Mitigating the Curse of Dimensionality of the Bayesian Beta-Bernoulli Bandit Problem

Peter Jacko*

Workshop on Multi-Armed Bandits and
Learning Algorithms, Rotterdam, 24-25 May 2018

*Dept. of Management Science, Lancaster University, UK

Multi-disciplinary Bandits

- Different terminology across disciplines

Anecdotic	strategy	choice	pull	arms
OR	policy	allocation	resource	projects
CS/ML	algorithm	decision	time step	actions
Biometrics	design	randomisation	patient	treatments
Telecom	scheduler	allocation	server	jobs
Universal (?)	design	allocation	subject	interventions

Clinical Trials

- The gold standard design: randomised controlled trial
  - 50% vs 50% fixed equal randomisation
  - avoids all types of biases
  - in use since 1948 (advocated since Hill 1937)
- Its main goal is to learn about intervention effectiveness with a view to prioritising future outside subjects
  - maximises power of an intervention effect difference
  - if approved, future subjects are, say, 95% confident that the novel intervention is better than the control
- A half of trial subjects gets the inferior intervention

Randomised Controlled Trial

- Statistical testing based on randomised equal allocation is a widespread state-of-the-art approach in the design of experiments, under different names:
  - randomised controlled trial in biostatistics
  - between-group design in social sciences
  - A/B testing in Internet marketing

Bayesian Decision-Theoretic Trial

"...there can be no objection to the use of data, however meagre, as a guide to action required before more can be collected ... Indeed, the fact that such objection can never be eliminated entirely—no matter how great the number of observations—suggested the possible value of seeking other modes of operation than that of taking a large number of observations before analysis or any attempt to direct our course... This would be important in cases where either the rate of accumulation of data is slow or the individuals treated are valuable, or both."

Bayesian Decision-Theoretic Trial

- Proposed in Thompson 1933 (pre-dates Hill 1937)
- The goal is to provide higher benefit to both in-trial subjects and after-trial subjects
  - ▷ as opposed to the RCT's learning goal of reliable intervention effect estimation
- It is done by deciding the allocation, i.e., the randomisation probabilities for every subject (or for a group of subjects)
  - ▷ response-adaptive: decisions based on the responses accumulated so far, i.e. Bayesian

Bayesian Decision-Theoretic Trial

- In theory, can be solved to optimality by decision theory
- In practice, optimal decisions are computed numerically
  - it is often believed tractable only for small trials
- Milestones IMHO
  - Thompson (Biometrika 1933)
  - Glazebrook (Biometrika 1978)
  - Gittins & Jones (Biometrika 1979)
  - Armitage (ISR 1985)
  - Cheng, Su & Berry (Biometrika 2003)
  - Berry (Nature 2006), Cheng & Berry (Biometrika 2007)
  - Villar, Bowden & Wason (Statistical Science 2015)

Health Benefit Approach

- Important because healing patients is the ultimate goal of new treatment development
- Bayesian decision-theoretic model
  - ▷ optimally solving learning/healing trade-off
  - ▷ both learning and healing takes place during the trial
- This kind of general problem became known as the multi-armed bandit problem

Multi-Armed Bandit Problem

Multi-armed Bandit Problem

- Addressing the learning/earning trade-off
- Studied in scientific disciplines including probability, statistics, biometrics, operational research, economics, marketing, econometrics, machine learning, computer simulation, computer science, and communications
- Many formulations: i.i.d., Markovian, Bayesian, etc.
- Many extensions, mainly in machine learning
- Appropriate model for trials: finite horizon
  - ▷ after the end of the horizon we will not be able to influence the allocation

Bayesian Bernoulli Bandit Model

- Finite horizon: n sequentially arriving subjects
- Two-armed: intervention A or B for each subject
- Binary endpoints: success (1) or failure (0)
- Let X_i and Y_i denote subject i's response from intervention A and B respectively (for $i = 1, ..., n$). Then,

$$X_i \sim \text{Bernoulli}(1,\ \theta_A) \ \text{ and } \ Y_i \sim \text{Bernoulli}(1,\ \theta_B),$$

where θ_A and θ_B are the unknown success probabilities of interventions A and B respectively

Bayesian Approach

- Beliefs $\widehat{\theta}_A$ and $\widehat{\theta}_B$ to be updated over the trial
- Prior Distribution: $\widehat{\theta}_A \sim \text{Beta}(a,\ b)$, $\widehat{\theta}_B \sim \text{Beta}(c,\ d)$ where we take $a = b = c = d = 1$ (uninformative)
- Posterior Distribution: After observing i (j) successes (failures) on intervention A, and k (l) successes (failures) on intervention B, the posterior distribution is represented by another Beta distribution (by conjugacy)

$$\widehat{\theta}_A \sim \text{Beta}(a+i, b+j),\ \widehat{\theta}_B \sim \text{Beta}(c+k, d+l)$$

