Sharper Rates for Separable Minimax and Finite Sum Optimization via Primal-Dual Extragradient Methods

Yujia Jin YUJIAJIN@STANFORD.EDU
Stanford, CA, US

Aaron Sidford SIDFORD@STANFORD.EDU
Stanford, CA, US

Kevin Tian KJTIAN@STANFORD.EDU
Stanford, CA, US

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We design accelerated algorithms with improved rates for several fundamental classes of optimization problems. Our algorithms all build upon techniques related to the analysis of primal-dual extragradient methods via relative Lipschitzness proposed recently by Cohen et al. (2021).

(1) **Separable minimax optimization.** We study separable minimax optimization problems of the form $\min_x \max_y f(x) - g(y) + h(x, y)$, where f and g have smoothness and strong convexity parameters $(L^{\mathsf{x}}, \mu^{\mathsf{x}})$, $(L^{\mathsf{y}}, \mu^{\mathsf{y}})$, and h is convex-concave with a $(\Lambda^{\mathsf{xx}}, \Lambda^{\mathsf{xy}}, \Lambda^{\mathsf{yy}})$-blockwise operator norm bounded Hessian. We provide an algorithm using $\widetilde{O}\left(\sqrt{\frac{L^{\mathsf{x}}}{\mu^{\mathsf{x}}}} + \sqrt{\frac{L^{\mathsf{y}}}{\mu^{\mathsf{y}}}} + \frac{\Lambda^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\Lambda^{\mathsf{xy}}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}} + \frac{\Lambda^{\mathsf{yy}}}{\mu^{\mathsf{y}}}\right)$ gradient queries. Notably, for convex-concave minimax problems with bilinear coupling (e.g. quadratics), where $\Lambda^{\mathsf{xx}} = \Lambda^{\mathsf{yy}} = 0$, our rate matches a lower bound of Zhang et al. (2019).

(2) **Finite sum optimization.** We study finite sum optimization problems of the form $\min_x \frac{1}{n}\sum_{i\in[n]} f_i(x)$, where each f_i is L_i-smooth and the overall problem is μ-strongly convex. We provide an algorithm using $\widetilde{O}\left(n + \sum_{i\in[n]} \sqrt{\frac{L_i}{n\mu}}\right)$ gradient queries. Notably, when the smoothness bounds $\{L_i\}_{i\in[n]}$ are non-uniform, our rate improves upon accelerated SVRG (Lin et al., 2015; Frostig et al., 2015) and Katyusha (Allen-Zhu, 2017) by up to a $\sqrt{n}$ factor.

(3) **Minimax finite sums.** We generalize our algorithms for minimax and finite sum optimization to solve a natural family of minimax finite sum optimization problems at an accelerated rate, encapsulating both above results up to a logarithmic factor.

Keywords: convex optimization, first-order methods, stochastic optimization, minimax optimization, acceleration

1. Introduction

We study several fundamental families of optimization problems, namely (separable) minimax optimization, finite sum optimization, and minimax finite sum optimization (which generalizes both). These families have received widespread recent attention from the optimization community due to their prevalence in modeling tasks arising in modern data science. For example, minimax optimization has been used in both convex-concave settings and beyond to model robustness to (possibly adversarial) noise in many training tasks (Madry et al., 2018; Rahimian and Mehrotra, 2019; Goodfellow et al., 2020). Moreover, finite sum optimization serves as a fundamental subroutine in many of the empirical risk minimization tasks of machine learning today (Bottou et al., 2018). Nonetheless, and perhaps surprisingly, there remain gaps in our understanding of the optimal rates for these problems. Toward closing these gaps, we provide new accelerated algorithms improving upon the state-of-the-art for each family of problems.

Our results build upon advances in using primal-dual extragradient methods to recover accelerated rates for smooth convex optimization in Cohen et al. (2021), which considered the problem[1]

$$\min_{x \in \mathcal{X}} f(x) + \frac{\mu}{2} \|x\|^2 \text{ for } L\text{-smooth and convex } f, \tag{1}$$

and its equivalent primal-dual formulation as an appropriate "Fenchel game"

$$\min_{x \in \mathcal{X}} \max_{x^* \in \mathcal{X}^*} \frac{\mu}{2} \|x\|^2 + \langle x^*, x \rangle - f^*(x^*), \text{ where } f^* \text{ is the convex conjugate of } f\,. \tag{2}$$

Cohen et al. (2021) showed that applying extragradient methods (Nemirovski, 2004; Nesterov, 2007) and analyzing them through a condition the paper refers to as *relative Lipschitzness* recovers an accelerated gradient query complexity for computing (1), known to be optimal (Nesterov, 2003).

Both the Fenchel game (Abernethy et al., 2018; Wang and Abernethy, 2018) and relative Lipschitzness (independently proposed in Stonyakina et al. (2020)) have a longer history, discussed in Appendix A. This work is particularly motivated by their synthesis in Cohen et al. (2021), which used these tools to give the following general recipe for designing accelerated methods.

(1) Choose a primal-dual formulation of an optimization problem and a regularizer, r.

(2) Bound iteration costs, i.e. the cost of implementing mirror steps with respect to r.

(3) Bound the relative Lipschitzness of the gradient operator of the problem with respect to r.

In Cohen et al. (2021), this recipe was applied with (2) as the primal-dual formulation and $r(x, x^*) := \frac{\mu}{2} \|x\|^2 + f^*(x^*)$. Further, it was shown that each iteration could be implemented (implicitly) with $O(1)$ gradient queries and that the gradient operator Φ of the objective (2) is $O(\sqrt{L/\mu})$-relatively Lipschitz with respect to r. Combining these ingredients gave the accelerated rate for (2); we note that additional tools were further developed in Cohen et al. (2021) for other settings including accelerated coordinate-smooth optimization (see Section 1.2).

In this paper, we broaden the primal-dual extragradient approach of Cohen et al. (2021) and add new recipes to the optimization cookbook. As a result, we obtain methods with improved rates for minimax optimization, finite sum optimization, and minimax finite sum optimization. We follow a similar recipe as Cohen et al. (2021) but change the ingredients with different primal-dual formulations, regularizers, extragradient methods, and analyses. In Sections 1.1, 1.2, and 1.3, we discuss each problem family, our results and approach, and situate them in the relevant literature. We discuss further related work not covered by this introduction in Appendix A.

1.1. Minimax optimization

In Section 2, we study separable convex-concave minimax optimization problems of the form[2]

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} F_{\text{mm}}(x, y) := f(x) + h(x, y) - g(y), \tag{3}$$

[1] Throughout, $\mathcal{X}, \mathcal{Y}$ are unconstrained, Euclidean spaces and $\|\cdot\|$ denotes the Euclidean norm (see Appendix B).

[2] Our results in Section 2 apply generally to non-twice differentiable, gradient Lipschitz h, but we use these assumptions for simplicity in the introduction. All norms are Euclidean (see Appendix B for relevant definitions).

where f is L^{x}-smooth and μ^{x}-strongly convex, g is L^{y}-smooth and μ^{y}-strongly convex, and h is convex-concave and twice-differentiable with $\left\|\nabla^2_{xx}h\right\| \le \Lambda^{\mathsf{xx}}$, $\left\|\nabla^2_{xy}h\right\| \le \Lambda^{\mathsf{xy}}$, and $\left\|\nabla^2_{yy}h\right\| \le \Lambda^{\mathsf{yy}}$. Our goal is to compute a pair of points (x, y) with bounded duality gap with respect to F_{mm}: $\mathrm{Gap}_{F_{\mathrm{mm}}}(x, y) \le \epsilon$ (we defer definitions used throughout the paper to Appendix B).

The problem family (3) contains as a special case the following family of convex-concave minimax optimization problems with bilinear coupling (with $\Lambda^{\mathsf{xx}} = \Lambda^{\mathsf{yy}} = 0$ and $\Lambda^{\mathsf{xy}} = \|A\|$):

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x) + \left(y^\top \mathbf{A}x - \langle b, y\rangle + \langle c, x\rangle\right) - g(y). \tag{4}$$

Problem (4) has been widely studied, dating at least to the classic work of Chambolle and Pock (2011), which used (4) to relax optimization with affine constraints related to imaging inverse problems. Problem (4) also encapsulates convex-concave quadratics and is used to model problems in reinforcement learning (Du et al., 2017) and decentralized optimization (Kovalev et al., 2020).

Our results. We give the following result on solving (3).

Theorem 1 (informal, cf. Theorem 14, Corollary 15) *Define* $\mathrm{Gap}_h(x, y) := \max_{y'\in\mathcal{Y}} h(x, y') - \min_{x'\in\mathcal{X}} h(x', y)$*, there is an algorithm that, given* $(x_0, y_0) \in \mathcal{X}\times\mathcal{Y}$ *satisfying* $\mathrm{Gap}_{F_{\mathrm{mm}}}(x_0, y_0) \le \epsilon_0$*, returns* (x, y) *with* $\mathrm{Gap}_{F_{\mathrm{mm}}}(x, y) \le \epsilon$ *using* T *gradient evaluations to* f*,* h*, and* g*, for*

$$T = O\left(\kappa_{\mathrm{mm}} \log\left(\frac{\kappa_{\mathrm{mm}}\epsilon_0}{\epsilon}\right)\right), \text{ with } \kappa_{\mathrm{mm}} := \sqrt{\frac{L^{\mathsf{x}}}{\mu^{\mathsf{x}}}} + \sqrt{\frac{L^{\mathsf{y}}}{\mu^{\mathsf{y}}}} + \frac{\Lambda^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\Lambda^{\mathsf{xy}}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}} + \frac{\Lambda^{\mathsf{yy}}}{\mu^{\mathsf{y}}}.$$

In the special case of (4), Theorem 1 matches a lower bound of Zhang et al. (2019), which applies to the family of quadratic minimax problems obeying our regularity bounds. More generally, Theorem 1 matches the lower bound whenever Λ^{xx} and Λ^{yy} are sufficiently small compared to other parameters, improving prior state-of-the-art rates (Wang and Li, 2020) in this regime.

By applying reductions based on explicit regularization used in Lin et al. (2020), Theorem 1 also yields analogous accelerated rates depending polynomially on the desired accuracy when either f, g, or both are not strongly convex. For conciseness, in this paper we focus on the strongly convex and strongly concave regime discussed previously in this section.

Our approach. Our algorithm for solving (3) is based on the simple observation that minimax problems with the separable structure can be effectively "decoupled" by using convex conjugation on the components f and g. In particular, following a similar recipe as the one in Cohen et al. (2021) for smooth convex optimization, we rewrite (an appropriate regularized formulation of) the problem (3) using convex conjugates as follows:

$$\min_{x\in\mathcal{X}, y^*\in\mathcal{Y}^*} \max_{y\in\mathcal{Y}, x^*\in\mathcal{X}^*} \frac{\mu^{\mathsf{x}}}{2}\|x\|^2 - \frac{\mu^{\mathsf{y}}}{2}\|y\|^2 + \langle x^*, x\rangle - \langle y^*, y\rangle + h(x, y) - f^*(x^*) + g^*(y^*).$$

This can be viewed as an equivalent reformulation of the problem (3) by simply replace $f \leftarrow f - \frac{\mu^{\mathsf{x}}}{2}\|x\|^2$ and $g \leftarrow g - \frac{\mu^{\mathsf{y}}}{2}\|y\|^2$ and using the definition of convex conjugates. Further, we define the regularizer $r(x, y, x^*, y^*) := \frac{\mu^{\mathsf{x}}}{2}\|x\|^2 + \frac{\mu^{\mathsf{y}}}{2}\|y\|^2 + f^*(x^*) + g^*(y^*)$. Finally, we apply an extragradient method for strongly monotone operators to our problem, using this regularizer. As in Cohen et al. (2021) we demonstrate efficient implementability, and analyze the relative Lipschitzness of the problem's gradient operator with respect to r, yielding Theorem 1. In the final gradient oracle complexity, our method obtains the accelerated trade-off between primal and dual blocks for $\frac{\mu^{\mathsf{x}}}{2}\left\|x\right\|^2 + \langle x^*, x\rangle - f^*(x^*)$ and $\frac{\mu^{\mathsf{y}}}{2}\left\|y\right\|^2 + \langle y^*, y\rangle - g^*(y^*)$, for the separable parts f and g respectively. It also obtains an unaccelerated rate for the h component, by bounding the relative Lipschitzness corresponding to h under our assumptions.

Prior work. Many recent works obtaining improved rates for minimax optimization under smoothness and strong convexity restrictions concentrate on a more general family of problems of the form:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} F(x,y). \tag{5}$$

Typically, these works assume for twice-differentiable F, $\nabla^2_{xx}F$ is bounded between $\mu^{\mathsf{x}}\mathbf{I}$ and $\Lambda^{\mathsf{xx}}\mathbf{I}$ everywhere, $\nabla^2_{yy}F$ is bounded between $\mu^{\mathsf{y}}\mathbf{I}$ and $\Lambda^{\mathsf{yy}}\mathbf{I}$ everywhere, and $\nabla^2_{xy}F$ is operator norm bounded by Λ^{xy}. It is straightforward to see that (5) contains (3) as a special case, by setting $f \leftarrow \frac{\mu^{\mathsf{x}}}{2}\left\|\cdot\right\|^2$, $g \leftarrow \frac{\mu^{\mathsf{y}}}{2}\left\|\cdot\right\|^2$, and $h \leftarrow F - f + g$.

For (5), under gradient access to F, the works of Lin et al. (2020); Wang and Li (2020); Cohen et al. (2021) presented different approaches yielding a variety of query complexities. Letting $\Lambda^{\max} := \max\left(\Lambda^{\mathsf{xx}}, \Lambda^{\mathsf{xy}}, \Lambda^{\mathsf{yy}}\right)$, these complexities scaled respectively as[3]

$$\widetilde{O}\left(\sqrt{\frac{\max\left(\Lambda^{\mathsf{xx}}, \Lambda^{\mathsf{xy}}, \Lambda^{\mathsf{yy}}\right)^2}{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}}\right),\ \widetilde{O}\left(\sqrt{\frac{\Lambda^{\mathsf{xx}}}{\mu^{\mathsf{x}}}} + \sqrt{\frac{\Lambda^{\mathsf{yy}}}{\mu^{\mathsf{y}}}} + \sqrt{\frac{\Lambda^{\mathsf{xy}}\Lambda^{\max}}{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}}\right),\ \widetilde{O}\left(\frac{\Lambda^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\Lambda^{\mathsf{yy}}}{\mu^{\mathsf{y}}} + \frac{\Lambda^{\mathsf{xy}}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}}\right).$$

The state-of-the-art rate (ignoring logarithmic factors) is due to Wang and Li (2020), which obtained the middle gradient query complexity above.

For the comparison, we first note that for quadratic minimax problems, i.e. $\min_x \max_y F(x,y)$, where F is convex-concave and $\nabla^2 F$ is constant, Theorem 1 obtains the optimal complexity (up to a logarithmic term). To see this, setting $f(x)$ and $g(y)$ to be quadratics in $\nabla^2_{xx}F$ and $-\nabla^2_{yy}F$, $h = F - f + g$ is bilinear and hence Theorem 1 matches the lower bound of Zhang et al. (2019) (since $\Lambda^{\mathsf{xx}} = \Lambda^{\mathsf{yy}} = 0$). Notably in this case we *improve* Wang and Li (2020) (Corollary 3, Section 5, NeurIPS version) by a $o(1)$ factor in the runtime exponent. Our method's optimality extends "for free" to cases when h is bilinear (but f and g may be non-quadratic). This setting naturally arises in (relaxed) affine-constrained optimization (and structured composite problems $f(\mathbf{A}x) + g(y)$), as well as applications in reinforcement learning and decentralized optimization. Further, if F can be decomposed as $f(x) - g(y) + h(x,y)$ where $\nabla^2_{xx}h \preceq \Lambda^{\mathsf{xx}}\mathbf{I}$ and $\nabla^2_{yy}h \preceq \Lambda^{\mathsf{yy}}\mathbf{I}$ for "small" $\Lambda^{\mathsf{xx}}, \Lambda^{\mathsf{yy}}$, i.e. $\frac{\Lambda^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\Lambda^{\mathsf{yy}}}{\mu^{\mathsf{y}}} = O(\sqrt{\frac{L^{\mathsf{x}}}{\mu^{\mathsf{x}}}} + \sqrt{\frac{L^{\mathsf{y}}}{\mu^{\mathsf{y}}}} + \frac{\Lambda^{\mathsf{xy}}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}})$, Theorem 1 matches Zhang et al. (2019), whereas Wang and Li (2020) does not (when $\max(L^{\mathsf{x}}, L^{\mathsf{y}}) \gg \Lambda^{\mathsf{xy}}$).

In the general regime where no such favorable decomposition exists and we may as well choose $f = \frac{\mu^{\mathsf{x}}}{2}\|\cdot\|^2, g = \frac{\mu^{\mathsf{y}}}{2}\|\cdot\|^2$, Theorem 1 recovers Cohen et al. (2021) but does not improve Wang and Li (2020) (short of saving logarithmic factors). This general application may improve Lin et al. (2020), e.g. in the setting when $\Lambda^{\mathsf{xx}} \gg \max(\Lambda^{\mathsf{yy}}, \Lambda^{\mathsf{xy}})$ and $\mu^{\mathsf{x}} \gg \mu^{\mathsf{y}}$ but $\frac{\Lambda^{\mathsf{xx}}}{\mu^{\mathsf{x}}} \approx \frac{\Lambda^{\mathsf{yy}}}{\mu^{\mathsf{y}}}$. Each work matches the lower bound of Zhang et al. (2019) in some (incomparable) parameter regimes.

From the algorithmic perspective, the method in Theorem 1 uses only a single loop, as opposed to the multi-loop methods in Lin et al. (2020); Wang and Li (2020) which lose logarithmic factors. It thus has an arguably simpler structure and may find advantage in practice.

[3] $\widetilde{O}$ hides logarithmic factors throughout, see Appendix B.

Concurrent work. A pair of independent and concurrent works (Kovalev et al., 2021; Thekumparampil et al., 2022) obtained variants of Theorem 1. Their results were stated under the restricted setting of bilinear coupling (4), but they each provided alternative results under (different) weakenings of strong convexity. The algorithm of Thekumparampil et al. (2022) is closer to the one developed in this paper (also going through a primal-dual lifting), although the ultimate methods and analyses are somewhat different. Though our results were obtained independently, our presentation was informed by a reading of Kovalev et al. (2021); Thekumparampil et al. (2022) for a comparison.

1.2. Finite sum optimization

In Appendix E, we study finite sum optimization problems of the form

$$\min_{x\in\mathcal{X}} F_{\text{fs}}(x) := \frac{1}{n}\sum_{i\in[n]} f_i(x), \tag{6}$$

where f_i is L_i-smooth for each $i \in [n]$, and $\frac{1}{n}\sum_{i\in[n]} f_i$ is μ-strongly convex. We focus on the strongly convex regime; through generic reductions (Zhu and Hazan, 2016), our results yield accelerated rates depending polynomially on the target accuracy, without strong convexity.

Methods for solving (6) have garnered substantial interest because of their widespread applicability to empirical risk minimization problems over a dataset of n points, which encapsulate a variety of (generalized) regression problems in machine learning (see Bottou et al. (2018)).

Our results. We give the following result on solving (6).

Theorem 2 (informal, cf. Theorem 25, Corollary 27) *There is an algorithm that, given $x_0 \in \mathcal{X}$ satisfying $F_{\text{fs}}(x_0) - F_{\text{fs}}(x_\star) \le \epsilon_0$ where $x_\star$ minimizes F_{fs}, returns $x \in \mathcal{X}$ with $\mathbb{E}F_{\text{fs}}(x) - F_{\text{fs}}(x_\star) \le \epsilon$ using T gradient evaluations (each to some f_i) for*

$$T = O\left(\kappa_{\text{fs}} \log\left(\frac{\kappa_{\text{fs}}\epsilon_0}{\epsilon}\right)\right), \text{ with } \kappa_{\text{fs}} := n + \sum_{i\in[n]} \frac{\sqrt{L_i}}{\sqrt{n\mu}}.$$

Our approach. Our algorithm for solving (6) builds upon an accelerated coordinate descent method developed in Cohen et al. (2021), for which it used an analysis of a randomized extragradient method. We consider an equivalent primal-dual formulation of (a regularized variant of) (6), inspired by analogous developments in the ERM literature (Shalev-Shwartz and Zhang, 2013, 2016):

$$\min_{x\in\mathcal{X}} \max_{\{x_i^*\}_{i\in[n]}\subset\mathcal{X}^*} \frac{\mu}{2}\|x\|^2 + \frac{1}{n}\sum_{i\in[n]} \left(\langle x_i^*, x\rangle - f_i^*(x_i^*)\right).$$

Our algorithm then solves this regularized primal-dual game to high precision.

A key building block of our method is a randomized extragradient method which is compatible with strongly monotone problems. To this end, we extend the randomized extragradient method in Cohen et al. (2021) to also obtain high-precision guarantees under strong monotonicity. We proceed as follows: for roughly κ_{fs} iterations (defined in Theorem 2) of our method, we run the non-strongly monotone randomized mirror prox method of Cohen et al. (2021) to obtain a regret bound. We then subsample a random iterate, which we show halves an appropriate potential in expectation via our regret bound and strong monotonicity; recursing this procedure yields our high-precision solver.

Prior work. Developing accelerated algorithms for (6) under our regularity assumptions has been the subject of a substantial amount of research effort in the community, see e.g. Lin et al. (2015); Frostig et al. (2015); Shalev-Shwartz and Zhang (2016); Allen-Zhu (2017); Song et al. (2020) and references therein. The particular approach of combining coordinate methods with primal-dual formulations and its application to the ERM problem has also appeared in a variety of literature (Zhang and Lin, 2015; Chambolle et al., 2018; Alacaoglu et al., 2020; Song et al., 2021). Previously, the state-of-the-art gradient query complexities (up to logarithmic factors) for (6) were obtained by Lin et al. (2015); Frostig et al. (2015); Allen-Zhu (2017); Song et al. (2020),[4] and scaled as

$$\widetilde{O}\left(n + \sqrt{\frac{\sum_{i\in[n]} L_i}{\mu}}\right). \tag{7}$$

Rates such as (7), which scale as functions of $\sum_{i\in[n]} \frac{L_i}{\mu}$, arise in known *variance reduction*-based approaches (Johnson and Zhang, 2013; Defazio et al., 2014; Schmidt et al., 2017; Allen-Zhu, 2017) due to their applications of a "dual strong convexity" lemma (e.g. Theorem 1, Johnson and Zhang (2013) or Lemma 2.4, Allen-Zhu (2017)) of the form

$$\left\|\nabla f_i(x) - \nabla f_i(\bar{x})\right\|^2 \le 2L_i \left(f_i(\bar{x}) - f_i(x) - \langle \nabla f_i(x), \bar{x} - x\rangle\right).$$

The analyses of e.g. Johnson and Zhang (2013); Allen-Zhu (2017) sample $i \in [n]$ proportional to L_i, yielding variance bounds on a gradient estimator by a quantity related to the F_{fs} divergence.

The rate in (7) is known to be optimal in the uniform smoothness regime (Woodworth and Srebro, 2016), but in a more general setting its optimality is unclear. Theorem 2 shows that the rate can be improved for sufficiently non-uniform L_i, which may happen e.g. in regression with a matrix $\mathbf{A}$ that has non-uniform row norms. In particular, Cauchy-Schwarz shows that the quantity κ_{fs} is never worse than (7), and improves upon it by a factor asymptotically between 1 and $\sqrt{n}$ when the $\{L_i\}_{i\in[n]}$ are non-uniform. The best improvement of $\sqrt{n}$ is achievable in, e.g. extreme cases when $\exists i \in [n]$ with $L_j \approx 0$, $\forall j \neq i$. Moreover, even in the uniform smoothness case, Theorem 2 matches the tightest rate in Allen-Zhu (2017) up to an additive $\log \kappa_{\text{fs}}$ term, as opposed to an additional multiplicative logarithmic overhead incurred by the reduction-based approaches of Lin et al. (2015); Frostig et al. (2015).

Our rate's improvement over (7) is comparable to a similar improvement that was achieved previously in the literature on coordinate descent methods. In particular, Lee and Sidford (2013) first obtained a (generalized) partial derivative query complexity comparable to (7) under coordinate smoothness bounds, which was later improved to a query complexity comparable to Theorem 2 by Zhu et al. (2016); Nesterov and Stich (2017). Due to connections between coordinate-smooth optimization and empirical risk minimization (ERM) previously noted in the literature (Shalev-Shwartz and Zhang, 2013, 2016), it is natural to conjecture that the rate in Theorem 2 is achieveable for finite sums (6) as well. However, prior to our work (to our knowledge) this rate was not known, except in special cases e.g. linear regression (Agarwal et al., 2020).

[4] There have been a variety of additional works which have also attained accelerated rates for either the problem (6) or its ERM specialization, see e.g. Defazio (2016); Zhang and Xiao (2017); Lan et al. (2019); Zhou et al. (2019). However, to the best of our knowledge these do not improve upon the state-of-the-art rate of Allen-Zhu (2017) in our setting.

From the algorithmic perspective, our basic Algorithm 4 and Algorithm 1 of Allen-Zhu (2017) both are "double loop" as they aggregate information every $\approx O(n)$ iterations; we acknowledge Algorithm 6 adds one loop, but point out the resulting complexity is only affected by a constant factor. We agree finding a more direct approach is an interesting future direction.

Our method is based on using a primal-dual formulation of (6) to design our gradient estimators. It attains Theorem 25 by sampling summands proportional to $\sqrt{L_i}$, trading off primal and dual variances through a careful coupling. It can be viewed as a modified dual formulation to the coordinate descent algorithm in Cohen et al. (2021), which used primal-dual couplings inspired by Zhu et al. (2016); Nesterov and Stich (2017). We believe our result sheds further light on the duality between coordinate-smooth and finite sum optimization, and gives an interesting new acceleration approach for finite sum problems via algorithmically leveraging their primal-dual formulations.

1.3. Minimax finite sum optimization

In Appendix F, we study a family of minimax finite sum optimization problems of the form

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} F_{\text{mmfs}}(x,y) := \frac{1}{n}\sum_{i\in[n]} \left(f_i(x) + h_i(x,y) - g_i(y)\right). \tag{8}$$

We assume f_i is L_i^{x}-smooth, g_i is L^{y}-smooth, and h_i is convex-concave and twice-differentiable with blockwise operator norm bounds Λ_i^{xx}, Λ_i^{xy}, and Λ_i^{yy} for each $i \in [n]$. We also assume the whole problem is μ^{x}-strongly convex and μ^{y}-strongly concave.

We propose the family (8) because it encapsulates (5) and (6), and is amenable to techniques from solving both. Moreover, (8) is a natural description of instances of (5) which arise from primal-dual formulations of ERM problems, e.g. Zhang and Xiao (2017); Wang and Xiao (2017). It also generalizes natural minimax finite sum problems previously considered in e.g. Carmon et al. (2019).

Our results. We give the following result on solving (8).

Theorem 3 (informal, cf. Theorem 39, Corollary 41) *There is an algorithm that, given* $(x_0, y_0) \in \mathcal{X}\times\mathcal{Y}$ *satisfying* $\text{Gap}_{F_{\text{mmfs}}}(x_0,y_0) \le \epsilon_0$, *returns* (x,y) *with* $\mathbb{E}\text{Gap}_{F_{\text{mmfs}}}(x,y) \le \epsilon$, *using* T *gradient evaluations, each to some* f_i, g_i, *or* h_i, *where*

$$T = O\left(\kappa_{\text{mmfs}} \log\left(\kappa_{\text{mmfs}}\right) \log\left(\frac{\kappa_{\text{mmfs}}\epsilon_0}{\epsilon}\right)\right),$$

$$\textit{with } \kappa_{\text{mmfs}} := n + \frac{1}{\sqrt{n}}\sum_{i\in[n]}\left(\sqrt{\frac{L_i^{\mathsf{x}}}{\mu^{\mathsf{x}}}} + \sqrt{\frac{L_i^{\mathsf{y}}}{\mu^{\mathsf{y}}}} + \frac{\Lambda_i^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\Lambda_i^{\mathsf{xy}}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}} + \frac{\Lambda_i^{\mathsf{yy}}}{\mu^{\mathsf{y}}}\right).$$

The rate in Theorem 3 captures (up to a logarithmic factor) both of the rates in Theorems 1 and 2, when (8) is appropriately specialized. It can be more generally motivated as follows. When n is not the dominant term in Theorem 2's bound, the remaining term is $\sqrt{n}$ times the average rate attained by Nesterov's accelerated gradient method (Nesterov, 1983) on each summand in (6). This improves upon the factor of n overhead which one might naively expect from computing full gradients. In similar fashion, Theorem 3 attains a rate (up to an additive n, and logarithmic factors) which is $\sqrt{n}$ times the average rate attained by Theorem 1 on each summand in (8).

Our approach. Our algorithm for solving (8) is a natural synthesis of the algorithms suggested in Sections 1.1 and 1.2. However, to obtain our results we apply additional techniques to bypass complications which arise from the interplay between the minimax method and the finite sum method, inspired by Carmon et al. (2019). In particular, to obtain our tightest rate we would like to subsample the components in our gradient operator corresponding to $\{f_i\}_{i\in[n]}, \{g_i\}_{i\in[n]}, \{h_i\}_{i\in[n]}$ all at different frequencies when applying the randomized extragradient method. These different sampling distributions introduce dependencies between iterates, making our randomized estimators no longer "unbiased" for the true gradient operator.

To circumvent this difficulty, we obtain our result via a partial decoupling, treating components corresponding to $\{f_i\}_{i\in[n]}$, $\{g_i\}_{i\in[n]}$ and those corresponding to $\{h_i\}_{i\in[n]}$ separately. For the first two aforementioned components, which are separable and hence do not interact, we pattern an expected relative Lipschitzness analysis for each block, similar to the finite sum optimization. For the remaining component $\{h_i\}_{i\in[n]}$, we develop a variance-reduced stochastic method which yields a relative variance bound. We put these pieces together in Proposition 29, a new randomized extragradient method analysis, to give a method with a convergence rate of roughly

$$n + \frac{1}{\sqrt{n}}\sum_{i\in[n]}\left(\sqrt{\frac{L_i^{\mathsf{x}}}{\mu^{\mathsf{x}}}} + \sqrt{\frac{L_i^{\mathsf{y}}}{\mu^{\mathsf{y}}}}\right) + (\kappa^h_{\text{mmfs}})^2, \text{ where } \kappa^h_{\text{mmfs}} := \frac{1}{n}\sum_{i\in[n]}\left(\frac{\Lambda_i^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\Lambda_i^{\mathsf{xy}}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}} + \frac{\Lambda_i^{\mathsf{yy}}}{\mu^{\mathsf{y}}}\right).$$

The dependence on all pieces above is the same as in Theorem 3, except for the term corresponding to the $\{h_i\}_{i\in[n]}$. We finally wrap our solver in an "outer loop" proximal point method which solves a sequence of γ-regularized variants of (8). This outer loop does not harm the accelerated rate obtained for $\{f_i\}_{i\in[n]}$ and $\{g_i\}_{i\in[n]}$ since the regularization does not change the relative condition number of the separable components. It further allows us to trade off the terms n and $(\kappa^h_{\text{mmfs}})^2$ through our choice of γ, which yields the accelerated convergence rate of Theorem 3.

