Statistical Complexity of Soft Bellman Residual Minimization

Enoch Hyunwook Kang
Foster School of Business
University of Washington
ehwkang@uw.edu

Kyoungseok Jang
Department of AI
Chung Ang University
ksjang@cau.ac.kr

Abstract

Offline reinforcement learning and offline inverse reinforcement learning aim to recover near-optimal value functions or reward models from a fixed batch of logged trajectories, yet current practice still struggles to enforce Bellman consistency scalably. Bellman residual minimization (BRM) has emerged as an attractive remedy. While globally convergent gradient-based soft (i.e., entropy-regularized) BRM methods for neural networks have recently been established, their statistical complexity under stochastic gradient descent remains largely unknown. In this paper, we address this theoretical gap for soft BRM. Through a Lyapunov-based analysis under a uniform PL condition in the auxiliary variable, we establish an $\mathcal{O}(1/n)$ primal argument stability bound up to optimization error and an $\mathcal{O}(1/n)$ statistical complexity for the soft BRM objective under the stated sampling conditions.

1 Introduction

Reinforcement learning (RL) provides a principled framework for sequential decision-making, formalizing dynamic problems as Markov decision processes (MDPs) and learning policies that optimize long-horizon rewards. In many high-stakes domains, however, such as adaptive clinical trials [Zhao et al., 2009], recommendation systems [Afsar et al., 2022], and online education [Bassen et al., 2020], online interaction is unethical, dangerous, or prohibitively expensive [Jiang and Xie, 2025]. This motivates offline RL, where the learner has access only to a fixed batch of data collected by a behavior policy, and its sister field, inverse RL (IRL), which infers reward functions from logged expert trajectories.

Beyond the lack of online interaction, these problems pose both statistical and computational challenges: realistic state spaces are often high-dimensional or continuous, making exact dynamic programming intractable. Following Jiang and Xie [2025], existing offline approaches can be broadly organized into three paradigms: approximate dynamic programming, such as fitted Q-iteration [Ernst et al., 2005]; marginalized importance sampling [Liu et al., 2018]; and Bellman Residual Minimization (BRM) [Baird, 1995]. In this work, we adopt the BRM as our primary framework.

BRM reformulates dynamic programming as fitting a value function to the Bellman equation and minimizes the mean-squared Bellman discrepancy under the offline data distribution. Despite its conceptual appeal and empirical effectiveness, BRM has historically had limited theory for global optimality and convergence. A notable exception is SBEED [Dai et al., 2018], which establishes stable convergence guarantees for the soft, i.e., entropy-regularized, BRM objective, replacing the Bellman max operator with a softmax[1].

Building on this direction, Kang et al. [2025] analyzed the optimization landscape of soft BRM. After a classical bi-conjugate transformation, they showed that, for common Q-function classes such

[1] A more comprehensive discussion of related work is deferred to Appendix D.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Statistical rates for prior works. BRM rates concern squared Bellman error or the corresponding population objective, subject to the qualifications below. The row of Wang et al. [2022] concerns general excess primal population risk. **Oracle** means the analysis assumes the exact empirical (ERM or minimax) solution is available. **SGDA** means it bounds the output of stochastic gradient descent–ascent.

Work	Hypothesis Class	Estimator	Key assumptions	Rate
Maillard et al. [2010]	linear	Oracle	generative model	$O(n^{-1/2})$
Dai et al. [2018]	general	Oracle	β-mixing path	$\tilde{O}(n^{-1/2})$
Chen and Jiang [2019]	finite	Oracle	realizability, completeness	$\tilde{O}(n^{-1})^{\dagger}$
Duan et al. [2021]	general	Oracle	completeness	$O(n^{-1/2})^{\ddagger}$
Hu et al. [2025]	linear	Oracle	completeness, feature cov.	$\tilde{O}(n^{-1})^{\dagger}$
Wang et al. [2022]	general	SGDA	convex–strongly concave	$\tilde{O}(n^{-1/2})$
Ours	general	SGDA	PL–PL, QG	$O(n^{-1})$

as neural networks[2], the mean-squared Bellman error (MSBE) admits a Polyak–Łojasiewicz (PL) structure[3]. Under suitable learning rates and the stated geometry, stochastic gradient descent–ascent (SGDA) enjoys global convergence [Yang et al., 2020]. While this line of work substantially clarifies the optimization behavior of soft BRM, it leaves open the question of how statistical errors arising from finite offline data affect the learned policy.[4] This raises the central question:

How many offline samples do we need for BRM to get a near-optimal value function, under SGDA?

Contributions. In this paper, we close this statistical gap for function classes satisfying the conditions below, including neural-network and linear parameterizations when their geometry is verified. Building on the bi-conjugate reformulation [Antos et al., 2008, Dai et al., 2018] and the PL structure identified by Kang et al. [2025], we develop a Lyapunov potential tailored to PL optimization and combine on-average primal argument stability with a direct statistical analysis. Our main theorem shows that, for n independent observations, SGDA attains an $\mathcal{O}(1/n + 1/T)$ excess primal risk after T iterations, with an MSBE interpretation under auxiliary-function realizability. Stationary dependent data enter through an explicit covariance factor. This complements the convex–concave minimax guarantees with Markov-sampled update indices in Wang et al. [2022]; Table 1 compares assumptions and rates, and Appendix C.5 explains why convex–concave stability arguments do not apply to soft BRM. In particular:

1. We prove an $\mathcal{O}(1/n)$ on-average primal argument-stability bound for SGDA under inner PL, up to optimization error, without requiring independent observations for this stability bound.
2. Using the same PL geometry, we derive the $\mathcal{O}(1/n)$ generalization guarantee for BRM under the stated sampling conditions.
3. Our analysis is constructive, requires no variance reduction or extra regularization, and applies to neural-network parametrizations satisfying the stated PL, QG, and domain conditions.

Technical Sketch. We couple two SGDA runs on neighboring datasets using the same initialization and the same minibatch index sequence. The Lyapunov potential contracts in expectation by a factor $(1 - c\eta_t)$, while stochastic perturbations enter through a term of order η_t^2. Comparing each run with its empirical primal minimizer gives the desired $O(1/n)$ stability term. Evaluating empirical gradients at a fixed population optimum then gives the statistical bound.

[2]Following Liu et al. [2022], this result uses neural networks whose width scales with $\mathrm{radius}^{\mathrm{depth}}$.

[3]The PL condition requires $\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \mu(f(x) - f^*)$ for some $\mu > 0$, and yields convergence guarantees comparable to strong convexity even without convexity. Here we impose PL on the outer minimization and the inner maximization, retaining the primal quadratic-growth condition.

[4]The main statistical guarantee of this paper was later referenced and used in the recent version of Kang et al. [2025]; see Appendix C.5 for details.

[†]Table 1: squared-loss bound behind the headline value-gap / regret guarantee.

[‡]Table 1: $O(n^{-1})$ via local Rademacher complexity for finite or linear classes.

2 Setup and Backgrounds

Markov Decision Process. Throughout, we focus on a single-agent decision-making problem interacting with a discounted Markov Decision Process (MDP) described by the tuple $\big(\mathcal{S}, \mathcal{A}, P, r, \beta, \nu_0\big)$. A state is an element of the measurable space $\mathcal{S}$ and the agent chooses actions from the finite set $\mathcal{A}$. For any state–action pair (s, a) the transition kernel $P(\cdot \mid s, a)$ gives a probability distribution over the next state and the reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ gives the immediate reward. We study the entropy-regularized control problem with regularization coefficient $\eta > 0$, where the Shannon entropy of a distribution $q \in \Delta_{\mathcal{A}}$ is $\mathcal{H}(q) := -\sum_{a \in \mathcal{A}} q(a) \log q(a)$. The scalar $\beta \in (0, 1)$ exponentially discounts rewards that occur further in the future and ν_0 denotes the distribution of the starting state s_0. Throughout, the unsubscripted symbol η denotes the entropy-regularization coefficient; later, the subscripted sequence $(\eta_t)_{t \geq 0}$ continues to denote the SGDA stepsizes.

Policy and value functions. A (stationary Markov) policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ assigns every state s to $\pi(\cdot \mid s)$, a distribution over actions $\mathcal{A}$. Combined with the initial draw $s_0 \sim \nu_0$, a policy induces a probability measure $\mathbb{P}_{\nu_0, \pi}$ on infinite trajectories $(s_0, a_0, s_1, a_1, \dots)$, and the corresponding expectation operator is written $\mathbb{E}_{\nu_0, \pi}$. Under this setup, we consider the entropy-regularized optimal policy and its corresponding value functions defined as

$$\pi^* := \operatorname{argmax}_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \mathbb{E}_{\nu_0, \pi} \left[\sum_{h=0}^{\infty} \beta^h \left(r(s_h, a_h) + \eta \mathcal{H}(\pi(\cdot \mid s_h)) \right) \right]$$

$$V^*(s) := \max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \mathbb{E}_{\nu_0, \pi} \left[\sum_{h=0}^{\infty} \beta^h \left(r(s_h, a_h) + \eta \mathcal{H}(\pi(\cdot \mid s_h)) \right) \mid s_0 = s \right]$$

$$Q^*(s, a) := \max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \mathbb{E}_{\nu_0, \pi} \left[r(s_0, a_0) + \sum_{h=1}^{\infty} \beta^h \left(r(s_h, a_h) + \eta \mathcal{H}(\pi(\cdot \mid s_h)) \right) \mid s_0 = s, a_0 = a \right].$$

One can show that the optimal policy π^* and the value functions satisfy the following entropy-regularized Bellman optimality equations [Kang et al., 2025, Section 3.1 and Appendix B]:

$$V^*(s) = \max_{q \in \Delta_{\mathcal{A}}} \left\{ \mathbb{E}_{a \sim q}\big[Q^*(s, a)\big] + \eta \mathcal{H}(q) \right\} = \eta \ln \left[\sum_{a \in \mathcal{A}} \exp\left(Q^*(s, a)/\eta\right) \right]$$

$$\pi^*(a \mid s) = \frac{\exp\left(Q^*(s, a)/\eta\right)}{\sum_{a' \in \mathcal{A}} \exp\left(Q^*(s, a')/\eta\right)} = \exp\left(\frac{Q^*(s, a) - V^*(s)}{\eta} \right) \text{ for } a \in \mathcal{A}$$

$$Q^*(s, a) = r(s, a) + \beta \cdot \mathbb{E}_{s' \sim P(\cdot \mid s, a)} \left[V^*(s') \mid s, a \right]$$

2.1 Bellman Residual Minimization

Bellman Error (Bellman Residual) and Temporal Difference Error. Let $\mathcal{Q}$ be the set of bounded functions on $\mathcal{S} \times \mathcal{A}$, which contains Q^* when rewards are bounded. We introduce the *(soft) Bellman optimality operator*, $\mathcal{T} : \mathcal{Q} \to \mathcal{Q}$, defined for any $Q \in \mathcal{Q}$ by:

$$(\mathcal{T}Q)(s, a) := r(s, a) + \beta \cdot \mathbb{E}_{s' \sim P(\cdot \mid s, a)} \left[V_Q(s') \mid s, a \right] \text{ where } V_Q(s) := \eta \ln \left[\sum_{a \in \mathcal{A}} \exp\left(Q(s, a)/\eta\right) \right].$$

Note that Q^* is the unique fixed point of $\mathcal{T}$ [Rust, 1994]; Q^* is the unique solution to the *Bellman optimality equation*: $\mathcal{T}Q^* = Q^*$. This uniqueness motivates the definition of the *Bellman error* (Bellman residual) of Q at (s, a), $(\mathcal{T}Q)(s, a) - Q(s, a)$. Computing $\mathcal{T}Q$ requires knowledge of the transition kernel P, which is typically unavailable in reinforcement learning. Instead, given a sampled transition (s, a, s'), we work with the *sampled Bellman operator*, $\hat{\mathcal{T}}$, and its corresponding *temporal-difference (TD) error*, δ_Q. Given a single transition tuple (s, a, s'), they are defined as:

$$\hat{\mathcal{T}}Q(s, a, s') := r(s, a) + \beta V_Q(s'), \qquad \delta_Q(s, a, s') := \hat{\mathcal{T}}Q(s, a, s') - Q(s, a)$$

Note that δ_Q is an unbiased, single-sample estimate of the Bellman error: $\mathbb{E}_{s' \sim P(\cdot \mid s, a)}[\delta_Q(s, a, s')] = (\mathcal{T}Q)(s, a) - Q(s, a)$.

Bellman Residual Minimization. Note that both Bellman error (Bellman residual) and its proxy, the TD error, are functions of (s, a). To find Q that minimizes the Bellman error for all (s, a), we can

instead find Q that minimizes expected square error on the offline data distribution. That is, we define the *Squared Bellman Error* at (s, a) as $\mathcal{L}_{\text{BE}}(Q)(s, a) := ((\mathcal{T}Q)(s, a) - Q(s, a))^2$ and minimize the *Mean Squared Bellman Error* (MSBE), defined as:

$$\overline{\mathcal{L}_{\text{BE}}}(Q) := \mathbb{E}_{(s,a)\sim\pi_D,\nu_0}\left[\mathcal{L}_{\text{BE}}(Q)(s, a)\right]$$

where π_D is the policy used for collecting data. The mean squared TD error (MSTDE), built from the squared TD error $\mathcal{L}_{\text{TD}}(Q)(s, a, s') := \delta_Q(s, a, s')^2$, is a biased proxy for MSBE because expectation and square do not commute. This is the *double sampling problem* [Antos et al., 2008]. The bias equals $\beta^2\,\mathbb{E}_{s'}\left[(V_Q(s') - \mathbb{E}_{s'}V_Q(s'))^2\right]$ [Kang et al., 2025, Appendix C.1] and cannot be removed without knowing P.

To resolve this issue, we employ the "Bi-Conjugate Trick" [Antos et al., 2008, Dai et al., 2018, Patterson et al., 2022], $\mathcal{L}_{\text{BE}}(Q)(s, a) = \max_{h\in\mathbb{R}}\{2h\,\mathbb{E}_{s'}[\delta_Q(s, a, s')] - h^2\}$, which after substituting $\zeta := (h - r(s, a) + Q(s, a))/\beta$ [Kang et al., 2025, Appendix C.1] reads

$$\mathcal{L}_{BE}(Q)(s, a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\mathcal{L}_{TD}(Q)\,(s, a, s')\right] - \beta^2 \min_{\zeta\in\mathbb{R}} \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\left(V_Q\,(s') - \zeta\right)^2\right] \tag{1}$$

Now define the per-sample objective

$$f(Q, \zeta; s, a, s') := \mathcal{L}_{\text{TD}}(Q)\,(s, a, s') - \beta^2\left(V_Q\,(s') - \zeta\right)^2. \tag{2}$$

If we slightly abuse notation such that $\mathbb{E}_{s,a,s'}[*] = \mathbb{E}_{s,a\sim\nu_0,\pi_D}\left[\mathbb{E}_{s'\sim P(\cdot|s,a)}[*]\right]$, then minimizing MSBE can be written as a min–max problem over (Q, ζ):

$$\min_{Q\in\mathcal{Q}} \mathbb{E}_{(s,a,s')}\left[f(Q, \widetilde{\zeta}(Q); s, a, s')\right] \quad \text{where } \widetilde{\zeta}(Q) \in \operatorname*{argmin}_{\zeta\in\mathbb{R}^{S\times A}} \mathbb{E}_{(s,a,s')}\left[\left(V_Q\,(s') - \zeta(s, a)\right)^2\right].$$

By parametrizing ζ as ζ_v and Q as Q_w using function classes such as neural networks and relaxing the definition of the per-sample objective

$$f(w, v; s, a, s') := f(Q_w, \zeta_v; s, a, s') = \mathcal{L}_{\text{TD}}(Q_w)\,(s, a, s') - \beta^2\left(V_{Q_w}\,(s') - \zeta_v\right)^2, \tag{3}$$

the resulting parametrized problem becomes

$$\min_{w\in\mathcal{W}} \mathbb{E}_{(s,a,s')}\left[f(w, \tilde{v}; s, a, s')\right] \quad \text{where } \tilde{v}(w) \in \operatorname*{argmin}_{v\in\mathcal{V}} \mathbb{E}_{(s,a,s')}\left[\left(V_{Q_w}\,(s') - \zeta_v(s, a)\right)^2\right]. \tag{4}$$

The population objective in (4) equals MSBE when the auxiliary class realizes $m_w(s, a) := \mathbb{E}[V_{Q_w}(S') \mid s, a]$. In general it equals MSBE minus $\beta^2 \inf_v \mathbb{E}[(m_w(S, A) - \zeta_v(S, A))^2]$; a uniform upper bound ϵ_{aux} on this approximation error adds $\beta^2\epsilon_{\text{aux}}$ to the excess-MSBE guarantee.

