Optimal Motivation Scheme Design using Machine Learning and Control Theory

Oksana Chernova[1], Sitadri Roy Choudhuri[2], Damian Jelito[3], Jedrzej Kardach[4], Kacper Kulczyck[5], Zofia Michalik[6], Kinga Tikosi[7], Agnieszka Rygiel[8], Anton Yurchenko-Tytarenko[1]

[1] *Taras Schevchenko National University of Kyiv, Ukraine* [2] *Indian Institute of Technology, Mumbai, India*
[3] *Jagiellonian University, Krakow, Poland*
[4] *Roomsage*
[5] *Avidata, Poland*
[6] *University of Warsaw*
[7] *Central European University Budapest, Hungary*
[8] *Krakow University of Economics, Poland*

(*Communicated to* MIIR *on 31 October 2021*)

Study Group: ECGI 144, 17–22 March, 2019, Polish Academy of Sciences, Warsaw, Poland

Communicated by: Kamil Kulesza

Industrial Partner: one2tribe

Presenter: Wojciech Ozimek

Team Members: Artur Chabowski, Mabion, Łodz; Oksana Chernova, Taras Schevchenko National University of Kyiv; Sitadri Roy Choudhuri, Indian Institute of Technology, Bombay; Damian Jelito, Jagiellonian University; Jedrzej Kardach, Roomsage; Krzysztof Kepczynski, University of Wrocław; Kacper Kulczycki, Avidata; Kamil Kulesza[9], Polish Academy of Sciences; Zofia Michalik, University of Warsaw; Adam Paszkiewicz, University of Łodz; Marcin Pitera, Jagiellonian University, Krakow; Dariusz Socha, Warsaw University of Technology; Leszek Staryszak; Kinga Tikosi, Central European University Budapest; Agnieszka Rygiel, Krakow University of Economics; Anton Yurchenko-Tytarenko, Taras Schevchenko National University of Kyiv; Aleksandra Zdunek, Warsaw University of Technology

Industrial Sector: Social

Key Words: Stress dynamics, Control theory, Machine learning

MSC2020 Codes: 93, 62

Summary

The aim is to construct a mathematical model of stress dynamics that takes into account stress level and response of employees to different motivation factors. It suggests an optimal motivational scenario that maximizes "profit", i.e. provides tradeoff between the best business outcome and the stress level. It consists of task-reward sequences (i.e. scenarios that include tasks and reinforcement delivered when they are completed). We propose a statistical inference method for the model based on data.

[9] Corresponding contributor: `kamil.kulesza@maths.com.pl`

Contents

1 Introduction

1.1 Problem formulation

Oxford dictionary defines stress as "*physiological disturbance or damage caused to an organism by adverse circumstances*". According to World Health Organization (WHO) nearly one in three European workers (more than 40 million people) report that they are affected by stress at work [8]. Although a certain small level of stress can help an employee to stay focused, energetic, and able to meet new challenges in the workplace, it becomes dangerous when stress is protracted. As it was stated in the problem description, "the losses are not only measured in billions of euro and dollars, but are a matter of something less tangible, but far more valuable – human lives".

People experience stress when they perceive that there is an imbalance between the demands made of them and the resources available to cope with those demands (e.g. knowledge, abilities, control, etc).

There exists a wide range of methods to measure employee's stress level. For instance, *Questionnaire on psychosocial working conditions* that was developed and piloted in 2007 by the Central Institute of Labor Protection as a means of supporting workers suffering from the negative consequences of work-related stress. In reality, an employee could be stressed due to some personal troubles which cannot be influenced by a company, but an employer still has to cope with such uncontrollable stress bursts as they directly influence the productivity.

Our aim is to construct a mathematical model of stress dynamics that takes into account stress level and response of employees to different motivation factors. It suggests an optimal motivational scenario that maximizes "profit", i.e. provides tradeoff between the best business outcome and the stress level. It consists of task-reward sequences (i.e. scenarios that include tasks and reinforcement delivered when they are completed). We propose a statistical inference method for the model based on data.

1.2 Psychological background

In this subsection we briefly describe the psychological literature research we made when analysing the problem. The proposed model is intended to be backed by the theory of psychology. However, it should be noted that none of the authors is a qualified psychologist.

The relationships between stress, motivation and performance seem to be a frequently discussed issue in the psychological literature, see e.g. [12] for a detailed review. However,

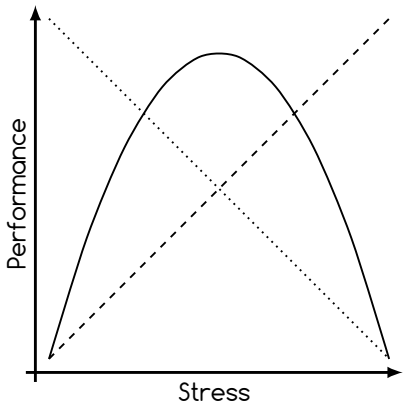


Figure 1: Stylised relationship between stress and performance.

we have not found one unambiguous and widely-accepted model of these relations. Based on the article [9] one can divide the models into three main groups:

- negative linear relationship - underpinned by the belief that high stress level decreases performance;
- positive linear relationship - stress and anxiety are challenges that improve performance;
- combination of the above (inverted-U, Yerkes-Dodson model) - there is the optimal stress level and deviation in both directions reduces the performance, see [15] as the seminal paper in this direction and [13] for a modern treatment.

The models are summarised in Figure 1. In fact, similar relationships may be used to describe dependence between stress and motivation (willinges to work) or motivation (reward) and performance. However, it should be noted that every model has been criticised based on the empirical data, see e.g. [14]. Moreover, problems with correct definition and methodology of measuring stress should be emphasised. One can measure subjective feelings that clearly may be biased or try to find objective measure. See e.g. [7, 12] for further discussion.

It is clear that the problems mentioned above are outside of the domain of mathematics. However, when using the proposed model one has to decide on the cost functional that reflects employer's belief on employees' performance and behaviour (see Section 3 for details). Hence, the thorough psychological study must be performed.

1.3 Solution and document overview

Our task was to create a tool for designing motivation system. We propose the solution that combines various techniques taken from statistics, machine learning and stochastic control theory. In Section 2 we describe changes of stress level in terms of a controlled Markov chain. By applying appropriate control (tasks and rewards) the employer affects the employee's stress level and the company's cost. We express the business goal in terms of various cost functions; see Section 3 for details. In order to use the proposed solution, the employer needs to have the transition probability matrix of the Markov chain. In Section 4 we propose two methods of estimation – logistic-type regression and Bayesian updating rule. Once the transition matrix is given, one can apply the algorithm from 5 to obtain optimal solution. In Section 6 we present illustrative implementation of the proposed motivation scheme. Further ideas are discussed in Section 7.

2 Model design

We want to model the stress level of an employee by a controlled Markov chain (see e.g. [1]). Here, the control reflects actions that can be done by the employer and affect the employee's stress level. This includes e.g. setting sales targets and possible rewards for meeting them. Using a Markov chain we are able to capture some form of inertia in stress level, i.e. the fact that it should depend on the stress in the past. Moreover, the random nature reflects some external circumstances that affect stress level, e.g. related to private life and not described by the model.

