Stochastic Gradient Descent for Hybrid Quantum-Classical Optimization

Frederik Wilde
frederikwil.de/jmc2021

Journées de la Matière Condensée
2021-08-25

Freie Universität Berlin

1) Hybrid Quantum-Classical Models
2) Optimization
3) Parameter Shift Rules
4) Stochastic Gradient Descent

Hybrid quantum-classical algorithms

$$|\psi(\boldsymbol{\theta})\rangle$$

$\boldsymbol{\theta} \in \mathbb{R}^d$

[1] A. Peruzzo et al., Nature Com. 2014 [2] Farhi et al., 1411.4028

Hybrid quantum-classical algorithms

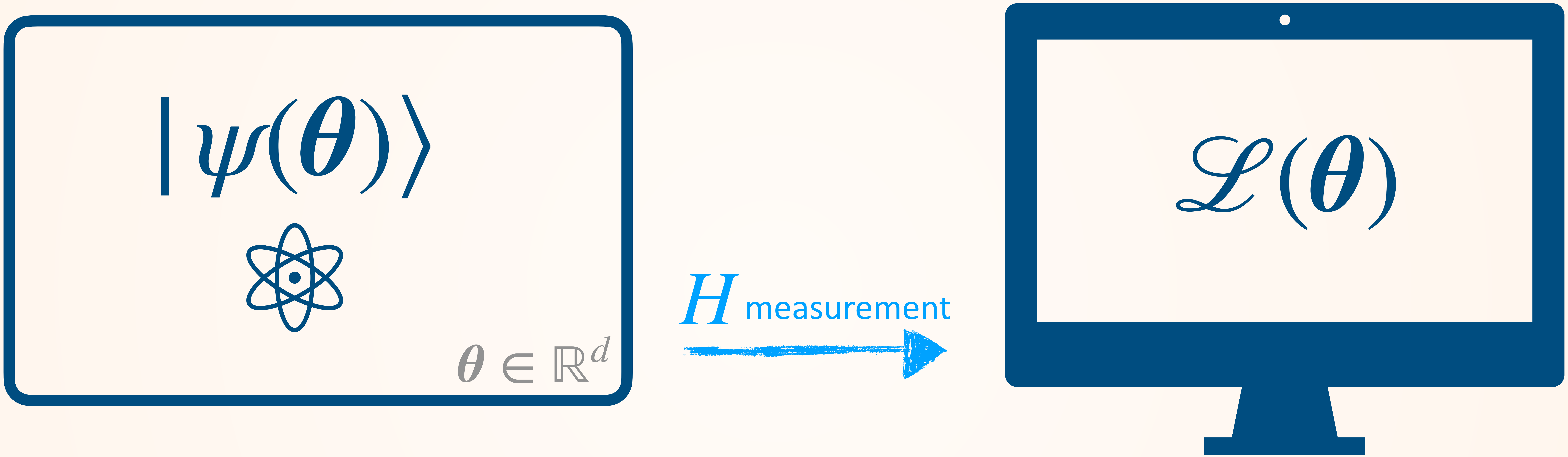
$$|\psi(\boldsymbol{\theta})\rangle$$

$\boldsymbol{\theta} \in \mathbb{R}^d$

H measurement

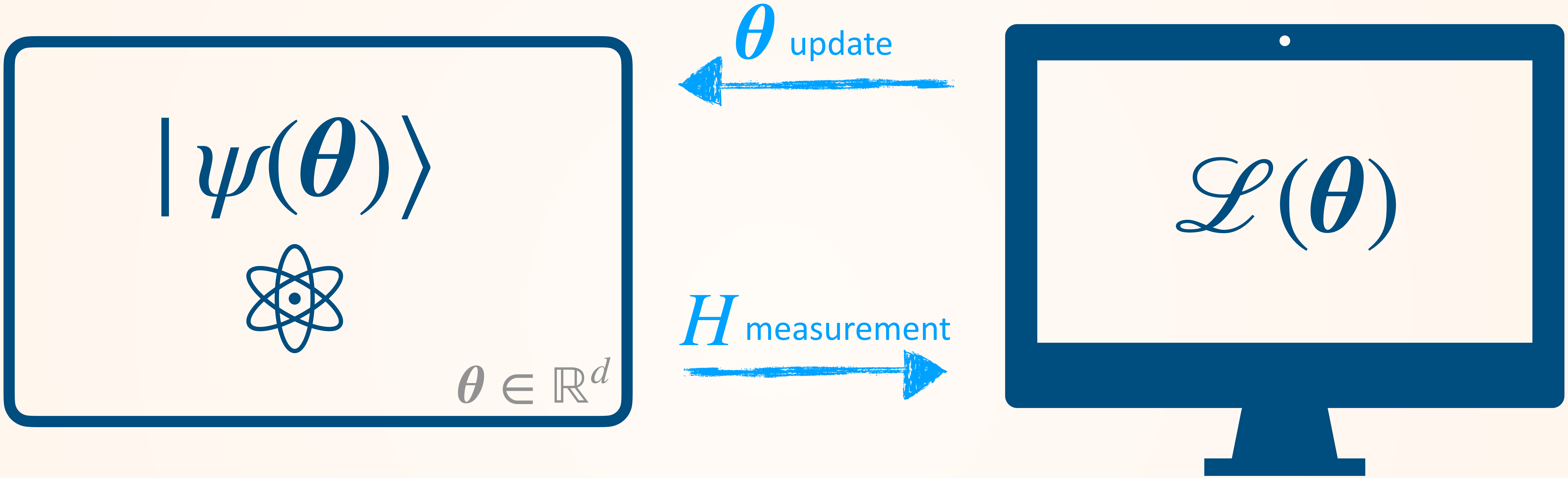
[1] A. Peruzzo et al., Nature Com. 2014 [2] Farhi et al., 1411.4028

Hybrid quantum-classical algorithms



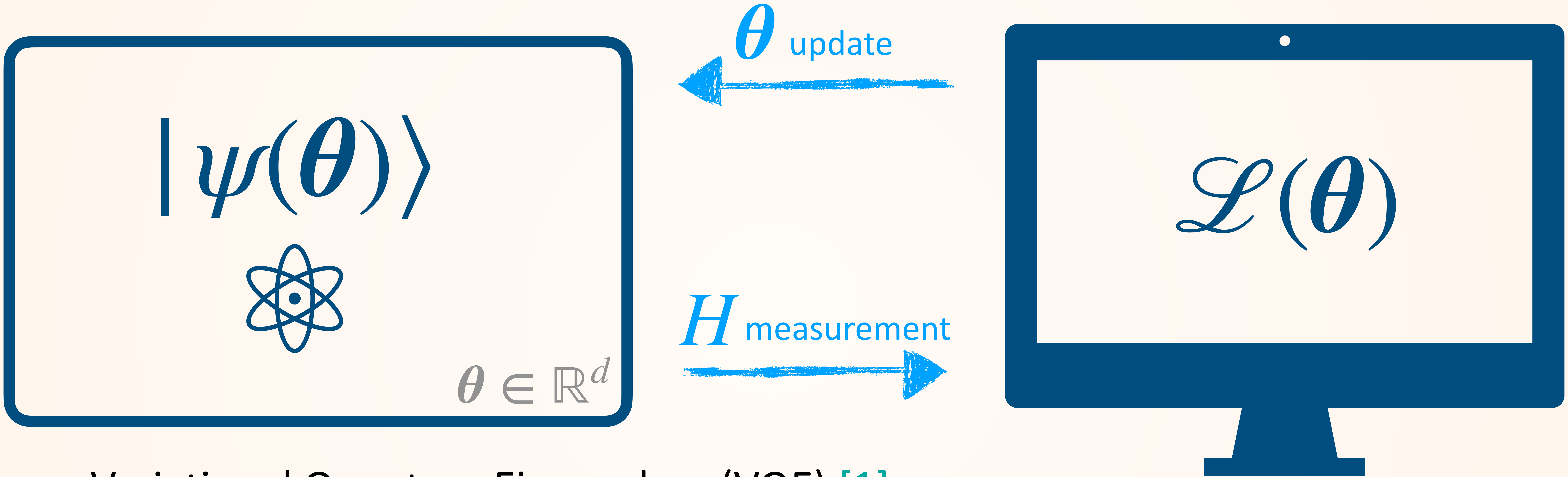
[1] A. Peruzzo et al., Nature Com. 2014 [2] Farhi et al., 1411.4028

Hybrid quantum-classical algorithms



[1] A. Peruzzo et al., Nature Com. 2014 [2] Farhi et al., 1411.4028

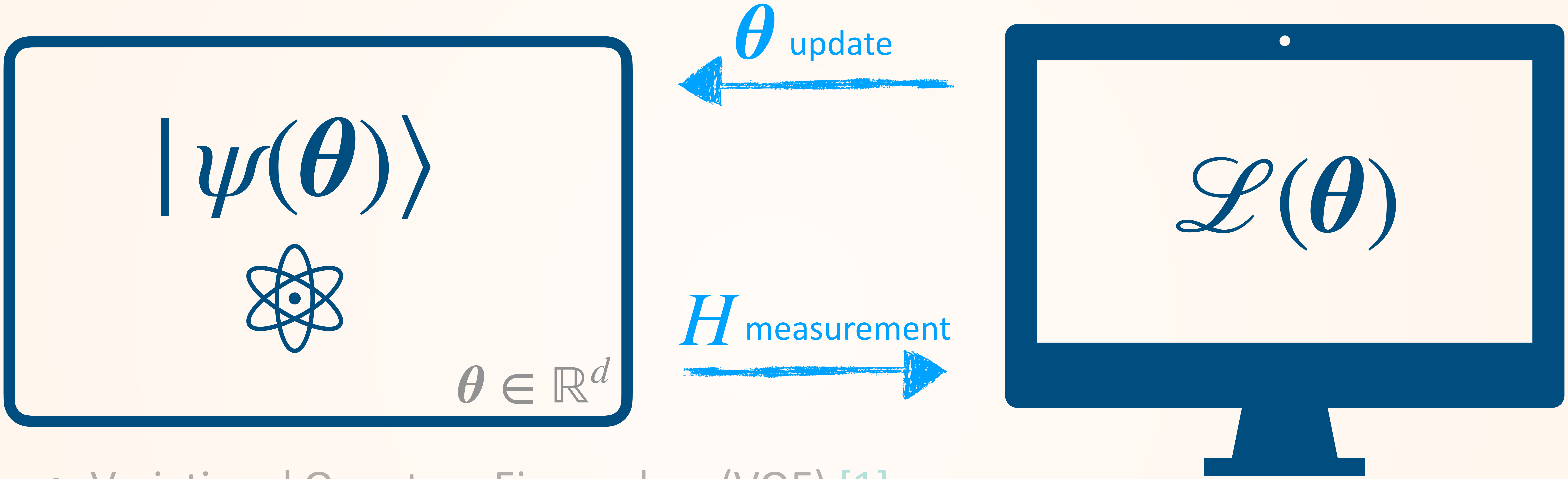
Hybrid quantum-classical algorithms



- Variational Quantum Eigensolver (VQE) [1]
  $\mathscr{L}(\boldsymbol{\theta}) = \langle \psi(\boldsymbol{\theta}) | H | \psi(\boldsymbol{\theta}) \rangle$

