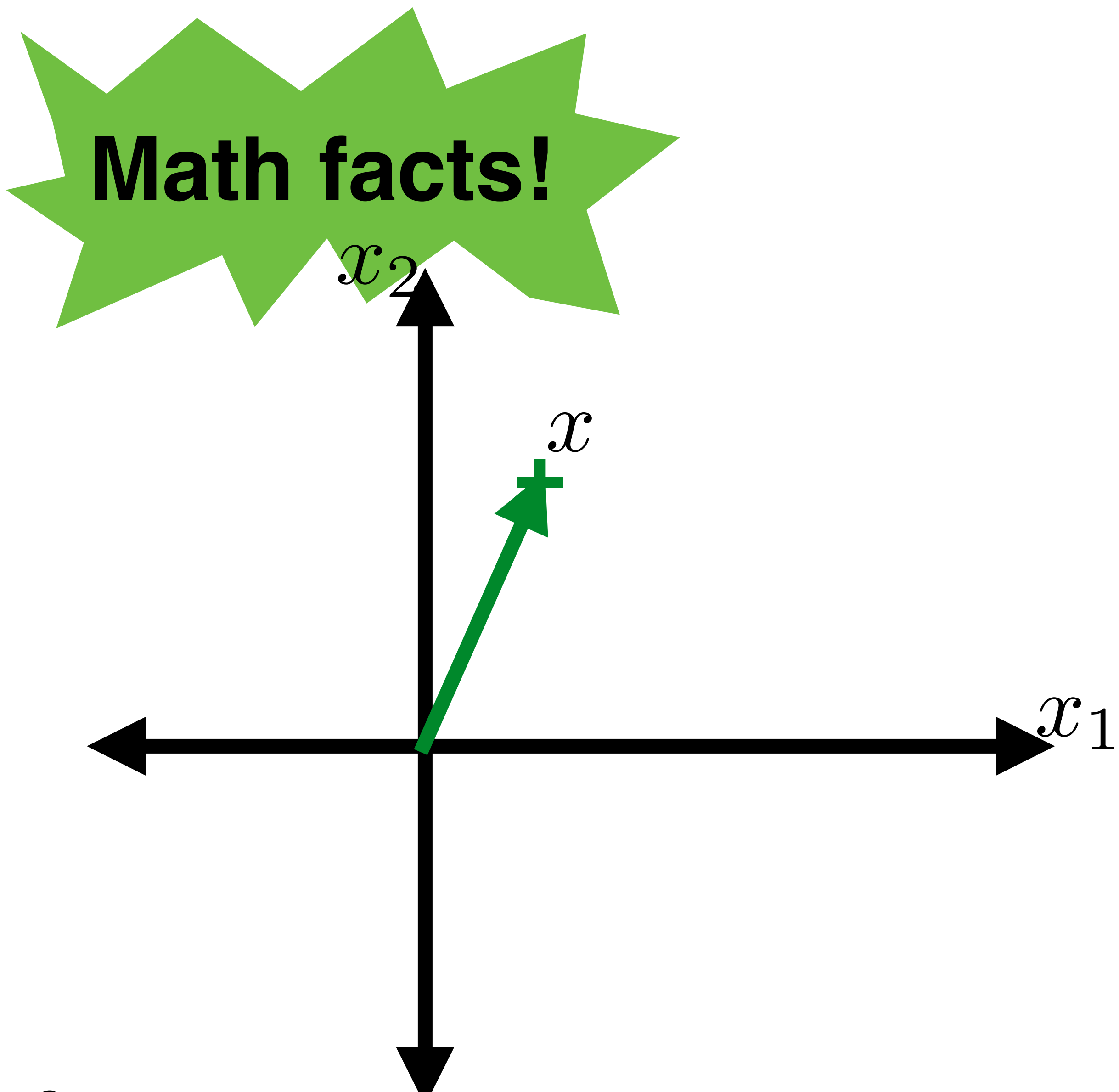




Linear classifiers

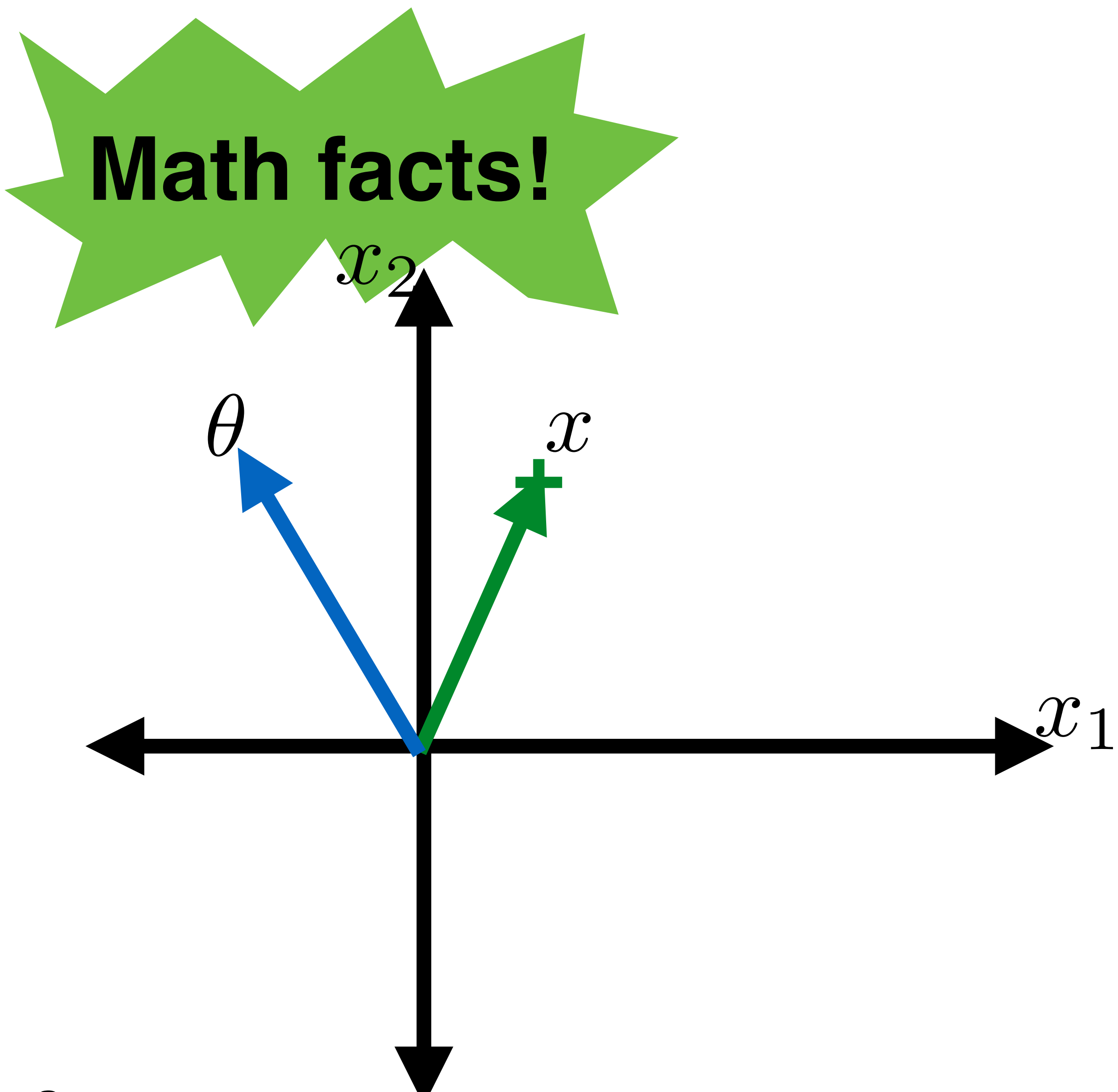
- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

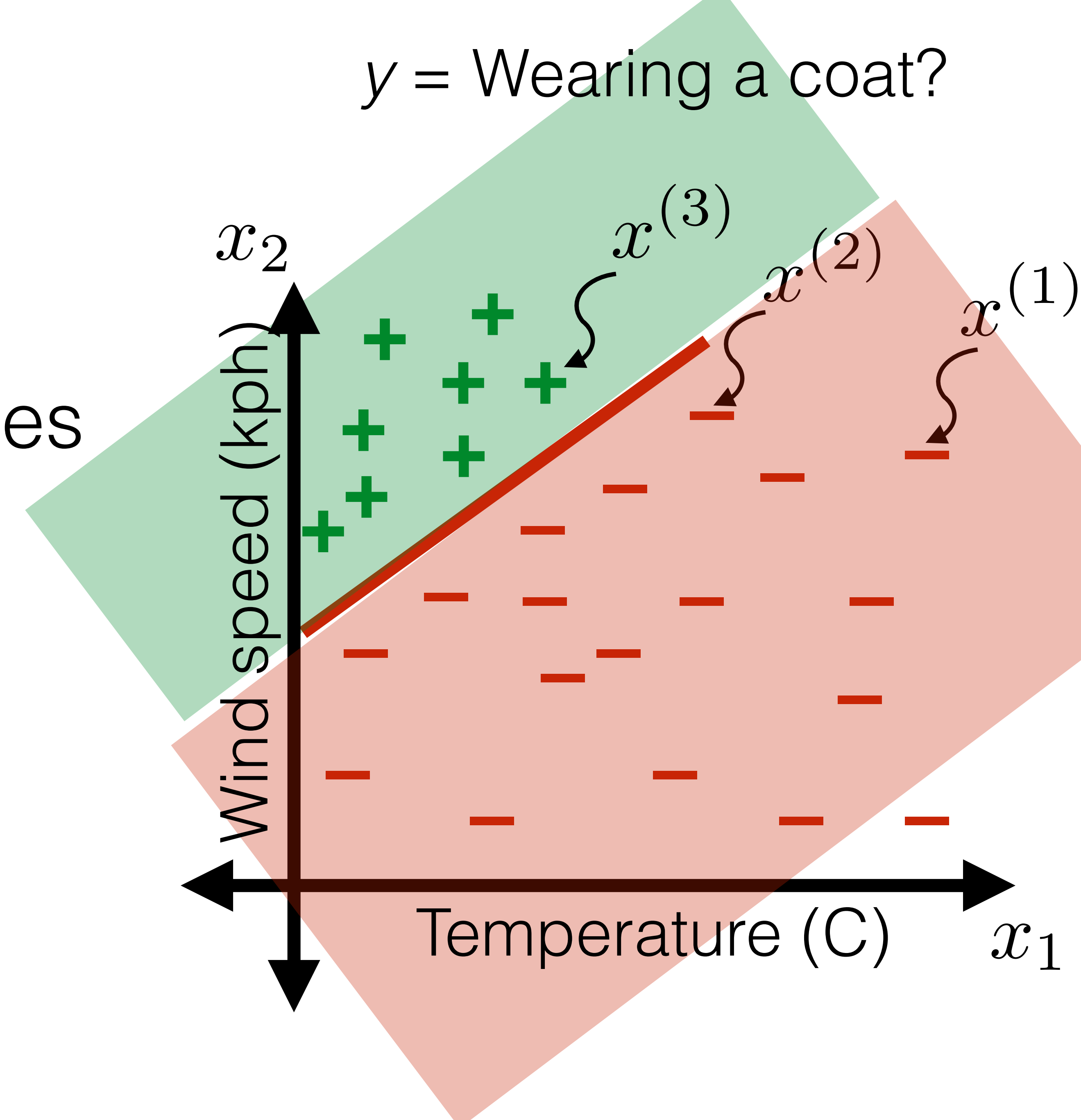




Linear classifiers

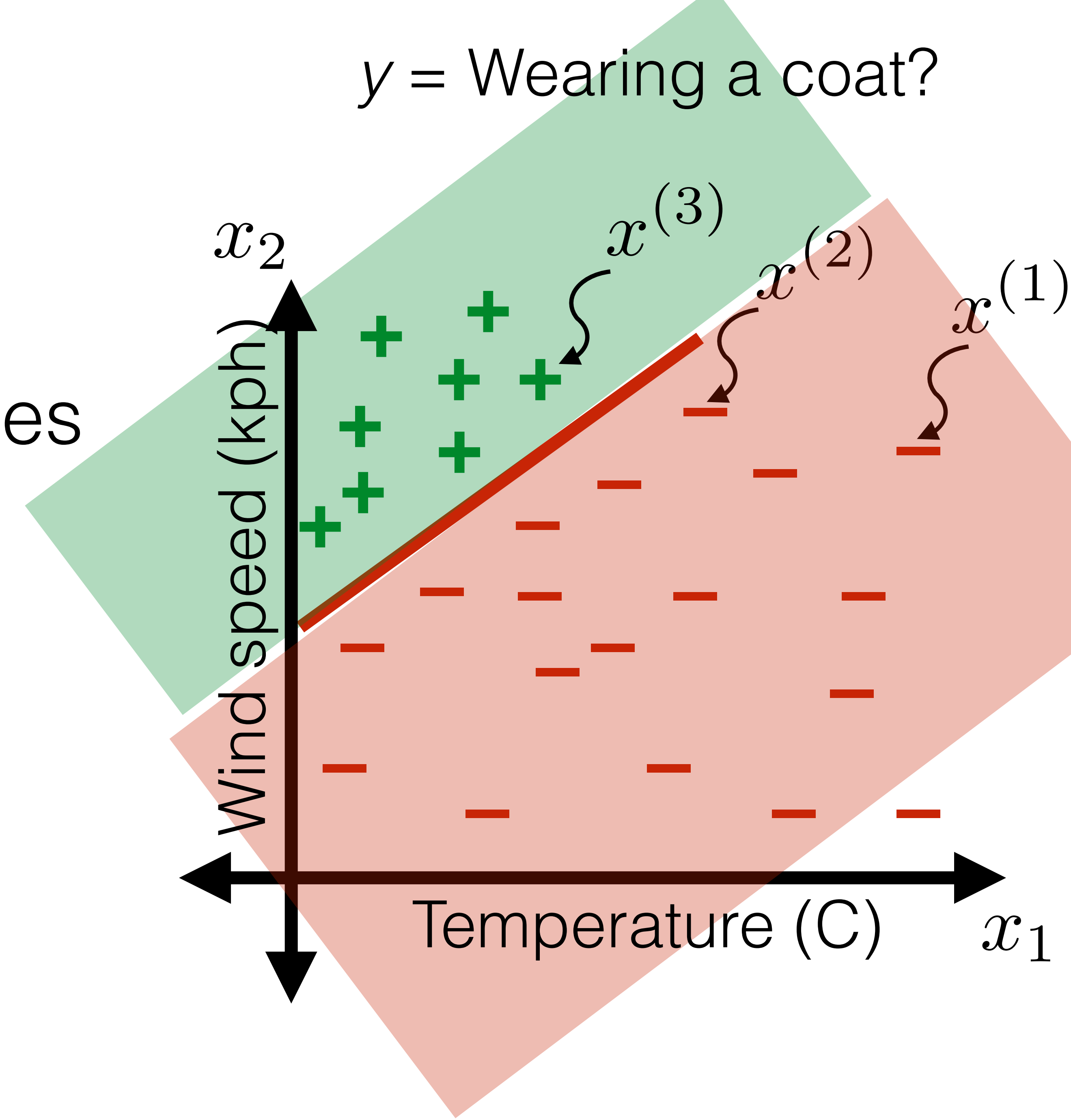
- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

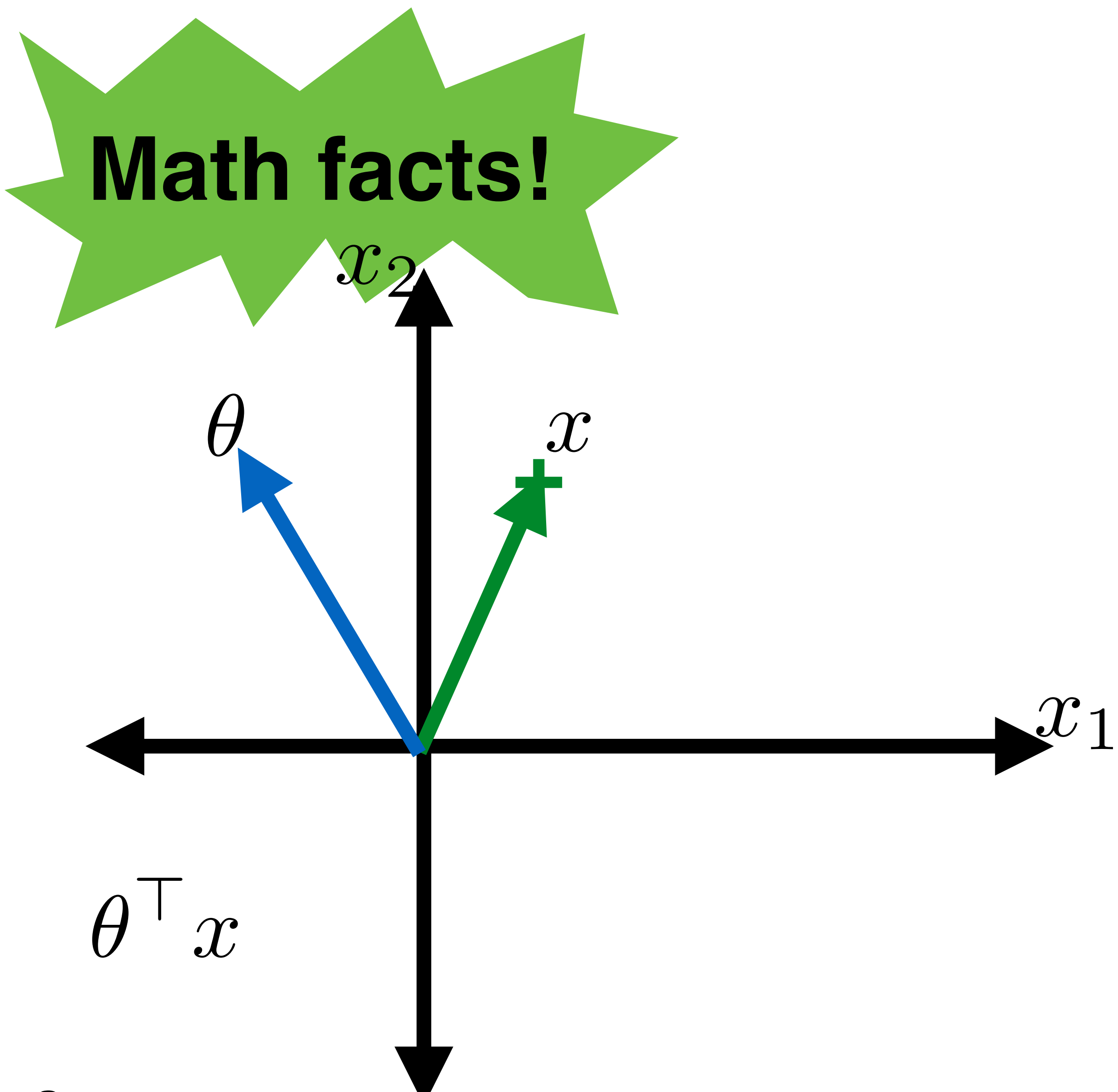




Linear classifiers

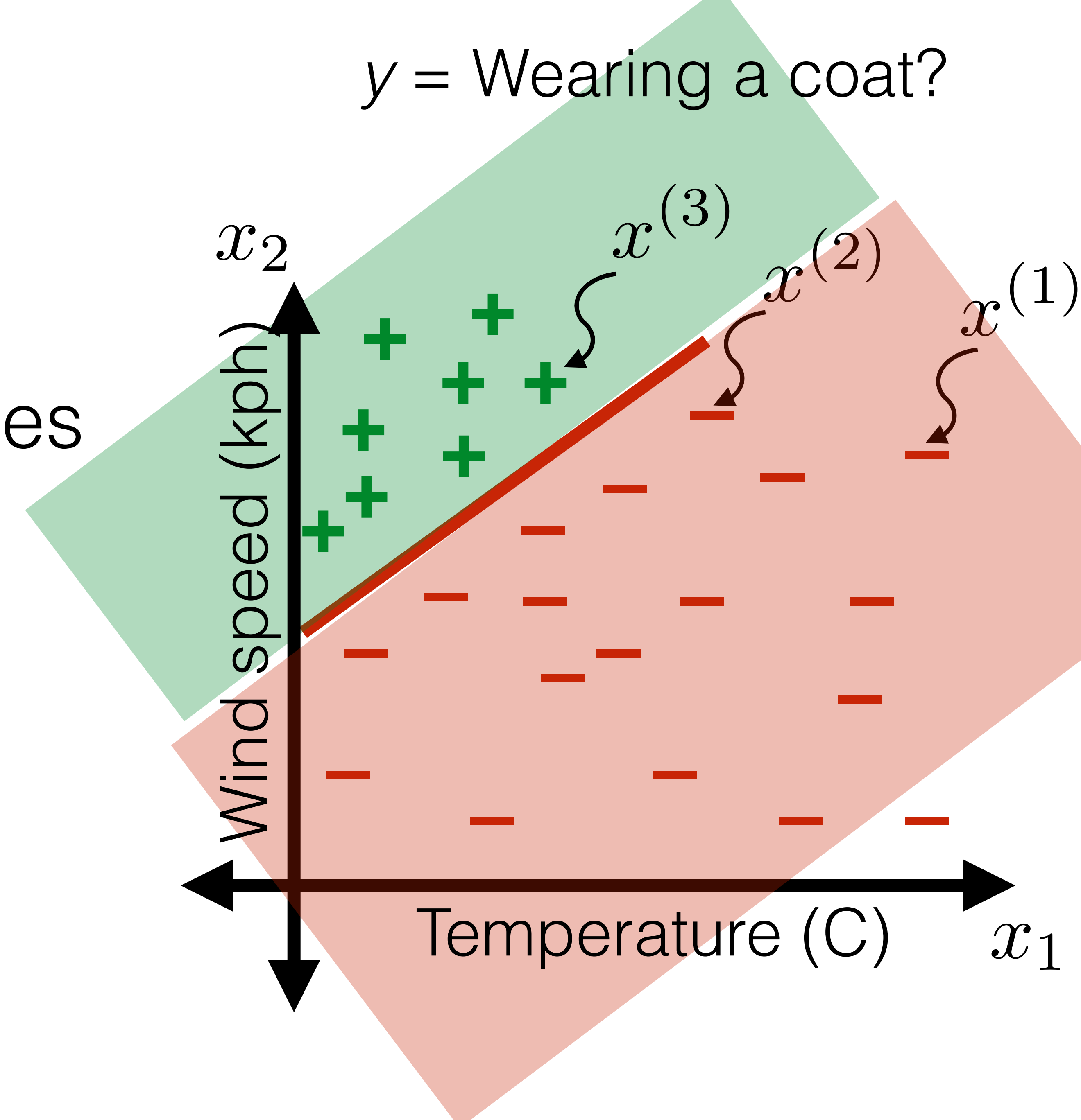
- Classification hypothesis:
  $$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

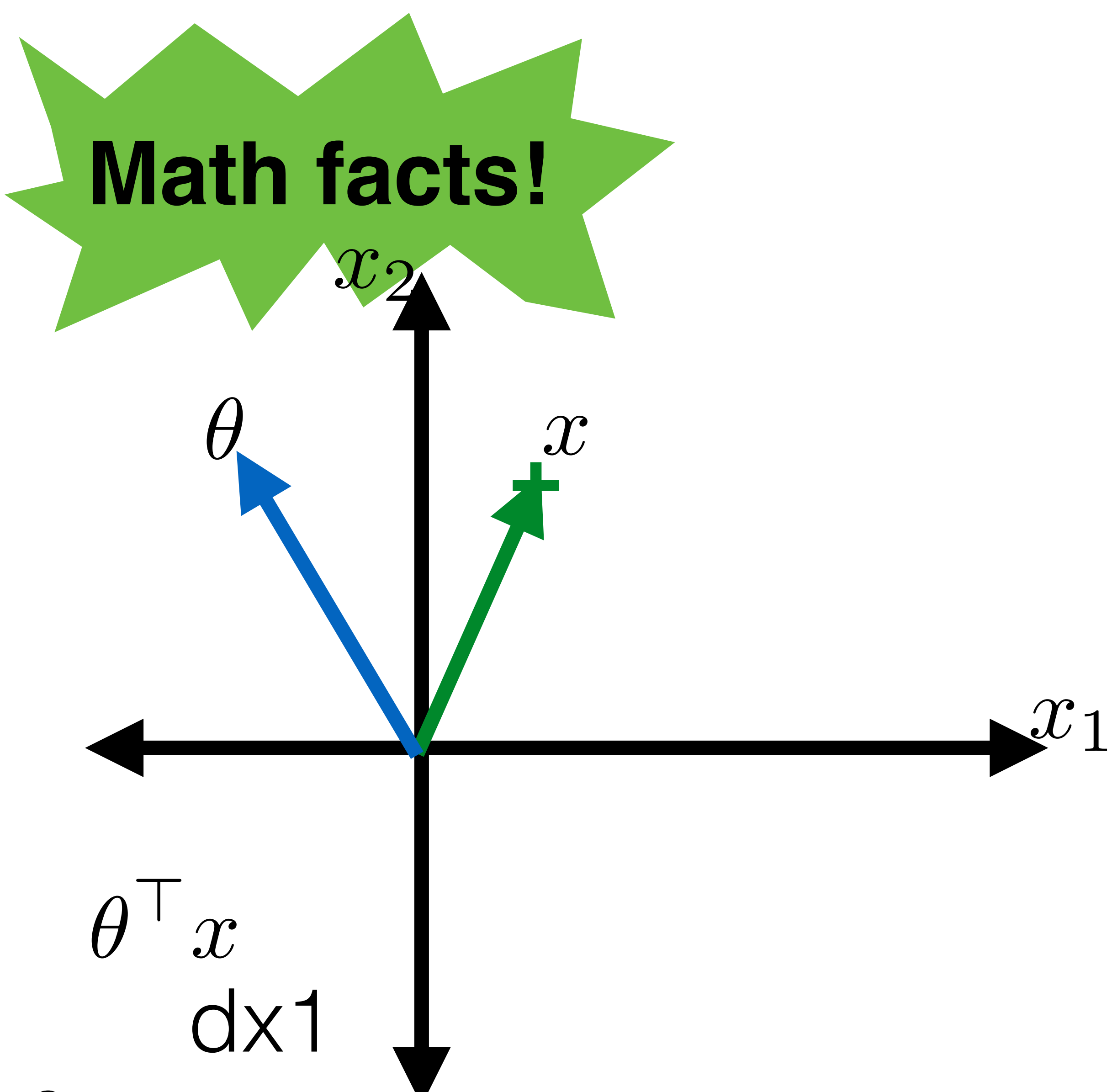




Linear classifiers

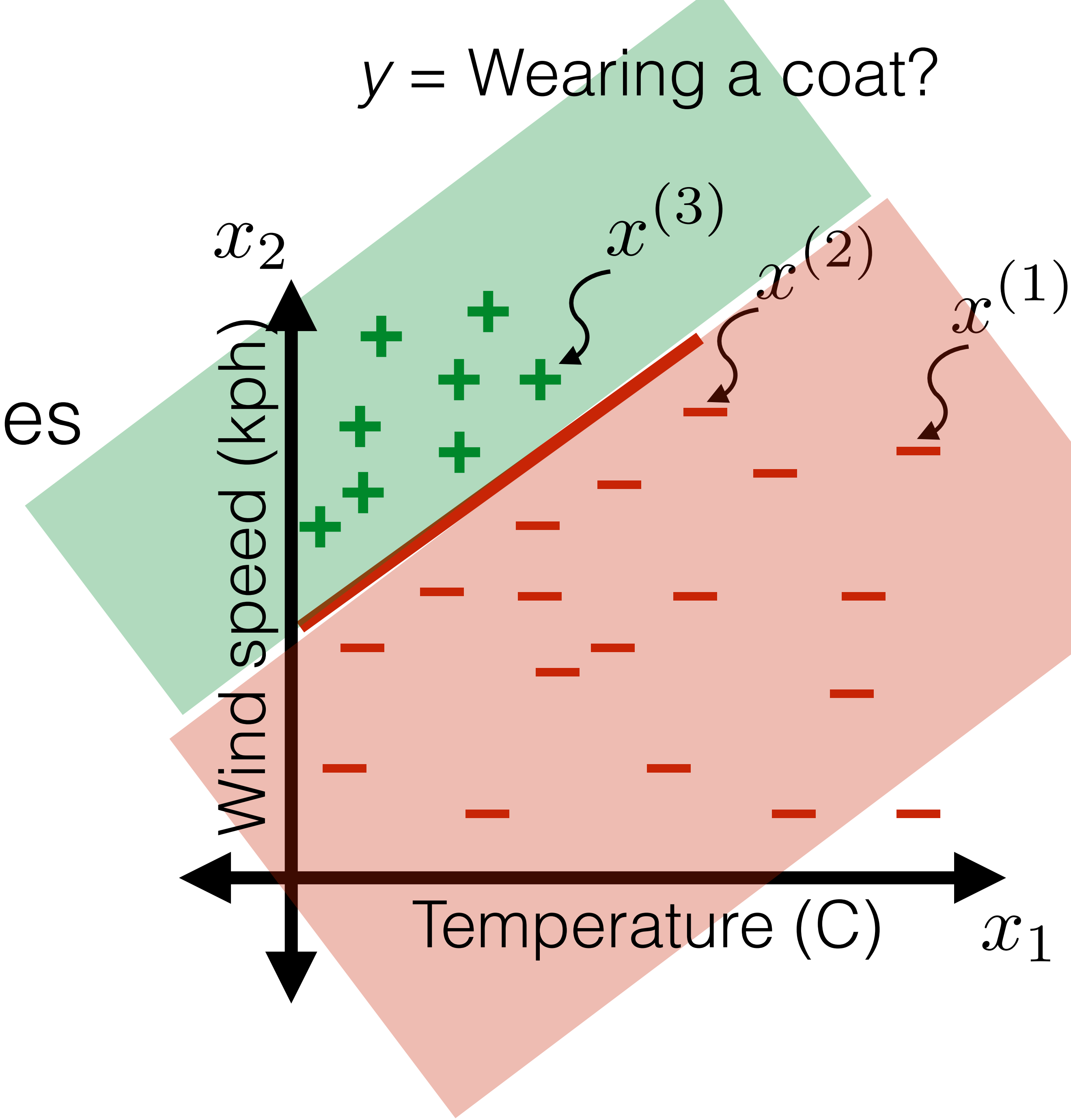
- Classification hypothesis:
  $$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



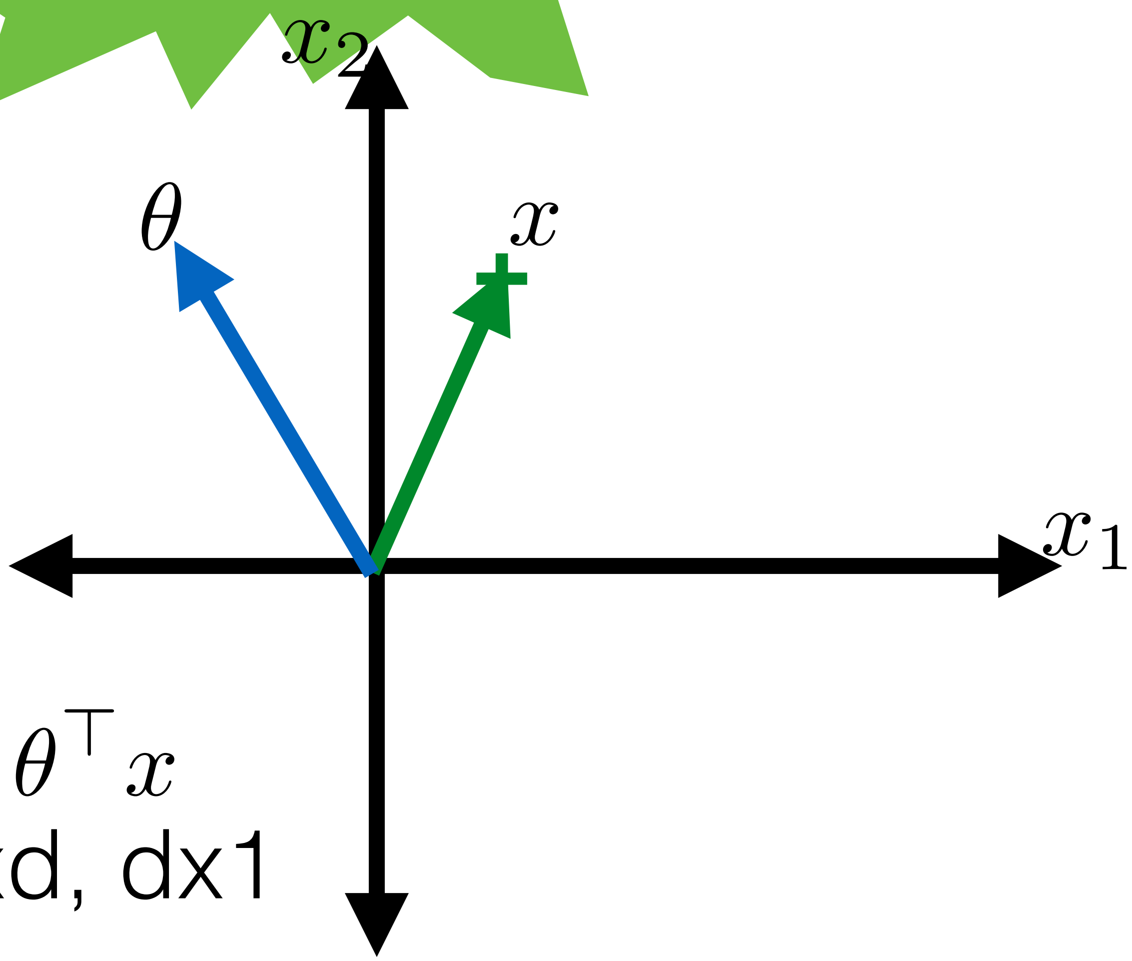


Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



Math facts!

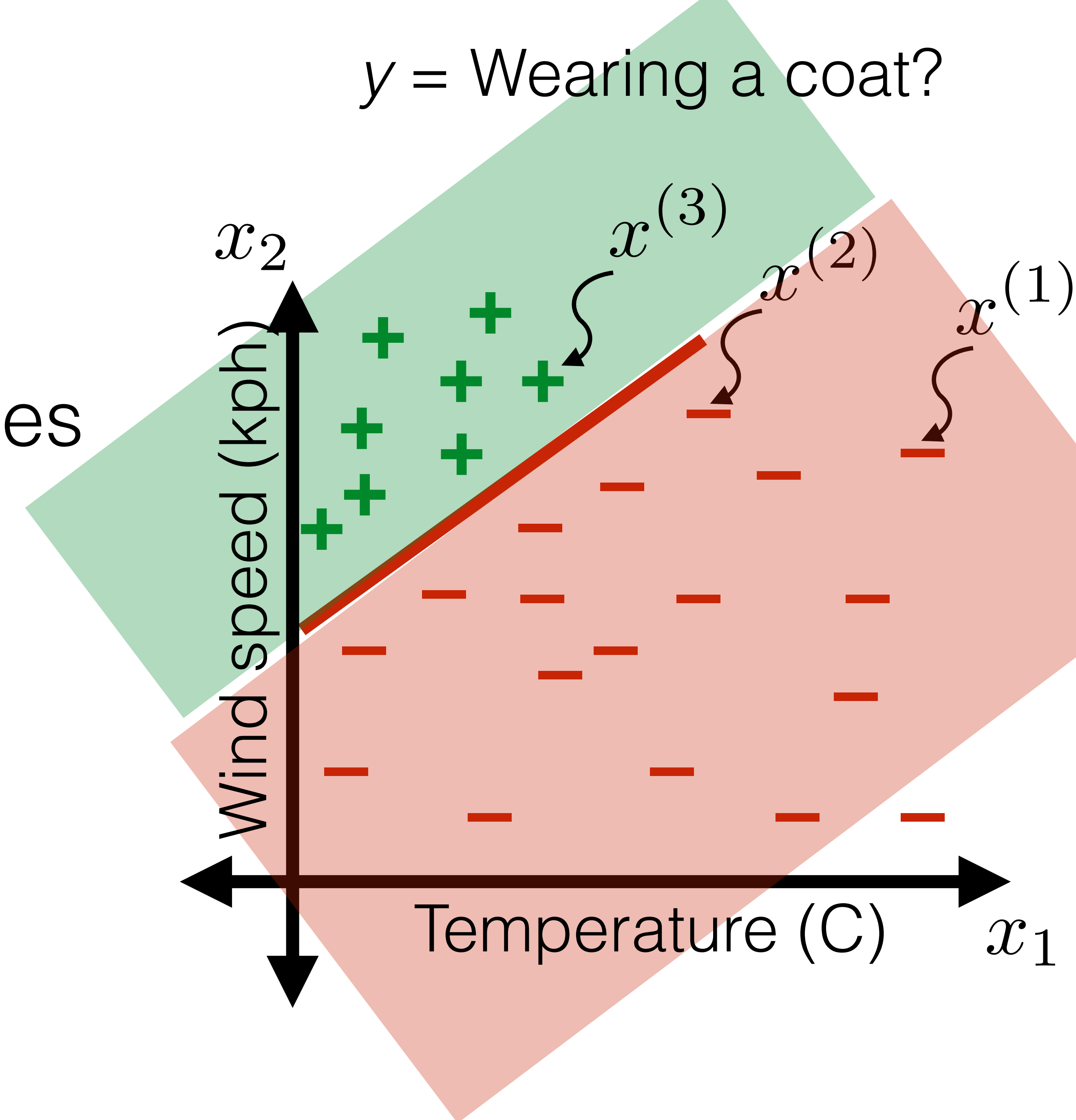


$\theta^\top x$

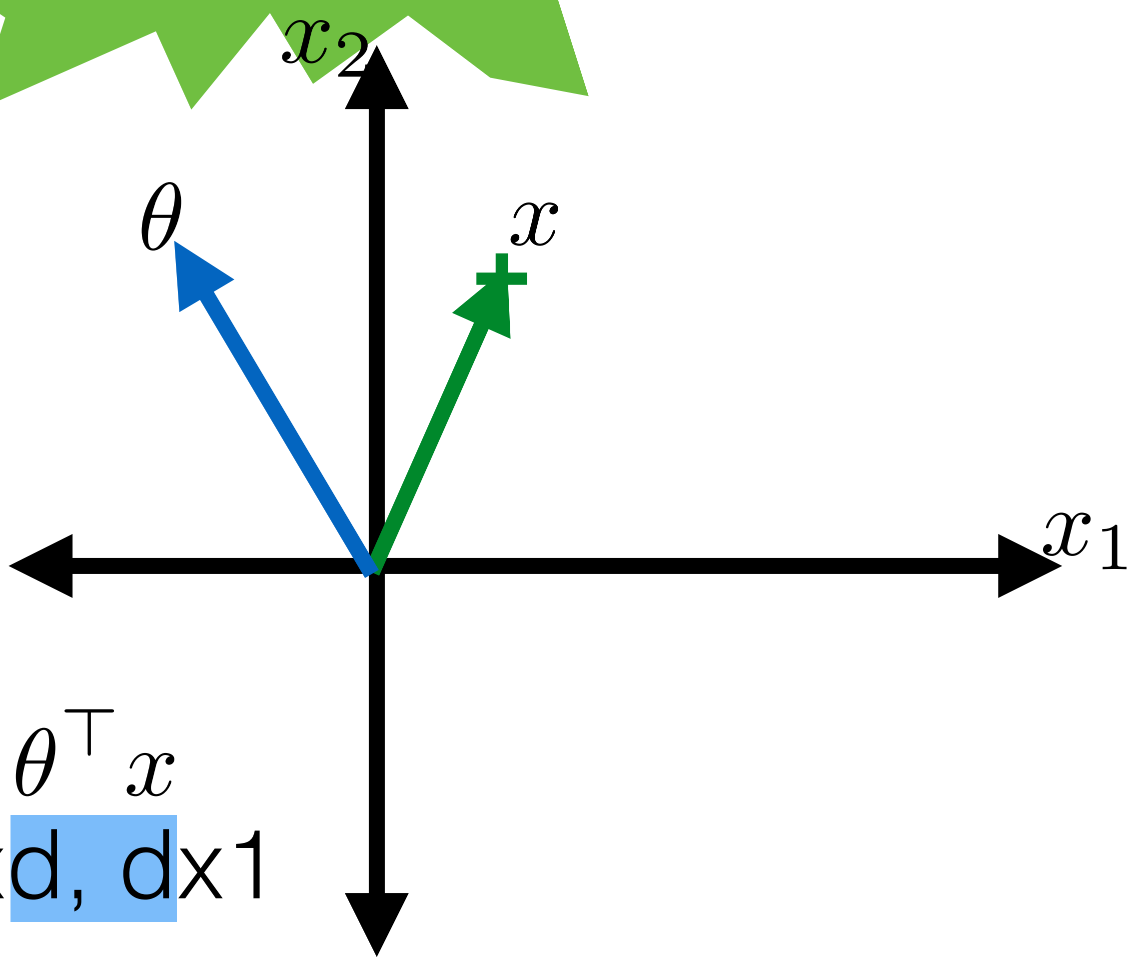
1xd, dx1

Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



Math facts!

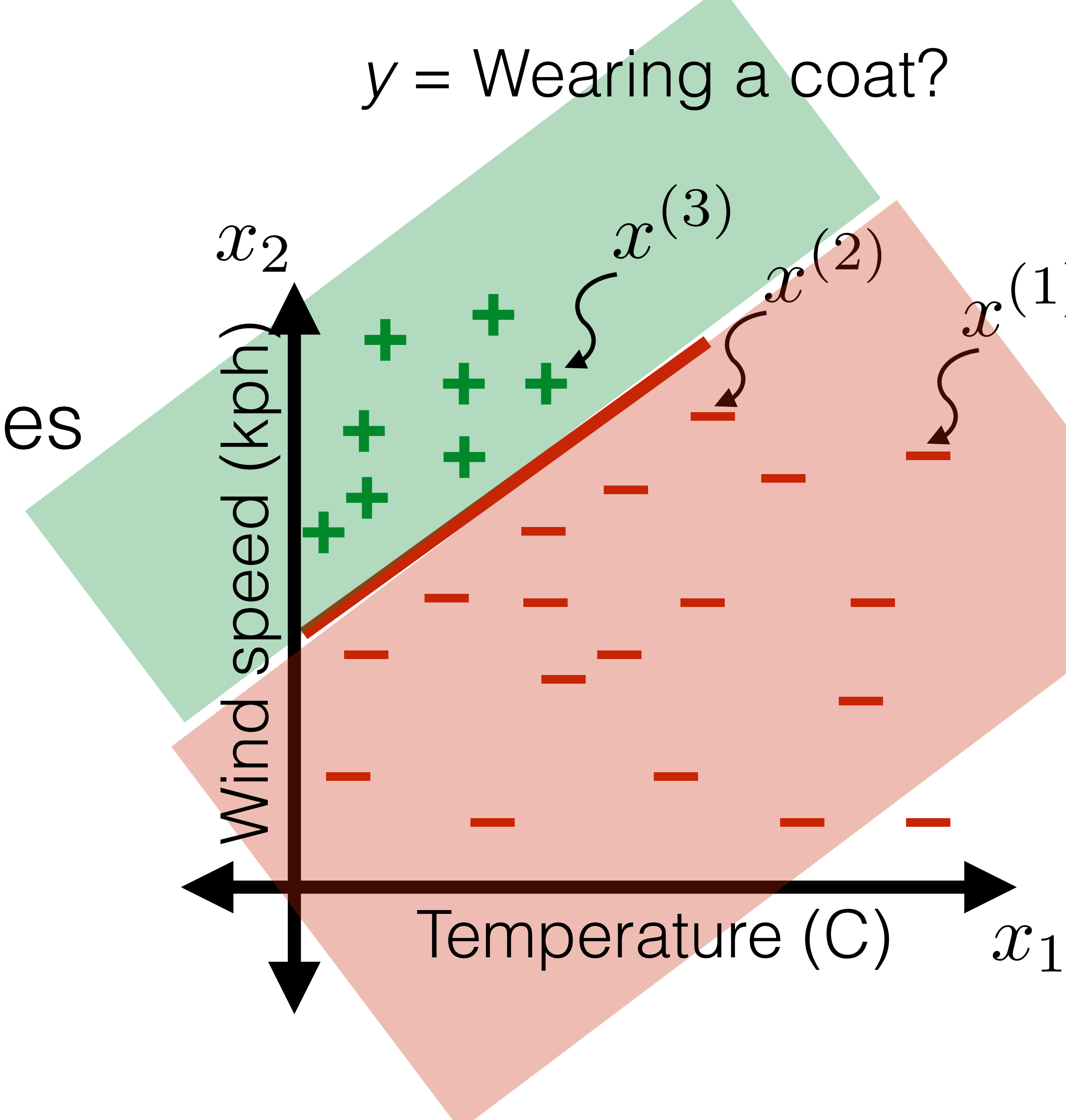


$\theta^\top x$

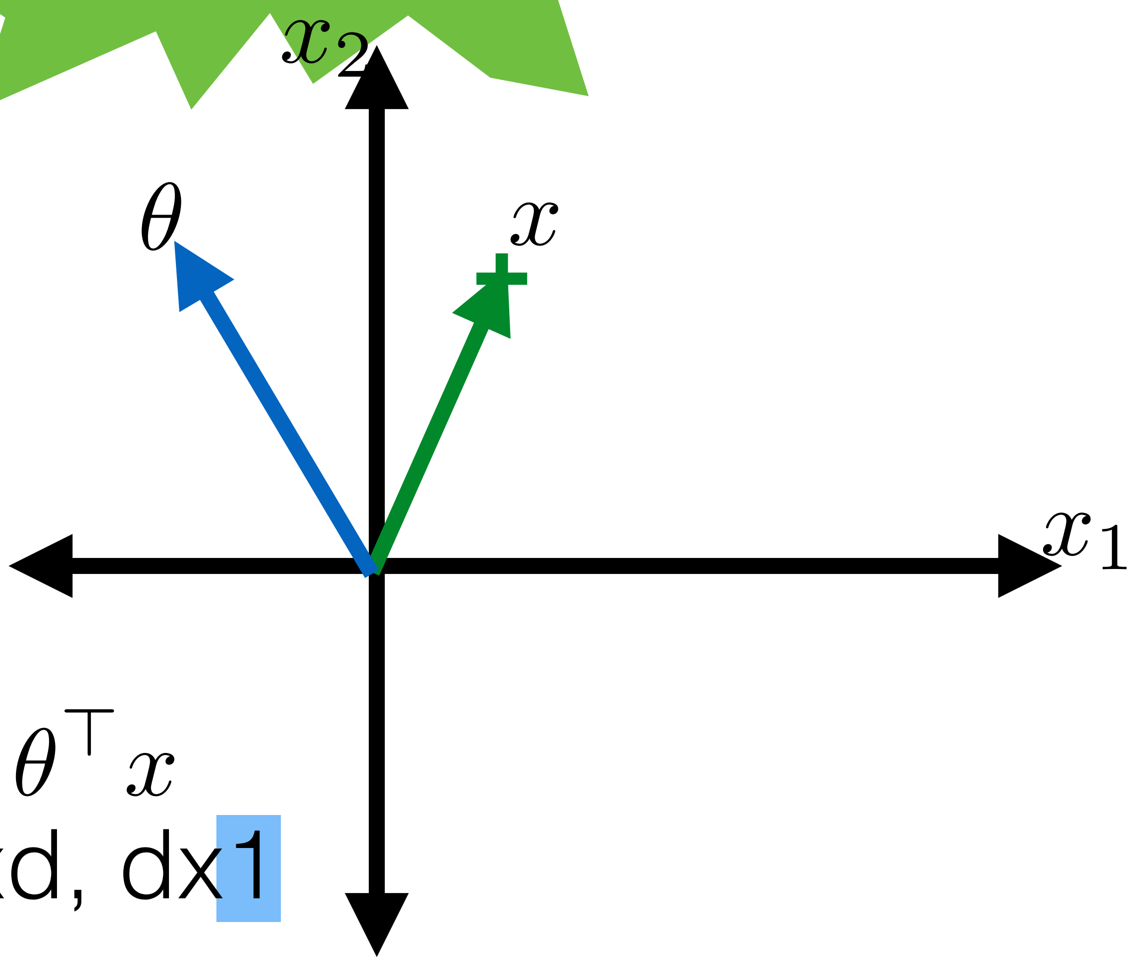
1xd, dx1

Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



Math facts!

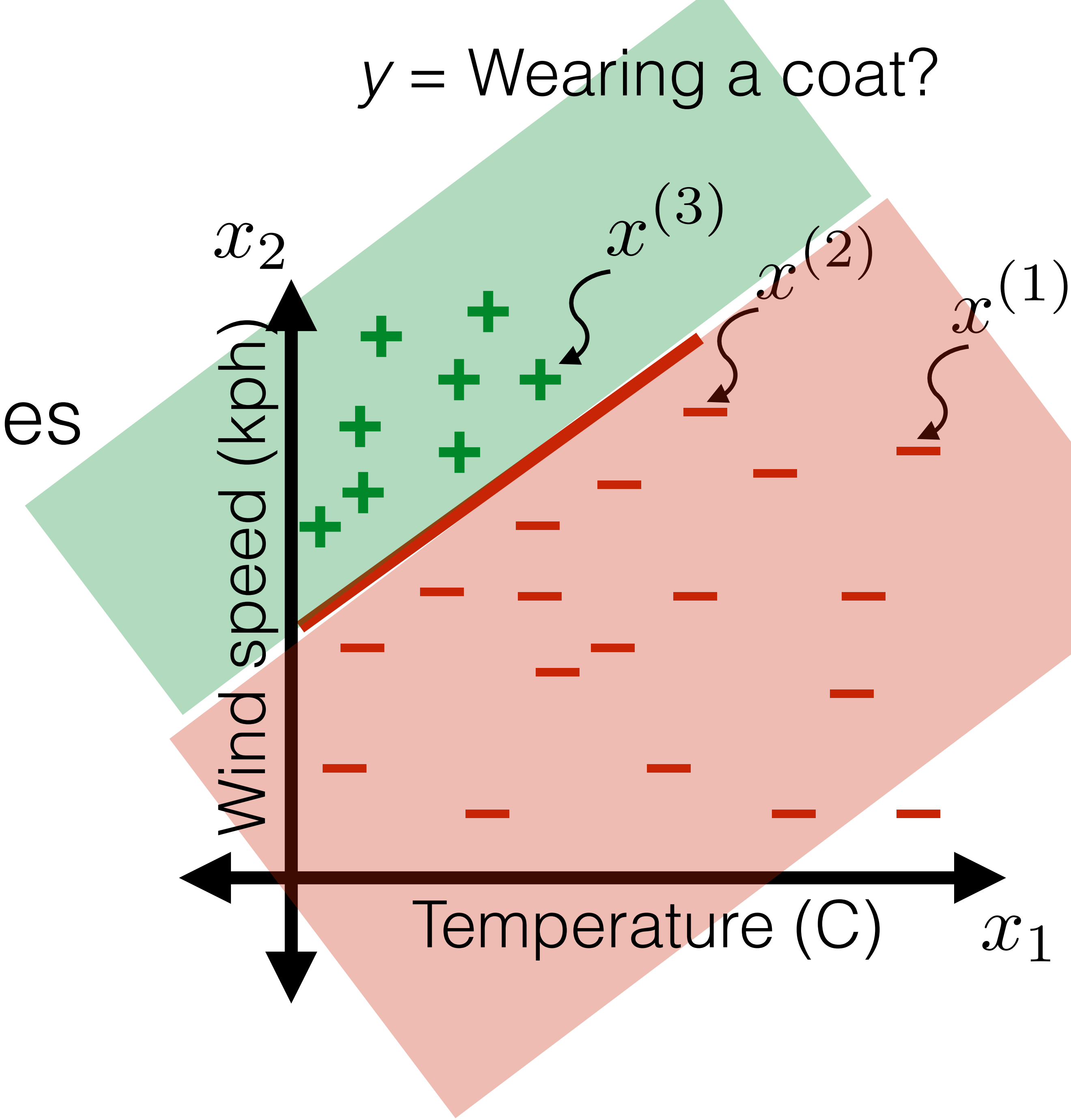


$\theta^\top x$

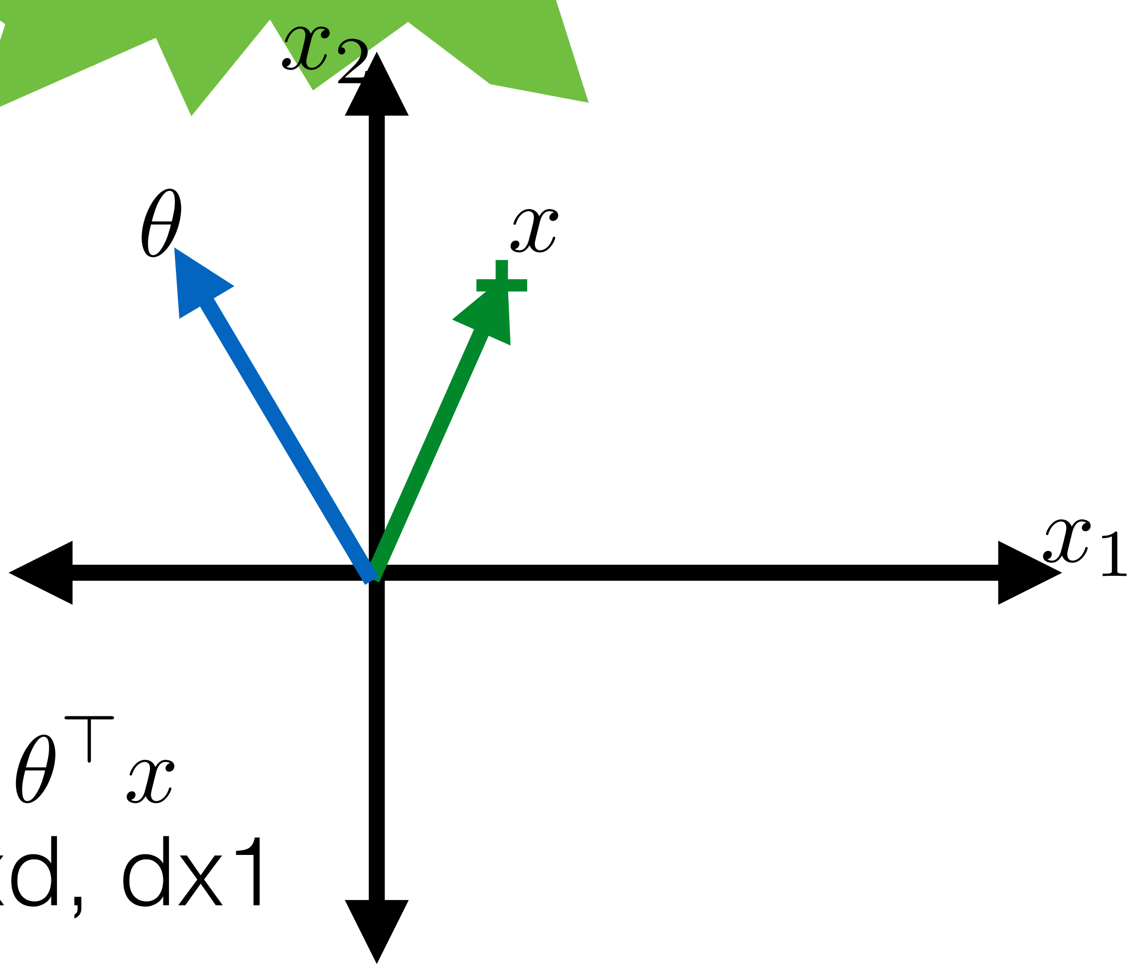
1xd, dx1

Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



Math facts!



$\theta^\top x$

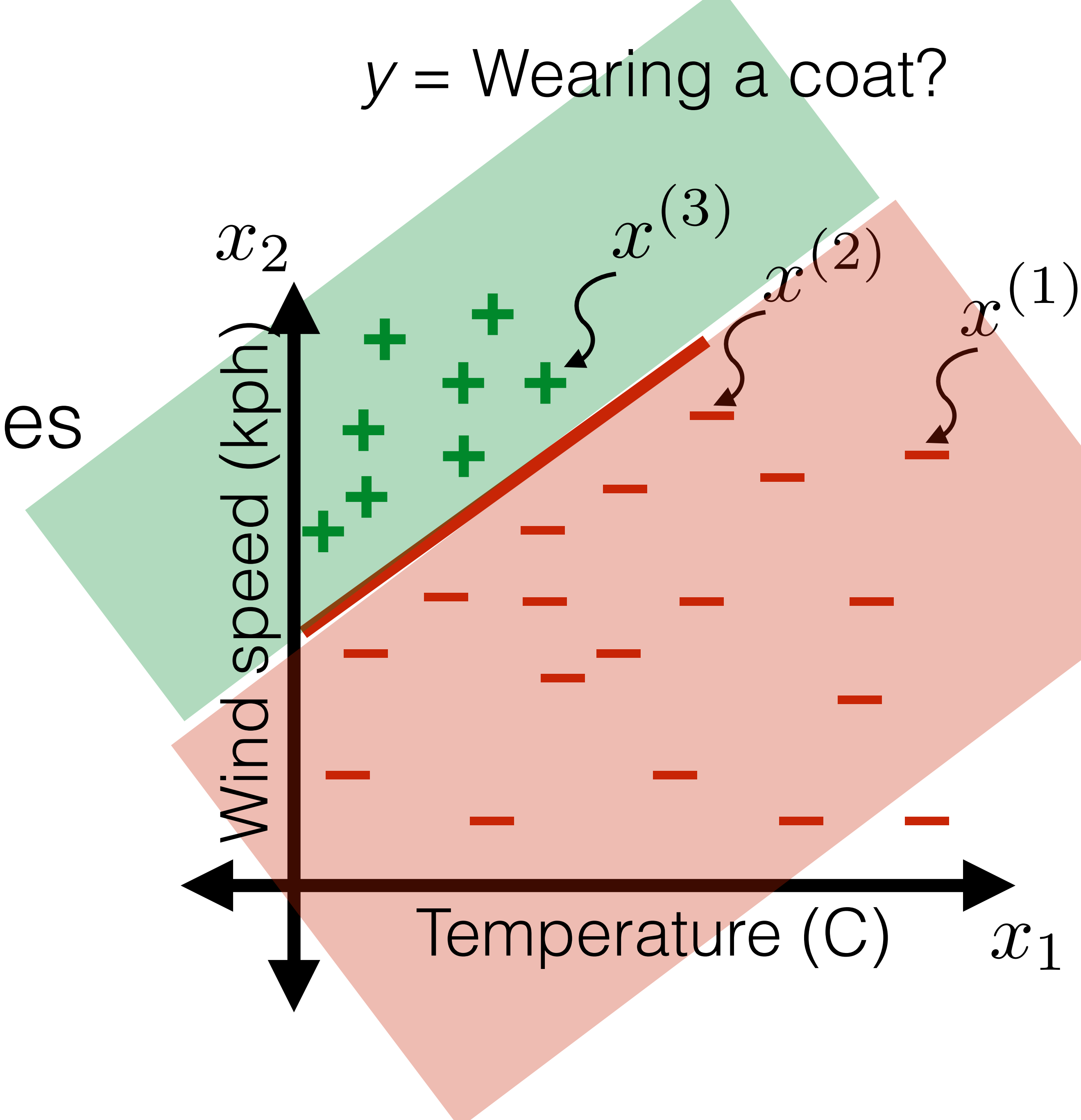
1xd, dx1

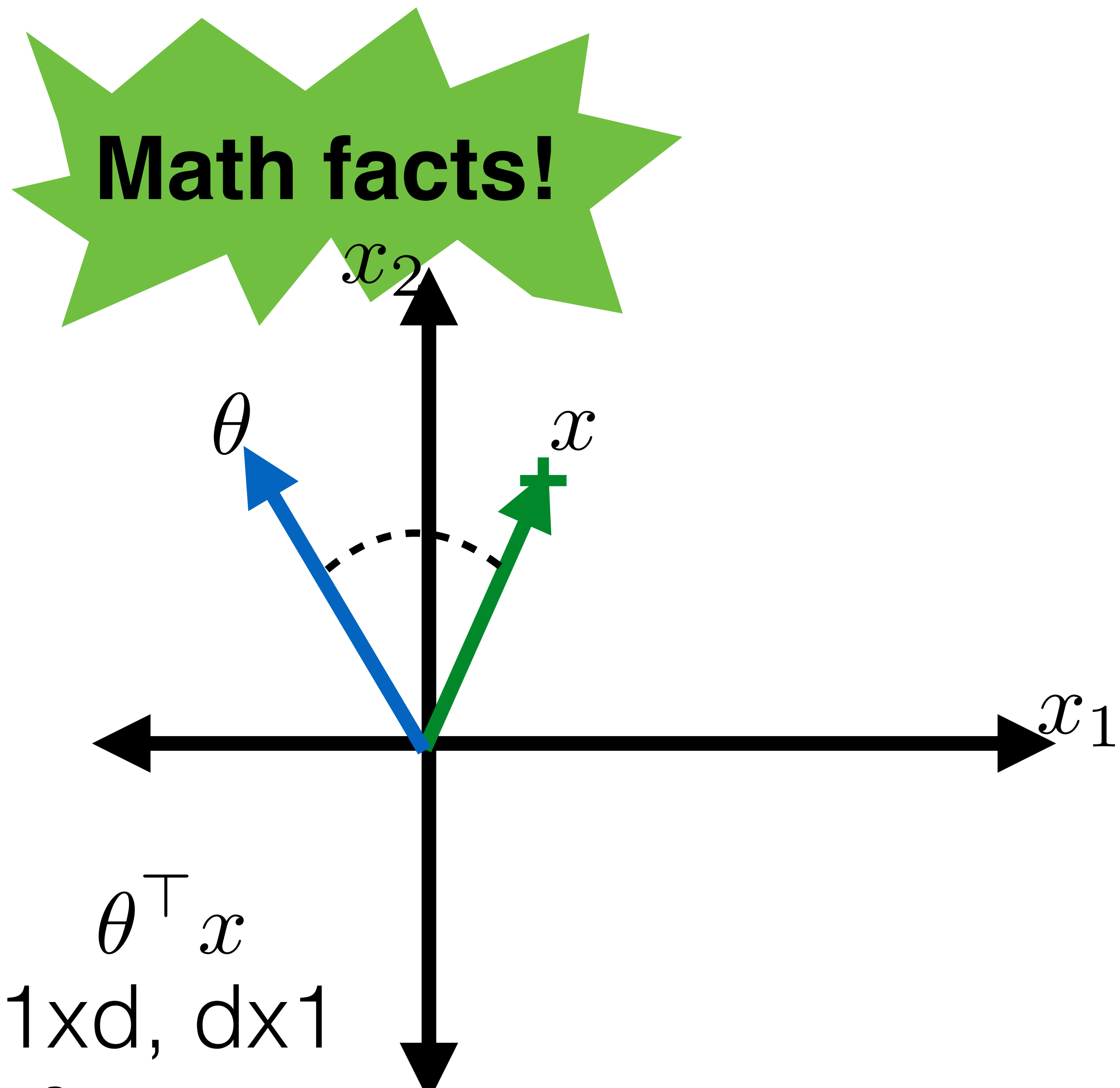
Linear classifiers

- Classification hypothesis:

$$h : \mathbb{R}^d \to \{-1, +1\}$$

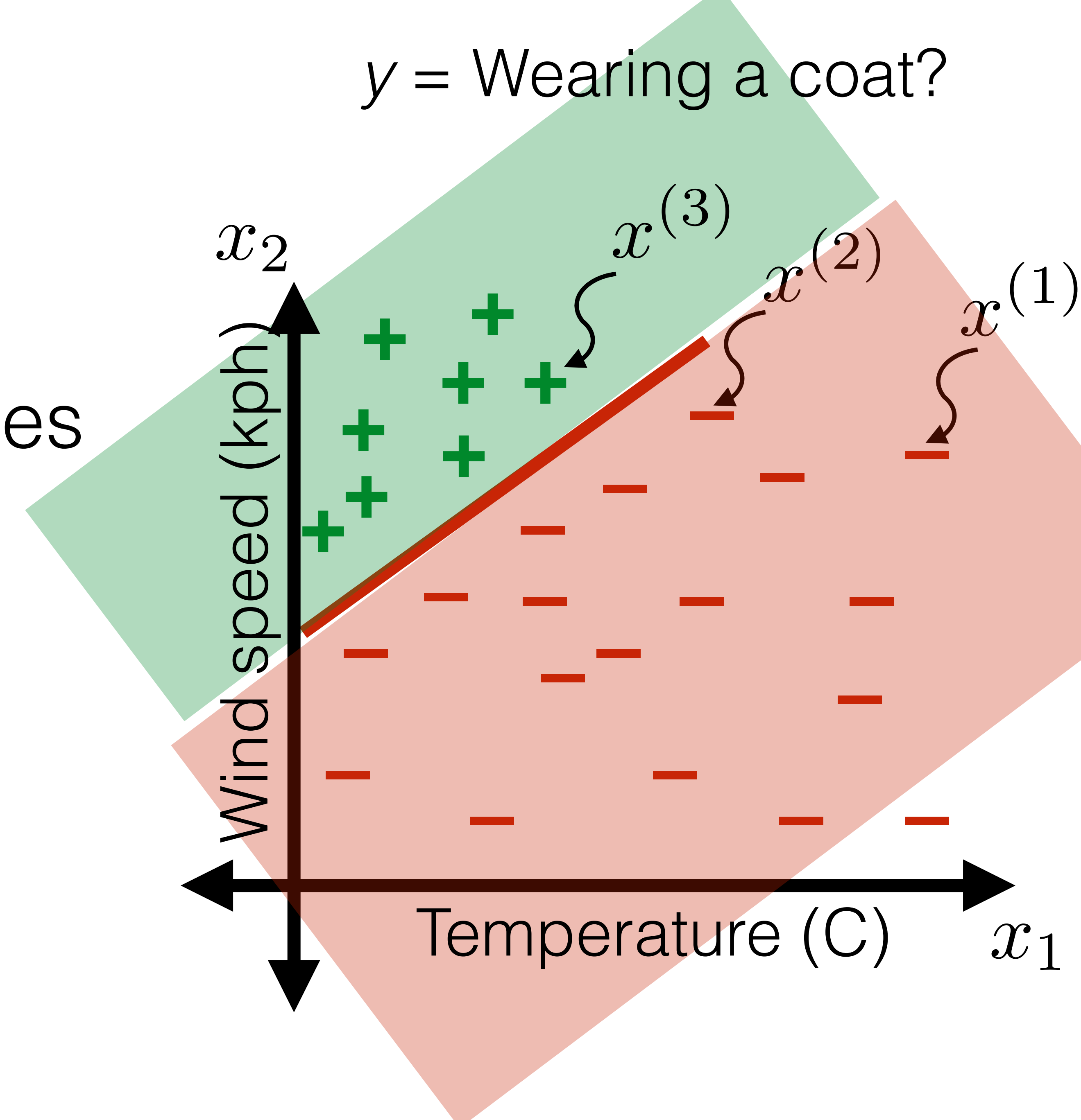
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



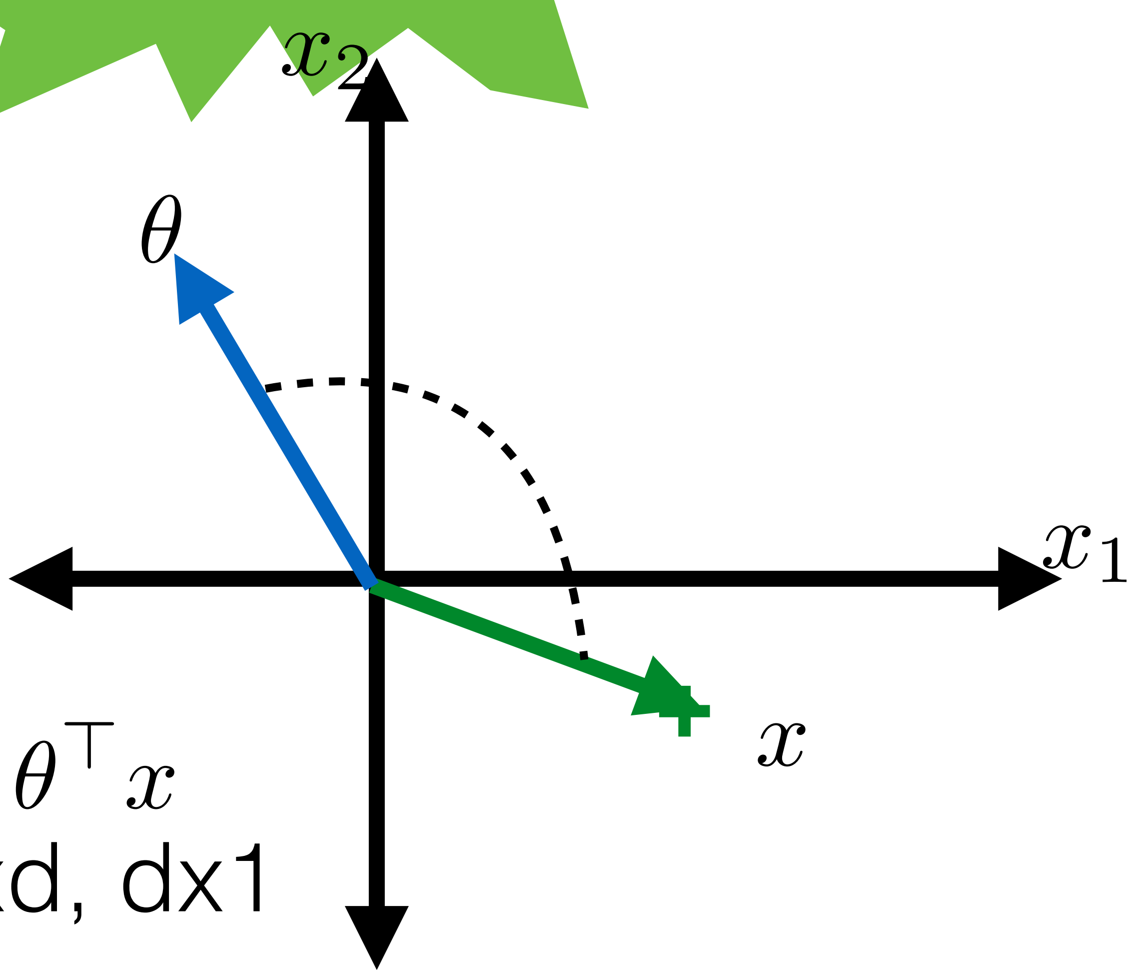


Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



Math facts!

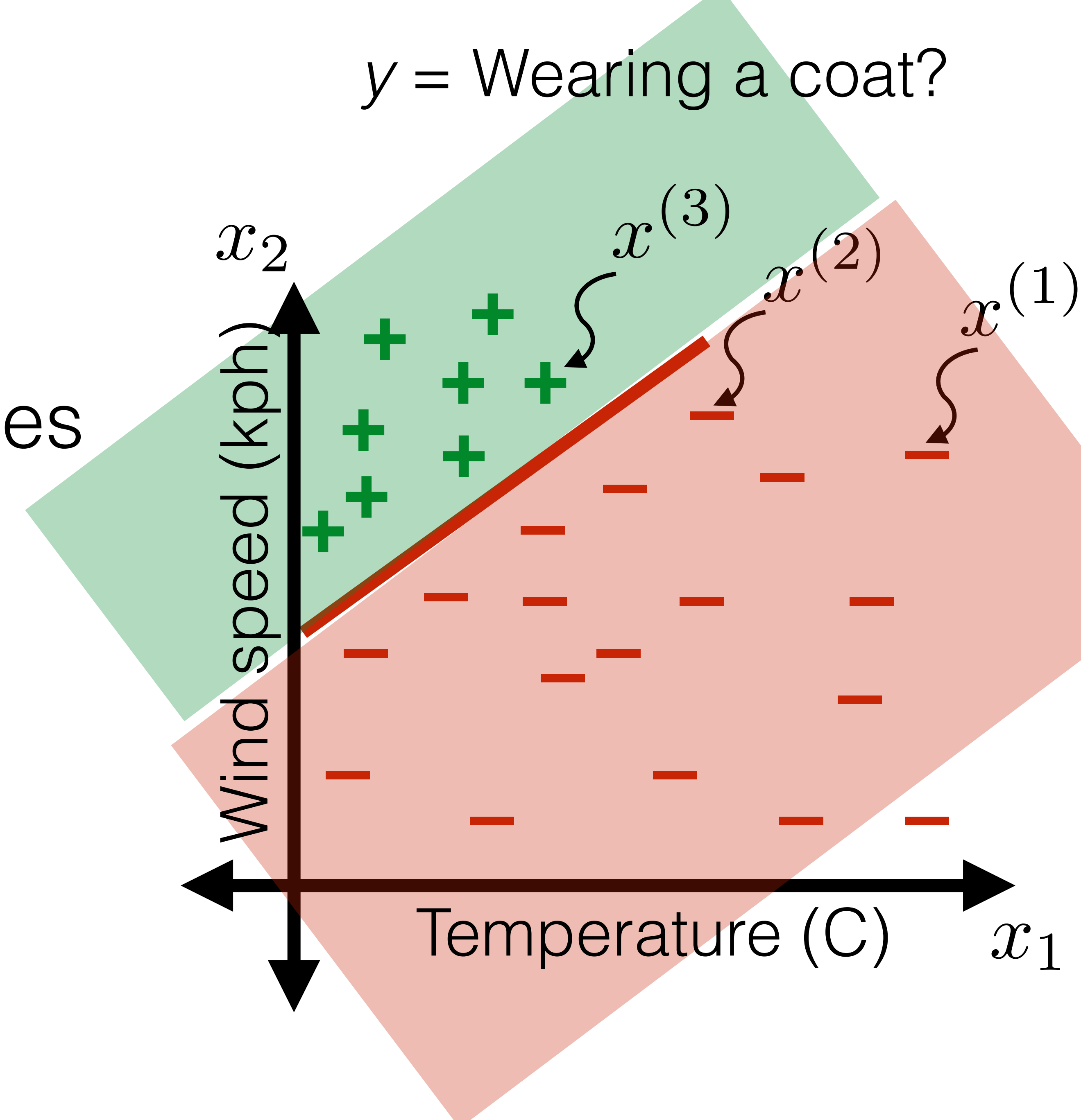


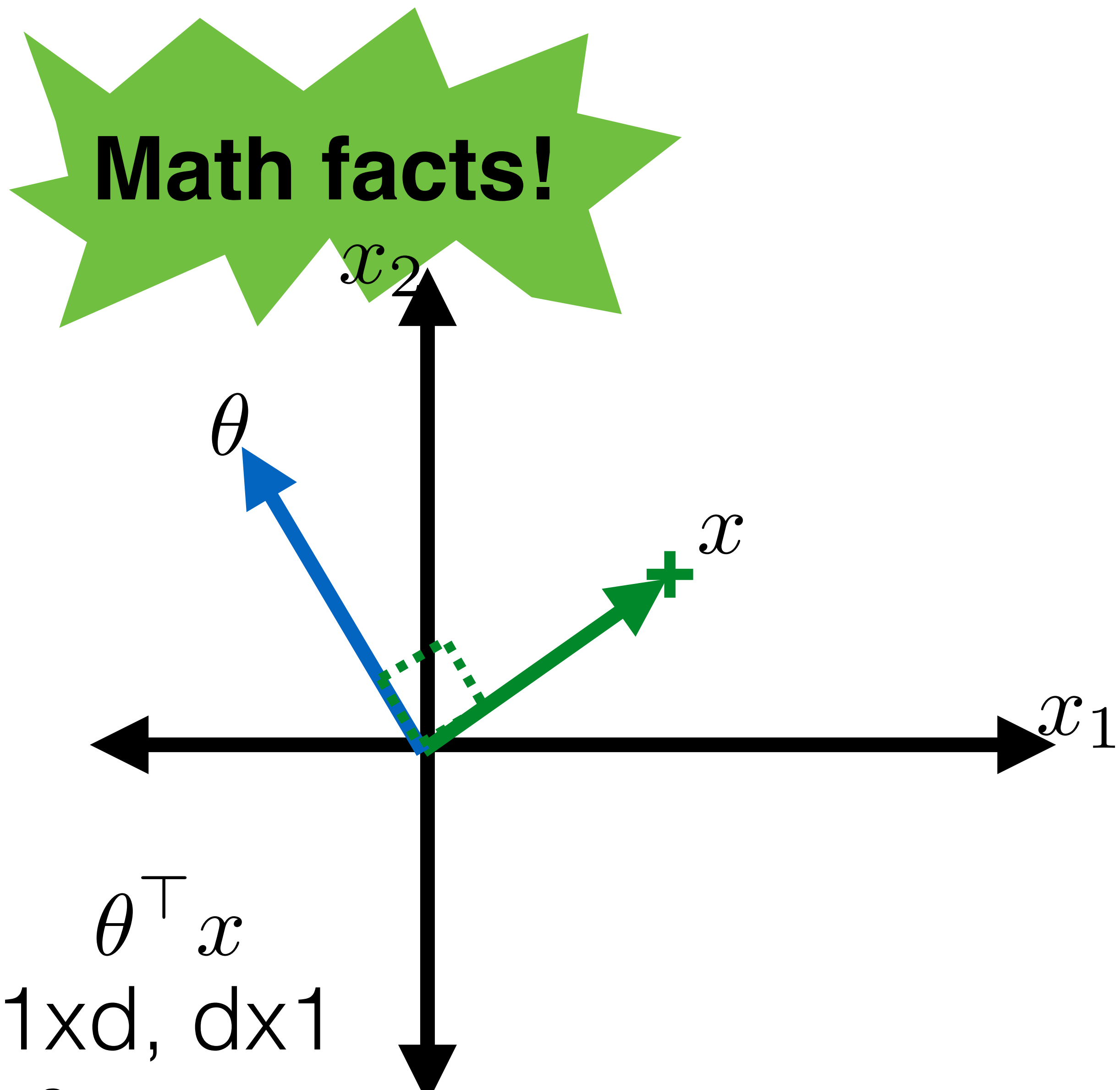
$\theta^\top x$

1xd, dx1

Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



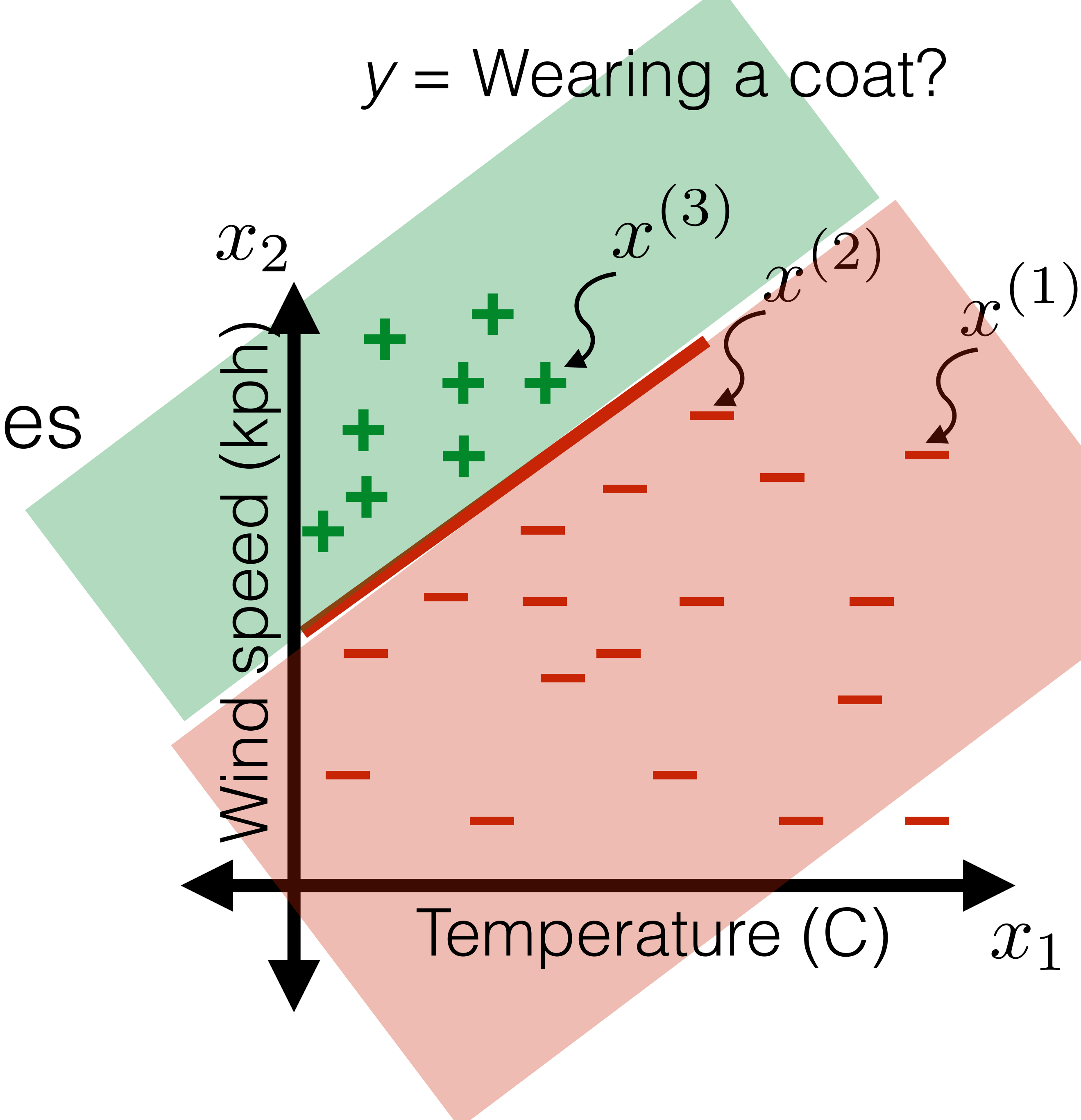


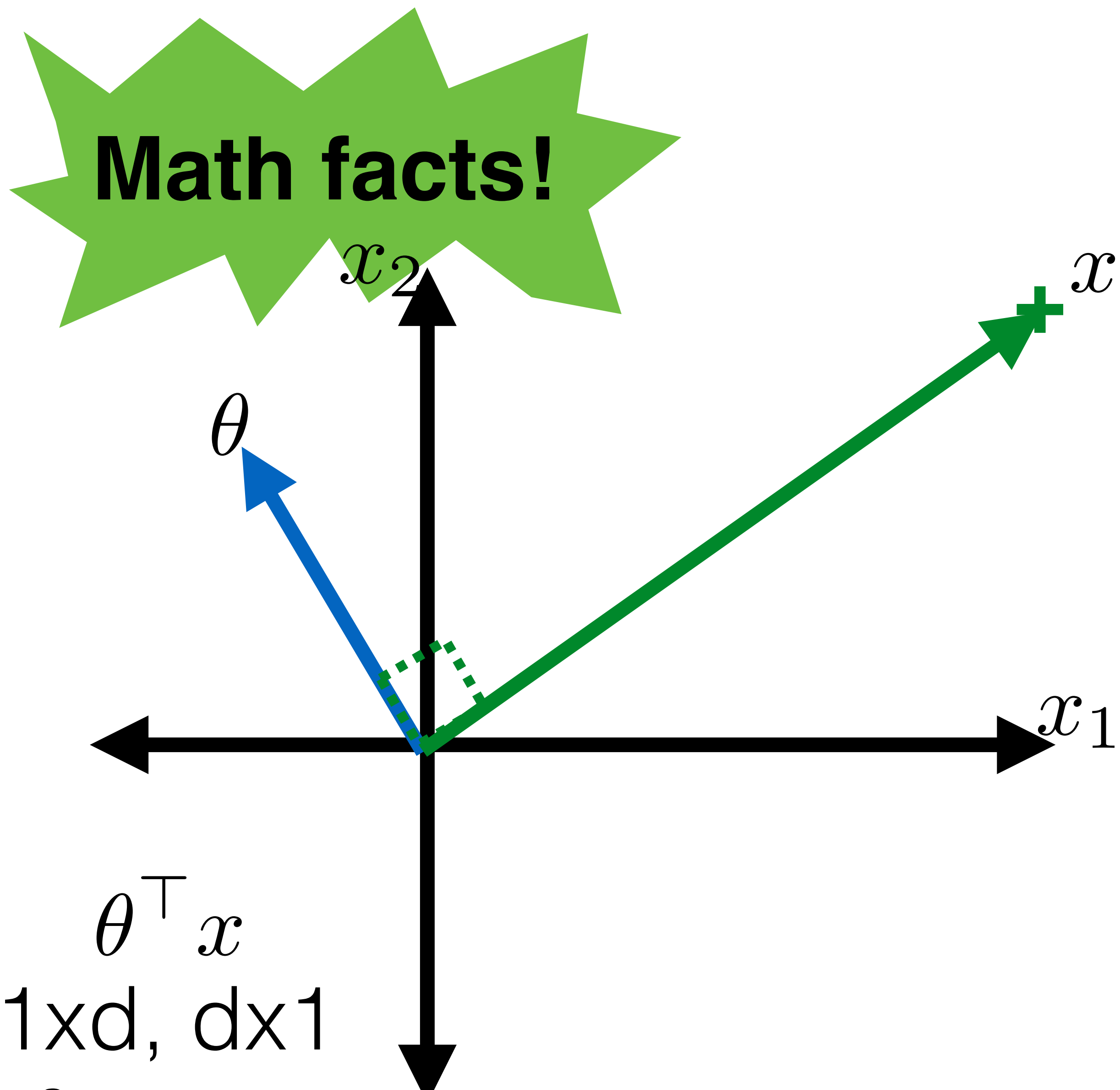
$\theta^\top x$

1xd, dx1

Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side





$\theta^\top x$

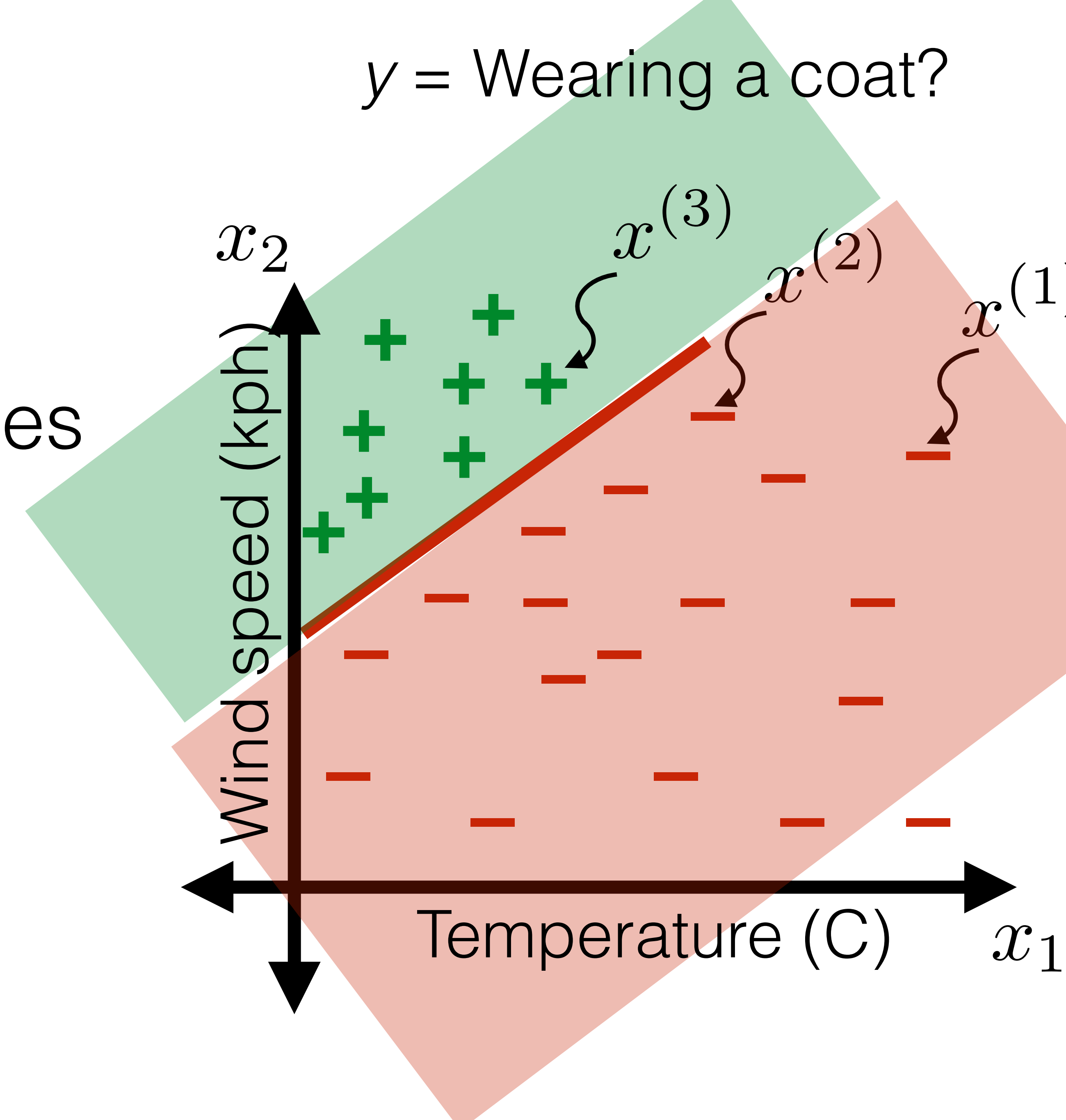
1xd, dx1

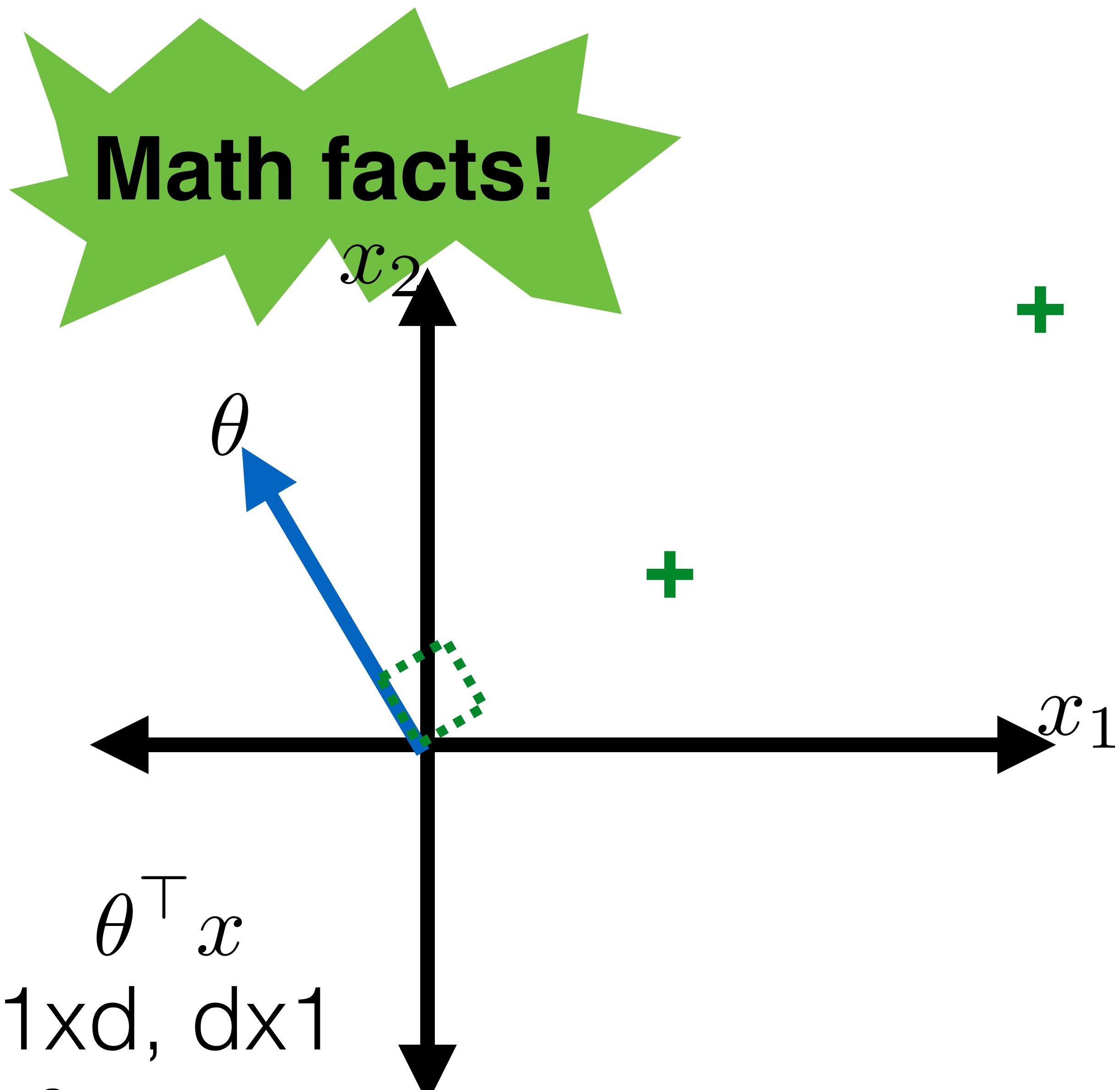
Linear classifiers

- Classification hypothesis:

$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$

- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



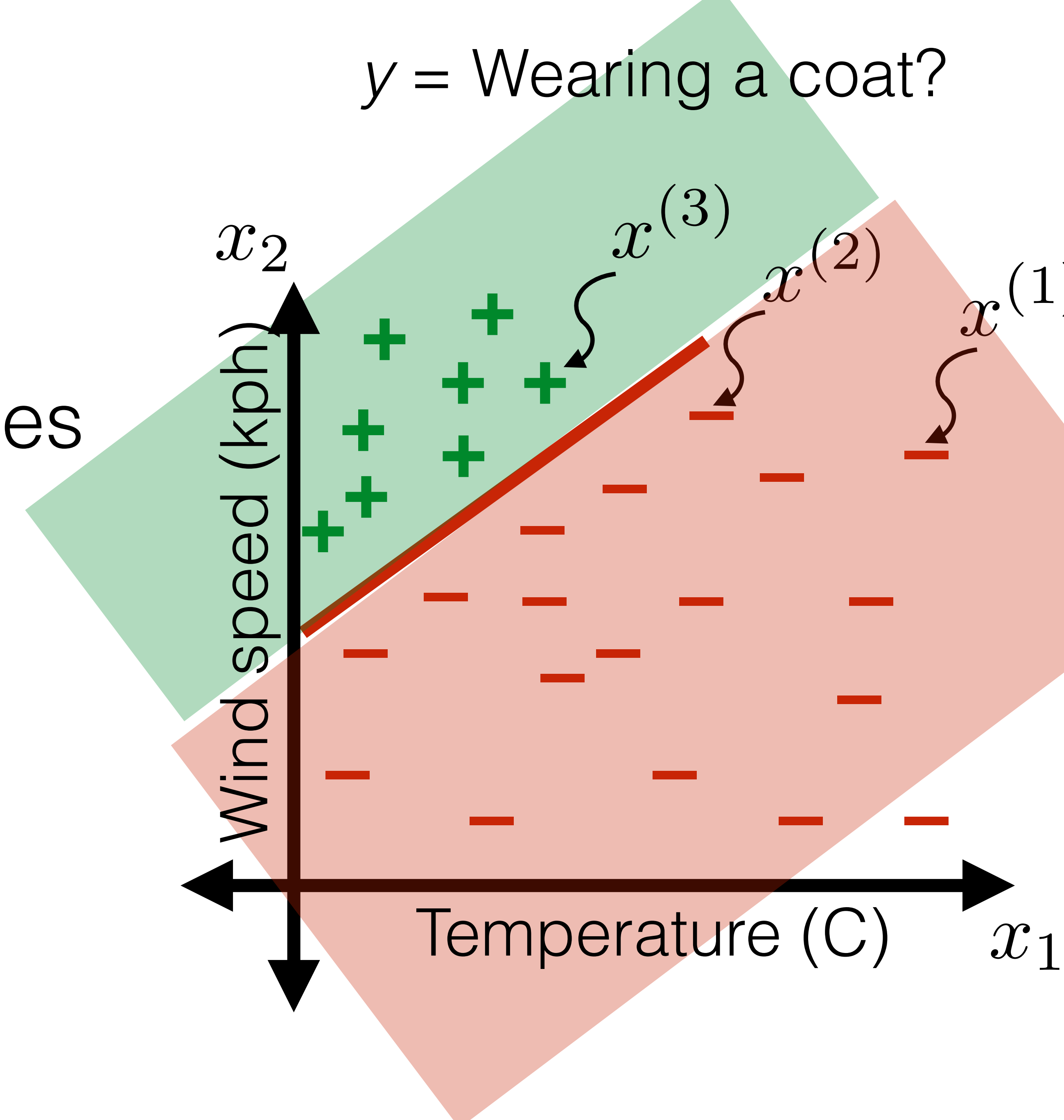


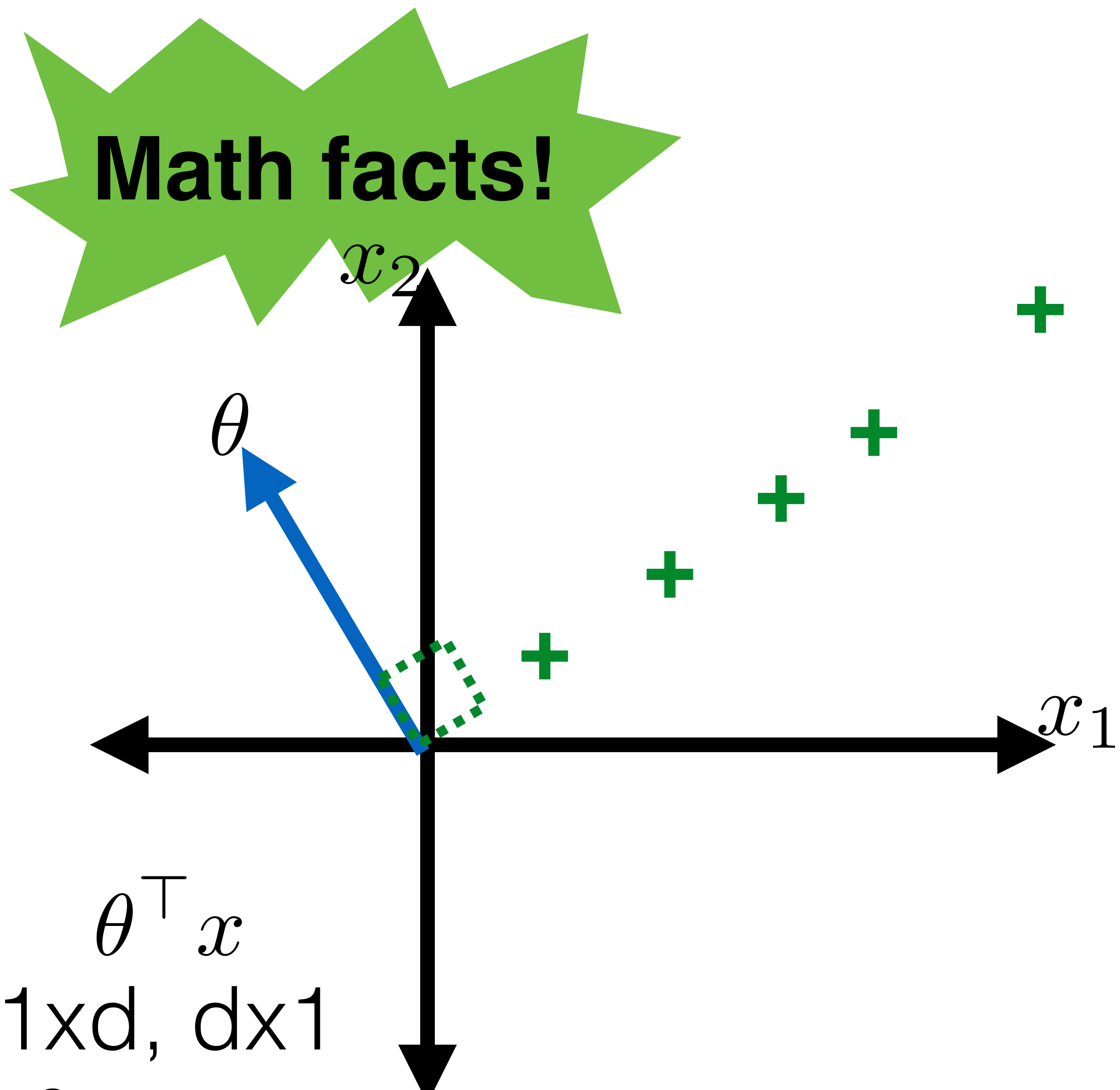
$\theta^\top x$

1xd, dx1

Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



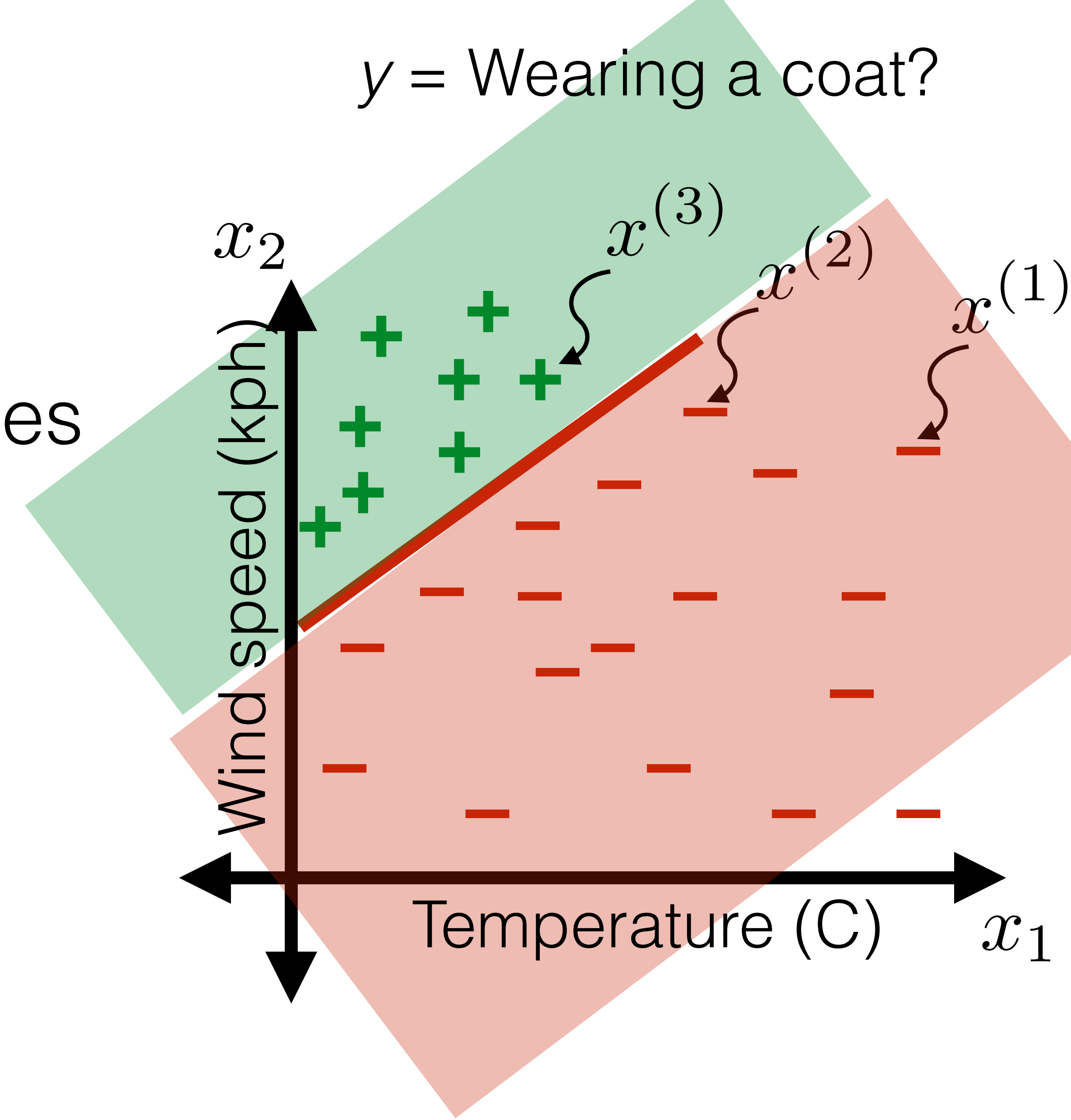


$\theta^\top x$

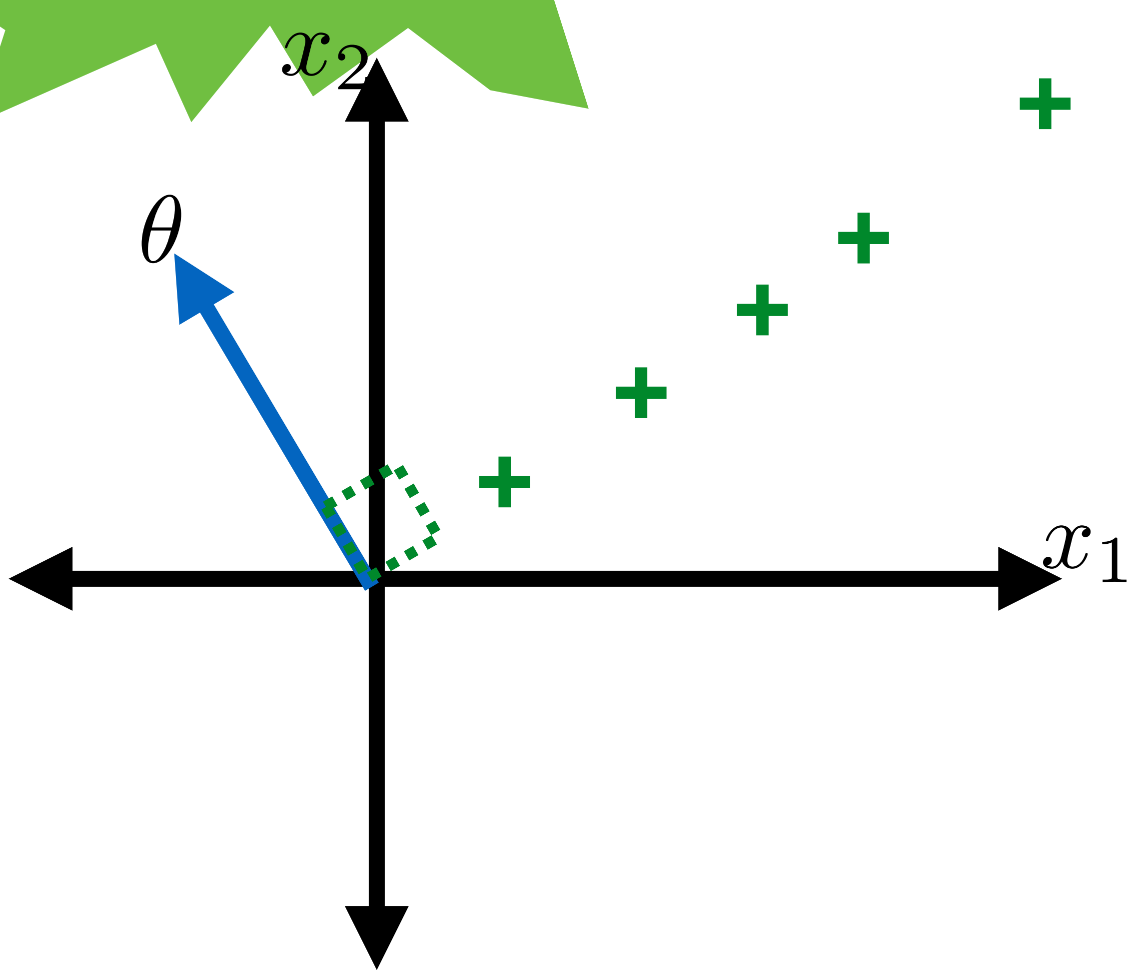
1xd, dx1

Linear classifiers

- Classification hypothesis:
  $$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

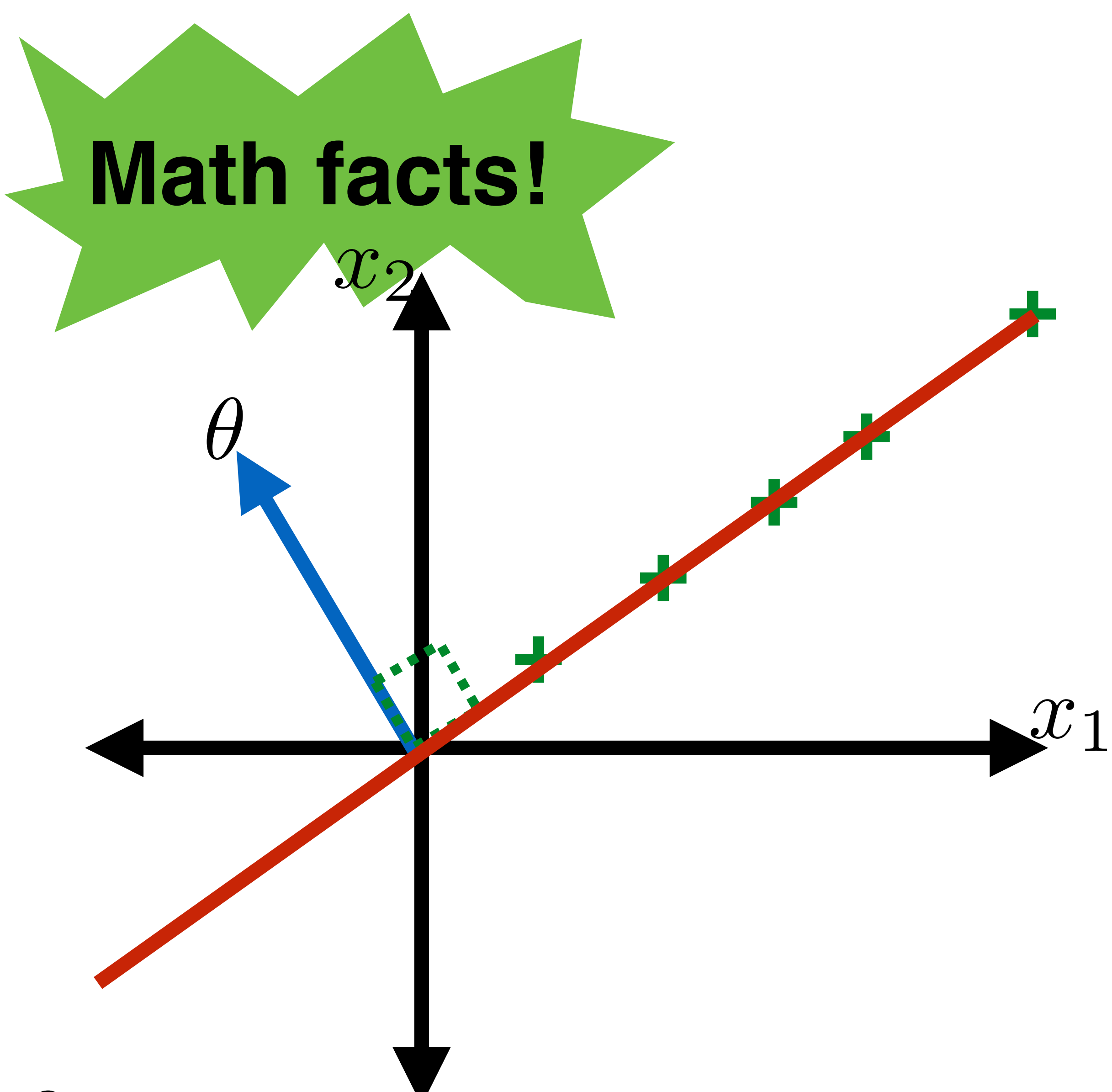


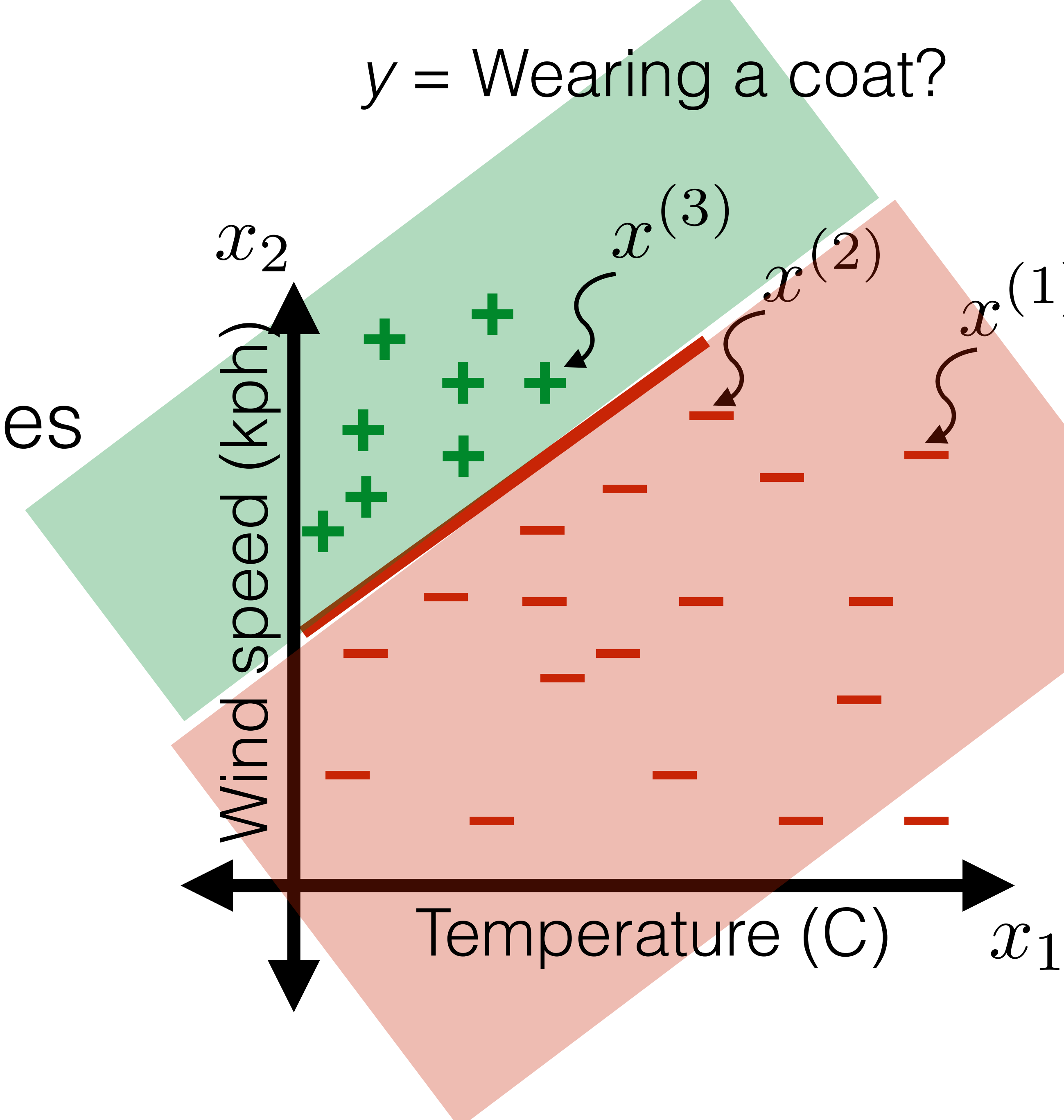
Math facts!



