Distributional Reinforcement Learning

Quantile Regression - Implicit Quantile Networks

February 2019

Pantelis R. Vlachas

Seminar on Deep Reinforcement Learning

ETH Zurich

Distribution over returns is NOT a new idea…

[1] Sobel, L.M. 1982

[2] Morimura, T. et. al. 2010

But achieved state-of-the-art results…

C51 [3] Bellemare, M.G. et. al. 2017

And triggered a lot of discussion…

[4] Barth-Maron, G. et. al. 2018

[5] Hessel, M. et. al. 2018

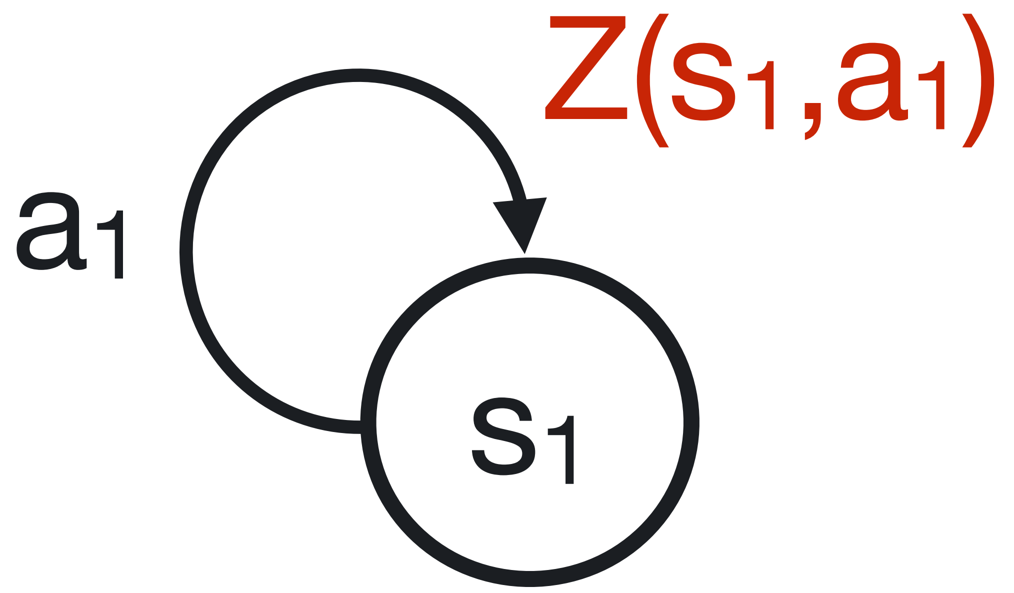
[6] Gruslys, A. et. al. 2018

[7] Rowland, M. et. al. 2018

In this session…

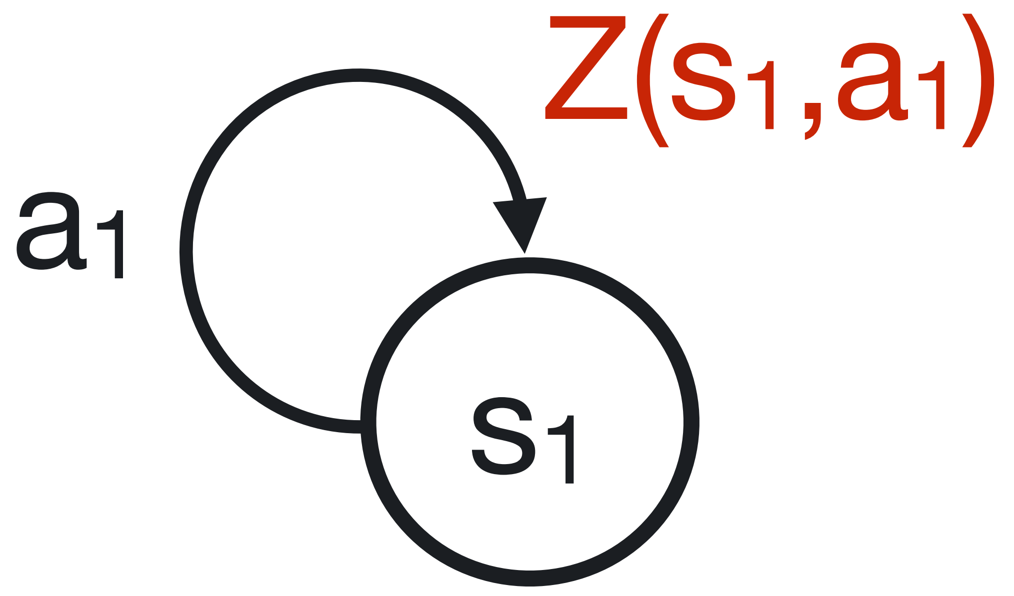
Distributional Reinforcement Learning with (Implicit) Quantile Regression

How to model distributions over returns ?



How to model distributions over returns ?

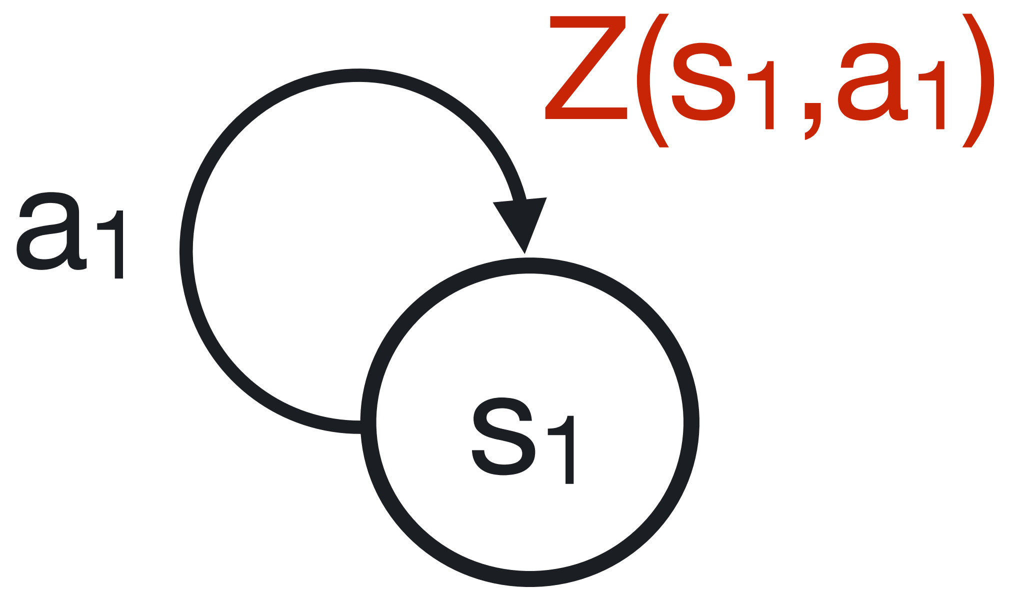
1. **Categorical** of PDF or CDF



How to model distributions over returns ?

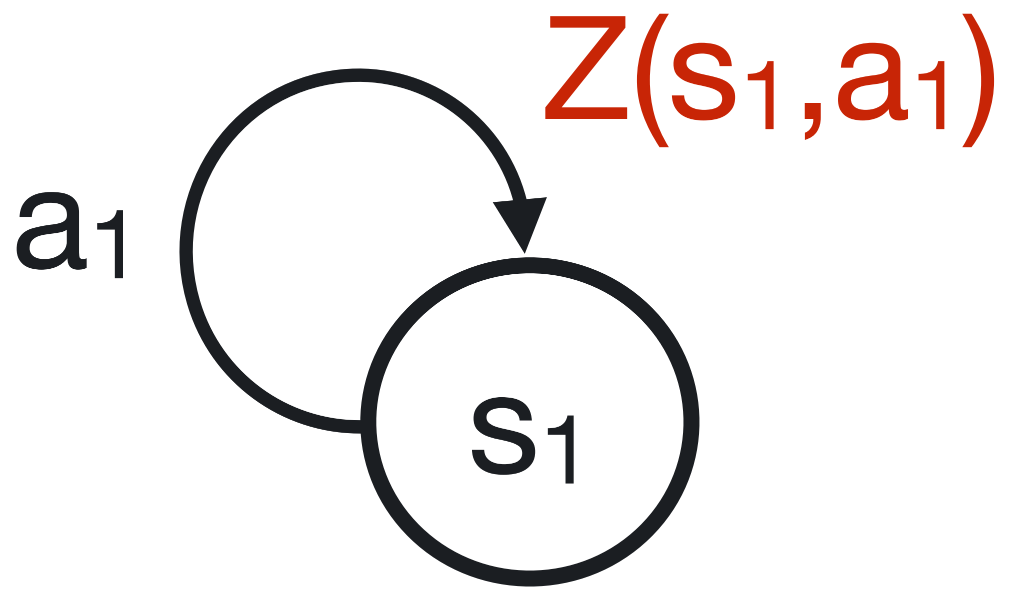
1. **Categorical** of PDF or CDF

2. **Quantiles** of Inverse CDF



How to model distributions over returns ?

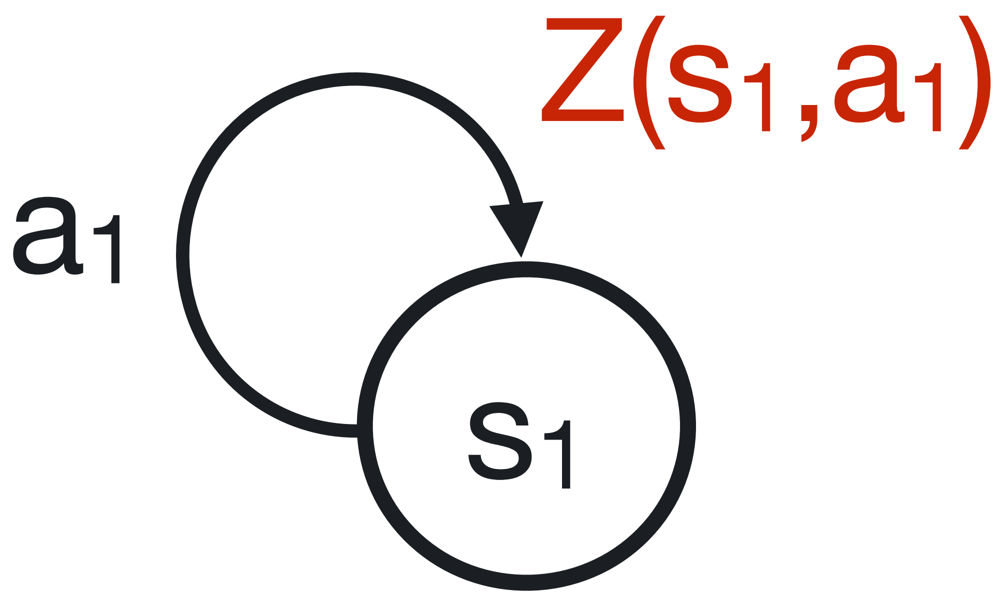
1. **Categorical** of PDF or CDF



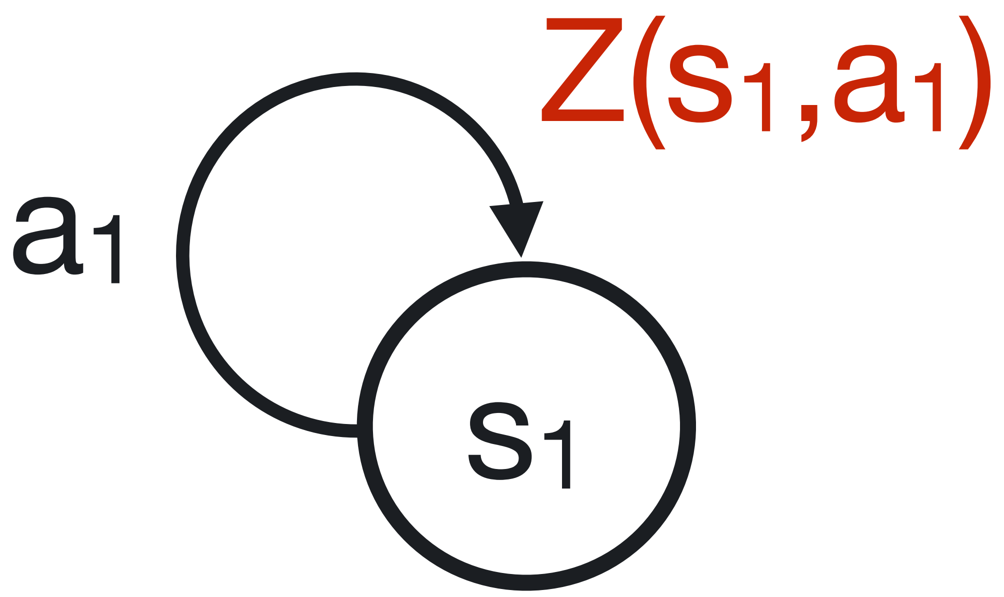
2. **Quantiles** of Inverse CDF

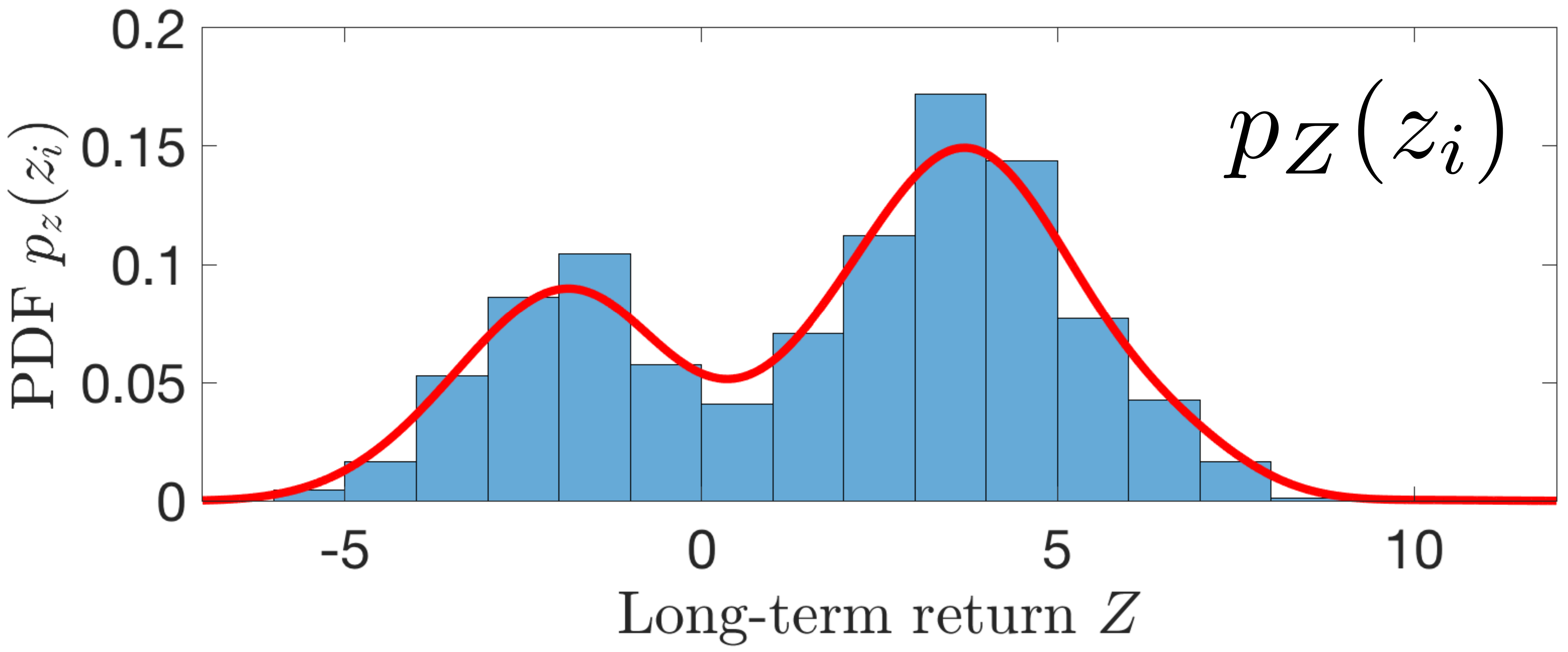
3. **Implicit Quantiles** of Inverse CDF

1. Categorical

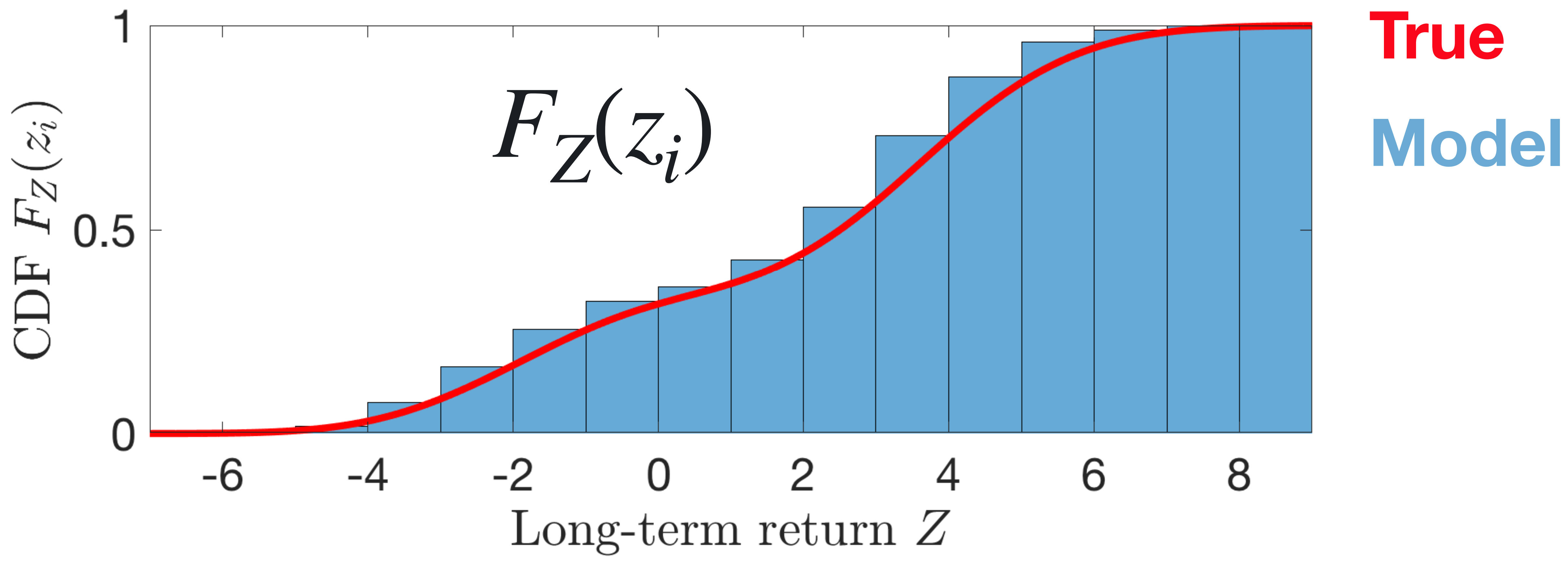


1. Categorical



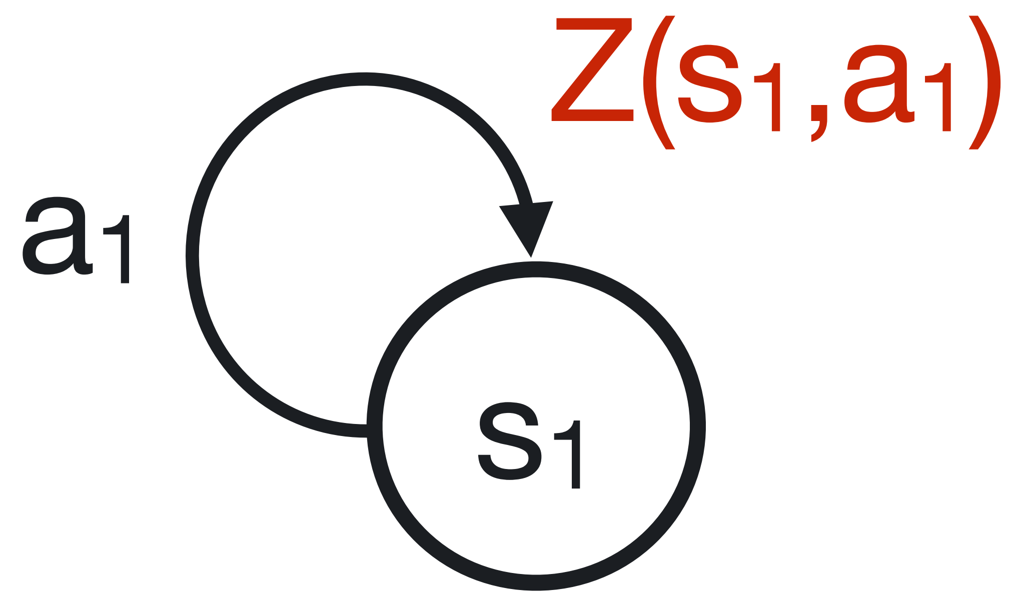


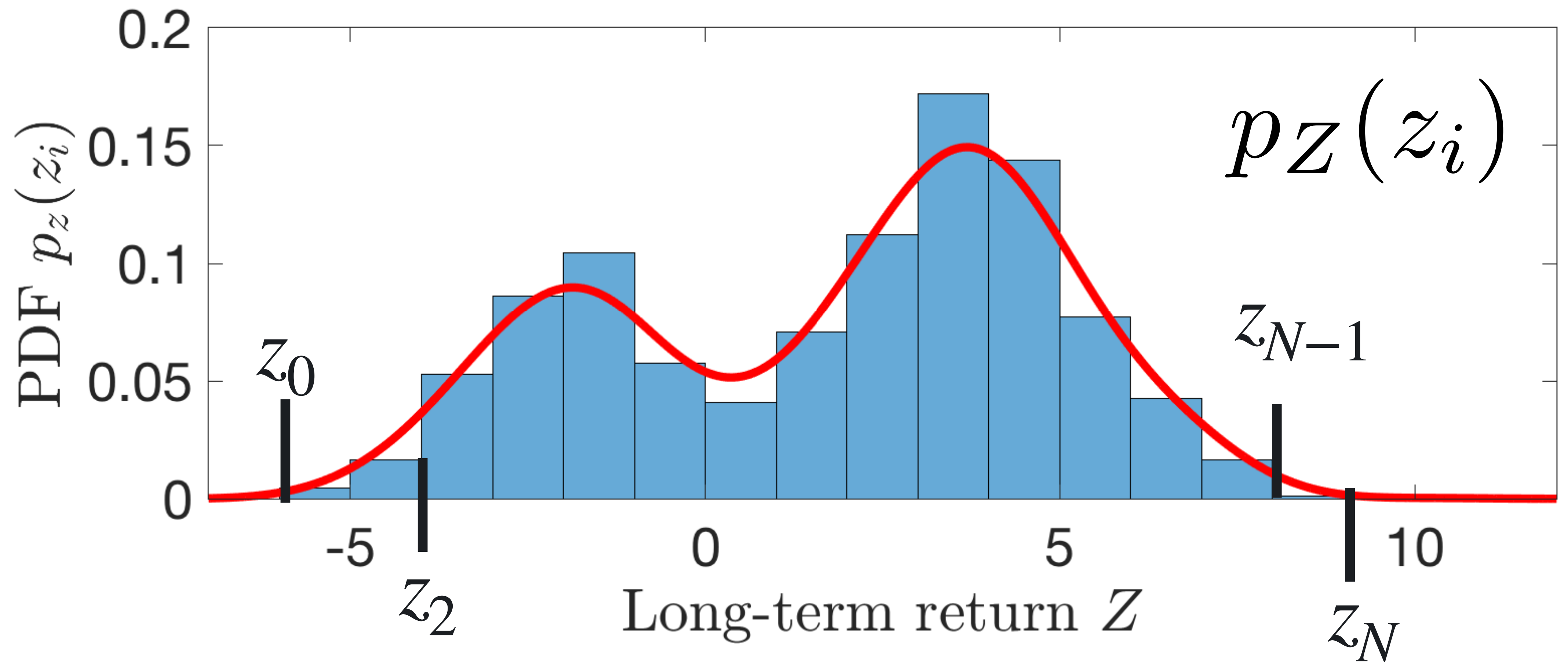
OR



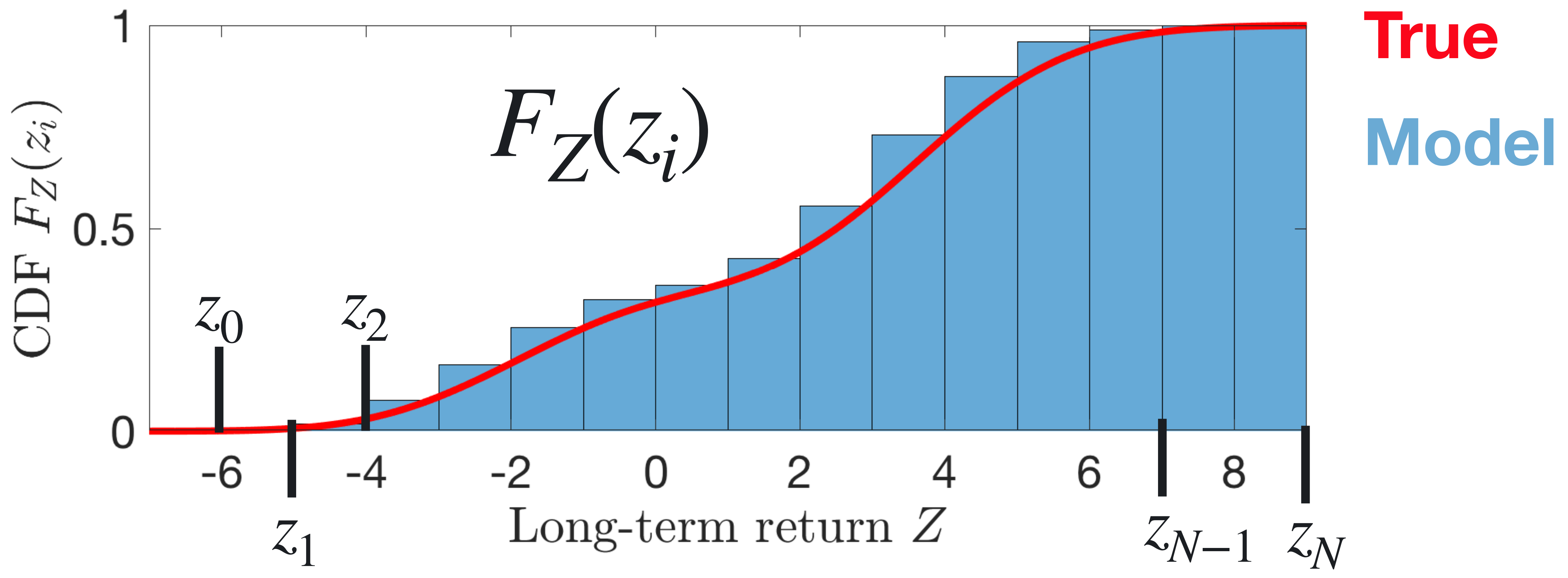
1. Categorical

- Fixed support bins $z_0, z_1, z_2, \ldots, z_N$



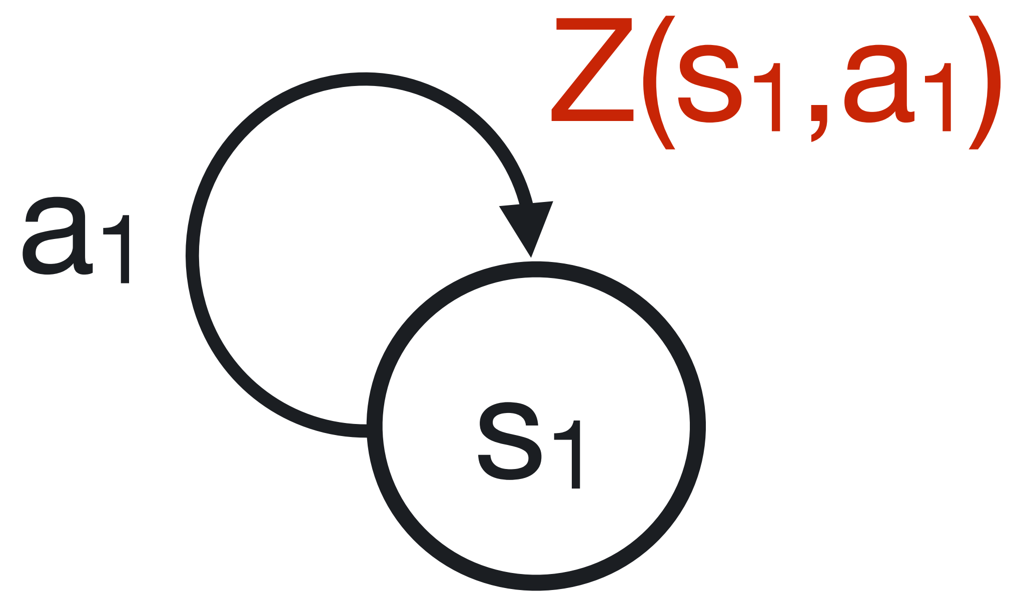


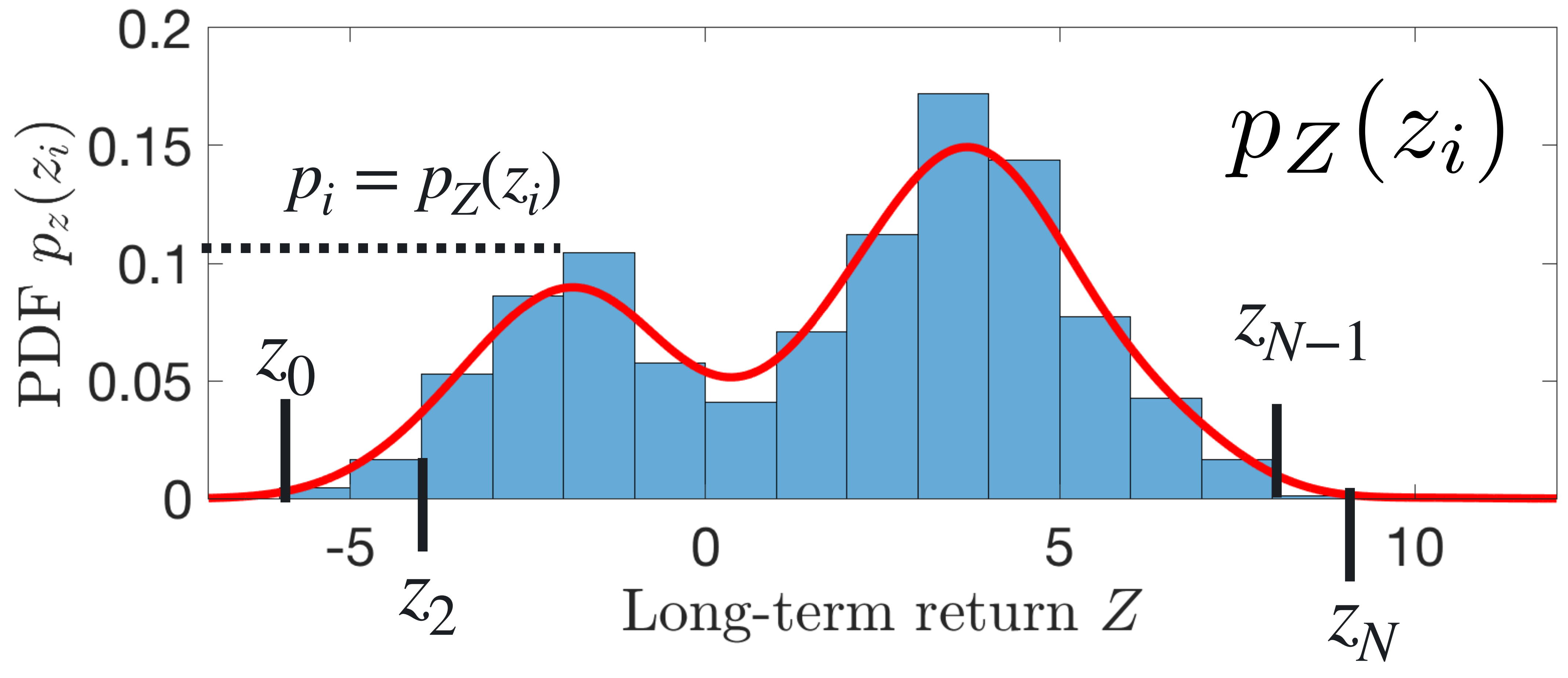
OR



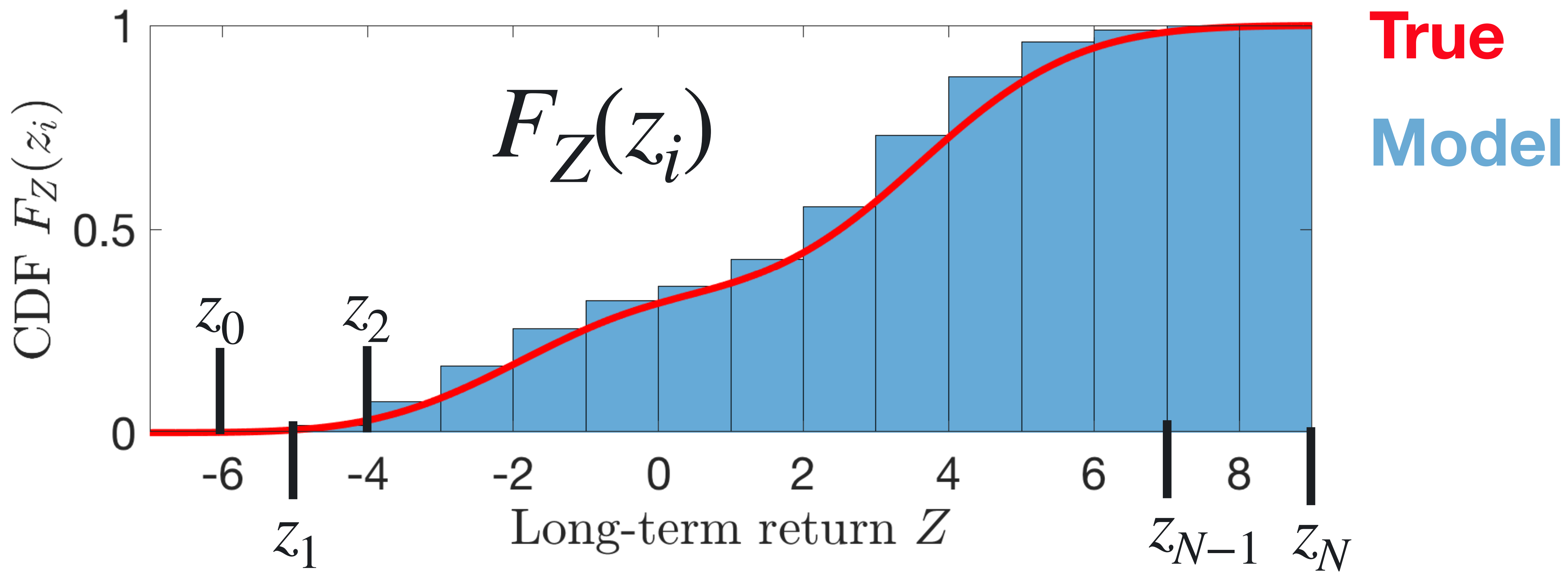
1. Categorical

- Fixed support bins $z_0, z_1, z_2, \ldots, z_N$
- Learn probabilities/quantiles p_i, τ_i



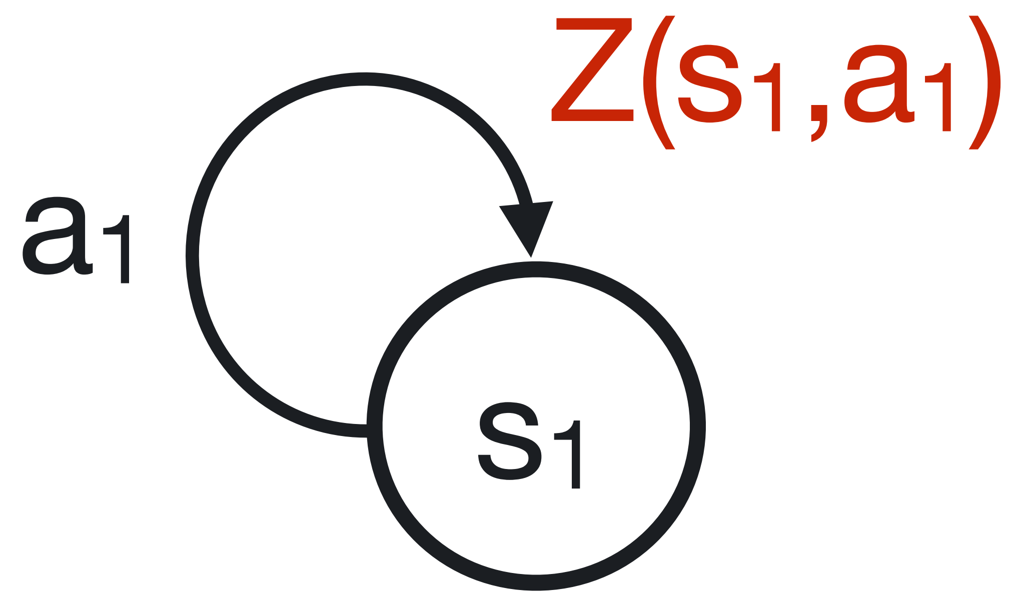


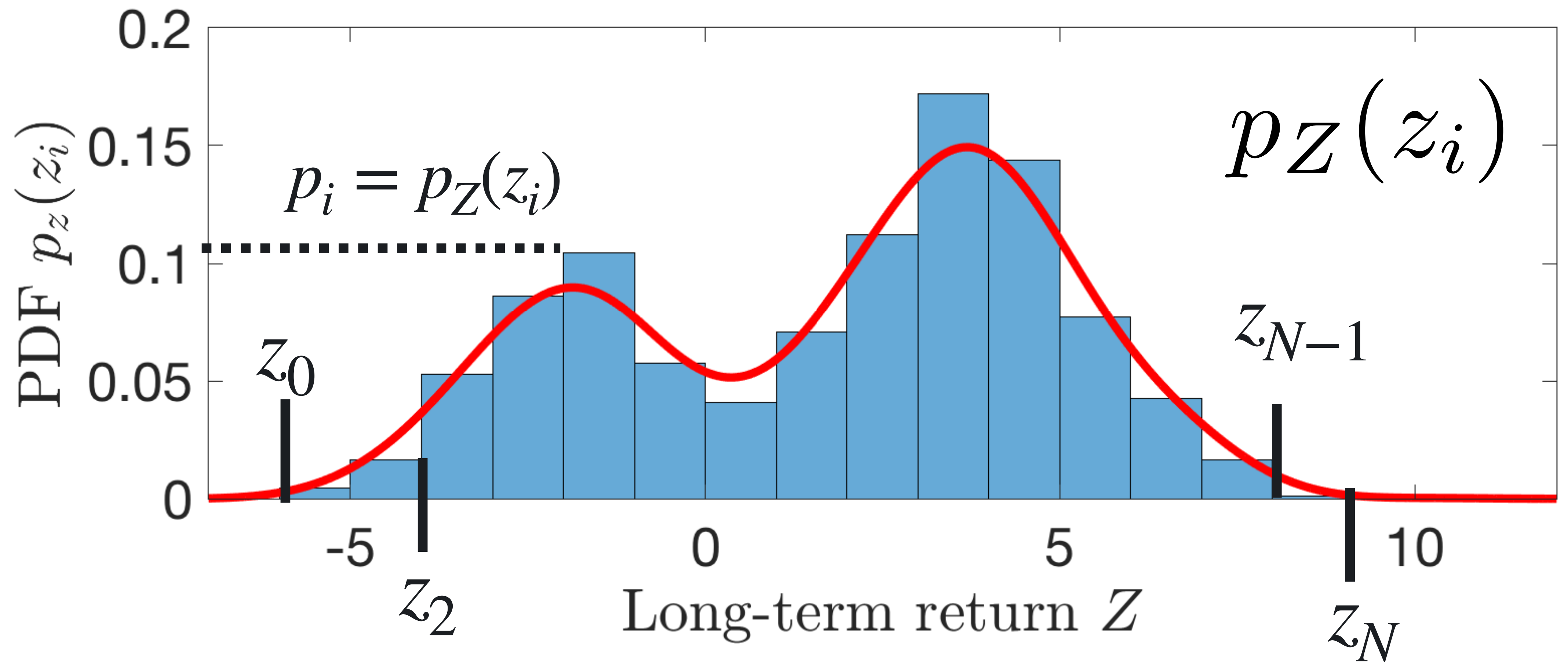
OR



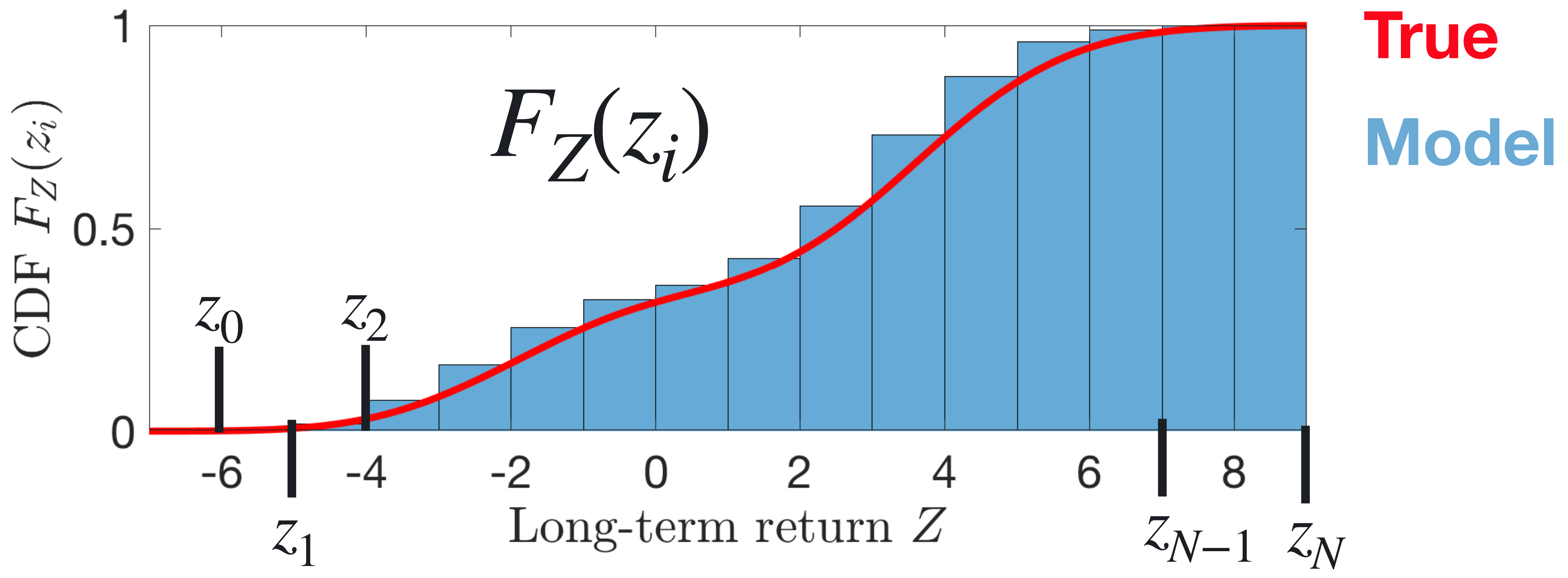
1. Categorical

- Fixed support bins $z_0, z_1, z_2, \ldots, z_N$
- Learn probabilities/quantiles p_i, τ_i
- Easy to program



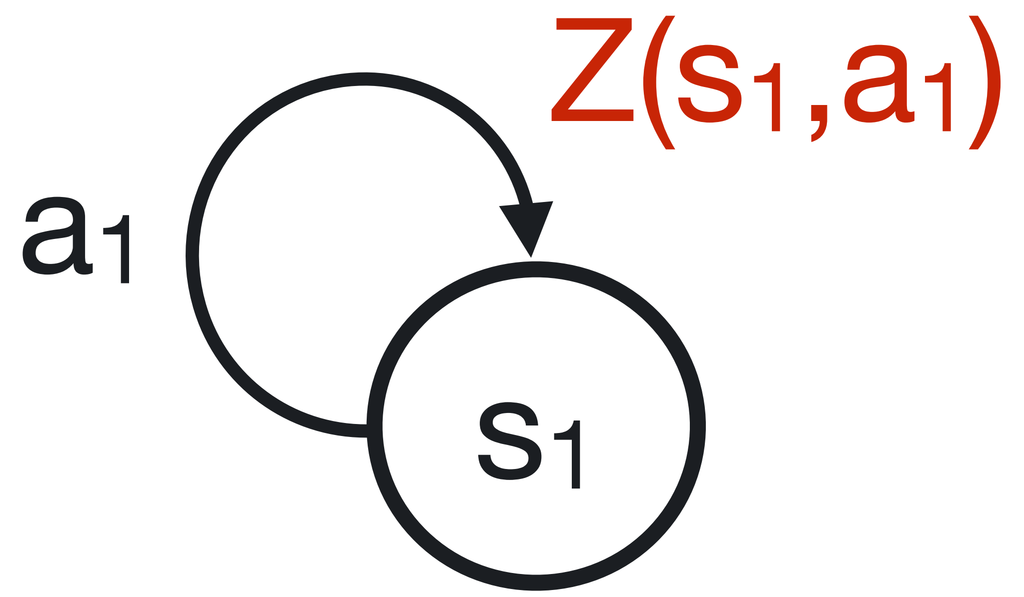


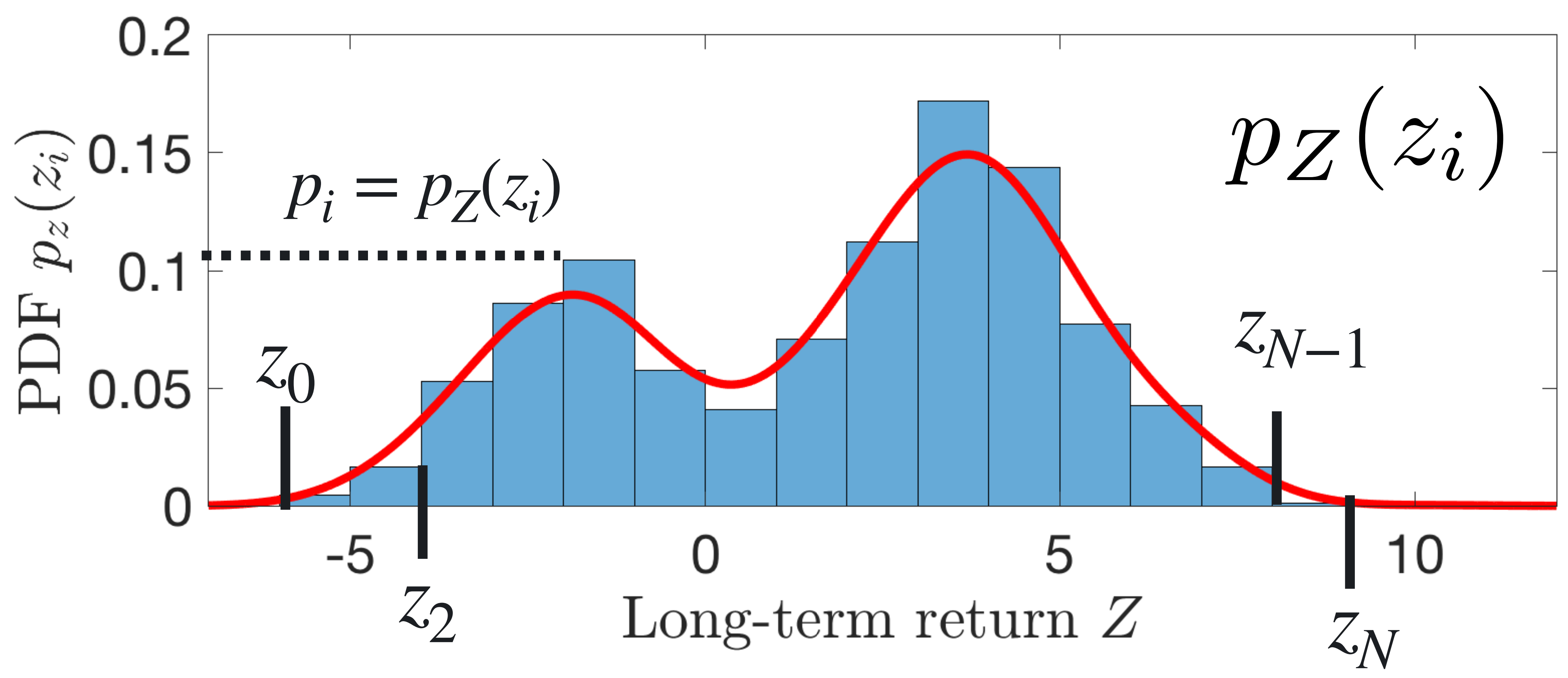
OR



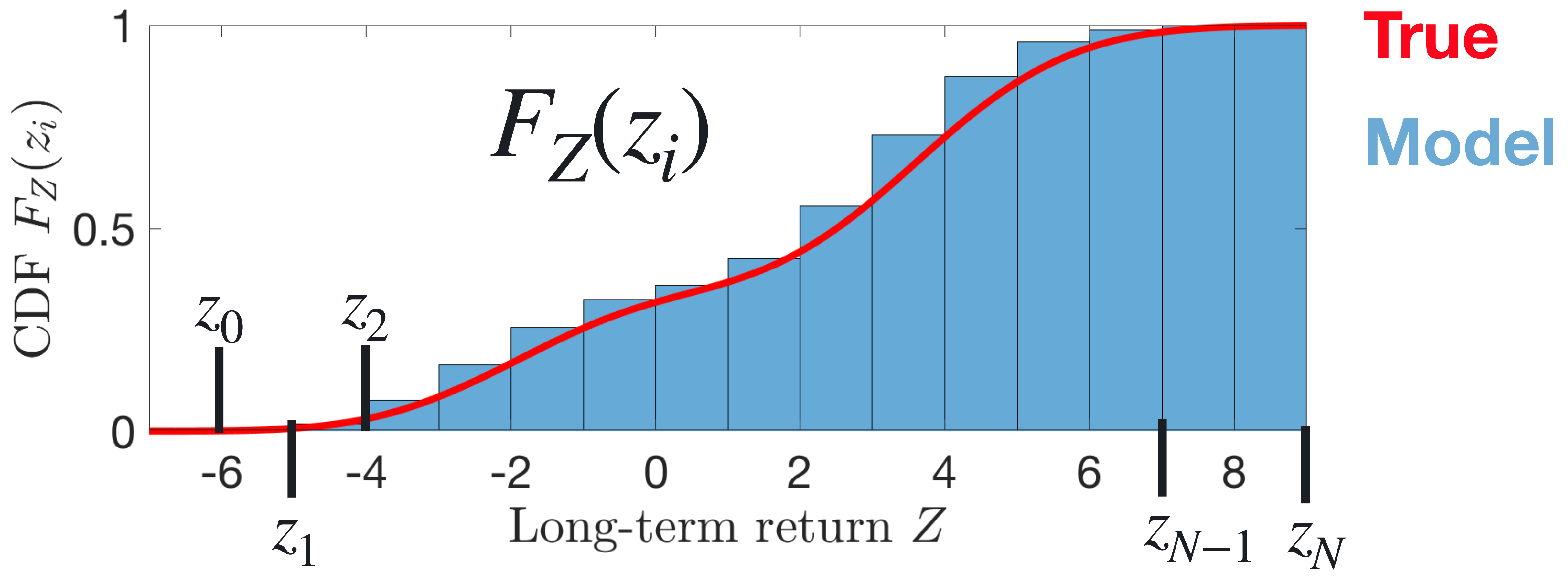
1. Categorical

- Fixed support bins $z_0, z_1, z_2, \ldots, z_N$
- Learn probabilities/quantiles p_i, τ_i
- Easy to program
- Value range of Z needs to be known!



