Generalization Bounds for Stochastic Saddle Point Problems

Junyu Zhang[1,4] junyuz@princeton.edu
Mingyi Hong[2] mhong@umn.edu
Mengdi Wang [1,3] mengdiw@princeton.edu
Shuzhong Zhang[2] zhangs@umn.edu
[1]Princeton University [2]University of Minnesota, Twin Cities [3]DeepMind [4]National University of Singapore

Abstract

This paper studies the generalization bounds for the empirical saddle point (ESP) solution to stochastic saddle point (SSP) problems. For SSP with Lipschitz continuous and strongly convex-strongly concave objective functions, we establish an $O(1/n)$ generalization bound by using a probabilistic stability argument. We also provide generalization bounds under a variety of assumptions, including the cases without strong convexity and without bounded domains. We illustrate our results in three examples: batch policy learning in Markov decision process, stochastic composite optimization problem, and mixed strategy Nash equilibrium estimation for stochastic games. In each of these examples, we show that a regularized ESP solution enjoys a near-optimal sample complexity. To the best of our knowledge, this is the first set of results on the generalization theory of ESP.

1 Introduction

Consider the stochastic saddle point (SSP) problem

$$\text{(SSP)} \qquad \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi(x,y) := \mathbf{E}\big[\Phi_\xi(x,y)\big], \tag{1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are compact and convex sets, and ξ is a random variable. We denote the optimal solution to (1) as (x^*, y^*). SSP problem finds a wide range of applications in machine learning, reinforcement learning, operations research and game theory. Many stochastic approximation (SA) algorithms have been proposed for approximating the SSP solution based on samples of ξ, see e.g. Natole et al. (2018); Nemirovski et al. (2009); Shalev-Shwartz and Zhang (2013); Xiao et al. (2019); Yan et al. (2019); Zhang and Xiao (2017); Zhao (2019). Most of the algorithms make primal-dual updates and enjoy convergence guarantees with appropriately chosen stepsizes.

In this paper, we provide understanding of (1) from a finite sample perspective. Consider the empirical counterpart of the SSP, which we refer to as the empirical saddle point (ESP) problem

$$\text{(ESP)} \qquad \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \hat{\Phi}_n(x,y) := \frac{1}{n}\sum_{i=1}^{n} \Phi_{\xi_i}(x,y)\,, \tag{2}$$

where $\Gamma := \{\xi_1, ..., \xi_n\}$ is a collection of n i.i.d. samples of ξ's. A natural way of estimating the optimal solution (x^*, y^*) to (1) is to solve instead its empirical approximation given by (2). This approach is also known as sample average approximation (SAA), see Shapiro et al. (2014). We denote by $(\hat{x}, \hat{y})$ the empirical saddle point (ESP) solution to problem (2). Based on a given set of samples, one can compute the ESP solution using any convergent algorithm for minimax optimization. In parallel to the generalization theory for empirical risk minimization Vapnik (1992, 2006, 2013), we aim to analyze the empirical saddle point and establish finite-sample generalization bounds.

1.1 Motivating Examples

Stochastic saddle points (1) are very common in machine learning, game theory, and operations research. A generalization theory for ESP would be useful for establishing generalization bounds for a number of machine learning tasks that are beyond empirical risk minimization. We will study three examples in this paper.

One example is batch policy learning for Markov Decision Process (MDP). For the infinite-horizon average-reward MDP, the policy optimization problem is equivalent to an SSP, known as the Bellman saddle point Puterman (2014); Wang (2017), given by

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi(x,y) := \langle y, r\rangle + \sum_{a\in\mathcal{A}} y_a^\top (P_a - I)x, \tag{3}$$

where x is the value function, $y = \{y_a\}_{a\in\mathcal{A}}$ is the state-action occupancy measure, a and $\mathcal{A}$ denote the action and action space respectively, P_a denotes the transition probability matrix under a, r is the reward function (see Section 3 for

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

details). In the batch policy learning problem, we want to solve the MDP without knowledge of r, P_a, instead we only have sample state transitions. This motivates us to study the empirical optimal policy that is equivalent to the ESP solution of the Bellman saddle point (3). Another example is the stochastic composite optimization, which finds applications in off-policy policy evaluation and risk-averse optimization, of the form

$$\min_{x\in\mathcal{X}} \; r(x) + f\left(\mathbf{E}\left[A_\xi x - b_\xi\right]\right), \tag{4}$$

where A_ξ is a random matrix, b_ξ is a random vector, f is a convex loss, r is some regularizer; see Nesterov (2007); Wang et al. (2017); Zhang and Xiao (2019) and references therein. By using the convex conjugate f^* of f, we can reformulate the problem as an SSP:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} r(x) + \mathbf{E}\left[\left(A_\xi x - b_\xi\right)^\top y\right] - f^*(y).$$

Thus a generalization theory for this SSP would lead to generalization bounds for the original composite optimization problem. The third example is the two-person stochastic matrix game

$$\min_{x\in\Delta_{N_1}}\max_{y\in\Delta_{N_2}} x^\top \mathbf{E}[A_\xi] y \tag{5}$$

where the constraint sets are defined as $\Delta_{N_i} := \{z \in \mathbf{R}^{N_i} : z \geq 0, \mathbf{1}^\top z = 1\}$, x, y are the mixed strategies of the two players, and $A_\xi \in \mathbf{R}^{N_1\times N_2}$ is a stochastic payoff matrix. Based on sample payoffs from past plays, we can estimate the mixed strategy Nash equilibrium by solving a regularized version of the empirical matrix game. Our generalization bound will be used to evaluate the quality of the empirical Nash equilibrium learned from data.

1.2 Weak and Strong Generalization Measures

We will study the generalization properties of the empirical saddle point (ESP) solution $(\hat{x}, \hat{y})$ via two metrics. The first metric, referred to as the *weak generalization measure* [1] *(WGM)*, is defined as

$$\Delta^w(\hat{x},\hat{y}) := \max_{y\in\mathcal{Y}} \mathbf{E}\big[\Phi(\hat{x},y)\big] - \min_{x\in\mathcal{X}} \mathbf{E}\big[\Phi(x,\hat{y})\big], \tag{6}$$

where the expectations are taken over the sample set $\{\xi_1,\ldots,\xi_n\}$. In some applications, one desires a stronger metric of optimality, which we refer to as the *strong generalization measure (SGM)*, given by

$$\Delta^s(\hat{x},\hat{y}) := \mathbf{E}\Big[\max_{y\in\mathcal{Y}} \Phi(\hat{x},y) - \min_{x\in\mathcal{X}} \Phi(x,\hat{y})\Big]. \tag{7}$$

The SGM is often referred to as the expected *duality gap* in the optimization literature. A third commonly used metric is $d^2(\hat{x},\hat{y}) := \|\hat{x}-x^*\|^2 + |||\hat{y}-y^*|||^2$, which is the squared distance between the ESP solution to the true saddle point solution. Note that here we allow the use of two different norms $\|\cdot\|$ and $|||\cdot|||$ to measure the distances in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Suppose that $\Phi(\cdot,\cdot)$ is μ_x-strongly convex and μ_y-strongly concave. Then

$$\begin{aligned}
\Delta^w(\hat{x},\hat{y}) &= \max_{y\in\mathcal{Y}} \mathbf{E}\left[\Phi(\hat{x},y)\right] - \min_{x\in\mathcal{X}} \mathbf{E}\left[\Phi(x,\hat{y})\right] \\
&\geq \mathbf{E}\left[\Phi(\hat{x},y^*) - \Phi(x^*,\hat{y})\right] \\
&= \mathbf{E}\left[\Phi(\hat{x},y^*) - \Phi(x^*,y^*) + \Phi(x^*,y^*) - \Phi(x^*,\hat{y})\right] \\
&\geq \mathbf{E}\left[\frac{\mu_x}{2}\|\hat{x}-x^*\|^2 + \frac{\mu_y}{2}|||\hat{y}-y^*|||^2\right] \\
&\geq \frac{\min\{\mu_x,\mu_y\}}{2}\cdot \mathbf{E}\left[d^2(\hat{x},\hat{y})\right].
\end{aligned} \tag{8}$$

