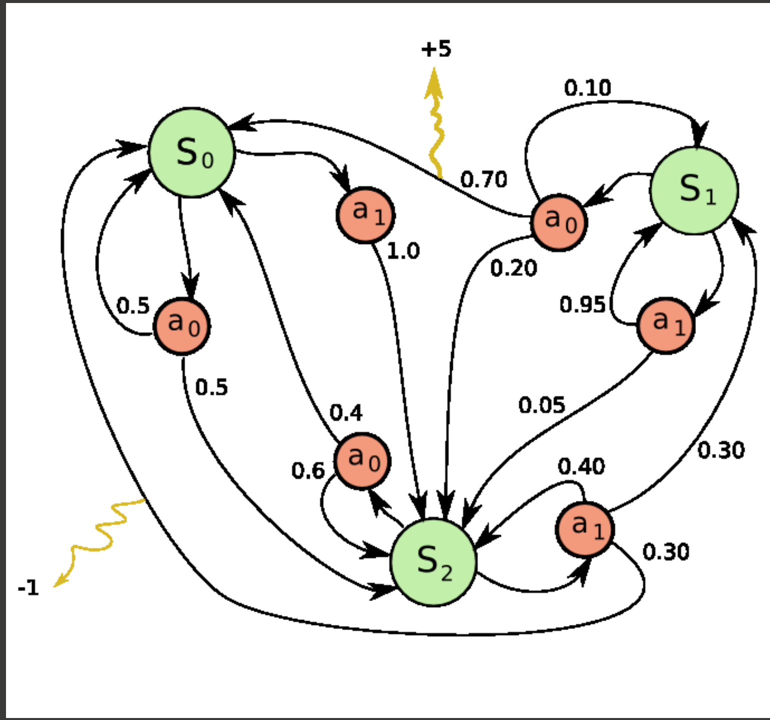
DeepThermal: Combustion Optimization for Thermal Power Generating Units Using Offline Reinforcement Learning

Presentation by Jacob Barkovitch

Background

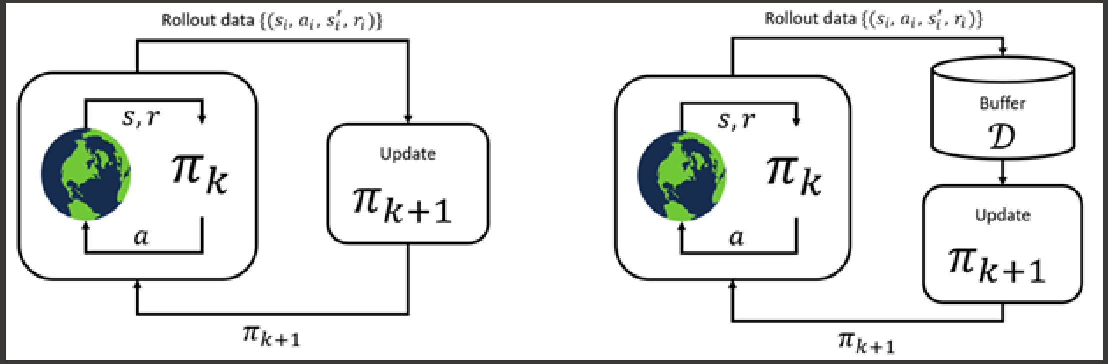
- Markov Decision Process (MDP)
  - Mathematical framework for modeling decision making
  - 4 tuple set (*S, A, P, R, y)* where:
    - S is set of states
    - A is set of actions
    - P is set of transition probabilities to the next state
    - R is reward after moving to the next state
    - y is discount factor applied on back-propagated future rewards



