Lecture 14: Aleatoric vs. Epistemic Uncertainty

CIS 7000: Trustworthy Machine Learning

Spring 2024

Homework 2

- **Logistics**
  - Due Monday, March 18
  - Minor typo fix
  - Algorithm descriptions can be high-level
- **Office hours**
  - Alaia will have office hours from 12:30-1:30pm on Friday, March 15
  - I will have office hours from 4-5pm on Friday, March 15

Agenda

- Aleatoric vs. epistemic uncertainty
- Linear regression example
- Bootstrapping ensembles for estimating epistemic uncertainty
- Application to active learning

Predictive Uncertainty

- **Goal:** What is the distribution of $y - f_{\hat{\beta}}(x)$?

- Useful for decision-making
  - Uncertain → patient should be seen by a doctor
  - Uncertain → robot should avoid potential obstacle

- However, aggregates multiple sources of uncertainty

Motivation: Active Learning

- **Goal:** Will obtaining additional information help make better decisions?

- **Example**
  - Robot is not sure if an object is a fork or a spoon
  - Is it worth moving closer to get a better look?

yes!
epistemic uncertainty

no!
aleatoric uncertainty

Aleatoric vs. Epistemic Uncertainty

- **Epistemic uncertainty**
  - Uncertainty due to limitations in our knowledge about the world
  - Can be eliminated by obtaining additional labels/information

- **Aleatoric uncertainty**
  - “Intrinsic” uncertainty that can’t be avoided
  - Not helpful to obtain additional labels/information

Another Example

- **Scenario:** Gave a loan to an individual, but they failed to repay; why?

- **Case 1:** They had bad credit score, but we didn't bother checking
  - Epistemic uncertainty
  - Gathering additional information would have helped

- **Case 2:** They were robbed
  - Aleatoric uncertainty
  - Gather additional information would not have helped
  - (What if they lived in a dangerous neighborhood?)

Aleatoric vs. Epistemic Uncertainty

- In general, the **residual error** decomposes as

$$y - f_{\hat{\beta}(Z)}(x)$$

Aleatoric vs. Epistemic Uncertainty

- In general, the **residual error** decomposes as

$$y - f_{\hat{\beta}(Z)}(x) = \left(y - f_{\beta^*}(x)\right)$$

Aleatoric vs. Epistemic Uncertainty

- In general, the **residual error** decomposes as

$$y - f_{\hat{\beta}(Z)}(x) = \underbrace{\left(y - f_{\beta^*}(x)\right)}_{\text{Aleatoric uncertainty}} + \underbrace{\left(f_{\beta^*}(x) - f_{\hat{\beta}(Z)}(x)\right)}_{\text{Epistemic uncertainty}}$$

  - **Aleatoric uncertainty:** Error of best possible model f_{β^*}
  - **Epistemic uncertainty:** Error of our model $f_{\hat{\beta}(Z)}$ vs. f_{β^*}

- How can we disentangle the two?

Agenda

- Aleatoric vs. epistemic uncertainty
- Linear regression example
- Bootstrapping ensembles for estimating epistemic uncertainty
- Application to active learning

Linear Regression

- **Model family:** Linear functions $f_\beta(x) = \beta^\top x$
- **Loss function:** Mean-squared error $L(\beta; Z) = n^{-1} \sum_{i=1}^{n} (y_i - \beta^\top x_i)^2$
- **Closed-form solution:** Compute using matrix operations

Vectorizing Linear Regression

Vectorizing Linear Regression

$$\begin{bmatrix} f_\beta(x_1) \\ \vdots \\ f_\beta(x_n) \end{bmatrix}$$

Vectorizing Linear Regression

$$\begin{bmatrix} f_\beta(x_1) \\ \vdots \\ f_\beta(x_n) \end{bmatrix} = \begin{bmatrix} \beta^\top x_1 \\ \vdots \\ \beta^\top x_n \end{bmatrix}$$

Vectorizing Linear Regression

$$\begin{bmatrix} f_\beta(x_1) \\ \vdots \\ f_\beta(x_n) \end{bmatrix} = \begin{bmatrix} \beta^\top x_1 \\ \vdots \\ \beta^\top x_n \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{d} \beta_j x_{1,j} \\ \vdots \\ \sum_{j=1}^{d} \beta_j x_{n,j} \end{bmatrix}$$

Vectorizing Linear Regression

$$\begin{bmatrix} f_\beta(x_1) \\ \vdots \\ f_\beta(x_n) \end{bmatrix} = \begin{bmatrix} \beta^\top x_1 \\ \vdots \\ \beta^\top x_n \end{bmatrix} = \begin{bmatrix} \boxed{\sum_{j=1}^{d} \beta_j x_{1,j}} \\ \vdots \\ \sum_{j=1}^{d} \beta_j x_{n,j} \end{bmatrix} = \begin{bmatrix} \boxed{x_{1,1} & \cdots & x_{1,d}} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{bmatrix} \boxed{\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}}$$

Vectorizing Linear Regression

$$\begin{bmatrix} f_\beta(x_1) \\ \vdots \\ f_\beta(x_n) \end{bmatrix} = \begin{bmatrix} \beta^\top x_1 \\ \vdots \\ \beta^\top x_n \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{d} \beta_j x_{1,j} \\ \vdots \\ \boxed{\sum_{j=1}^{d} \beta_j x_{n,j}} \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & \ddots & \vdots \\ \boxed{x_{n,1} \quad \cdots \quad x_{n,d}} \end{bmatrix} \boxed{\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}}$$

Vectorizing Linear Regression

$$\begin{bmatrix} f_\beta(x_1) \\ \vdots \\ f_\beta(x_n) \end{bmatrix} = \begin{bmatrix} \beta^\top x_1 \\ \vdots \\ \beta^\top x_n \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{d} \beta_j x_{1,j} \\ \vdots \\ \sum_{j=1}^{d} \beta_j x_{n,j} \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} = X\beta$$

Vectorizing Linear Regression

$$\begin{bmatrix} f_\beta(x_1) \\ \vdots \\ f_\beta(x_n) \end{bmatrix} = \begin{bmatrix} \beta^\top x_1 \\ \vdots \\ \beta^\top x_n \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{d} \beta_j x_{1,j} \\ \vdots \\ \sum_{j=1}^{d} \beta_j x_{n,j} \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} = X\beta$$

$$\approx$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Vectorizing Linear Regression

$$\begin{bmatrix} f_\beta(x_1) \\ \vdots \\ f_\beta(x_n) \end{bmatrix} = \begin{bmatrix} \beta^\top x_1 \\ \vdots \\ \beta^\top x_n \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{d} \beta_j x_{1,j} \\ \vdots \\ \sum_{j=1}^{d} \beta_j x_{n,j} \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} = X\beta$$

$$\wr\wr$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = Y$$

Summary: $Y \approx X\beta$

Vectorizing Linear Regression

$$Y \approx X\beta$$

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}$$

Vectorizing Mean Squared Error

Vectorizing Mean Squared Error

$$L(\beta; Z)$$

Vectorizing Mean Squared Error

$$L(\beta; Z) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta^\top x_i)^2$$