That is, we have $\Delta^w(\hat{x},\hat{y}) \geq \Omega\big(\mathbf{E}[d^2(\hat{x},\hat{y})]\big)$. Due to the Jensen's inequality, we also have $\Delta^w(\hat{x},\hat{y}) \leq \Delta^s(\hat{x},\hat{y})$. Therefore the SGM is strongest among the three, WGM is the second and $d^2(\hat{x},\hat{y})$ is the weakest, i.e.,

$$\Omega\left(\mathbf{E}[d^2(\hat{x},\hat{y})]\right) \leq \Delta^w(\hat{x},\hat{y}) \leq \Delta^s(\hat{x},\hat{y}).$$

1.3 Main Results

In this paper, we establish the generalization bounds for solving SSP using the empirical saddle point solution under various assumptions. See Table 1 for an overview of our technical results. Contributions of the paper are three-folded:

- We establish a uniform stability argument for the ESP solution by extending the technique of Shalev-Shwartz et al. (2009). For SSP over compact domains that are Lipschitz continuous and strongly convex-strongly concave (SC-SC), we provide an $O(1/n)$ bound for the WGM metric. With an additional assumption on gradient Lipschitz continuity, we also establish an $O(1/n)$ bound for the SGM metric. Especially, the WGM bound has matched the state-of-the-art result in Yan et al. (2020), though with slightly different measures and assumptions[2]. Especially, when $\mu_x \ll \mu_y$ and $\ell_x^w \ll \ell_y^w$, our result provides a tighter bound.
- Further, we extend the generalization theory to SSP problems with unbounded domains or without the SC-SC property. By using a different analysis, we establish an $O(1/n)$ generalization bound for the ESP solution even if the feasible regions are unbounded. We also provide a generalization bound for the regularized ESP problem.
- We illustrate the applications of our theory in three important learning tasks: the batch policy learning in MDP, the stochastic composite optimization, and the Nash equilibrium estimation. In each of these tasks, the empirical saddle point provides a conceptually simple estimator. Further, we use the generalization theory to show these estimators have

[1] We decide not to use "weak (strong) duality gap measure" to avoid confusion with the well known terminologies of weak (strong) duality.

[2] Yan et al. (2020) proved result under a slightly stronger measure, yet it also requires a stronger assumption on the tails of the distribution.

Table 1: Generalization bounds for stochastic saddle points (results from this paper)

	Assumption 1 & 2	Assumption 1, 2 & 3	Assumption 1, 3 & 4
$\mathbf{E}\left[d^2(\hat{x},\hat{y})\right]$	$O\left(\frac{1}{n\mu_{\min}}\cdot\left(\frac{(\ell_x^w)^2}{\mu_x}+\frac{(\ell_y^w)^2}{\mu_y}\right)\right)$	Same as left	$O\left(\frac{\kappa^2\cdot\mathbf{E}\left[\|\nabla\Phi_\xi(x^*,y^*)\|^2\right]}{n\cdot\min\{\mu_x^2,\mu_y^2\}}\right)$
$\Delta^w(\hat{x},\hat{y})$	$O\left(\frac{(\ell_x^w)^2}{n\mu_x}+\frac{(\ell_y^w)^2}{n\mu_y}\right)$	Same as left	$O\left(\frac{\kappa^4\cdot\mathbf{E}\left[\|\nabla\Phi_\xi(x^*,y^*)\|^2\right]}{n\cdot\min\{\mu_x,\mu_y\}}\right)$
$\Delta^s(\hat{x},\hat{y})$	NA	$O\left(\sqrt{1+\frac{L_{xy}^2}{\mu_x\mu_y}}\cdot\left(\frac{(\ell_x^s)^2}{n\mu_x}+\frac{(\ell_y^s)^2}{n\mu_y}\right)\right)$	$O\left(\frac{\kappa^4\cdot\mathbf{E}\left[\|\nabla\Phi_\xi(x^*,y^*)\|^2\right]}{n\cdot\min\{\mu_x,\mu_y\}}\right)$

provable sample complexities of $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{\epsilon^2}\right)$, $O\left(\frac{1}{\mu\epsilon}\right)$ and $\tilde{O}(N_1N_2\epsilon^{-2})$ for the three learning tasks.[3]. These sample complexities have tight dependencies on the parameters $|\mathcal{S}||\mathcal{A}|$, μ, N_1N_2 and ϵ, and they have not been studied before.

1.4 Related works

Let us review the stochastic approximation (SA) approaches for the SSP problem (1). When $\Phi(\cdot,\cdot)$ is only convex and concave, the seminal paper Nemirovski et al. (2009) established an $O(1/\sqrt{n})$ convergence in SGM for a stochastic mirror descent ascent algorithm. Similar $O(1/\sqrt{n})$ convergence in SGM are also obtained by Bach and Levy (2019), Zhao (2019) and Chen et al. (2014) under various assumptions. Specifically, such $O(1/\sqrt{n})$ convergence is also obtained for the class of general stochastic variational inequalities, which include the SSP problem as a special case, see Juditsky et al. (2011) and Chen et al. (2017). The interested readers are also referred to the Chapter 4 of the Lan (2020). When the SC-SC property is further assumed, a faster $O(1/n)$ convergence can be derived. For example, Natole et al. (2018) obtained an $O(1/n)$ convergence in terms of the squared distance metric. Yan et al. (2019) designed a stochastic gradient method with $O(1/n)$ convergence in SGM when the coupling between the primal and dual variable is linear. Yan et al. (2020) derived an epoch-wise stochastic gradient method that guarantees $O(1/n)$ in SGM. This result of Yan et al. (2020) does not rely on the linear coupling structure, but instead requires additional conditions on the tail distribution of sampling noise. There also exist research results for SSP with special structures such as finite-sum and bilinear coupling, etc, see e.g. Du and Hu (2018); Shalev-Shwartz and Zhang (2013); Xiao et al. (2019); Zhang and Xiao (2017).

There exist a rich body of literatures on the generalization theory for solving stochastic convex optimization (SCO) by empirical risk minimization (ERM), namely,

$$\begin{cases}\min_{x\in\mathcal{X}}\ \Phi(x) \ := \mathbf{E}_\xi\left[\Phi_\xi(x)\right] & \text{(SCO)}\\ \min_{x\in\mathcal{X}}\hat{\Phi}_n(x) := \frac{1}{|\Gamma|}\sum_{\xi\in\Gamma}\Phi_\xi(x) & \text{(ERM)}\end{cases}$$

In the seminal paper Shalev-Shwartz et al. (2009), Shalev et al. established an $O(1/n)$ ERM generalization bound for strongly convex problems and an $O(1/\sqrt{n})$ risk bound for general convex problems. Similar rates are also obtained in related works Sridharan et al. (2009); Gonen and Shalev-Shwartz (2017). The main technique used by Shalev-Shwartz et al. (2009) and our paper is the *uniform stability* argument, which was originally introduced by Bousquet and Elisseeff (2002), and later on studied in many papers, see e.g. Kearns and Ron (1999); Mukherjee et al. (2006); Shalev-Shwartz et al. (2009); Shalev-Shwartz and Zhang (2013); Hardt et al. (2015); Chen et al. (2018b), etc. With the tool of the Rademacher complexity R_n, Srebro et al. (2010) demonstrated an $O(R_n/\sqrt{n})$ risk bound for ERM, and many papers strengthened the theory further Bartlett and Mendelson (2002); Bartlett et al. (2005). For nonconvex but exp-concave objectives, Koren and Levy (2015) and Mehta (2016) derived a risk bound of $O(1/n)$. Under certain stronger conditions, a tighter $O(1/n^2)$ risk bound has been shown Zhang et al. (2017).

In a weakly related work Dikkala et al. (2020), the authors introduce a min-max approach for estimating models with conditional moment restrictions. Under a different setting, Dikkala et al. (2020) studies the statistical properties of the ESP solutions in terms of some distance between the emprical and true solutions in the minimization side, which is analogous to $\mathbf{E}[\|\hat{x}-x^*\|]$ in our setting. To the authors' best knowledge, there is no existing generalization bound for stochastic saddle point problems.

2 Generalization Bounds for Empirical Saddle Points

2.1 Assumptions

In most of our analysis, we require that Φ is strongly convex and strongly concave (SC-SC).