Consider a single employee or a homogeneous group of employees. Denote $\mathcal{A}$ – the set of possible actions that can be performed by the employer towards their employee for a given day (period). For computational reasons, we assume that the number of possible actions is finite, i.e. $|\mathcal{A}| = K$, but this assumption may be relaxed. Moreover, for convenience (both interpretational and mathematical), we will assume that each $a \in \mathcal{A}$ has the form $a = (a^{(d)}, a^{(r)})$, where $a^{(d)}$ is a quantified "business value" of the task (target, demand) and $a^{(r)}$ is a quantified "reward" that were given to an employee.

We assume that the employee always meets the target, possibly increasing their level of stress. Hence, the reward for meeting the target is always paid and from the perspective of the employer $a^{(r)}$ describes the cost of the action while $a^{(d)}$ denotes the income. We refer to Section 7 for possible modifications of this approach.

Let $S = (S_0, S_1, \ldots)$ be a time-homogeneous controlled Markov chain on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$. The chain S describes the stress level of the particular employee. Here

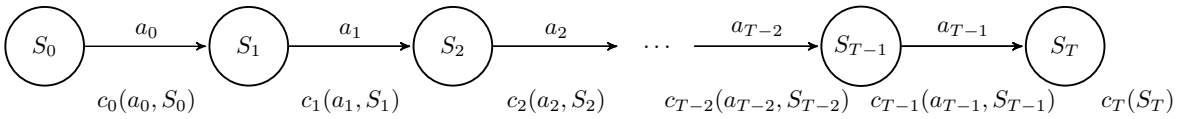


Figure 2: Dynamics of the model. At time t the control a_t affects the probability of transition from the stress level S_t to S_{t+1} and results in the cost $c_t(a_t, S_t)$.

the time corresponds to subsequent days (periods) of the considered motivation scheme. We assume that the stress level is measured in the discrete scale and can take only finite number of states. Hence, $\mathcal{S} := \{1, \dots, N\}$ will denote the state space for the Markov chain S with the convention that higher number corresponds to higher stress level. It should be explicitly stated that in our model the stress level is directly observable, i.e. we are able to reliably measure it.

To each action $a \in \mathcal{A}$, we associate a probability transition matrix

$$P(a) = \{P_{ij}(a)\}_{i,j=1}^{N}$$

of size $N \times N$ that represents the probabilities of changing the stress level from i to j given a particular action a of the employer, i.e.

$$P_{ij}(a) = \mathbb{P}\left(S_{t+1} = j \mid S_t = i, a_t = a\right), \quad t \in \mathbb{N}.$$

With the slight abuse of notation define $a_t : \mathcal{S} \mapsto \mathcal{A}$ as the mapping (decision rule) that specifies the action undertaken by the employer at time t depending on the current stress level. In other words, if the stress level at time t is given by s, then the employer apply control $a_t(s)$. By $\underline{a} = (a_0, a_1, a_2, \dots)$ we denote the policy of the employer, that is a sequence of decision rules. It should be clear from the context when a_t denotes the decision rule (function) and when it describes the control itself (the value of the function).

Let $T \in \mathbb{N}$ denote the horizon of the control problem, i.e. number of days (periods) to the end of the project and considered motivation scheme. In this case by $C_T(s_0, \underline{a})$ we denote the expected cost corresponding to the initial stress level s_0 of the employee and the policy $\underline{a}$. For simplicity of the argument we consider the cost functional in the additive form, i.e.

$$C_T(s_0, \underline{a}) := \mathbb{E}_{s_0}\left(\sum_{t=0}^{T-1} c_t(a_t, S_t) + c_T(S_T)\right), \tag{2.1}$$

where $\mathbb{E}_{s_0}$ denotes the conditional expectation given that $S_0 = s_0$. Here $c_t(a, s)$ is an intertemporal cost at time t of undertaking an action a, when the current level of employee's stress is s. At the final step the employer does not take any action, so the cost at time T depends only on the terminal stress level. For possible choices of intertemporal cost functions see Section 3. The dynamics of the model is summarised in Figure 2.

The objective of the employer is to apply a policy $\underline{a}$ that minimizes $C_T(s_0, \underline{a})$ given initial state s_0. With the slight abuse of notation we denote by $C_T(s_0)$ the optimal cost, i.e.

$$C_T(s_0) := \inf_{\underline{a}} C_T(s_0, \underline{a}). \tag{2.2}$$

The optimal policy, if exists, is then

$$\underline{a}^* = \operatorname{argmin} C_T(s_0, \underline{a}).$$

In addition to the problem (2.2) we consider infinite horizon functional given by

$$C_\infty(s_0) = \inf_{\underline{a}} \mathbb{E}_{s_0} \left(\sum_{t=0}^{\infty} c_t(a_t, S_t) \right), \tag{2.3}$$

assuming appropriate convergence of the sum. This approach may be useful if the duration of the project is unknown or the employer wants to take into account long-run balance between employee's wellness and the company standing. See Subsection 5.2 for further assumptions on c_t and Section 7 for generalisations of infinite horizon model.

Method for finding the optimal policy is described in Section 5.

3 Performance quantification

One of the most crucial points is the appropriate choice of the intertemporal cost function c_t. Recall that a possible action taken by the employer consists of $a^{(d)}$ – the business value of the task demanded by the employee – and $a^{(r)}$ – employee's reward for fulfilling the task. Here we should note that the task does not have to provide a direct profit, but it may also consist of learning or teaching activities. Nonetheless, we can also associate monetary value to such tasks.

The cost function should not only take into account the direct cost of an action a (that is, cost of $a^{(r)}$ and profit from $a^{(d)}$), but also help managing the stress level of an employee, for example by introducing a penalty for certain stress levels or adding the cost of hiring a new worker in case an employee decides to leave the company. Please note that while it is straightforward that the cost of an action and penalty for leaving have monetary values, it is less obvious how to interpret the penalty for stress as a real cost. However, by introducing this penalty we take into account the situation when the overstressed worker is less productive and therefore does not meet the target completely or needs more time to finish it. Hence the penalty for stress may be interpreted as the indirect cost of a lower efficiency of an employee. Therefore we assume the general form of the cost function:

$$c_t(s, a) = \text{cost of } a + \text{ penalty for } s \ (+ \text{ leaving cost }). \tag{3.1}$$

Here we briefly discuss each component of the formula (3.1).

1. **Cost of a.**
   Let us denote this cost by $g(a)$. The minimal requirement for g is being decreasing in $a^{(d)}$ and increasing in $a^{(r)}$. The choice of function g depends on the employer's preference, may be treated as a form of utility of $a^{(d)}$ and disutility $a^{(r)}$. Here we list a few functions, which are mostly inspired by the utility theory.

