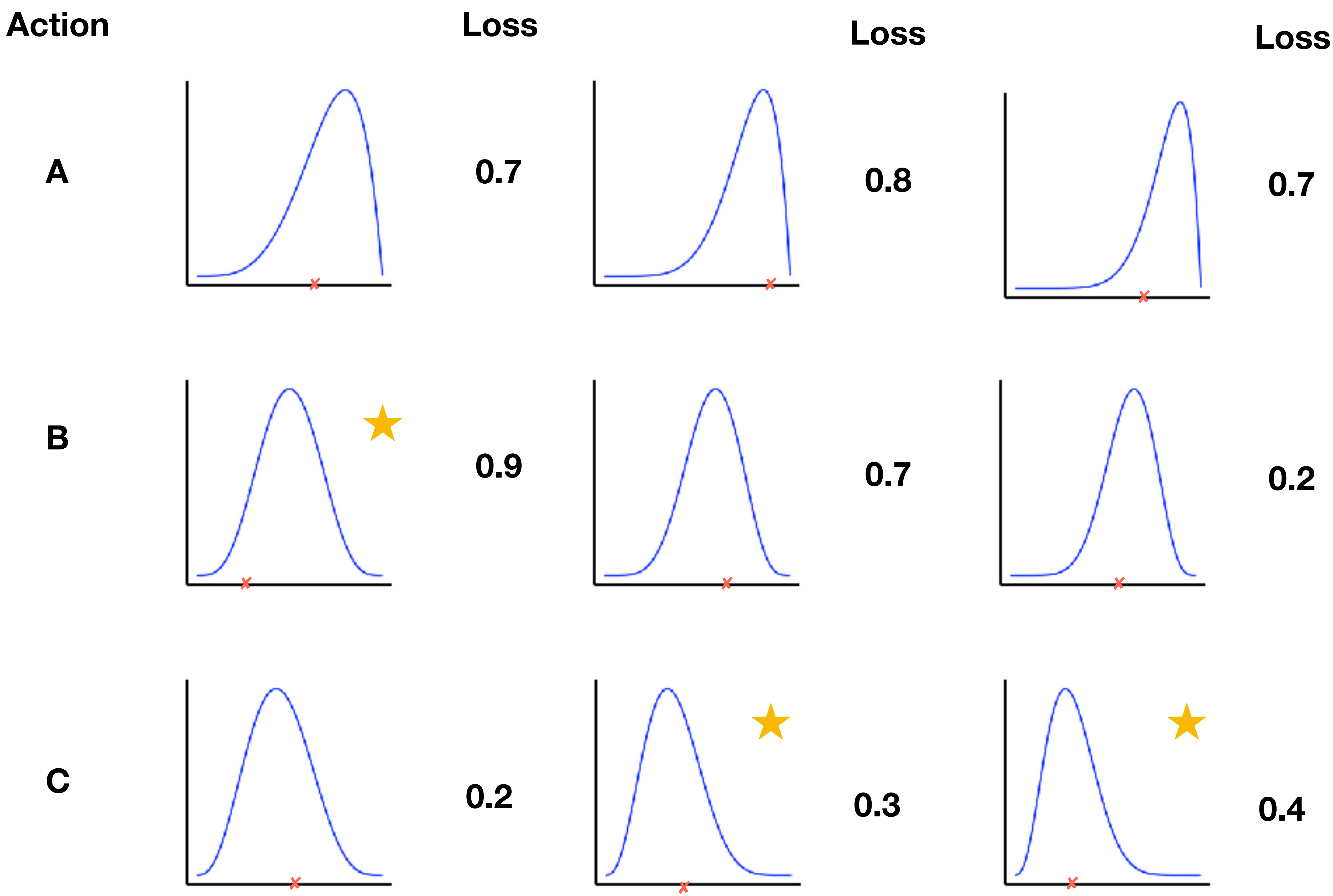
Thompson Sampling in Adversarial Environments



