A data-driven approach for capacity estimation of batteries based on voltage dependent health indicators

Jinzen Kong, Jie Liu, Yikai Chen and Dong Wang

The State Key Laboratory of Mechanical Systems and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

1 E-mail: dongwang4-c@sjtu.edu.cn

Abstract. Battery capacity estimation plays an important role in the normal operation of electric vehicles. In this work, we presented a data-driven approach for capacity estimation of batteries based on voltage dependent health indicators. A difference-based model of discharge voltage and capacity was built. Next, two health indicators are constructed from partial voltage curves, and correlations between capacity and health indicators are investigated. Afterward, the capacity estimation approach based on Gaussian process regression model is expounded. To validate the accuracy of the proposed method, a case study is carried out. Results demonstrate that RMSE and RMSPE of capacity estimation are lower than 1% compared with actual capacity.

1. Introduction

Electric vehicles (EVs) are expected to be a major force recently due to the increasingly serious issues of environmental pollution [1]. Batteries are critical components of EVs and the health states of batteries play a major role in EVs working normally. There have been many studies focusing on batteries, such as battery structure improvement [2], charged state prediction [3], battery thermal management [4-6], battery prognostics and health management [7, 8], etc. The accurate estimation of battery capacity has a remarkable significance in reducing accident risks and promoting the development of second-hand EVs markets. Many studies in recent years have focused on capacity estimation of batteries.

Physical-based approaches are conventional in battery capacity estimation [9]. This kind of method mainly relies on the establishment of an accurate physical model of the battery, including the electrochemical model estimation method and equivalent circuit model estimation method. Simplified electrochemical models such as Pseudo-two-Dimensional models [10] and single particle model [11] have been studied by some researchers. Data-driven based approaches for battery capacity estimation have gained growing attention these years [12]. They can be divided into statistical methods and Machine learning (ML) methods. Statistical methods use historical battery data to model battery aging and estimate parameters by filtering methods. Wang et al. [13] published a piecewise model combining with the expectation-maximization algorithm to estimate the capacity of batteries with capacity regeneration phenomena. Qiu et al. [14] improved the cuckoo search particle filter and presented a SOH estimation approach. ML methods such as neural networks (NN) [15], support vector machine (SVM) [16], random forest regression (RFR) [17] and extreme learning machine (ELM) [18], etc. have been applied to battery capacity estimation problem. For example, Zhang et al. [19] proposed a Gaussian process machine learning method combining electrochemical impedance spectroscopy (EIS) to predict the battery health states. Dai et al. [20] introduced a state of health (SOH) estimation approach utilizing NN and Markov chain.
A major current concern in battery capacity estimation field is how to construct health indicators by observable quantities such as voltages and currents. Yang et al. [21] extracted health indicators from discharge curves under the various depth of discharge. Liu et al. [18] constructed an energy-based health indicator from discharge voltage. However, current health indicators construction methods almost require fully charged/discharged data, which is limited in practice. The aims of the present work are building health indicators that are sensitive to capacity degradation trends, and conduct capacity estimation for batteries. In this paper, two voltage health indicators are constructed by the proposed difference model and Gaussian process regression (GPR) model is utilized to estimate capacity of batteries.

The structure of this work is summarized as follows. In Section 2, a data-driven approach for capacity estimation of batteries based on voltage curves is introduced. Then, a case study is carried out to validate the proposed approach and results are summarized in Section 3. Finally, Section 4 draws conclusions.

2. A data-driven capacity estimation approach of batteries based on discharge voltage curves

2.1. Data description

The battery dataset we used in this work was conducted by MIT, which contains voltages, currents, capacity, and other measurements during charging/discharging process. The batteries in the experiment were from A123 System, the type of them is APR18650M1A. Each cell had a rated capacity of 1.1 Ah and a rated voltage of 3.6V. The technical specification of batteries is listed in Table 1. The rule of charging process was defined as a multi-stage constant current and constant voltage. Firstly, the charging policy was “I(1S1)-I2”, which meant batteries were charged under I1 C (C-rate, which is defined as the measurement of the charge and discharge current concerning its nominal capacity) until SOC reached S1 and then charged under I2 C to 80% SOC, finally filled up batteries under 1 C. About discharging, all batteries were discharged under 4C. As an example, B101 was a battery following 5.3 C (54%)-4 C charging/discharging rule. Its voltage-discharge capacity curves can be shown in Figure 1, and the color bar represents charging/discharging cycles of B101. And the experiments were implemented at 30 °C in an environmental chamber, surface temperatures of batteries were measured by a Type T thermocouple.

