Lecture 12: Conformal Prediction

CIS 7000: Trustworthy Machine Learning

Spring 2024

Course Project

- **Goals**
  - Exposure to research ideas in trustworthy machine learning
  - Understand some aspect of trustworthy machine learning more deeply

- Course project is a major component of this class

Course Project

- We urge you to start thinking about the course project now
- The project can be done individually or in a group of two
- You are welcome to set up a meeting with one of us to discuss project ideas at any time

Possible Project Categories

- Implementing and rigorously evaluating a technique discussed in class, in a bit more in depth than homework
- Review a specific paper, implement the described technique, and evaluate it empirically
- Review two or three papers with a common theme, and summarize their techniques with relative strengths and weaknesses
- Intersection of your current research and the course theme

Tentative Project Timeline

- **Monday, March 25:** Decide on team and project topic
- **Monday, April 1:** Finalize a concrete project with approval from us
- **Monday, April 22:** Submit project report
  - 4-5 pages is typical length
- **April 22, 24, 29, May 1:** In-class project presentations
  - 15 min talk + 5 min Q&A

Homework 2

- Covers distribution shift and uncertainty quantification
  - Written homework focused on theoretical understanding
- Plan to release by Friday (March 1)
- Due Monday, March 11

Calibrated Prediction

- Predict a **probability** $\vec{p}(x)_y$ for each label y
- Probabilities are correct if conditioned on $\hat{p}(x) = p$, the accuracy is p

Why Calibration?

- Imagine you are making a decision with utility $U(y^*)$ (for $y^* \in \{0,1\}$)

- **Claim:** If making decisions purely based on $\hat{p}(x)$, you can act as if $\hat{p}(x)$ is the true probability of $y^* = 1$

- **"Proof":**
  - Among all x for which $\hat{p}(x) = p$, exactly p fraction of them satisfy $y^* = 1$
  - Thus, you obtain the payoff that you expected among these values of x

Shortcomings of Calibration

- Unintuitive/hard to reason about probabilities
  - Both for humans and for algorithms

- Structured prediction (e.g., sentences, object detection, etc.)
  - Probabilities of complex outputs quickly become small
  - Probabilities of different portions of the output can be highly correlated

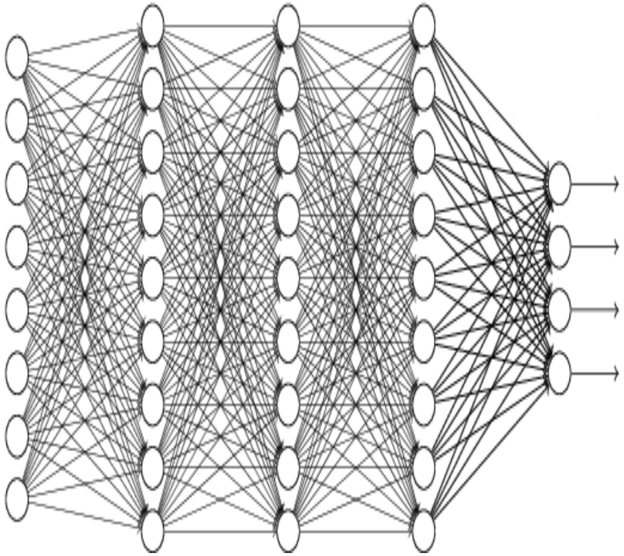
- **Conformal prediction**
  - Represents of uncertainty using **prediction sets**, which can be more intuitive
  - Also easier to reason about algorithmically

Agenda

- Conformal prediction problem
- Conformal prediction algorithm
- Correctness proof

Conformal Prediction

Image x



DNN f

"toilet seat"

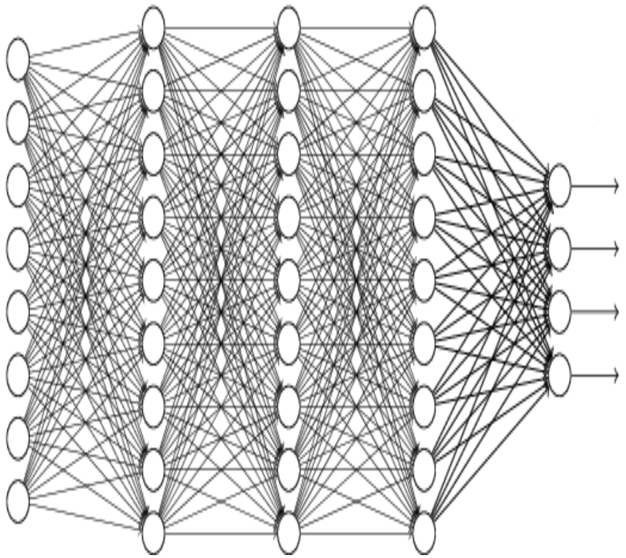
Prediction
$\hat{y} = \max_{y} f(y \mid x)$

Conformal Prediction

Incorrect!

(Ground truth label: $y^* =$ “plunger”)

Image x



DNN f

“toilet seat”

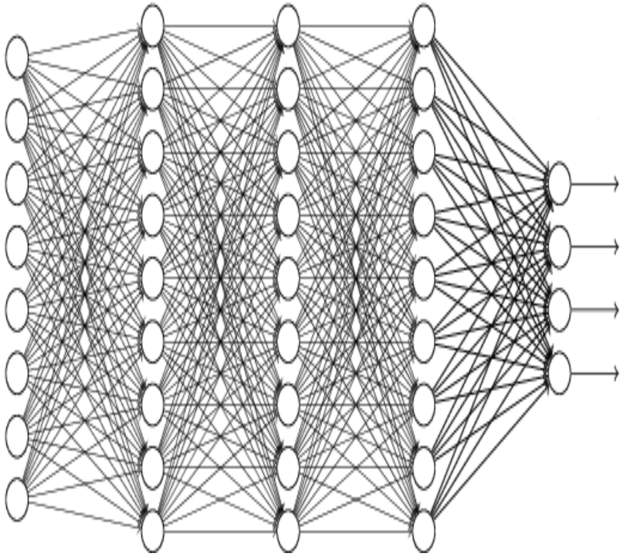
Prediction

$$\hat{y} = \max_{y} f(y \mid x)$$

Conformal Prediction

Incorrect!
(Ground truth label: $y^* =$ "plunger")

Image x



DNN f

"toilet seat"

Prediction
$$\hat{y} = \max_{y} f(y \mid x)$$

Idea: Modify DNN f to predict **sets of labels**

Conformal Prediction

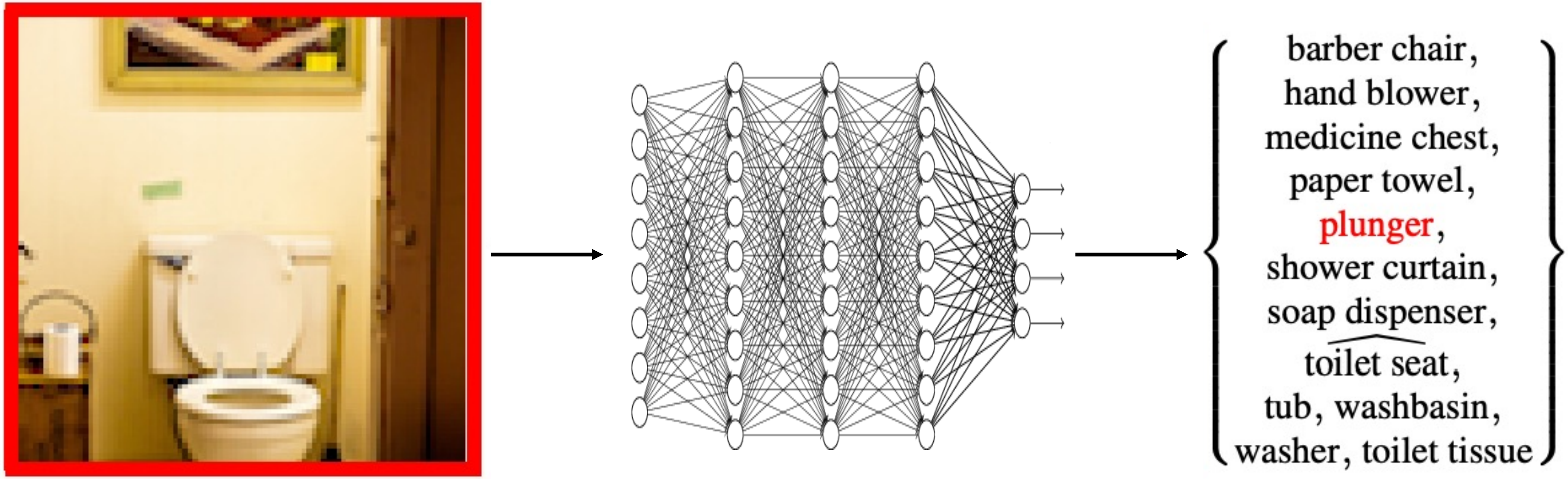
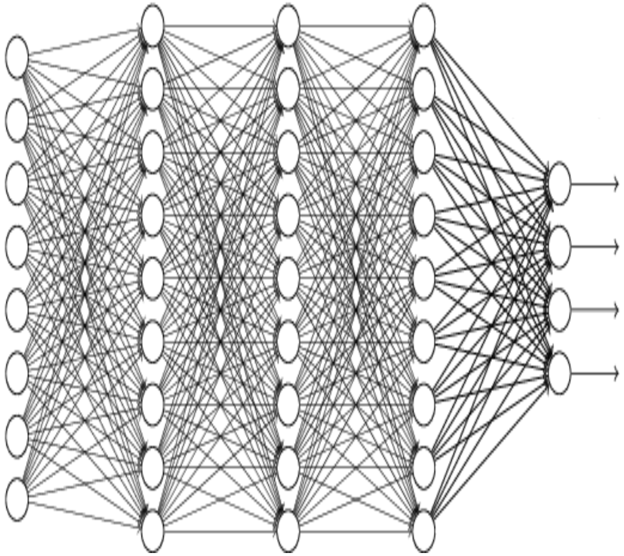


Image x Prediction Set $\tilde{f}$ Output $Y = \tilde{f}(x)$

Idea: Modify DNN f to predict **sets of labels**

Conformal Prediction

Now, we have $y^* \in \tilde{f}(x)$ (**coverage**)



{ barber chair, hand blower, medicine chest, paper towel, plunger, shower curtain, soap dispenser, toilet seat, tub, washbasin, washer, toilet tissue }

Image x

Prediction Set $\tilde{f}$

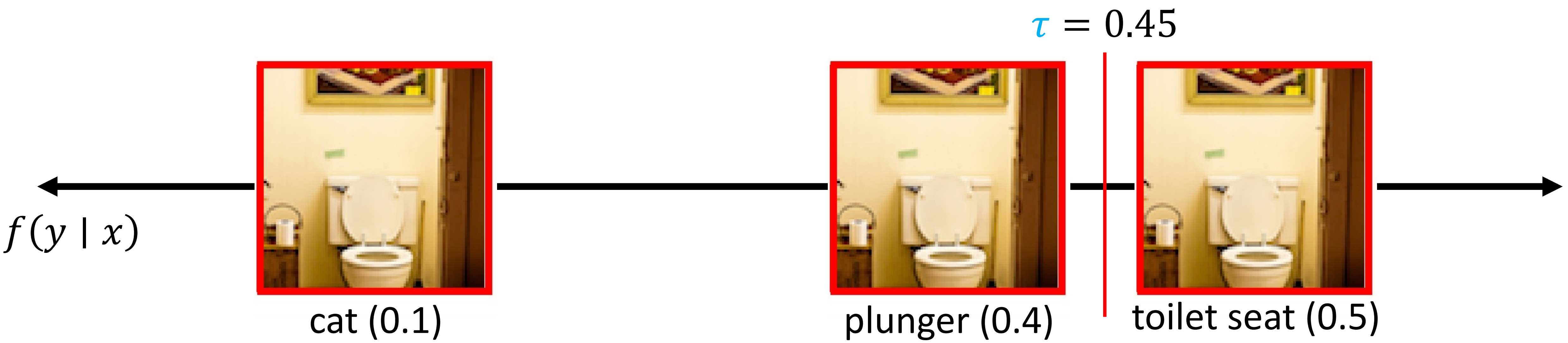
Output $Y = \tilde{f}(x)$

Idea: Modify DNN f to predict **sets of labels**

Conformal Prediction Problem

- **Parametric model family of prediction sets**
  - We construct prediction sets based on an **existing** DNN $f(y \mid x)$
  - Consider prediction sets that are **level sets** of f:

$$\tilde{f}_{\tau}(x) = \{\, y \mid f(y \mid x) \geq \tau \,\}$$



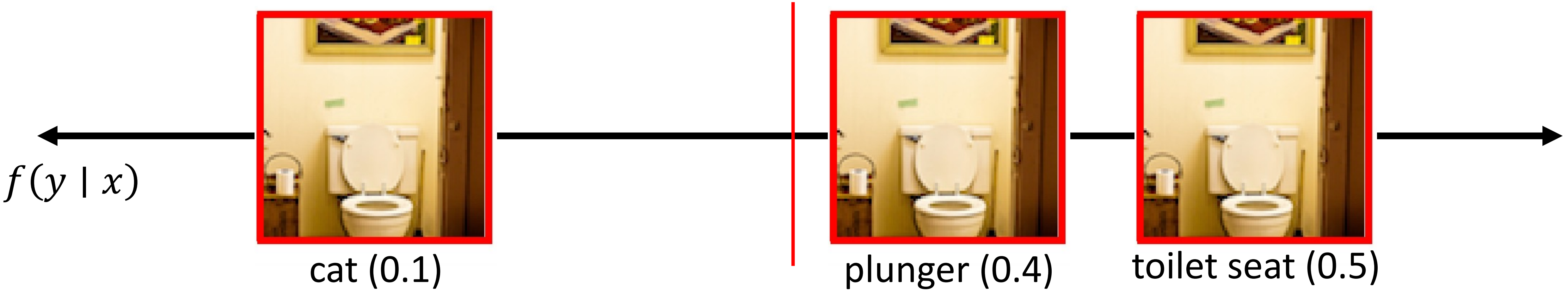
$$\tilde{f}_{\tau}(x) = \{\text{toilet seat}\}$$