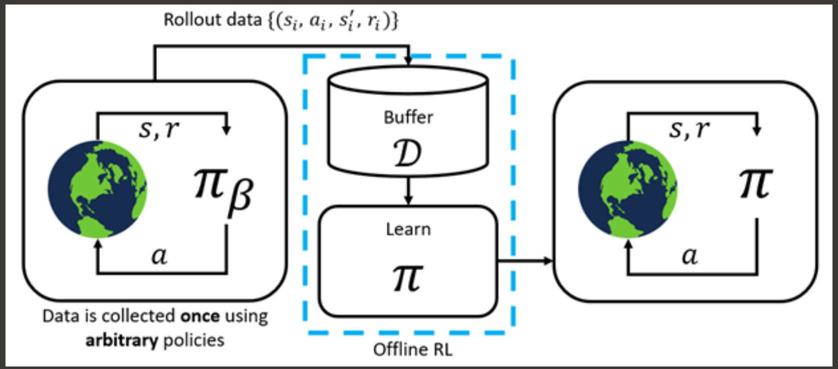
- Constrained Markov Decision Process (CMDP)
  - Adds an extra parameter *c*
    - Keeps costs of actions under a specified threshold

Background Cont.

- Reinforcement Learning (RL)
  - Agent that interacts with a MDP (environment)
  - Chooses actions based on probabilities
  - Updates probabilities based on rewards
  - Learns an optimal policy for the environment



- Offline Reinforcement Learning
  - Agent that interacts with only collected data from an MDP
  - Normally can only take actions that occurred in the data
  - Difficult to improve on the policy already present in the data



- Long Short-Term Memory Neural Network (LSTM)
  - Predicts future states based on time series data

An MDP Example

- Powerplant furnace environment with:
  - states: flame temp, efficiency, emissions, air intake, supplied fuel, heat demand
  - Actions: supply fuel (time), change air intake (%)
  - Rewards: limit fuel, lower emissions, keep temp high

- Normal RL approach would not work
  - RL involves learning from mistakes
  - Could blow up the furnace during training
  - Solution: offline approach on collected data

Time	RoofTem	EspTemp	EspTemp	AirHum	AirTemp	StackO2	EspOpac	AvgDraft	BoilEff	FanFlow	WaterFlo	HeatGen	Demand	DraftA	DraftB	DraftC	Shaker1	Shaker2	FlueOut	WaterTe	WaterTe	FlameTe	GatePos
3/28/2022 7:40	187.62	354.68	316.48	39.864	16.061	11.402	16.031	-0.5789	57.215	0.3129	1217.7	14.351	15.854	-0.1507	-0.1077	-0.1109	11.402	11.402	437.06	322.72	349.51	323.89	11.402
3/28/2022 7:40	187.57	354.61	316.49	39.919	16.061	11.402	16.009	-0.5786	57.385	0.3449	1234.2	14.378	15.836	-0.1507	-0.1077	-0.1109	11.402	11.402	437.07	322.71	349.52	323.89	11.402
3/28/2022 7:41	187.52	354.89	316.5	39.974	16.061	11.401	15.986	-0.5784	56.713	0.3769	1221.8	14.405	15.817	-0.1507	-0.1077	-0.1109	11.401	11.401	437.09	322.7	349.53	323.88	11.401
3/28/2022 7:41	187.47	354.89	316.52	39.973	16.061	11.401	15.963	-0.5782	56.733	0.4089	1228.7	14.431	15.798	-0.1507	-0.1077	-0.1109	11.401	11.401	437.11	322.69	349.54	323.88	11.401
3/28/2022 7:41	187.42	354.69	316.54	39.923	16.061	11.4	15.94	-0.5779	56.536	0.4409	1225.5	14.458	15.78	-0.1507	-0.1077	-0.1109	11.4	11.4	437.12	322.69	349.55	323.87	11.4
3/28/2022 7:41	187.37	354.17	316.55	39.873	16.061	11.4	15.918	-0.5777	56.421	0.4645	1221.6	14.484	15.761	-0.1507	-0.1077	-0.1109	11.4	11.4	437.14	322.68	349.56	323.87	11.4
3/28/2022 7:41	187.32	355.15	316.57	39.823	16.061	11.399	15.895	-0.5775	56.501	0.3447	1228	14.511	15.785	-0.1507	-0.1077	-0.1109	11.399	11.399	437.16	322.67	349.56	323.87	11.399

