Quantum Natural Gradient

The basic update rule of gradient descent is

$$\theta_{t+1} = \theta_t - \eta \overbrace{\nabla_\theta \mathcal{L}(\theta)\big|_{\theta=\theta_t}}^{\nabla \mathcal{L}(\theta_t) \text{ (shorthand)}}$$

Where does this come from?

At least two ways of thinking about it.

1) First is as a linear approximation of $\mathcal{L}(\theta)$ w/ penalty term.

Consider Taylor expansion of $\mathcal{L}(\theta)$ about $\mathcal{L}(\theta_t)$:

$$\mathcal{L}(\theta) = \mathcal{L}(\theta_t) + \nabla \mathcal{L}(\theta_t)^T (\theta - \theta_t) + \dots$$

we can neglect higher-order terms & minimize right-hand side.

However, if we do so, then the optimal choice is θ such that $\ell(\theta) = -\infty$.

So we instead impose a penalty on step sizes that are too large, by means of a distance measure.

- Then solve the minimization problem

$$\min_{\theta \in H} \ell(\theta_t) + \nabla \ell(\theta_t)^T(\theta - \theta_t) + \frac{\gamma}{2}\|\theta - \theta_t\|_2^2$$

where γ is a penalty parameter + will corresponding to learning rate of gradient descent.

$$= \ell(\theta_t) + \min_{\theta \in H} \nabla \ell(\theta_t)^T(\theta - \theta_t) + \frac{\gamma}{2}\|(\theta - \theta_t)\|_2^2$$

minimizing objective function

(consider that $\nabla \mathcal{L}(\theta_t)$ is a function of θ_t)

gives

$$\theta = \nabla_\theta (\cdots)$$

$$= \nabla \mathcal{L}(\theta_t) + \gamma(\theta - \theta_t)$$

$\Rightarrow$ optimal solution is

$$\theta = \theta_t - \frac{1}{\gamma} \nabla \mathcal{L}(\theta_t)$$

$\Rightarrow$ gradient descent update rule

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

where $\eta = \frac{1}{\gamma}$

$\Rightarrow$ larger penalty means smaller step size & vice versa

2) Another motivation for gradient descent is from the following inequality that holds for convex functions w/ L-Lipschitz gradients, i.e.,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2 \quad \forall x, y.$$

For such a loss function $\mathcal{L}(\theta)$, the following inequality holds

$$\mathcal{L}(\theta) \geq \mathcal{L}(\theta_t) + \nabla \mathcal{L}(\theta_t)^T (\theta - \theta_t) + \frac{1}{2} L \|\theta - \theta_t\|_2^2$$

Then can prove that gradient descent w/ step size $\frac{1}{L}$ converges to global optimum in $O\left(\frac{1}{\epsilon}\right)$ steps, where ϵ is desired error.

- Main issue w/ gradient descent when used for the VQE problem is that Euclidean metric $\|\theta - \theta_t\|_2$ for parameters is not adapted to <u>geometry</u> of the quantum state space.

- A natural distance measure between two pure states $|\psi_1\rangle$ & $|\psi_2\rangle$ is Bures distance:

$$d_B(|\psi_1\rangle, |\psi_2\rangle) = \min_{\phi \in [0, 2\pi]} \| |\psi_1\rangle - e^{i\phi} |\psi_2\rangle \|_2$$

(phase minimization term is present since $|\psi\rangle$ & $e^{i\phi}|\psi\rangle$ are physically indistinguishable & so should have zero distance)

can show that

$$d_B(|\psi_1\rangle, |\psi_2\rangle) = \sqrt{2(1 - |\langle\psi_1|\psi_2\rangle|)}$$

(clearly, $d_B = 0$ iff $|\psi_1\rangle = e^{i\phi}|\psi_2\rangle$)

The idea of quantum natural gradient is to add a penalty term related to this distance (natural distance for pure q. states) rather than Euclidean distance between parameters

So we change the optimization problem to be as follows:

$$\min_{\theta \in \Theta} \mathcal{L}(\theta_t) + \nabla \mathcal{L}(\theta_t)^T (\theta - \theta_t) + \frac{\gamma}{2} d_B^2(|\psi(\theta)\rangle, |\psi(\theta_t)\rangle)$$

where $\{|\psi(\theta)\rangle\}_{\theta \in \Theta}$ is the

parameterized family of states being considered.

