Gesellschaft für Informatik (Hrsg.): SKILL 2022,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2022 1

Implementierung und Analyse von Gradientenberechnung in Quantenalgorithmen

Moritz Schmidt[1]

Abstract: Quantencomputer bieten die theoretische Möglichkeit, verschiedenste Probleme präziser und schneller zu lösen als klassische Computer. Auch im Gebiet des maschinellen Lernens, welches in den letzten Jahren in einem immer größer werdenden Spektrum an Disziplinen Anwendung findet, hofft man, das Potential des Quantencomputers zu entfalten. Viele Algorithmen des maschinellen Lernens sind im Kern Optimierungsprobleme. Um eine möglichst genaue Lösung für diese Probleme zu finden, werden oft gradientenbasierte Verfahren als Kompromiss zwischen Rechenaufwand und Qualität der Lösung verwendet. In dieser Arbeit werden verschiedene Methoden zur Bestimmung von Gradienten von Funktionen, die durch Quantenschaltkreise implementiert werden, analysiert und verglichen. Die Ergebnisse zeigen, wie die inhärente Varianz von Messungen auf Quantencomputern zu einem Dilemma bei der Wahl von Hyperparametern von numerischen Verfahren führt, warum das analytische Parameter-Shift Verfahren einzelne Gradienten nicht nur exakt, sondern auch effizient berechnet und warum das SPSA Verfahren vor allem zur Gradientenberechnung auf großen Schaltkreisen mit vielen Parametern eine gute numerische Alternative sein kann. Dies kann als Entscheidungsgrundlage zur Gradientenberechnung für zukünftige Implementierungen von Algorithmen des maschinellen Lernens auf Quantencomputern dienen.

Keywords: Quanteninformatik; Quantum Machine Learning; Variationelle Quantenalgorithmen; Gradientenverfahren

1 Einleitung

Variationelle Quantenalgorithmen (VQAs) sind eine Klasse von Quantenalgorithmen, die Schaltkreise aus parametrisierten Gattern verwenden, sogenannte *Ansätze* oder Variational Forms, deren Parameter mit klassischen Methoden optimiert werden. Als Verlustfunktion fungiert dabei die Messung in einer problemspezifischen Messmatrix, der sogenannten *Observable*, beziehungsweise eine Funktion über mehrere solcher Messungen.

Ein als gut angesehener Ansatz kann mit wenig Gattern eine große Menge unitärer Operationen generieren [Du20]. Man muss sich also zu jedem Problem nicht spezifisch einen konkreten Schaltkreis überlegen, sondern hofft, dass die gewünschte unitäre Operation durch den Ansatz erzeugt und die dazugehörige Parameterkonfiguration durch Optimierung

[1] Universität Stuttgart, Institut für Architektur von Anwendungssystemen, Universitätsstraße 38, 70569 Stuttgart, Deutschland st165607@stud.uni-stuttgart.de

gefunden werden kann. Auf das maschinelle Lernen bezogen lassen sich Parallelen zum klassischen Feature bzw. Representation Learning erkennen: Beispielsweise soll ein neuronales Netzwerk bei einem Klassifizierungsproblem eine Funktion darstellen, die Eingabedaten verschiedenen Klassen zuordnet, wobei man hofft, dass die gewünschte Funktion durch die Netzwerkarchitektur möglichst akkurat dargestellt werden kann.

Bekannte VQAs sind der Quantum Approximate Optimization Algorithmus (QAOA), der beispielsweise angewandt auf das Max-Cut Problem in Clusteringalgorithmen Verwendung findet [FGG14] oder der Variational Quantum Eigensolver (VQE) [Pe14], der in modifizierter Form [Ce20] für PCA verwendet werden kann.

VQAs sind in der heutigen Noisy Intermediate-Scale Quantum (NISQ) Ära [Pr18], in der Gatter von Quantencomputern fehlerbehaftet sind und ungenauer werden, je länger die Berechnung andauert, besonders prominent, da in einer Iteration ein einzelner Schaltkreis meist kurz ist. Außerdem bestehen die Ansätze meist aus Gattern, die von vielen Quantencomputern als Basisgatter verwendet werden.

Da VQAs iterative Optimierungsalgorithmen sind, können je nach Verlustfunktion, Ansatz, Observable und Optimierungsalgorithmus viele Ausführungen nötig sein, bis das Minimum gefunden wurde. Falls die Verlustfunktion mehrere Minima besitzt, kann der Algorithmus auch in einem lokalen Minimum stecken bleiben.

