Data-driven Coordination of Distributed Energy Resources

Alejandro D. Domínguez-García

Coordinated Science Laboratory
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

PSERC Webinar Series
November 5, 2019

Outline

Impacts of Renewable-Based Power Generation Resources

- Deep penetration of renewable-based generation imposes additional requirements on ancillary services including:
  - Frequency regulation (in bulk power systems)
  - Reactive power support (in distribution systems)
- Frequency regulation in bulk power systems is typically achieved by controlling large synchronous generators
  - Resources in distribution systems are not utilized for this task
- Reactive power support in distribution systems is provided by devices such as load tap changers (LTCs) and fixed/switched capacitors
  - These devices are not designed to manage high variability in voltage fluctuations induced by renewable-based generation

The Solution

- An increasing number of DERs are being integrated into distribution systems
- DERs could potentially be utilized to provide ancillary services if properly coordinated by, e.g., an aggregator

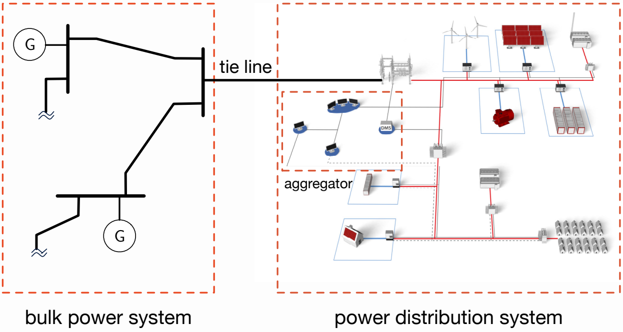
PV systems

Electric Vehicles

Fuel Cells

Residential Storage

Need for Data-Driven Coordination



- DER aggregators needs to develop appropriate coordination schemes so DERs can collectively provide services that meet certain requirements
- Model-based schemes may be infeasible due to the lack of accurate models
- Data-driven schemes that only rely on measurements provide a promising alternative for developing efficient coordination schemes

Presentation Overview

Objective

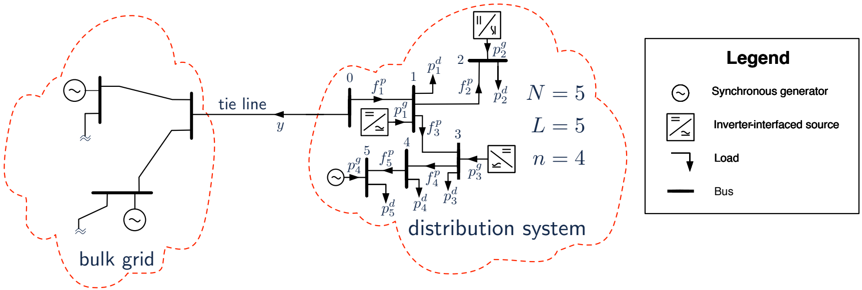
To develop data-driven coordination frameworks for assets (DERs, LTCs) in distribution systems

- **Part I. Active power provision problem**
  - total active power exchanged between the distribution and bulk systems needs to equal to some amount requested by the bulk system operator

- **Part II. Voltage regulation problem**
  - the voltage magnitude at each bus needs be maintained to stay close to some reference value

Outline

Optimal DER Coordination Problem (ODCP)



Determine the DER active power injection vector, $\boldsymbol{p}^g$, that minimizes total cost of operation while satisfying:

C1. The power exchanged with the bulk system, y, tracks some pre-specified value, $y^\star$

C2. The active power injection from each DER does not exceed its corresponding capacity limits, i.e., $\underline{\boldsymbol{p}}^g \leq \boldsymbol{p}^g \leq \overline{\boldsymbol{p}}^g$

C3. The power flow on each line does not exceed its maximum capacity, i.e., $-\overline{\boldsymbol{f}} \leq \boldsymbol{f} \leq \overline{\boldsymbol{f}}$

Input-Output System Model

- y can be written as a function of $\boldsymbol{p}^g, \boldsymbol{p}^d, \boldsymbol{q}^d$ as follows:

$$y = h(\boldsymbol{p}^g, \boldsymbol{p}^d, \boldsymbol{q}^d)$$

- h captures the impacts from both the physical laws as well as the effect of any reactive power control scheme

Assumption 1

H1. The rate of change in y w.r.t. $\boldsymbol{p}^g$ is bounded for bounded changes in the DER active power injections

H2. The total active power provided to the bulk power system will increase when more active power is injected in the power distribution system