Prior work. To our knowledge, there have been relatively few results for solving (8) under our fine-grained assumptions on problem regularity, although various stochastic minimax algorithms have been developed in natural settings (Juditsky et al., 2011; Palaniappan and Bach, 2016; Hsieh et al., 2019; Carmon et al., 2019; Chavdarova et al., 2019; Carmon et al., 2020; Alacaoglu and Malitsky, 2021; Zhao, 2022). For the general problem of solving $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\frac{1}{n}\sum_{i\in[n]}F_i(x,y)$ where F_i is L_i-smooth and convex-concave, and the whole problem is μ^{x}-strongly convex and μ^{y}-strongly concave, perhaps the most direct comparisons are Section 5.4 of Carmon et al. (2019) and Theorem 15 of Tominin et al. (2021). In particular, Carmon et al. (2019) provided a high-precision solver using roughly $\widetilde{O}\left(n + \frac{1}{\sqrt{n}}\sum_{i\in[n]}\frac{L_i}{\mu}\right)$ gradient queries, when $\mu^{\mathsf{x}} = \mu^{\mathsf{y}} = \mu$. This is recovered by Theorem 39 in the special setting of $f_i = g_i \leftarrow 0$, $\mu^{\mathsf{x}} = \mu^{\mathsf{y}} \leftarrow \mu$, and $\Lambda_i^{\mathsf{xx}} = \Lambda_i^{\mathsf{xy}} = \Lambda_i^{\mathsf{yy}} \leftarrow L_i$. More generally, Carmon et al. (2019) gave a result depending polynomially on the accuracy without the strongly convex and strongly concave assumptions, which follows from a variant of Theorem 39 after applying the explicit regularization in Lin et al. (2020) that reduces to the strongly convex-concave case.

Moreover, Theorem 15 of Tominin et al. (2021) provided a high-precision solver using roughly

$$\widetilde{O}\left(n + \frac{1}{\sqrt{n}}\sum_{i\in[n]}\frac{L_i}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}}\right)$$

gradient queries. Our work recovers (and sharpens dependences in) this result for minimax finite sum problems where each summand has the bilinear coupling (4). In the more general setting where each summand only has a uniform smoothness bound, one can interpret the Tominin et al. (2021) result as a finite sum analog of the main claim in Lin et al. (2020), which is incomparable to our Theorem 1. In a similar way, the rate of Tominin et al. (2021) is incomparable to Theorem 3, and each improves upon the other in different parameter regimes. We believe designing a single algorithm which obtains the best of both worlds for (8) is an interesting future direction.

Paper organization. In the remainder of the main body, we provide an overview of our techniques by presenting our main algorithm for proving Theorem 1 and its analysis for minimax optimization. We defer helper proofs used in Section 2, proofs of Theorem 2 for finite sum optimization, proofs of Theorem 3 for minimax finite sum optimization to the appendices. We provide abbreviated preliminaries here and defer more detailed preliminaries to Appendix B.

General notation. We use $\widetilde{O}$ to hide logarithmic factors in problem parameters, $\mathcal{X}$ and $\mathcal{Y}$ to represent Euclidean spaces, and $\|\cdot\|$ for the Euclidean norm. We refer to blocks of $z \in \mathcal{X} \times \mathcal{Y}$ by $(z^{\mathsf{x}}, z^{\mathsf{y}})$. The *Bregman divergence* in differentiable, convex r is $V_x^r(x') := r(x') - r(x) - \langle \nabla r(x), x' - x\rangle$, for any $x, x' \in \mathcal{X}$. When we omit superscripts, $r = \frac{1}{2}\|\cdot\|^2$ so $V_x(x') = \frac{1}{2}\|x - x'\|^2$.

Functions and operators. We say $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is convex-concave if $h(\cdot, y)$ and $h(x, \cdot)$ are respectively convex and concave, for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The duality gap of (x, y) is $\mathrm{Gap}_h(x, y) := \max_{y' \in \mathcal{Y}} h(x, y') - \min_{x' \in \mathcal{X}} h(x', y)$; a saddle point is $(x_\star, y_\star)$ with zero duality gap. We call operator $\Phi : \mathcal{Z} \to \mathcal{Z}^*$ monotone if $\langle \Phi(z) - \Phi(z'), z - z'\rangle \geq 0$ for all $z, z' \in \mathcal{Z}$. The convex conjugate of $f : \mathcal{X} \to \mathbb{R}$ is defined as $f^*(x^*) := \max_{x \in \mathcal{X}} \langle x, x^*\rangle - f(x)$. We define the proximal operation in r by

$$\mathrm{Prox}_x^r(\Phi) := \mathrm{argmin}_{x' \in \mathcal{X}} \left\{ \langle \Phi, x' \rangle + V_x^r(x') \right\}.$$

Regularity. Function $f : \mathcal{X} \to \mathbb{R}$ is L-smooth if $\|\nabla f(x) - \nabla f(x')\| \leq L\|x - x'\|$ for all $x, x' \in \mathcal{X}$. Differentiable $f : \mathcal{X} \to \mathbb{R}$ is μ-strongly convex if $V_x^f(x') \geq \frac{\mu}{2}\|x - x'\|^2$ for all $x, x' \in \mathcal{X}$. Operator $\Phi : \mathcal{Z} \to \mathcal{Z}^*$ is m-strongly monotone with respect to convex $r : \mathcal{Z} \to \mathbb{R}$ if for all $z, z' \in \mathcal{Z}$, $\langle \Phi(z) - \Phi(z'), z - z'\rangle \geq m\langle \nabla r(z) - \nabla r(z'), z - z'\rangle = m\left(V_z^r(z') + V_{z'}^r(z)\right)$.

2. Minimax optimization

In this section, we provide efficient algorithms for computing an approximate saddle point of

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} F_{\mathrm{mm}}(x, y) \text{ for } F_{\mathrm{mm}} := f(x) + h(x, y) - g(y). \tag{9}$$

Here and throughout this section $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ are differentiable, convex functions and $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a differentiable, convex-concave function. For the remainder, we focus on algorithms for solving the following regularized formulation of (9):

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} F_{\text{mm-reg}}(x, y) \text{ for } F_{\text{mm-reg}}(x, y) := f(x) + h(x, y) - g(y) + \frac{\mu^{\mathsf{x}}}{2}\|x\|^2 - \frac{\mu^{\mathsf{y}}}{2}\|y\|^2. \tag{10}$$

To instead solve an instance of (9) where f is μ^{x}-strongly convex and g is μ^{y}-strongly convex, we may instead equivalently solve (10) by reparameterizing $f \leftarrow f - \frac{\mu^{\mathsf{x}}}{2}\|\cdot\|^2$, $g \leftarrow g - \frac{\mu^{\mathsf{y}}}{2}\|\cdot\|^2$. As it is notationally convenient for our analysis, we focus on solving the problem (10) and then give the results for (9) at the end of this section in Corollary 15.

In designing methods for solving (10) we make the following additional regularity assumptions.

Assumption 1 (Minimax regularity) *We assume the following about* (10).

(1) f is L^{x}-smooth and g is L^{y}-smooth.

(2) h has the following blockwise-smoothness properties: for all $u, v \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned} \|\nabla_x h(u) - \nabla_x h(v)\| &\le \Lambda^{\mathsf{xx}} \|u^{\mathsf{x}} - v^{\mathsf{x}}\| + \Lambda^{\mathsf{xy}} \|u^{\mathsf{y}} - v^{\mathsf{y}}\|, \\ \|\nabla_y h(u) - \nabla_y h(v)\| &\le \Lambda^{\mathsf{xy}} \|u^{\mathsf{x}} - v^{\mathsf{x}}\| + \Lambda^{\mathsf{yy}} \|u^{\mathsf{y}} - v^{\mathsf{y}}\|. \end{aligned} \tag{11}$$

Note that when h is twice-differentiable, (11) equates to everywhere operator norm bounds on blocks of $\nabla^2 h$. Namely, for all $w \in \mathcal{X} \times \mathcal{Y}$,

$$\left\|\nabla^2_{xx} h(w)\right\| \le \Lambda^{\mathsf{xx}},\ \left\|\nabla^2_{xy} h(w)\right\| \le \Lambda^{\mathsf{xy}}, \text{ and } \left\|\nabla^2_{yy} h(w)\right\| \le \Lambda^{\mathsf{yy}}.$$

In the particular case when $h(x,y) = y^\top \mathbf{A} x - b^\top y + c^\top x$ is bilinear, clearly $\Lambda^{\mathsf{xx}} = \Lambda^{\mathsf{yy}} = 0$ (as remarked in the introduction). In this case, we may then set $\Lambda^{\mathsf{xy}} := \|\mathbf{A}\|$.

The remainder of this section is organized as follows. In Section 2.1, we state a primal-dual formulation of (10) which we apply our methods to, and prove its equivalence to (10). In Section 2.2, we give our algorithm and show it is efficiently implementable. In Section 2.3, we give the convergence rate of our algorithm. In Section 2.4, we state and prove our main result, Theorem 14. We defer all omitted proofs in this section to Appendix D.

2.1. Setup

To solve (10), we will instead find a saddle point to the expanded primal-dual function

$$F_{\text{mm-pd}}(z) := \left\langle z^{\mathsf{f}^*}, z^{\mathsf{x}} \right\rangle - \left\langle z^{\mathsf{g}^*}, z^{\mathsf{y}} \right\rangle + \frac{\mu^{\mathsf{x}}}{2} \|z^{\mathsf{x}}\|^2 - \frac{\mu^{\mathsf{y}}}{2} \|z^{\mathsf{y}}\|^2 + h(z^{\mathsf{x}}, z^{\mathsf{y}}) - f^*(z^{\mathsf{f}^*}) + g^*(z^{\mathsf{g}^*}). \tag{12}$$

We denote the domain of $F_{\text{mm-pd}}$ by $\mathcal{Z} := \mathcal{X} \times \mathcal{Y} \times \mathcal{X}^* \times \mathcal{Y}^*$. For $z \in \mathcal{Z}$, we refer to its blocks by $(z^{\mathsf{x}}, z^{\mathsf{y}}, z^{\mathsf{f}^*}, z^{\mathsf{g}^*})$. The primal-dual function $F_{\text{mm-pd}}$ is related to $F_{\text{mm-reg}}$ in the following way.

Lemma 4 *Let $z_\star$ be the saddle point to* (12)*. Then, $(z_\star^{\mathsf{x}}, z_\star^{\mathsf{y}})$ is a saddle point to* (10).

We next define Φ, the gradient operator of $F_{\text{mm-pd}}$. Before doing so, it will be convenient to define $r : \mathcal{Z} \to \mathbb{R}$, which combines the (unsigned) separable components of $F_{\text{mm-pd}}$:

$$r(z) := \frac{\mu^{\mathsf{x}}}{2} \|z^{\mathsf{x}}\|^2 + \frac{\mu^{\mathsf{y}}}{2} \|z^{\mathsf{y}}\|^2 + f^*(z^{\mathsf{f}^*}) + g^*(z^{\mathsf{g}^*}). \tag{13}$$

The function r will also serve as a regularizer in our algorithm. With this definition, we decompose Φ into three parts, roughly corresponding to the contributions from r, the bilinear portions of primal-dual representations, and h. In particular, we define

$$\begin{aligned} \Phi^r(z) &:= \nabla r(z) = \left(\mu^{\mathsf{x}} z^{\mathsf{x}}, \mu^{\mathsf{y}} z^{\mathsf{y}}, \nabla f^*(z^{\mathsf{f}^*}), \nabla g^*(z^{\mathsf{g}^*})\right) \\ \Phi^{\text{bilin}}(z) &:= (z^{\mathsf{f}^*}, z^{\mathsf{g}^*}, -z^{\mathsf{x}}, -z^{\mathsf{y}}), \\ \Phi^h(z) &:= (\nabla_x h(z^{\mathsf{x}}, z^{\mathsf{y}}), -\nabla_y h(z^{\mathsf{x}}, z^{\mathsf{y}}), 0, 0). \end{aligned} \tag{14}$$

It is straightforward to check that Φ, the gradient operator of $F_{\text{mm-pd}}$, satisfies

$$\Phi(z) := \Phi^r(z) + \Phi^{\text{bilin}}(z) + \Phi^h(z). \tag{15}$$

Finally, we note that by construction Φ is 1-strongly monotone with respect to r.

Lemma 5 (Strong monotonicity) *The operator Φ (as defined in* (15)*) is 1-strongly monotone with respect to the function $r : \mathcal{Z} \to \mathbb{R}$ as in* (13).

2.2. Algorithm

Our algorithm will be an instantiation of *strongly monotone mirror prox* (Cohen et al., 2021) stated as Algorithm 1, an alternative to the mirror prox algorithm in Nemirovski (2004) and the Halpern iteration method in Diakonikolas (2020).

Algorithm 1: SM-MIRROR-PROX(λ, T, z_0): Strongly monotone mirror prox Cohen et al. (2021)

```
Input: Convex $r : \mathcal{Z} \to \mathbb{R}$, $m$-strongly monotone $\Phi : \mathcal{Z} \to \mathcal{Z}^*$ (with respect to $r$), $z_0 \in \mathcal{Z}$
Parameter(s): $\lambda > 0, T \in \mathbb{N}$
for $0 \le t < T$ do
    $z_{t+1/2} \leftarrow \text{Prox}^r_{z_t}(\frac{1}{\lambda}\Phi(z_t))$
    $z_{t+1} \leftarrow \text{argmin}_{z \in \mathcal{Z}}\{\frac{1}{\lambda}\langle\Phi(z_{t+1/2}), z\rangle + \frac{m}{\lambda}V^r_{z_{t+1/2}}(z) + V^r_{z_t}(z)\}$
end
```

In order to analyze Algorithm 1, we need to introduce a definition from Cohen et al. (2021).

Definition 6 (Relative Lipschitzness) *We say operator $\Phi : \mathcal{Z} \to \mathcal{Z}^*$ is λ-relatively Lipschitz with respect to convex $r : \mathcal{Z} \to \mathbb{R}$ over $\mathcal{Z}_{\text{alg}} \subseteq \mathcal{Z}$ if for every three $z, w, u \in \mathcal{Z}_{\text{alg}}$,*

$$\langle \Phi(w) - \Phi(z), w - u \rangle \le \lambda \left(V^r_z(w) + V^r_w(u) \right).$$

As an example of the above definition, we have the following bound when $\Phi = \nabla r$, which follows directly from nonnegativity of Bregman divergences and (18).

Lemma 7 *Let $r : \mathcal{Z} \to \mathbb{R}$ be convex. Then, ∇r is 1-relatively Lipschitz with respect to r over $\mathcal{Z}$.*

As another example, Cohen et al. (2021) shows that if Φ is L-Lipschitz and r is μ-strongly convex (the setup considered in Nemirovski (2004)), then Φ is $\frac{L}{\mu}$-relatively Lipschitz with respect to r over $\mathcal{Z}$. This was generalized by Cohen et al. (2021) via Definition 6, who showed the following.

Proposition 8 (Proposition 3, Cohen et al. (2021)) *If Φ is λ-relatively Lipschitz with respect to r over $\mathcal{Z}_{\text{alg}}$ containing all iterates of Algorithm 1, and its VI is solved by $z_\star$, Algorithm 1 satisfies*

$$V^r_{z_t}(z_\star) \le \left(1 - \frac{m}{\lambda}\right)^t V^r_{z_0}(z_\star), \textit{ for all } t \in [T].$$

Our algorithm for minimax optimization, Algorithm 2, will simply apply Algorithm 1 to the operator-regularizer pair (Φ, r) defined in (15) and (13). Crucially, by using properties of convex conjugates, we demonstrate that one can efficiently implement the steps which solved linearized problems regularized by r. To do so, we implicitly maintain all dual iterates (in $\mathcal{X}^*, \mathcal{Y}^*$) as appropriate gradients of primal points (in $\mathcal{X}$, $\mathcal{Y}$). We give this implementation as pseudocode in Algorithm 2, and show that it is a correct implementation of Algorithm 1 in the following lemma.

Lemma 9 *Algorithm 2 implements Algorithm 1 with $m = 1$ on (Φ, r) defined in (15), (13).*

Algorithm 2: MINIMAX-SOLVE$(F_{\text{mm-reg}}, x_0, y_0)$: Separable minimax optimization

```
Input: (10) satisfying Assumption 1, $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$
Parameter(s): $\lambda > 0, T \in \mathbb{N}$
$(z_0^{\mathsf{x}}, z_0^{\mathsf{y}}) \leftarrow (x_0, y_0)$, $(z_0^{\mathsf{f}}, z_0^{\mathsf{g}}) \leftarrow (x_0, y_0)$
for $0 \le t < T$ do
    $\Phi^{\mathsf{x}} \leftarrow \mu^{\mathsf{x}} z_t^{\mathsf{x}} + \nabla f(z_t^{\mathsf{f}}) + \nabla_x h(z_t^{\mathsf{x}}, z_t^{\mathsf{y}}), \quad \Phi^{\mathsf{y}} \leftarrow \mu^{\mathsf{y}} z_t^{\mathsf{y}} + \nabla g(z_t^{\mathsf{g}}) - \nabla_y h(z_t^{\mathsf{x}}, z_t^{\mathsf{y}}).$
    $z_{t+1/2}^{\mathsf{x}} \leftarrow z^{\mathsf{x}} - \frac{1}{\lambda \mu^{\mathsf{x}}} \Phi^{\mathsf{x}}$ and $z_{t+1/2}^{\mathsf{y}} \leftarrow z^{\mathsf{y}} - \frac{1}{\lambda \mu^{\mathsf{y}}} \Phi^{\mathsf{y}}$
    $z_{t+1/2}^{\mathsf{f}} \leftarrow (1 - \frac{1}{\lambda}) z_t^{\mathsf{f}} + \frac{1}{\lambda} z_t^{\mathsf{x}}$ and $z_{t+1/2}^{\mathsf{g}} \leftarrow (1 - \frac{1}{\lambda}) z_t^{\mathsf{g}} + \frac{1}{\lambda} z_t^{\mathsf{y}}$
    $\Phi^{\mathsf{x}} \leftarrow \mu^{\mathsf{x}} z_{t+1/2}^{\mathsf{x}} + \nabla f(z_{t+1/2}^{\mathsf{f}}) + \nabla_x h(z_{t+1/2}^{\mathsf{x}}, z_{t+1/2}^{\mathsf{y}}),$
    $\Phi^{\mathsf{y}} \leftarrow \mu^{\mathsf{y}} z_{t+1/2}^{\mathsf{y}} + \nabla g(z_{t+1/2}^{\mathsf{g}}) - \nabla_y h(z_{t+1/2}^{\mathsf{x}}, z_{t+1/2}^{\mathsf{y}}).$
    $z_{t+1}^{\mathsf{x}} \leftarrow \frac{1}{1+\lambda} z_{t+1/2}^{\mathsf{x}} + \frac{\lambda}{1+\lambda} z_t^{\mathsf{x}} - \frac{1}{(1+\lambda)\mu^{\mathsf{x}}} \Phi^{\mathsf{x}}$ and $z_{t+1}^{\mathsf{y}} \leftarrow \frac{1}{1+\lambda} z_{t+1/2}^{\mathsf{y}} + \frac{\lambda}{1+\lambda} z_t^{\mathsf{y}} - \frac{1}{(1+\lambda)\mu^{\mathsf{y}}} \Phi^{\mathsf{y}}$
    $z_{t+1}^{\mathsf{f}} \leftarrow \frac{\lambda}{1+\lambda} z_t^{\mathsf{f}} + \frac{1}{1+\lambda} z_{t+1/2}^{\mathsf{x}}$ and $z_{t+1}^{\mathsf{g}} \leftarrow \frac{\lambda}{1+\lambda} z_t^{\mathsf{g}} + \frac{1}{1+\lambda} z_{t+1/2}^{\mathsf{y}}$
end
```

In particular, the proof of Lemma 9 shows that Algorithm 2 preserves the invariants that $z_t^{\mathsf{f}^*} = \nabla f(z_t^{\mathsf{f}})$ and $z_t^{\mathsf{g}^*} = \nabla g(z_t^{\mathsf{g}})$, where z_t^{f} and z_t^{g} are defined in Algorithm 2 (a similar invariant holds for each $z_{t+1/2}$). As a corollary, we have the following characterization of our iterates, recalling the definitions of $\mathcal{X}_f^*$ and $\mathcal{Y}_g^*$ from Appendix B.

Corollary 10 *Define the product space $\mathcal{Z}_{\text{alg}} := \mathcal{X} \times \mathcal{Y} \times \mathcal{X}_f^* \times \mathcal{Y}_g^*$, where $\mathcal{X}_f^* := \{\nabla f(x) \mid x \in \mathcal{X}\}$ and $\mathcal{Y}_g^* := \{\nabla g(y) \mid y \in \mathcal{Y}\}$. Then all iterates of Algorithm 2 lie in $\mathcal{Z}_{\text{alg}}$.*

More generally, for $z \in \mathcal{Z}_{\text{alg}}$, we define $z^{\mathsf{f}} := \nabla f^*(z^{\mathsf{f}^*})$ and $z^{\mathsf{g}} := \nabla g^*(z^{\mathsf{g}^*})$ (see Fact 1).

2.3. Convergence analysis

In order to use Proposition 8 to analyze Algorithm 2, we require a strong monotonicity bound and a relative Lipschitzness bound on the pair (Φ, r); the former is already given by Lemma 5. We state the latter bound, which we prove using consequences of Assumption 1 shown in Lemma 16.

Lemma 11 (Relative Lipschitzness) *Define $\Phi : \mathcal{Z} \to \mathcal{Z}^*$ as in (15), and define $r : \mathcal{Z} \to \mathbb{R}$ as in (13). Then Φ is λ-relatively Lipschitz with respect to r over $\mathcal{Z}_{\text{alg}}$ defined in Corollary 10 for*

$$\lambda = 1 + \sqrt{\frac{L^{\mathsf{x}}}{\mu^{\mathsf{x}}}} + \sqrt{\frac{L^{\mathsf{y}}}{\mu^{\mathsf{y}}}} + \frac{\Lambda^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\Lambda^{\mathsf{xy}}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}} + \frac{\Lambda^{\mathsf{yy}}}{\mu^{\mathsf{y}}}. \tag{16}$$

Finally, we provide simple bounds regarding initialization and termination of Algorithm 2.

Lemma 12 *Let $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$, and define $z_0 := (x_0, y_0, \nabla f(x_0), \nabla g(y_0))$. Suppose we have* $\text{Gap}_{F_{\text{mm-reg}}}(x_0, y_0) \le \epsilon_0$. *Then, letting $z_\star$ be the solution to (12),*

$$V_{z_0}^r(z_\star) \le \left(1 + \frac{L^{\mathsf{x}}}{\mu_x} + \frac{L^{\mathsf{y}}}{\mu^{\mathsf{y}}}\right) \epsilon_0.$$

Lemma 13 *Let $z \in \mathcal{Z}$ have*

$$V_z^r(z_\star) \le \left(\frac{\mu^{\mathsf{x}} + L^{\mathsf{x}} + \Lambda^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\mu^{\mathsf{y}} + L^{\mathsf{y}} + \Lambda^{\mathsf{yy}}}{\mu^{\mathsf{y}}} + \frac{(\Lambda^{\mathsf{xy}})^2}{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}\right) \cdot \frac{\epsilon}{2},$$

for $z_\star$ the solution to (12). Then,

$$\mathrm{Gap}_{F_{\text{mm-reg}}}(z^{\mathsf{x}}, z^{\mathsf{y}}) \le \epsilon.$$

2.4. Main result

We now state and prove our main claim.

Theorem 14 *Suppose $F_{\text{mm-reg}}$ in (10) satisfies Assumption 1, and suppose we have $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathrm{Gap}_{F_{\text{mm-reg}}}(x_0, y_0) \le \epsilon_0$. Algorithm 2 with λ as in (16) returns $(x, y) \in \mathcal{X} \times \mathcal{Y}$ with $\mathrm{Gap}_{F_{\text{mm-reg}}}(x, y) \le \epsilon$ in T iterations, using a total of $O(T)$ gradient calls to each of f, g, h, where*

$$T = O\left(\kappa_{\text{mm}} \log\left(\frac{\kappa_{\text{mm}}\epsilon_0}{\epsilon}\right)\right), \text{ for } \kappa_{\text{mm}} := \sqrt{\frac{L^{\mathsf{x}}}{\mu^{\mathsf{x}}}} + \sqrt{\frac{L^{\mathsf{y}}}{\mu^{\mathsf{y}}}} + \frac{\Lambda^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\Lambda^{\mathsf{xy}}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}} + \frac{\Lambda^{\mathsf{yy}}}{\mu^{\mathsf{y}}}. \tag{17}$$

Proof By Lemma 4, the points $x_\star$ and $y_\star$ are consistent between (10) and (12). The gradient complexity of each iteration follows from observation of Algorithm 2.

Next, by Lemma 9, Algorithm 2 implements Algorithm 1 on the pair (15), (13). By substituting the bounds on λ and m in Lemmas 11 and 5 into Proposition 8 (where we define $\mathcal{Z}_{\text{alg}}$ as in Corollary 10), it is clear that after T iterations (for a sufficiently large constant in the definition of T), we will have $V_{z_T}^r(z_\star)$ is bounded by the quantity in Lemma 13, where we use the initial bound on $V_{z_0}^r(z^\star)$ from Lemma 12. The conclusion follows from setting $(x, y) \leftarrow (z_T^{\mathsf{x}}, z_T^{\mathsf{y}})$. ■

As an immediate corollary, we have the following result on solving (9).

Corollary 15 *Suppose for F_{mm} in (9) solved by $(x_\star, y_\star)$, $(f - \frac{\mu^{\mathsf{x}}}{2}\|\cdot\|^2, g - \frac{\mu^{\mathsf{y}}}{2}\|\cdot\|^2, h)$ satisfies Assumption 1. There is an algorithm taking $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$ satisfying $\mathrm{Gap}_{F_{\text{mm}}}(x_0, y_0) \le \epsilon_0$, which performs T iterations for T in (17), returns $(x, y) \in \mathcal{X} \times \mathcal{Y}$ satisfying $\mathrm{Gap}_{F_{\text{mm}}}(x, y) \le \epsilon$, and uses a total of $O(T)$ gradient calls to each of f, g, h.*

Acknowledgments

We thank anonymous reviewers and Arun Jambulapati for helpful suggestions and pointing out typos. AS was supported in part by a Microsoft Research Faculty Fellowship, NSF CAREER Award CCF-1844855, NSF Grant CCF-1955039, a PayPal research award, and a Sloan Research Fellowship. KT was supported in part by a Google Ph.D. Fellowship, a Simons-Berkeley VMware Research Fellowship, a Microsoft Research Faculty Fellowship, NSF CAREER Award CCF-1844855, NSF Grant CCF-1955039, and a PayPal research award. YJ was supported on a Stanford Graduate Fellowship and the Dantzig-Lieberman Operations Research Fellowship.

References

Jacob D. Abernethy, Kevin A. Lai, Kfir Y. Levy, and Jun-Kun Wang. Faster rates for convex-concave games. In *31st Annual Conference on Computational Learning Theory (COLT)*, pages 1595–1625, 2018.

Naman Agarwal, Sham M. Kakade, Rahul Kidambi, Yin Tat Lee, Praneeth Netrapalli, and Aaron Sidford. Leverage score sampling for faster accelerated regression and ERM. In *Algorithmic Learning Theory, ALT 2020*, pages 22–47, 2020.

Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. *arXiv e-prints*, abs/2102.08352, 2021.

Ahmet Alacaoglu, Olivier Fercoq, and Volkan Cevher. Random extrapolation for primal-dual coordinate descent. In *International conference on machine learning*, pages 191–201. PMLR, 2020.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *J. Mach. Learn. Res.*, 18:221:1–221:51, 2017.

Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Math. Oper. Res.*, 42(2):330–348, 2017.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60(2):223–311, 2018.

Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Variance reduction for matrix games. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 11377–11388, 2019.

Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Coordinate methods for matrix games. In *61st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 283–293, 2020.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

Antonin Chambolle, Matthias J Ehrhardt, Peter Richtárik, and Carola-Bibiane Schonlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 391–401, 2019.

Michael B. Cohen, Aaron Sidford, and Kevin Tian. Relative lipschitzness in extragradient methods and a direct recipe for acceleration. In *12th Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 62:1–62:18, 2021.

Aaron Defazio. A simple practical accelerated method for finite sums. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29 (NeurIPS)*, pages 676–684, 2016.

Aaron Defazio, Francis R. Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems 27 (NeurIPS)*, pages 1646–1654, 2014.

Jelena Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *Conference on Learning Theory*, pages 1428–1451. PMLR, 2020.

Radu-Alexandru Dragomir, Adrien Taylor, Alexandre d'Aspremont, and Jérôme Bolte. Optimal complexity and certification of bregman first-order methods. *arXiv e-prints*, abs/1911.08510, 2019.

Simon S. Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *34th International Conference on Machine Learning (ICML)*, pages 1049–1058, 2017.

Roy Frostig, Rong Ge, Sham M. Kakade, and Aaron Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *32nd International Conference on Machine Learning (ICML)*, pages 2540–2548, 2015.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63 (11):139–144, 2020.

Filip Hanzely, Peter Richtarik, and Lin Xiao. Accelerated bregman proximal gradient methods for relatively smooth convex optimization. *arXiv e-prints*, abs/1808.03045, 2018.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 6936–6946, 2019.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26 (NeurIPS)*, pages 315–323, 2013.

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

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Applications of strong convexity–strong smoothness duality to learning with matrices. *arXiv e-prints*, abs/0910.0610, 2009.

G. M. Korpelevich. An extragradient method for finding saddle points and for other problems. *Ekonomika i Matematicheskie Metody*, 12(4):747–756, 1976.

Dmitry Kovalev, Adil Salim, and Peter Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 18342–18352, 2020.

Dmitry Kovalev, Alexander V. Gasnikov, and Peter Richtárik. Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling. *arXiv e-prints*, abs/2112.15199, 2021.

Guanghui Lan, Zhize Li, and Yi Zhou. A unified variance-reduced accelerated gradient method for convex optimization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 10462–10472, 2019.

Yin Tat Lee and Aaron Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 147–156, 2013.

Hongzhou Lin, Julien Mairal, and Zaïd Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*, pages 3384–3392, 2015.

