Deep learning algorithms for stochastic control and applications to energy storage problems

Huyên PHAM*

*University Paris Diderot, LPSM

Joint work with
A. Bachrouf, University of Oslo
C. Huré, University Paris Diderot, LPSM
N. Langrené, CSIRO, Data61, Risklab Australia

USC, October 15, 2018

Discrete-time stochastic control on finite horizon

Markov Decision Process (MDP)

- **State process** $X = (X_n)_n$ in $\mathcal{X} \subset \mathbb{R}^d$, $n = 0, \ldots, N$

- Controlled by $\alpha = (\alpha_n)_n$ **action/policy**: $\alpha_n = \pi_n(X_n)$ for some measurable sequence $\pi_n : \mathcal{X} \to \mathbb{A}$, $n = 0, \ldots, N-1$.

- **State dynamics** in a random environment: $X = X^\alpha$

$$X_{n+1} \quad = \quad F_n(X_n, \alpha_n, \varepsilon_{n+1})$$

$\leftrightarrow$ One-step transition probabilities:

$$\begin{aligned} P_n^a(x, dx') &= \mathbb{P}\big[X_{n+1}^\alpha \in dx' | X_n^\alpha = x, \alpha_n = a\big] \\ &= \mathbb{P}\big[F_n(x, a, \varepsilon_{n+1}) \in dx'\big] \end{aligned}$$

- **Reward**: running reward $f_n(x, a)$ and terminal reward $g(x)$

Performance criterion

$$J_n(x,\alpha) \quad = \quad \mathbb{E}\Big[\sum_{k=n}^{N-1} f_n(X_n^\alpha, \alpha_n) + g(X_N^\alpha) \Big| X_n^\alpha = x \Big]$$

▶ **Goal**: Find optimal performance V and optimal action/policy $\alpha^* \leftrightarrow \pi^* = (\pi_n^*)_n$ valued in $\mathbb{A}^{\mathcal{X}}$:

$$V_n(x) \quad := \quad \sup_\alpha J_n(x,\alpha) \; = \; J_n(x,\alpha^*), \quad n = 0, \ldots, N, \; x \in \mathcal{X}.$$

Remark:

- MDP can also be viewed as time discretization of continuous-time stochastic control problem $\leftrightarrow$ Bellman PDE

Dynamic Programming (DP) Bellman equation

From global to local optimization: Backward recursion on $V = (V_n)$ (Value function iteration)

$$\begin{cases} V_N(x) & = & g(x) \\ V_n(x) & = & \sup\limits_{a \in \mathbb{A}} \Big\{ \underbrace{f_n(x,a) + \mathbb{E}\big[V_{n+1}(X^\alpha_{n+1}) \big| X^\alpha_n = x, \alpha_n = a\big]}_{Q_n(x,a) := f_n(x,a) + P^a_n V_{n+1}(x)} \Big\}, \quad n = N-1, \ldots, 0. \end{cases}$$

$\longrightarrow$ Optimal policy: $\pi^* = (\pi^*_n)_n$ from the Q-value function

$$\pi^*_n(x) \quad \in \quad \arg\max_{a \in \mathbb{A}} Q_n(x,a), \quad n = N-1, \ldots, 0$$

Remark.
$\{V_n(X^*_n) + \sum_{k=0}^n f(X^*_n, \alpha^*_n),\ n = 0, \ldots, N\}$, is a martingale:

$$V_n(x) \quad = \quad f_n(x, \pi^*_n(x)) + P_n^{\pi^*_n(x)} V_{n+1}(x).$$

Numerical challenges

Two sources of curses of dimensionality:

- Computations of the conditional expectation operator $P_n^a V_{n+1}(x)$, $n = 0, \ldots, N-1$, for any $x \in \mathcal{X} \subset \mathbb{R}^d$, and $a \in \mathbb{A}$. Computational complexity in **high-dimension for the state space** $\mathbb{R}^d$ and also the control space $\mathbb{A}$!
- Computation of the optimal policy: Supremum in $a \in \mathbb{A}$ of $Q_n(x, a) = f_n(x, a) + P_n^a V_{n+1}(x)$, for each $x \in \mathcal{X}$: $\rightarrow$ optimal policy $\hat{\pi}(x)$. Computational complexity in **high dimension for the control space** $\mathbb{A}$!