Conformal Prediction Problem

- **Parametric model family of prediction sets**
  - We construct prediction sets based on an **existing** DNN $f(y \mid x)$
  - Consider prediction sets that are **level sets** of f:

$$\tilde{f}_{\tau}(x) = \{ y \mid f(y \mid x) \geq \tau \}$$

$$\tau = 0.35$$

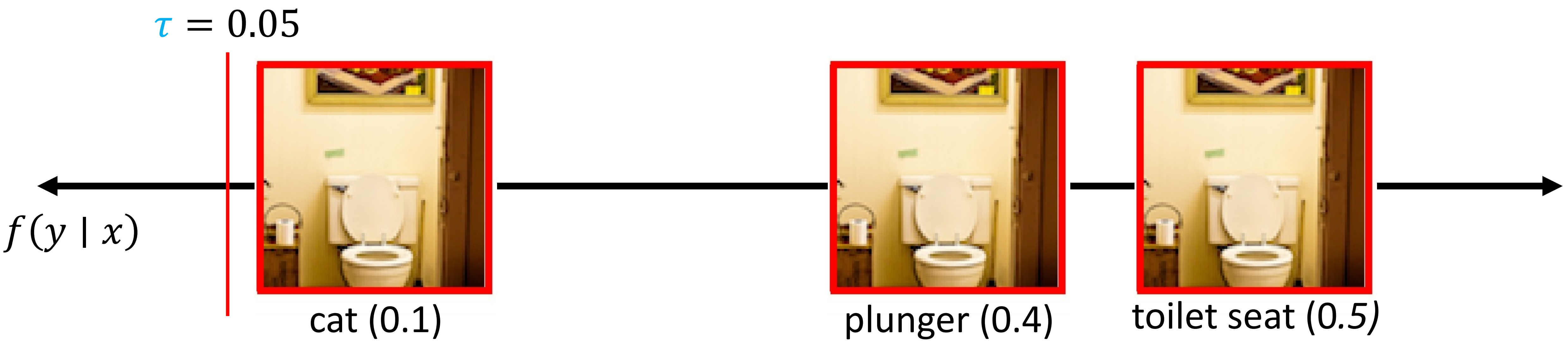


$$\tilde{f}_{\tau}(x) = \{\text{toilet seat, plunger}\}$$

Conformal Prediction Problem

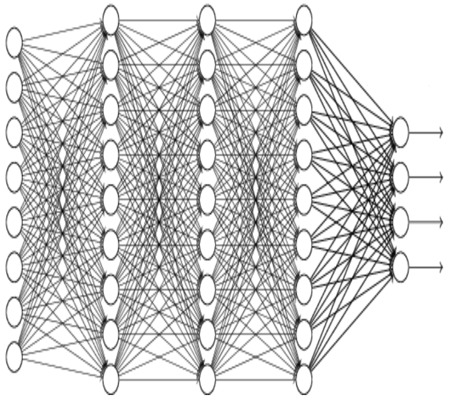
- **Parametric model family of prediction sets**
  - We construct prediction sets based on an **existing** DNN $f(y \mid x)$
  - Consider prediction sets that are **level sets** of f:

$$\tilde{f}_{\tau}(x) = \{ y \mid f(y \mid x) \geq \tau \}$$



$$\tilde{f}_{\tau}(x) = \{\text{toilet seat, plunger, cat}\}$$

Conformal Prediction Problem

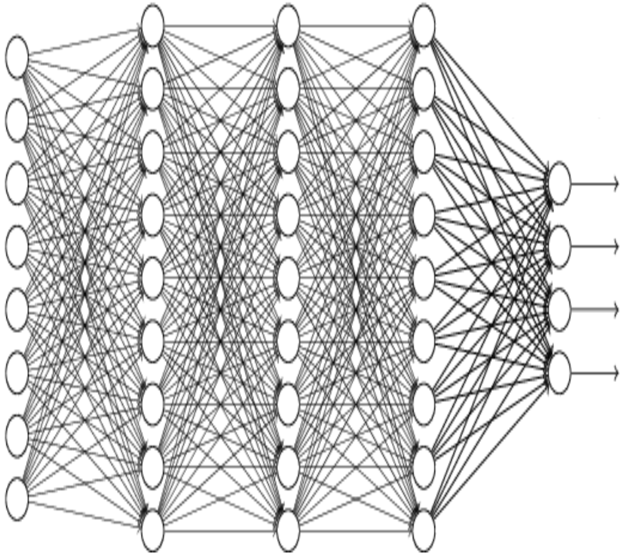


DNN f

chihuahua iguana shopping basket king penguin church

Validation Dataset $Z_{\text{val}} = \{(x, y^*)\}$

Our Algorithm



Prediction Set $\tilde{f}$

PAC Prediction Sets

- **IID assumption (standard in learning theory)**
  - Assume an "underlying distribution" $p(x, y^*)$
  - Validation examples are $Z_{\text{val}} \sim_{\text{iid}} p$

- Given τ, we say prediction set $\tilde{f}_\tau$ is ϵ *approximately correct (AC)* if

$$\Pr_{p(x,y^*)}\left[y^* \in \tilde{f}_\tau(x)\right] \geq 1 - \epsilon$$

- I.e., $\tilde{f}_\tau(x)$ contains true label y^* with probability $\geq 1 - \epsilon$ over $p(x, y^*)$

PAC Prediction Sets

- Consider a **learning algorithm** $\hat{\tau}(Z_{\mathrm{val}})$
  - **Input:** Validation dataset Z_{val} (and implicitly, DNN f)
  - **Output:** PAC prediction set $\tilde{f}_{\hat{\tau}(Z_{\mathrm{val}})}$
- We say $\hat{\tau}$ is (ϵ, δ) *probably approximately correct (PAC)* if

$$\Pr_{p(Z_{\mathrm{val}})}\left[\tilde{f}_{\hat{\tau}(Z_{\mathrm{val}})} \text{ is } \epsilon \text{ AC}\right] \geq 1 - \delta$$

- I.e., $\tilde{f}_{\hat{\tau}(Z_{\mathrm{val}})}$ is ϵ AC with probability $\geq 1 - \delta$ over $p(Z_{\mathrm{val}})$

PAC Prediction Set Problem

- Devise a prediction set algorithm $\hat{\tau}(Z_{\text{val}})$ satisfying the PAC property
- Can always take $\hat{\tau}(Z_{\text{val}}) = -\infty$ to satisfy PAC guarantee!
- **Goal:** Construct "smallest" PAC prediction sets

Aside: Types of Conformal Prediction

- **Traditional conformal prediction**
  - Guarantees $\Pr_{p(Z_{\text{val}}),p(x,y^*)}\left[y^* \in \tilde{f}_{\hat{\tau}(Z_{\text{val}})}(x)\right] \geq 1 - \alpha$
  - Combines ϵ and δ, called a **marginal guarantee**
  - Different algorithm and proof based on exchangeability argument

- **Training-conditional conformal prediction**
  - Same as PAC guarantee
  - Much more closely aligned with learning theory

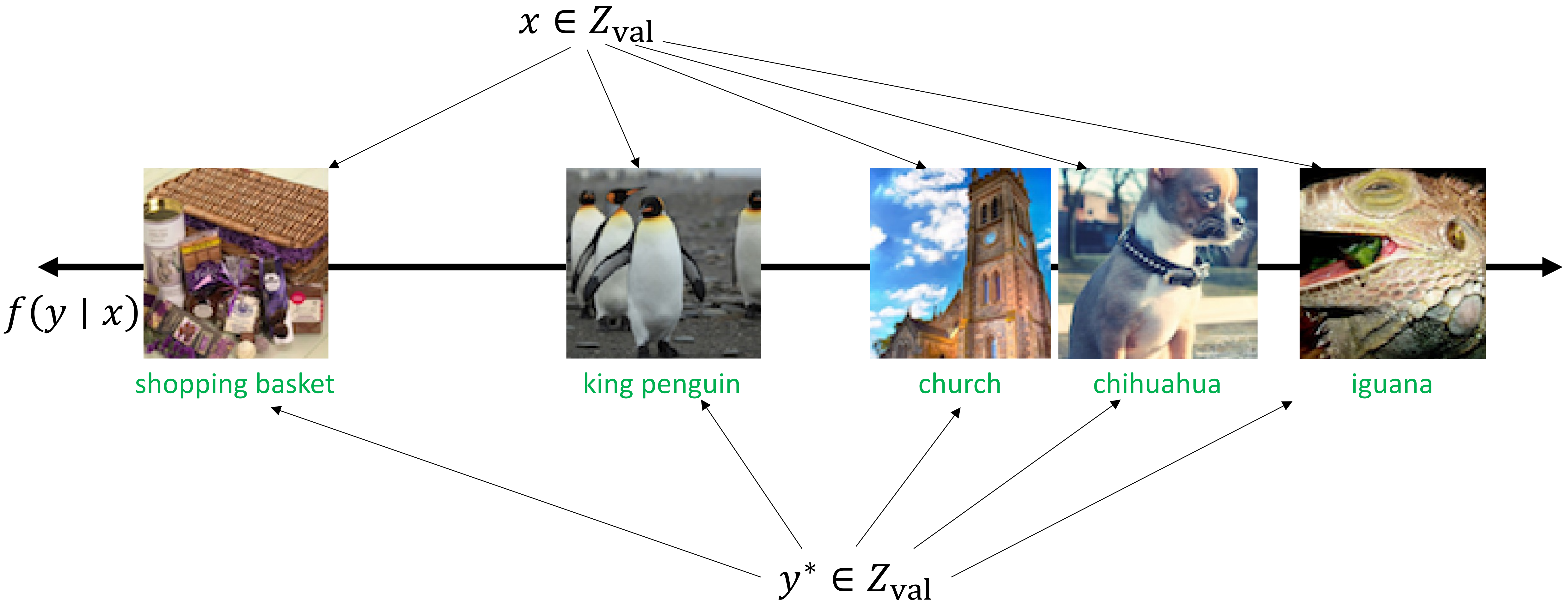
Agenda

- Conformal prediction problem
- Conformal prediction algorithm
- Correctness proof

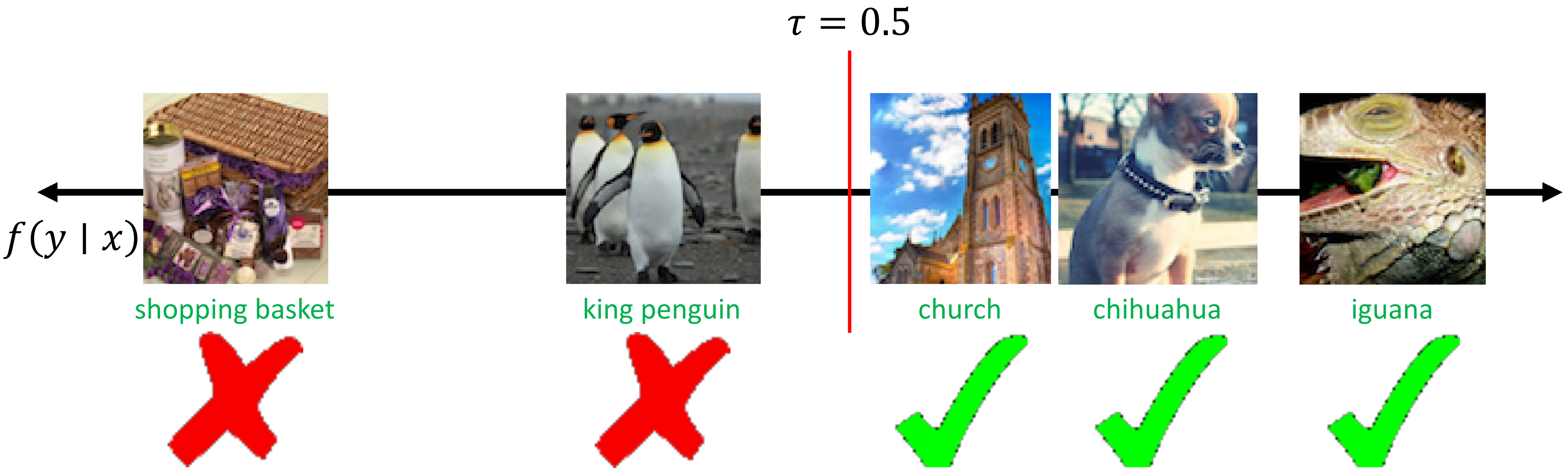
PAC Prediction Set Algorithm

- **Step 1:** Reduce problem to supervised learning problem
  - Binary classification
  - 1D covariate space
  - 1D parameter space
- **Step 2:** Devise a statistical learning algorithm to solve this problem

