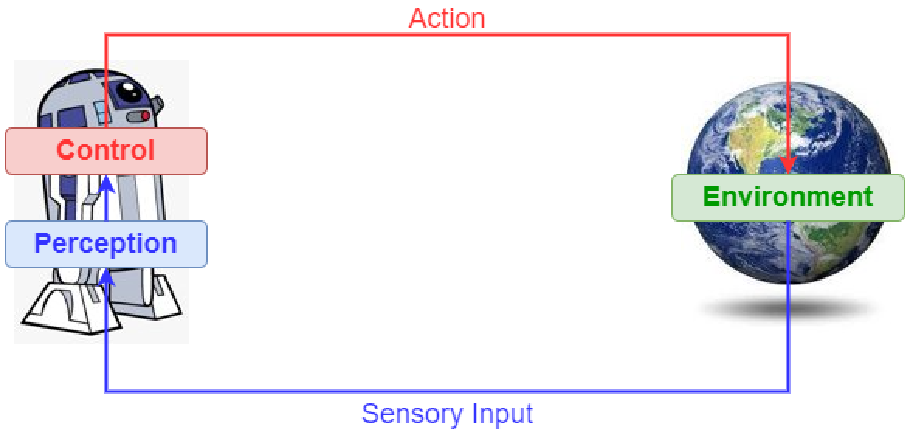
Adversarially-Robust TD Learning: Finite-Time Rates and Fundamental Limits

Sreejeet Maity Aritra Mitra
Department of Electrical and Computer Engineering
North Carolina State University

28th International Conference on Artificial Intelligence and Statistics 2025

The Standard RL Pipeline

- Agent takes action.
- Environment provides feedback (rewards).
- Agent learns to take "better" actions.
- **Goal:** Maximize long-term returns.



Q. How do we take "good" decisions under environmental uncertainty?

Towards Robust RL

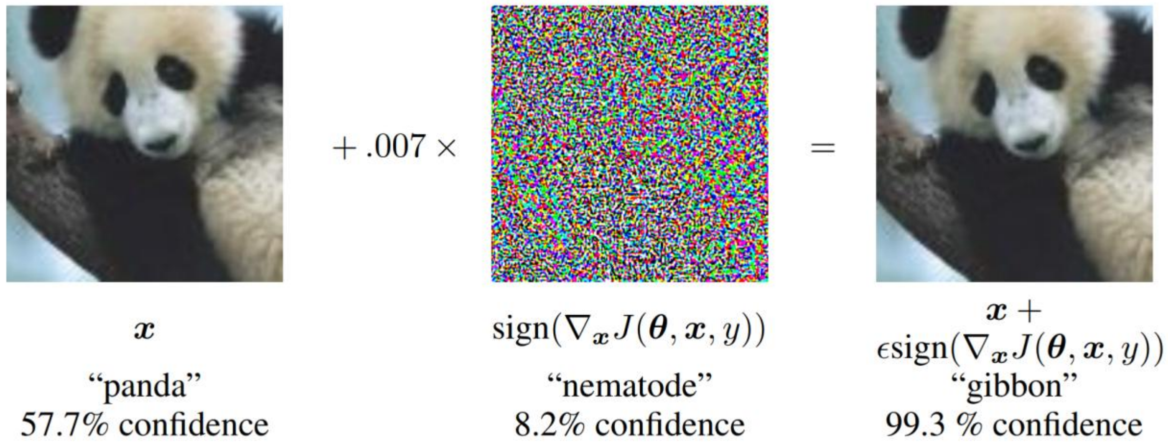
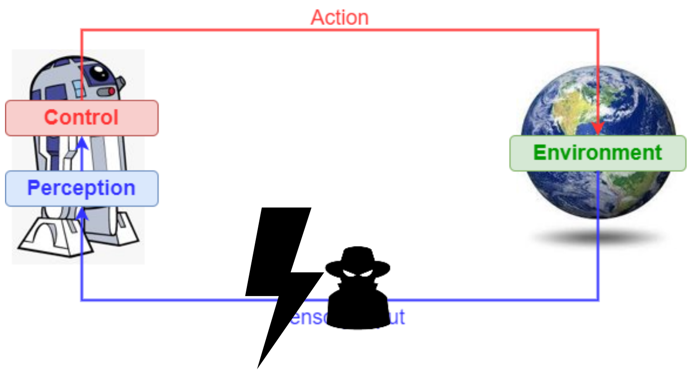


Image taken from Goodfellow et al., ICLR 2015



Q. How to make RL algorithms robust to **adversarially perturbed rewards**?

Q. Can **autonomous agents** (e.g., self-driving cars) make **reliable decisions** with **corrupted** data?

Basic RL Setup

- We consider an MDP $\mathcal{M} = (S, A, P, R, \gamma)$ with finite state and action spaces.
- $R(s, a)$ is the immediate expected reward at state-action pair (s, a).
- $P(s'|s, a)$ is the probability of transitioning from s to s' under action a.
- A deterministic policy $\pi: S \mapsto A$ induces a Markov Reward Process (MRP) with a reward function r_π and transition matrix P_π.

- The "goodness" of a policy π is captured by the value function V_π:

$$V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_\pi(s_t) | s_0 = s\right],\ \gamma \in (0,1).$$

- Goal: Find a policy π that maximizes $V_\pi(s)$ for all s.

The Policy Evaluation Problem

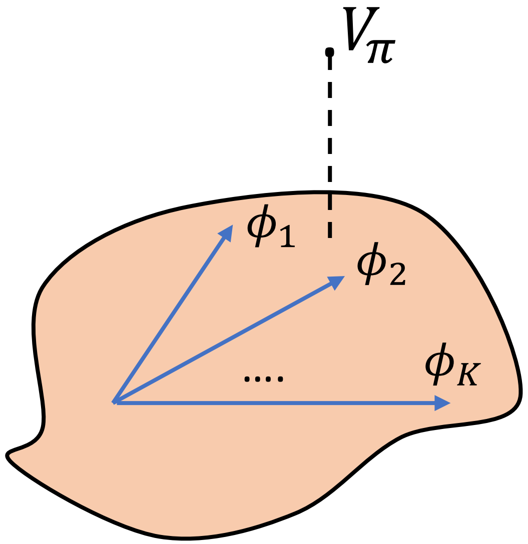
- **Policy Evaluation Goal:** Given a policy π, compute V_π.
- **Policy-Specific Bellman Operator:** We have $\boxed{\mathcal{T}_\pi V_\pi = V_\pi}$ where

$$\boxed{(\mathcal{T}_\pi \mathrm{V})(\mathrm{s}) = \mathrm{r}_\pi(\mathrm{s}) + \gamma \sum_{s' \in S} P_\pi(s, s') V(s')}$$

- **Dynamic Programming:** Run $\hat{V}_{t+1} = \mathcal{T}_\pi\left(\hat{V}_t\right)$, and use contractivity of $\mathcal{T}_\pi$.
- Q. But what if the MDP is unknown?
    - **Temporal Difference Learning**, Richard Sutton, Machine Learning, 1988.

