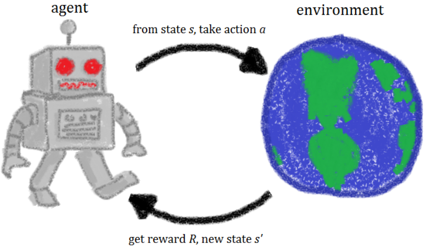
Nonparametric Stochastic Compositional Gradient Descent for Q-Learning in Continuous Markov Decision Problems

Ekaterina Tolstaya*, Alec Koppel†, Ethan Stump†, Alejandro Ribeiro*
* University of Pennsylvania
† U.S. Army Research Laboratory

American Control Conference, Milwaukee, WI
June 29th, 2018

Markov Decision Problems



- At time t, agent is in state $\mathbf{s}_t$, select action $\mathbf{a}_t$.
- Transition from state $\mathbf{s}_t$ to $\mathbf{s}_{t+1}$, with $\mathbf{s}_{t+1} \sim \mathbb{P}(\cdot \mid \mathbf{s}_t, \mathbf{a}_t)$
  - Reward $r_t := r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})$
- Markov Decision Process (MDP): $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$
  - State space $\mathcal{S}$, action space $\mathcal{A}$, discount factor $\gamma \in (0, 1)$
- Goal: find a policy $\pi \in \mathcal{S} \to \mathcal{A}$, a map from states to actions,
  - That maximizes the long-term reward accumulation

Reinforcement Learning

Goal: choose actions to maximize infinite discounted reward accumulation

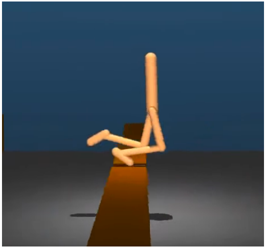
$$\max_{\{a_t\}} \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1})$$

Recent successes

- AlphaGo Zero (Silver, 2017)
- Bipedal walker on terrain (Heess, 2017)
- Personalized web services (Theocharous, 2015)

Remaining challenges in infinite spaces

- Reproducibility
- Lack of guarantees for function approx.



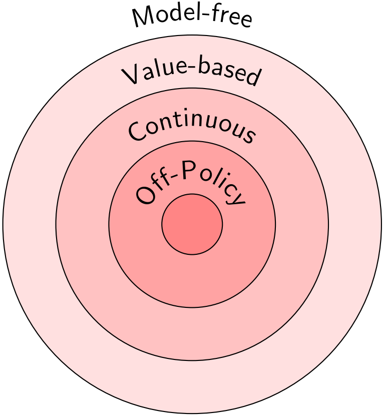
Our Approach

We develop the KQ-Learning algorithm:

- Formulate a stochastic program for off-policy Bellman loss
- Compute stochastic gradient
- Update the kernel model
- Sparsify the model

Our results provide:

- Convergence guarantees
- Experimental validation
- Low complexity solutions



Value Based Approaches

- Value function, expected reward accumulation given initial $\mathbf{s}$:
  - While following policy π

$$V^{\pi}(\mathbf{s}) := \mathbb{E}_{\mathbf{s}'}\left[\sum_{t=0}^{\infty}\gamma^{t} r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}'_t) \mid \mathbf{s}_0 = \mathbf{s}, \mathbf{a}_t = \pi(\mathbf{s}_t)\right] \quad (1)$$

- Action-value function, the reward accumulation given initial $\mathbf{s}, \mathbf{a}$

$$Q^{\pi}(\mathbf{s}, \mathbf{a}) := \mathbb{E}_{\mathbf{s}'}\left[\sum_{t=0}^{\infty}\gamma^{t} r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}'_t) \mid \mathbf{s}_0 = \mathbf{s}, \mathbf{a}_0 = \mathbf{a}, \mathbf{a}_t = \pi(\mathbf{s}_t)\right] \quad (2)$$

- Goal: learn the optimal action-value function
  - Satisfying the Bellman optimality equation

$$Q^{*}(\mathbf{s}, \mathbf{a}) = \mathbb{E}_{\mathbf{s}'}[r(\mathbf{s}, \mathbf{a}, \mathbf{s}') + \gamma \max_{\mathbf{a}'} Q^{*}(\mathbf{s}', \mathbf{a}') \mid \mathbf{s}_0 = \mathbf{s}, \mathbf{a}_0 = \mathbf{a}] \quad (3)$$