Vectorizing Mean Squared Error

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \begin{bmatrix} f_\beta(x_1) \\ \vdots \\ f_\beta(x_n) \end{bmatrix}$$

$$L(\beta; Z) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta^\top x_i)^2 = \frac{1}{n}\|Y - X\beta\|_2^2$$

$$\|z\|_2^2 = \sum_{i=1}^{n} z_i^2$$

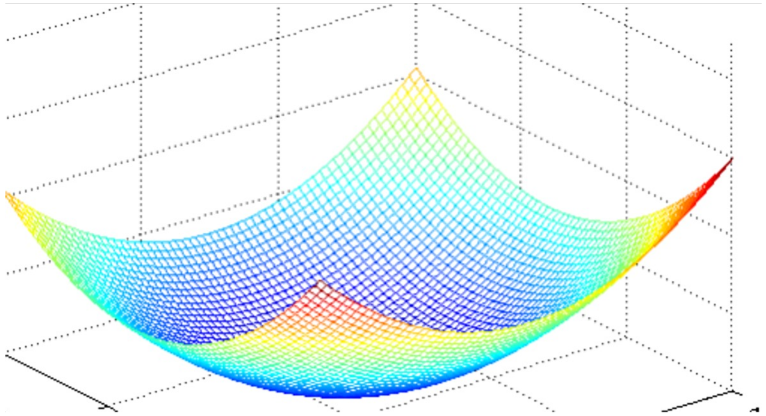
Closed-Form Solution

- Recall that linear regression minimizes the loss

$$L(\beta; Z) = \frac{1}{n} \|Y - X\beta\|_2^2$$

- Minimum solution has gradient equal to zero:

$$\nabla_\beta L\left(\hat{\beta}(Z); Z\right) = 0$$



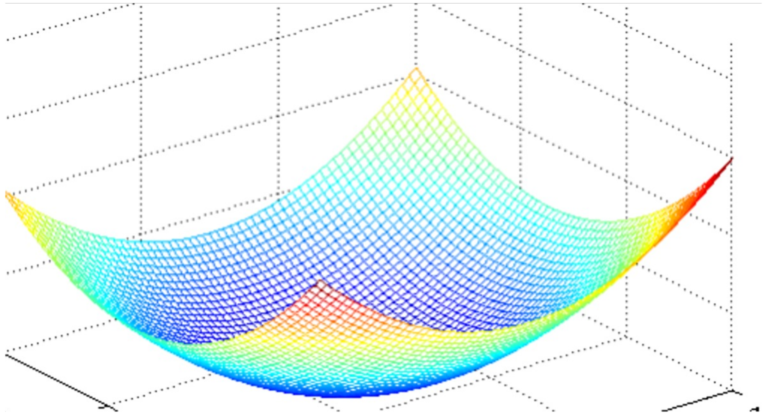
Closed-Form Solution

- Recall that linear regression minimizes the loss

$$L(\beta; Z) = \frac{1}{n} \|Y - X\beta\|_2^2$$

- Minimum solution has gradient equal to zero:

$$\nabla_\beta L(\hat{\beta}; Z) = 0$$



Closed-Form Solution

- The gradient is

$$\nabla_{\beta} L(\beta; Z)$$

Closed-Form Solution

- The gradient is

$$\nabla_{\beta} L(\beta; Z) = \nabla_{\beta} \frac{1}{n} \|Y - X\beta\|_2^2$$

Closed-Form Solution

- The gradient is

$$
\begin{aligned}
\nabla_\beta L(\beta; Z) = \nabla_\beta \frac{1}{n} \|Y - X\beta\|_2^2 &= \nabla_\beta \frac{1}{n} (Y - X\beta)^\top (Y - X\beta) \\
&= \frac{2}{n} \left[\nabla_\beta (Y - X\beta)^\top\right] (Y - X\beta) \\
&= -\frac{2}{n} X^\top (Y - X\beta) \\
&= -\frac{2}{n} X^\top Y + \frac{2}{n} X^\top X\beta
\end{aligned}
$$

Closed-Form Solution

- Thus, we have

$$-\frac{2}{n}X^\top Y + \frac{2}{n}X^\top X\hat{\beta} = 0$$

- Solving for $\hat{\beta}$ gives

$$\hat{\beta} = (X^\top X)^{-1}X^\top Y$$

True Data Generating Process

- Assume that the data is **actually** generated by some linear model:

$$y_i = \beta^{*\top} x_i + \epsilon_i$$

- **Vectorized form:** $Y = X\beta^* + E$, where $E = [\epsilon_1 \quad \cdots \quad \epsilon_n]^\top$

- Then, we have

$$\hat{\beta} - \beta^*$$

True Data Generating Process

- Assume that the data is **actually** generated by some linear model:

$$y_i = \beta^{*\top} x_i + \epsilon_i$$

- **Vectorized form:** $Y = X\beta^* + E$, where $E = [\epsilon_1 \quad \cdots \quad \epsilon_n]^\top$

- Then, we have

$$\hat{\beta} - \beta^* = (X^\top X)^{-1} X^\top Y - \beta^*$$

True Data Generating Process

- Assume that the data is **actually** generated by some linear model:

$$y_i = \beta^{*\top} x_i + \epsilon_i$$

- **Vectorized form:** $Y = X\beta^* + E$, where $E = [\epsilon_1 \quad \cdots \quad \epsilon_n]^\top$

- Then, we have

$$\begin{aligned} \hat{\beta} - \beta^* = (X^\top X)^{-1} X^\top Y - \beta^* &= (X^\top X)^{-1} X^\top (X\beta^* + E) - \beta^* \\ &= (X^\top X)^{-1} X^\top E \end{aligned}$$