TD Learning with Function Approximation

- **Linear Function Approximation:** Find a parametric approximation $\hat{V}_{\boldsymbol{\theta}}$ of V_π in the span of a feature matrix Φ.



$$\underbrace{\begin{pmatrix} \hat{V}_\theta(s_1) \\ \vdots \\ \hat{V}_\theta(s_N) \end{pmatrix}}_{\hat{V}_\theta} = \underbrace{\overbrace{\begin{pmatrix} | & | & & | \\ \phi_1 & \phi_2 & \cdots & \phi_K \\ | & | & & | \end{pmatrix}}^{K \text{ basis vectors, where } K << |S| = N}}_{\substack{\Phi \\ \text{Feature Matrix}}} \underbrace{\begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_K \end{pmatrix}}_{\theta}$$

For each state s,

$$\hat{V}_\theta(s) = \langle \phi(s), \theta \rangle$$

Feature vector for state s

TD Learning with Function Approximation

- **Linear Function Approximation:** $\hat{V}_\theta = \Phi\theta$.
- For each state s, $\hat{V}_\theta(s) = \langle \phi(s), \theta \rangle$.
- **TD Learning:** Play policy π. At each $t = 0,1, \ldots,$ observe $X_t = \big(s_t, s_{t+1}, r_\pi(s_t)\big)$.

$$\boxed{\textbf{TD(0) Update:} \quad \theta_{t+1} = \theta_t + \alpha g_t(\theta_t)}$$

Data tuples are **random**, and randomness is **Markovian**

$$g_t(\theta) := \Big(\underbrace{r_\pi(s_t) + \gamma \overset{\hat{V}_\theta(s_{t+1})}{\langle \phi(s_{t+1}), \theta \rangle}}_{\text{New estimate of } V_\pi(s_t)} - \underbrace{\overset{\hat{V}_\theta(s_t)}{\langle \phi(s_t), \theta \rangle}}_{\text{Old estimate of } V_\pi(s_t)}\Big)\phi(s_t), \forall \theta$$

Prior Analysis of TD with Function Approx.

- **Asymptotic Analysis:** Tsitsiklis & Van Roy, TAC, 1997.
  - Main Idea: View TD methods as instances of Stochastic Approximation.

  "Though temporal-difference learning is simple and elegant, a rigorous analysis of its behavior requires significant sophistication" – Tsitsiklis and Van Roy

- **Non-Asymptotic Analysis:** Bhandari, Russo, & Singal, COLT 2018; Srikant & Ying, COLT 2019.
  - Bhandari et al. → Connections to optimization; Assume a projection step.
  - Srikant & Ying → Use Lyapunov theory; No projection, but involved proof.

Q. Can we provide a **simple** convergence analysis of **unprojected** TD?

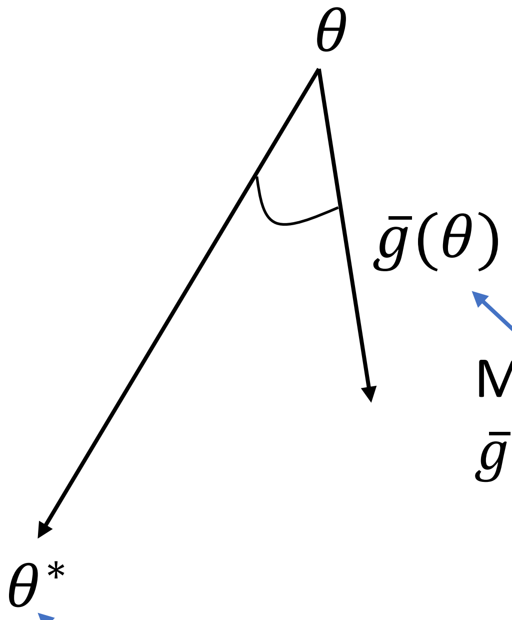
Main Technical Challenge

- TD(0) update direction can be expressed as: $g_t(\theta) = A_t\theta - b_t$, where randomness due to Markov sampling is contained in A_t and b_t.
- Suppose Markov chain is aperiodic and irreducible $\Rightarrow$ $A_t \to \bar{A}, b_t \to \bar{b}$.
- Define $\bar{g}(\theta) = \bar{A}\theta - \bar{b}$ as steady-state/mean-path TD direction.

- **Idea:** $\theta_{t+1} = \underbrace{\theta_t + \alpha\bar{g}(\theta_t)}_{\text{Steady-state dynamics}} + \underbrace{\alpha\big(g_t(\theta_t) - \bar{g}(\theta_t)\big)}_{\text{Disturbance} = w_t}$.

- **Key Challenge:** w_t depends on the magnitude of θ_t.
  - Projection can help control $||\theta_t||$. But we don't want to project!

Some Basic Facts about TD



Mean-Path TD(0) direction:
$\bar{g}(\theta) = \bar{A}\theta - \bar{b}$

TD(0) Fixed Point: $\theta^* = \bar{A}^{-1}\bar{b}$

Fact 1 (“Strong-Convexity”):

$$\langle \theta^* - \theta, \bar{g}(\theta) \rangle \geq \mu \|\theta^* - \theta\|^2, \forall \theta,$$

where $\mu > 0$.

Tsitsiklis & Roy, 97; Bhandari et al., 2018, …

Fact 2 (“Smoothness”):

$g_t(\theta)$ and $\bar{g}(\theta)$ are both 2-Lipschitz.

Step 1: Setting up the Main Recursion

- Let $d_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ be the mean-squared error. Then, using Facts 1 and 2,

$$d_{t+1} \leq \underbrace{(1-\alpha\mu)d_t}_{\text{Contractive term from steady-state dynamics}} + \underbrace{O(\alpha^2\sigma^2)}_{\text{Noise variance term}} + \underbrace{2\alpha\mathbb{E}[\langle\theta_t - \theta^*, g_t(\theta_t) - \bar{g}(\theta_t)\rangle]}_{\text{Markovian bias}}$$

Contractive term from steady-state dynamics

Noise variance term

(Depends on bound on rewards and $\|\theta^*\|$)

Markovian bias

(No such bias for SGD with i.i.d. noise)

Q. How do we control the Markovian bias **without projection?**

Step 2: Arguing Boundedness of Iterates

Theorem (Informal): Boundedness of Iterates

There exists a constant step-size $\alpha \propto \frac{\mu}{\tau_{mix}}$, such that with this step-size, $d_t \leq O(max\{\|\theta_0 - \theta^*\|^2, \sigma^2\})$

