Statistical Machine Learning (BE4M33SSU)
Lecture 2: Predictor evaluation and learning via using empirical risk

Czech Technical University in Prague
V. Franc

Definition of the prediction problem

- $\mathcal{X}$ is a set of input observations/features
- $\mathcal{Y}$ is a set of hidden states/labels
- $(x, y) \in \mathcal{X} \times \mathcal{Y}$ samples randomly drawn from r.v. with p.d.f. $p(x, y)$
- $h\colon \mathcal{X} \to \mathcal{Y}$ is a prediction strategy/hypothesis
- $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss function
- Task: find a strategy with the minimal true risk (expected loss)

$$R(h) = \int \sum_{y \in \mathcal{Y}} \ell(y, h(x))\, p(x, y)\, \mathrm{d}x = \mathbb{E}_{(x,y)\sim p}\Big(\ell(y, h(x))\Big)$$

- Optimal solution: Bayes predictor h^* attaining the minimal risk

$$R(h^*) = \inf_{h \in \mathcal{Y}^{\mathcal{X}}} R(h)$$

Example of a prediction problem

- The statistical model is known:
  - $\mathcal{X} = \mathbb{R}$, $\quad \mathcal{Y} = \{+1, -1\}$, $\quad \ell(y, y') = \begin{cases} 0 & \text{if} \quad y = y' \\ 1 & \text{if} \quad y \neq y' \end{cases}$
  - $p(x, y) = p(y)\, \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2}$, $\quad y \in \mathcal{Y}$.

Example of a prediction problem

- The statistical model is known:
  - $\mathcal{X} = \mathbb{R}$, $\quad \mathcal{Y} = \{+1, -1\}$, $\quad \ell(y, y') = \begin{cases} 0 & \text{if} \quad y = y' \\ 1 & \text{if} \quad y \neq y' \end{cases}$
  - $p(x, y) = p(y) \, \frac{1}{\sqrt{2\pi}\sigma} \, e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2}, \quad y \in \mathcal{Y}.$

- The optimal strategy (assuming $\mu_- < \mu_+$):

$$h(x) = \operatorname{argmax}_{y \in \mathcal{Y}} p(y \mid x) = \operatorname{sign}(x - \theta)$$

Example of a prediction problem

- The statistical model is known:
  - $\mathcal{X} = \mathbb{R}, \quad \mathcal{Y} = \{+1, -1\}, \quad \ell(y, y') = \begin{cases} 0 & \text{if} \quad y = y' \\ 1 & \text{if} \quad y \neq y' \end{cases}$
  - $p(x, y) = p(y) \, \frac{1}{\sqrt{2\pi}\sigma} \, e^{-\frac{1}{2\sigma^2}(x - \mu_y)^2}, \quad y \in \mathcal{Y}.$

- The optimal strategy (assuming $\mu_- < \mu_+$):

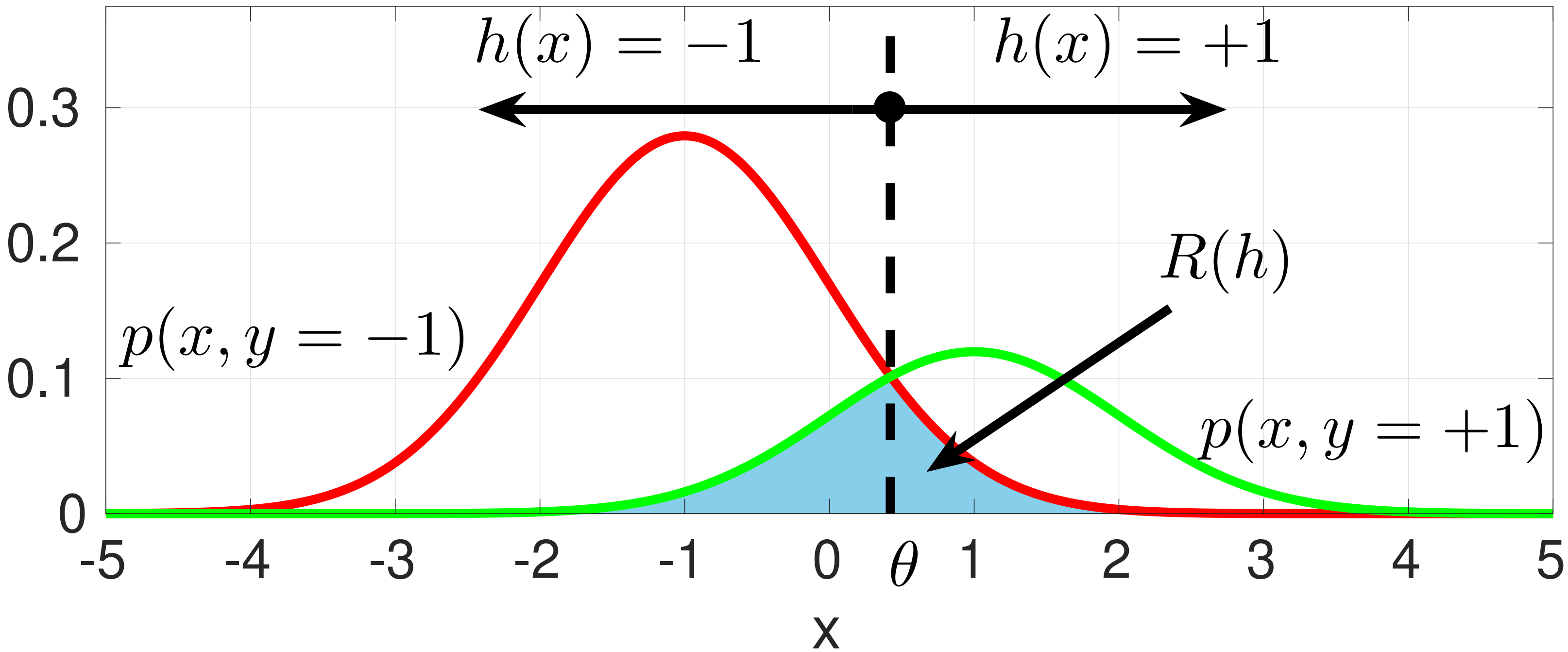
$$h(x) = \operatorname{argmax}_{y \in \mathcal{Y}} p(y \mid x) = \operatorname{sign}(x - \theta)$$

- The value of the true risk:

$$R(h) = \int_{-\infty}^{\theta} p(x, +1) \mathrm{d}x + \int_{\theta}^{\infty} p(x, -1) \mathrm{d}x$$

Example of a prediction problem

- The statistical model is known:
  - $\mathcal{X} = \mathbb{R}, \quad \mathcal{Y} = \{+1, -1\}, \quad \ell(y, y') = \begin{cases} 0 & \text{if} \quad y = y' \\ 1 & \text{if} \quad y \neq y' \end{cases}$
  - $p(x, y) = p(y) \, \frac{1}{\sqrt{2\pi}\sigma} \, e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2}, \quad y \in \mathcal{Y}.$



Predictor evaluation and learning based on examples

- Assumption: The true risk $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x)))$ is unknow due to unknown $p(x, y)$, however, we assume to have examples

$$(x^1, y^1), (x^2, y^2), \ldots, (x^n, y^n)$$

drawn from i.i.d. r.v. distributed according to $p(x, y)$.

Predictor evaluation and learning based on examples

- Assumption: The true risk $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x)))$ is unknow due to unknown $p(x,y)$, however, we assume to have examples

$$(x^1, y^1), (x^2, y^2), \dots, (x^n, y^n)$$

drawn from i.i.d. r.v. distributed according to $p(x,y)$.

- We will analyze two problems:

1. Evaluation: given $h\colon \mathcal{X} \to \mathcal{Y}$, estimate its $R(h)$ using test set

$$\mathcal{S}^l = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, l\} \quad \text{drawn i.i.d. from} \quad p(x,y)$$

2. Learning: find $h\colon \mathcal{X} \to \mathcal{Y}$ with small $R(h)$ using training set

$$\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, m\} \quad \text{drawn i.i.d. from} \quad p(x,y)$$

Predictor evaluation via empirical risk

- Given a predictor $h\colon \mathcal{X} \to \mathcal{Y}$ and a test set $\mathcal{S}^l$ draw i.i.d. from $p(x,y)$, compute the empirical risk

$$R_{\mathcal{S}^l}(h) = \frac{1}{l}\big(\ell(y^1, h(x^1)) + \cdots + \ell(y^l, h(x^l))\big) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(x^i))$$

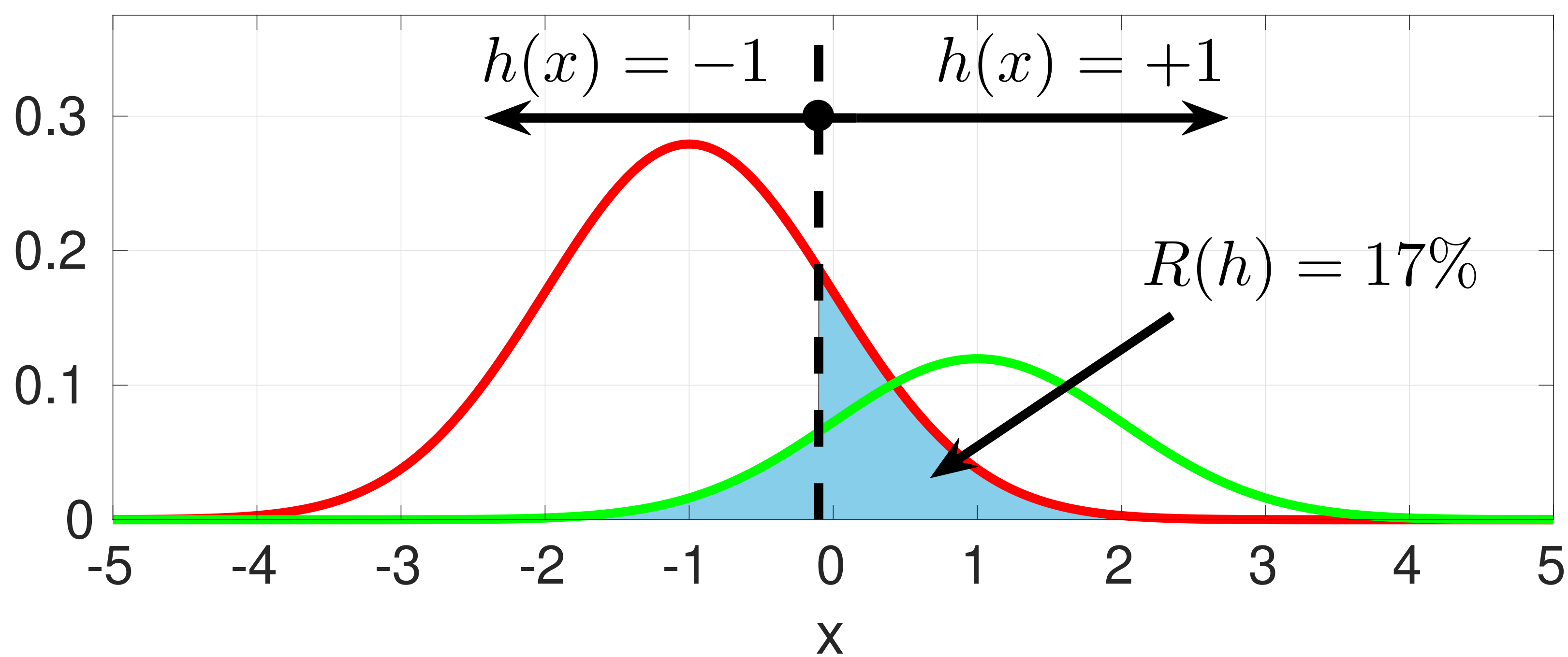
and use it as an estimate of the true risk $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x)))$.

Predictor evaluation via empirical risk

- Given a predictor $h\colon \mathcal{X} \to \mathcal{Y}$ and a test set $\mathcal{S}^l$ draw i.i.d. from $p(x,y)$, compute the empirical risk

$$R_{\mathcal{S}^l}(h) = \frac{1}{l}\big(\ell(y^1, h(x^1)) + \cdots + \ell(y^l, h(x^l))\big) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(x^i))$$

and use it as an estimate of the true risk $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x)))$.

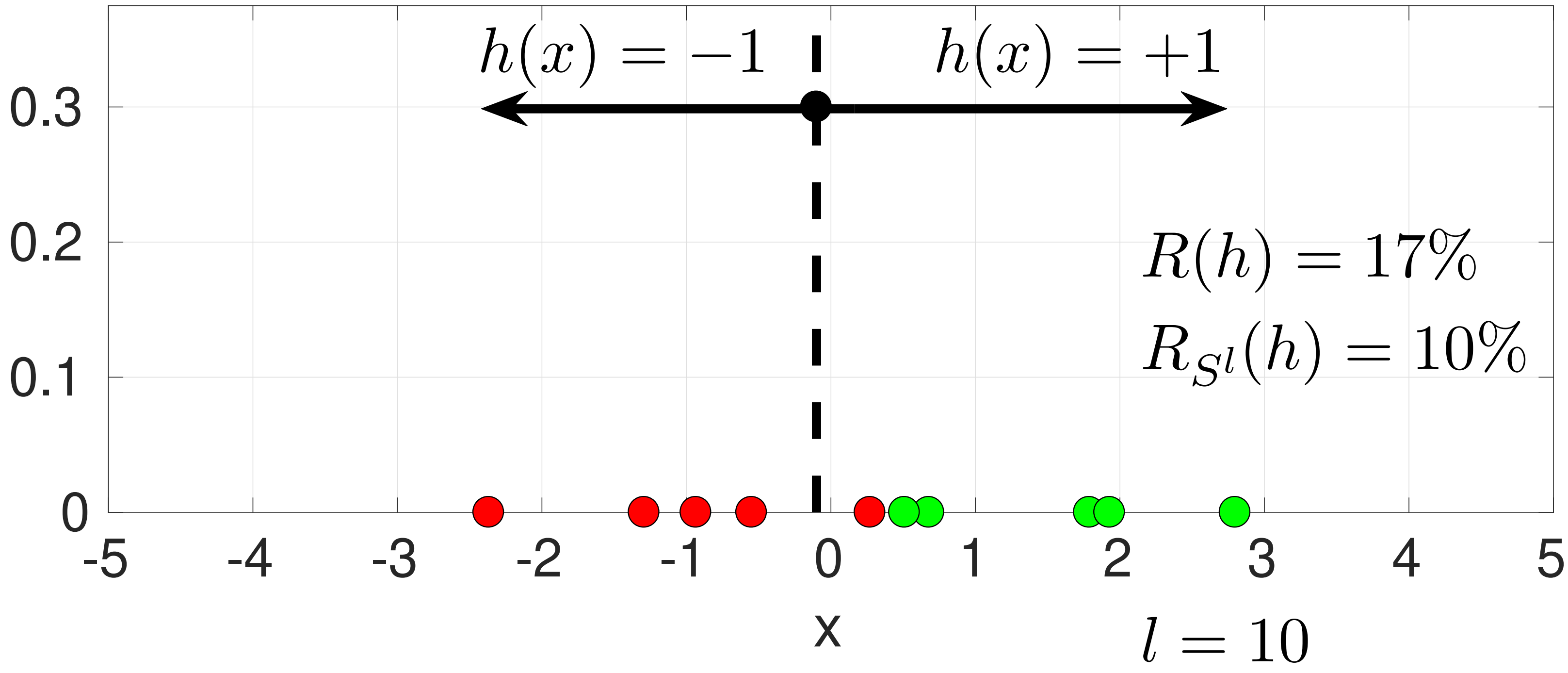


Predictor evaluation via empirical risk

- Given a predictor $h\colon \mathcal{X} \to \mathcal{Y}$ and a test set $\mathcal{S}^l$ draw i.i.d. from $p(x,y)$, compute the empirical risk

$$R_{\mathcal{S}^l}(h) = \frac{1}{l}\big(\ell(y^1, h(x^1)) + \cdots + \ell(y^l, h(x^l))\big) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(x^i))$$

and use it as an estimate of the true risk $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x)))$.

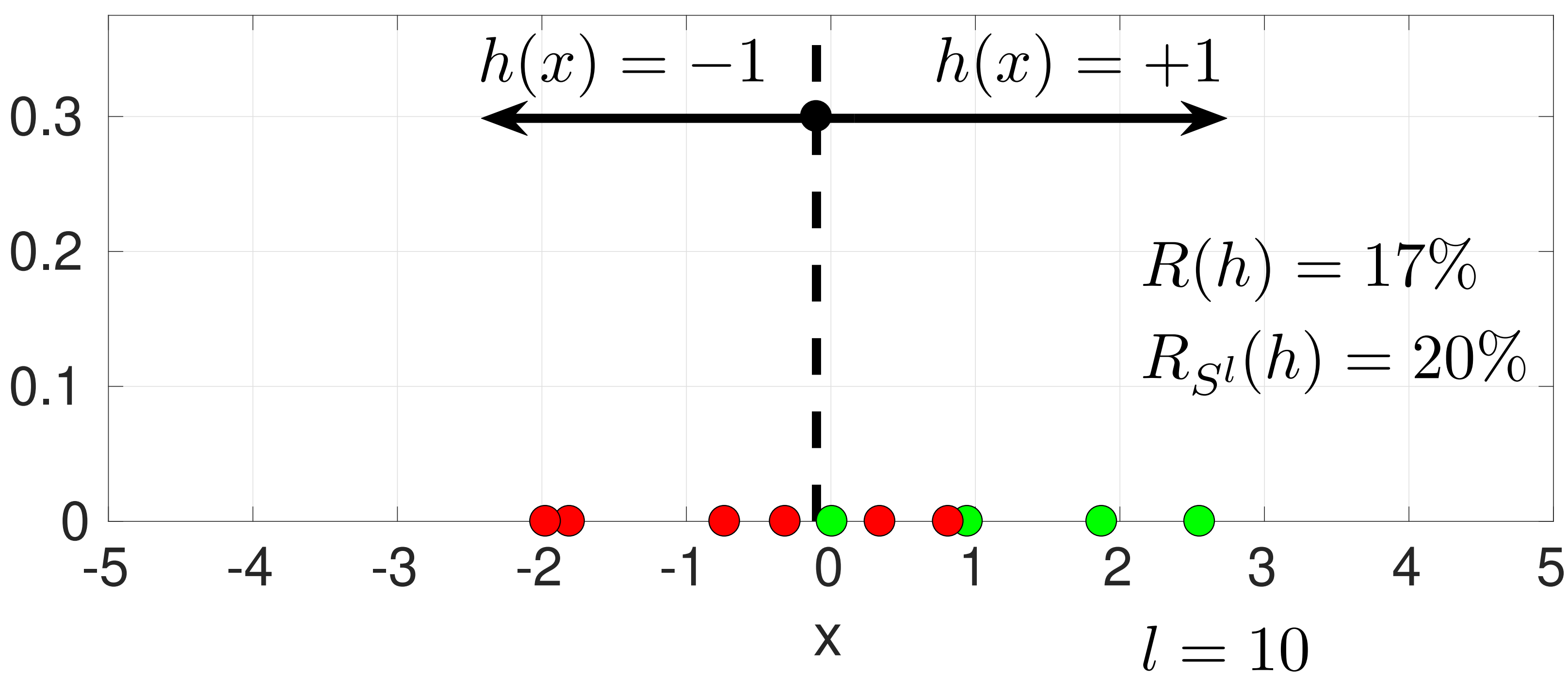


Predictor evaluation via empirical risk

- Given a predictor $h\colon \mathcal{X} \to \mathcal{Y}$ and a test set $\mathcal{S}^l$ draw i.i.d. from $p(x,y)$, compute the empirical risk

$$R_{\mathcal{S}^l}(h) = \frac{1}{l}\big(\ell(y^1, h(x^1)) + \cdots + \ell(y^l, h(x^l))\big) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(x^i))$$

and use it as an estimate of the true risk $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x)))$.