[1] A. Peruzzo et al., Nature Com. 2014 [2] Farhi et al., 1411.4028

Hybrid quantum-classical algorithms



- Variational Quantum Eigensolver (VQE) [1]

  $\mathscr{L}(\boldsymbol{\theta}) = \langle \psi(\boldsymbol{\theta}) | H | \psi(\boldsymbol{\theta}) \rangle$

- Quantum Approximate Optimization Algorithm (QAOA) [2]

  $|\psi(\boldsymbol{\beta},\boldsymbol{\gamma})\rangle = \mathrm{e}^{-\mathrm{i}\beta_p X}\,\mathrm{e}^{-\mathrm{i}\gamma_p H} \dots \mathrm{e}^{-\mathrm{i}\beta_1 X}\,\mathrm{e}^{-\mathrm{i}\gamma_1 H}\, |+\rangle$ $\qquad X = -\sum_i \sigma_i^x$

[1] A. Peruzzo et al., Nature Com. 2014 [2] Farhi et al., 1411.4028

Hybrid methods are promising for NISQ devices

[1] L. Cincio et al., 2007.01210
Review of NISQ algorithms: K. Barthi et al., 2101.08448
Review of variational algorithms: M. Cerezo et al., 2012.09265

Hybrid methods are promising for NISQ devices

- Ideally: Only do the quantum-easy-classically-hard part on the quantum device

[1] L. Cincio et al., 2007.01210
Review of NISQ algorithms: K. Barthi et al., 2101.08448
Review of variational algorithms: M. Cerezo et al., 2012.09265

Hybrid methods are promising for NISQ devices

- Ideally: Only do the quantum-easy-classically-hard part on the quantum device

- Variational principle has a long history

[1] L. Cincio et al., 2007.01210
Review of NISQ algorithms: K. Barthi et al., 2101.08448
Review of variational algorithms: M. Cerezo et al., 2012.09265

Hybrid methods are promising for NISQ devices

- Ideally: Only do the quantum-easy-classically-hard part on the quantum device
- Variational principle has a long history
- Short coherence times are OK due to iterative process

[1] L. Cincio et al., 2007.01210
Review of NISQ algorithms: K. Barthi et al., 2101.08448
Review of variational algorithms: M. Cerezo et al., 2012.09265

Hybrid methods are promising for NISQ devices

- Ideally: Only do the quantum-easy-classically-hard part on the quantum device
- Variational principle has a long history
- Short coherence times are OK due to iterative process
- (Implicit) error mitigation by the classical optimizer [1]

[1] L. Cincio et al., 2007.01210
Review of NISQ algorithms: K. Barthi et al., 2101.08448
Review of variational algorithms: M. Cerezo et al., 2012.09265

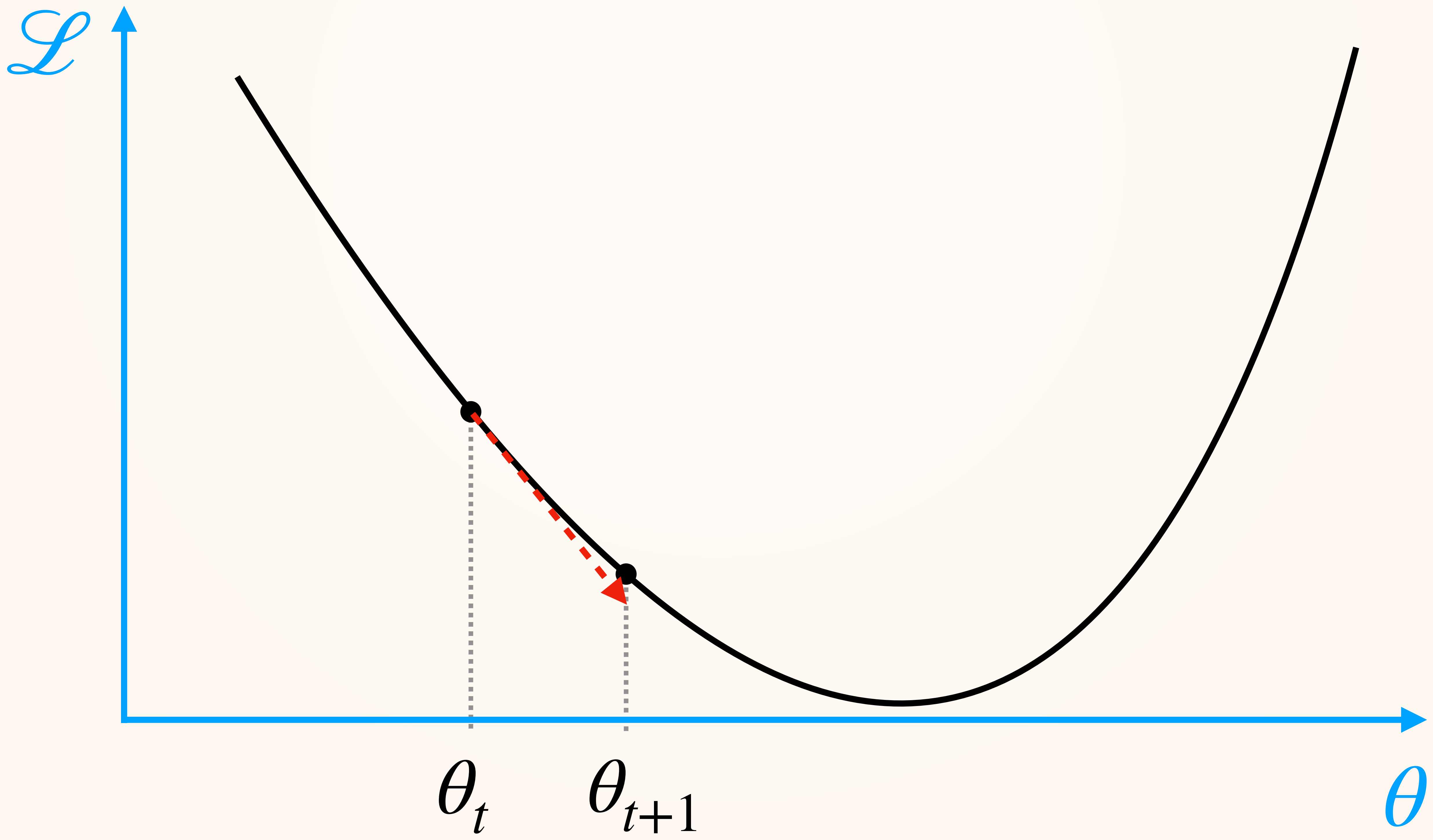
Zeroth-order Optimization

- Nelder-Mead method
- SPSA (simultaneous pertubation stochastic approximation) and RSGF combined with ADAM
- swarm optimization
- genetic algorithm
- scikit-quant.org [2]

Review: M. Benedetti et al., Quantum Sci. Technol. 2019
[2] W. Lavrijsen et al., 2004.03004

First-order Optimization

Gradient descent method: $\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathscr{L} \Big|_{\theta_t}$



First-order Optimization

[1] G. G. Guerreschi et al., 1701.01450 [3] L. Banchi et al., 2005.10299
[2] M. Schuld et al., PRA 2019

First-order Optimization

- gradient is expensive (no-cloning thm, no autograd)

[1] G. G. Guerreschi et al., 1701.01450   [3] L. Banchi et al., 2005.10299
[2] M. Schuld et al., PRA 2019

First-order Optimization

- gradient is expensive (no-cloning thm, no autograd)
- "exact" gradient is accessible

[1] G. G. Guerreschi et al., 1701.01450 [3] L. Banchi et al., 2005.10299
[2] M. Schuld et al., PRA 2019

First-order Optimization

- gradient is expensive (no-cloning thm, no autograd)
- "exact" gradient is accessible
- gradient as a circuit [1] $V e^{-i\theta X} \tilde{W} = VW, \quad |\psi\rangle = VW|0\rangle$

[1] G. G. Guerreschi et al., 1701.01450 [3] L. Banchi et al., 2005.10299
[2] M. Schuld et al., PRA 2019

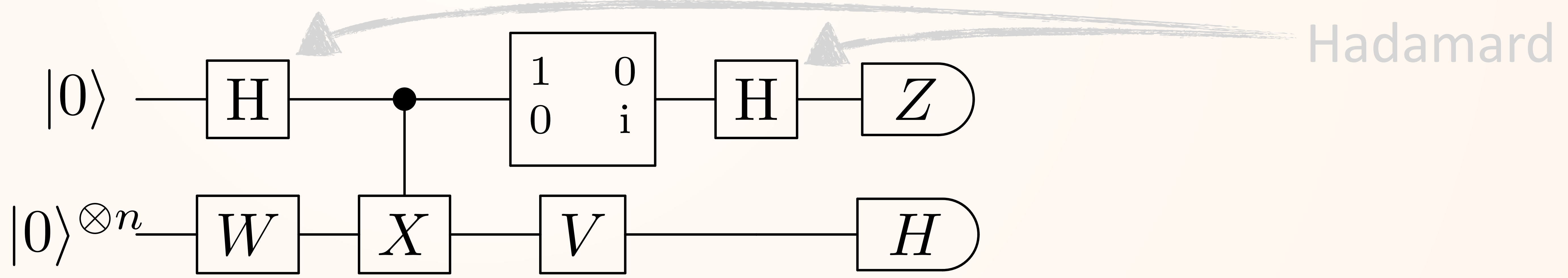
First-order Optimization

- gradient is expensive (no-cloning thm, no autograd)
- "exact" gradient is accessible
- gradient as a circuit [1] $Ve^{-i\theta X}\tilde{W} = VW, \quad |\psi\rangle = VW|0\rangle$
- $\partial_\theta \langle\psi|H|\psi\rangle \;=\; 2\,\Im\langle\psi|H\,VXW|0\rangle \;=\; 2\,\Im\langle\psi|H\,VXV^\dagger|\psi\rangle$

[1] G. G. Guerreschi et al., 1701.01450 [3] L. Banchi et al., 2005.10299
[2] M. Schuld et al., PRA 2019

First-order Optimization

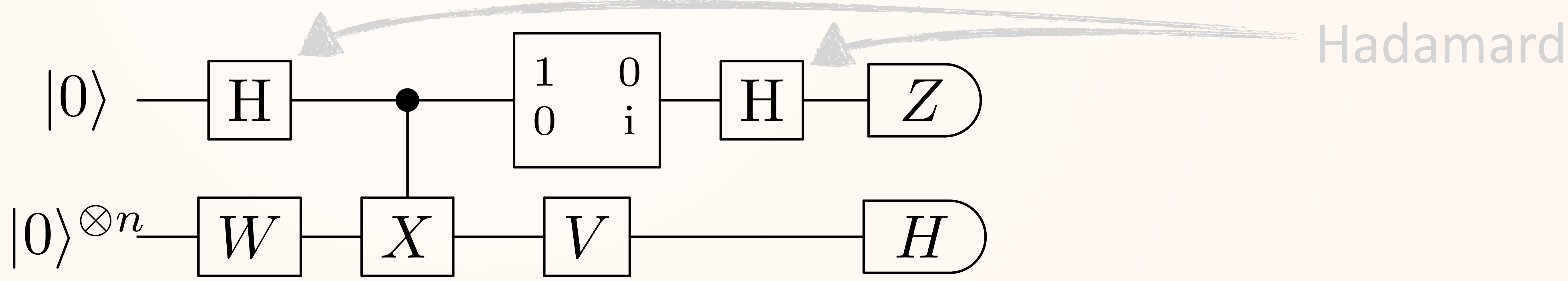
- gradient is expensive (no-cloning thm, no autograd)
- "exact" gradient is accessible
- gradient as a circuit [1] $Ve^{-i\theta X}\tilde{W} = VW, \quad |\psi\rangle = VW|0\rangle$
- $\partial_\theta \langle\psi|H|\psi\rangle = 2\,\Im\langle\psi|H\,VXW|0\rangle = 2\,\Im\langle\psi|H\,VXV^\dagger|\psi\rangle$