  - Q-learning is an off-policy approach

Guarantees for Value Based Approaches

- Tabular Q-Learning (Dayan, 1992)
  - Off-policy observations in discrete state and action spaces
  - Convergence w.p. 1 when all states, actions observed i.o.
- Policy evaluation (Tsitsiklis, 1997)
  - Continuous state space with linear function approximation
  - Convergence a.s.
- Gradient Temporal Difference (Sutton, 2009)
  - Off-policy updates with linear function approximation
  - Convergence w.p. 1
- Policy evaluation (Koppel, 2017)
  - Continuous state space with non-parametric kernel methods
  - Convergence to the Bellman fixed point w.p. 1

Optimizing Bellman Loss

- Bellman optimality equation (Bertsekas, 2004)

$$Q^*(\mathbf{s},\mathbf{a}) = \mathbb{E}_{\mathbf{s}'}[r(\mathbf{s},\mathbf{a},\mathbf{s}') + \gamma \max_{\mathbf{a}'} Q^*(\mathbf{s}',\mathbf{a}')] \tag{4}$$

- Temporal difference for an observation $(\mathbf{s}, \mathbf{a}, \mathbf{s}')$:

$$\delta := r(\mathbf{s},\mathbf{a},\mathbf{s}') + \gamma \max_{\mathbf{a}'} Q(\mathbf{s}',\mathbf{a}') - Q(\mathbf{s},\mathbf{a}) \tag{5}$$

- Define an auxiliary function for the expected temporal difference:

$$f(Q;\mathbf{s},\mathbf{a}) := \mathbb{E}_{\mathbf{s}'}\delta = \mathbb{E}_{\mathbf{s}'}\left[r(\mathbf{s},\mathbf{a},\mathbf{s}') + \gamma \max_{\mathbf{a}'} Q(\mathbf{s}',\mathbf{a}') - Q(\mathbf{s},\mathbf{a}) \,\middle|\, \mathbf{s},\mathbf{a}\right]$$

- Reformulate Bellman optimality equation as comp. stochastic prog:

$$L(Q) = \frac{1}{2}\mathbb{E}_{\mathbf{s},\mathbf{a}}\left[f^2(Q;\mathbf{s},\mathbf{a})\right]. \tag{6}$$

- Q^* is a function that satisfies $f(Q;\mathbf{s},\mathbf{a}) = 0$ for all $(\mathbf{s},\mathbf{a})$.
  - A solution to the non-convex optimization problem

Reproducing Kernel Hilbert Spaces (RKHS)

- We restrict $Q \in \mathcal{H}$, a **Reproducing Kernel Hilbert space**
- An RKHS over $\mathcal{S} \times \mathcal{A}$ is equipped with a reproducing kernel (Norkin, 2009; Argyriou, 2009)
  - An inner product-like map, $\kappa : (\mathcal{S} \times \mathcal{A}) \times (\mathcal{S} \times \mathcal{A}) \to \mathbb{R}$:

$$\text{(i)}\langle Q, \kappa((\mathbf{s}, \mathbf{a}), \cdot)\rangle_{\mathcal{H}} = Q((\mathbf{s}, \mathbf{a})), \qquad \text{(ii)}\mathcal{H} = \text{span}\{\kappa((\mathbf{s}, \mathbf{a}), \cdot)\} \quad (7)$$

- A continuous function over a compact set may be approx. uniformly
  - In an RKHS equipped with a universal kernel (Michelli, 2006)
- We solve the regularized problem:

$$Q^* = \arg\min_{Q \in \mathcal{H}} J(Q) = \arg\min_{Q \in \mathcal{H}} L(Q) + \frac{\lambda}{2}\|Q\|^2_{\mathcal{H}}. \qquad (8)$$

- Goal: optimize $J(Q)$ over $\mathcal{H}$ given samples $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}'_t)$
- First, differentiate $J(Q)$ w.r.t. Q. (Koppel, 2017)

$$\nabla_Q J(Q_t) = \mathbb{E}_{\mathbf{s}_t, \mathbf{a}_t}\Big[f(Q_t; \mathbf{s}_t, \mathbf{a}_t) \times \nabla_Q f(Q_t; \mathbf{s}_t, \mathbf{a}_t)\Big] + \lambda Q_t. \tag{9}$$

