Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Московский физико-технический институт (государственный университет)»

Факультет управления и прикладной математики

Кафедра «Интеллектуальные Системы»

Построение системы поиска похожих опухолей с помощью глубоких сверточных сетей

Выпускная квалификационная работа
(бакалаврская работа)

Направление подготовки: 03.03.01 Прикладные математика и физика

Выполнил:
студент 474 группы _____ Макарчук Глеб Игоревич

Научный руководитель:
к.ф.-м.н. _____ Беляев Михаил Геннадьевич

Москва 2018

Contents

Todo list

Chapter 1

Introduction

Energy efficiency turns to be the one of main challenge for humanity. Lack of resources together with rocketed energy demand across the planet urge peoples to increase efficiency of their energy systems. One of the most promising technologies in this field is Smart Grid. Traditionally, the term ”grid” denotes an electricity system, that supports electricity generation, transmission, distribution and control. Most of them use for direct energy delivery from several large generators to consumers. In contrast, a Smart Grid is an electricity grid which utilise two-way flows of energy and information to establish automated and distributed next-generation energy delivery network [1]. These intelligent technologies are incorporated across the entire system which improve its efficiency, safety and reliability [2].

The main advantage of Smart Grid systems is a potential for altering end-user consumption behaviour, which allows to shift peak demands to stabilise daily consumption profile. Such interaction between energy operators and consumers is called Demand-Response (henceforth DR). Plenty of DR schemes have been proposed recent years, and some of them are in use in US energy market. They differ in what they assume about the system, what type of grid architecture they utilise, what type of motivation for end users to participate into the program they propose, and what mathematical approach to emerge system's intelligence they apply. The main difficulty which most of the schemes are trying to tackle is a lack of communication between server and customers in real world, which prohibits to address each device individually. However, most of works assume two-sided information stream, which leads to significant infrastructure investment therefore detaining proliferation of such intelligent systems into the market.

In this study we propose a novel scheme which is aimed to minimise communication requirements without a lessening in functionality to curtail or adjust overall consumption in grid. In particular, it require to know only

aggregated energy consumption in each period, which is significantly more realistic then two-sided communication architecture. To pursue this we consider the set of consumers as an ensemble of devices, which behaviour is representable by finite Markov chains. This is an extension of [3] ideas, who proves this approach to be viable end efficient. To tackle optimisation subroutines we consider this problem as a Linear Contextual Bandit With Knapsack (BwK [4]) setup, applying the algorithms proposed in [5]. The main contribution of this work is proposed markov-chains-based ensemble model, which provides context for contextual bandit learning algorithms and has a naturally linear dependency of reward over this context. It enables to apply advanced bandit algorithms as [5] without loss of its convergence guarantees making no additional assumptions.

1.1 Demand-Response Overview

One of the distinguishing features of Smart Grid networks is demand manageability. This concept of the Demand-Side Management (DSM) includes all activities aiming to alter the consumer's demand profile to make it match the supply or to effectively incorporate renewable energy sources [6]. Nowadays the major activity in DSM is Demand-Response (DR)[7]. According to the United States Department of energy, Demand-Response is "a tariff or program established to motivate changes in electric use by end-use customers, in response to changes in the price of electricity over time, or to give incentive payments designed to induce lower electricity use at times of high market prices or when grid reliability is jeopardized" [8]. As was summarised in [9], the main objectives of the application of DR are:

- Reduction of the total energy consumption both on demand and transmission sides. Such overall consumption curtailment may help governments and energy providers to meet their pollution obligations [8, 10, 11].
- Reduction of the maximal needed power generation in order to eliminate the need of activating expensive-to-run power plants to meet peak demands.
- Efficient incorporation of renewable energy sources through making the demand follow the available local supply fluctuations. Such incorporation may significantly increase the overall system's reliability in regions with high penetration of wind farms and solar panels [12].

- Reduction or even elimination of overloads in distribution systems shifting peak demands' time for a subset of consumers.

