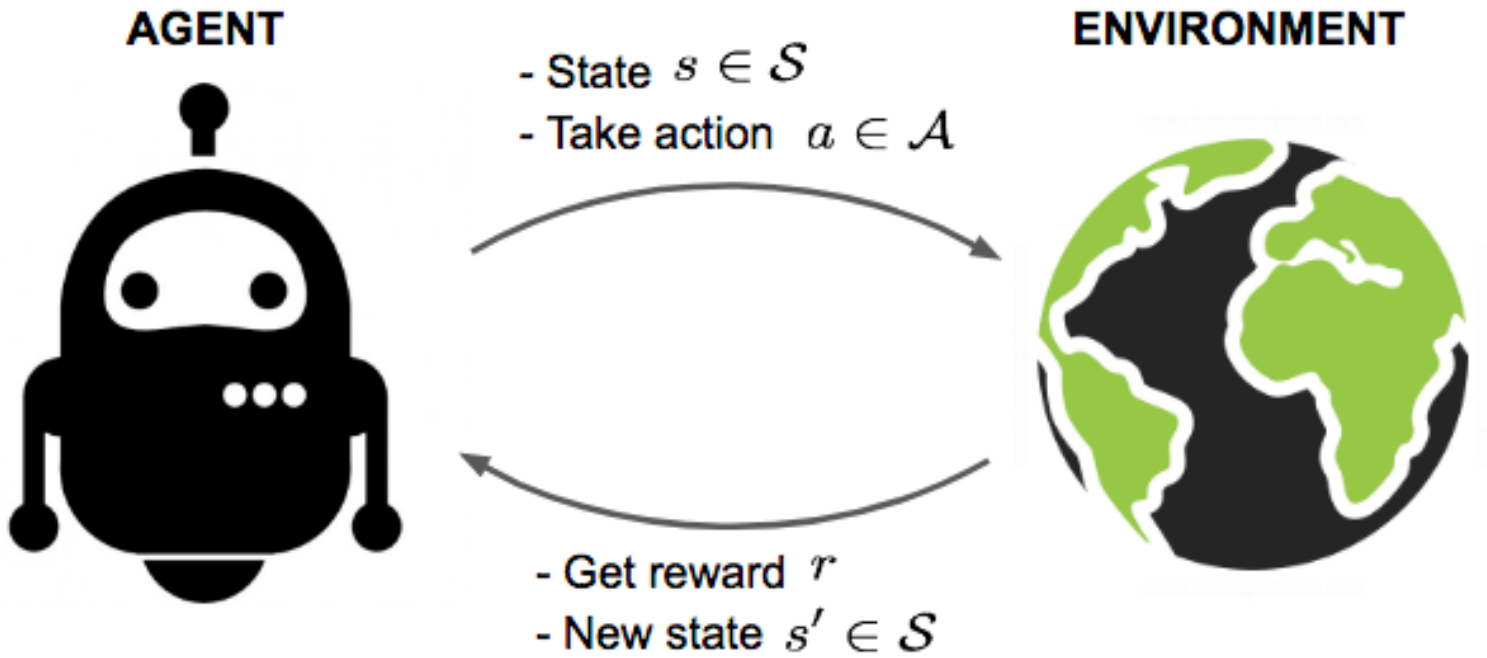
Statistical Inference of the Value Function for Reinforcement Learning in Infinite Horizon Settings

Chengchun Shi

Department of Statistics
North Carolina State University

Joint work with Sheng Zhang, Wenbin Lu and Rui Song

Reinforcement learning



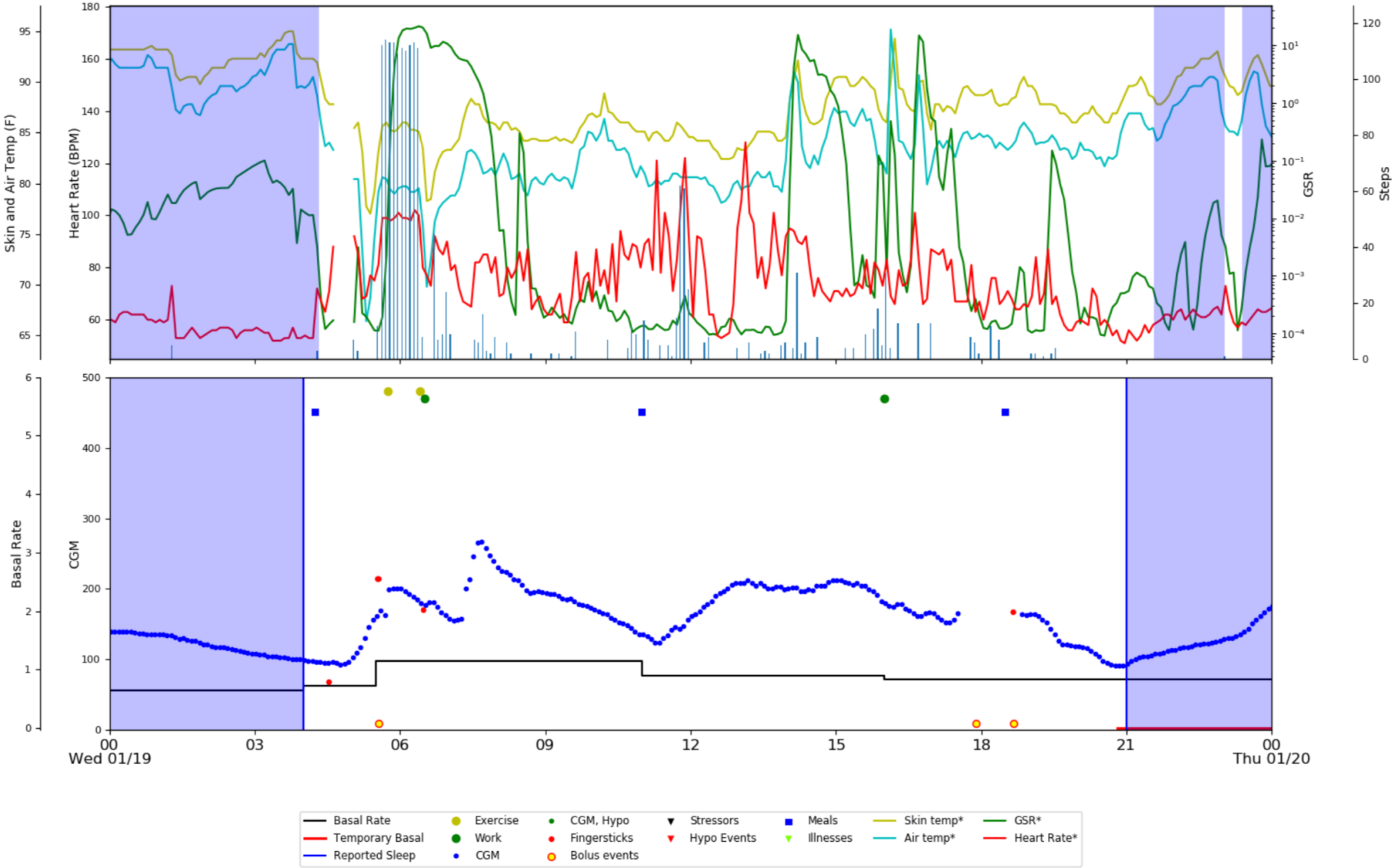
Applications

- video games (Silver et al., 2016)
- robotics (Kormushev et al., 2013)
- bidding (Jin et al., 2018)
- ridesharing (Xu et al., 2018)

The OhioT1DM dataset (Marling and Bunescu, 2018)

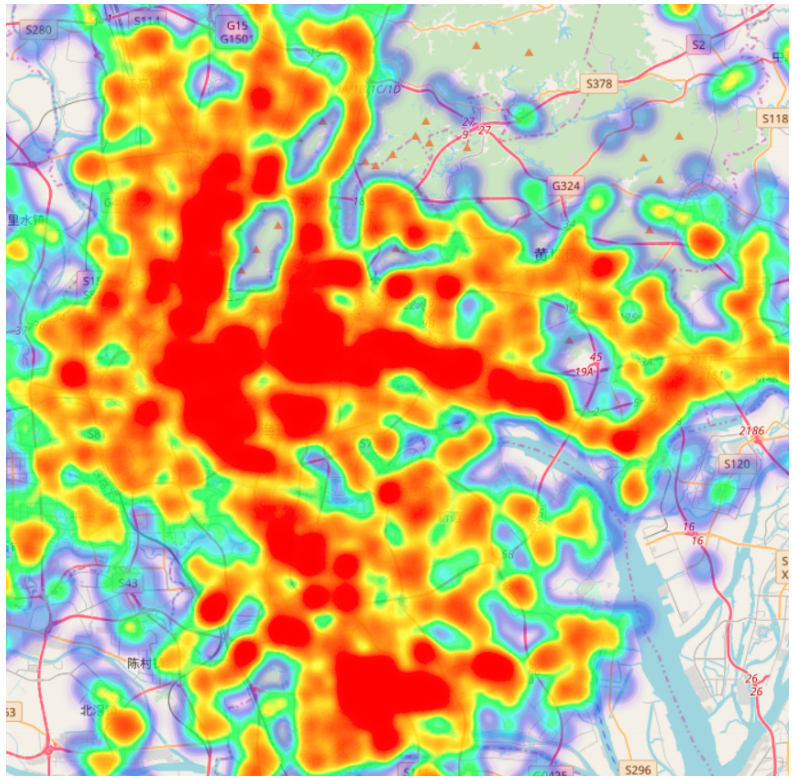
- 6 patients with type I diabetes
- For each patient, it contains eight weeks' worth of
  - continuous glucose monitoring blood glucose levels;
  - insulin doses including bolus and basal rates;
  - self-reported times of meals and exercises.
- Features of data:
  - small N, large T
  - heterogeneity across patients
- The optimal policy: a function that maps patient's status at each time point into the insulin doses

