R-GTD: A GEOMETRIC ANALYSIS OF GRADIENT TEMPORAL-DIFFERENCE LEARNING IN SINGULAR REGIMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient temporal-difference (GTD) learning provides a principled approach to off-policy policy evaluation with linear function approximation. However, existing convergence analyses rely on the restrictive assumption that the so-called feature interaction matrix (FIM) is nonsingular. In practice, the FIM can become singular and leads to instability or degraded performance. In this paper, we propose a regularized optimization objective by reformulating the mean-square projected Bellman error minimization. This formulation naturally yields a regularized GTD algorithms, referred to as R-GTD, which guarantees convergence to a unique solution even when the FIM is singular. We conduct a geometric analysis to establish theoretical convergence guarantees and explicit error bounds for the proposed method, and validate its effectiveness through empirical experiments.

1 INTRODUCTION

Temporal-difference (TD) learning (Sutton, 1988) constitutes a foundational approach for policy evaluation in reinforcement learning (RL) (Sutton & Barto, 1998). Despite its effectiveness, it is now well understood that instability and divergence may arise when TD learning is combined with function approximation and bootstrapping in an off-policy setting. This combination, commonly referred to as the *deadly triad*, has been identified as a fundamental source of instability in RL (Tsitsiklis & Van Roy, 1996). To resolve the convergence issue, gradient temporal-difference (GTD) learning was introduced as an alternative to TD learning (Sutton et al., 2008). Unlike TD learning, which relies on semi-gradient updates, GTD is derived from the true stochastic gradient of a well-defined objective function known as the mean-square Bellman error (MSBE), whose minimization is equivalent to solving the policy evaluation problem by enforcing the Bellman equation in a least-squares sense. Unlike the TD learning, GTD is known to guarantee convergence under off-policy training with linear function approximation. However, it still faces limitations with linear function approximation. Because the Bellman operator typically maps a value function outside the class of functions representable by the linear function approximation. As a result, even the optimal solution of the MSBE may exhibit a nonzero residual Bellman error. To address this issue, GTD2 (Sutton et al., 2009) considers the projected Bellman equation (PBE), where the Bellman update is projected onto the feature space. Specifically, GTD2 minimizes the mean-square projected Bellman error (MSPBE), which measures the squared residual of the PBE. By enforcing the Bellman equation within the feature space, the projected Bellman error can be driven to zero, and it allows GTD2 to obtain a more accurate approximation of the value function that can be represented by the chosen linear features.

From an optimization perspective, the MSPBE objective admits an equivalent saddle-point formulation, under which GTD2 can be interpreted as solving a convex–concave saddle-point problem (Macua et al., 2014). This perspective allows the use of analytical tools such as Fenchel duality (Boyd & Vandenberghe, 2004) and Lagrangian methods (Dai et al., 2018). Recent work in Lee et al. (2022) has extended this perspective by proposing new saddle-point formulations and further GTD2 variants that exploit these structural insights. Alongside these perspectives, Ghiassian et al. (2020) introduces regularization to improve robustness, and other general regularization techniques have been proposed to stabilize various GTD objectives (Yu, 2017) or to ensure stability of the critic in off-policy control settings (Zhang et al., 2020). However, existing GTD methods in the literature typically require a nonsingularity assumption on a certain matrix which is associated with feature interaction. Under this

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

assumption, the solution to the PBE is well defined and unique, which in turn ensures the existence of a unique minimizer of the MSPBE. In this paper, for notational convenience, we refer to this matrix as the *feature interaction matrix* (FIM). While some prior works have applied regularization to handle cases where FIM is singular, their theoretical guarantees inevitably rely on other restrictive conditions. For instance, they either restrict the optimal solution to the origin (Yu, 2017), or resort to the global nonsingularity assumption to derive error bounds (Zhang et al., 2020). Moreover, mere convergence leaves the geometric link to the original GTD2 solution completely unexplored. To further elucidate this theoretical gap, a detailed mathematical comparison with these regularization-based methods is provided in Section K of Appendix.

In this paper, we propose regularized gradient temporal-difference (R-GTD), which guarantees convergence to a unique solution even when the FIM is singular without introducing additional conditions. Furthermore, we provide a geometric analysis that explicitly characterizes the limiting behavior of the R-GTD solution within the GTD2 affine solution set, which deepens the theoretical understanding of GTD2.

Finally, the main contributions are summarized as follows:

(a) We propose R-GTD, which establishes rigorous convergence guarantees even when the FIM is singular.

(b) We study the theoretical properties of R-GTD from two complementary perspectives. First, we analyze the convergence of R-GTD toward the saddle point of the associated convex–concave formulation. Second, we provide a geometric analysis of the R-GTD solution across both nonsingular and singular regimes. We establish explicit error bounds between the R-GTD solution and the true projected solution in both nonsingular and singular settings, which precisely characterize the limiting behavior of the R-GTD solution within the null space of the FIM.

(c) We empirically validate the proposed algorithm in a representative setting where the FIM is singular. R-GTD converges consistently to the unique saddle-point solution with more stable behavior. Additional experiments in Section J of Appendix extend this evaluation to diverse scenarios.

2 PRELIMINARIES

2.1 MARKOV DECISION PROCESS

A Markov decision process (MDP) is defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces, respectively. The transition probability $P(s'|s,a)$ specifies the probability of moving from state s to s' when action a is taken, and $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function. The discount factor $\gamma \in (0,1)$ determines the relative importance of future rewards. When an agent in state s selects action a, it transits to the next state s' with probability $P(s'|s,a)$ and receives an immediate reward $r(s,a,s')$. A stochastic policy $\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$ specifies the probability $\pi(a|s)$ of selecting action a at state s. Let P^π denote the state transition matrix induced by policy π and D denote a diagonal matrix with positive diagonal elements $d(s)$ for $s \in \mathcal{S}$. Here, d can be any state distribution satisfying $d(s) > 0$ for all $s \in \mathcal{S}$. As a particular case, d may be chosen as the stationary state distribution d^β induced by a behavior policy β. The expected reward under policy π at state s is denoted by $R^\pi(s)$. The infinite-horizon discounted value function associated with π is defined as

$$J^\pi(s) := \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k, s_{k+1}) \,\middle|\, s_0 = s,\ \pi\right],$$

where $\mathbb{E}[\cdot]$ denotes the expectation with respect to the state–action trajectories generated by π. Given a set of basis (feature) functions $\phi_1, \ldots, \phi_q : \mathcal{S} \to \mathbb{R}$, we define the *feature matrix* $\Phi \in \mathbb{R}^{|\mathcal{S}| \times q}$ whose s-th row is given by $\phi(s) := [\phi_1(s) \ \cdots \ \phi_q(s)]$. Here, q is the number of feature functions and $|\mathcal{S}|$ denote the cardinality of the state. Throughout this paper, we assume that Φ has full column rank. Under this setting, the *policy evaluation problem* aims to estimate the value function J^π corresponding to a fixed policy π.

2.2 SADDLE-POINT PROBLEM

We briefly review the saddle-point problem (Nedić & Ozdaglar, 2009; Qu & Li, 2018). Let $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a convex–concave function. We begin with the standard definition of a saddle point.

Definition 2.1 (Saddle point)**.** A pair $(\theta^*, \lambda^*) \in \mathbb{R}^n \times \mathbb{R}^n$ is a saddle point of L if

$$L(\theta^*, \lambda) \le L(\theta^*, \lambda^*) \le L(\theta, \lambda^*) \quad \forall (\theta, \lambda).$$

When L is convex in θ and concave in λ, the saddle-point condition is equivalent to the following min–max problem:

Problem 1 (Min–max problem)*.*

$$\min_{\theta \in \mathbb{R}^n} \max_{\lambda \in \mathbb{R}^n} L(\theta, \lambda) = \max_{\lambda \in \mathbb{R}^n} \min_{\theta \in \mathbb{R}^n} L(\theta, \lambda).$$

This equivalence requires the convex–concave structure of L together with mild regularity conditions such as coercivity (Sion, 1958; Rockafellar, 1970). Under these conditions, a pair (θ^*, λ^*) is a saddle point if and only if it satisfies the first-order stationarity conditions

$$\nabla_\theta L(\theta^*, \lambda^*) = 0, \ \nabla_\lambda L(\theta^*, \lambda^*) = 0. \tag{1}$$

This characterization motivates gradient-based methods for finding such saddle points. A standard approach is the primal–dual gradient dynamic (PDGD) method (Nedić & Ozdaglar, 2009), whose discrete-time iteration is given as

$$\theta_{k+1} = \theta_k - \alpha_k \nabla_\theta L(\theta_k, \lambda_k), \quad \lambda_{k+1} = \lambda_k + \alpha_k \nabla_\lambda L(\theta_k, \lambda_k).$$

Moreover, we can also consider a continuous-time counterpart of PDGD given as

$$\dot{\theta}_t = -\nabla_\theta L(\theta_t, \lambda_t), \ \dot{\lambda}_t = \nabla_\lambda L(\theta_t, \lambda_t), \tag{2}$$

which converges to a saddle point under standard assumptions (Nedić & Ozdaglar, 2009; Qu & Li, 2018). When only stochastic approximations of the gradients are available, the corresponding stochastic formulation is given as follows:

$$\theta_{k+1} = \theta_k - \alpha_k \nabla_\theta \big(L(\theta_k, \lambda_k) + v_k\big), \quad \lambda_{k+1} = \lambda_k + \alpha_k \nabla_\lambda \big(L(\theta_k, \lambda_k) + w_k\big),$$

where $(v_k, w_k) \in \mathbb{R}^n \times \mathbb{R}^n$ are i.i.d. zero-mean noise terms. We refer to this version as stochastic PDGD. The stochastic PDGD is also known to converge to a saddle point in a probabilistic sense (Chen & Wang, 2016; Wang & Chen, 2016).

2.3 REVIEW OF GTD2

In this section, we provide a brief overview of the GTD2 algorithm introduced in Sutton et al. (2009) which tries to solve the policy evaluation problem. In the policy evaluation problem with linear function approximation, the Bellman equation $\Phi\theta = R^\pi + \gamma P^\pi \Phi\theta$ generally does not admit a solution within the feature subspace. To address this issue, GTD2 considers the projected Bellman equation (PBE):

$$\Phi\theta = \Pi_{\mathcal{R}(\Phi)}\big(R^\pi + \gamma P^\pi \Phi\theta\big), \tag{3}$$

which seeks a solution within the range space of Φ. Here, $\Pi_{\mathcal{R}(\Phi)}$ denotes the projection onto the range space of Φ, denoted by $\mathcal{R}(\Phi)$: $\Pi_{\mathcal{R}(\Phi)}(x) := \arg\min_{x' \in \mathcal{R}(\Phi)} \|x - x'\|^2_{D^\beta}$. GTD2 formulates this problem by minimizing the mean-square projected Bellman error (MSPBE), whose minimizer coincides with the solution of the PBE. It then seeks this solution in an incremental and model-free manner via stochastic gradient updates:

Problem 2*.* Solve for $\theta \in \mathbb{R}^q$ the optimization

$$\min_{\theta \in \mathbb{R}^q} \mathrm{MSPBE}(\theta) := \frac{1}{2} \|\Pi_{\mathcal{R}(\Phi)}(R^\pi + \gamma P^\pi \Phi\theta) - \Phi\theta\|^2_{D^\beta}.$$

Here, $\|x\|_{D^\beta}$ is defined as $\sqrt{x^\top D^\beta x}$ for any positive-definite matrix D^β. Note that minimizing the objective means minimizing the error of the PBE in (3) with respect to $\|\cdot\|_{D^\beta}$. Moreover, in the

objective of Problem 2, d^β depends on the behavior policy, β, while P^π and R^π depend on the target policy, π, that we want to evaluate. This structure allows us to obtain an off-policy learning algorithm through the importance sampling (Precup et al., 2001) or sub-sampling techniques (Sutton et al., 2008).

Although GTD2 is a promising algorithm with convergence guarantees under off-policy training, its theoretical analysis relies on the following assumption:

Assumption 2.2. $\Phi^\top D^\beta(\gamma P^\pi - I)\Phi$ is nonsingular, where I denotes the identity matrix with an appropriate dimension.

Note that Assumption 2.2 is common in the literature, and is adopted in Sutton et al. (2008; 2009); Ghiassian et al. (2020) for convergence of GTD methods. In this paper, we refer to $\Phi^\top D^\beta(\gamma P^\pi - I)\Phi$ as the *feature interaction matrix* (FIM), and we say the problem is *ill-conditioned* when the FIM is singular or nearly singular. Under Assumption 2.2, the MSPBE minimization in Problem 2 has a unique solution. In particular, the GTD2 solution is given by

$$\theta_{\mathrm{GTD2}} = -\big(\Phi^\top D^\beta(\gamma P^\pi - I)\Phi\big)^{-1}\Phi^\top D^\beta R^\pi, \tag{4}$$

as shown in Macua et al. (2014), which coincides with the unique solution of the PBE under Assumption 2.2. More recently, to provide a deeper optimization-theoretic perspective, Lee et al. (2022) showed that the MSPBE minimization in Problem 2 can be equivalently expressed as the following constrained optimization problem:

Problem 3. Solve for $\theta \in \mathbb{R}^q$ the optimization

$$\min_{\theta\in\mathbb{R}^q} 0 \quad \text{s.t.} \quad 0 = \Phi^\top D^\beta(R^\pi + \gamma P^\pi\Phi\theta - \Phi\theta).$$

It can be easily proved that the equality constraint in Problem 3 is equivalent to the PBE in (3). Specifically, this constraint can be rewritten as $\Phi^\top D^\beta(I - \gamma P^\pi)\Phi\theta = \Phi^\top D^\beta R^\pi$, which leads to the same solution as the GTD2 algorithm in (4). This constrained formulation not only characterizes the PBE solution but also naturally admits an equivalent saddle-point representation.

3 Regularized GTD

Before diving into the derivation of our algorithm, we provide a brief roadmap of the optimization problems introduced in this section and the next. We first propose a regularized optimization objective (Problem 4), which establishes the foundation for both the proposed algorithm and its saddle-point convergence analysis. To derive the closed-form solution of the proposed algorithm, we then recast this objective into a min-max saddle-point formulation (Problem 5). Finally, to facilitate the analysis of the error bounds, we present an equivalent unconstrained formulation (Problem 6).

Following this framework, we reformulate the policy evaluation problem by extending the constrained optimization in Problem 3. First, we observe that adding a quadratic term $\frac{1}{2}\theta^\top\Phi^\top D^\beta\Phi\theta$ to the objective of Problem 3 does not shift the optimal solution θ_{GTD2} as long as the FIM is nonsingular (Assumption 2.2). This term serves as a base for regularization without altering the solution. However, when the FIM is singular, the added quadratic term selects a particular solution among them, thereby altering the optimal solution. Next, to ensure the problem remains well-defined even when FIM is singular, we introduce an auxiliary variable $w \in \mathbb{R}^q$ into the constraint. This w acts as a slack variable that allows for a small error in the PBE. Finally, to prevent this error from becoming large, we incorporate a penalty term $\frac{c}{2}w^\top\Phi^\top D^\beta\Phi w$ into the objective, where $c > 0$ is a regularization coefficient. This leads to the following regularized optimization problem:

Problem 4. Solve for $\theta, w \in \mathbb{R}^q$ the optimization

$$\begin{aligned}\min_{\theta,w\in\mathbb{R}^q} \frac{c}{2}w^\top\Phi^\top D^\beta\Phi w + \frac{1}{2}\theta^\top\Phi^\top D^\beta\Phi\theta \\ \text{s.t.} \quad 0 = \Phi^\top D^\beta(R^\pi + \gamma P^\pi\Phi\theta - \Phi\theta + \Phi w).\end{aligned} \tag{5}$$

Here, we allow a weighted error term Φw in the constraint compared to Problem 3, which effectively introduces a small error in the PBE in (3). The variable w represents an auxiliary variable capturing the error of the PBE, and $c > 0$ is a regularization parameter that controls its penalty. Note that the

regularization parameter c controls the trade-off between the constraint satisfaction and the objective penalty. A detailed practical guidelines for selecting c in practice are provided in Section L of Appendix. Based on this formulation, we derive the R-GTD algorithm in this paper. The first step in this derivation is to recast Problem 4 as a min–max saddle-point problem in Problem 1, for which we now introduce the corresponding Lagrangian function

$$L(\theta, w, \lambda) = \frac{c}{2} w^\top \Phi^\top D^\beta \Phi w + \frac{1}{2}\theta^\top \Phi^\top D^\beta \Phi\theta + \lambda^\top \Phi^\top D^\beta (R^\pi + \gamma P^\pi \Phi\theta - \Phi\theta + \Phi w), \quad (6)$$

where we can prove that $L(\theta, w, \lambda)$ is strongly convex in the primal variable θ and w, concave in the dual variable λ, and coercive due to the positive definite weighting matrix. Therefore, by Lemma A.4 of Appendix, the problem satisfies the structural conditions of the min–max formulation in Problem 1 and admits a saddle point. Moreover, due to the strong convexity of $L(\theta, w, \lambda)$ in the primal variables (θ, w) and its concavity in the dual variable λ, the min–max problem admits a unique saddle-point solution. The corresponding min-max saddle-point problem is summarized below for convenience.