Assumption 1 (SC-SC objective function)**.** *$\exists\mu_x,\mu_y\geq 0$, s.t. for almost every ξ, $\Phi_\xi(\cdot,y)$ is μ_x-strongly convex under norm $\|\cdot\|$ and $\Phi_\xi(x,\cdot)$ is μ_y-strongly concave under the norm $|||\cdot|||$. Namely, denote $\partial_x\Phi_\xi(\cdot,y)$ and $\partial_y\Phi_\xi(x,\cdot)$ the subgradients and supergradients respecitively, then for $\forall x,x'\in\mathcal{X},\forall y,y'\in\mathcal{Y}$, $u\in\partial_x\Phi_\xi(x,y)$ and $v\in\partial_y\Phi_\xi(x,y)$,*

[3]The definition of these parameters can be found in the corresponding sections.

it holds that

$$\begin{cases}\Phi_\xi(x',y)\geq\Phi_\xi(x,y)+\langle u,x'-x\rangle+\frac{\mu_x}{2}\|x'-x\|^2,\\ \Phi_\xi(x,y')\leq\Phi_\xi(x,y)+\langle v,y'-y\rangle-\frac{\mu_y}{2}\|\|y'-y\|\|^2.\end{cases}\tag{9}$$

For convex analysis of strongly convex function under non-Euclidean norms, see Shalev-Shwartz and Singer (2007); Kakade et al. (2012) and references therein. We further assume that the feasible regions $\mathcal{X},\mathcal{Y}$ are bounded and the objective function is Lipschitz continuous.

Assumption 2 (Function Lipschitz continuity)**.** *The feasible regions $\mathcal{X}$ and $\mathcal{Y}$ are compact convex sets. For almost every ξ, there exist constants $\ell_x(\xi,y)$ and $\ell_y(\xi,x)$ s.t.*

$$\begin{cases}|\Phi_\xi(x',y)-\Phi_\xi(x,y)|\leq\ell_x(\xi,y)\|x'-x\|,\\ |\Phi_\xi(x,y')-\Phi_\xi(x,y)|\leq\ell_y(\xi,x)\|\|y'-y\|\|,\end{cases}\tag{10}$$

for $\forall x,x'\in\mathcal{X}$ and $\forall y,y'\in\mathcal{Y}$. To bound the WGM, we need to assume

$$\begin{cases}(\ell_x^w)^2:=\sup_{y\in\mathcal{Y}}\mathbf{E}_\xi\left[\ell_x^2(\xi,y)\right]<+\infty,\\ (\ell_y^w)^2:=\sup_{x\in\mathcal{X}}\mathbf{E}_\xi\left[\ell_y^2(\xi,x)\right]<+\infty.\end{cases}\tag{11}$$

To bound the SGM, we assume

$$\begin{cases}(\ell_x^s)^2:=\mathbf{E}_\xi\left[\sup_{y\in\mathcal{Y}}\ell_x^2(\xi,y)\right]<+\infty,\\ (\ell_y^s)^2:=\mathbf{E}_\xi\left[\sup_{x\in\mathcal{X}}\ell_y^2(\xi,x)\right]<+\infty.\end{cases}\tag{12}$$

Due to Jensen's inequality, $\ell_x^w\leq\ell_x^s$ and $\ell_y^w\leq\ell_y^s$ always hold.

In our analysis, Assumptions 1 and 2 only guarantee the $O(1/n)$ bound for the WGM metric. In order to prove an $O(1/n)$ bound for the stronger metric SGM, we will require additional smoothness of Φ.

Assumption 3 (Gradient Lipschitz continuity)**.** *There exist constants L_x, L_y and L_{xy} s.t. for $\forall x,x'\in\mathcal{X},\forall y,y'\in\mathcal{Y}$, it holds that*

$$\begin{cases}\|\nabla_x\Phi(x,y)-\nabla_x\Phi(x',y)\|_*\leq L_x\|x-x'\|\\ \|\|\nabla_y\Phi(x,y)-\nabla_y\Phi(x,y')\|\|_*\leq L_y\|\|y-y'\|\|\\ \|\nabla_x\Phi(x,y)-\nabla_x\Phi(x,y')\|_*\leq L_{xy}\|\|y-y'\|\|\\ \|\|\nabla_y\Phi(x,y)-\nabla_y\Phi(x',y)\|\|_*\leq L_{xy}\|x-x'\|\end{cases}$$

where $\|\cdot\|_*$ and $\|\|\cdot\|\|_*$ stand for the dual norms of $\|\cdot\|$ and $\|\|\cdot\|\|$ respectively.

Finally, we also study the case where $\mathcal{X}$ and $\mathcal{Y}$ are unbounded. Such unboundedness would invalidate Assumption 2 as well as the stability argument. To remedy this issue, we would replace Assumption 2 with the following assumption about the true optimal solution (x^*,y^*).

Assumption 4. $\exists C>0$ *s.t.* $\mathbf{E}_\xi\left[\|\nabla\Phi_\xi(x^*,y^*)\|_2^2\right]\leq C$.

2.2 Main Results

We use the leave-one-out technique in Shalev-Shwartz et al. (2009) to analyze the stability of the ESP solutions. Let $\Gamma:=\{\xi_1,...,\xi_n\}$ be a set of n i.i.d. samples, and let ξ_i' be another independent sample. We then define the perturbed sample set $\Gamma(i)=\Gamma\cup\{\xi_i'\}\backslash\{\xi_i\}$. That is, $\Gamma(i)$ is constructed by replacing just the i-th sample Γ with another i.i.d. sample ξ_i'. For the sake of generality, instead of the ESP solution, we will establish the stability property of the regularized ESP (R-ESP) solutions:

$$\text{(R-ESP)}\qquad\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\Phi_\Gamma(x,y)+\Psi(x,y),\tag{13}$$

where Ψ is a regularization function that is SC-SC and can be specified by the user. In particular, if we set $\Psi\equiv 0$, the R-ESP problem (13) is reduced to the ESP problem (2) where no regularization is applied.

Lemma 1 (Stability property for R-ESP solution)**.** *Suppose the regularization function $\Psi(\cdot,y)$ is ν_x-strongly convex under norm $\|\cdot\|$ and $\Psi(x,\cdot)$ is ν_y-strongly concave under norm $\|\|\cdot\|\|$, and there exists $R>0$ s.t. $|\Psi(x,y)|\leq R$, $\forall(x,y)\in\mathcal{X}\times\mathcal{Y}$. Under the Assumptions 1, 2, let $(\hat{x},\hat{y})$ and $(\hat{x}_{(i)},\hat{y}_{(i)})$ be the solution to the R-ESP problem* (13) *with sample set Γ and $\Gamma(i)$ respectively, and suppose $\mu_x+\nu_x>0$, $\mu_y+\nu_y>0$, then*

$$\begin{aligned}&\sqrt{(\mu_x+\nu_x)\|\hat{x}-\hat{x}_{(i)}\|^2+(\mu_y+\nu_y)\|\|\hat{y}-\hat{y}_{(i)}\|\|^2}\\ \leq\;&\frac{1}{n}\sqrt{\frac{(\ell_x(\xi_i,\hat{y}_{(i)})+\ell_x(\xi_i',\hat{y}))^2}{\mu_x+\nu_x}+\frac{(\ell_y(\xi_i,\hat{x}_{(i)})+\ell_y(\xi_i',\hat{x}))^2}{\mu_y+\nu_y}}.\end{aligned}$$