The OhioT1DM dataset (Marling and Bunescu, 2018)



Data from Didi Chuxing: The World's Leading Ridesharing Company

- Applying different order dispatching strategies at different time points :
- State: demand/supply
- Action: order dispatching strategies
- Reward: GMV/answer rate/average pickup distance
- 10^8 orders per week in Guangzhou (heat map at the right)



Value function in infinite horizon settings

- **State-value function** measures the goodness of a policy, starting from some initial states.
- **Integrated value function** is the aggregated state-value function over different initial states.
- **Inference of the state-value function**
  - helps a decision maker to evaluate the impact of implementing a policy when the environment is in a certain state.
- **Inference of the integrated value function**
  - helps a decision maker to evaluate the impact of implementing a policy in the population;
  - helps compare different policies.

Statistics literature on reinforcement learning

Finite horizon settings

- Estimation of the optimal policy
  - Murphy (2003); Zhang et al. (2013); Zhao et al. (2015); Shi et al. (2018); Zhang et al. (2018)
- Inference of the value
  - Chakraborty et al. (2013); Luedtke and van der Laan (2016); Shi et al. (2019)

Infinite horizon settings

- Estimation of the optimal policy
  - Gradient Q-learning (Ertefaie and Strawderman, 2018)
  - V-learning (Luckett et al., 2019)
- Inference of the value:
  - **Our proposal**

A general framework for inference of the value

Policies	Types of values	On/off-policy
Fixed: random ✓ deterministic ✓	CI for the value under a given state ✓	Off-policy ✓
Data-dependent: regular ✓ nonregular ✓	CI for the integrated value with respect to a reference function ✓	On-policy ✓

A general framework for inference of the value

Policies	Types of values	On/off-policy
Fixed: random ✓ deterministic ✓	CI for the value under a given state ✓	Off-policy ✓
Data-dependent: regular ✓ nonregular ✓	CI for the integrated value with respect to a reference function ✓	On-policy ✓

Propose a **S**equenti**AI** **V**alue **E**valuation (**SAVE**) method for handling data-dependent policies

A general framework for inference of the value

Policies	Types of values	On/off-policy
Fixed: random ✓ deterministic ✓	CI for the value under a given state ✓	Off-policy ✓
Data-dependent: regular ✓ nonregular ✓	CI for the integrated value with respect to a reference function ✓	On-policy ✓

A general framework for inference of the value

Policies	Types of values	On/off-policy
Fixed: random ✓ deterministic ✓	CI for the value under a given state ✓	Off-policy ✓
Data-dependent: regular ✓ nonregular ✓	CI for the integrated value with respect to a reference function ✓	On-policy ✓

Bidirectional asymptotics: a framework where either n or T grows to infinity

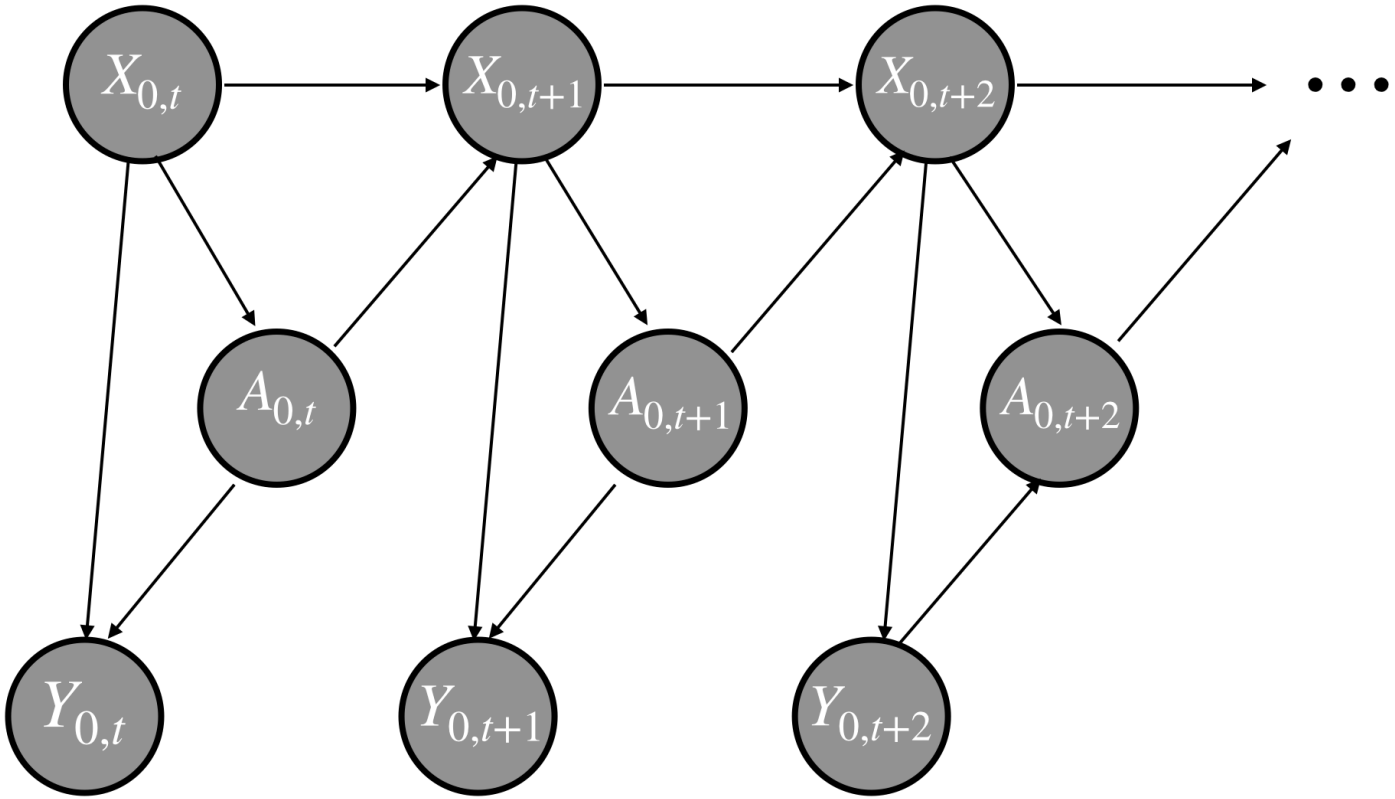
- large T, small N (mobile health, ridesharing)

- large N, small T

- large N, large T (games)

Optimal policy in infinite horizon settings

- At time point t: $X_{0,t}$, state variable; $A_{0,t}$, action; $Y_{0,t}$, reward.



Optimal policy in infinite horizon settings (Cont'd)

- Markovian assumption (MA):

$$\begin{aligned}&\Pr(X_{0,t+1} \in \mathcal{B} | X_{0,t} = x, A_{0,t} = a, \{X_{0,j}\}_{0 \le j < t}, \{A_{0,j}\}_{0 \le j < t}) \\ &\quad = \Pr(X_{0,t+1} \in \mathcal{B} | X_{0,t} = x, A_{0,t} = a) = \mathcal{P}(\mathcal{B} | x, a),\end{aligned}$$

  for some transition kernel $\mathcal{P}$.

- Conditional mean independence assumption (CMIA):

$$\begin{aligned}&\mathrm{E}(Y_{0,t} | X_{0,t} = x, A_{0,t} = a, \{X_{0,j}\}_{0 \le j < t}, \{A_{0,j}\}_{0 \le j < t}) \\ &\quad = \mathrm{E}(Y_{0,t} | X_{0,t} = x, A_{0,t} = a) = r(x, a),\end{aligned}$$

  for some function r.

- CMIA is implied by MA when $Y_{0,t}$ is a determinsitic function of $X_{0,t}$, $A_{0,t}$ and $X_{0,t+1}$.

Optimal policy in infinite horizon settings (Cont'd)

- The class of stationary policies $\pi : \mathbb{X} \to$ pmf on $\mathcal{A}$.
- Under π, agent will set $A_{0,t} = a$ with probability $\pi(a|X_{0,t})$ at time t.
- State-value function under π:

$$V(\pi; x) = \sum_{t=0}^{+\infty} \gamma^t \mathrm{E}^{\pi}(Y_{0,t}|X_{0,0} = x).$$

- Under MA and CMIA, there exists at least one optimal policy π^{opt} that satisfies $V(\pi^{opt}; x) \geq V(\pi; x)$ for all π and x (Puterman, 1994).