Problem 5. Solve for $\theta, w, \lambda \in \mathbb{R}^q$ the optimization

$$\min_{\theta, w \in \mathbb{R}^q} \max_{\lambda \in \mathbb{R}^q} L(\theta, w, \lambda) := \frac{c}{2} w^\top \Phi^\top D^\beta \Phi w + \frac{1}{2}\theta^\top \Phi^\top D^\beta \Phi\theta + \lambda^\top \Phi^\top D^\beta \times (R^\pi + \gamma P^\pi \Phi\theta - \Phi\theta + \Phi w).$$

Solving the stationary conditions in (1) yields a closed-form characterization of the optimal solution $(\theta_{\mathrm{RGTD}}, \lambda_{\mathrm{RGTD}}, w_{\mathrm{RGTD}})$ to the min-max problem in Problem 5. For clarity, the explicit expressions of the optimal solution are summarized below.

Proposition 3.1. *Under Assumption 2.2, let $(\theta_{\mathrm{RGTD}}, \lambda_{\mathrm{RGTD}}, w_{\mathrm{RGTD}})$ denote the optimal solution to the min-max problem in Problem 5. Then the optimal solution admits the following closed-form expressions:*

$$\theta_{\mathrm{RGTD}} = \underbrace{-\big(\Phi^\top D^\beta(\gamma P^\pi - I)\Phi\big)^{-1}\Phi^\top D^\beta R^\pi}_{\theta_{GTD2}} \underbrace{-\big(\Phi^\top D^\beta(\gamma P^\pi - I)\Phi\big)^{-1}\Phi^\top D^\beta \Phi\, w_{\mathrm{RGTD}}}_{error}, \quad (7)$$

$$\lambda_{\mathrm{RGTD}} = -\big(\Phi^\top D^\beta(\gamma P^\pi - I)\Phi\big)^{-\top}\Phi^\top D^\beta \Phi\, \theta_{\mathrm{RGTD}}, \quad (8)$$

$$w_{\mathrm{RGTD}} = -\frac{1}{c}\lambda_{\mathrm{RGTD}}. \quad (9)$$

The proof of Proposition 3.1 is given in Section B of Appendix. The characterization in Proposition 3.1 reveals that the R-GTD solution θ_{RGTD} does not exactly coincide with the GTD2 solution θ_{GTD2} for finite values of the parameter c. In particular, compared to θ_{GTD2} in (4), the θ_{RGTD} contains an additional bias term $-(\Phi^\top D^\beta(\gamma P^\pi - I)\Phi)^{-1}\Phi^\top D^\beta \Phi\, w_{\mathrm{RGTD}}$, which originates from the auxiliary variable w_{RGTD} introduced in (5). Moreover, by direct manipulations of (7)–(9), one can easily prove that θ_{RGTD} is explicitly given by the following equation:

$$\theta_{\mathrm{RGTD}} = -\Big((\Phi^\top D^\beta(\gamma P^\pi - I)\Phi)^\top (\Phi^\top D^\beta \Phi)^{-1}(\Phi^\top D^\beta(\gamma P^\pi - I)\Phi) + \tfrac{1}{c}\Phi^\top D^\beta \Phi\Big)^{-1} \times (\Phi^\top D^\beta(\gamma P^\pi - I)\Phi)^\top (\Phi^\top D^\beta \Phi)^{-1}\Phi^\top D^\beta R^\pi. \quad (10)$$

We can see that as $c \to \infty$, the solution (10) converges to the GTD2 solution in (4) provided that the FIM, $\Phi^\top D^\beta(\gamma P^\pi - I)\Phi$, is nonsingular. When FIM is singular, then still the term inside $\Big((\Phi^\top D^\beta(\gamma P^\pi - I)\Phi)^\top (\Phi^\top D^\beta \Phi)^{-1}(\Phi^\top D^\beta(\gamma P^\pi - I)\Phi) + \frac{1}{c}\Phi^\top D^\beta \Phi\Big)^{-1}$ is nonsingular. Therefore, the R-GTD solutions is still well defined in this case. Consequently, under Assumption 2.2, letting $c \to \infty$ eliminates the effect of the auxiliary variable, yielding $\theta_{\mathrm{RGTD}} \to \theta_{\mathrm{GTD2}}, w_{\mathrm{RGTD}} \to 0$, while $\lambda_{\mathrm{RGTD}} \to -\big(\Phi^\top D^\beta(\gamma P^\pi - I)\Phi\big)^{-\top}\Phi^\top D^\beta \Phi\, \theta_{\mathrm{GTD2}}$. Having characterized the optimal solution and its properties, we now derive the practical update rules for R-GTD. As discussed earlier, a standard approach for solving such convex–concave saddle-point problems is to apply the PDGD. Accordingly, we first consider the continuous-time PDGD in (2) associated with the Lagrangian $L(\theta, w, \lambda)$ in (6), given by

$$\dot{\theta}_t = -\nabla_\theta L(\theta_t, w_t, \lambda_t) = -\Phi^\top D^\beta \Phi\, \theta_t - \big((\gamma P^\pi \Phi)^\top D^\beta \Phi - \Phi^\top D^\beta \Phi\big)\lambda_t,$$

$$\dot{w}_t = -\nabla_w L(\theta_t, w_t, \lambda_t) = -c\,\Phi^\top D^\beta \Phi\, w_t - \Phi^\top D^\beta \Phi\, \lambda_t,$$

$$\dot{\lambda}_t = \nabla_\lambda L(\theta_t, w_t, \lambda_t) = \Phi^\top D^\beta (R^\pi + \gamma P^\pi \Phi\theta_t - \Phi\theta_t + \Phi w_t). \quad (11)$$

Algorithm 1 R-GTD

1: Set the step-size $(\alpha_k)_{k=0}^{\infty}$ and regularization parameter c. Initialize $(\theta_0, w_0, \lambda_0)$.
2: **for** $k \in \{0, \dots\}$ **do**
3: Observe $s_k \sim d^\beta$, $a_k \sim \beta(\cdot|s_k)$, and $s'_k \sim P(\cdot|s_k, a_k)$, $r_{k+1} := r(s_k, a_k, s'_k)$.
4: Update parameters according to

$$\theta_{k+1} = \theta_k + \alpha_k((\phi_k - \gamma\rho_k\phi'_k)(\phi_k^\top \lambda_k) - \phi_k(\phi_k^\top \theta_k)),$$
$$w_{k+1} = w_k + \alpha_k(-c\phi_k^\top w_k - \phi_k^\top \lambda_k)\phi_k,$$
$$\lambda_{k+1} = \lambda_k + \alpha_k(\delta_k + \phi_k^\top w_k)\phi_k,$$

where $\phi_k := \phi(s_k), \phi'_k := \phi(s'_k), \rho_k := \frac{\pi(a_k|s_k)}{\beta(a_k|s_k)}$, and $\delta_k = \rho_k r_{k+1} + \gamma\rho_k(\phi'_k)^\top\theta_k - \phi_k^\top\theta_k$.
5: **end for**

Its discrete-time counterpart (by Euler discretization) is

$$\begin{aligned}
\theta_{k+1} =& \theta_k + \alpha_k(-\Phi^\top D^\beta \Phi\theta_k - ((\gamma P^\pi \Phi)^\top D^\beta \Phi - \Phi^\top D^\beta \Phi)\lambda_k),\\
w_{k+1} =& w_k + \alpha_k(-c\Phi^\top D^\beta \Phi w_k - \Phi^\top D^\beta \Phi\lambda_k),\\
\lambda_{k+1} =& \lambda_k + \alpha_k\Phi^\top D^\beta(R^\pi + \gamma P^\pi \Phi\theta_k - \Phi\theta_k + \Phi w_k),
\end{aligned} \tag{12}$$

where the scalar $\alpha_k > 0$ denotes the step size at iteration k and $e_s \in \mathbb{R}^{|\mathcal{S}|}$ denotes the standard basis vector associated with state s, whose s-th entry is one and all other entries are zero. Considering $s_k \sim d^\beta$, $a_k \sim \pi(\cdot \mid s_k)$, and $s'_k \sim P(\cdot \mid s_k, a_k)$, The corresponding stochastic PDGD can then be obtained as follows:

$$\begin{aligned}
\theta_{k+1} =& \theta_k + \alpha_k(-\Phi^\top e_{s_k} e_{s_k}^\top \Phi\theta_k - ((\gamma e_{s_k} e_{s'_k} \Phi)^\top e_{s_k} e_{s_k}^\top \Phi - \Phi^\top e_{s_k} e_{s_k}^\top \Phi)\lambda_k),\\
w_{k+1} =& w_k + \alpha_k(-c\Phi^\top e_{s_k} e_{s_k}^\top \Phi w_k - \Phi^\top e_{s_k} e_{s_k}^\top \Phi\lambda_k),\\
\lambda_{k+1} =& \lambda_k + \alpha_k\Phi^\top e_{s_k} e_{s_k}^\top (e_{s_k} r(s_k, a_k, s'_k) + \gamma e_{s_k} e_{s'_k}^\top \Phi\theta_k - \Phi\theta_k + \Phi w_k),
\end{aligned}$$

where s_k is the state sampled at iteration k, and s'_k is the next state observed after taking an action a_k according to the behavior policy β. The resulting stochastic PDGD updates constitute the R-GTD algorithm, summarized in Algorithm 1, which incorporates importance sampling for off-policy data (Precup et al., 2001). In particular, it includes the ratio $\rho_k := \pi(a_k \mid s_k)/\beta(a_k \mid s_k)$ to correct for the mismatch between the behavior policy β, which generates the data, and the target policy π, whose value function is being evaluated. Intuitively, the importance sampling ratio reweights each sample so that, in expectation, the update direction corresponds to that of on-policy learning under π, as commonly done in off-policy temporal-difference methods (Precup et al., 2001; Sutton et al., 2008).

Finally, to facilitate the convergence analysis in the next section, we introduce an alternative but equivalent unconstrained optimization formulation of Problem 4 as follows:

Problem 6. Solve for $\theta \in \mathbb{R}^q$ the optimization

$$\min_{\theta\in\mathbb{R}^q} \frac{c}{2}\|\Pi_{\mathcal{R}(\Phi)}(R^\pi + \gamma P^\pi \Phi\theta - \Phi\theta)\|_{D^\beta}^2 + \frac{1}{2}\|\Phi\theta\|_{D^\beta}^2. \tag{13}$$

An important feature of the above formulation is that it is an unconstrained optimization while Problem 4 is a constrained optimization. This alternative formulation will serve as a convenient basis for analyzing the stability and convergence properties of R-GTD. Moreover, Problem 6 can be interpreted as a regularized MSPBE formulation, where the first term is the MSPBE which is the objective of GTD2, while the second term is a quadratic regularization term. As $c \to \infty$, Problem 6 (and equivalently Problem 4) yields the same solution as GTD2. The proof is given in Section C of Appendix.

4 Convergence of R-GTD

In this section, we analyze the theoretical properties of R-GTD in two parts. First, we establish convergence to a saddle point under the convex–concave formulation. Second, we derive error bounds relating the R-GTD solution to the true projected solution.

4.1 CONVERGENCE OF R-GTD TOWARDS SADDLE-POINT

In this section, we analyze the convergence of R-GTD towards a saddle point by studying its continuous-time dynamics and the associated discrete-time algorithm. Specifically, we leverage recent results on PDGD from Qu & Li (2018) to establish the convergence of the continuous-time PDGD, and subsequently apply the ordinary differential equation (ODE) method of Borkar & Meyn (2000) to prove the convergence of the stochastic R-GTD iterates. To analyze the convergence of R-GTD, we consider a constrained optimization in Problem 4. Specifically, we express this constrained optimization in Problem 4 in the following general form:

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{s.t.} \quad Ax = b, \tag{14}$$

where f is convex and continuously differentiable. As discussed earlier, a standard approach to solving (14) is to introduce the associated convex–concave Lagrangian function (Boyd & Vandenberghe, 2004), defined as

$$L(x,\lambda) = f(x) + \lambda^\top (Ax - b). \tag{15}$$

By invoking a key lemma from Qu & Li (2018), whose statement and proof are provided in Lemma A.5 of Appendix, we establish the convergence of the continuous-time PDGD associated with R-GTD. When the objective function f is strongly convex and the constraint matrix A has full row rank, Qu & Li (2018) shows that the continuous-time PDGD of (15),

$$\dot{x} = -\nabla f(x) - A^\top \lambda, \qquad \dot{\lambda} = Ax - b,$$

globally converges to the unique saddle point of the associated Lagrangian. However, in the optimization form of GTD2 in Problem 3, the constraint matrix is given by $A = \Phi^\top D^\beta(\gamma P^\pi - I)\Phi$, which refers to the FIM introduced earlier. This matrix satisfies the full row rank condition only under Assumption 2.2. If this assumption is violated, the constraint matrix becomes rank-deficient, and the convergence guarantee provided by Qu & Li (2018) no longer applies. When applying the R-GTD formulation, the constraint matrix A in Problem 4 takes the block form $A = \begin{bmatrix} \text{FIM} & \Phi^\top D^\beta \Phi \end{bmatrix}$, where $\Phi^\top D^\beta \Phi \succ 0$. As a consequence of the additional $\Phi^\top D^\beta \Phi$ term, the resulting constraint matrix A is always full row rank, regardless of whether Assumption 2.2 holds. Consequently, the conditions required by Qu & Li (2018) are always satisfied for R-GTD, which ensures global convergence of the associated continuous-time PDGD. For completeness, a detailed verification of this rank condition is provided in Section D of Appendix. Based on this continuous-time convergence result and Lemma A.3 of Appendix, convergence of Algorithm 1 using ODE approach can be proved as follows:

Theorem 4.1. *Let us consider Algorithm 1, and assume that the step-size satisfies* (24)*. Then,* $(\theta_k, w_k, \lambda_k) \to (\theta_{\text{RGTD}}, w_{\text{RGTD}}, \lambda_{\text{RGTD}})$ *as* $k \to \infty$ *with probability one (see Section E of Appendix for the proof).*

While Theorem 4.1 establishes asymptotic convergence, a finite-time analysis can also be carried out using standard arguments, but we do not pursue it here since it is not the main focus of this work.