Aleatoric vs. Epistemic Uncertainty

- The residual error decomposes as

$$y - \hat{\beta}^\top x$$

Aleatoric vs. Epistemic Uncertainty

- The residual error decomposes as

$$y - \hat{\beta}^{\top} x = (y - \beta^{*\top} x)$$

Aleatoric vs. Epistemic Uncertainty

- The residual error decomposes as

$$y - \hat{\beta}^\top x = \underbrace{(y - \beta^{*\top} x)}_{\text{Aleatoric uncertainty}} + \underbrace{(\beta^{*\top} x - \hat{\beta}^\top x)}_{\text{Epistemic uncertainty}}$$

- **Aleatoric uncertainty:** $y - \beta^{*\top} x = \epsilon$
- **Epistemic uncertainty:** $\beta^{*\top} x - \hat{\beta}^\top x = E^\top X (X^\top X)^{-1} x$

Epistemic Uncertainty

- Note that

$$E^\top X(X^\top X)^{-1}x$$

Epistemic Uncertainty

- Note that

$$E^\top X(X^\top X)^{-1}x = x^\top (X^\top X)^{-1}X^\top E$$

Epistemic Uncertainty

- Note that

$$E^\top X(X^\top X)^{-1}x = x^\top(X^\top X)^{-1}X^\top E = x^\top(X^\top X)^{-1}\sum_{i=1}^{n} x_i\epsilon_i$$

- Suppose that $\epsilon \sim_{\text{i.i.d.}} N(0, \sigma^2)$
- Then, we have

$$\mathbb{E}\left[x^\top(X^\top X)^{-1}\sum_{i=1}^{n} x_i\epsilon_i\right]$$

Epistemic Uncertainty

- Note that

$$E^\top X(X^\top X)^{-1}x = x^\top (X^\top X)^{-1} X^\top E = x^\top (X^\top X)^{-1} \sum_{i=1}^{n} x_i \epsilon_i$$

- Suppose that $\epsilon \sim_{\text{i.i.d.}} N(0, \sigma^2)$
- Then, we have

$$\mathbb{E}\left[x^\top (X^\top X)^{-1} \sum_{i=1}^{n} x_i \epsilon_i\right] = x^\top (X^\top X)^{-1} \sum_{i=1}^{n} x_i \mathbb{E}[\epsilon_i]$$

Epistemic Uncertainty

- Note that

$$E^\top X(X^\top X)^{-1}x = x^\top(X^\top X)^{-1}X^\top E = x^\top(X^\top X)^{-1}\sum_{i=1}^{n} x_i\epsilon_i$$

- Suppose that $\epsilon \sim_{\text{i.i.d.}} N(0, \sigma^2)$
- Then, we have

$$\mathbb{E}\left[x^\top(X^\top X)^{-1}\sum_{i=1}^{n} x_i\epsilon_i\right] = x^\top(X^\top X)^{-1}\sum_{i=1}^{n} x_i\mathbb{E}[\epsilon_i] = 0$$

Epistemic Uncertainty

- The variance satisfies

$$
\begin{aligned}
&\mathrm{Var}[x^\top (X^\top X)^{-1} \textstyle\sum_{i=1}^n x_i \epsilon_i] \\
&= x^\top (X^\top X)^{-1} \left(\textstyle\sum_{i=1}^n x_i \,\mathrm{Var}[\epsilon_i]\, x_i^\top \right) (X^\top X)^{-1} x \\
&= x^\top (X^\top X)^{-1} \left(\textstyle\sum_{i=1}^n x_i \sigma^2 x_i^\top \right) (X^\top X)^{-1} x \\
&= \sigma^2 x^\top (X^\top X)^{-1} x \\
&\approx \frac{\sigma^2 x^\top \Sigma x}{n}
\end{aligned}
$$

Aleatoric vs. Epistemic Uncertainty

- **Aleatoric uncertainty**

$$\text{Aleatoric}(x) = y - \beta^{*\top} x = \epsilon \sim_{\text{i.i.d.}} N(0, \sigma^2)$$

- **Epistemic uncertainty:**

$$\text{Epistemic}(x) = E^\top X (X^\top X)^{-1} x \sim_{\text{i.i.d.}} N\left(0, \frac{\sigma^2 x^\top \Sigma x}{n}\right)$$

  - $\frac{\sigma^2 x^\top \Sigma x}{n} = O\left(\frac{1}{n}\right)$, standard deviation is $O\left(\frac{1}{\sqrt{n}}\right)$

Agenda

- Aleatoric vs. epistemic uncertainty
- Linear regression example
- Bootstrapping ensembles for estimating epistemic uncertainty
- Application to active learning

Aleatoric vs. Epistemic Uncertainty

- In general, we have

$$y - f_{\hat{\beta}}(x) = \left(y - f_{\beta^*}(x)\right) + \left(f_{\beta^*}(x) - f_{\hat{\beta}(Z)}(x)\right)$$

- **Hard to disentangle**
  - We directly observe the predictive uncertainty $y - f_{\hat{\beta}}(x)$
  - But we don't know β^*
- **General strategy (statistics):** Pretend $\hat{\beta} = \beta^*$, and disentangle
  - Works in practice even if it feels circular