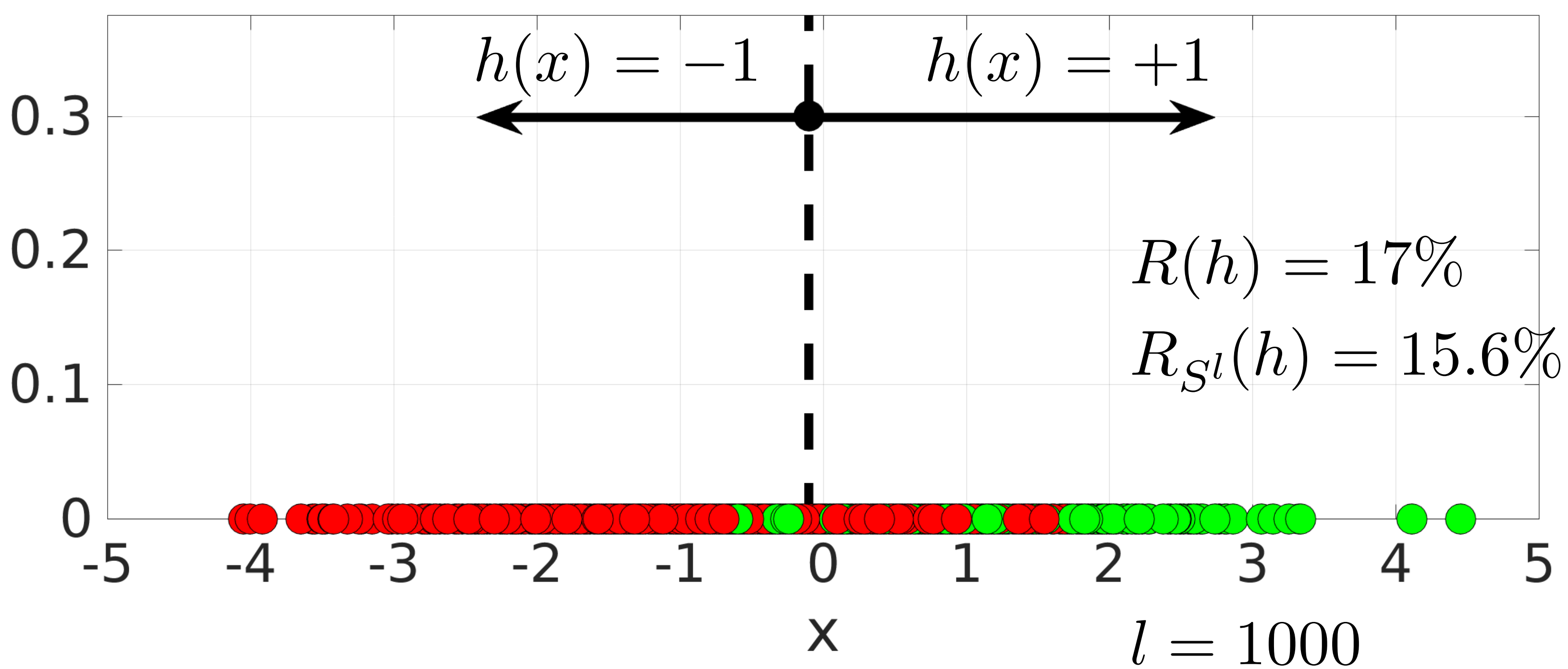


Predictor evaluation via empirical risk

- Given a predictor $h\colon \mathcal{X} \to \mathcal{Y}$ and a test set $\mathcal{S}^l$ draw i.i.d. from $p(x,y)$, compute the empirical risk

$$R_{\mathcal{S}^l}(h) = \frac{1}{l}\big(\ell(y^1, h(x^1)) + \cdots + \ell(y^l, h(x^l))\big) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(x^i))$$

and use it as an estimate of the true risk $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x)))$.



Predictor evaluation via empirical risk

- Given a predictor $h\colon \mathcal{X} \to \mathcal{Y}$ and a test set $\mathcal{S}^l$ draw i.i.d. from $p(x, y)$, compute the empirical risk

$$R_{\mathcal{S}^l}(h) = \frac{1}{l}\big(\ell(y^1, h(x^1)) + \cdots + \ell(y^l, h(x^l))\big) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(x^i))$$

and use it as an estimate of the true risk $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x)))$.

- $R_{\mathcal{S}^l}(h)$ is a random number with an unknown distribution.

- We will construct a confidence interval such that

$$R(h) \in (R_{\mathcal{S}^l(h)} - \varepsilon, R_{\mathcal{S}^l(h)} + \varepsilon) \;\text{ with probability (confidence) } \gamma \in (0, 1)$$

where ε is a deviation.

Law of large numbers

- Sample mean (arithmetic average) of the results of random trials gets closer to the expected value as more trials are performed.

Law of large numbers

- Sample mean (arithmetic average) of the results of random trials gets closer to the expected value as more trials are performed.

- Example: The expected value of a single roll of a fair die is

$$\mu = \mathbb{E}_{z\sim p}(z) = \sum_{z=1}^{6} z\, p(z) = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\hat{\mu} = \frac{1}{l}\sum_{i=1}^{l} z^i$$

$z^1 = 3$ $\quad z^2 = 1$ $\quad z^3 = 5$ $\quad z^l = 2$

Law of large numbers

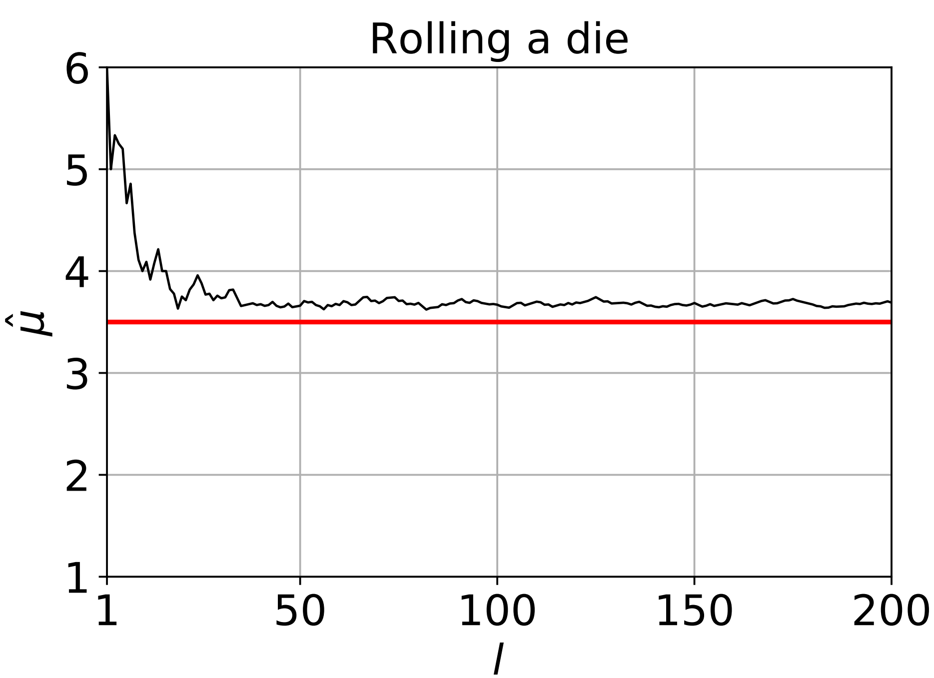
- Sample mean (arithmetic average) of the results of random trials gets closer to the expected value as more trials are performed.

- Example: The expected value of a single roll of a fair die is

$$\mu = \mathbb{E}_{z\sim p}(z) = \sum_{z=1}^{6} z\, p(z) = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\hat{\mu} = \frac{1}{l}\sum_{i=1}^{l} z^i$$

$z^1 = 3$ $z^2 = 1$ $z^3 = 5$ … $z^l = 2$



Law of large numbers

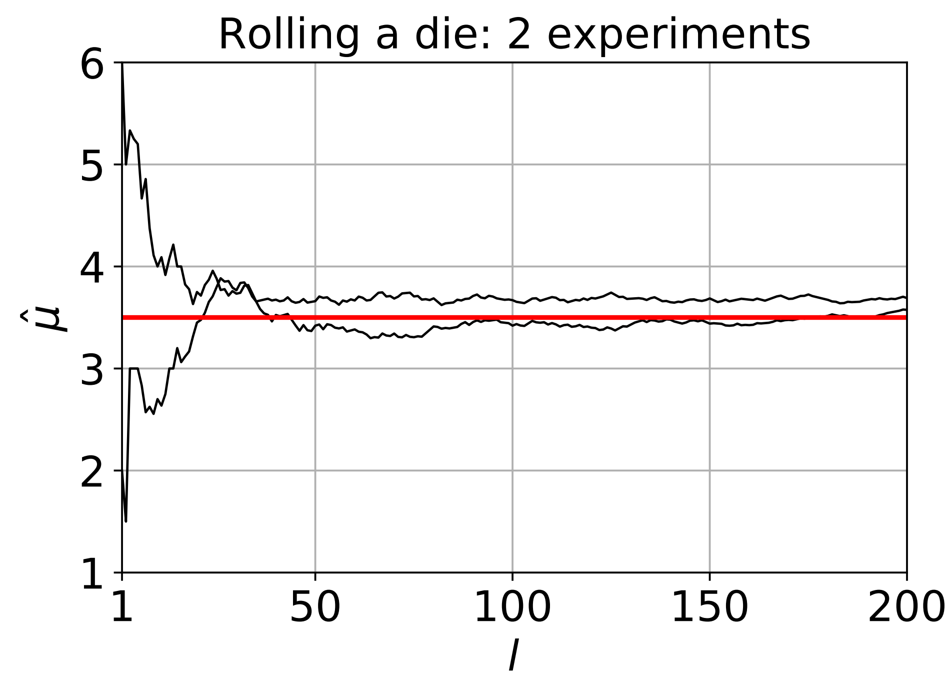
- Sample mean (arithmetic average) of the results of random trials gets closer to the expected value as more trials are performed.

- Example: The expected value of a single roll of a fair die is

$$\mu = \mathbb{E}_{z \sim p}(z) = \sum_{z=1}^{6} z\, p(z) = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\hat{\mu} = \frac{1}{l} \sum_{i=1}^{l} z^i$$

$z^1 = 3$ $z^2 = 1$ $z^3 = 5$ $\bullet\bullet\bullet$ $z^l = 2$



Law of large numbers

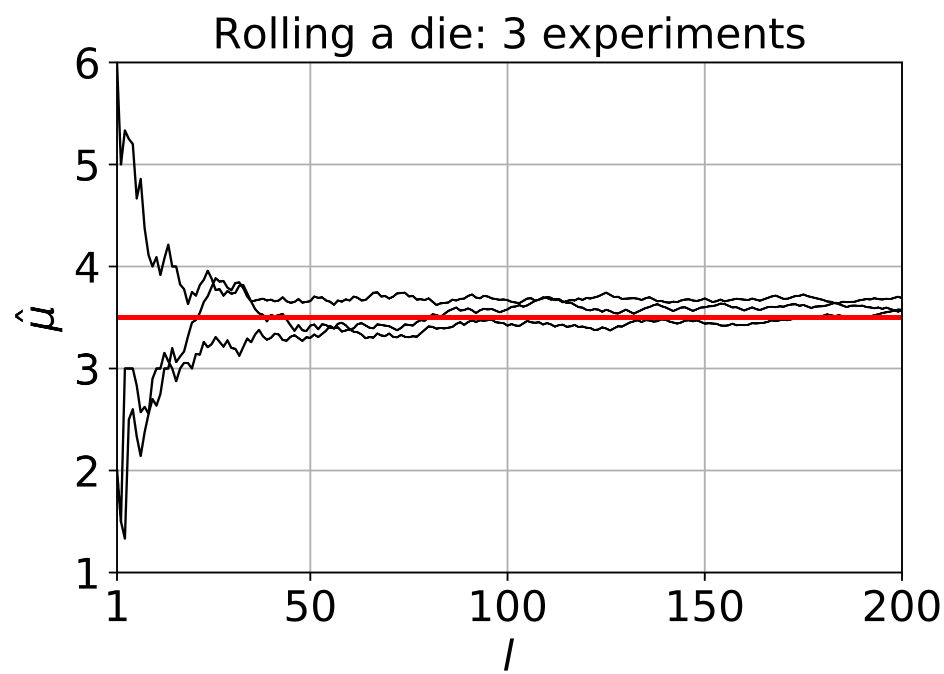
- Sample mean (arithmetic average) of the results of random trials gets closer to the expected value as more trials are performed.

- Example: The expected value of a single roll of a fair die is

$$\mu = \mathbb{E}_{z\sim p}(z) = \sum_{z=1}^{6} z\, p(z) = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\hat{\mu} = \frac{1}{l} \sum_{i=1}^{l} z^i$$

$z^1 = 3$ $z^2 = 1$ $z^3 = 5$ $z^l = 2$



Law of large numbers

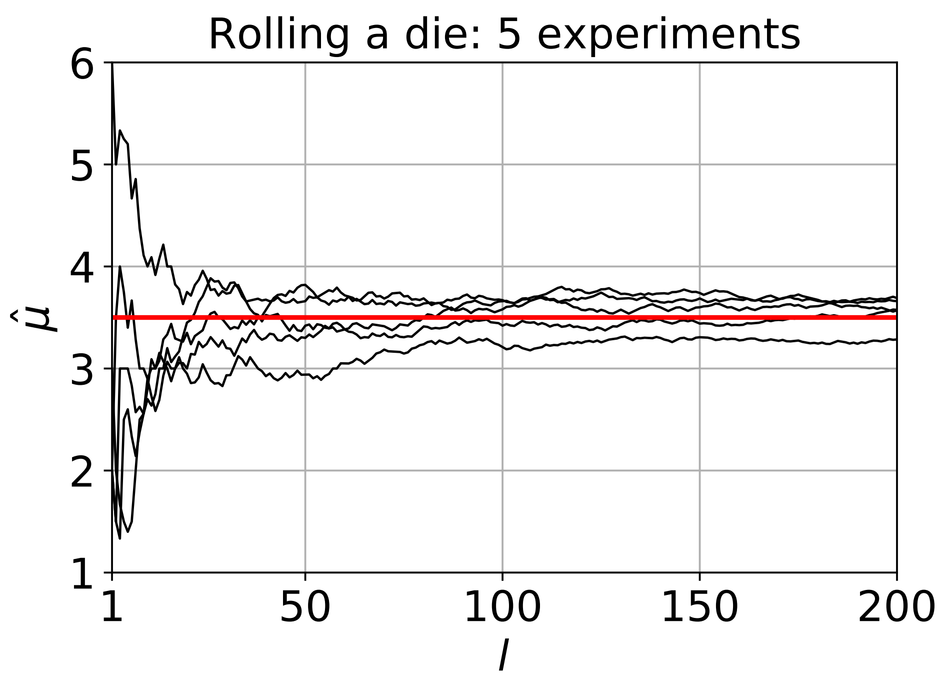
- Sample mean (arithmetic average) of the results of random trials gets closer to the expected value as more trials are performed.

