Gradient of the loss function used in QCML

Geet Rakala

April 18, 2025

1 Introduction

The loss function is defined in terms of observables that depend on the ground state of an error Hamiltonian. The challenge arises when computing the gradient of this loss, especially when the mapping has a complicated dependence on the Hamiltonian's eigenvectors. Here, we derive an expression for the gradient of the loss function using first-order perturbation theory. Another loss function – one involving a bias-variance tradeoff via quantum fluctuations – is not considered here; we will tackle that one later!

2 Loss Function Definition

The total loss is defined as the average over single-point losses:

$$L = \frac{1}{N_{\text{points}}} \sum_x L_{\text{single}}(x), \tag{1}$$

where the single-point loss is

$$L_{\text{single}}(x) = \sum_\mu \left(y_\mu(x) - x_\mu\right)^2. \tag{2}$$

The point mapping $y_\mu(x)$ depends on the ground state $|\psi_0(x)\rangle$ of the error Hamiltonian $\mathcal{H}(x)$ and is given by

$$y_\mu(x) = \mathrm{Re}\left[\langle\psi_0(x)|A_\mu|\psi_0(x)\rangle\right]. \tag{3}$$

Since A_μ is Hermitian, this simplifies to

$$y_\mu(x) = \langle\psi_0(x)|A_\mu|\psi_0(x)\rangle. \tag{4}$$

The error Hamiltonian is defined as

$$\mathcal{H}(x) = \frac{1}{2} \sum_\nu \left(A_\nu - x_\nu I\right)^\dagger \left(A_\nu - x_\nu I\right), \tag{5}$$

and, since A_ν is Hermitian and x_ν is real,

$$\mathcal{H}(x) = \frac{1}{2} \sum_\nu \left(A_\nu - x_\nu I\right)^2. \tag{6}$$

The ground state $|\psi_0(x)\rangle$ is defined by

$$\mathcal{H}(x)\,|\psi_0(x)\rangle = E_0(x)\,|\psi_0(x)\rangle, \tag{7}$$

where $E_0(x)$ is the smallest eigenvalue of $\mathcal{H}(x)$.

3 Gradient of the Single-Point Loss

To compute the gradient with respect to a specific matrix A_λ, note that the gradient $\nabla_{A_\lambda} L_{\text{single}}$ has the same dimensions as A_λ. For a small Hermitian perturbation δA_λ, the change in the loss is given by

$$\delta L_{\text{single}} = \operatorname{Tr}\left(G_\lambda\, \delta A_\lambda\right), \tag{8}$$

where the gradient matrix $G_\lambda = \nabla_{A_\lambda} L_{\text{single}}$ is also Hermitian.

The variation in the loss due to a change in y_μ is

$$\delta L_{\text{single}} = \sum_\mu 2\left(y_\mu(x) - x_\mu\right) \delta y_\mu(x). \tag{9}$$

Since $y_\mu = \langle\psi_0|A_\mu|\psi_0\rangle$, its variation under the perturbation δA_λ is

$$\begin{aligned} \delta y_\mu &= \langle\delta\psi_0|A_\mu|\psi_0\rangle + \langle\psi_0|\delta A_\mu|\psi_0\rangle + \langle\psi_0|A_\mu|\delta\psi_0\rangle \\ &= \langle\psi_0|\delta A_\mu|\psi_0\rangle + 2\operatorname{Re}\left[\langle\psi_0|A_\mu|\delta\psi_0\rangle\right]. \end{aligned} \tag{10}$$

where $\delta A_\mu = \delta_{\mu\lambda}\, \delta A_\lambda$ because only A_λ is perturbed.

3.1 Variation of the Ground State $|\delta\psi_0\rangle$

The change in the ground state due to a perturbation is given by first-order perturbation theory:

$$|\delta\psi_0\rangle = \sum_{n\neq 0} |\psi_n\rangle \frac{\langle\psi_n|\delta\mathcal{H}|\psi_0\rangle}{E_0 - E_n}, \tag{11}$$

where $|\psi_n\rangle$ and E_n are the eigenvectors and eigenvalues of $\mathcal{H}(x)$ excluding the ground state.

The perturbation $\delta\mathcal{H}$ in the error Hamiltonian is

$$\begin{aligned} \delta\mathcal{H} &= \frac{1}{2}\sum_\nu \Big[\delta\big(A_\nu - x_\nu I\big)\big(A_\nu - x_\nu I\big) + \big(A_\nu - x_\nu I\big)\delta\big(A_\nu - x_\nu I\big)\Big] \\ &= \frac{1}{2}\Big[\delta A_\lambda\left(A_\lambda - x_\lambda I\right) + \big(A_\lambda - x_\lambda I\big)\delta A_\lambda\Big]. \end{aligned} \tag{12}$$

since $\delta A_\nu = \delta_{\nu\lambda}\, \delta A_\lambda$.

