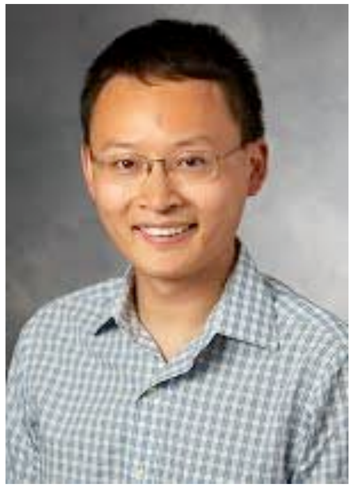
Adaptive Monte Carlo Multiple Testing via Multi-Armed Bandits

Martin Zhang

joint work with:

David Tse & James Zou
Stanford University

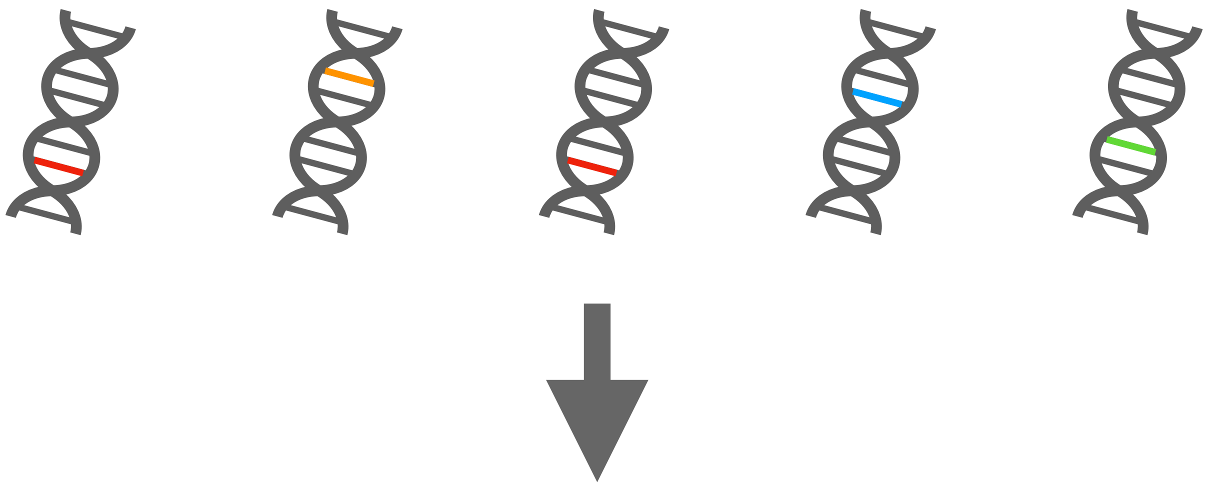


Problem | Monte Carlo Multiple Hypothesis Testing

SNP 1 SNP 2 $\cdots$ SNP m

Problem | Monte Carlo Multiple Hypothesis Testing

SNP 1 SNP 2 $\cdots$ SNP m



P_1 P_2 $\cdots$ P_m

Problem | Monte Carlo Multiple Hypothesis Testing

SNP 1 SNP 2 $\cdots$ SNP m

P_1 P_2 $\cdots$ P_m

$\times$ $\surd$ $\cdots$ $\times$

Problem | Monte Carlo Multiple Hypothesis Testing

SNP 1 SNP 2 $\cdots$ SNP m

Monte Carlo test

$$P_1 \sim \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\{T_{1,j}^{null} \geq t_1^{obs}\}$$

P_1 P_2 $\cdots$ P_m

$\times$ $\surd$ $\cdots$ $\times$

Problem | Monte Carlo Multiple Hypothesis Testing

SNP 1 SNP 2 $\cdots$ SNP m

Monte Carlo test

$$P_1 \sim \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\{T_{1,j}^{null} \geq t_1^{obs}\}$$