ODCP Formulation

- The ODCP can be formulated as:

$$\underset{\boldsymbol{p}^g \in [\underline{\boldsymbol{p}}^g, \overline{\boldsymbol{p}}^g]}{\text{minimize}} \; c(\boldsymbol{p}^g)$$

subject to

$$h(\boldsymbol{p}^g, \boldsymbol{p}^d, \boldsymbol{q}^d) = y^\star,$$
$$-\overline{\boldsymbol{f}} \leq \boldsymbol{M}^{-1}(\boldsymbol{C}\boldsymbol{p}^g - \boldsymbol{p}^d) \leq \overline{\boldsymbol{f}}$$

$\boldsymbol{p}^g$	DER active power injections
$\underline{\boldsymbol{p}}^g, \overline{\boldsymbol{p}}^g$	DER upper and lower capacity limits
$\boldsymbol{p}^d$, $\boldsymbol{q}^d$	Load active, reactive power demands
$y^\star$	Requested power to be exchanged with bulk grid
$\overline{\boldsymbol{f}}$	Line flow limits
$\boldsymbol{M}$	Reduced node-to-edge incidence matrix
$\boldsymbol{C}$	Matrix mapping DER indices to buses

Data-Driven DER Coordination Framework

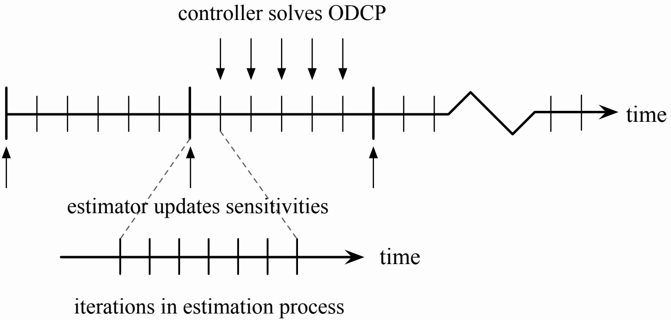
- Data defining the ODCP problem:
  - Cost function, $c(\cdot)$
  - DER capacity limits, $\overline{\boldsymbol{p}}^g, \underline{\boldsymbol{p}}^g$
  - Network topology and DER location, $\boldsymbol{M}, \boldsymbol{C}$
  - Load active and reactive power demand, $\boldsymbol{p}^d, \boldsymbol{q}^d$
  - Line flow limits, $\overline{\boldsymbol{f}}$
  - Input-output model, $h(\cdot, \cdot, \cdot) \leftarrow$ Assumed unknown

- Real-time measurements available:
  - DER active power injections, $\boldsymbol{p}^g[k],\ k = 1, 2, \ldots$
  - Active power exchanged with the bulk system, $\boldsymbol{y}[k],\ k = 1, 2, \ldots$

Framework Building Blocks

- An input-output (IO) model estimator that uses available real-time measurement data
- A controller that uses uses the identified IO model to solve the ODCP

Two-Timescale Coordination Framework



Estimation Process

$\boldsymbol{p}^d$ and $\boldsymbol{q}^d$ remain approximately constant between two time instants; therefore, changes in $y[k]$ that occur across time steps in the estimation process depend only on changes in $\boldsymbol{p}^g[k]$; thus,

$$y[k] = h(\boldsymbol{p}^g[k], \boldsymbol{p}^d, \boldsymbol{q}^d), \quad k = 0, 1, \ldots$$

Input-Output Model as a Linear Time-Varying System

- For notational simplicity, define $\boldsymbol{u}[k] = \boldsymbol{p}^g[k]$, $\underline{\boldsymbol{u}} = \underline{\boldsymbol{p}}^g$, $\overline{\boldsymbol{u}} = \overline{\boldsymbol{p}}^g$, and $\boldsymbol{\pi} = [(\boldsymbol{p}^d)^\top, (\boldsymbol{q}^d)^\top]^\top$; then, the IO model can be written as:

$$y[k] = h(\boldsymbol{u}[k], \boldsymbol{\pi}), \quad k = 0, 1, \dots,$$

- For $k > 1$, the above equation can be transformed into the following equivalent linear time-varying model:

$$y[k] = y[k-1] + \boldsymbol{\phi}[k]^\top (\boldsymbol{u}[k] - \boldsymbol{u}[k-1])$$

where $\boldsymbol{\phi}[k]^\top = [\phi_i[k]] = \left.\dfrac{\partial h}{\partial \boldsymbol{u}}\right|_{\tilde{\boldsymbol{u}}[k]}$, with $\tilde{\boldsymbol{u}}[k] = a_k \boldsymbol{u}[k] + (1 - a_k)\boldsymbol{u}[k]$