Proof. First, in parallel to $\hat{\Phi}_n$, we define

$$\hat{\Phi}_{n,i}(x,y)=\frac{1}{n}\sum_{\xi\in\Gamma(i)}\Phi_\xi(x,y).$$

Then we have

$$\begin{aligned}&\hat{\Phi}_n(\hat{x}_{(i)},\hat{y})+\Psi(\hat{x}_{(i)},\hat{y})-\hat{\Phi}_n(\hat{x},\hat{y}_{(i)})-\Psi(\hat{x},\hat{y}_{(i)})\\ =\;&\frac{1}{n}\sum_{j=1}^n\left(\Phi_{\xi_j}(\hat{x}_{(i)},\hat{y})-\Phi_{\xi_j}(\hat{x},\hat{y}_{(i)})\right)+\Psi(\hat{x}_{(i)},\hat{y})-\Psi(\hat{x},\hat{y}_{(i)})\\ =\;&\frac{1}{n}\left(\Phi_{\xi_i'}(\hat{x}_{(i)},\hat{y})-\Phi_{\xi_i'}(\hat{x},\hat{y}_{(i)})+\sum_{j=1,j\neq i}^n\left(\Phi_{\xi_j}(\hat{x}_{(i)},\hat{y})-\Phi_{\xi_j}(\hat{x},\hat{y}_{(i)})\right)\right)\\ &+\frac{1}{n}\left(\Phi_{\xi_i}(\hat{x}_{(i)},\hat{y})-\Phi_{\xi_i}(\hat{x}_{(i)},\hat{y}_{(i)})+\Phi_{\xi_i}(\hat{x}_{(i)},\hat{y}_{(i)})-\Phi_{\xi_i}(\hat{x},\hat{y}_{(i)})\right)\\ &-\frac{1}{n}\left(\Phi_{\xi_i'}(\hat{x}_{(i)},\hat{y})-\Phi_{\xi_i'}(\hat{x},\hat{y})+\Phi_{\xi_i'}(\hat{x},\hat{y})-\Phi_{\xi_i'}(\hat{x},\hat{y}_{(i)})\right)\\ &+\Psi(\hat{x}_{(i)},\hat{y})-\Psi(\hat{x},\hat{y}_{(i)})\\ \overset{(a)}{\leq}\;&\left(\hat{\Phi}_{n,i}(\hat{x}_{(i)},\hat{y})+\Psi(\hat{x}_{(i)},\hat{y})-\hat{\Phi}_{n,i}(\hat{x},\hat{y}_{(i)})-\Psi(\hat{x},\hat{y}_{(i)})\right)\\ &+\frac{\ell_x(\xi_i,\hat{y}_{(i)})+\ell_x(\xi_i',\hat{y})}{n}\|\hat{x}-\hat{x}_{(i)}\|+\frac{\ell_y(\xi_i,\hat{x}_{(i)})+\ell_y(\xi_i',\hat{x})}{n}\|\|\hat{y}-\hat{y}_{(i)}\|\|\end{aligned}\tag{14}$$

$$
\begin{aligned}
&= \Big(\hat{\Phi}_{n,i}\big(\hat{x}_{(i)},\hat{y}\big)+\Psi(\hat{x}_{(i)},\hat{y})-\hat{\Phi}_{n,i}\big(\hat{x}_{(i)},\hat{y}_{(i)}\big)-\Psi\big(\hat{x}_{(i)},\hat{y}_{(i)}\big)\Big)\\
&\quad+\Big(\hat{\Phi}_{n,i}\big(\hat{x}_{(i)},\hat{y}_{(i)}\big)+\Psi\big(\hat{x}_{(i)},\hat{y}_{(i)}\big)-\hat{\Phi}_{n,i}\big(\hat{x},\hat{y}_{(i)}\big)+\Psi\big(\hat{x},\hat{y}_{(i)}\big)\Big)\\
&\quad+\frac{\ell_x(\xi_i,\hat{y}_{(i)})+\ell_x(\xi_i',\hat{y})}{n}\big\|\hat{x}-\hat{x}_{(i)}\big\|+\frac{\ell_y(\xi_i,\hat{x}_{(i)})+\ell_y(\xi_i',\hat{x})}{n}|||\hat{y}-\hat{y}_{(i)}|||\\
&\overset{(b)}{\leq} -\frac{\mu_x+\nu_x}{2}\|\hat{x}-\hat{x}_{(i)}\|^2-\frac{\mu_y+\nu_y}{2}|||\hat{y}-\hat{y}_{(i)}|||^2\\
&\quad+\frac{\ell_x(\xi_i,\hat{y}_{(i)})+\ell_x(\xi_i',\hat{y})}{n}\big\|\hat{x}-\hat{x}_{(i)}\big\|+\frac{\ell_y(\xi_i,\hat{x}_{(i)})+\ell_y(\xi_i',\hat{x})}{n}|||\hat{y}-\hat{y}_{(i)}|||.
\end{aligned}
$$

The step (a) is due to Lipschitz continuity of $\Phi_{\xi_i}, \Phi_{\xi_i'}$. The step (b) is due the $(\mu_y+\nu_y)$-strong concavity of $\hat{\Phi}_{n,i}\big(\hat{x}_{(i)},\cdot\big)+\Psi\big(\hat{x}_{(i)},\cdot\big)$ and the fact that

$$
\hat{y}_{(i)}=\operatorname*{argmax}_{y\in\mathcal{Y}}\hat{\Phi}_{n,i}\big(\hat{x}_{(i)},y\big)+\Psi\big(\hat{x}_{(i)},y\big).
$$

Hence

$$
\begin{aligned}
&\hat{\Phi}_{n,i}\big(\hat{x}_{(i)},\hat{y}\big)+\Psi\big(\hat{x}_{(i)},\hat{y}\big)-\hat{\Phi}_{n,i}\big(\hat{x}_{(i)},\hat{y}_{(i)}\big)-\Psi\big(\hat{x}_{(i)},\hat{y}_{(i)}\big)\\
\leq\ & -\frac{\mu_y+\nu_y}{2}|||\hat{y}-\hat{y}_{(i)}|||^2.
\end{aligned}
$$

The other part of argument on $-\frac{\mu_x+\nu_x}{2}\|\hat{x}-\hat{x}_{(i)}\|^2$ is similar. On the other hand, similar to the argument of step (b) above, because $\hat{x},\hat{y}$ and solves the strongly convex and strongly concave R-ESP problem (13), we also have

$$
\begin{aligned}
&\hat{\Phi}_n\big(\hat{x}_{(i)},\hat{y}\big)+\Psi\big(\hat{x}_{(i)},\hat{y}\big)-\hat{\Phi}_n\big(\hat{x},\hat{y}_{(i)}\big)-\Psi\big(\hat{x},\hat{y}_{(i)}\big)\\
\geq\ & \frac{\mu_x+\nu_x}{2}\|\hat{x}-\hat{x}_{(i)}\|^2+\frac{\mu_y+\nu_y}{2}|||\hat{y}-\hat{y}_{(i)}|||^2. \qquad (15)
\end{aligned}
$$

Combining the (14) and (15) yields

$$
\begin{aligned}
&(\mu_x+\nu_x)\|\hat{x}-\hat{x}_{(i)}\|^2+(\mu_y+\nu_y)|||\hat{y}-\hat{y}_{(i)}|||^2\\
&\leq\frac{\ell_x(\xi_i,\hat{y}_{(i)})+\ell_x(\xi_i',\hat{y})}{n}\big\|\hat{x}-\hat{x}_{(i)}\big\|+\frac{\ell_y(\xi_i,\hat{x}_{(i)})+\ell_y(\xi_i',\hat{x})}{n}|||\hat{y}-\hat{y}_{(i)}|||\\
&\leq\frac{1}{n}\sqrt{\frac{\big(\ell_x(\xi_i,\hat{y}_{(i)})+\ell_x(\xi_i',\hat{y})\big)^2}{\mu_x+\nu_x}+\frac{\big(\ell_y(\xi_i,\hat{x}_{(i)})+\ell_y(\xi_i',\hat{x})\big)^2}{\mu_y+\nu_y}}\\
&\quad\times\sqrt{(\mu_x+\nu_x)\|\hat{x}-\hat{x}_{(i)}\|^2+(\mu_y+\nu_y)|||\hat{y}-\hat{y}_{(i)}|||^2}
\end{aligned}
$$

where the last row uses the Caucy-Schwartz inequality. Dividing both sides by $\sqrt{(\mu_x+\nu_x)\|\hat{x}-\hat{x}_{(i)}\|^2+(\mu_y+\nu_y)|||\hat{y}-\hat{y}_{(i)}|||^2}$ proves this lemma. □

For the important special case where $\mu_x,\mu_y>0$ and the regularization term $\Psi\equiv 0$, we have the following corollary.

Corollary 1 (Stability property)**.** *Let the Assumptions 1 and 2 hold, and $\mu_x,\mu_y>0$. Denote $(\hat{x},\hat{y})$ and $(\hat{x}_{(i)},\hat{y}_{(i)})$ as the solutions to the ESP problem* (2) *with sample sets Γ and $\Gamma(i)$ respectively. Then*

$$
\begin{aligned}
&\sqrt{\mu_x\|\hat{x}-\hat{x}_{(i)}\|^2+\mu_y\|\hat{y}-\hat{y}_{(i)}\|^2}\\
&\leq\frac{1}{n}\sqrt{\frac{(\ell_x(\xi_i,\hat{y}_{(i)})+\ell_x(\xi_i',\hat{y}))^2}{\mu_x}+\frac{(\ell_y(\xi_i,\hat{x}_{(i)})+\ell_y(\xi_i',\hat{x}))^2}{\mu_y}}.
\end{aligned}
$$

For the ESP problem, it is worth noting that by the Mcdiarmid's inequality McDiarmid (1989, 1998), the stability argument of Corollary 1 immediately results in an $\tilde{O}(1/\sqrt{n})$ generalization bound for SC-SC problems, which, however, is not tight. In Theorem 1, we establish a tighter $O(1/n)$ bound for SC-SC problems by using a more careful analysis.