Aleatoric vs. Epistemic Uncertainty

- The **epistemic uncertainty** is

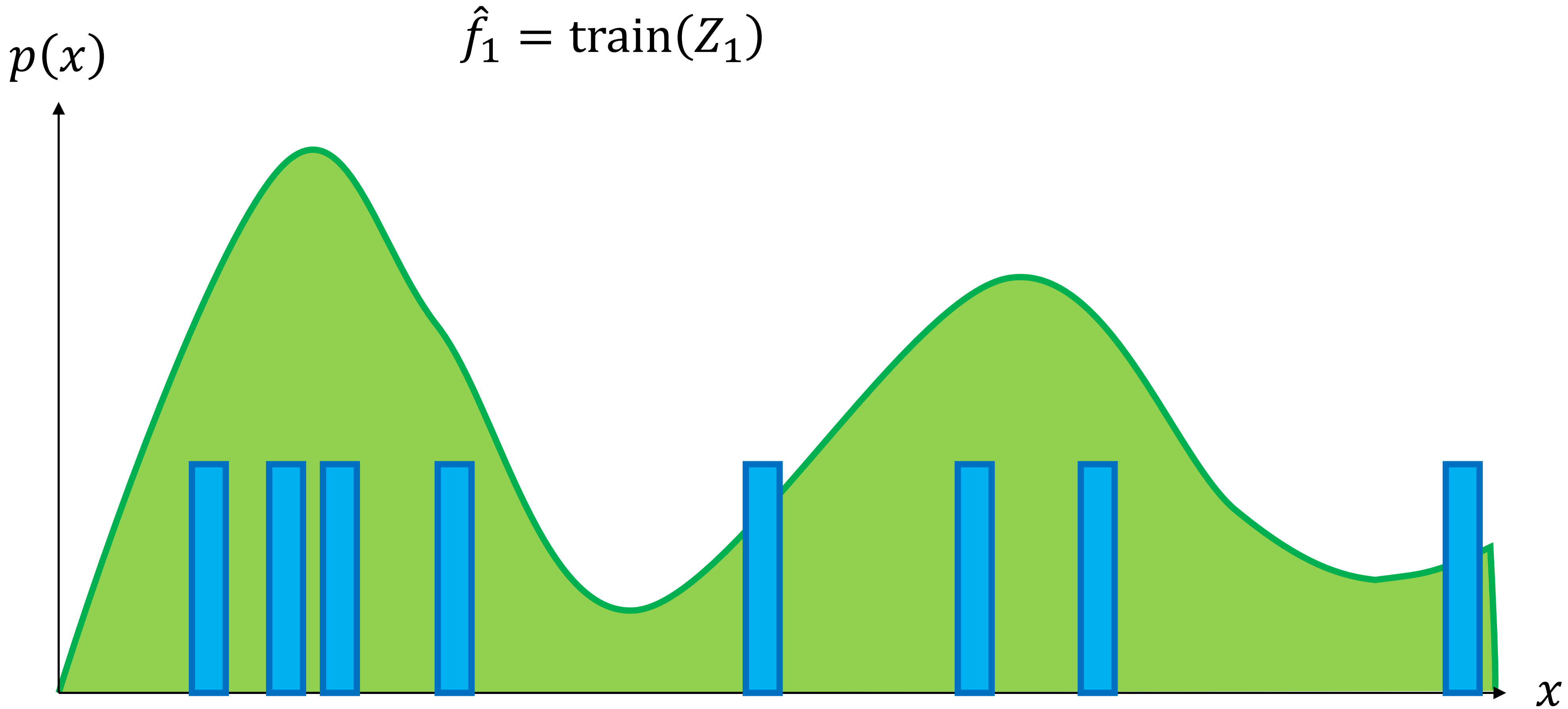
$$\mathrm{Epistemic}(x) = f_{\beta^*}(x) - f_{\hat{\beta}(Z)}(x)$$

  - Here, $\mathrm{Epistemic}(x)$ is a random function of the random variable $Z \sim p^n$
  - Thus, $\mathrm{Epistemic}(x)$ is itself a random variable

- **Goal:** Estimate the distribution of $\mathrm{Epistemic}(x)$

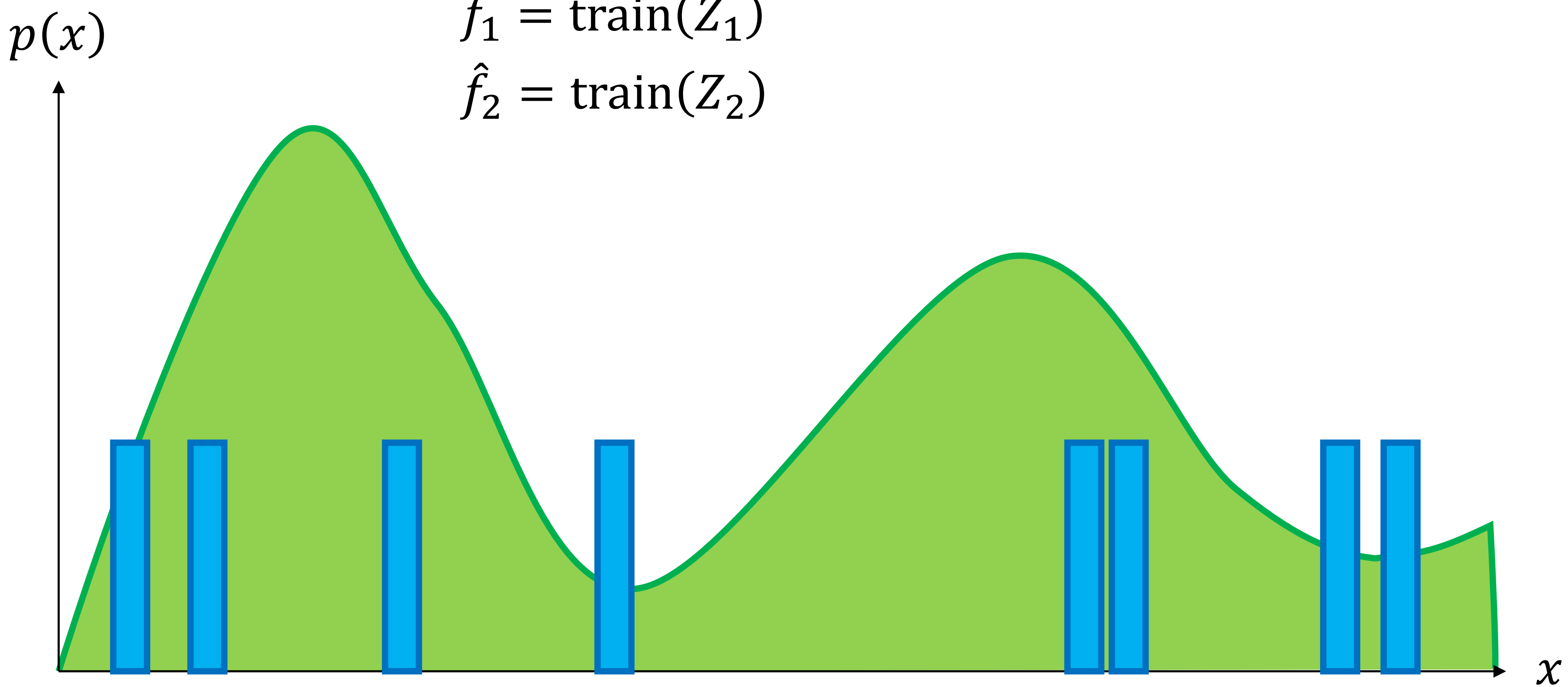
- **Assumption:** Our model is **unbiased**: $\mathbb{E}_Z\left[f_{\hat{\beta}(Z)}(x)\right] = f_{\beta^*}(x)$

What if we know $p(x)$?