[1] G. G. Guerreschi et al., 1701.01450 [3] L. Banchi et al., 2005.10299
[2] M. Schuld et al., PRA 2019

First-order Optimization

- gradient is expensive (no-cloning thm, no autograd)
- "exact" gradient is accessible
- gradient as a circuit [1] $Ve^{-i\theta X}\tilde{W} = VW, \quad |\psi\rangle = VW|0\rangle$
- $\partial_\theta \langle\psi|H|\psi\rangle = 2\,\mathfrak{I}\langle\psi|H\,VXW|0\rangle = 2\,\mathfrak{I}\langle\psi|H\,VXV^\dagger|\psi\rangle$

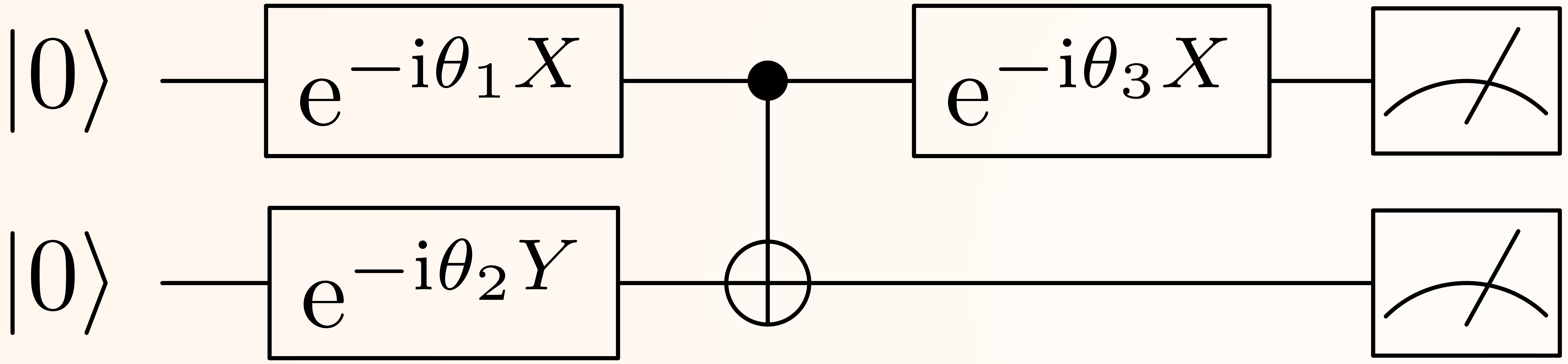


- Parameter-shift rule [2,3] $\partial_\theta \langle H\rangle = \langle H\rangle_{\theta+\pi/4} - \langle H\rangle_{\theta-\pi/4}$

[1] G. G. Guerreschi et al., 1701.01450 [3] L. Banchi et al., 2005.10299
[2] M. Schuld et al., PRA 2019

Parameter Shift Rule

$$\partial_\theta \langle H \rangle = \langle H \rangle_{\theta+\pi/4} - \langle H \rangle_{\theta-\pi/4}$$



Sample-estimator for $\langle H \rangle$

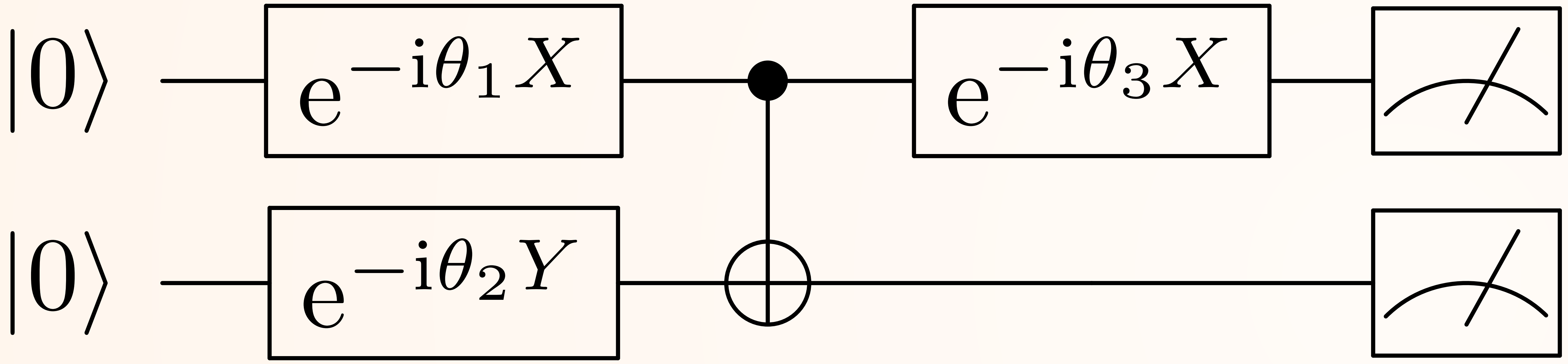
Generator can only have two eigenvalues (here +1 and -1)

Generalizations for larger gates:

[L. Banchi et al., 2005.10299]
[D. Wierichs et al., 2107.12390]

Parameter Shift Rule

$$\partial_\theta \langle H \rangle = \langle H \rangle_{\theta+\pi/4} - \langle H \rangle_{\theta-\pi/4}$$



Sample-estimator for $\langle H \rangle$

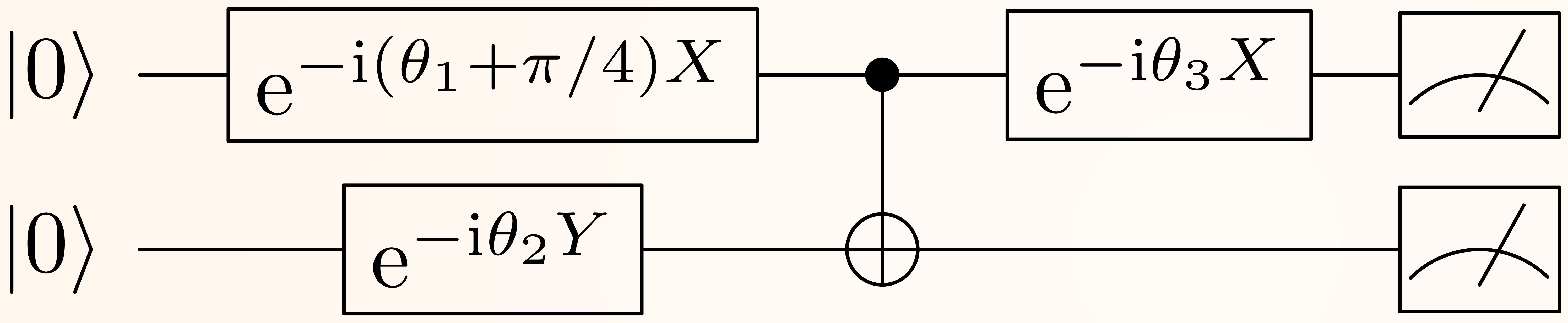
$$\nabla_{\boldsymbol{\theta}} \langle H \rangle = \left(\quad \right)$$

Generator can only have two eigenvalues (here +1 and -1)

Generalizations for larger gates:
[L. Banchi et al., 2005.10299]
[D. Wierichs et al., 2107.12390]

Parameter Shift Rule

$$\partial_\theta \langle H \rangle = \langle H \rangle_{\theta+\pi/4} - \langle H \rangle_{\theta-\pi/4}$$



Sample-estimator for $\langle H \rangle$

$$\nabla_{\boldsymbol{\theta}} \langle H \rangle = \left(\quad \right)$$

Generator can only have two eigenvalues (here +1 and -1)

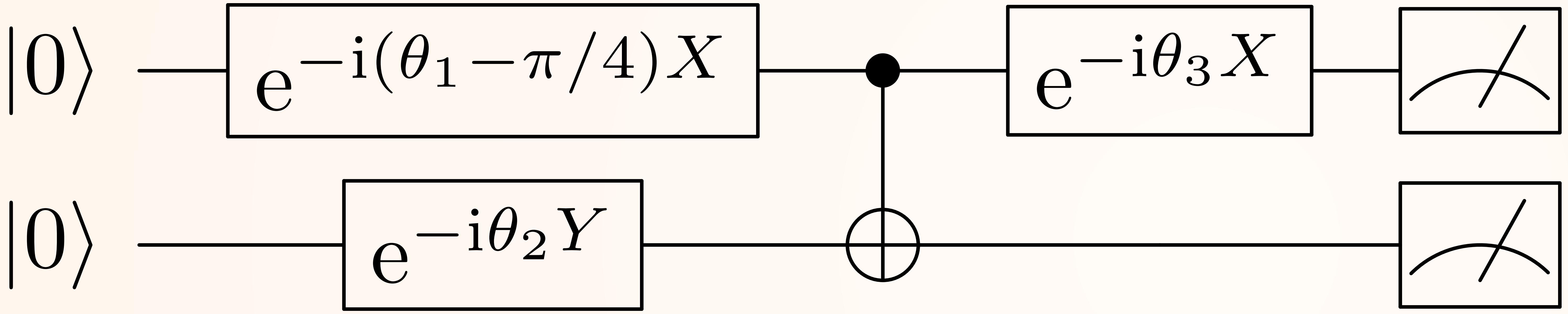
Generalizations for larger gates:

[L. Banchi et al., 2005.10299]

[D. Wierichs et al., 2107.12390]

Parameter Shift Rule

$$\partial_\theta \langle H \rangle = \langle H \rangle_{\theta+\pi/4} - \langle H \rangle_{\theta-\pi/4}$$



Sample-estimator for $\langle H \rangle$

$$\nabla_{\boldsymbol{\theta}} \langle H \rangle = \left(\quad \right)$$

Generator can only have two eigenvalues (here +1 and -1)