Optimal policy in infinite horizon settings (Cont'd)

- State-action value function (Q-function) under π:

$$Q(\pi; x, a) = \sum_{t=0}^{+\infty} \gamma^t \mathrm{E}^{\pi}(Y_{0,t} | X_{0,0} = x, A_{0,0} = a).$$

- Optimal Q-function:

$$Q^{opt}(x, a) = \sup_{\pi} Q(\pi; x, a), \quad \forall x, a.$$

- Any optimal policy π^{opt} satisfies

$$\pi^{opt}(a|x) = 0 \quad \text{if} \quad a \notin \arg\max_{a'} Q^{opt}(x, a'), \quad \forall x, a.$$

Value under a fixed policy in off-policy settings

- $\{(X_{0,t}, A_{0,t}, Y_{0,t})\}_{t\geq 0}$ generated according to a behavior policy $b(\cdot|\cdot)$:

  $\{(X_{1,t}, A_{1,t}, Y_{1,t})\}_{t\geq 0}, \{(X_{2,t}, A_{2,t}, Y_{2,t})\}_{t\geq 0}, \cdots, \{(X_{n,t}, A_{n,t}, Y_{n,t})\}_{t\geq 0},$

  are i.i.d copies of $\{(X_{0,t}, A_{0,t}, Y_{0,t})\}_{t\geq 0}$.
- **Objective**:
  - Construct CI for the state-value function $V(\pi; x)$;
  - Construct CI for the integrated value

$$V(\pi; \mathbb{G}) = \int_{x\in\mathbb{X}} V(\pi; x)\mathbb{G}(dx),$$

    for a given reference function $\mathbb{G}$.

Modelling value or Q-function?

- Bellman equation for the value:

$$V(\pi; x) = \sum_{a \in \mathcal{A}} \pi(a|x) \underbrace{\left\{ r(x, a) + \gamma \int_{x'} V(\pi; x') \mathcal{P}(dx'|x, a) \right\}}_{C(\pi; x, a)}.$$

- $C(\pi; \cdot, a)$ is continuous, under certain conditions. When $\pi(a|\cdot)$ is not continuous (such as the class of deterministic policies), so is $V(\pi; \cdot)$.
- Many nonparametric methods (series estimators, kernel smoothers, neural networks) require the underlying function to be smooth.
- Directly modelling the value poses significant challenges in performing inference to policies that are discontinuous functions of x.

Modelling value or Q-function? (Cont'd)

- Bellman equation for the Q-function:

$$Q(\pi; x, a) = r(x, a) + \gamma \sum_{a' \in \mathcal{A}} \int_{x'} Q(\pi; x', a') \pi(a'|x') \mathcal{P}(dx'|x, a).$$

Lemma

Suppose $\mathcal{P}(dx'|x, a) = q(x'|x, a)dx'$ for some transition density q, and $r(\cdot, a)$ and $q(x'|\cdot, a)$ are p-smooth for any a, x'. Then $Q(\pi; \cdot, a)$ is p-smooth for any π and a.

- Our procedure:
  - first estimate Q-function;
  - then derive value estimator based on the estimated Q-function.

Inference method

- Approximate $Q(\pi;\cdot,\cdot)$ by linear sieves $Q(\pi;x,a)\approx\Phi_L^\top(x)\beta_{\pi,a}$.
- Based on the Bellman equation,

$$\mathrm{E}\left[\left\{Y_{i,t}+\gamma\sum_{a\in\mathcal{A}}Q(\pi;X_{i,t+1},a)\pi(a|X_{i,t+1})-Q(\pi;X_{i,t},A_{i,t})\right\}\middle|X_{i,t},A_{i,t}\right]=0.$$

- Based on the above equation, compute $\widehat{\boldsymbol{\beta}}_\pi=(\widehat{\beta}_{\pi,1}^\top,\cdots,\widehat{\beta}_{\pi,m}^\top)^\top$ as the solution to

$$\sum_{i,t}\left\{Y_{i,t}+\gamma\sum_{a\in\mathcal{A}}\Phi_L^T(X_{i,t+1})\widehat{\beta}_{\pi,a}\pi(a|X_{i,t+1})-\Phi_L^T(X_{i,t})\widehat{\beta}_{\pi,a'}\right\}$$
$$\times\Phi_L(X_{i,t})\mathbb{I}(A_{i,t}=a')=0,\qquad\forall a'\in\{1,\cdots,m\}.$$

Inference method (Cont'd)

- Estimate $Q(\pi; x, a)$ by $\widehat{Q}(\pi; x, a) = \Phi_L^\top(x)\widehat{\beta}_{\pi,a}$.
- Derive estimator for the value based on the relation $V(\pi; x) = \sum_a Q(\pi; x, a)\pi(a|x)$.
- Set $\widehat{V}(\pi; x) = \sum_a \widehat{Q}(\pi; x, a)\pi(a|x)$ and

$$\widehat{V}(\pi; \mathbb{G}) = \int_x \widehat{V}(\pi; x)\mathbb{G}(dx).$$

- The proposed CI: $\widehat{V}(\pi; \mathbb{G}) \pm z_{\alpha/2}\widehat{\sigma}(\pi, \mathbb{G})$ for some $\widehat{\sigma}^2(\pi, \mathbb{G})$ that consistently estimates the variance of $\widehat{V}(\pi; \mathbb{G})$.
- Set $\mathbb{G} = \delta_x$ (the Dirac function), it yields CI for $V(\pi; x)$.

Validity of the proposed CI

A1 $r(\cdot, a)$ and $q(x'|\cdot, a)$ are p-smooth for any a, x'.

A2 Either tesnor-produc B-splines or Wavelet basis is used for $\Phi_L(\cdot)$.

A3 Densities of the initial distribution of $X_{0,0}$ and the limiting distribution of the Markov chain $\{X_{0,t}\}_{t\geq 0}$ are bounded away from 0 and ∞.

A4 Suppose (i) and (ii) hold when $T \to \infty$ and (iii) holds when T is bounded.

(i) $\liminf_T \mathrm{E}\widehat{\Sigma}_\pi(T) > 0$

(ii) The markov chain $\{X_{0,t}\}_{t\geq 0}$ is geometrically ergodic

(iii) $\mathrm{E}\widehat{\Sigma}_\pi(T) > 0$

Validity of the proposed CI (Cont'd)

Theorem (Bidirectional Asymptotics)

Under A1-A4 and some other conditions, the proposed CI is valid as long as either $n \to \infty$ *or* $T \to \infty$.

Major technical challenge

Characterize the convergence rate of some $L \times L$ random matrices (L allowed to grow with nT) as a function of both n and T.

Value under an (estimated) optimal policy in off-policy settings

- Consider an estimated policy $\widehat{\pi}$ that depends on the data $\{(X_{1,t}, A_{1,t}, Y_{1,t})\}_{t\geq 0}, \cdots, \{(X_{n,t}, A_{n,t}, Y_{n,t})\}_{t\geq 0}$.
- **Objective**:
  - Construct CI for the state-value function $V(\widehat{\pi}; x)$.
  - Construct CI for the integrated value $V(\widehat{\pi}; \mathbb{G})$.
- **Requirements on $\widehat{\pi}$**:
  - $V(\widehat{\pi}; \mathbb{G}) - V(\pi^*; \mathbb{G}) = O\{(nT)^{-b_*}\}$ for some $b_* > 1/2$.
  - When Q-learning type algorithms are used, the values can converge faster than the estimated Q-function so that the rate $(nT)^{-b_*}$ is achievable.
  - We do **not** require $\sum_a |\widehat{\pi}(a|x) - \pi^*(a|x)| \xrightarrow{P} 0$. This condition is violated in the nonregular cases.

Challenge of inference in the nonregular cases

- Suppose $\widehat{\pi}$ is computed by some Q-learning type algorithms,

$$\widehat{\pi}(a|x) = \begin{cases} 1, & \text{if } \ a = \operatorname{sarg\,max}_{a' \in \mathcal{A}} \widehat{Q}(x, a'), \\ 0, & \text{otherwise,} \end{cases}$$

  where $\widehat{Q}(\cdot, \cdot)$ denotes some consistent estimator for $Q^{opt}(\cdot, \cdot)$.
- In the nonregular cases where $\arg\max_a Q^{opt}(x, a)$ is not unique for some x, $\widehat{\pi}$ will not converge to a fixed quantity.
- The variance of $\widehat{V}(\widehat{\pi}; \mathbb{G})$ is difficult to estimate.

SAVE in off-policy settings

- Our procedure:

Step 1 Divide the data into $K_n \times K_T$ blocks.