Amount of DR service providers (also know as Aggregators) in US energy market soared up recent years due to emerging new intelligent solution and overall market liberalisation . A typical aggregator company represents several thousand households or a few dozens of commercial consumers (e.g. downtown office buildings). One of the most typical aggregator's business models is providing

is it true?

Add citation

some statistics

As shown in figure the principal DR-scheme consists of cooperation of four main participants: a) an Aggregator, b) a System Operator (SO), c) Power Generation Unit(s) d) and Power Consumer(s) [13]. Their interaction is a cyclic process typically started by the SO, which determines the preferred power consumption and sends it to the Aggregator. Next the Aggregator chooses participating loads from available, calculates possible change in demand and sends it back to the SO. And finally the Operator informs the most available substations about the upcoming demand. In such scheme the Aggregator provides the grid's intelligence executing optimisation procedures pr revealing problems in distribution system [9].

insert scheme

DR Schemes distinct in control architecture they utilise, and in motivation to participate which they provide to customers.

DR Schemes by its architecture DR schemes may be classified into centralised and distributed programs [14], according to where the decision for the execution program are made. In centralised schemes load activations are managed only by the central utility. Such schemes are easy to implement, but turn to be hard-headed in large and complex systems. However it remains an effective approach for controlling ensembles of termostatically controlled loads [15], charging systems for electro vehicles [16] and commercial consumers [17]. For example

Add example

DR Schemes by its motivation The proposed motivation schemes usually adopt either price-based or incentive-based approach.

1.2 Multiarmed Bandits

Reinforcement learning can be defined as a learning paradigm concerned with learning to control a complex system so to maximize a numerical performance measure that express some long-term objective [18]. The most typical setting where reinforcement learning operates is an iterative process of agent-environment interactions (see fig *somefig*). Formally, at the

moment t the agent performs an action a_t according to his current policy π_t and gets a reward r_t from the environment. The aim of the agent is to maximize the reward $\sum_{t=t_0}^{t_0+T} r_t$ in a given time horizon T, or, equally, to minimize the regret $\bar{R_T} = R^x *_T - \sum_{t=t_0}^{t_0+T} r_t$ where R_T^* is the best reward the agent can theoretically get with the best policy π^* available. However, in most of the real problem we have some additional information associated with each arm or with the situation in general. The vector that represents this information is usually called ¡¡the context¿¿ hence the related setup is called ¡¡contextual multi armed bandit¿¿. Under the described conditions the agent should take into account the context to achieve better performance.

Multiarmed bandit problem can be defined as a Reinforcement Learning setup with a fixed state of the environment, so the problem here is to learn the best policy.

The amount of papers devoted to Reinforcement Learning has soared up recent years due to significant advances in *some domains (link)*.

Plenty of papers is devoted to the general multi armed bandit setup. One of the most famous algorithms – ε-greedy – was proposed by [19]: it pulls an arbitrary arm with a probability of ε and the best arm according to the current policy otherwise. This is due tho the exploration-exploitation balance problem: without this arbitrary steps the agent may fail to gain enough statistics for determining the best policy. However, ε is a hyperparameter which should be appropriately tuned. Another approach here is so-called ¡¡optimism in the face of uncertainty¿¿ [20] according to which the learner should choose an arm with the best Upper Confidence Bound. A very successful algorithm which has implemented this technique is UCB1 [19], later analysed and improved by [21]. The later not only often outperforms UCB1, but also had shown to be essentially unimprovable under the assumption that the variance of the reward associated with some of the actions are small.

In contrast to the regular bandit problem, the contextual bandit setup tends to be tougher problem to solve. One approach is to assume the particular dependency model between the arms' context $x_{a,t}$ and the expected reward r_t. One possible model is a linear one:

$$r_{a,t} = \bar{r_{a,t}} + \varepsilon_{a,t} \tag{1.1}$$

$$s.t.\, \bar{r_{a,t}} = \mathbb{E}[r_{t,a}|x_{t,a}] = x_{t,a}^T \theta^* \tag{1.2}$$

$$\mathbb{E}\, \varepsilon_{a,t} = 0 \tag{1.3}$$

For this linear case [22] proposed an adaptation of UCB1 approach: LinUCB. Later [23] provided better theoretical analysis by eliminating the assumption that the reward is identically distributed over time and arms,

which is mostly far from true. The similar case with different regularization strategy was examined by [24]. Another heuristics for balancing exploration and exploitation is known as Thompson Sapling was adapted to the linear reward model by [25].

