Boosting Soft Q-Learning by Bounding

Jacob Adamczyk, *Volodymyr Makarenko*, *Stas Tiomkin*, *Rahul Kulkarni*

Soft Q-Learning

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} V^*(\mathrm{s}')$$

$$V^*(s) = \beta^{-1} \log \mathbb{E}_{a \sim \pi_0} \exp \beta Q^*(s', a')$$

New Bounds (Intuition)

$$\left|Q^*(s,a) - BQ(s,a)\right| \leq \mathcal{O}\left(H\sqrt{\mathcal{L}}\right)$$

New Bounds (Intuition)

Arbitrary function

$$\left|Q^*(s,a) - BQ(s,a)\right| \leq \mathcal{O}\left(H\sqrt{\mathcal{L}}\right)$$

Bellman iteration

One iteration of Bellman produces double-sided bounds on Q*, with error scaling as the Bellman residual

New Bounds

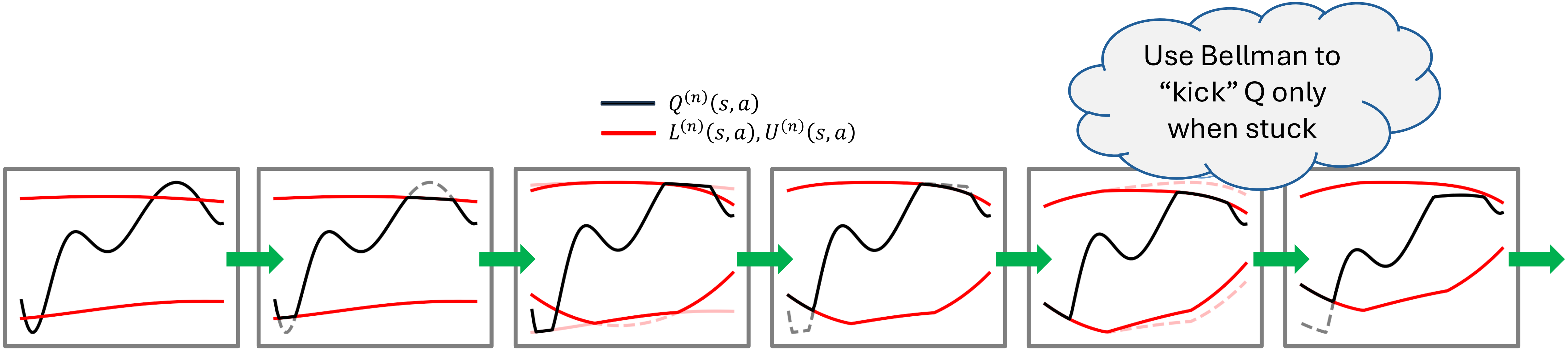
Theorem 1. *Consider an entropy-regularized MDP* $\langle \mathcal{S}, \mathcal{A}, p, r, \gamma, \beta, \pi_0 \rangle$ *with optimal value function* $Q^*(s,a)$*. Let any bounded function* $Q(s,a)$ *be given. Denote the corresponding state-value function as* $V(s) \doteq 1/\beta \log \mathbb{E}_{a\sim\pi_0} \exp \beta Q(s,a)$*. Then,* $Q^*(s,a)$ *is bounded by:*

$$r(s,a) + \gamma \left(\mathop{\mathbb{E}}_{s'\sim p} V(s') + \frac{\inf \Delta}{1-\gamma} \right) \leq Q^*(s,a) \leq r(s,a) + \gamma \left(\mathop{\mathbb{E}}_{s'\sim p} V(s') + \frac{\sup \Delta}{1-\gamma} \right) \quad (2)$$

