Differentially Private Algorithms for the Stochastic Saddle Point Problem with Optimal Rates for the Strong Gap

Raef Bassily BASSILY.1@OSU.EDU
Department of Computer Science & Engineering and the Translational Data Analytics Institute (TDAI), The Ohio State University

Cristóbal Guzmán CRGUZMANP@MAT.UC.CL
Institute for Mathematical and Computational Eng., Facultad de Matemáticas & Escuela de Ingeniería Pontificia Universidad Católica de Chile

Michael Menart MENART.2@OSU.EDU
Department of Computer Science & Engineering, The Ohio State University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We show that convex-concave Lipschitz stochastic saddle point problems (also known as stochastic minimax optimization) can be solved under the constraint of (ϵ, δ)-differential privacy with *strong (primal-dual) gap* rate of $\tilde{O}\big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\epsilon}\big)$, where n is the dataset size and d is the dimension of the problem. This rate is nearly optimal, based on existing lower bounds in differentially private stochastic convex optimization. Specifically, we prove a tight upper bound on the strong gap via novel implementation and analysis of the recursive regularization technique repurposed for saddle point problems. We show that this rate can be attained with $O\big(\min\big\{\frac{n^2\epsilon^{1.5}}{\sqrt{d}}, n^{3/2}\big\}\big)$ gradient complexity, and $\tilde{O}(n)$ gradient complexity if the loss function is smooth. As a byproduct of our method, we develop a general algorithm that, given a black-box access to a subroutine satisfying a certain α primal-dual accuracy guarantee with respect to the empirical objective, gives a solution to the stochastic saddle point problem with a strong gap of $\tilde{O}(\alpha + \frac{1}{\sqrt{n}})$. We show that this α-accuracy condition is satisfied by standard algorithms for the empirical saddle point problem such as the proximal point method and the stochastic gradient descent ascent algorithm. Finally, to emphasize the importance of the strong gap as a convergence criterion compared to the weaker notion of primal-dual gap, commonly known as the *weak gap*, we show that even for simple problems it is possible for an algorithm to have zero weak gap and suffer from $\Omega(1)$ strong gap. We also show that there exists a fundamental tradeoff between stability and accuracy. Specifically, we show that any Δ-stable algorithm has empirical gap $\Omega\big(\frac{1}{\Delta n}\big)$, and that this bound is tight. This result also holds also more specifically for empirical risk minimization problems and may be of independent interest.

Keywords: Differential Privacy, Stochastic Saddle Point Problem, Strong Gap, Stochastic Minimax Optimization, Algorithmic Stability

1. Introduction

Stochastic (convex-concave) saddle point problems (SSP)[1] (also referred to in the literature as stochastic minimax optimization problems) are an increasingly important model for modern machine learning, arising in areas such as stochastic optimization (Nemirovski et al., 2009; Juditsky

1. In this work, we will exclusively focus on the case where the function of interest for the stochastic saddle-point problem is convex-concave, and therefore we will omit it from the problem denomination.

et al., 2011; Zhang and Lin, 2015), robust statistics (Yu et al., 2021), and algorithmic fairness (Mohri et al., 2019; Williamson and Menon, 2019).

On the other hand, the reliance of modern machine learning on large datasets has led to concerns of user privacy. These concerns in turn have led to a variety of privacy standards, of which differential privacy (DP) has become the premier standard. However, for a variety of machine learning problems it is known that their differentially-private counterparts have provably worse rates. As such, characterizing the fundamental cost of differential privacy has become an important problem.

Currently, the theory of solving SSPs under differential privacy has major limitations, compared to its non-private counterpart. To illustrate this point, we need to discuss the notions of accuracy used in the literature. In SSPs, the goal is to find an approximate solution of the problem

$$\min_{w\in\mathcal{W}} \max_{\theta\in\Theta} \Big\{ F_{\mathcal{D}}(w,\theta) := \mathbb{E}_{x\sim\mathcal{D}}[f(w,\theta;x)] \Big\}, \tag{1}$$

where $\mathcal{D}$ is an unknown distribution for which we have access to an i.i.d. sample S. Given a (randomized) algorithm $\mathcal{A}$ with output $[\mathcal{A}_w(S), \mathcal{A}_\theta(S)] \in \mathcal{W}\times\Theta$, two studied measures of performance are the *strong and weak gap*[2], defined respectively as

$$\mathrm{Gap}(\mathcal{A}) = \mathop{\mathbb{E}}_{\mathcal{A},S}\left[\max_{\theta\in\Theta}\{F_{\mathcal{D}}(\mathcal{A}_w(S),\theta)\} - \min_{w\in\mathcal{W}}\{F_{\mathcal{D}}(w,\mathcal{A}_\theta(S))\}\right], \tag{2}$$

$$\mathrm{Gap}_{\mathsf{weak}}(\mathcal{A}) = \mathop{\mathbb{E}}_{\mathcal{A}}\left[\max_{\theta\in\Theta}\left\{\mathop{\mathbb{E}}_{S}[F_{\mathcal{D}}(\mathcal{A}_w(S),\theta)]\right\} - \min_{w\in\mathcal{W}}\left\{\mathop{\mathbb{E}}_{S}[F_{\mathcal{D}}(w,\mathcal{A}_\theta(S))]\right\}\right]. \tag{3}$$

It is easy to see that the strong gap upper bounds the weak gap, and thus it is a stronger accuracy measure. On the other hand, even for simple problems, the difference between these measures can be $\Omega(1)$; a fact we elaborate on in Section 5. We also note that the strong gap has a clear game-theoretic interpretation: if we consider $\mathcal{A}_w(S)$ and $\mathcal{A}_\theta(S)$ as the actions of two players in a (stochastic) zero-sum game, the strong gap upper bounds the most profitable unilateral deviation for either of the two players. In game theory this is known as an approximate Nash equilibrium. By contrast, there is no general guarantee associated with the weak gap.

Non-privately, it is known how to achieve optimal rates w.r.t. the strong gap, and those rates are similar to those established for stochastic convex optimization (SCO) (Nemirovski et al., 2009; Juditsky et al., 2011). However, for DP methods optimal rates are only known for the weak gap (Boob and Guzmán, 2023; Yang et al., 2022; Zhang et al., 2022). In a nutshell, the main limitation of these approaches is that –in order to amplify privacy– they make multiple passes over the data (e.g., by sampling with replacement stochastic gradients from the dataset), and the existing theory of generalization for SSPs is much more limited than it is for SCO (Zhang et al., 2021; Lei et al., 2021; Ozdaglar et al., 2022). Our approach largely circumvents the current limitations of generalization theory for SSPs, providing the first nearly-optimal rates for the strong gap in DP-SSP.

1.1. Contributions

In this work, we establish the optimal rates on the strong gap for DP-SSP. In the following, we let n be the number of samples, d be the dimension, and ϵ, δ be the privacy parameters. Our main result is an (ϵ,δ)-DP algorithm for SSP whose strong gap is $\tilde{O}\big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\epsilon}\big)$. This rate is nearly

2. The weak gap is sometimes stated with $\mathbb{E}_{\mathcal{A}}[\cdot]$ taken inside the max. However Boob and Guzmán (2023) showed this was not necessary to obtain the stability implies generalization result used in various works.

optimal, due to matching lower bounds for differentially private SCO (Bassily et al., 2014, 2019). These minimization lower bounds hold for saddle point problems since minimization problems are a special case of saddle point problems when Θ is constrained to be a singleton. For non-smooth loss function, we show this rate can be obtained in gradient complexity $O\big(\min\big\{\frac{n^2\epsilon^{1.5}}{\sqrt{d}}, n^{3/2}\big\}\big)$. This improves even upon the previous best known running time for achieving analogous rates on the *weak gap*, which was $n^{5/2}$ (Yang et al., 2022). Furthermore, we show that if the loss function is smooth, this rate can be achieved in nearly linear gradient complexity.

In order to obtain an upper bound for this problem, we present a novel analysis of the recursive regularization algorithm of Allen-Zhu (2018). Our work is the first to show how the sequential regularization approach can be repurposed to provide an algorithmic framework for attaining optimal strong gap guarantees for DP-SSP. As a byproduct of our analysis, we show that empirical saddle point solvers which satisfy a certain α accuracy guarantee can be used as a black box to obtain an $\tilde{O}\left(\alpha + 1/\sqrt{n}\right)$ guarantee on the strong (population) gap. This class of algorithms includes common techniques such as the proximal point method, the extragradient method, and stochastic gradient descent ascent (SGDA) (Mokhtari et al., 2020; Nemirovski, 2004; Juditsky et al., 2011). This fact may be of interest independent of differential privacy, as to the best of our knowledge, existing algorithms which achieve the optimal $1/\sqrt{n}$ rate on the strong population gap rely crucially on a one-pass structure which optimizes the population gap directly (Nemirovski et al., 2009).

Under the additional assumption that the loss function is smooth, we show that it is possible to use recursive regularization to obtain the optimal strong gap rate in nearly linear time. We here leverage accelerated algorithms for smooth and strongly convex/strongly concave loss functions (Palaniappan and Bach, 2016; Jin et al., 2022).

Our results stand in contrast to previous work on DP-SSPs, which has achieved optimal rates only for the weak gap and has crucially relied on "stability implies generalization" results for the weak gap. In this vein, we prove that even for simple problems, the strong and weak gap may differ by $\Theta(1)$. We also elucidate the challenges of extending existing techniques to strong gap guarantees by showing a fundamental tradeoff between stability and empirical accuracy. Specifically, we show that even for the more specific case of empirical risk minimization, any algorithm which is Δ-uniform argument stable algorithm must have empirical risk $\Omega\left(\frac{1}{\Delta n}\right)$. We also show this bound is tight, and note that it may be of independent interest. Such a tradeoff was also investigated by Chen et al. (2018), but their result only implies such a tradeoff for the specific case of $\Delta = \frac{1}{\sqrt{n}}$ and their proof technique is unrelated to ours.

1.2. Related Work

Differentially private stochastic optimization has been extensively studied for over a decade (Jain et al., 2012; Bassily et al., 2014; Jain and Thakurta, 2014; Talwar et al., 2015; Bassily et al., 2019; Feldman et al., 2020b; Asi et al., 2021; Bassily et al., 2021). Among such problems, stochastic convex minimization (where problem parameters are measured in the ℓ_2-norm) is perhaps the most widely studied, where it is known the optimal rate is $\tilde{O}(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\epsilon})$ (Bassily et al., 2019, 2014). Further, under smoothness assumptions such rates can be obtained in linear (in the sample size) gradient complexity (Feldman et al., 2020a). Without smoothness, no linear time algorithms which achieve the optimal rates are known (Kulkarni et al., 2021).

The study of stochastic saddle point problems under differential privacy is comparatively newer. In the non-private setting, optimal $O(1/\sqrt{n})$ guarantees on the strong gap have been known as far

back as Nemirovski and Yudin (1978). Under privacy (without strong convexity/strong concavity), optimal rates are known only for the *weak gap*. These rates $\tilde{O}(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\epsilon})$ have been obtained by several works (Boob and Guzmán, 2023; Yang et al., 2022; Zhang et al., 2022). The work of Zhang et al. (2022) additionally showed that under smoothness assumptions such a result could be obtained in near linear gradient complexity by leveraging accelerated methods (Jin et al., 2022; Palaniappan and Bach, 2016). All of these results are for the weak gap and they rely crucially on the fact that, for the weak gap, Δ-stability implies Δ-generalization Zhang et al. (2021).

By contrast, for the strong gap (without strong convexity/strong concavity assumptions), the best stability implies generalization result is a $\sqrt{\Delta}$ bound obtained by Ozdaglar et al. (2022) provided the loss is smooth. As a result of this discrepancy, known bounds on the strong gap under privacy are worse. The best known rates for the strong gap are $O\left(\min\left(\frac{d^{1/4}}{\sqrt{n\epsilon}}, \frac{1}{n^{1/3}} + \frac{\sqrt{d}}{n^{2/3}\epsilon}\right)\right)$ (Boob and Guzmán, 2023). This rate was obtained through of mixture of noisy stochastic extragradient and noisy inexact proximal point methods, avoiding stability arguments altogether and instead relying on one-pass algorithms which optimize the population loss directly. Without smoothness, we are not aware of any work which provides bounds on the strong gap under privacy, but one may note that a straightforward implementation of one-pass noisy SGDA leads to a rate of $O\big(\frac{\sqrt{d}}{\sqrt{n\epsilon}}\big)$ in this setting. We give these details in Appendix A.2 and note this same algorithm establishes the optimal rate for SSPs under local differential privacy.

Finally, under the stringent assumptions of μ-strong convexity/strong concavity (μ-SC/SC) and smoothness with constant condition number, κ, optimal rates on the strong gap have been obtained (Zhang et al., 2022). Under these assumptions, the optimal rate of $O\big(\frac{1}{\mu n} + \frac{d}{\mu n^2 \epsilon^2}\big)$ was achieved by leveraging the fact that Δ stability implies $\kappa\Delta$ generalization Zhang et al. (2021). The lower bound for this rate comes from lower bounds for the minimization setting (Hazan and Kale, 2014; Bassily et al., 2019).