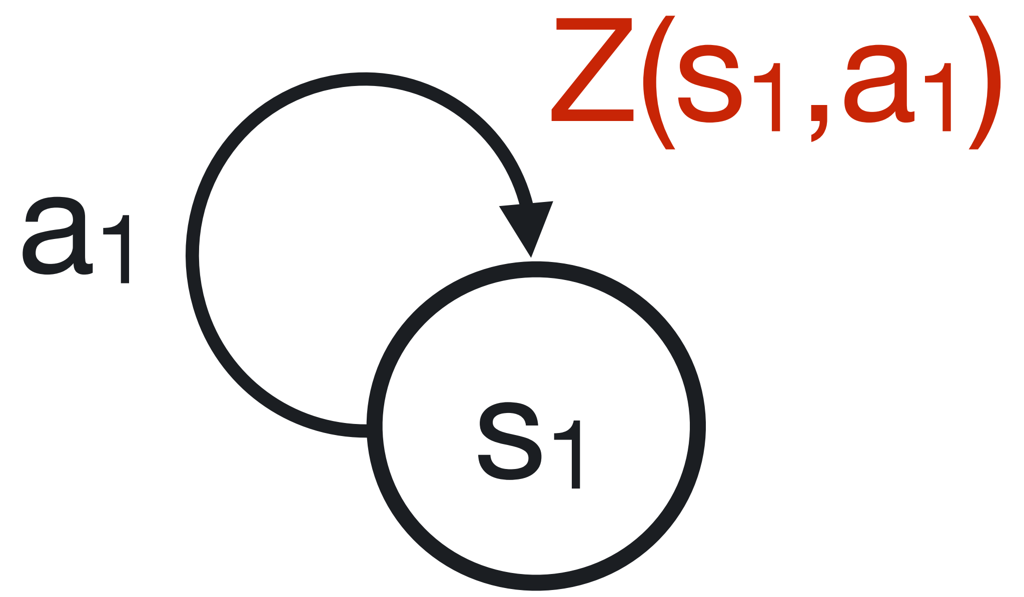


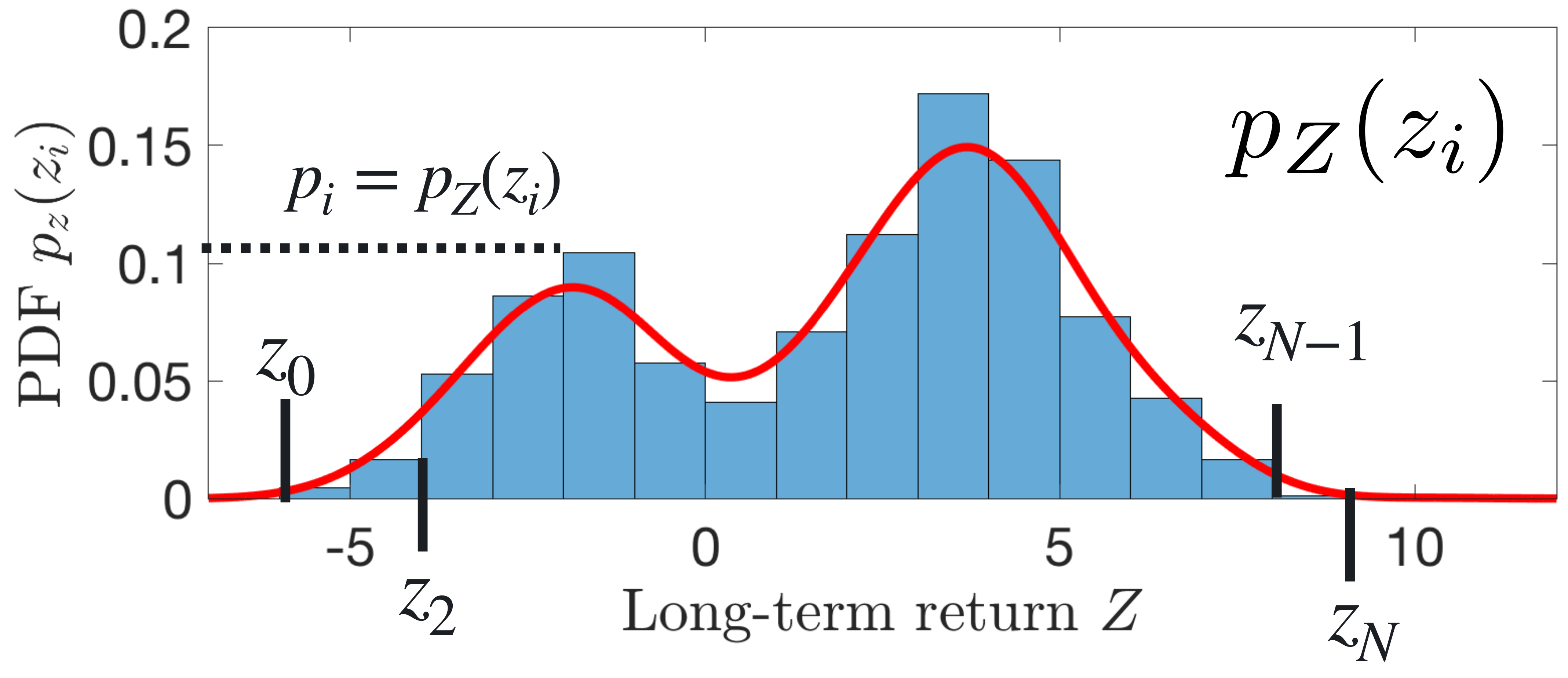
OR



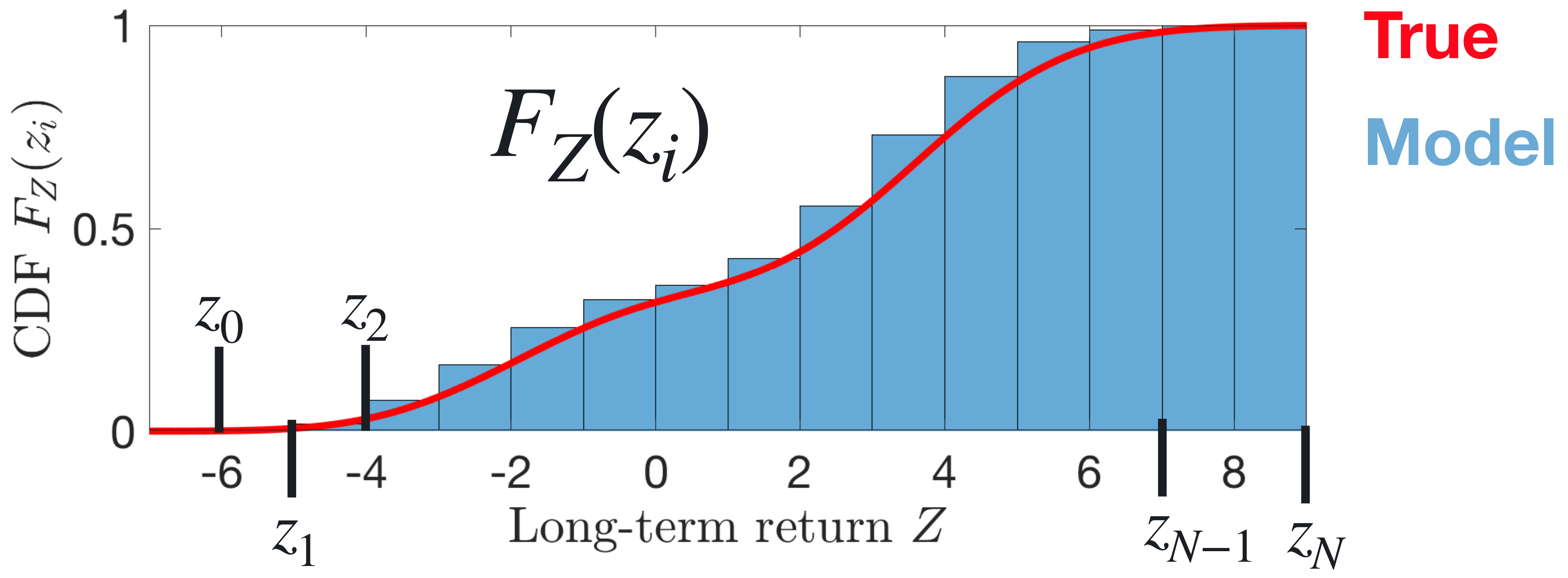
1. Categorical

- Fixed support bins $z_0, z_1, z_2, \ldots, z_N$
- Learn probabilities/quantiles p_i, τ_i
- Easy to program
- Value range of Z needs to be known!
- The bins are fixed (less expressive)

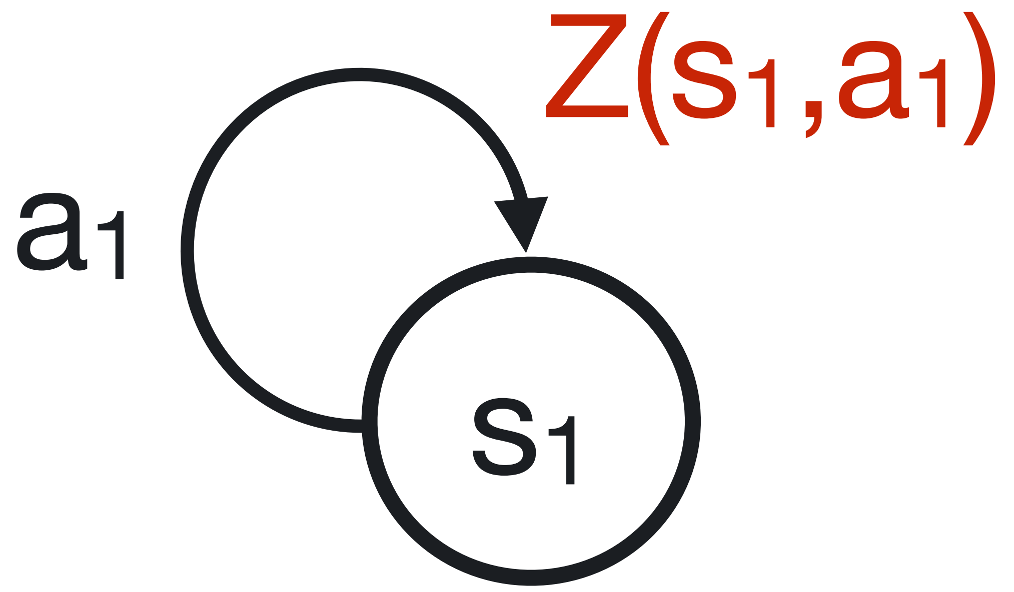




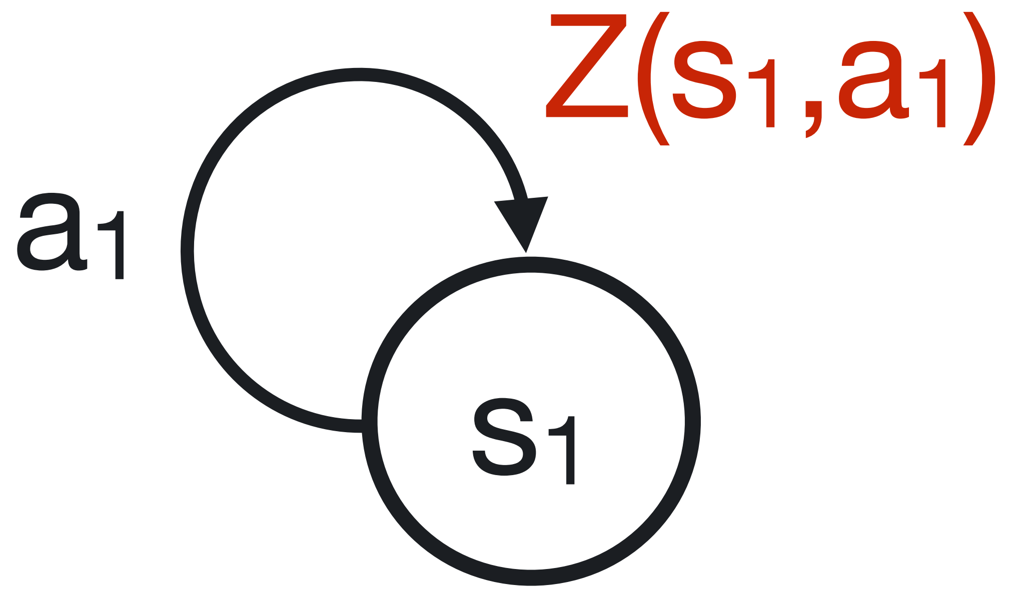
OR



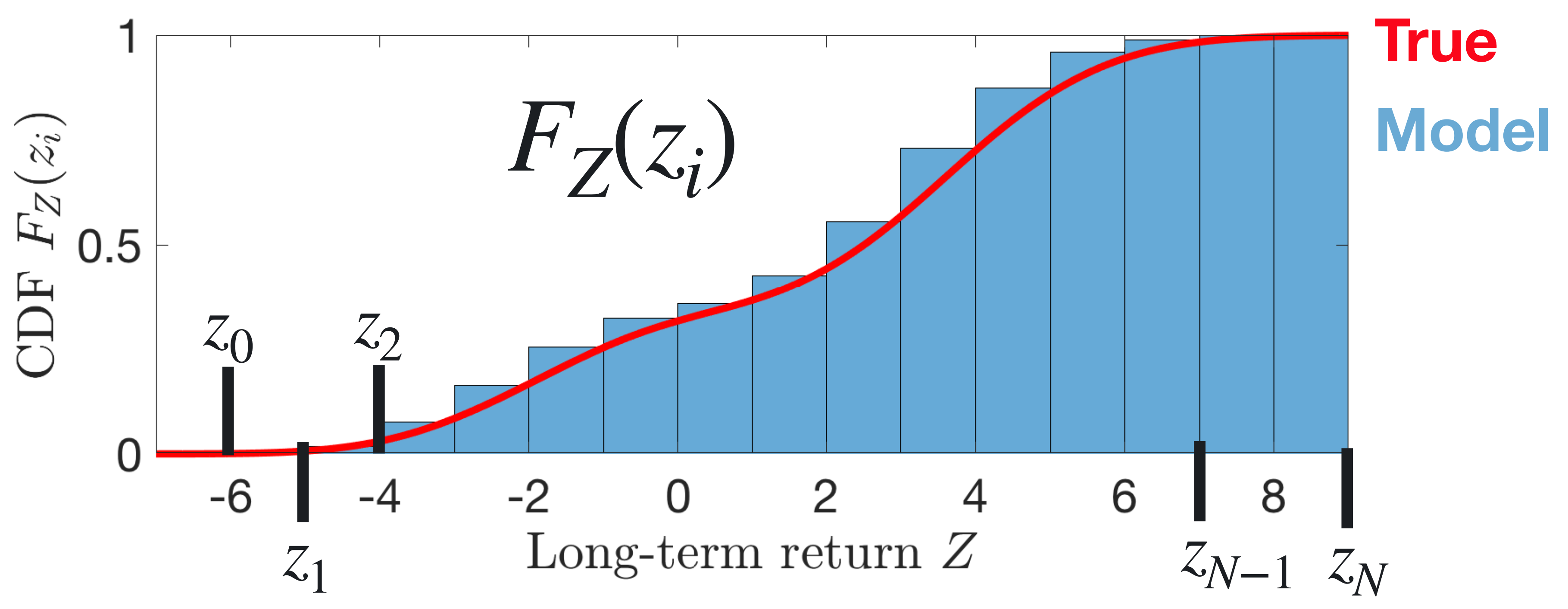
2. Quantiles of Inverse CDF



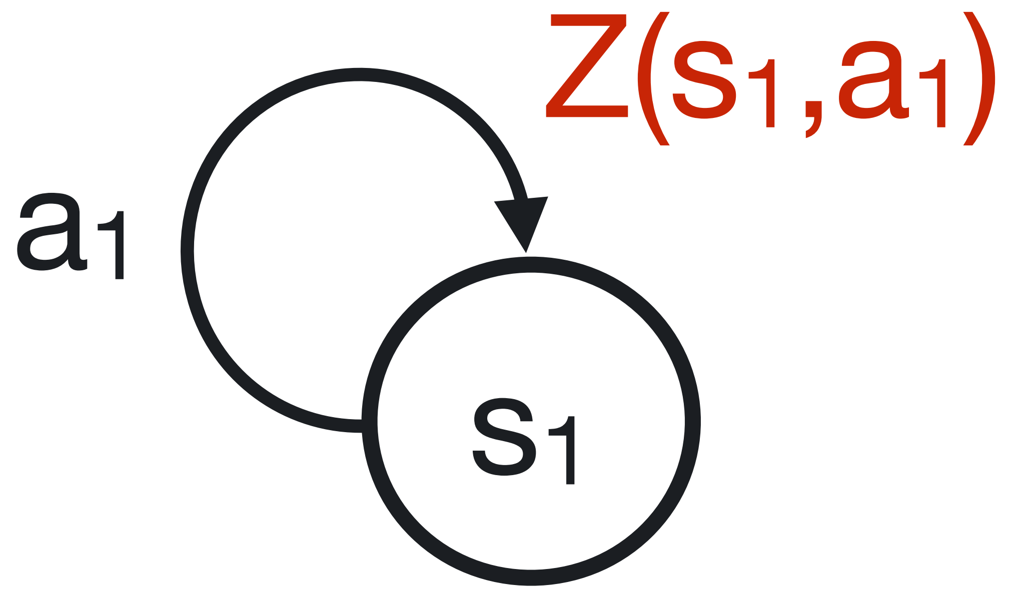
2. Quantiles of Inverse CDF



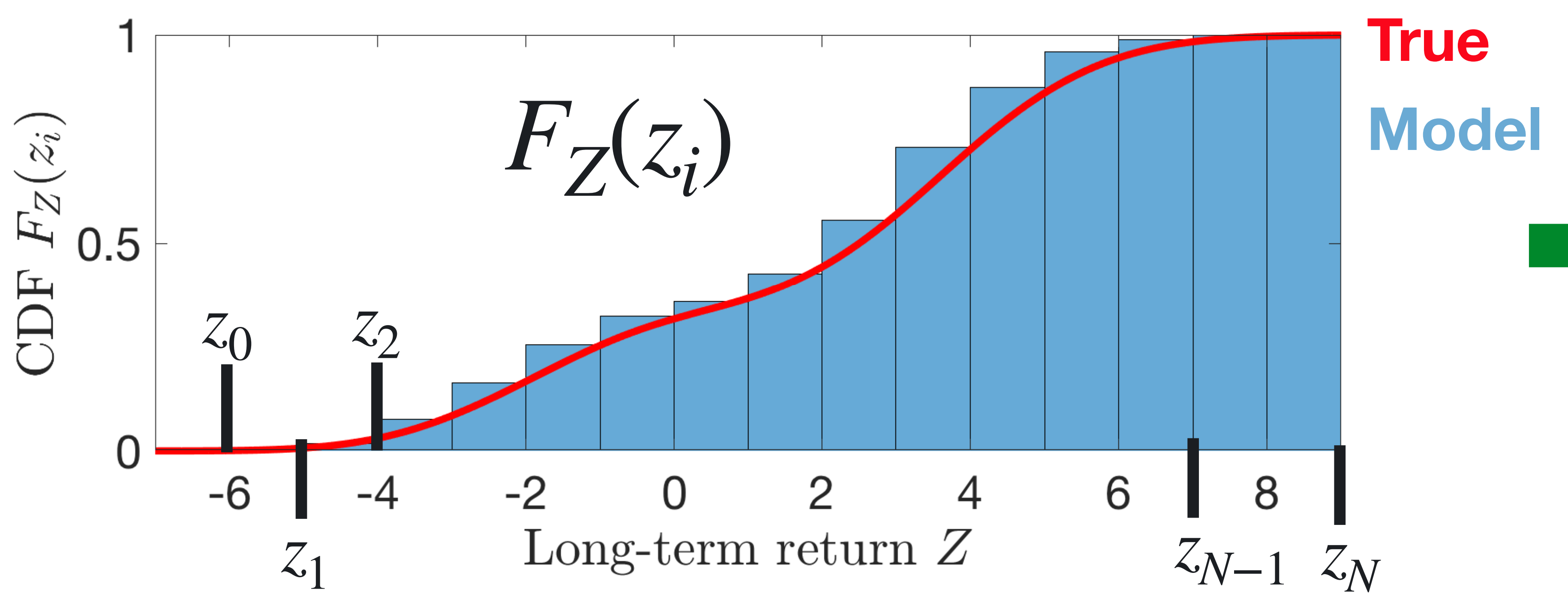
Categorical



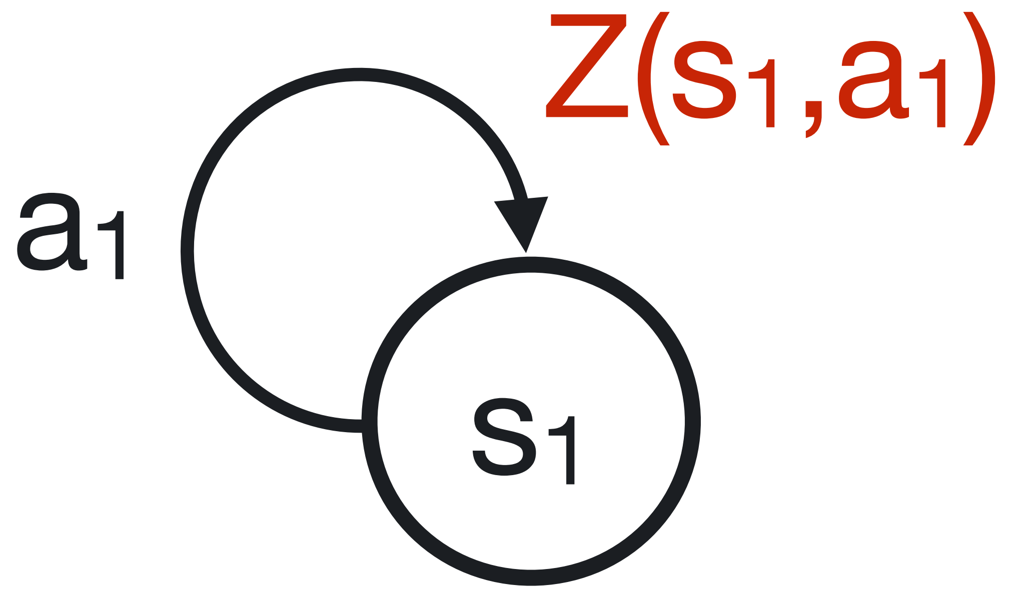
2. Quantiles of Inverse CDF



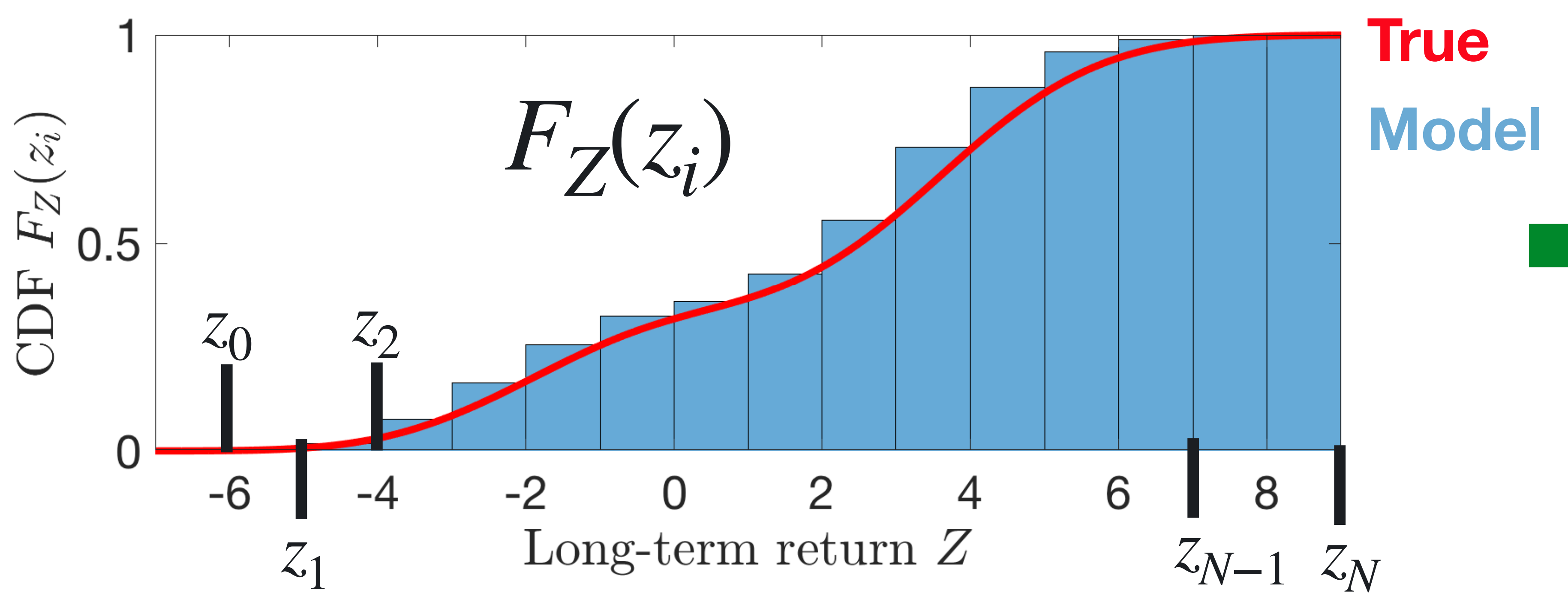
Categorical

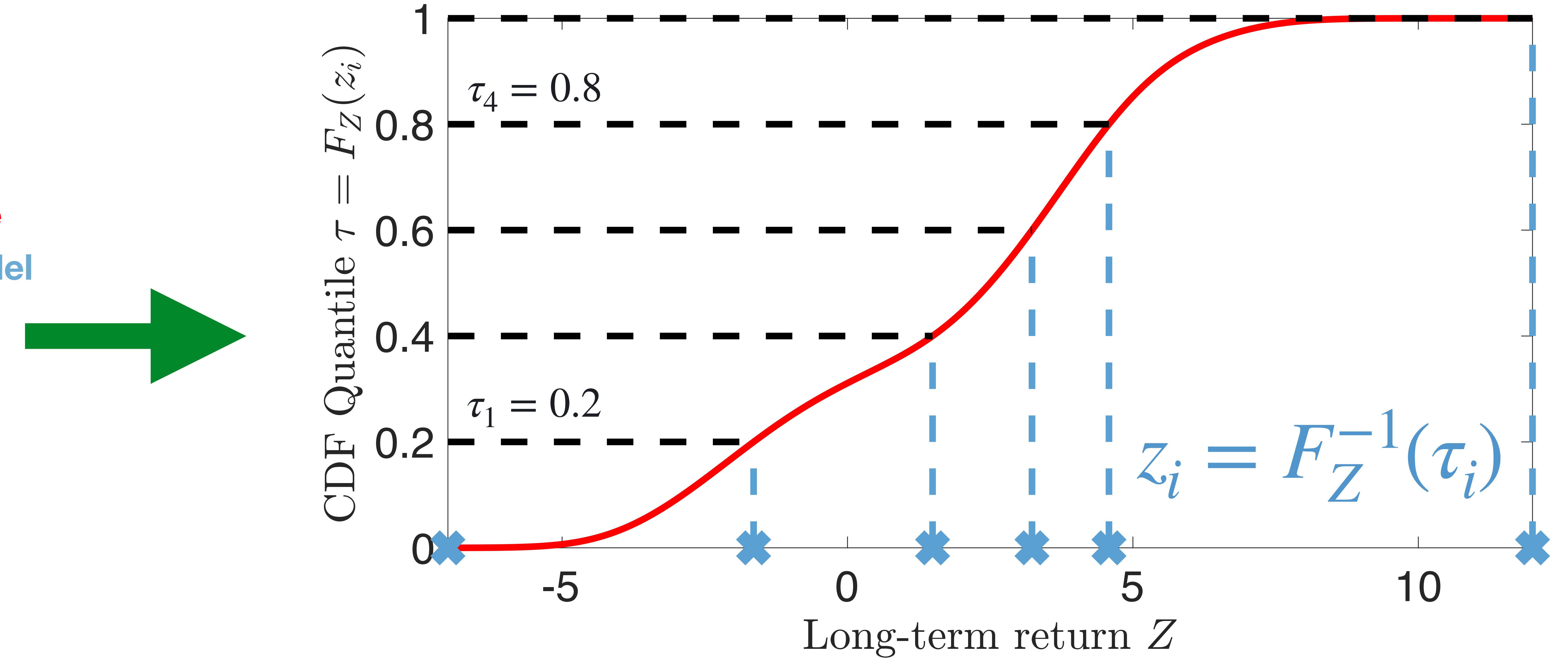


2. Quantiles of Inverse CDF



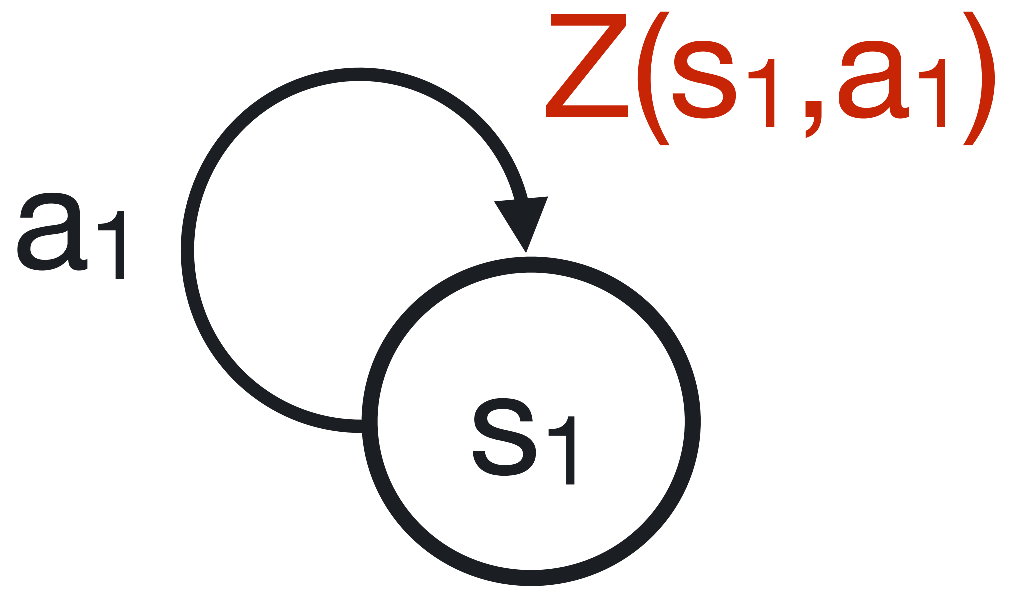
Categorical



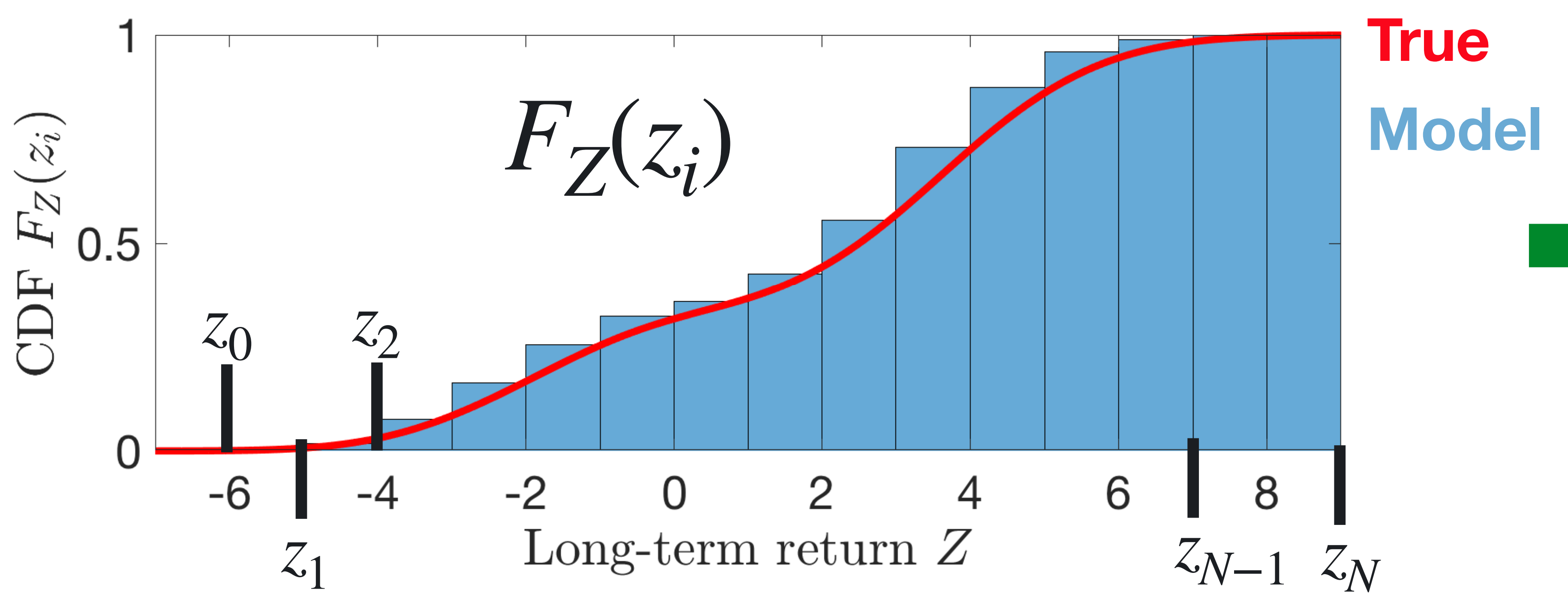


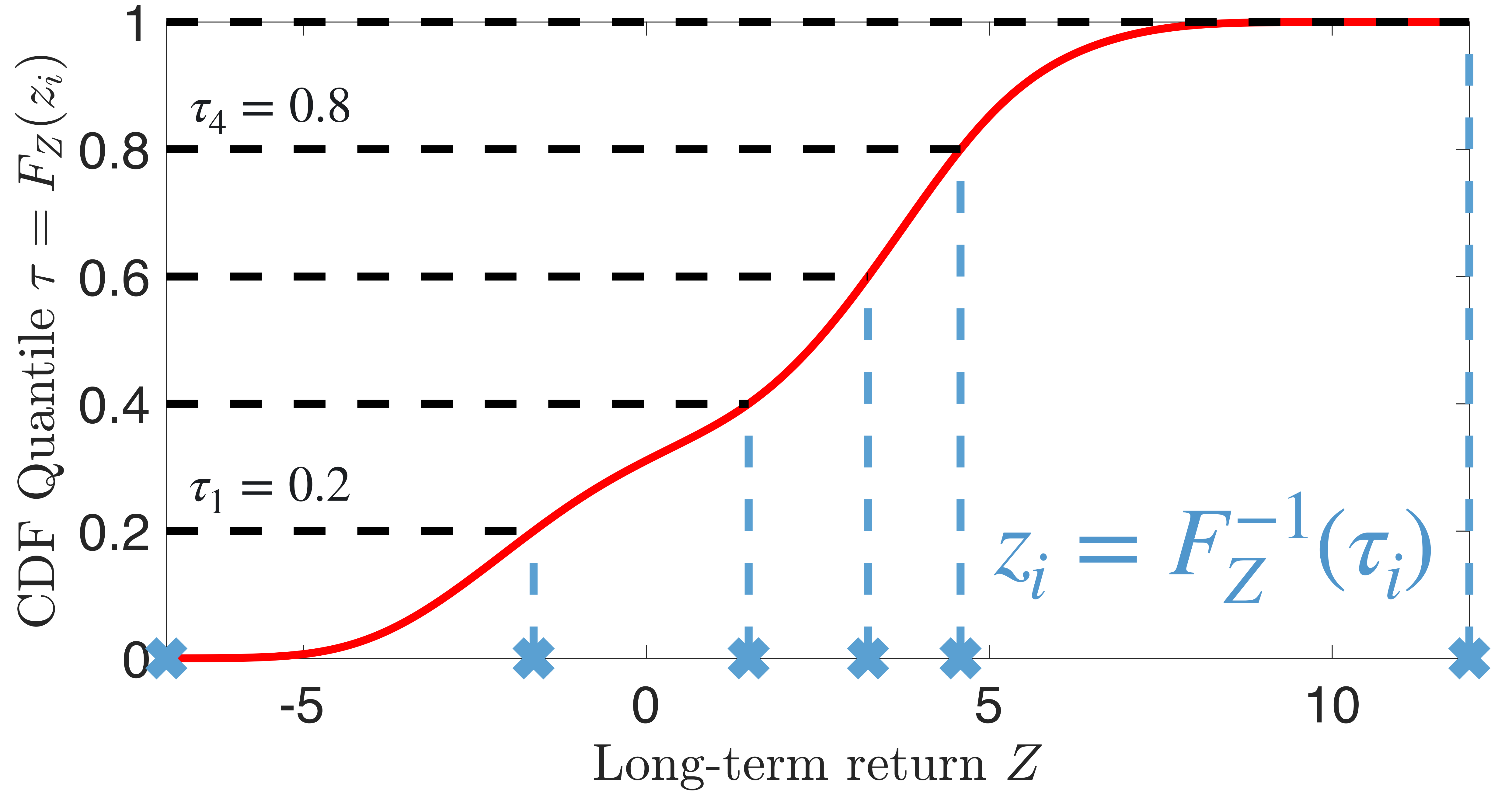
2. Quantiles of Inverse CDF

- Fixed quantile bins $\tau_0, \tau_1, \tau_2, \ldots, \tau_N$



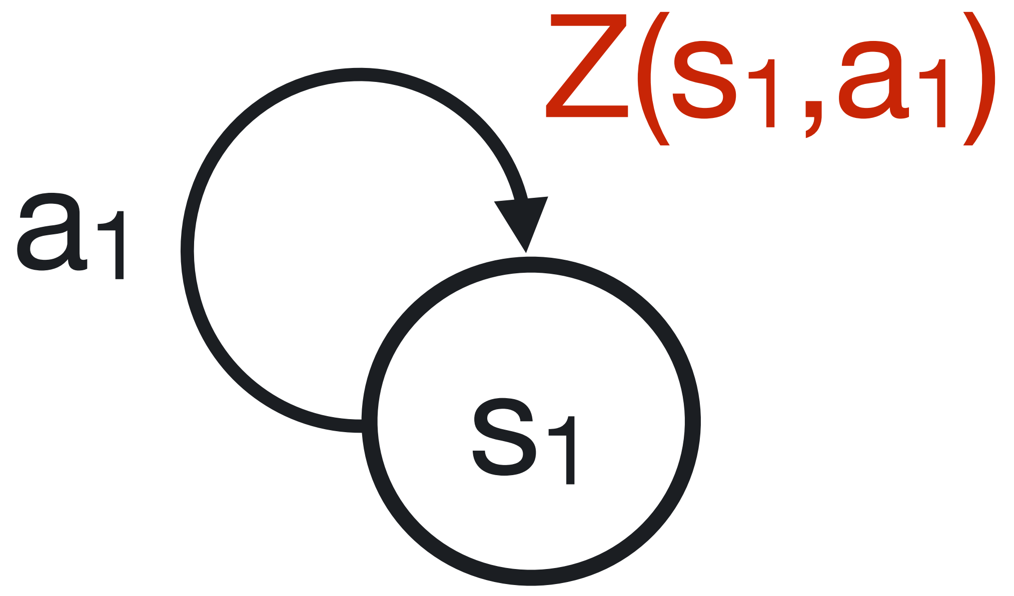
Categorical



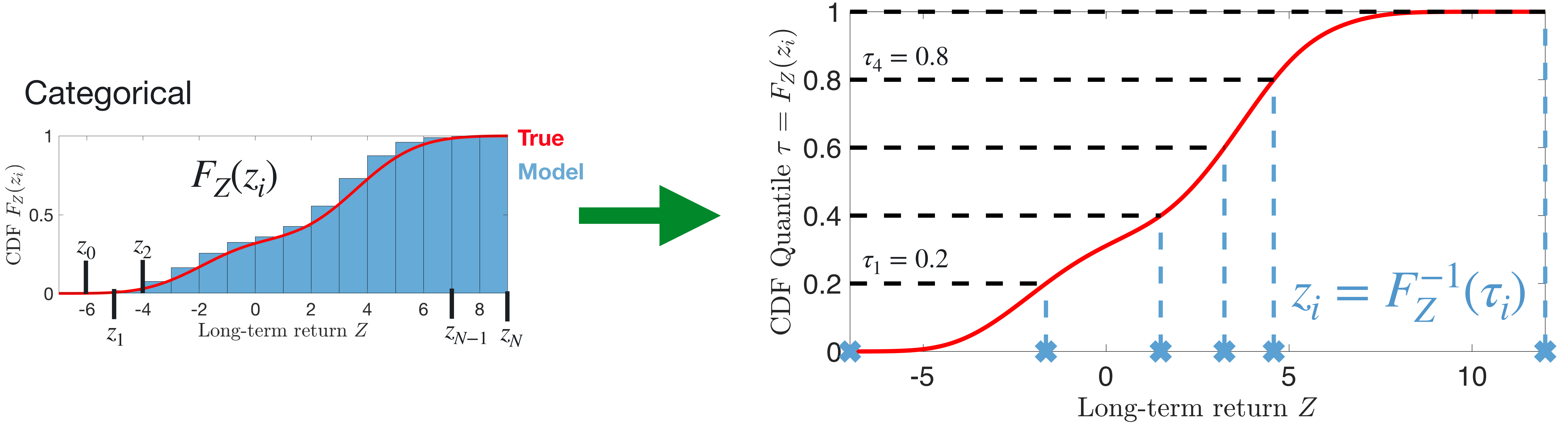


2. Quantiles of Inverse CDF

- Fixed quantile bins $\tau_0, \tau_1, \tau_2, \ldots, \tau_N$
- Learn support $z_i = F_Z^{-1}(\tau_i)$

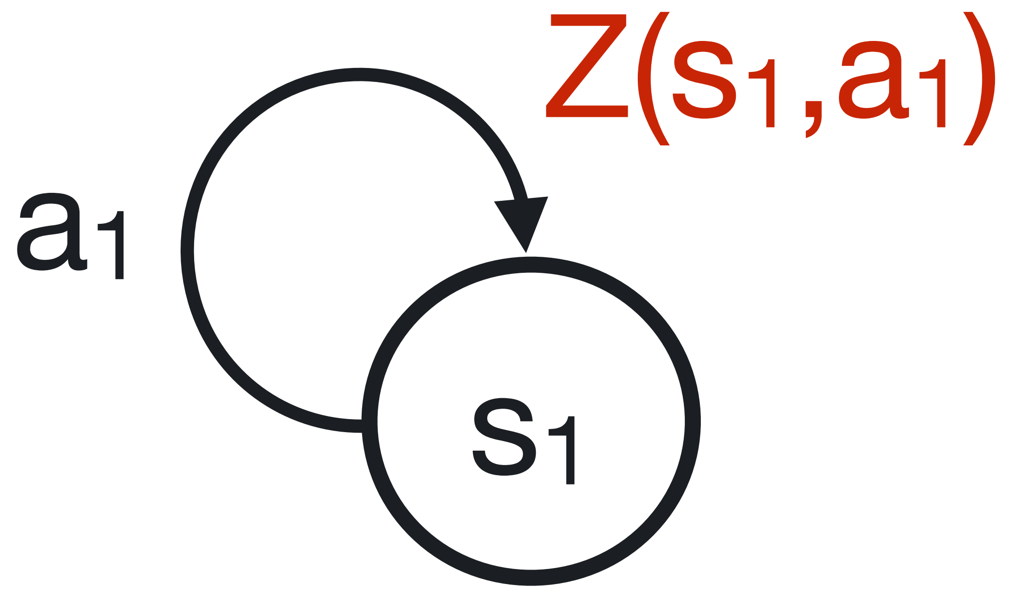


Categorical

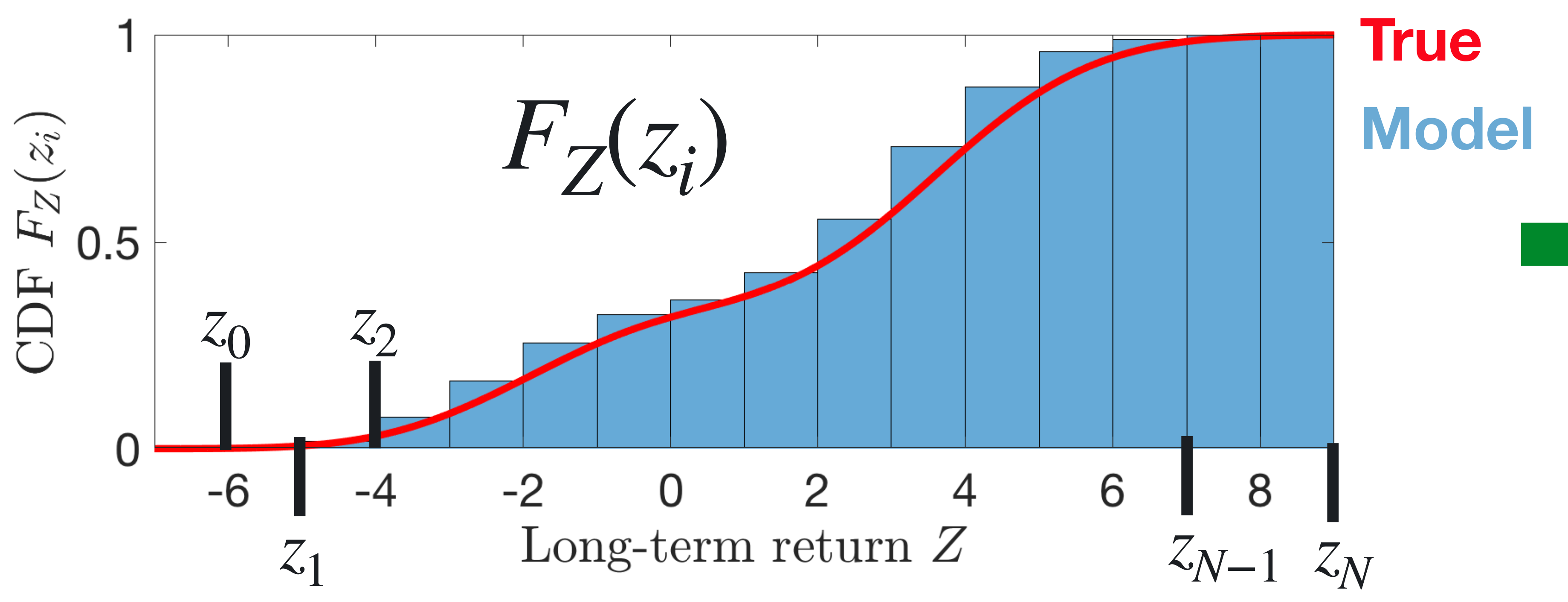


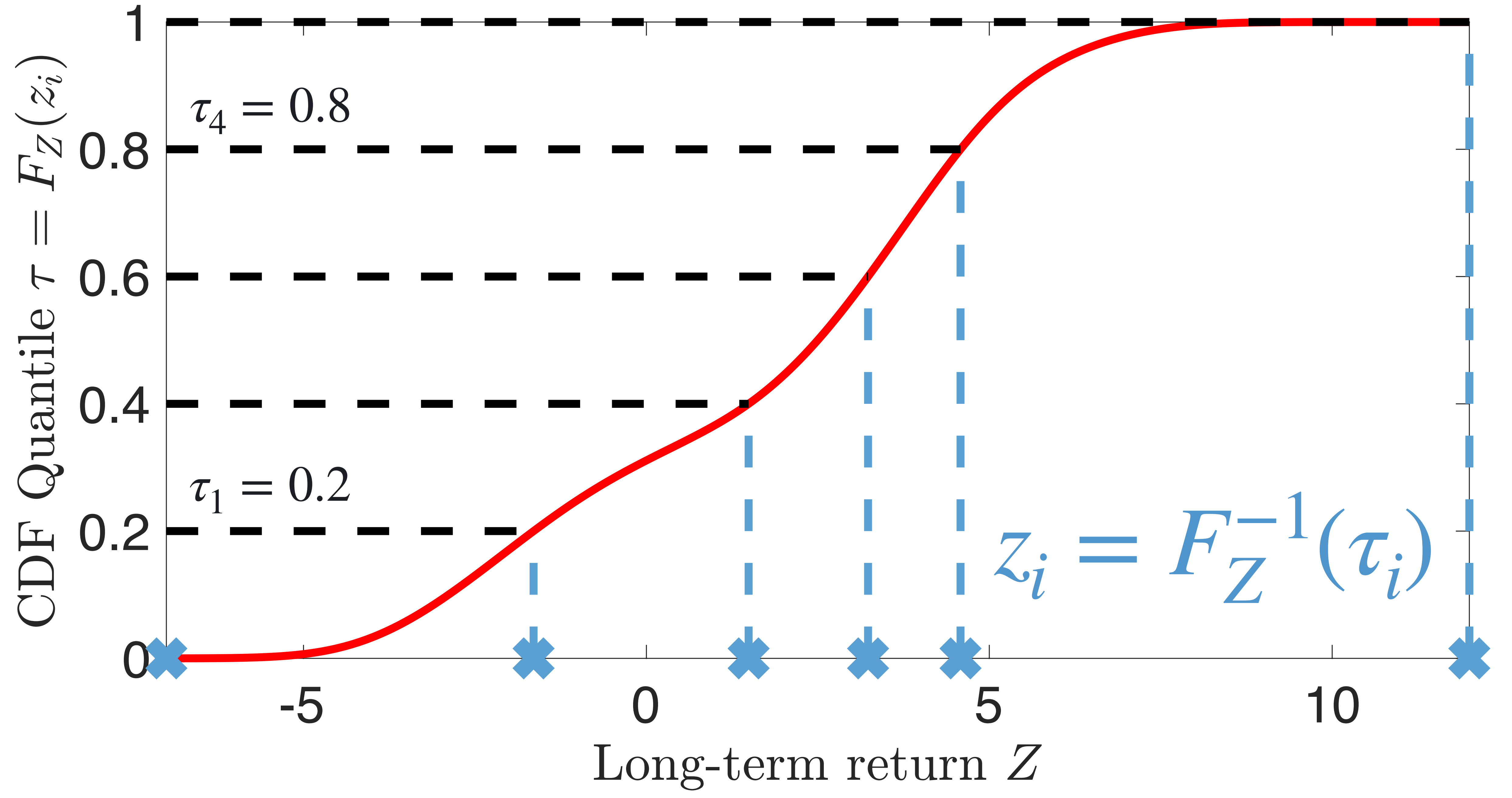
2. Quantiles of Inverse CDF

- Fixed quantile bins $\tau_0, \tau_1, \tau_2, \ldots, \tau_N$
- Learn support $z_i = F_Z^{-1}(\tau_i)$
- No need for value range of Z



