IEEE Control Systems Letters paper presented at
2023 American Control Conference (ACC)
San Diego, CA, USA. May 31 - June 2, 2023

Excitation Optimization for Estimating Battery Health Parameters using Reinforcement Learning considering Information Content and Bias

Rui Huang[1], Morgan Jones[2], and Xinfan Lin[1*]

***Abstract*— Accurate parameter estimation has been a long-pursued objective in battery modeling and control practice. To this end, optimization of excitation to improve the estimation accuracy has been an emerging topic, since the quality of data critically determines the accuracy of estimation. However, there are several major drawbacks with existing approaches. First, the commonly used criterion for optimization, e.g., Fisher information, is limited in performance due to not considering the estimation bias caused by inevitable system uncertainties. In addition, alternative existing methods rely on a good $a\,priori$ knowledge of the parameter to be estimated, which is intrinsically contradictory to the goal of estimation. To address these issues, we propose a reinforcement learning (RL) framework to learn the optimal policy for excitation generation that is robust to system uncertainties. In particular, the framework involves a non-additive objective/reward associated with the newly established optimization criterion, and a state augmentation technique is applied to address the ensuing challenge. It is shown that, when applied to estimate a key health-related battery electrochemical parameter, the RL-based approach achieves significantly higher objective value under nominal conditions, and reduces the estimation error by one-order-of-magnitude in the presence of uncertainties compared with the baseline in existing approaches.**

I. INTRODUCTION

The accurate estimation of model parameters is important to guarantee the efficacy of model-based diagnostics and management of lithium-ion batteries [1]. Typically, model parameters are estimated based on measured output data generated by an excitation input sequence. Therefore, the quality of parameter estimation depends strongly on the ability of the excitation input to generate output containing maximal information about the underlying system parameters. This practice can be formulated as an optimization problem where we optimize over the excitation sequence to maximize the information content of the target parameter. The problem is further complicated by inevitable output measurement bias and uncertainties in system model.

To date, research in this field has consisted of early works on maximizing statistical criteria related to estimation, including but not limited to the Fisher information (FI) [2] and the information gain produced by the data, such as the Kullback-Leibler divergence [3], [4] and others. Specifically regarding the estimation in battery applications, FI has been commonly used as the criterion for excitation design. Some early works studied the topic for the equivalent circuit model [5], [6] and the physics-based electrochemical model [7], with the goal of optimizing the features of an imposed input current pattern, e.g., the frequency and magnitude of sinusoidal or pulse. More recently, the research has evolved to discover the fundamentally optimal excitation for battery electrochemical parameters with no imposed current pattern [8], enabled by efficient sensitivity computation [9].

Despite significant progress on the topic, however, there are major drawbacks with the state of the art regarding both the criterion and approach for input optimization. First, regarding the criterion, the commonly used FI objective is subject to inherent limitations, as it is only related to the estimation error variance achievable by an unbiased estimator (through the Cramér-Rao bound) [10], and neglects the estimation bias. In practice, the bias is not only ubiquitous but also prominent due to inevitable system uncertainties, including those in model, measurement, and parameters. Our previous works have shown that input optimized solely based on FI could yield significant estimation error in the presence of system uncertainties [11], [12]. Second, regarding the approach, the existing practice typically formulates input design as optimizing an (open-loop) excitation sequence. It faces a fundamental dilemma: in order to optimize the excitation to estimate the target parameter, a good $a\,priori$ knowledge of the parameter itself is needed, since the optimization objective (e.g., FI) is dependent on its value. In fact, a moderate uncertainty in the $a\,priori$ knowledge of the target parameter may cause the optimized sequence to deviate far from optimality. In addition, the optimization procedure is computationally intensive, due to the large number of optimization variables (dependent on the length of the input sequence) and the need for repetitive computation (and inversion) of system dynamics and Jacobians. As a result, local and sub-optimality is hard to avoid because of the non-linear and non-convex essence of the problem.

In this work, we propose a reinforcement learning (RL) framework as a new approach for input optimization to address the fundamental challenges in the state of the art. We envision the generation of excitation as a Markov Decision Process (MDP) with dynamics described by certain state(s), and the goal is to find an optimal policy that maps the state(s) to an optimal excitation that maximizes the objective. The benefits of using RL are two-fold. First, the obtained policy, which uses states as feedback to generate input when implemented, could significantly reduce the impact of system uncertainties. Compared with the existing practice of optimizing an open-loop excitation sequence, the feedback of states in real-time would correct for the deviation caused by

[1]R. Huang and X. Lin are with the Department of Mechanical and Aerospace Engineering, University of California, Davis, CA 95616, USA, corresponding author e-mail: lxflin@ucdavis.edu

