Convergence Theory

Variational Quantum Algorithms · Module 6 — Optimization Dynamics

Node 6.5

Position in Knowledge Graph

System: Variational Quantum Algorithms **Structural Domain:** Optimization Dynamics **Node:** 6.5

This node is the **theoretical** node of Module 6. It collects the convergence guarantees for VQA optimizers — what we can prove about how fast they converge, under what conditions, and to what.

The honest answer is: **less than we'd like.** Most classical convergence theory assumes convexity or strong-convexity, which VQA cost landscapes do not satisfy. Most stochastic convergence theory assumes bounded-variance noise, which is violated by hardware noise (which can be biased) and is borderline for shot noise. The result is that VQA optimizer behavior is mostly understood empirically, with theory providing a (loose) skeleton.

That said, some theoretical results do apply, and they are worth knowing because they tell you what's *possible* and what's *not* in principle.

The Convergence Hierarchy

VQA convergence theory roughly stratifies into the following hierarchy, from strongest to weakest:

1. **Convex strongly-convex theory:** $O(1/k)$ or $O(\rho^k)$ (linear). Requires convex C. **Does not apply to VQAs in general.**
2. **Smooth non-convex theory:** $O(1/\sqrt{k})$ for $\mathbb{E}[\|\nabla C\|^2]$ under SGD. Requires L-smooth C and bounded gradient variance. **Applies to VQAs in shallow regimes.**
3. **PL-condition theory:** $O(\rho^k)$ under the Polyak-Łojasiewicz condition (a weakened convexity replacement). **Sometimes applies to VQAs near minima.**
4. **No-guarantee theory:** convergence is empirical, no formal rates. **The default assumption for VQAs in practice.**

The challenge of VQA convergence theory is figuring out which level of the hierarchy applies to which VQA instance, and what the relevant constants are.

L-Smoothness And Why It Holds For VQAs

A cost function is L**-smooth** if its gradient is L-Lipschitz:

$$\|\nabla C(\boldsymbol{\theta}_1) - \nabla C(\boldsymbol{\theta}_2)\| \leq L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

For VQAs, this is **always satisfied** with a calculable L, because the cost is built from cosines and sines of parameters, which have bounded second derivatives.

Specifically: for an ansatz built from Pauli rotations, each parameter contributes at most $\|H\|$ to the Lipschitz constant of the gradient (this is a consequence of the parameter-shift identity and the boundedness of H). For a P-parameter ansatz, $L \leq \|H\| \cdot P$ at most, often much less.

L-smoothness is what enables the standard SGD convergence rate $O(1/\sqrt{k})$ for non-convex objectives. This is the **only** clean convergence rate that applies to VQAs in general — it gives you a guarantee that, asymptotically, you're heading toward a critical point.

The catch: $O(1/\sqrt{k})$ for the *gradient norm*, not the cost. To convert "small gradient" into "small cost minus optimum" requires additional structure (convexity or PL).

The Polyak-Łojasiewicz (PL) Condition

The PL condition is a weakened form of convexity that often holds locally:

$$\frac{1}{2}\|\nabla C(\boldsymbol{\theta})\|^2 \geq \mu(C(\boldsymbol{\theta}) - C^*),$$

for some $\mu > 0$ and the global minimum C^*.

This says: the gradient is at least *some* fraction of the cost gap. Equivalently, you can't have a small gradient with a large cost gap.

Under PL: - SGD converges at the *deterministic* rate, not the slower stochastic rate (you can think of PL as saying "the noise floor doesn't matter as long as you're far from the minimum"). - The rate is $O(\rho^k)$ (linear) for the cost, not just for the gradient.

Does PL hold for VQAs? - Globally: very rarely. - Locally near a minimum: often, with μ related to the local Hessian eigenvalues. - In the barren plateau regime: definitely not (small gradient, large cost gap).

So PL is a **regional** property that you can sometimes invoke for late-stage convergence proofs.

SGD Convergence Under L-Smooth Non-Convex

The cleanest result that applies to VQAs:

Theorem (smooth non-convex SGD): For L-smooth C with bounded gradient variance σ^2, the iterates of stochastic GD with step size $\eta = c/\sqrt{k}$ satisfy

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\|\nabla C(\boldsymbol{\theta}_k)\|^2] \leq \frac{C(\boldsymbol{\theta}_0) - C^*}{c\sqrt{K}} + \frac{cL\sigma^2}{\sqrt{K}}.$$

In words: the *average squared gradient norm* over the trajectory is $O(1/\sqrt{K})$.