- Example: The expected value of a single roll of a fair die is

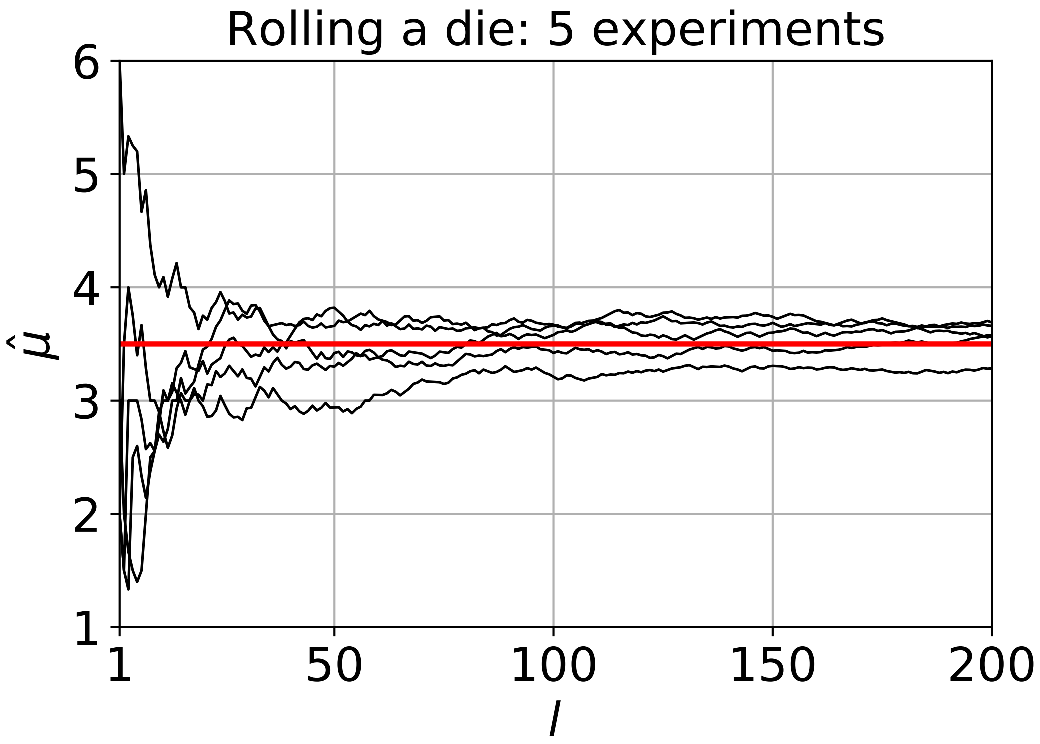
$$\mu = \mathbb{E}_{z \sim p}(z) = \sum_{z=1}^{6} z\, p(z) = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\hat{\mu} = \frac{1}{l} \sum_{i=1}^{l} z^i$$

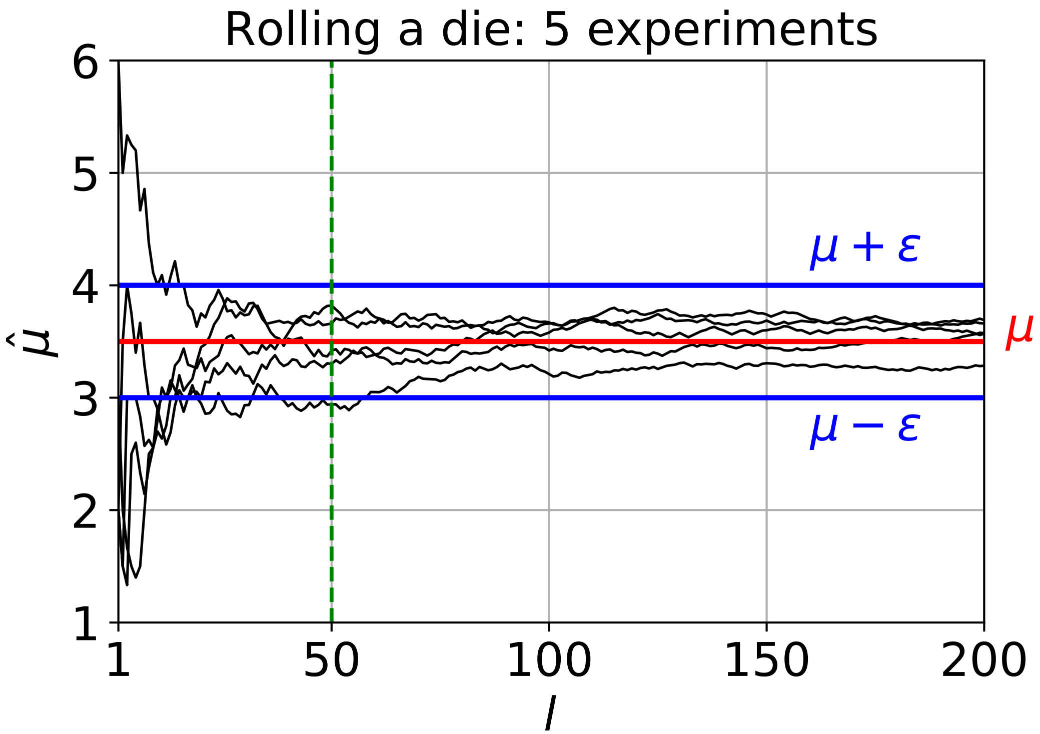
$z^1 = 3$ $z^2 = 1$ $z^3 = 5$ $z^l = 2$



Counting frequency of bad estimates

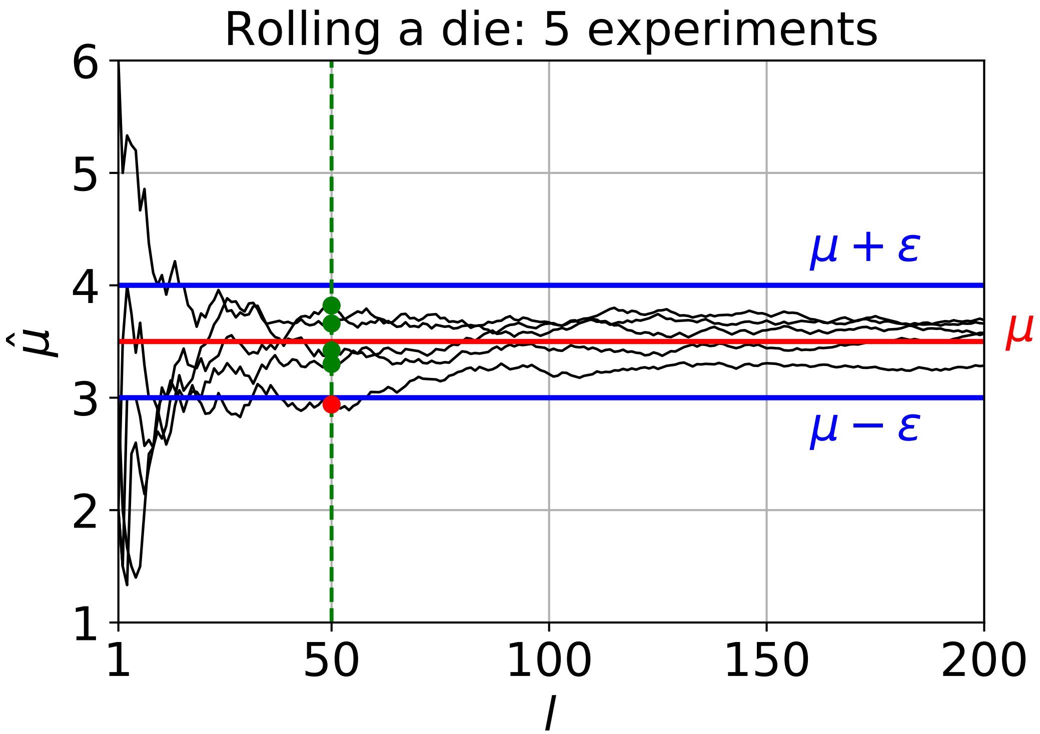


Counting frequency of bad estimates



sample size $l = 50$, deviation $\varepsilon = 0.5$

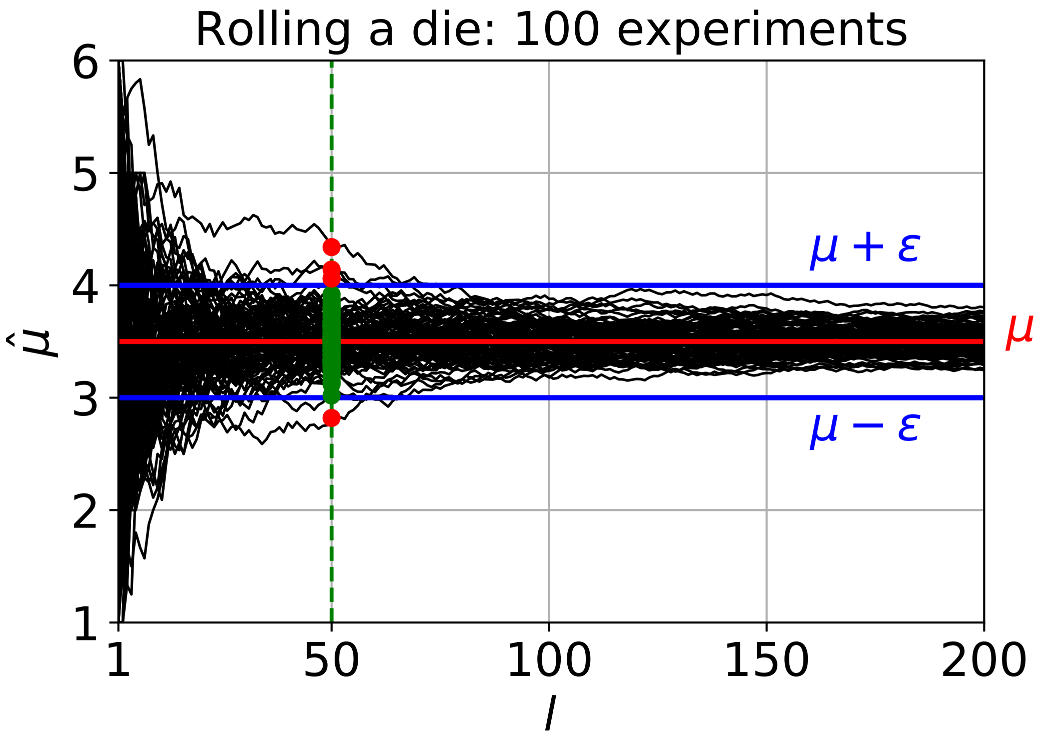
Counting frequency of bad estimates



sample size $l = 50$, deviation $\varepsilon = 0.5$

$$\frac{\#\left(|\hat{\mu} - \mu| \geq \varepsilon\right)}{\#\text{experiments}} = \frac{1}{5} = 0.2$$

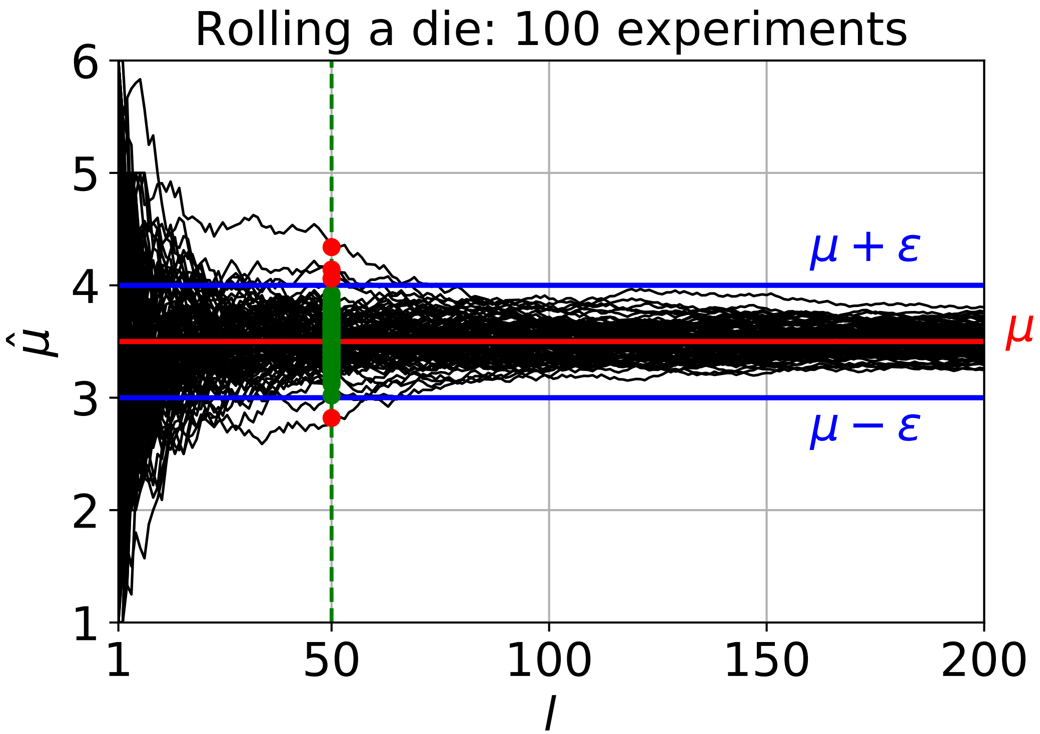
Counting frequency of bad estimates



sample size $l = 50$, deviation $\varepsilon = 0.5$

$$\frac{\#\big(|\hat{\mu}-\mu| \geq \varepsilon\,\big)}{\#\text{experiments}} = \frac{5}{100} = 0.05$$

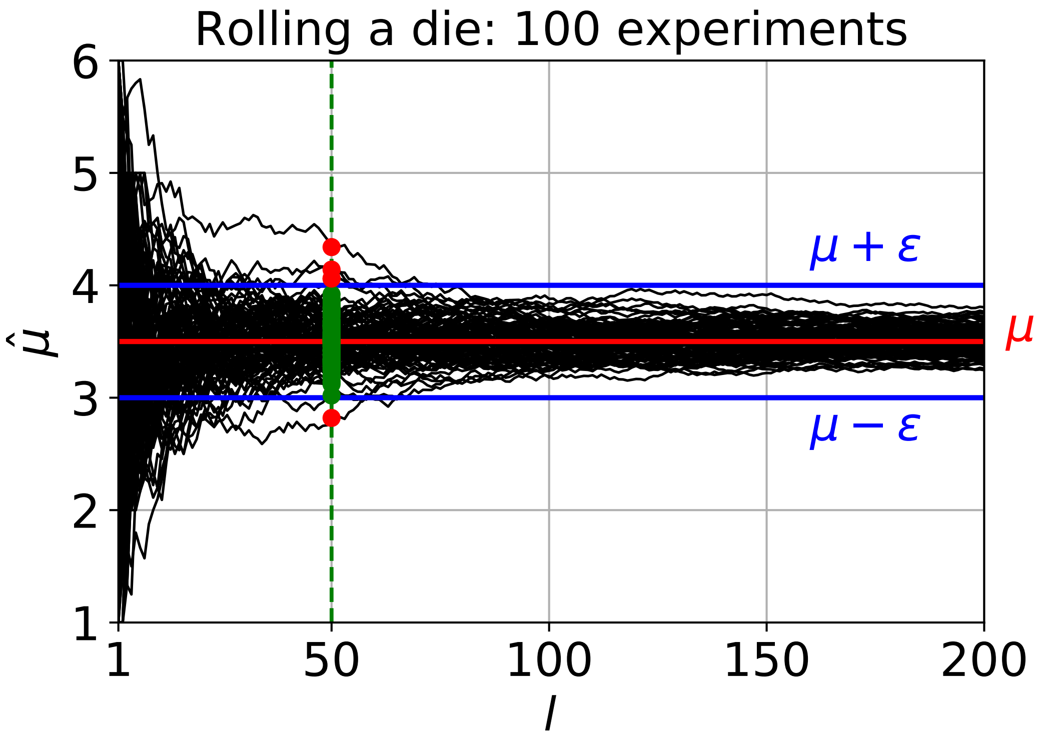
Counting frequency of bad estimates



sample size $l = 50$, deviation $\varepsilon = 0.5$

$$\frac{\#\Big(|\hat{\mu} - \mu| \geq \varepsilon\Big)}{\#\text{experiments}} = \frac{5}{100} = 0.05 \quad \rightarrow \quad \mathbb{P}\Big(|\hat{\mu} - \mu| \geq \varepsilon\Big)$$

Counting frequency of bad estimates



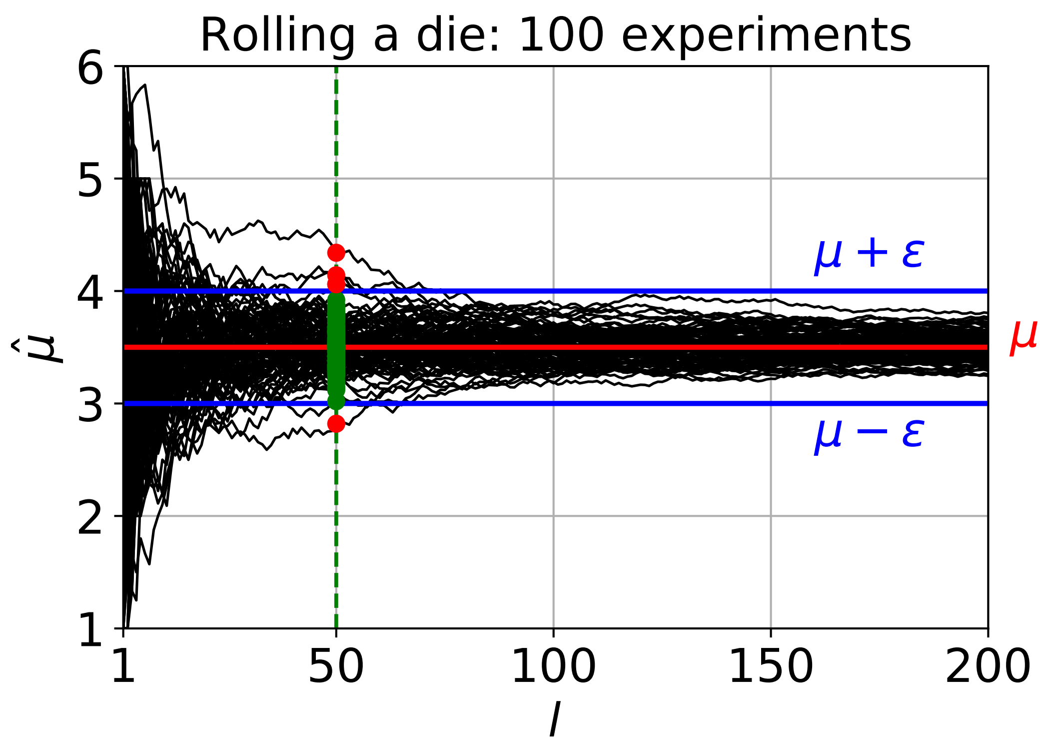
sample size $l = 50$, deviation $\varepsilon = 0.5$

