Computing Gradients on Quantum Computers

A prominent method for performing optimization is gradient descent.

Idea is to go in the direction of steepest descent. Update rule for parameter vector θ is

$$\theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}(\theta)$$

where θ_t is ~~value of~~ parameter vector @ time t,

η is step size or learning rate,

$\mathcal{L}(\theta)$ is the loss or cost function,

& $\nabla \mathcal{L}(\theta)$ is its gradient.

As part of this algorithm, it is necessary to evaluate the gradient.

One can do so by means of the finite, central difference approximation:

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta_i} \approx \frac{1}{2\varepsilon}\left[\mathcal{L}(\theta + \varepsilon \hat{e}_i) - \mathcal{L}(\theta - \varepsilon \hat{e}_i)\right]$$

where $\varepsilon > 0$ is small

& $\hat{e}_i$ is the unit vector along the ith component.

However it is actually possible to find an analytic (exact) expression for the gradient, which can be estimated on quantum computers & can be evaluated using the same parametrized circuit used to evaluate the loss function $\mathcal{L}(\theta)$.

- Such a result is known as a parameter-shift rule.

- It has also been proven that there are theoretical convergence advantages of analytic gradients over the finite difference approach (1901.05374)

The basic idea behind this is that

$$\frac{\partial}{\partial\theta}\left(\sin(\theta)\right) = \cos(\theta)$$

& $\quad \sin(\theta + \pi/2) = \cos(\theta)$

Recall that the function being optimized for the variational q. eigensolver is

$$\mathcal{L}(\theta) = \langle\psi(\theta)|\, H \,|\psi(\theta)\rangle \equiv \langle H(\theta)\rangle$$

& the optimization task is

$$\min_{\theta} \mathcal{L}(\theta),$$

where $|\psi(\theta)\rangle = U(\theta)\,|0\rangle^{\otimes n}$

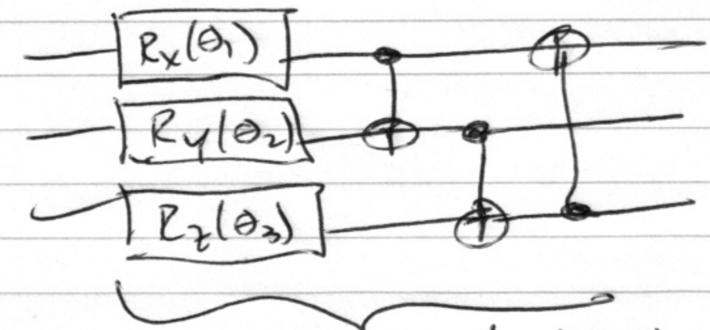
& $U(\theta) = V_L(\theta_L)\, W_L\, V_{L-1}(\theta_{L-1})\, W_{L-1} \cdots$

$\cdots V_\ell(\theta_\ell)\, W_\ell \cdots V_1(\theta_1)\, W_1$,

each $V_\ell(\theta_\ell)$ is a parameterized gate, typically taken to be Pauli rotations $R_X(\theta), R_Y(\theta), R_Z(\theta)$

& each W_ℓ is a fixed gate (not parameterized)

Recall that such circuits look something like



repeated some # of times.

The main result of the parameter-shift rule is that

$$\frac{\partial \langle H(\theta) \rangle}{\partial \theta_i} = \frac{1}{2}\left[\langle H(\theta + \tfrac{\pi}{2}\hat{e}_i) \rangle - \langle H(\theta - \tfrac{\pi}{2}\hat{e}_i) \rangle \right]$$

Thus, to evaluate the gradient analytically, one only needs to evaluate the two terms $\langle H(\theta + \frac{\pi}{2}\hat{e}_i) \rangle$ & $\langle H(\theta - \frac{\pi}{2}\hat{e}_i) \rangle$ & can do so using the same circuit as needed to evaluate $\langle H(\theta) \rangle$

Suppose we are interested in

$$\frac{\partial}{\partial\theta_i}\langle H(\theta)\rangle ,$$

one can show that, for a single paramater θ_i, the cost function can be written as a sine function:

$$\langle H(\theta)\rangle = \alpha \sin(\theta_i + \beta) + \gamma$$

where α, β, γ are functions of H + the other gates in the circuit. plausible b/c $\langle H(\theta)\rangle$ has to be real + rotation gate is $e^{-iX\theta_i/2}$

