Statistical Complexity of Soft Bellman Residual Minimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Bellman residual minimization (BRM) provides a scalable, gradient-based approach to offline reinforcement learning in large state spaces. 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 novel Lyapunov-based analysis, we establish an $\mathcal{O}(1/n)$ average argument stability bound, which translates directly into a $\mathcal{O}(1/n)$ statistical complexity for the soft BRM objective.

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, online interaction is unethical, dangerous, or prohibitively expensive [Jiang and Xie, 2024]. 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 [2024], 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 as neural networks[2], the mean-squared Bellman error (MSBE) induces a Polyak–Łojasiewicz (PL)–strongly-concave minimax landscape[3]. Consequently, plain stochastic gradient descent–ascent

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

[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. In a PL–strongly-concave minimax problem, the inner maximization is strongly concave and the outer minimization satisfies the PL condition.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

(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 we need for BRM to recover a near-optimal value function, under SGDA?

Contributions. In this paper, we close this statistical gap for popular function classes such as neural networks and linear functions. Building on the PL structure identified by Kang et al. [2025], we develop a Lyapunov potential tailored to PL–strongly-concave optimization and blend it with a modern on-average argument-stability analysis. Our main theorem shows that, for a dataset of size n, SGDA attains an $\mathcal{O}\big(1/n\big)$ excess MSBE loss, a rate that doubles the exponent enjoyed in convex–concave optimization with Markov-sampled data, where the best known rate is $\mathcal{O}\big(1/\sqrt{n}\big)$ [Wang et al., 2022]. In particular:

1. We prove $\mathcal{O}(1/n)$ on-average argument-stability bound for SGDA under PL–strong concavity, avoiding any independence assumptions on the minibatch sampling indices.
2. Leveraging this stability, we derive the $\mathcal{O}(1/n)$ generalization guarantee for BRM.
3. Our analysis is constructive, requires no variance reduction or extra regularisation, and applies verbatim to standard neural-network parameterisations commonly used in offline RL.

At a technical level, 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 the stochastic perturbations enter only through summable lower-order terms of order η_t^2 and η_t/n. Under Robbins–Monro stepsizes, the accumulated contraction dominates these perturbations, yielding the desired $O(1/n)$ stability rate.

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 $P(\cdot \,|\, s, a)$ 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 $\pi(\cdot \,|\, s)$, a distribution over actions $\mathcal{A}$; when the agent is in state s_h at time h it samples $a_h \sim \pi(\cdot \,|\, s_h)$. 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

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

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

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

[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.

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):

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

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

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

The normalized formulation recovered by setting $\eta = 1$ coincides with the mean-zero Gumbel-based softmax Bellman equations used in the econometrics literature [Rust, 1994, Kang et al., 2025].

2.1 Bellman Residual Minimization

Bellman Error (Bellman Residual) and Temporal Difference Error. Define the function space $\mathcal{Q}$ as the set of all bounded real-valued functions on the state-action space $\mathcal{S} \times \mathcal{A}$:

$$\mathcal{Q} := \{ Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R} \mid \|Q\|_\infty < \infty \}$$

Under standard discounted-MDP assumption, the optimal action-value function, Q^*, is an element of this space, i.e., $Q^* \in \mathcal{Q}$, provided that the discount factor β is in (0,1).

We introduce the *(soft) Bellman optimality operator*, $\mathcal{T} : \mathcal{Q} \to \mathcal{Q}$, defined for any $Q \in \mathcal{Q}$ by:

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

Note that Q^* is the unique fixed point of $\mathcal{T}$ [Rust, 1994]; Q^* is the unique solution to the Bellman optimality equation:

$$\mathcal{T}Q^* = Q^* \quad \text{or equivalently,} \quad (\mathcal{T}Q^*)(s,a) - Q^*(s,a) = 0.$$

The extent to which an arbitrary Q-function Q fails to satisfy the Bellman optimality equation motivates the following definitions of error.

Definition 1 (Bellman Error (Bellman Residual))**.** *For any function $Q \in \mathcal{Q}$, we define the Bellman error (or Bellman residual) at a state-action pair (s,a) as the difference:*

$$(\mathcal{T}Q)(s,a) - Q(s,a)$$

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*, $\hat{\mathcal{T}}$. Given a single transition tuple (s,a,s'), this operator is defined as:

$$\hat{\mathcal{T}}Q(s,a,s') := r(s,a) + \beta\eta \log \sum_{a' \in \mathcal{A}} \exp\left(Q\left(s',a'\right)/\eta\right)$$

Definition 2 (Temporal-Difference Error)**.** *Using the sampled operator, we can define the Temporal-Difference (TD) error for a given transition (s,a,s'):*

$$\delta_Q(s,a,s') := \hat{\mathcal{T}}Q(s,a,s') - Q(s,a)$$

The connection between the Bellman error and the TD error is established in the following lemma. It shows that the TD error is an unbiased, single-sample estimate of the Bellman error.

Lemma 1 (Relationship between Bellman and TD Errors)**.** *For any $Q \in \mathcal{Q}$ and any state-action pair (s,a), the expectation of the Sampled Bellman operator over the next state s' recovers the original Bellman operator:*

$$\mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\hat{\mathcal{T}}Q(s,a,s') \right] = (\mathcal{T}Q)(s,a)$$

Consequently, the expected TD error is equal to the Bellman error:

$$\mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\delta_Q(s,a,s') \right] = (\mathcal{T}Q)(s,a) - Q(s,a)$$

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 first 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. Furthermore, as a proxy for Squared Bellman Error, we define the *Squared TD Error*: $\mathcal{L}_{\text{TD}}(Q)(s, a, s') := \delta_Q(s, a, s')^2$ and minimize the *Mean Squared TD Error* (MSTDE) as a proxy for MSBE, defined as:

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

Unfortunately, MSTDE is a *biased* proxy for MSBE. This bias happens because $\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\delta_Q\left(s, a, s'\right) \mid s, a\right]^2 \neq \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\delta_Q\left(s, a, s'\right)^2 \mid s, a\right]$ i.e., expectation and square are not exchangeable. This issue is often called the *double sampling problem* [Antos et al., 2008]. Specifically, one can show that

$$\begin{aligned}\mathcal{L}_{BE}(Q)(s, a) &= \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\mathcal{L}_{TD}(Q)\left(s, a, s'\right)\right] - \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\left(\mathcal{T}Q(s, a) - \hat{\mathcal{T}}Q\left(s, a, s'\right)\right)^2\right] \\ &= \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\mathcal{L}_{TD}(Q)\left(s, a, s'\right)\right] - \beta^2\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\left(V_Q(s') - \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_Q(s')]\right)^2\right].\end{aligned}$$

(For the detailed derivation, see [Kang et al., 2025, Appendix C.1].) Since the bias term includes the $\mathbb{E}_{s'\sim P(\cdot|s,a)}$ part, correcting this bias term again remains challenging without full knowledge of the system's transition dynamics, P. To resolve this issue, we employ an approach often referred to as the "Bi-Conjugate Trick" [Antos et al., 2008, Dai et al., 2018, Patterson et al., 2022]:

$$\mathcal{L}_{BE}(Q)(s, a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\delta_Q\left(s, a, s'\right) \mid s, a\right]^2 = \max_{h\in\mathbb{R}} 2\cdot\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\delta_Q\left(s, a, s'\right) \mid s, a\right]\cdot h - h^2$$

According to [Kang et al., 2025, Appendix C.1], this bi-conjugate form can be re-parametrized using $\zeta := h - r(s, a) + Q(s, a)$ as:

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

Now define the per-sample objective

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

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 the following min-max problem:

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

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

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

the resulting parametrized problem becomes

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

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

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

A canonical way of solving the mini-max problem is to apply the Stochastic Gradient Ascent Descent (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$^{\text{depth}}$), and therefore the SGDA algorithm finds the (w, v) that are global minima of Equation (5). (For the detailed discussion on neural networks, see Appendix C.2.)

2.2 Stochastic Gradient Ascent–Descent Algorithm (SGDA)

As discussed earlier, Stochastic Gradient Ascent–Descent (SGDA) is the workhorse we use to solve the minimax problem (5). Given a function $f(w, v)$, at every iteration it performs a *descent* step on the primal variable w and an *ascent* step on the dual variable v using (possibly noisy) gradients computed from a minibatch of samples.

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 set $I_t \subseteq [n], |I_t| = B$, 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

$$g_t^w = \nabla_w F_{I_t}(w_t, v_t),\ g_t^v = \nabla_v F_{I_t}(w_t, v_t) \text{ where } F_I(w, v) := \frac{1}{|I|} \sum_{j \in I_t} f(w, v; z_j) \text{ for any } I \subset [n].$$

By a slight abuse of notation, when the subscript corresponds to the full dataset $\mathcal{D}$, we write $F_{\mathcal{D}}(w, v) := \frac{1}{n} \sum_{z \in \mathcal{D}} f(w, v; z)$.

Unbiasedness is preserved: $\mathbb{E}[g_t^w] = \nabla_w F_{\mathcal{D}}(w_t, v_t)$ and likewise for v, while the variance contracts by the usual $1/B$ factor. Using stepsize sequence $(\eta_t)_{t \geq 0}$, SGDA proceeds as

$$w_{t+1} = w_t - \eta_t\, g_t^w, \qquad v_{t+1} = v_t + \eta_t\, g_t^v.$$

The recursion can be written as follows, which is the form used in the stability proofs of Section 3.

$$(w_{t+1}, v_{t+1}) = (w_t, v_t) + \eta_t \left(-g_t^w,\, g_t^v\right),$$

Algorithm 1 Minibatch SGDA on the empirical objective (5)

Input: Dataset $\mathcal{D} = \{z_i\}_{i=1}^n$, minibatch size B, stepsizes (η_t), initial (w_0, v_0)
for $t = 0$ **to** $T - 1$ **do**
  Draw $I_t \subseteq [n]$ with $|I_t| = B$ uniformly at random
  $g_t^w \leftarrow \frac{1}{B} \sum_{i \in I_t} \nabla_w f(w_t, v_t; z_i)$
  $g_t^v \leftarrow \frac{1}{B} \sum_{i \in I_t} \nabla_v f(w_t, v_t; z_i)$
  $w_{t+1} \leftarrow w_t - \eta_t g_t^w$ {gradient *descent*}
  $v_{t+1} \leftarrow v_t + \eta_t g_t^v$ {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 2 (Global convergence of minibatch SGDA in parameter space [Yang et al., 2020, Kang et al., 2025])**.** *Let the iterates in* ALGORITHM 1 *be written as* $\{(w_t, v_t)\}_{t \geq 0}$*, where* w_t *parametrises the action–value function* Q_{w_t} *(primal variable) and* v_t *parametrises* ζ_{v_t} *(dual variable). Choose the harmonic step-sizes* $\eta_t = \frac{c_1}{c_2 + t}, t \geq 0$ *with some constants* $c_1 > 0$ *and* $c_2 \geq 1$ *such that* $\eta_t \leq \min\{1/(4L),\, 1/\rho\}$ *for every* t*. Then the* ALGORITHM 1*'s output sequence* $\{(w_t, v_t)\}$ *converges almost surely to a saddle point of the empirical objective* (5)*, where the suboptimality of empirical objective* (5) *is bounded by* $\frac{d_1}{d_2 + t}$ *for some constants* d_1 *and* d_2*.*

Lemma 2 clarifies the optimization side of empirical BRM: under the PL–strongly-concave landscape, SGDA globally converges to a saddle point of the empirical objective. What remains is the generalization side. The lemma does not address how errors induced by the finite offline dataset propagate to the population BRM objective, or whether the value function learned from empirical data generalizes to the underlying MDP distribution.

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. The key fact is that stability implies generalization: algorithms that are stable on neighbouring 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.

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 PL–strong-concave saddle-point problem. 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 $D = \{z_i\}_{i=1}^n$. Working in this slightly more abstract template keeps the proofs uncluttered; in Theorem 6 we then specialize the resulting stability and generalization bounds back to BRM.

In this section, for the sake of generality, we use notations that generalize the Bellman residual minimization problem. 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 neighbour* 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 $\mathcal{D}^{(i)}$ *neighbouring 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 $i \sim \mathrm{Unif}([n])$. This coupling is the key mechanism that allows the analysis to proceed without an i.i.d. assumption on the data. To see how it works, please check Appendix C.3.

Throughout, we present all proofs for the *single-sample* ("minibatch-of-one") variant of SGDA. The extension to a minibatch of size $B \geq 1$ (or the full-batch case. $B = n$) is mechanical. Every stochastic-gradient term is replaced by its averaged counterpart, which reduces all variance contributions by a factor $1/B$, while the probability that a particular data point appears in the update increases from $1/n$ to B/n. Consequently, every lemma and theorem below remains valid verbatim—with constants rescaled by these factors—and no new conceptual issues arise.

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. In this case, as in Wang et al. [2022], the concept of *on-average algorithmic stability* is useful.

Definition 3 (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 is over the algorithm's internal randomness and the choice of i.*

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–strongly concave saddle problems, following the stability framework for minimax optimization.

Let SGDA iterates start from the same initialization $(w_0, v_0) = (w_0', v_0')$:

$$w_{t+1} := w_t - \eta_t \nabla_w f(w_t, v_t; z_{i_t}), \qquad w'_{t+1} := w'_t - \eta_t \nabla_w f(w'_t, v'_t; z'_{i_t}),$$
$$v_{t+1} := v_t + \eta_t \nabla_v f(w_t, v_t; z_{i_t}), \qquad v'_{t+1} := v'_t + \eta_t \nabla_v f(w'_t, v'_t; z'_{i_t}),$$

with same-index coupling of datasets: $z'_j = z_j$ for $j \neq i$ and $z'_i = \tilde{z}_i$. For $D \in \{\mathcal{D}, \mathcal{D}^{(i)}\}$, define

$$F_D(w, v) := \frac{1}{n}\sum_{j=1}^n f(w, v; z_j^D), \quad \Phi_D(w) := \max_v F_D(w, v), \quad \Phi_D^\star := \min_w \Phi_D(w).$$

We let $\mathcal{F}_t := \sigma\big((w_s, v_s, w'_s, v'_s, i_s)_{0\leq s\leq t}\big)$ be the natural filtration and introduce a *ghost* index $\hat{i}_t \sim \mathrm{Unif}(\{1, \dots, n\})$ independent of $\mathcal{F}_t$, shared by both runs. The role of the ghost index is to decouple the sampling noise at time t from the past and from the coupling across datasets.

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[5]:

(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) **Strong concavity in v.** $F_D(\cdot, \cdot)$ is ρ-strongly concave in v uniformly in w.