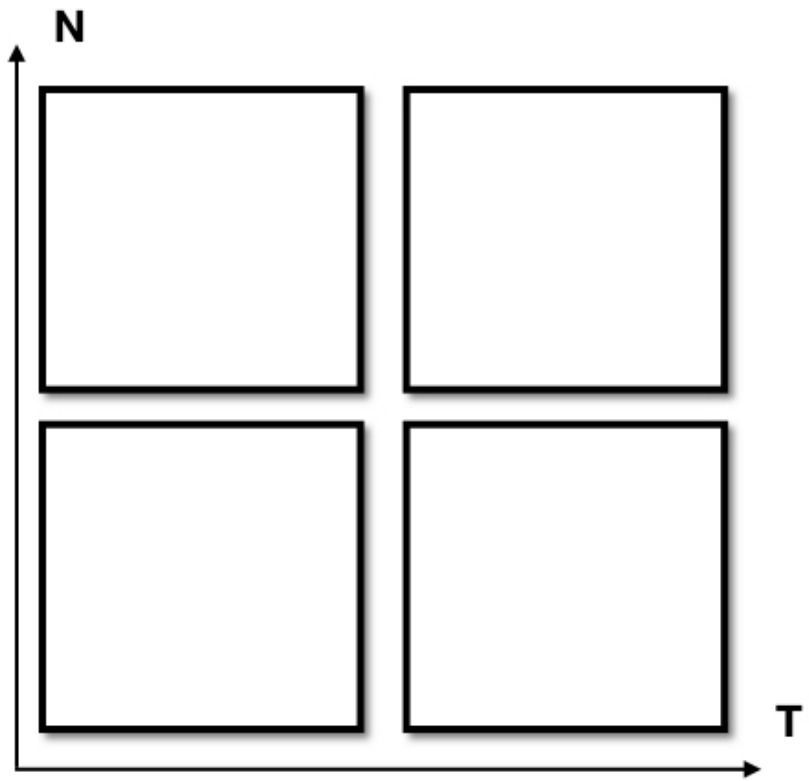
Step 2 Initialize $k = 1$. While $k < K_n K_T$:

1. Use the first k-th blocks of data to estimate the optimal policy and use the $k+1$-th block of data to evaluate its value;
2. Set $k \to k+1$.

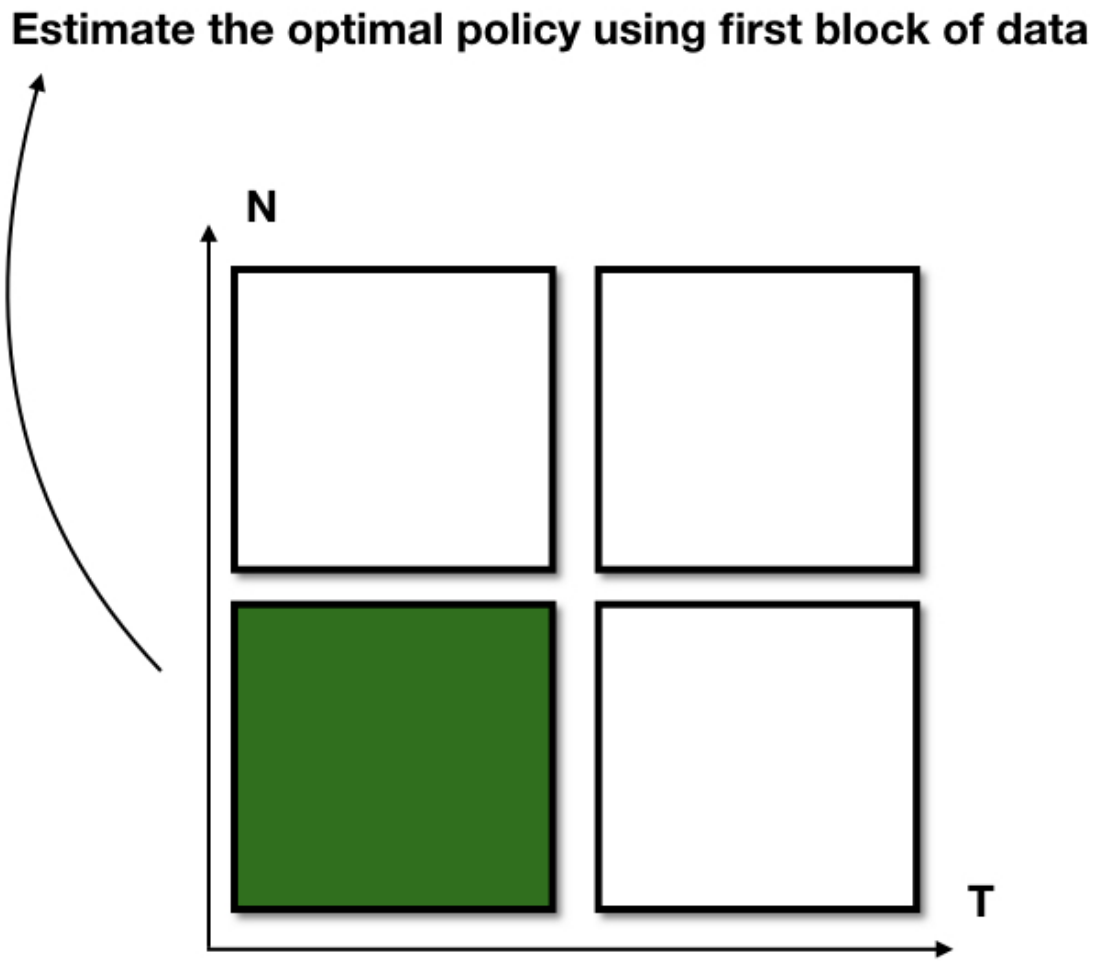
Step 3 Derive the final estimator as a weighted average of all $K-1$ value estimators.

- Orders of these blocks cannot be arbitrarily determined since observations are dependent