Tianyi Lin, Chi Jin, and Michael I. Jordan. Near-optimal algorithms for minimax optimization. In *33rd Annual Conference on Computational Learning Theory (COLT)*, pages 2738–2779, 2020.

Haihao Lu, Robert M. Freund, and Yurii E. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.*, 28(1):333–354, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations (ICLR)*, 2018.

Arkadi Nemirovski. Prox-method with rate of convergence $\mathrm{O}(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Yurii Nesterov. A method for solving a convex programming problem with convergence rate $\mathrm{O}(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course, volume I.* 2003.

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Math. Program.*, 109(2-3):319–344, 2007.

Yurii E. Nesterov and Sebastian U. Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM J. Optim.*, 27(1):110–123, 2017.

Balamurugan Palaniappan and Francis R. Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, pages 1408–1416, 2016.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv e-prints*, abs/1908.05659, 2019.

R.T̃yrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970a.

R Tyrrell Rockafellar. Monotone operators associated with saddle-functions and minimax problems. *Nonlinear functional analysis*, 18(part 1):397–407, 1970b.

Mark W. Schmidt, Nicolas Le Roux, and Francis R. Bach. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 162(1-2):83–112, 2017.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *J. Mach. Learn. Res.*, 14(1):567–599, 2013.

Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Math. Program.*, 155(1-2):105–145, 2016.

Jonah Sherman. Area-convexity, l_∞ regularization, and undirected multicommodity flow. In Hamed Hatami, Pierre McKenzie, and Valerie King, editors, *49th Annual ACM Symposium on Theory of Computing (STOC)*, pages 452–460. ACM, 2017.

Chaobing Song, Yong Jiang, and Yi Ma. Variance reduction via accelerated dual averaging for finite-sum optimization. *Advances in Neural Information Processing Systems*, 33:833–844, 2020.

Chaobing Song, Stephen J Wright, and Jelena Diakonikolas. Variance reduction via primal-dual accelerated dual averaging for nonsmooth convex finite-sums. In *International Conference on Machine Learning*, pages 9824–9834. PMLR, 2021.

Fedor Stonyakina, Alexander Tyurin, Alexander Gasnikov, Pavel Dvurechensky, Artem Agafonov, Darina Dvinskikh, Dmitry Pasechnyuk, Sergei Artamonov, and Victorya Piskunova. Inexact relative smoothness and strong convexity for optimization and variational inequalities by inexact model. *arXiv e-prints*, abs/2001.09013, 2020.

Kiran Koshy Thekumparampil, Niao He, and Sewoong Oh. Lifted primal-dual method for bilinearly coupled smooth minimax optimization. In *25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Vladislav Tominin, Yaroslav Tominin, Ekaterina Borodich, Dmitry Kovalev, Alexander Gasnikov, and Pavel Dvurechensky. On accelerated saddle-point problems with composite structure. *arXiv e-prints*, abs/2103.09344v2, 2021.

Jialei Wang and Lin Xiao. Exploiting strong convexity from data with primal-dual first-order algorithms. In *34th International Conference on Machine Learning (ICML)*, pages 3694–3702, 2017.

Jun-Kun Wang and Jacob D. Abernethy. Acceleration through optimistic no-regret dynamics. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 3828–3838, 2018.

Yuanhao Wang and Jian Li. Improved algorithms for convex-concave minimax optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 4800–4810, 2020.

Blake E. Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, pages 3639–3647, 2016.

Junyu Zhang, Minyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the saddle point problems. *arXiv e-prints*, abs/1912.07481, 2019.

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

Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *J. Mach. Learn. Res.*, 18:84:1–84:42, 2017.

Renbo Zhao. Accelerated stochastic algorithms for convex-concave saddle-point problems. *Mathematics of Operations Research*, 47(2):1443–1473, 2022.

Kaiwen Zhou, Qinghua Ding, Fanhua Shang, James Cheng, Danli Li, and Zhi-Quan Luo. Direct acceleration of SAGA using sampled negative momentum. In *22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1602–1610. PMLR, 2019.

Zeyuan Allen Zhu and Elad Hazan. Optimal black-box reductions between optimization objectives. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, pages 1606–1614, 2016.

Zeyuan Allen Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *33rd International Conference on Machine Learning (ICML)*, pages 1110–1119, 2016.

Appendix A. Additional related work

We give a brief discussion of several lines of work which our results build upon, and their connection with the techniques used in this paper, in addition to the works reviewed in Section 1.

Acceleration via primal-dual extragradient methods. Our algorithms are based on *extragradient methods*, a framework originally proposed by Korpelevich (1976) which was later shown to obtain optimal rates for solving Lipschitz variational inequalities in Nemirovski (2004); Nesterov (2007). There have been various implementations of extragradient methods including mirror prox (Nemirovski, 2004) and dual extrapolation (Nesterov, 2007); we focus on adapting the former in this work. Variations of extragradient methods have been studied in the context of primal-dual formulations of smooth convex optimization (Abernethy et al., 2018; Wang and Abernethy, 2018; Cohen et al., 2021), and are known to obtain optimal (accelerated) rates in this setting. In particular, the relative Lipschitzness analysis of acceleration in Cohen et al. (2021) is motivated by developments in the bilinear setting, namely the area convexity framework of Sherman (2017). We build upon these works by using primal-dual formulations to design accelerated algorithms in various settings beyond smooth convex optimization, namely (3), (6), and (8).

Acceleration under relative regularity assumptions. Our analysis builds upon a framework for analyzing extragradient methods known as *relative Lipschitzness*, proposed independently by Stonyakina et al. (2020); Cohen et al. (2021). We demonstrate that this framework (and randomized variants thereof) obtains improved rates for primal-dual formulations beyond those studied in prior works.

Curiously, our applications of the relative Lipschitzness framework reveal that the regularity conditions our algorithms require are weaker than standard assumptions of smoothness in a norm. In particular, several technical requirements of specific components of our algorithms are satisfied by setups with regularity assumptions generalizing and strengthening the *relative smoothness* assumption of Bauschke et al. (2017); Lu et al. (2018). This raises interesting potential implications in terms of the necessary regularity assumptions for non-Euclidean acceleration, because relative smoothness is known to be alone insufficient for obtaining accelerated rates in general (Dragomir et al., 2019). Notably, Hanzely et al. (2018) also developed an acceleration framework under a strengthened relative smoothness assumption, which requires strengthened bounds on divergences between three points. We further elaborate on these points in Appendix D, when deriving relative Lipschitzness bounds through weaker assumptions in Lemma 16. We focus on the Euclidean setup in this paper, but we believe an analogous study of non-Euclidean setups is interesting and merits future exploration.

Appendix B. Preliminaries

We provide detailed preliminaries, introducing notations and definitions used throughout the paper.

General notation. We use $\widetilde{O}$ to hide logarithmic factors in problem regularity parameters, initial radius bounds, and target accuracies when clear from context. We denote $[n] := \{i \in \mathbb{N} \mid i \le n\}$. Throughout the paper, $\mathcal{X}$ (and $\mathcal{Y}$, when relevant) represent Euclidean spaces, and $\|\cdot\|$ will mean the Euclidean norm in appropriate dimension when applied to a vector. For a variable on a product space, e.g. $z \in \mathcal{X} \times \mathcal{Y}$, we refer to its blocks as $(z^{\mathsf{x}}, z^{\mathsf{y}})$ when clear from context. For a bilinear operator $\mathbf{A} : \mathcal{X} \to \mathcal{Y}^*$, $\|\cdot\|$ will mean the (Euclidean) operator norm, i.e.

$$\|\mathbf{A}\| := \sup_{\|x\|=1} \|\mathbf{A}x\| = \sup_{\|x\|=1} \sup_{\|y\|=1} y^\top \mathbf{A}x.$$

Complexity model. Throughout the paper, we evaluate the complexity of methods by their gradient oracle complexity, and do not discuss the cost of vector operations (which typically are subsumed by the cost of the oracle). In Section 2, the gradient oracle returns ∇f, ∇g, or ∇h at any point; in Appendix E (respectively, Appendix F), the oracle returns ∇f_i at a point for some $i \in [n]$ (respectively, ∇f_i, ∇g_i, or ∇h_i at a point for some $i \in [n]$).

Divergences. The Bregman divergence induced by differentiable, convex r is $V_x^r(x') := r(x') - r(x) - \langle \nabla r(x), x' - x \rangle$, for any $x, x' \in \mathcal{X}$. For all x, V_x^r is nonnegative and convex. Whenever we use no superscript r, we assume $r = \frac{1}{2}\left\|\cdot\right\|^2$ so that $V_x(x') = \frac{1}{2}\left\|x - x'\right\|^2$. Bregman divergences satisfy the equality

$$\langle \nabla r(w) - \nabla r(z), w - u \rangle = V_z^r(w) + V_w^r(u) - V_z^r(u). \tag{18}$$

We define the proximal operation in r by

$$\mathrm{Prox}_x^r(\Phi) := \mathrm{argmin}_{x' \in \mathcal{X}} \left\{ \left\langle \Phi, x' \right\rangle + V_x^r(x') \right\}.$$

Functions and operators. We say $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is convex-concave if its restrictions $h(\cdot, y)$ and $h(x, \cdot)$ are respectively convex and concave, for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The duality gap of a pair (x, y) is $\mathrm{Gap}_h(x, y) := \max_{y' \in \mathcal{Y}} h(x, y') - \min_{x' \in \mathcal{X}} h(x', y)$; a saddle point is a pair $(x_\star, y_\star) \in \mathcal{X} \times \mathcal{Y}$ with zero duality gap.

We call operator $\Phi : \mathcal{Z} \to \mathcal{Z}^*$ monotone if $\langle \Phi(z) - \Phi(z'), z - z' \rangle \geq 0$ for all $z, z' \in \mathcal{Z}$. We say $z_\star$ solves the variational inequality (VI) in Φ if $\langle \Phi(z_\star), z_\star - z \rangle \leq 0$ for all $z \in \mathcal{Z}$. We equip differentiable convex-concave h with the "gradient operator" $\Phi(x, y) := (\nabla_x h(x, y), -\nabla_y h(x, y))$. The gradient of convex f and the gradient operator of convex-concave h are both monotone. Their VIs are respectively solved by any minimizers of f and saddle points of h.

Regularity. We say function $f : \mathcal{X} \to \mathbb{R}$ is L-smooth if $\left\|\nabla f(x) - \nabla f(x')\right\| \leq L \left\|x - x'\right\|$ for all $x, x' \in \mathcal{X}$; if f is twice-differentiable, this is equivalent to $(x' - x)^\top \nabla^2 f(x)(x' - x) \leq L \left\|x' - x\right\|^2$ for all $x, x' \in \mathcal{X}$. We say differentiable function $f : \mathcal{X} \to \mathbb{R}$ is μ-strongly convex if $V_x^f(x') \geq \frac{\mu}{2}\left\|x - x'\right\|^2$ for all $x, x' \in \mathcal{X}$; if f is twice-differentiable, this is equivalent to $(x' - x)^\top \nabla^2 f(x)(x' - x) \geq \mu \left\|x' - x\right\|^2$ for all $x, x' \in \mathcal{X}$. Finally, we say operator $\Phi : \mathcal{Z} \to \mathcal{Z}^*$ is m-strongly monotone with respect to convex $r : \mathcal{Z} \to \mathbb{R}$ if for all $z, z' \in \mathcal{Z}$,

$$\left\langle \Phi(z) - \Phi(z'), z - z' \right\rangle \geq m \left\langle \nabla r(z) - \nabla r(z'), z - z' \right\rangle = m \left(V_z^r(z') + V_{z'}^r(z) \right).$$

Convex conjugates. The (Fenchel dual) convex conjugate of a convex $f : \mathcal{X} \to \mathbb{R}$ is denoted

$$f^*(x^*) := \max_{x \in \mathcal{X}} \langle x, x^* \rangle - f(x).$$

We allow f^* to take the value ∞. We recall the following facts about convex conjugates.

Fact 1 *Let $f : \mathcal{X} \to \mathbb{R}$ be differentiable.*

(1) For all $x \in \mathcal{X}$, $\nabla f(x) \in \mathrm{argmax}_{x^* \in \mathcal{X}^*} \langle x^*, x \rangle - f^*(x^*)$.

(2) $(f^*)^* = f$.

(3) If f^* is differentiable, for all $x \in \mathcal{X}$, $\nabla f^*(\nabla f(x)) = x$.

(4) If f is L-smooth, then for all $x, x' \in \mathcal{X}$,

$$f(x') - f(x) - \left\langle \nabla f(x), x' - x \right\rangle \geq \frac{1}{2L} \left\| \nabla f(x') - \nabla f(x) \right\|^2 .$$

If f is μ-strongly convex, f^* is $\frac{1}{\mu}$-smooth.

Proof The first three items all follow from Chapter 11 of Rockafellar (1970a). The first part of the fourth item is shown in Appendix A of Cohen et al. (2021), and the second part is shown in Kakade et al. (2009). ■

For a function $f : \mathcal{X} \to \mathbb{R}$, we define the set $\mathcal{X}_f^* \subset \mathcal{X}^*$ to be the set of points realizable as a gradient, namely $\mathcal{X}_f^* := \{\nabla f(x) \mid x \in \mathcal{X}\}$. This will be come relevant in applications of Item 4 in Fact 1 throughout the paper, when ∇f is not surjective onto $\mathcal{X}^*$.

Appendix C. Helper facts

Here for completeness we state two helper facts that we use throughout the analysis. The first gives a few properties on monotone operators. We first recall by definition, an operator $\Phi : \mathcal{Z} \to \mathcal{Z}^*$ is monotone if

$$\left\langle \Phi(z) - \Phi(z'), z - z' \right\rangle \geq 0, \quad \text{for all } z, z' \in \mathcal{Z}.$$

An operator Φ is m-strongly monotone with respect to convex $r : \mathcal{Z} \to \mathbb{R}$ if for all $z, z' \in \mathcal{Z}$,

$$\left\langle \Phi(z) - \Phi(z'), z - z' \right\rangle \geq m \left\langle \nabla r(z) - \nabla r(z'), z - z' \right\rangle, \quad \text{for all } z, z' \in \mathcal{Z}.$$

We state the following standard facts about monotone operators and their specialization to convex-concave functions, and include references or proofs for completeness.

Fact 2 *The following facts about monotone operators hold true:*

(1) Given a convex function $f(x) : \mathcal{X} \to \mathbb{R}$, its induced operator $\Phi = \nabla f : \mathcal{X} \to \mathcal{X}^*$ is monotone.

(2) Given a convex-concave function $h(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, its induced operator $\Phi(x, y) = (\nabla_x h(x, y), -\nabla_y h(x, y)) : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}^* \times \mathcal{Y}^*$ is monotone.

(3) Given a convex function f, its induced operator $\Phi = \nabla f$ is 1-strongly monotone with respect to itself.

(4) Monotonicity is preserved under addition: For any $m, m' \geq 0$, if Φ is m-strongly monotone and Ψ is m'-strongly monotone with respect to convex r, then $\Phi + \Psi$ is $(m + m')$-strongly monotone with respect to r.

Proof The first two items are basic fact of convexity and minimax optimization (Rockafellar, 1970b). For the third item, we note that for any $x, x' \in \mathcal{X}$

$$\left\langle \Phi(x) - \Phi(x'), x - x' \right\rangle = \left\langle \nabla f(x) - \nabla f(x'), x - x' \right\rangle,$$

which satisfies 1-strong monotonicity with respect to f by definition.

For the fourth item, we note that for any $m, m' \geq 0$ and assumed Φ, Ψ,

$$\begin{aligned} \left\langle \Phi(z) - \Phi(z'), z - z' \right\rangle &\geq m \left\langle \nabla r(z) - \nabla r(z'), z - z' \right\rangle, \\ \left\langle \Psi(z) - \Psi(z'), z - z' \right\rangle &\geq m' \left\langle \nabla r(z) - \nabla r(z'), z - z' \right\rangle, \\ \implies \left\langle \Phi(z) + \Psi(z) - \left(\Phi(z') + \Psi(z')\right), z - z' \right\rangle &\geq (m + m') \left\langle \nabla r(z) - \nabla r(z'), z - z' \right\rangle. \end{aligned}$$

■

These facts about monotone operators find usage in proving (relative) strong monotonicity of our operators; see Lemma 5, 18 and 32.

The second fact bounds the smoothness of best-response function of some given convex-concave function $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. We refer readers to Fact 1 of Wang and Li (2020) for a complete proof.

Fact 3 (Fact 1, Wang and Li (2020)) *Suppose h satisfies the blockwise-smoothness properties: for all $u, v \in \mathcal{X} \times \mathcal{Y}$,*

$$\begin{aligned} \left\| \nabla_x h(u) - \nabla_x h(v) \right\| &\leq \Lambda^{\mathsf{xx}} \left\| u^{\mathsf{x}} - v^{\mathsf{x}} \right\| + \Lambda^{\mathsf{xy}} \left\| u^{\mathsf{y}} - v^{\mathsf{y}} \right\|, \\ \left\| \nabla_y h(u) - \nabla_y h(v) \right\| &\leq \Lambda^{\mathsf{xy}} \left\| u^{\mathsf{x}} - v^{\mathsf{x}} \right\| + \Lambda^{\mathsf{yy}} \left\| u^{\mathsf{y}} - v^{\mathsf{y}} \right\|, \end{aligned} \tag{19}$$

and suppose h is μ^{x}-strongly convex in x and μ^{y}-strongly concave in y. The best response function $h^{\mathsf{y}}(x) := \max_{y \in \mathcal{Y}} h(x, y)$ is μ^{x}-strongly convex and $\left(\Lambda^{\mathsf{xx}} + \frac{(\Lambda^{\mathsf{xy}})^2}{\mu^{\mathsf{y}}}\right)$-smooth, and $h^{\mathsf{x}}(y) := \min_{x \in \mathcal{Y}} h(x, y)$ is μ^{y}-strongly concave and $\left(\Lambda^{\mathsf{yy}} + \frac{(\Lambda^{\mathsf{xy}})^2}{\mu^{\mathsf{x}}}\right)$-smooth.

We use this fact when converting radius bounds to duality gap bounds in Lemma 12 and 13.

Appendix D. Proofs for Section 2

D.1. Proofs for Section 2.1

Lemma 4 *Let $z_\star$ be the saddle point to (12). Then, $(z_\star^{\mathsf{x}}, z_\star^{\mathsf{y}})$ is a saddle point to (10).*

Proof By performing the maximization over z^{f^*} and minimization over z^{g^*}, we see that the problem of computing a saddle point to the objective in (12) is equivalent to

$$\begin{aligned} \min_{z^{\mathsf{x}} \in \mathcal{X}} \max_{z^{\mathsf{y}} \in \mathcal{Y}} & \frac{\mu^{\mathsf{x}}}{2} \left\| z^{\mathsf{x}} \right\|^2 - \frac{\mu^{\mathsf{y}}}{2} \left\| z^{\mathsf{y}} \right\|^2 + h(z^{\mathsf{x}}, z^{\mathsf{y}}) \\ & + \left(\max_{z^{\mathsf{f}^*} \in \mathcal{X}^*} \left\langle z^{\mathsf{f}^*}, z^{\mathsf{x}} \right\rangle - f^*(z^{\mathsf{f}^*}) \right) - \left(\max_{z^{\mathsf{g}^*} \in \mathcal{Y}^*} \left\langle z^{\mathsf{g}^*}, z^{\mathsf{y}} \right\rangle - g^*(z^{\mathsf{g}^*}) \right). \end{aligned}$$

By Item 2 in Fact 1, this is the same as (10). ■

Lemma 5 (Strong monotonicity) *The operator Φ (as defined in (15)) is 1-strongly monotone with respect to the function $r : \mathcal{Z} \to \mathbb{R}$ as in (13).*

Proof Consider the decomposition of $\Phi = \Phi^r + \Phi^{\mathsf{bilin}} + \Phi^h$ defined in (14) and (15). By definition and Items (1) to (3) from Fact 2, we know the operators Φ^h and Φ^{bilin} are monotone, and $\Phi^r = \nabla r$ is 1-strongly monotone with respect to r. Combining the three operators and using additivity of monotonicity in Item (4) of Fact 2 yields the claim. ■

D.2. Proofs for Section 2.2

Lemma 9 *Algorithm 2 implements Algorithm 1 with $m = 1$ on (Φ, r) defined in (15), (13).*

Proof Let $\{z_t, z_{t+1/2}\}_{0 \le t \le T}$ be the iterates of Algorithm 1. We will inductively show that Algorithm 2 preserves the invariants

$$z_t = \left(z_t^{\mathsf{x}}, z_t^{\mathsf{y}}, \nabla f\left(z_t^{\mathsf{f}}\right), \nabla g\left(z_t^{\mathsf{g}}\right)\right),\ z_{t+1/2} = \left(z_{t+1/2}^{\mathsf{x}}, z_{t+1/2}^{\mathsf{y}}, \nabla f\left(z_{t+1/2}^{\mathsf{f}}\right), \nabla g\left(z_{t+1/2}^{\mathsf{g}}\right)\right),$$

for the iterates of Algorithm 2. Once we prove this claim, it is clear from inspection that Algorithm 2 implements Algorithm 1, upon recalling the definitions (15), (13).

The base case of our induction follows from our initialization so that $(\nabla f(z_0^{\mathsf{f}}), \nabla g(z_0^{\mathsf{g}})) \leftarrow (\nabla f(x_0), \nabla f(y_0))$. Next, suppose for some $0 \le t < T$, we have $z_t^{\mathsf{f}^*} = \nabla f(z_t^{\mathsf{f}})$ and $z_t^{\mathsf{g}^*} = \nabla g(z_t^{\mathsf{g}})$. By the updates in Algorithm 1,

$$\begin{aligned} z_{t+1/2}^{\mathsf{f}^*} \leftarrow \mathrm{argmin}_{z^{\mathsf{f}^*} \in \mathcal{X}^*} &\left\{\frac{1}{\lambda}\left\langle \nabla f^*(z_t^{\mathsf{f}^*}) - z_t^{\mathsf{x}}, z^{\mathsf{f}^*}\right\rangle + V^{f^*}_{z_t^{\mathsf{f}^*}}(z^{\mathsf{f}^*})\right\} \\ = \mathrm{argmin}_{z^{\mathsf{f}^*} \in \mathcal{X}^*} &\left\{\frac{1}{\lambda}\left\langle z_t^{\mathsf{f}} - z_t^{\mathsf{x}}, z^{\mathsf{f}^*}\right\rangle - \left\langle z_t^{\mathsf{f}}, z^{\mathsf{f}^*}\right\rangle + f^*(z^{\mathsf{f}^*})\right\} \\ = \mathrm{argmax}_{z^{\mathsf{f}^*} \in \mathcal{X}^*} &\left\{\left\langle \left(1 - \frac{1}{\lambda}\right) z_t^{\mathsf{f}} + \frac{1}{\lambda} z_t^{\mathsf{x}}, z^{\mathsf{f}^*}\right\rangle - f^*(z^{\mathsf{f}^*})\right\} = \nabla f\left(\left(1 - \frac{1}{\lambda}\right) z_t^{\mathsf{f}} + \frac{1}{\lambda} z_t^{\mathsf{x}}\right). \end{aligned}$$

The second line used our inductive hypothesis and Item 3 in Fact 1, and the last used Item 1 in Fact 1. Hence, the update to $z_{t+1/2}^{\mathsf{f}}$ in Algorithm 2 preserves our invariant; a symmetric argument yields $z_{t+1/2}^{\mathsf{g}^*} = \nabla g(z_{t+1/2}^{\mathsf{g}})$ where $z_{t+1/2}^{\mathsf{g}} := (1 - \frac{1}{\lambda}) z_t^{\mathsf{g}} + \frac{1}{\lambda} z_t^{\mathsf{y}}$.

Similarly, we show we may preserve this invariant for z_{t+1}:

$$\begin{aligned} z_{t+1}^{\mathsf{f}^*} \leftarrow \mathrm{argmin}_{z^{\mathsf{f}^*} \in \mathcal{X}^*} &\left\{\frac{1}{\lambda}\left\langle z_{t+1/2}^{\mathsf{f}} - z_{t+1/2}^{\mathsf{x}}, z^{\mathsf{f}^*}\right\rangle - \frac{1}{\lambda}\left\langle z_{t+1/2}^{\mathsf{f}}, z^{\mathsf{f}^*}\right\rangle - \left\langle z_t^{\mathsf{f}}, z^{\mathsf{f}^*}\right\rangle + \left(1 + \frac{1}{\lambda}\right) f^*(z^{\mathsf{f}^*})\right\} \\ = \mathrm{argmax}_{a \in \mathcal{X}^*} &\left\{\left\langle z_t^{\mathsf{f}} + \frac{1}{\lambda} z_{t+1/2}^{\mathsf{x}}, z^{\mathsf{f}^*}\right\rangle - \left(1 + \frac{1}{\lambda}\right) f^*(z^{\mathsf{f}^*})\right\} = \nabla f\left(\frac{\lambda}{1+\lambda} z_t^{\mathsf{f}} + \frac{1}{1+\lambda} z_{t+1/2}^{\mathsf{x}}\right). \end{aligned}$$

Hence, we may set $z_{t+1}^{\mathsf{f}} := \frac{\lambda}{1+\lambda} z_t^{\mathsf{f}} + \frac{1}{1+\lambda} z_{t+1/2}^{\mathsf{x}}$ and similarly, $z_{t+1}^{\mathsf{g}} := \frac{\lambda}{1+\lambda} z_t^{\mathsf{g}} + \frac{\lambda}{1+\lambda} z_{t+1/2}^{\mathsf{y}}$. ■

D.3. Proofs for Section 2.3

We build up to our relative Lipschitzness bound by first giving the following consequences of Assumption 1.

Lemma 16 (Minimax smoothness implications) *Let convex $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$, and convex-concave $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ satisfy Assumption 1. Then, the following hold.*

(1) $|\langle \nabla f(v) - \nabla f(w), x - y\rangle| \le \alpha L^{\mathsf{x}} V_v^f(w) + \alpha^{-1} V_x(y)$ *for all $v, w, x, y \in \mathcal{X}$ and $\alpha > 0$.*

(2) $|\langle \nabla g(v) - \nabla g(w), x - y\rangle| \le \alpha L^{\mathsf{y}} V_v^g(w) + \alpha^{-1} V_x(y)$ *for all $v, w, x, y \in \mathcal{Y}$ and $\alpha > 0$.*

(3) Φ^h is 1-relatively Lipschitz with respect to $r_\alpha^h : \mathcal{Z} \to \mathbb{R}$ defined for all $z \in \mathcal{Z}$ and $\alpha > 0$ by $r_\alpha^h(z) := \frac{1}{2}(\Lambda^{\mathsf{xx}} + \alpha \Lambda^{\mathsf{xy}}) \|z^{\mathsf{x}}\|^2 + \frac{1}{2}\left(\Lambda^{\mathsf{yy}} + \alpha^{-1}\Lambda^{\mathsf{xy}}\right) \|z^{\mathsf{y}}\|^2$.

Proof We will prove Items 1 and 3, as Item 2 follows symmetrically to Item 1.

Proof of Item (1). We compute:

$$\begin{aligned}|\langle \nabla f(v) - \nabla f(w), x - y\rangle| &\leq \|\nabla f(v) - \nabla f(w)\| \, \|x - y\| \\ &\leq \frac{\alpha}{2} \|\nabla f(v) - \nabla f(w)\|^2 + \frac{1}{2\alpha} \|x - y\|^2 \\ &\leq \alpha L^{\mathsf{x}} V^{f^*}_{\nabla f(w)}(\nabla f(v)) + \alpha^{-1} V_x(y) = \alpha L^{\mathsf{x}} V^f_v(w) + \alpha^{-1} V_x(y).\end{aligned}$$

The first inequality was Cauchy-Schwarz, the second was Young's inequality, and the third used Items 3 and 4 in Fact 1. The last equality follows from Fact 1.

Proof of Item (3). Let $w, v, z \in \mathcal{Z}$ be arbitrary. We have,

$$\begin{aligned}&\left\langle \Phi^h(w) - \Phi^h(z), w - v \right\rangle \\ &\quad = \langle \nabla_x h(w^{\mathsf{x}}, w^{\mathsf{y}}) - \nabla_x h(z^{\mathsf{x}}, z^{\mathsf{y}}), w^{\mathsf{x}} - v^{\mathsf{x}}\rangle - \langle \nabla_y h(w^{\mathsf{x}}, w^{\mathsf{y}}) - \nabla_y h(z^{\mathsf{x}}, z^{\mathsf{y}}), w^{\mathsf{y}} - v^{\mathsf{y}}\rangle .\end{aligned}$$

Applying Cauchy-Schwarz, Young's inequality, and Assumption 1 yields

$$\begin{aligned}\langle \nabla_x h(w^{\mathsf{x}}, w^{\mathsf{y}}) - \nabla_x h(z^{\mathsf{x}}, z^{\mathsf{y}}), w^{\mathsf{x}} - v^{\mathsf{x}}\rangle &\leq \|\nabla_x h(w^{\mathsf{x}}, w^{\mathsf{y}}) - \nabla_x h(z^{\mathsf{x}}, z^{\mathsf{y}})\| \, \|w^{\mathsf{x}} - v^{\mathsf{x}}\| \\ &\leq \left(\Lambda^{\mathsf{xx}} \|w^{\mathsf{x}} - z^{\mathsf{x}}\| + \Lambda^{\mathsf{xy}} \|w^{\mathsf{y}} - z^{\mathsf{y}}\|\right) \|w^{\mathsf{x}} - v^{\mathsf{x}}\| \\ &\leq \frac{\Lambda^{\mathsf{xx}}}{2} \|w^{\mathsf{x}} - z^{\mathsf{x}}\|^2 + \frac{\Lambda^{\mathsf{xx}}}{2} \|w^{\mathsf{x}} - v^{\mathsf{x}}\|^2 + \Lambda^{\mathsf{xy}} \|w^{\mathsf{y}} - z^{\mathsf{y}}\| \, \|w^{\mathsf{x}} - v^{\mathsf{x}}\| .\end{aligned}$$

Symmetrically,

$$\begin{aligned}&\langle \nabla_y h(w^{\mathsf{x}}, w^{\mathsf{y}}) - \nabla_y h(z^{\mathsf{x}}, z^{\mathsf{y}}), w^{\mathsf{y}} - v^{\mathsf{y}}\rangle \\ &\quad \leq \frac{\Lambda^{\mathsf{yy}}}{2} \|w^{\mathsf{y}} - z^{\mathsf{y}}\|^2 + \frac{\Lambda^{\mathsf{yy}}}{2} \|w^{\mathsf{y}} - v^{\mathsf{y}}\|^2 + \Lambda^{\mathsf{xy}} \|w^{\mathsf{x}} - z^{\mathsf{x}}\| \, \|w^{\mathsf{y}} - v^{\mathsf{y}}\| .\end{aligned}$$

Applying Young's inequality again yields

$$\begin{aligned}\Lambda^{\mathsf{xy}} \|w^{\mathsf{y}} - z^{\mathsf{y}}\| \, \|w^{\mathsf{x}} - v^{\mathsf{x}}\| &\leq \frac{\alpha\Lambda^{\mathsf{xy}}}{2} \|w^{\mathsf{x}} - v^{\mathsf{x}}\|^2 + \frac{\Lambda^{\mathsf{xy}}}{2\alpha} \|w^{\mathsf{y}} - z^{\mathsf{y}}\|^2 , \\ \text{and } \Lambda^{\mathsf{xy}} \|w^{\mathsf{x}} - z^{\mathsf{x}}\| \, \|w^{\mathsf{y}} - v^{\mathsf{y}}\| &\leq \frac{\alpha\Lambda^{\mathsf{xy}}}{2} \|w^{\mathsf{x}} - z^{\mathsf{x}}\|^2 + \frac{\Lambda^{\mathsf{xy}}}{2\alpha} \|w^{\mathsf{y}} - v^{\mathsf{y}}\|^2 .\end{aligned}$$

Combining these inequalities yields the desired bound of

$$\begin{aligned}&\left\langle \Phi^h(w) - \Phi^h(z), w - v \right\rangle \\ &\quad \leq \left(\Lambda^{\mathsf{xx}} + \alpha\Lambda^{\mathsf{xy}}\right)\left(V_{z^{\mathsf{x}}}(w^{\mathsf{x}}) + V_{w^{\mathsf{x}}}(v^{\mathsf{x}})\right) + \left(\Lambda^{\mathsf{yy}} + \alpha\Lambda^{\mathsf{xy}}\right)\left(V_{z^{\mathsf{y}}}(w^{\mathsf{y}}) + V_{w^{\mathsf{y}}}(v^{\mathsf{y}})\right) \\ &\quad = V^{r^h_\alpha}_z(w) + V^{r^h_\alpha}_w(v).\end{aligned}$$

■

Leveraging Lemma 16 and Lemma 7, we prove relative Lipschitzness of Φ with respect to r in Lemma 11. Interestingly, the implications in Lemma 16 are sufficient for this proof, and this serves as a (potentially) weaker replacement for Assumption 1 in yielding a convergence rate for our method.