| Cathode | Anode   | Nominal capacity | Nominal voltage | Discharge cut-off voltage |
|---------|---------|------------------|-----------------|--------------------------|
| LiFePO4 | Graphite| 1.1 Ah           | 3.3 V           | 2 V                      |

In this work, we extract features based on partial discharge voltage curves. In practice, sometimes the full curve of discharge voltage cannot be obtained so that some conventional methods which are required for fully discharge data such as ICA/DVA cannot be used. As we have seen in Figure 1, capacity can be considered as a function of voltage, and in different charging/discharging cycles, voltage curves drift in the same direction, where discharge capacity of B101 decreases accordingly. Besides, the drop rate of discharge voltage changes at different capacity internal and different charging/discharging cycles.

2.2. A difference-based model proposed for health indicators construction

For capturing the offset tendency character of discharge voltage curves during different charging/discharging cycles and further construct health indicators, a difference-based model is introduced in this section.

We investigate the difference capacity respect to voltage, which can be expressed as:

$$\Delta Q(V_i) = Q_{i+1}(V_{i+1}) - Q_i(V_i)$$

where $i$ is the index of voltage data, the voltage interval is set as 0.0015 V.
Figure 1. Voltage-discharge capacity curves of B101.

Figure 2. Fitting results of the difference-based model.

To linearize the difference capacity, the following difference-based model for capacity is put forward, illuminated by difference equation:

$$\Delta Q(V) = \omega(c)(-Q(V))Q(V) + b(c) + \epsilon(c)$$  \hspace{1cm} (2)

where $c$ is charging/discharging cycle; $\omega(c)$ and $b(c)$ are parameters of the model; $\epsilon(c)$ is a Gaussian error following $N(0, \sigma^2)$.

Using the model in Equation (2), the capacity respect to voltage can be turned into a linear relationship approximate. It is noted that only the central part (from 2.7 to 3.3 V roughly) of discharge voltage curves are used in this model, where has great differences among different cycles. We can get the original results, and fitting results by cftool in Matlab, illustrated in Figure 2. The R-square is 0.9974.

2.3. Voltage dependent health indicators construction

The slope characteristic $\omega(c)$ and intercept characteristic $b(c)$ are derived by the above proposed difference-based model. After smoothing by moving averaging algorithm, they can be drawn as Figure 3. As we can see in Figure 3, $\omega(c)$ and $b(c)$ are all monotonic basically and $\omega(c)$ increases with charge-discharge cycle while $b(c)$ has an opposite trend. In addition, we also draw the discharge capacity in Figure 3 in order to better compare the relationship between capacity and two characteristics.

After analysis, we attend to define the two characteristics related to voltage as health indicators and detailed demonstrations are elaborated as follows.

An analysis between discharge capacity and $\omega(c)$, $b(c)$ are given according to Pearson correlation coefficient (PCC) [22], which can be calculated by Equation (3):

$$PCC = \frac{\sum_{i=1}^{N}(HI_i - \overline{HI})(Q_i - \overline{Q})}{\sqrt{\sum_{i=1}^{N}(HI_i - \overline{HI})^2} \sqrt{\sum_{i=1}^{N}(Q_i - \overline{Q})^2}}$$  \hspace{1cm} (3)

where $HI$ is health indicator we derived before, $Q$ represents discharge capacity.