Total Loss = 0.9 + 0.3 + 0.4 = 1.6

Loss of Best Fixed Action = 0.2 + 0.3 + 0.4 = 0.9

Regret = 1.6 - 0.9 = 0.7

Thompson Sampling

Beta Bernoulli

$$\underbrace{P(\theta_{i,t}|\ell_{i,t})}_{beta}^{posterior} = \underbrace{P(\ell_{i,t}|\theta_{i,t})}_{bernoulli}^{liklihood} \underbrace{P(\theta_{i,t})}_{beta}^{prior}$$

Algorithm:

Set parameters for each action $\alpha_i = 1$, $\beta_i = 1$

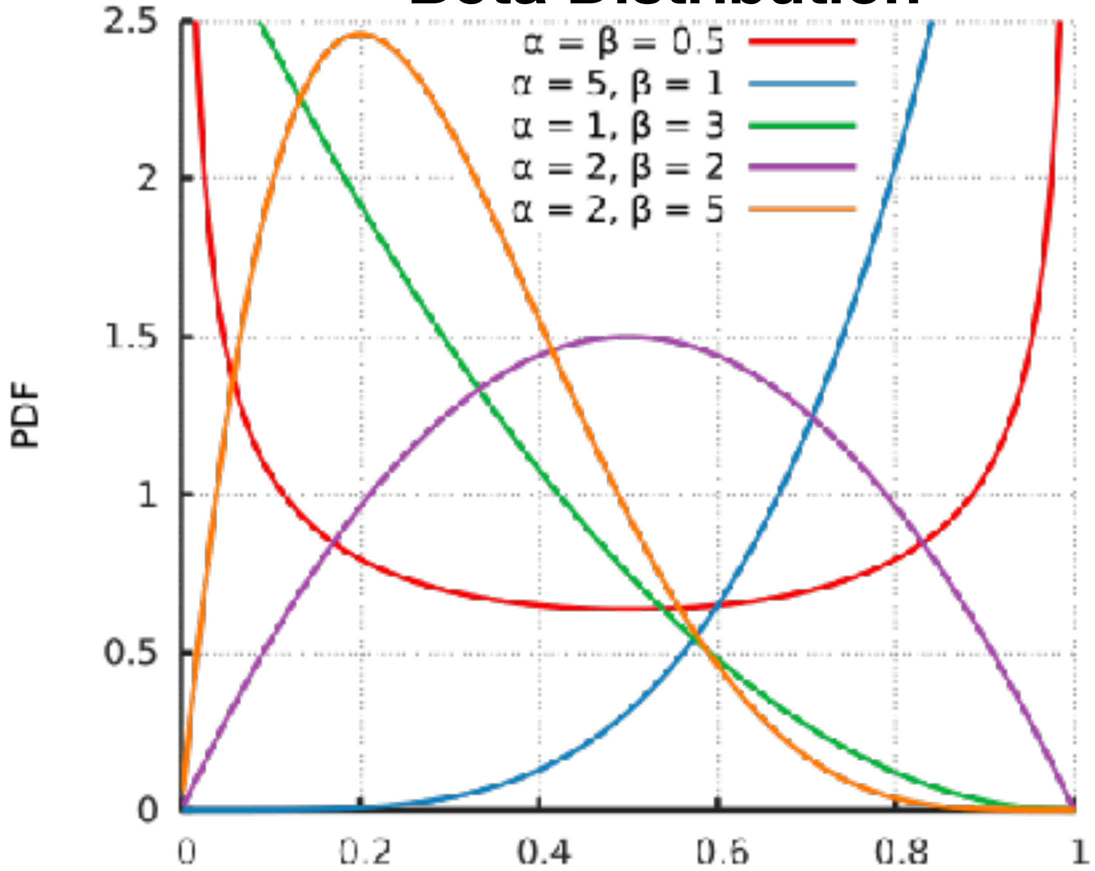
For each timestep $t = 1, 2, ..., T$

1. Sample from $\mathrm{Beta}(\alpha_i, \beta_i)$ for every action $i \in \{1, ..., N\}$ and chose the minimum
2. Observe losses $\ell_{i,t}$ for every action
3. Perform Bernoulli trial $\widetilde{\ell}_{i,t} \sim \mathrm{Bernoulli}(\ell_{i,t})$
4. Update parameters:

$$\alpha_i = \alpha_i + \widetilde{\ell}_{i,t}$$

$$\beta_i = \beta_i + 1 - \widetilde{\ell}_{i,t}$$

Beta Distribution



Other Algorithms

Exponential Weighted Average

$$P(a_t = i) = \frac{e^{-\eta_t L_{i,t-1}}}{\sum_{j=1}^{N} e^{-\eta_t L_{j,t-1}}}$$

Algorithm	Type	Adversarial Regret Bound
Exponential Weighted Average	$\eta_t = \sqrt{8(\ln N)/t}$	$\mathcal{O}(\sqrt{T \log N} + \log N)$
	$\eta = \sqrt{8(\ln N)/T}$	$\mathcal{O}(\sqrt{T \log N})$
	$\eta_t = \min\{1, C\sqrt{(\ln N)/\mathrm{Var}(\widehat{L}_t)}\}$	$\mathcal{O}(\sqrt{\mathrm{Var}(\widehat{L}_T) \log N} + \log N)$
	AdaHedge	$\mathcal{O}(\sqrt{L^* \log N} + \log N)$

Other Algorithms

Follow the Perturbed Leader

$$a_t = argmin_i L_{i,t-1} + Z_{i,t}$$

Algorithm	Type	Adversarial Regret Bound
Follow the Perturbed Leader	Uniform	$\mathcal{O}(\sqrt{TN})$
	Random Walk	$\mathcal{O}(\sqrt{T\ \log N} + \log T)$
	Exponential	$\mathcal{O}(\sqrt{L^*\ \log N} + \log N)$
	Dropout	$\mathcal{O}(\sqrt{L^*\ \log N} + \log N)$

How does Thompson Sampling compare?

If losses of each arm are independent and identically distributed between 0 and 1 the regret of Beta-Binomial Thompson Sampling is bounded in: $\mathcal{O}(\ln T)$

But what about adversarial losses?

goal: $\max_{\ell} E[R_T]$

where: $\ell \in [0,1]^{N \times T}$

Constant Sum Games

Payoff Matrix: $P \in [0,1]^{N \times N}$

Row Player v.s. Column Player

For every round:

Row player chooses i
Column player chooses j
Row player incurs loss $P_{i,j}$
Column player incurs loss $1 - P_{I,j}$

Let's duel!