P_1 P_2 $\cdots$ P_m

Benjamini Hochberg procedure

$\times$ $\surd$ $\cdots$ $\times$

Data-dependent # of discoveries

Control $FDR = \mathbb{E}\left[\frac{false\ discovery}{discovery}\right]$

Problem | Monte Carlo Multiple Hypothesis Testing

SNP 1 SNP 2 $\cdots$ SNP m

Computational cost: nm

$$P_1 \sim \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\{T_{1,j}^{null} \geq t_1^{obs}\}$$

Benjamini Hochberg procedure

$\times$ $\surd$ $\cdots$ $\times$

Data-dependent # of discoveries

Control $FDR = \mathbb{E}\left[\frac{false\ discovery}{discovery}\right]$

Problem | Monte Carlo Multiple Hypothesis Testing

SNP 1 SNP 2 $\cdots$ SNP m

Computational cost: nm

m hypothesis tests

$\times$ n MC samples per test

$$P_1 \sim \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\{T_{1,j}^{null} \geq t_1^{obs}\}$$

Benjamini Hochberg procedure

$\times$ $\surd$ $\cdots$ $\times$

Data-dependent # of discoveries

Control $FDR = \mathbb{E}\left[\frac{false\ discovery}{discovery}\right]$

Problem | Monte Carlo Multiple Hypothesis Testing

Genome-wide association studies

$$m = 500{,}000 \qquad n = 50{,}000{,}000$$

m hypothesis tests
n MC samples per test

Problem | Monte Carlo Multiple Hypothesis Testing

Genome-wide association studies

$$m = 500{,}000 \qquad n = 50{,}000{,}000$$

Total MC samples: $nm = 2.5 \times 10^{13}$

Typical computation time: ~2 months

m hypothesis tests
n MC samples per test

Problem | Monte Carlo Multiple Hypothesis Testing

Genome-wide association studies

$$m = 500{,}000 \qquad n = 50{,}000{,}000$$

Total MC samples: $nm = 2.5 \times 10^{13}$

Typical computation time: ~2 months

Can we make it faster?

m	hypothesis tests
n	MC samples per test

Results | **A**daptive **M**onte Carlo Multiple **T**esting (AMT)

Theorem (informal):

Expected # of MC samples: $\sqrt{nm}$

baseline: nm

same discoveries with high probability; information theoretically optimal

Results | **A**daptive **M**onte Carlo Multiple **T**esting (AMT)

Theorem (informal):

