Mathematical Programming
https://doi.org/10.1007/s10107-023-01953-5

FULL LENGTH PAPER

Series A

Optimal algorithms for differentially private stochastic monotone variational inequalities and saddle-point problems

Digvijay Boob[1] · **Cristóbal Guzmán**[2]

Received: 2 April 2021 / Accepted: 9 March 2023

Abstract
In this work, we conduct the first systematic study of stochastic variational inequality (SVI) and stochastic saddle point (SSP) problems under the constraint of differential privacy (DP). We propose two algorithms: Noisy Stochastic Extragradient (NSEG) and Noisy Inexact Stochastic Proximal Point (NISPP). We show that a stochastic approximation variant of these algorithms attains risk bounds vanishing as a function of the dataset size, with respect to the strong gap function; and a sampling with replacement variant achieves optimal risk bounds with respect to a weak gap function. We also show lower bounds of the same order on weak gap function. Hence, our algorithms are optimal. Key to our analysis is the investigation of algorithmic stability bounds, both of which are new even in the nonprivate case. The dependence of the running time of the sampling with replacement algorithms, with respect to the dataset size n, is n^2 for NSEG and $\widetilde{O}(n^{3/2})$ for NISPP.

Keywords Variational inequalities · Saddle-point problems · Differential privacy · Stochastic algorithms · Algorithmic stability

Mathematics Subject Classification 49M37 · 90C47 · 90C31 · 90C30 · 62L20

DB was partially supported by the NSF grants CCF 1909298, CCF 2245705. CG was partially supported by INRIA through the INRIA Associate Teams project, ANID—Millenium Science Initiative Program—NCN17_059, and FONDECYT 1210362 project.

✉ Digvijay Boob
dboob@smu.edu

Cristóbal Guzmán
crguzmanp@mat.uc.cl

[1] Department of Operations Research and Engineering Management, Southern Methodist University, Dallas, USA

[2] Institute for Mathematical and Computational Eng., Pontificia Universidad Católica de Chile, Santiago, Chile

Published online: 19 April 2023

1 Introduction

Stochastic variational inequalities (SVI) and stochastic saddle-point (SSP) problems have become a central part of the modern machine learning toolbox. The main motivation behind this line of research is the design of algorithms for multiagent systems and adversarial training, which are more suitably modeled by the language of games, rather than pure (stochastic) optimization. Applications that rely on these methods may often involve the use of sensitive user data, so it becomes important to develop algorithms for these problems with provable privacy-preserving guarantees. In this context, differential privacy (DP) has become the gold standard of privacy-preserving algorithms, thus a natural question is whether it is possible to design DP algorithms for SVI and SSP that attain high accuracy.

Motivated by these considerations, this work provides the first systematic study of differentially-private SVI and SSP problems. Before proceeding to the specific results, we present more precisely the problems of interest. The stochastic variational inequality (SVI) problem is: given a monotone operator $F : \mathcal{W} \mapsto \mathbb{R}^d$ in expectation form $F(w) = \mathbb{E}_{\boldsymbol{\beta}\sim\mathcal{P}}[F_{\boldsymbol{\beta}}(w)]$, find $w^* \in \mathcal{W}$ such that

$$\langle F(w^*), w - w^* \rangle \geqslant 0 \quad \forall w \in \mathcal{W}. \tag{VI(F)}$$

The closely related stochastic saddle point (SSP) problem is: given a convex-concave real-valued function $f : \mathcal{W} \mapsto \mathbb{R}$ (here $\mathcal{W} = \mathcal{X} \times \mathcal{Y}$ is a product space), given in expectation form $f(x, y) = \mathbb{E}_{\boldsymbol{\beta}\sim\mathcal{P}}[f_{\boldsymbol{\beta}}(x, y)]$, the goal is to find (x^*, y^*) that solves

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x, y). \tag{SP(f)}$$

In both of these problems, the input to the algorithm is an i.i.d. sample $\mathbf{S} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n) \sim \mathcal{P}^n$. Uncertainty introduced by a finite random sample renders the computation of exact solutions infeasible, so gap (a.k.a. population risk) functions are used to quantify the quality of solutions. Let $\mathcal{A} : \mathcal{Z}^n \mapsto \mathcal{W}$ be an algorithm for SVI problems (VI(F)).

We define the *strong VI-gap* associated with $\mathcal{A}$ as

$$\mathrm{Gap}_{\mathrm{VI}}(\mathcal{A}, F) := \mathbb{E}_{\mathcal{A},\mathbf{S}}\left[\sup_{w\in\mathcal{W}} \langle F(w), \mathcal{A}(\mathbf{S}) - w\rangle\right]. \tag{1.1}$$

We also define the *weak VI-gap* as

$$\mathrm{WeakGap}_{\mathrm{VI}}(\mathcal{A}, F) := \mathbb{E}_{\mathcal{A}} \sup_{w\in\mathcal{W}} \mathbb{E}_{\mathbf{S}}\left[\langle F(w), \mathcal{A}(\mathbf{S}) - w\rangle\right]. \tag{1.2}$$

Here, expectation is taken over both the sample data $\mathbf{S}$ and the internal randomization of $\mathcal{A}$. For SSP (SP(f)), given an algorithm $\mathcal{A} : \mathcal{Z}^n \mapsto \mathcal{X} \times \mathcal{Y}$, and letting $\mathcal{A}(\mathbf{S}) = (x(\mathbf{S}), y(\mathbf{S}))$, a natural gap function is the following *saddle-point (a.k.a. primal-dual)*

gap

$$\mathrm{Gap}_{\mathrm{SP}}(\mathcal{A}, f) := \mathbb{E}_{\mathcal{A},\mathbf{S}}\left[\sup_{x\in\mathcal{X},\, y\in\mathcal{Y}} [f(x(\mathbf{S}), y) - f(x, y(\mathbf{S}))]\right]. \tag{1.3}$$

Analogously as above, we define the *weak SSP gap* as[1]

$$\mathrm{WeakGap}_{\mathrm{SP}}(\mathcal{A}, f) := \mathbb{E}_{\mathcal{A}} \sup_{x\in\mathcal{X},\, y\in\mathcal{Y}} \mathbb{E}_{\mathbf{S}}[f(x(\mathbf{S}), y) - f(x, y(\mathbf{S}))]. \tag{1.4}$$

It is easy to see that in both cases the gap is always nonnegative, and any exact solution must have zero-gap. For examples and applications of SVI and SSP we refer to Sect. 2.1. Despite the fact that the strong VI is a more classical and well-studied quantity, the weak VI gap has been observed to be useful in various contexts. We refer the reader to [50] for more discussions on the weak VI gap.

On the other hand, we are interested in designing algorithms that are *differentially private*. These algorithms build a solution based on a given dataset $\mathbf{S}$ of random i.i.d. examples from the target distribution, and output a (randomized) feasible solution, $\mathcal{A}(\mathbf{S})$. We say that two datasets $\mathbf{S} = (\boldsymbol{\beta}_i)_i$, $\mathbf{S}' = (\boldsymbol{\beta}'_i)_i$ are neighbors, denoted $\mathbf{S} \simeq \mathbf{S}'$, if they only differ in a single entry i. We say that an algorithm $\mathcal{A}(\mathbf{S})$ is (ε, η)-differentially private if for every event E in the output space[2]

$$\mathbb{P}_{\mathcal{A}}[\mathcal{A}(\mathbf{S}) \in E] \leqslant e^{\varepsilon}\mathbb{P}_{\mathcal{A}}[\mathcal{A}(\mathbf{S}') \in E] + \eta \quad (\forall \mathbf{S} \simeq \mathbf{S}'). \tag{1.5}$$

Here $\varepsilon, \eta \geqslant 0$ are prescribed parameters that quantify the privacy guarantee. Designing DP algorithms for particular data analysis problems is an active area of research. Optimal risk algorithms for stochastic convex optimization have only very recently been developed, and it is unclear whether these methods are extendable to SVI and SSP settings.

1.1 Summary of contributions

Our work is the first to provide population risk bounds for DP-SVI and DP-SSP problems. Moreover, our algorithms attain provably optimal rates and are computationally efficient. We summarize our contributions as follows:

1. We provide two different algorithms for DP-SVI and DP-SSP: namely, the noisy stochastic extragradient method (NSEG) and a noisy inexact stochastic proximal-point method (NISPP). The NSEG method is a natural DP variant of the well-known stochastic extragradient method [30], where privacy is obtained

[1] The denominations of weak and strong gap functions used in this paper are not standard, but we believe are the most appropriate in this context. For example, in [50] used the terms *weak and strong generalization measure* for (1.4) and (1.3) respectively, but it is clear that these quantities do not refer to standard generalization measures used in stochastic optimization.

[2] Note that the probabilities in the definition of DP only involve the probability space of algorithmic randomization, and not of the datasets, which is emphasized by the notation $\mathbb{P}_{\mathcal{A}}$. The datasets must be neighbors, but they are otherwise arbitrary, and this is crucial to certify the privacy for any user.

by Gaussian noise addition; on the other hand, the NISPP method is an approximate proximal point algorithm [28, 43] in which every proximal iterate is made noisy to make it differentially private. Our more basic variants of both of these methods are based on iterations involving disjoint sets of datapoints (a.k.a. single pass method), which are known to typically lead to highly suboptimal rates in DP (see the Related Work Section for further discussion).

2. We derive novel uniform stability bounds for the NSEG and NISSP methods. For NSEG, our stability upper bounds are inspired by the interpretation of the extragradient method as a (second order) approximation of the proximal point algorithm. In particular, we provide expansion bounds for the extragradient iterates, and solve a (stochastic) linear recursion. The stability bounds for NISPP method are based on stability of the (unique) SVI solution in the strongly monotone case. Finally, we investigate the risk attained by multipass versions of the NSEG and NISPP methods, leveraging known generalization bounds for stable algorithms [35]. Here, we show that the optimal risk for DP-SVI and DP-SSP can be attained by running these algorithms with their sampling with replacement variant. In particular, NSEG method requires n^2 stochastic operator evaluations, and NISPP method requires much smaller $\widetilde{O}(n^{3/2})$ operator evaluations for both DP-SVI and DP-SSP problems. In particular, these upper bounds also show the dependence of the running time of each of these algorithms w.r.t. the dataset size.
3. Finally, we prove lower bounds on the weak gap function for any DP-SSP and DP-SVI algorithm, showing the risk optimality of the aforementioned multipass algorithms. The main challenge in these lower bounds is showing that existing constructions of lower bounds for DP convex optimization [5, 7, 46] lead to lower bounds on the weak gap of a related SP/VI problem.

The following table provides details of population risk and operator evaluation complexity.

1.2 Related work

We divide our discussion on related work in three main areas. Each of these areas has been extensively investigated, so a thorough description of existing work is not possible. We focus ourselves on the work which is more directly related to our own.

1. **Stochastic Variational Inequalities and Saddle-Point Problems:** Variational inequalities and saddle-point problems are classical topics in applied mathematics, operations research and engineering (e.g., [3, 18, 33, 38, 40–43, 45]). Their stochastic counterparts have only gained traction recently, mainly motivated by their applications in machine learning (e.g., [24, 25, 29, 30, 34] and references therein). For the stochastic version of (SP(f)), [39] proposed a robust stochastic approximation method. The first optimal algorithm for SVI with monotone Lipschitz operators was obtained by Juditsky, Nemirovski and Tauvel [30], and very recently Kotsalis, Lan and Li [34] developed optimal variants for the strongly monotone case (in terms of distance to the optimum criterion, rather than VI gap). It is important to note that naive adaptation of these methods to the DP setting

requires adding noise to the operator evaluations at every iteration, which substantially degrades the accuracy of the obtained solution. A careful privacy accounting and minibatch schedule can lead to optimal guarantees for single-pass methods [19], however this requires accuracy guarantees for the last iterate, which is currently an open problem for SVI and SSP (aside from specific cases, typically involving strong monotonicity conditions, e.g., [24, 34]). We circumvent this problem by providing population risk guarantees for *multipass methods*.

2. **Stability and Generalization:** Deriving generalization (or population risk) bounds for general-purpose algorithms is a challenging task, actively studied in theoretical machine learning. Bousquet and Elisseeff [8] provided a systematic treatment of this question for algorithms which are *stable*, with respect to changes of a single element in the training dataset, and a sequence of works have refined these generalization guarantees (see [9, 21] and references therein). This idea has been applied to investigate the generalization properties of regularized empirical risk minimization [8, 44], and more recently to iterative methods, such as stochastic gradient descent [4, 23].
   Using stability to obtain population risk bounds in SVI and SSP is substantially more challenging, due to the presence of a supremum in the accuracy measure (see Eqs. (1.1) and (1.3)). Recently, Zhang et al. [50], established stability implies generalization results for the strong SP gap under strong monotonicity assumptions. Their proof strategy applies analogously to address the SVI setting, although this is not carried out in their work. More recently, Lei et al. [35], proved generalization bounds on the weak SP gap without strong monotonicity assumptions. We leverage this result for our algorithms, and further elaborate on its implications for SVI in Sect. 2.2.
3. **Differential Privacy:** Differential privacy is the gold standard for private data analysis, and it has been studied for nearly 20 years [15, 16]. Beyond its classical definition, multiple variants have been introduced, including local [14, 31], concentrated [10], Rényi [37], and Gaussian [13]. Relevant to the optimization community are the applications of differential privacy to combinatorial optimization [22].
   Differentially private empirical risk minimization and stochastic convex optimiza-

Table 1 Different levels of risk and complexity achieved by NSEG and NISPP methods for (ε, η)-differentially private SVI/SSP. Here n is the dataset size, and d is the dimension of the solution search space. We omit the dependence on other problem parameters (e.g., Lipschitz constants and diameter), as well as the privacy parameter η

	Type of sampling single pass	multipass	Type of sampling single pass	multipass
Method	Criterion - Strong Gap	Criterion - Weak Gap	Number of operator evaluations	
NSEG	$O\left(\frac{d^{1/4}}{\sqrt{n\varepsilon}} + \frac{\sqrt{d}}{n\varepsilon}\right)$	$O\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\varepsilon}\right)$	n	n^2
NISPP (OE subroutine)	$O\left(\frac{1}{n^{1/3}} + \frac{\sqrt{d}}{n^{2/3}\varepsilon}\right)$	$O\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\varepsilon}\right)$	$O(n \log n)$	$O(n^{3/2} \log n)$

tion have been extensively studied for over a decade (see, e.g. [5, 7, 11, 12, 19, 26, 27, 32, 47]). Relevant to our work are the first optimal risk algorithms for DP-ERM [7] and DP-SCO [5]. Non-Euclidean extensions have also been obtained recently [2, 6]. To the best of our knowledge, our work is the first to address DP algorithms for SVI and SSP. Our approach for generalization of multipass algorithms is inspired by the noisy SGD analysis in [4]. However, our stability analysis differs crucially from [4]: in the case of NSEG, we need to carefully address the double operator evaluation of the extragradient step, which is done by using the fact that the extragradient operator is approximately nonexpansive. In the case of NISPP, we leverage the contraction properties of strongly monotone VI solutions. By contrast, SGD in the nonsmooth case is far from nonexpansive [4]. Alternative approaches to obtain optimal risk in DP-SCO, including privacy amplification by iteration [19, 20], and phased regularization or phased SGD [19], appear to run into fundamental limitations when applied to DP-SVI and DP-SSP. It is an interesting future research direction to obtain faster running times with optimal population risk in DP-SVI and DP-SSP, which may benefit from these alternative approaches.

The main body of this paper is organized as follows. In Sect. 2, we provide the necessary background information on SVI/SSP, uniform stability, and differential privacy, which are necessary for the rest of the paper. In Sect. 3 we introduce the NSEG method, together with its basic privacy and accuracy guarantee for a single pass version. Section 4 provides stability bounds for NSEG method along with the consequent optimal rates for SVI and SSP. In Sect. 5, we introduce the single-pass differentially private NISPP method with bound on expected SVI-gap. Section 6 presents stability analysis of NISPP, together with the resulting optimal rates for SVI/SSP gap. We conclude in Sect. 7 with lower bounds that prove the optimality of the obtained rates.

2 Notation and preliminaries

We work on the Euclidean space $(\mathbb{R}^d, \langle\cdot,\cdot\rangle)$, where $\langle\cdot,\cdot\rangle$ is the standard inner product, and $\|u\| = \sqrt{\langle u, u\rangle}$ is the ℓ_2-norm. Throughout, we consider a compact convex set $\mathcal{W} \subseteq \mathbb{R}^d$ with diameter $D > 0$. We denote the standard Euclidean projection operator on set $\mathcal{W}$ by $\Pi_{\mathcal{W}}(\cdot)$. The identity matrix on $\mathbb{R}^d$ is denoted by $\mathbb{I}_d$.

We let $\mathcal{P}$ denote an unknown distribution supported on an arbitrary set $\mathcal{Z}$, from which we have access to exactly n i.i.d. datapoints which we denote by sample set $\mathbf{S} \sim \mathcal{P}^n$. Throughout, we will use boldface characters to denote sources of randomness (coming from the data, or internal algorithmic randomization). We say that two datasets $\mathbf{S}, \mathbf{S}'$ are adjacent (or neighbors), denoted by $\mathbf{S} \simeq \mathbf{S}'$, if they differ in a single data point. We also denote subsets (a.k.a. batches), or single data points, of $\mathbf{S}$ or $\mathcal{P}$ by $\mathbf{B}$ and $\boldsymbol{\beta}$, respectively. Whether $\boldsymbol{\beta}$ or $\mathbf{B}$ is sampled from $\mathcal{P}$ or $\mathbf{S}$ is specified explicitly unless it is clear from the context. For a batch $\mathbf{B}$, we denote its size by $|\mathbf{B}|$. Therefore, we have $|\mathbf{S}| = n$. Throughout, we will denote Gaussian random variables by $\boldsymbol{\xi}$.

We say that $F : \mathcal{W} \to \mathbb{R}^d$ is a monotone operator if

$$\langle F(w_1) - F(w_2), w_1 - w_2\rangle \geqslant 0, \quad \forall w_1, w_2 \in \mathcal{W}.$$

Given $L > 0$, we say that F is L-Lipschitz continuous, if

$$\|F(w_1) - F(w_2)\| \leqslant L\|w_1 - w_2\|, \quad \forall w_1, w_2 \in \mathcal{W}.$$

Finally, we say that F is M-bounded if $\sup_{w\in\mathcal{W}} \|F(w)\| \leqslant M$. We denote the set of monotone, L-Lipschitz and M-bounded operators by $\mathcal{M}^1_{\mathcal{W}}(M, L)$. In this work, we will focus on the case where F is an expectation operator, i.e., $F(w) := \mathbb{E}_{\boldsymbol{\beta}\sim\mathcal{P}}[F_{\boldsymbol{\beta}}(w)]$, where $\mathcal{P}$ is an arbitrary distribution supported on $\mathcal{Z}$,

and for any $\boldsymbol{\beta}$ in $\mathcal{Z}$, $F_{\boldsymbol{\beta}}(\cdot) \in \mathcal{M}^1_{\mathcal{W}}(M, L)$, $\boldsymbol{\beta}$-a.s.[3]

In the stochastic saddle point problem (SP(f)), we modify the notation slightly. Here, $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_2}$ are compact convex sets, and we will assume that the saddle point functions $f_{\boldsymbol{\beta}}(\cdot, \cdot) : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$, satisfy the following conditions $\boldsymbol{\beta}$-a.s.

- $\nabla_x f_{\boldsymbol{\beta}}(\cdot, \cdot)$ is L_x-Lipschitz continuous, and $\nabla_y f_{\boldsymbol{\beta}}(\cdot, \cdot)$ is L_y-Lipschitz continuous, and;
- $f_{\boldsymbol{\beta}}(\cdot, y)$ is convex, for any given $y \in \mathcal{Y}$, and $f_{\boldsymbol{\beta}}(x, \cdot)$ is concave, for any given $x \in \mathcal{X}$ (we will say in this case the function is convex-concave).

If the assumptions above are met, we will denote $L \triangleq \sqrt{L_x^2 + L_y^2}$. Under the assumptions above, it is well-known that SSP [42, 45] (and SVI [18], respectively) have a solution.

In the case of saddle-point problems, given the convex-concave function $f_{\boldsymbol{\beta}}(\cdot, \cdot) : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$, it is well-known that the operator $F : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}^d \times \mathbb{R}^d$ below is monotone

$$F_{\boldsymbol{\beta}}(x, y) = (\nabla_x f_{\boldsymbol{\beta}}(x, y), -\nabla_y f(x, y)). \tag{2.1}$$

We will call this operator the *monotone operator associated with* $f_{\boldsymbol{\beta}}(\cdot, \cdot)$. Furthermore, if $\nabla_x f_{\boldsymbol{\beta}}(\cdot, y)$ has L_x-Lipschitz continuous gradient and $\nabla_y f_{\boldsymbol{\beta}}(x, \cdot)$ has L_y-Lipschitz continuous gradient, then F is $\sqrt{L_x^2 + L_y^2}$-Lipschitz continuous.

