QC-FQL(+EMA): Stabilized Chunk-based Flow Q-Learning for Long-Horizon Reinforcement Learning

Author: Ji-hye Kim (2025)

1. Introduction

Recent advances in offline and model-free reinforcement learning (RL) have focused on integrating value-based updates with generative or flow-based policy architectures. Flow Q-Learning (FQL) [Park et al., 2024] introduced a novel formulation where the policy is represented as a continuous vector field trained via flow-matching, bridging the gap between behavior cloning and Q-learning. Subsequently, Q-Chunked Flow Q-Learning (QC-FQL) [Li et al., 2025, Reinforcement Learning with Action Chunking] extended FQL to chunked (multi-step) action spaces, improving temporal consistency and long-horizon reasoning. However, as the chunk size H increases, the variance of TD targets and the instability of policy updates rise significantly. This report provides a theoretical and architectural comparison of FQL, QC-FQL, and a stabilized variant QC-FQL(+EMA) that incorporates an exponential-moving-average (EMA) target actor to mitigate instability in long-horizon settings.

2. Flow Q-Learning (FQL)

Flow Q-Learning models the policy as a deterministic transformation of Gaussian noise through a learned vector field. Rather than parameterizing $\pi(a|s)$directly, FQL defines a flow function $f\hat{}\zeta(s,x,t)$ whose time-integrated velocity field transports a base distribution $x_0 \sim \mathcal{N}(0, I)$ to an expert or optimal action a*. The flow is trained by minimizing a flow-matching loss, ensuring that the learned vector field approximates the gradient of the optimal transport map under the 2-Wasserstein metric.

Flow-matching loss (behavioral teacher):

$$L_{FM\,(\zeta)} = E\left[\left|\left|f_\zeta(s_t,t) - (a^* - x_0)\right|\right|^2\right], where\ x_t = (1-t)x_0 + ta^*$$

Student actor loss:

$$L_{actor(\psi)} = \alpha E[||\mu_\psi(s,z) - sg(f_\zeta(s,z))||^2] - E[Q^\theta(s,\mu^\psi(s,z))]$$

Critic loss:

$$L_{critic(\theta)} = E[(Q^\theta(s,a) - \left(r + \gamma\, \bar{Q}_\theta\left(s', \mu_\psi(s',z')\right)\right)^2]$$

3. Q-Chunked Flow Q-Learning (QC-FQL)

QC-FQL generalizes FQL to multi-step "action chunks," enabling the policy to output temporally coherent control sequences. Instead of learning $Q(s,a)$ for a single action, QC-FQL defines a chunked critic $Q(s, a_{\{t:t+H\}})$ that evaluates the value of an entire action subsequence of horizon length H. This permits longer temporal credit assignment and reduces bootstrapping bias common in n-step value estimates.

Chunked TD target:

$$L_{Q(\theta)} = E\left[\left(Q_\theta(s_t, a_{\{t:t+H\}}) - \sum_0^{\{H-1\}} \gamma^k r_{\{t+k\}} - \gamma^H\, \bar{Q}_\theta(s_{\{t+H\}}, a'_{\{t+H:t+2H\}})\right)^2\right]$$

QC-FQL significantly accelerates value propagation and captures long-range dependencies, but as H increases, target variance also increases due to rapidly changing policies.

4. Instability and EMA Stabilization

Experiments show that as chunk length H increases (>5 steps), QC-FQL training suffers from high target variance because the critic target depends on a rapidly updating actor πψ. To mitigate this, QC-FQL(+EMA) introduces a slowly moving target actor updated via exponential moving average (EMA):

$$\psi_{tgt} \leftarrow \tau_\psi + (1-\tau)\psi_{tgt}$$

Modified TD loss:

$$L_{Q^{EMA}(\theta)} = E\left[\left(Q_\theta(s_t, a_{\{t:t+H\}}) - \sum_0^{\{H-1\}} \gamma^k\, r_{\{t+k\}} - \gamma^H \bar{Q}_\theta\left(s_{\{t+H\}}, \mu_{\psi_{tgt}}(s_{\{t+H\}}, z')\right)\right)^2\right]$$

The EMA target actor smooths policy updates, acting as a low-pass filter to stabilize Q-value targets. This restores convergence for large-H training (H≥8) and reduces gradient oscillations.

5. Comparative Analysis

QC-FQL(+EMA) inherits the flow-matching regularization and chunked abstraction of QC-FQL, while enhancing stability.
The EMA actor complements the target critic, forming a dual-target mechanism similar to TD3. Both critics and actors maintain synchronized yet slowly evolving parameters, reducing instability in long-horizon settings.

Aspect	FQL	QC-FQL	QC-FQL(+EMA)
Action Representation	Single step	Chunked ($H \times Da$)	Chunked ($H \times Da$)
Critic Target	$r + \gamma \bar{Q}(s', \mu_\psi)$	$r^H + \gamma^H \bar{Q}(s_{\{t+H\}}, \mu_\psi)$	$r^H + \gamma^H \bar{Q}(s_{\{t+H\}}, \mu_{\psi_{tgt}})$
Target Networks	Critic	Critic	Critic+Actor(EMA)
Advantage	Smooth policy	Long-horizon learning	Stable long-horizon learning
Limitation	Short horizon	Variance with H	Bias from EMA

6. Theoretical Implications and Conclusion

From a theoretical viewpoint, introducing the EMA target actor imposes a temporal smoothness prior on the policy update process,
equivalent to a differential constraint $d\psi_{tgt}/dt = \tau(\psi - \psi_{tgt})$. This reduces high-frequency policy oscillations and prevents
target drift during multi-step bootstrapping. Though slightly biased, it significantly enhances convergence stability.
Empirically, QC-FQL(+EMA) achieves higher success rates and smoother learning curves on dual-arm manipulation tasks with long-horizon dependencies.

In summary, QC-FQL(+EMA) unifies flow-matching imitation, chunked temporal abstraction, and EMA-based target stabilization.
This synergy allows robust and scalable flow-based reinforcement learning for complex

robotic systems requiring long-horizon control.

References

[1] Park et al. (2024). Flow Q-Learning: Wasserstein-Regularized Offline-to-Online Reinforcement Learning.
[2] Li et al. (2025). Reinforcement Learning with Action Chunking. arXiv:2507.07969.