Expected # of MC samples: $\sqrt{nm}$

baseline: nm

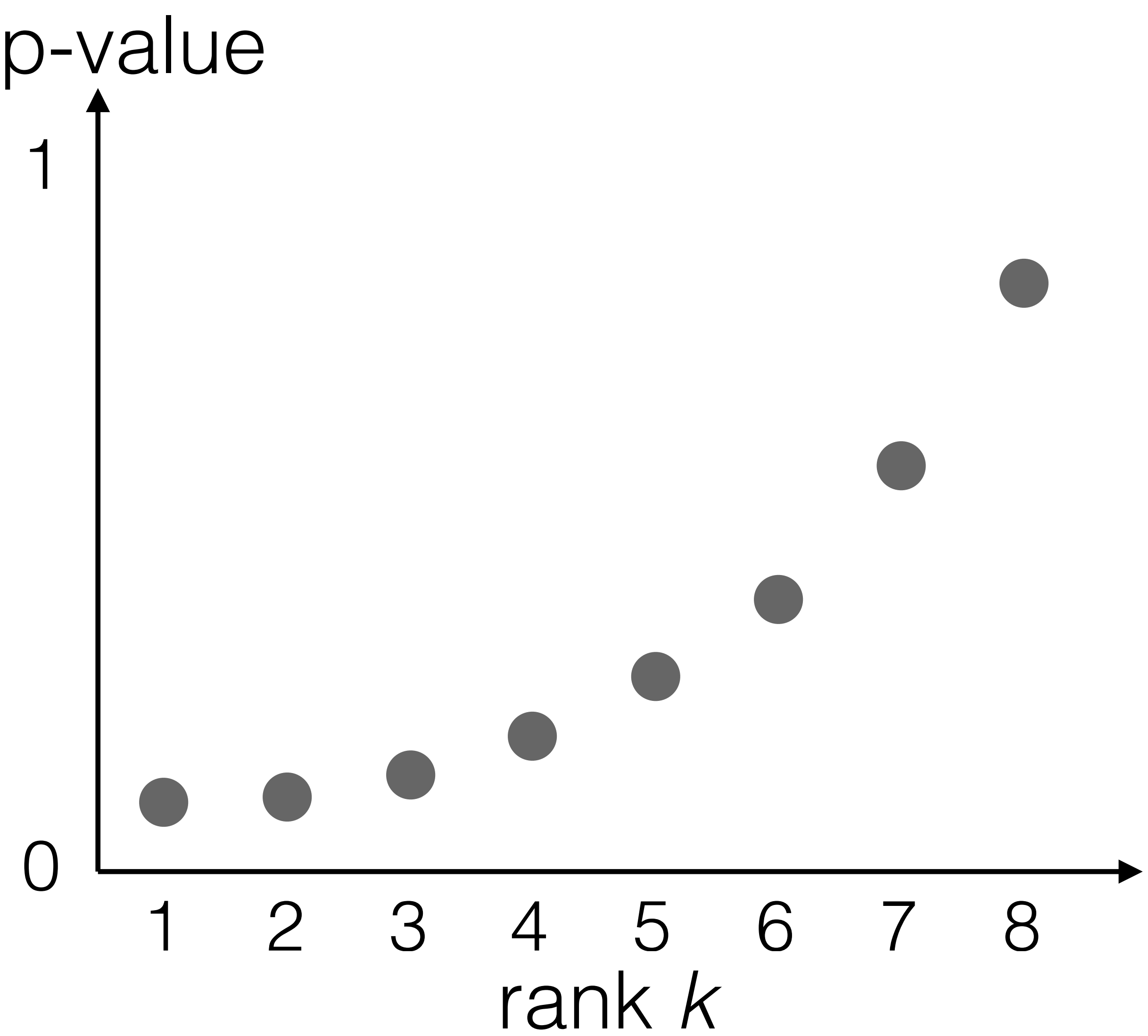
same discoveries with high probability; information theoretically optimal

GWAS example:

2 months $\longrightarrow$ 1 hour with the same discoveries

Results | **A**daptive **M**onte Carlo Multiple **T**esting (AMT)