Auch wenn einige Optimierungsverfahren, wie der Nelder-Mead Algorithmus [NM65], ohne Gradienteninformationen optimieren können, liefern gradientenbasierte Verfahren oft effizientere und bessere Lösungen. Um nun auch mit gradientenbasierten Optimierungsverfahren Probleme des maschinellen Lernens auf Quantencomputern zu lösen, gilt es zu klären, ob sich möglichst genaue Quantengradienten mit akzeptablem Aufwand berechnen lassen.

2 Quantengradientenverfahren

2.1 Problemdefinition

Gegeben sei ein Schaltkreis $U(\theta)$ auf beliebig vielen Qubits, der auf den Startzustand $|0\rangle$ (alle Qubits sind im Zustand 0) angewandt wird. Die Gatter des Schaltkreises können dabei von den k Parametern $\theta \in \mathbb{R}^k$ abhängen. Der resultierende Zustand nach Anwendung von $U(\theta)$ ist der Zustand $|\psi(\theta)\rangle$. Danach misst man in Observable $O(\omega)$, die von m Parametern $\omega \in \mathbb{R}^m$ abhängt, mit Erwartungswert $\langle\psi(\theta)| O(\omega) |\psi(\theta)\rangle$.

Man kann den Gradienten des Erwartungswerts bezüglich Observablenparametern ω oder Schaltkreisparametern θ berechnen. Der Fokus der Arbeit liegt auf Gradientenberechnung bezüglich Schaltkreisparametern θ.

In der Arbeit wird davon ausgegangen, dass der Erwartungswert der Messung in $O(\omega)$ als Verlustfunktion verwendet wird. Die Verlustfunktion kann auch eine Funktion eines oder mehrerer Erwartungswerte sein. Ein Beispiel dafür ist eine übliche Verlustfunktion des maschinellen Lernens, bei der für jedes Eingabedatum der Erwartungswert des Schaltkreises ausgewertet wird und der mittlere quadratische Fehler zu gewünschten Zieldatenpunkten als Verlustfunktion $L(\theta)$ fungiert. Die konkrete Ableitung kann dann über die Kettenregel bestimmt werden, wobei zusätzliche klassische Berechnungen hinzukommen. Da unsere Messergebnisse nur Schätzer der Erwartungswerte sind, kann es hier zu Schwierigkeiten kommen, falls nicht-lineare Funktionen verwendet werden [Sw20].

Im Weiteren werden Approximationen der partiellen Ableitung des i-ten Parameters θ_i als g_i bezeichnet. Zur einfacheren Notation wird im Weiteren bei der Berechnung von g_i am Parameterpunkt θ die Messung der Verlustfunktion $L(\theta + e_i h)$, also an einer Stelle, an der nur der Parameter, bezüglich dessen gerade die partielle Ableitung berechnet wird, um einen Skalar h vom Ausgangspunkt verschoben wurde, mit $L(\theta + h)$ abgekürzt.

2.2 Finite Differenzen

Finite Differenzen Methoden erlauben das Approximieren von Ableitungen und finden vor allem Anwendung in numerischen Methoden zur Lösung von Differenzialgleichungen [Sm85].

Aus der Taylor-Reihenentwicklung ergeben sich die zentralen finiten Differenzen:

$$g_{iCFD} = \frac{L(\theta + h) - L(\theta - h)}{2h} + O(h^2) \tag{1}$$

Übertragen auf den Quantencomputer lassen sich so mit zwei Messungen pro Parameter die einzelnen partiellen Ableitungen $\frac{\partial L}{\partial \theta_j}$ des Gradienten approximieren.

2.3 Simultaneous Perturbation Stochastic Approximation

Ein weiteres numerisches Verfahren zur Gradientenapproximation ist Simultaneous Perturbation Stochastic Approximation (SPSA) [Sp87]. Im Gegensatz zu zentralen Differenzen benötigt SPSA nur 2 Funktionsaufrufe für beliebig viele Parameter. Dazu wählt man einen zufällig gewählten Vektor Δ, der, mit einem festen Abstand h skaliert, als Abstandsvektor fungiert. Die einzelnen Gradiententerme werden dann ähnlich zu zentralen Differenzen berechnet, wobei die zwei Funktionswerte jedoch für alle Ableitungen gleich bleiben und nur die Skalierung mithilfe der einzelnen Elemente $h\Delta_i$ des Abstandsvektors variiert wird:

$$g_i(\Delta)_{SPSA} = \frac{L(\theta + h\Delta) - L(\theta - h\Delta)}{2h\Delta_i} \tag{2}$$

Dies ist meist eine gröbere Abschätzung des Gradienten, aber bei passend gewählter Wahrscheinlichkeitsverteilung für Δ ist g_{SPSA} ein erwartungstreuer Schätzer des Gradienten. Hier wurden die Einträge Δ_i unabhängig voneinander zufällig aus $\{\pm 1\}$ anhand einer Bernoulliverteilung mit Wahrscheinlichkeit $\frac{1}{2}$ für beide Fälle gezogen.