- Stoch. grad. unusuable since $\nabla_Q J(Q)$ has two expectations
- *Coupled* descent: estimate both terms in product-of-expectations
- Construct total mean of $\hat{\nabla}_Q f = [\gamma\kappa((\mathbf{s}'_t, \mathbf{a}'_t), \cdot) - \kappa((\mathbf{s}_t, \mathbf{a}_t), \cdot)]$?
  - Infinite complexity!
- Instead: build up expectation of scalar temporal difference δ

Stochastic Quasi-Gradient

- Goal: extend gradient temporal diff. (Sutton, 2009) to infinite MDPs
- Define a scalar fixed pt. recursion z_t to estimate average TD $\bar{\delta}$

$$\delta_t = r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}') + \gamma \max_{a'} Q(\mathbf{s}'_t, \mathbf{a}') - Q_t(\mathbf{s}_t, \mathbf{a}_t),$$
$$z_{t+1} = (1 - \beta_t) z_t + \beta_t \delta_t$$

- $\delta_t \Rightarrow$ temporal difference; $\beta_t \in (0, 1) \Rightarrow$ step-size.
- $\mathbf{a}'_t = \arg\max_{\mathbf{a}'} Q(\mathbf{s}'_t, \mathbf{a}')$ can be replaced with a softmax
  - In practice, evaluated via simulated annealing
- Stoch. descent step: replace 1st term in expectation w/ estimate
  - $[\gamma\kappa((\mathbf{s}'_t, \mathbf{a}'_t), \cdot) - \kappa((\mathbf{s}_t, \mathbf{a}_t), \cdot)]$, evaluated at $(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}'_t, \mathbf{a}'_t)$
  - replace δ_t by $z_{t+1} \Rightarrow$ stoch. quasi-gradient (Ermoliev '83)

$$\hat{Q}_{t+1} = (1 - \alpha_t \lambda) \hat{Q}_t - \alpha_t z_{t+1} \left[\gamma\kappa((\mathbf{s}'_t, \mathbf{a}'_t), \cdot) - \kappa((\mathbf{s}_t, \mathbf{a}_t), \cdot)\right]$$

  - α_t is a second step-size

RKHS Parameterization

- If $Q_0 = 0 \in \mathcal{H}$, inductively applying Representer Thm. yields

$$Q_t(\mathbf{s},\mathbf{a}) = \sum_{n=1}^{2(t-1)} w_n \kappa((\mathbf{s}_n,\mathbf{a}_n),(\mathbf{s},\mathbf{a})) = \mathbf{w}_t{}^T \kappa_{\mathbf{X}_t}((\mathbf{s},\mathbf{a})) \qquad (10)$$

$$\begin{aligned} \mathbf{w}_t &= [w_1,\ldots,w_{2(t-1)}] \,, \\ \mathbf{X}_t &= [(\mathbf{s}_1,\mathbf{a}_1),(\mathbf{s}'_1,\mathbf{a}'_1),\ldots,(\mathbf{s}_{t-1},\mathbf{a}_{t-1}),(\mathbf{s}'_{t-1},\mathbf{a}'_{t-1})], \end{aligned} \qquad (11)$$

- Kernel expansion + together with FSQG $\Rightarrow$ parametric updates:

$$\begin{aligned} \mathbf{X}_{t+1} &= [\mathbf{X}_t,(\mathbf{s}_t,\mathbf{a}_t),(\mathbf{s}_t{}',\mathbf{a}_t{}')] \,, \\ \mathbf{w}_{t+1} &= [(1-\alpha_t\lambda)\mathbf{w}_t, \alpha_t z_{t+1}, -\alpha_t\gamma z_{t+1}] \end{aligned} \qquad (12)$$

- Intractable complexity intrinsic to RKHS optimization: $M_t = \mathcal{O}(t)$
- Solve via Kernel Orthogonal Matching Pursuit (KOMP) (Koppel, 2016)

KQ-Learning Algorithm

Require: $\{\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}'_t, \alpha_t, \beta_t, \epsilon_t\}_{t=0,1,2,\ldots}$

initialize $Q_0(\cdot) = 0, \mathbf{D}_0 = [], \mathbf{w}_0 = [], z_0 = 0$

for $t = 0, 1, 2, \ldots$ **do**

Obtain trajectory realization $(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}'_t)$