(C5) **Bounded gradients on the effective domain.** There exists a compact convex set $\Omega \subset \mathcal{W} \times \mathcal{V}$ such that the sequence of iterates $\{(w_t, v_t)\}_{t=0}^T$ generated by the algorithm remains within Ω almost surely. 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) **Selected saddle point.** Each dataset D admits at least one saddle point, and the SGDA iterates initialized at (w_0, v_0) converge to an initialization-dependent saddle point $(x_D^\star, v_D^\star)$, which we call the selected saddle.

In addition, we impose the following settings for the stability analysis. (S1) is a relatively trivial setting, (S2) is the key mechanism that allows the analysis to proceed without an i.i.d. assumption on the data, and (S3) is standard in the stability literature: for more details, see Appendix C.4.

(S1) **Stepsizes.** $0 < \eta_t \leq \min\{\frac{1}{4L}, \frac{1}{\rho}\}$.

(S2) **Shared-index coupling (no i.i.d. needed).** The two coupled runs on $\mathcal{D}$ and $\mathcal{D}^{(i)}$ use the *same* index sequence $(i_t)_{t\geq 0}$.

(S3) **Uniformity of constants across datasets.** The constants $L, \rho, \mu_{\mathrm{PL}}, \mu_{\mathrm{QG}}, G$ are the same for D and $D^{(i)}$.

To establish stability in our PL–strongly-concave 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 $\nabla\Phi_D(w_t)$ only when the dual variable v_t is already at its inner maximizer $v_D^\star(w_t)$; away from this manifold there is a non-negligible "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* :

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

where $\alpha \in \left[\frac{4L^2}{\rho^2}, \infty\right)$. Our Lyapunov potential is designed precisely to couple these two effects in a single scalar quantity: the PL term measures how close the current value function is to the data-dependent BRM optimum, and the dual-gap term penalizes the inner mismatch strongly enough that the beneficial drift from strong concavity in v dominates the adverse contribution of Δ_t to the primal dynamics.

With a suitable choice of the weight α, this potential contracts in expectation at each SGDA step, and can then be translated back into a bound on the distance between coupled SGDA trajectories via PL and QG. Because the step sizes in Lemma 2 satisfy the Robbins–Monro conditions $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$, the contraction dominates these perturbations and keeps the two trajectories close for large t. The next theorem formalizes this as an $\mathcal{O}(1/n)$ bound on the on-average argument stability of SGDA, stated in a general PL–strongly-concave minimax setting that our BRM formulation instantiates.

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

Theorem 3 (On-average argument stability of SGDA without i.i.d. sampling)**.** *Suppose that the loss function f and datasets $\mathcal{D}, \mathcal{D}^{(i)}$ satisfy (C1)–(C6) and under the setting (S1)–(S3). Let $\varepsilon_T := \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\|w_T(\mathcal{D}) - w_T(\mathcal{D}^{(i)})\| + \|v_T(\mathcal{D}) - v_T(\mathcal{D}^{(i)})\|\big]$. With choices of $\beta := 2L$ and α above, the following on-average argument stability bound holds:*[6]

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

where $\Psi_{\alpha,0}^{\max} := \max\big\{\mathbb{E}\big[\Psi_{\alpha,\mathcal{D}}(w_0,v_0)\big],\ \max_{1\le i\le n}\mathbb{E}\big[\Psi_{\alpha,\mathcal{D}^{(i)}}(w_0,v_0)\big]\big\}$, and c, C'_{var}, C_{dist}, C'_{hit} are all constants depending only on $L, \rho, \mu_{\mathrm{PL}}, \mu_{\mathrm{QG}}, G$.[7]

To make the rate transparent, we now specialize the stepsizes to the harmonic Robbins–Monro rule $\eta_t = c_1/(c_2+t)$, which Kang et al. [2025] chooses for Algorithm 1 to prove Lemma 2. This schedule satisfies Setting (S1) for suitable $c_1 > 0$, $c_2 \ge 1$ and turns the kernel sums in Theorem 3 into closed forms. This yields the next corollary, which displays roughly $O(T^{-1/2})$ decay of the optimization term while keeping the $O(1/n)$ contribution explicit. See Appendix B.2 for details.

Corollary 4 (Informal bound under a harmonic stepsize schedule)**.** *Choose the Robbins–Monro stepsizes in* ALGORITHM 1 *as $\eta_t = \frac{c_1}{c_2+t}, t \ge 0$, with constants $c_1 > 0$ and $c_2 \ge 1$ small enough that satistying (S1). Let $c := \min\{\mu_{\mathrm{PL}}/2, \rho/2\}$. Then from Theorem 3, the stability constant ε_T in Theorem 3 admits the stability bound scales as*

$$\varepsilon_T \;=\; O\!\left((c_2+T)^{-\min\left\{\frac{1}{2},\,\frac{3c\,c_1}{8}\right\}}\right) \;+\; O\big(\tfrac{1}{n}\big)\,.$$

3.2 Generalization

In this section, we quantify generalization through algorithmic stability we derived in the previous section. Following the minimax stability framework of Wang et al. [2022], stability controls both the *primal function*, which is the Bellman residual, and the *weak primal–dual gap*. We first define these two risks, then invoke the transfer lemma that turns our stability bound from Theorem 3 into generalization guarantees. Specifically, our goal is to 1) bound the difference between population Bellman–residual risk and empirical Bellman–residual risk and 2) bound the population Bellman–residual risk of the SGDA output.

Definition 4 (Primal Risk)**.** *Under (C4), define the* value function

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

Definition 5 (Weak primal–dual risk)**.** *For $(w,v) \in \mathcal{W}\times\mathcal{V}$, define the* population *and* empirical *weak–PD risks by*

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

The key transfer principle is stability ⇒ generalization for minimax problems: if the SGDA iterate (w_T, v_T) has on-average argument stability ε_T on neighboring datasets, then the primal value-function gap $\mathbb{E}\big[R(w_T) - R_n(w_T)\big]$ and the weak primal–dual gap $\big|\mathbb{E}[\Delta^{\mathrm{PD}}(w_T,v_T) - \Delta_n^{\mathrm{PD}}(w_T,v_T)]\big|$ can be effectively upper bounded by constant times ε_T from Theorem 3, which is $\mathcal{O}(1/n)$. The remainder of this subsection formalizes this to apply this transfer with our stability bound (Theorem 3) to obtain Theorem 6, $\mathcal{O}(1/n)$ generalization for BRM under SGDA.

Lemma 5 (Theorem 5, Wang et al. [2022])**.** *Let Conditions (C1), (C4), and (C5) hold. Let (w_T, v_T) be the SGDA iterates produced on $\mathcal{D}_n$ and let*

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

[6] Note that 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].

[7] $c := \min\{\mu_{\mathrm{PL}}/2,\ \rho/2\}$, $C'_{\mathrm{var}} = C_{\mathrm{var}}\Big(L(1+L/\rho) + \alpha\,\frac{L^2}{\rho}\Big)\,G^2$ for some numerical constant $C_{\mathrm{var}} > 0$, $C_{\mathrm{dist}} = \sqrt{\Big(1+\frac{L}{\rho}\Big)^2\,\frac{2}{\mu_{\mathrm{QG}}} + \frac{2}{\alpha\rho}}$, and $C'_{\mathrm{hit}} = 2G\left(\frac{(1+L/\rho)^2}{\sqrt{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}}} + \frac{1}{\rho}\right) + \frac{G}{L}\left(9 + \frac{L}{\rho}\right)$.

be the on-average argument-stability constant from Theorem 3. Then

$$\mathbb{E}\big[R(w_T) - R_n(w_T)\big] \;\le\; (1 + L/\rho)\, G\, \varepsilon_T,$$
$$\big|\,\mathbb{E}\big[\,\Delta^{\mathrm{PD}}(w_T, v_T) \;-\; \Delta_n^{\mathrm{PD}}(w_T, v_T)\big]\big| \;\le\; G\,\varepsilon_T.$$

Combining Corollary 4 and Lemma 5 [Wang et al., 2022], we arrive at Theorem 6, the main result of this paper, i.e., the *generalization guarantee of Bellman residuals*. In words, the learned Q (i.e., the corresponding learned w) generalizes: its empirical Bellman residual on the offline dataset closely matches its expected Bellman residual on the true MDP distribution. This is direct from the fact that proving $\mathcal{O}(1/n)$ stability for SGDA immediately delivers $\mathcal{O}(1/n)$ generalization bounds for Bellman residual minimization [Wang et al., 2022].

Theorem 6 (Generalization for the empirical Bellman–residual objective)**.** *Let* $\big(\widehat{w}^{(T)}, \widehat{v}^{(T)}\big)$ *be the parameters returned by* ALGORITHM 1 *after* T *SGDA iterations on the empirical objective* (5)*. Define the population and empirical risks*

$$\mathcal{R}(w) := \mathbb{E}_{(s,a)\sim\pi_D,\nu_0}\,\mathbb{E}_{s'\sim P(\cdot|s,a)}\Big[f(w, \tilde{v}^*; s, a, s')\Big],\; \widehat{\mathcal{R}}_n(w) := \frac{1}{N}\sum_{(s,a,s')\in\mathcal{D}_{\pi_D},\nu_0}\Big[f(w, \tilde{v}^*; s, a, s')\Big],$$

where $\tilde{v}^\star$ *is the minimizer in* (5)*. Then for* ε_T *from Theorem 3, we have*

$$\mathbb{E}\Big[\mathcal{R}\big(\widehat{w}^{(T)}\big) - \widehat{\mathcal{R}}_n\big(\widehat{w}^{(T)}\big)\Big] \;\le\; (1 + L/\rho)\, G\, \varepsilon_T,$$
$$\Big|\mathbb{E}\Big[\Delta^{\mathrm{PD}}\big(\widehat{w}^{(T)}, \widehat{v}^{(T)}\big) - \Delta_n^{\mathrm{PD}}\big(\widehat{w}^{(T)}, \widehat{v}^{(T)}\big)\Big]\Big| \;\le\; G\,\varepsilon_T.$$

The generalization guarantee in Theorem 6 is a critical result, confirming that the empirical risk is a reliable proxy for the true population risk. However, the ultimate measure of success for a learning algorithm is its performance on the population distribution relative to the best possible model. This is quantified by the *population excess risk*, which measures the gap $\mathcal{R}(\widehat{w}^{(T)}) - \mathcal{R}(w^*)$.

Theorem 7 (Population excess risk)**.** *Let* $\big(\widehat{w}^{(T)}, \widehat{v}^{(T)}\big)$ *be the SGDA iterate outcome of Algorithm 1 after* T *steps and* $w^* \in \arg\min_{w\in\mathcal{W}} \mathcal{R}(w)$ *its population minimiser. Then, with* ε_T *from Theorem 3 and* d_1, d_2 *defined in Lemma 2,*

$$\mathbb{E}\Big[\mathcal{R}\big(\widehat{w}^{(T)}\big) - \mathcal{R}\big(w^*\big)\Big] \;\le\; \underbrace{(1 + L/\rho)\, G\, \varepsilon_T}_{\textit{stability / generalization}} \;+\; \underbrace{\frac{d_1}{d_2 + T}}_{\textit{optimization error}}$$

4 Conclusion and Limitation

We studied the statistical behavior of Bellman Residual Minimization (BRM) in offline RL/IRL through the lens of stability. Exploiting the PL–strongly-concave geometry of the bi-conjugate formulation, we coupled two SGDA trajectories on neighboring datasets with a single Lyapunov potential and a ghost-index decoupling device. This yielded an *on-average argument-stability* bound with $\mathcal{O}(1/n)$ rate (Theorem 3), which directly implies $\mathcal{O}(1/n)$ generalization for BRM (Theorem 6) and a population excess-risk bound that cleanly decomposes optimization and estimation errors (Theorem 7). The analysis is constructive, tracks explicit constants in $(L, \rho, \mu_{\mathrm{PL}}, \mu_{\mathrm{QG}}, G)$, accommodates minibatching, and requires neither variance reduction nor independence assumptions on the sampling indices. Together with the global convergence of SGDA in parameter space (Lemma 2), these results close the statistical gap for BRM and improve the sample-complexity exponent over the $\mathcal{O}(n^{-1/2})$ rates known for convex–concave saddle problems.

About the 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.

References

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.

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.

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.

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.

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*, 2024.

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.

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.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

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, 2018.

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.

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
$\Delta_{\mathcal{S}}^{\mathcal{A}}$	Set of stationary Markov policies
$\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}$	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
$\mathcal{D}^{(i)}$	neighbouring dataset
g_t^w, g_t^v	per-sample 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	Strong concavity 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 4)
$\Delta^{PD}, \Delta_n^{PD}$	Weak primal-dual risk (Definition 5)

Table 1: Notations

B Technical Proofs

B.1 Supporting Lemmas for proving Theorem 3

Lemma 8 (Mismatch control)**.** *Let $v_D^\star(w):=\arg\max_v F_D(w,v)$. Under (C1) and (C4),*

$$\Delta(w,v):=\nabla_w F_D(w,v)-\nabla\Phi_D(w)=\nabla_w F_D(w,v)-\nabla_w F_D\big(w,v_D^\star(w)\big)$$

satisfies $\|\Delta(w,v)\|\le L\,\|v-v_D^\star(w)\|$. Moreover,

$$\|v-v_D^\star(w)\|^2\ \le\ \frac{2}{\rho}\big(\Phi_D(w)-F_D(w,v)\big).$$

Proof. L-Lipschitzness of $\nabla_w F_D(w,\cdot)$ yields the first bound. For $g(\cdot):=F_D(w,\cdot)$, ρ-strong concavity implies $g(v^\star)-g(v)\ge(\rho/2)\|v-v^\star\|^2$ [Nesterov, 2004], which gives the second inequality. □

Lemma 9 (Smoothness of the value function Φ)**.** *Assume (C1) (joint L-smoothness) and (C4) (ρ-strong concavity in v). Then, for each dataset D:*

(i) the maximizer $v_D^\star(w) := \arg\max_v F_D(w,v)$ is well-defined and (L/ρ)-Lipschitz:

$$\|v_D^\star(w) - v_D^\star(u)\| \;\le\; \tfrac{L}{\rho}\,\|w-u\| \quad \forall\, w,u,$$

(ii) $\Phi_D(w) := \max_v F_D(w,v)$ is differentiable with $\nabla\Phi_D(w) = \nabla_w F_D\big(w, v_D^\star(w)\big)$ (Danskin),

(iii) and Φ_D is L_Φ-smooth with

$$\|\nabla\Phi_D(w) - \nabla\Phi_D(u)\| \;\le\; L\Big(1+\tfrac{L}{\rho}\Big)\,\|w-u\| \qquad \forall\, w,u.$$

Proof. (i) Lipschitzness of $v_D^\star(\cdot)$. By (C4), for each fixed w the map $v \mapsto F_D(w,v)$ is ρ-strongly concave, so $v_D^\star(w)$ is unique. Using the first-order conditions $\nabla_v F_D(w, v_D^\star(w)) = 0$ and $\nabla_v F_D(u, v_D^\star(u)) = 0$, write

$$\begin{aligned} 0 \;&=\; \nabla_v F_D(w, v_D^\star(w)) - \nabla_v F_D(u, v_D^\star(u)) \\ &= \underbrace{\big[\nabla_v F_D(w, v_D^\star(w)) - \nabla_v F_D(w, v_D^\star(u))\big]}_{(\mathsf{A})} + \underbrace{\big[\nabla_v F_D(w, v_D^\star(u)) - \nabla_v F_D(u, v_D^\star(u))\big]}_{(\mathsf{B})}. \end{aligned}$$

Strong concavity makes $\nabla_v F_D(w,\cdot)$ ρ-strongly *monotone*, so $\langle (\mathsf{A}),\, v_D^\star(w) - v_D^\star(u)\rangle \le -\rho\|v_D^\star(w) - v_D^\star(u)\|^2$. Joint L-smoothness gives $\|(\mathsf{B})\| \le L\|w-u\|$. Taking inner products with $v_D^\star(w) - v_D^\star(u)$ yields

$$\rho\|v_D^\star(w) - v_D^\star(u)\| \;\le\; \|(\mathsf{B})\| \;\le\; L\|w-u\| \qquad\Rightarrow\qquad \|v_D^\star(w) - v_D^\star(u)\| \;\le\; \tfrac{L}{\rho}\|w-u\|.$$

(ii) By joint smoothness, Danskin's theorem applies: $\nabla\Phi_D(w) = \nabla_w F_D\big(w, v_D^\star(w)\big)$.