- $\boldsymbol{\phi}[k]$ is referred to as the sensitivity vector at time step k

- The entries of $\boldsymbol{\phi}[k]$ are bounded for all k by Assumption 1

Estimator – Estimation Step

- At time step k, the objective of the estimator is to obtain an estimate of $\boldsymbol{\phi}[k]$, denoted by $\hat{\boldsymbol{\phi}}[k]$, using available measurements of $\boldsymbol{u}$ and y
- Problem P1:

$$\hat{\boldsymbol{\phi}}[k] = \underset{\hat{\boldsymbol{\phi}} \in \mathcal{Q} = [\underline{b}_1, \overline{b}_1]^n}{\arg\min} \quad J^e(\hat{\boldsymbol{\phi}}) = \frac{1}{2}(y[k-1] - \hat{y}[k-1])^2$$

subject to

$$\hat{y}[k-1] = y[k-2] + \hat{\boldsymbol{\phi}}^\top (\boldsymbol{u}[k-1] - \boldsymbol{u}[k-2])$$

Estimator – Control Step

- The objective of the controller during the estimation process is to ensure that the output tracks the target [Different from the ODCP]
- Problem P2:

$$\boldsymbol{u}[k] = \underset{\boldsymbol{u}\in\mathcal{U}=[\underline{\boldsymbol{u}},\overline{\boldsymbol{u}}]}{\arg\min} \; J^c(\boldsymbol{u}) = \frac{1}{2}(y^\star - \hat{y}[k])^2$$

subject to

$$\hat{y}[k] = y[k-1] + \hat{\boldsymbol{\phi}}[k]^\top(\boldsymbol{u} - \boldsymbol{u}[k-1])$$

- Note that $\hat{\boldsymbol{\phi}}[k]$ is used to predict the value of $y[k]$ for a given $\boldsymbol{u}$

Estimation Process Workflow

$$\cdots \boldsymbol{u}[k-1] \rightarrow \underbrace{\boldsymbol{y}[k-1] \rightarrow \hat{\boldsymbol{\phi}}}_{\text{estimation step}}\overbrace{[k] \rightarrow \boldsymbol{u}[k]}^{\text{control step}} \rightarrow \boldsymbol{y}[k] \rightarrow \hat{\boldsymbol{\phi}}[k+1] \cdots$$

- At the beginning of iteration k, $y[k-1]$ is used in Problem P1 to update the sensitivity vector estimate, $\hat{\boldsymbol{\phi}}[k]$
- The updated sensitivity vector estimate, $\hat{\boldsymbol{\phi}}[k]$, is then used in Problem P2 to determine the control, $\boldsymbol{u}[k]$
- Then, the DERs are instructed to change their active power injection set-points based on $\boldsymbol{u}[k]$
- Problems P1 and P2 are not solved to completion for each k
- Instead, we iterate the projected gradient descent algorithm that would solve them for one step at each iteration k

Simulation Setup

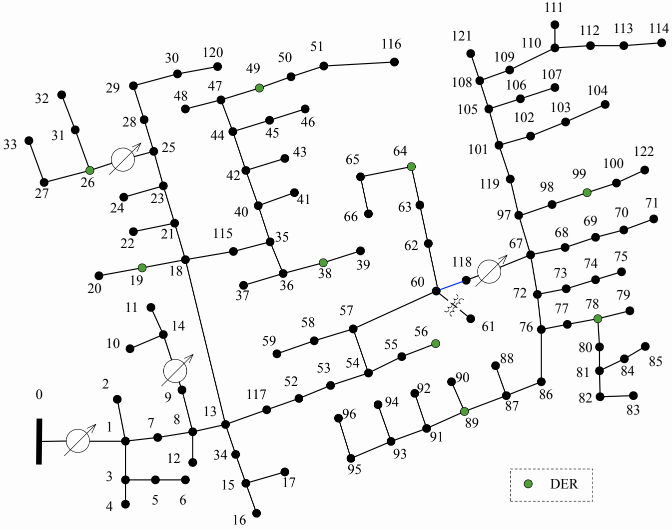


Figure 1: The IEEE 123-bus distribution test feeder.

Tracking Performance During Estimation

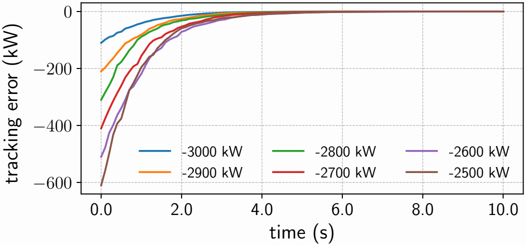


Figure 2: Tracking error for $\beta_k = 0.02$ under various tracking targets.