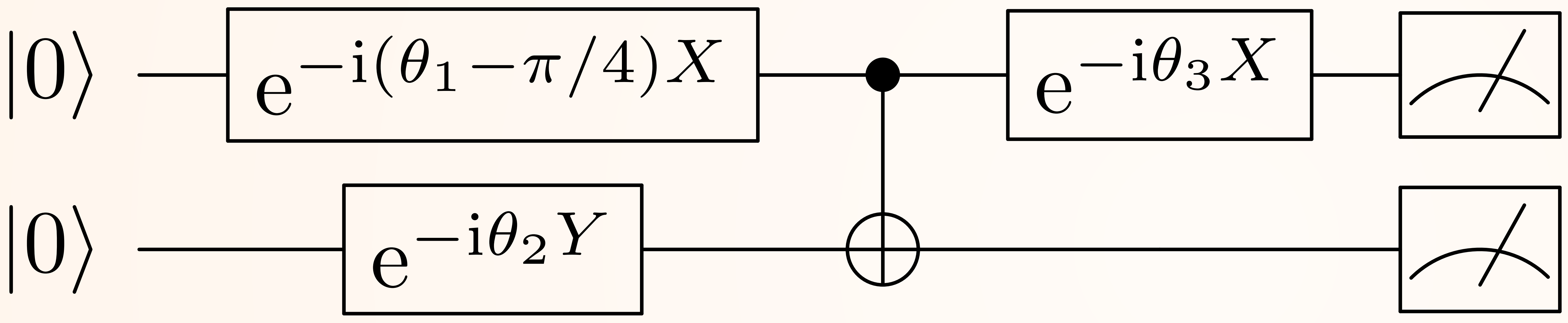
Generalizations for larger gates:

[L. Banchi et al., 2005.10299]

[D. Wierichs et al., 2107.12390]

Parameter Shift Rule

$$\partial_\theta \langle H \rangle = \langle H \rangle_{\theta+\pi/4} - \langle H \rangle_{\theta-\pi/4}$$



Sample-estimator for $\langle H \rangle$

$$\nabla_{\boldsymbol{\theta}} \langle H \rangle = \begin{pmatrix} \langle H \rangle_{\theta_1+\pi/4} - \langle H \rangle_{\theta_1-\pi/4} \\ \\ \\ \end{pmatrix}$$

Generator can only have two eigenvalues (here +1 and -1)

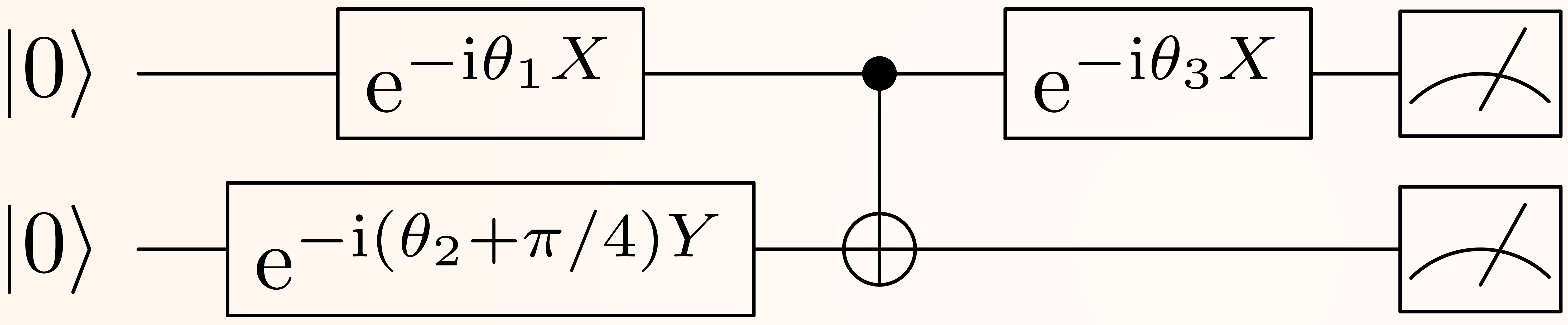
Generalizations for larger gates:

[L. Banchi et al., 2005.10299]

[D. Wierichs et al., 2107.12390]

Parameter Shift Rule

$$\partial_\theta \langle H \rangle = \langle H \rangle_{\theta+\pi/4} - \langle H \rangle_{\theta-\pi/4}$$



Sample-estimator for $\langle H \rangle$

$$\nabla_{\boldsymbol{\theta}} \langle H \rangle = \begin{pmatrix} \langle H \rangle_{\theta_1+\pi/4} - \langle H \rangle_{\theta_1-\pi/4} \\ \\ \\ \end{pmatrix}$$

Generator can only have two eigenvalues (here +1 and -1)

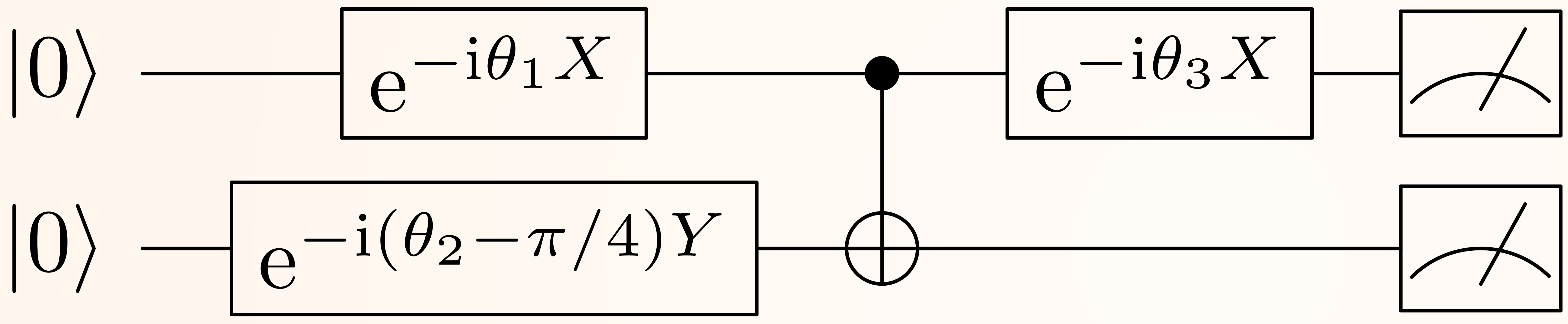
Generalizations for larger gates:

[L. Banchi et al., 2005.10299]

[D. Wierichs et al., 2107.12390]

Parameter Shift Rule

$$\partial_\theta \langle H \rangle = \langle H \rangle_{\theta+\pi/4} - \langle H \rangle_{\theta-\pi/4}$$



Sample-estimator for $\langle H \rangle$

$$\nabla_{\boldsymbol{\theta}} \langle H \rangle = \begin{pmatrix} \langle H \rangle_{\theta_1+\pi/4} - \langle H \rangle_{\theta_1-\pi/4} \\ \\ \\ \end{pmatrix}$$

Generator can only have two eigenvalues (here +1 and -1)

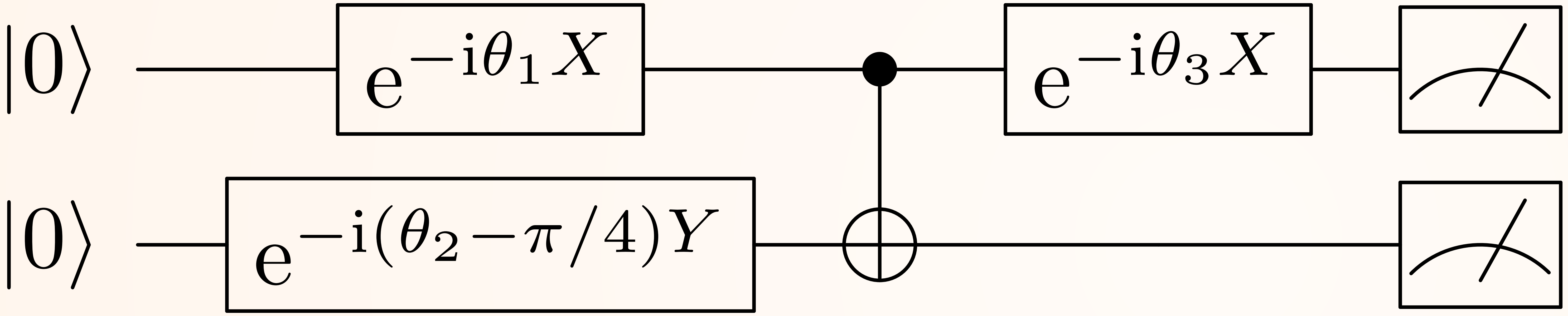
Generalizations for larger gates:

[L. Banchi et al., 2005.10299]

[D. Wierichs et al., 2107.12390]

Parameter Shift Rule

$$\partial_\theta \langle H \rangle = \langle H \rangle_{\theta+\pi/4} - \langle H \rangle_{\theta-\pi/4}$$



Sample-estimator for $\langle H \rangle$

$$\nabla_{\boldsymbol{\theta}} \langle H \rangle = \begin{pmatrix} \langle H \rangle_{\theta_1+\pi/4} - \langle H \rangle_{\theta_1-\pi/4} \\ \langle H \rangle_{\theta_2+\pi/4} - \langle H \rangle_{\theta_2-\pi/4} \\ \\ \end{pmatrix}$$

Generator can only have two eigenvalues (here +1 and -1)

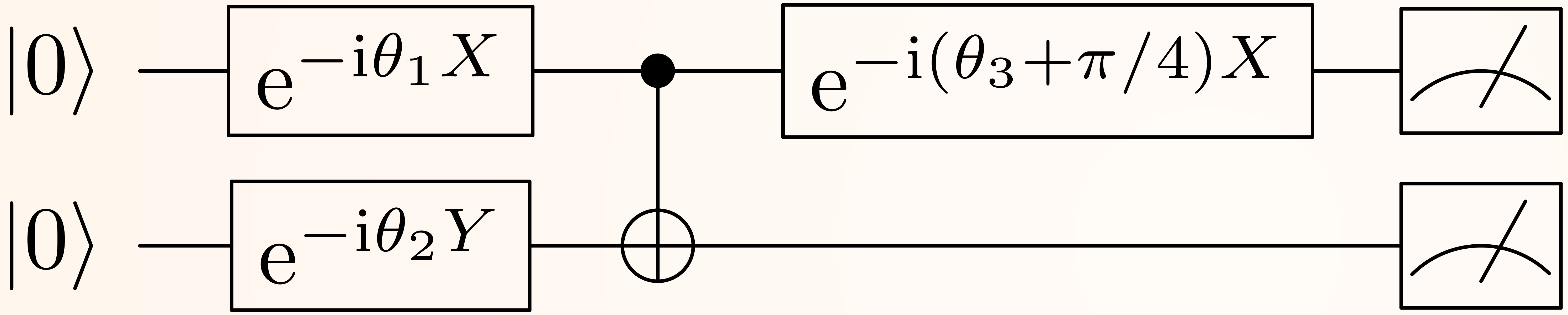
Generalizations for larger gates:

[L. Banchi et al., 2005.10299]

[D. Wierichs et al., 2107.12390]

Parameter Shift Rule

$$\partial_\theta \langle H \rangle = \langle H \rangle_{\theta+\pi/4} - \langle H \rangle_{\theta-\pi/4}$$