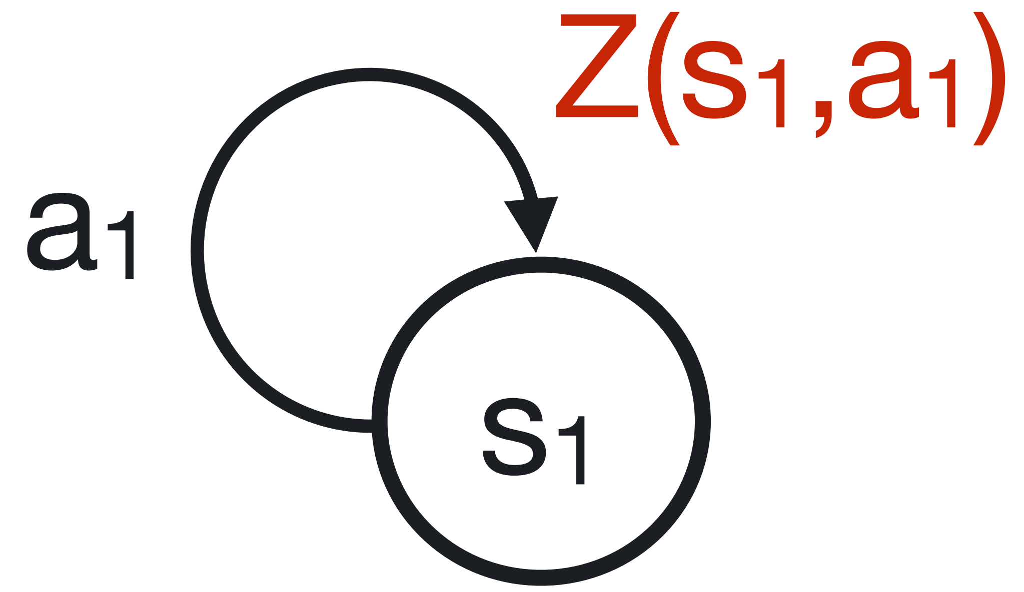
Categorical



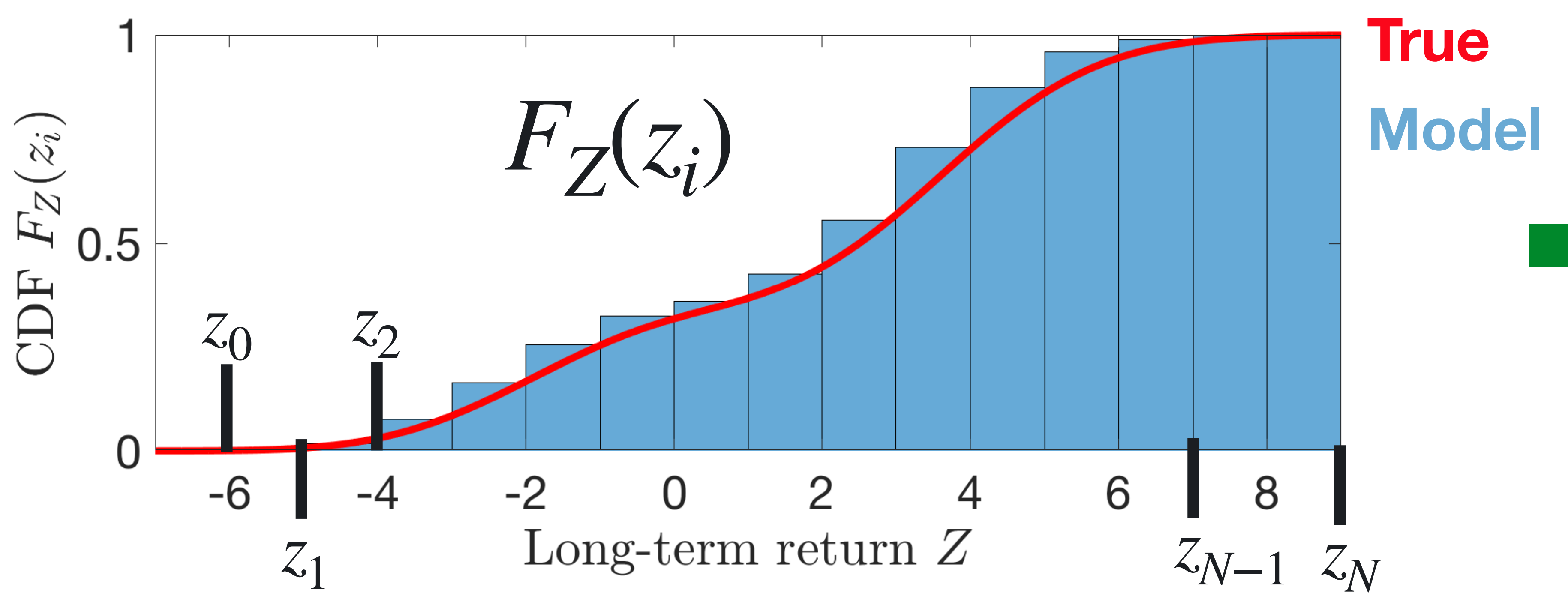


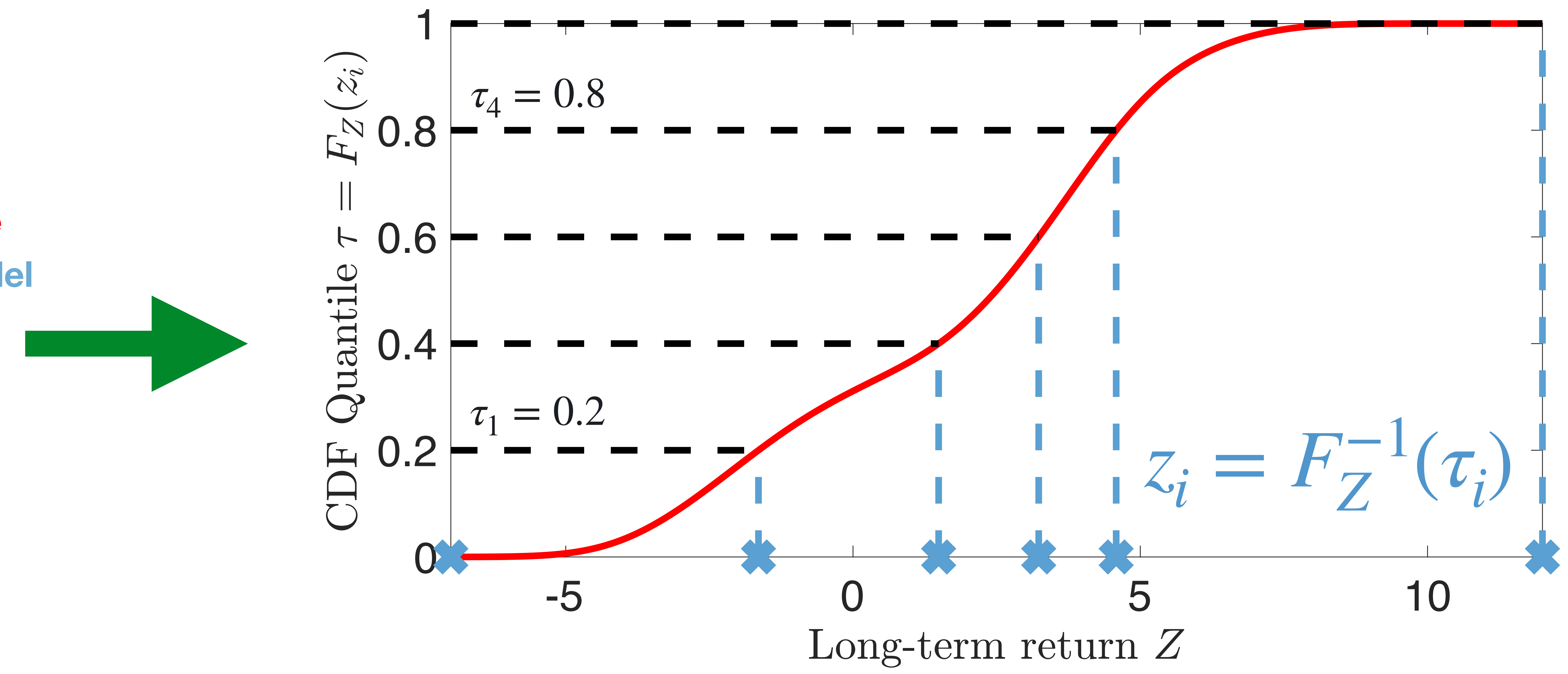
2. Quantiles of Inverse CDF

- Fixed quantile bins $\tau_0, \tau_1, \tau_2, \ldots, \tau_N$
- Learn support $z_i = F_Z^{-1}(\tau_i)$
- No need for value range of Z
- Training loss

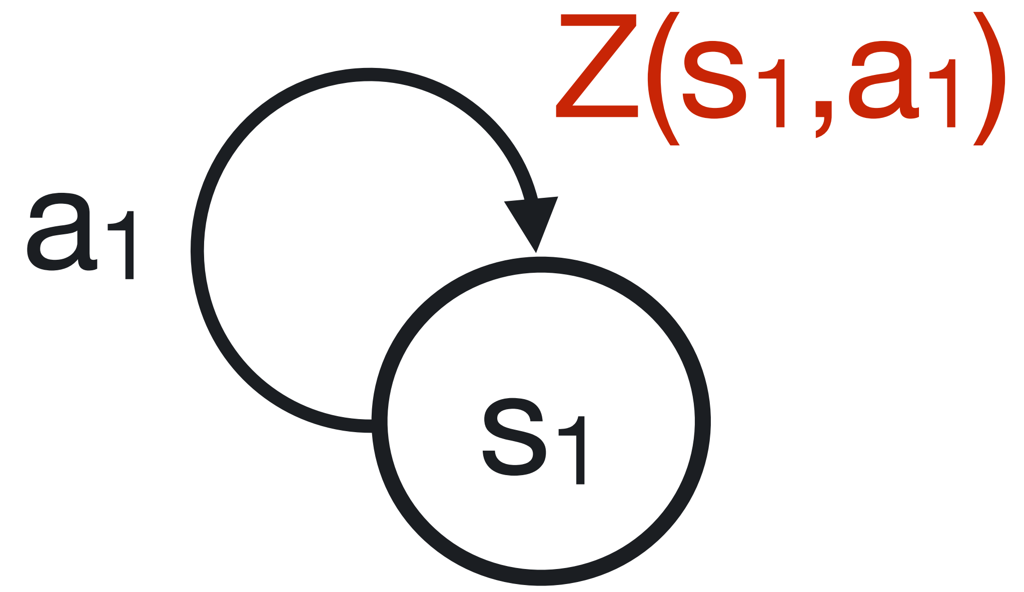


Categorical



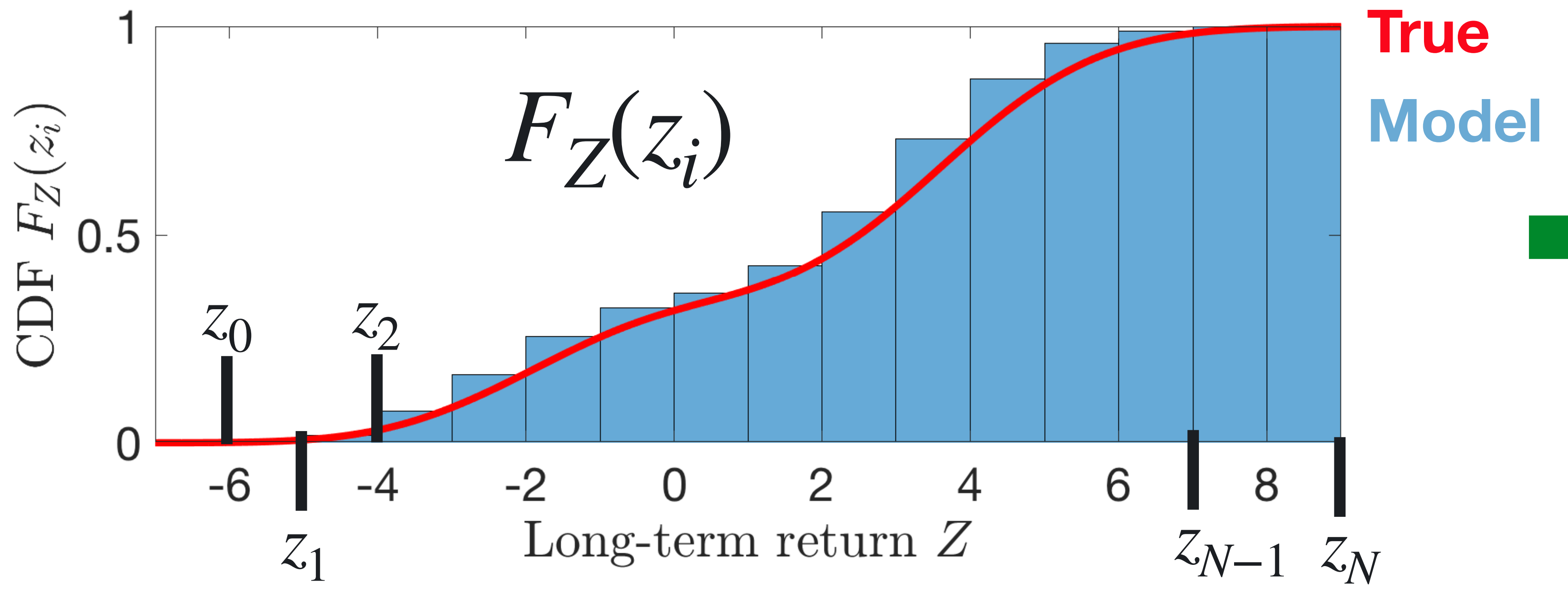


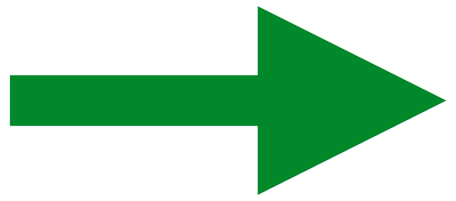
2. Quantiles of Inverse CDF

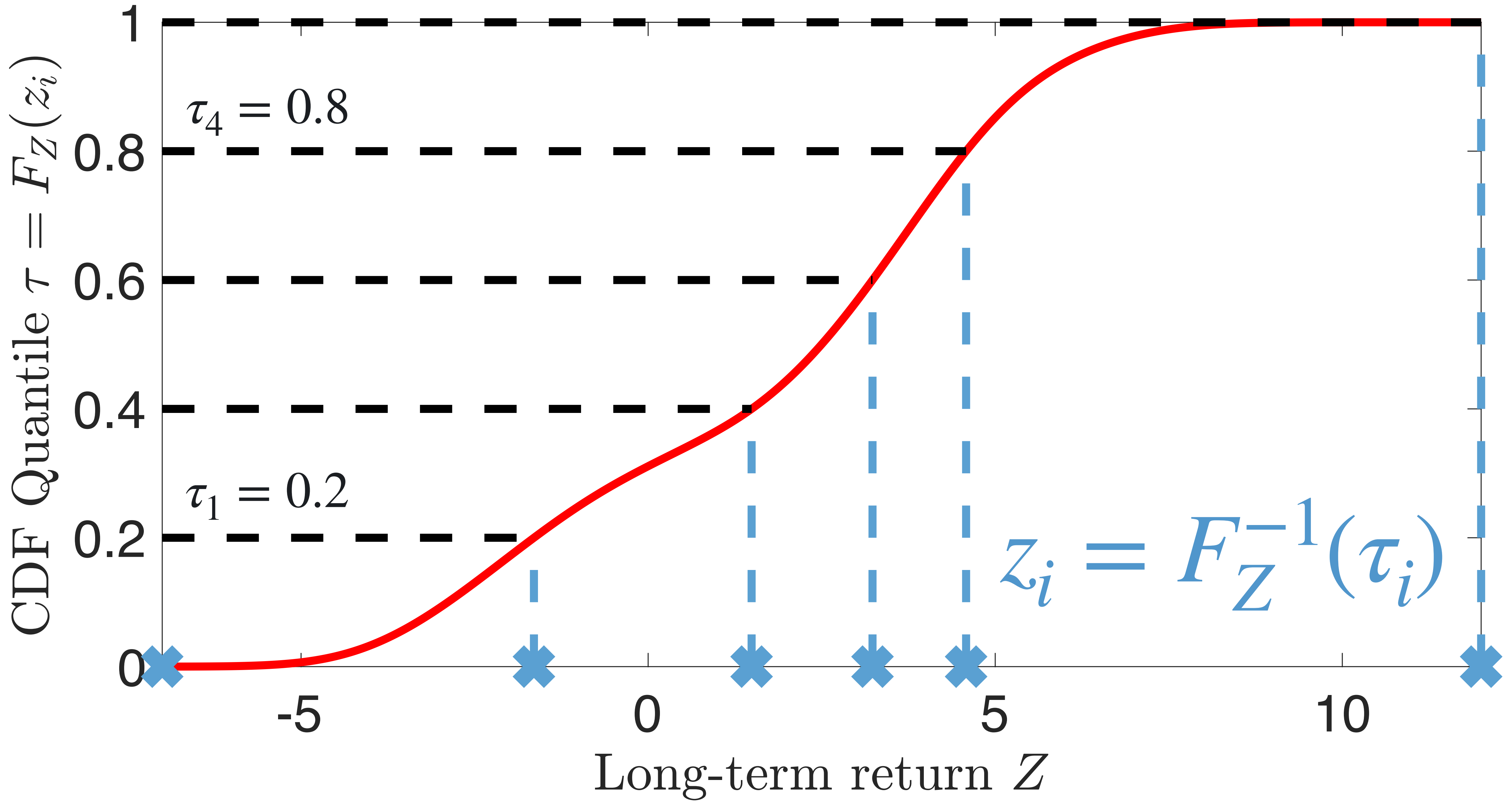


- Fixed quantile bins $\tau_0, \tau_1, \tau_2, \ldots, \tau_N$
- Learn support $z_i = F_Z^{-1}(\tau_i)$
- No need for value range of Z
- Training loss
- Fixed quantile bins

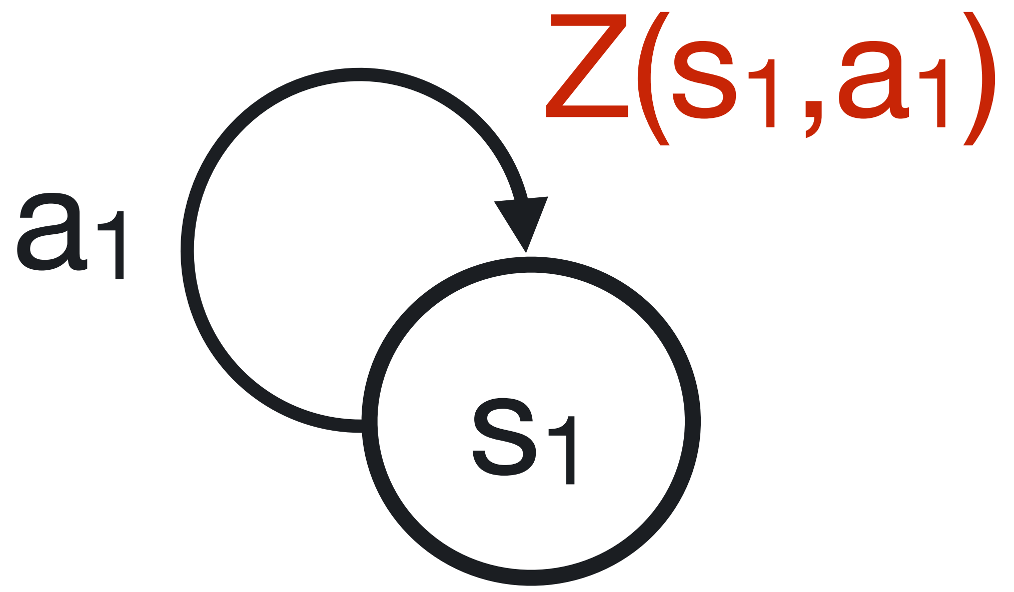
Categorical



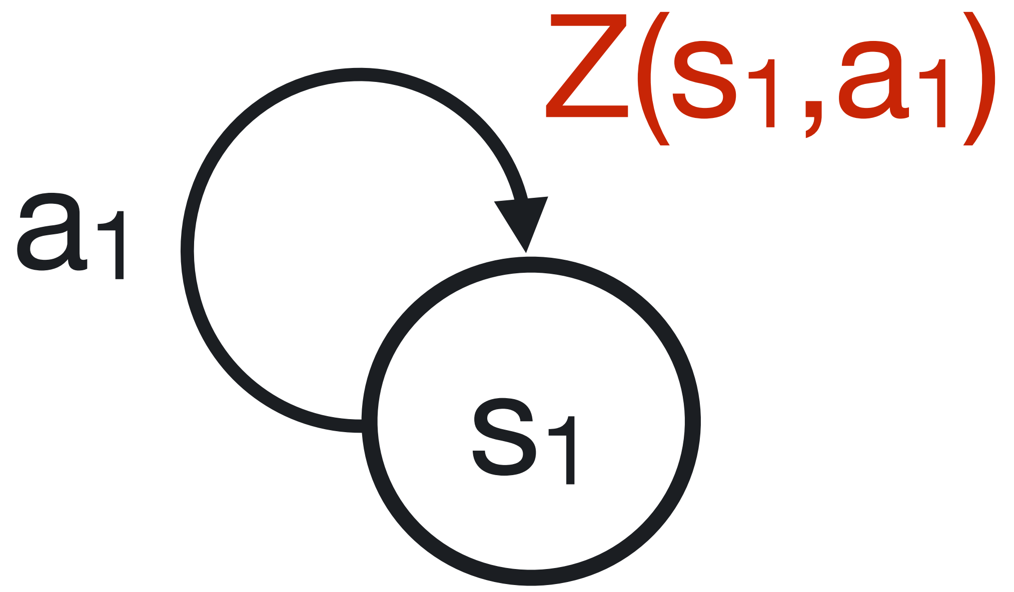


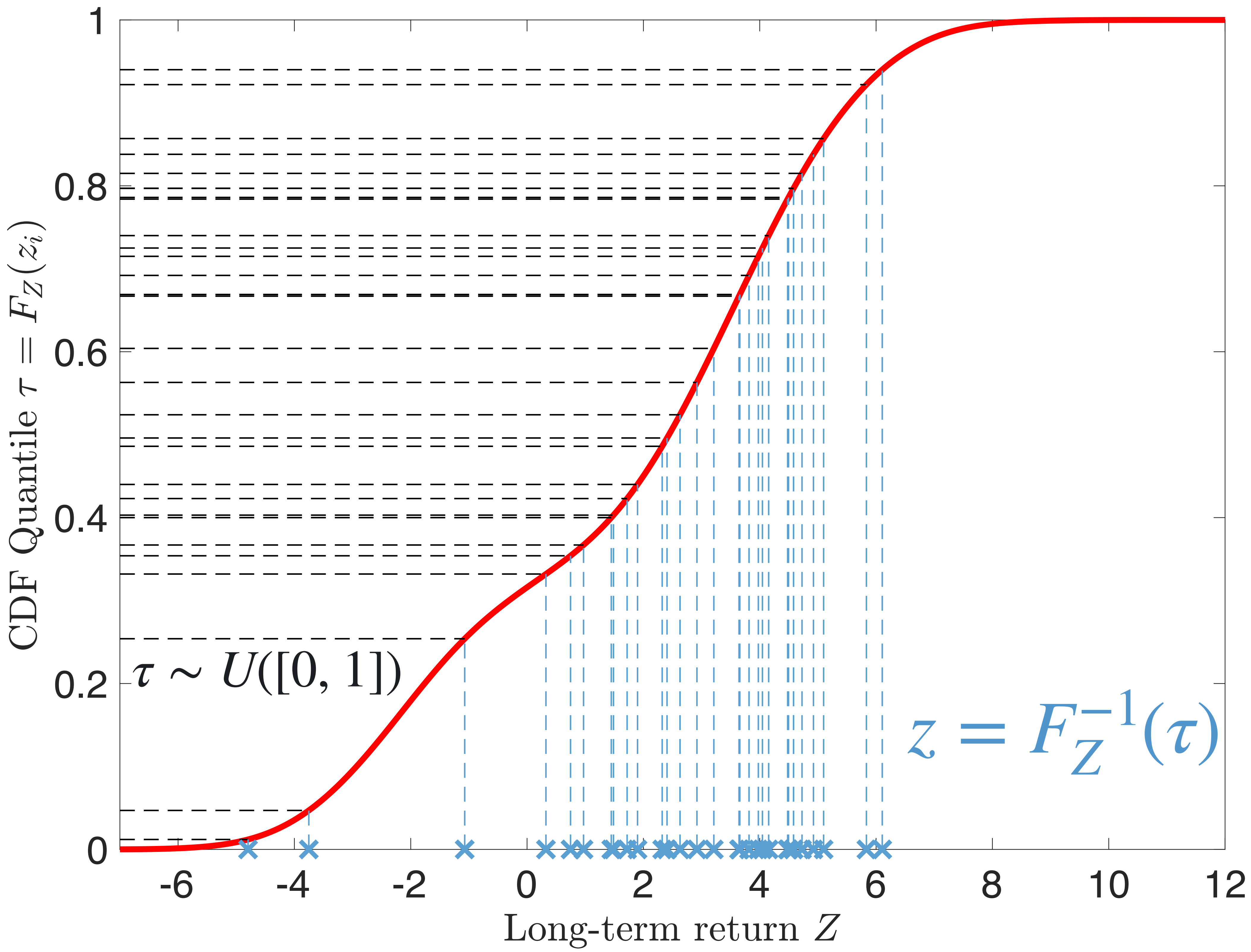


3. Implicit Quantile Inverse CDF



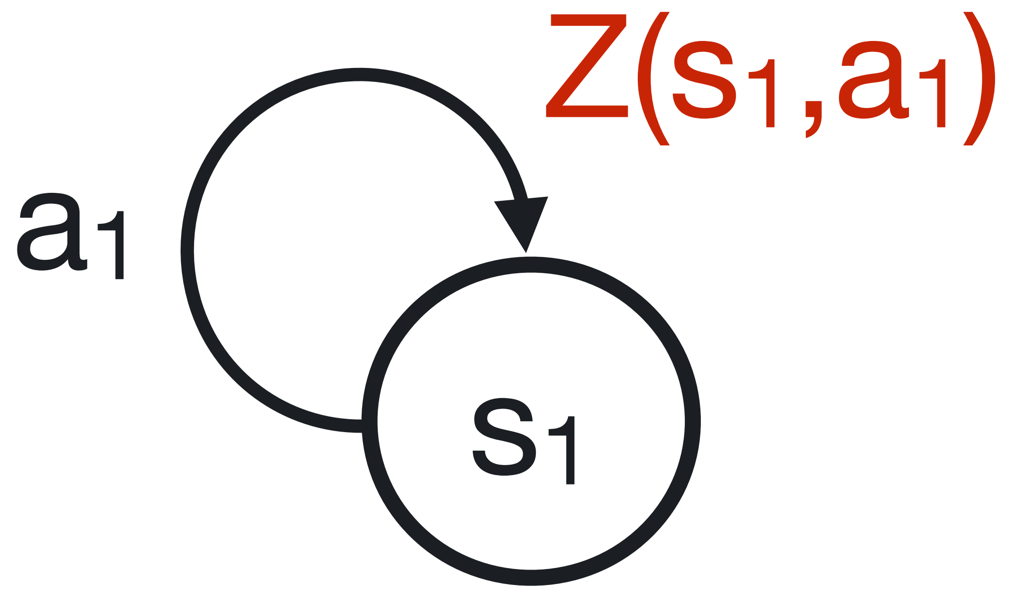
3. Implicit Quantile Inverse CDF

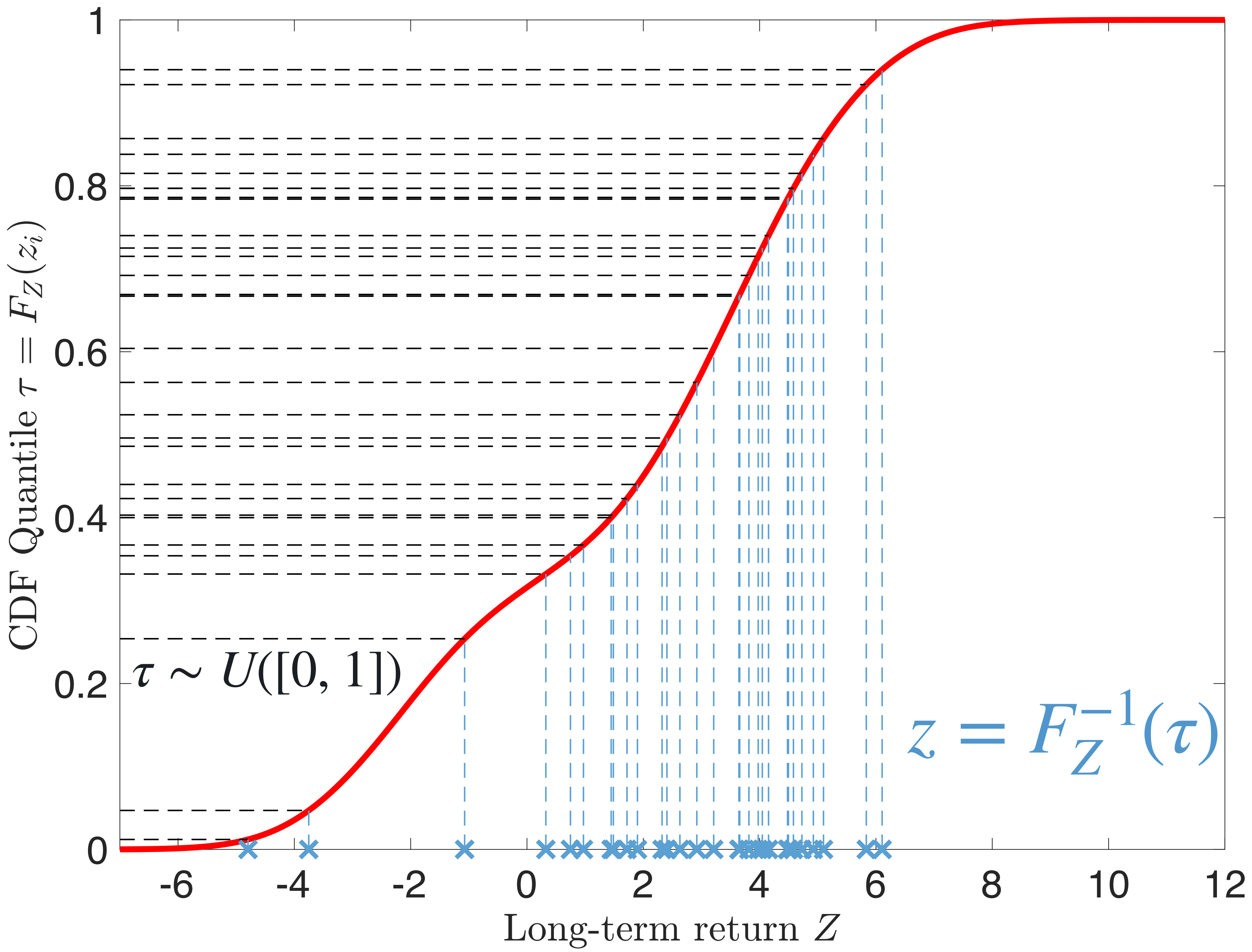




3. Implicit Quantile Inverse CDF

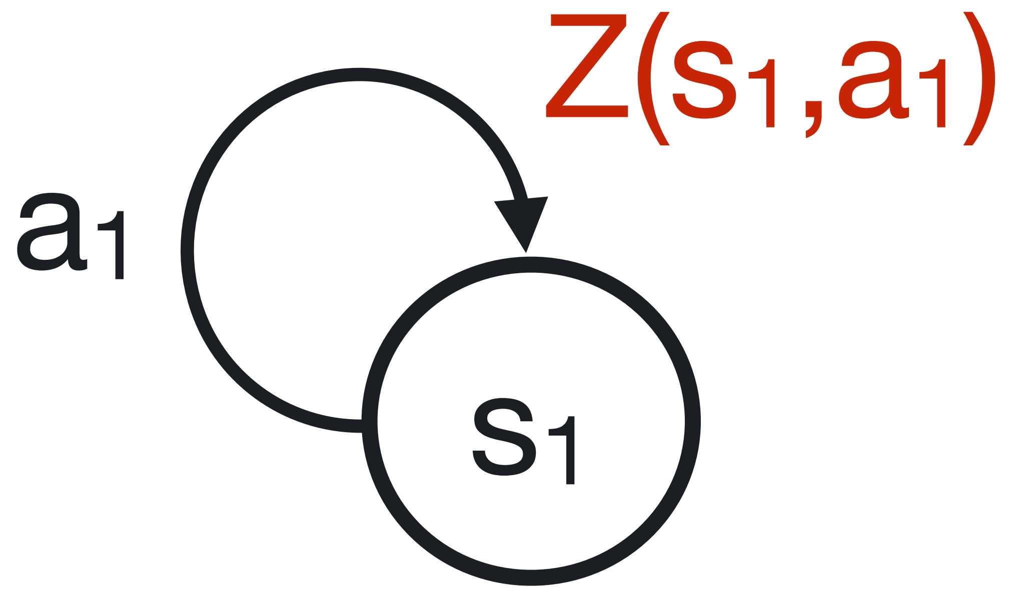
- Quantile sample (input) $\tau \sim U([0, 1])$

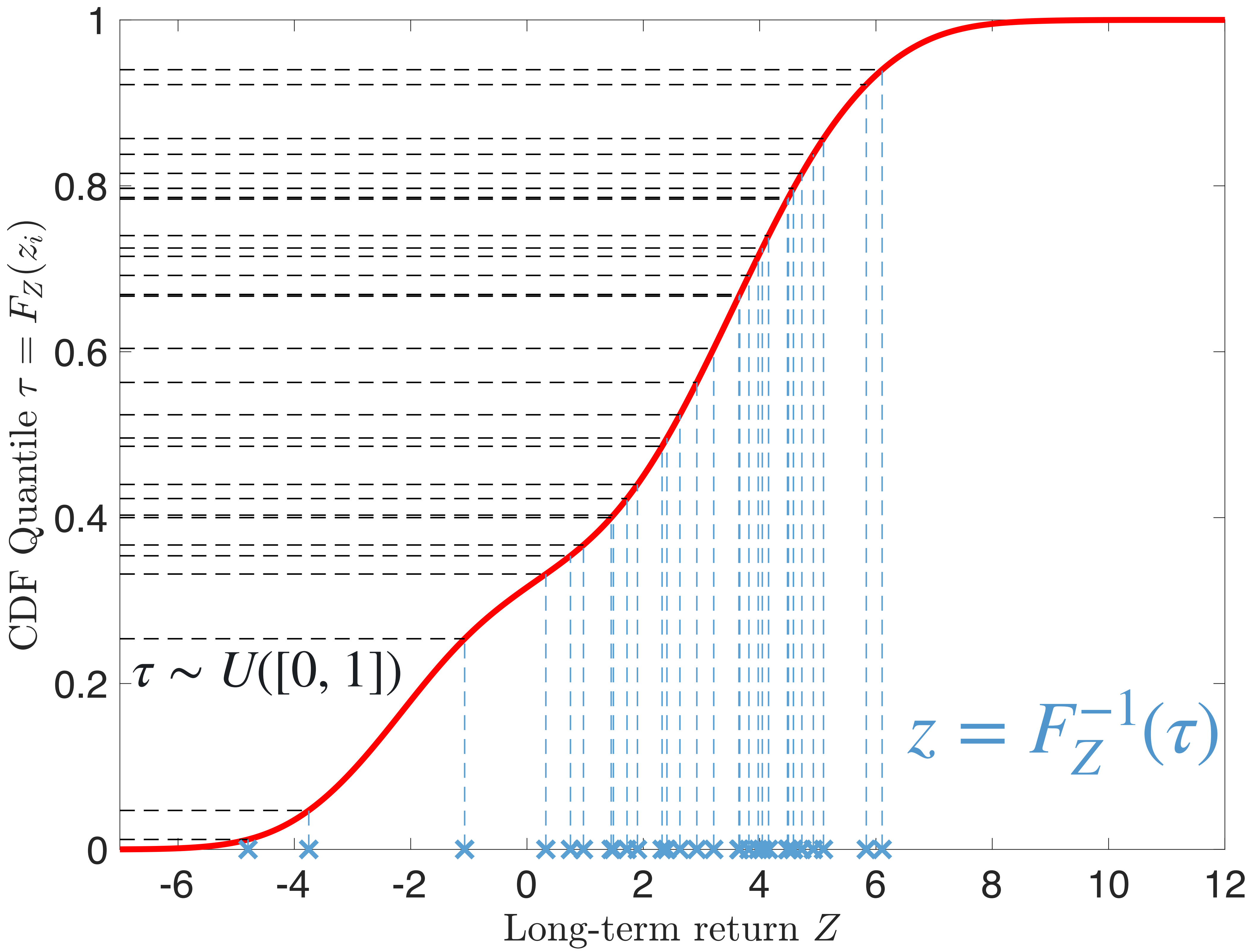




3. Implicit Quantile Inverse CDF

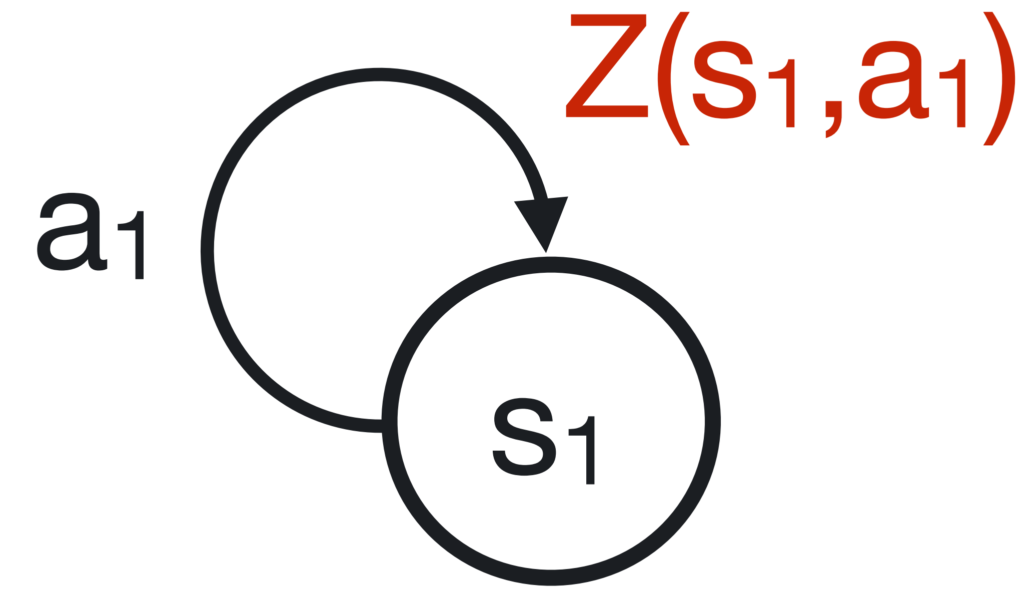
- Quantile sample (input) $\tau \sim U([0, 1])$
- Learn support $z = F_Z^{-1}(\tau)$

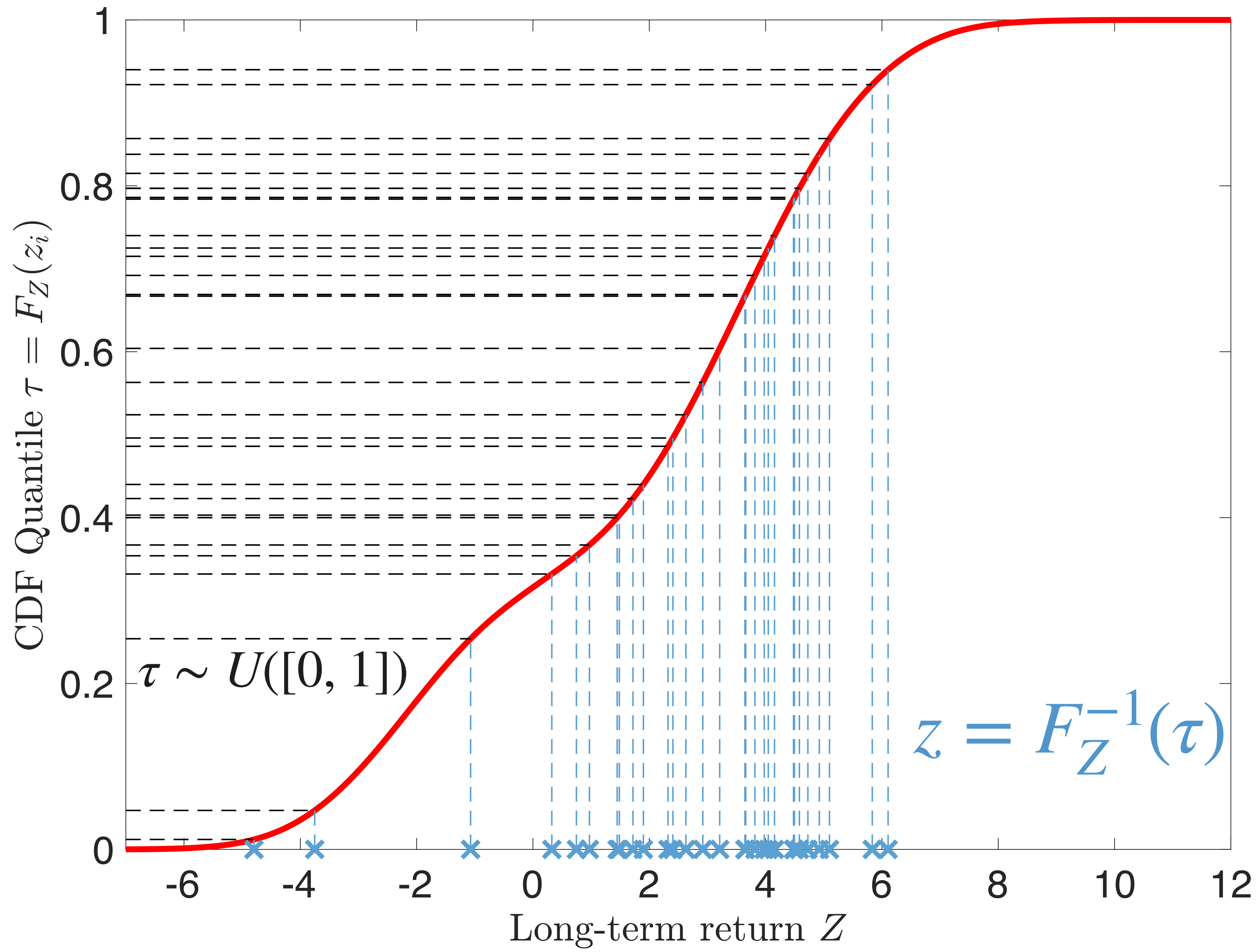


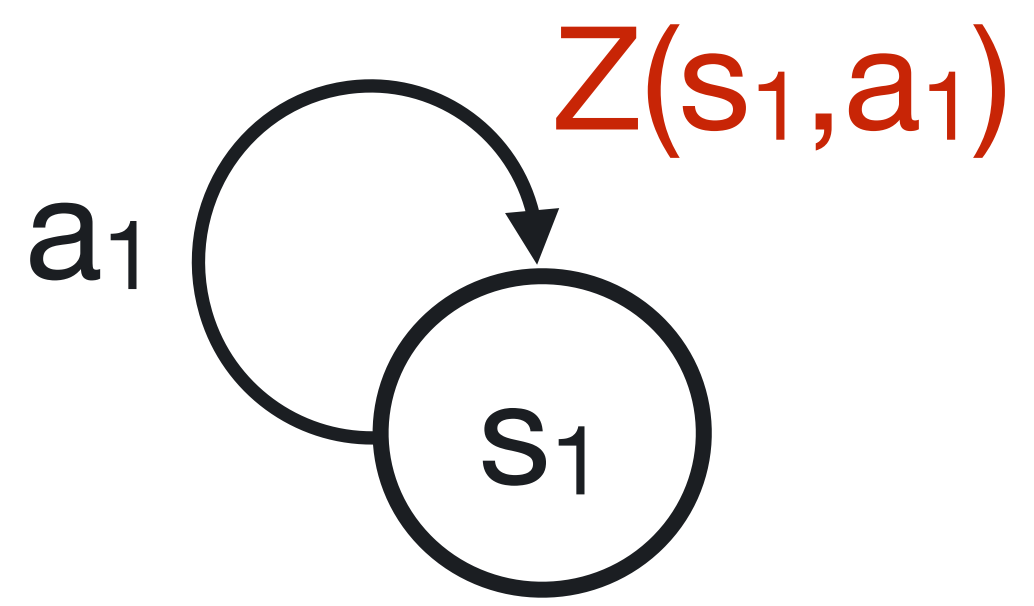


3. Implicit Quantile Inverse CDF

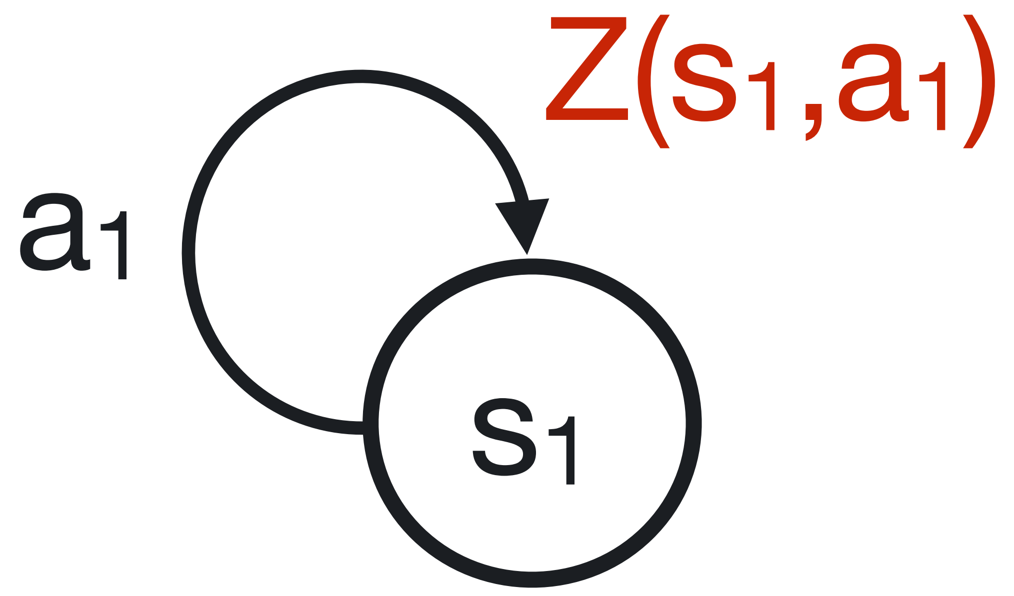
- Quantile sample (input) $\tau \sim U([0,1])$
- Learn support $z = F_Z^{-1}(\tau)$
- Training loss