Für verrauschte Funktionen, wie Messungen auf Quantencomputern, wird empfohlen, in jeder Iteration den Gradienten genauer zu approximieren, indem man den Durchschnitt mehrerer solcher Gradientenapproximationen berechnet.

2.4 Parameter-Shift

Das bekannteste Verfahren zur analytischen Gradientenberechnung auf Quantencomputern ist das Parameter-Shift Verfahren [Sc19].

Sei lediglich das Gatter $\mathbb{G}(\theta)$ vom Parameter θ_j abhängig. Jede unitäre Matrix lässt sich als Matrixexponential e^{iG} der hermiteschen Generatormatrix G schreiben. Wenn sich $\mathbb{G}(\theta_j)$ schreiben lässt als $e^{i\theta_j G}$ und für den Generator G gilt, dass $G^2 = r^2 I$, dann gilt Gleichung (3).

$$g_{iPS} = \frac{\partial L}{\partial \theta_j} = r[L(\theta_j + \frac{\pi}{4r}) - L(\theta_j - \frac{\pi}{4r})] \tag{3}$$

Der Name des Verfahrens ergibt sich daraus, dass nur die Parameterwerte verschoben werden, der sonstige Schaltkreis aber komplett identisch bleibt. Man kann in diesem Fall also den Gradienten eines Schaltkreises mit zwei Anwendungen desselben Schaltkreises berechnen.

Falls die Generatorbedingung $G^2 = r^2 I$ nicht erfüllt wird, aber G genau zwei Eigenwerte λ_1, λ_2 mit $2r' = \lambda_1 - \lambda_2$ besitzt, können mittels zusätzlicher Phasenverschiebung die Eigenwerte auf $\lambda_1 = -\lambda_2 = r'$ gesetzt werden, sodass das Verfahren angewandt werden kann. Alle einzelnen Paulistrings erfüllen diese Eigenschaft als Generator, somit lässt sich beispielsweise der Gradient von den 1-Qubit Rotationsmatrizen R_x, R_y, R_z mittels Parameter-Shift berechnen.

Interessant ist die Ähnlichkeit zur zentralen finiten Differenzen Gleichung (1). Die beiden Verfahren unterscheiden sich nur im Bezug auf den Skalierungsfaktor der Differenz der beiden Messwerte.

Wenn es Verfahren ähnlich zu zentralen Differenzen gäbe, wären Verfahren ähnlich zu vorwärts oder rückwärts Differenzen von großem Vorteil, da der Messwert $L(\theta)$ an der Samplestelle dann für alle partiellen Ableitungen verwendet werden kann, also man sich die Hälfte aller Messungen sparen könnte. In [Hu22] wurde jedoch gezeigt, dass ein solches Verfahren nicht existieren kann.

3 Ergebnisse

3.1 Experimente

Als Experimente wurden die Gradientenverfahren in Qiskit [AN21] implementiert und mittels QASM-Simulator getestet. Viele verschiedene Faktoren spielen bei der Gradientenbestimmung eine Rolle, wie die Observable und der Ansatz, die verwendet werden. Diese können auch wieder Hyperparameter haben. VQE-Ansätze zum Beispiel sind häufig als sich wiederholende Schichten aufgebaut, mit denen man die Tiefe des Schaltkreises skalieren kann. Bei der Observable ist die Dimensionalität selbst relevant, welche die Mindestanzahl an benötigten Qubits impliziert, aber auch die Anzahl und Orientierungen der Eigenvektoren, sowie die Größenordnungen der Eigenwerte.

In den Experimenten wurde der in Qiskit als EfficientSU2 bezeichnete Schaltkreis verwendet, der konzeptionell dem Hardware Efficient Ansatz aus [Ka17] ähnelt. Der Schaltkreis besteht aus alternierenden 1-Qubit Gatter- und Verschränkungsschichten. Als 1-Qubit Gatter wurden die Rotationsgatter R_x und R_z , sowie CNOT-Gatter als Verschränkungsgatter angewandt (siehe Abb. 1). Da als parametrisierte Gatter nur einfache Rotationsgatter verwendet wurden, lässt sich auf allen Gattern das Parameter-Shift Verfahren anwenden.