Sample-estimator for $\langle H \rangle$

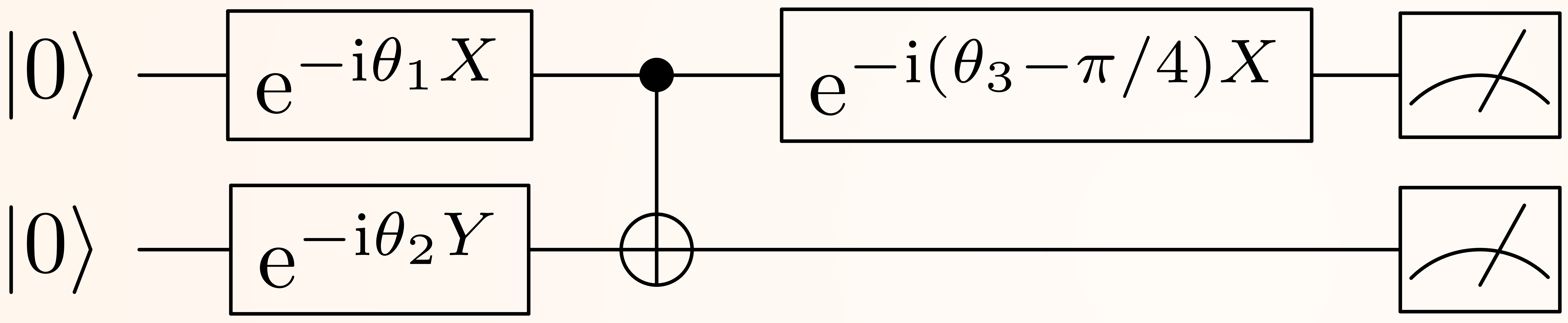
$$\nabla_{\boldsymbol{\theta}} \langle H \rangle = \begin{pmatrix} \langle H \rangle_{\theta_1+\pi/4} - \langle H \rangle_{\theta_1-\pi/4} \\ \langle H \rangle_{\theta_2+\pi/4} - \langle H \rangle_{\theta_2-\pi/4} \\ \\ \end{pmatrix}$$

Generator can only have two eigenvalues (here +1 and -1)

Generalizations for larger gates:
[L. Banchi et al., 2005.10299]
[D. Wierichs et al., 2107.12390]

Parameter Shift Rule

$$\partial_\theta \langle H \rangle = \langle H \rangle_{\theta+\pi/4} - \langle H \rangle_{\theta-\pi/4}$$



Sample-estimator for $\langle H \rangle$

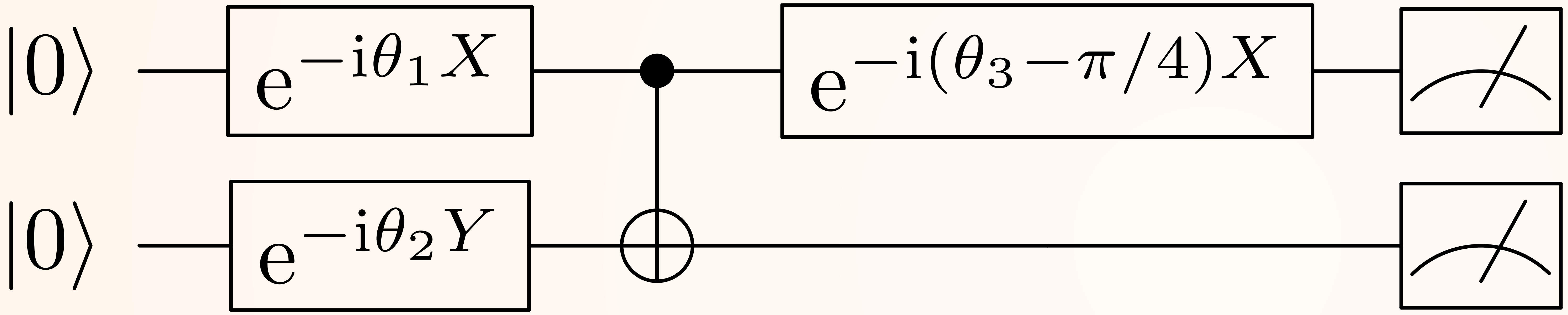
$$\nabla_{\boldsymbol{\theta}} \langle H \rangle = \begin{pmatrix} \langle H \rangle_{\theta_1+\pi/4} - \langle H \rangle_{\theta_1-\pi/4} \\ \langle H \rangle_{\theta_2+\pi/4} - \langle H \rangle_{\theta_2-\pi/4} \\ \\ \end{pmatrix}$$

Generator can only have two eigenvalues (here +1 and -1)

Generalizations for larger gates:
[L. Banchi et al., 2005.10299]
[D. Wierichs et al., 2107.12390]

Parameter Shift Rule

$$\partial_\theta \langle H \rangle = \langle H \rangle_{\theta+\pi/4} - \langle H \rangle_{\theta-\pi/4}$$



Sample-estimator for $\langle H \rangle$

$$\nabla_{\boldsymbol{\theta}} \langle H \rangle = \begin{pmatrix} \langle H \rangle_{\theta_1+\pi/4} - \langle H \rangle_{\theta_1-\pi/4} \\ \langle H \rangle_{\theta_2+\pi/4} - \langle H \rangle_{\theta_2-\pi/4} \\ \langle H \rangle_{\theta_3+\pi/4} - \langle H \rangle_{\theta_3-\pi/4} \end{pmatrix}$$

Generator can only have two eigenvalues (here +1 and -1)

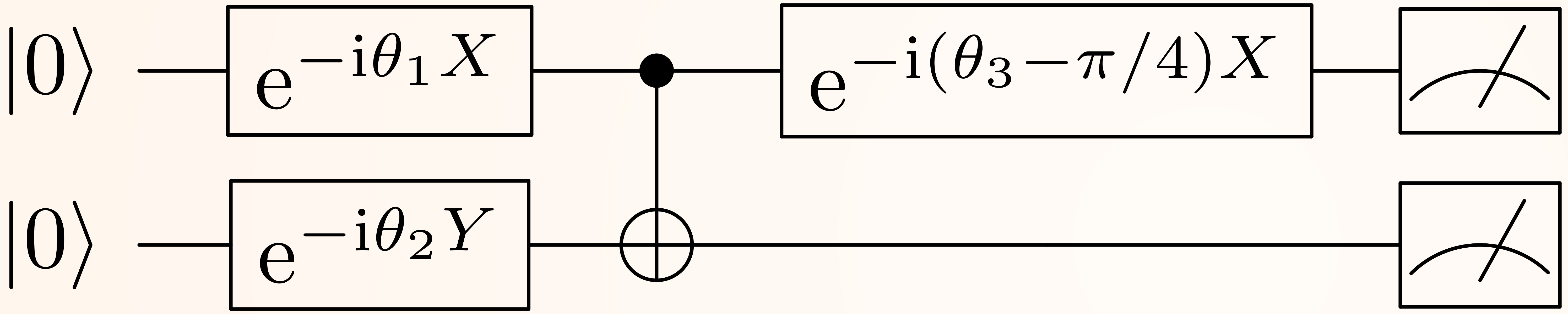
Generalizations for larger gates:

[L. Banchi et al., 2005.10299]

[D. Wierichs et al., 2107.12390]

Parameter Shift Rule

$$\partial_\theta \langle H \rangle = \langle H \rangle_{\theta+\pi/4} - \langle H \rangle_{\theta-\pi/4}$$



Sample-estimator for $\langle H \rangle$

$$\nabla_{\boldsymbol{\theta}} \langle H \rangle = \begin{pmatrix} \langle H \rangle_{\theta_1+\pi/4} - \langle H \rangle_{\theta_1-\pi/4} \\ \langle H \rangle_{\theta_2+\pi/4} - \langle H \rangle_{\theta_2-\pi/4} \\ \langle H \rangle_{\theta_3+\pi/4} - \langle H \rangle_{\theta_3-\pi/4} \end{pmatrix}$$

Generator can only have two eigenvalues (here +1 and -1)

Generalizations for larger gates:
[L. Banchi et al., 2005.10299]
[D. Wierichs et al., 2107.12390]

A single-shift rule is impossible!
[T. Hubregtsen et al., 2106.01388]

Sources of stochasticity

$$\mathscr{L}(\theta) = \langle \; H \; \rangle_\theta$$

1. Measurement shots

Sources of stochasticity

$$\mathscr{L}(\theta) = \left\langle \sum_j h_j \right\rangle_\theta$$

1. Measurement shots
2. Observable components

Sources of stochasticity

$$\mathscr{L}(\theta) = \sum_i \left[\left\langle \sum_j h_j \right\rangle_{(x_i,\, \theta)} - y_i \right]^2$$

$x_i \in \mathbb{R}^{3N}$	$y_i \in \{0,1\}$
	0
	1
	0

1. Measurement shots
2. Observable components
3. Data

Sources of stochasticity

$$\mathcal{L}(\theta) = \sum_i \left[\left\langle \sum_j h_j \right\rangle_{(x_i, \theta)} - y_i \right]^2$$

1. Measurement shots
2. Observable components
3. Data
4. Parameter-shift terms

Sources of stochasticity

$$\mathscr{L}(\theta) = \sum_i \left[\left\langle \sum_j h_j \right\rangle_{(x_i,\, \theta)} - y_i \right]^2$$

hardware-dependent
(shots might actually be cheap)

1. Measurement shots
2. Observable components
3. Data
4. Parameter-shift terms

Sources of stochasticity

$$\mathscr{L}(\theta) = \sum_i \left[\left\langle \sum_j h_j \right\rangle_{(x_i,\, \theta)} - y_i \right]^2$$

1. Measurement shots ← hardware-dependent (shots might actually be cheap)
2. Observable components ← non-commuting components
3. Data
4. Parameter-shift terms

Sources of stochasticity

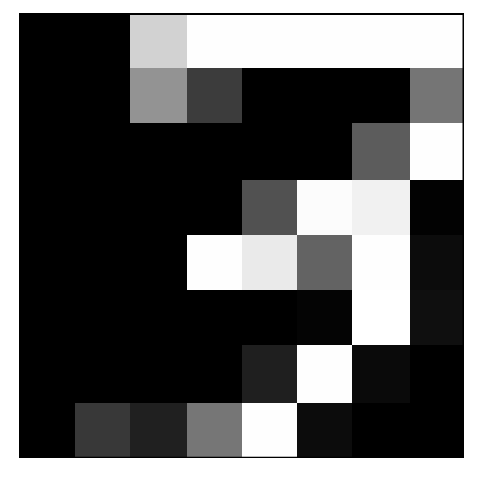
$$\mathscr{L}(\theta) = \sum_i \left[\left\langle \sum_j h_j \right\rangle_{(x_i, \theta)} - y_i \right]^2$$

1. Measurement shots ← hardware-dependent (shots might actually be cheap)
2. Observable components ← non-commuting components
3. Data
4. Parameter-shift terms ← dependent on circuit architecture (at least factor of 2)

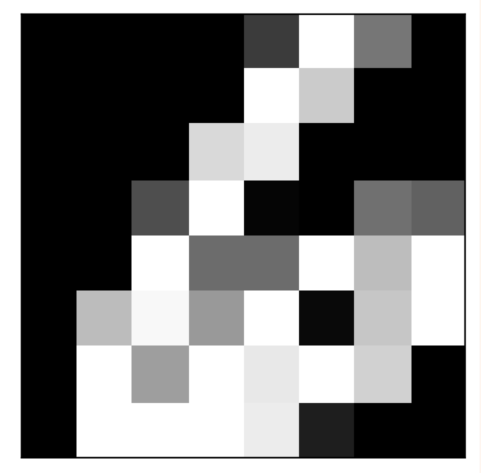
Sources of stochasticity

$$\mathscr{L}(\theta) = \sum_i \left[\left\langle \sum_j h_j \right\rangle_{(x_i,\, \theta)} - y_i \right]^2$$