[2]M. Jones is with the Department of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield, UK.

system uncertainties and disturbances. Second, RL could also substantially improve the tractability of optimization. During the training process, the policy is learned using the reward generated at each time instant, and the procedures only involve the forward computation of the rewards and states based on the model. The associated computational complexity is much more favorable than both the direct optimization of the time sequence and the traditional optimal control, which essentially need to not only compute but also invert the system dynamics and the Jacobians of the design objective. Our prior works have demonstrated promising results when optimizing the FI criterion for input optimization [13], [14]. Furthermore, a new criterion is adopted for optimization, which takes model/measurement bias into account. It was discovered in our prior work that, apart from FI, if the data satisfy certain structure in terms of parameter sensitivity, the error caused by system uncertainties can be minimized [12]. Leveraging the results, a new objective for optimization is adopted by augmenting FI with an additional term accounting for constant model/measurement bias. Other types of uncertainties can be addressed in a similar way. It is noted that a major challenge for RL under the new objective is that the latter is non-additive, resulting in difficulty to formulate the single-step reward for learning (as in the case of sole FI). A state augmentation technique is introduced to address the challenge [15]. It will be shown that when applied to estimate a key health-related battery electrochemical parameter, the RL-based approach achieves significantly higher objective values, and reduces the estimation error by near one-order-of-magnitude in the presence of model/measurement bias compared with the baseline existing approach.

II. Input Optimization Objective

In this section, the objective of input optimization is formulated. Specifically, we introduce a previously derived formula that captures the parameter estimation error caused by system uncertainties under the commonly used least-square algorithm [11]. Based on the error formula, a desirable objective function is formulated by supplementing the FI with sensitivity-based data structures that could minimize the estimation error caused by constant model/measurement bias. In addition, the computation of battery parameter sensitivity, which is needed to evaluate the objective during optimization, is also discussed [9].

A. Estimation Error Formula

Consider a discrete-time single-input-single-output system model,

$$\begin{aligned} \boldsymbol{x}_k &= \boldsymbol{f}_k(\boldsymbol{x}_{k-1}, \boldsymbol{\theta}, \boldsymbol{\phi}, u_{k-1}) \\ y_k &= g_k(\boldsymbol{x}_k, \boldsymbol{\theta}, \boldsymbol{\phi}, u_k), \end{aligned} \tag{1}$$

where $\boldsymbol{x}$, u, and y are the states (vector), input, and output of the system respectively, and $\boldsymbol{f}$ and g are the nonlinear state and output equations. The system is parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, with $\boldsymbol{\theta} = [\theta_1, ..., \theta_{n_\theta}]$ representing the target parameters to be estimated, and $\boldsymbol{\phi} = [\phi_1, ..., \phi_{n_\phi}]$ representing other system parameters. In this work, we focus on the estimation of a single-variate system parameter with θ being a scalar ($n_\theta = 1$). The goal of estimation is to determine θ based a sequence of N output data $[y_1^m, y_2^m, ..., y_N^m]$ measured over consecutive time steps excited by a sequence of input $[u_1, u_2, ..., u_N]$. Moreover, by considering various system uncertainties, each output measurement y_k^m is expressed as

$$y_k^m = y_k(\theta^*, \boldsymbol{\phi}^*, \boldsymbol{u}_k) + \Delta y + \delta y_k, \tag{2}$$

where $y_k(\theta^*, \boldsymbol{\phi}^*, \boldsymbol{u}_k)$ is the output of the system based on the true parameters θ^* and $\boldsymbol{\phi}^*$, Δy represents the constant bias between $y_k(\theta^*, \boldsymbol{\phi}^*, \boldsymbol{u}_k)$ and the output measurement including that caused by the measurement noise and/or model uncertainty, and δy_k is the varying uncertainty between the model and measurement. It is noted that $y_k(\theta, \boldsymbol{\phi}, \boldsymbol{u}_k)$ denotes the mapping from θ, $\boldsymbol{\phi}$, and $\boldsymbol{u}_k$ to y_k with the state dynamics contained implicitly and $\boldsymbol{u}_k = [u_1, u_2, ..., u_k]$. A least-squares optimization problem can be formulated to determine the estimate of target variable, $\hat{\theta}$, by minimizing the sum of squared error between the measurement and model output,

$$\min_{\hat{\theta}} J = \frac{1}{2} \sum_{k=1}^{N} \left(y_k^m(\theta^*, \boldsymbol{\phi}^*, \boldsymbol{u}_k) - y_k(\hat{\theta}, \hat{\boldsymbol{\phi}}, \boldsymbol{u}_k) \right)^2. \tag{3}$$

It is noted that the estimation problem further includes the parameter uncertainty in $\boldsymbol{\phi}$, whose exact values are unknown and hence denoted as $\hat{\boldsymbol{\phi}}$ in (3).