It is easy to see that, given a SSP problem with function $f_{\boldsymbol{\beta}}(\cdot, \cdot)$ and sets $\mathcal{X}$, $\mathcal{Y}$, an (exact) SVI solution (VI(F)) for the monotone operator associated to $f(x, y) = \mathbb{E}_{\boldsymbol{\beta}}[f_{\boldsymbol{\beta}}(x, y)]$ over the set $\mathcal{W} = \mathcal{X} \times \mathcal{Y}$, yields an exact SSP solution for the starting problem. Unfortunately, such reduction does not directly work for approximate solutions to (1.1) and (1.3), so the analysis must be done separately for both problems.

For batch $\mathbf{B}$, we denote the empirical (a.k.a. sample average) operator $F_{\mathbf{B}}(w) := \frac{1}{|\mathbf{B}|}\sum_{\boldsymbol{\beta}\in\mathbf{B}} F_{\boldsymbol{\beta}}(w)$. On the other hand, for a batch $\mathbf{B}$, the empirical saddle point function is denoted as $f_{\mathbf{B}}(x, y) = \frac{1}{|\mathbf{B}|}\sum_{\boldsymbol{\beta}\in\mathbf{B}} f_{\boldsymbol{\beta}}(x, y)$. Given a distribution $\mathcal{P}$, the expectation operator and function are denoted by $F_{\mathcal{P}}(w) := \mathbb{E}_{\boldsymbol{\beta}\sim\mathcal{P}}[F_{\boldsymbol{\beta}}(w)]$, and $f_{\mathcal{P}}(x, y) = \mathbb{E}_{\boldsymbol{\beta}\sim\mathcal{P}}[f_{\boldsymbol{\beta}}(x, y)]$, respectively. For brevity, whenever it is clear from context we will drop the dependence on $\mathcal{P}$.

[3] Here, we mean that for almost every $\boldsymbol{\beta}$, we have $F_{\boldsymbol{\beta}} \in \mathcal{M}^1_{\mathcal{W}}(M, L)$.

2.1 Examples and applications of SVI and SSP

An interesting problem which can be formulated as a SSP-problem is the minimization of a max-type convex function:

$$\min_{x\in\mathcal{X}}\left\{\phi(x) := \max_{1\leqslant j\leqslant m}\phi_j(x)\right\},$$

where $\phi_j : \mathcal{X} \to \mathbb{R}$ is a stochastic convex function $\phi_j(x) := \mathbb{E}_{\zeta_j\sim\mathcal{P}_j}[\phi_{j,\zeta_j}(x)]$ for all $j \in [m]$. This problem is essentially a structured nonsmooth optimization problem which can be reformulated into a convex-concave saddle point problem:

$$\min_{x\in\mathcal{X}}\max_{y\in\Delta_m}\mathbb{E}_{\zeta_1\ldots\zeta_m}[\textstyle\sum_{j=1}^m y_i\phi_{j,\zeta_j}(x)]$$

Here, $\boldsymbol{\beta} = (\zeta_j)_{j=1}^m$ is the random input to the saddle point problem: $f_{\boldsymbol{\beta}}(x, y) = \sum_{j=1}^m y_j\phi_{j,\zeta_j}(x)$. Note that a substantial generalization of the max-type problem above is the so called compositional optimization problem:

$$\min_{x\in\mathcal{X}}\phi(x) := \Phi(\phi_1(x), \ldots, \phi_m(x)),$$

where $\phi_j(x)$ are convex maps and $\Phi(u_1, \ldots, u_m)$ is a real-valued convex function whose Fenchel-type representation is assumed to have the form

$$\Phi(u_1, \ldots, u_m) = \max_{y\in\mathcal{Y}}\textstyle\sum_{j=1}^m\langle u_j, A_j y + b_j\rangle - \Phi_*(y),$$

where Φ_* is a convex, Lipschitz and smooth. Then, overall optimization problem can be reformulated as a convex-concave saddle point problem:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\textstyle\sum_{j=1}^m\langle\phi_j(x), A_j y + b_j\rangle - \Phi_*(y),$$

where stochasticity is introduced due to constituent functions $\phi_j(x) = \mathbb{E}_{\zeta_j}[\phi_{j,\zeta_j}(x)]$.

To conclude, we remark that these types of models have been recently proposed in machine learning to address *approximate fairness* [49] and federated learning on heterogeneous populations [36]. In these examples, the different indices $j \in [m]$ may denote different subgroups from a population, and we are interested in bounding the (excess) population risk on these subgroups uniformly (with the motivation of preventing discrimination against any subgroup). This clearly cannot be achieved by a stochastic convex program, and a stochastic saddle-point formulation is effective in certifying accuracy across the different subgroups separately.

For further examples and applications of stochastic variational inequalities and saddle-point problems, we refer the reader to [29, 30, 50].

2.2 Algorithmic stability

In general, an algorithm is a randomized function mapping datasets to candidate solutions, $\mathcal{A} : \mathcal{Z}^n \mapsto \mathbb{R}^d$, which is measurable w.r.t. the dataset. Two datasets, $\mathbf{S} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n)$, $\mathbf{S}' = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_n) \in \mathcal{Z}^n$ are said to be neighbors (denoted $\mathbf{S} \simeq \mathbf{S}'$) if they only differ in at most one data point, namely

$$\boldsymbol{\beta}_j = \boldsymbol{\beta}'_j \qquad (\exists i \in [n])(\forall j \neq i).$$

Algorithmic stability is a notion of sensitivity analysis of an algorithm under neighboring datasets. Of particular interest to our work is the notion of *uniform argument stability* (UAS).

Definition 1 (Uniform Argument Stability) Let $\mathcal{A} : \mathcal{Z}^n \mapsto \mathbb{R}^d$ be a randomized mapping and $\delta > 0$. We say that $\mathcal{A}$ is δ-uniformly argument stable (for short, δ-UAS) if

$$\sup_{\mathbf{S} \simeq \mathbf{S}'} \mathbb{E}_{\mathcal{A}} \|\mathcal{A}(\mathbf{S}) - \mathcal{A}(\mathbf{S}')\| \leqslant \delta.$$

Occasionally, we may denote $\delta_{\mathcal{A}}(\mathbf{S}, \mathbf{S}') \triangleq \|\mathcal{A}(\mathbf{S}) - \mathcal{A}(\mathbf{S}')\|$, for convenience. The importance of algorithmic stability in machine learning comes from the fact that *stability implies generalization* in stochastic optimization and stochastic saddle point (SSP) problems [8, 9, 50]. Below, we restate existing results on stability implies generalization for SSP problems below. Before doing so we need to briefly introduce the *(strong) empirical gap function*: given a dataset $\mathbf{S}$ and an algorithm $\mathcal{A}$, we define the empirical gap function for a saddle point and variational inequality problem respectively as

$$\mathrm{EmpGap}_{\mathrm{SP}}(\mathcal{A}, f_{\mathbf{S}}) := \mathbb{E}_{\mathcal{A}}[\sup_{x,y} f_{\mathbf{S}}(x(\mathbf{S}), y) - f_{\mathbf{S}}(x, y(\mathbf{S}))] \tag{2.2}$$

$$\mathrm{EmpGap}_{\mathrm{VI}}(\mathcal{A}, F_{\mathbf{S}}) := \mathbb{E}_{\mathcal{A}}[\sup_{w} \langle F_{\mathbf{S}}(w), \mathcal{A}(\mathbf{S}) - w \rangle]. \tag{2.3}$$

Notice that in these definitions the dataset $\mathbf{S}$ is fixed.

Proposition 1 [35, 50] *Consider the stochastic saddle point problem (SP(f)) with functions $f_{\boldsymbol{\beta}}(\cdot, y)$ and $f_{\boldsymbol{\beta}}(x, \cdot)$ being M-Lipschitz for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $\boldsymbol{\beta}$-a.s.. Let $\mathcal{A} : \mathcal{Z}^n \to \mathcal{X} \times \mathcal{Y}$ be an algorithm, where $\mathcal{A}(\mathbf{S}) = (x(\mathbf{S}), y(\mathbf{S}))$. If $x(\cdot)$ is δ_x-UAS and $y(\cdot)$ is δ_y-UAS, and both are integrable, then*

$$\mathit{WeakGap}_{SP}(\mathcal{A}, f) \leqslant \mathbb{E}_{\mathbf{S}}[\mathit{EmpGap}_{SP}(\mathcal{A}, f_{\mathbf{S}})] + M[\delta_x + \delta_y]. \tag{2.4}$$

This result can be extended for SVI problems as well. We provide a formal statement below and prove it in Appendix A.

Proposition 2 *Consider a stochastic variational inequality with M-bounded operators $F_{\boldsymbol{\beta}}(\cdot) : \mathcal{W} \mapsto \mathbb{R}^d$. If $\mathcal{A} : \mathcal{Z}^n \mapsto \mathcal{W}$ is integrable and δ-UAS, then*

$$\mathit{WeakGap}_{VI}(\mathcal{A}, F) \leqslant \mathbb{E}_{\mathbf{S}}[\mathit{EmpGap}_{VI}(\mathcal{A}, F_{\mathbf{S}})] + M\delta. \tag{2.5}$$

2.3 Background on differential privacy

Differential privacy is an algorithmic stability type of guarantee for randomized algorithms, that certifies that the output distribution of the algorithm "does not change too much" by changes in a single element from the dataset. The formal definition is provided in Eq. (1.5). Next we provide some basic results in differential privacy, which we will need for our work. For further information on the topic, we refer the reader to the monograph [16].

2.3.1 Basic privacy guarantees

In this work, most of our privacy guarantees will be obtained by the well-known *Gaussian mechanism*, which performs Gaussian noise addition on a function with bounded sensitivity. Given a function $\mathcal{A} : \mathcal{Z}^n \mapsto \mathbb{R}^d$, we define its ℓ_2-sensitivity as

$$\sup_{\mathbf{S} \simeq \mathbf{S}'} \|\mathcal{A}(\mathbf{S}) - \mathcal{A}(\mathbf{S}')\|. \tag{2.6}$$

If $\mathcal{A}$ is randomized, then the supremum must hold with high-probability over the randomization of $\mathcal{A}$ (this will not be a problem in this work, since our randomized algorithms enjoy sensitivity bounds w.p. 1). The Gaussian mechanism (associated to function $\mathcal{A}$) is defined as $\mathcal{A}_{\mathcal{G}}(\mathbf{S}) \sim \mathcal{N}(\mathcal{A}(\mathbf{S}), \sigma^2 I)$.

Proposition 3 *Let $\mathcal{A} : \mathcal{Z}^n \mapsto \mathbb{R}^d$ be a function with ℓ_2-sensitivity $s > 0$. Then, for $\sigma^2 = 2s^2 \ln(1/\eta)/\varepsilon^2$, the Gaussian mechanism is (ε, η)-DP.*

Our algorithms will adaptively use a DP mechanism such as the above. Certifying privacy of a composition can be achieved in different ways. The most basic result establishes that if we use disjoint batches of data at each iteration, then the composition will preserve the largest privacy parameter among its building blocks. This result is known as *parallel composition*, and its proof is a direct application of the post-processing property of DP.

Proposition 4 (Parallel composition of differential privacy) *Let $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_K) \in \mathcal{Z}^n$ be a dataset partitioned on blocks of sizes $n_1, \ldots, n_K$, respectively. $\mathcal{A}_k : \mathcal{Z}^{n_k} \times \mathbb{R}^{d \times (k-1)} \mapsto \mathbb{R}^d$, $k = 1, \ldots, K$, be a sequence of mechanisms, and let $\mathcal{A} : \mathcal{Z}^n \mapsto \mathbb{R}^d$ be given by*

$$\begin{aligned}
\mathcal{B}_1(\mathbf{S}) &= \mathcal{A}_1(\mathbf{S}_1) \\
\mathcal{B}_k(\mathbf{S}) &= \mathcal{A}_k(\mathbf{S}_k, \mathcal{B}_1(\mathbf{S}), \mathcal{B}_2(\mathbf{S}), \ldots, \mathcal{B}_{k-1}(\mathbf{S})) \quad (\forall k = 2, \ldots, K-1) \\
\mathcal{A}(\mathbf{S}) &= \mathcal{A}_K(\mathbf{S}_K, \mathcal{B}_1(\mathbf{S}), \mathcal{B}_2(\mathbf{S}), \ldots, \mathcal{B}_{K-1}(\mathbf{S})).
\end{aligned}$$

Then, If each $\mathcal{A}_k$ is (ε_k, η_k)-DP in its first argument (i.e., w.r.t. $\mathbf{S}_k$) then $\mathcal{A}$ is $(\max_k \varepsilon_k, \max_k \eta_k)$-DP.

Some of the algorithms we develop in this work make repeated use of the data, and certifying privacy for these algorithms requires the use of adaptive composition results

in DP (see, e.g. [16, 17]). For our algorithms, it is particularly important to leverage the sampling with replacement procedure to select the data that is used at each iteration, for which sharp bounds on DP can be obtained by the *moments accountant method* [1]. Below we summarize a specific version of this method that suffices for our purposes.[4]

Theorem 1 [1] *Consider sequence of functions* $\mathcal{A}_1, \ldots, \mathcal{A}_K$, *where* $\mathcal{A}_k : \mathcal{Z}^{n_k} \times \mathbb{R}^{d\times(k-1)} \mapsto \mathbb{R}^d$ *is a function with sensitivity bounded as a function of the last data batch size, as follows*

$$\sup_{L\in\mathbb{R}^{d\times(k-1)}, \mathbf{S}_k \simeq \mathbf{S}'_k} \|\mathcal{A}_k(\mathbf{S}_k, L) - \mathcal{A}_k(\mathbf{S}'_k, L)\| \leqslant s.$$

Consider the mechanism obtained by sampling a random subset of size m from the dataset, i.e., letting $\mathbf{S}_k \sim (Unif([\mathbf{S}]))^m$, *and composing it with a Gaussian mechanism with noise* σ^2, *i.e.*

$$\begin{aligned}\mathcal{B}_1(\mathbf{S}) &= (\mathcal{A}_1)_{\mathcal{G}}(\mathbf{S}_1)\\ \mathcal{B}_k(\mathbf{S}) &= (\mathcal{A}_k)_{\mathcal{G}}(\mathbf{S}_k, \mathcal{B}_1(\mathbf{S}), \mathcal{B}_2(\mathbf{S}), \ldots, \mathcal{B}_{k-1}(\mathbf{S})) \quad (\forall k = 2, \ldots, K).\end{aligned}$$

There exists an absolute constant $c_1 > 0$, *such that if* $\varepsilon < c_1 K (m/n)^2$ *and the noise parameter* $\sigma \geqslant \sqrt{2K\ln(1/\eta)}sm/[n\varepsilon]$, *then* $\mathcal{A}(\mathbf{S}) := \{\mathcal{B}_1(\mathbf{S}), \ldots, \mathcal{B}_K(\mathbf{S})\}$ *is* (ε, η)*-differentially private.*

3 The noisy stochastic extragradient method

To solve the DP-SVI problem we propose a noisy stochastic extragradient method (NSEG) in Algorithm 1.

Algorithm 1 Noisy Stochastic Extragradient (NSEG) Method

Input: Starting point $u_0 \in \mathcal{W}$, dataset $\mathbf{S} = (\boldsymbol{\beta}_i)_{i\in[n]} \sim \mathcal{P}^n$, stepsizes $(\gamma_t)_{t\in[T]}$
for $t = 1, \ldots, T$ **do**
  $F_{1,t}(\cdot) = F_{\mathbf{B}_t^1}(\cdot) + \boldsymbol{\xi}_t^1$, where $\mathbf{B}_t^1 \subseteq \mathbf{S}$ and $\boldsymbol{\xi}_t^1 \sim \mathcal{N}(0, \sigma_t^2)$
  $w_t = \Pi_{\mathcal{W}}(u_{t-1} - \gamma_t F_{1,t}(u_{t-1}))$
  $F_{2,t}(\cdot) = F_{\mathbf{B}_t^2}(\cdot) + \boldsymbol{\xi}_t^2$, where $\mathbf{B}_t^2 \subseteq \mathbf{S}$ and $\boldsymbol{\xi}_t^1 \sim \mathcal{N}(0, \sigma_t^2)$
  $u_t = \Pi_{\mathcal{W}}(u_{t-1} - \gamma_t F_{2,t}(w_t))$
end for
return $\overline{w}^T = (\sum_{t=1}^T \gamma_t)^{-1} \sum_{t=1}^T \gamma_t w_t$

The name noisy and stochastic in Algorithm 1 is justified by the sequence of operators $F_{1,t}$, $F_{2,t}$ we use:

$$F_{1,t}(\cdot) \triangleq F_{\mathbf{B}_t^1}(\cdot) + \boldsymbol{\xi}_t^1, \quad F_{2,t}(\cdot) \triangleq F_{\mathbf{B}_t^2}(\cdot) + \boldsymbol{\xi}_t^2. \tag{3.1}$$

[4] In our case we use uniform sampling on each iteration, as opposed to the Poisson sampling of [1]; however, it is possible to verify that similar moment estimates lead to our stated result.

where $\mathbf{B}_t^1, \mathbf{B}_t^2$ are batches extracted from dataset $\mathbf{S}$, and $\boldsymbol{\xi}_t^1, \boldsymbol{\xi}_t^2 \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_t^2)$. We will denote the batch size of batch $\mathbf{B}_t^1$ and $\mathbf{B}_t^2$ as $B_t := |\mathbf{B}_t^1| = |\mathbf{B}_t^2|$. The exact details of the sampling method for $\mathbf{B}_t$ will depend on the variant of the algorithm. Here, we detail some key features of the above algorithm. Stochastic extragradient was proposed in [30] where they do not have any noise addition in $F_{1,t}$, $F_{2,t}$ (stochasticity only arises from the dataset randomness), and where disjoint batches are used for all iterations, as well as within iterations. This choice is motivated by the goal of extracting population risk bounds for their algorithm.

Another important consideration is that this algorithm can also be applied to an SSP problem by using as stochastic oracle the monotone operator associated to the stochastic convex-concave function (2.1), over the set $\mathcal{W} = \mathcal{X} \times \mathcal{Y}$. From here onwards, when we say that a certain SVI algorithm is applied to an SSP, we mean using the choices above for the operator and feasible set, respectively.

We start by stating the convergence guarantees for the single-pass NSEG method. This is obtained as a direct corollary of [30, Thm. 1], where we use an explicit bound on the oracle error with the variance of the Gaussian.

Theorem 2 [30] *Consider a stochastic variational inequality (VI(F)), with operators $F_{\boldsymbol{\beta}}$ in $\mathcal{M}^1(L, M)$. Let $\mathcal{A}$ be the NSEG method (Algorithm 1) where $0 < \gamma_t \leqslant 1/[\sqrt{3}L]$ and $(\mathbf{B}_t^1, \mathbf{B}_t^2)_t$ are independent random variables from a product distribution $\mathbf{B}_t^1, \mathbf{B}_t^2 \sim \mathcal{P}^{B_t}$, satisfies*

$$Gap_{VI}(\mathcal{A}, F) \leqslant \frac{K_0(T)}{\Gamma_T},$$

where $K_0(T) \triangleq \left(D^2 + 7\sum_{t\in[T]} \gamma_t^2 [M^2/2 + d\sigma_t^2]\right)$, $\Gamma_T = \sum_{t=1}^T \gamma_t$, $\mathcal{A}(\mathbf{S})$ is the output of Algorithm 1 on the dataset $\mathbf{S} = \bigcup_t \mathbf{B}_t \sim \mathcal{P}^n$ and expectation in the left hand side is taken over the dataset draws, random sample batch choices, as well as noise $\boldsymbol{\xi}_1^t, \boldsymbol{\xi}_2^t$.

On the other hand, $\mathcal{A}$ applied to a stochastic (SP(f)) problem attains saddle point gap

$$Gap_{SP}(\mathcal{A}, f) \leqslant \frac{K_0(T)}{\Gamma_T}.$$

3.1 Differential privacy analysis of NSEG method

We now proceed to establish the privacy guarantees for the single-pass variant of Algorithm 1. This is a direct consequence of Propositions 3 and 4, and the fact that each operator evaluation has sensitivity bounded by $2M/B_t$.

Proposition 5 *Algorithm 1 with batch sizes $(B_t)_{t\in[T]}$ and variance $\sigma_t^2 = \frac{8M^2}{B_t^2}\frac{\ln(1/\eta)}{\varepsilon^2}$ is (ε, η)-differentially private.*

We now apply the previous results to obtain population risk bounds for DP-SVI by the NSEG method.