Lemma 2 (Generalization bound for R-ESP)**.** *Under the settings of Lemma 1, the R-ESP solution $(\hat{x},\hat{y})$ satisfies*

$$
\Delta^w(\hat{x},\hat{y})\leq\frac{2\sqrt{2}}{n}\cdot\left(\frac{(\ell_x^w)^2}{\mu_x+\nu_x}+\frac{(\ell_y^w)^2}{\mu_y+\nu_y}\right)+2R,
$$

where the WGM $\Delta^w(\cdot,\cdot)$ is defined for the original unregularized SSP problem.

Proof. By the function Lipschitz continuity of Assumption 2, for any $1\leq i\leq n$, and for any $x\in\mathcal{X}$ and $y\in\mathcal{Y}$,

$$
\begin{aligned}
\Phi_{\xi_i}\big(\hat{x}_{(i)},y\big)-\Phi_{\xi_i}\big(x,\hat{y}_{(i)}\big)\leq\ & \Phi_{\xi_i}(\hat{x},y)-\Phi_{\xi_i}\big(x,\hat{y}\big) \qquad (16)\\
&\underbrace{+\ell_x\big(\xi_i,y\big)\|\hat{x}-\hat{x}_{(i)}\|+\ell_y\big(\xi_i,x\big)|||\hat{y}-\hat{y}_{(i)}|||}_{T(i)}.
\end{aligned}
$$

As a result, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n\mathbf{E}\big[\Phi(\hat{x}_{(i)},y)-\Phi\big(x,\hat{y}_{(i)}\big)\big]+\Psi\big(\hat{x},y\big)-\Psi\big(x,\hat{y}\big) \qquad (17)\\
&\overset{(a)}{=}\frac{1}{n}\sum_{i=1}^n\mathbf{E}\big[\Phi_{\xi_i}\big(\hat{x}_{(i)},y\big)-\Phi_{\xi_i}\big(x,\hat{y}_{(i)}\big)\big]+\Psi\big(\hat{x},y\big)-\Psi\big(x,\hat{y}\big)\\
&\overset{(b)}{\leq}\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n\big(\Phi_{\xi_i}\big(\hat{x},y\big)-\Phi_{\xi_i}\big(x,\hat{y}\big)\big)\right]+\Psi\big(\hat{x},y\big)-\Psi\big(x,\hat{y}\big)+\frac{1}{n}\sum_{i=1}^n\mathbf{E}[T(i)].
\end{aligned}
$$

The step (a) is due to the fact that $(\hat{x}_{(i)},\hat{y}_{(i)})$ is independent from ξ_i, and hence one can take the expectation over the ξ_i's first. And the step (b) is due to (16). Then, because the distribution of $\big(\hat{x}_{(i)},\hat{y}_{(i)}\big)$ are the same as that of $\big(\hat{x},\hat{y}\big)$ for any $1\leq i\leq n$. Therefore, the expectation term on the LHS of (17) can be simplified to

$$
\frac{1}{n}\sum_{i=1}^n\mathbf{E}\big[\Phi\big(\hat{x}_{(i)},y\big)-\Phi\big(x,\hat{y}_{(i)}\big)\big]=\mathbf{E}\big[\Phi\big(\hat{x},y\big)-\Phi\big(x,\hat{y}\big)\big]. \qquad (18)
$$

Second, the first term on the RHS of (17) is actually

$$
\begin{aligned}
&\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n(\Phi_{\xi_i}(\hat{x},y)-\Phi_{\xi_i}(x,\hat{y}))\right]+\Psi(\hat{x},y)-\Psi(x,\hat{y})\\
=\ &\mathbf{E}\left[\hat{\Phi}_n(\hat{x},y)-\hat{\Phi}_n(x,\hat{y})+\Psi(\hat{x},y)-\Psi(x,\hat{y})\right] \qquad (19)\\
\leq\ & 0,
\end{aligned}
$$

for $\forall(x,y)\in\mathcal{X}$ and $y\in\mathcal{Y}$, which is because $(\hat{x},\hat{y})$ solves the R-ESP problem (13). Third, because the distributions of $T(i)$'s are the same, for the second term on the RHS of (17),

we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}[T(i)] \qquad (20)\\
&=\mathbf{E}\left[\ell_x(\xi_i,y)\|\hat{x}-\hat{x}_{(i)}\|+\ell_y(\xi_i,x)\|\hat{y}-\hat{y}_{(i)}\|\right]\\
&\overset{(a)}{\leq}\mathbf{E}\left[\sqrt{\frac{\ell_x^2(\xi_i,y)}{\mu_x+\nu_x}+\frac{\ell_y^2(\xi_i,x)}{\mu_y+\nu_y}}\right.\\
&\left.\times\sqrt{(\mu_x+\nu_x)\|\hat{x}-\hat{x}_{(i)}\|^2+(\mu_y+\nu_y)\|\hat{y}-\hat{y}_{(i)}\|^2}\right]\\
&\overset{(b)}{\leq}\sqrt{\mathbf{E}\left[\frac{\ell_x^2(\xi_i,y)}{\mu_x+\nu_x}+\frac{\ell_y^2(\xi_i,x)}{\mu_y+\nu_y}\right]}\\
&\times\sqrt{\mathbf{E}\left[(\mu_x+\nu_x)\|\hat{x}-\hat{x}_{(i)}\|^2+(\mu_y+\nu_y)\|\hat{y}-\hat{y}_{(i)}\|^2\right]}\\
&\overset{(c)}{\leq}\sqrt{\mathbf{E}\left[\frac{\ell_x^2(\xi_i,y)}{\mu_x+\nu_x}+\frac{\ell_y^2(\xi_i,x)}{\mu_y+\nu_y}\right]}\\
&\times\frac{1}{n}\sqrt{\mathbf{E}\left[\frac{(\ell_x(\xi_i,\hat{y}_{(i)})+\ell_x(\xi_i',\hat{y}))^2}{\mu_x+\nu_x}+\frac{(\ell_y(\xi_i,\hat{x}_{(i)})+\ell_y(\xi_i',\hat{x}))^2}{\mu_y+\nu_y}\right]}\\
&\overset{(d)}{\leq}\sqrt{\frac{2(\ell_x^w)^2}{\mu_x+\nu_x}+\frac{2(\ell_y^w)^2}{\mu_y+\nu_y}}\cdot\frac{1}{n}\sqrt{\frac{2(\ell_x^w)^2+2(\ell_x^w)^2}{\mu_x+\nu_x}+\frac{2(\ell_y^w)^2+2(\ell_y^w)^2}{\mu_y+\nu_y}}\\
&=\frac{2\sqrt{2}}{n}\cdot\left(\frac{(\ell_x^w)^2}{\mu_x+\nu_x}+\frac{(\ell_y^w)^2}{\mu_y+\nu_y}\right).
\end{aligned}
$$

The step (a) uses the vector Cauchy-Schwartz inequality $a^\top b \leq \|a\|_2\cdot\|b\|_2$ for some vectors a and b. The step (b) uses the expectation version of Cauchy-Schwartz inequality $\mathbf{E}[ab]\leq\sqrt{\mathbf{E}[a^2]}\cdot\sqrt{\mathbf{E}[b^2]}$ for some random variables a and b. The step (c) is due to Lemma 1. And the step (d) is due to Assumption 2, and the fact that $(\hat{x}_{(i)},\hat{y}_{(i)})$ is independent from ξ_i and $(\hat{x},\hat{y})$ is independent from ξ_i'. Finally, substituting (18), (19), and (20) into (17) provides the following result:

$$
\begin{aligned}
&\mathbf{E}\big[\Phi(\hat{x},y)\big]-\mathbf{E}\big[\Phi(x,\hat{y})\big]+\Psi(\hat{x},y)-\Psi(x,\hat{y})\\
&\leq\frac{2\sqrt{2}}{n}\cdot\left(\frac{(\ell_x^w)^2}{\mu_x+\nu_x}+\frac{(\ell_y^w)^2}{\mu_y+\nu_y}\right),\quad\text{for}\quad\forall x,y.
\end{aligned}
$$

Due to the bound the regularizer, we know $|\Psi(\hat{x},y)-\Psi(x,\hat{y})|\leq 2R$. Note that the above inequality is true for any x and y. Therefore, we prove the overal result that

$$
\begin{aligned}
&\max_{y\in\mathcal{Y}}\mathbf{E}\big[\Phi(\hat{x},y)\big]-\min_{x\in\mathcal{X}}\mathbf{E}\big[\Phi(x,\hat{y})\big]\\
&\leq\frac{2\sqrt{2}}{n}\cdot\left(\frac{(\ell_x^w)^2}{\mu_x+\nu_x}+\frac{(\ell_y^w)^2}{\mu_y+\nu_y}\right)+2R.
\end{aligned}
$$

This completes the proof. □

Given the above lemma, let us present our first theorem about the generalization of the ESP solution for an SC-SC stochstic saddle point problem.