identity

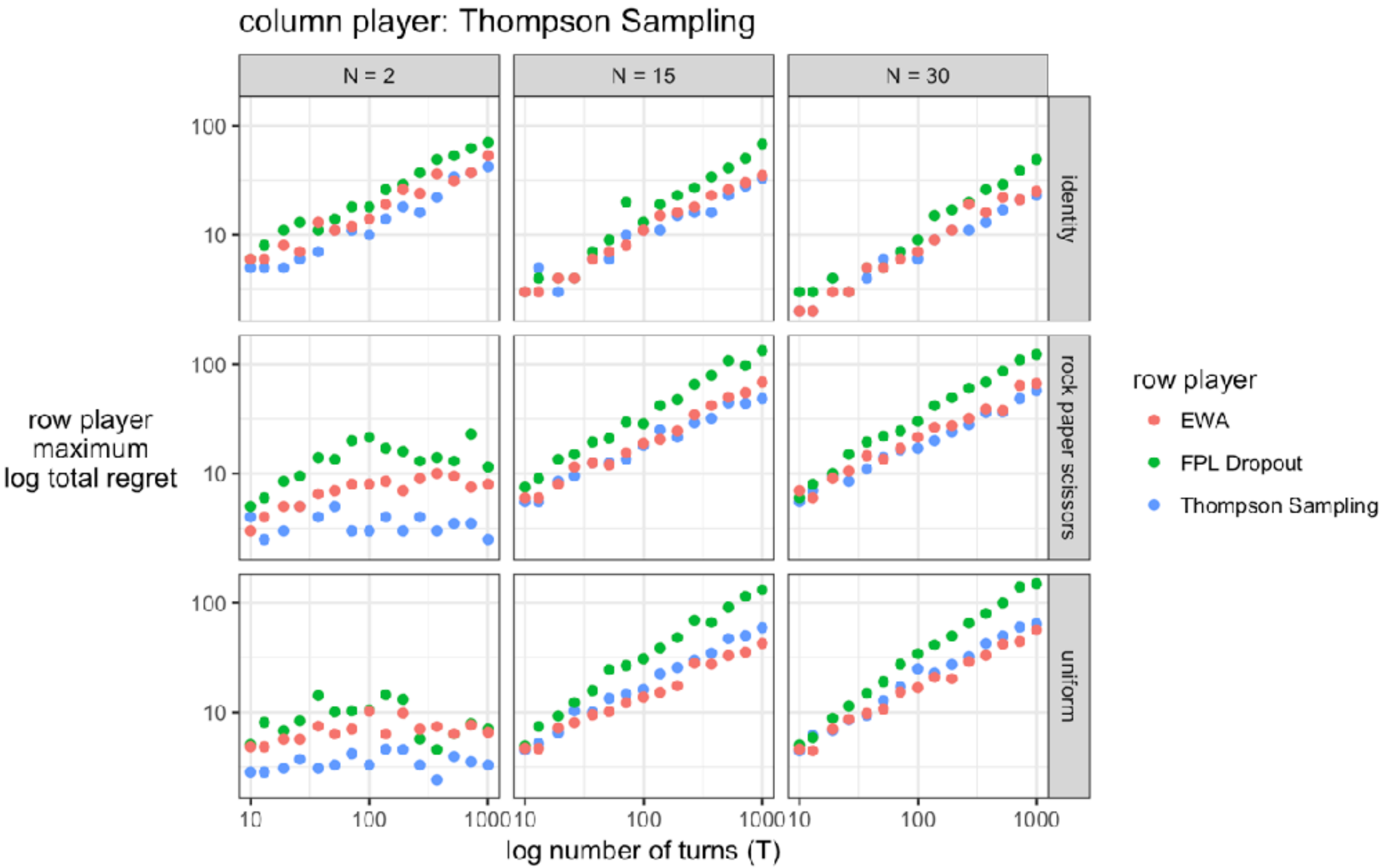
$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