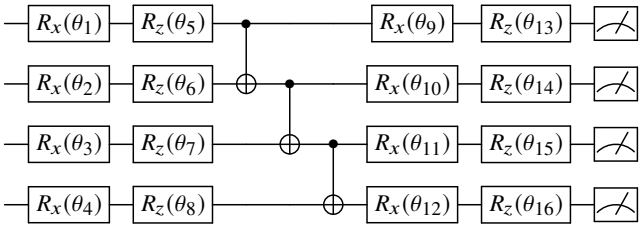

Abb. 1: EfficientSU2 Schaltkreis mit CNOT-Gattern in der Verschränkungsschicht, sowie R_x und R_z Gattern in der SU2-Schicht

Als Observable wurde in den Experimenten eine einfache Diagonalmatrix, wie in Gleichung (4) dargestellt, verwendet. Die Eigenwerte wurden als $\lambda_i = 2i$ gewählt, ähnlich zur globalen Observable im Variational Quantum State Eigensolver [Ce20], bei der die Eigenwerte ebenfalls monoton steigen ($\lambda_i < \lambda_{i+1}$).

$$O_{Diag}(\lambda_1, \ldots, \lambda_N) = \begin{pmatrix} \lambda_1 & 0 & \ldots & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & \ldots & 0 & \lambda_N \end{pmatrix} \tag{4}$$

Der Gradient ist immer abhängig von der Parameterposition an der gemessen wird. Eine Abtastung des gesamten Parameterraumes mit einem feinen Gitter, um allgemeine Aussagen

treffen zu können, ist anhand der Laufzeiten einzelner Messungen praktisch nicht möglich. Daher wurden in dieser Arbeit gleichverteilte zufällige Parameterpunkte für die Experimente genutzt.

Da die Experimente als Grundlage zur Anwendung auf NISQ-Quantencomputern dienen sollen, kommen noch zusätzliche Faktoren wie begrenzte Anzahl Shots und verrauschtes Ausführen der Schaltkreise hinzu. Da alle Experimente auf dem Simulator ausgeführt wurden, spielt Rauschen für die Experimente keine Rolle. In der Praxis sollten die einfachen Verfahren ähnlich von Rauschen und ungenauen Gattern betroffen sein, da die untersuchten Methoden die gleichen Schaltkreise verwenden und sich nur in Parameterwerten und klassischen Nachberechnungen unterscheiden.

In den Experimenten wurde im Detail untersucht, wie sich verschiedene Shotzahlen und die Wahl der Hyperparameter der Verfahren auf die Gradientenapproximation auswirken.

Für SPSA und zentrale finite Differenzen wurde ein Abstand von $h = 0.01$ gewählt. Da SPSA weniger Shots als die anderen Verfahren verwendet, wurde der Mittelwert über 10 Ausführungen gebildet.

Die hier besprochenen Ergebnisse wurden mit 4 Qubits und dem Ansatz aus Abb. 1 mit einer zusätzlichen Verschränkungs- und 1-Qubit Gatter Schicht generiert. Es wurden verschiedene Anzahlen an Schichten mit unterschiedlichen Arten von Verschränkungsschichten und 1-Qubit Gattern getestet. Dies schien aber keinen signifikanten Einfluss auf das Verhältnis zwischen den Fehlern der verschiedenen Gradientenverfahren zu haben.

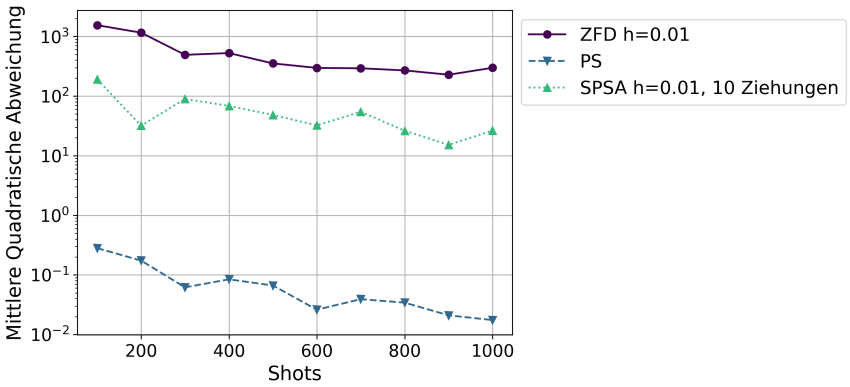

Abb. 2: Durchschnittlicher quadratischer Fehler über alle partiellen Ableitungen an zufälligem Parameterpunkt für Parameter-Shift (PS), zentrale finite Differenzen (ZFD) und SPSA (gemittelt über 10 Approximationen). Die Anzahl an Shots wurde für jede individuelle Messung verwendet.

Aus den in Abb. 2 gezeigten Ergebnissen der Experimente lassen sich einige Hypothesen über den allgemeinen Einsatz der Verfahren formulieren:

Hypothese 1: Parameter-Shift berechnet deutlich bessere Approximationen als die numerischen Verfahren.

Hypothese 2: SPSA berechnet bessere Approximationen als finite Differenzen.