This is particularly interesting when the condition in Item (1) is replaced with a non-Euclidean divergence, namely $|\langle \nabla f(v) - \nabla f(w), x - y\rangle| \le \alpha L^{\mathsf{x}} V_v^f(w) + \alpha^{-1} V_x^{\omega}(y)$ for some convex $\omega : \mathcal{X} \to \mathbb{R}$. Setting, setting $v = y, w = x, \alpha = \frac{1}{L^{\mathsf{x}}}$ in this condition yields $V_x^f(y) \le L V_x^{\omega}(y)$. Hence, this extension to Item (1) generalizes *relative smoothness* between f and ω, a condition introduced by Bauschke et al. (2017); Lu et al. (2018). It has been previously observed (Hanzely et al., 2018; Dragomir et al., 2019) that relative smoothness alone does not suffice for accelerated rates. Item (1) provides a new strengthening of relative smoothness which, as shown by its (implicit) use in Cohen et al. (2021), suffices for acceleration. We believe a more thorough investigation comparing these conditions is an interesting avenue for future work.

Lemma 11 (Relative Lipschitzness) *Define $\Phi : \mathcal{Z} \to \mathcal{Z}^*$ as in (15), and define $r : \mathcal{Z} \to \mathbb{R}$ as in (13). Then Φ is λ-relatively Lipschitz with respect to r over $\mathcal{Z}_{\mathsf{alg}}$ defined in Corollary 10 for*

$$\lambda = 1 + \sqrt{\frac{L^{\mathsf{x}}}{\mu^{\mathsf{x}}}} + \sqrt{\frac{L^{\mathsf{y}}}{\mu^{\mathsf{y}}}} + \frac{\Lambda^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\Lambda^{\mathsf{xy}}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}} + \frac{\Lambda^{\mathsf{yy}}}{\mu^{\mathsf{y}}}. \tag{16}$$

Proof Let $w, v, z \in \mathcal{Z}_{\mathsf{alg}}$. We wish to show (cf. Definition 6)

$$\langle \Phi(w) - \Phi(z), w - v\rangle \le \lambda \left(V_z^r(w) + V_w^r(v)\right).$$

Since $\Phi = \Phi^r + \Phi^{\mathsf{bilin}} + \Phi^h$ (cf. (15)), we bound the contribution of each term individually. The conclusion follows from combining (20), (21), and (22).

Bound on Φ^r: By applying Lemma 7 to r,

$$\langle \Phi^r(w) - \Phi^r(z), w - v\rangle = \langle \nabla r(w) - \nabla r(z), w - v\rangle \le V_z^r(w) + V_w^r(v). \tag{20}$$

Bound on Φ^{bilin}: For all $a \in \mathcal{Z}_{\mathsf{alg}}$, we may write for some $a^{\mathsf{f}} \in \mathcal{X}$ and $a^{\mathsf{g}} \in \mathcal{Y}$,

$$\begin{aligned} \Phi^{\mathsf{bilin}}(a) &= (a^{\mathsf{f}^*}, a^{\mathsf{g}^*}, -a^{\mathsf{x}}, -a^{\mathsf{y}}) = (\nabla f(a^{\mathsf{f}}), \nabla g(a^{\mathsf{g}}), -a^{\mathsf{x}}, -a^{\mathsf{y}}) \\ \text{and } a &= (a^{\mathsf{x}}, a^{\mathsf{y}}, a^{\mathsf{f}^*}, a^{\mathsf{g}^*}) = (a^{\mathsf{x}}, a^{\mathsf{y}}, \nabla f(a^{\mathsf{f}}), \nabla g(a^{\mathsf{g}})). \end{aligned}$$

Consequently,

$$\begin{aligned} \left\langle \Phi^{\mathsf{bilin}}(w) - \Phi^{\mathsf{bilin}}(z), w - v\right\rangle &= \left\langle \nabla f(w^{\mathsf{f}}) - \nabla f(z^{\mathsf{f}}), w^{\mathsf{x}} - v^{\mathsf{x}}\right\rangle + \left\langle \nabla g(w^{\mathsf{f}}) - \nabla g(z^{\mathsf{f}}), w^{\mathsf{y}} - v^{\mathsf{y}}\right\rangle \\ &\quad - \left\langle w^{\mathsf{x}} - z^{\mathsf{x}}, \nabla f(w^{\mathsf{f}}) - \nabla f(v^{\mathsf{f}})\right\rangle - \langle w^{\mathsf{y}} - z^{\mathsf{y}}, \nabla g(w^{\mathsf{g}}) - \nabla g(v^{\mathsf{g}})\rangle. \end{aligned}$$

Applying Lemma 16 (Item (1) and Item (2)) to each term, with $\alpha = (\mu^{\mathsf{x}} L^{\mathsf{x}})^{-\frac{1}{2}}$ for terms involving f and $\alpha = (\mu^{\mathsf{y}} L^{\mathsf{y}})^{-\frac{1}{2}}$ for terms involving g yields

$$\begin{aligned} \left\langle \Phi^{\mathsf{bilin}}(w) - \Phi^{\mathsf{bilin}}(z), w - v\right\rangle &\le \sqrt{\frac{L^{\mathsf{x}}}{\mu^{\mathsf{x}}}} \left(V_{w^{\mathsf{f}}}^f(z^{\mathsf{f}}) + V_{v^{\mathsf{f}}}^f(w^{\mathsf{f}})\right) + \sqrt{\frac{L^{\mathsf{x}}}{\mu^{\mathsf{x}}}} \left(\mu^{\mathsf{x}} V_{w^{\mathsf{x}}}(v^{\mathsf{x}}) + \mu^{\mathsf{x}} V_{z^{\mathsf{x}}}(w^{\mathsf{x}})\right) \\ &\quad + \sqrt{\frac{L^{\mathsf{y}}}{\mu^{\mathsf{y}}}} \left(V_{w^{\mathsf{g}}}^g(z^{\mathsf{g}}) + V_{v^{\mathsf{g}}}^g(w^{\mathsf{g}})\right) + \sqrt{\frac{L^{\mathsf{y}}}{\mu^{\mathsf{y}}}} \left(\mu^{\mathsf{y}} V_{w^{\mathsf{y}}}(v^{\mathsf{y}}) + \mu^{\mathsf{y}} V_{z^{\mathsf{y}}}(w^{\mathsf{y}})\right). \end{aligned}$$

Applying Item 3 in Fact 1 and recalling the definition of r (13) yields

$$\left\langle \Phi^{\mathsf{bilin}}(w) - \Phi^{\mathsf{bilin}}(z), w - v\right\rangle \le \left(\sqrt{\frac{L^{\mathsf{x}}}{\mu^{\mathsf{x}}}} + \sqrt{\frac{L^{\mathsf{y}}}{\mu^{\mathsf{y}}}}\right) \left(V_z^r(w) + V_w^r(v)\right). \tag{21}$$

Bound on Φ^h: Applying Lemma 16 (Item (3) with $\alpha = \sqrt{\mu^{\mathsf{x}}/\mu^{\mathsf{y}}}$), we have that Φ^h is 1-relatively Lipschitz with respect to $r^h_\alpha : \mathcal{Z} \to \mathbb{R}$ defined for all $z \in \mathcal{X}$ and $\alpha > 0$ by

$$\begin{aligned} r^h_\alpha(z) &:= \frac{1}{2}\left(\Lambda^{\mathsf{xx}} + \alpha\Lambda^{\mathsf{xy}}\right)\left\|z^{\mathsf{x}}\right\|^2 + \frac{1}{2}\left(\Lambda^{\mathsf{yy}} + \alpha^{-1}\Lambda^{\mathsf{xy}}\right)\left\|z^{\mathsf{y}}\right\|^2 \\ &= \left(\frac{\Lambda^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\Lambda^{\mathsf{xy}}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}}\right) \cdot \frac{\mu^{\mathsf{x}}}{2}\left\|z^{\mathsf{x}}\right\|^2 + \left(\frac{\Lambda^{\mathsf{yy}}}{\mu^{\mathsf{y}}} + \frac{\Lambda^{\mathsf{xy}}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}}\right) \cdot \frac{\mu^{\mathsf{y}}}{2}\left\|z^{\mathsf{y}}\right\|^2. \end{aligned}$$

Leveraging the nonnegativity of Bregman divergences, we conclude

$$\begin{aligned} \left\langle \Phi^h(w) - \Phi^h(z), w - v \right\rangle &\le V_z^{r^h_\alpha}(w) + V_w^{r^h_\alpha}(v) \\ &\le \left(\frac{\Lambda^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\Lambda^{\mathsf{xy}}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}} + \frac{\Lambda^{\mathsf{yy}}}{\mu^{\mathsf{y}}}\right)\left(V_z^r(w) + V_w^r(v)\right). \end{aligned} \tag{22}$$

■

Lemma 12 *Let $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$, and define $z_0 := (x_0, y_0, \nabla f(x_0), \nabla g(y_0))$. Suppose we have* $\mathrm{Gap}_{F_{\text{mm-reg}}}(x_0, y_0) \le \epsilon_0$. *Then, letting $z_\star$ be the solution to (12),*

$$V_{z_0}^r(z_\star) \le \left(1 + \frac{L^{\mathsf{x}}}{\mu_x} + \frac{L^{\mathsf{y}}}{\mu^{\mathsf{y}}}\right)\epsilon_0.$$

Proof By the characterization in Lemma 4, we have by Item 1 in Fact 1:

$$z_\star = (x_\star, y_\star, \nabla f(x_\star), \nabla g(y_\star)).$$

Hence, we bound

$$\begin{aligned} V_{z_0}^r(z_\star) &= \mu^{\mathsf{x}} V_{x_0}(x_\star) + V_{x_\star}^f(x_0) + \mu^{\mathsf{y}} V_{y_0}(y_\star) + V_{y_\star}^g(y_0) \\ &\le \mu^{\mathsf{x}} V_{x_0}(x_\star) + \frac{L^{\mathsf{x}}}{2}\left\|x_0 - x_\star\right\|_{\mathcal{X}}^2 + \mu^{\mathsf{y}} V_{y_0}(y_\star) + \frac{L^{\mathsf{y}}}{2}\left\|y_0 - y_\star\right\|_{\mathcal{Y}}^2 \\ &= \left(\frac{L^{\mathsf{x}}}{\mu^{\mathsf{x}}} + 1\right)\mu^{\mathsf{x}} V_{x_0}(x_\star) + \left(\frac{L^{\mathsf{y}}}{\mu^{\mathsf{y}}} + 1\right)\mu^{\mathsf{y}} V_{y_0}(y_\star) \\ &\le \left(\frac{L^{\mathsf{x}}}{\mu^{\mathsf{x}}} + \frac{L^{\mathsf{y}}}{\mu^{\mathsf{y}}} + 1\right)\epsilon_0. \end{aligned}$$

The first line used Item 3 in Fact 1, and the second used smoothness of f and g (Assumption 1). To obtain the last line, define the functions

$$F^{\mathsf{x}}_{\text{mm-reg}}(x) := \max_{y \in \mathcal{Y}} F_{\text{mm-reg}}(x, y) \text{ and } F^{\mathsf{y}}_{\text{mm-reg}}(y) := \min_{x \in \mathcal{X}} F_{\text{mm-reg}}(x, y).$$

Fact 3 shows $F^{\mathsf{x}}_{\text{mm-reg}}$ is μ^{x}-strongly convex and $F^{\mathsf{y}}_{\text{mm-reg}}$ is μ^{y}-strongly concave, so

$$\begin{aligned} \mathrm{Gap}_{F_{\text{mm-reg}}}(x_0, y_0) &= \left(F^{\mathsf{x}}_{\text{mm-reg}}(x_0) - F^{\mathsf{x}}_{\text{mm-reg}}(x_\star)\right) + \left(F^{\mathsf{y}}_{\text{mm-reg}}(y_\star) - F^{\mathsf{y}}_{\text{mm-reg}}(y_0)\right) \\ &\ge \mu^{\mathsf{x}} V_{x_0}(x^\star) + \mu^{\mathsf{y}} V_{y_0}(y^\star). \end{aligned}$$

■

Lemma 13 *Let $z \in \mathcal{Z}$ have*

$$V_z^r(z_\star) \le \left(\frac{\mu^{\mathsf{x}} + L^{\mathsf{x}} + \Lambda^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\mu^{\mathsf{y}} + L^{\mathsf{y}} + \Lambda^{\mathsf{yy}}}{\mu^{\mathsf{y}}} + \frac{(\Lambda^{\mathsf{xy}})^2}{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}\right) \cdot \frac{\epsilon}{2},$$

for $z_\star$ the solution to (12). Then,

$$\mathrm{Gap}_{F_{\text{mm-reg}}}(z^{\mathsf{x}}, z^{\mathsf{y}}) \le \epsilon.$$

Proof We follow the notation of Lemma 12. From Fact 3 we know $F^{\mathsf{x}}_{\text{mm-reg}}$ is $\mathcal{L}^{\mathsf{x}}$-smooth and $F^{\mathsf{y}}_{\text{mm-reg}}$ is $\mathcal{L}^{\mathsf{y}}$-smooth, where

$$\mathcal{L}^{\mathsf{x}} := \mu^{\mathsf{x}} + L^{\mathsf{x}} + \Lambda^{\mathsf{xx}} + \frac{(\Lambda^{\mathsf{xy}})^2}{\mu^{\mathsf{y}}} \text{ and } \mathcal{L}^{\mathsf{y}} := \mu^{\mathsf{y}} + L^{\mathsf{y}} + \Lambda^{\mathsf{yy}} + \frac{(\Lambda^{\mathsf{xy}})^2}{\mu^{\mathsf{x}}},$$

under Assumption 1. Moreover, by Lemma 4 and the definition of saddle points, $x_\star := z^{\mathsf{x}}_\star$ is the minimizer to $F^{\mathsf{x}}_{\text{mm-reg}}$, and $y_\star := z^{\mathsf{y}}_\star$ is the maximizer to $F^{\mathsf{y}}_{\text{mm-reg}}$. We conclude via

$$\begin{aligned}
\mathrm{Gap}_{F_{\text{mm-reg}}}(z^{\mathsf{x}}, z^{\mathsf{y}}) &= \left(F^{\mathsf{x}}_{\text{mm-reg}}(x) - F^{\mathsf{x}}_{\text{mm-reg}}(x_\star)\right) + \left(F^{\mathsf{y}}_{\text{mm-reg}}(y_\star) - F^{\mathsf{y}}_{\text{mm-reg}}(z^{\mathsf{y}})\right) \\
&\le \left(\mu^{\mathsf{x}} + L^{\mathsf{x}} + \Lambda^{\mathsf{xx}} + \frac{(\Lambda^{\mathsf{xy}})^2}{\mu^{\mathsf{y}}}\right)\|x - x_\star\|^2 \\
&+ \left(\mu^{\mathsf{y}} + L^{\mathsf{y}} + \Lambda^{\mathsf{yy}} + \frac{(\Lambda^{\mathsf{xy}})^2}{\mu^{\mathsf{x}}}\right)\|y - y_\star\|^2 \\
&\le 2\left(\frac{\mu^{\mathsf{x}} + L^{\mathsf{x}} + \Lambda^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\mu^{\mathsf{y}} + L^{\mathsf{y}} + \Lambda^{\mathsf{yy}}}{\mu^{\mathsf{y}}} + \frac{(\Lambda^{\mathsf{xy}})^2}{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}\right) V_z^r(z_\star) \le \epsilon.
\end{aligned}$$

The first inequality was smoothness of $F^{\mathsf{x}}_{\text{mm-reg}}$ and $F^{\mathsf{y}}_{\text{mm-reg}}$ (where we used that the gradients at $x_\star$ and $y_\star$ vanish because the optimization problems they solve are over unconstrained domains), and the last inequality was nonnegativity of Bregman divergences. ■

Appendix E. Finite sum optimization

In this section, we give an algorithm for efficiently finding an approximate minimizer of the following finite sum optimization problem:

$$F_{\text{fs}}(x) := \frac{1}{n}\sum_{i\in[n]} f_i(x). \tag{23}$$

Here and throughout this section $f_i : \mathcal{X} \to \mathbb{R}$ is a differentiable, convex function for all $i \in [n]$. For the remainder, we focus on algorithms for solving the following regularized formulation of (23):

$$\min_{x\in\mathcal{X}} F_{\text{fs-reg}}(x) \text{ for } F_{\text{fs-reg}}(x) := \frac{1}{n}\sum_{i\in[n]} f_i(x) + \frac{\mu}{2}\|x\|^2. \tag{24}$$

As in Section 2, to solve an instance of (23) where each f_i is μ-strongly convex, we may instead equivalently solve (24) by reparameterizing $f_i \leftarrow f_i - \frac{\mu}{2}\|\cdot\|^2$ for all $i \in [n]$. We further remark that our algorithms extend to solve instances of (23) where F_{fs} is μ-strongly convex in $\|\cdot\|$, but individual summands are not. We provide this result at the end of the section in Corollary 27.

In designing methods for solving (24) we make the following additional regularity assumptions.

Assumption 2 *For all $i \in [n]$, f_i is L_i-smooth.*

The remainder of this section is organized as follows.

(1) In Appendix E.1, we state a primal-dual formulation of (24) which we will apply our methods to, and prove that its solution also yields a solution to (24).

(2) In Appendix E.2, we give our algorithm and prove it is efficiently implementable.

(3) In Appendix E.3, we prove the convergence rate of our algorithm.

(4) In Appendix E.4, we state and prove our main result, Theorem 25.

E.1. Setup

To solve (24), we instead find a saddle point to the primal-dual function

$$F_{\text{fs-pd}}\left(z\right) := \frac{1}{n}\sum_{i\in[n]}\left(\left\langle z^{\mathsf{f}_i^*}, z^{\mathsf{x}}\right\rangle - f_i^*(z^{\mathsf{f}_i^*})\right) + \frac{\mu}{2}\left\|z^{\mathsf{x}}\right\|^2. \tag{25}$$

We denote the domain of $F_{\text{fs-pd}}$ by $\mathcal{Z} := \mathcal{X} \times (\mathcal{X}^*)^n$. For $z \in \mathcal{Z}$, we refer to its blocks by $(z^{\mathsf{x}}, \{z^{\mathsf{f}_i^*}\}_{i\in[n]})$. The primal-dual function $F_{\text{fs-pd}}$ is related to $F_{\text{fs-reg}}$ in the following way.

Lemma 17 *Let $z_\star$ be the saddle point to (25). Then, $z_\star^{\mathsf{x}}$ is a minimizer of (24).*

Proof By performing the maximization over each $z^{\mathsf{f}_i^*}$, we see that the problem of computing a minimizer to the objective in (25) is equivalent to

$$\min_{z^{\mathsf{x}}\in\mathcal{X}} \frac{\mu}{2}\left\|z^{\mathsf{x}}\right\|^2 + \frac{1}{n}\sum_{i\in[n]}\left(\max_{z^{\mathsf{f}_i^*}\in\mathcal{X}^*}\left\langle z^{\mathsf{f}_i^*}, z^{\mathsf{x}}\right\rangle - f_i^*(z^{\mathsf{f}_i^*})\right).$$

By Item 2 in Fact 1, this is the same as (24). ■

As in Section 2.1, it will be convenient to define the convex function $r : \mathcal{Z} \to \mathbb{R}$, which combines the (unsigned) separable components of $F_{\text{fs-pd}}$:

$$r\left(z\right) := \frac{\mu}{2}\left\|z^{\mathsf{x}}\right\|^2 + \frac{1}{n}\sum_{i\in[n]} f_i^*(z^{\mathsf{f}_i^*}). \tag{26}$$

Again, r serves as a regularizer in our algorithm. We next define Φ, the gradient operator of $F_{\text{fs-pd}}$:

$$\Phi(z) := \left(\frac{1}{n}\sum_{i\in[n]} z^{\mathsf{f}_i^*} + \mu z^{\mathsf{x}}, \left\{\frac{1}{n}\left(\nabla f_i^*(z^{\mathsf{f}_i^*}) - z^{\mathsf{x}}\right)\right\}_{i\in[n]}\right). \tag{27}$$

By construction, Φ is 1-strongly monotone with respect to r.

Lemma 18 (Strong monotonicity) *Define $\Phi : \mathcal{Z} \to \mathcal{Z}^*$ as in (27), and define $r : \mathcal{Z} \to \mathbb{R}$ as in (26). Then Φ is 1-strongly-monotone with respect to r.*

Proof The proof is identical to Lemma 5 without the Φ^h term: the bilinear component cancels in the definition of strong monotonicity, and the remaining part is exactly the gradient of r. ■

E.2. Algorithm

Our algorithm is an instantiation of *randomized mirror prox* Cohen et al. (2021) stated as Algorithm 3 below, an extension to mirror prox allowing for randomized gradient estimators. We note that the operators Φ_i need only be defined on iterates of the algorithm.

Algorithm 3: RAND-MIRROR-PROX($\{\Phi_i\}_{i\in[n]}, w_0$): Randomized mirror prox Cohen et al. (2021)

```
Input: Convex $r : \mathcal{Z} \to \mathbb{R}$, probability distribution $p : [n] \to \mathbb{R}_{\geq 0}$ with $\sum_{i\in[n]} p_i = 1$,
operators $\{\Phi_i\}_{i\in[n]} : \mathcal{Z} \to \mathcal{Z}^*$, $z_0 \in \mathcal{Z}$;
Parameter(s): $\lambda > 0$, $S \in \mathbb{N}$ ;
for $0 \leq s < S$ do
    Sample $i \sim p$;
    $w_{s+1/2} \leftarrow \text{Prox}^r_{w_t}(\frac{1}{\lambda}\Phi_i(w_s))$;
    $w_{s+1} \leftarrow \text{Prox}^r_{w_t}(\frac{1}{\lambda}\Phi_i(w_{s+1/2}))$;
end
```

We provide the following result from Cohen et al. (2021) giving a guarantee on Algorithm 3.

Proposition 19 (Proposition 2, Cohen et al. (2021)) *Suppose $\{\Phi_i\}_{i\in[n]}$ are defined so that in each iteration s, for all $u \in \mathcal{Z}$, there exists a point $\bar{w}_s \in \mathcal{Z}$ and a monotone operator $\Phi : \mathcal{Z} \to \mathcal{Z}^*$ such that (where all expectations fix w_s, and condition only on the randomness in iteration s)*

$$\begin{gathered}\mathbb{E}_{i\sim p}\left[\left\langle \Phi_i(w_{s+1/2}), w_{s+1/2} - u\right\rangle\right] = \langle \Phi(\bar{w}_s), \bar{w}_s - u\rangle \text{ for all } u \in \mathcal{Z},\\ \mathbb{E}_{i\sim p}\left[\left\langle \Phi_i(w_{s+1/2}) - \Phi_i(w_s), w_{s+1/2} - w_{s+1}\right\rangle\right] \leq \lambda \mathbb{E}_{i\sim p}\left[V^r_{w_s}(w_{s+1/2}) + V^r_{w_{s+1/2}}(w_{s+1})\right].\end{gathered} \tag{28}$$

Then (where the expectation below is taken over the randomness of the entire algorithm):

$$\mathbb{E}\left[\frac{1}{S}\sum_{0\leq s<S} \langle \Phi(\bar{w}_s), \bar{w}_s - u\rangle\right] \leq \frac{\lambda V^r_{w_0}(u)}{S}, \text{ for all } u \in \mathcal{Z}.$$

The first condition in (28) is an "unbiasedness" requirement on the operators $\{\Phi_i\}_{i\in[n]}$ with respect to the operator Φ, for which we wish to conclude a regret guarantee. The second posits that relative Lipschitzness (Definition 6) holds in an expected sense. We recall that Algorithm 3 requires us to specify a set of sampling probabilities $\{p_i\}_{i\in[n]}$. We define

$$p_i := \frac{\sqrt{L_i}}{2\sum_{j\in[n]}\sqrt{L_j}} + \frac{1}{2n} \text{ for all } i \in [n]. \tag{29}$$

This choice crucially ensures that all $p_i \geq \frac{1}{2n}$, and that all $\frac{\sqrt{L_i}}{p_i} \leq 2\sum_{j\in[n]}\sqrt{L_j}$.

Our algorithm, Algorithm 4, recursively applies Algorithm 3 to the operator-pair (Φ, r) defined in (27) and (26), for an appropriate specification of $\{\Phi_i\}_{i\in[n]}$. We give this implementation as pseudocode in Algorithms 4 and 5 below, and show that Algorithm 5 is a correct implementation of Algorithm 3 with respect to our specified $\{\Phi_i\}_{i\in[n]}$ in the remainder of the section.

Algorithm 4: FINITE-SUM-SOLVE($F_{\text{fs-reg}}, x_0$): Finite sum optimization

```
Input: (24) satisfying Assumption 2, $x_0 \in \mathcal{X}$;
Parameter(s): $T \in \mathbb{N}$;
$z_0^{\mathsf{x}} \leftarrow x_0$, $z_0^{\mathsf{f_i}} \leftarrow x_0$, $z_0^{\mathsf{f_i^*}} \leftarrow \nabla f_i(x_0)$ for all $i \in [n]$;
for $0 \le t < T$ do
    $z_{t+1} \leftarrow$ FINITE-SUM-ONE-PHASE($F_{\text{fs-reg}}, z_t$);
end
```

Algorithm 5: FINITE-SUM-ONE-PHASE($F_{\text{fs-reg}}, w_0$): Finite sum optimization subroutine

```
Input: (24) satisfying Assumption 2, $w_0 \in \mathcal{Z}$ specified by $w_0^{\mathsf{x}}$, $\{w_0^{\mathsf{f_i}}\}_{i\in[n]} \in \mathcal{X}$;
Parameter(s): $\lambda \ge 2$, $S \in \mathbb{N}$;
Sample $0 \le \sigma < S$ uniformly at random;
for $0 \le s \le \sigma$ do
    Sample $j \in [n]$ according to $p$ defined in (29);
    $w_{s+1/2}^{\mathsf{x}} \leftarrow w_s^{\mathsf{x}} - \frac{1}{\lambda\mu}(\mu w_s^{\mathsf{x}} + \frac{1}{n}\sum_{i\in[n]} \nabla f_i(w_s^{\mathsf{f_i}}))$;
    $w_{s+1/2}^{\mathsf{f_j}} \leftarrow (1 - \frac{1}{\lambda p_j})w_s^{\mathsf{f_j}} + \frac{1}{\lambda p_j} w_s^{\mathsf{x}}$;
    $w_{s+1/2}^{\mathsf{f_i}} \leftarrow w_s^{\mathsf{f_i}}$ for all $i \ne j$;
    $\Delta_s \leftarrow \nabla f_j(w_{s+1/2}^{\mathsf{f_j}}) - \nabla f_j(w_s^{\mathsf{f_j}})$;
    $w_{s+1}^{\mathsf{x}} \leftarrow w_s^{\mathsf{x}} - \frac{1}{\lambda\mu}(\mu w_{s+1/2}^{\mathsf{x}} + \frac{1}{n}\sum_{i\in[n]} \nabla f_i(w_s^{\mathsf{f_i}}) + \frac{1}{np_j}\Delta_s)$;
    $w_{s+1}^{\mathsf{f_j}} \leftarrow w_s^{\mathsf{f_j}} + \frac{1}{\lambda p_j}(w_{s+1/2}^{\mathsf{x}} - w_{s+1/2}^{\mathsf{f_j}})$;
    $w_{s+1}^{\mathsf{f_i}} \leftarrow w_s^{\mathsf{f_i}}$ for all $i \ne j$;
end
Return: $(w_{\sigma+1/2}^{\mathsf{x}}, \{\nabla f_i((1 - \frac{1}{\lambda p_i})w_\sigma^{\mathsf{f_i}} + \frac{1}{\lambda p_i} w_\sigma^{\mathsf{x}})\}_{i\in[n]})$
```

We next describe the operators $\{\Phi_i\}_{i\in[n]}$ used in our implementation of Algorithm 3. Fix some $0 \le s < S$, and consider some iterates $\{w, w_{\mathsf{aux}}(j)\} := \{w_s, w_{s+1/2}\}$ of Algorithm 3 (where we use the notation (j) to mean the iterate that would be taken if $j \in [n]$ was sampled in iteration s, and we drop the subscript s for simplicity since we only focus on one iteration). We denote the $\mathcal{X}$ block of $w_{\mathsf{aux}}(j)$ by $w_{\mathsf{aux}}^{\mathsf{x}}$, since (as made clear in the following) conditioned on w, $w_{\mathsf{aux}}^{\mathsf{x}}$ is always the same regardless of the sampled $j \in [n]$. For all $j \in [n]$, we then define the operators

$$\Phi_j(w) := \left(\frac{1}{n}\sum_{i\in[n]} w^{\mathsf{f_i^*}} + \mu w^{\mathsf{x}}, \left\{\frac{1}{np_i}(\nabla f_i^*(w^{\mathsf{f_i^*}}) - w^{\mathsf{x}}) \cdot \mathbf{1}_{i=j}\right\}\right),$$

$$\Phi_j(w_{\mathsf{aux}}(j)) := \left(\frac{1}{n}\sum_{i\in[n]} w^{\mathsf{f_i^*}} + \frac{1}{np_j}\left(w_{\mathsf{aux}}^{\mathsf{f_j^*}}(j) - w^{\mathsf{f_j^*}}\right) + \mu w_{\mathsf{aux}}^{\mathsf{x}}, \left\{\frac{1}{np_i}\left(\nabla f_i^*\left(w_{\mathsf{aux}}^{\mathsf{f_i^*}}(i)\right) - w_{\mathsf{aux}}^{\mathsf{x}}\right) \cdot \mathbf{1}_{i=j}\right\}\right), \tag{30}$$

where $\mathbf{1}_{i=j}$ is a zero-one indicator. In other words, $\Phi_j(w)$ and $\Phi_j(w_{\mathsf{aux}}(j))$ both only have two nonzero blocks, corresponding to the $\mathcal{X}$ and j^{th} $\mathcal{X}^*$ blocks. We record the following useful observation about our randomized operators (30), in accordance with the first condition in (28). To give a brief interpretation of our "aggregate point" defined in (31), the $\mathcal{X}$ coordinate is updated deterministically from w^{x} according to the corresponding block of Φ, and every dual block $j \in [n]$ of $\bar{w}$ is set to the corresponding dual block had j been sampled in that step.