Corollary 1 *Algorithm 1 with disjoint batches of size $B_t = B = \min\{\sqrt{d(\ln(1/\eta)}/\varepsilon, n\}$, constant stepsize $\gamma_t \equiv \gamma = D/\left[M\sqrt{7T\left(1 + \frac{8d}{B^2}\frac{\ln(1/\eta)}{\varepsilon^2}\right)}\right]$ and variance $\sigma_t^2 = \sigma^2 =$*

$\frac{8M^2}{B^2}\frac{\ln(1/\eta)}{\varepsilon^2}$ *is* (ε, η)*-differentially private and achieves* $Gap_{VI}(\mathcal{A}, F)$ *(for SVI) or* $Gap_{SP}(\mathcal{A}, f)$ *(for SSP) of*

$$O\left(MD\max\left\{\frac{[d\ln(1/\eta)]^{1/4}}{\sqrt{n\varepsilon}}, \frac{\sqrt{d\ln(1/\eta)}}{n\varepsilon}\right\}\right).$$

Remark 1 Notice that in the corollary above, the gap is nontrivial iff $\sqrt{d\ln(1/\eta)}/[n\varepsilon] < 1$, which means that the left hand side attains the max on the range where the gap is nontrivial.

Proof Consider a SVI or SSP problem. Let us recall that by Theorem 2, Algorithm 1 achieves expected gap

$$\frac{D^2}{\gamma T} + 7M^2\gamma\left(1 + \frac{8d}{B^2}\frac{\ln(1/\eta)}{\varepsilon^2}\right).$$

Choosing $\gamma = D/\left[M\sqrt{7T\left(1 + \frac{8d}{B^2}\frac{\ln(1/\eta)}{\varepsilon^2}\right)}\right]$, we obtain an expected gap

$$\frac{2\sqrt{7}MD}{\sqrt{T}}\left(1 + \frac{\sqrt{8d}}{B\varepsilon}\sqrt{\ln(1/\eta)}\right) = \frac{2\sqrt{14}MD\sqrt{B}}{\sqrt{n}} + \frac{8\sqrt{7}MD\sqrt{d}}{\sqrt{nB}\varepsilon}\sqrt{\ln(1/\eta)},$$

where we used that for a single-pass algorithm, $n = 2TB$ (this choice of T exhausts the data when disjoint batches are chosen).

Recalling that $B = \min\{\sqrt{d\ln(1/\eta)}/\varepsilon, n\}$. Then the expected gap is bounded by

$$O\left(MD\max\left\{\frac{[d\ln(1/\eta)]^{1/4}}{\sqrt{n\varepsilon}}, \frac{\sqrt{d\ln(1/\eta)}}{n\varepsilon}\right\}\right).$$

Hence, we conclude the prrof. □

We observe that excess risk bounds of the same order for DP-SCO based on noisy SGD and the uniform stability of differential privacy have been established [7]. Improving these bounds in DP-SCO required substantial efforts, which was only achieved recently [4, 5, 19]. Furthermore, to the best of our knowledge, the upper bounds on the risk above are the first of their type for DP-SVI and DP-SSP, respectively. To improve upon them, we will follow the approach of [4], based on a multi-pass empirical error convergence, combined with weak gap generalization bounds based on uniform stability.

4 Stability of NSEG and optimal risk for DP-SVI and DP-SSP

The bounds established for DP-SVI are potentially suboptimal, and many of the past approaches used to attain optimal rates for DP-SCO, such as privacy amplification by iteration, phased regularization, etc. appear to encounter substantial barriers for their

application to DP-SVI. In order to resolve this gap, we show that for both DP-SVI and DP-SSP we can indeed obtain optimal rates, which match those of DP-SCO. In order to achieve this we develop a *multi-pass* variant of the NSEG method, which enjoys generalization performance due to its stability.

4.1 Stability of NSEG method

To analyze the stability of NSEG it is useful to interpret the extragradient method as an approximation of the proximal point algorithm. This connection has been established at least since [38]. Given a monotone and 1-Lipschitz operator $G : \mathbb{R}^d \mapsto \mathbb{R}^d$, we define the s-extragradient operator inductively as follows. First, $R_0(\cdot; G) : \mathbb{R}^d \mapsto \mathbb{R}^d$ is defined as $R_0(u; G) = \Pi_{\mathcal{W}}(u)$. Then, for $s \geqslant 0$

$$R_{s+1}(u; G) = \Pi_{\mathcal{W}}(u - G(R_s(u; G))). \tag{4.1}$$

Given such operator, the (deterministic) extragradient method [33] corresponds to, starting from $u_0 \in \mathcal{W}$, iterating

$$u_{t+1} = R_2(u_t; \gamma F) \qquad (\forall t \in [T-1]).$$

It is known that if G is contractive, the recursion (4.1) leads to a fixed point $R(u; G)$, satisfying

$$R(u; G) = \Pi_{\mathcal{W}}(u - G(R(u; G))). \tag{4.2}$$

It is also easy to see that $R(\cdot; G) : \mathbb{R}^d \mapsto \mathcal{W}$ is nonexpansive.

Proposition 6 (Near nonexpansiveness of the extragradient operator) *Let $F \in \mathcal{M}^1_{\mathcal{W}}(L, M)$ and $\mathcal{W} \subseteq \mathbb{R}^d$ compact convex set with diameter $D > 0$. Then, for all s nonnegative integer, and $u, v \in \mathbb{R}^d$,*

$$\|R_s(u; \gamma F) - R(u; \gamma F)\| \leqslant (\gamma L)^s \|R_0(u; \gamma F) - R(u; \gamma F)\| \tag{4.3}$$

and

$$\|R_s(u; \gamma F) - R_s(v; \gamma F)\| \leqslant \|u - v\| + 2D(\gamma L)^s. \tag{4.4}$$

Proof The first part, Eq. (4.3), is proved by induction on s. The result clearly holds for $s = 0$, and if $s \geqslant 1$, we use (4.1) and (4.2) to obtain

$$\begin{aligned}
&\|R_s(u; \gamma F) - R(u; \gamma F)\| \\
&\quad = \|\Pi_{\mathcal{W}}(u - \gamma F(R_{s-1}(u; \gamma F)) - \Pi_{\mathcal{W}}(u - \gamma F(R(u; \gamma F)))\| \\
&\quad \leqslant \gamma \|F(R_{s-1}(u; \gamma F)) - F(R(u; \gamma F))\| \ \leqslant \ \gamma L \|R_{s-1}(u; \gamma F) - R(u; \gamma F)\| \\
&\quad \leqslant (\gamma L)^s \|R_0(u; \gamma F) - R(u; \gamma F)\|,
\end{aligned}$$

where in the first inequality we used the nonexpansiveness of the projection operator, next we used the L-Lipschitzness of F, and finally we used the inductive hypothesis to conclude.

The second part, Eq. (4.4), is a direct consequence of (4.3), the triangle inequality, and that $R_0(u;\gamma F), R(u;\gamma F), R_0(v;\gamma F), R(v;\gamma F) \in \mathcal{W}$. □

The next lemma shows an expansion upper bound for extragradient iterations. This type of bound will be later used to establish the uniform argument stability of the NSEG algorithm.

Lemma 1 (Expansion of the extragradient iteration) *Let $F_1, F_2 : \mathbb{R}^d \mapsto \mathbb{R}^d$ monotone L-Lipschitz operators, and $0 \leqslant \gamma < 1/L$. Let $u, v \in \mathcal{W}$, and $w, z, u', v' \in \mathcal{W}$ such that*

$$\begin{aligned} w &= \Pi_{\mathcal{W}}(u - \gamma F_1(u)) \qquad & z &= \Pi_{\mathcal{W}}(v - \gamma F_1(v)) \\ u' &= \Pi_{\mathcal{W}}(u - \gamma F_2(w)) \qquad & v' &= \Pi_{\mathcal{W}}(v - \gamma F_2(z)). \end{aligned}$$

Then,

$$\|w - z\| \leqslant \|u - v\| + 2LD\gamma, \tag{4.5}$$

$$\|u' - v'\| \leqslant \|u - v\| + (\widetilde{M}_1 + \widetilde{M}_2 + 2LD)L\gamma^2, \tag{4.6}$$

where $\widetilde{M}_1 \triangleq \|F_1(R(u;\gamma F_1)) - F_2(R(u;\gamma F_2))\|$ and $\widetilde{M}_2 \triangleq \|F_1(R(v;\gamma F_1)) - F_2(R(v;\gamma F_2))\|$.

Proof By definition of w and z, we have,

$$\begin{aligned} \|w - z\| &= \|\Pi_{\mathcal{W}}(u - \gamma F_1(u)) - \Pi_{\mathcal{W}}(v - \gamma F_1(v))\| \\ &\leqslant \|R(u;\gamma F_1) - R(v;\gamma F_1)\| + \|\Pi_{\mathcal{W}}(u - \gamma F_1(u)) - R(u;\gamma F_1)\| \\ &\quad + \|\Pi_{\mathcal{W}}(v - \gamma F_1(v)) - R(v;\gamma F_1)\| \\ &\leqslant \|u - v\| + \gamma L[\|R_0(u;\gamma F_1) - R(u;\gamma F_1)\| + \|R_0(v;\gamma F_1) - T\gamma F_1 v\|], \end{aligned}$$

where we used the nonexpansiveness of the operator $R(\cdot;\gamma F_1)$ and Proposition 6. Moreover, since $u, v, R_0(u;\gamma F_1), R_0(v;\gamma F_1) \in \mathcal{W}$, we have $\|w - z\| \leqslant \|u - v\| + 2LD\gamma$, proving (4.5).

Next, to prove (4.6), we proceed as follows:

$$\begin{aligned} \|u' - v'\| &= \|\Pi_{\mathcal{W}}(u - \gamma F_2(w)) - \Pi_{\mathcal{W}}(v - \gamma F_2(z))\| \\ &\leqslant \|R(u;\gamma F_2) - R(v;\gamma F_2)\| \\ &\quad + \|\Pi_{\mathcal{W}}(u - \gamma F_2(w)) - \Pi_{\mathcal{W}}(u - \gamma F_2(R(u;\gamma F_2)))\| \\ &\quad + \|\Pi_{\mathcal{W}}(v - \gamma F_2(z)) - \Pi_{\mathcal{W}}(v - \gamma F_2(R(v;\gamma F_2)))\| \\ &\leqslant \|u - v\| + \gamma L\|w - R(u;\gamma F_2)\| + \gamma L\|z - R(v;\gamma F_2)\|. \end{aligned}$$

Using again Proposition 6, we have that

$$\begin{aligned} \|w - R(u;\gamma F_2)\| &= \|R_1(u;\gamma F_1) - R(u;\gamma F_2)\| \\ &\leqslant \|R_1(u;\gamma F_1) - R(u;\gamma F_1)\| + \|R(u;\gamma F_1) - R(u;\gamma F_2)\| \end{aligned}$$

$$
\begin{aligned}
&\leqslant LD\gamma + \|\Pi_{\mathcal{W}}\big(u - \gamma F_1(R(u;\gamma F_1))\big) \\
&\quad -\Pi_{\mathcal{W}}\big(u - \gamma F_2(R(u;\gamma F_2))\big)\| \\
&\leqslant LD\gamma + \gamma\|F_1(R(u;\gamma F_1)) - F_2(R(u;\gamma F_2))\| \\
&\leqslant LD\gamma + \widetilde{M}_1\gamma.
\end{aligned}
$$

An analog bound can be obtained for $\|z - R(v;\gamma F_2)\|$:

$$
\|z - R(v;\gamma F_2)\| \leqslant LD\gamma + \widetilde{M}_2\gamma,
$$

concluding the claimed bound (4.6):

$$
\|u' - v'\| \leqslant \|u - v\| + L[\widetilde{M}_1 + \widetilde{M}_2 + 2LD]\gamma^2.
$$

□

The Expansion Lemma above allows us to bound how much would two trajectories of the NSEG method may deviate, given two pairs of sequences of operators $F_{1,t}$, $F_{2,t}$ and $F'_{1,t}$, $F'_{2,t}$. The bounds we will obtain from this analysis will give us direct bounds on the UAS for the NSEG method.

Lemma 2 *Let $F_{1,t}$, $F_{2,t}$ and $F'_{1,t}$, $F'_{2,t}$ be L-Lipschitz operators, and $0 \leqslant \gamma_t < 1/L$ for all $t \in [T]$. Let $\{(u_t, w_t)\}_{t\in[T]}$ and $\{(v_t, z_t)\}_{t\in[T]}$ be the sequences resulting from Algorithm 1, with operators $\{(F_{1,t}, F_{2,t})\}_{t\in[T]}$ and $\{(F'_{1,t}, F'_{2,t})\}_{t\in[T]}$, respectively; and starting from $u^0 = v^0$. Let*

$$
\begin{aligned}
\Delta_{1,t} &\triangleq \sup_{u\in\mathcal{W}} \|F_{1,t}(u) - F'_{1,t}(u)\|, \\
\Delta_{2,t} &\triangleq \sup_{u\in\mathcal{W}} \|F_{2,t}(u) - F'_{2,t}(u)\|, \\
\widetilde{M}_{1,t} &\triangleq \|F_{1,t}(R(u_{t-1};\gamma F_{1,t})) - F_{2,t}(R(u_{t-1};\gamma F_{2,t}))\|, \text{ and} \\
\widetilde{M}_{2,t} &\triangleq \|F_{1,t}(R(v_{t-1};\gamma F_{1,t})) - F_{2,t}(R(v_{t-1};\gamma F_{2,t}))\|;
\end{aligned}
$$

then, for al $t = 0, \ldots, T$,

$$
v_t \triangleq \|u_t - v_t\| \leqslant \sum_{s=1}^{t} \Big([\widetilde{M}_{1,t} + \widetilde{M}_{2,t} + 2LD]L\gamma_s^2 + L\Delta_{1,s}\gamma_s^2 + \Delta_{2,s}\gamma_s\Big) \tag{4.7}
$$

$$
\begin{aligned}
\delta_t \triangleq \|w_t - z_t\| \leqslant &\sum_{s=1}^{t-1} \Big([\widetilde{M}_{1,t} + \widetilde{M}_{2,t} + 2LD]L\gamma_s^2 + L\Delta_{1,s}\gamma_s^2 + \Delta_{2,s}\gamma_s\Big) \\
&+\Delta_{1,t}\gamma_t + 2LD\gamma_t.
\end{aligned} \tag{4.8}
$$

Proof Clearly, $v_0 = 0$. Let us now derive a recurrence for both v_t and δ_t.

$$
\delta_t = \|R_1(u_{t-1};\gamma_t F_{1,t}) - R_1(v_{t-1};\gamma_t F'_{1,t})\|
$$

$$\leqslant \|R_1(u_{t-1};\gamma_t F_{1,t}) - R_1(v_{t-1};\gamma_t F_{1,t})\| + \|R_1(v_{t-1};\gamma_t F_{1,t}) - R_1(v_{t-1};\gamma_t F'_{1,t})\|$$
$$\leqslant \|u_{t-1} - v_{t-1}\| + 2LD\gamma_t + \|R_1(v_{t-1};\gamma_t F_{1,t}) - R_1(v_{t-1};\gamma_t F'_{1,t})\|,$$

where in the last inequality we used inequality (4.5). Let us bound now the rightmost term above,

$$\begin{aligned}
&\|R_1(v_{t-1};\gamma_t F_{1,t}) - R_1(v_{t-1};\gamma_t F'_{1,t})\| \\
&= \|\Pi_{\mathcal{W}}(v_{t-1} - \gamma_t F_{1,t}(v_{t-1})) - \Pi_{\mathcal{W}}(v_{t-1} - \gamma_t F'_{1,t}(v_{t-1}))\| \\
&\leqslant \gamma_t \|F_{1,t}(v_{t-1}) - F'_{1,t}(v_{t-1})\| \\
&\leqslant \Delta_{1,t}\gamma_t. \qquad (4.9)
\end{aligned}$$

We conclude that

$$\delta_t \leqslant \nu_{t-1} + 2LD\gamma_t + \Delta_{1,t}\gamma_t. \qquad (4.10)$$

Now,

$$\begin{aligned}
\nu_t =& \|u_t - v_t\| \leqslant \|\Pi_{\mathcal{W}}(u_{t-1} - \gamma_t F_{2,t}(w_t)) - \Pi_{\mathcal{W}}(v_{t-1} - \gamma_t F'_{2,t}(z_t))\| \\
=& \|\Pi_{\mathcal{W}}(u_{t-1} - \gamma_t F_{2,t}(w_t)) - \Pi_{\mathcal{W}}(v_{t-1} - \gamma_t F_{2,t}(z_t))\| \\
&+ \|\Pi_{\mathcal{W}}(v_{t-1} - \gamma_t F_{2,t}(z_t)) - \Pi_{\mathcal{W}}(v_{t-1} - \gamma_t F'_{2,t}(z_t))\| \\
\leqslant& \|\Pi_{\mathcal{W}}(u_{t-1} - \gamma_t F_{2,t}(R_1(u_{t-1};\gamma_t F_{1,t}))) - \Pi_{\mathcal{W}}(v_{t-1} - \gamma_t F_{2,t}(R_1(v_{t-1};\gamma_t F'_{1,t})))\| \\
&+ \gamma_t \|F_{2,t}(z_t) - F'_{2,t}(z_t)\| \\
\leqslant& \|\Pi_{\mathcal{W}}(u_{t-1} - \gamma_t F_{2,t}(R_1(u_{t-1};\gamma_t F_{1,t}))) - \Pi_{\mathcal{W}}(v_{t-1} - \gamma_t F_{2,t}(R_1(v_{t-1};\gamma_t F_{1,t})))\| \\
&+ \|\Pi_{\mathcal{W}}(v_{t-1} - \gamma_t F_{2,t}(R_1(v_{t-1};\gamma_t F_{1,t}))) \\
&- \Pi_{\mathcal{W}}(v_{t-1} - \gamma_t F_{2,t}(R_1(v_{t-1};\gamma_t F'_{1,t})))\| + \gamma_t \Delta_{2,t} \\
\overset{\text{(i)}}{\leqslant}& \|u_{t-1} - v_{t-1}\| + [\widetilde{M}_{1,t} + \widetilde{M}_{2,t} + 2LD]L\gamma_t^2 + \gamma_t L\|R_1(v_{t-1};\gamma_t F_{1,t}) \\
&- R_1(v_{t-1};\gamma_t F'_{1,t})\| + \gamma_t \Delta_{2,t} \\
\overset{\text{(ii)}}{\leqslant}& \nu_{t-1} + [\widetilde{M}_{1,t} + \widetilde{M}_{2,t} + 2LD]L\gamma_t^2 + L\Delta_{1,t}\gamma_t^2 + \Delta_{2,t}\gamma_t,
\end{aligned}$$

where in inequality (i), we used Lemma 1 (more precisely, inequality (4.6)), and in inequality (ii), we used (4.9). Unraveling the above recursion, we get that for all $t \in [T]$,

$$\nu_t \leqslant \sum_{s=1}^{t} \Big([\widetilde{M}_{1,t} + \widetilde{M}_{2,t} + 2LD]L\gamma_s^2 + L\Delta_{1,s}\gamma_s^2 + \Delta_{2,s}\gamma_s \Big).$$

Finally, we combine the bound above with (4.10), to conclude that for all $t \in [T]$:

$$\delta_t \leqslant \sum_{s=1}^{t-1} \Big([\widetilde{M}_{1,t} + \widetilde{M}_{2,t} + 2LD]L\gamma_s^2 + L\Delta_{1,s}\gamma_s^2 + \Delta_{2,s}\gamma_s \Big) + \Delta_{1,t}\gamma_t + 2LD\gamma_t.$$

□

The next theorem provides in-expectation and high probability upper bounds for the NSEG method. Despite the fact that we will not particularly apply the latter bounds, we believe these may be of independent interest.