(iii) Smoothness of Φ_D. For any w, u,

$$\begin{aligned} &\|\nabla\Phi_D(w) - \nabla\Phi_D(u)\| \\ &= \big\|\nabla_w F_D\big(w, v_D^\star(w)\big) - \nabla_w F_D\big(u, v_D^\star(u)\big)\big\| \\ &\le \underbrace{\big\|\nabla_w F_D\big(w, v_D^\star(w)\big) - \nabla_w F_D\big(w, v_D^\star(u)\big)\big\|}_{\le L\|v_D^\star(w) - v_D^\star(u)\|} + \underbrace{\big\|\nabla_w F_D\big(w, v_D^\star(u)\big) - \nabla_w F_D\big(u, v_D^\star(u)\big)\big\|}_{\le L\|w-u\|} \\ &\le L\Big(\tfrac{L}{\rho}+1\Big)\|w-u\|. \end{aligned}$$

The first two inequalities use joint L-smoothness (of $\nabla_w F_D$ in both arguments); the last uses part (i). Thus $L_\Phi \le L(1 + L/\rho)$. □

Lemma 10 (Cross-dataset gradient sensitivity for Φ)**.** *Under (C1), (C4) and (C5), for any w, u and any 1-sample replacement $D \to D^{(i)}$,*

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

Proof. By Lemma 9, $v_D^\star(\cdot)$ is (L/ρ)-Lipschitz, Φ_D is L_Φ-smooth, and $\nabla\Phi_D(w) = \nabla_w F_D\big(w, v_D^\star(w)\big)$. Hence

$$\|\nabla\Phi_D(w) - \nabla\Phi_{D^{(i)}}(u)\| \;\le\; \underbrace{\|\nabla\Phi_D(w) - \nabla\Phi_D(u)\|}_{\le L_\Phi\|w-u\|} \;+\; \underbrace{\|\nabla\Phi_D(u) - \nabla\Phi_{D^{(i)}}(u)\|}_{(\mathsf{A})}.$$

We bound (A) by inserting and subtracting $v_{D^{(i)}}^\star(u)$ and using joint L-smoothness of F:

$$\begin{aligned} (\mathsf{A}) &= \big\|\nabla_w F_D\big(u, v_D^\star(u)\big) - \nabla_w F_{D^{(i)}}\big(u, v_{D^{(i)}}^\star(u)\big)\big\| \\ &\le \underbrace{\big\|\nabla_w F_D\big(u, v_D^\star(u)\big) - \nabla_w F_D\big(u, v_{D^{(i)}}^\star(u)\big)\big\|}_{\le L\,\|v_D^\star(u) - v_{D^{(i)}}^\star(u)\|} + \underbrace{\big\|\nabla_w F_D\big(u, v_{D^{(i)}}^\star(u)\big) - \nabla_w F_{D^{(i)}}\big(u, v_{D^{(i)}}^\star(u)\big)\big\|}_{\le \frac{2G}{n} \text{ by Condition (C5)}}, \end{aligned}$$

so that

$$(\mathsf{A}) \;\le\; L\,\|v_D^\star(u) - v_{D^{(i)}}^\star(u)\| + \frac{2G}{n}.$$

Since $v \mapsto F_D(u,v)$ is ρ-strongly concave, the gradient map is ρ-strongly monotone, which gives the error bound with the normal cone $N_{\mathcal{V}}(\cdot)$:

$$\|v - v_D^\star(u)\| \;\le\; \frac{1}{\rho}\,\mathrm{dist}\big(\nabla_v F_D(u,v),\, N_{\mathcal{V}}(v)\big) \qquad \forall v \in \mathcal{V}.$$

Apply this at $v = v_{D^{(i)}}^\star(u)$. There are two cases.

(i) Unconstrained (or interior) maximizer. Then $N_{\mathcal{V}}\big(v_{D^{(i)}}^\star(u)\big) = \{0\}$, so

$$\|v_D^\star(u) - v_{D^{(i)}}^\star(u)\| \;\le\; \frac{1}{\rho}\,\|\nabla_v F_D(u, v_{D^{(i)}}^\star(u))\| \;=\; \frac{1}{\rho}\,\|\nabla_v F_D(u, v_{D^{(i)}}^\star(u)) - \nabla_v F_{D^{(i)}}(u, v_{D^{(i)}}^\star(u))\|.$$

(ii) Constrained boundary maximizer. By KKT optimality,

$$0 \;\in\; -\nabla_v F_{D^{(i)}}(u, v_{D^{(i)}}^\star(u)) + N_{\mathcal{V}}(v_{D^{(i)}}^\star(u)) \iff \nabla_v F_{D^{(i)}}(u, v_{D^{(i)}}^\star(u)) \in N_{\mathcal{V}}(v_{D^{(i)}}^\star(u)).$$

Hence, choosing $\xi := \nabla_v F_{D^{(i)}}(u, v_{D^{(i)}}^\star(u)) \in N_{\mathcal{V}}(v_{D^{(i)}}^\star(u))$, we obtain

$$\mathrm{dist}\big(\nabla_v F_D(u, v_{D^{(i)}}^\star(u)),\, N_{\mathcal{V}}(v_{D^{(i)}}^\star(u))\big) \;\le\; \big\|\nabla_v F_D(u, v_{D^{(i)}}^\star(u)) - \nabla_v F_{D^{(i)}}(u, v_{D^{(i)}}^\star(u))\big\|.$$

In either case, using the single-sample replacement identity

$$\nabla_v F_D(u,v) - \nabla_v F_{D^{(i)}}(u,v) \;=\; \frac{1}{n}\Big(\nabla_v f(u,v;z_i) - \nabla_v f(u,v;\tilde{z}_i)\Big),$$

the triangle inequality and Condition (C5) give

$$\begin{aligned}
\|v_D^\star(u) - v_{D^{(i)}}^\star(u)\| &\le \frac{1}{\rho}\,\mathrm{dist}\big(\nabla_v F_D(u, v_{D^{(i)}}^\star(u)),\, N_{\mathcal{V}}(v_{D^{(i)}}^\star(u))\big) \\
&\le \frac{1}{\rho}\,\big\|\nabla_v F_D(u, v_{D^{(i)}}^\star(u)) - \nabla_v F_{D^{(i)}}(u, v_{D^{(i)}}^\star(u))\big\| \\
&= \frac{1}{\rho}\cdot\frac{1}{n}\,\big\|\nabla_v f(u, v_{D^{(i)}}^\star(u); z_i) - \nabla_v f(u, v_{D^{(i)}}^\star(u); \tilde{z}_i)\big\| \\
&\le \frac{1}{\rho}\cdot\frac{1}{n}\,\Big(\|\nabla_v f(u, v_{D^{(i)}}^\star(u); z_i)\| + \|\nabla_v f(u, v_{D^{(i)}}^\star(u); \tilde{z}_i)\|\Big) \\
&\le \frac{2G}{\rho n}\,.
\end{aligned}$$

Plugging this into the bound for (A) yields

$$(\mathsf{A}) \;\le\; L\cdot\frac{2G}{\rho n} + \frac{2G}{n} \;=\; \frac{2G}{n}\Big(1 + \frac{L}{\rho}\Big),$$

and therefore

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

□

Lemma 11 (Deterministic potential ⇒ distance)**.** *Under (C3) and (C4), for any $\alpha > 0$ and any (w,v),*

$$\|w - x_D^\star\| + \|v - v_D^\star\| \;\le\; \sqrt{\Big(1+\tfrac{L}{\rho}\Big)^2\tfrac{2}{\mu_{\mathrm{QG}}} \;+\; \tfrac{2}{\alpha\rho}}\;\cdot\sqrt{\Psi_{\alpha,D}(w,v)}.$$

Proof. Using $\|v - v_D^\star\| \le \|v - v_D^\star(w)\| + \|v_D^\star(w) - v_D^\star\|$ and the (L/ρ)-Lipschitzness of $w \mapsto v_D^\star(w)$,

$$\|w - x_D^\star\| + \|v - v_D^\star\| \le (1+\tfrac{L}{\rho})\|w - x_D^\star\| + \|v - v_D^\star(w)\|.$$

$$\|w - x_D^\star\| \le \sqrt{\tfrac{2}{\mu_{\mathrm{QG}}}\big(\Phi_D(w) - \Phi_D^\star\big)}, \qquad \|v - v_D^\star(w)\| \le \sqrt{\tfrac{2}{\rho}\big(\Phi_D(w) - F_D(w,v)\big)},$$

where $v_D^\star(w) \in \arg\max_u F_D(w,u)$ and $\Phi_D(w) = \max_u F_D(w,u)$.

Using the weighted Cauchy–Schwarz inequality with any $\alpha > 0$,

$$\begin{aligned}&(1+\tfrac{L}{\rho})\sqrt{\tfrac{2}{\mu_{\mathrm{QG}}}\big(\Phi_D(w) - \Phi_D^\star\big)} + \sqrt{\tfrac{2}{\rho}\big(\Phi_D(w) - F_D(w,v)\big)}\\ &\le \sqrt{\big(\Phi_D(w) - \Phi_D^\star\big) + \alpha\big(\Phi_D(w) - F_D(w,v)\big)}\,\sqrt{\tfrac{2(1+L/\rho)^2}{\mu_{\mathrm{QG}}} + \tfrac{2}{\alpha\rho}}.\end{aligned}$$

Noting that $\Psi_{\alpha,D}(w,v) := \big(\Phi_D(w) - \Phi_D^\star\big) + \alpha\big(\Phi_D(w) - F_D(w,v)\big)$, we obtain

$$(1+\tfrac{L}{\rho})\sqrt{\tfrac{2}{\mu_{\mathrm{QG}}}\big(\Phi_D(w) - \Phi_D^\star\big)} + \sqrt{\tfrac{2}{\rho}\big(\Phi_D(w) - F_D(w,v)\big)} \;\le\; \sqrt{\Psi_{\alpha,D}(w,v)}\,\sqrt{\tfrac{2(1+L/\rho)^2}{\mu_{\mathrm{QG}}} + \tfrac{2}{\alpha\rho}}.$$

□

Lemma 12 (Saddle-point sensitivity)**.** *Let $D^{(i)}$ be obtained from D by replacing one sample. Assume (C1) (joint L-smoothness), (C4) (ρ-strong concavity in v), (C5) (per-sample gradients bounded by G), (C2)–(C3) (PL and QG for Φ_D), and (S3) (the same constants hold for D and $D^{(i)}$). Let $(x_D^\star, v_D^\star)$ and $\big(x_{D^{(i)}}^\star, v_{D^{(i)}}^\star\big)$ be the selected saddle points from Condition (C6) for D and $D^{(i)}$, respectively. Then Then*

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

Proof. At $(x_{D^{(i)}}^\star, v_{D^{(i)}}^\star)$, only one summand in F_D differs from $F_{D^{(i)}}$, and per-sample gradients are bounded by G. Hence,

$$\begin{aligned}\|\nabla_w F_D(x_{D^{(i)}}^\star, v_{D^{(i)}}^\star)\| &= \|\nabla_w F_D(x_{D^{(i)}}^\star, v_{D^{(i)}}^\star) - \nabla_w F_D^{(i)}(x_{D^{(i)}}^\star, v_{D^{(i)}}^\star)\| \le \tfrac{2G}{n}\\ \|\nabla_v F_D(x_{D^{(i)}}^\star, v_{D^{(i)}}^\star)\| &= \|\nabla_v F_D(x_{D^{(i)}}^\star, v_{D^{(i)}}^\star) - \nabla_v F_D^{(i)}(x_{D^{(i)}}^\star, v_{D^{(i)}}^\star)\| \le \tfrac{2G}{n}.\end{aligned}$$

Using Lemma 8,

$$\|\nabla\Phi_D(x_{D^{(i)}}^\star)\| \le \|\nabla_w F_D(x_{D^{(i)}}^\star, v_{D^{(i)}}^\star)\| + L\,\|v_{D^{(i)}}^\star - v_D^\star(x_{D^{(i)}}^\star)\| \le \tfrac{2G}{n}(1+L/\rho),$$

where the last step uses ρ-strong concavity to get $\|v_{D^{(i)}}^\star - v_D^\star(x_{D^{(i)}}^\star)\| \le \frac{1}{\rho}\|\nabla_v F_D(x_{D^{(i)}}^\star, v_{D^{(i)}}^\star)\|$. Since Φ_D satisfies PL and QG,

$$\|x_D^\star - x_{D^{(i)}}^\star\| \le \frac{1}{\sqrt{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}}}\,\|\nabla\Phi_D(x_{D^{(i)}}^\star)\| \le \tfrac{2G}{n}\cdot\frac{1+L/\rho}{\sqrt{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}}}.$$