Linear classifiers

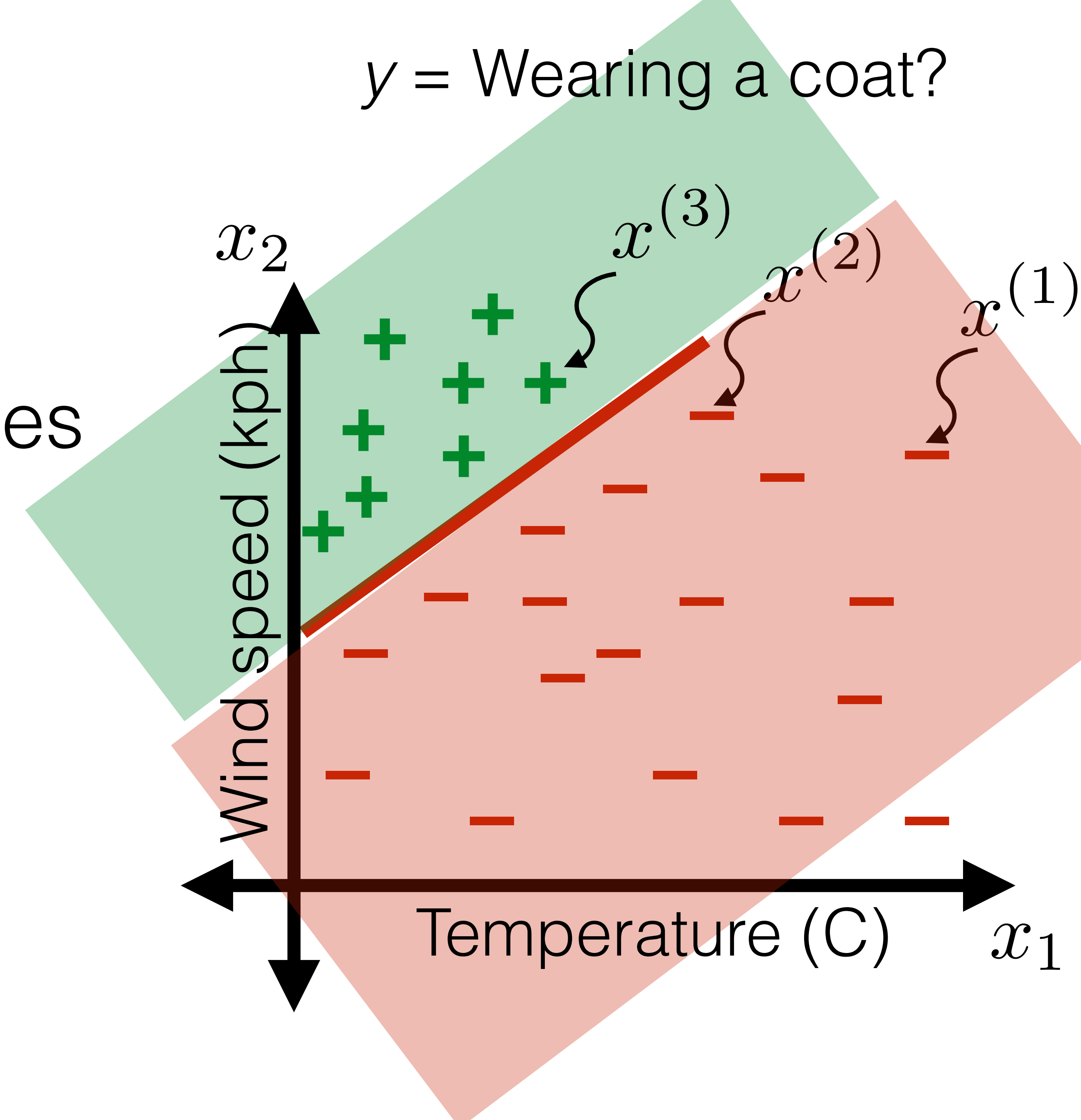
- Classification hypothesis:
$$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

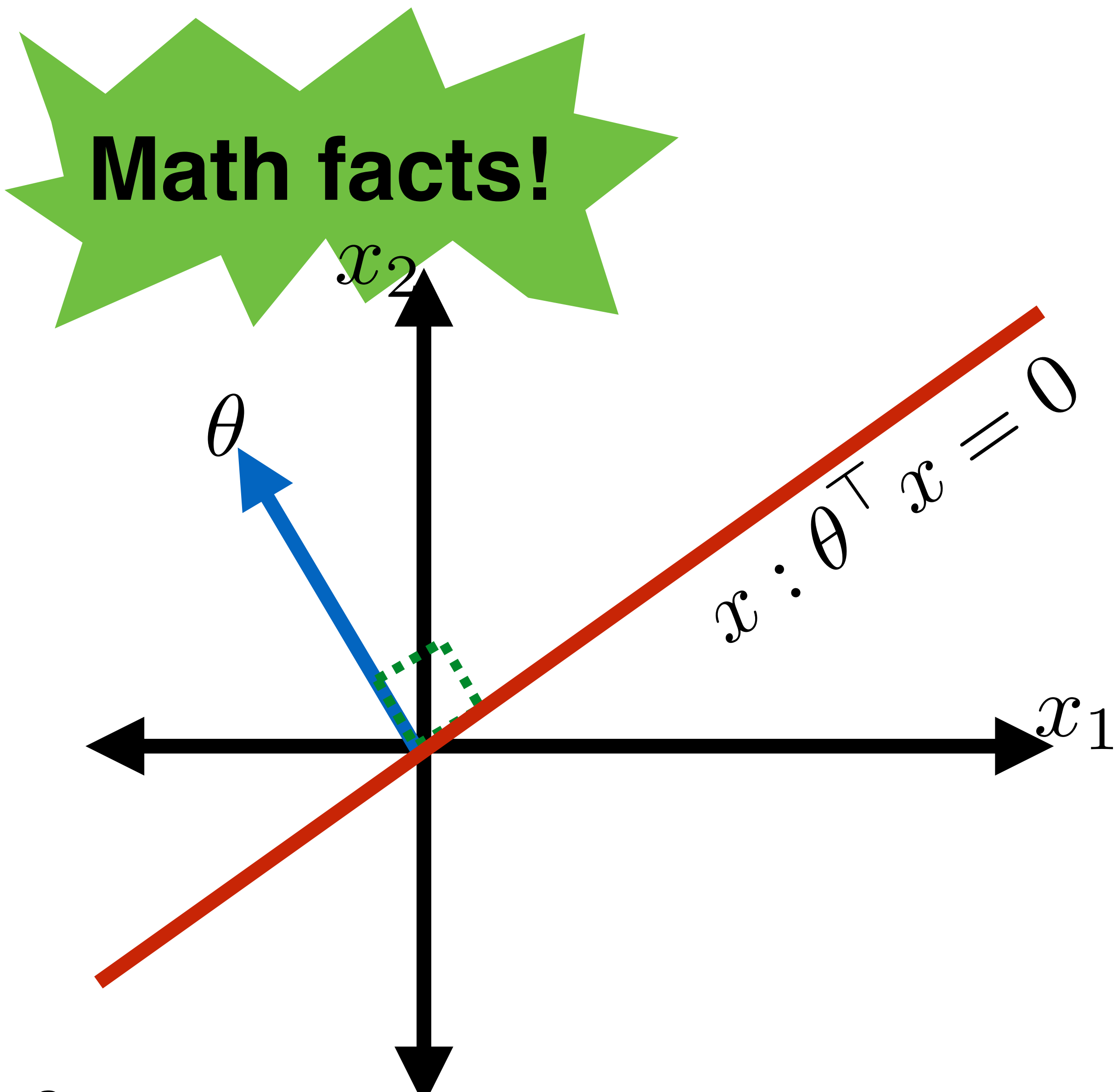




Linear classifiers

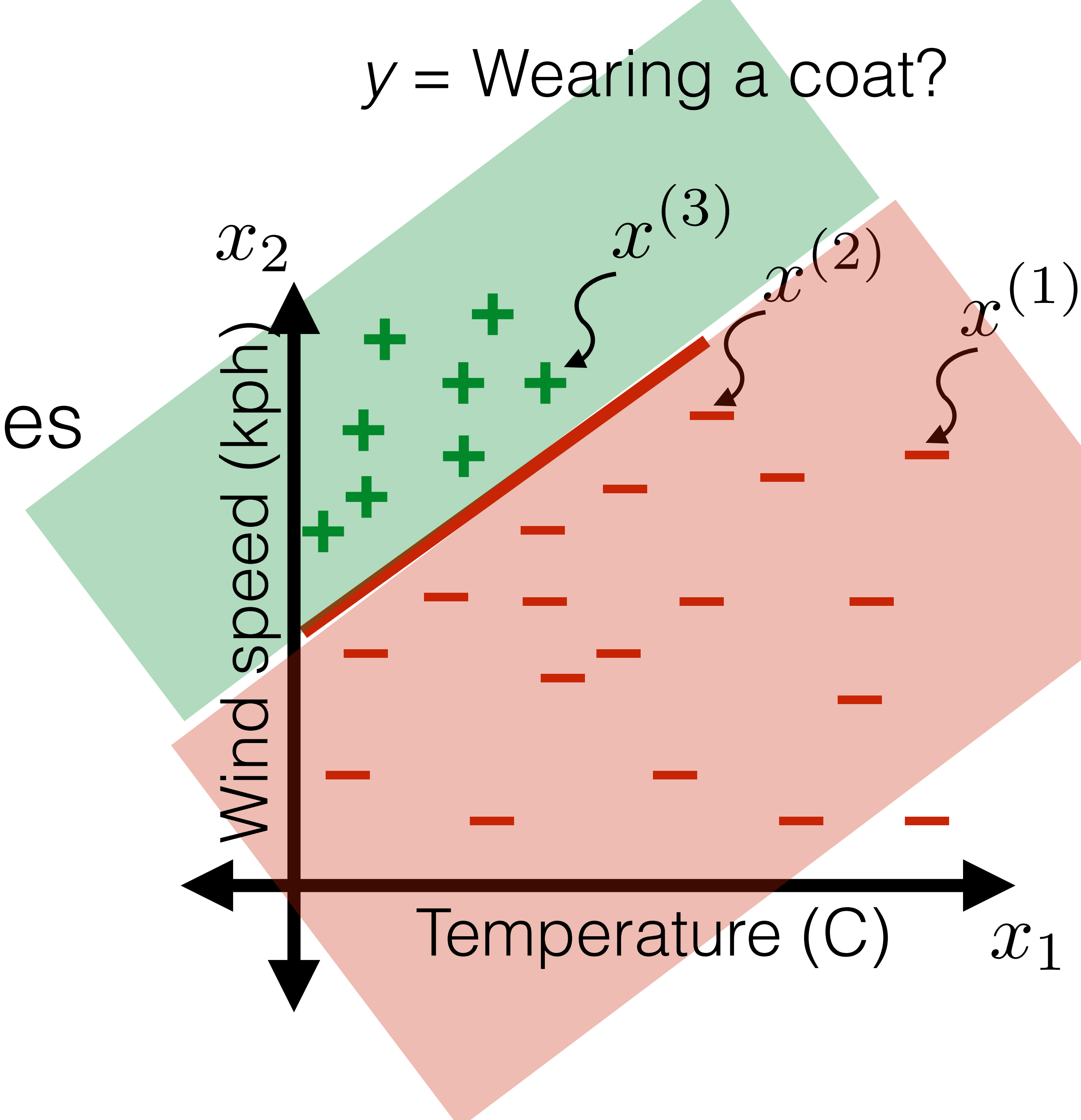
- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

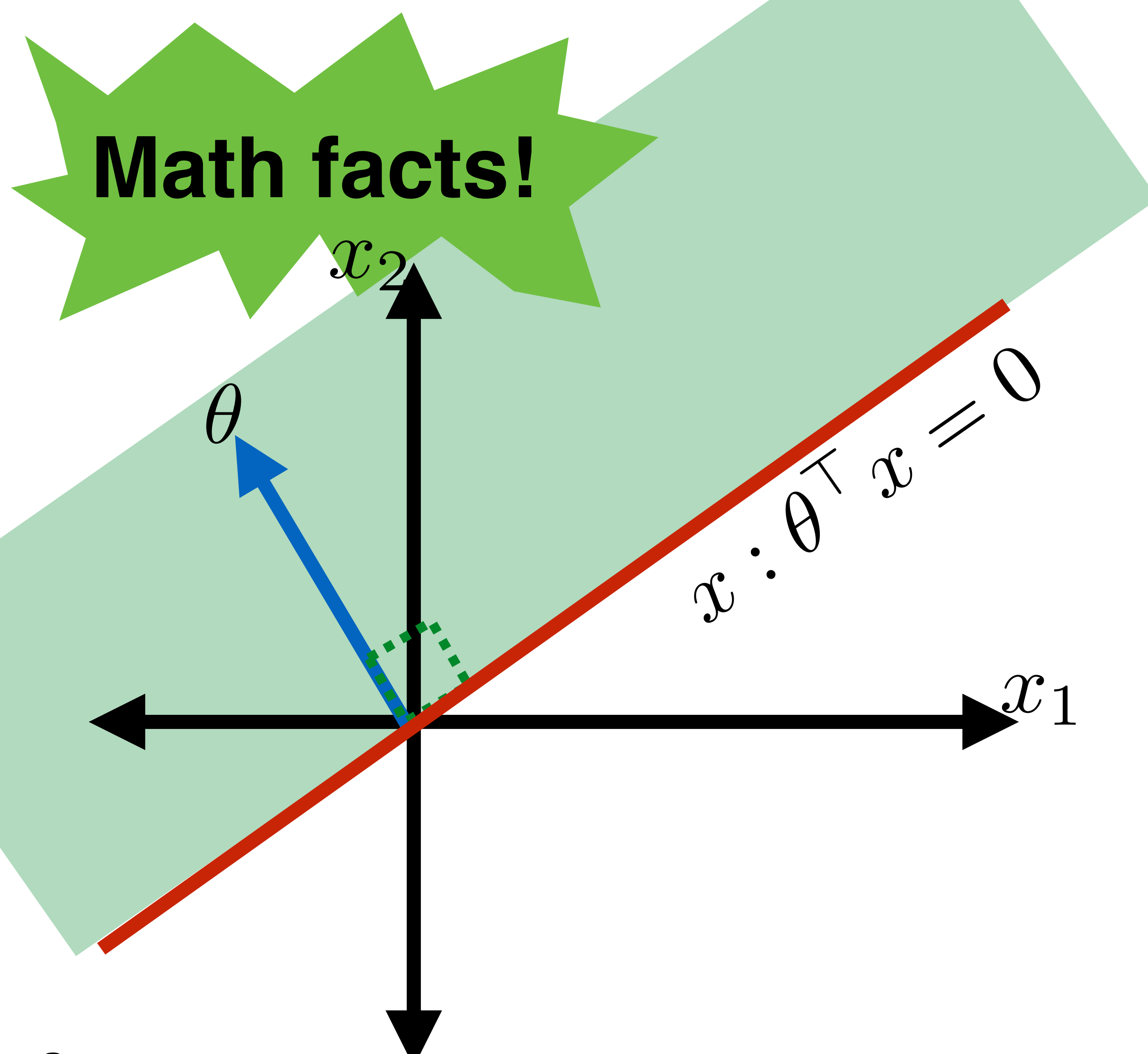




Linear classifiers

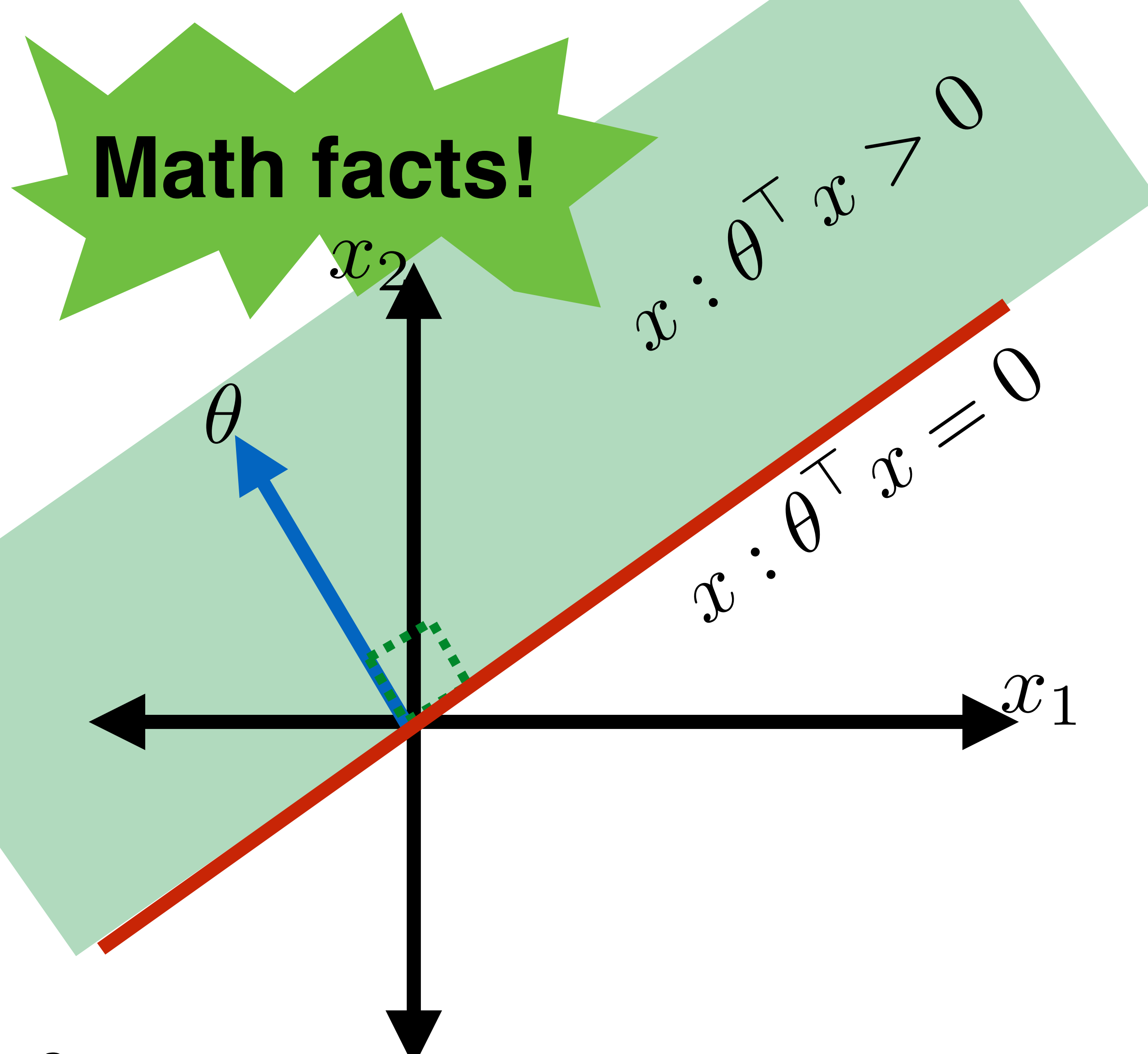
- Classification hypothesis:
  $$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

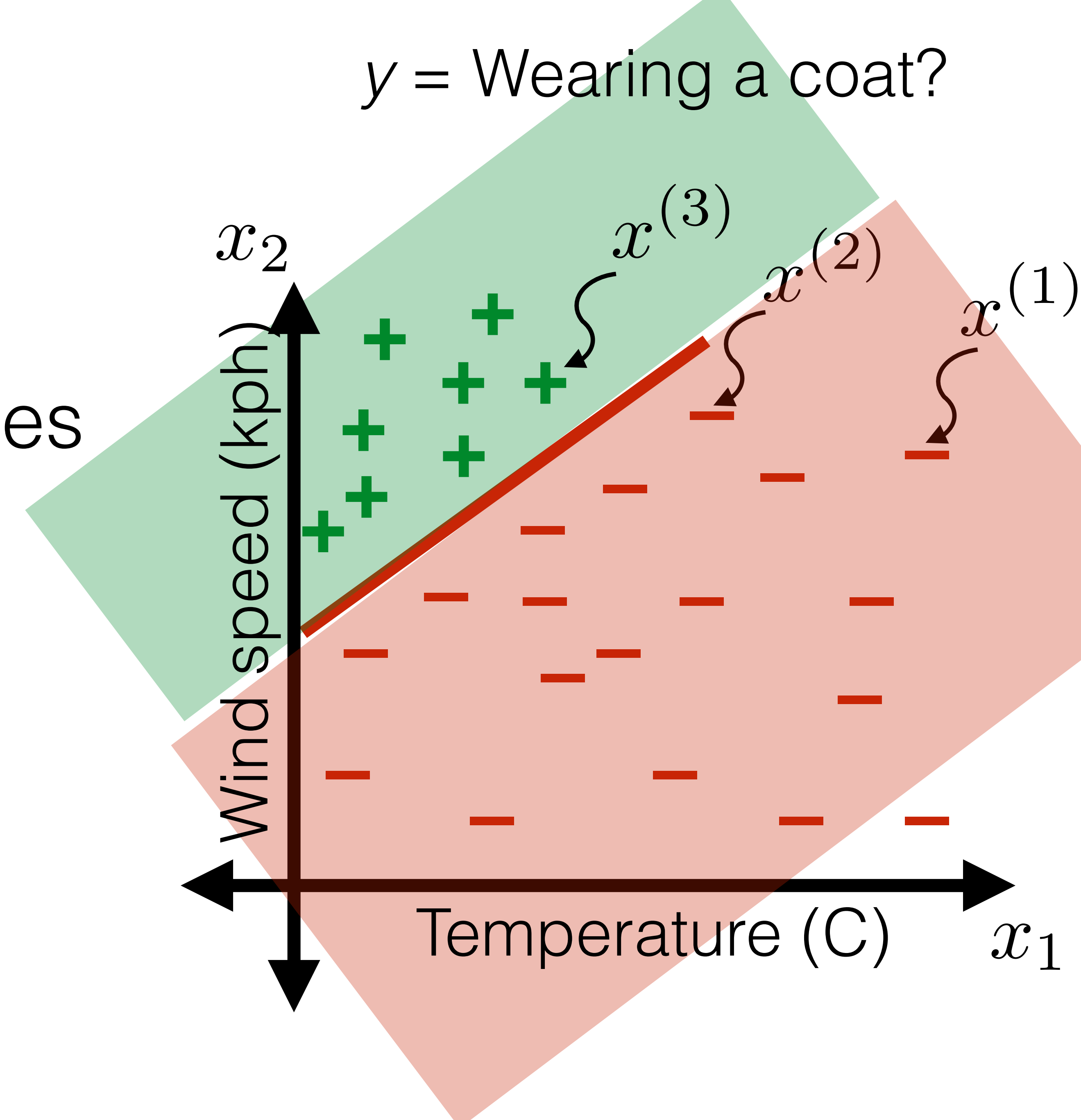




Linear classifiers

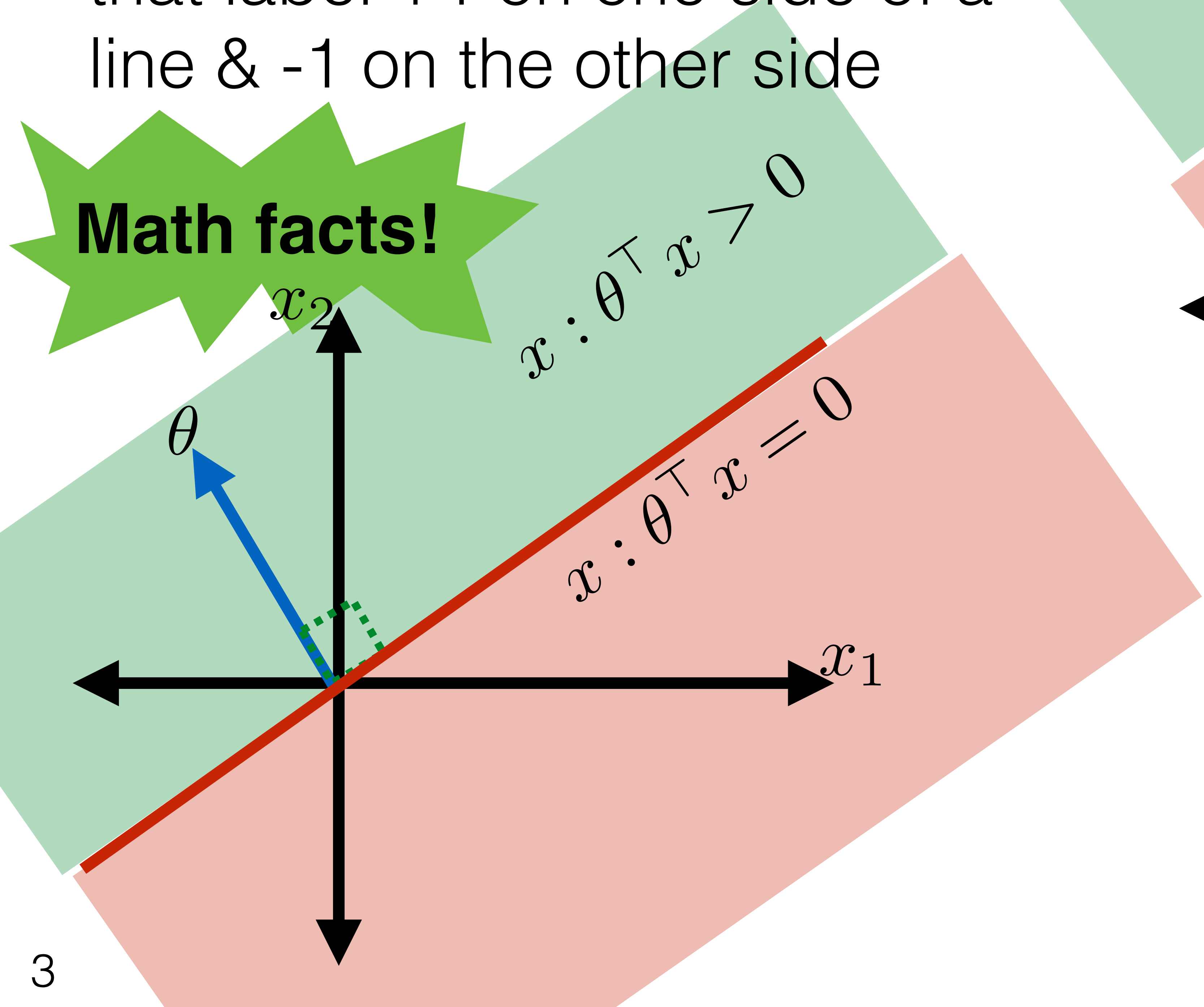
- Classification hypothesis:
$$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

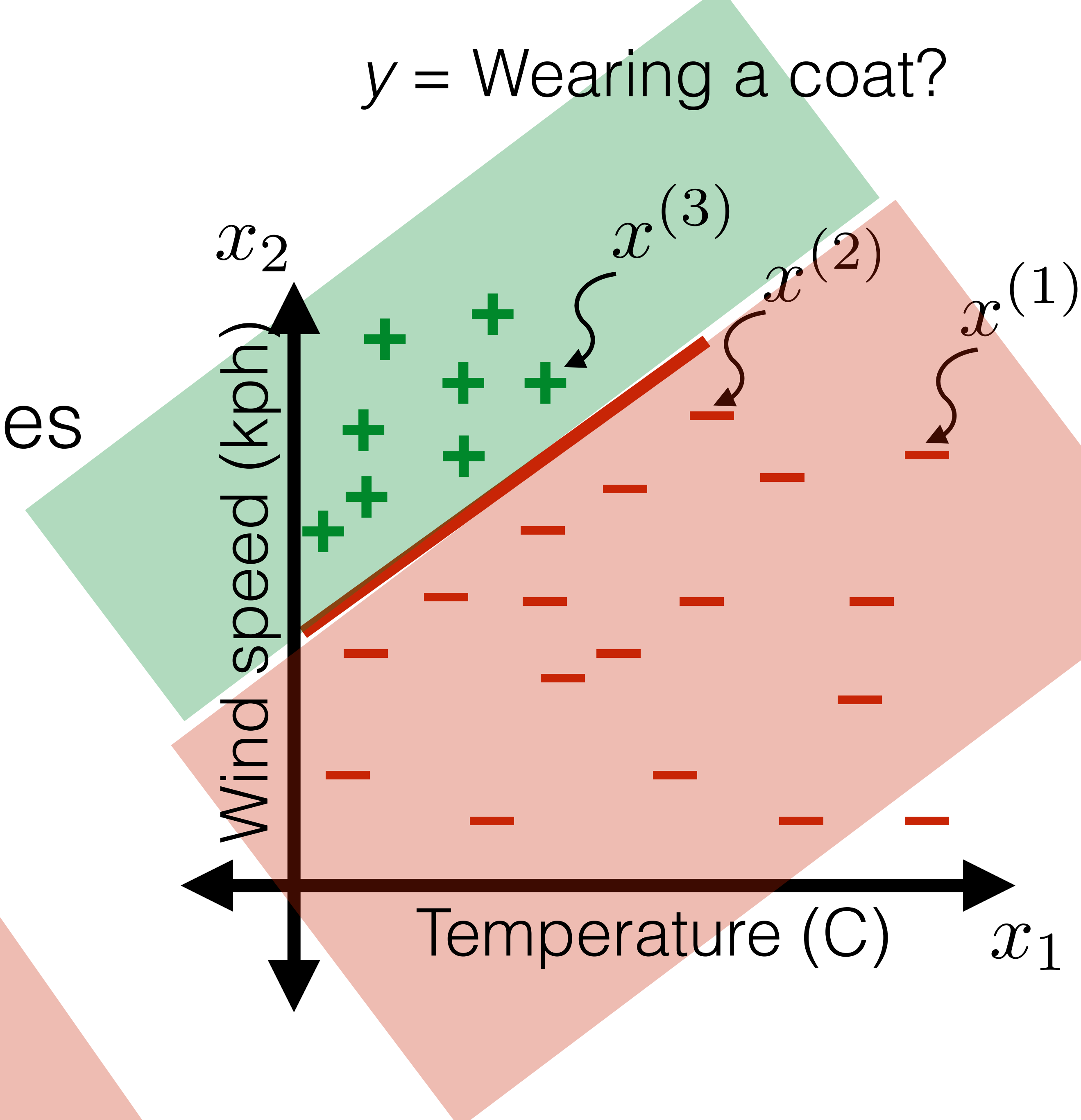




Linear classifiers

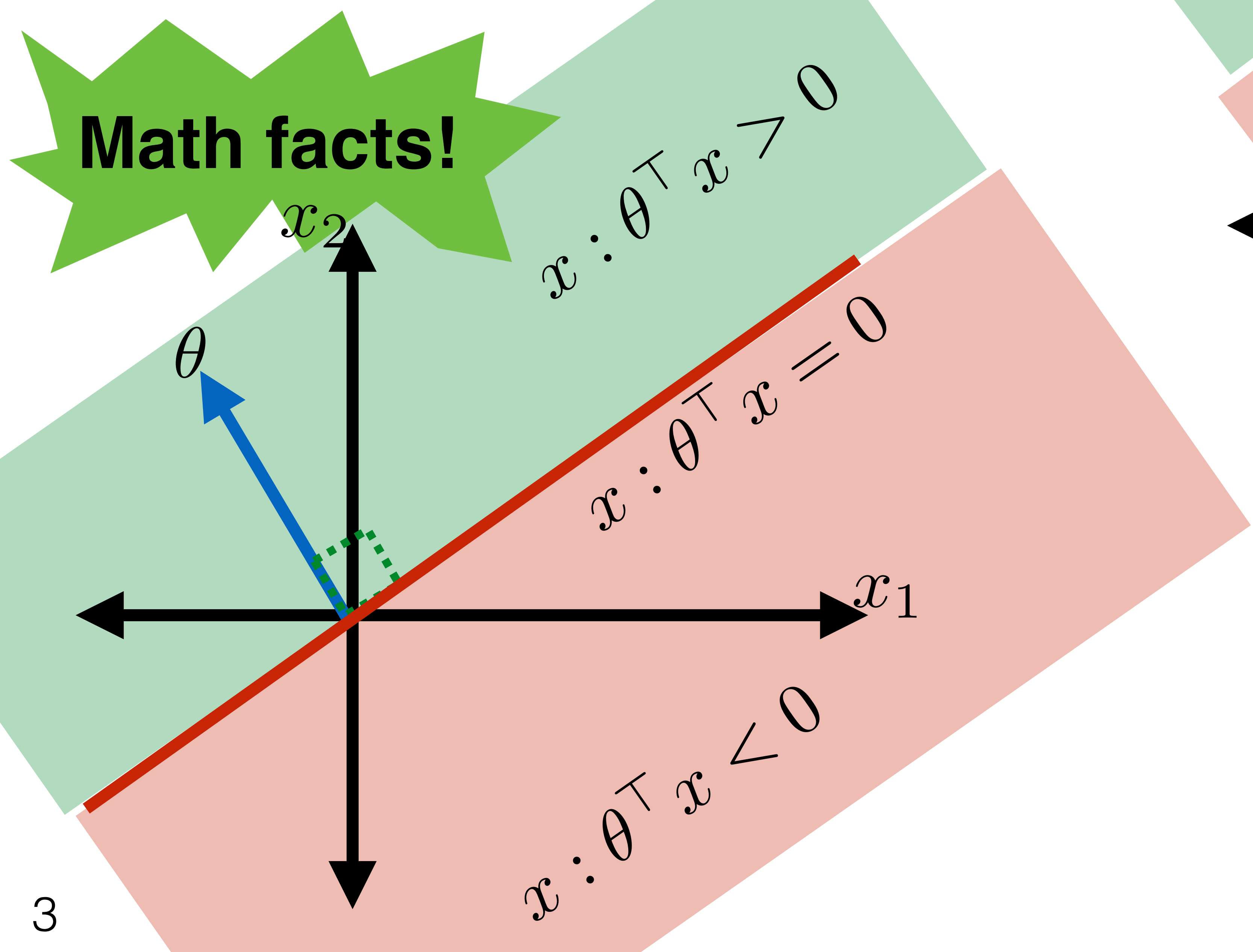
- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

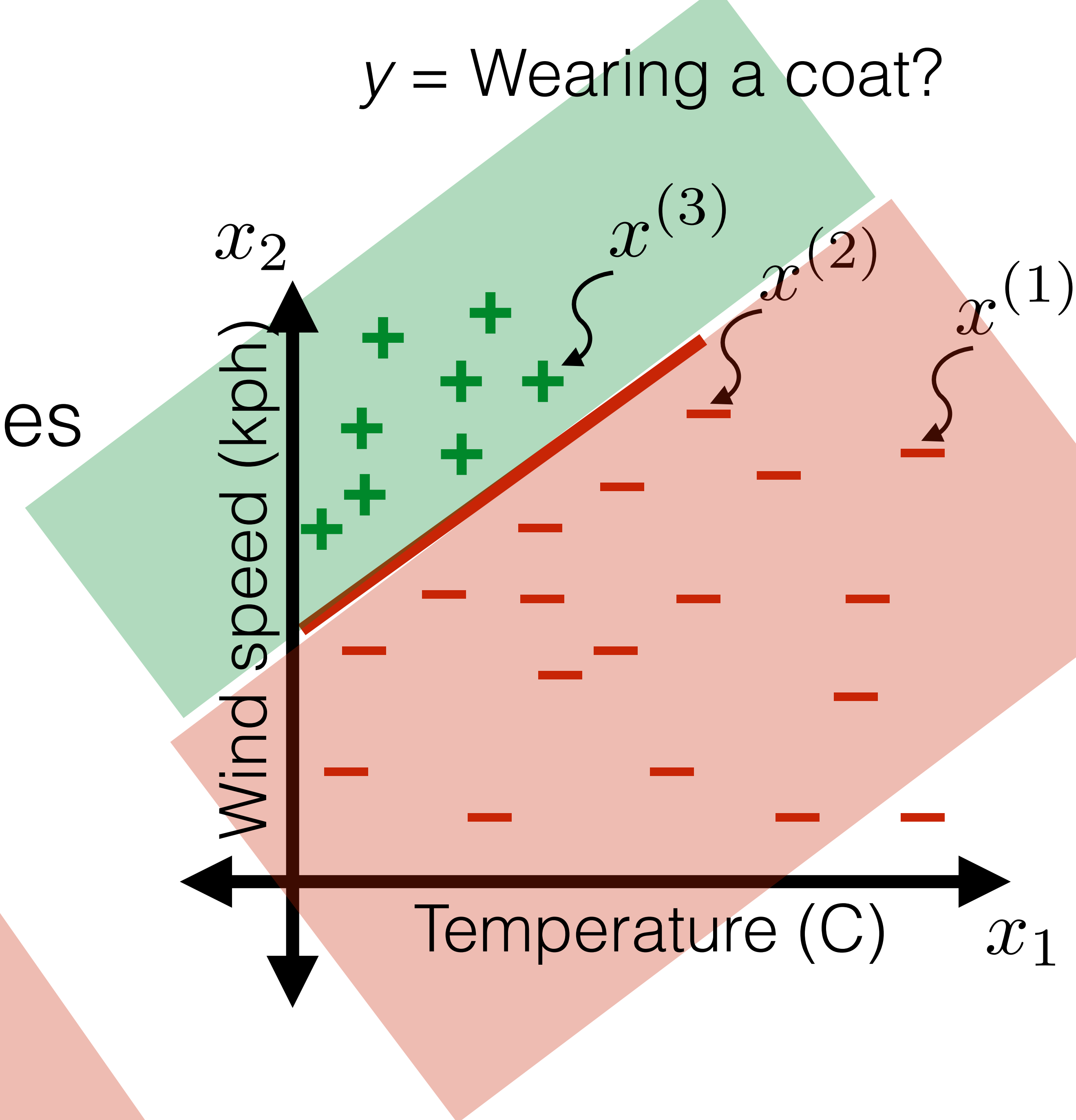




Linear classifiers

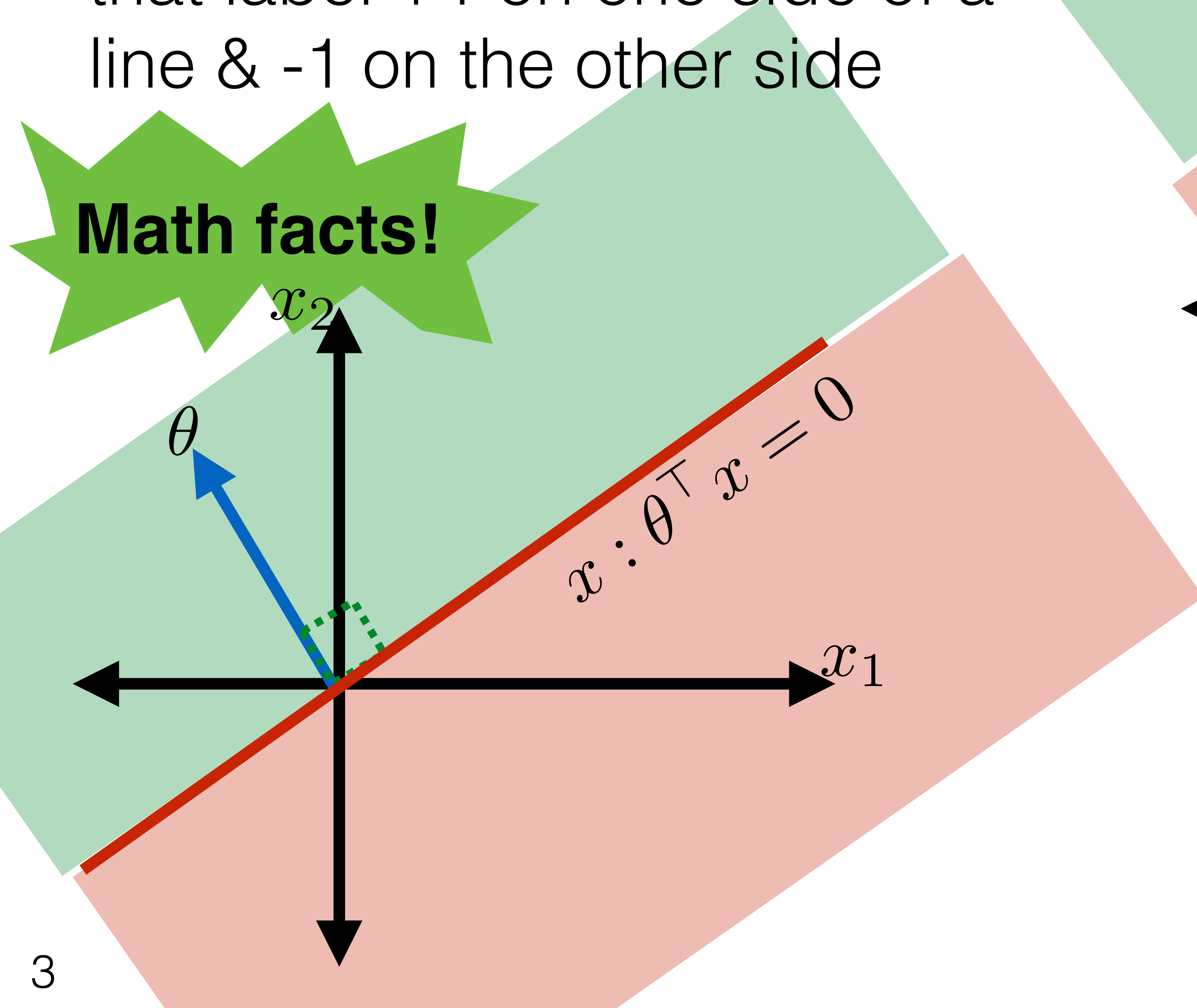
- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

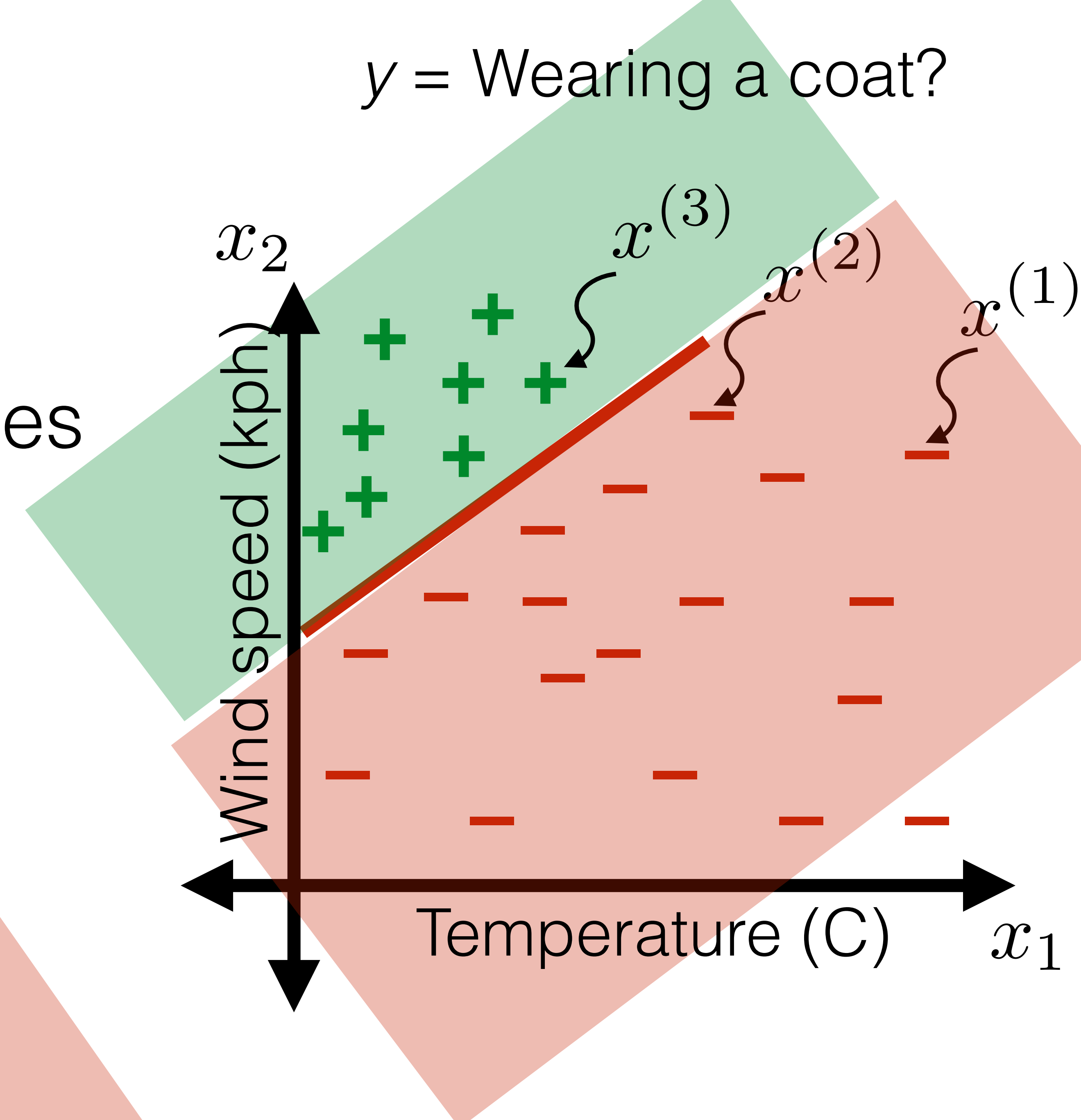




Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



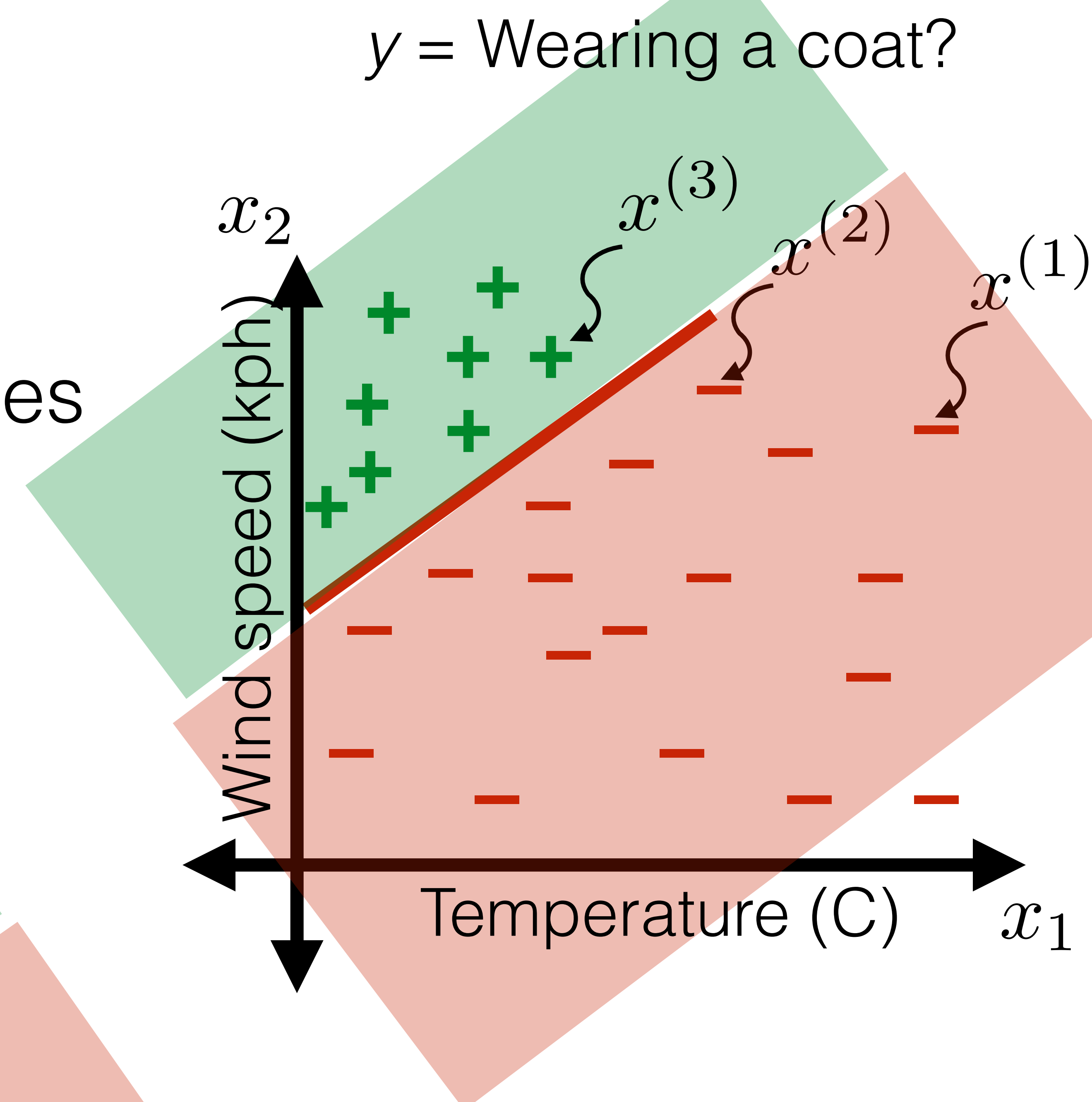


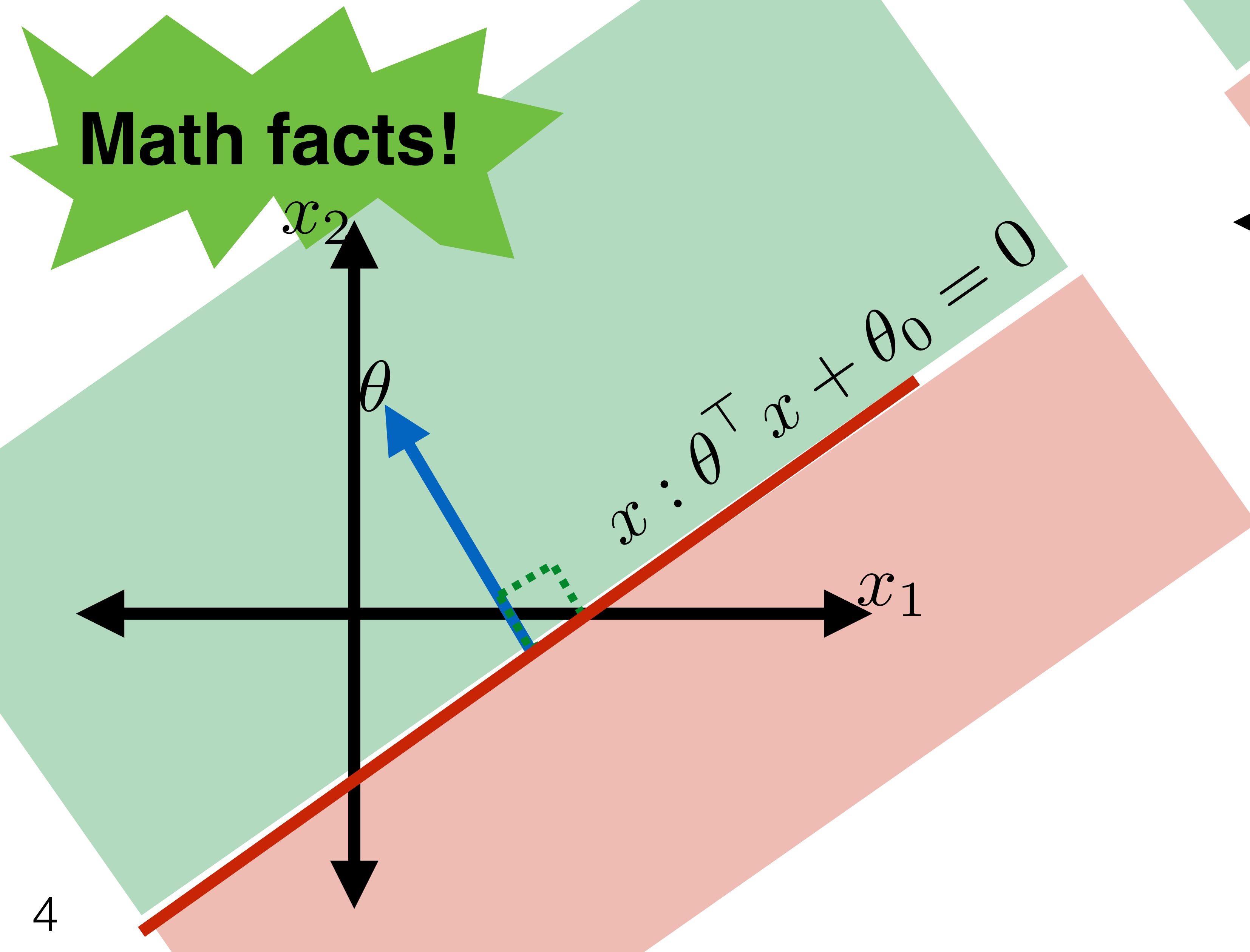
Linear classifiers

- Classification hypothesis:

$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$

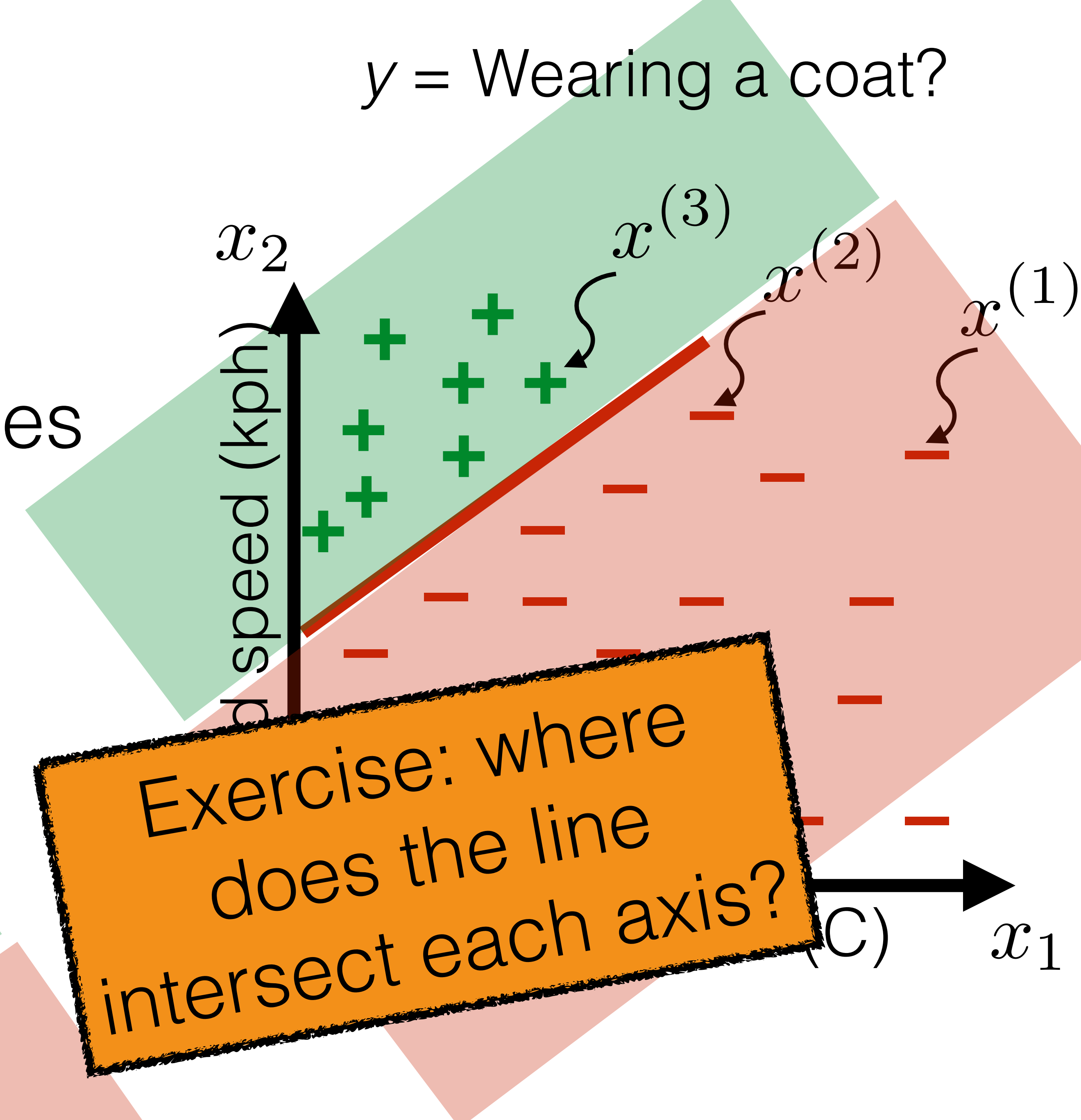
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

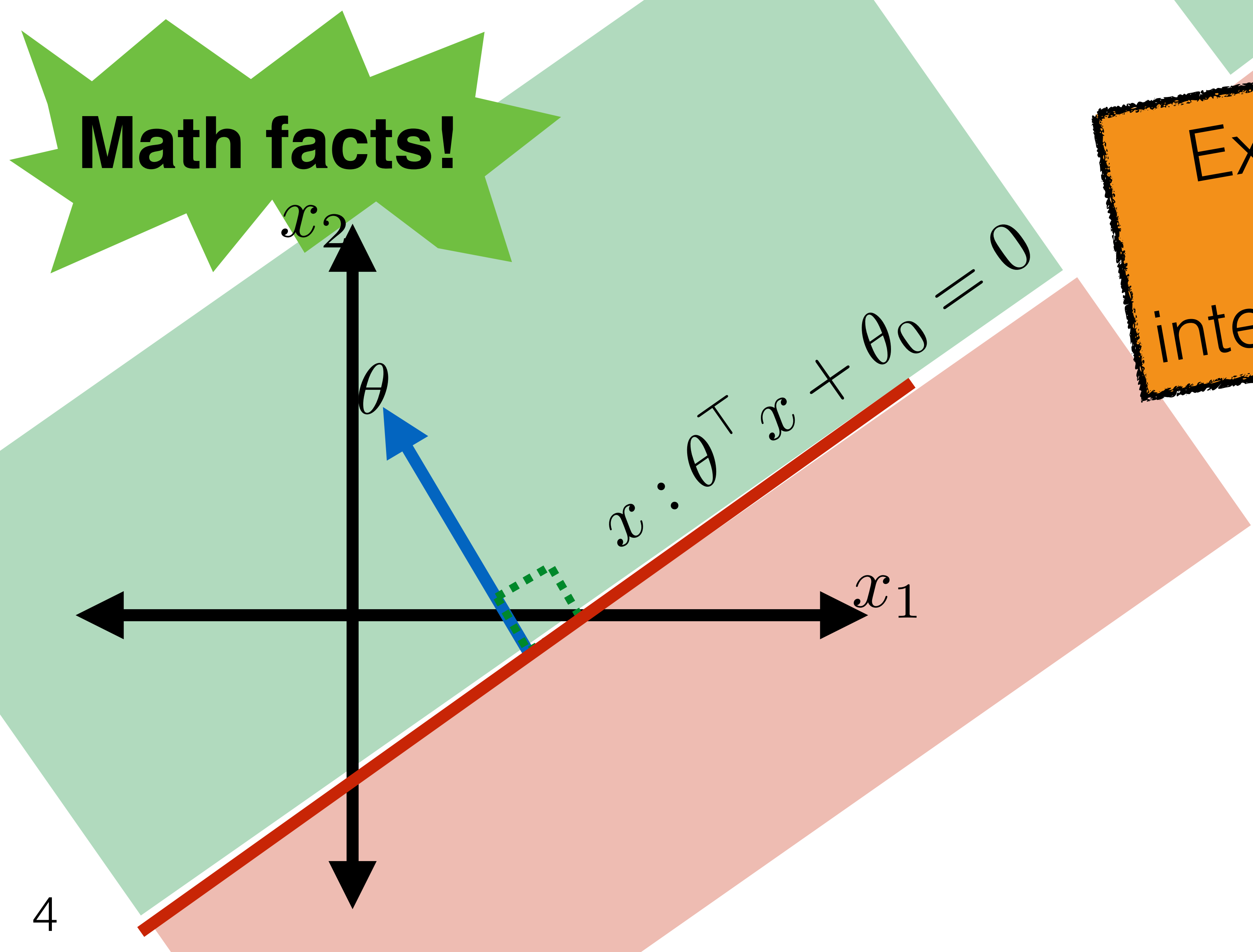




Linear classifiers

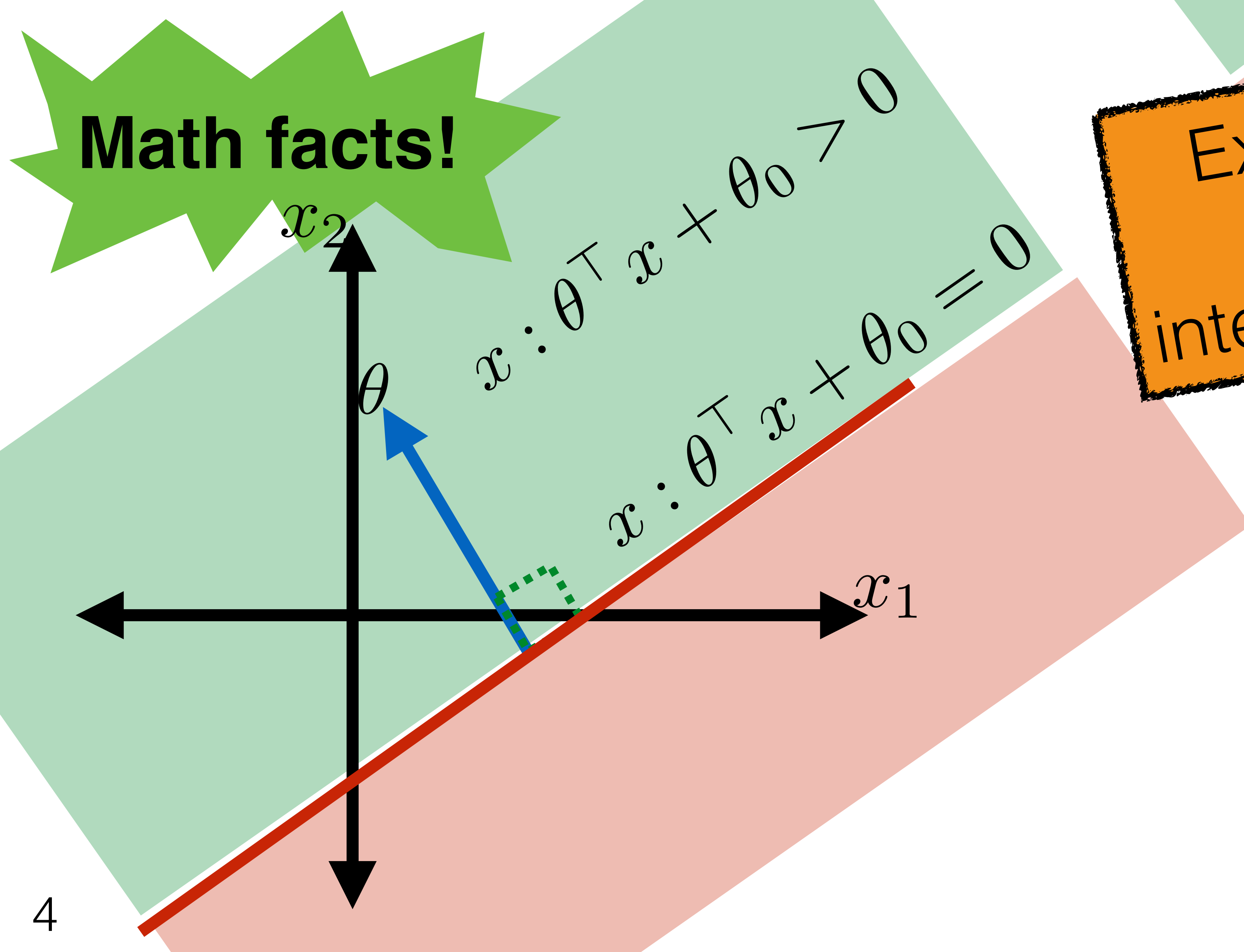
- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

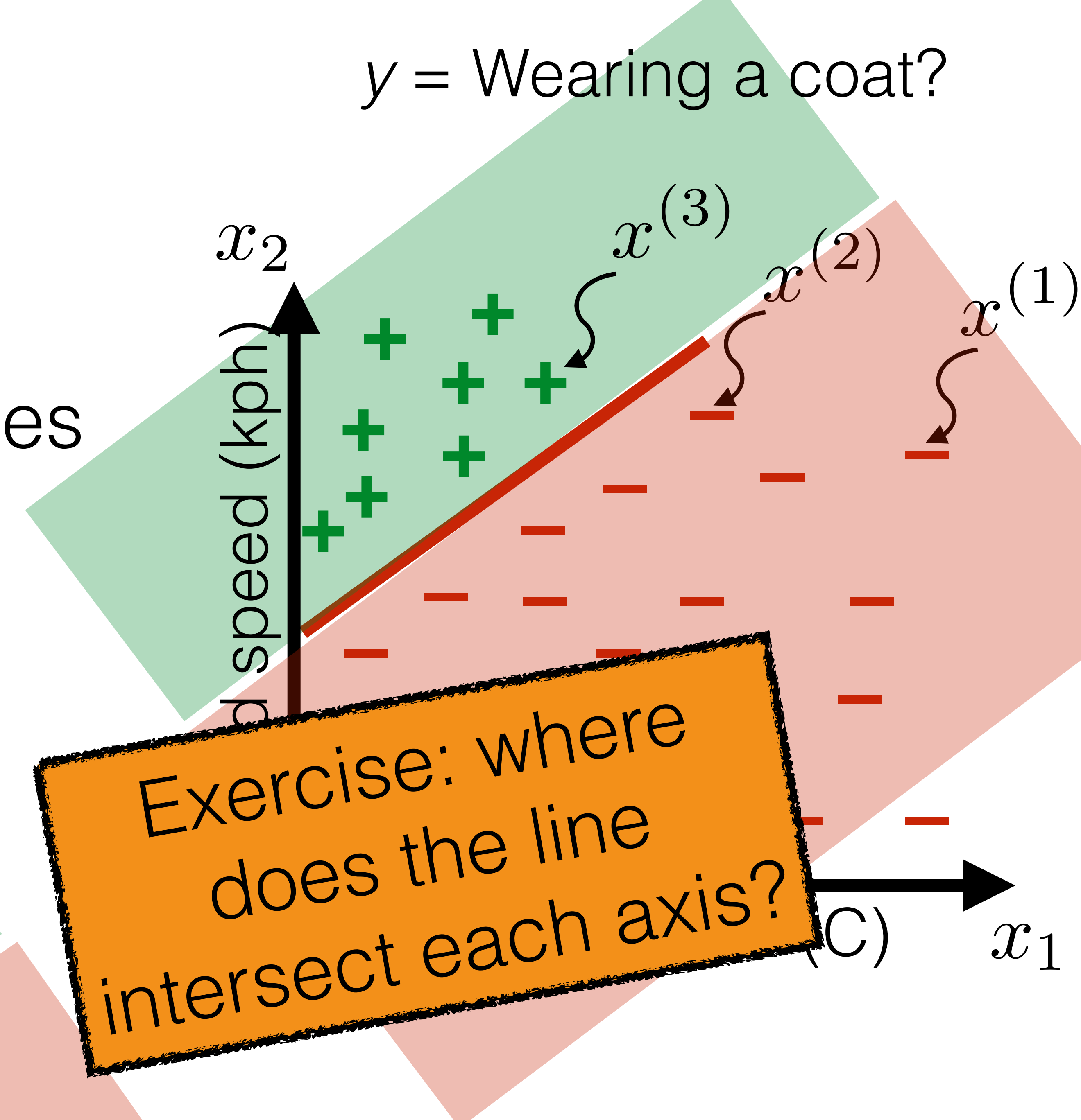




Linear classifiers

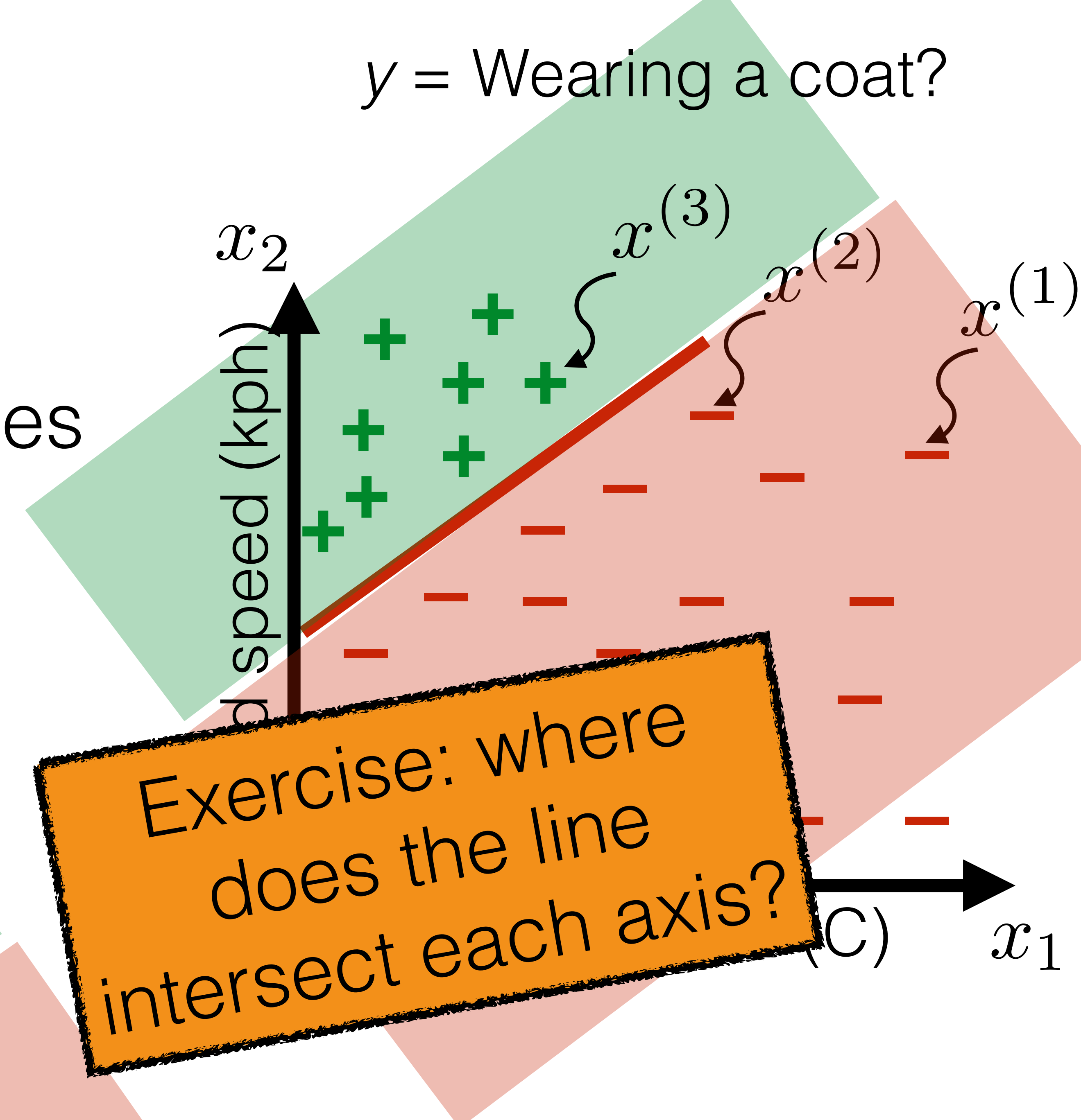
- Classification hypothesis:
  $$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

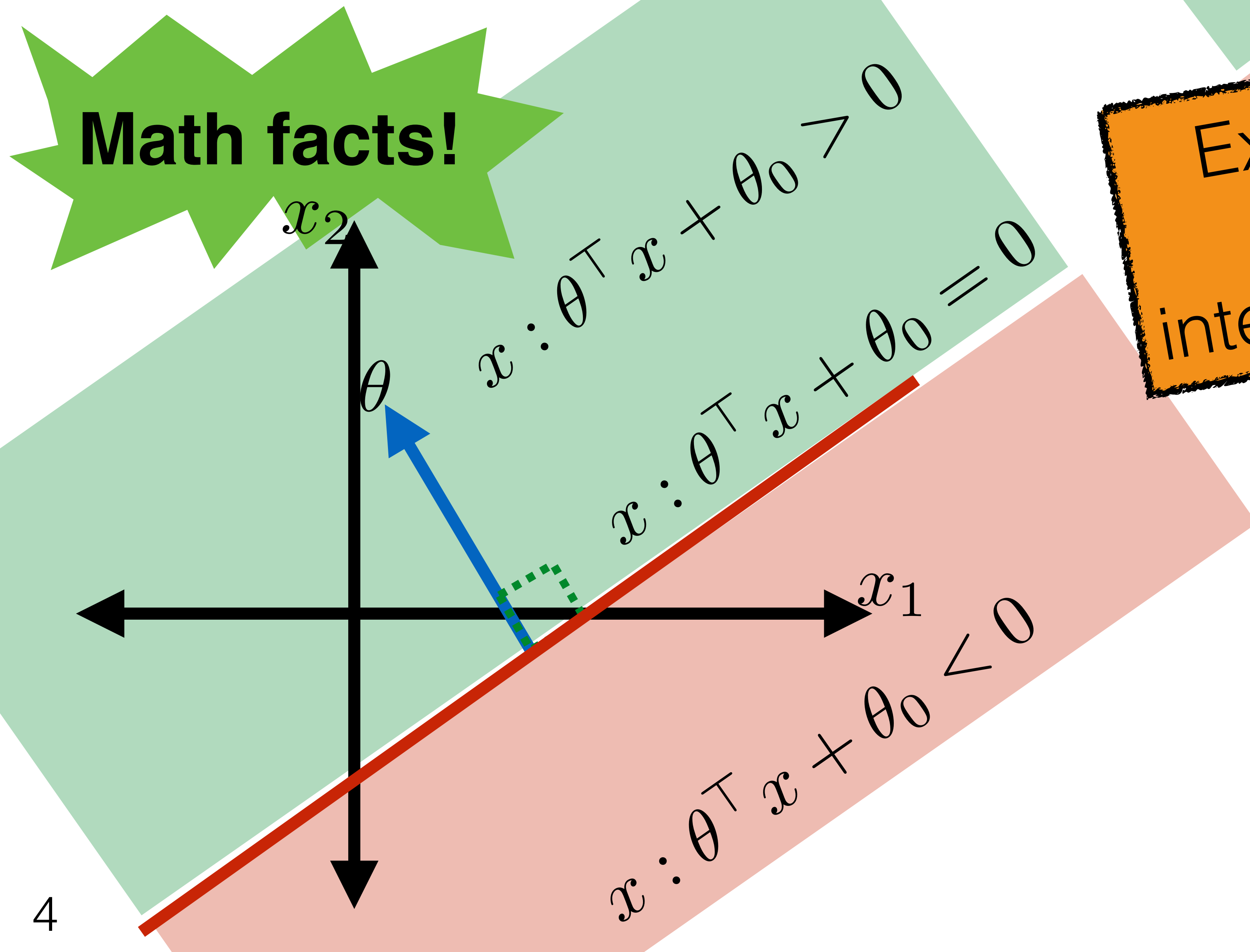




Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side





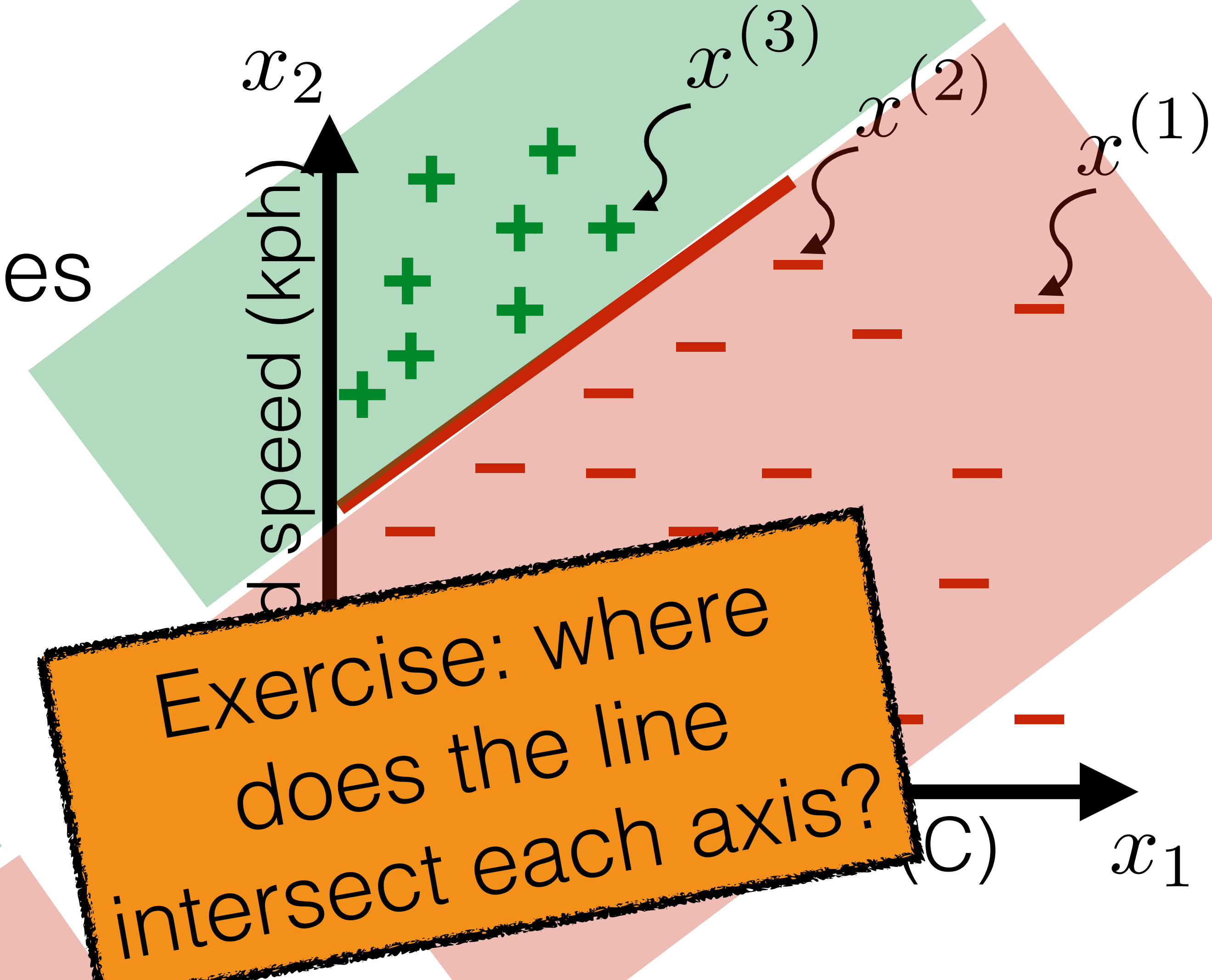
Linear classifiers

y = Wearing a coat?

- Classification hypothesis:

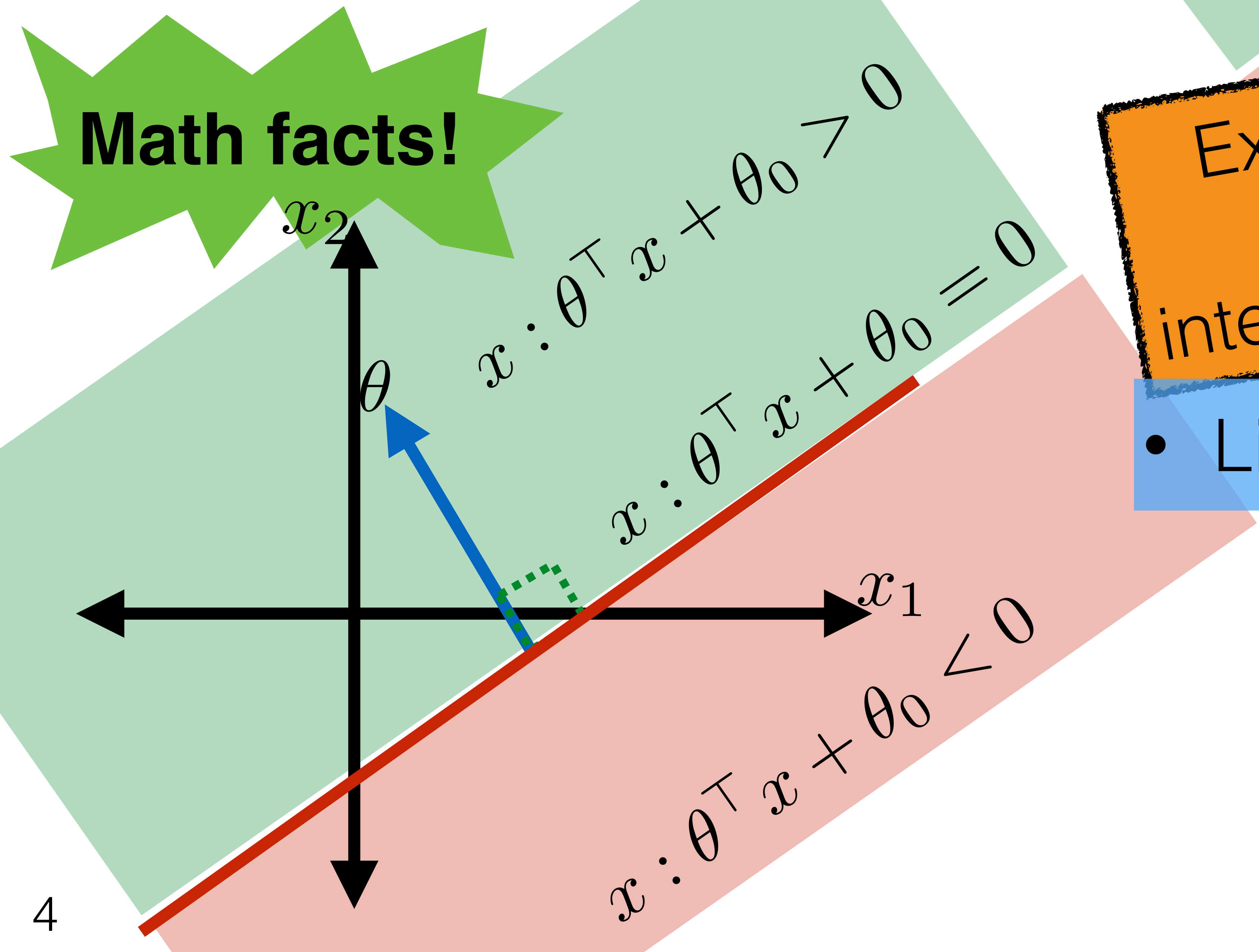
$$h : \mathbb{R}^d \to \{-1, +1\}$$

- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



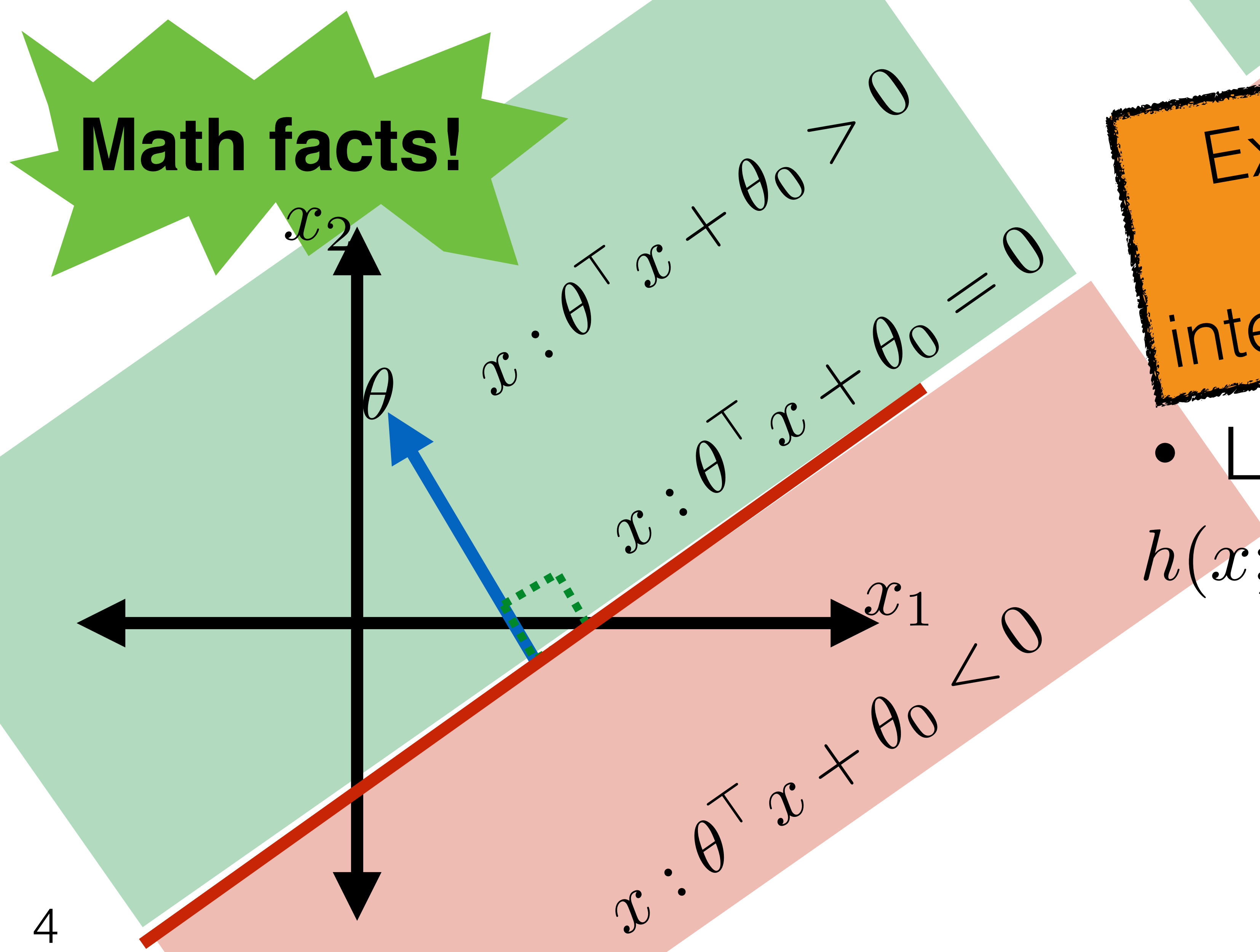
Exercise: where does the line intersect each axis?

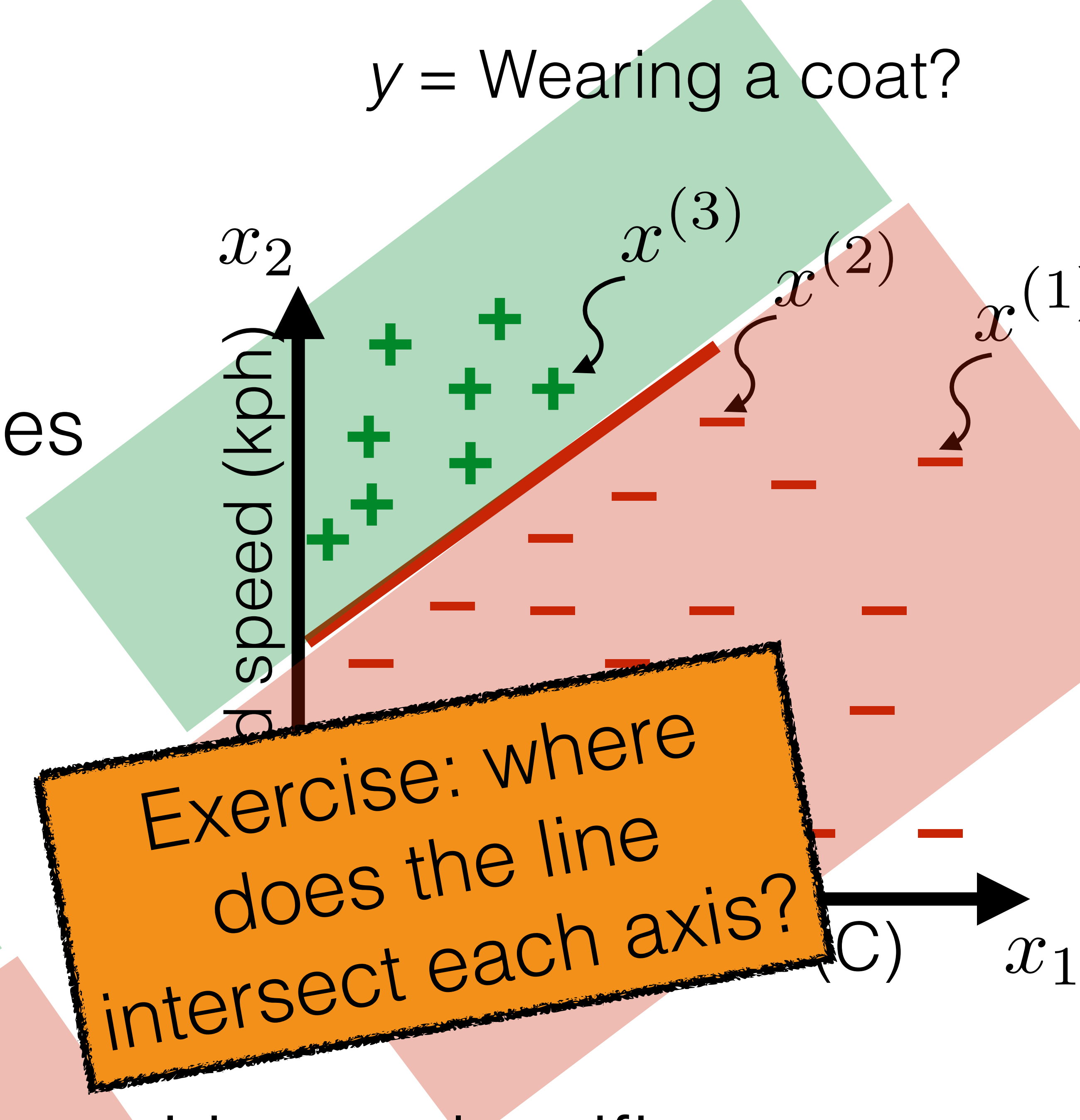
- Linear classifier:



Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side





- Linear classifier:
$$h(x; \theta, \theta_0) = \text{sign}(\theta^\top x + \theta_0)$$

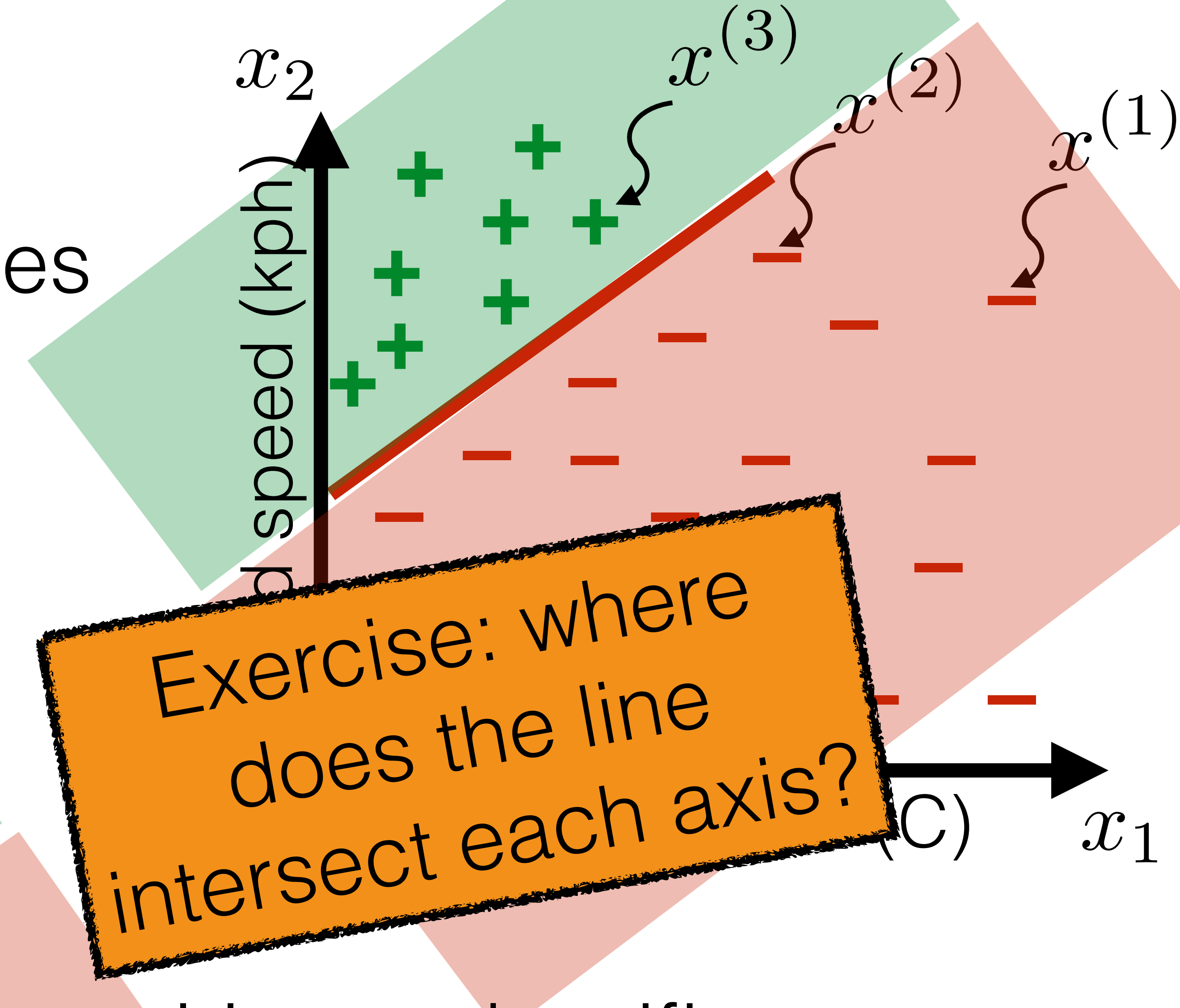
Linear classifiers

- Classification hypothesis:

$$h : \mathbb{R}^d \to \{-1, +1\}$$

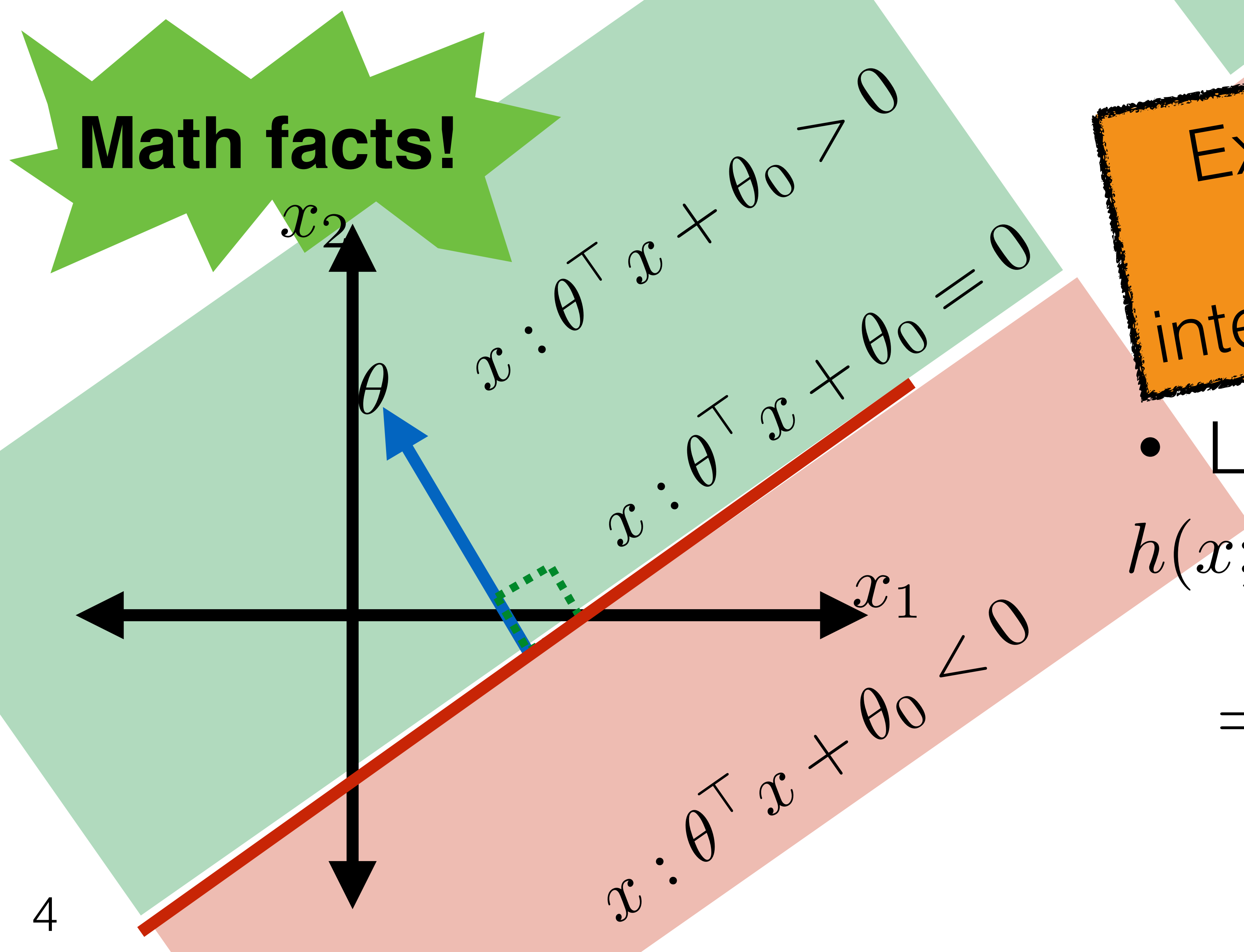
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

y = Wearing a coat?



Exercise: where does the line intersect each axis?

Math facts!