$d_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$

Mixing time of underlying Markov chain

Key Idea in Proof: Use **Induction** + Contraction Properties of Operator + Mixing Properties of Markov Chain

Step 2 (Continued): Controlling Markovian Bias

Theorem (Informal): Boundedness of Iterates

There exists a constant step-size $\alpha \propto \frac{\mu}{\tau_{mix}}$, such that with this step-size, $d_t \leq O(max\{\|\theta_0 - \theta^*\|^2, \sigma^2\})$

Corollary (Informal): Markovian Bias $\leq O(\alpha\tau_{mix}B)$, where $B = 10\, max\{\|\theta_0 - \theta^*\|^2, \sigma^2\}$

Step 3: Simplifying the Main Recursion

- From Steps 1 and 2,

$$\boxed{d_{t+1} \leq (1 - \alpha\mu)d_t + O(\alpha^2 \tau_{mix} B)}$$

Uniformly bounded perturbation

- With constant non-diminishing step-size, exponential convergence to a noise ball.
- With step-size $\alpha \propto \frac{\log(T)}{T}$, can obtain $O(1/T)$ rate.

Summary of General Recipe

- **Step 1:** Use contraction + smoothness properties of operator to establish a basic MSE recursion.

- **Step 2:** Use induction to argue uniform boundedness of iterates and Markovian bias.

- **Step 3:** Use bound from Step 2 to refine the MSE recursion.

A Closer look at the Induction Analysis

- **Lemma 1:** Suppose $\alpha \leq \frac{1}{8\tau_{mix}}$, and let $B = 10\, max\{\|\theta_0 - \theta^*\|^2, \sigma^2\}$.
  Then, we have:
  $$\|\theta_k - \theta^*\|^2 \leq B, \forall k \in [\tau_{mix}].$$

  Comments:
  - Unroll TD recursion and use $\|g_t(\theta)\| \leq 2\,\|\theta\| + 2\,\sigma, \forall \theta$.
  - Lemma 1 serves as base case of induction.

A Closer look at the Induction Analysis

- Recall $d_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$.
- **Lemma 2:** Consider any $t \geq \tau_{mix}$. Suppose $d_k \leq B, \forall k \in [t]$. Then,

$$\mathbb{E}\left[\left\|\theta_t - \theta_{t-\tau_{mix}}\right\|^2\right] \leq O\left(\alpha^2 \tau_{mix}^2 B\right).$$

Proof idea: Observe $\left\|\theta_t - \theta_{t-\tau_{mix}}\right\| \leq \sum_{k=t-\tau_{mix}}^{t-1} \|\theta_{k+1} - \theta_k\|$

$$\leq \alpha \sum_{k=t-\tau_{mix}}^{t-1} \|g_k(\theta_k)\| \quad \text{(From update rule)}$$

$$\leq O(\alpha) \sum_{k=t-\tau_{mix}}^{t-1} (\|\theta_k - \theta^*\| + \sigma)$$

Now use $d_k \leq B$

A Closer look at the Induction Step

- Recall Markovian bias term $e_t = \mathbb{E}[\langle \theta_t - \theta^*, g_t(\theta_t) - \bar{g}(\theta_t) \rangle]$.
- **Lemma 3:** Consider any $t \geq \tau_{mix}$. Suppose $d_k \leq B, \forall k \in [t]$. Then,

$$\boxed{e_t \leq O(\alpha \tau_{mix} B)}$$

- Comments on proof:
  - Condition sufficiently into the past, use Lemma 2, and exploit geometric mixing.
  - The assumption $d_k \leq B$ considerably simplifies the proof.
- Plugging in bound from Lemma 3 in coarse recursion from Step 1, it is easy to show that $d_t \leq B, \forall t$.
- Key Point: The above step shows that the requirements for Lemmas 2 and 3 will be met at all time steps.

The Heavy-Tailed Noise Model

- Recall we wish to evaluate the value function V_π corresponding to a policy π.
- With each state $s \in \mathcal{S}$, we associate a conditional distribution $\mathcal{D}_\pi(\cdot\,|s)$ s.t. whenever $\pi(s)$ is played in state s, we observe a noisy reward $r(s) \sim \mathcal{D}_\pi(\cdot\,|s)$.

- Statistics of $r(s)$:
  - **Unbiasedness**: $\mathbb{E}_{r(s)\sim\mathcal{D}_\pi(\cdot|s)}[r(s)] = r_\pi(s)$. Assume $|r_\pi(s)| \leq \bar{r}, \forall s \in \mathcal{S}$.
  - **Bounded variance:** $Var\big(r(s)\big) \leq \rho^2$.

- **Note:** We do not assume that the noisy reward random variables are sub-Gaussian.

The Corruption Model

- **Corruption Model:** At each time-step t, learner observes $\tilde{r}_t$ generated as follows.
  - Toss a biased coin with probability of heads $(1-\varepsilon)$, where $\varepsilon \in \left[0, \frac{1}{2}\right)$.
  - If coin lands heads, learner observes $\tilde{r}_t \sim \mathcal{D}_\pi(\cdot\,|s_t)$.
  - If coin lands tails, learner observes $\tilde{r}_t \sim \mathcal{Q}$, where $\mathcal{Q}$ is an **unknown and unconstrained** error distribution controlled by an **adversary**.

Huber-contaminated reward model

$$\boxed{\tilde{r}_t \sim (1-\varepsilon)\mathcal{D}_\pi(\cdot\,|s_t) + \varepsilon\mathcal{Q}}$$