Probabilistic numerical methods based on DP

- **Approximate dynamic programming (ADP)**

(i) Approximate the Q-value function (conditional expectation) by Monte-Carlo regression on: basis functions, neural networks, SVM, etc

- MC regression in the spirit of Longstaff-Schwartz (for optimal stopping problems)
- Main issue: simulation of the endogenous controlled process

(ii) Optimal control is then "computed" from $\arg\max_{a \in \mathbb{A}} \hat{Q}_n(x, a)$ where $\hat{Q}$ is an approximation of the Q-value function. Typically:

- $\mathbb{A}$ finite set, or discretize $\mathbb{A}$
- Newton method for the search of extremum

Numerical methods by direct approximation (without DP)

• **Control approximation**: Focus directly on the (parametric) approximation $\pi = (\pi_n)$ of the policy on the whole period

$$\pi_n(x) \quad = \quad A(x;\theta_n), \quad n = 0, \ldots, N-1,$$

for some given function $A(.,\theta)$ with parameters $\theta = (\theta_0, \ldots, \theta_{N-1}) \in \mathbb{R}^{q\times N}$ $\rightarrow$ maximize over θ

$$J_0\big(x_0, A(.;\theta_0), \ldots, A(.;\theta_{N-1})\big) \quad = \quad \mathbb{E}\Big[\sum_{n=0}^{N-1} f_n(X_n, A(X_n;\theta_n)) + g(X_N)\Big].$$

- Kou, Peng, Xu (16): E-M algorithm with basis functions for A
- J. Han, W. E, A. Jentzen (17): Deep neural network (DNN) for A and global optimization by stochastic gradient descent (SGD), see also P. Henry-Labordère (18)

Our approach and contributions

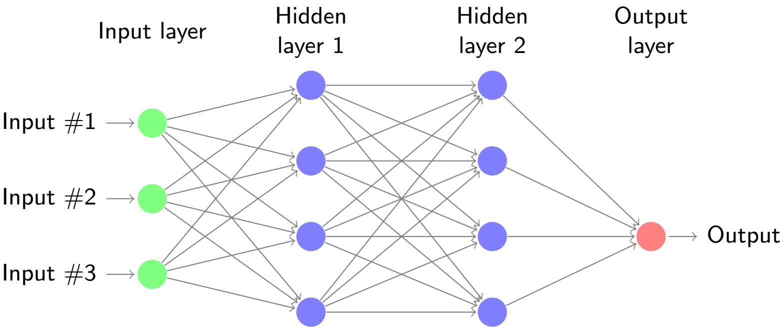
- Combine different ideas from maths (numerical probability) and computer science communities (reinforcement learning) to propose (and compare) three algorithms based on:
  - Dynamic programming (DP)
  - Deep Neural Networks (DNN) for the approximation/learning of
    - (i) Optimal policy
    - (ii) Value function
  - Monte-Carlo regressions with different characteristics:
    - Performance/policy iteration (PI) or hybrid iteration (HI)
    - Now or later/quantization

- Convergence analysis

- Numerical tests and an application to energy storage problems

Deep Neural networks (DNN): multilayer perceptron

Architecture of a DNN: composed of layers and neurons (units)



(Feedforward artificial NN)

Mathematical representation of DNN

• DNN: composition of simple functions to approximate complicated ones $\neq$ usual additive approximation theory

▶ Represented by parametrized function:

$$\begin{aligned} x \in \mathbb{R}^d \quad \mapsto \quad \Phi(x;\theta) &= \mathcal{L}^{out} \circ \mathcal{L}^L \circ \ldots \mathcal{L}^1(x), \\ \Phi_\ell \quad = \quad \mathcal{L}^\ell \Phi_{\ell-1} &:= \sigma(w_\ell \Phi_{\ell-1} + b_\ell) \in \mathbb{R}^{d_\ell}, \end{aligned}$$

with L hidden layers (layer ℓ with d_ℓ units), activation function σ (Sigmoid, ReLu, ELU, etc), and weights $\theta = (w_\ell, b_\ell)_\ell$.

Mathematical representation of DNN

• DNN: composition of simple functions to approximate complicated ones $\neq$ usual additive approximation theory

▶ Represented by parametrized function:

$$\begin{aligned} x \in \mathbb{R}^d \quad \mapsto \quad \Phi(x;\theta) &= \mathcal{L}^{out} \circ \mathcal{L}^L \circ \ldots \mathcal{L}^1(x), \\ \Phi_\ell \quad &= \quad \mathcal{L}^\ell \Phi_{\ell-1} := \sigma(w_\ell \Phi_{\ell-1} + b_\ell) \in \mathbb{R}^{d_\ell}, \end{aligned}$$

with L hidden layers (layer ℓ with d_ℓ units), activation function σ (Sigmoid, ReLu, ELU, etc), and weights $\theta = (w_\ell, b_\ell)_\ell$.