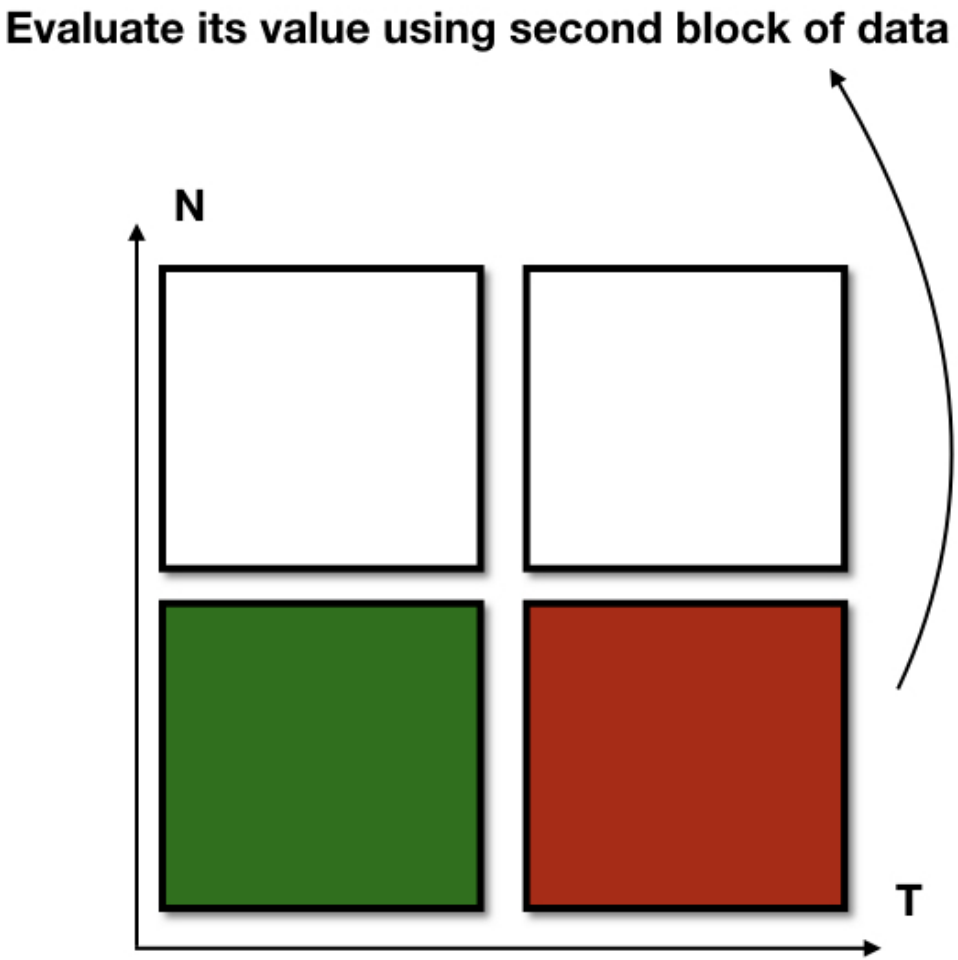
An illustration of SAVE



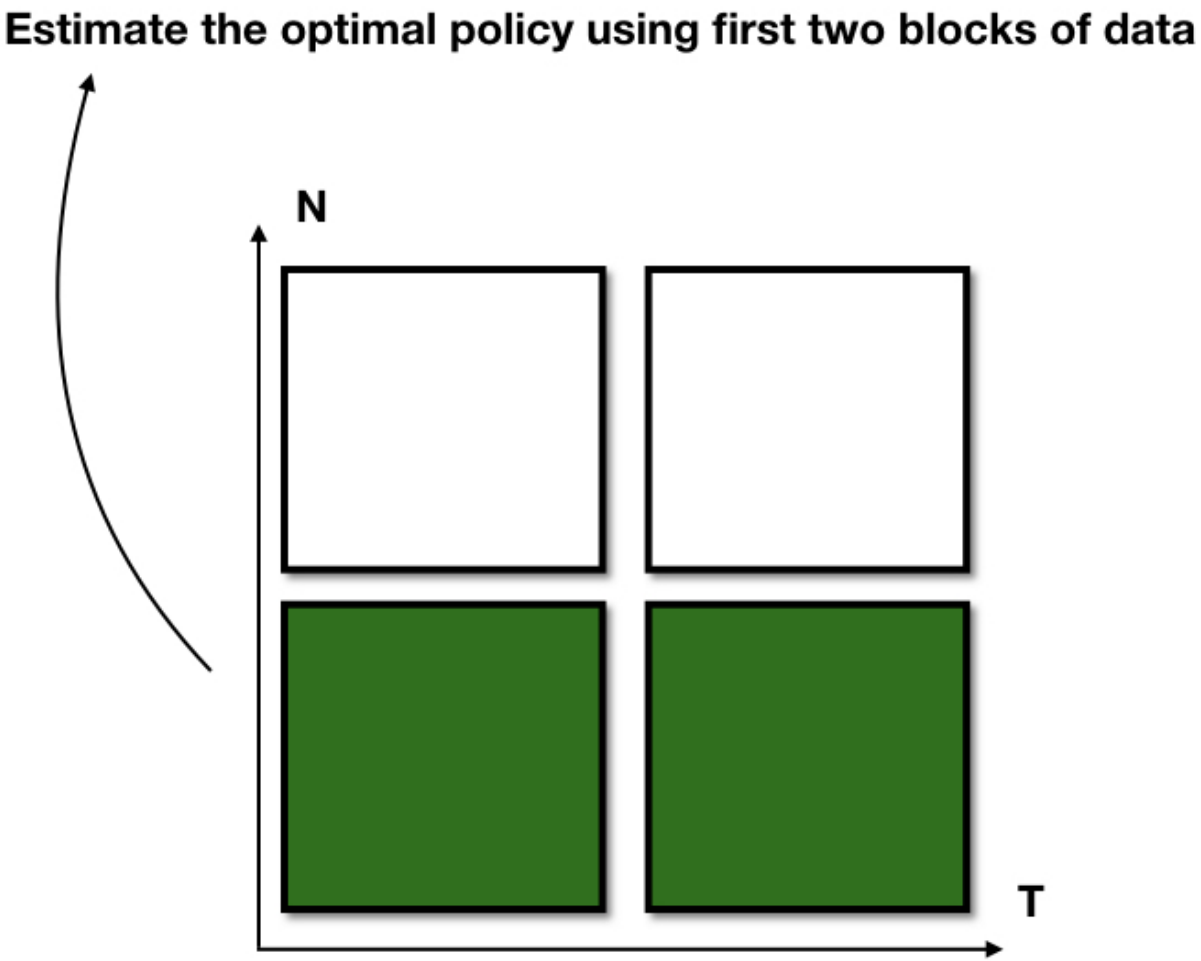
An illustration of SAVE



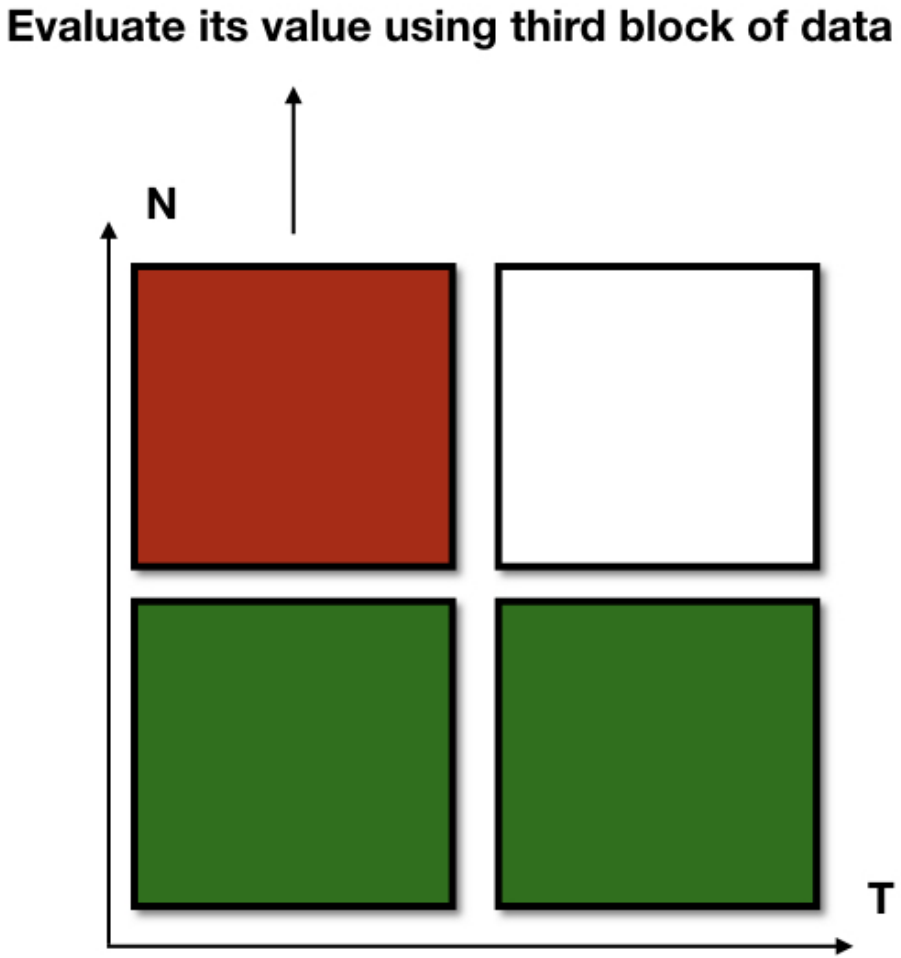
An illustration of SAVE



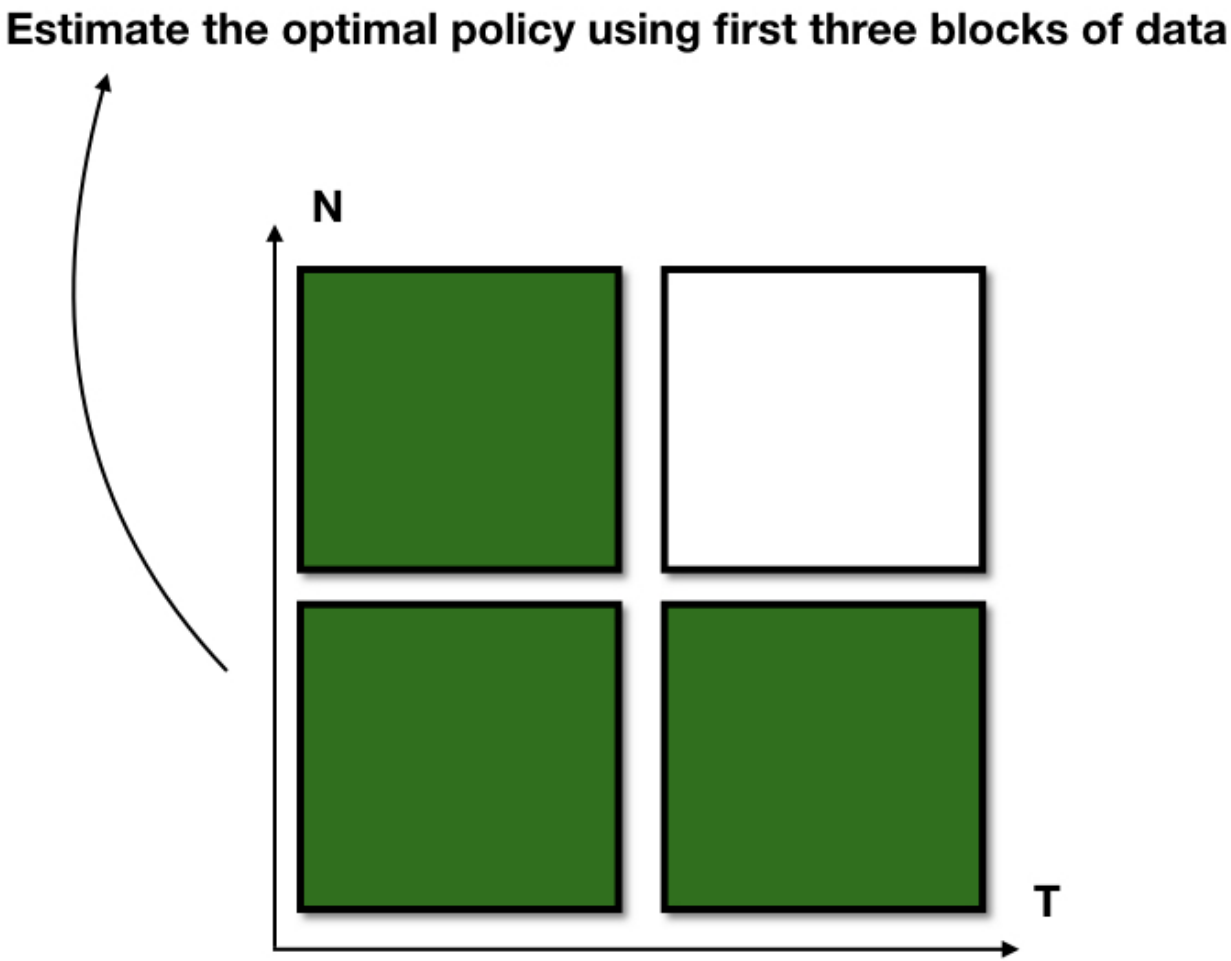
An illustration of SAVE