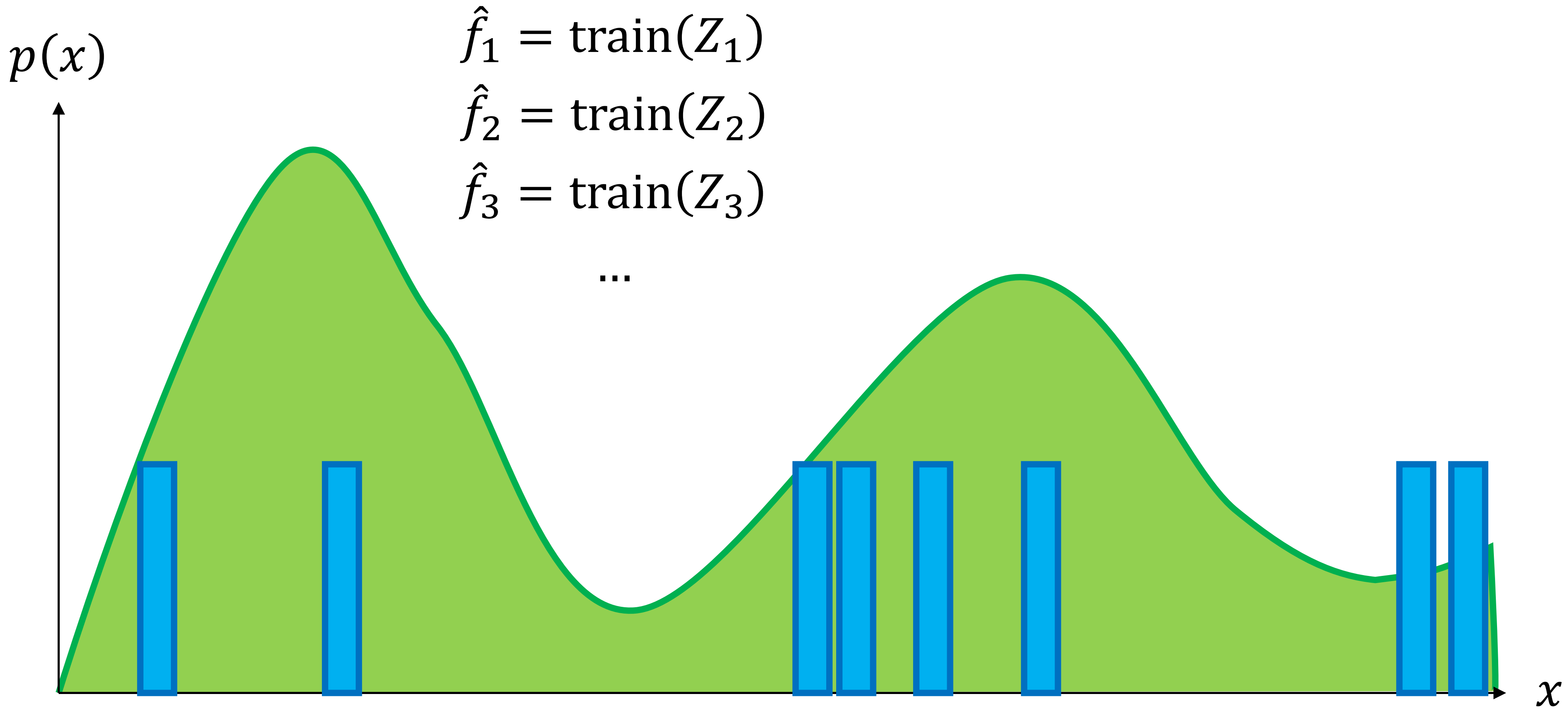


What if we know $p(x)$?

$$\hat{f}_1 = \text{train}(Z_1)$$
$$\hat{f}_2 = \text{train}(Z_2)$$



What if we know $p(x)$?



What if we know $p(x)$?

- By our unbiasedness assumption:

$$f_{\beta^*}(x) = \mathbb{E}_Z\left[f_{\widehat{\beta}(Z)}(x)\right]$$

What if we know $p(x)$?

- $\left\{\hat{f}_i(x) - f_{\beta^*}(x)\right\}_{i=1}^{k}$ are i.i.d. samples from $\mathrm{Epistemic}(x)$
- By our unbiasedness assumption:

$$f_{\beta^*}(x) = \mathbb{E}_Z\left[f_{\widehat{\beta}(Z)}(x)\right]$$

What if we know $p(x)$?

- $\left\{\hat{f}_i(x) - f_{\beta^*}(x)\right\}_{i=1}^{k}$ are i.i.d. samples from $\mathrm{Epistemic}(x)$
- By our unbiasedness assumption:

$$f_{\beta^*}(x) = \mathbb{E}_Z\left[f_{\widehat{\beta}(Z)}(x)\right] \approx k^{-1} \sum_{i=1}^{k} \hat{f}_i(x)$$

What if we know $p(x)$?

- $\{\hat{f}_i(x) - f_{\beta^*}(x)\}_{i=1}^{k}$ are i.i.d. samples from $\text{Epistemic}(x)$
- By our unbiasedness assumption:

$$f_{\beta^*}(x) = \mathbb{E}_Z\left[f_{\hat{\beta}(Z)}(x)\right] \approx k^{-1}\sum_{i=1}^{k}\hat{f}_i(x) \coloneqq \hat{\mu}(x)$$

- $\{\hat{f}_i(x) - \hat{\mu}(x)\}_{i=1}^{k}$ are approximately i.i.d. samples from $\text{Epistemic}(x)$
    - **Problem:** We cannot take unlimited samples from P
    - Only have a single training dataset Z!