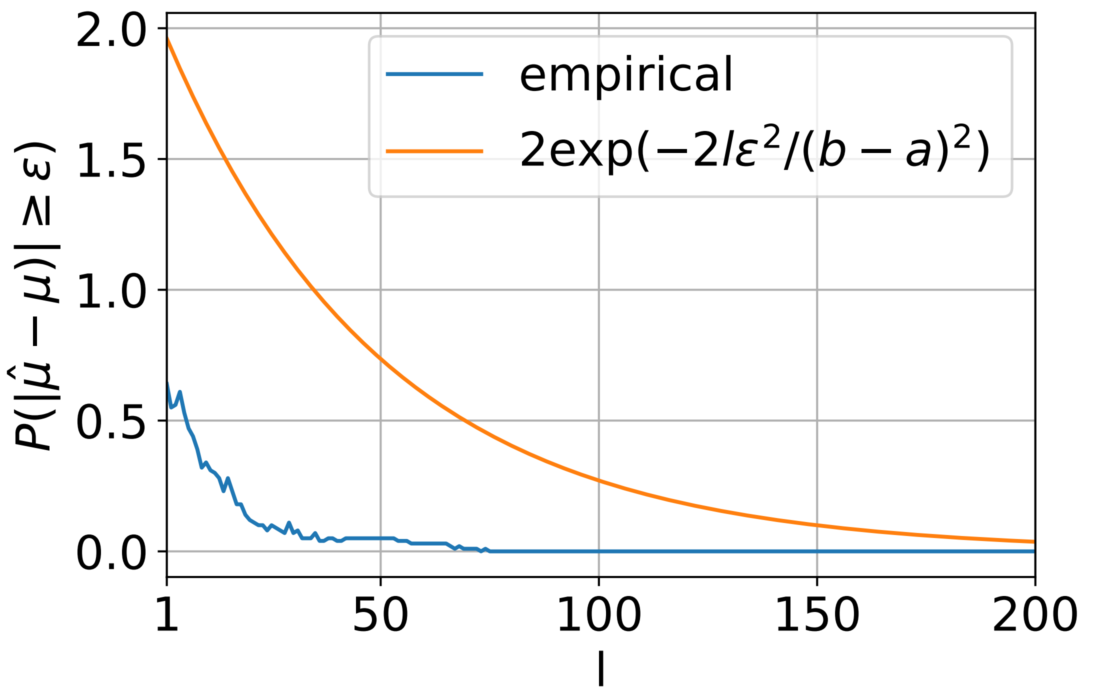
Hoeffding inequality

$$\frac{\#\big(|\hat{\mu}-\mu| \geq \varepsilon\big)}{\#\text{experiments}} = \frac{5}{100} = 0.05 \quad \rightarrow \quad \mathbb{P}\Big(|\hat{\mu}-\mu| \geq \varepsilon\Big) \leq 2e^{-\frac{2l\,\varepsilon^2}{(b-a)^2}}$$

$$a = 1,\; b = 6$$

Counting frequency of bad estimates





sample size $l = 50$, deviation $\varepsilon = 0.5$

Hoeffding inequality

$$\frac{\#\big(|\hat{\mu} - \mu| \geq \varepsilon\big)}{\#\text{experiments}} = \frac{5}{100} = 0.05 \quad \to \quad \mathbb{P}\Big(|\hat{\mu} - \mu| \geq \varepsilon\Big) \leq 2e^{-\frac{2l\,\varepsilon^2}{(b-a)^2}}$$

$$a = 1,\ b = 6$$

Hoeffding inequality

Theorem: Let $\{z^1, \ldots, z^l\}$ be a sample from independent r.v. from $[a, b]$ with expected value μ. Let $\hat{\mu} = \frac{1}{l}\sum_{i=1}^{l} z^i$. Then for any $\varepsilon > 0$ it holds that

$$\mathbb{P}\Big(|\hat{\mu} - \mu| \geq \varepsilon\Big) \leq 2e^{-\frac{2l\,\varepsilon^2}{(b-a)^2}}$$

Hoeffding inequality

Theorem: Let $\{z^1, \ldots, z^l\}$ be a sample from independent r.v. from $[a, b]$ with expected value μ. Let $\hat{\mu} = \frac{1}{l}\sum_{i=1}^{l} z^i$. Then for any $\varepsilon > 0$ it holds that

$$\mathbb{P}\Big(|\hat{\mu} - \mu| \geq \varepsilon\Big) \leq 2e^{-\frac{2l\,\varepsilon^2}{(b-a)^2}}$$

Properties:

- Conservative: the bound may not be tight.
- General: the bound holds for any distribution.
- Cheap: The bound is simple and easy to compute.

Confidence intervals

$(\boldsymbol{l}, \boldsymbol{\gamma}) \rightarrow \boldsymbol{\varepsilon}$

- Let $\hat{\mu} = \frac{1}{l}\sum_{i=1}^{l} z^i$ be the sample mean computed from $\{z^1, \ldots, z^l\} \in [a, b]^l$ sampled from r.v. with expected value μ.

- Find ε such that $\mu \in (\hat{\mu} - \varepsilon, \hat{\mu} + \varepsilon)$ with probability at least γ.

Confidence intervals
$(l, \gamma) \to \varepsilon$

- Let $\hat{\mu} = \frac{1}{l}\sum_{i=1}^{l} z^i$ be the sample mean computed from $\{z^1, \ldots, z^l\} \in [a, b]^l$ sampled from r.v. with expected value μ.

- Find ε such that $\mu \in (\hat{\mu} - \varepsilon, \hat{\mu} + \varepsilon)$ with probability at least γ.

  Using the Hoeffding inequality we can write

$$\mathbb{P}\Big(|\hat{\mu} - \mu| < \varepsilon\Big) = 1 - \mathbb{P}\Big(|\hat{\mu} - \mu| \geq \varepsilon\Big) \geq 1 - 2e^{-\frac{2\,l\;\varepsilon^2}{(b-a)^2}} = \gamma$$

  and solving the last equation for ε yields

$$\varepsilon = |b - a|\sqrt{\frac{\log(2) - \log(1-\gamma)}{2\,l}}$$

Confidence intervals
$(\varepsilon, \gamma) \rightarrow l$

- Let $\hat{\mu} = \frac{1}{l}\sum_{i=1}^{l} z^i$ be the sample mean computed from $\{z^1, \ldots, z^l\} \in [a, b]^l$ sampled from r.v. with expected value μ.

- Given a fixed $\varepsilon > 0$ and $\gamma \in (0, 1)$, what is the minimal number of examples l such that $\mu \in (\hat{\mu} - \varepsilon, \hat{\mu} + \varepsilon)$ with probability γ at least ?

Confidence intervals
$(\varepsilon, \gamma) \to l$

- Let $\hat{\mu} = \frac{1}{l}\sum_{i=1}^{l} z^i$ be the sample mean computed from $\{z^1, \ldots, z^l\} \in [a,b]^l$ sampled from r.v. with expected value μ.

- Given a fixed $\varepsilon > 0$ and $\gamma \in (0,1)$, what is the minimal number of examples l such that $\mu \in (\hat{\mu} - \varepsilon, \hat{\mu} + \varepsilon)$ with probability γ at least ?

  Starting from

$$\mathbb{P}\Big(|\hat{\mu} - \mu| < \varepsilon\Big) = 1 - \mathbb{P}\Big(|\hat{\mu} - \mu| \geq \varepsilon\Big) \geq 1 - 2e^{-\frac{2\,l\,\varepsilon^2}{(b-a)^2}} = \gamma$$

  and solving for l yields

$$l = \frac{\log(2) - \log(1-\gamma)}{2\,\varepsilon^2}\,(b-a)^2$$

Back to the problem: Estimation of the true risk by using confidence intervals

- Given $h \colon \mathcal{X} \to \mathcal{Y}$ estimate the true risk $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x)))$ by the empirical risk $R_{\mathcal{S}^l}(h) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(x^i))$ using the test set $\mathcal{S}^l$.

- Confidence interval:

$$R(h) \in \big(R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon\big) \quad \text{with probability} \quad \gamma \in (0,1)$$

Back to the problem: Estimation of the true risk by using confidence intervals

- Given $h\colon \mathcal{X} \to \mathcal{Y}$ estimate the true risk $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x)))$ by the empirical risk $R_{\mathcal{S}^l}(h) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(x^i))$ using the test set $\mathcal{S}^l$.

- Confidence interval:

$$R(h) \in \big(R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon\big) \quad \text{with probability} \quad \gamma \in (0,1)$$

- For fixed l and $\gamma \in (0,1)$ compute interval width

$$\varepsilon = (\ell_{\max} - \ell_{\min})\sqrt{\frac{\log(2) - \log(1-\gamma)}{2\,l}}\,.$$

- For fixed ε and $\gamma \in (0,1)$ compute number of test examples

$$l = \frac{\log(2) - \log(1-\gamma)}{2\,\varepsilon^2}\,(\ell_{\max} - \ell_{\min})^2$$

Learning: the definition

- **The goal:** Find a strategy $h\colon \mathcal{X} \to \mathcal{Y}$ minimizing $R(h)$ using the training set of examples

$$\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$$

drawn from i.i.d. rv. with unknown $p(x, y)$.

- **Hypothesis class (space):**

$$\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}} = \{h\colon \mathcal{X} \to \mathcal{Y}\}$$

- **Learning algorithm:** a function

$$A\colon \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$$

which returns a strategy $h_m = A(\mathcal{T}^m)$ for a training set $\mathcal{T}^m$

Learning: Empirical Risk Minimization approach

- The expected risk $R(h)$, i.e. the true but unknown objective, is replaced by the empirical risk computed from the training examples $\mathcal{T}^m$,

$$R_{\mathcal{T}^m}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(y^i, h(x^i))$$

- The ERM based algorithm returns h_m such that

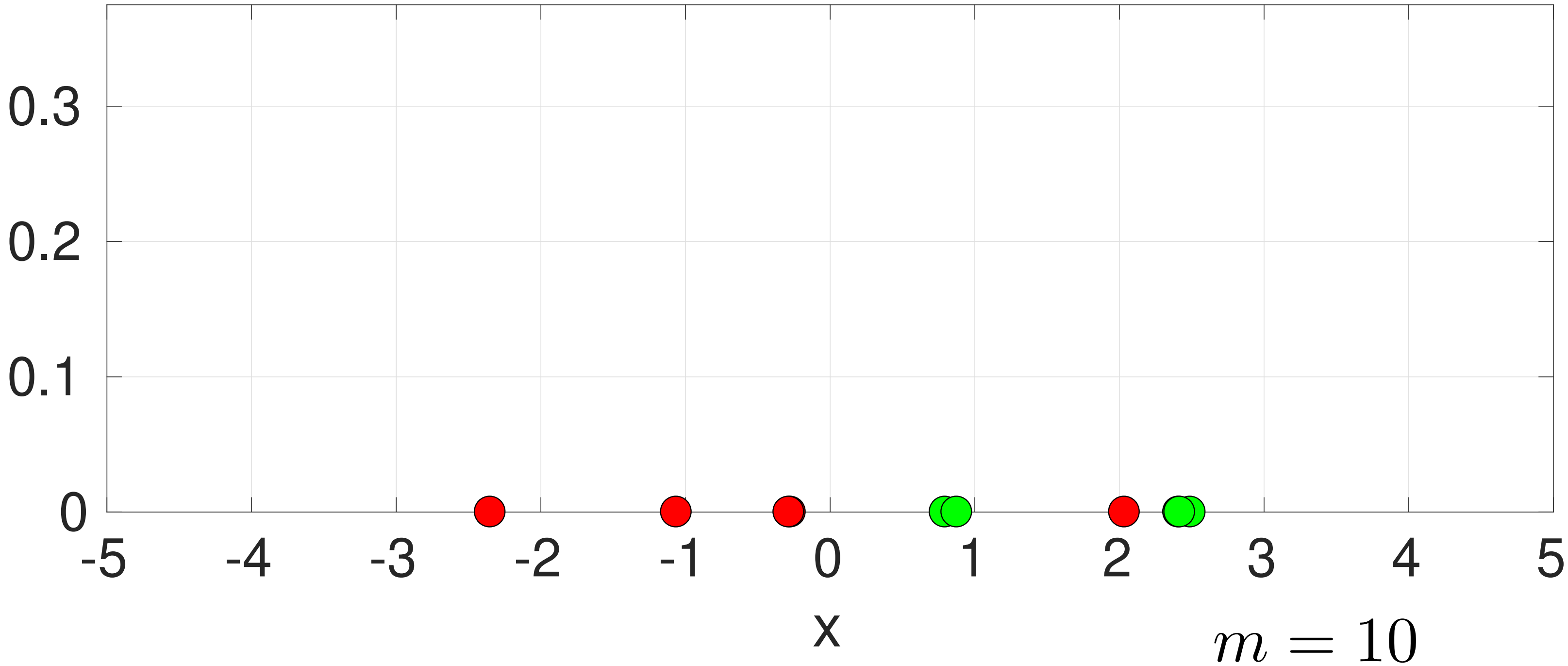
$$h_m \in \operatorname*{Argmin}_{h\in\mathcal{H}} R_{\mathcal{T}^m}(h) \tag{1}$$

Learning: Empirical Risk Minimization approach

- The expected risk $R(h)$, i.e. the true but unknown objective, is replaced by the empirical risk computed from the training examples $\mathcal{T}^m$,

$$R_{\mathcal{T}^m}(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(y^i, h(x^i))$$

$$\mathcal{H} = \{h(x) = \mathrm{sign}(x - \theta) \mid \theta \in \mathbb{R}\} \;,\;\; \ell(y, y') = [\![y \neq y']\!]$$



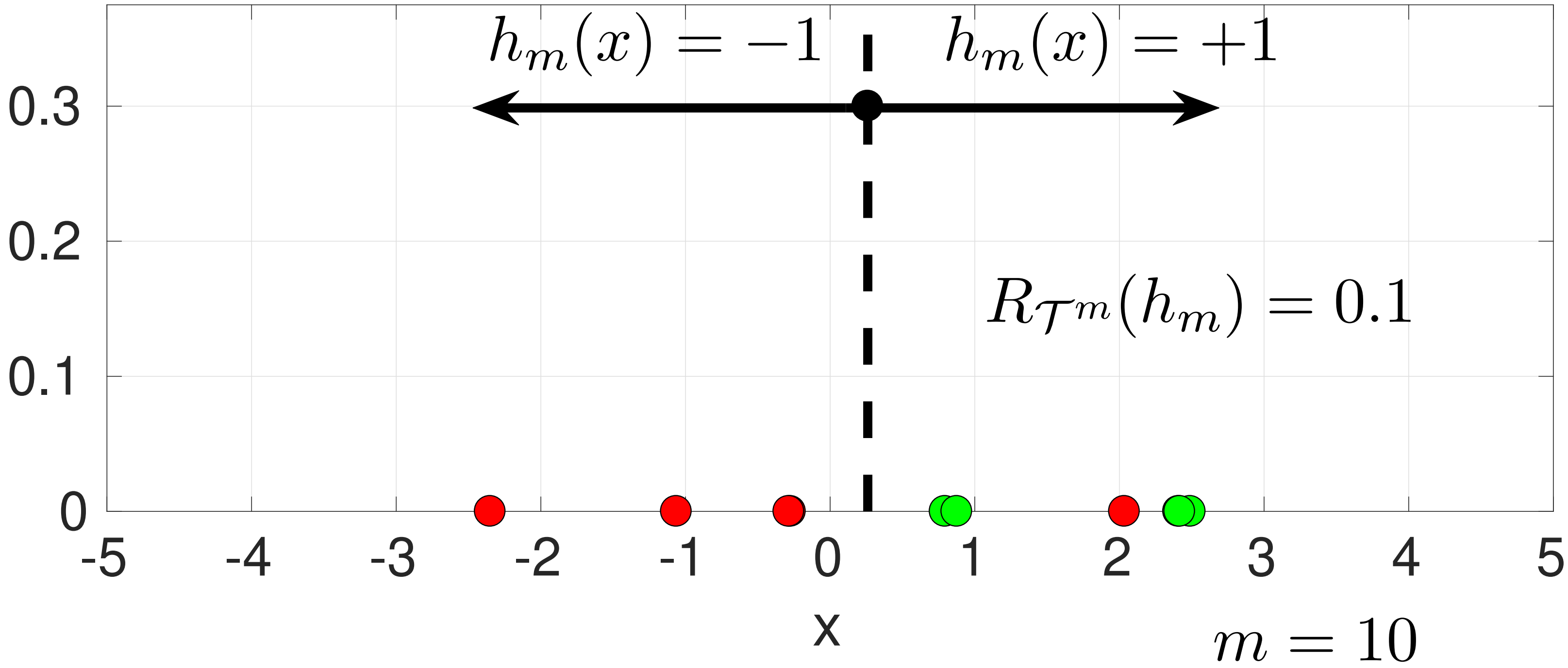
$m = 10$

Learning: Empirical Risk Minimization approach

- The expected risk $R(h)$, i.e. the true but unknown objective, is replaced by the empirical risk computed from the training examples $\mathcal{T}^m$,

$$R_{\mathcal{T}^m}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(y^i, h(x^i))$$

$$\mathcal{H} = \{h(x) = \text{sign}(x - \theta) \mid \theta \in \mathbb{R}\} \; , \;\; \ell(y, y') = [\![y \neq y']\!]$$



Learning: Empirical Risk Minimization approach

- The expected risk $R(h)$, i.e. the true but unknown objective, is replaced by the empirical risk computed from the training examples $\mathcal{T}^m$,

$$R_{\mathcal{T}^m}(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(y^i, h(x^i))$$

- The ERM based algorithm returns h_m such that

$$h_m \in \operatorname*{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h) \qquad (1)$$

- Depending on the choince of $\mathcal{H}$ and ℓ and algorithm solving (1) we get individual instances e.g. Support Vector Machines, Linear Regression, Logistic Regression, Neural Networks learned by back-propagation, AdaBoost, Gradient Boosted Trees, ...

Example of ERM failure

- Let $\mathcal{X} = [a, b] \subset \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = [\![y \neq y']\!]$, $p(x \mid y = +1)$ and $p(x \mid y = -1)$ be uniform distributions on $\mathcal{X}$ and $p(y = +1) = 0.8$.

Example of ERM failure

- Let $\mathcal{X} = [a, b] \subset \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = [\![y \neq y']\!]$, $p(x \mid y = +1)$ and $p(x \mid y = -1)$ be uniform distributions on $\mathcal{X}$ and $p(y = +1) = 0.8$.