Step 1: Reduction to Binary Classification



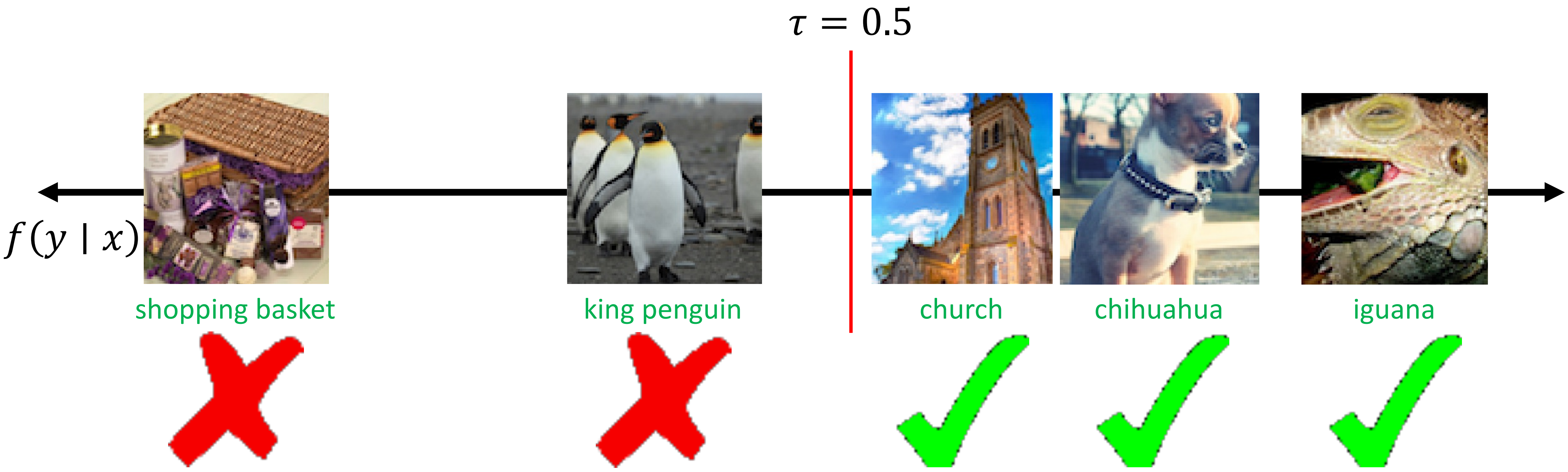
Step 1: Reduction to Binary Classification



$y^* \notin \tilde{f}_\tau(x)$ $\quad$ $y^* \in \tilde{f}_\tau(x)$

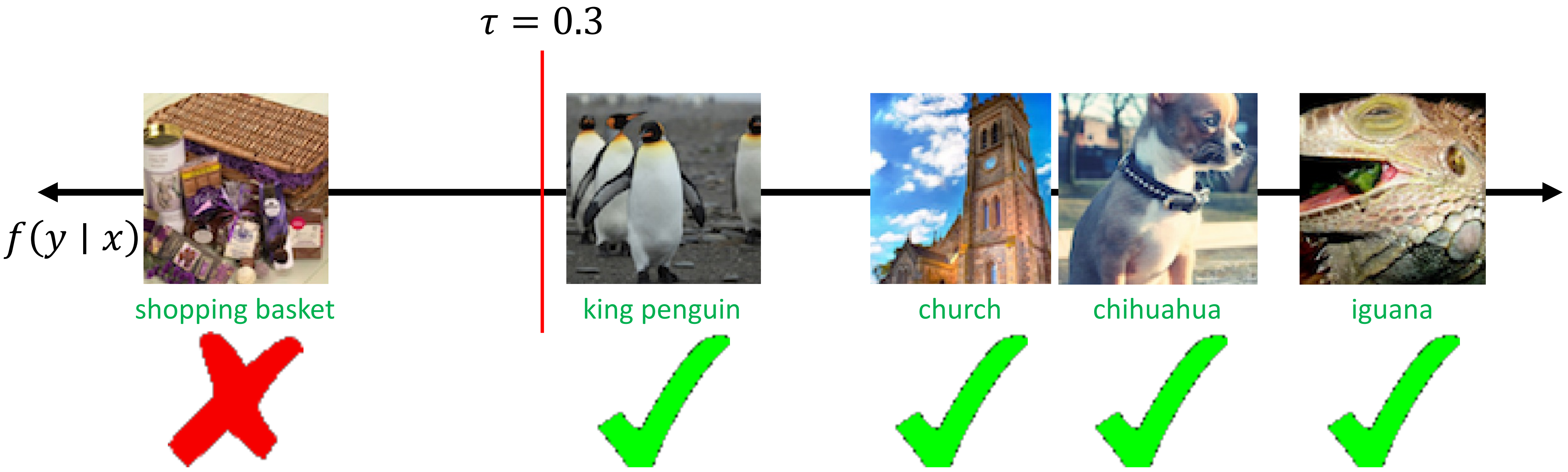
$$\tilde{f}_\tau(x) = \{\, y \in Y \mid f(y \mid x) \geq \tau \,\}$$

Step 1: Reduction to Binary Classification



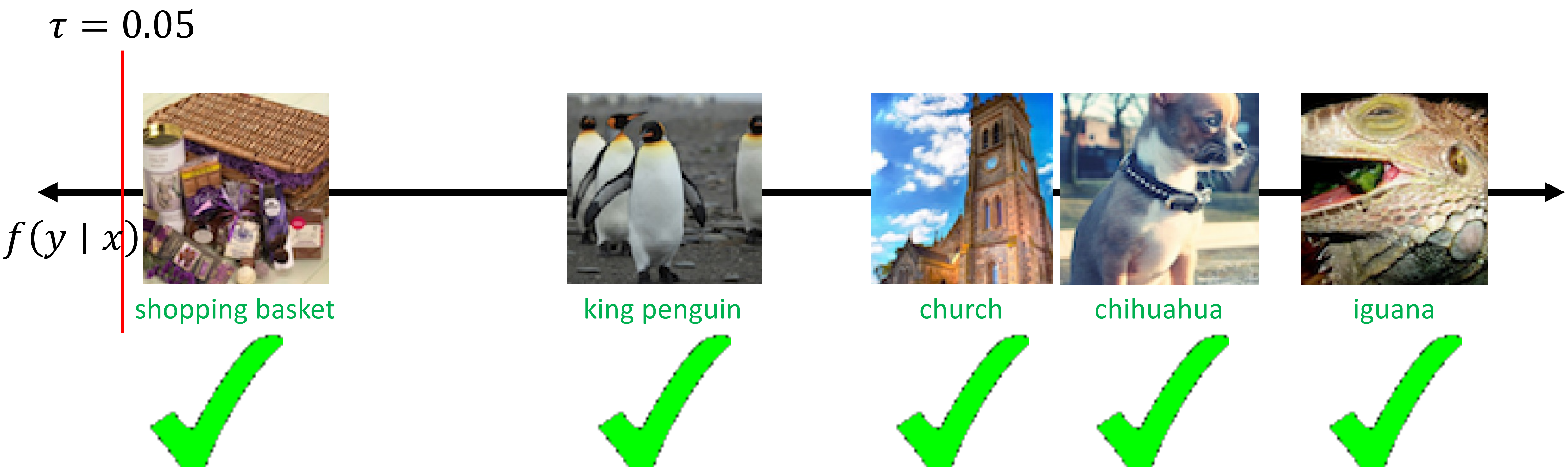
$$\mathrm{err}(\tau = 0.5; Z_{\mathrm{val}}) = 2$$

Step 1: Reduction to Binary Classification



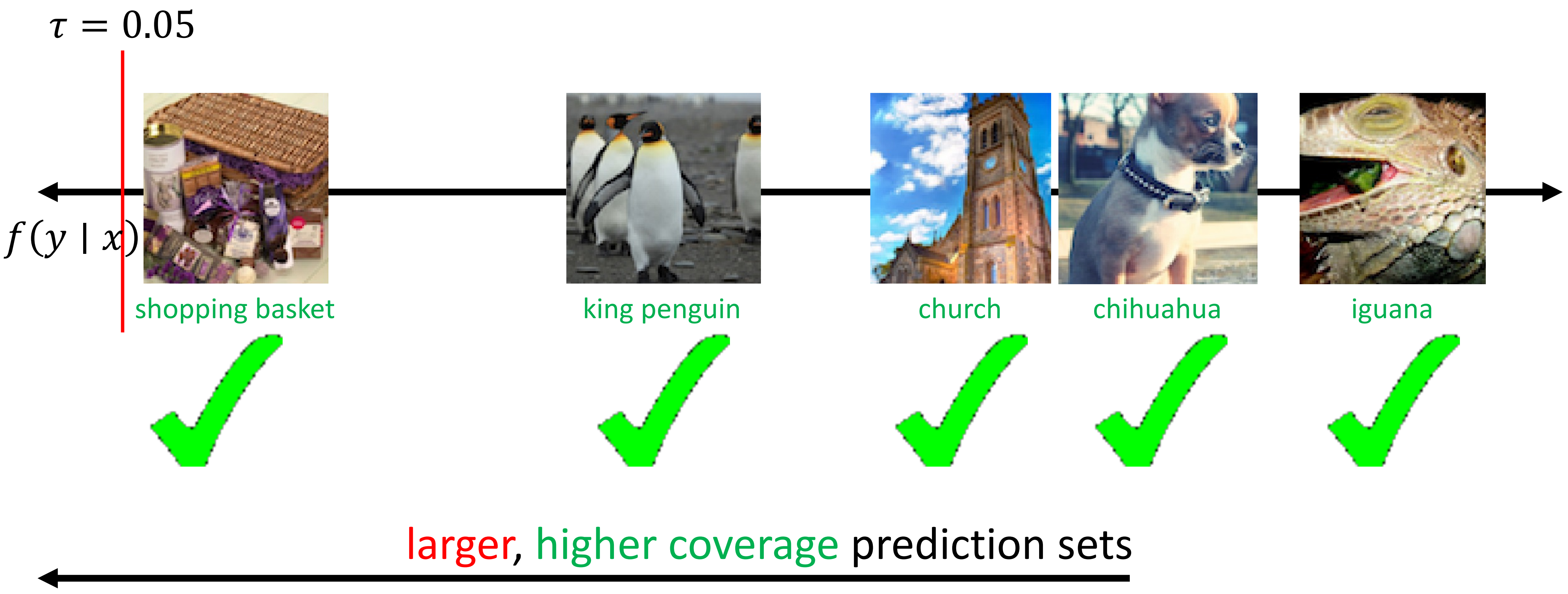
$$\mathrm{err}(\tau = 0.3; Z_{\mathrm{val}}) = 1$$

Step 1: Reduction to Binary Classification



$$\mathrm{err}(\tau = 0.05; Z_{\mathrm{val}}) = 0$$

Step 1: Reduction to Binary Classification



Step 2: Statistical Learning Algorithm

minimize prediction set size

- 1D, so optimize using binary search

$$\hat{\tau}(Z_{\text{cal}}) = \begin{array}{l} \arg\max_{\tau} \ \tau \\ \text{subj. to} \ \ \text{err}(\tau; Z_{\text{val}}) \leq k \end{array}$$

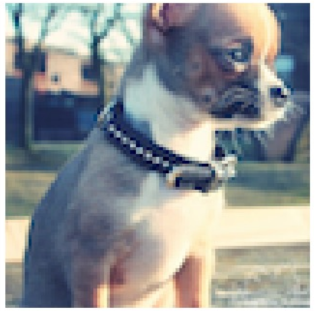
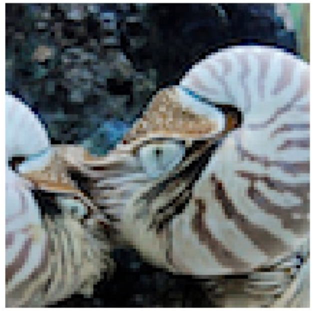
subject to **constraint** on prediction set error

- k chosen to satisfy the (ϵ, δ) PAC property

Theoretical Guarantees

Theorem: $\tilde{f}_{\hat{\tau}(Z_{\text{val}})}$ is an (ϵ, δ) PAC prediction set

Examples on ImageNet

$1 \leq \lvert C(x)\rvert < 5$	$5 \leq \lvert C(x)\rvert < 10$	$10 \leq \lvert C(x)\rvert < 20$
 {king penguin}	 {Chihuahua, toy terrier, Italian greyhound, Boston bull, miniature pinscher}	 {banded gecko, common iguana, American chameleon, whiptail, agama, frilled lizard, alligator lizard, green lizard, African chameleon, Komodo dragon}
 {shopping basket}	 {English springer, Welsh springer spaniel, collie, boxer, Saint Bernard, Leonberg}	 {altar, analog clock, bell cote, castle, church, cinema, dome, monastery, palace, vault, wall clock}
 {chambered nautilus}	 {face powder, hamper, lotion, packet, shopping basket}	 {barber chair, hand blower, medicine chest, paper towel, plunger, shower curtain, soap dispenser, toilet seat, tub, washbasin, washer, toilet tissue}

Examples on Object Detection

Examples on Code Generation

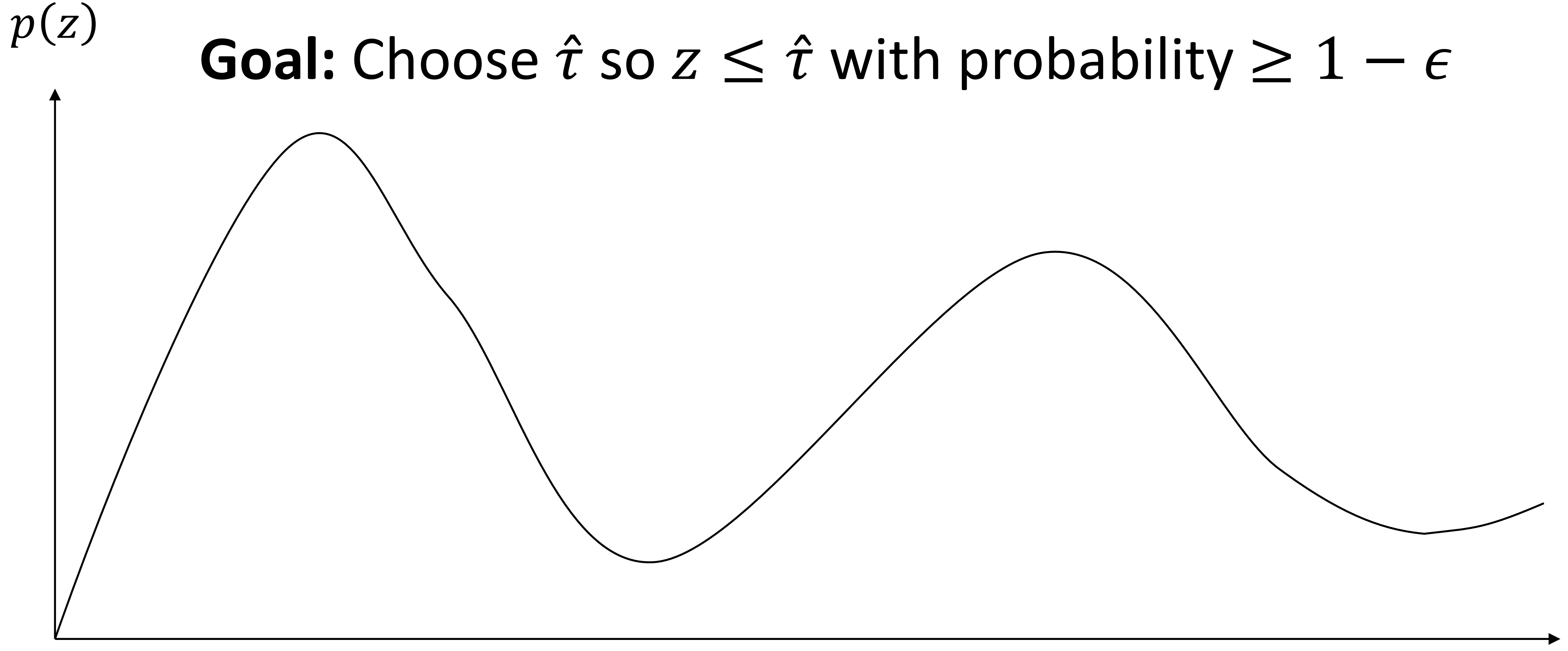
```
SELECT COUNT(*) FROM countries AS t1
JOIN car_makers as t2 on t1.countryid = t2.country
WHERE t1.countryname = "usa";
```