Let's consider Equation (4) as the expected risk minimization problem. Define $\mathcal{D}_{\pi_D,\nu_0} = \{z_i\}_{i=1}^N$ as the offline data collected from following π_D starting from ν_0, where each $z_i = (s_i, a_i, s_i')$ denotes a sample consisting of the current state, the action taken, and the resulting next state. Then, the corresponding empirical risk minimization problem can be written as

$$\min_{w\in\mathcal{W}} \frac{1}{N} \sum_{z\in\mathcal{D}_{\pi_D,\nu_0}} f(w, \tilde{v}; z), \text{ where } \tilde{v}(w) \in \operatorname*{argmin}_{v\in\mathcal{V}} \frac{1}{N} \sum_{z\in\mathcal{D}_{\pi_D,\nu_0}} \left[\left(V_{Q_w}\,(s') - \zeta_v(s, a)\right)^2\right]. \tag{5}$$

A canonical way of solving the mini-max problem is to apply the Stochastic Gradient Descent–Ascent (SGDA) algorithm [Yang et al., 2020]. Kang et al. [2025] proved that both (4) and (5) satisfy the Polyak-Łojasiewicz condition within a large enough ball around the initialization point for a Neural Network parametrization of Q with a sufficient network width (width scaling with radius^{depth}). The SGDA guarantee used here is stated below under explicit inner PL, primal QG, and domain conditions. (For the detailed discussion on neural networks, see Appendix C.2.)

2.2 Stochastic Gradient Descent–Ascent Algorithm (SGDA)

As discussed earlier, SGDA solves (5) by a descent step in w and an ascent step in v using minibatch gradients. Let $\mathcal{D} = \{z_i\}_{i=1}^n$ denote the dataset, where each $z_i = (s_i, a_i, s_i')$ denotes a sample consisting of the current state, the action taken, and the resulting next state. Fix a minibatch size

$B \in \{1, \dots, n\}$. At round t we draw an index batch I_t of size B (a multiset for sampling with replacement and a subset of $[n]$ otherwise), either *with* or *without* replacement (our theory does not depend on this choice). With f standing for the relaxed per–sample saddle objective introduced in (3), the averaged stochastic gradients are $g_t^w = \nabla_w F_{I_t}(w_t, v_t), \quad g_t^v = \nabla_v F_{I_t}(w_t, v_t),$ where $\quad F_I(w, v) := \frac{1}{|I|} \sum_{j \in I} f(w, v; z_j)$ for any minibatch I. By a slight abuse of notation, when the subscript corresponds to the full dataset $\mathcal{D}$, we write $F_{\mathcal{D}}(w, v) := \frac{1}{n} \sum_{z \in \mathcal{D}} f(w, v; z)$.

Conditionally on the dataset and the past iterates, the minibatch is drawn uniformly and independently of the past. Thus $\mathbb{E}[g_t^w \mid \mathcal{F}_t] = \nabla_w F_{\mathcal{D}}(w_t, v_t)$ and likewise for v. Write σ_B^2 for a common conditional variance bound for the two gradient blocks; $\sigma_B^2 \le G^2/B$, and $\sigma_B^2 = 0$ for an exact full-batch update. Set $\kappa_v := 1 + (L/\rho)^2$ and use primal and dual stepsizes η_t and $\kappa_v \eta_t$, respectively. Here L, ρ and G are the smoothness, inner-PL and gradient-bound constants of Conditions (C1), (C4) and (C5) in Section 3. Algorithm 1 is stated for a generic per-sample saddle objective f: with f from (3) it solves the BRM problem (5), and Section 3 analyzes it for any f satisfying Conditions (C1)–(C6).

Algorithm 1 Minibatch SGDA on an empirical saddle objective $F_{\mathcal{D}}$

Input: Dataset $\mathcal{D} = \{z_i\}_{i=1}^n$, per-sample objective f, minibatch size B, stepsizes (η_t), dual ratio κ_v, initial (w_0, v_0)
for $t = 0$ **to** $T - 1$ **do**
  Draw a fresh uniform index batch I_t of size B
  $g_t^w \leftarrow \frac{1}{B} \sum_{i \in I_t} \nabla_w f(w_t, v_t; z_i)$
  $g_t^v \leftarrow \frac{1}{B} \sum_{i \in I_t} \nabla_v f(w_t, v_t; z_i)$
  $w_{t+1} \leftarrow w_t - \eta_t g_t^w$ ▷ gradient *descent*
  $v_{t+1} \leftarrow v_t + \kappa_v \eta_t g_t^v$ ▷ gradient *ascent*
end for
Output: (w_T, v_T)

With harmonic stepsizes, we can state the global convergence guarantee of Algorithm 1 in the parameter space:

Lemma 1 (Global convergence of minibatch SGDA in parameter space)**.** *Suppose Conditions (C1)–(C6) hold on the effective domain Ω of Condition (C5) (Section 3), and use the conditionally unbiased minibatch updates above. Choose $\eta_t = c_1/(c_2 + t)$ with $cc_1 > 1$ and $c_2 \ge \max\{1, c_1/\bar{\eta}\}$, where c and $\bar{\eta}$ depend only on $L, \rho, \mu_{\mathrm{PL}}$ and κ_v*[5]*. Then, with $\Phi_D(w) := \max_v F_D(w, v)$ and $\Phi_D^\star := \min_w \Phi_D(w)$, the iterates converge almost surely to an empirical minimax pair $(x_D^\star, v_D^\star)$, meaning $x_D^\star \in \arg\min_w \Phi_D(w)$ and $v_D^\star \in \arg\max_v F_D(x_D^\star, v)$. Moreover, with $\Psi_{1,D} = \Phi_D - \Phi_D^\star + \Phi_D - F_D$ and $d_2 := c_2$,*

$$\mathbb{E}[\Phi_D(w_t) - \Phi_D^\star] \le \mathbb{E}\Psi_{1,D}(w_t, v_t) \le \frac{d_1}{c_2 + t}.$$

The proof is given after Lemma 12. Lemma 1 settles optimization; the rest of the paper bounds how the finite-sample error propagates to the population BRM objective.

3 Stability and Generalization for Bellman Residual Minimization

We quantify generalization through *algorithmic stability* for minimax learning. Algorithmic stability formalizes how sensitive a learning algorithm is to small changes in the training set: if replacing a single training example only slightly perturbs the algorithm's output, then the algorithm is said to be stable. Under the corresponding data-sampling assumptions, stability implies generalization: algorithms that are stable on neighboring datasets exhibit small discrepancies between their empirical risk and population risk. As shown in Wang et al. [2022], this principle carries over to minimax optimization. Here generalization refers to the gap between the empirical and population BRM objectives, as in statistical learning theory [Bousquet and Elisseeff, 2002, Hardt et al., 2016]; policy-level guarantees follow under coverage assumptions (Appendix C.6).

[5] $L_\Phi = L(1 + L/\rho)$, $H = L_\Phi + L/2$, $c = \min\{\mu_{\mathrm{PL}}/2, \rho/2\}$, $\bar{\eta} = \min\{\frac{1}{4H}, \frac{1}{4H(L/\rho)^2 + L\kappa_v^2}, \frac{1}{c}\}$, $C'_{\mathrm{var}} = (H + L\kappa_v^2/2)\sigma_B^2$, and $d_1 := \max\{c_2 \mathbb{E}\Psi_{1,D}(w_0, v_0), C'_{\mathrm{var}} c_1^2/(cc_1 - 1)\}$. For comparisons and population bounds, take d_1 uniformly over the datasets under consideration.

Notations. For the stability analysis, it is convenient to momentarily step back from the specific Bellman–residual objective and view our problem as a generic empirical minimax problem with primal and dual PL conditions. In the BRM formulation, the primal variable w parametrizes the action–value function Q_w, the dual variable v parametrizes the auxiliary function ζ_v, and the empirical objective in (5) is an average over samples $z = (s, a, s')$ drawn from the offline dataset. In this section we abstract this structure and write $f(w, v; z)$ for the per–sample saddle loss and $F_D(w, v)$ for its empirical average over a dataset $D = \{z_i\}_{i=1}^n$. Write $F(w, v) := \mathbb{E}[f(w, v; Z)]$ for the population objective under the target marginal distribution. In Theorem 4 we then specialize the resulting stability and generalization bounds back to BRM. Let $\mathcal{D} = (z_1, \dots, z_n)$ be a dataset with $z_i \in \mathcal{Z}$. For any $i \in [n]$ and any $\tilde{z}_i \in \mathcal{Z}$, define the *replace-one neighbor* of $\mathcal{D}$ by $\mathcal{D}^{(i)} := (z_1, \dots, z_{i-1}, \tilde{z}_i, z_{i+1}, \dots, z_n)$. We call $\mathcal{D}$ and $\mathcal{D}^{(i)}$ *neighboring datasets*. When comparing SGDA runs on $\mathcal{D}$ and $\mathcal{D}^{(i)}$, we couple them using the same minibatch index sequence $(I_t)_{t\geq 0}$ (shared-index coupling). Expectations averaged over i are taken with $i \sim \mathrm{Unif}([n])$. The conditional stability bound does not require an i.i.d. assumption on the data; population generalization is treated separately below. Appendix C.3 explains how the coupling works.

Throughout, we present the coupled updates for the *single-sample* variant of SGDA. The extension to a minibatch of size $B \geq 1$ replaces each sample gradient by its average. The conditional variance bound satisfies $\sigma_B^2 \leq G^2/B$, and it is zero for exact full-batch updates. The constants below retain this dependence.

3.1 Stability

We consider an offline setting where the data may be dependently sampled (e.g., from a single trajectory in a Markov Decision Process), violating the standard i.i.d. assumption. We use *on-average algorithmic stability* to compare outputs on neighboring datasets; dependence enters the population bound through the covariance term below.

Definition 1 (On-average algorithmic stability)**.** *Let $\mathcal{A}$ be a randomized learning algorithm that maps a dataset $\mathcal{D} = (z_1, \dots, z_n)$ to a parameter output $\mathcal{A}(\mathcal{D})$. For each $i \in [n]$, let $\mathcal{D}^{(i)}$ denote the replace-one neighbor of $\mathcal{D}$, where z_i is replaced by an independent copy $\tilde{z}_i$. Then the* on-average argument stability *of $\mathcal{A}$ after T iterations is $\varepsilon_T := \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\,\|\mathcal{A}(\mathcal{D}) - \mathcal{A}(\mathcal{D}^{(i)})\|\,\big]$, where the expectation includes the algorithm's internal randomness and the replacement; population statements also average over the dataset.*

Intuitively, ε_T measures how much the algorithm's output changes when a single training point is replaced on average. We analyze the on-average argument stability of SGDA (i.e., $\mathcal{A}$ =SGDA) for smooth minimax problems satisfying the stated PL conditions, following the stability framework for minimax optimization.

We run Algorithm 1 on $\mathcal{D}$ and on $\mathcal{D}^{(i)}$ (where $z_j' = z_j$ for $j \neq i$ and $z_i' = \tilde{z}_i$) from the same initialization and with the same fresh uniform minibatches, so each stochastic gradient is conditionally unbiased for its own empirical objective given $\mathcal{F}_t := \sigma(\mathcal{D}, \mathcal{D}^{(i)}, w_0, v_0, I_0, \dots, I_{t-1})$. For $D \in \{\mathcal{D}, \mathcal{D}^{(i)}\}$ write $F_D(w, v) := \frac{1}{n}\sum_{j=1}^n f(w, v; z_j^D)$, $\Phi_D(w) := \max_v F_D(w, v)$ and $\Phi_D^\star := \min_w \Phi_D(w)$.

Throughout, we impose the following conditions (C1)–(C6) (satisfied by the BRM objective [Kang et al., 2025] under Section 2 when the neural network parametrization's width is enough[6]):

- (C1) **Smoothness.** F_D is L-smooth in the joint variable (w, v).
- (C2) **PL for Φ_D.** Φ_D satisfies the Polyak–Łojasiewicz (PL) inequality with parameter $\mu_{\mathrm{PL}} > 0$: $\frac{1}{2}\|\nabla\Phi_D(w)\|^2 \geq \mu_{\mathrm{PL}}(\Phi_D(w) - \Phi_D^\star)$.
- (C3) **QG for Φ_D.** Φ_D satisfies Quadratic Growth (QG) with parameter $\mu_{\mathrm{QG}} > 0$: $\Phi_D(w) - \Phi_D^\star \geq \frac{\mu_{\mathrm{QG}}}{2}\|w - x_D^\star\|^2$.
- (C4) **Inner PL: PL in v.** F_D satisfies $\frac{1}{2}\|\nabla_v F_D(w, v)\|^2 \geq \rho\big(\Phi_D(w) - F_D(w, v)\big)$ uniformly in (w, v).
- (C5) **Bounded gradients on the effective domain.** There exists a compact convex set $\Omega \subset \mathcal{W} \times \mathcal{V}$ containing the iterates almost surely, the fixed population comparison pair, and the empirical

[6]For details, See Appendix C.2.

maximizers and comparison segments used in the proofs. We define $G < \infty$ as the uniform gradient bound on this set:

$$G := \sup_{(w,v)\in\Omega, z\in\mathcal{Z}} \|\nabla_w f(w,v;z)\| \vee \|\nabla_v f(w,v;z)\|.$$

(C6) **Attainment.** The inner maxima and empirical primal minima are attained, and the population primal risk has an attained minimizer. We write $x_D^\star \in \arg\min_w \Phi_D(w)$ and $V_D(w) := \arg\max_v F_D(w,v)$.

Remark 3.1 (Clarifications on the conditions)**.** *The parameters are unconstrained, so Ω in (C5) is a region that contains the iterates rather than a constraint set; because PL and smoothness hold only locally, the auxiliary ascent paths and segments used in the proofs are also taken in Ω. Appendix C.1 shows that such a set Ω exists under two-sided PL.*

In addition, we impose the following settings for the stability analysis. Setting (S2) concerns the algorithm's random indices, whereas the observations in the offline dataset may be dependent. Uniformity is discussed further in Appendix C.4.