Generally, if the absolute value of PCC between two quantities closes to 1, then the two quantities have a high correlation. According to PCC, PCCs of two health indicators are 0.9947 and $-0.9360$, respectively. The correlations between the two proposed health indicators and discharge capacity are high. Thus, the two health indicators are reliable to reflect the health state of battery.
2.4. Degradation modeling and capacity estimation of batteries based on GPR

GPR [23] methods are generally utilized as a nonlinear mapping way which can achieve the goal of modeling or exploring functions, solving exploration-exploitation problems, and so on. It is a Bayesian-related method combining statistics knowledge and machine learning. In this work, we consider using GPR based model to estimate capacity of batteries.

We assume the underlying model follows the form \( y = g(x) + \epsilon \), where \( x \) represents health indicators; error term \( \epsilon \) follows \( \epsilon_i \sim N(0, \sigma_i^2) \). \( g(HI) \) is defined as a Gaussian process (GP), which follows \( g = (g_1, \ldots, g_n)^T \sim GP(m(x), k(x, x')) \). \( m(x) \) and \( k(x, x') \) are mean function and covariance function as shown in Equation (4):

\[
\begin{align*}
    m(x) &= E(g(x)) \\
    k(x, x') &= E[(m(x) - g(x))(m(x') - g(x'))]
\end{align*}
\] (4)

where \( E(\cdot) \) represents the expectation of functions.

In this work, we set \( m(x) = 0 \) and select Matérn covariance function as \( k(x, x') \) in Equation (5), which can be indicated as

\[
K(x, x') = \sigma_n^2 \frac{1}{\Gamma(v)\nu^2} \left[ \frac{\sqrt{2\nu}}{\rho} |x - x'| \right]^{\nu} K_v \left( \frac{\sqrt{2\nu}}{\rho} |x - x'| \right)
\] (5)

where \( \Gamma \) denotes a gamma function; \( K_v \) indicates the improved Bessel function.

Using GPR, the joint prior distribution of measured capacity \( y \) and estimated capacity \( y' \) is derived:

\[
\begin{pmatrix} y \\ y' \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ \mathbf{K}(x, x) + \sigma_n^2 \mathbf{I}_n \end{pmatrix}, \begin{pmatrix} \mathbf{K}(x', x) & \mathbf{K}(x', x') \end{pmatrix} \right)
\] (6)

Using the modified maximum likelihood estimation algorithm, the posterior conditional distribution of estimated capacity \( y' \) and new given dataset can be calculated:

\[
p(y'|x', x, y) \sim N(y'|\bar{y}', \text{cov}(y'))
\] (7)

where \( \bar{y}' \) is the mean of estimation results, \( \text{cov}(y') \) is variance matrix, and they can be calculated by:

\[
\begin{align*}
    \bar{y}' &= \mathbf{K}(x, x')^T \left[ \mathbf{K}(x, x) + \sigma_n^2 \mathbf{I}_n \right]^{-1} y \\
    \text{cov}(y') &= \mathbf{K}(x', x') - \mathbf{K}(x', x)^T \left[ \mathbf{K}(x, x) + \sigma_n^2 \mathbf{I}_n \right]^{-1} \mathbf{K}(x, x')
\end{align*}
\] (8)

Based on GPR, we conduct capacity estimation of batteries. Firstly, model training is implemented by inputting health indicators and capacity of training dataset, which are derived from the proposed difference-based model. Then, the GPR degradation model is built and hypermeters of GPR model are...
optimized with the help of training dataset. The inputs of GPR are two voltage dependent health indicators and outputs are battery capacity. Subsequently, we can estimate capacity of batteries and uncertainty of estimation while new data comes.

3. Case study and results
We use a battery dataset from MIT [24] to verify the effectiveness of the proposed approach. We select B93, B113, B116 as training datasets and B108 as testing datasets, their charging/discharging rules are all following 5.3 C (54%)-4 C. The data of the initial 300 cycles is also considered as historical operation data. Detailed results of capacity estimation are listed as follows.

3.1. Error analysis
In this work, Root mean square error (RMSE) and Root mean square percentage error (RMSPE) metrics are chosen to measure the estimated error.

RMSE can be calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}.$$  \hspace{1cm} (9)

RMSPE can be calculated as:

$$\text{RMSPE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}.$$  \hspace{1cm} (10)

where $n$ denotes the estimated samples; $y_i$ and $\hat{y}_i$ represent the actual capacity and estimated capacity.

