Controlling Overestimation Bias

with **Truncated Mixture** of **Continuous Distributional Quantile Critics**




Arsenii Kuznetsov[1]   Pavel Shvechikov[1,2]   Alexander Grishin[1,3]   Dmitry Vetrov[1,3]

NATIONAL RESEARCH UNIVERSITY

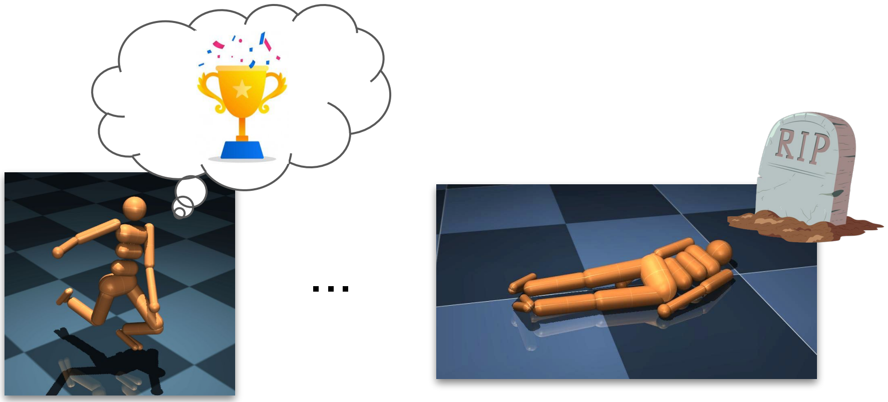
SAMSUNG Research Russia

p(B|A)yesgroup.ru

1 Samsung AI center, Moscow
2 Higher School of Economics, Moscow
3 Samsung HSE Laboratory

Overestimation bias in off-policy learning

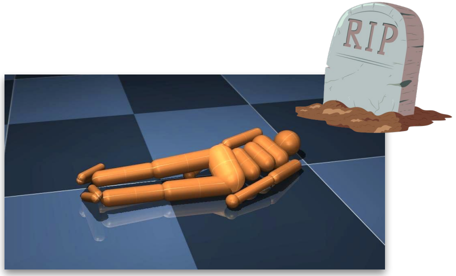
1. Value estimates are imprecise
2. Agent pursues erroneous estimates
3. Errors propagate through time
4. Performance degrades

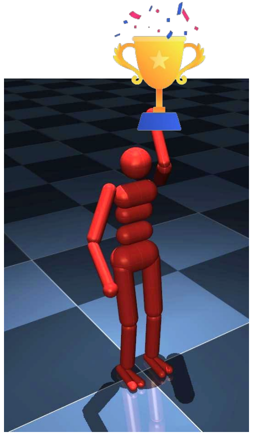


Overestimation bias in off-policy learning

1. Value estimates are imprecise
2. Agent pursues erroneous estimates
3. Errors propagate through time
4. Performance degrades

...



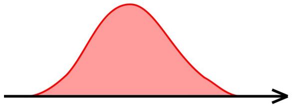


We propose a novel method:

Truncated Quantile Critics (TQC)

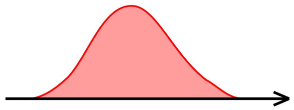
Key elements of TQC

1. Distributional critics
    - Impressive empirical performance
    - Captures info about return variance

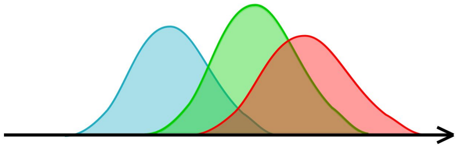


Key elements of TQC

1. Distributional critics
    - Impressive empirical performance
    - Captures info about return variance

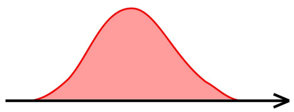


2. Ensembling of the critics
    - Increases performance and stability

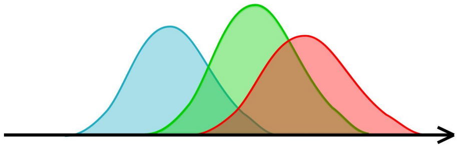


Key elements of TQC

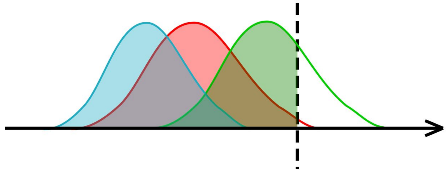
1. Distributional critics
    - Impressive empirical performance
    - Captures info about return variance



2. Ensembling of the critics
    - Increases performance and stability



3. Truncating the mixture of distributions
    - Alleviates overestimation



TQC’s novelties

1. Incorporates **stochasticity of returns** into the overestimation control

TQC’s novelties

1. Incorporates **stochasticity of returns** into the overestimation control

2. Provides **fine-grained** and **adjustable** level of the overestimation control

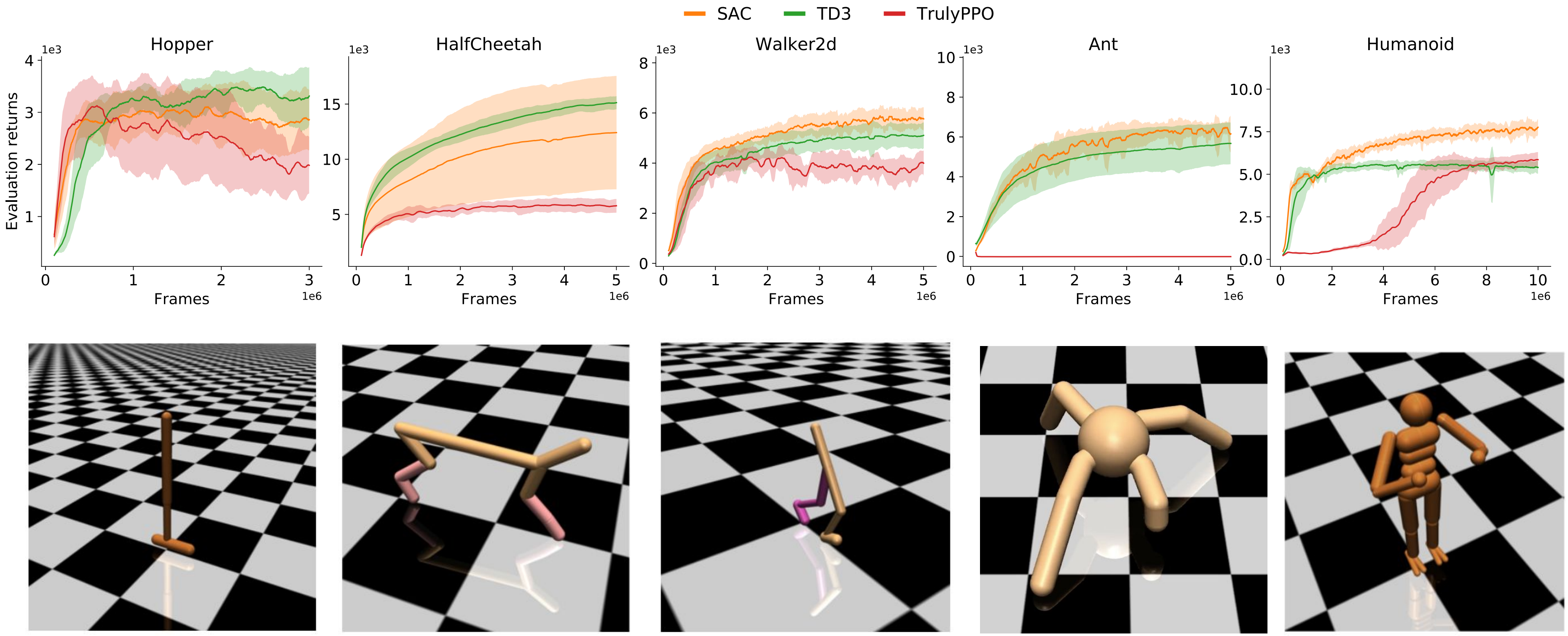
TQC's novelties

1. Incorporates **stochasticity of returns** into the overestimation control

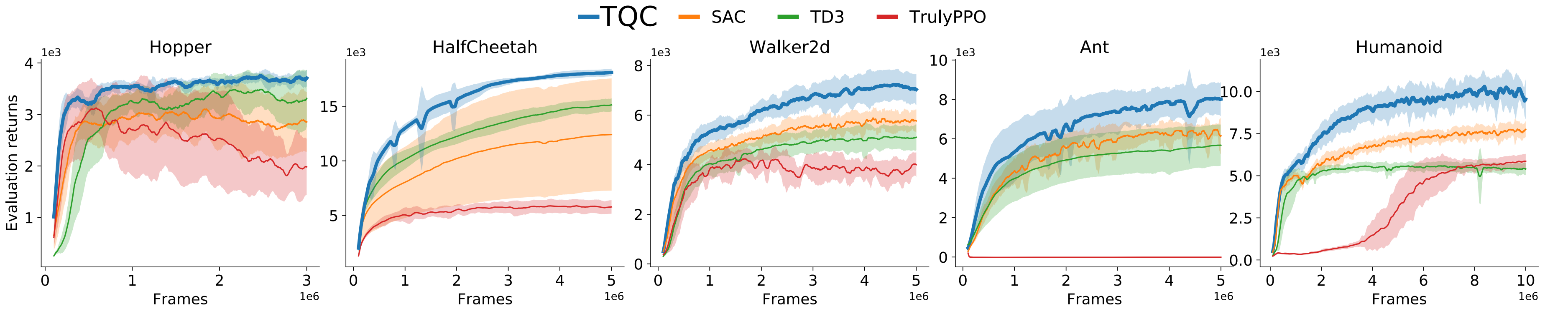
2. Provides **fine-grained** and **adjustable** level of the overestimation control