In our prior work [11], a formula quantifying the estimation error $\Delta\theta = \theta^* - \hat{\theta}$ induced by the system uncertainties, i.e., constant bias Δy, dynamic uncertainty δy_k, and parameter uncertainty $\Delta\boldsymbol{\phi}$ has been derived as in (4). Several important insights can be made from the formula. On one hand, the denominator is actually the FI of the target parameter θ (under the assumption of i.i.d. Gaussian noises), which has been the predominant objective function adopted in existing studies. Maximizing FI could indeed help decrease the estimation error by increasing the denominator. One the other hand, there are terms in the numerator that indicate the propagation of each uncertainty source to the estimation error. For example, for constant bias Δy, its impact on the estimation error is dependent on the sum of the sensitivity of y to θ over the sequence, $\sum_{k=1}^{N} \frac{\partial y_k}{\partial \theta}(\hat{\theta}, \hat{\boldsymbol{\phi}}, \boldsymbol{u}_k)$. Therefore, to mitigate such error, the input can be designed to achieve minimal $|\sum_{k=1}^{N} \frac{\partial y_k}{\partial \theta}|$. Other types of uncertainties can be minimized similarly based on their associated sensitivity terms. These insights will be used to guide the formulation of the objective for input optimization in the next subsection.

B. New Excitation Optimization Objective Considering Bias

Based on the analysis of the error formula in (4), a new input optimization objective can be established to minimize the impact of system uncertainties on the estimation results. In this work, we focus on countering the constant bias Δy, but other types of uncertainties can be incorporated in a similar way. Specifically, a new objective function is formulated as

$$\max_{\boldsymbol{u}} \sum_{k=1}^{N} \left(\frac{\partial y_k}{\partial \theta}(\hat{\theta}, \hat{\boldsymbol{\phi}}, \boldsymbol{u}_k) \right)^2 - \beta \left| \sum_{k=1}^{N} \frac{\partial y_k}{\partial \theta}(\hat{\theta}, \hat{\boldsymbol{\phi}}, \boldsymbol{u}_k) \right|. \tag{5}$$

$$\Delta\theta \approx -\frac{\left(\sum_{k=1}^{N}\frac{\partial y_k}{\partial\theta}(\hat{\theta},\hat{\boldsymbol{\phi}},\boldsymbol{u}_k)\right)\Delta y+\left(\sum_{k=1}^{N}\frac{\partial y_k}{\partial\theta}(\hat{\theta},\hat{\boldsymbol{\phi}},\boldsymbol{u}_k)\delta y_k\right)+\sum_{i=1}^{n_\phi}\left(\sum_{k=1}^{N}\frac{\partial y_k}{\partial\phi_i}(\hat{\theta},\hat{\boldsymbol{\phi}},\boldsymbol{u}_k)\frac{\partial y_k}{\partial\theta}(\hat{\theta},\hat{\boldsymbol{\phi}},\boldsymbol{u}_k)\right)\Delta\phi_i}{\sum_{k=1}^{N}\left(\frac{\partial y_k}{\partial\theta}(\hat{\theta},\hat{\boldsymbol{\phi}},\boldsymbol{u}_k)\right)^2} \quad (4)$$

The first term is the FI that has already been widely adopted in practice, while the second term is a soft penalty with a scaling factor β to penalize $|\sum_{k=1}^{N}\frac{\partial y_k}{\partial\theta}|$. The new criterion is based on the insight from the error formula (4) that maximizing FI could reduce the overall estimation error, while minimizing $\left|\sum_{k=1}^{N}\frac{\partial y_k}{\partial\theta}\right|$ could mitigate the estimation error caused by the constant modeling/measurement bias Δy. The optimization variables are the sequence of inputs applied over discrete time instances $\boldsymbol{u}=[u_1,...,u_N]^T$.

In the context of battery parameter estimation problem, the output is the voltage, and input is the current, i.e., $y=V$ and $u=I$.

C. Parameter Sensitivity for Battery Electrochemical Parameters

As shown in (5), the objective function involves the calculation of the parameter sensitivity $\frac{\partial y_k}{\partial\theta}(\hat{\theta},\hat{\boldsymbol{\phi}},\boldsymbol{u}_k)$. For the battery electrochemical parameters to be estimated in this research, a method for efficient computation of sensitivity has been developed in our previous work [9], based on the single particle model with electrolyte dynamics (SPMe), which is a reduced order battery electrochemical model widely used for battery control and estimation [1]. A series of techniques and approximations, e.g., Laplace Transform and Padé approximation, have been applied to derive a closed-form expression based on the sensitivity transfer function (STF), which maps system input to sensitivity.