- Theoretical justification by universal approximation theorem (Hornick 91). Rate of convergence not yet well understood (partial results in the case of one hidden layer, see Bach 17).
- Key feature: automatic differentiation for computing derivatives of Φ used in SGD to find the "optimal" parameters.

Algo NNContPI: control learning by performance iteration

A combination of DP and Han, E, Jentzen algo:

• For $n = N-1, \ldots, 0$: keep track of the approximated optimal policies $\hat{\pi}_k$, $k = n+1, \ldots, N-1$, and compute

$$\hat{\pi}_n \quad \in \quad \arg\max_{\pi} \mathbb{E}\Big[f_n(X_n, \pi(X_n)) + \underbrace{\sum_{k=n+1}^{N-1} f_k(\hat{X}_k, \hat{\pi}_k(\hat{X}_k)) \; + g(\hat{X}_N)}_{\hat{Y}^{\pi}_{n+1}} \Big]$$

where $X_n \rightsquigarrow \mu$ (probability distribution on $\mathcal{X}$), $(\hat{X}_k)_{k=n+1,\ldots,N}$, generated from X_n, with control $(\pi, \hat{\pi}_k)_{k=n+1,\ldots,N-1}$. $\rightarrow$ Practical implementation:

- DNN for policy: $\pi(x) = A(x;\beta) \rightarrow$ optimization over parameter β
- SGD based on training samples $X_n^{(m)}$, $(\hat{X}_k^{(m)})_k$, $m = 1, \ldots, M \rightarrow \hat{\pi}_n^M = A(.;\hat{\beta}_n^M)$.

Algo NNContPI: control learning by performance iteration

A combination of DP and Han, E, Jentzen algo:

- For $n = N-1, \ldots, 0$: keep track of the approximated optimal policies $\hat{\pi}_k$, $k = n+1, \ldots, N-1$, and compute

$$\hat{\pi}_n \in \arg\max_{\pi} \mathbb{E}\Big[f_n(X_n, \pi(X_n)) + \underbrace{\sum_{k=n+1}^{N-1} f_k(\hat{X}_k, \hat{\pi}_k(\hat{X}_k)) + g(\hat{X}_N)}_{\hat{Y}^{\pi}_{n+1}} \Big]$$

where $X_n \rightsquigarrow \mu$ (probability distribution on $\mathcal{X}$), $(\hat{X}_k)_{k=n+1,\ldots,N}$, generated from X_n, with control $(\pi, \hat{\pi}_k)_{k=n+1,\ldots,N-1}$. $\rightarrow$ Practical implementation:

- DNN for policy: $\pi(x) = A(x;\beta) \rightarrow$ optimization over parameter β
- SGD based on training samples $X_n^{(m)}$, $(\hat{X}_k^{(m)})_k$, $m = 1, \ldots, M \rightarrow \hat{\pi}_n^M = A(.;\hat{\beta}_n^M)$.

Remarks.
1) **No value function iteration**: $\hat{V}_n^M$ is simply computed as the gain functional associated to controls $(\hat{\pi}_k^M)_{k=n,\ldots,N-1}$.
2) Low bias estimate, but possibly high variance estimate and large complexity, especially when N is large.

Algo Hybrid: control learning by hybrid iteration

- Initialization: $\hat{V}_N = g$
- For $n = N-1, \ldots, 0$:

(i) Compute the approximated optimal policy

$$\hat{\pi}_n \quad \in \quad \arg\max_{\pi} \mathbb{E}\big[f_n(X_n, \pi(X_n)) + \hat{V}_{n+1}(X^{\pi}_{n+1})\big]$$

where $X_n \rightsquigarrow \mu$, $X^{\pi}_{n+1} \rightsquigarrow P_n^{\pi(X_n)}(X_n, dx')$. Implemented by

- DNN for policy: $\pi(x) = A(x; \beta) \rightarrow$ optimization over parameter β
- SGD method based on training samples $X_n^{(m)}$, $m = 1, \ldots, M \rightarrow \hat{\pi}_n^M = A(.; \hat{\beta}_n^M)$.