Agenda

- Conformal prediction problem
- Conformal prediction algorithm
- Correctness proof

Proof Sketch

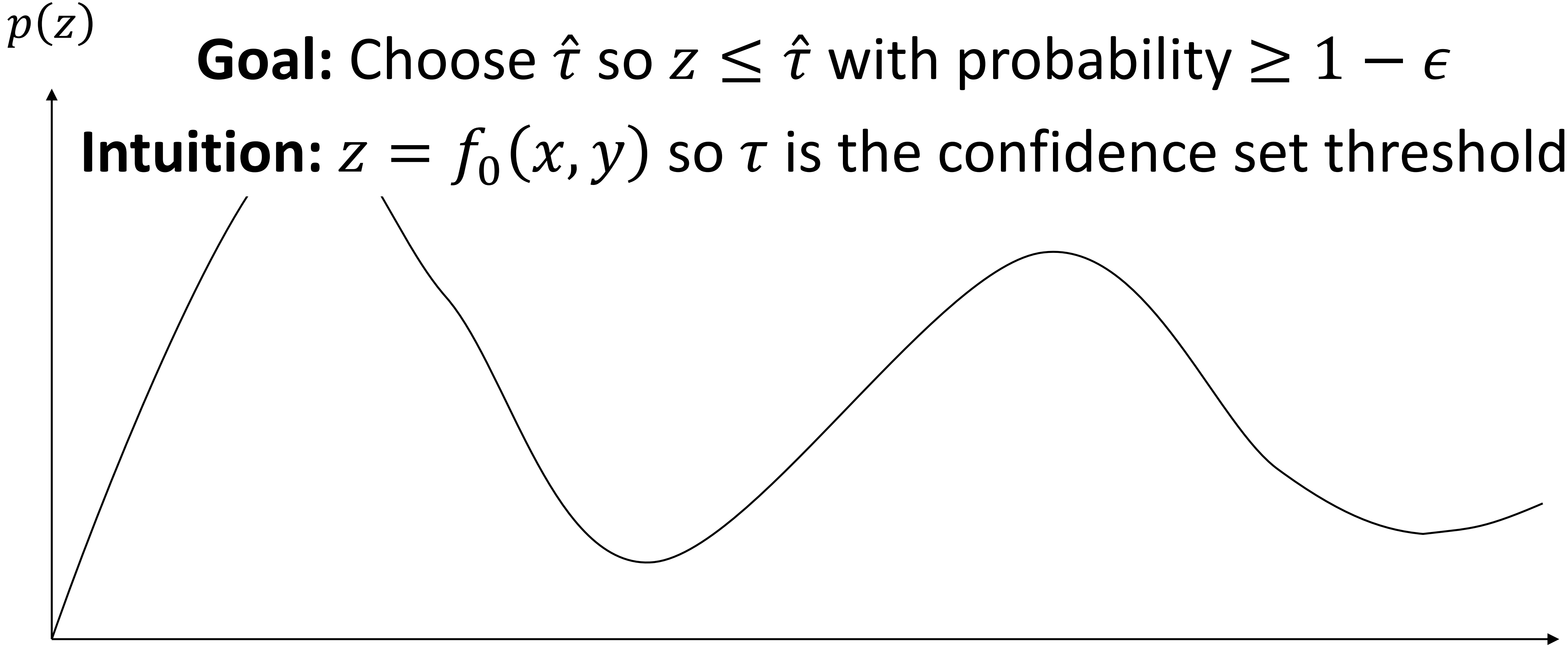
Goal: Choose $\hat{\tau}$ so $z \leq \hat{\tau}$ with probability $\geq 1 - \epsilon$



Proof Sketch

Goal: Choose $\hat{\tau}$ so $z \leq \hat{\tau}$ with probability $\geq 1 - \epsilon$

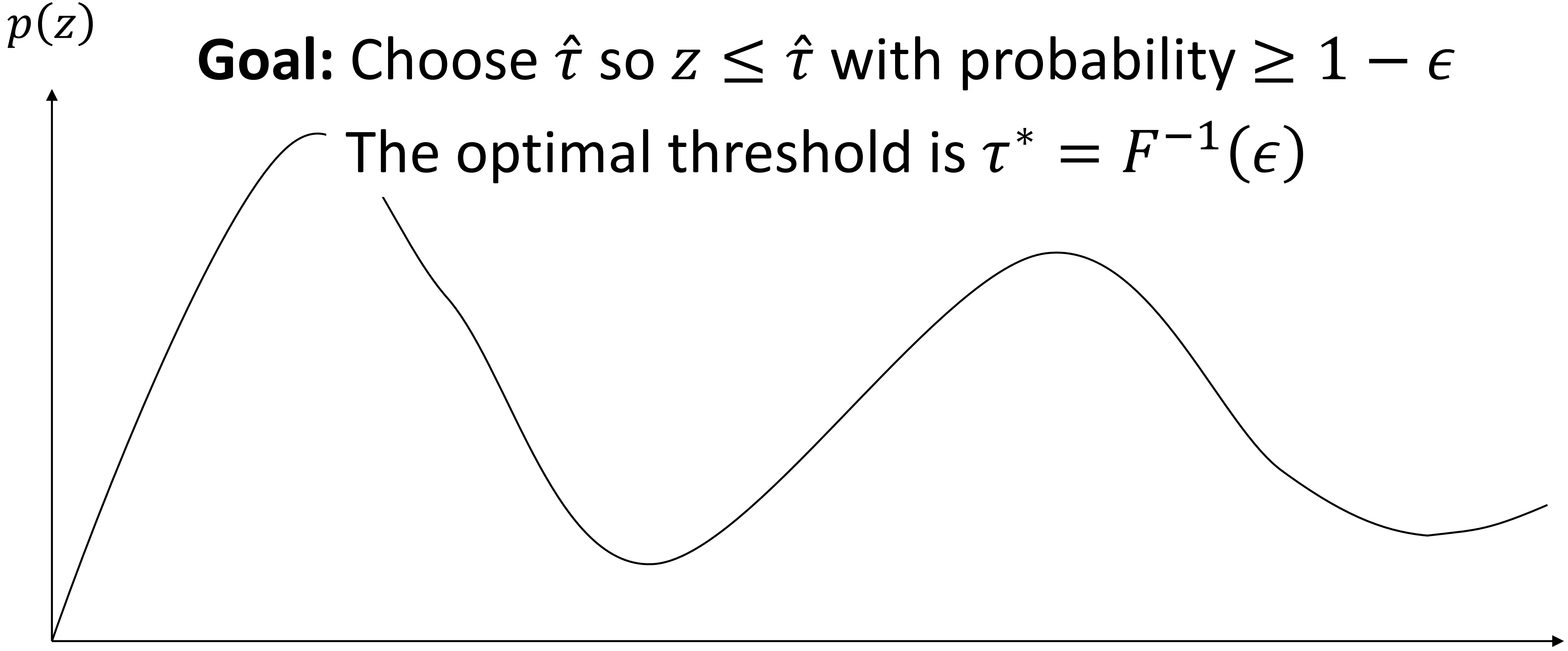
Intuition: $z = f_0(x, y)$ so τ is the confidence set threshold



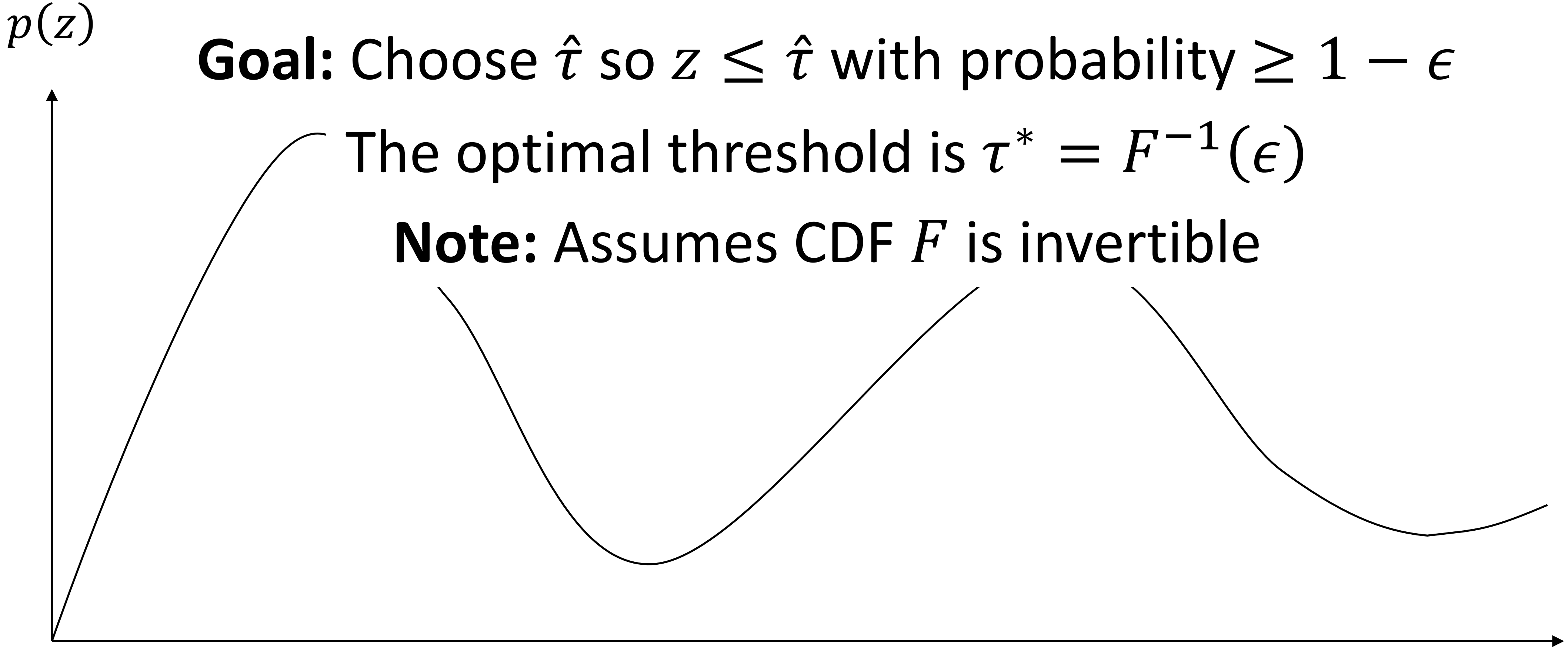
Proof Sketch

Goal: Choose $\hat{\tau}$ so $z \leq \hat{\tau}$ with probability $\geq 1 - \epsilon$

The optimal threshold is $\tau^* = F^{-1}(\epsilon)$



Proof Sketch



Goal: Choose $\hat{\tau}$ so $z \leq \hat{\tau}$ with probability $\geq 1 - \epsilon$

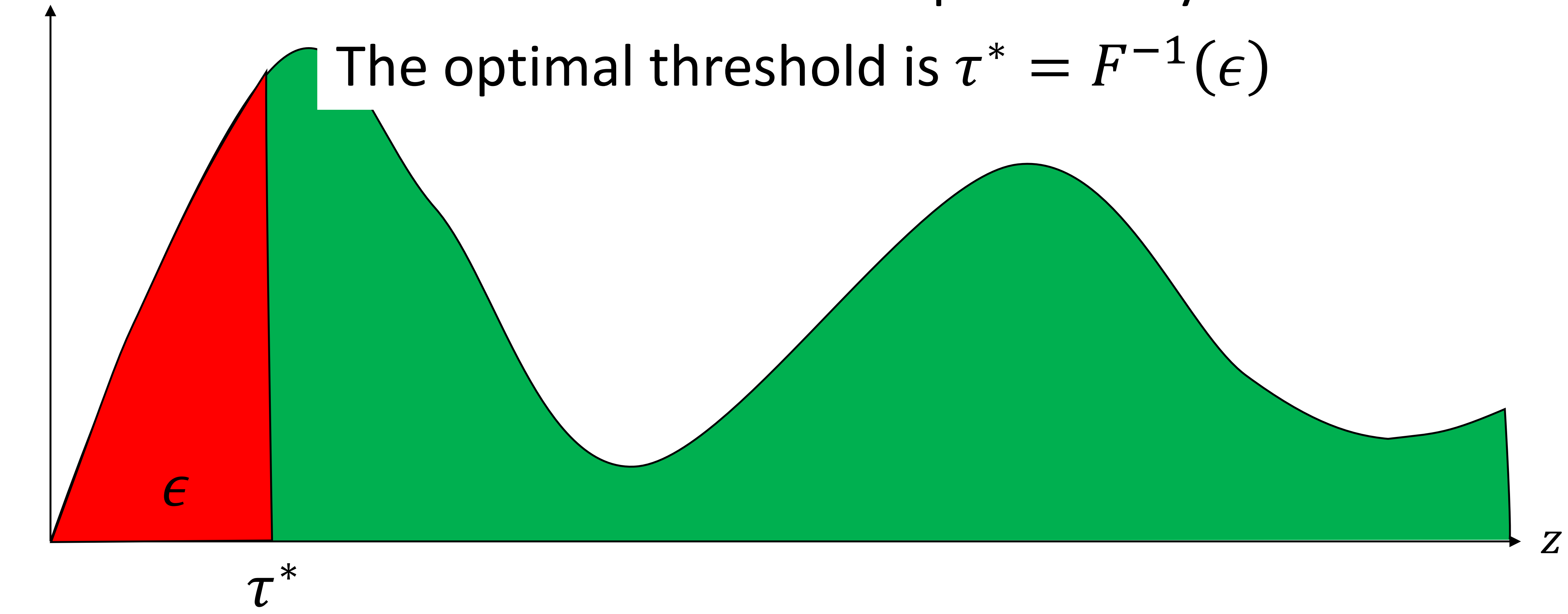
The optimal threshold is $\tau^* = F^{-1}(\epsilon)$