Solution

- DeepThermal
  - Propose Model-based Offline RL with Restrictive Exploration (MORE)
  - Offline RL approach to thermal power generating unit (TPGU) optimization
    - a constrained Markov decision process
    - Problem: not all possible actions are taken by human agents
  - Solve offline RL problems with a simulated system
    - LSTM trained on dataset to predict unknown future states!
    - Can approximate unseen states based on unseen actions

- MORE can learn optimal actions to lower harmful emissions!

Time	RoofTem	EspTemp	EspTemp	AirHum	AirTemp	StackO2	EspOpac	AvgDraft	BoilEff	FanFlow	WaterFlc	HeatGen	Demand	DraftA	DraftB	DraftC	Shaker1	Shaker2	FlueOut	WaterTe	WaterTe	FlameTe	GatePos
3/28/2022 7:40	187.62	354.68	316.48	39.864	16.061	11.402	16.031	-0.5789	57.215	0.3129	1217.7	14.351	15.854	-0.1507	-0.1077	-0.1109	11.402	11.402	437.06	322.72	349.51	323.89	11.402
3/28/2022 7:40	187.57	354.61	316.49	39.919	16.061	11.402	16.009	-0.5786	57.385	0.3449	1234.2	14.378	15.836	-0.1507	-0.1077	-0.1109	11.402	11.402	437.07	322.71	349.52	323.89	11.402
3/28/2022 7:41	187.52	354.89	316.5	39.974	16.061	11.401	15.986	-0.5784	56.713	0.3769	1221.8	14.405	15.817	-0.1507	-0.1077	-0.1109	11.401	11.401	437.09	322.7	349.53	323.88	11.401
3/28/2022 7:41	187.47	354.89	316.52	39.973	16.061	11.401	15.963	-0.5782	56.733	0.4089	1228.7	14.431	15.798	-0.1507	-0.1077	-0.1109	11.401	11.401	437.11	322.69	349.54	323.88	11.401
3/28/2022 7:41	187.42	354.69	316.54	39.923	16.061	11.4	15.94	-0.5779	56.536	0.4409	1225.5	14.458	15.78	-0.1507	-0.1077	-0.1109	11.4	11.4	437.12	322.69	349.55	323.87	11.4
3/28/2022 7:41	187.37	354.17	316.55	39.873	16.061	11.4	15.918	-0.5777	56.421	0.4645	1221.6	14.484	15.761	-0.1507	-0.1077	-0.1109	11.4	11.4	437.14	322.68	349.56	323.87	11.4
3/28/2022 7:41	187.32	355.15	316.57	39.823	16.061	11.399	15.895	-0.5775	56.501	0.3447	1228	14.511	15.785	-0.1507	-0.1077	-0.1109	11.399	11.399	437.16	322.67	349.56	323.87	11.399

Solution Details

- DeepThermal Constrained Markov Decision Process
  - States:
    - Chemical property of fuel (fuel strength)
    - Sensor data (temp, pressure, humidity, demand)
  - Actions:
    - Adjustments of control variables
    - Valves, intakes, baffles, shakers (continuous)
  - Rewards:
    - Increase efficiency *Effi*
    - Reduce NOx *Emi*
    - $r_t = a_r Effi_t + (1 - a_r) Emi_t$
  - Costs:
    - Safety constraints
    - Load, internal pressure, temperature

Solution Details Cont.