Algo Hybrid: control learning by hybrid iteration

- Initialization: $\hat{V}_N = g$
- For $n = N-1, \ldots, 0$:

(i) Compute the approximated optimal policy

$$\hat{\pi}_n \quad \in \quad \arg\max_{\pi} \mathbb{E}\big[f_n(X_n, \pi(X_n)) + \hat{V}_{n+1}(X_{n+1}^{\pi})\big]$$

where $X_n \rightsquigarrow \mu$, $X_{n+1}^{\pi} \rightsquigarrow P_n^{\pi(X_n)}(X_n, dx')$. Implemented by

- DNN for policy: $\pi(x) = A(x;\beta) \rightarrow$ optimization over parameter β
- SGD method based on training samples $X_n^{(m)}$, $m = 1, \ldots, M \rightarrow \hat{\pi}_n^M = A(.;\hat{\beta}_n^M)$.

(ii) Updating: compute the approximated value function

$$\begin{aligned}\hat{V}_n(x) &= \mathbb{E}\Big[f_n(X_n, \hat{\pi}_n(X_n)) + \hat{V}_{n+1}(X_{n+1}^{\hat{\pi}_n})\Big|X_n = x\Big] \\ &= f_n(x, \hat{\pi}_n(x)) + P_n^{\hat{\pi}_n(x)}\hat{V}_{n+1}(x)\end{aligned}$$

by Monte-Carlo regression: now or later

Algo Hybrid-Now

Regress now on a set $\mathcal{F}$ of functions on $\mathcal{X}$ (from $n+1$ to n)

$$\hat{V}_n \quad \in \quad \arg\min_{\Phi\in\mathcal{F}} \mathbb{E}\Big| f_n(X_n, \hat{\pi}_n(X_n)) + \hat{V}_{n+1}(X_{n+1}^{\hat{\pi}_n}) \; - \; \Phi(X_n)\Big|^2$$

- For instance, $\mathcal{F}$ class of DNN: $x \mapsto \Psi(x;\theta)$
- Optimization over θ by SGD based on training samples $X_n^{(m)} \rightsquigarrow \mu$, $m = 1, \ldots, M$, $\rightarrow$ $\hat{V}_n^M = \Psi(.;\hat{\theta}_n^M)$.

Algo Hybrid-LaterQ

(a) Later (at time $n+1$) interpolation of the value function:

$$\tilde{V}_{n+1} \quad \in \quad \arg\min_{\Phi \in \mathcal{F}} \mathbb{E}\Big[\ell\big(\hat{V}_{n+1}(X_{n+1}^{\hat{\pi}_n}) \; - \; \Phi(X_{n+1}^{\hat{\pi}_n})\big)\Big]$$

for some loss function ℓ on $\mathbb{R}$, e.g., $\ell(y) = y^2$, and for instance, $\mathcal{F}$ class of DNN: $x \mapsto \Psi(x;\theta)$.

(b) Update the value function at time n by approximating **analytically by quantization** the conditional expectation

$$\begin{aligned} \hat{V}_n(X_n) \quad &:= \quad f_n(X_n, \hat{\pi}_n(X_n)) + \hat{P}_n^{\hat{\pi}_n(X_n)} \tilde{V}_{n+1}(X_n) \\ &:= \quad f_n(X_n, \hat{\pi}_n(X_n)) + \sum_{j=1}^{J} p_j \hat{V}_{n+1}\big(F_n(X_n, \hat{\pi}_n(X_n), e_j)\big) \end{aligned}$$

where $\hat{\varepsilon}_{n+1} \rightsquigarrow \sum_{j=1}^{J} p_j \delta_{e_j}$ is a J-quantizer of ε_{n+1}.

Remark. Compared to Regress now, Regress Later MC reduces the variance of the estimated $\hat{V}_n^M$.

Case of finite control space: classification

- Card($\mathbb{A}$) $= L < \infty$: $\mathbb{A} = \{a_1, \ldots, a_L\}$

- Randomize the control: given a state value x, the controller chooses a_ℓ with a probability $p_\ell(x)$
  - (Deep) Neural Network for the probability vector $p = (p_\ell)_\ell$ with softmax output layer:

$$z \quad \longmapsto \quad p_\ell(z;\beta) \;=\; \frac{\exp(\beta_\ell.z)}{\sum_{\ell=1}^{L}\exp(\beta_\ell.z)}, \quad \ell = 1, \ldots, L.$$