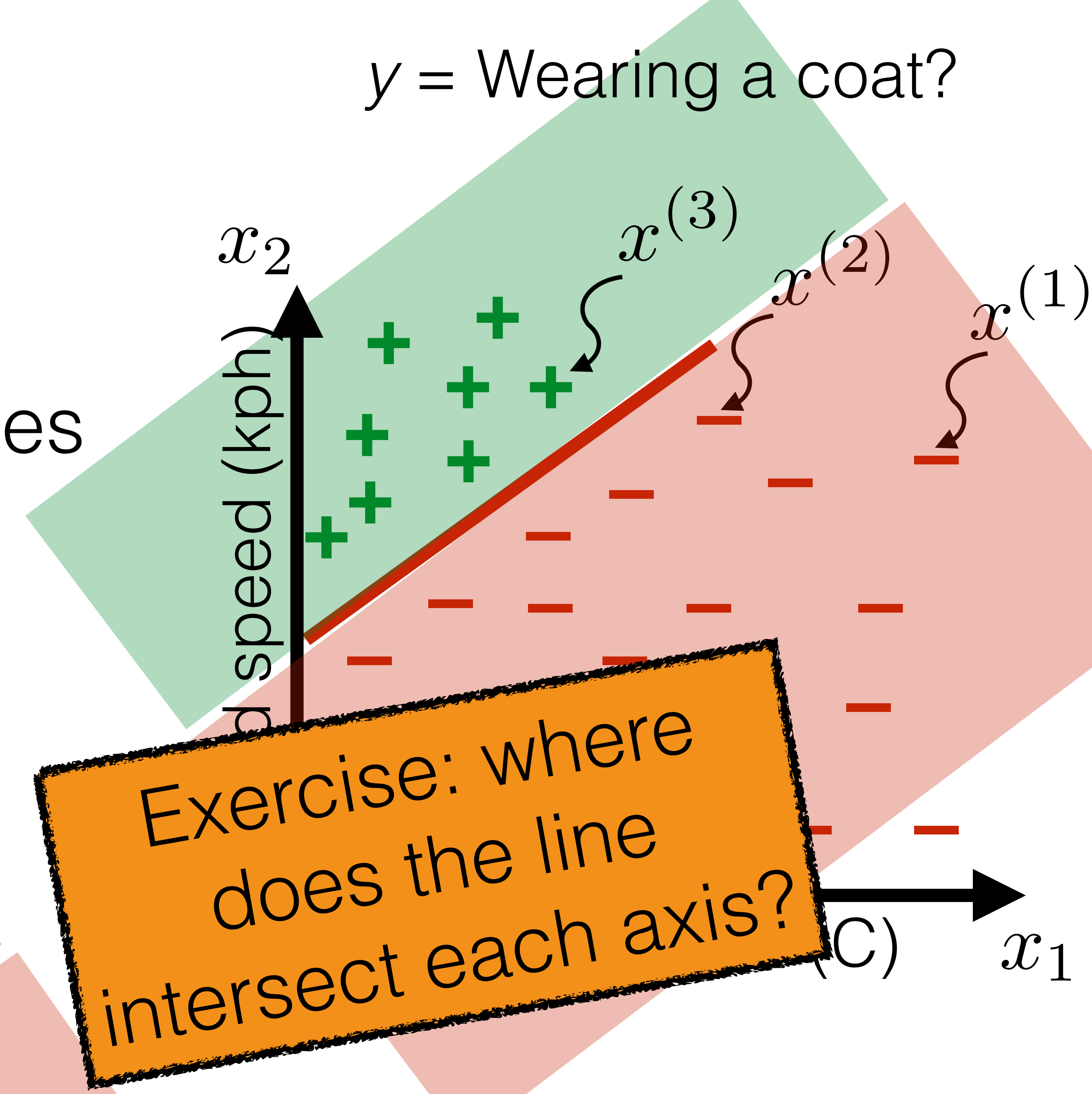


- Linear classifier:

$$h(x; \theta, \theta_0) = \text{sign}(\theta^\top x + \theta_0)$$

$$= \begin{cases} +1 \text{ if } \theta^\top x + \theta_0 > 0 \\ -1 \text{ if } \theta^\top x + \theta_0 < 0 \end{cases}$$

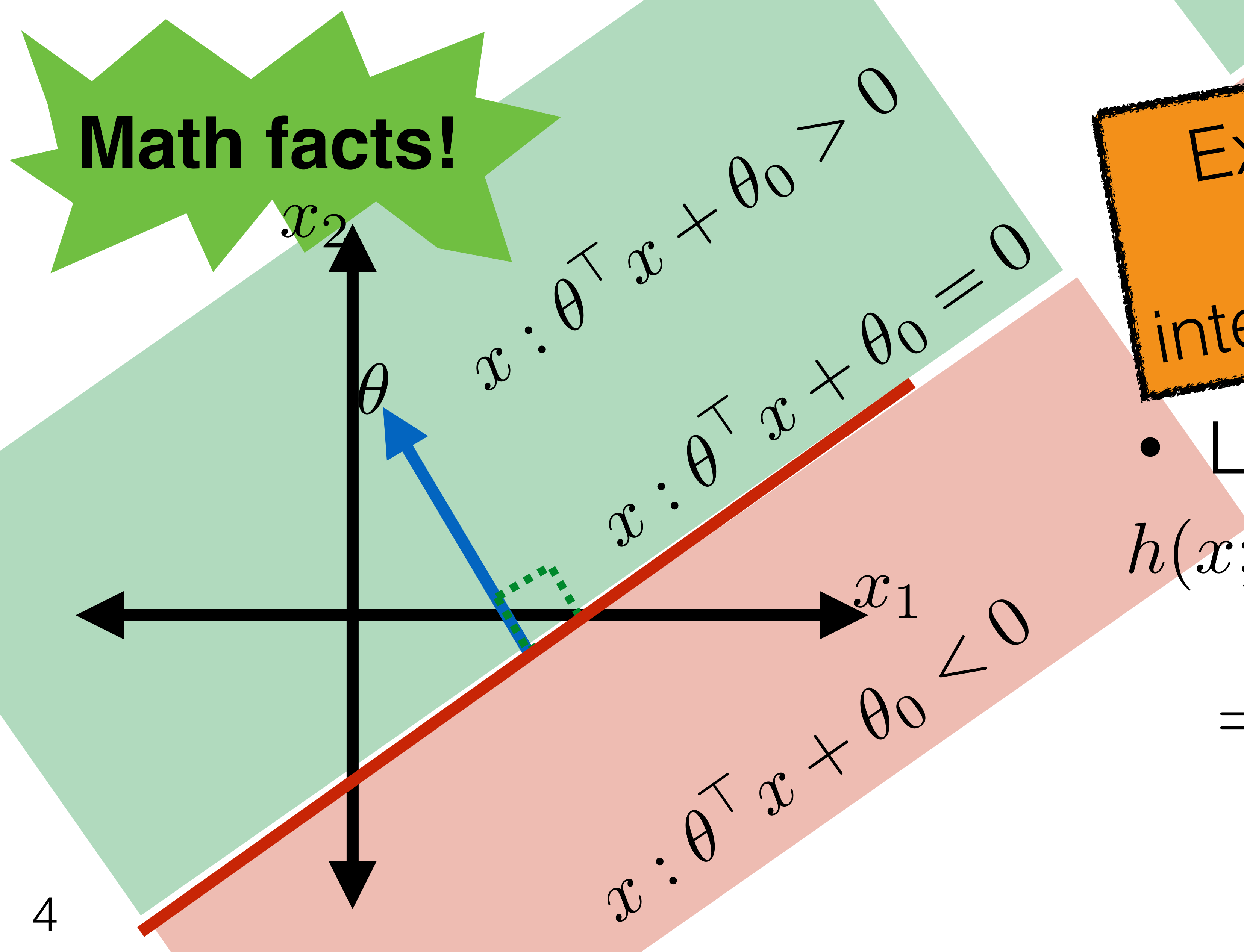
Linear classifiers



- Classification hypothesis:

$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$

- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side



- Linear classifier:

$$h(x; \theta, \theta_0) = \text{sign}(\theta^\top x + \theta_0)$$

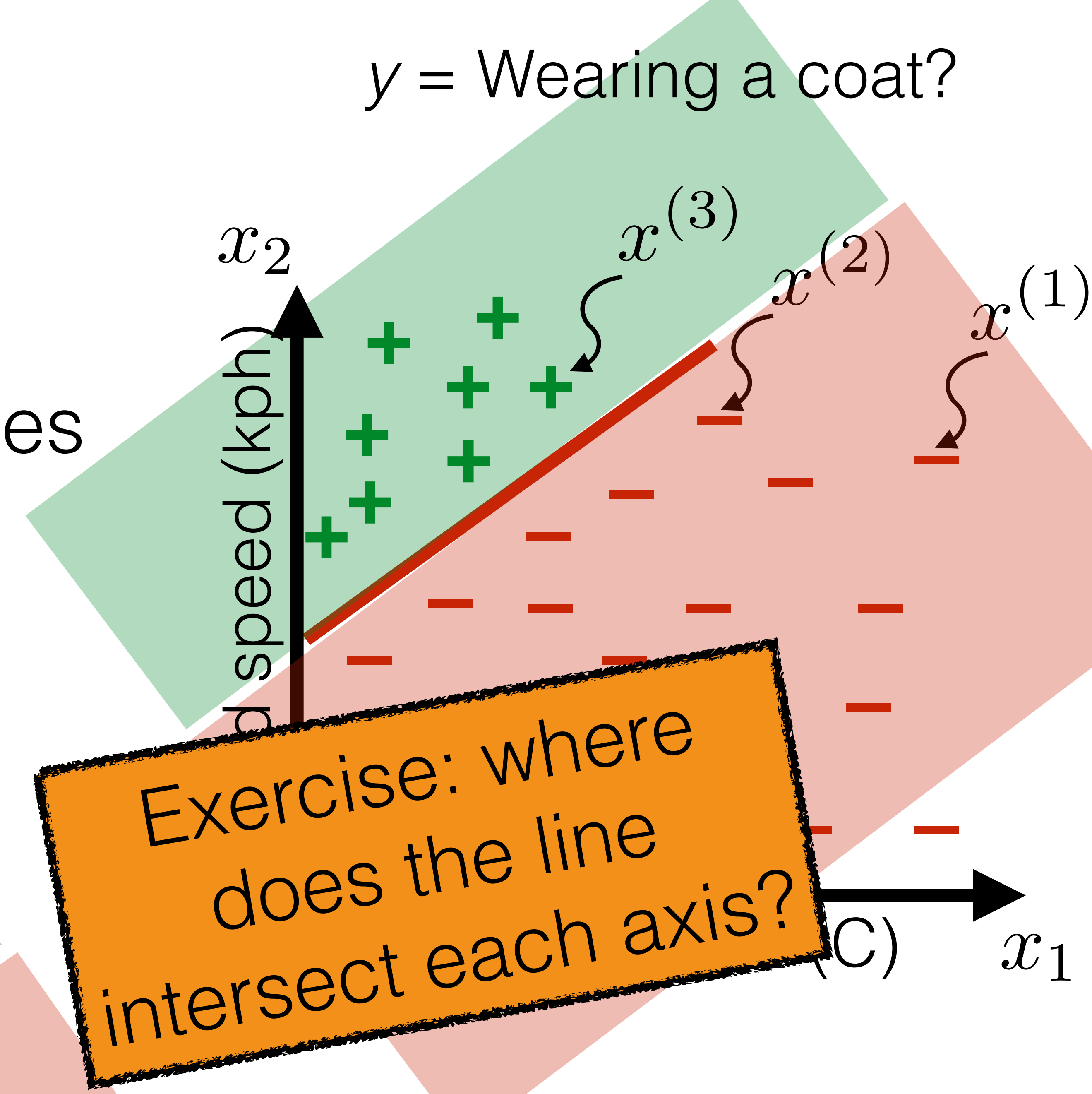
$$= \begin{cases} +1 \text{ if } \theta^\top x + \theta_0 > 0 \\ -1 \text{ if } \theta^\top x + \theta_0 < 0 \end{cases}$$

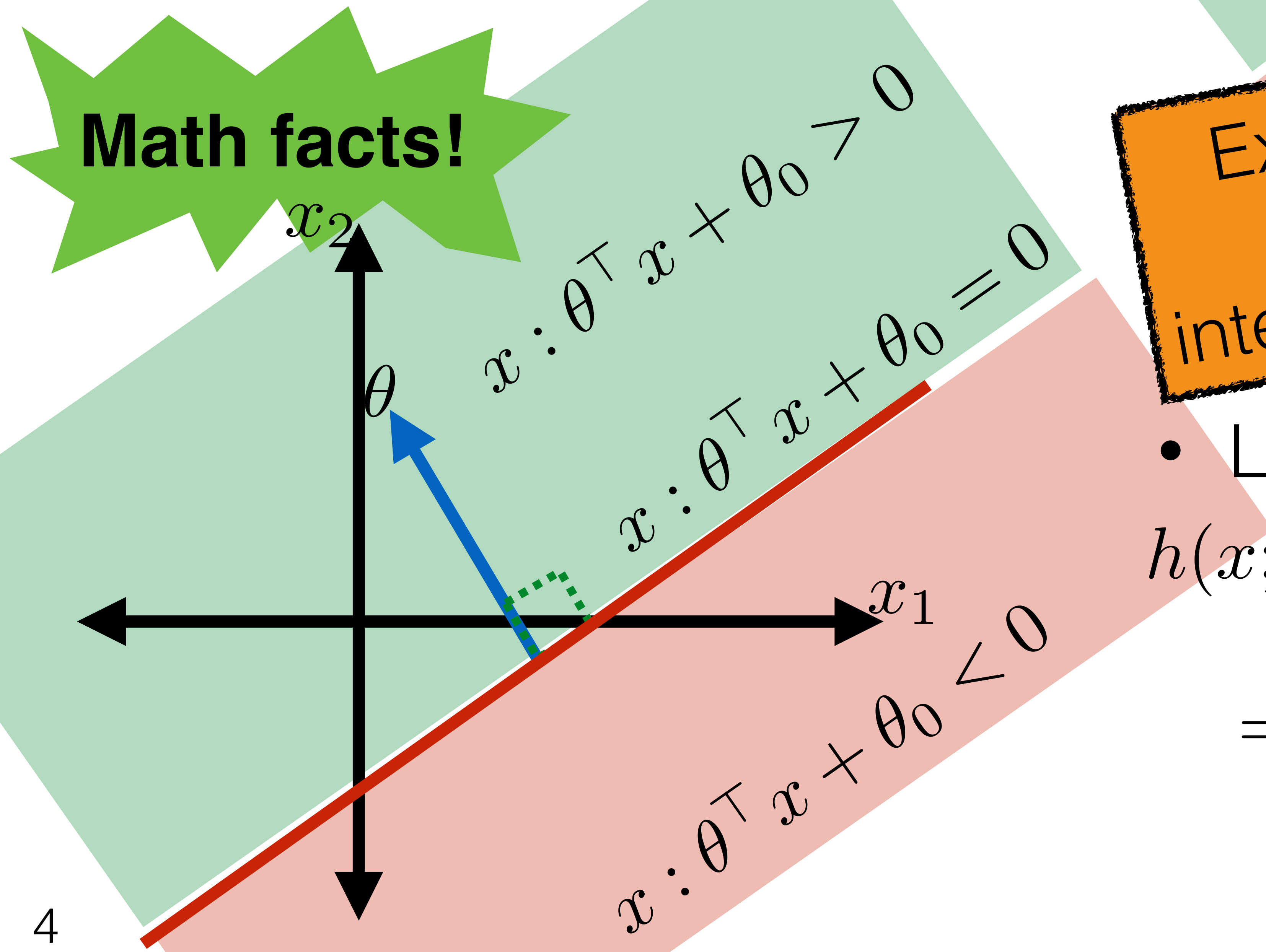
Linear classifiers

- Classification hypothesis:

$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$

- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side





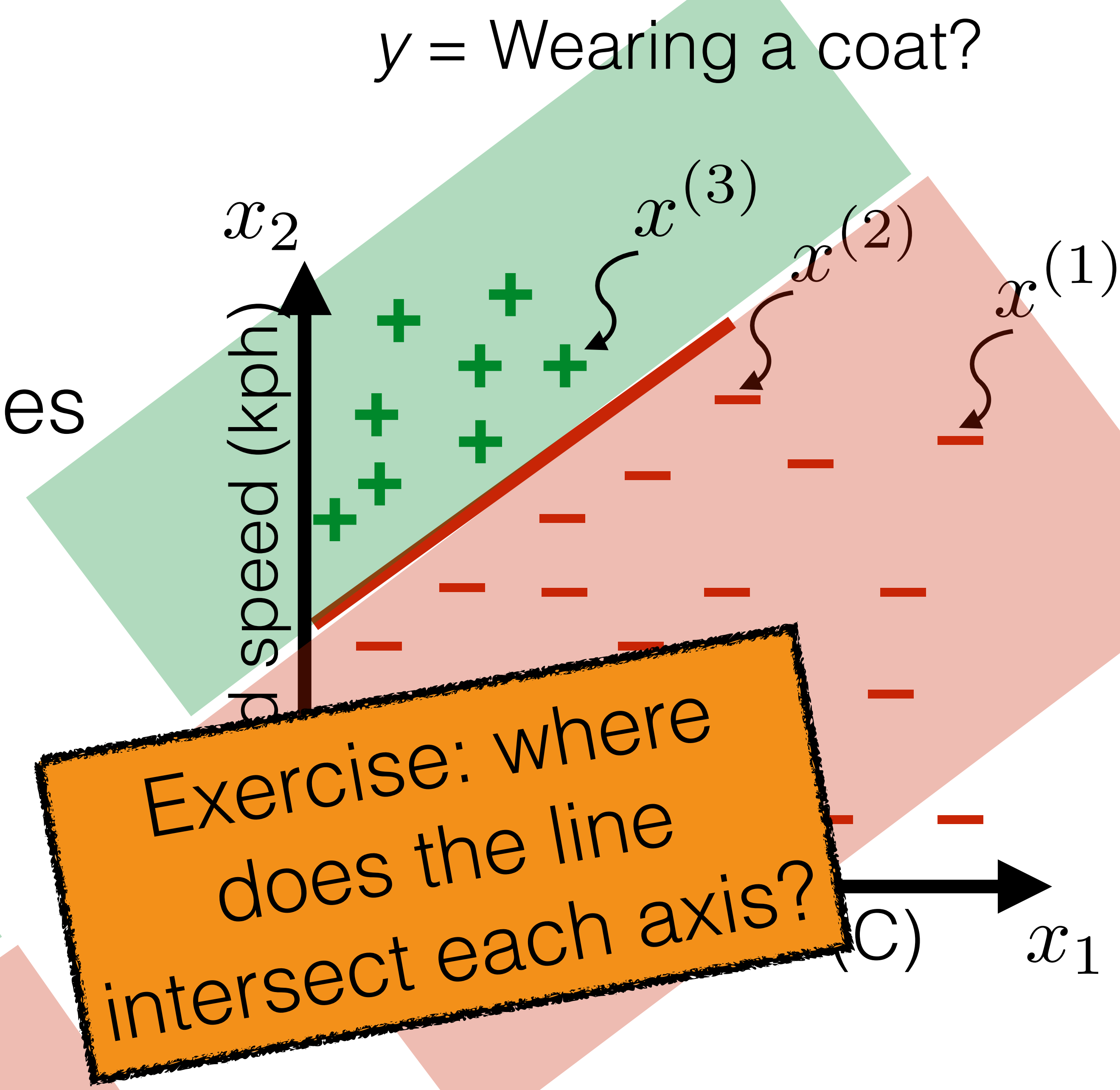
- Linear classifier:

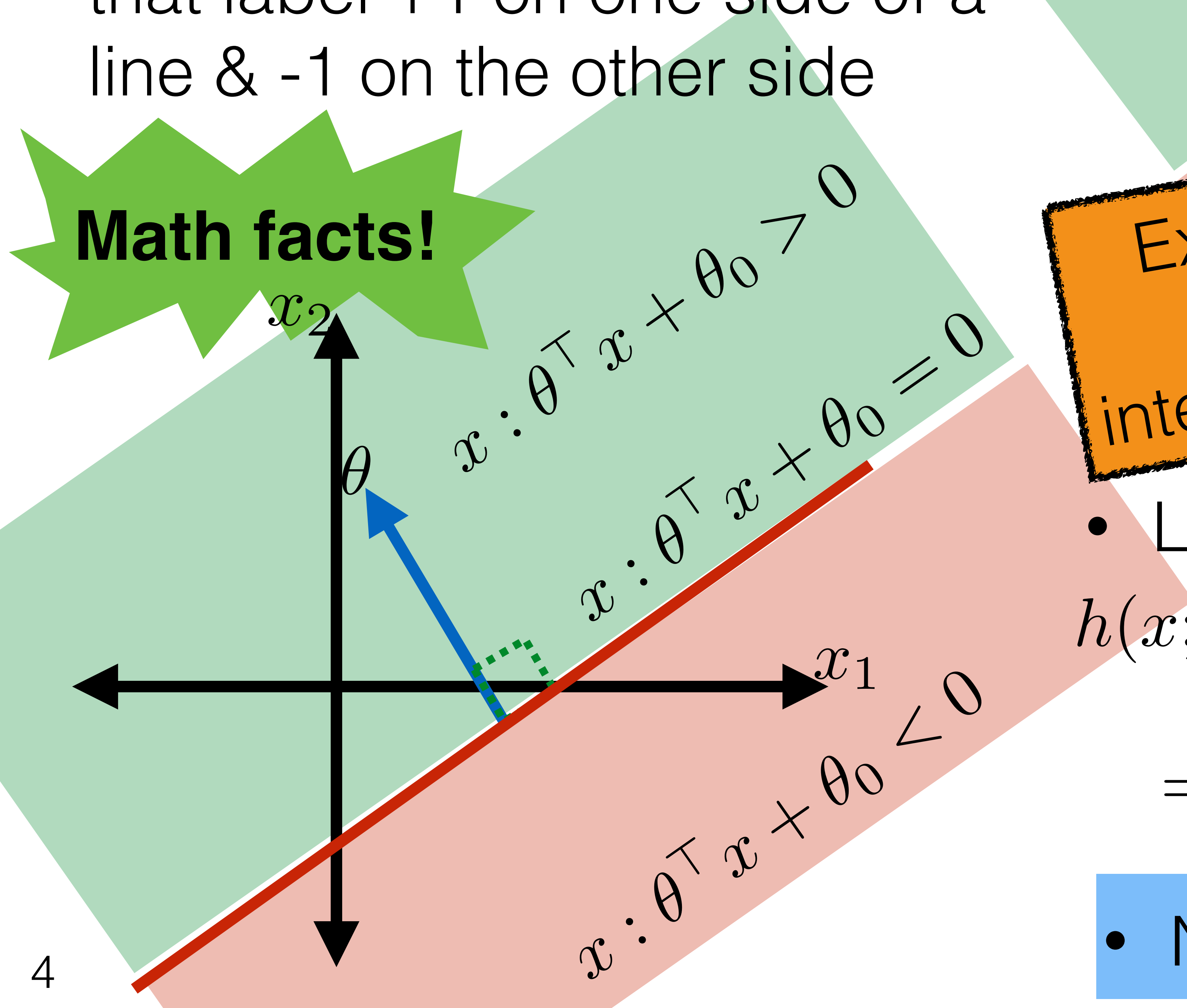
$$h(x; \theta, \theta_0) = \text{sign}(\theta^\top x + \theta_0)$$

$$= \begin{cases} +1 \text{ if } \theta^\top x + \theta_0 > 0 \\ -1 \text{ if } \theta^\top x + \theta_0 \leq 0 \end{cases}$$

Linear classifiers

- Classification hypothesis:
$$h : \mathbb{R}^d \to \{-1, +1\}$$
- Linear classifiers $\mathcal{H}$: Hypotheses that label +1 on one side of a line & -1 on the other side

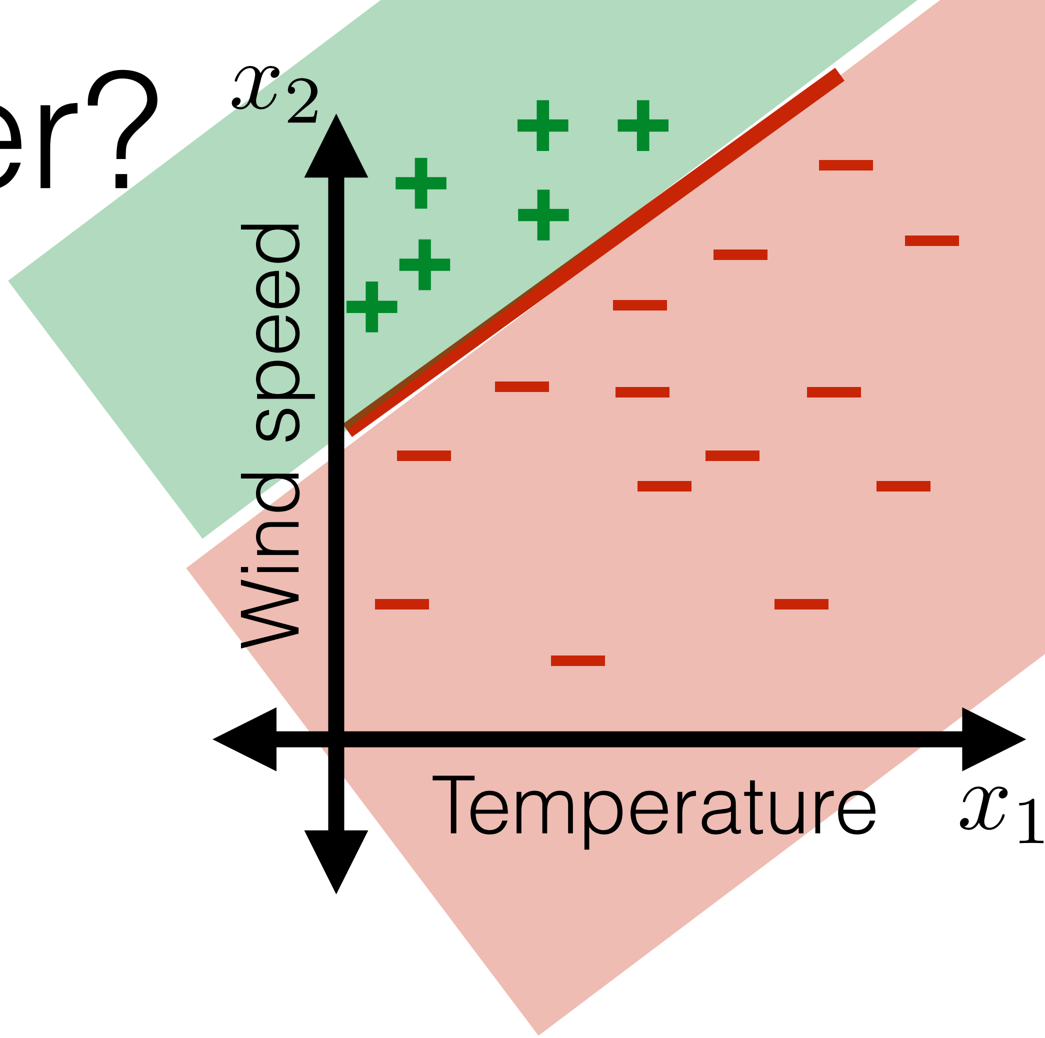




- Linear classifier:
$$h(x; \theta, \theta_0) = \text{sign}(\theta^\top x + \theta_0)$$
$$= \begin{cases} +1 \text{ if } \theta^\top x + \theta_0 > 0 \\ -1 \text{ if } \theta^\top x + \theta_0 \leq 0 \end{cases}$$

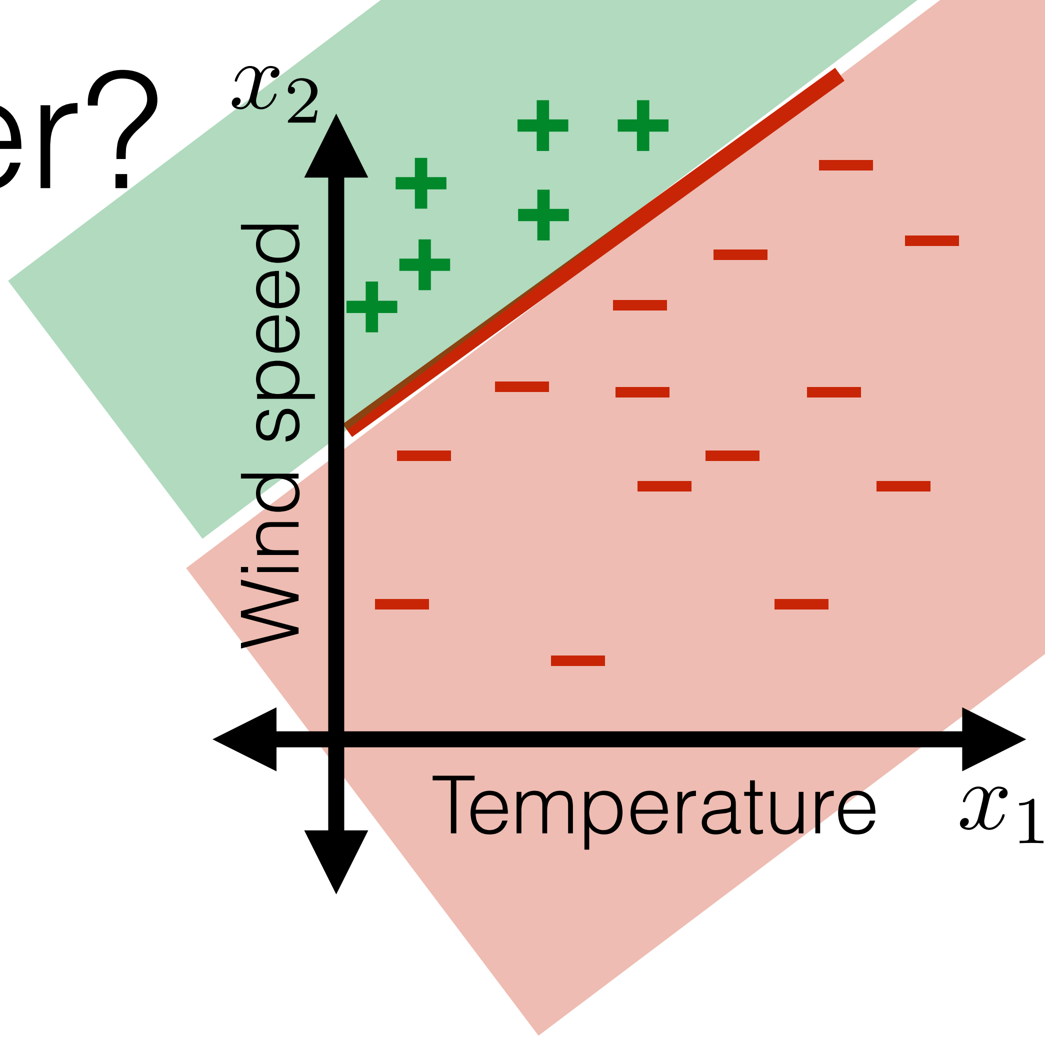
- Note: θ tells us direction

How good is a classifier?



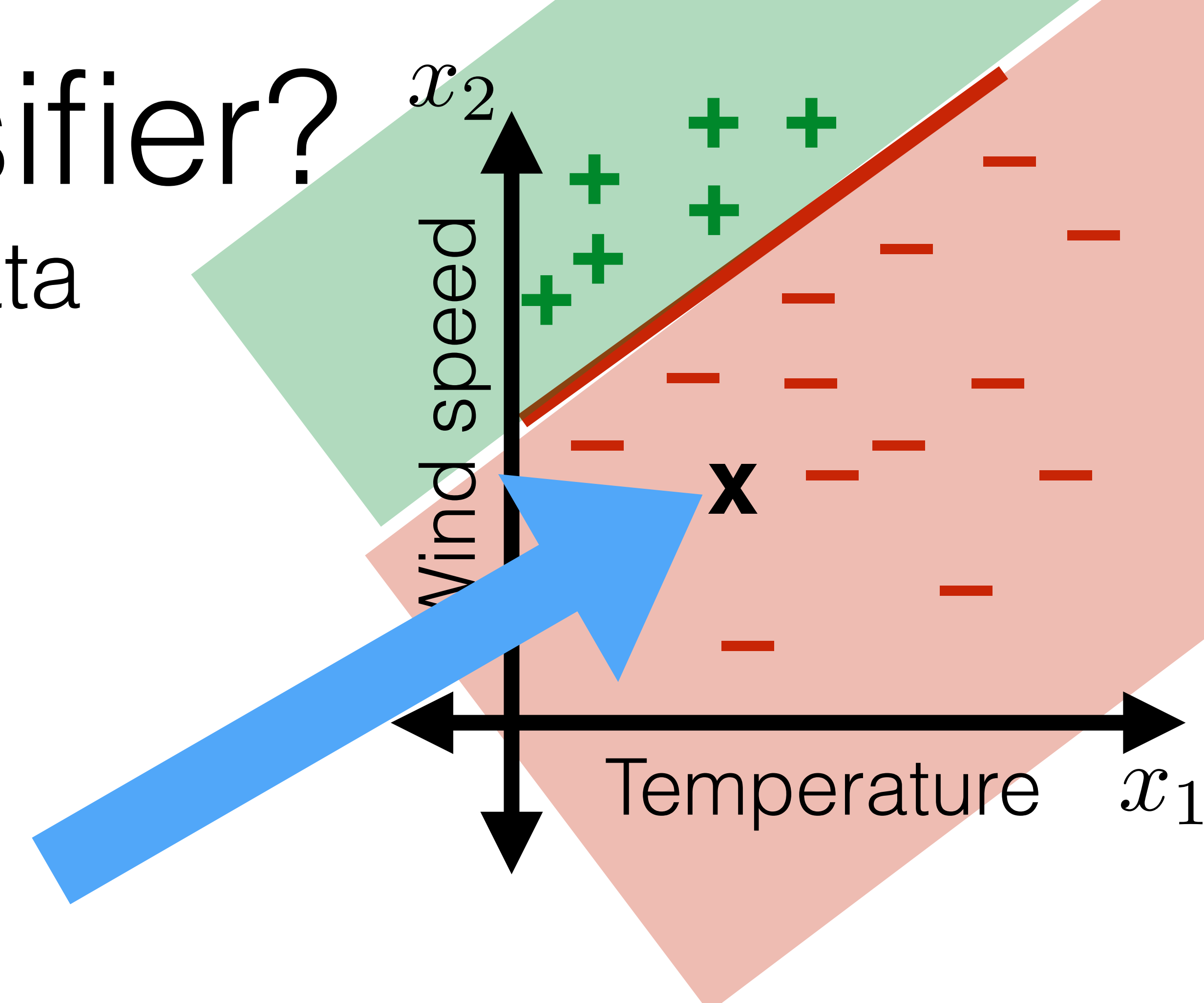
How good is a classifier?

- Should predict well on future data



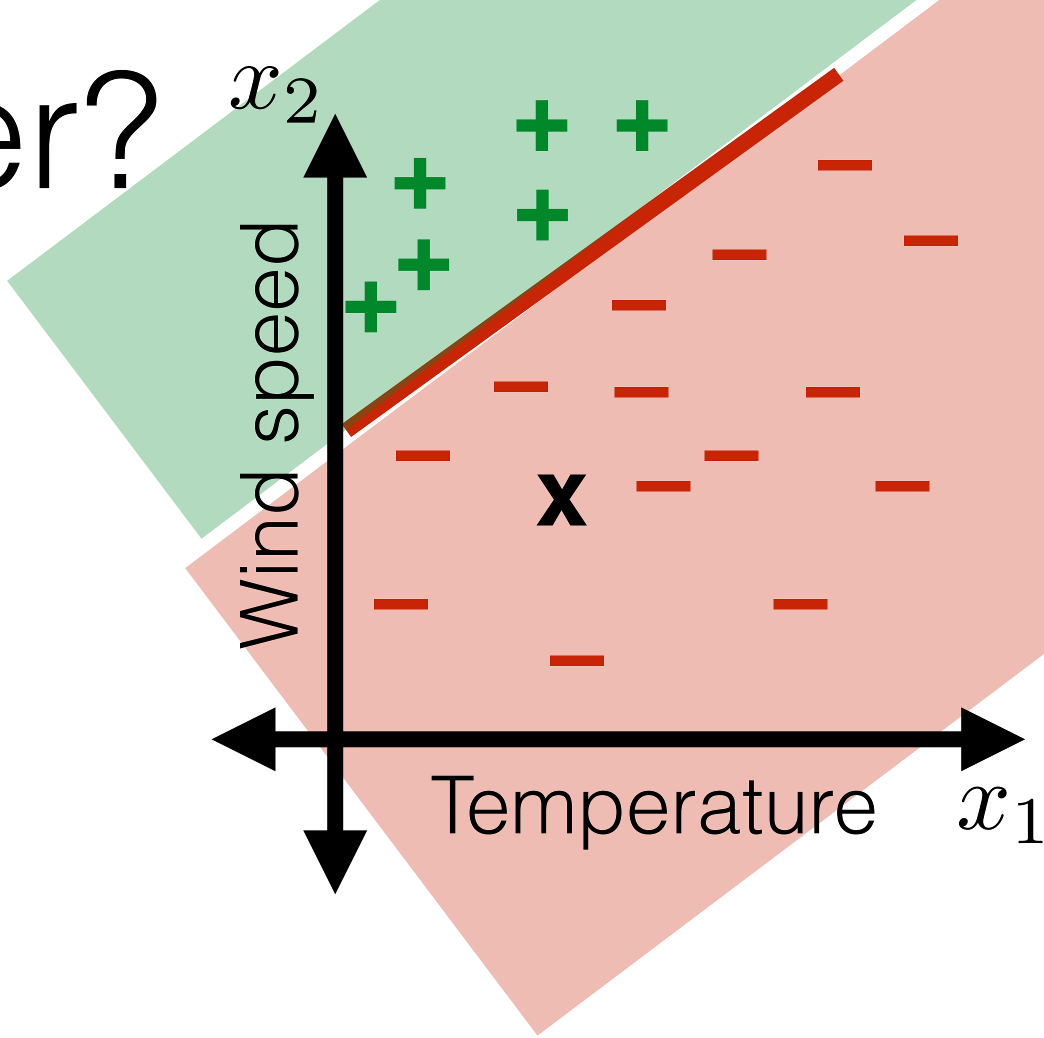
How good is a classifier?

- Should predict well on future data



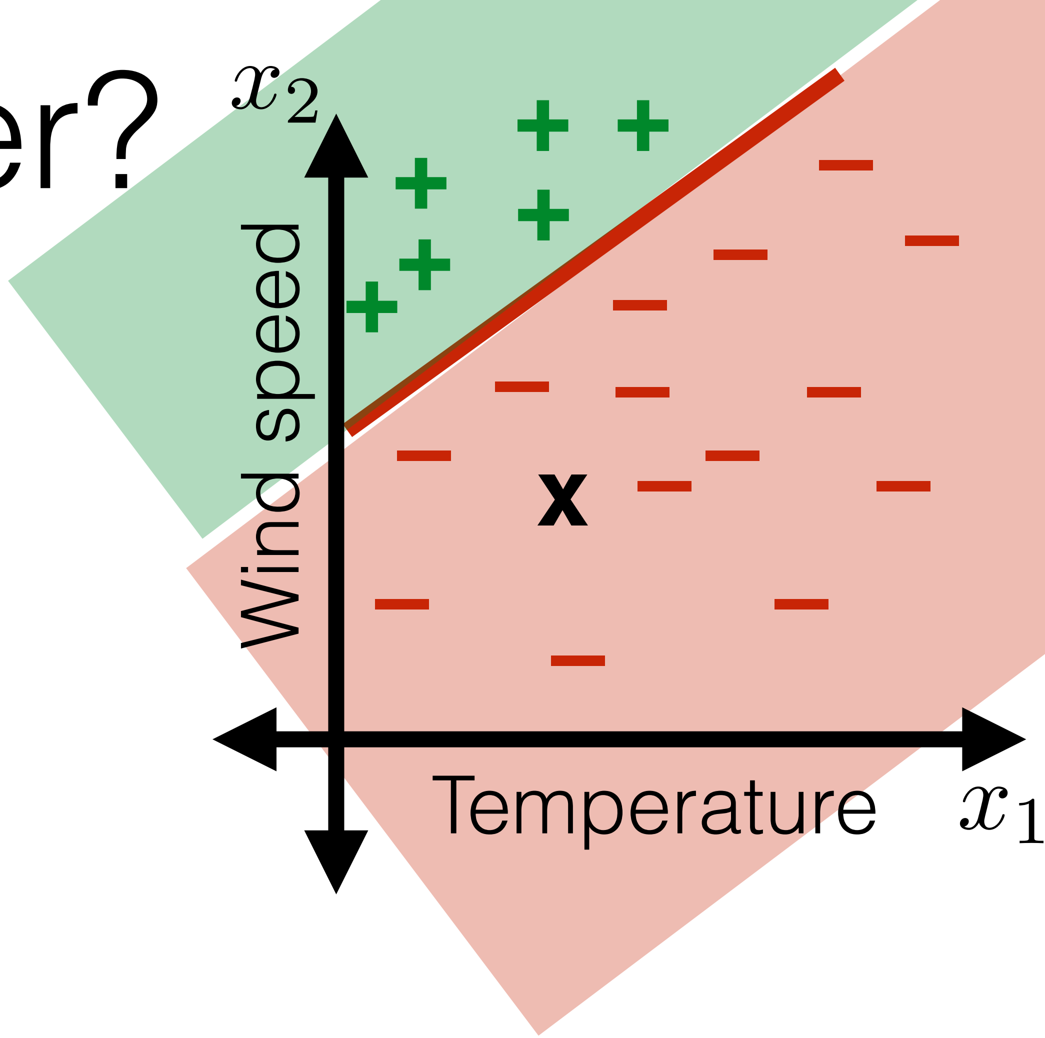
How good is a classifier?

- Should predict well on future data



How good is a classifier?

- Should predict well on future data
  - Example: 0-1 loss

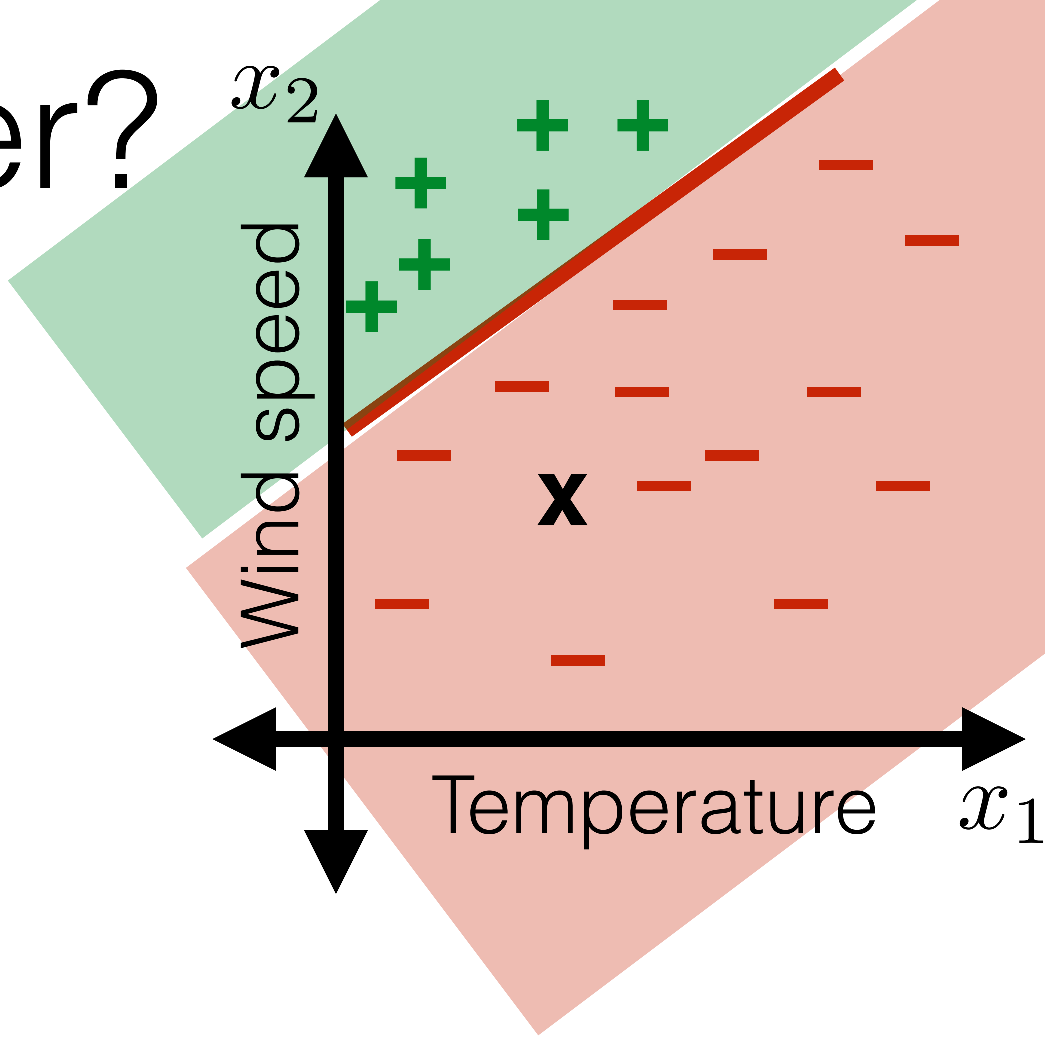


How good is a classifier?

- Should predict well on future data
  - Example: 0-1 loss

$$L(g, a) = \begin{cases} 0 \text{ if } g = a \\ 1 \text{ else} \end{cases}$$

g: guess,
a: actual



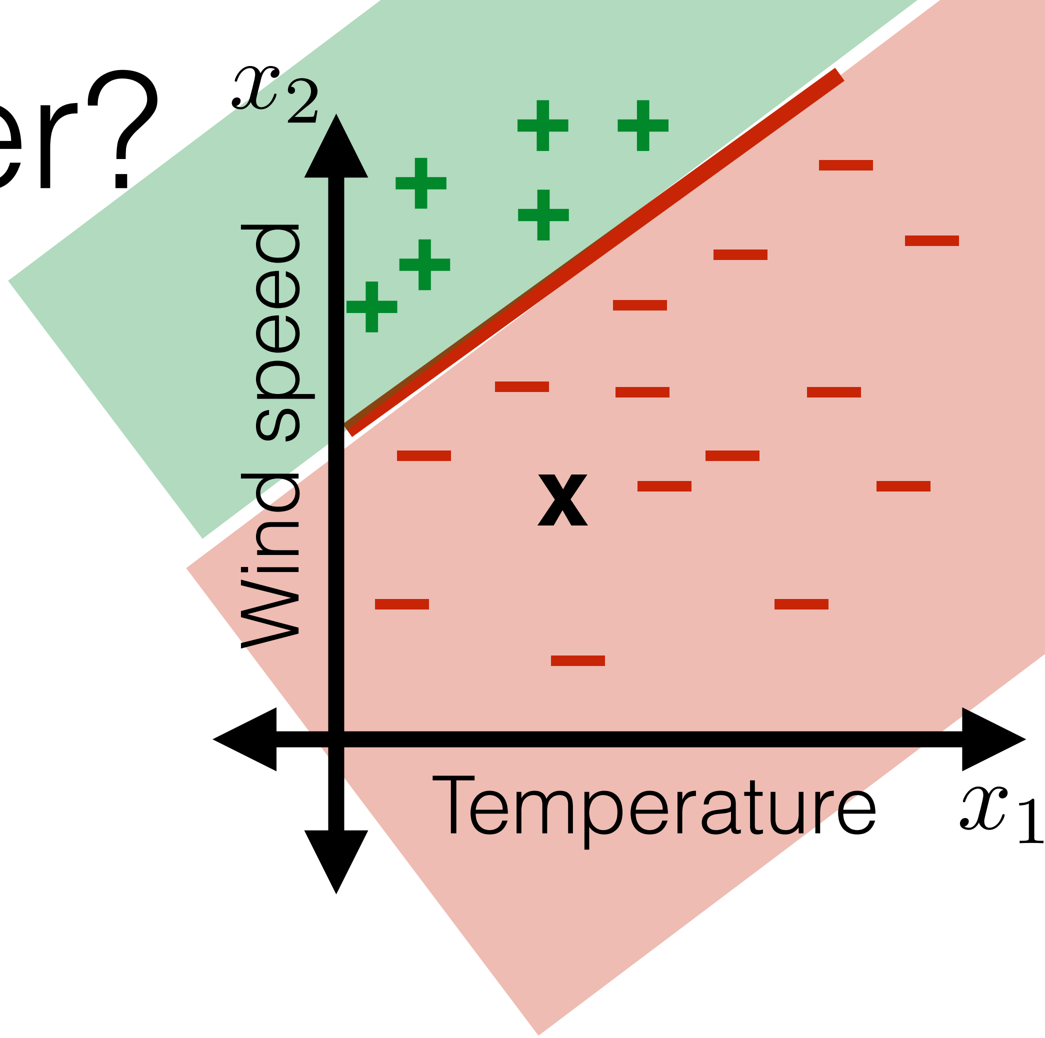
How good is a classifier?

- Should predict well on future data
  - Example: 0-1 loss

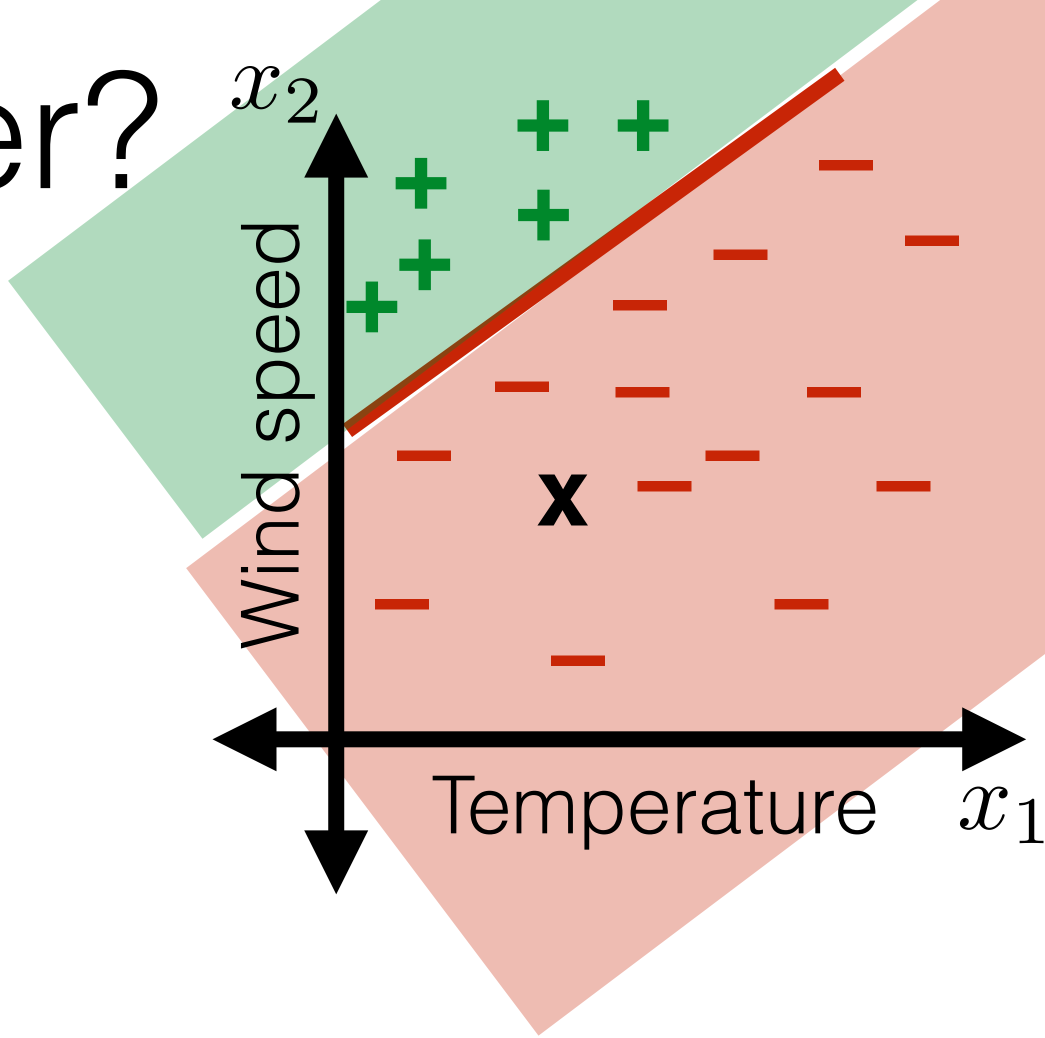
$$L(g, a) = \begin{cases} 0 \text{ if } g = a \\ 1 \text{ else} \end{cases}$$

g: guess,
a: actual

  - Example: asymmetric loss



How good is a classifier?



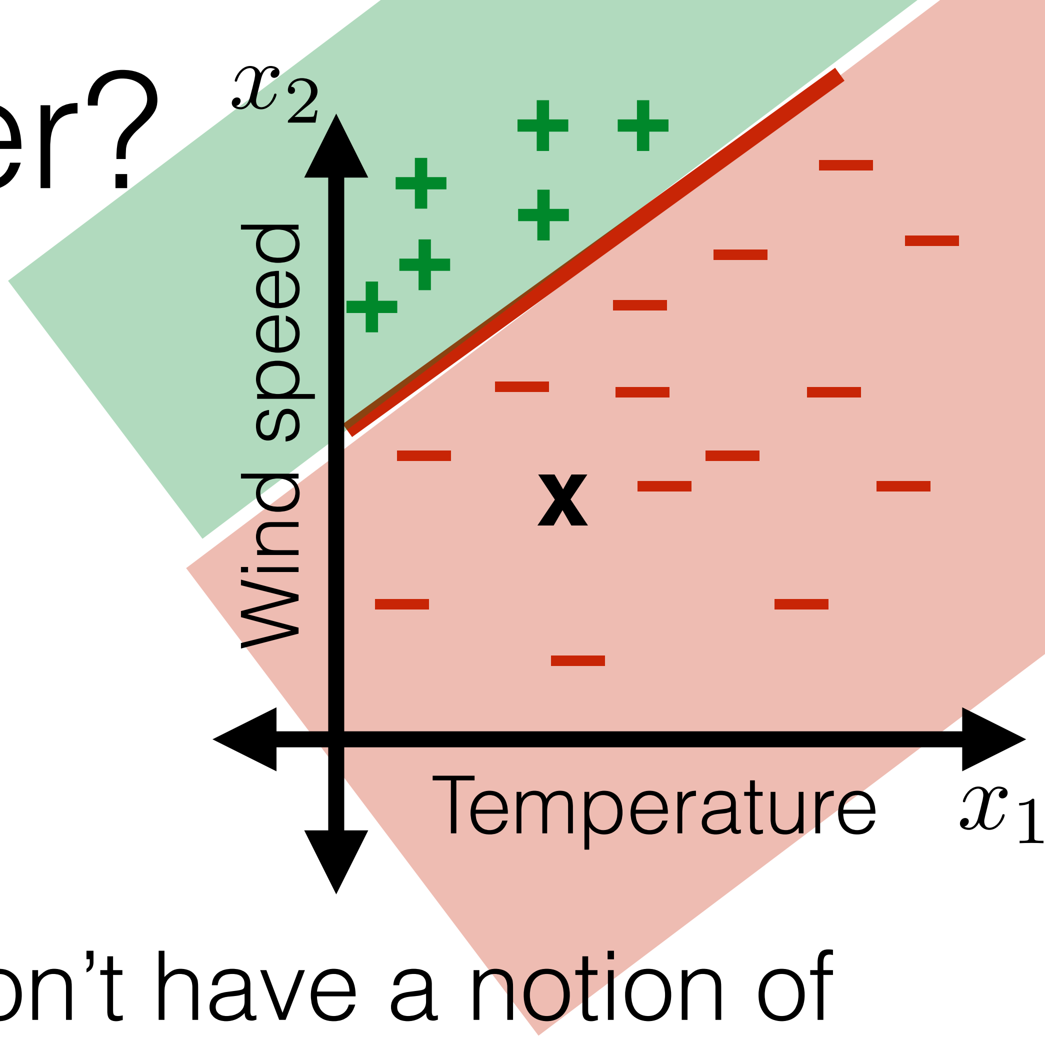
- Should predict well on future data
  - Example: 0-1 loss

$$L(g, a) = \begin{cases} 0 \text{ if } g = a \\ 1 \text{ else} \end{cases}$$

g: guess,
a: actual

  - Example: asymmetric loss
  - But:

How good is a classifier?



- Should predict well on future data
  - Example: 0-1 loss

$$L(g, a) = \begin{cases} 0 \text{ if } g = a \\ 1 \text{ else} \end{cases}$$

g: guess,
a: actual

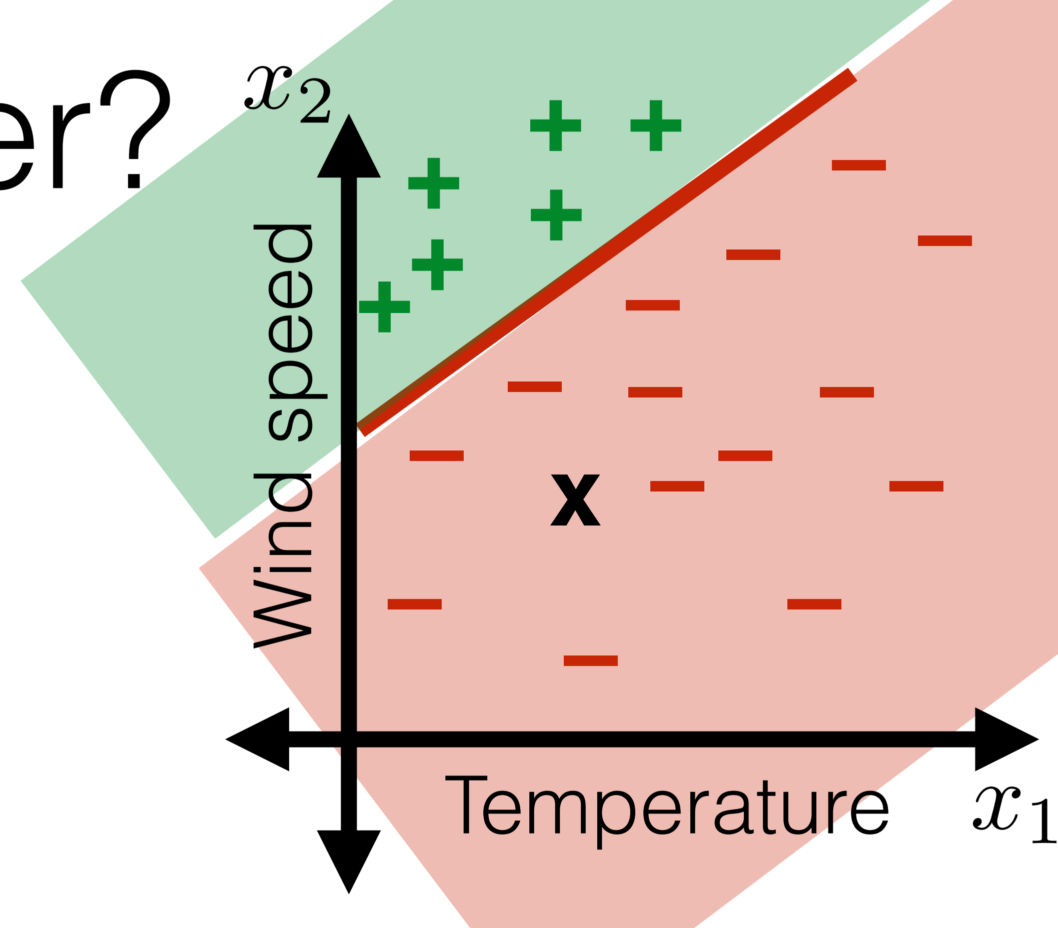
  - Example: asymmetric loss
  - But: 0-1 loss & linear classifiers don't have a notion of uncertainty (how well do we know what we know?)

How good is a classifier?

- Should predict well on future data
  - Example: 0-1 loss

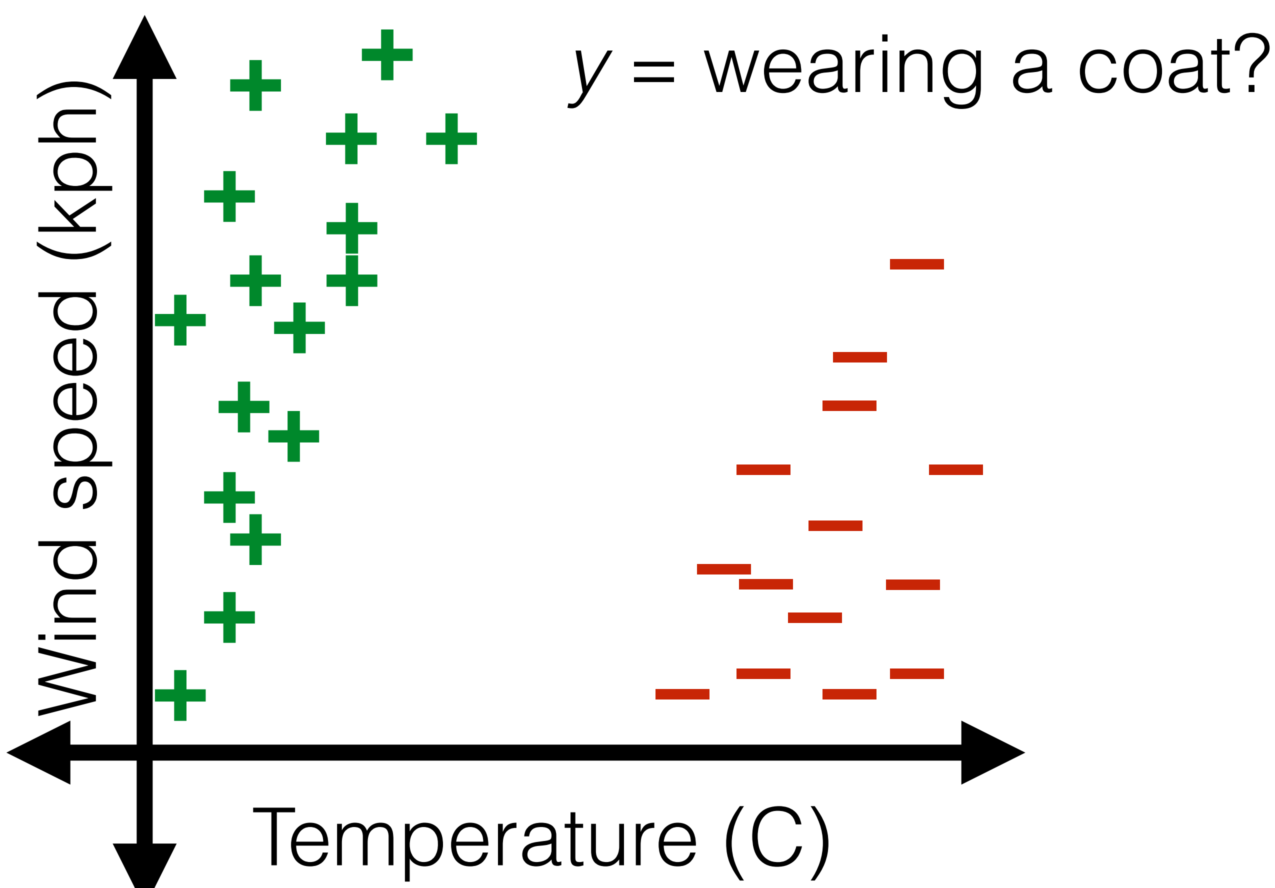
$$L(g, a) = \begin{cases} 0 \text{ if } g = a \\ 1 \text{ else} \end{cases}$$

g: guess,
a: actual

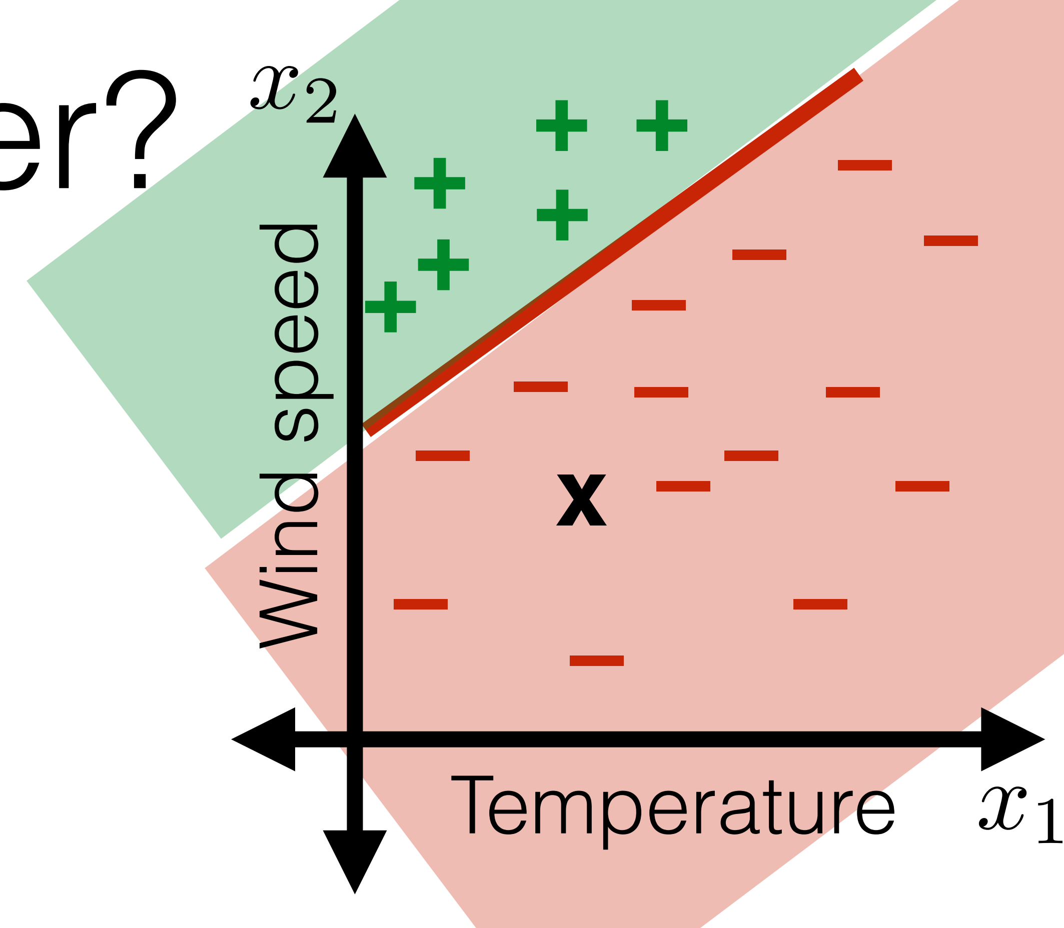


  - Example: asymmetric loss

  - But: 0-1 loss & linear classifiers don't have a notion of uncertainty (how well do we know what we know?)



How good is a classifier?

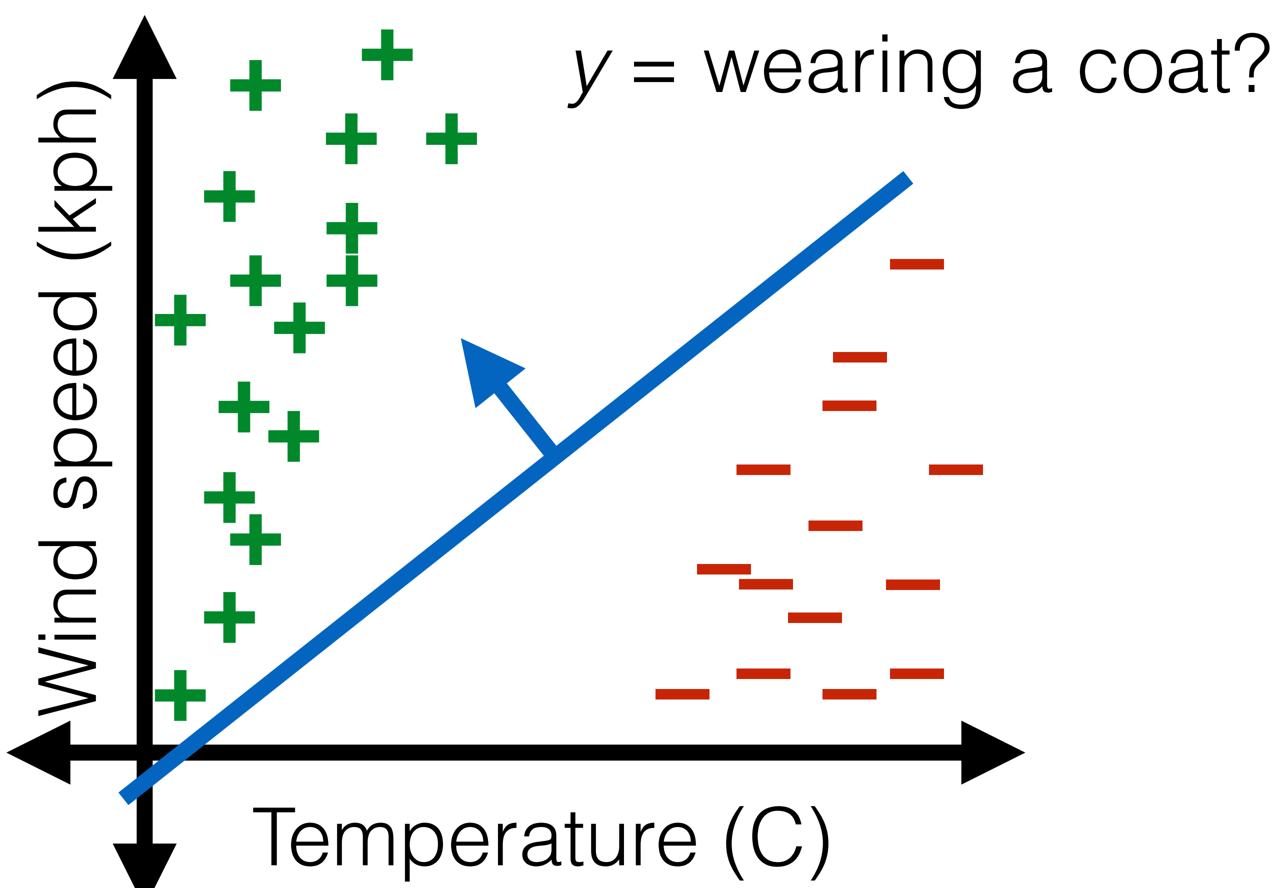


- Should predict well on future data
  - Example: 0-1 loss

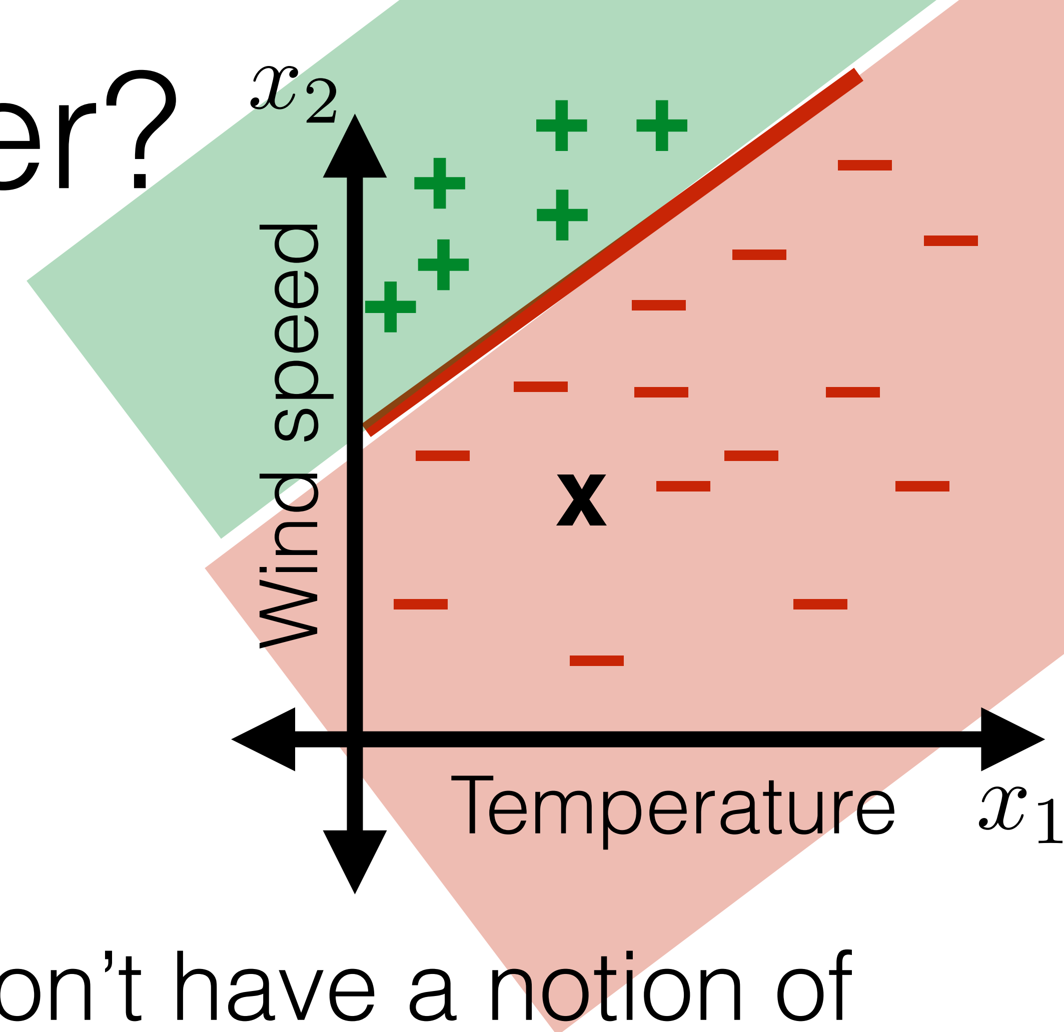
$$L(g, a) = \begin{cases} 0 \text{ if } g = a \\ 1 \text{ else} \end{cases}$$

g: guess,
a: actual

  - Example: asymmetric loss
  - But: 0-1 loss & linear classifiers don't have a notion of uncertainty (how well do we know what we know?)



How good is a classifier?

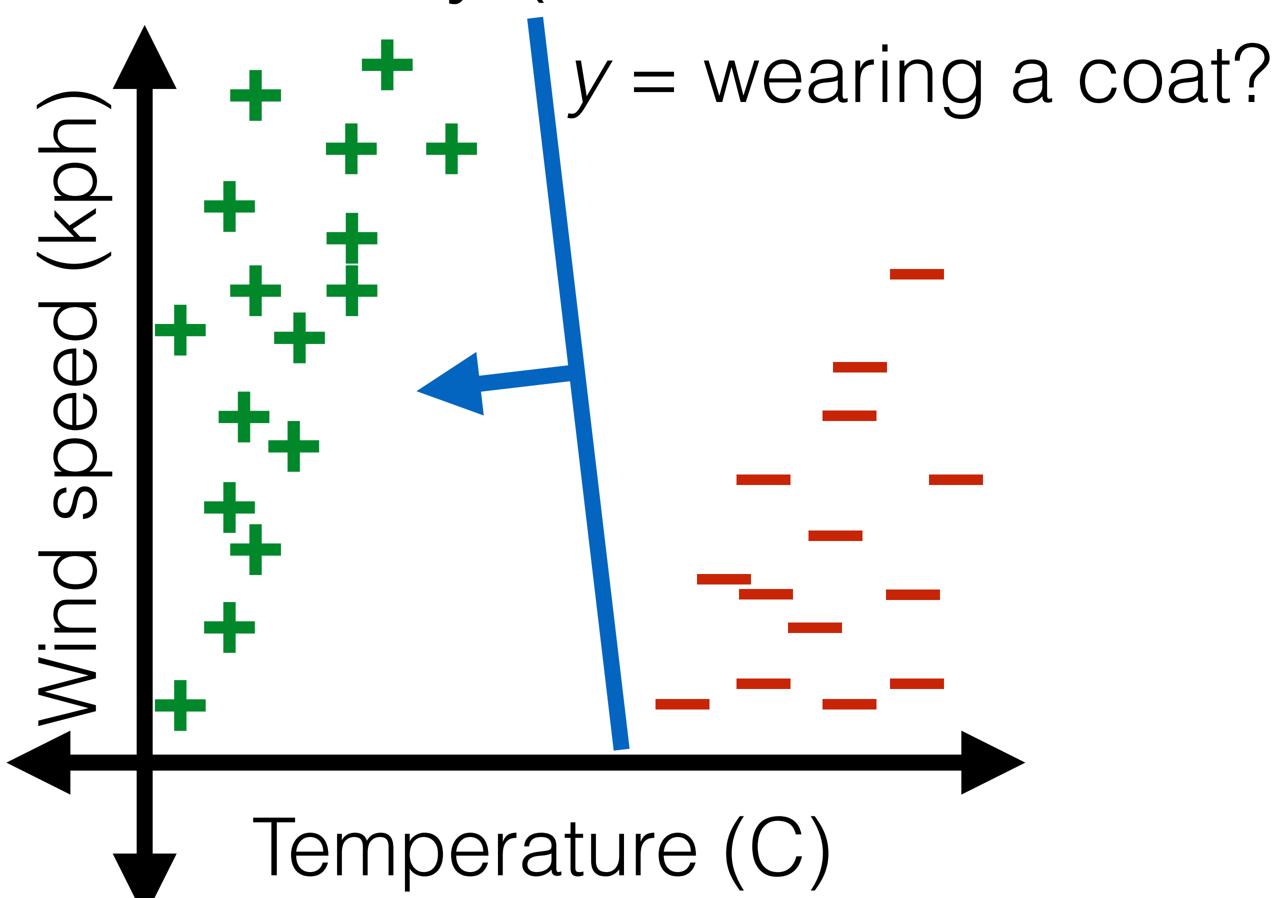


- Should predict well on future data
  - Example: 0-1 loss

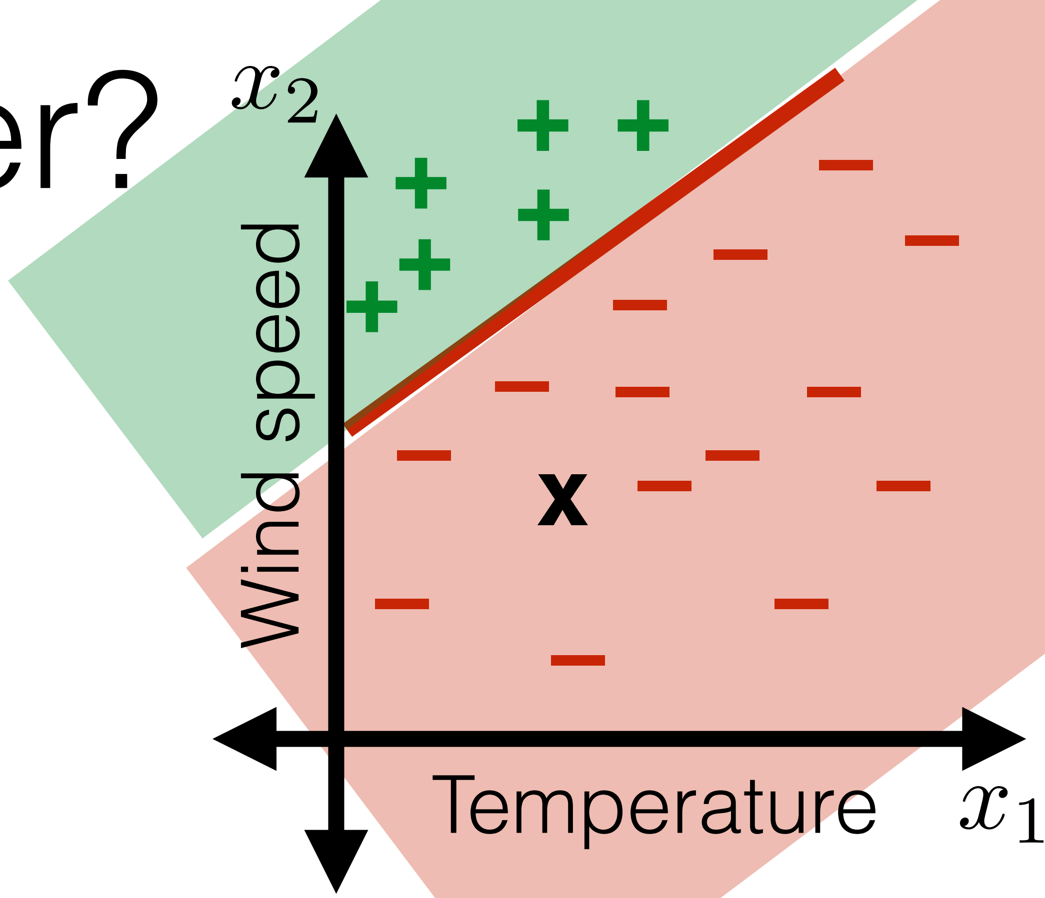
$$L(g, a) = \begin{cases} 0 \text{ if } g = a \\ 1 \text{ else} \end{cases}$$

g: guess,
a: actual

  - Example: asymmetric loss
  - But: 0-1 loss & linear classifiers don't have a notion of uncertainty (how well do we know what we know?)



How good is a classifier?

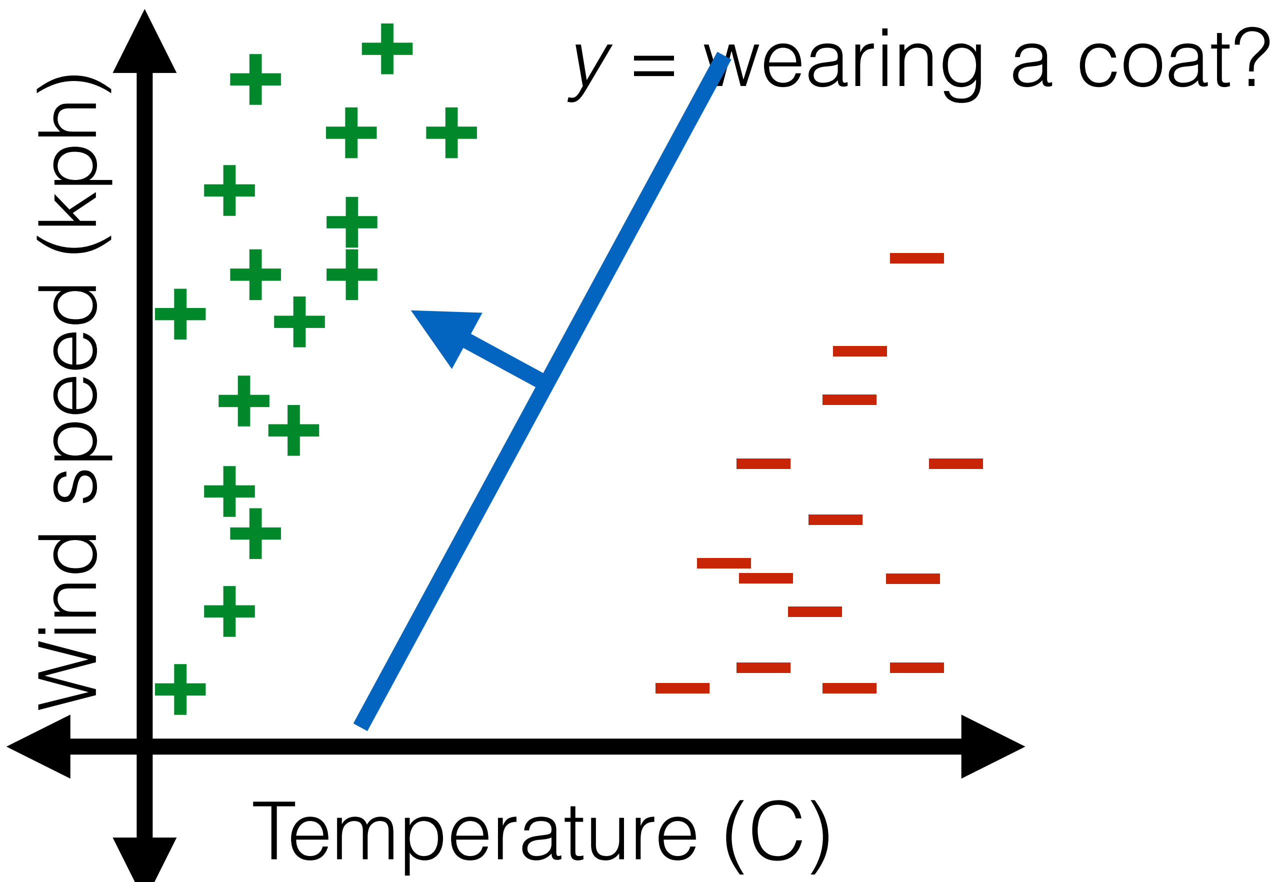


- Should predict well on future data
  - Example: 0-1 loss

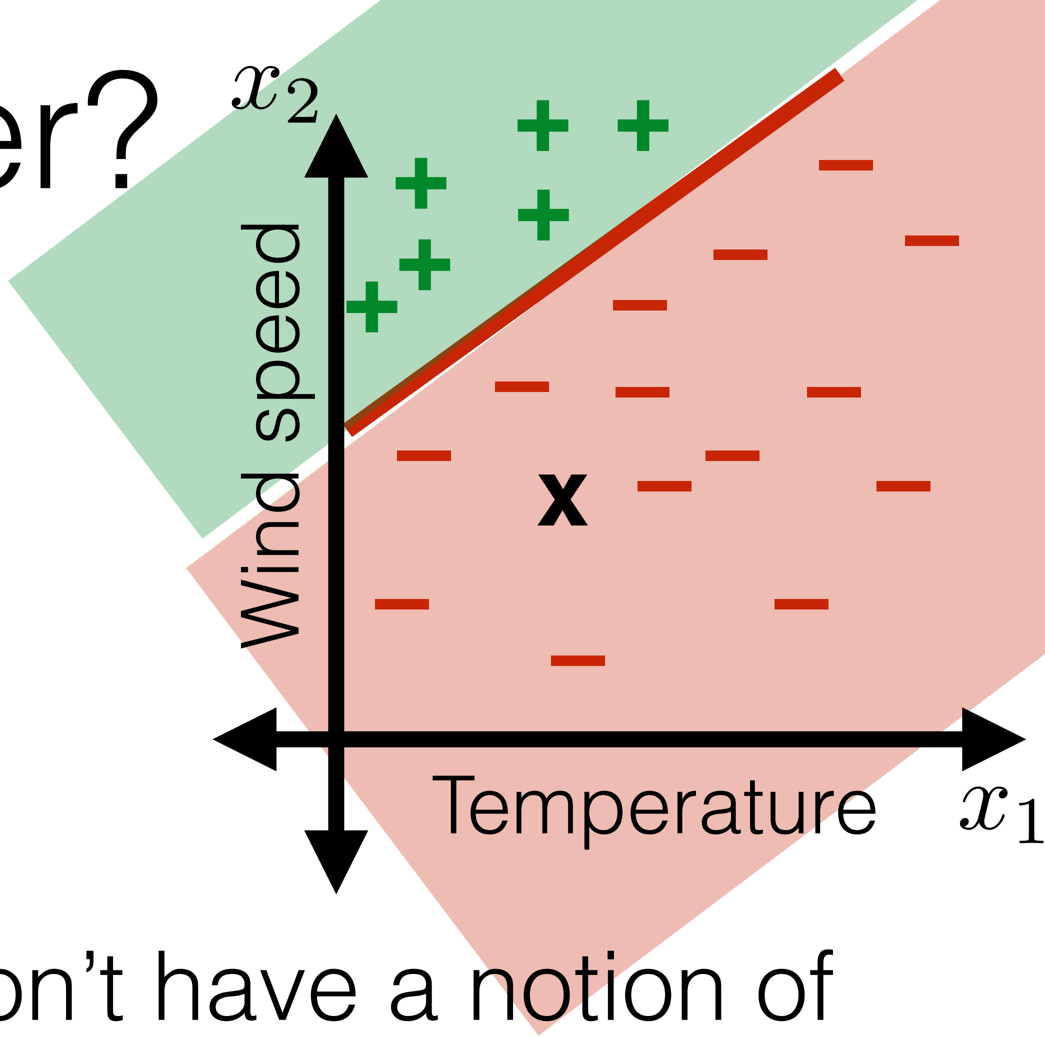
$$L(g, a) = \begin{cases} 0 \text{ if } g = a \\ 1 \text{ else} \end{cases}$$

g: guess,
a: actual

  - Example: asymmetric loss
  - But: 0-1 loss & linear classifiers don't have a notion of uncertainty (how well do we know what we know?)



How good is a classifier?

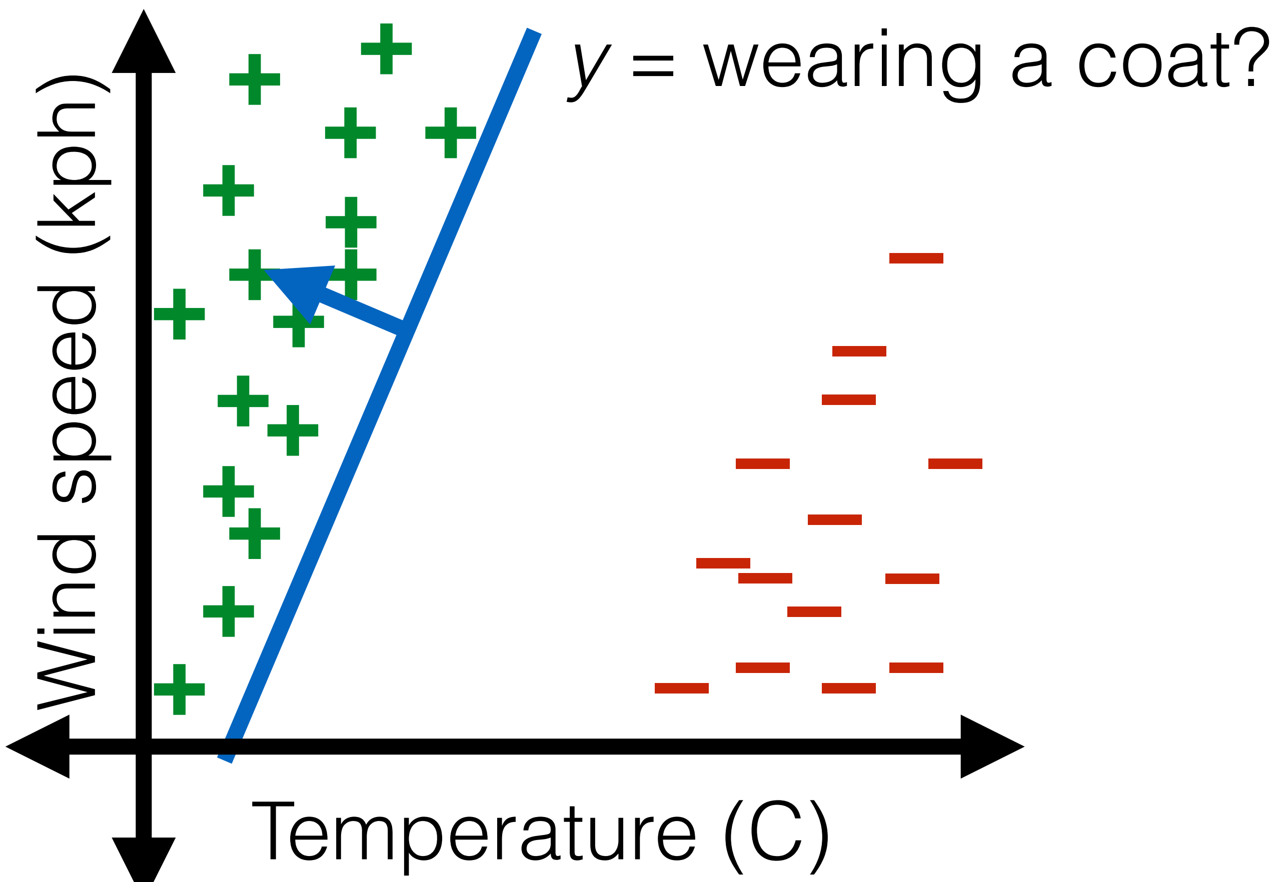


- Should predict well on future data
  - Example: 0-1 loss

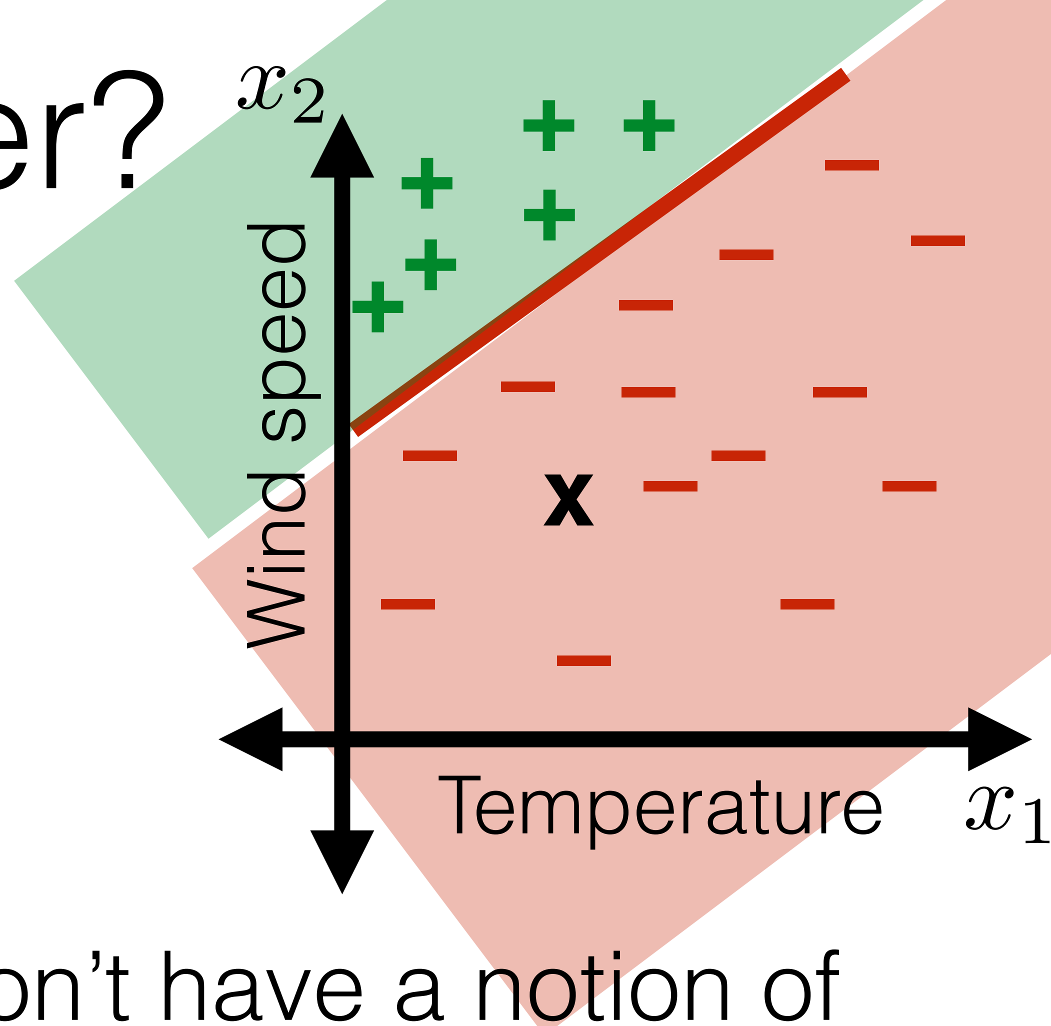
$$L(g,a) = \begin{cases} 0 \text{ if } g = a \\ 1 \text{ else} \end{cases}$$

g: guess,
a: actual

  - Example: asymmetric loss
  - But: 0-1 loss & linear classifiers don't have a notion of uncertainty (how well do we know what we know?)



How good is a classifier?



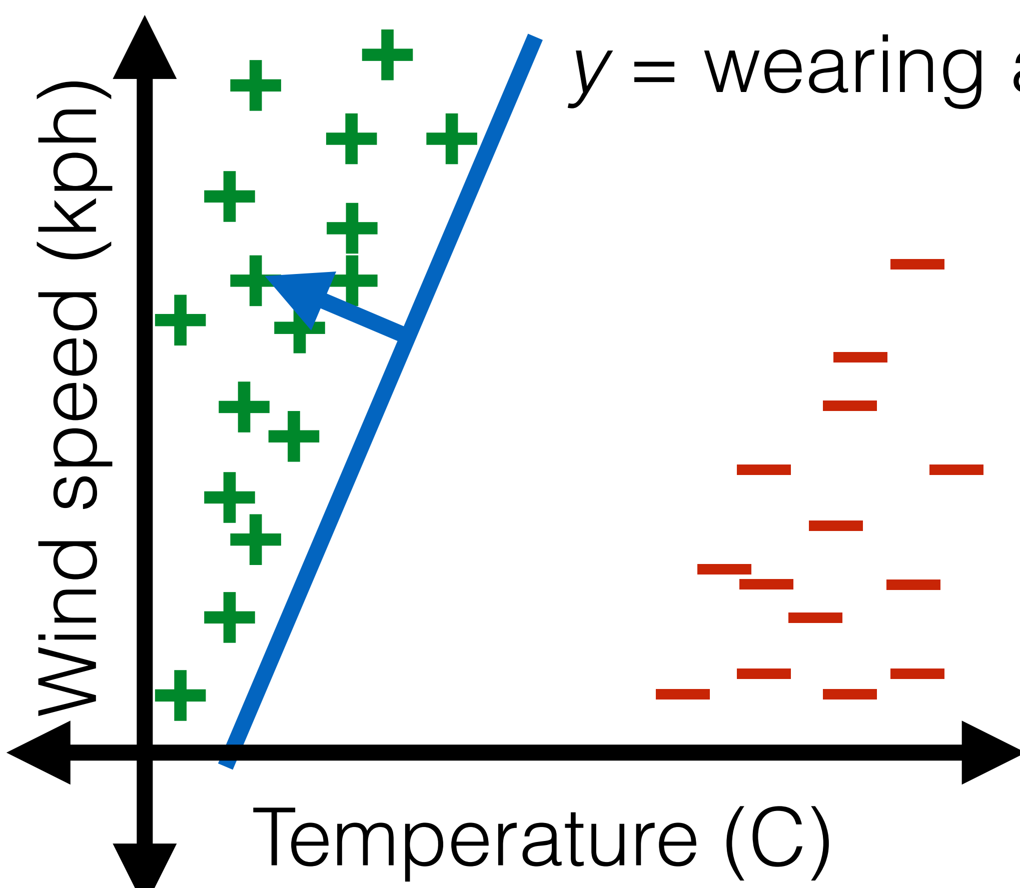
- Should predict well on future data
  - Example: 0-1 loss

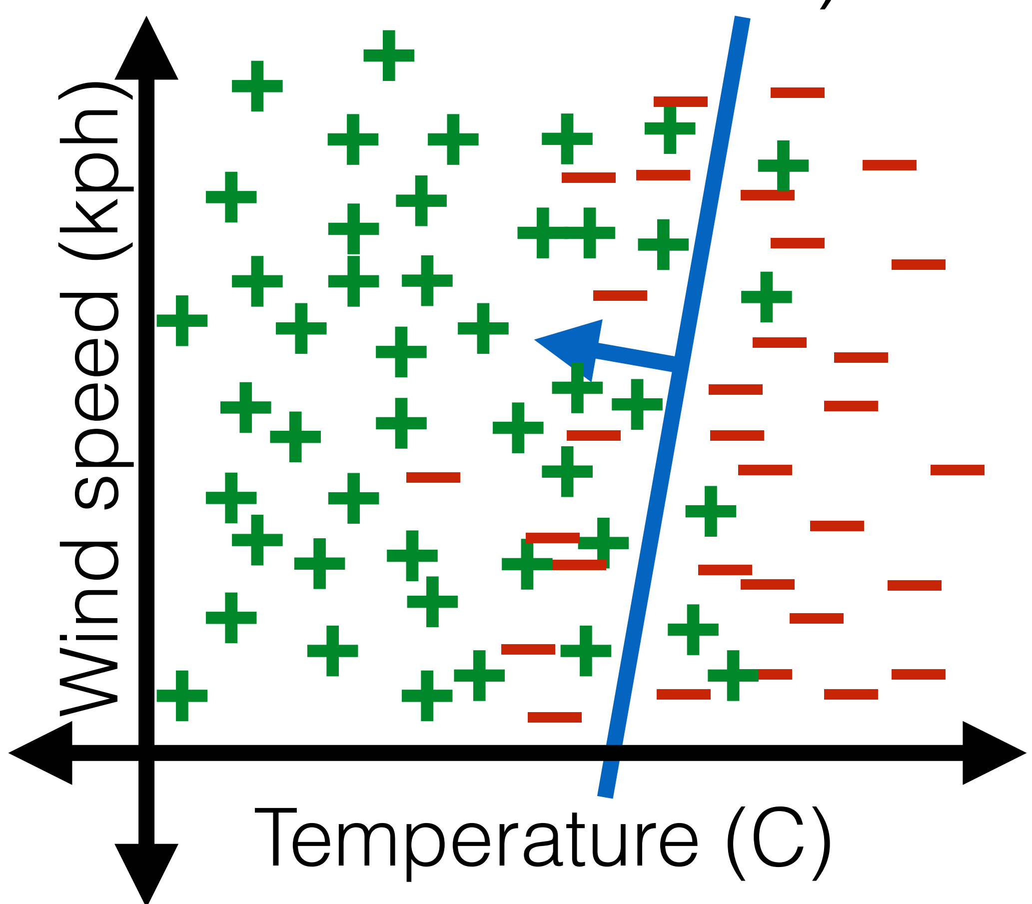
$$L(g, a) = \begin{cases} 0 \text{ if } g = a \\ 1 \text{ else} \end{cases}$$

g: guess,
a: actual

  - Example: asymmetric loss

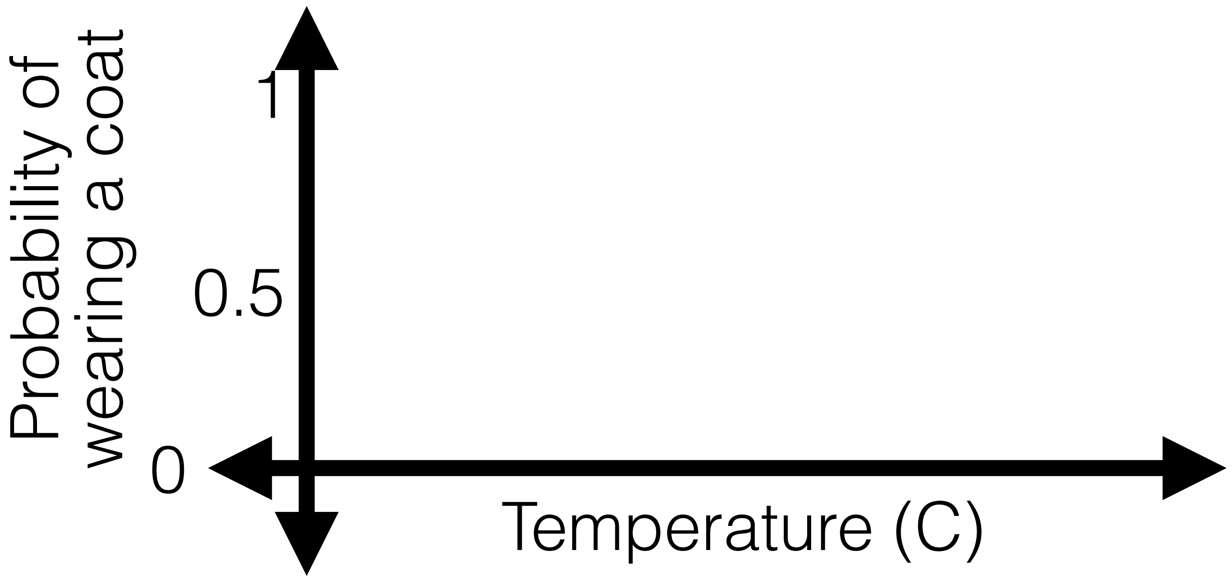
  - But: 0-1 loss & linear classifiers don't have a notion of uncertainty (how well do we know what we know?)