1.3 Reinforcement Learning in Demand-Response

Chapter 2

Proposed Framework

2.1 Notation and Problem Setup

We consider the system of n thermostatically controlled loads, indexed by $i = 1, 2, \ldots, n$. We suppose that there is an aggregator A which send matrices $\bar{P}(t), t = 1, \ldots, T$.

We refer $P_i(t) \in \mathbb{R}^{N \times N}$ as a transition matrix of load i, $i = 1, \ldots, n$, and N is a number of possible states of any load (all loads have the same number of states). Each load has its own behaviour defined as a function $P_i(t)$. All these behaviours are independent. We also suppose that any $\bar{P}(t)$ and $P_i(t) \in \mathcal{P} = \{P^{(1)}, P^{(2)}, \ldots, P^{(m)}\}$. Each load is allowed to accept or reject a transition matrix $\bar{P}(t)$ according to the following rule:

$$\bar{P}_i(t) = \begin{cases} \bar{P}(t), & \text{if } \Omega(\bar{P}(t), P_i(t), t) = 1 \\ P_i(t), & \text{otherwise,} \end{cases}$$

where $\Omega(P_1, P_2, t)$ is some decision rule which meets all legal and customer requirements. For the sake of simplicity we may assume this rule is known (i.e. a part of the contract), however, it is not required: we may learn it on-the-fly. Also, as we have only m different matrices, Ω is a time-dependent matrix where $\Omega_{ij}(t) = 1$ if having own matrix $P_k(t) = P^{(i)}$ a load k accepts proposed matrix $\bar{P}^{(j)}$ at time t.

We refer $\pi_i(t)$ as a unit vector corresponding to the state of load i, $1 \leq i \leq n$, at time t, so that

$$\pi(t) = \frac{1}{n} \sum_{i=1}^{n} \pi_i(t).$$

We also assume the the power consumption q at each state is known, so that the total power consumption $s(t)$ is

$$s(t) = n \cdot \pi(t)^\top q.$$

We refer $\bar{s}(t)$ as the power requested by the system operator at time t.

We also assume the loss function of the system operator to be

$$\sum_{i=1}^{T} c(t)|\bar{s}(t) - s(t)|,$$

where $c(t) > 0$ is non-negative cost function.

In this research we implement an incentive-based model of users motivation to participate in the curtailment program. In particular we consider the Emergency Demand-Response Program (EDRP) setting here, in which consumers get incentive payments for reducing their power consumption during reliability triggered events (see [26]). Consumers may choose not to curtail and therefore to forgo the payments, which are usually specified beforehand [9]. Due to privacy protection reasons ([27]) the aggregator is forbidden to observe exact loads accepting a curtailment request, but it is important to know the total amount of these loads for two reasons:

1. As the aggregator's budget is strictly limited, it must estimate expenses for incentive payments.

2. Most of the DR incentive-based programs limit the total amount of curtailment hours to avoid user disturbance (typically 200 hours/year [28])

To sum up, the exact objective will be:

$$\min \sum_{i=1}^{T} c(t)|\bar{s}(t) - s(t)| \tag{2.1}$$

$$s.t. \sum_{t=1}^{T} k(t) \leq K \tag{2.2}$$

2.2 Solution

In the following section we treat 2.1 as a **linear contextual bandit with knapsack** problem [4]. Define the policy as some mapping from the context $f(t)$ to the matrix number $i \in \{0, 1, \ldots, m\}$ (a.k.a. arm). Define the regret $\mathcal{R}_T^\nu$ as the difference between performances under the best policy available and under some policy ν:

$$\mathcal{R}_T^\nu = \sum_{t=1}^{T} (r^*(t) - r^\nu(t)) \tag{2.3}$$

The goal is to learn the policy ν which minimizes the regret:

$$\begin{aligned} &\min_{\nu} \mathcal{R}_T^{\nu} = \min_{\nu} \sum_{t=1}^{T} (r^{\nu}(t) - r^*(t)) \\ &s.t. \sum_{t=1}^{T} k(t) \leq K \end{aligned} \tag{2.4}$$

where the penalty is defined as

$$r^{\nu}(t) = c(t)|s^{\nu}(t) - \bar{s}(t)|$$

and $k(t)$ is the number of loads who accepted the curtailment request at the time t, K is the contract limit for user disturbance and ν is the policy being used. Hereafter we refer to $r^{\nu}(t)$ as $r(t)$.

In this form of bandits setup we face two major troubles:

1. General bandit with knapsack problem setup [4] consider penalties as a random independent variables over time and arms. Both assumptions here are false: our choices from the past strongly affects the current reward distributions.

2. Variables $k(t)$ are unobservable, hence we can only ensure thresholds with some probability.

To deal with these issues consider the the mechanics of how $s(t)$ depends on aggregator's actions. Suppose that by the time t we have a (possibly unknown) state-distribution $\pi(t)$. It is easy to see that if each device chooses its own matrix, and we denote the number of devices which have chosen matrix P_j as n_j, then the next-moment consumption with no control will be:

$$s(t|0) := \sum_{j=1}^{m} n_j(t) q^T P_j \pi(t) = \sum_{j=1}^{m} n_j(t) f_j(t|0) \tag{2.5}$$

and if we pull i^{th} arm, the consumption will be:

$$s(t|i) := \sum_{j=1}^{m} n_j(t) q^T (\Omega_{ij}(t) P_i + (1 - \Omega_{ij}(t)) P_j) \pi(t) = \sum_{j=1}^{m} n_j(t) f_j(t|i) \tag{2.6}$$

The model appears to be linear over unknown variables $n_j(t)$ and some vector $f(t|i) := \{f_j(t|i)\}_{j=1}^{m}$, $i \in [0, 1, \dots m]$ which we know if we have

$\pi(t)$ and $\Omega_{ij}(t)$. Therefore it is surprisingly convenient to consider $f(t|i)$ as a feature vector of an arm i at the moment t. Note, that the features $f_j(i|t)$ have the natural interpretation as the amount of energy being consumed by an average device which has $\bar{P} = P_j$ at the time t if we pull the i-th arm. Moreover, we also have an estimator for budgeted variable: $k(t|i) = \sum_{i=1}^{m} \Omega_{ij} n_j$.

The Bandit The above mentioned arm feature vectors (a.k.a. arm-contexts) $f(i|t)$ reveals a straightforward "linear contextual bandit with knapsack" setup:

$$\begin{aligned} \min_{\nu:\, t\to\{0,1\ldots m\}} & \sum_{t=1}^{T} \mathcal{L}_t \left(n(t)^T f(t|\nu(t)) \right) \\ s.t. \, \mathbb{E} & \sum_{t=1}^{T} k(t|\nu(t)) \leq K \\ & \sum_{j=1}^{m} n_j(t) = n \, \forall t \\ & n_j(t) \geq 0 \, \forall t \end{aligned}$$

where $n(t) = \{n_j(t)\}_{j=1}^{m}$ (do not miss it with n – the total amount of devices), and $\mathcal{L}_t(x)$ is a loss function. In this work two classical demand-response loss functions are considered. The first one is a trivial mapping:

$$\mathcal{L}_t(x) = x \tag{2.7}$$

which gives the consumption minimization setup:

$$\begin{aligned} \min_{\nu:\, t\to\{0,1\ldots m\}} & \sum_{t=1}^{T} n(t)^T f(t|\nu(t)) = \min_{\nu} \sum_{t=1}^{T} s^{\nu}(t) \\ s.t. \, \mathbb{E} & \sum_{t=1}^{T} k(t|\nu(t)) \leq K \\ & \sum_{j=1}^{m} n_j(t) = n \, \forall t \\ & n_j(t) \geq 0 \, \forall t \end{aligned} \tag{2.8}$$