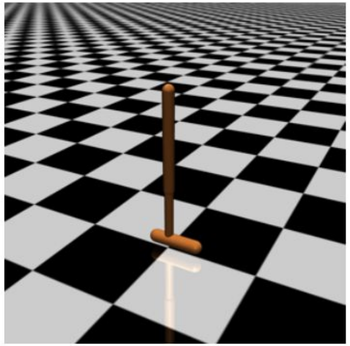
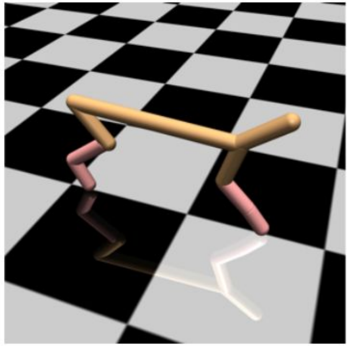
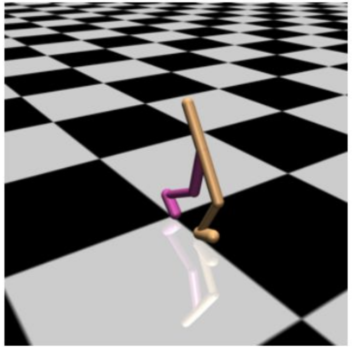
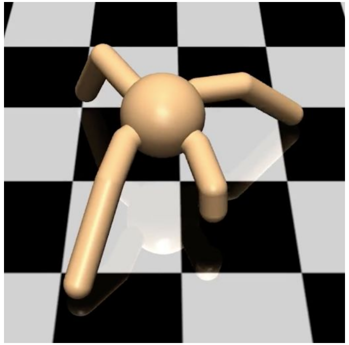
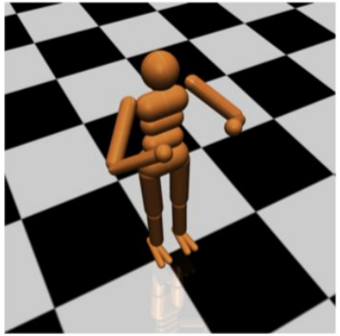
3. **Decouples** the overestimation control and the number of critics

TQC is a new SOTA on MuJoCo

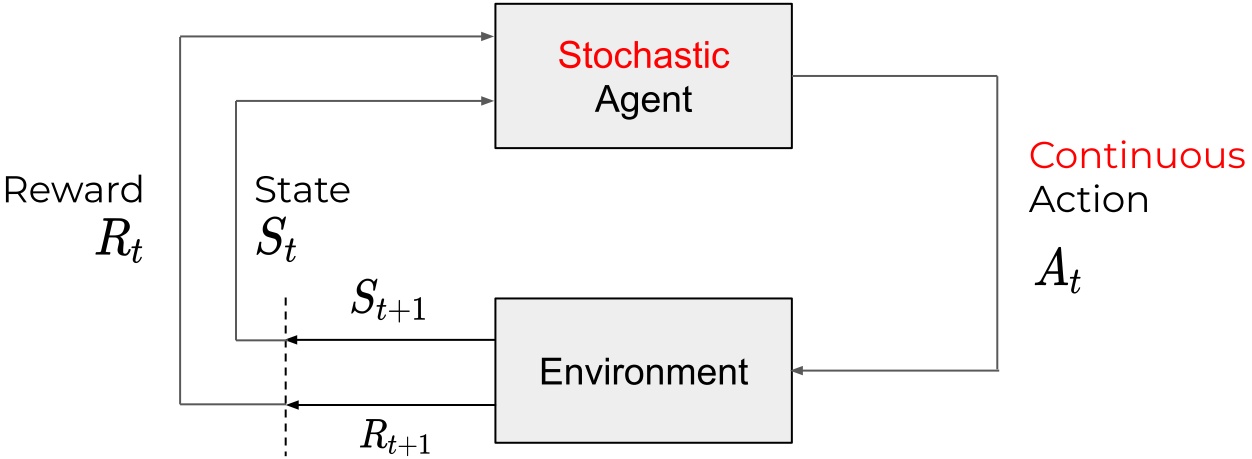


TQC is a new SOTA on MuJoCo

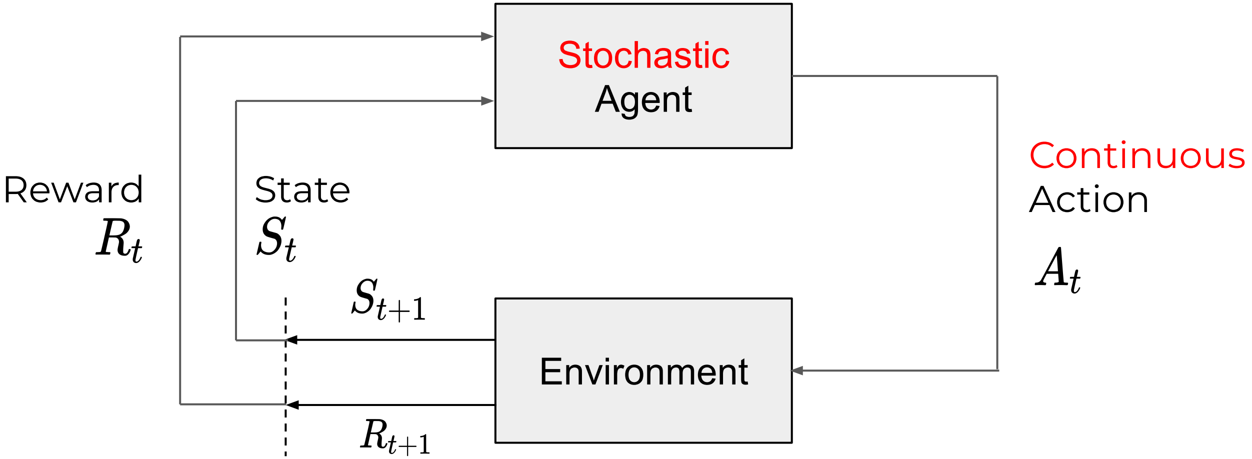


Stochastic Continuous Control in MDP

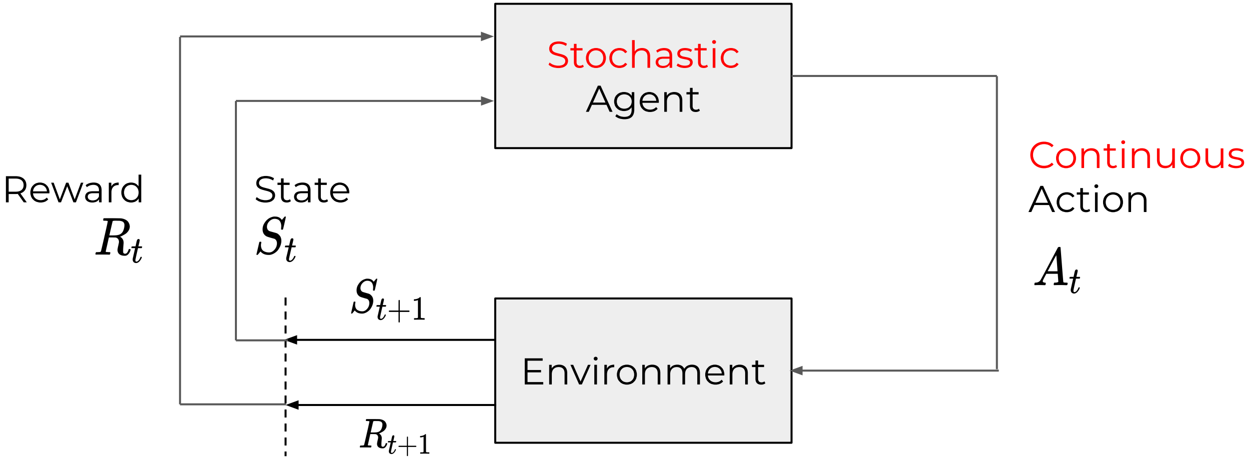


Stochastic Continuous Control in MDP



$$Q^{\pi}(s,a) = \mathbb{E}[\sum_{\tau=t}^{T} \gamma^{\tau-t} R_{\tau} | A_t = a, S_t = s]$$