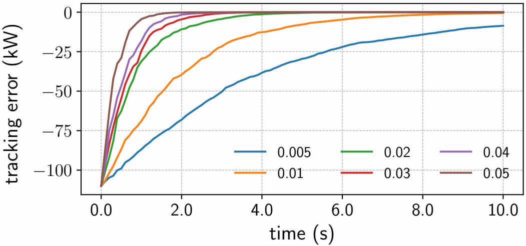


Figure 3: Tracking error for $y^\star = -3000$ kW and various constant control step sizes.

Estimation Accuracy

- Mean absolute error (MAE) of estimation errors:

$$\text{MAE} = \frac{\sum_{i=1}^{n} \left| \hat{\phi}_i[k] - \phi_i[k] \right|}{n},$$

where n is the number of DERs

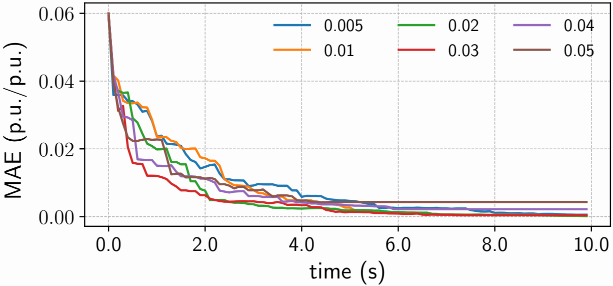


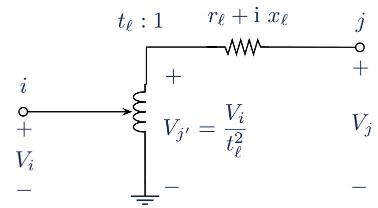
Figure 4: Estimation error under various control step sizes.

Outline

Background

- Voltage regulation transformers—also referred to as Load Tap Changers (LTCs)—are widely utilized in power distribution systems to regulate voltage magnitudes along a feeder

- Model for a load tap changer on a line connecting buses i and j:



V_i	Primary side voltage magnitude
$V_{j'}$	Secondary side voltage magnitude
V_j	Line receiving end voltage magnitude
t_ℓ	Tap ratio
r_ℓ	Transformer + line resistance
x_ℓ	Transformer + line reactance

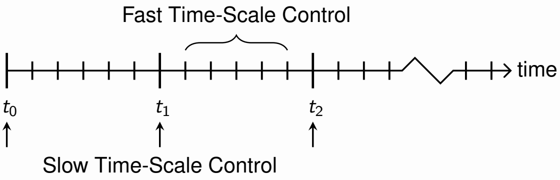
- The tap ratio, t_l, typically takes 33 discrete values ranging from 0.9 to 1.1, by an increment of $5/8\%$, i.e.,

$$t_l \in \mathcal{T} = \{0.9, 0.90625, \cdots, 1.09375, 1.1\}$$

Problem Motivation

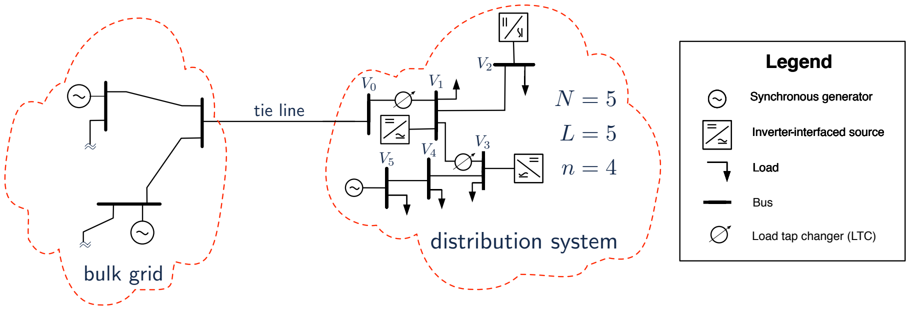
- Current LTC control schemes are myopic:
  - Based on local voltage measurements
  - Do not account for future uncertainty effects on current control actions
- These schemes are no longer suitable because of increased variability and uncertainty in uncontrolled power injections arising from:
  - Residential PV installations
  - Electric vehicles
- In addition, the use of controlled DERs for providing frequency regulation to the bulk grid has an impact on voltage regulation

Volt/VAR Control Architecture



- **Slow Time-Scale**: LTCs are periodically dispatched so as to reduce mechanical wear
- **Fast Time-Scale Control**: Power-electronic-interfaced DERs with reactive power provision capability

Optimal LTC Dispatch Problem



Objective

Find a policy for determining the LTC tap positions based on measurements of current

- tap ratios
- bus voltage magnitudes

so as to minimize bus voltages deviations from some reference value as power injections change as time evolves