and the second one is the absolute deviation from a given series $\bar{s}(t)$:

$$\mathcal{L}_t(x) = |x - \bar{s}(t)| \tag{2.9}$$

which gives the consumption stabilization setup:

$$\begin{aligned} \min_{\nu:\, t\to\{0,1\dots m\}} \sum_{t=1}^{T} |n(t)^T f(t|\nu(t)) - \bar{s}(t)| &= \min_{\nu} \sum_{t=1}^{T} |s^{\nu}(t) - \bar{s}(t)| \\ s.t.\, \mathbb{E} \sum_{t=1}^{T} k(t|\nu(t)) &\le K \\ \sum_{j=1}^{m} n_j(t) &= n\, \forall t \\ n_j(t) &\ge 0\, \forall t \end{aligned} \tag{2.10}$$

It remains to notice that $n(t)$ is a periodic function over time with a period of 24 hours (at least within one season of a year), so we can consider 2.8 and 2.10 as a set of 24 bandits: each one is learning to make a decision at the assigned hour. Note that these problems are in exact form for recently emerged budgeted bandit solvers ([4] and more (add later)), and all the requirements are met: the divergence between predicted and real consumption may appear only regarding to stochatical fluctuations due to the finiteness of the ensemble, hence we have independence of reward as a function of a context through arms and time (*of course we need a proof here*).

The Algorithm with known $\Omega_{ij}(t)$ and without a knapsack: Consider a simpler setup when $K = \infty$ i.e. we have no knapsack constraints in our problem. Here τ is the number of steps which a device performs each hour. We also assume that we do not know neither an initial state distribution $\pi(0)$ nor $n_j(t)$. In this case 1 is how a dummy algorithm may look like (we assume here a 2.9 loss function).

One may notice that in 1 the greedy policy was used, and no exploration steps were proposed. In fact, the greedy policy here turns to be optimal: all the reward variances depend only on unknown vector $n(t)$ and when we learn it we decrease reward variances for all arms simultaneously. Hence, in contrast to the following case, no exploration-exploitation balance is required here.

The Algorithm without a knapsack and without $\Omega_{ij}(t)$ Note that it is not crucial for the previous algorithm to know the oracle $\Omega_{ij}(t)$, as we

Algorithm 1 Non-budgeted TCL-control with oracle

Require: $\mathcal{P}$, n, $\tau, \Omega_{ij}(t)$

$n_i(0) := 1/n \; \forall i$

$\pi_0 := \frac{1}{n} \sum_{i=1}^{m} n_i(0) u_i$, where u_i is a stationary distribution of the matrix P_i

$A(t) = \{\emptyset\} \; \forall t$ where A(t) is an array of all feature vectors through time which we got at the hour t. It will be used as a learning dataset for the t-th regressor.

$S(t) = \{\emptyset\} \; \forall t$ where S(t) is an array of all s(t) through time which we got at the hour t. These are the targets for the t-th regressor.

for $t := 1 \dots T$ **do**

  Generate the arm contexts $\{f_j(t|i)\}_{j=1}^{m}$ for all $i \in \{0, 1 \dots m\}$:

  $f_j(t|i) := \sum_{\xi=1}^{\tau} q^T (\Omega_{ij}(t) P_i^{\xi} + (1 - \Omega_{ij}(t)) P_j^{\xi}) \pi(t) \quad i \in \{1 \dots m\}$

  $f_j(t|0) := \sum_{\xi=1}^{\tau} q^T P_j^{\tau} \pi(t)$

  if $\bar{s}(t) == 0$ **then**

    Set $i := 0$ (no control) and send it to the ensemble

  else

    $i := \arg\max_{i \in \{1 \dots m\}} \mathcal{L}_t(s(t|i)) := \arg\max_{i \in \{1 \dots m\}} \mathcal{L}_t(n(t)^T f(t|i))$.