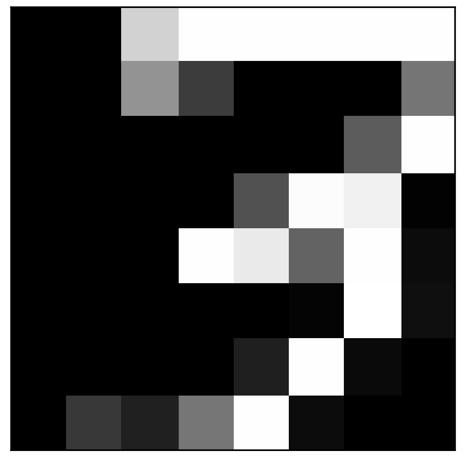
$$\partial_\theta \mathscr{L} = \sum_i 2\left[\sum_j \langle h_j \rangle_{(x_i,\, \theta)} - y_i \right] \sum_j \frac{1}{2}\left(\langle h_j \rangle_{(x_i,\, \theta + \frac{\pi}{2})} - \langle h_j \rangle_{(x_i,\, \theta - \frac{\pi}{2})} \right)$$



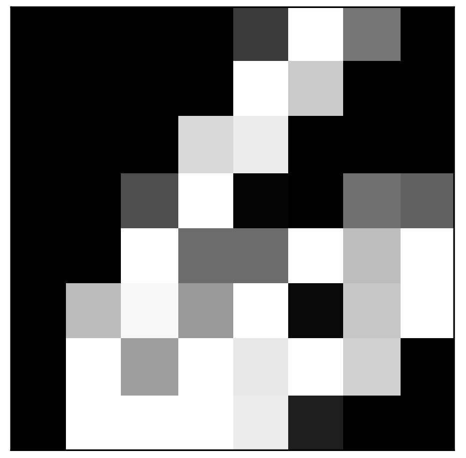
MNIST Classifier

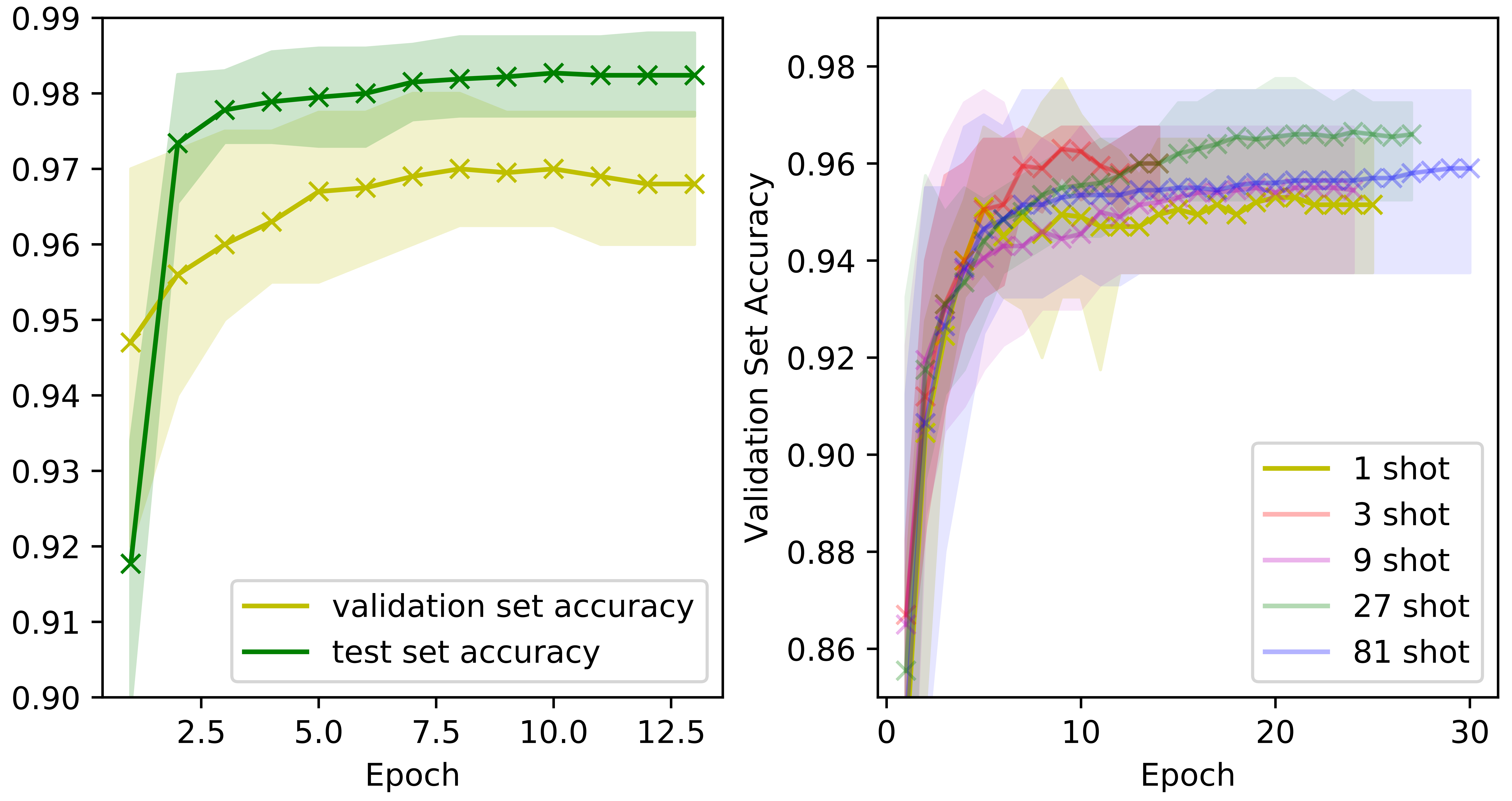


8 qubits, 400 parameters, batch size = 1



MNIST Classifier

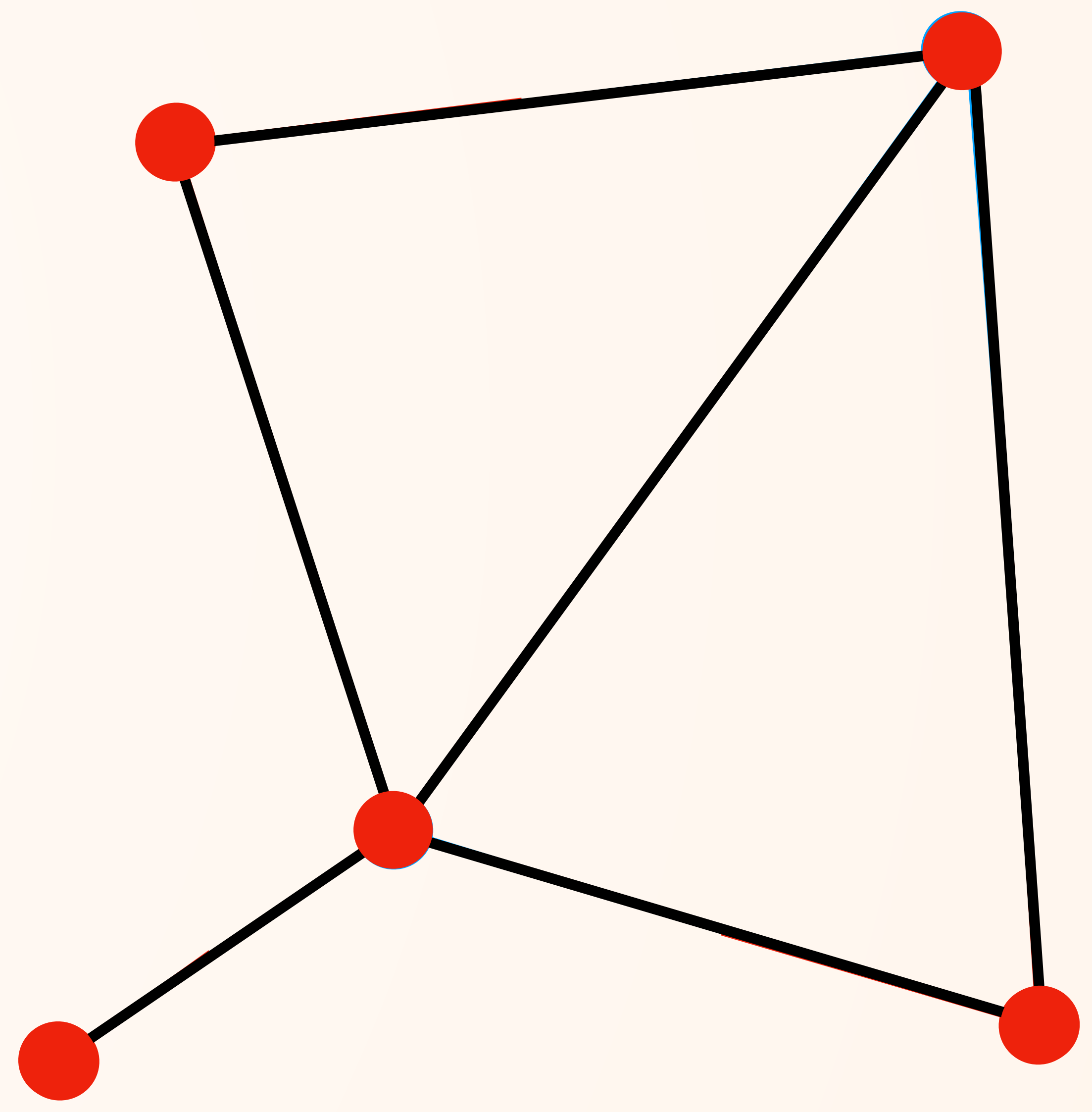




8 qubits, 400 parameters, batch size = 1

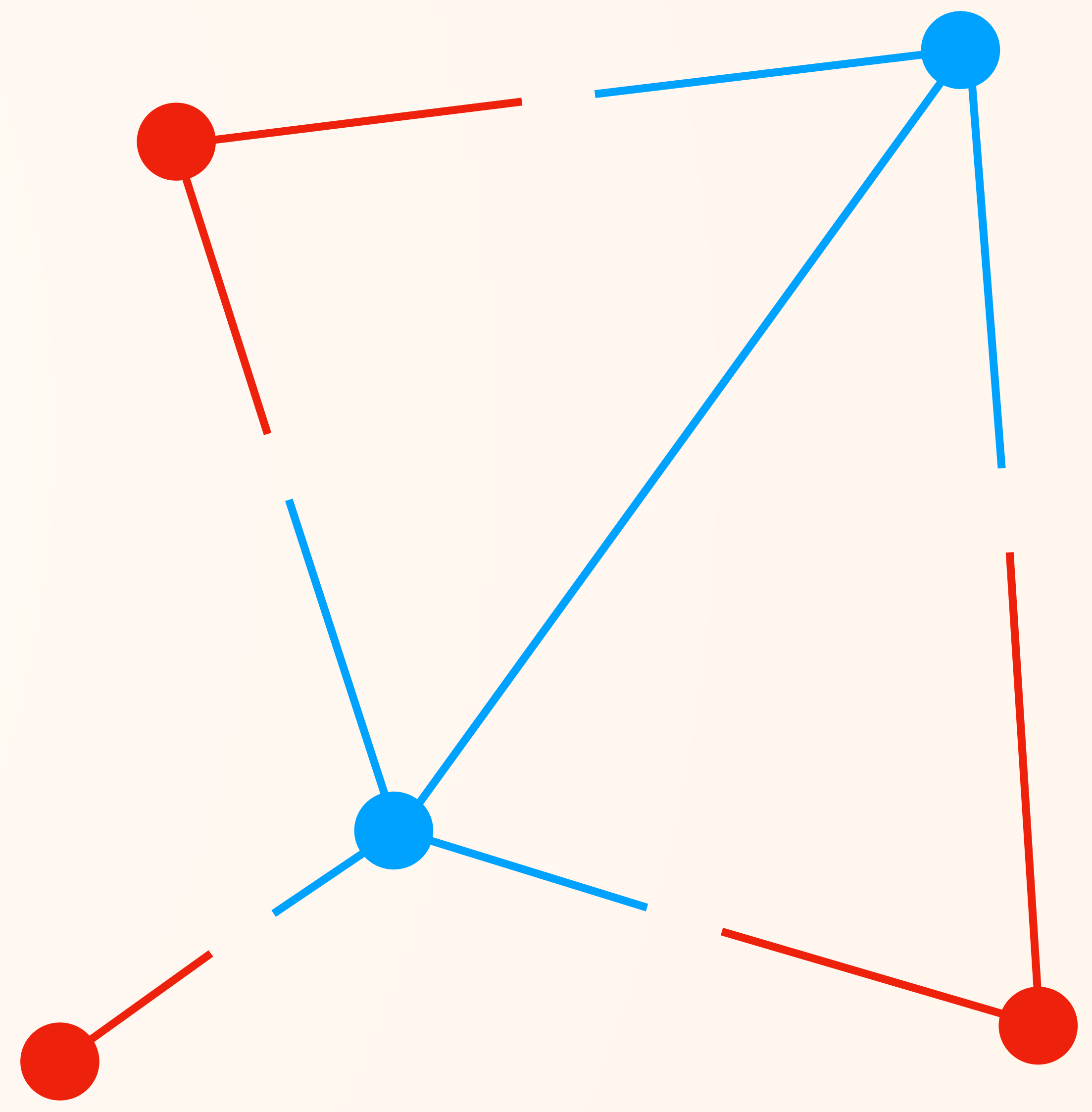
MAXCUT with QAOA

- $G = (V, E)$ Problem: Divide V into two subsets, s.t. the number of edges between them is maximal. $\leftarrow$ NP-hard