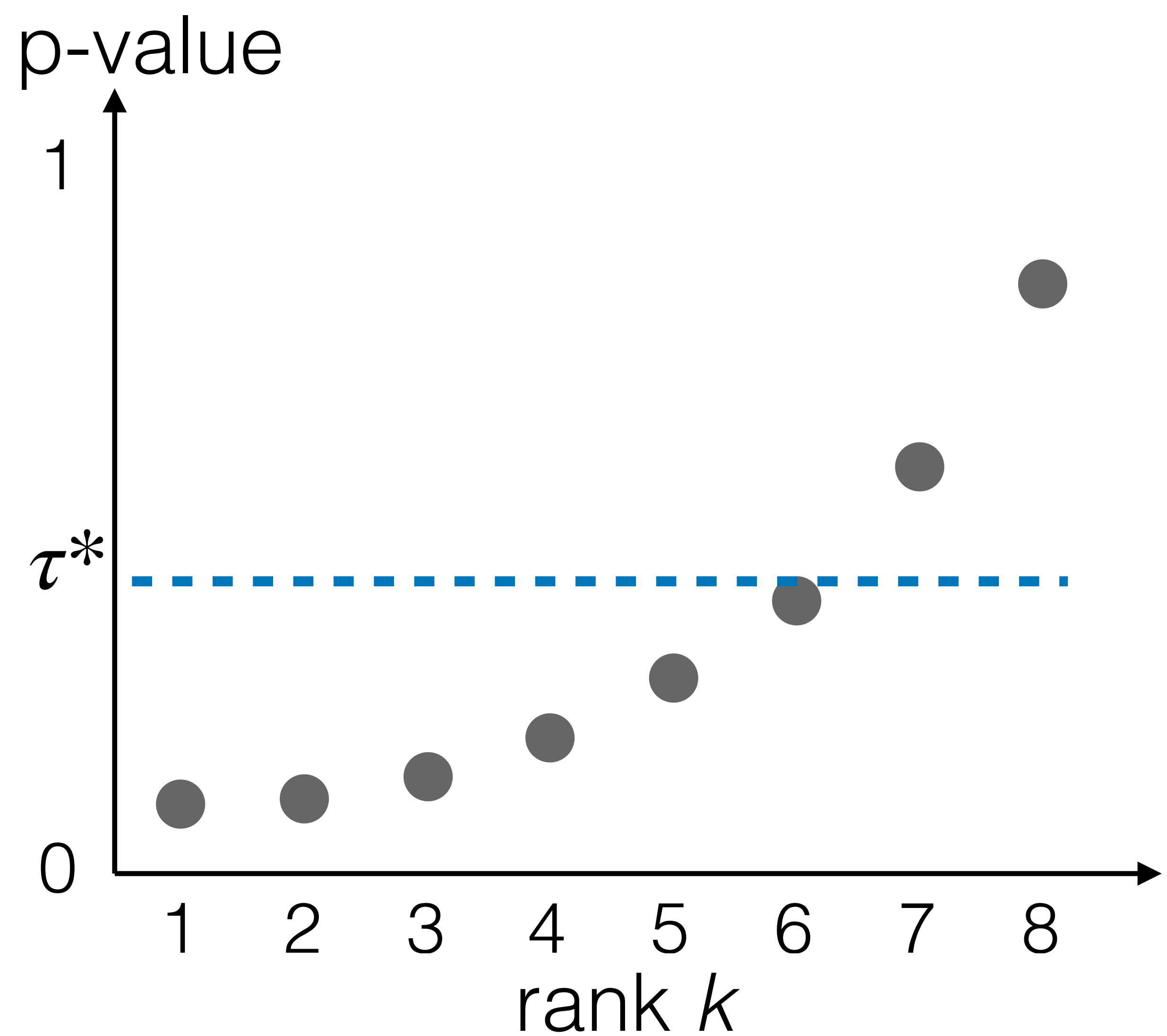
Quantities to estimate:



Results | **A**daptive **M**onte Carlo Multiple **T**esting (AMT)

Quantities to estimate:

BH threshold τ^*

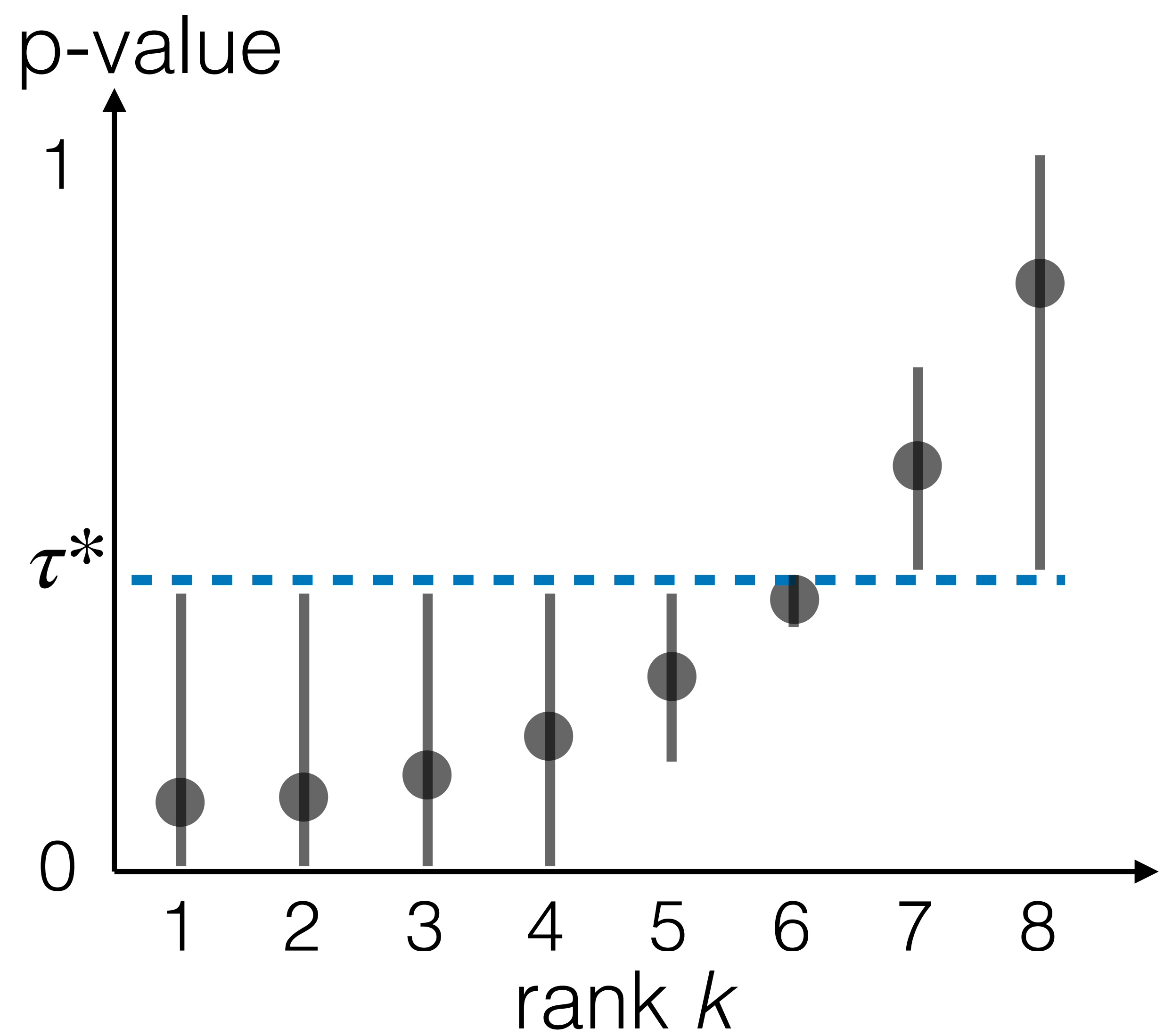


Results | **A**daptive **M**onte Carlo Multiple **T**esting (AMT)

Quantities to estimate:

BH threshold τ^*

How each p-value compares with τ^*

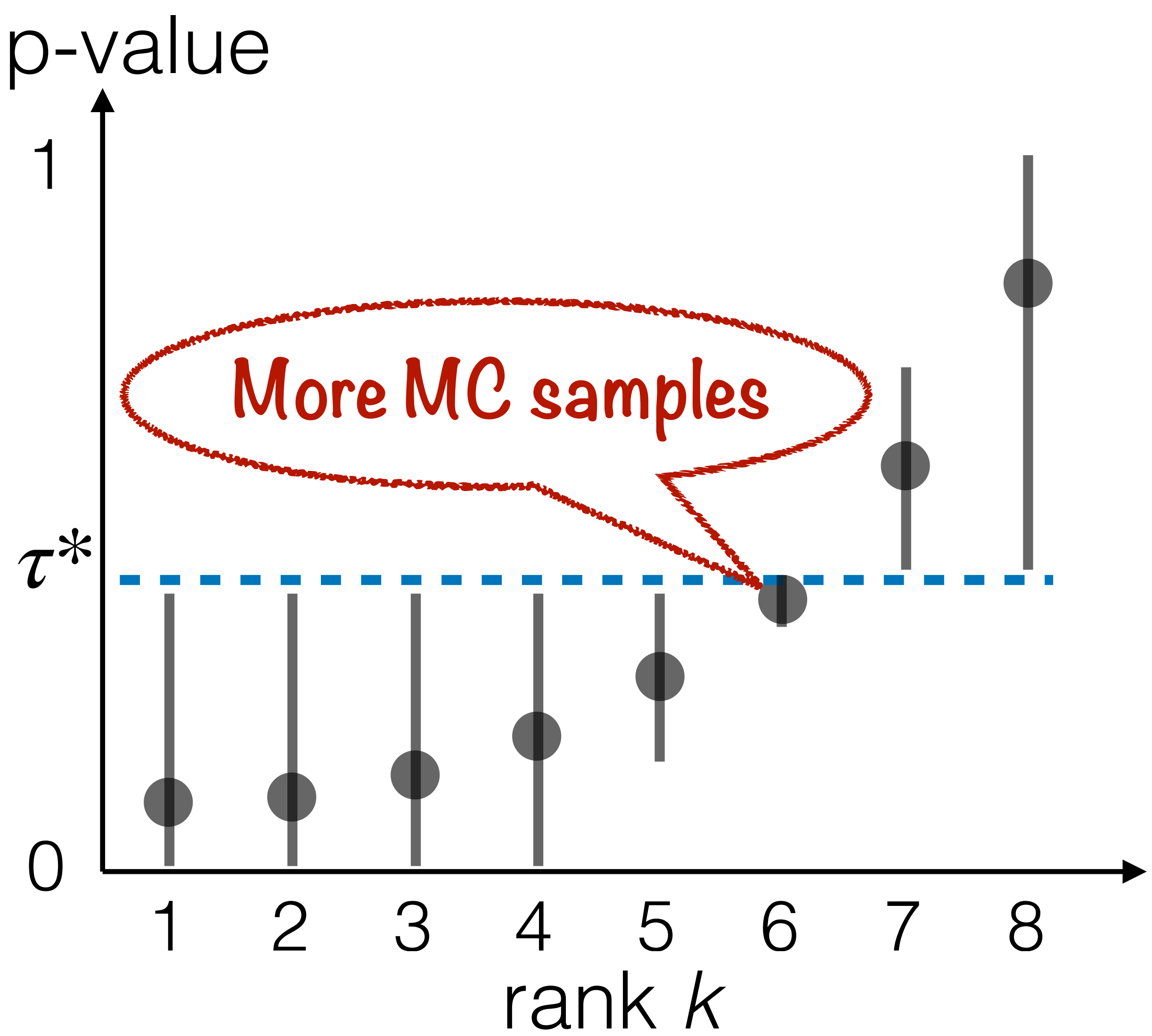


Results | **A**daptive **M**onte Carlo Multiple **T**esting (AMT)

Quantities to estimate:

BH threshold τ^*

How each p-value compares with τ^*

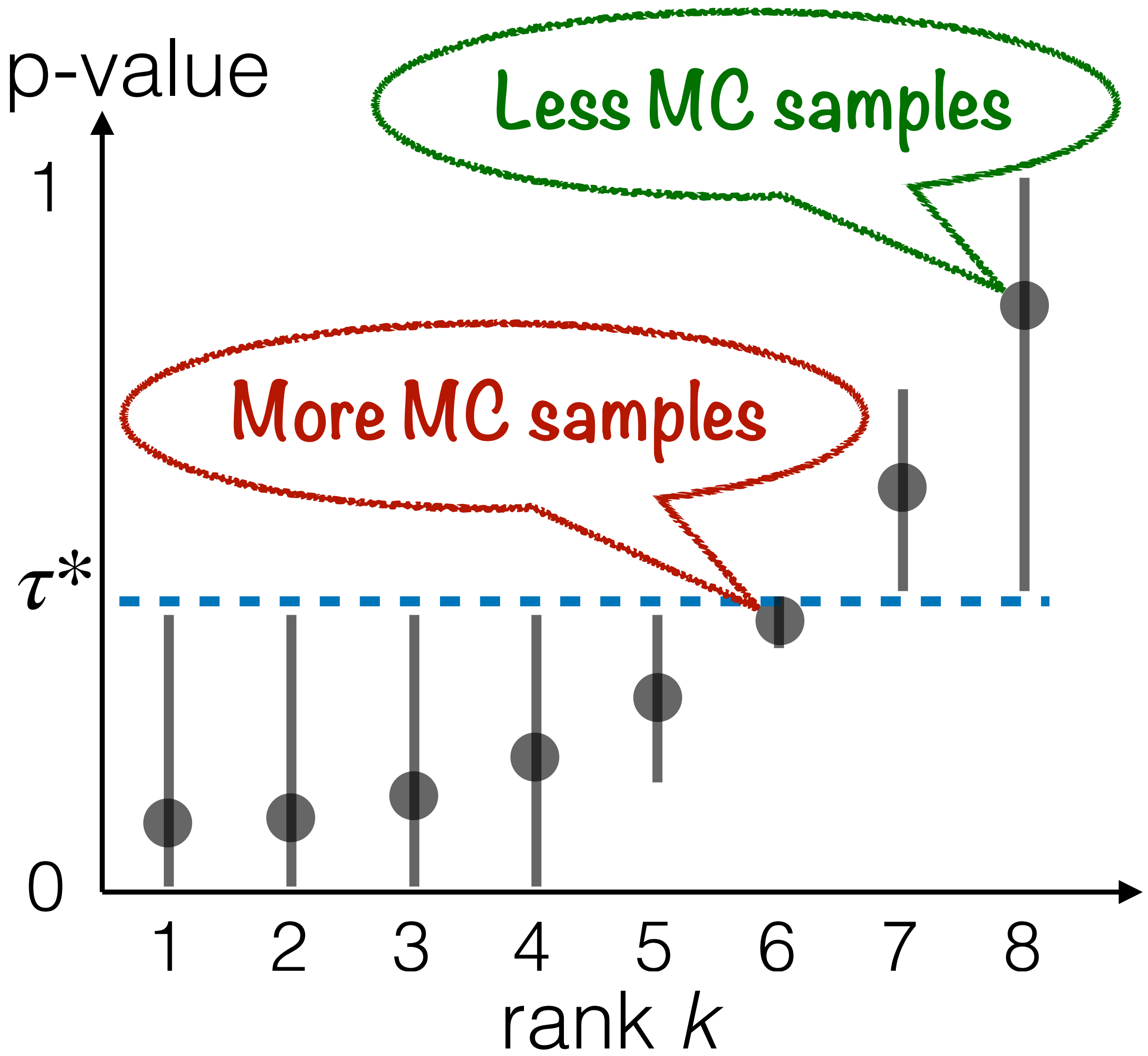


Results | **A**daptive **M**onte Carlo Multiple **T**esting (AMT)

Quantities to estimate:

BH threshold τ^*

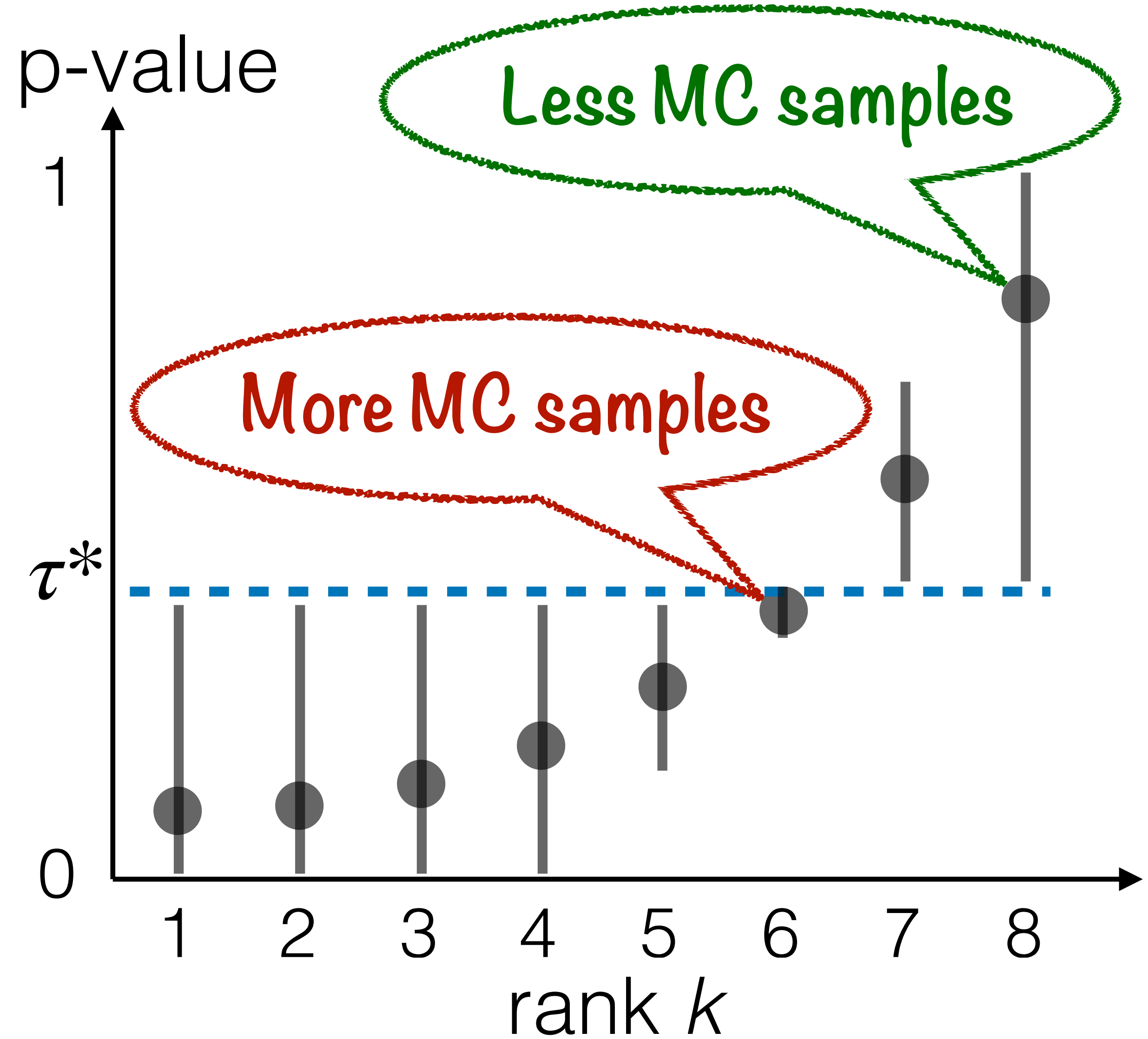
How each p-value compares with τ^*



Results | **A**daptive **M**onte Carlo Multiple **T**esting (AMT)

Quantities to estimate:

- BH threshold τ^*
- How each p-value compares with τ^*



Adaptive Estimation via Multi-Armed Bandits