Then

$$\frac{\partial}{\partial\theta_i}\langle H(\theta)\rangle = \alpha \cos(\theta_i + \beta)$$

Goal is to express this in terms of original cost function. Then consider that

$$\langle H(\theta + \pi/2\,\hat{e}_i)\rangle = \alpha \sin(\theta_i + \pi/2 + \beta) + \gamma$$

$$= \alpha \cos(\theta + \beta) + \gamma$$

Additionally,

$$\langle H(\theta - \pi/2\, \hat{e}_i) \rangle = \alpha \sin(\theta_i - \pi/2 + \beta) + \gamma$$

$$= -\alpha \cos(\theta_i + \beta) + \gamma$$

$$\Rightarrow \frac{1}{2}\left[\langle H(\theta + \pi/2\, \hat{e}_i) \rangle - \langle H(\theta - \pi/2\, \hat{e}_i) \rangle \right]$$

$$= \alpha \cos(\theta_i + \beta)$$

$$= \frac{\partial \langle H(\theta) \rangle}{\partial \theta_i}$$

Now let us prove the claim that

$$\langle H(\theta) \rangle = \alpha \sin(\theta_i + \beta) + \gamma$$

Recall that

$$\langle H(\theta) \rangle = \langle \psi(\theta) | H | \psi(\theta) \rangle$$

where

$$|\psi(\theta)\rangle = V_L(\theta_L) W_L \cdots V_1(\theta_1) W_1 |0\rangle^{\otimes n}$$

dependence on θ_i is then

$$|\psi(\theta)\rangle = U_2 \; V_i(\theta_i) \; U_1 \; |0\rangle^{\otimes n}$$

ignore dependence on $\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_L$

since we want $\frac{\partial}{\partial \theta_i}$

then taking

$$V_i(\theta_i) = e^{-i\theta_i X/2} \quad (\text{WLOG})$$

+ expanding as $\sum_{j \in \{0,1\}} e^{-i\theta_i (-1)^j/2} |\phi_j\rangle\langle\phi_j|$

we find that

$$\langle \psi(\theta) | H | \psi(\theta) \rangle$$

$$= \langle 0 | U_1^\dagger e^{i\theta_i X/2} U_2^\dagger H U_2 e^{-i\theta_i X/2} U_1 | 0 \rangle$$

$$= \langle 0 | U_1^\dagger \sum_j e^{+i\theta_i (-1)^j/2} |\phi_j\rangle\langle\phi_j| U_2^\dagger H U_2$$

$$\sum_k e^{-i\theta_i (-1)^k/2} |\phi_k\rangle\langle\phi_k| U_1 |0\rangle$$

$$= \sum_{j,k\in\{0,1\}} e^{i\theta_i [(-1)^j - (-1)^k]/2} \langle 0|U_1^\dagger|\phi_j\rangle \langle\phi_j|U_2^\dagger H U_2|\phi_k\rangle \langle\phi_k|U_1|0\rangle$$

$$= \sum_{j,k\in\{0,1\}} e^{i\theta_i \{(-1)^j - (-1)^k\}/2} c_{jk}$$

where $c_{jk} = \langle 0| U_1^\dagger |\phi_j\rangle \langle\phi_j| U_2^\dagger H U_2 |\phi_k\rangle \cdot \langle\phi_k|U_1|0\rangle$

then expand as

$$c_{00} + c_{11} + c_{01} e^{i\theta_i} + c_{10} e^{-i\theta_i}$$

observe that $c_{01} = c_{10}^*$

$\Rightarrow$ $\langle \psi(\theta) | H | \psi(\theta) \rangle$

$$= c_{00} + c_{11} + \mathrm{RE}\left[c_{01} e^{i\theta_i} \right]$$

Set $c_{01} = \alpha e^{i\beta'}$ ← phase

↑ magnitude

$\Rightarrow$

$$= c_{00} + c_{11} + \mathrm{RE}\left[\alpha e^{i(\theta_i + \beta')} \right]$$

$$= \underbrace{c_{00} + c_{11}}_{\text{set } \gamma} + \alpha \cos(\theta_i + \beta')$$

offset β' by $\pi/2$ again & we get the original claim.