(S1) **Stepsizes.** $0 < \eta_t \le \bar{\eta}$, with $\bar{\eta}$ as in Lemma 1; the dual stepsize is $\kappa_v \eta_t$.

(S2) **Shared-index coupling.** The two runs use the same fresh uniform minibatch at each iteration, independently of the past. No independence of the observations within the offline dataset is needed for the conditional optimization and stability bounds.

(S3) **Uniformity of constants across datasets.** The constants $L, \rho, \mu_{\mathrm{PL}}, \mu_{\mathrm{QG}}, G$ are the same for D and $D^{(i)}$. Population bounds use these constants and a common initial-potential bound uniformly over the sampled datasets; the bounds remain controlled as n varies.

To establish stability in our PL setting, a key difficulty is that SGDA is not optimizing the value function $\Phi_D(w) := \max_v F_D(w,v)$ directly, but rather the saddle objective $F_D(w,v)$. The gradient used for the primal update, $\nabla_w F_D(w_t, v_t)$, coincides with $\nabla\Phi_D(w_t)$ whenever the dual variable v_t is already in its inner maximizer set $V_D(w_t)$; away from this set there may be a "mismatch" term $\Delta_t := \nabla_w F_D(w_t,v_t) - \nabla\Phi_D(w_t)$. As a result, tracking the PL suboptimality $\Phi_D(w_t) - \Phi_D^\star$ alone does not yield a contracting recursion under SGDA, while tracking only the dual gap $\Phi_D(w_t) - F_D(w_t,v_t)$ ignores how far the primal iterate is from the BRM minimizer.

To address this, we define the *Lyapunov potential* :

$$\Psi_{\alpha,D}(w,v) := \underbrace{\Phi_D(w) - \Phi_D^\star}_{A(w)} + \alpha \cdot \underbrace{\big(\Phi_D(w) - F_D(w,v)\big)}_{\Gamma(w,v)}.$$

The choice $\alpha = 1$ cancels the mixed term in the one-step expansion, and the dual stepsize ratio $\kappa_v = 1 + (L/\rho)^2$ of Algorithm 1 lets dual ascent absorb the remaining squared gradient mismatch. The two PL inequalities then give contraction up to the additive stochastic-gradient error.

The next theorem gives an $O(1/n)$ bound on the on-average primal argument stability of SGDA under the harmonic stepsizes of Lemma 1; Proposition B.1 in Appendix B.2 gives the bound for general stepsizes.

Theorem 2 (On-average argument stability of SGDA)**.** *Suppose that the loss function f and datasets $\mathcal{D}, \mathcal{D}^{(i)}$ satisfy (C1)–(C6) under the setting (S1)–(S3), with $\alpha = 1$, the dual stepsize multiplier κ_v above and the harmonic stepsizes of Lemma 1. Let $\varepsilon_T^w := \frac{1}{n}\sum_{i=1}^n \mathbb{E}\|w_T(\mathcal{D}) - w_T(\mathcal{D}^{(i)})\|$. Then, for d_1, d_2 from Lemma 1 taken uniformly over the compared datasets,*

$$\varepsilon_T^w \le 2C_{\mathrm{dist}}\sqrt{\frac{d_1}{d_2+T}} + \frac{C'_{\mathrm{hit}}}{n} = O\big((d_2+T)^{-1/2}\big) + O(1/n),$$

where C_{dist} and C'_{hit} depend only on $L, \rho, \mu_{\mathrm{PL}}, \mu_{\mathrm{QG}}, G$.[7] *If, in addition, $V_D(x_D^\star) = \{v_D^\star\}$ for every compared dataset, the same bound holds for $\varepsilon_T^{\mathrm{joint}} := n^{-1}\sum_i \mathbb{E}[\|w_T - w'_T\| + \|v_T - v'_T\|]$; this condition is not needed for the primal stability or risk guarantees.*

[7] $C_{\mathrm{dist}} = \sqrt{2(1+L/\rho)^2/\mu_{\mathrm{QG}} + 2/(\alpha\rho)}$ and $C'_{\mathrm{hit}} = 2G\big((1+L/\rho)^2/\sqrt{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}} + 1/\rho\big)$.

Proof sketch. Insert the two empirical primal minimizers between the coupled outputs:

$$\|w_T - w'_T\| \le \underbrace{\|w_T - x^\star_{\mathcal{D}}\| + \|w'_T - x^\star_{\mathcal{D}^{(i)}}\|}_{\text{optimization}} + \underbrace{\|x^\star_{\mathcal{D}} - x^\star_{\mathcal{D}^{(i)}}\|}_{\text{sensitivity}}.$$

Three steps control these terms; the corresponding lemmas and proofs are in Appendix B.1.

Step 1 (contraction of the potential). Write $g_t := \nabla\Phi_D(w_t)$, $p_t := \nabla_w F_D(w_t, v_t) = g_t + \Delta_t$ and $q_t := \nabla_v F_D(w_t, v_t)$. In the one-step expansion of $\Psi_{1,D}$, the first-order term is $-2\langle g_t, p_t\rangle + \|p_t\|^2 - \kappa_v\|q_t\|^2$, and the choice $\alpha = 1$ cancels the mixed term: $-2\langle g_t, p_t\rangle + \|p_t\|^2 = -\|g_t\|^2 + \|\Delta_t\|^2$. Inner PL bounds the mismatch by $\|\Delta_t\| \le (L/\rho)\|q_t\|$, so the dual ratio $\kappa_v = 1 + (L/\rho)^2$ reduces the first-order term to $-\|g_t\|^2 - \|q_t\|^2$. The two PL inequalities then give (Lemma 12)

$$\mathbb{E}[\Psi_{1,D}(w_{t+1}, v_{t+1}) \mid \mathcal{F}_t] \le (1 - c\eta_t)\Psi_{1,D}(w_t, v_t) + C'_{\mathrm{var}}\eta_t^2, \qquad c = \min\{\mu_{\mathrm{PL}}/2, \rho/2\},$$

and the harmonic stepsizes turn this recursion into $\mathbb{E}\Psi_{1,D}(w_T, v_T) \le d_1/(d_2 + T)$ (Lemma 1).

Step 2 (from potential to distance). Since $\Psi_{1,D} \ge \Phi_D - \Phi^\star_D$, QG gives (Lemma 10)

$$\|w - x^\star_D\| \le \sqrt{2\Psi_{1,D}(w, v)/\mu_{\mathrm{QG}}},$$

so by Jensen's inequality each optimization term is at most $\sqrt{2d_1/(\mu_{\mathrm{QG}}(d_2 + T))}$ in expectation.

Step 3 (minimizer sensitivity). Replacing one sample moves each empirical gradient average by at most $2G/n$; with the inner PL error bound this gives $\|\nabla\Phi_{\mathcal{D}}(w) - \nabla\Phi_{\mathcal{D}^{(i)}}(w)\| \le 2G(1 + L/\rho)/n$ (Lemma 9). At $u = x^\star_{\mathcal{D}^{(i)}}$, where $\nabla\Phi_{\mathcal{D}^{(i)}}(u) = 0$, PL and QG give (Lemma 11)

$$\frac{\mu_{\mathrm{QG}}}{2}\|u - x^\star_{\mathcal{D}}\|^2 \le \Phi_{\mathcal{D}}(u) - \Phi^\star_{\mathcal{D}} \le \frac{\|\nabla\Phi_{\mathcal{D}}(u)\|^2}{2\mu_{\mathrm{PL}}},$$

hence $\|x^\star_{\mathcal{D}} - x^\star_{\mathcal{D}^{(i)}}\| \le S_w/n$ with $S_w := 2G(1 + L/\rho)/\sqrt{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}}$. Averaging over i gives the theorem, since $C_{\mathrm{dist}} \ge \sqrt{2/\mu_{\mathrm{QG}}}$ and $C'_{\mathrm{hit}} \ge S_w$. The full proof is in Appendix B.2.

Remark 3.2 (Where the conditions enter)**.** *Inner PL (C4) gives the error bound for the auxiliary problem and the mismatch bound of Step 1, which the dual ratio κ_v absorbs. Primal PL (C2) and inner PL give the contraction in Step 1, QG (C3) gives Step 2, and primal PL and QG together give the minimizer sensitivity in Step 3. PL therefore plays two roles, linear drift in the optimization recursion and control of minimizer sensitivity, and smoothness alone provides neither: the smooth nonconvex SGD stability bound of Hardt et al. [2016] grows with the number of iterations.*

3.2 Generalization

In this section, we quantify generalization using the optimization and stability bounds derived in the previous section. We bound the population BRM risk of the SGDA output through the primal risk below, which is the Bellman residual under auxiliary-function realizability.

Definition 2 (Primal Risk)**.** *Define the* value functions

$$R(w) := \max_{v\in\mathcal{V}} F(w, v), \qquad R_n(w) := \max_{v\in\mathcal{V}} F_{\mathcal{D}}(w, v) = \Phi_{\mathcal{D}}(w).$$

From stability to generalization. The generalization bound reuses the mechanism of Theorem 2, with the population objective in place of the neighboring dataset. In Theorem 2, replacing one sample perturbs $\nabla\Phi$ by $O(1/n)$, and Step 3 of its proof turns this into an $O(1/n)$ move of the empirical minimizer. The same step at the population minimizer $w^\star \in \arg\min_w R(w)$ gives $\|x^\star_{\mathcal{D}} - w^\star\|^2 \le \|\nabla\Phi_{\mathcal{D}}(w^\star)\|^2/(\mu_{\mathrm{PL}}\mu_{\mathrm{QG}})$, which is $O(1/n)$ in expectation because $\nabla\Phi_{\mathcal{D}}(w^\star)$ is controlled by averages of n mean-zero gradients. The optimization part is also shared: Steps 1 and 2 bound $\|w_T - x^\star_{\mathcal{D}}\|^2$. Since $\nabla R(w^\star) = 0$, the excess risk is quadratic in $\|w_T - w^\star\|$, which gives the $O(1/n + 1/T)$ rate and allows multiple dual maximizers. The resulting generalization bound is as follows.

Lemma 3 (Generalization under inner PL)**.** *Let Conditions (C1)–(C6) hold uniformly for the empirical objectives, and let the population objective F also satisfy (C1), (C4) and inner attainment, so that R is smooth and $\nabla R(w^\star) = 0$. For the population guarantees, assume that the observations are*

independently drawn from the population distribution[8]*. Choose a fixed interior population minimizer* $w^\star$ *of* R *and an interior maximizer* $v^\star \in \arg\max_v F(w^\star, v)$*. The common comparison domain must contain this pair, the required empirical maximizers, and the segments used below. Put*

$$a_T := \mathbb{E}[\Phi_{\mathcal{D}}(w_T) - \Phi^\star_{\mathcal{D}}], \qquad r_n := \mathbb{E}\|\nabla\Phi_{\mathcal{D}}(w^\star)\|^2, \qquad L_\Phi := L(1 + L/\rho).$$

Then $r_n \le (1 + L/\rho)^2 G^2/n$ *and*

$$\mathbb{E}[R(w_T) - R_n(w_T)] \le \frac{2L_\Phi}{\mu_{\mathrm{QG}}} a_T + \left(\frac{L_\Phi}{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}} + \frac{1}{2\mu_{\mathrm{PL}}}\right) r_n.$$

Combining Lemma 1 and Lemma 3, we arrive at Theorem 4, the *generalization guarantee of Bellman residuals*. In words, the learned Q (i.e., the corresponding learned w) generalizes: its expected population risk is bounded by its expected empirical risk plus the optimization error and an $\mathcal{O}(1/n)$ statistical term.

Theorem 4 (Generalization for the empirical Bellman–residual objective)**.** *Let* $(\widehat{w}^{(T)}, \widehat{v}^{(T)})$ *be the parameters returned by Algorithm 1 after* T *iterations on* (5)*, with the harmonic stepsizes of Lemma 1 and under the conditions of Lemma 3. Let* $\mathcal{R}(w) := \max_v \mathbb{E}_{(s,a,s')} f(w, v; s, a, s')$ *and* $\widehat{\mathcal{R}}_n(w) := \max_v \frac{1}{n}\sum_{z\in\mathcal{D}_{\pi_D},\nu_0} f(w, v; z)$ *with* $n = N$*. Then, for* d_1, d_2 *from Lemma 1,*

$$\mathbb{E}[\mathcal{R}(\widehat{w}^{(T)}) - \widehat{\mathcal{R}}_n(\widehat{w}^{(T)})] \le \frac{C_{\mathrm{opt}}\, d_1}{d_2 + T} + \frac{C_{\mathrm{gen}}}{n},$$

where C_{opt} *and* C_{gen} *depend only on* $L, \rho, \mu_{\mathrm{PL}}, \mu_{\mathrm{QG}}, G$*.*[9]

The generalization guarantee in Theorem 4 bounds the expected population risk by the expected empirical risk plus the displayed error. The ultimate measure of success is the population risk relative to the best possible model. This is quantified by the *population excess risk*, $\mathcal{R}(\widehat{w}^{(T)}) - \mathcal{R}(w^\star)$.

Theorem 5 (Population excess risk)**.** *Under the conditions of Theorem 4, let* $w^\star \in \arg\min_w \mathcal{R}(w)$*. Then*

$$\mathbb{E}[\mathcal{R}(\widehat{w}^{(T)}) - \mathcal{R}(w^\star)] \le \frac{C_{\mathrm{opt}}\, d_1}{d_2 + T} + \frac{C_{\mathrm{stat}}}{n}, \qquad C_{\mathrm{stat}} := \frac{L_\Phi(1 + L/\rho)^2 G^2}{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}},$$

where the first term is the optimization error and the second the statistical error.

Remark 3.3 (Weak primal–dual gap)**.** *The primal-risk bounds above borrow the structure of the proof of Theorem 2. The weak primal–dual gap of Wang et al. [2022],* $\Delta^{\mathrm{PD}}(w_T, v_T) := \max_{v'} \mathbb{E}F(w_T, v') - \min_{w'} \mathbb{E}F(w', v_T)$*, compares the output with fixed comparators on each side, and it follows directly from stability: for independent observations, the exchange argument of Bousquet and Elisseeff [2002] gives* $|\Delta^{\mathrm{PD}} - \Delta^{\mathrm{PD}}_n| \le G\varepsilon^{\mathrm{joint}}_T$*, where* Δ^{PD}_n *is its empirical counterpart, and the joint statement of Theorem 2 makes this* $O((d_2 + T)^{-1/2}) + O(1/n)$*. See Appendix B.3.*

4 Numerical Illustration

We check the n^{-1} rates of Theorems 2 and 5 on a small soft MDP ($|\mathcal{S}| = 5$, $|\mathcal{A}| = 3$, $\beta = 0.9$) in which all population quantities are exact. Two-layer softplus networks of width $H \in \{8, 16, 32\}$ are trained by Algorithm 1 on $n \in \{100, \dots, 6400\}$ transitions, i.i.d. or from a single trajectory, and each run on $\mathcal{D}$ is coupled with a run on a neighbor $\mathcal{D}^{(i)}$ through the same initialization and minibatch indices (details in Appendix E). Code is available at `https://github.com/jajajang/soft-brm-stability`.

Both stability and excess MSBE decay as n^{-1} (Figure 1): over $n \ge 200$ the fitted slopes are -1.08 to -1.05 for ε^w_T and -1.01 to -0.97 for the excess MSBE, and every 95% bootstrap interval excludes $-1/2$. A single dependent trajectory gives the same slopes, as the stability bound needs no independence. The SGDA output is within 10% of the exact empirical minimizer for $200 \le n \le 3200$, and the width changes the constant[10] but not the rate. In short, both the stability of SGDA and the excess MSBE of its output follow the n^{-1} rates predicted by Theorems 2 and 5.