3.2. Capacity estimation results of batteries
Capacity estimation of B108 is shown in Figure 4 and Figure 5. As we can see, the estimated efficiency is well and estimated uncertainty is low. Quantificationally, the max error estimation is - 0.0268; RMSE of estimation is 0.82% and RMSPE of estimation is 0.85%, which are accuracy.

4. Conclusions
Battery health management is a popular topic in the EVs business. However, the capacity estimation accuracy of existing methods should be improved. In this study, we constructed voltage dependent health indicators and proposed a data-driven approach for capacity estimation of batteries. A difference-based model was presented to linearize the capacity v.s. voltage curves and two health indicators were extracted from partial voltage curves, then correlations between capacity and health indicators were analyzed. Besides, capacity estimation approach of batteries was introduced. To verify the accuracy of the proposed method, a case study was conducted. Results showed that RMSE and RMSPE of capacity estimation are 0.82% and 0.85%, which are accuracy.
estimation were both lower than 1%, which were more accurate than existing methods. In future work, more efforts will be put to explore the battery capacity estimation approaches under different working conditions.

Acknowledgments
This work was supported by National Natural Science Foundation of China under Grant No. 51975355 and Ministry of Education-China Mobile Research Foundation under Grant No. CMHQ-JS-201900003.

Appendix

### Nomenclature

| Symbol | Meaning |
|--------|---------|
| $Q$ | Discharge capacity in the $i^{th}$ data |
| $\Delta Q(V)$ | Difference capacity |
| $c$ | Cycle number of a battery |
| $\sigma^2$ | Variance of normal errors |
| $m_{HF}$, $m_{HF}$ | Mean function, mean matrix |
| $k_{HF, HF}$, $K_{HF, HF}$ | Covariance function, covariance matrix |
| $\mathbf{y}', \mathbf{y}$ | Predicted/actual health states |
| $\bar{\mathbf{y}}$ | Mean of estimation results |
| $\Theta$ | Hyperparameters of the GPR model |

### Abbreviations

| Abbreviation | Meaning |
|--------------|---------|
| GPR | Gaussian process regression |
| ICA | Incremental capacity analysis |
| SoH | State of health |
| DVA | Differential voltage analysis |
| SoC | State of charge |
| PHM | Prognostics and health management |
| HIs | Health indicators |
| PCC | Pearson correlation coefficient |
| EVs | Electric vehicles |
| RMSE | Root mean square error |
| P2D | Pseudo Two-Dimensional |
| RMSPE | Root mean square percentage error |
| ML | Machine learning |
| SVM | Support vector machine |
| RFR | Random forest regression |
| ELM | Extreme learning machine |

### References

[1] Wang Y J, Tian J Q, Sun Z D, Wang L, Xu R L, Li M C and Chen Z H 2020 A comprehensive review of battery modeling and state estimation approaches for advanced battery management systems *Renew Sust Energ Rev* **131** 110015

[2] O'Malley R, Liu L and Depcik C 2018 Comparative study of various cathodes for lithium ion batteries using an enhanced Peukert capacity model *Journal of Power Sources* **396** 621-31

[3] Wang S, Fernandez C, Yu C, Fan Y and Cao W Stroe D-I 2020 A novel charged state prediction method of the lithium ion battery packs based on the composite equivalent modeling and improved splice Kalman filtering algorithm *Journal of Power Sources* **471** 228450

[4] Jilte R, Afzal A and Panchal S 2021 A novel battery thermal management system using nano-enhanced phase change materials *Energy* **219** 119564

[5] Panchal S, Dincer I, Agelin-Chaab M, Fraser R and Fowler M 2016 Experimental and theoretical investigations of heat generation rates for a water cooled LiFePO4 battery *International Journal of Heat and Mass Transfer* **101** 1093-102

[6] Liu S, Liu X, Dou R, Zhou W, Wen Z and Liu L 2020 Experimental and simulation study on thermal characteristics of 18,650 lithium–iron–phosphate battery with and without spot-welding tabs *Applied Thermal Engineering* **166** 114648