2. Preliminaries

Throughout, we consider the space $\mathbb{R}^d$ endowed with the standard ℓ_2 norm $\|\cdot\|$. Let the primal parameter space $\mathcal{W}$ and the dual parameter space Θ be compact convex sets such that $\mathcal{W} \times \Theta \subset \mathbb{R}^d$ for some $d > 0$. Let $\mathcal{D}$ be some distribution over data domain $\mathcal{X}$. Consider the *stochastic saddle-point problem* given in equation (1) for some loss function f that is convex w.r.t. w and concave w.r.t. θ. We define the corresponding population loss and empirical loss functions as $F_{\mathcal{D}}(w,\theta) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[f(w,\theta;x)]$ and $F_S(w,\theta) = \frac{1}{n}\sum_{x\in S} f(w,\theta;x)$ respectively. For some $B > 0$ we assume that $\max_{u,u'\in\mathcal{W}\times\Theta} \|u - u'\| \leq B$. To simplify notation, for vectors $w \in \mathcal{W}$ and $\theta \in \Theta$, we will use $[w,\theta]$ to denote their concatenation, noting $[w,\theta]$ is a vector in $\mathbb{R}^d$. We primarily consider the case where f is L-Lipschitz, but will also consider the additional assumption of β-smoothness for certain results[3]. Specifically, these assumptions are that $\forall w_1, w_2 \in \mathcal{W}$ and $\forall \theta_1, \theta_2 \in \Theta$:

$$\begin{aligned} &\text{Lipschitzness:} && |f(w_1,\theta_1;x) - f(w_2,\theta_2;x)| \leq L\,\|[w_1,\theta_1] - [w_2,\theta_2]\| \\ &\text{Smoothness:} && \left\|\nabla_{[w,\theta]} f(w_1,\theta_1;x) - \nabla_{[w,\theta]} f(w_2,\theta_2;x)\right\| \leq \beta\,\|[w_1,\theta_1] - [w_2,\theta_2]\|\,. \end{aligned}$$

Under such assumptions (in fact, smoothness is not necessary), a solution for problem (1) always exists (Sion, 1958), which we will call as a *saddle point* onwards. Further, given an SSP (1), we will denote a saddle point as $[w^*,\theta^*]$.

3. Throughout, any properties for f are considered as a function of $[w,\theta]$. No assumptions about f w.r.t. x are made.

Gap functions In addition to the strong and weak gap functions defined in equations (2) and (3), it will be useful to define the following *gap function* expressed as a function of the parameter vector instead of the algorithm, $\widehat{\mathrm{Gap}}(\bar{w}, \bar{\theta}) = \max_{\theta \in \Theta} \{F_{\mathcal{D}}(\bar{w}, \theta)\} - \min_{w \in \mathcal{W}} \left\{F_{\mathcal{D}}(w, \bar{\theta})\right\}$.

We have the following useful fact regarding $\widehat{\mathrm{Gap}}$ (see Appendix A for a proof).

Fact 1 *If f is L-Lipschitz then $\widehat{\mathrm{Gap}}$ is $\sqrt{2}L$-Lipschitz.*

Note the strong gap can be written as an expectation of the gap function. Further, since the gap function is zero if and only if $(\bar{w}, \bar{\theta})$ is a solution for problem (1), the strong gap is considered the most suitable measure of accuracy for SSPs (Nemirovski et al., 2010; Juditsky et al., 2011). We also define the empirical gap as, $\mathrm{Gap}_S(\mathcal{A}) = \mathop{\mathbb{E}}_{\mathcal{A}}\left[\max_{\theta \in \Theta} \{F_S(\mathcal{A}_w(S), \theta)\} - \min_{w \in \mathcal{W}} \{F_S(w, \mathcal{A}_\theta(S))\}\right]$. We will consider at various points the notion of *generalization error* with respect to the strong/weak gap, which refers to difference between the strong/weak gap and the empirical gap. Note that because the empirical gap treats the dataset as a fixed quantity, there are not differing strong and weak versions of the empirical gap.

Saddle Operator Define the *saddle operator* as $g(w, \theta; x) = [\nabla_w f(w, \theta; x), -\nabla_\theta f(w, \theta; x)]$. Similarly define $G_{\mathcal{D}}(w, \theta) = \mathbb{E}_{x \sim \mathcal{D}}[g(w, \theta; x)]$ and $G_S(w, \theta) = \frac{1}{n}\sum_{x \in S} g(w, \theta; x)$. Note that the assumption on the smoothness of f implies the Lipschitzness of g. We note that since the saddle operator can be computed using one computation of the gradient, we refer indistinctly to saddle operator complexity or gradient complexity when discussing the running time of our algorithms.

Stability We will also use the notion of uniform argument stability frequently in our analysis (Bousquet and Elisseeff, 2002).

Definition 1 *A randomized algorithm $\mathcal{A} : \mathcal{X}^n \mapsto \mathcal{W} \times \Theta$ satisfies Δ-uniform argument stability if for any pair of adjacent datasets $S, S' \in \mathcal{X}^n$ it holds that $\mathop{\mathbb{E}}_{\mathcal{A}}\left[\|\mathcal{A}(S) - \mathcal{A}(S')\|\right] \leq \Delta$.*

A fact we will use is that the (constrained) regularized saddle-point is stable. Specifically, for some $\hat{w} \in \mathcal{W}$, $\hat{\theta} \in \Theta$, and $\lambda \geq 0$ consider the regularized objective function

$$(w, \theta) \mapsto \frac{1}{n}\sum_{z \in S} f(w, \theta; z) + \frac{\lambda}{2}\|w - \hat{w}\|^2 - \frac{\lambda}{2}\|\theta - \hat{\theta}\|^2. \tag{4}$$

It is easy to see that his problem has a unique saddle point. The mapping which selects its output according the unique solution of (4) has the following stability property.

Lemma 2 *(Zhang et al., 2021, Lemma 1) The algorithm which outputs the regularized saddle point with parameters $\lambda > 0$, $\hat{w} \in \mathcal{W}$ and $\hat{\theta} \in \Theta$, is $\left(\frac{2L}{\lambda n}\right)$-uniform argument stable w.r.t. S.*

In addition to the stability of the regularized saddle point, we will also frequently use the following fact.

Lemma 3 *(Zhang et al., 2021, Theorem 1) Let $h : \mathcal{W} \times \Theta \mapsto \mathbb{R}$ be λ-SC/SC with saddle point $[w^*, \theta^*]$ and gap function $\widehat{\mathrm{Gap}}^h$. For any $[w, \theta] \in \mathcal{W} \times \Theta$ it holds that $\|[w, \theta] - [w^*, \theta^*]\|^2 \leq \frac{2(h(w, \theta^*) - h(w^*, \theta))}{\lambda} \leq \frac{2}{\lambda}\widehat{\mathrm{Gap}}^h(w, \theta)$.*

Differential Privacy (DP) Dwork et al. (2006): An algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if for all datasets S and S' differing in one data point and all events $\mathcal{E}$ in the range of the $\mathcal{A}$, we have, $\mathbb{P}\left(\mathcal{A}(S) \in \mathcal{E}\right) \leq e^{\epsilon}\mathbb{P}\left(\mathcal{A}(S') \in \mathcal{E}\right) + \delta$.

3. From Empirical Saddle Point to Strong Gap Guarantee via Recursive Regularization

Our approach for obtaining near optimal rates on the strong gap leverages the recursive regularization technique of Allen-Zhu (2018). In addition to adapting this algorithm to fit SSP problems, we also provide a novel analysis which differs substantially from the analysis presented in previous work (Foster et al., 2019; Arora et al., 2022).

Algorithm 1 Recursive Regularization: $\mathcal{R}$

Require: Dataset $S \in \mathcal{X}^n$, loss function f, subroutine $\mathcal{A}_{\mathsf{emp}}$, regularization parameter $\lambda \geq \frac{L}{B\sqrt{n}}$, constraint set diameter B, Lipschitz constant L.

Let $n' = n/\log_2(n)$, and $T = \log_2(\frac{L}{B\lambda})$.
Let $S_1, ..., S_T$ be a disjoint partition of S with each S_t of size n' *(which is always possible due to the condition on λ)*
Let $[\bar{w}_0, \bar{\theta}_0]$ be any point in $\mathcal{W} \times \Theta$
Define function $(w, \theta, x) \mapsto f^{(1)}(w, \theta; x) = f(w, \theta; x) + 2\lambda \left\|w - \bar{w}_0\right\|^2 - 2\lambda \left\|\theta - \bar{\theta}_0\right\|^2$
for $t = 1$ to T **do**
  $[\bar{w}_t, \bar{\theta}_t] = \mathcal{A}_{\mathsf{emp}}\left(S_t, f^t, [\bar{w}_{t-1}, \bar{\theta}_{t-1}], \frac{B}{2^t}\right)$
  Define $(w, \theta, x) \mapsto f^{(t+1)}(w, \theta; x) = f^{(t)}(w, \theta; x) + 2^{t+1}\lambda \left\|w - \bar{w}_t\right\|^2 - 2^{t+1}\lambda \left\|\theta - \bar{\theta}_t\right\|^2$
end for
Output: $[\bar{w}_T, \bar{\theta}_T]$

Our recursive regularization algorithm works by solving a series of regularized objectives, $f^{(1)}, ..., f^{(T)}$, with increasingly large regularization parameters. Specifically, after solving the t'th objective to obtain $[\bar{w}_t, \bar{\theta}_t]$, the algorithm creates a new objective which is $f^{(t+1)}(w, \theta; x) = f^{(t)}(w, \theta; x) + 2^{t+1}\lambda \left\|w - \bar{w}_t\right\|^2 - 2^{t+1}\lambda \left\|\theta - \bar{\theta}_t\right\|^2$ for the subsequent round. Notice that each subsequent objective is easier in the sense that the strong convexity parameter is larger.

Our analysis will leverage the fact that approximate solutions to intermediate objectives do not need to obtain good bounds on the strong gap for the regularization parameter to be increased. This is in contrast to, for example, the *iterative* regularization technique of Zhang et al. (2022), which finds $[w, \theta]$ that satisfies a near optimal (weak) gap bound before adding noise.

Empirical Subroutine Recursive regularization utilizes a subroutine, $\mathcal{A}_{\mathsf{emp}}$, which is roughly an approximate empirical saddle point solver. In addition to a dataset and Lipschitz loss function, $\mathcal{A}_{\mathsf{emp}}$ takes as input an initial point and a bound, $\hat{D}$, on the expected distance between the initial point and the saddle point of the empirical loss defined over the input dataset. At round $t \in [T]$ this distance is bounded by $\frac{B}{2^t}$, allowing the algorithm to obtain increasingly strong accuracy guarantees for each subproblem. Note also it can be verified that for all $t \in [T]$, $f^{(t)}$ is $O(L)$-Lipschitz due the scaling of the regularization. Specifically, the accuracy guarantee of interest is the following.

Definition 4 ($\hat{\alpha}$-relative accuracy) *Given a dataset $S' \in \mathcal{X}^{n'}$, loss function f', and an initial point $[w', \theta']$, we say that $\mathcal{A}_{\mathsf{emp}}$ satisfies $\hat{\alpha}$-relative accuracy w.r.t. the empirical saddle point $[w^*_{S'}, \theta^*_{S'}]$ of $F'_{S'}(w, \theta) = \frac{1}{n}\sum_{x \in S'} f'(w, \theta; x)$ if, $\forall \hat{D} > 0$, whenever $\mathbb{E}\left[\left\|[w', \theta'] - [w^*_{S'}, \theta^*_{S'}]\right\|\right] \leq \hat{D}$, the output $[\bar{w}, \bar{\theta}]$ of $\mathcal{A}_{\mathsf{emp}}$ satisfies $\mathbb{E}\left[F'_{S'}(\bar{w}, \theta^*_{S'}) - F'_{S'}(w^*_{S'}, \bar{\theta})\right] \leq \hat{D}\hat{\alpha}$.*

The relative accuracy guarantee for $\mathcal{A}_{\mathsf{emp}}$ differs from the more standard gap guarantee, and is not necessarily implied by a bound on the empirical gap. The motivation for this notion of accuracy is twofold. First, when the loss function is additionally SC/SC, this guarantee is sufficient to provide a bound on the distance between the *output* of $\mathcal{A}_{\mathsf{emp}}$ and the saddle point, which will play a crucial role in our convergence proof for Algorithm 1. Second, while it is certainly true that a bound on the empirical gap implies the same bound on $\mathbb{E}\left[F_S(\bar{w},\theta) - F_S(w,\bar{\theta})\right]$, for any given $[w,\theta]$, it is not necessarily the case that the gap itself may enjoy a bound that is proportional to the initial distance to the saddle point[4]. The reason is that the gap function is defined by a supremum that is taken w.r.t. the whole feasible set $\mathcal{W}\times\Theta$, and thus the information of the evaluation of the objective w.r.t. particular points is lost. However, it is usually the case that saddle point solvers provide a bound of the form $F_S(\bar{w},\theta) - F_S(w,\bar{\theta}) \leq \|[w,\theta]-[w',\theta']\|\,\hat{\alpha}$, for all $[w,\theta]\in\mathcal{W}\times\Theta$, and some initial point $[w',\theta']\in\mathcal{W}\times\Theta$. Algorithms such as the proximal point method, extragradient method, and SGDA (with appropriately tuned learning rate) satisfy this condition, and thus satisfy the condition for relative accuracy (Mokhtari et al., 2020; Nemirovski, 2004; Juditsky et al., 2011).

Guarantees of Recursive Regularization Given such an algorithm, recursive regularization achieves the following guarantee.