4.2 ERROR BOUNDS BETWEEN THE R-GTD AND THE TRUE PROJECTED SOLUTION

While we have established convergence to the R-GTD saddle point, it is essential to quantify its proximity to the true value function V^π. To this end, we introduce the true projected solution θ_*^π, as considered in Lee (2024), as the vector which satisfies $\Phi\theta_*^\pi = \Pi_{\mathcal{R}(\Phi)} V^\pi$, where V^π admits the representation $V^\pi = \sum_{k=0}^\infty \gamma^k (P^\pi)^k R^\pi$. This θ_*^π corresponds to the minimizer of the projection error $f(\theta) = \frac{1}{2}\|V^\pi - \Phi\theta\|_{D^\beta}^2$. To analyze the error bound, we first introduce the following notations, which will be used repeatedly in the sequel.

Notation Recall that the FIM is defined as $\text{FIM} := \Phi^\top D^\beta(\gamma P^\pi - I)\Phi$. For notational convenience, we denote this matrix by $M := \Phi^\top D^\beta(\gamma P^\pi - I)\Phi$, and define

$$B := \Phi^\top D^\beta \Phi,\ b := \Phi^\top D^\beta R^\pi,\ G := M^\top B^{-1} M,$$

where $B \succ 0$. When M is nonsingular, we further define $K := G^{-1} B G^{-1} M^\top B^{-1} b$. Using this notation, the MSPBE objective admits the quadratic form $\text{MSPBE}(\theta) = \frac{1}{2}(M\theta + b)^\top B^{-1}(M\theta + b)$. Therefore, the corresponding first-order optimality condition can be compactly written as follows:

$$\nabla_\theta \text{MSPBE}(\theta) = M^\top B^{-1}(M\theta + b) = 0. \tag{16}$$

The structure of the GTD2 solution implied by this condition is summarized in the following lemma:

Lemma 4.2. *The GTD2 stationary condition* (16) *admits an affine solution set*

$$\Theta_{\mathrm{GTD2}} := \{\theta \in \mathbb{R}^q : M^\top B^{-1}(M\theta + b) = 0\}. \tag{17}$$

When M (FIM) is nonsingular, this set reduces to the singleton

$$\theta_{\mathrm{GTD2}} = -(M^\top B^{-1} M)^{-1} M^\top B^{-1} b. \tag{18}$$

To derive the subsequent error bounds, we rewrite the optimality condition of the R-GTD problem in (13) as an equivalent linear system. This reformulation yields $\left(M^\top B^{-1} M + \frac{1}{c} B\right)\theta_{\mathrm{RGTD}} = -M^\top B^{-1} b$. The derivation of this linear system form is provided in Lemma F.1 of Appendix. Consequently, the R-GTD solution admits the closed-form expression

$$\theta_{\mathrm{RGTD}} := -\left(M^\top B^{-1} M + \tfrac{1}{c} B\right)^{-1} M^\top B^{-1} b. \tag{19}$$

Note that the expression in (19) coincides with the closed-form solution given in (10).

We now characterize the asymptotic relationship between θ_{RGTD} and θ_{GTD2} as the regularization parameter c increases in the following lemma:

Lemma 4.3. *Let M, B, b, G, K be defined as above, and let θ_{RGTD} be given by* (19)*. Let $\Pi^B_{\mathcal{N}(G)}$ denote the projection onto $\mathcal{N}(G)$ under the B-norm, i.e., $\Pi^B_{\mathcal{N}(G)}(\theta_{\mathrm{GTD2}}) = \arg\min_{v \in \mathcal{N}(G)} \|\theta_{\mathrm{GTD2}} - v\|_B$. If M is singular, then there exists a constant $c_0 := \left\|G^\dagger B \left(I - (\Pi_{\mathcal{N}(G)} B \Pi_{\mathcal{N}(G)})^\dagger B\right)\right\|_2$ such that for all $c > c_0$, every $\theta_{\mathrm{GTD2}} \in \Theta_{\mathrm{GTD2}}$ defined in* (17) *admits the expansion*

$$\theta_{\mathrm{RGTD}} = \left(\theta_{\mathrm{GTD2}} - \Pi^B_{\mathcal{N}(G)}(\theta_{\mathrm{GTD2}})\right) + O\left(\frac{1}{c}\right). \tag{20}$$

If M is nonsingular, then there exists a constant $c_0 := \|G^{-1}B\|_2$ such that for all $c > c_0$, the unique GTD2 solution θ_{GTD2} given in (4.2) *admits the expansion*

$$\theta_{\mathrm{RGTD}} = \theta_{\mathrm{GTD2}} + \frac{1}{c} K + O\left(\frac{1}{c^2}\right). \tag{21}$$

The proof is given in Section G of Appendix. While the nonsingular case is relatively straightforward as θ_{RGTD} converges to a unique point, the singular case involves a solution set Θ_{GTD2} that is an affine subspace. We therefore focus on the singular case to clarify how the regularization influences the limiting behavior of θ_{RGTD} within this solution set. Figure 1 provides a geometric illustration of the expansion in (20). In particular, as $c \to \infty$, the $O(1/c)$ term vanishes and the right-hand side of (20) reduces to unique B-minimum-norm element $(I - \Pi^B_{\mathcal{N}(G)})\theta_{\mathrm{GTD2}}$ of the GTD2 solution set, as depicted in the figure. To build intuition for the singular-case asymptotic behavior in (20), we provide a toy example in Section M of Appendix. Building on these properties, we further quantify the Euclidean distance between $\Phi\theta_{\mathrm{RGTD}}$ and the GTD2 solution set in Lemma H.1 of Appendix. These results lead to our main theorem, which establishes explicit error bounds between the R-GTD solution and the true projected solution.

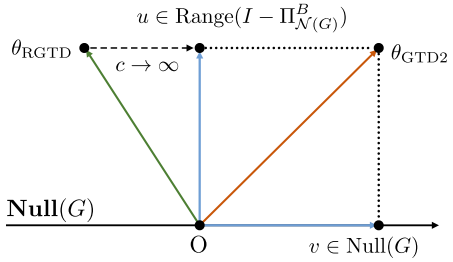


Figure 1: As $c \to \infty$, the R-GTD solution θ_{RGTD} converges to the GTD2 solution θ_{GTD2}. θ_{GTD2} decomposes uniquely into two components: $v \in \mathrm{Null}(G)$ along the null space of G, and $u \in \mathrm{Range}(I - \Pi^B_{\mathcal{N}(G)})$.

Theorem 4.4. *Let $\mathrm{dist}(x, S) := \inf_{y \in S} \|x - y\|_2$ denote the Euclidean distance from a point x to a set S. Then, the prediction error satisfies*

$$\|\Phi\theta_{\mathrm{RGTD}} - \Phi\theta^\pi_*\|_2 \le \mathrm{dist}(\Phi\theta^\pi_*,\ \Phi\Theta_{\mathrm{GTD2}}) + \left\|\Pi_{\Phi\Theta_{\mathrm{GTD2}}}(\Phi\theta_{\mathrm{RGTD}}) - \Pi_{\Phi\Theta_{\mathrm{GTD2}}}(\Phi\theta^\pi_*)\right\|_2 + O\left(\tfrac{1}{c}\right).$$

If M is nonsingular, then $\Theta_{\mathrm{GTD2}} = \{\theta_{\mathrm{GTD2}}\}$. In this case, the bound simplifies to

$$\|\Phi\theta_{\mathrm{RGTD}} - \Phi\theta^\pi_*\|_2 \le \|\Phi\theta_{\mathrm{GTD2}} - \Phi\theta^\pi_*\|_2 + \frac{\|\Phi\|_2 \|K\|_2}{c} + O\left(\tfrac{1}{c^2}\right).$$

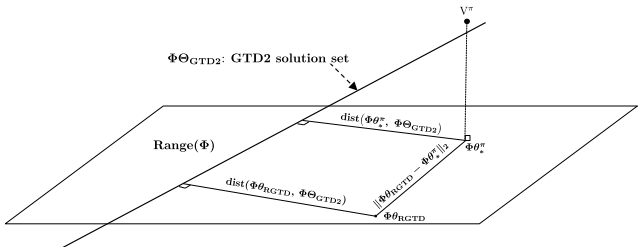


Figure 2: Geometric illustration of the singular case. The error bounds of R-GTD and GTD2 are compared through the projection and distance terms in Theorem 4.4.

The proof is given in Section I of Appendix. Theorem 4.4 highlights the geometric structure of the prediction error by decomposing the error bound into two distinct components. In the *ill-conditioned case*, where Assumption 2.2 does not hold, the first term, $dist(\Phi\theta^\pi_*, \Phi\Theta_{GTD2})$, represents the intrinsic approximation error of GTD2 arising from its affine solution set. The second term reflects the structural difference of the specific representative chosen by R-GTD within this set. Crucially, because the original MSPBE objective lacks strong convexity in the singular setting, the limiting behavior of specific stochastic GTD2 trajectories cannot be uniquely characterized, which necessitates a set-based theoretical comparison. In contrast, the regularization in R-GTD ensures a strongly convex objective, guaranteeing that all stochastic trajectories reliably converge to a unique point. Figure 2 visualizes this singular case decomposition. In the *nonsingular case*, GTD2 already provides the optimal approximation, but R-GTD converges to the same solution as $c \to \infty$. Thus, even in nonsingular settings, R-GTD does not degrade performance: its error is at most an $O(1/c)$ bias away from GTD2.

5 Experiments

We consider an illustrative singular setting example designed to highlight the effect of R-GTD. Specifically, we construct a scenario where the matrix FIM is ill-conditioned, with its smallest singular value on the order of 10^{-13}. We initialize θ_0 with a null-space component of FIM. The regularization coefficients are chosen as $c \in \{0.2, 0.4, 1\}$. We compare R-GTD with regularization-based baselines Ghiassian et al. (2020); Yu (2017). Note that while Zhang et al. (2020) also employs a similar regularization, it is fundamentally an off-policy control algorithm. Therefore, we restrict our empirical comparison to Ghiassian et al. (2020); Yu (2017), which directly target the same problem setting. Each experiment is repeated 30 times, and the distribution of $\|\theta_k - \theta^\pi_*\|_2$ is summarized using box-and-whisker plots. Detailed experimental settings are provided in Section J.2 of Appendix. As shown in Figure 3, R-GTD suppresses the null-space component and converges stably toward θ^π_*. Additional experiment results are provided in Section J of Appendix.

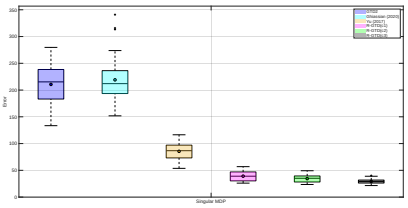


Figure 3: Box-and-whisker representation of $\|\theta_k - \theta^\pi_*\|_2$ over 30 independent runs with step-size $\alpha_k = 1/(k+30)$.

6 Conclusion and Limitations

This paper introduced R-GTD, a regularized variant of GTD2 derived from a convex–concave saddle-point formulation. The proposed approach relaxes the convergence assumptions of GTD2 while preserving the core theoretical structure of GTD2. Beyond convergence, we further provided a geometric analysis that characterizes the limiting behavior of the R-GTD solution relative to the GTD2 solution set across both nonsingular and singular regimes. As a result, the method achieves improved stability in ill-conditioned settings. We proved asymptotic convergence to the saddle point and established error bounds with respect to the true projected solution. However, the current analysis is limited to linear function approximation. Extending the framework to nonlinear function approximation remain important directions for future work.

AI USE STATEMENT

In this work, we did not use generative AI tools for tasks requiring mandatory disclosure under the ICLR 2027 AI Policy. We used generative AI tools to aid and polish the writing of the manuscript and for retrieval and discovery of relevant literature. We reviewed and verified all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

This work primarily makes theoretical contributions. The assumptions, the R-GTD algorithm, and the main theoretical results are stated in the main text, while complete proofs and supporting derivations are provided in Sections A to I of Appendix. The experimental settings and additional results, including nonsingular and singular MDPs, Baird's counterexample, Frozen Lake, and a condition-number sweep, are documented in Section J of Appendix. The sensitivity analysis and practical guidelines for selecting the regularization parameter c are provided in Section L, and Section M of Appendix presents a fully specified toy example illustrating the singular regime.

REFERENCES

Leemon Baird et al. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the twelfth international conference on machine learning*, pp. 30–37, 1995.

S.. Bhatnagar, HL Prasad, and LA Prashanth. *Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods*. Springer, 2013.

Åke Björck. *Numerical methods for least squares problems*. SIAM, 2024.

Vivek S Borkar and Sean P Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Yichen Chen and Mengdi Wang. Stochastic primal-dual methods and sample complexity of reinforcement learning. *arXiv preprint arXiv:1612.02516*, 2016.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International conference on machine learning*, pp. 1125–1134. PMLR, 2018.

Sina Ghiassian, Andrew Patterson, Shivam Garg, Dhawal Gupta, Adam White, and Martha White. Gradient temporal-difference learning with regularized corrections. In *International Conference on Machine Learning*, pp. 3524–3534. PMLR, 2020.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Hassan K Khalil and Jessy W Grizzle. *Nonlinear systems*, volume 3. Prentice hall Upper Saddle River, NJ, 2002.

Harold J Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*. Springer, 2003.

Donghwan Lee. Analysis of off-policy multi-step td-learning with linear function approximation. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pp. 1153–1160. IEEE, 2024.

Donghwan Lee, Han-Dong Lim, Jihoon Park, and Okyong Choi. New versions of gradient temporal-difference learning. *IEEE Transactions on Automatic Control*, 68(8):5006–5013, 2022.

Sergio Valcarcel Macua, Jianshu Chen, Santiago Zazo, and Ali H Sayed. Distributed policy evaluation under multiple behavior strategies. *IEEE Transactions on Automatic Control*, 60(5):1260–1274, 2014.

Carl D Meyer. *Matrix analysis and applied linear algebra*. SIAM, 2023.

Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228, 2009.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006.

Doina Precup, Richard S Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In *ICML*, pp. 417–424, 2001.

Guannan Qu and Na Li. On the exponential stability of primal-dual gradient dynamics. *IEEE Control Systems Letters*, 3(1):43–48, 2018.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, USA, 1970.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3 (1):9–44, 1988.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Richard S Sutton, Hamid 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*, 21, 2008.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th annual international conference on machine learning*, pp. 993–1000, 2009.

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

Mengdi Wang and Yichen Chen. An online primal-dual method for discounted markov decision processes. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 4516–4521. IEEE, 2016.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

Huizhen Yu. On convergence of some gradient-based temporal-differences algorithms for off-policy learning. *arXiv preprint arXiv:1712.09652*, 2017.

Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. In *International Conference on Machine Learning*, pp. 11204–11213. PMLR, 2020.

A MATHEMATICAL PRELIMINARIES

To facilitate the convergence analysis of the proposed regularized GTD (R-GTD) algorithm, this appendix summarizes several mathematical preliminaries that are used throughout the main text. We collect the basic notions from nonlinear system theory, the ordinary differential equation (ODE)-based stochastic approximation framework, and key results on primal–dual gradient dynamics (PDGD). These tools provide the technical foundation for the stability proofs, ODE limiting arguments, and saddle-point convergence results appearing in Section 4. For completeness, we restate the assumptions and lemmas in a self-contained manner so that readers can verify the theoretical arguments without referring to external sources.

A.1 BASICS OF NONLINEAR SYSTEM THEORY

We briefly review basic concepts from nonlinear system theory, which will serve as a foundation for the convergence analysis and stochastic approximation methods developed later. Consider the nonlinear dynamical system

$$\frac{d}{dt}x_t = f(x_t), \quad x_0 \in \mathbb{R}^n, \quad t \geq 0, \tag{22}$$

where $x_t \in \mathbb{R}^n$ denotes the state, $t \geq 0$ is the time variable, $x_0 \in \mathbb{R}^n$ is the initial condition, and $f : \mathbb{R}^n \to \mathbb{R}^n$ is a nonlinear mapping. For simplicity, we assume that the solution to (22) exists and is unique, which holds whenever f is globally Lipschitz continuous.