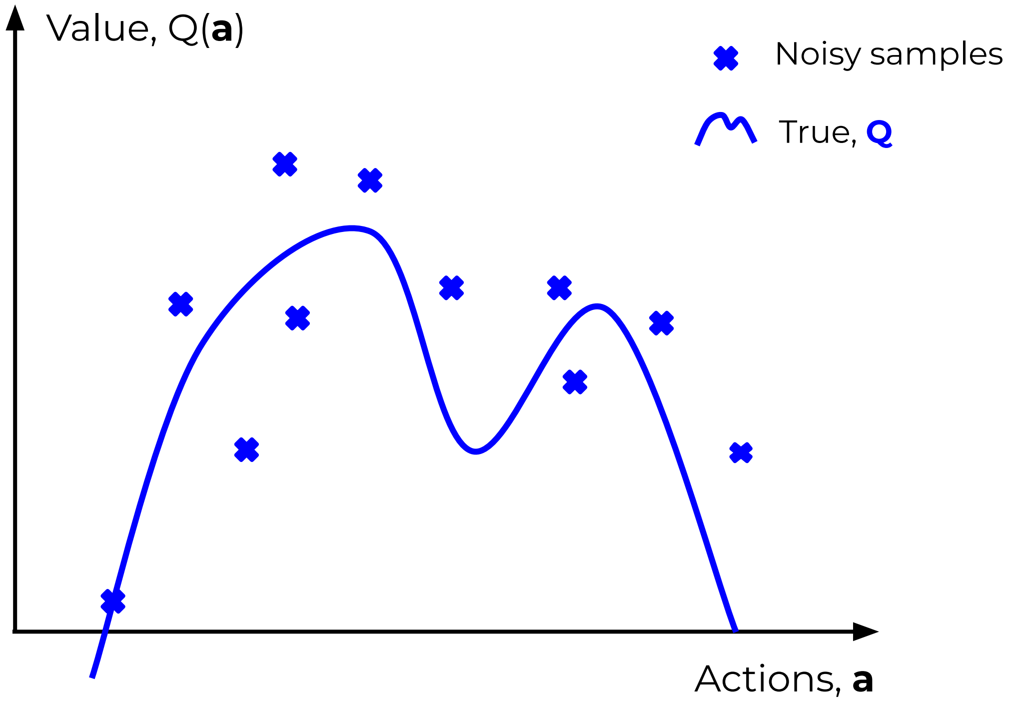
Stochastic Continuous Control in MDP



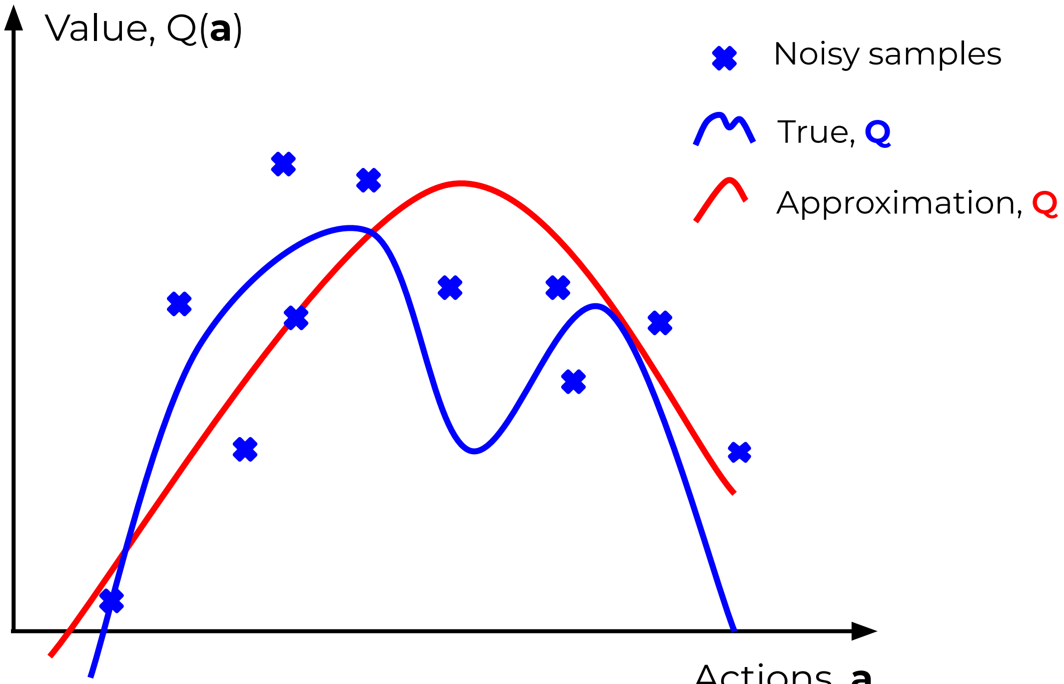
$$Q^{\pi}(s,a) = \mathbb{E}[\sum_{\tau=t}^{T} \gamma^{\tau-t} R_{\tau} | A_t = a, S_t = s]$$

$$Q^{\pi}_{\theta}(s,a) \approx Q^{\pi}(s,a) \quad \forall s, a$$

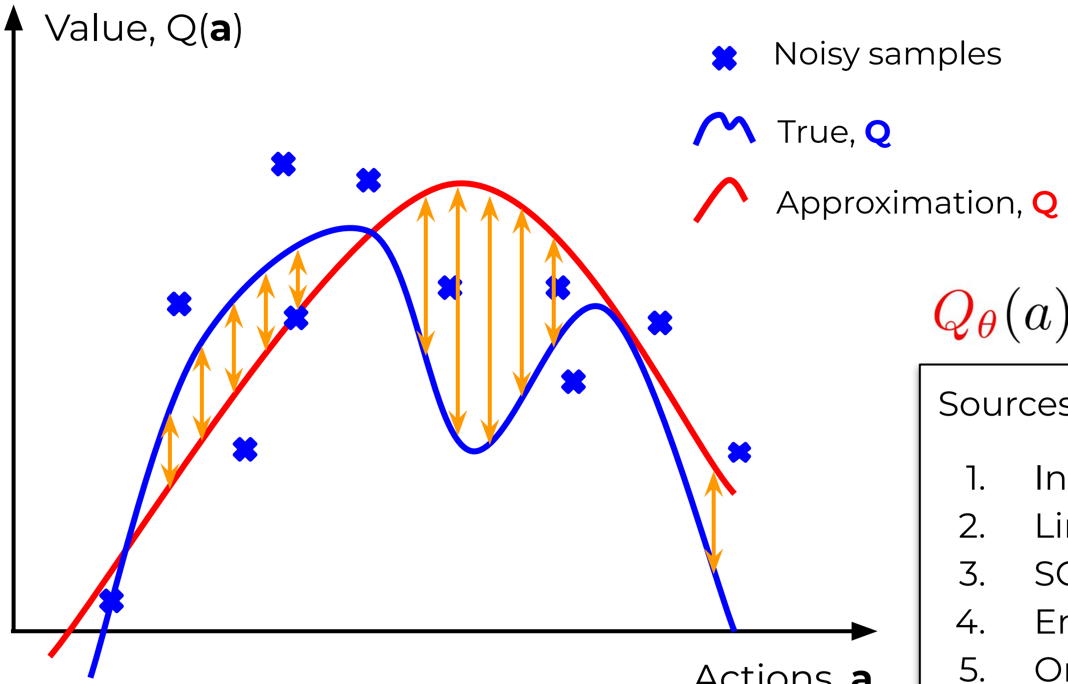
Overestimation: intuition



Overestimation: intuition



Overestimation: intuition

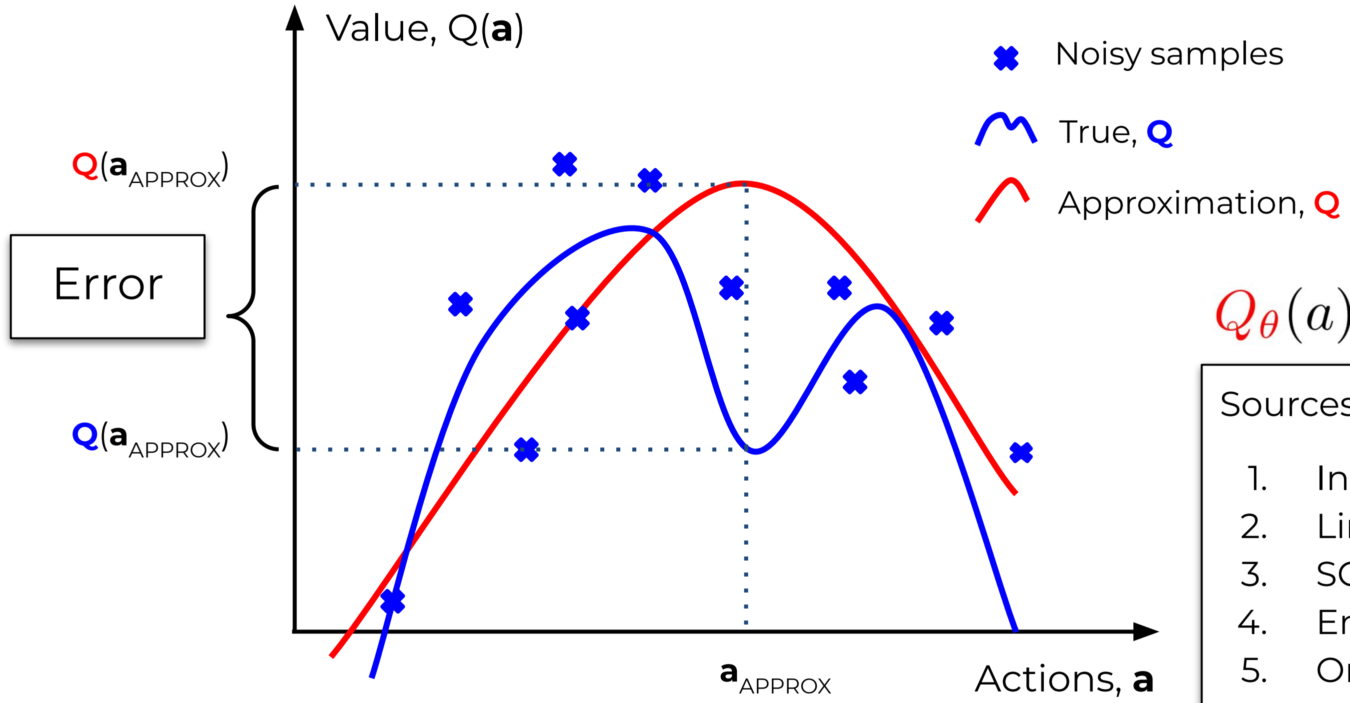


$$Q_\theta(a) = Q(a) + U(a)$$

Sources of distortion, U:

1. Insufficient data
2. Limited model capacity
3. SGD noise
4. Env's stochasticity
5. Ongoing policy changes

Overestimation: intuition



$$Q_\theta(a) = Q(a) + U(a)$$

Sources of distortion, U:

1. Insufficient data
2. Limited model capacity
3. SGD noise
4. Env's stochasticity
5. Ongoing policy changes

Overestimation: mathematical model[1]

Predicted maximum:

$$\max_a \{Q(a) + U(a)\}$$

[1]: Thrun, Sebastian, and Anton Schwartz. "Issues in using function approximation for reinforcement learning."

Overestimation: mathematical model[1]

Predicted maximum averaged over zero mean distortion:

$$\mathbb{E}_{U}\left[\max_{a}\{Q(a)+U(a)\}\right]$$

[1]: Thrun, Sebastian, and Anton Schwartz. "Issues in using function approximation for reinforcement learning."

Overestimation: mathematical model[1]

Predicted maximum averaged over zero mean distortion:

$$\mathbb{E}_U\left[\max_a \{Q(a) + U(a)\}\right] \geq \max_a \mathbb{E}_U\left[Q(a) + U(a)\right]$$

Jensen inequality

[1]: Thrun, Sebastian, and Anton Schwartz. "Issues in using function approximation for reinforcement learning."

Overestimation: mathematical model[1]

Predicted maximum averaged over zero mean distortion:

$$\mathbb{E}_U\left[\max_a\{Q(a)+U(a)\}\right] \geq \max_a \mathbb{E}_U\left[Q(a)+U(a)\right] = \max_a Q(a)$$

[1]: Thrun, Sebastian, and Anton Schwartz. "Issues in using function approximation for reinforcement learning."