Theorem 5 *Let $\mathcal{A}_{\mathsf{emp}}$ satisfy $\hat{\alpha}$-relative accuracy for any $(5L)$-Lipschitz loss function and dataset of size $n' = \frac{n}{\log(n)}$. Then Algorithm 1, run with $\mathcal{A}_{\mathsf{emp}}$ as a subroutine and $\lambda = \frac{48}{B}\left(\hat{\alpha} + \frac{L}{\sqrt{n'}}\right)$, satisfies*

$$\mathrm{Gap}(\mathcal{R}) = O\left(\log(n)B\hat{\alpha} + \frac{\log^{3/2}(n)BL}{\sqrt{n}}\right).$$

Recall that B is a bound on the diameter of the constraint set. In the following, we will sketch the proof of this theorem and highlight key lemmas. We defer the full proof to Appendix B.2. For simplicity, let us here consider the case where $\hat{\alpha}=0$. A crucial aspect of our proof is that we avoid the need to bound the strong gap of the actual iterates, $\{\bar{w}_t\}_{t=1}^{T-1}$. Instead, we bound the strong gap of the *expected* iterates, where the expectation is taken with respect to S_t. More concretely, consider some $t\in[T]$ and let $\mathcal{B}$ be the algorithm which on input $[\bar{w}_{t-1},\bar{\theta}_{t-1}]$ outputs $\mathop{\mathbb{E}}_{S_t,\mathcal{A}_{\mathsf{emp}}}\left[\mathcal{A}_{\mathsf{emp}}(S_t,f^t,[\bar{w}_{t-1},\bar{\theta}_{t-1}],\frac{B}{2^t})\right]$. Note $\mathcal{B}$ is deterministic and data independent. As a result, it is possible to prove bounds on the strong gap of $\mathcal{B}$.

Lemma 6 *Let $S\sim\mathcal{D}^n$. For any Δ-uniform argument stable algorithm $\mathcal{A}$, it holds that*

$$\widehat{\mathrm{Gap}}\left(\mathop{\mathbb{E}}_{\mathcal{A},S}\left[\mathcal{A}_w(S)\right], \mathop{\mathbb{E}}_{\mathcal{A},S}\left[\mathcal{A}_\theta(S)\right]\right) \leq \mathrm{Gap}_{\mathsf{weak}}(\mathcal{A}) \leq \mathop{\mathbb{E}}_{S}\left[\mathrm{Gap}_S(\mathcal{A})\right] + \Delta L.$$

The proof follows straightforwardly from an application of Jensen's inequality and the "stability implies generalization" result for the weak gap (Lei et al., 2021, Theorem 1). We give full details in Appendix B.1. Note that, for this discussion, the LHS of the above is equal to $\mathrm{Gap}(\mathcal{B})$ when we apply this lemma to the data batch S_t and subroutine $\mathcal{A}_{\mathsf{emp}}$.

In fact, running $\mathcal{B}$ is infeasible. Instead, we show that the output $\mathcal{A}_{\mathsf{emp}}$ is close to the output of $\mathcal{B}$. This in turn can be accomplished using the fact that bounded stability implies bounded variance. Concretely, we use the vector valued version of McDiarmid's inequality.

4. (Farnia and Ozdaglar, 2020, Theorem 4) claims such a bound on the primal risk, but this is due to a misapplication of (Mokhtari et al., 2020, Lemma 2).

Lemma 7 *(Rivasplata et al., 2018, Lemma 6)* [5] *Let $\mathcal{A}$ be deterministic Δ-uniform argument stable stable with respect to $S \sim \mathcal{D}^n$. Then its output satisfies $\mathbb{E}\left[\left\|\mathcal{A}(S) - \mathbb{E}_{\hat{S}\sim\mathcal{D}^n}\left[\mathcal{A}(\hat{S})\right]\right\|^2\right] \leq n\Delta^2$.*

Observe that the exact empirical saddle point is a deterministic quantity conditioned on the randomness of the t'th empirical objective. Using the fact that $(2^t\lambda)$-regularization implies $\left(\frac{L}{2^t\lambda n'}\right)$-stability of the empirical saddle point in conjunction with the above lemma, we obtain a (conditional) variance bound of $\frac{L^2}{2^{2t}\lambda^2 n'}$. Under the setting of $\lambda = \Omega(\frac{L}{B\sqrt{n'}})$, we can ultimately prove that the distance between the output of $\mathcal{A}_{\mathsf{emp}}$ and $\mathcal{B}$ (at round t) is $O(\frac{B}{2^t})$. Since the strong gap of $\mathcal{B}$ with respect to $F_{\mathcal{D}}^{(t)}(w,\theta) := \mathbb{E}_{x\sim\mathcal{D}}[f^{(t)}(w,\theta;x)]$ is at most $\Delta L = \frac{L^2}{2^t\lambda n'}$ by Lemma 6 (recall we here assume $\hat{\alpha} = 0$ for simplicity) and $F_{\mathcal{D}}^{(t)}$ is $(2^{t+1}\lambda)$-SC/SC, the output of $\mathcal{B}$ must in turn be close to the population saddle point. Specifically, this distance is also bounded as $\left(\frac{\Delta L}{2^t\lambda}\right)^{1/2} = \frac{L}{\sqrt{2^t\lambda n'}}\frac{1}{\sqrt{2^t\lambda}} = O(\frac{B}{2^t})$. Thus we ultimately have that the distance between $[\bar{w}_t, \bar{\theta}_t]$ and the population saddle point of $F_{\mathcal{D}}^{(t)}$, $[w_t^*, \theta_t^*]$, satisfies $\mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w_t^*, \theta_t^*]\right\|\right] = O(\frac{B}{2^t})$. These ideas also lead to a bound $\mathbb{E}\left[\left\|[w_{t+1}^*, \theta_{t+1}^*] - [\bar{w}_t, \bar{\theta}_t]\right\|\right] = O(\frac{B}{2^t})$, although the argument in this case is more technical and thus deferred to the full proof.

The upshot of this analysis is that as the level of regularization increases, the distance of the iterates to the their respective population minimizers decreases in kind. One consequence of this fact is that $\left\|[\bar{w}_T, \bar{\theta}_T] - [w_T^*, \theta_T^*]\right\| = \tilde{O}\left(\frac{B}{\sqrt{n}}\right)$, and thus by the Lipschitzness of the gap function, the output of recursive regularization has a gap bound close to that of $[w_T^*, \theta_T^*]$. Turning now towards the utility of $[w_T^*, \theta_T^*]$, using the fact that $F_{\mathcal{D}}$ is convex-concave we have

$$\widehat{\mathrm{Gap}}(w_T^*, \theta_T^*) \leq \max_{w'\in\mathcal{W},\theta'\in\Theta}\left\{\left\langle G_{\mathcal{D}}(w_T^*, \theta_T^*), [w_T^*, \theta_T^*] - [w', \theta']\right\rangle\right\}.$$

Further, an expression for $G_{\mathcal{D}}$ be obtained using the definition of $F_{\mathcal{D}}^{(T)}$:

$$G_{\mathcal{D}}(w_T^*, \theta_T^*) = G_{\mathcal{D}}^{(T)}(w_T^*, \theta_T^*) - 2\lambda\textstyle\sum_{t=0}^{T-1} 2^{t+1}([w_T^*, -\theta_T^*] - [\bar{w}_t, -\bar{\theta}_t]),$$

where $G_{\mathcal{D}}^{(T)}$ is the saddle operator of $F_{\mathcal{D}}^{(T)}$. Plugging the latter into the former and using Cauchy-Schwarz inequality, the triangle inequality, and the fact that $[w_T^*, \theta_T^*]$ is the exact saddle point of $F_{\mathcal{D}}^{(T)}$, one can obtain a bound on the gap in terms of the distances discussed previously.

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mathrm{Gap}}(w_T^*, \theta_T^*)\right] &\leq 4B\cdot\mathbb{E}\left[\lambda\sum_{t=0}^{T-1} 2^t\left\|[w_T^*, \theta_T^*] - [\bar{w}_t, \bar{\theta}_t]\right\|\right] \\
&\overset{(i)}{\leq} 4B\cdot\mathbb{E}\left[\lambda\sum_{t=0}^{T-1} 2^t\left(\left\|[w_{t+1}^*, \theta_{t+1}^*] - [\bar{w}_t, \bar{\theta}_t]\right\| + \sum_{r=t+1}^{T-1}\left\|[w_{r+1}^*, \theta_{r+1}^*] - [w_r^*, \theta_r^*]\right\|\right)\right] \\
&\overset{(ii)}{=} O\left(B\sum_{t=0}^{T-1} 2^t\lambda\mathbb{E}\left[\left\|[w_{t+1}^*, \theta_{t+1}^*] - [\bar{w}_t, \bar{\theta}_t]\right\|\right] + B\sum_{t=1}^{T-1} 2^t\lambda\mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w_t^*, \theta_t^*]\right\|\right]\right) \\
&= O\left(B\sum_{t=0}^{T-1} 2^t\lambda\frac{B}{2^t} + B\sum_{r=1}^{T-1} 2^t\lambda\frac{B}{2^t}\right) = O\left(T\lambda B^2\right) = O\left(\frac{\log_2(n)BL}{\sqrt{n'}}\right),
\end{aligned}$$

5. Although stated therein for the distance, the last step of their proof shows a squared distance bound can be obtained.

where step (i) comes from a triangle inequality and step (ii) is obtained from a series of algebraic manipulations which are expanded upon in the full proof. Finally, in the case where $\hat{\alpha} > 0$, extra steps are required to bound the distance of output of $\mathcal{A}_{\mathsf{emp}}$ to the exact saddle point of $F_S^{(t)}(w,\theta) := \frac{1}{n'}\sum_{x\in S_t} f^{(t)}(w,\theta;x)$. This is accomplished using the SC/SC property of $F_S^{(t)}$ and the $\hat{\alpha}$-relative accuracy guarantee of $\mathcal{A}_{\mathsf{emp}}$.

4. Optimal Strong Gap Rate for DP-SSP

With the guarantees of recursive regularization established, what remains is to show there exist (ϵ,δ)-DP algorithms which achieve a sufficient accuracy on the empirical objective. Note this suffices to make the entire recursive regularization algorithm private.

Theorem 8 *Let $\mathcal{A}_{\mathsf{emp}}$ used in Algorithm 1 be (ϵ,δ)-DP. Then Algorithm 1 is (ϵ,δ)-DP.*

This follows simply from post processing the parallel composition theorem for differential privacy, since each run of $\mathcal{A}_{\mathsf{emp}}$ is run on a disjoint partition of the dataset.

4.1. Efficient algorithm for the non-smooth setting

In the non-smooth setting, one can obtain optimal rates on the empirical gap using noisy stochastic gradient descent ascent (noisy SGDA). We give this algorithm in detail in Appendix C.2. More briefly, noisy SGDA starts at $[w_0,\theta_0] \in \mathcal{W}\times\Theta$ and takes parameters $T,\eta > 0$, where T is the number of iterations and η is the learning rate. New iterates are obtained via the update rule $[w_{t+1},\theta_{t+1}] = [w_t,\theta_t] - \frac{\eta}{|M_t|}\sum_{x\in M_t} g(w_t,\theta_t;x) + \xi_t$, where $\xi_0,...,\xi_{T-1}$ are i.i.d. Gaussian noise vectors and M_t is a minibatch sampled uniformly with replacement from S. The algorithm then returns the average iterate, $\frac{1}{T}\sum_{t=0}^{T-1}[w_t,\theta_t]$. Noisy SGDA can be used to obtain the following result.

Lemma 9 *There exists an (ϵ,δ)-DP algorithm which satisfies $\hat{\alpha}$-relative accuracy with $\hat{\alpha} = O\left(\frac{\log(n)L\sqrt{d\log(1/\delta)}}{n\epsilon}\right)$ and runs in $O\left(\min\left\{\frac{n^2\epsilon^{1.5}}{\log^2(n)\sqrt{d\log(1/\delta)}}, \frac{n^{3/2}}{\log^{3/2}(n)}\right\}\right)$ gradient evaluations.*

Applying Theorem 5 then yields a near optimal rate on the strong gap.

Corollary 10 *There exists an Algorithm, $\mathcal{R}$, which is (ϵ,δ)-DP, has gradient evaluations bounded by $O\big(\min\big\{\frac{n^2\epsilon^{1.5}}{\log(n)\sqrt{d\log(1/\delta)}}, \frac{n^{3/2}}{\sqrt{\log(n)}}\big\}\big)$, and satisfies*

$$\mathrm{Gap}(\mathcal{R}) = O\left(\frac{\log^{3/2}(n)BL}{\sqrt{n}} + \frac{\log^2(n)BL\sqrt{d\log(1/\delta)}}{n\epsilon}\right).$$

4.2. Near linear time algorithm for the smooth setting

In the smooth setting, we can achieve the optimal rate in nearly linear time. Our result leverages accelerated algorithms for smooth and strongly convex-strongly concave saddle point problems (Jin et al., 2022; Palaniappan and Bach, 2016).

Lemma 11 *(Jin et al. (2022, Theorem 3, Corollary 41)) Let $f : \mathcal{W} \times \Theta \times \mathcal{X} \mapsto \mathbb{R}$ be β-smooth and $\alpha > 0$. Let both $h_w : \mathcal{W} \mapsto \mathbb{R}$ and $h_\theta : \Theta \mapsto \mathbb{R}$ be $c_1\mu$-strongly convex and $c_2\mu$-smooth functions for some $\mu > 0$ and constants c_1, c_2. Consider the objective $F_h(w, \theta; S) = \sum_{t=1}^{T} f(w, \theta; S) + h_w(w) - h_\theta(\theta)$. Then there exists an algorithm which finds an approximate saddle point of F_h with empirical gap at most α in $O\left(\kappa \log(\kappa) \log(\frac{\kappa BL}{\alpha})\right)$ gradient evaluations, where $\kappa = O(n + \sqrt{n}(1 + \beta/\mu))$.*

Given this, we consider the following implementation of $\mathcal{A}_{\mathsf{emp}}$. Define $[w^*_{S,t}, \theta^*_{S,t}]$ to be the saddle point of $F^{(t)}(w, \theta) = \frac{1}{n}\sum_{x \in S_t} f^{(t)}(w, \theta; x)$ for all $t \in [T]$. At round $t \in [T]$, find a point $[\hat{w}_t, \hat{\theta}_t]$ such that $\mathbb{E}\left[\|[\hat{w}_t, \hat{\theta}_t] - [w^*_{S,t}, \theta^*_{S,t}]\|^2\right] \leq \left(\frac{\delta}{5} \cdot \frac{L}{2^t \lambda n'}\right)^2$. We can find this point efficiently using the algorithm from Jin et al. (2022) referenced above. Then output $[\bar{w}_t, \bar{\theta}_t] = [\hat{w}_t, \hat{\theta}_t] + \xi_t$ where $\xi_t \sim \mathcal{N}(0, \mathbb{I}_d \sigma_t^2)$ and $\sigma_t = \frac{8L\sqrt{\log(2/\delta)}}{2^t \lambda n' \epsilon}$. This implementation gives us the following result.