DP Design

- We use dynamic programming (DP) to obtain an optimal adaptive intervention allocation sequence
- Optimal in the sense of maximising the expected total number of successes in the trial
- Specifically, we use backward induction algorithm
- Let $\mathcal{F}_m(i, j, k, l)$ be the expected total number of successes under an optimal policy after m subjects
- If $m = n$, there is nothing to do: $\mathcal{F}_n(i, j, k, l) = 0$ $\forall i, j, k, l$

Backward Induction

- If $m = n - 1$ (one subject left):

  1. If intervention A, we compute the expectation

$$\mathcal{F}^A_{n-1}(i,j,k,l) = \frac{i}{i+j} \cdot 1 + \frac{j}{i+j} \cdot 0$$

  2. If intervention B, we compute the expectation

$$\mathcal{F}^B_{n-1}(i,j,k,l) = \frac{k}{k+l} \cdot 1 + \frac{l}{k+l} \cdot 0$$

- We wish to choose the optimal allocation such that

$$\mathcal{F}_{n-1}(i,j,k,l) = \max\{\mathcal{F}^A_{n-1}(i,j,k,l),\ \mathcal{F}^B_{n-1}(i,j,k,l)\}$$

DP Design

- Problem? This design is not suitable to implement in practice because it is completely deterministic
- As a result, there is a risk of introducing bias into the trial through the intentional selection of subjects (selection bias)
- Therefore, we modify the DP design by forcing actions to be randomised
  - ▷ see also Cheng & Berry (Biometrika, 2007)
- Helps to maintain blinding and reduce the risk of bias

Forcing Randomised Actions

- Action 1: intervention A is allocated with probability p
- Action 2: intervention B is allocated with probability p
- The expected total number of successes under Action 1

$$\mathcal{V}_m^1(i,j,k,l) = p \cdot \mathcal{F}_m^A(i,j,k,l) + (1-p) \cdot \mathcal{F}_m^B(i,j,k,l)$$

- The objective function becomes

$$\mathcal{V}_m(i,j,k,l) = \max\left\{\mathcal{V}_m^1(i,j,k,l),\ \mathcal{V}_m^2(i,j,k,l)\right\}$$

- Lower selection bias, but lower controllability

Randomised Variant

- Problems? After running simulations, we found:
  - ▷ this design is very underpowered for high p
  - ▷ in some of the runs (only a few out of 10,000), all subjects were allocated to only one of the interventions
- This means we cannot be confident about the results
- ...we cannot calculate important performance measures
- Therefore, we lower-limit the number of observations on each intervention

Constrained Variant

- We modify the optimal randomised DP policy by adding a constraint to ensure that we obtain $\geq \ell$ observations from each intervention
- To do this, we assign a large penalty to every terminal state that has $< \ell$ observations on any intervention
- The undesirable states will now be avoided (as much as possible) by the optimal policy
- We tried a range of values for ℓ, i.e. $0.05n$, $0.10n$, $0.15n$, $0.20n$ and $0.25n$. (Note that $0.50n$ corresponds to equal, fixed randomisation)

Simulation Study

- We evaluate the performance of proposed designs by
  - Bias of the intervention effect estimator
  - ...and its mean squared error (MSE)
  - Statistical power
  - Expected proportion of successes (EPS)
  - Subjects allocated to the superior arm (On sup)
- For each configuration, we replicate 10,000 trials

Simulation Results: Randomised Variant

Example. $n = 75, \theta_A = 0.2, \theta_B = 0.6$

p	Bias	MSE	Power	EPS	On sup
50%	0.001	0.004	0.938	40.0%	50.0%
60%	0.002	0.005	0.935	43.7%	59.1%
70%	0.002	0.007	0.910	47.3%	68.2%
80%	0.005	0.009	0.830	50.9%	77.3%
90%	0.015	0.015	0.636	54.4%	86.0%
100%	0.089	0.030	0.070	57.7%	94.2%

- The Power (almost) does not change if p increased from 50% to 60% or 70%. Room for increasing EPS!