Evaluate instantaneous maximizing action $\mathbf{a}'_t = \arg\max_{\mathbf{a}'} Q(\mathbf{s}'_t, \mathbf{a}')$

Compute temporal difference

$\delta_t = r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}'_t) + \gamma \max_{\mathbf{a}'} Q_t(\mathbf{s}_t, \mathbf{a}') - Q_t(\mathbf{s}_t, \mathbf{a}_t)$

Update auxiliary sequence $z_{t+1} = (1 - \beta_t) z_t + \beta_t \delta_t$

Compute functional stochastic quasi-gradient step

$$\hat{Q}_{t+1}(\cdot) = (1 - \alpha_t \lambda) Q_t(\cdot) - \alpha_t z_{t+1} [\gamma \kappa((\mathbf{s}'_t, \mathbf{a}'_t), \cdot) - \kappa((\mathbf{s}_t, \mathbf{a}_t), \cdot)]$$

Revise dictionary $\hat{\mathbf{D}}_{t+1} = [\mathbf{D}_t, (\mathbf{s}_t, \mathbf{a}_t), (\mathbf{s}'_t, \mathbf{a}'_t)]$,

and weights $\hat{\mathbf{w}}_{t+1} = [(1 - \alpha_t \lambda) \mathbf{w}_t, \;\; \alpha_t z_{t+1}, -\alpha_t \gamma z_{t+1}]$

Project function $(Q_{t+1}, \mathbf{D}_{t+1}, \mathbf{w}_{t+1}) = \textbf{KOMP}(\hat{Q}_{t+1}, \hat{\mathbf{D}}_{t+1}, \hat{\mathbf{w}}_{t+1}, \epsilon_t)$

end for

Convergence w.p. 1 for Decreasing Step Sizes

Theorem

Consider the sequence z_t and $\{Q_t\}$ as stated in the KQ-Learning algorithm. Assume the regularizer is positive $\lambda > 0$, Assumptions 15-19 hold, and the step-size conditions hold, with $C > 0$ a positive constant:

$$\sum_{t=1}^{\infty} \alpha_t = \infty, \sum_{t=1}^{\infty} \beta_t = \infty, \sum_{t=1}^{\infty} \alpha_t^2 + \beta_t^2 + \frac{\alpha_t^2}{\beta_t} < \infty, \epsilon_t = C\alpha_t^2 \qquad (13)$$

Then $\|\nabla_Q J(Q)\|_{\mathcal{H}}$ **converges to null with probability 1***, and Q_t attains a stationary point of (8).*

- *The limit of Q_t achieves a Bellman fixed point in the RKHS.*
- *Proof uses existing results for compositional stochastic gradient descent (Wang, 2017)*

Convergence for Constant Step Sizes

Theorem

Consider the sequence z_t and $\{Q_t\}$ as stated in the KQ-Learning algorithm. Assume the regularizer is positive $\lambda > 0$, Assumptions 15-19 hold, and the step-sizes are chosen as constant such that $0 < \alpha < \beta < 1$, with $\epsilon = C\alpha^2$ and the parsimony constant $C > 0$ is positive.
Then **Bellman error converges to a neighborhood in expectation***:*

$$\liminf_{t\to\infty} \mathbb{E}[J(Q_t)] \leq \mathcal{O}\left(\frac{\alpha\beta}{\beta-\alpha}\left[1+\sqrt{1+\frac{\beta-\alpha}{\alpha\beta}\left(\frac{1}{\beta}+\frac{\beta^2}{\alpha^2}\right)}\right]\right) \quad (14)$$

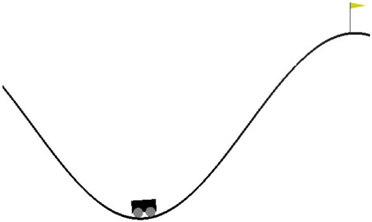
Summary of Results

	Diminishing	Constant
Learning rate	$\sum_{t=1}^{\infty} \alpha_t^2 + \beta_t^2 + \frac{\alpha_t^2}{\beta_t} < \infty$	$0 < \alpha < \beta < 1$
Compression	$\epsilon_t = \mathcal{O}(\alpha_t^2)$	$\epsilon = \mathcal{O}(\alpha^2)$
Regularization	$0 < \lambda$	$0 < \lambda$
Convergence	$\|\nabla_Q J(Q_t)\|_{\mathcal{H}} \to 0$ a.s.	$\liminf_t \mathbb{E}[J(Q_t)] = R(\alpha, \beta)$
Model Order	Infinite	Finite