Theorem 12 *Let $\mathcal{A}_{\mathsf{emp}}$ be as described above. Then Algorithm 1 is (ϵ, δ)-DP and when run with $\lambda = \frac{48}{B}\left(\frac{L}{\sqrt{n'}} + \frac{L\sqrt{d\log(2/\delta)}}{n'\epsilon}\right)$ satisfies*

$$\mathrm{Gap}(\mathcal{R}) = O\left(\frac{\log^{3/2}(n) BL}{\sqrt{n}} + \frac{\log^2(n) BL\sqrt{d\log(1/\delta)}}{n\epsilon}\right),$$

and runs in at most $O\left(\kappa \log(\kappa) \log(\kappa n/\delta) \log(n)\right)$ gradient evaluations with $\kappa = O\left(n + n\beta B/L\right)$.

Proof [proof of Theorem 12] In the following, we start by proving the privacy guarantee. Then, we prove the utility guarantee, and finish by verifying the running time of the algorithm.

Privacy Guarantee: Consider any $t \in [T]$ and fix $[w_1, \theta_1], ..., [w_{t-1}, \theta_{t-1}]$. The stability of the regularized saddle point at round t, $[w^*_{S,t}, \theta^*_{S,t}]$, is then $\frac{L}{2^t \lambda n'}$ by Lemma 2. Since $\mathcal{A}_{\mathsf{emp}}$ guarantees that $\mathbb{E}\left[\|[\hat{w}_t, \hat{\theta}_t] - [w^*_{S,t}, \theta^*_{S,t}]\|\right] \leq \frac{\delta}{5} \cdot \frac{L}{2^t \lambda n'}$, we have by Markov's inequality that with probability at least $1 - \frac{\delta}{2}$ that $\|[\hat{w}_t, \hat{\theta}_t] - [w^*_{S,t}, \theta^*_{S,t}]\| \leq \frac{L}{2^t \lambda n'}$. Thus with probability at least $1 - \frac{\delta}{2}$, generating $[\hat{w}_t, \hat{\theta}_t]$ satisfies $\frac{2L}{2^t \lambda n'}$ uniform argument stability. Thus Gaussian noise of scale $\sigma_t = \frac{8L\sqrt{\log(2/\delta)}}{2^t \lambda n' \epsilon}$ ensures the round is (ϵ, δ)-DP. Parallel composition then ensures the entire algorithm is (ϵ, δ)-DP since each phase acts on a disjoint partition of the dataset.

Utility Guarantee: We now turn to the accuracy guarantee. Specifically, we leverage the generalized convergence guarantee of Algorithm 1 given by Theorem 18 in Appendix B. This theorem guarantees that so long as the distance condition $\mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w^*_{S,t}, \theta^*_{S,t}]\right\|^2\right] \leq \frac{B^2}{12 \cdot 2^{2t}}$ is satisfied for all $t \in [T]$, one obtains convergence guarantee $\mathrm{Gap}(\mathcal{R}) = O(\log(n) B^2 \lambda)$. That is, after the distance guarantee is established, the rest of the analysis (i.e. the proof of Theorem 18) follows the same lines as in the non-smooth case. Note under the setting of λ in Theorem 12 we have

$$\mathrm{Gap}(\mathcal{R}) = O(\log(n) B^2 \lambda) = O\left(\frac{\log^{3/2}(n) BL}{\sqrt{n}} + \frac{\log^2(n) BL\sqrt{d\log(2/\delta)}}{n\epsilon}\right).$$

Thus all that remains is to show that the distance condition, $\mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w^*_{S,t}, \theta^*_{S,t}]\right\|^2\right] \leq \frac{B^2}{12 \cdot 2^{2t}}$, is satisfied for all $t \in [T]$. In this regard we have,

$$\begin{aligned}
\mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w^*_{S,t}, \theta^*_{S,t}]\right\|^2\right] &\leq \mathbb{E}\left[\|[\bar{w}_t, \bar{\theta}_t] - [\hat{w}_t, \hat{\theta}_t]\|^2 + \|[\hat{w}_t, \hat{\theta}_t] - [w^*_{S,t}, \theta^*_{S,t}]\|^2\right] \\
&\leq d\sigma_t^2 + \left(\frac{\delta}{5} \cdot \frac{L}{2^t \lambda n'}\right)^2 \\
&\leq \frac{64 d L^2 \log(2/\delta)}{2^{2t} \lambda^2 (n')^2 \epsilon^2} + \frac{B^2}{25 \cdot 2^{2t}} \leq \frac{B^2}{12 \cdot 2^{2t}}.
\end{aligned}$$

For the first inequality, observe that the noise vector is uncorrelated with the vectors, $[\hat{w}_t, \hat{\theta}_t]$ and $[w^*_{S,t}, \theta^*_{S,t}]$. For the second inequality note $\mathbb{E}\left[\|[\bar{w}_t, \bar{\theta}_t] - [\hat{w}_t, \hat{\theta}_t]\|^2\right] = \mathbb{E}\left[\|\xi_t\|^2\right] = d\sigma_t^2$. Further, $\mathbb{E}\left[\|[\hat{w}_t, \hat{\theta}_t] - [w^*_{S,t}, \theta^*_{S,t}]\|^2\right]$ is bounded due to the chosen implementation of $\mathcal{A}_{\mathsf{emp}}$. The third inequality comes from the settings of σ_t and the fact that $\lambda > \frac{48L}{B\sqrt{n'}}$. The last inequality uses the fact that $\lambda > \frac{48L\sqrt{d\log(2/\delta)}}{Bn'\epsilon}$.

Running Time: One can ensure that overall algorithm runs in nearly linear time by leveraging accelerated methods to find the point $[\hat{w}, \hat{\theta}_t]$. The description of $\mathcal{A}_{\mathsf{emp}}$ requires that at each phase $t \in [T]$, one has $\mathbb{E}\left[\|[\hat{w}_t, \hat{\theta}_t] - [w^*_{S,t}, \theta^*_{S,t}]\|^2\right] \leq \left(\frac{\delta}{5} \cdot \frac{L}{2^t \lambda n'}\right)^2$, which by Lemma 3 is satisfied if the empirical gap is at most $\lambda\left(\frac{\delta}{5} \cdot \frac{L}{2^t \lambda n'}\right)^2 = \frac{\delta^2}{25} \cdot \frac{L^2}{2^{2t}\lambda (n')^2}$. For simplicity, we observe that

$$\frac{\delta^2}{25} \cdot \frac{L^2}{2^{2t}\lambda n'^2} = \Omega\left(\frac{\delta^2 L^2}{2^{2T}\lambda (n')^2}\right) = \Omega\left(\frac{B^2\lambda^2}{L^2}\frac{\delta^2 L^2}{\lambda (n')^2}\right) = \Omega\left(\frac{\delta^2 BL}{n^{2.5}}\right)$$

We now apply Lemma 11 with $h_w(w) = \lambda \sum_{k=0}^{t-1} 2^{k+1} \|w - \bar{w}_k\|^2$, $h_\theta(\theta) = \lambda \sum_{k=0}^{t-1} 2^{k+1} \|w - \bar{w}_k\|^2$, $\mu = 2^t\lambda$ and $\alpha = \frac{c_3 \delta^2 BL}{n^{2.5}}$ for some sufficiently small constant c_3. This gives that the running time of phase t is $O\left(\kappa_t \log(\kappa_t) \log(\kappa_t n^{2.5}/\delta^2]\right)$, where $\kappa_t = O\left(n + \sqrt{n}\beta/[2^t\lambda])\right) = O\left(n + n\beta B/L\right)$. Running this implementation of $\mathcal{A}_{\mathsf{emp}}$ each phase incurs an extra factor of $T = \log(\frac{L}{B\lambda}) = O(\log(n))$, giving the claimed running time bound of $O\left(\kappa \log(\kappa) \log(\kappa n/\delta] \log(n)\right)$, where $\kappa = O\left(n + n\beta B/L\right)$. ■

5. On the Limitations of Previous Approaches

Prior work into DP SSPs has largely focused on the weak gap criteria. In this section, we provide further investigation into both the importance and challenges of bounding the strong gap over the weak gap. We start by considering a natural question. Do there exist cases where the strong and weak gap differ substantially? We answer this question affirmatively in the following.

Proposition 13 *There exists a convex-concave function f with range $[-1, +1]$ and algorithm $\mathcal{A}$ such that* $\mathrm{Gap}(\mathcal{A}) - \mathrm{Gap}_{\mathsf{weak}}(\mathcal{A}) = 2$.

Our construction shows that this result holds even for a simple one dimensional bilinear problem.

Proof Consider the loss function $f(w,\theta;x) = w\theta$, where $w,\theta,x \in [-1,1]$. Let $\mathcal{D}$ be the uniform distribution over $\{\pm 1\}$. For $\{x_1,\dots,x_n\} \sim \mathcal{D}^n$ consider the algorithm $\mathcal{A}$ which outputs $\bar{w}$ as the mode of the first half of the samples in S and similarly $\bar{\theta}$ is set as the mode of the second half of the samples in S [6]. Note $\bar{w}$ and $\bar{\theta}$ are independent and distributed uniformly over $\{\pm 1\}$ (under the randomness from $\mathcal{D}$).

Now, since $\mathcal{A}$ is a deterministic function of the dataset, the randomness in $\bar{w},\bar{\theta}$ comes only from S. Thus for the weak gap we have $\max_{\theta\in[-1,1]}\{\mathop{\mathbb{E}}_{S}[\bar{w}\theta]\} - \min_{w\in[-1,1]}\{\mathop{\mathbb{E}}_{S}[w\bar{\theta}]\}$ which evaluates to $\max_{\theta\in[-1,1]}\{\mathop{\mathbb{E}}_{S}[\bar{w}]\,\theta\} - \min_{w\in[-1,1]}\{w\mathop{\mathbb{E}}_{S}[\bar{\theta}]\} = 0$. However, one can see for the strong gap we have $\mathop{\mathbb{E}}_{S}\left[\max_{\theta\in[-1,1]}\{\bar{w}\theta\} - \min_{w\in[-1,1]}\{w\bar{\theta}\}\right] = \mathop{\mathbb{E}}_{S}\left[|\bar{w}| + |\bar{\theta}|\right] = 2$, where the first equality comes from evaluating $\theta = \mathsf{sgn}(\bar{w})$ and $w = -\mathsf{sgn}(\bar{\theta})$ in the maximization and minimization operators. ■

Observe that the generalization error w.r.t. the strong gap of this algorithm is always 0 because the loss function does not depend on the random sample from $\mathcal{D}$. The discrepancy between the gaps instead comes from the fact that having the expectation w.r.t. S inside the max/min changes the function over which the dual/primal adversary is maximizing/minimizing. Specifically, note here that the weak gap measures the ability of θ to maximize the function $\theta \mapsto \bar{w}\theta$ for $\bar{w} = 0$, but note $\bar{w} = 0$ does not occur for *any* realization of the dataset S.

One might further observe that a key attribute of this construction is the high variance of the parameter vectors. One can show such behavior is in fact necessary to see such a separation; the full proof of the following is statement is given in Appendix D.1.

Proposition 14 *Let $\mathcal{A}$ be an algorithm such that $\mathop{\mathbb{E}}_{\mathcal{A},S}\left[\left\|\mathcal{A}(S) - \mathbb{E}_{\hat{S}\sim\mathcal{D}^n,\mathcal{A}}\mathcal{A}(\hat{S})\right\|^2\right] \le \tau^2$, then if f is L-Lipschitz it holds that $\mathrm{Gap}(\mathcal{A}) - \mathrm{Gap}_{\mathsf{weak}}(\mathcal{A}) \le L\tau$.*

Tradeoff between Accuracy and Stability An additional consequence of Proposition 14 (in conjunction with Lemma 7) is that Δ-uniform argument stability implies $\sqrt{n}\Delta L$ generalization bound w.r.t. the strong gap that does not rely on smoothness (in contrast to the $\sqrt{L\beta\Delta}$ bound of Ozdaglar et al. (2022) which does). We leave determining tight bounds for stability implies generalization on the strong gap as an interesting direction for future work. In this section however, we show that stronger upper bounds are likely necessary to obtain a more direct algorithm for DP-SSPs. In fact, our key result holds even for empirical risk minimization (ERM) problems. That is, for $f : \mathcal{W}\times\mathcal{X} \mapsto \mathbb{R}$ and $S \in \mathcal{X}^n$, consider the problem of minimizing the excess empirical risk $F_S(w) - \min_{w\in\mathcal{W}}\{F_S(w)\}$, where $F_S(w) = \frac{1}{n}\sum_{x\in S} f(w;x)$. We have the following.