Lemma A.1. *Khalil & Grizzle (2002) Consider the nonlinear system* (22)*, and suppose that f is globally Lipschitz continuous, i.e., $\|f(x) - f(y)\| \leq l\|x - y\|$ for all $x, y \in \mathbb{R}^n$ and some constant $l > 0$ with respect to a norm $\|\cdot\|$. Then, for any $x_0 \in \mathbb{R}^n$, the system admits a unique solution x_t for all $t \geq 0$.*

An equilibrium point is a fundamental notion in nonlinear system theory. A point $x_\infty \in \mathbb{R}^n$ is called an equilibrium of (22) if the state remains constant at x_∞ whenever $x_0 = x_\infty$ (Khalil & Grizzle, 2002). Equivalently, equilibrium points correspond to the real roots of $f(x) = 0$. An equilibrium x_∞ is said to be *globally asymptotically stable* if $x_t \to x_\infty$ as $t \to \infty$ for all initial states $x_0 \in \mathbb{R}^n$.

A.2 ODE-BASED STOCHASTIC APPROXIMATION

Owing to its broad applicability, the convergence of many reinforcement learning (RL) algorithms is commonly analyzed using the ODE method (Bhatnagar et al., 2013; Kushner & Yin, 2003). This approach investigates the stability of an associated limiting ODE to understand the behavior of stochastic recursions with diminishing step sizes, which asymptotically follow the trajectory of the ODE. Among such techniques, one of the most influential is the ODE method proposed by Borkar & Meyn (2000). In the following, we provide a brief overview of their ODE-based convergence analysis framework for a general stochastic recursion of the form

$$\theta_{k+1} = \theta_k + \alpha_k\big(f(\theta_k) + \varepsilon_{k+1}\big), \tag{23}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a nonlinear mapping.

The Borkar–Meyn theorem establishes that under Assumption A.2, the stochastic process $(\theta_k)_{k=0}^{\infty}$ generated by (23) is almost surely bounded and converges to the equilibrium point θ_∞. This result forms the theoretical basis for proving the convergence of the proposed R-GTD algorithm. The following technical assumptions are standard in stochastic approximation theory:

Assumption A.2.

1. The mapping $f : \mathbb{R}^n \to \mathbb{R}^n$ is globally Lipschitz continuous, and there exists a function $f_\infty : \mathbb{R}^n \to \mathbb{R}^n$ such that $\lim_{r\to\infty} \frac{f(rx)}{r} = f_\infty(x)$ for all $x \in \mathbb{R}^n$.
2. The origin is an asymptotically stable equilibrium of the ODE $\dot{\theta}_t = f_\infty(\theta_t)$.
3. The ODE $\dot{\theta}_t = f(\theta_t)$ admits a unique globally asymptotically stable equilibrium $\theta_\infty \in \mathbb{R}^n$, i.e., $\theta_t \to \theta_\infty$ as $t \to \infty$.

4. The sequence $\{\varepsilon_k, \mathcal{G}_k, k \geq 1\}$, with $\mathcal{G}_k = \sigma(\theta_i, \varepsilon_i, i \leq k)$, is a Martingale difference sequence. Furthermore, there exists a constant $C_0 < \infty$ such that for any initial $\theta_0 \in \mathbb{R}^n$,
$$\mathbb{E}\big[\|\varepsilon_{k+1}\|_2^2 \mid \mathcal{G}_k\big] \leq C_0\big(1 + \|\theta_k\|_2^2\big), \quad \forall k \geq 0,$$
where $\|\cdot\|_2$ denotes the Euclidean norm.

5. The step-size sequence $\{\alpha_k\}$ satisfies
$$\alpha_k > 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{24}$$

Lemma A.3 (Borkar and Meyn theorem, Borkar & Meyn (2000))**.** *Suppose that Assumption A.2 holds. Then, the following statements are true:*

1. *For any initial condition $\theta_0 \in \mathbb{R}^n$, the iterates remain bounded, i.e., $\sup_{k\geq 0} \|\theta_k\|_2 < \infty$ with probability one.*
2. *Moreover, $\theta_k \to \theta_\infty$ as $k \to \infty$ with probability one.*

The preceding assumptions, together with the Borkar–Meyn theorem, provide the ODE-based foundation needed to analyze the stochastic approximation aspects of the R-GTD algorithm. In addition, to ensure that the limiting ODE admits a well-defined equilibrium, we first establish the existence of a saddle point for the associated min–max formulation. Beyond existence, our convergence analysis further requires understanding the stability properties of the continuous-time PDGD induced by this saddle-point structure. The following lemmas summarize key results on the existence of saddle points and the exponential stability of the corresponding PDGD:

Lemma A.4. *Rockafellar (1970) Let $L(\theta, \lambda)$ be a function that is convex in θ and concave in λ. Assume that*

- *for every λ, the function $L(\cdot, \lambda)$ admits a minimizer, and*
- *the dual function $g(\lambda) := \min_\theta L(\theta, \lambda)$ is upper bounded and attains its maximum.*

Then the saddle value exists and satisfies
$$\min_\theta \max_\lambda L(\theta, \lambda) = \max_\lambda \min_\theta L(\theta, \lambda).$$

Lemma A.5. *Qu & Li (2018) Consider the equality constrained optimization in* (14)*, and suppose that f is twice differentiable, μ-strongly convex, and l-smooth, i.e, for all $x, y \in \mathbb{R}^n$, $\mu\|x-y\|_2^2 \leq (\nabla f(x) - \nabla f(y))^\top (x-y) \leq l\|x-y\|_2^2$. Moreover, suppose that A is full row rank. Consider the corresponding Lagrangian function* (15)*. Then, the corresponding saddle-point (θ^*, λ^*) is unique, and the corresponding continuous-time PDGD,*
$$\dot{x}_t = -\nabla_\theta L(\theta_t, \lambda_t) = -\nabla_\theta f(\theta_t) - A^\top \lambda_t, \quad \dot{\lambda}_t = \nabla_\lambda L(\theta_t, \lambda_t) = A\theta_t - b,$$
exponentially converges to (θ^*, λ^*).

Beyond the stability properties of the primal–dual dynamics, our convergence analysis of the R-GTD algorithm toward the true value function V^π also relies on a structural property of a key matrix arising in the analysis.

Recall that
$$M = \Phi^\top D^\beta(\gamma P^\pi - I)\Phi, \quad B = \Phi^\top D^\beta \Phi, \quad b = \Phi^\top D^\beta R^\pi,$$
where $B \succ 0$, and define $G := M^\top B^{-1} M$ which is positive semidefinite.

To facilitate the singular-case analysis in Section 4.2, we first collect several standard auxiliary results from linear algebra. These properties will be used to characterize the limiting R-GTD solution when G is singular. Note that the following results do not rely on any new technical insight; they follow directly from standard results in linear algebra and are stated here for completeness and later use.

Lemma A.6. *(Nocedal & Wright, 2006, p. 603) Let $G \in \mathbb{R}^{d\times d}$ be a symmetric matrix. Then the space admits the orthogonal decomposition*
$$\mathbb{R}^d = \mathrm{Range}(G) \ \oplus \ \mathrm{Null}(G),$$
where $\oplus$ denotes the orthogonal direct sum*, i.e., the sum of two mutually orthogonal subspaces.*

Lemma A.7. *(Boyd & Vandenberghe, 2004, p. 649) Let $G \in \mathbb{R}^{m\times n}$ have singular value decomposition $G = U\Sigma V^\top$, and let $G^\dagger$ denote its Moore–Penrose pseudoinverse. Then the matrices*

$$GG^\dagger = UU^\top, \qquad G^\dagger G = VV^\top$$

are the orthogonal projection matrices onto $\mathrm{Range}(G)$ *and* $\mathrm{Range}(G^\top)$*, respectively.*

Lemma A.8. *Let $G \in \mathbb{R}^{n\times n}$ be a symmetric matrix and $\Pi_{\mathcal{N}(G)}$ be the orthogonal projection onto* $\mathrm{Null}(G)$*. Then, the following identities hold:*

$$\Pi_{\mathcal{N}(G)} G^\dagger = 0$$

Proof. By the properties of the Moore-Penrose inverse, the range of $G^\dagger$ is equal to the range of $G^\top$. Since G is symmetric, we have $\mathrm{Range}(G^\dagger) = \mathrm{Range}(G)$ (Meyer, 2023, p. 424). By Lemma A.6, the range and null space of a symmetric matrix are orthogonal, i.e., $\mathrm{Range}(G) \perp \mathrm{Null}(G)$. Since $\Pi_{\mathcal{N}(G)}$ is the orthogonal projection onto $\mathrm{Null}(G)$, it maps any vector in $\mathrm{Range}(G)$ to zero. Consequently, for any vector x, $G^\dagger x \in \mathrm{Range}(G)$, which implies $\Pi_{\mathcal{N}(G)}(G^\dagger x) = 0$. Thus, $\Pi_{\mathcal{N}(G)} G^\dagger = 0$.

□

Lemma A.9. *Let $B \succ 0$, and let $U \in \mathbb{R}^{n\times r}$ have full column rank with*

$$\mathrm{Range}(U) = \mathcal{N}(G).$$

Then the B-orthogonal projection onto $\mathcal{N}(G)$, where $\langle x, y\rangle_B := x^\top By$, is given by

$$\Pi^B_{\mathcal{N}(G)} = U(U^\top BU)^{-1}U^\top B.$$

Proof. For any $x \in \mathbb{R}^n$, its B-orthogonal projection onto $\mathcal{N}(G)$ is defined as

$$\Pi^B_{\mathcal{N}(G)} x = \arg\min_{z\in\mathcal{N}(G)} \|x - z\|_B^2, \qquad \|y\|_B^2 := y^\top By.$$

Since $\mathrm{Range}(U) = \mathcal{N}(G)$, every $z \in \mathcal{N}(G)$ can be written as $z = Ua$. Hence, the above problem is equivalent to

$$\min_{a\in\mathbb{R}^r} (x - Ua)^\top B(x - Ua).$$

By the normal-equation characterization of weighted least squares (Björck, 2024), the optimal coefficient $a^\star$ satisfies

$$U^\top BUa^\star = U^\top Bx.$$

Since $B \succ 0$ and U has full column rank, $U^\top BU \succ 0$, and therefore

$$a^\star = (U^\top BU)^{-1}U^\top Bx.$$

Thus,

$$\Pi^B_{\mathcal{N}(G)} x = Ua^\star = U(U^\top BU)^{-1}U^\top Bx.$$

Since this holds for every $x \in \mathbb{R}^n$,

$$\Pi^B_{\mathcal{N}(G)} = U(U^\top BU)^{-1}U^\top B.$$

□

Lemma A.10. *Let $B \succ 0$, and let U have orthonormal columns. Define $P := UU^\top$. Then*

$$(PBP)^\dagger = U(U^\top BU)^{-1}U^\top.$$

Proof. Since

$$PBP = U(U^\top BU)U^\top,$$

and $B \succ 0$ with U having full column rank, we have $U^\top BU \succ 0$, so $(U^\top BU)^{-1}$ exists.

Using $U^\top U = I$, we obtain

$$(PBP)\,U(U^\top BU)^{-1}U^\top = UU^\top = P,$$

and similarly,

$$U(U^\top BU)^{-1}U^\top(PBP) = P.$$

Therefore,

$$(PBP)\,U(U^\top BU)^{-1}U^\top(PBP) = P(PBP) = PBP,$$

and

$$U(U^\top BU)^{-1}U^\top(PBP)U(U^\top BU)^{-1}U^\top = P\,U(U^\top BU)^{-1}U^\top = U(U^\top BU)^{-1}U^\top.$$

Moreover, since $P = P^\top$, both products above are symmetric. Hence, by the Penrose characterization of the Moore–Penrose pseudoinverse (Björck, 2024),

$$(PBP)^\dagger = U(U^\top BU)^{-1}U^\top.$$

□

Lemma A.11. *Let $B \succ 0$, and let $P := \Pi_{\mathcal{N}(G)}$ denote the Euclidean orthogonal projection onto $\mathcal{N}(G)$. Then the B-orthogonal projection onto $\mathcal{N}(G)$ is*

$$\Pi^B_{\mathcal{N}(G)} = (PBP)^\dagger B = (PBP)^\dagger PB.$$

Proof. Let U have orthonormal columns spanning $\mathcal{N}(G)$, so that $P = UU^\top$. By Lemma A.9, the B-orthogonal projection onto $\mathrm{Range}(U) = \mathcal{N}(G)$ is

$$\Pi^B_{\mathcal{N}(G)} = U(U^\top BU)^{-1}U^\top B.$$

Moreover,

$$PBP = U(U^\top BU)U^\top,$$

and since $U^\top BU \succ 0$ and by Lemma A.10

$$(PBP)^\dagger = U(U^\top BU)^{-1}U^\top.$$

Therefore,

$$\Pi^B_{\mathcal{N}(G)} = (PBP)^\dagger B.$$

Furthermore, using $P = UU^\top$ and $U^\top U = I$,

$$\begin{aligned}(PBP)^\dagger P &= U(U^\top BU)^{-1}U^\top UU^\top \\ &= U(U^\top BU)^{-1}U^\top \\ &= (PBP)^\dagger.\end{aligned}$$

Therefore, multiplying both sides by B from the right gives

$$(PBP)^\dagger PB = (PBP)^\dagger B.$$

Hence,

$$\Pi^B_{\mathcal{N}(G)} = (PBP)^\dagger B = (PBP)^\dagger PB.$$

□

B Proof of Proposition 3.1

Proposition 3.1 (restated). Under Assumption 2.2, let $(\theta_{\mathrm{RGTD}}, \lambda_{\mathrm{RGTD}}, w_{\mathrm{RGTD}})$ denote the optimal solution to the min-max problem in Problem 5. Then the optimal solution admits the following closed-form expressions:

$$\theta_{\mathrm{RGTD}} = \underbrace{-\big(\Phi^\top D^\beta(\gamma P^\pi - I)\Phi\big)^{-1}\Phi^\top D^\beta R^\pi}_{\theta_{GTD2}} \underbrace{-\big(\Phi^\top D^\beta(\gamma P^\pi - I)\Phi\big)^{-1}\Phi^\top D^\beta\Phi\, w_{\mathrm{RGTD}}}_{\text{error}},$$

$$\lambda_{\mathrm{RGTD}} = -\big(\Phi^\top D^\beta(\gamma P^\pi - I)\Phi\big)^{-\top}\Phi^\top D^\beta\Phi\,\theta_{\mathrm{RGTD}},$$

$$w_{\mathrm{RGTD}} = -\frac{1}{c}\,\lambda_{\mathrm{RGTD}}.$$

The optimal solution $(\theta_{\mathrm{RGTD}}, w_{\mathrm{RGTD}}, \lambda_{\mathrm{RGTD}})$of Problem 5 is characterized by the first-orderstationarity conditions with respect to the primal variables $(\theta, w)$and the dual variable λ. Taking gradients of the Lagrangian

$$L(\theta, w, \lambda) = \frac{c}{2} w^\top \Phi^\top D^\beta \Phi w + \frac{1}{2} \theta^\top \Phi^\top D^\beta \Phi \theta + \lambda^\top \Phi^\top D^\beta (R^\pi + \gamma P^\pi \Phi\theta - \Phi\theta + \Phi w)$$

and setting them to zero yields

$$\nabla_\theta L = \Phi^\top D^\beta \Phi\theta + (\gamma P^\pi \Phi - \Phi)^\top D^\beta \Phi\lambda = 0, \tag{25}$$

$$\nabla_w L = c\,\Phi^\top D^\beta \Phi w + \Phi^\top D^\beta \Phi\lambda = 0, \tag{26}$$

$$\nabla_\lambda L = \Phi^\top D^\beta (R^\pi + \gamma P^\pi \Phi\theta - \Phi\theta + \Phi w) = 0. \tag{27}$$

From (26), we directly obtain

$$w = -\frac{1}{c}\,\lambda. \tag{28}$$

From (27), we have:

$$\Phi^\top D^\beta (\gamma P^\pi - I)\Phi\theta + \Phi^\top D^\beta R^\pi + \Phi^\top D^\beta \Phi w = 0.$$

Under the assumption that the feature interaction matrix $\Phi^\top D^\beta(\gamma P^\pi - I)\Phi$ is nonsingular, we can solve this equation directly for θ, yielding:

$$\theta = -\big(\Phi^\top D^\beta(\gamma P^\pi - I)\Phi\big)^{-1}\Phi^\top D^\beta R^\pi - \big(\Phi^\top D^\beta(\gamma P^\pi - I)\Phi\big)^{-1}\Phi^\top D^\beta \Phi w,$$

which exactly yields the expression for θ_{RGTD} in (7). Next, from (25), we can rearrange the terms as:

$$(\gamma P^\pi \Phi - \Phi)^\top D^\beta \Phi\lambda = -\Phi^\top D^\beta \Phi\theta.$$

Since the matrix $\Phi^\top D^\beta(\gamma P^\pi - I)\Phi$ is nonsingular, its transpose is also nonsingular. Solving for λ directly yields:

$$\lambda = -\big(\Phi^\top D^\beta(\gamma P^\pi - I)\Phi\big)^{-\top}\Phi^\top D^\beta \Phi\theta,$$

which gives the expression for λ_{RGTD}. Finally, the expression for w_{RGTD} follows directly from (28), completing the proof.