Theorem 1. *(Upper bound on WGM) For the SSP problem* (1)*, let* $(\hat{x},\hat{y})$ *be the solution to the ESP problem* (2)*. And let* $n=|\Gamma|$ *be sample size. Under Assumptions 1, 2, and suppose* $\mu_x,\mu_y>0$*, we have*

$$
\begin{cases}
\mathbf{E}\left[d^2(\hat{x},\hat{y})\right]\leq\frac{2\Delta^w(\hat{x},\hat{y})}{\min\{\mu_x,\mu_y\}},\\
\Delta^w(\hat{x},\hat{y})\leq\frac{2\sqrt{2}}{n}\left(\frac{(\ell_x^w)^2}{\mu_x}+\frac{(\ell_y^w)^2}{\mu_y}\right).
\end{cases}
\qquad (21)
$$

The first inequality of (21) is directly due to (8). The second inequality of (21) can be derived from Lemma 2 by setting the regularizer Ψ to be 0. Namely, $R=\nu_x=\nu_y=0$.

The generalization bounds given in Theorem 1 have tight dependence on the sample size n, as well as the problem's parameters ℓ_x^w,ℓ_y^w and μ_x,μ_y. To see this, we can simply consider the special case of SCO. When $\Phi_\xi(x,\cdot)\equiv f_\xi(x)$, i.e., the objective function is constant in y, the SSP and ESP reduce to the classical SCO and ERM respectively. In this case, the difference between ℓ_x^w and ℓ_x^s vanishes, and we denote them as $\ell_x:=\ell_x^w=\ell_x^s$. The WGM also reduces to

$$
\begin{aligned}
\Delta^w(\hat{x},\hat{y}) &= \max_{y\in\mathcal{Y}}\mathbf{E}[f(\hat{x})]-\min_{x\in\mathcal{X}}\mathbf{E}[f(x)]\\
&= \mathbf{E}\big[f(\hat{x})-\min_{x\in\mathcal{X}}f(x)\big].
\end{aligned}
$$

The generalization bound (21) becomes $O\left(\frac{\ell_x^2}{n\mu_x}\right)$ and matches the generalization bound for ERM Shalev-Shwartz et al. (2009).

Note that Theorem 1 requires $\mu_x,\mu_y>0$, i.e., the problem is strongly convex and strongly concave. When the problem is general convex and concave with $\mu_x=\mu_y=0$, we will need to choose the regularizer optimally and establish a generalization bound that depends only on the diameters of $\mathcal{X},\mathcal{Y}$ and constants of Lipschitz continuity of Φ.

Theorem 2. *Suppose the* Φ_ξ*'s are convex concave but not SC-SC. Namely,* $\mu_x=\mu_y=0$*. Suppose the* Φ_ξ*'s satisfy Assumption 2 under the* L_2 *norm. Then we can set the regularizer to be* $\Psi(x,y)=\frac{\alpha_x}{2}\|x\|_2^2-\frac{\alpha_y}{2}\|y\|_2^2$*. Consequently,* $R=\frac{\alpha_x}{2}D_x^2+\frac{\alpha_y}{2}D_y^2$*, where* D_x *and* D_y *denote the diameters of* $\mathcal{X}$ *and* $\mathcal{Y}$ *under* L_2*-norm respectively. If we set* $\alpha_x=\frac{\ell_x^w}{\sqrt{n}D_x}$ *and* $\alpha_y=\frac{\ell_y^w}{\sqrt{n}D_y}$*, Lemma 2 implies*

$$
\Delta^w(\hat{x},\hat{y})\leq O\left(\frac{\ell_x^w D_x+\ell_y^w D_y}{\sqrt{n}}\right).
$$

Similar to Theorem 1, Theorem 2 is also tight for the special case of general convex stochastic optimization problem.

The above two theorems characterize the generalization bounds under the measure of weak generalization measure. By utilizing additional smoothness of the objective function, we provide an $O(1/n)$ bound on SGM, for the SC-SC case where $\mu_x,\mu_y>0$. We provide the proof in Appendix A.1.

Theorem 3. *(Upper bound on SGM) Under the settings of Theorem 1, if Assumption 3 holds in addition, we have*

$$\Delta^s(\hat{x},\hat{y}) \le \frac{2\sqrt{2}}{n}\cdot\sqrt{\frac{L_{xy}^2}{\mu_x\mu_y}+1}\cdot\left(\frac{(\ell_x^s)^2}{\mu_x}+\frac{(\ell_y^s)^2}{\mu_y}\right). \quad (22)$$

We remark that the bound (22) has an additional multiplicative $O\big(L_{xy}/\sqrt{\mu_x\mu_y}\big)$ factor compared to bound (21). It remains open whether this dependence can be improved, as a question for future work. It is worth noting that for SA type method, Yan et al. (2020) has derived a similar bound for SGM of $O\Big(\frac{\max\{\ell_x^s,\ell_y^s\}^2}{n\min\{\mu_x,\mu_y\}}\Big)$, yet with stronger requirements on tails of the stochastic gradients. When the parameters are unbalanced, i.e. $\mu_x \ll \mu_y$ and $\ell_x^s \ll \ell_y^s$, our result can be better.

Note that Theorem 1, 2 and 3 are all based on the stability argument in Lemma 1, which relies heavily on the Lipschitz continuity of the objective function. Next we study SSP problem over unbounded domains. In this case, the SC-SC property and function Lipschitz continuity are mutually exclusive. In the next theorem, we provide a generalization bound for this class of problem.

Theorem 4 (Generalization error for unbounded problems)**.** *Let Assumptions 1, 3 and 4 hold, and let the SSP be unconstrained. Let* $\|\cdot\| = |||\cdot||| = \|\cdot\|_2$. *Then*

$$\begin{cases}\mathbf{E}\left[d^2(\hat{x},\hat{y})\right] \le O\left(\frac{C\kappa^2}{n\mu^2}\right),\\ \Delta^s(\hat{x},\hat{y}) \le O\left(\frac{C\kappa^4}{n\mu}\right),\end{cases} \quad (23)$$

where $\mu = \min\{\mu_x,\mu_y\}$ *and* $\kappa = \frac{\max\{L_x,L_y,L_{xy}\}}{\min\{\mu_x,\mu_y\}}$ *is the condition number.*

3 Application to Batch Policy Learning for MDP

3.1 Saddle Point Formulation of MDP

Consider the policy learning problem for an infinite-horizon average-reward Markov Decision Process (MDP). The MDP instance is specified by $M = (\mathcal{S},\mathcal{A},P,r)$ where $\mathcal{S}$ is a finite state space. $\mathcal{A}$ is a finite action space, $P = \{P_a\}$ are state transition matrices with $P_a(s,s') = \mathbf{Prob}(s_{t+1} = s' \mid s_t = s, a_t = a)$, for $\forall s,s' \in \mathcal{S}$ and $a \in \mathcal{A}$. r is the reward function with $r_{sa} \in [0,1]$ being the reward received after taking action a at state s. A policy $\pi : \mathcal{S} \mapsto \Delta_{\mathcal{A}}$ maps a state s to a distribution over the action space $\mathcal{A}$, where we denote the probability of taking action a at state s as $\pi(a|s)$. The objective is to maximize the long-term average reward, defined as

$$\hat{v}^* := \max_{\pi}\lim_{T\to\infty}\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_{s_t a_t}\,\middle|\,\pi, s_0 = s\right]. \quad (24)$$

The optimal Bellman equation has an equivalent saddle point formulation (3) Puterman (2014); Chen and Wang (2016)

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\Phi(x,y) := \langle y, r\rangle + \sum_{a\in\mathcal{A}} y_a^\top(P_a - I)x,$$

where $x \in \mathbf{R}^{|\mathcal{S}|}$ is the *difference-of-value* vector, $y \in \mathbf{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ stands for the stationary state-action distribution under certain policy π. $y_a = [y_{1,a},...,y_{|\mathcal{S}|,a}]'$ is the a-th column of y. Under the assumption of fast mixing time and uniform ergodicity (Assumptions 2,3 of Chen et al. (2018a)), there exists constant t_{mix} and τ such that one can set the feasible regions $\mathcal{X}$ and $\mathcal{Y}$ as

$$\mathcal{X} := \{x \in \mathbf{R}^{|\mathcal{S}|} : \|x\|_\infty \le 2t_{mix}\}, \quad (25)$$

and

$$\mathcal{Y} := \Big\{y \in \mathbf{R}^{|\mathcal{S}|\times|\mathcal{A}|} : y \ge 0, \|y\|_1 = 1, \frac{\mathbf{1}}{\sqrt{\tau}|\mathcal{S}|} \le \sum_{a\in\mathcal{A}} y_a \le \frac{\sqrt{\tau}\cdot\mathbf{1}}{|\mathcal{S}|}\Big\}, \quad (26)$$

(see Appendix B.1 for details). In the policy learning setting, we do not know either P or r. Instead, we want to estimate the optimal policy π^* based on sample transitions.