Overestimation: mathematical model[1]

Predicted maximum averaged over zero mean distortion:

$$\underbrace{\mathbb{E}_U\left[\max_a \{Q(a) + U(a)\}\right]}_{\textbf{Predicted}} \geq \max_a \mathbb{E}_U[Q(a) + U(a)] = \underbrace{\max_a Q(a)}_{\textbf{True}}$$

Predicted **≥** **True**

[1]: Thrun, Sebastian, and Anton Schwartz. "Issues in using function approximation for reinforcement learning."

Overestimation: mathematical model[1]

Predicted maximum averaged over zero mean distortion:

$$\underbrace{\mathbb{E}_U\left[\max_a\{Q(a)+U(a)\}\right]}_{\textbf{Predicted}} \geq \max_a \mathbb{E}_U\left[Q(a)+U(a)\right] = \underbrace{\max_a Q(a)}_{\textbf{True}}$$

Predicted **≥** **True**

1. Policy exploits critic's erroneous estimates
2. TD-learning propagates estimation errors
3. Positive feedback loop may occur

[1]: Thrun, Sebastian, and Anton Schwartz. "Issues in using function approximation for reinforcement learning."

Soft Actor Critic[2]

[2]: Haarnoja, Tuomas, et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor."

Soft Actor Critic[2]

Soft Policy Evaluation:

$$(s, a, r, s') \sim \mathcal{D}, \quad a' \sim \pi(s')$$

[2]: Haarnoja, Tuomas, et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor."

Soft Actor Critic[2]

Soft Policy Evaluation:

$$(s, a, r, s') \sim \mathcal{D}, \quad a' \sim \pi(s')$$

$$y(s,a) = r + \gamma\Big(Q_{\bar{\theta}}(s', a') + \alpha\mathcal{H}[\pi(s')]\Big)$$

[2]: Haarnoja, Tuomas, et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor."

Soft Actor Critic[2]

Soft Policy Evaluation:

$$(s, a, r, s') \sim \mathcal{D}, \quad a' \sim \pi(s')$$

$$y(s,a) = r + \gamma\Big(Q_{\bar{\theta}}(s', a') + \alpha\mathcal{H}[\pi(s')]\Big)$$

$$[Q_\theta(s,a) - y(s,a)]^2 \to \min_\theta$$

[2]: Haarnoja, Tuomas, et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor."

Soft Actor Critic[2]

Overestimation alleviation
(Clipped Double Estimate[3]):

Soft Policy Evaluation:

$$(s, a, r, s') \sim \mathcal{D}, \quad a' \sim \pi(s')$$

$$y(s,a) = r + \gamma\Big(Q_{\bar{\theta}}(s', a') + \alpha\mathcal{H}[\pi(s')]\Big)$$

$$[Q_\theta(s,a) - y(s,a)]^2 \to \min_\theta$$

[2]: Haarnoja, Tuomas, et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor."
[3]: Scott Fujimoto, Herke van Hoof, David Meger “Addressing Function Approximation Error in Actor-Critic Methods”

Soft Actor Critic[2]

Soft Policy Evaluation:

$$(s, a, r, s') \sim \mathcal{D}, \quad a' \sim \pi(s')$$

$$y(s,a) = r + \gamma\Big(Q_{\bar{\theta}}(s', a') + \alpha\mathcal{H}[\pi(s')]\Big)$$

$$[Q_\theta(s,a) - y(s,a)]^2 \to \min_\theta$$

Overestimation alleviation
(Clipped Double Estimate[3]):

1.

$$\Big[Q_{\theta_1}(s,a) - y(s,a)\Big]^2 \to \min_{\theta_1}$$

$$\Big[Q_{\theta_2}(s,a) - y(s,a)\Big]^2 \to \min_{\theta_2}$$

[2]: Haarnoja, Tuomas, et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor."
[3]: Scott Fujimoto, Herke van Hoof, David Meger “Addressing Function Approximation Error in Actor-Critic Methods”

Soft Actor Critic[2]

Soft Policy Evaluation:

$$(s, a, r, s') \sim \mathcal{D}, \quad a' \sim \pi(s')$$

$$y(s, a) = r + \gamma\Big(Q_{\bar{\theta}}(s', a') + \alpha \mathcal{H}[\pi(s')]\Big)$$

$$[Q_\theta(s, a) - y(s, a)]^2 \to \min_\theta$$

Overestimation alleviation (Clipped Double Estimate[3]):

1. $$\Big[Q_{\theta_1}(s, a) - y(s, a)\Big]^2 \to \min_{\theta_1}$$
$$\Big[Q_{\theta_2}(s, a) - y(s, a)\Big]^2 \to \min_{\theta_2}$$

2. $$y(s, a) = r + \gamma\Big(\cancel{Q_{\bar{\theta}}(s', a')} + \alpha \mathcal{H}[\pi(s')]\Big)$$

$$\min\left(Q_{\bar{\theta}_1}(s', a'), Q_{\bar{\theta}_2}(s', a')\right)$$

[2]: Haarnoja, Tuomas, et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor."
[3]: Scott Fujimoto, Herke van Hoof, David Meger “Addressing Function Approximation Error in Actor-Critic Methods”

Limitations:

- **Coarse** bias control
- **Wasteful** aggregation

Solution:

Truncated Quantile Critics

Overestimation alleviation (Clipped Double Estimate[3]):

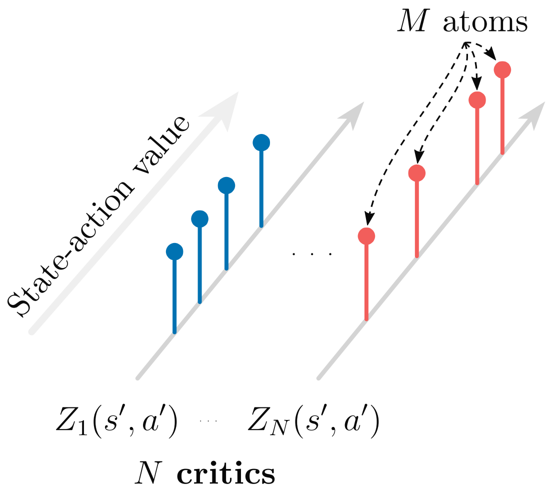
1. $$\left[Q_{\theta_1}(s,a) - y(s,a)\right]^2 \to \min_{\theta_1}$$
   $$\left[Q_{\theta_2}(s,a) - y(s,a)\right]^2 \to \min_{\theta_2}$$

2. $$y(s,a) = r + \gamma\Big(\cancel{Q_{\bar{\theta}}(s',a')} + \alpha\mathcal{H}[\pi(s')]\Big)$$

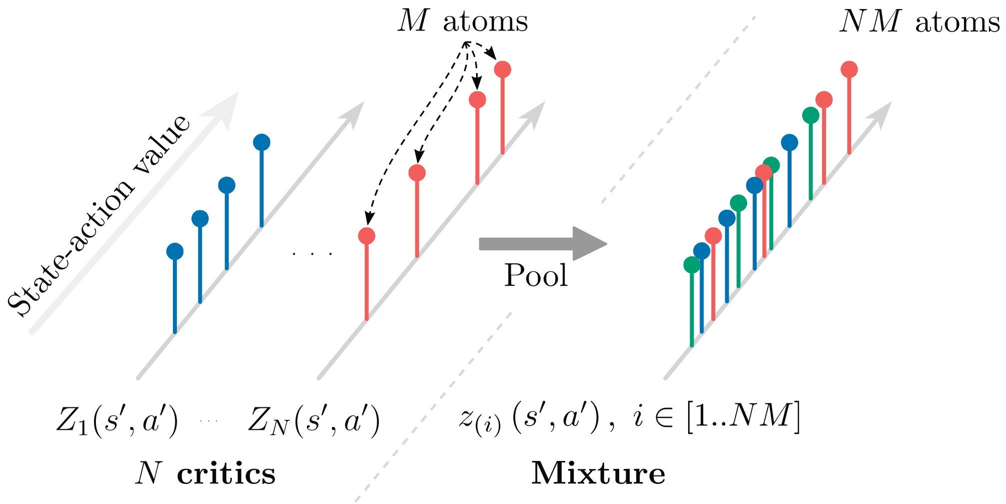
$$\min\left(Q_{\bar{\theta}_1}(s',a'), Q_{\bar{\theta}_2}(s',a')\right)$$

[2]: Haarnoja, Tuomas, et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor."
[3]: Scott Fujimoto, Herke van Hoof, David Meger “Addressing Function Approximation Error in Actor-Critic Methods”

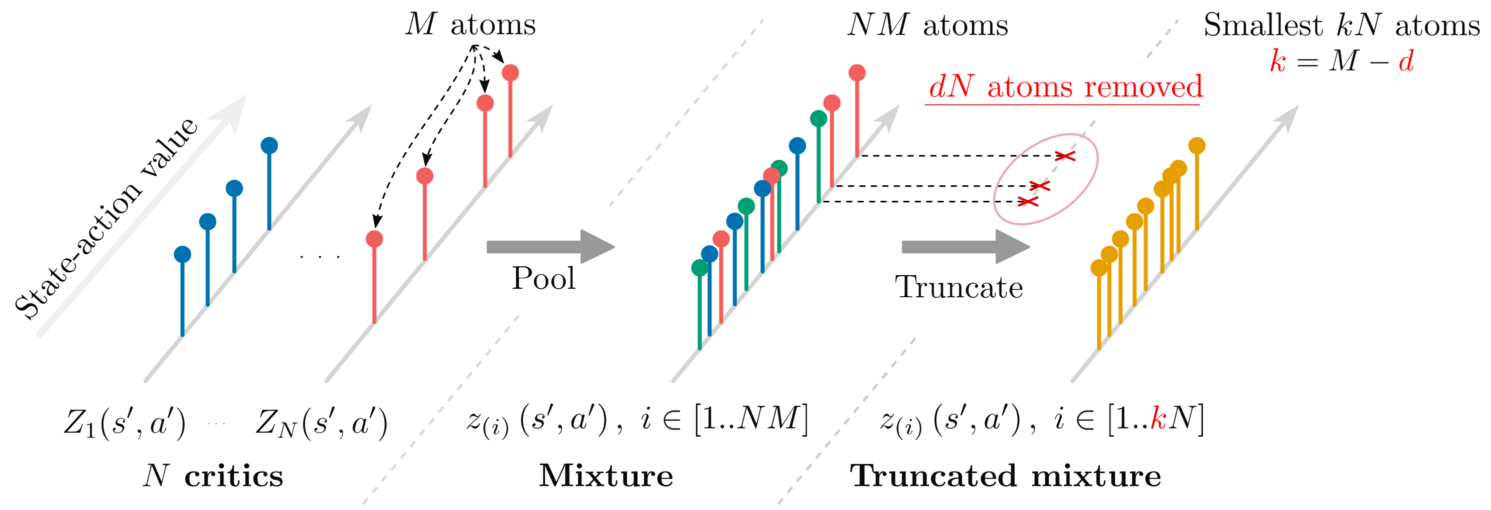
TQC step 1: Prediction of N distributions