Theorem 15 *For any (possibly randomized) algorithm $\mathcal{A} : \mathcal{X}^n \mapsto \mathcal{W}$ which is Δ-uniform argument stable, there exists a 0-smooth L-Lipschitz loss function, $f : \mathcal{W}\times\mathcal{X} \mapsto \mathbb{R}$, and dataset $S \in \mathcal{X}^n$ such that $\mathbb{E}[F_S(\mathcal{A}(S)) - \min_{w\in\mathcal{W}}\{F_S(w)\}] = \Omega\left(\frac{B^2L}{\Delta n}\right)$ provided $\Delta \ge \frac{B}{\sqrt{\min\{n,d\}}}$.*

The proof can be found in Appendix D.2. Lemma 2 shows this bound is tight for both ERM and empirical saddle point problems. Generalization bounds are only useful when it is possible to obtain good empirical performance. Thus, the implication of this bound is that generalization error

6. Without much loss of generality, we assume that n is divisible by 2 but not by 4, so that the mode of each half of the data are well-defined and belong to $\{-1,+1\}$.

which is $O(\Delta)$ is necessary to obtain the optimal $O\left(1/\sqrt{n}\right)$ statistical rate. To elaborate, let $H(\Delta)$ characterize some (potentially suboptimal) generalization bound for Δ stable algorithms and assume $H(\Delta) = \omega(\Delta)$. To then bound the sum of empirical risk and generalization error, Theorem 15 implies $F_S(\mathcal{A}(S)) - F_S(w^*) + H(\Delta) = \Omega\left(\frac{1}{\Delta n} + H(\Delta)\right) = \omega\left(\frac{1}{\Delta n} + \Delta\right)$. Note the RHS is asymptotically larger than $\frac{1}{\sqrt{n}}$ (i.e. not optimal) for any Δ.

Acknowledgments

RB's and MM's research is supported by NSF CAREER Award 2144532 and NSF Award AF-1908281. CG's research was partially supported by INRIA Associate Teams project, FONDECYT 1210362 grant, ANID Anillo ACT210005 grant, and National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. CCS '16, page 308–318, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex sgd. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/996a7fa078cc36c46d02f9af3bef918b-Paper.pdf.

Raman Arora, Raef Bassily, Cristóbal Guzmán, Michael Menart, and Enayat Ullah. Differentially private generalized linear models revisited. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in ℓ_1 geometry. In *International Conference on Machine Learning*, 2021.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS 2014). (arXiv preprint arXiv:1405.7085)*, pages 464–473. 2014.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11279–11288, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3bd8fdb090f1f5eb66a00c84dbc5ad51-Abstract.html.

Raef Bassily, Cristobal Guzman, and Anupama Nandi. Non-euclidean differentially private stochastic convex optimization. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty*

Fourth Conference on Learning Theory, volume 134 of *Proceedings of Machine Learning Research*, pages 474–499. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/bassily21a.html.

Digvijay Boob and Cristóbal Guzmán. Optimal algorithms for differentially private stochastic monotone variational inequalities and saddle-point problems. *Mathematical Programming*, pages 1–43, 2023. doi: 10.1007/s10107-023-01953-5.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms, 2018. URL https://arxiv.org/abs/1804.01619.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438, 2013. doi: 10.1109/FOCS.2013.53.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Farzan Farnia and Asuman E. Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning*, 2020.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 439–449, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450369794. doi: 10.1145/3357713.3384335. URL https://doi.org/10.1145/3357713.3384335.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020b.

Dylan J. Foster, Ayush Sekhari, Ohad Shamir, Nathan Srebro, Karthik Sridharan, and Blake Woodworth. The complexity of making the gradient small in stochastic convex optimization. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1319–1345. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/foster19b.html.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(71):2489–2512, 2014. URL http://jmlr.org/papers/v15/hazan14a.html.

Prateek Jain and Abhradeep Thakurta. (near) dimension independent risk bounds for differentially private learning. In *ICML*, 2014.

Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta. Differentially private online learning. In *25th Annual Conference on Learning Theory (COLT)*, pages 24.1–24.34, 2012.

Yujia Jin, Aaron Sidford, and Kevin Tian. Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4362–4415. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/jin22b.html.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17 – 58, 2011. doi: 10.1214/10-SSY011. URL https://doi.org/10.1214/10-SSY011.

Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth erm and sco in subquadratic steps. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4053–4064. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/211c1e0b83b9c69fa9c4bdede203c1e3-Paper.pdf.

Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In Marina Meila 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, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/lei21b.html.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4615–4625. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/mohri19a.html.

Aryan Mokhtari, Asuman E. Ozdaglar, and Sarath Pattathil. Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020. doi: 10.1137/19M127375X. URL https://doi.org/10.1137/19M127375X.

Arkadi Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004. doi: 10.1137/S1052623403425629. URL https://doi.org/10.1137/S1052623403425629.

Arkadi Nemirovski and D Yudin. On cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. In *Soviet Mathematics. Doklady*, volume 19, pages 258–269, 1978.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and And Shapiro. Robust stochastic approximation approach to stochastic programming. *Society for Industrial and Applied Mathematics*, 19:1574–1609, 01 2009. doi: 10.1137/070704277.

Arkadi Nemirovski, Shmuel Onn, and Uriel G. Rothblum. Accuracy certificates for computational problems with convex structure. *Math. Oper. Res.*, 35(1):52–78, 2010.

Asuman Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. What is a good metric to study generalization of minimax learners? In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022. doi: 10.48550/ARXIV.2206.04502. URL https://arxiv.org/abs/2206.04502.

Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/1aa48fc4880bb0c9b8a3bf979d3b917e-Paper.pdf.

Omar Rivasplata, Emilio Parrado-Hernandez, John S Shawe-Taylor, Shiliang Sun, and Csaba Szepesvari. Pac-bayes bounds for stable algorithms with instance-dependent priors. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/386854131f58a556343e056f03626e00-Paper.pdf.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171 – 176, 1958. doi: pjm/1103040253. URL https://doi.org/.

Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Nearly optimal private lasso. In *NIPS*, 2015.

Robert Williamson and Aditya Menon. Fairness risk measures. In *International Conference on Machine Learning*, pages 6786–6797. PMLR, 2019.

Zhenhuan Yang, Shu Hu, Yunwen Lei, Kush R Vashney, Siwei Lyu, and Yiming Ying. Differentially private sgda for minimax problems. In James Cussens and Kun Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 2192–2202. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/yang22a.html.

Yaodong Yu, Tianyi Lin, Eric Mazumdar, and Michael I. Jordan. Fast distributionally robust learning with variance reduced min-max optimization. *CoRR*, abs/2104.13326, 2021.

Junyu Zhang, Mingyi Hong, Mengdi Wang, and Shuzhong Zhang. Generalization bounds for stochastic saddle point problems. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 568–576. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/zhang21a.html.

Liang Zhang, Kiran Koshy Thekumparampil, Sewoong Oh, and Niao He. Bring your own algorithm for optimal differentially private stochastic minimax optimization. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

Yuchen Zhang and Xiao Lin. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *International Conference on Machine Learning*, pages 353–361. PMLR, 2015.

Appendix A. Supporting Proofs from Preliminaries

A.1. Lipschitzness of the Gap Function

Proof [proof of Fact 1] For any $[\bar{w}, \bar{\theta}], [\bar{w}', \bar{\theta}'] \in \mathcal{W} \times \Theta$ we have

$$\begin{aligned}
\widehat{\mathrm{Gap}}(\bar{w}, \bar{\theta}) - \widehat{\mathrm{Gap}}(\bar{w}', \bar{\theta}') &= \sup_{w,\theta} \left\{ F_{\mathcal{D}}(\bar{w}, \theta) - F_{\mathcal{D}}(w, \bar{\theta}) \right\} - \sup_{w,\theta} \left\{ F_{\mathcal{D}}(\bar{w}', \theta) - F_{\mathcal{D}}(w, \bar{\theta}') \right\} \\
&\leq \sup_{w,\theta} \left\{ F_{\mathcal{D}}(\bar{w}, \theta) - F_{\mathcal{D}}(\bar{w}', \theta) + F_{\mathcal{D}}(w, \bar{\theta}') - F_{\mathcal{D}}(w, \bar{\theta}) \right\} \\
&\leq L \sup_{w,\theta} \left\{ \|\bar{w} - \bar{w}'\| + \|\bar{\theta}' - \bar{\theta}\| \right\} \\
&\leq \sqrt{2} L \|[\bar{w}, \bar{\theta}] - [\bar{w}', \bar{\theta}']\|,
\end{aligned}$$

where we used in the last inequality that $a + b \leq \sqrt{2}\sqrt{a^2 + b^2}$. ■

A.2. Local Privacy

In the case of local differential privacy (LDP), a simple implementation of noisy SGDA (see Appendix C.1) suffices to obtain the optimal rate. We defer the reader to Duchi et al. (2013) for a discussion of LDP and the matching lower bound. Consider the implementation of SGDA which defines the saddle estimator as

$$\nabla_t = g(w_{t-1}, \theta_{t-1}; x_t) + \xi_t$$

where $\xi_t \sim \mathcal{N}(0, \mathbb{I}_d \sigma)$ and $\sigma = \frac{L\sqrt{\log(1/\delta)}}{\epsilon}$ and x_t is sampled without replacement from S. By Lemma 20 we have the following.

Corollary 16 *Let $T = n$. Then the algorithm described above, denoted as $\mathcal{A}$, is (ϵ, δ)-LDP and if $\eta = \frac{B}{\sqrt{nd\log(1/\delta)}L\epsilon}$ the average iterate, $[\bar{w}, \bar{\theta}]$, satisfies* $\mathrm{Gap}(\mathcal{A}) = O\left(\frac{BL\sqrt{d\log(1/\delta)}}{\sqrt{n}\epsilon}\right)$.

Appendix B. Missing Results from Section 3

B.1. Proof of Lemma 6

The first inequality follows from an application of Jensen's inequality.

$$\begin{aligned}
&\widehat{\mathrm{Gap}}\left(\mathop{\mathbb{E}}_{\mathcal{A},S}\left[\mathcal{A}_w(S)\right], \mathop{\mathbb{E}}_{\mathcal{A},S}\left[\mathcal{A}_\theta(S)\right] \right) \\
&= \max_{\theta\in\Theta} \left\{ F_{\mathcal{D}}\Big(\mathop{\mathbb{E}}_{\hat{S}\sim\mathcal{D}^n, \mathcal{A}_w} \left[\mathcal{A}_w(\hat{S})\right], \theta \Big) \right\} - \min_{w\in\mathcal{W}} \left\{ F_{\mathcal{D}}\Big(w, \mathop{\mathbb{E}}_{\hat{S}\sim\mathcal{D}^n, \mathcal{A}_\theta} \left[\mathcal{A}_\theta(\hat{S})\right] \Big) \right\} \\
&\leq \max_{\theta\in\Theta} \left\{ \mathop{\mathbb{E}}_{\hat{S}\sim\mathcal{D}^n, \mathcal{A}_w} \left[F_{\mathcal{D}}(\mathcal{A}_w(\hat{S}), \theta) \right] \right\} - \min_{w\in\mathcal{W}} \left\{ \mathop{\mathbb{E}}_{\hat{S}\sim\mathcal{D}^n, \mathcal{A}_\theta} \left[F_{\mathcal{D}}(w, \mathcal{A}_\theta(\hat{S})) \right] \right\} \\
&= \mathrm{Gap}_{\mathsf{weak}}(\mathcal{A}).
\end{aligned}$$

The second inequality in the theorem statement then follows from stability implies generalization result for the weak gap, for which we provide a restatement below.

Lemma 17 *(Lei et al., 2021, Theorem 1), (Boob and Guzmán, 2023, Proposition 2.1) Let the loss function f be L-Lipschitz and the algorithm $\mathcal{A}$ be Δ-uniform argument stable. Then* $\mathrm{Gap}_{\mathsf{weak}}(\mathcal{A}) \leq \mathop{\mathbb{E}}_{S}[\mathrm{Gap}_S(\mathcal{A})] + \Delta L$.

B.2. Convergence of Recursive Regularization

In this section we prove the following more general statement of Theorem 5, which will be useful later.

Theorem 18 *Let $\lambda \geq \frac{48L}{B\sqrt{n'}}$ and $\mathcal{A}_{\mathsf{emp}}$ be such that for all $t \in [T]$ it holds that $\mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w^*_{S,t}, \theta^*_{S,t}]\right\|^2\right] \leq \frac{B^2}{12\cdot 2^{2t}}$. Then Recursive Regularization satisfies*

$$\mathrm{Gap}(\mathcal{R}) = O\Big(\log(n)B^2\lambda\Big)$$

To prove this result, it will be helpful to first show several intermediate results. We start by defining several useful quantities. Define $\{\mathcal{F}_t\}_{t=0}^T$ as the filtration where $\mathcal{F}_t$ is the sigma algebra induced by all randomness up to $[\bar{w}_t, \bar{\theta}_t]$. For every $t \in [T]$ we define

- $[w^*_t, \theta^*_t]$: saddle point of $F^{(t)}_{\mathcal{D}}(w,\theta) := \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[f^{(t)}(w,\theta;x)\right]$;
- $[w^*_{S,t}, \theta^*_{S,t}]$: saddle point of $F^{(t)}_{S}(w,\theta) := \frac{1}{n}\sum_{x\in S} f^{(t)}(w,\theta;x)$;
- $[\widetilde{w}_t, \widetilde{\theta}_t] := \mathbb{E}\left[[w^*_{S,t}, \theta^*_{S,t}]\middle|\mathcal{F}_{t-1}\right]$;
- $\widehat{\mathrm{Gap}}^{(t)}(\bar{w},\bar{\theta}) := \max_{\theta\in\Theta}\left\{F^{(t)}_{\mathcal{D}}(\bar{w},\theta)\right\} - \min_{w\in\mathcal{W}}\left\{F^{(t)}_{\mathcal{D}}(w,\bar{\theta})\right\}$: the gap function w.r.t. $F^{(t)}_{\mathcal{D}}$; and,
- $\widehat{\mathrm{Gap}}^{(t)}_S(\bar{w},\bar{\theta}) := \max_{\theta\in\Theta}\left\{F^{(t)}_{S_t}(\bar{w},\theta)\right\} - \min_{w\in\mathcal{W}}\left\{F^{(t)}_{S_t}(w,\bar{\theta})\right\}$: the empirical gap function.