- The optimal strategy is $h(x) = +1$ with the Bayes risk $R^* = 0.2$.

Example of ERM failure

- Let $\mathcal{X} = [a, b] \subset \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = [\![y \neq y']\!]$, $p(x \mid y = +1)$ and $p(x \mid y = -1)$ be uniform distributions on $\mathcal{X}$ and $p(y = +1) = 0.8$.

- The optimal strategy is $h(x) = +1$ with the Bayes risk $R^* = 0.2$.

- Consider learning algorithm which for a given training set $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\}$ returns memorizing strategy

$$h_m(x) = \begin{cases} y^j & \text{if } x = x^j \text{ for some } j \in \{1, \ldots, m\} \\ -1 & \text{otherwise} \end{cases}$$

Example of ERM failure

- Let $\mathcal{X} = [a, b] \subset \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = [\![y \neq y']\!]$, $p(x \mid y = +1)$ and $p(x \mid y = -1)$ be uniform distributions on $\mathcal{X}$ and $p(y = +1) = 0.8$.

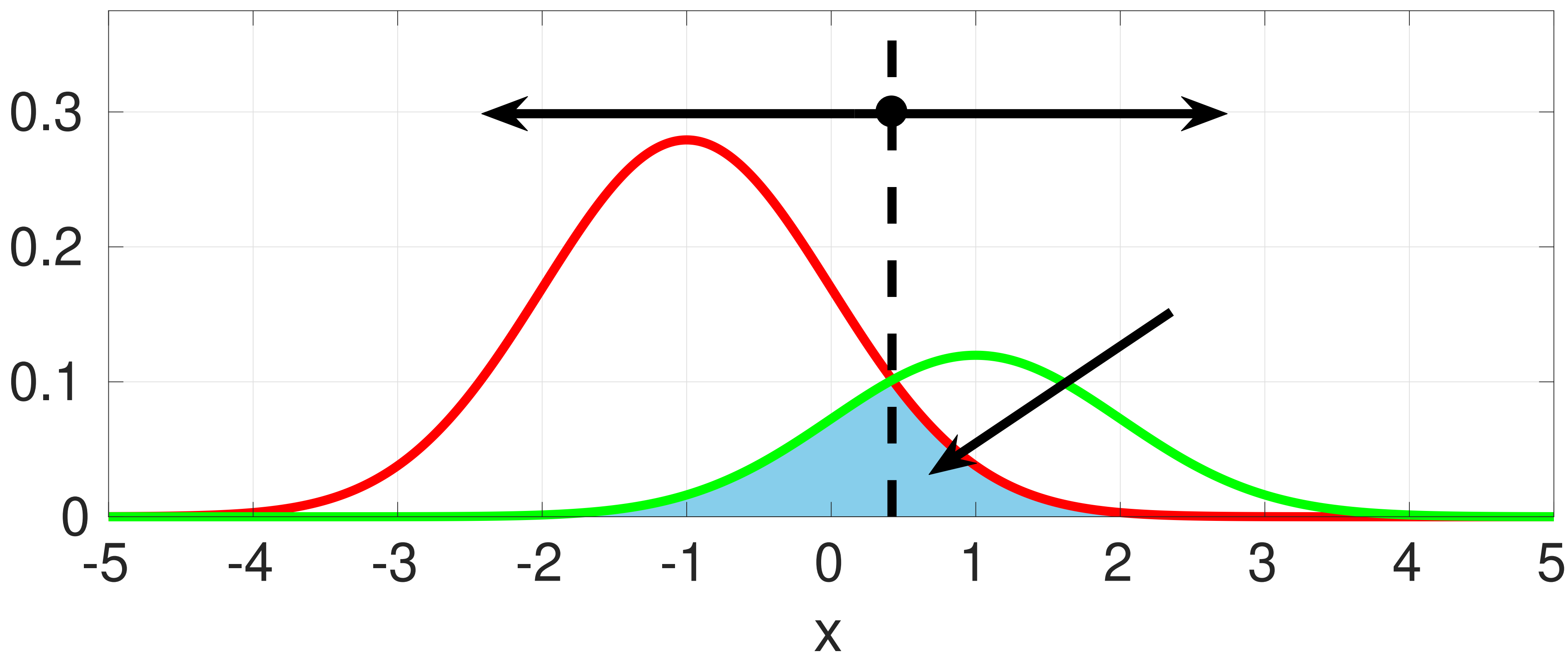
- The optimal strategy is $h(x) = +1$ with the Bayes risk $R^* = 0.2$.

- Consider learning algorithm which for a given training set $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\}$ returns memorizing strategy

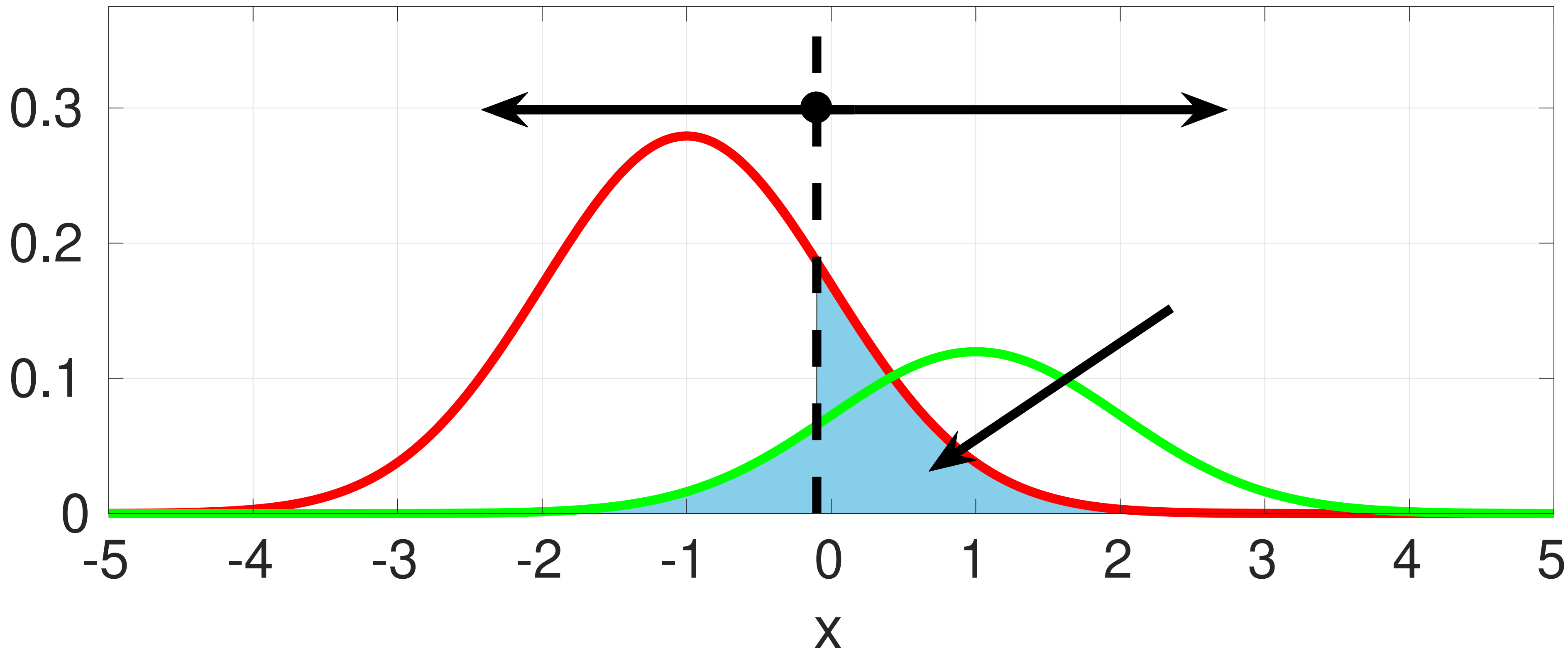
$$h_m(x) = \begin{cases} y^j & \text{if } \ x = x^j \ \text{ for some } \ j \in \{1, \ldots, m\} \\ -1 & \text{otherwise} \end{cases}$$

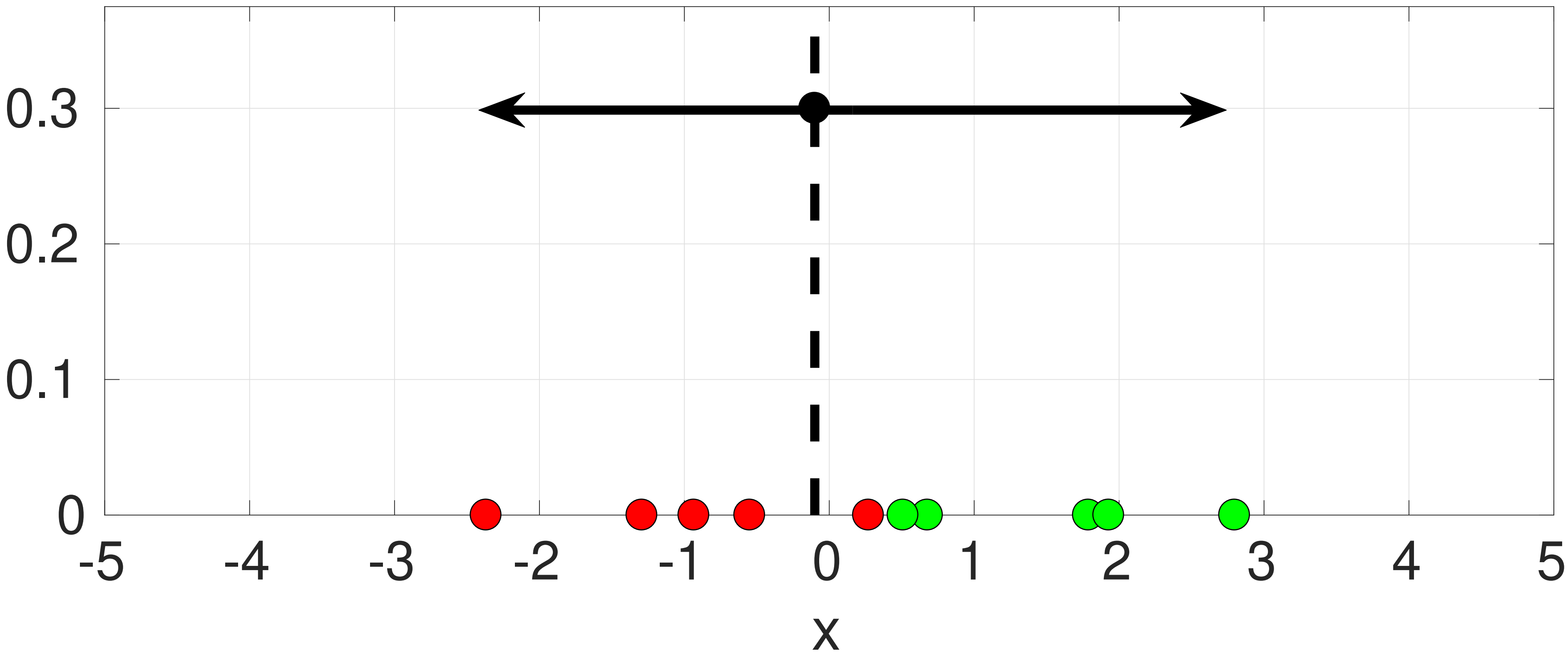
- The empirical risk is $R_{\mathcal{T}^m}(h_m) = 0$ with probability 1 for any m.

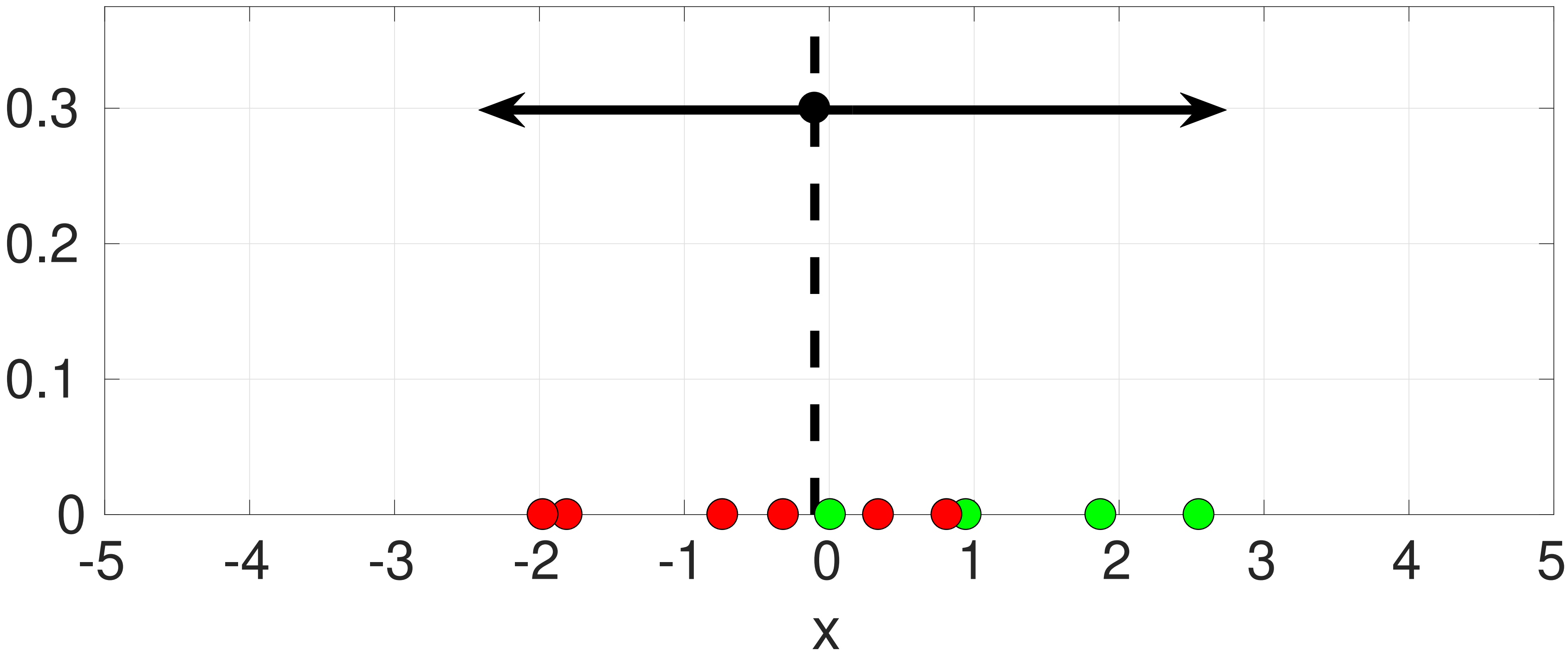
- The expected risk is $R(h_m) = 0.8$ for any m.



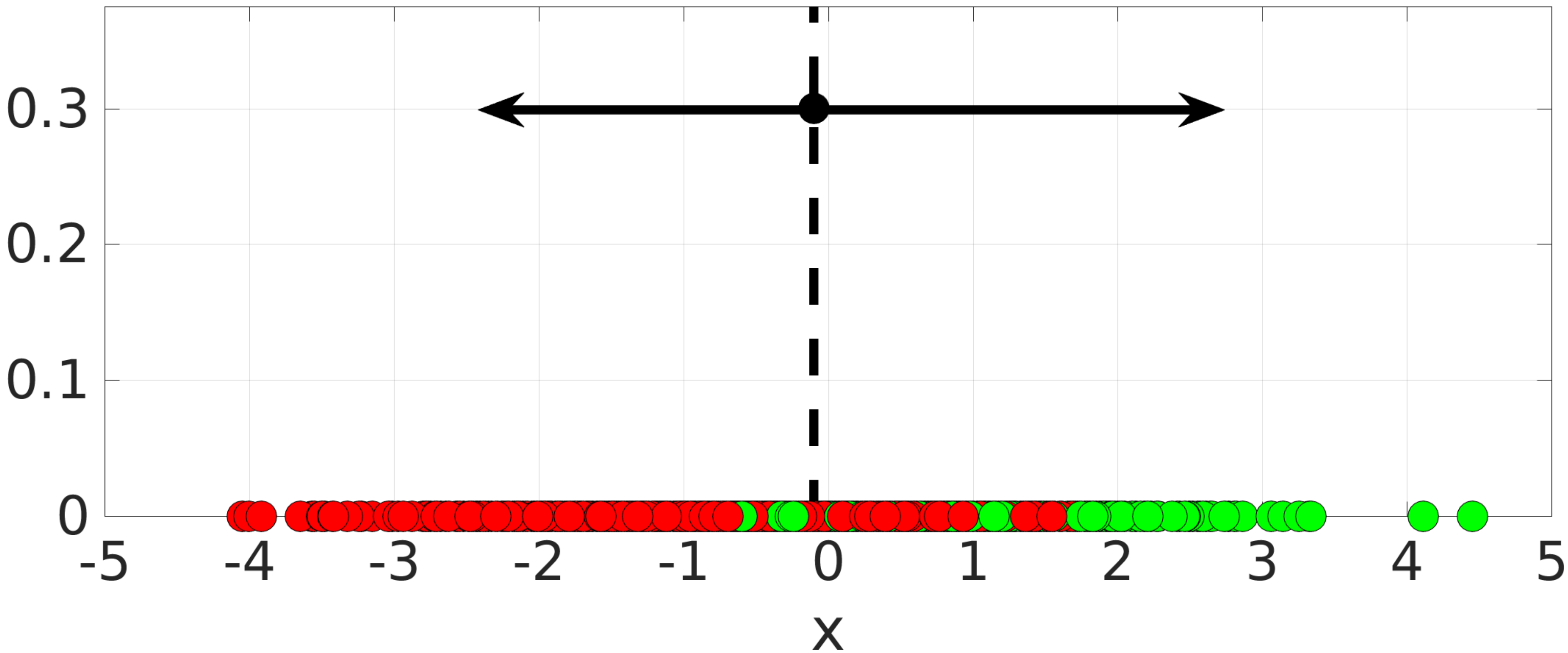
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
x