Simulation Results: Constrained Randomised Variant

Example. $n = 75, \ell = 0.15n, \theta_A = 0.2, \theta_B = 0.6$

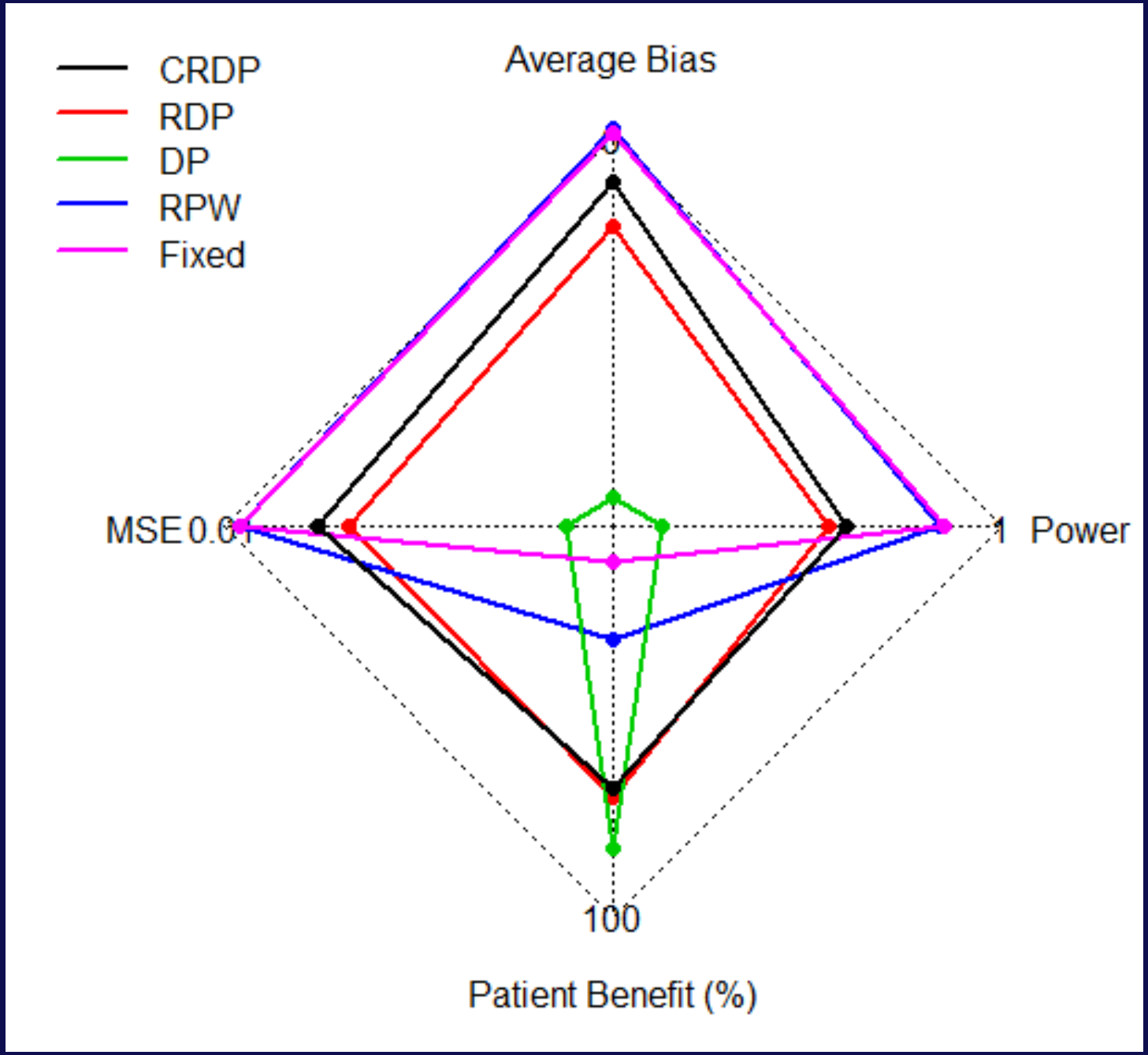
p	Bias	MSE	Power	EPS	On sup
60%	0.002	0.005	0.935	43.7%	59.1%
70%	0.002	0.007	0.910	47.3%	68.2%
80%	0.005	0.009	0.834	50.9%	77.2%
90%	0.008	0.013	0.724	53.6%	84.0%

- The Power is quite high even if p increased to 80% or 90%. Also bias diminishes!

Simulation Study

- We compare our proposed constrained randomised variant of DP (CRDP) design to the following designs:
  - ▷ Fixed randomisation (RCT)
  - ▷ Randomised play-the-winner rule (RPW)
  - ▷ Optimal dynamic programming policy (DP)
  - ▷ Whittle index policy (WI)
  - ▷ Randomised variant of the DP policy (RDP)
- We suggest to set $p = 90\%, \ell = 0.15n$ in CRDP