Theorem 3 *The NSEG method (Algorithm 1) for closed and convex domain $\mathcal{W} \subseteq \mathbb{R}^d$ with diameter D, operators in $\mathcal{M}^1_{\mathcal{W}}(L, M)$ and stepsizes $0 < \gamma_t \leqslant 1/L$, satisfies the following uniform argument stability bounds:*

1. *Let $\mathcal{A}_{batch\text{-}EG}$ denote the* Batch method *where given dataset $\mathbf{S}$, $F_{1,t} = F_{\mathbf{S}} + \boldsymbol{\xi}_1^t$, and $F_{2,t} = F_{\mathbf{S}} + \boldsymbol{\xi}_2^t$. Then, in expectation,*

$$\begin{aligned}
&\sup_{\mathbf{S} \simeq \mathbf{S}'} \mathbb{E}_{\mathcal{A}_{batch\text{-}EG}}[\delta_{\mathcal{A}_{batch\text{-}EG}}(\mathbf{S}, \mathbf{S}')] \\
&\quad \leqslant \sum_{t=0}^{T-1} \left([4M + 2LD + 4\sqrt{d}\sigma] L\gamma_t^2 + \frac{2ML}{n}\gamma_t^2 + \frac{2M}{n}\gamma_t \right) \\
&\qquad + \frac{1}{T} \sum_{t=1}^{T} \left(\frac{2ML}{n} + 2LD \right) \gamma_t,
\end{aligned}$$

and for constant stepsize $\gamma_t \equiv \gamma$, there exists a universal constant $K > 0$, such that for any $0 < \theta < 1$, with probability $1 - \theta$:

$$\begin{aligned}
\sup_{\mathbf{S} \simeq \mathbf{S}'} \delta_{\mathcal{A}_{batch\text{-}EG}}(\mathbf{S}, \mathbf{S}') &\leqslant 4[T\sqrt{d}\sigma + \sigma\sqrt{Kd\ln(1/\theta)}]L\gamma^2 + [4M + 2LD]LT\gamma^2 \\
&\quad + \frac{2ML}{n} T\gamma^2 + \frac{2M}{n} T\gamma + \left(\frac{2ML}{n} + 2LD \right) \gamma. \qquad (4.11)
\end{aligned}$$

2. *Let $\mathcal{A}_{repl\text{-}EG}$ denote* Sampled with replacement method *where given dataset $\mathbf{S}$, $F_{1,t} = F_{\boldsymbol{\beta}_{i(1,t)}} + \boldsymbol{\xi}_1^t$, and $F_{2,t} = F_{\boldsymbol{\beta}_{i(2,t)}} + \boldsymbol{\xi}_2^t$, for $i(1,t), i(2,t) \sim$ Unif$([n])$, independently. Then, in expectation,*

$$\begin{aligned}
&\sup_{\mathbf{S} \simeq \mathbf{S}'} \mathbb{E}[\delta_{\mathcal{A}_{repl\text{-}EG}}(\mathbf{S}, \mathbf{S}')] \\
&\quad \leqslant \sum_{t=0}^{T-1} \left([4M + 2LD + 4\sqrt{d}\sigma] L\gamma_t^2 + \frac{2ML}{n}\gamma_t^2 + \frac{2M}{n}\gamma_t \right) \\
&\qquad + \frac{1}{T} \sum_{t=1}^{T} \left(\frac{2ML}{n} + 2LD \right) \gamma_t. \qquad (4.12)
\end{aligned}$$

And for constant stepsize $\gamma_t \equiv \gamma$, there exists a universal constant $K > 0$, such that for any $0 < \theta < 1/[2n]$, with probability $1 - \theta$:

$$\begin{aligned}
\sup_{\mathbf{S} \simeq \mathbf{S}'} \delta_{\mathcal{A}_{repl\text{-}EG}}(\mathbf{S}, \mathbf{S}') &\leqslant 4[T\sqrt{d}\sigma + \sigma\sqrt{Kd\ln(2/\theta)}]L\gamma^2 \\
&\quad + [4M + 2LD]LT\gamma^2 + 2LD\gamma
\end{aligned}$$

$$+\big(1+3\log\big(\frac{2n}{\theta}\big)\big)\frac{2MT}{n}(L\gamma^2+\gamma/T+\gamma). \quad (4.13)$$

Proof Let $\mathbf{S} \simeq \mathbf{S}'$. Then

1. **Batch method.** Notice that for the batch case $F_{1,t} = F_{\mathbf{S}} + \boldsymbol{\xi}_1^t$, and $F'_{1,t} = F_{\mathbf{S}'} + \boldsymbol{\xi}_1^t$; and $F_{2,t} = F_{\mathbf{S}} + \boldsymbol{\xi}_2^t$, and $F'_{2,t} = F_{\mathbf{S}'} + \boldsymbol{\xi}_2^t$. Then, it is easy to see that $\Delta_{1,t} \leqslant 2M/n$ and $\Delta_{2,t} \leqslant 2M/n$. On the other hand, since the operators are M bounded and since noise addition is Gaussian

$$\begin{aligned}\mathbb{E}[\widetilde{M}_{1,t}] &= \mathbb{E}[\|F_{1,t}(R(u_{t-1};\gamma F_{1,t})) - F_{2,t}(R(u_{t-1};\gamma F_{2,t}))\|] \\ &\leqslant \mathbb{E}[\|F_{\mathbf{B}_t^1}(R(u_{t-1};\gamma F_{1,t})) + \boldsymbol{\xi}_1^t\|] + \mathbb{E}[\|F_{\mathbf{B}_t^2}(R(u_{t-1};\gamma F_{2,t})) + \boldsymbol{\xi}_2^t\|] \\ &\leqslant 2M + \mathbb{E}[\|\boldsymbol{\xi}_1^t\| + \|\boldsymbol{\xi}_2^t\|] \leqslant 2[M+\sqrt{d}\sigma], \end{aligned} \quad (4.14)$$

and an analog bound holds for $\mathbb{E}[\widetilde{M}_{2,t}]$. Hence, by Lemma 2:

$$\begin{aligned}\mathbb{E}_{\mathcal{A}_{\text{batch-EG}}}[\delta_{\mathcal{A}_{\text{batch-EG}}}(S,S')] \leqslant \sum_{t=0}^{T-1}\Big([4M+2LD+4\sqrt{d}\sigma]L\gamma_t^2 + \frac{2ML}{n}\gamma_t^2 + \frac{2M}{n}\gamma_t\Big) \\ +\frac{1}{T}\sum_{t=1}^{T}\big(\frac{2ML}{n}+2LD\big)\gamma_t,\end{aligned}$$

which proves the claimed bound.
For the high probability bound, we use that the norm of a Gaussian vector is $Kd\sigma^2$-subgaussian, for a universal constant $K > 0$ (see, e.g. [48, Thm. 3.1.1]), and therefore $\mathbb{E}[\exp\{\lambda(\|\boldsymbol{\xi}_i^t\| - \sigma\sqrt{d})\}] \leqslant \exp\{Kd\sigma^2\lambda^2\}$; hence by the Chernoff-Crámer bound, for any $\alpha > 0$

$$\begin{aligned}\mathbb{P}\left[\sum_{t\in[T]}\big(\|\boldsymbol{\xi}_1^t\| + \|\boldsymbol{\xi}_2^t\|\big) > (2+\alpha)T\sqrt{d}\sigma\right] &\leqslant \exp\{-\lambda\alpha T\sqrt{d}\sigma\}\Big(\exp\{2Kd\sigma^2\lambda^2\}\Big)^T \\ &= \exp\{T(2Kd\sigma^2\lambda^2 - \alpha\sqrt{d}\sigma\lambda)\}.\end{aligned}$$

Choosing $\lambda = \alpha/[4K\sqrt{d}\sigma]$ and $\alpha = \frac{2\sqrt{K}}{T}\sqrt{\ln(1/\theta)}$, we get

$$\mathbb{P}\left[\sum_{t\in[T]}\big(\|\boldsymbol{\xi}_1^t\| + \|\boldsymbol{\xi}_2^t\|\big) > 2T\sqrt{d}\sigma + 2\sigma\sqrt{Kd\ln(1/\theta)}\right] \leqslant \theta. \quad (4.15)$$

This guarantee, together with the rest of the terms appearing in our previous stability bound (which hold w.p. 1) proves (4.11).

2. **Sampled with replacement.** Let $i \in [n]$ be the coordinate where $\mathbf{S}$ and $\mathbf{S}'$ may differ. Let $i(1,t), i(2,t) \sim \text{Unif}([n])$ i.i.d., for $t \in [T]$. Now we apply Lemma 2 with $F_{1,t} = F_{\boldsymbol{\beta}_{i(1,t)}} + \boldsymbol{\xi}_1^t$, and $F'_{1,t} = F_{\boldsymbol{\beta}'_{i(1,t)}} + \boldsymbol{\xi}_1^t$; and $F_{2,t} = F_{\boldsymbol{\beta}_{i(2,t)}} + \boldsymbol{\xi}_2^t$, and $F'_{2,t} = F_{\boldsymbol{\beta}'_{i(2,t)}} + \boldsymbol{\xi}_2^t$. Hence we have that $(\Delta_{1,t})_{t\in[T]}$ and $(\Delta_{2,t})_{t\in[T]}$ are sequences

of independent r.v. with expectation bounded by $2M/n$. Therefore, by Lemma 2 (Eq. (4.8)), and following the steps that lead to inequality (4.14), we have:

$$\mathbb{E}[\delta_{\mathcal{A}}(\mathbf{S},\mathbf{S}')] \leqslant \sum_{t=0}^{T-1}\left([4M+2LD+4\sqrt{d}\sigma]L\gamma_t^2+\frac{2ML}{n}\gamma_t^2+\frac{2M}{n}\gamma_t\right) + \frac{1}{T}\sum_{t=1}^{T}\left(\frac{2ML}{n}+2LD\right)\gamma_t.$$

Finally for the high-probability bound, note that for any realization of the algorithm randomness, we have

$$\delta_{\mathcal{A}}(\mathbf{S},\mathbf{S}') \leqslant \sum_{t=1}^{T}[4M+2(\|\boldsymbol{\xi}_1^t\|+\|\boldsymbol{\xi}_2^t\|)+2LD]L\gamma_t^2 + \frac{2LD}{T}\sum_{t=1}^{T}\gamma_t + L\sum_{t=1}^{T}\gamma_t^2\Delta_{1,t}+\frac{1}{T}\sum_{t=1}^{T}\Delta_{1,t}\gamma_t+\sum_{t=1}^{T}\Delta_{2,t}\gamma_t.$$

We additionally assume constant stepsize, $\gamma_t \equiv \gamma > 0$. Hence, we can resort on concentration of sums of Bernoulli random variables, which guarantees that

$$\mathbb{P}\left[\sum_{t=1}^{T}\Delta_{1,t} > (1+3\log(2/\theta))\frac{2MT}{n}\right] \leqslant \exp\{-\log(2/\theta)\} = \frac{\theta}{2}.$$

An analog bound can be established for $\Delta_{2,t}$, which together with bound (4.15) leads to

$$\mathbb{P}_{\mathcal{A}_{\text{repl-EG}}}\Big[\delta_{\mathcal{A}_{\text{repl-EG}}}(\mathbf{S},\mathbf{S}') > 4[T\sqrt{d}\sigma+\sigma\sqrt{Kd\ln(1/\theta)}]L\gamma^2 + [4M+2LD]LT\gamma^2+2LD\gamma + \big(1+3\log\big(\frac{2}{\theta}\big)\big)\frac{2MT}{n}(L\gamma^2+\gamma/T+\gamma)\Big] \leqslant 2\theta.$$

Notice this bound only depends on our choice of i, and it is otherwise uniform over all $\mathbf{S} \simeq \mathbf{S}'$. Finally, by a union bound on $i \in [n]$ (together with a renormalization of θ), we have that

$$\mathbb{P}_{\mathcal{A}_{\text{repl-EG}}}\Big[\sup_{\mathbf{S}\simeq\mathbf{S}'}\delta_{\mathcal{A}_{\text{repl-EG}}}(\mathbf{S},\mathbf{S}') > 4[T\sqrt{d}\sigma+\sigma\sqrt{Kd\ln(2/\theta)}]L\gamma^2 + [4M+2LD]LT\gamma^2+2LD\gamma + \big(1+3\log\big(\frac{4n}{\theta}\big)\big)\frac{2MT}{n}(L\gamma^2+\gamma/T+\gamma)\Big] \leqslant \theta.$$

□

4.2 Optimal risk for DP-SVI and DP-SSP by NSEG method

Now we use our stability and risk bounds for NSEG to derive optimal risk bounds for DP-SSP. For this, we use the sampled with replacement variant, $\mathcal{A}_{\text{repl-EG}}$.

$$F_{1,t}(\cdot) = F_{\boldsymbol{\beta}_{i(1,t)}}(\cdot) + \boldsymbol{\xi}_1^t; \qquad F_{2,t}(\cdot) = F_{\boldsymbol{\beta}_{i(2,t)}}(\cdot) + \boldsymbol{\xi}_2^t. \tag{4.16}$$

Using the moments accountant method (Theorem 1) one can show the following.

Proposition 7 (Privacy of sampled with replacement NSEG) *Algorithm 1 with operators given by Eq. (4.16) and $\sigma_t^2 = 8M^2\log(1/\eta)/\varepsilon^2$, is (ε, η)-differentially private.*

Theorem 4 (Excess risk of sampled with replacement NSEG) *Consider an instance of the (VI(F)) or (SP(f)) problem. Let $\mathcal{A}$ be the sampled with replacement variant (4.16) of NSEG method (Algorithm 1), with $\gamma_t = \gamma = \min\{D/M, 1/L\}/[n\max\{\sqrt{n}, \sqrt{d\ln(1/\eta)}/\varepsilon\}]$, $\sigma_t^2 = 8M^2\log(1/\eta)/\varepsilon^2$, $T = n^2$. Then, $\text{WeakGap}_{VI}(\mathcal{A}, F)$ (for SVI) or $\text{WeakGap}_{SP}(\mathcal{A}, f)$ (for SSP) are bounded by*

$$O\Big((MD + LD^2)\max\Big\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d\ln(1/\eta)}}{n\varepsilon}\Big\} + \frac{MLD}{n^{5/2}}\Big).$$

Remark 2 Notice that assuming $n = \Omega(\min\{\sqrt{L}, \sqrt{M/D}\})$ the bound of the Theorem simplifies to

$$O\left((MD + LD^2)\max\left\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d\ln(1/\eta)}}{n\varepsilon}\right\}\right).$$

This is quite a mild sample size requirement. In this range, when $M \geqslant LD$, our upper bound matches the excess risk bounds for DP-SCO [5], and we will show these rates are indeed optimal for DP-SVI and DP-SSP as well

Proof Given that our bounds for SVI and SSP are analogous, we proceed indistinctively for both problems.

First, let us bound the empirical accuracy of the method. By Theorem 2, together with the fact that sampling with replacement is an unbiased stochastic oracle for the empirical operator:

$$\begin{aligned}
\mathbb{E}_{\mathbf{S}}\Big[\text{EmpGap}(\mathcal{A}, \mathbf{S})\Big] &\leqslant \frac{1}{\gamma T}\left(\frac{D^2}{2} + 7M^2T\gamma^2\left(1 + \frac{8d\log(1/\eta)}{\varepsilon^2}\right)\right) \\
&\leqslant \frac{D^2}{2n}\max\{M/D, L\}\max\{\sqrt{n}, \sqrt{d\ln(1/\eta)}/\varepsilon\} \\
&\quad + \frac{7M^2\min\{D/M, 1/L\}}{n\max\{\sqrt{n}, \sqrt{d\ln(1/\eta)}/\varepsilon\}}\frac{9d\ln(1/\eta)}{\varepsilon^2} \\
&= O\left((MD + LD^2)\max\left\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d\ln(1/\eta)}}{n\varepsilon}\right\}\right),
\end{aligned}$$

where $\mathrm{EmpGap}(\mathcal{A}, \mathbf{S})$ is $\mathrm{EmpGap}_{\mathrm{VI}}(\mathcal{A}, F_{\mathbf{S}})$ or $\mathrm{EmpGap}_{\mathrm{SP}}(\mathcal{A}, f_{\mathbf{S}})$ for an SVI or SSP problem, respectively.

Next, by Theorem 3, we have that $\mathcal{A}$ (or $x(\mathbf{S})$ and $y(\mathbf{S})$, for the SSP case) are UAS with parameter

$$\begin{aligned}\delta &= [4M + 2LD + 4\sqrt{d}\sigma]LT\gamma^2 + \frac{2ML}{n}T\gamma^2 + \frac{2M}{n}T\gamma + \left(\frac{2ML}{n} + 2LD\right)\gamma \\ &\leqslant \left(\frac{4LD^2}{M} + 2D\right)\frac{1}{n} + \frac{8LD^2\sqrt{2d\ln(1/\eta)}}{M\varepsilon n} + \frac{2LD^2}{M}\frac{1}{n^{3/2}} + \frac{2D}{\sqrt{n}} + \frac{2LD}{n^{5/2}} + \frac{2D}{n^{3/2}} \\ &= O\Big(\frac{1}{M}\cdot\Big(\frac{MD + LD^2}{n} + \frac{LD^2\sqrt{d\ln(1/\eta)}}{\varepsilon n} + \frac{MLD}{n^{5/2}}\Big)\Big).\end{aligned}$$

Hence, noting that empirical risk upper bounds weak empirical gap and using Proposition 1 or Proposition 2 (depending on whether the problem is an SSP or SVI, respectively), we have that the risk is upper bounded by its empirical risk plus $M\delta$, where δ is the UAS parameter of the algorithm; in particular, is bounded by

$$\begin{aligned}\mathrm{WeakGap}_{\mathrm{VI}}(\mathcal{A}, F) &\leqslant \mathbb{E}_{\mathbf{S}}[\mathrm{EmpGap}_{\mathrm{VI}}(\mathcal{A}, F_{\mathbf{S}})] + M\delta \\ &= O\Big((MD + LD^2)\max\Big\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d\ln(1/\eta)}}{n\varepsilon}\Big\} + \frac{MLD}{n^{5/2}}\Big),\end{aligned}$$

Similar claims can be made $\mathrm{WeakGap}_{\mathrm{SP}}(\mathcal{A}, f_{\mathbf{S}})$. Hence, we conclude the proof. □

5 The noisy inexact stochastic proximal point method

In the previous sections, we presented NSEG method with its single-pass and multipass variants and provided optimal risk guarantees for DP-SVI and DP-SSP problems in $O(n^2)$ stochastic operator evaluations. In the rest of the paper, our aim is to provide another algorithm that can achieve the optimal risk for both of these problems with much less computational effort. Towards that end, consider the following algorithm:

Algorithm 2 Noisy Inexact Stochastic Proximal Point (NISPP) Method

Input: $w_0 \in \mathcal{W}$
for $k = 0, 1, \ldots, K$ **do**
  Sample a batch $\mathbf{B}_{k+1} \subseteq \mathbf{S}$.
  $u_{k+1} \leftarrow \mathrm{VI}_v(\mathcal{W}, F_{\mathbf{B}_{k+1}}(\cdot) + \lambda_k(\cdot - w_k))$.
  $w_{k+1} \leftarrow u_{k+1} + \boldsymbol{\xi}_{k+1}$, where $\boldsymbol{\xi}_{k+1} \sim \mathcal{N}(0, \sigma_{k+1}^2\mathbb{I}_d)$
end for
$\overline{w}_K := (\sum_{k=0}^K \gamma_k w_{k+1})/(\sum_{k=0}^K \gamma_k)$
Output: $\Pi_{\mathcal{W}}(\overline{w}_K)$

In the above algorithm, we leave a few things unspecified which will be stated later during convergence and privacy analysis. Here, we detail some key features of the

above algorithm. In line 3, we sample a batch $\mathbf{B}_{k+1}$ of size $B_{k+1} = |\mathbf{B}_{k+1}|$. Similar to the NSEG, we will look at two different variants of NISPP method: single-pass and multi-pass. Depending on the type of the method, we will specify the sampling mechanism. In line 4 of Algorithm 2, we have u_{k+1} is a ν-approximate strong VI solution of the mentioned VI problem for some $\nu \geqslant 0$, i.e.,

$$\langle F_{\mathbf{B}_{k+1}}(u_{k+1}) + \lambda_k(u_{k+1} - w_k), w - u_{k+1}\rangle \geqslant -\nu \qquad (\forall w \in \mathcal{W}). \tag{5.1}$$

Note that if $\nu = 0$ then this is an exact solution satisfying (VI(F)) with operator F replaced by $F(\cdot) + \lambda_k(\cdot - w_k)$. For $\nu > 0$, we obtain that u_{k+1} is an inexact solution satisfying solution criterion up to ν additive error. In line 5, we add a Gaussian noise to u_{k+1} in order to preserve privacy. The resulting iterate w_{k+1} can be potentially outside the set $\mathcal{W}$. Hence, in line 7, the ergodic average $\overline{w}_K$ can be outside $\mathcal{W}$. In order to preserve feasibility of the solution, we project $\overline{w}_K$ onto set $\mathcal{W}$ and output it as a solution in line 8. Projection of the average in line 8, as opposed to projection individual w_{k+1} in line 5 is crucial for convergence guarantee of Algorithm 2.