We construct an unbiased sample of $P = \{P_a\}$ by generating one sample transition from every (s,a), i.e., $\xi := \big\{(s,a,s',\hat{r}_{sa}) : \forall s \in \mathcal{S}, a \in \mathcal{A}, s' \sim P_a(s,\cdot)\big\}$. In other words, each ξ consists of $|\mathcal{S}||\mathcal{A}|$ sample transitions. Thus we obtain a sample transition matrix $P_\xi = \{P_a\}$ where $P_{\xi,a}(s,s') = 1$ if s' is sampled and $P_{\xi,a}(s,s') = 0$ otherwise. Thus, we can define a stochastic sample of the objective function of (3) as

$$\Phi_\xi(x,y) = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} y_{sa}\hat{r}_{sa} - \sum_{a\in\mathcal{A}} y_a^T(P_{\xi,a} - I)x. \quad (27)$$

It is easy to see that $\Phi(x,y) = \mathbf{E}_\xi\left[\Phi_\xi(x,y)\right]$.

3.2 Efficiency of the Empirical Optimal Policy

To handle the bilinear objective function, we consider the regularized empirical saddle point (R-ESP) problem, given by

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\frac{\alpha_x}{2}\|x\|_2^2 + \Phi_\Gamma(x,y) - \alpha_y\sum_{s,a} y_{sa}\log(y_{sa}), \quad (28)$$

where $\Phi_\Gamma(x,y) := \frac{1}{|\Gamma|}\sum_{\xi\in\Gamma}\Phi_\xi(x,y)$ is the empirical objective, α_x,α_y are to be chosen later. Let $(\bar{x},\bar{y})$ be the solution to the R-ESP problem (28). Then we obtain the empirical optimal policy $\bar{\pi}$ given by

$$\bar{\pi}(a|s) := \bar{y}_{sa}/\Big(\sum_{a'\in\mathcal{A}}\bar{y}_{sa'}\Big), \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

The entropy regularizer $\sum_{s,a} y_{sa}\log(y_{sa})$ plays an important role in the analysis. It is 1-strongly convex in L_1-norm due to the Pinsker's inequality. To analyze the efficiency of $\bar{\pi}$, we will apply the generalization theory for SSP problem by choosing the norms as $\|\cdot\| := \|\cdot\|_2$ and $|||\cdot||| := \|\cdot\|_1$, respectively.

Theorem 5 (Sample Efficiency of Empirical Optimal Policy)**.** *Let* $\alpha_x = \frac{\tau^{3/2}}{\sqrt{n|\mathcal{S}|t_{mix}}}$, $\alpha_y = \frac{t_{mix}}{\sqrt{n\log(|\mathcal{S}||\mathcal{A}|)}}$. *Then the empirical optimal policy* $\bar{\pi}$ *satisfies*

$$\mathbf{E}\left[\hat{v}^* - v^{\bar{\pi}}\right] \le O\left(\frac{t_{mix}\tau}{\sqrt{n}} \cdot \left(\tau^{1.5} + \sqrt{\log(|\mathcal{S}||\mathcal{A}|)}\right)\right).$$

Consequently, to guarantee that $\mathbf{E}\left[\hat{v}^* - v^{\bar{\pi}}\right] \le \epsilon$, *we need* $n = \Omega\left(\frac{t_{mix}^2}{\epsilon^2}\left(\tau^5 + \tau^3\log(|\mathcal{S}||\mathcal{A}|)\right)\right)$. *Since each* ξ *consists of* $|\mathcal{S}||\mathcal{A}|$ *samples of state transitions, the total sample complexity will be* $|\mathcal{S}||\mathcal{A}| \cdot n = \tilde{O}\left(\frac{t_{mix}^2\tau^5|\mathcal{S}||\mathcal{A}|}{\epsilon^2}\right)$. *The* $|\mathcal{S}||\mathcal{A}|/\epsilon^2$ *dependence in this bound is optimal.*

Theorem 5 has several implications:

• The regularized empirical optimal policy $\bar{\pi}$ achieves a near-optimal sample complexity, which matches known upper/lower bounds in their dependences on $|\mathcal{S}|, |\mathcal{A}|, \epsilon$ Chen and Wang (2017). This result is somewhat surprising: It means that one can simply compute an empirical MDP and solve it for estimating the optimal policy. This approach is conceptually simple, yet has satisfying error bound.

• Also note that the transition matrix P contains $|\mathcal{S}|^2|\mathcal{A}|$ unknown variables, but the policy error of $\bar{\pi}$ scales with $|\mathcal{S}||\mathcal{A}|$ which is significantly smaller. Namely, one does not need to estimate the full matrix P but can still get a good policy estimator by solving the R-ESP.

• The proof of Theorem 5 is nontrivial because we want to evaluate $v^{\bar{\pi}}$, which is the average reward of the state-transition process if $\bar{\pi}$ is implemented. The first step of the proof is to apply the result of Lemma 2 to the R-ESP (28), so that we obtain a WGM upper bound as

$$\Delta^w(\bar{x}, \bar{y}) \le O\left(t_{mix} \cdot \left(\tau^{1.5} + \sqrt{\log(|\mathcal{S}||\mathcal{A}|)}\right) \cdot n^{-\frac{1}{2}}\right) \quad (29)$$

Then, we exploit the stationarity condition of the MDP (3) and prove that

$$\begin{aligned} \hat{v}^* - & \mathbf{E}\left[\sum_{a\in\mathcal{A}} \bar{y}_a^\top((P_a - I)x^* + r_a)\right] \\ = \quad & \mathbf{E}\big[\Phi(\bar{x}, y^*) - \Phi(x^*, \bar{y})\big] \\ \le \quad & \Delta^w(\bar{x}, \bar{y}). \end{aligned} \quad (30)$$

Finally, we use the uniform ergodicity property of the MDP to show that $\mathbf{E}\left[\hat{v}^* - v^{\bar{\pi}}\right] \le \tau\mathbf{E}\big[\hat{v}^* - \sum_{a\in\mathcal{A}} \bar{y}_a^\top((P_a - I)x^* + r_a)\big]$ (Chen et al. (2018a)), which further leads to our theorem. Proofs of (29), (30) are given in Appdices B.2,B.3.

4 Application to Stochastic Composite Optimization

Another example is stochastic composite optimization of the general form

$$\min_{x\in\mathcal{X}} F(x) := f\big(\mathbf{E}_\xi[g_\xi(x)]\big) + r(x).$$

Such problems have been studied in many literatures, see Nesterov (2007); Wang et al. (2017); Zhang and Xiao (2019) and references therein. Let us consider the case where f is convex and has Lipschitz continuous gradient and $g_\xi(x) = A_\xi x - b_\xi$ is linear, where A_ξ is a random matrix and b_ξ is a random vector. This case applies to inverse problems and policy evaluation with function approximation Wang et al. (2017).This problem is equivalent to an SSP problem

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi(x, y) = r(x) + \mathbf{E}_\xi\left[y^\top(A_\xi x - b_\xi)\right] - f^*(y), \quad (31)$$

where $f^*(y) = \sup_{x\in\mathcal{X}} y^\top x - f(x)$ is the convex conjugate of f. Suppose $\mathcal{X}$ is a convex compact set and both f and r are convex functions. To guarantee the SC-SC property of Φ, we assume the function r is μ-strongly convex and ℓ_r-Lipschtiz continuous and f is smooth and has L-Lipschitz continuous gradient. we also assume that the L_2 norms of the sample A_ξ and b_ξ are almost surely bounded.