Simulation Results: Designs Comparison



Simulation Results: Designs Comparison

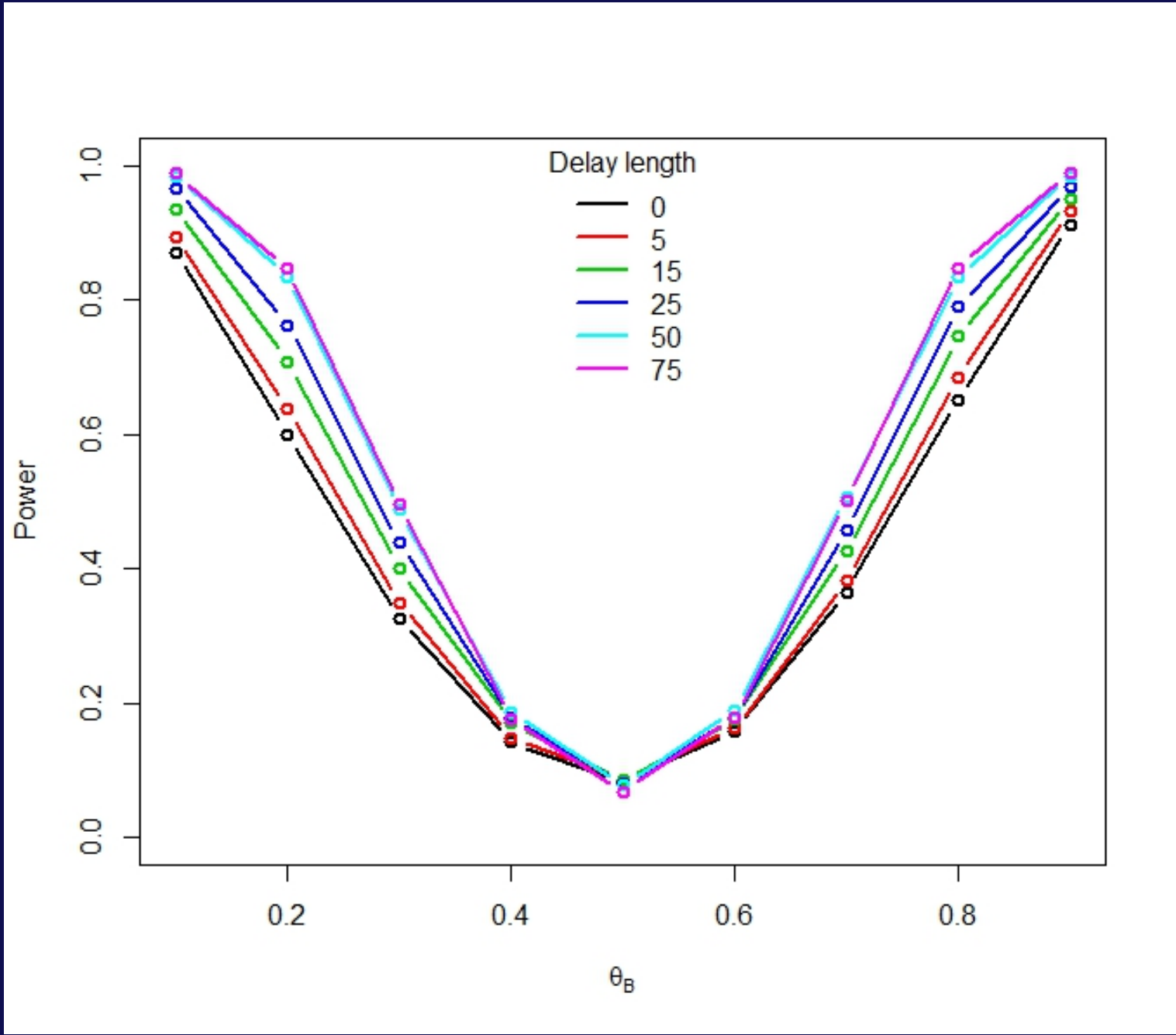
Example. $n = 75, \theta_A = 0.2, \theta_B = 0.6$

Design	Bias	Power	EPS	On sup
RCT	0.000	0.935	40.0%	50.0%
RPW	0.002	0.928	46.2%	65.4%
WI	0.092	0.066	57.8%	94.4%
DP	0.088	0.074	57.7%	94.1%
RDP	0.015	0.636	54.4%	86.0%
CRDP	0.008	0.724	53.6%	84.0%

Simulation Results: Designs Comparison

- In our suggested CRDP design
  - the % expected proportion of successes is much higher than in the traditional fixed and RPW designs
  - the % allocated to the superior arm is much higher than in the traditional fixed and RPW designs
  - the power is largely improved upon relative to the other bandit designs
  - the bias is negligible, opposed to large bias of other bandit designs