0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
x


0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
x



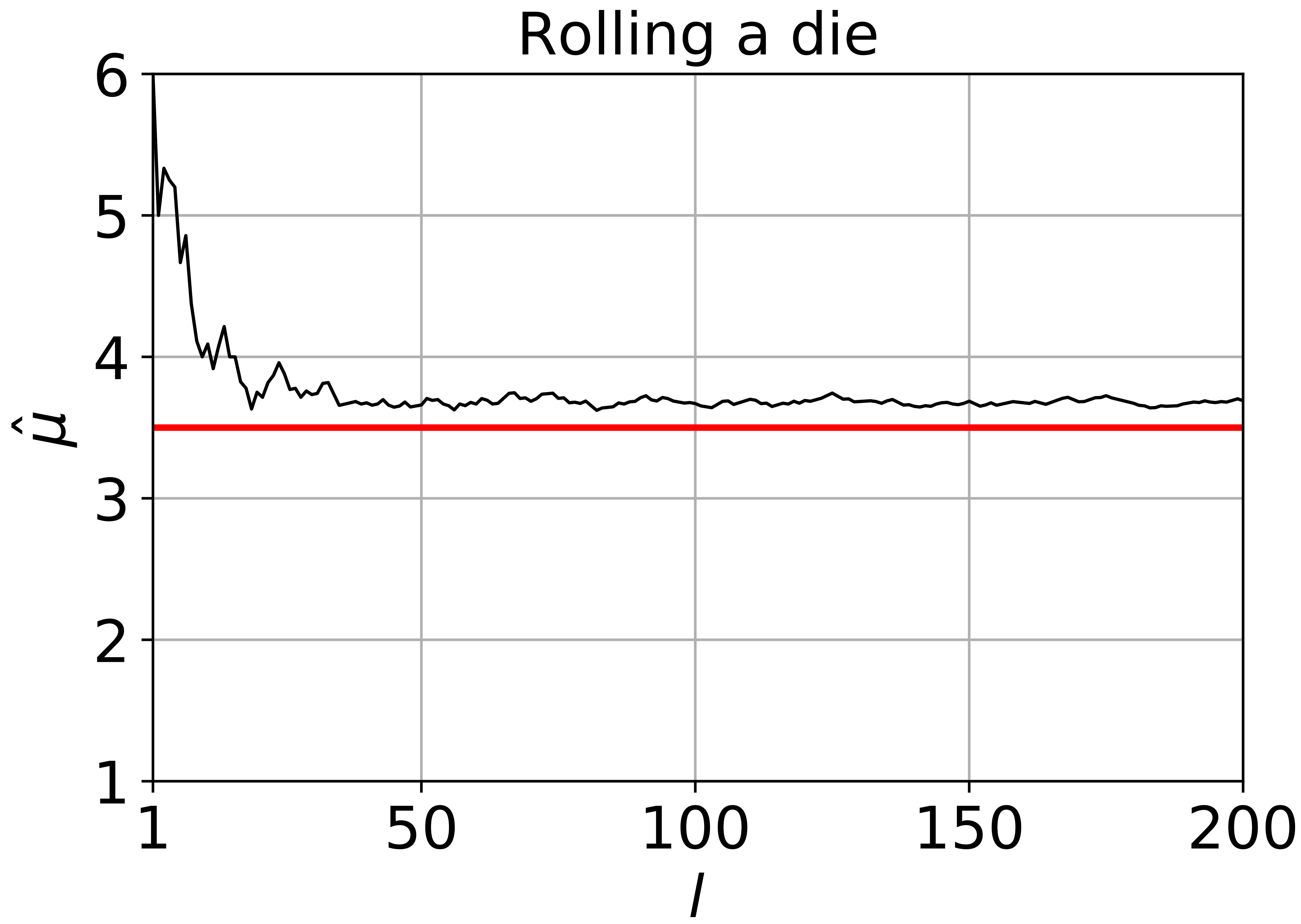
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
x



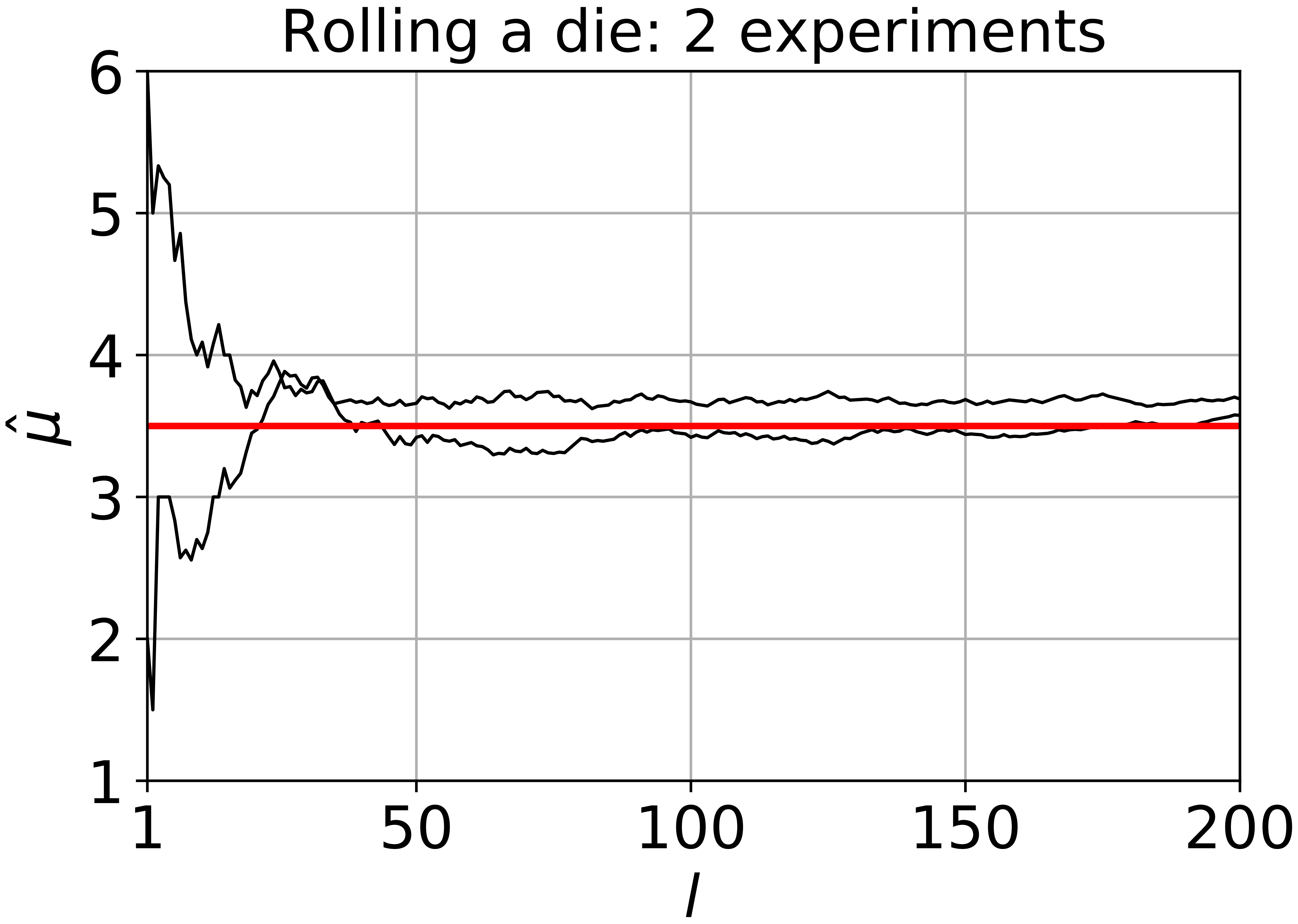
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
x

$z^1 = 3 \quad z^2 = 1 \quad z^3 = 5 \quad \cdots \quad z^l = 2$

$z^1 = 3 \quad z^2 = 1 \quad z^3 = 5 \quad \cdots \quad z^l = 2$

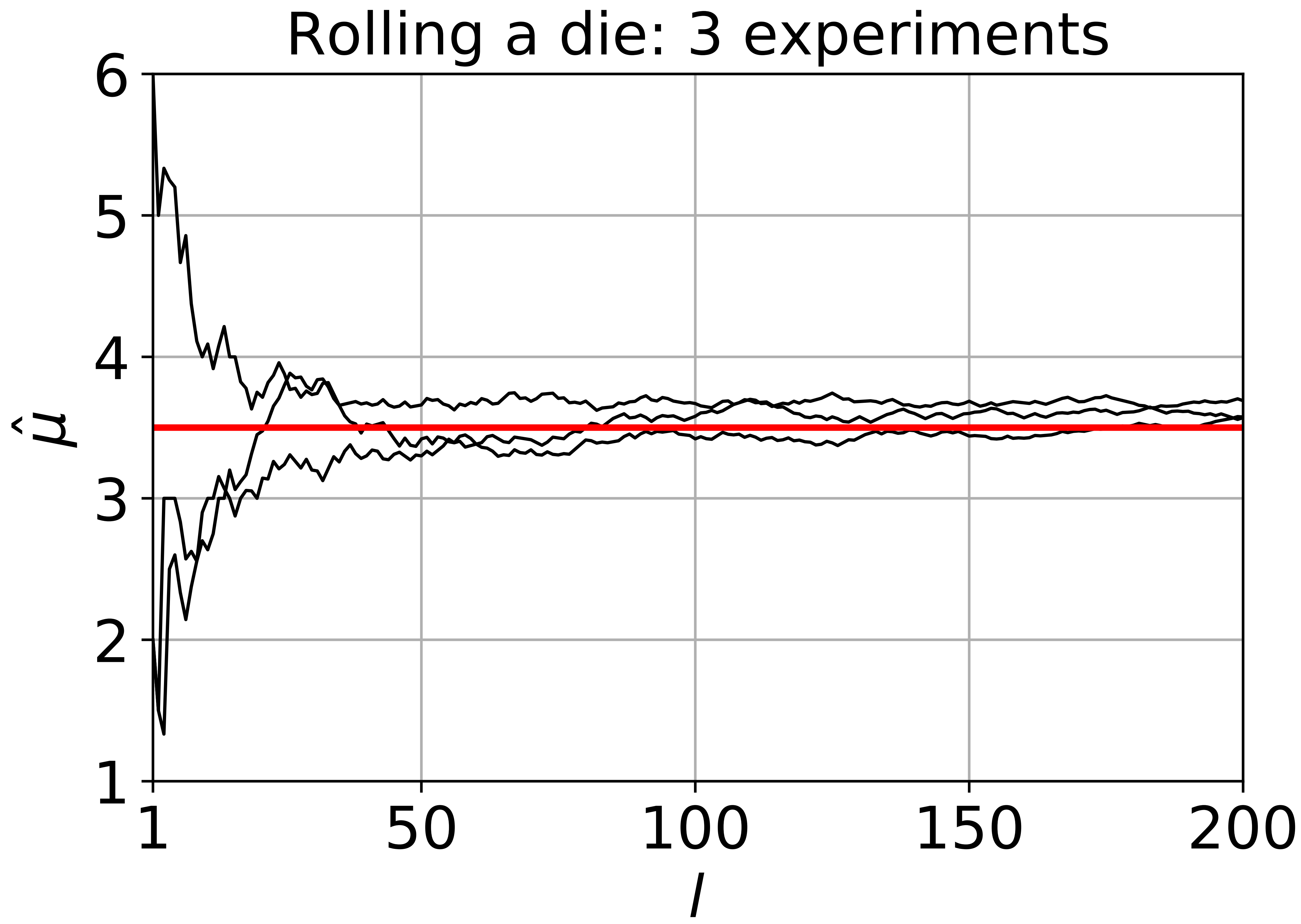

Rolling a die
6
5
4
3
2
1
$\hat{\mu}$
1
50
100
150
200
l

$z^1 = 3$ $z^2 = 1$ $z^3 = 5$ $\cdots$ $z^l = 2$



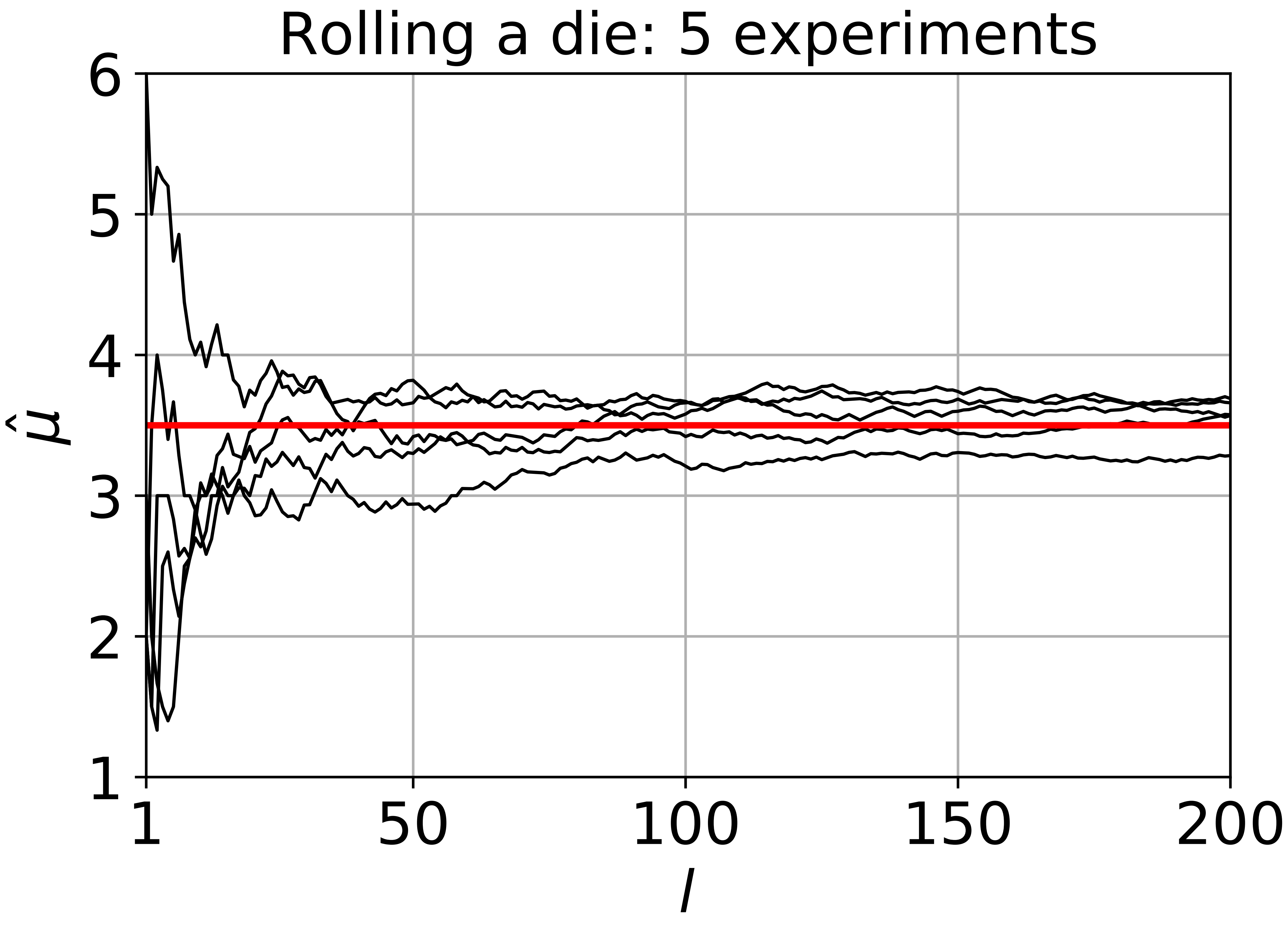
Rolling a die: 2 experiments
6
5
4
3
2
1
$\hat{\mu}$
1
50
100
150
200
l