- $R(\alpha, \beta) = \mathcal{O}\left(\frac{\alpha\beta}{\beta-\alpha}\left[1+\sqrt{1+\frac{\beta-\alpha}{\alpha\beta}\left(\frac{1}{\beta}+\frac{\beta^2}{\alpha^2}\right)}\right]\right)$
- Exact solution requires infinite memory
- **Approximate, but accurate solution with finite memory**

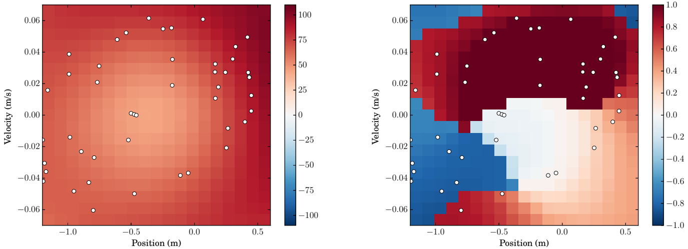
Continuous Mountain Car

- Open AI Gym benchmark problem
- State is 2-dimensional: position and velocity
- Action is 1-dimensional: force within a continuous interval $[-1, 1]$
- Reward is 100 at goal position, and $-0.1a^2$ for actions a



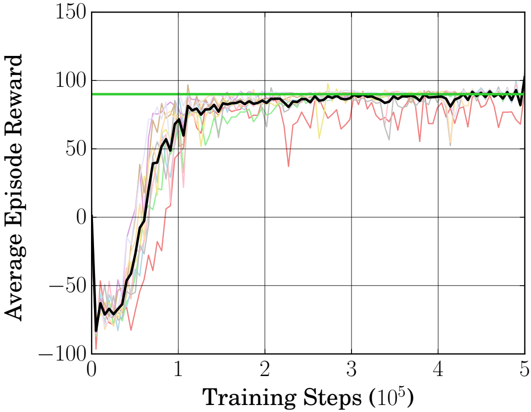
Continuous Mountain Car

- Open AI Gym benchmark problem
- State is 2-dimensional: position and velocity
- Action is 1-dimensional: force within a continuous interval $[-1, 1]$
- Reward is 100 at goal position, and $-0.1a^2$ for actions a



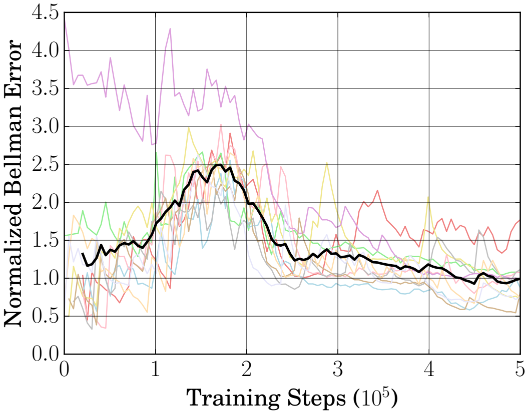
- Visualization of the learned value function and policy
  - Grid color - value,policy
  - White circles - kernel dictionary elements

Continuous Mountain Car



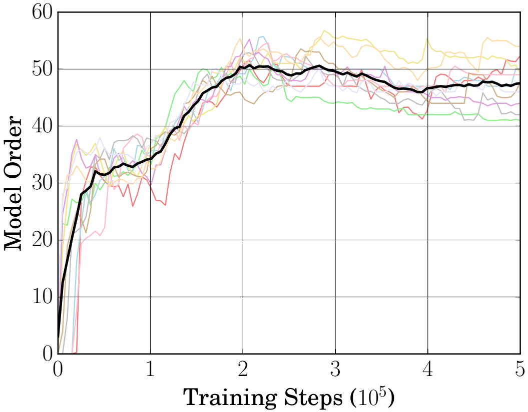
- Reward above 90 (green line) is considered solved

Continuous Mountain Car



- Corroborated by Q function converging to stationarity
- Bellman error $J(Q)$ normalized by $\|Q\|_{\mathcal{H}}$

Continuous Mountain Car



- This is done with an automatic sparse parameterization of Q
  - Directly in a continuous space
  - Model order stabilizes between 45-55

Conclusions

- Contributions
  - KQ-learning approach using non-parametric RKHS representations
  - Convergence guarantees for the KQ-Learning algorithm
- Demonstration on the Mountain Car benchmark problem
  - High reproducibility of results
  - Low complexity of solutions
- Future work
  - Applications in higher-dimensional problems
  - Robotics applications
  - Policy and actor-critic based approaches

Thank you!