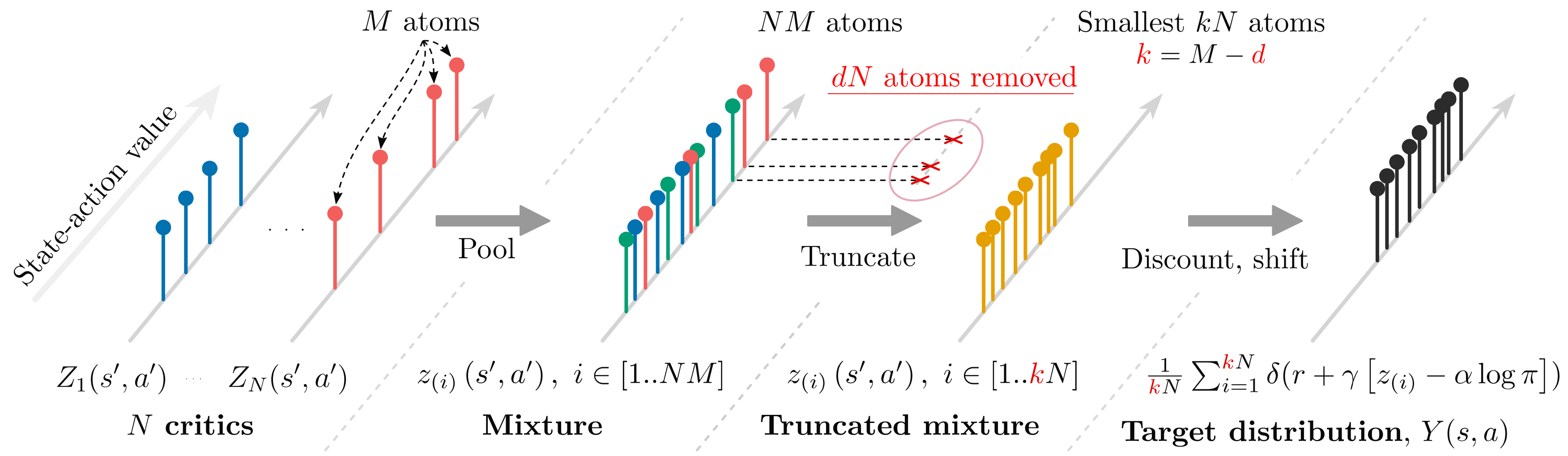
TQC step 2: Pooling



TQC step 3: Truncation

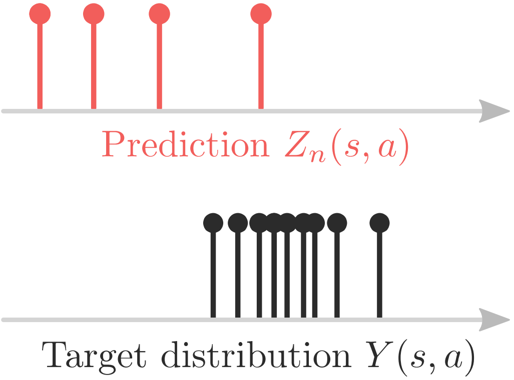


TQC step 4: Discounting and Shifting



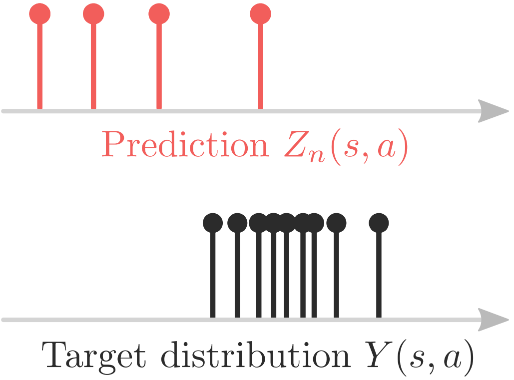
Training

For each Z-network:



Training

For each Z-network:



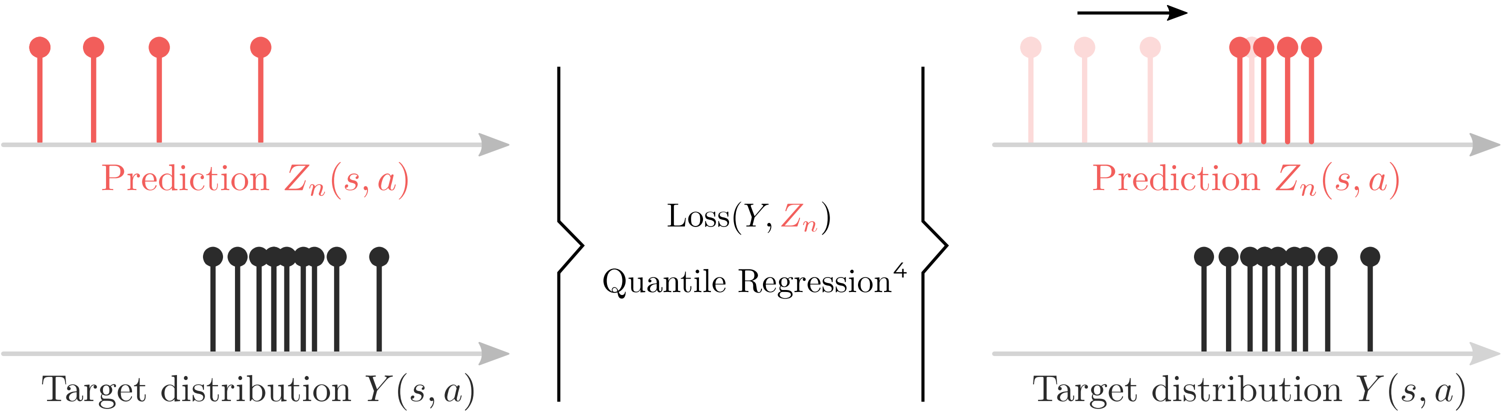
$\text{Loss}(Y, Z_n)$

Quantile Regression[4]

[4]: Dabney, Will, et al. "Distributional reinforcement learning with quantile regression." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

Training

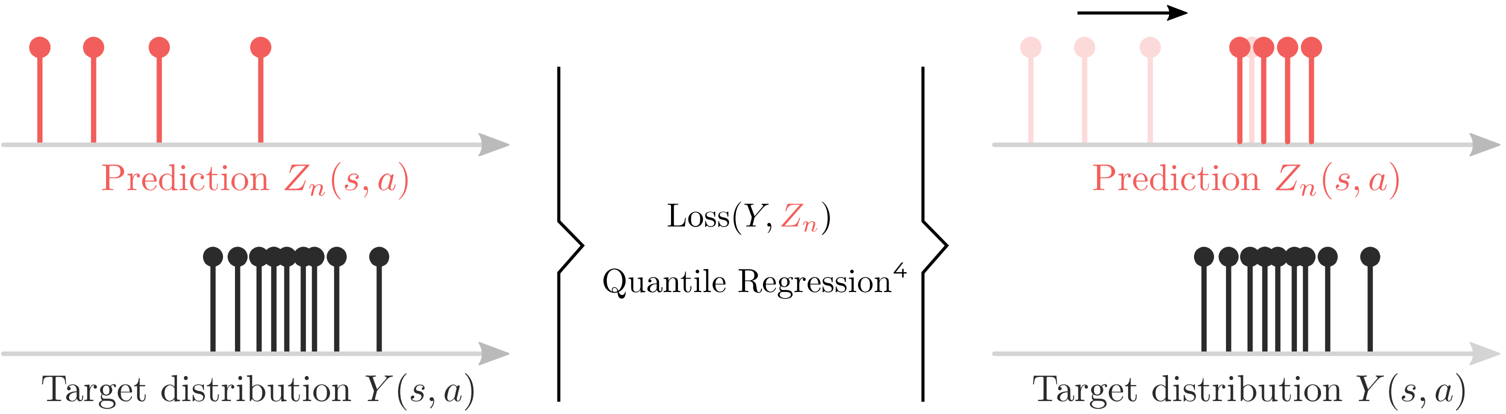
For each Z-network:



[4]: Dabney, Will, et al. "Distributional reinforcement learning with quantile regression." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

Training

For each Z-network:



Policy:

Maximizes nontruncated average of all atoms of the mixture

[4]: Dabney, Will, et al. "Distributional reinforcement learning with quantile regression." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

Our contribution: Truncated Quantile Critics

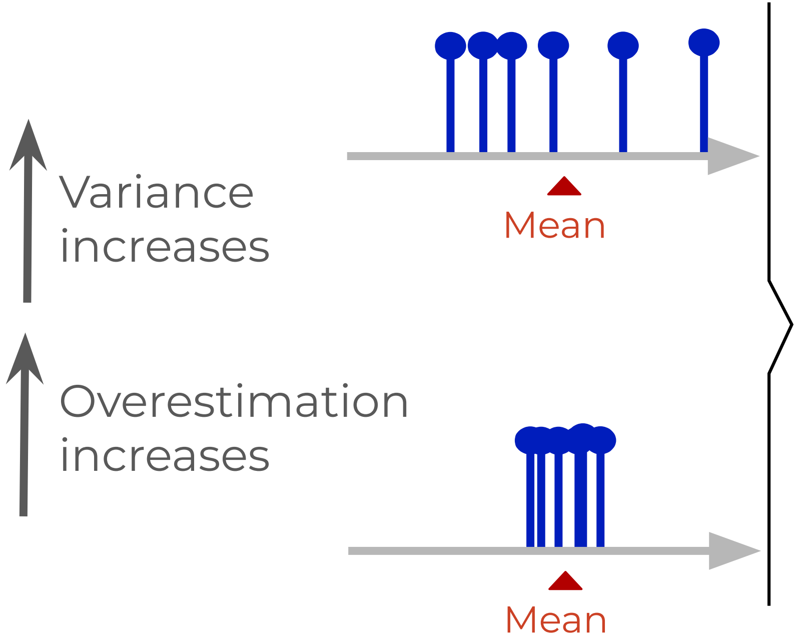
1. **Uses return stochasticity for overestimation control**