- Dataset
  - β = (s, a, s', r, c)
  - Generated by unknown behavior policies (humans)

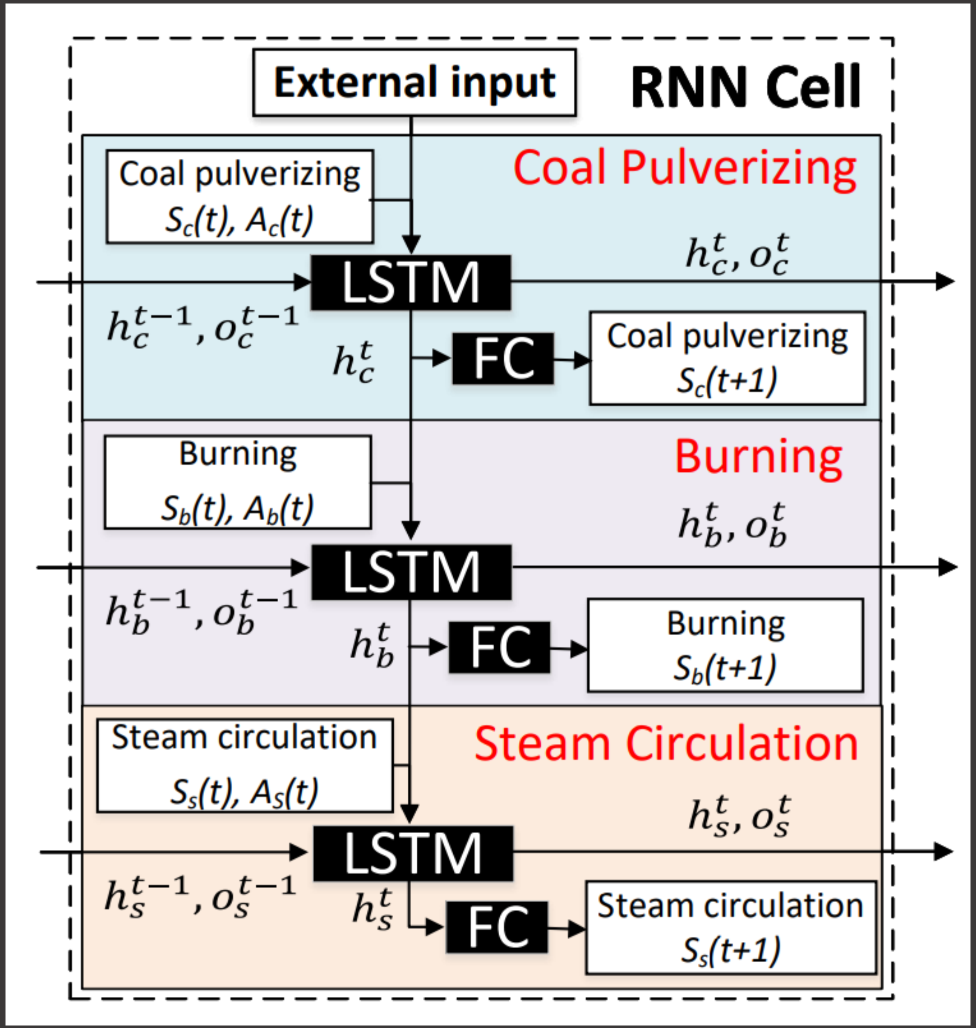
- Simulator
  - Add simulated data to allow exploration

- Goal:
  - Learn policy $\pi^*(s)$ from β that maximizes reward $R(\pi)$
  - Control costs $C(\pi)$ below a threshold l
  - $\pi^* = argmax\ R(\pi) \quad \text{s.t.} \quad C(\pi) \leq l$

Combustion Process Simulator

- Customized deep recurrent neural network
  - Structured like physical combustion process
  - LSTM to capture temporal correlations
  - Predicts future states
  - Techniques:
    - MSE, Seq2seq, Scheduled sampling, Data augmentation

- Drawbacks
  - Loses accuracy over increased time
  - Loses accuracy based on unseen actions



MORE Framework

- MORE policy optimization uses Q-functions
  - Q_r: reward maximization
  - Q_c: cost evaluation
  - Q(x): probability that a normal random variable takes a value larger than x

- Hybrid training approach
  - Mostly trained on original dataset
  - Occasionally interacts with simulator
    - Restrictive exploration
  - Reduces out-of-distribution states and actions