[8] The stability bound of Theorem 2 needs no independence: the minibatch indices are drawn uniformly by the algorithm, so no mixing time enters it, unlike the Markov-chain indices of Wang et al. [2022]. For stationary observations whose marginal matches the target distribution and whose fixed-pair gradient covariances are summable, the primal-risk statements here and in Theorems 4–5 hold with G^2 replaced by $G^2 C_{\mathrm{cov}}$ (Appendix B.2); C_{cov} plays the role of a mixing time.

[9] $C_{\mathrm{opt}} = 2L_\Phi/\mu_{\mathrm{QG}}$ and $C_{\mathrm{gen}} = (1 + L/\rho)^2 G^2 \big(L_\Phi/(\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}) + 1/(2\mu_{\mathrm{PL}})\big)$.

[10] Appendix E also compares the constants in these theorems with the observed values.

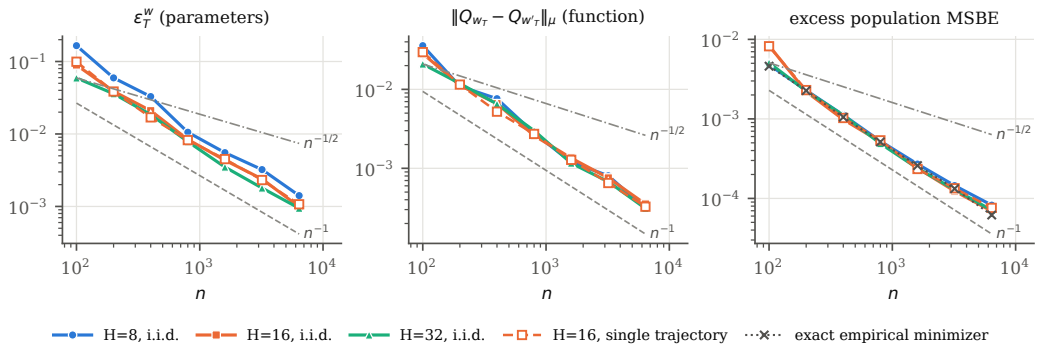

Figure 1: Primal argument stability, function-space stability and excess population MSBE of the SGDA output against n (log–log), averaged over 50 repetitions. Solid lines: i.i.d. data; dashed: a single trajectory; dotted: exact empirical minimizer. Grey guides show slopes -1 and $-1/2$.

5 Conclusion and Limitations

We studied the statistical behavior of Bellman Residual Minimization (BRM) in offline RL/IRL through the lens of stability. Using the primal and dual PL geometry of the bi-conjugate formulation, a single Lyapunov potential and empirical-minimizer sensitivity give an $O(1/n)$ on-average primal argument-stability term for SGDA (Theorem 2), $O(1/n)$ generalization for BRM (Theorem 4), and a population excess-risk bound that separates optimization and statistical error (Theorem 5). The analysis needs no variance reduction, and stationary dependence enters only through a covariance factor.

Implications for offline RL. Theorem 5 answers the question in the Introduction: an excess MSBE of ϵ requires $n = O(C_{\text{stat}}/\epsilon)$ transitions and a number of SGDA iterations of order n. Beyond that point, longer optimization on the same dataset cannot improve the guarantee, which is then limited by the statistical term.

Limitations and future work. This work focuses on finite-sample stability and generalization guarantees for smooth, entropy-regularized BRM, and therefore does not address genuinely non-smooth BRM variants or the conversion from Bellman-residual guarantees to policy-performance guarantees without additional coverage assumptions. We discuss these topics further in Appendix C.6, and view them as natural next steps toward a more complete statistical theory of BRM-based offline reinforcement learning. In addition, our theory guarantees the PL conditions only for relatively wide neural networks (Appendix C.2), and the geometry of soft BRM at practical network sizes requires further study.

Acknowledgments and Disclosure of Funding

K. Jang was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)] and by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT). The authors declare no competing interests.

References

M. Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71 (1):89–129, 2008.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier, 1995.

Jonathan Bassen, Bharathan Balaji, Michael Schaarschmidt, Candace Thille, Jay Painter, Dawn Zimmaro, Alex Games, Ethan Fast, and John C. Mitchell. Reinforcement learning for the adaptive scheduling of educational activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 610–626. PMLR, 2020.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International conference on machine learning*, pages 745–754. PMLR, 2018.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1042–1051, 2019.

Yixuan Chen, Yubin Shi, Mingzhi Dong, Xiaochen Yang, Dongsheng Li, Yujiang Wang, Robert P. Dick, Qin Lv, Yingying Zhao, Fan Yang, Ning Gu, and Li Shang. Over-parameterized model optimization with polyak-łojasiewicz condition. In *International Conference on Learning Representations*, 2023.

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*, pages 1125–1134. PMLR, 2018.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

Yaqi Duan, Chi Jin, and Zhiyuan Li. Risk bounds and Rademacher complexity in batch reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2892–2902, 2021.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

Farzan Farnia and Asuman Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *Proceedings of the 38th International Conference on Machine Learning*, pages 3174–3185. PMLR, 2021.

Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

Guillaume Garrigos and Robert M. Gower. Handbook of convergence theorems for (stochastic) gradient methods. arXiv preprint arXiv:2301.11235, 2023. URL `https://arxiv.org/abs/2301.11235`.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

Yichun Hu, Nathan Kallus, and Masatoshi Uehara. Fast rates for the regret of offline reinforcement learning. *Mathematics of Operations Research*, 50(1):633–655, 2025. doi: 10.1287/moor.2021.0167.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statistical Science*, 2025. arXiv:2510.04088.

Enoch H. Kang, Hema Yoganarasimhan, and Lalit Jain. An empirical risk minimization approach for offline inverse rl and dynamic discrete choice model, 2025. URL `https://arxiv.org/abs/2502.14131`.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. *arXiv preprint arXiv:1608.04636*, 2016. URL `https://arxiv.org/abs/1608.04636`.

Ilja Kuzborskij and Christoph H. Lampert. Data-dependent stability of stochastic gradient descent. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2815–2824. PMLR, 2018.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. *Proceedings of the 37th International Conference on Machine Learning*, pages 5809–5819, 2020.

Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In Marina Meilă and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6175–6186. PMLR, 2021. URL `https://proceedings.mlr.press/v139/lei21b.html`.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59: 85–116, 2022.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, 2018.

Odalric-Ambrym Maillard, Rémi Munos, Alessandro Lazaric, and Mohammad Ghavamzadeh. Finite-sample analysis of Bellman residual minimization. In *Proceedings of the 2nd Asian Conference on Machine Learning*, volume 13 of *Proceedings of Machine Learning Research*, pages 299–314, 2010.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Andrew Patterson, Adam White, and Martha White. A generalized projected bellman error for off-policy value estimation in reinforcement learning. *Journal of Machine Learning Research*, 23 (145):1–61, 2022.

Boris T. Polyak. Gradient methods for minimizing functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

John Rust. Structural estimation of markov decision processes. *Handbook of econometrics*, 4: 3081–3143, 1994.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2019.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

John N Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22(1):59–94, 1996.

Hado Van Hasselt, Yotam Doron, Florian Strub, Matteo Hessel, Nicolas Sonnerat, and Joseph Modayil. Deep reinforcement learning and the deadly triad. *arXiv preprint arXiv:1812.02648*, 2018.

Puyu Wang, Yunwen Lei, Yiming Ying, and Ding-Xuan Zhou. Stability and generalization for markov chain stochastic gradient methods. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f61538f83b0f19f9306d9d801c15f41c-Abstract-Conference.html.

Ziqing Xu, Hancheng Min, Salma Tarmoun, Enrique Mallada, and René Vidal. A local polyak-łojasiewicz and descent lemma of gradient descent for overparametrized linear models. *Transactions on Machine Learning Research*, 2025.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems, 2020. URL https://arxiv.org/abs/2002.09621.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2020.

Siqi Zhang, Yifan Hu, Liang Zhang, and Niao He. Generalization bounds of nonconvex-(strongly)-concave stochastic minimax optimization. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 694–702. PMLR, 2024.

Yufan Zhao, Michael R. Kosorok, and Donglin Zeng. Reinforcement learning design for cancer clinical trials. *Statistics in Medicine*, 28(26):3294–3315, 2009.

Miaoxi Zhu, Li Shen, Bo Du, and Dacheng Tao. Stability and generalization of the decentralized stochastic gradient descent ascent algorithm. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Appendix

A Notations

Symbol	Meaning
$\mathcal{S}$	State space
$\mathcal{A}$	Action space
$P(\cdot\|s,a)$	Transition kernel over the next state from the state-action pair (s,a)
$r:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$	Deterministic reward function
$\Delta_{\mathcal{A}}^{\mathcal{S}}$	Set of stationary Markov policies
$\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}$	Policy
$\mathcal{H}(q)$	Shannon entropy of an action distribution q
β	Discount factor
η	Entropy regularization coefficient
$\mathcal{T}$	Bellman optimality operator
$\hat{\mathcal{T}}$	Sampled Bellman operator
$\delta_Q(s,a,s')$	Temporal-Difference error
Q^*	(The unique) solution to Bellman optimality equation
$z_i=(s_i,a_i,s_i')$	i-th sample
$\mathcal{D}=\{z_i\}_{i=1}^n$	Dataset
$f(w,v;z_i)$	Per-sample objective
$\mathcal{D}^{(i)}$	neighboring dataset
g_t^w, g_t^v	Minibatch gradients
η_t	SGDA learning rate at time t
$F_D(w,v)$	$\frac{1}{n}\sum_{j=1}^n f(w,v;z_j^D)$
$\Phi_D(w)$	$\max_v F_D(w,v)$
$\Phi_D^\star$	$\min_w \Phi_D(w)$
L	Smoothness (Condition (C1))
μ_{PL}	Polyak–Łojasiewicz condition constant (Condition (C2))
μ_{QG}	Quadratic Growth constant (Condition (C3))
ρ	Inner PL constant (Condition (C4))
G	Per-sample gradient bound (Condition (C5))
$\Psi_{\alpha,D}(w,v)$	Lyapunov potential, $\Phi_D(w)-\Phi_D^\star+\alpha\cdot\big(\Phi_D(w)-F_D(w,v)\big)$
$A(w)$	$\Phi_D(w)-\Phi_D^\star$
$\Gamma(w,v)$	$\Phi_D(w)-F_D(w,v)$
$R(w), R_n(w)$	Primal Risk (Definition 2)
Δ^{PD}, Δ_n^{PD}	Weak primal–dual risk (Definition 3)

Table 2: Notations

B Technical Proofs

B.1 Supporting Lemmas for proving Theorem 2

Lemma 6 (PL error bound)**.** *Under Conditions (C1) and (C4), let*

$$V_D(w):=\arg\max_v F_D(w,v),\qquad \Gamma_D(w,v):=\Phi_D(w)-F_D(w,v).$$

Then

$$\operatorname{dist}(v,V_D(w))\le\sqrt{\frac{2\Gamma_D(w,v)}{\rho}}\le\frac{\|\nabla_v F_D(w,v)\|}{\rho}.$$

Proof. Fix w and follow $\dot v=\nabla_v F_D(w,v)$ from $v(0)=v$. Writing $q(t)=\Gamma_D(w,v(t))$, inner PL gives $\dot q=-\|\dot v\|^2\le-2\rho q$ and $-\frac{d}{dt}\sqrt{q}\ge\sqrt{\rho/2}\,\|\dot v\|$. Integration shows that the trajectory has length at most $\sqrt{2q(0)/\rho}$ and converges to a point of $V_D(w)$. This proves the first inequality; inner PL

proves the second. The domain conditions ensure that this auxiliary ascent trajectory remains in the region on which the inequalities hold. □

Lemma 7 (Smoothness of the value function Φ)**.** *Assume Conditions (C1) and (C4). Then:*

(i) For every $p \in V_D(w)$ there is $q \in V_D(u)$ such that $\|p - q\| \le (L/\rho)\|w - u\|$.

(ii) Φ_D is differentiable and $\nabla\Phi_D(w) = \nabla_w F_D(w, p)$ for every $p \in V_D(w)$.

(iii) Φ_D is L_Φ-smooth, where $L_\Phi := L(1 + L/\rho)$.

Proof. Since $\nabla_v F_D(w, p) = 0$, Lemma 6 gives

$$\operatorname{dist}(p, V_D(u)) \le \rho^{-1}\|\nabla_v F_D(u, p)\| \le (L/\rho)\|u - w\|,$$

proving (i). Choose such a q, and put $h = u - w$. Using $F_D(w, p) = \Phi_D(w)$ and $F_D(w, q) \le \Phi_D(w)$, smoothness yields

$$\begin{aligned}\langle \nabla_w F_D(w, p), h\rangle - \tfrac{L}{2}\|h\|^2 &\le \Phi_D(u) - \Phi_D(w)\\ &\le \langle \nabla_w F_D(w, p), h\rangle + (\tfrac{L}{2} + L^2/\rho)\|h\|^2.\end{aligned}$$

Thus (ii) holds for every p. Finally,

$$\|\nabla\Phi_D(w) - \nabla\Phi_D(u)\| \le L(\|w - u\| + \|p - q\|) \le L_\Phi\|w - u\|.$$

□

Lemma 8 (Mismatch control)**.** *Under Conditions (C1) and (C4),*

$$\|\Delta(w, v)\| := \|\nabla_w F_D(w, v) - \nabla\Phi_D(w)\| \le L \operatorname{dist}(v, V_D(w)) \le L\sqrt{\frac{2\Gamma_D(w, v)}{\rho}}.$$

Proof. Choose a closest $p \in V_D(w)$, apply Lemma 7(ii) and joint smoothness, and then apply Lemma 6. □

Lemma 9 (Cross-dataset gradient sensitivity for Φ)**.** *Under Conditions (C1), (C4), (C5), and (S3), a one-sample replacement $D \to D^{(i)}$ gives*

$$\|\nabla\Phi_D(w) - \nabla\Phi_{D^{(i)}}(u)\| \le L_\Phi\|w - u\| + \frac{2G}{n}(1 + L/\rho).$$

Proof. Fix $p' \in V_{D^{(i)}}(u)$. The replacement identity and Condition (C5) give $\|\nabla_v F_D(u, p')\| \le 2G/n$. By Lemma 6, there is $p \in V_D(u)$ with $\|p - p'\| \le 2G/(\rho n)$. Therefore

$$\|\nabla\Phi_D(u) - \nabla\Phi_{D^{(i)}}(u)\| \le L\|p - p'\| + 2G/n \le \frac{2G}{n}(1 + L/\rho).$$

Add $\|\nabla\Phi_D(w) - \nabla\Phi_D(u)\| \le L_\Phi\|w - u\|$. □

Lemma 10 (Deterministic potential ⇒ distance)**.** *Under Condition (C3), for every $\alpha > 0$,*

$$\|w - x_D^\star\| \le \sqrt{\frac{2\Psi_{\alpha,D}(w, v)}{\mu_{\mathrm{QG}}}}.$$