   A novel direction: interplay between overestimation and stochasticity

Our contribution: Truncated Quantile Critics

1. **Uses return stochasticity for overestimation control**

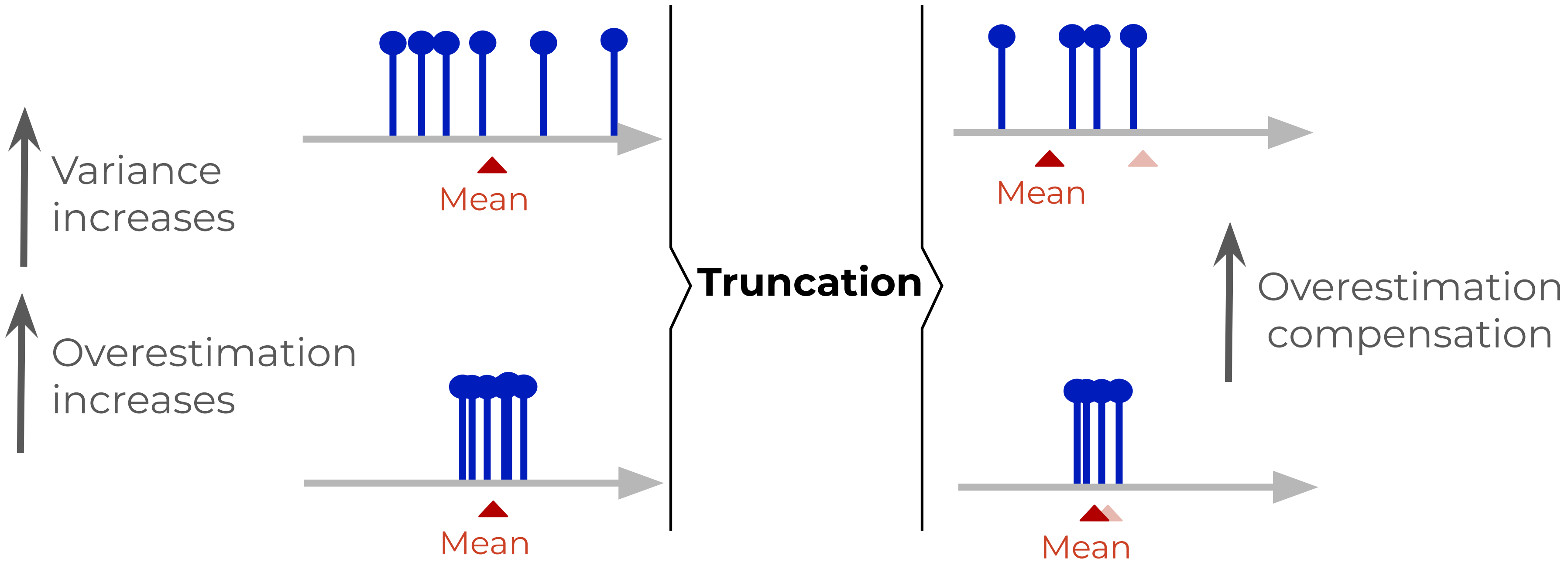
   A novel direction: interplay between overestimation and stochasticity



Our contribution: Truncated Quantile Critics

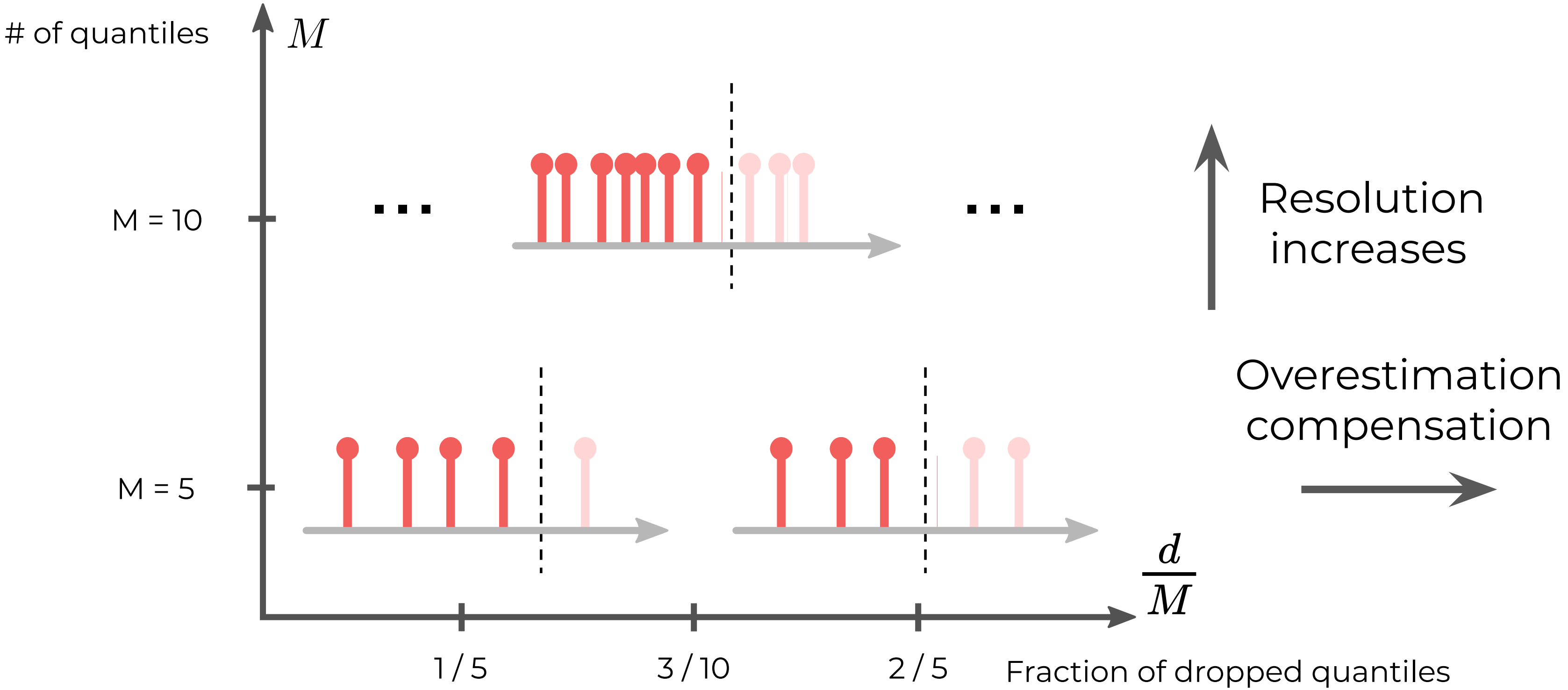
1. **Uses return stochasticity for overestimation control**

   A novel direction: interplay between overestimation and stochasticity



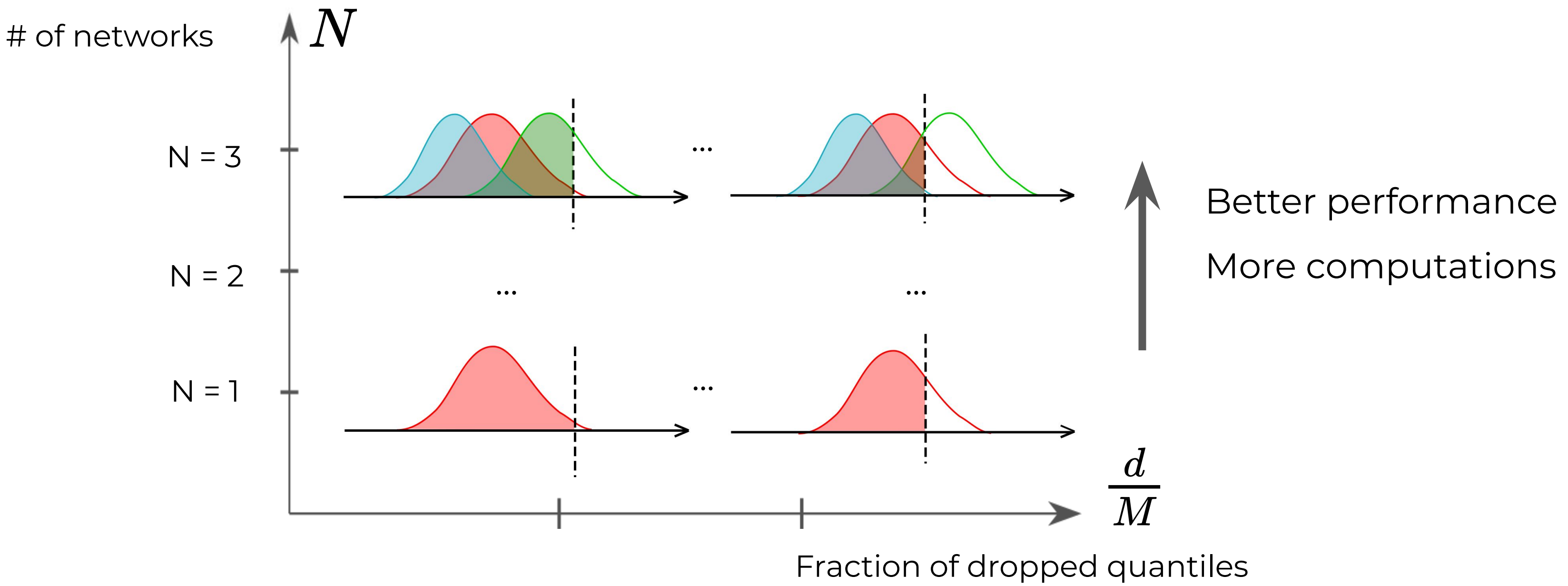
Our contribution: Truncated Quantile Critics

1. Uses return stochasticity for overestimation control
2. **Method provides adjustable and fine-grained overestimation bias control**