Note: Assumes CDF F is invertible

Proof Sketch

Goal: Choose $\hat{\tau}$ so $z \leq \hat{\tau}$ with probability $\geq 1 - \epsilon$

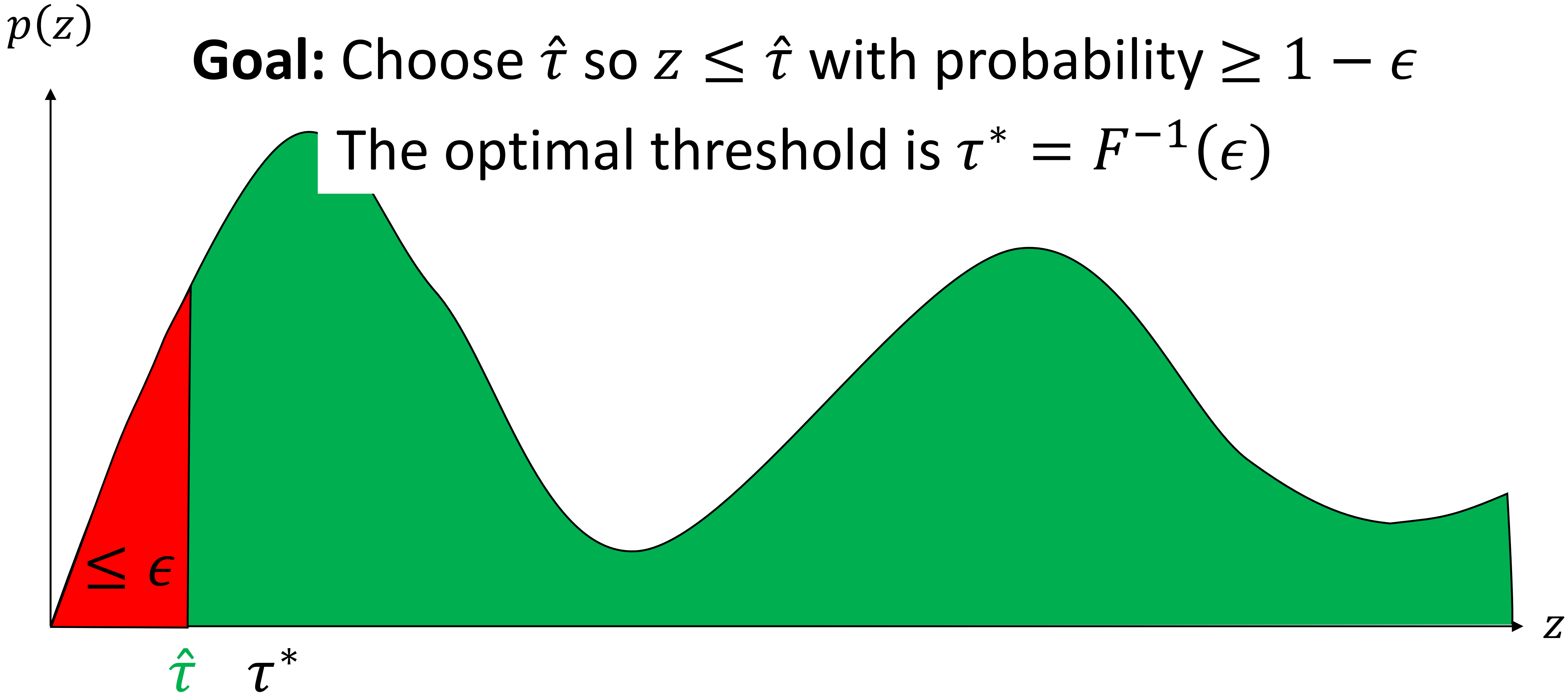
The optimal threshold is $\tau^* = F^{-1}(\epsilon)$



Proof Sketch

Goal: Choose $\hat{\tau}$ so $z \leq \hat{\tau}$ with probability $\geq 1 - \epsilon$

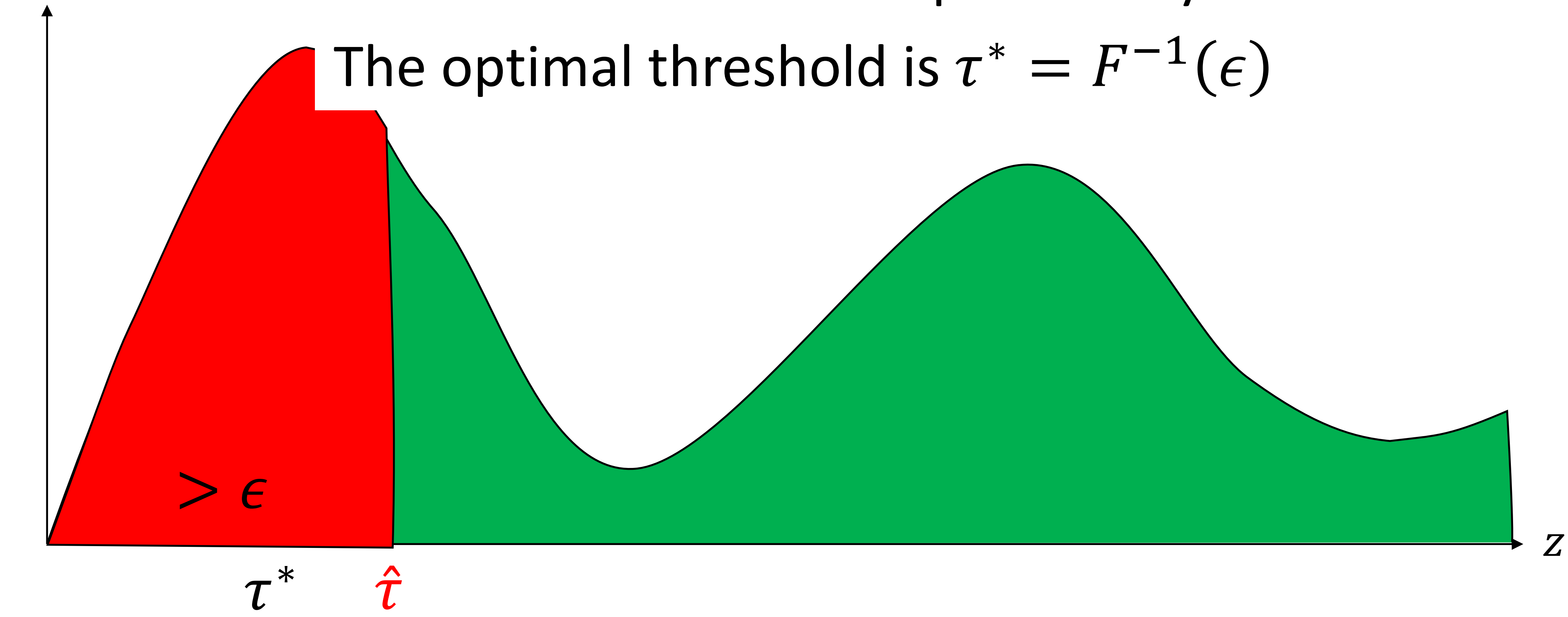
The optimal threshold is $\tau^* = F^{-1}(\epsilon)$



Proof Sketch

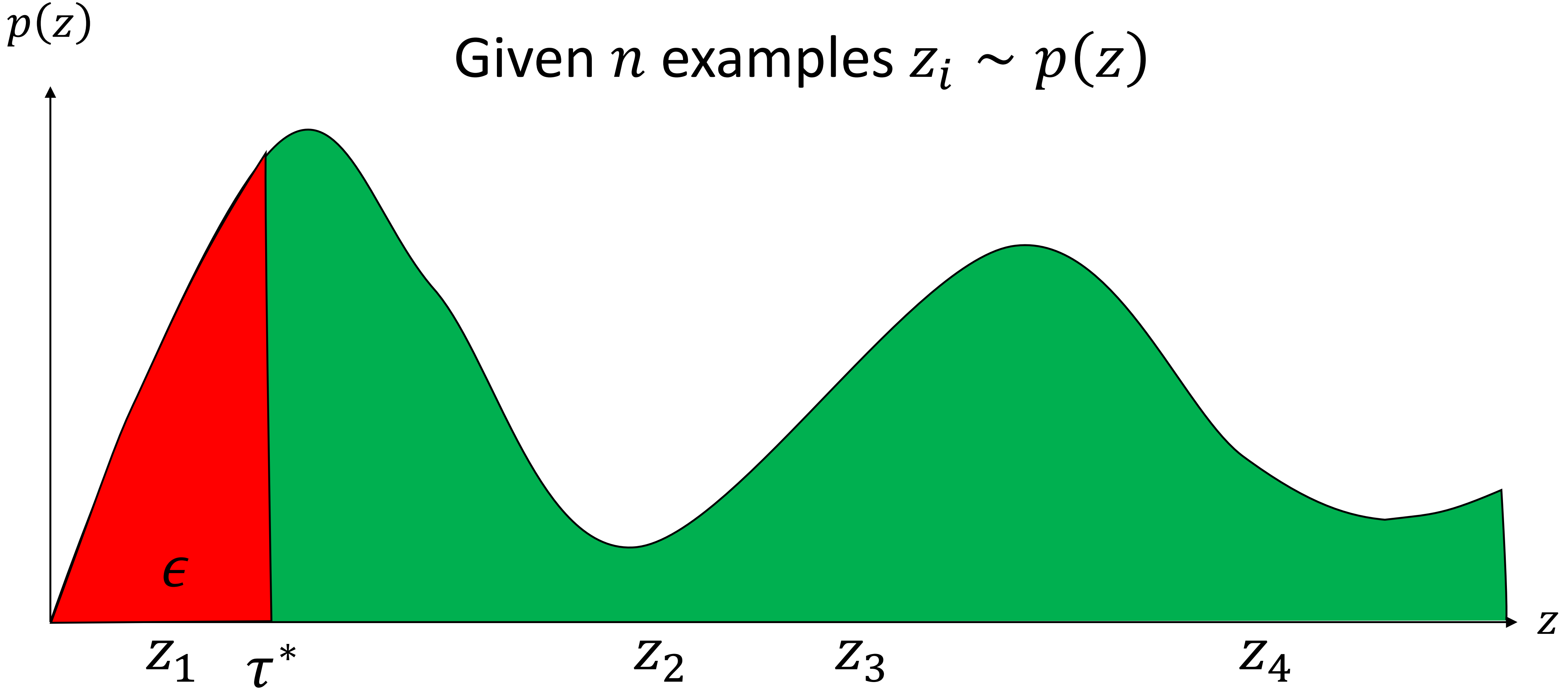
Goal: Choose $\hat{\tau}$ so $z \leq \hat{\tau}$ with probability $\geq 1 - \epsilon$