Implications:

- Average gradient eventually decreases — we are converging toward a critical point.
- The rate $O(1/\sqrt{K})$ is slow but nontrivial.
- Doubling K only reduces the bound by $\sqrt{2}$, not 2. So 4x more iterations halves the bound.
- The constant depends on σ^2 — the *gradient variance* — which for VQAs is set by shot noise plus barren plateau effects.

For shallow VQAs without barren plateau, this gives a useful (if loose) guarantee. For deep VQAs in the plateau, σ^2 explodes and the bound is vacuous.

Convergence Rates For Different Optimizers

A summary table for the optimizers covered in Module 6:

Optimizer	Rate (cost)	Rate (grad norm)	Cost per iteration	Where it applies
Vanilla GD (deterministic)	$O(1/k)$	$O(1/k)$	$2P$	Smooth convex
Vanilla SGD	$O(1/\sqrt{k})$	$O(1/\sqrt{k})$	$2P$	Smooth non-convex
Natural Gradient	$O(1/k)$ (better constants)	$O(1/k)$	$2P + O(P^2)$	Smooth convex
Adam	$O(1/\sqrt{k})$	$O(1/\sqrt{k})$	$2P$	Smooth non-convex
SPSA	$O(1/\sqrt{k})$	$O(1/\sqrt{k})$	2	Smooth non-convex
Population (CMA-ES)	Empirically $O(1/k)$	N/A (zero-order)	O(pop size)	Multi-modal

The **rates** are roughly the same across optimizers (for non-convex problems, everyone gets $O(1/\sqrt{k})$ at best). The differences are in:

1. **Constants** (Natural Gradient and Adam have better constants in practice).
2. **Per-iteration cost** (SPSA has $O(1)$, Natural Gradient has $O(P^2)$).
3. **Robustness regime** (Adam handles ill-conditioning, SPSA handles many parameters cheaply).

The **total cost** to reach a fixed accuracy is a product of (rate) and (per-iteration cost), and the best optimizer depends on which factor dominates.

What The Theory Doesn't Cover

Things that VQA convergence theory does *not* address well:

1. Barren plateaus. No clean convergence rate when the gradient is below the noise floor. The theorems either give vacuous bounds or assume away the plateau.

2. Non-stationary noise. Hardware noise drifts over hours/days, but stochastic-optimization theory assumes i.i.d. noise. Drift breaks the analysis.

3. Biased gradients. Hardware noise can bias the gradient estimator (the noisy expected gradient differs from the true gradient). Standard theory assumes unbiased estimators.

4. Discrete decisions. Many VQA optimizers make discrete choices — which Pauli grouping to measure, which parameter to update next, etc. — that aren't covered by smooth-optimization theory.

5. Population dynamics. Population-based optimizers (HOPSO, CMA-ES) have their own convergence theories but they don't fit the SGD framework.

For these reasons, VQA optimizer practice usually relies on **empirical** convergence behavior with theory as a sanity check, not the other way around.

A Useful Theoretical Frame: Implicit Regularization

A more recent theoretical perspective: **the choice of optimizer implicitly regularizes the solution.**

In classical deep learning, it has been shown that SGD has an implicit bias toward "flat minima" (minima with small Hessian eigenvalues), which generalize better. This implicit bias is partly responsible for why SGD outperforms full-batch GD on overparameterized neural networks, even though both should converge to the same minima.

The same idea applies to VQAs: different optimizers find different minima, even when starting from the same initialization. Adam tends to find flatter minima than vanilla GD; natural gradient finds minima that are flat in the Fubini-Study metric; SPSA finds minima robust to per-coordinate perturbations.

This is the bridge to **regularization** (Module 5): the optimizer is not just a tool for finding the minimum — it's a mechanism for *selecting* among the minima a given cost landscape contains. This is the thesis's "dynamics-side regularization" view in another guise.

Stochastic Convergence Theorems For VQA-Specific Settings

A few results worth knowing that are tailored to VQA:

Sweke et al. (2020): "Stochastic gradient descent for hybrid quantum-classical optimization." Showed that even with single-shot gradient estimates ($N = 1$ per parameter-shift evaluation), SGD on VQAs converges in the smooth non-convex sense, provided the learning rate decays appropriately. This is theoretically reassuring — you don't need many shots per estimate, you can use single shots and rely on the optimizer to average.

Harrow and Napp (2021): "Low-depth gradient measurements can improve convergence in variational hybrid quantum-classical algorithms." Showed that more efficient gradient measurements (jointly measuring multiple gradient components) can reduce variance and accelerate convergence.

Skolik et al. (2021): "Layerwise learning for quantum neural networks." Provided the theoretical justification for layerwise training as a way to avoid the barren plateau regime, with explicit convergence guarantees in the layerwise setting.