In this work, the target parameter for estimation is the electrode (cathode) active material volume fraction, $\theta=\varepsilon_s$, which is a critical battery parameter related to capacity and reflecting the battery degradation progress. It has been frequently used as the subject for estimation to monitor the battery state of health [16]. The sensitivity of the output, i.e., battery voltage V, to ε_s takes the form

$$\frac{\partial V}{\partial\varepsilon_s}(t)=\frac{\partial\eta}{\partial\varepsilon_s}(t)+\left(\frac{\partial\eta}{\partial c_{se}}+\frac{\partial U}{\partial c_{se}}\right)\cdot\frac{\partial c_{se}}{\partial\varepsilon_s}(t). \quad (6)$$

The first term accounts for the impact of ε_s on the overpotential η, which drives the (de)intercalation reaction and the flow of lithium ions between the electrode and electrolyte. This term is non-dynamic and can be computed easily based on the model. The second term captures the impact of ε_s on the open circuit potential (OCP) U and overpotential η through lithium diffusion. Specifically, OCP is the battery electrode equilibrium potential dependent on the lithium concentration at the electrode particle surface c_{se}, which will change during battery operation as lithium ions diffuse in the electrode. For this term, $\frac{\partial U}{\partial c_{se}}$ and $\frac{\partial\eta}{\partial c_{se}}$ are the slopes of the U and η with respective to c_{se}, which can both be computed easily. Meanwhile, $\frac{\partial c_{se}}{\partial\varepsilon_s}$ is the sensitivity of the diffusion state, i.e., the particle surface concentration c_{se}, to ε_s, for which an STF has been derived as

$$\frac{\partial c_{se}}{\partial\varepsilon_s}(s)=\frac{7R_s^4s^2+420D_sR_s^2s+3465D_s^2}{s(R_s^4s^2+189D_sR_s^2s+3465D_s^2)}\cdot\frac{I(s)}{F\varepsilon_s^2A\delta}, \quad (7)$$

where R_s, A, δ, and F represent the particle radius, electrode area, electrode thickness, and Faraday constant respectively.

We can implement the derived STF in time domain, e.g., through state-space representation, to efficiently compute the state sensitivity $\frac{\partial c_{se}}{\partial\varepsilon_s}(t)$ and then obtain the output sensitivity $\frac{\partial V}{\partial\varepsilon_s}(t)$ based on (6), which can be used for excitation optimization, as demonstrated in [12]. More details about the battery model and sensitivity computation can be found in [9].

III. Reinforcement Learning Framework

A new input optimization objective has been formulated in (5), which aims to maximize the FI while minimizing the estimation error induced by constant bias. This objective has been applied to optimize the open-loop input sequence in [12]. A major issue with that existing approach is that the objective is often dependent on the target parameter that needs to be estimated. As an example, for $\theta=\varepsilon_s$, as seen from (7), its sensitivity includes ε_s itself. Therefore, in order to attain the optimum of the true objective, a good $a\,priori$ knowledge of θ is needed, which is typically impossible because θ itself is the unknown target of estimation. Deviation in the $a\,priori$ knowledge could significantly affect the optimality of the obtained sequence. In addition, the existing approach is computationally intensive and prone to local/sub-optimality due to the large number of optimization variables and nonlinearity and complexity of the battery model. Specifically, regarding the battery current optimization problem, for a 30-min input sequence with a sampling rate of 1 second, the total number of input values to be optimized is 1800.

RL provides a fundamentally different way to address the challenges facing the existing approach. In RL, the generation of optimal excitation is viewed as a Markov Decision Process (MDP) with dynamics described by certain state (s) S, which are driven by underlying dynamics related to battery physics. The goal is to find an optimal control policy $\pi(S)$ mapping the state S to action a (which is the current I in the context of battery estimation problem) that can maximize the objective of input optimization. The schematic of the RL-based excitation optimization framework is shown in Fig. 1. During the training process, the policy is learned through a reward r_k generated at each step related to the objective, by leveraging the RL principle of balancing between exploration and exploitation to solve the Bellman equation, the form of which most optimal control problems can be formulated. During implementation, the

optimal excitation is generated by the trained policy in a closed-loop manner based on the feedback of state or its estimate by an observer.