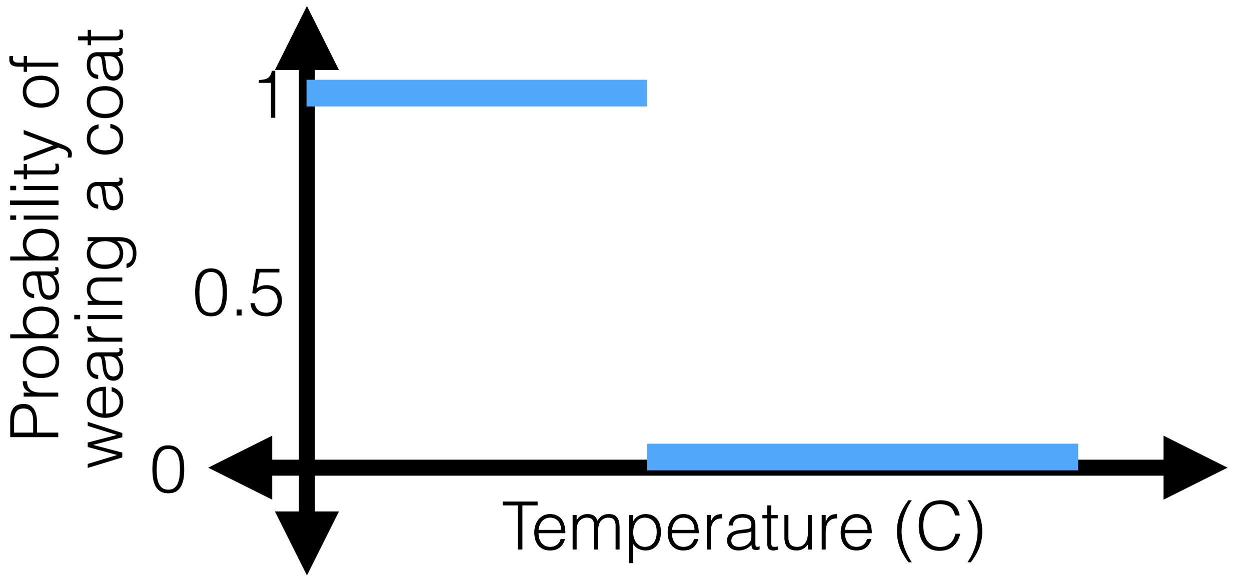


Capturing uncertainty

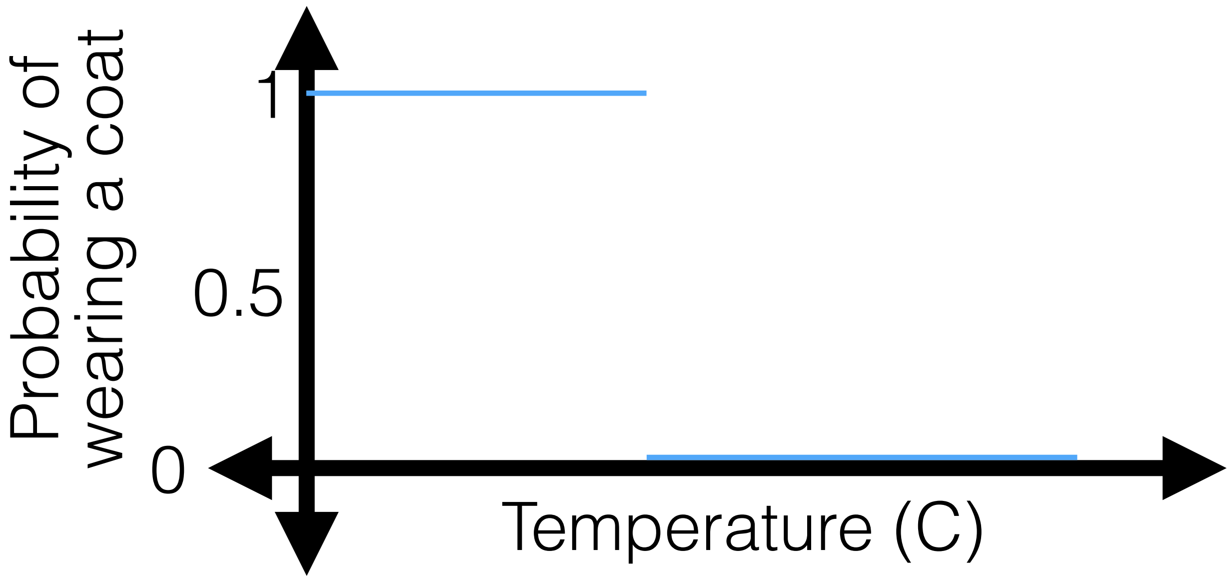
Capturing uncertainty



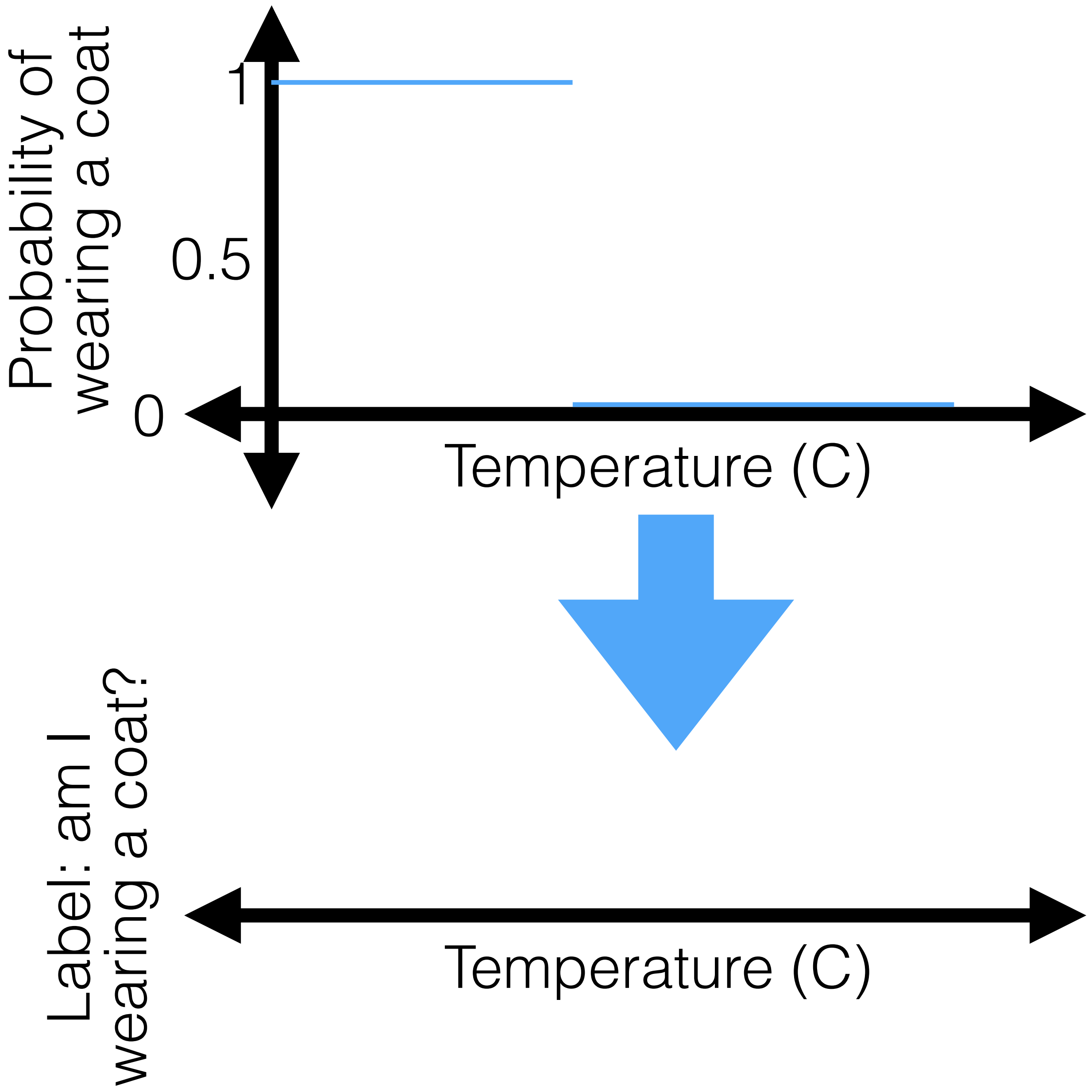
Capturing uncertainty



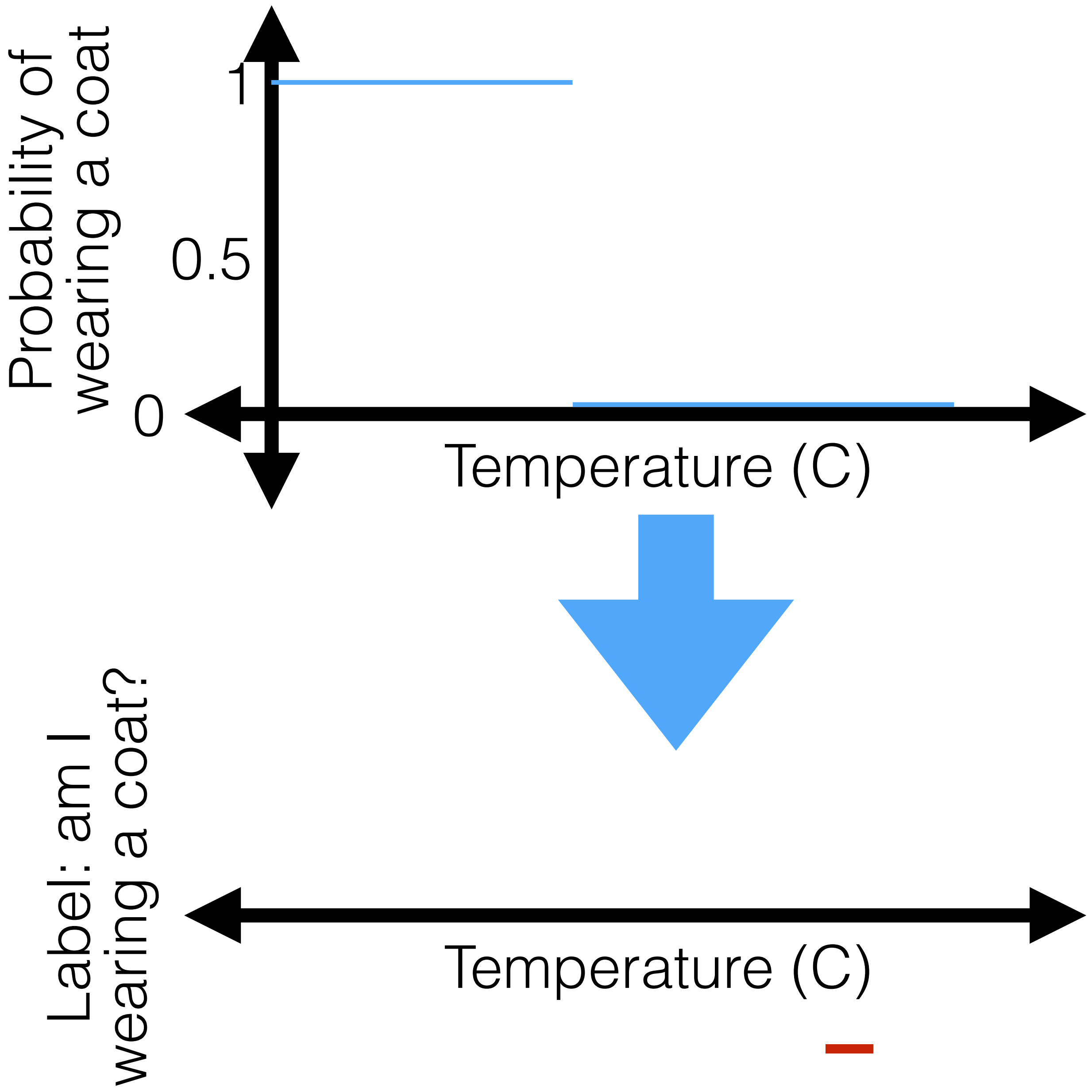
Capturing uncertainty



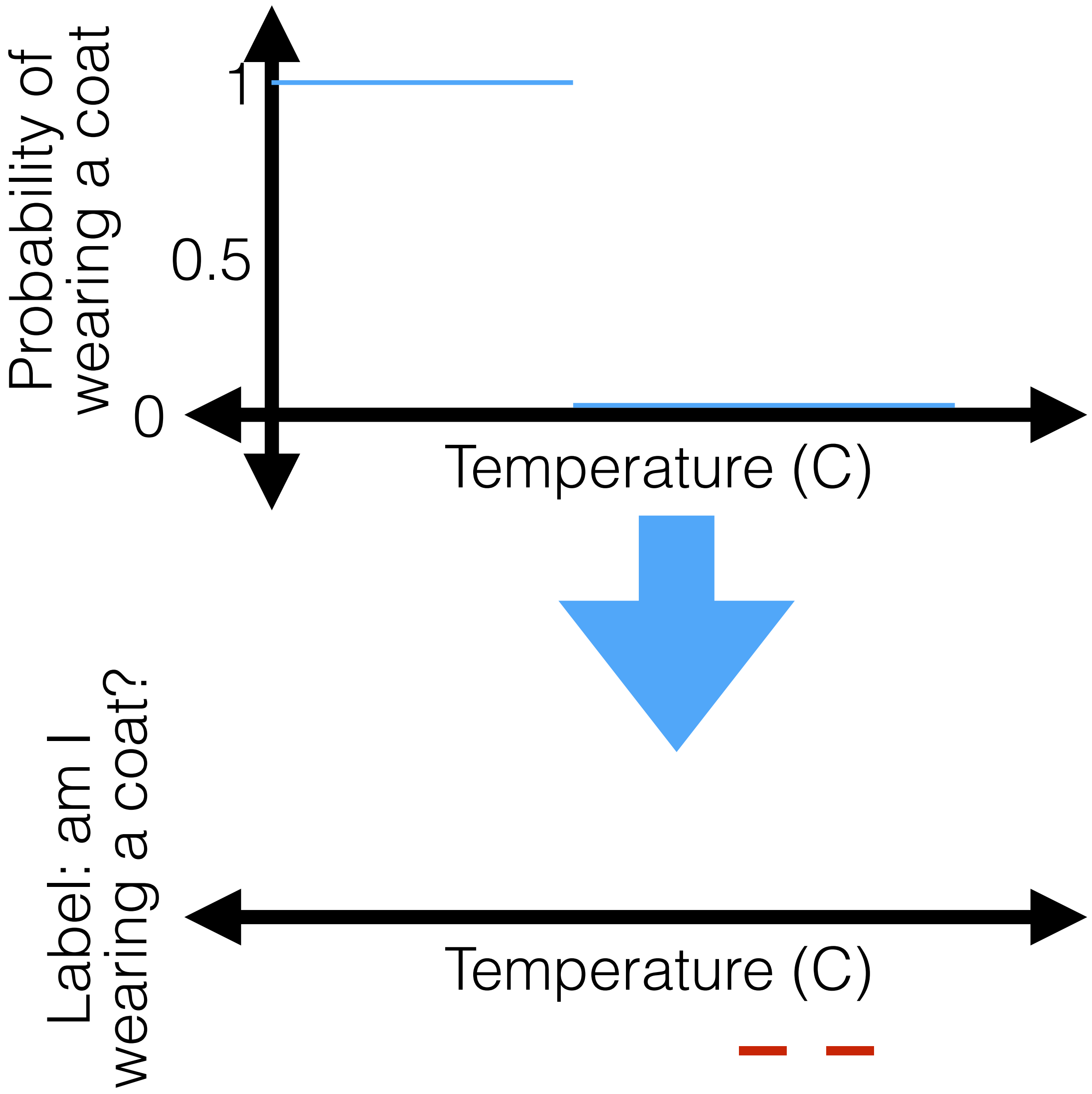
Capturing uncertainty



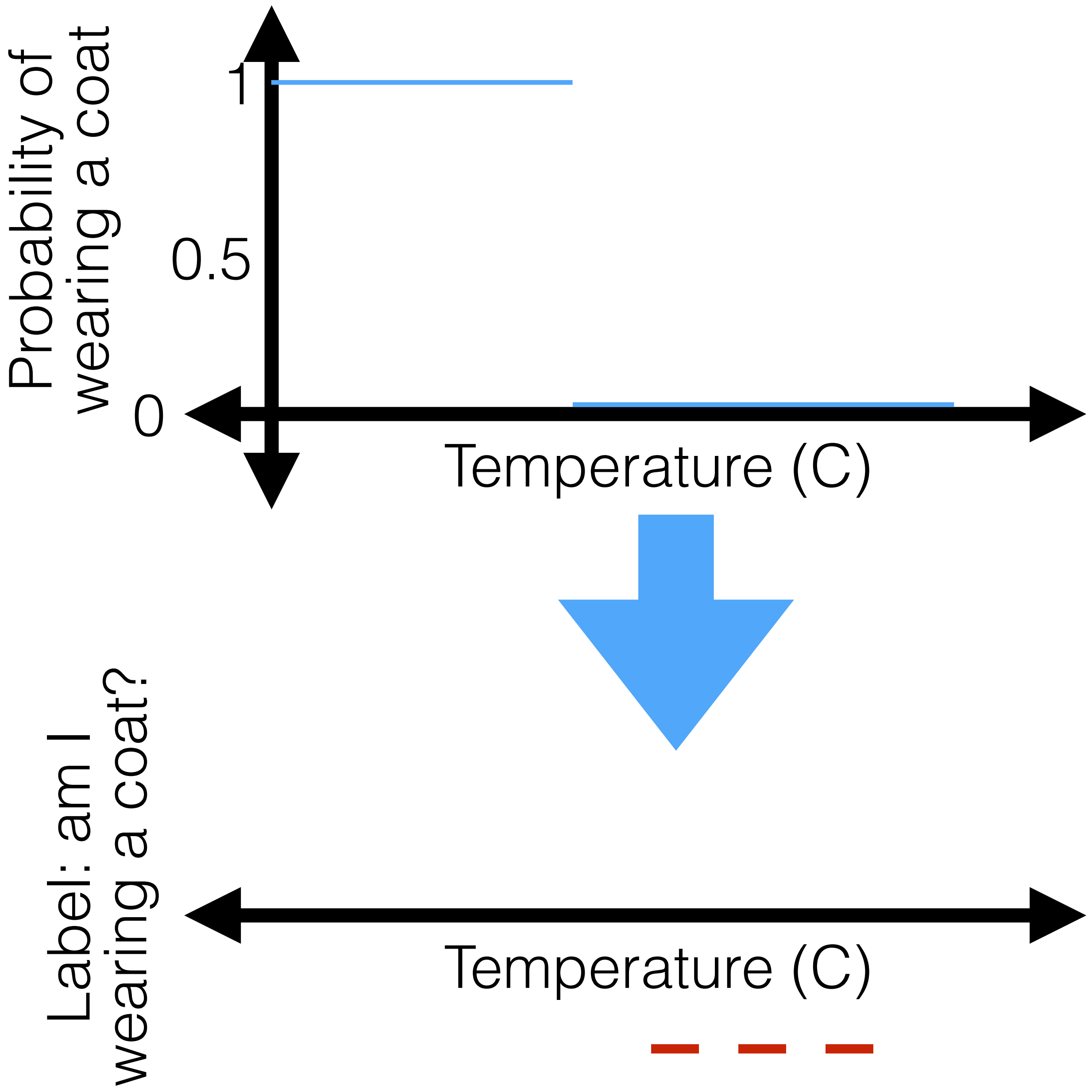
Capturing uncertainty



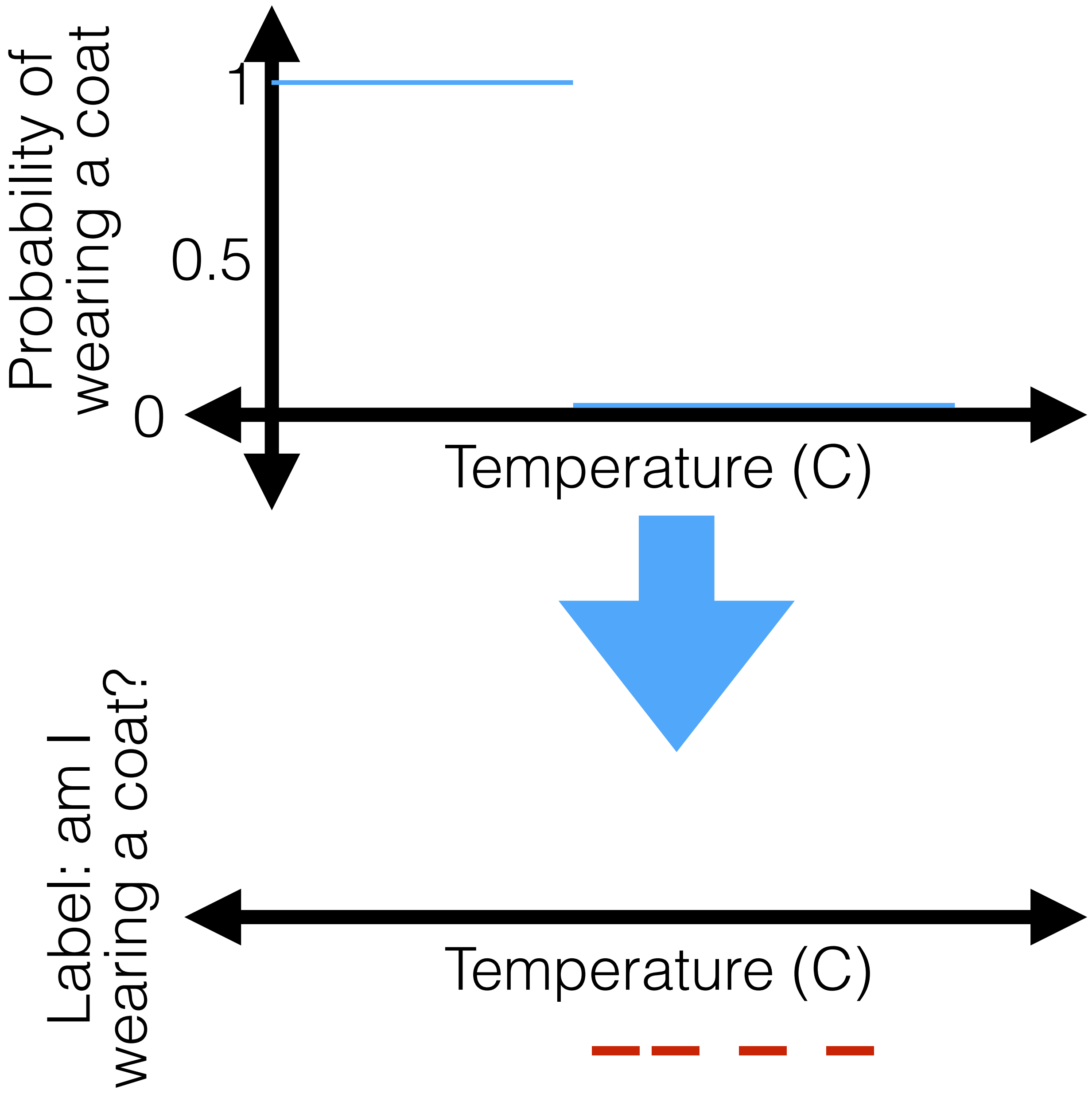
Capturing uncertainty



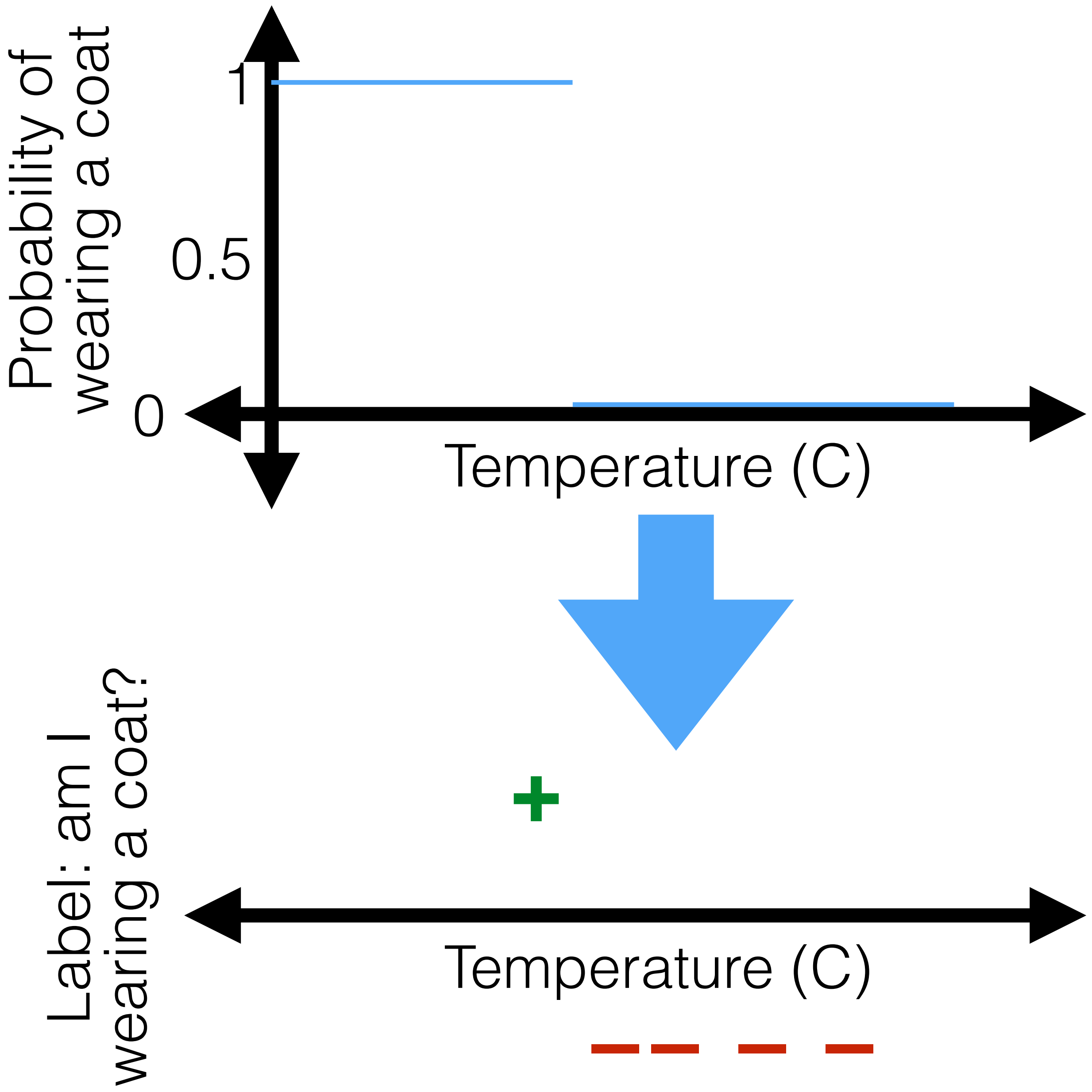
Capturing uncertainty



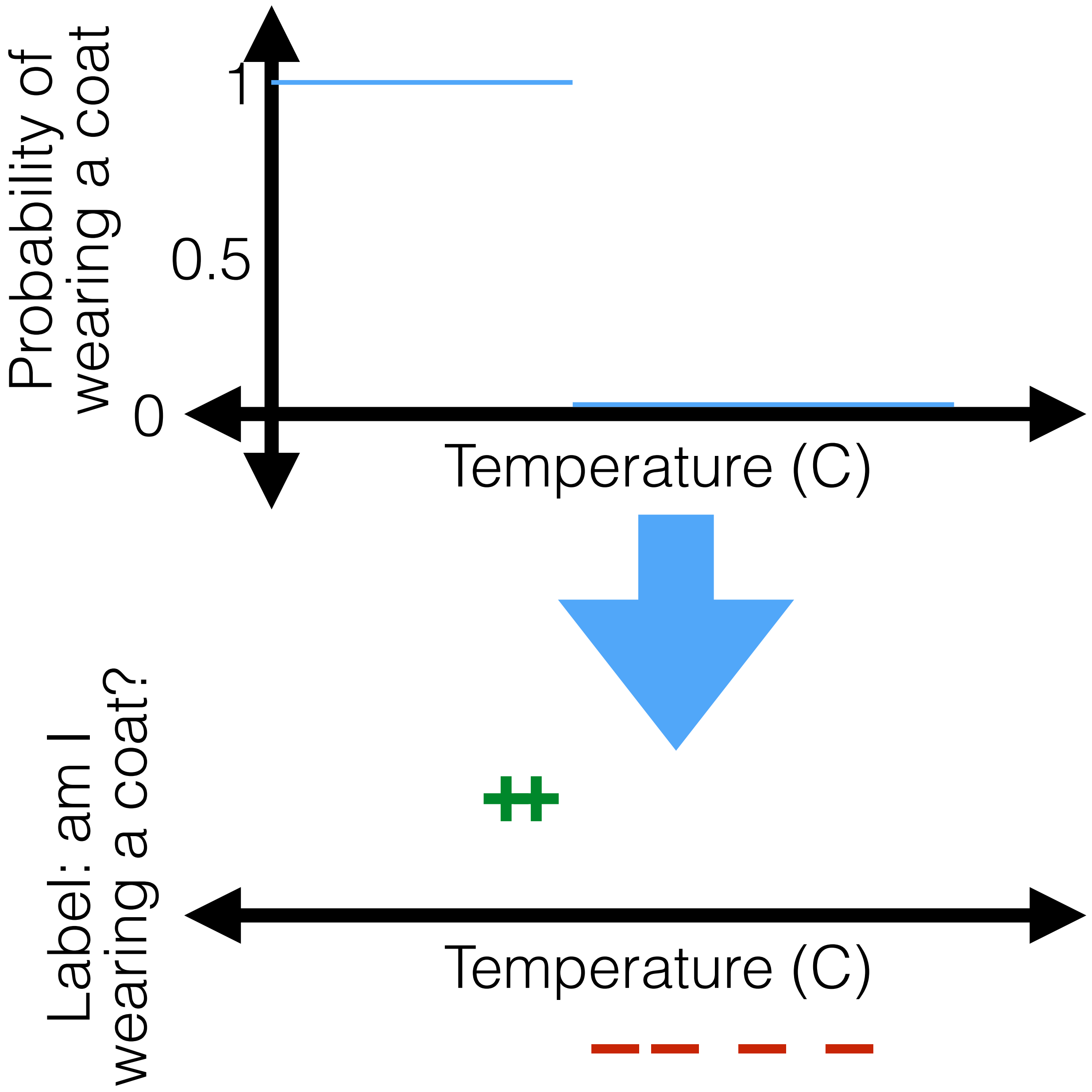
Capturing uncertainty



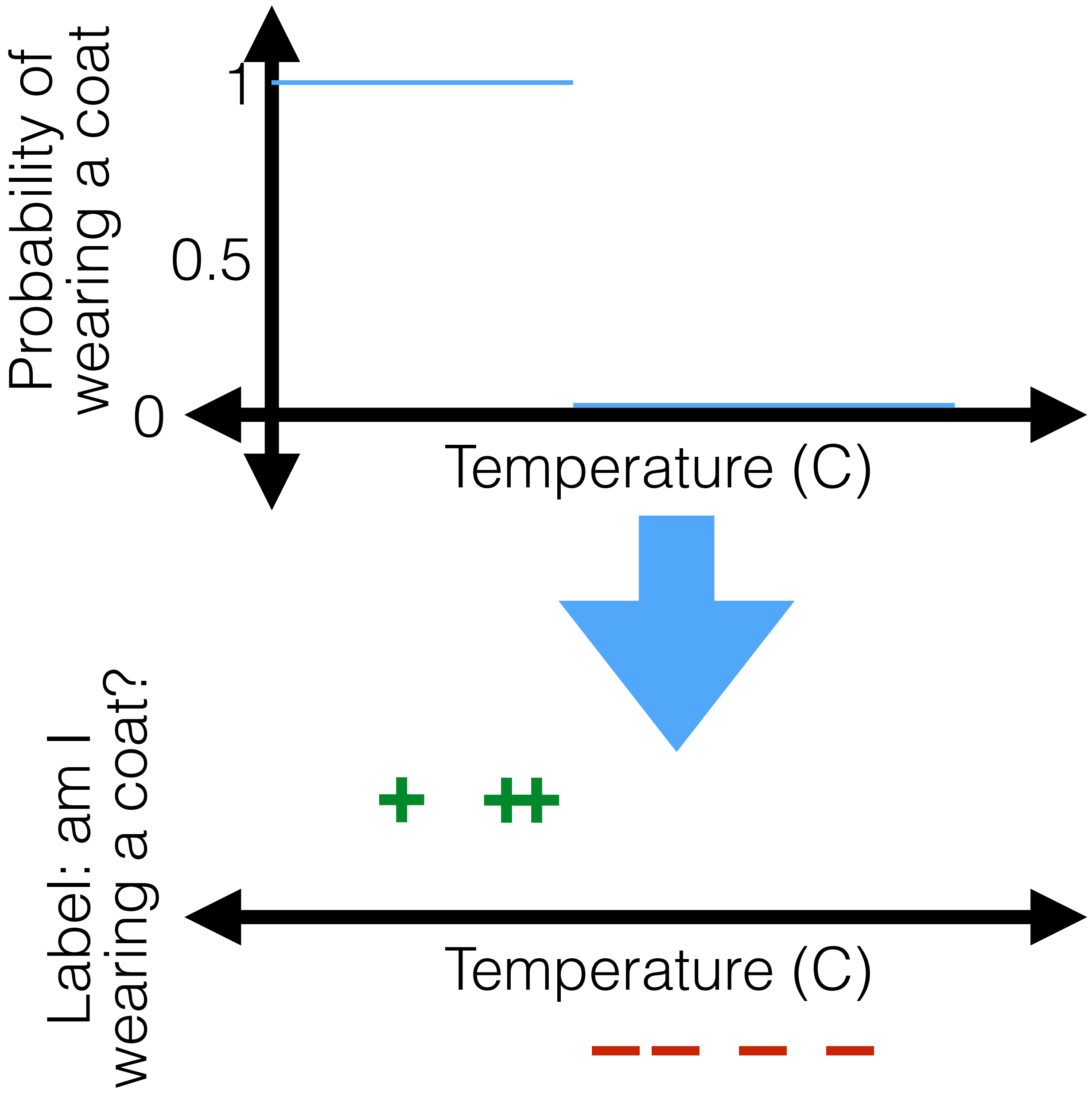
Capturing uncertainty



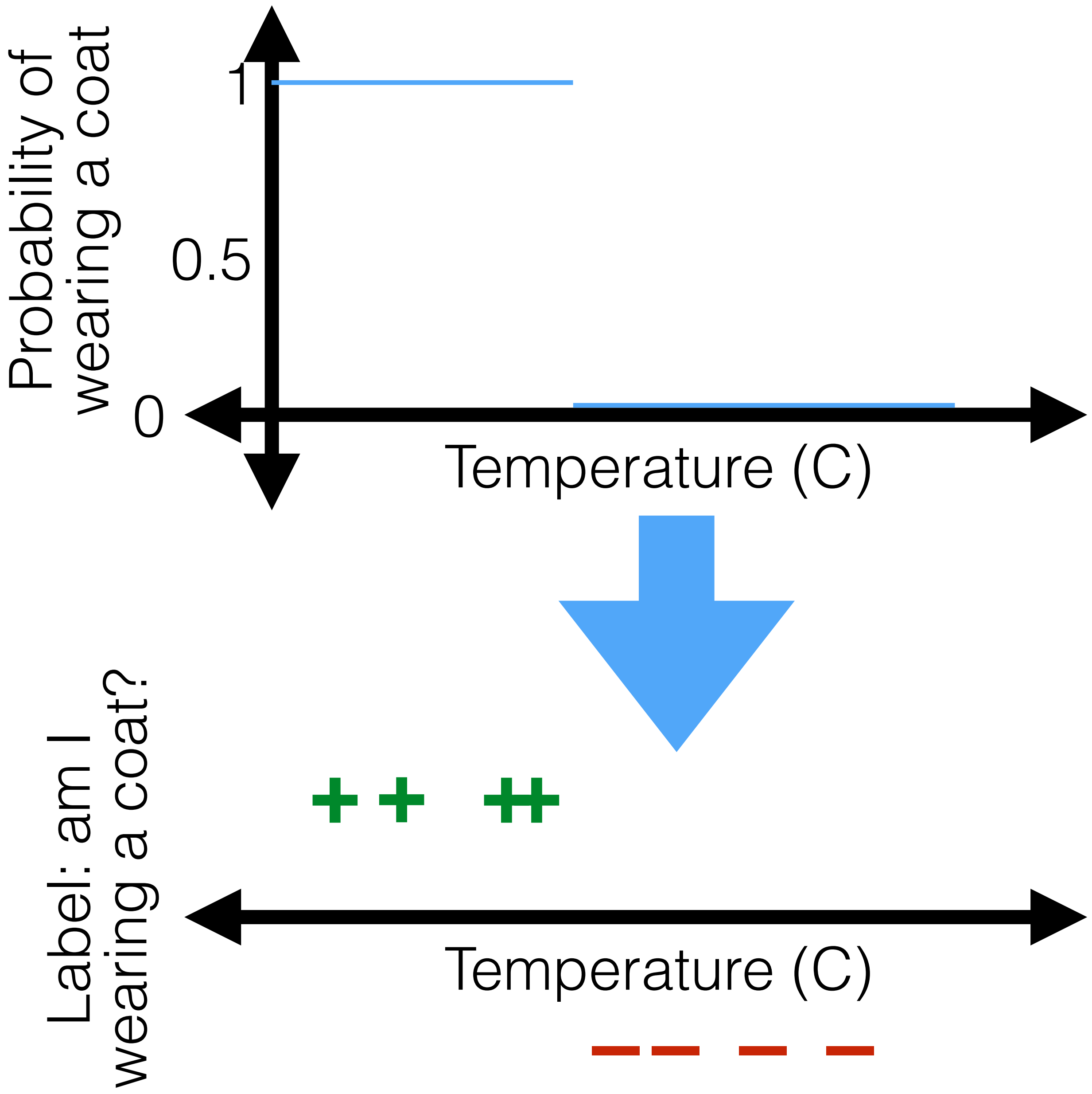
Capturing uncertainty



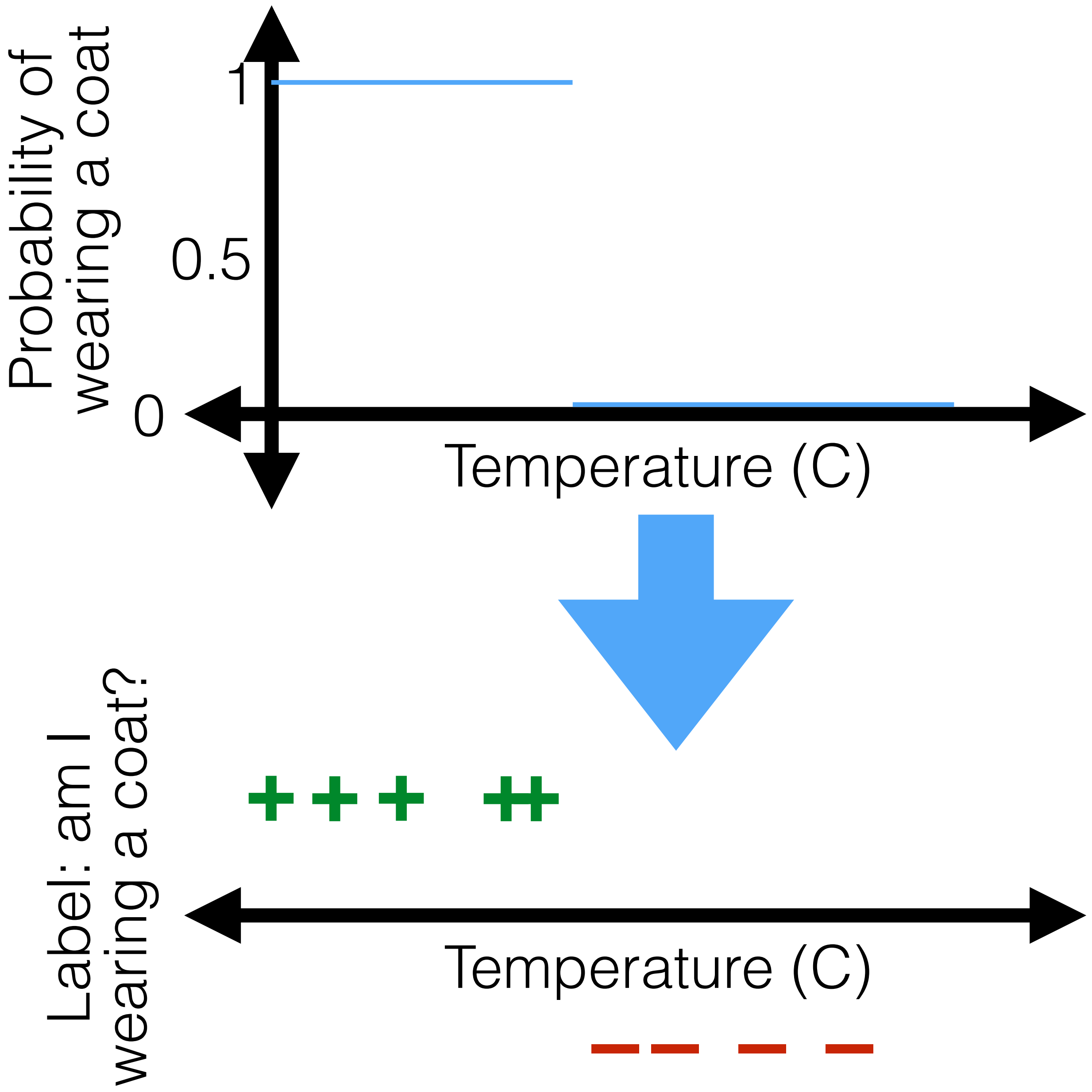
Capturing uncertainty



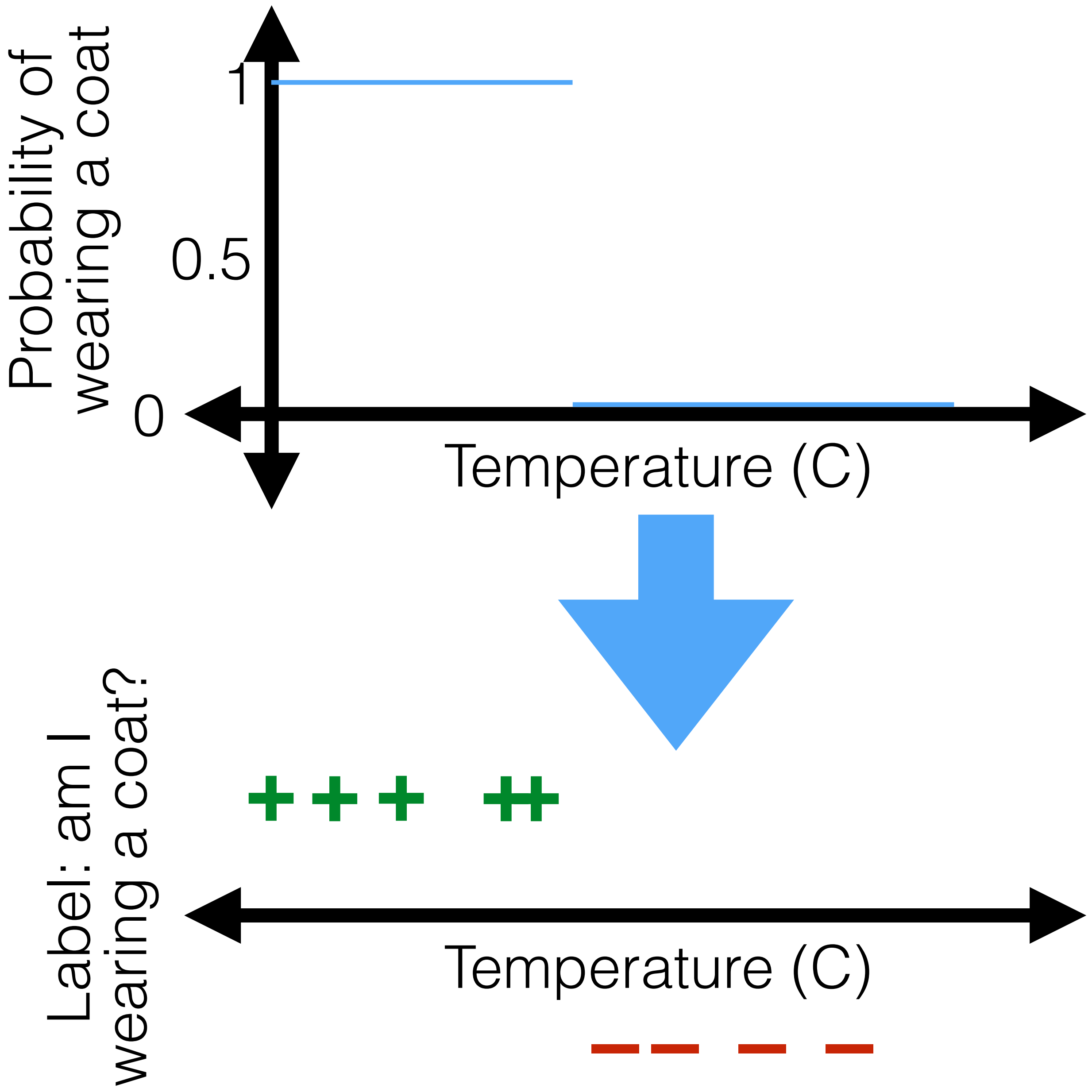
Capturing uncertainty



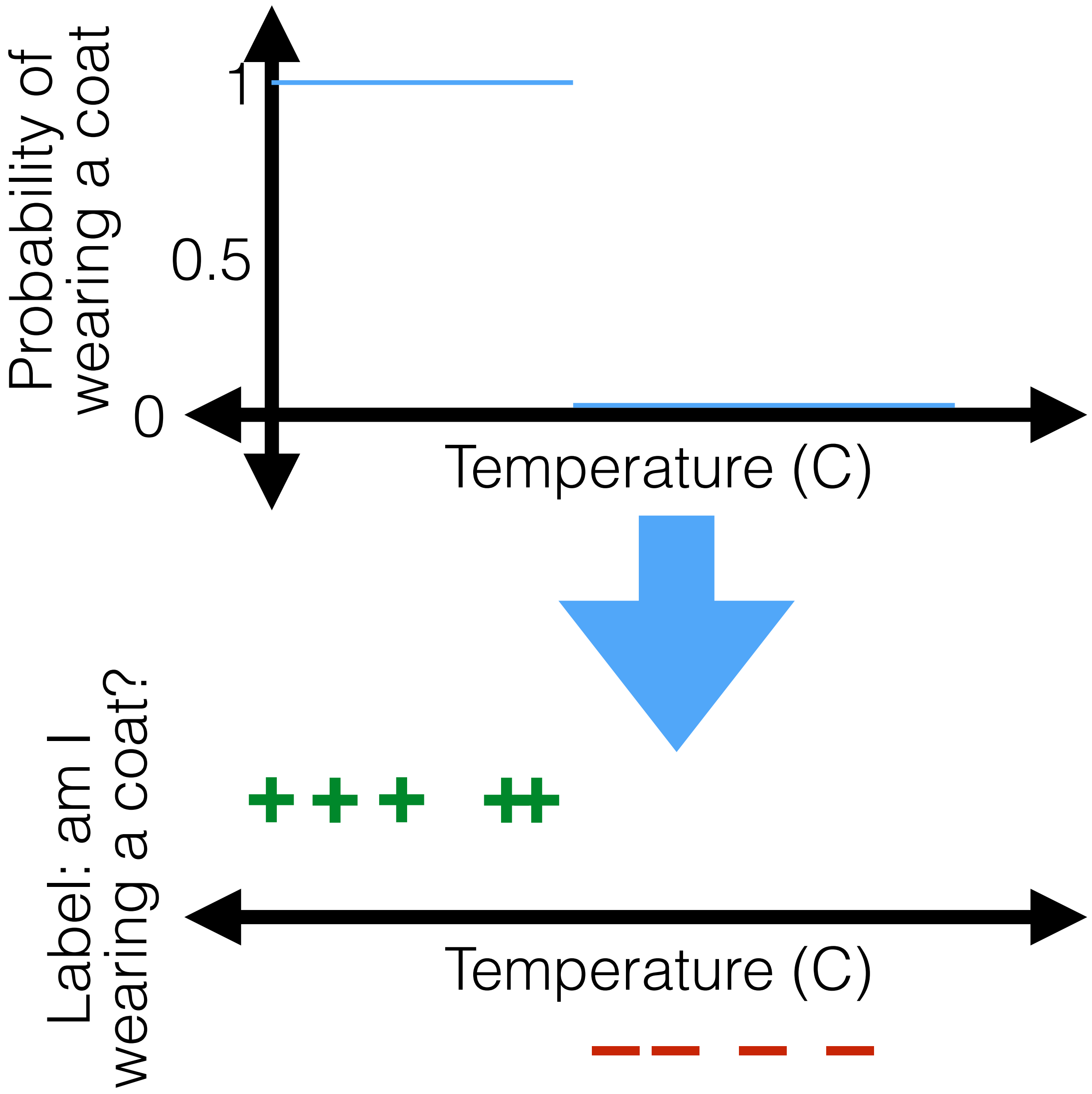
Capturing uncertainty

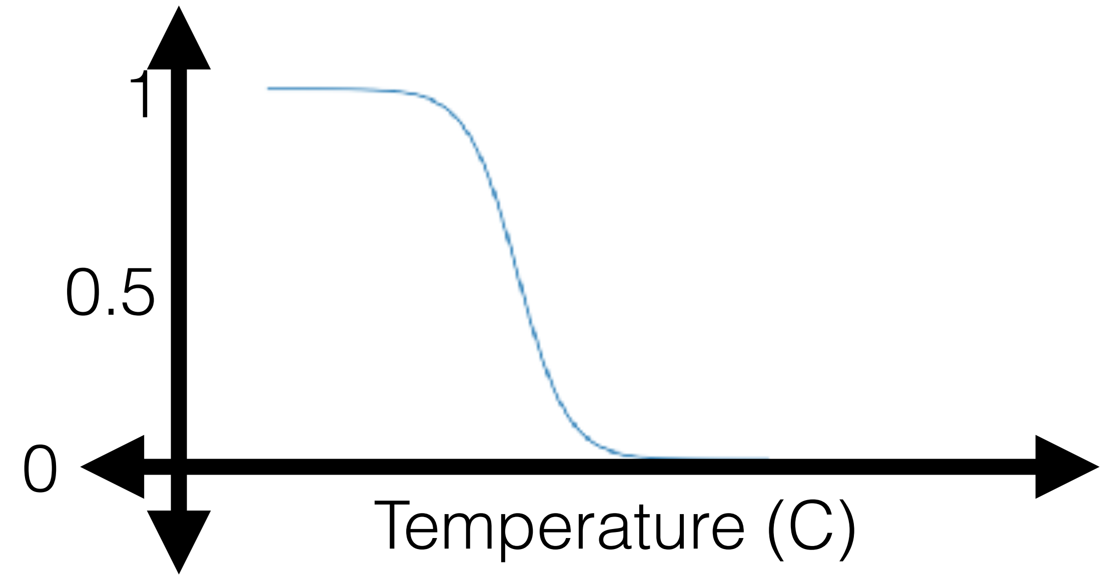


Capturing uncertainty

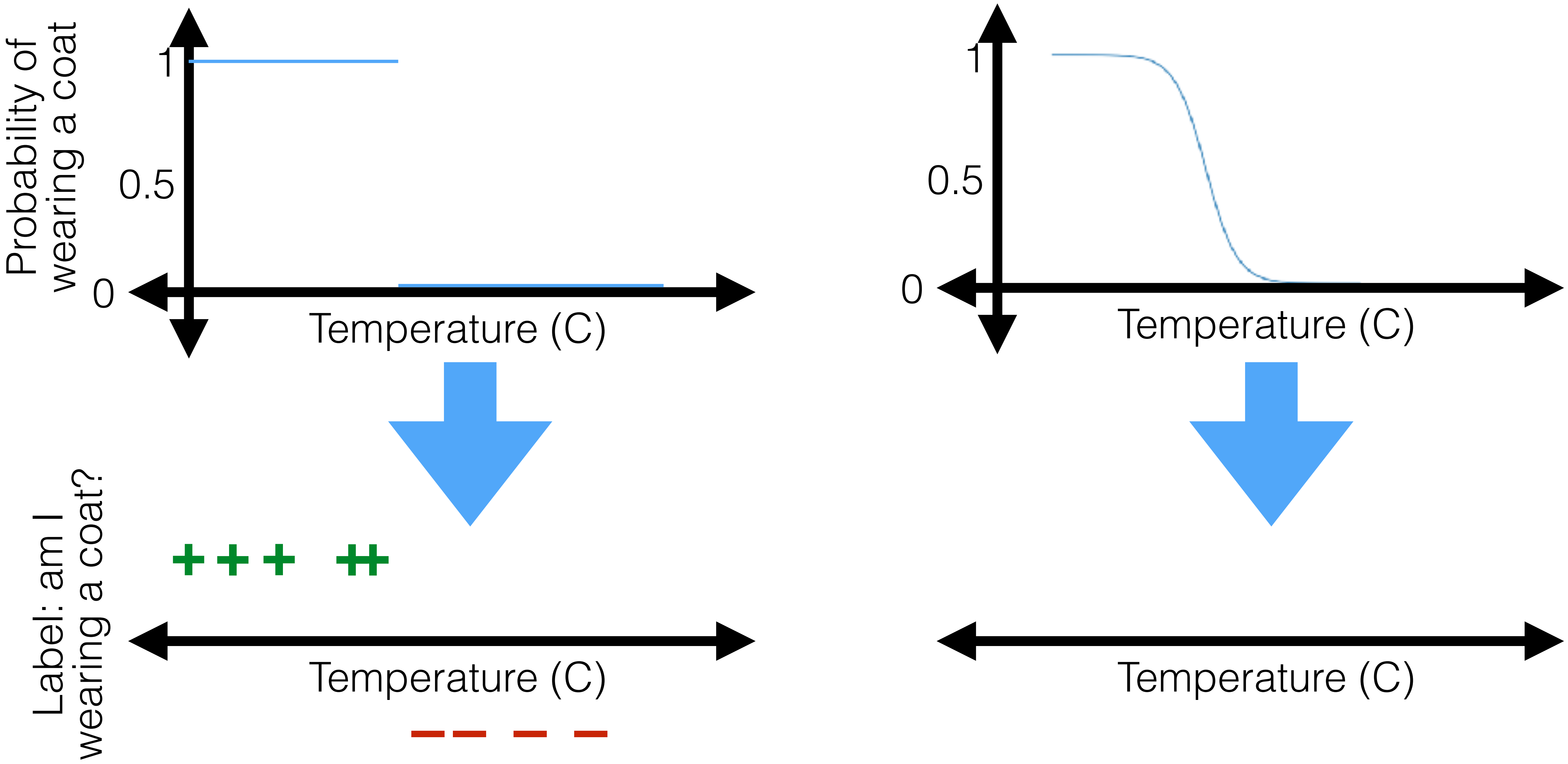


Capturing uncertainty

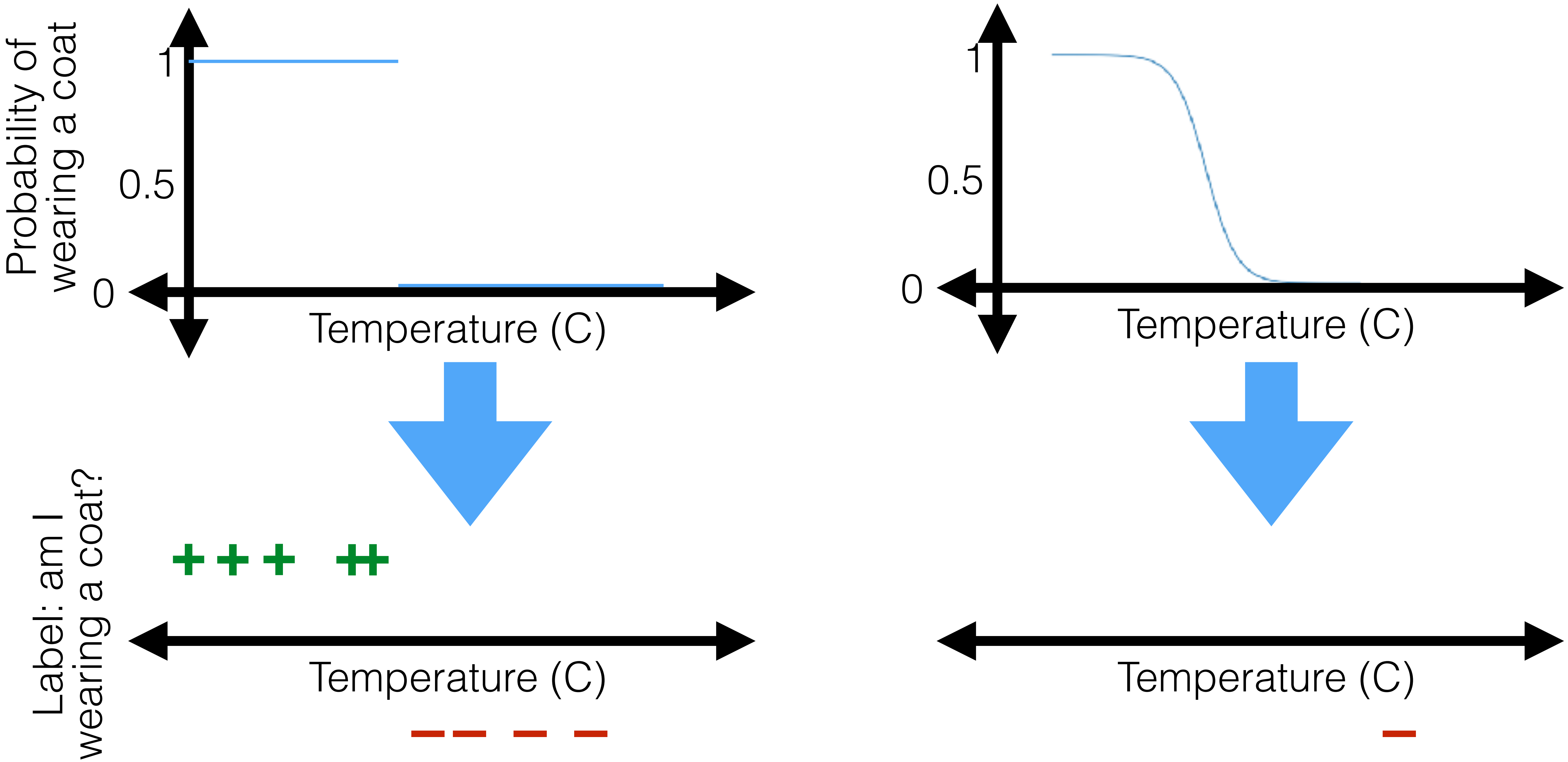




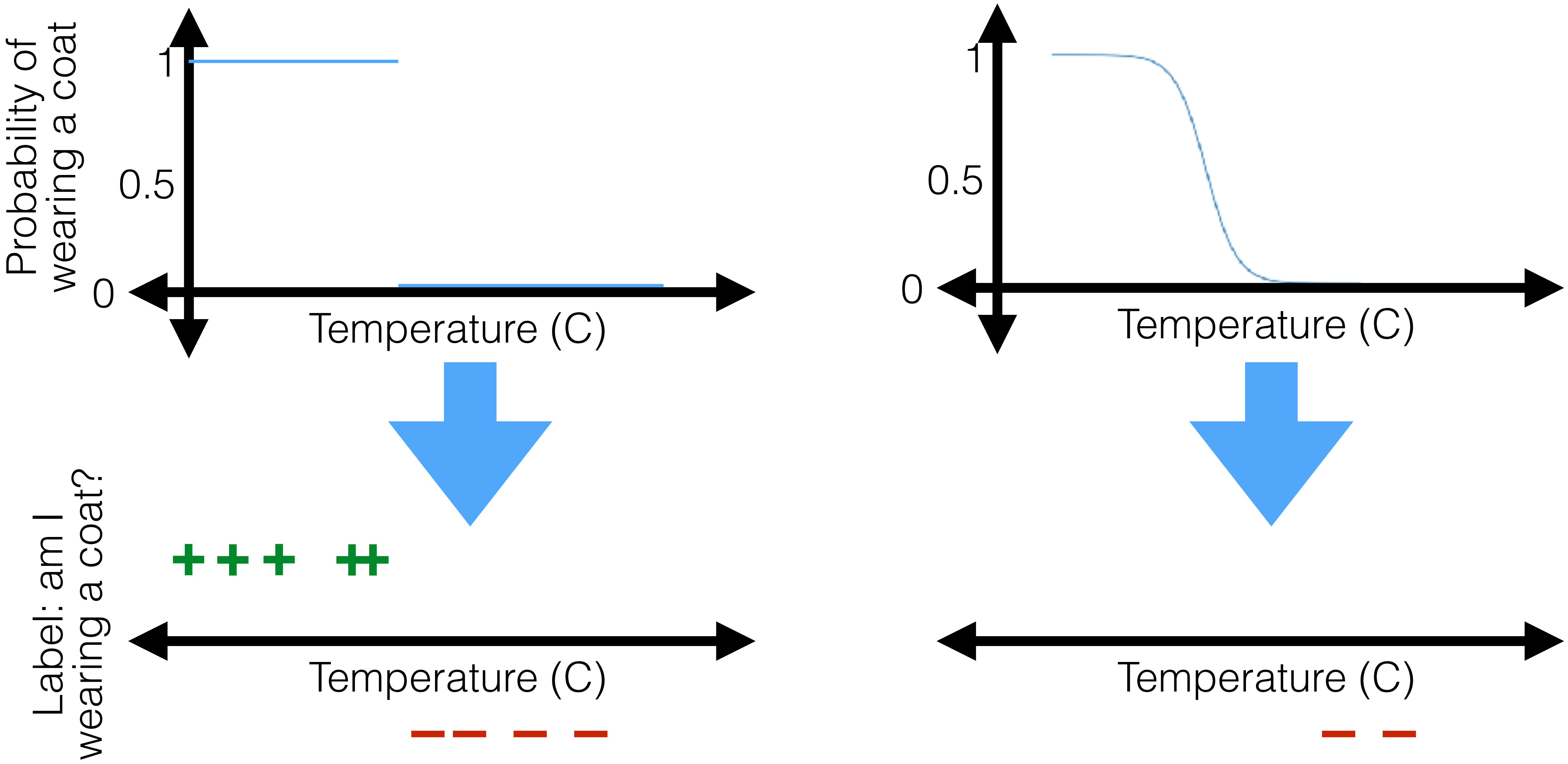
Capturing uncertainty



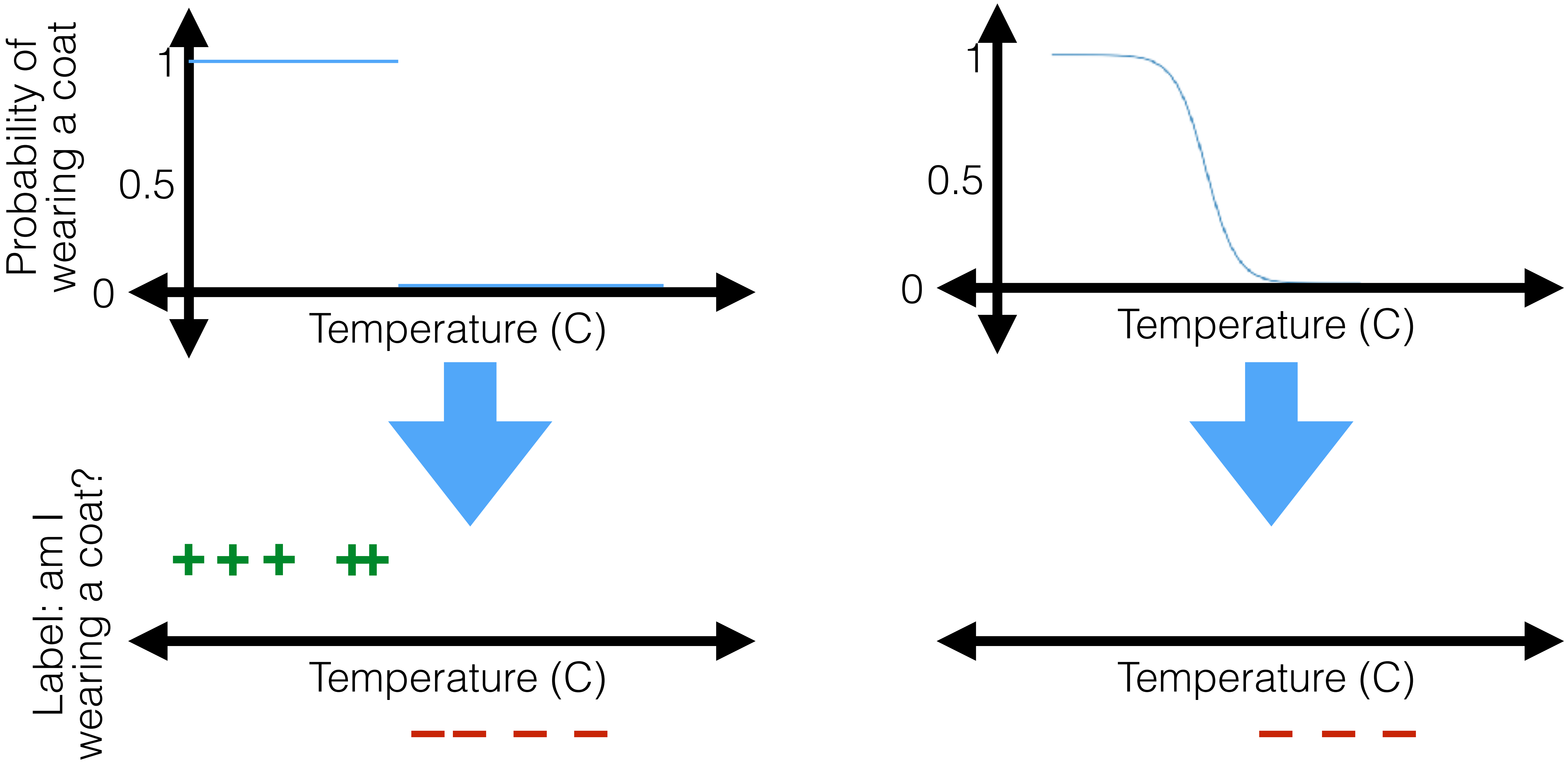
Capturing uncertainty



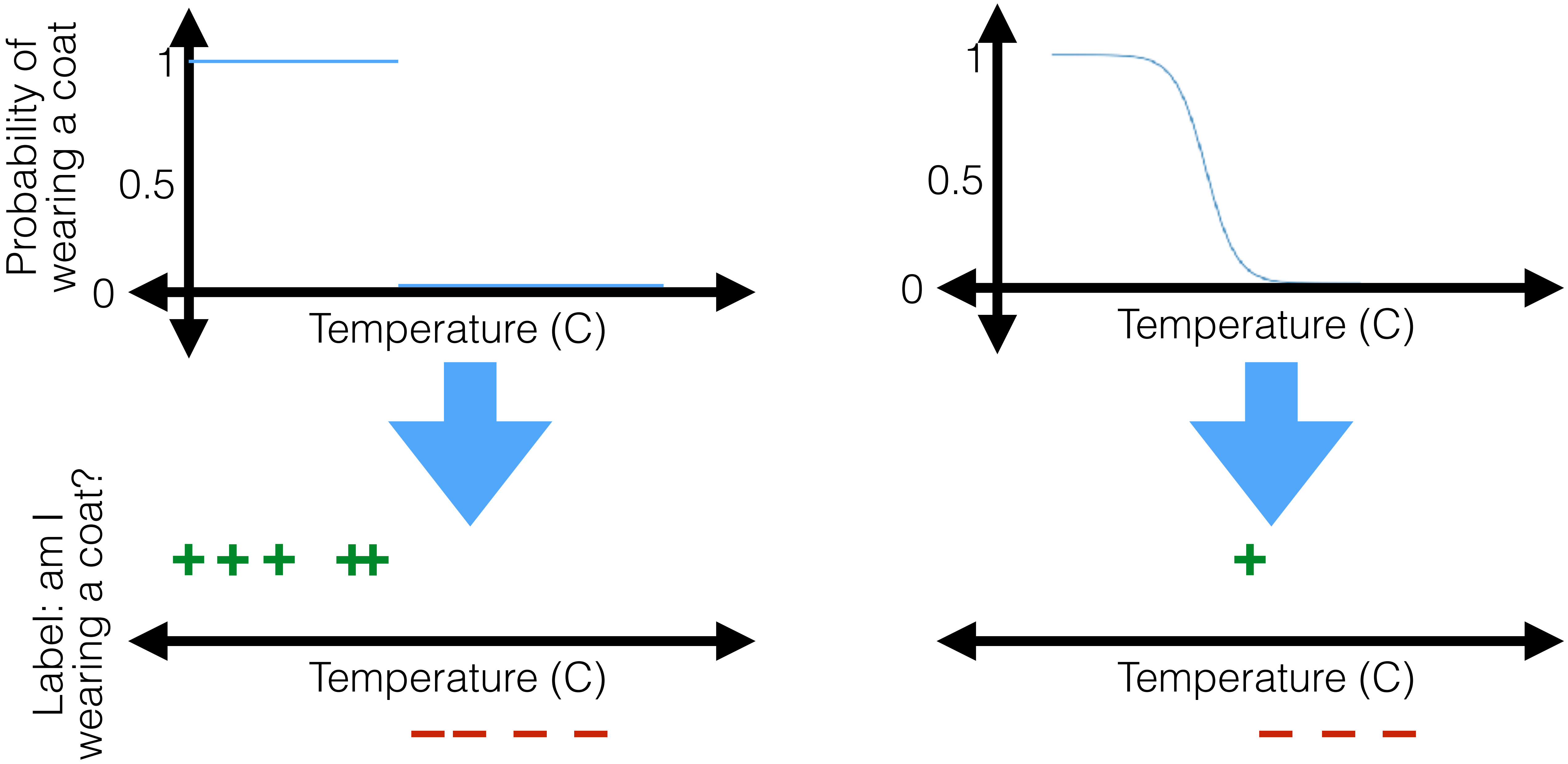
Capturing uncertainty



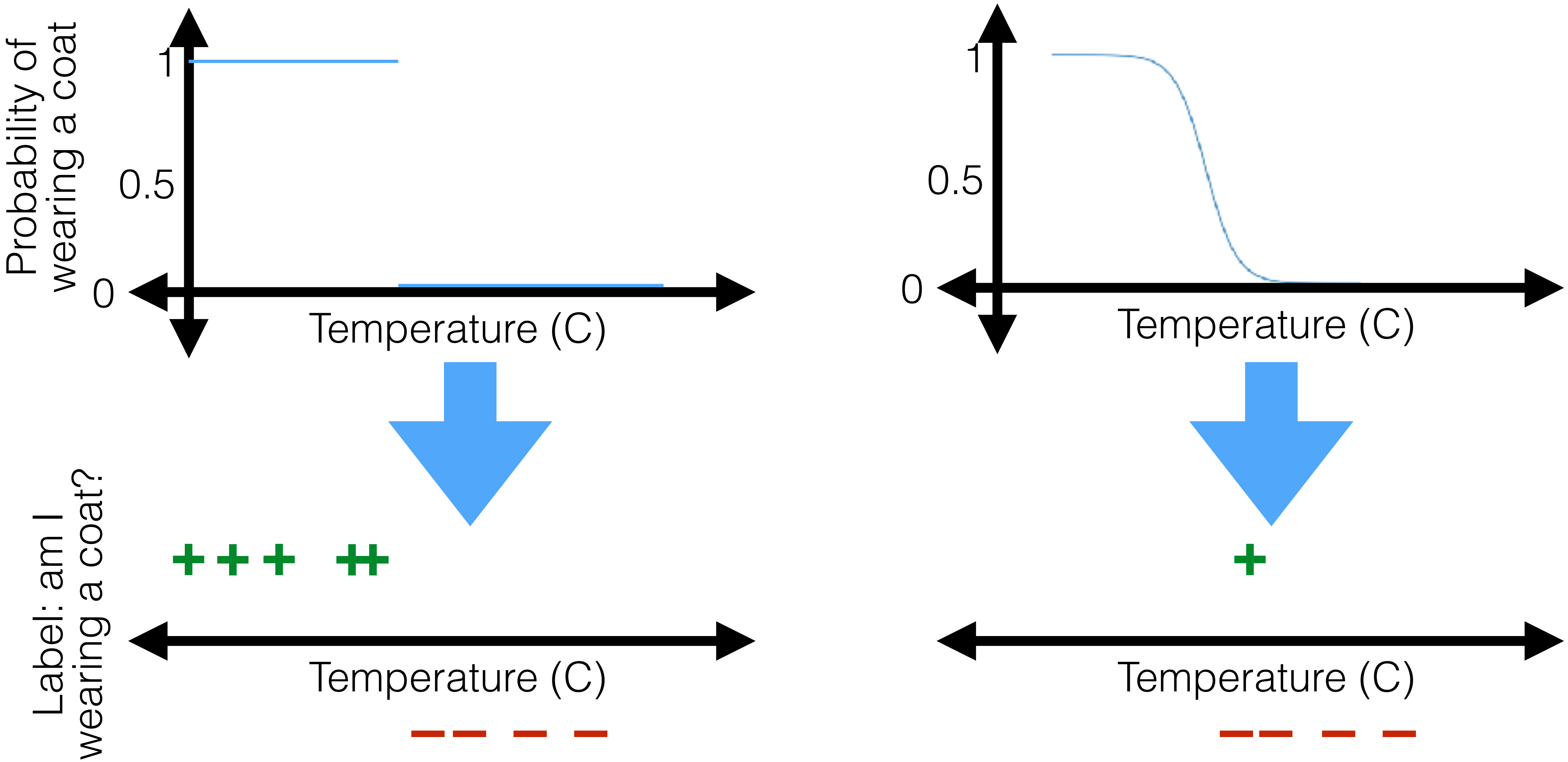
Capturing uncertainty



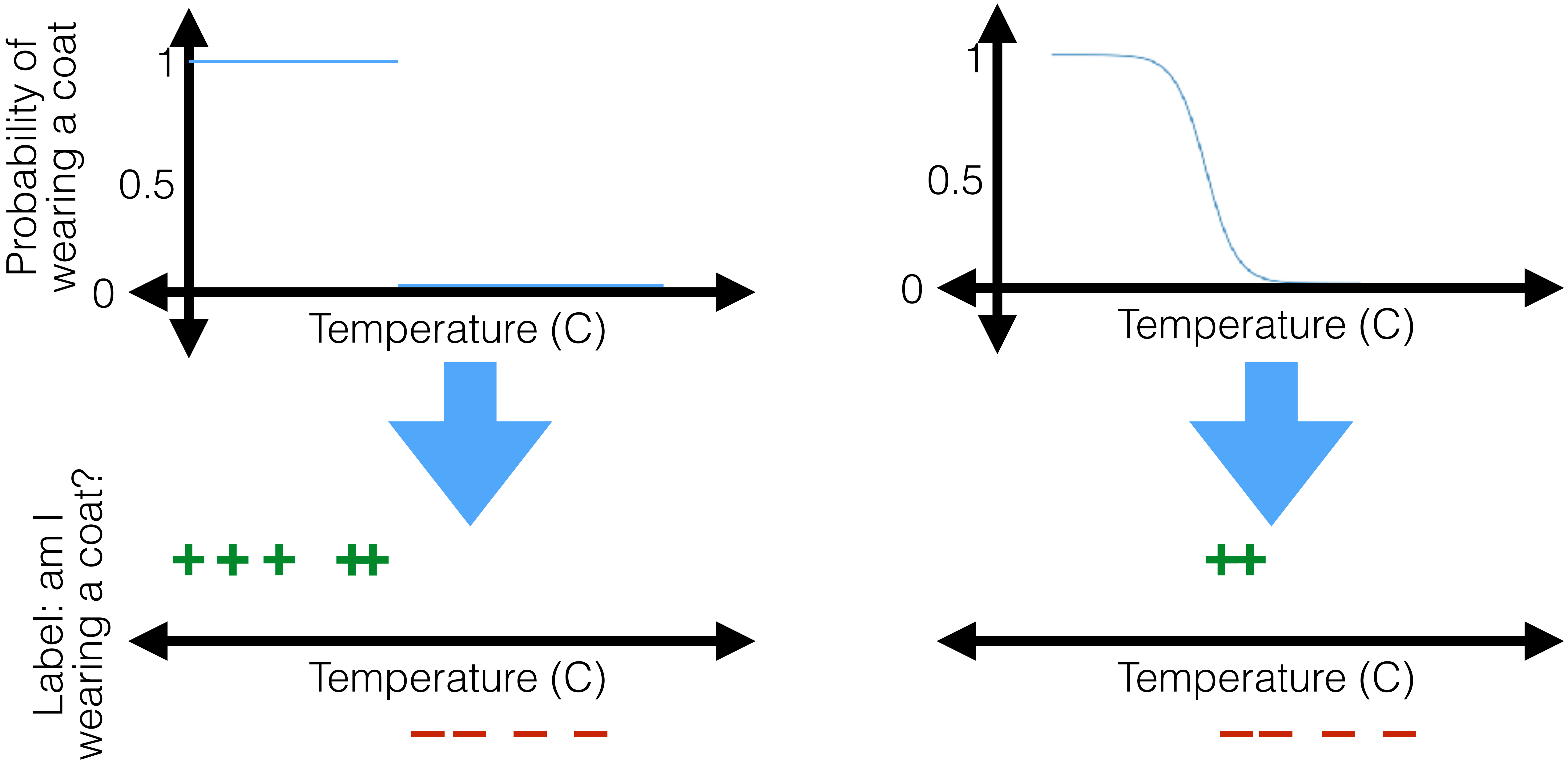
Capturing uncertainty



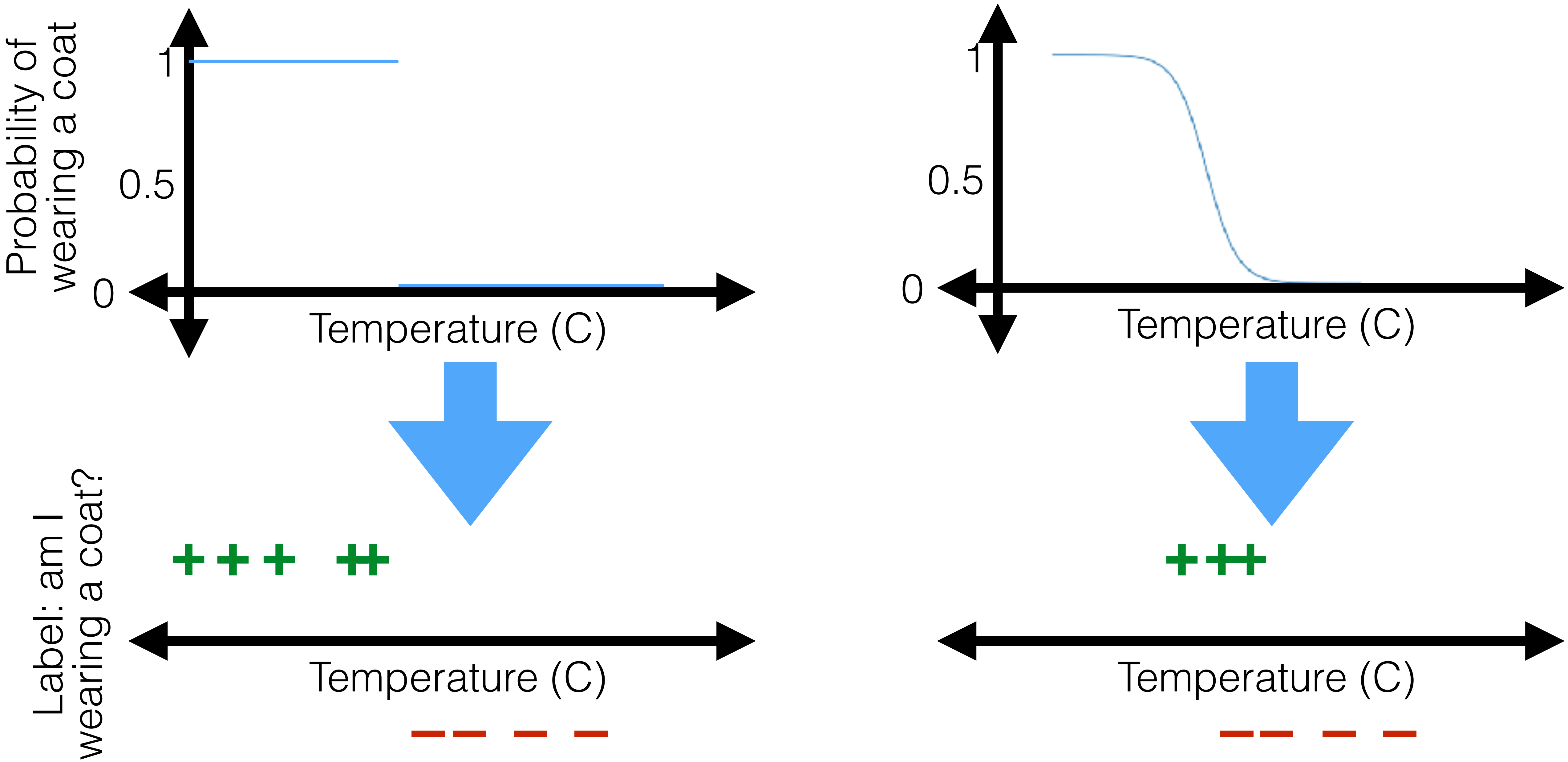
Capturing uncertainty



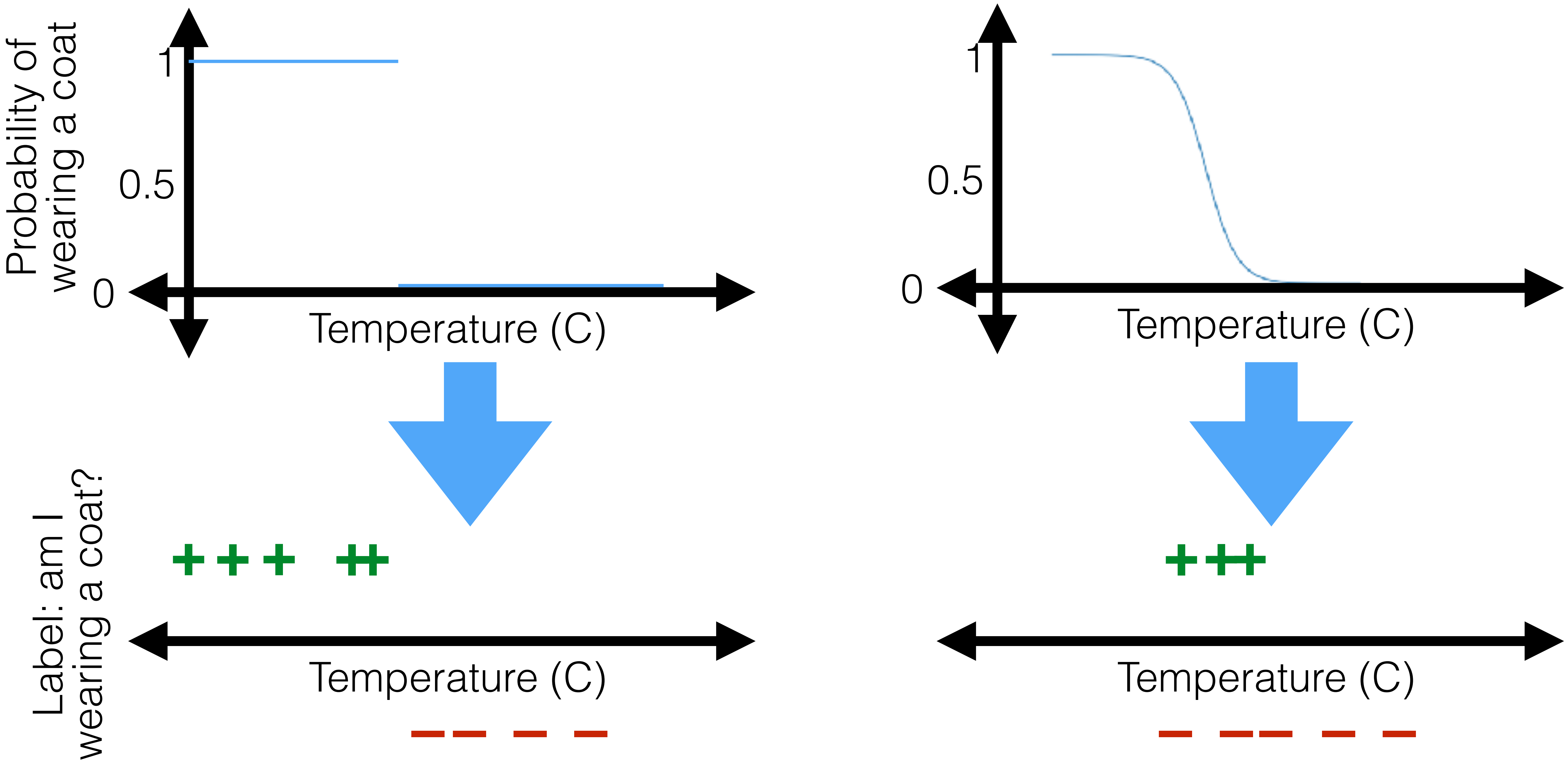
Capturing uncertainty



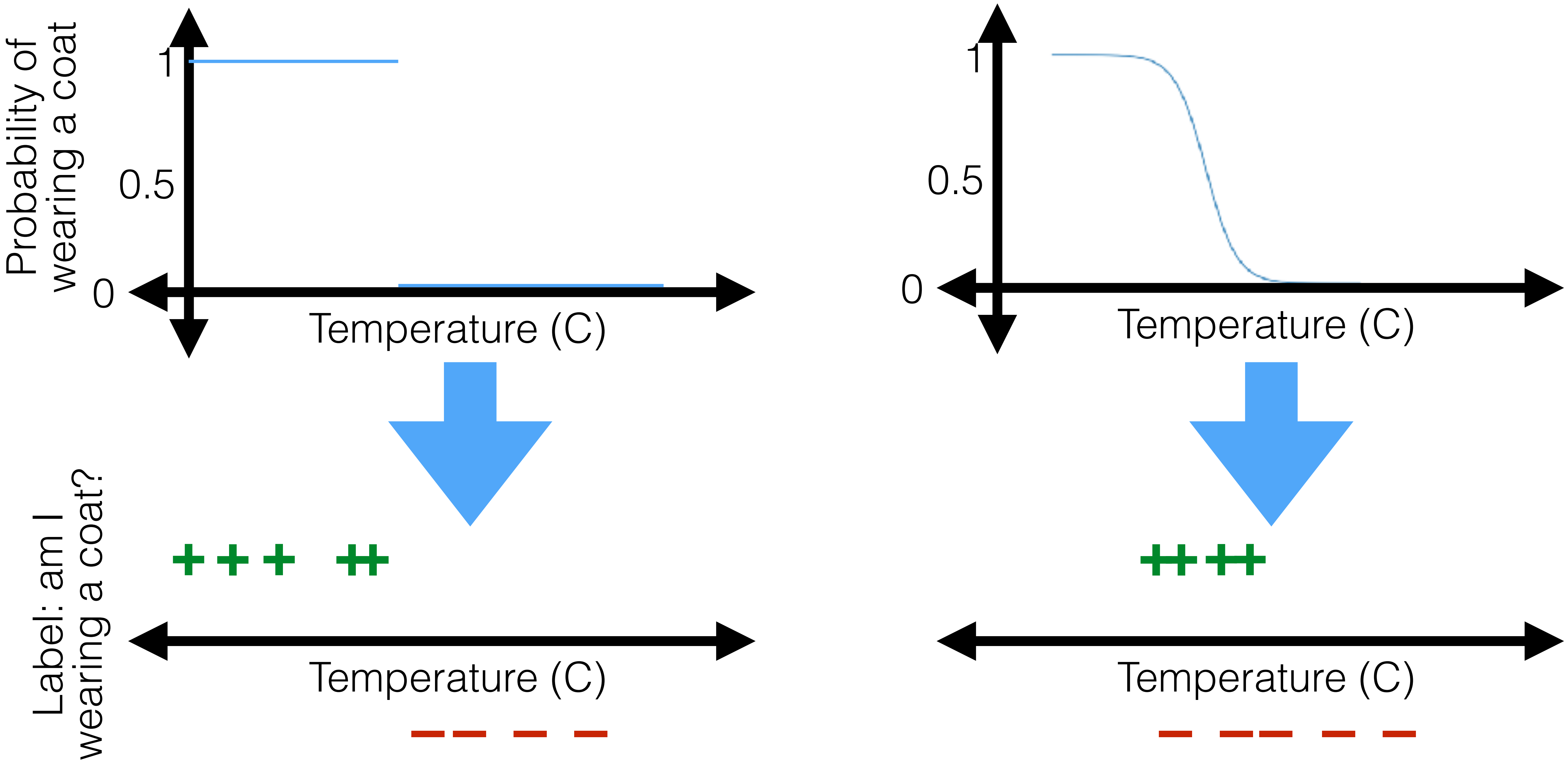
Capturing uncertainty



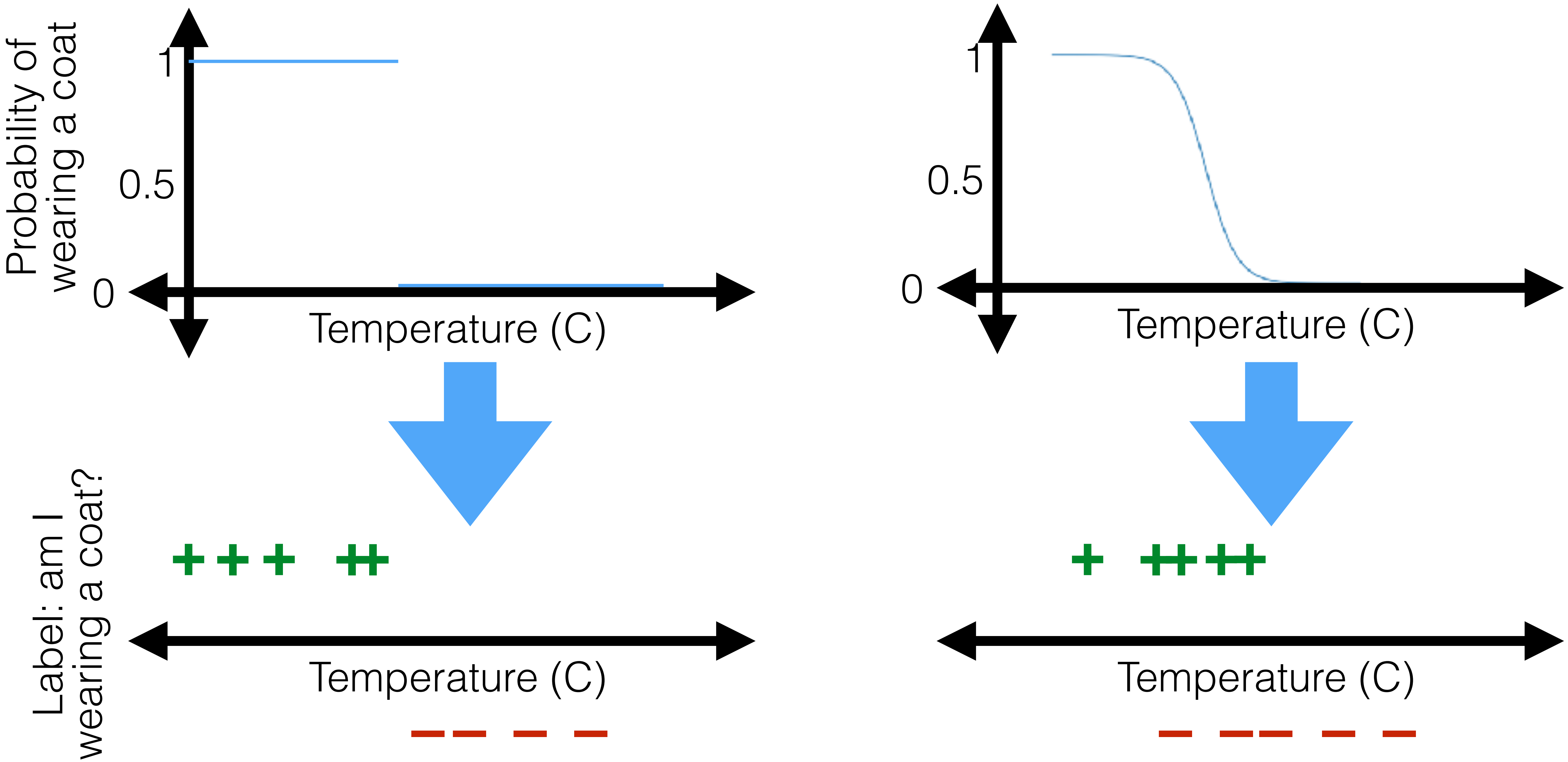
Capturing uncertainty



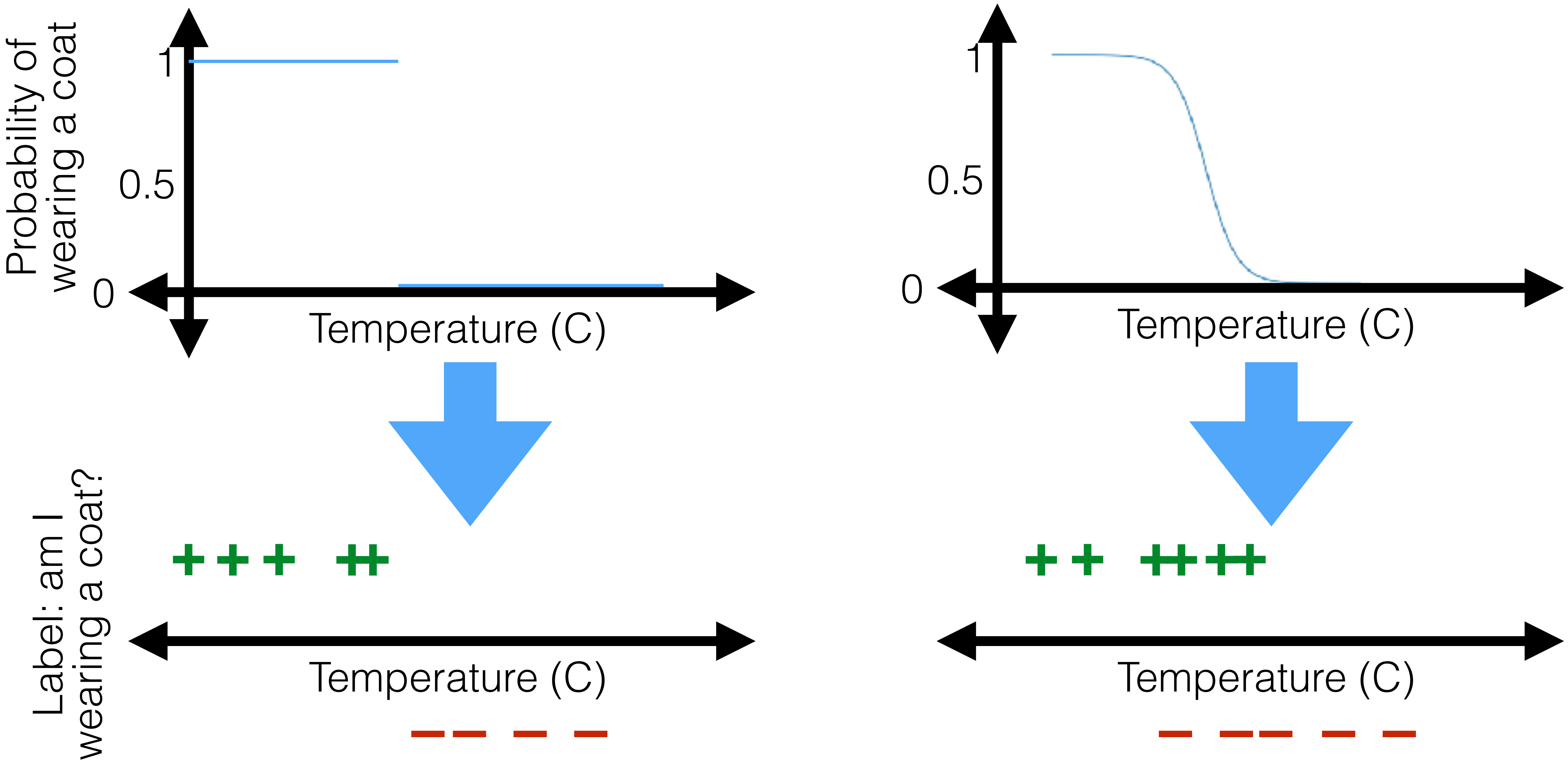
Capturing uncertainty



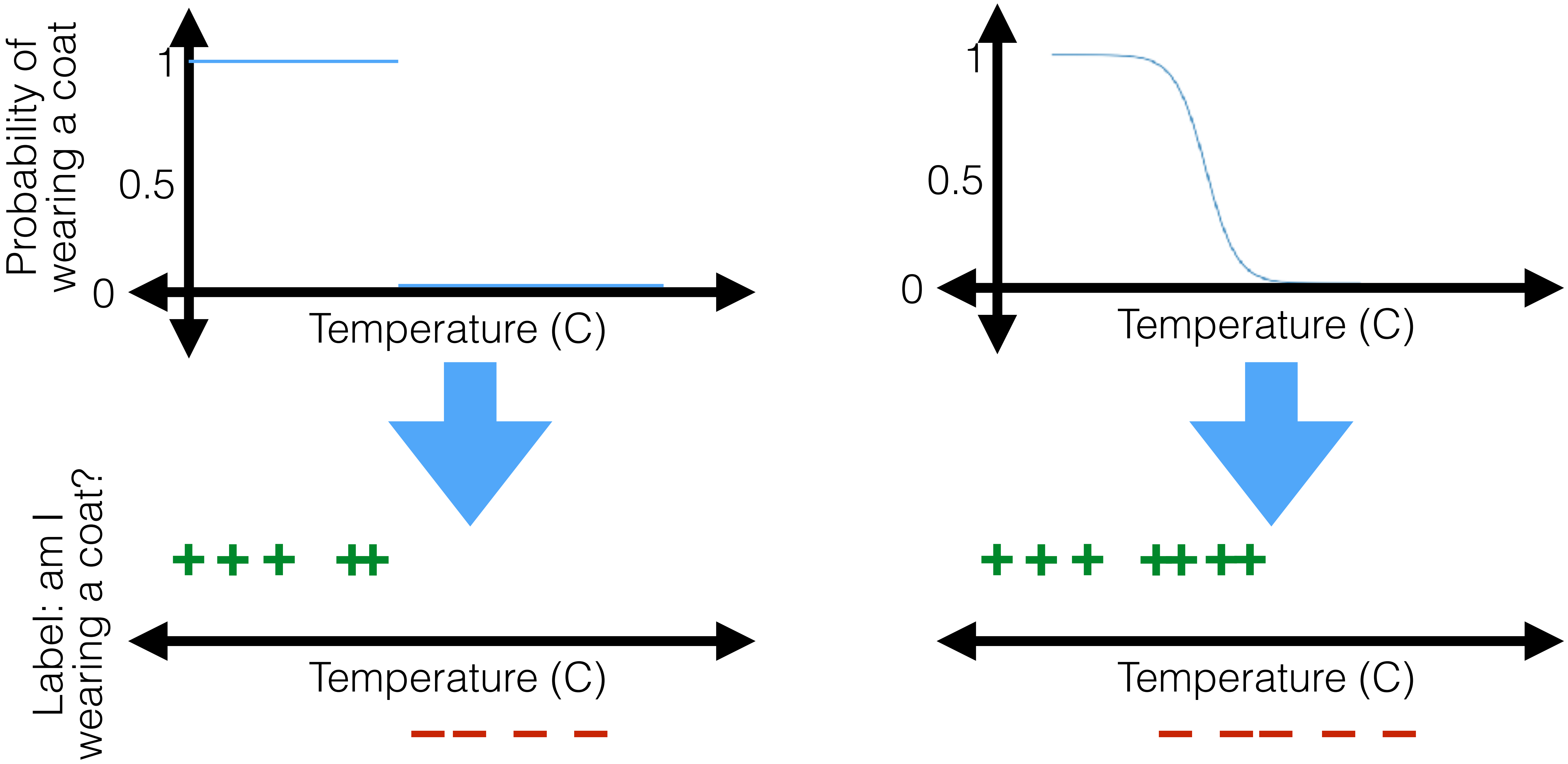
Capturing uncertainty



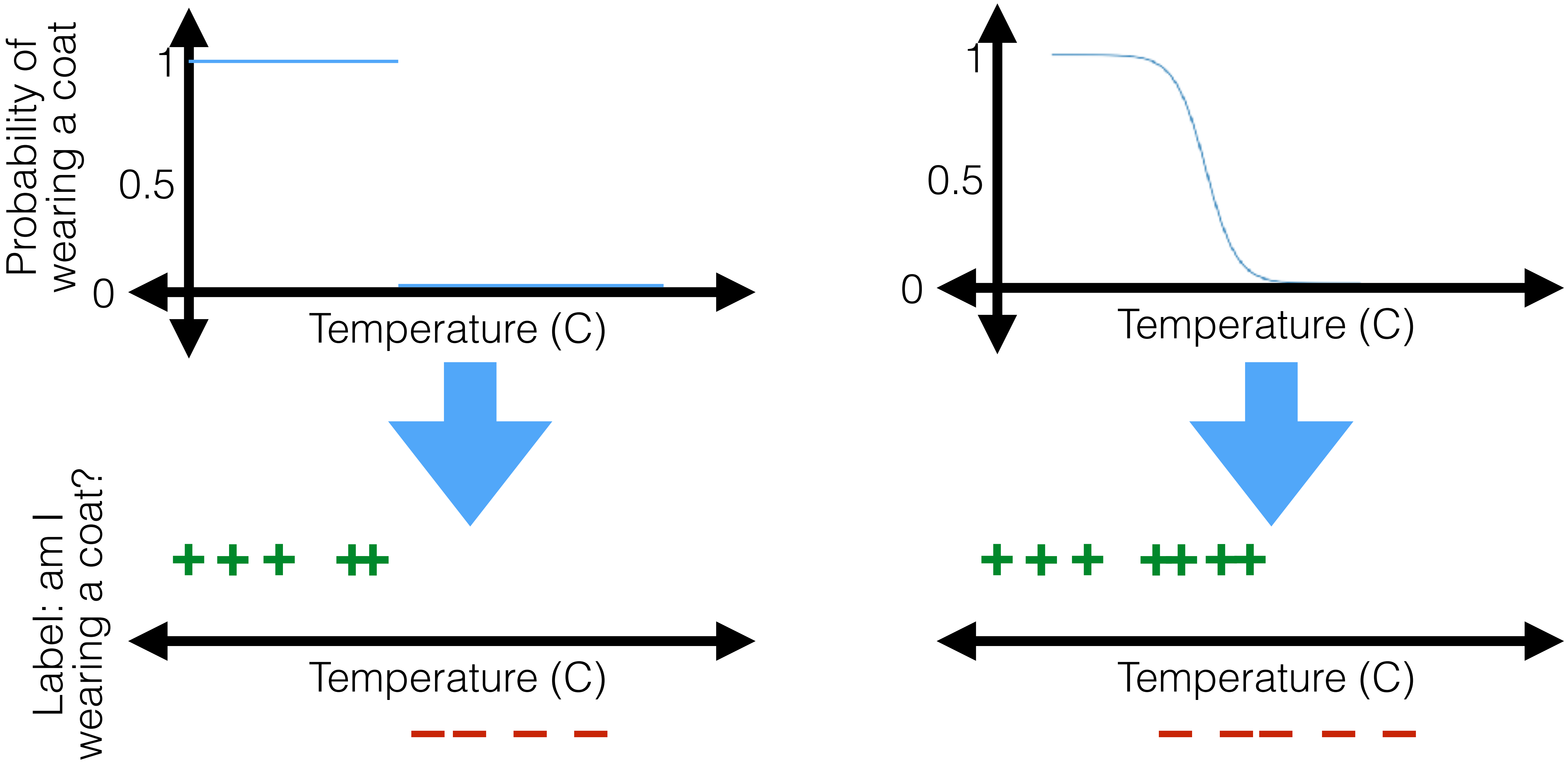
Capturing uncertainty



Capturing uncertainty

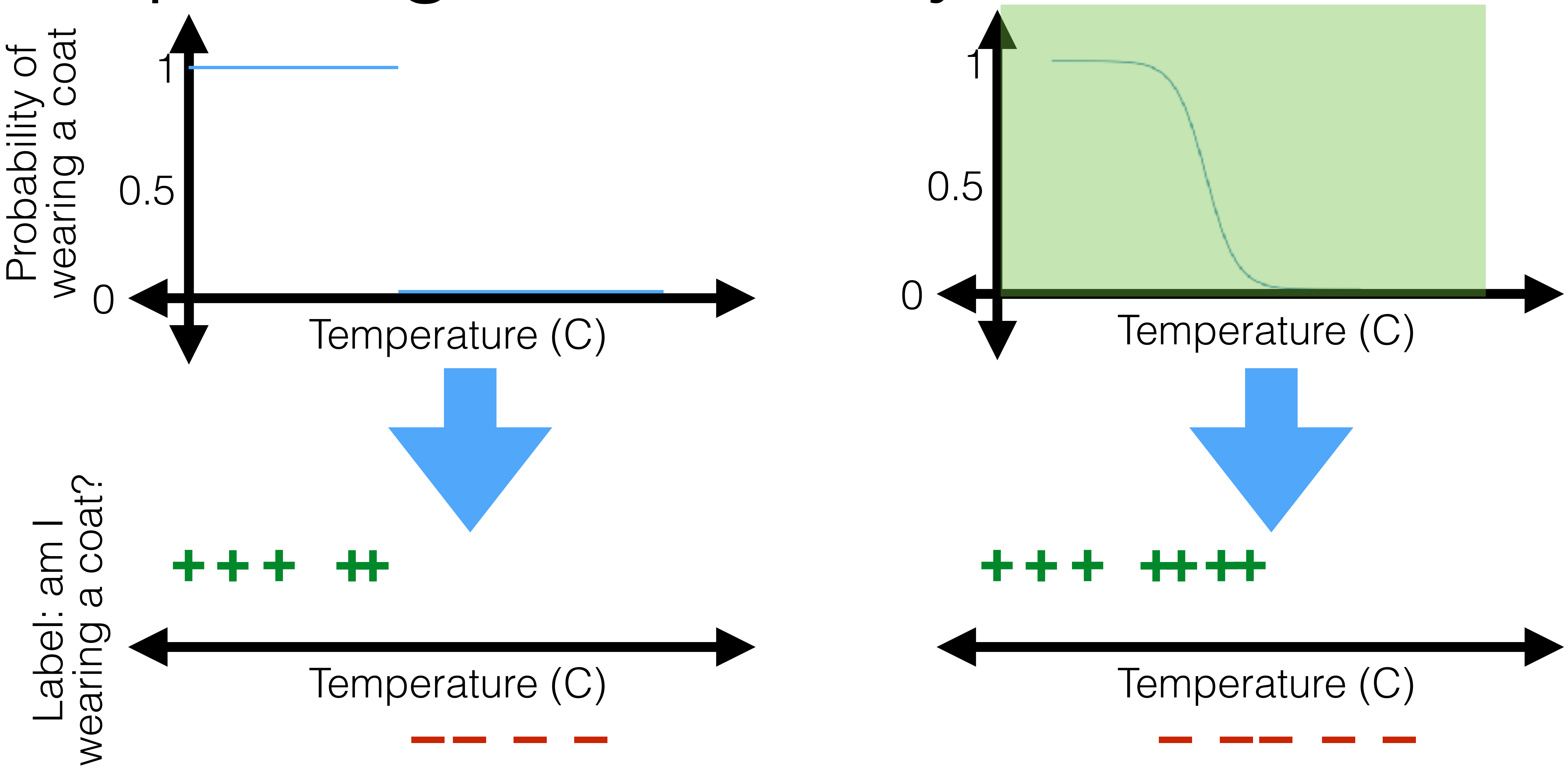


Capturing uncertainty



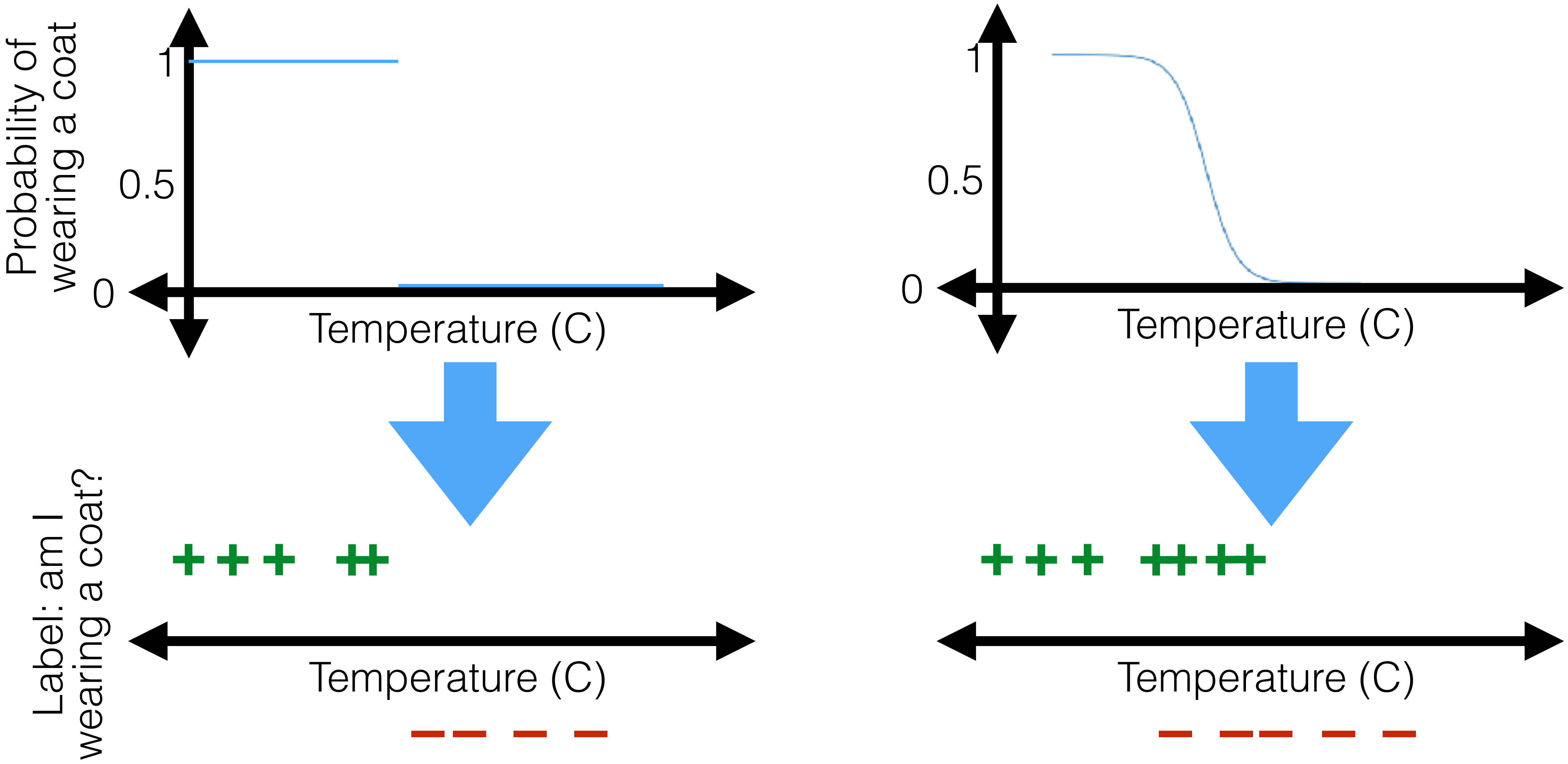
- How to make this shape?

Capturing uncertainty



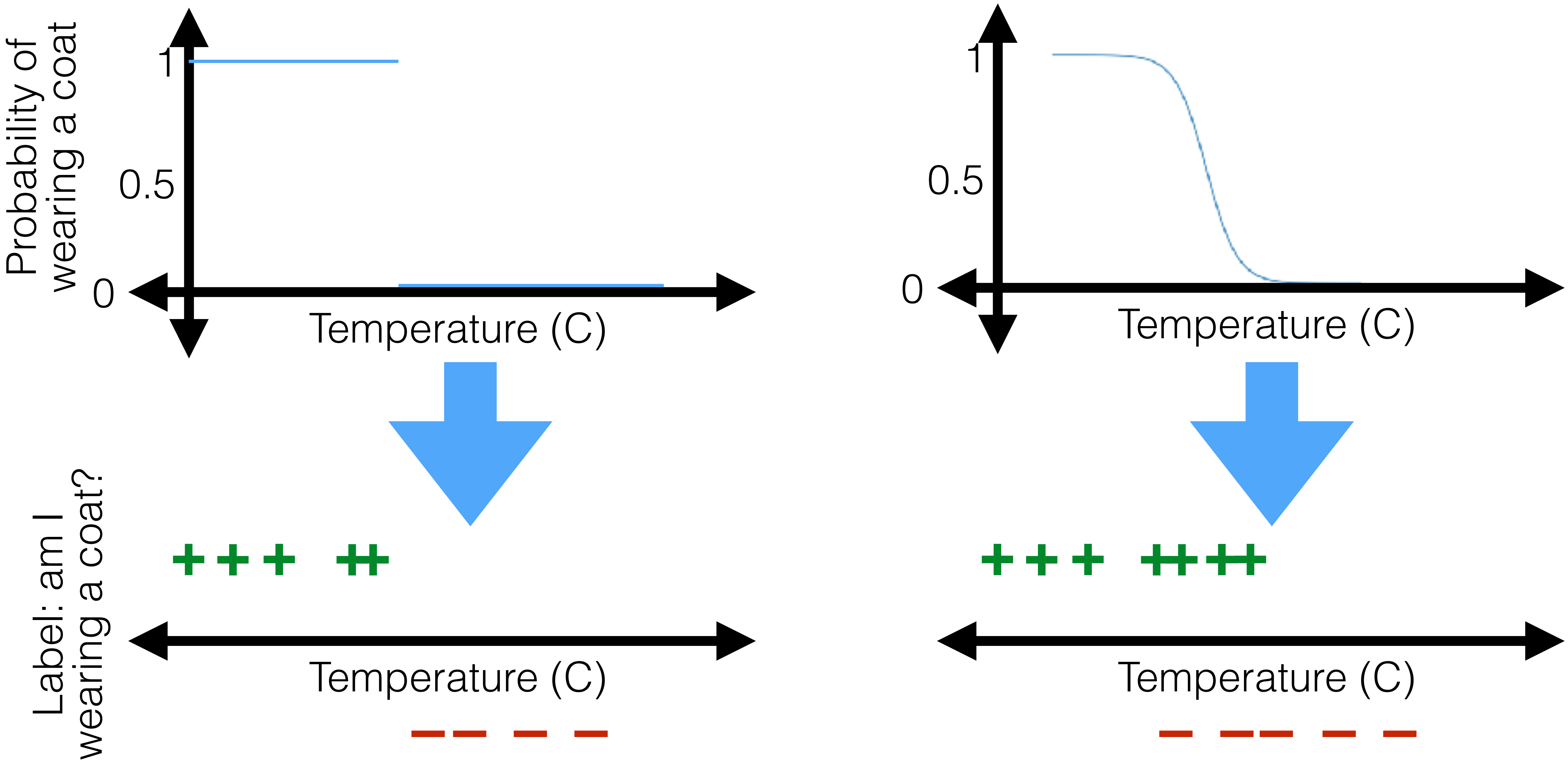
- How to make this shape?