As stated, this optimization problem is too hard to solve.

So the next idea is to perform a Taylor expansion of

$d_B^2(|\psi(\theta)\rangle, |\psi(\theta_t)\rangle)$ (function of θ)

about θ_t.

After a long calculation, this is found to be

$$d_B^2(|\psi(\theta)\rangle, |\psi(\theta_t)\rangle)$$
$$= (\theta - \theta_t)^T I_F(\theta_t)(\theta - \theta_t) + O(\|\theta - \theta_t\|^3)$$

where $I_F(\theta_t)$ is the Fisher information matrix

w/ matrix elements given by

$$I_F(\theta) = \mathrm{RE}\left[\langle \partial_i \psi(\theta) | \left(I - |\psi(\theta)\rangle\langle\psi(\theta)|\right) | \partial_j \psi(\theta)\rangle\right]$$

where $\partial_i \equiv \frac{\partial}{\partial \theta_i}$

$I_F(\theta)$ is also known as Riemannian metric tensor & applies a transformation to parameters to account for how they change distances between q. states.

so the modified optimization problem for q. natural gradient is

$$\min_{\theta_t \in \Theta} \mathcal{L}(\theta_t) + \nabla \mathcal{L}(\theta_t)^T (\theta - \theta_t)$$

$$+ \frac{\gamma}{2} (\theta - \theta_t)^T I_F(\theta_t) (\theta - \theta_t)$$

Solving it gives the equation

$$0 = \nabla \mathcal{L}(\theta_t) + \gamma I_F(\theta_t)(\theta - \theta_t)$$

$\Rightarrow$ choose optimal θ to be solution to

$$I_F(\theta_t)(\theta - \theta_t) = -\frac{1}{\gamma} \nabla \mathcal{L}(\theta_t)$$

$$= -\eta \nabla \mathcal{L}(\theta_t)$$

If $I_F(\theta_t)$ is invertible, then pick next parameter vector θ_{t+1} to be

$$\theta_{t+1} = \theta_t - \eta \left[I_F(\theta_t)\right]^{-1} \nabla \mathcal{L}(\theta_t).$$

Can interpret this equation as rescaling the gradient vector in order to account for geometry of q. states.

Remaining question:

How to evaluate Fisher information matrix?

Suppose a parameterized circuit of the form

$$V(\theta) = U_L(\theta_L) W_L U_{L-1}(\theta_{L-1}) W_{L-1} \cdots U_2(\theta_2) W_2 U_1(\theta_1) W_1$$

acting on $|0\rangle^{\otimes n}$

$$\Rightarrow |\psi(\theta)\rangle = V(\theta)|0\rangle^{\otimes n}$$

take $U_i(\theta_i) = e^{-iH_i\theta_i}$

$$\Rightarrow |\partial_i \psi(\theta)\rangle$$

$$= U_L(\theta_L) W_L \cdots (-iH_i) U_i(\theta_i) W_i \cdots U_1(\theta_1) W_1$$

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

$$= \Big[\langle 0|^{\otimes n} W_1^\dagger U_1^\dagger(\theta_1) \cdots W_L^\dagger U_L^\dagger(\theta_L) \Big]$$

$$\Big\{ U_L(\theta_L) W_L \cdots (-iH_i) U_i(\theta_i) W_i$$

$$\cdots U_1(\theta_1) W_1 |0\rangle^{\otimes n} \Big]$$

$$= \langle 0|^{\otimes n} W_1^\dagger U_1^\dagger(\theta_1) \cdots U_{i-1}^\dagger(\theta_{i-1}) W_i^\dagger$$