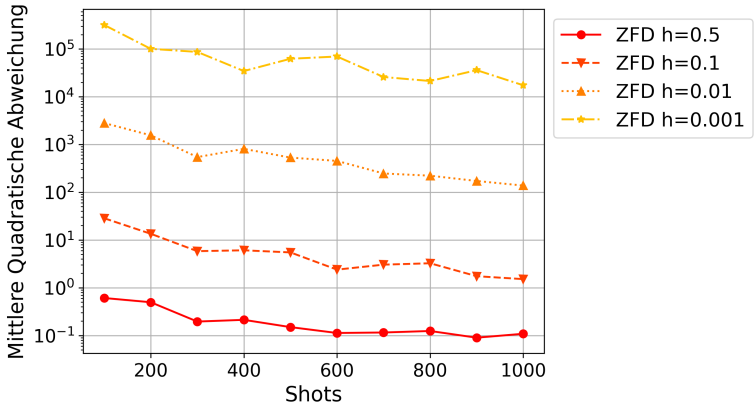

Abb. 3: Durchschnittlicher quadratischer Fehler über alle partiellen Ableitungen an zufälligem Parameterpunkt für zentrale finite Differenzen (ZFD) mit verschiedenen Abstandswerten h. Die Anzahl an Shots wurde für jede individuelle Messung verwendet.

Außerdem ließ sich aus weiteren Experimenten, deren Ergebnisse in Abb. 3 dargestellt werden, die folgende Hypothese folgern:

Hypothese 3: Finite Differenzen mit kleinerem h berechnet schlechtere Approximationen.

Hypothese 2 entspricht dabei nicht den Erwartungen zu SPSA in anderen Problemstellungen, da der eigentliche Vorteil von SPSA darin besteht, schlechtere, aber dafür weniger aufwendigere Approximationen iterativ auszugleichen. Auch hat SPSA bei den Versuchen insgesamt weniger Shots benötigt als alle anderen Verfahren. Bei $k = 24$ Parametern benötigten die anderen Verfahren $2k = 48$ Messungen. SPSA benötigt zwei Messungen pro Gradientenapproximation. Da über 10 Approximationen der Mittelwert gebildet wurde, wurde insgesamt also nur 20 gemessen.

Auch Hypothese 3 scheint unintuitiv, da man normalerweise $0 < h < 1$ möglichst klein wählt, um den Fehlerterm $O(h^2)$ in Gleichung (1) zu verringern. Wenn man spezifisch zentrale finite Differenzen als den Differenzenquotienten der Ableitungsdefinition betrachtet, ergibt die Wahl eines kleineren Abstands ebenfalls Sinn, da ein kleineres h einem besseren Annähern des Grenzwertes, also der exakten Ableitung entspricht. Auf einem Computer entstehen ab einer gewissen Nachkommastelle, aufgrund der begrenzten Darstellungsmöglichkeit von Fließkommazahlen, Rundungsfehler. Wenn Abstände h diese Größenordnung erreichen, können diese Rundungsfehler größere Approximationsfehler zur Folge haben. In den

Experimenten führten kleinere h aber schon zu immer schlechteren Ergebnissen, bevor Rundungsfehler eine Rolle spielten.

Auffällig ist, dass dieser Fehler kontinuierlich mit mehr Shots sinkt. Weitere Experimente zeigten, dass finite Differenzen mit kleineren Abständen deutlich mehr Shots benötigt um zu konvergieren.

3.2 Analyse

Aufgrund des durch die Messungen entstehenden Zufalls lassen sich die konkreten Gradientenberechnungen als Zufallsvariablen auffassen und deren Fehler in Bias und Varianz aufteilen. Der Bias entspricht dabei dem Fehler des Erwartungswertes des Verfahrens. Die Varianz beschreibt die zu erwartende quadratische Abweichung vom Bias. Messungen mit mehr Shots verringern die Varianz des Messergebnisses. Für unendlich Shots geht die Varianz gegen 0; man misst also sicher den Erwartungswert und es bleibt lediglich der Bias als Restfehler.

In den Experimenten mit NISQ-üblichen Shotzahlen von bis zu 1000 Shots war jedoch nicht der Bias, sondern die Varianz für den Großteil des Fehlers verantwortlich. Der Bias verhält sich den Erwartungen entsprechend: SPSA hat dort den größten Fehler, Parameter-Shift ist exakt und für finite Differenzen sinkt der Bias mit geringerem Abstand h.