True reward distribution

Adversarial distribution

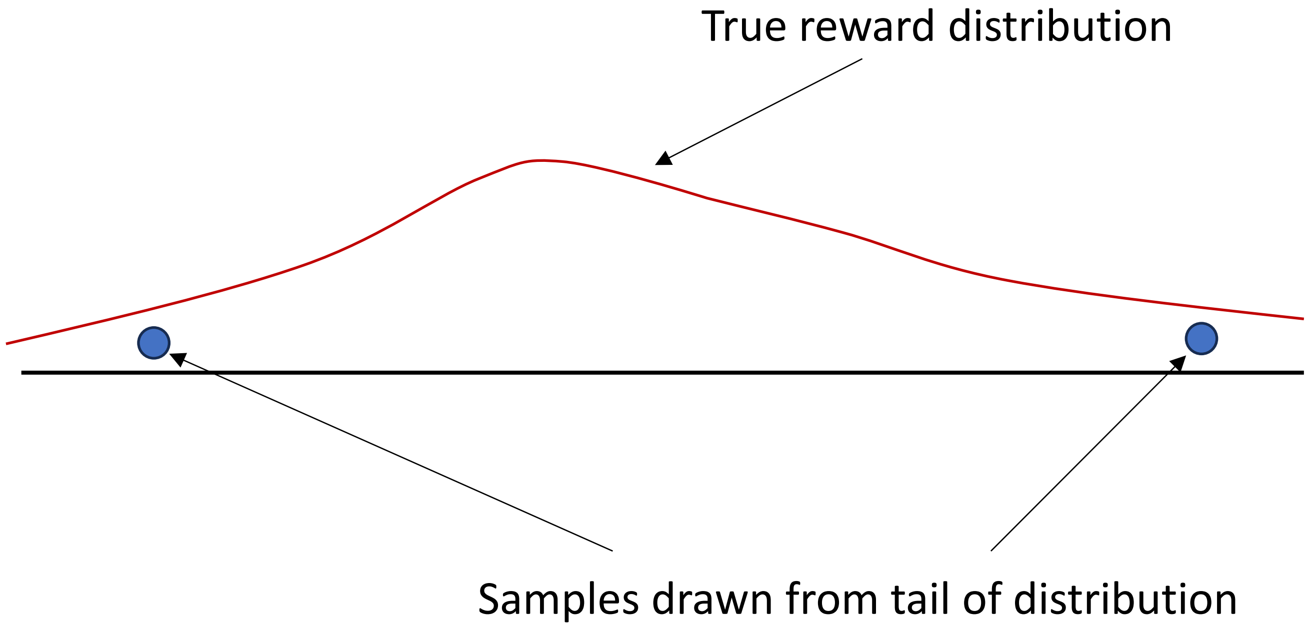
Key Questions of Interest

- Under the Huber-contaminated reward model,
  - What can be said of the vanilla TD learning algorithm?
  - Can we still hope to obtain a **reliable** estimate of V_π?
  - What are the **fundamental limits** on performance?

Challenges

- **Challenge 1:** Inliers are **heavy-tailed**.
  - Hard to distinguish from outliers.

- **Challenge 2:** Data are **correlated** over time.
  - Robust statistics deals with i.i.d. data.



Vulnerability of Vanilla TD

- Suppose standard TD(0) algorithm is run with step-size sequence $\{\alpha_t\}$.
- Assume Markov chain induced by π is aperiodic and irreducible.
- Recall in the absence of corruptions, TD(0) converges to $\theta^* = \bar{A}^{-1}\bar{b}$.

Theorem (Informal): Vulnerability of TD(0)

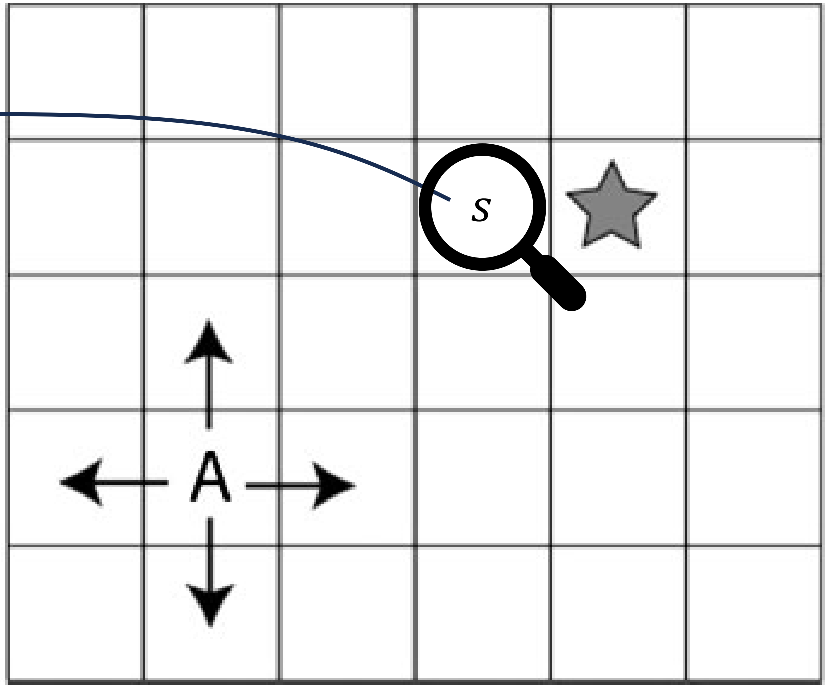
Suppose $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$. With probability 1, the iterates of TD(0) can be made to converge to $\tilde{\theta}^* = (1-\varepsilon)\theta^* + \varepsilon C$, where C is a *corruption vector that can be controlled by the adversary.*

Note: For every $w \in \mathbb{R}^K$, there exists a feasible attack s.t. $\tilde{\theta}^* = w$.

Towards a Robust TD Algorithm

- **Idea 1:** Use historical data to build robust estimates of reward means.
  - What about temporal correlations?
  - What about rare events?

- **Idea 2:** Reject estimates that deviate "too much" from expected bounds.
  - How to design rejection threshold?



Building Intuition - 1

- Recall TD(0) update direction (without corruption) is of the form
$$g_t(\theta) = A_t\theta - b_t, \text{ where } b_t = -\phi(s_t)r(s_t).$$

- **Observation 1:** Rewards only affect the term b_t, not A_t.

- **Observation 2:** Eventually, b_t will approach its stationary value $\bar{b}$, given by
$$\bar{b} = -\sum_{s\in\mathcal{S}} \mu(s)\phi(s)r_\pi(s),$$
where μ is the stationary distribution of the Markov chain induced by π.

Goal: Maintain a robust estimate of $\bar{b}$.

Building Intuition - 2

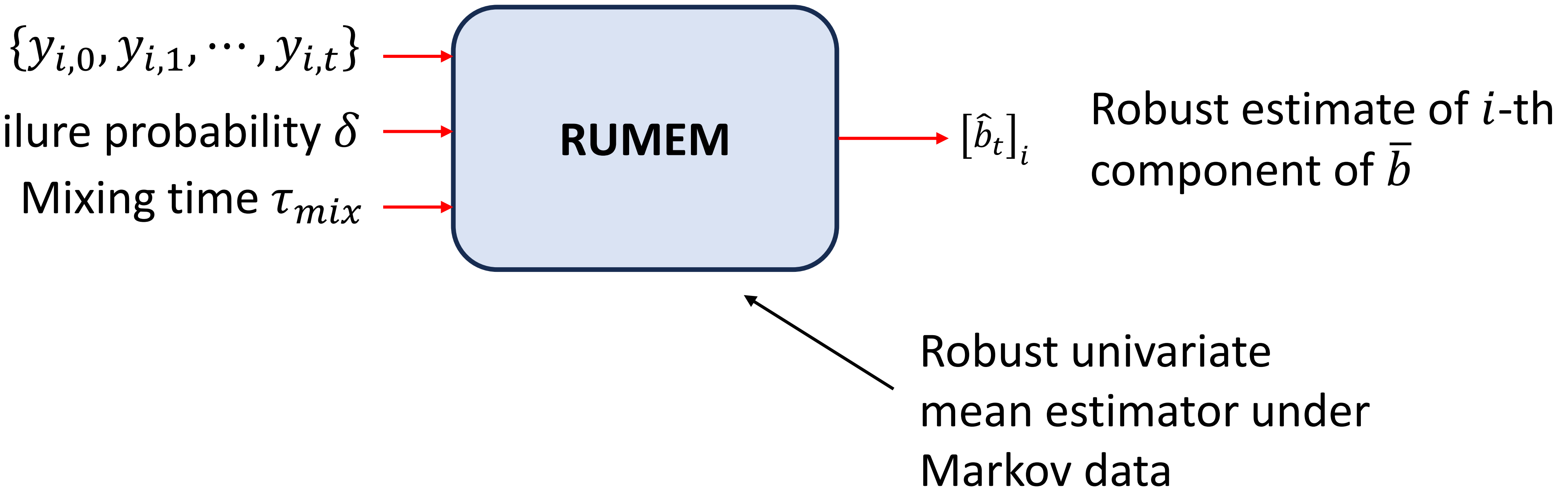
- **Goal:** Maintain estimate of $\bar{b} = -\sum_{s\in\mathcal{S}} \mu(s)\phi(s)r_\pi(s)$.