$$(-iH_i) W_i \cdots U_1(\theta_1) W_1 |0\rangle^{\otimes n}$$

$$= -i \langle \psi_i | H_i | \psi_i \rangle$$

where $|\psi_i\rangle = W_i U_{i-1}(\theta_{i-1}) W_{i-1}$

$$\cdots U_1(\theta_1) W_1 |0\rangle^{\otimes n}$$

$$\Rightarrow \langle \partial_j \psi(\theta) | \psi(\theta) \rangle = -i \langle \psi_j | H_j | \psi_j \rangle$$

$$\Rightarrow \langle \partial_i \psi(\theta) | \psi(\theta) \rangle \langle \psi(\theta) | \partial_j \psi(\theta) \rangle$$

$$= \Big(\langle \psi_i | \otimes \langle \psi_j | \Big) \Big(H_i \otimes H_j \Big) \Big(|\psi_i\rangle \otimes |\psi_j\rangle \Big)$$

↑ can be evaluated as expectation of observables

Then consider that for $i > j$

$$\langle \partial_i \psi(\theta) \,|\, \partial_j \psi(\theta) \rangle$$

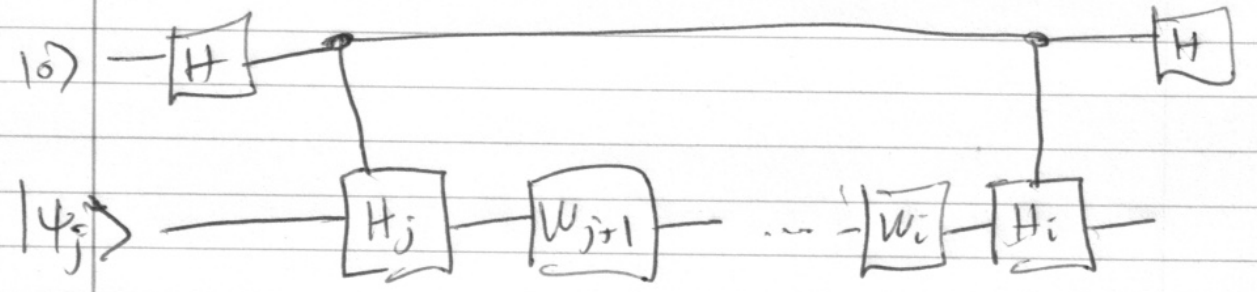
$$= \Big[\langle 0|^{\otimes n} W_1^\dagger U_1^\dagger(\theta_1) \cdots W_i^\dagger (+iH_i) U_i^\dagger(\theta_i) \cdots W_2^\dagger U_2^\dagger(\theta_2) \Big] \times$$

$$\Big\{ U_2(\theta_2) W_2 \cdots U_j(\theta_j)(-iH_j) W_j \cdots U_1(\theta_1) W_1 |0\rangle^{\otimes n} \Big]$$

$$= \langle \psi_i | H_i W_i U_{i-1}(\theta_{i-1}) W_{i-1} \cdots U_{j+1}(\theta_{j+1}) W_{j+1} H_j |\psi_j\rangle$$

$$= \langle \psi_j | W_{j+1}^\dagger U_{j+1}^\dagger(\theta_{j+1}) \cdots U_{i-1}^\dagger(\theta_{i-1}) W_i^\dagger H_i W_i U_{i-1}(\theta_{i-1}) \cdots W_{j+1} H_j |\psi_j\rangle$$

we want real part of this,
can evaluate it using the
Hadamard test when H_i & H_j are unitary



this follows b/c, controlled on $|0\rangle$ state is $W_i \cdots W_{j+1} |\psi_j\rangle$

$= |\psi_i\rangle$

+ controlled on $|1\rangle$ it is

$H_i W_i \cdots W_{j+1} H_j |\psi_j\rangle$

thus estimates

$$\mathrm{RE}\left[\langle \partial_i \psi(\theta) | \partial_j \psi(\theta)\rangle\right]$$

- when using a layered ansatz, easier to estimate

$\mathrm{RE}\left[\langle \partial_i \psi(\theta) | \partial_j \psi(\theta)\rangle\right]$ when parameters are in the same layer.

see 1909.07108