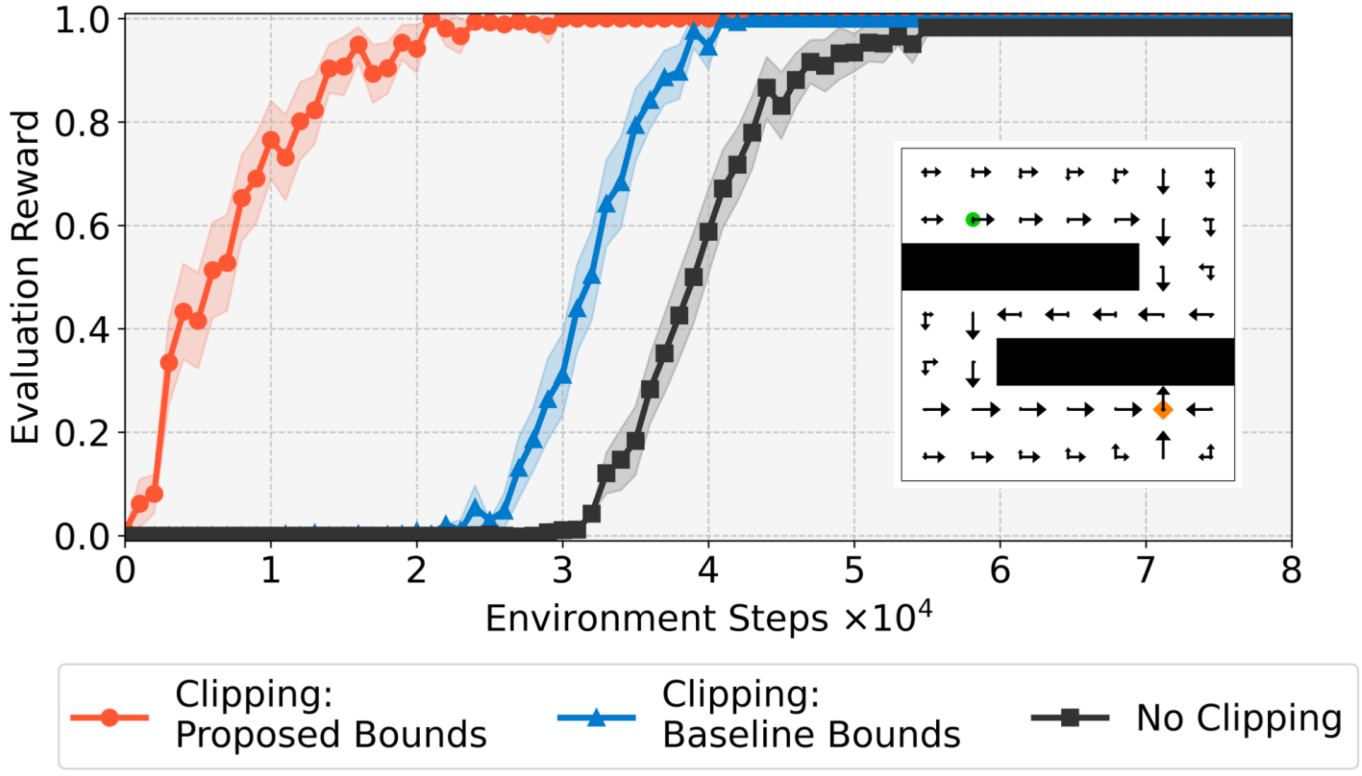
where

$$\Delta(s,a) \doteq r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim p} V(s') - Q(s,a).$$