Lemma 20 (Expected regret) *Define $\{\Phi_j\}_{j\in[n]} : \mathcal{Z} \to \mathcal{Z}^*$ as in (30), and the "aggregate point"*

$$\bar{w} := \left(w^{\mathsf{x}}_{\mathsf{aux}}, \left\{w^{\mathsf{f}^*_{\mathsf{j}}}_{\mathsf{aux}}(j)\right\}_{j\in[n]}\right). \tag{31}$$

Then, for all $u \in \mathcal{Z}$, defining Φ as in (27),

$$\mathbb{E}_{j\sim p}\left[\langle \Phi_j(w_{\mathsf{aux}}(j)), w_{\mathsf{aux}}(j) - u\rangle\right] = \langle \Phi(\bar{w}), \bar{w} - u\rangle .$$

Proof We expand the expectation, using (30) and taking advantage of the sparsity of Φ_j:

$$\begin{aligned}
&\mathbb{E}_{j\sim p}\left[\langle \Phi_j(w_{\mathsf{aux}}(j)), w_{\mathsf{aux}}(j) - u\rangle\right] \\
&\quad= \left\langle \sum_{j\in[n]} p_j \left(\frac{1}{n}\sum_{i\in[n]} w^{\mathsf{f}^*_{\mathsf{i}}} + \frac{1}{np_j}\left(w^{\mathsf{f}^*_{\mathsf{j}}}_{\mathsf{aux}}(j) - w^{\mathsf{f}^*_{\mathsf{j}}}\right) + \mu w^{\mathsf{x}}_{\mathsf{aux}}\right), w^{\mathsf{x}}_{\mathsf{aux}} - u^{\mathsf{x}}\right\rangle \\
&\qquad+ \sum_{j\in[n]} p_j \left\langle \frac{1}{np_j}\left(\nabla f^*_j\left(w^{\mathsf{f}^*_{\mathsf{j}}}_{\mathsf{aux}}(j)\right) - w^{\mathsf{x}}_{\mathsf{aux}}\right), w^{\mathsf{f}^*_{\mathsf{j}}}_{\mathsf{aux}}(j) - u^{\mathsf{f}^*_{\mathsf{j}}}\right\rangle \\
&\quad= \left\langle \frac{1}{n}\sum_{j\in[n]} w^{\mathsf{f}^*_{\mathsf{j}}}_{\mathsf{aux}}(j) + \mu w^{\mathsf{x}}_{\mathsf{aux}}, w^{\mathsf{x}}_{\mathsf{aux}} - u^{\mathsf{x}}\right\rangle \\
&\qquad+ \sum_{j\in[n]} \left\langle \frac{1}{n}\left(\nabla f^*_j\left(w^{\mathsf{f}^*_{\mathsf{j}}}_{\mathsf{aux}}(j)\right) - w^{\mathsf{x}}_{\mathsf{aux}}\right), w^{\mathsf{f}^*_{\mathsf{j}}}_{\mathsf{aux}}(j) - u^{\mathsf{f}^*_{\mathsf{j}}}\right\rangle = \langle \Phi(\bar{w}), \bar{w} - u\rangle .
\end{aligned}$$

■

We conclude this section by demonstrating that Algorithm 5 is an appropriate implementation of Algorithm 3.

Lemma 21 (Implementation) *Algorithm 5 implements Algorithm 3 on $(\{\Phi_i\}_{i\in[n]}, r)$ defined in (30), (26), for σ iterations, and returns $\bar{w}_\sigma$, following the definition (31). Each iteration $s > 0$ is implementable in $O(1)$ gradient calls to some f_i, and $O(1)$ vector operations on $\mathcal{X}$.*

Proof Let $\{w_s, w_{s+1/2}\}_{0\le s\le\sigma}$ be the iterates of Algorithm 3. We will inductively show that Algorithm 5 preserves the invariants

$$w_s = \left(w^{\mathsf{x}}_s, \left\{\nabla f_i(w^{\mathsf{f}_{\mathsf{i}}}_s)\right\}_{i\in[n]}\right),\quad w_{s+1/2} = \left(w^{\mathsf{x}}_s, \left\{\nabla f_i(w^{\mathsf{f}_{\mathsf{i}}}_{s+1/2})\right\}_{i\in[n]}\right)$$

for all $0 \le s \le \sigma$. Once we prove this claim, it is clear from inspection that Algorithm 5 implements Algorithm 3 and returns $\bar{w}_\sigma$, upon recalling the definitions (30), (26), and (31).

The base case of our induction follows from the initialization guarantee of Algorithm 4 in Algorithm 5. Next, suppose for some $0 \le s \le \sigma$, we have $w_s^{\mathsf{f}_\mathsf{i}^*} = \nabla f(w_s^{\mathsf{f}_\mathsf{i}})$ for all $i \in [n]$. By the updates in Algorithm 3, if $j \in [n]$ was sampled on iteration s,

$$\begin{aligned} w_{s+1/2}^{\mathsf{f}_\mathsf{j}^*} &\leftarrow \text{argmin}_{w^{\mathsf{f}_\mathsf{j}^*} \in \mathcal{X}^*} \left\{ \frac{1}{\lambda n p_j} \left\langle w_s^{\mathsf{f}_\mathsf{j}^*} - w_s^{\times}, w^{\mathsf{f}_\mathsf{j}^*} \right\rangle - \frac{1}{n} \left\langle w_s^{\mathsf{f}_\mathsf{j}^*}, w^{\mathsf{f}_\mathsf{j}^*} \right\rangle + \frac{1}{n} f_j^* \left(w^{\mathsf{f}_\mathsf{j}^*} \right) \right\} \\ &= \text{argmax}_{w^{\mathsf{f}_\mathsf{j}^*} \in \mathcal{X}^*} \left\{ \left\langle \left(1 - \frac{1}{\lambda p_j}\right) w_s^{\mathsf{f}_\mathsf{j}^*} + \frac{1}{\lambda p_j} w_s^{\times}, w^{\mathsf{f}_\mathsf{j}^*} \right\rangle - f_j^* \left(w^{\mathsf{f}_\mathsf{j}^*} \right) \right\} \\ &= \nabla f_j \left(\left(1 - \frac{1}{\lambda p_j}\right) w_s^{\mathsf{f}_\mathsf{j}^*} + \frac{1}{\lambda p_j} w_s^{\times} \right). \end{aligned}$$

Here, we used the first item in Fact 1 in the last line. Hence, the update to $w_{s+1/2}^{\mathsf{f}_\mathsf{j}^*}$ in Algorithm 5 preserves our invariant, and all other $w_{s+1/2}^{\mathsf{f}_\mathsf{i}^*}$, $i \neq j$ do not change by sparsity of Φ_j. An analogous argument shows the update to each $w_{s+1}^{\mathsf{f}_\mathsf{i}^*}$ preserves our invariant. Finally, in every iteration $s > 0$, the updates to $w_{s+1/2}^{\times}$ and $w_{s+1}^{\times}$ only require evaluating one new gradient each, by 1-sparsity of the dual block updates in the prior iteration. ■

E.3. Convergence analysis

In this section, we prove a convergence result on Algorithm 5 via an application of Proposition 19. To begin, we require a bound on the quantity λ in (28).

Lemma 22 (Expected relative Lipschitzness) *Define $\{\Phi_j\}_{j \in [n]} : \mathcal{Z} \to \mathcal{Z}^*$ as in (30), and define $r : \mathcal{Z} \to \mathbb{R}$ as in (26). Letting $w_+(j)$ be w_{s+1} in Algorithm 3 if $j \in [n]$ was sampled in iteration s,*

$$\mathbb{E}_{j \sim p} \left[\langle \Phi_j(w_{\mathsf{aux}}(j)) - \Phi_j(w), w_{\mathsf{aux}}(j) - w_+(j) \rangle \right] \le \mathbb{E}_{j \sim p} \left[V_w^r \left(w_{\mathsf{aux}}(j) \right) + V_{w_{\mathsf{aux}}(j)}^r \left(w_+(j) \right) \right]$$

for

$$\lambda = 2n + \frac{2 \sum_{j \in [n]} \sqrt{L_j}}{\sqrt{n\mu}}. \tag{32}$$

Proof We begin by expanding the expectation of the left-hand side:

$$\begin{aligned} \mathbb{E}_{j \sim p} \left[\langle \Phi_j(w_{\mathsf{aux}}(j)) - \Phi_j(w), w_{\mathsf{aux}}(j) - w_+(j) \rangle \right] &= \mathbb{E}_{j \sim p} \left[\left\langle \mu w_{\mathsf{aux}}^{\times} - \mu w^{\times}, w_{\mathsf{aux}}^{\times} - w_+^{\times}(j) \right\rangle \right] \\ &+ \mathbb{E}_{j \sim p} \left[\frac{1}{n p_j} \left\langle \nabla f_j^* \left(w_{\mathsf{aux}}^{\mathsf{f}_\mathsf{j}^*}(j) \right) - \nabla f_j^* \left(w^{\mathsf{f}_\mathsf{j}^*} \right), w_{\mathsf{aux}}^{\mathsf{f}_\mathsf{j}^*}(j) - w_+^{\mathsf{f}_\mathsf{j}^*}(j) \right\rangle \right] \\ &+ \mathbb{E}_{j \sim p} \left[\frac{1}{n p_j} \left\langle w_{\mathsf{aux}}^{\mathsf{f}_\mathsf{j}^*}(j) - w^{\mathsf{f}_\mathsf{j}^*}, w_{\mathsf{aux}}^{\times} - w_+^{\times}(j) \right\rangle \right] \\ &+ \mathbb{E}_{j \sim p} \left[\frac{1}{n p_j} \left\langle w^{\times} - w_{\mathsf{aux}}^{\times}, w_{\mathsf{aux}}^{\mathsf{f}_\mathsf{j}^*}(j) - w_+^{\mathsf{f}_\mathsf{j}^*}(j) \right\rangle \right]. \end{aligned} \tag{33}$$

To bound the first two lines of (33), fix some $j \in [n]$. We apply Lemma 7 to the functions $\frac{\mu}{2} \|\cdot\|^2$ and $\frac{1}{n} \nabla f_j^*$, and use nonnegativity of Bregman divergences, to conclude

$$\begin{aligned} &\left\langle \mu w_{\mathsf{aux}}^{\times} - \mu w^{\times}, w_{\mathsf{aux}}^{\times} - w_+^{\times}(j) \right\rangle + \frac{1}{n p_j} \left\langle \nabla f_j^* \left(w_{\mathsf{aux}}^{\mathsf{f}_\mathsf{j}^*}(j) \right) - \nabla f_j^* \left(w^{\mathsf{f}_\mathsf{j}^*} \right), w_{\mathsf{aux}}^{\mathsf{f}_\mathsf{j}^*}(j) - w_+^{\mathsf{f}_\mathsf{j}^*}(j) \right\rangle \\ &\le 2n \left(V_w^r \left(w_{\mathsf{aux}}(j) \right) + V_{w_{\mathsf{aux}}(j)}^r \left(w_+(j) \right) \right). \end{aligned}$$

In particular, we used $\frac{1}{p_j} \le 2n$ by assumption, and noted we only need to handle the case where the second inner product term above is positive (in the other case, the above inequality is clearly true). Hence, taking expectations the first two lines in (33) contribute $2n$ to λ in the final bound.

To bound the last two lines of (33), fix $j \in [n]$. By applying Item (1) in Lemma 16 to the pair $(\frac{\mu}{2}\left\|\cdot\right\|^2, nf_i)$, we have

$$\frac{1}{n}\left\langle w_{\textsf{aux}}^{\textsf{f}_\textsf{j}^*}(j) - w^{\textsf{f}_\textsf{j}^*}, w_{\textsf{aux}}^{\textsf{x}} - w_+^{\textsf{x}}(j)\right\rangle + \frac{1}{n}\left\langle w^{\textsf{x}} - w_{\textsf{aux}}^{\textsf{x}}, w_{\textsf{aux}}^{\textsf{f}_\textsf{j}^*}(j) - w_+^{\textsf{f}_\textsf{j}^*}(j)\right\rangle$$

$$\le \frac{1}{n}\sqrt{\frac{nL_j}{\mu}}\left(\mu V_{w_{\textsf{aux}}^{\textsf{x}}}\left(w_+^{\textsf{x}}(j)\right) + V^{f_j^*}_{w^{\textsf{f}_\textsf{j}^*}}\left(w_{\textsf{aux}}^{\textsf{f}_\textsf{j}^*}(j)\right)\right) + \frac{1}{n}\sqrt{\frac{nL_j}{\mu}}\left(\mu V_{w^{\textsf{x}}}\left(w_{\textsf{aux}}^{\textsf{x}}\right) + V^{f_j^*}_{w_{\textsf{aux}}^{\textsf{f}_\textsf{j}^*}(j)}\left(w_+^{\textsf{f}_\textsf{j}^*}(j)\right)\right)$$

$$= \sqrt{\frac{L_j}{n\mu}}\left(V_w^r\left(w_{\textsf{aux}}(j)\right) + V_{w_{\textsf{aux}}(j)}^r\left(w_+(j)\right)\right).$$

Using $\frac{\sqrt{L_i}}{p_i} \le 2\sum_{j\in[n]}\sqrt{L_j}$ and taking expectations over the above display,

$$\mathbb{E}_{j\sim p}\left[\frac{1}{np_j}\left\langle w_{\textsf{aux}}^{\textsf{f}_\textsf{j}^*}(j) - w^{\textsf{f}_\textsf{j}^*}, w_{\textsf{aux}}^{\textsf{x}} - w_+^{\textsf{x}}(j)\right\rangle + \frac{1}{np_j}\left\langle w^{\textsf{x}} - w_{\textsf{aux}}^{\textsf{x}}, w_{\textsf{aux}}^{\textsf{f}_\textsf{j}^*}(j) - w_+^{\textsf{f}_\textsf{j}^*}(j)\right\rangle\right]$$

$$\le \frac{2\sum_{j\in[n]}\sqrt{L_j}}{\sqrt{n\mu}}\mathbb{E}_{j\sim p}\left[V_w^r\left(w_{\textsf{aux}}(j)\right) + V_{w_{\textsf{aux}}(j)}^r\left(w_+(j)\right)\right].$$

Hence, the last two lines in (33) contribute $\frac{2\sum_{j\in[n]}\sqrt{L_j}}{\sqrt{n\mu}}$ to λ in the final bound. ■

We next apply Proposition 19 to analyze the convergence of Algorithm 5.

Lemma 23 *Let $w_0 := (w_0^{\textsf{x}}, \{\nabla f_i(w_0^{\textsf{f}_\textsf{i}})\}_{i\in[n]})$, which is the input z_t to Algorithm 5 at iteration t. If $S \ge 2\lambda$ in Algorithm 5 with λ as in (32), then Algorithm 5 returns $\widetilde{w} \leftarrow \bar{w}_\sigma$ as defined in (31) such that for $z_\star$ as the saddle point to (25),*

$$\mathbb{E}V_{\widetilde{w}}^r(z_\star) \le \frac{1}{2}V_{w_0}^r(z_\star).$$

Proof We apply Proposition 19, where (28) is satisfied via Lemmas 20 and 22. By Proposition 19 with $u = z_\star$ and $S \ge 2\lambda$,

$$\mathbb{E}\left[\frac{1}{S}\sum_{0\le s<S}\left\langle \Phi(\bar{w}_s), \bar{w}_s - z_\star\right\rangle\right] \le \frac{1}{2}V_{w_0}^r(z_\star).$$

Moreover, since σ is uniformly chosen in $[0, S-1]$, we have

$$\mathbb{E}\left[\left\langle \Phi(\bar{w}_\sigma), \bar{w}_\sigma - z_\star\right\rangle\right] \le \frac{1}{2}V_{w_0}^r(z_\star).$$

Finally, Lemma 21 shows that (an implicit representation of) $\bar{w}_\sigma$ is indeed returned. We conclude by applying Lemma 18 and using that $z_\star$ solves the VI in Φ, yielding

$$\mathbb{E}\left[\left\langle \Phi(\bar{w}_\sigma), \bar{w}_\sigma - z_\star\right\rangle\right] \ge \mathbb{E}\left[\left\langle \Phi(\bar{w}_\sigma) - \Phi(z_\star), \bar{w}_\sigma - z_\star\right\rangle\right] \ge V_{\bar{w}_\sigma}^r(z_\star).$$

■

Finally, we provide a simple bound regarding initialization of Algorithm 4.

Lemma 24 *Let $x_0 \in \mathcal{X}$, and define*

$$z_0 := \left(x_0, \{\nabla f_i(x_0)\}_{i \in [n]} \right). \tag{34}$$

Moreover, suppose that for $x_\star$ the solution to (24), $F_{\text{fs-reg}}(x_0) - F_{\text{fs-reg}}(x_\star) \le \epsilon_0$. Then, letting $z_\star$ be the solution to (25), we have

$$V_{z_0}^r(z_\star) \le \left(1 + \frac{\sum_{i \in [n]} L_i}{n\mu}\right) \epsilon_0.$$

Proof By the characterization in Lemma 17, we have by Item 1 in Fact 1:

$$z_\star = \left(x_\star, \{\nabla f_i(x_\star)\}_{i \in [n]} \right).$$

Hence, we bound analogously to Lemma 12:

$$\begin{aligned}
V_{z_0}^r(z_\star) &\le \mu V_{x_0}(x_\star) + V_{x_\star}^{\frac{1}{n}\sum_{i\in[n]} f_i}(x_0) \\
&\le \mu V_{x_0}(x_\star) + \frac{\sum_{i \in [n]} L_i}{2n} \|x_0 - x_\star\|^2 \\
&\le \left(1 + \frac{\sum_{i \in [n]} L_i}{n\mu}\right) \mu V_{x_0}(x_\star) \le \left(1 + \frac{\sum_{i \in [n]} L_i}{n\mu}\right) \epsilon_0.
\end{aligned}$$

The last line applied strong convexity of $F_{\text{fs-reg}}$. ■

E.4. Main result

We now state and prove our main claim.

Theorem 25 *Suppose $F_{\text{fs-reg}}$ satisfies Assumption 2 and has minimizer $x_\star$, and suppose we have $x_0 \in \mathcal{X}$ such that $F_{\text{fs-reg}}(x_0) - F_{\text{fs-reg}}(x_\star) \le \epsilon_0$. Algorithm 4 using Algorithm 5 with λ as in (32) returns $x \in \mathcal{X}$ with $\mathbb{E}F_{\text{fs-reg}}(x) - F_{\text{fs-reg}}(x_\star) \le \epsilon$ in N_{tot} iterations, using a total of $O(N_{\text{tot}})$ gradient calls each to some f_i for $i \in [n]$, where*

$$N_{\text{tot}} = O\left(\kappa_{\text{fs}} \log\left(\frac{\kappa_{\text{fs}}\epsilon_0}{\epsilon}\right)\right), \textit{ for } \kappa_{\text{fs}} := n + \frac{\sum_{i \in [n]} \sqrt{L_i}}{\sqrt{n\mu}}. \tag{35}$$

Proof By Lemma 17, the point $x_\star$ is consistent between (23) and (25). We run Algorithm 4 with

$$T = O\left(\log\left(\frac{\kappa_{\text{fs}}\epsilon_0}{\epsilon}\right)\right).$$

By recursively applying Lemma 23 for T times, we obtain a point z such that

$$\mathbb{E}V_z^r(z_\star) \le \frac{\epsilon\mu}{\mathcal{L}} \text{ for } \mathcal{L} = \mu + \frac{1}{n}\sum_{i \in [n]} L_i,$$

and hence applying $\mathcal{L}$-smoothness of $F_{\text{fs-reg}}$ and optimality of $z_\star^\times$ yields the claim. The complexity follows from Lemma 9, and spending $O(n)$ gradient evaluations on the first and last iterates of each call to Algorithm 5 (which is subsumed by the fact that $S = \Omega(n)$). ■

Algorithm 6: REDX-CONVEX: Strongly convex optimization reduction

```
Input: $\mu$-strongly convex $f : \mathcal{X} \to \mathbb{R}$, $x_0 \in \mathcal{X}$
Parameter(s): $K \in \mathbb{N}$
for $0 \le k < K$ do
    $x_{k+1} \leftarrow$ any (possibly random) point satisfying
        $\mathbb{E}V_{x_{k+1}}(x^\star_{k+1}) \le \frac{1}{4}V_{x_k}(x^\star_{k+1})$, where $x^\star_{k+1} := \text{argmin}_{x\in\mathcal{X}} f(x) + \frac{\mu}{4}V_{x_k}(x)$
end
```

We now revisit the problem (23), and design a method which applies when F_{fs} is strongly convex but no summand necessarily is. To do so, we give the following generic reduction for strongly convex optimization in the form of an algorithm. Similar reductions are standard in the literature (Frostig et al., 2015), but we include the algorithm and full analysis here for completeness.

Lemma 26 *In Algorithm 6, letting $x_\star$ minimize f, we have for every $k \in [K]$:*

$$\mathbb{E}V_{x_k}(x_\star) \le \frac{1}{2^k}V_{x_0}(x_\star).$$

Proof By applying the optimality condition on $x^\star_{k+1}$, strong convexity of f, and (18),

$$\begin{aligned}
\left\langle \nabla f(x^\star_{k+1}), x^\star_{k+1} - x_\star \right\rangle &\le \frac{\mu}{4}\left\langle x_k - x^\star_{k+1}, x^\star_{k+1} - x_\star \right\rangle \\
\implies \mu V_{x^\star_{k+1}}(x_\star) &\le f(x^\star_{k+1}) - f(x_\star) \\
&\le \left\langle \nabla f(x^\star_{k+1}), x^\star_{k+1} - x_\star \right\rangle \\
&\le \frac{\mu}{4}V_{x_k}(x_\star) - \frac{\mu}{4}V_{x^\star_{k+1}}(x_\star) - \frac{\mu}{4}V_{x_k}(x^\star_{k+1}).
\end{aligned}$$

Further by the triangle inequality and $(a+b)^2 \le 2a^2 + 2b^2$, we have

$$V_{x_{k+1}}(x_\star) \le 2V_{x_{k+1}}(x^\star_{k+1}) + 2V_{x^\star_{k+1}}(x_\star).$$

Hence, combining these pieces,

$$\begin{aligned}
\mathbb{E}V_{x_{k+1}}(x_\star) &\le 2V_{x^\star_{k+1}}(x_\star) + 2\mathbb{E}V_{x_{k+1}}(x^\star_{k+1}) \\
&\le 2V_{x^\star_{k+1}}(x_\star) + \frac{1}{2}V_{x_k}(x^\star_{k+1}) \\
&\le \frac{1}{2}V_{x_k}(x_\star) - \frac{1}{2}V_{x^\star_{k+1}}(x_\star) \le \frac{1}{2}V_{x_k}(x_\star).
\end{aligned}$$

■

We apply this reduction in order to prove Corollary 27.

Corollary 27 *Suppose the summands $\{f_i\}_{i\in[n]}$ in (23) satisfy Assumption 2, and F_{fs} is μ-strongly convex with minimizer $x_\star$. Further, suppose we have $x_0 \in \mathcal{X}$ such that $F_{\text{fs}}(x_0) - F_{\text{fs}}(x_\star) \le \epsilon_0$. Algorithm 6 using Algorithm 4 to implement steps returns $x \in \mathcal{X}$ with $\mathbb{E}F_{\text{fs}}(x) - F_{\text{fs}}(x_\star) \le \epsilon$ in N_{tot} iterations, using a total of $O(N_{\text{tot}})$ gradient calls each to some f_i for $i \in [n]$, where*

$$N_{\text{tot}} = O\left(\kappa_{\text{fs}} \log\left(\frac{\kappa_{\text{fs}}\epsilon_0}{\epsilon}\right)\right), \text{ for } \kappa_{\text{fs}} := n + \sum_{i\in[n]} \frac{\sqrt{L_i}}{\sqrt{n\mu}}.$$

Proof The overhead K is asymptotically the same here as the parameter T in Theorem 25, by analogous smoothness and strong convexity arguments. Moreover, we use Theorem 25 to solve each subproblem required by Algorithm 6; in particular, the subproblem is equivalent to approximately minimizing $F_{\text{fs}} + \frac{\mu}{8}\|\cdot\|^2$, up to a linear shift which does not affect any smoothness bounds, and a constant in the strong convexity. We note that we will initialize the subproblem solver in iteration k with x_k. We hence can set $T = 2$ and $S = O(\kappa_{\text{fs}})$, yielding the desired iteration bound. ■

Appendix F. Minimax finite sum optimization

In this section, we provide efficient algorithms for computing an approximate saddle point of the following minimax finite sum optimization problem:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} F_{\text{mmfs}}(x,y) := \frac{1}{n}\sum_{i\in[n]} \left(f_i(x) + h_i(x,y) - g_i(y)\right). \tag{36}$$

Here and throughout this section $\{f_i : \mathcal{X} \to \mathbb{R}\}_{i\in[n]}$, $\{g_i : \mathcal{Y} \to \mathbb{R}\}_{i\in[n]}$ are differentiable convex functions, and $\{h_i : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}\}_{i\in[n]}$ are differentiable convex-concave functions. For the remainder, we focus on algorithms for solving the following regularized formulation of (36):

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} F_{\text{mmfs-reg}}(x,y) := \frac{1}{n}\sum_{i\in[n]} \left(f_i(x) + h_i(x,y) - g_i(y)\right) + \frac{\mu^{\mathsf{x}}}{2}\|x\|^2 - \frac{\mu^{\mathsf{y}}}{2}\|y\|^2. \tag{37}$$

As in Section 2 and Appendix E, to instead solve an instance of (36) where each f_i is $2\mu^{\mathsf{x}}$-strongly convex and each g_i is $2\mu^{\mathsf{y}}$-strongly convex, we may instead equivalently solve (37) by reparameterizing $f_i \leftarrow f_i - \mu^{\mathsf{x}}\|\cdot\|^2$, $g_i \leftarrow g_i - \mu^{\mathsf{y}}\|\cdot\|^2$ for each $i \in [n]$. The extra factor of 2 is so we can make a strong convexity assumption in Assumption 3 about separable summands, which only affects our final bounds by constants. We further remark that our algorithms extend to solve instances of (36) where f, g is μ^{x} and μ^{y}-strongly convex in $\|\cdot\|$, but individual summands are not. We provide this result at the end of the section in Corollary 41.

In designing methods for solving (37) we make the following additional regularity assumptions.

Assumption 3 *We assume the following about (37) for all $i \in [n]$.*

(1) f_i is L_i^{x}-smooth and μ_i^{x}-strongly convex and g_i is L_i^{y}-smooth and μ_i^{y}-strongly convex.

(2) h_i has the following blockwise-smoothness properties: for all $u, v \in \mathcal{X}\times\mathcal{Y}$,

$$\begin{aligned} \|\nabla_x h_i(u) - \nabla_x h_i(v)\| &\le \Lambda_i^{\mathsf{xx}}\|u^{\mathsf{x}} - v^{\mathsf{x}}\| + \Lambda_i^{\mathsf{xy}}\|u^{\mathsf{y}} - v^{\mathsf{y}}\| \text{ and} \\ \|\nabla_y h_i(u) - \nabla_y h_i(v)\| &\le \Lambda_i^{\mathsf{xy}}\|u^{\mathsf{x}} - v^{\mathsf{x}}\| + \Lambda_i^{\mathsf{yy}}\|u^{\mathsf{y}} - v^{\mathsf{y}}\|. \end{aligned} \tag{38}$$

The remainder of this section is organized as follows.

(1) In Appendix F.1, we state a primal-dual formulation of (37) which we will apply our methods to, and prove that its solution also yields a solution to (37).

(2) In Appendix F.2, we give our algorithm, which is composed of an outer loop and an inner loop, and prove it is efficiently implementable.

(3) In Appendix F.3, we prove the convergence rate of our inner loop.

(4) In Appendix F.4, we prove the convergence rate of our outer loop.

(5) In Appendix F.5, we state and prove our main result, Theorem 39.

F.1. Setup

To solve (37), we will instead find a saddle point to the primal-dual function

$$\begin{aligned} F_{\text{mmfs-pd}}(z) := &\frac{\mu^{\mathsf{x}}}{2}\|z^{\mathsf{x}}\|^2 - \frac{\mu^{\mathsf{y}}}{2}\|z^{\mathsf{y}}\|^2 \\ &+ \frac{1}{n}\sum_{i\in[n]}\left(h_i(z^{\mathsf{x}}, z^{\mathsf{y}}) + \left\langle z^{\mathsf{f}_i^*}, z^{\mathsf{x}}\right\rangle - \left\langle z^{\mathsf{g}_i^*}, z^{\mathsf{y}}\right\rangle - f_i^*\left(z^{\mathsf{f}_i^*}\right) + g_i^*(z^{\mathsf{g}_i^*})\right). \end{aligned} \tag{39}$$

We denote the domain of $F_{\text{mmfs-pd}}$ by $\mathcal{Z} := \mathcal{X} \times \mathcal{Y} \times (\mathcal{X}^*)^n \times (\mathcal{Y}^*)^n$. For $z \in \mathcal{Z}$, we refer to its blocks by $(z^{\mathsf{x}}, z^{\mathsf{y}}, \{z^{\mathsf{f}_i^*}\}_{i\in[n]}, \{z^{\mathsf{g}_i^*}\}_{i\in[n]})$. The primal-dual function $F_{\text{mmfs-pd}}$ is related to the original function F_{mmfs} in the following way; we omit the proof, as it follows analogously to the proofs of Lemmas 4 and 17.