C Proof of Problem 6

Problem 6 (restated). Solve for $\theta \in \mathbb{R}^q$ the optimization

$$\min_{\theta\in\mathbb{R}^q} \frac{c}{2}\|\Pi_{\mathcal{R}(\Phi)}(R^\pi + \gamma P^\pi \Phi\theta - \Phi\theta)\|^2_{D^\beta} + \frac{1}{2}\|\Phi\theta\|^2_{D^\beta}.$$

This problem is equivalent to Problem 4 in the sense that both formulations admit the same unique optimal solution for θ.

We begin by restating the optimization problem in Problem 4 for clarity. We aim to find $(\theta, w) \in \mathbb{R}^q$ that solve

$$\begin{aligned} \min_{\theta, w\in\mathbb{R}^q} \quad & \frac{c}{2} w^\top \Phi^\top D^\beta \Phi w + \frac{1}{2}\theta^\top \Phi^\top D^\beta \Phi\theta \\ \text{subject to} \quad & 0 = \Phi^\top D^\beta \left(R^\pi + \gamma P^\pi \Phi\theta - \Phi\theta + \Phi w\right). \end{aligned} \tag{29}$$

We now analyze the constraint (29). The constraint can be written as

$$0 = \Phi^\top D^\beta \left(R^\pi + \gamma P^\pi \Phi\theta - \Phi\theta + \Phi w\right).$$

Rearranging the terms yields

$$\Phi^\top D^\beta \Phi w = -\Phi^\top D^\beta \left(R^\pi + \gamma P^\pi \Phi\theta - \Phi\theta\right). \tag{30}$$

Since Φ has full column rank, the matrix $\Phi^\top D^\beta \Phi$ is positive definite. Therefore, it is invertible, and we can solve for w as

$$w = -(\Phi^\top D^\beta \Phi)^{-1} \Phi^\top D^\beta (R^\pi + \gamma P^\pi \Phi\theta - \Phi\theta).$$

This implies:

$$\Phi w = -\Pi_{\mathcal{R}(\Phi)} (R^\pi + \gamma P^\pi \Phi\theta - \Phi\theta)$$

where $\Pi_{\mathcal{R}(\Phi)} := \Phi(\Phi^\top D^\beta \Phi)^{-1} \Phi^\top D^\beta$ is the projection matrix onto Range(Φ). Substituting this expression for Φw back into the objective function of Problem 4 yields:

$$\frac{c}{2} \|\Pi_{\mathcal{R}(\Phi)} (R^\pi + \gamma P^\pi \Phi\theta - \Phi\theta)\|^2_{D^\beta} + \frac{1}{2} \|\Phi\theta\|^2_{D^\beta},$$

which matches exactly the objective function of Problem 6. Therefore, minimizing Problem 4 with respect to (θ, w) under the given constraint is equivalent to directly minimizing Problem 6 with respect to θ.

D Verification of the full row rank condition

Proposition D.1. *Consider the trajectory (θ_t, w, λ_t) of the PDGD in (11), and let $(\theta_{\mathrm{RGTD}}, w_{\mathrm{RGTD}}, \lambda_{\mathrm{RGTD}})$ be the corresponding saddle point. Then, $(\theta_t, w_t, \lambda_t) \to (\theta_{\mathrm{RGTD}}, w_{\mathrm{RGTD}}, \lambda_{\mathrm{RGTD}})$ exponentially as $t \to \infty$.*

Proof. Note that (5) has a strongly convex, smooth, and twice differentiable objective function. Moreover, constraint term can be presented as $0 = \Phi^\top D^\beta R^\pi + \begin{bmatrix} \gamma \Phi^\top D^\beta P^\pi \Phi - \Phi^\top D^\beta \Phi & \Phi^\top D^\beta \Phi \end{bmatrix} \begin{bmatrix} \theta \\ w \end{bmatrix}$. which is full row rank by assumption. These conditions ensure that all assumptions of Lemma A.5 hold. Hence, by Lemma A.5, the PDGD of Problem 4, given in (11), is globally asymptotically stable, and converges to its equilibrium point $(\theta_{RGTD}, w_{RGTD}, \lambda_{RGTD})$. □

E Proof of Theorem 4.1

Theorem 4.1 (restated). Let us consider Algorithm 1, and assume that the step-size satisfies (24). Then, $(\theta_k, w_k, \lambda_k) \to (\theta_{\mathrm{RGTD}}, w_{\mathrm{RGTD}}, \lambda_{\mathrm{RGTD}})$ as $k \to \infty$ with probability one.

The algorithm (12) can be rewritten by

$$\begin{bmatrix} \theta_{k+1} \\ w_{k+1} \\ \lambda_{k+1} \end{bmatrix} = \begin{bmatrix} \theta_k \\ w_k \\ \lambda_k \end{bmatrix} + \alpha_k \left(f\left(\begin{bmatrix} \theta_k \\ w_k \\ \lambda_k \end{bmatrix} \right) + \varepsilon_{k+1} \right),$$

where

$$f\left(\begin{bmatrix} \theta \\ w \\ \lambda \end{bmatrix} \right) := A \begin{bmatrix} \theta \\ w \\ \lambda \end{bmatrix} + b,$$

$$A = \begin{bmatrix} -\Phi^\top D^\beta \Phi & 0 & -\Phi^\top (\gamma P^\pi - I)^\top D^\beta \Phi \\ 0 & -c\,\Phi^\top D^\beta \Phi & -\Phi^\top D^\beta \Phi \\ \Phi^\top D^\beta (\gamma P^\pi - I)\Phi & \Phi^\top D^\beta \Phi & 0 \end{bmatrix}, \tag{31}$$

$$b = \begin{bmatrix} 0 \\ 0 \\ \Phi^\top D^\beta R^\pi \end{bmatrix}. \tag{32}$$

and

$$\varepsilon_{k+1} = \begin{bmatrix} -\Phi^\top e_{s_k} e_{s_k}^\top \Phi\theta_k - ((\gamma\rho_k e_{s_k} e_{s'_k}^\top \Phi)^\top e_{s_k} e_{s_k}^\top \Phi - \Phi^\top e_{s_k} e_{s_k}^\top \Phi)\lambda_k) \\ -c\Phi^\top e_{s_k} e_{s_k}^\top \Phi w_k - \Phi^\top e_{s_k} e_{s_k}^\top \Phi\lambda_k \\ \Phi^\top e_{s_k} e_{s_k}^\top (\rho_k e_{s_k} r(s_k, a_k, s'_k) + \gamma\rho_k e_{s_k} e_{s'_k}^\top \Phi\theta_k - \Phi\theta_k + \Phi w_k) \end{bmatrix} - f\left(\begin{bmatrix} \theta_k \\ w_k \\ \lambda_k \end{bmatrix}\right).$$

The proof is completed by examining all the statements in Assumption A.2. To prove the first statement of Assumption A.2, we have

$$\lim_{c\to\infty} f\left(r\begin{bmatrix} \theta \\ w \\ \lambda \end{bmatrix}\right)/r = f_\infty\left(\begin{bmatrix} \theta \\ w \\ \lambda \end{bmatrix}\right) = A\begin{bmatrix} \theta \\ w \\ \lambda \end{bmatrix}.$$

Moreover, since f is affine in its argument, it is globally Lipschitz continuous. The second statement of Assumption A.2: The PDGD of Problem 4 can be written as

$$\frac{d}{dt}\begin{bmatrix} \theta_t - \theta_\infty \\ w_t - w_\infty \\ \lambda_t - \lambda_\infty \end{bmatrix} = \bar{A}\begin{bmatrix} \theta_t - \theta_\infty \\ w_t - w_\infty \\ \lambda_t - \lambda_\infty \end{bmatrix},$$

where $\bar{A}$ is as follows

$$\bar{A} = \begin{bmatrix} -\Phi^\top D^\beta \Phi & 0 & -\Phi^\top(\gamma P^\pi - I)^\top D^\beta \Phi \\ 0 & -c\,\Phi^\top D^\beta \Phi & -\Phi^\top D^\beta \Phi \\ \Phi^\top D^\beta(\gamma P^\pi - I)\Phi & \Phi^\top D^\beta \Phi & 0 \end{bmatrix},$$

by Lemma A.5 and Proposition D.1 the origin is the globally asymptotically stable equilibrium point. Now, one can observe that it is identical to the ODE

$$\frac{d}{dt}\begin{bmatrix} \theta - \theta_\infty \\ w - w_\infty \\ \lambda - \lambda_\infty \end{bmatrix} = f_\infty\left(\begin{bmatrix} \theta_t - \theta_\infty \\ w_t - w_\infty \\ \lambda_t - \lambda_\infty \end{bmatrix}\right).$$

Therefore, its origin is the globally asymptotically stable equilibrium point. The third statement of Assumption A.2: The ODE, $\frac{d}{dt}\begin{bmatrix} \theta_t \\ w_t \\ \lambda_t \end{bmatrix} = f\left(\begin{bmatrix} \theta_t \\ w_t \\ \lambda_t \end{bmatrix}\right)$, is identical to the PDGD of Problem 4. Therefore, it admits a unique globally asymptotically stable equilibrium point by Lemma A.5. Next, we prove the remaining parts. Recall that the R-GTD update. Define the history $\mathcal{G}_k$ as

$$\mathcal{G}_k := (\varepsilon_k, \varepsilon_{k-1}, \ldots, \varepsilon_1, \theta_k, \theta_{k-1}, \ldots, \theta_0, w_k, w_{k-1}, \ldots, w_0, \lambda_k, \lambda_{k-1}, \ldots, \lambda_0)$$

and the process $(M_k)_{k=0}^\infty$ with $M_k := \sum_{i=1}^k \varepsilon_i$. Then, we can prove that $(M_k)_{k=0}^\infty$ is Martingale. To do so, we first prove $\mathbb{E}[\varepsilon_{k+1}|\mathcal{G}_k] = 0$ by

$$\begin{aligned}
&\mathbb{E}[\varepsilon_{k+1}|\mathcal{G}_k] \\
&= \mathbb{E}\left[\begin{bmatrix} -\Phi^\top e_{s_k} e_{s_k}^\top \Phi\theta_k - ((\gamma\rho_k e_{s_k} e_{s'_k}^\top \Phi)^\top e_{s_k} e_{s_k}^\top \Phi - \Phi^\top e_{s_k} e_{s_k}^\top \Phi)\lambda_k) \\ -c\Phi^\top e_{s_k} e_{s_k}^\top \Phi w_k - \Phi^\top e_{s_k} e_{s_k}^\top \Phi\lambda_k \\ \Phi^\top e_{s_k} e_{s_k}^\top (\rho_k e_{s_k} r(s_k, a_k, s'_k) + \gamma\rho_k e_{s_k} e_{s'_k}^\top \Phi\theta_k - \Phi\theta_k + \Phi w_k) \end{bmatrix} \middle| \mathcal{G}_k\right] - \mathbb{E}\left[f\left(\begin{bmatrix} \theta_k \\ w_k \\ \lambda_k \end{bmatrix}\right) \middle| \mathcal{G}_k\right] \\
&= \mathbb{E}\left[f\left(\begin{bmatrix} \theta_k \\ w_k \\ \lambda_k \end{bmatrix}\right) \middle| \mathcal{G}_k\right] - \mathbb{E}\left[f\left(\begin{bmatrix} \theta_k \\ w_k \\ \lambda_k \end{bmatrix}\right) \middle| \mathcal{G}_k\right] = 0,
\end{aligned}$$

where the second equality is due to the i.i.d. assumption of samples. Using this identity, we have

$$\begin{aligned}
\mathbb{E}[M_{k+1}|\mathcal{G}_k] =& \mathbb{E}\left[\sum_{i=1}^{k+1} \varepsilon_i \middle| \mathcal{G}_k\right] = \mathbb{E}[\varepsilon_{k+1}|\mathcal{G}_k] + \mathbb{E}\left[\sum_{i=1}^{k} \varepsilon_i \middle| \mathcal{G}_k\right] \\
=& \mathbb{E}\left[\sum_{i=1}^{k} \varepsilon_i \middle| \mathcal{G}_k\right] = \sum_{i=1}^{k} \varepsilon_i = M_k.
\end{aligned}$$

Therefore, $(M_k)_{k=0}^\infty$ is a Martingale sequence, and $\varepsilon_{k+1} = M_{k+1} - M_k$ is a Martingale difference. Moreover, it can be easily proved that the second statement of the fourth condition of Assumption A.2 is satisfied by algebraic calculations. Therefore, the fourth condition is met.

F Deriving the normal equation for the R-GTD optimality condition

We begin by recalling the notational conventions used throughout the main text:

$$M = \Phi^\top D^\beta(\gamma P^\pi - I)\Phi, \quad B = \Phi^\top D^\beta \Phi, \quad b = \Phi^\top D^\beta R^\pi,$$

where $B \succ 0$, and define $G := M^\top B^{-1} M$, $K := G^{-1} B G^{-1} M^\top B^{-1} b$.

This appendix provides the identity that allows the optimality condition of the R-GTD problem to be rewritten in a linear normal-equation form. Before deriving closed-form expressions of the R-GTD solution and establishing error bounds, it is necessary to show that the optimal parameter θ satisfies

$$\big(M^\top B^{-1} M + \tfrac{1}{c} B\big)\theta = -M^\top B^{-1} b.$$

The lemma below formally establishes this relation, which serves as the starting point for the analysis developed in the main text.