Capturing uncertainty



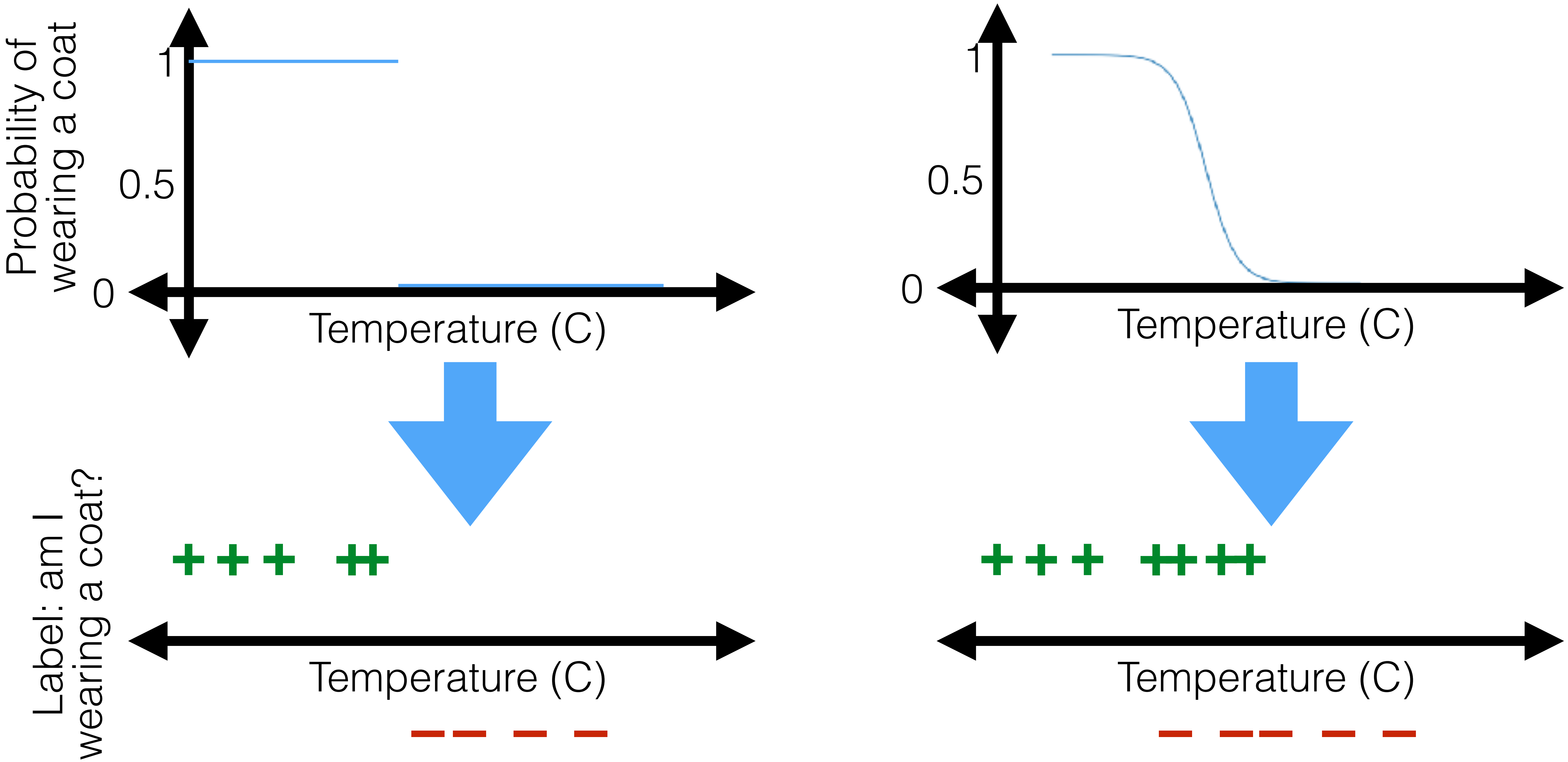
- How to make this shape?

Capturing uncertainty



- How to make this shape?
    - Sigmoid/logistic function

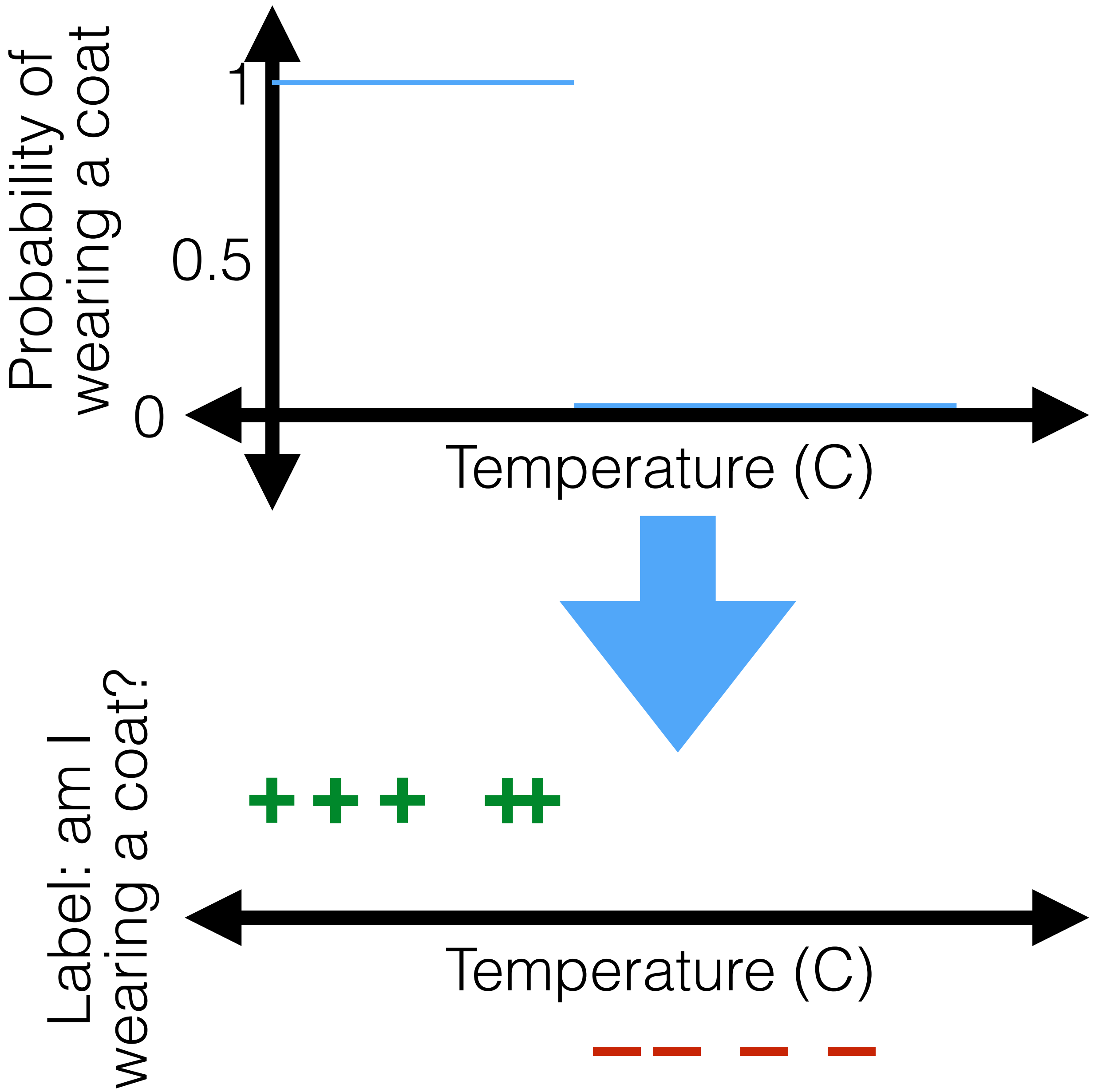
Capturing uncertainty

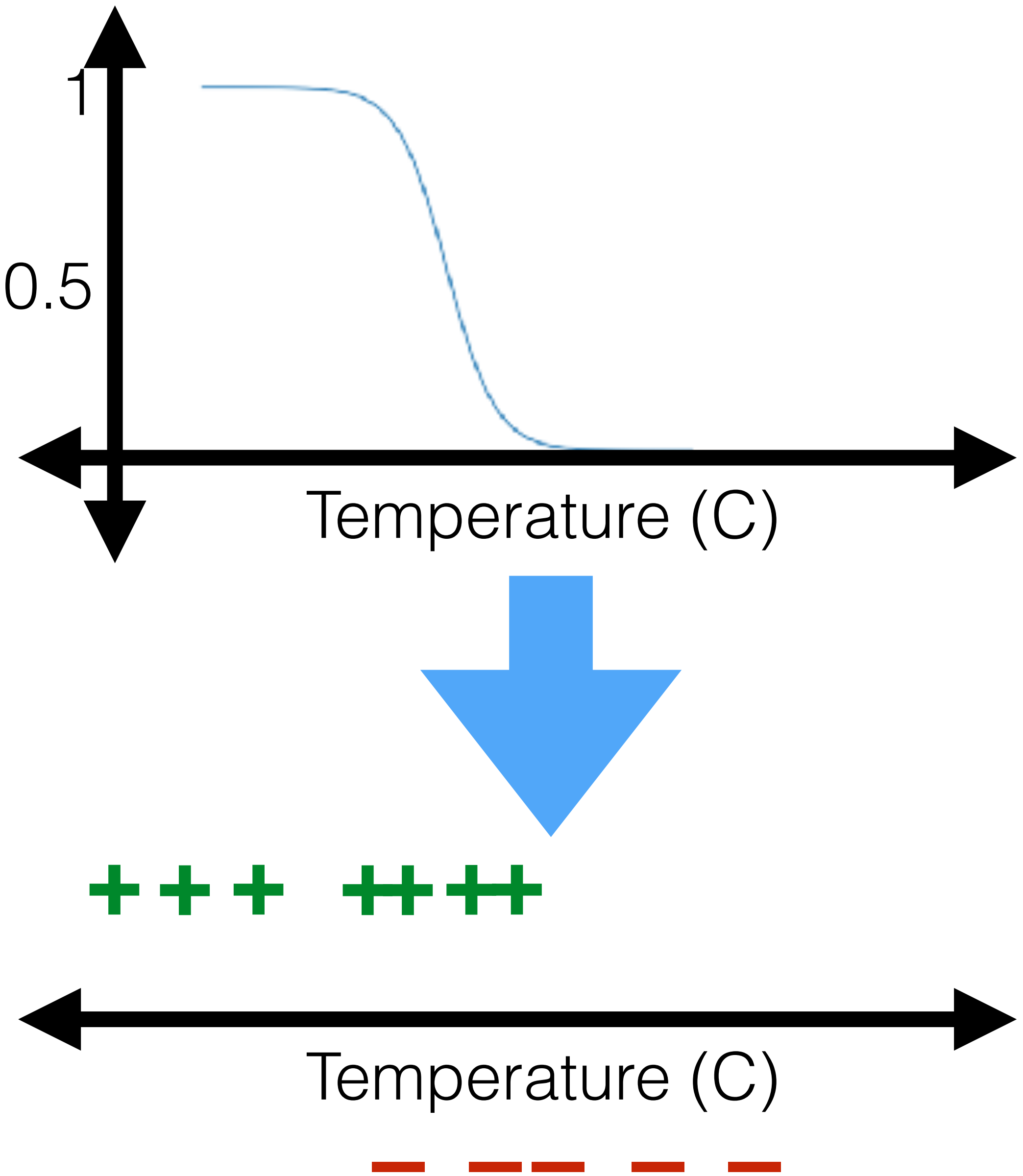


- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

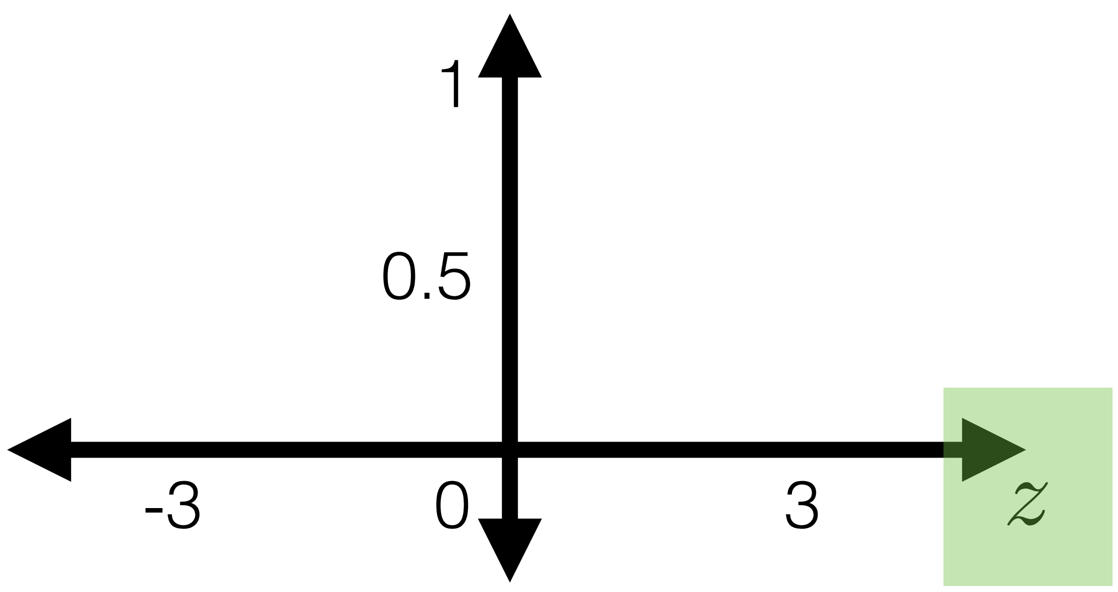
Capturing uncertainty



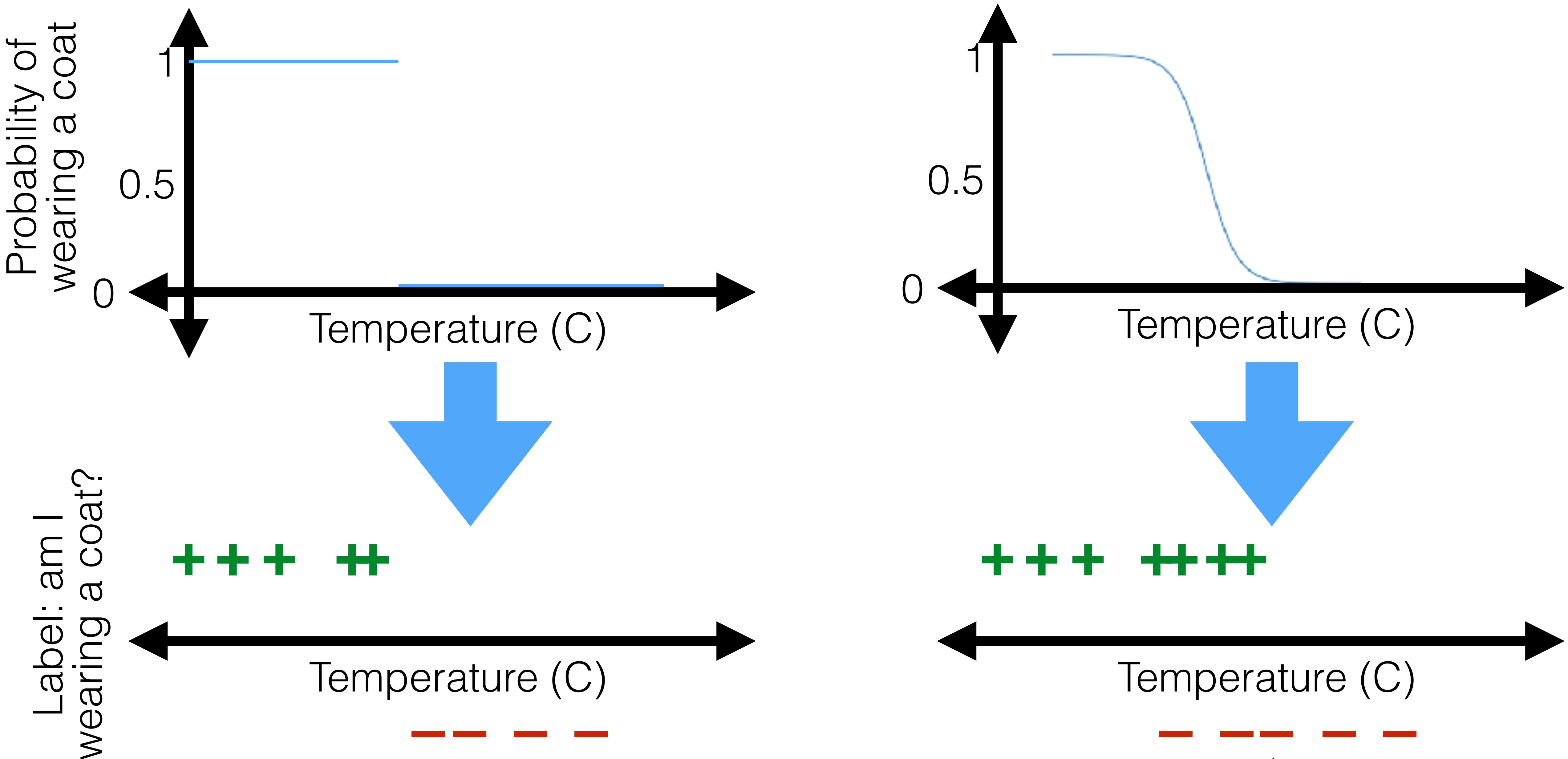


- How to make this shape?
    - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

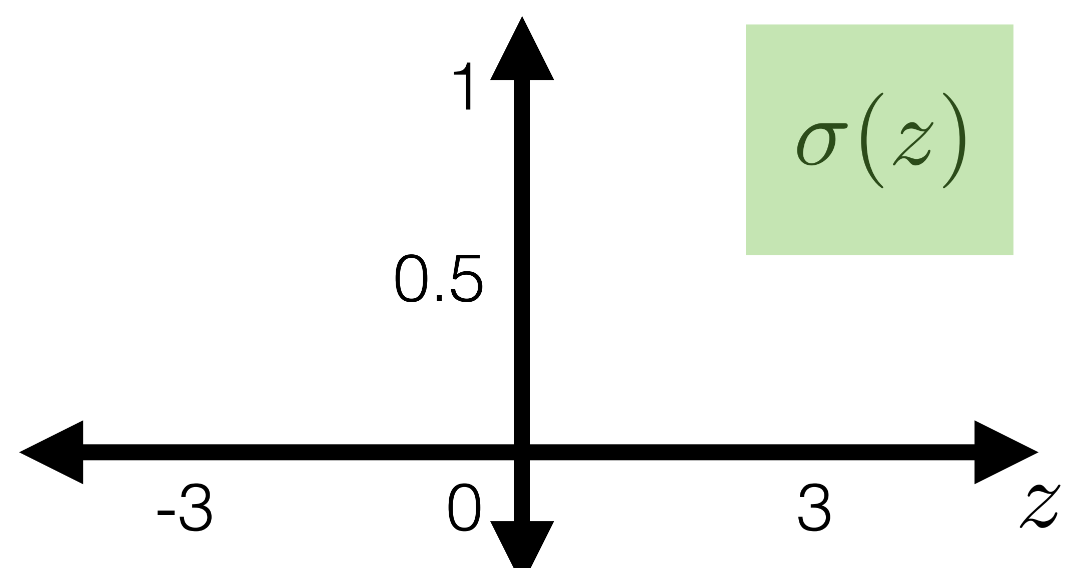


Capturing uncertainty

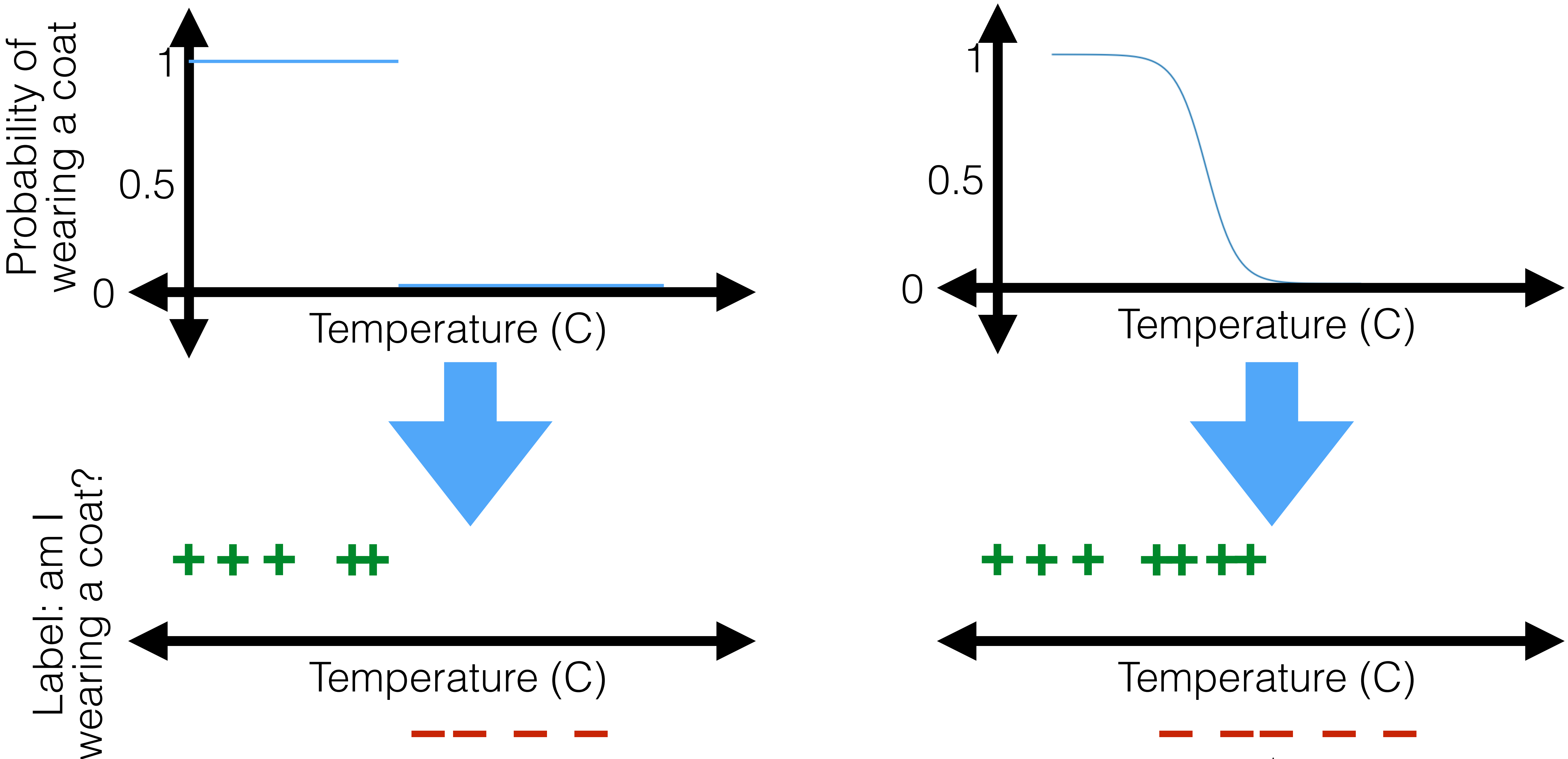


- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

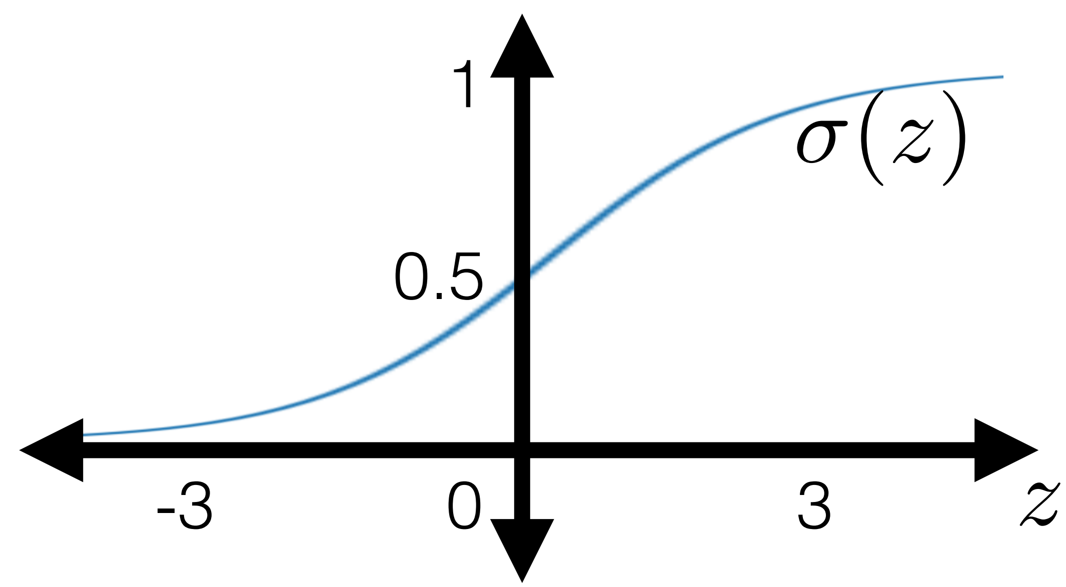


Capturing uncertainty

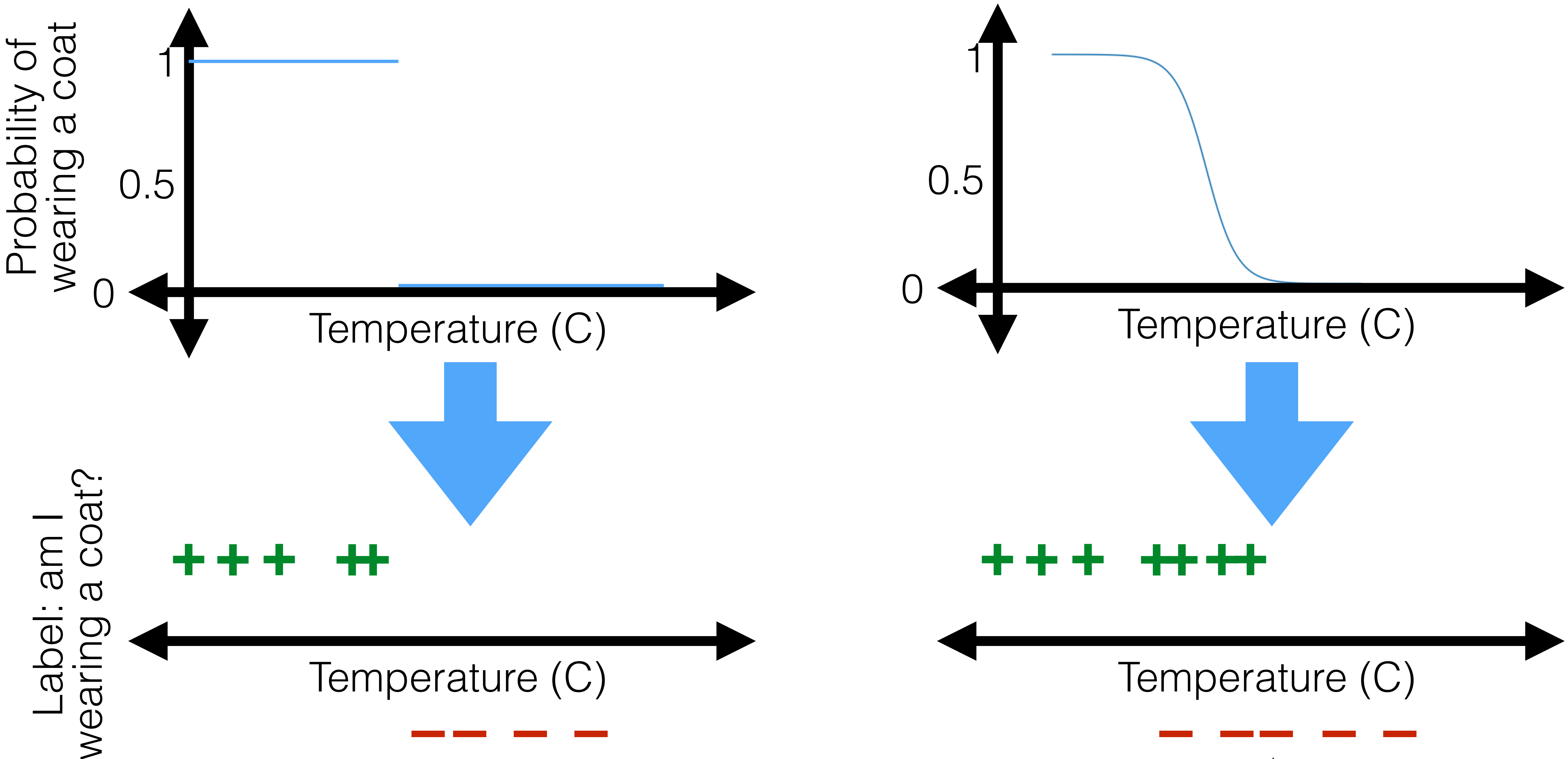


- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1+\exp(-z)}$$

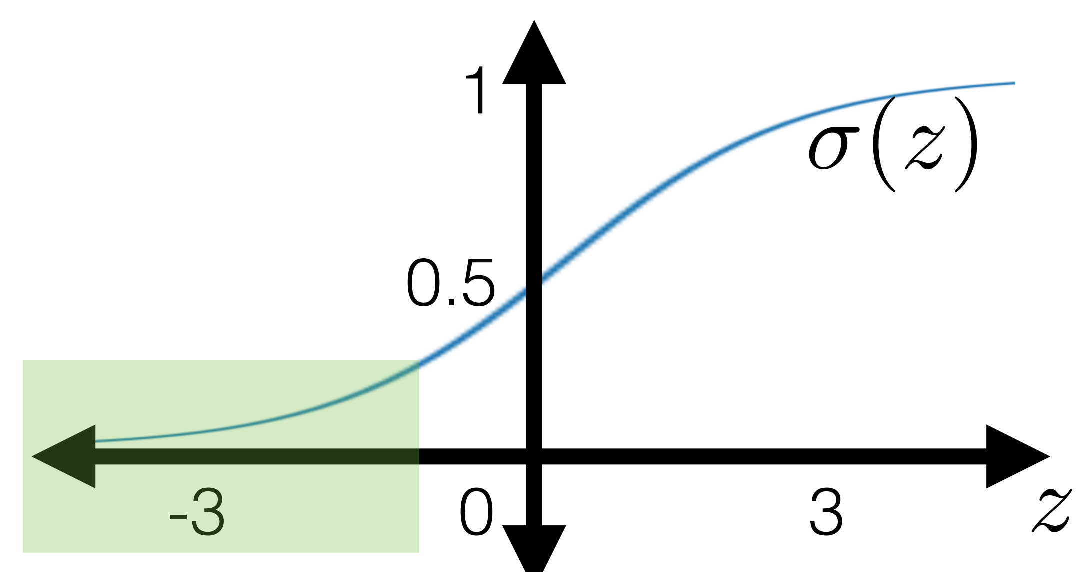


Capturing uncertainty

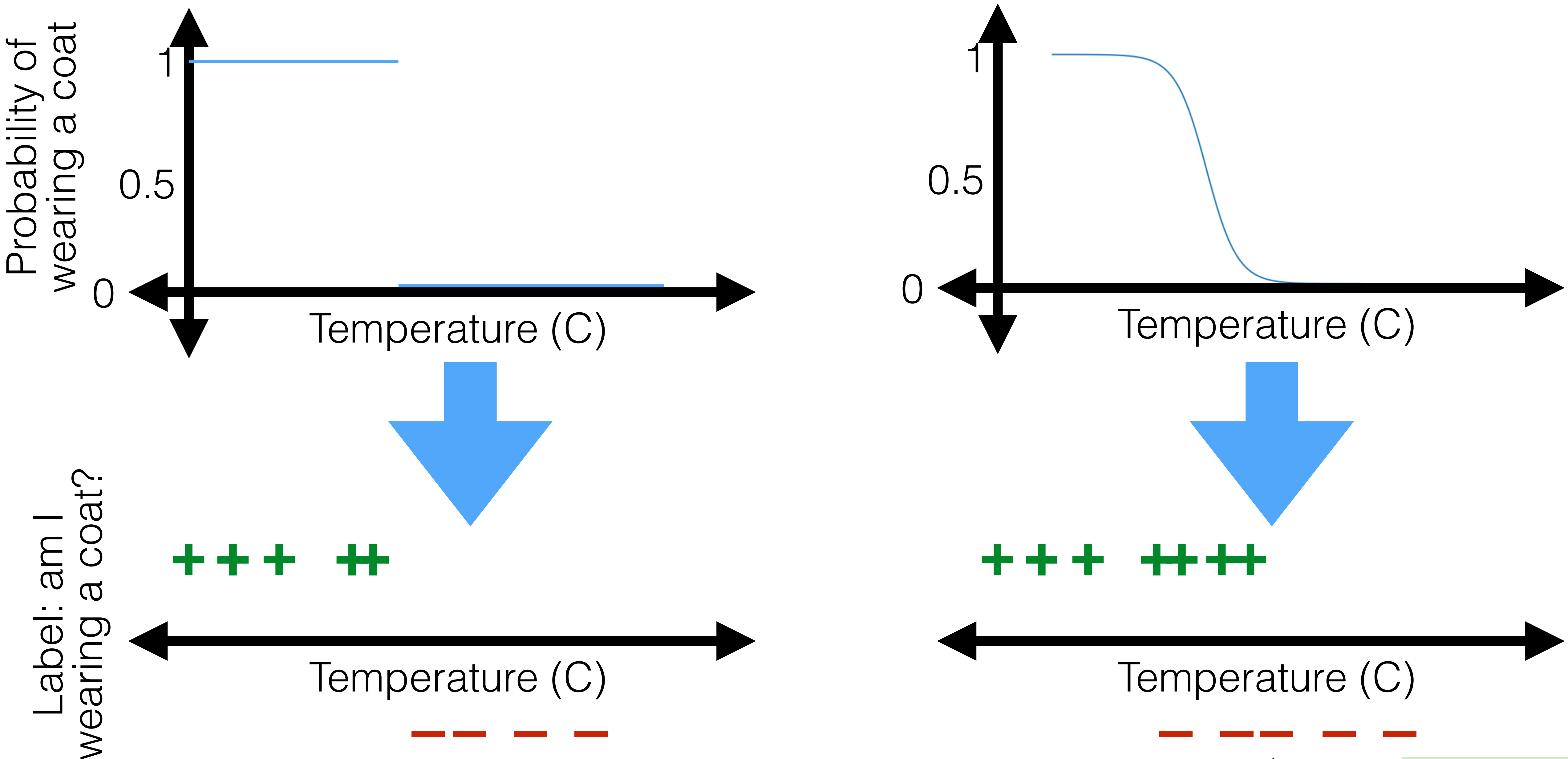


- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

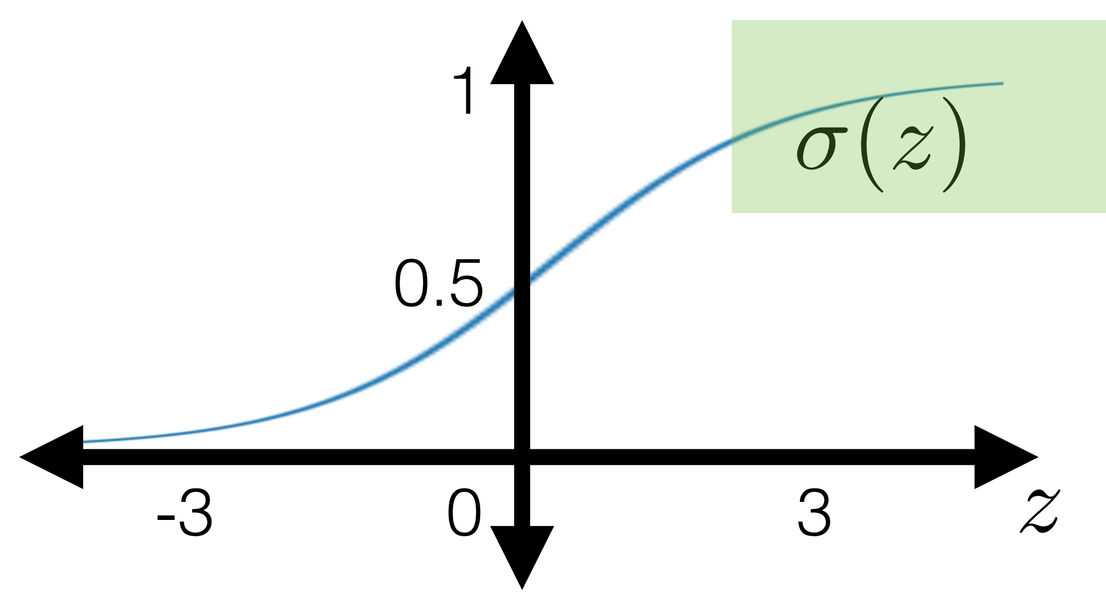


Capturing uncertainty

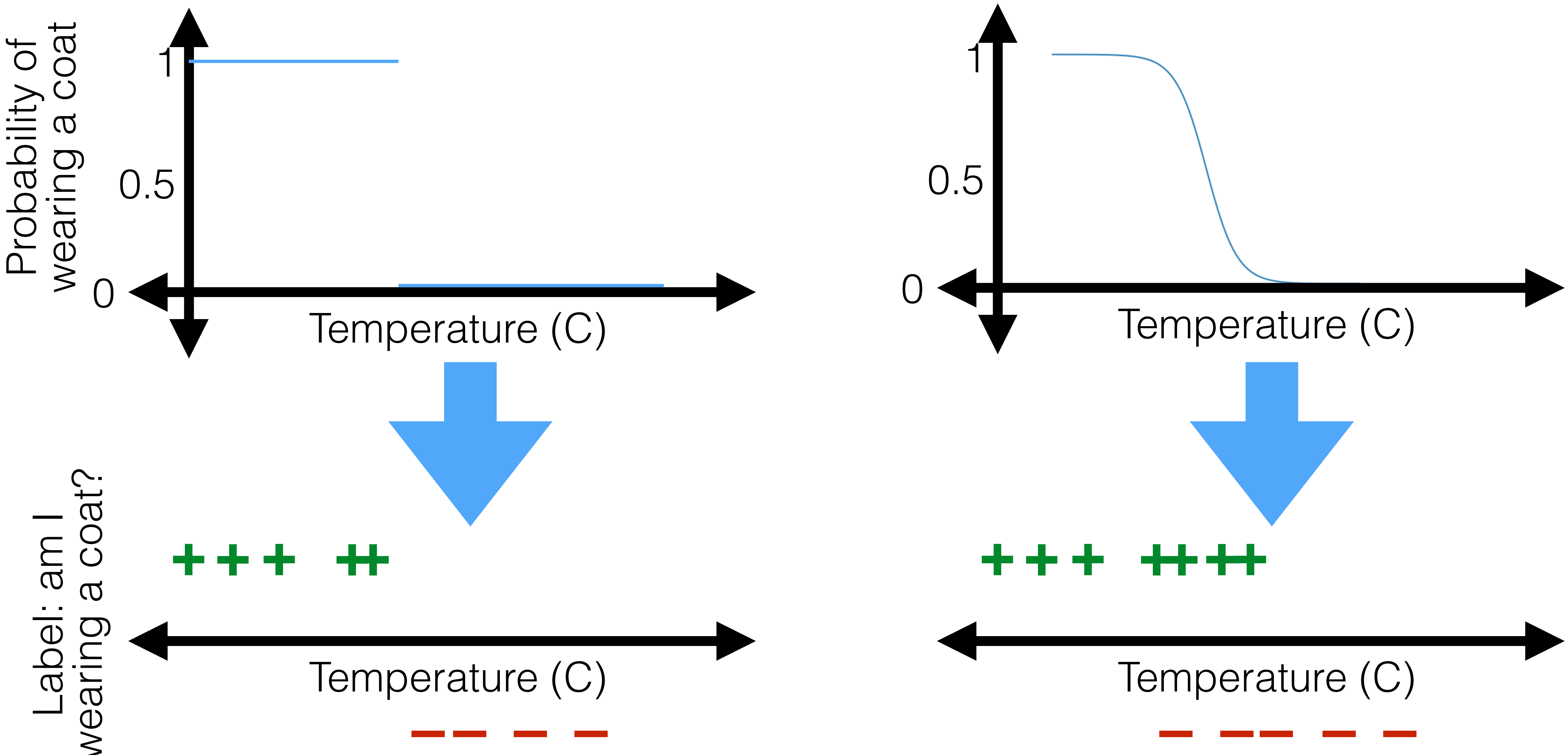


- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

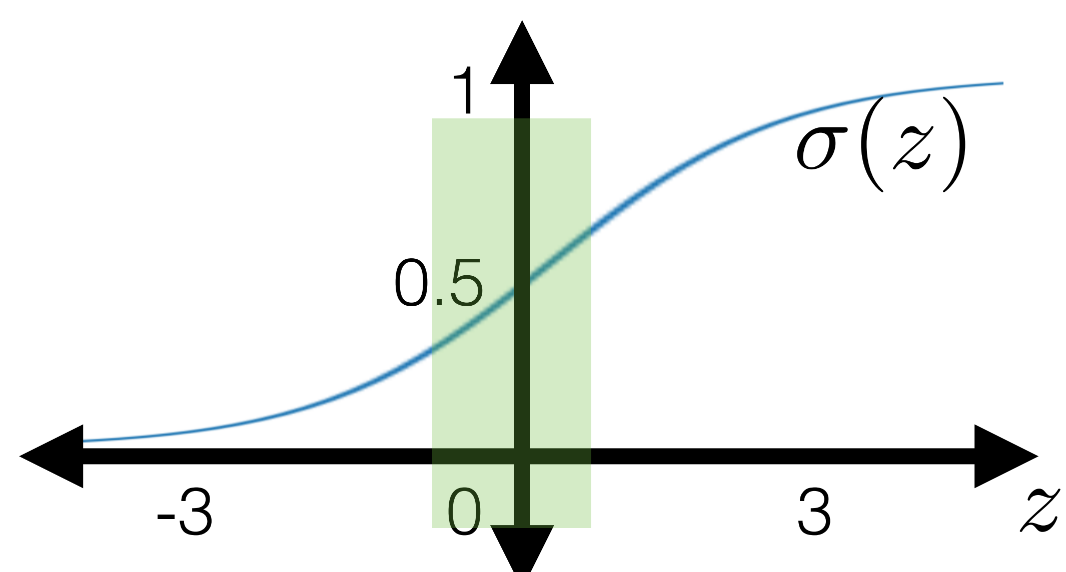


Capturing uncertainty

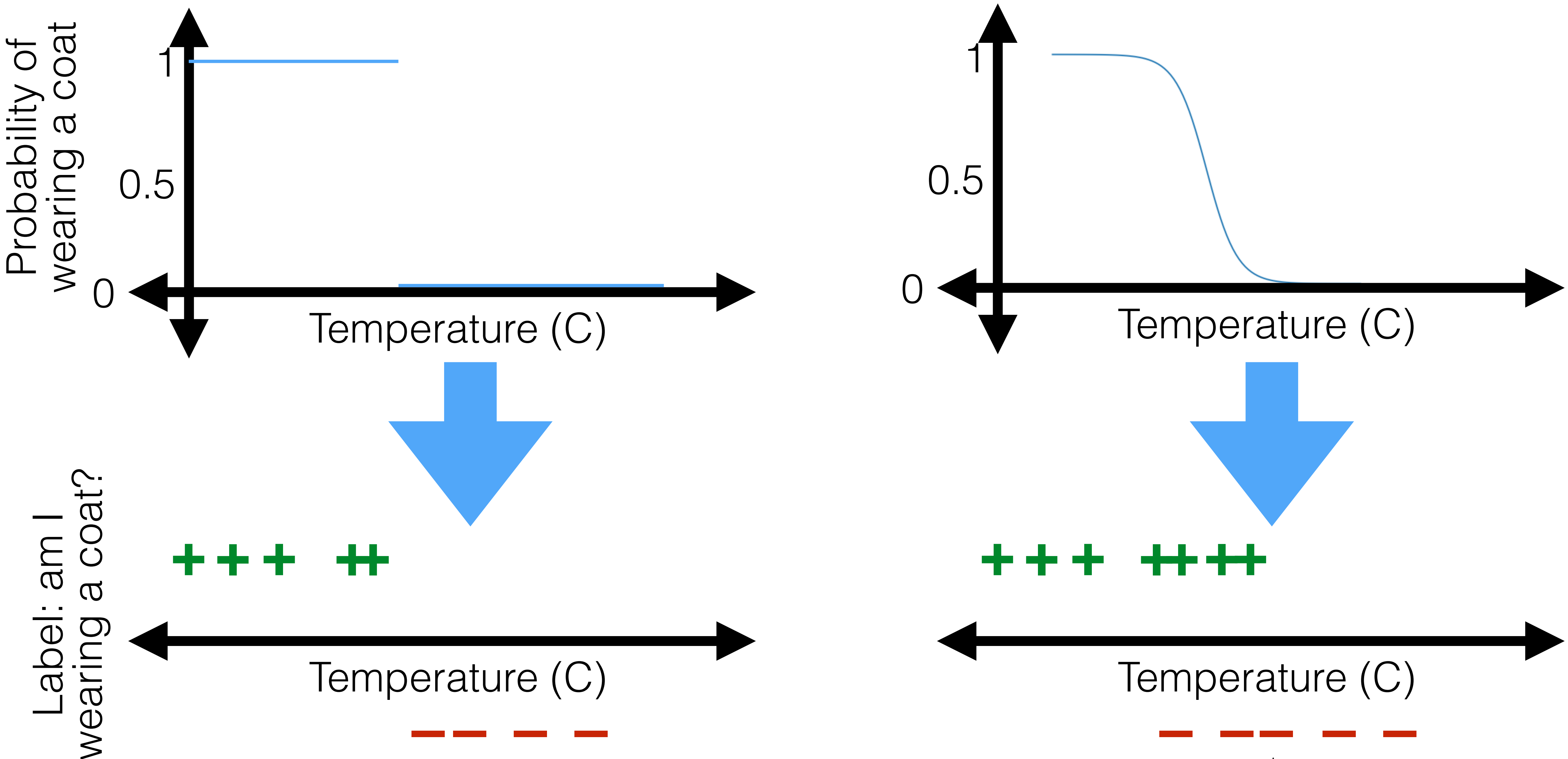


- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

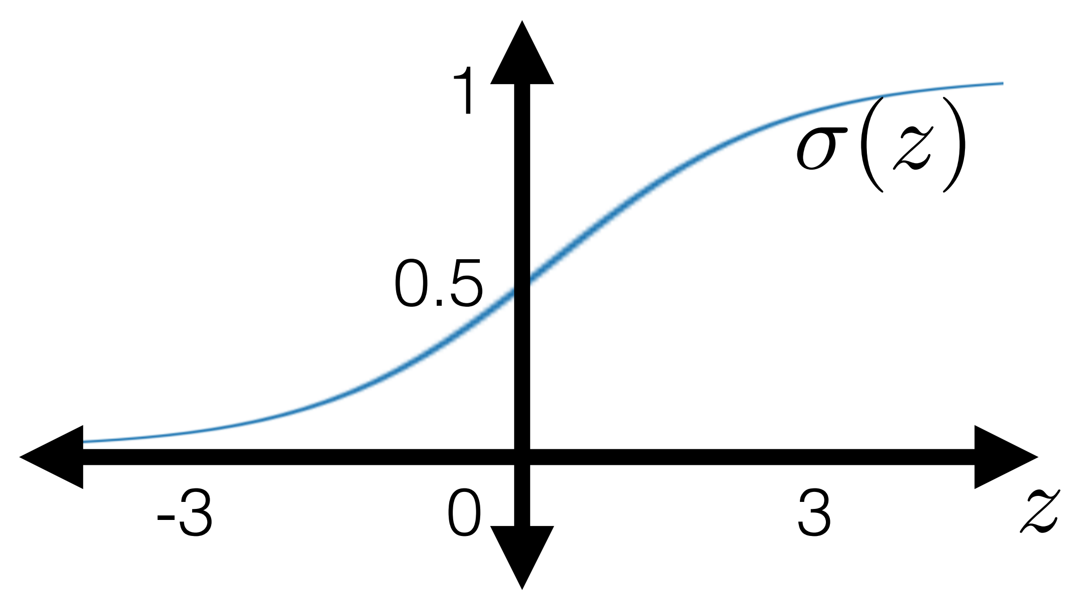


Capturing uncertainty

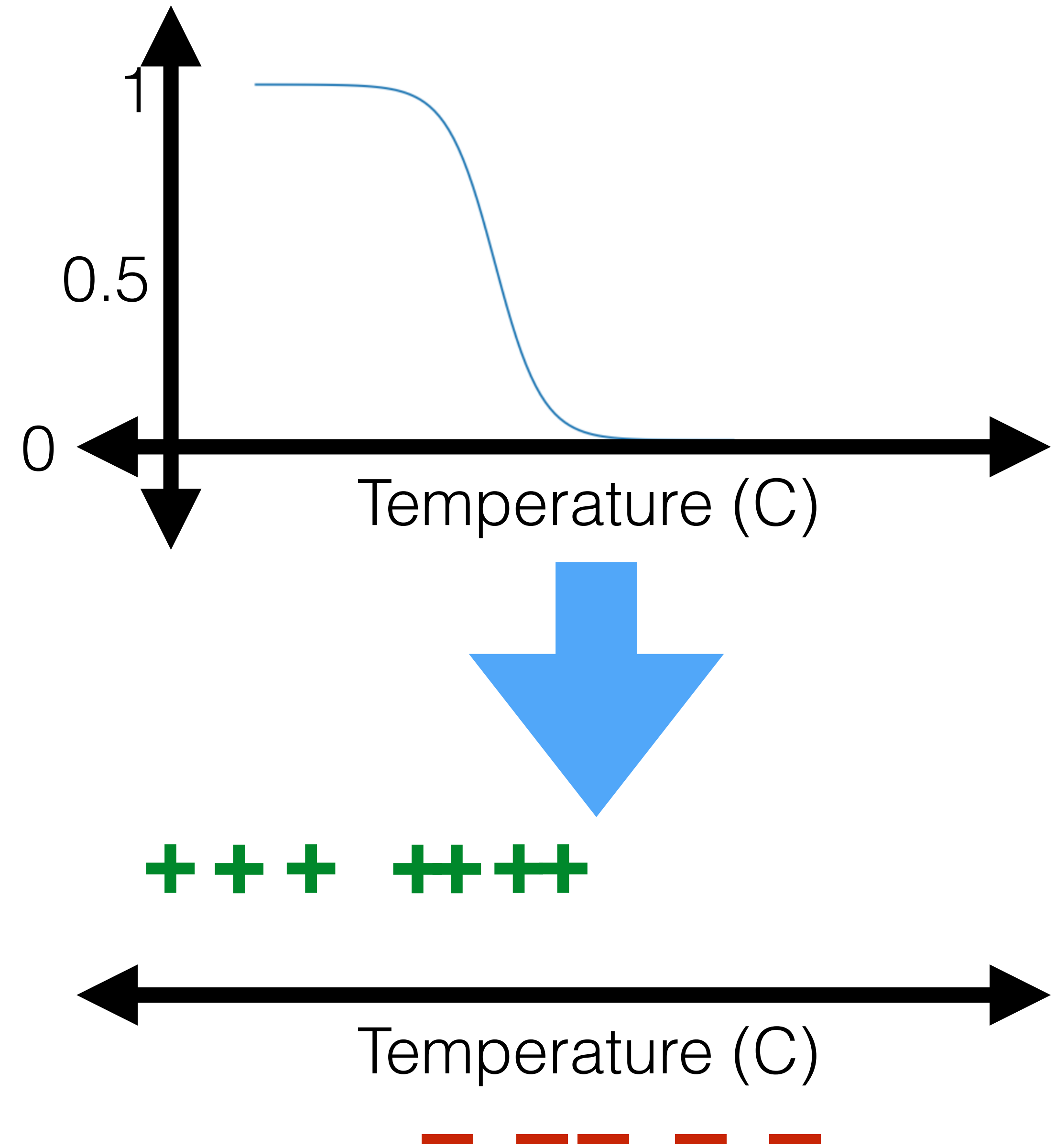


- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

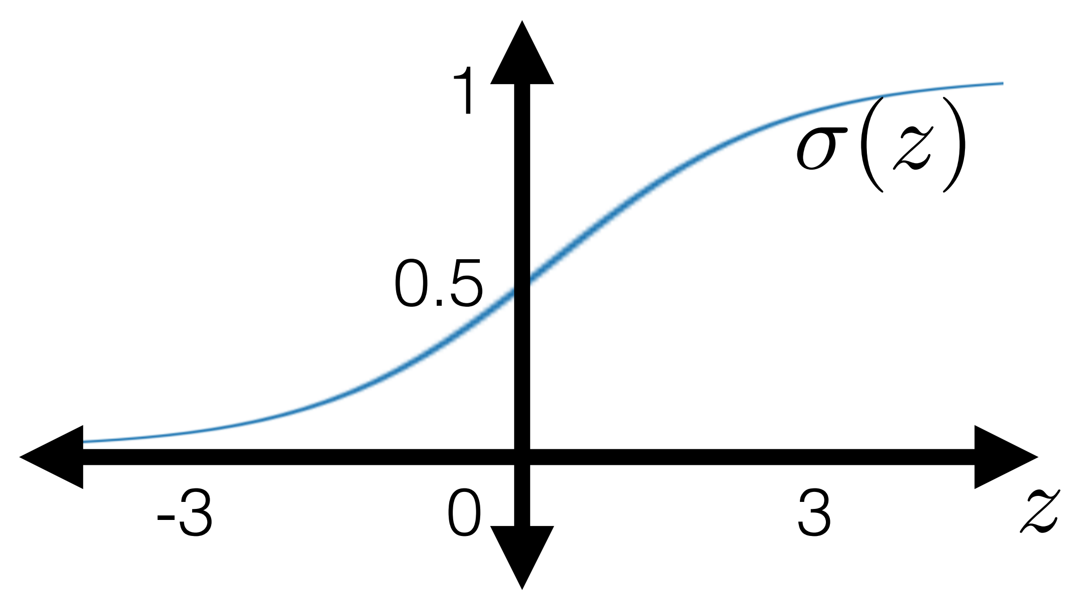


Capturing uncertainty

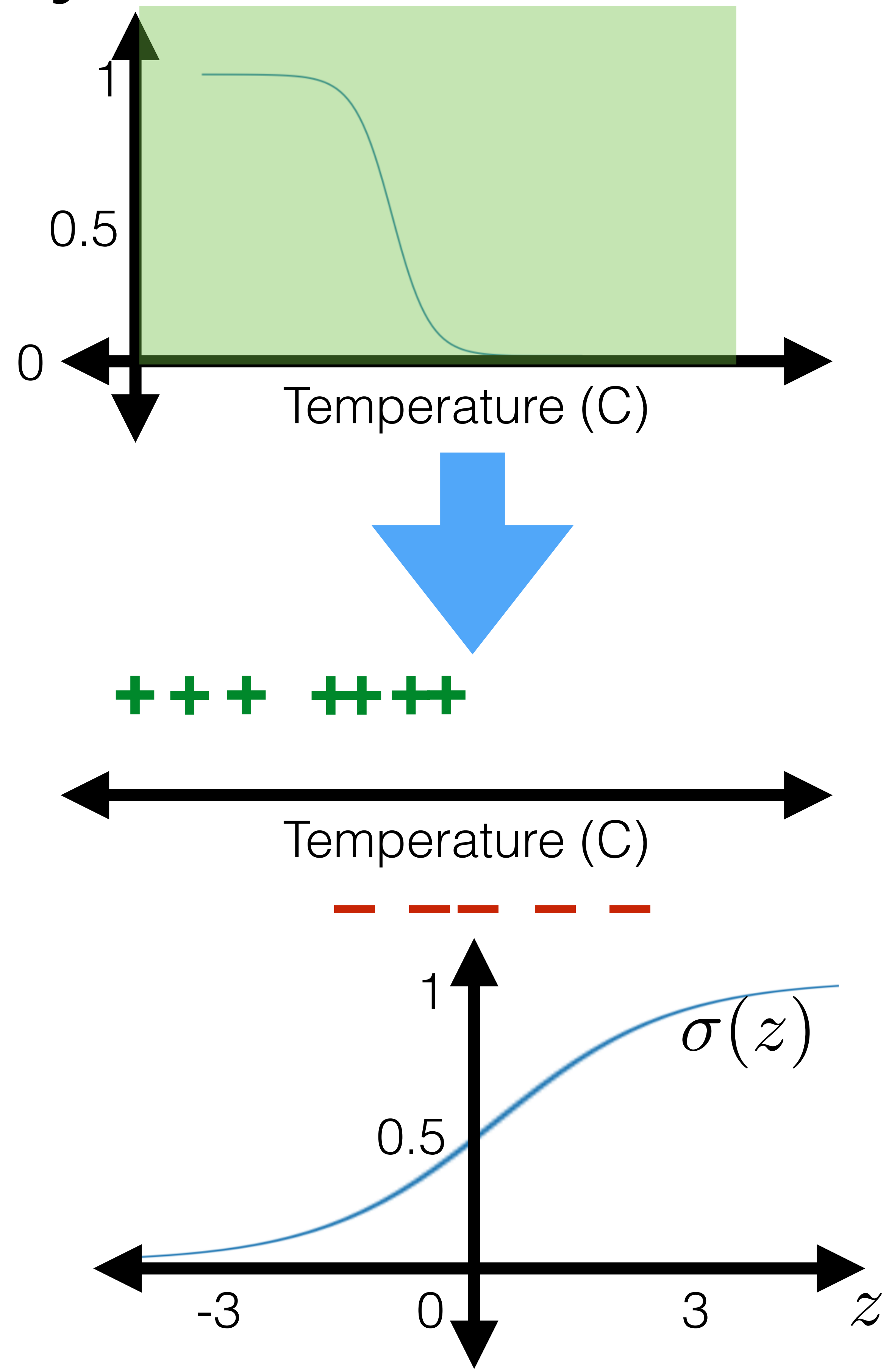


- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$



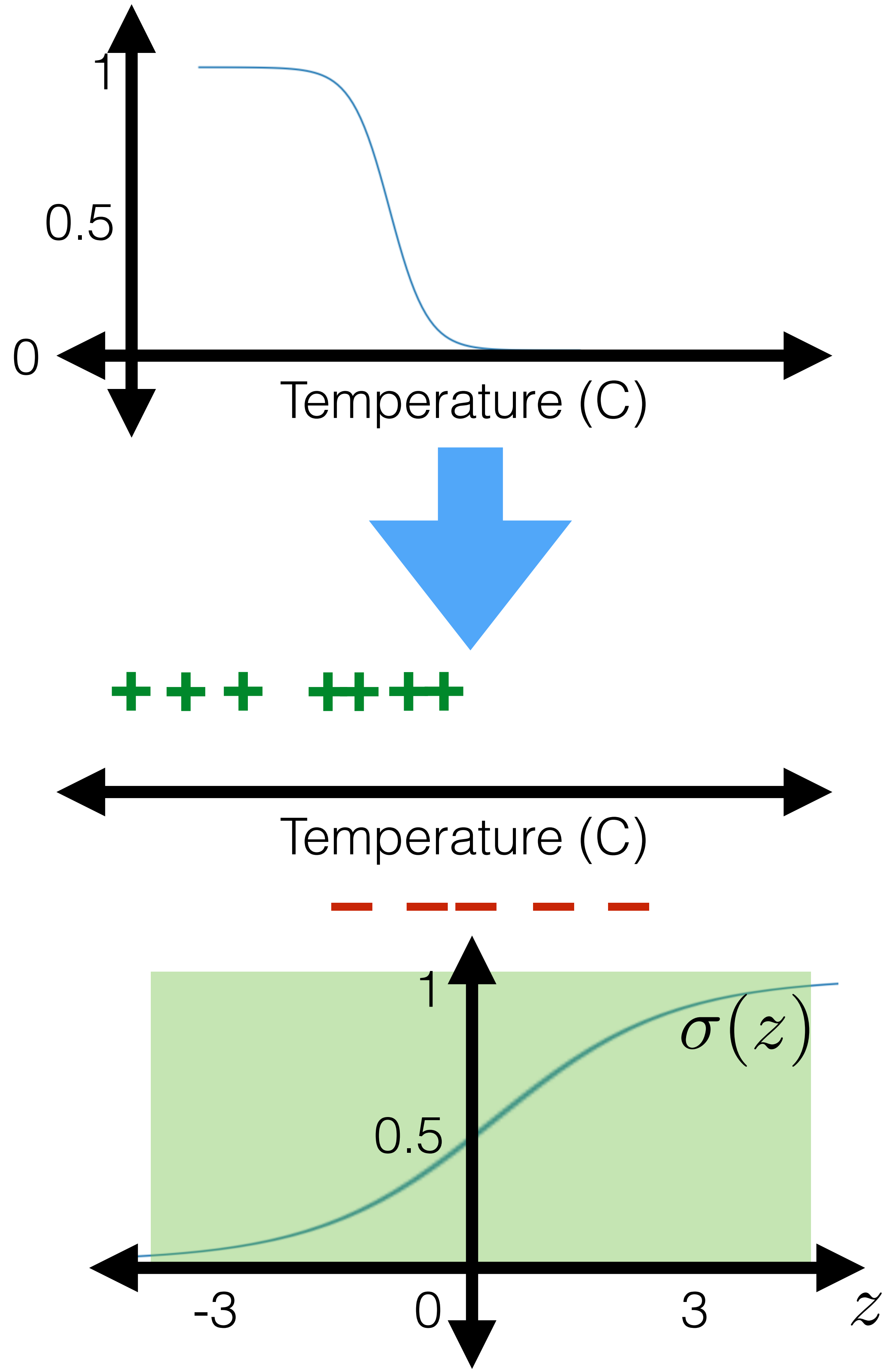
Capturing uncertainty



- How to make this shape?
    - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

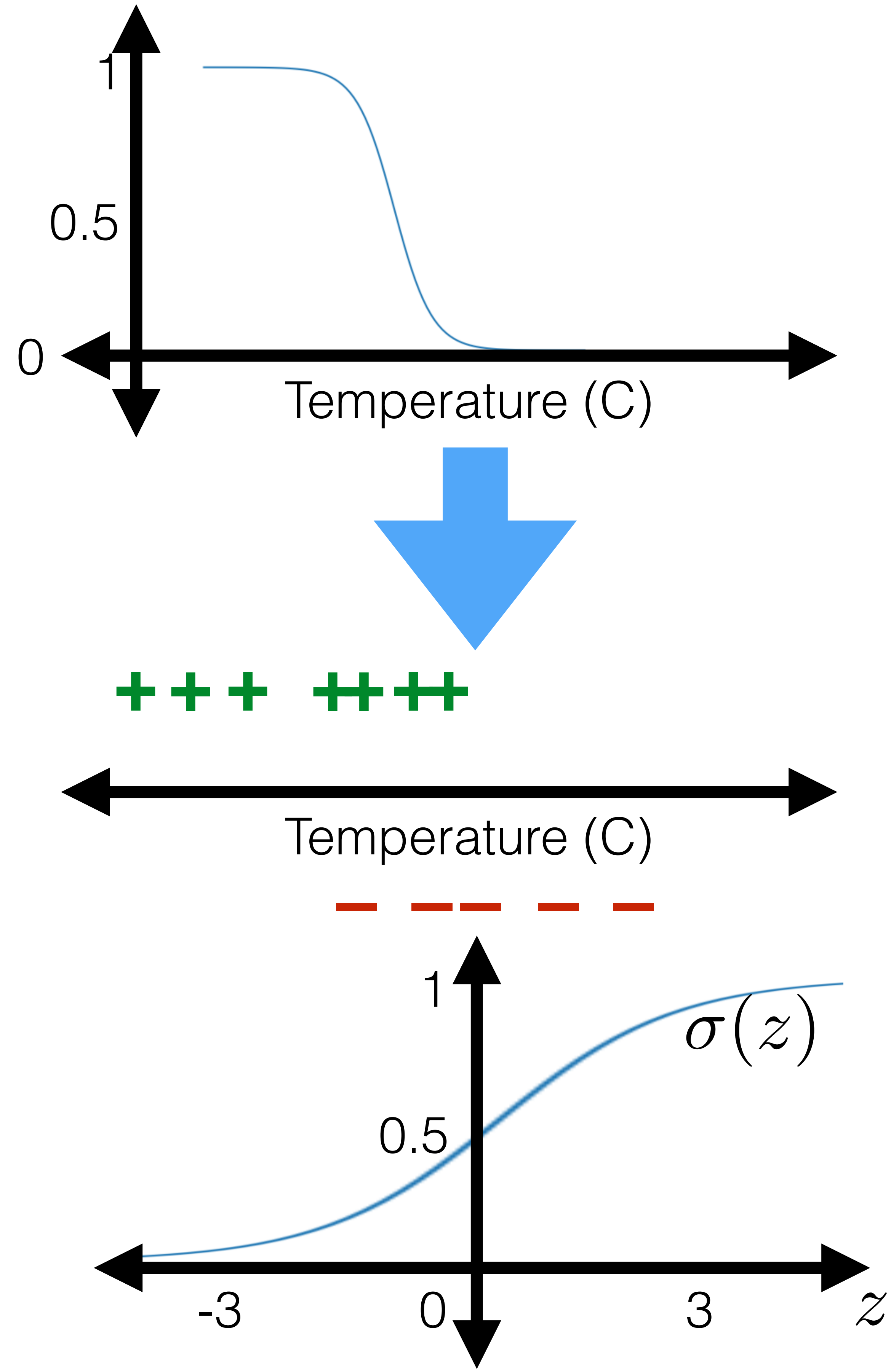
Capturing uncertainty



- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

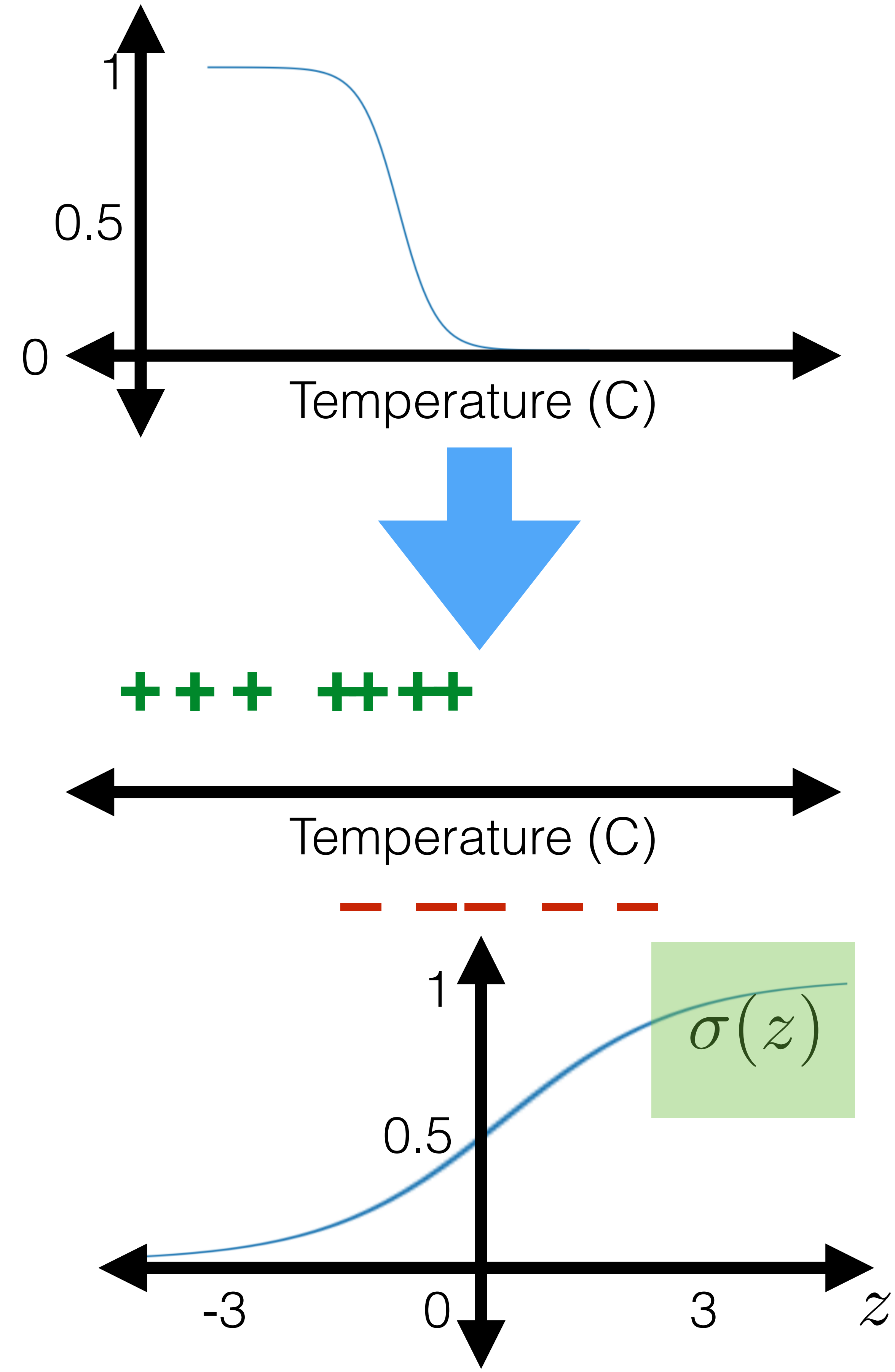
Capturing uncertainty



- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

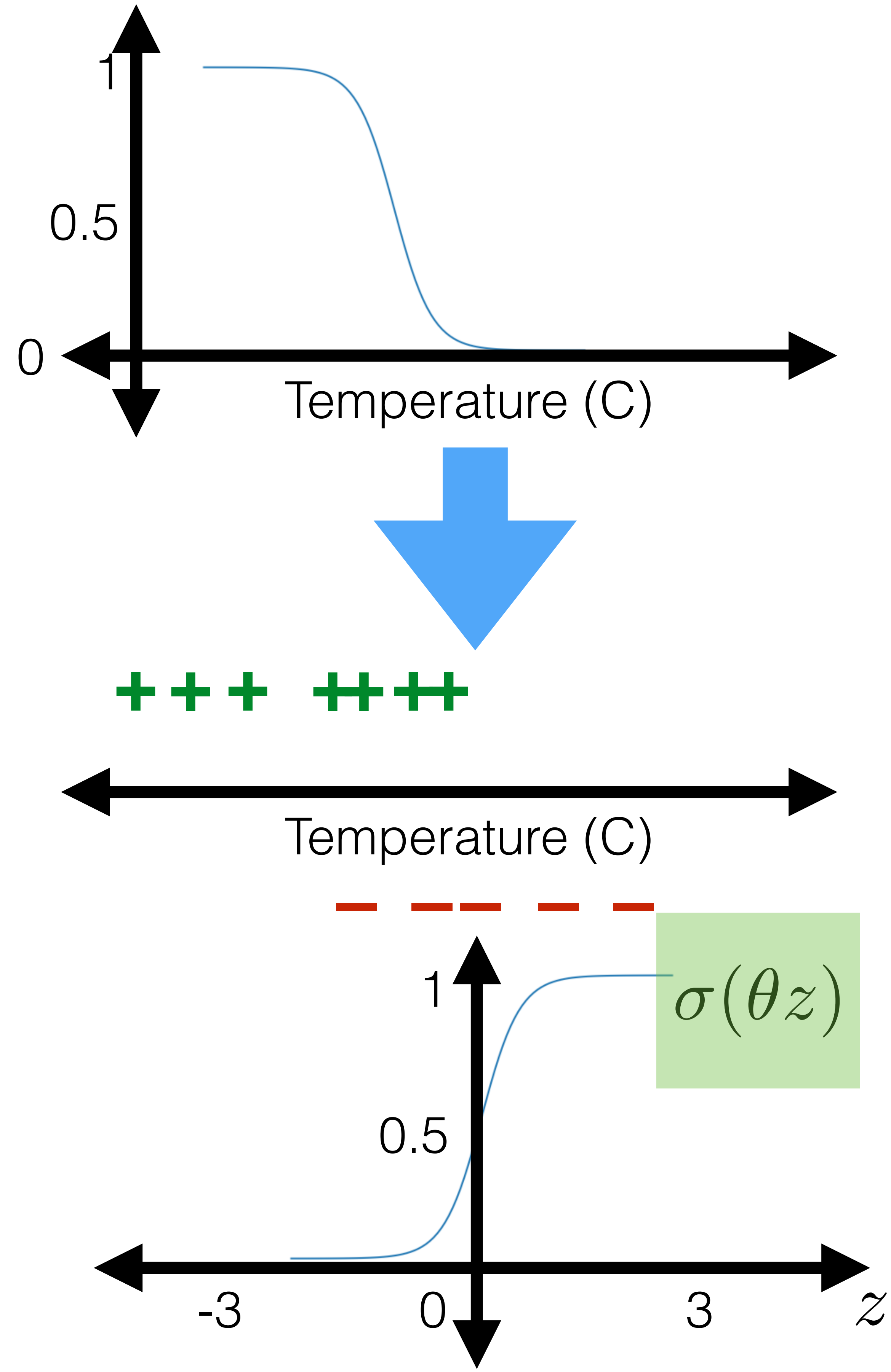
Capturing uncertainty



- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

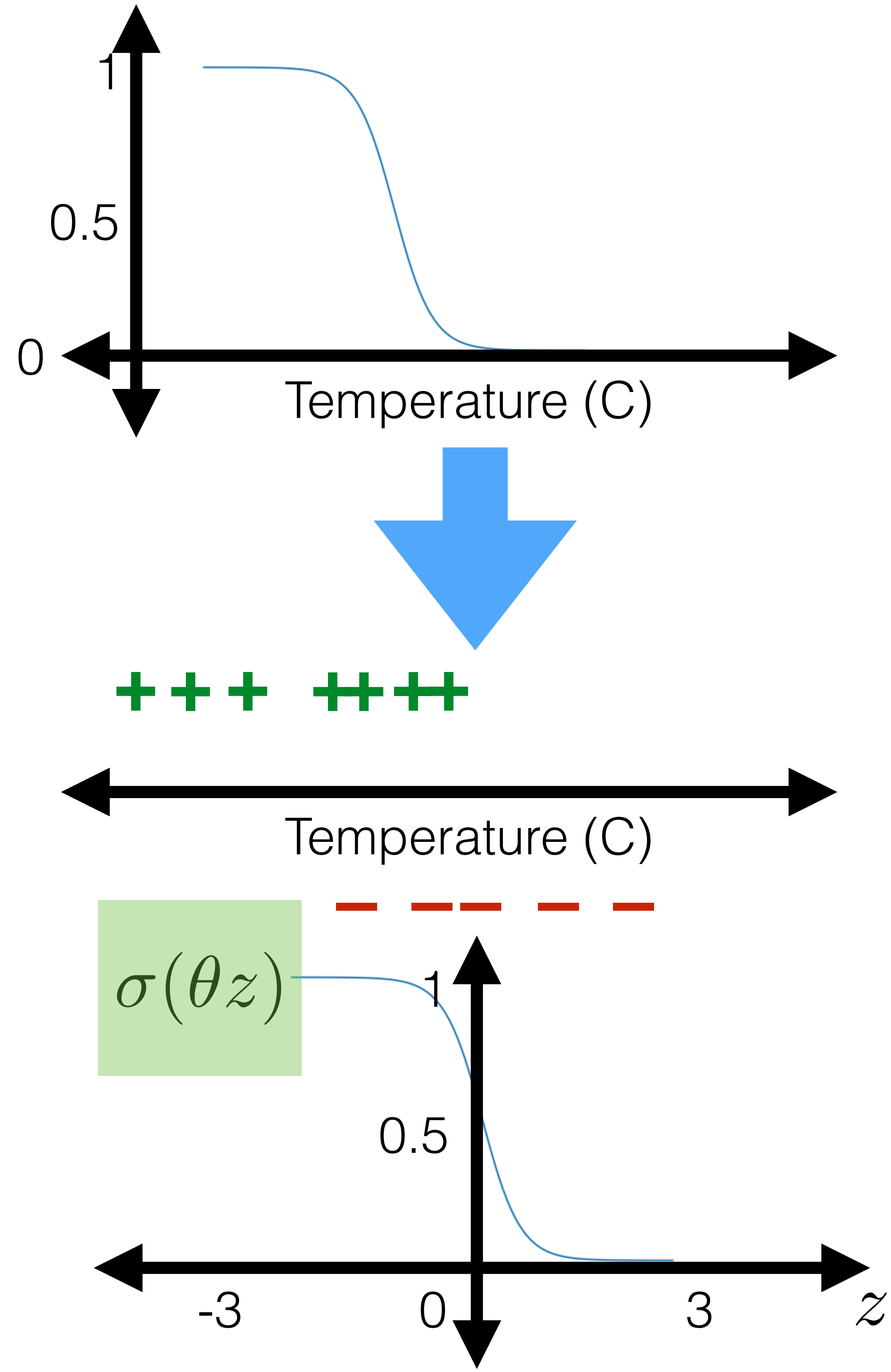
Capturing uncertainty



- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

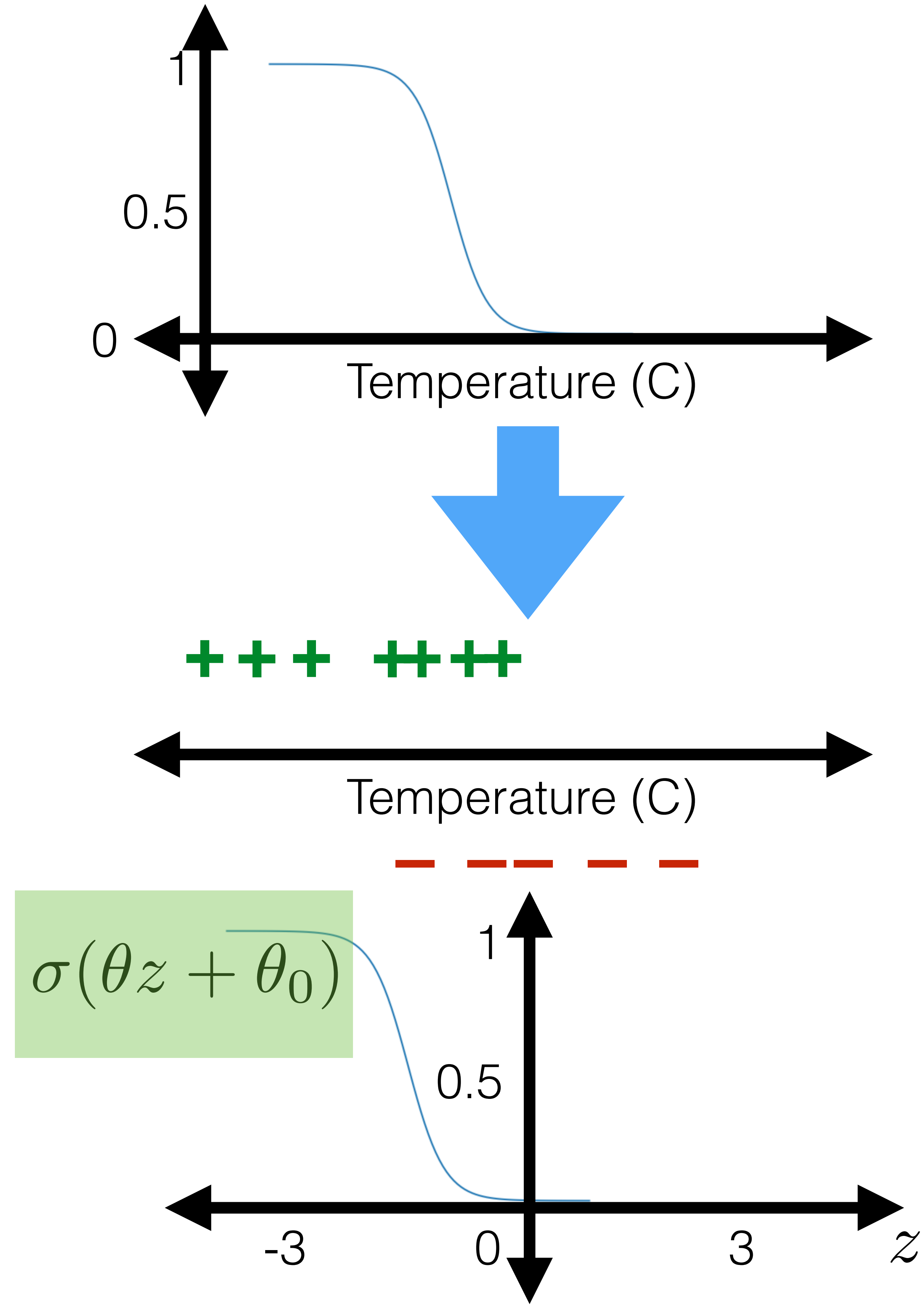
Capturing uncertainty



- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

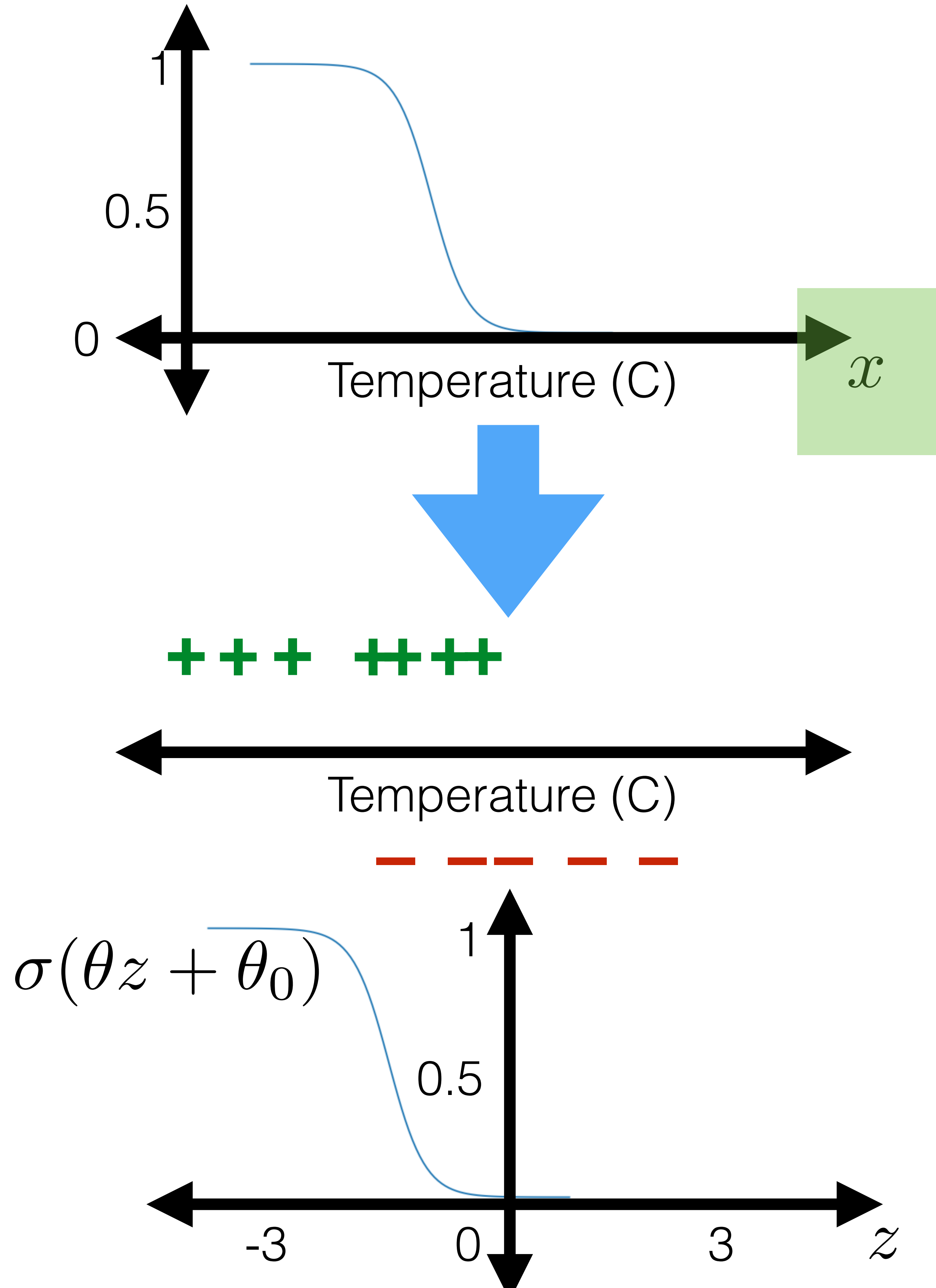
Capturing uncertainty



- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

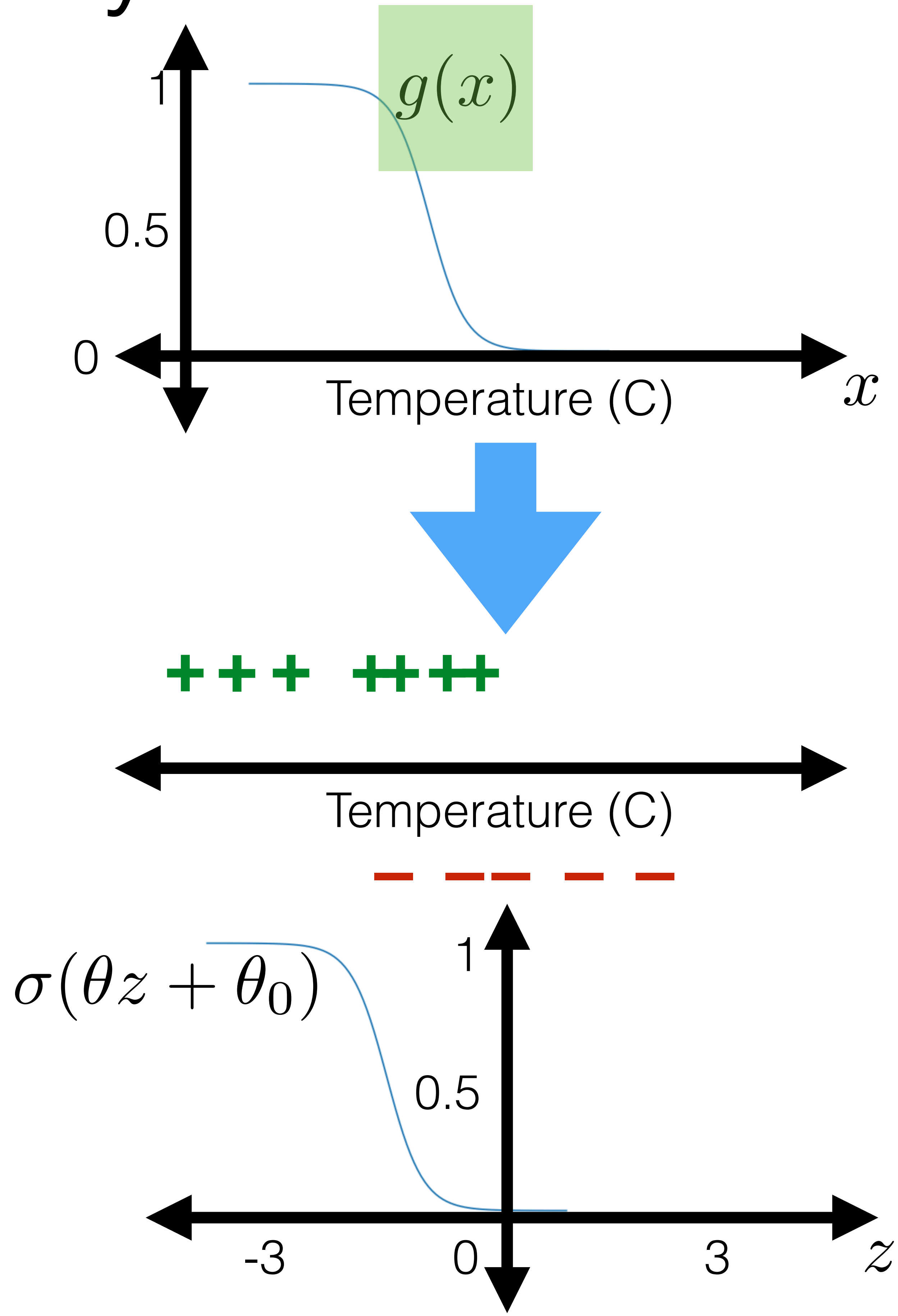
Capturing uncertainty



- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Capturing uncertainty

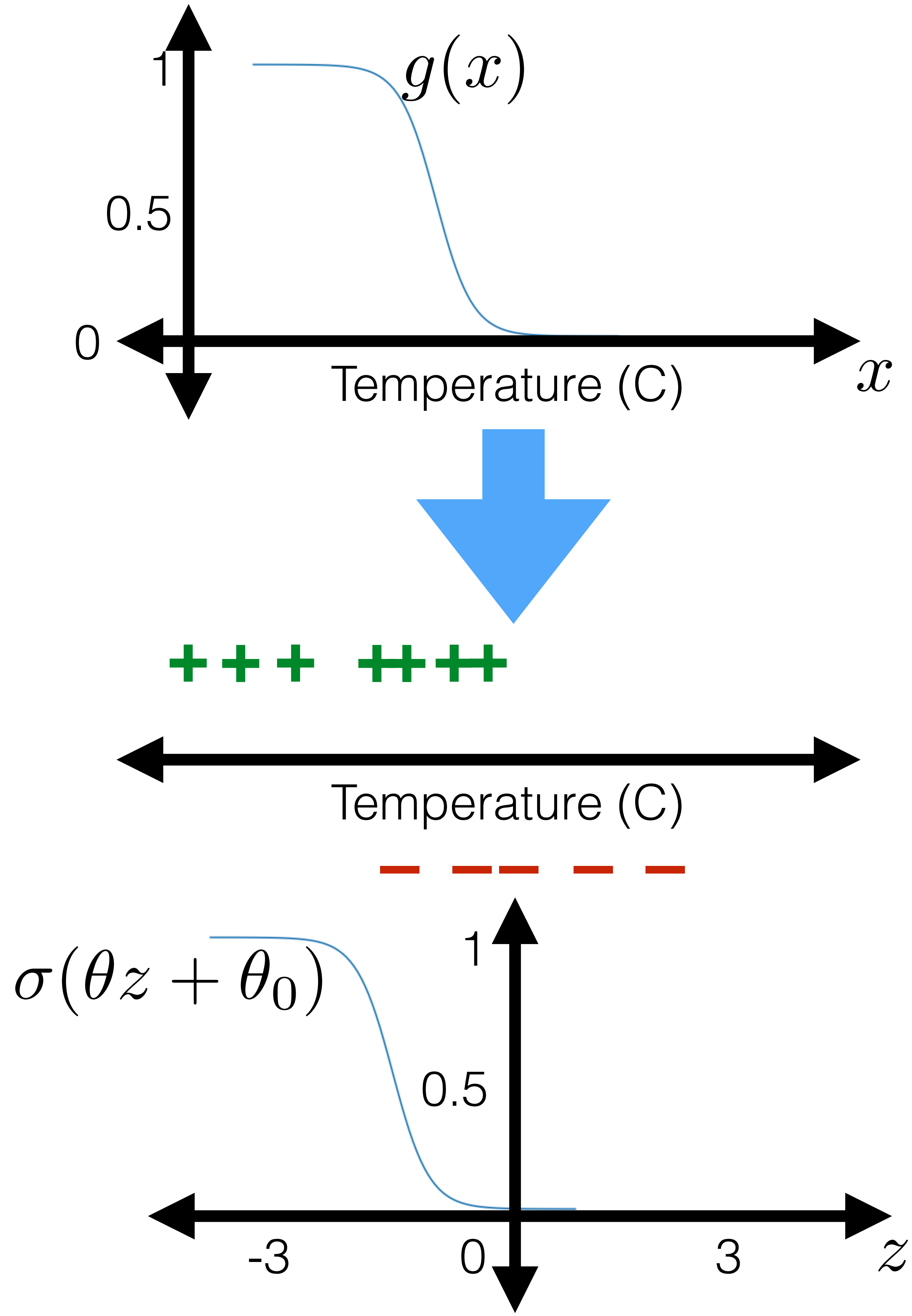


- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Capturing uncertainty

$$g(x) = \sigma(\theta x + \theta_0)$$

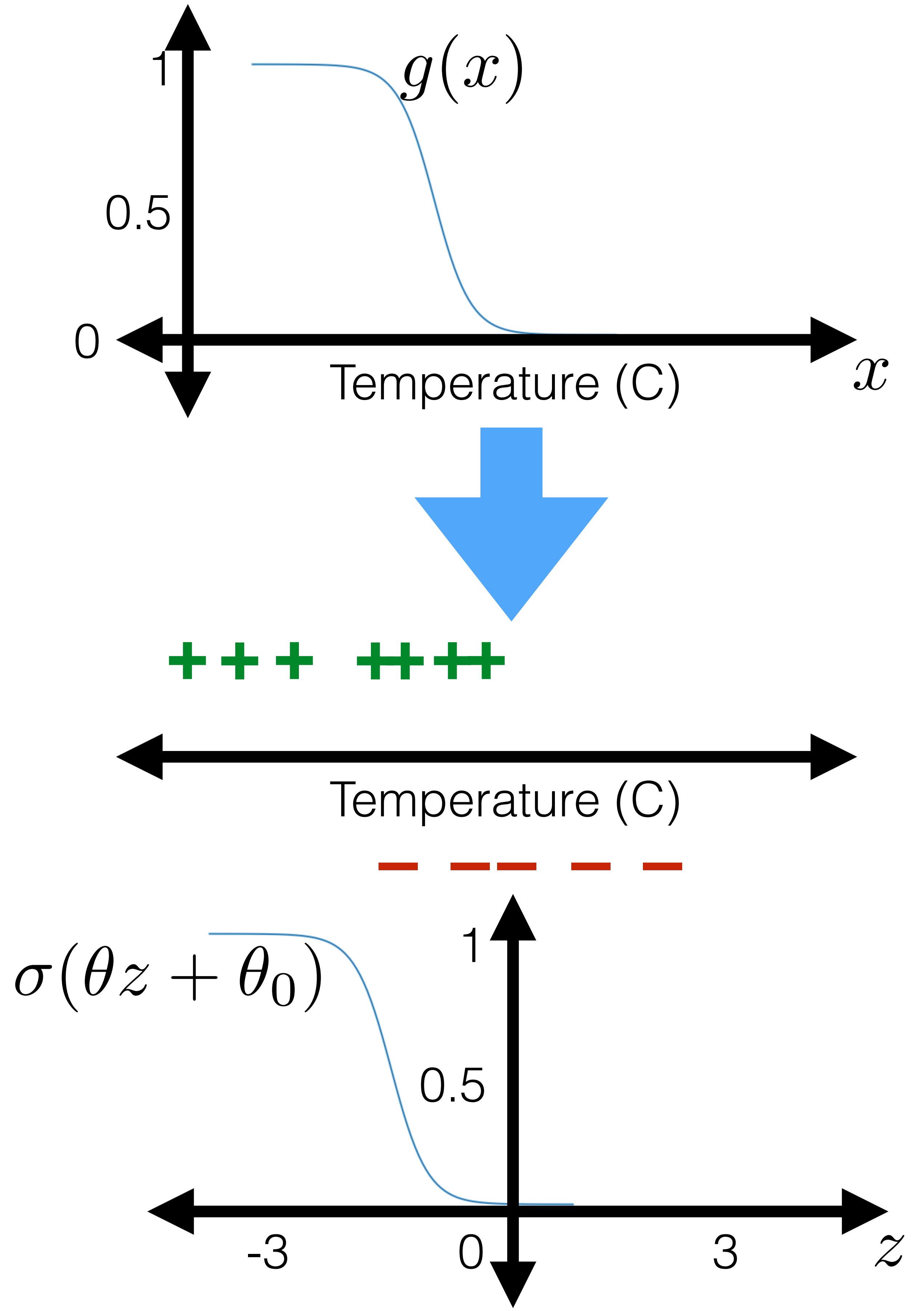


- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Capturing uncertainty

$$g(x) = \sigma(\theta x + \theta_0) = \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$



- How to make this shape?
  - Sigmoid/logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Capturing uncertainty

Capturing uncertainty

1 feature:

Capturing uncertainty

1 feature:

$$g(x) = \sigma(\theta x + \theta_0)$$

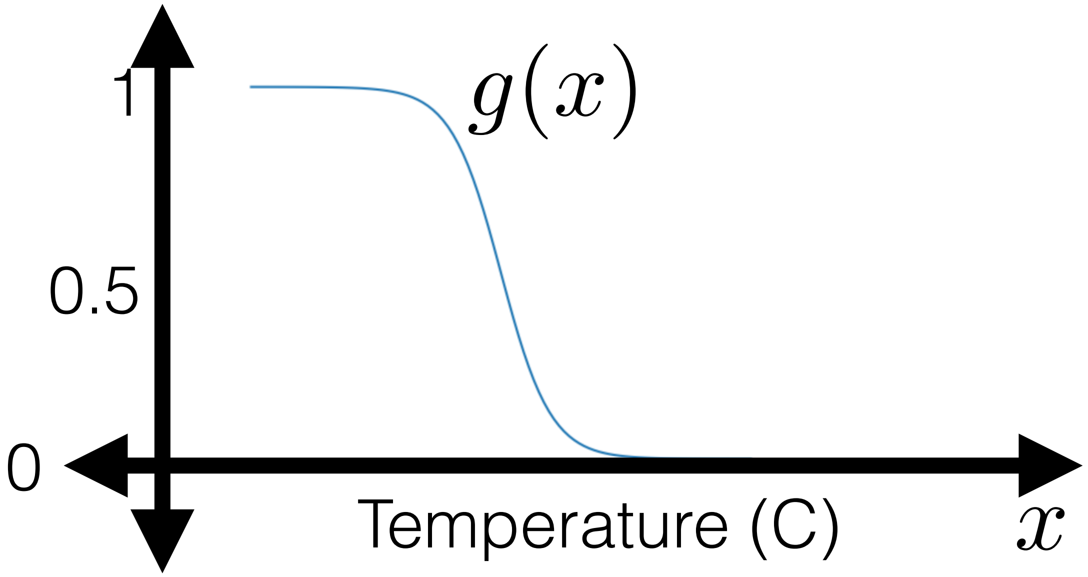
$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$

Capturing uncertainty

1 feature:

$$g(x) = \sigma(\theta x + \theta_0)$$

$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$

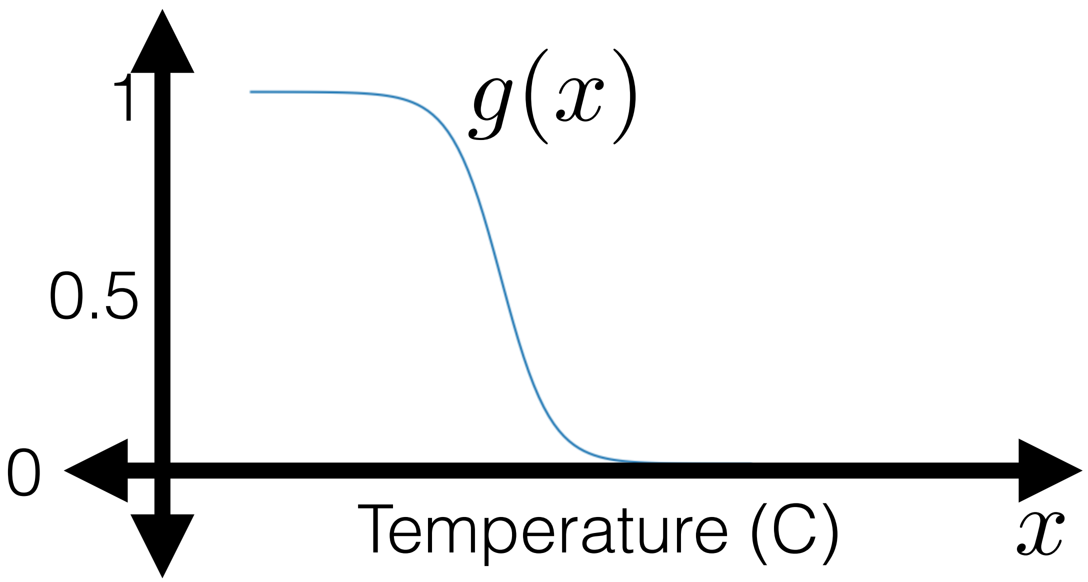


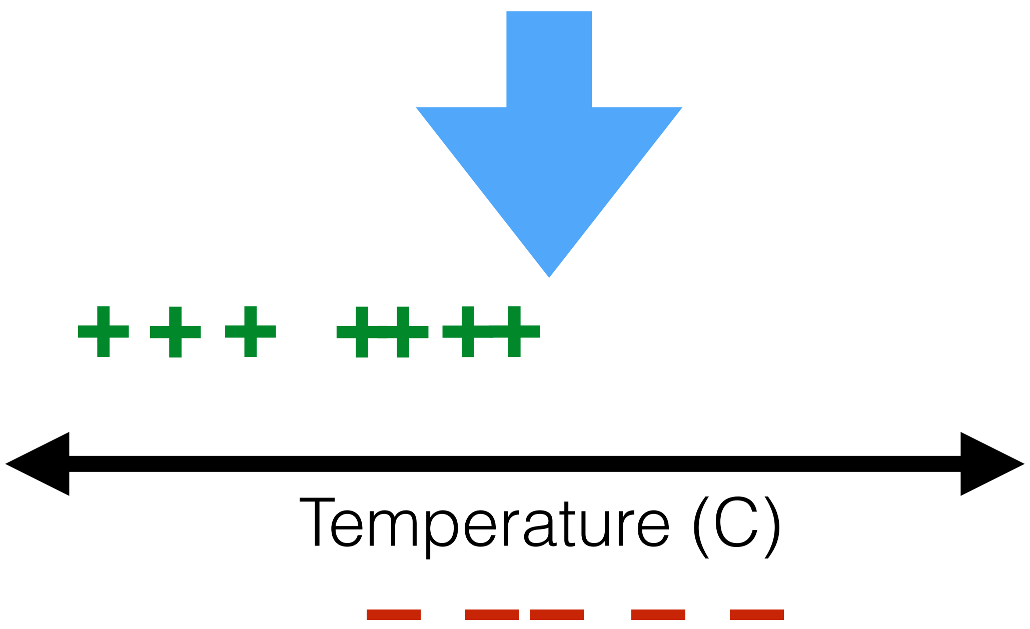
Capturing uncertainty

1 feature:

$$g(x) = \sigma(\theta x + \theta_0)$$

$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$





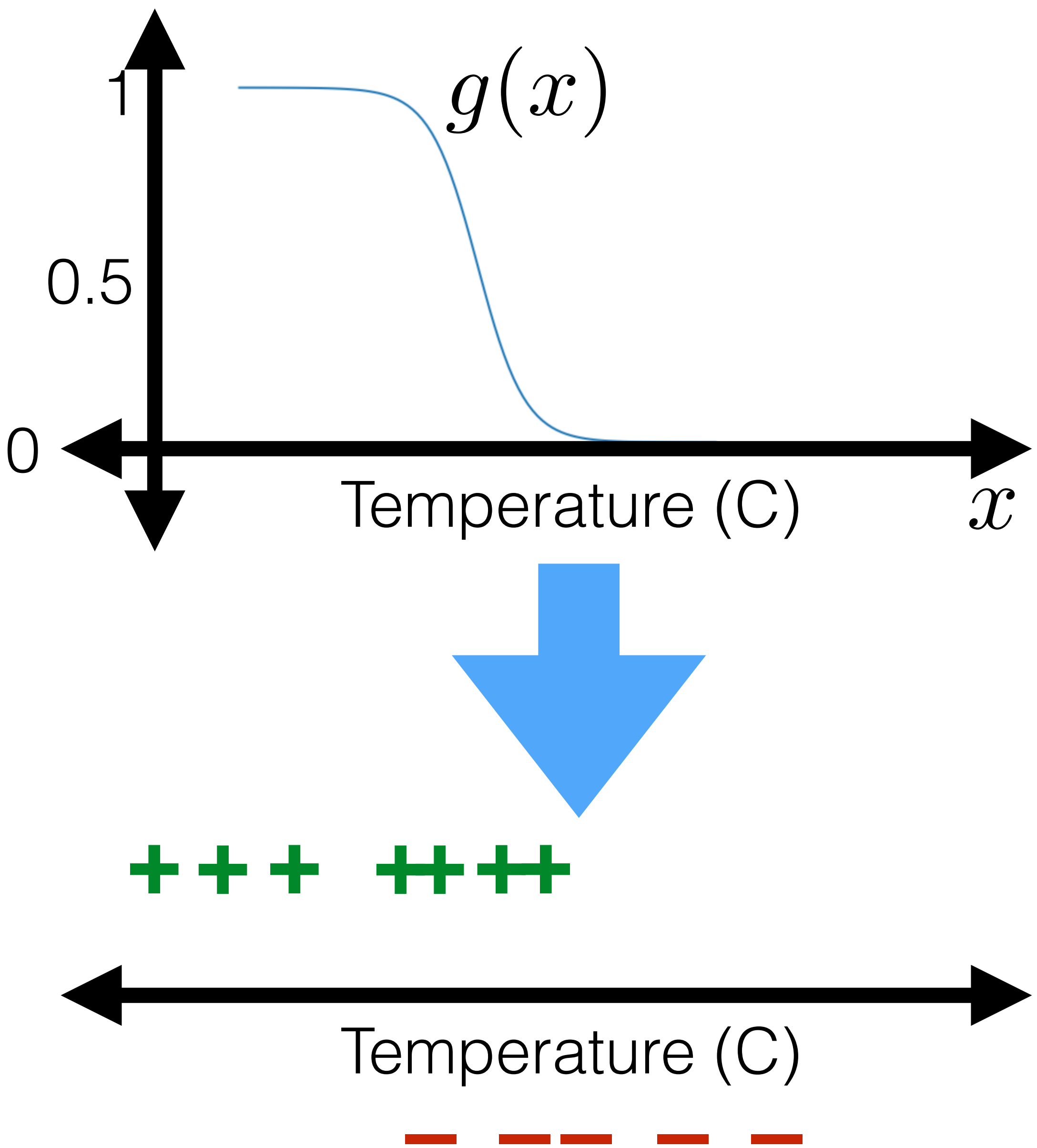
Capturing uncertainty

2 features:

1 feature:

$$g(x) = \sigma(\theta x + \theta_0)$$

$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$



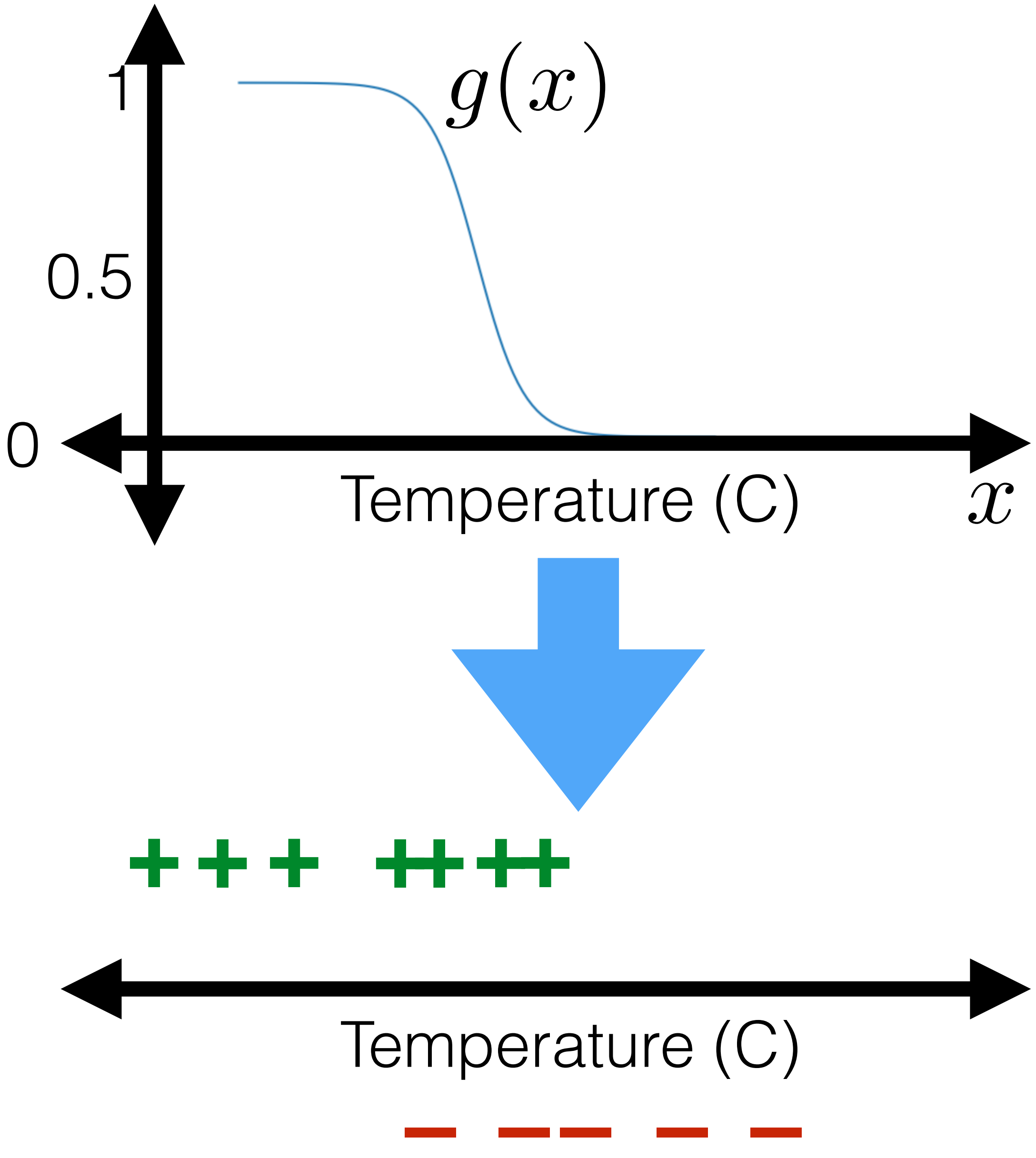
Capturing uncertainty

2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0)$$
$$= \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$

1 feature:

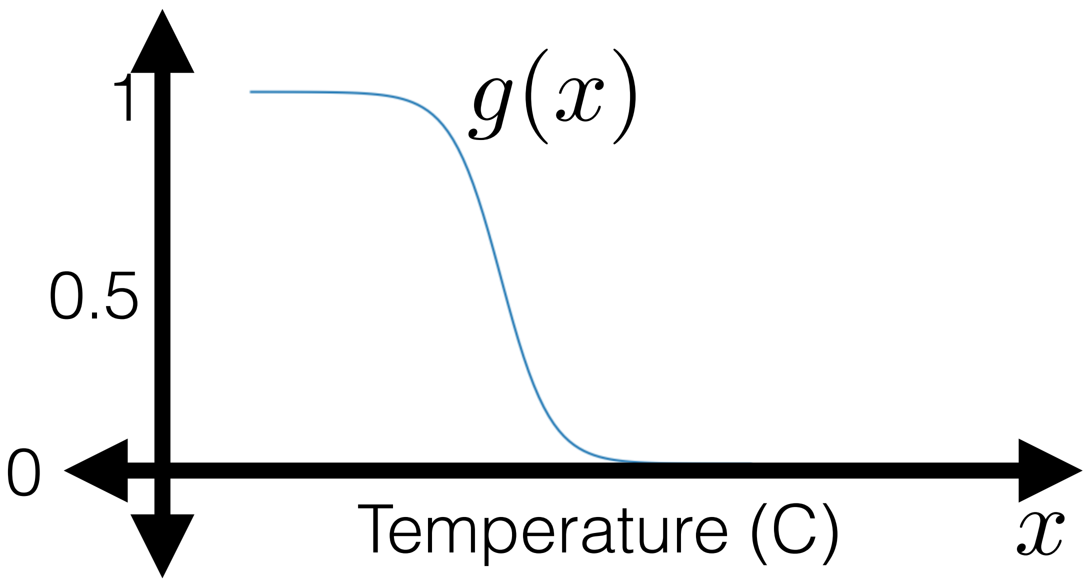
$$g(x) = \sigma(\theta x + \theta_0)$$
$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$

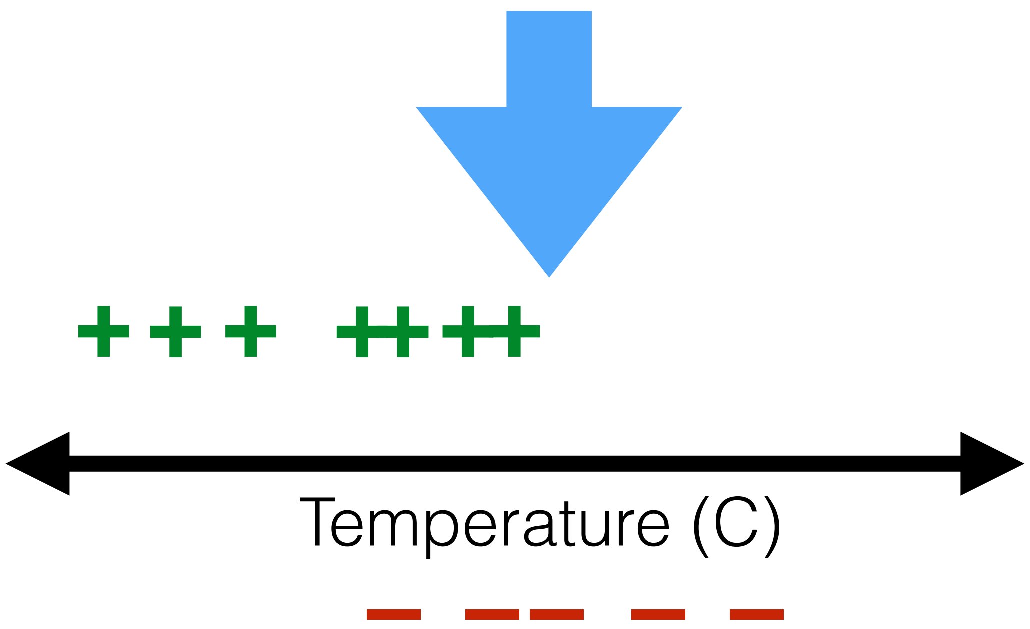


Capturing uncertainty

1 feature:

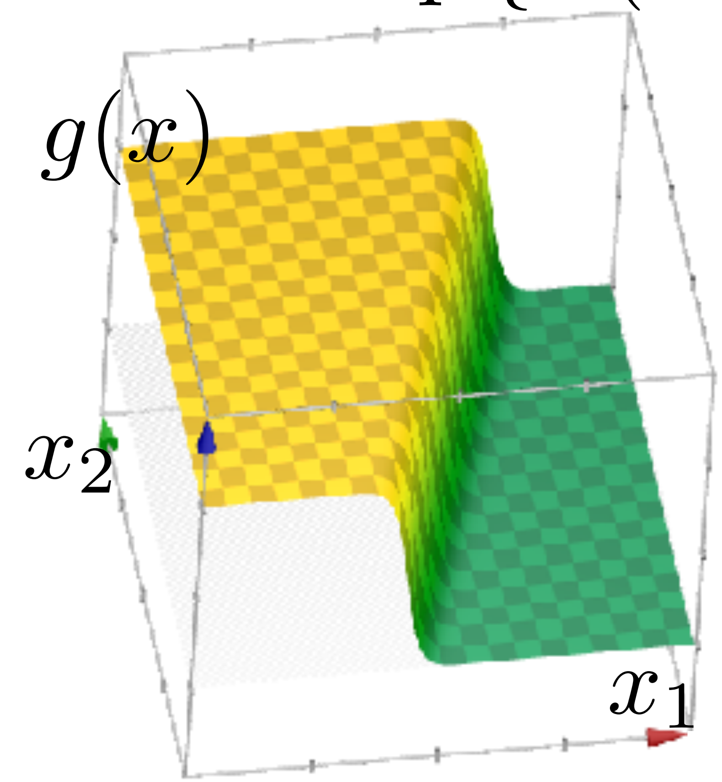
$$g(x) = \sigma(\theta x + \theta_0)$$
$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$





2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0)$$
$$= \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$

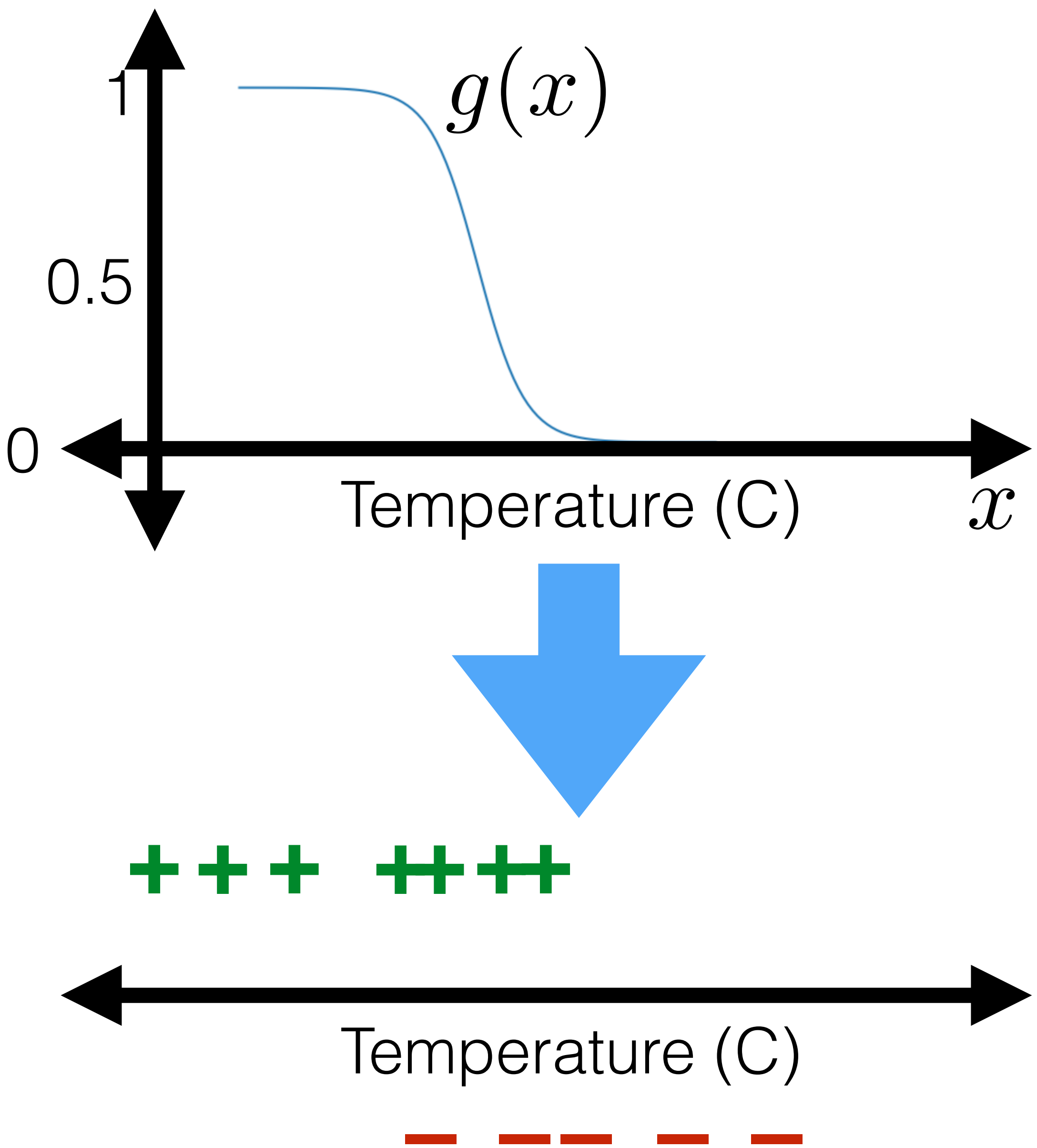


Capturing uncertainty

1 feature:

$$g(x) = \sigma(\theta x + \theta_0)$$

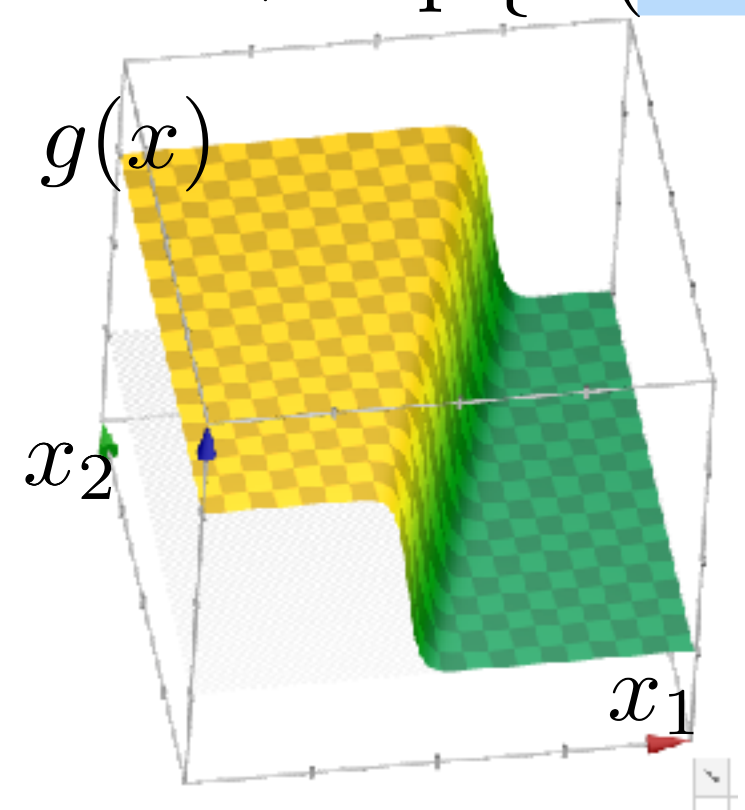
$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$



2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0)$$

$$= \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$

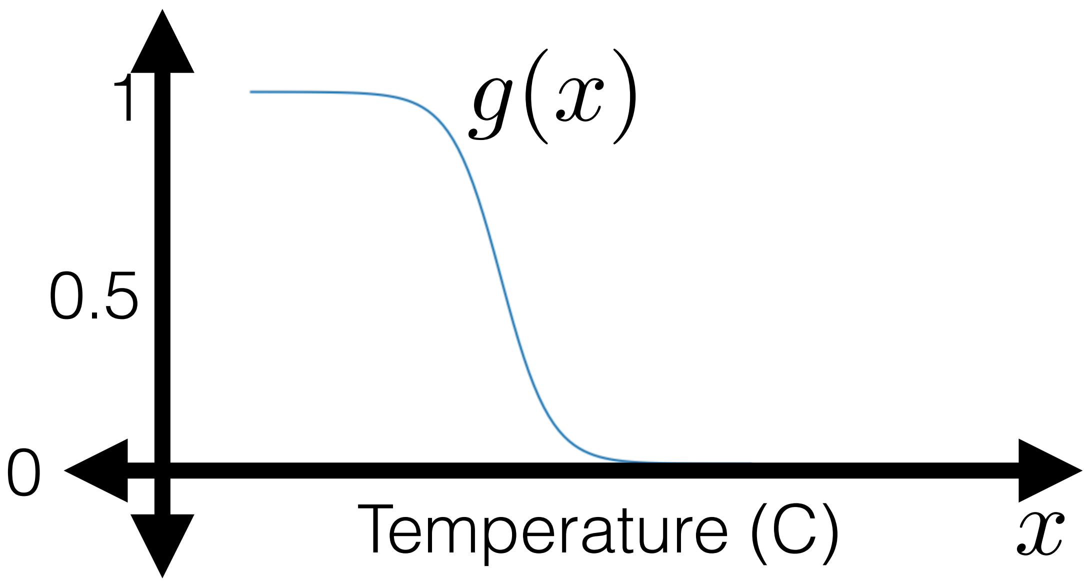


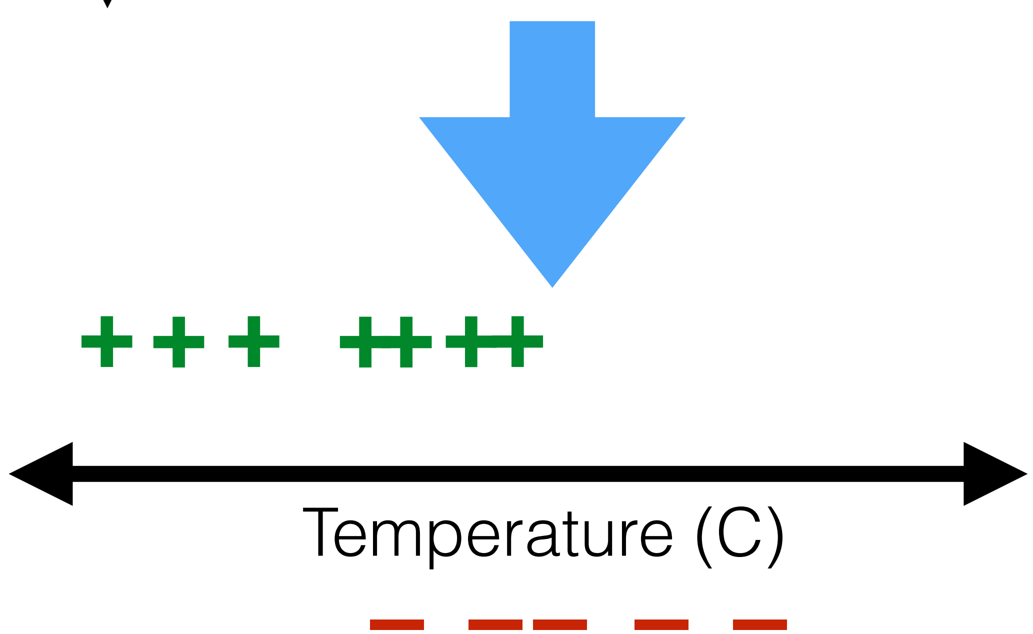
Capturing uncertainty

1 feature:

$$g(x) = \sigma(\theta x + \theta_0)$$

$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$

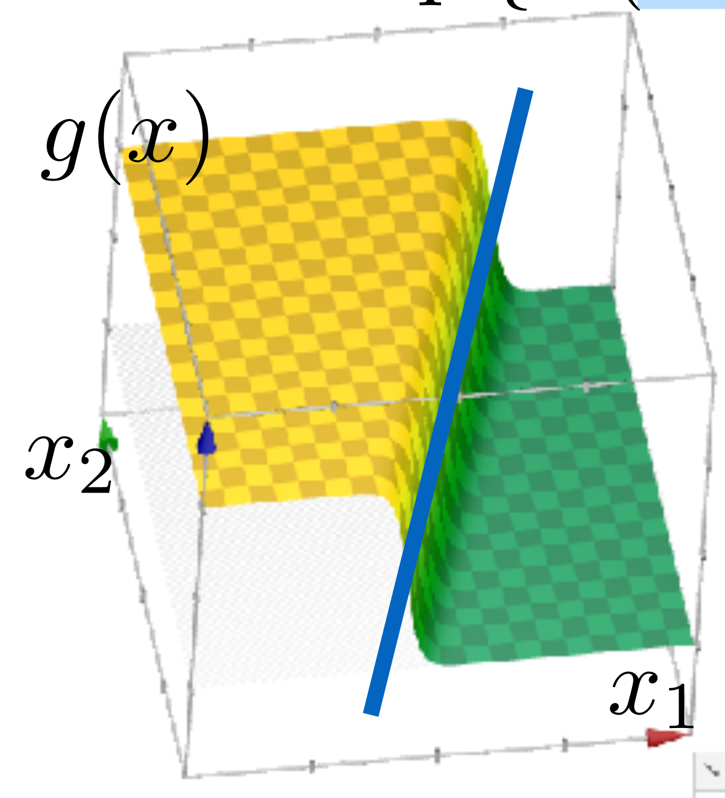




2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0)$$

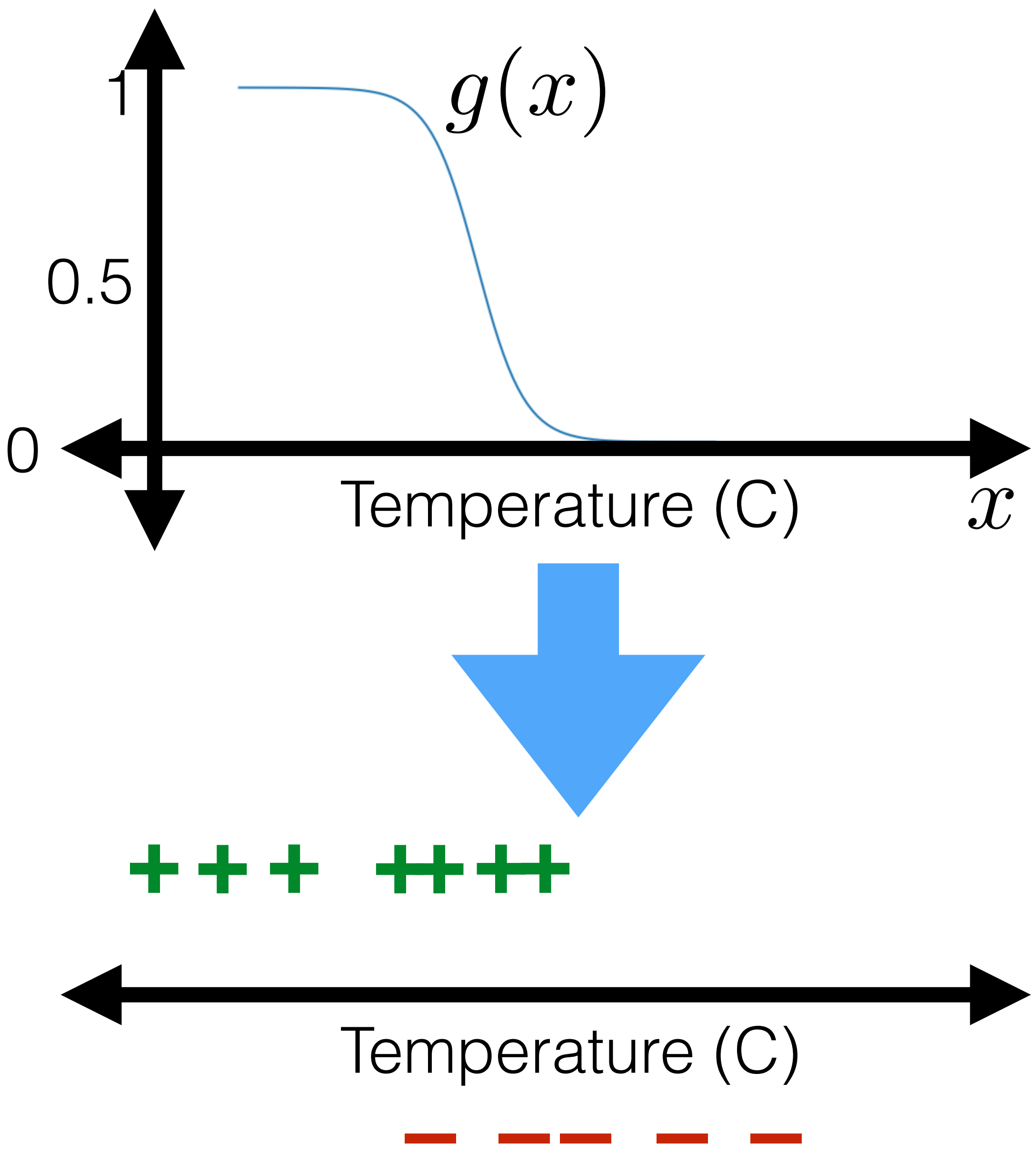
$$= \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$



Capturing uncertainty

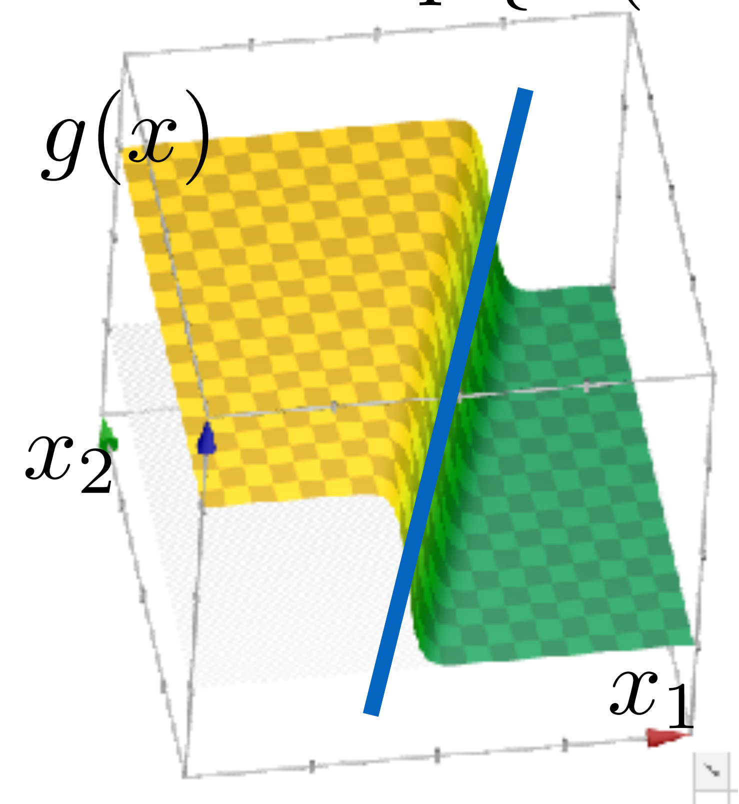
1 feature:

$$g(x) = \sigma(\theta x + \theta_0)$$
$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$



2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0)$$
$$= \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$

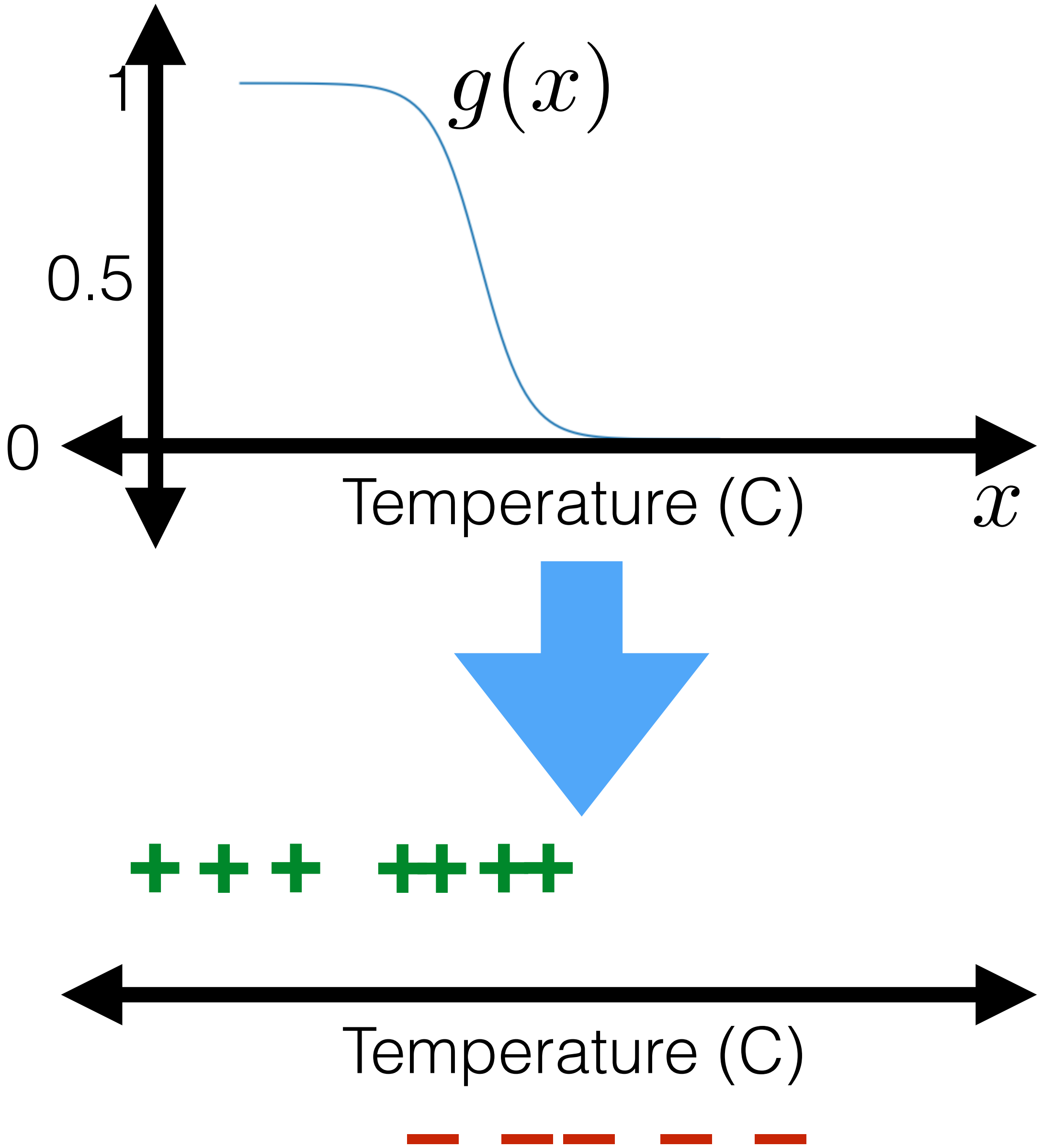


Capturing uncertainty

1 feature:

$$g(x) = \sigma(\theta x + \theta_0)$$

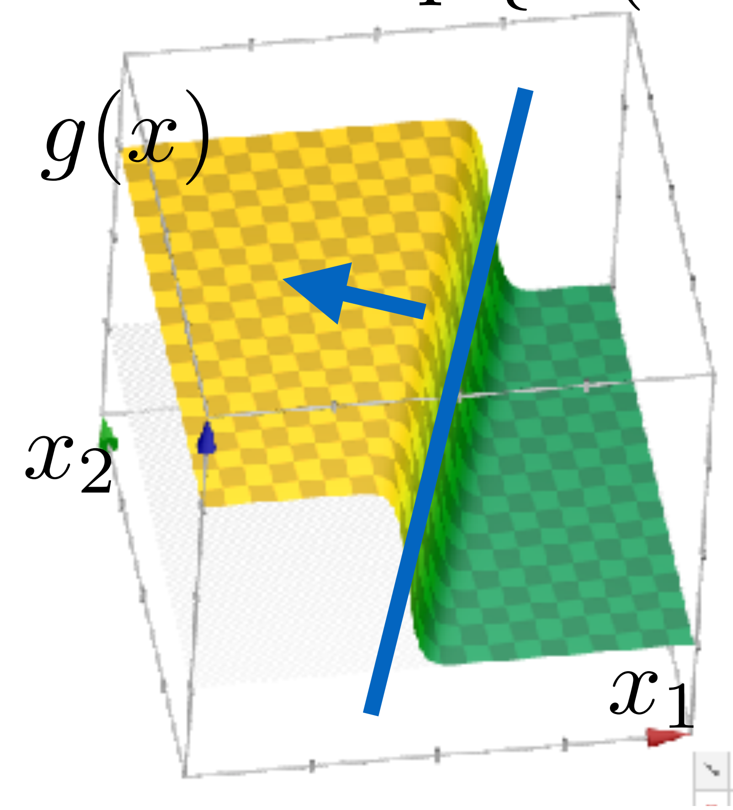
$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$



2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0)$$

$$= \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$

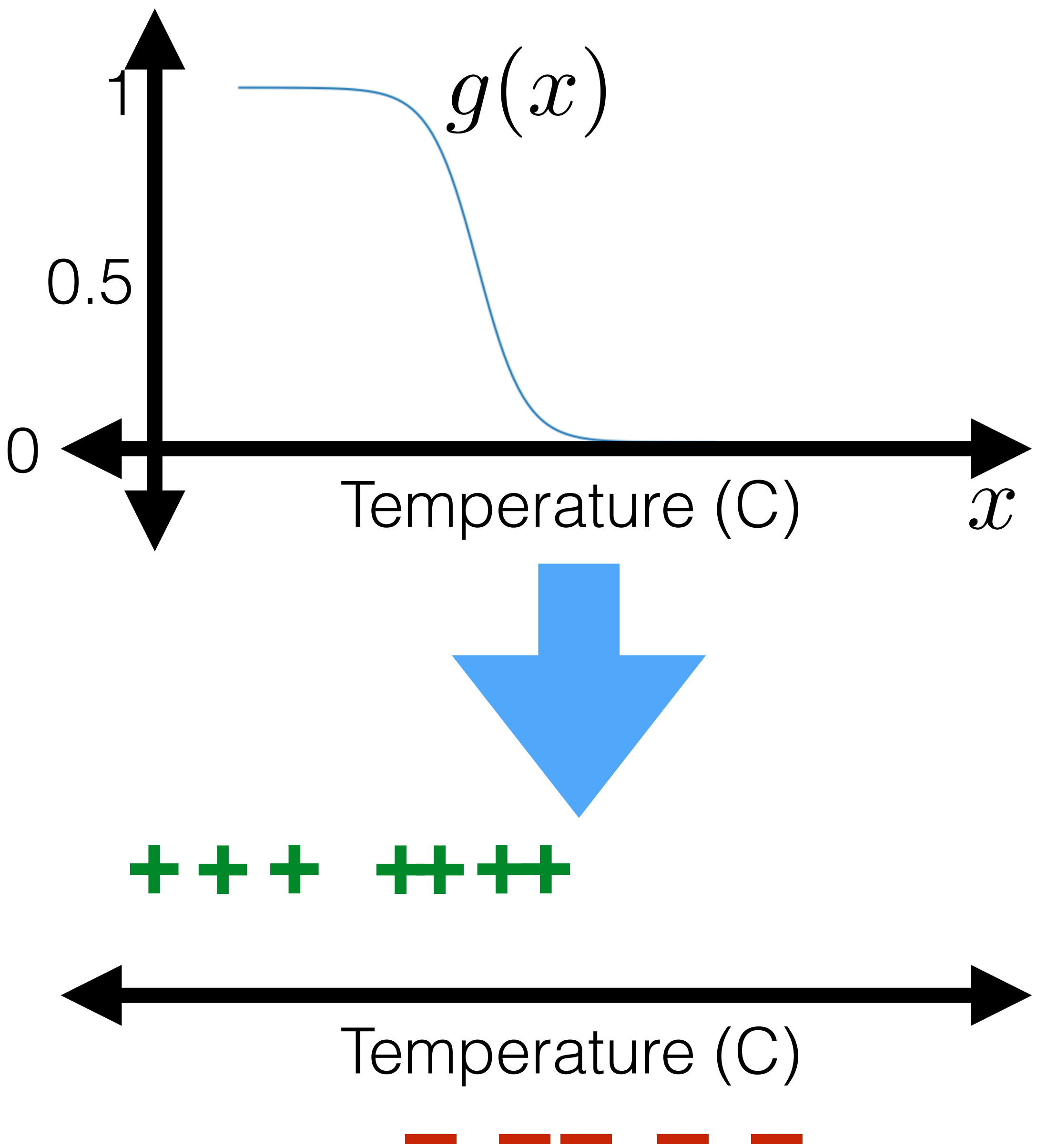


Capturing uncertainty

1 feature:

$$g(x) = \sigma(\theta x + \theta_0)$$

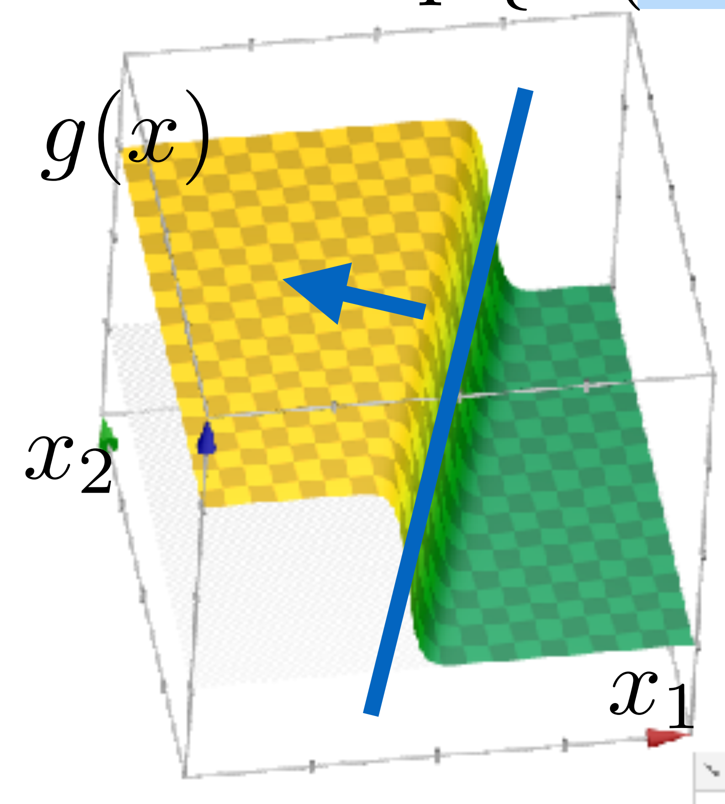
$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$



2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0)$$

$$= \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$



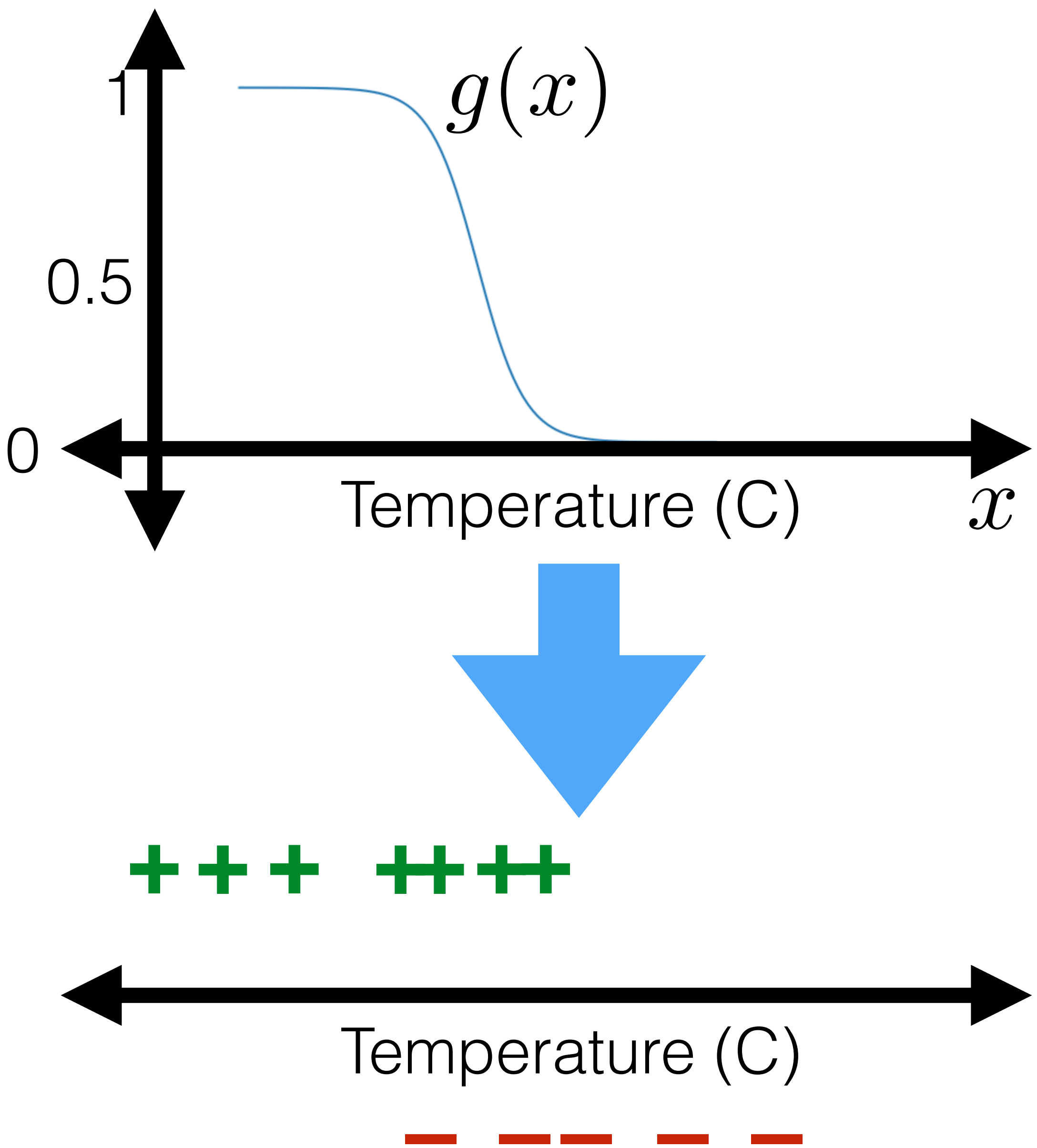
Capturing uncertainty

1 feature:

$$g(x) = \sigma(\theta x + \theta_0)$$

$$g(x) =$$

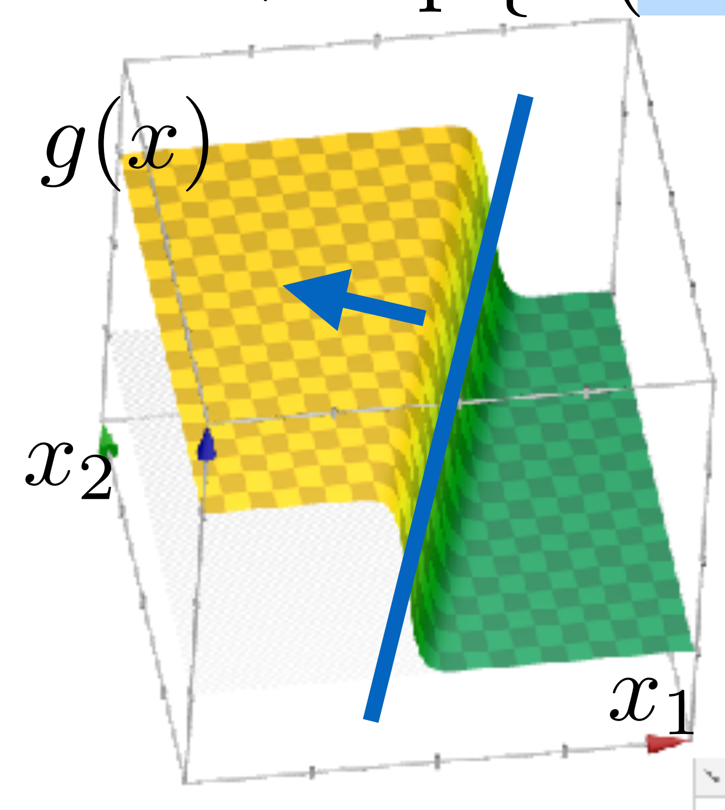
$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$



2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0)$$

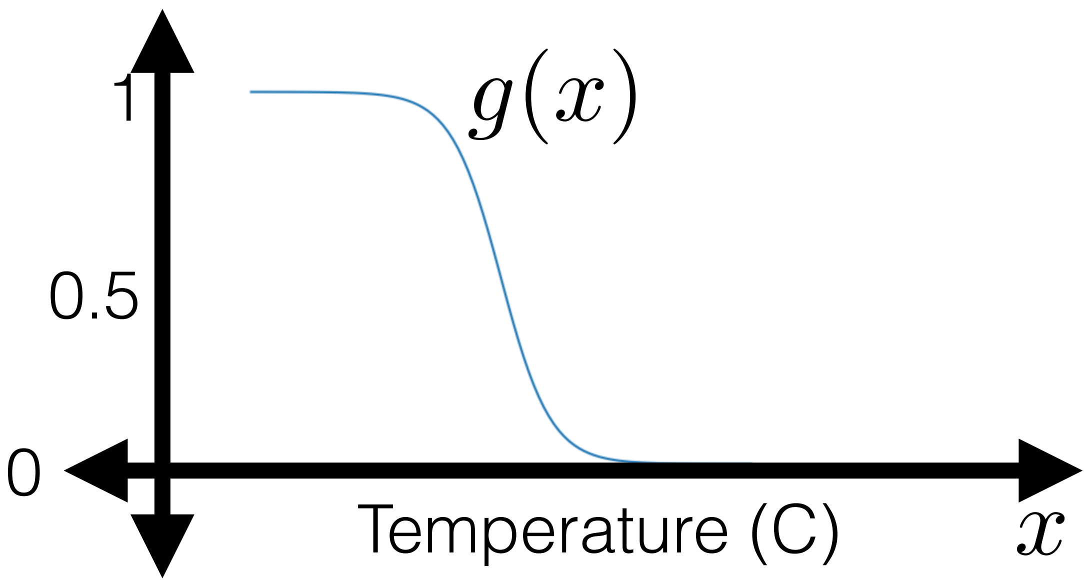
$$= \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$

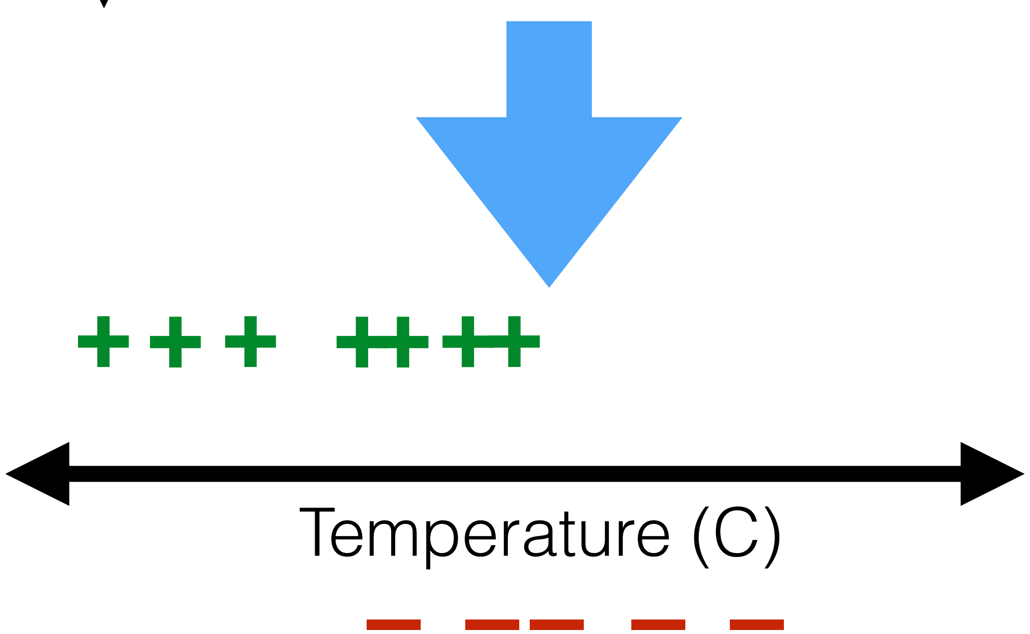


Capturing uncertainty

1 feature:

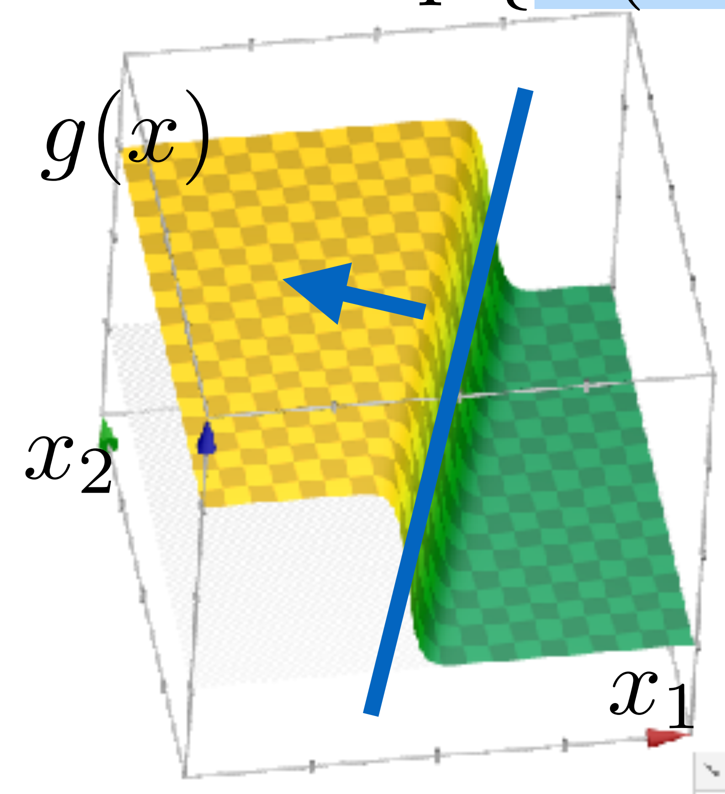
$$g(x) = \sigma(\theta x + \theta_0) = \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$





2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0) = \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$

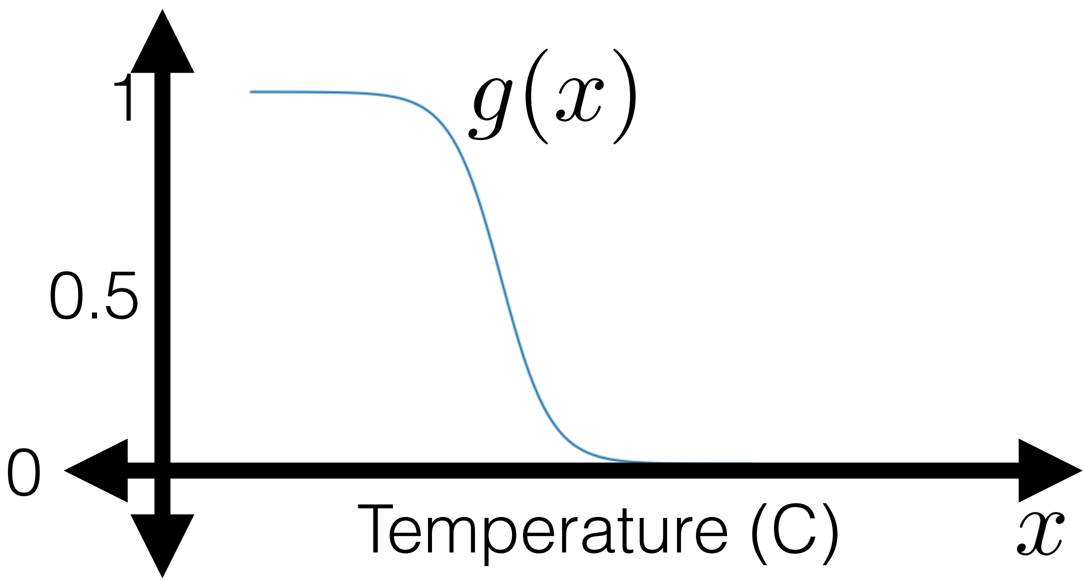


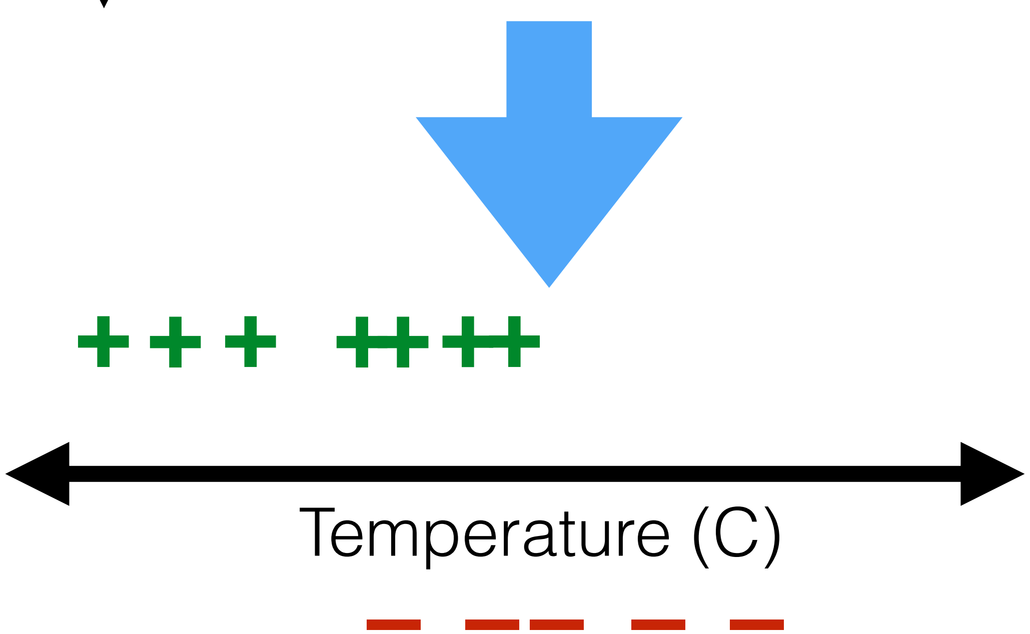
Capturing uncertainty

1 feature:

$$g(x) = \sigma(\theta x + \theta_0)$$

$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$

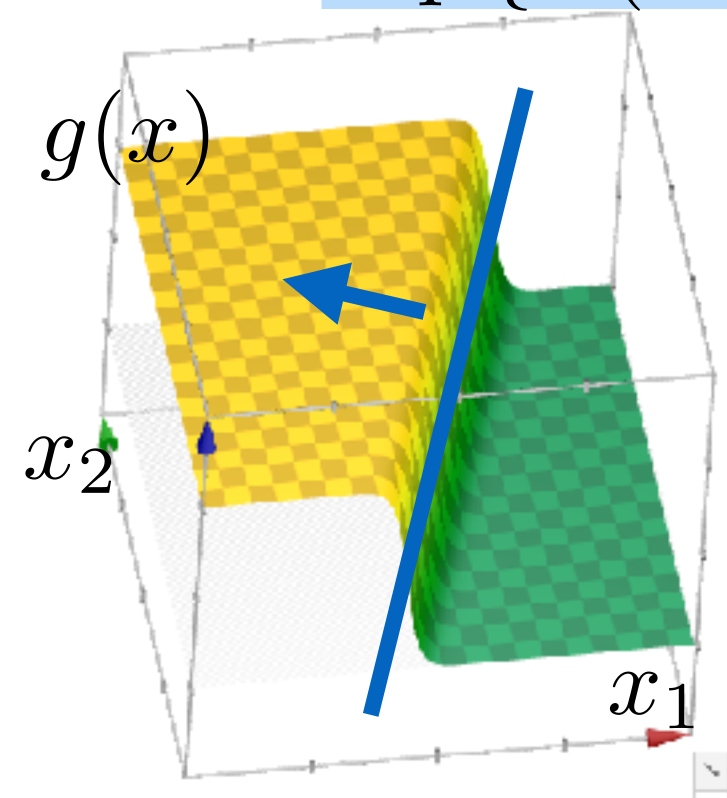




2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0)$$

$$= \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$



Capturing uncertainty

1 feature:

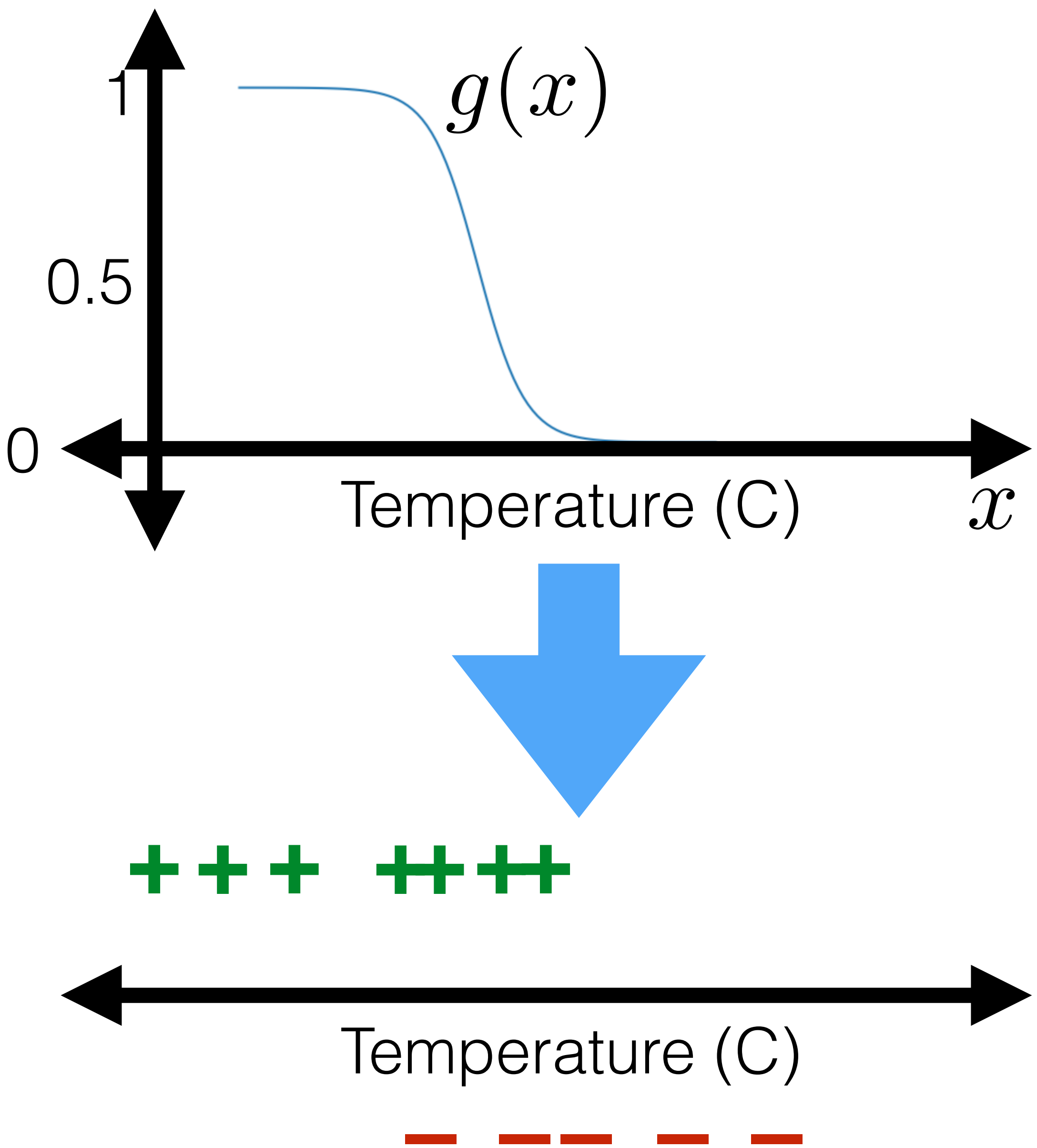
$$g(x) = \sigma(\theta x + \theta_0)$$

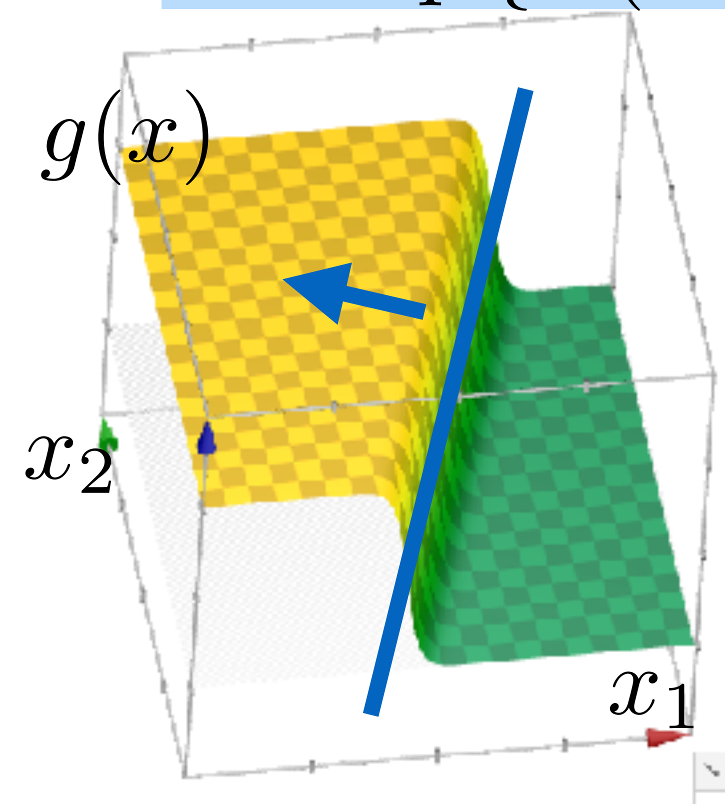
$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$

2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0)$$

$$= \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$

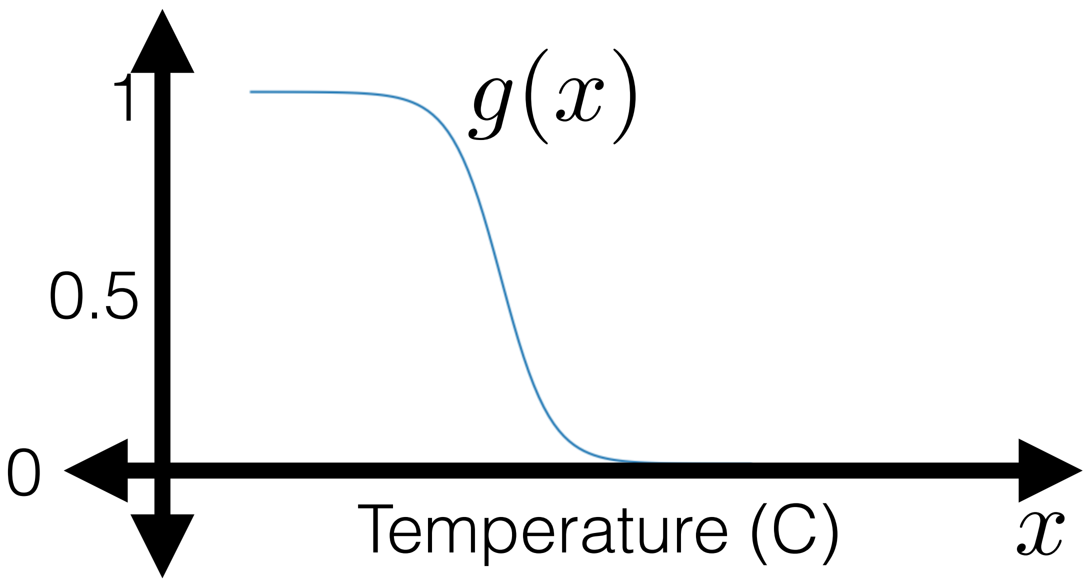


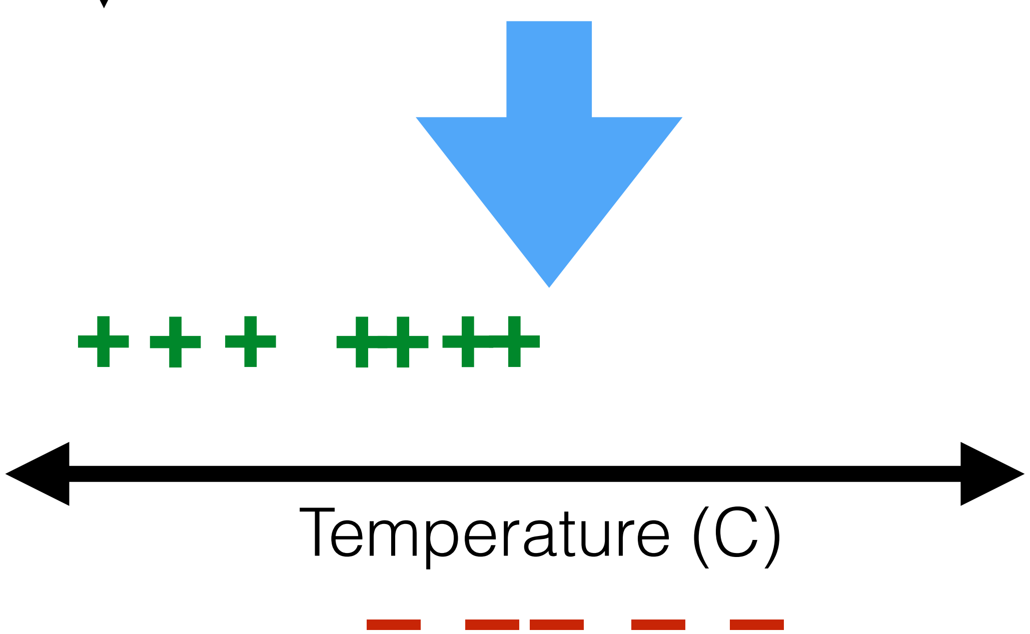


Capturing uncertainty

1 feature:

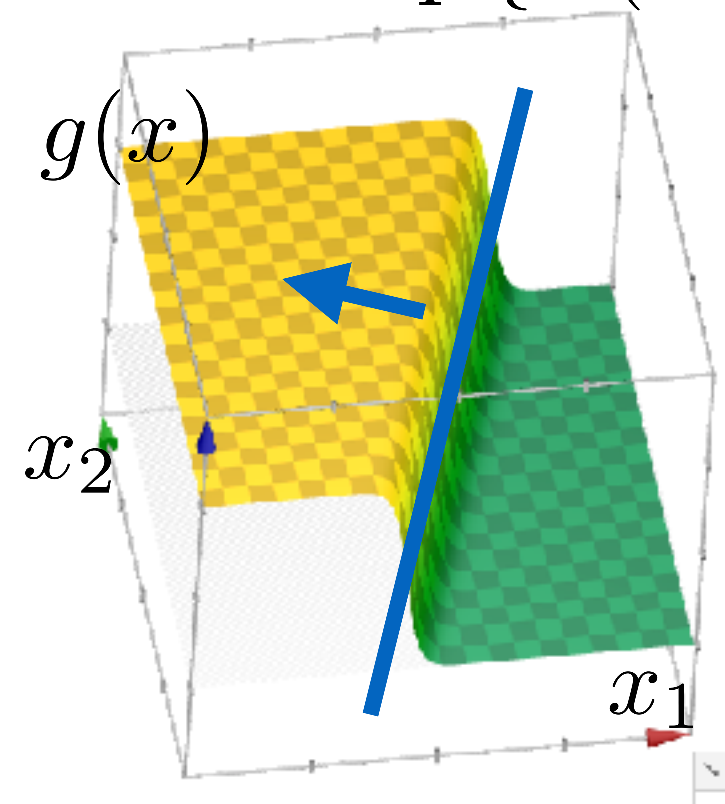
$$g(x) = \sigma(\theta x + \theta_0)$$
$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$





2 features:

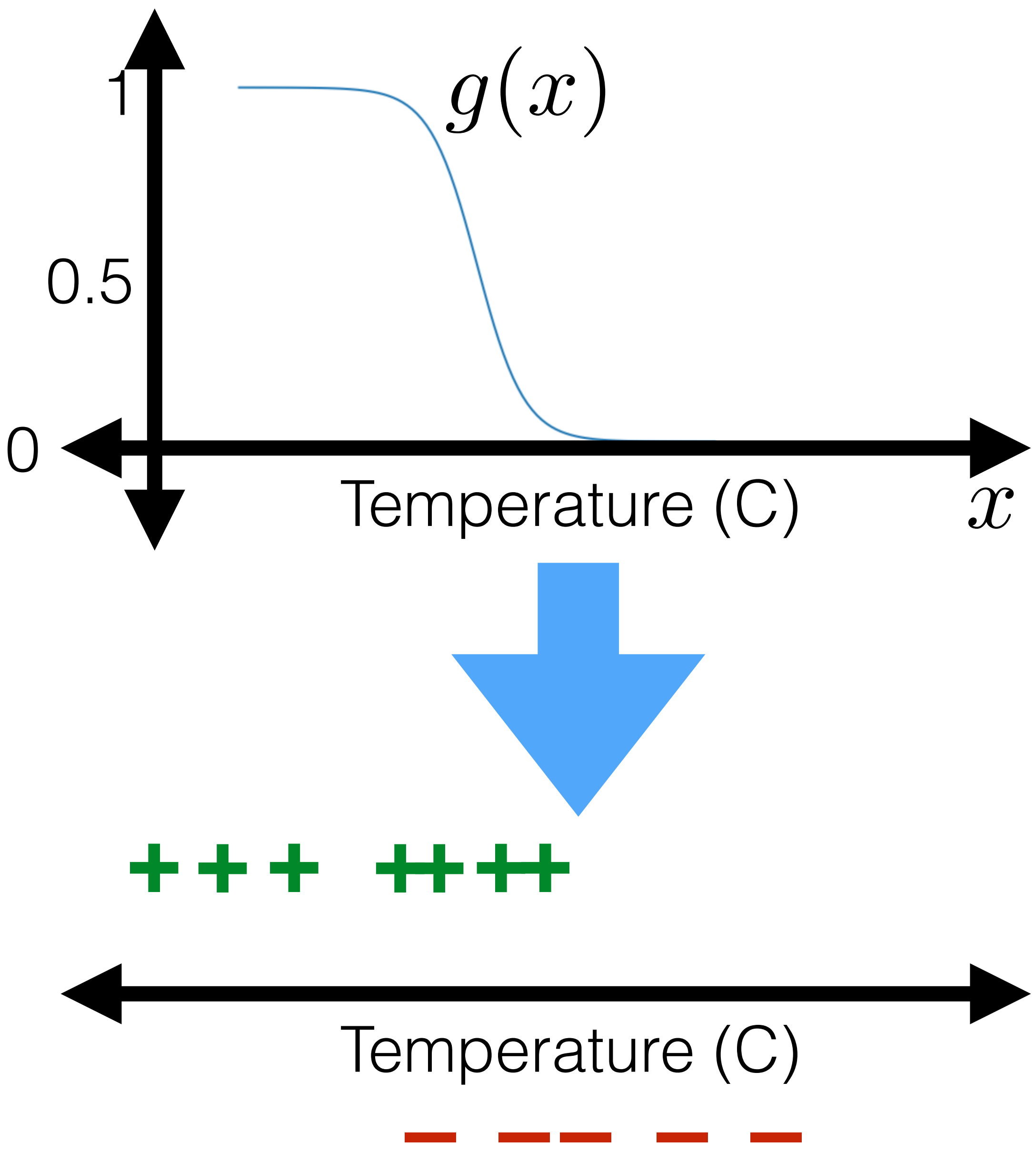
$$g(x) = \sigma(\theta^\top x + \theta_0)$$
$$= \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$



Capturing uncertainty

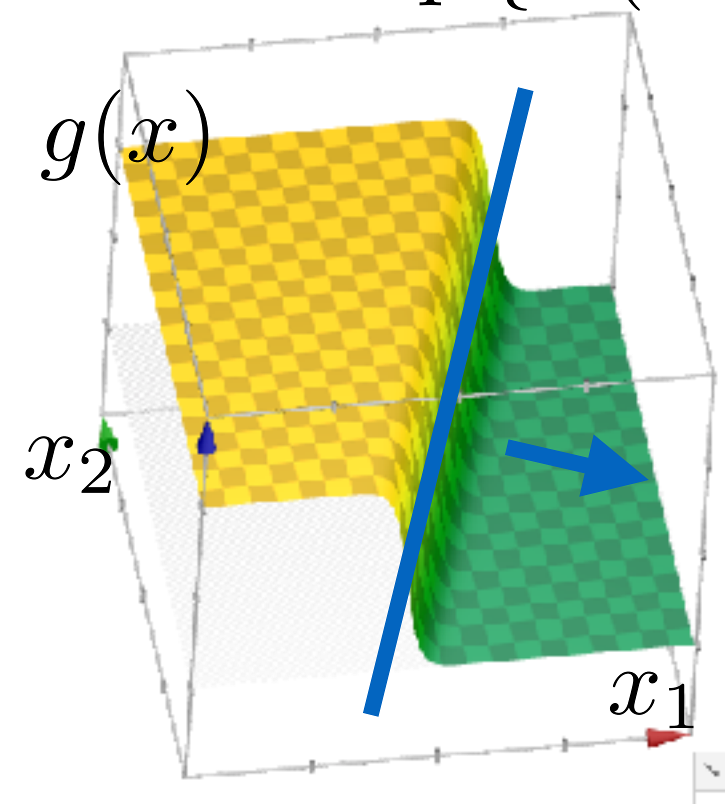
1 feature:

$$g(x) = \sigma(\theta x + \theta_0) = \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$



2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0) = \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$

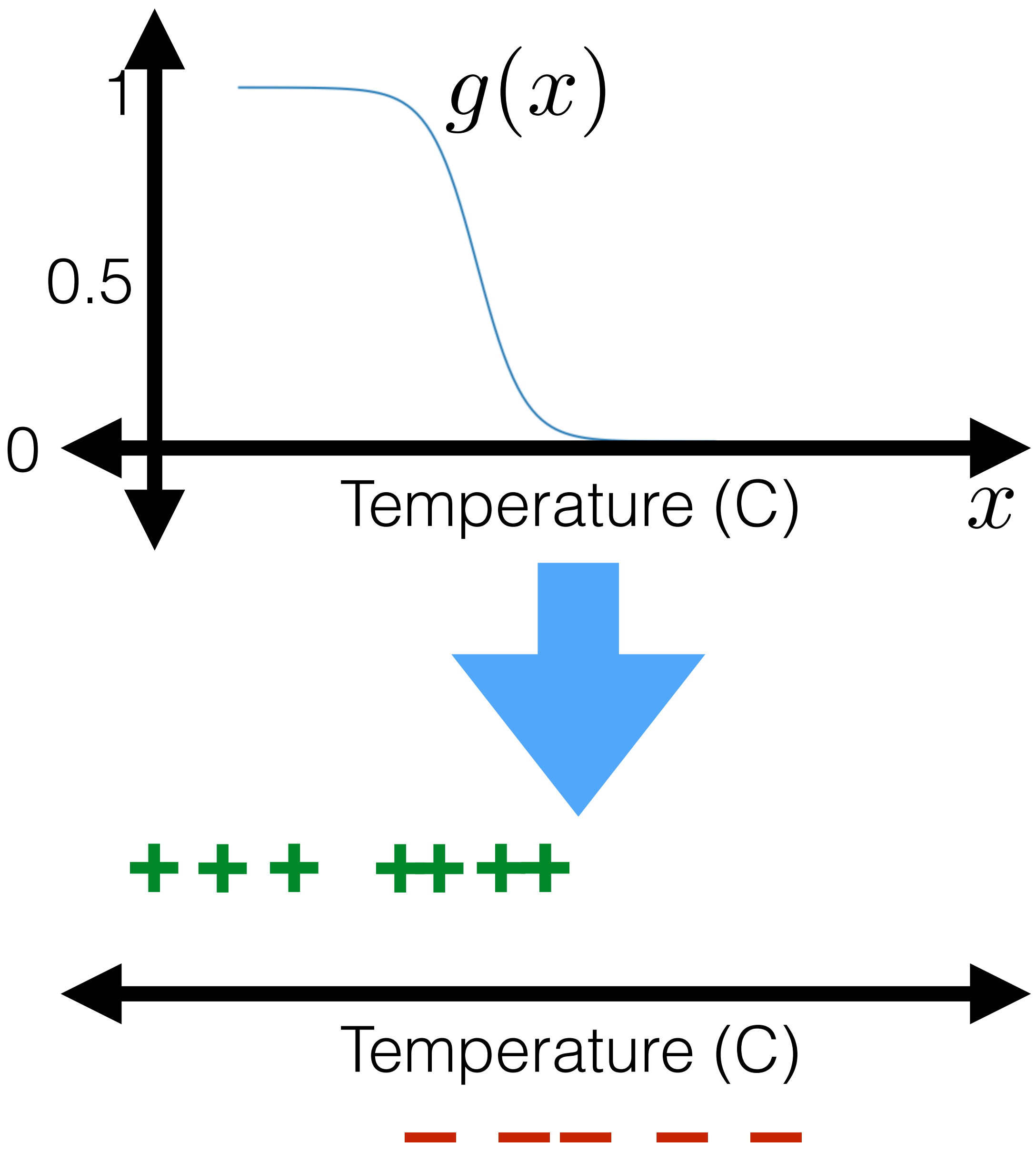


Capturing uncertainty

1 feature:

$$g(x) = \sigma(\theta x + \theta_0)$$

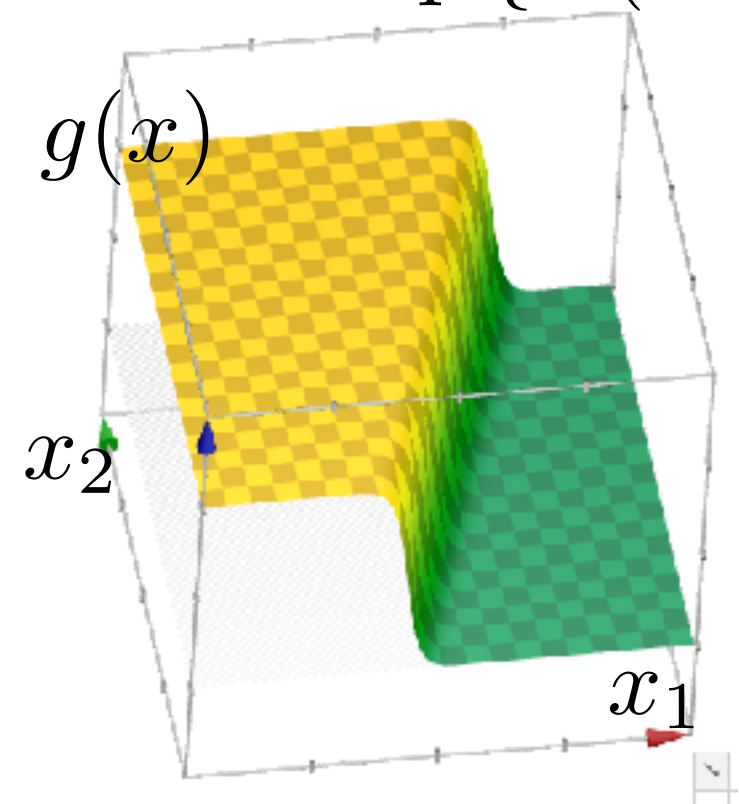
$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$



2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0)$$

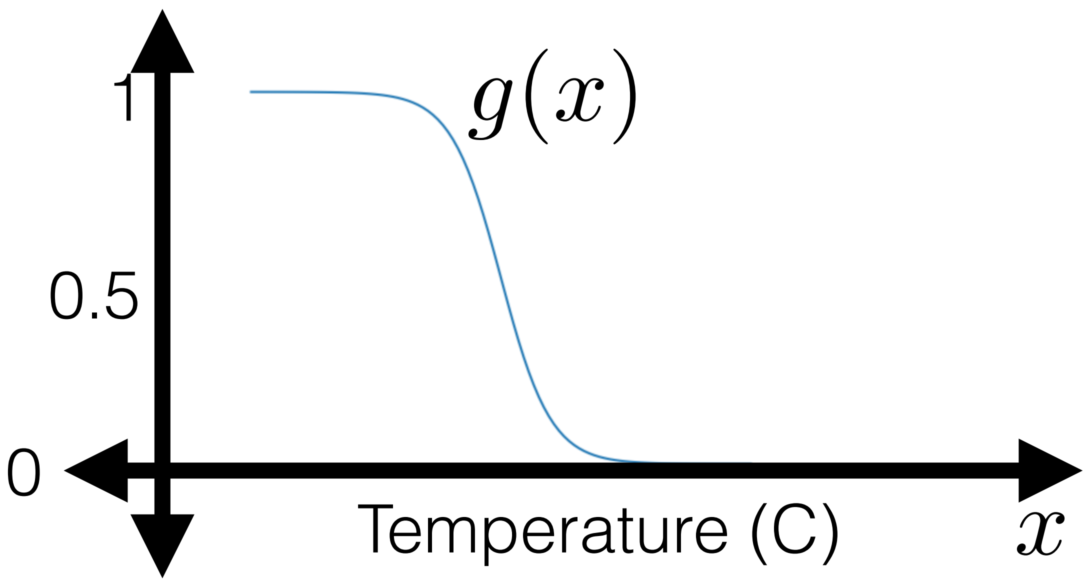
$$= \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$

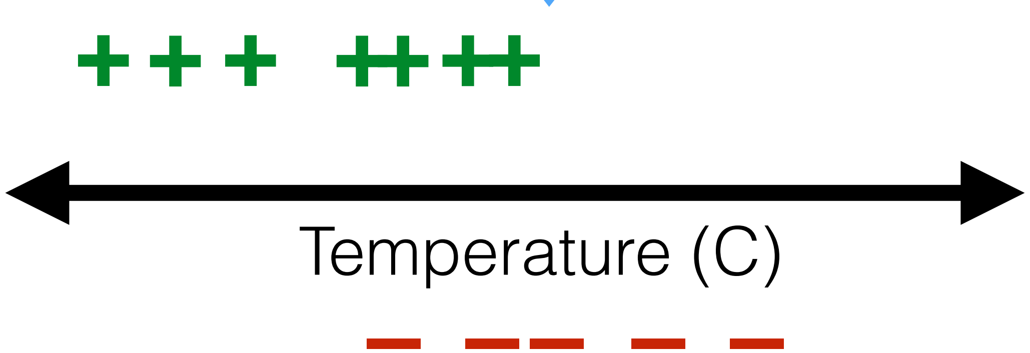


Capturing uncertainty

1 feature:

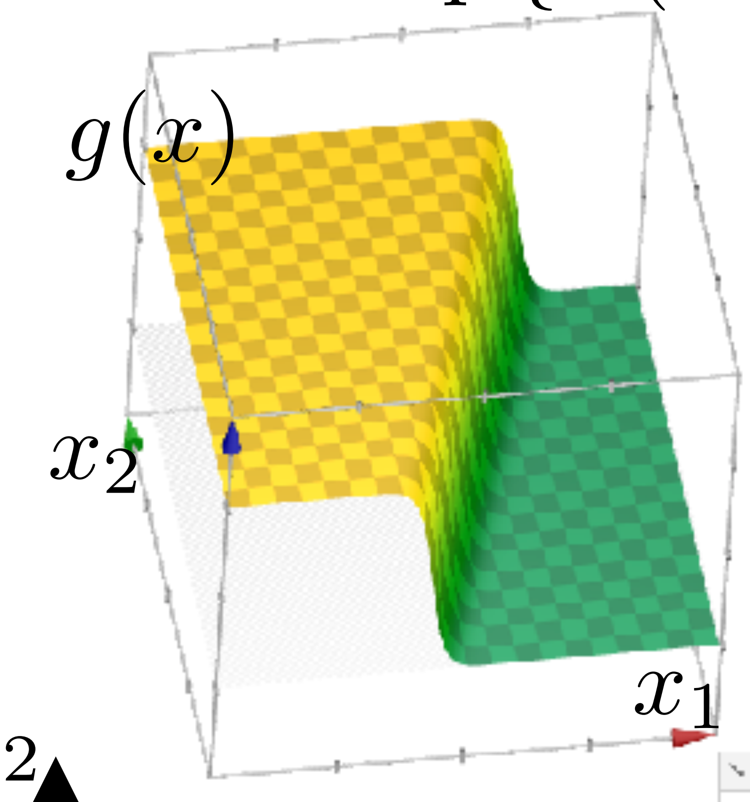
$$g(x) = \sigma(\theta x + \theta_0) = \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$

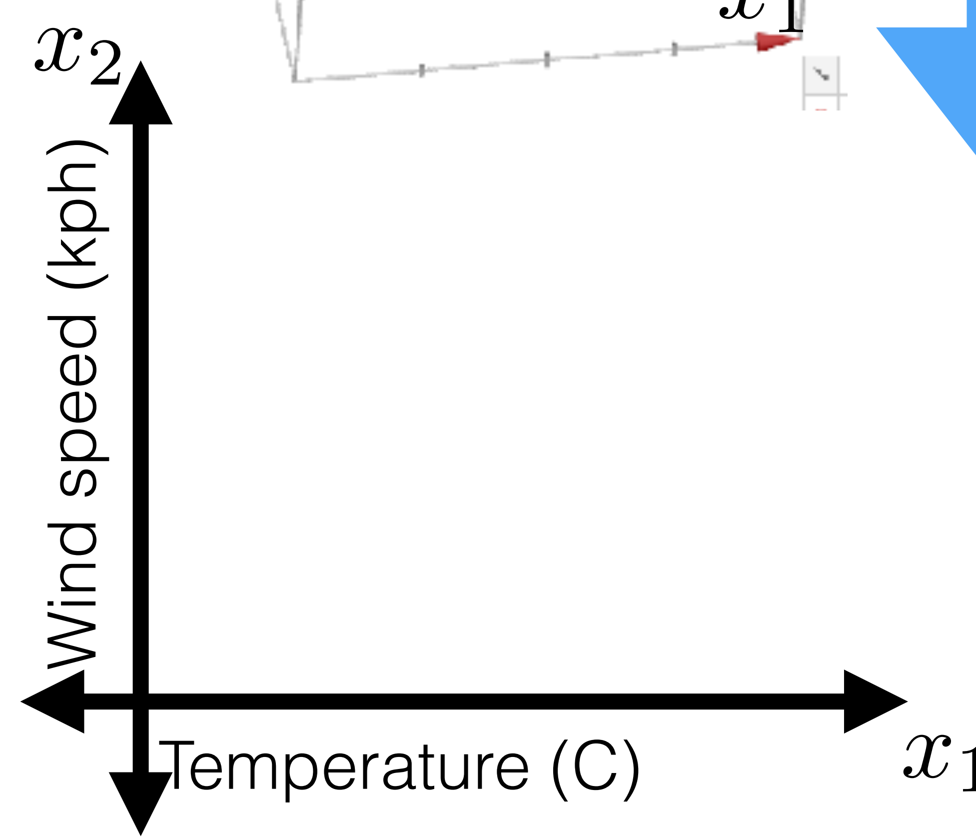




2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0) = \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$

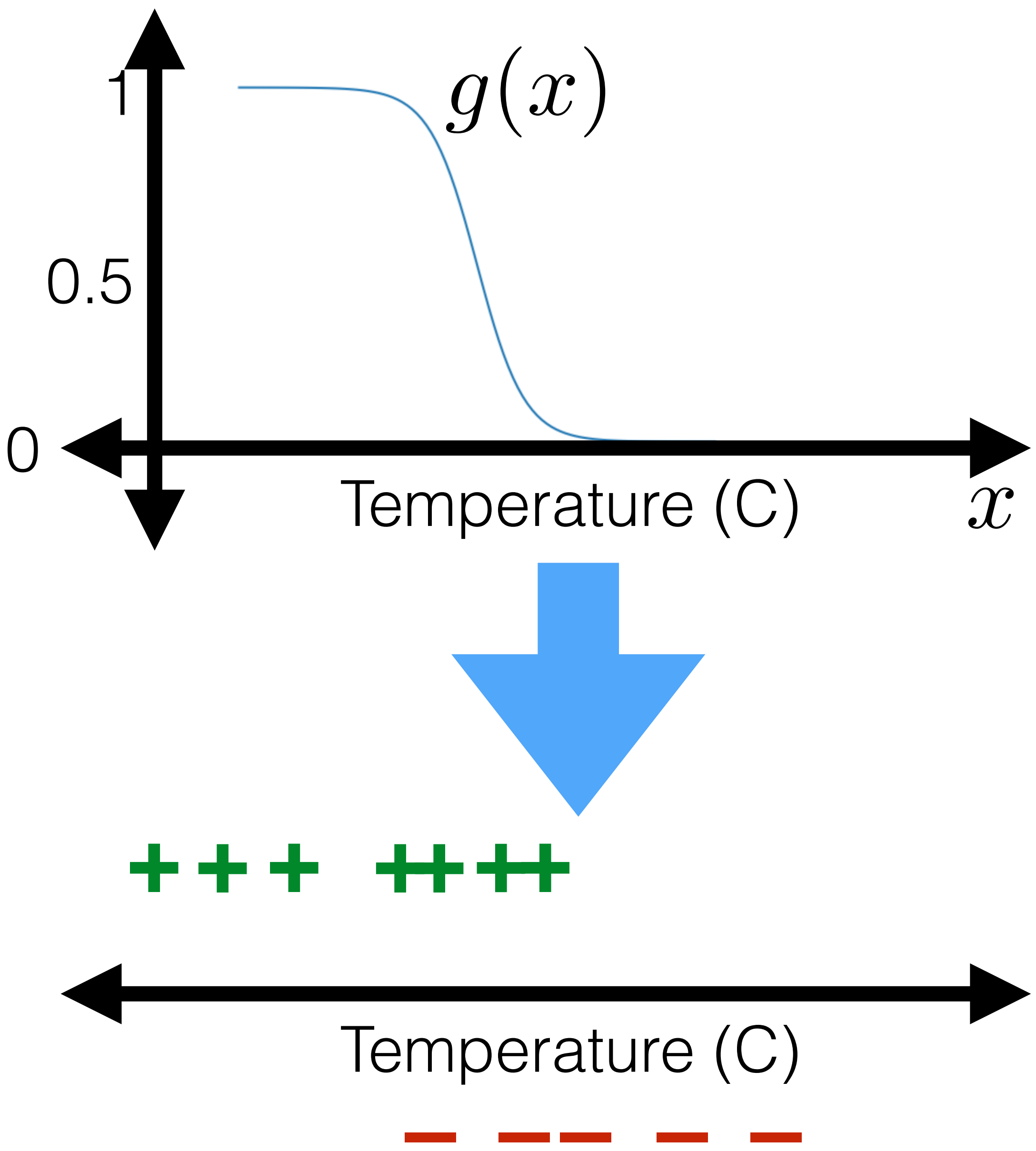




Capturing uncertainty

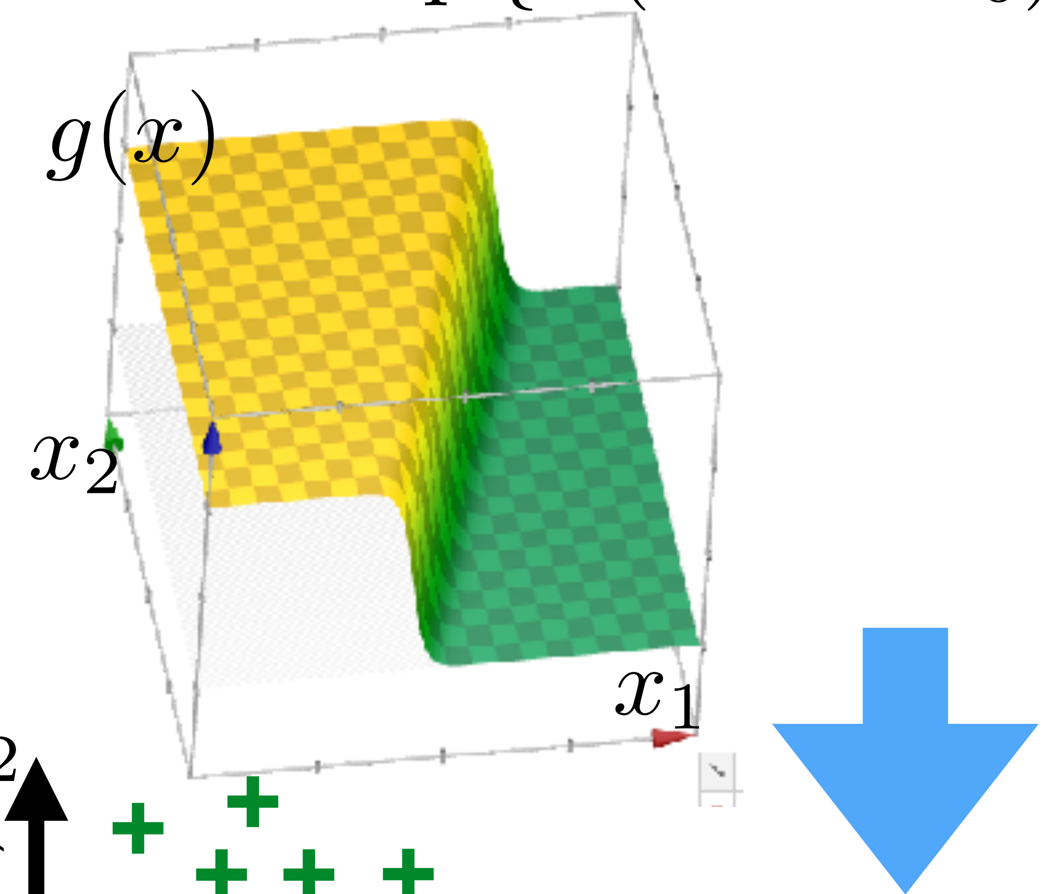
1 feature:

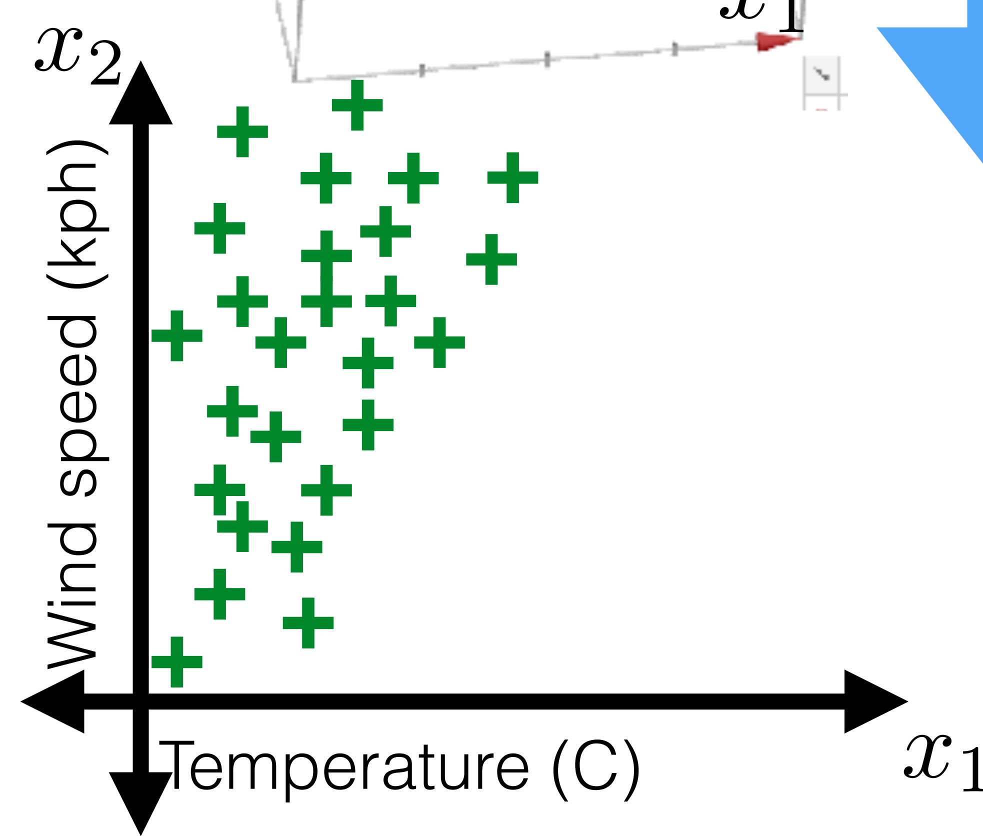
$$g(x) = \sigma(\theta x + \theta_0)$$
$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$



2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0)$$
$$= \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$

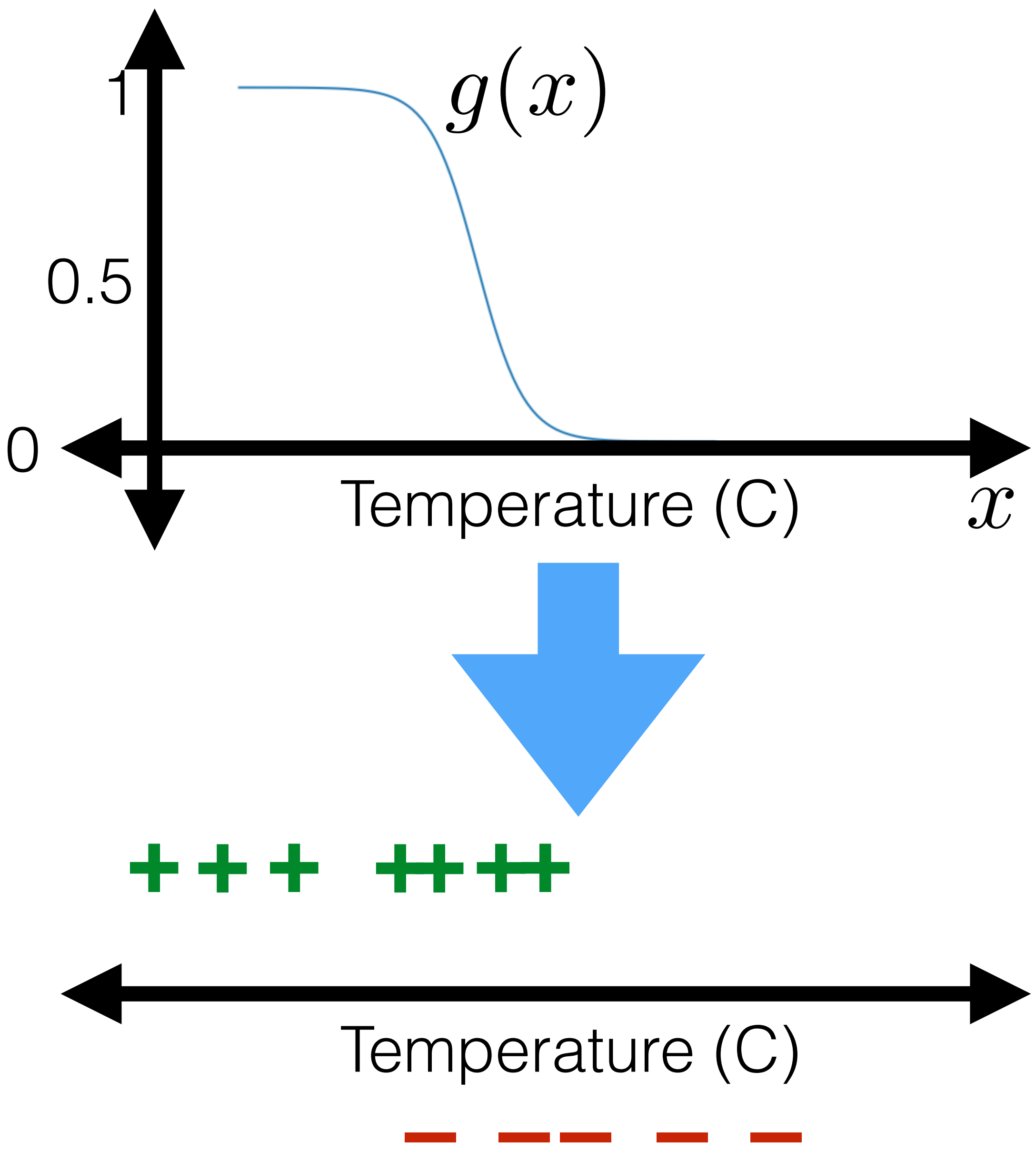




Capturing uncertainty

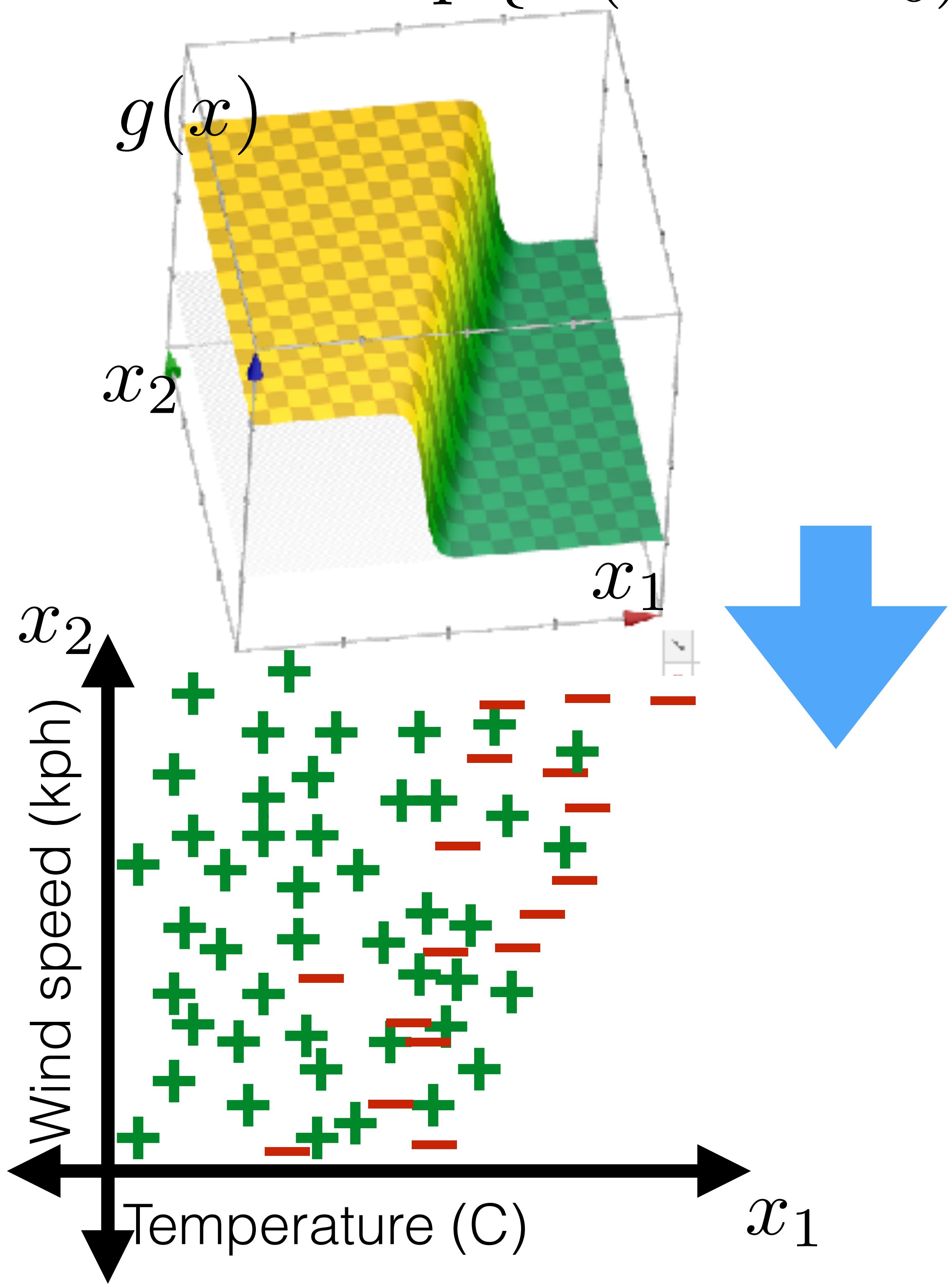
1 feature:

$$g(x) = \sigma(\theta x + \theta_0) = \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$



2 features:

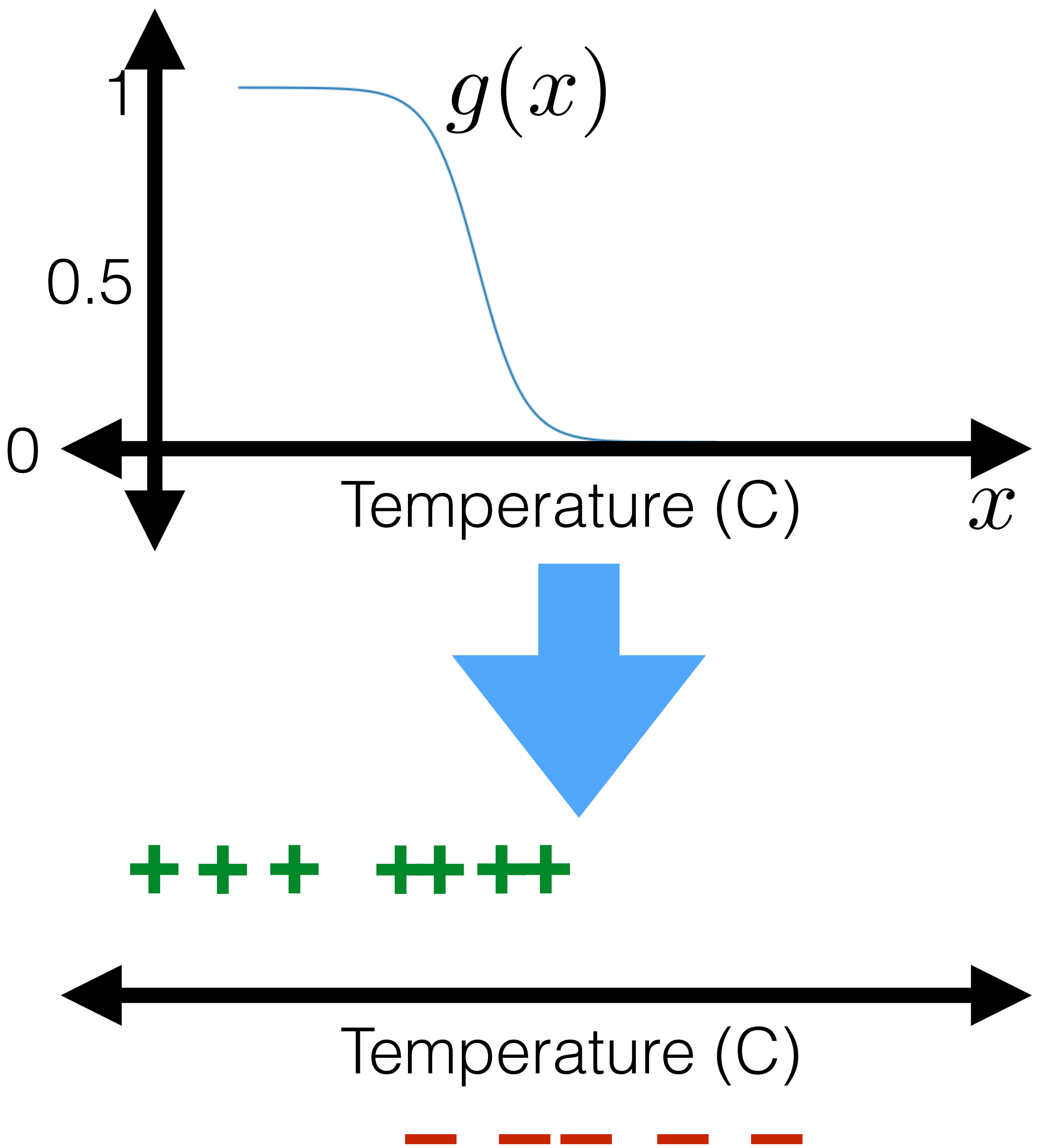
$$g(x) = \sigma(\theta^\top x + \theta_0) = \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$



Capturing uncertainty

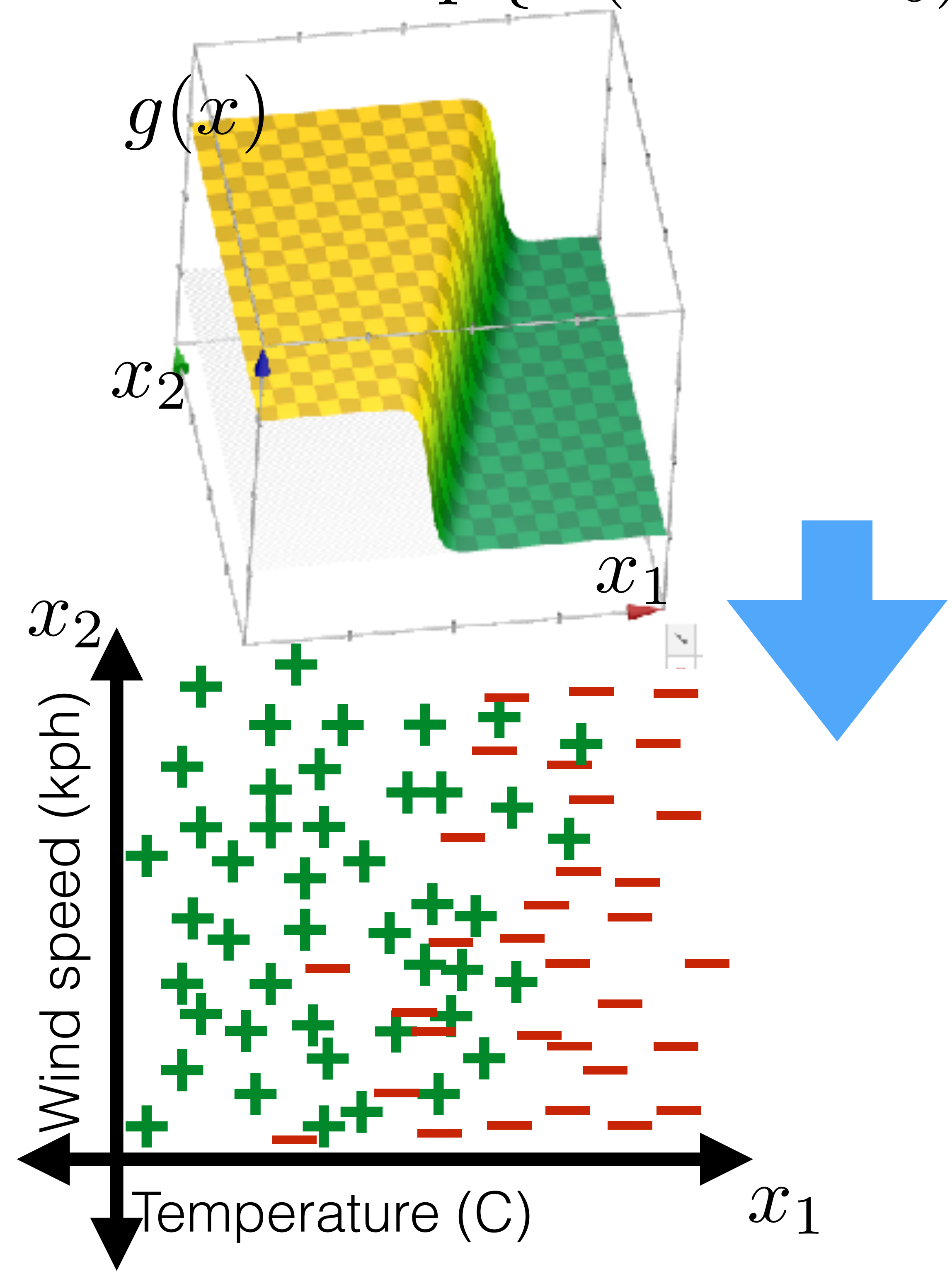
1 feature:

$$g(x) = \sigma(\theta x + \theta_0)$$
$$= \frac{1}{1 + \exp\{-(\theta x + \theta_0)\}}$$



2 features:

$$g(x) = \sigma(\theta^\top x + \theta_0)$$
$$= \frac{1}{1 + \exp\{-(\theta^\top x + \theta_0)\}}$$



Linear logistic classification

aka logistic regression

Linear logistic classification

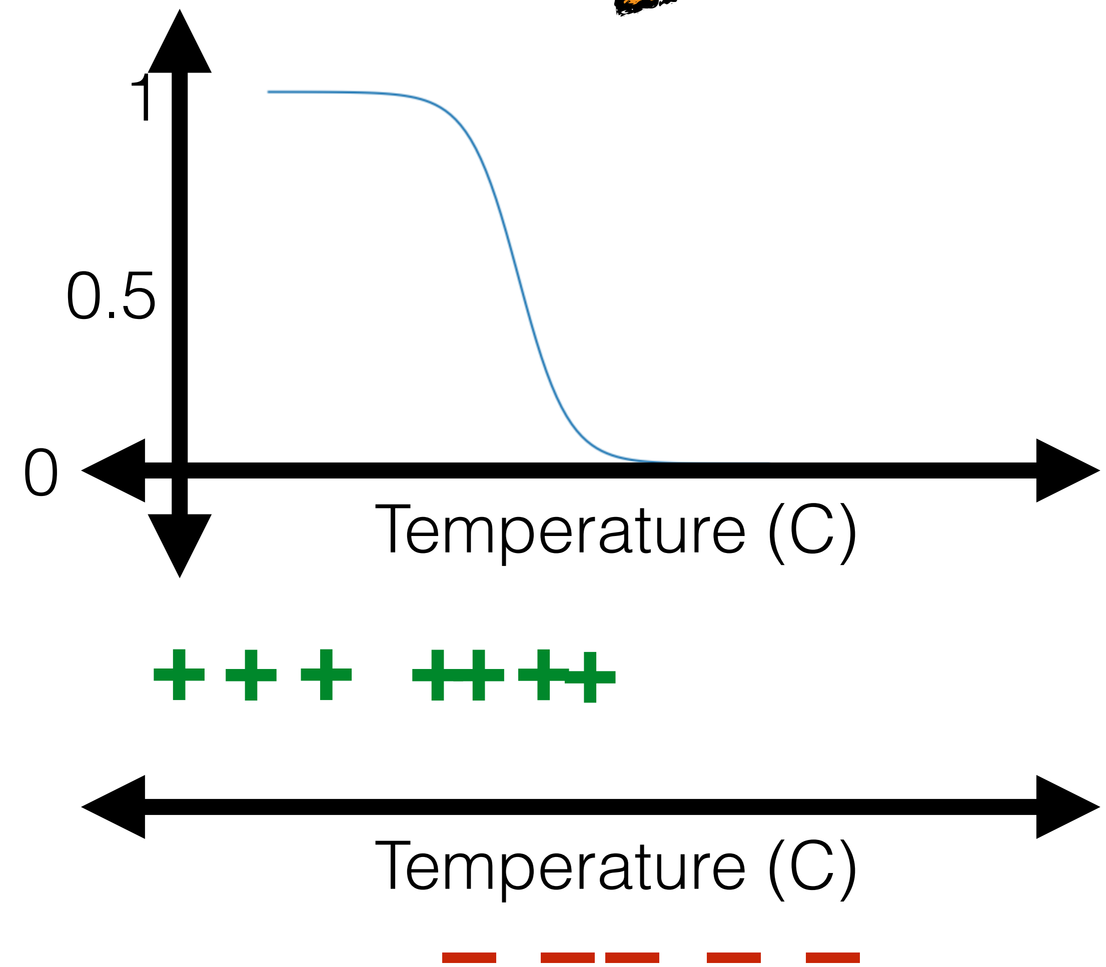
aka logistic regression

- What's an appropriate loss for this guess?