In the rest of this section, we exclusively deal with the single-pass version of NISPP method, i.e., we assume that batches $\{\mathbf{B}_{k+1}\}_{k=0,\dots,K}$ are disjoint subsets of the dataset $\mathbf{S}$. We start with the convergence guarantees of single-pass NISPP method. In order to prove convergence, we show a useful bound on $\mathrm{dist}_{\mathcal{W}}(\overline{w}_K) := \min_{w\in\mathcal{W}}\|w - \overline{w}_K\|$.

Proposition 8 *Let* $\overline{u}_K := \frac{1}{\Gamma_K}\sum_{k=0}^K \gamma_k u_{k+1}$. *Then,*

$$\mathrm{dist}_{\mathcal{W}}(\overline{w}_K)^2 \leqslant \|\overline{u}_K - \overline{w}_K\|^2 = \frac{1}{\Gamma_K^2}\|\textstyle\sum_{k=0}^K \gamma_k \boldsymbol{\xi}_{k+1}\|^2. \tag{5.2}$$

Moreover, we have

$$\mathbb{E}[\mathrm{dist}_{\mathcal{W}}(\overline{w}_K)^2] \leqslant \frac{1}{\Gamma_K^2}\textstyle\sum_{k=0}^K \gamma_k^2 \mathbb{E}\|\boldsymbol{\xi}_{k+1}\|^2 \tag{5.3}$$

Proof Note that $\overline{u}_K \in \mathcal{W}$. Hence, first relation in (5.2) follows by definition of $\mathrm{dist}_{\mathcal{W}}(\cdot)$ function. Equality follows from definition of $\overline{u}_K$ and $\overline{w}_K$. To obtain (5.3), note that $\{\boldsymbol{\xi}_k\}_{k=1}^{K+1}$ are i.i.d. random variable with mean 0. Expanding $\|\sum_{k=0}^K \gamma_k \boldsymbol{\xi}_{k+1}\|^2$, using linearity of expectation and noting $\mathbb{E}[\boldsymbol{\xi}_i^T \boldsymbol{\xi}_j] = 0$ for all $i \neq j$, we conclude (5.3). Hence, we conclude the proof. □

We prove the following convergence rate result for Algorithm 2 for the risk of SVI/SSP problem. In particular, we assume that the algorithm performs a single-pass over the dataset $\mathbf{S} \sim \mathcal{P}^n$ containing n i.i.d. datapoints.

Theorem 5 *Consider the stochastic (VI(F)) problem with operators* $F_{\boldsymbol{\beta}} \in \mathcal{M}^1_{\mathcal{W}}(L, M)$. *Let* $\mathcal{A}$ *be the single-pass NISPP method (Algorithm 2) where sequence* $\{\gamma_k\}_{k\geqslant 0}$, $\{\lambda_k\}_{k\geqslant 0}$ *satisfy*

$$\gamma_k \lambda_k = \gamma_0 \lambda_0 \tag{5.4}$$

for all $k \geqslant 0$. *Moreover,* $\mathbf{B}_{k+1}$ *are independent samples from a product distribution* $\mathbf{B}_{k+1} \sim \mathcal{P}^{B_{k+1}}$ *and* $\mathbf{B}_{k+1} \subset \mathbf{S}$. *Then, we have*

$$Gap_{VI}(\mathcal{A}, F) \leqslant \nu + \frac{Z_0(K)}{\Gamma_K} + M\sqrt{\frac{1}{\Gamma_K^2}\sum_{k=0}^K \gamma_k^2 \sigma_{k+1}^2 d}, \tag{5.5}$$

where, $Z_0(K) := \frac{3\gamma_0\lambda_0}{2}D^2 + \frac{4M^2+3L^2D^2}{\gamma_0\lambda_0}\sum_{k=0}^K \gamma_k^2 + \frac{5\gamma_0\lambda_0 d}{2}\sum_{k=1}^K \sigma_{k+1}^2$ *and* $\Gamma_K := \sum_{k=0}^K \gamma_k$.

Similarly, $\mathcal{A}$ *applied to stochastic* (SP(f)) *problem achieves*

$$Gap_{SP}(\mathcal{A}, f) \leqslant \nu + \frac{Z_0(K)}{\Gamma_K} + M\sqrt{\frac{1}{\Gamma_K^2}\sum_{k=0}^K \gamma_k^2 \sigma_{k+1}^2 d}. \tag{5.6}$$

Proof Let $w \in \mathcal{W}$. Then

$$\langle F(w), w_{k+1} - w\rangle = \langle F(w), u_{k+1} - w\rangle + \langle F(w), w_{k+1} - u_{k+1}\rangle. \tag{5.7}$$

We will analyze each term above separately. First, note that

$$\langle F(w), w_{k+1} - u_{k+1}\rangle \leqslant \frac{1}{2\lambda_k}\|F(w)\|^2 + \frac{\lambda_k}{2}\|\boldsymbol{\xi}_{k+1}\|^2 \leqslant \frac{M^2}{2\lambda_k} + \frac{\lambda_k}{2}\|\boldsymbol{\xi}_{k+1}\|^2. \tag{5.8}$$

Note that

$$\begin{aligned}
&\langle F(w), u_{k+1} - w\rangle \\
&\quad\leqslant \langle F(u_{k+1}), u_{k+1} - w\rangle \\
&\quad= \langle F_{\mathbf{B}_{k+1}}(u_{k+1}), u_{k+1} - w\rangle + \langle F(u_{k+1}) - F_{\mathbf{B}_{k+1}}(u_{k+1}), u_{k+1} - w\rangle \\
&\quad\leqslant \lambda_k\langle u_{k+1} - w_k, w - u_{k+1}\rangle + \nu + \langle F(u_{k+1}) - F_{\mathbf{B}_{k+1}}(u_{k+1}), u_{k+1} - w\rangle \\
&\quad= \frac{\lambda_k}{2}\left[\|w - w_k\|^2 - \|w - u_{k+1}\|^2 - \|u_{k+1} - w_k\|^2\right] + \nu \\
&\qquad + \langle F(u_{k+1}) - F_{\mathbf{B}_{k+1}}(u_{k+1}), u_{k+1} - w\rangle,
\end{aligned} \tag{5.9}$$

where first inequality follows from monotonicity and second inequality follows from (5.1). Now note that

$$\begin{aligned}
&\langle F(u_{k+1}) - F_{\mathbf{B}_{k+1}}(u_{k+1}), u_{k+1} - w\rangle \\
&\quad= \langle F(u_{k+1}) - F_{\mathbf{B}_{k+1}}(u_{k+1}) - [F(w_k) - F_{\mathbf{B}_{k+1}}(w_k)], u_{k+1} - w\rangle \\
&\qquad + \langle F(w_k) - F_{\mathbf{B}_{k+1}}(w_k), u_{k+1} - w_k\rangle + \langle F(w_k) - F_{\mathbf{B}_{k+1}}(w_k), w_k - w\rangle \\
&\quad\leqslant \frac{\lambda_k}{6L^2}\|F(u_{k+1}) - F(w_k)\|^2 + \frac{3L^2}{2\lambda_k}\|u_{k+1} - w\|^2 \\
&\qquad + \frac{\lambda_k}{6L^2}\|F_{\mathbf{B}_{k+1}}(u_{k+1}) - F_{\mathbf{B}_{k+1}}(w_k)\|^2 + \frac{3L^2}{2\lambda_k}\|u_{k+1} - w\|^2
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{3}{2\lambda_k}\|F(w_k) - F_{\mathbf{B}_{k+1}}(w_k)\|^2 \\
&+ \frac{\lambda_k}{6}\|u_{k+1} - w_k\|^2 + \langle F(w_k) - F_{\mathbf{B}_{k+1}}(w_k), w_k - w\rangle \\
&\leqslant \frac{\lambda_k}{2}\|u_{k+1} - w_k\|^2 + \frac{3L^2}{\lambda_k}\|u_{k+1} - w\|^2 + \frac{3}{2\lambda_k}\|F(w_k) - F_{\mathbf{B}_{k+1}}(w_k)\|^2 \\
&+ \langle F(w_k) - F_{\mathbf{B}_{k+1}}(w_k), w_k - w\rangle,
\end{aligned}
$$

where last inequality follows from L-Lipschitz continuity of F and $F_{\mathbf{B}_{k+1}}$. Noting that $\|u_{k+1} - w\| \leqslant D$ for all $w \in \mathcal{W}$ and using the above bound in (5.9), we have

$$
\begin{aligned}
&\langle F(w), u_{k+1} - w\rangle \\
&\quad\leqslant \langle F(u_{k+1}), u_{k+1} - w\rangle \\
&\quad\leqslant \frac{\lambda_k}{2}\left[\|w - w_k\|^2 - \|w - u_{k+1}\|^2\right] \\
&\quad+ \underbrace{\nu + \frac{3L^2D^2}{\lambda_k} + \frac{3}{2\lambda_k}\|F(w_k) - F_{\mathbf{B}_{k+1}}(w_k)\|^2 + \langle F(w_k) - F_{\mathbf{B}_{k+1}}(w_k), w_k - w\rangle}_{E_k}
\end{aligned}
\tag{5.10}
$$

Letting $u_0 := w_0$ and consequently $\boldsymbol{\xi}_0 = \mathbf{0}$, we have from (5.10)

$$
\begin{aligned}
&\langle F(w), u_{k+1} - w\rangle \\
&\quad\leqslant \langle F(u_{k+1}), u_{k+1} - w\rangle \\
&\quad\leqslant \frac{\lambda_k}{2}\left[\|w - u_k\|^2 - \|w - u_{k+1}\|^2 + 2\langle w - u_k, u_k - w_k\rangle + \|u_k - w_k\|^2\right] + E_k \\
&\quad= \frac{\lambda_k}{2}\left[\|w - u_k\|^2 - \|w - u_{k+1}\|^2\right] + \lambda_k\langle \boldsymbol{\xi}_k, u_k - w\rangle + \frac{1}{2}\lambda_k\|\boldsymbol{\xi}_k\|^2 + E_k
\end{aligned}
\tag{5.11}
$$

Let us define an auxiliary sequence $\{z_k\}_{k\geqslant 0}$, where $z_0 = w_0$ and for all $k \geqslant 1$, we have

$$
z_k = \Pi_{\mathcal{W}}[z_{k-1} - \boldsymbol{\xi}_k].
$$

Then, due to the mirror-descent bound, we have

$$
\sum_{k=1}^K \langle \boldsymbol{\xi}_k, z_k - w\rangle \leqslant \tfrac{1}{2}\|w - w_0\|^2 - \tfrac{1}{2}\|w - z_K\|^2 + \sum_{k=1}^K \|\boldsymbol{\xi}_k\|^2. \tag{5.12}
$$

Moreover, noting that

$$
\begin{aligned}
\langle \boldsymbol{\xi}_k, u_k - z_k\rangle &= \langle \boldsymbol{\xi}_k, u_k - z_{k-1}\rangle + \langle \boldsymbol{\xi}_k, z_{k-1} - z_k\rangle \leqslant \langle \boldsymbol{\xi}_k, u_k - z_{k-1}\rangle + \|\boldsymbol{\xi}_k\|\|z_k - z_{k-1}\| \\
&\leqslant \langle \boldsymbol{\xi}_k, u_k - z_{k-1}\rangle + \|\boldsymbol{\xi}_k\|^2.
\end{aligned}
$$

Combining above relation with (5.12), we have

$$\sum_{k=1}^{K}\langle \boldsymbol{\xi}_k, u_k - w\rangle \leqslant \tfrac{1}{2}\|w - w_0\|^2 + 2\sum_{k=1}^{K}\|\boldsymbol{\xi}_k\|^2 + \sum_{k=1}^{K}\langle \boldsymbol{\xi}_k, u_k - z_{k-1}\rangle. \quad (5.13)$$

Now multiplying (5.7) by γ_k then summing from $k = 0$ to K; noting the definition of $\overline{w}_K$ and Γ_K; and using (5.8), (5.11) and (5.13) along with assumption (5.4) implies

$$\begin{aligned}
&\Gamma_k\langle F(w), \overline{w}_K - w\rangle \\
&\quad\leqslant \gamma_0\lambda_0\|w - w_0\|^2 \\
&\quad+ \sum_{k=0}^{K}\left[\tfrac{\gamma_k M^2}{2\lambda_k} + \tfrac{\gamma_0\lambda_0}{2}\|\boldsymbol{\xi}_{k+1}\|^2 + \tfrac{5\gamma_0\lambda_0}{2}\|\boldsymbol{\xi}_k\|^2 + \gamma_k E_k + \gamma_0\lambda_0\langle \boldsymbol{\xi}_k, u_k - z_{k-1}\rangle\right] \\
&\quad\Rightarrow \Gamma_k\langle F(w), \Pi_{\mathcal{W}}[\overline{w}_K] - w\rangle \\
&\quad\leqslant \gamma_0\lambda_0\|w - w_0\|^2 + \Gamma_K\langle F(w), \Pi_{\mathcal{W}}[\overline{w}_K] - \overline{w}_K\rangle \\
&\quad+ \sum_{k=0}^{K}\left[\tfrac{\gamma_k M^2}{2\lambda_k} + \tfrac{\gamma_0\lambda_0}{2}\|\boldsymbol{\xi}_{k+1}\|^2 + \tfrac{5\gamma_0\lambda_0}{2}\|\boldsymbol{\xi}_k\|^2 + \gamma_k E_k + \gamma_0\lambda_0\langle \boldsymbol{\xi}_k, u_k - z_{k-1}\rangle\right] \\
&\quad\leqslant \gamma_0\lambda_0\|w - w_0\|^2 + \Gamma_K M\,\mathrm{dist}_{\mathcal{W}}(\overline{w}_K) \\
&\quad+ \sum_{k=0}^{K}\left[\tfrac{\gamma_k M^2}{2\lambda_k} + \tfrac{\gamma_0\lambda_0}{2}\|\boldsymbol{\xi}_{k+1}\|^2 + \tfrac{5\gamma_0\lambda_0}{2}\|\boldsymbol{\xi}_k\|^2 + \gamma_k E_k + \gamma_0\lambda_0\langle \boldsymbol{\xi}_k, u_k - z_{k-1}\rangle\right]
\end{aligned} \quad (5.14)$$

Now note that

$$\begin{aligned}
\sum_{k=0}^{K}\gamma_k E_k = \Gamma_K\nu + \frac{3L^2D^2}{\gamma_0\lambda_0}\sum_{k=0}^{K}\gamma_k^2 + \tfrac{3}{2\gamma_0\lambda_0}\sum_{k=0}^{K}\gamma_k^2\|F_{\mathbf{B}_{k+1}}(w_k) - F(w_k)\|^2 \\
+ \gamma_0\lambda_0\sum_{k=0}^{K}\langle \tfrac{1}{\lambda_k}(F(w_k) - F_{\mathbf{B}_{k+1}}(w_k)), w_k - w\rangle
\end{aligned} \quad (5.15)$$

Define $\boldsymbol{\Delta}_k := \frac{1}{\lambda_k}(F(w_k) - F_{\mathbf{B}_{k+1}}(w_k))$. Note that $\mathbb{E}_{\mathbf{B}_{k+1}}[\boldsymbol{\Delta}_k|w_k] = 0$. Moreover, define an auxiliary sequence $\{h_k\}_{k\geqslant 0}$ with $h_0 := w_0$ and

$$h_{k+1} := \Pi_{\mathcal{W}}[h_k - \boldsymbol{\Delta}_k].$$

Then due to mirror descent bound, we have

$$\sum_{k=0}^{K}\langle \boldsymbol{\Delta}_k, h_{k+1} - w\rangle \leqslant \tfrac{1}{2}\|w - w_0\|^2 - \tfrac{1}{2}\|w - h_{K+1}\|^2 + \sum_{k=0}^{K}\|\boldsymbol{\Delta}_k\|^2. \quad (5.16)$$

Moreover,

$$\langle \boldsymbol{\Delta}_k, h_k - h_{k+1}\rangle \leqslant \|\boldsymbol{\Delta}_k\|\|h_k - h_{k+1}\| \leqslant \|\boldsymbol{\Delta}_k\|^2.$$

Using above relation along with (5.16), we have

$$\begin{aligned}
&\sum_{k=0}^{K}\langle \boldsymbol{\Delta}_k, h_k - w\rangle \\
&\quad= \sum_{k=0}^{K}[\langle \boldsymbol{\Delta}_k, h_{k+1} - w\rangle + \langle \boldsymbol{\Delta}_k, h_k - h_{k+1}\rangle]
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \frac{1}{2}\|w-w_0\|^2 - \frac{1}{2}\|w-h_{K+1}\|^2 + 2\textstyle\sum_{k=0}^K \|\boldsymbol{\Delta}_k\|^2 \\
&\Rightarrow \textstyle\sum_{k=0}^K \langle \boldsymbol{\Delta}_k, w_k - w\rangle \\
&= \textstyle\sum_{k=0}^K [\langle \boldsymbol{\Delta}_k, w_k - h_k\rangle + \langle \boldsymbol{\Delta}_k, h_k - w\rangle] \\
&\leqslant \textstyle\sum_{k=0}^K \langle \boldsymbol{\Delta}_k, w_k - h_k\rangle + \frac{1}{2}\|w-w_0\|^2 - \frac{1}{2}\|w-h_{K+1}\|^2 + 2\textstyle\sum_{k=0}^K \|\boldsymbol{\Delta}_k\|^2.
\end{aligned}
$$

Using the above relation in (5.15), we have

$$
\begin{aligned}
\textstyle\sum_{k=0}^K \gamma_k E_k &= \Gamma_K \nu + \frac{3L^2D^2}{\gamma_0\lambda_0}\textstyle\sum_{k=0}^K \gamma_k^2 + \frac{3}{2\gamma_0\lambda_0}\textstyle\sum_{k=0}^K \gamma_k^2 \|F_{\mathbf{B}_{k+1}}(w_k) - F(w_k)\|^2 \\
&\quad + \gamma_0\lambda_0 \textstyle\sum_{k=0}^K \langle \boldsymbol{\Delta}_k, w_k - w\rangle \\
&\leqslant \frac{\gamma_0\lambda_0}{2}\|w-w_0\|^2 + \Gamma_K \nu \frac{3L^2D^2}{\gamma_0\lambda_0}\textstyle\sum_{k=0}^K \gamma_k^2 \\
&\quad + \frac{3}{2\gamma_0\lambda_0}\textstyle\sum_{k=0}^K \gamma_k^2 \|F_{\mathbf{B}_{k+1}}(w_k) - F(w_k)\|^2 \\
&\quad + 2\gamma_0\lambda_0 \textstyle\sum_{k=0}^K \|\boldsymbol{\Delta}_k\|^2 + \gamma_0\lambda_0 \textstyle\sum_{k=0}^K \langle \boldsymbol{\Delta}_k, w_k - h_k\rangle. \qquad (5.17)
\end{aligned}
$$

Finally, note that for all valid k, we have

$$
\begin{aligned}
\mathbb{E}[\|\boldsymbol{\xi}_k\|^2] &= \sigma_k^2 d, \qquad (5.18)\\
\mathbb{E}[\|\boldsymbol{\Delta}_k\|^2] &= \mathbb{E}_{w_k}[\mathbb{E}_{\mathbf{B}_{k+1}}[\|\boldsymbol{\Delta}_k\|^2 | w_k]] \\
&= \frac{1}{\lambda_k^2}\mathbb{E}_{w_k}\mathbb{E}_{\mathbf{B}_{k+1}}\left[\|F_{\mathbf{B}_{k+1}}(w_k) - F(w_k)\|^2 | w_k\right] \\
&\leqslant \frac{1}{\lambda_k^2}\mathbb{E}_{w_k}\mathbb{E}_{\mathbf{B}_{k+1}}\|F_{\mathbf{B}_{k+1}}(w_k)\|^2 \leqslant \frac{M^2}{\lambda_k^2}, \qquad (5.19)\\
\mathbb{E}[\langle \boldsymbol{\Delta}_k, w_k - h_k\rangle] &= \mathbb{E}[\langle \mathbb{E}[\boldsymbol{\Delta}_k | w_k, h_k], w_k - h_k\rangle] = 0, \qquad (5.20)\\
\mathbb{E}[\langle \boldsymbol{\xi}_k, u_k - z_{k-1}\rangle] &= \mathbb{E}[\langle \mathbb{E}[\boldsymbol{\xi}_k | u_k, z_{k-1}], u_k - z_{k-1}\rangle] = 0, \qquad (5.21)
\end{aligned}
$$

where, in (5.19), we used the fact that $F_{\boldsymbol{\beta}})$ is M-bounded for all $\boldsymbol{\beta} \in \mathbf{S}$.