Lemma 28 *Let $z_\star = (z_\star^{\mathsf{x}}, z_\star^{\mathsf{y}}, \{z_\star^{\mathsf{f}_i^*}\}_{i\in[n]}, \{z_\star^{\mathsf{g}_i^*}\}_{i\in[n]})$ be the saddle point to (39). Then, $(z_\star^{\mathsf{x}}, z_\star^{\mathsf{y}})$ is a saddle point to (37).*

As in Section 2.1, it will be convenient to define the convex function $r : \mathcal{Z} \to \mathbb{R}$, which combines the (unsigned) separable components of $F_{\text{mmfs-pd}}$:

$$r\left(z^{\mathsf{x}}, z^{\mathsf{y}}, \left\{z^{\mathsf{f}_i^*}\right\}_{i\in[n]}, \left\{z^{\mathsf{g}_i^*}\right\}_{i\in[n]}\right) := \frac{\mu^{\mathsf{x}}}{2}\|z^{\mathsf{x}}\|^2 + \frac{\mu^{\mathsf{y}}}{2}\|z^{\mathsf{y}}\|^2 + \frac{1}{n}\sum_{i\in[n]} f_i^*\left(z^{\mathsf{f}_i^*}\right) + \frac{1}{n}\sum_{i\in[n]} g_i^*\left(z^{\mathsf{g}_i^*}\right). \tag{40}$$

Again, r serves as a regularizer in our algorithm. We next define $\Phi^{\text{mmfs-pd}}$, the gradient operator of $F_{\text{mmfs-pd}}$. We decompose $\Phi^{\text{mmfs-pd}}$ into three parts, roughly corresponding to the contribution from r, the contributions from the primal-dual representations of $\{f_i\}_{i\in[n]}$ and $\{g_i\}_{i\in[n]}$, and the contribution from $\{h_i\}_{i\in[n]}$. In particular, we define

$$\begin{aligned}
\Phi^{\text{mmfs-pd}}(z) &:= \nabla r(z) + \Phi^h(z) + \Phi^{\text{bilin}}(z),\\
\nabla r\,(z) &:= \left(\mu^{\mathsf{x}} z^{\mathsf{x}}, \mu^{\mathsf{y}} z^{\mathsf{y}}, \left\{\frac{1}{n}\nabla f_i^*\left(z^{\mathsf{f}_i^*}\right)\right\}_{i\in[n]}, \left\{\frac{1}{n}\nabla g_i^*\left(z^{\mathsf{g}_i^*}\right)\right\}_{i\in[n]}\right),\\
\Phi^h\,(z) &:= \left(\frac{1}{n}\sum_{i\in[n]}\nabla_x h_i(z^{\mathsf{x}}, z^{\mathsf{y}}), -\frac{1}{n}\sum_{i\in[n]}\nabla_y h_i(z^{\mathsf{x}}, z^{\mathsf{y}}), \{0\}_{i\in[n]}, \{0\}_{i\in[n]}\right),\\
\Phi^{\text{bilin}}\,(z) &:= \left(\frac{1}{n}\sum_{i\in[n]} z^{\mathsf{f}_i^*}, \frac{1}{n}\sum_{i\in[n]} z^{\mathsf{g}_i^*}, \left\{-\frac{1}{n}z^{\mathsf{x}}\right\}_{i\in[n]}, \left\{-\frac{1}{n}z^{\mathsf{y}}\right\}_{i\in[n]}\right).
\end{aligned} \tag{41}$$

F.2. Algorithm

In this section we present our algorithm which consists of the following two parts; its design is inspired by a similar strategy used in prior work (Carmon et al., 2019, 2020).

(1) Our "outer loop" is based on a proximal point method (Algorithm 7, adapted from Nemirovski (2004)).

(2) Our "inner loop" solves each proximal subproblem to high accuracy via a careful analysis of randomized mirror prox (Algorithm 8, adapted from Algorithm 3).

At each iteration t of the outer loop (Algorithm 7), we require an accurate approximation

$$z_{t+1} \approx z_{t+1}^\star \text{ which solves the VI in } \Phi := \Phi^{\text{mmfs-pd}}(z) + \gamma\left(\nabla r(z) - \nabla r(z_t)\right), \tag{42}$$

where we recall the definitions of g_{tot} and r from (41) and (40), and when z_t is clear from context (i.e. we are analyzing a single implementation of the inner loop).

To implement our inner loop (i.e. solve the VI in Φ), we apply randomized mirror prox (Algorithm 3) with a new analysis. In particular, we will not be able to obtain the expected relative Lipschitzness bound required by Proposition 19 for our randomized gradient estimators, so we develop a new "partial variance" analysis of Algorithm 3 to obtain our rate. We use this terminology because we use variance bounds on a component of Φ for which we cannot directly obtain expected relative Lipschitzness bounds.

Proposition 29 (Partial variance analysis of randomized mirror prox) *Suppose (possibly random) $\widetilde{\Phi}$ is defined so that in each iteration s, for all $u \in \mathcal{Z}$ and all $\rho > 0$, there exists a (possibly random) point $\bar{w}_s \in \mathcal{Z}$ and a γ-strongly monotone operator $\Phi : \mathcal{Z} \to \mathcal{Z}^*$ (with respect to r) such that*

$$\begin{aligned}
\mathbb{E}\left[\left\langle \widetilde{\Phi}(w_{s+1/2}), w_{s+1/2} - w_\star\right\rangle\right] &= \mathbb{E}\left[\langle \Phi(\bar{w}_s), \bar{w}_s - w_\star\rangle\right],\\
\mathbb{E}\left[\left\langle \widetilde{\Phi}(w_{s+1/2}) - \widetilde{\Phi}(w_s), w_{s+1/2} - w_{s+1}\right\rangle\right] &\le \left(\lambda_0 + \frac{1}{\rho}\right)\mathbb{E}\left[V^r_{w_s}(w_{s+1/2}) + V^r_{w_{s+1/2}}(w_{s+1})\right]\\
&\quad + \rho\lambda_1\mathbb{E}\left[V^r_{w_0}(w_\star) + V^r_{\bar{w}_s}(w_\star)\right],
\end{aligned} \tag{43}$$

where $w_\star$ solves the VI in Φ. Then by setting

$$\rho \leftarrow \frac{\gamma}{5\lambda_1},\ \lambda \leftarrow \lambda_0 + \frac{1}{\rho},\ T \leftarrow \frac{5\lambda}{\gamma} = \frac{5\lambda_0}{\gamma} + \frac{25\lambda_1}{\gamma^2},$$

in Algorithm 3, and returning $\bar{w}_\sigma$ for $0 \le \sigma < S$ sampled uniformly at random,

$$\mathbb{E}\left[V^r_{\bar{w}_\sigma}(w_\star)\right] \le \frac{1}{2} V^r_{w_0}(w_\star).$$

Proof First, consider a single iteration $0 \le s < S$, and fix the point w_s in Algorithm 3. By the optimality conditions on $w_{s+1/2}$ and w_{s+1}, we have

$$\begin{aligned}\frac{1}{\lambda}\left\langle \widetilde{\Phi}(w_s), w_{s+1/2} - w_{s+1}\right\rangle &\le V^r_{w_s}(w_{s+1}) - V^r_{w_{s+1/2}}(w_{s+1}) - V^r_{w_s}(w_{s+1/2}),\\ \frac{1}{\lambda}\left\langle \widetilde{\Phi}(w_{s+1/2}), w_{s+1} - w_\star\right\rangle &\le V^r_{w_s}(w_\star) - V^r_{w_{s+1}}(w_\star) - V^r_{w_s}(w_{s+1}).\end{aligned}$$

Summing the above, rearranging, and taking expectations yields

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\lambda}\left\langle \Phi(\bar{w}_s), \bar{w}_s - w_\star\right\rangle\right] &= \mathbb{E}\left[\frac{1}{\lambda}\left\langle \widetilde{\Phi}(w_{s+1/2}), w_{s+1/2} - w_\star\right\rangle\right]\\ &\le \mathbb{E}\left[V^r_{w_s}(w_\star) - V^r_{w_{s+1}}(w_\star)\right]\\ &\quad + \mathbb{E}\left[\frac{1}{\lambda}\left\langle \widetilde{\Phi}(w_{s+1/2}) - \widetilde{\Phi}(w_s), w_{s+1/2} - w_{s+1}\right\rangle - V^r_{w_s}(w_{s+1/2}) + V^r_{w_{s+1/2}}(w_{s+1})\right]\\ &\le \mathbb{E}\left[V^r_{w_s}(w_\star) - V^r_{w_{s+1}}(w_\star)\right] + \frac{\rho\lambda_1}{\lambda}\mathbb{E}\left[V^r_{w_0}(w_\star) + V^r_{\bar{w}_s}(w_\star)\right].\end{aligned}$$

In the last line we used the assumption (43). Since $w_\star$ solves the VI in Φ, adding $\mathbb{E}\frac{1}{\lambda}\left\langle \Phi(w_\star), w_\star - \bar{w}_s\right\rangle$ to the left-hand side above and applying strong monotonicity of g in r yields

$$\mathbb{E}\left[\frac{1}{\lambda} V^r_{\bar{w}_s}(w_\star)\right] \le \mathbb{E}\left[V^r_{w_s}(w_\star) - V^r_{w_{s+1}}(w_\star)\right] + \frac{\rho\lambda_1}{\lambda}\mathbb{E}\left[V^r_{w_0}(w_\star) + V^r_{\bar{w}_s}(w_\star)\right].$$

Telescoping the above for $0 \le s < S$ and using nonnegativity of Bregman divergences yields

$$(\gamma - \rho\lambda_1)\,\mathbb{E}\left[\frac{1}{T}\sum_{0 \le t < T} V^r_{\bar{w}_s}(w_\star)\right] \le \left(\frac{\lambda}{T} + \rho\lambda_1\right) V^r_{w_0}(w_\star).$$

Substituting our choices of $\bar{w}_s$, ρ, λ, and T yields the claim. ■

For simplicity in the following we denote $\bar{z} := z_t$ whenever we discuss a single proximal subproblem. We next introduce the gradient estimator $\widetilde{\Phi}$ we use in each inner loop, i.e. finding a solution to the VI in Φ defined in (42). We first define three sampling distributions p, q, r, via

$$p_j := \frac{\sqrt{L^{\mathsf{x}}_j}}{2\sum_{i\in[n]}\sqrt{L^{\mathsf{x}}_i}} + \frac{1}{2n} \text{ for all } j \in [n],\ \ q_k := \frac{\sqrt{L^{\mathsf{y}}_k}}{2\sum_{i\in[n]}\sqrt{L^{\mathsf{y}}_i}} + \frac{1}{2n} \text{ for all } k \in [n],$$

$$\text{and } r_\ell := \frac{\Lambda^{\text{tot}}_\ell}{2\sum_{i\in[n]}\Lambda^{\text{tot}}_i} + \frac{1}{2n} \text{ for all } \ell \in [n],\ \text{where } \Lambda^{\text{tot}}_i := \frac{\Lambda^{\mathsf{xx}}_i}{\mu^{\mathsf{x}}} + \frac{\Lambda^{\mathsf{xy}}_i}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}} + \frac{\Lambda^{\mathsf{yy}}_i}{\mu^{\mathsf{y}}} \text{ for all } i \in [n]. \tag{44}$$

Algorithm 8 will run in logarithmically many phases, each initialized at an "anchor point" w_0 (cf. Algorithm 8). We construct gradient estimators for Algorithm 3 of $\Phi(w) = \Phi^{\text{mmfs-pd}}(w) + \gamma(\nabla r(w) - \nabla r(\bar{z}))$ as defined in (42) as follows. In each iteration, for a current anchor point w_0, we sample four coordinates $j \sim p$, $k \sim q$, and $\ell, \ell' \sim r$, all independently. We believe that it is likely that other sampling schemes, e.g. sampling j and k non-independently, will also suffice for our method but focus on the independent scheme for simplicity. We use g^{xy} to refer to the $\mathcal{X} \times \mathcal{Y}$ blocks of a vector g in $\mathcal{Z}^*$, and $^{\mathsf{f}^*\mathsf{g}^*}$ to refer to all other blocks corresponding to $(\mathcal{X}^*)^n \times (\mathcal{Y}^*)^n$. Then we define for an iterate $w = w_s$ of Algorithm 8 (where Φ^h is as in (41)):

$$
\begin{aligned}
\widetilde{\Phi}(w) &:= \Phi_{jk\ell}(w) := \Phi^h_{jk\ell}(w) + \Phi^{\text{sep}}_{jk\ell}(w) + \Phi^{\text{bilin}}_{jk\ell}(w), \\
\left[\Phi^h_{jk\ell}(w)\right]^{\mathsf{x}} &:= \left[\Phi^h(w_0)\right]^{\mathsf{x}} + \frac{1}{nr_\ell}\left(\nabla_x h_\ell(w^{\mathsf{x}}, w^{\mathsf{y}}) - \nabla_x h_\ell(w_0^{\mathsf{x}}, w_0^{\mathsf{y}})\right), \\
\left[\Phi^h_{jk\ell}(w)\right]^{\mathsf{y}} &:= \left[\Phi^h(w_0)\right]^{\mathsf{y}} - \frac{1}{nr_\ell}\left(\nabla_y h_\ell(w^{\mathsf{x}}, w^{\mathsf{y}}) - \nabla_y h_\ell(w_0^{\mathsf{x}}, w_0^{\mathsf{y}})\right), \\
\left[\Phi^h_{jk\ell}(w)\right]^{\mathsf{f}^*\mathsf{g}^*} &:= \left(\{0\}_{i\in[n]}, \{0\}_{i\in[n]}\right), \\
\left[\Phi^{\text{sep}}_{jk\ell}(w)\right]^{\mathsf{xy}} &:= (1+\gamma)\left(\mu^{\mathsf{x}} w^{\mathsf{x}}, \mu^{\mathsf{y}} w^{\mathsf{y}}\right) - \gamma\left(\mu^{\mathsf{x}} \bar{z}^{\mathsf{x}}, \mu^{\mathsf{y}} \bar{z}^{\mathsf{y}}\right), \\
\left[\Phi^{\text{sep}}_{jk\ell}(w)\right]^{\mathsf{f}^*\mathsf{g}^*} &:= (1+\gamma)\left(\left\{\frac{1}{np_j}\nabla f_j^*\left(w^{\mathsf{f}_\mathsf{j}^*}\right)\cdot \mathbf{1}_{i=j}\right\}_{i\in[n]}, \left\{\frac{1}{nq_k}\nabla g_k^*\left(w^{\mathsf{f}_\mathsf{k}^*}\right)\cdot \mathbf{1}_{i=k}\right\}_{i\in[n]}\right) \\
&\quad - \gamma\left(\left\{\frac{1}{np_j}\nabla f_j^*\left(\bar{z}^{\mathsf{f}_\mathsf{j}^*}\right)\cdot \mathbf{1}_{i=j}\right\}_{i\in[n]}, \left\{\frac{1}{nq_k}\nabla g_k^*\left(\bar{z}^{\mathsf{g}_\mathsf{k}^*}\right)\cdot \mathbf{1}_{i=k}\right\}_{i\in[n]}\right), \\
\left[\Phi^{\text{bilin}}_{jk\ell}(w)\right]^{\mathsf{xy}} &:= \left(\frac{1}{n}\sum_{i\in[n]} w^{\mathsf{f}_\mathsf{i}^*}, \frac{1}{n}\sum_{i\in[n]} w^{\mathsf{g}_\mathsf{i}^*}\right), \\
\left[\Phi^{\text{bilin}}_{jk\ell}(w)\right]^{\mathsf{f}^*\mathsf{g}^*} &:= \left(\left\{-\frac{1}{np_j} w^{\mathsf{x}}\cdot \mathbf{1}_{i=j}\right\}_{i\in[n]}, \left\{-\frac{1}{nq_k} w^{\mathsf{y}}\cdot \mathbf{1}_{i=k}\right\}_{i\in[n]}\right).
\end{aligned}
\tag{45}
$$

In particular, the estimator $\Phi_{jk\ell}(w)$ only depends on the sampled indices j, k, ℓ, and not ℓ'. Next, consider taking the step $w_{\mathsf{aux}}(jk\ell) \leftarrow \text{Prox}^r_w(\frac{1}{\lambda} g_{jk\ell}(w))$ as in Algorithm 3, where we use the shorthand $w_{\mathsf{aux}}(jk\ell) = w_{s+1/2}$ to indicate the iterate of Algorithm 3 taken from w_s assuming j, k, ℓ were sampled. Observing the form of $g_{jk\ell}$, we denote the blocks of $w_{\mathsf{aux}}(jk\ell)$ by

$$
w_{\mathsf{aux}}(jk\ell) := \left(w^{\mathsf{x}}_{\mathsf{aux}}(\ell), w^{\mathsf{y}}_{\mathsf{aux}}(\ell), \left\{w^{\mathsf{f}_\mathsf{i}^*}_{\mathsf{aux}}(j)\right\}_{i\in[n]}, \left\{w^{\mathsf{g}_\mathsf{i}^*}_{\mathsf{aux}}(k)\right\}_{i\in[n]}\right),
$$

where we write $w^{\mathsf{x}}_{\mathsf{aux}}(\ell)$ to indicate that it only depends on the random choice of ℓ (and not j or k); we use similar notation for the other blocks. We also define

$$
\Delta^{\mathsf{x}}(j) := w^{\mathsf{f}_\mathsf{j}^*}_{\mathsf{aux}}(j) - w^{\mathsf{f}_\mathsf{j}^*}(j),\ \Delta^{\mathsf{y}}(k) := w^{\mathsf{g}_\mathsf{k}^*}_{\mathsf{aux}}(k) - w^{\mathsf{g}_\mathsf{k}^*}(k),
$$

and then set (where we use the notation $\Phi_{jk\ell'}$ to signify its dependence on j, k, ℓ', and not ℓ):

$$\begin{aligned}
\widetilde{\Phi}(w_{\mathsf{aux}}(jk\ell)) &:= \Phi_{jk\ell'}(w_{\mathsf{aux}}(jk\ell)) := \Phi^{h}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell)) + \Phi^{\mathsf{sep}}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell)) + \Phi^{\mathsf{bilin}}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell)),\\
\left[\Phi^{h}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell))\right]^{\mathsf{x}} &:= \left[\Phi^{h}(w_0)\right]^{\mathsf{x}} + \frac{1}{nr_{\ell'}}\left(\nabla_x h_{\ell'}(w^{\mathsf{x}}_{\mathsf{aux}}(\ell), w^{\mathsf{y}}_{\mathsf{aux}}(\ell)) - \nabla_x h_{\ell'}(w^{\mathsf{x}}_0, w^{\mathsf{y}}_0)\right),\\
\left[\Phi^{h}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell))\right]^{\mathsf{y}} &:= \left[\Phi^{h}(w_0)\right]^{\mathsf{y}} - \frac{1}{nr_{\ell'}}\left(\nabla_y h_{\ell'}(w^{\mathsf{x}}_{\mathsf{aux}}(\ell), w^{\mathsf{y}}_{\mathsf{aux}}(\ell)) - \nabla_y h_{\ell'}(w^{\mathsf{x}}_0, w^{\mathsf{y}}_0)\right),\\
\left[\Phi^{h}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell))\right]^{\mathsf{f}^*\mathsf{g}^*} &:= \left(\{0\}_{i\in[n]}, \{0\}_{i\in[n]}\right),\\
\left[\Phi^{\mathsf{sep}}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell))\right]^{\mathsf{xy}} &:= (1+\gamma)\left(\mu^{\mathsf{x}} w^{\mathsf{x}}_{\mathsf{aux}}(\ell), \mu^{\mathsf{y}} w^{\mathsf{y}}_{\mathsf{aux}}(\ell)\right) - \gamma\left(\mu^{\mathsf{x}}\bar{z}^{\mathsf{x}}, \mu^{\mathsf{y}}\bar{z}^{\mathsf{y}}\right),\\
\left[\Phi^{\mathsf{sep}}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell))\right]^{\mathsf{f}^*\mathsf{g}^*} &:= (1+\gamma)\left(\left\{\frac{1}{np_j}\nabla f^*_j\left(w^{\mathsf{f}^*_\mathsf{j}}_{\mathsf{aux}}\right)\cdot \mathbf{1}_{i=j}\right\}_{i\in[n]}, \left\{\frac{1}{nq_k}\nabla g^*_k\left(w^{\mathsf{f}^*_\mathsf{k}}_{\mathsf{aux}}\right)\cdot \mathbf{1}_{i=k}\right\}_{i\in[n]}\right)\\
&\quad - \gamma\left(\left\{\frac{1}{np_j}\nabla f^*_j\left(\bar{z}^{\mathsf{f}^*_\mathsf{j}}\right)\cdot \mathbf{1}_{i=j}\right\}_{i\in[n]}, \left\{\frac{1}{nq_k}\nabla g^*_k\left(\bar{z}^{\mathsf{g}^*_\mathsf{k}}\right)\cdot \mathbf{1}_{i=k}\right\}_{i\in[n]}\right),\\
\left[\Phi^{\mathsf{bilin}}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell))\right]^{\mathsf{xy}} &:= \left(\frac{1}{n}\sum_{i\in[n]} w^{\mathsf{f}^*_\mathsf{i}} + \frac{1}{np_j}\Delta^{\mathsf{x}}(j), \frac{1}{n}\sum_{i\in[n]} w^{\mathsf{g}^*_\mathsf{i}} + \frac{1}{nq_k}\Delta^{\mathsf{y}}(k)\right),\\
\left[\Phi^{\mathsf{bilin}}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell))\right]^{\mathsf{f}^*\mathsf{g}^*} &:= \left(\left\{-\frac{1}{np_j} w^{\mathsf{x}}_{\mathsf{aux}}(\ell)\cdot \mathbf{1}_{i=j}\right\}_{i\in[n]}, \left\{-\frac{1}{nq_k} w^{\mathsf{y}}_{\mathsf{aux}}(\ell)\cdot \mathbf{1}_{i=k}\right\}_{i\in[n]}\right).
\end{aligned} \tag{46}$$

We also define the random "aggregate point" we will use in Proposition 29:

$$\bar{w}(\ell) := w + \left(w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}, w^{\mathsf{y}}_{\mathsf{aux}}(\ell) - w^{\mathsf{y}}, \{\Delta^{\mathsf{x}}(j)\}_{j\in[n]}, \{\Delta^{\mathsf{y}}(k)\}_{k\in[n]}\right). \tag{47}$$

Notably, $\bar{w}(\ell)$ depends only on the randomly sampled ℓ. We record the following useful observation about our randomized operators (45), (46), in accordance with the first condition in (43).

Lemma 30 *Define $\{\Phi_{jk\ell}, \Phi_{jk\ell'}\} : \mathcal{Z} \to \mathcal{Z}^*$ as in (45), (46), and the random "aggregate point" $\bar{w}(\ell)$ as in (47). Then, for all $u \in \mathcal{Z}$, recalling the definition of $\Phi = \Phi^{\mathsf{mmfs\text{-}pd}} + \gamma(\nabla r - \nabla r(\bar{z}))$ from (42),*

$$\mathbb{E}\left[\left\langle \Phi_{jk\ell'}(w_{\mathsf{aux}}(jk\ell)), w_{\mathsf{aux}}(jk\ell) - u\right\rangle\right] = \mathbb{E}_{\ell\sim r}\left[\langle \Phi(\bar{w}(\ell)), \bar{w}(\ell) - u\rangle\right].$$

Proof We demonstrate this equality for the $\mathcal{X}$ and $(\mathcal{X}^*)^n$ blocks; the others (the $\mathcal{Y}$ and $(\mathcal{Y}^*)^n$ blocks) follow symmetrically. We will use the definitions of Φ^h and Φ^{bilin} from (41).

$\mathcal{X}$ block. Fix $\ell \in [n]$. We first observe that

$$\begin{aligned}
\mathbb{E}_{\ell'\sim r}\left[\left[\Phi^{h}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell))\right]^{\mathsf{x}}\right] &= \left[\Phi^{h}(\bar{w}(\ell))\right]^{\mathsf{x}},\\
\mathbb{E}_{\ell'\sim r}\left[\left[\Phi^{\mathsf{sep}}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell))\right]^{\mathsf{x}}\right] &= (1+\gamma)\left[\nabla r(\bar{w}(\ell))\right]^{\mathsf{x}} - \gamma\left[\nabla r(\bar{z})\right]^{\mathsf{x}}.
\end{aligned}$$

Moreover, by expanding the expectation over $j \sim p$,

$$\begin{aligned}\mathbb{E}_{j\sim p}\left[\left\langle\left[\Phi^{\mathsf{bilin}}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell))\right]^{\mathsf{x}}, w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - u^{\mathsf{x}}\right\rangle\right] &= \left\langle \frac{1}{n}\sum_{j\in[n]}(w^{\mathsf{f}^*_{\mathsf{j}}} + \Delta^{\mathsf{x}}(j)), w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - u^{\mathsf{x}}\right\rangle \\ &= \left\langle\left[\Phi^{\mathsf{bilin}}(\bar{w}(\ell))\right]^{\mathsf{x}}, w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - u^{\mathsf{x}}\right\rangle.\end{aligned}$$

Summing, we conclude that for fixed ℓ and taking expectations over j, k, ℓ',

$$\mathbb{E}\left[\left\langle\left[\Phi_{jk\ell'}(w_{\mathsf{aux}}(jk\ell))\right]^{\mathsf{x}}, w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - u^{\mathsf{x}}\right\rangle\right] = \left\langle[\Phi(\bar{w}(\ell))]^{\mathsf{x}}, w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - u^{\mathsf{x}}\right\rangle.$$

The conclusion for the $\mathcal{X}$ block follows by taking expectations over ℓ.

$\mathcal{X}^*$ **blocks.** Note that the $[\Phi^h_{jk\ell'}]^{\mathsf{f}^*}$ blocks are always zero. Next, for the $[\Phi^{\mathsf{sep}}_{jk\ell'}]^{\mathsf{f}^*}$ component, by expanding the expectation over $j \sim p$ and taking advantage of sparsity, for any $\ell \in [n]$,

$$\begin{aligned}&\mathbb{E}_{j\sim p}\left[\left\langle\left[\Phi^{\mathsf{sep}}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell))\right]^{\mathsf{f}^*}, w^{\mathsf{f}^*}_{\mathsf{aux}}(jk\ell) - u^{\mathsf{f}^*}\right\rangle\right] \\ &= (1+\gamma)\sum_{j\in[n]}\left\langle\frac{1}{n}\nabla f^*_j\left(w^{\mathsf{f}^*_{\mathsf{j}}}_{\mathsf{aux}}\right), w^{\mathsf{f}^*_{\mathsf{j}}}_{\mathsf{aux}} - u^{\mathsf{f}^*_{\mathsf{j}}}\right\rangle - \gamma\sum_{j\in[n]}\left\langle\frac{1}{n}\nabla f^*_j\left(\bar{z}^{\mathsf{f}^*_{\mathsf{j}}}\right), w^{\mathsf{f}^*_{\mathsf{j}}}_{\mathsf{aux}} - u^{\mathsf{a}_{\mathsf{j}}}\right\rangle \\ &= \left\langle(1+\gamma)\left[\nabla r(\bar{w}(\ell))\right]^{\mathsf{f}^*} - \gamma\left[\nabla r(\bar{z})\right]^{\mathsf{f}^*}, \bar{w}^{\mathsf{f}^*}(\ell) - u^{\mathsf{f}^*}\right\rangle.\end{aligned}$$

Here, we recall $\mathsf{f}^*_{\mathsf{j}}$ denotes the block corresponding to the j^{th} copy of $\mathcal{X}^*$. Finally, for the $[\Phi^{\mathsf{bilin}}_{jk\ell'}]^{\mathsf{f}^*}$ component, fix $\ell \in [n]$. Expanding the expectation over $j \sim p$ and taking advantage of sparsity,

$$\mathbb{E}_{j\sim p}\left[\left\langle\left[\Phi^{\mathsf{bilin}}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell))\right]^{\mathsf{f}^*}, \left[w_{\mathsf{aux}}(jk\ell)\right]^{\mathsf{f}^*} - u^{\mathsf{f}^*}\right\rangle\right] = \left\langle\left[\Phi^{\mathsf{bilin}}(\bar{w}(\ell))\right]^{\mathsf{f}^*}, \bar{w}^{\mathsf{f}^*}(\ell) - u^{\mathsf{f}^*}\right\rangle.$$

Summing, we conclude that for fixed ℓ and taking expectations over j, k, ℓ',

$$\mathbb{E}\left\langle\left[g_{jk\ell'}(w_{\mathsf{aux}}(jk\ell))\right]^{\mathsf{f}^*}, \left[w_{\mathsf{aux}}(jk\ell)\right]^{\mathsf{f}^*} - u^{\mathsf{f}^*}\right\rangle = \left\langle\left[g_{\mathsf{tot}}(\bar{w}(\ell))\right]^{\mathsf{f}^*}, \bar{w}^{\mathsf{f}^*}(\ell) - u^{\mathsf{f}^*}\right\rangle.$$

The conclusion for the $\mathcal{X}^*$ blocks follows by taking expectations over ℓ. ■

Finally, we give a complete implementation of our method as pseudocode below in Algorithms 7 (the outer loop) and 8 (the inner loop). We also show that it is a correct implementation in the following Lemma 31.