Power Distribution System Model

- The relation between square voltage magnitudes and active/reactive power injections and LTC tap ratios at instant k can be written as

$$\boldsymbol{V}[k] = g(\boldsymbol{p}[k], \boldsymbol{q}[k], \boldsymbol{t}[k])$$

$\boldsymbol{V}[k]$	Vector of voltage magnitudes at instant k
$\boldsymbol{p}[k], \boldsymbol{q}[k]$	Vector of active, reactive, power injections at instant k
$\boldsymbol{t}[k]$	Vector of tap ratios at instant k

- Network topology and transmission line parameters are encapsulated into g

Assumptions

T1. Changes in active and reactive power injections at instant k, $\Delta \boldsymbol{p}[k] := \boldsymbol{p}[k+1] - \boldsymbol{p}[k]$ and $\Delta \boldsymbol{q}[k] := \boldsymbol{q}[k+1] - \boldsymbol{q}[k]$, are random

T2. Active and reactive power injections k, $\boldsymbol{p}[k]$ and $\boldsymbol{q}[k]$, are not measured and their joint probability distribution is unknown

T3. Network topology is known, i.e., $\boldsymbol{M}$ is known

T4. Line parameters are unknown, i.e., $\boldsymbol{r}$ and $\boldsymbol{x}$ are unknown

- Because of Assumption T1 is natural to formulate the problem as a Markov Decision Process (MDP)
- Because of Assumptions T2 – T4, we will have to resort to reinforcement learning (RL) techniques to solve this MDP

LTC Coordination Problem as an MDP

- State space, $\mathcal{S}$: the state, $\boldsymbol{s}$, is composed of tap ratio and squared voltage magnitude vector, i.e., $\boldsymbol{s} = (\boldsymbol{t}, \boldsymbol{v})$, $\boldsymbol{t} \in \mathcal{T}^{L^t}$, $\boldsymbol{v} \in \mathbb{R}^N$; thus,

$$\mathcal{S} \subseteq \mathcal{T}^{L^t} \times \mathbb{R}^N$$

- Action space, $\mathcal{A}$: the action, $\boldsymbol{a}$, is the change in LTC tap ratio between two consecutive time instants, i.e., $\boldsymbol{a} = \Delta \boldsymbol{t}$, $\Delta \boldsymbol{t} \in \Delta \mathcal{T}^{L^t} =: \mathcal{A}$, where

$$\Delta \mathcal{T} = \{0, \pm 0.00625, \cdots, \pm 0.19375, \pm 0.2\}$$

is the set set of feasible tap ratio changes

- Reward function, $\mathcal{R}$: the reward when the system transitions from state $\boldsymbol{s} = (\boldsymbol{t}, \boldsymbol{v})$ into state $\boldsymbol{s}' = (\boldsymbol{t}', \boldsymbol{v}')$ after taking action $\boldsymbol{a} = \Delta \boldsymbol{t}$ is

$$\mathcal{R}(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}') = -\frac{1}{N} \|\boldsymbol{v}' - \boldsymbol{v}^\star\|,$$

i.e., we are penalizing voltage deviation from some reference value $\boldsymbol{v}^\star$

LTC Coordination Problem as an MDP

- State transitions: governed by random changes in active and reactive power injection vectors, $\Delta \boldsymbol{p} := \boldsymbol{p}' - \boldsymbol{p}$ and $\Delta \boldsymbol{q} := \boldsymbol{q}' - \boldsymbol{q}$:

$$\boldsymbol{s}' = h(\boldsymbol{s}, \boldsymbol{a}, \Delta \boldsymbol{p}, \Delta \boldsymbol{q})$$

  - Network topology and transmission line paramters are encapsulated into h
  - Probability of transitioning from $\boldsymbol{s}$ to $\boldsymbol{s}'$ under action $\boldsymbol{a}$, $\mathcal{P}(\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a})$, could be computed if h and the joint pdf of $\Delta \boldsymbol{p}$ and $\Delta \boldsymbol{q}$ were know

Objective

Find a policy $\boldsymbol{\pi} : (\boldsymbol{t}, \boldsymbol{v}) \mapsto \Delta \boldsymbol{t},\ \boldsymbol{t} \in \mathcal{T}^{L^t},\ \boldsymbol{v} \in \mathbb{R}^N, \Delta \boldsymbol{t} \in \Delta \mathcal{T}^{L^t}$ so that

$$-\frac{1}{N} \sum_{k=0}^{\infty} \gamma^k \mathbb{E}\Big[\|\boldsymbol{v}[k+1] - \boldsymbol{v}^\star\| \;\Big|\; \boldsymbol{t}[0] = \boldsymbol{t}_0, \boldsymbol{v}[0] = \boldsymbol{v}_0 \Big]$$

is maximized (equivalent to minimizing voltage deviations)