$z^1 = 3 \quad z^2 = 1 \quad z^3 = 5 \quad \cdots \quad z^l = 2$

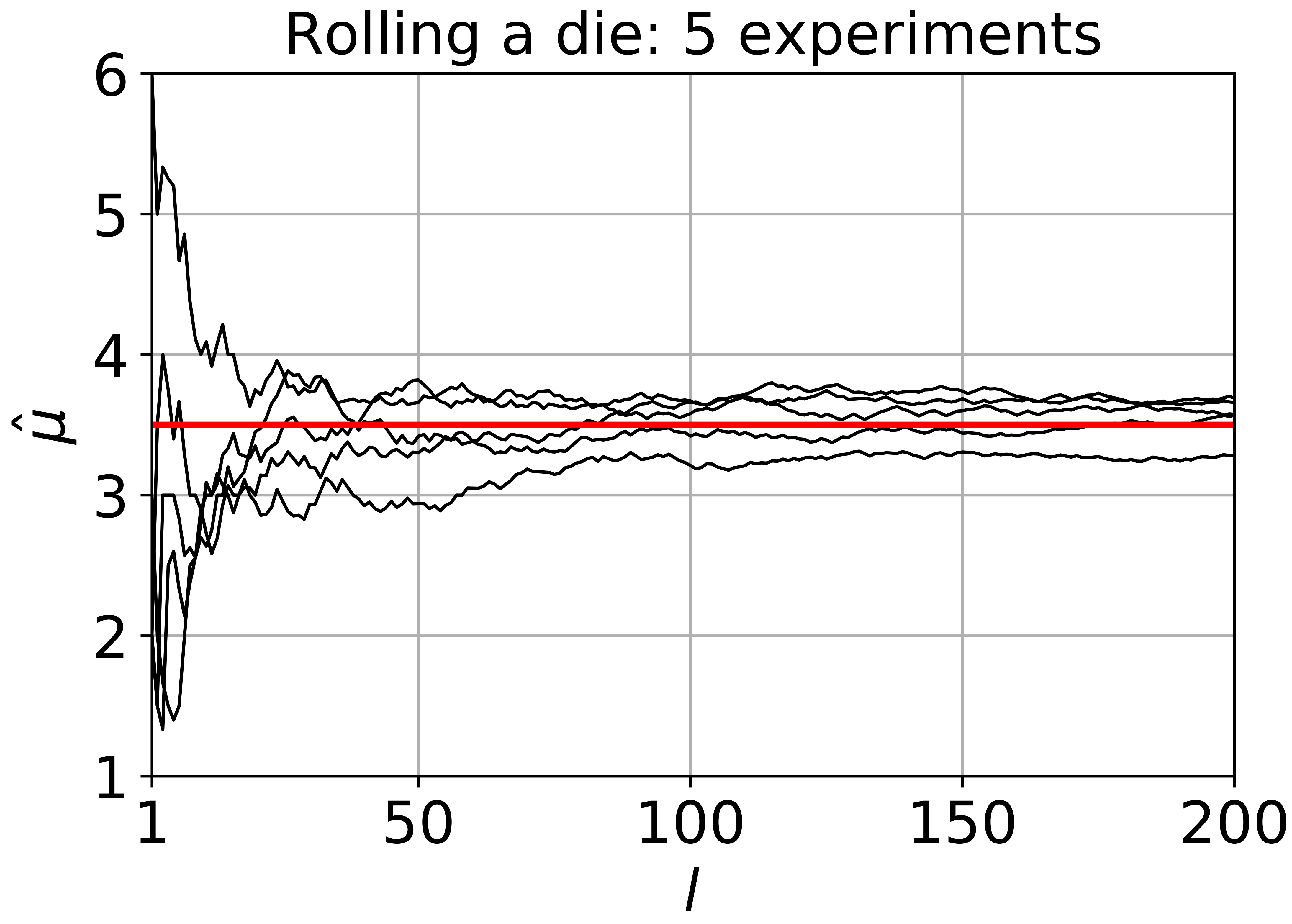


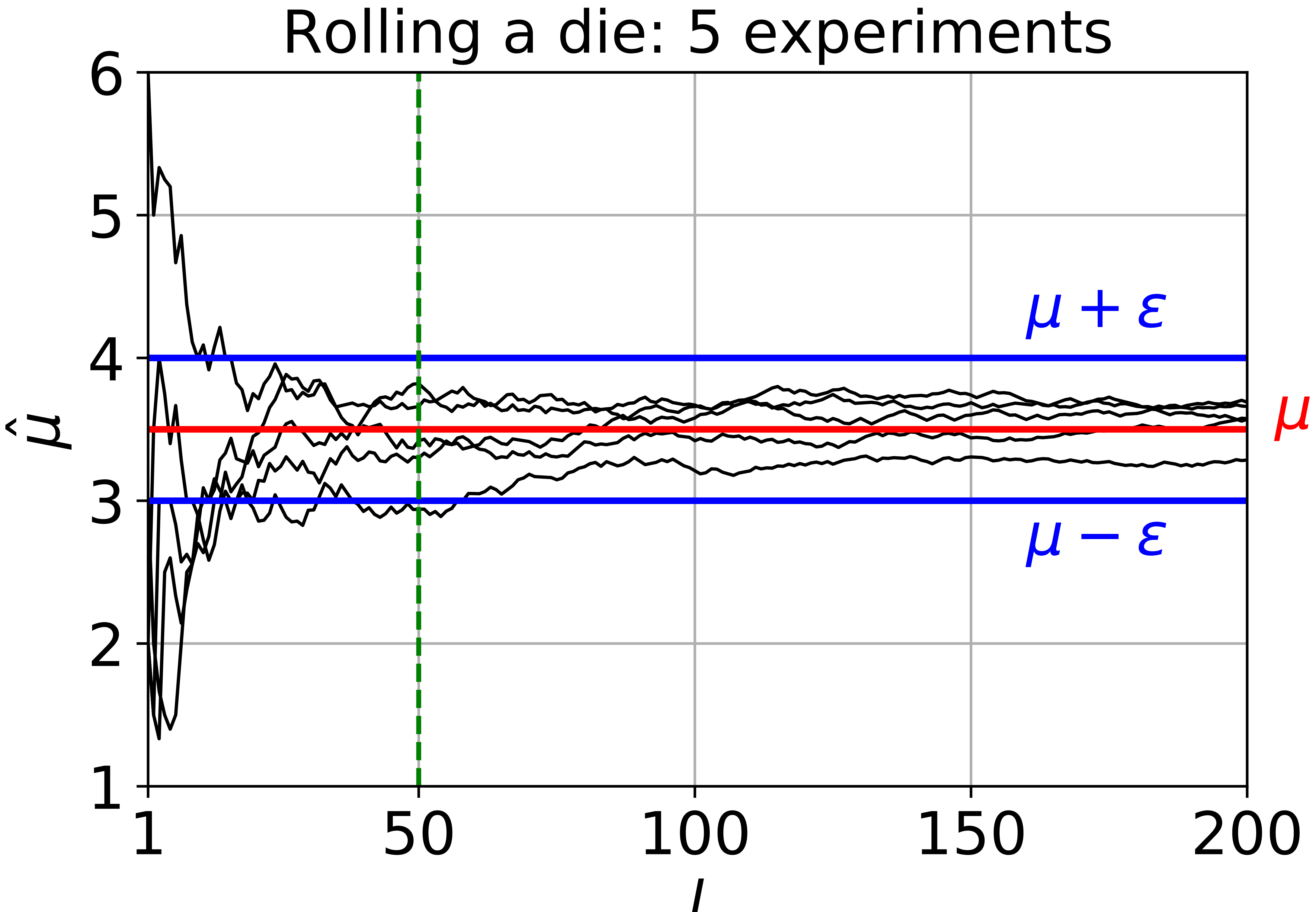
Rolling a die: 3 experiments
6
5
4
3
2
1
$\hat{\mu}$
1
50
100
150
200
l

$z^1 = 3$ $z^2 = 1$ $z^3 = 5$ $\cdots$ $z^l = 2$

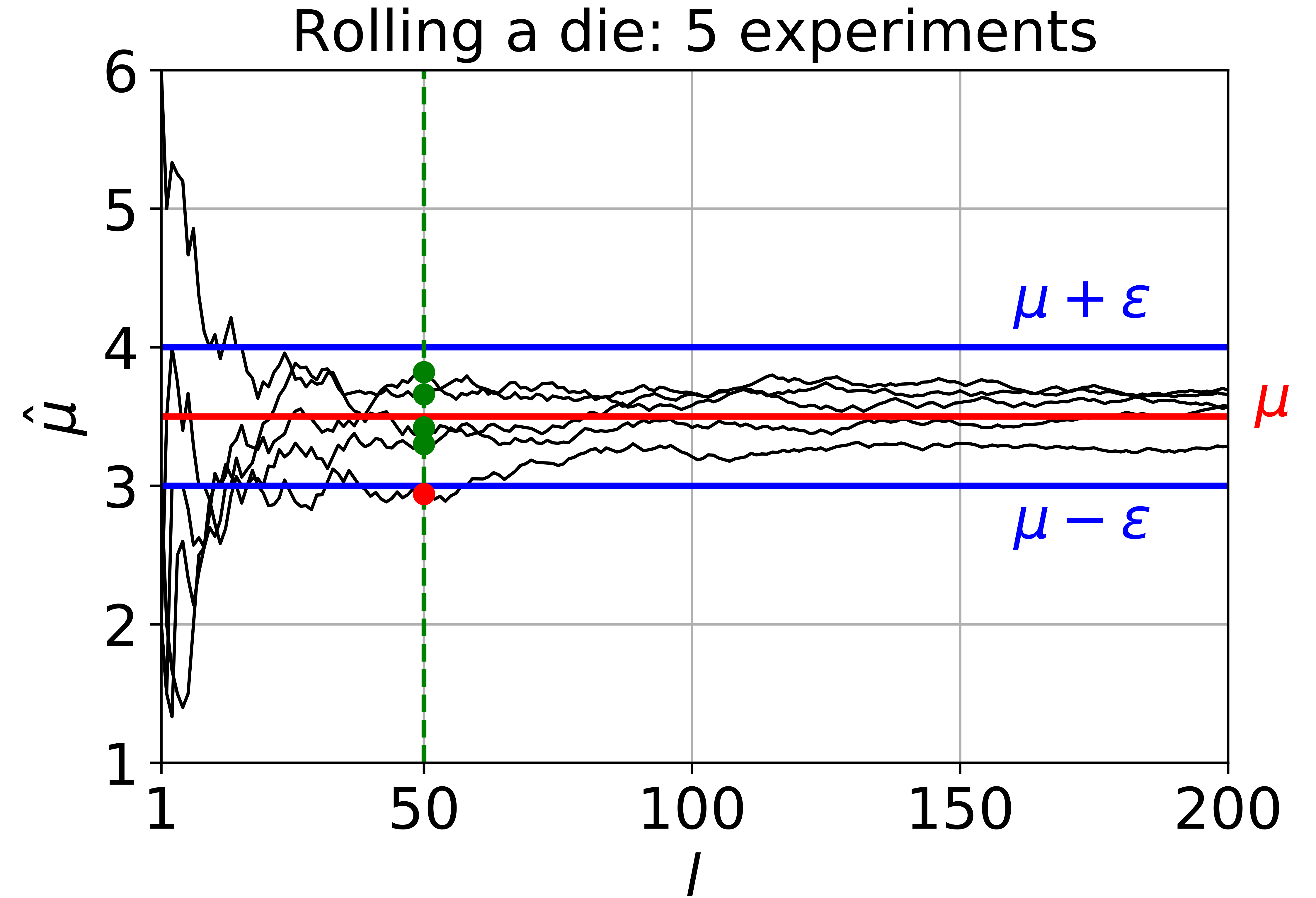


Rolling a die: 5 experiments
6
5
4
3
2
1
$\hat{\mu}$
1
50
100
150
200
l

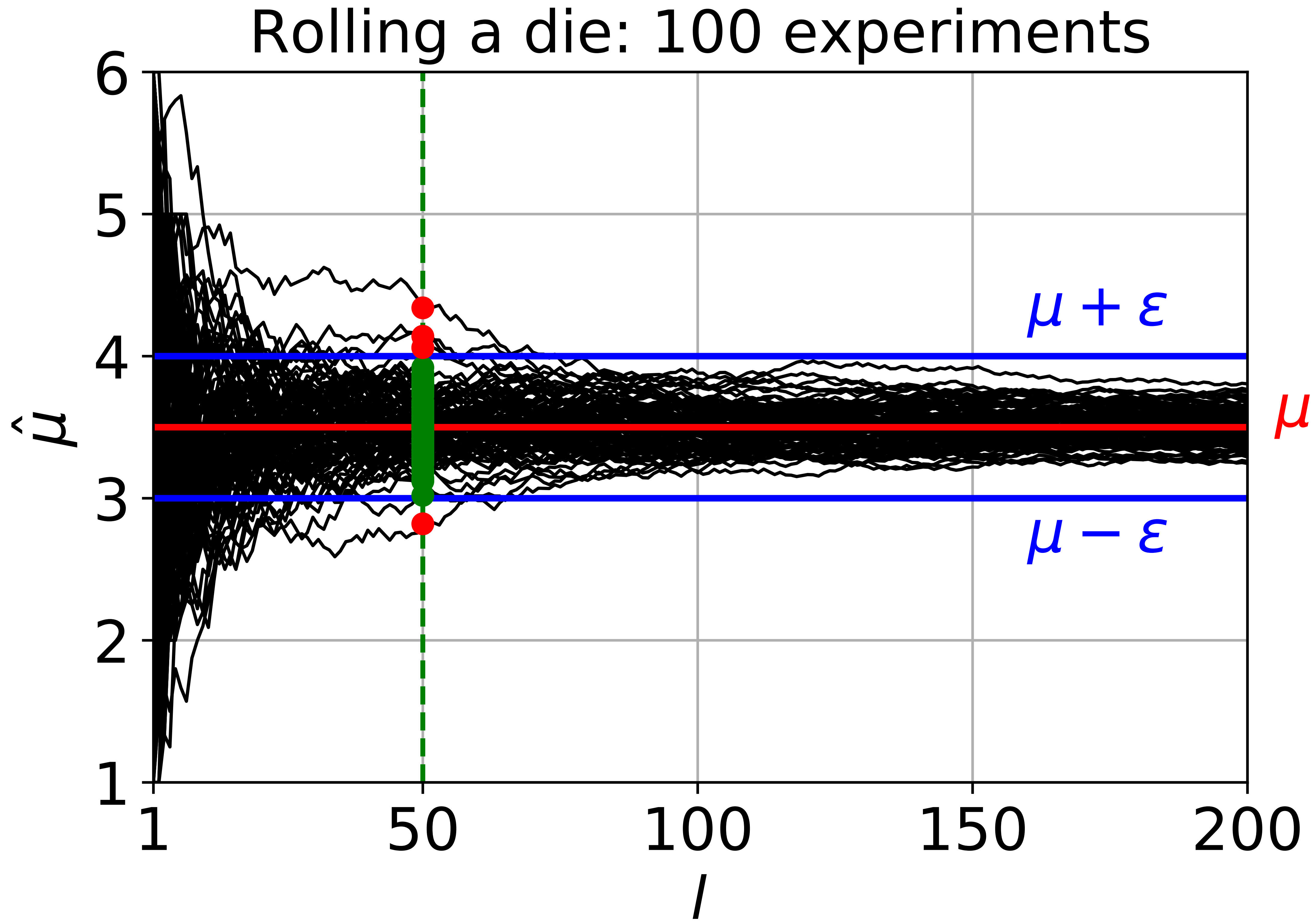

Rolling a die: 5 experiments
6
5
4
3
2
1
$\hat{\mu}$
1
50
100
150
200
l



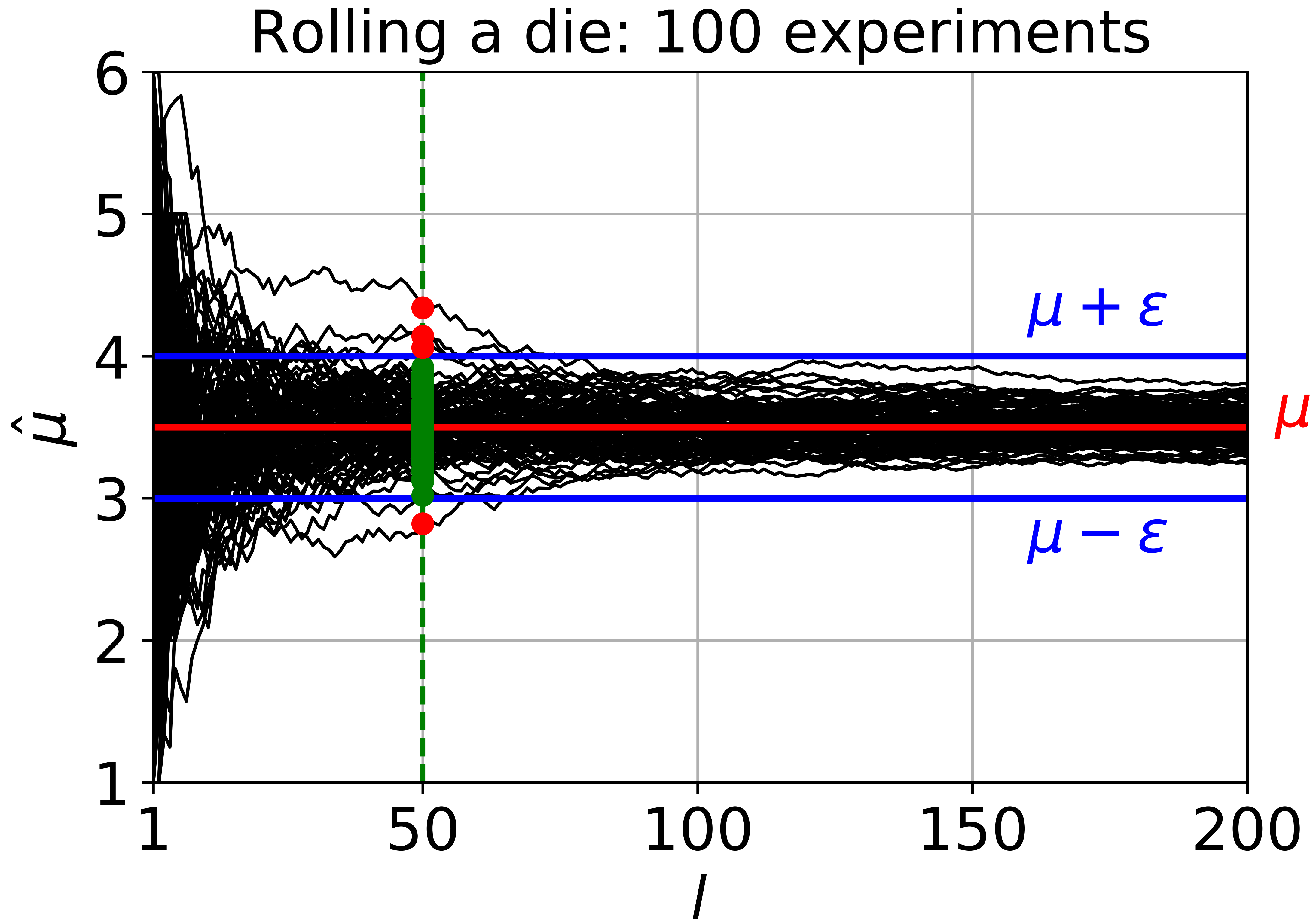
Rolling a die: 5 experiments
$\hat{\mu}$
6
5
4
3
2
1
1
50
100
150
200
l
$\mu + \varepsilon$
μ
$\mu - \varepsilon$