   - Linear function
     $$g(a) = \alpha a^{(r)} - \beta a^{(d)}, \text{ where } \alpha, \beta > 0.$$
     The main advantage of this function is its simplicity and computational feasibility.
   - Assuming $a^{(r)}, a^{(d)} > 0$, we may apply logarithmic function
     $$g(a) = \alpha \ln a^{(r)} - \beta \ln a^{(d)}, \text{ for } \alpha, \beta > 0.$$
     It is concave in $a^{(r)}$ and convex in $a^{(d)}$ – both properties reflect the following fact: When the value of $a^{(r)}$ (respectively $a^{(d)}$) is high, a perturbation of this value implies a smaller change of the cost than the same perturbation when $a^{(r)}$ (respectively $a^{(d)}$) is small. If $\alpha = \beta$, then the function depends only on the ratio $\frac{a^{(r)}}{a^{(d)}}$. The choice of logarithmic utility corresponds to Kelly criterion of maximisation long run investment return.
   - Under the same assumptions on $a^{(r)}$, $a^{(d)}$, we can use power function
     $$g(a) = \alpha \left(\frac{a^{(r)}}{a^{(d)}} \right)^{\beta} \text{ for } \alpha \geq 0, \ \beta \in (0, 1).$$
     As in the previous example, this function is concave in $a^{(r)}$ and convex in $a^{(d)}$.

2. **Penalty for stress**
   The penalty for stress will be denoted by $p(s)$. The choice of p depends directly on the employer's view on how stress affects the efficiency of an employee. As it was discussed in Subsection 1.2, there are three main hypotheses about the relation between stress and performance, each of whom may be reflected in the penalty $p(s)$.

   - In order to follow the "negative linear" approach mentioned in [9], one should choose p to be strictly increasing in s. The simplest (and least computationally demanding) example is a linear function $p(s) = \alpha s$ for some $\alpha > 0$.
   - In order to be consistent with the "inverted-U" model from [15] and [13], one needs to choose a preferable stress level $\hat{s}$ (usually in the middle of scale) and define e.g. $p(s) = \alpha |s - \hat{s}|$ or $p(s) = \alpha (s - \hat{s})^2$ for $\alpha > 0$. These functions are symmetric with respect to $\hat{s}$ and penalize the deviation from the optimal level of stress in both directions.

- A slight modification of the approach proposed above is choosing $p(s) = \alpha(s - \hat{s})^+$, $\alpha > 0$. Here $x^+ := \max(0, x)$ stands for the positive part of x. Such a function penalizes only exceeding $\hat{s}$ and is indifferent to stress levels below $\hat{s}$. This behaviour resembles the properties of semivariance.
- Another variation of these approaches is to penalise deviations from $\hat{s}$ in an asymmetric way. This may be reflected in a function $p(s) = \alpha(s - \hat{s})^+ + \beta(s - \hat{s})^-$ for positive $\alpha \neq \beta$. The notion $x^- := \max(0, -x)$ stands for the negative part of x.

3. **Leaving cost.**
   We may distinguish the state $\Delta := N$ as a situation where the employee decides to leave the company (either due to being overstressed or to external factors, such as a better offer from another company). Then, the new employee may be immediately hired hence we do not need to assume that the state Δ is absorbing. However such situation generates a substantial cost (which we denote by L) for the employer and is rather undesirable. Therefore we may add $L\mathbb{1}_{\{s=\Delta\}}$ to the cost function.

   A remark should be stated at this point – if we penalise high stress levels using the function p mentioned in point 2, adding a leaving cost may lead to penalising stress twice: once (explicitly) via $p(s)$, and second time by increasing the probability of reaching level Δ, which increases the conditional expectation of $L\mathbb{1}_{\{s=\Delta\}}$. Secondly, the event of an employee leaving the company is rather rare and it might be difficult to estimate the probabilities of reaching the state Δ.

Remark 3.1. All functions proposed above are time-independent. However, this property may be easily relaxed, for example by allowing the multiplicative constants that appear in the formulas to depend on t. This enables us to capture the fact that the goals of the company may evolve during the time period of the project – e.g. the employer might be more likely to encourage more work at the beginning or in the middle of the project. On the other hand, in the infinite horizon case we usually assume that $c_t(a, s) = \gamma^t c(a, s)$, for $\gamma \in (0, 1)$, in order to ensure convergence of the sum in (2.3), see Subsection 5.2.

Remark 3.2. As mentioned in Section 2, at the final step the employer does not perform any action, so the terminal cost c_T does not depend on a. Therefore c_T consist only of the penalty for stress and possibly of leaving cost. Both may be constructed as proposed in points 2 and 3.

Remark 3.3. Please note that using penalty for stress (that is measured in some abstract scale) we may lose direct interpretability of the cost functional in monetary units. However, if p is piecewise linear, the multiplicative constants α and β may be interpreted as the cost of changing stress level by one. We may associate monetary value to this according to the interpretation of the penalty as the cost of lower productivity of an employee. The direct interpretability of cost functional in monetary units is lost also when introducing non-linear

cost of control (e.g. power function). Note that we do not face this problem in case of penalty for leaving.

4 Bayesian learning for transition probabilities estimation

In this section, we present a Bayesian method for estimating Markov transition probabilities. We split it into three parts. In Subsection 4.1 we recall several definitions and notation; in Subsection 4.2 we introduce general framework and discuss necessary updating algorithms; in Subsection 4.3, we describe one of possible ways to estimate priors.

For more details on Bayesian statistics in general and its applications in machine learning, see [3] or [10].

4.1 Definitions and notation

For reader's convenience, we start with recalling several definitions and notation that will be used in what follows.

Definition 4.1. Random vector $\mathbf{X} = (X_1, ..., X_d)$ has **multinomial** distribution with vector of parameters $(m; p_1, .., p_d)$, $m \in \mathbb{N}$, $p_i > 0$, $\sum_{i=1}^d p_i = 1$, if

$$\mathbb{P}(X_1 = x_1, ..., X_d = x_d) = \begin{cases} \frac{d!}{x_1! \dots x_d!} p_1^{x_1} \dots p_d^{x_d}, & \text{when } \sum_{i=1}^d x_i = m, \\ 0 & \text{otherwise}, \end{cases}$$

for non-negative integers x_1, ..., x_d.

Remark 4.2. Multinomial distribution is a generalization of binomial distribution in case of d possible outcomes. In this case parameter m has an interpretation of the number of trials while p_1,..., p_d are probabilities of different outcomes in each trial.

Definition 4.3. The **Dirichlet distribution** of order $d \geq 2$ with parameters α_1, ..., $\alpha_d > 0$ is a continuous distribution with probability density

$$f(x_1, ..., x_d) = \left(\frac{\prod_{i=1}^d \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^d \alpha_i)} \prod_{i=1}^d x_i^{\alpha_i - 1} \right) \mathbb{1}_{\{\sum_{i=1}^d x_i = 1 \text{ and } x_i \geq 0,\ i=1,\dots,d\}}.$$

Remark 4.4. We shall denote random vector $\mathbf{X}$ with multinomial distribution with parameters $(m, p_1, ..., p_d)$ as

$$\mathbf{X} \sim \text{Multinom}(m; p_1, ..., p_d)$$

and the random vector $\mathbf{P}$ with Dirichlet distribution with parameters $(\alpha_1, ..., \alpha_d)$ as

$$\mathbf{P} \sim \text{Dir}(\alpha_1, ..., \alpha_d).$$

Denote

$$P_{i*}(a) := (P_{i1}(a), P_{i2}(a), ..., P_{iN}(a)), \quad i = 1, ..., N, \quad a \in \mathcal{A}$$

the i-th row of transition matrix $P(a)$ that contains probabilities of changing the stress level *from* i given a certain action a of the employer.

4.2 Bayesian model of stress profile

Consider day t and an employee with the stress level i with employer's action a on this day. Hereafter by a stress profile we mean the family of the transition probabilities $P(a), a \in \mathcal{A}$.