  - Optimization over the probability vector p via the parameter β

Remark. In practice, we then use pure control strategies: given a state value x, choose $a_{\ell^*(x)}$ with

$$\ell^*(x) \quad \in \quad \arg\max_{\ell=1,\ldots,L} p_\ell(x).$$

Convergence of the algo NNcontPI

- M number of training samples
- Neural Network for policy:
  - $\mathcal{A}_K^\gamma$: class of NN with one hidden layer, K **neurons**, and **total variation norm** smaller than γ

Theorem. Under suitable conditions, and assuming the existence of an optimal policy $(\pi_k^*)_k$, we have for all $n = 0, \ldots, N-1$:

$$\mathbb{E}_M \big| V_n(X_n) - \hat{V}_n^M(X_n) \big| = \mathcal{O}_{\mathbb{P}} \Big(\gamma_M^{N-n} \sqrt{\frac{\ln M}{M}} + \underbrace{\sup_{k=n,\ldots,N-1} \inf_{A \in \mathcal{A}_K^\gamma} \big\| A(X_k) - \pi_k^*(X_k) \big\|_{L^1}}_{\varepsilon_n^{NN}(A)} \Big), \tag{1}$$

where $\mathbb{E}_M$ stands for the expectation conditioned on the training set used for computing the approximated optimal policies $\hat{\pi}_k^M$, and $(X_k)_k$ is the corresponding controlled process starting from $X_n \rightsquigarrow \mu$.

Proof. Arguments from statistical learning theory: Györfi et al. (02)

Convergence of the algo Hybrid

- M number of training samples
- Neural Network for policy and value function:
  - $\mathcal{A}_K^\gamma$: class of NN (valued in $\mathbb{A}$) with one hidden layer, K **neurons**, and **total variation norm** smaller than γ
  - $\mathcal{F}_K^\gamma$: class of NN (valued in $\mathbb{R}$) with one hidden layer, K **neurons**, and **total variation norm** smaller than γ

Theorem. Under suitable conditions, and assuming the existence of an optimal policy $(\pi_k^*)_k$, we have for all $n = 0, \ldots, N-1$:

$$\mathbb{E}_M\left|V_n(X_n) - \hat{V}_n^M(X_n)\right| = \mathcal{O}_{\mathbb{P}}\Bigg(\left(\gamma^4 K \frac{\ln M}{M}\right)^{\frac{1}{2(N-n)}}$$

$$+ \underbrace{\sup_{k=n,\ldots,N-1} \inf_{A \in \mathcal{A}_K^\gamma} \left\| A(X_k) - \pi_k^*(X_k) \right\|_{L^1}^{\frac{1}{2(N-n)}}}_{\varepsilon_n^{NN}(A)} + \underbrace{\sup_{k=n,\ldots,N} \inf_{\Psi \in \mathcal{F}_K^\gamma} \left\| \Psi(X_k) - V_k(X_k) \right\|_{L^2}^{\frac{1}{2(N-n)}}}_{\varepsilon_n^{NN}(V)} \Bigg).$$

Convergence of the algo LaterQ

- M number of training samples
- Neural Network for policy and value function:
  - $\mathcal{A}_K^\gamma$: class of NN (valued in $\mathbb{A}$) with one hidden layer, K **neurons**, and **total variation norm** smaller than γ
  - $\mathcal{F}_K^\gamma$: class of NN (valued in $\mathbb{R}$) with one hidden layer, K **neurons**, and **total variation norm** smaller than γ
  - L points for the **quantization** of the exogenous noise ε.

Theorem. Under suitable conditions, and assuming the existence of an optimal policy $(\pi_k^*)_k$, we have for all $n = 0, \ldots, N-1$:

$$\mathbb{E}_M\big|V_n(X_n) - \hat{V}_n^M(X_n)\big| = \mathcal{O}_{\mathbb{P}}\Bigg(\gamma^2\sqrt{K\frac{\ln M}{M}} + \frac{\gamma}{L^{1/d}}$$

$$+ \underbrace{\sup_{k=n,\ldots,N-1}\inf_{A\in\mathcal{A}_K^\gamma}\big\|A(X_k) - \pi_k^*(X_k)\big\|_{L^1}}_{\varepsilon_n^{NN}(A)} + \underbrace{\sup_{k=n,\ldots,N}\inf_{\Psi\in\mathcal{F}_K^\gamma}\big\|\Psi(X_k) - V_k(X_k)\big\|_{L^2}}_{\varepsilon_n^{NN}(V)}\Bigg).$$

A semi-linear PDE with quadratic gradient term

$$\begin{cases} \frac{\partial v}{\partial t} + \Delta_x v - |D_x v|^2 &= 0, \quad (t,x) \in [0,T) \times \mathbb{R}^d \\ v(T,x) &= g(x) \end{cases}$$