The optimal threshold is $\tau^* = F^{-1}(\epsilon)$



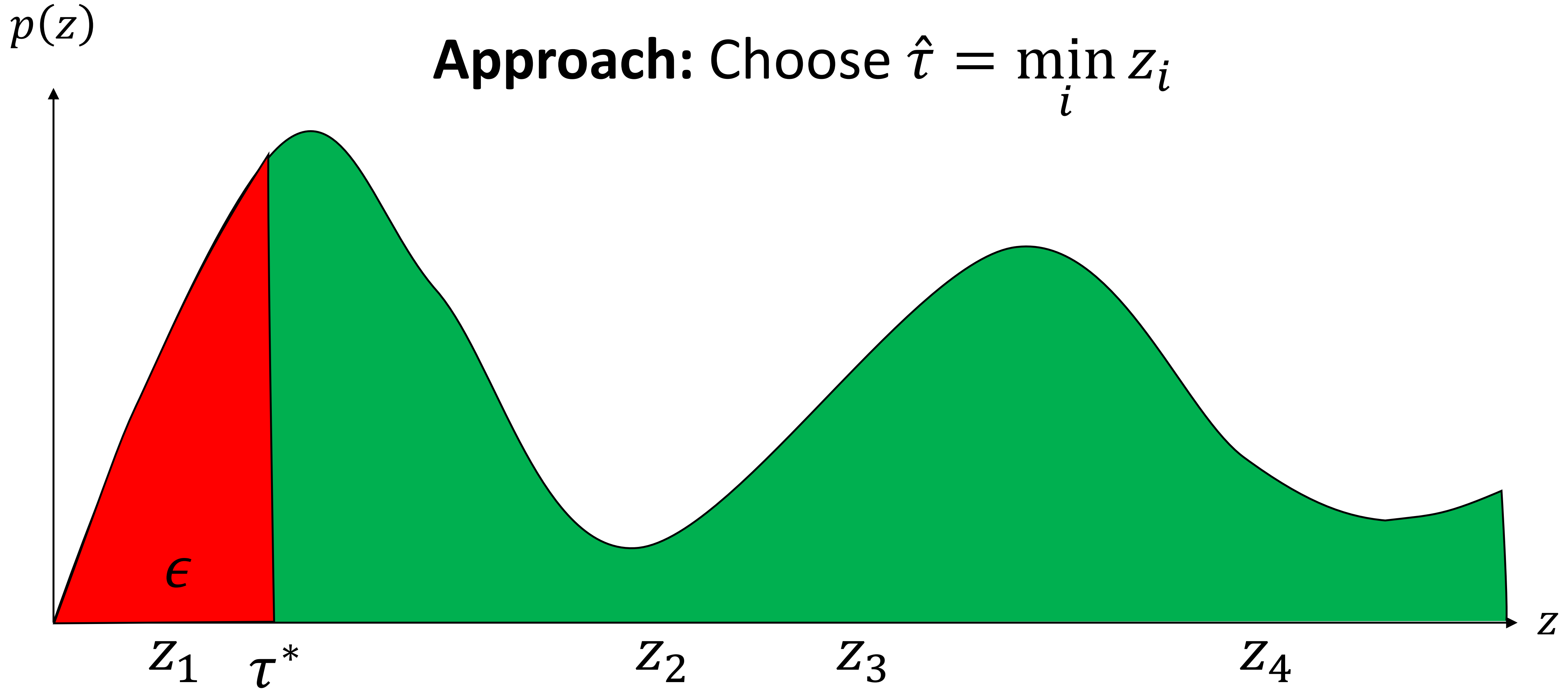
Proof Sketch

Given n examples $z_i \sim p(z)$



Proof Sketch

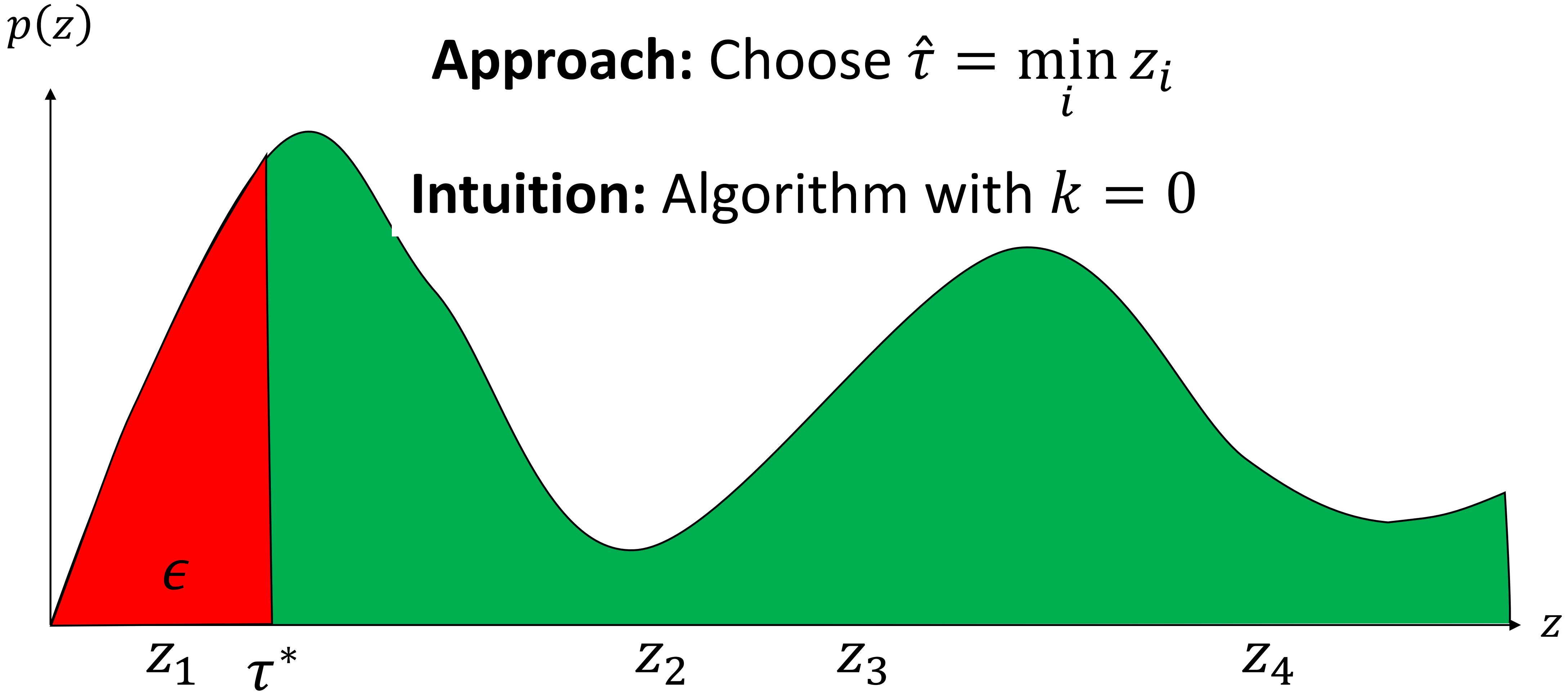
Approach: Choose $\hat{\tau} = \min_i z_i$



Proof Sketch

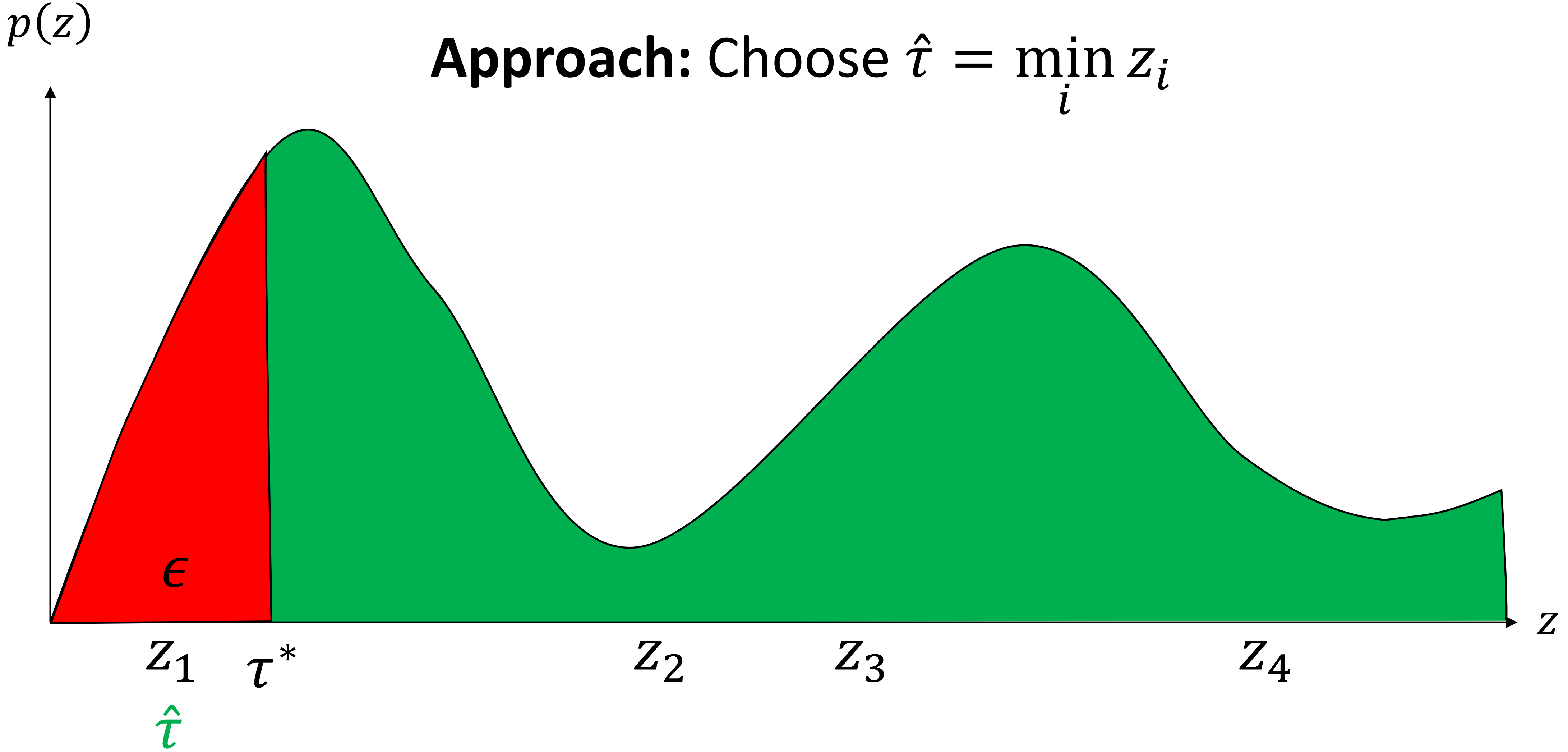
Approach: Choose $\hat{\tau} = \min_i z_i$