For the dual variable, split and use the (L/ρ)-Lipschitzness of $w \mapsto v_D^\star(w)$ (Lemma 9) and strong concavity:

$$\begin{aligned}\|v_D^\star - v_{D^{(i)}}^\star\| &\le \|v_D^\star(x_D^\star) - v_D^\star(x_{D^{(i)}}^\star)\| + \|v_D^\star(x_{D^{(i)}}^\star) - v_{D^{(i)}}^\star(x_{D^{(i)}}^\star)\|\\ &\le \tfrac{L}{\rho}\|x_D^\star - x_{D^{(i)}}^\star\| + \tfrac{1}{\rho}\|\nabla_v F_D(x_{D^{(i)}}^\star, v_{D^{(i)}}^\star)\|\\ &\le \tfrac{2G}{n}\left(\tfrac{L}{\rho}\cdot\frac{1+L/\rho}{\sqrt{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}}} + \tfrac{1}{\rho}\right),\end{aligned}$$

where the last inequality uses the same single-sample replacement and gradient bound as above (cf. Lemma 10). Adding the two bounds yields the claim. □

Lemma 13 (Coupled one-step bounds)**.** *Assume (C1)–(S3). Define*

$$\Xi_t \;:=\; \Psi_{\alpha,\mathcal{D}}(w_t, v_t) + \Psi_{\alpha,\mathcal{D}^{(i)}}(w_t', v_t'), \qquad D_t \;:=\; \|w_t - w_t'\| + \|v_t - v_t'\|.$$

Let $c := \min\{\mu_{\mathrm{PL}}/2, \rho/2\}$ and $L_\Phi \le L(1 + L/\rho)$ (Lemma 9). Then, for $\tilde{B}_0 = \tilde{B}_1 = 8$ and a constant $C_{\mathrm{var}} > 0$ that depends only on L, ρ, we have

$$(\mathbf{P}) \quad \mathbb{E}\big[\Xi_{t+1} \mid \mathcal{F}_t\big] \;\le\; (1 - c\eta_t)\,\Xi_t \;+\; A\,\eta_t^2 \\ +\; C_{\mathrm{leak}}\,\eta_t\,D_t \;+\; C_*\,\tfrac{G}{n}\,\eta_t \;+\; C_{\mathrm{lin}}\,G^2\,\eta_t \;+\; \eta_t\big(\mathbf{1}\{i_t = i\} + \mathbf{1}\{\hat{i}_t = i\}\big)\tilde{B}_0 G^2, \tag{6}$$

$$(\mathbf{D}_{\mathrm{weak}}) \quad \mathbb{E}\big[D_{t+1} \mid \mathcal{F}_t\big] \;\le\; (1 + \kappa\eta_t)\,D_t \;+\; a\,\eta_t\sqrt{\Xi_t} \;+\; \eta_t\,\mathbf{1}\{i_t = i\}\,\tilde{B}_1 G \;+\; \eta_t\,\frac{2G}{n}\Big(1 + \frac{L}{\rho}\Big), \tag{7}$$

where

$$A := C_{\mathrm{var}}\Big(L(1{+}L/\rho) + \alpha L^2/\rho\Big)G^2, \quad C_{\mathrm{leak}} := 2(2L + L_\Phi),$$

$$a := 2C_{\mathrm{dist}}\big(L + L_\Phi\big), \quad \kappa := 2L, \quad C_* := 4\big(1 + L/\rho\big), \quad C_{\mathrm{dist}} := \sqrt{\Big(1 + \tfrac{L}{\rho}\Big)^2 \frac{2}{\mu_{\mathrm{QG}}} + \frac{2}{\alpha\rho}},$$

$$\textit{and} \qquad C_{\mathrm{lin}} \;:=\; 12 + \frac{10L^2}{\rho^2}\,.$$

Why introduce a ghost index? Writing and conditioning on the past, the primal descent term $-\eta_t\langle\nabla\Phi(w_t), \nabla_w f(w_t, v_t; z_{i_t})\rangle$ has no sign we can control, since i_t is already revealed in $\mathcal{F}_t$. We therefore add–subtract a ghost gradient and take $\mathbb{E}_{\hat{i}_t}[\cdot \mid \mathcal{F}_t]$:

$$-\eta_t\langle\nabla\Phi(w_t), g_t^w\rangle = -\eta_t\langle\nabla\Phi(w_t), \hat{g}_t^w\rangle - \eta_t\langle\nabla\Phi(w_t), g_t^w - \hat{g}_t^w\rangle.$$

The first term becomes the correct drift $-\eta_t\langle\nabla\Phi(w_t), \nabla_w F(w_t, v_t)\rangle$, which contracts by PL for Φ, while the second is a centered correction that we bound by Young's inequality and the gradient bound $\|g_t^w - \hat{g}_t^w\| \le 2G$, yielding an $O(\eta_t G^2)$ remainder absorbed into the variance term. An identical maneuver applies to the dual-gap part Γ. This device avoids any i.i.d. sampling assumption and yields the one-step recursion (P) with a contractive factor and small, explicit noise.

Proof. All expectations are conditional on $\mathcal{F}_t := \sigma\big((w_s, v_s, w'_s, v'_s, i_s)_{0 \le s \le t}\big)$. Write $\gamma := \eta_t$. Fix $D \in \{\mathcal{D}, \mathcal{D}^{(i)}\}$ and abbreviate $F := F_D$, $\Phi := \Phi_D$, $\Psi_\alpha := \Psi_{\alpha,D}$, $\Gamma(w, v) := \Phi(w) - F(w, v)$. Set $g^w := \nabla_w f(w_t, v_t; z_{i_t})$, $g^v := \nabla_v f(w_t, v_t; z_{i_t})$,

$$w^+ = w_t - \gamma g^w, \qquad v^+ = v_t + \gamma g^v,$$

and introduce a ghost index $\hat{i}_t$ independent of $\mathcal{F}_t$ with $\hat{g}^w := \nabla_w f(w_t, v_t; z_{\hat{i}_t})$, $\hat{g}^v := \nabla_v f(w_t, v_t; z_{\hat{i}_t})$.

(a) Primal part Φ. Recall $w^+ = w_t - \gamma\, g^w$ with $g^w := \nabla_w f(w_t, v_t; z_{i_t})$ and let $\gamma := \eta_t$. By L_Φ-smoothness of Φ,

$$\Phi(w^+) - \Phi(w_t) \;\le\; -\gamma\,\langle\nabla\Phi(w_t),\, g^w\rangle \;+\; \tfrac{L_\Phi}{2}\,\gamma^2\,\|g^w\|^2.$$

Insert and subtract a ghost gradient $\hat{g}^w := \nabla_w f(w_t, v_t; z_{\hat{i}_t})$ with $\hat{i}_t \perp \mathcal{F}_t$, then take $\mathbb{E}_{\hat{i}_t}[\cdot \mid \mathcal{F}_t]$. Using $\mathbb{E}_{\hat{i}_t}[\hat{g}^w \mid \mathcal{F}_t] = \nabla_w F(w_t, v_t)$ and $\|g^w\| \le G$,

$$\mathbb{E}_{\hat{i}_t}\big[\Phi(w^+) - \Phi(w_t) \mid \mathcal{F}_t\big] \le -\gamma\,\langle\nabla\Phi(w_t),\, \nabla_w F(w_t, v_t)\rangle \;-\; \gamma\,\langle\nabla\Phi(w_t),\, g^w - \nabla_w F(w_t, v_t)\rangle \;+\; \tfrac{L_\Phi}{2}\gamma^2 G^2. \tag{8}$$

For the centered correction, Young's inequality with $\|g^w - \nabla_w F(w_t, v_t)\| \le 2G$ yields

$$\gamma\,\big|\langle\nabla\Phi(w_t),\, g^w - \nabla_w F(w_t, v_t)\rangle\big| \;\le\; \tfrac{1}{4}\,\gamma\,\|\nabla\Phi(w_t)\|^2 \;+\; 4\,\gamma\,G^2. \tag{9}$$

For the drift, write $\Delta_t := \nabla_w F(w_t, v_t) - \nabla\Phi(w_t)$. Then

$$-\langle\nabla\Phi(w_t),\, \nabla_w F(w_t, v_t)\rangle = \;-\|\nabla\Phi(w_t)\|^2 \;-\; \langle\nabla\Phi(w_t),\, \Delta_t\rangle \;\le\; -\tfrac{1}{2}\|\nabla\Phi(w_t)\|^2 \;+\; \tfrac{1}{2}\|\Delta_t\|^2. \tag{10}$$

By Lemma 8 and ρ-strong concavity in v,

$$\|\Delta_t\| \;\le\; L\,\|v_t - v^\star(w_t)\| \;\le\; L\sqrt{\tfrac{2}{\rho}\,\Gamma(w_t, v_t)} \qquad \Rightarrow \qquad \tfrac{1}{2}\|\Delta_t\|^2 \;\le\; \tfrac{L^2}{\rho}\,\Gamma(w_t, v_t), \tag{11}$$

where $\Gamma(w, v) := \Phi(w) - F(w, v)$. Combining (8), (9), (10), and (11),

$$\mathbb{E}_{\hat{i}_t}\big[\Phi(w^+) - \Phi(w_t) \mid \mathcal{F}_t\big] \;\le\; -\tfrac{1}{4}\,\gamma\,\|\nabla\Phi(w_t)\|^2 \;+\; \tfrac{L^2}{\rho}\,\gamma\,\Gamma(w_t, v_t) \;+\; \tfrac{L_\Phi}{2}\gamma^2 G^2 \;+\; 4\,\gamma\,G^2.$$

Finally, by the PL inequality $\frac{1}{2}\|\nabla\Phi(w_t)\|^2 \ge \mu_{\mathrm{PL}}\big(\Phi(w_t) - \Phi^\star\big)$,

$$-\tfrac{1}{4}\,\gamma\,\|\nabla\Phi(w_t)\|^2 \;\le\; -\tfrac{\mu_{\mathrm{PL}}}{2}\,\gamma\,\big(\Phi(w_t) - \Phi^\star\big),$$

hence

$$\mathbb{E}_{\hat{i}_t}\big[\Phi(w^+) - \Phi(w_t) \mid \mathcal{F}_t\big] \;\le\; -\tfrac{\mu_{\mathrm{PL}}}{2}\,\gamma\,\big(\Phi(w_t) - \Phi^\star\big) + \tfrac{L^2}{\rho}\,\gamma\,\Gamma(w_t, v_t) + \tfrac{L_\Phi}{2}\gamma^2 G^2 + 4\,\gamma\,G^2.$$

Since $\Phi(w_t) - \Phi^\star$ is $\mathcal{F}_t$-measurable, adding it to both sides yields

$$\mathbb{E}_{\hat{i}_t}\big[\Phi(w^+) - \Phi^\star \mid \mathcal{F}_t\big] \;\le\; \big(1 - \tfrac{\mu_{\mathrm{PL}}}{2}\gamma\big)\big(\Phi(w_t) - \Phi^\star\big) + \tfrac{L^2}{\rho}\,\gamma\,\Gamma(w_t, v_t) + \tfrac{L_\Phi}{2}\gamma^2 G^2 + 4\,\gamma\,G^2.$$

(b) Dual-gap part Γ. Recall $\Gamma(w, v) := \Phi_D(w) - F_D(w, v)$ and write $v^+ = v_t + \gamma\,g_t^v$ with $g_t^v := \nabla_v f(w_t, v_t; z_{i_t})$. By Lemma 14, for fixed w_t,

$$\Gamma(w_t, v^+) \le (1 - 2\rho\gamma + \rho L\gamma^2)\,\Gamma(w_t, v_t). \tag{12}$$

Let $w^+ := w_t - \gamma g_t^w$ with $g_t^w := \nabla_w f(w_t, v_t; z_{i_t})$. By L_Φ-smoothness of Φ_D and joint L-smoothness of F_D,

$$\Gamma(w^+, v^+) \le \Gamma(w_t, v^+) - \gamma\langle\nabla\Phi_D(w_t), g_t^w\rangle + \gamma\langle\nabla_w F_D(w_t, v^+), g_t^w\rangle + \tfrac{L_\Phi + L}{2}\gamma^2 G^2.$$

Insert and subtract a ghost gradient $\hat{g}_t^w := \nabla_w f(w_t, v_t; z_{\hat{i}_t})$, take $\mathbb{E}_{\hat{i}_t}[\cdot \mid \mathcal{F}_t]$, and use $\mathbb{E}_{\hat{i}_t}[\hat{g}_t^w \mid \mathcal{F}_t] = \nabla_w F_D(w_t, v_t)$. Then

$$\begin{aligned}\mathbb{E}_{\hat{i}_t}\big[\Gamma(w^+, v^+) \mid \mathcal{F}_t\big] \le\; & \Gamma(w_t, v^+) - \gamma\langle\nabla\Phi_D(w_t),\, \nabla_w F_D(w_t, v_t)\rangle \\ & + \gamma\big\langle\nabla_w F_D(w_t, v^+) - \nabla\Phi_D(w_t),\, \nabla_w F_D(w_t, v_t)\big\rangle \\ & + \gamma\big\langle\nabla_w F_D(w_t, v^+) - \nabla\Phi_D(w_t),\, g_t^w - \nabla_w F_D(w_t, v_t)\big\rangle \\ & + \tfrac{L_\Phi + L}{2}\gamma^2 G^2.\end{aligned}$$

Decompose $\nabla_w F_D(w_t, v^+) - \nabla\Phi_D(w_t) = \Delta_t + \big(\nabla_w F_D(w_t, v^+) - \nabla_w F_D(w_t, v_t)\big)$, where $\Delta_t := \nabla_w F_D(w_t, v_t) - \nabla\Phi_D(w_t)$. Using Lemma 8, $\|\Delta_t\| \le L\sqrt{2\Gamma(w_t, v_t)/\rho}$, joint L-smoothness, $\|\nabla_w F_D(w_t, v^+) - \nabla_w F_D(w_t, v_t)\| \le L\gamma G$, and $\|g_t^w - \nabla_w F_D(w_t, v_t)\| \le 2G$,

$$\begin{aligned}&\gamma\big|\langle\Delta_t,\, g_t^w - \nabla_w F_D(w_t, v_t)\rangle\big| \;\le\; \tfrac{\rho}{4}\gamma\,\Gamma(w_t, v_t) + \tfrac{8L^2}{\rho^2}\gamma G^2,\\ &\gamma\big|\langle\nabla_w F_D(w_t, v^+) - \nabla_w F_D(w_t, v_t),\, g_t^w - \nabla_w F_D(w_t, v_t)\rangle\big| \;\le\; 2L\,\gamma^2 G^2,\\ &\gamma\big|\langle\Delta_t,\, \nabla_w F_D(w_t, v_t)\rangle\big| \;\le\; \tfrac{\rho}{4}\gamma\,\Gamma(w_t, v_t) + \tfrac{2L^2}{\rho^2}\gamma G^2,\\ &\gamma\big|\langle\nabla_w F_D(w_t, v^+) - \nabla_w F_D(w_t, v_t),\, \nabla_w F_D(w_t, v_t)\rangle\big| \;\le\; L\,\gamma^2 G^2.\end{aligned}$$

Combining the bounds with the ascent contraction (12), we can group the contributions as follows:

(i) From the primal descent step we obtain

$$-\tfrac{1}{4}\,\gamma\,\|\nabla\Phi_D(w_t)\|^2 \;\le\; -\tfrac{\mu_{\mathrm{PL}}}{2}\,\gamma\big(\Phi_D(w_t) - \Phi_D^\star\big),$$

where the inequality uses the PL condition. From the dual ascent we obtain

$$(1 - 2\rho\gamma + \rho L\gamma^2)\,\Gamma(w_t, v_t) = \Gamma(w_t, v_t) - 2\rho\gamma\,\Gamma(w_t, v_t) + O(\gamma^2),$$

so the leading negative part is $-2\rho\gamma\,\Gamma(w_t, v_t)$. Together, these two negative terms contract the Lyapunov potential $\Psi_\alpha(w_t, v_t) = (\Phi_D(w_t) - \Phi_D^\star) + \alpha\Gamma(w_t, v_t)$, yielding a multiplicative

shrinkage factor $(1 - c\gamma)\Psi_\alpha(w_t, v_t)$ with $c := \min\{\mu_{\mathrm{PL}}/2, \rho/2\}$. Furthermore, the positive $\gamma\Gamma$ pieces produced by the algebra consist of

$$\alpha \cdot \tfrac{\rho}{2}\,\gamma\,\Gamma(w_t, v_t) \quad \text{and} \quad \tfrac{L^2}{\rho}\,\gamma\,\Gamma(w_t, v_t).$$

Together with the dual drift they appear as

$$-2\alpha\rho\,\gamma\,\Gamma \;+\; \alpha \cdot \tfrac{\rho}{2}\,\gamma\,\Gamma \;+\; \tfrac{L^2}{\rho}\,\gamma\,\Gamma \;=\; -\tfrac{3}{2}\alpha\rho\,\gamma\,\Gamma \;+\; \tfrac{L^2}{\rho}\,\gamma\,\Gamma.$$

Choose α so that $\frac{L^2}{\rho} \le \frac{\alpha\rho}{2}$ (i.e. $\alpha \ge 2L^2/\rho^2$). Then

$$-\tfrac{3}{2}\alpha\rho\,\gamma\,\Gamma \;+\; \tfrac{L^2}{\rho}\,\gamma\,\Gamma \;\le\; -\tfrac{\alpha\rho}{2}\,\gamma\,\Gamma,$$

so these $\gamma\Gamma$ remainders are dominated by the dual drift and absorbed into the contraction factor. A simple sufficient standing choice is $\alpha \ge 4L^2/\rho^2$.

(ii) The mismatch bounds, ghost-correction terms, and gradient–difference terms contribute constants times γG^2. All such pieces are nonnegative and can be grouped into a single term $C_{\mathrm{lin}}\,G^2\,\gamma$.

(iii) Smoothness corrections (e.g. $(L_\Phi + L)/2\,\gamma^2 G^2$, $2L\,\gamma^2 G^2$, $L\,\gamma^2 G^2$) contribute constants times $\gamma^2 G^2$. These are also nonnegative and can be collected into $A\,\gamma^2$.

Putting the three groups together, the one-step recursion takes the form

$$\mathbb{E}\big[\Psi_\alpha(w^+, v^+)\,\big|\,\mathcal{F}_t\big] \;\le\; (1 - c\gamma)\,\Psi_\alpha(w_t, v_t) \;+\; C_{\mathrm{lin}}G^2\,\gamma \;+\; A\,\gamma^2,$$

which is exactly the recursion $(\mathbf{P})$ used in the stability analysis.

(c) Two-run recursion. Recall $\Xi_t := \Psi_{\alpha,\mathcal{D}}(w_t, v_t) + \Psi_{\alpha,\mathcal{D}^{(i)}}(w'_t, v'_t)$ and $D_t := \|w_t - w'_t\| + \|v_t - v'_t\|$. Apply the one-run bound from part (b) to $\mathcal{D}$ and to $\mathcal{D}^{(i)}$ (same stepsize; shared index i_t), then sum:

$$\begin{aligned}\mathbb{E}[\Xi_{t+1} \mid \mathcal{F}_t] \le{}& (1 - c\eta_t)\,\Xi_t \;+\; A\,\eta_t^2 \;+\; C_{\mathrm{lin}}G^2\,\eta_t \\ &+\; \underbrace{\eta_t\Big(\big\langle \nabla\Phi_{\mathcal{D}}(w_t) - \nabla\Phi_{\mathcal{D}^{(i)}}(w'_t),\, g_t^w\big\rangle \;-\; \big\langle \nabla_w F_{\mathcal{D}}(w_t, v_t) - \nabla_w F_{\mathcal{D}^{(i)}}(w'_t, v'_t),\, g_t^w\big\rangle\Big)}_{=:\,\mathsf{Leak}_t^{(w)}} \\ &+\; \eta_t\Big(\mathbf{1}\{i_t = i\} + \mathbf{1}\{\hat{i}_t = i\}\Big)\,\tilde{B}_0\,G^2.\end{aligned} \tag{13}$$

Bounding the leak term by Cauchy–Schwarz, joint L-smoothness, and Lemma 10 gives

$$\begin{aligned}|\mathsf{Leak}_t^{(w)}| &= \eta_t\Big|\big\langle \nabla\Phi_{\mathcal{D}}(w_t) - \nabla\Phi_{\mathcal{D}^{(i)}}(w'_t),\, g_t^w\big\rangle - \big\langle \nabla_w F_{\mathcal{D}}(w_t, v_t) - \nabla_w F_{\mathcal{D}^{(i)}}(w'_t, v'_t),\, g_t^w\big\rangle\Big| \\ &\le \eta_t\Big(\big|\big\langle \nabla\Phi_{\mathcal{D}}(w_t) - \nabla\Phi_{\mathcal{D}^{(i)}}(w'_t),\, g_t^w\big\rangle\big| + \big|\big\langle \nabla_w F_{\mathcal{D}}(w_t, v_t) - \nabla_w F_{\mathcal{D}^{(i)}}(w'_t, v'_t),\, g_t^w\big\rangle\big|\Big) \\ &\le \eta_t\Big(\big\|\nabla\Phi_{\mathcal{D}}(w_t) - \nabla\Phi_{\mathcal{D}^{(i)}}(w'_t)\big\|\|g_t^w\| + \big\|\nabla_w F_{\mathcal{D}}(w_t, v_t) - \nabla_w F_{\mathcal{D}^{(i)}}(w'_t, v'_t)\big\|\|g_t^w\|\Big) \\ &\qquad\qquad \text{(Cauchy-Schwarz)} \\ &\le \eta_t\Big(L\|w_t - w'_t\|\|g_t^w\| + L_\Phi\|w_t - w'_t\|\|g_t^w\| + 2L\|v_t - v'_t\|\|g_t^w\|\Big) \\ &\qquad\qquad \text{(Joint } L\text{-smoothness of gradients)} \\ &\le \eta_t G\Big((2L + 2L_\Phi)\|w_t - w'_t\| + 2L\|v_t - v'_t\| + \frac{2G}{n}\left(2 + \frac{L}{\rho}\right)\Big). \qquad \text{(Lemma 10)}\end{aligned}$$

hence, with $D_t = \|w_t - w'_t\| + \|v_t - v'_t\|$,

$$\mathsf{Leak}_t := \mathsf{Leak}_t^{(w)} \;\le\; C_{\mathrm{leak}}\,G\,\eta_t\,D_t \;+\; C_*\,\frac{G^2}{n}\,\eta_t, \qquad C_{\mathrm{leak}} := 2(2L + L_\Phi), \quad C_* := 4\Big(2 + \frac{L}{\rho}\Big).$$

Therefore

$$\mathbb{E}[\Xi_{t+1} \mid \mathcal{F}_t] \;\le\; (1-c\eta_t)\,\Xi_t + A\,\eta_t^2 + C_{\text{leak}}\,G\,\eta_t\,D_t + C_*\,\tfrac{G^2}{n}\,\eta_t + C_{\text{lin}}\,G^2\,\eta_t + \eta_t\Big(\mathbf{1}\{i_t = i\}+\mathbf{1}\{\hat{i}_t = i\}\Big)\tilde{B}_0 G^2, \tag{14}$$

which matches (6) up to explicit constants.

Weak distance recursion. A single SGDA step gives

$$\|w_{t+1} - w'_{t+1}\| \le \|w_t - w'_t\| + \eta_t\,\|\nabla_w f(w_t, v_t; z_{i_t}) - \nabla_w f(w'_t, v'_t; z_{i_t})\| + \eta_t\,\mathbf{1}\{i_t = i\}\cdot 2G,$$

and similarly for v (with ascent). Joint L-smoothness yields

$$\|\nabla_w f(w_t, v_t; z) - \nabla_w f(w'_t, v'_t; z)\| \le L(\|w_t - w'_t\| + \|v_t - v'_t\|), \quad \|\nabla_v f(\cdot) - \nabla_v f(\cdot)\| \le L(\|w_t - w'_t\| + \|v_t - v'_t\|).$$

Hence

$$\mathbb{E}[D_{t+1} \mid \mathcal{F}_t] \;\le\; (1 + \kappa\eta_t)\,D_t \;+\; \eta_t\,\mathbf{1}\{i_t = i\}\,\tilde{B}_1 G \;+\; \eta_t\,\Upsilon_t,$$

with $\kappa := 2L$ and $\tilde{B}_1 := 8$.

To bound the term Υ_t, which represents the expected difference in stochastic gradients, we first analyze the difference of the full-batch gradients. We decompose this difference for the primal and dual variables separately.

For the primal variable, we use the triangle inequality to introduce the value function Φ and the mismatch term $\Delta_t := \nabla_w F_{\mathcal{D}}(w_t, v_t) - \nabla\Phi_{\mathcal{D}}(w_t)$:

$$\|\nabla_w F_{\mathcal{D}}(w_t, v_t) - \nabla_w F_{\mathcal{D}^{(i)}}(w'_t, v'_t)\| \le \underbrace{\|\nabla\Phi_{\mathcal{D}}(w_t) - \nabla\Phi_{\mathcal{D}^{(i)}}(w'_t)\|}_{\le L_\Phi\|w_t - w'_t\| + \frac{2G}{n}(1+\frac{L}{\rho}) \text{ by Lem. 10}} + \underbrace{\|\Delta_t\|}_{\le L\sqrt{2\Gamma_t/\rho} \text{ by Lem. 8}} + \underbrace{\|\Delta'_t\|}_{\le L\sqrt{2\Gamma'_t/\rho} \text{ by Lem. 8}}$$

For the dual variable, we use joint L-smoothness and the single-sample replacement identity:

$$\begin{aligned}\|\nabla_v F_{\mathcal{D}}(w_t, v_t) - \nabla_v F_{\mathcal{D}^{(i)}}(w'_t, v'_t)\| &\le \|\nabla_v F_{\mathcal{D}}(w_t, v_t) - \nabla_v F_{\mathcal{D}}(w'_t, v'_t)\| + \|\nabla_v F_{\mathcal{D}}(w'_t, v'_t) - \nabla_v F_{\mathcal{D}^{(i)}}(w'_t, v'_t)\| \\ &\le L(\|w_t - w'_t\| + \|v_t - v'_t\|) + \frac{2G}{n} = LD_t + \frac{2G}{n}.\end{aligned}$$

Combining the bounds on the primal and dual components gives a complete bound on the full-batch gradient difference. Summing the two inequalities and using $D_t \ge \|w_t - w'_t\|$, we have:

$$\|\nabla F_{\mathcal{D}}(w_t, v_t) - \nabla F_{\mathcal{D}^{(i)}}(w'_t, v'_t)\|_1 \le (L_\Phi + L)D_t + L\sqrt{2/\rho}(\sqrt{\Gamma_t} + \sqrt{\Gamma'_t}) + \frac{2G}{n}\left(2 + \frac{L}{\rho}\right).$$

We then convert the geometric quantities on the right-hand side (D_t and Γ_t) into the Lyapunov potential Ξ_t. First, using the triangle inequality along with Lemmas 11 and 12, we bound the distance D_t:

$$\begin{aligned}D_t &= \|w_t - w'_t\| + \|v_t - v'_t\| \\ &\le (\|w_t - x^\star_D\| + \|v_t - v^\star_D\|) + (\|x^\star_D - x^\star_{D^{(i)}}\| + \|v^\star_D - v^\star_{D^{(i)}}\|) + (\|x^\star_{D^{(i)}} - w'_t\| + \|v^\star_{D^{(i)}} - v'_t\|) \\ &\le C_{\text{dist}}\sqrt{\Psi_{\alpha,\mathcal{D}}(w_t, v_t)} + \frac{S_{\text{sens}}}{n} + C_{\text{dist}}\sqrt{\Psi_{\alpha,\mathcal{D}^{(i)}}(w'_t, v'_t)} \\ &\le C_{\text{dist}}\left(\sqrt{\Psi_{\alpha,\mathcal{D}}} + \sqrt{\Psi_{\alpha,\mathcal{D}^{(i)}}}\right) + \frac{S_{\text{sens}}}{n} \le \sqrt{2}C_{\text{dist}}\sqrt{\Xi_t} + \frac{S_{\text{sens}}}{n}.\end{aligned}$$

Second, from the definition of the potential, we have $\sqrt{\Gamma_t} + \sqrt{\Gamma'_t} \le 2\sqrt{\Xi_t/\alpha}$. Substituting these into our main inequality gives:

$$\|\nabla F_{\mathcal{D}}(w_t, \cdot) - \nabla F_{\mathcal{D}^{(i)}}(w'_t, \cdot)\|_1 \le (L_\Phi + L)\left(\sqrt{2}C_{\text{dist}}\sqrt{\Xi_t} + \frac{S_{\text{sens}}}{n}\right) + L\sqrt{2/\rho}\left(2\sqrt{\Xi_t/\alpha}\right) + O(G/n).$$

The terms involving $\sqrt{\Xi_t}$ can be collected and absorbed into a single term $a\sqrt{\Xi_t}$, where a is a constant that depends on the problem parameters (e.g., $a := 2C_{\text{dist}}(L + L_\Phi)$ serves as a valid, convenient

upper bound). The remaining terms are of order $O(1/n)$. Therefore, we state the final bound on Υ_t as:

$$\Upsilon_t \;\le\; 2C_{\mathrm{dist}}(L+L_\Phi)\,\sqrt{\Xi_t} \;+\; \frac{2G}{n}\Big(1+\frac{L}{\rho}\Big),$$

and therefore

$$\mathbb{E}[D_{t+1}\mid\mathcal{F}_t] \;\le\; (1+\kappa\eta_t)\,D_t + a\,\eta_t\,\sqrt{\Xi_t} + \eta_t\,\mathbf{1}\{i_t=i\}\,\tilde{B}_1 G + \eta_t\,\frac{2G}{n}\Big(1+\frac{L}{\rho}\Big), \qquad a := 2C_{\mathrm{dist}}(L+L_\Phi). \tag{15}$$