Our contribution: Truncated Quantile Critics

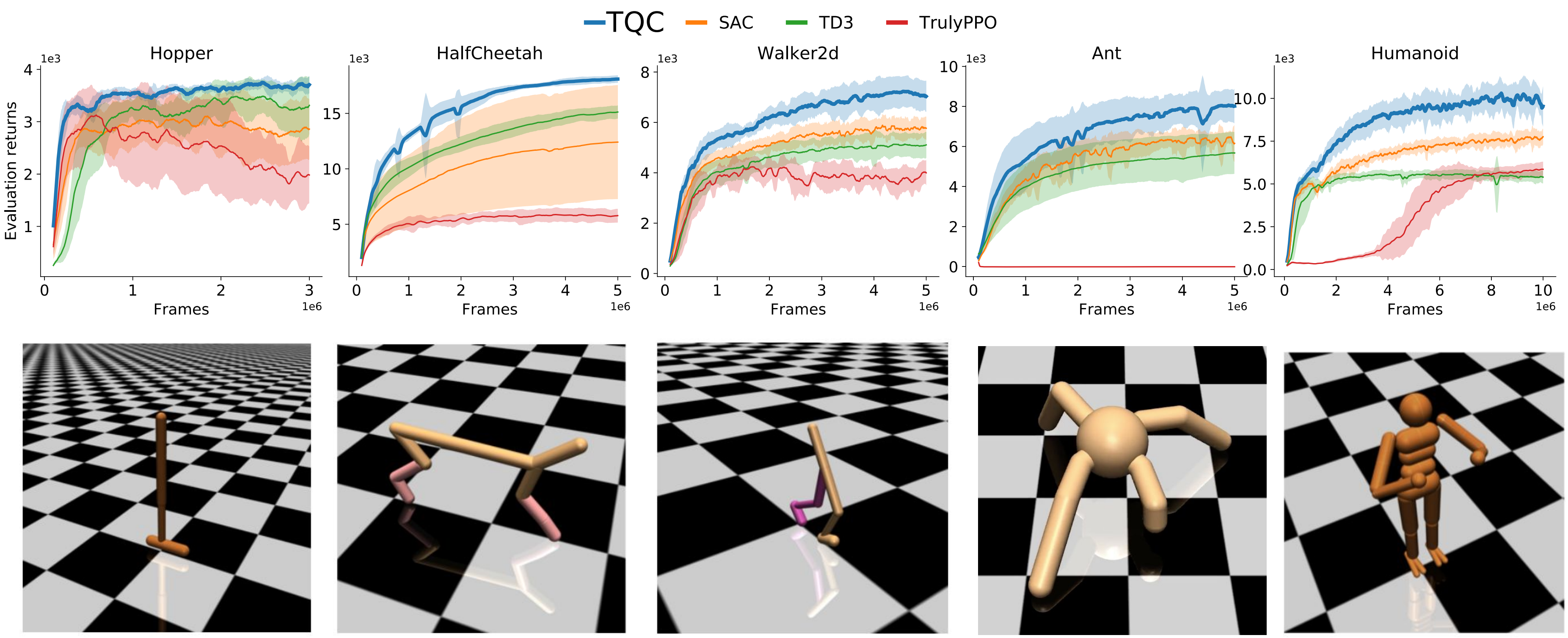
1. Uses return stochasticity for overestimation control
2. Method provides adjustable and fine-grained overestimation bias control
3. **Decouples overestimation control and number of approximators**



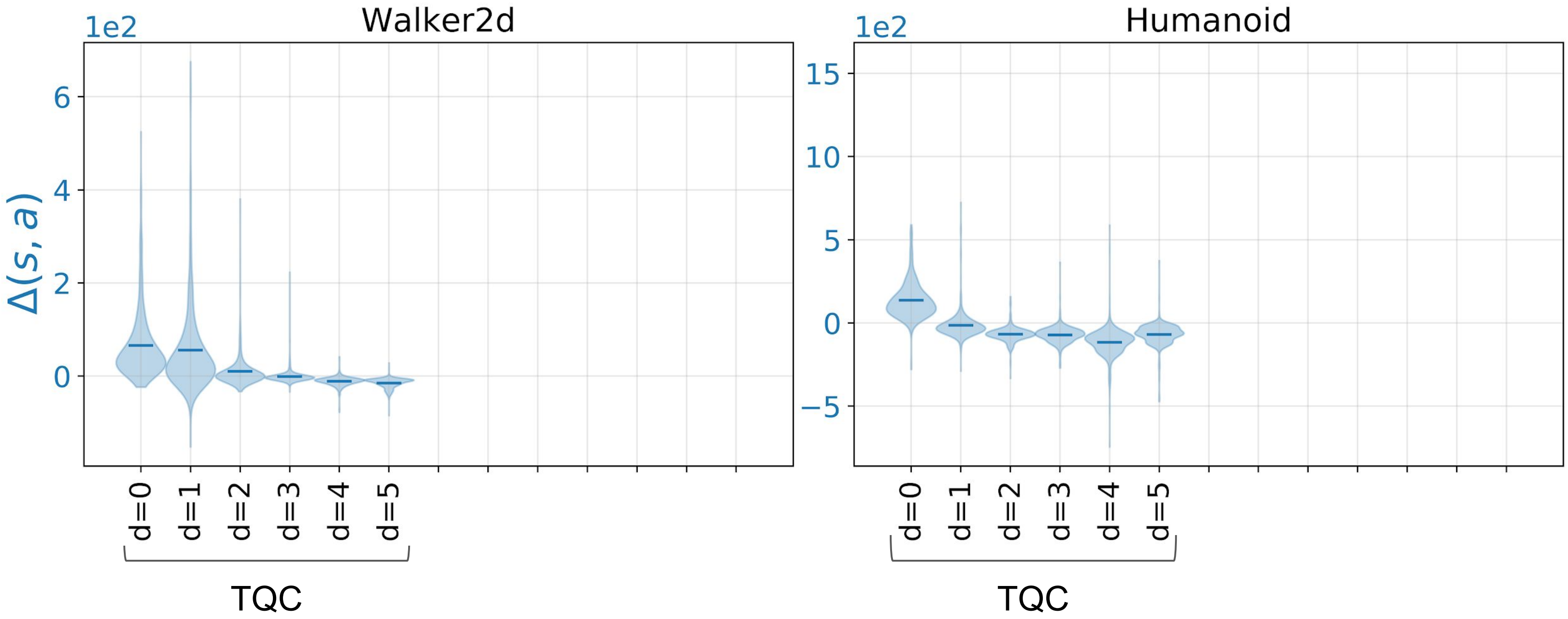
Our contribution: Truncated Quantile Critics

1. Uses return stochasticity for overestimation control
2. Method provides fine-grained overestimation bias control
3. Decouples overestimation control and multiplicity of approximators
4. **New SOTA on MuJoCo locomotion suite**