Note that stress level i of the employee can be represented as an N-dimensional vector with 1 at the i-th coordinate and zeroes at all the others. Such representation is convenient as in this case the stress level the next day is a multinomial random variable with "number of trials" parameter equal to 1 and vector of outcome probabilities $P_{i*}(a)$ that depends on current stress level and action performed by the employer. Denote this random variable $S_{t+1,i,a}$, i.e.

$$S_{t+1,i,a} := S_{t+1} \mid S_t = i,\ a.$$

Assume that vector $P_{i*}(a)$ is random with Dirichlet distribution with parameters $\alpha_1(a,i)$, ..., $\alpha_N(a,i)$, i.e. consider the following Bayesian model of the stress profile:

$$\begin{aligned} S_{t+1,i,a} &\sim \text{Multinom}(1; P_{i*}(a)), \\ P_{i*}(a) &\sim \text{Dir}(\alpha_1(a,i), ..., \alpha_N(a,i)). \end{aligned} \tag{4.1}$$

Assume that $S_{t+1,i,a} = \mathbf{y}$, where $\mathbf{y} = (y_1, ..., y_N)$ is a vector with all zeroes except for the position (which is 1) where the stress level was on day $t+1$. In this case, denoting $\mathbf{x} = (x_1, ..., x_N)$, according to the Bayes' theorem,

$$\begin{aligned} f_{P_{i*}(a)|S_{t+1,i,a}=\mathbf{y}}(\mathbf{x}) \propto f_{S_{t+1,i,a}|P_{i*}(a)=\mathbf{x}}\,(\mathbf{y}) f_{P_{i*}(a)}(\mathbf{x}) \propto \\ \propto \prod_{j=1}^{N} x_j^{y_j} \prod_{j=1}^{N} x_j^{\alpha_j(a,i)-1} = \prod_{j=1}^{N} x_j^{y_j+\alpha_j(a,i)-1}, \end{aligned} \tag{4.2}$$

where

$$f_{P_{i*}(a)|S_{t+1,i,a}=\mathbf{y}}(\mathbf{x})$$

is a conditional density of $P_{i*}(a)$ given $S_{t+1,i,a} = \mathbf{y}$,

$$f_{S_{t+1,i,a}|P_{i*}(a)=\mathbf{x}}(\mathbf{y})$$

is a conditional density of $S_{t+1,i,a}$ given $P_{i*}(a) = \mathbf{x}$ and

$$f_{P_{i*}(a)}(\mathbf{x})$$

is a density of $P_{i*}(a)$.

Therefore, due to the form of the Dirichlet distribution density,

$$P_{i*}(a) \mid S_{t+1,i,a} = \mathbf{y} \sim \mathrm{Dir}(\alpha_1(a,i) + y_1, ..., \alpha_N(a,i) + y_N).$$

This gives us a simple and efficient online algorithm to update our beliefs concerning the transition probability vectors in the stress profile.

Updating Rule 4.5. Given

- stress level i on day t,
- action a on day t,
- Dirichlet distribution parameters for the day t:

$$(\alpha_1^t(a,i), ..., \alpha_N^t(a,i)),$$

- stress level j on day $t+1$,

the **updated Dirichlet distribution parameters** for the day $t+1$

$$(\alpha_1^{t+1}(a,i), ..., \alpha_N^{t+1}(a,i))$$

are computed as follows:

$$\alpha_j^{t+1}(a,i) = \alpha_j^t(a,i) + 1,$$
$$\alpha_k^{t+1}(a,i) = \alpha_k^t(a,i), \quad \forall k \neq j.$$

Remark 4.6. Note that at each step we update only one row of the matrix $P(a)$ for the given action a. It is easy to see that, in general, there are $K \times N$ of such vectors in the stress profile.

Remark 4.7. As a result of each update, we obtain a *distribution* of the transition probability vector $P_{i*}(a)$. In order to obtain the point estimates for $P_{i*}(a)$, we can use, for example, **posterior mean** or **maximum a posteriori** (MAP), if all parameters of the Dirichlet distribution are greater than 1.

Namely, if

$$P_{i*}(a) = (P_{i1}(a), ..., P_{i,N}(a)) \sim \mathrm{Dir}(\alpha_1(a,i), ..., \alpha_N(a,i)).$$

then

- $\mathbb{E}\left[P_{ij}(a)\right] = \frac{\alpha_j(a,i)}{\sum_{k=1}^N \alpha_k(a,i)}$,
- $\mathrm{MAP}\left(P_{ij}(a)\right) = \frac{\alpha_j(a,i)-1}{\sum_{k=1}^N \alpha_k(a,i)-N}$.

4.3 Estimating priors

As it is stated in Remark 4.6, there are $K \times N$ vectors to estimate, which requires commensurate database size to train the model to an appropriate level.

One of possible approaches lies in clustering the employees into homogenuous groups and use the data from all the group members to update the probabilities. In such case, if d employees from the group have the same stress level i and are asked to perform the same task a, the update formulas for $P_{i*}(a)$ should be applied d times.

However, in order to make the model more convenient on early stages of learning, it is useful to choose meaningful initial values for parameters

$$(\alpha_1(a,i),...,\alpha_N(a,i)), \quad i=1,...,N, \quad a \in \mathcal{A},$$

to plug in into the model in the beginning.

One may collect initial dataset consisting of the stress level measurement before application of the control, the control itself and the stress level after the control. Then an N-variate logistic regression-type model of the following form may be fitted

$$\begin{aligned} P_{ij}(a^{(d)},a^{(r)}) &= \mathbb{P}\left\{S_1=j \mid S_0=i,\ a_0=(a^{(d)},a^{(r)})\right\} \\ &= \frac{\exp\left\{\beta_0^j+\beta_1^j a^{(d)}+\beta_2^j a^{(r)}+\beta_3^j|j-i|^+ +\beta_4^j|j-i|^-\right\}}{1+\sum_{k=1}^{N-1}\exp\left\{\beta_0^k+\beta_1^k a^{(d)}+\beta_2^k a^{(r)}+\beta_3^k|k-i|^+ +\beta_4^k|k-i|^-\right\}}, \quad j=1,...,N-1; \\ P_{i,N}(a^{(d)},a^{(r)}) &= 1-\sum_{j=1}^{N-1} P_{i,j}(a^{(d)},a^{(r)}), \end{aligned} \tag{4.3}$$

with some parameters $\beta_u^j, u=0,\dots,4,\ j=1,\dots,N-1$ to be found. Here the terms $(j-i)^+ := \max\{j-i,0\}$ and $(j-i)^- := \max\{i-j,0\}$ correspond to up and down movement in the stress level respectively and penalize drastic day-to-day changes in stress.

Note that in (4.3) we used logistic transform mapping $\mathbb{R}$ to $[0,1]$. In general one may use some different function from the theory of generalised linear models, see e.g. [10, Chapter 9] .

After obtaining estimates $\hat{P}_{ij}(a)$ of transition probabilities given the state a, we set initial parameters of the Dirichlet distributions $\mathrm{Dir}(\alpha_1^0(a,i),...,\alpha_N^0(a,i))$ as follows:

$$\alpha_j^0(a,i) = \hat{P}_{ij}(a) \times L_{P_{i*}(a)},$$

where $L_{P_{i*}(a)} > 0$ are positive constants that should be chosen heuristically. We suggest to set

$$L_{P_{i*}(a)} = N,$$

but further data-based tuning is required.