Lemma 31 *The inner for loop of Algorithm 8 implements Algorithm 3 on $(\{\widetilde{\Phi}\}, r)$ defined in (45), (46), (40), for σ iterations, and returns $\bar{w}_\sigma$, following the definition (47). Each iteration $s > 0$ is implementable in $O(1)$ gradient calls to some $\{f_j, g_k, h_l\}$, and $O(1)$ vector operations on $\mathcal{X}$ and $\mathcal{Y}$.*

Proof Let $\{w_s, w_{s+1/2}\}_{0\le s\le\sigma}$ be the iterates of Algorithm 3. We will inductively show that some run of the inner for loop in Algorithm 8 preserves the invariants

$$\begin{aligned}w_s &= \left(w^{\mathsf{x}}_s, w^{\mathsf{y}}_s, \left\{\nabla f_i(w^{\mathsf{f}_{\mathsf{i}}}_s)\right\}_{i\in[n]}, \left\{\nabla f_i(w^{\mathsf{g}_{\mathsf{i}}}_s)\right\}_{i\in[n]}\right), \\ w_{s+1/2} &= \left(w^{\mathsf{x}}_{s+1/2}, w^{\mathsf{y}}_{s+1/2}, \left\{\nabla f_i(w^{\mathsf{f}_{\mathsf{i}}}_{s+1/2})\right\}_{i\in[n]}, \left\{\nabla f_i(w^{\mathsf{g}_{\mathsf{i}}}_{s+1/2})\right\}_{i\in[n]}\right)\end{aligned}$$

```
Algorithm 7: MINIMAX-FINITESUM-SOLVE($F_{\text{mmfs-reg}}, x_0, y_0$): Minimax finite sum optimization
Input: (37) satisfying Assumption 3, $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$
Parameter(s): $T \in \mathbb{N}$
$z_0^{\mathsf{x}} \leftarrow x_0$, $z_0^{\mathsf{y}} \leftarrow y_0$, $z_0^{\mathsf{f_i}} \leftarrow x_0$, $z_0^{\mathsf{f_i^*}} \leftarrow \nabla f_i(x_0)$, $z_0^{\mathsf{g_i}} \leftarrow y_0$, $z_0^{\mathsf{g_i^*}} \leftarrow \nabla g_i(y_0)$ for all $i \in [n]$
for $0 \le t < T$ do
    $z_{t+1} \leftarrow$ MINIMAX-FINITESUM-INNER($F_{\text{mmfs-reg}}, \{z_t^{\mathsf{x}}, z_t^{\mathsf{y}}, \{z_t^{\mathsf{f_i}}\}_{i\in[n]}, \{z_t^{\mathsf{g_i}}\}_{i\in[n]}\}$)
end
Return: $(z_T^{\mathsf{x}}, z_T^{\mathsf{y}})$
```

for all $0 \le s \le \sigma$. Once we prove this claim, it is clear that the inner for loop in Algorithm 8 implements Algorithm 3 and returns $\bar{w}_\sigma$, upon recalling the definitions (45), (46), (40), and (47).

The base case of our induction follows from the way w_0 is initialized. Next, suppose for some $0 \le s \le \sigma$, our inductive claim holds. By the update for $w_{s+1/2}^{\mathsf{f_j^*}}$, if $j \in [n]$ was sampled in iteration s, using the first item in Fact 1,

$$\begin{aligned} w_{s+1/2}^{\mathsf{f_j^*}} &\leftarrow \text{argmin}_{w^{\mathsf{f_j^*}} \in \mathcal{X}^*} \left\{ \left\langle \frac{1}{n\lambda p_j} \left((1+\gamma) w_s^{\mathsf{f_j}} - \gamma \bar{z}^{\mathsf{f_j}} - w_s^{\mathsf{x}} \right), w^{\mathsf{f_j^*}} \right\rangle + V_{w_s^{\mathsf{f_j^*}}}^{f_j^*} \left(w^{\mathsf{f_j^*}} \right) \right\} \\ &= \nabla f_j \left(w_s^{\mathsf{f_j}} - \frac{1}{n\lambda p_j} \left((1+\gamma)\, w_s^{\mathsf{f_j}} - \gamma \bar{z}^{\mathsf{f_j}} - w_s^{\mathsf{x}} \right) \right). \end{aligned}$$

Similarly, if $k \in [n]$ was sampled in iteration s,

$$\begin{aligned} w_{s+1/2}^{\mathsf{g_k^*}} &\leftarrow \text{argmin}_{w^{\mathsf{g_k^*}}} \left\langle \frac{1}{n\lambda q_k} \left((1+\gamma) w_s^{\mathsf{g_k}} - \gamma \bar{z}^{\mathsf{g_k}} - w_s^{\mathsf{y}} \right), w^{\mathsf{g_k^*}} \right\rangle + V_{w_s^{\mathsf{g_k^*}}}^{g_k^*} \left(w^{\mathsf{g_k^*}} \right). \\ &= \nabla g_k \left(w_s^{\mathsf{g_k}} - \frac{1}{n\lambda q_k} \left((1+\gamma)\, w_s^{\mathsf{g_k}} - \gamma \bar{z}^{\mathsf{g_k}} - w_s^{\mathsf{y}} \right) \right). \end{aligned}$$

Hence, the updates to $w_{s+1/2}^{\mathsf{f_j^*}}$ and $w_{s+1/2}^{\mathsf{g_k^*}}$ preserve our invariant, and all other $w_{s+1/2}^{\mathsf{f_i^*}}$, $i \neq j$ and $w_{s+1/2}^{\mathsf{g_i^*}}$, $i \neq k$ do not change by sparsity of $\Phi_{jk\ell}$. Analogously the updates to each $w_{s+1}^{\mathsf{f_i^*}}$ and $w_{s+1}^{\mathsf{g_i^*}}$ preserve our invariant. Finally, in every iteration $s > 0$, the updates to $w_{s+1/2}^{\mathsf{xy}}$ and w_{s+1}^{xy} only require evaluating $O(1)$ new gradients each, by 1-sparsity of the dual block updates. ■

F.3. Inner loop convergence analysis

We give a convergence guarantee on Algorithm 8 for solving the VI in $\Phi := g_{\text{tot}} + \gamma(\nabla r - \nabla r(\bar{z}))$. In order to use Proposition 29 to solve our problem, we must prove strong monotonicity of Φ and specify the parameters λ_0, λ_1 and ρ in (43); note that Lemma 30 handles the first condition in (43). To this end we give the following properties on Φ, $\widetilde{\Phi}$ as defined in (45) and (46).

Strong monotonicity. We begin by proving strong monotonicity of Φ.

Lemma 32 (Strong monotonicity) *Define $\Phi : \mathcal{Z} \to \mathcal{Z}^*$ as in (42), and define $r : \mathcal{Z} \to \mathbb{R}$ as in (40). Then Φ is $(1+\gamma)$-strongly monotone with respect to r.*

Algorithm 8: MINIMAX-FINITESUM-INNER$(F_{\text{mmfs-reg}}, \bar{z}^{\mathsf{x}}, \bar{z}^{\mathsf{y}}, \{\bar{z}^{\mathsf{f_i}}\}_{i\in[n]}, \{\bar{z}^{\mathsf{g_i}}\}_{i\in[n]})$: Minimax finite sum optimization subroutine

```
Input: (37) satisfying Assumption 3, $\bar{z}^{\mathsf{x}}, \{\bar{z}^{\mathsf{f_i}}\}_{i\in[n]} \in \mathcal{X}$, $\bar{z}^{\mathsf{y}}, \{\bar{z}^{\mathsf{g_i}}\}_{i\in[n]} \in \mathcal{Y}$
Parameter(s): $\gamma \geq 1$, $\lambda > 0$, $N, S \in \mathbb{N}$
$w_0 \leftarrow \bar{z}$
for $0 \leq \tau < N$ do
    Sample $0 \leq \sigma < S$ uniformly at random
    for $0 \leq s \leq \sigma$ do
        Sample $j, k, \ell, \ell' \in [n]$ independently according to $p, q, r, r$ respectively defined in (44), and define
            $[\Phi^{\mathsf{sep}}]^{\mathsf{x}} := (1+\gamma)\,\mu^{\mathsf{x}} w_s^{\mathsf{x}} - \gamma\mu^{\mathsf{x}}\bar{z}^{\mathsf{x}}, \quad [\Phi^{\mathsf{sep}}]^{\mathsf{y}} = (1+\gamma)\,\mu^{\mathsf{y}} w_s^{\mathsf{y}} - \gamma\mu^{\mathsf{y}}\bar{z}^{\mathsf{y}},$
            $\left[\Phi^{\mathsf{bilin}}\right]^{\mathsf{x}} := \frac{\sum_{i\in[n]} \nabla f_i\left(w_s^{\mathsf{f_i}}\right)}{n}, \quad \left[\Phi^{\mathsf{bilin}}\right]^{\mathsf{y}} := \frac{\sum_{i\in[n]} \nabla g_i\left(w_s^{\mathsf{g_i}}\right)}{n},$
            $\Phi^{\mathsf{x}} := \left[\Phi^h(w_0)\right]^{\mathsf{x}} + \frac{\nabla_x h_\ell(w_s^{\mathsf{x}}, w_s^{\mathsf{y}}) - \nabla_x h_\ell(w_0^{\mathsf{x}}, w_0^{\mathsf{y}})}{n r_\ell} + [\Phi^{\mathsf{sep}}]^{\mathsf{x}} + \left[\Phi^{\mathsf{bilin}}\right]^{\mathsf{x}},$
            $\Phi^{\mathsf{y}} := \left[\Phi^h(w_0)\right]^{\mathsf{y}} - \frac{\nabla_y h_\ell(w_s^{\mathsf{x}}, w_s^{\mathsf{y}}) - \nabla_y h_\ell(w_0^{\mathsf{x}}, w_0^{\mathsf{y}})}{n r_\ell} + [\Phi^{\mathsf{sep}}]^{\mathsf{y}} + \left[\Phi^{\mathsf{bilin}}\right]^{\mathsf{y}}$
        $w_{s+1/2}^{\mathsf{x}} \leftarrow w_s^{\mathsf{x}} - \frac{1}{\lambda\mu^{\mathsf{x}}}\Phi^{\mathsf{x}}$, $w_{s+1/2}^{\mathsf{y}} \leftarrow w_s^{\mathsf{y}} - \frac{1}{\lambda\mu^{\mathsf{y}}}\Phi^{\mathsf{y}}$
        $w_{s+1/2}^{\mathsf{f_j}} \leftarrow w_s^{\mathsf{f_j}} - \frac{1}{n\lambda p_j}\left((1+\gamma)\,w_s^{\mathsf{f_j}} - \gamma\bar{z}^{\mathsf{f_j}} - w_s^{\mathsf{x}}\right)$
        $w_{s+1/2}^{\mathsf{g_k}} \leftarrow w_s^{\mathsf{f_k}} - \frac{1}{n\lambda p_k}\left((1+\gamma)\,w_s^{\mathsf{g_k}} - \gamma\bar{z}^{\mathsf{g_k}} - w_s^{\mathsf{y}}\right)$
        Define
            $[\Phi^{\mathsf{sep}}]^{\mathsf{x}} := (1+\gamma)\,\mu^{\mathsf{x}} w_{s+1/2}^{\mathsf{x}} - \gamma\mu^{\mathsf{x}}\bar{z}^{\mathsf{x}}, \quad [\Phi^{\mathsf{sep}}]^{\mathsf{y}} := (1+\gamma)\,\mu^{\mathsf{y}} w_{s+1/2}^{\mathsf{y}} - \gamma\mu^{\mathsf{y}}\bar{z}^{\mathsf{y}},$
            $\left[\Phi^{\mathsf{bilin}}\right]^{\mathsf{x}} := \frac{\sum_{i\in[n]} \nabla f_i\left(w_s^{\mathsf{f_i}}\right)}{n} + \frac{\nabla f_j\left(w_{s+1/2}^{\mathsf{f_j}}\right) - \nabla f_j\left(w_s^{\mathsf{f_j}}\right)}{n p_j},$
            $\left[\Phi^{\mathsf{bilin}}\right]^{\mathsf{y}} := \frac{\sum_{i\in[n]} \nabla g_i\left(w_s^{\mathsf{g_i}}\right)}{n} + \frac{\nabla g_k\left(w_{s+1/2}^{\mathsf{g_k}}\right) - \nabla g_k\left(w_s^{\mathsf{g_k}}\right)}{n q_k},$
            $\Phi^{\mathsf{x}} := \left[\Phi^h(w_0)\right]^{\mathsf{x}} + \frac{\nabla_x h_{\ell'}(w_{s+1/2}^{\mathsf{x}}, w_{s+1/2}^{\mathsf{y}}) - \nabla_x h_{\ell'}(w_0^{\mathsf{x}}, w_0^{\mathsf{y}})}{n r_{\ell'}} + [\Phi^{\mathsf{sep}}]^{\mathsf{x}} + \left[\Phi^{\mathsf{bilin}}\right]^{\mathsf{x}},$
            $\Phi^{\mathsf{y}} := \left[\Phi^h(w_0)\right]^{\mathsf{y}} - \frac{\nabla_y h_{\ell'}(w_{s+1/2}^{\mathsf{x}}, w_{s+1/2}^{\mathsf{y}}) - \nabla_y h_{\ell'}(w_0^{\mathsf{x}}, w_0^{\mathsf{y}})}{n r_{\ell'}} + [\Phi^{\mathsf{sep}}]^{\mathsf{y}} + \left[\Phi^{\mathsf{bilin}}\right]^{\mathsf{y}}$
        $w_{s+1}^{\mathsf{x}} \leftarrow w_s^{\mathsf{x}} - \frac{1}{\lambda\mu^{\mathsf{x}}}\Phi^{\mathsf{x}}$, $w_{s+1}^{\mathsf{y}} \leftarrow w_s^{\mathsf{y}} - \frac{1}{\lambda\mu^{\mathsf{y}}}\Phi^{\mathsf{y}}$
        $w_{s+1}^{\mathsf{f_j}} \leftarrow w_s^{\mathsf{f_j}} - \frac{1}{n\lambda p_j}\left((1+\gamma)\,w_{s+1/2}^{\mathsf{f_j}} - \gamma\bar{z}^{\mathsf{f_j}} - w_{s+1/2}^{\mathsf{x}}\right)$
        $w_{s+1}^{\mathsf{g_k}} \leftarrow w_s^{\mathsf{f_k}} - \frac{1}{n\lambda p_k}\left((1+\gamma)\,w_{s+1/2}^{\mathsf{g_k}} - \gamma\bar{z}^{\mathsf{g_k}} - w_{s+1/2}^{\mathsf{y}}\right)$
    end
    $\bar{w}^{\mathsf{f_i}} \leftarrow w_\sigma^{\mathsf{f_i}} - \frac{1}{n\lambda p_i}\left((1+\gamma)\,w_\sigma^{\mathsf{f_i}} - \gamma\bar{z}^{\mathsf{f_i}} - w_\sigma^{\mathsf{x}}\right)$ for each $i \in [n]$
    $\bar{w}^{\mathsf{g_i}} \leftarrow w_\sigma^{\mathsf{g_i}} - \frac{1}{n\lambda q_i}\left((1+\gamma)\,w_\sigma^{\mathsf{g_i}} - \gamma\bar{z}^{\mathsf{g_i}} - w_\sigma^{\mathsf{y}}\right)$ for each $i \in [n]$
    $w_0^{\mathsf{xy}} \leftarrow w_{\sigma+1/2}^{\mathsf{xy}}$, $w_0^{\mathsf{f_i}} \leftarrow \bar{w}^{\mathsf{f_i}}$, $w_0^{\mathsf{g_i}} \leftarrow \bar{w}^{\mathsf{g_i}}$ for all $i \in [n]$
end
Return: $(w_0^{\mathsf{x}}, w_0^{\mathsf{y}}, \{\nabla f_i(w_0^{\mathsf{f_i}})\}_{i\in[n]}, \{\nabla g_i(w_0^{\mathsf{g_i}})\}_{i\in[n]})$
```

Proof We decompose $\Phi(z) = (1+\gamma)\nabla r(z) + \Phi^{\text{bilin}}(z) + \Phi^h(z) - \gamma\nabla r(\bar{z})$, using the definitions in (41). By a similar argument as Lemma 5, we obtain the claim. ■

Expected relative Lipschitzness. We next provide bounds on the components of (43) corresponding to Φ^{sep} and Φ^{bilin}, where we use the shorthand $\Phi^{\text{sep}} := (1+\gamma)\nabla r - \gamma\nabla r(\bar{z})$ in the remainder of this section. In particular, we provide a partial bound on the quantity λ_0.

Lemma 33 *Define $\{\Phi_{jk\ell}, \Phi_{jk\ell'}\} : \mathcal{Z} \to \mathcal{Z}^*$ as in (45), (46), and define $r : \mathcal{Z} \to \mathbb{R}$ as in (40). Letting $w_+(jk\ell\ell')$ be w_{s+1} in Algorithm 8 if j, k, ℓ, ℓ' were sampled in iteration s, defining*

$$\Phi^{fg}_{jk\ell}(w) := \Phi^{\text{sep}}_{jk\ell}(w) + \Phi^{\text{bilin}}_{jk\ell}(w),$$
$$\Phi^{fg}_{jk\ell'}(w_{\text{aux}}(jk\ell)) := \Phi^{\text{sep}}_{jk\ell'}(w_{\text{aux}}(jk\ell)) + \Phi^{\text{bilin}}_{jk\ell'}(w_{\text{aux}}(jk\ell)),$$

we have

$$\mathbb{E}\left[\left\langle \Phi^{fg}_{jk\ell'}(w_{\text{aux}}(jk\ell)) - \Phi^{fg}_{jk\ell}(w), w_{\text{aux}}(jk\ell) - w_+(jk\ell\ell')\right\rangle\right]$$
$$\le \lambda^{fg}\mathbb{E}\left[V^r_w\left(w_{\text{aux}}(jk\ell)\right) + V^r_{w_{\text{aux}}(jk\ell)}\left(w_+(jk\ell\ell')\right)\right],$$

for

$$\lambda^{fg} = 2n(1+\gamma) + \frac{\sum_{i\in[n]}\sqrt{L^{\mathsf{x}}_i}}{\sqrt{n\mu^{\mathsf{x}}}} + \frac{\sum_{i\in[n]}\sqrt{L^{\mathsf{y}}_i}}{\sqrt{n\mu^{\mathsf{y}}}}.$$

Proof This is immediate upon combining the following Lemmas 34 and 35. ■

Lemma 34 *Following notation of Lemma 33, for $\lambda^{\text{sep}} := 2n(1+\gamma)$, we have*

$$\mathbb{E}\left[\left\langle \Phi^{\text{sep}}_{jk\ell'}(w_{\text{aux}}(jk\ell)) - \Phi^{\text{sep}}_{jk\ell}(w), w_{\text{aux}}(jk\ell) - w_+(jk\ell\ell')\right\rangle\right]$$
$$\le \lambda^{\text{sep}}\mathbb{E}\left[V^r_w\left(w_{\text{aux}}(jk\ell)\right) + V^r_{w_{\text{aux}}(jk\ell)}\left(w_+(jk\ell\ell')\right)\right].$$

Proof The proof is similar to (part of) the proof of Lemma 22. We claim that for any j, k, ℓ, ℓ',

$$\left\langle \Phi^{\text{sep}}_{jk\ell'}(w_{\text{aux}}(jk\ell)) - \Phi^{\text{sep}}_{jk\ell}(w), w_{\text{aux}}(jk\ell) - w_+(jk\ell\ell')\right\rangle$$
$$\le \lambda^{\text{sep}}\left(V^r_w\left(w_{\text{aux}}(jk\ell)\right) + V^r_{w_{\text{aux}}(jk\ell)}\left(w_+(jk\ell\ell')\right)\right).$$

Fix j, k, ℓ, ℓ'. Since all p_j and q_k are lower bounded by $\frac{1}{2n}$ by assumption, applying Lemma 7 to the relevant blocks of r and nonnegativity of Bregman divergences proves the above display. ■

Lemma 35 *Following notation of Lemma 33, for*

$$\lambda^{\text{cross}} := \frac{2\sum_{i\in[n]}\sqrt{L^{\mathsf{x}}_i}}{\sqrt{n\mu^{\mathsf{x}}}} + \frac{2\sum_{i\in[n]}\sqrt{L^{\mathsf{y}}_i}}{\sqrt{n\mu^{\mathsf{y}}}},$$

we have

$$\mathbb{E}\left[\left\langle \Phi^{\text{bilin}}_{jk\ell'}(w_{\text{aux}}(jk\ell)) - \Phi^{\text{bilin}}_{jk\ell}(w), w_{\text{aux}}(jk\ell) - w_+(jk\ell\ell')\right\rangle\right]$$
$$\le \lambda^{\text{cross}}\mathbb{E}\left[V^r_w\left(w_{\text{aux}}(jk\ell)\right) + V^r_{w_{\text{aux}}(jk\ell)}\left(w_+(jk\ell\ell')\right)\right].$$

Proof The proof is similar to (part of) the proof of Lemma 22. We claim that for any j, k, ℓ, ℓ',

$$\begin{aligned}&\left\langle \Phi^{\text{bilin}}_{jk\ell'}(w_{\mathsf{aux}}(jk\ell)) - \Phi^{\text{bilin}}_{jk\ell}(w), w_{\mathsf{aux}}(jk\ell) - w_+(jk\ell\ell')\right\rangle \\ &\qquad\le \lambda^{\text{cross}} \left(V^r_w\left(w_{\mathsf{aux}}(jk\ell)\right) + V^r_{w_{\mathsf{aux}}(jk\ell)}\left(w_+(jk\ell\ell')\right)\right).\end{aligned}$$

Fix j, k, ℓ, ℓ'. By applying Item (1) in Lemma 16 with $f = f_j$, $\alpha = (L^{\mathsf{x}}_j \mu^{\mathsf{x}})^{-\frac{1}{2}}$,

$$\begin{aligned}&\mathbb{E}_j\left[\frac{1}{np_j}\left\langle w^{\mathsf{f}_j^*}_{\mathsf{aux}} - w^{\mathsf{f}_j^*}, w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_+(jk\ell\ell')\right\rangle + \frac{1}{np_j}\left\langle w^{\mathsf{x}} - w^{\mathsf{x}}_{\mathsf{aux}}(\ell), w^{\mathsf{f}_j^*}_{\mathsf{aux}} - w^{\mathsf{f}_j^*}_+(jk\ell\ell')\right\rangle\right] \\ &\qquad\le \frac{2\sum_{i\in[n]}\sqrt{L^{\mathsf{x}}_i}}{\sqrt{n\mu^{\mathsf{x}}}}\left(V^r_w\left(w_{\mathsf{aux}}(jk\ell)\right) + V^r_{w_{\mathsf{aux}}(jk\ell)}\left(w_+(jk\ell\ell')\right)\right).\end{aligned}$$

Similarly, by applying Item (1) in Lemma 16 with $f = g_k$, $\alpha = (L^{\mathsf{y}}_k \mu^{\mathsf{y}})^{-\frac{1}{2}}$,

$$\begin{aligned}&\mathbb{E}_j\left[\frac{1}{nq_k}\left\langle w^{\mathsf{g}_k^*}_{\mathsf{aux}} - w^{\mathsf{g}_k^*}, w^{\mathsf{y}}_{\mathsf{aux}}(\ell) - w^{\mathsf{y}}_+(jk\ell\ell')\right\rangle + \frac{1}{nq_k}\left\langle w^{\mathsf{y}} - w^{\mathsf{y}}_{\mathsf{aux}}(\ell), w^{\mathsf{g}_k^*}_{\mathsf{aux}} - w^{\mathsf{g}_k^*}_+(jk\ell\ell')\right\rangle\right] \\ &\qquad\le \frac{2\sum_{i\in[n]}\sqrt{L^{\mathsf{y}}_i}}{\sqrt{n\mu^{\mathsf{y}}}}\left(V^r_w\left(w_{\mathsf{aux}}(jk\ell)\right) + V^r_{w_{\mathsf{aux}}(jk\ell)}\left(w_+(jk\ell\ell')\right)\right).\end{aligned}$$

Summing the above displays yields the desired claim. ■

Partial variance bound. Finally, we provide bounds on the components of (43) corresponding to Φ^h. Namely, we bound the quantity λ_1, and complete the bound on λ_0 within Proposition 29.

Lemma 36 *Following notation of Lemma 33, and recalling the definition (48), for*

$$\lambda_1 := 32(\lambda^h)^2,$$

where we define

$$\lambda^h := \frac{1}{n}\sum_{i\in[n]}\left(\frac{\Lambda^{\mathsf{xx}}_i}{\mu^{\mathsf{x}}} + \frac{\Lambda^{\mathsf{xy}}_i}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}} + \frac{\Lambda^{\mathsf{yy}}_i}{\mu^{\mathsf{y}}}\right). \tag{48}$$

we have for any $\rho > 0$,

$$\begin{aligned}&\mathbb{E}\left[\left\langle \Phi^h_{jk\ell'}(w_{\mathsf{aux}}(jk\ell)) - \Phi^h_{jk\ell}(w), w_{\mathsf{aux}}(jk\ell) - w_+(jk\ell\ell')\right\rangle\right] \\ &\qquad\le \left(2\lambda^h + \frac{1}{\rho}\right)\mathbb{E}\left[V^r_w\left(w_{\mathsf{aux}}(jk\ell)\right) + V^r_{w_{\mathsf{aux}}(jk\ell)}\left(w_+(jk\ell\ell')\right)\right] + \rho\lambda_1\mathbb{E}\left[V^r_{w_0}(w^\star) + V^r_{\bar{w}(\ell)}(w_\star)\right].\end{aligned} \tag{49}$$

Proof The proof is similar to (part of) the proof of Lemma 11. Fix j, k, ℓ, ℓ'. By definition,

$$\begin{aligned}&\left[\Phi^h_{jk\ell'}(w_{\mathsf{aux}}(jk\ell)) - \Phi^h_{jk\ell}(w)\right]^{\mathsf{xy}} \\ &= \frac{1}{nr_{\ell'}}\left(\nabla_x h_{\ell'}(w^{\mathsf{x}}_{\mathsf{aux}}(\ell), w^{\mathsf{y}}_{\mathsf{aux}}(\ell)) - \nabla_x h_{\ell'}(w^{\mathsf{x}}_0, w^{\mathsf{y}}_0), \nabla_y h_{\ell'}(w^{\mathsf{x}}_0, w^{\mathsf{y}}_0) - \nabla_y h_{\ell'}(w^{\mathsf{x}}_{\mathsf{aux}}(\ell), w^{\mathsf{y}}_{\mathsf{aux}}(\ell))\right) \\ &\qquad -\frac{1}{nr_\ell}\left(\nabla_x h_\ell(w^{\mathsf{x}}, w^{\mathsf{y}}) - \nabla_x h_\ell(w^{\mathsf{x}}_0, w^{\mathsf{y}}_0), \nabla_y h_\ell(w^{\mathsf{x}}_0, w^{\mathsf{y}}_0) - \nabla_y h_\ell(w^{\mathsf{x}}, w^{\mathsf{y}})\right).\end{aligned}$$

We decompose the x blocks of the left-hand side of (49) as

$$\begin{aligned}
\left\langle \left[\Phi^h_{jk\ell'}(w_{\mathsf{aux}}(jk\ell)) - \Phi^h_{jk\ell}(w)\right]^{\mathsf{x}}, w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_{+}(jk\ell\ell')\right\rangle &= \textcircled{1} + \textcircled{2} + \textcircled{3}, \\
\textcircled{1} := \frac{1}{nr_{\ell'}} \left\langle \nabla_x h_{\ell'}(w^{\mathsf{x}}_{\mathsf{aux}}(\ell), w^{\mathsf{y}}_{\mathsf{aux}}(\ell)) - \nabla_x h_{\ell'}(w^{\mathsf{x}}_0, w^{\mathsf{y}}_0), w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_{+}(jk\ell\ell')\right\rangle, \\
\textcircled{2} := \frac{1}{nr_{\ell}} \left\langle \nabla_x h_{\ell}(w^{\mathsf{x}}_0, w^{\mathsf{y}}_0) - \nabla_x h_{\ell}(w^{\mathsf{x}}_{\mathsf{aux}}(\ell), w^{\mathsf{y}}_{\mathsf{aux}}(\ell)), w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_{+}(jk\ell\ell')\right\rangle, \\
\textcircled{3} := \frac{1}{nr_{\ell}} \left\langle \nabla_x h_{\ell}(w^{\mathsf{x}}_{\mathsf{aux}}(\ell), w^{\mathsf{y}}_{\mathsf{aux}}(\ell)) - \nabla_x h_{\ell}(w^{\mathsf{x}}, w^{\mathsf{y}}), w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_{+}(jk\ell\ell')\right\rangle.
\end{aligned}$$

By the Lipschitzness bounds in (38) and Young's inequality,

$$\begin{aligned}
\textcircled{1} &\le \frac{1}{nr_{\ell'}} \left\|\nabla_x h_{\ell'}(w^{\mathsf{x}}_{\mathsf{aux}}(\ell), w^{\mathsf{y}}_{\mathsf{aux}}(\ell)) - \nabla_x h_{\ell'}(w^{\mathsf{x}}_0, w^{\mathsf{y}}_0)\right\| \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_{+}(jk\ell\ell')\right\| \\
&\le \frac{1}{nr_{\ell'}} \Lambda^{\mathsf{xx}}_{\ell'} \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_0\right\| \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_{+}(jk\ell\ell')\right\| \\
&\qquad + \frac{1}{nr_{\ell'}} \Lambda^{\mathsf{xy}}_{\ell'} \left\|w^{\mathsf{y}}_{\mathsf{aux}}(\ell) - w^{\mathsf{y}}_0\right\| \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_{+}(jk\ell\ell')\right\| \\
&\le \frac{2\rho(\Lambda^{\mathsf{xx}}_{\ell'})^2}{\mu^{\mathsf{x}} n^2 r^2_{\ell'}} \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_0\right\|^2 + \frac{2\rho(\Lambda^{\mathsf{xy}}_{\ell'})^2}{\mu^{\mathsf{x}} n^2 r^2_{\ell'}} \left\|w^{\mathsf{y}}_{\mathsf{aux}}(\ell) - w^{\mathsf{y}}_0\right\|^2 + \frac{\mu^{\mathsf{x}}}{4\rho} \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_{+}(jk\ell\ell')\right\|^2.
\end{aligned}$$

Symmetrically, we bound

$$\textcircled{2} \le \frac{2\rho(\Lambda^{\mathsf{xx}}_{\ell})^2}{\mu^{\mathsf{x}} n^2 r^2_{\ell}} \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_0\right\|^2 + \frac{2\rho(\Lambda^{\mathsf{xy}}_{\ell})^2}{\mu^{\mathsf{x}} n^2 r^2_{\ell}} \left\|w^{\mathsf{y}}_{\mathsf{aux}}(\ell) - w^{\mathsf{y}}_0\right\|^2 + \frac{\mu^{\mathsf{x}}}{4\rho} \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_{+}(jk\ell\ell')\right\|^2.$$

Finally, we have

$$\begin{aligned}
\textcircled{3} &\le \frac{1}{nr_{\ell}} \left\|\nabla_x h_{\ell}(w^{\mathsf{x}}_{\mathsf{aux}}(\ell), w^{\mathsf{y}}_{\mathsf{aux}}(\ell)) - \nabla_x h_{\ell}(w^{\mathsf{x}}, w^{\mathsf{y}})\right\| \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_{+}(jk\ell\ell')\right\| \\
&\le \frac{1}{nr_{\ell}} \Lambda^{\mathsf{xx}}_{\ell} \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}\right\| \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_{+}(jk\ell\ell')\right\| \\
&\qquad + \frac{1}{nr_{\ell}} \Lambda^{\mathsf{xy}}_{\ell} \left\|w^{\mathsf{y}}_{\mathsf{aux}}(\ell) - w^{\mathsf{y}}\right\| \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_{+}(jk\ell\ell')\right\| \\
&\le \frac{1}{nr_{\ell}} \left(\frac{\Lambda^{\mathsf{xx}}_{\ell}}{\mu^{\mathsf{x}}} \left(\frac{\mu^{\mathsf{x}}}{2} \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}\right\|^2 + \frac{\mu^{\mathsf{x}}}{2} \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_{+}(jk\ell\ell')\right\|^2\right)\right) \\
&\qquad + \frac{1}{nr_{\ell}} \left(\frac{\Lambda^{\mathsf{xy}}_{\ell}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}} \left(\frac{\mu^{\mathsf{y}}}{2} \left\|w^{\mathsf{y}}_{\mathsf{aux}}(\ell) - w^{\mathsf{y}}\right\|^2 + \frac{\mu^{\mathsf{x}}}{2} \left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell) - w^{\mathsf{x}}_{+}(jk\ell\ell')\right\|^2\right)\right).
\end{aligned}$$