Solving the LTC Coordination Problem

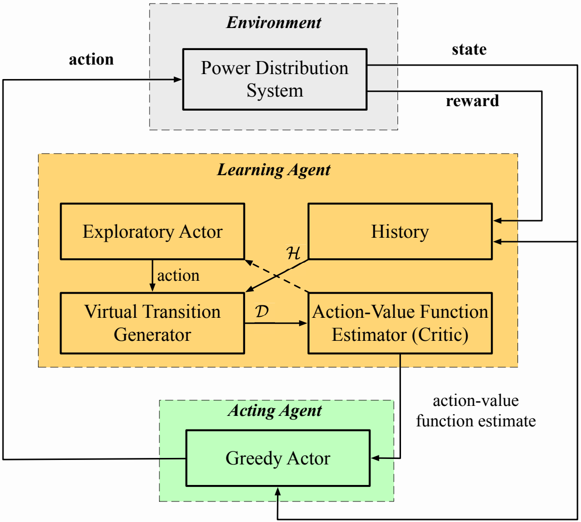
- Let $\bar{R}(\boldsymbol{s}, \boldsymbol{a})$ denote the expected reward for the pair $(\boldsymbol{s}, \boldsymbol{a})$
- The MDP is solved when we find $Q^*(\boldsymbol{s}, \boldsymbol{a}), \boldsymbol{s} \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}$, and $\boldsymbol{\pi}^*(s), \boldsymbol{s} \in \mathcal{S}$, satisfying

$$Q^*(\boldsymbol{s}, \boldsymbol{a}) = \bar{R}(\boldsymbol{s}, \boldsymbol{a}) + \gamma \sum_{\boldsymbol{s'} \in \mathcal{S}} \mathcal{P}(\boldsymbol{s'}|\boldsymbol{s}, \boldsymbol{a}) \max_{\boldsymbol{a}' \in \mathcal{A}} Q^*(\boldsymbol{s}', \boldsymbol{a}')$$

$$\boldsymbol{\pi}^*(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} Q^*(\boldsymbol{s}, \boldsymbol{a})$$

- Two issues in our setting:
  - I1. We do not know $\mathcal{P}(\cdot \mid \cdot, \cdot)$
  - I2. Even if we knew $\mathcal{P}(\cdot \mid \cdot, \cdot)$, we could not solve for $Q^*(\boldsymbol{s}, \boldsymbol{a})$ efficiently because of the curse of dimensionality in the state and action spaces
- To circumvent Issue I1, we apply a model-free RL algorithm that utilizes transition samples obtained via a virtual transition generator
- To circumvent Issue I2, we use function approximation and a learning scheme for sequential estimation of the action-value function

LTC Coordination Framework Building Blocks



Action-Value Function Value Estimator

- Let $\hat{Q}(\cdot, \cdot)$ denote an approximation of the optimal action-value function $Q(\cdot, \cdot)$

- Methods for obtaining $\hat{Q}(\cdot, \cdot)$ include:
  - Parametric functions
  - Neural networks

- Here we consider a linear parametrization of $\hat{Q}(\cdot, \cdot)$:

$$\hat{Q}(\boldsymbol{s}, \boldsymbol{a}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{s}, \boldsymbol{a}),$$

  where $\boldsymbol{w} \in \mathbb{R}^f$ is the parameter vector and $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^f$ is some basis function

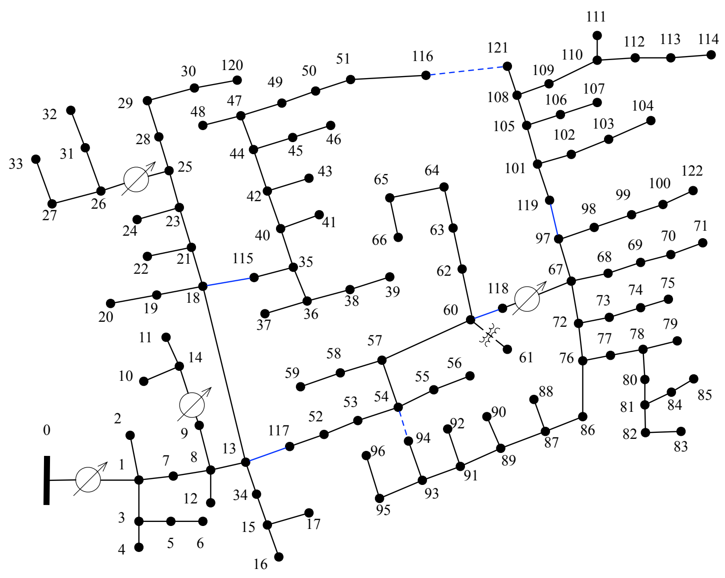
- Let $\mathcal{D} = \{(\boldsymbol{s}, \boldsymbol{a}, r, \boldsymbol{s}') : \boldsymbol{s}, \boldsymbol{s}' \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}\}$ denote a set (batch) of transition samples obtained via observation or simulation