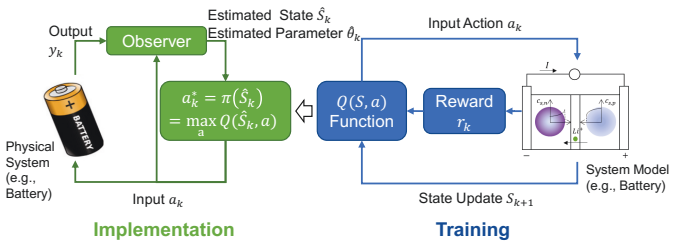

Fig. 1. Schematic of RL for Input Optimization and Generation

One key step in formulating the RL framework is reward design. For the commonly used objective of maximizing the FI, i.e., $\sum_{k=1}^{N}\left(\frac{\partial y_k}{\partial \theta}\right)^2$, which is additive over the sequence, the reward (or equivalently the one-step cost) can be conveniently set as $r_k = \left(\frac{\partial y_k}{\partial \theta}\right)^2$ at each time step. The new objective in (5), however, is non-additive, as it further includes a term $\beta\left|\sum_{k=1}^{N}\frac{\partial y_k}{\partial \theta}\right|$, which needs to be evaluated over the whole sequence and cannot be distributed to each step [17]. In order to address this challenge, a state augmentation technique is applied in this study [15]. Specifically, a state w is defined as

$$w_k = \sum_{i=1}^{k}\frac{\partial y_i}{\partial \theta}(\hat{\theta}, \hat{\boldsymbol{\phi}}, \boldsymbol{u}_i). \tag{8}$$

In this way, the reward r_k can be designed as

$$r_k = \begin{cases} (\frac{\partial y_k}{\partial \theta}(\hat{\theta}, \hat{\boldsymbol{\phi}}, \boldsymbol{u}_k))^2, & k < N \\ (\frac{\partial y_k}{\partial \theta}(\hat{\theta}, \hat{\boldsymbol{\phi}}, \boldsymbol{u}_k))^2 - |w(k)|, & k = N \end{cases} \tag{9}$$

to account for the non-additive term in the objective function. It has been proven that RL performed with the augmented states and designed reward satisfies the principle of optimality of the original objective function [15].

Regarding the rest of the states, we choose one to be a battery physical state, i.e., state of charge (SOC), which is related to c_{se} and reflects the underlying battery dynamics. In addition, the time index k is also adopted as a state, because the input optimization is a finite horizon RL problem and the optimal action depends on the time progression over the sequence. The complete set of states S is hence

$$S(k) = \begin{bmatrix}\text{SOC}(k) & w_k & k\end{bmatrix}^T. \tag{10}$$

In this research, the policy is trained by the classical Q-learning method for its simplicity, while other RL algorithms can be adopted alternatively under the framework. Specifically, a Q function $Q(S,a)$ is used to represent the expected total reward to be received after taking an action a when at state S (until the end of the sequence). This Q function needs to be learned during the training phase. Specifically, learning is performed in repeated episodes, which consist of a series of time steps. At each time step k, an action a_k is taken with the state at S_k. The state transition S_{k+1} and the associated reward r_k are observed, and the Q function $Q(S_k, a_k)$ is updated by iteratively applying the following update rule until convergence [18],

$$\begin{aligned} Q(S_k, a_k) \leftarrow & Q(S_k, a_k) + \alpha[r_k + \\ & \gamma \max_a Q(S_{k+1}, a) - Q(S_k, a_k)], \end{aligned} \tag{11}$$

$$a_k \leftarrow \begin{cases} \text{argmax}_a\, Q(S_k, a) & \text{with probability } 1-\epsilon \\ \text{a random action} & \text{with probability } \epsilon \end{cases} \tag{12}$$

in which α is the learning rate and γ is the discounting factor. The action a_k is chosen using the ϵ-greedy strategy described in (12). With probability of $1-\epsilon$, the strategy exploits the maximum return by selecting the action that maximizes the Q function, while with probability of ϵ, it explores a random action for the purpose of learning. Following training, the obtained Q function is used to generate the excitation in the implementation stage, according to the greedy policy (with ϵ = 0), with the aim of maximizing the return based on the feedback from the states.

The proposed framework takes a model-based RL approach, where the states and reward are computed based on the SPMe during training. The framework addresses the challenges faced by the existing approach of direct sequence optimization. On one hand, the closed-loop policy is more robust to uncertainties since it modifies the excitation based on the feedback of actual states during implementation, rather than sticking with a fixed pre-determined excitation sequence. On the other hand, the training process of RL, which only involves forward propagation of states and reward based on the model, is more tractable than direct sequence optimization, which requires repetitive computation and inversion of the system dynamics and Jacobians.