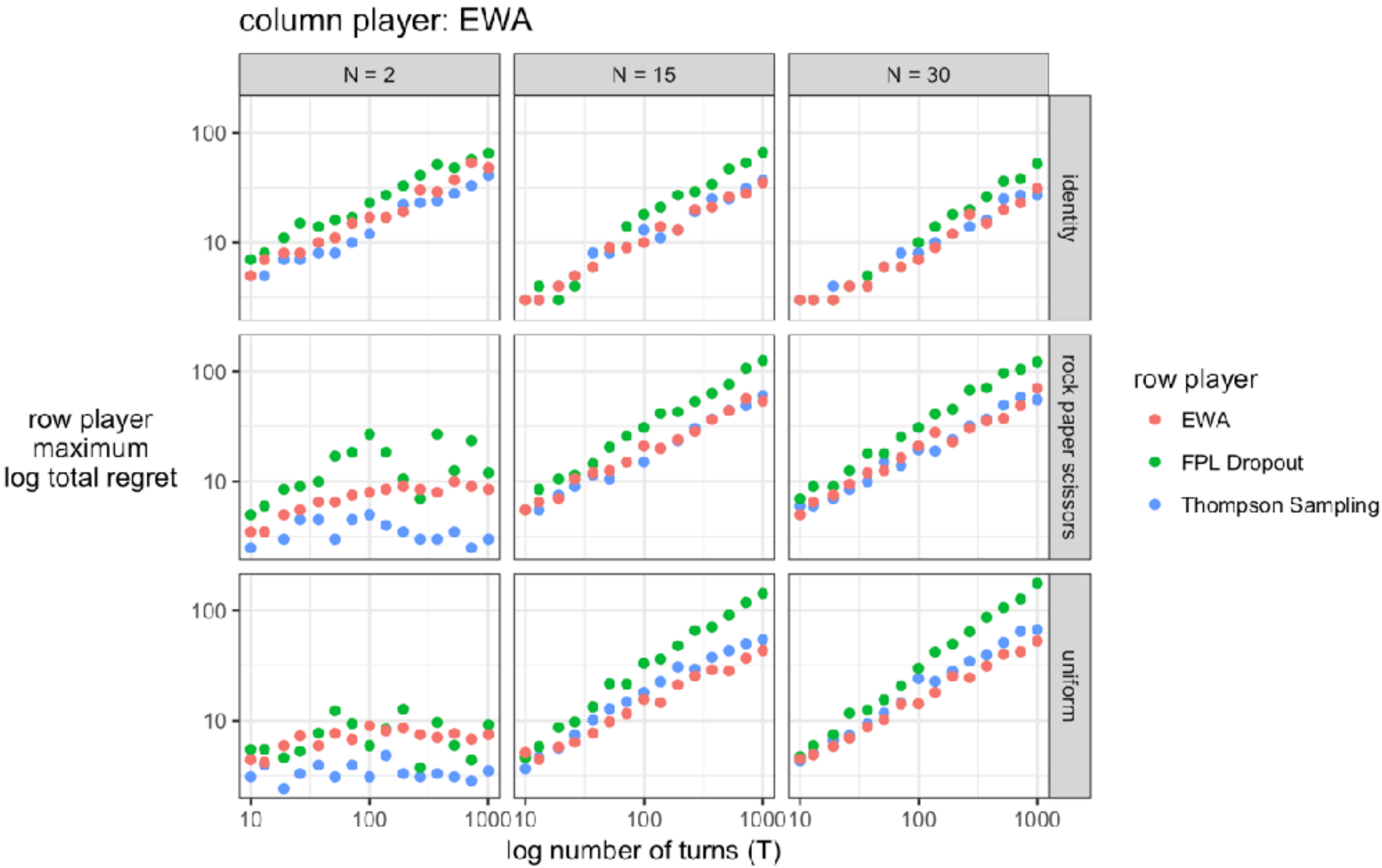
rock paper scissors

$$\begin{bmatrix} 0.5 & 0 & 1 \\ 1 & 0.5 & 0 \\ 0 & 1 & 0.5 \end{bmatrix}$$

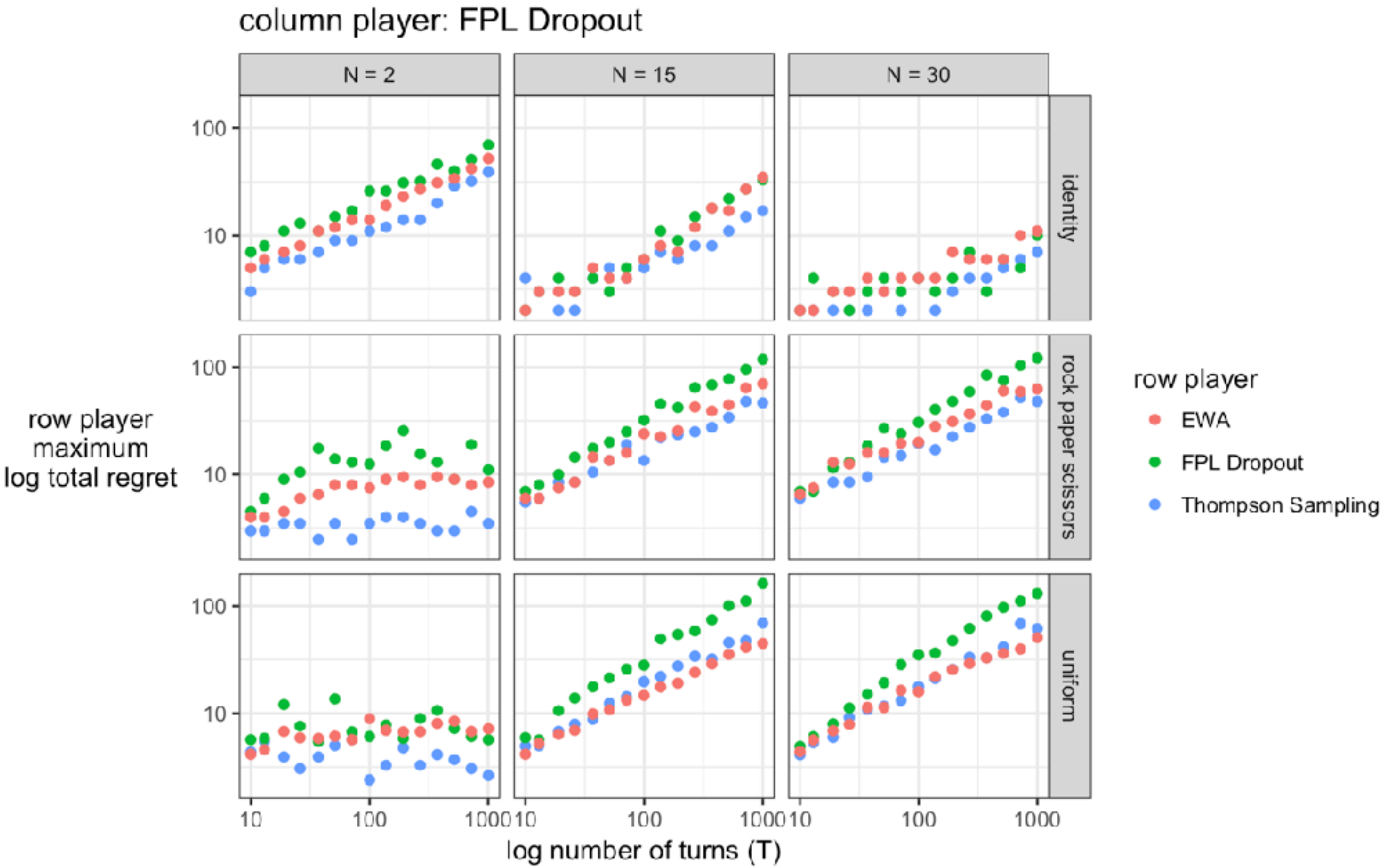
uniform

$$\begin{bmatrix} 0.9 & 0.8 & 0.8 \\ 0.5 & 0.6 & 0.4 \\ 0.3 & 0.4 & 0.6 \end{bmatrix}$$


column player: Thompson Sampling
N = 2
N = 15
N = 30
identity
rock paper scissors
uniform
row player maximum log total regret
log number of turns (T)
row player
EWA
FPL Dropout
Thompson Sampling
10
100
1000



column player: EWA
N = 2
N = 15
N = 30
identity
rock paper scissors
uniform
row player maximum log total regret
log number of turns (T)
10
100
1000
row player
EWA
FPL Dropout
Thompson Sampling


column player: FPL Dropout
N = 2
N = 15
N = 30
identity
rock paper scissors
uniform
row player
maximum
log total regret
100
10
10
100
1000
log number of turns (T)
row player
EWA
FPL Dropout
Thompson Sampling