□

Lemma 14 (One-step ascent for L-smooth, ρ-strongly concave g)**.** *Let $g:\mathbb{R}^d\to\mathbb{R}$ be L-smooth and ρ-strongly concave, and let $v^+ = v+\gamma\,\nabla g(v)$ with $\gamma\ge 0$. Denote $\theta(v) := g(v^\star)-g(v)$, where $v^\star\in\arg\max g$. Then*

$$\theta(v^+) \;\le\; \big(1-2\rho\gamma+\rho L\gamma^2\big)\,\theta(v).$$

Consequently,

$$g(v^\star)-g(v^+) \;\le\; \big(1-2\rho\gamma+\rho L\gamma^2\big)\,[g(v^\star)-g(v)].$$

Proof. For any L-smooth differentiable function h one has

$$h(y) \;\ge\; h(x)+\langle\nabla h(x),\,y-x\rangle - \tfrac{L}{2}\|y-x\|^2.$$

Apply this to $h=g$, $x=v$, $y=v^+=v+\gamma\nabla g(v)$:

$$g(v^+) \;\ge\; g(v)+\gamma\|\nabla g(v)\|^2 - \tfrac{L}{2}\gamma^2\|\nabla g(v)\|^2.$$

Rearranging gives

$$g(v^\star)-g(v^+) \;\le\; g(v^\star)-g(v) \;-\; \big(\gamma-\tfrac{L}{2}\gamma^2\big)\,\|\nabla g(v)\|^2. \tag{16}$$

Since g is ρ-strongly concave, $f:=-g$ is ρ-strongly convex; hence f satisfies the Polyak–Łojasiewicz (PL) inequality

$$\tfrac{1}{2}\|\nabla f(x)\|^2 \;\ge\; \rho\big(f(x)-f^\star\big),$$

with the *same* constant ρ [Karimi et al., 2016]. Translating back to g gives

$$\tfrac{1}{2}\|\nabla g(v)\|^2 \;\ge\; \rho\,\big(g(v^\star)-g(v)\big) \;=\; \rho\,\theta(v). \tag{17}$$

Step 3 (Combine). Insert (17) into (16):

$$\theta(v^+) \;\le\; \theta(v) \;-\; 2\rho\gamma\Big(1-\tfrac{L}{2}\gamma\Big)\,\theta(v) \;=\; \big(1-2\rho\gamma+\rho L\gamma^2\big)\,\theta(v),$$

□

Lemma 15 (Damping the weak distance recursion)**.** *Let $S_t := \sum_{s=0}^{t-1}\eta_s$ and define the damped distance $\widetilde{D}_t := e^{-2LS_t}D_t$ from* (7)*. Set $\lambda := 2L$ for this lemma. Then, taking expectations and averaging over i,*

$$\bar{\widetilde{D}}_{t+1} \;\le\; \bar{\widetilde{D}}_t \;+\; a\,\eta_t\,e^{-2LS_{t+1}}\mathbb{E}\big[\sqrt{\bar{\Xi}_t}\big] \;+\; \frac{1}{n}\left(2\tilde{B}_1 G + 2G\Big(1+\frac{L}{\rho}\Big)\right)\eta_t\,e^{-\lambda S_{t+1}}.$$

Consequently, summing from $t=0$ to $T-1$,

$$\bar{D}_T \;\le\; a\sum_{t=0}^{T-1}\eta_t\,e^{-2L\sum_{s=t+1}^{T-1}\eta_s}\,\mathbb{E}\big[\sqrt{\bar{\Xi}_t}\big] \;+\; \frac{1}{\lambda n}\left(2\tilde{B}_1 G + 2G\Big(1+\frac{L}{\rho}\Big)\right). \tag{18}$$

Proof. Start from the weak distance recursion (7):

$$\mathbb{E}[D_{t+1}\mid\mathcal{F}_t] \;\le\; (1+2L\eta_t)D_t + a\,\eta_t\,\sqrt{\Xi_t} + \eta_t\,\mathbf{1}\{i_t=i\}\,\tilde{B}_1\,G + \eta_t\,\frac{2G}{n}\Big(1+\frac{L}{\rho}\Big).$$

Multiply both sides by $e^{-2LS_{t+1}} = e^{-2L(S_t+\eta_t)}$ to get

$$e^{-2LS_{t+1}}\mathbb{E}[D_{t+1} \mid \mathcal{F}_t] \;\le e^{-2LS_t}\big(1+2L\eta_t\big)e^{-2L\eta_t}D_t \;+\; a\,\eta_t\,e^{-2LS_{t+1}}\sqrt{\Xi_t} \\ +\; \eta_t\,e^{-2LS_{t+1}}\,\mathbf{1}\{i_t=i\}\,\tilde{B}_1 G \;+\; \eta_t\,e^{-2LS_{t+1}}\,\frac{2G}{n}\Big(1+\frac{L}{\rho}\Big).$$

Since $(1+x)e^{-x} \le 1$ for all $x \ge 0$, the first term is $\le e^{-2LS_t}D_t = \widetilde{D}_t$. Taking total expectation and then averaging over the replacement index i, we make explicit that the hit $\mathbf{1}\{i_t = i\}$ contributes to *both* coordinates in $D_t = \|w_t - w_t'\| + \|v_t - v_t'\|$. For a fixed i,

$$\mathbb{E}[D_{t+1} \mid \mathcal{F}_t] \le (1+2L\eta_t)D_t \;+\; a\,\eta_t\,\sqrt{\Xi_t} \;+\; \eta_t\,\mathbf{1}\{i_t=i\}\,c_w G \;+\; \eta_t\,\mathbf{1}\{i_t=i\}\,c_v G + \eta_t\,\frac{2G}{n}\Big(1+\frac{L}{\rho}\Big) \\ \le (1+2L\eta_t)D_t \;+\; a\,\eta_t\,\sqrt{\Xi_t} \;+\; \eta_t\,\mathbf{1}\{i_t=i\}\,(2\tilde{B}_1)G + \eta_t\,\frac{2G}{n}\Big(1+\frac{L}{\rho}\Big),$$

where we bundle constants so that $c_w = c_v = \tilde{B}_1$. Multiplying both sides by $e^{-2LS_{t+1}}$ with $S_t := \sum_{s=0}^{t-1}\eta_s$ and using $(1+2L\eta_t)e^{-2L\eta_t} \le 1$ yields

$$\mathbb{E}\Big[\widetilde{D}_{t+1} \mid \mathcal{F}_t\Big] \;\le\; \widetilde{D}_t + a\,\eta_t\,e^{-2LS_{t+1}}\,\sqrt{\Xi_t} + \eta_t\,e^{-2LS_{t+1}}\,\mathbf{1}\{i_t=i\}\,(2\tilde{B}_1)G + \eta_t e^{-2LS_{t+1}}\frac{2G}{n}\Big(1+\frac{L}{\rho}\Big),$$

where $\widetilde{D}_t := e^{-2LS_t}D_t$. Taking total expectation and then averaging over i (so that $\mathbb{E}[\mathbf{1}\{i_t=i\}] = 1/n$) gives

$$\bar{\widetilde{D}}_{t+1} \;\le\; \bar{\widetilde{D}}_t \;+\; a\,\eta_t\,e^{-2LS_{t+1}}\,\mathbb{E}\big[\sqrt{\Xi_t}\big] \;+\; \frac{1}{n}\left(2\tilde{B}_1 G + 2G\Big(1+\frac{L}{\rho}\Big)\right)\eta_t\,e^{-2LS_{t+1}}.$$

Summing the inequality from $t=0$ to $T-1$ and noting that both runs start from the same initialization $(w_0, v_0) = (w_0', v_0')$, we have $D_0 = 0$ and hence $\widetilde{D}_0 = 0$ and $\bar{\widetilde{D}}_0 = 0$. Therefore,

$$\bar{\widetilde{D}}_T - \bar{\widetilde{D}}_0 \;=\; \bar{\widetilde{D}}_T \;\le\; a\sum_{t=0}^{T-1}\eta_t e^{-2LS_{t+1}}\,\mathbb{E}\big[\sqrt{\Xi_t}\big] + \frac{1}{n}\left(2\tilde{B}_1 G + 2G\Big(1+\frac{L}{\rho}\Big)\right)\sum_{t=0}^{T-1}\eta_t e^{-2LS_{t+1}}.$$

Applying Lemma 16 with $\gamma = \lambda = 2L$ to the last sum and noting that $e^{2LS_T}e^{-2LS_{t+1}} = e^{2L\sum_{s=t+1}^{T-1}\eta_s}$ yields

$$\bar{D}_T \;=\; e^{2LS_T}\bar{\widetilde{D}}_T \;\le\; a\sum_{t=0}^{T-1}\eta_t e^{-2L\sum_{s=t+1}^{T-1}\eta_s}\,\mathbb{E}\big[\sqrt{\Xi_t}\big] \;+\; \frac{1}{\lambda n}\left(2\tilde{B}_1 G + 2G\Big(1+\frac{L}{\rho}\Big)\right),$$

which is (18). □

Lemma 16. *Let (η_t) satisfy the Robbins–Monro conditions. For $S_t := \sum_{s=0}^{t-1}\eta_s$ and any $\gamma > 0$, any $T \ge 1$, set $\eta_{\max,T} := \max_{0\le t<T}\eta_t$. Then*

$$\sum_{t=0}^{T-1}\eta_t\,e^{-\gamma\sum_{s=t+1}^{T-1}\eta_s} \;\le\; \frac{e^{\gamma\eta_{\max,T}}}{\gamma}\left(1-e^{-\gamma S_T}\right) \;\le\; \frac{e^{\gamma\eta_{\max,T}}}{\gamma},$$

and

$$\sum_{t=0}^{T-1}\eta_t^2\,e^{-\gamma\sum_{s=t+1}^{T-1}\eta_s} \xrightarrow[T\to\infty]{} 0.$$

In particular, under Condition (S1) and with $\gamma = 2L$, one has $\eta_{\max,T} \le 1/(4L)$ and therefore

$$\sum_{t=0}^{T-1}\eta_t\,e^{-2L\sum_{s=t+1}^{T-1}\eta_s} \;\le\; \frac{e^{1/2}}{2L}.$$

Proof. Let $S_t := \sum_{s=0}^{t-1}\eta_s$ and note $S_t \uparrow \infty$ under Robbins–Monro. For the first display, fix t and any $u \in [S_t, S_{t+1}]$. Then

$$e^{-\gamma(S_T - S_{t+1})} = e^{-\gamma(S_T-u)}\,e^{\gamma(S_{t+1}-u)} \;\le\; e^{\gamma\eta_{\max,T}}\,e^{-\gamma(S_T-u)},$$

since $0 \le S_{t+1} - u \le S_{t+1} - S_t = \eta_t \le \eta_{\max,T}$. Hence

$$\eta_t \, e^{-\gamma(S_T - S_{t+1})} \;\le\; e^{\gamma \eta_{\max,T}} \int_{S_t}^{S_{t+1}} e^{-\gamma(S_T - u)} \, \mathrm{d}u.$$

Summing over $t = 0, \dots, T-1$ and using $\sum_t \int_{S_t}^{S_{t+1}} (\cdot) = \int_0^{S_T} (\cdot)$ gives

$$\sum_{t=0}^{T-1} \eta_t \, e^{-\gamma \sum_{s=t+1}^{T-1} \eta_s} \;\le\; e^{\gamma \eta_{\max,T}} \int_0^{S_T} e^{-\gamma(S_T - u)} \, \mathrm{d}u \;=\; \frac{e^{\gamma \eta_{\max,T}}}{\gamma} \left(1 - e^{-\gamma S_T}\right) \;\le\; \frac{e^{\gamma \eta_{\max,T}}}{\gamma}.$$

For the second display, fix $\varepsilon > 0$. Since $\sum_t \eta_t^2 < \infty$, pick T_0 with $\sum_{t \ge T_0} \eta_t^2 < \varepsilon$. Then split

$$\sum_{t=0}^{T-1} \eta_t^2 \, e^{-\gamma \sum_{s=t+1}^{T-1} \eta_s} = \sum_{t=0}^{T_0-1} \eta_t^2 \, e^{-\gamma(S_T - S_{t+1})} + \sum_{t=T_0}^{T-1} \eta_t^2 \, e^{-\gamma(S_T - S_{t+1})} \;\le\; e^{-\gamma(S_T - S_{T_0})} \sum_{t=0}^{T_0-1} \eta_t^2 + \varepsilon.$$

Let $T \to \infty$ to send the first term to 0 (since $S_T \to \infty$), and then let $\varepsilon \downarrow 0$. Finally, under Condition (S1) and $\gamma = 2L$, we have $\eta_{\max,T} \le 1/(4L)$, hence $e^{\gamma \eta_{\max,T}} \le e^{1/2}$, which yields the stated corollary. □

Lemma 17 (Closing the potential recursion)**.** *After averaging* (6) *over i,*

$$\bar{\Xi}_{t+1} \;\le\; \Big(1 - \tfrac{c}{2}\eta_t\Big)\bar{\Xi}_t \;+\; A\,\eta_t^2 \;+\; \frac{b^2}{2c}\,\eta_t \;+\; \frac{\tilde{H}}{n}\,\eta_t, \tag{19}$$

with $b := 2C_{\mathrm{leak}} C_{\mathrm{dist}}$ *and* $\tilde{H} := 2\tilde{B}_0 G^2 + C_{\mathrm{leak}} S_{\mathrm{sens}} + C_* G$*, where*

$$S_{\mathrm{sens}} \;:=\; 2G\left(\frac{(1 + L/\rho)^2}{\sqrt{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}}} + \frac{1}{\rho}\right)$$

is the constant from Lemma 12.