This PDE can be written as an HJB equation associated to a stochastic control problem whose discrete-time version (time step $h = T/N$) is:

$$\begin{aligned} V_0(x_0) &= \inf_{\alpha} \mathbb{E}\Big[\sum_{n=0}^{N-1} |\alpha_n|^2 h \; + \; g(X_N^\alpha) \Big] \\ X_{n+1}^\alpha &= X_n^\alpha + 2\alpha_n h + \sqrt{2}\,\Delta W_{(n+1)h}, \quad X_0^\alpha = x_0. \end{aligned}$$

$\rightarrow$ Explicit solution (via Hopf-Cole transformation):

$$V_0(x_0) \;=\; -\ln\Big(\mathbb{E}\Big[\exp\big(-g(x_0 + \sqrt{2}W_T)\big)\Big]\Big).$$

Implementation

Algo Hybrid-Now

- $N = 30$ time steps, $T = 1$, $h = 1/30$.

- DNN for policy (resp. value function): function from $\mathcal{X} = \mathbb{R}^d$ into $\mathbb{A} = \mathbb{R}^d$ (resp. $\mathbb{R}$):
  - Input layer with d neurons
  - 3 hidden layers with $d + 10$ neurons each
  - Output layer with d neurons (resp. 1 neuron)

- Exponential Linear Unit (ELU) activation function

- Optimization with Adam method in TensorFlow
  - Training distribution $\mu \rightsquigarrow \mathcal{N}\big(x_0, \sqrt{2h} I_d\big)$.

Test 1a: from Han, E, Jentzen (17)

- $d = 100$, $g(x) = \ln\left(\frac{1}{2}(1+|x|^2)\right)$.

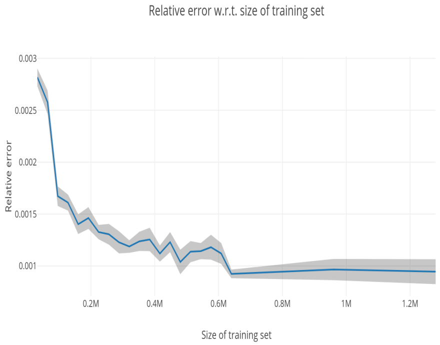


Figure: Relative error of the Algo Hybrid-Now for $V_0(x_0 = 0)$.
RelError $= 0.092\%$; Standard deviation of $\hat{V}_0^M(0) = 0.00191\%$

Test 1b: comparison with quadratic BSDE methods (Richou's thesis)

• $d = 1$, $g(x) = -x^p 1_{0 \leq x \leq 1} - 1_{1 < x}$, $p \in (0, 1]$ (Non-Lipschitz at $x = 0$ when $p < 1$)

- $N = 40$, 3 hidden layers with $10 + 5 + 5$ neurons
- Estimation of $V_0(0)$:

p	Richou	Hybrid-LaterQ	Hybrid-Now	Bench
1	- 0.402	- 0.456	-0.460	-0.464
0.5	-0.466	- 0.495	- 0.507	-0.509

Model setup

Real-options valuation of gas storage (discrete-time version of the Carmona-Ludkovski model)

- Gas (random) price $(P_n)_n$
- Gas inventory $(C_n)_n$ controlled by the decision α_n to inject, do nothing, or withdraw gas:

$$C_{n+1} = \begin{cases} C_n + b_{in} & \text{if } \alpha_n = +1 \quad \text{(injection/buy gas)} \\ C_n & \text{if } \alpha_n = 0 \quad \text{(do nothing)} \\ C_n - s_{out} & \text{if } \alpha_n = -1 \quad \text{(withdraw/sell gas)} \end{cases}$$

with b_{in}, $s_{out} > 0$.

- Physical inventory constraint:

$$C_n \in [C_{min}, C_{max}].$$

Control problem

- Maximize over α on finite horizon N:

$$\mathbb{E}\Big[\sum_{k=0}^{N-1} f(P_n, C_n, \alpha_n) \; + \; g(P_N, C_N)\Big]$$

with

- Revenue at any time n:

$$f(p,c,a) \quad = \quad \begin{cases} -b_{in}p - \kappa c & \text{if } a \; = \; +1 \quad \text{(injection/buy gas)} \\ -\kappa c & \text{if } a \; = \; 0 \quad \text{(do nothing)} \\ s_{out}p - \kappa c & \text{if } a \; = \; -1 \quad \text{(withdraw/sell gas)} \end{cases}$$

  with storage cost $\kappa > 0$.