Simulation Results: Delayed Observations



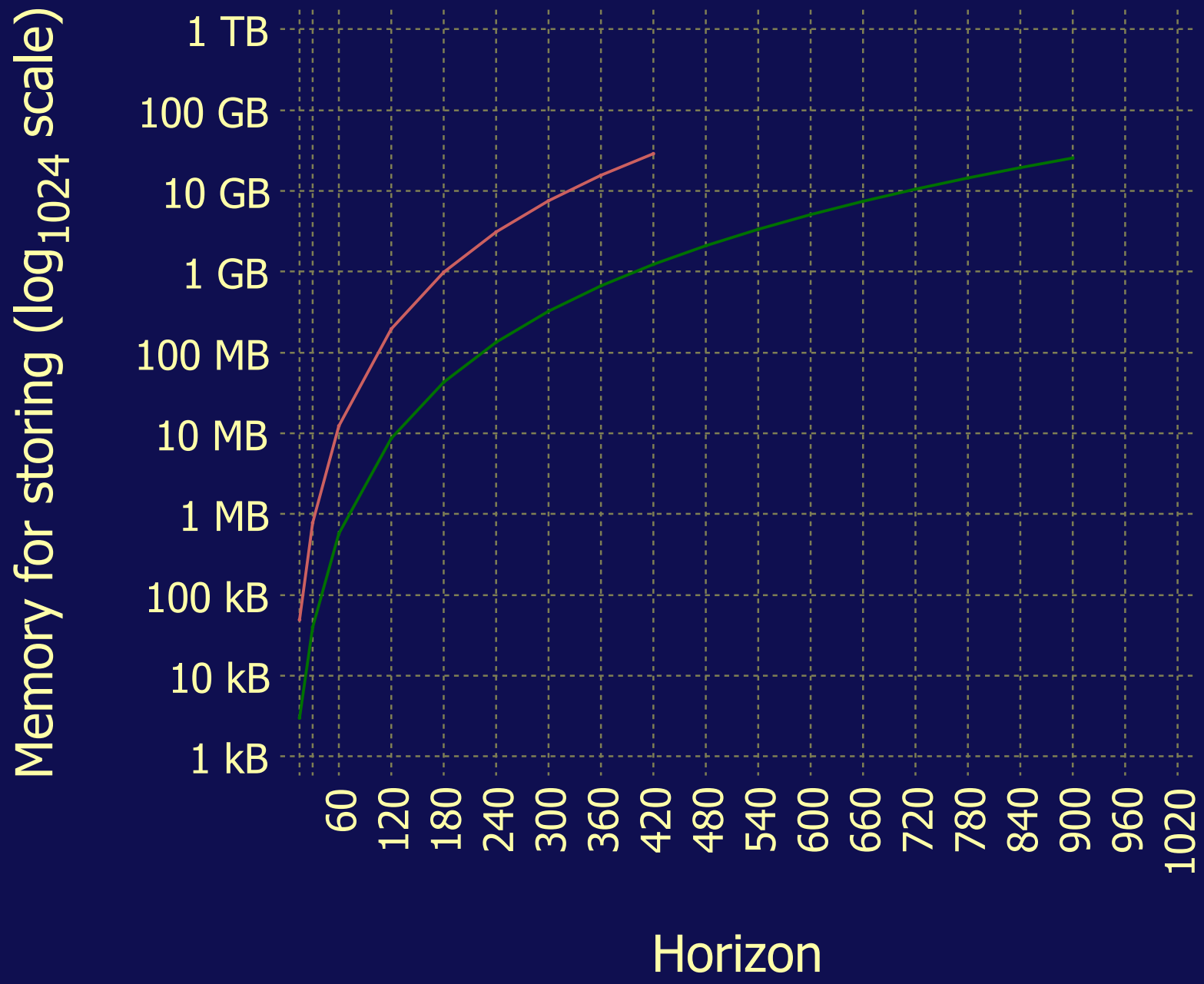
Conclusion about CRDP

- We address some of the key issues preventing bandit models from being implemented in clinical trial practice
  - ▷ lack of randomisation
  - ▷ insufficient statistical power
  - ▷ biased estimates of the intervention effect
- We need to talk to (bio)statisticians and clinicians about bandit models
  - ▷ give me randomisation probability and desired power
  - ▷ I tell how to randomise treatments to heal patients
- See Williamson et al. (CSDA 2017)