If, additionally, $V_D(x_D^\star) = \{v_D^\star\}$ and Conditions (C1) and (C4) hold, then

$$\|w - x_D^\star\| + \|v - v_D^\star\| \le C_{\mathrm{dist}}\sqrt{\Psi_{\alpha,D}(w, v)}, \qquad C_{\mathrm{dist}} := \sqrt{\frac{2(1 + L/\rho)^2}{\mu_{\mathrm{QG}}} + \frac{2}{\alpha\rho}}.$$

Proof. The primal claim follows from QG and $\Psi_{\alpha,D} \ge A_D := \Phi_D - \Phi_D^\star$. For the joint claim choose a closest $p \in V_D(w)$ to v. Lemma 7(i) and the singleton condition give $\|p - v_D^\star\| \le (L/\rho)\|w - x_D^\star\|$. Hence

$$\|w - x_D^\star\| + \|v - v_D^\star\| \le (1 + L/\rho)\sqrt{2A_D(w)/\mu_{\mathrm{QG}}} + \sqrt{2\Gamma_D(w, v)/\rho}.$$

Weighted Cauchy–Schwarz gives the displayed constant. □

Lemma 11 (Primal and saddle-point sensitivity)**.** *Under Conditions (C1)–(C5) and Setting (S3),*

$$\|x_D^\star - x_{D^{(i)}}^\star\| \le \frac{S_w}{n}, \qquad S_w := \frac{2G(1+L/\rho)}{\sqrt{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}}}.$$

If, additionally, $V_D(x_D^\star) = \{v_D^\star\}$ for both compared datasets, then

$$\|x_D^\star - x_{D^{(i)}}^\star\| + \|v_D^\star - v_{D^{(i)}}^\star\| \le \frac{S_{\mathrm{sens}}}{n}, \qquad S_{\mathrm{sens}} := 2G\left(\frac{(1+L/\rho)^2}{\sqrt{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}}} + \frac{1}{\rho}\right).$$

Proof. At $u = x_{D^{(i)}}^\star$, Lemma 9 gives $\|\nabla\Phi_D(u)\| \le 2G(1+L/\rho)/n$. Combining outer PL with QG yields

$$\frac{\mu_{\mathrm{QG}}}{2}\|u - x_D^\star\|^2 \le \Phi_D(u) - \Phi_D^\star \le \frac{\|\nabla\Phi_D(u)\|^2}{2\mu_{\mathrm{PL}}},$$

which proves the primal bound. For the joint bound choose $p \in V_D(u)$ with $\|p - v_{D^{(i)}}^\star\| \le 2G/(\rho n)$, as in the proof of Lemma 9. The singleton condition and Lemma 7(i) imply

$$\|v_D^\star - v_{D^{(i)}}^\star\| \le (L/\rho)\|x_D^\star - x_{D^{(i)}}^\star\| + 2G/(\rho n).$$

Add the primal bound. □

Lemma 12 (Coupled one-step bounds)**.** *Set $\alpha = 1$ and $\Xi_t := \Psi_{1,\mathcal{D}}(w_t, v_t) + \Psi_{1,\mathcal{D}^{(i)}}(w_t', v_t')$. Under the geometric and sampling conditions of Lemma 1 and any stepsizes $0 < \eta_t \le \bar\eta$,*

$$\mathbb{E}[\Psi_{1,D}(w_{t+1}, v_{t+1}) \mid \mathcal{F}_t] \le (1 - c\eta_t)\Psi_{1,D}(w_t, v_t) + C'_{\mathrm{var}}\eta_t^2, \tag{6}$$

$$\mathbb{E}[\Xi_{t+1} \mid \mathcal{F}_t] \le (1 - c\eta_t)\Xi_t + 2C'_{\mathrm{var}}\eta_t^2. \tag{7}$$

Proof. Fix D and suppress its subscript. Put $g = \nabla\Phi(w_t)$, $p = \nabla_w F(w_t, v_t)$, $q = \nabla_v F(w_t, v_t)$, and $\Delta = p - g$. Smoothness of Φ and F, followed by conditional unbiasedness, gives

$$\begin{aligned}\mathbb{E}[\Psi_1(w_{t+1}, v_{t+1}) \mid \mathcal{F}_t] - \Psi_1(w_t, v_t) \le \eta_t(-2\langle g, p\rangle + \|p\|^2 - \kappa_v\|q\|^2) \\ + \eta_t^2\big(H\|p\|^2 + (L\kappa_v^2/2)\|q\|^2 + C'_{\mathrm{var}}\big).\end{aligned}$$

The mixed term cancels: $-2\langle g, p\rangle + \|p\|^2 = -\|g\|^2 + \|\Delta\|^2$. The inner PL error bound and Lemma 8 imply $\|\Delta\| \le (L/\rho)\|q\|$. Hence the first-order term is at most $-\eta_t(\|g\|^2 + \|q\|^2)$. Also $\|p\|^2 \le 2\|g\|^2 + 2(L/\rho)^2\|q\|^2$. The definition of $\bar\eta$ therefore bounds the mean-gradient Taylor remainder by $\eta_t(\|g\|^2 + \|q\|^2)/2$. Finally, the outer and inner PL inequalities yield

$$-\tfrac{1}{2}\|g\|^2 - \tfrac{1}{2}\|q\|^2 \le -\mu_{\mathrm{PL}}(\Phi(w_t) - \Phi^\star) - \rho(\Phi(w_t) - F(w_t, v_t)) \le -c\Psi_1(w_t, v_t).$$

This proves (6); adding the two runs proves (7). □

Proof of Lemma 1. By (6), writing $r = c_2 + t$, induction gives

$$\mathbb{E}\Psi_{1,D}(w_{t+1}, v_{t+1}) \le \frac{d_1}{r} - \frac{cc_1d_1 - C'_{\mathrm{var}}c_1^2}{r^2} \le \frac{d_1}{r} - \frac{d_1}{r^2} \le \frac{d_1}{r+1}.$$

The almost-supermartingale theorem gives $\Psi_{1,D}(w_t, v_t) \to 0$ almost surely because $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$. To establish parameter convergence, smoothness, anchored QG, and Lemmas 8 and 6 give

$$\begin{aligned}\|(-\nabla_w F_D, \kappa_v\nabla_v F_D)(w_t, v_t)\|^2 \le C_H\Psi_{1,D}(w_t, v_t), \\ C_H = \max\{4L_\Phi^2/\mu_{\mathrm{QG}}, (4 + 2\kappa_v^2)L^2/\rho\}.\end{aligned}$$

Thus the expected sum of the norms of the true-gradient increments is finite, since $\sum_t \eta_t(c_2 + t)^{-1/2} < \infty$. Their sum converges absolutely almost surely. The noise increments form a martingale series with summable second moments, so their sum also converges almost surely. The iterates consequently have a finite limit. Continuity and $\Psi_{1,D} \to 0$ show that this limit is a minimax pair. □

Lemma 13 (One-step ascent under PL)**.** *Let g be L-smooth, attain $g^\star = \max_v g(v)$, and satisfy $\frac{1}{2}\|\nabla g(v)\|^2 \ge \rho(g^\star - g(v))$. For $v^+ = v + \gamma\nabla g(v)$ and $0 \le \gamma \le 2/L$,*

$$g^\star - g(v^+) \le (1 - 2\rho\gamma + \rho L\gamma^2)(g^\star - g(v)).$$

Proof. Smoothness gives $g(v^+) \geq g(v) + (\gamma - L\gamma^2/2)\|\nabla g(v)\|^2$. The coefficient is nonnegative, so applying PL proves the claim. □

Lemma 14 (Closing the potential recursion)**.** *For either dataset D, define*

$$B_T := e^{-c\sum_{s=0}^{T-1}\eta_s}\Psi_{1,0}^{\max} + C'_{\text{var}}\sum_{t=0}^{T-1}\eta_t^2 e^{-c\sum_{s=t+1}^{T-1}\eta_s}.$$

Then $\mathbb{E}\Psi_{1,D}(w_T, v_T) \leq B_T$ and $\mathbb{E}\Xi_T \leq 2B_T$.

Proof. Iterate (6) and use $1 - c\eta_s \leq e^{-c\eta_s}$. □

B.2 Proof of Main Result

Proposition B.1 (On-average argument stability of SGDA for general stepsizes)**.** *Suppose that the loss function f and datasets $\mathcal{D}, \mathcal{D}^{(i)}$ satisfy (C1)–(C6) under the setting (S1)–(S3). Let $\varepsilon_T^w := \frac{1}{n}\sum_{i=1}^n \mathbb{E}\|w_T(\mathcal{D}) - w_T(\mathcal{D}^{(i)})\|$. With $\alpha = 1$ and the dual stepsize multiplier κ_v above, the following on-average argument stability bound holds:*[11]

$$\varepsilon_T^w \leq 2\,C_{\text{dist}}\left[e^{-\frac{3c}{4}\sum_{s=0}^{T-1}\eta_s}\,\Psi_{\alpha,0}^{\max} + C'_{\text{var}}\sum_{t=0}^{T-1}\eta_t^2\,e^{-\frac{3c}{4}\sum_{s=t+1}^{T-1}\eta_s}\right]^{1/2} + \frac{C'_{\text{hit}}}{n},$$

where $\Psi_{\alpha,0}^{\max} := \max\{\mathbb{E}[\Psi_{\alpha,\mathcal{D}}(w_0, v_0)],\ \max_{1\leq i\leq n}\mathbb{E}[\Psi_{\alpha,\mathcal{D}^{(i)}}(w_0, v_0)]\}$, $c := \min\{\mu_{\text{PL}}/2,\ \rho/2\}$, $C'_{\text{var}} = (L_\Phi + L/2 + L\kappa_v^2/2)\sigma_B^2$, and C_{dist}, C'_{hit} are as in Theorem 2. If, in addition, $V_D(x_D^\star) = \{v_D^\star\}$ for every compared dataset, the same bound holds for $\varepsilon_T^{\text{joint}} := n^{-1}\sum_i \mathbb{E}[\|w_T - w'_T\| + \|v_T - v'_T\|]$. This additional condition is not needed for the primal stability or risk guarantees.

Proof of Proposition B.1 and Theorem 2. Let q_T denote the bracketed quantity in Proposition B.1. The optimization bound gives $\mathbb{E}\Psi_{1,D}(w_T, v_T) \leq q_T$ for $D \in \{\mathcal{D}, \mathcal{D}^{(i)} : i \in [n]\}$. Insert the two empirical primal minimizers between the coupled outputs. Lemmas 10 and 11 give

$$\|w_T - w'_T\| \leq \sqrt{\frac{2\Psi_{1,\mathcal{D}}(w_T, v_T)}{\mu_{\text{QG}}}} + \sqrt{\frac{2\Psi_{1,\mathcal{D}^{(i)}}(w'_T, v'_T)}{\mu_{\text{QG}}}} + \frac{S_w}{n}.$$

Take expectations, apply Jensen's inequality, and average over i:

$$\varepsilon_T^w \leq 2\sqrt{\frac{2q_T}{\mu_{\text{QG}}}} + \frac{S_w}{n}.$$

Since $C_{\text{dist}} \geq \sqrt{2/\mu_{\text{QG}}}$ and $C'_{\text{hit}} = S_{\text{sens}} \geq S_w$, this proves the displayed bound with the original constants. Under the harmonic schedule with $cc_1 > 1$, the sharper optimization estimate $\mathbb{E}\Psi_{1,D}(w_T, v_T) \leq d_1/(c_2 + T)$ gives Theorem 2 by the same argument. If $V_D(x_D^\star) = \{v_D^\star\}$ for every compared dataset, the joint conclusions of Lemmas 10 and 11 instead give

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\|w_T - w'_T\| + \|v_T - v'_T\|\big] \leq 2C_{\text{dist}}\sqrt{q_T} + \frac{S_{\text{sens}}}{n}.$$

This proves the optional joint statement. □

Proof of Lemma 3. Fix the population pair $(w^\star, v^\star)$ from the statement. The envelope identity and interior optimality give $\nabla_w F(w^\star, v^\star) = \nabla_v F(w^\star, v^\star) = 0$. The inner PL error bound supplies $u_D \in \arg\max_v F_D(w^\star, v)$ with

$$\|u_D - v^\star\| \leq \rho^{-1}\|\nabla_v F_D(w^\star, v^\star)\|.$$

Consequently, the envelope identity and smoothness imply

$$\|\nabla\Phi_D(w^\star)\| \leq \|\nabla_w F_D(w^\star, v^\star)\| + (L/\rho)\|\nabla_v F_D(w^\star, v^\star)\|.$$

[11] Under the Robbins–Monro conditions $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$, the optimization term (the square root term) vanishes as $T \to \infty$ [Garrigos and Gower, 2023].

For independent data, each empirical block gradient at this fixed pair is an average of independent, mean-zero vectors of norm at most G. Its second moment is at most G^2/n. The L^2 triangle inequality therefore gives $r_n \le (1 + L/\rho)^2 G^2/n$.

Write $x_D^\star$ for the empirical primal minimizer. Primal PL and the retained anchored QG condition yield

$$\|x_D^\star - w^\star\|^2 \le \frac{\|\nabla \Phi_D(w^\star)\|^2}{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}}, \qquad \mathbb{E}\|w_T - x_D^\star\|^2 \le \frac{2a_T}{\mu_{\mathrm{QG}}}.$$

By L_Φ-smoothness of R and $\nabla R(w^\star) = 0$,

$$\begin{aligned}\mathbb{E}[R(w_T) - R(w^\star)] &\le \frac{L_\Phi}{2}\mathbb{E}\|w_T - w^\star\|^2 \\ &\le \frac{2L_\Phi}{\mu_{\mathrm{QG}}} a_T + \frac{L_\Phi r_n}{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}}. \end{aligned} \tag{8}$$

For generalization, use

$$\mathbb{E}[R(w_T) - \Phi_D(w_T)] = \mathbb{E}[R(w_T) - R(w^\star)] - a_T + \mathbb{E}[R(w^\star) - \Phi_D^\star].$$

Since $\mathbb{E}\Phi_D(w^\star) \ge R(w^\star)$ and primal PL gives $\mathbb{E}[\Phi_D(w^\star) - \Phi_D^\star] \le r_n/(2\mu_{\mathrm{PL}})$, the last term is at most $r_n/(2\mu_{\mathrm{PL}})$. Combining this with (8) and dropping $-a_T$ proves the primal-risk claim. In particular, the empirical maximum need not be unbiased at a fixed $w^\star$; the displayed inequality accounts for its optimism.