Proof. Starting from (6),

$$\mathbb{E}[\Xi_{t+1} \mid \mathcal{F}_t] \;\le\; (1 - c\eta_t)\,\Xi_t + A\,\eta_t^2 + C_{\mathrm{leak}}\,\eta_t\,D_t + C_*\,\tfrac{G}{n}\,\eta_t + \eta_t\big(\mathbf{1}\{i_t = i\} + \mathbf{1}\{\hat{i}_t = i\}\big)\,\tilde{B}_0\,G^2.$$

By Lemma 11, for each run $\|w_t - x_D^\star\| + \|v_t - v_D^\star\| \le C_{\mathrm{dist}} \sqrt{\Psi_{\alpha,\mathcal{D}}(w_t, v_t)}$ and analogously for the primed run. Using the triangle inequality and Lemma 12 for the cross-dataset saddle shift, we have

$$\begin{aligned}
D_t &= \|w_t - w_t'\| + \|v_t - v_t'\| \\
&\le \|w_t - x_D^*\| + \|x_D^* - x_{D^{(i)}}^*\| + \|x_{D^{(i)}}^* - w_t'\| + \|v_t - v_D^*\| + \|v_D^* - v_{D^{(i)}}^*\| - \|v_{D^{(i)}}^* - v_t'\| \\
&\qquad\qquad \text{(Triangle inequality)} \\
&\le C_{\mathrm{dist}} \left(\sqrt{\Psi_{\alpha,\mathcal{D}}(w_t, v_t)} + \sqrt{\Psi_{\alpha,\mathcal{D}^{(i)}}(w_t', v_t')}\right) + \|x_D^* - x_{D^{(i)}}^*\| - \|v_{D^{(i)}}^* - v_t'\| \\
&\qquad\qquad \text{(Lemma 12)} \\
&\le C_{\mathrm{dist}} \left(\sqrt{\Psi_{\alpha,\mathcal{D}}(w_t, v_t)} + \sqrt{\Psi_{\alpha,\mathcal{D}^{(i)}}(w_t', v_t')}\right) + \frac{S_{\mathrm{sens}}}{n} \qquad \text{(Lemma 11)}
\end{aligned}$$

Since $\sqrt{\Psi_{\alpha,\mathcal{D}}} \le \sqrt{\Xi_t}$ and $\sqrt{\Psi_{\alpha,\mathcal{D}^{(i)}}} \le \sqrt{\Xi_t}$,

$$D_t \;\le\; 2C_{\mathrm{dist}} \sqrt{\Xi_t} \;+\; \frac{S_{\mathrm{sens}}}{n}.$$

Plug this into (6):

$$\mathbb{E}[\Xi_{t+1} \mid \mathcal{F}_t] \;\le\; (1 - c\eta_t)\,\Xi_t + A\,\eta_t^2 + \underbrace{2C_{\mathrm{leak}} C_{\mathrm{dist}}}_{=:b}\,\eta_t\,\sqrt{\Xi_t} + \frac{C_{\mathrm{leak}} S_{\mathrm{sens}}}{n}\,\eta_t + C_*\,\tfrac{G}{n}\,\eta_t + \eta_t\big(\mathbf{1}\{i_t = i\} + \mathbf{1}\{\hat{i}_t = i\}\big)\,\tilde{B}_0\,G^2.$$

Apply the inequality $uv \le \frac{\gamma}{2}u^2 + \frac{1}{2\gamma}v^2$ with $u = \sqrt{\eta_t \Xi_t}$, $v = b\sqrt{\eta_t}$, and $\gamma = c$ to the mixed term:

$$b\,\eta_t\,\sqrt{\Xi_t} \;\le\; \frac{c}{2}\,\eta_t\,\Xi_t \;+\; \frac{b^2}{2c}\,\eta_t.$$

Therefore,

$$\mathbb{E}[\Xi_{t+1} \mid \mathcal{F}_t] \;\le\; \Big(1-\tfrac{c}{2}\eta_t\Big)\Xi_t + A\,\eta_t^2 + \frac{b^2}{2c}\,\eta_t + \frac{C_{\mathrm{leak}} S_{\mathrm{sens}}}{n}\,\eta_t + C_*\,\tfrac{G}{n}\,\eta_t + \eta_t\big(\mathbf{1}\{i_t = i\} + \mathbf{1}\{\hat{i}_t = i\}\big)\,\tilde{B}_0\,G^2.$$

Taking total expectation and averaging over i (both indicators have mean $1/n$) gives

$$\bar{\Xi}_{t+1} \;\le\; \Big(1 - \tfrac{c}{2}\eta_t\Big)\bar{\Xi}_t \;+\; A\,\eta_t^2 \;+\; \frac{b^2}{2c}\,\eta_t \;+\; \frac{1}{n}\Big(2\tilde{B}_0 G^2 + C_{\mathrm{leak}} S_{\mathrm{sens}} + C_* G\Big)\eta_t.$$

With $\tilde{H} := 2\tilde{B}_0 G^2 + C_{\mathrm{leak}} S_{\mathrm{sens}} + C_* G$ and $b = 2C_{\mathrm{leak}} C_{\mathrm{dist}}$ this is (19). □

Lemma 18 (Bounded weighted sum of potentials)**.** *For any $\theta \in (\frac{1}{2}, 1)$ and $S_t := \sum_{s=0}^{t-1} \eta_s$,*

$$\sum_{t=0}^{T-1} e^{-\theta c \sum_{s=t+1}^{T-1} \eta_s}\,\bar{\Xi}_t \;\le\; \frac{2}{1 - e^{-(\theta-\frac{1}{2})c\,\eta_{\min,T}}}\, e^{-\frac{c}{2}\sum_{s=0}^{T-1}\eta_s}\,\bar{\Xi}_0$$
$$+\; \frac{2}{1 - e^{-(\theta-\frac{1}{2})c\,\eta_{\min,T}}}\, A \sum_{t=0}^{T-1} \eta_t^2\, e^{-\frac{c}{2}\sum_{s=t+1}^{T-1}\eta_s} \;+\; \frac{8}{c\big(1 - e^{-(\theta-\frac{1}{2})c\,\eta_{\min,T}}\big)}\left(\frac{b^2}{2c} + \frac{\tilde{H}}{n}\right).$$

In particular, since $c\,\eta_{\min,T} \le \frac{1}{2}$ (by Condition (S1)) and hence $1 - e^{-x} \ge x/2$ for $x \in [0, \frac{1}{2}]$, we have

$$\sum_{t=0}^{T-1} e^{-\theta c \sum_{s=t+1}^{T-1} \eta_s}\,\bar{\Xi}_t \;\le\; \frac{4}{(\theta - \frac{1}{2})c\,\eta_{\min,T}}\, e^{-\frac{c}{2}\sum_{s=0}^{T-1}\eta_s}\,\bar{\Xi}_0$$
$$+\; \frac{4}{(\theta - \frac{1}{2})c\,\eta_{\min,T}}\, A \sum_{t=0}^{T-1} \eta_t^2\, e^{-\frac{c}{2}\sum_{s=t+1}^{T-1}\eta_s} \;+\; \frac{16}{(\theta - \frac{1}{2})c^2\,\eta_{\min,T}}\left(\frac{b^2}{2c} + \frac{\tilde{H}}{n}\right).$$

Proof. Start from (19) and unroll using $1 - \frac{c}{2}\eta_k \le e^{-\frac{c}{2}\eta_k}$:

$$\bar{\Xi}_t \;\le\; e^{-\frac{c}{2}S_t}\bar{\Xi}_0 + \sum_{k=0}^{t-1} e^{-\frac{c}{2}(S_t - S_{k+1})}\big(A\eta_k^2 + q\,\eta_k\big), \qquad q := \tfrac{b^2}{2c} + \tfrac{\tilde{H}}{n}.$$

Multiply by the weight $w_t := e^{-\theta c(S_T - S_{t+1})}$ and sum over $t = 0, \dots, T-1$:

$$\sum_{t=0}^{T-1} w_t \bar{\Xi}_t \;\le\; \underbrace{\sum_{t=0}^{T-1} w_t e^{-\frac{c}{2}S_t}}_{\mathsf{I}}\,\bar{\Xi}_0 + \underbrace{A\sum_{t=0}^{T-1}\sum_{k=0}^{t-1} w_t e^{-\frac{c}{2}(S_t - S_{k+1})}\eta_k^2}_{\mathsf{II}} + \underbrace{q\sum_{t=0}^{T-1}\sum_{k=0}^{t-1} w_t e^{-\frac{c}{2}(S_t - S_{k+1})}\eta_k}_{\mathsf{III}}.$$

Throughout, Condition (S1) implies $c\,\eta_t \le \frac{1}{2}$ (since $c \le \rho/2$ and $\eta_t \le 1/\rho$), hence $1 - e^{-x} \ge x/2$ for $x \in [0, \frac{1}{2}]$.

Claim A. For $\theta \in (\frac{1}{2}, 1)$,

$$\mathsf{I} = \sum_{t=0}^{T-1} e^{-\theta c(S_T - S_{t+1})}\, e^{-\frac{c}{2}S_t} \;\le\; \frac{2\,e^{-\frac{c}{2}S_T}}{1 - \exp\big(-(\theta-\frac{1}{2})c\,\eta_{\min,T}\big)} \;\le\; \frac{4}{(\theta - \frac{1}{2})c\,\eta_{\min,T}}\, e^{-\frac{c}{2}S_T},$$

where $\eta_{\min,T} := \min_{0 \le t < T} \eta_t$.

Proof of Claim A. Using $S_t = S_{t+1} - \eta_t$ and $S_T = S_{t+1} + (S_T - S_{t+1})$,

$$e^{-\theta c(S_T - S_{t+1})}\, e^{-\frac{c}{2}S_t} = e^{-\frac{c}{2}S_T}\, e^{-(\theta-\frac{1}{2})c(S_T - S_{t+1})}\, e^{\frac{c}{2}\eta_t}.$$

Since $c\,\eta_t \le \frac{1}{2}$, $e^{\frac{c}{2}\eta_t} \le e^{1/4} \le 2$. Therefore,

$$\mathsf{I} \le 2\,e^{-\frac{c}{2}S_T} \sum_{t=0}^{T-1} e^{-(\theta-\frac{1}{2})c(S_T - S_{t+1})}.$$

Because $S_T - S_{t+1} \geq (T-1-t)\,\eta_{\min,T}$, the sum is bounded by a geometric series, giving the first display; the second follows from $1 - e^{-x} \geq x/2$ for $x \in [0,1]$. $\triangle$

Claim B. Let $\alpha := (\theta - \frac{1}{2})c > 0$. For any fixed $k \in \{0, \ldots, T-2\}$,

$$\sum_{t=k+1}^{T-1} w_t\, e^{-\frac{c}{2}(S_t - S_{k+1})} \;\leq\; \frac{2}{1 - e^{-\alpha\,\eta_{\min,T}}}\; e^{-\frac{c}{2}(S_T - S_{k+1})} \;\leq\; \frac{4}{(\theta - \frac{1}{2})c\,\eta_{\min,T}}\; e^{-\frac{c}{2}(S_T - S_{k+1})}.$$

Proof of Claim B. As above,

$$w_t\, e^{-\frac{c}{2}(S_t - S_{k+1})} = e^{-\frac{c}{2}(S_T - S_{k+1})}\, e^{-\alpha(S_T - S_{t+1})}\, e^{\frac{c}{2}\eta_t} \;\leq\; 2\, e^{-\frac{c}{2}(S_T - S_{k+1})}\, e^{-\alpha(S_T - S_{t+1})}.$$

Let $I_t := \int_{S_{t+1}}^{S_{t+2}} e^{-\alpha(S_T - u)}\,\mathrm{d}u$. A direct computation gives

$$I_t = \frac{1}{\alpha}\, e^{-\alpha(S_T - S_{t+1})}\big(1 - e^{-\alpha\eta_{t+1}}\big) \quad \Rightarrow \quad e^{-\alpha(S_T - S_{t+1})} = \frac{\alpha}{1 - e^{-\alpha\eta_{t+1}}}\, I_t \leq \frac{\alpha}{1 - e^{-\alpha\eta_{\min,T}}}\, I_t,$$

since $\eta_{t+1} \geq \eta_{\min,T}$. Summing over $t = k+1, \ldots, T-1$ and using $\sum_{t=k+1}^{T-1} I_t = \int_{S_{k+1}}^{S_T} e^{-\alpha(S_T - u)}\,\mathrm{d}u \leq 1/\alpha$ yields the claim; the explicit bound uses $1 - e^{-x} \geq x/2$ with $x = \alpha\eta_{\min,T} \leq \frac{1}{2}$. $\triangle$

With Claim A and $1 - e^{-x} > x/2$,

$$\mathsf{I} \;\leq\; \frac{2}{1 - e^{-(\theta - \frac{1}{2})c\,\eta_{\min,T}}}\, e^{-\frac{c}{2}S_T},$$

and, applying Claim B for each fixed k inside the double sum and then summing in t,

$$\begin{aligned}
\mathsf{II} &= \sum_{t=0}^{T-1} \sum_{k=0}^{t-1} w_t e^{-\frac{c}{2}(S_t - S_{k+1})} \eta_k^2 \\
&= \sum_{k=0}^{T-2} \sum_{t=k+1}^{T-1} w_t e^{-\frac{c}{2}(S_t - S_{k+1})} \eta_k^2 && \text{(Fubini)} \\
&\leq \sum_{k=0}^{T-2} \frac{2}{1 - e^{-\alpha\,\eta_{\min,T}}}\; e^{-\frac{c}{2}(S_T - S_{k+1})} \eta_k^2 && \text{(Claim B)}
\end{aligned}$$

For III, combine Claim B with the kernel bound from Lemma 16

$$\sum_{k=0}^{T-2} \eta_k\, e^{-\frac{c}{2}(S_T - S_{k+1})} \;\leq\; \frac{2\, e^{\frac{c}{2}\eta_{\max,T}}}{c} \;\leq\; \frac{4}{c} \qquad (\text{since } c\,\eta_{\max,T} \leq \tfrac{1}{2}),$$

to obtain

$$\mathsf{III} \;\leq\; \frac{8}{c\big(1 - e^{-(\theta - \frac{1}{2})c\,\eta_{\min,T}}\big)}\; q.$$

Putting the three pieces together and reindexing the middle sum gives the first displayed bound in the lemma. Putting the three pieces together and reindexing the middle sum gives the first displayed bound in the lemma. The "in particular" version follows by $1 - e^{-x} \geq x/2$ with $x = (\theta - \frac{1}{2})c\,\eta_{\min,T} \leq \frac{1}{2}$. $\square$

B.2 Proof of Main Result

Proof of Theorem 3.
From Lemma 17, for $c := \min\{\mu_{\mathrm{PL}}/2, \rho/2\}$ and every t,

$$\bar{\Xi}_{t+1} \;\leq\; \Big(1 - \tfrac{c}{2}\eta_t\Big)\bar{\Xi}_t \;+\; A\,\eta_t^2 \;+\; \frac{b^2}{2c}\,\eta_t \;+\; \frac{\tilde{H}}{n}\,\eta_t, \tag{20}$$

with the constants $A, b, \tilde{H}$ defined in Lemma 17. Unrolling (20) and using $1 - \frac{c}{2}\eta_k \le e^{-\frac{c}{2}\eta_k}$ yields

$$\bar{\Xi}_T \;\le\; e^{-\frac{c}{2}\sum_{s=0}^{T-1}\eta_s}\,\bar{\Xi}_0 \;+\; A\sum_{t=0}^{T-1}\eta_t^2\, e^{-\frac{c}{2}\sum_{s=t+1}^{T-1}\eta_s} \;+\; e^{\frac{c}{2}\eta_{\max,T}}\left(\frac{b^2}{c^2} \;+\; \frac{2\tilde{H}}{cn}\right), \qquad (\star)$$

where $\eta_{\max,T} := \max_{0\le t<T}\eta_t$.