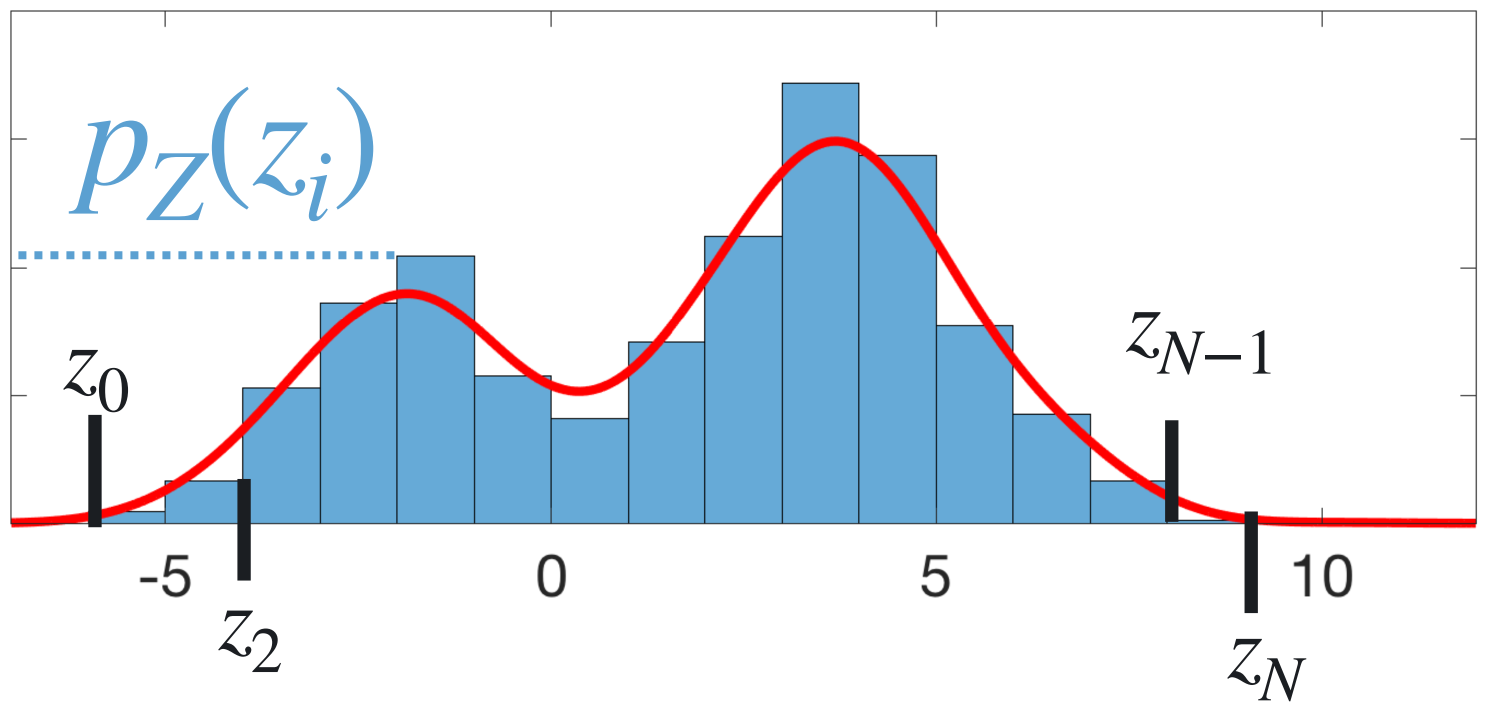


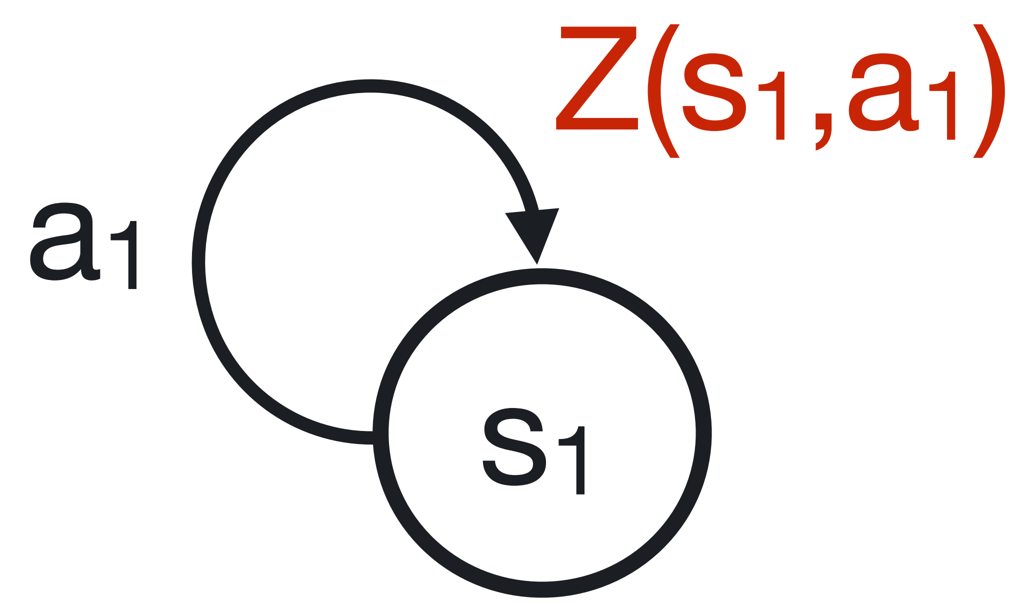
Z(s1,a1)
a1
s1



1. Categorical PDF

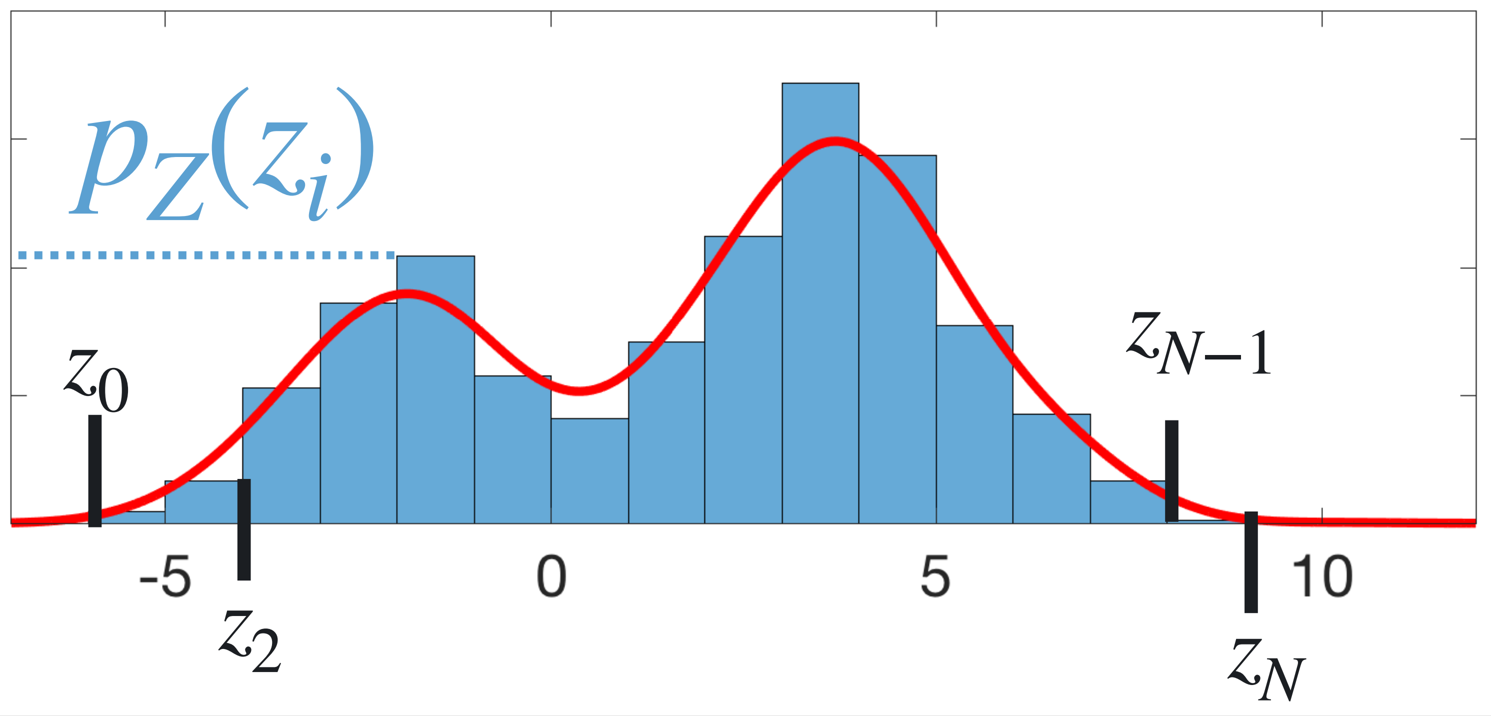
- Fixed support bins $z_0, \ldots, z_N$
- Learn probabilities $p_Z(z_i)$





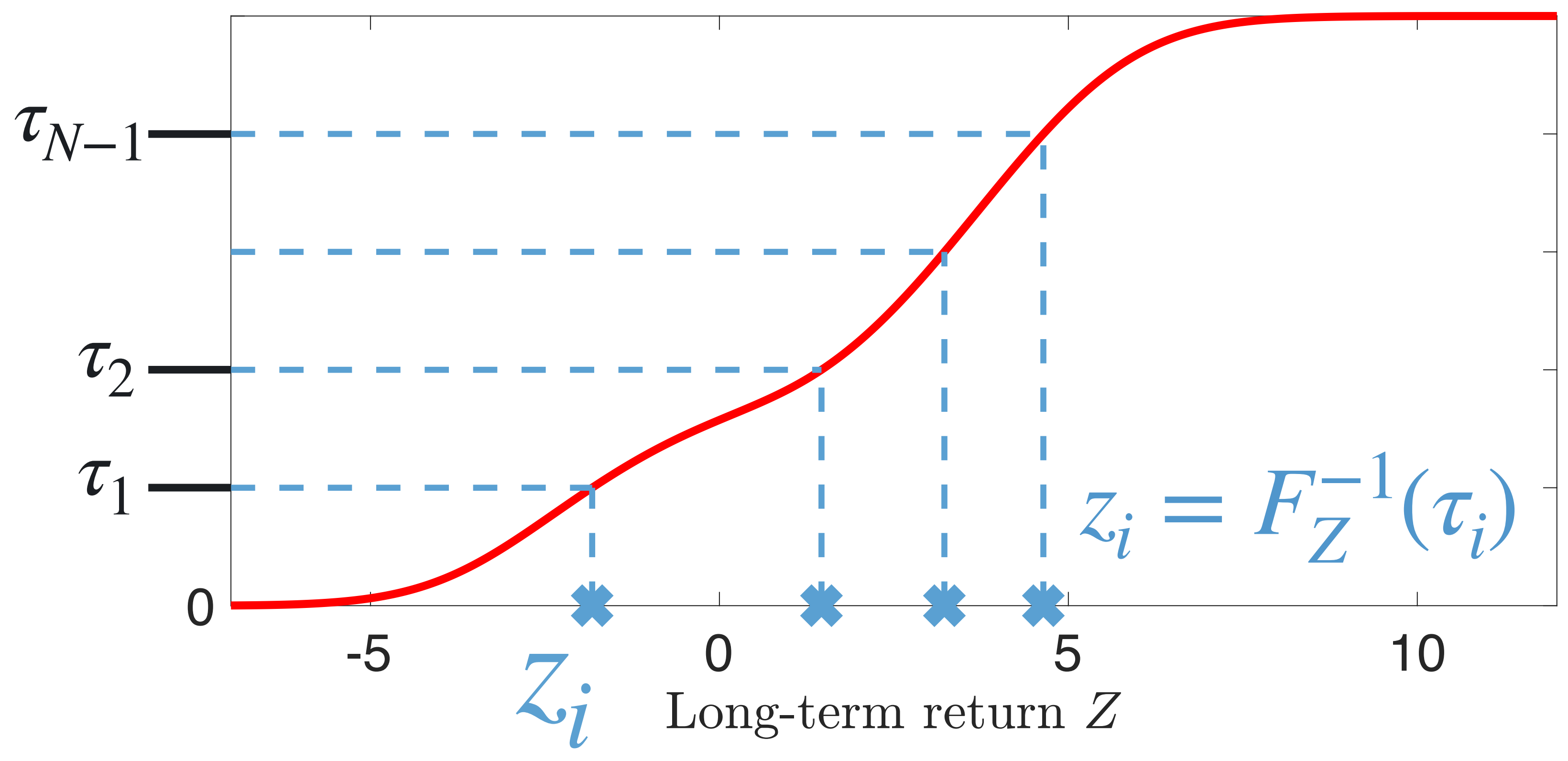
1. Categorical PDF

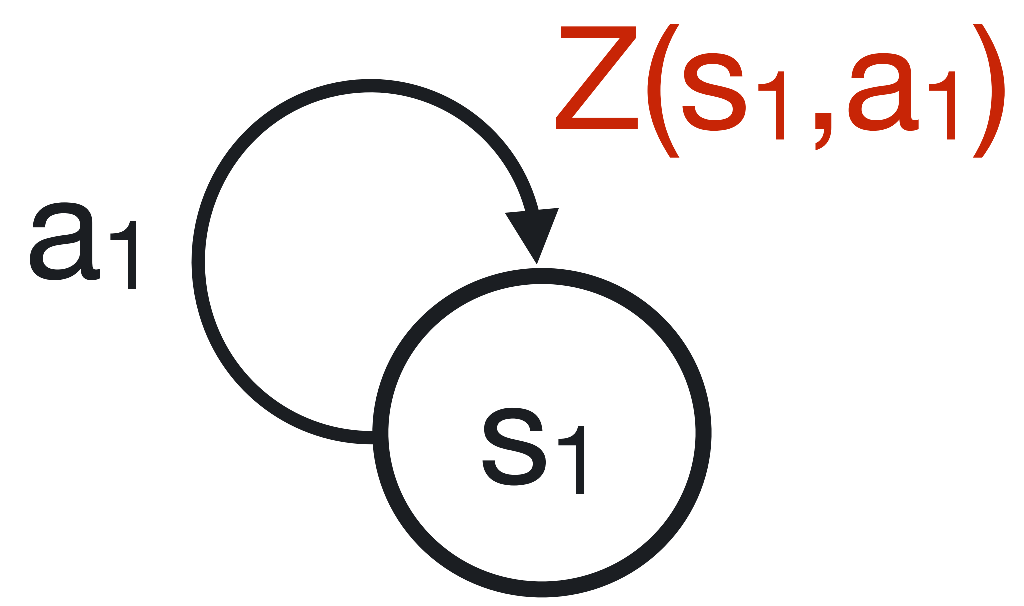
- Fixed support bins $z_0, \ldots, z_N$
- Learn probabilities $p_Z(z_i)$



2. Quantile Inverse CDF

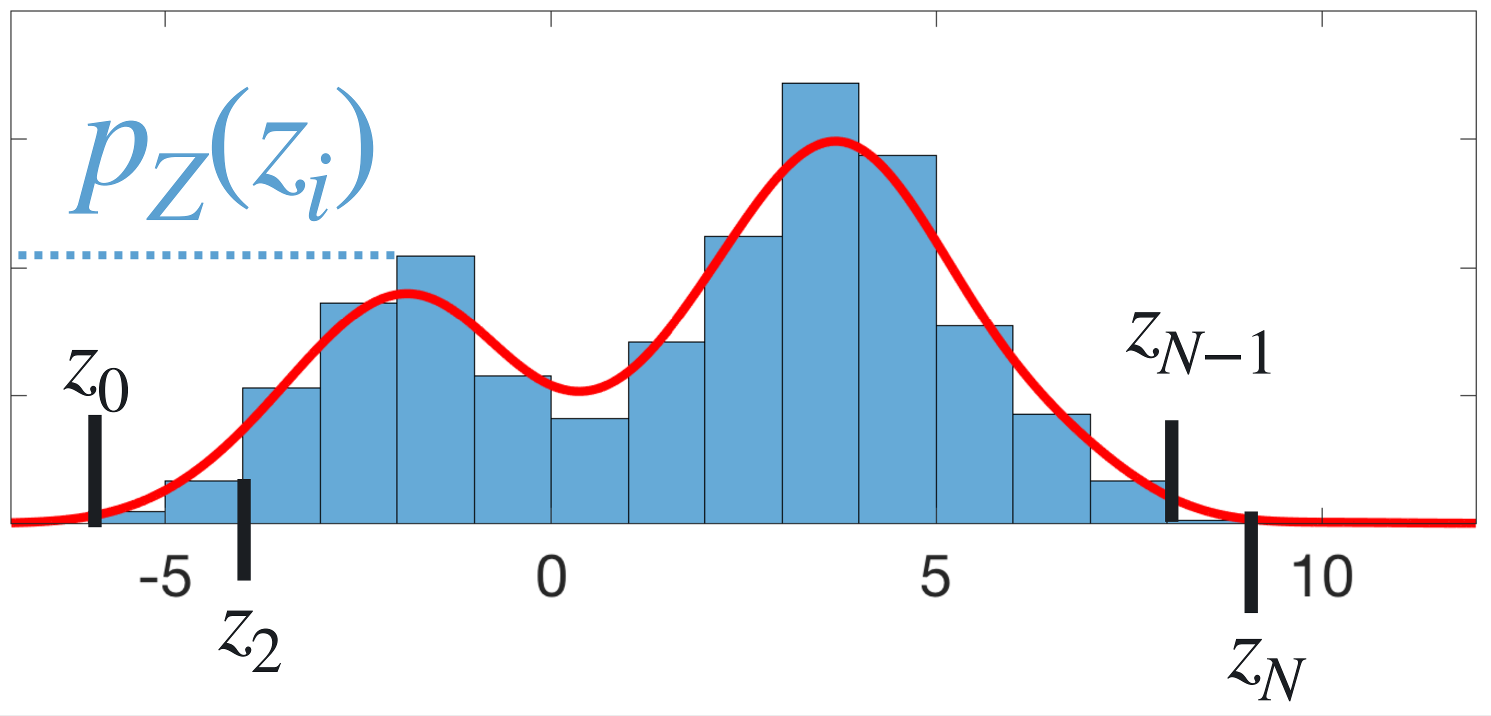
- Fixed quantile bins $\tau_0, \ldots, \tau_N$
- Learn support values $z_i = F_Z^{-1}(\tau_i)$





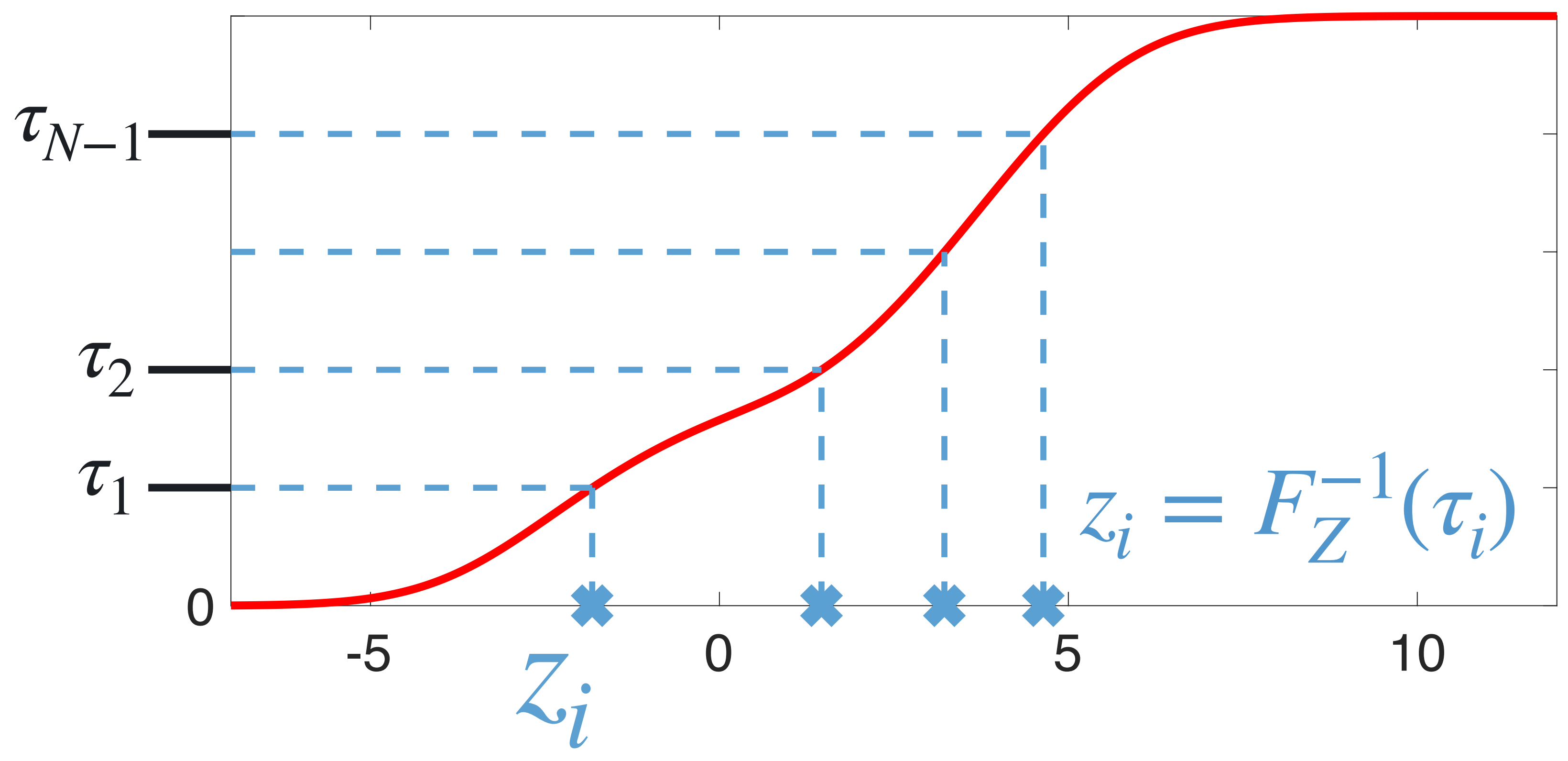
1. Categorical PDF

- Fixed support bins $z_0, \ldots, z_N$
- Learn probabilities $p_Z(z_i)$



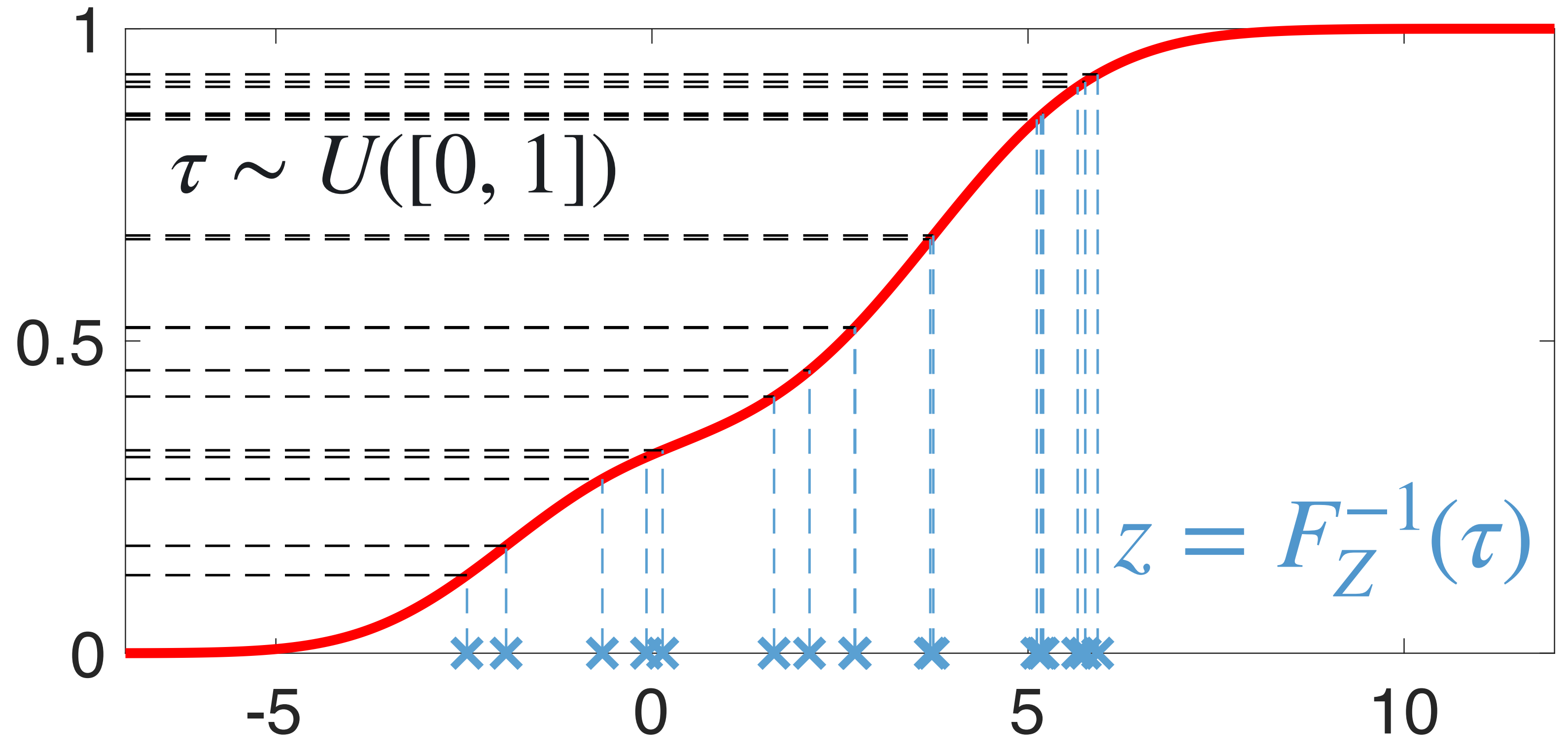
2. Quantile Inverse CDF

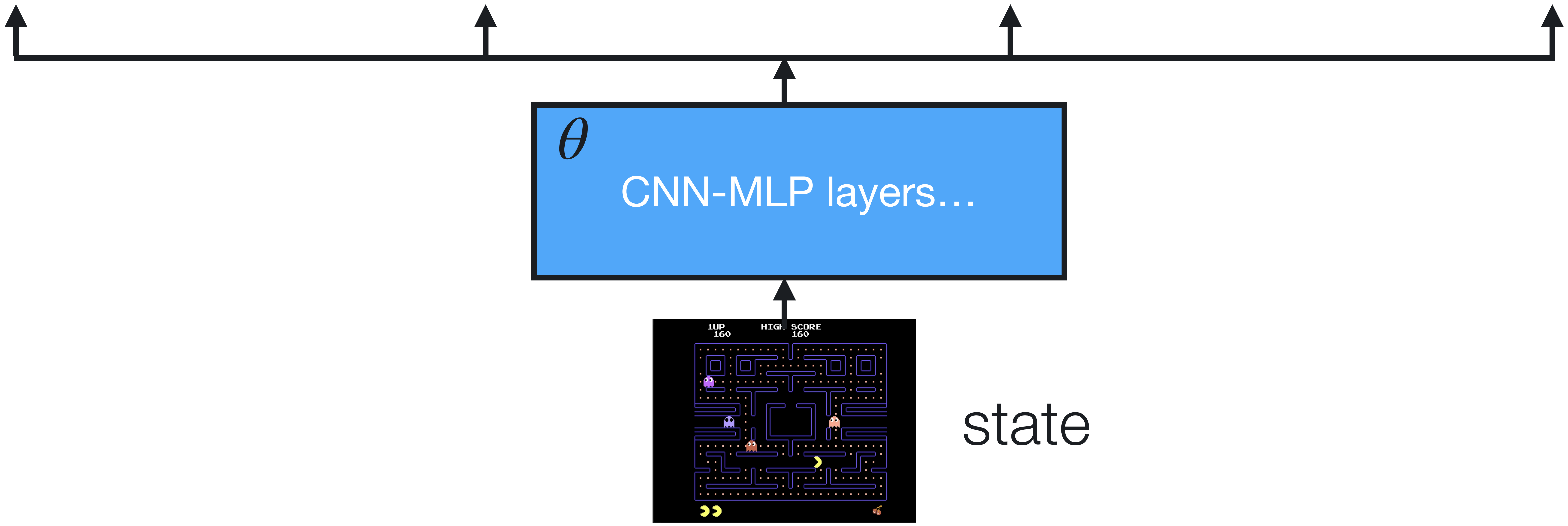
- Fixed quantile bins $\tau_0, \ldots, \tau_N$
- Learn support values $z_i = F_Z^{-1}(\tau_i)$



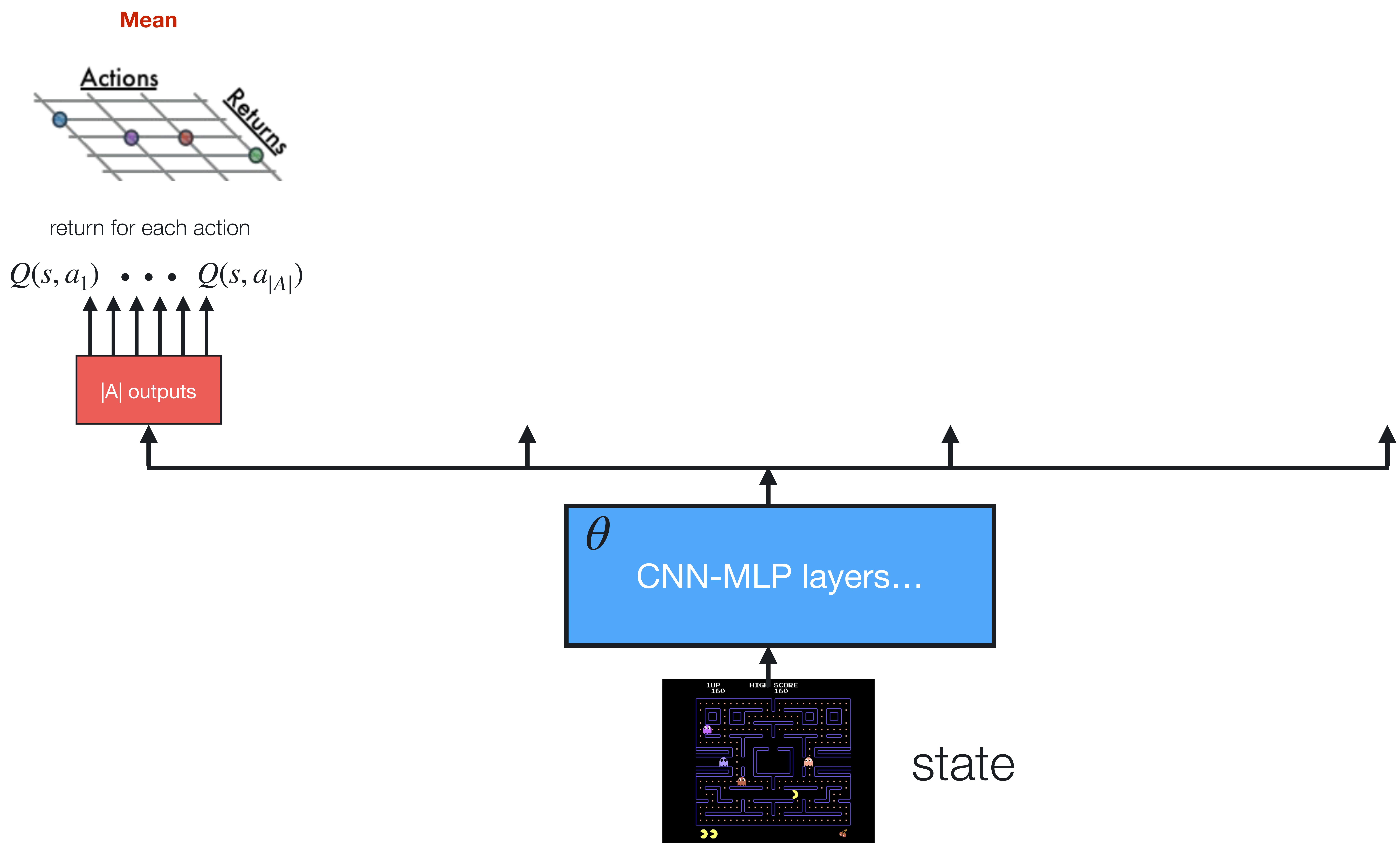
3. Implicit Quantile Inverse CDF

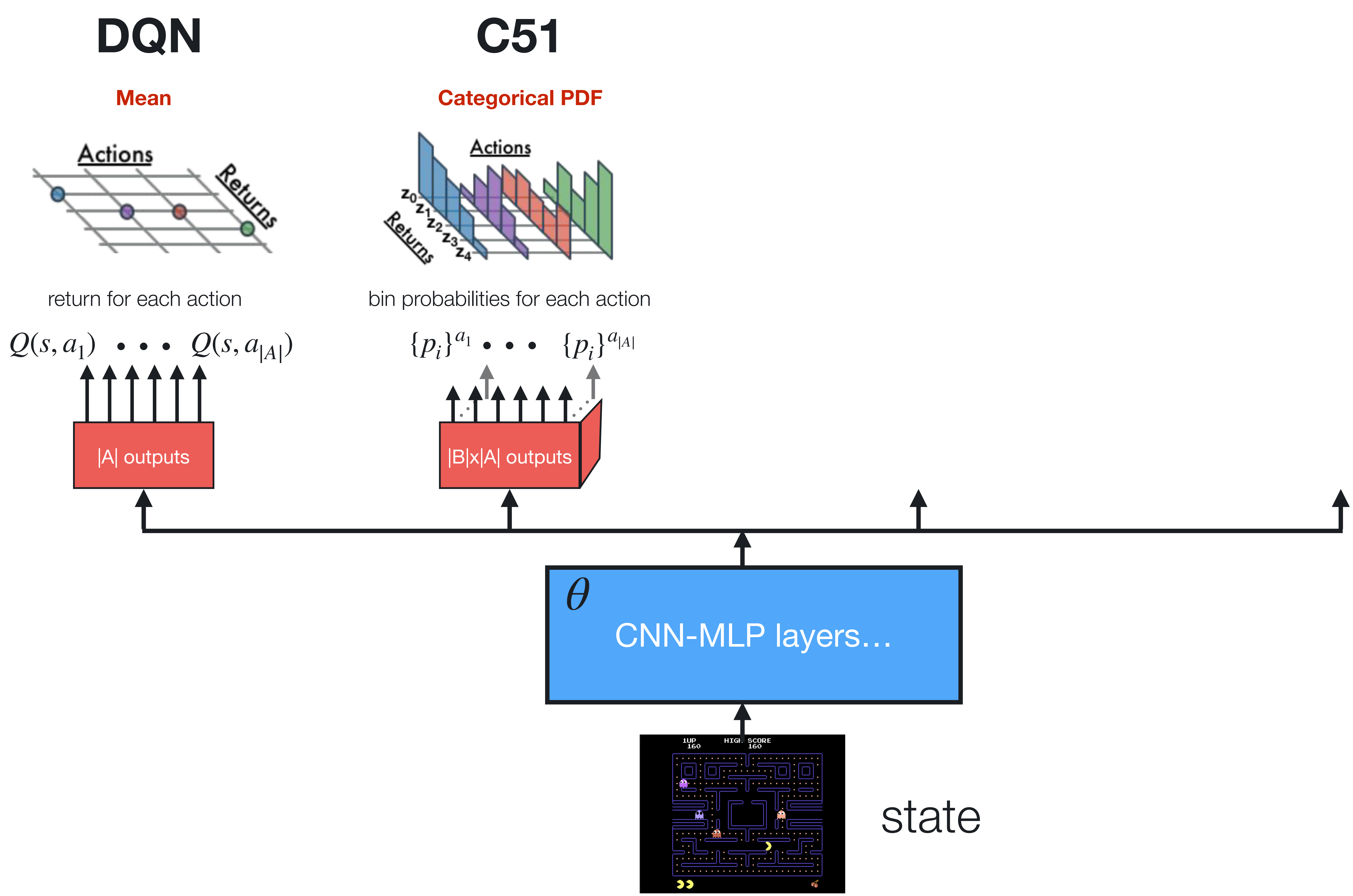
- Quantile sampled from uniform $\tau \sim U([0, 1])$
- Learn support values $z = F_Z^{-1}(\tau)$

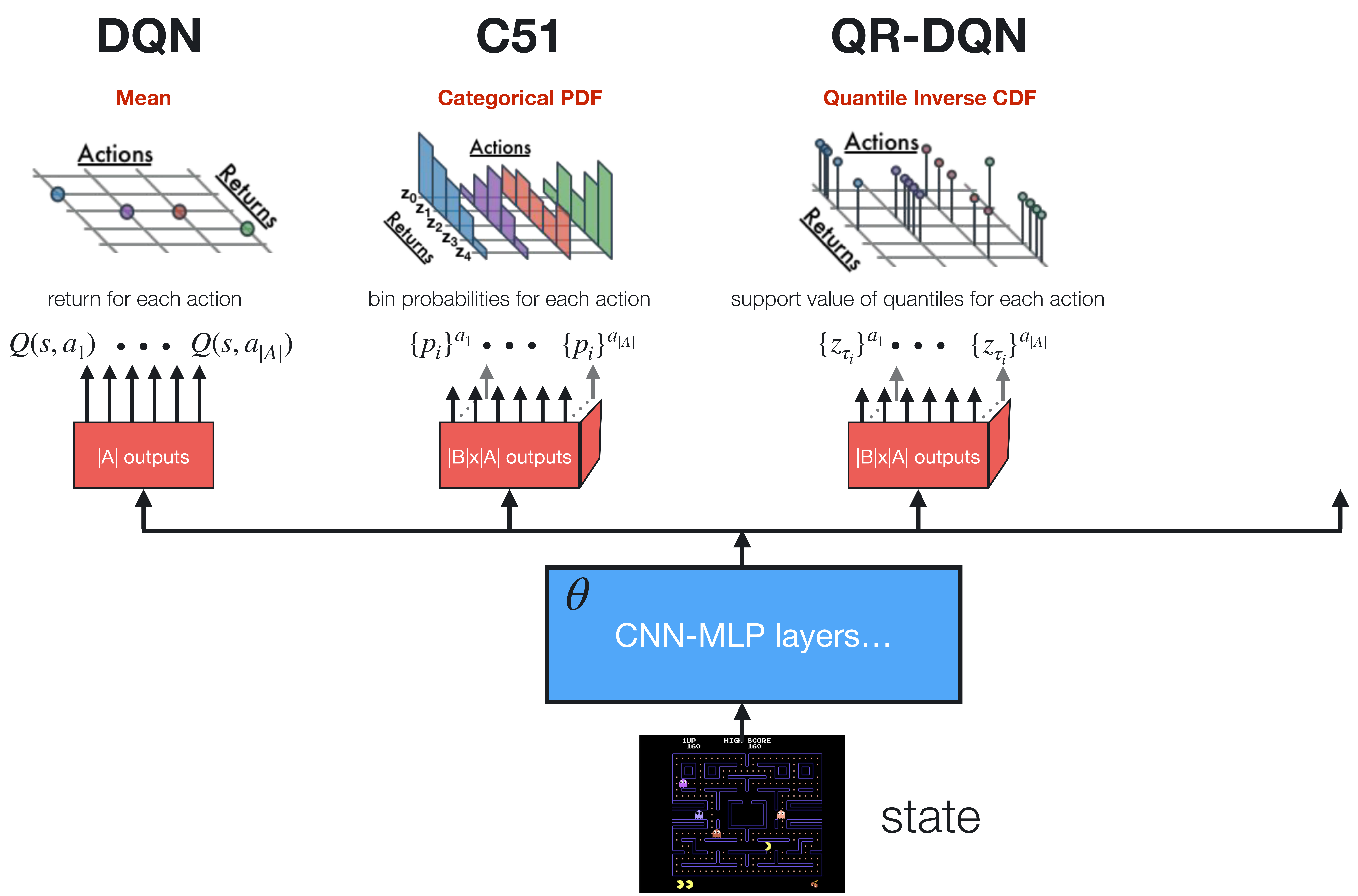


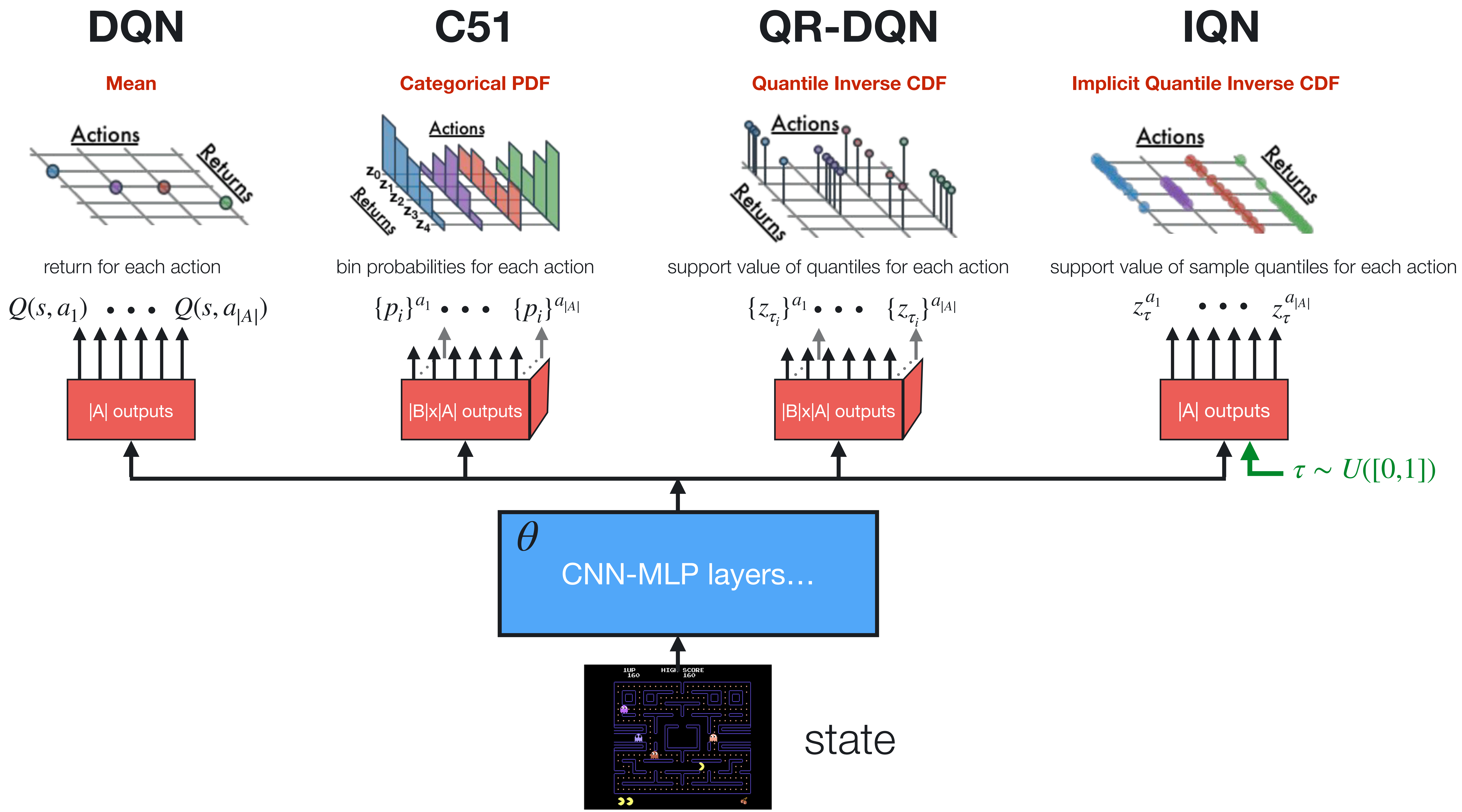

θ
CNN-MLP layers…
state

DQN


Mean
Actions
Returns
return for each action
$Q(s, a_1)$ • • • $Q(s, a_{|A|})$
|A| outputs
θ
CNN-MLP layers…
state

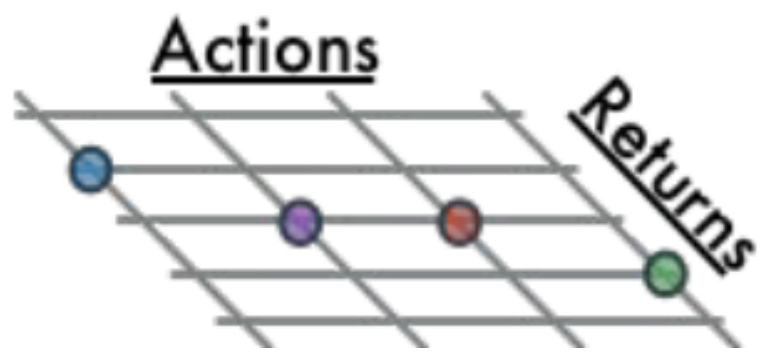

DQN
C51
Mean
Categorical PDF
Actions
Returns
Actions
Returns
return for each action
bin probabilities for each action
$Q(s, a_1) \cdots Q(s, a_{|A|})$
$\{p_i\}^{a_1} \cdots \{p_i\}^{a_{|A|}}$
|A| outputs
|B|x|A| outputs
θ
CNN-MLP layers…
state


DQN
Mean
Actions
Returns
return for each action
$Q(s, a_1) \bullet \bullet \bullet Q(s, a_{|A|})$
|A| outputs
C51
Categorical PDF
Actions
z_0 z_1 z_2 z_3 z_4
Returns
bin probabilities for each action
$\{p_i\}^{a_1} \bullet \bullet \bullet \{p_i\}^{a_{|A|}}$
|B|x|A| outputs
QR-DQN
Quantile Inverse CDF
Actions
Returns
support value of quantiles for each action
$\{z_{\tau_i}\}^{a_1} \bullet \bullet \bullet \{z_{\tau_i}\}^{a_{|A|}}$
|B|x|A| outputs
θ
CNN-MLP layers…
state

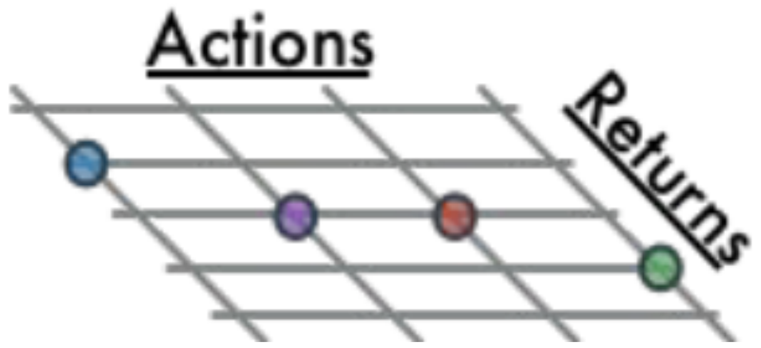

DQN
C51
QR-DQN
IQN
Mean
Categorical PDF
Quantile Inverse CDF
Implicit Quantile Inverse CDF
Actions
Returns
return for each action
bin probabilities for each action
support value of quantiles for each action
support value of sample quantiles for each action
$Q(s, a_1)$ • • • $Q(s, a_{|A|})$
$\{p_i\}^{a_1}$ • • • $\{p_i\}^{a_{|A|}}$
$\{z_{\tau_i}\}^{a_1}$ • • • $\{z_{\tau_i}\}^{a_{|A|}}$
$z_\tau^{a_1}$ • • • $z_\tau^{a_{|A|}}$
|A| outputs
|B|x|A| outputs
|B|x|A| outputs
|A| outputs
$\tau \sim U([0,1])$
θ
CNN-MLP layers…
state

How do we train these networks ?