- **Idea 1:** For each $s \in \mathcal{S}$, maintain separate estimates of $\mu(s)$ and $r_\pi(s)$.
  - Issue: Defeats the purpose of function approximation.

- **Idea 2:** Apply a robust mean estimator to set of reward observations $\{\tilde{r}_k\}$.
  - Issue: Provides estimate of $-\sum_{s\in\mathcal{S}} \mu(s)r_\pi(s)$.

Step 1: Estimating $\bar{b}$

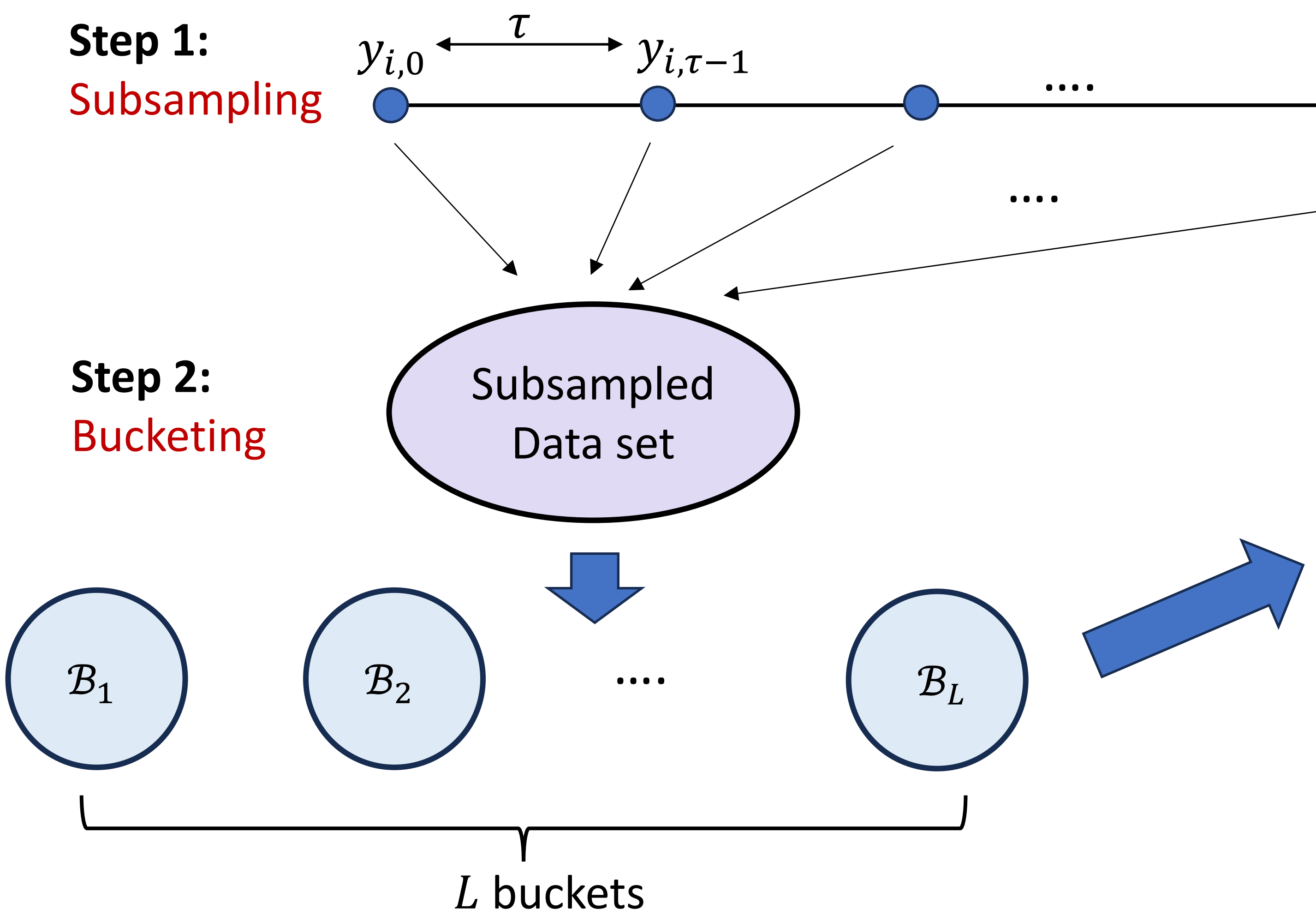
- Define $y_{i,k} = [\phi(s_k)]_i \tilde{r}_k$.



Key insight: If Markov chain is stationary, each uncorrupted $y_{i,k}$ provides an unbiased estimate of $[\bar{b}]_i$

RUMEM Estimator

- **Goal:** Data set $\{y_{i,0}, y_{i,1}, \cdots, y_{i,t}\}$. Estimate $[\bar{b}]_i$.



Step 3: Compute Median of Means:
$[\hat{b}_t]_i = Median\{\hat{z}_1, \dots \hat{z}_L\}$, where $\hat{z}_j$ is the mean of the data in j-th bucket.

Robust Markovian Mean Estimation

Theorem (Informal):

Under appropriate choices of the subsampling gap τ and number of buckets L, the output of RUMEM satisfies the following w.p. $1-\delta$:

$$\left|\left[\hat{b}_t\right]_i - \left[\bar{b}\right]_i\right| \le \max\{\bar{r},\rho\}\,\tilde{O}\left(\sqrt{\varepsilon} + \sqrt{\frac{\tau_{mix}}{t}\log\left(\frac{t}{\delta}\right)}\right)$$

Bounds on reward mean and variance

Corruption fraction

Mixing time of Markov chain

Note: First guarantees of robust mean estimation under both Markovian and adversarial data. Proof uses a coupling technique (Dorfman and Levy, ICML 22).