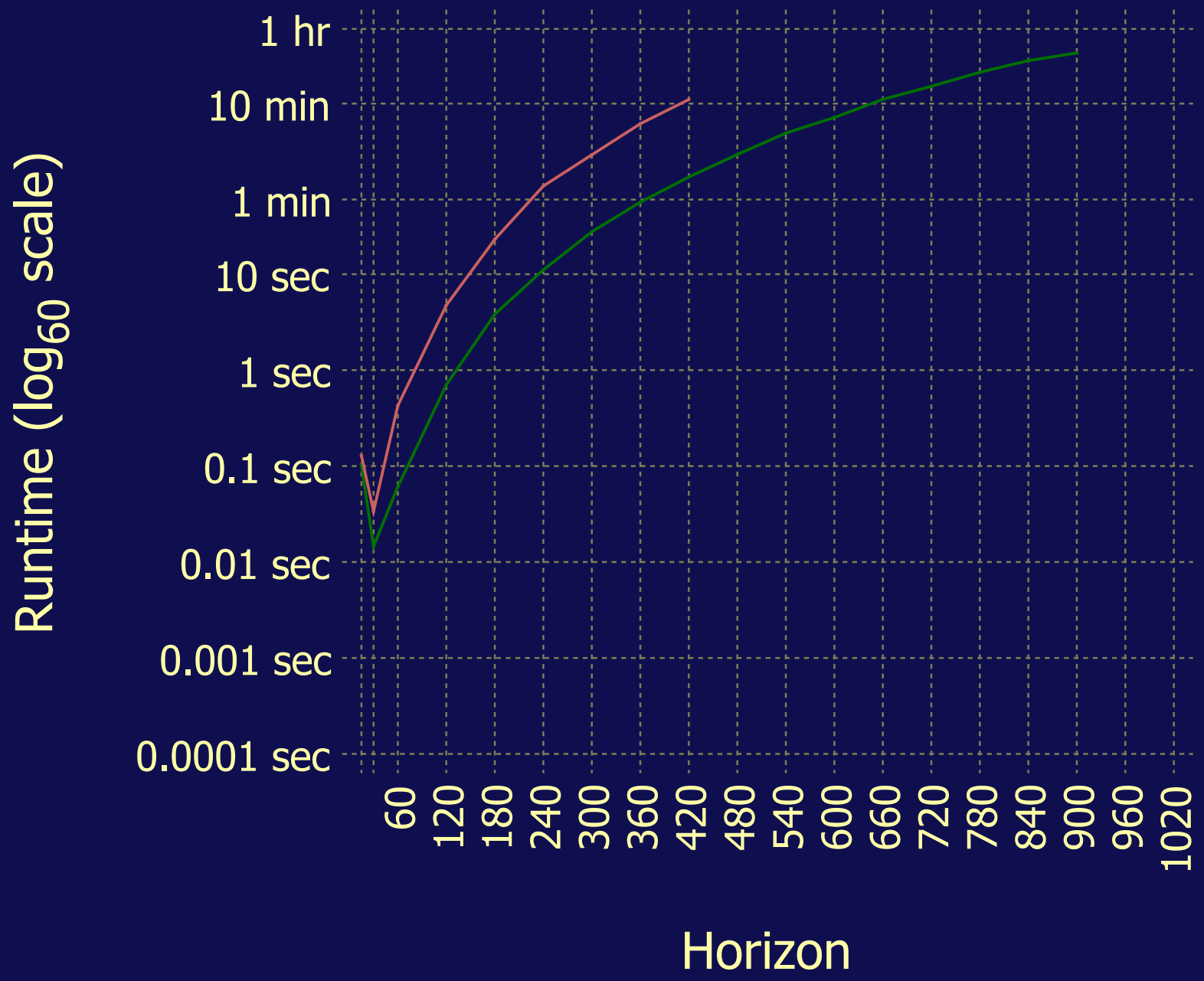
Optimal Designs

- The designs computed using DP (DP, RDP, CRDP) are optimal, i.e. provide the maximum benefit given their respective restrictions
- For the two-armed case (on a standard laptop):
  - ▷ a basic R code can design trials of size up to 200
  - ▷ an efficient Julia code up to $1,000$
  - ▷ a during-the-trial computation allows even larger trials
- Longer trials can be designed on a workstation/cloud
- More complex trials with much smaller sizes

Optimal Designs: Memory



Optimal Designs: Runtime



Near-optimal Designs

- Near-optimal designs can be obtained using index rules
  - idea: for every subject, allocate the intervention which has the highest value of an adaptive priority index
  - these indices are obtained in a particular way (Gittins index, Whittle index), and are asymptotically optimal
  - suboptimality is negligible
  - do not suffer from the curse of dimensionality
  - suitable for more complex trials
- Unfortunately, not enough time to explain in detail
  - requires understanding of Markov decision processes

Near-optimal Designs

- See e.g.,
  - Glazebrook (Biometrika 1978)
  - Gittins & Jones (Biometrika 1979)
  - Coad (Biometrika 1991)
  - Wang (Biometrika 1991)
  - Villar, Bowden & Wason (Statistical Science 2015): survey
  - Villar, Wason & Bowden (Biometrics 2015): group-sequential
  - Villar (PEIS 2017): Whittle index rule
  - Villar & Rosenberger (Biometrics 2017): covariates
  - Villar, Bowden & Wason (Pharmaceutical S. 2018): time trends
  - Smith & Villar (Applied S. 2018): Gaussian responses
  - Williamson & Villar (2018+): Gaussian responses & group-sequential

Thompson's Posterior Sampling Designs

- Thompson 1933 proposed a heuristic:
  - ▷ randomise according to the posterior probability of being the best arm
- This can be done by exact calculation or by sampling
- Recently, several trials have been designed in this way
  - ▷ Don Berry (MD Anderson) and Berry Consultants
  - ▷ e.g. I SPY-2, GBM Agile
- Several recent papers by a group at Harvard
- "Easy" to use, but quite suboptimal

Related Designs

- To overcome the computational difficulty of optimal designs and to add randomisation, improve estimation, etc., several designs have been proposed
  - ▷ not in Bayesian framework
- Urn designs (e.g. Randomised Play-the-Winner)
  - ▷ based on the property of the optimal design to allocate the intervention again if a success was observed on the last subject
  - ▷ there are several variants (e.g. Drop-the-Loser)
  - ▷ well studied theoretically (e.g. Rosenberger)
  - ▷ randomisation probabilities not too far from 50%

Related Designs

- Biased-coin designs
  - ▷ well studied theoretically
  - ▷ randomisation probabilities not too far from 50%
- UCB designs
  - ▷ optimal asymptotically in infinite horizon

Comparison of Designs

Table 1: $n = 148, \theta_A = 0.3, \theta_B = 0.5$.