Algorithm 2: Complete algorithm of MORE

Require: Offline dataset $\mathcal{B}$

Pre-train actor π_θ, reward critic ensemble $\{Q_{r_i}(s,a|\phi_{r_i})\}_{i=1,2}$ and cost critic $Q_c(s,a|\phi_c)$ with real data. Initialize target networks $\{Q'_{r_i}\}_{i=1}^2$ and Q'_c with $\phi'_{r_i} \leftarrow \phi_{r_i}$ and $\phi'_c \leftarrow \phi_c$

for Training step: $t = 1, ..., T$ **do**

  Random sample mini-batch transitions τ_n from $\mathcal{B}$

  Obtain (τ^+, τ^-) using restrictive exploration (Alg. 1)

  Construct local buffer $\mathcal{R} = \{(s,a,r,c,s')\}$ using τ^+, τ^- and τ_n, as well as Eq.8

  Set $y = \min_{i=1,2} Q'_{r_i}(s', \pi(s'))$, $z = Q'_c(s', \pi(s'))$

  Update Q_{r_i} by minimizing $(Q_{r_i} - (r + \gamma y))^2$

  Update Q_c by minimizing $(Q_c - (c + \gamma z))^2$

  Update policy π_θ by Eq.3 using policy gradient

  Update λ by Eq.4 using dual gradient ascent

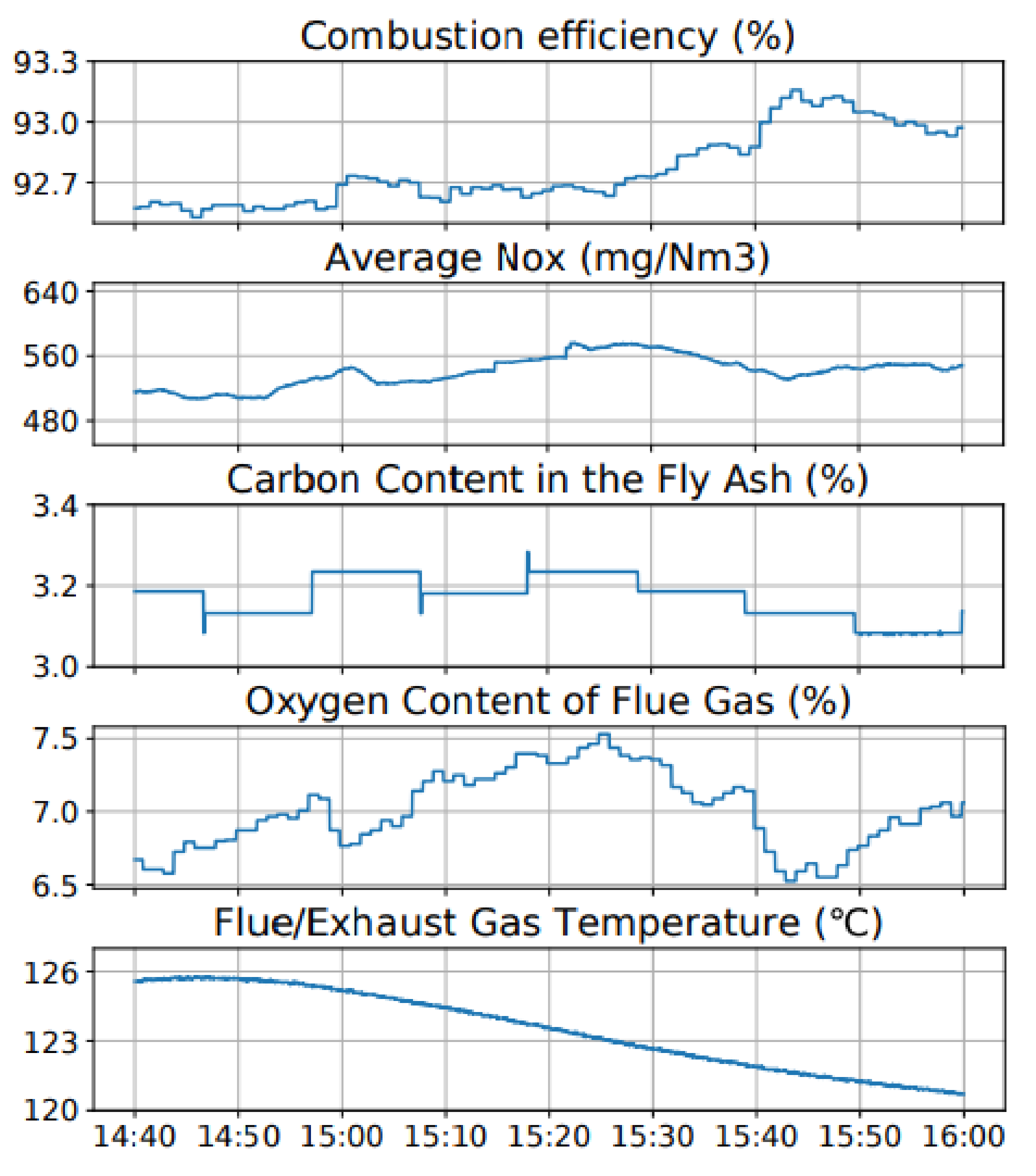
  Update target cost critic: $\phi'_c \leftarrow \rho\phi_c + (1-\rho)\phi'_c$