Substantial improvement on all environments



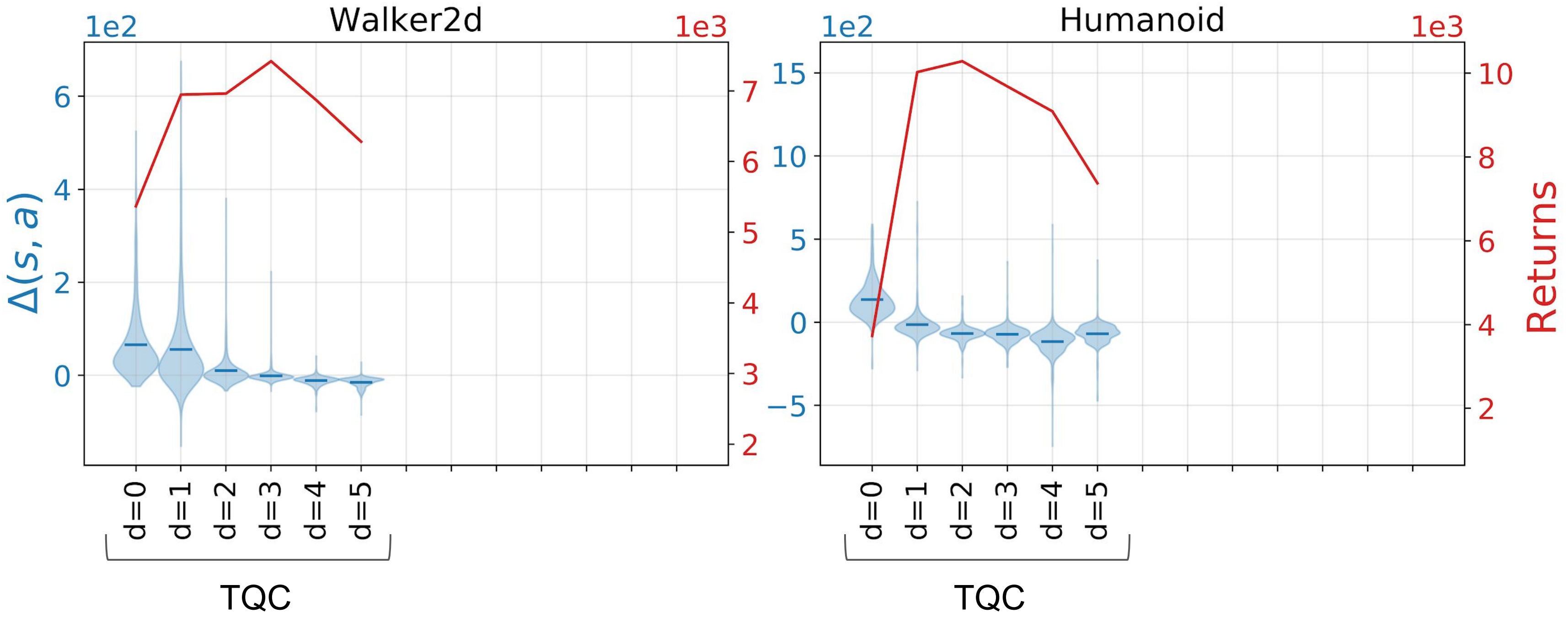
Overestimation measurement



$$\Delta(s,a) = Q_\theta(s,a) - \hat{Q}(s,a) \qquad s, a \sim \rho^\pi$$

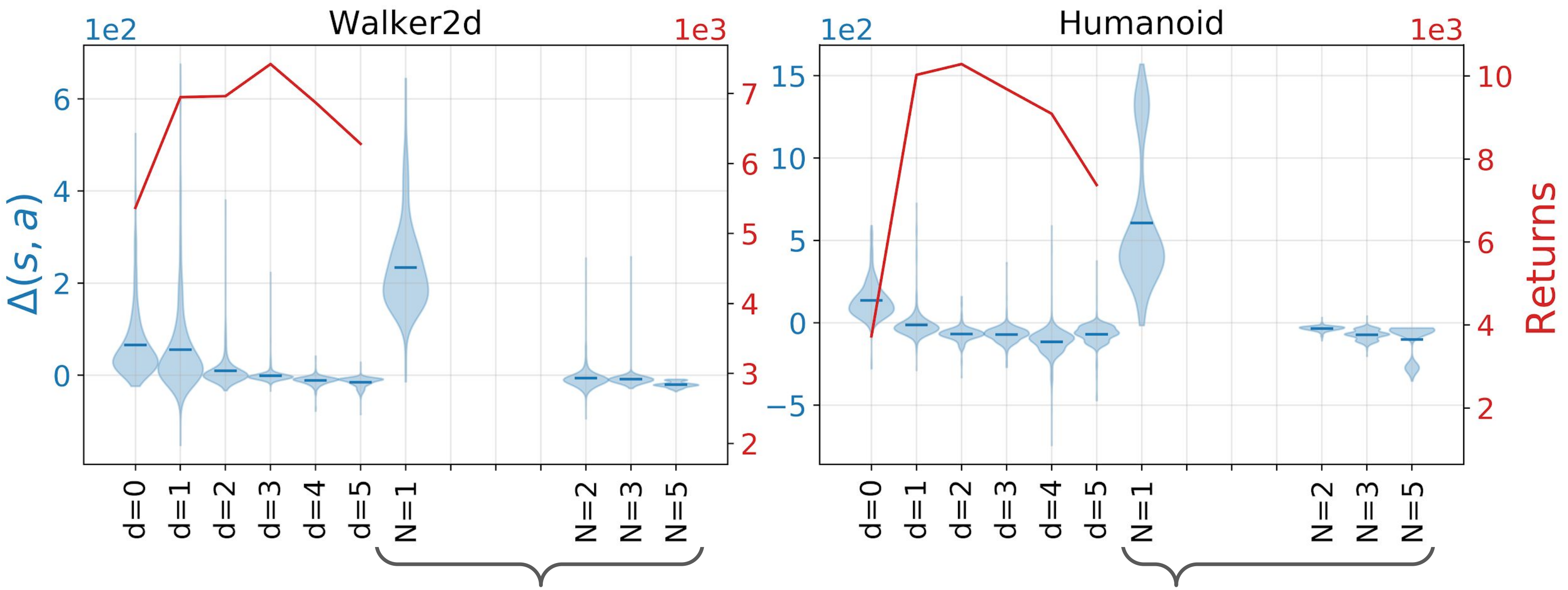
$\hat{Q}(s,a)$ – Monte-Carlo estimate

Overestimation measurement



There is a clear optimum in terms of performance.

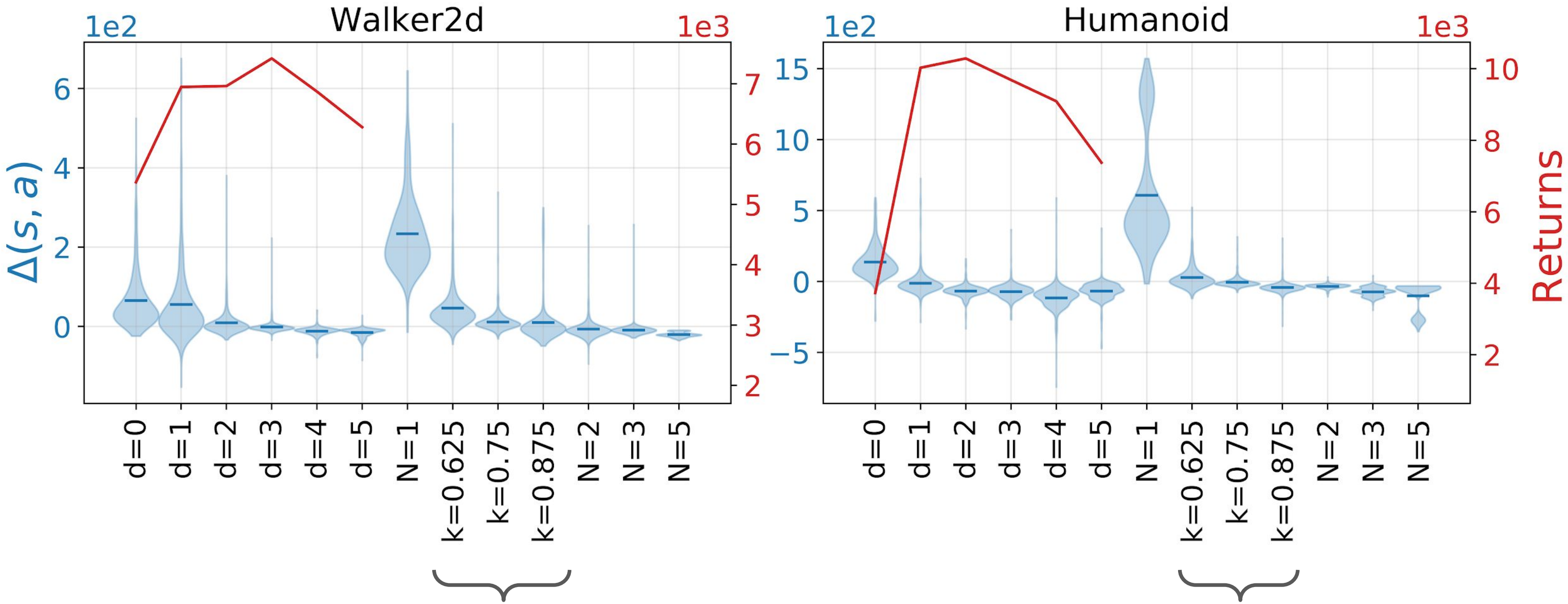
Overestimation measurement



SAC with min over multiple critics N=1..5

$$y(s,a) = r + \gamma \min(Q_1, \ldots, Q_N)$$

Overestimation measurement

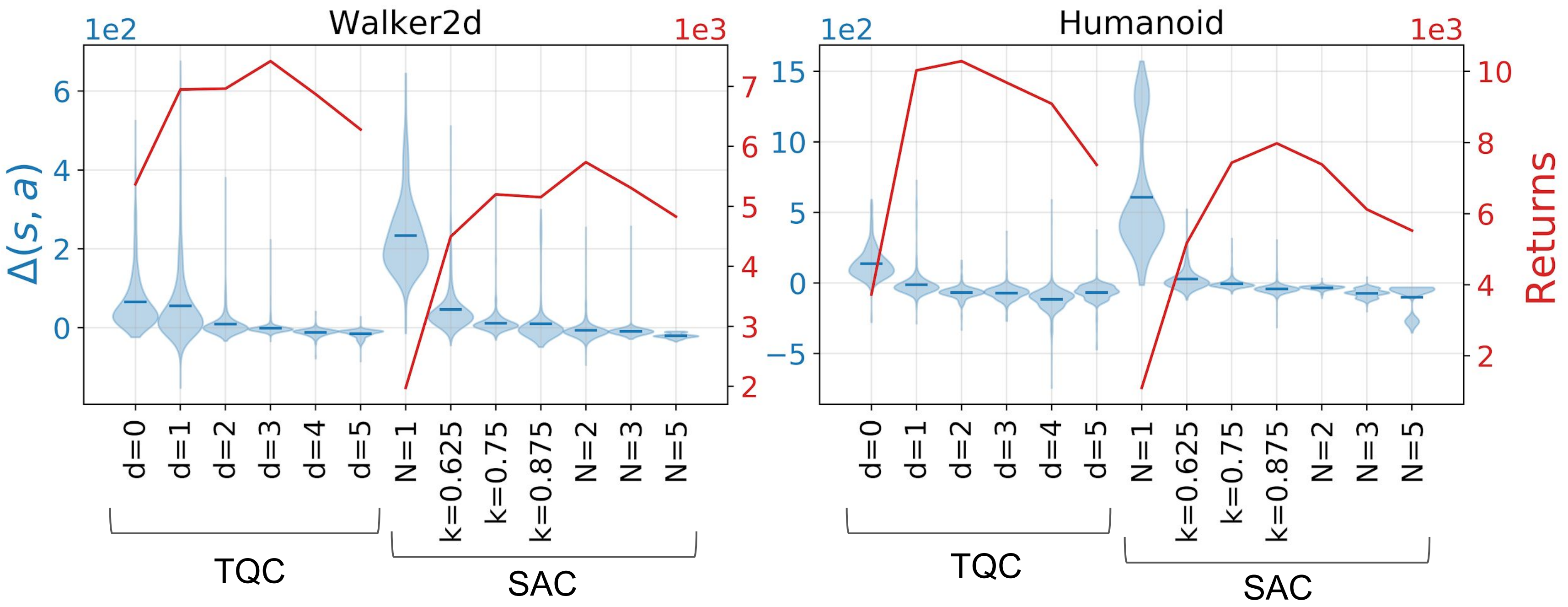


SAC with linear combination of minimum and maximum Q-functions[4]:

$$y(s, a) = r + \gamma(k \min(Q_1, Q_2) + (1 - k) \max(Q_1, Q_2))$$

[4] Fujimoto, Scott, David Meger, and Doina Precup. "Off-Policy Deep Reinforcement Learning without Exploration."

Overestimation measurement



Sources of TQC performance:

- Distributional critics
- Different procedure of overestimation correction
- Benefits from larger networks

Links

1. Arxiv Paper https://arxiv.org/abs/2005.04269
2. Method Page https://bayesgroup.github.io/tqc/
3. Tensorflow Code https://github.com/bayesgroup/tqc
4. Pytorch Code https://github.com/bayesgroup/tqc_pytorch
5. Video https://www.youtube.com/watch?v=idp4k1L9UhM