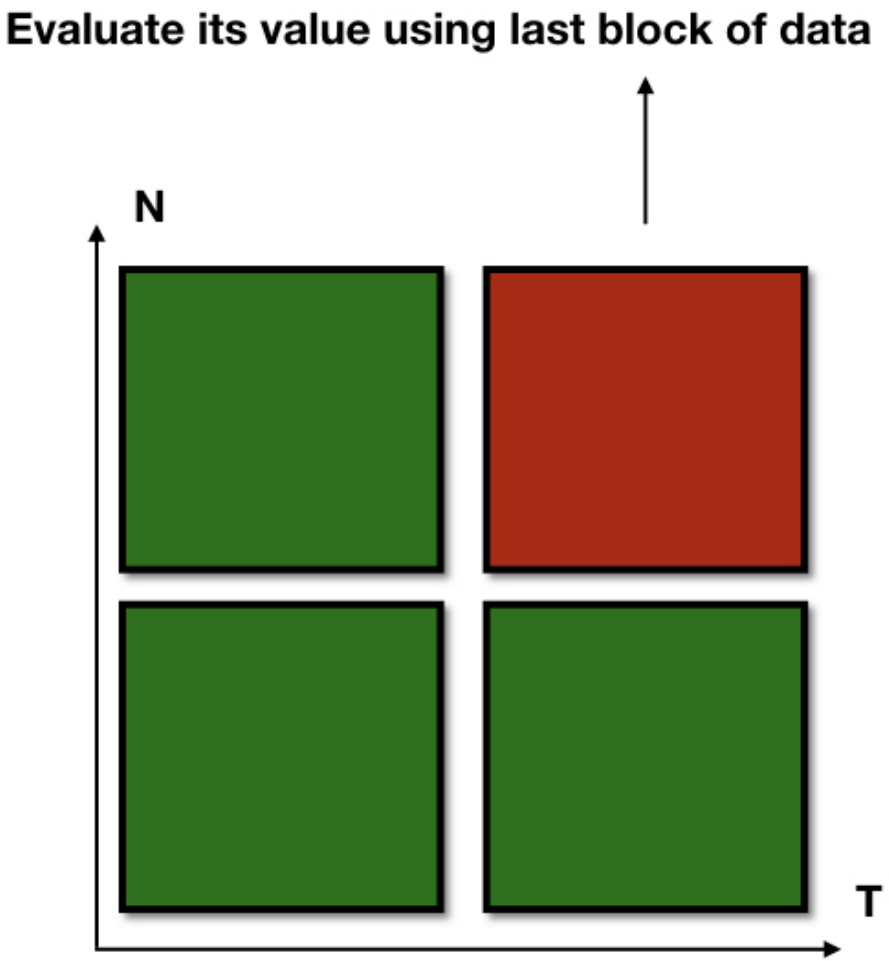
An illustration of SAVE



An illustration of SAVE

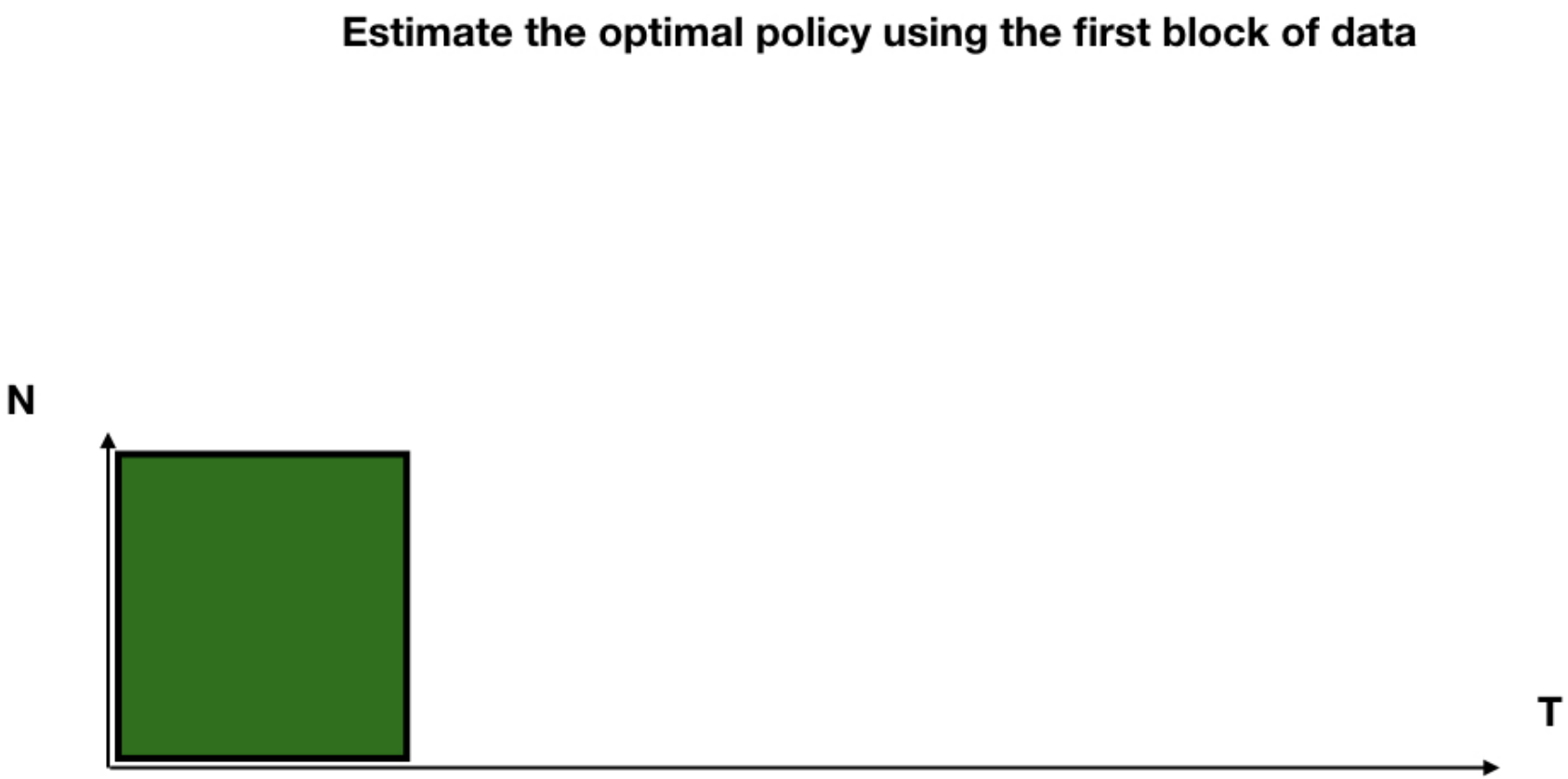


An illustration of SAVE

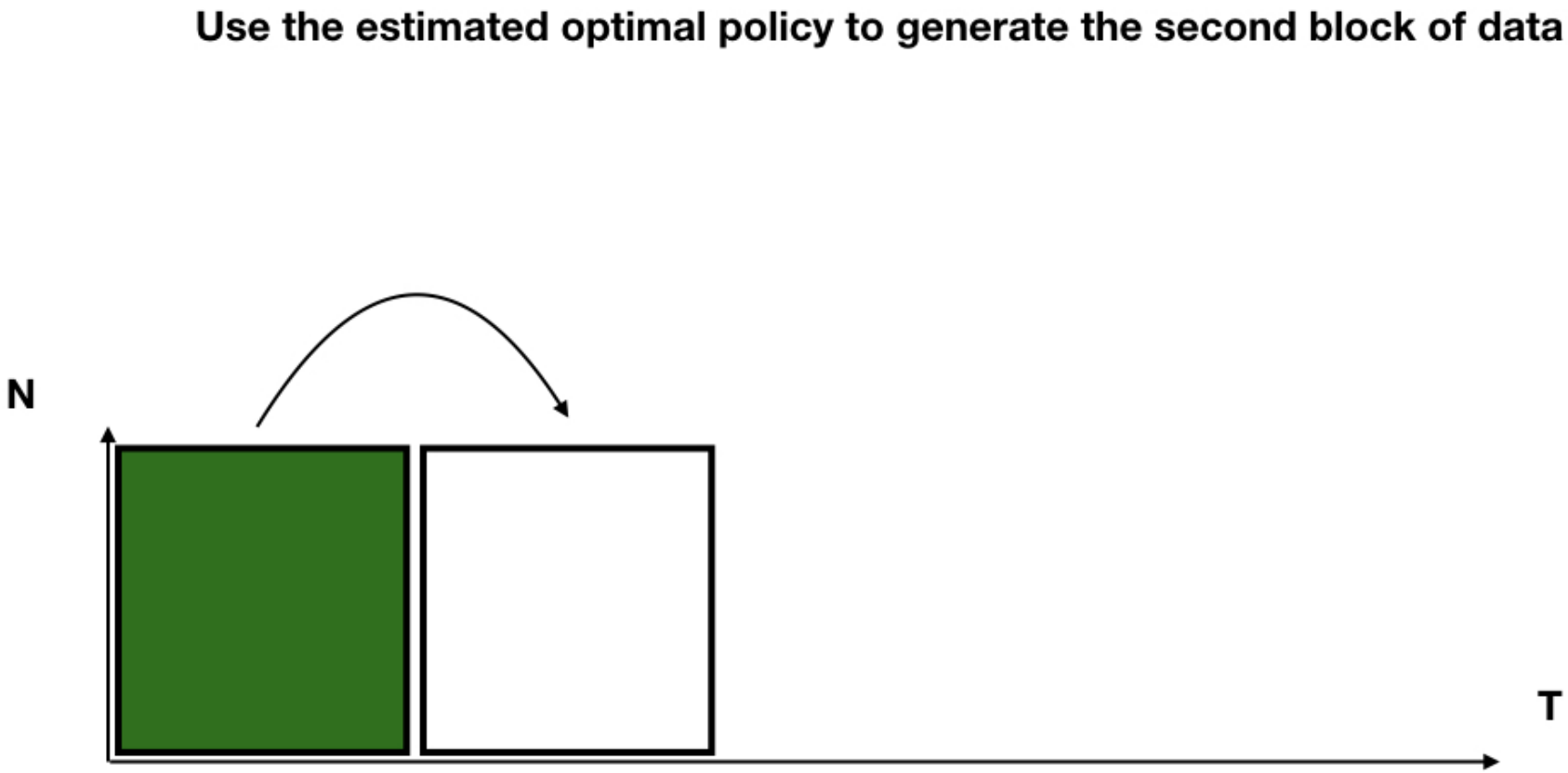


Value under an (estimated) optimal policy in on-policy settings: an illustration