Remark 4.8. The parameters of the Dirichlet distribution should be positive, therefore the estimates $\hat{P}_{ij}(a)$ should be positive as well. The estimates obtained by logistic regression-type method described above have this property.

The interpretation of $L_{P_{i*}(a)}$ is as follows. Let

$$P_{i*}(a) = (P_{i1}(a), ..., P_{i,N}(a)) \sim \text{Dir}(\alpha_1(a,i), ..., \alpha_N(a,i)).$$

Recall the posteriori mean point estimates of $P_{ij}(a)$, $j = 1, ..., N$, given in Remark 4.7. If we equate them to estimates $\hat{P}_{ij}(a)$ obtained with any other method (for instance, logistic regression described above) we get:

$$\frac{\alpha_j^0(a,i)}{\sum_{k=1}^N \alpha_k^0(a,i)} = \hat{P}_{ij}(a),$$
$$\alpha_j^0(a,i) = \hat{P}_{ij}(a) \times \left(\sum_{k=1}^N \alpha_k^0(a,i) \right) =: \hat{P}_{ij}(a) \times L_{P_{i*}(a)},$$

i.e. $L_{P_{i*}(a)} := \sum_{k=1}^N \alpha_k^0(a,i)$.

Note that

$$\text{Var}\left[P_{ij}(a)\right] = \frac{\frac{\alpha_j^0(a,i)}{\sum_{k=1}^N \alpha_k^0(a,i)} \left(1 - \frac{\alpha_j^0(a,i)}{\sum_{k=1}^N \alpha_k^0(a,i)}\right)}{\sum_{k=1}^N \alpha_k(a,i)} = \frac{\frac{\alpha_j^0(a,i)}{\sum_{k=1}^N \alpha_k^0(a,i)} \left(1 - \frac{\alpha_j^0(a,i)}{\sum_{k=1}^N \alpha_k^0(a,i)}\right)}{L_{P_{i*}(a)}},$$

so the constants $L_{P_{i*}(a)}$ in the denominator tune marginal variances of the Dirichlet distributions. In oher words, the greater the values of $L_{P_{i*}(a)}$ are, the smaller the variances of marginals become.

We can use this fact to take into account the stress profile changes in time: division of parameters of the Dirichlet distribution by a constant $C > 1$ does not change posteriori mean point estimates but increases variance of the transition probabilities. Therefore, weekly or monthly division of model parameters by a "relatively big" constant (chosen, again, heuristically) makes the model emphasize on new observations and re-train faster, adapting to new changes swiftly.

5 Solution of the Markov decision model

The solution of the Markov decision model stated in the Section 2 will be based on the Bellman equations and the so-called reward iteration. For the sake of convenience, we sketch this approach below. Interested reader may consult e.g. [1], [5], [6] for a thorough description and mathematical details of the problem.

5.1 Finite horizon case

For function $v : \mathcal{S} \mapsto \mathbb{R}$ define the minimal reward operators

$$\mathcal{T}_t v(s) := \inf_{a \in \mathcal{A}} \left(c_t(a, s) + \mathbb{E}\left(v(S_{t+1}) | S_t = s, a_t = a \right) \right) \tag{5.1}$$

where $t = 0, \ldots, T-1$. Consider in addition the system of recursive equations

$$\begin{cases} V_T = c_T, \\ V_t = \mathcal{T}_t V_{t+1} \quad \text{for} \quad t = 0, \ldots, T-1, \end{cases} \tag{5.2}$$

where we look for functions $V_0, \ldots, V_T : \mathcal{S} \mapsto \mathbb{R}$.

Optimal strategy for (2.2) may be found using Theorem 5.1 below. In a nutshell, if we can find a solution to the system of equations (5.2) and minimisers for (5.1), then we can construct an optimal strategy for the problem (2.2). For the proof and further discussion see e.g. [1, Theorem 2.3.7].

Theorem 5.1. *Suppose* (V_t) *are solutions to the system of equations* (5.2)*. Moreover, assume* (a_t^*) *are minimisers for* (5.1) *and* (V_t)*, i.e.*

$$V_t(s) = c_t(a_t^*(s), s) + \mathbb{E}\left(V_{t+1}(S_{t+1}) | S_t = s, a_t = a_t^*(s) \right), \quad s \in \mathcal{S}.$$

Then the policy $\underline{a}^* := (a_t^*)$ *is optimal for* (2.2) *and*

$$C_T(s_0) = C_T(s_0, \underline{a}^*) = V_0(s_0).$$

Note that in our case (5.2) reads as

$$\begin{cases} V_T(s) = c_T(s), \\ V_t(s) = \min\limits_{a \in \mathcal{A}} \left(c_t(a, s) + \sum\limits_{j=1}^{N} V_{t+1}(j) P_{ij}(a) \right) \quad \text{for} \quad t = 0, \ldots, T-1 \end{cases}$$

for any $s \in \mathcal{S}$. Since we assumed finiteness of the control set $\mathcal{A}$, existence of the minimisers is guaranteed. However, the Theorem can be applied in the more general framework; see [1] for details and necessary technical conditions. Moreover, based on Theorem 5.1, one can easily construct optimisation algorithm.

Algorithm 5.2. Backward induction

1. Set $t \leftarrow T$. For any $s \in \mathcal{S}$ compute

$$V_T(s) \leftarrow c_T(s).$$

2. Set $t \leftarrow t-1$. For any $s \in \mathcal{S}$ compute

$$V_t(s) \leftarrow \min_{a \in \mathcal{A}} \left(c_t(a,s) + \sum_{j=1}^{N} V_{t+1}(j) P_{ij}(a) \right).$$

   Denote by $a_t^*(s)$ the minimiser.

3. Repeat step 2 until $t = 0$. Then $\underline{a}^* := (a_t^*)_{t=0}^{T-1}$ is the optimal policy and $C_T(s) = V_T(s)$.

We refer to Section 6 for computational example.

5.2 Stationary and infinite horizon case

Now we restrict our attention to the intertemporal cost functions of the form

$$c_t(s,a) = \gamma^t c(a,s) \tag{5.3}$$

for some function $c : \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$ and a constant $\gamma \in (0,1)$. Hence, the cost function depends on time only through the "discount factor" γ. Assumption $\gamma \in (0,1)$ guarantees convergence of the sum in (2.3).

In analogy to (5.1) we introduce the following operator

$$\mathcal{T}v(s) := \inf_{a \in \mathcal{A}} \left(c(a,s) + \gamma \mathbb{E}\left(v(S_1) | S_0 = s, a_0 = a \right) \right). \tag{5.4}$$

Moreover, let

$$\begin{cases} V_0 = 0, \\ V_t = \mathcal{T} V_{t-1} \quad \text{for} \quad t = 1, 2, \dots. \end{cases} \tag{5.5}$$

Note that in contrast to (5.2), in (5.5) we used forward recursion.

The following theorem provides a method for finding optimal strategy for problem (2.3). For proof, see e.g. [1, Theorem 7.1.7, Theorem 7.1.8].