Evolutionary Methods

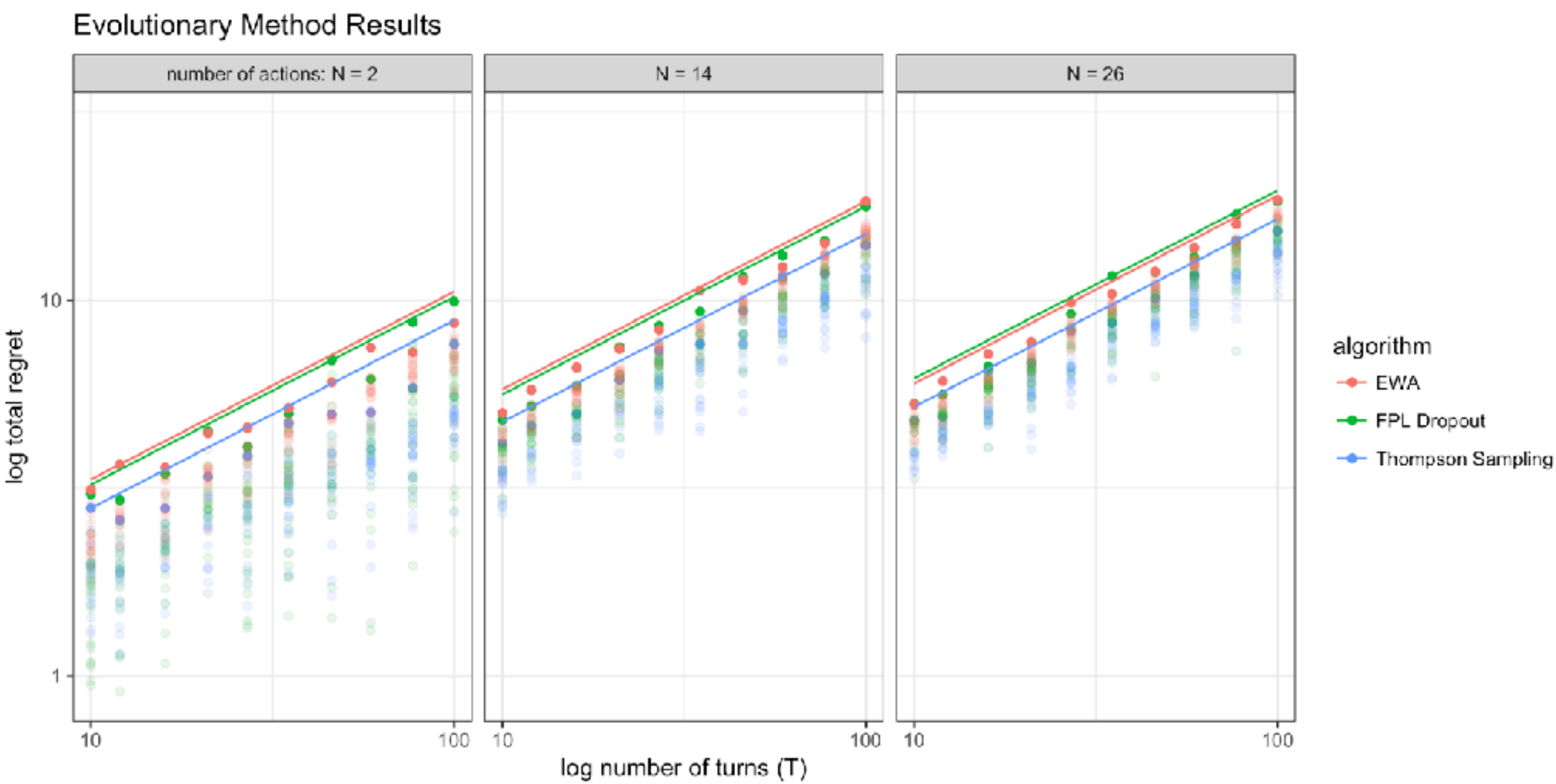
goal: $\max_\ell E[R_T]$

where: $\ell \in [0,1]^{N \times T}$

Vocabulary:
'individual' = loss matrix: $\ell \in [0,1]^{N \times T}$
'population' = group of 100 individuals

Algorithm:
Initialize population uniformly at random
For each generation:
- Estimate regret of each individual
- Remove two thirds individuals with lowest regret
- Remaining individuals have two children each
- Repeat

Note: we maintain separate populations for each algorithm and values of N and T



Evolutionary Method Results
number of actions: N = 2
N = 14
N = 26
log total regret
10
1
log number of turns (T)
10
100
algorithm
EWA
FPL Dropout
Thompson Sampling