- We use least-square policy iteration (LSPI) algorithm [Lagoudakis and Parr, 2003] to find $\boldsymbol{w}$ that best fits the transition samples in $\mathcal{D}$

Challenges

- The LSPI algorithm requires adequate transition samples that spread over $\mathcal{S} \times \mathcal{A}$
- This is challenging in power systems since the system operational reliability might be jeopardized when exploring randomly
  - We address this by developing by generating samples via a virtual transition generator that leverages historical system operational data
- The LSPI also suffers from the curse of dimensionality when the action space is large—the case in the LTC coordination problem
  - We address this by using a sequential scheme that breaks the learning problem into smaller problems and uses the LSPI algorithm on those

IEEE 123-bus Test Feeder



IEEE 123-bus Test Feeder Results

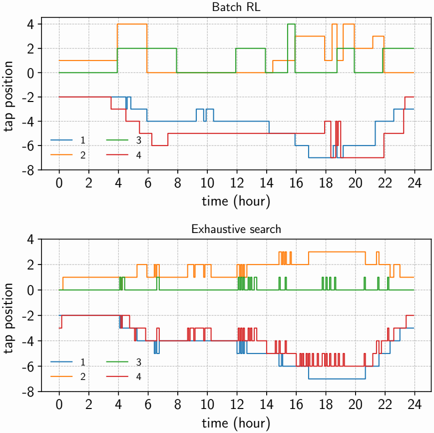


Figure 5: Tap positions.

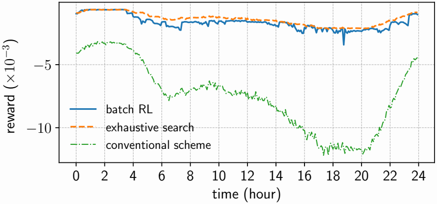


Figure 6: Immediate rewards.

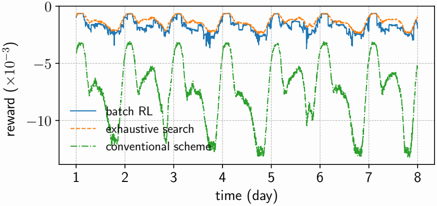


Figure 7: Rewards over 7 days.

Outline

Conclusions

- We developed a data-driven coordination framework for coordinating assets in distribution systems to provide ancillary services:
- The proposed framework:
  - It assumes no prior information on the distribution system model, except knowledge of network topology
  - It mainly relies on measurements
  - It is adaptive and robust to changes in operating conditions
- Refer to the following papers for more details:

1. H. Xu, A. Domínguez-García, and P. Sauer, "Data-driven Coordination of Distributed Energy Resources for Active Power Provision," *IEEE Transactions on Power System*, 2019.
2. H. Xu, A. Domínguez-García, and P. W. Sauer, "Optimal tap setting of voltage regulation transformers using batch reinforcement learning," *IEEE Transactions on Power System*, 2019.

Questions?

Alejandro D. Domínguez-García
e-mail: aledan@ILLINOIS.EDU
url: https://aledan.ece.illinois.edu

References I

[Lagoudakis and Parr, 2003] Lagoudakis, M. G. and Parr, R. (2003).
Least-squares policy iteration.
Journal of Machine Learning Research, 4:1107–1149.

A Primer on Markov Decision Processes

- A Markov Decision Process (MDP) is defined as a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ where
  - $\mathcal{S}$ is a finite set of states
  - $\mathcal{A}$ is a finite set of actions
  - $\mathcal{P}$ is a Markovian transition model
  - $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is a reward function
  - $\gamma \in [0, 1)$ is a discount factor

- Let $\boldsymbol{s}[k]$ and $\boldsymbol{a}[k]$ denote random variables (r.v.'s) respectively describing the value the state and action take at time instant k

- Let $r[k]$ denote a r.v. describing the reward received after taking action $\boldsymbol{a}[k]$ in state $\boldsymbol{s}[k]$ and transitioning to state $\boldsymbol{s}[k+1]$; then

$$r[k] = \mathcal{R}\big(\boldsymbol{s}[k], \boldsymbol{a}[k], \boldsymbol{s}[k+1]\big)$$