DQN

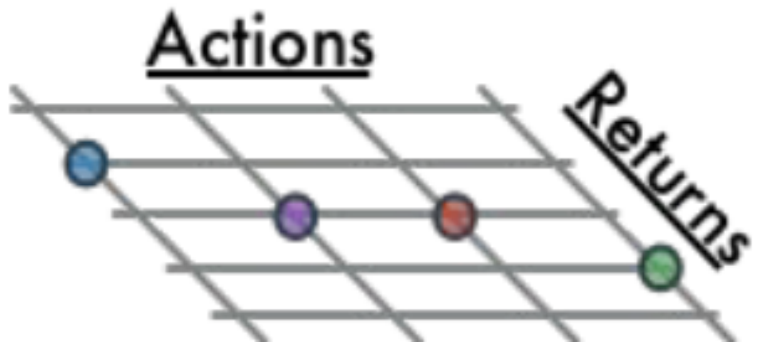


DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

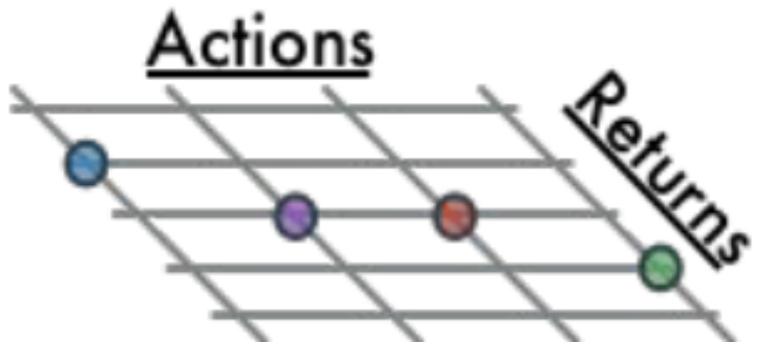
DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\operatorname{argmax}} \; Q^{\theta}(s_{t+1}, a)$$

DQN

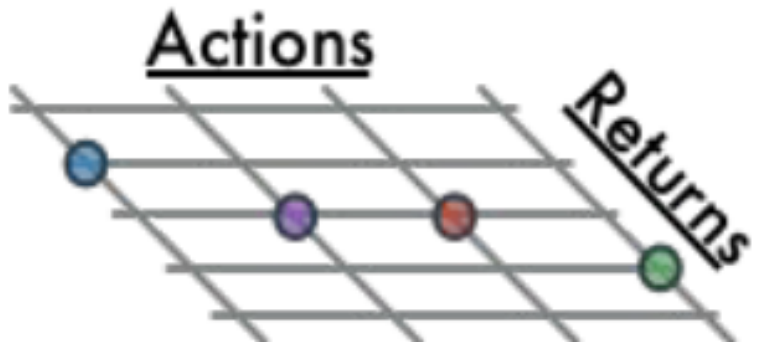


$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\operatorname{argmax}}\ Q^{\theta}(s_{t+1}, a)$$

$$q' = Q^{\theta}(s_{t+1}, a^\star)$$

DQN



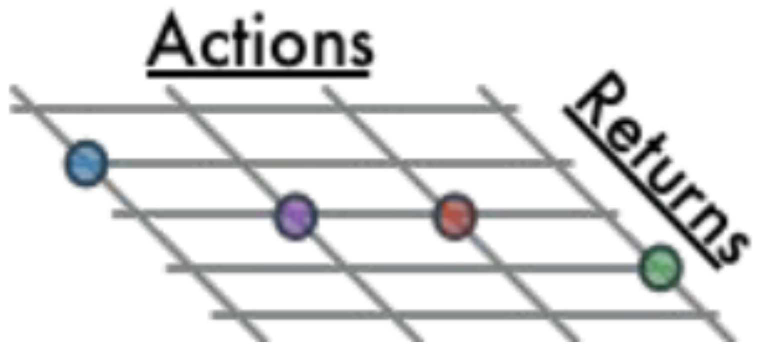
$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}}\ Q^{\theta}(s_{t+1}, a)$$

$$q' = Q^{\theta}(s_{t+1}, a^\star)$$

$$q = Q^{\theta}(s_t, a_t)$$

DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\operatorname{argmax}}\ Q^{\theta}(s_{t+1}, a)$$

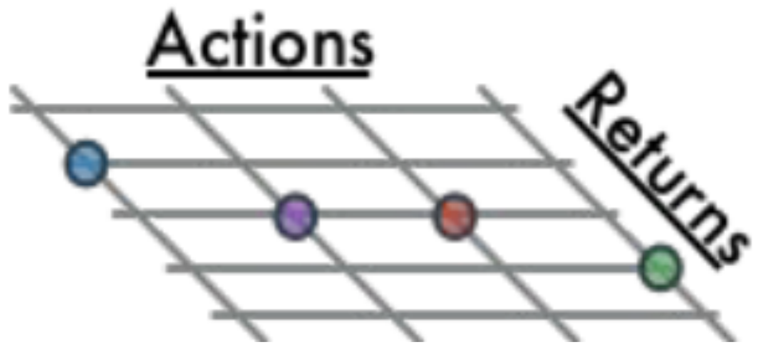
$$q' = Q^{\theta}(s_{t+1}, a^\star)$$

$$q = Q^{\theta}(s_t, a_t)$$

$$\delta_t = r_t + \gamma q' - q$$

temp. diff.

DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\operatorname{argmax}} \; Q^{\theta}(s_{t+1}, a)$$

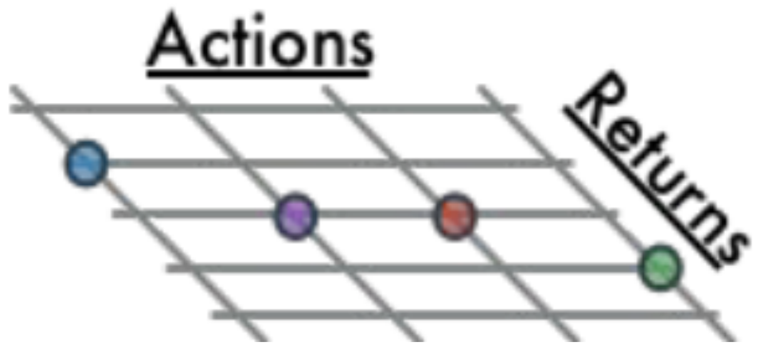
$$q' = Q^{\theta}(s_{t+1}, a^\star)$$

$$q = Q^{\theta}(s_t, a_t)$$

$$\delta_t = r_t + \gamma q' - q$$ **temp. diff.**

$$\mathscr{L}_{DQN} = \delta_t^2$$

DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}} \; Q^{\theta}(s_{t+1}, a)$$

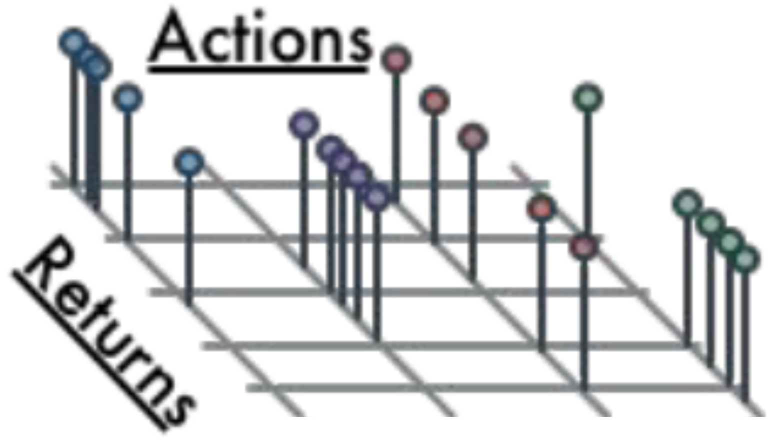
$$q' = Q^{\theta}(s_{t+1}, a^\star)$$

$$q = Q^{\theta}(s_t, a_t)$$

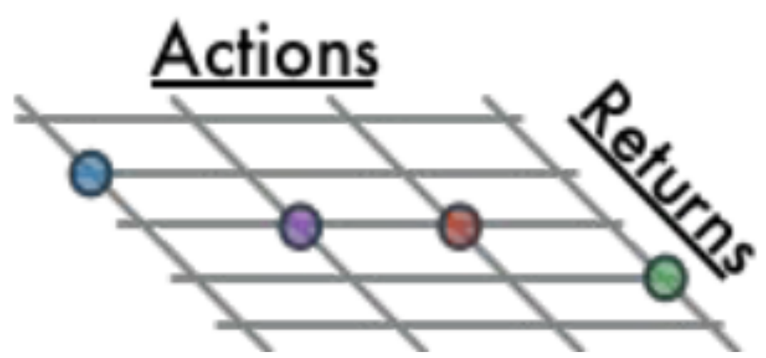
$$\delta_t = r_t + \gamma q' - q$$ **temp. diff.**

$$\mathscr{L}_{DQN} = \delta_t^2$$

QR-DQN



DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\operatorname{argmax}} \; Q^{\theta}(s_{t+1}, a)$$

$$q' = Q^{\theta}(s_{t+1}, a^\star)$$

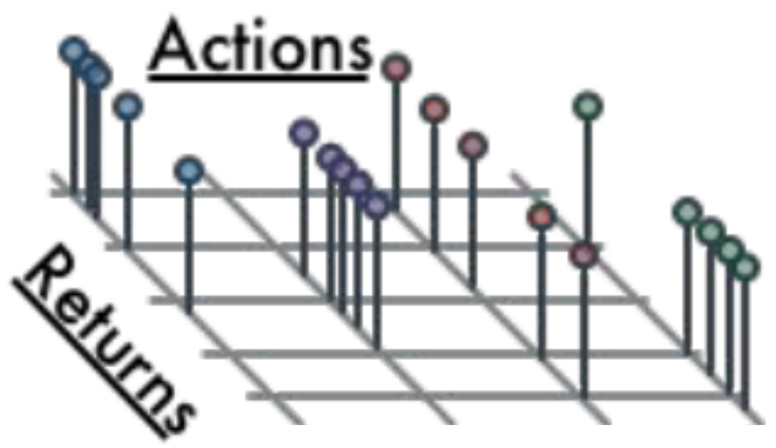
$$q = Q^{\theta}(s_t, a_t)$$

$$\delta_t = r_t + \gamma q' - q$$

temp. diff.

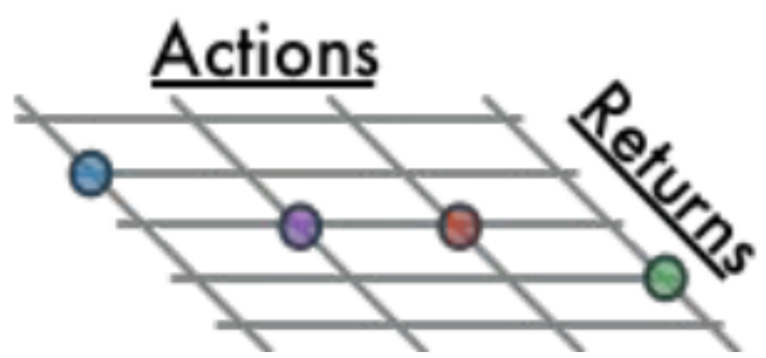
$$\mathcal{L}_{DQN} = \delta_t^2$$

QR-DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}} \; Q^\theta(s_{t+1}, a)$$

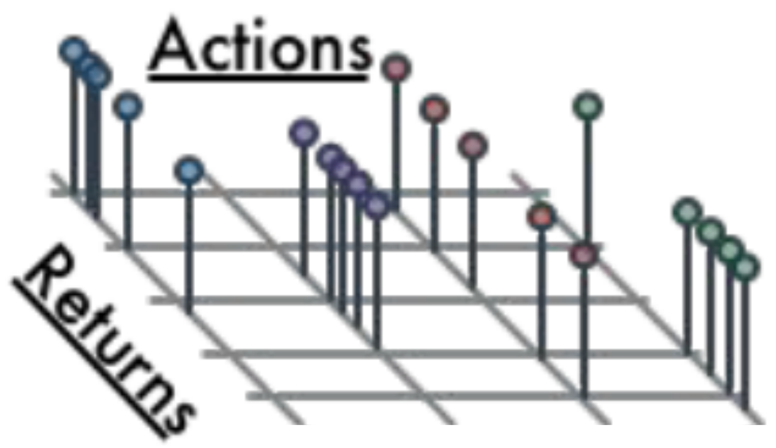
$$q' = Q^\theta(s_{t+1}, a^\star)$$

$$q = Q^\theta(s_t, a_t)$$

$$\delta_t = r_t + \gamma q' - q$$ **temp. diff.**

$$\mathscr{L}_{DQN} = \delta_t^2$$

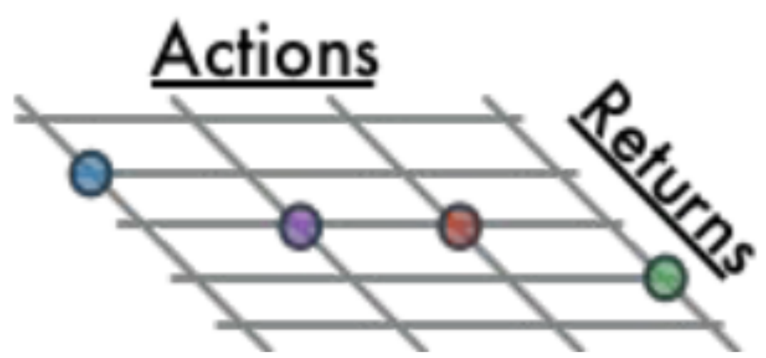
QR-DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}} \; \mathbb{E}\left[Z_\tau^\theta(s_{t+1}, a)\right]$$

DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}} \; Q^{\theta}(s_{t+1}, a)$$

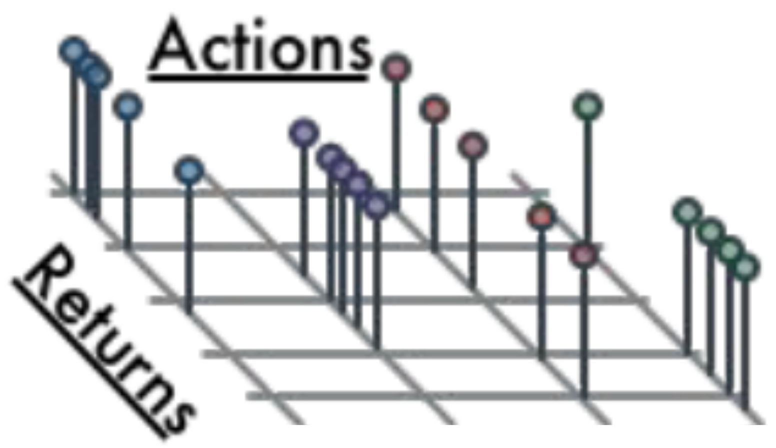
$$q' = Q^{\theta}(s_{t+1}, a^\star)$$

$$q = Q^{\theta}(s_t, a_t)$$

$$\delta_t = r_t + \gamma q' - q$$ **temp. diff.**

$$\mathcal{L}_{DQN} = \delta_t^2$$

QR-DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}} \; \mathbb{E}\left[Z_\tau^{\theta}(s_{t+1}, a)\right]$$

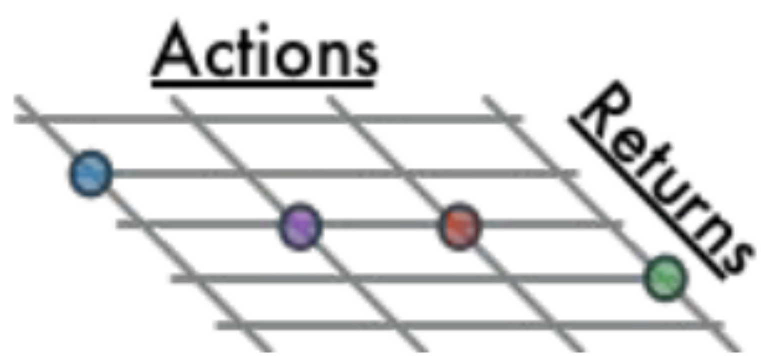
$\forall \tau, \tau'$

$$z' = Z_{\tau'}^{\theta}(s_{t+1}, a^\star)$$

$$z = Z_\tau^{\theta}(s_t, a_t)$$

$$\delta_t^{\tau,\tau'} = r_t + \gamma z' - z$$

DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}} \; Q^{\theta}(s_{t+1}, a)$$

$$q' = Q^{\theta}(s_{t+1}, a^\star)$$

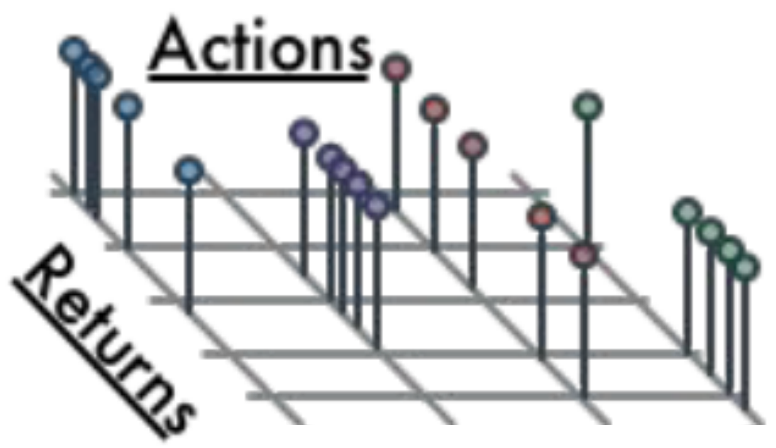
$$q = Q^{\theta}(s_t, a_t)$$

$$\delta_t = r_t + \gamma q' - q$$

temp. diff.

$$\mathscr{L}_{DQN} = \delta_t^2$$

QR-DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}} \; \mathbb{E}\left[Z_\tau^{\theta}(s_{t+1}, a)\right]$$

$\forall \tau, \tau'$

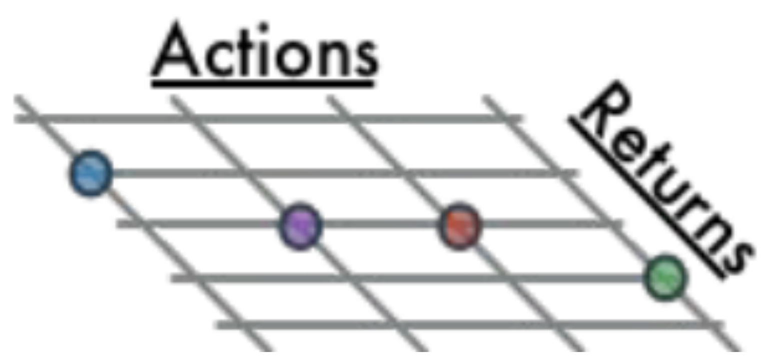
$$z' = Z_{\tau'}^{\theta}(s_{t+1}, a^\star)$$

$$z = Z_\tau^{\theta}(s_t, a_t)$$

$$\delta_t^{\tau,\tau'} = r_t + \gamma z' - z$$

$$\mathscr{L}_{QR-DQN} = ?$$

DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}}\ Q^{\theta}(s_{t+1}, a)$$

$$q' = Q^{\theta}(s_{t+1}, a^\star)$$

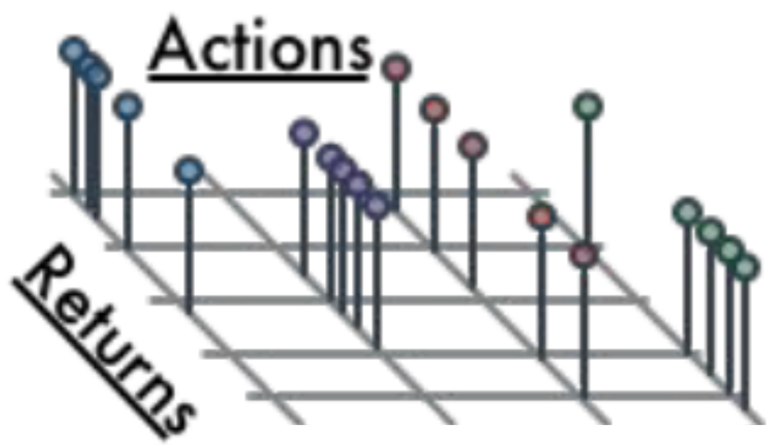
$$q = Q^{\theta}(s_t, a_t)$$

$$\delta_t = r_t + \gamma q' - q$$

temp. diff.

$$\mathscr{L}_{DQN} = \delta_t^2$$

QR-DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}}\ \mathbb{E}\left[Z_\tau^{\theta}(s_{t+1}, a)\right]$$

$\forall \tau, \tau'$

$$z' = Z_{\tau'}^{\theta}(s_{t+1}, a^\star)$$

$$z = Z_\tau^{\theta}(s_t, a_t)$$

$$\delta_t^{\tau,\tau'} = r_t + \gamma z' - z$$

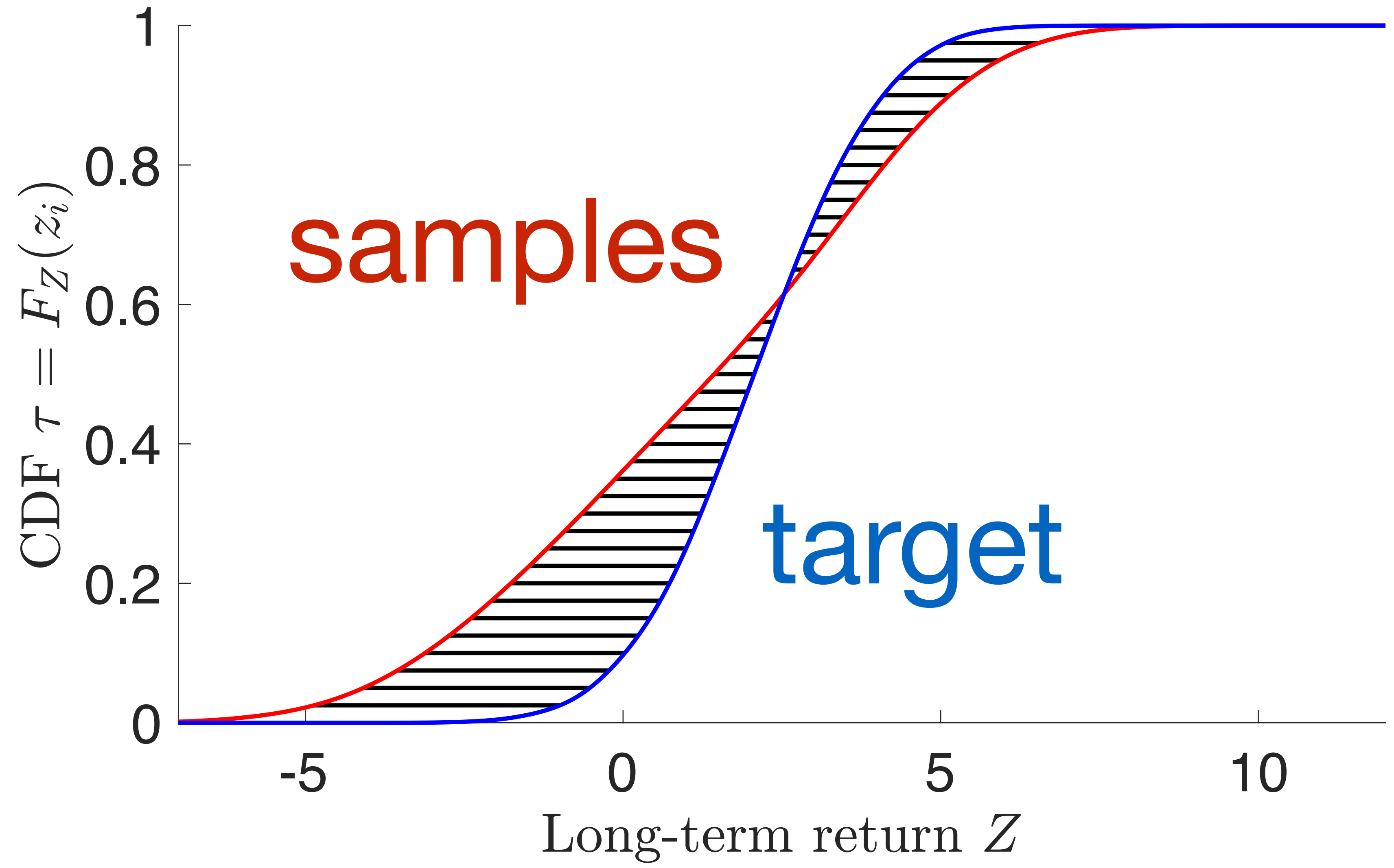
$$\mathscr{L}_{QR-DQN} = ?$$

Projection to Wasserstein metric!

The distributional Bellman Operator is a contraction on the Wasserstein metric

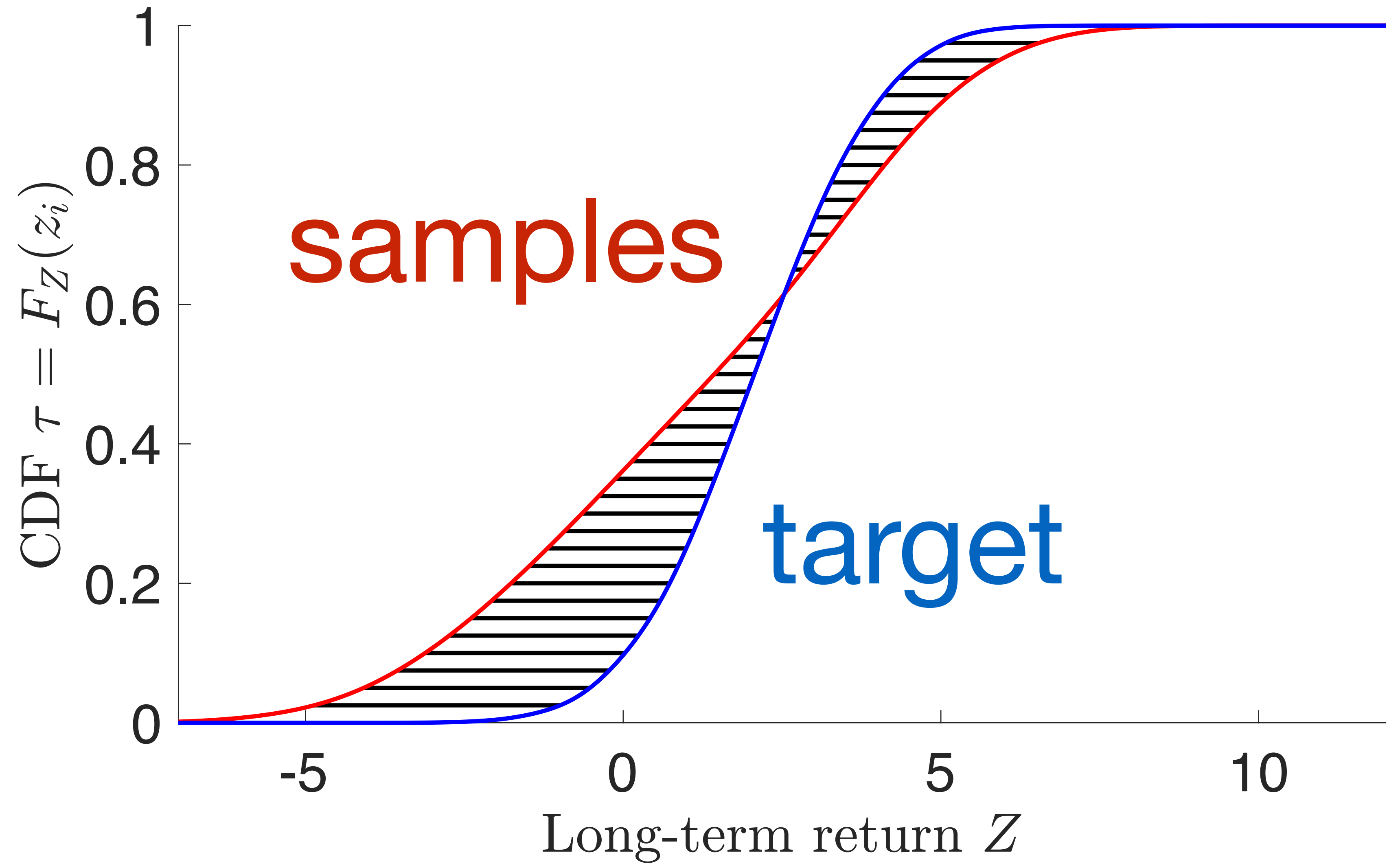
The distributional Bellman Operator is a contraction on the Wasserstein metric

$$w_1(X,Y) = \int_0^1 |F_X^{-1}(\tau) - F_Y^{-1}(\tau)|\, d\tau$$



The distributional Bellman Operator is a contraction on the Wasserstein metric

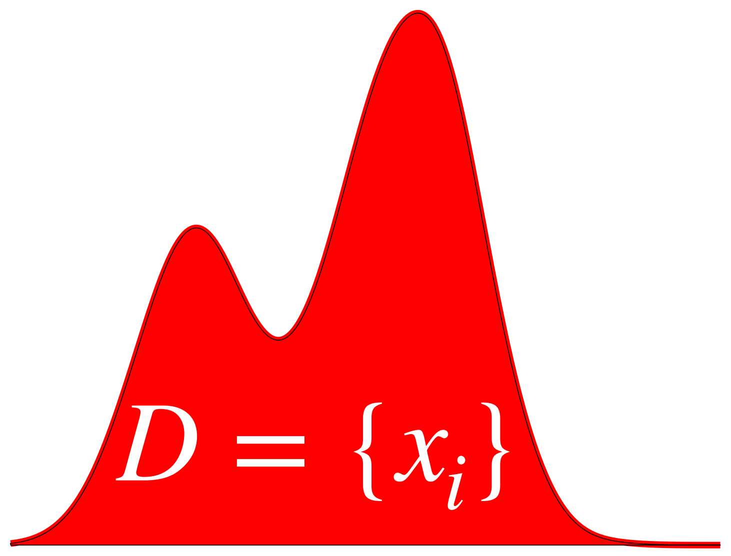
$$w_1(X, Y) = \int_0^1 |F_X^{-1}(\tau) - F_Y^{-1}(\tau)| \, d\tau$$



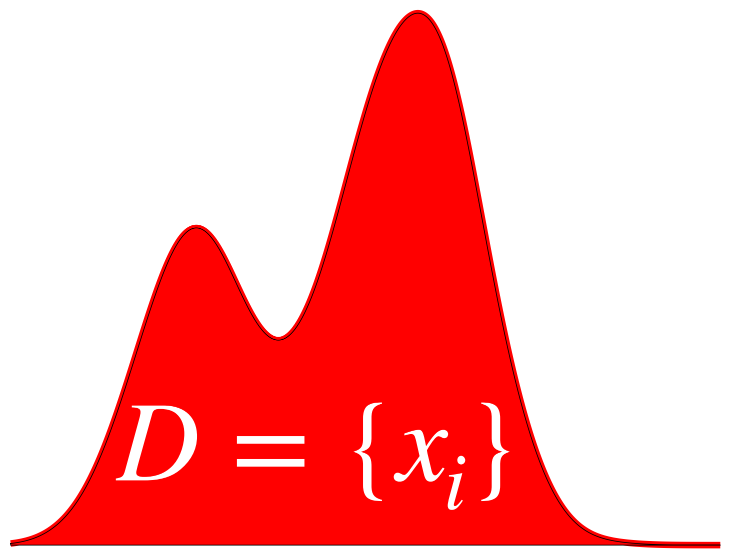
Projection to Wasserstein = **Quantile regression**

(uniform quantile grid)

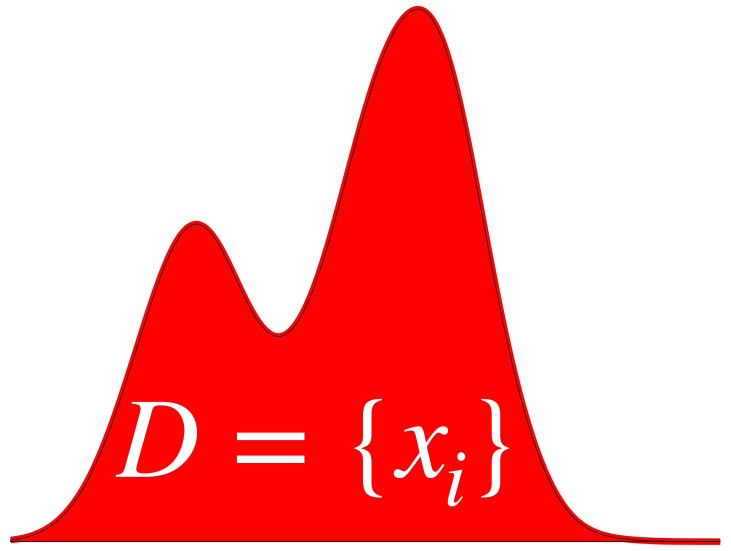
How can we **learn** from data ?

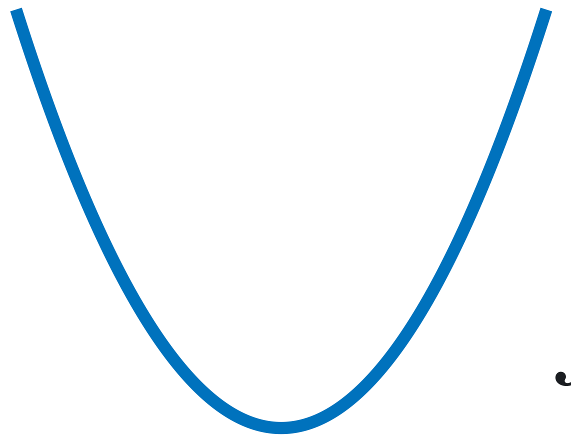


How can we **learn** from data ?

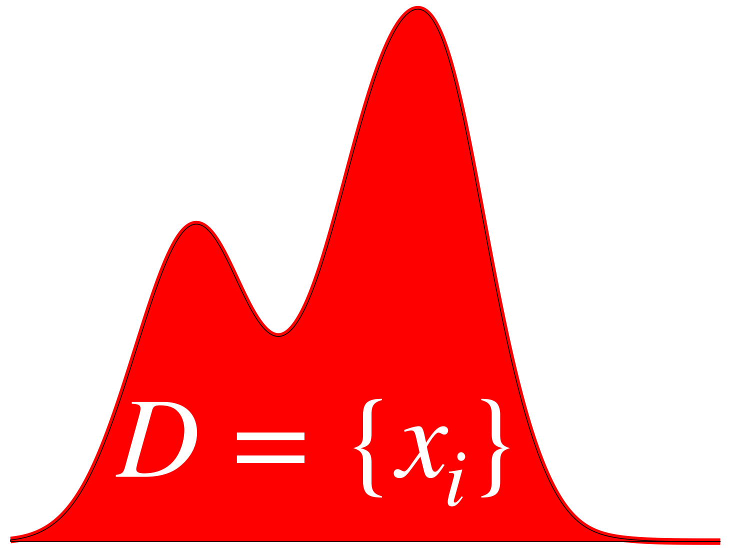


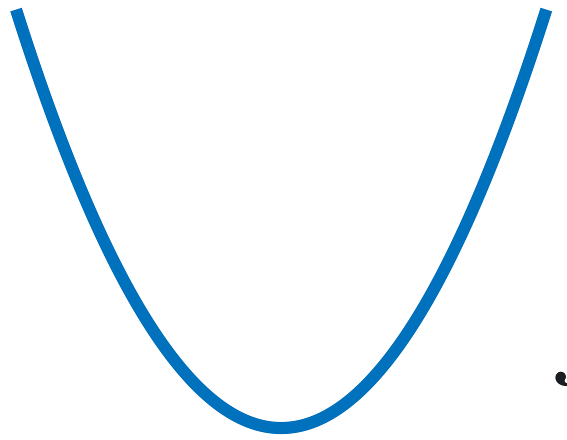
How can we **learn** from data ?



	$f(\delta)$	
$\mathscr{L}_2$ regression	 $f(\delta) = \delta^2$	

How can we **learn** from data ?



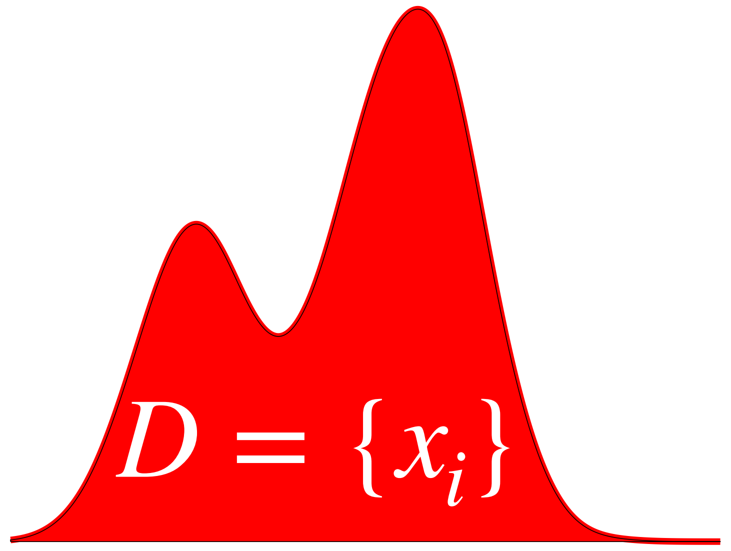
	$f(\delta)$	$x^\star = \underset{x}{\text{argmin}} \sum_i f(x_i - x)$
$\mathcal{L}_2$ regression	 $f(\delta) = \delta^2$	

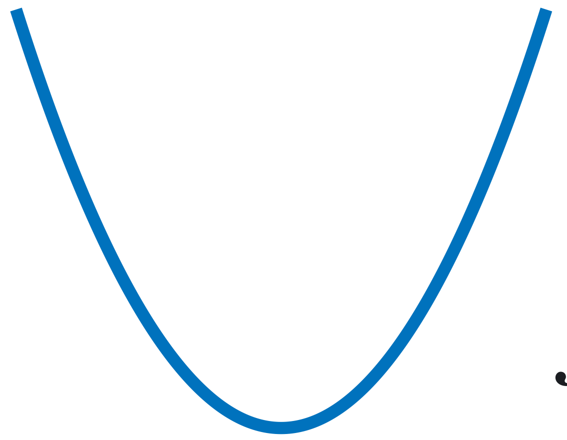
$$\frac{\partial \sum_i f(x_i - x)}{\partial x} = \frac{\partial \sum_i (x_i - x)^2}{\partial x} = \sum_i 2(x_i - x)$$

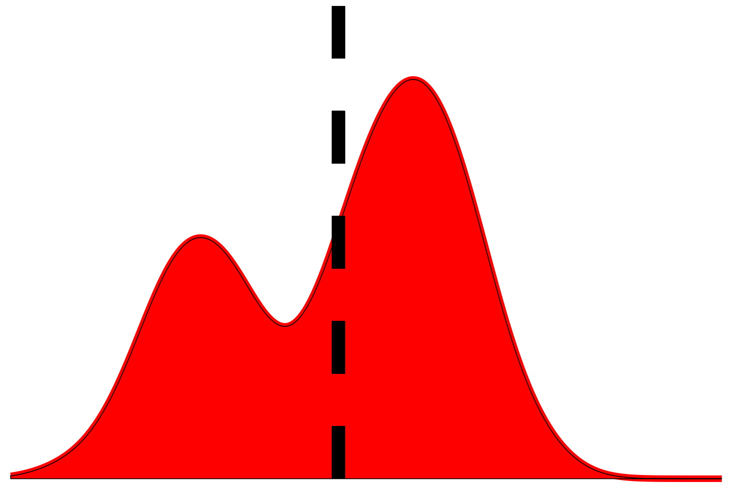
$$\sum_i 2(x_i - x)|_{x^\star} \hat{=} 0 \implies x^\star = \frac{1}{N}\sum_i x_i$$

Mean

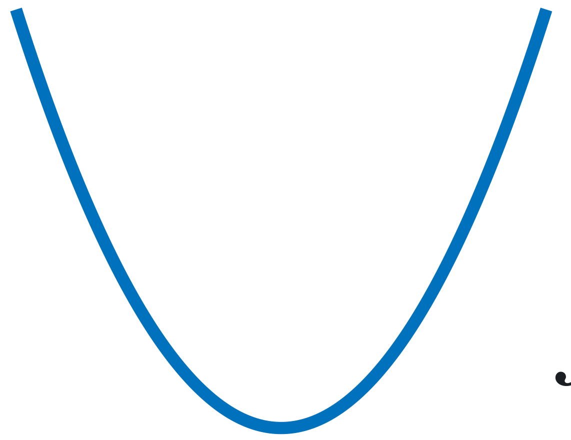
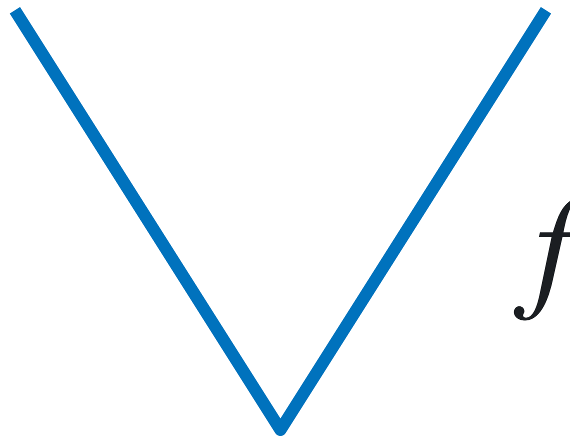
How can we **learn** from data ?

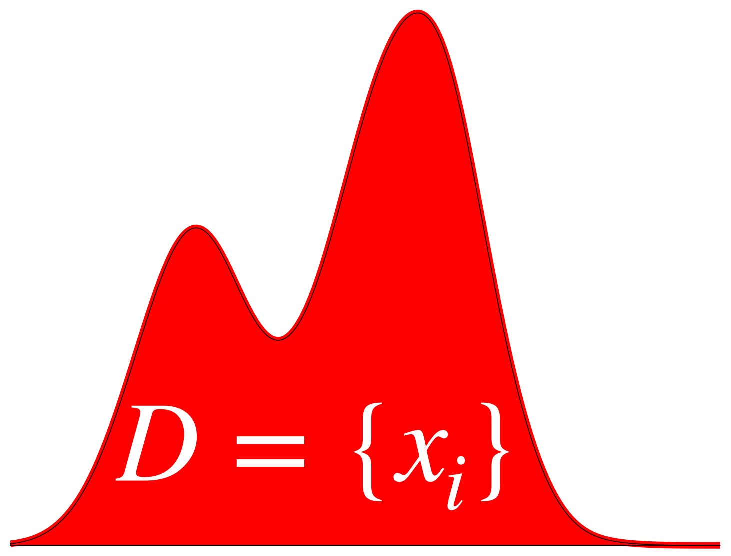


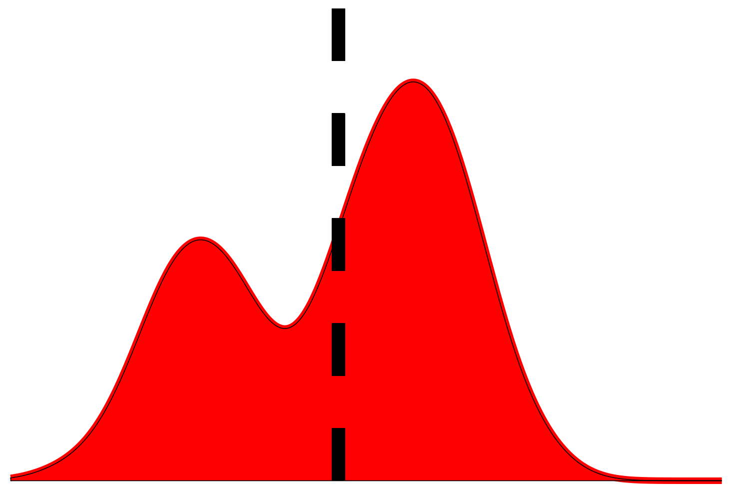
	$f(\delta)$	$x^\star = \underset{x}{\text{argmin}} \sum_i f(x_i - x)$
$\mathscr{L}_2$ regression	 $f(\delta) = \delta^2$	$x^\star = \frac{1}{N}\sum_i x_i$ Mean



How can we **learn** from data ?