IV. SIMULATION VALIDATION

In this section, the optimization results for estimating $\theta = \varepsilon_s$ considering the constant model/measurement bias are presented. The performance of the proposed RL framework will be analyzed and compared with the existing approach, both with and without parameter uncertainties.

The optimization task aims to generate an 1800-second current excitation sequence that maximizes the new criterion in (5) with $\beta = 1$ and a sampling interval of 1 second, starting from 0.5 SOC. During RL training, each episode starts from SOC $= 0.5$ and ends when time reaches 1800 seconds. The states and input are digitized to discrete values to index the Q function, i.e., the input, SOC and w are discretized by 11, 101 and 200 respectively. The time horizon is discretized by 18. It is noted that having more discretization points for the time horizon is favorable for control accuracy, but at the cost of proportionately larger discretization nodes for other states to discern the effect of different inputs on state transition, thus significantly complicating the problem. The hyper-parameters of the RL are calibrated as $\alpha = 0.5$ and $\gamma = 1$, while the ϵ decays lineally from 1 to 0 over 10 million episodes.

As a benchmark, direct sequence optimization, i.e., solving (5) using the conventional optimization approach, is implemented using the state-of-the-art General Purpose OPtimal Control Software based on the same model under the same conditions. The SPMe parameters are adopted from [19].

A. Nominal Optimization Results

We will first evaluate the optimization results of the RL and the conventional approach under nominal conditions, i.e., assuming we have perfect knowledge of ε_s during optimization. Specifically, optimization is performed with three target parameter values, i.e., $\varepsilon_s = 0.5, 0.45, 0.4$, and the results are summarized in Table I, including the objective value of the obtained policy/sequence, the FI, and $|\sum_{k=1}^{N} \frac{\partial y_k}{\partial \theta}|$ which represents the impact of constant bias. The obtained current excitation, resulting voltage, and SOC for the two approaches under $\varepsilon_s = 0.5$ are shown in Fig. 2 as an example.

It is seen that the RL policy achieves significantly higher objective value than the conventional direct optimization approach as shown in Table I for all values of ε_s, i.e., 41.0 versus 24.1 (70% higher) for $\varepsilon_s = 0.5$, 64% higher for $\varepsilon_s = 0.45$, and 43% higher for $\varepsilon_s = 0.4$. More specifically, the RL policy yields both consistently higher FI and lower $|\sum_{k=1}^{N} \frac{\partial y_k}{\partial \theta}|$, which are the exact goals of optimizing the new objective. These results indicate the potential advantage of the RL-based approach in finding the true goal optimum by reducing the computational complexity, which presents a major challenge for the conventional approach.

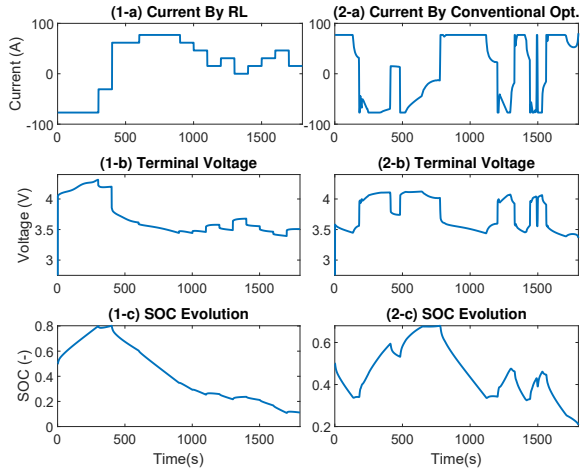

Fig. 2. Optimized current, resultant voltage and SOC for ε_s by RL (1-a,b,c) and conventional approach (2-a,b,c)

B. Simulation Validation subject to Parameter Uncertainty

In this part, the performance of the optimized current excitations is validated by applying to ε_s estimation. The validation is conducted in simulation, where the voltage data are generated by SPMe, and the value of ε_s can be controlled to benchmark the estimation results. Furthermore, to emulate the practical scenario where ε_s is unknown for input design, the uncertainty in ε_s is considered during validation. Namely, the policy/sequences optimized with $a\,priori$ knowledge of ε_s at 0.4 and 0.45 are applied to an actual battery (model) with $\varepsilon_s = 0.5$. The generation of excitation by the RL policy is informed by the feedback of states, while the conventionally optimized sequences are applied directly.

After the voltage data are generated, a constant bias of $0.05\,\mathrm{V}$ is injected into the data to emulate the model/measurement bias in practice. The estimation of ε_s is performed using the least-squares method based on the data generated by both optimization approaches, as well as several other heuristic excitations frequently used for battery parameter identification, e.g., 1C constant current discharging, 1C Pulse and the Federal Urban Driving Schedule (FUDS) [11], [20]. The results are listed in Table II for comparison.