$\Rightarrow$ eig@seas.upenn.edu

References I

- Andreas Argyriou, Charles A Micchelli, and Massimiliano Pontil, *When is there a representer theorem? vector versus matrix regularizers*, Journal of Machine Learning Research **10** (2009), no. Nov, 2507–2529.
- Dimitir P Bertsekas and Steven Shreve, *Stochastic optimal control: the discrete-time case*, 2004.
- Peter Dayan, *The convergence of td (λ) for general λ*, Machine learning **8** (1992), no. 3-4, 341–362.
- *Policy evaluation in continuous mdps with efficient kernelized gradient temporal difference*.
- Alec Koppel, Ekaterina Tolstaya, Ethan Stump, and Alejandro Ribeiro, *Nonparametric stochastic compositional gradient descent for q-learning in continuous markov decision problems*, IEEE Transactions on Automatic Control (under preparation) (2017).

References II

Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang, *Universal kernels*, Journal of Machine Learning Research **7** (2006), no. Dec, 2651–2667.

Vladimir Norkin and Michiel Keyzer, *On stochastic optimization and statistical learning in reproducing kernel hilbert spaces by support vector machines (svm)*, Informatica **20** (2009), no. 2, 273–292.

Richard S Sutton, Hamid R Maei, and Csaba Szepesvári, *A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation*, Advances in neural information processing systems, 2009, pp. 1609–1616.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al., *Mastering the game of go without human knowledge*, Nature **550** (2017), no. 7676, 354.

Georgios Theocharous, Philip S Thomas, and Mohammad Ghavamzadeh, *Personalized ad recommendation systems for life-time value optimization with guarantees.*

John N Tsitsiklis and Benjamin Van Roy, *An analysis of temporal-difference learning with function approximation*, IEEE transactions on automatic control **42** (1997), no. 5, 674–690.

Technical Setting for KQ-Learning

- The state space $\mathcal{S} \subset \mathbb{R}^p$ and action space $\mathcal{A} \subset \mathbb{R}^q$ are compact.
- The reproducing kernel map is **bounded**:

$$\sup_{\mathbf{s} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}} \sqrt{\kappa((\mathbf{s}, \mathbf{a}), (\mathbf{s}, \mathbf{a}))} = S < \infty \tag{15}$$

- The temporal difference δ and z satisfy, for $\bar{\delta} = \mathbb{E}[\delta | \mathbf{s}, \mathbf{a}]$,

$$\mathbb{E}[\delta | \mathbf{s}, \mathbf{a}] = \bar{\delta}, \ \ \mathbb{E}[(\delta - \bar{\delta})^2] \leq \sigma_\delta^2, \ \ \ \mathbb{E}[z^2 | \mathbf{s}, \mathbf{a}] \leq G_\delta^2 \tag{16}$$

- The quasi-gradient is an **unbiased estimate** for $\nabla_Q J(Q)$:

$$\mathbb{E}[(\gamma \kappa((\mathbf{s}_t', \mathbf{a}_t'), \cdot) - \kappa((\mathbf{s}_t, \mathbf{a}_t), \cdot))\bar{\delta}] = \nabla_Q J(Q) \tag{17}$$

- The difference of reproducing kernels has **finite cond. variance**:

$$\mathbb{E}[\| \gamma \kappa((\mathbf{s}_t', \mathbf{a}_t'), \cdot) - \kappa((\mathbf{s}_t, \mathbf{a}_t), \cdot) \|_{\mathcal{H}}^2 | \mathcal{F}_t] \leq G_Q^2 \tag{18}$$

- The projected functional gradient has **finite cond. 2$^{\text{nd}}$ moments**:

$$\mathbb{E}[\| \tilde{\nabla}_Q J(Q_t z_{t+1}; \mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_t') \|_{\mathcal{H}}^2 \mid \mathcal{F}_t] \leq \sigma_Q^2 \tag{19}$$