For Lemma 15, use a common random seed in the compared runs and fix a deterministic comparator v'. Exchanging the independent observations Z_i, Z_i' gives

$$\mathbb{E}[F(w_T(D), v') - F_D(w_T(D), v')] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big[f(w_T(D^{(i)}), v'; Z_i) - f(w_T(D), v'; Z_i)\Big].$$

Its absolute value is at most $Gn^{-1}\sum_i \mathbb{E}\|w_T(D^{(i)}) - w_T(D)\|$, uniformly in the fixed comparator. Taking maxima preserves this bound. The same argument for a fixed primal comparator bounds the difference of the two minima by $Gn^{-1}\sum_i \mathbb{E}\|v_T(D^{(i)}) - v_T(D)\|$. The triangle inequality gives $G\varepsilon_T^{\mathrm{joint}}$. The gradient bounds must hold along these comparisons with the fixed comparator; a common bounded product domain is sufficient. □

Proofs of Theorems 4 and 5. The BRM primal risks equal R and Φ_D in Lemma 3. Lemma 1 gives $a_T \le d_1/(d_2 + T)$. Substituting this inequality and $r_n \le (1 + L/\rho)^2 G^2/n$ into Lemma 3 proves Theorem 4. Substituting the same bounds into (8) proves Theorem 5. □

Stationary dependent data. Let $g_{b,j} := \nabla_b f(w^\star, v^\star; Z_j)$ for $b \in \{w, v\}$. The preceding fixed-pair argument requires only its second moment, and gives $r_n \le (\sqrt{V_{w,n}} + (L/\rho)\sqrt{V_{v,n}})^2$, where $V_{b,n} := \mathbb{E}\|n^{-1}\sum_j g_{b,j}\|^2$. For stationary observations with the target marginal distribution,

$$V_{b,n} = \frac{1}{n}\left[\mathbb{E}\|g_{b,0}\|^2 + 2\sum_{k=1}^{n-1}(1 - k/n)\mathbb{E}\langle g_{b,0}, g_{b,k}\rangle\right].$$

Thus, if

$$C_{\mathrm{cov}} := 1 + \frac{2}{G^2}\max_{b\in\{w,v\}}\sum_{k=1}^{\infty}\big|\mathbb{E}\langle g_{b,0}, g_{b,k}\rangle\big| < \infty,$$

then $V_{b,n} \le G^2 C_{\mathrm{cov}}/n$ and $r_n \le (1 + L/\rho)^2 G^2 C_{\mathrm{cov}}/n$. The primal generalization and excess-risk proofs therefore apply with this dependence factor. Their target distribution must match the stationary marginal; arbitrary dependence or a different target requires the corresponding variance or distribution-shift term.

B.3 Weak primal–dual risk

Following Wang et al. [2022], the stability-to-generalization principle also controls the weak primal–dual gap whenever the joint output has on-average argument stability.

Definition 3 (Weak primal–dual risk)**.** *For a random algorithm output (w_T, v_T), define the population and empirical weak–PD risks by*

$$\Delta^{\mathrm{PD}}(w_T, v_T) := \max_{v' \in \mathcal{V}} \mathbb{E}F(w_T, v') - \min_{w' \in \mathcal{W}} \mathbb{E}F(w', v_T),$$
$$\Delta_n^{\mathrm{PD}}(w_T, v_T) := \max_{v' \in \mathcal{V}} \mathbb{E}F_{\mathcal{D}}(w_T, v') - \min_{w' \in \mathcal{W}} \mathbb{E}F_{\mathcal{D}}(w', v_T).$$

The expectations include the data and algorithmic randomness.

Lemma 15 (Weak primal–dual transfer)**.** *Assume the conditions of Lemma 3 with independent observations. If the joint output has on-average argument stability*

$$\varepsilon_T^{\mathrm{joint}} := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\big[\|w_T(\mathcal{D}) - w_T(\mathcal{D}^{(i)})\| + \|v_T(\mathcal{D}) - v_T(\mathcal{D}^{(i)})\|\big],$$

and the finite extrema in Definition 3 can be taken over fixed bounded comparator sets containing the outputs, on whose product the block-gradient bounds hold, then

$$\big|\Delta^{\mathrm{PD}}(w_T, v_T) - \Delta_n^{\mathrm{PD}}(w_T, v_T)\big| \le G\varepsilon_T^{\mathrm{joint}}.$$

The optional condition $\arg\max_v F_D(x_D^\star, v) = \{v_D^\star\}$ *gives the joint stability bound through the potential-to-distance and saddle-sensitivity estimates; it is not needed for the primal-risk bound.*

In particular, under the conditions of Theorem 4 and the joint-stability conditions above, the output of Algorithm 1 satisfies $|\Delta^{\mathrm{PD}}(\widehat{w}^{(T)}, \widehat{v}^{(T)}) - \Delta_n^{\mathrm{PD}}(\widehat{w}^{(T)}, \widehat{v}^{(T)})| \le G\varepsilon_T^{\mathrm{joint}}$. The proof of Lemma 15 is part of the proof of Lemma 3 above.

C Further discussions

C.1 Justification of Condition (C5)

For Algorithm 1, Condition (C5) is retained as an assumption on the effective domain. The following lemma gives a sufficient bounded-domain condition for the deterministic AGDA recursion stated in it.

Lemma 16 (Bounded effective domain and bounded per-sample gradients for AGDA)**.** *Fix a dataset $\mathcal{D} = \{z_i\}_{i=1}^n$ and a per-sample objective $f(\cdot, \cdot; z)$. Let the empirical saddle objective be*

$$F_{\mathcal{D}}(w, v) \;:=\; \frac{1}{n} \sum_{i=1}^{n} f(w, v; z_i), \qquad g_{\mathcal{D}}(w) \;:=\; \max_v F_{\mathcal{D}}(w, v), \qquad g_{\mathcal{D}}^\star := \min_w g_{\mathcal{D}}(w).$$

Assume that $F_{\mathcal{D}}$ satisfies the following conditions stated using the notation and constants of Yang et al. [2020]: there exist constants $l_{\mathrm{Y}} > 0$ and $\mu_{1,\mathrm{Y}}, \mu_{2,\mathrm{Y}} > 0$ such that

(Y1) *(Lipschitz gradient) $F_{\mathcal{D}}$ has l_{Y}-Lipschitz gradients in the joint variable (w, v) in the sense of Yang et al. [2020, Assumption 1].*

(Y2) *(Existence of saddle point) $F_{\mathcal{D}}$ has at least one saddle point as in Yang et al. [2020, Assumption 2].*

(Y3) *(Two-sided PL) $F_{\mathcal{D}}$ satisfies the two-sided PL condition with constants $\mu_{1,\mathrm{Y}}, \mu_{2,\mathrm{Y}}$ as in Yang et al. [2020, Definition 2].*

Run deterministic *AGDA (i.e., Algorithm 1 of Yang et al. [2020] with $\sigma^2 = 0$) with constant step sizes*

$$\tau_1 \;=\; \frac{\mu_{2,\mathrm{Y}}^2}{18\, l_{\mathrm{Y}}^3}, \qquad \tau_2 \;=\; \frac{1}{l_{\mathrm{Y}}},$$

initialized at (w_0, v_0)*:*

$$w_{t+1} = w_t - \tau_1 \nabla_w F_{\mathcal{D}}(w_t, v_t), \qquad v_{t+1} = v_t + \tau_2 \nabla_v F_{\mathcal{D}}(w_{t+1}, v_t).$$

Define the potential (same form as in Yang et al. [2020, Eq. (8)])

$$P_t \;:=\; \big(g_{\mathcal{D}}(w_t) - g_{\mathcal{D}}^\star\big) \;+\; \frac{1}{10}\,\big(g_{\mathcal{D}}(w_t) - F_{\mathcal{D}}(w_t, v_t)\big).$$

Then there exist a saddle point $(w_{\mathcal{D}}^\star, v_{\mathcal{D}}^\star)$ *of* $F_{\mathcal{D}}$ *and a constant* $\alpha_{\mathrm{Y}} > 0$ *(depending only on* $l_{\mathrm{Y}}, \mu_{1,\mathrm{Y}}, \mu_{2,\mathrm{Y}}$*) such that for all* $t \geq 0$*,*

$$\|w_t - w_{\mathcal{D}}^\star\|_2^2 + \|v_t - v_{\mathcal{D}}^\star\|_2^2 \;\leq\; \alpha_{\mathrm{Y}} \left(1 - \frac{\mu_{1,\mathrm{Y}} \mu_{2,\mathrm{Y}}^2}{36\, l_{\mathrm{Y}}^3}\right)^t P_0.$$

In particular, all iterates remain in the closed Euclidean ball

$$\Omega \;:=\; \Big\{(w, v) : \|w - w_{\mathcal{D}}^\star\|_2^2 + \|v - v_{\mathcal{D}}^\star\|_2^2 \leq \alpha_{\mathrm{Y}} P_0\Big\},$$

which is compact and convex.

Moreover, assume additionally that the sample space $\mathcal{Z}$ *is compact (e.g., finite) and that* $(w, v, z) \mapsto \nabla_w f(w, v; z)$ *and* $(w, v, z) \mapsto \nabla_v f(w, v; z)$ *are continuous. Then the constant*

$$G \;:=\; \sup_{(w,v)\in\Omega,\; z\in\mathcal{Z}} \max\Big\{\|\nabla_w f(w, v; z)\|_2,\; \|\nabla_v f(w, v; z)\|_2\Big\}$$

is finite for these AGDA iterates. Condition (C5) additionally requires the comparison points used in the stability and population-risk proofs to lie in the bounded domain.

Proof. Under (Y1)–(Y3) and the specific choice of step sizes $\tau_1 = \mu_{2,\mathrm{Y}}^2/(18 l_{\mathrm{Y}}^3)$ and $\tau_2 = 1/l_{\mathrm{Y}}$, Theorem 3.2 of Yang et al. [2020] applies to the objective $F_{\mathcal{D}}$ and to the deterministic AGDA recursion stated above. Therefore, there exist a saddle point $(w_{\mathcal{D}}^\star, v_{\mathcal{D}}^\star)$ of $F_{\mathcal{D}}$ and a constant $\alpha_{\mathrm{Y}} > 0$ (depending only on $l_{\mathrm{Y}}, \mu_{1,\mathrm{Y}}, \mu_{2,\mathrm{Y}}$) such that for all $t \geq 0$,

$$\|w_t - w_{\mathcal{D}}^\star\|_2^2 + \|v_t - v_{\mathcal{D}}^\star\|_2^2 \;\leq\; \alpha_{\mathrm{Y}} \left(1 - \frac{\mu_{1,\mathrm{Y}} \mu_{2,\mathrm{Y}}^2}{36\, l_{\mathrm{Y}}^3}\right)^t P_0.$$

Since the contraction factor satisfies $0 < 1 - \frac{\mu_{1,\mathrm{Y}}\mu_{2,\mathrm{Y}}^2}{36\, l_{\mathrm{Y}}^3} < 1$, we have $\big(1 - \frac{\mu_{1,\mathrm{Y}}\mu_{2,\mathrm{Y}}^2}{36\, l_{\mathrm{Y}}^3}\big)^t \leq 1$ for all $t \geq 0$. Hence

$$\|w_t - w_{\mathcal{D}}^\star\|_2^2 + \|v_t - v_{\mathcal{D}}^\star\|_2^2 \;\leq\; \alpha_{\mathrm{Y}} P_0 \qquad \forall t \geq 0.$$

This shows $(w_t, v_t) \in \Omega$ for all t. The set Ω is a closed Euclidean ball in a finite-dimensional space, hence compact, and it is convex by construction.

By assumption, the maps $(w, v, z) \mapsto \nabla_w f(w, v; z)$ and $(w, v, z) \mapsto \nabla_v f(w, v; z)$ are continuous, and $\Omega \times \mathcal{Z}$ is compact. Therefore, by the extreme value theorem, the function

$$(w, v, z) \mapsto \max\Big\{\|\nabla_w f(w, v; z)\|_2,\; \|\nabla_v f(w, v; z)\|_2\Big\}$$

attains its maximum on $\Omega \times \mathcal{Z}$, and the supremum

$$G = \sup_{(w,v)\in\Omega,\; z\in\mathcal{Z}} \max\Big\{\|\nabla_w f(w, v; z)\|_2,\; \|\nabla_v f(w, v; z)\|_2\Big\}$$

is finite. This proves the stated gradient bound for the deterministic AGDA domain. □

C.2 Neural network parametrization and PL/QG conditions

Width/radius and high-probability regimes for uniform constants. This subsection records the neural-network conditions relevant to Conditions (C1)–(C5), particularly the PL conditions for the primal value function and the auxiliary regression. Fix a dataset $D = \{z_i\}_{i=1}^n$ and write, as in Section 2.2,

$$F_D(w, v) := \frac{1}{n}\sum_{i=1}^n f(w, v; z_i), \qquad \Phi_D(w) := \max_v F_D(w, v), \qquad \Phi_D^\star := \min_w \Phi_D(w).$$

Here w parametrizes the action-value network Q_w and v parametrizes the auxiliary function ζ_v in the bi-conjugate BRM objective (3). The PL/QG statements below are local: fix a failure level $\delta \in (0,1)$ and a radius $R > 0$, and take the effective domain in Condition (C5) to be the initialization neighborhood

$$\Omega_R := \{(w,v) : \|(w,v) - (w_0,v_0)\| \le R\} \cap (\mathcal{W} \times \mathcal{V}).$$

All constants used in the stability theorem are then understood on this same set: L is the joint smoothness constant of F_D, ρ is the inner PL constant in v, G is the per-sample gradient bound, and $\mu_{\mathrm{PL}}, \mu_{\mathrm{QG}}$ are the PL and QG constants for Φ_D.

Consider a depth $(H+1)$ fully connected network of width m, initialized with i.i.d. Gaussian weights. Let $\lambda_0 > 0$ denote the minimum eigenvalue of the empirical tangent kernel at initialization on the n sample inputs, and let $\rho_\sigma > 0$ be the lower bound on the derivative of the last-layer activation used in Liu et al. [2022]. Choose the target PL constant μ_{PL} so that

$$0 < \mu_{\mathrm{PL}} < \rho_\sigma^2 \lambda_0.$$

Combining the over-parameterized neural-network PL result of Liu et al. [2022] with the BRM reduction of Kang et al. [2025], there are problem-dependent constants, with logarithmic factors hidden in $\widetilde{\Omega}(\cdot)$, such that, if

$$m = \widetilde{\Omega}\left(\frac{n\, R^{6H+2}}{\left(\lambda_0 - \mu_{\mathrm{PL}}/\rho_\sigma^2\right)^2} \right),$$

then with probability at least $1-\delta$ the value function Φ_D satisfies the PL inequality throughout the primal projection of Ω_R:

$$\frac{1}{2}\|\nabla \Phi_D(w)\|^2 \ge \mu_{\mathrm{PL}}\big(\Phi_D(w) - \Phi_D^\star\big).$$

Thus the constant appearing in Condition (C2) and in the contraction factor $c = \min\{\mu_{\mathrm{PL}}/2, \rho/2\}$ in Theorem 2 can be taken to be any admissible $\mu_{\mathrm{PL}} \in (0, \rho_\sigma^2\lambda_0)$ for which the displayed width condition is imposed.

The local geometric statements hold on the intersection of this PL event with the event that the SGDA iterates and the auxiliary ascent trajectories used in the analysis remain inside Ω_R; see Theorem 7 of Liu et al. [2022] for the corresponding radius/width regime for stochastic gradient methods. The optimization recursion uses stochastic-gradient moments in the original pre-update filtration, without conditioning on future confinement, and unconditional risk bounds require control of the complementary event. On this event, the softmax Bellman operator is smooth because $\eta > 0$. For the auxiliary network, the squared regression in (3) gives an inner PL condition when its residual Jacobian is uniformly nondegenerate, as follows.