MAXCUT with QAOA

- $G = (V, E)$ Problem: Divide V into two subsets, s.t. the number of edges between them is maximal. $\leftarrow$ NP-hard



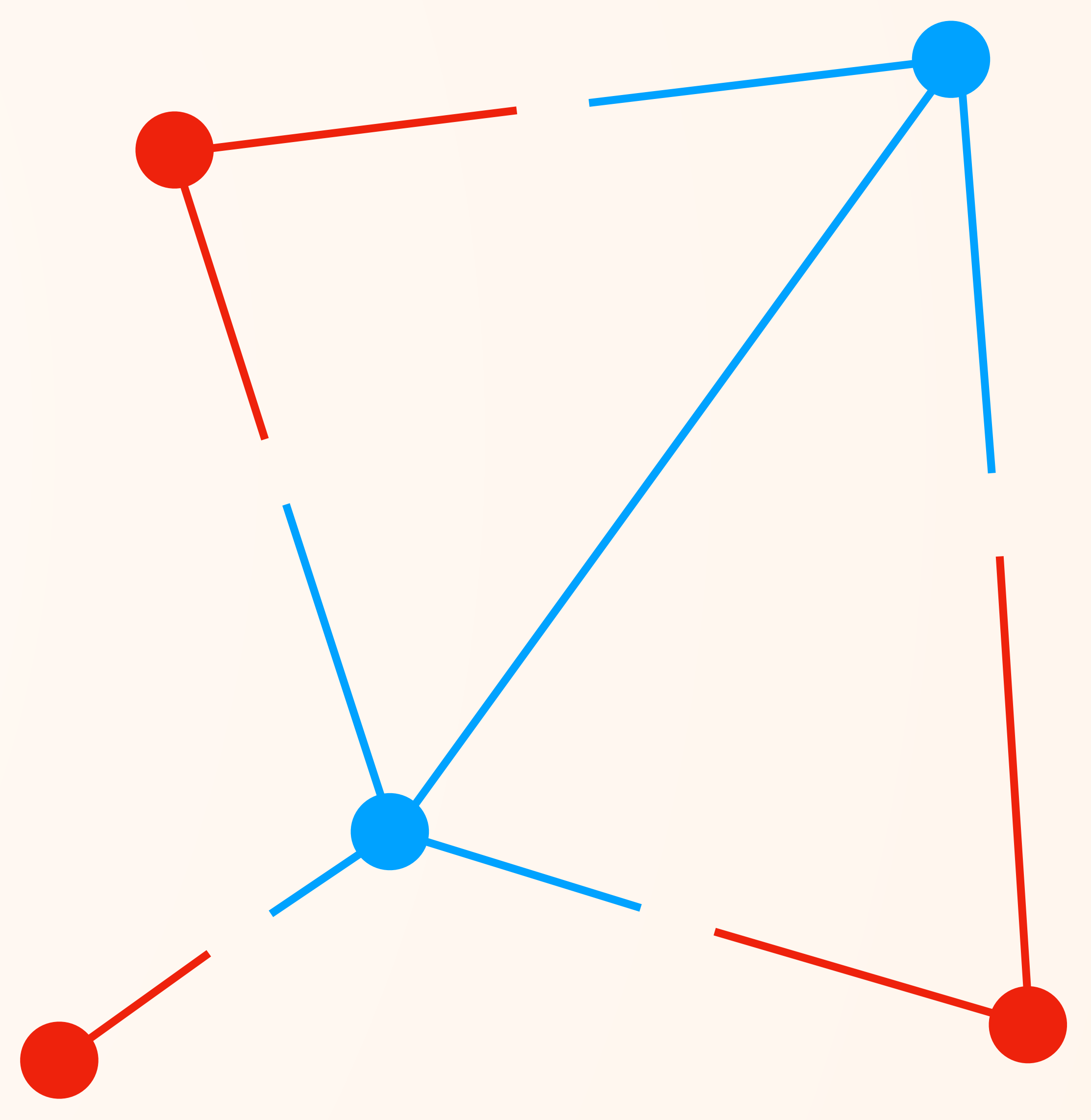
MAXCUT with QAOA

- $G = (V, E)$ Problem: Divide V into two subsets, s.t. the number of edges between them is maximal. $\leftarrow$ NP-hard

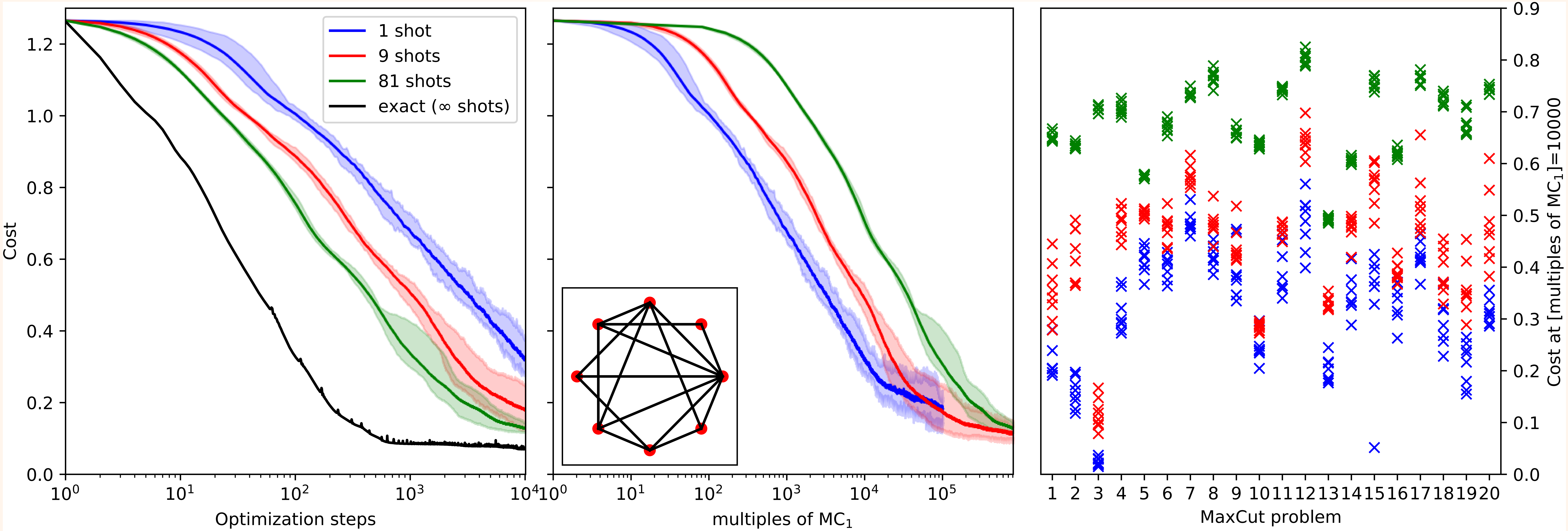
$$H = \sum_{(i,j) \in E} \sigma_i^z \sigma_j^z$$

- QAOA:

$$|\psi(\boldsymbol{\beta}, \boldsymbol{\gamma})\rangle = \mathrm{e}^{-\mathrm{i}\beta_p X} \mathrm{e}^{-\mathrm{i}\gamma_p H} \dots \mathrm{e}^{-\mathrm{i}\beta_1 X} \mathrm{e}^{-\mathrm{i}\gamma_1 H} \, |+\rangle$$