- A deterministic policy $\boldsymbol{\pi}$ is a mapping from $\mathcal{S}$ to $\mathcal{A}$, i.e., $\boldsymbol{a} = \boldsymbol{\pi}(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}$

A Primer on Markov Decision Processes

Objective

Given some initial state, s_0, we want to find a deterministic policy, $\boldsymbol{\pi}^*$, that maximizes the expected value of the cumulative discounted reward, i.e.,

$$\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi}} \sum_{k=0}^{\infty} \gamma^k \mathbb{E}\Big[r[k] \mid \boldsymbol{s}[0] = \boldsymbol{s}_0\Big]$$

A Primer on Markov Decision Processes

- The action-value function under policy $\boldsymbol{\pi}$ is defined as

$$Q^{\boldsymbol{\pi}}(\boldsymbol{s},\boldsymbol{a}) = \sum_{k=0}^{\infty} \gamma^k \mathbb{E}\Big[r[k] \mid \boldsymbol{s}[k] = \boldsymbol{s}, \boldsymbol{a}[k] = \boldsymbol{a}; \boldsymbol{\pi}\Big], \quad \boldsymbol{s} \in \mathcal{S}, \quad \boldsymbol{a} \in \mathcal{A}$$

- The optimal action-value function, $Q^*(\boldsymbol{s},\boldsymbol{a}), \boldsymbol{s} \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}$ is the maximum action-value function over all policies, i.e.,

$$Q^*(\boldsymbol{s},\boldsymbol{a}) = \max_{\boldsymbol{\pi}} Q^{\boldsymbol{\pi}}(\boldsymbol{s},\boldsymbol{a})$$

- $Q^*(\boldsymbol{s},\boldsymbol{a}), \boldsymbol{s} \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}$, satisfies the following Bellman equation:

$$Q^*(\boldsymbol{s},\boldsymbol{a}) = \mathbb{E}\left[r \mid \boldsymbol{s},\boldsymbol{a}\right] + \gamma \sum_{\boldsymbol{s}' \in \mathcal{S}} \mathcal{P}(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \max_{\boldsymbol{a}' \in \mathcal{A}} Q^*(\boldsymbol{s}',\boldsymbol{a}')$$

- The optimal policy, $\boldsymbol{\pi}^*(s), \boldsymbol{s} \in \mathcal{S}$, is obtained as follows:

$$\boldsymbol{\pi}^*(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} Q^*(\boldsymbol{s},\boldsymbol{a})$$

- The MDP is solved if we find $Q^*(\cdot,\cdot)$ and the corresponding $\boldsymbol{\pi}^*(\cdot)$

LSPI Algorithm Iteration

- Let $\boldsymbol{w}_i$ denote the estimate of $\boldsymbol{w}$ at the beginning of iteration i
- For each transition sample $(\boldsymbol{s}, \boldsymbol{a}, r, \boldsymbol{s}') \in \mathcal{D}$, compute

$$\boldsymbol{a}' = \arg\max_{\boldsymbol{\alpha} \in \mathcal{A}} \boldsymbol{w}_i^\top \boldsymbol{\phi}(\boldsymbol{s}', \boldsymbol{\alpha})$$

- Update the estimate of $\boldsymbol{w}$ as follows:

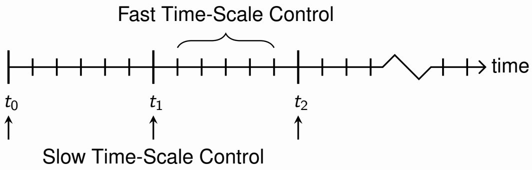
$$\boldsymbol{w}_{i+1} = \boldsymbol{B}^{-1}\boldsymbol{b}$$

where

$$\boldsymbol{B} = \sum_{(\boldsymbol{s},\boldsymbol{a},r,\boldsymbol{s}') \in \mathcal{D}} \underbrace{\boldsymbol{\phi}(\boldsymbol{s}, \boldsymbol{a})\Big(\boldsymbol{\phi}(\boldsymbol{s}, \boldsymbol{a}) - \gamma \boldsymbol{\phi}(\boldsymbol{s}', \boldsymbol{a}')\Big)^\top}_{\text{rank-1 matrix}}$$

$$\boldsymbol{b} = \sum_{(\boldsymbol{s},\boldsymbol{a},r,\boldsymbol{s}') \in \mathcal{D}} \boldsymbol{\phi}(\boldsymbol{s}, \boldsymbol{a}) r$$

Timeline



- Policy updated every $K\Delta T$ units of time, e.g., 2 hours
- Updated policy used for tap setting for K time instants