- Terminal condition: penalization for having less gas than initially

$$g(p,c) \quad = \quad -\mu p(C_0 - c)_+.$$

  with $\mu > 0$.

Numerical results

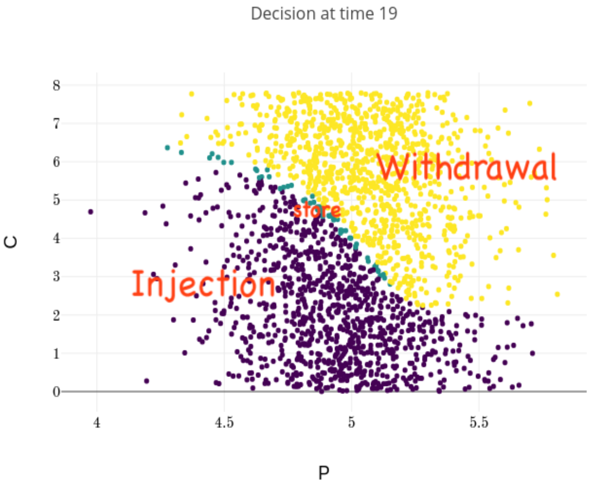
- **Model parameters**:
  - Mean-reverting gas price around $\bar{p} = 5$, with rate $\beta = 0.5$

$$P_{n+1} \quad = \quad \bar{p}(1-\beta) + \beta P_n + \xi_{n+1}, \;\; \xi_n \rightsquigarrow \mathcal{N}(0, \sigma^2 = 0.05), \;\; P_0 = 4,$$

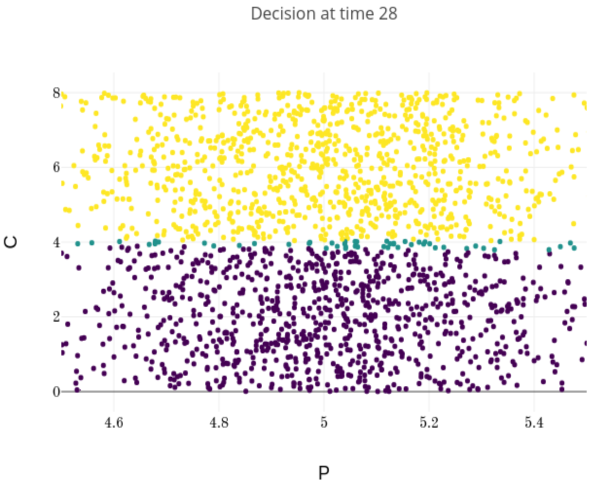
  - $b_{in} = 0.06, \; s_{out} = 0.25, \; \kappa = 0.01,$
  - $N = 30, \; \mu = 2, \; C_0 = 4, \; C_{min} = 0, \; C_{max} = 8.$

- Implementation by Algo NNContPI with DNN classification:
  - 3 hidden layers with $15 + 15 + 5$ neurons, output layer with 3 neurons
  - ELU activation function
  - Training samples of size $M = 250000$

Optimal policy regions



Optimal policy regions near maturity



Model description

Model from Heynmann et al. (17), see also Alasseur et al. (18)

- Microgrid for satisfying a power demand (D_n):
  - Photovoltaic (PV) $\rightarrow$ intermittent electricity production $(P_n)_n$
  - Generator (G) $\rightarrow$ power control α_n when turn on
  - Battery storage with capacity (C_n) in $[0, C_{max}]$
- Residual demand $R_n = D_n - P_n$:
  - $R_n < 0$: one can store the surplus power in the battery for later use
  - $R_n > 0$: one should provide power through diesel or battery

Control problem

Minimize over α valued in $\{0\} \times [A_{min}, A_{max}]$

$$\mathbb{E}\Big[\sum_{n=0}^{N-1} |\alpha_n|^2 + \kappa 1_{M_n \neq M_{n-1}}\Big] \quad \text{subject to satisfying demand constraint,}$$