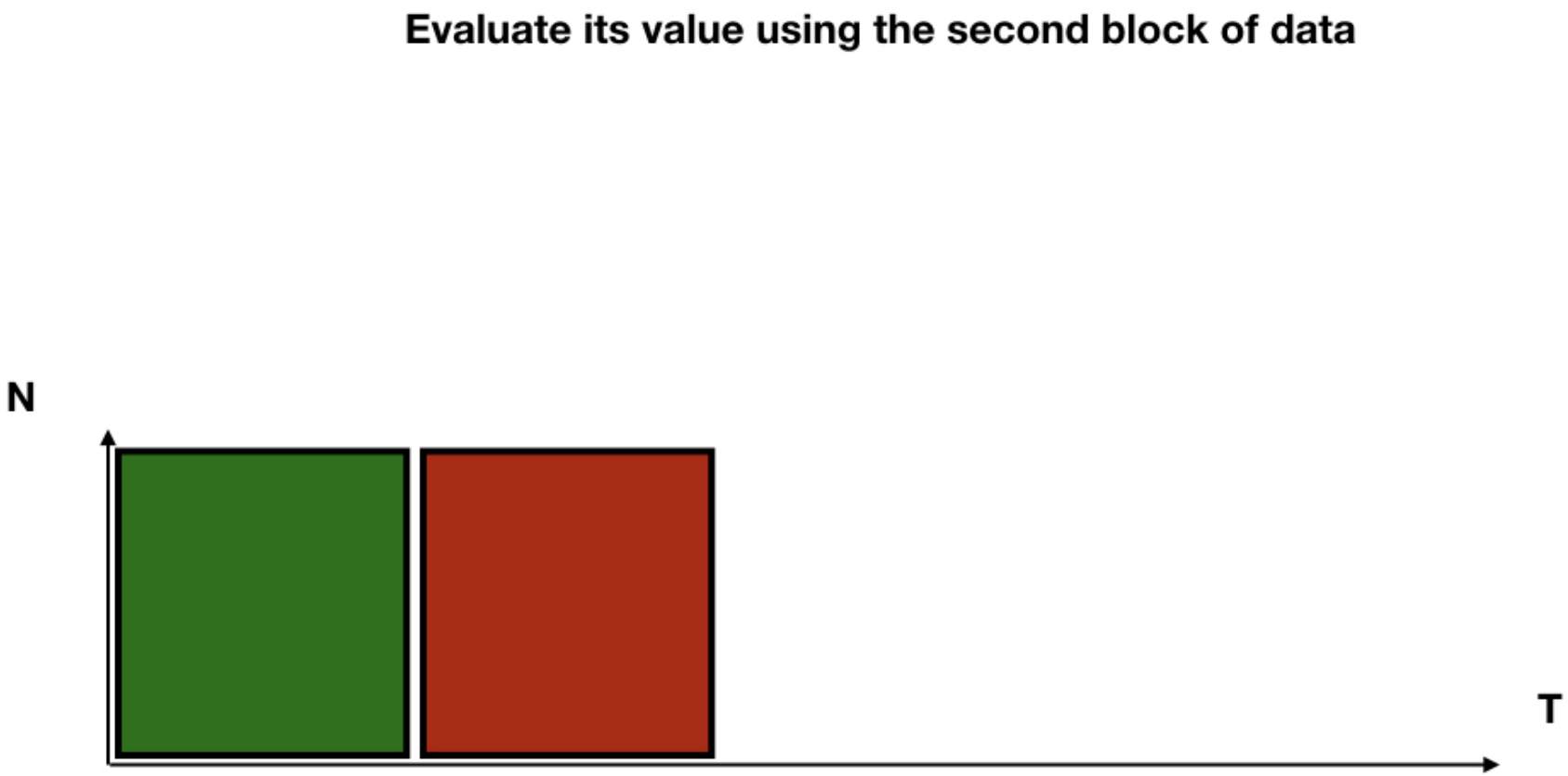
Value under an (estimated) optimal policy in on-policy settings: an illustration for SAVE



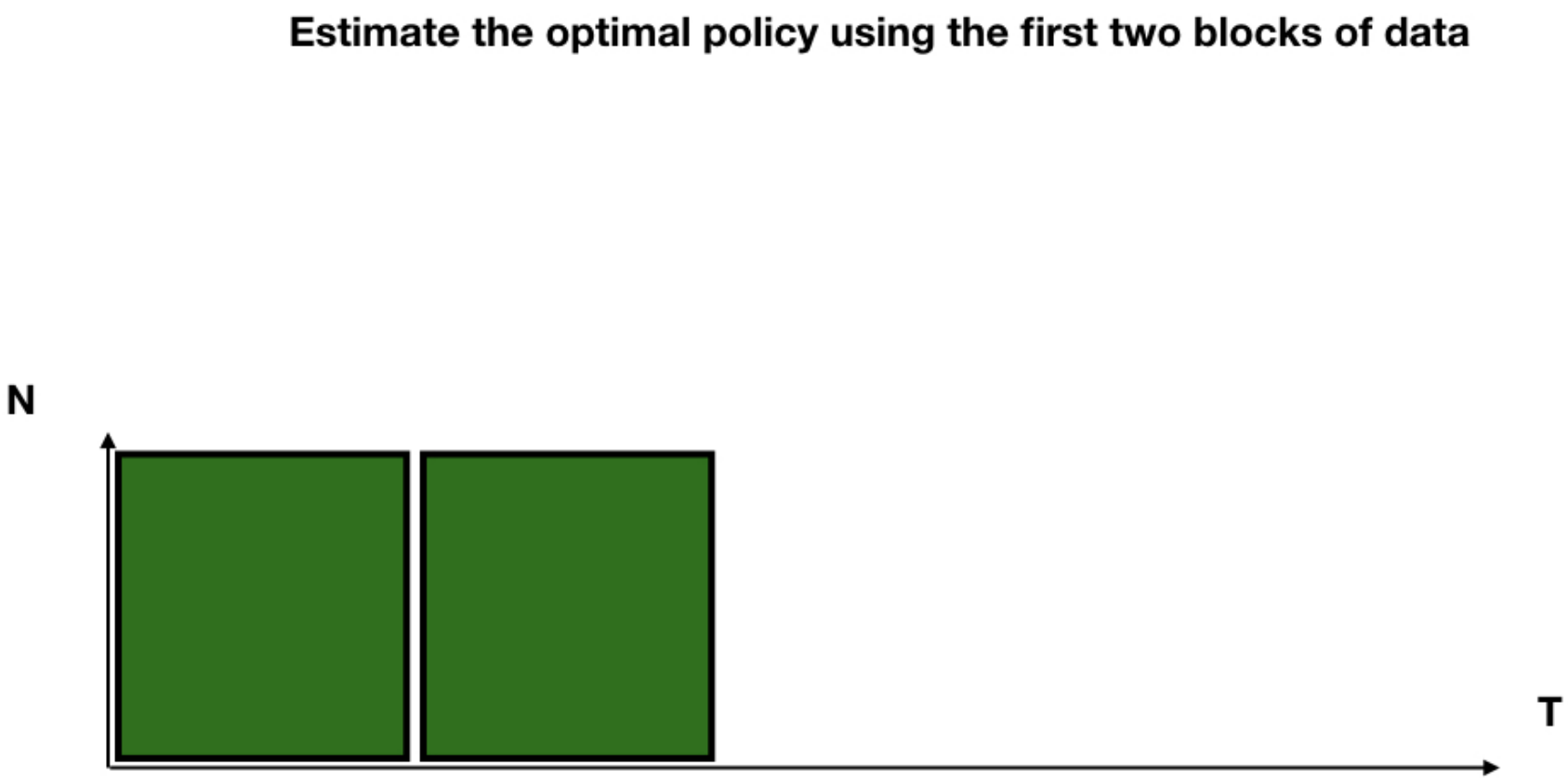
Value under an (estimated) optimal policy in on-policy settings: an illustration for SAVE