Now, using (5.17) in relation (5.14), noting the bound on $\mathrm{dist}_{\mathcal{W}}(\overline{w}_K)$ from Proposition 8 (in particular (5.3)), taking supremum with respect to $w \in \mathcal{W}$, then taking expectation and noting (5.18)-(5.21), we have

$$
\begin{aligned}
&\Gamma_k \mathbb{E} \sup_{w\in\mathcal{W}} \langle F(w), \Pi_{\mathcal{W}}[\overline{w}_K] - w\rangle \\
&\leqslant \frac{3\gamma_0\lambda_0}{2}D^2 + \textstyle\sum_{k=0}^K \left[\frac{\gamma_k(4M^2+3L^2D^2)}{\lambda_k} + \frac{5\gamma_0\lambda_0}{2}\sigma_{k+1}^2 d + \gamma_k \nu\right] \\
&\quad + \Gamma_K M \sqrt{\frac{1}{\Gamma_K^2}\textstyle\sum_{k=0}^K \gamma_k^2 \sigma_{k+1}^2 d}. \\
&\Rightarrow \mathbb{E} \sup_{w\in\mathcal{W}} \langle F(w), \Pi_{\mathcal{W}}[\overline{w}_K] - w\rangle
\end{aligned}
$$

$$\leqslant \nu + \frac{1}{\Gamma_K}\left[\frac{3\gamma_0\lambda_0}{2}D^2 + \frac{4M^2+3L^2D^2}{\gamma_0\lambda_0}\sum_{k=0}^K\gamma_k^2 + \frac{5\gamma_0\lambda_0 d}{2}\sum_{k=1}^K\sigma_{k+1}^2\right]$$
$$+ M\sqrt{\frac{1}{\Gamma_K^2}\sum_{k=0}^K\gamma_k^2\sigma_{k+1}^2 d},$$

where in the first inequality, we used the fact that $\mathbb{E}[\text{dist}_{\mathcal{W}}(\overline{w}_K)] \leqslant \sqrt{\mathbb{E}[\text{dist}_{\mathcal{W}}(\overline{w}_K)^2]}$. Hence, we conclude the proof of (5.5).

Now, we extend this for (SP(f)). Denote $u^{k+1} = (\widetilde{x}^{k+1}, \widetilde{y}^{k+1})$. Then, we have

$$\begin{aligned}
&\langle F(u_{k+1}), u_{k+1} - w\rangle \\
&\quad = \langle \nabla_x f(\widetilde{x}_{k+1}, \widetilde{y}_{k+1}), \widetilde{x}_{k+1} - x\rangle + \langle -\nabla_y f(\widetilde{x}_{k+1}, \widetilde{y}_{k+1}), \widetilde{y}_{k+1} - y\rangle \\
&\quad \geqslant f(\widetilde{x}_{k+1}, \widetilde{y}_{k+1}) - f(x, \widetilde{y}_{k+1}) + [-f(\widetilde{x}_{k+1}, \widetilde{y}_{k+1}) + f(\widetilde{x}_{k+1}, y)] \\
&\quad = f(\widetilde{x}_{k+1}, y) - f(x, \widetilde{y}_{k+1}).
\end{aligned}$$

Using the above in (5.11), we obtain,

$$\begin{aligned}
f(\widetilde{x}_{k+1}, y) - f(x, \widetilde{y}_{k+1}) \leqslant \frac{\lambda_k}{2}\left[\|w - u_k\|^2 - \|w - u_{k+1}\|^2\right] + \lambda_k\langle \boldsymbol{\xi}_k, u_k - w\rangle \\
+ \frac{1}{2}\lambda_k\|\boldsymbol{\xi}_k\|^2 + E_k.
\end{aligned}$$

Now, using Proposition 8 to bound the distance between points $\frac{1}{\Gamma_K}(\sum_{k=0}^K\gamma_k\widetilde{x}_{k+1}, \sum_{k=0}^K\gamma_k\widetilde{y}_{k+1})$ and $(\Pi_{\mathcal{X}}[\overline{x}_K], \Pi_{\mathcal{Y}}[\overline{y}_K])$ and using Jensen's inequality to conclude that

$$\begin{aligned}
&\frac{1}{\Gamma_K}\sum_{k=0}^K\gamma_k[f(\widetilde{x}_{k+1}, y) - f(x, \widetilde{y}_{k+1})] \\
&\quad \geqslant f(\tfrac{1}{\Gamma_K}\textstyle\sum_{k=0}^K\gamma_k\widetilde{x}_{k+1}, y) - f(x, \tfrac{1}{\Gamma_K}\sum_{k=0}^K\gamma_k\widetilde{y}_{k+1})] \\
&\quad \geqslant f(\Pi_{\mathcal{X}}[\overline{x}_K], y) - f(x, \Pi_{\mathcal{Y}}[\overline{y}_K]) - M\|*\|\begin{bmatrix}\frac{1}{\Gamma_K}\sum_{k=0}^K\gamma_k\widetilde{x}_{k+1} - \Pi_{\mathcal{X}}[\overline{x}_K] \\ \frac{1}{\Gamma_K}\sum_{k=0}^K\gamma_k\widetilde{y}_{k+1} - \Pi_{\mathcal{Y}}[\overline{y}_K]\end{bmatrix}
\end{aligned}$$

and retracing the steps of this proof from (5.11), we obtain (5.6). Hence, we conclude the proof. □

5.1 Differential privacy of the NISPP method

First, we show a simple bound on ℓ_2-sensitivity for updates of NISPP method.

Proposition 9 *Suppose* $\nu \leqslant \frac{2M^2}{\lambda_k B_{k+1}^2}$ *then* ℓ_2*-sensitivity of updates of Algorithm 2 is at most* $\frac{4M}{\lambda_k B_{k+1}}$

where $B_{k+1} = |\mathbf{B}_{k+1}|$ *is the batch size of* k*-th iteration.*

Proof Let w_k be an iterate in the start of k-th iteration of Algorithm 2. Suppose $\mathbf{B}_{k+1}$ and $\mathbf{B}'_{k+1}$ be two different batches used in k-th iteration to obtain u_{k+1} and u'_{k+1}, respectively. Also note that $\mathbf{B}_{k+1}$ and $\mathbf{B}'_{k+1}$ differ in only single datapoint. Then, due to (5.1), we have for all $w \in \mathcal{W}$

$$\begin{aligned}\langle F_{\mathbf{B}_{k+1}}(u_{k+1}) + \lambda_k(u_{k+1} - w_k), w - u_{k+1}\rangle &\geqslant -\nu\\ \langle F_{\mathbf{B}'_{k+1}}(u'_{k+1}) + \lambda_k(u'_{k+1} - w_k), w - u'_{k+1}\rangle &\geqslant -\nu\end{aligned}$$

Using $w = u'_{k+1}$ in the first relation and $w = u_{k+1}$ in the second relation above and then summing, we obtain

$$\begin{aligned}&\langle F_{\mathbf{B}_{k+1}}(u_{k+1}) - F_{\mathbf{B}'_{k+1}}(u'_{k+1}), u_{k+1} - u'_{k+1}\rangle\\ &\quad\leqslant 2\nu - \lambda_k\|u_{k+1} - u'_{k+1}\|^2\\ &\Rightarrow \langle F_{\mathbf{B}_{k+1}}(u_{k+1}) - F_{\mathbf{B}_{k+1}}(u'_{k+1}), u_{k+1} - u'_{k+1}\rangle\\ &\quad\leqslant \langle F_{\mathbf{B}'_{k+1}}(u'_{k+1}) - F_{\mathbf{B}_{k+1}}(u'_{k+1}), u_{k+1} - u'_{k+1}\rangle\\ &\quad + 2\nu - \lambda_k\|u_{k+1} - u'_{k+1}\|^2\end{aligned}$$

Now, noting that $F_{\mathbf{B}_{k+1}}$ is a monotone operator and denoting $\mathbf{a}_{k+1} := \|F_{\mathbf{B}'_{k+1}}(u'_{k+1}) - F_{\mathbf{B}_{k+1}}(u'_{k+1})\|$, $p_{k+1} := \|w_{k+1} - w'_{k+1}\| = \|u_{k+1} - u'_{k+1}\|$ we have

$$\begin{aligned}0 &\leqslant \langle F_{\mathbf{B}'_{k+1}}(u'_{k+1}) - F_{\mathbf{B}_{k+1}}(u'_{k+1}), u_{k+1} - u'_{k+1}\rangle + 2\nu - \lambda_k\|u_{k+1} - u'_{k+1}\|^2\\ &\leqslant \mathbf{a}_{k+1}p_{k+1} - \lambda_k p_{k+1}^2 + 2\nu. \end{aligned} \tag{5.22}$$

Finally noting that if $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ are the differing datapoints in $\mathbf{B}_{k+1}$ and $\mathbf{B}'_{k+1}$, then

$$\mathbf{a}_{k+1} = \frac{1}{B_{k+1}}\|F_{\boldsymbol{\beta}'}(u'_{k+1}) - F_{\boldsymbol{\beta}}(u'_{k+1})\| \leqslant \frac{2M}{B_{k+1}}.$$

Using the above relation in (5.22) and noting that ℓ_2-sensitivity $p_{k+1} = \|w_{k+1} - w'_{k+1}\| = \|u_{k+1} - u'_{k+1}\|$, we have, p_{k+1} satisfies

$$p_{k+1}^2 - \frac{2M}{\lambda_k B_{k+1}}p_{k+1} - \frac{2\nu}{\lambda_k} \leqslant 0.$$

This implies

$$p_{k+1} \leqslant \frac{M}{\lambda_k B_{k+1}} + \sqrt{\frac{M^2}{\lambda_k^2 B_{k+1}^2} + \frac{2\nu}{\lambda_k}} \leqslant \frac{2M}{\lambda_k B_{k+1}} + \sqrt{\frac{2\nu}{\lambda_k}}.$$

Setting $\nu \leqslant \frac{2M^2}{\lambda_k B_{k+1}^2}$, we have $p_{k+1} \leqslant \frac{4M}{\lambda_k B_{k+1}}$. Hence, we conclude the proof. □

Using the ℓ_2-sensitivity result above along with Proposition 3 and 4, we immediately obtain the following:

Proposition 10 *Algorithm 2 with batch sizes* $(B_{k+1})_{k\in[K]_0}$, *parameters* $(\lambda_k)_{k\in[K]_0}$, *variance* $\sigma_{k+1}^2 = \frac{32\,M^2}{\lambda_k^2 B_{k+1}^2}\frac{\ln(1/\eta)}{\varepsilon^2}$ *and* ν *satisfying assumptions of Proposition 9 is* (ε, η)*-differentially private.*

Now, we provide a policy for setting γ_k, λ_k and B_{k+1} to obtain population risk bounds for DP-SVI and DP-SSP problem by the NISPP method.

Corollary 2 *Algorithm 2 with disjoint batches* $\mathbf{B}_{k+1}$ *is of size* $B_{k+1} = B := n^{1/3}$ *for all* $k \geqslant 0$ *and the following parameters*

$$\gamma_k = 1, \qquad \lambda_k = \lambda_0 := \max\left\{\frac{M}{D}, L\right\}\max\left\{n^{1/3}, \frac{\sqrt{d\ln(1/\eta)}}{\varepsilon}\right\},$$
$$\sigma_{k+1}^2 = \frac{32M^2}{B\lambda_0}\frac{\ln(1/\eta)}{\varepsilon^2}, \qquad \nu = \frac{2M^2}{\lambda_0 B^2},$$

is (ε, η)*-differentially private and achieves expected SVI-gap (SSP-gap, respectively)*

$$O\left((M + LD)D\left[\frac{1}{n^{1/3}} + \frac{\sqrt{d\ln(1/\eta)}}{\varepsilon n^{2/3}}\right]\right).$$

Proof Note that values of ν, σ_{k+1} and other required conditions proposed in Propositions 9 and 10 are satisfied. Hence, this algorithm is (ε, η)-differentially private.

Moreover, all requirements of Theorem 5 are satisfied. In order to maintain single pass over the dataset, we require $K = \frac{n}{B} = n^{2/3}$ iterations. Then, we provide individual bounds on the terms of (5.5) ((5.6), respectively) and conclude the corollary using Theorem 5.

Note that we are using a constant parameter policy. Hence, $\sigma_{k+1} = \sigma = \frac{4M}{\rho B\lambda_0}$ for all $k \geqslant 0$. Substituting appropriate parameter values, we have

$$\nu = \frac{2MD}{n^{2/3}\max\{n^{1/3}, \sqrt{d\ln(1/\eta)}/\varepsilon\}} \leqslant \frac{2MD}{n},$$
$$M\sqrt{\frac{1}{\Gamma_K^2}\sum_{k=0}^K \gamma_k^2\sigma_{k+1}^2 d} = \frac{M\sqrt{d}\sigma_{k+1}}{\sqrt{K}} = \frac{4M^2\sqrt{2d\ln(1/\eta)}}{\varepsilon n^{2/3}\lambda_0} \leqslant \frac{4\sqrt{2}MD}{n^{2/3}},$$
$$\frac{3\lambda_0 D^2}{2K} \leqslant \frac{3(M + LD)D}{2}\left(\frac{1}{n^{1/3}} + \frac{\sqrt{d\ln(1/\eta)}}{\varepsilon n^{2/3}}\right),$$
$$\frac{4M^2 + 3L^2D^2}{\lambda_0} \leqslant \frac{4MD}{n^{1/3}} + \frac{3LD^2}{n^{1/3}},$$
$$\frac{5\lambda_0 d\sigma^2}{2} = \frac{40M^2 d\ln(1/\eta)}{\varepsilon^2 B^2\lambda_0} \leqslant \frac{40MD\sqrt{d\ln(1/\eta)}}{\varepsilon n^{2/3}}.$$

Substituting these bounds in Theorem 5, we conclude the proof. □

Remark 3 We have the following remarks for NISPP method:

1. In order to obtain ν-approximate solution of the subproblem of NISPP method satisfying (5.1), we can use the Operator Extrapolation (OE) method (see Theorem 2.3 [34]). OE method outputs a solution u_{k+1} satisfying $\|u_{k+1} - w^*_{k+1}\| \leqslant \zeta$ in $\frac{L+\lambda_0}{\lambda_0} \ln(\frac{D}{\zeta})$ iterations, where w^*_{k+1} is an exact SVI solution for problem (5.1). Furthermore, we have for all $w \in \mathcal{W}$,

$$\begin{aligned}
0 &\leqslant \langle F(w^*_{k+1}) + \lambda_k(w^*_{k+1} - w_k), w - w^*_{k+1}\rangle \\
&= \langle F(u_{k+1}) + F(w^*_{k+1}) - F(u_{k+1}) + \lambda_k(u_{k+1} - w_k) \\
&\quad + \lambda_k(w^*_{k+1} - u_{k+1}), w - w^*_{k+1}\rangle \\
&\leqslant \langle F(u_{k+1}) + \lambda_k(u_{k+1} - w_k), w - w^*_{k+1}\rangle \\
&\quad + (L + \lambda_k)\|u_{k+1} - w^*_{k+1}\|\,\|w - w^*_{k+1}\| \\
&\leqslant \langle F(u_{k+1}) + \lambda_k(u_{k+1} - w_k), w - w^*_{k+1}\rangle \\
&\quad + (L + \lambda_k)D\|u_{k+1} - w^*_{k+1}\| \\
&= \langle F(u_{k+1}) + \lambda_k(u_{k+1} - w_k), w - u_{k+1} + u_{k+1} - w^*_{k+1}\rangle \\
&\quad + (L + \lambda_k)D\|u_{k+1} - w^*_{k+1}\| \\
&\leqslant \langle F(u_{k+1}) + \lambda_k(u_{k+1} - w_k), w - u_{k+1}\rangle + \|F(u_{k+1}) \\
&\quad + \lambda_k(u_{k+1} - w_k)\|\,\|u_{k+1} - w^*_{k+1}\| \\
&\quad + (L + \lambda_k)D\|u_{k+1} - w^*_{k+1}\| \\
&\leqslant \langle F(u_{k+1}) + \lambda_k(u_{k+1} - w_k), w - u_{k+1}\rangle \\
&\quad + [LD + M + 2\lambda_k D]\|u_{k+1} - w^*_{k+1}\|
\end{aligned}$$

Setting $\zeta = \nu/[LD + M + 2\lambda_k D]$, we obtain that u_{k+1} is a ν-approximate solution satisfying (5.1). Using the convergence rate above, we require $\frac{L+\lambda_0}{\lambda_0} \ln \frac{MD+LD^2+2\lambda_k D^2}{\nu}$ operator evaluations.
Note that since, $\lambda_0 \geqslant L$, we have $\frac{L+\lambda_0}{\lambda_0} \leqslant 2$. Moreover,

$$\begin{aligned}
\ln \frac{MD + LD^2 + 2\lambda_k D^2}{\nu} &\leqslant \ln \frac{4\lambda_k D^2}{\nu} \\
&= \ln\left(\frac{2\lambda_0^2 D^2 B^2}{M^2}\right) \\
&= \ln\left(n^{2/3} \max\left\{n^{2/3}, \frac{d \ln(1\eta)}{\varepsilon^2}\right\} \max\left\{1, \frac{L^2 D^2}{M^2}\right)\right\}
\end{aligned} \tag{5.23}$$

Hence, each iteration of NISPP method requires $O(\log n)$ iterations of OE method for solving the subproblem. Moreover, each iteration of the OE method requires $2B$ stochastic operator evaluations. Hence, we require $O(KB \log n)$ stochastic operator evaluations in the entire run of NISPP (Algorithm 2). Noting that $KB = n$, we conclude that this is a near linear time algorithm and also performs only a single pass over the data in the stochastic outer-loop. We provide the details of OE method in the Appendix B.

2. For non-DP version of NISPP method, i.e., $\sigma_k = 0$ for all k, we can easily obtain population risk bound of $O(\frac{MD}{\sqrt{n}})$ by setting $\lambda_0 = \frac{M}{D}\sqrt{n}$, $B = 1$ (or $K = n$) and $\nu = \frac{MD}{\sqrt{n}}$ in Corollary 2.

In view of Corollary 2, it seems that running NISPP method for $n^{3/2}$ stochastic operator evaluations may provide optimal risk bounds. However, running that many stochastic operator evaluations requires multi-pass over the dataset so, in principle, this would only provide bounds in the empirical risk. In order to compute the population risk of this multi-pass version, we analyze the stability of NISPP and provide generalization guarantees which result in optimal population risk.

6 Stability of NISPP and optimal risk for DP-SVI and DP-SSP

In this section, we develop a multi-pass variant of NISPP method, and prove its stability to extrapolate empirical performance to population risk bounds.

6.1 Stability of NISPP method

Let us start with two adjacent datasets $\mathbf{S} \simeq \mathbf{S}'$. Suppose we run NISPP method on both datasets starting from the same point $w_0 \in \mathcal{W}$. Then, in the following lemma, we provide bound on the how far apart trajectories of these two runs can drift.

Lemma 3 *Let $(u_{k+1}, w_{k+1})_{k\geqslant 0}$ and $(u'_{k+1}, w'_{k+1})_{k\geqslant 0}$ be two trajectories of the NISPP method (Algorithm 2) for any adjacent datasets $\mathbf{S} \simeq \mathbf{S}'$ whose batches are denotes by $\mathbf{B}_{k+1}, \mathbf{B}'_{k+1}$ respectively. Moreover, denote $\mathbf{a}_{k+1} := \|F_{\mathbf{B}_{k+1}}(u_{k+1}) - F_{\mathbf{B}'_{k+1}}(u_{k+1})\|$ and $\delta_{k+1} := \|u_{k+1} - u'_{k+1}\| (= \|w_{k+1} - w'_{k+1}\|)$ for k-th iteration of Algorithm 2. Then, if $i = \inf\{k : \mathbf{B}_{k+1} \neq \mathbf{B}'_{k+1}\}$,*

$$\delta_{j+1} \begin{cases} = 0 & \text{if } j+1 \leqslant i \\ \leqslant \sum_{k=i}^{j} \frac{2\mathbf{a}_{k+1}}{\lambda_k} + \sqrt{\frac{4\nu}{\lambda_k}} & \text{otherwise.} \end{cases} \tag{6.1}$$

Proof It is clear from the definition i that $\mathbf{B}_j = \mathbf{B}'_j$ for all $j \leqslant i$. This implies $u_j = u'_j$ and $w_j = w'_j$ for all $j \leqslant i$. Hence, we conclude first case of (6.1).