Q-Learning by Bounding



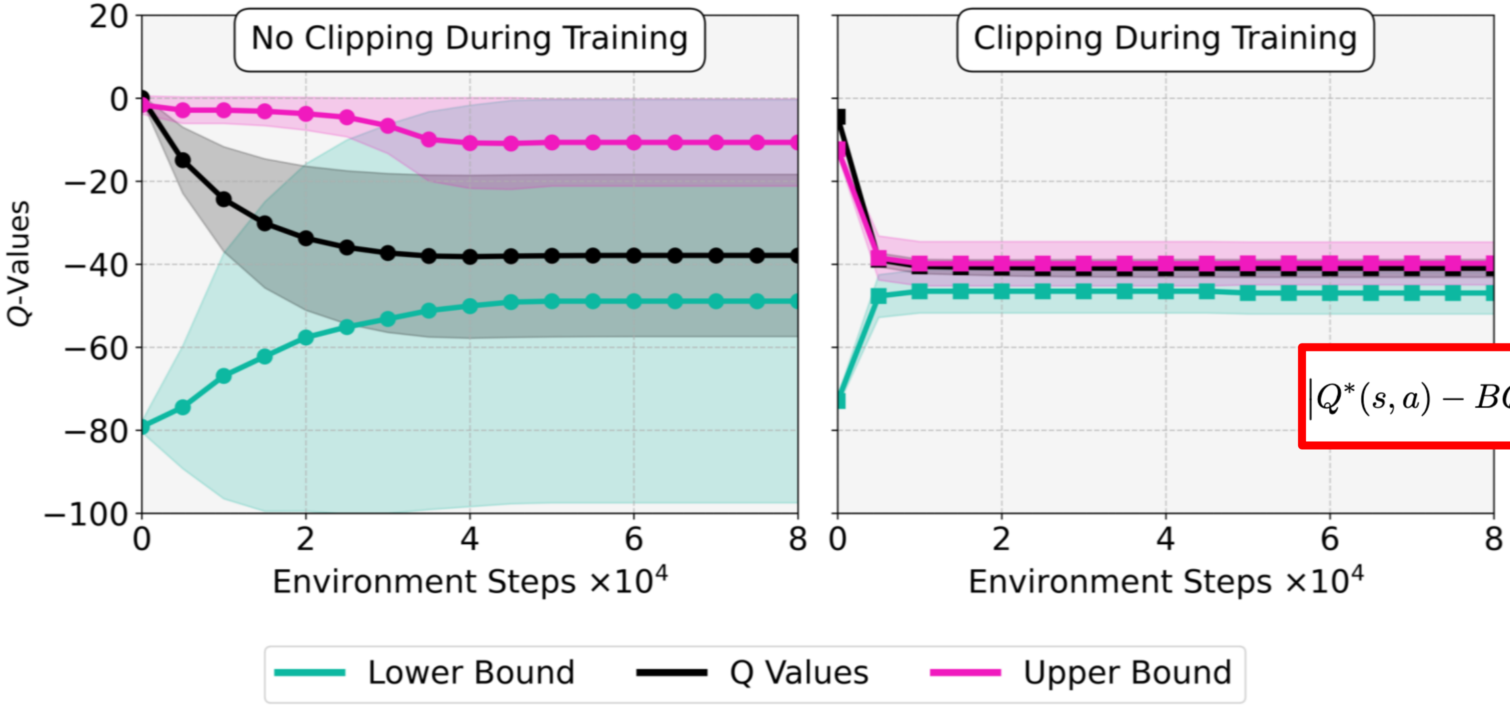
Clipping During Training



$$\left|Q^*(s,a) - BQ(s,a)\right| \leq \mathcal{O}\left(H\sqrt{\mathcal{L}}\right)$$