It is seen that generating an optimal excitation subject to uncertainty in the target parameter is indeed a major challenge for conventional sequence optimization. As shown in II, when optimized at $\varepsilon_s = 0.45$, which amounts to a 10% deviation from the true value of 0.5, the obtained open-loop sequence gives an estimation error of -6.46%. When the uncertainty in the $a\,priori$ knowledge is further increased to 20% (with the sequence optimized at $\varepsilon_s = 0.4$), the estimation error is as high as -20.42%, which basically means that the estimation using the open-loop sequence did not improve the knowledge of ε_s at all. The RL policy, on the other hand, manages to achieve highly accurate estimation results with one-order-of-magnitude smaller errors, namely 0.54% when trained at $\varepsilon_s = 0.45$ and 2.96% at $\varepsilon_s = 0.4$. The substantial difference in performance can be explained by the values of the optimization objective in (5) when evaluated at the true $\varepsilon_s = 0.5$, also shown in Table II. It can be seen that the objective value of the open-loop sequence drops significantly from 29.5 (with perfect knowledge of ε_s shown in Table. I) to 7.9 under 10% uncertainty, and from 34.1 to -15.8 under 20% uncertainty. This indicates that a deviation in the $a\,priori$ knowledge of the target parameter value could significantly affect the optimality of the generated open-loop sequence. Meanwhile, the RL policy is capable of maintaining substantially higher objective values under the same uncertainties, namely 32.1 under 10% uncertainty and 16.5 under 20%, thanks to the feedback of states during implementation which renders robustness to input generation. It is noted that even though the estimated states are also subject to errors caused by uncertainties, such error is typically minimal due to use of output (voltage) feedback for correction. In addition, the RL policy also yields much better estimation accuracy than the heuristic profiles, which have errors between -6.4% and -93.8%, indicating the benefits of using a designed input excitation for estimation.

V. CONCLUSIONS

In this paper, we propose a RL framework to optimize the current excitation for estimating battery parameters subject to constant model/measurement bias. The framework aims at maximizing a new objective for optimization, which considers both the FI of the data and the impact of system uncertainty. The optimization is achieved by leveraging the

TABLE I

OPTIMIZATION RESULTS OF TWO APPROACHES UNDER DIFFERENT ε_s

Approach	$\varepsilon_s = 0.5$			$\varepsilon_s = 0.45$			$\varepsilon_s = 0.4$		
	FI	$\lvert\sum_{k=1}^{N} \frac{\partial y_k}{\partial \theta}\rvert$	Obj. Value	FI	$\lvert\sum_{k=1}^{N} \frac{\partial y_k}{\partial \theta}\rvert$	Obj. Value	FI	$\lvert\sum_{k=1}^{N} \frac{\partial y_k}{\partial \theta}\rvert$	Obj. Value
Conventional	24.9	0.8	24.1	29.6	0.1	29.5	34.3	0.2	34.1
RL	41.4	0.4	41.0	48.4	0.1	48.3	48.7	0.1	48.6

TABLE II

ESTIMATION RESULTS OF VARIOUS EXCITATIONS WITH TRUE $\varepsilon_s = 0.5$

Excitation	Estimation Error	Obj. Value
Open-loop sequence (optimized at $\varepsilon_s = 0.4$)	-20.42%	-15.8
RL policy (trained at $\varepsilon_s = 0.4$)	2.96%	16.5
Open-loop sequence (optimized at $\varepsilon_s = 0.45$)	-6.46%	7.9
RL policy (trained at $\varepsilon_s = 0.45$)	0.54%	32.1
1C CC	-66.12%	-159.9
1C Pulse	-6.40%	-1.5
FUDS	-93.80%	-93.3

RL principle to train an optimal policy, which could guide the generation of the excitation based on feedback of the state estimates. To address the non-additive property of the new objective, state augmentation and reward re-structuring techniques are applied to enable RL. The framework is then applied to input optimization for estimating a key battery health-related electrochemical parameter ε_s under constant bias. The results have been validated by simulation, demonstrating the advantages of the RL-based new approach over the direct optimization of input sequence using a conventional approach. First, the optimization results by RL achieves more than 43% higher objective value under the nominal condition (i.e., with perfect knowledge of ε_s), indicating the potential of the new approach in finding the true optimum by reducing the complexity of optimization. More importantly, when applied to the practical estimation scenario with uncertainty in target parameter, the closed-loop policy renders the RL-based approach significantly improved robustness, with one-order-of-magnitude reduction in estimation error. In ongoing works, this framework is being extended to incorporate more types of system uncertainties and applied to parameter estimation for more complicated models (e.g., the full-order Doyle-Fuller-Newman model).