Die Varianz der Messung eines Quantenschaltkreises $L(\theta)$ ist bekannt und wird auch als One-Shot Varianz [Hu22] bezeichnet, da es die Varianz bei einer Messung mithilfe eines einzigen Shots beschreibt. Daraus lassen sich direkt die Varianzen von Parameter-Shift und zentralen finiten Differenzen erschließen, wenn in den Messungen nur ein einzelner Shot verwendet wird.

$$Var(g_{iPS}) = r^2(Var(L(\theta + \frac{\pi}{4r})) + Var(L(\theta - \frac{\pi}{4r}))) \tag{5}$$

$$Var(g_{iCFD}) = \frac{1}{4h^2}(Var(L(\theta + h)) + Var(L(\theta - h))) \tag{6}$$

Die Varianz nach n Shots pro Messung lässt sich daraus als $\frac{Var(g_i)}{n}$ berechnen. Solange nicht $h = \frac{\pi}{4r}$ gewählt wird, messen Parameter-Shift und zentrale finite Differenzen an unterschiedlichen Parameterpunkten. Man kann dann die beiden Varianzen daher eigentlich nicht direkt vergleichen, da sie abhängig von den Varianzen an den Messpunkten sind. Folgende Annahme aus [MBK21] erlaubt jedoch eine andere Perspektive:

Annahme 1: Die Varianz der Messung in einer Observable hängt nur schwach von der Verschiebung des Parameterwertes ab, sodass $Var(L(\theta+h)) + Var(L(\theta-h)) \backsimeq 2Var(L(\theta))$ für alle Werte von h gilt.

Für diese Annahme lässt sich leicht ein Gegenbeispiel konstruieren. Eine empirische Untersuchung in Form einer Stichprobe an verschiedenen zufälligen Parameterpunkten unterstützt jedoch die Annahme, gerade bei kleinen Abständen. Oft unterscheiden sich $Var(L(\theta + h))$, $Var(L(\theta - h))$ und $Var(L(\theta))$ kaum voneinander.

Eine mögliche Erklärung für dieses Verhalten könnte hier das Barren Plateau Phänomen [Mc18] sein. Die Anzahl an Zuständen, deren Erwartungswerte sich vom durchschnittlichen Erwartungswert signifikant unterscheiden, sinkt exponentiell mit der Anzahl an Qubits. In Plateaus aus Zuständen mit durchschnittlichen Erwartungswerten ist dann der Gradient sehr gering, jedoch kein Extremum vorhanden.

Eine Idee ist daher, auch wenn das Problem mit 4 Qubits noch relativ klein ist, das dieses Phänomen hier bereits Auswirkungen hat. Erst wenn man sich an dem Punkt θ, beziehungsweise den Verschiebungen um h, einem Eigenzustand der Observablen nahe genug annähert, unterscheidet sich die Varianz der Messung deutlich von der durchschnittlichen Varianz.

Wenn man davon ausgeht, dass ein gleich verteilter zufälliger Parameterpunkt circa zu einem gleich verteilten zufälligen Zustand führt, ist die Chance deutlich höher einen Zustand mit durchschnittlicher Varianz zu treffen, als einen Zustand in der Nähe eines Eigenvektors. Aufgrund der empirischen Ergebnisse wird die Annahme als zutreffend angenommen und für die weitere Analyse verwendet.

$$Var(g_{iPS}) \simeq 2\sigma^2 r^2 \tag{7}$$

$$Var(g_{iCFD}) \simeq \frac{\sigma^2}{2h^2} \tag{8}$$

Gleichung (7) und Gleichung (8) zeigen nun die Varianzen der Verfahren mit fester Varianz σ^2 für alle Messungen. Wie zu sehen, hängt die Varianz nun nur von den Faktoren r und $\frac{1}{2h}$, die in den ursprünglichen Gradientenformeln (Gleichung (1), Gleichung (3)) die Messwerte skalieren, ab. Für die numerischen Verfahren wird die Varianz mit $O(h^{-2})$ skaliert. Da üblicherweise ein Abstand $0 < h \ll 1$ gewählt wird, um einen geringeren Bias zu produzieren, wird die Varianz aufgebläht. Wenn man die Parameter aus den Experimenten ($h = 0.01$, $r = \frac{1}{2}$ für Rotationsgatter) einsetzt, erzielt das zentrale finite Differenzen Verfahren erst mit 10000 Shots die Varianz, die Parameter-Shift mit einem einzigen Shot generiert.

Dadurch lassen sich die Hypothesen 1 und 3 erklären. Kleinere Abstände führen zu größeren Varianzen, die mehr Shots benötigen, um sich dem Erwartungswert anzunähern. Da in den Experimenten der Fehler aus der Varianz dem des Bias überwiegt, lieferten die größeren Abstände bessere Ergebnisse. Die guten Resultate von Parameter-Shift lassen sich auf das kleine r der Rotationsgatter zurückführen. Das Verfahren hat keine allgemein geringere Varianz als die numerischen Verfahren.