Theorem 5.3. *Suppose $V := \lim_{t \to \infty} V_t$ is a bounded function such that*

$$V(s) = \mathcal{T} V(s).$$

Assume in addition that there exists $a^* : \mathcal{S} \mapsto \mathcal{A}$ satisfying for any $s \in \mathcal{S}$

$$\mathcal{T} V(s) = c(a^*(s), s) + \gamma \mathbb{E}\left(V(S_1) | S_0 = s, a_0 = a^*(s) \right).$$

Then, the policy $\underline{a}^* = (a^*)$ is optimal and

$$C_\infty(s_0) = V(s_0).$$

One can show that operator $\mathcal{T}$ is monotone, hence $\lim_{t\to\infty} V_t$ from Theorem 5.3 exists. By assumption, the function V is a fixed point of the operator $\mathcal{T}$ and a^* is a minimiser for (5.4) and V. Moreover, the optimal control depends on time only through the current state S_t, i.e. for any time t it is optimal to apply policy $a^*(S_t)$.

Let us consider again the finite horizon problem (2.2). Assume that the intertemporal cost functions is of the form (5.3) and the terminal cost is given by

$$c_T(s) = \gamma^T r(s)$$

for some $r : \mathcal{S} \mapsto \mathbb{R}$. Special structure of the cost functions allows us to replace Algorithm 5.2 by the following:

Algorithm 5.4. Forward induction

1. Set $t \leftarrow 0$. For any $s \in \mathcal{S}$ compute
$$V_0(s) \leftarrow r(s).$$
2. Set $t \leftarrow t + 1$. For any $s \in \mathcal{S}$ compute
$$V_t(s) \leftarrow \min_{a\in\mathcal{A}} \left(c(a, s) + \gamma \sum_{j=1}^{N} V_{t-1}(j) P_{ij}(a) \right).$$
Denote by $a_t^*(s)$ the minimiser.
3. Repeat step 2 until $t = T$. Then $\underline{a}^* := (a_{T-t}^*)_{t=0}^{T-1}$ is the optimal policy and $C_T(s) = V_T(s)$.

Note that Algorithm 5.4 combined with (5.5) and Theorem 5.3 provides a method for approximating solution of the infinite horizon problem. One simply needs to take $r \equiv 0$ and repeat step 2 of the Algorithm 5.4 for sufficiently large T. For other methods see [1, Section 7.5].

6 Numerical example

6.1 Setup

Our goal is to determine optimal actions for each time epoch by solving the Bellman equations as described in Section 5. In what follows we present an illustrational example of transition probabilities and cost functions without any references to any databases or needs of particular companies.

Assume that there are five different stress levels $\{1, 2, 3, 4, 5\}$ and only one type of task difficulty, i.e. $a_t^{(d)} = 1$. We define five different reward levels, namely $a_t^{(r)} \in \{1, 2, 3, 4, 5\}$. For the ease of notation a_t will stand for $a_t^{(r)}$.

Remark 6.1. The higher reward levels are associated with an attempt to reduce stress of the employee, which can be interpreted not only by getting a higher reward by the employee but also by having more time to finish the specific task.

We will use the following cost functions:

$$c_t^1(a,s) = 0.99^t(a + (s-1)^+), \qquad c_T^1(s) = 0.99^T(s-1)^+; \tag{6.1}$$

$$c_t^2(a,s) = 0.99^t(a + (s-1)^+ + 10\mathbb{1}_{\{s=5\}}), \quad c_T^2(s) = 0.99^T((s-1)^+ + 10\mathbb{1}_{\{s=5\}}); \tag{6.2}$$

$$c_t^3(a,s) = 0.99^t(a + 10\mathbb{1}_{\{s=5\}}), \qquad c_T^3(s) = 0.99^T(10\mathbb{1}_{\{s=5\}}). \tag{6.3}$$

For more information about choosing the suitable cost function, see Section 3.

The transition probabilities in general need to be estimated from empirical data, see Section 4 for more details. For this implementation we used the transition matrices included in the Appendix A. The matrices we use are meant to capture our belief that the lower rewards tend to shift stress level upwards, while the higher rewards will shift them downwards.

6.2 Results

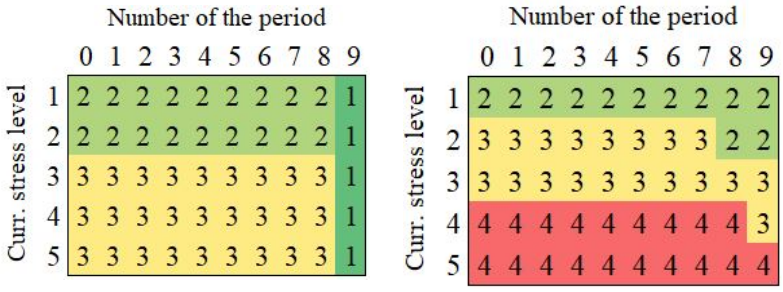
Figure 3 illustrates the optimal policies. The tables show that as time passes (horizontal line) which policy should be applied by the employer, i.e. which reward should be given (the numbers in the boxes) to the employees in the different stress levels (vertical line). E.g. the first column of Figure 3a is

$$\begin{pmatrix} 2 \\ 2 \\ 3 \\ 3 \\ 3 \end{pmatrix},$$

which means that in the first time period the optimal allocation is reward level 2 for stress level 1, reward level 2 for stress level 2, reward level 3 for stress level 3 etc. The second column stands for the optimal policy in the second time period and so on. Note that applying the optimal control for the specific employee one has to take into account that the stress level changes randomly. Hence, as time passes it may be necessary to switch between different rows of the table.

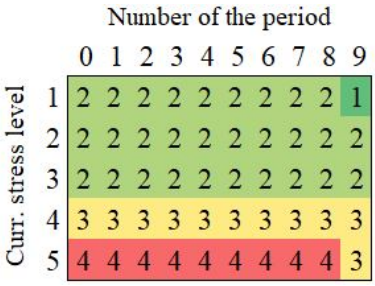
First let us consider cost function c^1 as defined in (6.1). The optimal policy depicted in Figure 3a is to give higher reward to the employees with high stress levels, and lower rewards in case of low stress levels. However, the optimal reward for the last period is to pay the minimum possible amount. If we add an additional penalty for the highest stress level, as in function c^2 from (6.2), we see that the rewards in the last periods grow, which reflects the

fact that in this case the employer tries even harder to avoid this level, see Figure 3b. We can also consider the function c^3 from (6.3) that penalises only the highest stress level. As expected, the results are in between the outcome for c^1 and c^2, see Figure 3c.



(a) Cost function c^1

(b) Cost function c^2



(c) Cost function c^3

Figure 3: Optimal control with cost functions c^1, c^2 and c^3 given by (6.1), (6.2) and (6.3), respectively. The number in the i-th row and t-th column of the table shows the optimal policy for the t-th period, assuming that the current stress level satisfies $S_t = i$.

Remark 6.2. In the examples, in the last periods the employer tends to decrease the reward as they are interested in the short-horizon result only. In this sense the infinite time horizon version is more meaningful. For more references on the topic, see Subsections 5.2 and 7.1.

7 Generalisations and further ideas

During the work, several ideas were proposed for the model. This includes completely different approaches as well as furher extensions of the main model. For various reasons the concepts have not been developed. However, we decided to briefly describe them in this section.