ACKNOWLEDGMENT

We appreciate the funding support from the NSF CAREER Program (Grant No.2046292) and the NASA HOME Space Technology Research Institute (Grant No.80NSSC19K1052).

REFERENCES

[1] X. Lin, Y. Kim, S. Mohan, J. B. Siegel, and A. G. Stefanopoulou, "Modeling and estimation for advanced battery management," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 393–426, 2019.

[2] E. L. Lehmann and G. Casella, *Theory of point estimation*. New York, NY, USA: Springer Science & Business Media, 2006.

[3] D. V. Lindley, "On a measure of the information provided by an experiment," *The Annals of Mathematical Statistics*, vol. 27, no. 4, pp. 986–1005, 1956.

[4] X. Huan and Y. M. Marzouk, "Sequential bayesian optimal experimental design via approximate dynamic programming," *arXiv preprint arXiv:1604.08320*, 2016.

[5] M. J. Rothenberger, D. J. Docimo, M. Ghanaatpishe, and H. K. Fathy, "Genetic optimization and experimental validation of a test cycle that maximizes parameter identifiability for a li-ion equivalent-circuit battery model," *Journal of Energy Storage*, vol. 4, pp. 156–166, 2015.

[6] Z. Song, H. Hofmann, X. Lin, X. Han, and J. Hou, "Parameter identification of lithium-ion battery pack for different applications based on cramer-rao bound analysis and experimental study," *Applied energy*, vol. 231, pp. 1307–1318, 2018.

[7] S. Park, D. Kato, Z. Gima, R. Klein, and S. Moura, "Optimal experimental design for parameterization of an electrochemical lithium-ion battery model," *Journal of The Electrochemical Society*, vol. 165, no. 7, p. A1309, 2018.

[8] Q. Lai, H. J. Ahn, G. Kim, W. T. Joe, and X. Lin, "Optimization of current excitation for identification of battery electrochemical parameters based on analytic sensitivity expression," in *2020 American Control Conference (ACC)*. IEEE, 2020, pp. 346–351.

[9] Q. Lai, S. Jangra, H. J. Ahn, G. Kim, W. T. Joe, and X. Lin, "Analytical derivation and analysis of parameter sensitivity for battery electrochemical dynamics," *Journal of Power Sources*, vol. 472, p. 228338, 2020.

[10] H. Cramér, *Mathematical Methods of Statistics (PMS-9), Volume 9*. Princeton, NJ, USA: Princeton university press, 2016.

[11] Q. Lai, H. J. Ahn, Y. Kim, Y. N. Kim, and X. Lin, "New data optimization framework for parameter estimation under uncertainties with application to lithium-ion battery," *Applied Energy*, vol. 295, p. 117034, 2021.

[12] Q. Lai, W. T. Joe, G. Kim, and X. Lin, "Data optimization for parameter estimation under system uncertainties with application to li-ion battery," in *2021 American Control Conference (ACC)*. IEEE, 2021, pp. 4408–4413.

[13] R. Huang, J. Fogelquist, and X. Lin, "Input excitation optimization for estimating battery electrochemical parameters using reinforcement learning," in *2022 IEEE Vehicle Power and Propulsion Conference (VPPC)*. IEEE, 2022, pp. 1–6.

[14] ——, "Reinforcement learning of optimal input excitation for parameter estimation with application to li-ion battery," *IEEE Transactions on Industrial Informatics*, 2023.

[15] M. Jones and M. M. Peet, "Extensions of the dynamic programming framework: Battery scheduling, demand charges, and renewable integration," *IEEE Transactions on Automatic Control*, vol. 66, no. 4, pp. 1602–1617, 2020.

[16] A. P. Schmidt, M. Bitzer, Á. W. Imre, and L. Guzzella, "Model-based distinction and quantification of capacity loss and rate capability fade in li-ion batteries," *Journal of Power Sources*, vol. 195, no. 22, pp. 7634–7638, 2010.

[17] R. Wang, P. Zhong, S. S. Du, R. R. Salakhutdinov, and L. Yang, "Planning with general objective functions: Going beyond total rewards," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 486–14 497, 2020.

[18] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[19] S. Moura, "scott-moura/spmet: The full spmet," Dec. 2016. [Online]. Available: https://doi.org/10.5281/zenodo.221376

[20] L. Xu, X. Lin, Y. Xie, and X. Hu, "Enabling high-fidelity electrochemical p2d modeling of lithium-ion batteries via fast and non-destructive parameter identification," *Energy Storage Materials*, vol. 45, pp. 952–968, 2022.