Note that in general, we should choose $\mathcal{Y}$ to be the whole space. However, However, in (31), because we assume f is convex and L-smooth in the whole space, f^* is strongly convex. Since $\mathcal{X}, A_\xi, b_\xi$ are all required to be bounded, $\text{argmax}_y\, y^\top\mathbf{E}[A_\xi x - b_\xi] - f^*(y)$ is bounded for all $x \in \mathcal{X}$. Consequently, we assume the domain $\mathcal{Y}$ is convex and compact without hurting the optimality.

Due to the compactness of $\mathcal{Y}$, we also assume that f^* is ℓ_{f^*}-Lipschitz continuous in $\mathcal{Y}$. Since μ is often very small, for simplify the result, we assume that $L\mu = O(1)$.

Let $(\bar{x}, \bar{y})$ solve the ESP version of (31). Under these conditions, if directly applying Theorem 3, we would be able to prove

$$\mathbf{E}\left[F(\bar{x})\right] - \min_{x\in\mathcal{X}} F(x) \le O\left(n^{-1}\mu^{-1.5}\right), \quad (32)$$

which corresponds to an $O\left(\frac{1}{\mu^{1.5}\epsilon}\right)$ sample complexity (see its proof in Appendix C.1). In the next theorem, we show that the ESP solution is actually more sample efficient and has smaller error. The proof is given in Appendix C.2.

Theorem 6. *Let* $(\bar{x}, \bar{y})$ *be the ESP solution to* (31) *based on n i.i.d. samples. Then*

$$\mathbf{E}\left[F(\bar{x})\right] - \min_{x\in\mathcal{X}} F(x) \le \tilde{\Delta}(\bar{x}, \bar{y}) \le O\left(\frac{L\ell_{f^*} + \ell_r}{\mu \cdot n}\right) = O\left(\frac{1}{\mu \cdot n}\right),$$

where $\tilde{\Delta}(\bar{x}, \bar{y}) = \mathbf{E}\big[\max_{y\in\mathcal{Y}} \Phi(\bar{x}, y)\big] - \min_{x\in\mathcal{X}} \mathbf{E}\big[\Phi(x, \bar{y})\big]$ *is a hybrid duality gap measure between WGM and SGM.*

In other words, the ESP solution leads to an ϵ-optimal solution to the original composite problem, when the sample size satisfies $n \ge O(\frac{1}{\mu\epsilon})$. In terms of the dependence on n and μ, this result matches the best known sample complexity for this problem given by Zhao (2019). However, we should also point out that the dependence on L and σ_A is not tight yet. How the dependence on these constants can be improved remains a future task.

5 Application to Stochastic Games

Consider the two-player stochastic matrix game problem (5):

$$\min_{x\in\Delta_{N_1}} \max_{y\in\Delta_{N_2}} x^\top \mathbf{E}_\xi[A_\xi]y,$$

with $\Delta_{N_i} := \{z \in \mathbf{R}^{N_i} : z \geq 0, \mathbf{1}^\top z = 1\}$, $i = 1, 2$, and x, y denote the mixed strategies of players 1 and 2, respectively. Based on n i.i.d. samples of the payoff matrix (with nN_1N_2 individual sample payoffs), we estimate the Nash equilibrium (x^*, y^*) by constructing the following R-ESP

$$\min_{x\in\Delta_{N_1}} \max_{y\in\Delta_{N_2}} \frac{\sum_i x_i \log x_i}{\sqrt{n \log N_1}} + x^\top \Big(\frac{1}{n}\sum_{i=1}^n A_{\xi_i}\Big) y - \frac{\sum_j y_j \log y_j}{\sqrt{n \log N_2}}.$$

Let $(\bar{x}, \bar{y})$ be the solution to the preceding R-ESP, which is referred to as the *empirical Nash equilibrium*. Then the following theorem holds.

Theorem 7. *Assume* $\max_{i,j} |A_\xi(i, j)| \leq 1$ *almost surely. Therefore,*

$$\begin{aligned}\mathbf{E}[x^\top A\bar{y}] - O\left(\sqrt{\log(N_1N_2)/n}\right) &\leq \mathbf{E}[\bar{x}^\top A\bar{y}] \\ &\leq \mathbf{E}[\bar{x}^\top Ay] + O\left(\sqrt{\log(N_1N_2)/n}\right),\end{aligned}$$

for $\forall x \in \Delta_{N_1}, y \in \Delta_{N_2}$.

The theorem means that the empirical strategy $(\bar{x}, \bar{y})$ is an ϵ-Nash equilibrium with high probability, as long as the total number of sample payoffs is greater than $\tilde{O}(N_1N_2\epsilon^{-2})$. This sample complexity is statistically optimal.

References

Francis Bach and Kfir Y Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. *arXiv preprint arXiv:1902.01637*, 2019.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

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

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

Yichen Chen and Mengdi Wang. Lower bound on the computational complexity of discounted markov decision problems. *arXiv preprint arXiv:1705.07312*, 2017.

Yichen Chen, Lihong Li, and Mengdi Wang. Scalable bilinear pi learning using state and action features. *arXiv preprint arXiv:1804.10328*, 2018a.

Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018b.

Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165(1):113–149, 2017.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *arXiv preprint arXiv:2006.07201*, 2020.

Simon S Du and Wei Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. *arXiv preprint arXiv:1802.01504*, 2018.

Alon Gonen and Shai Shalev-Shwartz. Average stability is invariant to data preconditioning: Implications to exp-concave empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):8245–8257, 2017.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13(Jun):1865–1890, 2012.

Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural computation*, 11(6):1427–1453, 1999.

Tomer Koren and Kfir Levy. Fast rates for exp-concave empirical risk minimization. In *Advances in Neural Information Processing Systems*, pages 1477–1485, 2015.

Guanghui Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer, 2020.

Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

Nishant A Mehta. Fast rates with high probability in exp-concave statistical learning. *arXiv preprint arXiv:1605.01288*, 2016.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.

Michael Natole, Yiming Ying, and Siwei Lyu. Stochastic proximal algorithms for auc maximization. In *International Conference on Machine Learning*, pages 3710–3719, 2018.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Yu Nesterov. Modified Gauss–Newton scheme with worst case guarantees for global performance. *Optimisation methods and software*, 22(3):469–483, 2007.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Maziar Sanjabi, Meisam Razaviyayn, and Jason D Lee. Solving non-convex non-concave min-max games under Polyak-Lojasiewicz condition. *arXiv preprint arXiv:1812.02878*, 2018.

Shai Shalev-Shwartz and Yoram Singer. Online learning: Theory, algorithms, and applications. 2007.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, pages 2199–2207, 2010.

Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. In *Advances in neural information processing systems*, pages 1545–1552, 2009.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992.

Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Mengdi Wang. Primal-dual pi learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017.

Mengdi Wang, Ji Liu, and Ethan X Fang. Accelerating stochastic composition optimization. *The Journal of Machine Learning Research*, 18(1):3721–3743, 2017.

Lin Xiao, Adams Wei Yu, Qihang Lin, and Weizhu Chen. Dscovr: Randomized primal-dual block coordinate algorithms for asynchronous distributed optimization. *Journal of Machine Learning Research*, 20(43):1–58, 2019.

Yan Yan, Yi Xu, Qihang Lin, Lijun Zhang, and Tianbao Yang. Stochastic primal-dual algorithms with faster convergence than $O(1/\sqrt{T})$ for problems without bilinear structure. *arXiv preprint arXiv:1904.10112*, 2019.

Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Sharp analysis of epoch stochastic gradient descent ascent methods for min-max optimization. *arXiv preprint arXiv:2002.05309*, 2020.

Junyu Zhang and Lin Xiao. A stochastic composite gradient method with incremental variance reduction. In *Advances in Neural Information Processing Systems*, pages 9075–9085, 2019.

Lijun Zhang, Tianbao Yang, and Rong Jin. Empirical risk minimization for stochastic convex optimization: $o(1/n)$ and $o(1/n^2)$-type of risk bounds. *arXiv preprint arXiv:1702.02030*, 2017.

Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):2939–2980, 2017.

Renbo Zhao. Optimal algorithms for stochastic three-composite convex-concave saddle point problems. *arXiv preprint arXiv:1903.01687*, 2019.