	$f(\delta)$	$x^\star = \underset{x}{\operatorname{argmin}} \sum_i f(x_i - x)$
$\mathscr{L}_2$ regression	 $f(\delta) = \delta^2$	$x^\star = \frac{1}{N} \sum_i x_i$ Mean
$\mathscr{L}_1$ regression	 $f(\delta) = \lvert\delta\rvert$	





$$\frac{\partial \sum_i f(x_i - x)}{\partial x} = \frac{\partial \sum_i |x_i - x|}{\partial x} =$$

$$\frac{\partial \sum_i f(x_i - x)}{\partial x} = \frac{\partial \sum_i |x_i - x|}{\partial x} = \sum_i \left(\mathbf{I}_{x_i \leq x} - \mathbf{I}_{x_i \geq x} \right) = \sum_i \mathbf{I}_{x_i \leq x} - \sum_i \mathbf{I}_{x_i \geq x}$$

$$\frac{\partial \sum_i f(x_i - x)}{\partial x} = \frac{\partial \sum_i |x_i - x|}{\partial x} = \sum_i \left(\mathbf{I}_{x_i \leq x} - \mathbf{I}_{x_i \geq x} \right) = \sum_i \mathbf{I}_{x_i \leq x} - \sum_i \mathbf{I}_{x_i \geq x}$$

$$\left. \sum_i \mathbf{I}_{x_i \leq x} - \sum_i \mathbf{I}_{x_i \geq x} \right|_{x^\star} \hat{=}\ 0$$

$$\frac{\partial \sum_i f(x_i - x)}{\partial x} = \frac{\partial \sum_i |x_i - x|}{\partial x} = \sum_i \left(\mathbf{I}_{x_i \le x} - \mathbf{I}_{x_i \ge x} \right) = \sum_i \mathbf{I}_{x_i \le x} - \sum_i \mathbf{I}_{x_i \ge x}$$

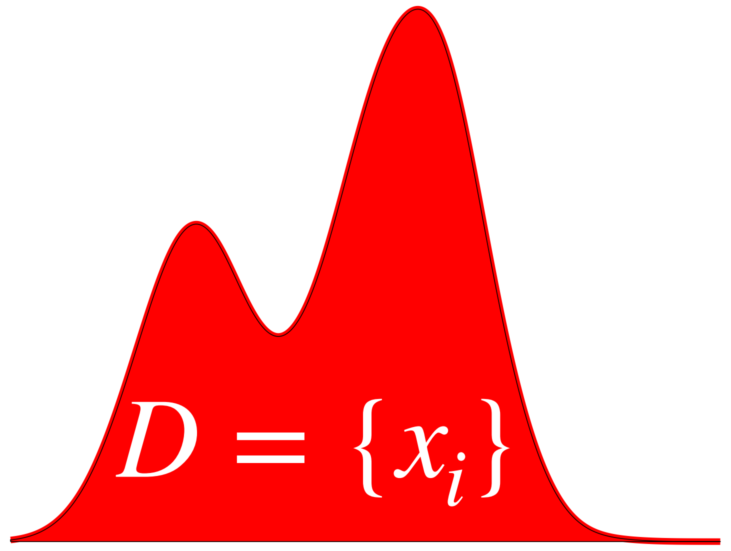
$$\left. \sum_i \mathbf{I}_{x_i \le x} - \sum_i \mathbf{I}_{x_i \ge x} \right|_{x^\star} \hat{=}\ 0 \implies \sum_i \mathbf{I}_{x_i \le x^\star} = \sum_i \mathbf{I}_{x_i \ge x^\star} \implies$$

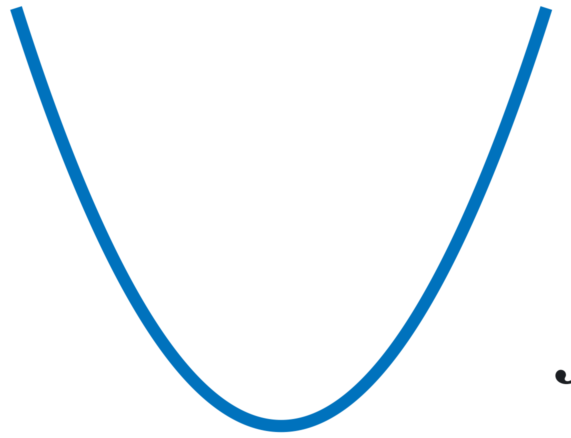
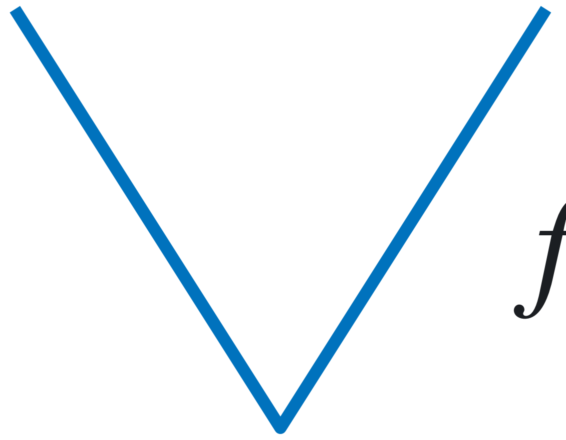
$$\frac{\partial \sum_i f(x_i - x)}{\partial x} = \frac{\partial \sum_i |x_i - x|}{\partial x} = \sum_i \left(\mathbf{I}_{x_i \leq x} - \mathbf{I}_{x_i \geq x}\right) = \sum_i \mathbf{I}_{x_i \leq x} - \sum_i \mathbf{I}_{x_i \geq x}$$

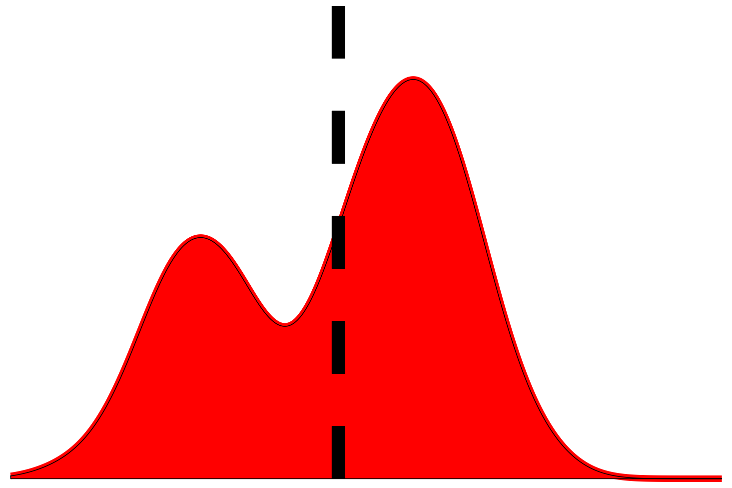
$$\sum_i \mathbf{I}_{x_i \leq x} - \sum_i \mathbf{I}_{x_i \geq x} \Bigg|_{x^\star} \hat{=}\ 0 \implies \sum_i \mathbf{I}_{x_i \leq x^\star} = \sum_i \mathbf{I}_{x_i \geq x^\star} \implies$$

$$\implies \frac{\sum_i \mathbf{I}_{x_i \leq x^\star}}{\sum_i \mathbf{I}_{x_i \geq x^\star}} = 1$$ Median

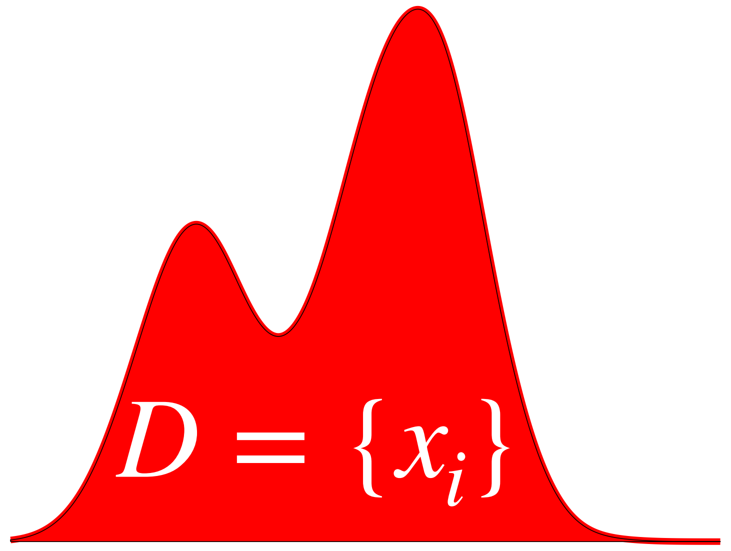
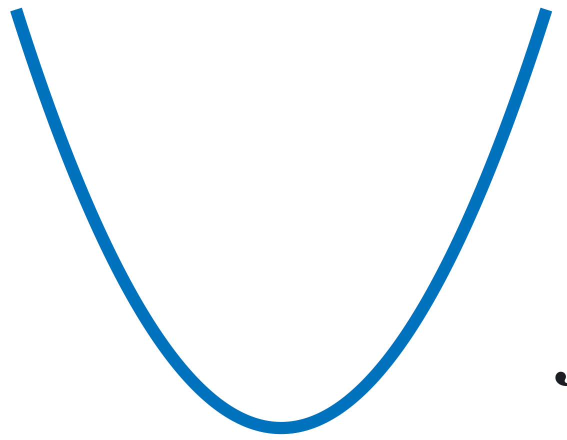
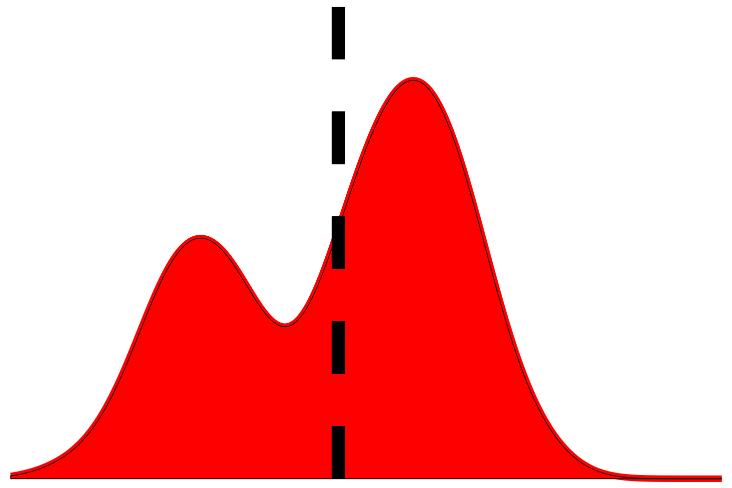
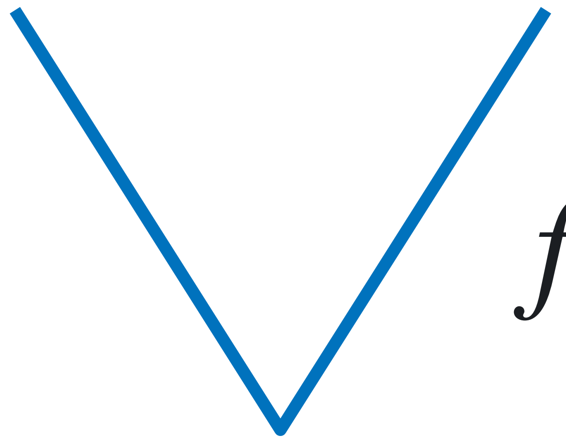
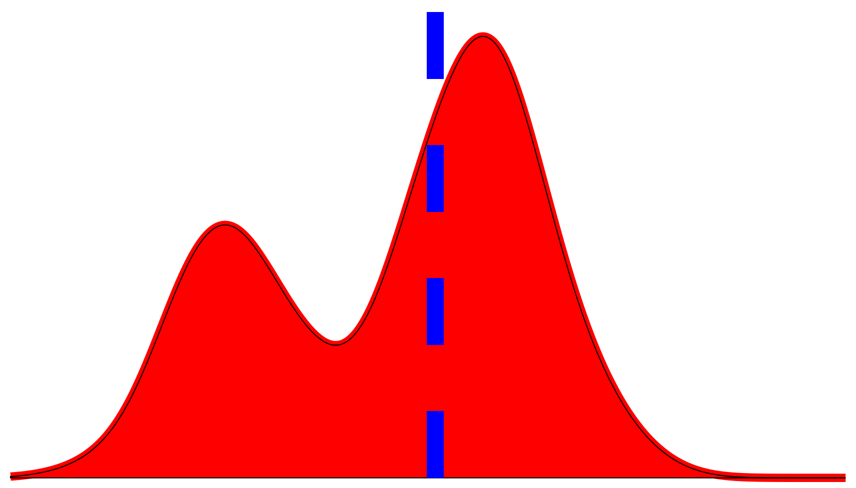
How can we **learn** from data ?



	$f(\delta)$	$x^\star = \underset{x}{\mathrm{argmin}} \sum_i f(x_i - x)$	
$\mathcal{L}_2$ regression	 $f(\delta) = \delta^2$	$x^\star = \frac{1}{N} \sum_i x_i$	Mean
$\mathcal{L}_1$ regression	 $f(\delta) = \vert\delta\vert$		



How can we **learn** from data ?

	$f(\delta)$	$x^\star = \underset{x}{\mathrm{argmin}} \sum_i f(x_i - x)$		
$\mathscr{L}_2$ regression	 $f(\delta) = \delta^2$	$x^\star = \frac{1}{N} \sum_i x_i$	Mean	
$\mathscr{L}_1$ regression	 $f(\delta) = \lvert\delta\rvert$	$x^\star = F_X^{-1}(0.5)$	Median $\tau = 0.5$	

$$\frac{\partial \sum_i f(x_i - x)}{\partial x} = \sum_i \left(\mathbf{I}_{x_i \le x} - \mathbf{I}_{x_i \ge x}\right) = \sum_i \mathbf{I}_{x_i \le x} - \sum_i \mathbf{I}_{x_i \ge x}$$

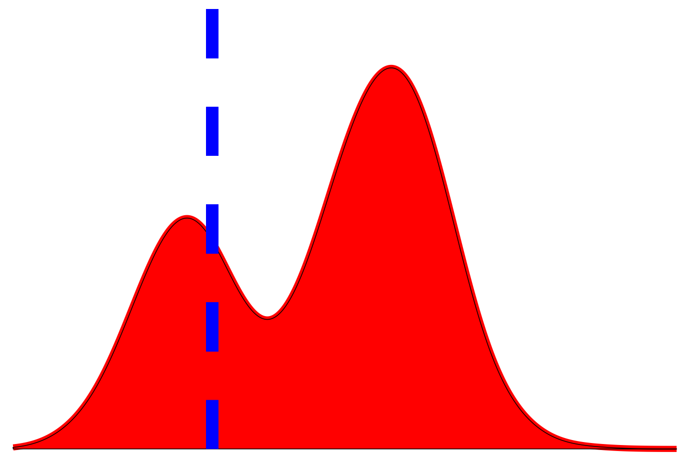
$$\sum_i \mathbf{I}_{x_i \le x} - \sum_i \mathbf{I}_{x_i \ge x}\Bigg|_{x^\star} \hat{=}\ 0 \implies \sum_i \mathbf{I}_{x_i \le x^\star} = \sum_i \mathbf{I}_{x_i \ge x^\star} \implies$$

$$\implies \frac{\sum_i \mathbf{I}_{x_i \le x^\star}}{\sum_i \mathbf{I}_{x_i \ge x^\star}} = 1$$ Median

$$\frac{\partial \sum_i f(x_i - x)}{\partial x} = \sum_i \left(\mathbf{I}_{x_i \leq x} - \mathbf{I}_{x_i \geq x} \right) = \sum_i \mathbf{I}_{x_i \leq x} - \sum_i \mathbf{I}_{x_i \geq x}$$

$$\sum_i \mathbf{I}_{x_i \leq x} - \sum_i \mathbf{I}_{x_i \geq x} \Bigg|_{x^\star} \hat{=}\ 0 \implies \sum_i \mathbf{I}_{x_i \leq x^\star} = \sum_i \mathbf{I}_{x_i \geq x^\star} \implies$$

$$\implies \frac{\sum_i \mathbf{I}_{x_i \leq x^\star}}{\sum_i \mathbf{I}_{x_i \geq x^\star}} = 1$$

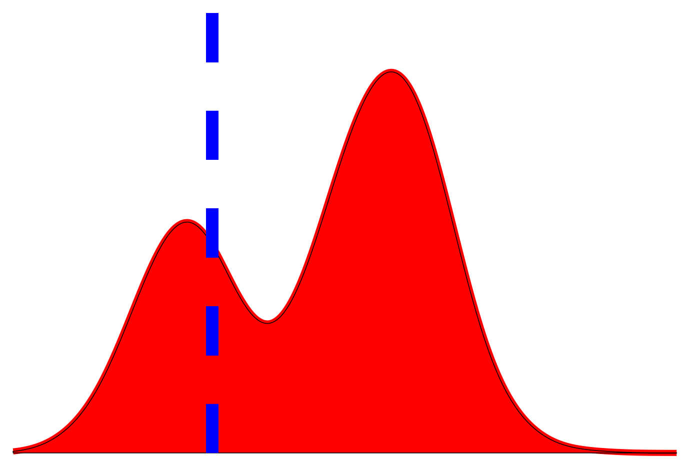


1/4 Quantile?

$$\frac{\partial \sum_i f(x_i - x)}{\partial x} = \sum_i \left(\mathbf{I}_{x_i \le x} - \mathbf{I}_{x_i \ge x} \right) = \sum_i \mathbf{I}_{x_i \le x} - \sum_i \mathbf{I}_{x_i \ge x}$$

$$\sum_i \mathbf{I}_{x_i \le x} - \sum_i \mathbf{I}_{x_i \ge x} \Big|_{x^\star} \hat{=}\, 0 \implies \sum_i \mathbf{I}_{x_i \le x^\star} = \sum_i \mathbf{I}_{x_i \ge x^\star} \implies$$

$$\implies \frac{\sum_i \mathbf{I}_{x_i \le x^\star}}{\sum_i \mathbf{I}_{x_i \ge x^\star}} = \frac{\frac{1}{4}}{\frac{3}{4}}$$

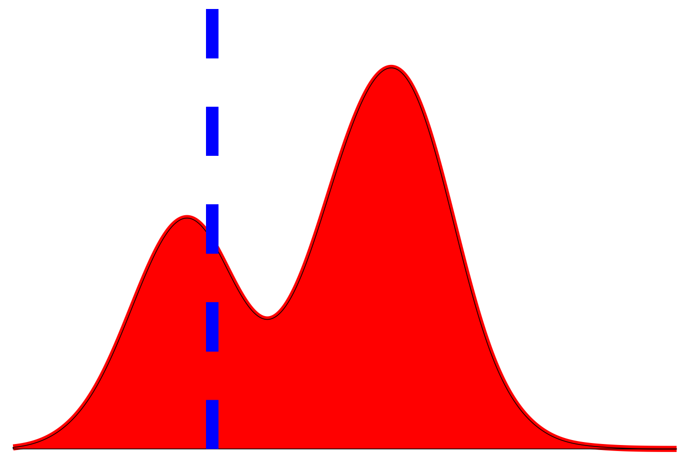


1/4 Quantile?

$$\frac{\partial \sum_i f(x_i - x)}{\partial x} = \sum_i \left(\mathbf{I}_{x_i \le x} - \mathbf{I}_{x_i \ge x} \right) = \sum_i \mathbf{I}_{x_i \le x} - \sum_i \mathbf{I}_{x_i \ge x}$$

$$\sum_i \mathbf{I}_{x_i \le x} - \sum_i \mathbf{I}_{x_i \ge x} \Big|_{x^\star} \hat{=}\ 0 \implies \sum_i \frac{3}{4} \mathbf{I}_{x_i \le x^\star} = \sum_i \frac{1}{4} \mathbf{I}_{x_i \ge x^\star} \implies$$

$$\implies \frac{\sum_i \mathbf{I}_{x_i \le x^\star}}{\sum_i \mathbf{I}_{x_i \ge x^\star}} = \frac{\frac{1}{4}}{\frac{3}{4}}$$

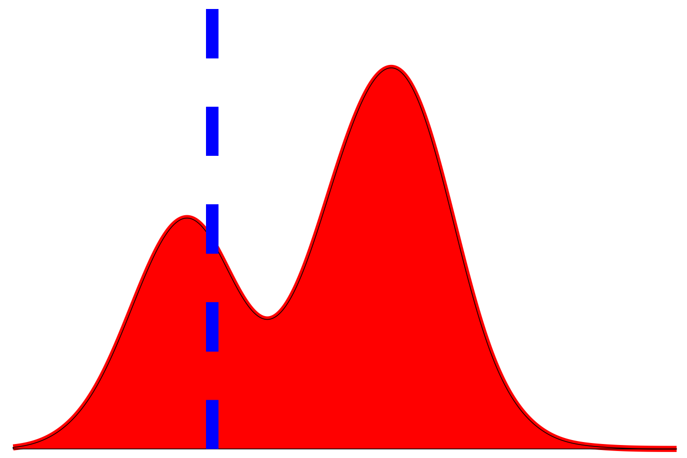


1/4 Quantile?

$$\frac{\partial \sum_i f(x_i - x)}{\partial x} = \sum_i \left(\mathbf{I}_{x_i \le x} - \mathbf{I}_{x_i \ge x}\right) = \sum_i \mathbf{I}_{x_i \le x} - \sum_i \mathbf{I}_{x_i \ge x}$$

$$\sum_i \frac{3}{4}\mathbf{I}_{x_i \le x} - \sum_i \frac{1}{4}\mathbf{I}_{x_i \ge x}\Bigg|_{x^\star} \mathrel{\hat{=}} 0 \implies \sum_i \frac{3}{4}\mathbf{I}_{x_i \le x^\star} = \sum_i \frac{1}{4}\mathbf{I}_{x_i \ge x^\star} \implies$$

$$\implies \frac{\sum_i \mathbf{I}_{x_i \le x^\star}}{\sum_i \mathbf{I}_{x_i \ge x^\star}} = \frac{\frac{1}{4}}{\frac{3}{4}}$$

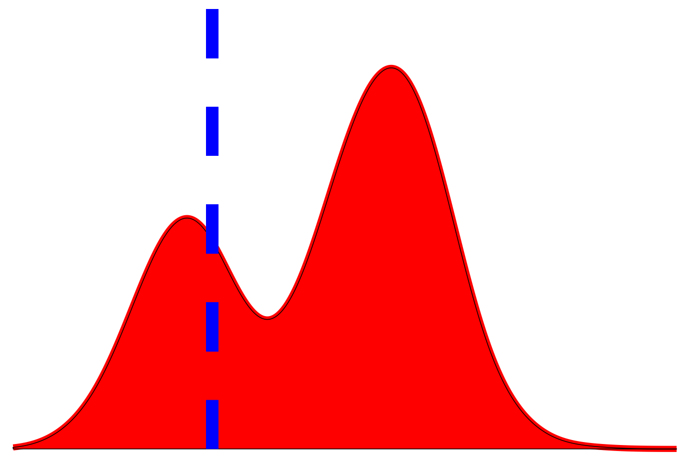


1/4 Quantile?

$$\frac{\partial \sum_i f(x_i - x)}{\partial x} = \sum_i \left(\frac{3}{4} \mathbf{I}_{x_i \le x} - \frac{1}{4} \mathbf{I}_{x_i \ge x} \right) = \sum_i \frac{3}{4} \mathbf{I}_{x_i \le x} - \sum_i \frac{1}{4} \mathbf{I}_{x_i \ge x}$$

$$\sum_i \frac{3}{4} \mathbf{I}_{x_i \le x} - \sum_i \frac{1}{4} \mathbf{I}_{x_i \ge x} \Bigg|_{x^\star} \hat{=}\ 0 \implies \sum_i \frac{3}{4} \mathbf{I}_{x_i \le x^\star} = \sum_i \frac{1}{4} \mathbf{I}_{x_i \ge x^\star} \implies$$

$$\implies \frac{\sum_i \mathbf{I}_{x_i \le x^\star}}{\sum_i \mathbf{I}_{x_i \ge x^\star}} = \frac{\frac{1}{4}}{\frac{3}{4}}$$



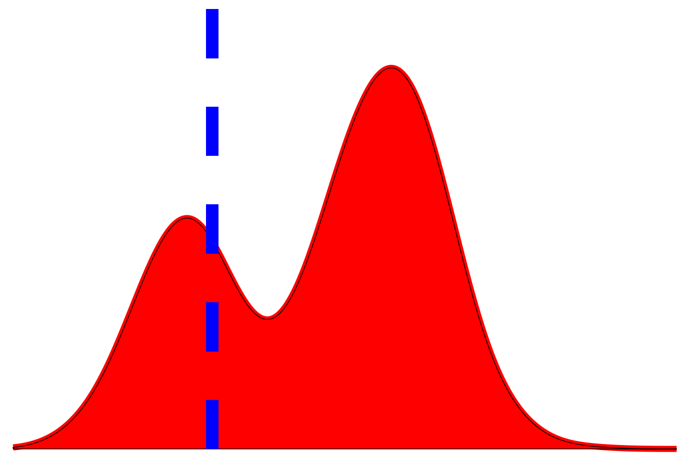
1/4 Quantile?

$$\frac{\partial \sum_i f(x_i - x)}{\partial x} = \sum_i \left(\frac{3}{4} \mathbf{I}_{x_i \le x} - \frac{1}{4} \mathbf{I}_{x_i \ge x} \right) = \sum_i \frac{3}{4} \mathbf{I}_{x_i \le x} - \sum_i \frac{1}{4} \mathbf{I}_{x_i \ge x}$$

$$\sum_i \frac{3}{4} \mathbf{I}_{x_i \le x} - \sum_i \frac{1}{4} \mathbf{I}_{x_i \ge x} \Bigg|_{x^\star} \hat{=}\ 0 \quad \Longrightarrow \quad \sum_i \frac{3}{4} \mathbf{I}_{x_i \le x^\star} = \sum_i \frac{1}{4} \mathbf{I}_{x_i \ge x^\star} \quad \Longrightarrow$$

$$\Longrightarrow \frac{\sum_i \mathbf{I}_{x_i \le x^\star}}{\sum_i \mathbf{I}_{x_i \ge x^\star}} = \frac{\frac{1}{4}}{\frac{3}{4}}$$

$$f(x_i - x) = \begin{cases} 1/4\,(x_i - x), & x_i \ge x \\ -3/4\,(x_i - x), & x_i < x \end{cases}$$



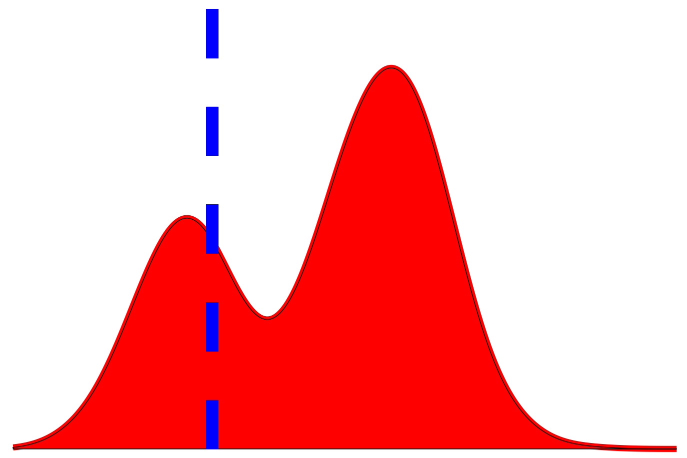
1/4 Quantile?

$$\frac{\partial \sum_i f(x_i - x)}{\partial x} = \sum_i \left(\frac{3}{4} \mathbf{I}_{x_i \le x} - \frac{1}{4} \mathbf{I}_{x_i \ge x} \right) = \sum_i \frac{3}{4} \mathbf{I}_{x_i \le x} - \sum_i \frac{1}{4} \mathbf{I}_{x_i \ge x}$$

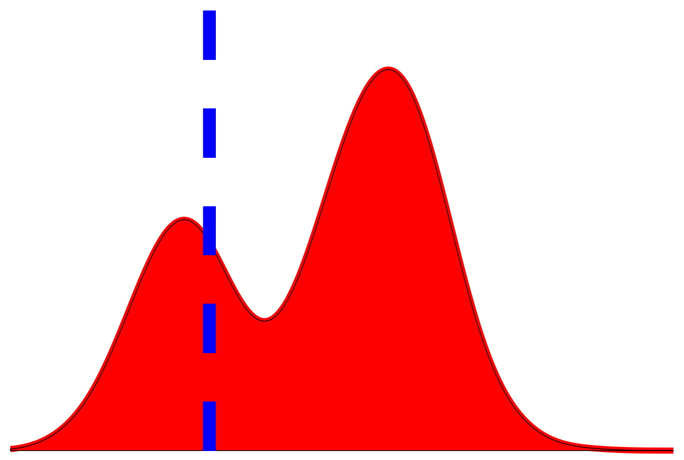
$$\left. \sum_i \frac{3}{4} \mathbf{I}_{x_i \le x} - \sum_i \frac{1}{4} \mathbf{I}_{x_i \ge x} \right|_{x^\star} \hat{=}\, 0 \implies \sum_i \frac{3}{4} \mathbf{I}_{x_i \le x^\star} = \sum_i \frac{1}{4} \mathbf{I}_{x_i \ge x^\star} \implies$$

$$\implies \frac{\sum_i \mathbf{I}_{x_i \le x^\star}}{\sum_i \mathbf{I}_{x_i \ge x^\star}} = \frac{\frac{1}{4}}{\frac{3}{4}}$$

$$f(\delta) = \begin{cases} 1/4\,\delta, & \delta \ge 0 \\ -3/4\,\delta, & \delta < 0 \end{cases}$$

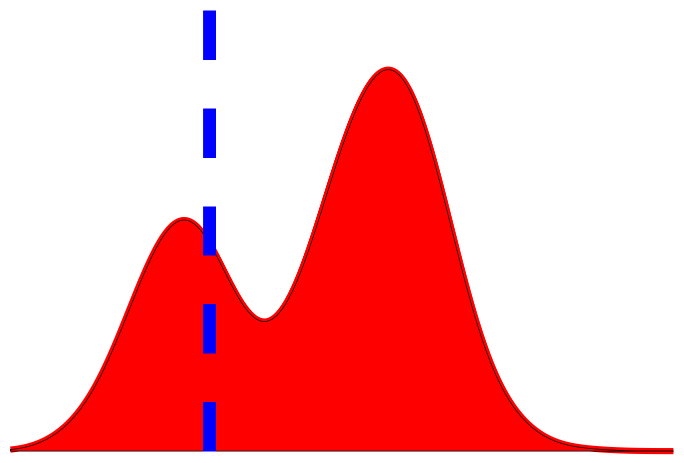


1/4 Quantile?



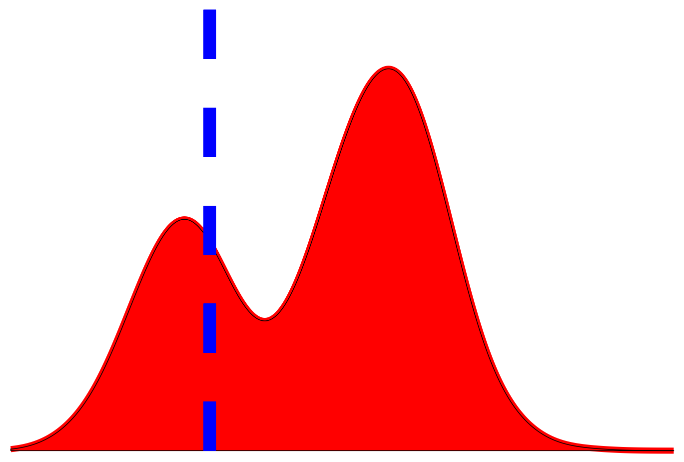
1/4 Quantile?

$$f(\delta) = \begin{cases} 1/4\,\delta, & \delta \geq 0 \\ -3/4\,\delta, & \delta < 0 \end{cases}$$



1/4 Quantile?

$$f(\delta) = \begin{cases} 1/4\,\delta, & \delta \geq 0 \\ -3/4\,\delta, & \delta < 0 \end{cases}$$

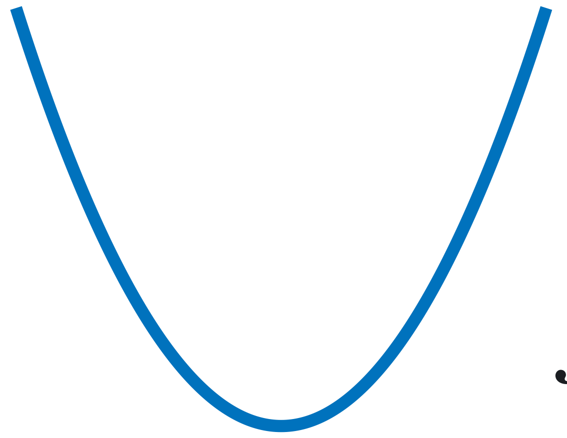
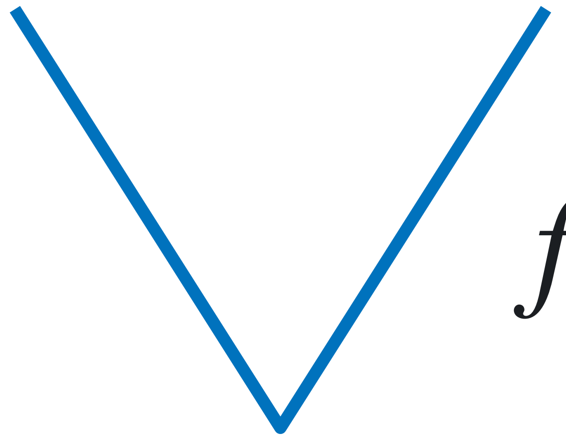


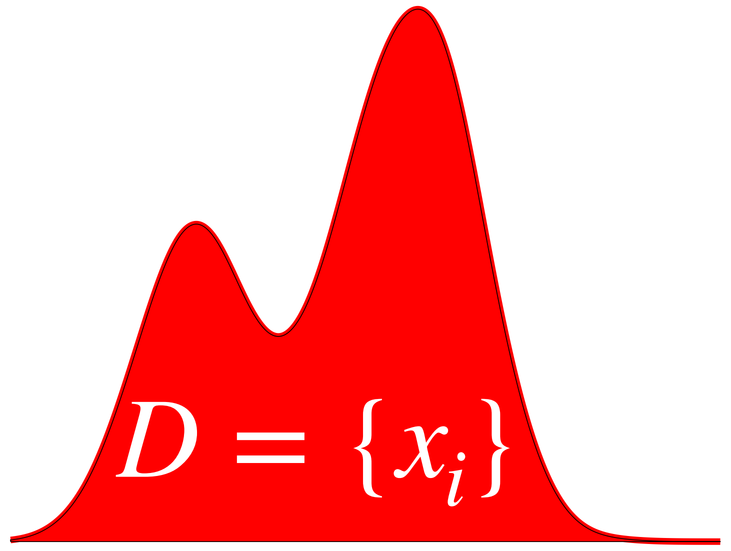
τ-Quantile?

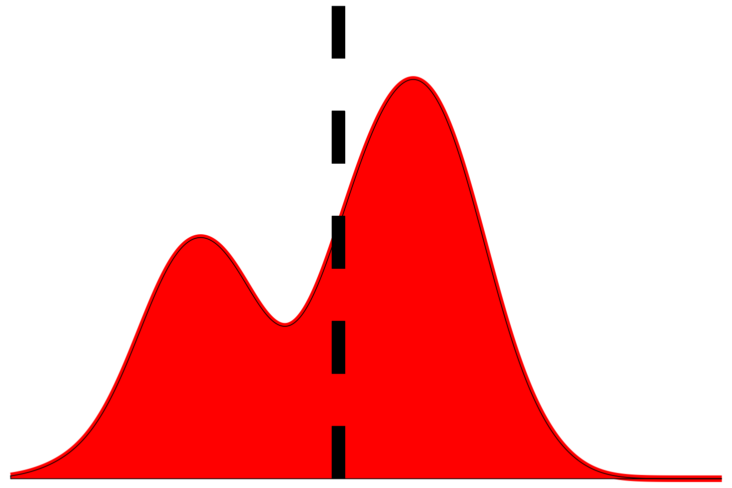
$$f_\tau(\delta) = \begin{cases} \tau\,\delta, & \delta \geq 0 \\ (\tau - 1)\,\delta, & \delta < 0 \end{cases}$$

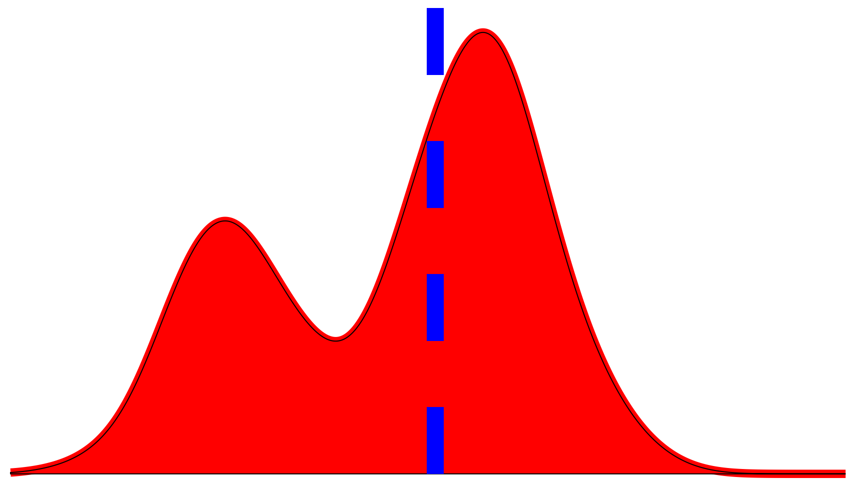
$$f_\tau(\delta) = \delta\left(\tau - \mathbf{I}_{\delta<0}\right)$$

How can we **learn** from data ?

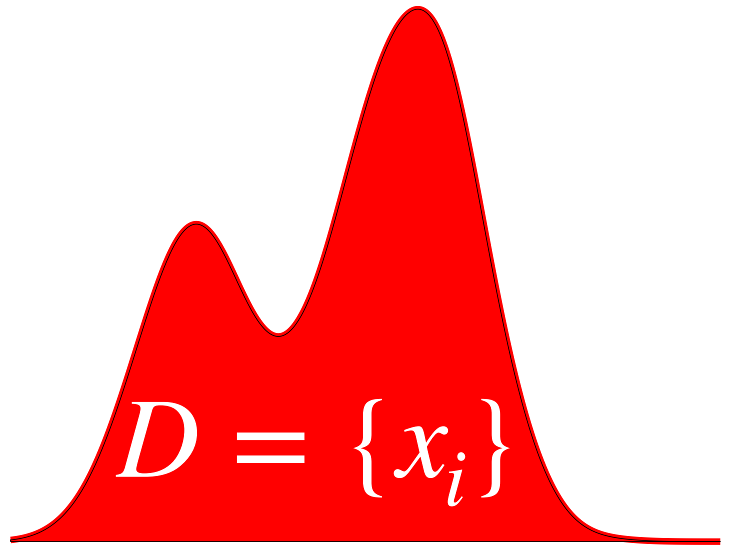
	$f(\delta)$	$x^\star = \underset{x}{\text{argmin}} \sum_i f(x_i - x)$	
$\mathscr{L}_2$ regression	 $f(\delta) = \delta^2$	$x^\star = \frac{1}{N} \sum_i x_i$	Mean
$\mathscr{L}_1$ regression	 $f(\delta) = \lvert \delta \rvert$	$x^\star = F_X^{-1}(0.5)$	Median $\tau = 0.5$

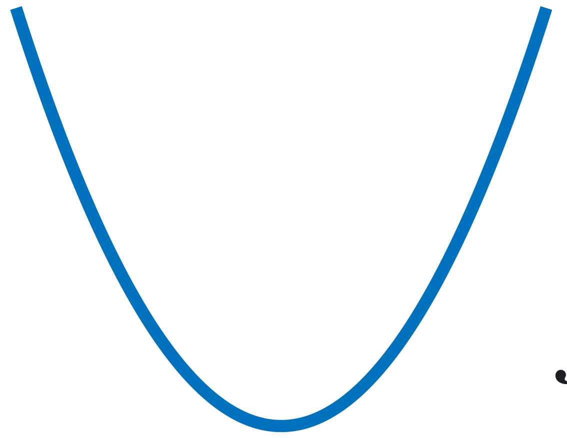
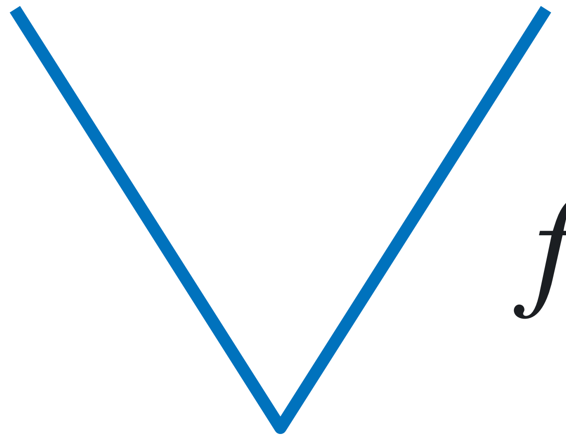
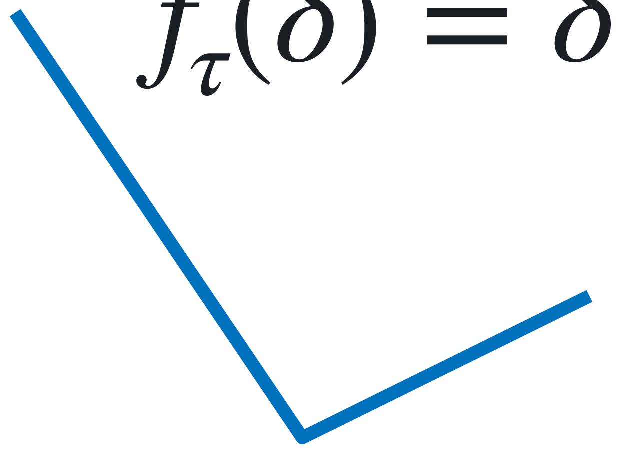


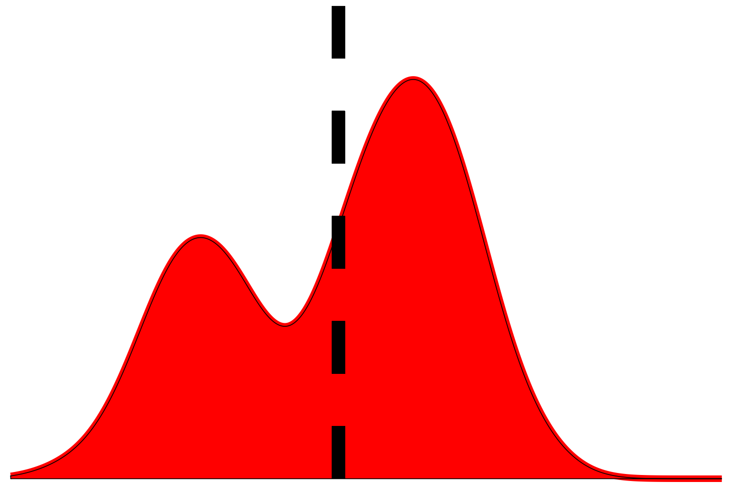


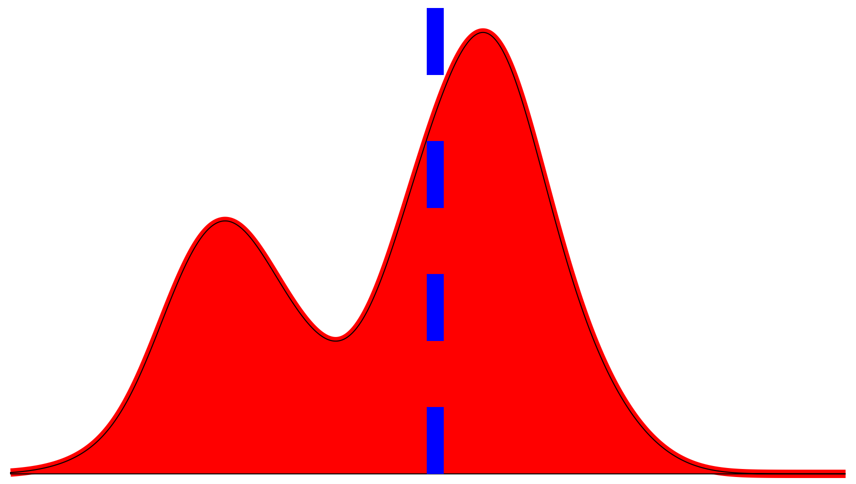


How can we **learn** from data ?

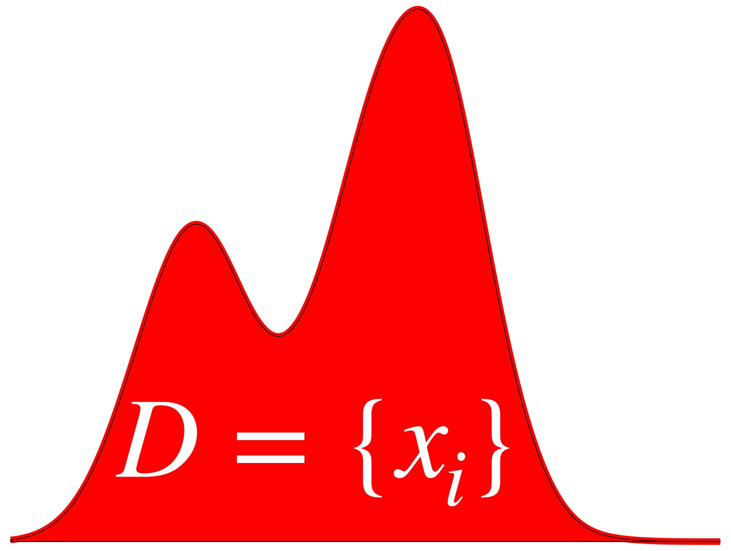


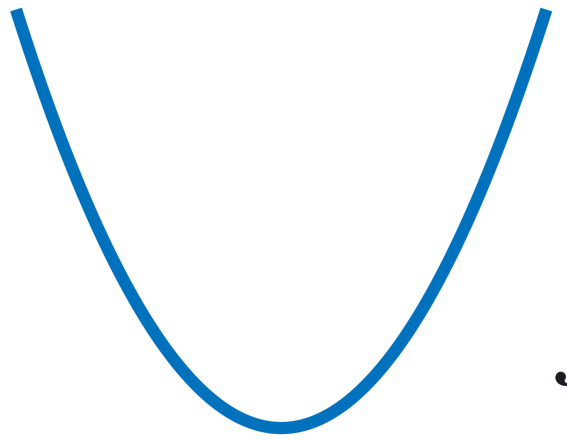
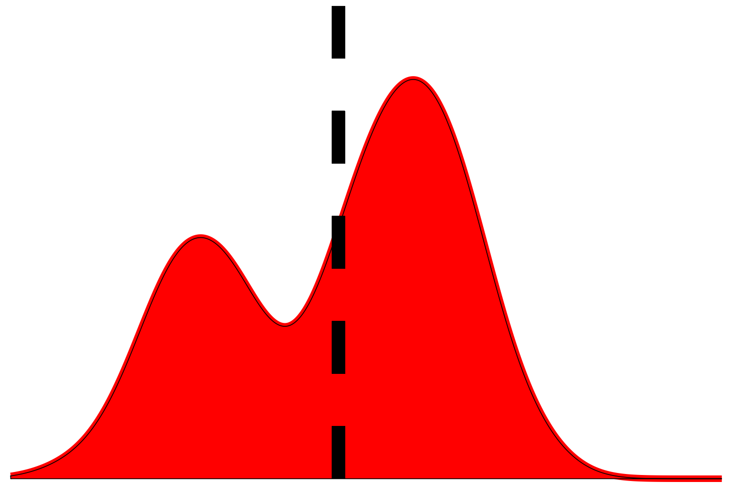
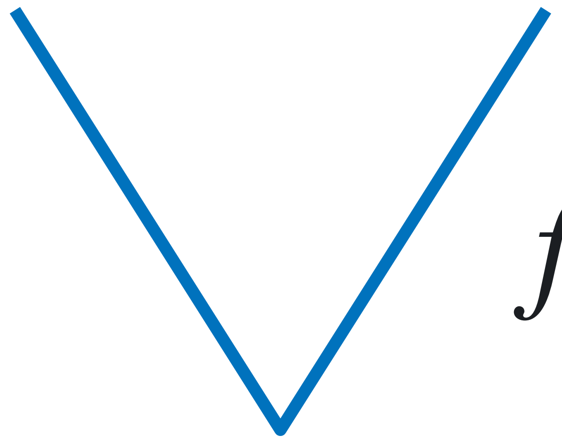
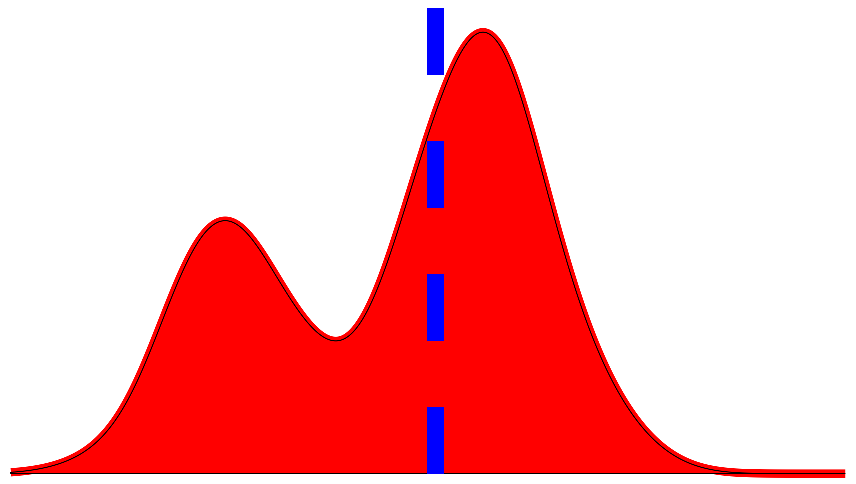
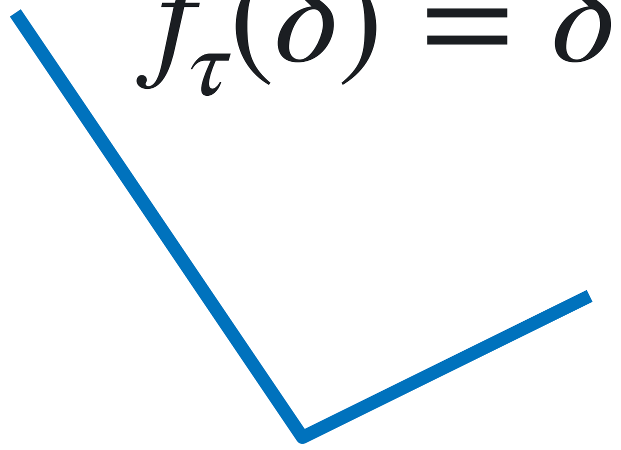
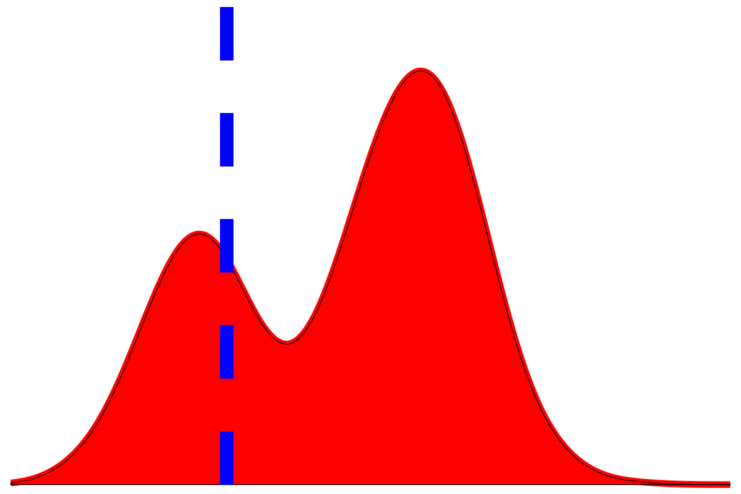
	$f(\delta)$	$x^\star = \underset{x}{\operatorname{argmin}} \sum_i f(x_i - x)$	
$\mathscr{L}_2$ regression	 $f(\delta) = \delta^2$	$x^\star = \frac{1}{N} \sum_i x_i$	Mean
$\mathscr{L}_1$ regression	 $f(\delta) = \lvert \delta \rvert$	$x^\star = F_X^{-1}(0.5)$	Median $\tau = 0.5$
	$f_\tau(\delta) = \delta \left(\tau - \mathbf{I}_{\delta<0}\right)$ 		



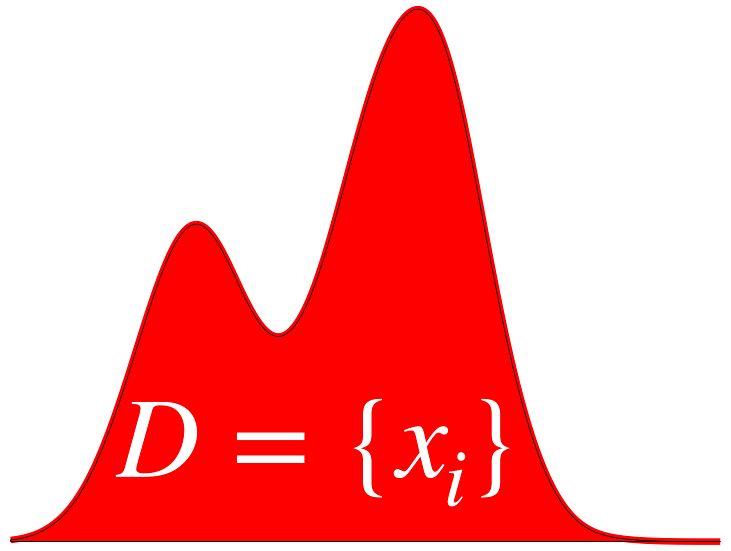
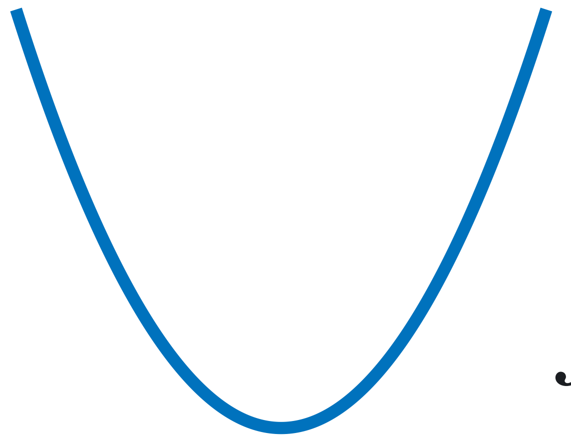
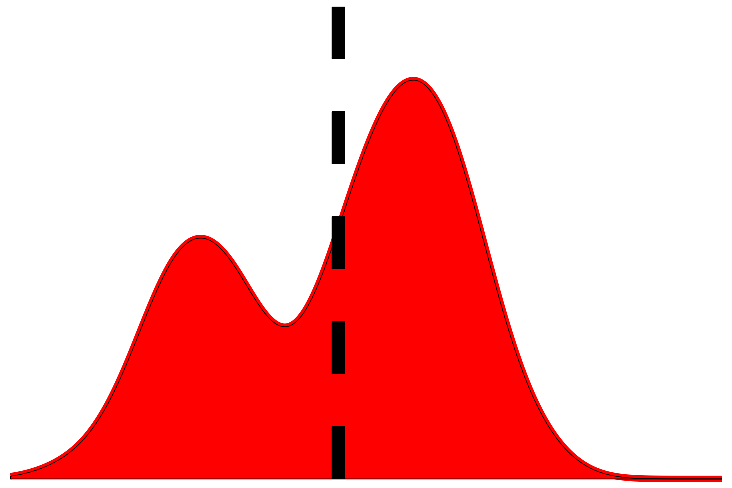
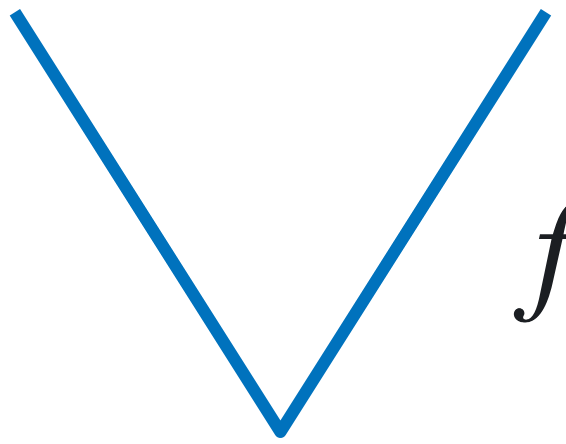
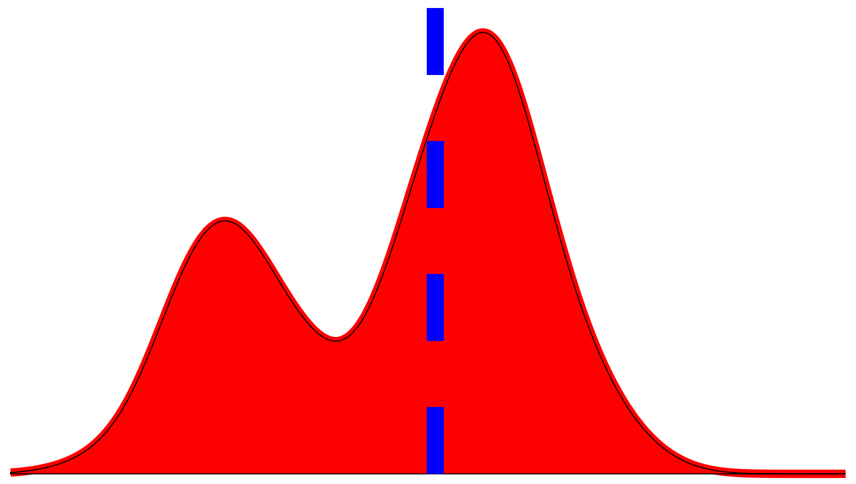
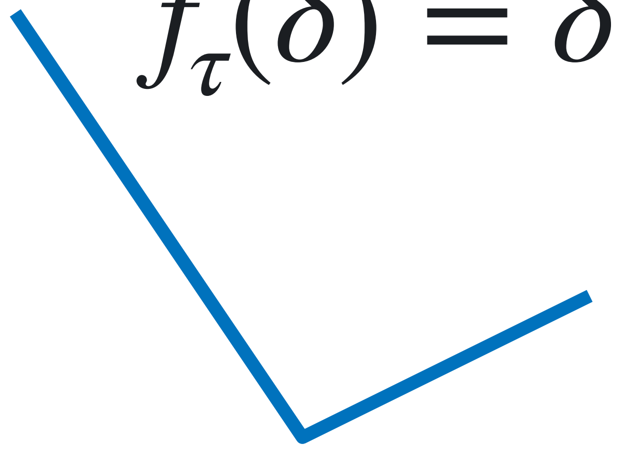
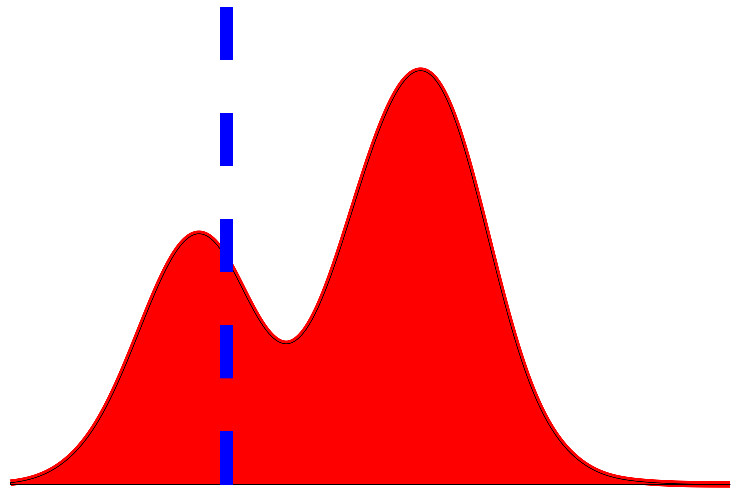


How can we **learn** from data ?