  Update target reward critics: $\phi'_{r_i} \leftarrow \rho\phi_{r_i} + (1-\rho)\phi'_{r_i}$

end for

Experiments

- Time-series simulator comparisons
  - ARIMA
  - GBRT
  - DNN
  - Stacked LSTM
- MORE vs human policy on TPGUs
  - With different load settings
- MORE vs state-of-the-art offline RL models
  - Standard offline RL benchmark D4RL (Fu et al. 2020)
  - Other models:
    - BCQ, BEAR, BRAC-v, MOPO, MBPO



Results

- Simulator results:

Model	ARIMA	GBRT	DNN	LSTM	Ours
RMSE	3.05e-1	1.97e-1	2.05e-2	1.69e-3	**6.54e-4**
MAE	2.66e-1	2.65e-1	2.73e-2	2.50e-2	**1.55e-3**

- MORE results

Dataset	Batch Mean	Batch Max	BC	BEAR	BRAC-v	BCQ	MBPO	MOPO	MORE (Ours)
halfcheetah-medium	3953	4410.7	4202.7	4513.0	5369.5	4767.9	3234.4	4972.3	**5970**
hopper-medium	1021.7	3254.3	924.1	**1674.5**	1031.4	**1752.4**	139.9	891.5	1264
walker2d-medium	498.4	3752.7	302.6	2717.0	**3733.4**	2441.3	582.8	817.0	**3649**
halfcheetah-mixed	2300.6	4834.2	4488.2	4215.1	5419.2	4463.9	5593.0	**6313.0**	5790
hopper-mixed	470.5	1377.9	364.4	331.9	9.7	688.7	1600.8	**2176.8**	**2100**
walker2d-mixed	358.4	1956.5	518.5	1161.4	36.2	1057.8	1019.1	1790.7	**1947**

Conclusion

- MORE outperforms state-of-the-art offline RL models
  - Leverages generalizability of imperfect models
  - Avoids exploitation errors on out-of-distribution samples

- DeepThermal is the first offline RL model deployed on real world tasks
  - More has been successfully deployed in four real world powerplants in China

- Technology could be used in many different mission-critical industries
  - Self driving cars
  - Robotics
  - Healthcare

References

- https://en.wikipedia.org/wiki/Markov_decision_process#:~:text=Constrained%20Markov%20decision%20processes%20(CMDPs,dynamic%20programming%20does%20not%20work
- https://towardsdatascience.com/the-power-of-offline-reinforcement-learning-5e3d3942421c
- https://ojs.aaai.org/index.php/AAAI/article/view/20393