Linear logistic classification

aka logistic regression

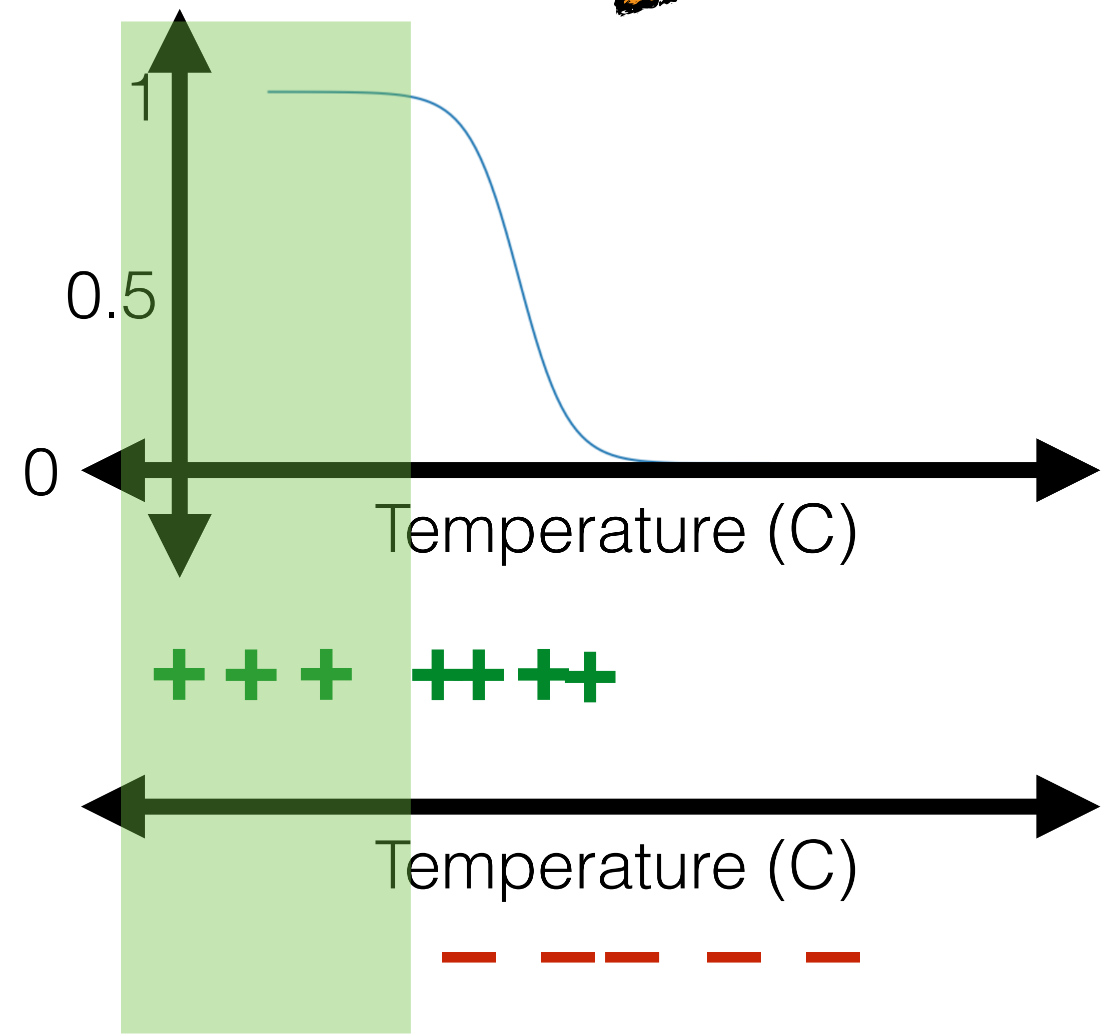
- What's an appropriate loss for this guess?



Linear logistic classification

aka logistic regression

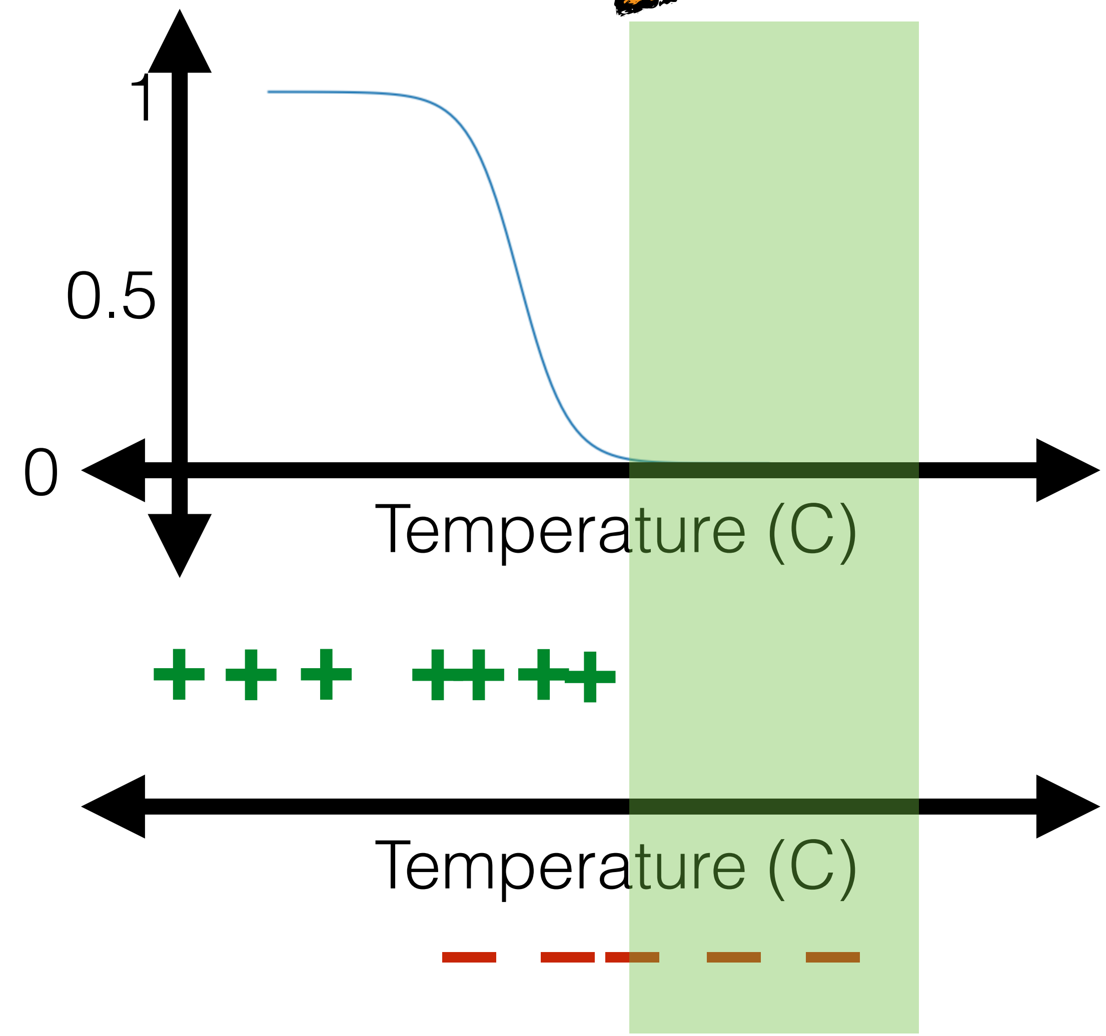
- What's an appropriate loss for this guess?



Linear logistic classification

aka logistic regression

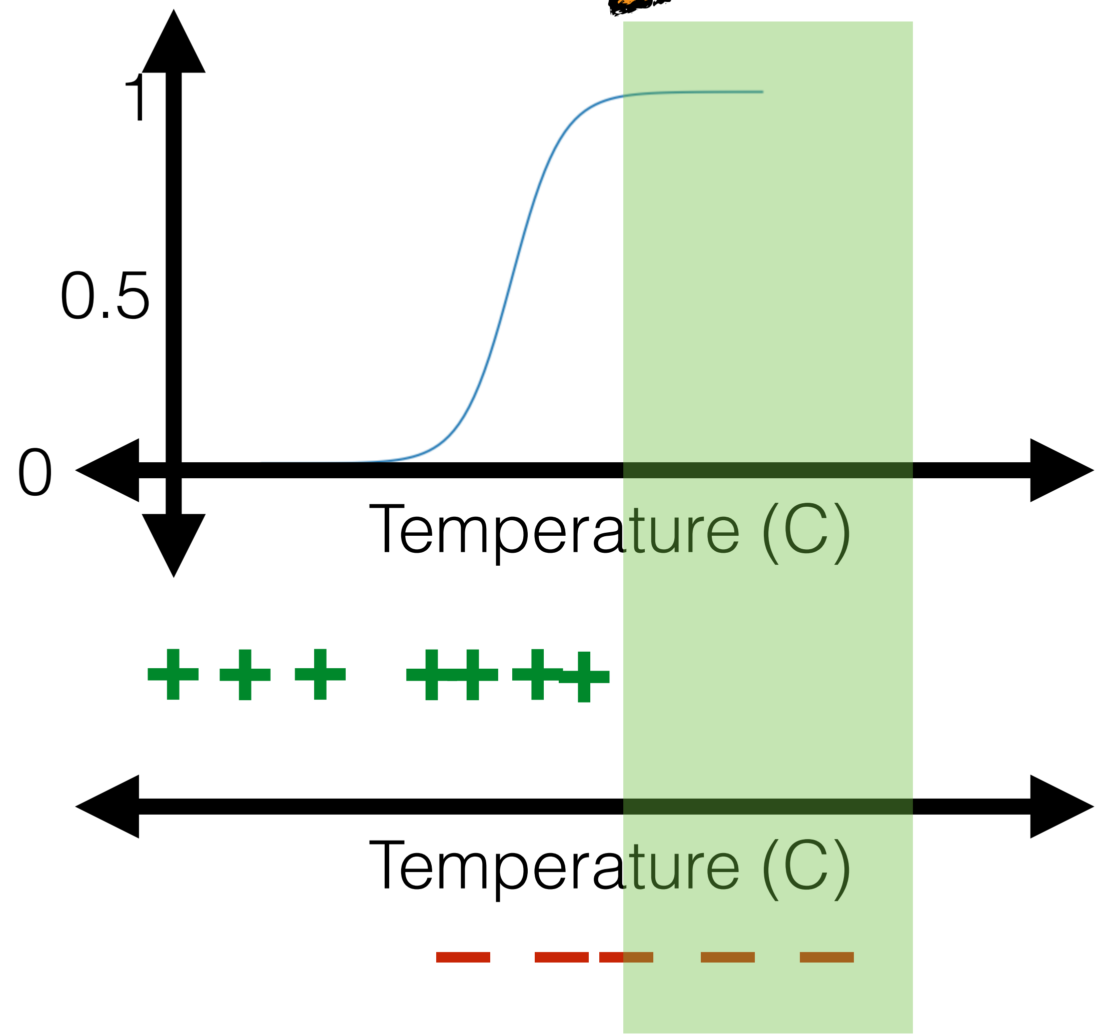
- What's an appropriate loss for this guess?



Linear logistic classification

aka logistic regression

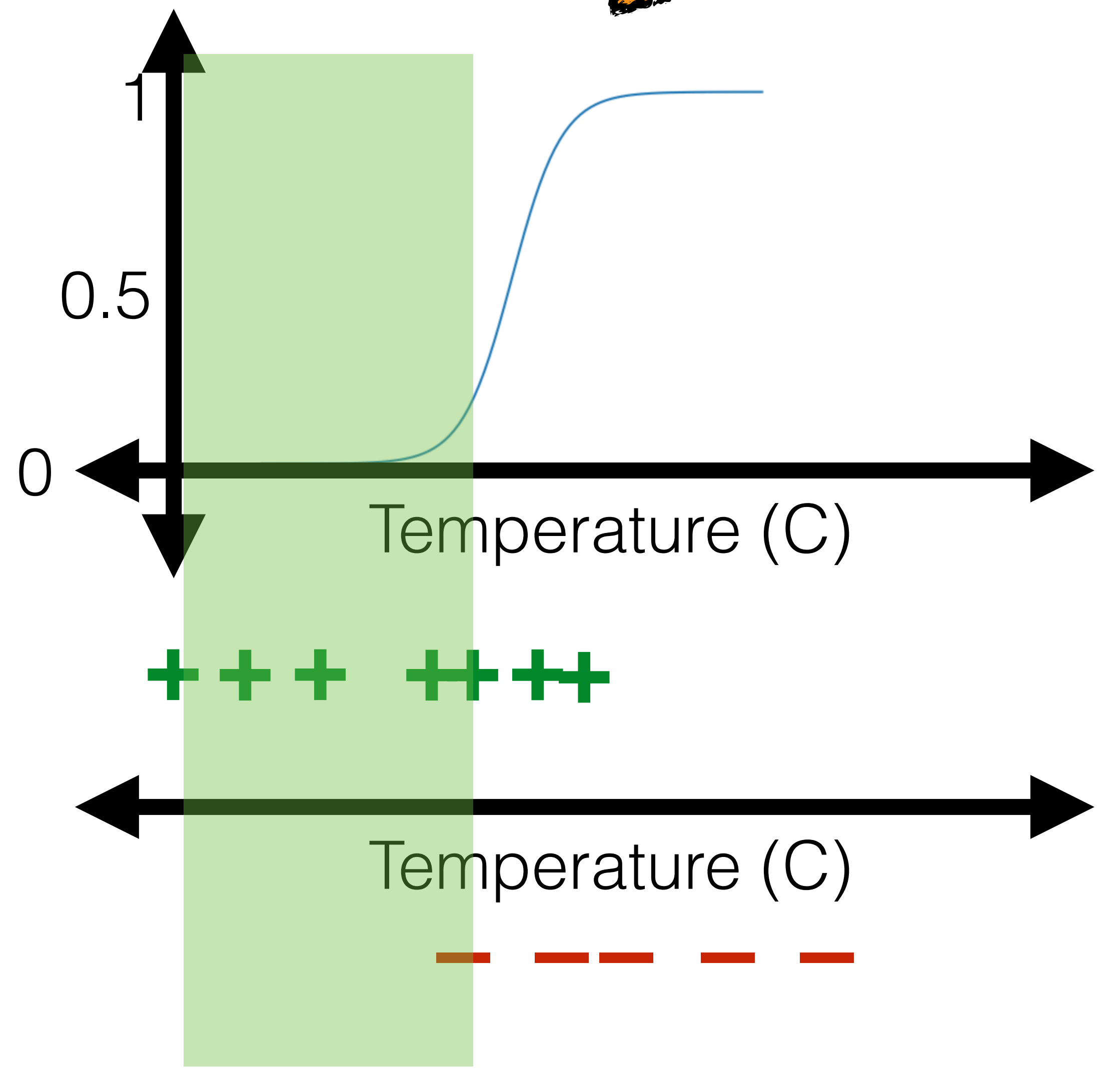
- What's an appropriate loss for this guess?



Linear logistic classification

aka logistic regression

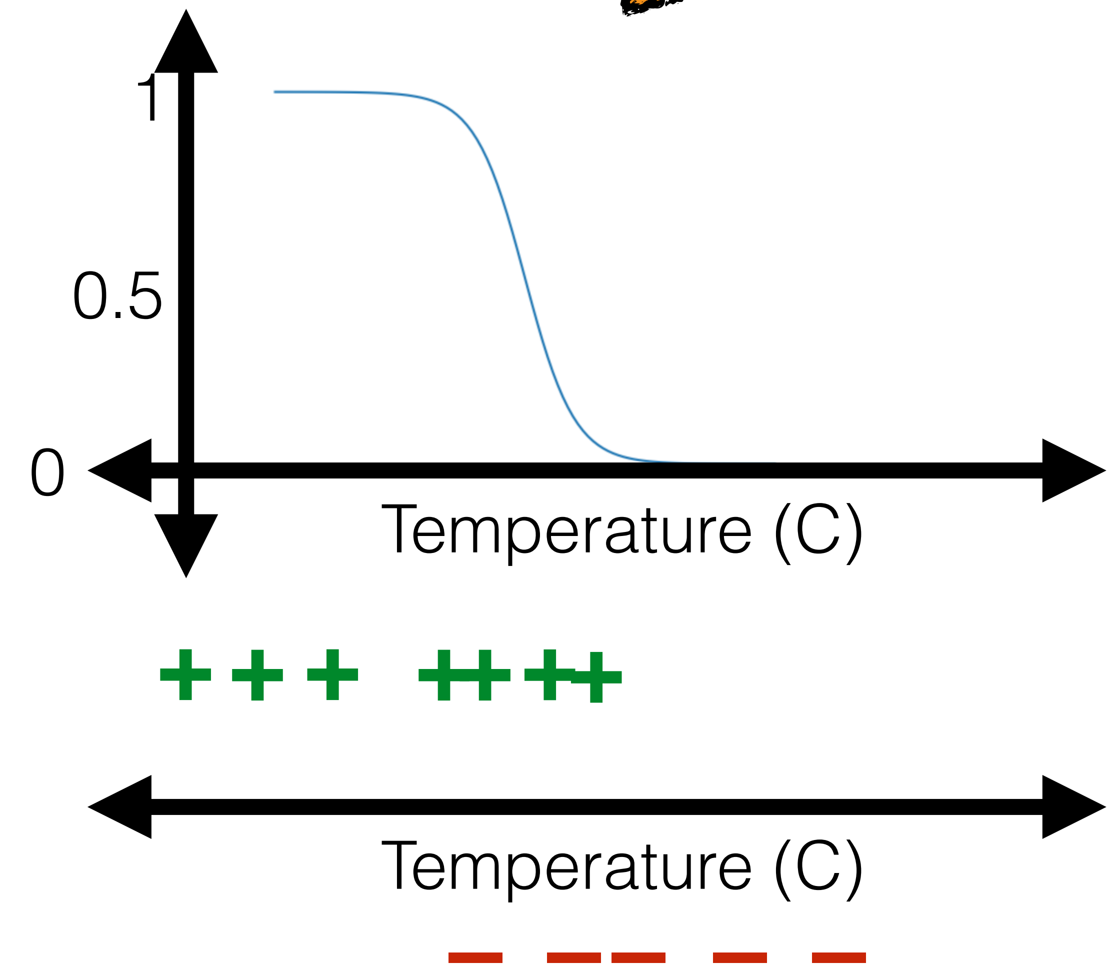
- What's an appropriate loss for this guess?



Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

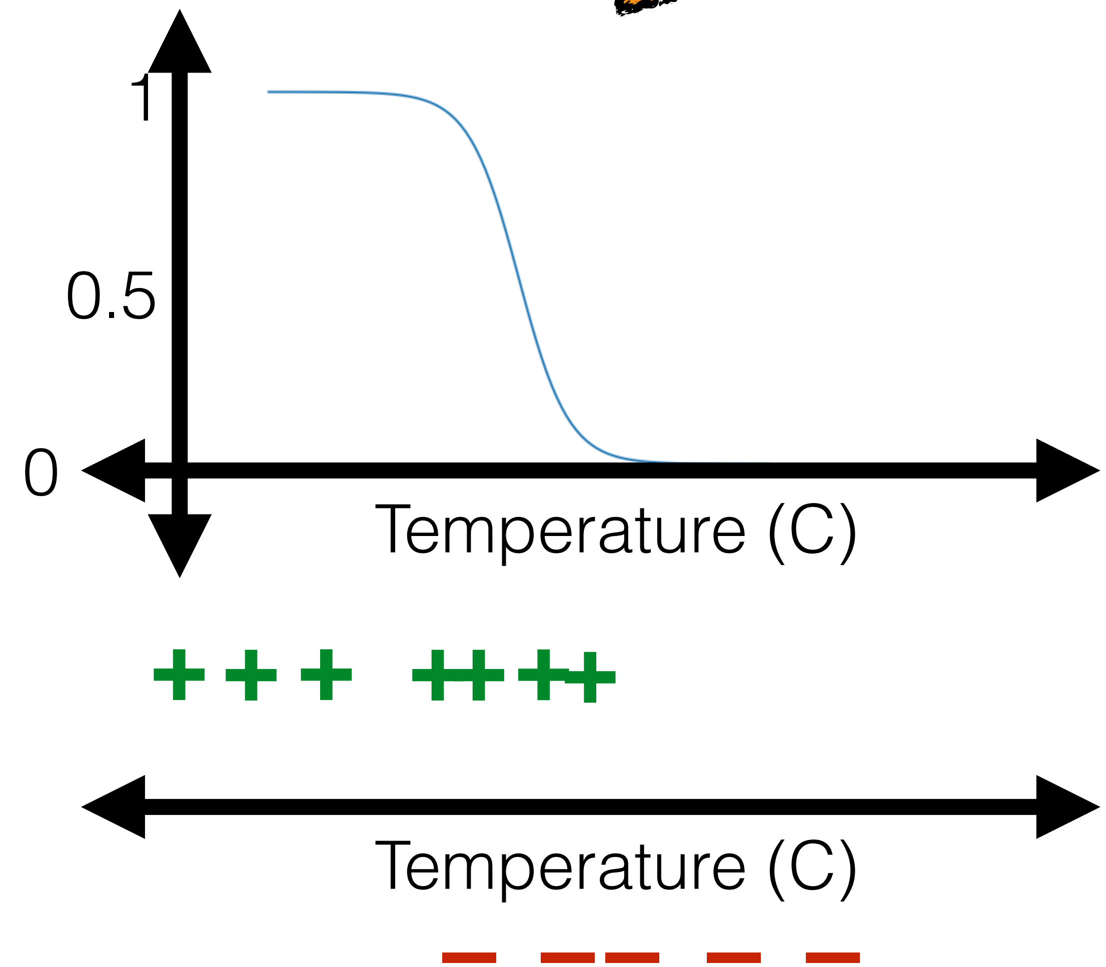


Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

Probability(data)



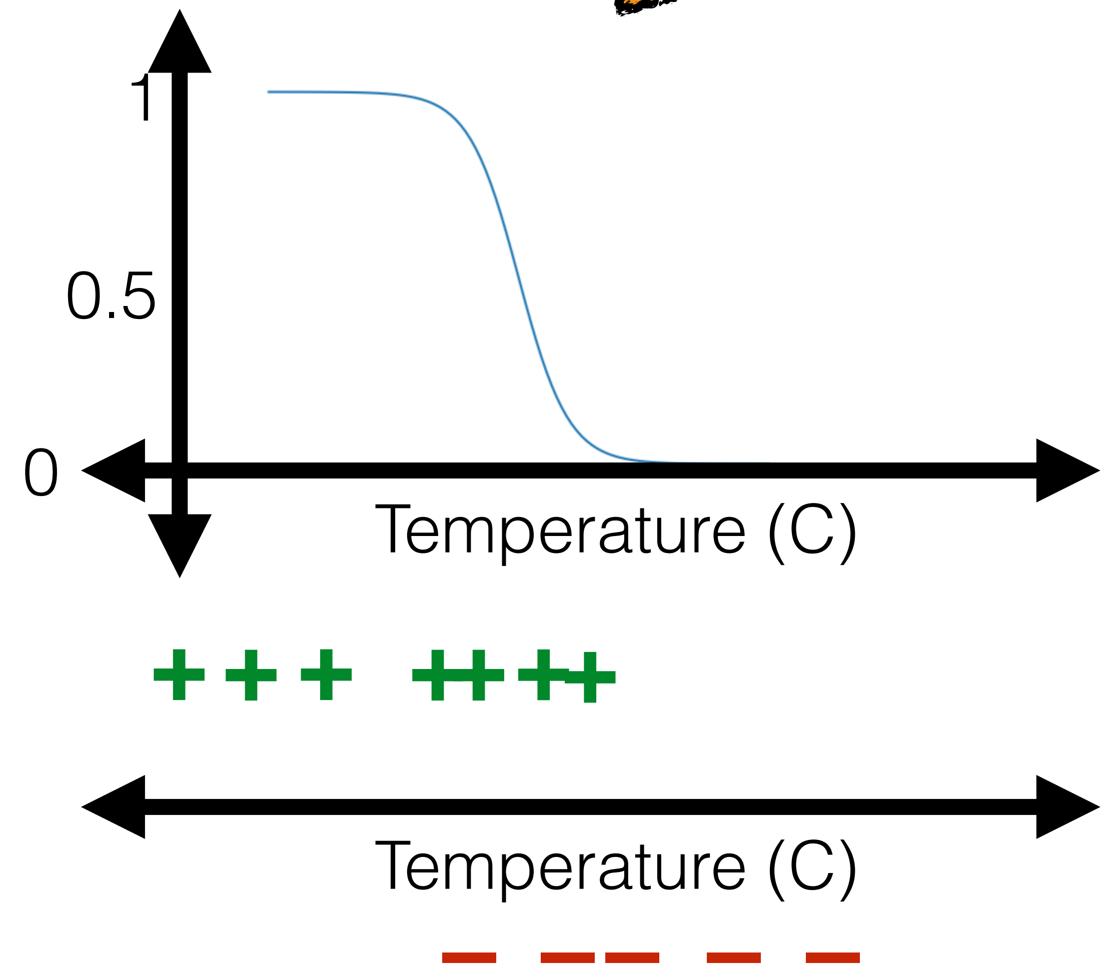
Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

Probability(data)

$$= \prod_{i=1}^{n} \text{Probability}(\text{data point } i)$$



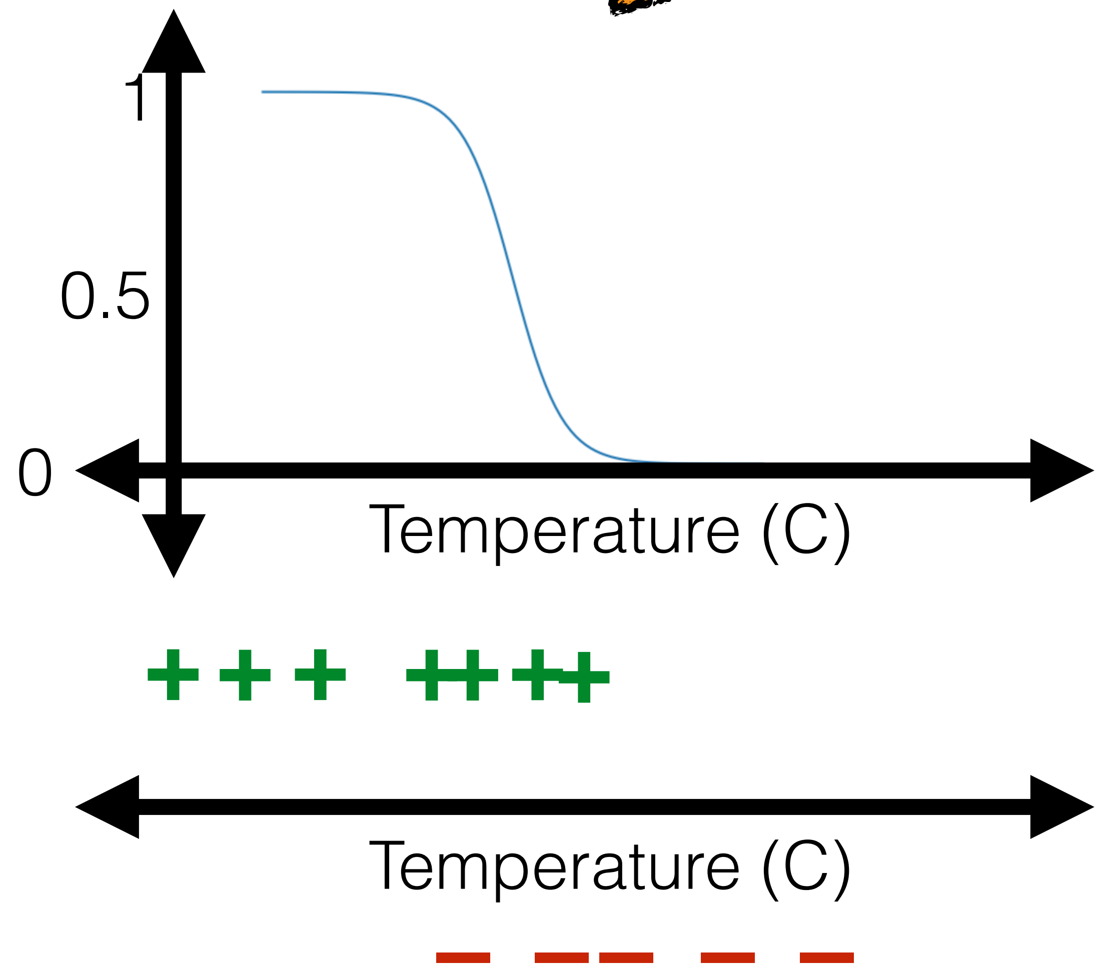
Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

Probability(data)

$$= \prod_{i=1}^{n} \text{Probability}(\text{data point } i)$$



Linear logistic classification

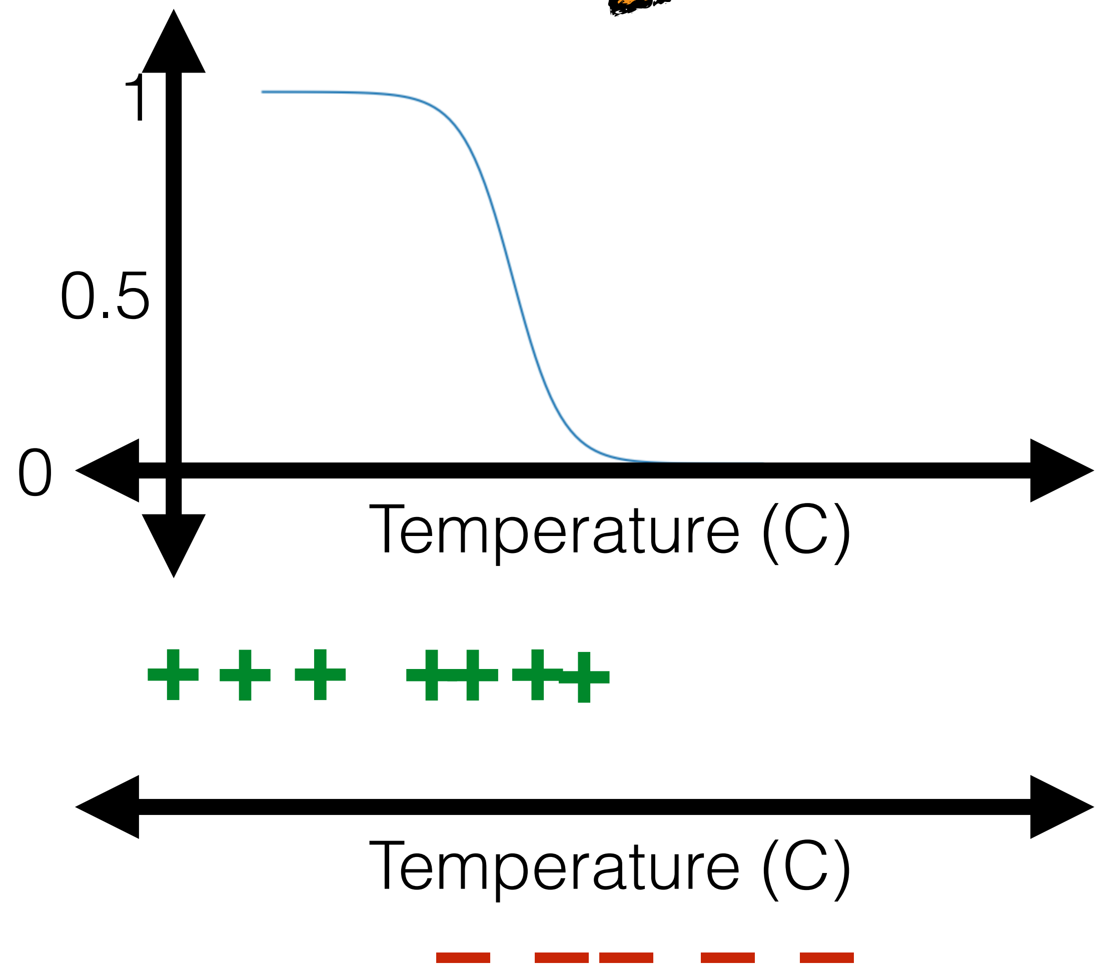
aka logistic regression

- What's an appropriate loss for this guess?

Probability(data)

$$= \prod_{i=1}^{n} \text{Probability}(\text{data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]



Linear logistic classification

- What's an appropriate loss for this guess?

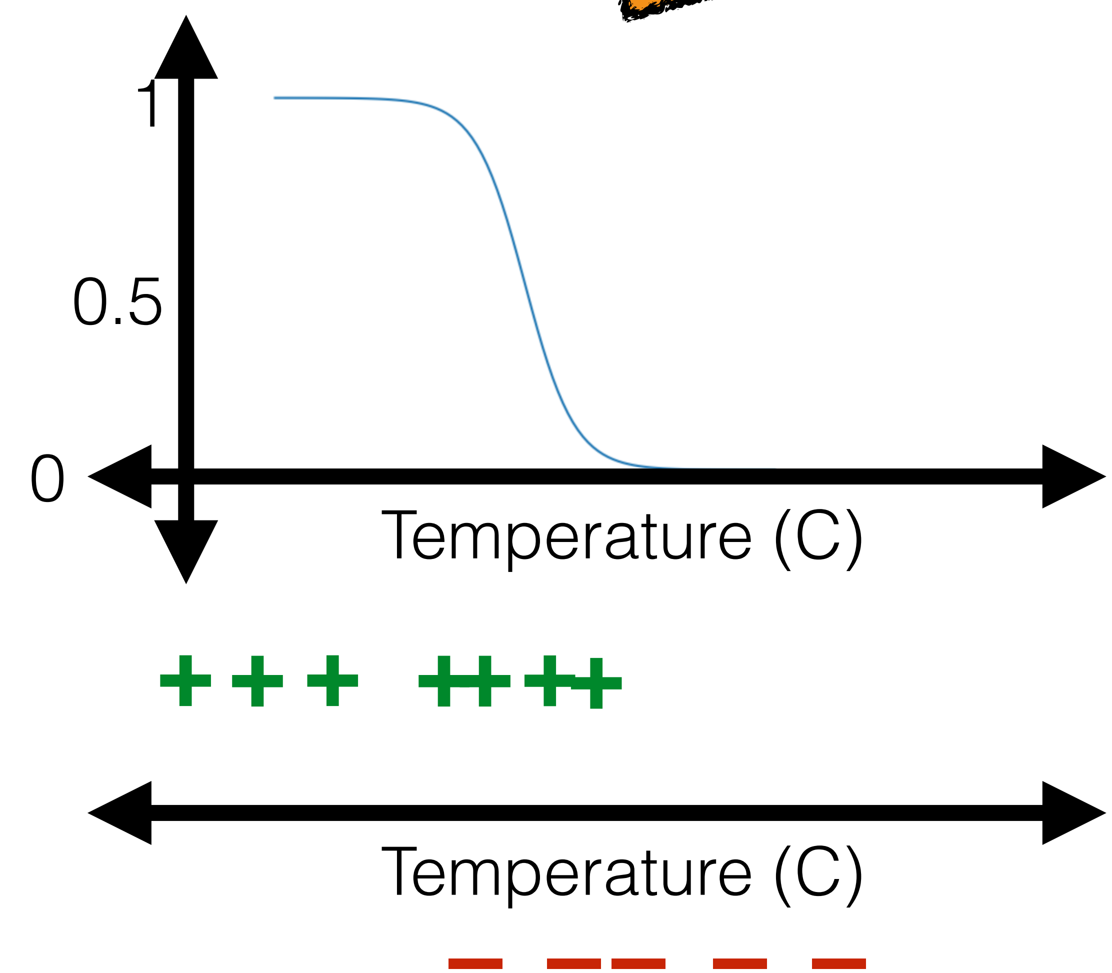
aka logistic regression

Probability(data)

$$= \prod_{i=1}^{n} \text{Probability(data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$



Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

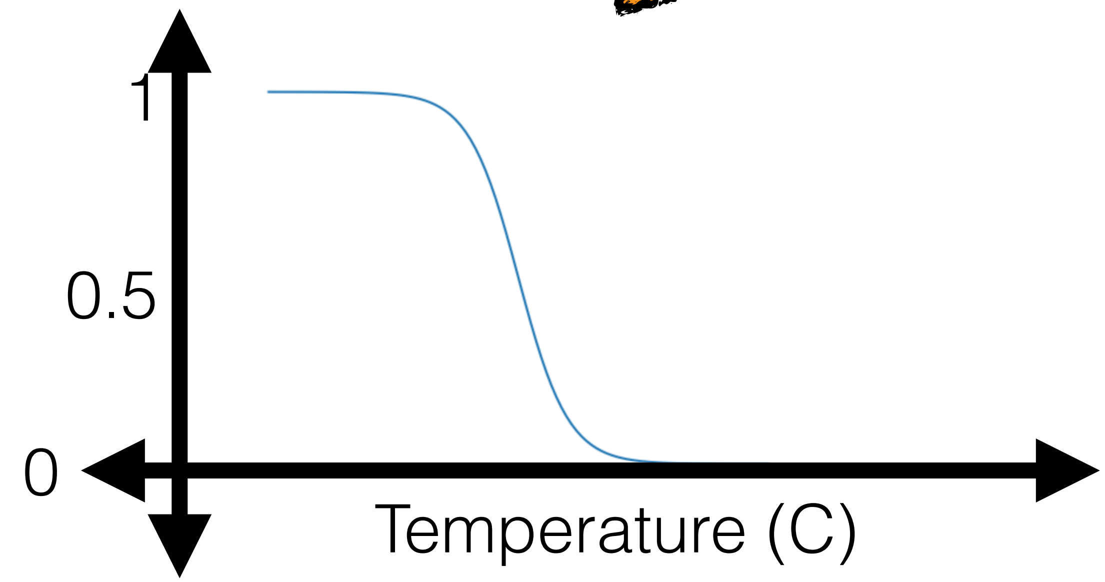
Probability(data)

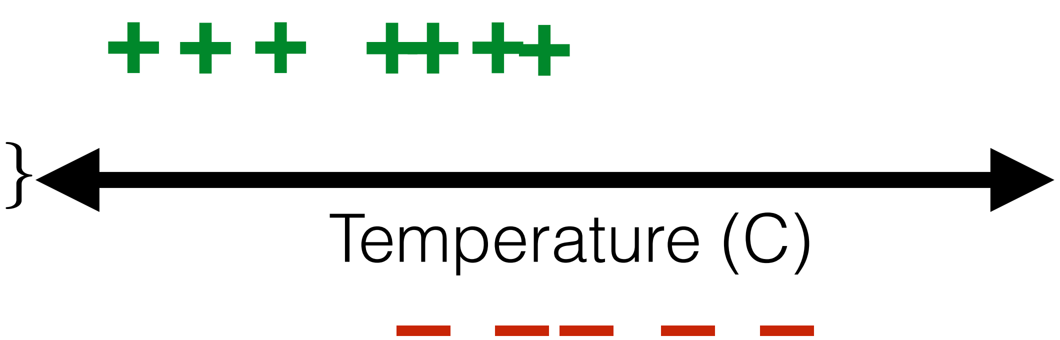
$$= \prod_{i=1}^{n} \text{Probability(data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$

$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)}\neq+1\}}$$





Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

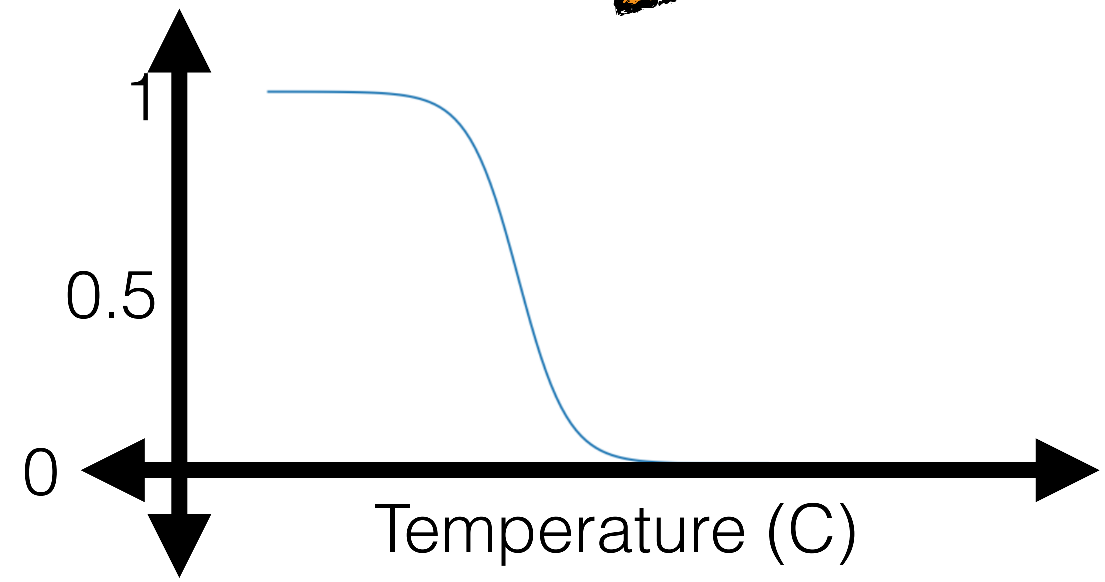
Probability(data)

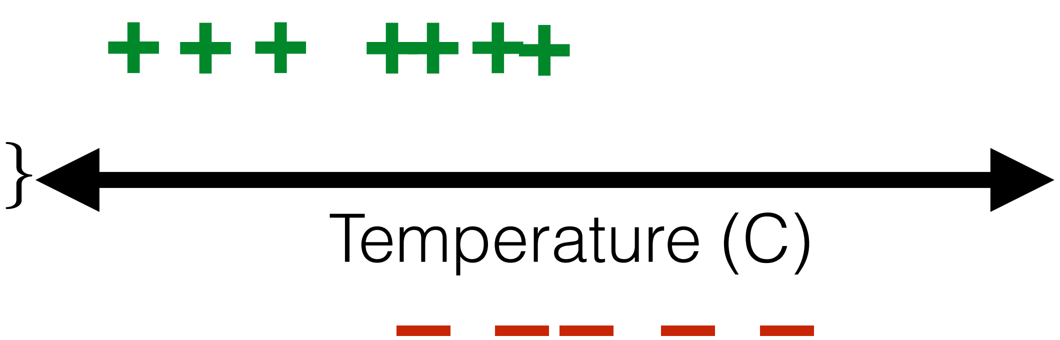
$$= \prod_{i=1}^{n} \text{Probability(data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$

$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)} \neq +1\}}$$





Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

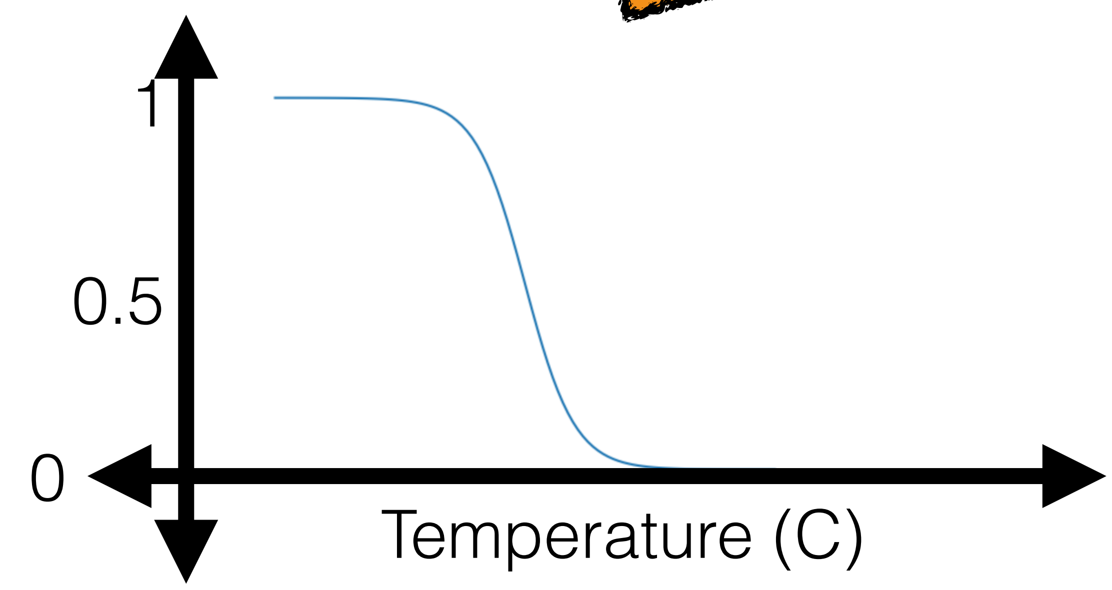
Probability(data)

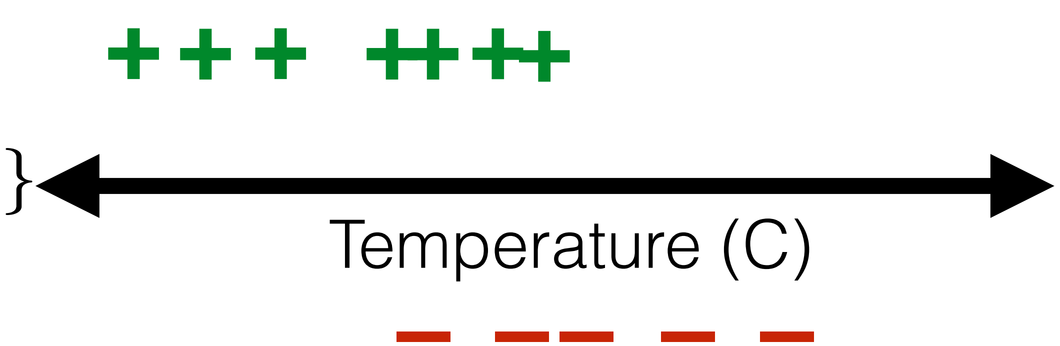
$$= \prod_{i=1}^{n} \text{Probability(data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$

$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)}\neq+1\}}$$





Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

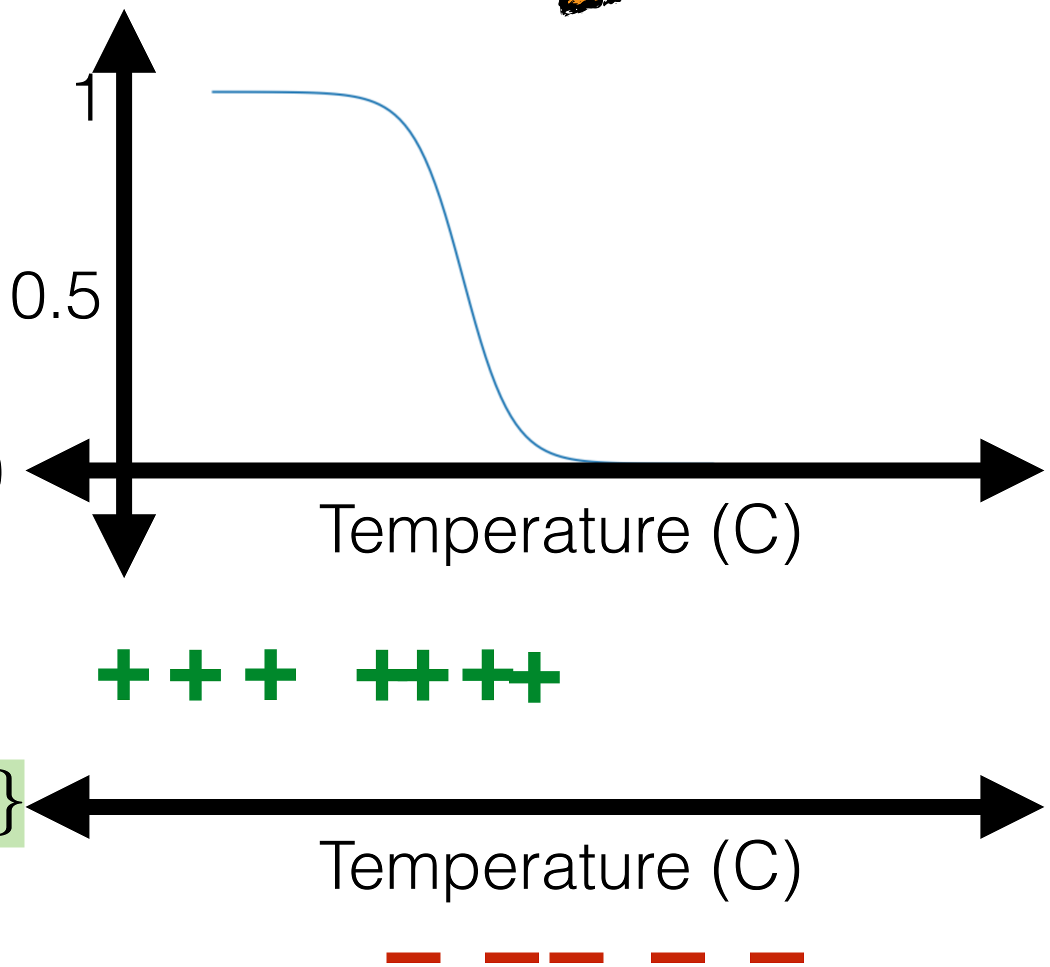
Probability(data)

$$= \prod_{i=1}^{n} \text{Probability(data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$

$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)}\neq+1\}}$$



Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

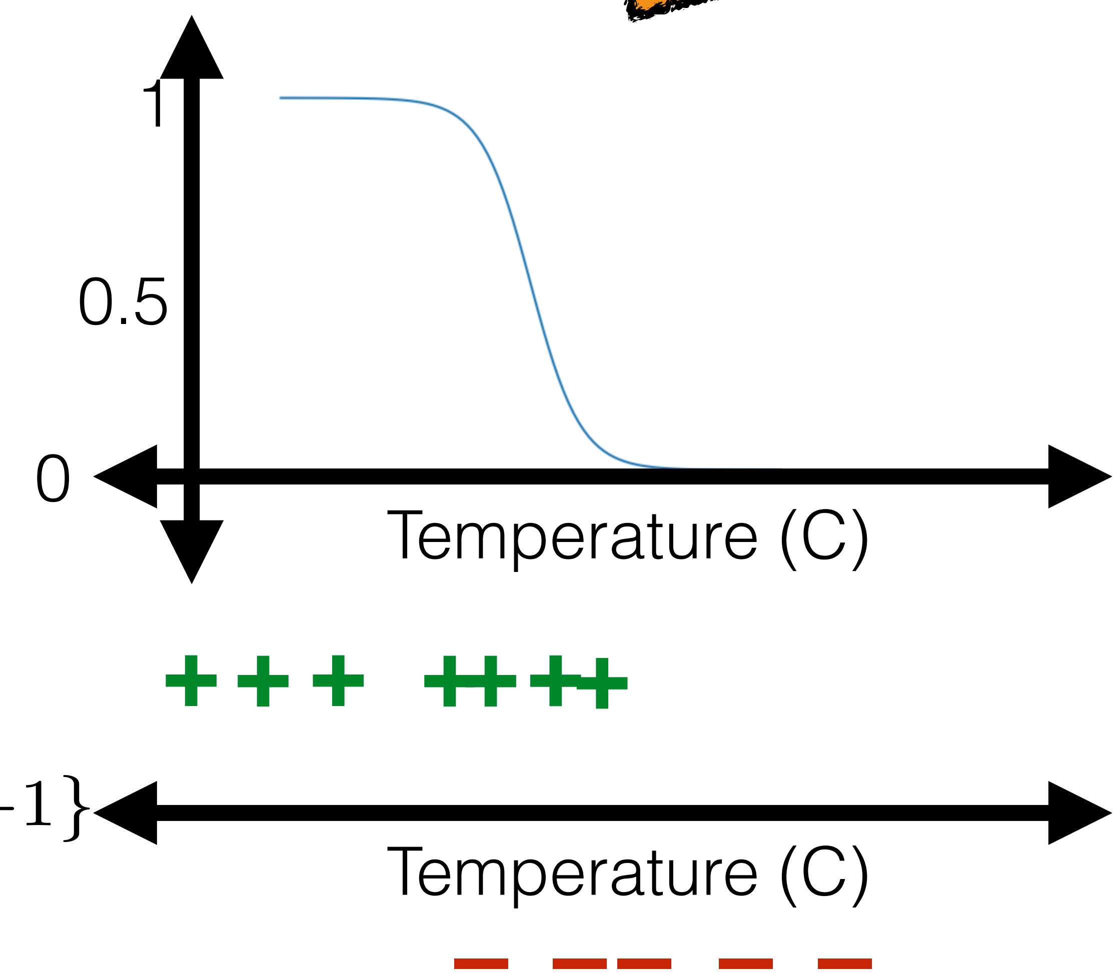
Probability(data)

$$= \prod_{i=1}^{n} \text{Probability(data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$

$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)}\neq+1\}}$$



Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

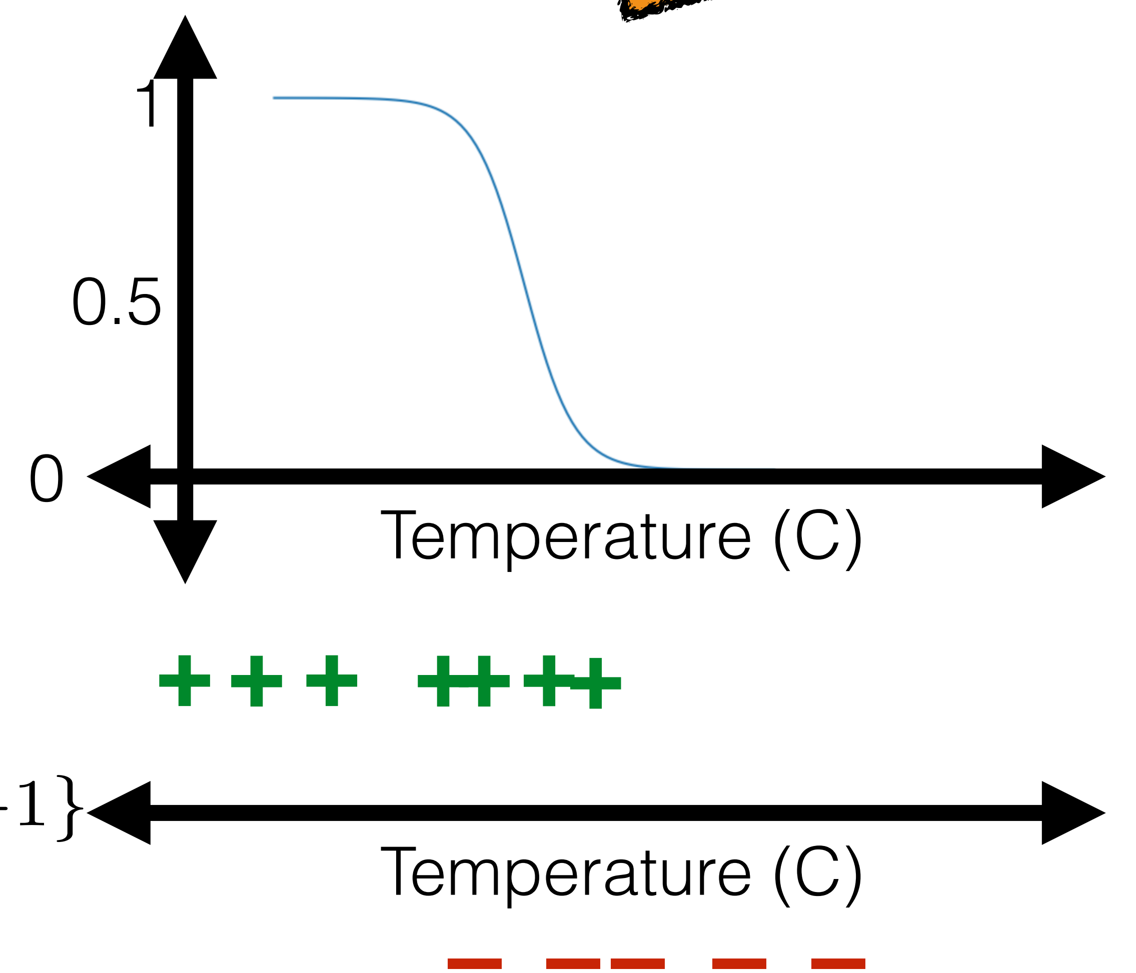
Probability(data)

$$= \prod_{i=1}^{n} \text{Probability(data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$

$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)}\neq+1\}}$$



log probability(data)

Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

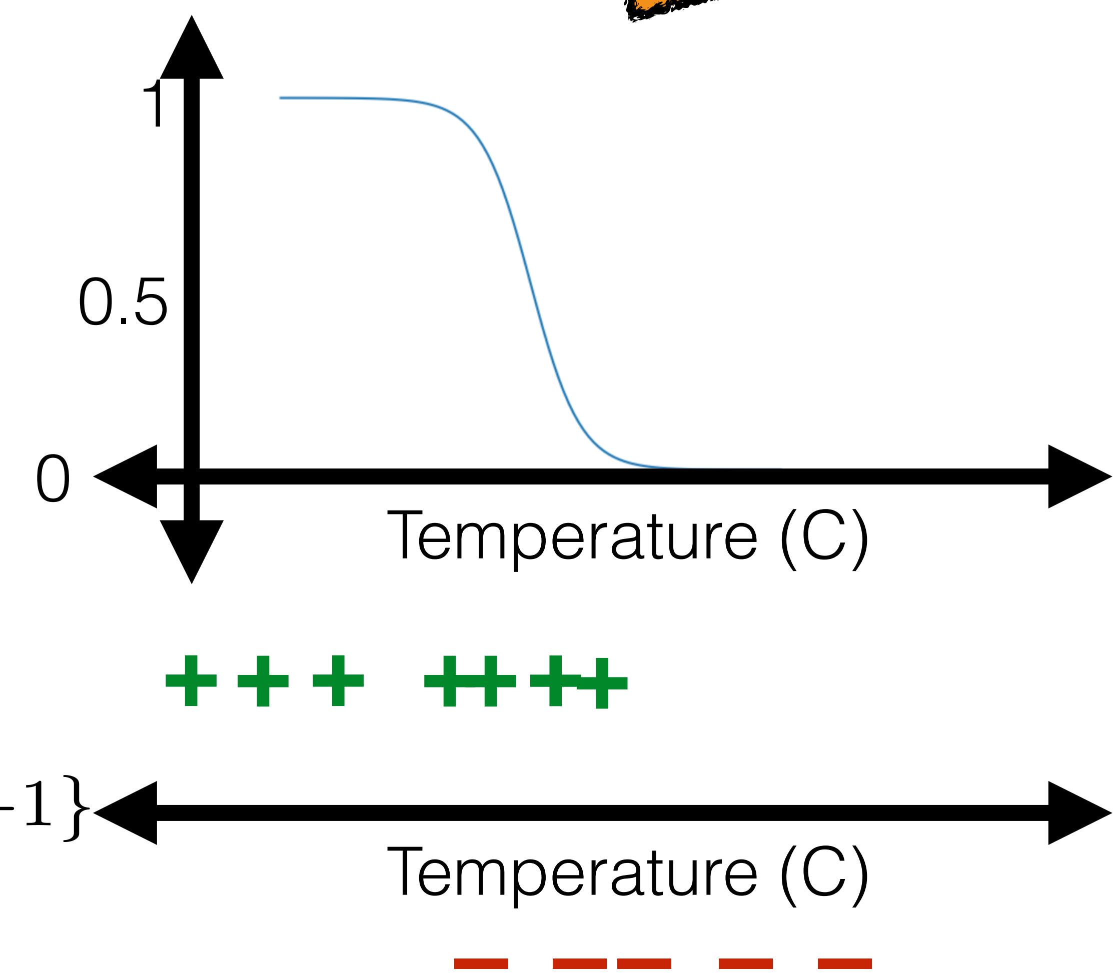
Probability(data)

$$= \prod_{i=1}^{n} \text{Probability}(\text{data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$

$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)} \neq +1\}}$$



Loss(data) = - log probability(data)

Linear logistic classification

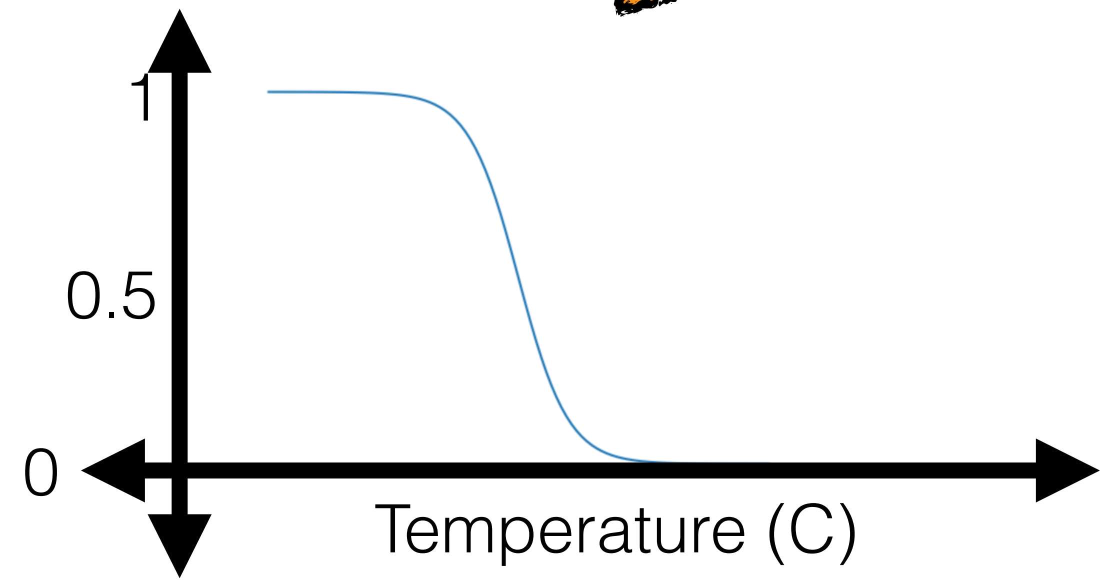
aka logistic regression

- What's an appropriate loss for this guess?

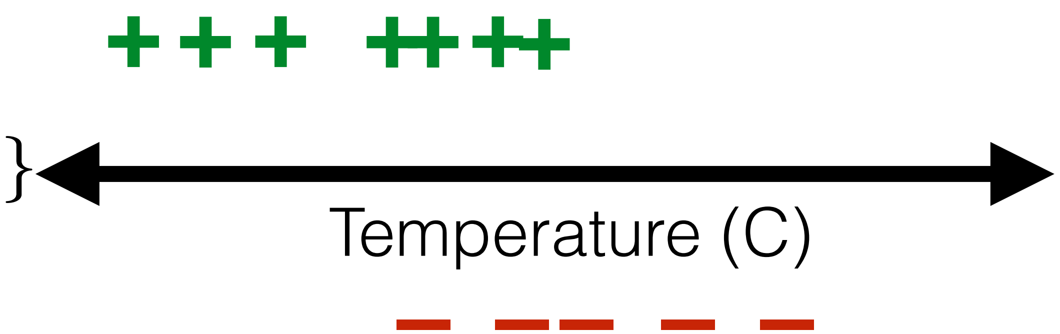
Probability(data)

$$= \prod_{i=1}^{n} \text{Probability(data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]



$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} & \text{if } y^{(i)} = +1 \\ (1 - g^{(i)}) & \text{else} \end{cases}$$



$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)}\neq+1\}}$$

Loss(data) = - log probability(data)

$$= \sum_{i=1}^{n} -\left(\mathbf{1}\{y^{(i)} = +1\} \log g^{(i)} + \mathbf{1}\{y^{(i)} \neq +1\} \log(1 - g^{(i)})\right)$$

Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

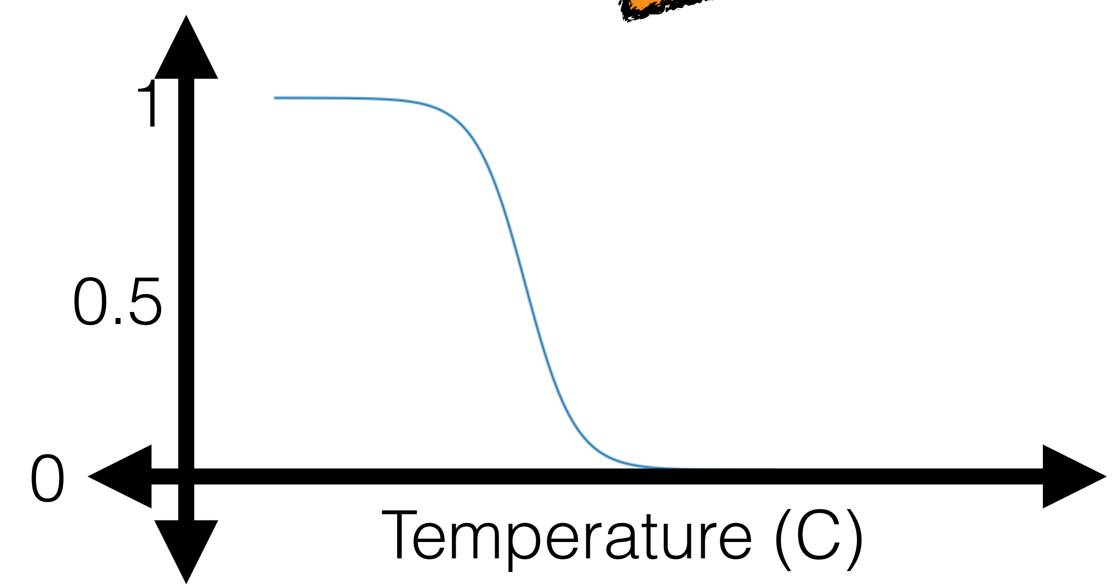
Probability(data)

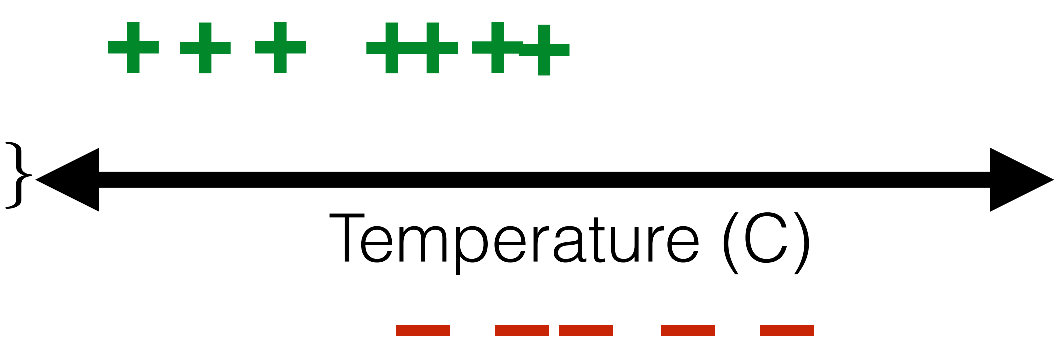
$$= \prod_{i=1}^{n} \text{Probability(data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$

$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)}\neq+1\}}$$





Loss(data) = - log probability(data)

$$= \sum_{i=1}^{n} -\left(\mathbf{1}\{y^{(i)} = +1\} \log g^{(i)} + \mathbf{1}\{y^{(i)} \neq +1\} \log(1 - g^{(i)})\right.$$

Linear logistic classification

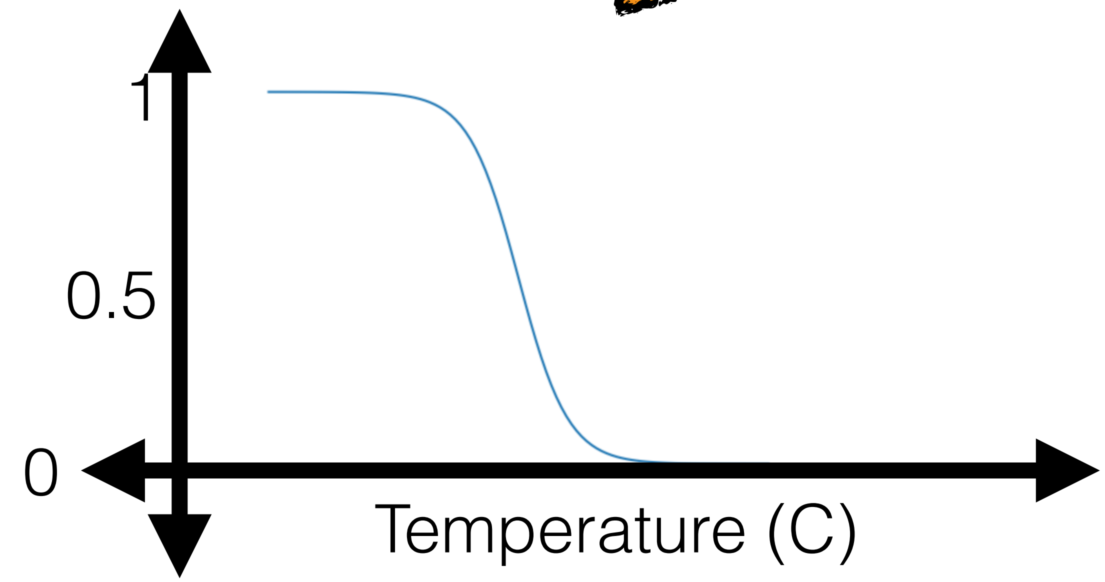
aka logistic regression

- What's an appropriate loss for this guess?

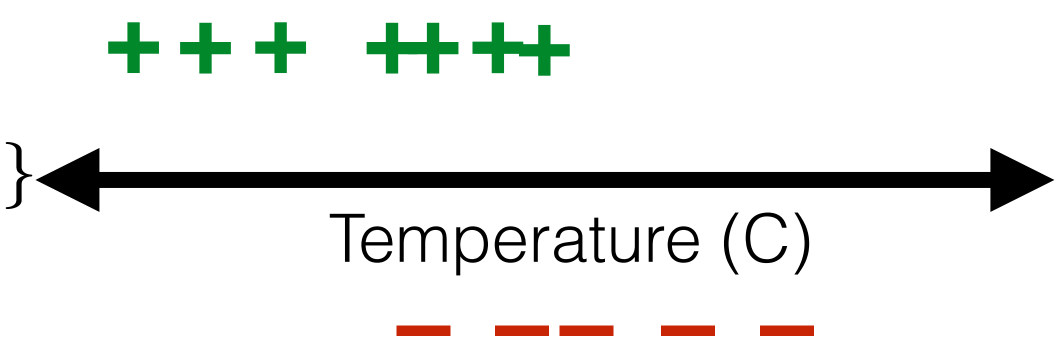
Probability(data)

$$= \prod_{i=1}^{n} \text{Probability}(\text{data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]



$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$



$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)} \neq +1\}}$$

Loss(data) = - log probability(data)

$$= \sum_{i=1}^{n} -\left(\mathbf{1}\{y^{(i)} = +1\} \log g^{(i)} + \mathbf{1}\{y^{(i)} \neq +1\} \log(1 - g^{(i)}) \right.$$

Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

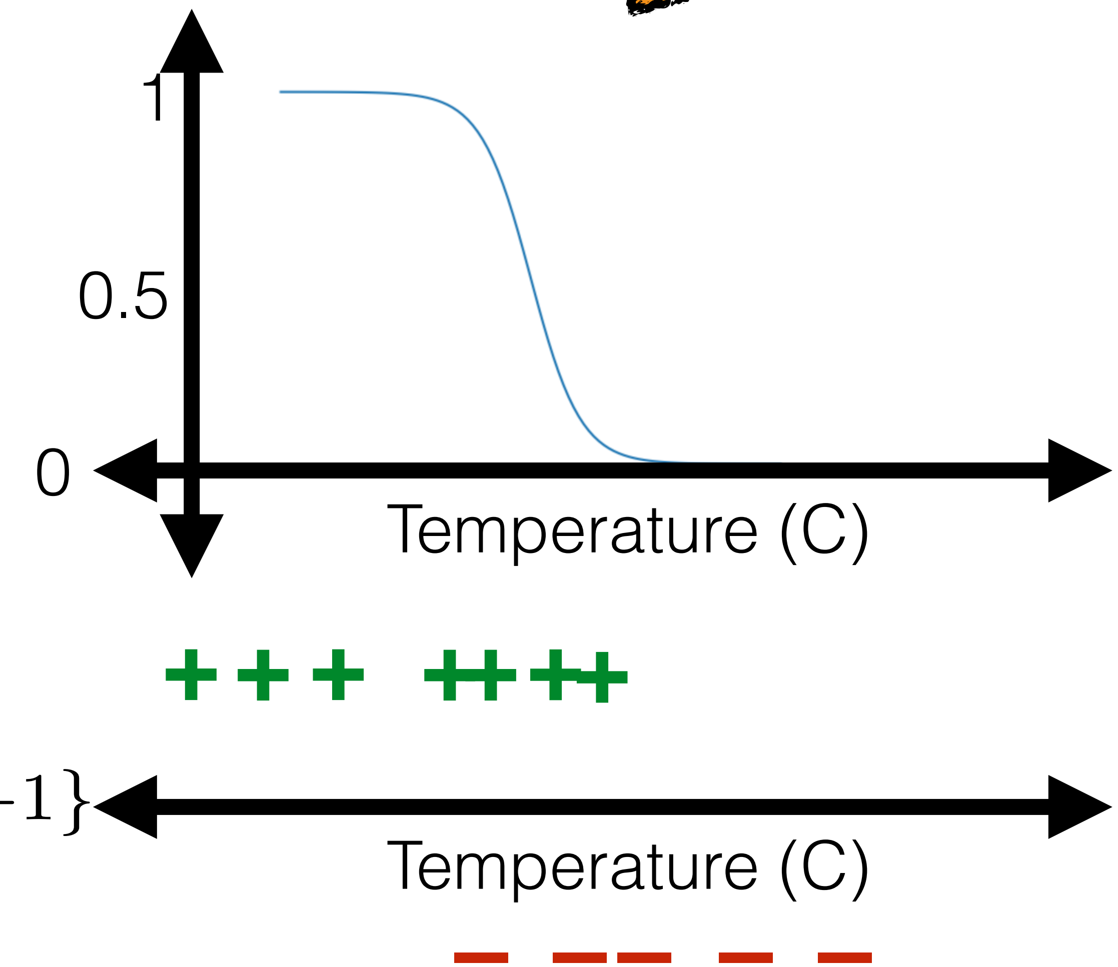
Probability(data)

$$= \prod_{i=1}^{n} \text{Probability(data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$

$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)}\neq+1\}}$$



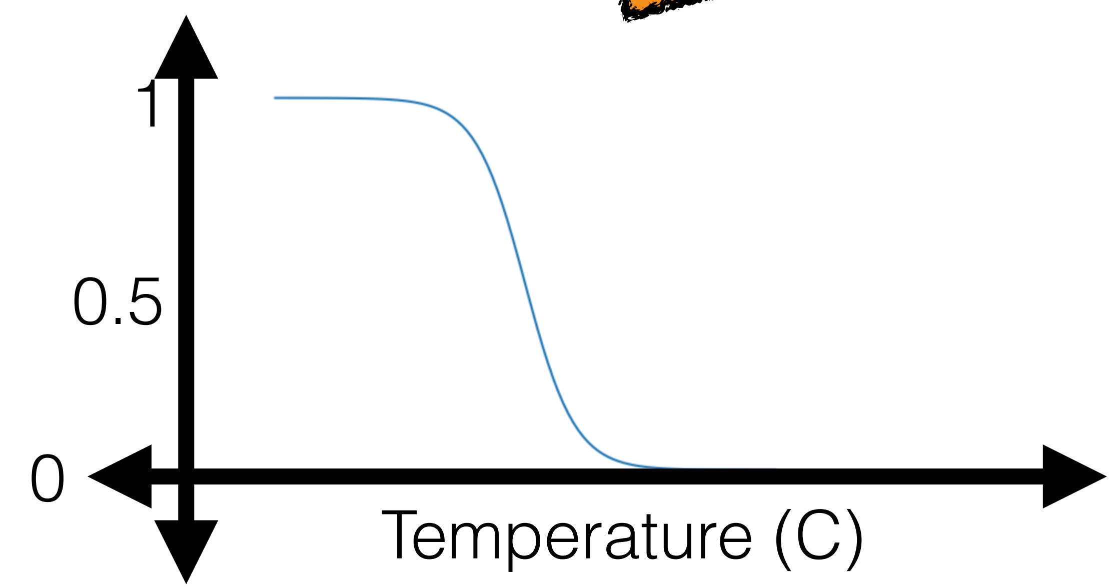
Loss(data) = - log probability(data)

$$= \sum_{i=1}^{n} - \left(\mathbf{1}\{y^{(i)} = +1\} \log g^{(i)} + \mathbf{1}\{y^{(i)} \neq +1\} \log(1 - g^{(i)}) \right)$$

Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

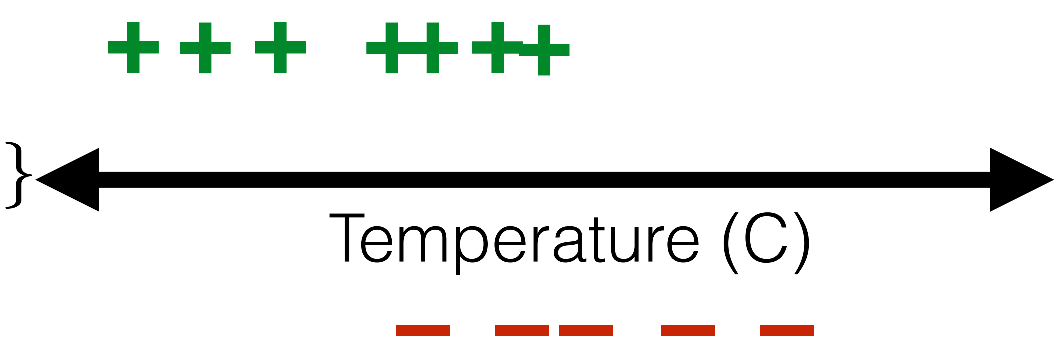


Probability(data)

$$= \prod_{i=1}^{n} \text{Probability(data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$



$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)}\neq+1\}}$$

Loss(data) = - log probability(data)

$$= \sum_{i=1}^{n} -\left(\mathbf{1}\{y^{(i)} = +1\} \log g^{(i)} + \mathbf{1}\{y^{(i)} \neq +1\} \log(1 - g^{(i)})\right.$$

Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

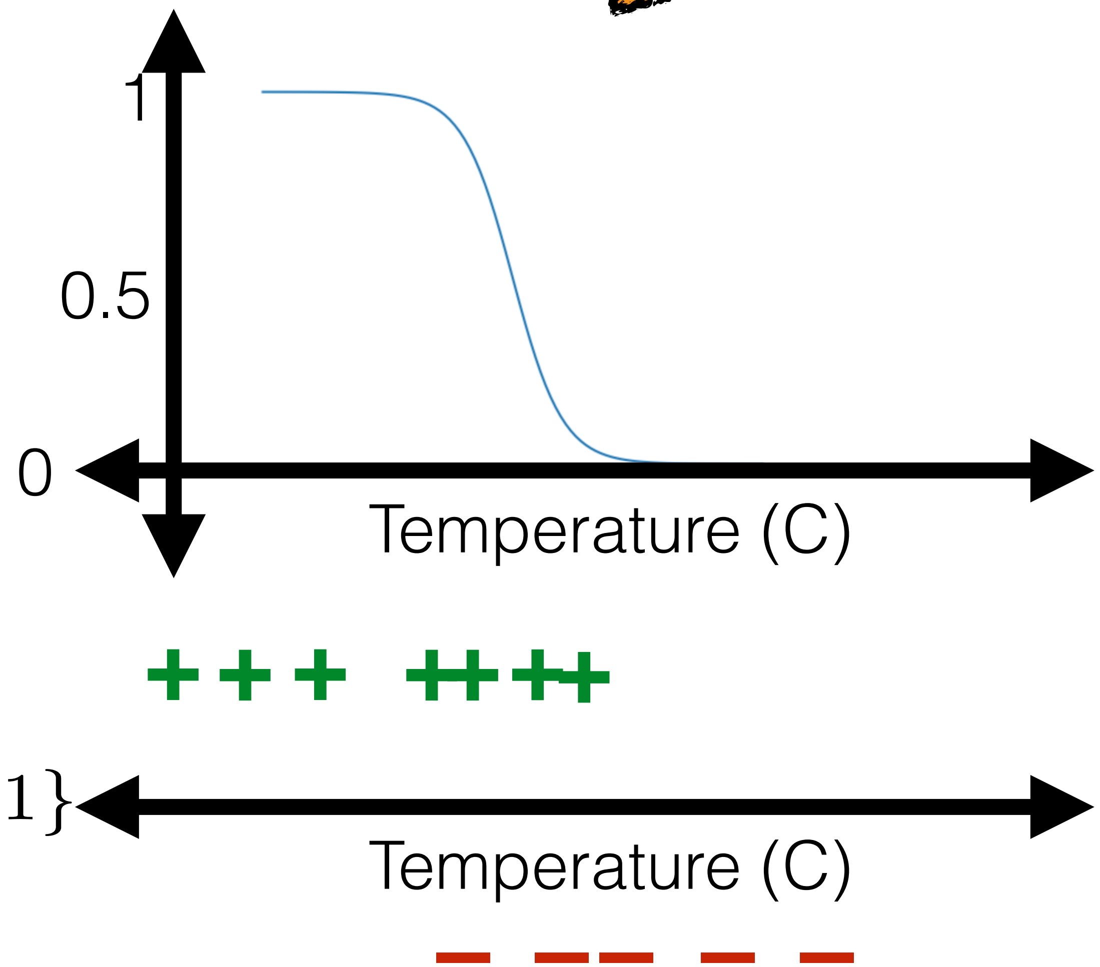
Probability(data)

$$= \prod_{i=1}^{n} \text{Probability(data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$

$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)} \neq +1\}}$$



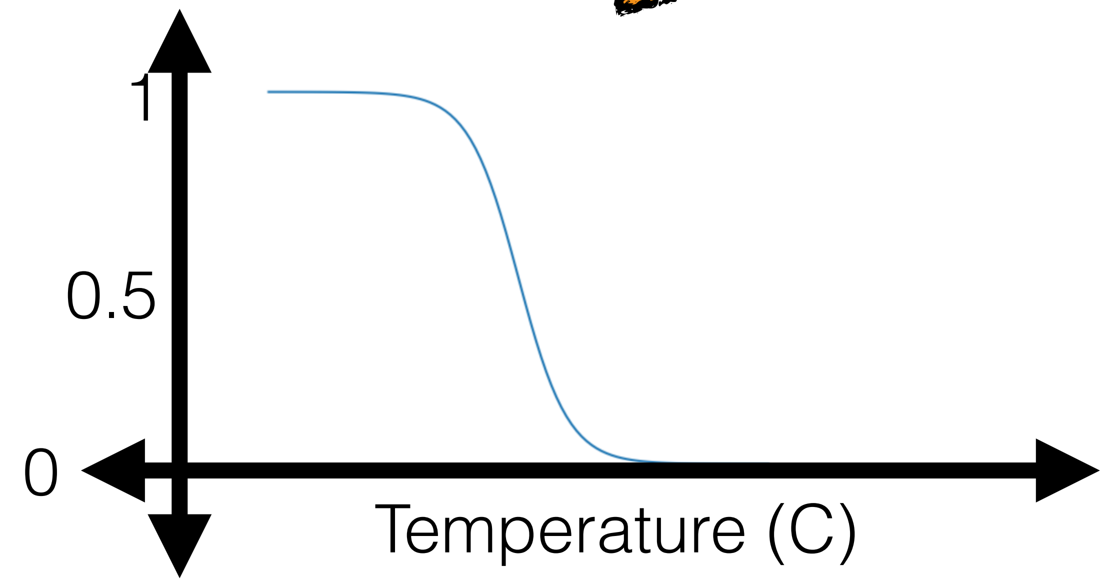
Loss(data) = - log probability(data)

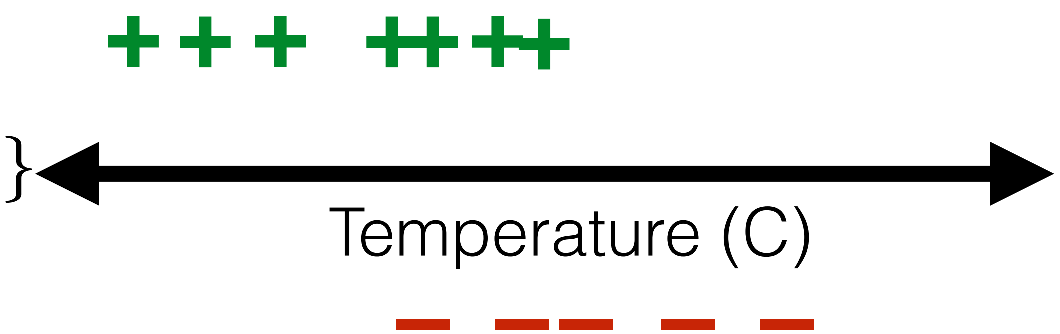
$$= \frac{1}{n} \sum_{i=1}^{n} - \left(\mathbf{1}\{y^{(i)} = +1\} \log g^{(i)} + \mathbf{1}\{y^{(i)} \neq +1\} \log(1 - g^{(i)}) \right.$$

Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?





Probability(data)

$$= \prod_{i=1}^{n} \text{Probability}(\text{data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$

$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)} \neq +1\}}$$

Loss(data) = -(1/n) * log probability(data)

$$= \frac{1}{n} \sum_{i=1}^{n} - \Big(\mathbf{1}\{y^{(i)} = +1\} \log g^{(i)} + \mathbf{1}\{y^{(i)} \neq +1\} \log(1 - g^{(i)})$$

Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

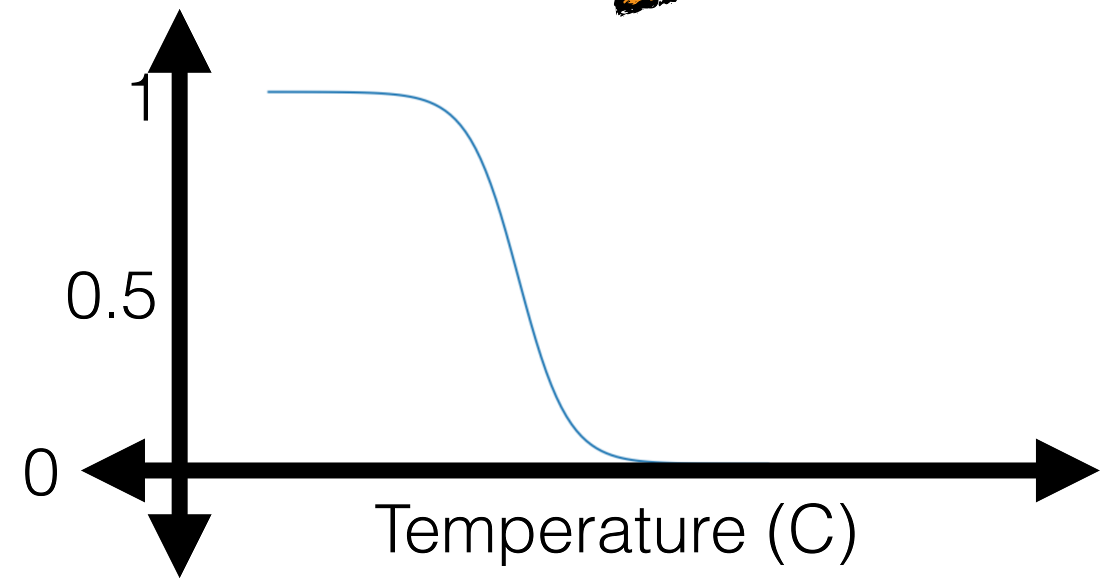
Probability(data)

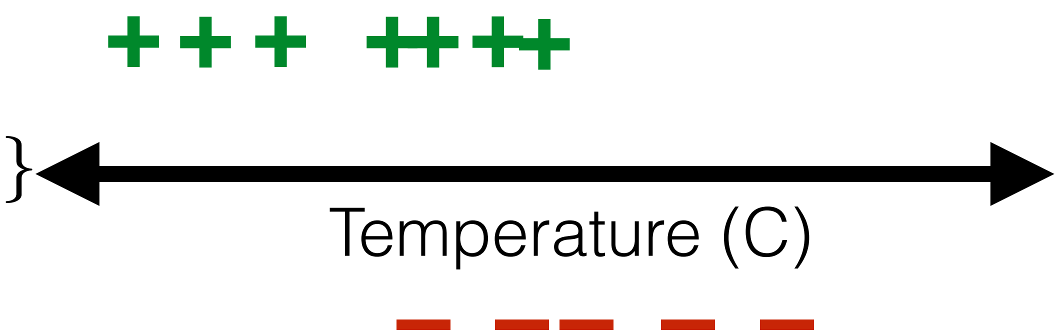
$$= \prod_{i=1}^{n} \text{Probability}(\text{data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$

$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)}\neq+1\}}$$





Loss(data) = -(1/n) * log probability(data)

$$= \frac{1}{n} \sum_{i=1}^{n} -\left(\mathbf{1}\{y^{(i)} = +1\} \log g^{(i)} + \mathbf{1}\{y^{(i)} \neq +1\} \log(1 - g^{(i)})\right)$$

Linear logistic classification

aka logistic regression

- What's an appropriate loss for this guess?

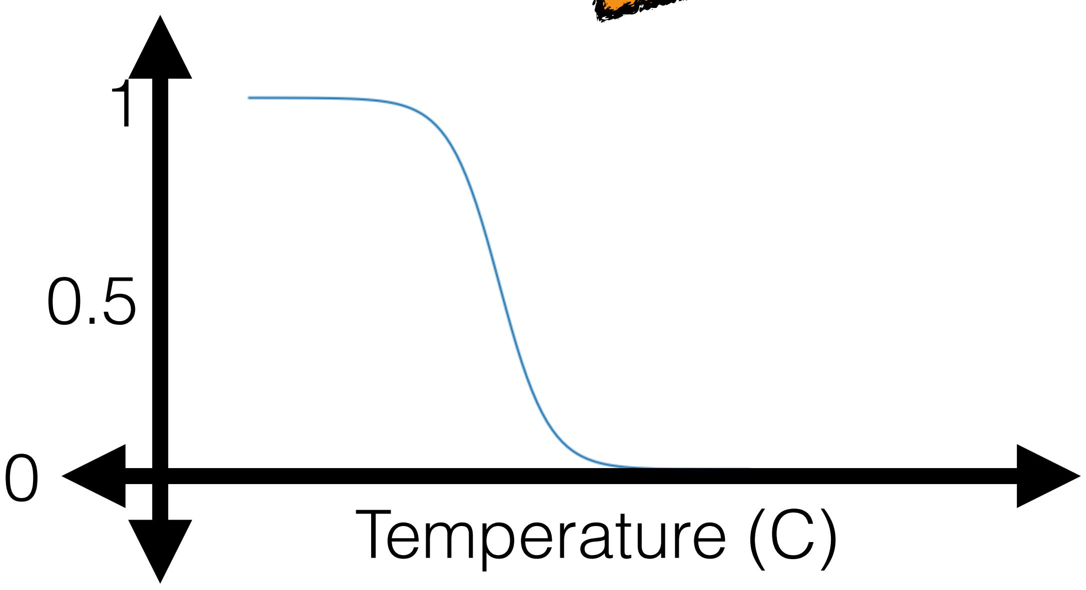
Probability(data)

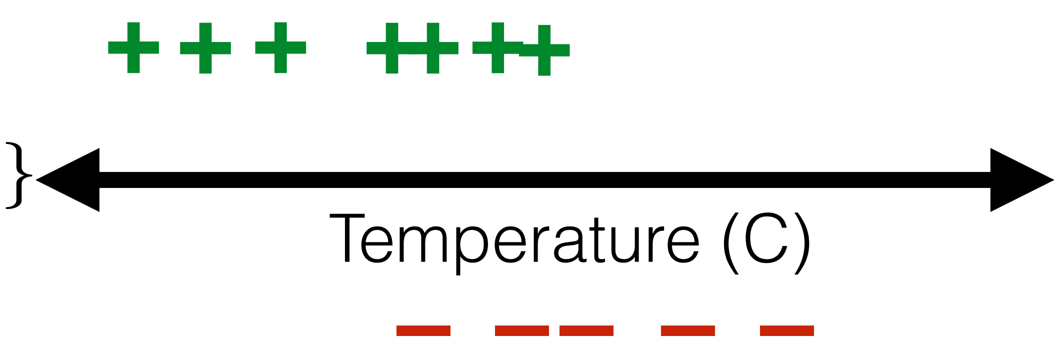
$$= \prod_{i=1}^{n} \text{Probability(data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$

$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)} \neq +1\}}$$





Loss(data) = -(1/n) * log probability(data)

$$= \frac{1}{n} \sum_{i=1}^{n} -\left(\mathbf{1}\{y^{(i)} = +1\} \log g^{(i)} + \mathbf{1}\{y^{(i)} \neq +1\} \log(1 - g^{(i)})\right)$$

Negative log likelihood loss (*g* for guess, *a* for actual):

Linear logistic classification

aka logistic regression

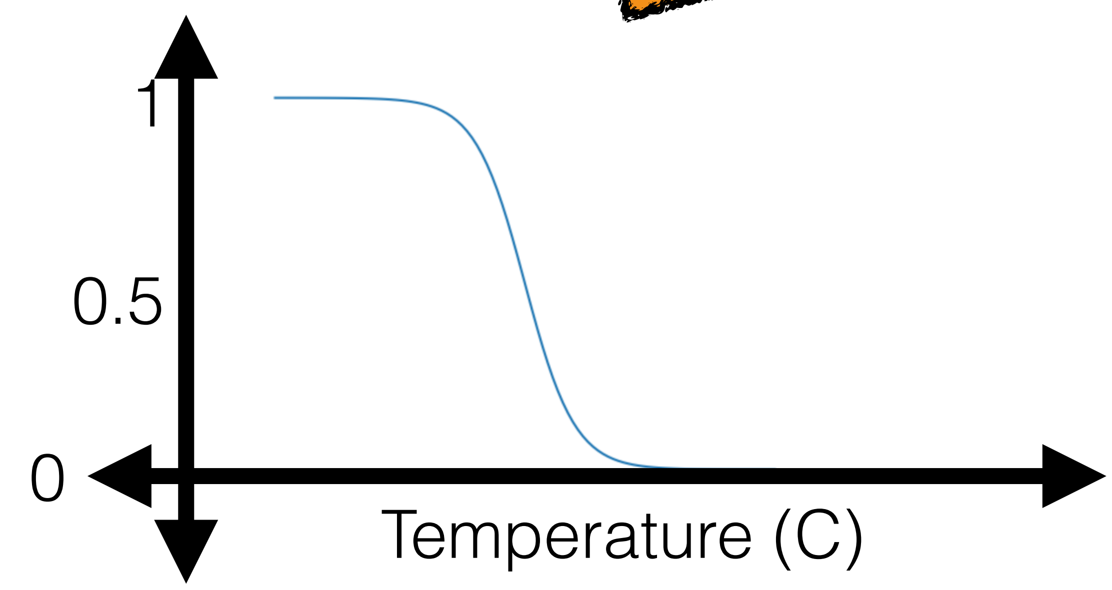
- What's an appropriate loss for this guess?