Group repeated state–action inputs and let $r_{D,w}(v)$ collect the differences between ζ_v and the empirical conditional means of $V_{Q_w}(s')$, weighted by the square roots of their empirical frequencies. Suppose these conditional means are realizable and write $J_{D,v} := \nabla_v r_{D,w}(v)$. If $J_{D,v} J_{D,v}^{\mathsf{T}} \succeq \kappa_{\mathrm{aux}} I$ throughout the effective domain, then

$$\Gamma_D(w,v) = \beta^2 \|r_{D,w}(v)\|^2, \qquad \frac{1}{2}\|\nabla_v F_D(w,v)\|^2 = 2\beta^4 \|J_{D,v}^{\mathsf{T}} r_{D,w}(v)\|^2 \ge 2\beta^2 \kappa_{\mathrm{aux}} \Gamma_D(w,v).$$

Thus Condition (C4) holds with $\rho = 2\beta^2\kappa_{\mathrm{aux}}$. The lower bound is a condition on the auxiliary network's tangent kernel, alongside the primal kernel condition above [Liu et al., 2022, Kang et al., 2025]. It permits redundant parameters and does not require concavity in v. For the population objective, the same argument applies to the conditional-mean regression when the corresponding residual-space Jacobian lower bound and realizability hold.

The outer PL condition yields quadratic growth relative to $\arg\min_w \Phi_D(w)$, by the same argument as Lemma 6. The parameter-distance results use Condition (C3) as stated: on the effective domain, the primal minimizer is identified, so this set is $\{x_D^\star\}$ and the growth bound takes the form

$$\Phi_D(w) - \Phi_D^\star \ge \frac{\mu_{\mathrm{QG}}}{2}\|w - x_D^\star\|^2.$$

This identification condition is imposed separately from network width.

For a neighboring dataset $D^{(i)}$, the same argument is applied to $F_{D^{(i)}}$ and $\Phi_{D^{(i)}}$; applying a union bound and then taking the larger upper-bound constants for L, G and the smaller lower-bound constants for $\rho, \mu_{\mathrm{PL}}, \mu_{\mathrm{QG}}$ gives the uniform constants required by Condition (S3).

For the remaining conditions, joint smoothness follows under the regularity conditions in Lemma 29 of Kang et al. [2025], and the bounded gradient condition is discussed in Appendix C.1. The primal stability and risk guarantees allow multiple dual maximizers. The joint parameter bound in Theorem 2 additionally requires a singleton dual maximizer at the empirical primal optimum; this property is checked separately when that bound is used.

C.3 How index pairing works

By synchronizing the minibatch selection, we use the same algorithmic randomness in the two runs. The proof compares each output with its empirical primal minimizer and bounds the change in that minimizer after one observation is replaced. Fresh uniform minibatches give the conditional optimization recursion even when the offline observations are dependent. This conditional stability statement is distinct from the population-risk calculation, where dependence is measured by the fixed-point gradient covariance.

C.4 Uniformity setting

The setting (S3) is standard in the stability literature: for example, Hardt et al. [2016] assumed each per-example loss $f(\cdot; z)$ is L-Lipschitz and β-smooth uniformly in z, and Wang et al. [2022] assumed the gradients and smoothness of $f(w, v; z)$ are bounded by global constants G and L for all z. These conditions give common smoothness and gradient bounds across datasets; the PL, QG, and initial-potential constants must also be controlled as required by Setting (S3).

C.5 Discussion of contributions

What Kang et al. [2025] give us, and what they do not. Kang et al. [2025] establish that, after a bi-conjugate transformation, the population and empirical BRM objectives admit PL structure under their function-class conditions. They analyze global convergence to the empirical optimum under their auxiliary-block assumptions. However, they:

- do not analyze algorithmic stability,
- do not analyze generalization (population vs empirical BRM). In fact, Lemma 28 in the latest version of Kang et al. [2025] uses our result for the BRM sample complexity: we use their PL structure for BRM, and they use our sample-complexity result.

What Wang et al. [2022] give us, and what they do not. Wang et al. [2022] develop on-average argument stability for Markov chain SGD/SGDA and show that for smooth convex–concave losses, the weak primal–dual population gap in their Theorem 9 has rate $\widetilde{O}(n^{-1/2})$. Their Theorem 10 gives this rate for excess primal population risk with the additional assumption of population inner strong concavity, with explicit dependence on the mixing parameter of the Markov chain over indices. We summarize the difference between Wang et al. [2022] and our paper in Table 3. We reuse from Wang et al. [2022]:

- The concept of on-average argument stability and its connection to weak primal–dual generalization, used in Remark 3.3 and Lemma 15.
- The two-run coupling on neighboring datasets. Our proof then uses empirical-minimizer sensitivity and the PL potential.

However, there are two crucial structural differences:

- Geometry: Wang et al. [2022] use convex–concave geometry for weak primal–dual risk and additionally require strong concavity of the population inner objective for their excess primal-risk theorem. Our BRM objective is nonconvex in the primal parameters; only the value function $\Phi_D(w) := \max_v F_D(w, v)$ is PL (and QG), while the auxiliary maximization satisfies PL in v. The algorithm, however, is run on the bilevel saddle objective $F_D(w, v)$, not on Φ_D. The convexity-based coupled-distance argument used by Wang et al. [2022] therefore does not directly supply a stability bound for this BRM geometry.
- Object of interest: Wang et al. [2022] work directly with the original minimax F and directly analyze its risk/generalization. Our primary object is the value function Φ_D (Bellman

residual); the weak primal–dual gap is an additional consequence when joint stability holds. The primal gradients used by SGDA at time t can differ from $\nabla\Phi_D(w_t)$ when the dual variable is not at an inner maximizer:

$$\Delta_t := \nabla_w F_D\left(w_t, v_t\right) - \nabla\Phi_D\left(w_t\right)$$

Controlling this mismatch under the PL geometry is precisely what drives the need for our Lyapunov potential.

Table 3: Comparison with prior work. Our statistical rate uses independent observations or a controlled stationary covariance factor. Markov update indices and dependent training observations are distinct settings. Table 4 gives the detailed assumptions and metrics.

Work	**Geometry**	**Setting**	**Stat. Rate**	**Opt. Rate**
This paper	Primal/dual PL	Offline BRM	$\mathcal{O}(1/n)$	$\mathcal{O}(1/T)$
Wang et al. [2022]	Convex-concave	minimax	$\mathcal{O}(1/\sqrt{n})$	$\mathcal{O}(1/\sqrt{T})$

C.6 Discussion of potential extensions

On non-smooth BRM. Our analysis, which is based on *smooth* Bellman–residual objectives, does not lose any generality for the entropy-regularized formulation adopted in Section 2. In particular, when $\eta > 0$, the optimality equations involve the entropic log-sum-exp (softmax) operator

$$V_Q(s) = \eta \log \sum_{a \in \mathcal{A}} \exp(Q(s,a)/\eta),$$

so the Bellman operator and the resulting residual objective are differentiable and (under standard boundedness conditions on the function class) admit Lipschitz gradients. By contrast, genuinely non-smooth BRM variants (e.g., the hard $\max$ Bellman operator obtained in the zero-temperature limit $\eta \to 0$, or objectives with non-differentiable losses/parameterizations) require different tools, such as proximal or sub-gradient dynamics, smoothing/Moreau-envelope arguments, and stability analyses tailored to non-smooth saddle problems, and are therefore orthogonal to the paper's goal.

From Bellman residuals to policy performance. Our main theorem is stated for the Bellman Residual Minimization (BRM) objective. In offline RL, such a result is often converted to a control-performance guarantee that is standard once one imposes a coverage (a.k.a. concentrability) assumption that relates the state–action occupancy of the target/learned policy to the data distribution. Under such coverage, a bound on the population Bellman residual under the dataset distribution immediately yields a bound on the Bellman error along the policy's visitation distribution, which then translates to policy suboptimality via standard approximate dynamic programming arguments. In fact, Kang et al. [2025] uses our result to instantiate this pipeline for offline inverse reinforcement learning and related entropy-regularized control problems.

D Related Works

Bellman Residual Minimization (BRM). Bellman Residual Minimization (BRM) can be viewed as a regression-style approach to offline reinforcement learning, in which the squared Bellman residual plays the role of a regression loss. It has been identified as a particularly direct way to enforce Bellman consistency under flexible function approximation, including nonparametric or neural-network parameterizations [Jiang and Xie, 2025]. A central obstacle, however, is the double-sampling problem [Antos et al., 2008]: the squared Bellman residual involves the square of a conditional expectation, which cannot be unbiasedly estimated from a single next-state sample. An influential remedy is to introduce a dual, or debiasing, correction through a bi-conjugate reformulation [Antos et al., 2008, Dai et al., 2018, Patterson et al., 2022]. While this removes the statistical bias caused by double sampling, it transforms BRM into a minimax optimization problem, whose primal component is generally nonconvex under nonlinear function approximation. Consequently, BRM-style methods have long been viewed as computationally delicate in offline RL, despite their conceptual appeal.

Table 4: Detailed comparison of the statistical bounds in Table 1. The notes below the table explain the risk measures, the rates, and the data assumptions.

Work	BRM	Class	Estimator	Data	Assumptions	Risk	Rate
Maillard et al. [2010]	Yes	linear	Oracle for empirical BRM	Generative model, two next states per query	–	Excess $\overline{\mathcal{L}_{\mathrm{BE}}}$ for $\mathcal{T}^{\pi}$ (Thm. 7)	$O(n^{-1/2})$
Dai et al. [2018]	Yes	general	Oracle for empirical BRM	Single β-mixing path	–	Excess $\mathcal{R}$-type objective (Thm. 6)	$\widetilde{O}(n^{-1/2})$
Chen and Jiang [2019]	Yes	finite	Oracle for empirical BRM	Independent transitions	Realizability, completeness	$\overline{\mathcal{L}_{\mathrm{BE}}}(\widehat{Q})$ (proof of Thm. 3)	$\widetilde{O}(n^{-1})$
Duan et al. [2021]	Yes	general	Oracle for empirical BRM or FQI	Independent transitions	Completeness	Excess $\overline{\mathcal{L}_{\mathrm{BE}}}$	$O(n^{-1/2})$
Hu et al. [2025]	Yes	linear	Oracle for empirical BRM	Independent transitions	Realizability, modified completeness, feature coverage	$\overline{\mathcal{L}_{\mathrm{BE}}}(\widehat{Q})$ via $\|\widehat{Q} - Q^{\star}\|_{\infty}$	$\widetilde{O}(n^{-1})$
Wang et al. [2022]	No	general	Averaged MC-SGDA	Independent observations, Markov update indices	Convex–concave (Thm. 10: population inner strong concavity)	Δ^{PD} (Thm. 9); $R(w) - R(w^{\star})$ (Thm. 10)	$\widetilde{O}(n^{-1/2})$
Ours	Yes	general	Last iterate of SGDA, $T \asymp n$	Independent transitions or one stationary path	Primal and inner PL, QG, (C1)–(C6)	$\mathcal{R}(\widehat{w}^{(T)}) - \mathcal{R}(w^{\star})$	$O(n^{-1})$

Risk. $\overline{\mathcal{L}_{\mathrm{BE}}}(\widehat{Q})$ is the MSBE of Section 2 at the learned $\widehat{Q}$, with the soft operator $\mathcal{T}$ replaced by the hard-max operator for Chen and Jiang [2019] and Hu et al. [2025] and by the policy-evaluation operator $\mathcal{T}^{\pi}$ for Maillard et al. [2010]. "Excess" subtracts the best value in the hypothesis class, $\min_{Q \in \mathcal{F}} \overline{\mathcal{L}_{\mathrm{BE}}}(Q)$, up to approximation terms. The objective of Dai et al. [2018] is a smoothed bi-conjugate Bellman error, the analogue of $\mathcal{R}$ in Theorem 4. Δ^{PD} is the weak primal–dual gap (Definition 3) and R the primal risk (Definition 2); for our BRM objective, $\mathcal{R}(\widehat{w}^{(T)}) - \mathcal{R}(w^{\star})$ equals the excess MSBE under auxiliary-function realizability.

Rates. Rates show the dependence on n for fixed problem parameters, and $\widetilde{O}$ hides logarithmic factors. For Chen and Jiang [2019] and Hu et al. [2025], the headline guarantees concern the value gap or regret, and the table reports the squared-loss bound behind them. For Hu et al. [2025], Bellman contraction and realizability give $\|Q - \mathcal{T}Q\|_{\infty} \le (1+\beta)\|Q - Q^{\star}\|_{\infty}$, so their sup-norm bound controls $\overline{\mathcal{L}_{\mathrm{BE}}}$; their regret result additionally uses a margin condition. The $O(n^{-1/2})$ rate of Duan et al. [2021] holds for general classes, and their local Rademacher analysis gives $O(n^{-1})$ for finite or linear classes when the approximation terms vanish. Our bound is $O(1/n + 1/T)$ (Theorem 5), so $T \asymp n$ iterations give $O(n^{-1})$; a uniform auxiliary approximation error ϵ_{aux} adds $\beta^2 \epsilon_{\mathrm{aux}}$, as in (4).

Data. From the strongest to the weakest requirement: a generative model that returns two independent next states for the same query [Maillard et al., 2010] removes the double-sampling problem by assumption and cannot be built from logged data. Independent transitions, each drawn from a fixed state–action distribution followed by P, are the standard idealization of an offline dataset [Chen and Jiang, 2019, Duan et al., 2021, Hu et al., 2025, Wang et al., 2022]; in Wang et al. [2022] the Markov chain drives the update indices, while the observations themselves are independent. A single stationary trajectory is what a behavior policy produces in practice. Dai et al. [2018] assume exponential β-mixing for it, whereas our population bounds need only summable gradient covariances at the population optimum ($C_{\mathrm{cov}} < \infty$, Appendix B.2), which exponential β-mixing implies for bounded gradients; our stability bound (Theorem 2) needs no condition on the observations.

Recently, Kang et al. [2025] showed that, for entropy-regularized BRM and under neural-network approximation satisfying suitable over-parameterization conditions, the bi-conjugate objective admits PL geometry and gradient-based minimax optimization converges globally under their auxiliary-block conditions. However, their result is primarily an optimization guarantee: it does not directly establish algorithmic stability or a statistical generalization bound comparing the empirical BRM objective with its population counterpart. This statistical question is the main focus of the present work.

Algorithmic stability and generalization. Algorithmic stability is a classical approach to generalization analysis that controls how sensitively the output of a learning algorithm responds to small perturbations of the training dataset [Bousquet and Elisseeff, 2002]. This framework has been developed through several notions of stability, including uniform stability, hypothesis stability, argument stability, and on-average stability [Bousquet and Elisseeff, 2002, Hardt et al., 2016, Kuzborskij and Lampert, 2018, Lei and Ying, 2020]. For gradient-based learning, Hardt et al. [2016] established influential stability bounds for stochastic gradient descent, and subsequent works refined these bounds using data-dependent or on-average notions of stability [Kuzborskij and Lampert, 2018, Lei and Ying, 2020, Lei et al., 2021]. Stability has also been used to study algorithms that converge to global optima [Charles and Papailiopoulos, 2018] and to obtain refined high-probability generalization guarantees for uniformly stable algorithms [Feldman and Vondrak, 2018, Bousquet et al., 2020].

More recently, stability-based generalization has been extended from minimization to minimax optimization. This direction is particularly relevant for BRM because the bi-conjugate correction turns the squared Bellman-residual objective into a saddle-point problem. Prior works studied the stability and generalization of SGDA under various geometries, with recent extensions to decentralized, distributed, and differentially private SGDA variants [Farnia and Ozdaglar, 2021, Lei et al., 2021, Zhu et al., 2023, Zhang et al., 2024]. In parallel, Wang et al. [2022] developed an on-average argument-stability framework for stochastic gradient methods with Markov-chain sampling and established stability-to-generalization guarantees for both minimization and minimax problems. Their Markov chain samples update indices from an independent training dataset; dependence among trajectory observations requires a separate statistical argument. However, these existing minimax stability results do not directly apply to the BRM problem studied here. The key difficulty is geometric: the BRM saddle objective is not convex in the primal parameter, and the algorithm is run on $F_D(w, v)$ rather than directly on the primal value function $\Phi_D(w) := \max_v F_D(w, v)$. Thus the primal update contains a mismatch term

$$\nabla_w F_D(w_t, v_t) - \nabla \Phi_D(w_t),$$

unless the dual variable is already maximized. Our analysis addresses this mismatch through a Lyapunov potential that couples the PL suboptimality of Φ_D with the dual gap, yielding an $O(1/n)$ primal argument-stability term and, through the same geometry, the BRM generalization guarantee.