Intuition: Algorithm with $k = 0$



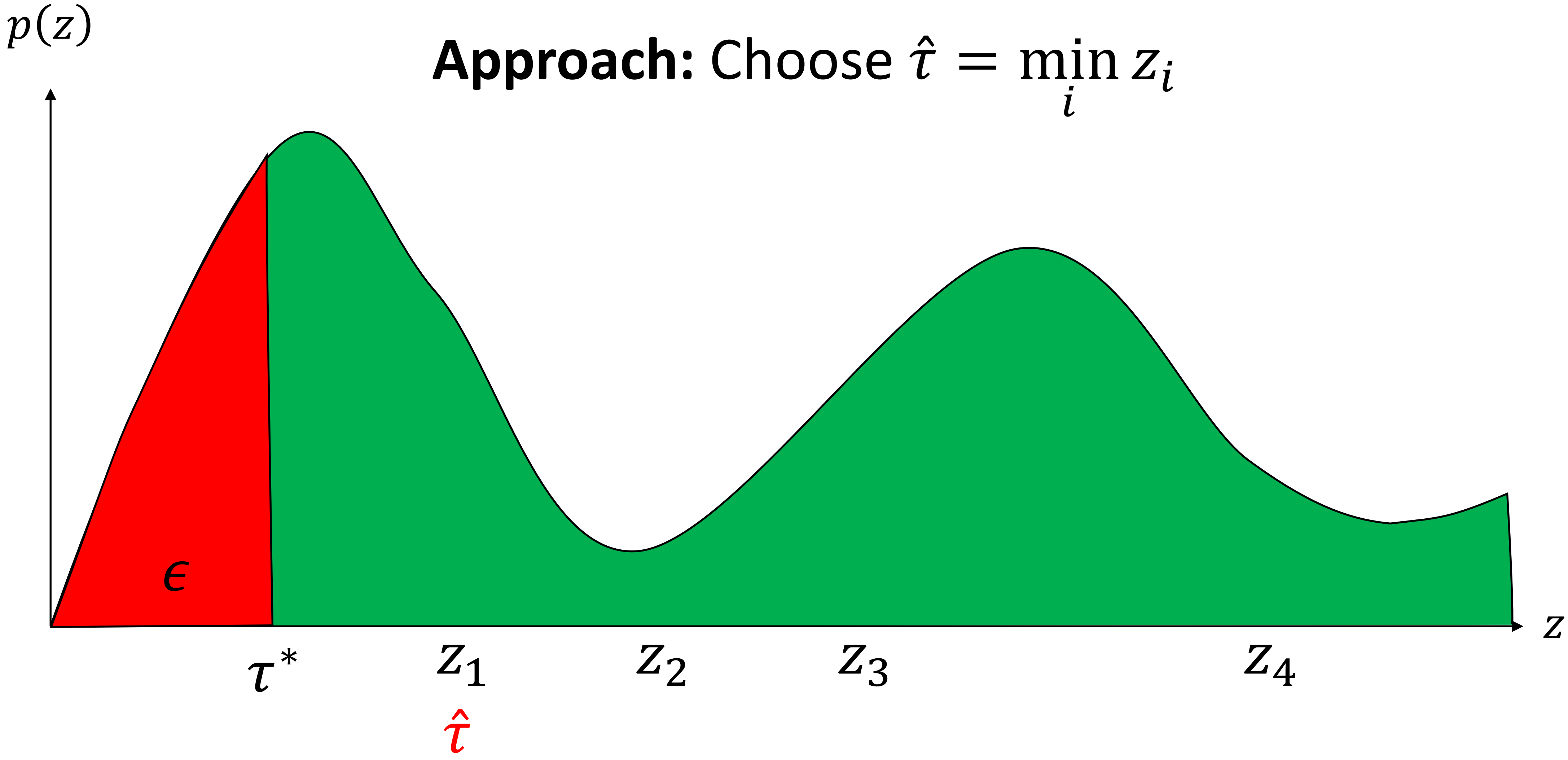
Proof Sketch

Approach: Choose $\hat{\tau} = \min_i z_i$



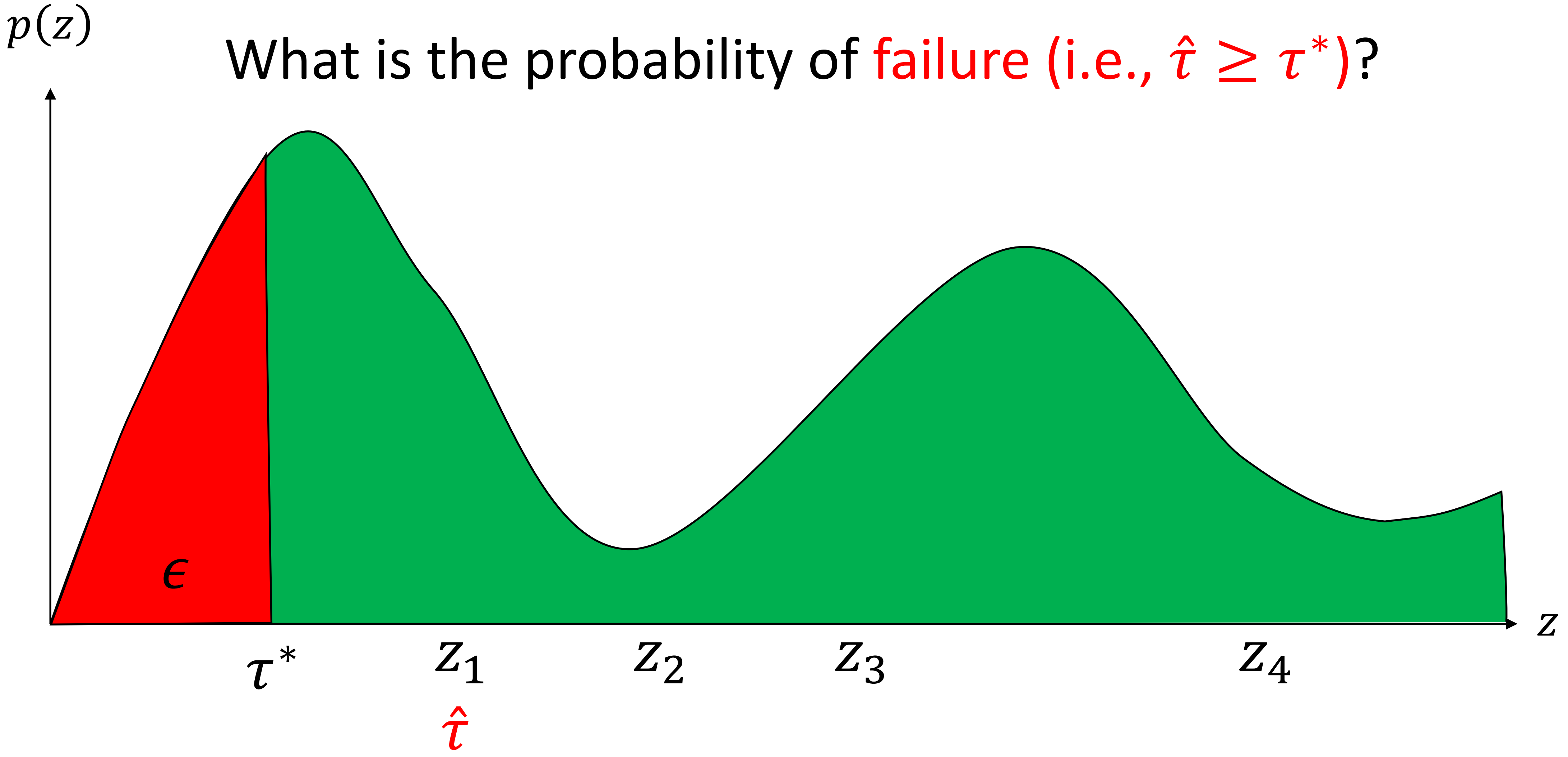
Proof Sketch

Approach: Choose $\hat{\tau} = \min_i z_i$



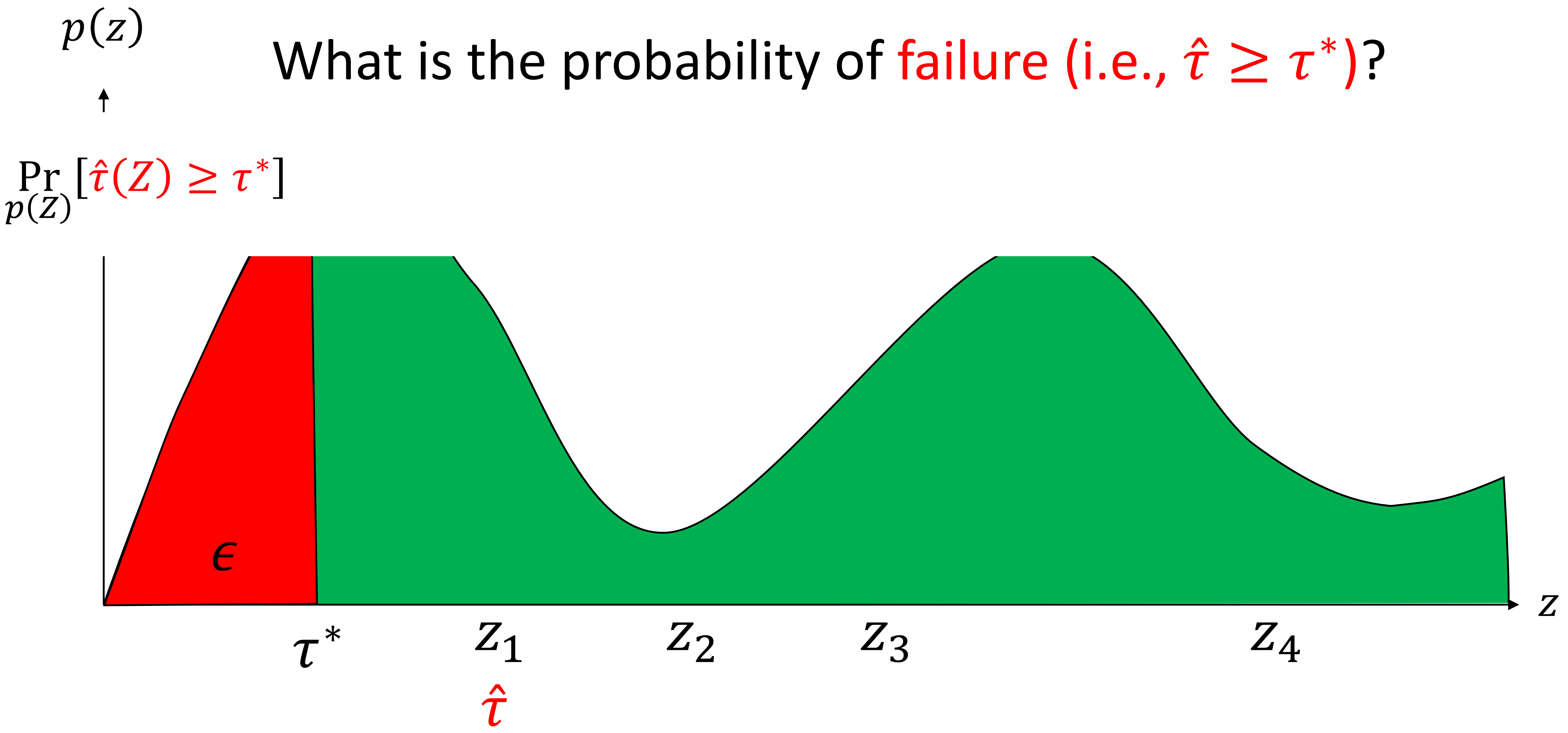
Proof Sketch

What is the probability of failure (i.e., $\hat{\tau} \geq \tau^*$)?



Proof Sketch

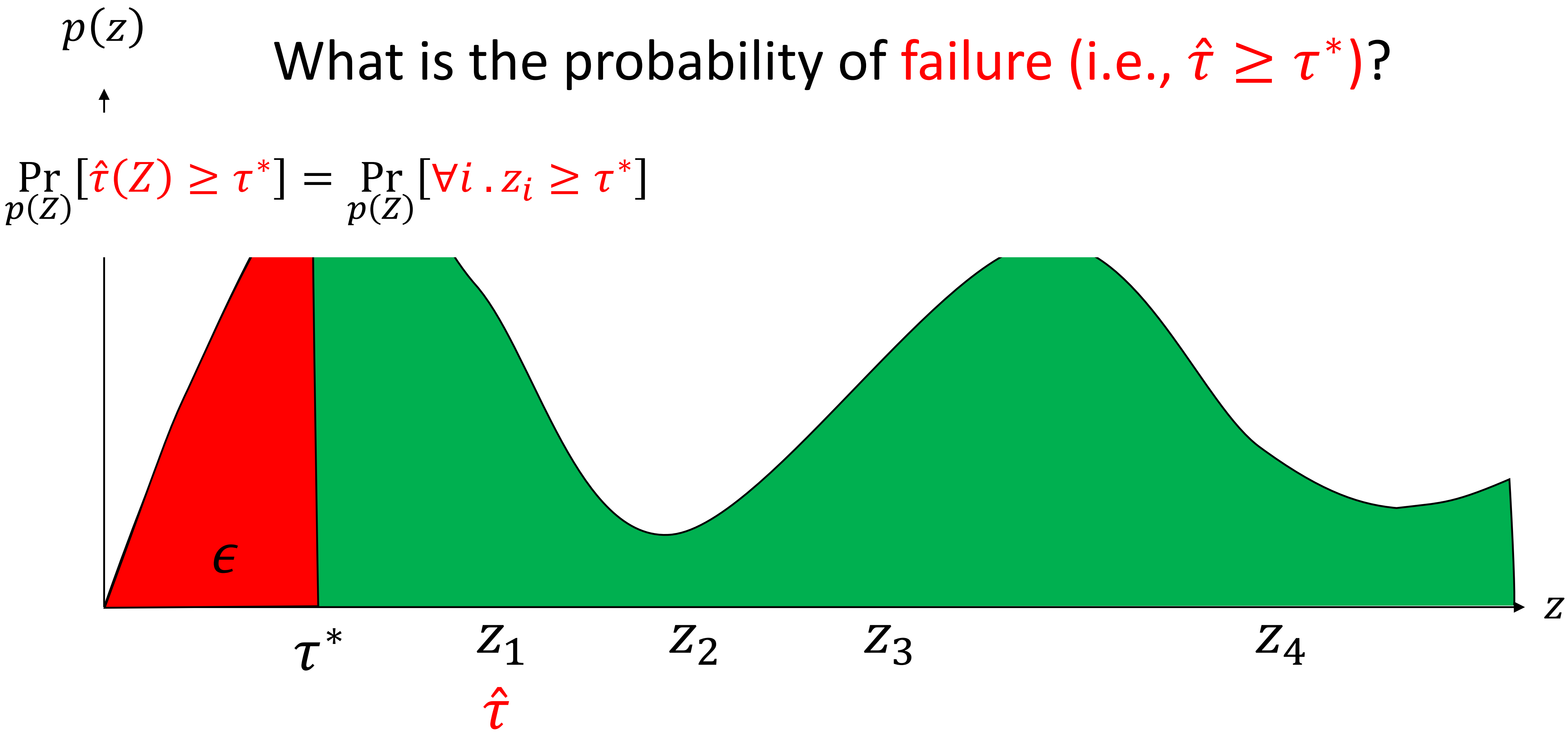
What is the probability of failure (i.e., $\hat{\tau} \geq \tau^*$)?



Proof Sketch

What is the probability of failure (i.e., $\hat{\tau} \geq \tau^*$)?

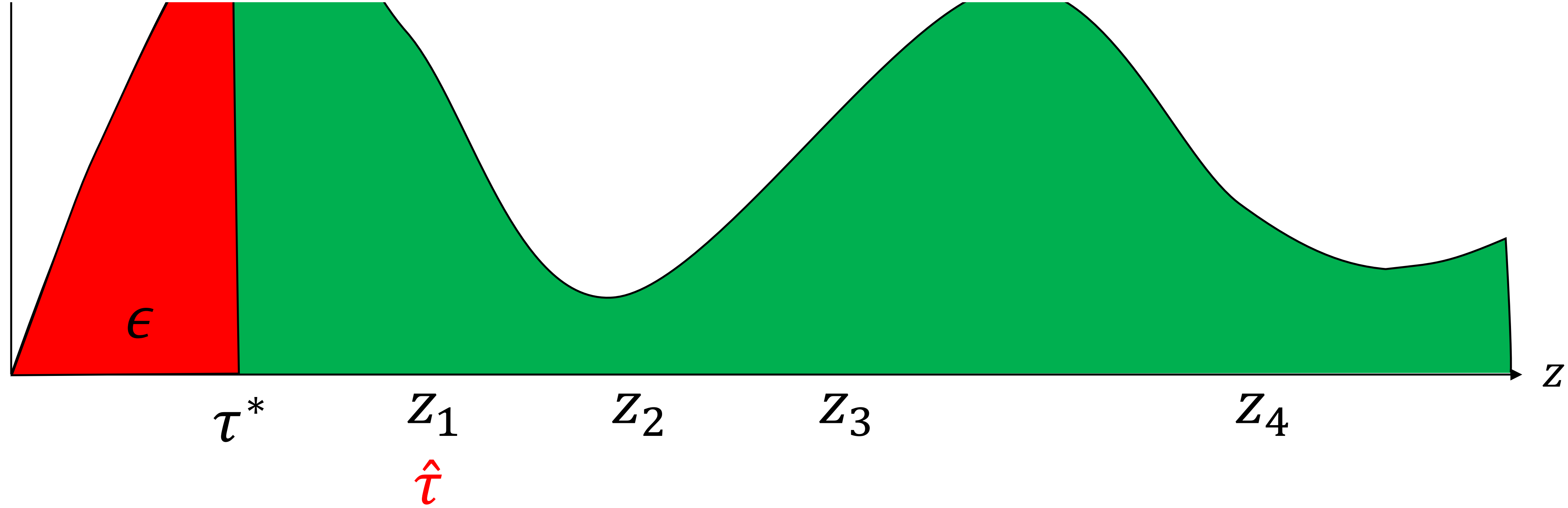
$$\Pr_{p(Z)}[\hat{\tau}(Z) \geq \tau^*] = \Pr_{p(Z)}[\forall i \,.\, z_i \geq \tau^*]$$



Proof Sketch

What is the probability of failure (i.e., $\hat{\tau} \geq \tau^*$)?