3.2 Combining the Components

Substituting $|\delta\psi_0\rangle$ into the expression for δy_μ gives

$$\delta y_\mu = \delta_{\mu\lambda}\langle\psi_0|\delta A_\lambda|\psi_0\rangle + 2\ \operatorname{Re}\left[\sum_{n\neq 0} \frac{\langle\psi_0|A_\mu|\psi_n\rangle\,\langle\psi_n|\delta E|\psi_0\rangle}{E_0 - E_n}\right]. \tag{13}$$

Substituting the expression for $\delta\mathcal{H}$ into the above expression, one obtains

$$\begin{aligned} \delta y_\mu = {}& \delta_{\mu\lambda}\langle\psi_0|\delta A_\lambda|\psi_0\rangle \\ & + 2\ \operatorname{Re}\left[\sum_{n\neq 0} \frac{\langle\psi_0|A_\mu|\psi_n\rangle}{E_0 - E_n} \left\langle \psi_n \left| \frac{1}{2}\left[\delta A_\lambda(A_\lambda - x_\lambda I) \right.\right.\right.\right. \\ & \left.\left.\left.\left. + (A_\lambda - x_\lambda I)\delta A_\lambda\right] \right| \psi_0 \right\rangle\right]. \end{aligned} \tag{14}$$

Define the operator

$$M_{mn,\lambda} = \frac{1}{2}\left[(A_\lambda - x_\lambda I)\left|\psi_n\right\rangle\left\langle\psi_m\right| + \left|\psi_n\right\rangle\left\langle\psi_m\right|(A_\lambda - x_\lambda I)\right], \tag{15}$$

so that

$$\langle\psi_n|\delta\mathcal{H}|\psi_0\rangle = \mathrm{Tr}\left(M_{n0,\lambda}\,\delta A_\lambda\right). \tag{16}$$

Using the cyclic property of the trace and the identity

$$\langle\psi_0|\delta A_\lambda|\psi_0\rangle = \mathrm{Tr}\left(\left|\psi_0\right\rangle\left\langle\psi_0\right|\,\delta A_\lambda\right), \tag{17}$$

the variation δy_μ can be written as

$$\delta y_\mu = \mathrm{Tr}\left[\left(\delta_{\mu\lambda}\left|\psi_0\right\rangle\left\langle\psi_0\right| + \sum_{n\neq 0} 2\,\mathrm{Re}\left(\frac{\langle\psi_0|A_\mu|\psi_n\rangle}{E_0 - E_n}\right) M_{n0,\lambda}\right)\delta A_\lambda\right]. \tag{18}$$

This shows that the gradient of y_μ with respect to A_λ is

$$\nabla_{A_\lambda} y_\mu = \delta_{\mu\lambda}\left|\psi_0\right\rangle\left\langle\psi_0\right| + \sum_{n\neq 0} 2\,\mathrm{Re}\left(\frac{\langle\psi_0|A_\mu|\psi_n\rangle}{E_0 - E_n}\right) M_{n0,\lambda}. \tag{19}$$

4 Final Gradient Expression

Substitute $\nabla_{A_\lambda} y_\mu$ into the gradient for L_{single}:

$$\nabla_{A_\lambda} L_{\text{single}} = \sum_\mu 2\left(y_\mu(x) - x_\mu\right)\nabla_{A_\lambda} y_\mu. \tag{20}$$

Expanding this expression, we have

$$\begin{aligned}\nabla_{A_\lambda} L_{\text{single}} &= 2\big(y_\lambda - x_\lambda\big)\left|\psi_0\right\rangle\left\langle\psi_0\right| \\ &\quad + \sum_\mu \sum_{n\neq 0} 2\big(y_\mu - x_\mu\big)\ \mathrm{Re}\left(\frac{\langle\psi_0|A_\mu|\psi_n\rangle}{E_0 - E_n}\right) \\ &\quad \times \Big[\big(A_\lambda - x_\lambda I\big)\left|\psi_0\right\rangle\left\langle\psi_n\right| + \left|\psi_0\right\rangle\left\langle\psi_n\right|\big(A_\lambda - x_\lambda I\big)\Big].\end{aligned} \tag{21}$$

Define a real scalar coefficient

$$C_{n\lambda}(x) = \sum_\mu 2\left(y_\mu(x) - x_\mu\right)\mathrm{Re}\left(\frac{\langle\psi_0(x)|A_\mu|\psi_n(x)\rangle}{E_0(x) - E_n(x)}\right), \tag{22}$$

so that the gradient for a single data point x becomes

$$\begin{aligned}\nabla_{A_\lambda} L_{\text{single}}(x) &= 2\big(y_\lambda(x) - x_\lambda\big)\ \left|\psi_0(x)\right\rangle\left\langle\psi_0(x)\right| \\ &\quad + \sum_{n\neq 0} C_{n\lambda}(x)\,\Big(\big(A_\lambda - x_\lambda I\big)\left|\psi_0(x)\right\rangle\left\langle\psi_n(x)\right| + \left|\psi_n(x)\right\rangle\left\langle\psi_0(x)\right|\big(A_\lambda - x_\lambda I\big)\Big).\end{aligned} \tag{23}$$

Averaging over all data points, the total gradient is

$$\begin{aligned}\nabla_{A_\lambda} L &= \frac{1}{N_{\text{points}}}\sum_x \Big[2\big(y_\lambda(x) - x_\lambda\big)\ \left|\psi_0(x)\right\rangle\left\langle\psi_0(x)\right| \\ &\quad + \sum_{n\neq 0} C_{n\lambda}(x)\,\Big((A_\lambda - x_\lambda I)\ \left|\psi_0(x)\right\rangle\left\langle\psi_n(x)\right| + \left|\psi_n(x)\right\rangle\left\langle\psi_0(x)\right|\ (A_\lambda - x_\lambda I)\Big)\Big].\end{aligned} \tag{24}$$

5 Conclusion

1. **Pseudo Gradient:** The first term in the final expression for the gradient [Eq. (24)]:

$$\frac{1}{N_{\text{points}}} \sum_x \big(2 \left(y_\lambda(x) - x_\lambda \right) |\psi_0(x)\rangle \langle \psi_0(x)| \big) \tag{25}$$

   is equivalent to the "pseudo gradient" used in the simplified training routines.

2. **Corrections from Excited States:** The remaining terms, involving the sum over $n \neq 0$, capture the corrections arising from the change in the ground state when A_λ is perturbed.