SPSA muss gesondert betrachtet werden, da es zwei Arten von Varianzen hat. Je nach Abstandsvektor Δ unterscheidet sich der Erwartungswert und die Varianz von g_i, da an unterschiedlichen Parameterpunkten gemessen wird. Analog zum Vorgehen im vorherigen Abschnitt lässt sich die Varianz einer Approximation g_i bestimmen, sobald ein spezifisches Δ gewählt wurde:

$$Var(g_i(\Delta)_{SPSA}) = \frac{Var(L(\theta + h\Delta)) + Var(L(\theta - h\Delta))}{4h^2} \tag{9}$$

Die erste Art von Varianz ist $\sigma_M^2 = \mathbb{E}_\Delta[Var(g_i(\Delta)_{SPSA})]$, die zu erwartende Varianz eines $g_i(\Delta)$ aufgrund der inhärenten Varianz der Messungen.

Aber auch die Erwartungswerte der Approximationen $\mathbb{E}[g_i(\Delta)]$ schwanken um den gesamten Erwartungswert $\mathbb{E}_\Delta[\mathbb{E}[g_i(\Delta)]]$ über alle möglichen Δ. Diese zweite Art der Varianz $\sigma_\Delta^2 = \mathbb{E}_\Delta[(\mathbb{E}[g_i|\Delta] - \mathbb{E}_\Delta[\mathbb{E}[g_i(\Delta)]])^2]$ sagt aus, wie groß die zu erwartende quadratische Abweichung zum Erwartungswert ist, wenn alle Messungen perfekt wären.

Die Gesamtvarianz von SPSA ergibt sich dann als:

$$Var(g_{iSPSA}) = \sigma_\Delta^2 + \sigma_M^2 \tag{10}$$

In den Experimenten ging der überwiegende Teil der Varianz von der durchschnittlichen Messvarianz σ_M^2 aus. Unter der vorherigen Annahme, dass die Varianz einzelner Messungen konstant mit σ^2 ist folgt:

$$\sigma_M^2 = \mathbb{E}_\Delta[Var(g_i(\Delta)_{SPSA})] = \sum_{j=0}^{2^k} p(\Delta_j) Var(g_i(\Delta_j)_{SPSA}) \simeq \frac{1}{2^k} \sum_{j=0}^{2^k} \frac{2\sigma^2}{4h^2} = \frac{\sigma^2}{2h^2} \tag{11}$$

σ_M^2 ist somit identisch zur Varianz von zentralen finiten Differenzen. Der entscheidende Punkt, warum SPSA eine geringere Varianz besitzt, besteht darin, dass jede Messung die Varianz in allen Dimensionen verringert. Während bei zentralen finiten Differenzen $2k$ Shots benötigt werden, um die Messvarianz σ_M^2 in jedem Parameter zu erzeugen, braucht SPSA nur 2 Messungen und generiert zusätzlich nur die Varianz σ_Δ^2, welche basierend auf den Experimenten vernachlässigbar ist. Somit lässt sich schlussendlich auch Hypothese 2 erklären.

4 Zusammenfassung und Ausblick

In dieser Arbeit wurden verschiedene Verfahren zur Gradientenberechnung auf Quantencomputern untersucht. Als numerische Verfahren wurden finite Differenzen, sowie das

Gradientenapproximationsverfahren des SPSA Algorithmus und als quantencomputerspezifisches analytisches Verfahren das Parameter-Shift Verfahren vorgestellt.

In Experimenten wurden die Verfahren verglichen und Beobachtungen als Hypothesen formuliert und weiter analysiert. Zuerst fiel auf, dass der größte Fehler (bei NISQ üblichen Shotzahlen) durch die Varianz entsteht und von den Hyperparametern der Verfahren abhängt. Bei finiten Differenzen führt dies zu einem Dilemma bei der Wahl des Abstands h, da dort die Varianz mit $O(h^{-2})$, der Bias jedoch mit $O(h^2)$ skaliert. Diese Behauptungen stützen sich auf eine Annahme aus [MBK21], wonach die Varianzen von Messungen an einem Parameterpunkt ähnlich sind zu denen an Parameterpunkten nach beliebigen Shifts.

Im Detail wurde die Zusammensetzung der Varianz von SPSA besprochen. Auch hier überwiegt die inhärente Varianz der Messungen, welche ebenfalls mit $O(h^{-2})$ skaliert. SPSA erzeugt diese Varianz unabhängig von der Anzahl an Parametern mit nur zwei Messungen, wodurch im Verhältnis mehr Shots auf einzelne Messungen angewandt werden können, was die Varianz in allen Parametern reduziert. Diese Unabhängigkeit von der Parameteranzahl kann gerade für zukünftige, größere, variationelle Quantenalgorithmen hilfreich sein.