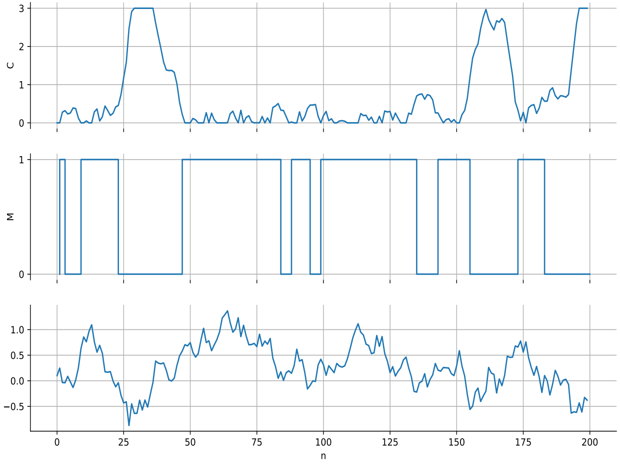
where

- $(M_n)_n$ mode of the diesel generator: 1 if turn on, 0 if turn off,
- $\kappa > 0$: switching cost for turning on/off the generator.

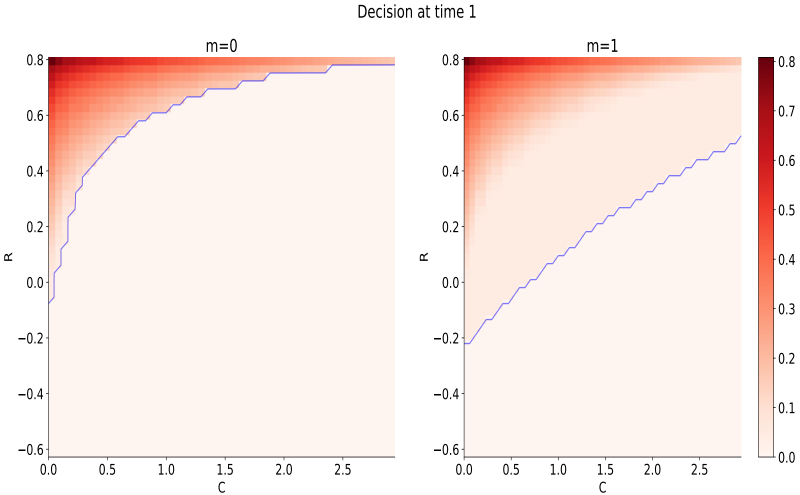
Numerical results

- **Model parameters**
  - Mean-reverting model for (R_n) around $\bar{R} = 0.1$, $R_0 = 0.1$
  - $A_{min} = 0.05$, $A_{max} = 10$, $C_{max} = 3$, $C_0 = 0$ s
  - switching cost $\kappa = 0.1$
  - $N = 200$,
- Implementation by Algo NNContPI:
  - 3 hidden layers with $100 + 50 + 50$ neurons,
  - ELU activation function
  - Training samples of size $M = 2^{16}$

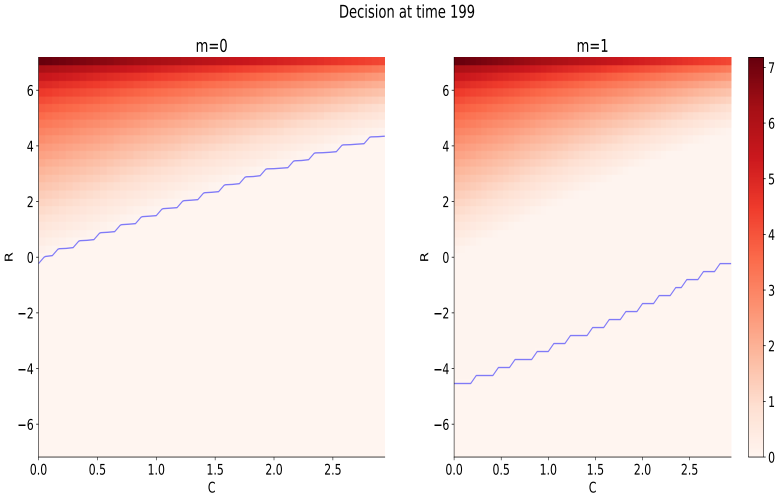
Evolution of battery capacity and diesel mode for one trajectory of residual demand



Diesel generator policy at time $n = 1$



Diesel generator policy near maturity $n = 199$



Concluding remarks

- Machine learning meets stochastic control
  - Neural network regression
  - Control learning
- We analyzed and tested three algorithms

Algo	Bias estimate	Variance	Complexity	Dimension	Time steps
NNContPI	+	-	-	+	−
Hybrid-Now	-	+	+	+	+
Hybrid-LaterQ	-	++	+	-	++

- Future work:
  - Extension to mean-field control problems ...