Step 2: Dynamic Thresholding

- What happens on **rare events** where robust mean estimation guarantees **do not** hold?

- Define a threshold $G_t = C\sqrt{K}\sigma\left(\sqrt{\varepsilon} + \log(KT)\sqrt{\frac{\tau_{mix}}{t}}\right)$

$\max\{\bar{r}, \rho\}$

Bounds on reward mean and variance

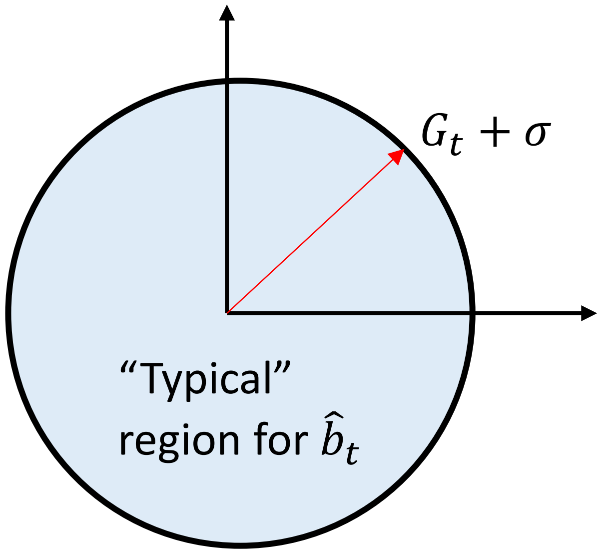
Number of samples

Corruption fraction

Note: Design of G_t is based on guarantees from robust mean estimation under Markov data.

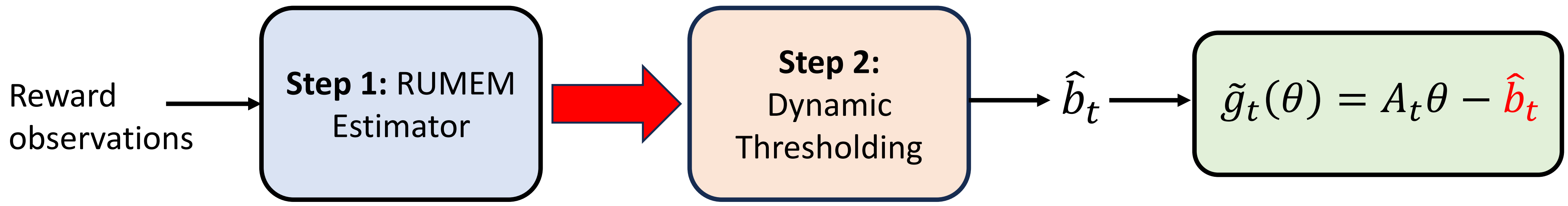
Step 2: Dynamic Thresholding

- Let $\hat{b}_t$ be the output from Step 1 (robust mean estimation).
- **Thresholding:** If $\|\hat{b}_t\|_2 > G_t + \sigma$, set $\hat{b}_t \leftarrow 0$.



Note: Should only threshold on rare events. Else, bounds would be vacuous.

Putting the pieces together



Robust TD algorithm $\boxed{\theta_{t+1} = \theta_t + \alpha\tilde{g}_t(\theta_t)}$

Performance of Robust TD

Theorem (Informal): Guarantees for Robust TD

Under a suitable step-size α, and for T large enough, we have

$$\mathbb{E}[\|\theta_T - \theta^*\|_2^2] \leq \underbrace{\tilde{O}\left(\frac{\tau_{mix} G}{T}\right)}_{\text{Standard TD(0) bound}} + \underbrace{O(\varepsilon \sigma^2 G)}_{\text{Effect of Corruption}}, \text{ where } G = \frac{K}{\omega^2 (1-\gamma)^2}$$

- When $\varepsilon = 0$, result matches existing bounds (e.g., Bhandari et al., 2018) up to multiplicative $O(K)$ factor.
- When $\varepsilon \neq 0$, additive corruption term is consistent with similar results for bandits with reward corruption (Lykouris et al., 2018, Gupta et al., 2019, and Kapoor et al., 2019).

Lower Bound

- Consider a simplified tabular setting where learner observes T i.i.d. samples, i.e., $s_t \sim \mu, s_{t+1} \sim P_\pi(\cdot\,|s_t), \tilde{r}(s_t) \sim (1-\varepsilon)\mathcal{D}_\pi(\cdot\,|s_t) + \varepsilon \mathcal{Q}$.
- Let $\mathcal{M}(\varepsilon, \rho, \mathcal{Q})$ represent class of MRPs where reward noise has variance at most ρ^2, and corruption fraction is ε.

Theorem: Lower Bound

There exists a universal constant $\tilde{c} > 0$ s.t.

$$inf_{\hat{V}_T}\ sup_{V \in \mathcal{M}(\varepsilon,\rho,\mathcal{Q})} \mathbb{P}\left(\left\|\hat{V}_T - V\right\|_2 \geq \frac{\tilde{c}\rho\sqrt{\varepsilon}}{(1-\gamma)}\right) \geq \frac{1}{2}.$$

Main Message: Dependencies of our upper-bound on corruption fraction ε, noise variance ρ, and discount factor γ are tight.

Main Ideas for Proving Upper Bound

- **Challenge:** We have two bias terms now: Markovian noise and adversarial corruption. Bias terms are **coupled**.
  - Still need to ensure boundedness of iterates (no projection).

- **Step 1:** Establish bounds for robust Markovian mean estimation.
- **Step 2:** Based on Step 1, show that on a "good-event", no thresholding will take place.
- **Step 3:** Establish: $\mathbb{E}\left[\left\|\hat{b}_t - \bar{b}\right\|^2\right] = \tilde{O}\left(\varepsilon + \frac{\tau_{mix}}{t}\right) K\sigma^2$.
- **Step 4:** Analyze how adversarial corruption error propagates through bounds.

Ideas behind Lower Bound

- **Reduction to Robust Mean Estimation:** Consider trivial MDP with just one state and action. Only randomness comes from noisy reward observations.