Probability(data)

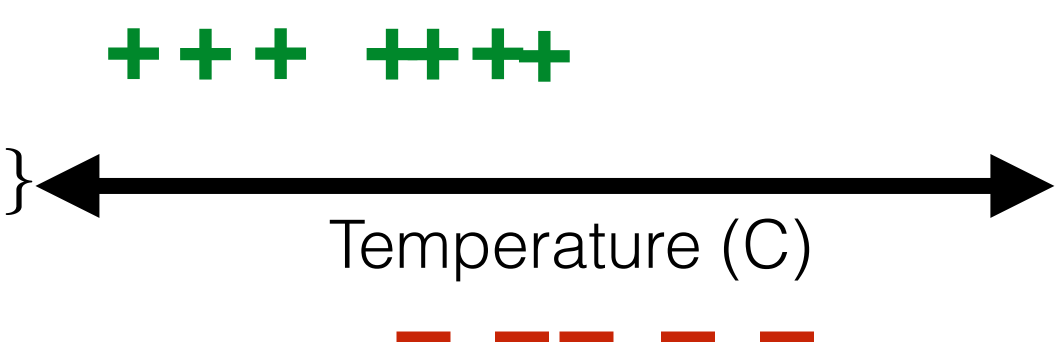
$$= \prod_{i=1}^{n} \text{Probability}(\text{data point } i)$$

[Let $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$]

$$= \prod_{i=1}^{n} \begin{cases} g^{(i)} \text{ if } y^{(i)} = +1 \\ (1 - g^{(i)}) \text{ else} \end{cases}$$



$$= \prod_{i=1}^{n} (g^{(i)})^{\mathbf{1}\{y^{(i)}=+1\}} (1 - g^{(i)})^{\mathbf{1}\{y^{(i)}\neq+1\}}$$



Loss(data) = -(1/n) * log probability(data)

$$= \frac{1}{n} \sum_{i=1}^{n} -\left(\mathbf{1}\{y^{(i)} = +1\} \log g^{(i)} + \mathbf{1}\{y^{(i)} \neq +1\} \log(1 - g^{(i)})\right)$$

Negative log likelihood loss (*g* for guess, *a* for actual):

$$-L_{\text{nll}}(g, a) = (1\{a = +1\} \log g + 1\{a \neq +1\} \log(1 - g))$$

Gradient descent for logistic regression

Gradient descent for logistic regression

- Want to minimize average (negative log likelihood) loss across the data

$$J_{\mathrm{lr}}(\Theta) = J_{\mathrm{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\mathrm{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

Gradient descent for logistic regression

- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

Gradient descent for logistic regression

- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

$$J_{\mathrm{lr}}(\Theta) = J_{\mathrm{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\mathrm{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

Gradient descent for logistic regression

- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

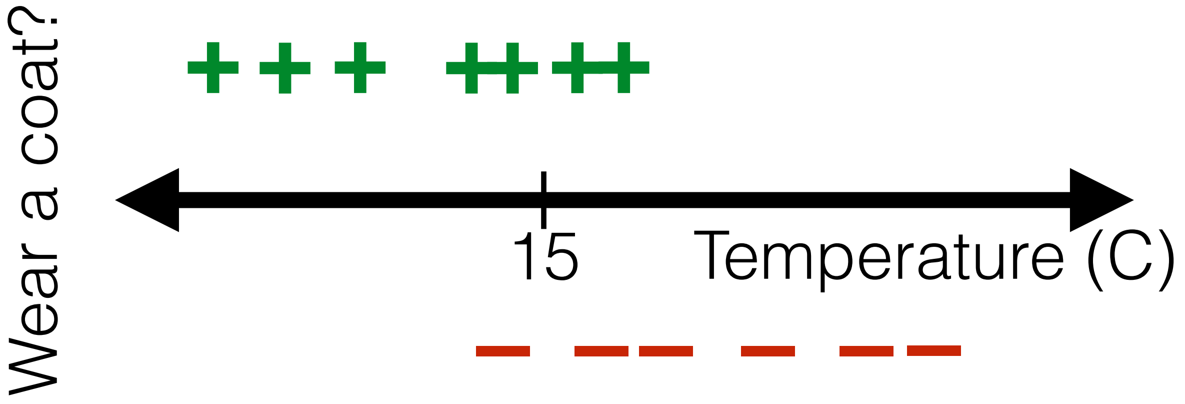
- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`

Gradient descent for logistic regression

- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`

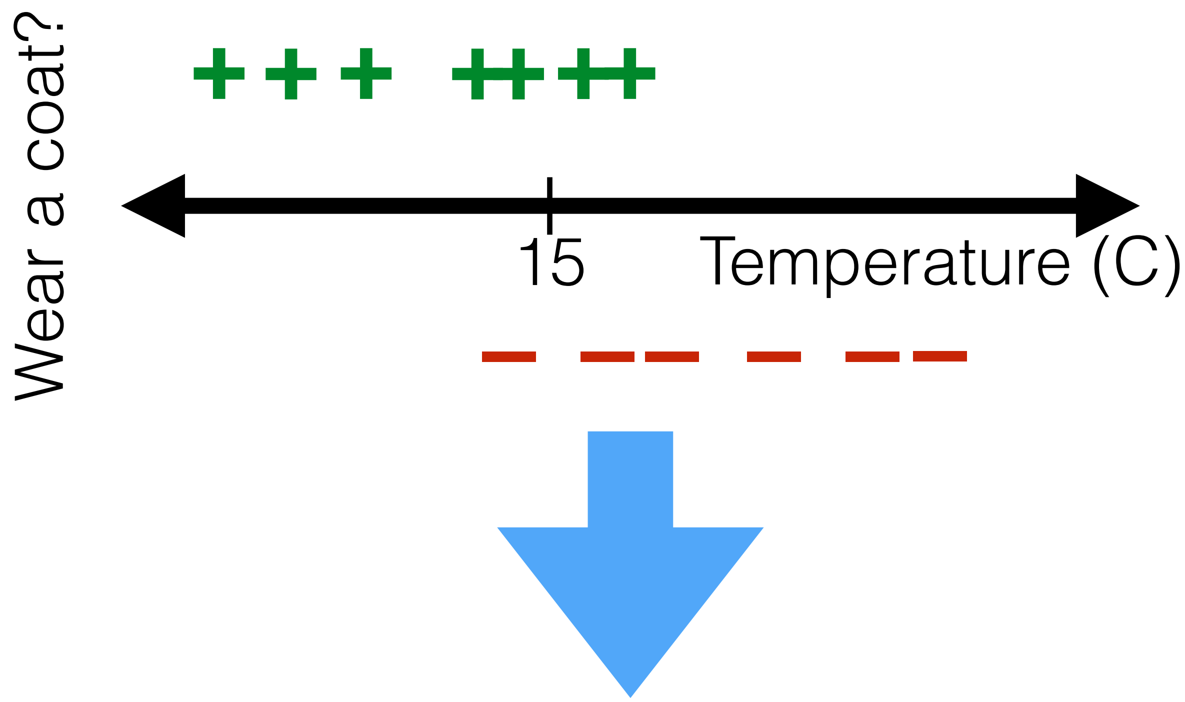


Gradient descent for logistic regression

- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`

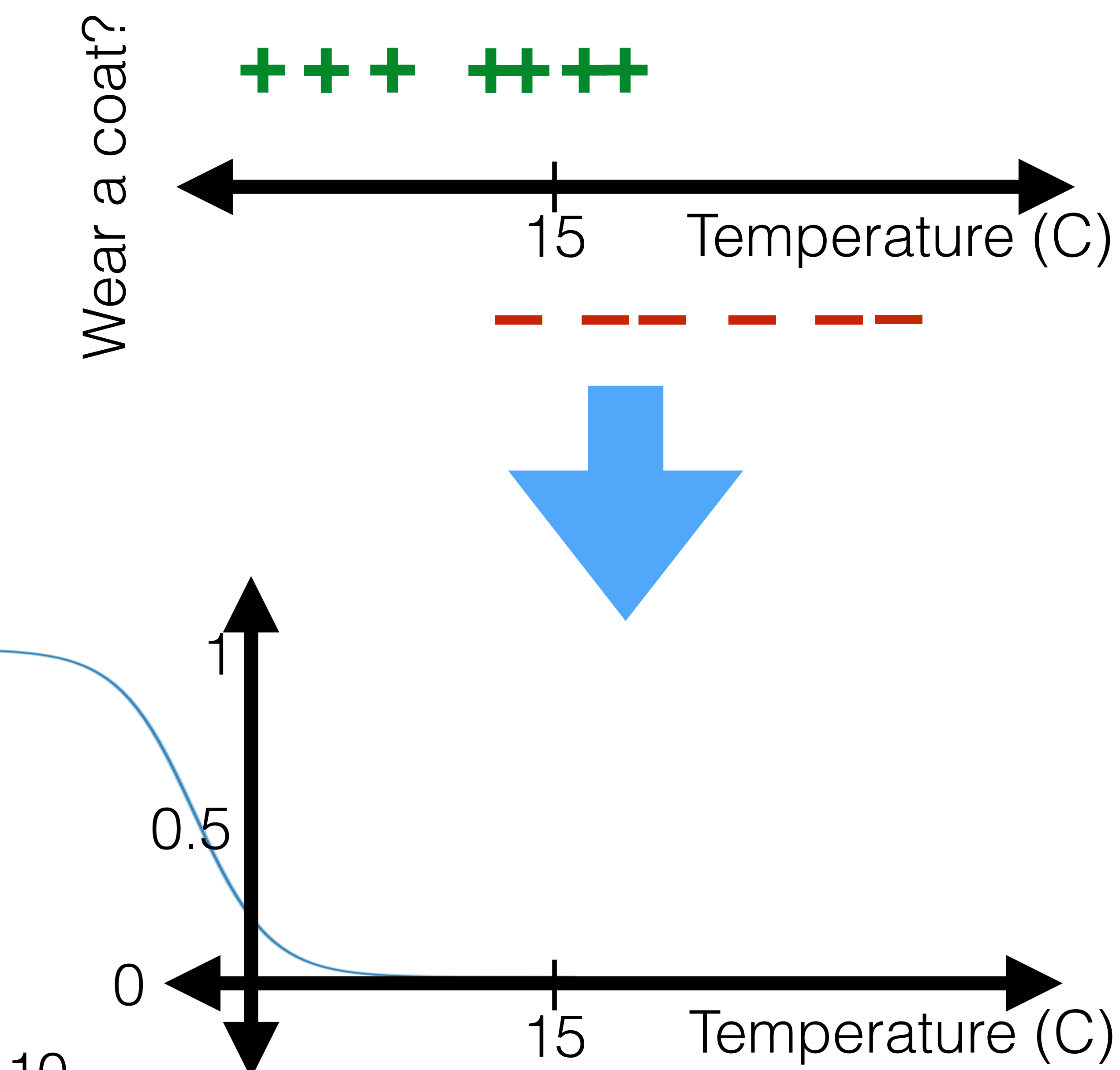


Gradient descent for logistic regression

- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

$$J_{\mathrm{lr}}(\Theta) = J_{\mathrm{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\mathrm{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

- Run `Gradient-Descent(` $\Theta_{\mathrm{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`

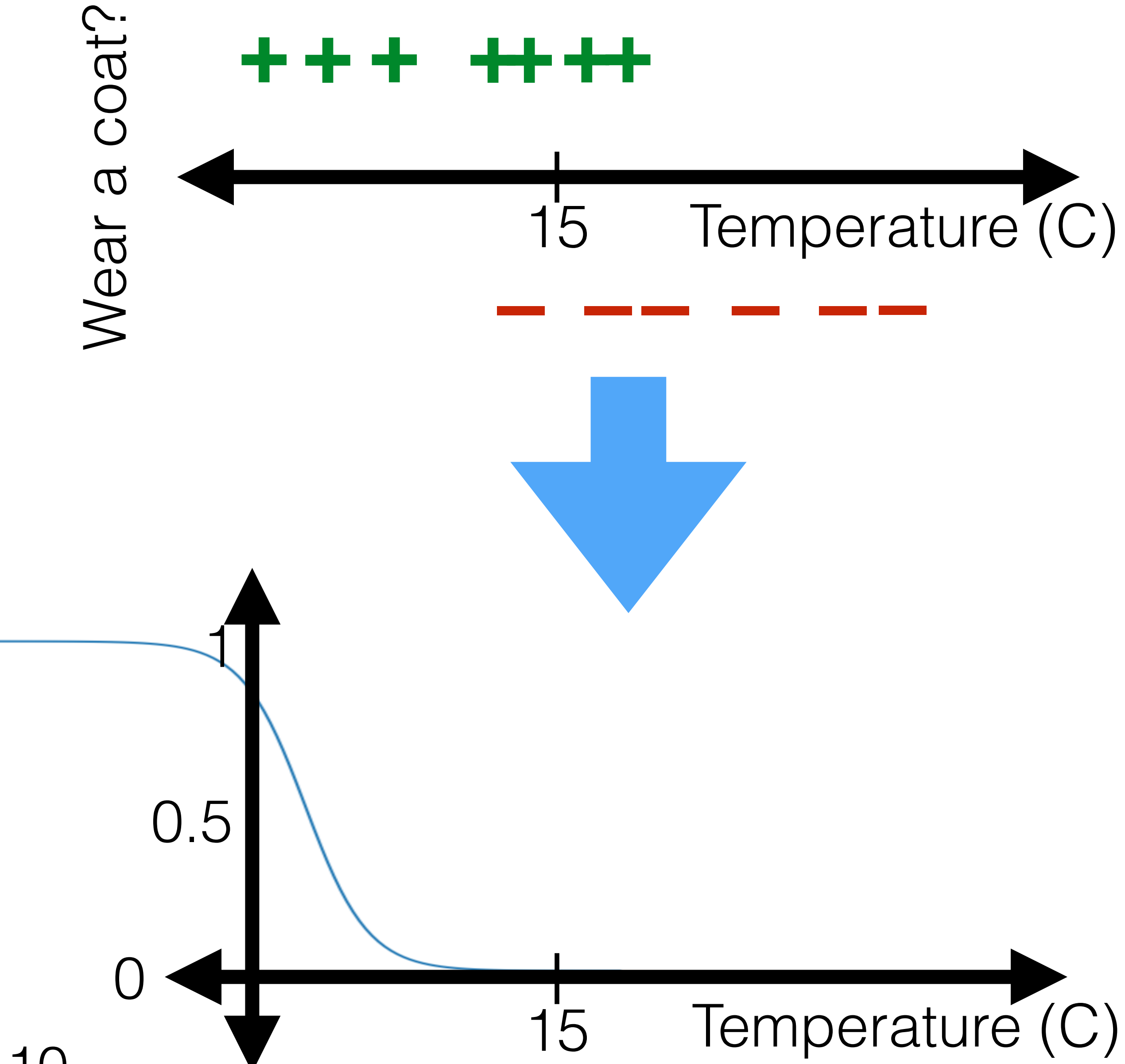


Gradient descent for logistic regression

- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

$$J_{\mathrm{lr}}(\Theta) = J_{\mathrm{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\mathrm{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

- Run `Gradient-Descent(` $\Theta_{\mathrm{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`

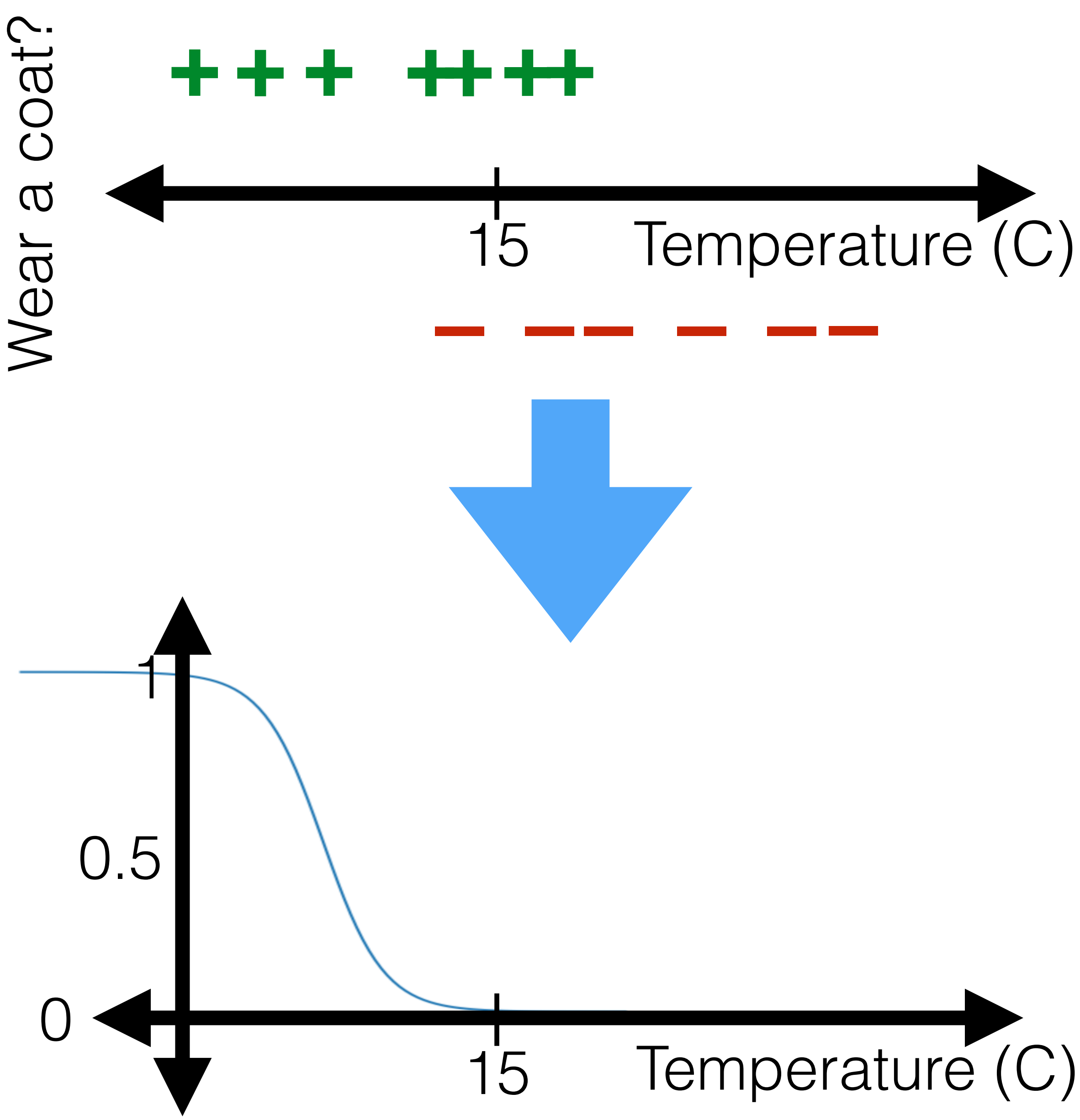


Gradient descent for logistic regression

- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

$$J_{\mathrm{lr}}(\Theta) = J_{\mathrm{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\mathrm{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

- Run `Gradient-Descent(` $\Theta_{\mathrm{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`

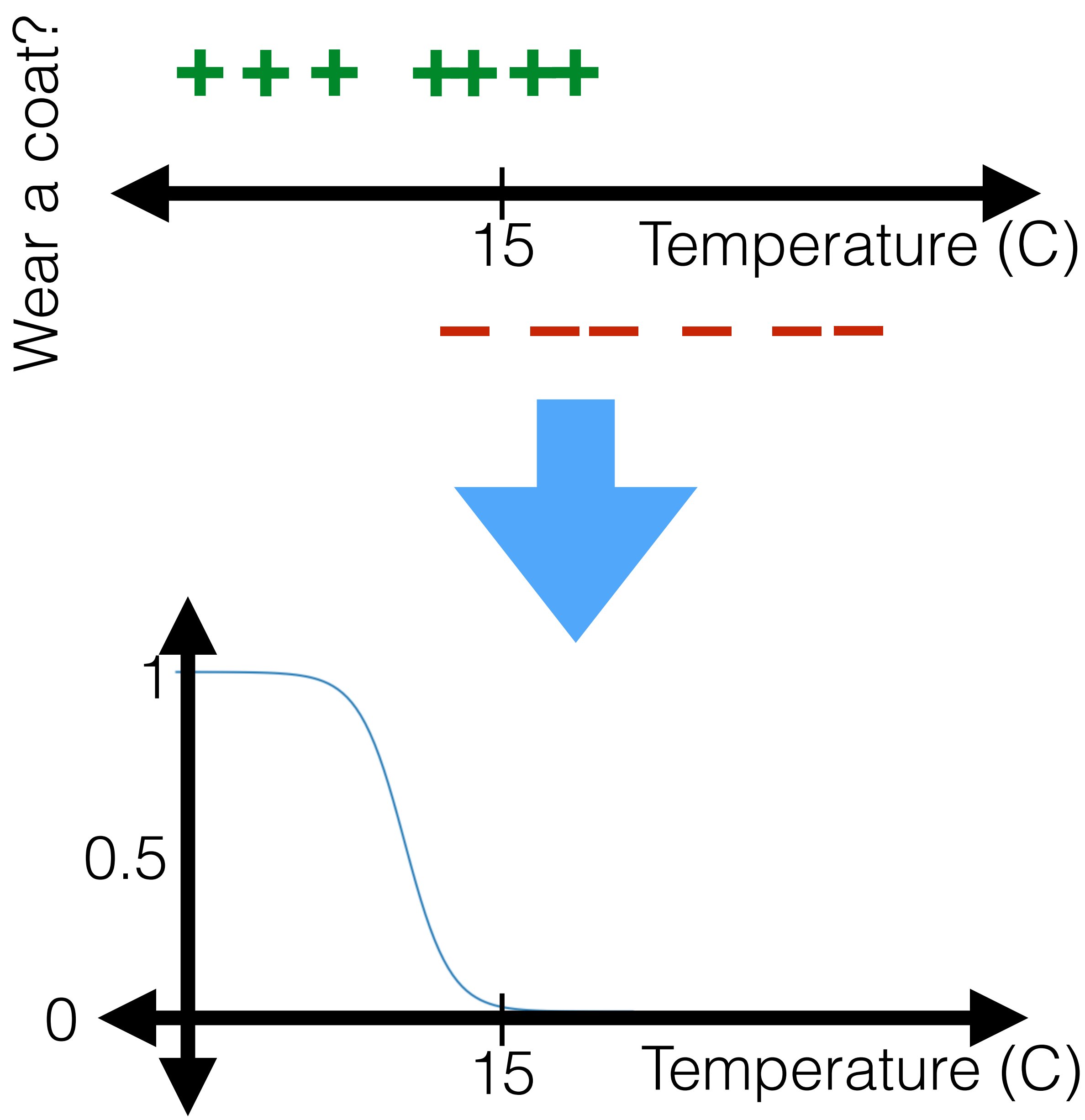


Gradient descent for logistic regression

- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`

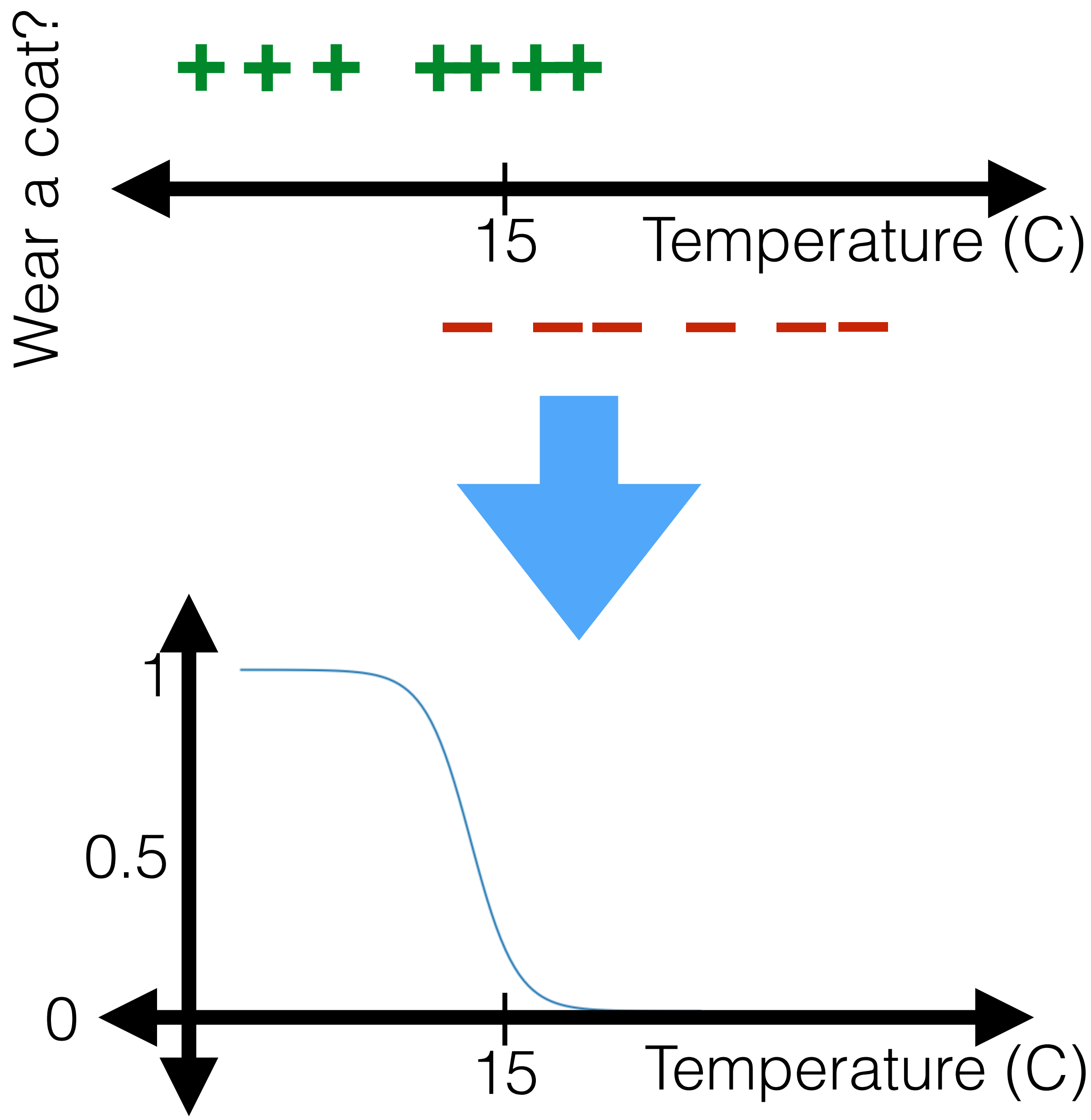


Gradient descent for logistic regression

- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`

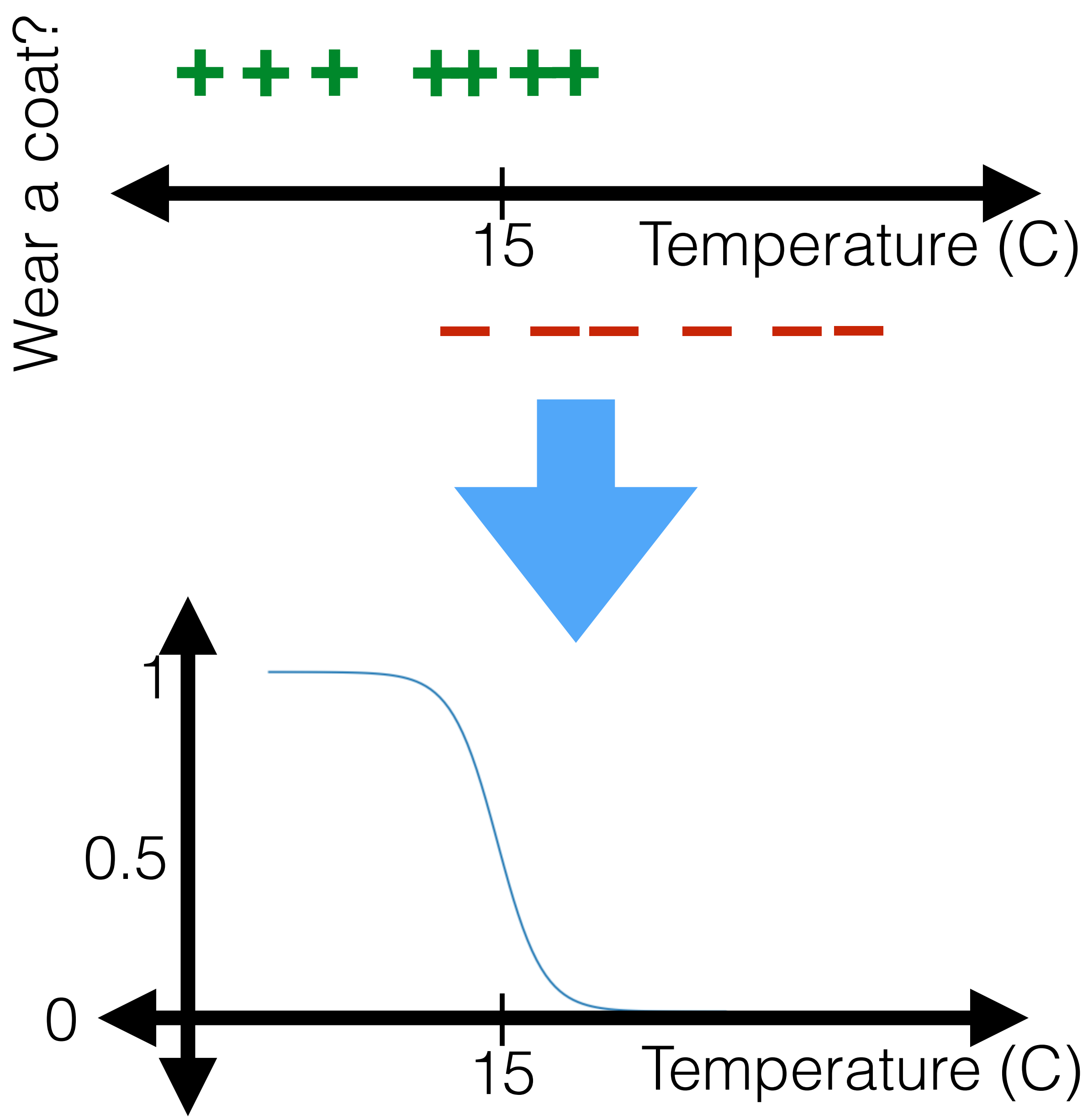


Gradient descent for logistic regression

- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`

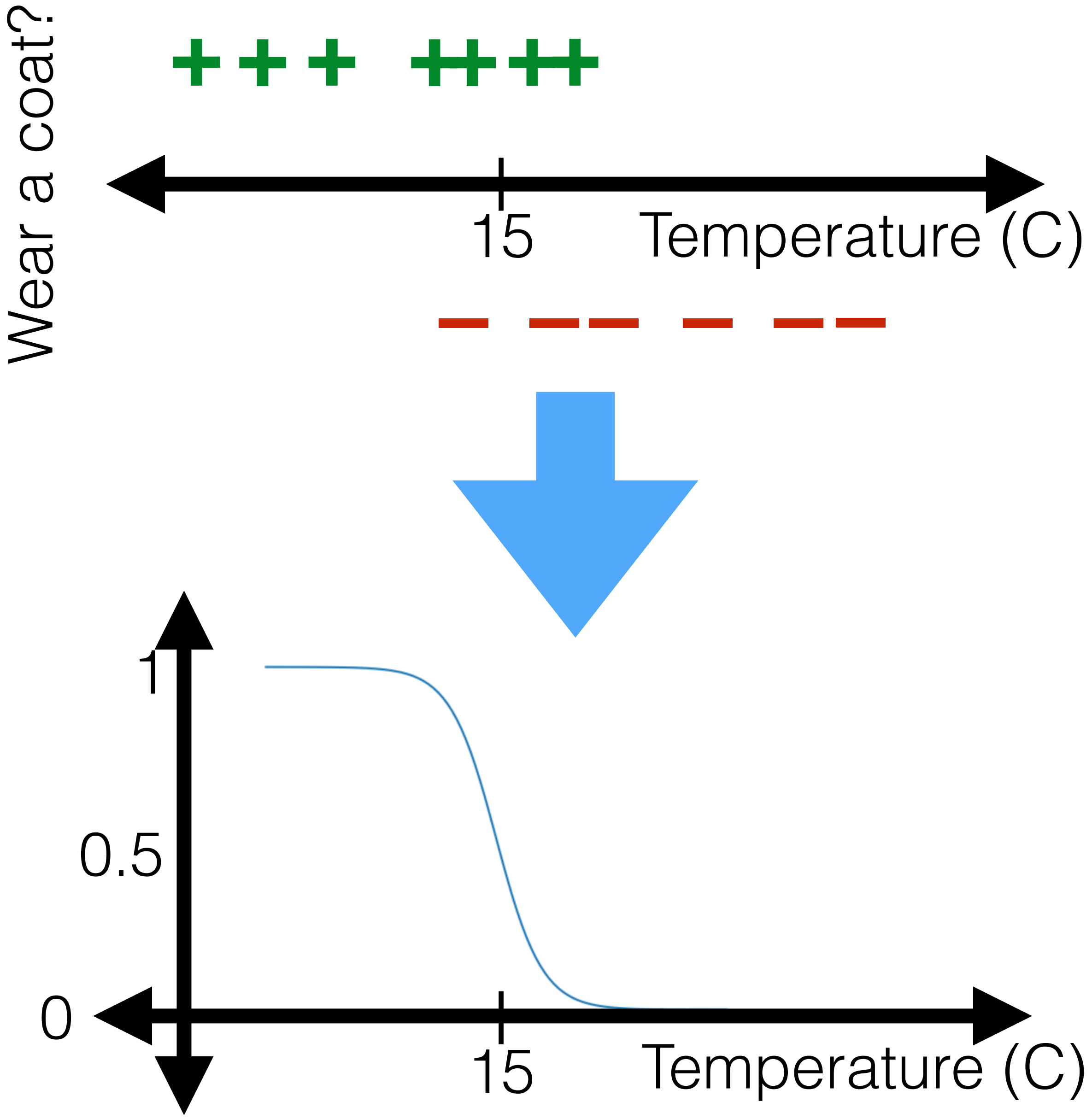


Gradient descent for logistic regression

- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`

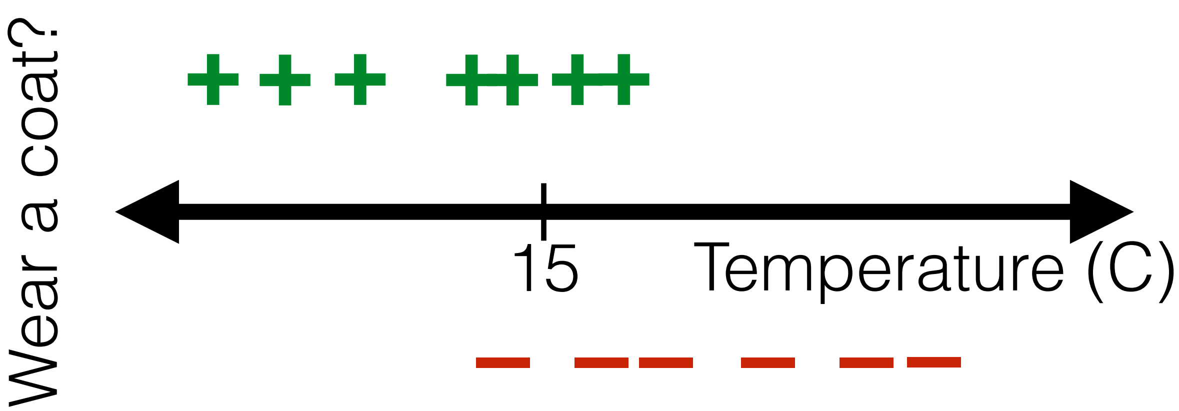


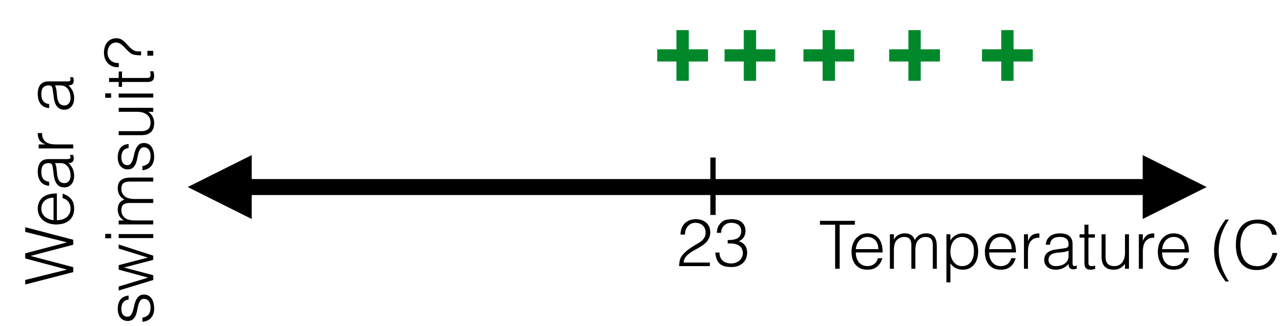
Gradient descent for logistic regression

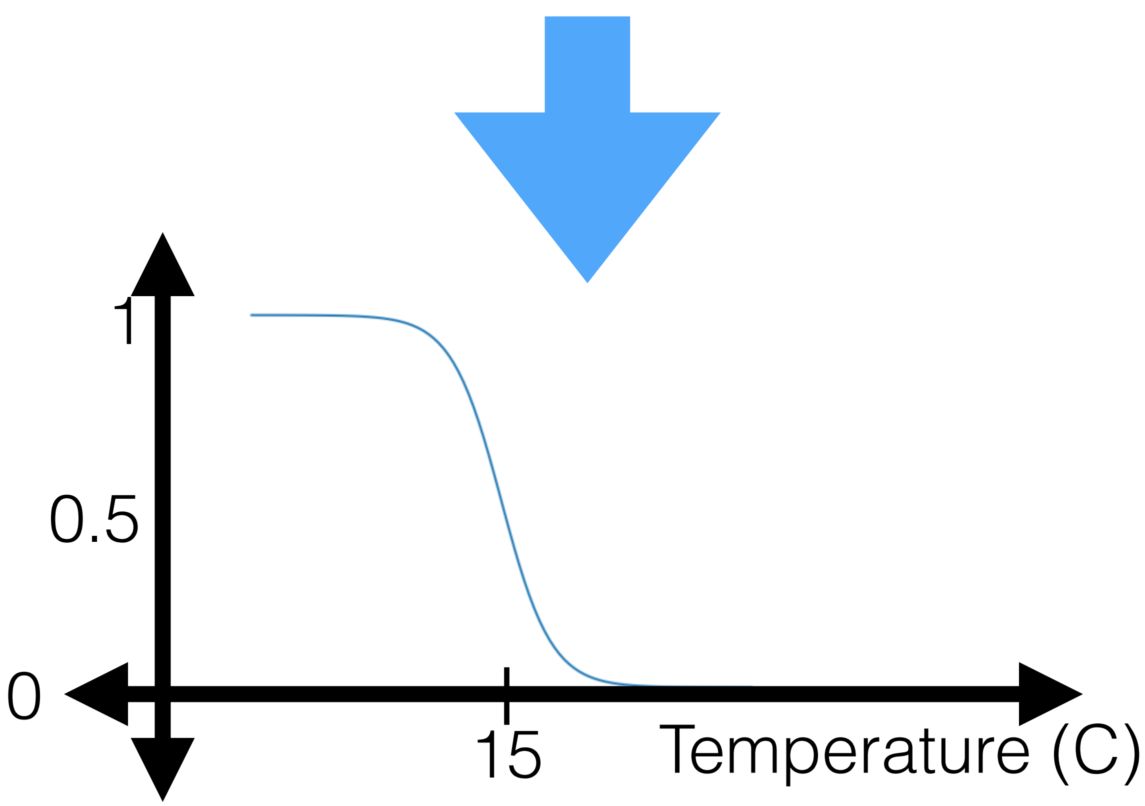
- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

$$J_{\mathrm{lr}}(\Theta) = J_{\mathrm{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\mathrm{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

- Run `Gradient-Descent(` $\Theta_{\mathrm{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`





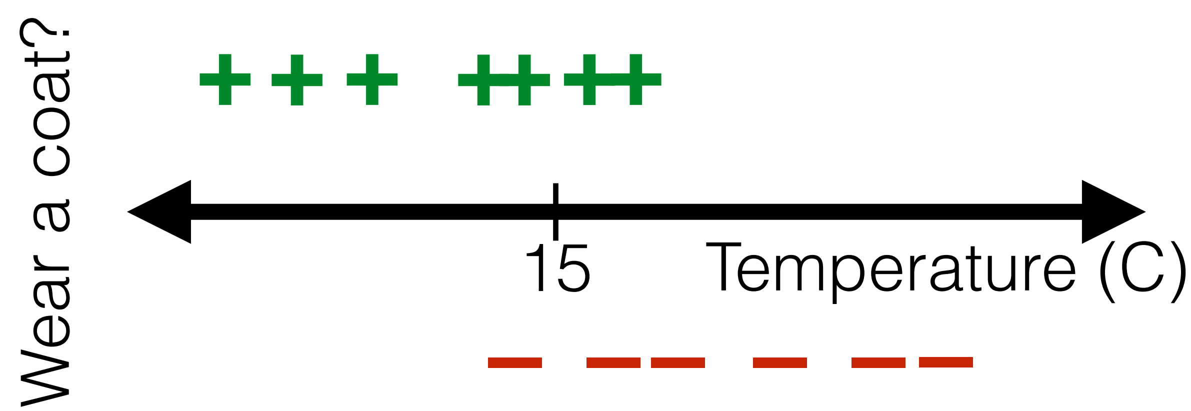


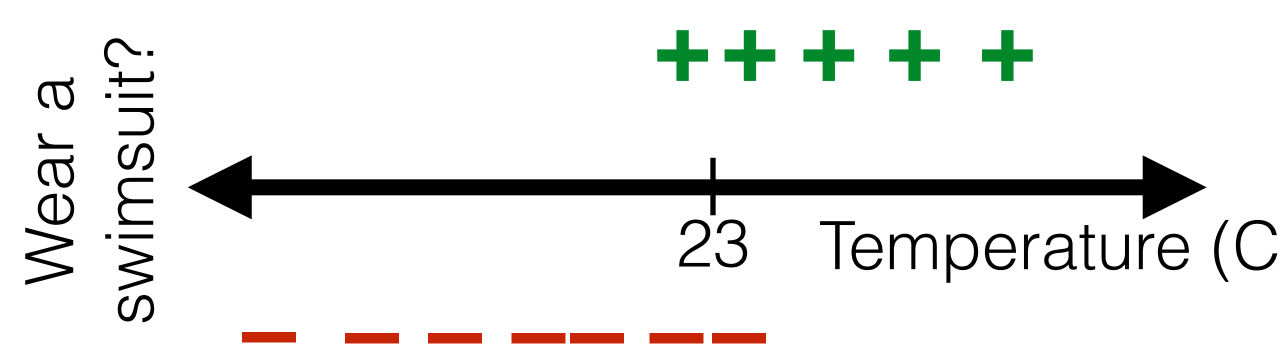
Gradient descent for logistic regression

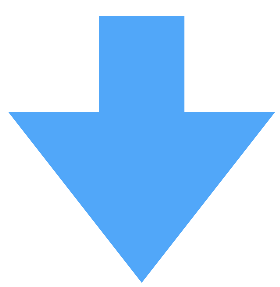
- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

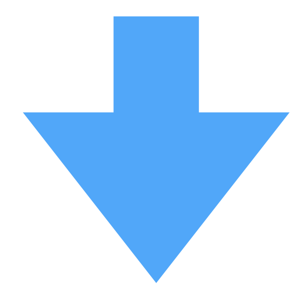
$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

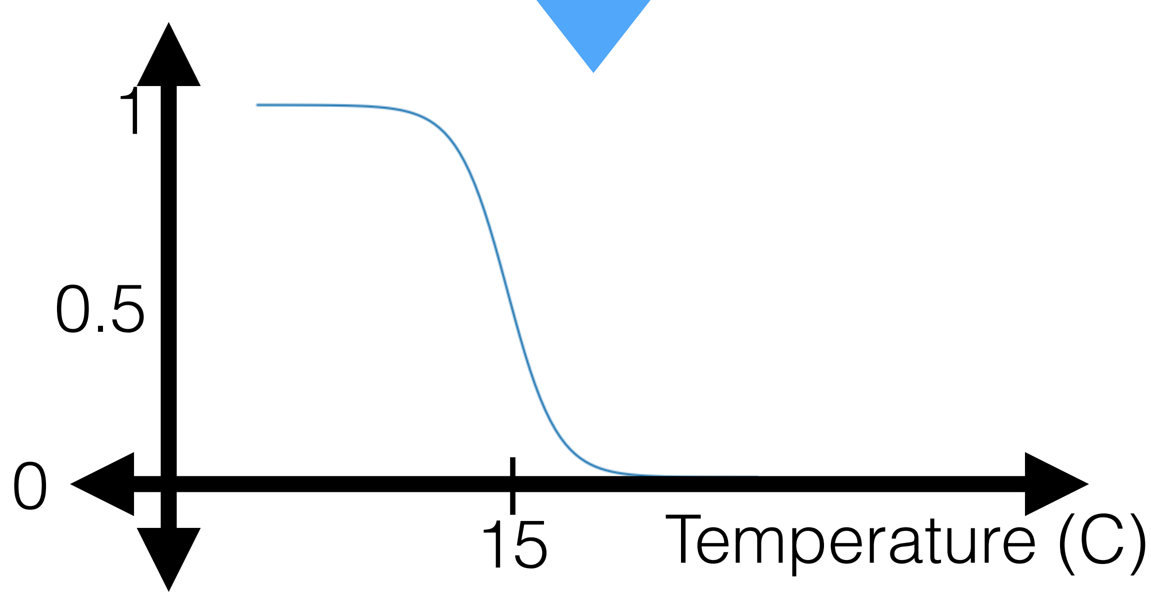
- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`









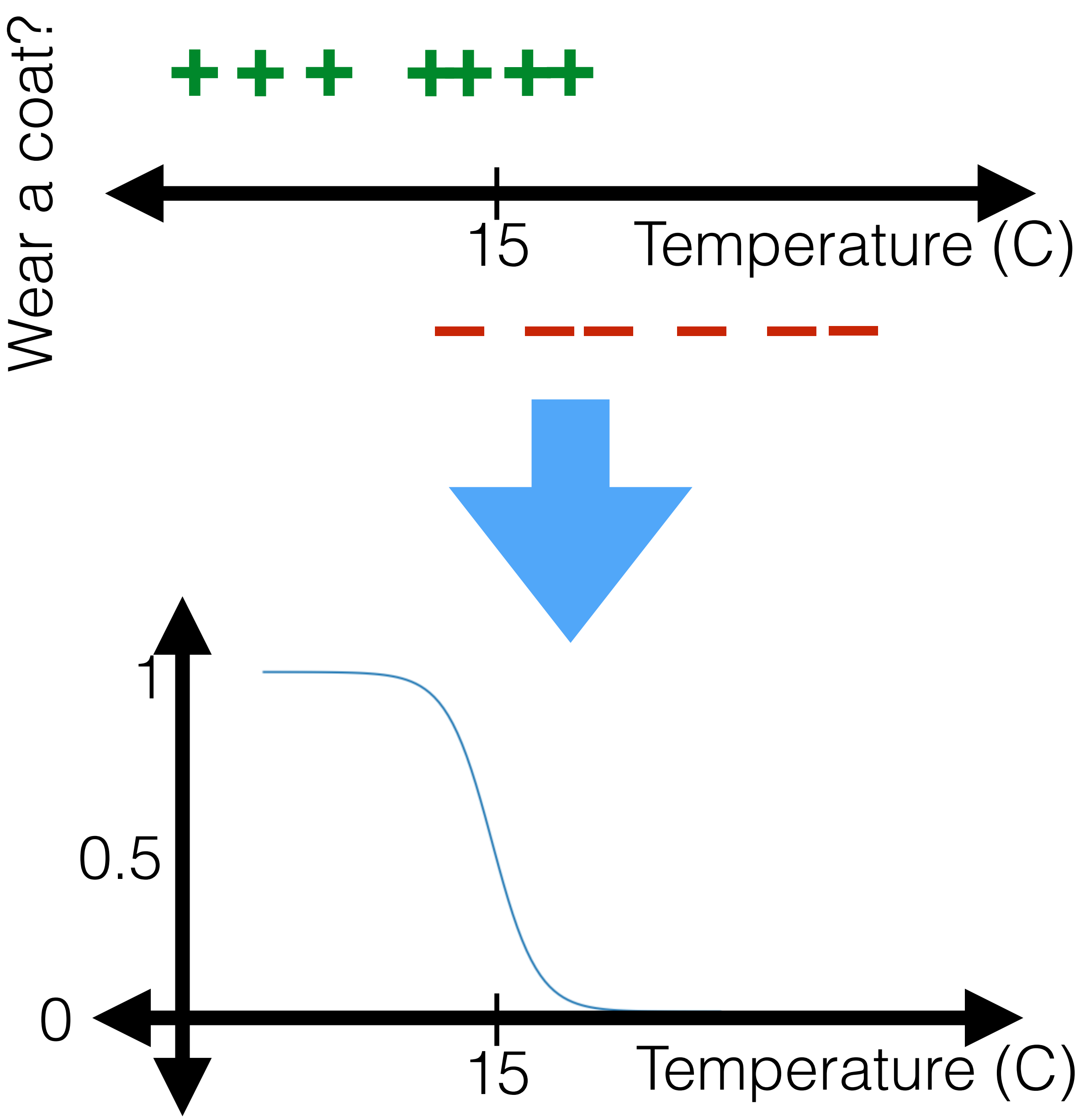


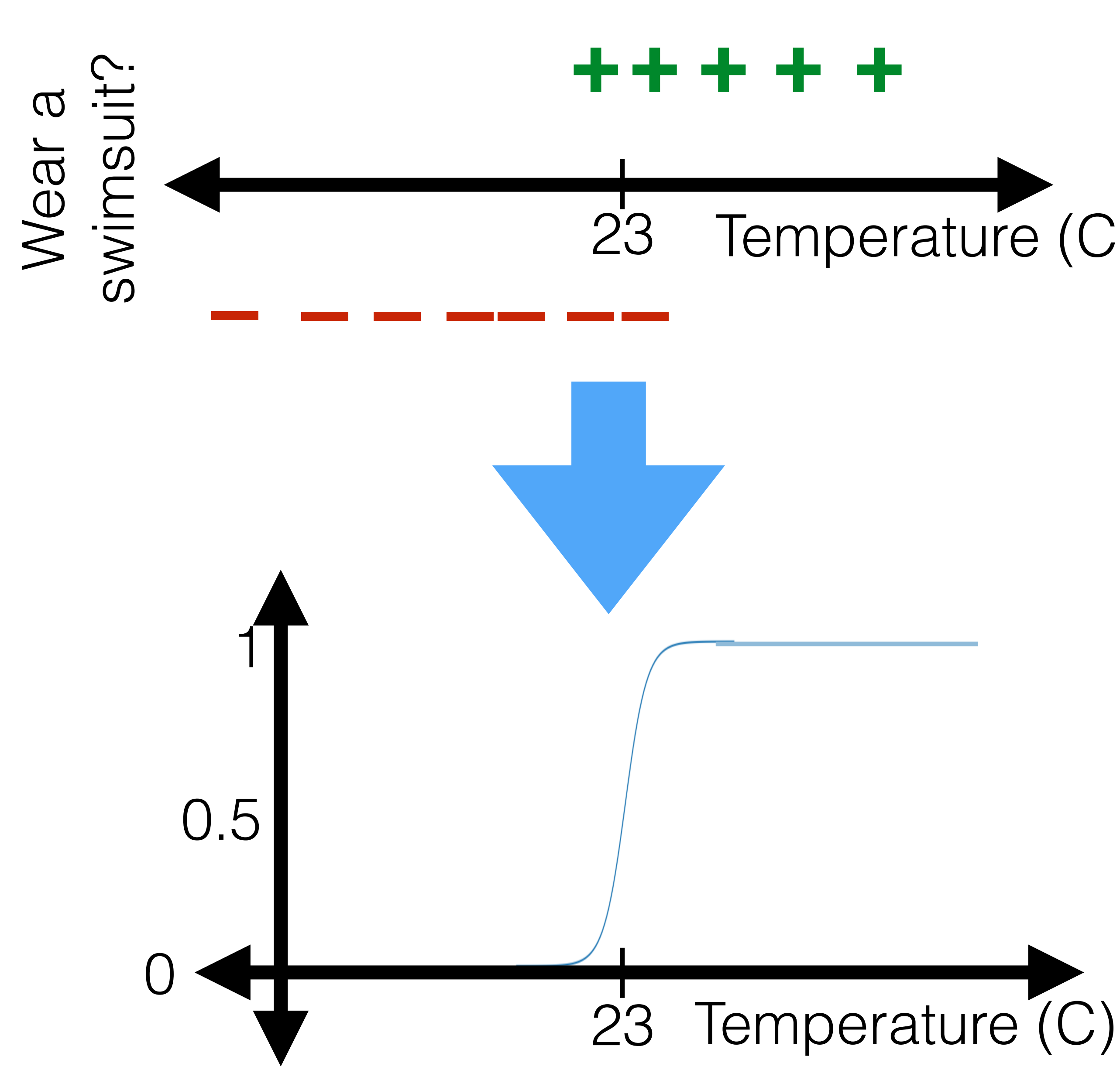
Gradient descent for logistic regression

- Want to minimize average (negative log likelihood) loss across the data (objective is differentiable and convex)

$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)})$$

- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`





Gradient descent for logistic regression

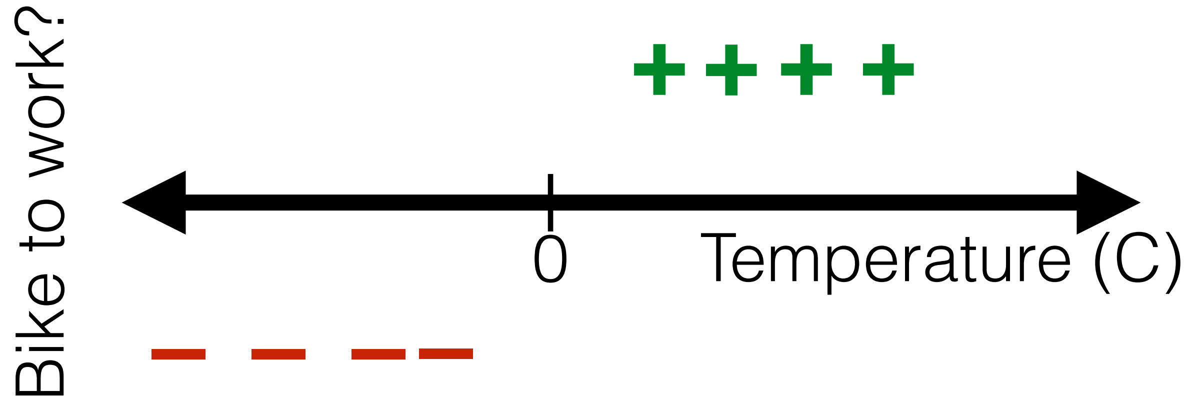
- Can still have practical issues though!

Gradient descent for logistic regression

- Can still have practical issues though!
- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`

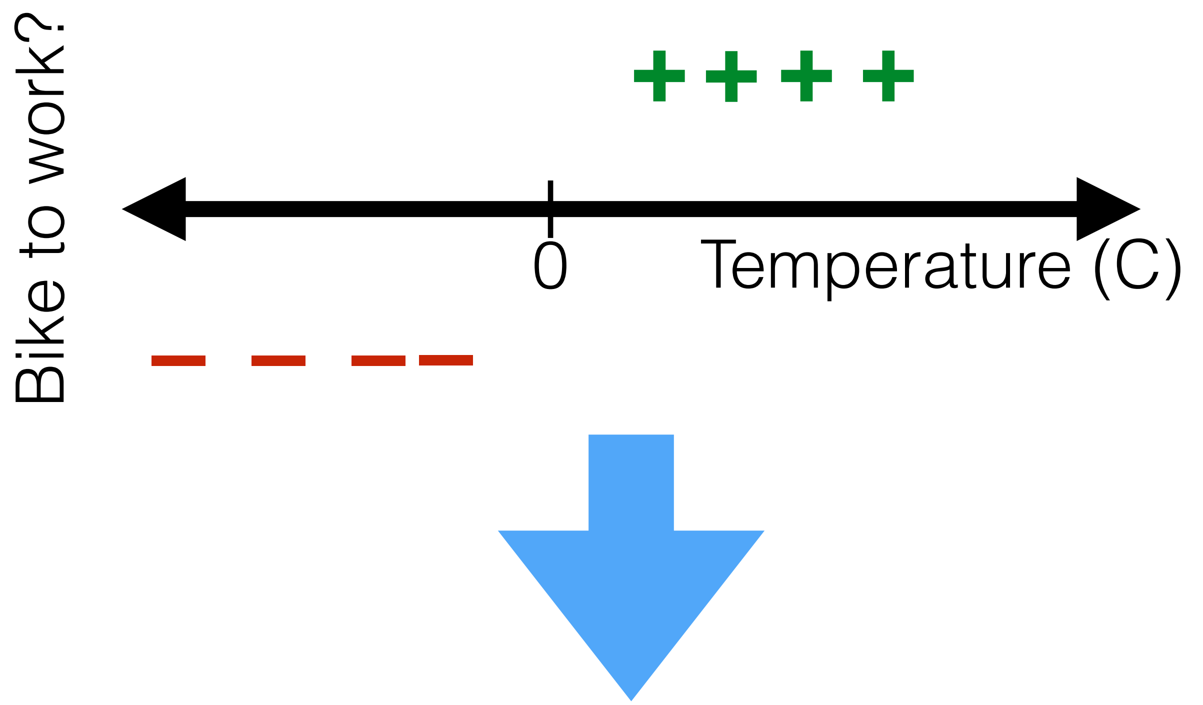
Gradient descent for logistic regression

- Can still have practical issues though!
- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`



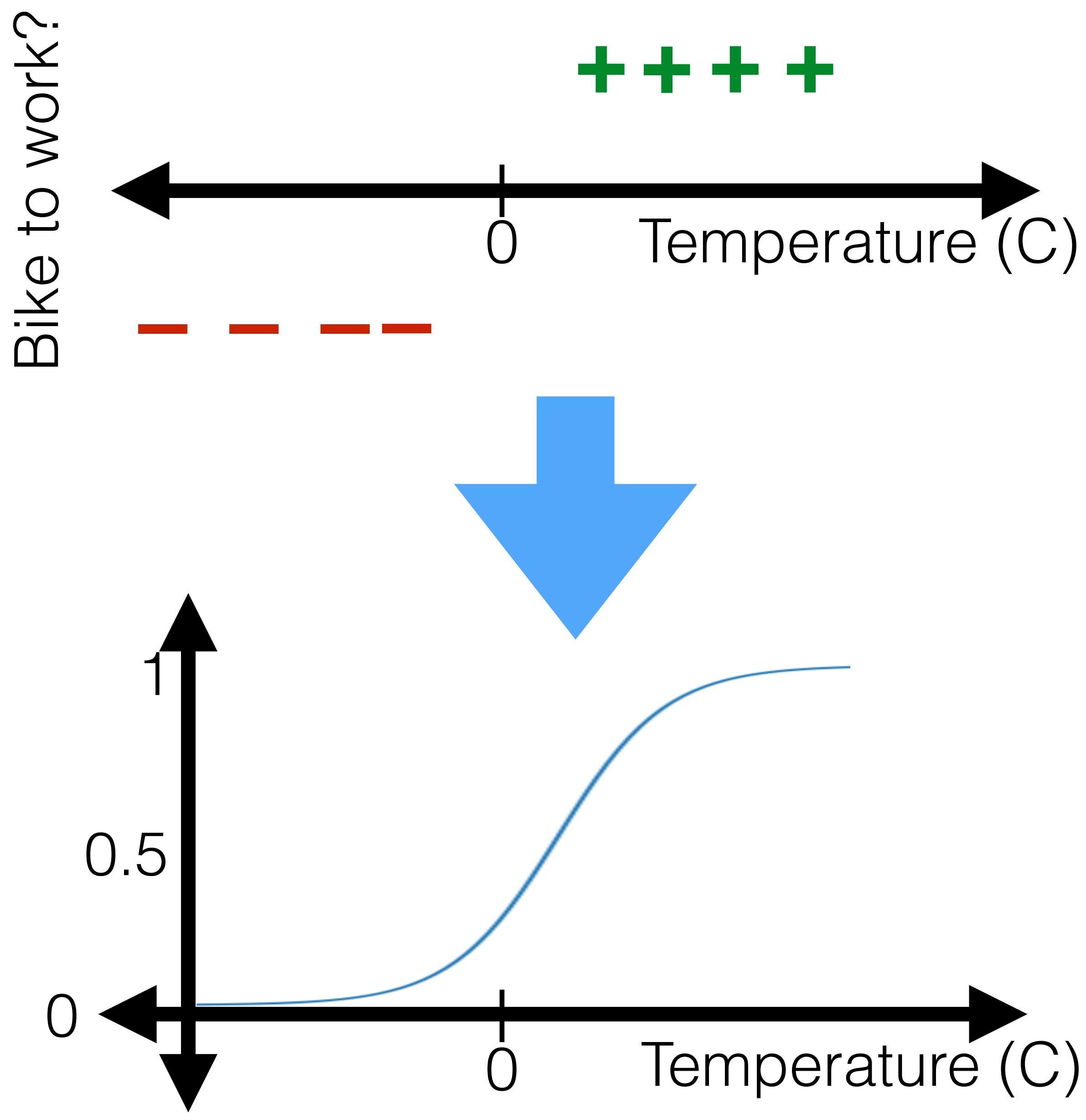
Gradient descent for logistic regression

- Can still have practical issues though!
- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`



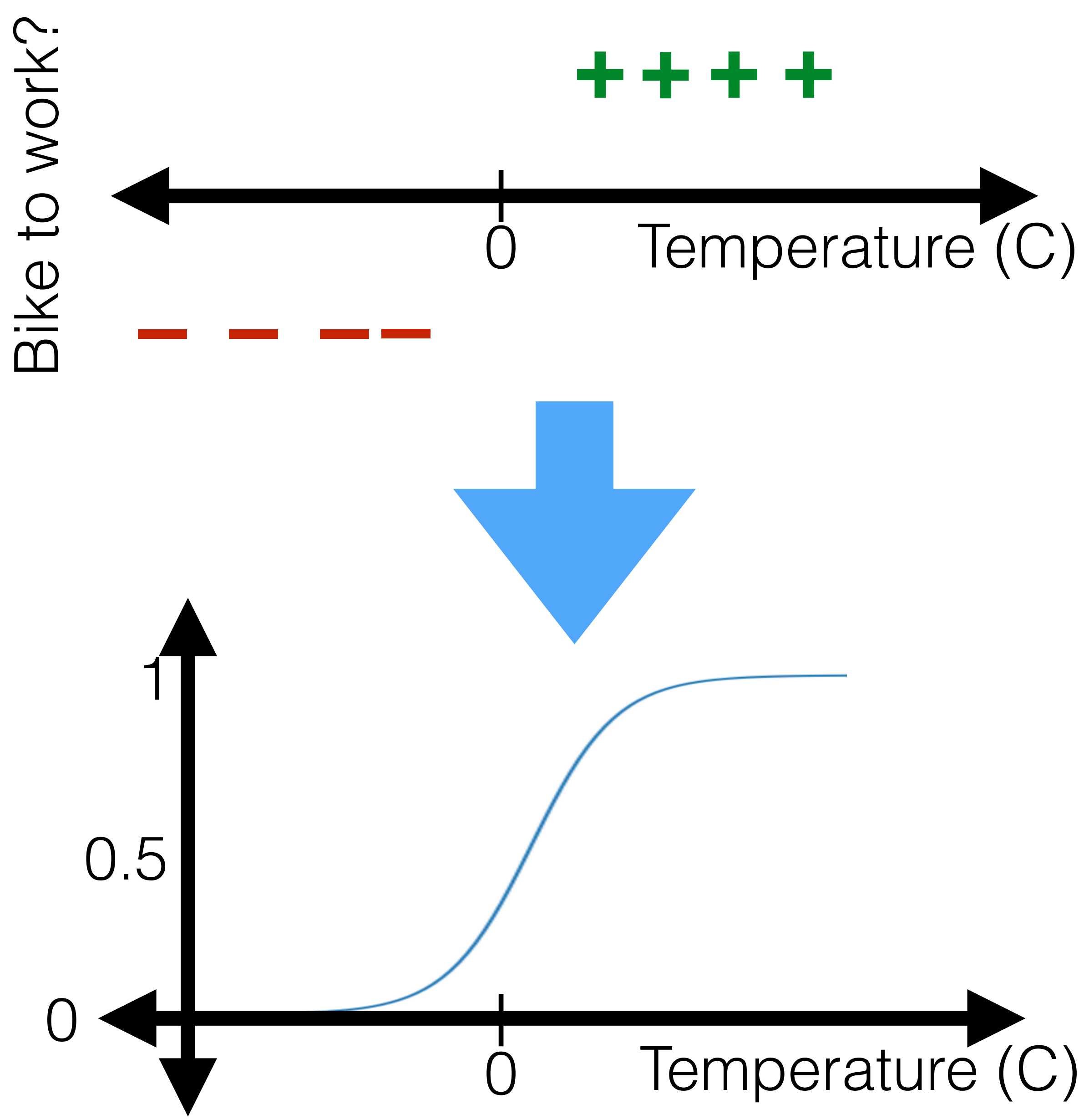
Gradient descent for logistic regression

- Can still have practical issues though!
- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_{\Theta} J_{lr}, \epsilon$ `)`



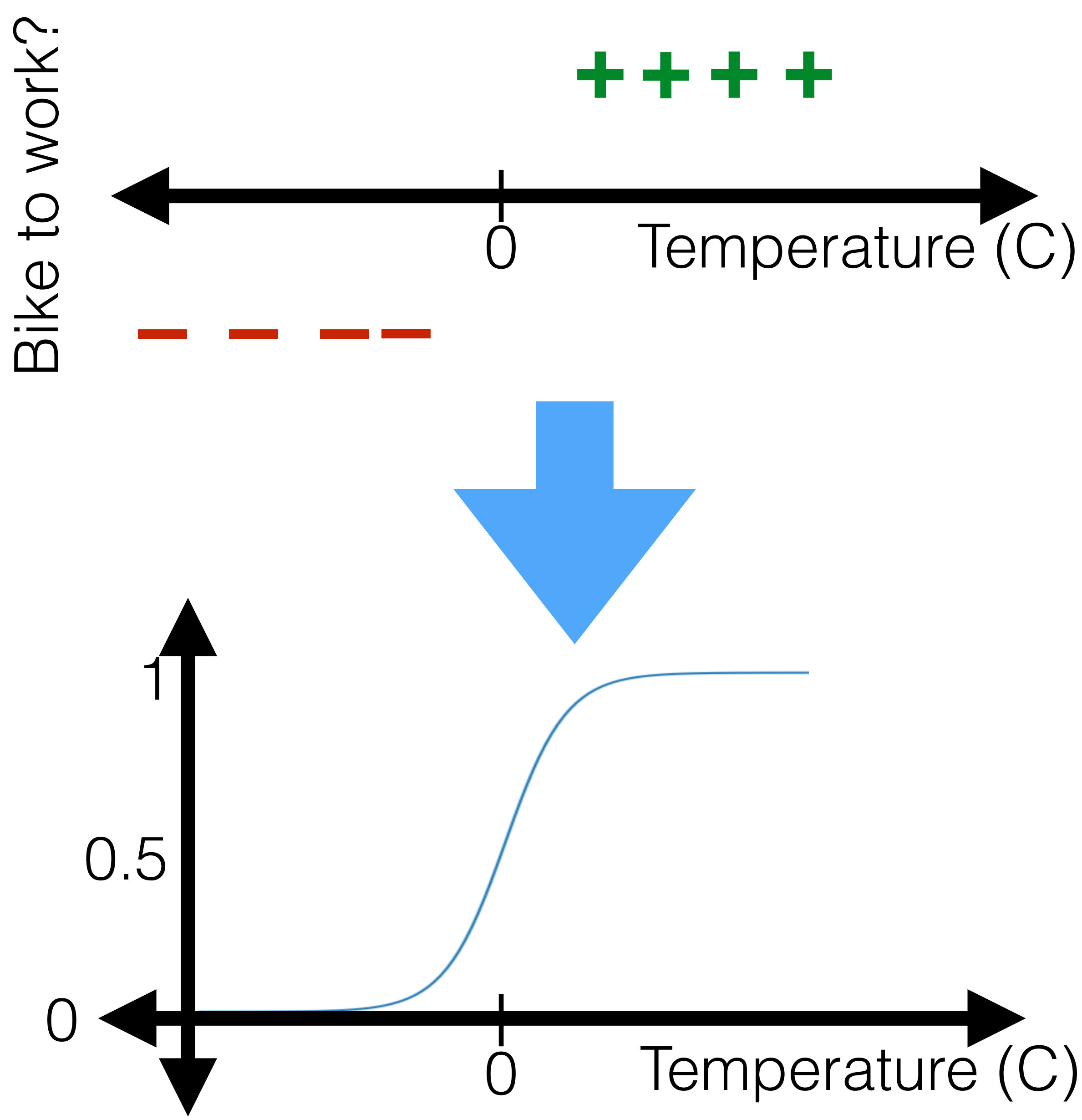
Gradient descent for logistic regression

- Can still have practical issues though!
- Run `Gradient-Descent` ($\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$)



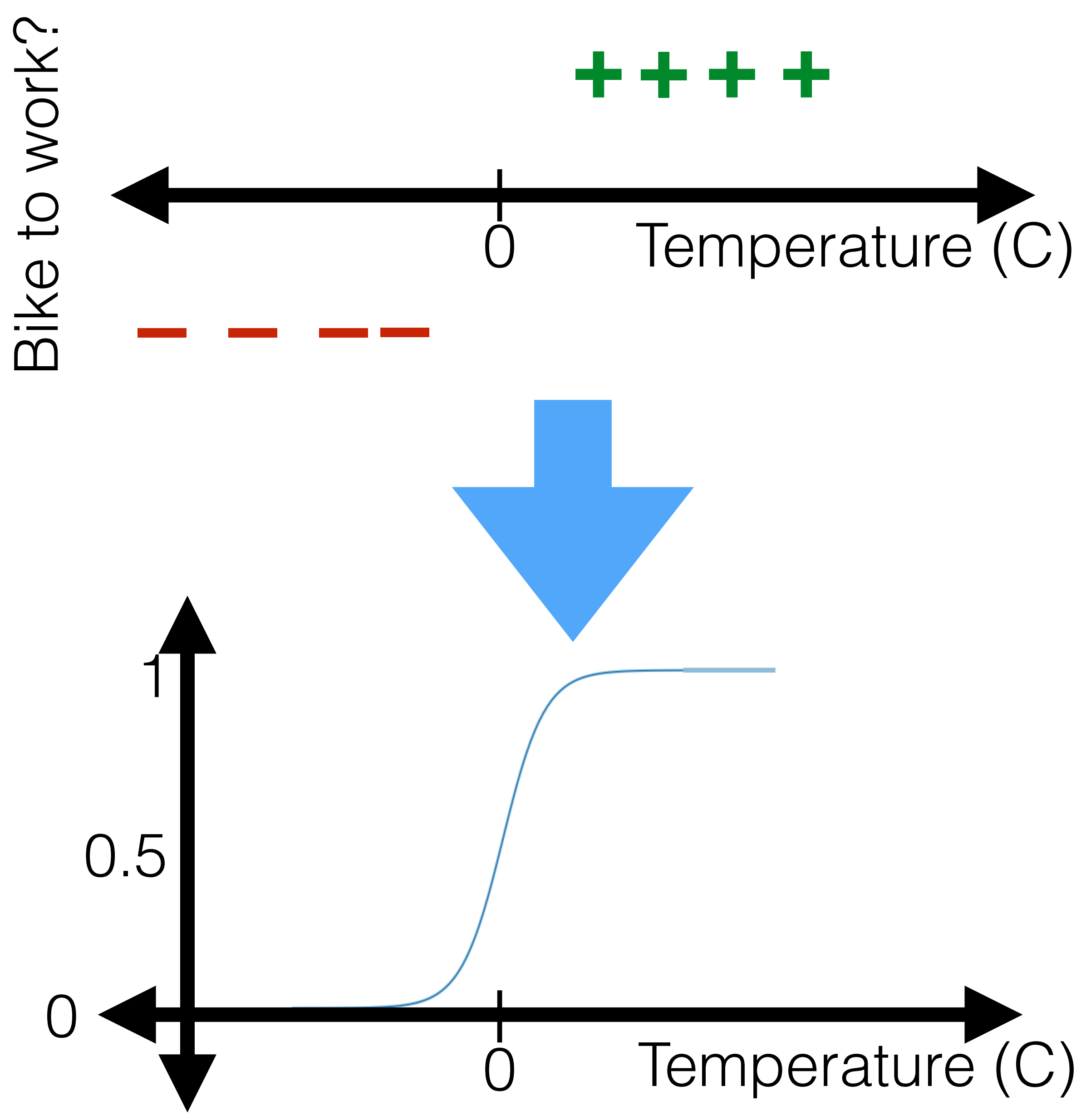
Gradient descent for logistic regression

- Can still have practical issues though!
- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_{\Theta} J_{lr}, \epsilon$ `)`



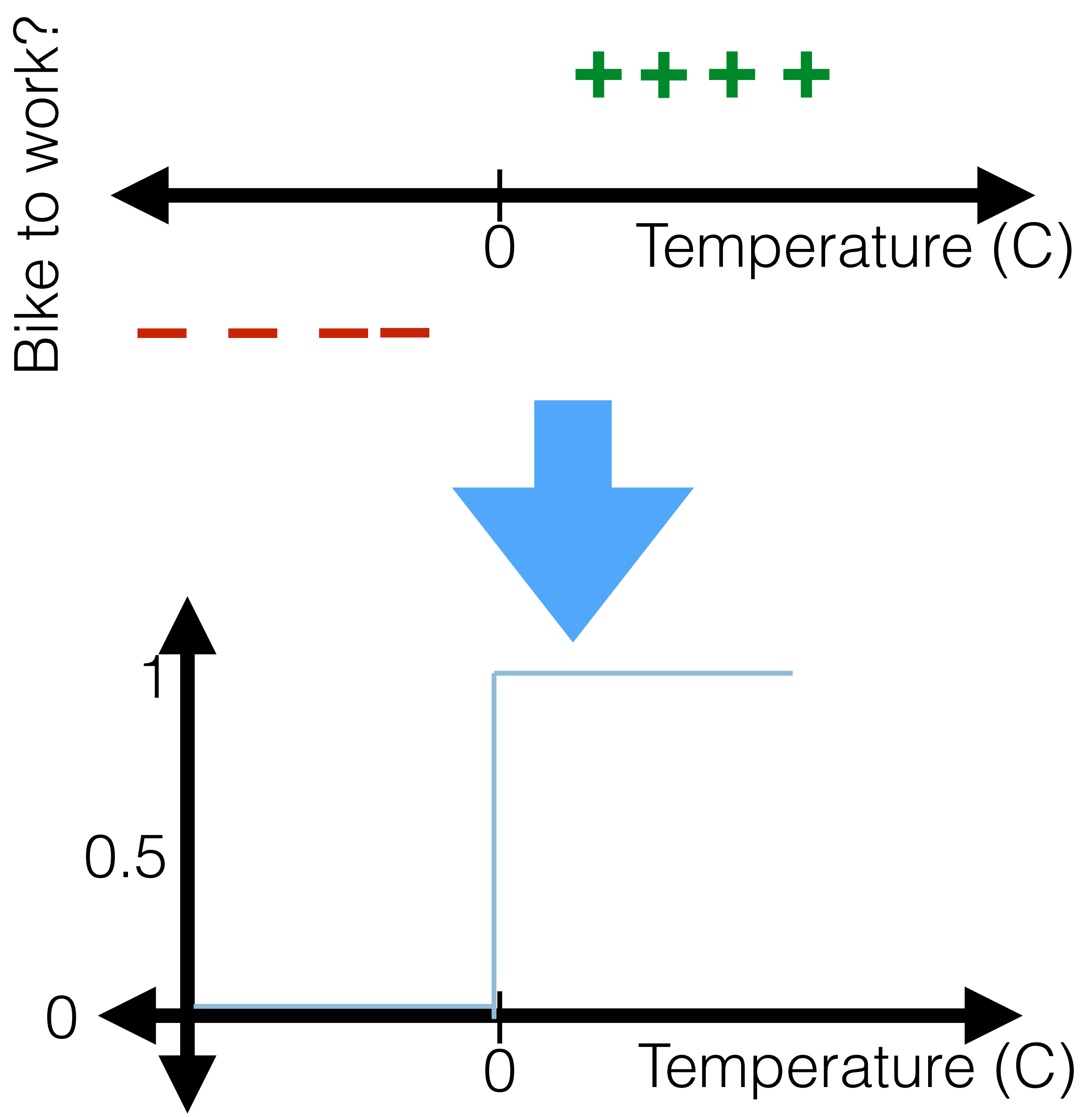
Gradient descent for logistic regression

- Can still have practical issues though!
- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`



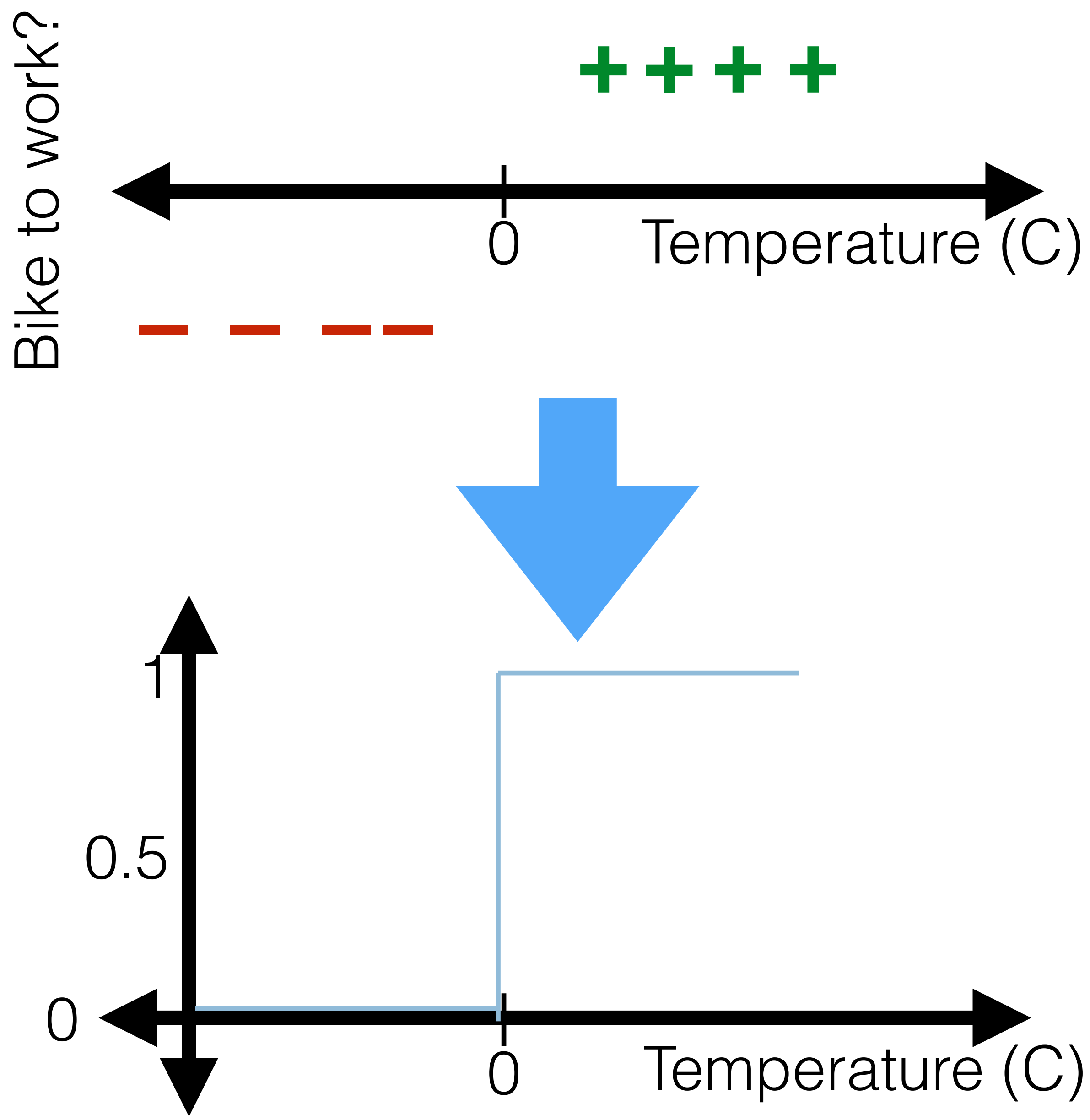
Gradient descent for logistic regression

- Can still have practical issues though!
- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`



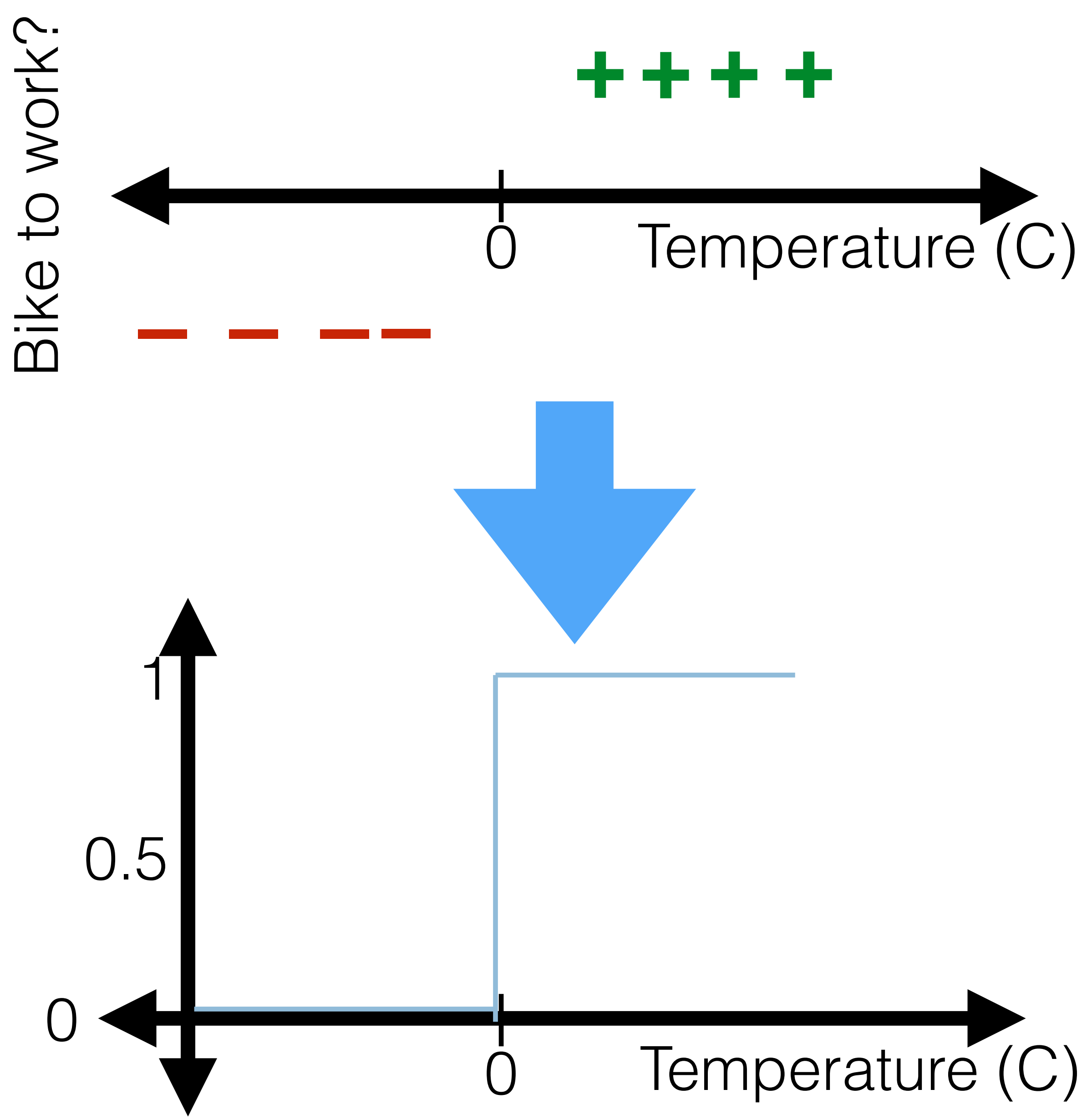
Gradient descent for logistic regression

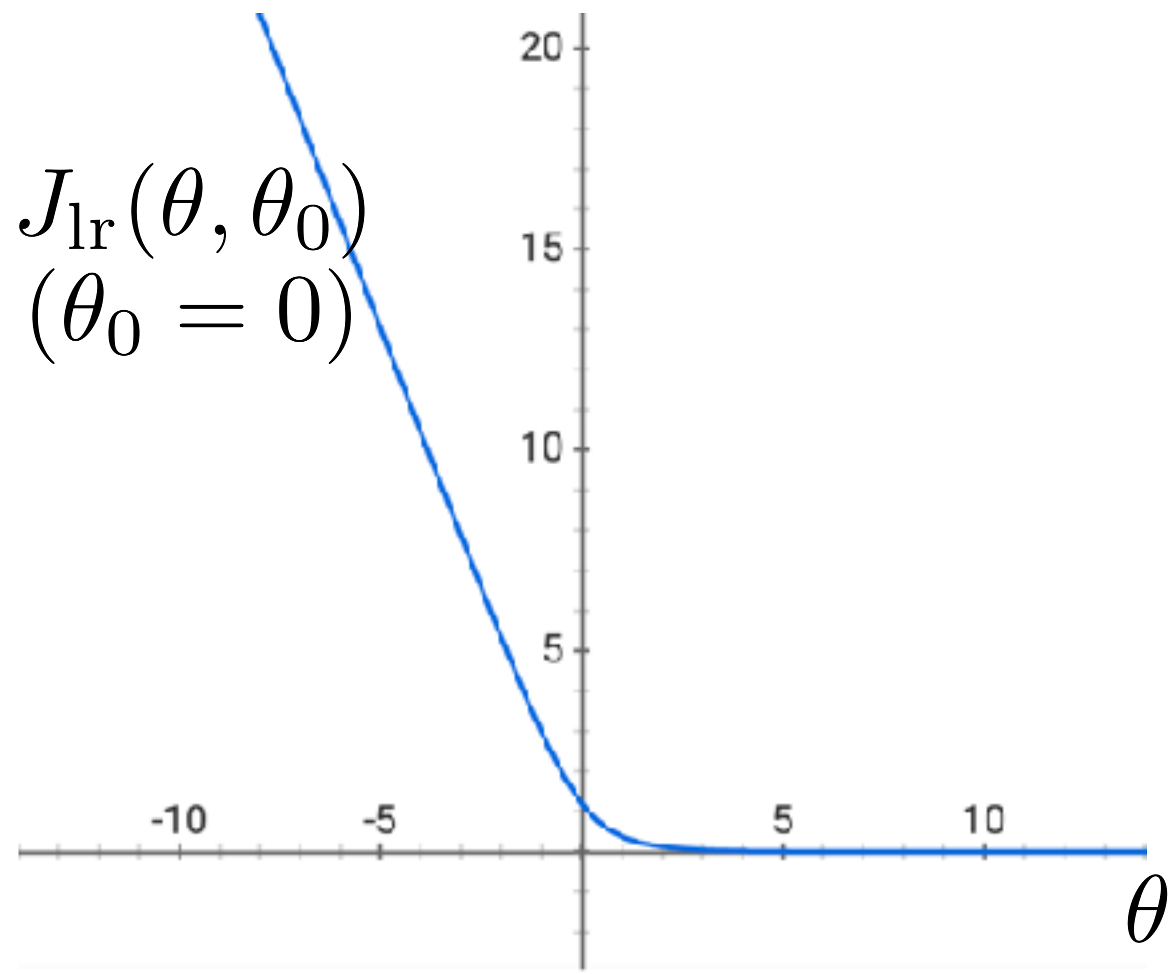
- Can still have practical issues though!
- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`



Gradient descent for logistic regression

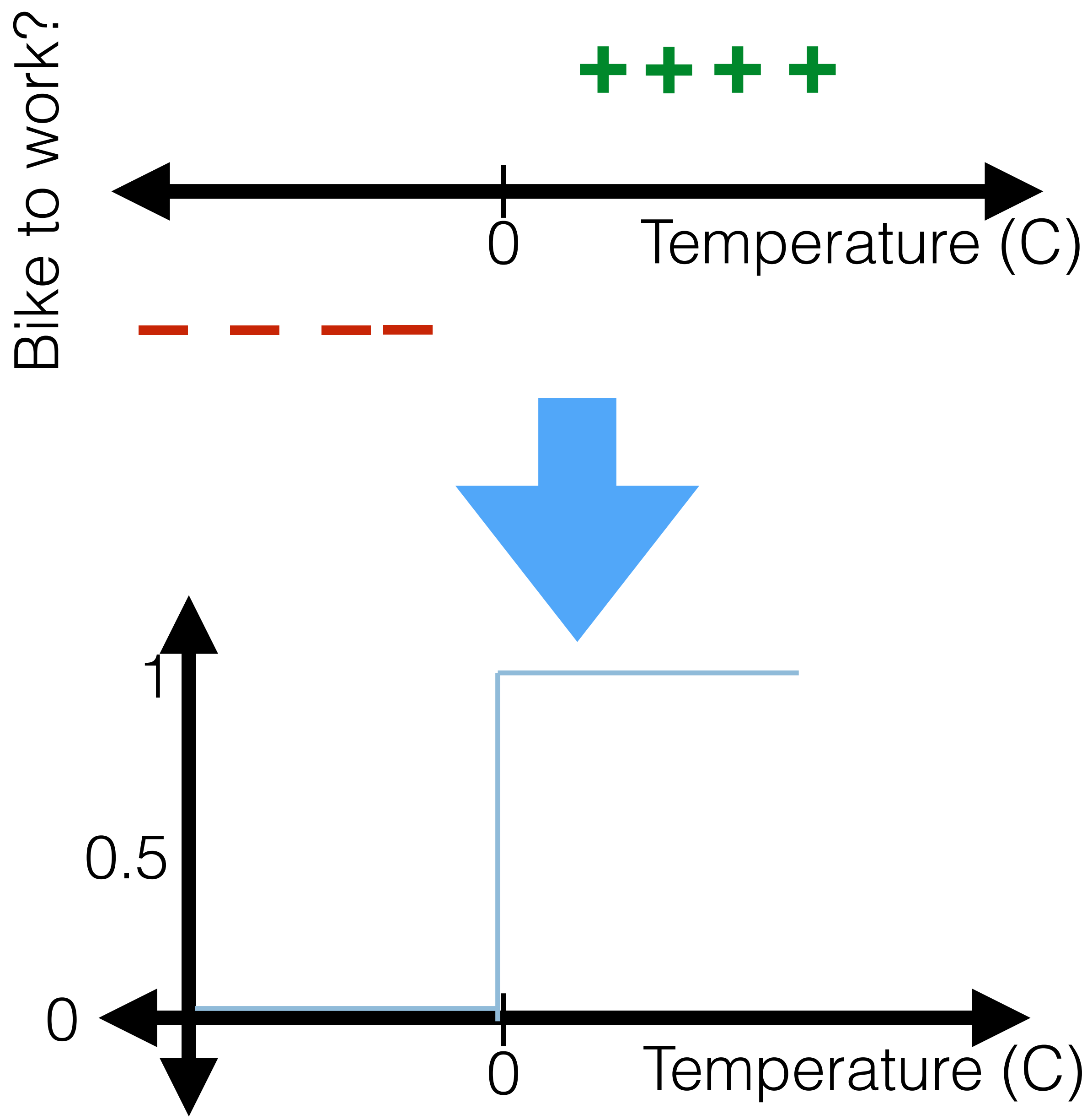
- Can still have practical issues though!
- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`

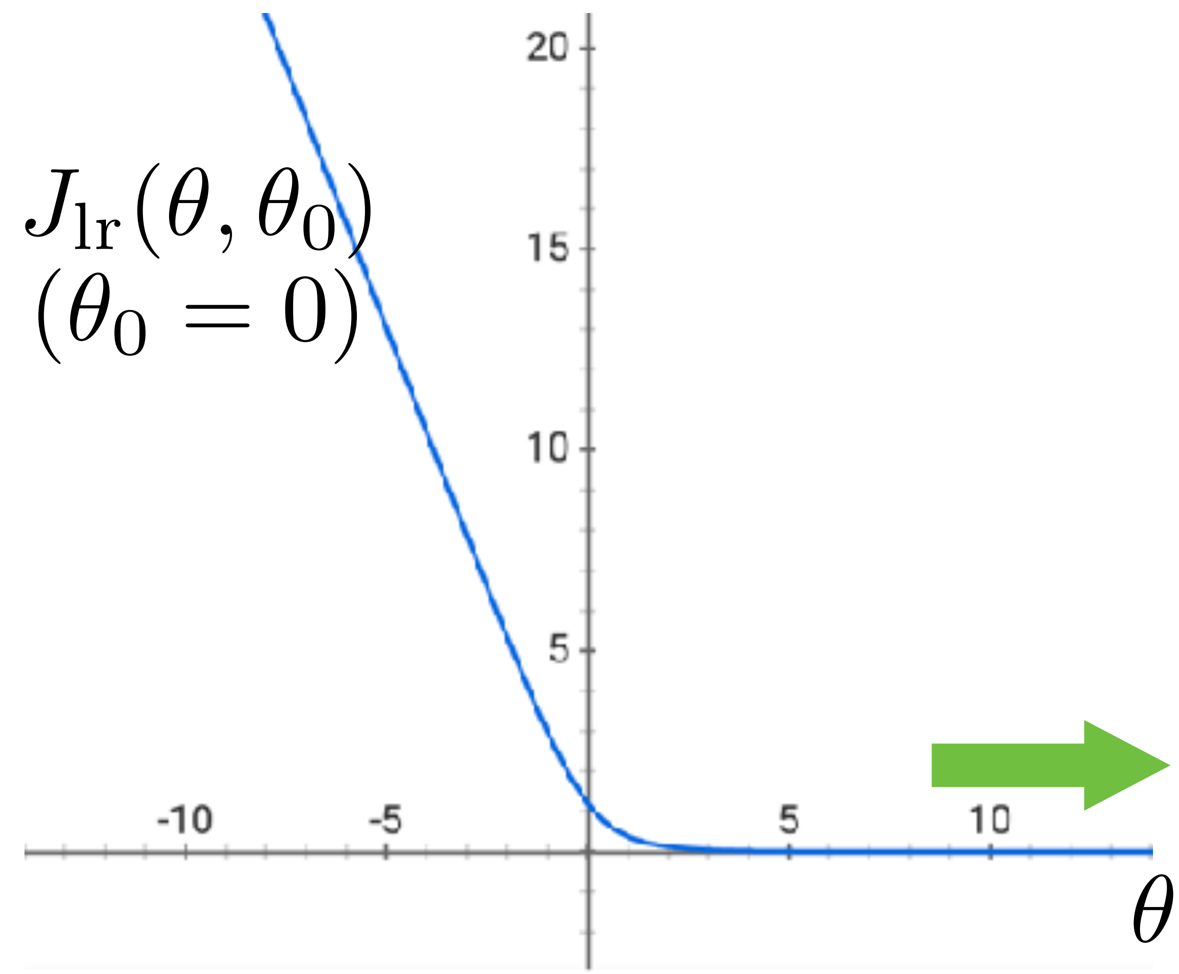




Gradient descent for logistic regression

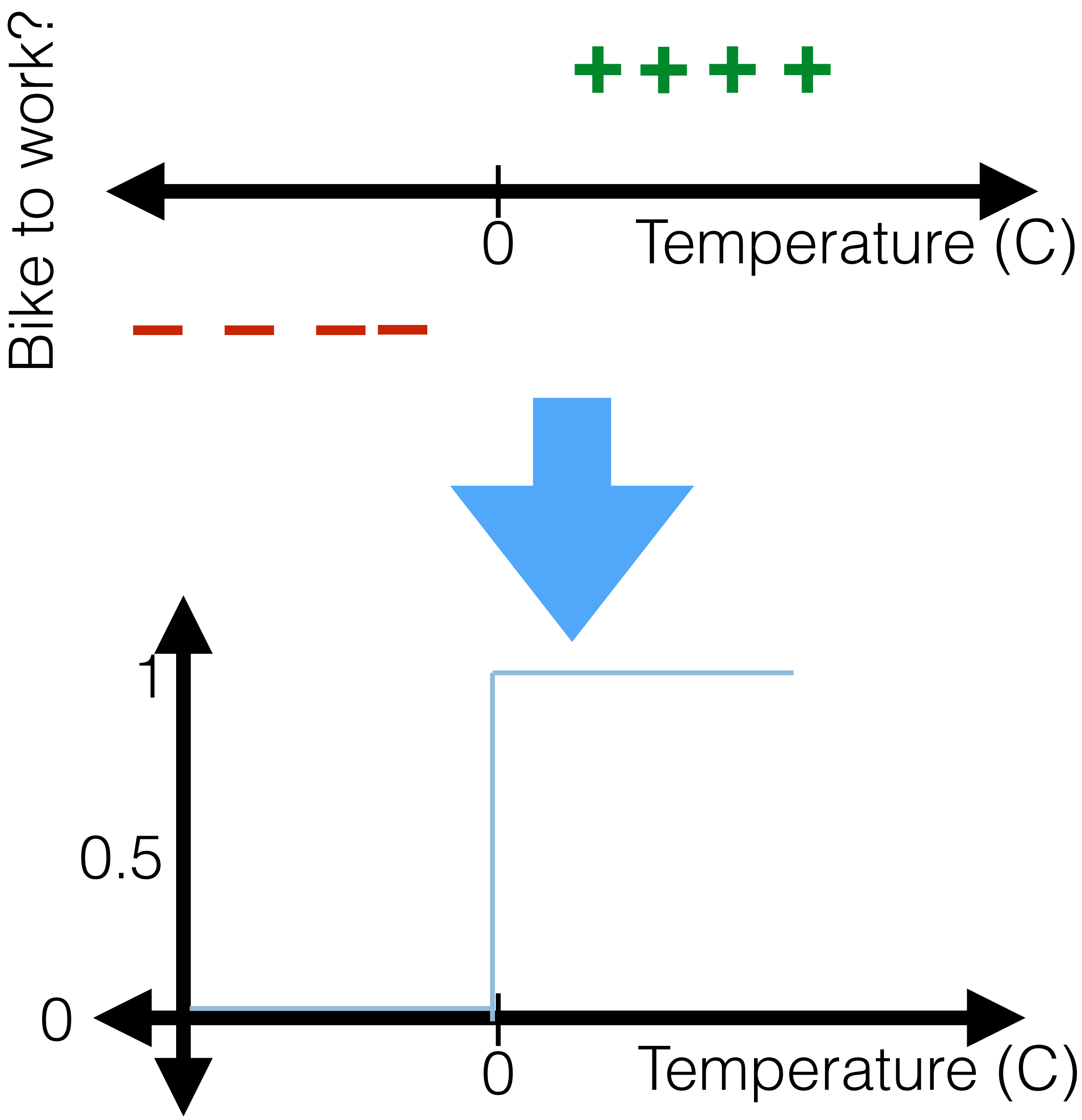
- Can still have practical issues though!
- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`

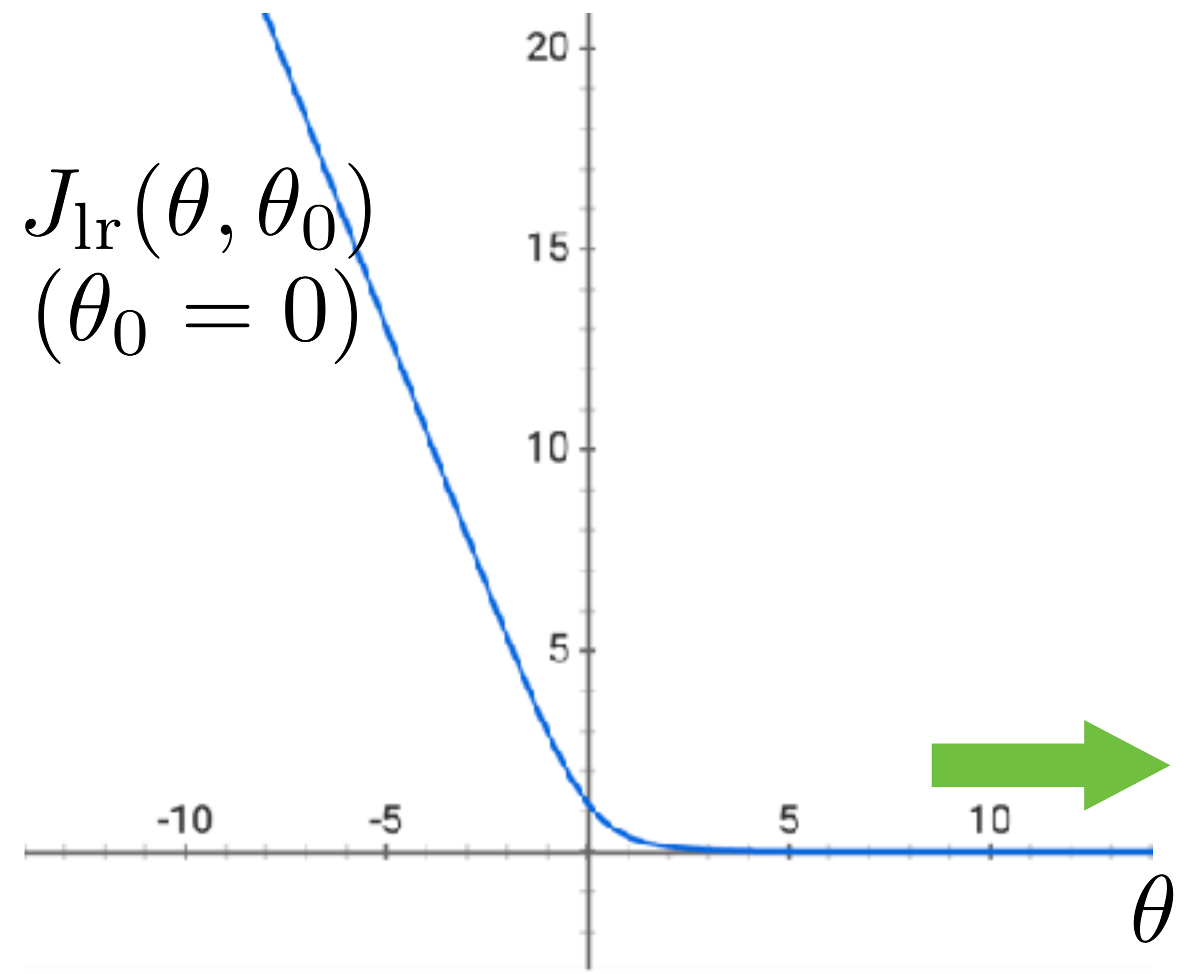




Gradient descent for logistic regression

- Can still have practical issues though!
- Run `Gradient-Descent(` $\Theta_{\text{init}}, \eta, J_{lr}, \nabla_\Theta J_{lr}, \epsilon$ `)`





no global optimum

Logistic regression loss revisited

$$\begin{aligned} J_{\text{lr}}(\Theta) &= J_{\text{lr}}(\theta, \theta_0) \\ &= \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)}) \end{aligned}$$

Logistic regression loss revisited

$$
\begin{aligned}
J_{\mathrm{lr}}(\Theta) &= J_{\mathrm{lr}}(\theta, \theta_0) \\
&= \frac{1}{n} \sum_{i=1}^{n} L_{\mathrm{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)}) + \lambda \|\theta\|^2 \qquad (\lambda \geq 0)
\end{aligned}
$$

Logistic regression loss revisited

$$\begin{aligned} J_{\text{lr}}(\Theta) &= J_{\text{lr}}(\theta, \theta_0) \\ &= \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)}) + \lambda\|\theta\|^2 \qquad (\lambda \geq 0) \end{aligned}$$

- A "regularizer" or "penalty" $R(\theta) = \lambda\|\theta\|^2$

Logistic regression loss revisited

$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0)$$
$$= \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)}) + \lambda\|\theta\|^2 \qquad (\lambda \geq 0)$$

- A “regularizer” or “penalty” $R(\theta) = \lambda\|\theta\|^2$
- Penalizes being overly certain

Logistic regression loss revisited

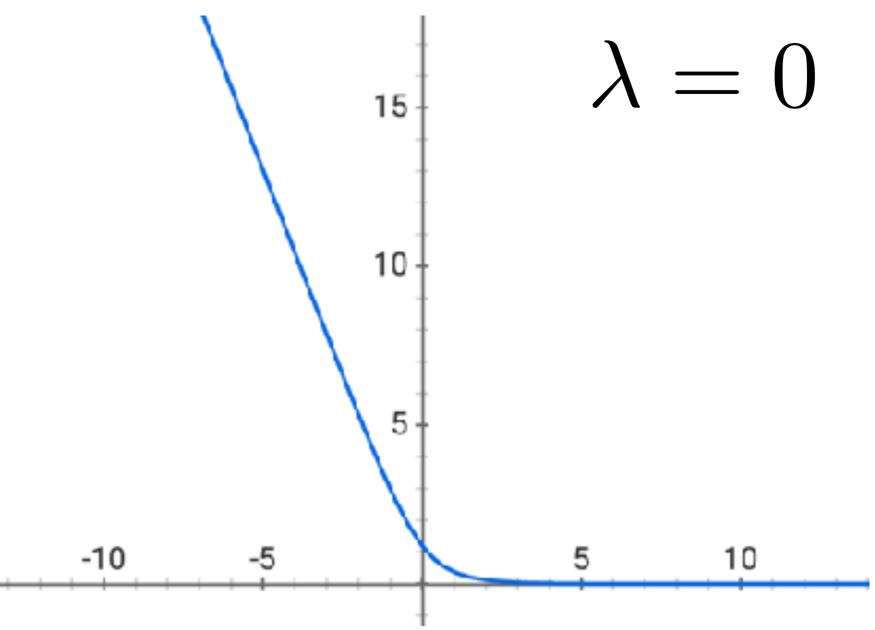
$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0)$$
$$= \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)}) + \lambda\|\theta\|^2 \qquad (\lambda \geq 0)$$

- A “regularizer” or “penalty” $R(\theta) = \lambda\|\theta\|^2$
- Penalizes being overly certain
- Objective is still differentiable & convex (gradient descent)

Logistic regression loss revisited

$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0)$$
$$= \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)}) + \lambda\|\theta\|^2 \qquad (\lambda \geq 0)$$

- A "regularizer" or "penalty" $R(\theta) = \lambda\|\theta\|^2$
- Penalizes being overly certain
- Objective is still differentiable & convex (gradient descent)

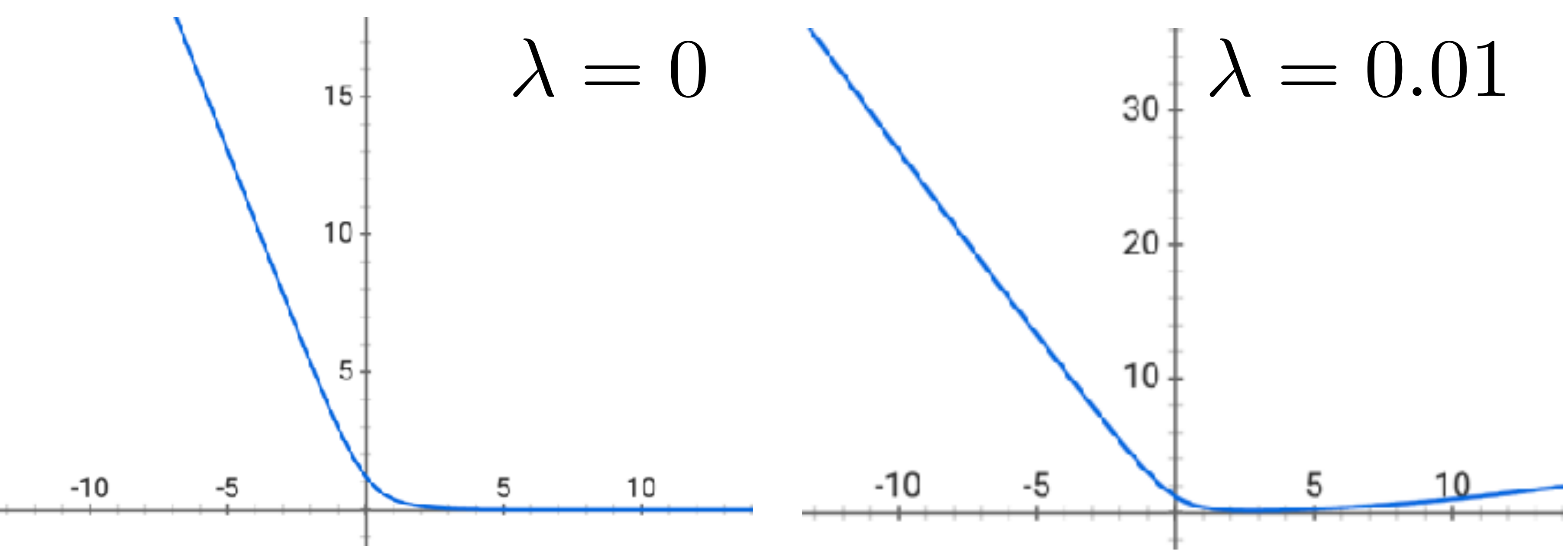


$\lambda = 0$

Logistic regression loss revisited

$$
\begin{aligned}
J_{\text{lr}}(\Theta) &= J_{\text{lr}}(\theta, \theta_0) \\
&= \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)}) + \lambda\|\theta\|^2 \qquad (\lambda \geq 0)
\end{aligned}
$$

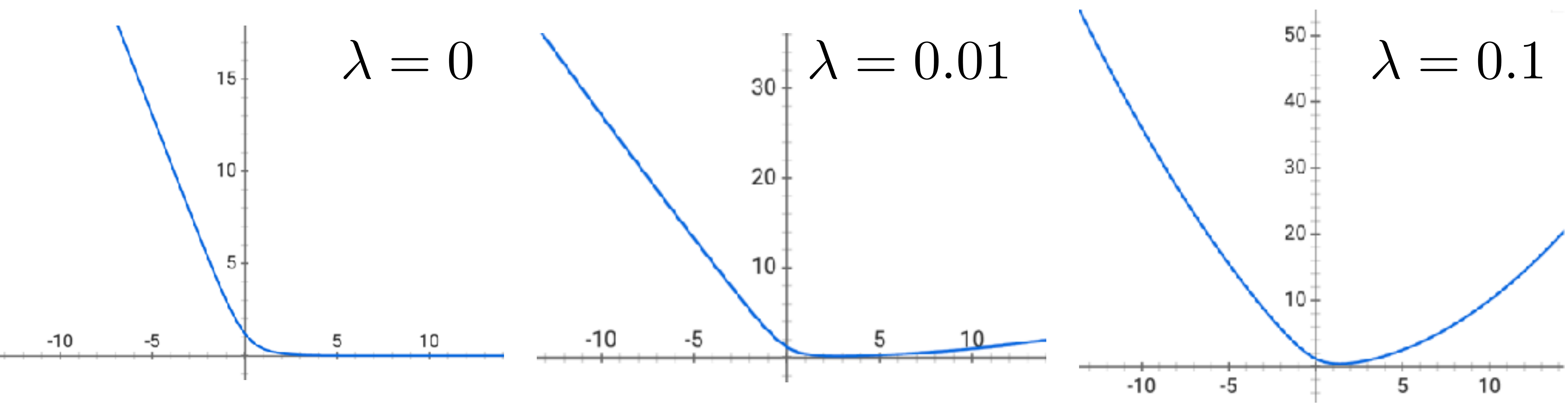
- A "regularizer" or "penalty" $R(\theta) = \lambda\|\theta\|^2$
- Penalizes being overly certain
- Objective is still differentiable & convex (gradient descent)



Logistic regression loss revisited

$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)}) + \lambda\|\theta\|^2 \qquad (\lambda \geq 0)$$

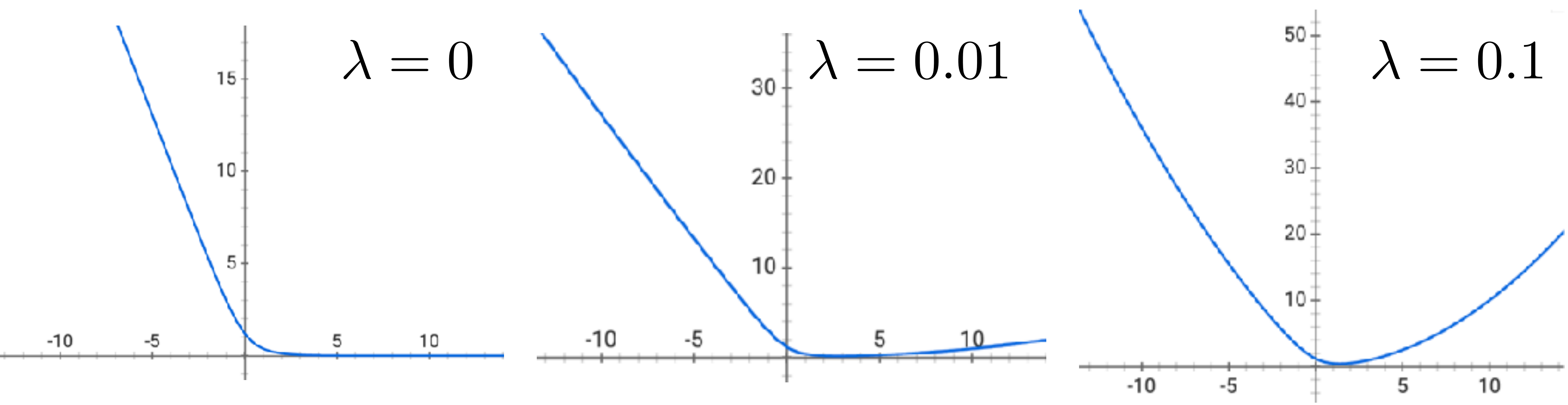
- A "regularizer" or "penalty" $R(\theta) = \lambda\|\theta\|^2$
- Penalizes being overly certain
- Objective is still differentiable & convex (gradient descent)



Logistic regression loss revisited

$$J_{\text{lr}}(\Theta) = J_{\text{lr}}(\theta, \theta_0)$$
$$= \frac{1}{n}\sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)}) + \lambda\|\theta\|^2 \qquad (\lambda \geq 0)$$

- A "regularizer" or "penalty" $R(\theta) = \lambda\|\theta\|^2$
- Penalizes being overly certain
- Objective is still differentiable & convex (gradient descent)



- How to choose hyperparameter? One option: consider a handful of possible values and compare via CV