Neben Parameter-Shift gibt es Verfahren, die bezüglich beliebiger Gatter ableiten können [Cr19][Sc19][BC21], welche jedoch in dieser Arbeit nicht weiter behandelt wurden. Es gilt die Qualität und den Aufwand dieser Verfahren mit den Resultaten der in dieser Arbeit besprochenen numerischen Verfahren in Vergleich zu setzen.

Es bleibt offen die Ergebnisse dieser Arbeit und der unterliegenden Annahme [MBK21] mit dem Barren Plateau Phänomen in Kontext zu setzen. Außerdem gibt es auch Ideen, wie man Barren Plateaus bei der Optimierung umgehen kann. So kann beispielsweise durch Entropiemessungen festgestellt werden, ob man sich einem Barren Plateau nähert und so versucht werden, diesen auszuweichen [Sa22]. Eine andere Idee sind spezialisierte, problemspezifische Ansätze, ähnlich zu problemspezifischen Architekturen von neuronalen Netzen. Kombiniert mit einem problemspezifischen Initialzustand, könnte dies zu einer lokaleren Optimierung führen, die nicht von Barren Plateaus betroffen ist.

Literatur

[AN21] ANIS, M. S. et al.: Qiskit: An Open-source Framework for Quantum Computing, 2021.

[BC21] Banchi, L.; Crooks, G. E.: Measuring Analytic Gradients of General Quantum Evolution with the Stochastic Parameter Shift Rule. Quantum 5/, S. 386, 2021.

[Ce20] Cerezo, M.; Sharma, K.; Arrasmith, A.; Coles, P. J.: Variational quantum state eigensolver. arXiv preprint arXiv:2004.01372/, 2020.

[Cr19] Crooks, G. E.: Gradients of parameterized quantum gates using the parameter-shift rule and gate decomposition. arXiv preprint arXiv:1905.13311/, 2019.

[Du20] Du, Y.; Hsieh, M.-H.; Liu, T.; Tao, D.: Expressive power of parametrized quantum circuits. Phys. Rev. Research 2/, Juli 2020.

[FGG14] Farhi, E.; Goldstone, J.; Gutmann, S.: A quantum approximate optimization algorithm. arXiv preprint arXiv:1411.4028/, 2014.

[Hu22] Hubregtsen, T.; Frederik, W.; Qasim, S.; Eisert, J.: Single-component gradient rules for variational quantum algorithms. Quantum Science and Technology 7/, Apr. 2022.

[Ka17] Kandala, A.; Mezzacapo, A.; Temme, K.; Takita, M.; Brink, M.; Chow, J.; Gambetta, J.: Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. Nature 549/, S. 242–246, Sep. 2017.

[MBK21] Mari, A.; Bromley, T. R.; Killoran, N.: Estimating the gradient and higher-order derivatives on quantum hardware. Phys. Rev. A 103/, Jan. 2021.

[Mc18] McClean, J. R.; Boixo, S.; Smelyanskiy, V. N.; Babbush, R.; Neven, H.: Barren plateaus in quantum neural network training landscapes. Nature communications 9/1, 2018.

[NM65] Nelder, J. A.; Mead, R.: A simplex method for function minimization. The computer journal 7/4, S. 308–313, 1965.

[Pe14] Peruzzo, A.; Mcclean, J.; Shadbolt, P.; Yung, M. H.; Zhou, X.; Love, P.; Aspuru-Guzik, A.; O'Brien, J.: A variational eigenvalue solver on a photonic quantum processor. Nature communications 5/1, 2014.

[Pr18] Preskill, J.: Quantum Computing in the NISQ era and beyond. Quantum 2/, S. 79, Aug. 2018.

[Sa22] Sack, S. H.; Medina, R. A.; Michailidis, A. A.; Kueng, R.; Serbyn, M.: Avoiding barren plateaus using classical shadows. arXiv preprint arXiv:2201.08194/, 2022.

[Sc19] Schuld, M.; Bergholm, V.; Gogolin, C.; Izaac, J.; Killoran, N.: Evaluating analytic gradients on quantum hardware. Phys. Rev. A 99/, März 2019.

[Sm85] Smith, G. D.: Numerical solution of partial differential equations: finite difference methods. Oxford university press, 1985.

[Sp87] Spall, J. C.: A stochastic approximation technique for generating maximum likelihood parameter estimates. In: 1987 American control conference. IEEE, S. 1161–1167, 1987.

[Sw20] Sweke, R.; Wilde, F.; Meyer, J.; Schuld, M.; Fährmann, P.; Meynard-Piganeau, B.; Eisert, J.: Stochastic gradient descent for hybrid quantum-classical optimization. Quantum 4/, S. 314, Aug. 2020.