7.1 Modifications of the Markov model

Functional (2.2) takes into account the expected outcome of the scheme performance. However, it does not guarantee stability of the results. We can introduce penalty for big variability of the income including variance in the cost functional, i.e. defining the mean-variance functional

$$C^{MV}(s_0) = \inf_{\underline{a}} \left[\mathbb{E}_{s_0} \left(\sum_{t=0}^{T} c_t(a_t, S_t) \right) + \theta D^2 \left(\sum_{t=0}^{T} c_t(a_t, S_t) \right) \right] \tag{7.1}$$

for some parameter $\theta > 0$. This idea follows the well known Markowitz approach for portfolio selection, see e.g. [1, Section 4.6].

The next extension is based on the cost per time criterion. In this case the problem reads as follows

$$\overline{C}^{\infty}(s_0) = \inf_{\underline{a}} \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_{s_0} \left(\sum_{t=0}^{T} c_t(a_t, S_t) \right). \tag{7.2}$$

Please note that in general, limit in (7.2) may not exist. In this case we should replace it by e.g. limit superior that corresponds to the worst case scenario. Moreover, note that the value of the functional (7.2) does not change if we modify the strategy for finitely many t. See e.g. [5, Chapter 3] for discussion and details.

Combining long-term idea of (7.2) and variability penalisation from (7.1) we can also consider the so-called risk sensitive criterion, i.e.

$$C^{RS}(s_0) = \inf_{\underline{a}} \lim_{T\to\infty} \frac{1}{T\theta} \ln \left[\mathbb{E}_{s_0} e^{\sum_{t=0}^{T} \theta c_t(a_t, S_t)} \right]. \tag{7.3}$$

Using Taylor expansion in (7.3), we recover idea standing behind (7.1), but we take into account also high order moments. Risk-sensitive functional (7.3) and the so-called entropic utility measure gained substantial popularity in the financial context, see e.g. [2].

In the model introduced in Section 2 we assumed that the employee always meets the goal set by the employer. In general, we can describe employee's performance at period t by a random variable p_t. It is natural to assume that distribution of p_t depends on the stress level and the motivation scheme (at least, $a^{(d)}$, i.e. difficulty of the task). Then, we should include dependence on productivity in the cost functions, hence we replace $c_t(a_t, S_t)$ by $c_t(a_t, S_t, p_t)$. Moreover, we should reflect in the model design the possibility for paying partial reward for the goal that is not completely met. Due to technical difficulties the idea has not been developed. However, the similar approach based on the theory of optimal learning is described in Subsection 7.2.

The model described in Section 2 allows the company to find the motivation scheme for one employee or one homogeneous group of employees. However, the proposed approach can be directly extended to the case with several motivation schemes. One needs to consider

one Markov chain for each stratum of employees and include the sum over all strata in (2.2). Clearly, this change leads to higher computational complexity of the model.

7.2 Bayesian Learning framework

To apply the model from Section 2 one has to specify the transition matrix describing the evolution of stress with given control. Therefore, the decision maker firstly performs statistical inference as described in Section 4 and then applies algorithms from 5. However, one can combine the process of estimating the transition probabilities and finding the optimal control using the theory of optimal learning (see [11]). On the other hand, in this case we can find the optimal policy only for one step ahead. Below we sketch this approach.

Recall that the decision maker has a finite set of possible actions $a \in \mathcal{A}$ from which they can choose. Each of these actions, when performed, results in collecting reward W_a by the decision maker. Here W_a is a random variable that comes from a known probability distribution with unknown parameters. The decision maker will try to estimate these parameters as they make subsequent decisions and learn about the system.

Consider a single employee or a homogeneous group of employees. Assume that the rewards resulting from performing action a, namely W_a, are normally distributed with true mean μ_a and standard deviation σ_a^2. For the sake of this argument, we assume that σ_a^2 is a deterministic value that is known to the decision maker. However the decision maker does not know the true mean μ_a and they will try to estimate this value by performing action a and observing the outcome (receiving the reward W_a). Furthermore, we assume that μ_a is itself normally distributed with unknown mean $\bar{\mu}_a$ and variance $\bar{\sigma}_a^2$. We define the precision of our estimate of μ_a to be $\beta_a = \frac{1}{\bar{\sigma}_a^2}$.

At the end of each time step t, the decision maker chooses to perform an action $a \in \mathcal{A}$ and at the beginning of time step $t+1$ they observe an outcome of this action W_a^{t+1}. They use this information to update their belief about action a. Updating equations are specific to the assumed probability distribution and in the case of presented normal-normal model they are (see [11, Section 2.2]):

$$\bar{\mu}_a^{t+1} = \frac{\beta^t \bar{\mu}_a^t + \beta^{W_a} W^{t+1}}{\beta_a^t + \beta^{W_a}}, \quad \beta_a^{t+1} = \beta_a^t + \beta^{W_a}, \tag{7.4}$$

where $\beta^{W_a} = \sigma_a^2$ is deterministic as is a precision of a single observation.

Let us now extend this model to the setting of the problem defined earlier. In addition to the dynamics of the revenue W_a, we take into account that performing the task changes the employee's stress level. We operate in the stress model introduced earlier in Section 2 where stress levels are modelled using a Markov chain. Therefore our observation collected when a worker performs task $a \in \mathcal{A}$ consists of two components: the revenue W_a^{t+1}

and a response vector $S_{t+1,i,a} = \mathbf{y}$. Recall that $\mathbf{y}$ is a vector with all zeroes except for the position (which is 1) where the stress level was on day $t+1$, i.e. the position corresponding to the stress level of the worker to which they transitioned from stress state i because of performing the task $a \in \mathcal{A}$ that was assigned to them at the end of t.

For each $a \in \mathcal{A}$ we will maintain a belief which will represent our State of Knowledge of that worker with respect to performing task $a \in \mathcal{A}$ which consists of:

- $\bar{\mu}_a^t$ - the vector of expected revenues resulting from performing the task at different stress levels. Vector $\bar{\mu}_a^t$ has N entries, where N is the number of stress level states at which the task can be performed. For example, expected revenue from assigning task $a \in \mathcal{A}$ at the stress level i at time step t is stored at the i-th position of vector $\bar{\mu}_a^t$
- β_a^t - the vector of precisions of estimates $\bar{\mu}_a^t$. The situation is analogous to the one introduced in earlier chapter with a difference that we extend a scalar value to a vector of values.
- P_a^t - a two-dimensional matrix of parameters from Dirichlet distribution indicating a probability of transitioning between certain stress levels due to performing task $a \in \mathcal{A}$.

We define an expected profit of performing assigned task at time step t to our worker being at that time at stress level i as:

$$q_{a,i}^t = \bar{\mu}_{a,i}^t, \tag{7.5}$$

where $\bar{\mu}_{a,i}^t$ is expected revenue from assigning a task $a \in \mathcal{A}$ at stress level i as introduced above.

Assume that we are at time t and we stop learning now, that is, we will be still collecting our rewards for every time period until infinity, however we will not update our state of knowledge. Then our optimal choice would be to indefinitely choose an action with the highest expected profit. Therefore we define the value of being in a state described by the state of knowledge $(\bar{\mu}_a^t, \beta_a^t, P_a^t)_{a \in \mathcal{A}}$ as:

$$V((\bar{\mu}_a^t, \beta_a^t, P_a^t)_{a \in \mathcal{A}}) = \max_{a \in \mathcal{A}} q_{a,i}^t. \tag{7.6}$$

For that specified online learning problem, there are plethora of policies (decision-making algorithms) that can be implemented depending on the goal of the business that is doing the implementation. Most of the literature focuses on the concept of balancing exploration and exploitation, that is, balancing immediate rewards with knowledge that can be obtained from testing yet not so well known alternatives.