Using (5.1) for ν-approximate strong VI solution, we have,

$$\langle F_{\mathbf{B}_{k+1}}(u_{k+1}) + \lambda_k(u_{k+1} - w_k), w - u_{k+1}\rangle \geqslant -\nu, \tag{6.2}$$
$$\langle F_{\mathbf{B}'_{k+1}}(u'_{k+1}) + \lambda_k(u'_{k+1} - w'_k), w - u'_{k+1}\rangle \geqslant -\nu. \tag{6.3}$$

Then, adding (6.2) with $w = u'_{k+1}$ and (6.3) with $w = u_{k+1}$, we have

$$\begin{aligned} &\langle F_{\boldsymbol{\beta}_{k+1}}(u_{k+1}) - F_{\boldsymbol{\beta}'_{k+1}}(u'_{k+1}), u_{k+1} - u'_{k+1}\rangle \\ &\quad\leqslant 2\nu - \lambda_k\langle u_{k+1} - u'_{k+1}, (u_{k+1} - w_k) - (u'_{k+1} - w'_k)\rangle \end{aligned}$$

$$
\begin{aligned}
&= 2\nu - \lambda_k \delta_{k+1}^2 + \lambda_k \langle u_{k+1} - u'_{k+1}, w_k - w'_k \rangle \\
&\leqslant 2\nu - \lambda_k \delta_{k+1}^2 + \frac{\lambda_k}{2}[\delta_k^2 + \delta_{k+1}^2] \\
&\leqslant 2\nu - \frac{\lambda_k}{2}\delta_{k+1}^2 + \frac{\lambda_k}{2}\delta_k^2. \qquad (6.4)
\end{aligned}
$$

Also note that

$$
\begin{aligned}
&\langle F_{\mathbf{B}_{k+1}}(u_{k+1}) - F_{\mathbf{B}'_{k+1}}(u'_{k+1}), u_{k+1} - u'_{k+1} \rangle \\
&\quad = \langle F_{\mathbf{B}'_{k+1}}(u_{k+1}) - F_{\mathbf{B}'_{k+1}}(u'_{k+1}), u_{k+1} - u'_{k+1} \rangle \\
&\quad + \langle F_{\mathbf{B}_{k+1}}(u_{k+1}) - F_{\mathbf{B}'_{k+1}}(u_{k+1}), u_{k+1} - u'_{k+1} \rangle \\
&\quad \geqslant \langle F_{\mathbf{B}_{k+1}}(u_{k+1}) - F_{\mathbf{B}'_{k+1}}(u_{k+1}), u_{k+1} - u'_{k+1} \rangle
\end{aligned}
$$

where the last inequality follows from monotonicity of $F_{\mathbf{B}'_{k+1}}$. Using above relation along with (6.4), we obtain

$$
\begin{aligned}
&\frac{\lambda_k}{2}\delta_{k+1}^2 \leqslant \frac{\lambda_k}{2}\delta_k^2 + 2\nu + \langle F_{\mathbf{B}'_{k+1}}(u_{k+1}) - F_{\mathbf{B}_{k+1}}(u_{k+1}), u_{k+1} - u'_{k+1} \rangle \\
&\Rightarrow \delta_{k+1}^2 \leqslant \delta_k^2 + \frac{4\nu}{\lambda_k} + \frac{2}{\lambda_k}\mathbf{a}_{k+1}\delta_{k+1},
\end{aligned}
$$

where we used the definition a_k along with Cauchy-Schwarz inequality. Solving for the quadratic inequality in δ_{k+1}, we obtain the following recursion

$$
\delta_{k+1} \leqslant \frac{\mathbf{a}_{k+1}}{\lambda_k} + \sqrt{\frac{\mathbf{a}_{k+1}^2}{\lambda_k^2} + \delta_k^2 + \frac{4\nu}{\lambda_k}}
$$

which can be further simplified to

$$
\delta_{k+1} \leqslant \delta_k + \frac{2\mathbf{a}_{k+1}}{\lambda_k} + \sqrt{\frac{4\nu}{\lambda_k}}.
$$

Solving this recursion and noting the base case that $\delta_i = 0$, we obtain (6.1). □

A direct consequence of the previous analysis are in-expectation and high probability uniform argument stability upper bounds for the sampling with replacement variant of Algorithm 2.

Theorem 6 *Let $\mathcal{A}$ denote the sampling with replacement NISPP method (Algorithm 2) where $\mathbf{B}_k$ is chosen uniformly at random from subsets of $\mathbf{S}$ of a given size B_k. Then $\mathcal{A}$ satisfies the following uniform argument stability bounds:*

$$
\sup_{\mathbf{S} \simeq \mathbf{S}'} \mathbb{E}_{\mathcal{A}}[\delta_{\mathcal{A}}(\mathbf{S}, \mathbf{S}')] \leqslant \sum_{k=1}^{K} \left(\frac{2M}{n\lambda_k} + \sqrt{\frac{4\nu}{\lambda_k}} \right).
$$

Furthermore, if $|\mathbf{B}_k| = B$ *and* $\lambda_k = \lambda$ *for all* k *(i.e., constant batch size and regularization parameter throughout iterations) then w.p. at least* $1 - n\exp\{-KB/[4n]\}$ *(over both sampling and noise addition)*

$$\sup_{\mathbf{S}\simeq\mathbf{S}'}[\delta_{\mathcal{A}}(\mathbf{S},\mathbf{S}')] \leqslant \frac{4MK}{\lambda n} + K\sqrt{\frac{4\nu}{\lambda}}.$$

Proof Let $\mathbf{S} \simeq \mathbf{S}'$ and $(u_{k+1})_k$, $(u'_{k+1})_k$ the trajectories of the algorithm on $\mathbf{S}$ and $\mathbf{S}'$, respectively. By Lemma 3, letting $\delta_{k+1} = \|\tilde{w}_{k+1} - \tilde{w}'_{k+1}\|$, we get $\delta_j \leqslant \sum_{k=1}^{j}\left(\frac{2\mathbf{a}_k}{\lambda_k} + \sqrt{\frac{4\nu}{\lambda_k}}\right)$, where $\mathbf{a}_k = \|F_{\mathbf{B}_{k+1}}(u_{k+1}) - F_{\mathbf{B}'_{k+1}}(u'_{k+1})\|$ is a random variable. By the law of total probability, $\mathbb{E}[\mathbf{a}_k] \leqslant \frac{|\mathbf{B}_{k+1}|}{n}\frac{2M}{|\mathbf{B}_{k+1}|} + \left(1 - \frac{|\mathbf{B}_{k+1}|}{n}\right)\cdot 0 = \frac{2M}{n}$. Hence, $\mathbb{E}[\delta_j] \leqslant \sum_{k=1}^{j}\left(\frac{2M}{n\lambda_k} + \sqrt{\frac{4\nu}{\lambda_k}}\right) \leqslant \sum_{k=1}^{K}\left(\frac{2M}{n\lambda_k} + \sqrt{\frac{4\nu}{\lambda_k}}\right)$. Since $\|\Pi_{\mathcal{W}}(\overline{w}_K) - \Pi_{\mathcal{W}}(\overline{w}'_K)\| \leqslant \|\overline{w}_K - \overline{w}'_K\| \leqslant \max_{k\in[K]}\delta_k$, and since $\mathbf{S} \simeq \mathbf{S}'$ are arbitrary,

$$\sup_{\mathbf{S}\simeq\mathbf{S}'}\mathbb{E}_{\mathcal{A}}[\delta_{\mathcal{A}}(\mathbf{S},\mathbf{S}')] \leqslant \sum_{k=1}^{K}\left(\frac{2M}{n\lambda_k} + \sqrt{\frac{4\nu}{\lambda_k}}\right).$$

We proceed now to the high-probability bound. Let $\boldsymbol{r}_k \sim \mathrm{Ber}(p)$, for $k \in [K]$, with $Kp < 1$. Then, for any $0 < \theta < 1/2$,

$$\mathbb{P}\left[\sum_{k=1}^{K}\boldsymbol{r}_k \geqslant Kp + \tau\right] \leqslant \exp\big(-\theta(\tau + Kp)\big)\Big[1 + p(e^{\theta} - 1)\Big]^K \leqslant \exp\{Kp\theta^2 - \theta\tau\}.$$

Choosing $\theta = \tau/[2Kp] < 1/2$, we get that the probability above is upper bounded by $\exp\{-\tau^2/[4Kp]\}$. Finally, choosing $\tau = Kp$, we get

$$\mathbb{P}\left[\sum_{k=1}^{K}\boldsymbol{r}_k \geqslant 2Kp\right] \leqslant \exp\{-Kp/4\}.$$

Next, fix the coordinate i where $S \simeq S'$ may differ. Noticing that $\mathbf{a}_k$ is a.s. upper bounded by $(2M/B)\boldsymbol{r}_k$ with $\boldsymbol{r}_k \sim \mathrm{Ber}(p)$, with $p = B/n$, we get

$$\mathbb{P}\left[\sum_{k=1}^{K}\frac{2\mathbf{a}_k}{\lambda} \geqslant \frac{2}{\lambda}\frac{2M}{n}\right] \leqslant \exp\{-\frac{KB}{4n}\}.$$

In particular, w.p. at least $1 - \exp\{-\frac{KB}{4n}\}$, we have $\mathbf{a}_k \leqslant \frac{4M}{\lambda n} + \sqrt{\frac{4\nu}{\lambda}}$. Using a union bound over $i \in [n]$ (and noticing that averaging preserves the stability bound), we conclude that

$$\mathbb{P}\left[\sup_{\mathbf{S}\simeq\mathbf{S}'}\delta_{\mathcal{A}}(\mathbf{S},\mathbf{S}') > \frac{4MK}{\lambda n} + K\sqrt{\frac{4\nu}{\lambda}}\right] \leqslant n\exp\{-KB/4n\}.$$

Hence, we conclude the proof. □

Remark 4 An important observation, for the high-probability guarantee to be useful, is necessary that the algorithm is run for sufficiently many iterations; in particular, we require $K = \omega(n/B)$. Whether this assumption can be completely avoided is an interesting question. Nevertheless, as we will see in the section, our policy for DP-SVI and DP-SSP problem satisfies this requirement.

6.2 Optimal risk for DP-SVI and DP-SSP by the NISPP method

In previous three sections, we provided bounds on optimization error, generalization error and value of σ for obtaining (ε, η)-differential privacy. In this section, we specify a policy for selecting λ_k, B_k, γ_k, σ_k and ν such that requirement in the previous three sections are satisfied and we can get optimal risk bounds while maintaining (ε, η)-privacy. In particular, consider the multi-pass NISPP method where each sample batch $\mathbf{B}_k$ is chosen randomly from subsets of $\mathbf{S}$ with replacement. Then, we have the following theorem:

Theorem 7 *Let $\mathcal{A}$ be the multi-pass NISPP method (Algorithm 2). Set the following constant stepsize and batchsize policy for $\mathcal{A}$:*

$$\gamma_k = 1, \qquad \lambda_k = \lambda_0 = \max\left\{\frac{M}{D}, L\right\} \max\left\{\sqrt{n}, \frac{\sqrt{d \log 1/\delta}}{\theta}\right\}, \qquad B_k = B = \sqrt{n},$$

$$K = n, \qquad \nu = \frac{M^2}{\lambda_0 n^2}, \qquad \sigma_{k+1} = \frac{8M}{B\lambda_0}\frac{\sqrt{\ln(1/\eta)}}{\varepsilon}.$$

Then, Algorithm 2 is (ε, η)-differential private. Moreover, output $\mathcal{A}(\mathbf{S})$ satisfies the following bound on $\mathbb{E}_{\mathcal{A}}[WeakGap_{VI}(\mathcal{A}(\mathbf{S}), F)]$ for SVI problem (or $\mathbb{E}_{\mathcal{A}}[WeakGap_{SP}(\mathcal{A}(\mathbf{S}), f)]$ for SSP problem)

$$O\left((M + LD)D\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d \ln 1/\eta}}{\varepsilon n}\right)\right),$$

Moreover, such solution is obtained in total of $\widetilde{O}(n\sqrt{n})$ stochastic operator evaluations.

Proof Note that since ν satisfies assumption in Proposition 9, we have ℓ_2-sensitivity of the update of u_{k+1} is $\frac{4M}{\lambda_0 B_{k+1}}$. Then, in view of Theorem 1 along with value of σ_{k+1}, we conclude that Algorithm 2 is (ε, η)-differential private.

Now, for convergence, we first bound the empirical gap. Given that our bounds for (VI(F)) and (SP(f)) are analogous, we proceed indistinctively for both problems. By Theorem 5, along with the fact that sampling with replacement is an unbiased stochastic oracle for the empirical operator, we have for any $\mathbf{S}$

$$\mathbb{E}_{\mathcal{A}}[\text{EmpGap}(\mathcal{A}, F_{\mathbf{S}})] \leqslant \nu + \frac{\lambda_0 D^2}{n} + \frac{7M^2}{\lambda_0} + \frac{160M^2 d}{\varepsilon^2 B^2 \lambda_0} \ln\frac{1}{\eta} + M\sqrt{2d}\frac{8M}{Bn\varepsilon\lambda_0}$$

$$\leqslant O\left(MD\left(\frac{1}{\sqrt{n}}+\frac{\sqrt{d\ln 1/\eta}}{\varepsilon n}\right)\right).$$

Similar claims can be made for empirical gap for (SP(f)) problem: $\mathbb{E}_{\mathcal{A}}\big[\text{EmpGap}(\mathcal{A}, f_{\mathbf{S}})\big]$ where output of $\mathcal{A}$ is $(x(\mathbf{S}), y(\mathbf{S}))$.

Next, by Theorem 6, we have that $\mathcal{A}(\mathbf{S})$ (or $x(\mathbf{S})$ and $y(\mathbf{S})$ for the SSP case) are UAS with parameter

$$\delta = \frac{2M}{\lambda_0}+n\sqrt{\frac{4\nu}{\lambda_0}} = \frac{4M}{\lambda_0} \leqslant \frac{4D}{\sqrt{n}}.$$

Hence, noting that empirical risk upper bounds weak VI or SP gap, i.e., using Proposition 1 or Proposition 2 (depending on whether the problem is an SSP or SVI, respectively), we have that the risk is upper bounded by its empirical risk plus $M\delta$, where δ is the UAS parameter of the algorithm; in particular, if $\text{WeakGap}(\mathcal{A}; \mathbf{S})$ is the (SVI or SSP, respectively) gap function for the expectation objective, then

$$\mathbb{E}_{\mathcal{A},\mathbf{S}}\big[\text{WeakGap}_{\text{VI}}(\mathcal{A}, F)\big] \leqslant O\Big(MD\Big(\frac{1}{\sqrt{n}}+\frac{\sqrt{d\ln(1/\eta)}}{\varepsilon n}\Big)\Big)$$
$$+\frac{20MD}{\sqrt{n}} = O\Big(MD\Big(\frac{1}{\sqrt{n}}+\frac{\sqrt{d\ln(1/\eta)}}{\varepsilon n}\Big)\Big)$$

Similar claim can be made for $\text{WeakGap}_{\text{SP}}(\mathcal{A}, f)$.

Finally, we analyze the running time performance. As in Remark 3, number of OE method iterations for obtaining ν-approximate solution is $O\big(\frac{L+\lambda_0}{\lambda_0}\ln\big(\frac{LD^2+MD+\lambda_0 D^2}{\nu}\big)\big)$. Now note that $\frac{L+\lambda_0}{\lambda_0} \leqslant \frac{\sqrt{n}+1}{\sqrt{n}} \leqslant 2$ since $n \geqslant 1$. Moreover, in view of (5.23), we have $\ln\big(\frac{LD^2+MD+\lambda_0 D^2}{\nu}\big) \leqslant \ln\Big(\frac{4\lambda_0 D^2}{\nu}\Big) \leqslant \ln\Big(n^2\max\{1, \frac{L^2D^2}{M^2}\}\max\{n, \frac{d\ln(1/\eta)}{\varepsilon^2}\}\Big)$. Each iteration of OE method costs B stochastic operator evaluations and we run outer-loop of NISPP method K times. Hence, total stochastic operator evaluations (after ignoring the ln-term) $\widetilde{O}(KB) = \widetilde{O}(n\sqrt{n})$. Hence, we conclude the proof. □

7 Lower bounds and optimality of our algorithms

In this section, we show the optimality of our obtained rates from Sects. 4.1 and 6.1. The first observation is that, since DP-SCO corresponds to a DP-SSP problem where $\mathcal{Y}$ is a singleton, the complexity of DP-SSP is lower bounded by $\Omega(MD\big(\frac{1}{\sqrt{n}}+\min\big\{1, \frac{\sqrt{d}}{\varepsilon n}\big\}\big))$: this is a known lower bound for DP-SCO [5]. It is important to note as well that this reduction applies to the weak generalization gap, as defined in (1.4), as in the case $\mathcal{Y} = \{\bar{y}\}$ is a singleton:

$$\text{WeakGap}_{\text{SP}}(\mathcal{A}, f) = \mathbb{E}_{\mathcal{A}}[\sup_{y\in\mathcal{Y}}\mathbb{E}_{\mathbf{S}}[f(x(\mathbf{S}), y)] - \inf_{x\in\mathcal{X}}\mathbb{E}_{\mathbf{S}}[f(x, y(\mathbf{S}))]]$$

$$= \mathbb{E}_{\mathcal{A}}\mathbb{E}_{\mathbf{S}}[f(x(\mathbf{S}), \bar{y})] - \inf_{x\in\mathcal{X}} f(x, \bar{y})$$
$$= \mathbb{E}_{\mathcal{A},\mathbf{S}}[f(x(\mathbf{S}), \bar{y}) - \inf_{x\in\mathcal{X}} f(x, \bar{y})],$$

which is simply the expected optimality gap. Using this reduction, together with a lower bound for DP-SCO [5], we conclude that

Proposition 11 *Let $n, d \in \mathbb{N}$, $L, M, D, \varepsilon > 0$ and $\delta = o(1/n)$. The class of problems DP-SSP with gradient operators within the class $\mathcal{M}^1_{\mathcal{W}}(L, M)$, and domain $\mathcal{W}$ containing an Euclidean ball of diameter $D/2$, satisfies the lower bound*

$$\Omega\Big(MD\Big(\frac{1}{\sqrt{n}} + \min\Big\{1, \frac{\sqrt{d}}{\varepsilon n}\Big\}\Big)\Big).$$

Next, we study the case of DP-SVI. The situation is more subtle here. Our approach is to first prove a reduction from population weak VI gap to empirical strong VI gap for the case where operators are constant w.r.t. w. In fact, it seems unlikely this reduction works for more general monotone operators, however this suffices for our purposes, as we will later prove a lower bound construction used for DP-ERM [7] leads to a lower bound for strong VI gap with constant operators.

The formal reduction to the empirical version of the problem is presented in the following lemma. Its proof follows closely the reduction from DP-SCO to DP-ERM from [5]. Below, given a dataset $\mathbf{S} \in \mathcal{Z}^n$, let $\mathcal{P}_{\mathbf{S}} = \frac{1}{n}\sum_{\boldsymbol{\beta}\in\mathbf{S}} \delta_{\boldsymbol{\beta}}$ be the empirical distribution associated with $\mathbf{S}$.

Lemma 4 *Let $\mathcal{A}$ be an $(\varepsilon/[4\log(1/\eta)], e^{-\varepsilon}\eta/[8\log(1/\eta)])$-DP algorithm for SVI problems. Then there exists an (ε, η)-DP algorithm $\mathcal{B}$ such that for any empirical VI problem with constant operators,*

$$\mathit{EmpGap}_{VI}(\mathcal{B}, F_{\mathbf{S}}) \leqslant \mathit{WeakGap}_{VI}(\mathcal{A}, F_{\mathcal{P}_{\mathbf{S}}}) \qquad (\forall \mathbf{S} \in \mathcal{Z}^n).$$

Proof Consider the algorithm $\mathcal{B}$ that does the following: first, it extracts a sample $\mathbf{T} \sim (\mathcal{P}_{\mathbf{S}})^n$; next, executes $\mathcal{A}$ on $\mathbf{T}$; and finally, outputs $\mathcal{A}(\mathbf{T})$. We claim that this algorithm is (ε, η)-DP w.r.t. $\mathbf{S}$, which follows easily by bounding the number of repeated examples with high probability, together with the group privacy property applied to $\mathcal{A}$ (for a more detailed proof, see Appendix C in [5]). Now, given a constant operator $F_{\boldsymbol{\beta}}(w)$, let $R(\boldsymbol{\beta}) \in \mathbb{R}^d$ be its unique evaluation. Let now $R_{\mathbf{S}}$ be the unique evaluation of $F_{\mathbf{S}}$, and given a distribution $\mathcal{P}$ let $R_{\mathcal{P}}$ be the unique evaluation of $F_{\mathcal{P}}(w) = \mathbb{E}_{\boldsymbol{\beta}\sim\mathcal{P}}[F_{\boldsymbol{\beta}}(w)]$.