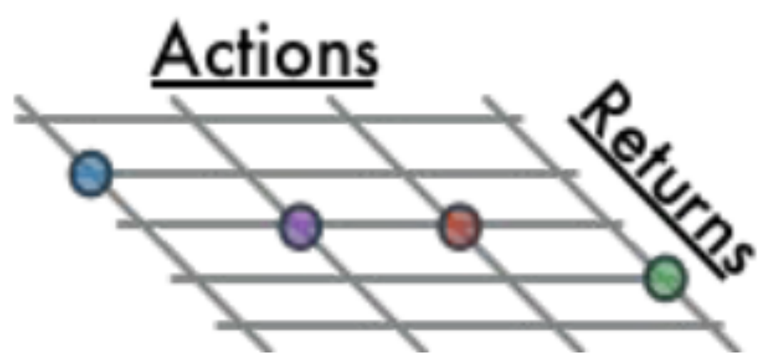


	$f(\delta)$	$x^\star = \underset{x}{\text{argmin}} \sum_i f(x_i - x)$		
$\mathcal{L}_2$ regression	 $f(\delta) = \delta^2$	$x^\star = \frac{1}{N} \sum_i x_i$	Mean	
$\mathcal{L}_1$ regression	 $f(\delta) = \lvert \delta \rvert$	$x^\star = F_X^{-1}(0.5)$	Median $\tau = 0.5$	
	$f_\tau(\delta) = \delta \left(\tau - \mathbf{I}_{\delta<0}\right)$ 	$x^\star = F_X^{-1}(\tau)$	τ-Quantile	

How can we **learn** from data ?

	$f(\delta)$	$x^\star = \underset{x}{\operatorname{argmin}} \sum_i f(x_i - x)$		
$\mathscr{L}_2$ regression	 $f(\delta) = \delta^2$	$x^\star = \frac{1}{N}\sum_i x_i$	Mean	
$\mathscr{L}_1$ regression	 $f(\delta) = \|\delta\|$	$x^\star = F_X^{-1}(0.5)$	Median $\tau = 0.5$	
Quantile regression	$f_\tau(\delta) = \delta\left(\tau - \mathbf{I}_{\delta<0}\right)$ 	$x^\star = F_X^{-1}(\tau)$	τ-Quantile	

DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}}\; Q^{\theta}(s_{t+1}, a)$$

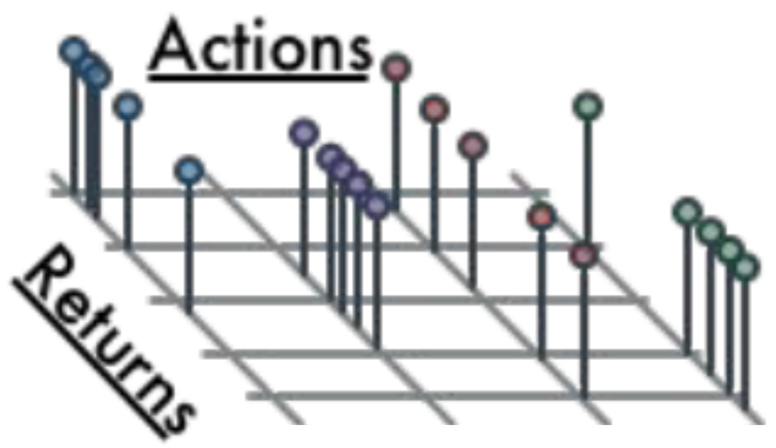
$$q' = Q^{\theta}(s_{t+1}, a^\star)$$

$$q = Q^{\theta}(s_t, a_t)$$

$$\delta_t = r_t + \gamma q' - q$$

$$\mathscr{L}_{DQN} = \delta_t^2$$

QR-DQN

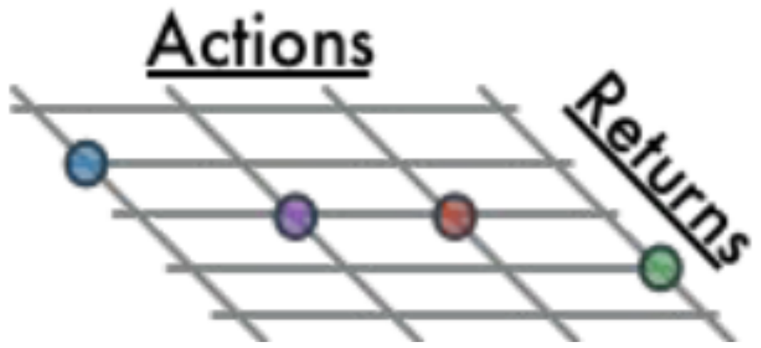
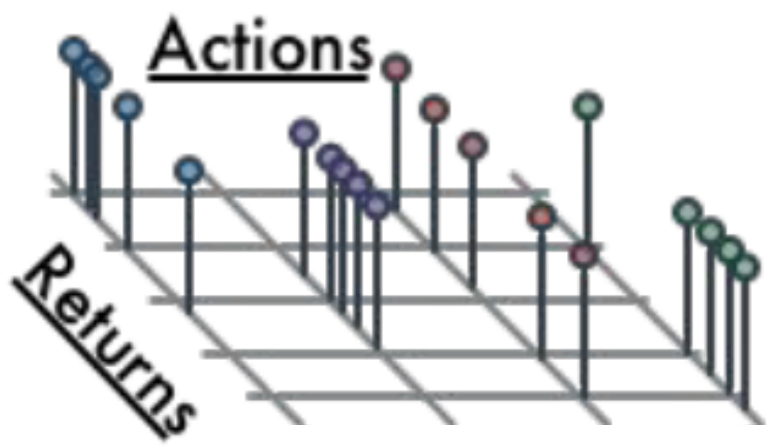


$$\{s_t, a_t, s_{t+1}, r_t\}$$

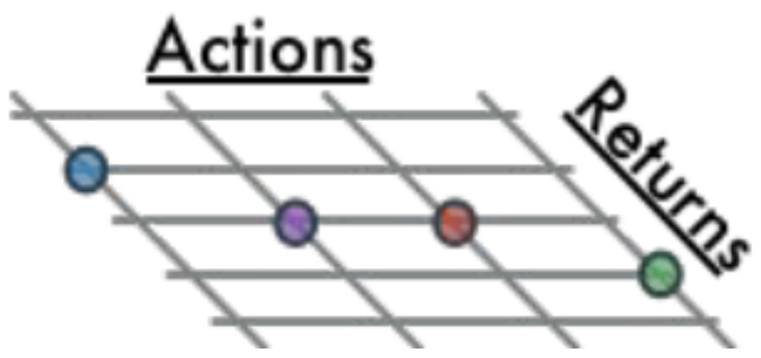
$$a^\star = \underset{a}{\text{argmax}}\; \mathbb{E}\left[Z_\tau^{\theta}(s_{t+1}, a)\right]$$

$$\forall \tau, \tau' \;\Bigg|\; \begin{aligned} z' &= Z_{\tau'}^{\theta}(s_{t+1}, a^\star) \\ z &= Z_\tau^{\theta}(s_t, a_t) \\ \delta_t^{\tau,\tau'} &= r_t + \gamma z' - z \end{aligned}$$

$$\mathscr{L}_{QR-DQN} = \,?$$

DQN	QR-DQN
	
$\{s_t, a_t, s_{t+1}, r_t\}$	$\{s_t, a_t, s_{t+1}, r_t\}$
$a^\star = \underset{a}{\text{argmax}}\ Q^{\theta}(s_{t+1}, a)$	$a^\star = \underset{a}{\text{argmax}}\ \mathbb{E}\left[Z^{\theta}_{\tau}(s_{t+1}, a)\right]$
$q' = Q^{\theta}(s_{t+1}, a^\star)$	$\forall \tau, \tau' \;\vert\; z' = Z^{\theta}_{\tau'}(s_{t+1}, a^\star)$
$q = Q^{\theta}(s_t, a_t)$	$z = Z^{\theta}_{\tau}(s_t, a_t)$
$\delta_t = r_t + \gamma q' - q$	$\delta_t^{\tau,\tau'} = r_t + \gamma z' - z$
$\mathscr{L}_{DQN} = \delta_t^2$	$\mathscr{L}_{QR-DQN} = \sum_{\tau} \sum_{\tau'} \delta_t^{\tau,\tau'} \left(\tau - \mathbf{I}_{\delta_t^{\tau,\tau'} < 0}\right)$

DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}} \; Q^{\theta}(s_{t+1}, a)$$

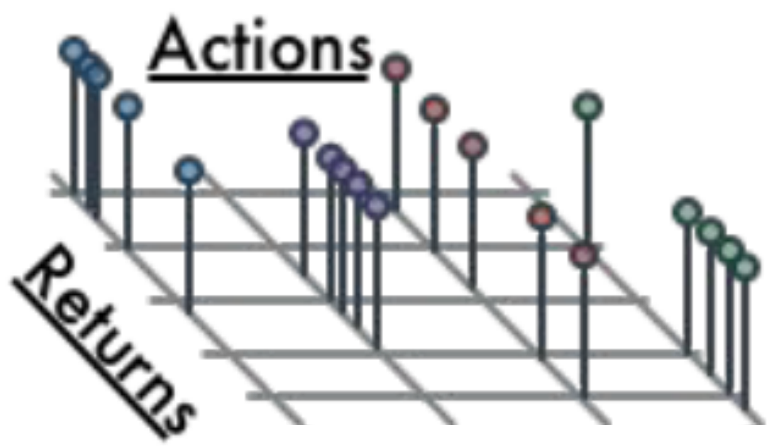
$$q' = Q^{\theta}(s_{t+1}, a^\star)$$

$$q = Q^{\theta}(s_t, a_t)$$

$$\delta_t = r_t + \gamma q' - q$$

$$\mathscr{L}_{DQN} = \delta_t^2$$

QR-DQN



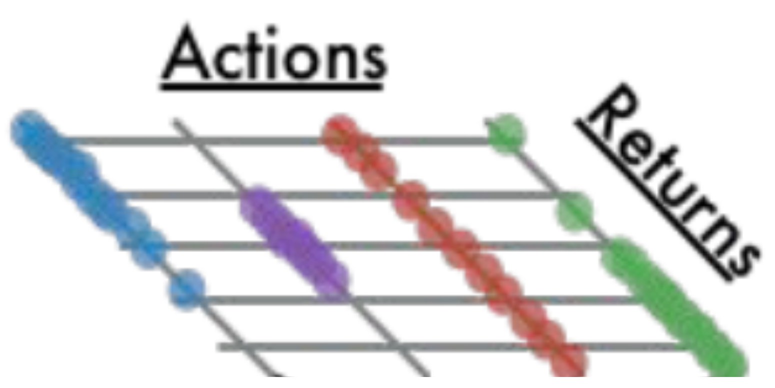
$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}} \; \mathbb{E}\left[Z_\tau^{\theta}(s_{t+1}, a)\right]$$

$$\forall \tau, \tau' \left| \begin{array}{l} z' = Z_{\tau'}^{\theta}(s_{t+1}, a^\star) \\ z = Z_\tau^{\theta}(s_t, a_t) \\ \delta_t^{\tau,\tau'} = r_t + \gamma z' - z \end{array} \right.$$

$$\mathscr{L}_{QR-DQN} = \sum_{\tau} \sum_{\tau'} \delta_t^{\tau,\tau'} \left(\tau - \mathbf{I}_{\delta_t^{\tau,\tau'} < 0}\right)$$

IQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}} \; \mathbb{E}\left[Z_\tau^{\theta}(s_{t+1}, a)\right]$$

$$\begin{array}{l} \forall \tau \sim U([0,1]) \\ \forall \tau' \sim U([0,1]) \end{array} \left| \begin{array}{l} z' = Z_{\tau'}^{\theta}(s_{t+1}, a^\star) \\ z = Z_\tau^{\theta}(s_t, a_t) \\ \delta_t^{\tau,\tau'} = r_t + \gamma z' - z \end{array} \right.$$

$$\mathscr{L}_{IQN} = \sum_{\tau} \sum_{\tau'} \delta_t^{\tau,\tau'} \left(\tau - \mathbf{I}_{\delta_t^{\tau,\tau'} < 0}\right)$$

Human normalised score (HNS)

$$\text{score} = \frac{\text{agent} - \text{random}}{\text{human} - \text{random}}$$

Human gap

$$\text{gap} = 1 - \text{clip}(\text{score},\ 0,\ 1) = \begin{cases} 1, & \text{random play} \\ 0, & \text{super-human} \end{cases}$$

RESULTS - **HNS** on Atari-57

	Mean	**Median**	**(human starts) Median**	**Human Gap**
DQN	228 %	79 %	68 %	0.334
Prio. Duel.	592 %	172 %	128 %	0.178
C51	701 %	178 %	125 %	0.152

RESULTS - **HNS** on Atari-57

smaller is better!

	Mean	Median	(human starts) Median	Human Gap
DQN	228 %	79 %	68 %	0.334
Prio. Duel.	592 %	172 %	128 %	0.178
C51	701 %	178 %	125 %	0.152

RESULTS - **HNS** on Atari-57

smaller is better!

	Mean	Median	(human starts) Median	Human Gap
DQN	228 %	79 %	68 %	0.334
Prio. Duel.	592 %	172 %	128 %	0.178
C51	701 %	178 %	125 %	0.152
QR-DQN	**864 %**	**193 %**	**153 %**	0.165

RESULTS - **HNS** on Atari-57

smaller is better!

	Mean	Median	(human starts) Median	Human Gap
DQN	228 %	79 %	68 %	0.334
Prio. Duel.	592 %	172 %	128 %	0.178
C51	701 %	178 %	125 %	0.152
QR-DQN	**864 %**	**193 %**	**153 %**	0.165
IQN	**1019 %**	**218 %**	**162 %**	**0.141**

RESULTS - **HNS** on Atari-57

smaller is better!

	Mean	Median	(human starts) Median	Human Gap
DQN	228 %	79 %	68 %	0.334
Prio. Duel.	592 %	172 %	128 %	0.178
C51	701 %	178 %	125 %	0.152
QR-DQN	**864 %**	**193 %**	**153 %**	0.165
IQN	**1019 %**	**218 %**	**162 %**	**0.141**
Rainbow (QR-DQN based)	**1189 %**	**230 %**	**125 %**	**0.144**

RESULTS - **HNS** on Atari-57

smaller is better!

	Mean	Median	(human starts) Median	Human Gap
DQN	228 %	79 %	68 %	0.334
Prio. Duel.	592 %	172 %	128 %	0.178
C51	701 %	178 %	125 %	0.152
QR-DQN	**864 %**	**193 %**	**153 %**	0.165
IQN	**1019 %**	**218 %**	**162 %**	**0.141**
Rainbow (QR-DQN based)	**1189 %**	**230 %**	**125 %**	**0.144**

RESULTS - **HNS** on Atari-57

smaller is better!

	Mean	Median	(human starts) Median	Human Gap
DQN	228 %	79 %	68 %	0.334
Prio. Duel.	592 %	172 %	128 %	0.178
C51	701 %	178 %	125 %	0.152
QR-DQN	**864 %**	**193 %**	**153 %**	0.165
IQN	**1019 %**	**218 %**	**162 %**	**0.141**
QR-DQN based **Rainbow**	**1189 %**	**230 %**	**125 %**	**0.144**

RESULTS - **HNS** on Atari-57

smaller is better!

	Mean	Median	(human starts) Median	Human Gap
DQN	228 %	79 %	68 %	0.334
Prio. Duel.	592 %	172 %	128 %	0.178
C51	701 %	178 %	125 %	0.152
QR-DQN	**864 %**	**193 %**	**153 %**	0.165
IQN	**1019 %**	**218 %**	**162 %**	**0.141**
Rainbow (QR-DQN based)	**1189 %**	**230 %**	**125 %**	**0.144**

RESULTS - **HNS** on Atari-57

smaller is better!

	Mean	Median	(human starts) Median	Human Gap
DQN	228 %	79 %	68 %	0.334
Prio. Duel.	592 %	172 %	128 %	0.178
C51	701 %	178 %	125 %	0.152
QR-DQN	**864 %**	**193 %**	**153 %**	0.165
IQN	**1019 %**	**218 %**	**162 %**	**0.141**
QR-DQN based **Rainbow**	**1189 %**	**230 %**	**125 %**	**0.144**

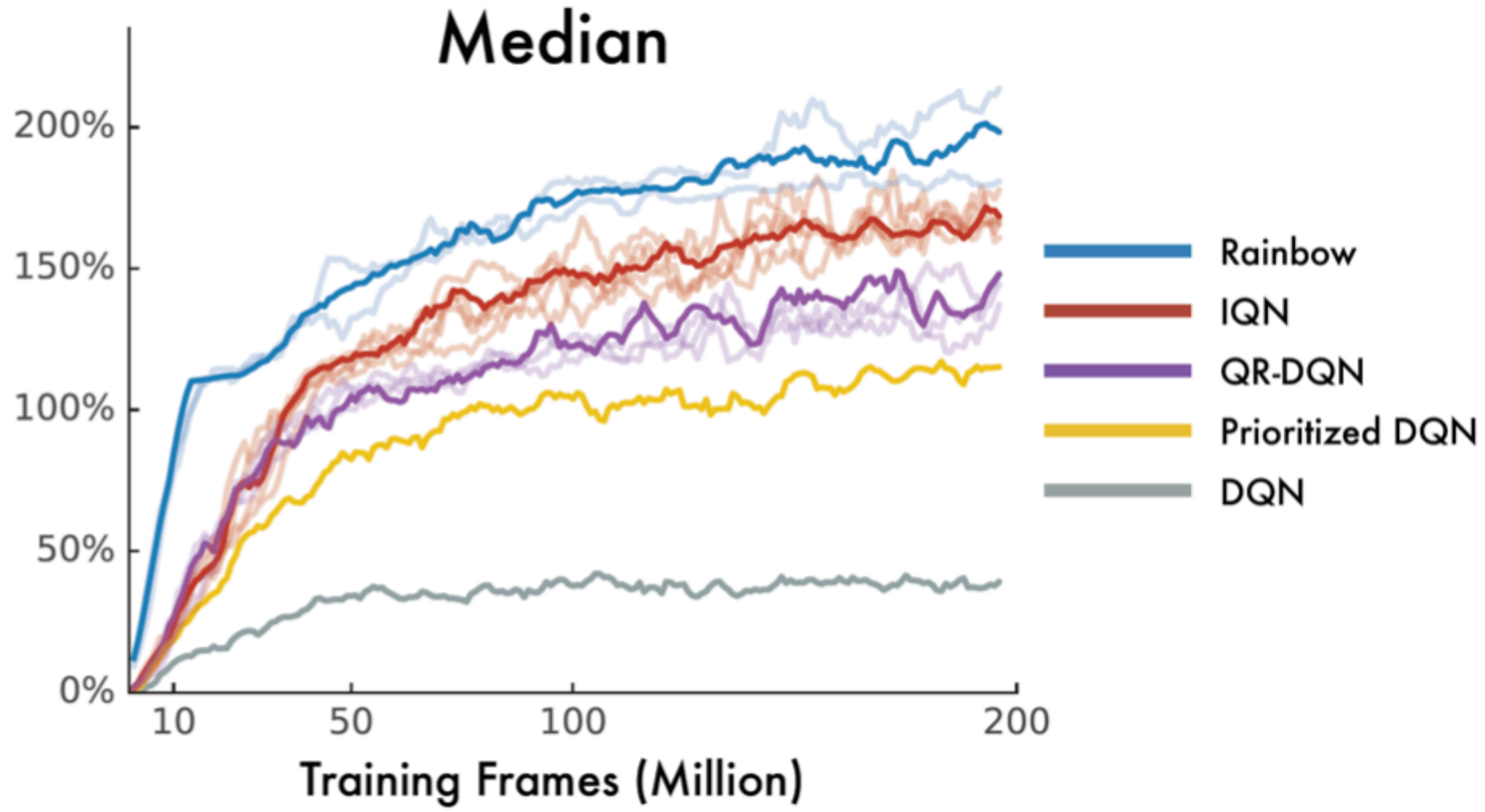
IQN outperforms Rainbow on **hardest** Atari gameas!

RESULTS - **HNS** on Atari-57

smaller is better!

	Mean	Median	(human starts) Median	Human Gap
DQN	228 %	79 %	68 %	0.334
Prio. Duel.	592 %	172 %	128 %	0.178
C51	701 %	178 %	125 %	0.152
QR-DQN	**864 %**	**193 %**	**153 %**	0.165
IQN	**1019 %**	**218 %**	**162 %**	**0.141**
QR-DQN based **Rainbow**	**1189 %**	**230 %**	**125 %**	**0.144**

IQN outperforms Rainbow on **hardest** Atari gameas!

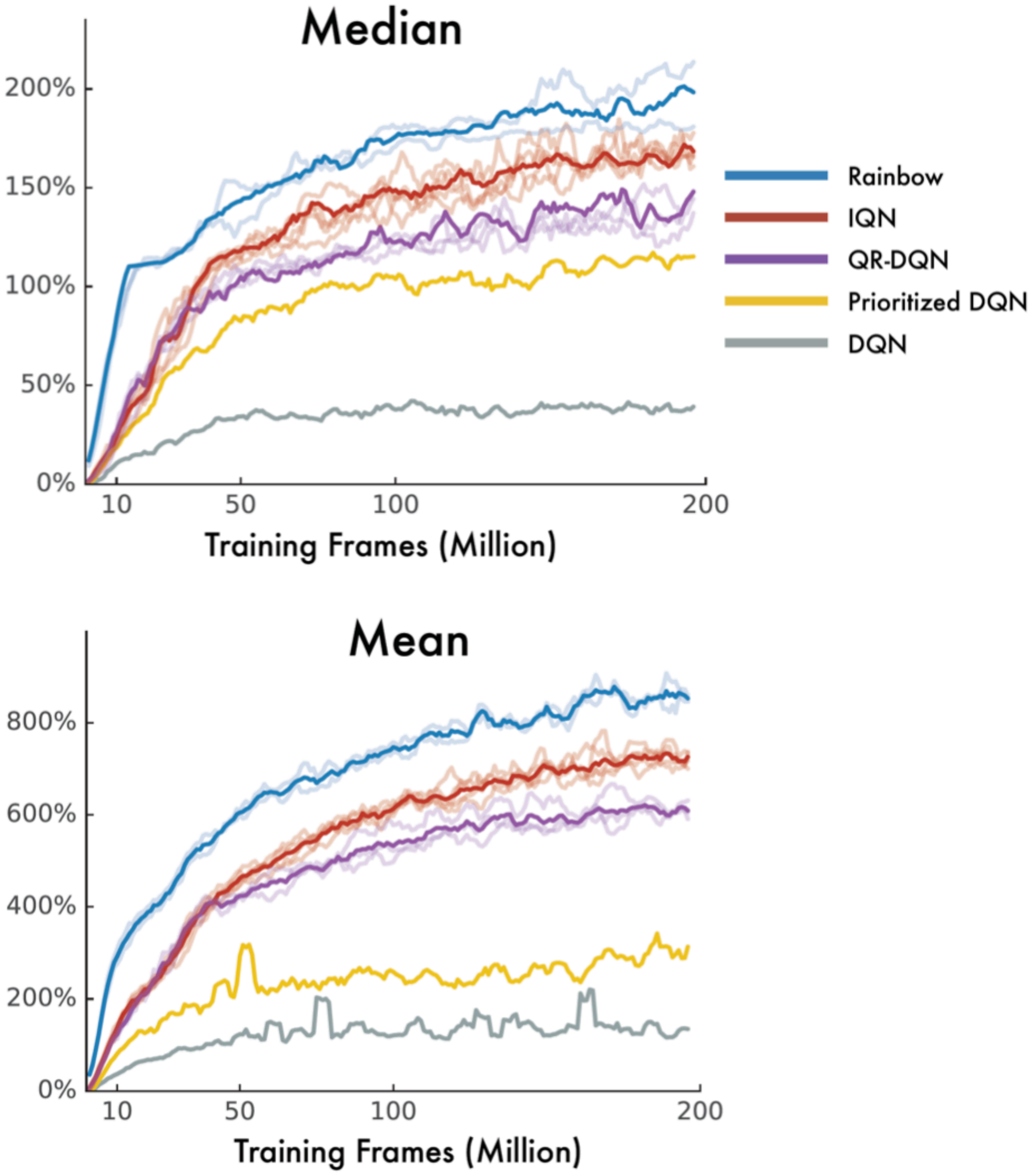


RESULTS - **HNS** on Atari-57

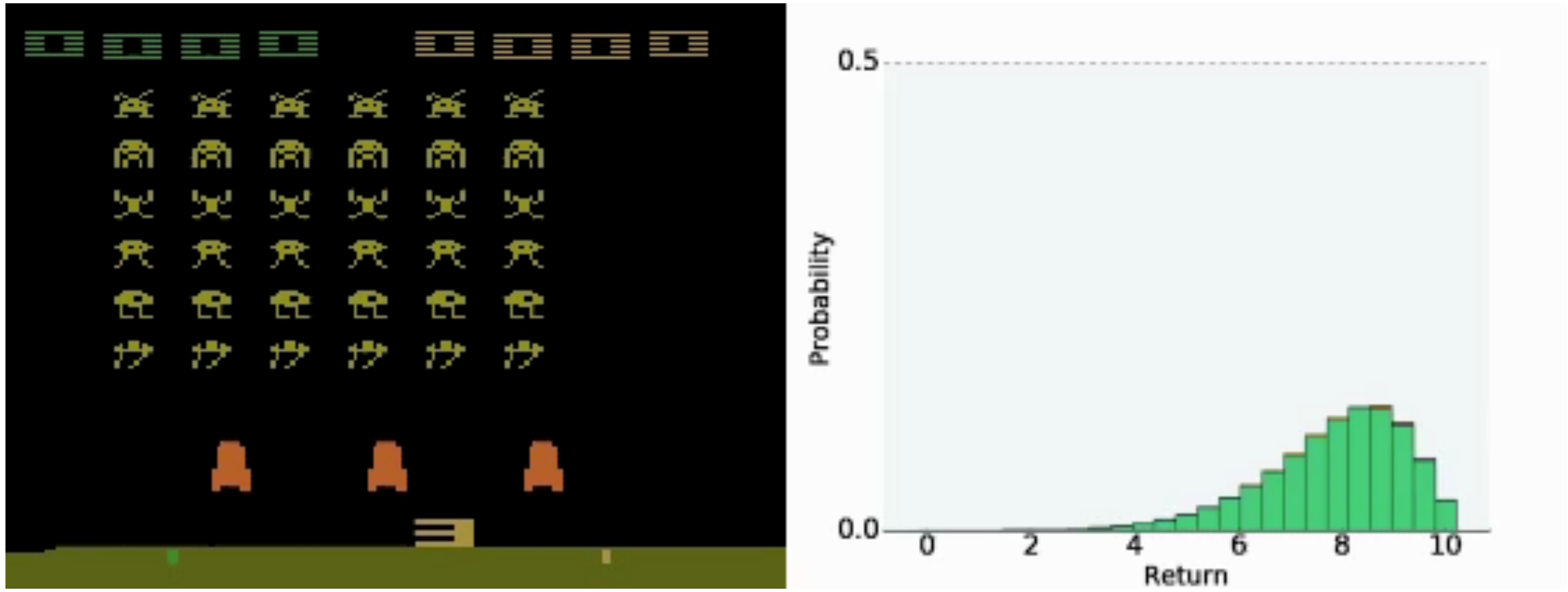
smaller is better!

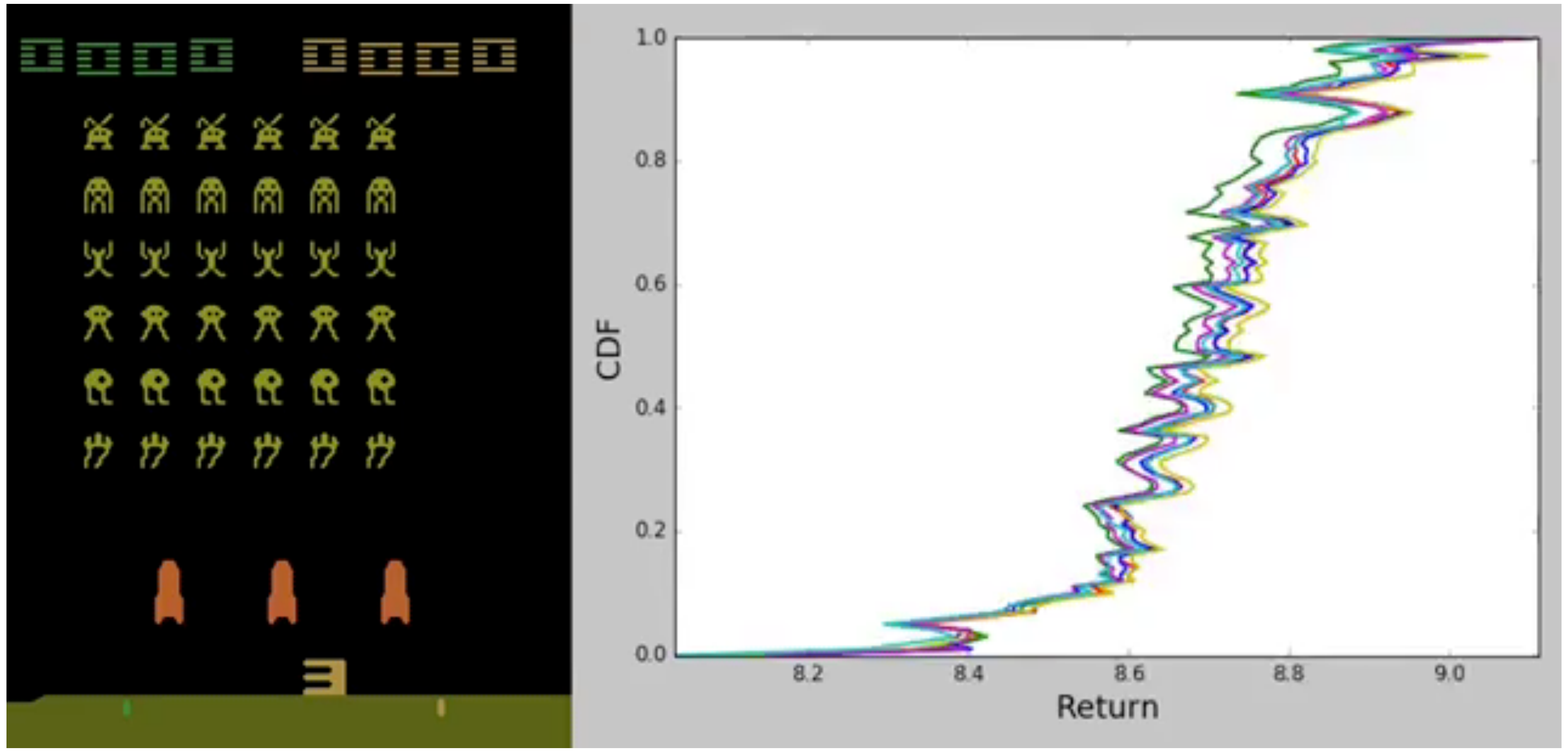
	Mean	Median	(human starts) Median	Human Gap
DQN	228 %	79 %	68 %	0.334
Prio. Duel.	592 %	172 %	128 %	0.178
C51	701 %	178 %	125 %	0.152
QR-DQN	**864 %**	**193 %**	**153 %**	0.165
IQN	**1019 %**	**218 %**	**162 %**	**0.141**
Rainbow (QR-DQN based)	**1189 %**	**230 %**	**125 %**	**0.144**

IQN outperforms Rainbow on **hardest** Atari gameas!

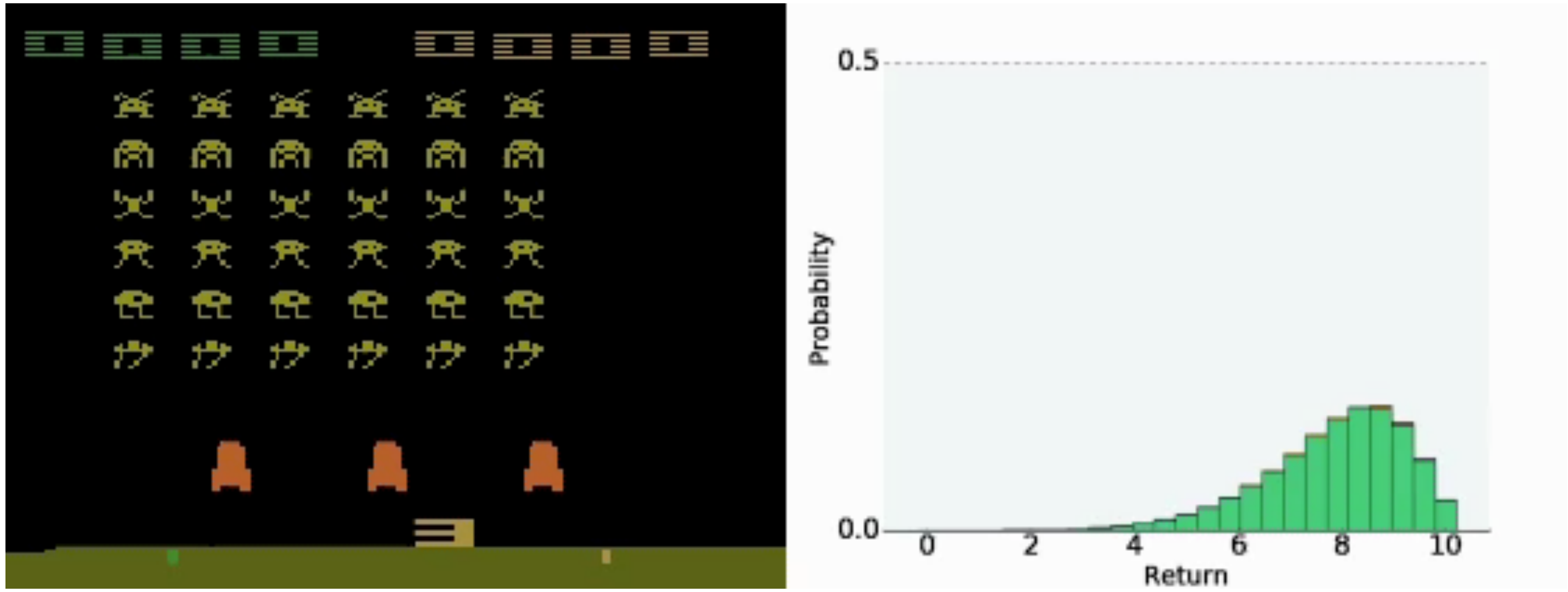


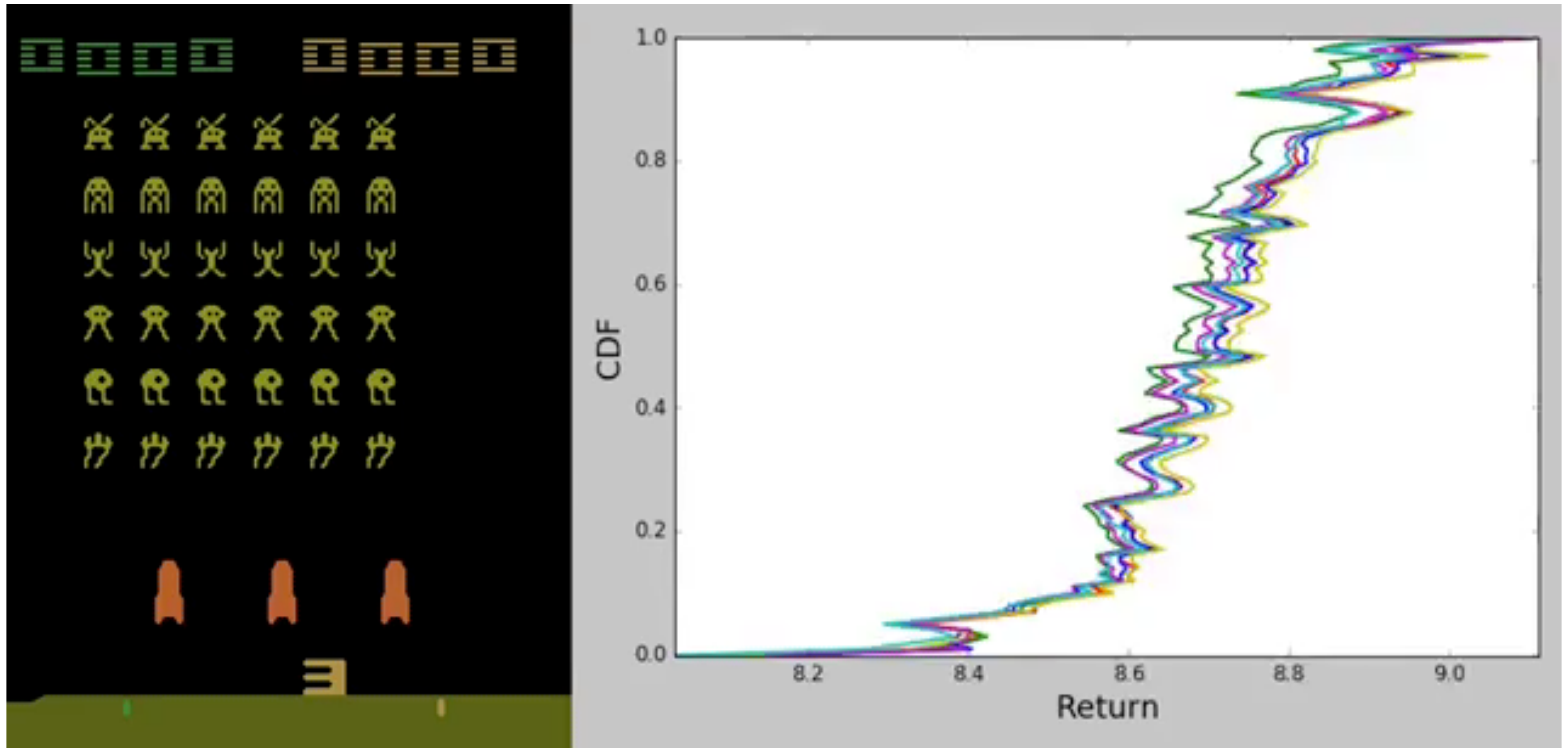
Demo


0.5
0.0
Probability
0
2
4
6
8
10
Return

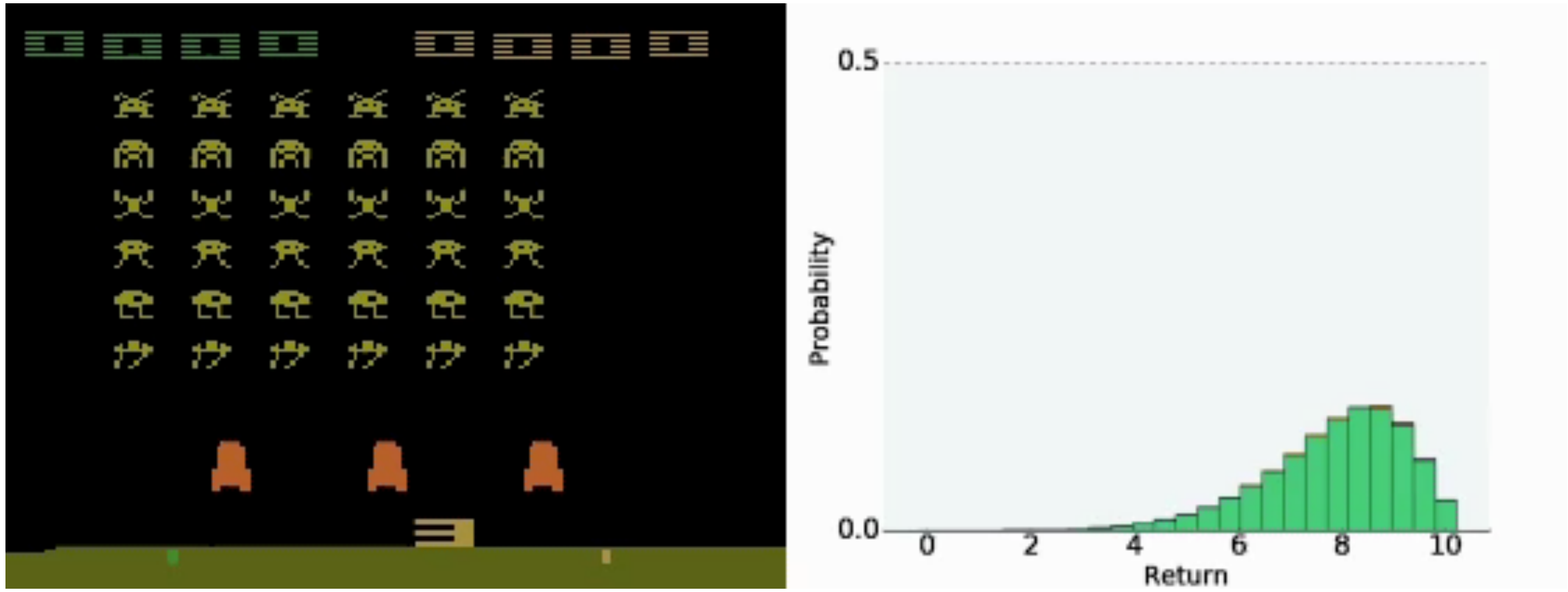

1.0
0.8
0.6
0.4
0.2
0.0
CDF
8.2
8.4
8.6
8.8
9.0
Return

Demo

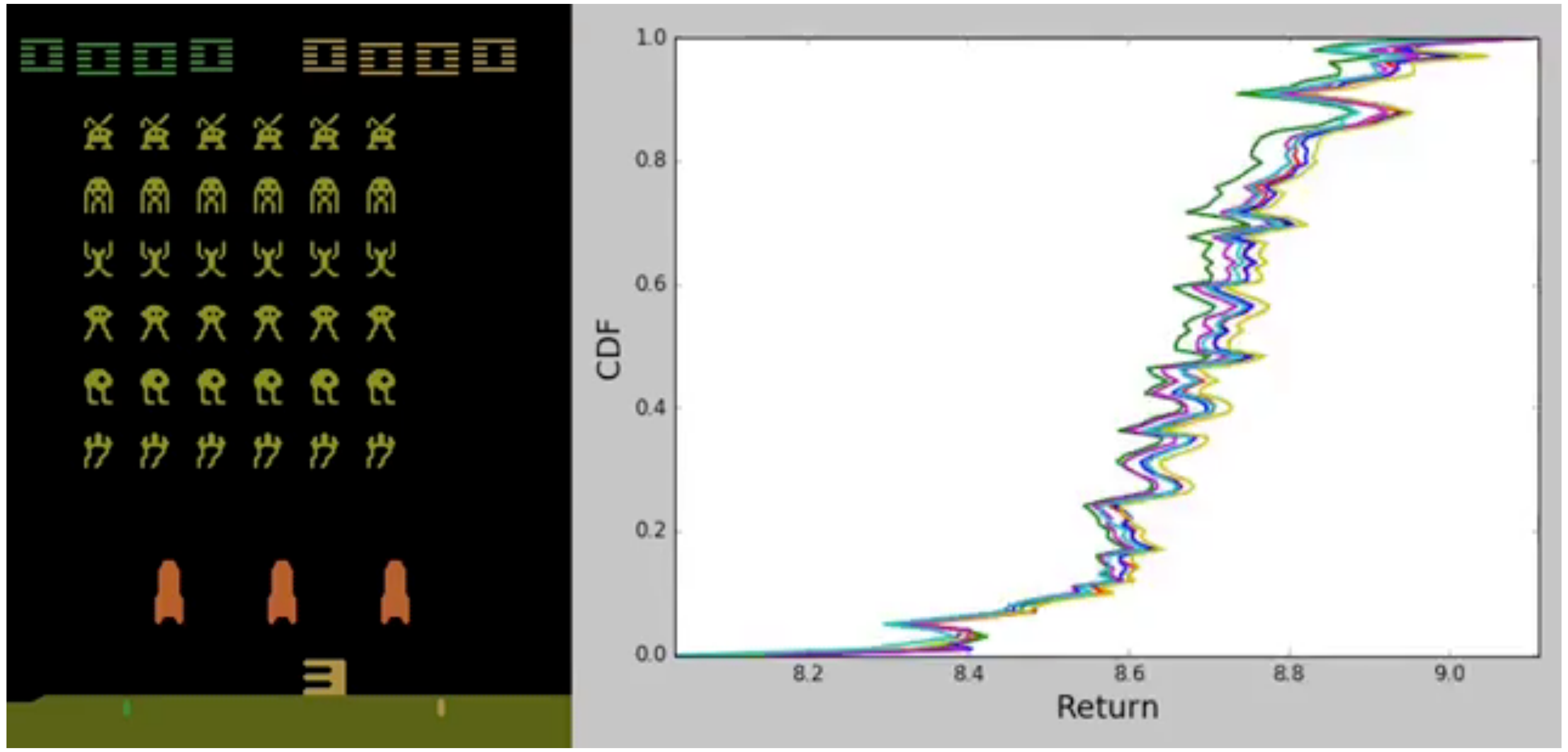

0.5
0.0
Probability
0
2
4
6
8
10
Return


1.0
0.8
0.6
0.4
0.2
0.0
CDF
8.2
8.4
8.6
8.8
9.0
Return

Demo



0.5
0.0
Probability
0
2
4
6
8
10
Return



1.0
0.8
0.6
0.4
0.2
0.0
CDF
8.2
8.4
8.6
8.8
9.0
Return

Pitfalls and future outlook

- Pitfalls
  - Only the policy **mean** used for evaluation
  - Convergence in theory only for **fixed** quantiles
  - Evaluated only in **discrete** action environments
  - No guarantee that quantiles are ordered
- Future
  - Distribution over policies
  - Continuous action domains
  - Rainbow-IQN
  - Solve the quantile ordering problem