Environment 1
V_1

Noise + Huber Contamination

$\mathcal{P}_1 = (1-\varepsilon)\mathcal{D}_1 + \varepsilon \mathcal{Q}_1$

$\mathcal{P}_1 = \mathcal{P}_2$

Learner

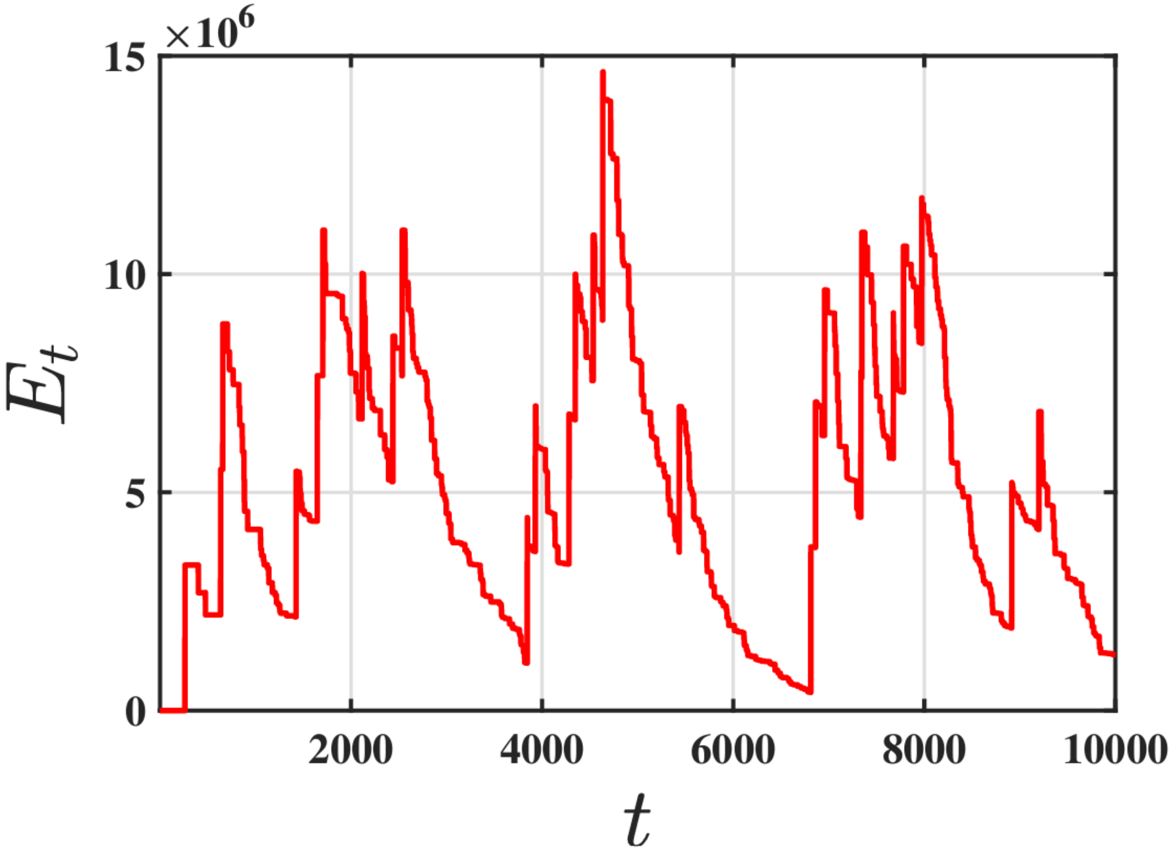
$$|V_1 - V_2| \geq \frac{\rho\sqrt{\varepsilon}}{2(1-\gamma)}$$

Environment 2
V_2

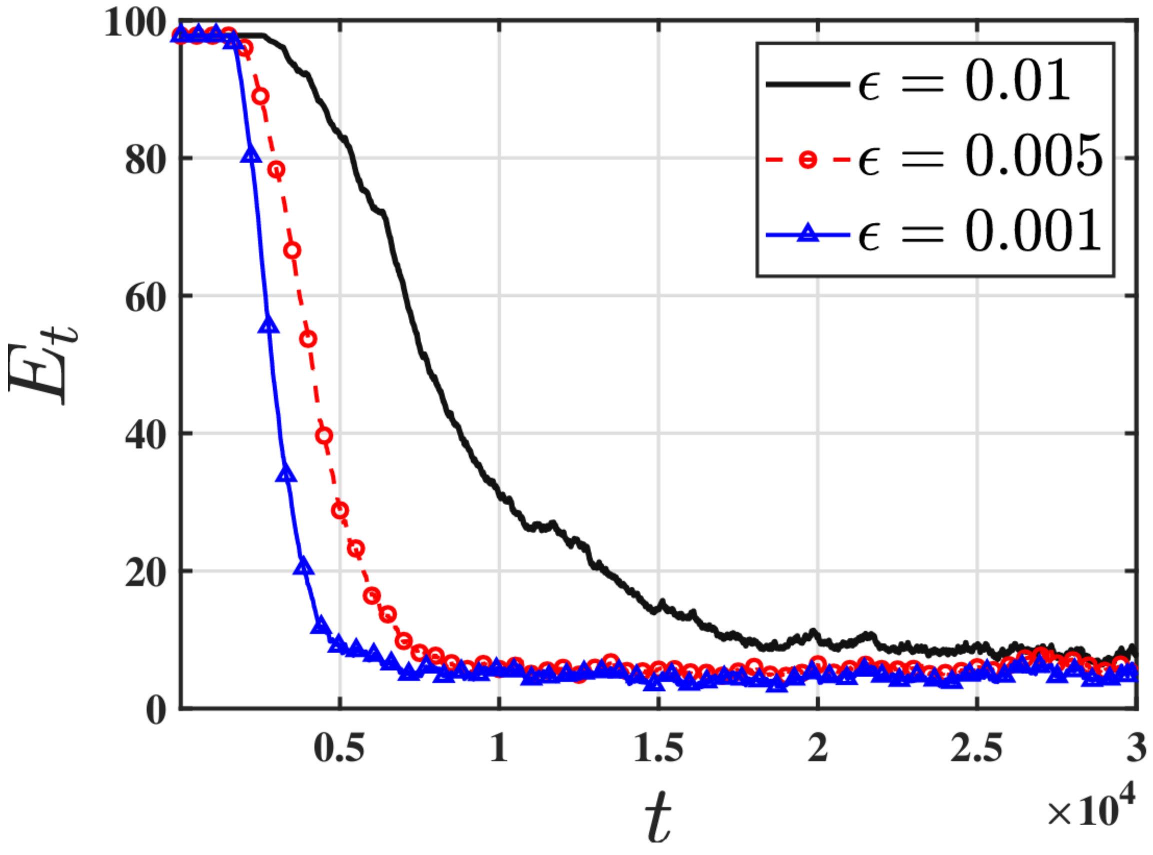
Noise + Huber Contamination

$\mathcal{P}_2 = (1-\varepsilon)\mathcal{D}_2 + \varepsilon \mathcal{Q}_2$

Simulation Study



Vanilla TD(0) with $\varepsilon = 0.001$



Robust TD(0)

Simulations on an MDP with 100 states, and 10 features. E_t is the MSE at time t.