Design	Power	On sup	ENS
RCT	0.81	50%	59.2
TS	0.80	69%	64.9
UCB	0.80	72%	66.0
GI	0.36	86%	70.2
WI	0.28	88%	70.7
*	-	100%	74.0

Source: Table 5 of Villar et al. (2015)

Conclusion about Designs

- Both the Thompson's posterior sampling and the above related designs are myopic
  - randomisation decisions are bases on the assumption that the next subject is the last one
  - relatively easy to use, but too suboptimal
- Optimal designs and index rules are not myopic
  - randomisation decisions take into account the remaining trial size and the after-trial population size
  - this is the reason why they provide a significantly higher benefit

Work in Progress

- With F. Williamson and T. Jaki, we are studying the CRDP design if there are delayed responses
- With J. Wason, we are looking at when it is optimal to add in a novel intervention to a platform trial
- I am working on a Julia package to compute the optimal designs

Conclusion about Bandits

- Look at papers in other disciplines, look at old papers
  - there are good ideas and useful techniques out there (reusable!)
  - hard (but not impossible) to read mainly due to different terminology and mindsets
- It would be beneficial to follow a common terminology across disciplines
  - to have quicker and wider impact
  - my suggestion: designs that prescribe subjects' allocation to interventions

Thank you for your attention

Clinical Trials

- Two treatments: control (existing one) and novel (not approved yet)
- Is the novel treatment better than the control?
  - ▷ clinically relevant treatment effect difference
  - ▷ if not, it will not be approved!
- One sets up a clinical trial of n patients
- A (clinical trial) design is an allocation policy that specifies which treatment the ith subject will receive

Randomised Controlled Trial

- Advantages of randomised controlled trial
  - ▷ best possible for estimating the intervention effect
  - ▷ "easy" to understand by trial statisticians, physicians
    - basic statistical knowledge
  - ▷ "quick" to design
    - trial size formulae available
  - ▷ "straightforward" to implement
    - no computation needed during the trial
  - ▷ "easy" to understand by in-trial subjects (?)
  - ▷ "easy" to interpret the results by regulators
  - ▷ "easy" to interpret the results by physicians (?)
  - ▷ "easy" to interpret the results by subjects (?)

Randomised Controlled Trial

- Disadvantages of randomised controlled trial
  - ▷ cost: 20% error of not approving a better intervention
    - — development and approval processes: $ billions
  - ▷ faith: once approved, no (simple) way to retract the intervention
    - — worse intervention approved by 5% chance
    - — unforeseen long-term secondary effects
  - ▷ feasibility: requires hundreds of subjects for a trial
  - ▷ stationarity: approval process takes years
    - — inappropriate for new diseases and epidemics
  - ▷ ethics: subjects join a trial expecting to get a possibly better (unapproved yet) intervention

Simulation Results: Randomised Variant

Example. $n = 75, \theta_A = 0.2, \theta_B = 0.8$

p	Bias	MSE	Power	EPS	On sup
50%	0.001	0.004	1.000	50.0%	50.0%
60%	0.001	0.005	1.000	55.7%	59.6%
70%	0.001	0.007	0.999	61.5%	69.2%
80%	0.004	0.010	0.995	67.2%	78.8%
90%	0.009	0.019	0.937	73.0%	88.3%
100%	0.100	0.043	0.118	78.6%	97.6%

- The Power (almost) does not change if p increased from 50% to 60% or 70%. Room for increasing EPS!