We now establish two distance inequalities which will be used when analyzing the final gap bound in Theorem 18. The first inequality above bounds the distance of the output of the t-th round to the minimizer of $F^{(t)}_{\mathcal{D}}$. The second inequality bounds the distance of the minimizer of $F^{(t)}_{\mathcal{D}}$ to the most recent regularization point.

Lemma 19 *Assume the conditions of Theorem 18 hold. Then for every $t \in [T]$, the following holds*

P.1 $\mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w^*_t, \theta^*_t]\right\|\right]^2 \leq \mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w^*_t, \theta^*_t]\right\|^2\right] \leq \frac{B^2}{2^{2t}}$*; and,*

P.2 $B_t^2 := \mathbb{E}\left[\left\|[w^*_t, \theta^*_t] - [\bar{w}_{t-1}, \bar{\theta}_{t-1}]\right\|\right]^2 \leq \mathbb{E}\left[\left\|[w^*_t, \theta^*_t] - [\bar{w}_{t-1}, \bar{\theta}_{t-1}]\right\|^2\right] \leq \frac{B^2}{2^{2(t-1)}}$.

Proof We will prove both properties via induction on $B_1, ..., B_T$. Specifically, for each $t \in [T]$ we will introduce three terms E_t, F_t, G_t, and show that these terms are bounded if the bound on B_t holds and that B_t holds if $E_{t-1}, F_{t-1}, G_{t-1}$ are bounded. Property **P.1** is then established as a result of the fact that $\mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w^*_t, \theta^*_t]\right\|^2\right] \leq 3(E_t + F_t + G_t)$. Note that B_1 holds as the base case because $\mathbb{E}\left[\left\|[w^*_1, \theta^*_1] - [\bar{w}_0, \bar{\theta}_0]\right\|^2\right] \leq B^2$.

Property P.1: We here prove that if B_t is sufficiently bounded, then E_t, F_t, G_t are bounded where for $t \in [T]$ we define

$$E_t = \mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w^*_{S,t}, \theta^*_{S,t}]\right\|^2\right], \quad F_t = \mathbb{E}\left[\left\|[w^*_{S,t}, \theta^*_{S,t}] - [\widetilde{w}_t, \widetilde{\theta}_t]\right\|^2\right], \quad G_t = \frac{1}{2^t\lambda}\mathbb{E}\left[\widehat{\mathrm{Gap}}^{(t)}\left(\widetilde{w}_t, \widetilde{\theta}_t\right)\right]. \tag{5}$$

Additionally, this will establish property **P.1** because for any $t \in [T]$ it holds that,

$$\begin{aligned}
&\mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w^*_t, \theta^*_t]\right\|^2\right] \\
&\leq 3\left(\mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w^*_{S,t}, \theta^*_{S,t}]\right\|^2\right] + \mathbb{E}\left[\left\|[w^*_{S,t}, \theta^*_{S,t}] - [\widetilde{w}_t, \widetilde{\theta}_t]\right\|^2\right] + \mathbb{E}\left[\left\|[\widetilde{w}_t, \widetilde{\theta}_t] - [w^*_t, \theta^*_t]\right\|^2\right]\right) \\
&\leq 3\Bigg(\underbrace{\mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w^*_{S,t}, \theta^*_{S,t}]\right\|^2\right]}_{E_t} + \underbrace{\mathbb{E}\left[\left\|[w^*_{S,t}, \theta^*_{S,t}] - [\widetilde{w}_t, \widetilde{\theta}_t]\right\|^2\right]}_{F_t} + \underbrace{\frac{1}{2^t\lambda}\mathbb{E}\left[\widehat{\mathrm{Gap}}^{(t)}\left(\widetilde{w}_t, \widetilde{\theta}_t\right)\right]}_{G_t}\Bigg).
\end{aligned} \tag{6}$$

The second inequality comes from the strong convexity-strong concavity of the loss.

Bounding E_t: We have that E_t is bounded by the assumption made in the statement of Theorem 18.

Bounding F_t:

$$\mathbb{E}\left[\left\|[w^*_{S,t}, \theta^*_{S,t}] - [\widetilde{w}_t, \widetilde{\theta}_t]\right\|^2\right] \leq \frac{L^2}{2^{2t}\lambda^2 n'} \leq \frac{B^2L^2}{2304 \cdot 2^{2t}(L/\sqrt{n'})^2 n'} = \frac{B^2}{2304 \cdot 2^{2t}}. \tag{7}$$

The first inequality comes from the stability of the regularized minimizer and Lemma 7. The second inequality comes from the setting of $\lambda \geq \frac{48L}{B\sqrt{n'}}$.

Bounding G_t: We have

$$\begin{aligned}
\frac{1}{2^t\lambda}\,\mathbb{E}\left[\widehat{\mathrm{Gap}}^{(t)}\left(\widetilde{w}_t, \widetilde{\theta}_t\right)\right] &= \frac{1}{2^t\lambda}\,\mathbb{E}\left[\mathbb{E}\left[\widehat{\mathrm{Gap}}^{(t)}\left(\mathbb{E}\left[w^*_{S,t}|\mathcal{F}_{t-1}\right], \mathbb{E}\left[\theta^*_{S,t}|\mathcal{F}_{t-1}\right]\right)\Big|\mathcal{F}_{t-1}\right]\right] \\
&\leq \frac{1}{2^t\lambda}\left(\mathbb{E}\left[\mathbb{E}\left[\widehat{\mathrm{Gap}}^{(t)}_S\left(w^*_{S,t}, \theta^*_{S,t}\right)\Big|\mathcal{F}_{t-1}\right]\right] + \frac{L^2}{2^t\lambda n'}\right) \\
&= \frac{L^2}{2^{2t}\lambda^2 n'} \leq \frac{B^2}{2304 \cdot 2^{2t}}.
\end{aligned}$$

The first equality comes from the definition of $[\widetilde{w}_t, \widetilde{\theta}_t]$. The first inequality comes from Lemma 6, where we consider the algorithm stated in the lemma to be the algorithm which outputs the *exact* regularized minimizer. Note this algorithm is $\frac{L^2}{2^t\lambda n'}$ stable. The second equality comes from the fact that $[w^*_{S,t}, \theta^*_{S,t}]$ is the exact empirical saddle point. The final inequality uses the same analysis as in Eqn. (7).

We thus have a final bound $3(E_t + F_t + G_t) \leq \frac{B^2}{2^{2t}}$.

Property P.2: Now assume B_{t-1} holds. We have

$$\begin{aligned}
\mathbb{E}\left[\left\|[w^*_t, \theta^*_t] - [\bar{w}_{t-1}, \bar{\theta}_{t-1}]\right\|^2\right] &\leq 2\mathbb{E}\left[\left\|[w^*_t, \theta^*_t] - [\widetilde{w}_{t-1}, \widetilde{\theta}_{t-1}]\right\|^2\right] + 2\mathbb{E}\left[\left\|[\widetilde{w}_{t-1}, \widetilde{w}_{t-1}] - [\bar{w}_{t-1}, \bar{\theta}_{t-1}]\right\|\right]^2 \\
&\leq 2\mathbb{E}\left[\left\|[w^*_t, \theta^*_t] - [\widetilde{w}_{t-1}, \widetilde{\theta}_{t-1}]\right\|^2\right] + 4E_{t-1} + 4F_{t-1}.
\end{aligned} \tag{8}$$

Above E_{t-1} and F_{t-1} are as defined in (5). We bound the remaining squared distance term in the following. First, note that the primal function $F^{(t)}(\cdot, \theta_t^*)$ is strongly convex and $\forall w \in \mathcal{W}$ it holds that $\left\langle \nabla_w F_{\mathcal{D}}^{(t)}(w_t^*, \theta_t^*), w_t^* - w \right\rangle \leq 0$. Similar facts hold for $-F^{(t)}(w_t^*, \cdot)$. Thus we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|[w_t^*, \theta_t^*] - [\widetilde{w}_{t-1}, \widetilde{\theta}_{t-1}]\right\|^2\right] = \mathbb{E}\left[\|\widetilde{w}_{t-1} - w_t^*\|^2 + \|\theta_t^* - \widetilde{\theta}_{t-1}\|^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{2^t\lambda}\left(F_{\mathcal{D}}^{(t)}(\widetilde{w}_{t-1}, \theta_t^*) - F_{\mathcal{D}}^{(t)}(w_t^*, \theta_t^*) + F_{\mathcal{D}}^{(t)}(w_t^*, \theta_t^*) - F_{\mathcal{D}}^{(t)}(w_t^*, \widetilde{\theta}_{t-1})\right)\right] \\
&= \mathbb{E}\Big[\frac{1}{2^t\lambda}\left(F_{\mathcal{D}}^{(t-1)}(\widetilde{w}_{t-1}, \theta_t^*) - F_{\mathcal{D}}^{(t-1)}(w_t^*, \widetilde{\theta}_{t-1})\right) + \|\widetilde{w}_{t-1} - \bar{w}_{t-1}\|^2 - \left\|\theta_t^* - \bar{\theta}_{t-1}\right\|^2 \\
&\quad - \|w_t^* - \bar{w}_{t-1}\|^2 + \|\widetilde{\theta}_{t-1} - \bar{\theta}_{t-1}\|^2\Big] \\
&\leq \mathbb{E}\left[\frac{1}{2^t\lambda}\left(F_{\mathcal{D}}^{(t-1)}(\widetilde{w}_{t-1}, \theta_t^*) - F_{\mathcal{D}}^{(t-1)}(w_t^*, \widetilde{\theta}_{t-1})\right) + \left\|[\widetilde{w}_{t-1}, \widetilde{\theta}_{t-1}] - [\bar{w}_{t-1}, \bar{\theta}_{t-1}]\right\|^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{2^t\lambda}\left(F_{\mathcal{D}}^{(t-1)}(\widetilde{w}_{t-1}, \theta_t^*) - F_{\mathcal{D}}^{(t-1)}(w_t^*, \widetilde{\theta}_{t-1})\right)\right] + 2E_{t-1} + 2F_{t-1} \\
&\leq \mathbb{E}\left[\frac{1}{2\cdot 2^{t-1}\lambda}\left(\widehat{\mathrm{Gap}}^{(t-1)}(\widetilde{w}_{t-1}, \widetilde{\theta}_{t-1})\right)\right] + 2E_{t-1} + 2F_{t-1} \\
&\leq \frac{1}{2}G_{t-1} + 2E_{t-1} + 2F_{t-1}.
\end{aligned}$$

The second inequality comes from removing the negative norm terms. The third inequality comes from the definition of E_{t-1} and F_{t-1}. The second to last inequality comes from the definition of G_{t-1}, as given in Eqn. (5). Plugging this result into (8) and using the previously established bounds on $E_{t-1}, F_{t-1}, G_{t-1}$ (which hold under the assumed bound on B_{t-1}) we have

$$\mathbb{E}\left[\left\|[w_t^*, \theta_t^*] - [\bar{w}_{t-1}, \bar{\theta}_{t-1}]\right\|^2\right] \leq \frac{1}{2}G_{t-1} + 6E_{t-1} + 6F_{t-1} \leq \frac{B^2}{2^{2(t-1)}}.$$

■

We now turn to analyzing the utility of the algorithm to complete the proof.

Proof [proof of Theorem 18] Using the fact that $\widehat{\mathrm{Gap}}$ is $\sqrt{2}L$-Lipschitz and property **P.1**, we have

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mathrm{Gap}}(\bar{w}_T, \bar{\theta}_T) - \widehat{\mathrm{Gap}}(w_T^*, \theta_T^*)\right] &\leq \sqrt{2}L\mathbb{E}\left[\left\|[\bar{w}_T, \bar{\theta}_T] - [w_T^*, \theta_T^*]\right\|\right] \\
&\leq \frac{\sqrt{2}BL}{2^T} \leq \sqrt{2}B^2\lambda. \qquad (9)
\end{aligned}$$

What remains is showing $\mathbb{E}\left[\widehat{\mathrm{Gap}}(w_T^*, \theta_T^*)\right]$ is $\tilde{O}(B\hat{\alpha} + \frac{BL}{\sqrt{n'}})$. Let $w' = \arg\min_{\theta\in\Theta} F_{\mathcal{D}}(w, \theta_T^*)$ and $\theta' = \arg\max_{w\in\mathcal{W}} F_{\mathcal{D}}(w_T^*, \theta)$. Using the fact that $F_{\mathcal{D}}$ is convex-concave we have

$$\widehat{\mathrm{Gap}}(w_T^*, \theta_T^*) = F_{\mathcal{D}}(w_T^*, \theta') - F_{\mathcal{D}}(w', \theta_T^*) \leq \left\langle G_{\mathcal{D}}(w_T^*, \theta_T^*), [w_T^*, \theta_T^*] - [w', \theta'] \right\rangle \qquad (10)$$

where $G_{\mathcal{D}}$ is the population loss saddle operator. Further by the definition of $F^{(T)}$ and denoting $G_{\mathcal{D}}^{(T)}$ as the saddle operator for $F_{\mathcal{D}}^{(T)}$ we have

$$G_{\mathcal{D}}(w_T^*, \theta_T^*) = G_{\mathcal{D}}^{(T)}(w_T^*, \theta_T^*) - 2\lambda \sum_{t=0}^{T-1} 2^{t+1}([w_T^*, -\theta_T^*] - [\bar{w}_t, -\bar{\theta}_t])$$

Thus plugging the above into Eqn. (10) we have

$$\begin{aligned}
\widehat{\mathrm{Gap}}(w_T^*, \theta_T^*) &\leq \left\langle G_{\mathcal{D}}^{(T)}(w_T^*, \theta_T^*), [w_T^*, \theta_T^*] - [w', \theta'] \right\rangle \\
&\quad - \left\langle 2\lambda \sum_{t=0}^{T-1} 2^{t+1}([w_T^*, -\theta_T^*] - [\bar{w}_t, -\bar{\theta}_t]), [w_T^*, \theta_T^*] - [w', \theta'] \right\rangle \\
&\leq - \left\langle 2\lambda \sum_{t=0}^{T-1} 2^{t+1}([w_T^*, -\theta_T^*] - [\bar{w}_t, -\bar{\theta}_t]), [w_T^*, \theta_T^*] - [w', \theta'] \right\rangle \\
&\leq 2B\lambda \sum_{t=0}^{T-1} 2^{t+1} \left\| [w_T^*, -\theta_T^*] - [\bar{w}_t, -\bar{\theta}_t] \right\| \\
&= 2B\lambda \sum_{t=0}^{T-1} 2^{t+1} \left\| [w_T^*, \theta_T^*] - [\bar{w}_t, \bar{\theta}_t] \right\| .
\end{aligned}$$

Above, the second inequality comes from the first order optimally conditions for $[w_T^*, \theta_T^*]$, the third from Cauchy Schwartz and a triangle inequality. The final equality uses the definition of the Euclidean norm and the fact that for any $a, b \in \mathbb{R}$, $(-a - (-b))^2 = (a - b)^2$.