Summary and Open Problems

- Considered policy evaluation with TD learning under corrupted rewards.
  - Vanilla TD(0) algorithm can incur arbitrarily large errors.
  - Proposed a new robust TD algorithm.
  - Showed that proposed algorithm nearly recovers performance of TD(0) without corruptions.
  - Proved a fundamental lower bound.
  - Result on robust Markovian mean estimation might be of independent interest.
- Open Problems:
  - Extending ideas to more general stochastic approximation problems, and data-driven control settings with continuous state-action spaces.
  - Other attack models? (e.g., state attacks?)
  - More refined lower bounds?
  - Getting rid of knowledge of mixing time, and bounds on mean and variance of reward distribution. Lesser memory requirements on algorithm?

References

- **Short Finite-Time Proof:**
  - *A Simple Finite-Time Analysis of TD Learning with Linear Function Approximation*, A. Mitra, **IEEE Transactions on Automatic Control**, 2024.
- **Adversarial Robustness in RL:**
  - *Adversarially-Robust TD Learning with Markovian Data: Finite-Time Rates and Fundamental Limits*, S. Maity & A. Mitra, **AISTATS 2025**.
  - *Robust Q-Learning with Corrupted Rewards*, S. Maity & A. Mitra, **CDC 2024**
- **Communication Constraints in RL:**
  - *TD Learning with Compressed Updates: Error-Feedback meets Reinforcement Learning*, A. Mitra, G. Pappas, and H. Hassani, **TMLR 2024.**
  - *Stochastic Approximation with Delayed Updates*, Adibi et al. & A. Mitra, **AISTATS 2024**

Robustness of Iterative RL Algorithms

- SGD is **robust** to various structured perturbations (e.g., noise, delays, biased compression).

Sparsified SGD with Memory

Sebastian U. Stich   Jean-Baptiste Cordonnier   Martin Jaggi

SIGNSGD: Compressed Optimisation for Non-Convex Problems

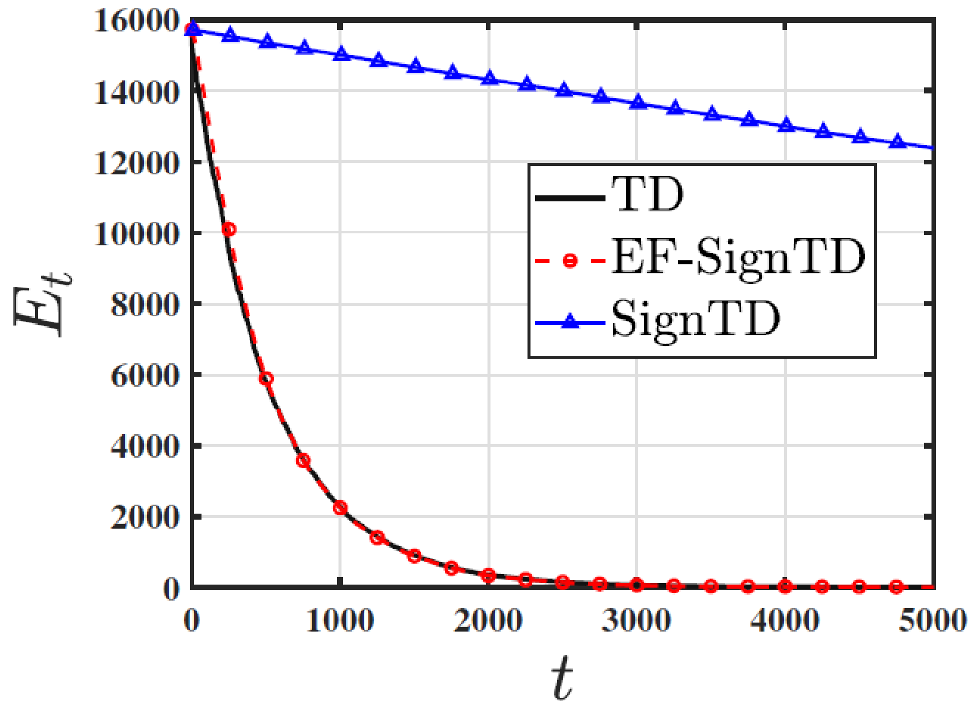
Jeremy Bernstein [1,2] Yu-Xiang Wang [2,3] Kamyar Azizzadenesheli [4] Anima Anandkumar [1,2]

Error Feedback Fixes SignSGD and other Gradient Compression Schemes

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, Martin Jaggi

Q. Are commonly used RL algorithms also robust to structured perturbations?

TD Learning with Perturbations



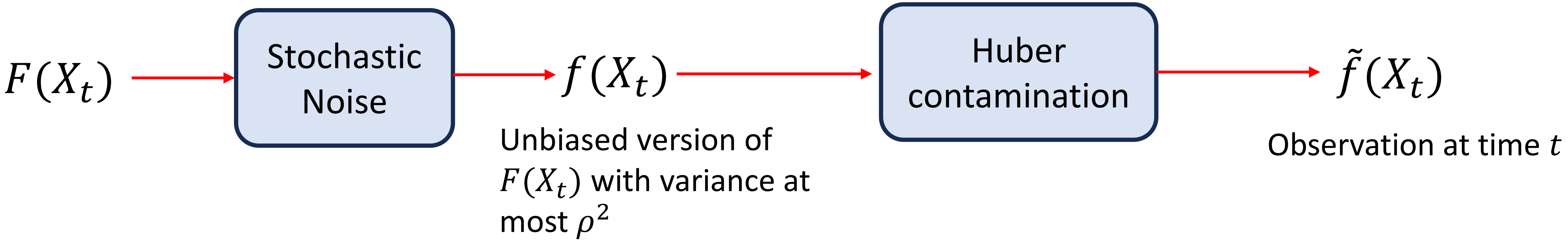
Error between iterates and solution to projected Bellman equation

- Analyze the behavior of TD algorithms with distorted update directions.
- Function approximation+Markovian sampling+Distortion+ErrorFeedback.
- **Main Message:** SignTD with error feedback retains the **same** finite-time convergence rates as TD.

Ref: *TD Learning with Compressed Updates: Error-Feedback meets RL*, A. Mitra, G. Pappas, and H. Hassani, TMLR (under review)

A Closer Look at Step 1

- Suppose $X_1, X_2, \ldots,$ is an ergodic, time-homogeneous and stationary Markov chain (on finite state space $\mathcal{X}$) with stationary distribution μ.
- Let $F: \mathcal{X} \to \mathbb{R}$ be a bounded function s.t. $|F(x)| \leq B, \forall x \in \mathcal{X}$.
- Define $\bar{F} = \mathbb{E}_{X\sim\mu}[F(X)]$.



Noisy and Huber-contaminated single-trajectory Markovian data