These are examples of theory tailored to VQA structure rather than borrowed wholesale from classical optimization.

Structural Role

This node is the **theoretical scaffolding** of Module 6. It provides the convergence rate framework that subsequent sections use, identifies the regimes where theory does and doesn't apply, and connects to the implicit-regularization view that bridges to Module 5.

Relations to Other Concepts

- **Robbins-Monro 1951** — foundational convergence theorem for stochastic approximation.
- **L-smoothness / Lipschitz gradients** — the standard structural assumption.
- **PL condition** — weakened convexity sufficient for fast rates.
- **Implicit regularization (deep learning)** — the optimizer-as-regularizer view.
- **Sweke et al. 2020** — single-shot SGD convergence for VQA.

Worked Example — Convergence Rate For A Smooth Convex VQE Toy

Consider $C(\theta) = (1 - \cos(\theta))^2/4$, a smooth convex cost on a single parameter with minimum at $\theta = 0$ and $C^* = 0$.

Compute the smoothness constant. $C'(\theta) = (1 - \cos\theta)\sin\theta/2$. $|C''(\theta)| \leq 1$ (verifiable by differentiating). So $L = 1$.

Vanilla GD with $\eta = 0.5$: the convergence rate for smooth convex is $O(1/k)$, so $C(\theta_k) - C^* \leq C(\theta_0)/(1 + k\eta)$.

Starting from $\theta_0 = \pi/2$: $C(\pi/2) = 1/4$. Predicted bound at $k = 100$: $(1/4)/51 \approx 0.005$. The actual error after 100 steps is similar (the bound is loose but in the right ballpark).

SGD with shot noise $\sigma = 0.05$: the bound becomes $O(\sigma^2/\sqrt{k})$ instead of $O(1/k)$, because the noise floor dominates after enough iterations. At $k = 100$, the bound is $0.05^2/10 = 0.00025$ — actually *better* than the deterministic bound! This is because the $O(1/k)$ deterministic bound is loose, and the SGD bound only "pays" for the noise asymptotically.

In practice, deterministic GD reaches 10^{-4} precision in ~100 iterations. SGD with the same step size and $\sigma = 0.05$ reaches 10^{-3} precision in ~100 iterations and stagnates near the noise floor. To reach 10^{-4}, you'd need either much more shots (lower σ) or many more iterations.

This is the canonical "rate vs noise floor" tradeoff that shapes all stochastic optimization.

Visualization

Pending. A log-log plot of $C(\boldsymbol{\theta}_k) - C^*$ vs iteration k for several optimizers (deterministic GD, stochastic GD, Adam, SPSA, natural gradient). Each shows a different rate and a different noise floor. Caption: "Rates differ; noise floors differ; total budget differs."

Exercises

1. Verify that $C(\theta) = \cos(\theta)$ is L-smooth with $L = 1$. Compute the convergence rate of GD with $\eta = 1$ starting at $\theta_0 = 0.1$, targeting the maximum at $\theta = 0$. (Note: this is a maximum, not a minimum — but the same theorem applies.)
2. The PL condition is $\frac{1}{2}\|\nabla C\|^2 \geq \mu(C - C^*)$. For $C(\theta) = \cos(\theta) - \cos(\theta^*)$ near $\theta = \theta^*$, what is μ?
3. (VQA) For a 30-parameter HEA at a fixed shot budget of 10^6 total shots, estimate which optimizer (vanilla GD, Adam, SPSA, natural gradient) will give the lowest final cost. State your assumptions.

Research Extensions

- **VQA-specific PL conditions** — finding regions of VQA cost landscapes where PL holds and quantifying the constant.
- **Convergence under hardware noise drift** — extending stochastic optimization theory to non-stationary noise.
- **Convergence under biased gradients** — VQA with biased estimators.
- **Implicit regularization theory for VQA optimizers** — formalizing which optimizers prefer which kinds of minima.

Cross-links to Other Courses

- **Convex Optimization (Boyd-Vandenberghe)** — foundational convergence results.
- **Stochastic Approximation (Kushner-Yin)** — asymptotic theory of stochastic recursions.
- **Module 4 (Barren Plateaus)** — the regime where convergence theory fails to give useful bounds.
- **Module 5 (Regularization)** — the implicit regularization perspective on optimizer choice.

Variational Quantum Algorithms · Module 6 — Optimization Dynamics · Node 6.5
Concepts: convergence, convexity, stochastic-optimization, gradient-descent

Prerequisites: 6.1, 6.4
Enables: 6.6

hbar.university — own.your.cognition