$$Q^* \in \left(\frac{r_{min}}{1-\gamma}, \frac{r_{max}}{1-\gamma}\right)$$

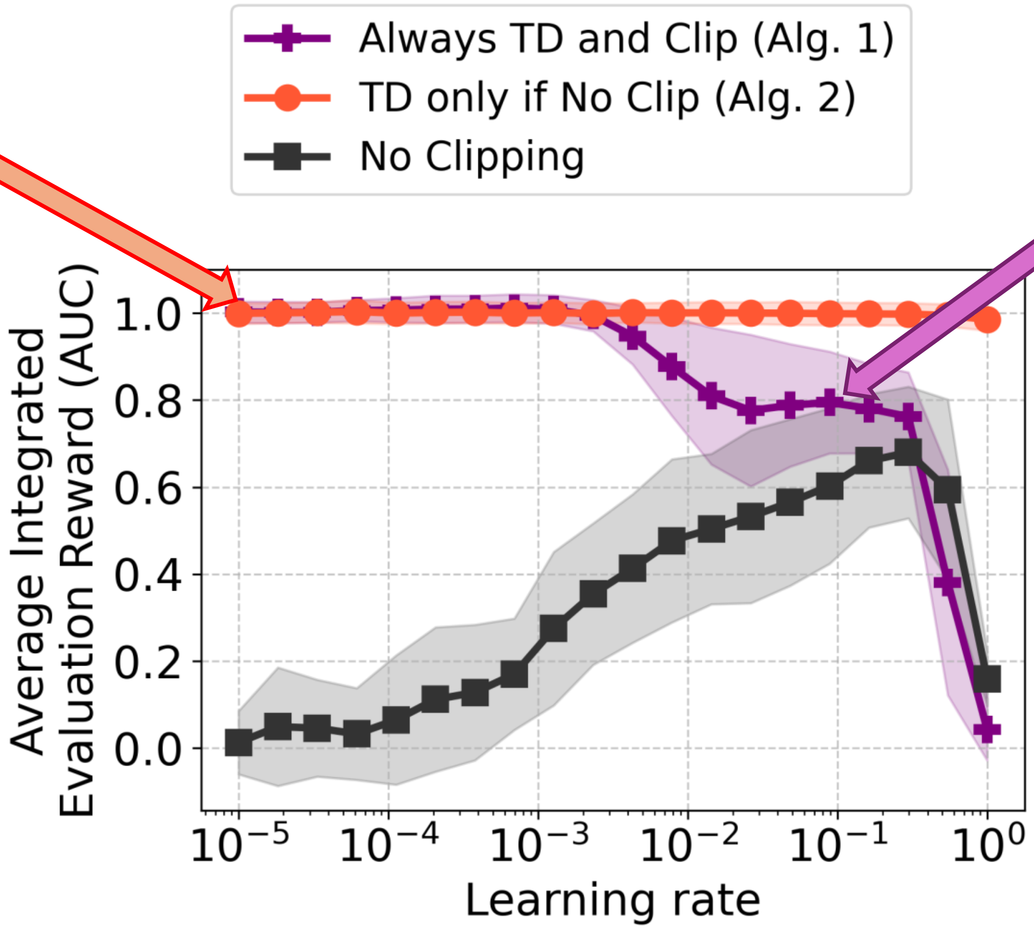
Clipping During Training



$$\left|Q^*(s,a) - BQ(s,a)\right| \leq \mathcal{O}\left(H\sqrt{\mathcal{L}}\right)$$