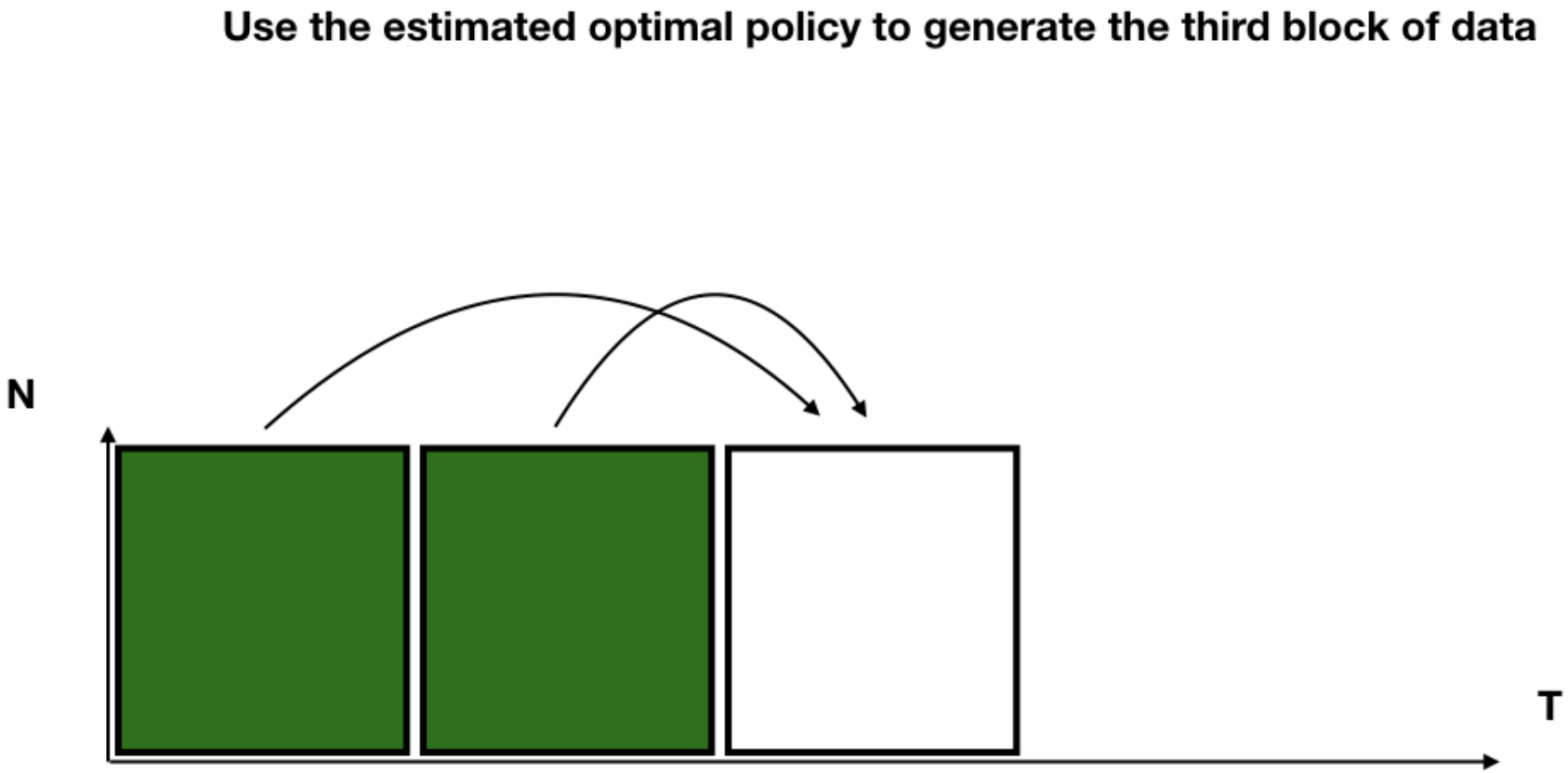
Value under an (estimated) optimal policy in on-policy settings: an illustration for SAVE



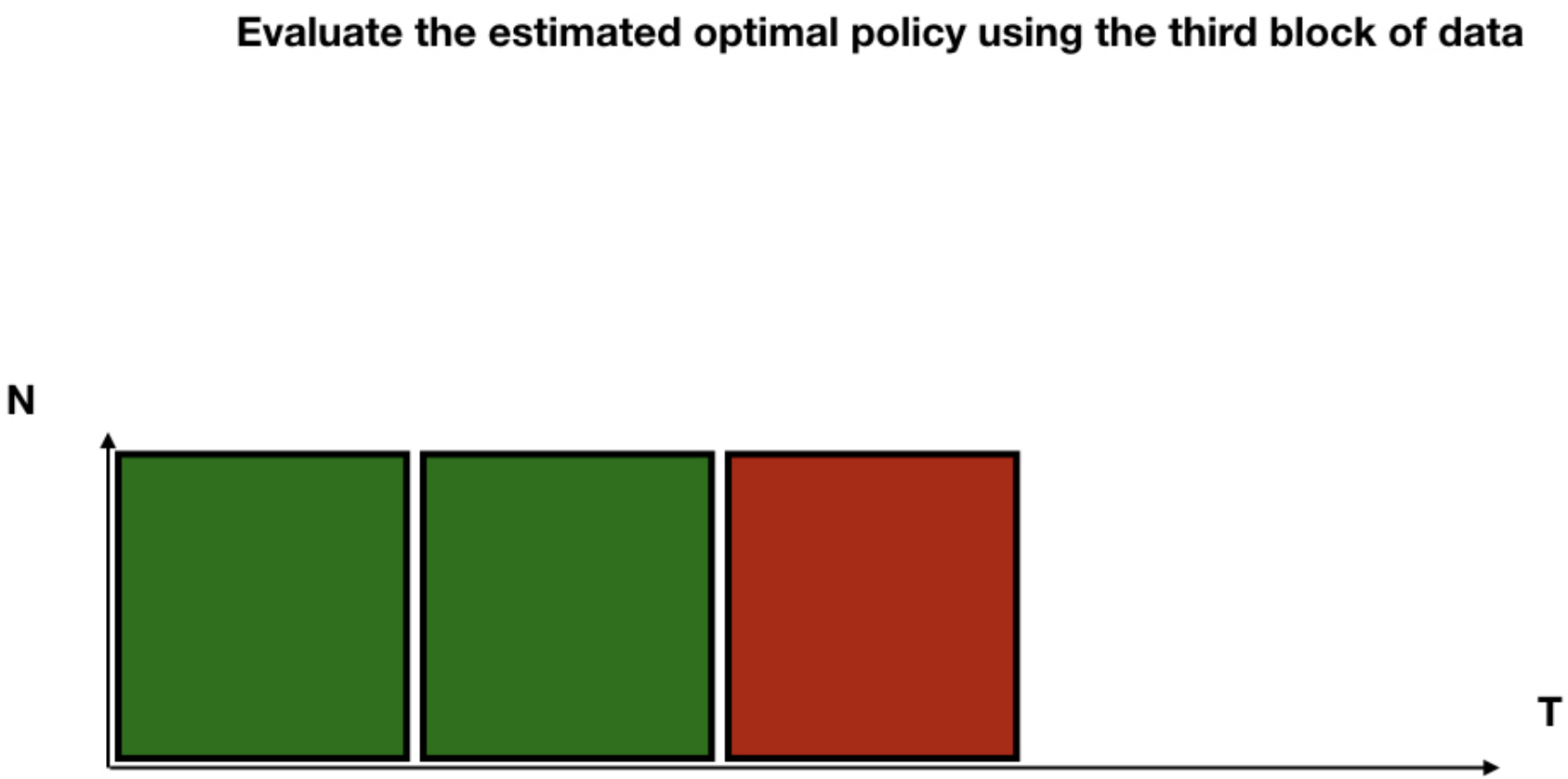
Value under an (estimated) optimal policy in on-policy settings: an illustration for SAVE



Value under an (estimated) optimal policy in on-policy settings: an illustration for SAVE



Value under an (estimated) optimal policy in on-policy settings: an illustration for SAVE



Thank you!