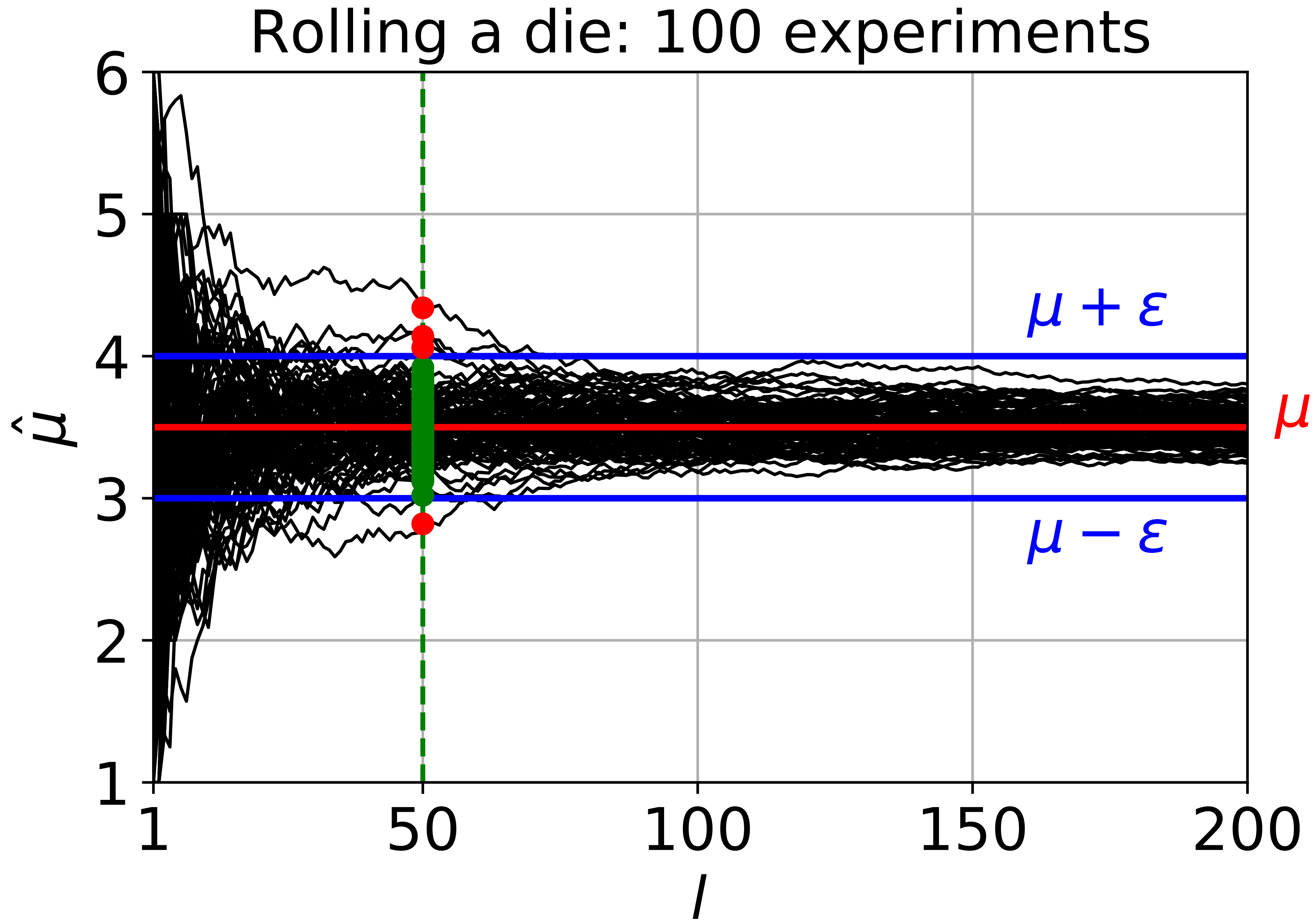


Rolling a die: 5 experiments
6
5
4
3
2
1
$\hat{\mu}$
$\mu + \varepsilon$
μ
$\mu - \varepsilon$
1
50
100
150
200
l

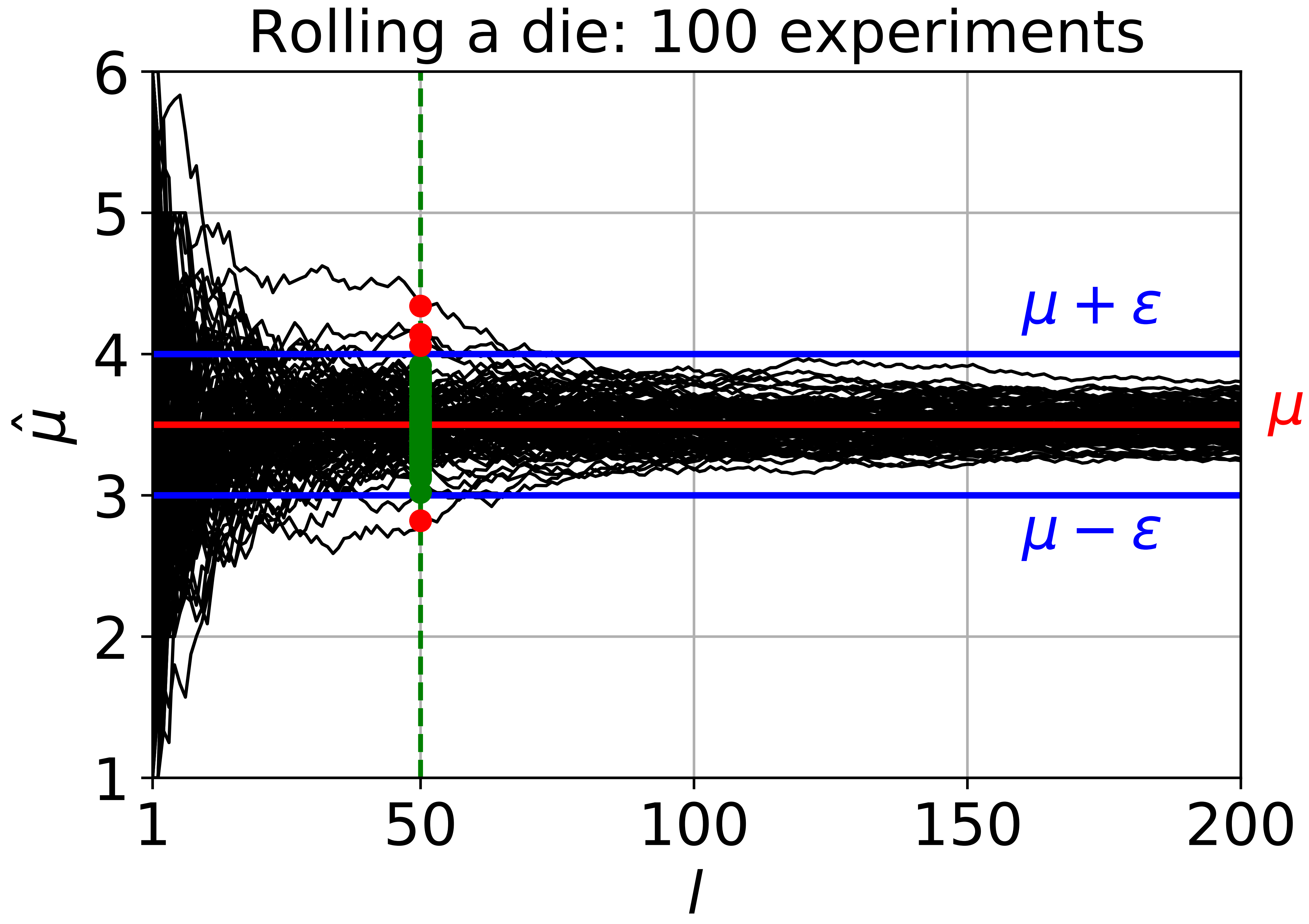


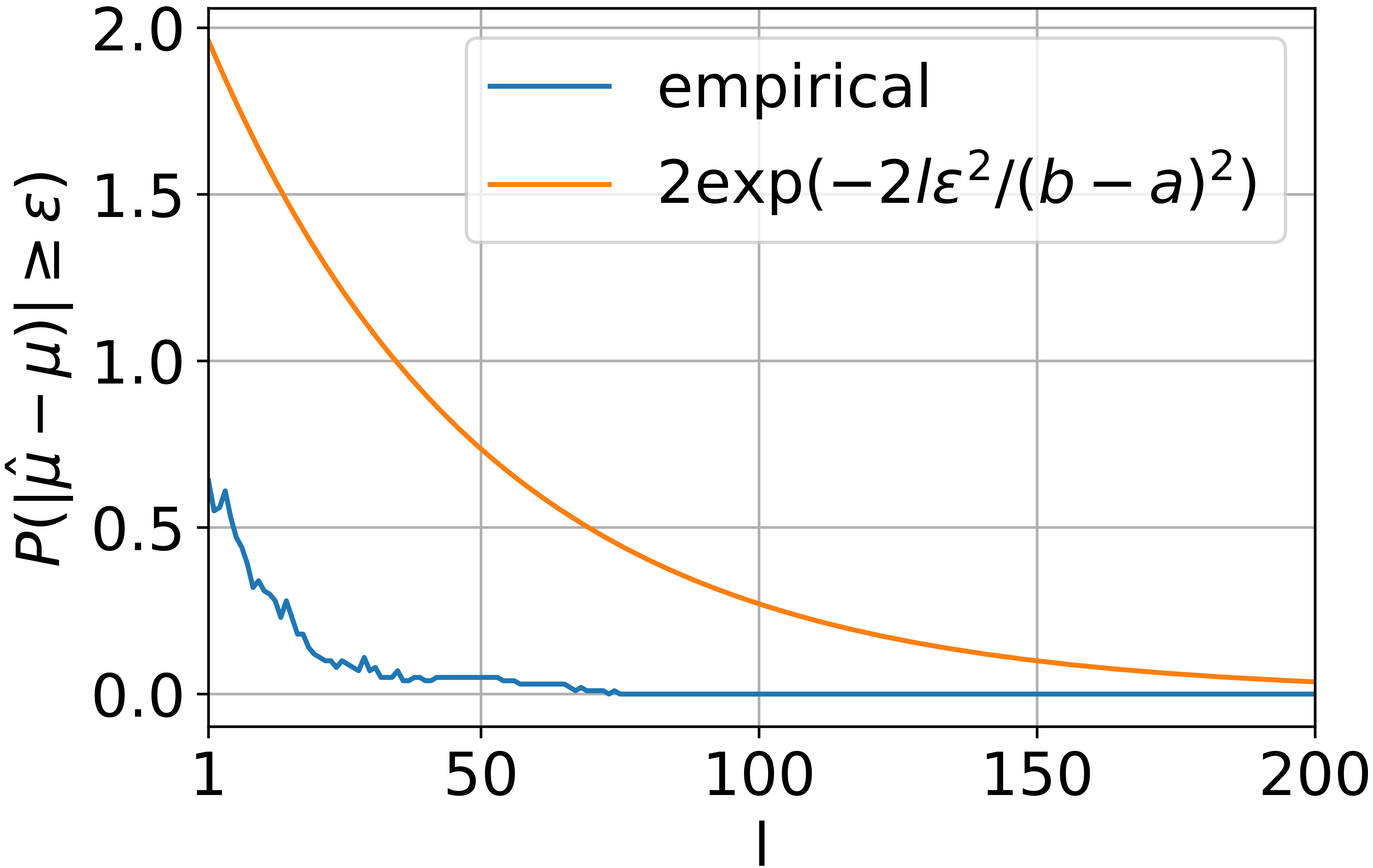
Rolling a die: 100 experiments
6
5
4
3
2
1
$\hat{\mu}$
1
50
100
150
200
l
$\mu + \varepsilon$
μ
$\mu - \varepsilon$


Rolling a die: 100 experiments
6
5
4
3
2
1
1
50
100
150
200
$\hat{\mu}$
l
$\mu + \varepsilon$
μ
$\mu - \varepsilon$

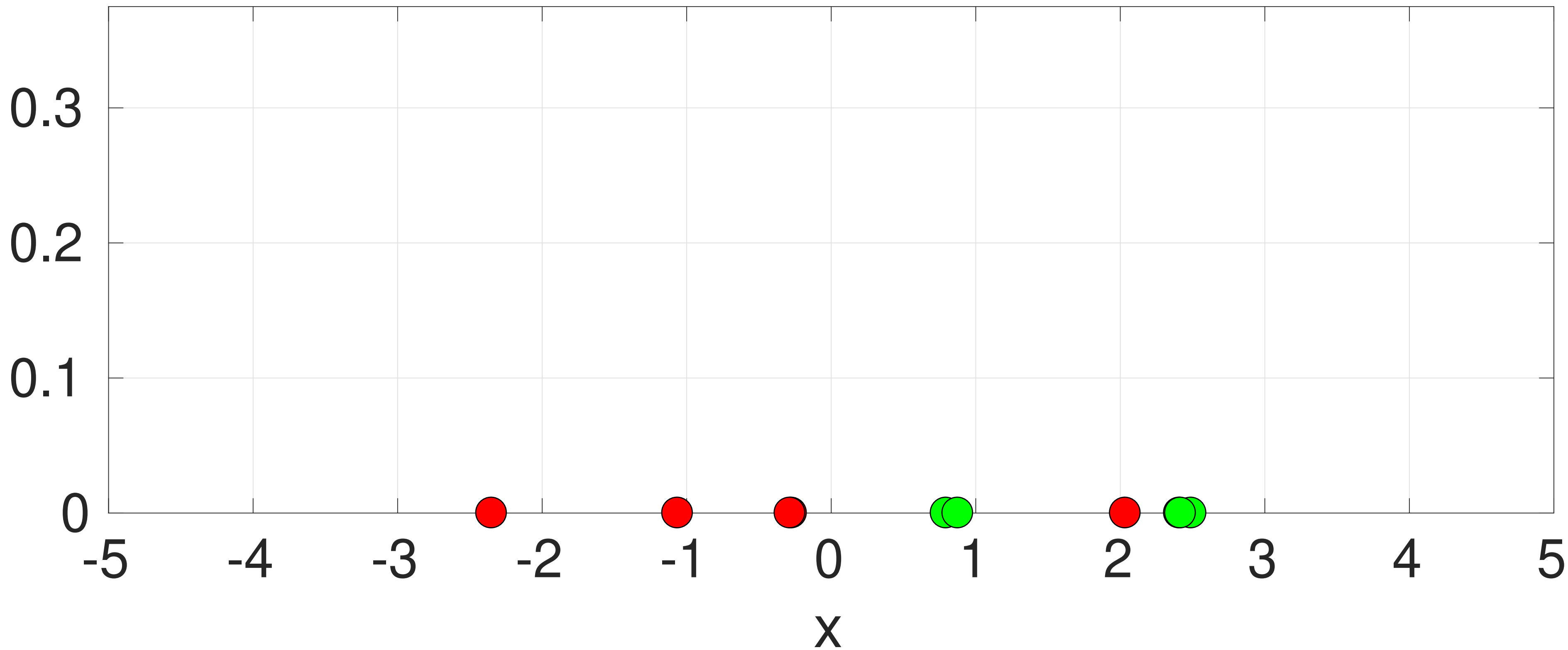


Rolling a die: 100 experiments
6
5
4
3
2
1
1
50
100
150
200
l
$\hat{\mu}$
$\mu + \varepsilon$
μ
$\mu - \varepsilon$

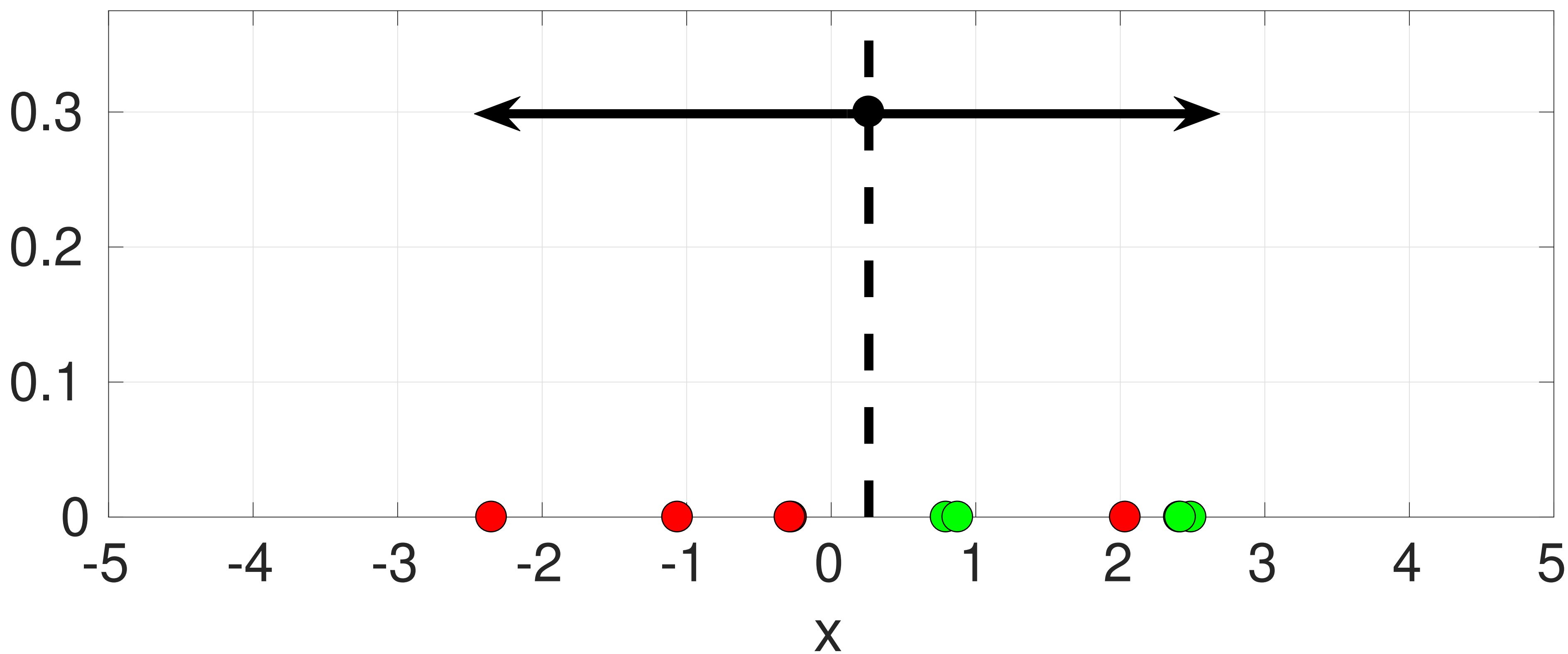

Rolling a die: 100 experiments
6
5
4
3
2
1
1
50
100
150
200
l
$\hat{\mu}$
$\mu + \varepsilon$
μ
$\mu - \varepsilon$



empirical
$2\exp(-2l\varepsilon^2/(b-a)^2)$
$P(|\hat{\mu}-\mu)| \geq \varepsilon)$
2.0
1.5
1.0
0.5
0.0
1
50
100
150
200
l



0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
x


0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
x