Taking the expectation on both sides of the above we have the following derivation,

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{\mathrm{Gap}}(w_T^*, \theta_T^*)\right] \leq 2B\mathbb{E}\left[\lambda \sum_{t=0}^{T-1} 2^{t+1} \left\|[w_T^*, \theta_T^*] - [\bar{w}_t, \bar{\theta}_t]\right\|\right] \\
&\overset{(i)}{\leq} 4B\mathbb{E}\left[\lambda \sum_{t=0}^{T-1} 2^t \left(\left\|[w_{t+1}^*, \theta_{t+1}^*] - [\bar{w}_t, \bar{\theta}_t]\right\| + \sum_{r=t+1}^{T-1} \left\|[w_{r+1}^*, \theta_{r+1}^*] - [w_r^*, \theta_r^*]\right\|\right)\right] \\
&\leq 4B\mathbb{E}\left[\lambda \sum_{t=0}^{T-1} 2^t \left(\left\|[w_{t+1}^*, \theta_{t+1}^*] - [\bar{w}_t, \bar{\theta}_t]\right\| + \sum_{r=t+1}^{T-1} \left\|[w_{r+1}^*, \theta_{r+1}^*] - [\bar{w}_r, \bar{\theta}_r]\right\| + \left\|[\bar{w}_r, \bar{\theta}_r] - [w_r^*, \theta_r^*]\right\|\right)\right] \\
&= 4B\mathbb{E}\left[\lambda \sum_{t=0}^{T-1} 2^t \left\|[w_{t+1}^*, \theta_{t+1}^*] - [\bar{w}_t, \bar{\theta}_t]\right\| + \lambda \sum_{t=0}^{T-1} 2^t \sum_{r=t+1}^{T-1} \left(\left\|[w_{r+1}^*, \theta_{r+1}^*] - [\bar{w}_r, \bar{\theta}_r]\right\| + \left\|[\bar{w}_r, \bar{\theta}_r] - [w_r^*, \theta_r^*]\right\|\right)\right] \\
&\overset{(ii)}{=} 4B\mathbb{E}\left[\lambda \sum_{t=0}^{T-1} 2^t \left\|[w_{t+1}^*, \theta_{t+1}^*] - [\bar{w}_t, \bar{\theta}_t]\right\| + \lambda \sum_{r=1}^{T-1} \sum_{t=0}^{r-1} 2^t \left(\left\|[w_{r+1}^*, \theta_{r+1}^*] - [\bar{w}_r, \bar{\theta}_r]\right\| + \left\|[\bar{w}_r, \bar{\theta}_r] - [w_r^*, \theta_r^*]\right\|\right)\right] \\
&= 4B\mathbb{E}\left[\lambda \sum_{t=0}^{T-1} 2^t \left\|[w_{t+1}^*, \theta_{t+1}^*] - [\bar{w}_t, \bar{\theta}_t]\right\| + \lambda \sum_{r=1}^{T-1} \left(\left\|[w_{r+1}^*, \theta_{r+1}^*] - [\bar{w}_r, \bar{\theta}_r]\right\| + \left\|[\bar{w}_r, \bar{\theta}_r] - [w_r^*, \theta_r^*]\right\|\right) \sum_{t=0}^{r-1} 2^t\right] \\
&\overset{(iii)}{\leq} 4B\left(\lambda \sum_{t=0}^{T-1} 2^t \left(\frac{B}{2^t}\right) + \lambda \sum_{r=1}^{T-1} \left(\frac{2B}{2^r}\right) \sum_{t=0}^{r-1} 2^t\right) \\
&\leq 4B\left(\lambda \sum_{t=0}^{T-1} 2^t \left(\frac{B}{2^t}\right) + \lambda \sum_{r=1}^{T-1} \left(\frac{B}{2^{r-1}}\right) 2 \cdot 2^{r-1}\right) \\
&= 4\lambda \sum_{t=0}^{T-1} B^2 + 8\lambda \sum_{r=1}^{T-1} B^2 \\
&\leq 12T\lambda B^2 \qquad (11)
\end{aligned}
$$

Above, (i) and the following inequality both come from the triangle inequality. Equality (ii) is obtained by rearranging the sums. Inequality (iii) comes from applying properties **P.1** and **P.2** proved above. The last equality comes from the setting of λ and T.

Now using this result in conjunction with Eqn. (9) we have

$$\mathrm{Gap}(\mathcal{R}) = \sqrt{2}\lambda B^2 + 12T\lambda B^2 = O\left(\log(n) B^2 \lambda\right).$$

Above we use the fact that $T = \log(\frac{L}{B\lambda})$ and $\lambda \geq \frac{L}{B\sqrt{n'}}$, and thus $T = O(\log(n))$. ■

Finally, we prove Theorem 5 leveraging the relative accuracy assumption.

Proof [Proof of Theorem 5] First, observe that under the setting of $\lambda = \frac{48}{B}\left(\hat{\alpha} + \frac{L}{\sqrt{n'}}\right)$ used in the theorem statement that $\log(n)B^2\lambda = O\left(\log(n)B\hat{\alpha} + \frac{\log^{3/2}(n)BL}{\sqrt{n}}\right)$. Thus what remains is to show that the distance condition required by Theorem 18 holds. That is, we now show that if $\mathcal{A}_{\mathsf{emp}}$ satisfies $\hat{\alpha}$-relative accuracy, then for all $t \in [T]$ it holds that $\mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w_{S,t}^*, \theta_{S,t}^*]\right\|^2\right] \leq \frac{B^2}{12 \cdot 2^{2t}}$.

To prove this property, we must leverage the induction argument made by Lemma 19. Specifically, to prove the condition holds for some $t \in [T]$, assume $B_t^2 = \mathbb{E}\left[\left\|[w_t^*, \theta_t^*] - [\bar{w}_{t-1}, \bar{\theta}_{t-1}]\right\|\right]^2 \leq$

$\frac{B^2}{2^{2(t-1)}}$ (recall the base case for $t = 1$ trivially holds). As shown in the proof of Lemma 19, this implies that the quantities F_t, G_t (as defined in 5) are bounded by $\frac{B^2}{2304 \cdot 2^{2t}}$. We thus have

$$\begin{aligned}
\mathbb{E}\left[\left\|[\bar{w}_t, \bar{\theta}_t] - [w^*_{S,t}, \theta^*_{S,t}]\right\|^2\right] &\overset{(i)}{\leq} \frac{\mathbb{E}\left[F_S^{(t)}(\bar{w}_t, \theta^*_{S,t}) - F_S^{(t)}(w^*_{S,t}, \bar{\theta}_t)\right]}{2^t\lambda} \\
&\overset{(ii)}{\leq} \frac{\hat{\alpha}\mathbb{E}\left[\|[w^*_{S,t}, \theta^*_{S,t}] - [\bar{w}_{t-1}, \bar{\theta}_{t-1}]\|\right]}{2^t\lambda} \\
&\leq \frac{\hat{\alpha}\mathbb{E}\left[\|[w^*_{S,t}, \theta^*_{S,t}] - [w^*_t, \theta^*_t]\| + \|[w^*_t, \theta^*_t] - [\bar{w}_{t-1}, \bar{\theta}_{t-1}]\|\right]}{2^t\lambda} \\
&\overset{(iii)}{\leq} \frac{(\sqrt{F_t} + \sqrt{G_t} + B_t)\hat{\alpha}}{2^t\lambda} \overset{(iv)}{\leq} \frac{2B\hat{\alpha}}{2^t 2^{t-1}\lambda} \leq \frac{B^2}{12 \cdot 2^{2t}},
\end{aligned} \tag{12}$$

where B_t is as defined in property **P.2**. Inequality (i) comes from Lemma 3. Inequality (ii) comes from the $\hat{\alpha}$-relative accuracy assumption on $\mathcal{A}_{\mathsf{emp}}$, and the fact that each $f^{(t)}$ is $2L$-Lipschitz. That is, observe

$$\max_{w,\theta \in \mathcal{W}\times\Theta} \left\|\nabla f^{(t)}(w, \theta, x)\right\| \leq L + 2\sum_{k=0}^{t-1} B2^{k+1}\lambda \leq L + 4B2^T\lambda \leq 5L$$

Inequality (iii) comes from a triangle inequality and the definition of F_t, G_t and B_t. Inequality (iv) comes from the induction hypothesis (specifically property **P.2**) and the bounds on F_t and G_t established above. The last inequality in Eqn. (12) comes from the setting $\lambda \geq 48\hat{\alpha}/B$. ■

Appendix C. Missing Results from Section 4

C.1. Stochastic Gradient Descent Ascent (SGDA)

Let $F : \mathcal{W} \times \Theta \mapsto \mathbb{R}$ have saddle operator $G : \mathcal{W} \times \Theta \mapsto \mathbb{R}^d$ and associated strong gap Gap^F. We define the SGDA algorithm in the following manner. Let $T, \eta \geq 0$. Let $[w_0, \theta_0]$ be any vector in $\mathcal{W}\times\Theta$. SGDA uses the following update rule. For $t \in [T-1]$ let ∇_t be a random vector (which may depend on $\nabla_1, ..., \nabla_{t-1}$ and $[w_0, \theta_0], ..., [w_{t-1}, \theta_{t-1}]$) that is a unbiased estimate of $G(w_{t-1}, \theta_{t-1})$ conditional on $[w_{t-1}, \theta_{t-1}]$ and has bounded variance. We define

$$[w_t, \theta_t] = \Pi_{\mathcal{W}\times\Theta}\left([w_{t-1}, \theta_{t-1}] - \eta\nabla_t\right), \quad t \in [T-1] \tag{13}$$

where $\Pi_{\mathcal{W}\times\Theta}$ is the orthogonal projection onto $\mathcal{W} \times \Theta$. The output of SGDA is defined to be

$$[\bar{w}, \bar{\theta}] = \frac{1}{T}\sum_{t=0}^{T-1}[w_t, \theta_t]. \tag{14}$$

We have the following result for the convergence of SGDA.

Lemma 20 *Assume* $\forall t \in [T-1]$ *that* $\mathbb{E}[\nabla_t] = G(w_t, \theta_t)$ *and* $\mathbb{E}\left[\|\nabla_t - G(w_t, \theta_t)\|^2\right] \leq \tau^2$*, then the algorithm,* $\mathcal{A}$*, that is SGDA run with parameters* $T, \eta > 0$ *satisfies for any* $w \in \mathcal{W}$ *and* $\theta \in \Theta$*,*

$$\mathbb{E}\left[F(\bar{w}, \theta) - F(w, \bar{\theta})\right] \leq \frac{\|[w_0, \theta_0] - [w, \theta]\|^2}{2\eta T} + \frac{\eta}{2}\left(L^2 + \tau^2\right)$$

This result is somewhat implicit in Yang et al. (2022, Lemma 3), but for completeness we provide a short proof here.

Proof By the convexity-concavity of F we have for any $[w, \theta] \in \mathcal{W} \times \Theta$ that

$$F(w_t, \theta) - F(w, \theta_t) \leq \langle G(w_t, \theta_t), [w_t, \theta_t] - [w, \theta] \rangle$$

and thus taking the expectation (conditional on $[w_t, \theta_t]$) and using the fact that each ∇_t is unbiased we have

$$\mathbb{E}\left[F(w_t, \theta) - F(w, \theta_t)\right] \leq \left\langle \mathbb{E}\left[\nabla_t\right], [w_t, \theta_t] - [w, \theta] \right\rangle.$$

Using $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$ and the fact that the projection is nonexpansive, we have

$$\begin{aligned}
&\mathbb{E}\left[F(w_t, \theta) - F(w, \theta_t)\right] \\
&\leq \mathbb{E}\left[\frac{1}{2\eta}\left(\|[w_t, \theta_t] - [w, \theta]\|^2 - \|[w_{t+1}, \theta_{t+1}] - [w, \theta]\|^2\right) + \frac{\eta}{2}\|\nabla_t\|^2\right] \\
&= \mathbb{E}\left[\frac{1}{2\eta}\left(\|[w_t, \theta_t] - [w, \theta]\|^2 - \|[w_{t+1}, \theta_{t+1}] - [w, \theta]\|^2\right) + \frac{\eta}{2}\left(\|G(w_t, \theta_t)\|^2 + \|G(w_t, \theta_t) - \nabla_t\|^2\right)\right] \\
&\leq \mathbb{E}\left[\frac{1}{2\eta}\left(\|[w_t, \theta_t] - [w, \theta]\|^2 - \|[w_{t+1}, \theta_{t+1}] - [w, \theta]\|^2\right)\right] + \frac{\eta}{2}\left(L^2 + \tau^2\right),
\end{aligned}$$

where in the first equality we use that $\mathbb{E}[\langle G(w_t, \theta_t), G(w_t, \theta_t) - \nabla_t\rangle] = 0$, due to the unbiasedness of the stochastic oracle.