CDF
Return
1.0
0.8
0.6
0.4
0.2
0.0
19
20
21
22
23
24
25
26
0.34
0.36
0.38
0.40
0.42
000 5 1
2.0
2.5
3.0
3.5
000000
SECTOR 01
15
2.1
2.2
2.3
2.4
2.5
2.6


CDF
Return
1.0
0.8
0.6
0.4
0.2
0.0
19
20
21
22
23
24
25
26
0.34
0.36
0.38
0.40
0.42
000 5 1
2.0
2.5
3.0
3.5
000000
SECTOR 01
15
2.1
2.2
2.3
2.4
2.5
2.6

References

[1] Sobel, L.M. 1982. "*The variance of discounted markov decision processes*"

[2] Morimura T. et. al. 2010. "*Parametric Return Density Estimation for Reinforcement Learning*"

[3] Bellemare, M.G. et. al. 2017. "*A Distributional Perspective on Reinforcement Learning*"

[4] Barth-Maron, G. et. al. 2018. "*Distributed Distributional Deterministic Policy Gradients*"

[5] Hessel, M. et. al. 2018. "*Rainbow: combining improvements in deep reinforcement learning*"

[6] Gruslys, A. et. al. 2018. "*The Reactor: a fast and sampleefficient actor-critic agent for reinforcement learning*"

[7] Rowland, M. et. al. 2018. "*An Analysis of Categorical Distributional Reinforcement Learning*"

[8] Dabney, W. et. al. 2017. "*Distributional Reinforcement Learning with Quantile Regression*"

[9] Dabney, W. et. al. 2018. "*Implicit Quantile Networks for Distributional Reinforcement Learning*"

IQN - Data efficiency vs. Computation

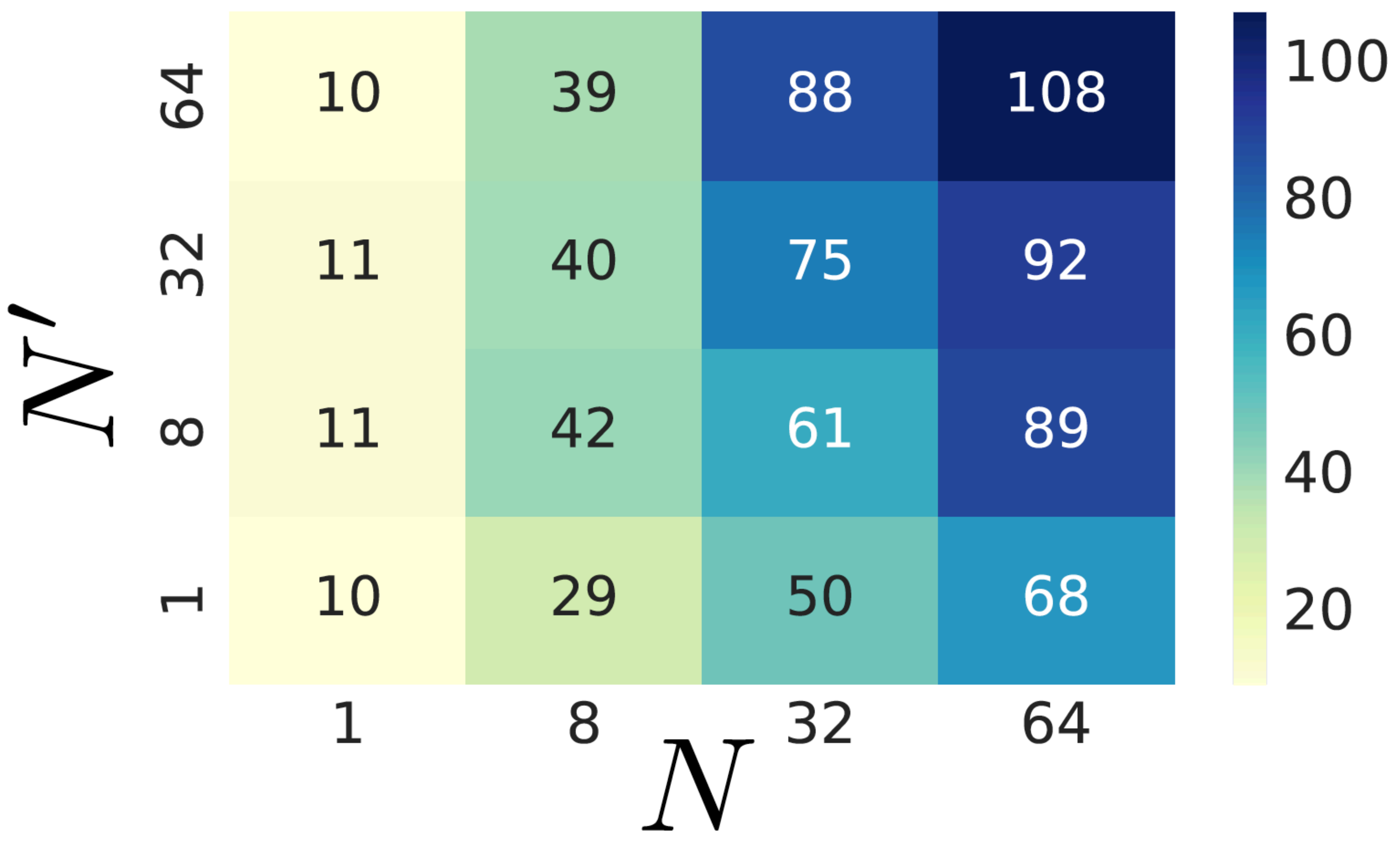
IQN - Data efficiency vs. Computation

$$\mathcal{L}_{IQN} = \sum_{\tau=\tau_1}^{\tau_N} \sum_{\tau'=\tau_1}^{\tau_{N'}} \delta_t^{\tau,\tau'} \left(\tau - \mathbf{I}_{\delta_t^{\tau,\tau'}<0}\right)$$

IQN - Data efficiency vs. Computation

$$\mathscr{L}_{IQN} = \sum_{\tau=\tau_1}^{\tau_N} \sum_{\tau'=\tau_1}^{\tau_{N'}} \delta_t^{\tau,\tau'} \left(\tau - \mathbf{I}_{\delta_t^{\tau,\tau'}<0}\right)$$

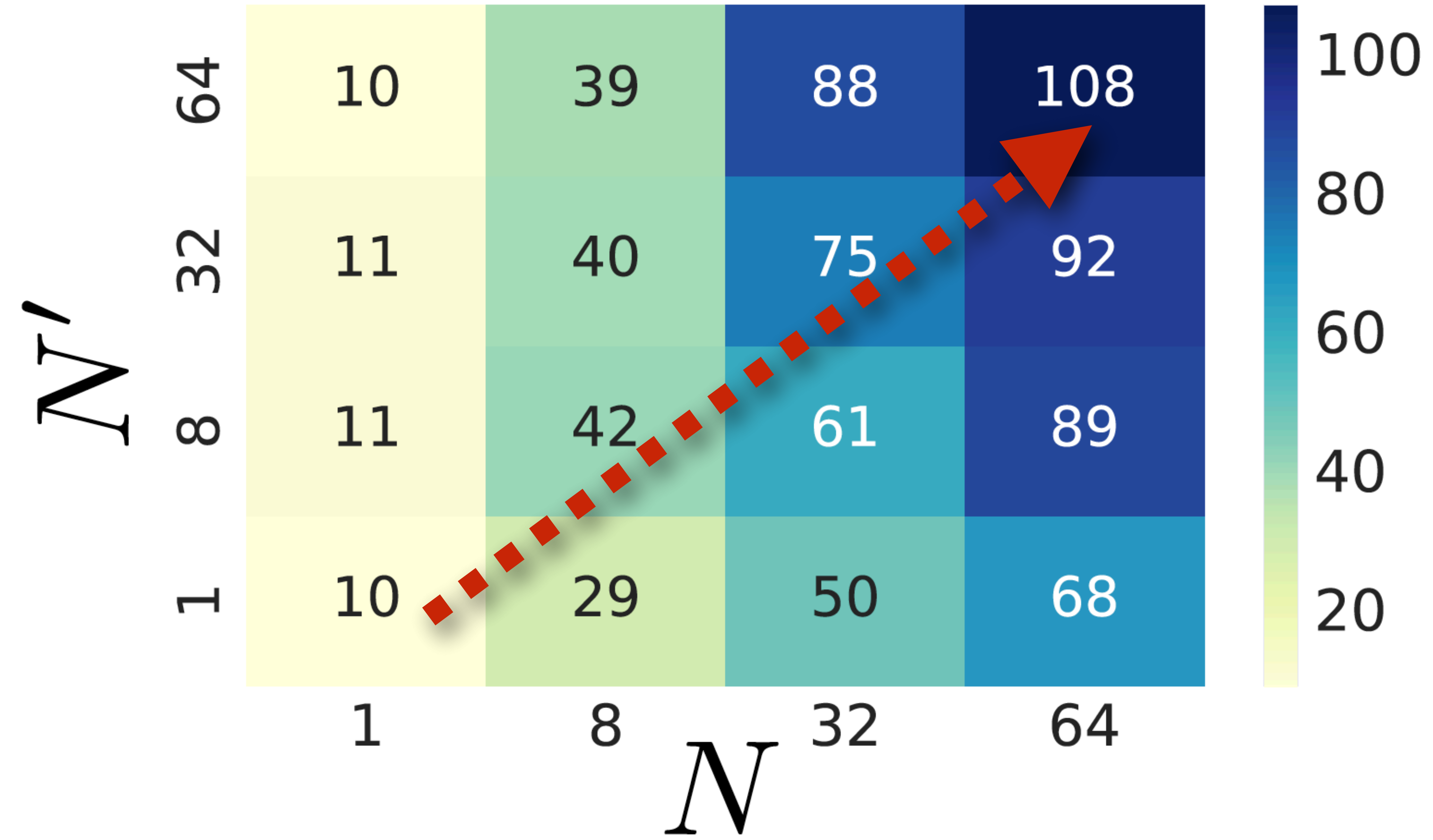
HNS on **first** 10 million frames



IQN - Data efficiency vs. Computation

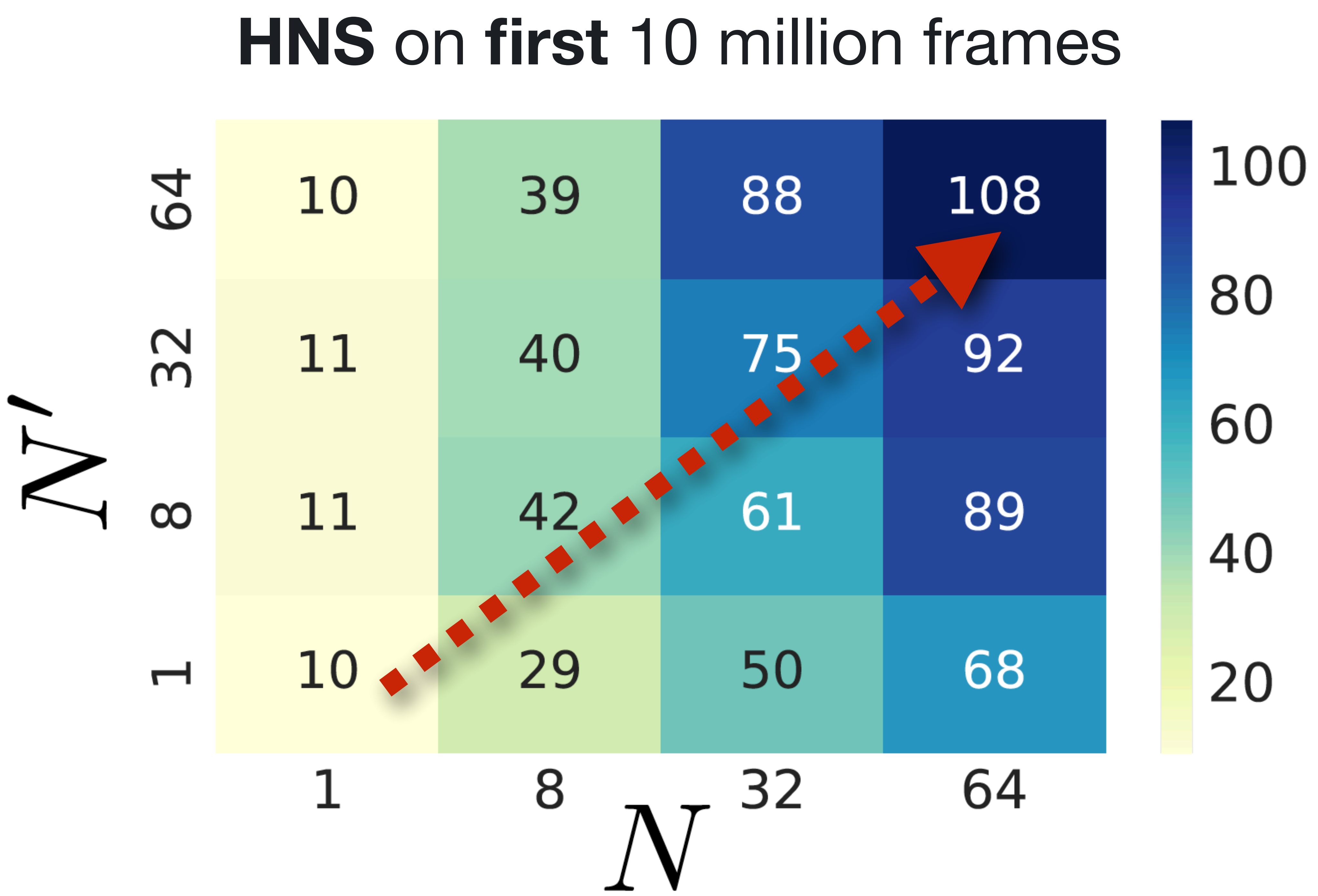
$$\mathscr{L}_{IQN} = \sum_{\tau=\tau_1}^{\tau_N} \sum_{\tau'=\tau_1}^{\tau_{N'}} \delta_t^{\tau,\tau'} \left(\tau - \mathbf{I}_{\delta_t^{\tau,\tau'}<0}\right)$$

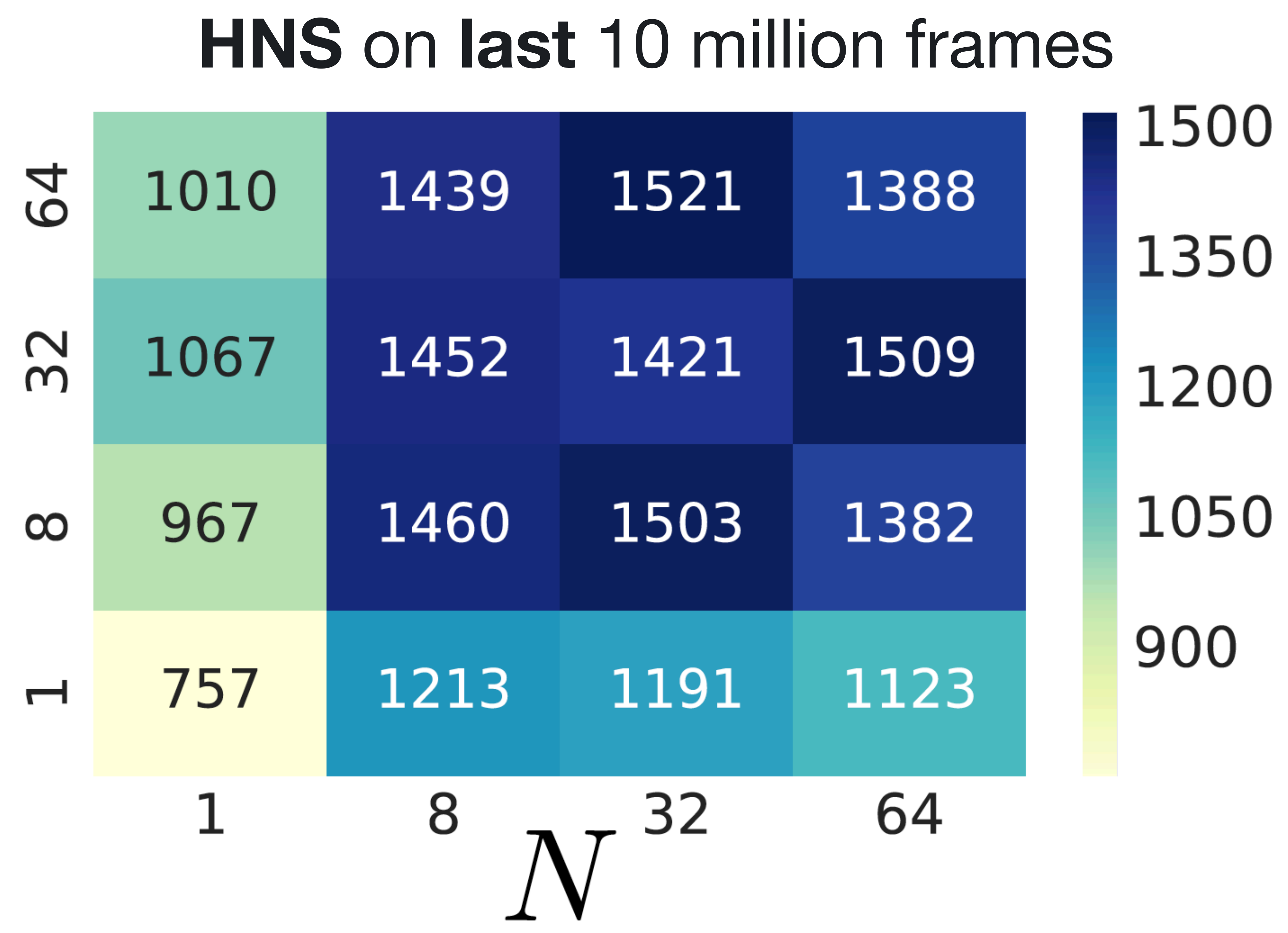
HNS on **first** 10 million frames



IQN - Data efficiency vs. Computation

$$\mathcal{L}_{IQN} = \sum_{\tau=\tau_1}^{\tau_{\color{red}N}} \sum_{\tau'=\tau_1}^{\tau_{\color{red}N'}} \delta_t^{\tau,\tau'} \left(\tau - \mathbf{I}_{\delta_t^{\tau,\tau'}<0}\right)$$

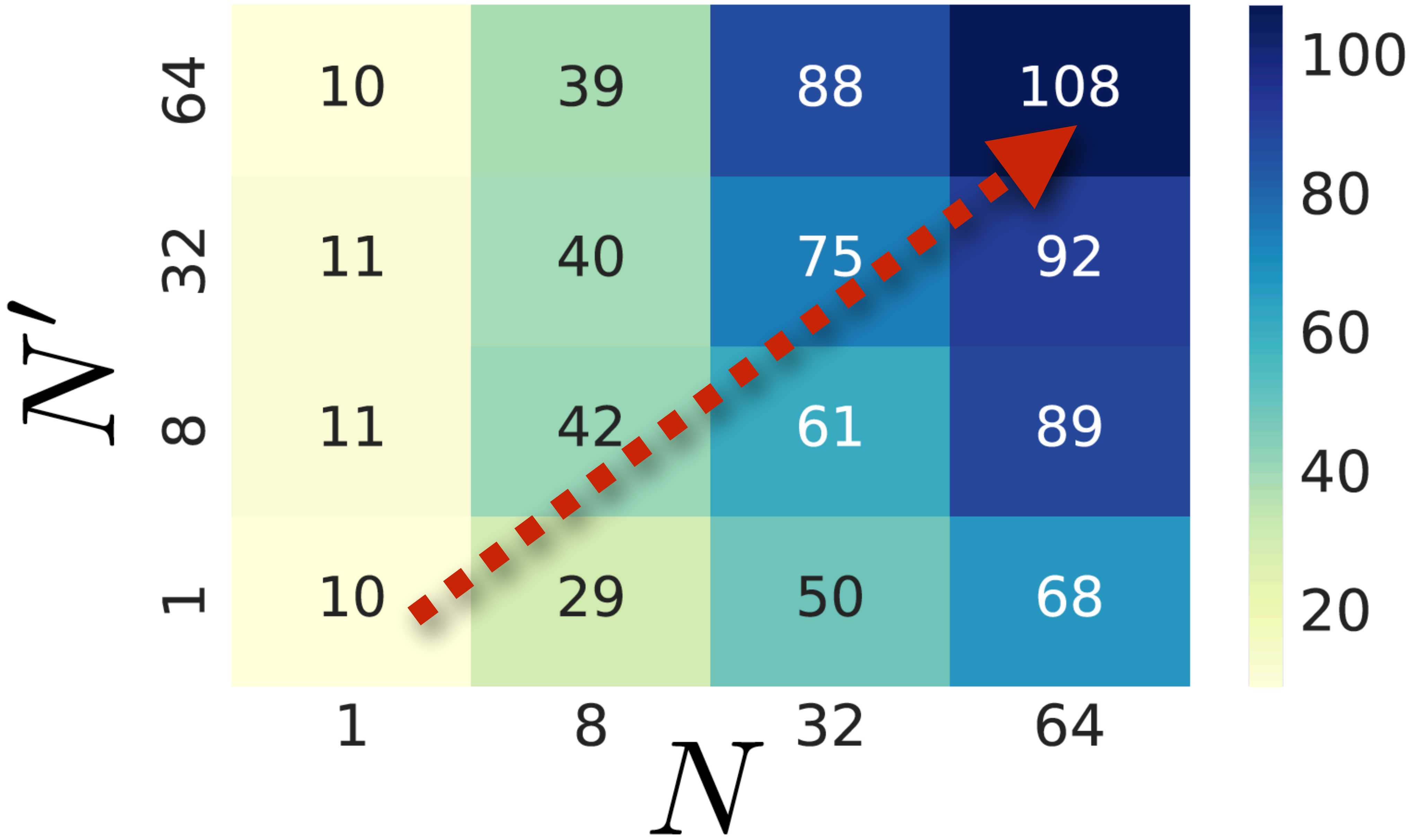




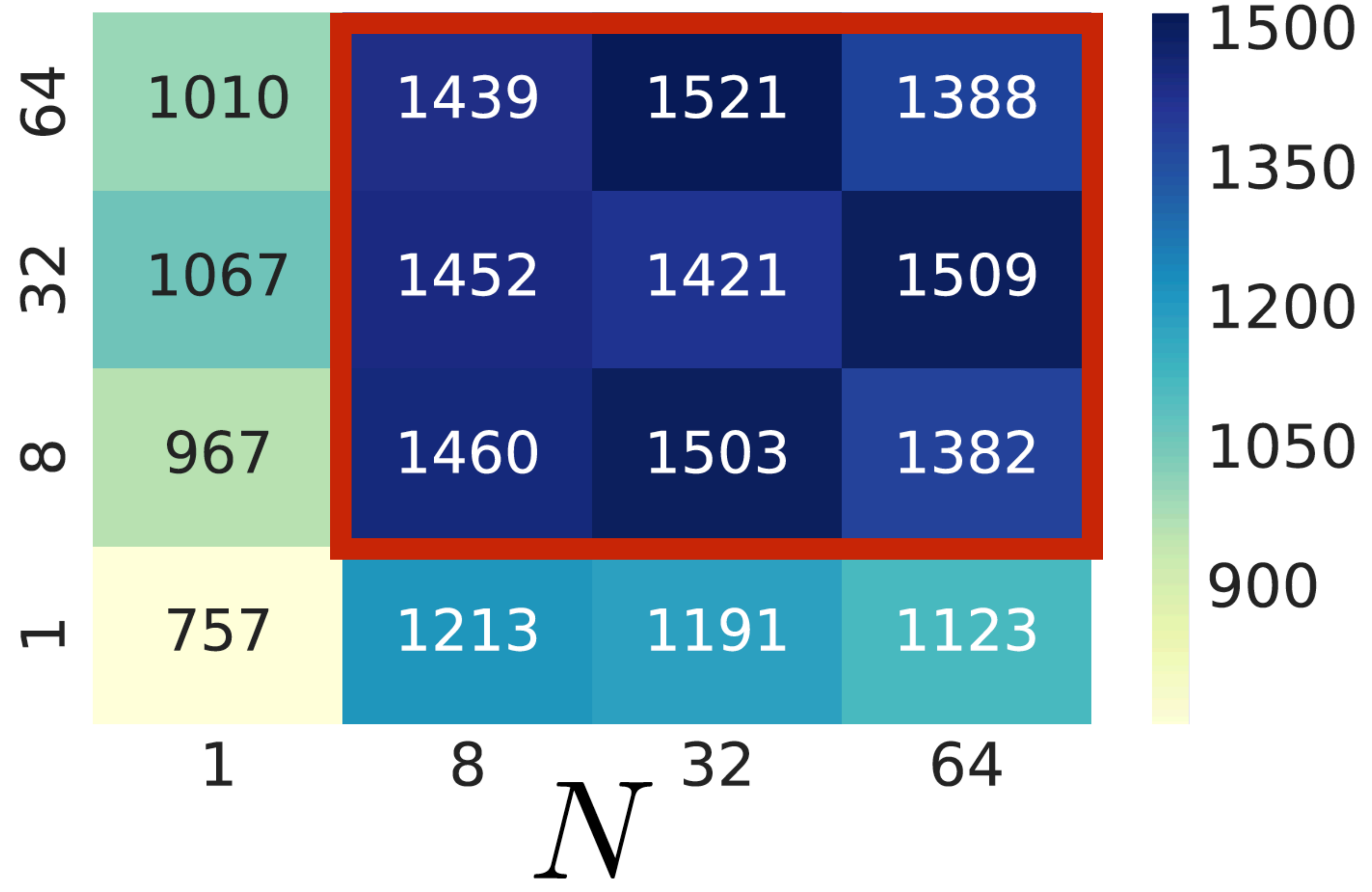
IQN - Data efficiency vs. Computation

$$\mathscr{L}_{IQN} = \sum_{\tau=\tau_1}^{\tau_N} \sum_{\tau'=\tau_1}^{\tau_{N'}} \delta_t^{\tau,\tau'} \left(\tau - \mathbf{I}_{\delta_t^{\tau,\tau'}<0}\right)$$

HNS on **first** 10 million frames



HNS on **last** 10 million frames

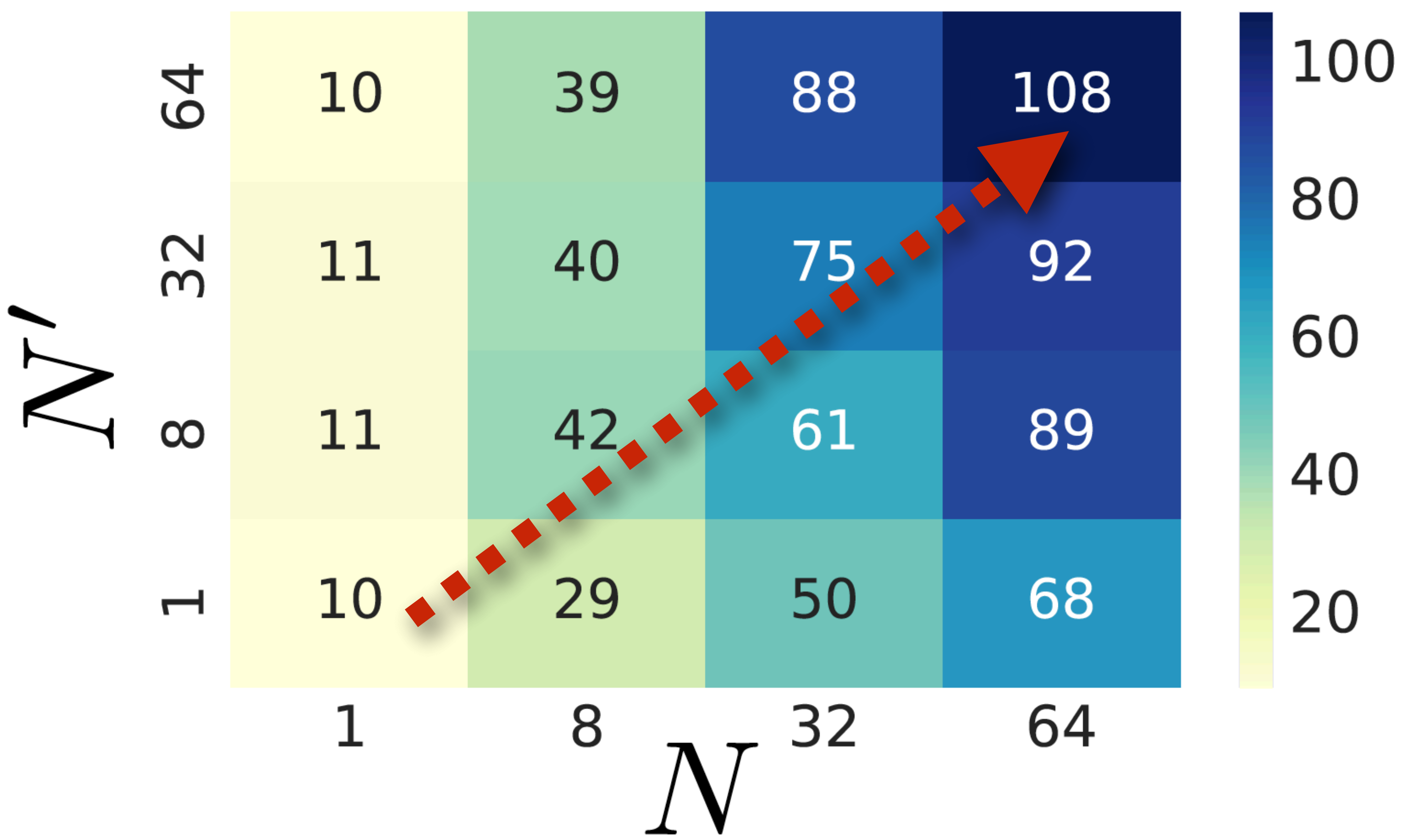


IQN - Data efficiency vs. Computation

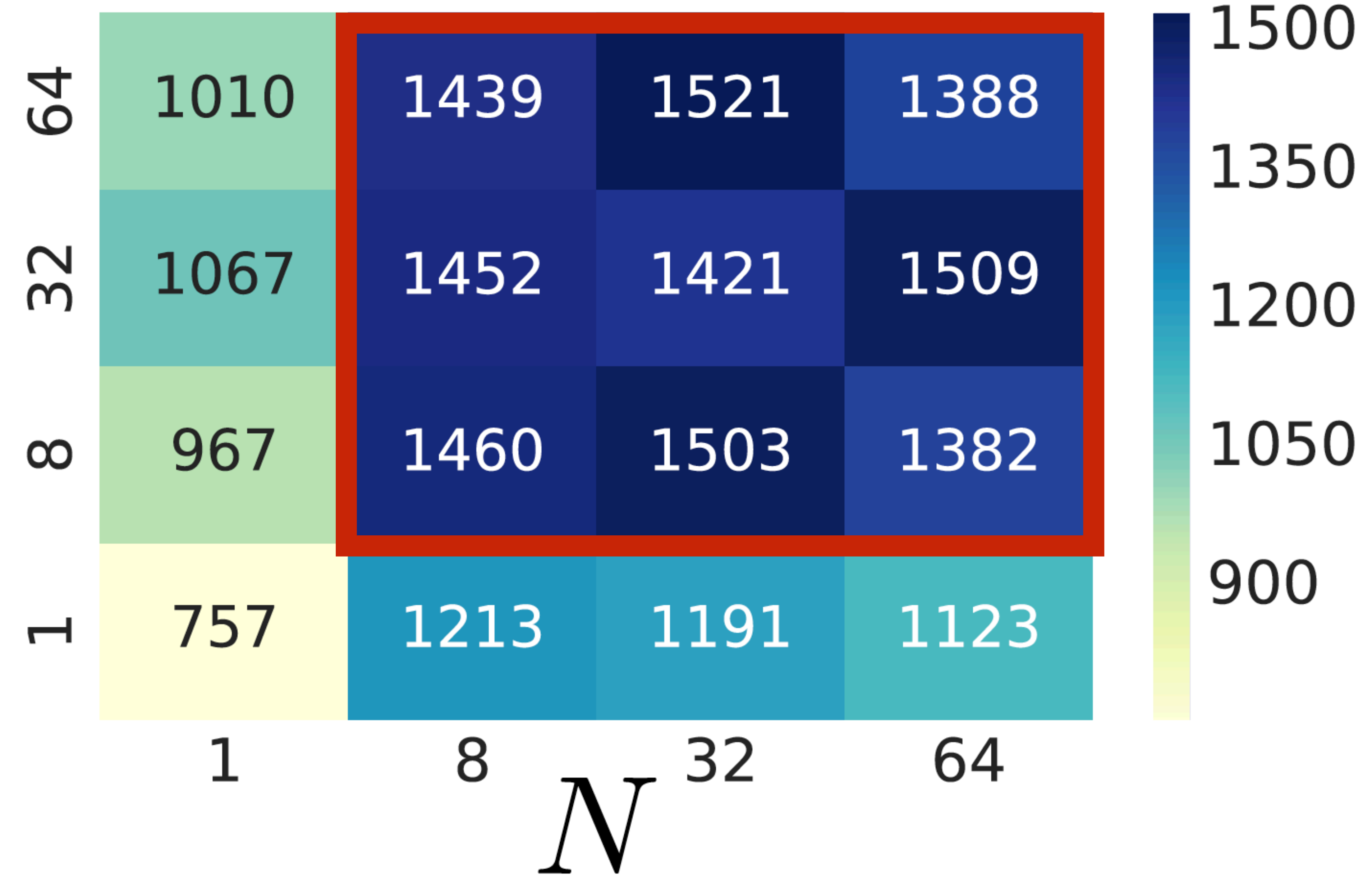
$$\mathcal{L}_{IQN} = \sum_{\tau=\tau_1}^{\tau_N} \sum_{\tau'=\tau_1}^{\tau_{N'}} \delta_t^{\tau,\tau'} \left(\tau - \mathbf{I}_{\delta_t^{\tau,\tau'}<0}\right)$$

- Total frames 200 million
- Averaged over 6 Atari games
- DQN **(32, 253)**
- QR-DQN (144, 1243)

HNS on **first** 10 million frames



HNS on **last** 10 million frames

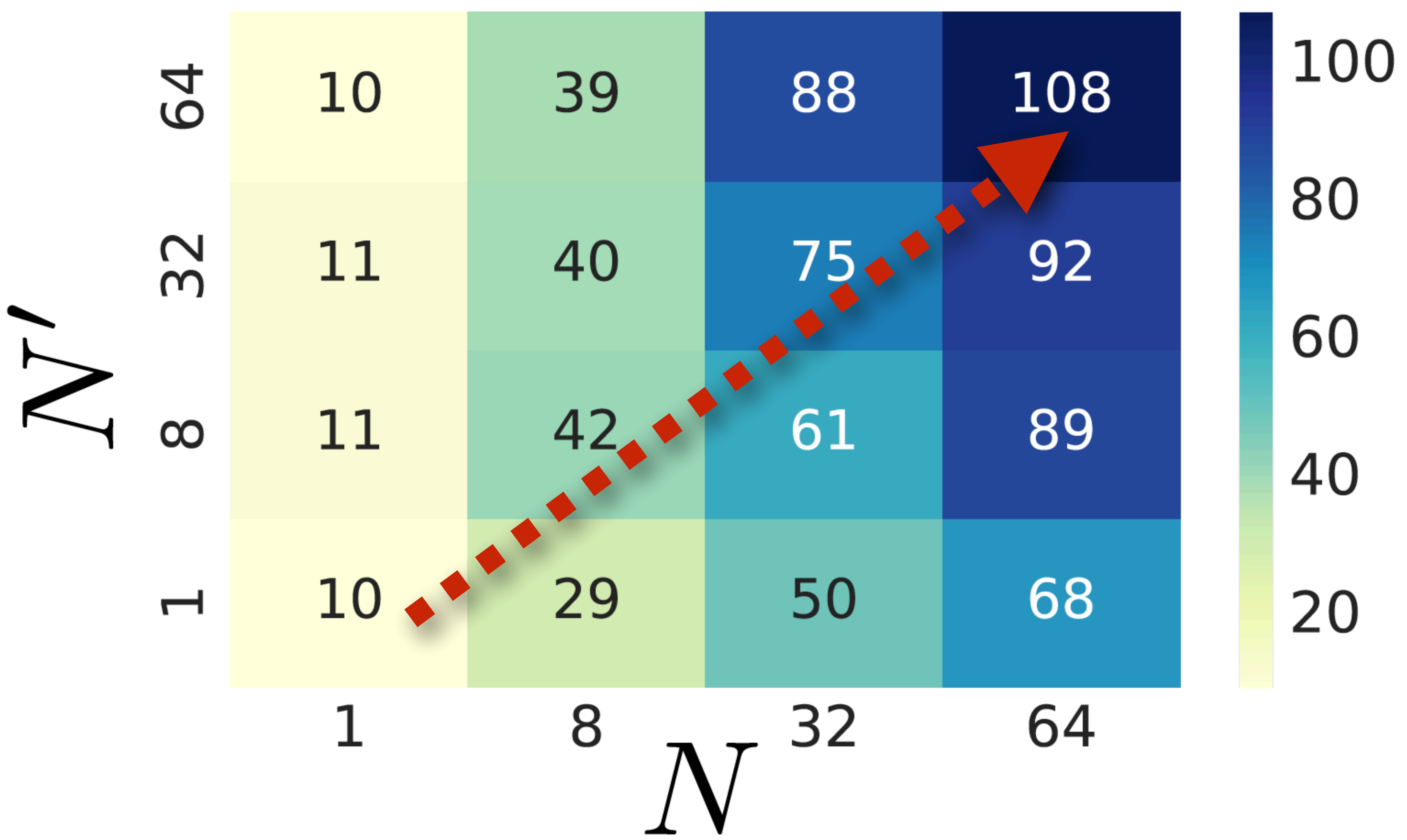


IQN - Data efficiency vs. Computation

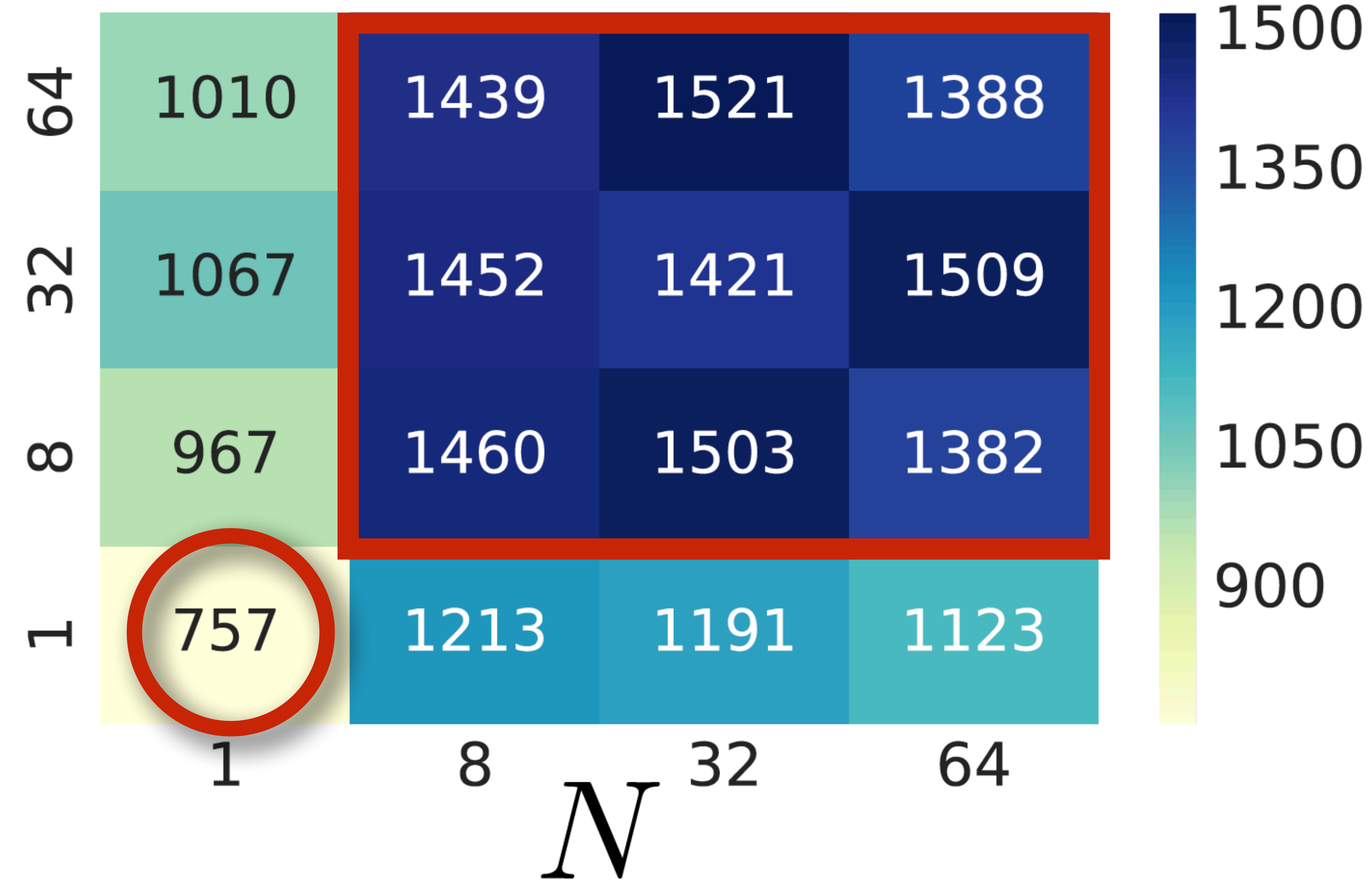
$$\mathscr{L}_{IQN} = \sum_{\tau=\tau_1}^{\tau_N} \sum_{\tau'=\tau_1}^{\tau_{N'}} \delta_t^{\tau,\tau'} \left(\tau - \mathbf{I}_{\delta_t^{\tau,\tau'}<0}\right)$$

- Total frames 200 million
- Averaged over 6 Atari games
- DQN **(32, 253)**
- QR-DQN (144, 1243)

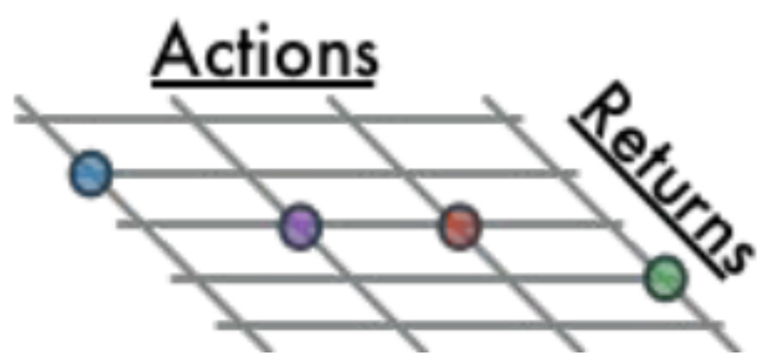
HNS on **first** 10 million frames



HNS on **last** 10 million frames



DQN



$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}} \; Q^{\theta}(s_{t+1}, a)$$

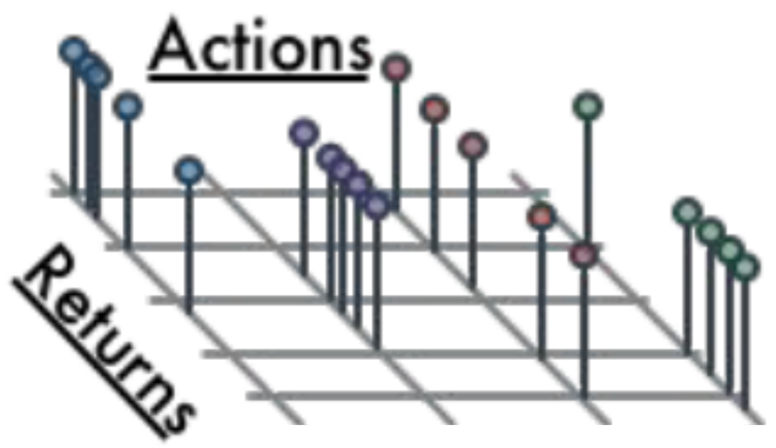
$$q' = Q^{\theta}(s_{t+1}, a^\star)$$

$$q = Q^{\theta}(s_t, a_t)$$

$$\delta_t = r_t + \gamma q' - q$$

$$\mathscr{L}_{DQN} = \delta_t^2$$

QR-DQN

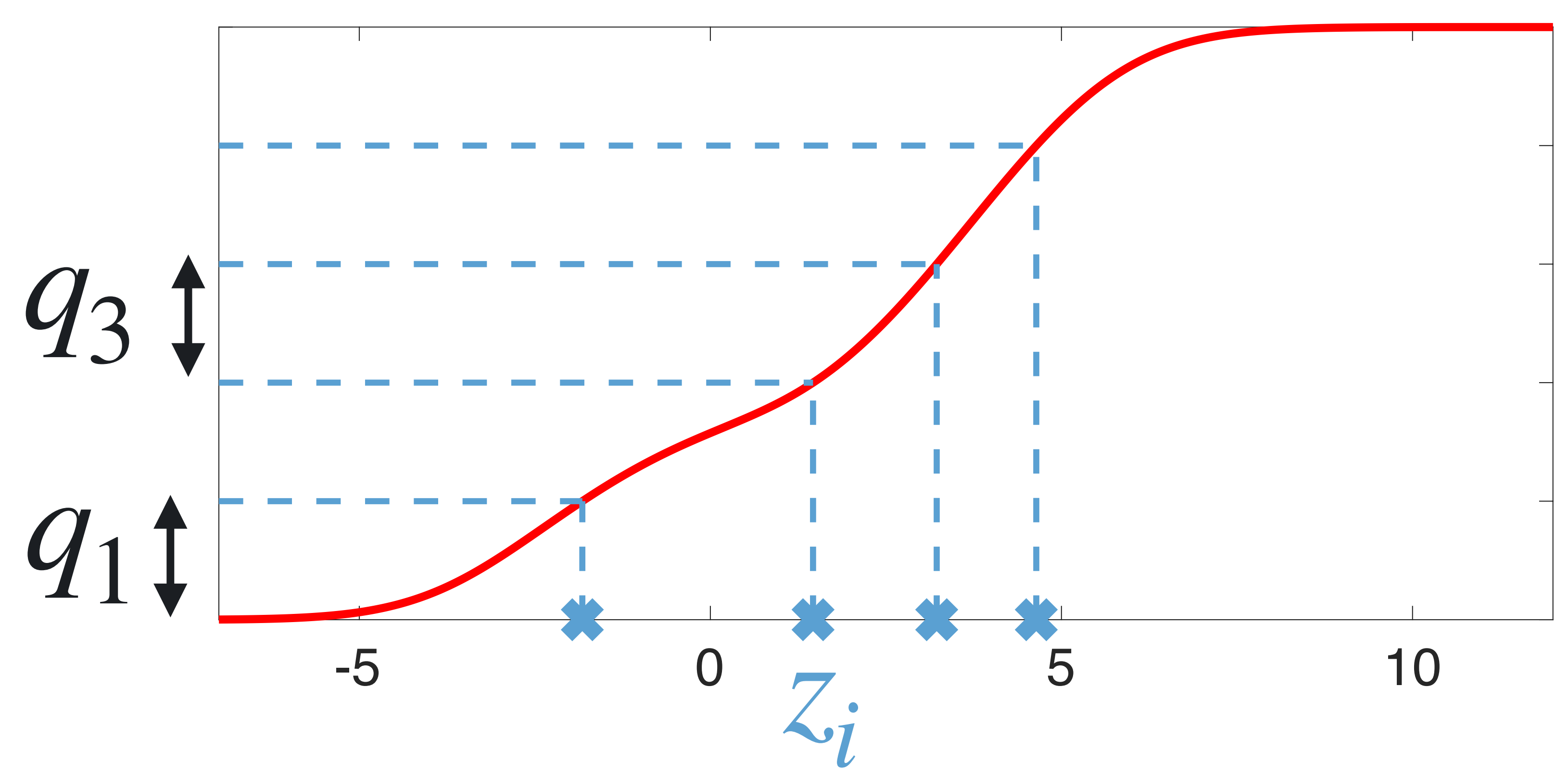


$$\{s_t, a_t, s_{t+1}, r_t\}$$

$$a^\star = \underset{a}{\text{argmax}} \; \mathbb{E}\left[Z_\tau^{\theta}(s_{t+1}, a)\right]$$

$$\forall \tau, \tau' \;\Bigg|\; \begin{aligned} z' &= Z_{\tau'}^{\theta}(s_{t+1}, a^\star) \\ z &= Z_{\tau}^{\theta}(s_t, a_t) \\ \delta_t^{\tau,\tau'} &= r_t + \gamma z' - z \end{aligned}$$

$$\mathscr{L}_{QR-DQN} = ?$$



$$Q(x', a') = \sum_j q_j \, z_j^{\theta}(x', a'), \quad \forall a'$$

$$a^\star = \underset{a'}{\text{argmax}} \; Q(x', a')$$