    Send the chosen arm i to the ensemble

  end if

  Wait and receive $s(t|i)$ – the actual ensemble consumption

  $A(t)$.append($f(t|i)$)

  $S(t)$.append($s(t|i)$)

  Solve a constrained regression problem to learn real $n_j(t)$:

  $n(t) = \arg\min_{n \in R_+^n} \|A(t)n(t) - S(t)\|^2, \; s.t. \sum_{j=1}^{m} n_j(t) = n, \; n_j(t) \geq 0,$ start from $n(t)$

  Calculate next π:

  $\pi(t+1) = \frac{1}{n} \sum_{j=1}^{m} n_j(t) (\Omega_{ij}(t) P_i^{\tau} + (1 - \Omega_{ij}(t)) P_j^{\tau})) \pi(t)$

  if $n_j(t+1)$ are undefined yet **then**

    $n_j(t+1) = n_j(t)$

  end if

end for

may learn in the same way we learn $n(t)$. Formally, let i be the arm the aggregator pulled at the moment t, then the consumption will be:

$$s(t|i) = n(t)^T f(t|i) = \sum_{j=1}^{m} n_j(t) \sum_{\xi=1}^{\tau} q^T(\Omega_{ij}(t)P_i^\xi + (1-\Omega_{ij}(t))P_j^\xi)\pi(t) =$$
$$= \sum_{j=1}^{m} \Omega_{ij}(t) n_j \sum_{\xi=1}^{\tau} q^T(P_i^\xi - P_j^\xi)\pi(t) + \underbrace{\sum_{j=1}^{m} n_j \sum_{\xi=1}^{\tau} q^T P_j^\xi \pi(t)}_{s(t|0)} \tag{2.11}$$

and we have a discrete regularized regression problem over the vector $\Omega_i(t) \in \{0,1\}^m$ here:

$$\underbrace{s(t|i) - s(t|0)}_{target} = \sum_{j=1}^{m} \underbrace{\Omega_{ij}(t)}_{variables} \underbrace{n_j(t) \sum_{\xi=1}^{\tau} q^T(P_i^\xi - P_j^\xi)\pi(t)}_{features} \tag{2.12}$$

If we also have observations from i-th arm at t-th hour from the past (i.e. this is not the first time we pull this arm at this hour), then we may solve the least-squared optimization problem over $\Omega_{ij}(t)|_{j=j} \in \{0,1\}^m$ which might reduce the fluctuations of Ω.

The algorithm 2 for this case is the essentially the same: the new parts are marked with red.

The Algorithm with a knapsack and without $\Omega_{ij}(t)$ This one is our final result. Note that we just need to apply not a regular UCB1 bandit inside the abovementioned algorithm, but the "bandit-with-knapsack" solver from [4] or something similar.

Algorithm 2 Non-budgeted TCL-control without oracle

Require: $\mathcal{P}$, n, τ, ~~$\Omega_{ij}(t)$~~

$n_i(0) := 1/n \; \forall i$

$\pi_0 := \frac{1}{n}\sum_{i=1}^{m} n_i(0) u_i$

$A(t) = \{\emptyset\} \; \forall t$

$S(t) = \{\emptyset\} \; \forall t$

$\Omega_{ij}(0) = \{0,1\}^{m \times m}$ – random matrix

for $t := 1 \dots T$ **do**

  Generate arm contexts $\{f_j(t|i)\}_{j=1}^{m}$ for all $i \in \{0, 1 \dots m\}$:

  $f_j(t|i) := \sum_{\xi=1}^{\tau} q^T (\Omega_{ij}(t) P_i^{\xi} + (1 - \Omega_{ij}(t)) P_j^{\xi}) \pi(t) \quad i \in \{1 \dots m\}$

  $f_j(t|0) := \sum_{\xi=1}^{\tau} q^T P_j^{\tau} \pi(t)$

  if $\bar{s}(t) == 0$ **then**

    Set $i := 0$ (no control) and send it to the ensemble

    Wait and receive $s(t|0)$ – the actual ensemble consumption

    When we have no control signal, we learn $n_j(t)$

    $A(t)$.append($f(t|0)$)

    $S(t)$.append($s(t|0)$)

    Solve a constrained regression problem to learn real $n_j(t)$:

    $n(t) = \arg\min_{n \in R_+^n} \|A(t)n(t) - S(t)\|^2, \; s.t. \sum_{j=1}^{m} n_j(t) = n, \; n_j(t) \geq 0,$ start from $n(t)$

    Calculate next π:

    $\pi(t+1) = \sum_{j=1}^{m} n_j(t) P_j^{\tau} \pi(t)$

  else

    Calculate $s(t|i) = n(t)^T f(t|i)$ for all $i \in \{0, 1, \dots m\}$

    Calculate variance estimations $\Delta s(t|i)$

    $i := \arg\min_{j \in \{0,1,\dots m\}} \mathcal{L}_t(s(t|j) + B\sqrt{\Delta s(t|i)})$

    Send the chosen arm i to the ensemble

    Wait and receive $s(t|i)$ – the actual ensemble consumption

    Learn the vector $\Omega_i(t)$ solving optimization problem:

    $\Omega_i(t) := \arg\min_{\omega \in \{0,1\}^m} \Big(\underbrace{(s(t|i) - s(t|0))}_{s} - \sum_{j=1}^{m} \omega_j \underbrace{n_j(t) \sum_{\xi=1}^{\tau} q^T (P_i^{\xi} - P_j^{\xi}) \pi}_{g_j} \Big)$

    $= \arg\min_{\omega \in \{0,1\}^m} \left(s - w^T g\right)^2$

    I suppose we need some kind of smooth relaxation here

    Calculate next π:

    $\pi(t+1) = \frac{1}{n}\sum_{j=1}^{m} n_j(t)(\Omega_{ij}(t) P_i^{\tau} + (1 - \Omega_{ij}(t)) P_j^{\tau})) \pi(t)$

  end if

  if $n(t+1)$ is undefined yet **then**

    $n(t+1) = n(t)$

  end if

  if $\Omega_{ij}(t+1)$ is undefined yet **then**

    $\Omega_{ij}(t+1) = \Omega_{ij}(t)$

  end if

Chapter 3

Numerical Experiments

Chapter 4

Conclusion

Bibliography

[1] Xi Fang, Satyajayant Misra, Guoliang Xue, and Dejun Yang, "Smart Grid — The New and Improved Power Grid: A Survey", *IEEE Communications Surveys & Tutorials*, vol. 14, no. 4, pp. 944–980, 2012.

[2] Jingcheng Gao, Yang Xiao, Jing Liu, Wei Liang, and C.L. Philip Chen, "A survey of communication/networking in Smart Grids", *Future Generation Computer Systems*, vol. 28, no. 2, pp. 391–404, feb 2012.

[3] Michael Chertkov, Vladimir Y Chernyak, and Deepjyoti Deka, "Ensemble Control of Cycling Energy Loads: Markov Decision Approach", jan 2017.

[4] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins, "Bandits with Knapsacks", in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. oct 2013, number May 2013, pp. 207–216, IEEE.

[5] Shipra Agrawal and Nikhil R. Devanur, "Linear Contextual Bandits with Knapsacks", jul 2015, number Nips.

[6] Mahnoosh Alizadeh, Xiao Li, Zhifang Wang, Anna Scaglione, and Ronald Melton, "Demand-Side Management in the Smart Grid: Information Processing for the Power Switch", *IEEE Signal Processing Magazine*, vol. 29, no. 5, pp. 55–67, sep 2012.

[7] Peter Palensky and Dietmar Dietrich, "Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads", *IEEE Transactions on Industrial Informatics*, vol. 7, no. 3, pp. 381–388, aug 2011.

[8] Department of Energy USA, "Benefits of Demand Response in Electricity Markets and Recommendations for Achieving them", Tech. Rep. February, 2006.

[9] John S. Vardakas, Nizar Zorba, and Christos V. Verikoukis, "A Survey on Demand Response Programs in Smart Grids: Pricing Methods and Optimization Algorithms", *IEEE Communications Surveys and Tutorials*, vol. 17, no. 1, pp. 152–178, 2015.