Lemma F.1. *Let M, B, and b be defined as in the notation above. Then the optimal solution θ of Problem 6 satisfies the linear system*

$$\big(M^\top B^{-1} M + \tfrac{1}{c} B\big)\theta = -M^\top B^{-1} b.$$

Proof. From Problem 6, using the projection $\Pi_{\mathcal{R}(\Phi)} = \Phi B^{-1}\Phi^\top D^\beta$ and the definitions $M = \Phi^\top D^\beta(\gamma P^\pi - I)\Phi$, $B = \Phi^\top D^\beta \Phi$, and $b = \Phi^\top D^\beta R^\pi$, we can rewrite the objective as

$$J(\theta) = \frac{c}{2}\theta^\top M^\top B^{-1} M\theta + c\,\theta^\top M^\top B^{-1} b + \frac{1}{2}\theta^\top B\theta + \text{const.}$$

Since $B \succ 0$, the problem is strongly convex. Taking the gradient with respect to θ yields

$$\nabla_\theta J(\theta) = c\,M^\top B^{-1} M\theta + c\,M^\top B^{-1} b + B\theta.$$

Setting $\nabla_\theta J(\theta) = 0$ gives

$$(c\,M^\top B^{-1} M + B)\theta = -c\,M^\top B^{-1} b,$$

and dividing both sides by c leads to

$$(M^\top B^{-1} M + \tfrac{1}{c} B)\theta = -M^\top B^{-1} b.$$

□

G Proof of Lemma 4.3

Lemma 4.3 (restated). Let M, B, b, G, K be defined as above, and let θ_{RGTD} be given by (19). Let $\Pi^B_{\mathcal{N}(G)}$ denote the B-orthogonal projection onto the null space of G, which gives the closest point in the null space under the B-norm, i.e., $\Pi^B_{\mathcal{N}(G)}(\theta_{\mathrm{GTD2}}) = \arg\min_{v\in\mathcal{N}(G)} \|\theta_{\mathrm{GTD2}} - v\|_B$. If M is singular, then there exists a constant $c_0 := \big\|G^\dagger B\left(I - (\Pi_{\mathcal{N}(G)} B \Pi_{\mathcal{N}(G)})^\dagger B\right)\big\|_2$ such that for all $c > c_0$, every $\theta_{\mathrm{GTD2}} \in \Theta_{\mathrm{GTD2}}$ defined in (17) admits the expansion

$$\theta_{\mathrm{RGTD}} = \big(\theta_{\mathrm{GTD2}} - \Pi^B_{\mathcal{N}(G)}(\theta_{\mathrm{GTD2}})\big) + O\left(\frac{1}{c}\right). \tag{33}$$

If M is nonsingular, then there exists a constant $c_0 := \|G^{-1}B\|_2$ such that for all $c > c_0$, the unique GTD2 solution θ_{GTD2} given in (4.2) admits the expansion

$$\theta_{\mathrm{RGTD}} = \theta_{\mathrm{GTD2}} + \frac{1}{c}K + O\left(\frac{1}{c^2}\right). \tag{34}$$

We prove the two cases separately.

(a) Nonsingular case. Let $G := M^\top B^{-1} M$. If G is nonsingular, the R-GTD solution can be written as

$$\theta_{RGTD} = -(G + \tfrac{1}{c}B)^{-1} M^\top B^{-1} b.$$

To expand the inverse of $G + \frac{1}{c}B$, we first rewrite the matrix as a perturbation of G:

$$G + \tfrac{1}{c}B = G\Big(I + \tfrac{1}{c}G^{-1}B\Big).$$

Since $G \succ 0$ in the nonsingular case, the matrix $G^{-1}B$ is well-defined. Moreover, for sufficiently large c, the operator norm satisfies

$$\big\|\tfrac{1}{c}G^{-1}B\big\|_2 < 1.$$

Define $c_0 := \|G^{-1}B\|_2$, then, for all $c > c_0$, we have $\|c^{-1}G^{-1}B\|_2 < 1$, and the Neumann series expansion is valid.

Thus, the inverse of $I + X$ with

$$X := \tfrac{1}{c}G^{-1}B$$

can be expanded using the classical Neumann series (see Horn & Johnson (2012)):

$$(I + X)^{-1} = I - X + X^2 - X^3 + \cdots, \qquad \|X\|_2 < 1.$$

Applying this expansion yields

$$(G + \tfrac{1}{c}B)^{-1} = \Big(I + \tfrac{1}{c}G^{-1}B\Big)^{-1} G^{-1} = \big(I - \tfrac{1}{c}G^{-1}B + \tfrac{1}{c^2}\,G^{-1}BG^{-1}B + \cdots\big)\, G^{-1}.$$

Multiplying out gives the first-order perturbation expansion

$$(G + \tfrac{1}{c}B)^{-1} = G^{-1} - \tfrac{1}{c}G^{-1}BG^{-1} + \tfrac{1}{c^2}\,G^{-1}BG^{-1}BG^{-1} + \cdots$$

Therefore,

$$\theta_{\mathrm{RGTD}} = -\Big(G^{-1} - \tfrac{1}{c}G^{-1}BG^{-1} + O(1/c^2)\Big) M^\top B^{-1} b.$$

Now, let us define

$$\theta_{\mathrm{GTD2}} := -G^{-1} M^\top B^{-1} b,$$

which coincides with the unique GTD2 solution in the nonsingular case (cf. (4.2) with $G = M^\top B^{-1} M$) and also define $K := G^{-1}BG^{-1}M^\top B^{-1}b$. Then, we obtain

$$\theta_{\mathrm{RGTD}} = \theta_{\mathrm{GTD2}} + \frac{1}{c}K + O\big(\tfrac{1}{c^2}\big)\,.$$

(b) Singular case. Suppose that M is singular, and hence $G = M^\top B^{-1} M$ is positive semidefinite but not invertible. Although G^{-1} does not exist, the regularized matrix $G + \frac{1}{c}B$ is invertible for every $c > 0$ because $B \succ 0$ implies $G + \frac{1}{c}B \succeq \frac{1}{c}B \succ 0$. From Lemma F.1, the exact normal equation is given by

$$(G + \frac{1}{c}B)\theta_{\mathrm{RGTD}} = -M^\top B^{-1} b.$$

Since $-M^\top B^{-1} b \in \mathrm{Range}(G)$, there exists a unique component of the GTD2 solution set lying in $\mathrm{Range}(G)$, which we denote by $\theta^{\perp}_{GTD2} = G^\dagger(-M^\top B^{-1} b)$. Here, we use Lemma A.7, which gives $G^\dagger G = \Pi_{\mathrm{Range}(G)}$, since G is symmetric. Let $P := \Pi_{\mathcal{N}(G)}$ be the orthogonal projection onto the null space of G. We decompose the R-GTD solution as

$$\theta_{\mathrm{RGTD}} = u_c + v_c, \quad u_c \in \mathrm{Range}(G), \quad v_c \in \mathrm{Null}(G).$$

Multiplying the normal equation by P from the left yields $P(G + \frac{1}{c}B)(u_c + v_c) = P(-M^\top B^{-1} b) = 0$. Since $PG = 0$, this simplifies to $PB(u_c + v_c) = 0$, which implies

$$PBPv_c = -PBu_c.$$

Because $B \succ 0$, the operator PBP is strictly positive definite on $\mathrm{Null}(G)$, allowing us to uniquely solve for

$$v_c = -(PBP)^\dagger PBu_c = -(PBP)^\dagger Bu_c,$$

where the second equality follows from Lemma A.11.

Thus, the R-GTD solution can be expressed purely in terms of u_c as $\theta_{\mathrm{RGTD}} = [I - (PBP)^\dagger B]u_c$. Next, multiplying the normal equation by $G^\dagger$ yields $u_c + \frac{1}{c}G^\dagger B\theta_{\mathrm{RGTD}} = G^\dagger(-M^\top B^{-1}b) = \theta^\perp_{\mathrm{GTD2}}$ where we use Lemma A.7, which gives $G^\dagger G = \Pi_{\mathrm{Range}(G)}$, together with $u_c \in \mathrm{Range}(G)$ and $v_c \in \mathrm{Null}(G)$. Substituting our expression for θ_{RGTD} into this equation gives:

$$\left[I + \frac{1}{c}G^\dagger B\left(I - (PBP)^\dagger B\right)\right] u_c = \theta^\perp_{\mathrm{GTD2}}.$$

Let $c_0 := \left\|G^\dagger B\left(I - (PBP)^\dagger B\right)\right\|_2$. For any $c > c_0$, the bracketed operator is invertible via the Neumann series, which rigorously establishes that $u_c = \theta^\perp_{\mathrm{GTD2}} + \mathcal{O}(1/c)$. Substituting this back into the expression for v_c yields the full expansion:

$$\theta_{\mathrm{RGTD}} = \theta^\perp_{\mathrm{GTD2}} - (PBP)^\dagger B\theta^\perp_{\mathrm{GTD2}} + \mathcal{O}(1/c).$$

We can interpret this algebraically via the B-inner product, defined as $\langle x, y\rangle_B = x^\top By$. Under this inner product, the B-orthogonal projection onto $\mathrm{Null}(G)$ is explicitly given by $\Pi^B_{\mathcal{N}(G)} := (PBP)^\dagger B$ as established in Lemma A.11. Consequently, the expansion becomes:

$$\theta_{\mathrm{RGTD}} = (I - \Pi^B_{\mathcal{N}(G)})\theta^\perp_{\mathrm{GTD2}} + \mathcal{O}(1/c).$$

Note that for any arbitrary GTD2 solution $\theta_{\mathrm{GTD2}} \in \Theta_{\mathrm{GTD2}}$, it decomposes as $\theta_{\mathrm{GTD2}} = \theta^\perp_{\mathrm{GTD2}} + v$ for some $v \in \mathrm{Null}(G)$. Since $\Pi^B_{\mathcal{N}(G)}v = v$, we have $(I - \Pi^B_{\mathcal{N}(G)})\theta_{\mathrm{GTD2}} = (I - \Pi^B_{\mathcal{N}(G)})\theta^\perp_{\mathrm{GTD2}}$. Thus, we obtain the final expansion:

$$\theta_{\mathrm{RGTD}} = \theta_{\mathrm{GTD2}} - \Pi^B_{\mathcal{N}(G)}\theta_{\mathrm{GTD2}} + \mathcal{O}(1/c).$$

Finally, taking the limit $c \to \infty$ yields $\lim_{c\to\infty}\theta_{\mathrm{RGTD}} = \theta_{\mathrm{GTD2}} - \Pi^B_{\mathcal{N}(G)}\theta_{\mathrm{GTD2}}$, which completes the proof.

H Auxiliary lemma for error bounds

Lemma H.1. *Let* $\mathrm{dist}(x, S) := \inf_{y\in S}\|x - y\|_2$ *denote the Euclidean distance from a point* x *to a convex set* S*. Then*

(a) Nonsingular case. *If* M *is nonsingular, the GTD2 solution* θ_{GTD2} *is exist. Define* $\Gamma := \|\Phi\|_2\|G^{-1}BG^{-1}BG^{-1}\|_2\|M^\top B^{-1}b\|_2$.

$$\|\Phi\theta_{\mathrm{RGTD}} - \Phi\theta_{\mathrm{GTD2}}\|_2 \le \frac{\|\Phi\|_2\,\|K\|_2}{c} + O\left(\tfrac{\Gamma}{c^2}\right).$$

(b) Singular case. *If* M *is singular, the GTD2 solution set is*

$$\Theta_{\mathrm{GTD2}} = \theta_{\mathrm{GTD2}} + \mathrm{Null}(G), \tag{35}$$

where θ_{GTD2} *denotes an arbitrary but fixed element of* Θ_{GTD2}*. Then*

$$\mathrm{dist}\left(\Phi\theta_{\mathrm{RGTD}},\, \Phi\Theta_{\mathrm{GTD2}}\right) = O\left(\tfrac{1}{c}\right). \tag{36}$$

Proof. Recall that

$$M = \Phi^\top D^\beta(\gamma P^\pi - I)\Phi, \quad B = \Phi^\top D^\beta\Phi \succ 0, \quad b = \Phi^\top D^\beta R^\pi,$$

and define $G := M^\top B^{-1}M$.

We consider the two cases separately.

(a) Nonsingular case. Let $G = M^\top B^{-1}M$, which is nonsingular. From Lemma 4.3, the R-GTD solution admits the expansion

$$\theta_{\mathrm{RGTD}} = \theta_{\mathrm{GTD2}} + \frac{1}{c}K + O\left(\tfrac{1}{c^2}\right), \quad K = G^{-1}BG^{-1}M^\top B^{-1}b.$$

Multiplying both sides by Φ, we get

$$\Phi\theta_{\mathrm{RGTD}} - \Phi\theta_{\mathrm{GTD2}} = \frac{1}{c}\Phi K + O\big(\tfrac{1}{c^2}\big)\,.$$

Taking the norm, we have

$$\begin{aligned}\|\Phi\theta_{\mathrm{RGTD}} - \Phi\theta_{\mathrm{GTD2}}\|_2 &\le \frac{1}{c}\|\Phi K\|_2 + \|\Phi \cdot O(1/c^2)\|_2 \\ &\le \frac{\|\Phi\|_2\,\|K\|_2}{c} + O\big(\tfrac{\Gamma}{c^2}\big)\,,\end{aligned}$$

where $\Gamma := \|\Phi\|_2\|G^{-1}BG^{-1}BG^{-1}\|_2\|M^\top B^{-1}b\|_2$. To see this, recall the higher-order expansion of the inverse from Lemma 4.3:

$$(G + \tfrac{1}{c}B)^{-1} = G^{-1} - \tfrac{1}{c}G^{-1}BG^{-1} + \tfrac{1}{c^2}G^{-1}BG^{-1}BG^{-1} + \cdots\,.$$

The term corresponding to $1/c^2$ in the expansion is $\frac{1}{c^2}\big(G^{-1}BG^{-1}BG^{-1}\big)M^\top B^{-1}b$. By the definition of Γ, the norm of this second-order term multiplied by Φ is bounded by Γ/c^2, which justifies the expansion.

(b) Singular case. When M is singular, the GTD2 solution set is the affine space

$$\Theta_{\mathrm{GTD2}} = \theta_{\mathrm{GTD2}} + \mathrm{Null}(G), \qquad \theta_{\mathrm{GTD2}} \in \Theta_{\mathrm{GTD2}}.$$

From the singular case analysis in Section G, the R-GTD solution satisfies

$$\theta_{\mathrm{RGTD}} = \theta_{\mathrm{GTD2}} - \Pi^B_{\mathcal{N}(\mathcal{G})}(\theta_{\mathrm{GTD2}}) + O\big(\tfrac{1}{c}\big)\,,$$

where the vector $\theta^\perp_{\mathrm{GTD2}} = \theta_{\mathrm{GTD2}} - \Pi^B_{\mathcal{N}(G)}(\theta_{\mathrm{GTD2}})$ is the unique component of θ_{GTD2} that is B-orthogonal to $\mathrm{Null}(G)$. Since $\theta^\perp_{\mathrm{GTD2}} \in \Theta_{\mathrm{GTD2}}$, by the definition of the distance to a set, we have

$$\mathrm{dist}(\Phi\theta_{\mathrm{RGTD}}, \Phi\Theta_{\mathrm{GTD2}}) \le \|\Phi(\theta_{\mathrm{RGTD}} - \theta^\perp_{\mathrm{GTD2}})\|_2.$$

Using the expansion above yields

$$\theta_{\mathrm{RGTD}} - \theta^\perp_{\mathrm{GTD2}} = O\big(\tfrac{1}{c}\big)\,,$$

and therefore

$$\mathrm{dist}(\Phi\theta_{\mathrm{RGTD}}, \Phi\Theta_{\mathrm{GTD2}}) \le \|\Phi\|_2\, O\big(\tfrac{1}{c}\big) = O\big(\tfrac{1}{c}\big)\,.$$

□

I Proof of Theorem 4.4

Theorem 4.4 (restated). The prediction error satisfies

$$\begin{aligned}\|\Phi\theta_{\mathrm{RGTD}} - \Phi\theta^\pi_*\|_2 \le \mathrm{dist}(\Phi\theta^\pi_*,\ \Phi\Theta_{\mathrm{GTD2}}) + \big\|\Pi_{\Phi\Theta_{\mathrm{GTD2}}}(\Phi\theta_{\mathrm{RGTD}}) - \Pi_{\Phi\Theta_{\mathrm{GTD2}}}(\Phi\theta^\pi_*)\big\|_2 \\ + O\big(\tfrac{1}{c}\big)\,.\end{aligned}$$

If M is nonsingular, then $\Theta_{\mathrm{GTD2}} = \{\theta_{\mathrm{GTD2}}\}$. In this case, the bound simplifies to

$$\|\Phi\theta_{\mathrm{RGTD}} - \Phi\theta^\pi_*\|_2 \le \|\Phi\theta_{\mathrm{GTD2}} - \Phi\theta^\pi_*\|_2 + \frac{\|\Phi\|_2\|K\|_2}{c} + O\big(\tfrac{1}{c^2}\big)\,.$$

We prove the two cases separately.