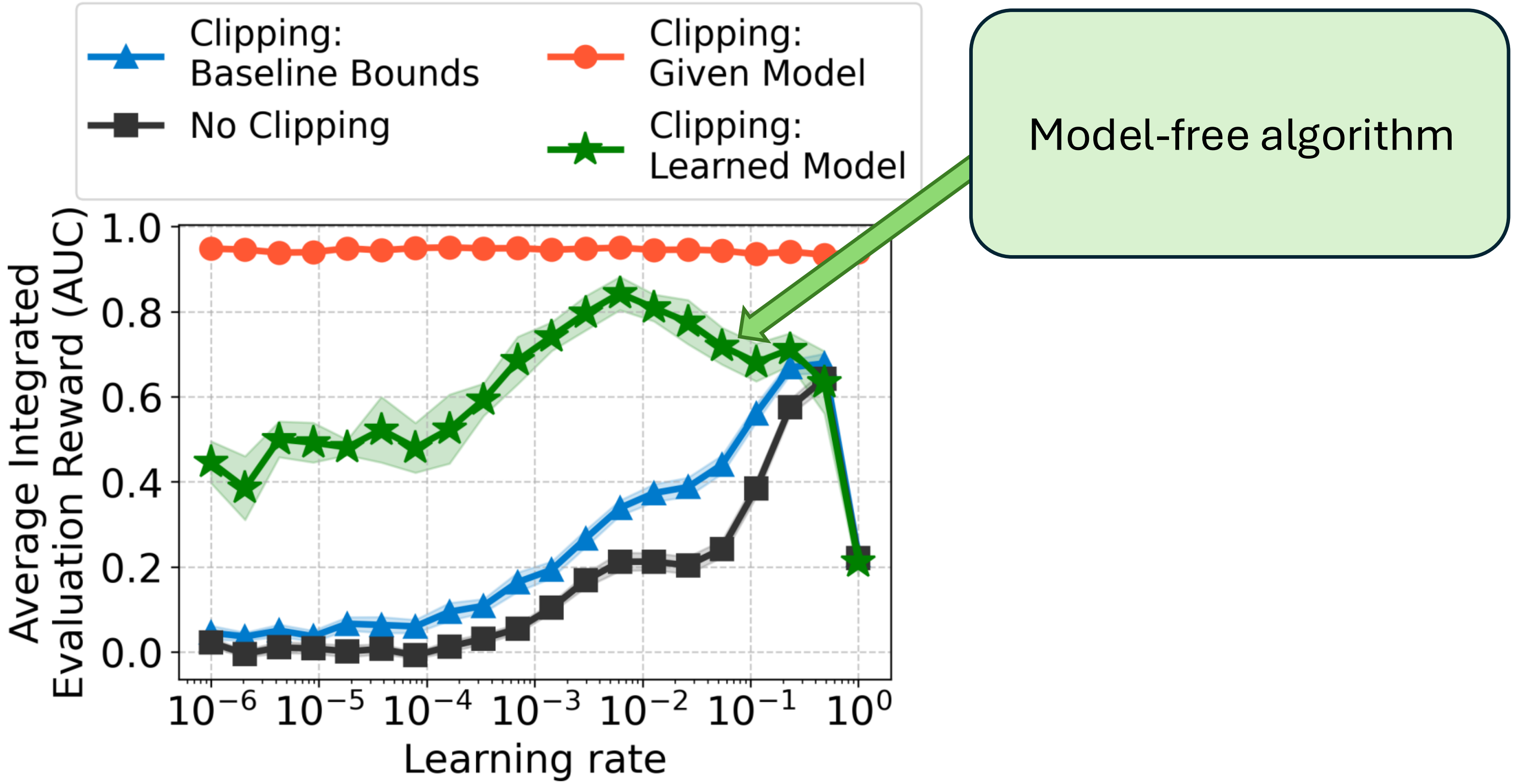
Clipping is All You Need*

Conditional clipping performs near-optimal for all learning rates



Always clipping with Bellman performs much worse

Clipping is All You Need*



Future Work

- Use model-based techniques for extending advantage in deep RL
- Derive even tighter bounds

Thank you!

Pseudocode

Take action a: observe reward r, next state s', and termination signal

Compute state value function: $V(s') = \beta^{-1} \log \mathbb{E}_{a' \sim \pi_0} \exp \beta Q(s', a')$

Calculate new bounds $\{\mathrm{L}'(s,a),\ \mathrm{U}'(s,a)\}$ using Q' in Equation 2.

Tighten lower bounds: $\mathrm{L}'(s,a) = \max\{\mathrm{L}'(s,a),\ \mathrm{L}(s,a)\}$

Tighten upper bounds: $\mathrm{U}'(s,a) = \min\{\mathrm{U}'(s,a),\ \mathrm{U}(s,a)\}$

Clip the Q-values: $Q'(s,a) = \mathrm{clamp}\left(Q(s,a), \min = \mathrm{L}'(s,a), \max = \mathrm{U}'(s,a)\right)$

if $Q' == Q$ **then**

    // No clipping has been applied, resort to TD-update:

    Compute the TD error: $\delta = r + \gamma \cdot (1 - \mathrm{terminated}) \cdot V(s') - Q(s,a)$

    Update Q-table: $Q'(s,a) \leftarrow Q'(s,a) + \alpha\delta$

end if

Theorem 2 (Informal). *Consider an MDP with a bounded continuous state and action space,* $\mathcal{S} \times \mathcal{A} \subset \mathbb{R}^d$*, with stochastic dynamics. Suppose an* L_Q*-Lipschitz function* $Q(s,a)$ *is given to generate double-sided bounds on the optimal value function, denoted* $Q^*(s,a)$*. Let* $\varepsilon > 0, \delta > 0$ *be given and define the horizon* $H = (1-\gamma)^{-1}$*, and sample budgets:* $|\mathcal{B}| \geq \mathcal{O}\left(\varepsilon^{-d} \log \delta^{-1}\right)$, $n_{\mathcal{S}} \geq \mathcal{O}\left(H^2 \varepsilon^{-2} \log \delta^{-1}\right)$, $n_{\mathcal{A}} \geq \mathcal{O}\left(e^{2\beta(H-\varepsilon)} \log \delta^{-1}\right)$.
Suppose $n_{\mathcal{S}}$ *samples are used to estimate the expectation over next-states and* $n_{\mathcal{A}}$ *samples are used to estimate the expectation over next-actions in the soft state-value function. Denoting* $\hat{V}, \hat{\Delta}$ *as the quantities estimated from samples, the following bounds*

$$Q^*(s,a) \leq r(s,a) + \gamma \left(\frac{1}{n_{\mathcal{S}}} \sum_{i=1}^{n_{\mathcal{S}}} \hat{V}(s_i') + \frac{\max_{(s,a)\in\mathcal{B}} \hat{\Delta}(s,a) + \varepsilon}{1-\gamma} \right) \tag{4}$$

$$Q^*(s,a) \geq r(s,a) + \gamma \left(\frac{1}{n_{\mathcal{S}}} \sum_{i=1}^{n_{\mathcal{S}}} \hat{V}(s_i') + \frac{\min_{(s,a)\in\mathcal{B}} \hat{\Delta}(s,a) - \varepsilon}{1-\gamma} \right) \tag{5}$$

hold with probability at least $1-\delta$*.*