[10] Igor Shishlov, Romain Morel, and Valentin Bellassen, "Compliance of the Parties to the Kyoto Protocol in the first commitment period", *Climate Policy*, vol. 16, no. 6, pp. 768–782, aug 2016.

[11] United Nations/Framework Convention on Climate Change, "Paris Agreement", *21st Conference of the Parties*, p. 3, 2015.

[12] Enrique Santacana, Gary Rackliffe, Le Tang, and Xiaoming Feng, "Getting Smart", *IEEE Power and Energy Magazine*, vol. 8, no. 2, pp. 41–48, mar 2010.

[13] Jose Medina, Nelson Muller, and Ilya Roytelman, "Demand Response and Distribution Grid Operations: Opportunities and Challenges", *IEEE Transactions on Smart Grid*, vol. 1, no. 2, pp. 193–198, sep 2010.

[14] Liang Zhou, Joel J P C Rodrigues, and Luís M. Oliveira, "QoE-driven power scheduling in smart grid: Architecture, strategy, and methodology", *IEEE Communications Magazine*, vol. 50, no. 5, pp. 136–141, may 2012.

[15] He Hao, Borhan M. Sanandaji, Kameshwar Poolla, and Tyrone L. Vincent, "Aggregate Flexibility of Thermostatically Controlled Loads", *IEEE Transactions on Power Systems*, vol. 30, no. 1, pp. 189–198, jan 2015.

[16] Hitoshi Yano, Koji Kudo, Takashi Ikegami, Hiroto Iguchi, Kazuto Kataoka, and Kazuhiko Ogimoto, "A novel charging-time control method for numerous EVs based on a period weighted prescheduling for power supply and demand balancing", in *2012 IEEE PES Innovative Smart Grid Technologies (ISGT)*. jan 2012, pp. 1–6, IEEE.

[17] Naoya Motegi, M.A. Piette, D.S. Watson, Sila Kiliccote, and P. Xu, "Introduction to Commercial Building Control Strategies and Techniques for Demand Response – Appendices", Tech. Rep. 500, Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA (United States), may 2007.

[18] Csaba Szepesvári, "Algorithms for Reinforcement Learning", *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 4, no. 1, pp. 1–103, jan 2010.

[19] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem", *Machine Learning*, vol. 47, no. 2/3, pp. 235–256, 2002.

[20] T. L. Lai and Herbert Robbins, "Asymptotically efficient adaptive allocation rules", *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[21] Jean Yves Audibert, Rémi Munos, and Csaba Szepesvári, "Exploration-exploitation tradeoff using variance estimates in multi-armed bandits", *Theoretical Computer Science*, vol. 410, no. 19, pp. 1876–1902, 2009.

[22] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire, "A contextual-bandit approach to personalized news article recommendation", in *Proceedings of the 19th international conference on World wide web - WWW '10*, New York, New York, USA, 2010, p. 661, ACM Press.

[23] Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvári, "Improved Algorithms for Linear Stochastic Bandits", *Neural Information Processing Systems*, pp. 1–19, 2011.

[24] Peter Auer, "Using Confidence Bounds for Exploitation-Exploration Trade-offs", *Journal of Machine Learning Research*, vol. 3, no. 3, pp. 397–422, 2003.

[25] Shipra Agrawal and Navin Goyal, "Thompson Sampling for Contextual Bandits with Linear Payoffs", *Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA*, vol. 28, sep 2013.

[26] H.A. Aalami, M. Parsa Moghaddam, and G.R. Yousefi, "Modeling and prioritizing demand response programs in power markets", *Electric Power Systems Research*, vol. 80, no. 4, pp. 426–435, apr 2010.

[27] Mikhail A. Lisovich, Deirdre K. Mulligan, and Stephen B. Wicker, "Inferring Personal Information from Demand-Response Systems", *IEEE Security & Privacy Magazine*, vol. 8, no. 1, pp. 11–20, jan 2010.

[28] H.A. Aalami, M. Parsa Moghaddam, and G.R. Yousefi, "Demand response modeling considering Interruptible/Curtailable loads and capacity market programs", *Applied Energy*, vol. 87, no. 1, pp. 243–250, jan 2010.