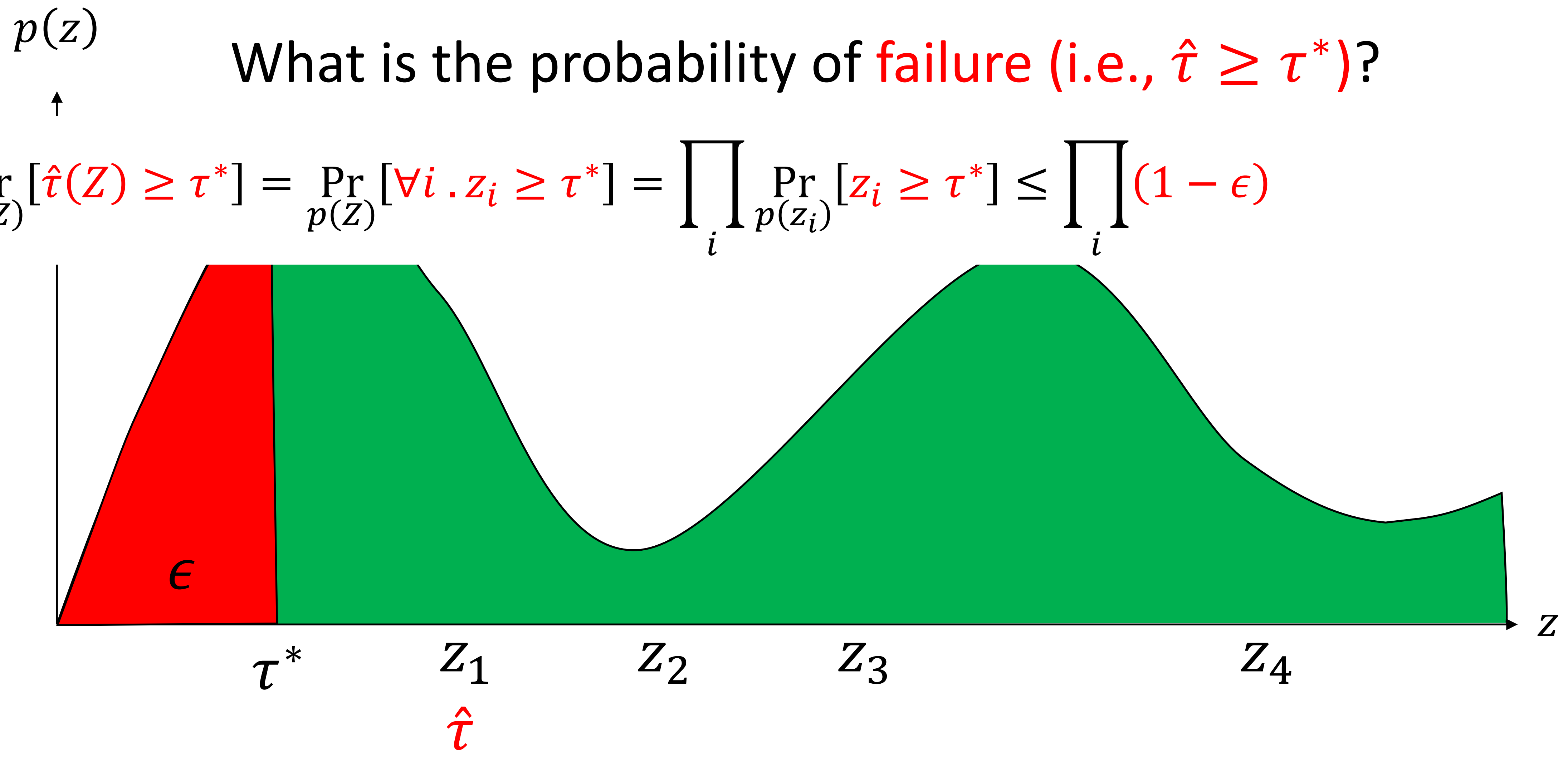
$$\Pr_{p(Z)}[\hat{\tau}(Z) \geq \tau^*] = \Pr_{p(Z)}[\forall i \, . \, z_i \geq \tau^*] = \prod_i \Pr_{p(z_i)}[z_i \geq \tau^*]$$



Proof Sketch

What is the probability of failure (i.e., $\hat{\tau} \geq \tau^*$)?

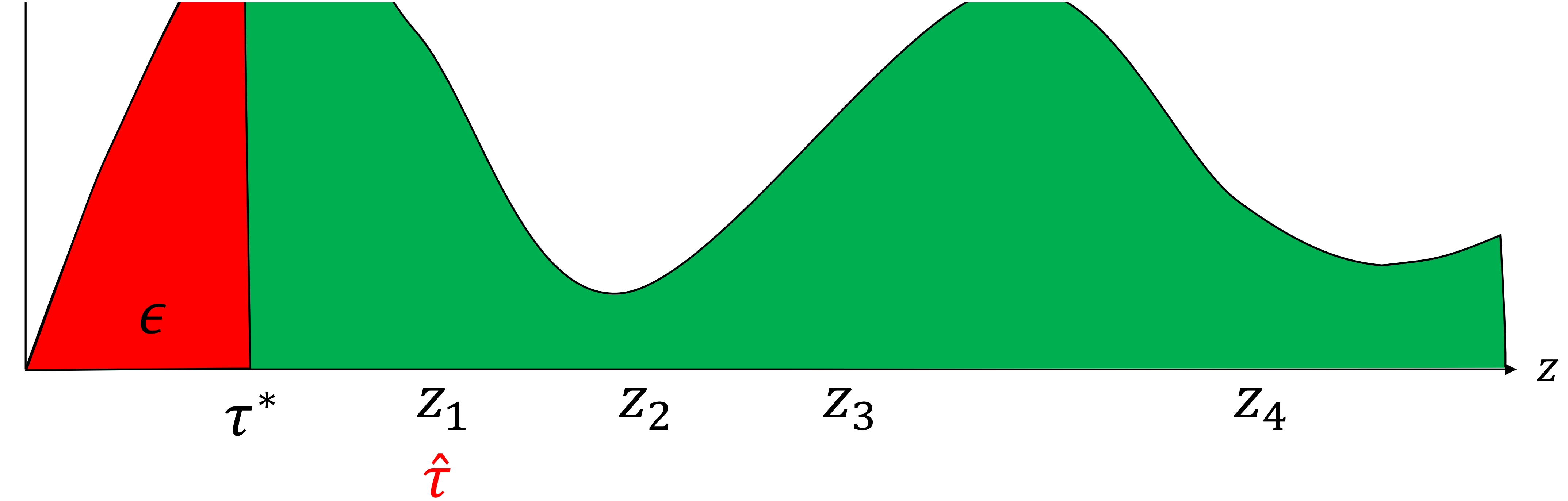
$$\Pr_{p(Z)}[\hat{\tau}(Z) \geq \tau^*] = \Pr_{p(Z)}[\forall i \,.\, z_i \geq \tau^*] = \prod_i \Pr_{p(z_i)}[z_i \geq \tau^*] \leq \prod_i (1-\epsilon)$$



Proof Sketch

What is the probability of failure (i.e., $\hat{\tau} \geq \tau^*$)?

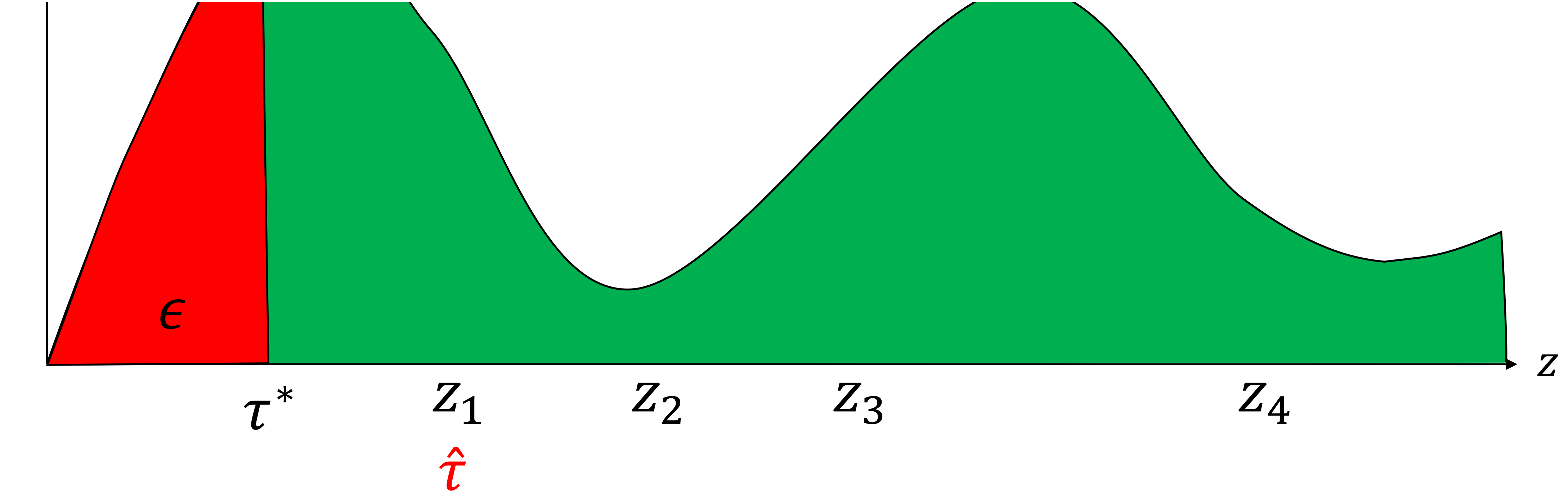
$$\Pr_{p(Z)}[\hat{\tau}(Z) \geq \tau^*] = \Pr_{p(Z)}[\forall i \,.\, z_i \geq \tau^*] = \prod_i \Pr_{p(z_i)}[z_i \geq \tau^*] \leq \prod_i (1-\epsilon) = (1-\epsilon)^n$$



Proof Sketch

What is the probability of failure (i.e., $\hat{\tau} \geq \tau^*$)?

$$\Pr_{p(Z)}[\hat{\tau}(Z) \geq \tau^*] = \Pr_{p(Z)}[\forall i\,.\, z_i \geq \tau^*] = \prod_i \Pr_{p(z_i)}[z_i \geq \tau^*] \leq \prod_i (1-\epsilon) = (1-\epsilon)^n \leq e^{-n\epsilon}$$

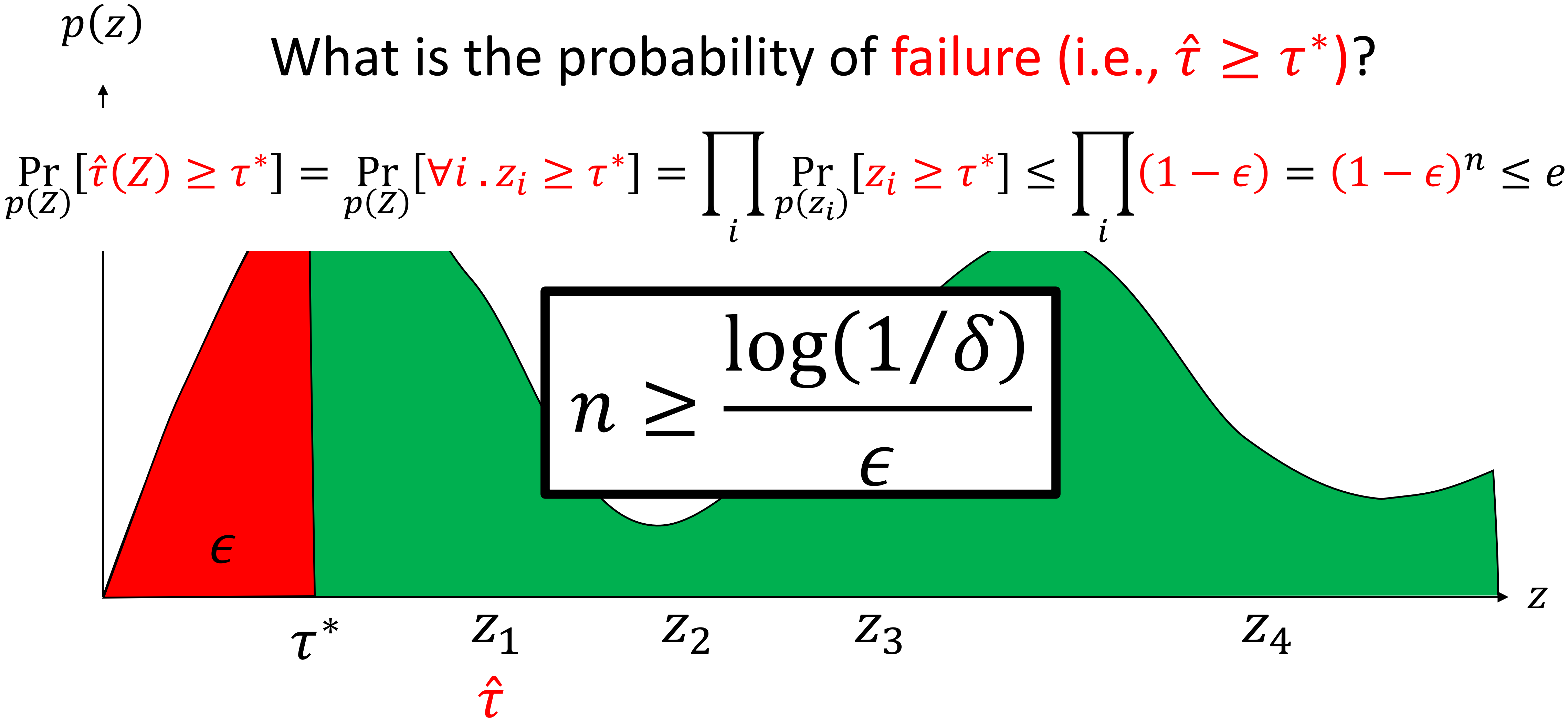


Proof Sketch

What is the probability of failure (i.e., $\hat{\tau} \geq \tau^*$)?

$$\Pr_{p(Z)}[\hat{\tau}(Z) \geq \tau^*] = \Pr_{p(Z)}[\forall i\,.\, z_i \geq \tau^*] = \prod_i \Pr_{p(z_i)}[z_i \geq \tau^*] \leq \prod_i (1-\epsilon) = (1-\epsilon)^n \leq e^{-n\epsilon}$$

$$\boxed{n \geq \frac{\log(1/\delta)}{\epsilon}}$$



Agenda

- Conformal prediction problem
- Conformal prediction algorithm
- Correctness proof