Lemma 15 (obtained from Lemma 13 ($\mathrm{D}_{\mathrm{weak}}$)) gives

$$\bar{D}_T - \frac{C_{\mathrm{hit}}}{n} \le a\cdot \underbrace{\Big(\sum_{t=0}^{T-1}\eta_t^2\, e^{-2(2L-\frac{c}{2})\sum_{s=t+1}^{T-1}\eta_s}\Big)^{1/2}}_{\sqrt{S_u}} \cdot \underbrace{\Big(\sum_{t=0}^{T-1} e^{-c\sum_{s=t+1}^{T-1}\eta_s}\,\mathbb{E}[\bar{\Xi}_t]\Big)^{1/2}}_{\sqrt{S_v}} \qquad (21)$$

where $C_{\mathrm{hit}} := \frac{1}{\lambda}\left(2\tilde{B}_1 G + 2G\Big(1+\frac{L}{\rho}\Big)\right)$. The potential sum, S_v, is bounded by Lemma 18. To combine the terms into the final compact form, we bound the products that arise after substitution. Since S_u is a convergent sum, it is bounded by a constant,

$$C_S \;:=\; \sum_{t=0}^{\infty}\eta_t^2,$$

which depends only on the stepsize schedule $\{\eta_t\}$. Under the harmonic rule $\eta_t = \frac{c_1}{c_2+t}$, $C_S = c_1^2\sum_{t=0}^{\infty}(c_2+t)^{-2} \le c_1^2\frac{\pi^2}{6}$, and for $c_2 > 1$ one may also use $C_S \le \frac{c_1^2}{c_2-1}$. Moreover $S_u \to 0$ as $T\to\infty$, while S_v need not vanish; in the product we will use only the decaying initialization term and the variance component

$$S_v^{\mathrm{var}}(T) \;:=\; \sum_{t=0}^{T-1}\eta_t^2\, e^{-\frac{c}{2}\sum_{s=t+1}^{T-1}\eta_s}.$$

This allows for two key bounds:

(i) The product involving the term $e^{-\frac{c}{2}\sum_{s=0}^{T-1}\eta_s}\bar{\Xi}_0$ is bounded by absorbing S_u into the constant: $S_u\cdot e^{-\frac{c}{2}\sum\eta_s}\,\Psi_{\alpha,0}^{\max} \le C_S\cdot e^{-\frac{c}{2}\sum\eta_s}\,\Psi_{\alpha,0}^{\max}$.

(ii) The product of variance sums, $S_u\cdot S_v^{\mathrm{var}}(T)$, is bounded by a multiple of their sum: $S_u\cdot S_v^{\mathrm{var}}(T) \le C_S\cdot S_v^{\mathrm{var}}(T) \le C_S\cdot\big(S_u + S_v^{\mathrm{var}}(T)\big)$.

Substituting these bounds, grouping all constants (a^2, C_S, etc.) into C_{var}, and relaxing the exponent[8] in S_u by defining $\kappa = \min\{\frac{3c}{4},\, 2L-\frac{c}{2}\}$ yields the compact bound:

$$\bar{D}_T \;\le\; 2\,C_{\mathrm{dist}}\sqrt{e^{-\frac{c}{2}\sum_{s=0}^{T-1}\eta_s}\,\Psi_{\alpha,0}^{\max} \;+\; C_{\mathrm{var}}\Big(L(1+L/\rho)+\alpha\,\tfrac{L^2}{\rho}\Big)\,G^2\left[\sum_{t=0}^{T-1}\eta_t^2\, e^{-2\kappa\sum_{s=t+1}^{T-1}\eta_s} \;+\; \sum_{t=0}^{T-1}\eta_t^2\, e^{-\frac{c}{2}\sum_{s=t+1}^{T-1}\eta_s}\right]} \qquad (22)$$
$$+\; \frac{2G}{n}\left(\frac{(1+L/\rho)^2}{\sqrt{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}}} + \frac{1}{\rho}\right) \;+\; \frac{C_{\mathrm{hit}}}{n},$$

with $\kappa = \min\{\frac{3c}{4},\, 2L-\frac{c}{2}\}$. □

[8] Since $x \mapsto e^{-\gamma x}$ is decreasing in $\gamma > 0$, replacing the larger decay parameter by the smaller $\kappa := \min\{\,2L-\theta, \frac{3}{4}c\,\}$ only increases the weighted sums, hence preserves a valid upper bound; we refer to this monotone weakening as "relaxing the exponent."

Proof of Corollary 4. Let $\eta_t = \frac{c_1}{c_2+t}$ with $c_1 > 0$, $c_2 \geq 1$, and $c_1 < \min\{\frac{1}{2\kappa}, \frac{2}{c}\}$ so that the geometric constants below are finite. Then:

$$\sum_{s=0}^{T-1} \eta_s = c_1 \sum_{s=0}^{T-1} \frac{1}{c_2+s} \;\geq\; c_1 \log\Big(\frac{c_2+T}{c_2}\Big),$$

$$e^{-\frac{c}{2}\sum_{s=0}^{T-1}\eta_s} \leq \Big(\frac{c_2}{c_2+T}\Big)^{\frac{c\,c_1}{2}}, \qquad e^{-\kappa\sum_{s=0}^{T-1}\eta_s} \leq \Big(\frac{c_2}{c_2+T}\Big)^{\kappa\,c_1},$$

$$\sum_{t=0}^{T-1} \eta_t^2\, e^{-2\kappa\sum_{s=t+1}^{T-1}\eta_s} \leq \frac{c_1^2}{\big(1-2\kappa c_1\big)\,(c_2+T)}, \qquad \sum_{t=0}^{T-1} \eta_t^2\, e^{-\frac{c}{2}\sum_{s=t+1}^{T-1}\eta_s} \leq \frac{c_1^2}{\big(1-\frac{c}{2}c_1\big)\,(c_2+T)}.$$

Substituting these into the general bound above yields the explicit finite–T rate:

$$\varepsilon = \bar{D}_T \leq 2\,C_{\mathrm{dist}} \left[\Big(\frac{c_2}{c_2+T}\Big)^{\frac{c_1 c}{2}} \Psi_{\alpha,0}^{\max} + \frac{C_{\mathrm{var}}\,c_1^2\big(L(1+L/\rho)+\alpha L^2/\rho\big)G^2}{(c_2+T)}\left(\frac{1}{1-2\kappa c_1}+\frac{1}{1-\frac{c}{2}c_1}\right)\right]^{1/2} + \frac{2G}{n}\left(\frac{(1+L/\rho)^2}{\sqrt{\mu_{\mathrm{PL}}\mu_{\mathrm{QG}}}}+\frac{1}{\rho}\right) + \frac{C_{\mathrm{hit}}}{n}. \tag{23}$$

Thus, for fixed n and feasible (c_1, c_2), the optimization/stochastic term in (23) decays at rate $O\big((c_2+T)^{-1/2}\big)$, while the initialization bias decays polynomially as $\big(\frac{c_2}{c_2+T}\big)^{\frac{c_1 c}{4}}$ inside the square root. □

Proof of Theorem 7. Write $\widehat{w} := \widehat{w}^{(T)}$, $w_{\mathrm{ERM}} := w^{\mathrm{ERM}}$ and w^* for the population minimizer. Add and subtract the empirical risks of these three parameters:

$$\mathcal{R}(\widehat{w}) - \mathcal{R}(w^*) = \underbrace{\big[\mathcal{R}(\widehat{w}) - \widehat{\mathcal{R}}_n(\widehat{w})\big]}_{\mathrm{(A)}} + \underbrace{\big[\widehat{\mathcal{R}}_n(\widehat{w}) - \widehat{\mathcal{R}}_n(w_{\mathrm{ERM}})\big]}_{\mathrm{(C)}} + \underbrace{\big[\widehat{\mathcal{R}}_n(w_{\mathrm{ERM}}) - \widehat{\mathcal{R}}_n(w^*)\big]}_{\mathrm{(D)}\,\leq 0} + \underbrace{\big[\widehat{\mathcal{R}}_n(w^*) - \mathcal{R}(w^*)\big]}_{\mathrm{(B)}}.$$

Theorem 6 gives

$$\mathbb{E}\big[(\mathrm{A})\big] \;\leq\; (1+L/\rho)G\varepsilon_T.$$

Because w^* is deterministic, $\mathbb{E}[(\mathrm{B})] = 0$. From Lemma 2, we have

$$(\mathrm{C}) \;=\; \widehat{\mathcal{R}}_n(\widehat{w}) - \widehat{\mathcal{R}}_n(w_{\mathrm{ERM}}) \;\leq\; \frac{d_1}{d_2+T} \tag{†}$$

$(\mathrm{D}) \leq 0$ deterministically (by definition of the ERM) and thus can be discarded when taking an upper bound.

Taking expectations and using $\mathbb{E}[(\mathrm{B})] = 0$ and (†):

$$\mathbb{E}\big[\mathcal{R}(\widehat{w}) - \mathcal{R}(w^*)\big] \;\leq\; (1+L/\rho)G\varepsilon_T \;+\; \frac{d_1}{d_2+T}.$$

□

C Further discussions

C.1 Justification of Condition (C5)

Lemma 19 (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 (which are all shown to be satisfied in Kang et al. [2025]), 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, and hence Condition (C5) in our paper holds for the AGDA iterates.

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 is exactly Condition (C5) in our paper. □

C.2 Neural network parametrization and PL/QG condition.

Width/radius and high-probability regimes for uniform constants. This subsection records the neural-network instantiation of Conditions (C1)–(C6), mainly (C2)–(C3). 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 strong concavity 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 3 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.

We work on the intersection of this PL event with the event that the SGDA iterates remain inside Ω_R; see Theorem 7 of Liu et al. [2022] for the corresponding radius/width regime for stochastic gradient methods. On this event, the softmax Bellman operator is smooth because $\eta > 0$, and the bi-conjugate BRM objective has a ρ-strongly concave inner maximization in the auxiliary block [Kang et al., 2025, Yang et al., 2020]. Since Φ_D is $\mathcal{C}^2$ on the local domain, the local PL–QG equivalence used in Section 2.2 yields a constant $\mu_{\mathrm{QG}} > 0$ such that

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

throughout the same domain, satisfying Condition (C3). 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 rest of the conditions, under the same neural network Kang et al. [2025] proved that Conditions (C1) hold (Lemma 29), (C4) is relatively obvious since f in our BRM is simply a quadratic function of ζ with negative leading term, Yang et al. [2020] proved Condition (C6) holds for the problem of minimising Equation (5). Condition (C5) is in Appendix C.1.

C.3 How index paring worked

By synchronizing the minibatch selection, we neutralize it as a source of difference between the two runs. Consequently, the parameter trajectories diverge only on the infrequent steps where the replaced index i is sampled (a "hit"). On all other steps, the updates are identical, and the optimization dynamics tend to pull the trajectories closer. The stability analysis thus bounds the cumulative effect of these rare "hits" by balancing their small, infrequent perturbations against the constant, contractive force of the optimization dynamics. This argument hinges entirely on the randomness of the sampling process, which makes the "hits" probabilistic, and not on the statistical independence of the data points, whose potential correlations are rendered irrelevant by the coupling.

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 immediately imply that the corresponding constants are identical for any dataset.

C.5 Discussions on contributions

What Kang et al. [2025] give us, and what they do not: Kang et al. [2025] establish that, after a bi-conjugate transformation, both the population and empirical BRM objectives enjoy PL-strongly-concave structure in the parameterization of Q and ζ. They then prove global convergence of SGDA to the empirical minimizer, and use PL to translate optimization error into parameter error. However, they:

- do not analyze algorithmic stability,
- do not analyze generalization (population vs empirical BRM). In fact, in the latest version of Kang et al., Lemma 28 references and uses our present result for the BRM sample complexity. Thus, on their side, we only inherit the PL result and strong-concavity constants for BRM; on our side, they use our sample-complexity result.

What Wang et al. [2022] give us, and what they do not Wang et al. [2022] develops on-average argument stability for Markov chain SGD/SGDA and shows that for convex-concave or strongly-convex-strongly-concave objectives, the excess population risk scales as $O(1/\sqrt{n})$, 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 2. We reuse from Wang et al. [2022]:

- The concept of on-average argument stability and its connection to generalization (Lemma 5 in our notation).
- Some proof templates: two-run coupling on neighboring datasets and the idea of counting "hits" of the replaced data point via an indicator $\mathbf{1}\{i_t = i\}$.

However, there are two crucial structural differences:

- Geometry: Wang et al. assume convex-concave or ρ-strongly-convex-strongly-concave objectives. 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), and only in w. The algorithm, however, is run on the bilevel saddle objective $F_D(w, v)$, not on Φ_D. This destroys the standard convexity-based distance contraction used in Wang et al.
- Object of interest: Wang et al. work directly with the original minimax F and directly analyze its risk/generalization. We instead are interested in the value function Φ_D (Bellman residual) and the primal-dual gap built on Φ_D. The gradients used by SGDA at time t are biased w.r.t. $\nabla\Phi_D(w_t)$ because the dual variable is not at its maximizer:

$$\Delta_t := \nabla_w F_D(w_t, v_t) - \nabla\Phi_D(w_t) \neq 0$$

  This mismatch term Δ_t is absent in Wang et al. and is precisely what drives the need for our Lyapunov potential.

Table 2: Comparison with prior work.

Work	Geometry	Setting	Stat. Rate	Opt. Rate
This paper	PL-strongly concave	Offline BRM	$\mathcal{O}(1/n)$	$\mathcal{O}((c_2+T)^{-\min\{1/2,\, 3cc_1/8\}})$
Wang et al. (2022)	Convex-concave	minimax	$\mathcal{O}(1/\sqrt{n})$	$\mathcal{O}(1/\sqrt{T})$

C.6 Discussion on potential extensions

Including experiments. It is not a usual practice for theoretical papers discussing stability bounds and corresponding generalization guarantees to include experiments, and therefore, not including experiments is not a weakness of this paper. Please refer to Wang et al. [2022], Bousquet and Elisseeff [2002], Charles and Papailiopoulos [2018], Feldman and Vondrak [2018].

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.

Extending Bellman residual minimization to value function estimation. 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] utilizes 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 on 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, 2024]. 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.

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 a PL–strongly-concave geometry and SGDA converges globally to an empirical saddle point. 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. This is especially relevant to offline RL, where data are often generated by trajectories and therefore need not be i.i.d. 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 $O(1/n)$ on-average argument-stability bound and the corresponding 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., 2018, 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 inherit a PL–strongly-concave structure under the neural-network parametrization analyzed by Kang et al. [2025]. 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 each other out, potentially 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.