Noting that $\mathbb{E}_{\mathbf{T}}[R_{\mathbf{T}}] = R_{\mathbf{S}}$, we have that

$$\begin{aligned}
\mathrm{EmpGap}_{\mathrm{VI}}(\mathcal{B}(\mathbf{S}), F_{\mathbf{S}}) &= \mathbb{E}_{\mathcal{B}}[\sup_{w\in\mathcal{W}} \langle R_{\mathbf{S}}, \mathcal{B}(\mathbf{S}) - w\rangle] \\
&= \mathbb{E}_{\mathcal{A},\mathbf{T}}[\langle R_{\mathbf{S}}, \mathcal{A}(\mathbf{T})\rangle - \inf_{w\in\mathcal{W}} \langle R_{\mathbf{S}}, w\rangle] \\
&= \mathbb{E}_{\mathcal{A}} \sup_{w\in\mathcal{W}} \mathbb{E}_{\mathbf{T}}[\langle R_{\mathbf{S}}, \mathcal{A}(\mathbf{T}) - w\rangle] \\
&= \mathrm{WeakGap}_{\mathrm{VI}}(\mathcal{A}, F_{\mathcal{P}_{\mathbf{S}}}),
\end{aligned}$$

where third equality holds since the optimal choice of w is independent of $\mathbf{T}$, and the last equality holds by definition of the weak gap function and the fact that $\mathbf{T} \sim (\mathcal{P}_{\mathbf{S}})^n$. □

Next, we prove a lower bound for the empirical VI problem over constant operators.

Proposition 12 *Let $n, d \in \mathbb{N}$, $L, M, D, \varepsilon > 0$ and $2^{-o(n)} \leqslant \delta \leqslant o(1/n)$. The class of DP empirical VI problems with constant operators within the class $\mathcal{M}^1_{\mathcal{W}}(L, M)$, and domain $\mathcal{W}$ containing an Euclidean ball of diameter $D/2$ satisfies the lower bound*

$$\Omega\Big(MD\Big(\min\Big\{1, \frac{\sqrt{d\log(1/\eta)}}{\varepsilon n}\Big\}\Big)\Big).$$

Proof Consider the following empirical VI problem: $F_{\boldsymbol{\beta}}(u) = M\boldsymbol{\beta}, \mathcal{W} = \mathcal{B}(0, D)$ and dataset $\mathbf{S}$ with points contained in $\{-1/\sqrt{d}, +1/\sqrt{d}\}^d$. Notice that, since the operator in this case is constant, the VI gap coincides with the excess risk of the associated convex optimization problem

$$(P) \quad \min_{u\in\mathcal{W}} \Big\langle \frac{M}{n}\sum_{i\in[n]} \boldsymbol{\beta}_i, u\Big\rangle.$$

Indeed, for any $u \in \mathcal{W}$,

$$\begin{aligned}\mathrm{EmpGap}_{\mathrm{VI}}(u, F_{\mathbf{S}}) &= \sup_{v\in\mathcal{B}(0,D)} \Big\langle \frac{M}{n}\sum_{i\in[n]} \boldsymbol{\beta}_i, u - v\Big\rangle = \Big\langle \frac{M}{n}\sum_{i\in[n]} \boldsymbol{\beta}_i, u + \frac{D\sum_i \boldsymbol{\beta}_i}{\|\sum_i \boldsymbol{\beta}_i\|}\Big\rangle \\ &= \Big\|\frac{MD}{n}\sum_{i\in[n]} \boldsymbol{\beta}_i\Big\| + MD\langle \frac{u}{D}, \frac{1}{n}\sum_i \boldsymbol{\beta}_i\rangle.\end{aligned}$$

This, together with the lower bounds on excess risk proved for this problem in [7, Appendix C] and [46, Theorem 5.1] show that any (ε, η)-DP algorithm for this problem must incur in worst-case VI gap $\Omega(MD\min\{1, \frac{\sqrt{d\log(1/\eta)}}{\varepsilon n}\})$, which proves the result. □

The two results above provide the claimed lower bound for the weak SVI gap of any differentially private algorithm.

Theorem 8 *Let $n, d \in \mathbb{N}$, $L, M, D, \varepsilon > 0$ and $2^{-o(n)} \leqslant \delta \leqslant o(1/n)$. The class of DP-SVI problems with operators within the class $\mathcal{M}^1_{\mathcal{W}}(L, M)$, and domain $\mathcal{W}$ containing an Euclidean ball of diameter $D/2$ satisfies a lower bound for the weak VI gap*

$$\tilde{\Omega}\Big(MD\Big(\frac{1}{\sqrt{n}} + \min\Big\{1, \frac{\sqrt{d}}{\varepsilon n}\Big\}\Big)\Big).$$

Before we prove the result, let us observe that the presented lower bound shows the optimality of our algorithms in the range where $M \geqslant LD$. Obtaining a matching

lower bound for any choice of M, L, D is an interesting question, which unfortunately our proof technique does not address: this is a limitation that the lower bound is based on constant operators, and therefore their Lipschitz constants are always zero.

Proof Let $\mathcal{A}$ be any algorithm for SVI. By the classical (nonprivate) lower bounds for SVI [30, 40], we have that the minimax SVI gap attainable is lower bounded by $\Omega(MD/\sqrt{n})$. On the other hand, using Lemma 4 the accuracy of any (ε, η)-DP algorithm for weak SVI gap is lower bounded by the strong gap achieved by $(4\varepsilon \ln(1/\eta), e^{\varepsilon}\tilde{O}(\eta))$-DP algorithms on empirical VI problems with constant operators. Finally, by Proposition 12, the latter class of problems enjoys a lower bound $\Omega(\min\{1, \sqrt{d \ln(1/[e^{\varepsilon}\tilde{O}(\eta)])/[\varepsilon n \ln(1/\eta)]}\}) = \tilde{\Omega}(\min\{1, \sqrt{d}/[n\varepsilon]\})$, which implies a lower bound on the former class of this order. We conclude by combining both the private and nonprivate lower bounds established above. □

Acknowledgements CG would like to thank Roberto Cominetti for valuable discussions on stochastic variational inequalities and nonexpansive iterations. Part of this work was done while CG was at the University of Twente.

Funding Open access funding provided by SCELC, Statewide California Electronic Library Consortium

A Proof of Proposition 2

Let $\mathbf{S}' = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_n)$ be independent of $\mathbf{S}$. For $i \in [n]$, we denote $\mathbf{S}^i = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{i-1}, \boldsymbol{\beta}'_i, \boldsymbol{\beta}_{i+1}, \ldots, \boldsymbol{\beta}_n)$. Then, for any $w \in \mathcal{W}$, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{S}}\langle F(w), \mathcal{A}(\mathbf{S}) - w\rangle \\
&\quad = \mathbb{E}_{\mathbf{S},\mathbf{S}'}\frac{1}{n}\sum_{i=1}^{n}\langle F_{\boldsymbol{\beta}'_i}(w), \mathcal{A}(\mathbf{S}) - w\rangle \\
&\quad = \mathbb{E}_{\mathbf{S},\mathbf{S}'}\frac{1}{n}\sum_{i=1}^{n}\langle F_{\boldsymbol{\beta}_i}(w), \mathcal{A}(\mathbf{S}^i) - w\rangle \\
&\quad = \mathbb{E}_{\mathbf{S},\mathbf{S}'}\frac{1}{n}\sum_{i=1}^{n}\langle F_{\boldsymbol{\beta}_i}(w), \mathcal{A}(\mathbf{S}) - w\rangle + \langle F_{\boldsymbol{\beta}_i}(w), \mathcal{A}(\mathbf{S}^i) - \mathcal{A}(\mathbf{S})\rangle \\
&\quad \leqslant \mathbb{E}_{\mathbf{S},\mathbf{S}'}\frac{1}{n}\sum_{i=1}^{n}\langle F_{\boldsymbol{\beta}_i}(w), \mathcal{A}(\mathbf{S}) - w\rangle + \|F_{\boldsymbol{\beta}_i}(w)\|\,\|\mathcal{A}(\mathbf{S}^i) - \mathcal{A}(\mathbf{S})\|
\end{aligned}
$$

$$\leqslant \mathbb{E}_{\mathbf{S},\mathbf{S}'} \frac{1}{n} \sum_{i=1}^{n} \langle F_{\boldsymbol{\beta}_i}(w), \mathcal{A}(\mathbf{S}) - w \rangle + M \|\mathcal{A}(\mathbf{S}^i) - \mathcal{A}(\mathbf{S})\|$$

$$= \mathbb{E}_{\mathbf{S}} \langle F_{\mathbf{S}}(w), \mathcal{A}(\mathbf{S}) - w \rangle + \mathbb{E}_{\mathbf{S},\mathbf{S}'} \frac{1}{n} \sum_{i=1}^{n} M \|\mathcal{A}(\mathbf{S}^i) - \mathcal{A}(\mathbf{S})\| \tag{A.1}$$

Now, taking supremum over $w \in \mathcal{W}$ and taking expectation over $\mathcal{A}$ which is δ-UAS, we have,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}[\mathrm{WeakGap}_{\mathrm{VI}}(\mathcal{A}(\mathbf{S}), F)] \\
&\leqslant \mathbb{E}_{\mathcal{A}}[\sup_{w \in \mathcal{W}} \mathbb{E}_{\mathbf{S}} \langle F_{\mathbf{S}}(w), \mathcal{A}(\mathbf{S}) - w \rangle] + \mathbb{E}_{\mathbf{S},\mathbf{S}'} \frac{1}{n} \sum_{i=1}^{n} M \mathbb{E}_{\mathcal{A}} \|\mathcal{A}(\mathbf{S}^i) - \mathcal{A}(\mathbf{S})\| \\
&\leqslant \mathbb{E}_{\mathcal{A},\mathbf{S}}[\sup_{w \in \mathcal{W}} \langle F_{\mathbf{S}}(w), \mathcal{A}(\mathbf{S}) - w \rangle] + \mathbb{E}_{\mathbf{S},\mathbf{S}'} \frac{1}{n} \sum_{i=1}^{n} M \mathbb{E}_{\mathcal{A}} \|\mathcal{A}(\mathbf{S}^i) - \mathcal{A}(\mathbf{S})\| \\
&\leqslant \mathbb{E}_{\mathbf{S}}[\mathrm{EmpGap}_{\mathrm{VI}}(\mathcal{A}, F_{\mathbf{S}})] + M\delta.
\end{aligned}$$

B Operator extrapolation method [34]

Suppose we want to solve VI problem associated with operator $F_k(\cdot) = F(\cdot) + \lambda_k(\cdot - w_k)$ whose (unique) solution be w^*_{k+1}. It is clear that F_k is an $(L + \lambda_k)$-Lipschitz continuous operator which is λ_k-strongly monotone as well. Denote $\kappa := \frac{\lambda_k}{L+\lambda_k} + 1$ and consider the following algorithm for solving this problem:

Algorithm 3 Operator Extrapolation (OE) method

1: Let $z_0 = z_1 = w_k$ be given.
2: **for** $t = 1, \ldots, T$ **do**
3: $\quad z_{t+1} = \mathrm{argmin}_{w \in \mathcal{W}} \frac{1}{2(L+\lambda_k)} \langle F_k(z_t) + \frac{1}{\kappa}\big[F_k(z_t) - F_k(z_{t-1})\big], w \rangle + \frac{1}{2}\|w - z_t\|^2$
4: **end for**

We have the following convergence guarantee for this algorithm:

$$\|z_T - w^*_{k+1}\|^2 \leqslant \kappa^{-T} \|z_1 - w^*_{k+1}\|^2$$

In particular, in order to ensue that $\|z_T - w^*_{k+1}\| \leqslant \frac{\nu}{LD^2 + MD + 2\lambda_k D^2}$, we require

$$T = 2\kappa \ln\Big(\frac{LD^2 + MD + 2\lambda_k D^2}{\nu}\Big) = \frac{2(L + 2\lambda_k)}{L + \lambda_k} \ln\Big(\frac{LD^2 + MD + 2\lambda_k D^2}{\nu}\Big)$$

itearations.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pp. 308–318. ACM (2016)
2. Asi, H., Feldman, V., Koren, T., Talwar, K.: Private stochastic convex optimization: optimal rates in l1 geometry. In: M. Meila, T. Zhang (eds.) Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 139, pp. 393–403. PMLR (2021)
3. Auslender, A., Teboulle, M.: Interior projection-like methods for monotone variational inequalities. Math. Program. **104**(1), 39–68 (2005)
4. Bassily, R., Feldman, V., Guzmán, C., Talwar, K.: Stability of stochastic gradient descent on nonsmooth convex losses. In: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 4381–4391. Curran Associates, Inc. (2020). https://proceedings.neurips.cc/paper/2020/file/2e2c4bf7ceaa4712a72dd5ee136dc9a8-Paper.pdf
5. Bassily, R., Feldman, V., Talwar, K., Guha Thakurta, A.: Private stochastic convex optimization with optimal rates. In: Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. (2019)
6. Bassily, R., Guzman, C., Nandi, A.: Non-euclidean differentially private stochastic convex optimization. In: M. Belkin, S. Kpotufe (eds.) Proceedings of Thirty Fourth Conference on Learning Theory, Proceedings of Machine Learning Research, vol. 134, pp. 474–499. PMLR (2021)
7. Bassily, R., Smith, A., Thakurta, A.: Differentially private empirical risk minimization: efficient algorithms and tight error bounds (2014)
8. Bousquet, O., Elisseeff, A.: Stability and generalization. J. Mach. Learn. Res. **2**, 499–526 (2002). https://doi.org/10.1162/153244302760200704
9. Bousquet, O., Klochkov, Y., Zhivotovskiy, N.: Sharper bounds for uniformly stable algorithms. In: J. Abernethy, S. Agarwal (eds.) Proceedings of Thirty Third Conference on Learning Theory, Proceedings of Machine Learning Research, vol. 125, pp. 610–626. PMLR (2020)
10. Bun, M., Steinke, T.: Concentrated differential privacy: Simplifications, extensions, and lower bounds. In: Theory of Cryptography Conference, pp. 635–658. Springer (2016)
11. Chaudhuri, K., Monteleoni, C.: Privacy-preserving logistic regression. In: NIPS (2008)
12. Chaudhuri, K., Monteleoni, C., Sarwate, A.D.: Differentially private empirical risk minimization. J. Mach. Learn. Res. **12**(Mar), 1069–1109 (2011)
13. Dong, J., Roth, A., Su, W.J., et al.: Gaussian differential privacy. J. R. Stat. Soc. B **84**(1), 3–37 (2022)
14. Duchi, J.C., Ruan, F.: The right complexity measure in locally private estimation: it is not the fisher information. arXiv preprint arXiv:1806.05756 (2018)
15. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Proceedings of the 3rd Conference on Theory of Cryptography, TCC'06, pp. 265–284. Springer, Berlin (2006). https://doi.org/10.1007/11681878_14
16. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3–4), 211–407 (2014). https://doi.org/10.1561/0400000042
17. Dwork, C., Rothblum, G.N., Vadhan, S.P.: Boosting and differential privacy. In: 51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23–26, 2010, Las Vegas, Nevada, USA, pp. 51–60. IEEE Computer Society (2010). https://doi.org/10.1109/FOCS.2010.12
18. Facchinei, F., Pang, J.S.: Finite-Dimensional Variational Inequalities and Complementarity Problems. Springer, Berlin (2007)
19. Feldman, V., Koren, T., Talwar, K.: Private stochastic convex optimization: optimal rates in linear time. In: Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, pp. 439–449. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3357713.3384335
20. Feldman, V., Mironov, I., Talwar, K., Thakurta, A.: Privacy amplification by iteration. In: M. Thorup (ed.) 59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7–9, 2018, pp. 521–532. IEEE Computer Society (2018). https://doi.org/10.1109/FOCS.2018.00056
21. Feldman, V., Vondrak, J.: High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In: A. Beygelzimer, D. Hsu (eds.) Proceedings of the Thirty-Second Conference on Learning Theory, Proceedings of Machine Learning Research, vol. 99, pp. 1270–1279. PMLR, Phoenix, USA (2019)

22. Gupta, A., Ligett, K., McSherry, F., Roth, A., Talwar, K.: Differentially private combinatorial optimization. In: M. Charikar (ed.) Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010, pp. 1106–1125. SIAM (2010)
23. Hardt, M., Recht, B., Singer, Y.: Train faster, generalize better: Stability of stochastic gradient descent. In: M.F. Balcan, K.Q. Weinberger (eds.) Proceedings of The 33rd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 48, pp. 1225–1234. PMLR, New York (2016)
24. Hsieh, Y., Iutzeler, F., Malick, J., Mertikopoulos, P.: On the convergence of single-call stochastic extra-gradient methods. In: H.M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E.B. Fox, R. Garnett (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8–14 December 2019, Vancouver, BC, Canada, pp. 6936–6946 (2019). http://papers.nips.cc/paper/8917-on-the-convergence-of-single-call-stochastic-extra-gradient-methods
25. Iusem, A.N., Jofré, A., Oliveira, R.I., Thompson, P.: Variance-based extragradient methods with line search for stochastic variational inequalities. SIAM J. Optim. **29**(1), 175–206 (2019). https://doi.org/10.1137/17M1144799
26. Jain, P., Kothari, P., Thakurta, A.: Differentially private online learning. In: 25th Annual Conference on Learning Theory (COLT), pp. 24.1–24.34 (2012)
27. Jain, P., Thakurta, A.: (near) dimension independent risk bounds for differentially private learning. In: ICML (2014)
28. Jalilzadeh, A., Shanbhag, U.V.: A proximal-point algorithm with variable sample-sizes (ppawss) for monotone stochastic variational inequality problems. In: 2019 Winter Simulation Conference (WSC), pp. 3551–3562 (2019). https://doi.org/10.1109/WSC40007.2019.9004836
29. Juditsky, A., Nemirovski, A.: First order methods for nonsmooth convex large-scale optimization, i: General Purpose Methods. MIT Press (2012)
30. Juditsky, A., Nemirovski, A.S., Tauvel, C.: Solving variational inequalities with Stochastic Mirror-Prox algorithm. Stoch. Syst. **1**(1), 17–58 (2011). https://doi.org/10.1214/10-SSY011
31. Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A.: What can we learn privately? In: 2008 49th Annual IEEE Symposium on Foundations of Computer Science, pp. 531–540. IEEE (2008)
32. Kifer, D., Smith, A., Thakurta, A.: Private convex empirical risk minimization and high-dimensional regression. In: Conference on Learning Theory, pp. 25–1 (2012)
33. Korpelevich, G.: The extragradient method for finding saddle points and other problems. Matecon **12**, 747–756 (1976)
34. Kotsalis, G., Lan, G., Li, T.: Simple and optimal methods for stochastic variational inequalities, i: operator extrapolation (2020)
35. Lei, Y., Yang, Z., Yang, T., Ying, Y.: Stability and generalization of stochastic gradient methods for minimax problems. In: International Conference on Machine Learning, pp. 6175–6186. PMLR (2021)
36. Li, T., Sanjabi, M., Beirami, A., Smith, V.: Fair resource allocation in federated learning. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020. OpenReview.net (2020). https://openreview.net/forum?id=ByexElSYDr
37. Mironov, I.: Rényi differential privacy. In: Proceedings of 30th IEEE Computer Security Foundations Symposium (CSF), pp. 263–275 (2017). https://arxiv.org/abs/1702.07476
38. Nemirovski, A.: 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 J. Optim. **15**(1), 229–251 (2004). https://doi.org/10.1137/S1052623403425629
39. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
40. Nemirovsky, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley, New York (1983)
41. Nesterov, Y.: Dual extrapolation and its applications to solving variational inequalities and related problems. Math. Program. **109**(2), 319–344 (2007)
42. Neumann, Jv.: Zur theorie der gesellschaftsspiele. Math. Ann. **100**, 295–320 (1928)
43. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control. Optim. **14**(5), 877–898 (1976)

44. Shalev-Shwartz, S., Shamir, O., Srebro, N., Sridharan, K.: Learnability, stability and uniform convergence. J. Mach. Learn. Res. **11**, 2635–2670 (2010)
45. Sion, M.: On general minimax theorems. Pac. J. Math. **8**(1), 171–176 (1958)
46. Steinke, T., Ullman, J.R.: Between pure and approximate differential privacy. J. Priv. Confident. **7**(2) (2016)
47. Ullman, J.: Private multiplicative weights beyond linear queries. In: Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, pp. 303–312. ACM (2015)
48. Vershynin, R.: High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press (2018). https://doi.org/10.1017/9781108231596
49. Williamson, R., Menon, A.: Fairness risk measures. In: K. Chaudhuri, R. Salakhutdinov (eds.) Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 97, pp. 6786–6797. PMLR (2019)
50. Zhang, J., Hong, M., Wang, M., Zhang, S.: Generalization bounds for stochastic saddle point problems. In: Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, vol. 130, pp. 568–576. PMLR (2021)