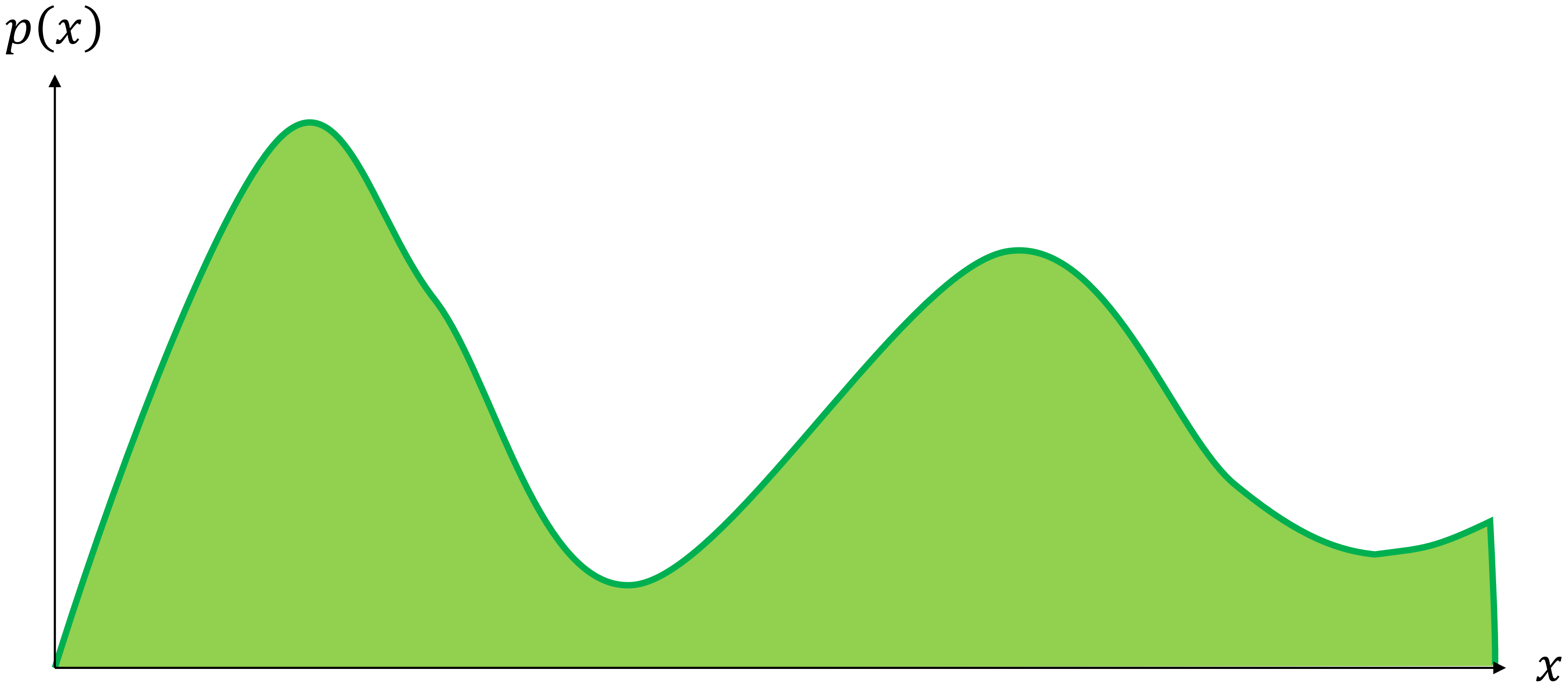
Bootstrap

- **Idea:** Given samples $x_1, \dots, x_n \sim P$, we can "approximate" the probability distribution P by

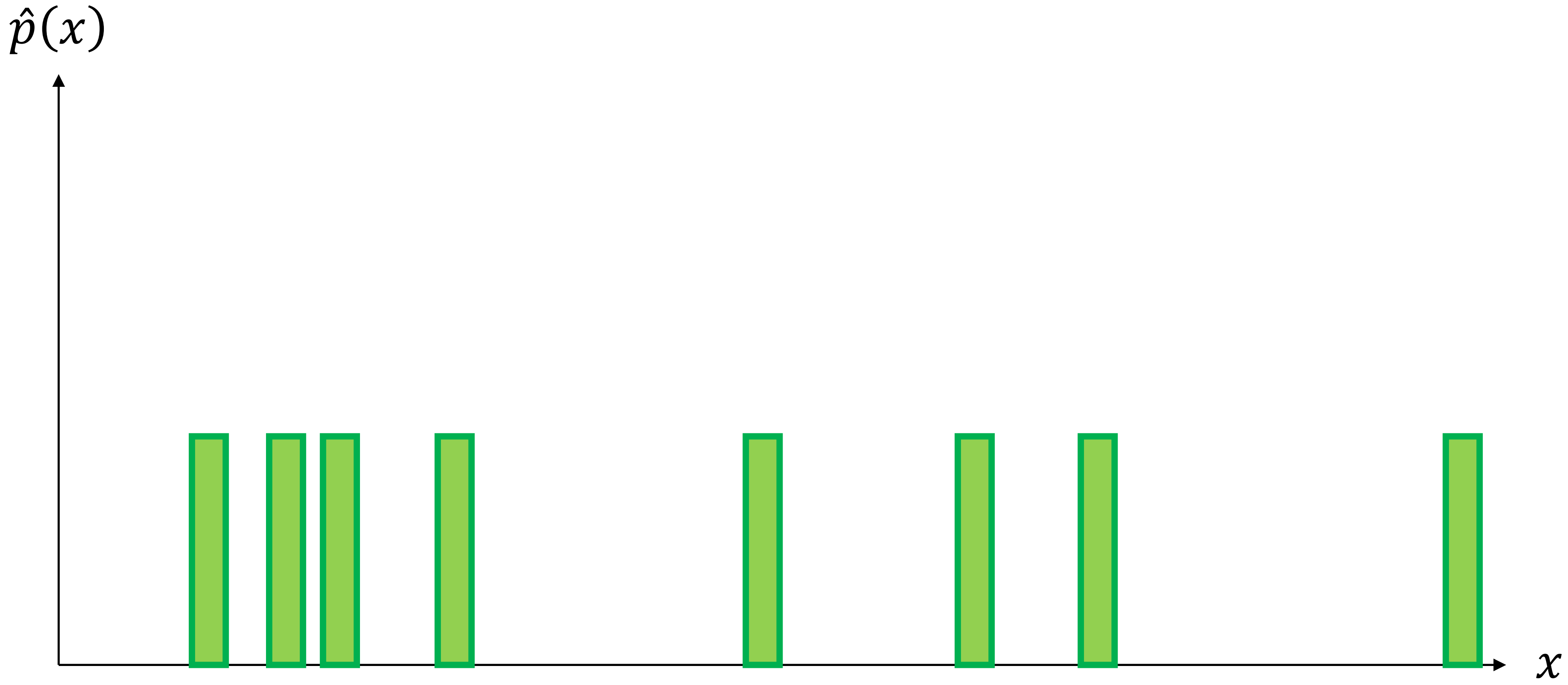
$$P(x) = \Pr[X = x] \approx \frac{1}{n} \sum_{i=1}^{n} 1(x = x_i) = \hat{P}(x)$$

  - This can be made to work for continuous distributions by using the probability density function: $p(x) \approx \hat{p}(x) = n^{-1} \sum_{i=1}^{n} \delta(x - x_i)$
  - For $\mathbb{R}$, uniform convergence of CDF by DKW inequality