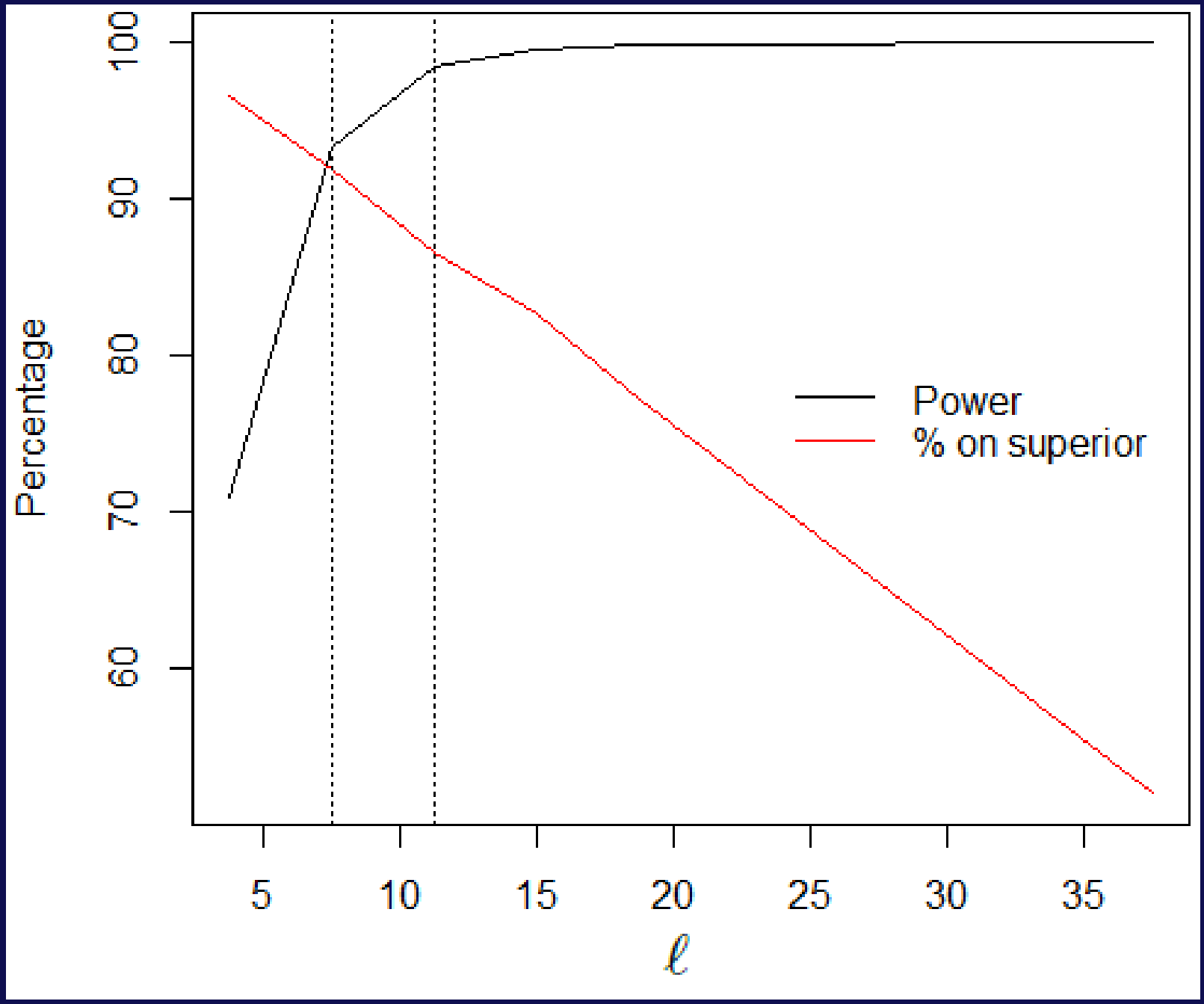
Simulation Results: Constrained Variant

Example. $n = 75, \theta_A = 0.2, \theta_B = 0.8$

ℓ	Power	EPS	On sup
$0.05n$	0.442	78.0%	96.6%
$0.10n$	0.884	75.2%	91.9%
$0.15n$	0.964	72.1%	86.7%
$0.20n$	0.985	69.7%	82.7%
$0.25n$	0.997	66.5%	77.3%
$0.50n$	1.000	51.2%	52.0%

- As ℓ increases, the power of the design increases hyperbolically, but the EPS and % allocated to the superior arm decreases linearly

Simulation Results: Constrained Variant



Simulation Results: Constrained Randomised Variant

Example. $n = 75, \ell = 0.15n, \theta_A = 0.2, \theta_B = 0.8$

p	Bias	MSE	Power	EPS	On sup
60%	0.001	0.005	1.000	55.7%	59.6%
70%	0.001	0.007	0.999	61.5%	69.2%
80%	0.003	0.010	0.996	67.2%	78.7%
90%	0.003	0.014	0.977	71.3%	85.5%

- The Power is quite high even if p increased to 80% or 90%. Also bias diminishes!

Simulation Results: Designs Comparison

Example. $n = 75, \theta_A = 0.2, \theta_B = 0.8$

Design	Bias	Power	EPS	On sup
RCT	0.000	1.000	50.0%	50.0%
RPW	0.008	0.998	66.2%	76.9%
WI	0.098	0.108	78.6%	97.6%
DP	0.100	0.118	78.6%	97.5%
RDP	0.009	0.937	73.0%	88.3%
CRDP	0.003	0.977	71.3%	85.5%