PL geometry and over-parameterized neural networks. The Polyak–Łojasiewicz (PL) condition is a central relaxation of strong convexity that still yields global convergence of gradient-based methods despite nonconvexity [Polyak, 1963, Karimi et al., 2016]. It has become particularly important in the analysis of over-parameterized models, where classical convexity-based arguments are often inappropriate. In modern neural-network theory, over-parameterized loss landscapes have been studied through several complementary perspectives, including the absence of strict spurious local minima, connectivity of sublevel sets or global-minimum manifolds, neural tangent kernel (NTK) dynamics, and local or global PL-type conditions [Du et al., 2019, Jacot et al., 2018, Lee et al., 2019, Allen-Zhu et al., 2019, Arora et al., 2019, Soltanolkotabi et al., 2019, Liu et al., 2022, Chen et al., 2023, Xu et al., 2025]. In particular, Liu et al. [2022] argue that sufficiently over-parameterized nonlinear systems, including wide neural networks, satisfy a PL*-type condition in a neighborhood of the initialization, and relate this property to tangent-kernel conditioning and Hessian norm control. This line of work explains why gradient-based methods can converge globally in highly nonconvex models without relying on convexity.

Our paper uses this perspective in the BRM setting: the relevant population and empirical soft Bellman-residual objectives are analyzed under primal and dual PL conditions, with the additional QG and domain requirements stated explicitly; see Kang et al. [2025] for the BRM geometry. We then complement this optimization geometry with an algorithmic-stability analysis to obtain finite-sample statistical guarantees.

Projected Mean-Squared Bellman Error (MSPBE) Another approach is to minimize the projected mean-squared Bellman error (MSPBE), which first projects the Bellman residual onto the approximation space and then measures its squared norm [Patterson et al., 2022]. This differs from BRM, which directly minimizes the mean-squared Bellman error (MSBE) and therefore penalizes the full Bellman inconsistency. MSPBE instead evaluates only the component of the Bellman residual that lies within the representable function space. From an optimization perspective, this projection can make the objective more tractable, since it only requires the residual to be small along directions captured by the function class. However, this advantage comes at a cost: MSPBE ignores the component of the Bellman residual orthogonal to the approximation space, and therefore does not control the full Bellman inconsistency.

Fitted Q-Iteration (FQI) Fitted Q-Iteration (FQI), introduced by Ernst et al. [2005], is a canonical value-based method for offline reinforcement learning. Although conceptually simple, iterative value-based procedures are known to suffer from instability. In particular, approximate value iteration / fitted value iteration can be unstable under off-policy sampling and function approximation, including the case of infinite data and exact regression with linear function approximation. This instability arises because FQI repeatedly fits value functions to targets that themselves depend on the current estimate. Thus, the regression target evolves over iterations, creating an inherently unstable learning process. More broadly, this phenomenon reflects the well-known deadly triad: the combination of function approximation, bootstrapping, and off-policy learning [Tsitsiklis and Van Roy, 1996, Van Hasselt et al., 2018, Sutton and Barto, 2018].

Marginalized Importance Sampling (MIS) Marginalized importance sampling (MIS) methods cast offline reinforcement learning as a minimax problem by learning marginalized density ratios, often interpreted as discriminators, to control worst-case reweighted Bellman errors [Nachum et al., 2019, Zhang et al., 2020, Liu et al., 2018]. Unlike BRM, these methods optimize linear Bellman-error objectives rather than squared residuals, and therefore avoid the double-sampling problem. However, MIS objectives typically involve signed average Bellman errors. As a result, positive and negative residuals can cancel, so these objectives may fail to penalize pointwise Bellman inconsistency [Nachum et al., 2019, Zhang et al., 2020]. This distinction motivates our focus on BRM as a direct residual-minimization approach, while addressing its statistical and optimization challenges through the bi-conjugate formulation and stability analysis.

E Details of the numerical illustration

This appendix gives the full setting and additional results for Section 4. Code and raw results are available at `https://github.com/jajajang/soft-brm-stability`.

Setup. The MDP has $|\mathcal{S}| = 5$, $|\mathcal{A}| = 3$, $\beta = 0.9$ and $\eta = 1$. Rewards are drawn from $\mathrm{Unif}[0, 1]$, and $P(\cdot \mid s, a) = \frac{1}{2}\delta_s + \frac{1}{2}p_{s,a}$ with $p_{s,a} \sim \mathrm{Dirichlet}(\mathbf{1})$ (fixed seed). The behavior policy is uniform, and μ denotes its stationary state–action distribution (smallest mass 0.046). We use two data regimes: (i) n i.i.d. transitions with $(s, a) \sim \mu$ and $s' \sim P(\cdot \mid s, a)$; (ii) a single stationary trajectory of length n under the behavior policy, whose state chain has second eigenvalue of modulus 0.59, so consecutive transitions are dependent. The neighbor $\mathcal{D}^{(i)}$ replaces a uniformly chosen z_i by an independent copy $\tilde{z}_i \sim \mu \otimes P$.

Q_w maps a one-hot state to $|\mathcal{A}|$ outputs through H softplus units, and ζ_v maps a one-hot state–action pair to a scalar through H softplus units, with $H \in \{8, 16, 32\}$. Both use the NTK parameterization (output layer scaled by $H^{-1/2}$) with standard Gaussian initialization, and the output biases start at the constant value scale. Softplus keeps $F_{\mathcal{D}}$ smooth, as required by (C1). We run Algorithm 1 with minibatch size $B = 32$, stepsizes $\eta_t = 400/(2000 + t)$, dual multiplier $\kappa_v = 2$ and $T = 40{,}000$ iterations, from the same initialization and with the same minibatch index sequence on $\mathcal{D}$ and $\mathcal{D}^{(i)}$. Each (n, H) pair uses 50 repetitions, each with a fresh dataset, initialization, index sequence and replaced index, for $n \in \{100, 200, \ldots, 6400\}$. All runs use NumPy on CPU, and the full study (both data regimes and the constant estimates below) takes about one hour on two CPU cores.

We report three quantities: the primal argument stability ε_T^w of Theorem 2; the function-space distance $\|Q_{w_T} - Q_{w'_T}\|_\mu$; and the excess population MSBE of the output on $\mathcal{D}$, computed exactly from P. With one-hot inputs both networks can represent any function of their inputs, so $Q^\star$ is representable

Table 5: n times the primal argument stability ε_T^w (mean $\pm$ standard error over 50 repetitions, $T = 40{,}000$). A constant column corresponds to an n^{-1} rate. The last two rows give the fitted slope of $\log \varepsilon_T^w$ against $\log n$ and its 95% bootstrap interval.

	i.i.d. data			single trajectory
n	$H = 8$	$H = 16$	$H = 32$	$H = 16$
100	16.54 ± 2.93	8.95 ± 1.38	5.93 ± 0.65	9.93 ± 1.20
200	11.88 ± 1.33	7.74 ± 0.57	7.18 ± 0.61	7.70 ± 1.06
400	13.13 ± 2.93	8.38 ± 0.78	7.39 ± 0.65	6.77 ± 0.55
800	8.45 ± 0.56	6.72 ± 0.59	6.27 ± 0.44	6.58 ± 0.68
1600	8.84 ± 0.73	6.95 ± 0.62	5.59 ± 0.55	7.20 ± 0.53
3200	10.32 ± 1.14	7.65 ± 0.70	5.75 ± 0.48	7.40 ± 0.85
6400	9.03 ± 0.77	6.33 ± 0.66	6.05 ± 0.54	6.86 ± 0.74
slope ($n \geq 200$)	-1.08	-1.05	-1.07	-1.01
95% interval	$[-1.16, -1.00]$	$[-1.11, -0.99]$	$[-1.13, -1.01]$	$[-1.09, -0.93]$

and the auxiliary class contains the conditional means; so $\mathcal{R}(w^\star) = 0$ and the excess MSBE equals the left-hand side of Theorem 5. As a reference we also compute the population MSBE of the exact empirical minimizer, the fixed point of the empirical soft Bellman operator. Slopes are least-squares fits of $\log(\text{mean})$ on $\log n$ over $n \geq 200$, with 95% bootstrap intervals over repetitions.

Results. Figure 1 in Section 4 and Tables 5 and 6 report the results. Both quantities decay as n^{-1}: the fitted slopes are between -1.08 and -1.05 for ε_T^w and between -1.01 and -0.97 for the excess MSBE, every 95% interval excludes $-1/2$, and $n\,\varepsilon_T^w$ and $n \cdot \mathrm{MSBE}$ stay roughly constant over a 32-fold range of n. The same holds for a single dependent trajectory (slopes -1.01 and -0.99), in line with the fact that the stability bound needs no independence of the observations. The generalization gap $\mathcal{R}(w_T) - \widehat{\mathcal{R}}_n(w_T)$ of Theorem 4 behaves in the same way (n times the gap lies in $[0.34, 0.45]$ for $n \geq 200$). No width shows saturation; the narrowest network has a larger constant but the same rate.

The optimization error is small at $T = 40{,}000$: ε_T^w changes by at most 15% between $T = 20{,}000$ and $T = 40{,}000$, and the excess MSBE of the SGDA output is within 10% of that of the exact empirical minimizer for $200 \leq n \leq 3200$ (within 33% at $n = 6400$, where residual optimization error becomes visible). At $n = 100$ some datasets miss a state–action pair, which inflates the error; this point is excluded from the fits.

Magnitude of the constants. For $H = 16$ and $n = 800$ we estimated the constants along five SGDA trajectories on their own datasets: μ_{PL} and ρ as the smallest observed ratios in (C2) and (C4), μ_{QG} from (C3) relative to the final iterate of a run four times longer (a proxy for the empirical minimizer), G as the largest per-sample gradient norm, and L by power iteration on the joint Hessian of $F_{\mathcal{D}}$. Over the whole trajectories the worst values are $\mu_{\mathrm{PL}} = 0.018$, $\rho = 0.014$, $\mu_{\mathrm{QG}} = 0.012$, $G = 2.5$ and $L = 4.2$. With these, the constant S_w of Lemma 11, which multiplies $1/n$ in the primal stability bound, is about 1.0×10^5, against an observed $n\,\varepsilon_T^w \approx 7.3$, and the statistical constant of Theorem 5 is about 3.0×10^{12}, against an observed $n \cdot \mathrm{MSBE} \approx 0.41$: with uniform constants the bounds are vacuous at this scale. The worst values come from early iterates, where the dual variable is far from its maximizer. Using the median constants after $t = 5000$ ($\mu_{\mathrm{PL}} \approx 0.95$, $\rho \approx 2.2$, $\mu_{\mathrm{QG}} \approx 0.11$) gives $S_w \approx 25$, within a factor of about 3 of the observed value, while the constant of Theorem 5 (about 1.7×10^3) remains loose because of its $(1 + L/\rho)^2 L_\Phi/(\mu_{\mathrm{PL}}\mu_{\mathrm{QG}})$ dependence. The experiment therefore supports the n^{-1} rate, not the size of the constants.

Table 6: n times the excess population MSBE of the SGDA output (mean $\pm$ standard error), and of the exact empirical minimizer on the datasets of the i.i.d. $H = 16$ runs.

	i.i.d. data			single trajectory	exact empirical
n	$H = 8$	$H = 16$	$H = 32$	$H = 16$	minimizer (i.i.d.)
100	0.480 ± 0.035	0.834 ± 0.127	0.508 ± 0.036	0.819 ± 0.274	0.672
200	0.457 ± 0.031	0.440 ± 0.039	0.448 ± 0.025	0.461 ± 0.038	0.436
400	0.447 ± 0.037	0.393 ± 0.020	0.446 ± 0.025	0.411 ± 0.026	0.391
800	0.420 ± 0.026	0.412 ± 0.022	0.391 ± 0.021	0.430 ± 0.029	0.404
1600	0.433 ± 0.029	0.385 ± 0.020	0.383 ± 0.020	0.375 ± 0.022	0.358
3200	0.464 ± 0.029	0.396 ± 0.024	0.414 ± 0.025	0.429 ± 0.029	0.361
6400	0.527 ± 0.033	0.443 ± 0.026	0.455 ± 0.022	0.486 ± 0.023	0.350
slope ($n \geq 200$)	-0.97	-1.00	-1.01	-0.99	-1.06
95% interval	$[-1.01, -0.92]$	$[-1.05, -0.95]$	$[-1.04, -0.97]$	$[-1.03, -0.94]$	$[-1.11, -1.02]$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main claims: an $O(1/n)$ on-average primal argument-stability bound for SGDA (Theorem 2) and $O(1/n + 1/T)$ generalization and excess-risk bounds for soft BRM (Theorems 4 and 5), under the primal and inner PL, QG and domain conditions of Section 3. Section 4 illustrates the rates numerically.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 (Limitations and future work) discusses the width requirement, the conditions that are hardest to verify, the size of the constants, and the restriction to smooth entropy-regularized BRM; Appendix C.6 discusses non-smooth BRM and policy-performance guarantees.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions are Conditions (C1)–(C6) and Settings (S1)–(S3). The results are Lemma 1, Theorem 2 (with a proof sketch in Section 3), Lemma 3, Theorems 4 and 5, and Proposition B.1; complete proofs are in Appendix B.1–B.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 and Appendix E specify the MDP, the two data regimes, the networks and their initialization, the SGDA hyperparameters, the number of repetitions, and how slopes and intervals are computed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often

one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is available at `https://github.com/jajajang/soft-brm-stability`; the data are synthetic and generated by the released code (Appendix E).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix E lists all settings: MDP, data regimes, network widths and parameterization, minibatch size, stepsizes, dual multiplier, number of iterations, and sample sizes.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tables 5 and 6 report means with standard errors over 50 repetitions, and every fitted slope comes with a 95% bootstrap interval over repetitions (Appendix E).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E states that all runs use NumPy on CPU and that the full study takes about one hour on two CPU cores.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research is theoretical and does not involve human subjects, sensitive data, deployed systems, or data/model releases with misuse risk. To the best of our knowledge, it conforms to the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The paper provides foundational theoretical guarantees for soft BRM and does not propose a deployed system, collect data, release a model, or target a specific application domain. Any societal impact would be indirect and mediated by future applications of offline reinforcement learning.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release data, models, or systems that pose a high risk of misuse. Therefore, safeguards for responsible release are not applicable.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The experiments use a synthetic MDP and standard open-source numerical libraries (NumPy, Matplotlib); no existing datasets, models or third-party code are used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The experiment code is released at `https://github.com/jajajang/soft-brm-stability` with a README that describes the files and the commands that reproduce every result.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, surveys, or any research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects research. Therefore, IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method, theoretical results, and proofs do not use LLMs as an important, original, or non-standard component. Any use of LLMs, if any, was limited to writing, editing, or formatting assistance and did not affect the scientific methodology, rigor, or originality of the work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.