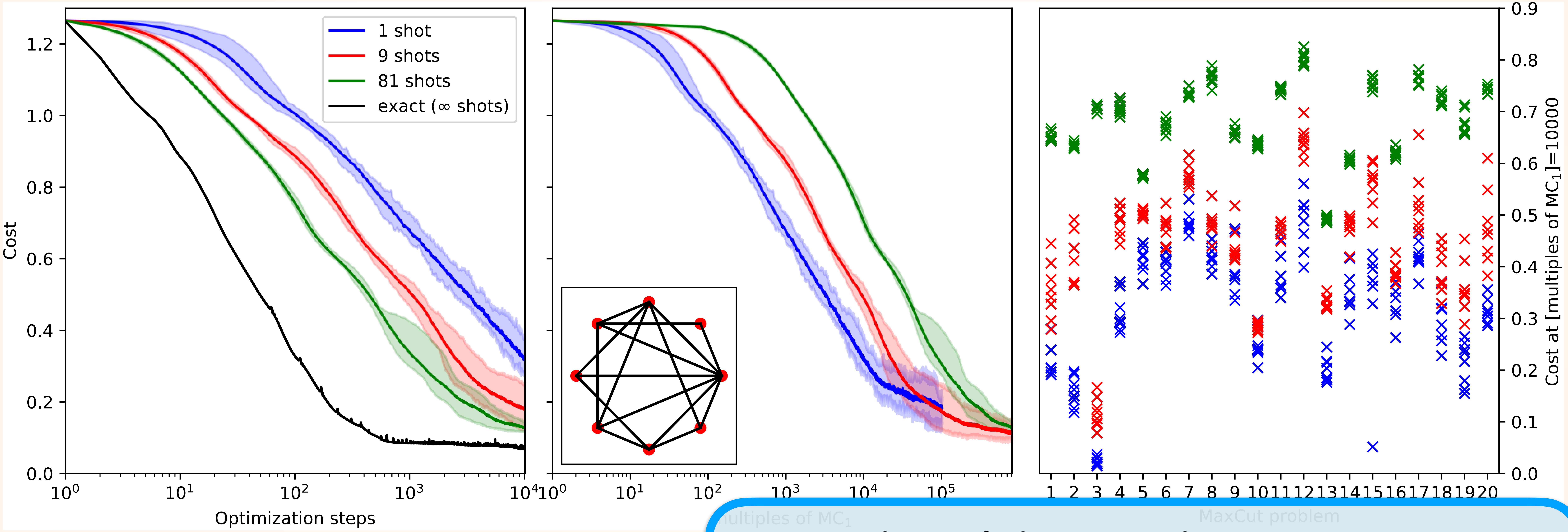


MAXCUT with QAOA



p=50, random graphs $|V| = 8, |E| = 16$

MAXCUT with QAOA



Number of shots is a hyper parameter.

p=50, random graphs $|V| = 8, |E| = 16$

Related work

[1] J. Napp et al., 1901.05374
[2] J. M. Kübler et al., Quantum 2020
[3] A. Arrasmith et al., 2004.06252
[4] B. van Straaten et al., PRX Quant. 2021

Related work

- Rigorous separation between 0th and 1st order optimization [1]

[1] J. Napp et al., 1901.05374
[2] J. M. Kübler et al., Quantum 2020
[3] A. Arrasmith et al., 2004.06252
[4] B. van Straaten et al., PRX Quant. 2021

Related work

- Rigorous separation between 0th and 1st order optimization [1]

- iCANS [2]
  individual Coupled Adaptive Number of Shots

$$\#(\text{shots}) = \frac{2L\alpha}{2 - L\alpha} \frac{\hat{\text{Var}}(g_i)}{g_i^2}$$

[1] J. Napp et al., 1901.05374
[2] J. M. Kübler et al., Quantum 2020
[3] A. Arrasmith et al., 2004.06252
[4] B. van Straaten et al., PRX Quant. 2021

Related work

- Rigorous separation between 0th and 1st order optimization [1]
- iCANS [2]
  individual Coupled Adaptive Number of Shots

$$\#(\text{shots}) = \frac{2L\alpha}{2 - L\alpha} \frac{\hat{\mathrm{Var}}(g_i)}{g_i^2}$$

- Rosalin [3]
  Cleverly distribute and adapt "shot budget" amongst Hamiltonian terms. → Weighted Random Sampling

[1] J. Napp et al., 1901.05374
[2] J. M. Kübler et al., Quantum 2020
[3] A. Arrasmith et al., 2004.06252
[4] B. van Straaten et al., PRX Quant. 2021

Related work

- Rigorous separation between 0th and 1st order optimization [1]
- iCANS [2]
  individual Coupled Adaptive Number of Shots

  $$\#(\text{shots}) = \frac{2L\alpha}{2 - L\alpha} \frac{\hat{\text{Var}}(g_i)}{g_i^2}$$

- Rosalin [3]
  Cleverly distribute and adapt "shot budget" amongst Hamiltonian terms. $\rightarrow$ Weighted Random Sampling
- Rigorous bounds on required number of shots [4]

[1] J. Napp et al., 1901.05374
[2] J. M. Kübler et al., Quantum 2020
[3] A. Arrasmith et al., 2004.06252
[4] B. van Straaten et al., PRX Quant. 2021

Convergence

[1] M. M. Wolf, “Mathematical foundations of supervised learning”

Convergence

Problem: $\mathbb{E}(\mathscr{L}(X)) \neq \mathscr{L}(\mathbb{E}(X))$

[1] M. M. Wolf, “Mathematical foundations of supervised learning”

Convergence

Problem: $\mathbb{E}(\mathscr{L}(X)) \neq \mathscr{L}(\mathbb{E}(X))$

- Solved for polynomial loss functions

[1] M. M. Wolf, "Mathematical foundations of supervised learning"

Convergence

Problem: $\mathbb{E}(\mathscr{L}(X)) \neq \mathscr{L}(\mathbb{E}(X))$

- Solved for polynomial loss functions

$$\mathscr{L}(X) = \sum_{j=0}^{k} a_j X^j = \mathscr{L}(X_1, \ldots, X_k)$$

[1] M. M. Wolf, "Mathematical foundations of supervised learning"

Convergence

Problem: $\mathbb{E}(\mathscr{L}(X)) \neq \mathscr{L}(\mathbb{E}(X))$

- Solved for polynomial loss functions

$$\mathscr{L}(X) = \sum_{j=0}^{k} a_j X^j = \mathscr{L}(X_1, \ldots, X_k)$$

$$e_k\big(\mathscr{L}(X)\big) = \frac{1}{k!} \sum_{i_1,\ldots,i_k \in \mathrm{P}\{1,\ldots,k\}} \mathscr{L}(x_{i_1}, \ldots, x_{i_k})$$

[1] M. M. Wolf, "Mathematical foundations of supervised learning"

Convergence

Problem: $\mathbb{E}(\mathscr{L}(X)) \neq \mathscr{L}(\mathbb{E}(X))$

- Solved for polynomial loss functions

$$\mathscr{L}(X) = \sum_{j=0}^{k} a_j X^j = \mathscr{L}(X_1, \ldots, X_k)$$

$$e_k\big(\mathscr{L}(X)\big) = \frac{1}{k!} \sum_{i_1,\ldots,i_k \in \mathrm{P}\{1,\ldots,k\}} \mathscr{L}(x_{i_1}, \ldots, x_{i_k})$$

- Provable convergence under certain assumptions about $\mathscr{L}$ [1]
  - Polyak-Lojasiewicz (PL) inequality "no local minima"
  - Lipschitz continuity

[1] M. M. Wolf, "Mathematical foundations of supervised learning"

Thanks for your attention

Co-authors:

Ryan Sweke

Freie Universität Berlin

Johannes Jakob Meyer

Freie Universität Berlin

Maria Schuld

Xanadu Inc. and University of KwaZulu-Natal South Africa

Paul K. Fährmann

Freie Universität Berlin

Barthélémy Meynard-Piganeau

Ecole Polytechnique Paris

Jens Eisert

Freie Universität Berlin

Quantum 4, 314 (2020)

Slides at:

frederikwil.de/jmc2021

What now?

[1] W. Lavrijsen et al., 2004.03004
[2] J. J. Meyer et al., 2006.06303

What now?

- Non-polynomial loss functions

[1] W. Lavrijsen et al., 2004.03004
[2] J. J. Meyer et al., 2006.06303

What now?

- Non-polynomial loss functions
- #(measurement shots) and barren plateaus

[1] W. Lavrijsen et al., 2004.03004
[2] J. J. Meyer et al., 2006.06303

What now?

- Non-polynomial loss functions
- #(measurement shots) and barren plateaus
- Impact of noise [1]
  Robustness through the parameter shift rule? [2]

[1] W. Lavrijsen et al., 2004.03004
[2] J. J. Meyer et al., 2006.06303

Fast QAOA optimization

Fast QAOA optimization

$$|\psi(\boldsymbol{\beta}, \boldsymbol{\gamma})\rangle = \mathrm{e}^{-\mathrm{i}\beta_p X}\, \mathrm{e}^{-\mathrm{i}\gamma_p H} \;\cdots\; \mathrm{e}^{-\mathrm{i}\beta_1 X}\, \mathrm{e}^{-\mathrm{i}\gamma_1 H} \; |+\rangle$$

Fast QAOA optimization

$$|\psi(\boldsymbol{\beta},\boldsymbol{\gamma})\rangle = e^{-i\beta_p X} e^{-i\gamma_p H} \dots e^{-i\beta_1 X} e^{-i\gamma_1 H} \;|+\rangle$$

$$e^{i\beta_p \sigma_1^x} \dots e^{i\beta_p \sigma_n^x}$$

Fast QAOA optimization

$$|\psi(\boldsymbol{\beta}, \boldsymbol{\gamma})\rangle = \mathrm{e}^{-\mathrm{i}\beta_p X} \mathrm{e}^{-\mathrm{i}\gamma_p H} \dots \mathrm{e}^{-\mathrm{i}\beta_1 X} \mathrm{e}^{-\mathrm{i}\gamma_1 H} \,|+\rangle$$

$$\mathrm{e}^{\mathrm{i}\beta_p \sigma_1^x} \dots \mathrm{e}^{\mathrm{i}\beta_p \sigma_n^x}$$

$2n$ parameter shift terms

Fast QAOA optimization

$$|\psi(\boldsymbol{\beta},\boldsymbol{\gamma})\rangle = e^{-i\beta_p X} e^{-i\gamma_p H} \dots e^{-i\beta_1 X} e^{-i\gamma_1 H} |+\rangle$$

$$e^{i\beta_p \sigma_1^x} \dots e^{i\beta_p \sigma_n^x}$$

$2m$ parameter shift terms

$2n$ parameter shift terms

Fast QAOA optimization

$$|\psi(\boldsymbol{\beta},\boldsymbol{\gamma})\rangle = e^{-i\beta_p X} e^{-i\gamma_p H} \dots e^{-i\beta_1 X} e^{-i\gamma_1 H} \, |+\rangle$$

$$e^{i\beta_p \sigma_1^x} \dots e^{i\beta_p \sigma_n^x}$$

$2m$ parameter shift terms

$2n$ parameter shift terms

Sampling PS terms alone reduces #(circuit evaluations) by a factor of $\dfrac{2p}{2pn + 2pm} = \dfrac{1}{n+m}$ per optimization step.

Scaling

QAOA on Erdős–Rényi graphs (edge probability 30%) ~20 samples/data-point (unpublished)

Corrections for polynomial loss functions

- Let X represent the measurement
- Expand $\mathscr{L}(X)$ around $\mathbb{E}(X) = x_0$
- $\mathscr{L}(X) = \mathscr{L}(x_0) + \mathscr{L}'(x_0)\,(X - x_0) + \sum_{n=2}^{s} \frac{1}{n!} \mathscr{L}^{(n)}(x_0)\,(X - x_0)^n$
- $\mathbb{E}\left[\mathscr{L}(X)\right] = \mathbb{E}\left[\mathscr{L}(x_0)\right] + \sum_{n=2}^{s} \frac{1}{n!} \mathscr{L}^{(n)}(x_0)\,\mathbb{E}\left[(X - x_0)^n\right]$