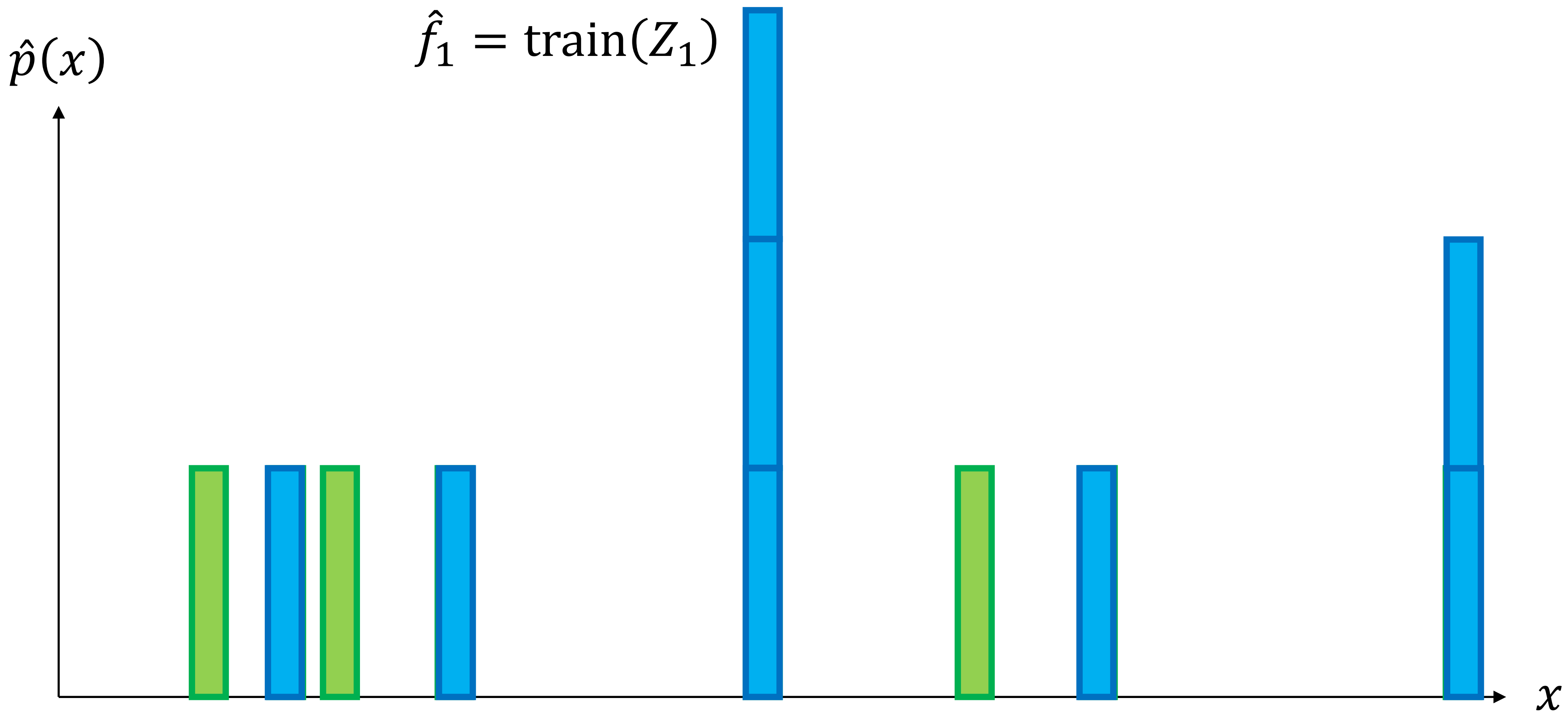
Bootstrap



Bootstrap

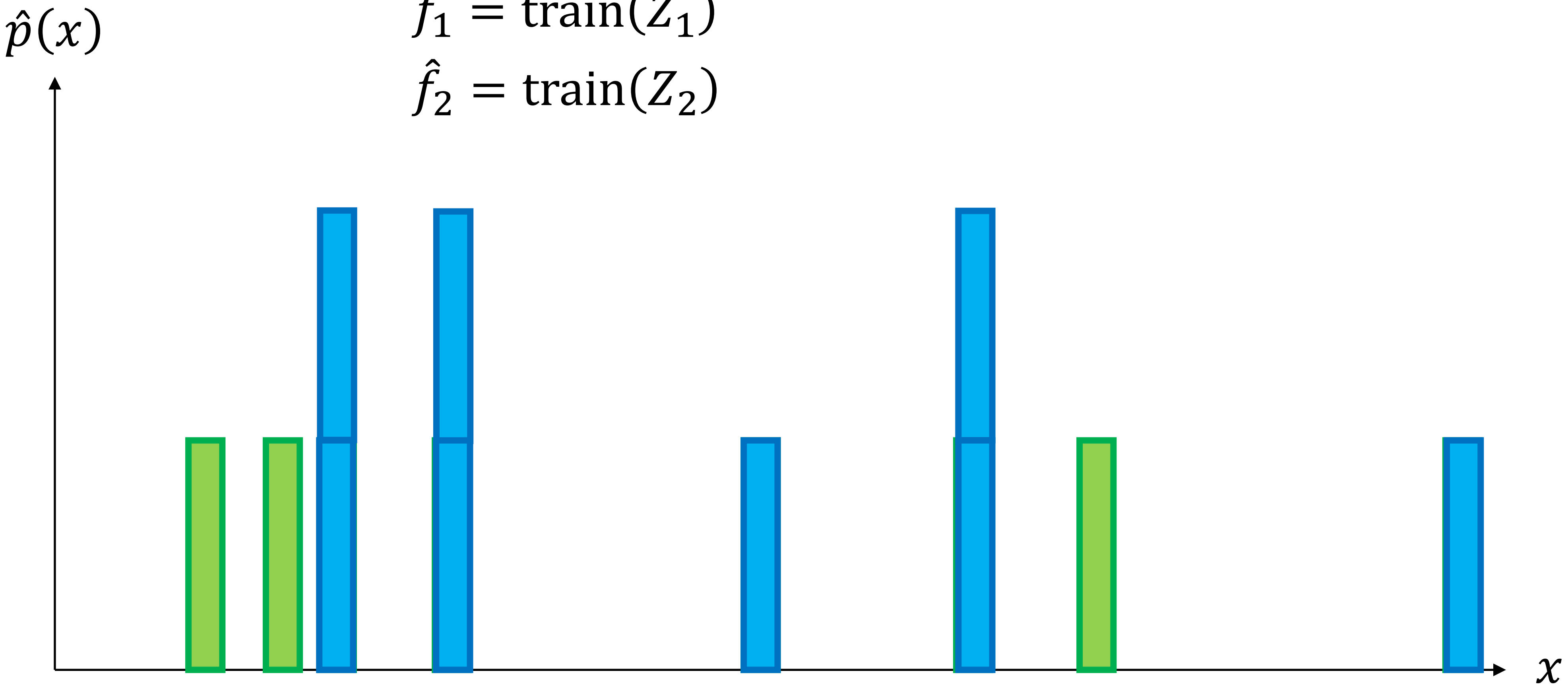


Bootstrap

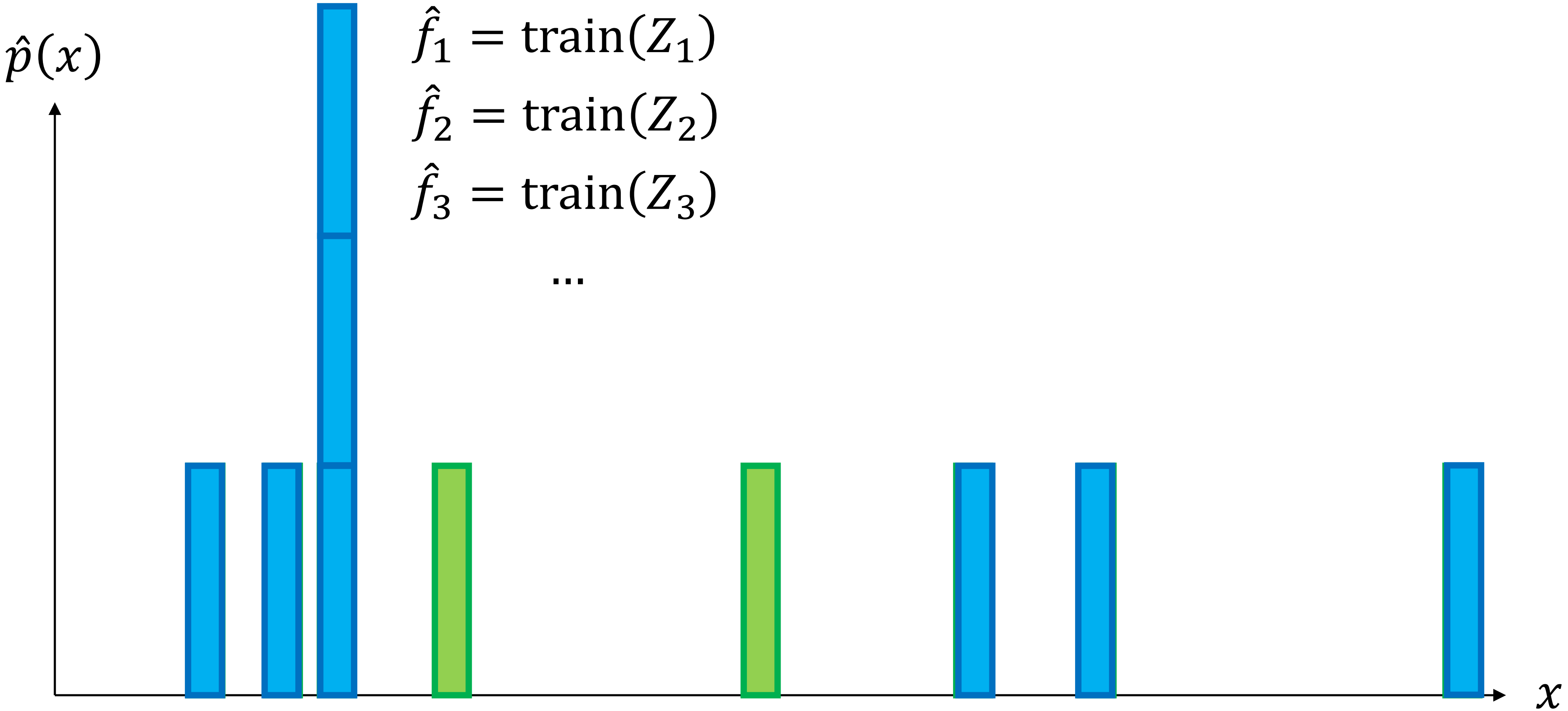


Bootstrap

$$\hat{f}_1 = \text{train}(Z_1)$$
$$\hat{f}_2 = \text{train}(Z_2)$$



Bootstrap



$\hat{f}_1 = \text{train}(Z_1)$

$\hat{f}_2 = \text{train}(Z_2)$

$\hat{f}_3 = \text{train}(Z_3)$

...

Bootstrap

- Subsample examples $\{(x, y)\}$ **with replacement**

- How do the new samples Z_i differ from the original sample Z?
  - They exclude $\approx \left(1 - \frac{1}{n}\right)^n$ of the training examples
  - As $n \to \infty$, excludes $\to \frac{1}{e} \approx 36.8\%$ examples

- Produces valid confidence intervals in many settings

Estimating Epistemic Uncertainty

bootstrap$(Z = \{(x_i, y_i^*)\}_{i=1}^n)$

$\hat{p} \leftarrow n^{-1} \sum_{i=1}^n \delta(x - x_i)$

for $i \in \{1, \ldots, k\}$:

$Z_i \sim_{\text{i.i.d.}} \hat{p}^n$

$\hat{f}_i \leftarrow \text{train}(Z_i)$

return $x \mapsto \left\{\hat{f}_i(x) - \hat{\mu}(x)\right\}_{i=1}^k$, where $\hat{\mu}(x) = k^{-1} \sum_{i=1}^k \hat{f}_i(x)$

Application to Active Learning

- Train bootstrapped ensemble of models $\left\{\hat{f}_i(x)\right\}_{i=1}^{k}$
- Label example where the ensemble has the highest **disagreement:**

$$x^* = \arg\max_x \frac{1}{k^2} \sum_{i,j=1}^{k} 1\left(\hat{f}_i(x) \neq \hat{f}_j(x)\right)$$

Application to Active Learning

- Train bootstrapped ensemble of models $\{\hat{f}_i(x)\}_{i=1}^{k}$

- Label example where the ensemble has the highest **disagreement:**

$$x^* = \arg\max_x \frac{1}{k^2} \sum_{i,j=1}^{k} 1\left(\hat{f}_i(x) \neq \hat{f}_j(x)\right) \approx \Pr_{Z,Z'}\left[f_{\hat{\beta}(Z)}(x) \neq f_{\hat{\beta}(Z')}(x)\right]$$

  - Other metrics based on epistemic uncertainty can also be used
  - Also commonly used for guiding exploration in reinforcement learning
  - More generally, decision-making with opportunity to gather information

Agenda

- Aleatoric vs. epistemic uncertainty
- Linear regression example
- Bootstrapping ensembles for estimating epistemic uncertainty
- Application to active learning