Pure exploration policy will always choose $a \in \mathcal{A}$ at random with equal probability. On the other hand, pure exploitation will choose $a \in \mathcal{A}$ according to $a^t = \arg\max_{a \in \mathcal{A}} q_{a,i}^t$, were a^t is

the action to be executed at time t. Examples of other policies include: upper confidence bounding, excitation policy, Gittins indices policy or quite recently developed knowledge gradient policy which cleverly balances immediate rewards with rewards that can be obtained in the future due to information collection process. The interested reader may consult [11] as an introduction to the field.

8 Conclusion

In this report, we present our solution of given stress management problem based on combination of stochastic processes theory, statistics and machine learning. Namely,

- the temporal stress changes are modeled in a framework of a discrete controlled Markov process, where an employer can directly influence the transition probabilities of moving from one stress level to another;
- the optimal control (i.e. a sequence of actions performed by the employer) is calculated with respect to some cost function that depicts business goals of the employer using Bellman approach;
- Bayesian learning algorithm is presented for estimation of model parameters.

Note that, in order to use this approach, the employer needs to determine an appropriate cost function that, in general, comprises a compromise between an urge to push the employees for profit maximization and an acceptable level of their stress. Moreover, as high stress levels directly influence productivity of the staff (and therefore achievement of business goals of the company), the cost function should depict the dependance between stress and efficiency.

In order to define a reasonable cost function, as well as to implement the Bayesian learning approach, a decent dataset is required. It should contain, on the one hand, information on how different actions of the employer change the stress levels (to estimate transition probabilities in the Bayesian scheme) and, on the other hand, data that "catches" performance under certain stress conditions (in order to find a cost function).

We also propose several extensions of the given model. The mean-variance, cost-per-time or risk-sensitive modifications of the cost function allow to obtain far more stable results and Bayesian learning framework gives another "greedy"-type way to get an optimal strategy with simultaneous estimation of model parameters.

As it has been shown, we can model policies that can improve process planning, leading to efficient use of resources, and, what is also important, incentivizing personnel by efficient

stress level control. Mainly, the goal in such approach lies in keeping the balance between willingness to work (stability of stress level) and low labour costs.

References

[1] N. Bäuerle, U. Rieder, *Markov Decision Processes with Applications to Finance*, Springer 2011.

[2] T. Bielecki, S. Pliska *Economic Properties of the Risk Sensitive Criterion for Portfolio Management*, Review of Accounting and Finance, 2, 3-17, 2003.

[3] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin, *Bayesian Data Analysis*, Chapman & Hall/CRC 2013.

[4] D. Hebb, *Drives and the C. N. S. (conceptual nervous system)*, Psychological Review, 62, 243-254, July 1955.

[5] O. Hernandez-Lerma, *Adaptive Markov Control Processes*, Springer 1989.

[6] O. Hernandez-Lerma, J. Lasserre, *Discrete-Time Markov Control Processes*, Springer 1996.

[7] A. Korman, *The psychology of motivation*, Englewood Cliffs NJ Prentice-Hall, September 1974.

[8] S. Leka, A. Jain *Health impact of psychosocial hazards at work: an overview*, WHO Press, World Health Organization 2010.

[9] L. Muse , S. Harris, H. Feild, *Has the Inverted-U Theory of Stress and Job Performance Had a Fair Test?*, Human Performance, 16, 349-364, October 2003.

[10] K. Murphy, *Machine learning: a probabilistic perspective*, The MIT Press 2012.

[11] W. Powell, I. Ryzhov, *Optimal Learning*, Wiley 2012

[12] M. Staal, *Stress, Cognition, and Human Performance: A Literature Review and Conceptual Framework*, NASA Technical Reports Server (NTRS), NASA Technical Memorandum, NASA/TM-2004-212824, August 2004.

[13] K. Teigen, *Yerkes-Dodson: A Law for all Seasons*, Theory & Psychology, 4, 525-547, November 1994.

[14] M. Westman, D. Eden, *The inverted-U relationship between stress and performance: A field study*, Work & Stress, 10, 165-173, April 1996.

[15] R. Yerkes, J. Dodson, *The Relation of Strength of Stimulus to Rapidity of Habit Formation*, Journal of Comparative Neurology and Psychology, 18, 459-482, November 1908.

A Transition matrices used in the numerical example

$a^{(r)}=1$	1	2	3	4	5
1	34.5%	26.1%	18.6%	12.6%	8.2%
2	7.6%	34.8%	26.3%	18.7%	12.7%
3	1.2%	8.6%	39.4%	29.7%	21.2%
4	0.2%	1.5%	10.9%	49.8%	37.6%
5	0%	0.3%	2.4%	17.4%	79.9%

Table 1: Exemplary transition probabilities $P_{ij}(a)$ for $a^{(r)}=1$. Due to rounding the numbers may not add up to 100%.

$a^{(r)}=2$	1	2	3	4	5
1	67.3%	24.5%	6.4%	1.5%	0.3%
2	25.2%	50.5%	18.4%	4.8%	1.1%
3	9.2%	23.1%	46.3%	16.9%	4.4%
4	3.3%	9.3%	23.4%	46.8%	17.1%
5	1.4%	4%	11.1%	27.8%	55.7%

Table 2: Exemplary transition probabilities $P_{ij}(a)$ for $a^{(r)}=2$. Due to rounding the numbers may not add up to 100%.

$a^{(r)}=3$	1	2	3	4	5
1	85.7%	13%	1.1%	0.1%	0%
2	44.9%	47.3%	7.2%	0.6%	0.1%
3	29.9%	31.5%	33.2%	5%	0.4%
4	22.1%	23.4%	24.7%	26%	3.9%
5	17.8%	18.9%	20%	21.1%	22.2%

Table 3: Exemplary transition probabilities $P_{ij}(a)$ for $a^{(r)}=3$. Due to rounding the numbers may not add up to 100%.

$a^{(r)} = 4$	1	2	3	4	5
1	94.3%	5.5%	0.2%	0%	0%
2	57.3%	40.3%	2.4%	0.1%	0%
3	40.9%	33.9%	23.8%	1.4%	0%
4	30.9%	28.3%	23.4%	16.5%	1%
5	24.5%	23.5%	21.6%	17.9%	12.6%

Table 4: Exemplary transition probabilities $P_{ij}(a)$ for $a^{(r)} = 4$. Due to rounding the numbers may not add up to 100%.

$a^{(r)} = 5$	1	2	3	4	5
1	97.8%	2.1%	0%	0%	0%
2	63%	36.2%	0.8%	0%	0%
3	41.5%	36.9%	21.2%	0.5%	0%
4	29.7%	29.1%	25.9%	14.9%	0.3%
5	23%	22.9%	22.5%	20%	11.5%

Table 5: Exemplary transition probabilities $P_{ij}(a)$ for $a^{(r)} = 5$. Due to rounding the numbers may not add up to 100%.