Summing over all T iterations and taking the average we obtain for the average iterate, $\bar{w}, \bar{\theta}$, and any $[w, \theta] \in \mathcal{W} \times \Theta$ that

$$\begin{aligned}
\mathbb{E}\left[F\Big(\frac{1}{T}\sum_{t=0}^{T-1} w_t, \theta\Big) - F\Big(w, \frac{1}{T}\sum_{s=1}^{T} \theta_t\Big)\right] &\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}[F(w_t, \theta) - F(w, \theta_t)]\right] \\
&\leq \frac{\|[w_0, \theta_0] - [w, \theta]\|^2}{2\eta T} + \frac{\eta}{2}\left(L^2 + \tau^2\right)
\end{aligned}$$

■

C.2. Private algorithm for the empirical gap (Noisy SGDA)

We here provide an implementation of SGDA (see Appendix C.1 above) which is differentially private and yields convergence guarantees for the empirical gap. Let $M_1, ..., M_T$ each be a batch of $m = \max\left\{n\sqrt{\frac{\epsilon}{4T}}, 1\right\}$ samples, each sampled uniformly with replacement from S. Let $\sigma^2 =$

$\frac{c_0 TL^2 \log(1/\delta)}{n^2\epsilon^2}$ for some universal constant c_0 and $\xi_1, \ldots, \xi_T$ each be sampled i.i.d. from $\mathcal{N}(0, \mathbb{I}_d\sigma^2)$. We define

$$\nabla_t = \frac{1}{m}\sum_{x\in M_t} g(w_{t-1}, \theta_{t-1}; x) + \xi_t.$$

Notice that ∇_t as defined above satisfies the assumptions for Lemma 20 with respect to the empirical saddle operator, G_S, for some finite τ.

We have the following result for SGDA run with this stochastic oracle.

Theorem 21 *Let $[w, \theta] \in \mathcal{W} \times \Theta$ such that $\mathbb{E}\left[\|[w_0, \theta_0] - [w, \theta]\|\right] \leq \hat{D}$. Let $\mathcal{A}$ be the algorithm SGDA run with $\nabla_1, \ldots, \nabla_T$ as described above, $T = \min\left\{\frac{n}{8}, \frac{n^2\epsilon^2}{32d\log(1/\delta)}\right\}$, and $\eta = \frac{\hat{D}}{L\sqrt{T}}$. Algorithm $\mathcal{A}$ is (ϵ, δ)-DP, has gradient complexity $O\left(\min\left\{\frac{n^2\epsilon^{1.5}}{\sqrt{d\log(1/\delta)}}, n^{3/2}\right\}\right)$, and satisfies*

$$\mathbb{E}\left[F_S(\bar{w}, \theta) - F_S(w, \bar{\theta})\right] = O\left(\frac{\hat{D}L\sqrt{d\log(1/\delta)}}{n\epsilon} + \frac{\hat{D}L}{\sqrt{n}}\right).$$

The proof of the utility guarantee follows directly from applying Lemma 20 with $\tau = O(L + \sqrt{d}\sigma) = O(L)$. The proof of the privacy guarantee relies on the moments accountant analysis, for which we provide the following restatement.

Theorem 22 (Abadi et al. (2016); Kulkarni et al. (2021)) *Let $\epsilon, \delta \in (0, 1]$ and c be a universal constant. Let $D \in \mathcal{Y}^n$ be a dataset over some domain $\mathcal{Y}$, and let $h_1, ..., h_T : \mathcal{Y} \mapsto \mathbb{R}^d$ be a series of (possibly adaptive) queries such that for any $y \in \mathcal{Y}$, $t \in [T]$, $\|h_t(y)\|_2 \leq L$. Let $\sigma \geq \frac{cL\sqrt{T\log(1/\delta)}}{n\epsilon}$ and $T \geq \frac{n^2\epsilon}{b^2}$. Then the algorithm which samples batches of size $B_1, .., B_t$ of size b uniformly at random and outputs $\frac{1}{b}\sum_{y\in B_t} h_t(y) + g_t$ for all $t \in [T]$ where $g_t \sim \mathcal{N}(0, \mathbb{I}\sigma^2)$, is (ϵ, δ)-DP.*

It can be verified for the described noisy SGDA implementation that $\sigma \geq \frac{c_1 L\sqrt{T\log(1/\delta)}}{n\epsilon}$ and $T \geq \frac{n^2\epsilon}{m^2}$ and thus the algorithm is (ϵ, δ)-DP.

Appendix D. Missing Result from Section 5

D.1. Low variance and weak gap implies strong gap

Proof [proof of Proposition 14] Consider the virtual algorithm, $\mathcal{B}(\mathcal{A}, \mathcal{D}) = \mathop{\mathbb{E}}_{\hat{S}\sim\mathcal{D}^n, \mathcal{A}}\left[\mathcal{A}(\hat{S})\right] = [\widetilde{w}, \widetilde{\theta}]$. Note this algorithm is deterministic and does not depend on any specific dataset drawn from $\mathcal{D}$. We first show that gap function at the output of $\mathcal{B}$ is bounded by the weak gap of $\mathcal{A}$. We have

$$\begin{aligned}
\widehat{\mathrm{Gap}}(\mathcal{B}(\mathcal{A}, \mathcal{D})) &= \max_{\theta\in\Theta}\left\{F_{\mathcal{D}}(\mathcal{B}_w(\mathcal{A}, \mathcal{D}), \theta)\right\} - \min_{w\in\mathcal{W}}\left\{F_{\mathcal{D}}(w, \mathcal{B}_\theta(\mathcal{A}, \mathcal{D}))\right\} \\
&= \max_{\theta\in\Theta}\left\{F_{\mathcal{D}}\Big(\mathop{\mathbb{E}}_{\hat{S}\sim\mathcal{D}^n, \mathcal{A}_w}\left[\mathcal{A}_w(\hat{S})\right], \theta\Big)\right\} - \min_{w\in\mathcal{W}}\left\{F_{\mathcal{D}}\Big(w, \mathop{\mathbb{E}}_{\hat{S}\sim\mathcal{D}^n, \mathcal{A}_\theta}\left[\mathcal{A}_\theta(\hat{S})\right]\Big)\right\} \\
&\leq \max_{\theta\in\Theta}\left\{\mathop{\mathbb{E}}_{\hat{S}\sim\mathcal{D}^n, \mathcal{A}_w}\left[F_{\mathcal{D}}(\mathcal{A}_w(\hat{S}), \theta)\right]\right\} - \min_{w\in\mathcal{W}}\left\{\mathop{\mathbb{E}}_{\hat{S}\sim\mathcal{D}^n, \mathcal{A}_\theta}\left[F_{\mathcal{D}}(w, \mathcal{A}_\theta(\hat{S}))\right]\right\} \\
&= \mathrm{Gap}_{\mathsf{weak}}(\mathcal{A}),
\end{aligned} \tag{15}$$

where the second equality follows from the definition of $\mathcal{B}$ and the inequality follows from Jensen's inequality.

Now by the assumption that $\mathcal{A}$ is low variance, we have

$$\mathop{\mathbb{E}}_{\mathcal{A},S}\left[\|\mathcal{A}(S)-\mathcal{B}(\mathcal{A},\mathcal{D})\|^2\right]=\mathop{\mathbb{E}}_{\mathcal{A},S}\left[\left\|\mathcal{A}(S)-\mathop{\mathbb{E}}_{\hat{S}\sim\mathcal{D}^n,\mathcal{A}}\left[\mathcal{A}(\hat{S})\right]\right\|^2\right]\leq\tau^2. \tag{16}$$

■

Thus using the Lipschitzness of $\widehat{\text{Gap}}$ we obtain

$$\begin{aligned}\text{Gap}(\mathcal{A})-\text{Gap}_{\mathsf{weak}}(\mathcal{A})&=\mathop{\mathbb{E}}_{S,\mathcal{A}}\left[\widehat{\text{Gap}}(\mathcal{A}_w(S),\mathcal{A}_\theta(S))\right]-\text{Gap}_{\mathsf{weak}}(\mathcal{A})\\&\leq\mathop{\mathbb{E}}_{S,\mathcal{A}}\left[\widehat{\text{Gap}}(\mathcal{A}_w(S),\mathcal{A}_\theta(S))\right]-\widehat{\text{Gap}}(\mathcal{B}(\mathcal{A},\mathcal{D}))\\&\leq L\mathop{\mathbb{E}}_{S,\mathcal{A}}\left[\|\mathcal{A}(S)-\mathcal{B}(\mathcal{A},\mathcal{D})\|\right]\leq L\tau.\end{aligned}$$

The first inequality comes from Eqn. (15). The second inequality comes from the Lipschitzness of the gap function. The third inequality comes from Eqn. (16). Thus we ultimately have

$$\text{Gap}(\mathcal{A})\leq\text{Gap}_{\mathsf{weak}}(\mathcal{A})+L\tau. \tag{17}$$

D.2. Stability-Risk Tradeoff

Proof [proof of Theorem 15] Let $f(w;x)=\langle w,x\rangle$. Let $0<K<\min\{n,d\}$ be a parameter to be chosen later and define $U=\{\pm1\}^K$. For any $\boldsymbol{\sigma}\in U$ define $S_{\boldsymbol{\sigma}}=\{L\boldsymbol{\sigma}_1e_1,...,L\boldsymbol{\sigma}_Ke_K,0,...,0\}$, where e_j is the j'th standard basis vector. We will denote $F(w;S_{\boldsymbol{\sigma}})=\frac{1}{n}\sum_{x\in S_{\boldsymbol{\sigma}}}f(w;x)$. Note that

$$w^*_{\boldsymbol{\sigma}}=\mathop{\arg\min}_{w\in\mathcal{W}}\{F(w;S_{\boldsymbol{\sigma}})\}=\frac{B}{\sqrt{K}}\sum_{j\in[K]}-\boldsymbol{\sigma}_je_j.$$

Further, for any $\boldsymbol{\sigma}\in U$, $F(w^*_{\boldsymbol{\sigma}};S_{\boldsymbol{\sigma}})=-\frac{BL\sqrt{K}}{n}$.

By Yao's minimax principle, it suffices to consider deterministic algorithms and lower bound the expected risk w.r.t. some distribution over the packing. Considering the uniform distribution over the packing and setting $K=\frac{B^2}{\Delta^2}$ we have

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim\text{Unif}(U)}\left[F(\mathcal{A}(S_{\boldsymbol{\sigma}});S_{\boldsymbol{\sigma}})-F(w^*_{\boldsymbol{\sigma}};S_{\boldsymbol{\sigma}})\right] &= \frac{1}{|U|}\sum_{\boldsymbol{\sigma}\in U}F(\mathcal{A}(S_{\boldsymbol{\sigma}});S_{\boldsymbol{\sigma}})+\frac{BL\sqrt{K}}{n} \\
&\overset{(i)}{=} \frac{1}{|U|}\sum_{\boldsymbol{\sigma}\in U}\left[\frac{1}{n}\sum_{j\in[K]}L\boldsymbol{\sigma}_j\mathcal{A}(S_{\boldsymbol{\sigma}})_j+\frac{1}{n}\sum_{j\in[K]}\frac{BL}{\sqrt{K}}\right] \\
&= \frac{1}{n|U|}\sum_{j\in[K]}\sum_{\boldsymbol{\sigma}\in U}L\boldsymbol{\sigma}_j\mathcal{A}(S_{\boldsymbol{\sigma}})_j+\frac{BL}{\sqrt{K}} \\
&= \frac{1}{n|U|}\sum_{j\in[K]}\sum_{\boldsymbol{\sigma}\in U:\boldsymbol{\sigma}_j=1}L\left(\mathcal{A}(S_{\boldsymbol{\sigma}})_j-\mathcal{A}(S_{\boldsymbol{\sigma}_{-j}})_j\right)+\frac{2BL}{\sqrt{K}} \\
&\overset{(ii)}{\geq} \frac{1}{n|U|}\sum_{j\in[K]}\sum_{\boldsymbol{\sigma}\in U:\boldsymbol{\sigma}_j=1}-L\Delta+\frac{2BL}{\sqrt{K}} \\
&= \frac{1}{n|U|}\sum_{j\in[K]}\sum_{\boldsymbol{\sigma}\in U:\boldsymbol{\sigma}_j=1}\frac{BL}{\sqrt{K}} \\
&= \frac{BL\sqrt{K}}{2n}
\end{aligned}
$$

where (i) comes from the definition of the loss function and the fact that the dataset consists of K standard basis vectors (up to sign) and $n-K$ zero vectors and (ii) comes from the $\Delta=\frac{B}{\sqrt{K}}$ stability property of $\mathcal{A}$ (i.e. $\mathcal{A}(S_{\boldsymbol{\sigma}_{-j}})_j-\mathcal{A}(S_{\boldsymbol{\sigma}})_j\leq\Delta \implies \mathcal{A}(S_{\boldsymbol{\sigma}})_j-\mathcal{A}(S_{\boldsymbol{\sigma}_{-j}})_j\geq-\Delta$). Finally, note that by the setting of K that $\frac{BL\sqrt{K}}{n}=\frac{B^2L}{\Delta n}$. ■