(a) Nonsingular case. From Lemma 4.3, the R-GTD solution admits the first-order expansion

$$\Phi\theta_{\mathrm{RGTD}} = \Phi\theta_{\mathrm{GTD2}} + \frac{1}{c}\Phi K + O\big(\tfrac{1}{c^2}\big)\,,$$

where

$$K := G^{-1}BG^{-1}M^\top B^{-1}b, \quad G := M^\top B^{-1}M.$$

Hence, subtracting $\Phi\theta_*^\pi$ and taking the 2-norm yields the asymptotic relation

$$\|\Phi\theta_{\mathrm{RGTD}} - \Phi\theta_*^\pi\|_2 = \|\Phi\theta_{\mathrm{GTD2}} - \Phi\theta_*^\pi + \tfrac{1}{c}\Phi K + O\big(\tfrac{1}{c^2}\big)\|_2 \le \|\Phi\theta_{\mathrm{GTD2}} - \Phi\theta_*^\pi\|_2 + \frac{\|\Phi\|_2\,\|K\|_2}{c} + O\big(\tfrac{1}{c^2}\big)\,.$$

This proves the non singular case.

(b) Singular case. When M is singular, the GTD2 solution set is the affine space $\Theta_{\mathrm{GTD2}} = \theta_{\mathrm{GTD2}} + \mathrm{Null}(G)$, and the distance is presented as (36). To relate θ_{RGTD} to the projected fixed point, insert and subtract the projection onto $\Phi\Theta_{\mathrm{GTD2}}$:

$$\|\Phi\theta_{\mathrm{RGTD}} - \Phi\theta_*^\pi\|_2 \le \mathrm{dist}(\Phi\theta_{\mathrm{RGTD}}, \Phi\Theta_{\mathrm{GTD2}}) + \mathrm{dist}(\Phi\theta_*^\pi, \Phi\Theta_{\mathrm{GTD2}}) + \|\Pi_{\Phi\Theta_{\mathrm{GTD2}}}(\Phi\theta_{\mathrm{RGTD}}) - \Pi_{\Phi\Theta_{\mathrm{GTD2}}}(\Phi\theta_*^\pi)\|_2.$$

Using the estimate $\mathrm{dist}(\Phi\theta_{\mathrm{RGTD}}, \Phi\Theta_{\mathrm{GTD2}}) = O(1/c)$ from above yields

$$\|\Phi\theta_{\mathrm{RGTD}} - \Phi\theta_*^\pi\|_2 \le \mathrm{dist}(\Phi\theta_*^\pi, \Phi\Theta_{\mathrm{GTD2}}) + \|\Pi_{\Phi\Theta_{\mathrm{GTD2}}}(\Phi\theta_{\mathrm{RGTD}}) - \Pi_{\Phi\Theta_{\mathrm{GTD2}}}(\Phi\theta_*^\pi)\|_2 + O\big(\tfrac{1}{c}\big)\,,$$

which establishes the singular case.

J ADDITIONAL EXPERIMENTS

This appendix provides additional experimental results that complement the main findings by evaluating the behavior of R-GTD against several baseline algorithms across a broader range of settings, including non-singular random MDPs, singular case, canonical off-policy counterexamples, and stochastic tabular environments. In addition, we include a systematic robustness analysis via a condition number sweep, which examines the performance of the algorithms across a wide spectrum of ill-conditioned regimes.

All experiments, including those presented in the main paper, were conducted on a single workstation equipped with an Intel Core i7-11800H CPU, an NVIDIA GeForce RTX 3060 GPU, and 16GB RAM. Due to the lightweight nature of the environments, each experimental run required only a few minutes.

J.1 NON-SINGULAR MDP

We first examine convergence in the non-singular setting where Assumption 2.2 holds. The randomly generated MDPs with 100 states, 10 actions, and $\gamma = 0.9$ were used. Rewards $r(s, a, s')$ were drawn from a uniform distribution over $[-1, 1]$ and then sparsified by setting elements with $|r(s, a, s')| \leq 0.2$ to zero. The feature matrix Φ was full column rank with 10 features, and regularization coefficients $c \in 0.2, 0.4, 1$. For each MDP, the experiment was repeated 30 times, and the distribution of $\|\theta_k - \theta_*^\pi\|_2$ across runs was summarized using box-and-whisker plots. Figure 4 shows that R-GTD exhibits convergence comparable to GTD2 in this setting. Moreover, larger c accelerates convergence, consistent with theoretical predictions.

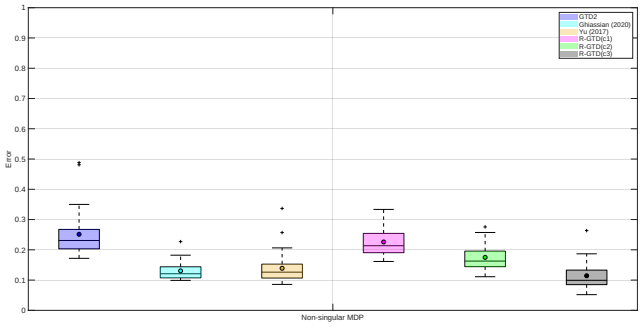


Figure 4: Box-and-whisker representation of $\|\theta_k - \theta_*^\pi\|_2$ over 30 independent runs.

Table 1: Final error of algorithms on non-singular MDP Environment

Algorithm	Final Error ($mean \pm std$)
GTD2	0.2514 ± 0.0768
Ghiassian (2020)	0.1304 ± 0.0299
Yu (2017)	0.1390 ± 0.0529
R-GTD(c1)	0.2259 ± 0.0445
R-GTD(c2)	0.1750 ± 0.0420
R-GTD(c3)	$\mathbf{0.1142 \pm 0.0441}$

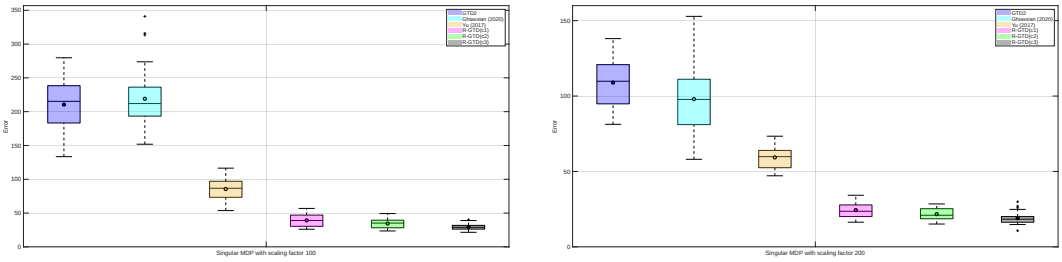
J.2 Robustness to initialization in the singular case

To further investigate the robustness of the proposed method in the singular regime, we evaluate its sensitivity to the magnitude of the initial null-space component. In particular, we consider different scaling factors for the initialization, which directly control the extent of the null-space error.

We conduct experiments with scaling factors of 100, 200, and 1000. The case with scaling factor 1000 is already presented in the main text, and here we provide additional quantitative results in terms of mean and variance to complement the visual comparison.

The experiments are conducted on a randomly generated MDP with 100 states, 10 actions, and discount factor $\gamma = 0.99$. Rewards $r(s, a, s')$ are drawn uniformly from $[-1, 1]$ and sparsified by setting entries with $|r(s, a, s')| \leq 0.2$ to zero. The feature matrix Φ has full column rank with 10 features.

As shown in Table 2, R-GTD consistently achieves stable convergence across all initialization scales, while GTD2 exhibits significant variance. These results demonstrate that the stability of R-GTD is not due to favorable initialization, but rather an inherent property of the regularized formulation, which effectively suppresses the null-space component regardless of its magnitude.



scaling factor 100 scaling factor 200

Figure 5: Box-and-whisker representation of $\|\theta_k - \theta_*^\pi\|_2$ over 30 independent runs.

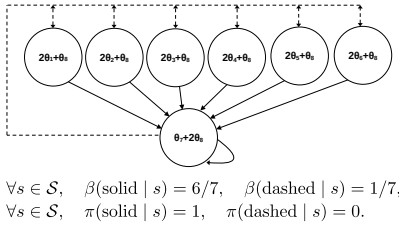
Table 2: Final error of algorithms under different magnitudes of null-space component initialization.

Algorithm	Final Error (mean $\pm$ std)		
	Scale = 100	Scale = 200	Scale = 1000
GTD2	131.3479 ± 12.9018	108.9250 ± 15.8008	210.5517 ± 33.0405
Ghiassian (2020)	126.8904 ± 17.3706	97.9613 ± 20.9245	219.0772 ± 45.9638
Yu (2017)	49.4193 ± 5.9066	59.3214 ± 7.3198	85.5006 ± 15.1741
R-GTD(c1)	19.4368 ± 5.0751	24.3974 ± 4.9124	39.1171 ± 9.3873
R-GTD(c2)	17.5592 ± 4.0874	21.7006 ± 4.0498	34.4717 ± 6.6838
R-GTD(c3)	$\mathbf{15.7314 \pm 3.2652}$	$\mathbf{18.9793 \pm 3.9212}$	$\mathbf{29.1966 \pm 4.1442}$

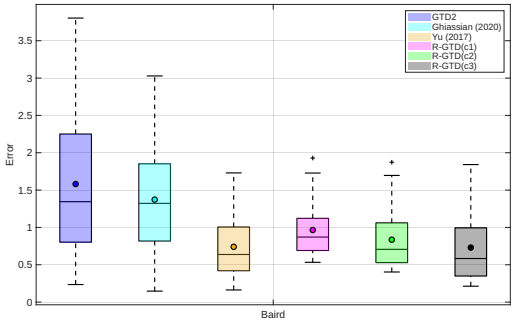
J.3 BAIRD'S COUNTEREXAMPLE

We next evaluate off-policy stability using Baird's counterexample in Baird et al. (1995), whose structure is illustrated in Figure 6 (Left). Quantitative results in Table 3 reveal that R-GTD consistently achieves lower final error than GTD2 and Ghiassian (2020), and performs competitively with Yu (2017).

These results indicate that R-GTD not only ensures stable convergence in off-policy settings but also provides reliable and accurate solutions that are competitive with strong baseline algorithms.



(a) Baird's counterexample diagram



(b) Performance comparison

Figure 6: Box-and-whisker representation of $\|\theta_k - \theta_*^\pi\|_2$ over 30 independent runs.

Table 3: Final error of algorithms on Baird's counterexample.

Algorithm	Final Error ($mean \pm std$)
GTD2	1.5810 ± 0.9377
Ghiassian (2020)	1.3725 ± 0.7256
Yu (2017)	0.7403 ± 0.4114
R-GTD(c1)	0.9652 ± 0.3710
R-GTD(c2)	0.8357 ± 0.4057
R-GTD(c3)	$\mathbf{0.7302 \pm 0.4531}$

J.4 FROZEN LAKE

We next evaluate the performance in a stochastic tabular environment, Frozen Lake (Brockman et al., 2016). The target policy π is pre-trained via on-policy Q-learning (Watkins & Dayan, 1992), and the behavior policy is defined as an ϵ-greedy variant of π with $\epsilon = 0.2$. Experiments were conducted using five different random seeds to ensure robustness.

Quantitative results in Table 4 show that R-GTD consistently achieves lower final error compared to GTD2 and Ghiassian (2020), and performs competitively with Yu (2017). In particular, while Yu (2017) achieves the lowest mean final error, R-GTD with a larger regularization coefficient also attains favorable performance, demonstrating improved accuracy over several baselines across runs. These results indicate that R-GTD provides reliable and accurate value estimation even in stochastic environments with sparse rewards and transition uncertainty.

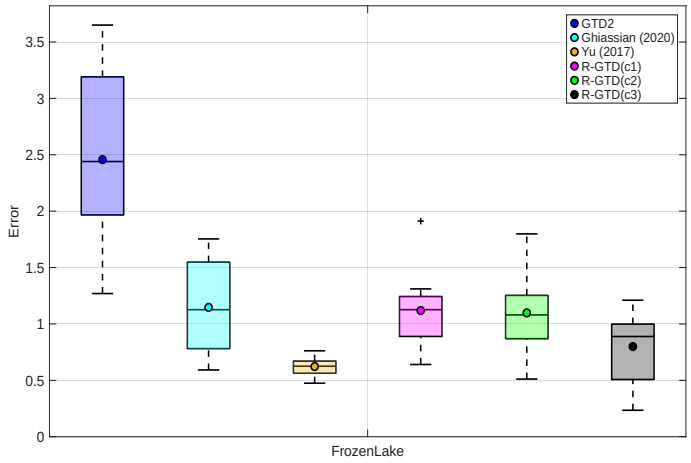


Figure 7: Convergence of $\|\theta_k - \theta_*^\pi\|_2$ on the Frozen Lake environment with step-size $\alpha_k = 1/(k+30)$ and regularization coefficients $c_1 = 0.2$, $c_2 = 0.4$, and $c_3 = 1$.

Table 4: Final error of algorithms on Frozen Lake environment.

Algorithm	Final Error ($mean \pm std$)
GTD2	2.4570 ± 0.7865
Ghiassian (2020)	1.1469 ± 0.4201
Yu (2017)	$\mathbf{0.6224 \pm 0.0857}$
R-GTD(c1)	1.1191 ± 0.3527
R-GTD(c2)	1.0981 ± 0.3739
R-GTD(c3)	0.7997 ± 0.3071

J.5 Robustness across varying condition numbers

We also conducted a comprehensive condition number sweep analysis. This analysis systematically evaluates the robustness of R-GTD and various baselines under a wide spectrum of ill-conditioned regimes.

We systematically varied the condition number, allowing the feature interaction matrix (FIM) to smoothly transition from benign to highly ill-conditioned regimes. To rigorously evaluate the robustness of the algorithms, we adopted an initialization strategy that injects error specifically along the minimum singular vector. Each configuration was evaluated over 5 independent runs, with each run consisting of 100,000 iterations.

The empirical results are summarized in Tables below. The findings demonstrate two main conclusions. First, R-GTD is consistently competitive with all evaluated baselines across the full spectrum of condition numbers. In particular, R-GTD achieves the lowest error and maintains stable variance. Second, the performance advantage of R-GTD remains consistent as the condition number increases. This confirms that the stabilizing benefits of the proposed regularization are highly generalizable.