[7] Meng H X and Li Y F 2019 A review on prognostics and health management (PHM) methods of
lithium-ion batteries

[8] Liu L and Zhu M 2014 Modeling of SEI Layer Growth and Electrochemical Impedance Spectroscopy Response using a Thermal-Electrochemical Model of Li-ion Batteries ECS Transactions 61(27) 43-61

[9] Tian H X, Qin P L, Li K and Zhao Z 2020 A review of the state of health for lithium-ion batteries: Research status and suggestions Journal of Cleaner Production 261 120813

[10] Jokar A, Rajabloo B, Desilets M and Lacroix M 2016 Review of simplified Pseudo-two-Dimensional models of lithium-ion batteries Journal of Power Sources 327 44-55

[11] Li J, Adewuyi K, Lotfi N, Landers R G and Park J 2018 A single particle model with chemical/mechanical degradation physics for lithium-ion battery State of Health (SOH) estimation Applied Energy 212 1178-90

[12] Ng M F, Zhao J, Yan Q Y, Conduit G J and Seh Z W 2020 Predicting the state of charge and health of batteries using data-driven machine learning Nature Machine Intelligence 2(3) 161-70

[13] Wang D, Kong J Z, Zhao Y and Tsui K L 2019 Piecewise model based intelligent prognostics for state of health prediction of rechargeable batteries with capacity regeneration phenomena Measurement 147

[14] Qiu X, Wu W and Wang S 2020 Remaining useful life prediction of lithium-ion battery based on improved cuckoo search particle filter and a novel state of charge estimation method Journal of Power Sources 450 227700

[15] Shen S, Sadoughi M, Li M, Wang Z D and Hu C 2020 Deep convolutional neural networks with ensemble learning and transfer learning for capacity estimation of lithium-ion batteries Applied Energy 260

[16] Shu X, Li G, Shen J W, Lei Z Z, Chen Z and Liu Y G 2020 A uniform estimation framework for state of health of lithium-ion batteries considering feature extraction and parameters optimization Energy 204 117957

[17] Li Y, Zou C, Bercicibar M, Nanini-Maury E, Chan J C W, van den Bossche P, Van Mierlo J and Omar N 2018 Random forest regression for online capacity estimation of lithium-ion batteries Applied Energy 232 197-210

[18] Liu W and Xu Y 2020 Data-Driven Online Health Estimation of Li-Ion Batteries Using A Novel Energy-Based Health Indicator IEEE Transactions on Energy Conversion 35(3) 1715-8

[19] Zhang Y, Tang Q, Zhang Y, Wang J, Stimming U and Lee A A 2020 Identifying degradation patterns of lithium-ion batteries from impedance spectroscopy using machine learning Nat Commun. 11(1) 1706

[20] Dai H, Zhao G, Lin M, Wu J and Zheng G 2019 A Novel Estimation Method for the State of Health of Lithium-Ion Battery Using Prior Knowledge-Based Neural Network and Markov Chain IEEE Transactions on Industrial Electronics 66(10) 7706-16

[21] Yang J, Du C Y, Liu W, Wang T, Yan L Q, Gao Y Z, Cheng X Q, Zuo P J, Ma Y L, Yin G P and Xie J Y 2020 State-of-health estimation for satellite batteries based on the actual operating parameters - Health indicator extraction from the discharge curves and state estimation Journal of Energy Storage 31 101490

[22] Feng J, Kvam P and Tang Y 2016 Remaining useful lifetime prediction based on the damage-marker bivariate degradation model: A case study on lithium-ion batteries used in electric vehicles Engineering Failure Analysis 70 323-42

[23] Schulz E, Speekenbrink M and Krause A 2018 A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions Journal of Mathematical Psychology 85 1-16

[24] Severson K A, Attia P M, Jin N, Perkins N, Jiang B, Yang Z, Chen M H, Aykol M, Herring P K, Fraggedakis D, Bazan M Z, Harris S J, Chueh W C and Braatz R D 2019 Data-driven prediction of battery cycle life before capacity degradation Nature Energy 4(5) 383-91