Table 5: Scale = 100, Final error of algorithms across varying condition numbers ($mean \pm std$)

$\kappa(M)$	GTD2	Ghiassian (2020)	Yu (2017)	R-GTD(c1)	R-GTD(c2)	R-GTD(c3)
1.8e+04	93.38 ± 21.95	70.35 ± 20.67	56.27 ± 10.66	31.00 ± 2.93	29.23 ± 2.34	$\mathbf{27.93 \pm 3.14}$
8.5e+03	84.19 ± 22.01	86.81 ± 16.40	51.89 ± 10.37	32.53 ± 8.32	31.01 ± 7.97	$\mathbf{29.79 \pm 7.53}$
4.0e+03	142.38 ± 21.23	135.23 ± 35.33	53.21 ± 8.71	45.70 ± 10.40	44.18 ± 9.01	$\mathbf{42.71 \pm 6.38}$
1.8e+03	113.38 ± 13.81	108.65 ± 10.68	$\mathbf{39.49 \pm 6.45}$	42.95 ± 4.64	41.26 ± 3.34	39.76 ± 2.64
8.5e+02	148.44 ± 25.93	159.25 ± 73.62	57.13 ± 12.20	40.45 ± 2.90	39.34 ± 3.06	$\mathbf{38.78 \pm 3.64}$
4.0e+02	124.41 ± 24.59	125.65 ± 20.37	49.16 ± 11.40	38.65 ± 1.82	36.69 ± 1.68	$\mathbf{34.96 \pm 2.99}$
1.8e+02	95.75 ± 27.73	83.01 ± 31.06	51.76 ± 13.99	24.17 ± 8.32	22.36 ± 8.17	$\mathbf{20.76 \pm 8.03}$
8.5e+01	120.94 ± 34.18	116.86 ± 25.05	52.42 ± 14.22	27.81 ± 5.42	25.82 ± 3.51	$\mathbf{24.27 \pm 2.44}$
3.9e+01	101.46 ± 15.93	94.23 ± 8.55	44.90 ± 7.08	19.31 ± 5.75	17.85 ± 4.06	$\mathbf{17.45 \pm 2.25}$
1.8e+01	114.31 ± 17.13	73.41 ± 18.11	53.05 ± 8.45	21.31 ± 5.74	18.39 ± 4.70	$\mathbf{15.11 \pm 4.00}$

Table 6: Scale = 200, Final error of algorithms across varying condition numbers ($mean \pm std$)

$\kappa(M)$	GTD2	Ghiassian (2020)	Yu (2017)	R-GTD(c1)	R-GTD(c2)	R-GTD(c3)
1.8e+04	182.08 ± 49.61	145.66 ± 61.35	97.03 ± 23.94	39.45 ± 6.25	35.70 ± 5.53	$\mathbf{34.14 \pm 5.00}$
8.5e+03	205.04 ± 43.69	154.73 ± 50.47	107.58 ± 19.84	45.66 ± 15.33	38.68 ± 12.25	$\mathbf{30.76 \pm 9.84}$
4.0e+03	234.23 ± 24.42	212.84 ± 29.62	95.72 ± 10.38	61.13 ± 8.28	57.22 ± 6.96	$\mathbf{53.84 \pm 5.39}$
1.8e+03	248.99 ± 35.18	248.38 ± 39.32	102.79 ± 16.63	64.51 ± 4.95	59.31 ± 4.41	$\mathbf{54.36 \pm 5.17}$
8.5e+02	229.78 ± 43.61	241.77 ± 62.74	94.45 ± 20.35	62.02 ± 11.65	57.29 ± 10.67	$\mathbf{51.22 \pm 8.44}$
4.0e+02	253.45 ± 52.43	233.49 ± 39.57	109.10 ± 22.21	60.10 ± 4.40	56.71 ± 5.33	$\mathbf{54.96 \pm 6.18}$
1.8e+02	184.45 ± 30.04	182.95 ± 51.00	91.85 ± 12.64	39.27 ± 14.66	32.50 ± 11.71	$\mathbf{26.32 \pm 9.11}$
8.5e+01	213.12 ± 38.16	219.50 ± 52.73	93.35 ± 17.91	46.12 ± 16.51	41.94 ± 14.06	$\mathbf{37.27 \pm 12.44}$
3.9e+01	173.03 ± 17.73	145.80 ± 35.54	75.85 ± 8.57	43.63 ± 6.78	37.61 ± 5.60	$\mathbf{31.01 \pm 5.05}$
1.8e+01	198.77 ± 24.50	142.29 ± 26.72	92.11 ± 11.49	37.58 ± 3.37	32.93 ± 4.70	$\mathbf{28.14 \pm 10.17}$

Table 7: Scale = 1000, Final error of algorithms across varying condition numbers ($mean \pm std$)

$\kappa(M)$	GTD2	Ghiassian (2020)	Yu (2017)	R-GTD(c1)	R-GTD(c2)	R-GTD(c3)
1.8e+04	1181.67 ± 282.87	1074.19 ± 578.73	558.50 ± 129.08	203.37 ± 66.93	178.66 ± 56.74	$\mathbf{154.03 \pm 50.56}$
8.5e+03	1114.69 ± 216.56	933.68 ± 191.36	531.04 ± 101.50	199.60 ± 78.71	171.64 ± 70.38	$\mathbf{149.06 \pm 68.87}$
4.0e+03	1093.49 ± 111.99	959.53 ± 197.04	504.59 ± 64.55	237.01 ± 76.21	215.00 ± 69.00	$\mathbf{199.50 \pm 68.96}$
1.8e+03	1137.22 ± 304.68	1031.44 ± 424.40	507.83 ± 141.80	238.80 ± 87.30	208.44 ± 75.97	$\mathbf{172.91 \pm 61.96}$
8.5e+02	1133.01 ± 266.62	865.24 ± 253.03	521.64 ± 116.49	222.76 ± 69.06	206.70 ± 53.29	$\mathbf{199.30 \pm 56.95}$
4.0e+02	1159.39 ± 226.08	1151.81 ± 284.92	529.01 ± 95.58	263.20 ± 89.91	238.32 ± 79.64	$\mathbf{218.26 \pm 65.28}$
1.8e+02	1229.50 ± 98.08	1117.51 ± 217.79	574.89 ± 39.56	274.72 ± 57.09	235.77 ± 35.80	$\mathbf{187.19 \pm 38.63}$
8.5e+01	1093.71 ± 232.09	987.45 ± 248.50	507.03 ± 114.99	218.03 ± 70.60	199.07 ± 55.70	$\mathbf{181.59 \pm 57.01}$
3.9e+01	1170.70 ± 262.15	1011.09 ± 230.20	544.23 ± 121.52	205.55 ± 50.54	183.31 ± 45.25	$\mathbf{172.34 \pm 42.13}$
1.8e+01	1150.72 ± 232.68	1045.75 ± 300.45	539.49 ± 110.71	227.89 ± 66.93	207.26 ± 63.80	$\mathbf{197.27 \pm 64.15}$

K EXPLICIT COMPARISON WITH PRIOR REGULARIZED METHODS

To clarify the novelty of R-GTD, we compare it with prior regularized GTD methods. The key distinction is not merely whether regularization is used, but *where* it is applied and *what geometry* it preserves in the singular feature-interaction regime.

1. TDRC: Regularization on the auxiliary variable. TDRC adds an L_2 penalty to the secondary variable h. In our notation, this gives

$$h_\beta = (B + \beta I)^{-1}(M\theta + b).$$

The induced fixed-point condition for θ becomes

$$(-M + \beta I)^\top (B + \beta I)^{-1}(M\theta + b) = 0,$$

with system matrix

$$(-M + \beta I)^\top (B + \beta I)^{-1} M.$$

When M is singular, this induced system matrix can remain singular. Thus, TDRC may improve stability through regularization of the auxiliary variable, but it does not directly guarantee uniqueness of the primary parameter θ.

2. Euclidean ridge methods: Regularization in parameter space. Yu (2017) studies regularized GTD objectives of the form

$$J_p(\theta) = J(\theta) + p(\theta),$$

where $p(\theta)$ is a smooth convex regularizer. A common special case is the Euclidean ridge penalty

$$p(\theta) = \frac{\eta}{2}\|\theta\|_2^2.$$

Similarly, Zhang et al. (2020) use ridge regularization in their GQ2 critic. For this Euclidean ridge choice, the induced regularized normal equation has the form

$$\left(M^\top B^{-1} M + \eta I\right)\theta = -M^\top B^{-1} b,$$

This makes the surrogate problem well posed, but the regularization acts on the coordinate vector θ. Hence, the selected solution is tied to the Euclidean parameter norm and can depend on the chosen feature parameterization.

This is different from analyzing the original singular GTD2 system. In particular, these ridge-based formulations establish convergence or well-posedness of a modified objective, but they do not explicitly characterize how the regularized solution relates to the affine solution set of the unregularized GTD2 normal equation when M is singular. For instance, Zhang et al. (2020) invokes a global nonsingularity-type condition, such as $\inf_\theta |\det(A(\theta))| > 0$, in its bias analysis, while Yu (2017) imposes an origin condition for unconstrained convergence. These assumptions contrast with our direct geometric analysis of the strictly singular GTD2 normal equation.

3. R-GTD: Regularization in value-function space. R-GTD is derived by introducing a slack variable directly into the PBE constraint. The resulting unconstrained form is

$$\min_{\theta \in \mathbb{R}^q} \frac{c}{2}\left\|\Pi_{\mathcal{R}(\Phi)}(R^\pi + \gamma P^\pi \Phi\theta - \Phi\theta)\right\|_{D^\beta}^2 + \frac{1}{2}\|\Phi\theta\|_{D^\beta}^2.$$

Since

$$\|\Phi\theta\|_{D^\beta}^2 = \theta^\top B\theta,$$

the normal equation becomes

$$\left(M^\top B^{-1} M + \frac{1}{c}B\right)\theta = -M^\top B^{-1} b.$$

Because $B \succ 0$, the matrix

$$M^\top B^{-1} M + \frac{1}{c}B$$

is strictly positive definite for every finite $c > 0$, even when M is singular. Thus, R-GTD gives a unique solution by regularizing the represented value function $\Phi\theta$, rather than the coordinate vector θ.

4. Solution-selection geometry in the singular regime. The difference between

$$M^\top B^{-1} M + \eta I \qquad \text{and} \qquad M^\top B^{-1} M + c^{-1} B$$

is not merely a different choice of regularization matrix. It changes the geometry of the selected solution.

When M is singular, the unregularized GTD2 normal equation admits the affine solution set

$$\Theta_{\mathrm{GTD2}} = \{\theta : M^\top B^{-1}(M\theta + b) = 0\}.$$

Euclidean ridge methods select a solution according to the Euclidean norm of the parameter vector. This yields a well-posed ridge surrogate, but it does not by itself explain which component of the original GTD2 solution set is identifiable from the projected Bellman equation.

By contrast, R-GTD regularizes the value function in the D^β-weighted feature space. This allows us to explicitly characterize the singular limiting solution. In particular, as $c \to \infty$, our analysis shows that

$$\theta_{\mathrm{RGTD}} = \theta_{\mathrm{GTD2}} - \Pi^B_{\mathcal{N}(G)}(\theta_{\mathrm{GTD2}}) + O(1/c), \qquad \theta_{\mathrm{GTD2}} \in \Theta_{\mathrm{GTD2}}.$$

Thus, R-GTD suppresses the null-space component while preserving the component determined by the projected Bellman equation. This solution-selection characterization is the main theoretical distinction from standard ridge-based regularized GTD methods, and it enables explicit error bounds in strictly singular settings.

L SENSITIVITY ANALYSIS AND PRACTICAL GUIDELINES FOR SELECTING c

As discussed in Section 4, the regularized matrix in R-GTD remains positive definite for any $c > 0$, ensuring a unique solution even when the feature interaction matrix (FIM) is singular. However, the choice of the regularization parameter c introduces a fundamental bias-stability trade-off:

- **Small c:** Strong regularization yields a highly well-conditioned and stable system, but it can introduce excessive bias, potentially degrading the empirical accuracy.
- **Intermediate c:** Provides the optimal trade-off between regularization-induced stability and bias reduction.
- **Large c:** Reduces the bias and allows the solution to approach that of GTD2, but it weakens the regularization effect, making the algorithm more susceptible to instability in ill-conditioned settings.

To provide a practical guideline for selecting c, we conducted a comprehensive sensitivity analysis across four different environments: a Nonsingular MDP, a Singular MDP, Baird's counterexample, and Frozen Lake. As shown in Table 8, intermediate values—specifically in the range of $c \in [1, 4]$—consistently achieve the best balance, yielding the lowest final errors across all tested scenarios. Based on these empirical results, we recommend starting with $c = 1$ or $c = 2$ as a robust default for practical applications, as it effectively stabilizes the learning process while keeping the residual bias minimal.

Table 8: Sensitivity analysis of R-GTD across different values of the regularization parameter c. The table reports the Final Error (mean $\pm$ std) across four diverse environments. Intermediate values ($c \in [1, 4]$) consistently strike the best balance between stability and bias.

Algorithm	Nonsingular MDP	Singular MDP	Baird's Counterexample	Frozen Lake
R-GTD (c=0.2)	0.2254 ± 0.0256	43.8394 ± 10.7775	1.3156 ± 0.5559	1.1024 ± 0.3450
R-GTD (c=0.4)	0.1981 ± 0.0217	40.5704 ± 9.0733	1.2156 ± 0.5742	1.1257 ± 0.3619
R-GTD (c=1)	0.1501 ± 0.0193	$\mathbf{37.1896 \pm 7.9214}$	1.1207 ± 0.6186	$\mathbf{0.8381 \pm 0.2262}$
R-GTD (c=2)	0.1197 ± 0.0237	37.8707 ± 8.9138	$\mathbf{1.0982 \pm 0.6349}$	1.0654 ± 0.4473
R-GTD (c=4)	$\mathbf{0.1073 \pm 0.0288}$	40.6424 ± 10.4395	1.1037 ± 0.6285	0.9186 ± 0.3209
R-GTD (c=8)	0.1082 ± 0.0316	43.3760 ± 11.7065	1.1262 ± 0.6130	1.0455 ± 0.4399
R-GTD (c=16)	0.1125 ± 0.0329	45.2561 ± 12.5310	1.1561 ± 0.5979	1.1277 ± 0.2326

M SINGULAR CASE TOY EXAMPLE

Illustrative example (singular case). To illustrate the singular case, we consider a simple example where FIM is singular. Specifically, we consider a simple Markov chain with three states and a single action. We set the discount factor $\gamma = 0.9$ and choose the state distribution matrix $D = \mathrm{diag}\Big(1/3, 1/3, 1/3\Big)$ where $\mathrm{diag}(\cdot)$ denotes a diagonal matrix with the given elements on its main diagonal. The transition matrix, reward vector, and feature matrix are given by

$$P = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \\ \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \\ \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{bmatrix}, \quad R = \begin{bmatrix} 1 \\ 5 \\ 5 \end{bmatrix}, \quad \Phi = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

A direct computation shows that the matrix FIM is singular. As a consequence, the MSPBE admits an affine solution set Θ_{GTD2} defined in (17). In particular, the solution set can be written as $\Theta_{\mathrm{GTD2}} = \theta_{\mathrm{GTD2}} + \mathrm{Null}(G)$, where θ_{GTD2} is an arbitrary reference solution and $\mathrm{Null}(G)$ denotes the null space of G. Figure 8 visualizes the closed-form trajectory of the θ_{RGTD}, projected onto the value-function space $\Phi\theta \in \mathbb{R}^3$, as the regularization parameter c increases. As c grows, the trajectory of θ_{RGTD} converges to the unique GTD2 solution $(I - \Pi^{B}_{\mathcal{N}(G)})\theta_{\mathrm{GTD2}}$ lying in the B-orthogonal complement of $\mathrm{Null}(G)$, as characterized in Lemma 4.3. This behavior matches the singular-case expansion in (20). (Note: For visual simplicity, this illustrative example is explicitly constructed such that the B-orthogonal projection aligns with the standard Euclidean projection onto $\mathrm{Null}(G)^{\perp}$.)

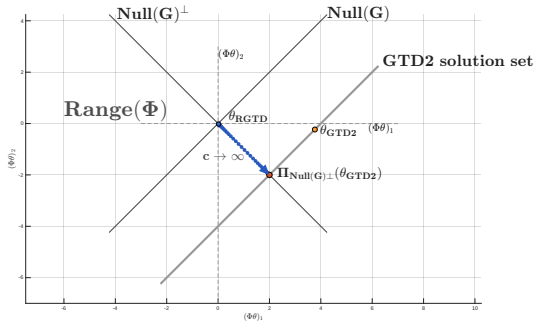


Figure 8: Solution trajectory illustrating the convergence of θ_{RGTD} to θ_{GTD2} as c increases.