We may similarly decompose the y blocks of the left-hand side of (49) as $\textcircled{4}+\textcircled{5}+\textcircled{6}$, where symmetrically, we have

$$\begin{aligned}
\textcircled{4} &\le \frac{2\rho(\Lambda^{\mathsf{yy}}_{\ell'})^2}{\mu^{\mathsf{y}} n^2 r_{\ell'}^2}\left\|w^{\mathsf{y}}_{\mathsf{aux}}(\ell)-w^{\mathsf{y}}_0\right\|^2 + \frac{2\rho(\Lambda^{\mathsf{xy}}_{\ell'})^2}{\mu^{\mathsf{y}} n^2 r_{\ell'}^2}\left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell)-w^{\mathsf{x}}_0\right\|^2 + \frac{\mu^{\mathsf{y}}}{4\rho}\left\|w^{\mathsf{y}}_{\mathsf{aux}}(\ell)-w^{\mathsf{y}}_+(jk\ell\ell')\right\|^2, \\
\textcircled{5} &\le \frac{2\rho(\Lambda^{\mathsf{yy}}_{\ell})^2}{\mu^{\mathsf{y}} n^2 r_{\ell}^2}\left\|w^{\mathsf{y}}_{\mathsf{aux}}(\ell)-w^{\mathsf{y}}_0\right\|^2 + \frac{2\rho(\Lambda^{\mathsf{xy}}_{\ell})^2}{\mu^{\mathsf{y}} n^2 r_{\ell}^2}\left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell)-w^{\mathsf{x}}_0\right\|^2 + \frac{\mu^{\mathsf{y}}}{4\rho}\left\|w^{\mathsf{y}}_{\mathsf{aux}}(\ell)-w^{\mathsf{y}}_+(jk\ell\ell')\right\|^2, \\
\textcircled{6} &\le \frac{1}{nr_\ell}\left(\frac{\Lambda^{\mathsf{yy}}_{\ell}}{\mu^{\mathsf{y}}}\left(\frac{\mu^{\mathsf{y}}}{2}\left\|w^{\mathsf{y}}_{\mathsf{aux}}(\ell)-w^{\mathsf{y}}\right\|^2+\frac{\mu^{\mathsf{y}}}{2}\left\|w^{\mathsf{y}}_{\mathsf{aux}}(\ell)-w^{\mathsf{y}}_+(jk\ell\ell')\right\|^2\right)\right) \\
&\quad + \frac{1}{nr_\ell}\left(\frac{\Lambda^{\mathsf{xy}}_{\ell}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}}\left(\frac{\mu^{\mathsf{x}}}{2}\left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell)-w^{\mathsf{x}}\right\|^2+\frac{\mu^{\mathsf{y}}}{2}\left\|w^{\mathsf{y}}_{\mathsf{aux}}(\ell)-w^{\mathsf{y}}_+(jk\ell\ell')\right\|^2\right)\right).
\end{aligned}$$

We first observe that by definition of r and nonnegativity of Bregman divergences,

$$\begin{aligned}
\textcircled{3}+\textcircled{6} &\le \frac{1}{nr_\ell}\left(\frac{\Lambda^{\mathsf{xx}}_{\ell}}{\mu^{\mathsf{x}}}+\frac{\Lambda^{\mathsf{xy}}_{\ell}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}}+\frac{\Lambda^{\mathsf{yy}}_{\ell}}{\mu^{\mathsf{y}}}\right)\left(V^r_w(w_{\mathsf{aux}}(jk\ell))+V^r_{w_{\mathsf{aux}}(jk\ell)}(w_+(jk\ell\ell'))\right) \\
&\le 2\lambda^h\left(V^r_w(w_{\mathsf{aux}}(jk\ell))+V^r_{w_{\mathsf{aux}}(jk\ell)}(w_+(jk\ell\ell'))\right).
\end{aligned}$$

Moreover, since by the triangle inequality and $(a+b)^2 \le 2a^2+2b^2$,

$$\begin{aligned}
\left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell)-w^{\mathsf{x}}_0\right\|^2 &\le 2\left\|w^{\mathsf{x}}_{\mathsf{aux}}(\ell)-w^{\mathsf{x}}_\star\right\|^2+2\left\|w^{\mathsf{x}}_0-w^{\mathsf{x}}_\star\right\|^2, \\
\left\|w^{\mathsf{y}}_{\mathsf{aux}}(\ell)-w^{\mathsf{y}}_0\right\|^2 &\le 2\left\|w^{\mathsf{y}}_{\mathsf{aux}}(\ell)-w^{\mathsf{y}}_\star\right\|^2+2\left\|w^{\mathsf{y}}_0-w^{\mathsf{y}}_\star\right\|^2,
\end{aligned}$$

we have by definition of r and λ_1,

$$\begin{aligned}
\textcircled{1}+\textcircled{2}+\textcircled{4}+\textcircled{5} &\le \frac{1}{\rho}\left(V^r_w(w_{\mathsf{aux}}(jk\ell))+V^r_{w_{\mathsf{aux}}(jk\ell)}(w_+(jk\ell\ell'))\right) \\
&\quad + \rho\lambda_1\left(V^r_{w_0}(w_\star)+V^r_{\bar{w}(\ell)}(w_\star)\right).
\end{aligned}$$

Summing the above displays and taking expectations yields the claim. ■

Combining the properties we prove above with Proposition 29, we obtain the following convergence guarantee for each loop $0 \le \tau < N$ in Algorithm 8.

Proposition 37 *Consider a run of the inner for loop in Algorithm 8 initialized at $w_0 \in \mathcal{Z}$, with*

$$\lambda \leftarrow \left(2n(1+\gamma)+\frac{2\sum_{i\in[n]}\sqrt{L^{\mathsf{x}}_i}}{\sqrt{n\mu^{\mathsf{x}}}}+\frac{2\sum_{i\in[n]}\sqrt{L^{\mathsf{y}}_i}}{\sqrt{n\mu^{\mathsf{y}}}}+2\lambda^h\right)+\frac{160(\lambda^h)^2}{\gamma},\ S \leftarrow \frac{5\lambda}{\gamma}, \tag{50}$$

where λ^h is defined in (48). Letting $\widetilde{w}$ be the new setting of w_0 in Line 8 at the end of the run,

$$\mathbb{E}\left[V^r_{\widetilde{w}}(w^\star)\right] \le \frac{1}{2}V^r_{w_0}(w^\star),$$

where $w^\star$ solves the VI in Φ (defined in (42)).

F.4. Outer loop convergence analysis

We state the following convergence guarantee on our outer loop, Algorithm 7. The analysis is a somewhat technical modification of the standard proximal point analysis for solving VIs (Nemirovski, 2004), to handle approximation error.

Proposition 38 *Consider a single iteration $0 \le t < T$ of Algorithm 7, and let $z_\star$ is the saddle point to $F_{\text{mmfs-pd}}$ (defined in (39)). Setting S as in (50) and*

$$N := O\left(\log\left(\gamma\lambda\right)\right), \tag{51}$$

for an appropriately large constant in our implementation of Algorithm 8 and λ as in (50), we have

$$\mathbb{E}V^r_{z_{t+1}}(z_\star) \le \frac{4\gamma}{1+4\gamma}V^r_{z_t}(z_\star).$$

Proof Fix an iteration $t \in [T]$ of Algorithm 7, and let $z^\star_{t+1}$ be the exact solution to the VI in $\Phi^{\text{mmfs-pd}} + \gamma\nabla r - \nabla r(z_t)$. By the guarantee of Proposition 37, after the stated number of NS iterations in Algorithm 8 (for an appropriately large constant), we obtain a point z_{t+1} such that

$$\mathbb{E}\left[V^r_{z_{t+1}}(z^\star_{t+1})\right] \le \frac{1}{1+3\gamma\widetilde{\kappa}}V^r_{z_t}(\hat{z}_{t+1}),\ \text{where } \widetilde{\kappa} := 10\sum_{i\in[n]}\left(\frac{L^{\mathsf{x}}_i + \Lambda^{\mathsf{xx}}_i}{\mu^{\mathsf{x}}} + \frac{L^{\mathsf{y}}_i + \Lambda^{\mathsf{yy}}_i}{\mu^{\mathsf{y}}} + \frac{\Lambda^{\mathsf{xy}}_i}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}}\right)^2. \tag{52}$$

The optimality condition on $z^\star_{t+1}$ yields

$$\left\langle \Phi^{\text{mmfs-pd}}\left(z^\star_{t+1}\right), z^\star_{t+1} - z_\star\right\rangle \le \gamma V^r_{z_t}\left(z_\star\right) - \gamma V^r_{z^\star_{t+1}}\left(z_\star\right) - \gamma V_{z_t}\left(z^\star_{t+1}\right).$$

Rearranging terms then gives:

$$\begin{aligned}
&\left\langle \Phi^{\text{mmfs-pd}}\left(z_{t+1}\right), z_{t+1} - z_\star\right\rangle \\
&\quad\le \gamma V^r_{z_t}\left(z_\star\right) - \gamma V^r_{z_{t+1}}\left(z_\star\right) - \gamma V^r_{z_t}\left(z^\star_{t+1}\right) + \gamma\left(V^r_{z_{t+1}}\left(z_\star\right) - V^r_{z^\star_{t+1}}\left(z_\star\right)\right) \\
&\qquad + \left\langle \Phi^{\text{mmfs-pd}}\left(z_{t+1}\right) - \Phi^{\text{mmfs-pd}}\left(z^\star_{t+1}\right), z^\star_{t+1} - z_\star\right\rangle \\
&\qquad + \left\langle \Phi^{\text{mmfs-pd}}\left(z_{t+1}\right), z_{t+1} - z^\star_{t+1}\right\rangle \\
&\quad= \gamma V^r_{z_t}\left(z_\star\right) - \gamma V^r_{z_{t+1}}\left(z_\star\right) - \gamma V^r_{z_t}\left(z^\star_{t+1}\right) + \gamma V^r_{z_{t+1}}\left(z^\star_{t+1}\right) \\
&\qquad + \gamma\left\langle \nabla r\left(z_{t+1}\right) - \nabla r\left(z^\star_{t+1}\right), z^\star_{t+1} - z_\star\right\rangle \\
&\qquad + \left\langle \Phi^{\text{mmfs-pd}}\left(z_{t+1}\right) - \Phi^{\text{mmfs-pd}}\left(z^\star_{t+1}\right), z^\star_{t+1} - z_\star\right\rangle + \left\langle \Phi^{\text{mmfs-pd}}\left(z_{t+1}\right), z_{t+1} - z^\star_{t+1}\right\rangle \\
&\quad\le \gamma V^r_{z_t}\left(z_\star\right) - \gamma V^r_{z_{t+1}}\left(z_\star\right) - \gamma V^r_{z_t}\left(z^\star_{t+1}\right) + \gamma V^r_{z_{t+1}}\left(z^\star_{t+1}\right) \\
&\qquad + \gamma\left\langle \nabla r\left(z_{t+1}\right) - \nabla r\left(z^\star_{t+1}\right), z_{t+1} - z_\star\right\rangle \\
&\qquad + \left\langle \Phi^{\text{mmfs-pd}}\left(z_{t+1}\right) - \Phi^{\text{mmfs-pd}}\left(z^\star_{t+1}\right), z_{t+1} - z_\star\right\rangle \\
&\qquad + \left\langle \Phi^{\text{mmfs-pd}}\left(z_{t+1}\right) - \Phi^{\text{mmfs-pd}}\left(z_\star\right), z_{t+1} - z^\star_{t+1}\right\rangle.
\end{aligned} \tag{53}$$

In the only equality, we used the identity (18). The last inequality used monotonicity of the operators $\gamma\nabla r$ and $\Phi^{\text{mmfs-pd}}$, as well as $\Phi^{\text{mmfs-pd}}(z_\star) = 0$ because it is an unconstrained minimax optimization problem. In the remainder of the proof, we will bound the last three lines of (53).

First, for any $\alpha > 0$, we bound:

$$\begin{aligned}
&\left\langle \nabla r\left(z_{t+1}\right) - \nabla r\left(z^\star_{t+1}\right), z_{t+1} - z_\star \right\rangle \\
&\quad = \mu^{\mathsf{x}} \left\langle z^{\mathsf{x}}_{t+1} - (z^\star_{t+1})^{\mathsf{x}}, z^{\mathsf{x}}_{t+1} - z^{\mathsf{x}}_\star \right\rangle + \mu^{\mathsf{y}} \left\langle z^{\mathsf{y}}_{t+1} - (z^\star_{t+1})^{\mathsf{y}}, z^{\mathsf{y}}_{t+1} - z^{\mathsf{y}}_\star \right\rangle \\
&\qquad + \frac{1}{n} \sum_{i \in [n]} \left\langle \nabla f_i^*(z^{\mathsf{f}_\mathsf{i}^*}_{t+1}) - \nabla f_i^*((z^\star_{t+1})^{\mathsf{f}_\mathsf{i}^*}), z^{\mathsf{f}_\mathsf{i}^*}_{t+1} - z^{\mathsf{f}_\mathsf{i}^*}_\star \right\rangle \\
&\qquad + \frac{1}{n} \sum_{i \in [n]} \left\langle \nabla g_i^*(z^{\mathsf{g}_\mathsf{i}^*}_{t+1}) - \nabla g_i^*((z^\star_{t+1})^{\mathsf{g}_\mathsf{i}^*}), z^{\mathsf{g}_\mathsf{i}^*}_{t+1} - z^{\mathsf{g}_\mathsf{i}^*}_\star \right\rangle \\
&\quad \le 2\alpha\mu^{\mathsf{x}} \left\| z^{\mathsf{x}}_{t+1} - (z^\star_{t+1})^{\mathsf{x}} \right\|^2 + \frac{\mu^{\mathsf{x}}}{8\alpha} \left\| z^{\mathsf{x}}_{t+1} - z^{\mathsf{x}}_\star \right\|^2 + 2\alpha\mu^{\mathsf{y}} \left\| z^{\mathsf{y}}_{t+1} - (z^\star_{t+1})^{\mathsf{y}} \right\|^2 + \frac{\mu^{\mathsf{y}}}{8\alpha} \left\| z^{\mathsf{y}}_{t+1} - z^{\mathsf{y}}_\star \right\|^2 \\
&\qquad + \frac{1}{n} \sum_{i \in [n]} \left(\frac{2\alpha L^{\mathsf{x}}_i}{(\mu^{\mathsf{x}})^2} \left\| z^{\mathsf{f}_\mathsf{i}^*}_{t+1} - (z^\star_{t+1})^{\mathsf{f}_\mathsf{i}^*} \right\|^2 + \frac{1}{8\alpha L^{\mathsf{x}}_i} \left\| z^{\mathsf{f}_\mathsf{i}^*}_{t+1} - z^{\mathsf{f}_\mathsf{i}^*}_\star \right\|^2 \right) \\
&\qquad + \frac{1}{n} \sum_{i \in [n]} \left(\frac{2\alpha L^{\mathsf{y}}_i}{(\mu^{\mathsf{y}})^2} \left\| z^{\mathsf{g}_\mathsf{i}^*}_{t+1} - (z^\star_{t+1})^{\mathsf{g}_\mathsf{i}^*} \right\|^2 + \frac{1}{8\alpha L^{\mathsf{y}}_i} \left\| z^{\mathsf{g}_\mathsf{i}^*}_{t+1} - z^{\mathsf{g}_\mathsf{i}^*}_\star \right\|^2 \right) \\
&\quad \le \frac{1}{4\alpha} V^r_{z_{t+1}}(z_\star) + \widetilde{\kappa}\alpha V^r_{z_{t+1}}(z^\star_{t+1}). \qquad (54)
\end{aligned}$$

The equality used the definition of r in (40). The first inequality used Young's and Cauchy-Schwarz on the $\mathcal{X} \times \mathcal{Y}$ blocks, as well as $\frac{1}{\mu_i^{\mathsf{x}}}$-smoothness of the f_i^* from Assumption 3 and Item 4 in Fact 1 (and similar bounds on each g_i^*). The last inequality used strong convexity of each piece of r.

Similarly, by definition of $\Phi^{\text{mmfs-pd}}$ (41) which we denote for Φ for brevity in the following:

$$\begin{aligned}
&\left\langle \Phi\left(z_{t+1}\right) - \Phi\left(z^\star_{t+1}\right), z_{t+1} - z_\star \right\rangle \\
&\quad \le \frac{1}{8} V^r_{z_{t+1}}(z_\star) + 2\widetilde{\kappa} V^r_{z_{t+1}}(z^\star_{t+1}) + \frac{1}{n} \sum_{i \in [n]} \left\langle \nabla_x h_i(z^{\mathsf{x}}_{t+1}, z^{\mathsf{y}}_{t+1}) - \nabla_x h_i((z^\star_{t+1})^{\mathsf{x}}, (z^\star_{t+1})^{\mathsf{y}}), z^{\mathsf{x}}_{t+1} - z^{\mathsf{x}}_\star \right\rangle \\
&\qquad + \frac{1}{n} \sum_{i \in [n]} \left\langle \nabla_y h_i(z^{\mathsf{x}}_{t+1}, z^{\mathsf{y}}_{t+1}) - \nabla_y h_i((z^\star_{t+1})^{\mathsf{x}}, (z^\star_{t+1})^{\mathsf{y}}), z^{\mathsf{y}}_{t+1} - z^{\mathsf{y}}_\star \right\rangle \\
&\qquad + \frac{1}{n} \sum_{i \in [n]} \left(\left\langle z^{\mathsf{f}_\mathsf{i}^*}_{t+1} - (z^\star_{t+1})^{\mathsf{f}_\mathsf{i}^*}, z^{\mathsf{x}}_{t+1} - z^{\mathsf{x}}_\star \right\rangle + \left\langle z^{\mathsf{g}_\mathsf{i}^*}_{t+1} - (z^\star_{t+1})^{\mathsf{g}_\mathsf{i}^*}, z^{\mathsf{y}}_{t+1} - z^{\mathsf{y}}_\star \right\rangle \right) \\
&\qquad - \frac{1}{n} \sum_{i \in [n]} \left(\left\langle z^{\mathsf{x}}_{t+1} - (z^\star_{t+1})^{\mathsf{x}}, z^{\mathsf{f}_\mathsf{i}^*}_{t+1} - z^{\mathsf{f}_\mathsf{i}^*}_\star \right\rangle + \left\langle z^{\mathsf{y}}_{t+1} - (z^\star_{t+1})^{\mathsf{y}}, z^{\mathsf{g}_\mathsf{i}^*}_{t+1} - z^{\mathsf{g}_\mathsf{i}^*}_\star \right\rangle \right)
\end{aligned}$$

where we used (54) to bound the ∇r terms. Consequently,

$$\begin{aligned}
&\left\langle \Phi\left(z_{t+1}\right) - \Phi\left(z^\star_{t+1}\right), z_{t+1} - z_\star \right\rangle \\
&\quad\le \frac{1}{8} V^r_{z_{t+1}}(z_\star) + 2\widetilde{\kappa} V^r_{z_{t+1}}(z^\star_{t+1}) + \frac{1}{n}\sum_{i\in[n]} \left(\frac{\mu^{\mathsf{x}}}{16} V_{z^{\mathsf{x}}_{t+1}}(z^{\mathsf{x}}_\star) + \frac{\mu^{\mathsf{y}}}{16} V_{z^{\mathsf{y}}_{t+1}}(z^{\mathsf{y}}_\star) \right) \\
&\qquad + \frac{1}{n}\sum_{i\in[n]} \left(16\left(\frac{(\Lambda^{\mathsf{xx}}_i)^2}{\mu^{\mathsf{x}}} + \frac{(\Lambda^{\mathsf{xy}}_i)^2}{\mu^{\mathsf{y}}} \right) V_{z^{\mathsf{x}}_{t+1}}((z^\star_{t+1})^{\mathsf{x}}) \right) \\
&\qquad + \frac{1}{n}\sum_{i\in[n]} \left(16\left(\frac{(\Lambda^{\mathsf{xy}}_i)^2}{\mu^{\mathsf{x}}} + \frac{(\Lambda^{\mathsf{yy}}_i)^2}{\mu^{\mathsf{y}}} \right) V_{z^{\mathsf{y}}_{t+1}}((z^\star_{t+1})^{\mathsf{y}}) \right) \\
&\qquad + \frac{1}{n}\sum_{i\in[n]} \left(\frac{\mu^{\mathsf{x}}}{16} V_{z^{\mathsf{x}}_{t+1}}(z^{\mathsf{x}}_\star) + \frac{\mu^{\mathsf{y}}}{16} V_{z^{\mathsf{y}}_{t+1}}(z^{\mathsf{y}}_\star) \right) \\
&\qquad + \frac{1}{n}\sum_{i\in[n]} \left(\frac{8}{\mu^{\mathsf{x}}} \left\| z^{\mathsf{f}_\mathsf{i}^*}_{t+1} - (z^\star_{t+1})^{\mathsf{f}_\mathsf{i}^*} \right\|^2 + \frac{8}{\mu^{\mathsf{y}}} \left\| z^{\mathsf{g}_\mathsf{i}^*}_{t+1} - (z^\star_{t+1})^{\mathsf{g}_\mathsf{i}^*} \right\|^2 \right) \\
&\qquad + \frac{1}{n}\sum_{i\in[n]} \left(\frac{1}{8} V^{f_i^*}_{z^{\mathsf{f}_\mathsf{i}^*}_{t+1}}\left(z^{\mathsf{f}_\mathsf{i}^*}_\star\right) + \frac{1}{8} V^{g_i^*}_{z^{\mathsf{g}_\mathsf{i}^*}_{t+1}}\left(z^{\mathsf{g}_\mathsf{i}^*}_\star\right) \right) \\
&\qquad + \frac{1}{n}\sum_{i\in[n]} \left(8L^{\mathsf{x}}_i V_{z^{\mathsf{x}}_{t+1}}((z^\star_{t+1})^{\mathsf{x}}) + 8L^{\mathsf{y}}_i V_{z^{\mathsf{y}}_{t+1}}((z^\star_{t+1})^{\mathsf{y}}) \right) \\
&\quad\le \frac{1}{4} V^r_{z_{t+1}}(z_\star) + \widetilde{\kappa} V^r_{z_{t+1}}(z^\star_{t+1}). \qquad (55)
\end{aligned}$$

In the first inequality, we used Cauchy-Schwarz, Young's, and our various smoothness assumptions (as well as strong convexity of each f_i^* and g_i^*). The last inequality used strong convexity of each piece of r.

For the last term, by a similar argument as in the previous bounds, we have

$$\left\langle \Phi(z_{t+1}) - \Phi(z_\star), z_{t+1} - z^\star_{t+1} \right\rangle \le \frac{1}{4} V^r_{z_{t+1}}(z_\star) + \widetilde{\kappa} V^r_{z_{t+1}}(z^\star_{t+1}). \qquad (56)$$

Plugging the inequalities (54) with $\alpha = \gamma$, (55) and (56) back into (53), this implies

$$\begin{aligned}
\left\langle \Phi^{\text{mmfs-pd}}\left(z_{t+1}\right), z_{t+1} - z_\star \right\rangle &\le \gamma V^r_{z_t}\left(z_\star\right) - \gamma V^r_{z_{t+1}}\left(z_\star\right) - \gamma V^r_{z_t}\left(z^\star_{t+1}\right) + \gamma V^r_{z_{t+1}}\left(z^\star_{t+1}\right) \qquad (57) \\
&\quad + \frac{3}{4} V^r_{z_{t+1}}(z_\star) + 3\widetilde{\kappa}\gamma^2 V^r_{z_{t+1}}(z^\star_{t+1}). \qquad (58)
\end{aligned}$$

By strong monotonicity of $\Phi^{\text{mmfs-pd}}$ with respect to r, we also have

$$\left\langle \Phi^{\text{mmfs-pd}}\left(z_{t+1}\right), z_{t+1} - z_\star \right\rangle \ge \left\langle \Phi^{\text{mmfs-pd}}\left(z_{t+1}\right) - \Phi^{\text{mmfs-pd}}\left(z_\star\right), z_{t+1} - z_\star \right\rangle \ge V^r_{z_{t+1}}\left(z_\star\right). \qquad (59)$$

Combining (58) and (59) with the assumption (52), and taking expectations, we obtain

$$\left(\frac{1}{4} + \gamma\right) \mathbb{E} V^r_{z_{t+1}}(z_\star) \le \gamma V^r_{z_t}(z_\star) \implies \mathbb{E} V^r_{z_{t+1}}(z_\star) \le \frac{4\gamma}{1+4\gamma} V^r_{z_t}(z_\star).$$

■

F.5. Main result

We now state and prove our main claim.

Theorem 39 *Suppose F_{mmfs} in (37) satisfies Assumption 3, and has saddle point $(x_\star, y_\star)$. Further, suppose we have $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$ such that $\text{Gap}_{F_{\text{mmfs-reg}}}(x_0, y_0) \le \epsilon_0$. Algorithm 7 using Algorithm 8 with λ as in (50) returns $(x, y) \in \mathcal{X} \times \mathcal{Y}$ with $\mathbb{E}\text{Gap}_{F_{\text{mmfs-reg}}}(x, y) \le \epsilon$ in N_{tot} iterations, using a total of $O(N_{\text{tot}})$ gradient calls each to some f_i, g_i, or h_i for $i \in [n]$, where*

$$N_{\text{tot}} = O\left(\kappa_{\text{mmfs}} \log\left(\kappa_{\text{mmfs}}\right) \log\left(\frac{\kappa_{\text{mmfs}}\epsilon_0}{\epsilon}\right)\right),$$
$$\text{for } \kappa_{\text{mmfs}} := n + \frac{1}{\sqrt{n}} \sum_{i\in[n]} \left(\sqrt{\frac{L_i^{\mathsf{x}}}{\mu^{\mathsf{x}}}} + \sqrt{\frac{L_i^{\mathsf{y}}}{\mu^{\mathsf{y}}}} + \frac{\Lambda_i^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\Lambda_i^{\mathsf{xy}}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}} + \frac{\Lambda_i^{\mathsf{yy}}}{\mu^{\mathsf{y}}}\right). \tag{60}$$

In particular, we use $N_{\text{tot}} = NTS$ for

$$T = O\left(\gamma \log\left(\frac{\kappa_{\text{fs}}\epsilon_0}{\epsilon}\right)\right),\ N = O\left(\log\left(\kappa_{\text{mmfs}}\right)\right),\ S = O\left(n + \frac{\kappa_{\text{mmfs}}}{\gamma} + \frac{(\lambda^h)^2}{\gamma^2}\right),\ \gamma = \frac{\lambda^h}{\sqrt{n}}.$$

Proof By Lemma 28, the point $(x_\star, y_\star)$ is consistent between (37) and (39). The complexity of each iteration follows from observation of Algorithm 7 and 8.

Next, by Proposition 37 and Proposition 38, and our choices of T, N, and S for appropriately large constants, we obtain a point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ such that

$$\mathbb{E}V^r_{(x,y)}(z_\star) \le \frac{\epsilon}{4}\left(\frac{1}{\kappa_{\text{mmfs}}}\right)^2.$$

Here we used an analogous argument to Lemma 12 to bound the initial divergence. We then use a similar bound as in Lemma 13 to obtain the desired duality gap bound. ■

We now revisit the problem (36). We apply a generic reduction framework for minimax optimization to develop a solver for this problem under a relaxed version of Assumption 3, without requiring strong convexity of individual summands.

Assumption 4 *We assume the following about (36) for all $i \in [n]$.*

(1) f_i is L_i^{x}-smooth, and g_i is L_i^{y}-smooth.

(2) h has the following blockwise-smoothness properties: for all $u, v \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned} \|\nabla_x h_i(u) - \nabla_x h_i(v)\| &\le \Lambda_i^{\mathsf{xx}} \|u^{\mathsf{x}} - v^{\mathsf{x}}\| + \Lambda_i^{\mathsf{xy}} \|u^{\mathsf{y}} - v^{\mathsf{y}}\|, \\ \|\nabla_y h_i(u) - \nabla_y h_i(v)\| &\le \Lambda_i^{\mathsf{xy}} \|u^{\mathsf{x}} - v^{\mathsf{x}}\| + \Lambda_i^{\mathsf{yy}} \|u^{\mathsf{y}} - v^{\mathsf{y}}\|. \end{aligned} \tag{61}$$

First, we give the following generic reduction for strongly convex-concave optimization in the form of an algorithm. For simplicity we define for $z = (z^{\mathsf{x}}, z^{\mathsf{y}}) \in \mathcal{X} \times \mathcal{Y}$,

$$\omega(z) := \frac{\mu^{\mathsf{x}}}{2}\|z^{\mathsf{x}}\|^2 + \frac{\mu^{\mathsf{y}}}{2}\|z^{\mathsf{y}}\|^2.$$

Algorithm 9: REDX-MINIMAX: Reduction for minimax

```
Input: $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that $F(\cdot, y)$ is $\mu^{\mathsf{x}}$-strongly convex for all $y \in \mathcal{Y}$ and $F(x, \cdot)$
$\mu^{\mathsf{y}}$-strongly concave for all $x \in \mathcal{X}$, $z_0 \in \mathcal{X} \times \mathcal{Y}$
Parameter(s): $K \in \mathbb{N}$
for $0 \le k < K$ do
    $z_{k+1} \leftarrow$ any (possibly random) point satisfying
        $$\mathbb{E}\left[V^{\omega}_{z_{k+1}}\left(z^{\star}_{k+1}\right)\right] \le \frac{1}{4}\left(V^{\omega}_{z_k}\left(z^{\star}_{k+1}\right)\right),$$
        where $z^{\star}_{k+1} := \text{argmin}_{z^{\mathsf{x}} \in \mathcal{X}} \text{argmax}_{z^{\mathsf{y}} \in \mathcal{Y}} F(z^{\mathsf{x}}, z^{\mathsf{y}}) + \frac{\mu^{\mathsf{x}}}{4} V_{z^{\mathsf{x}}_k}\left(z^{\mathsf{x}}\right) - \frac{\mu^{\mathsf{y}}}{4} V_{z^{\mathsf{y}}_k}\left(z^{\mathsf{y}}\right)$
end
```

Lemma 40 *In Algorithm 9, letting $(x_\star, y_\star)$ be the saddle point of F, we have for every $k \in [K]$:*

$$\mathbb{E}\left[V^{\omega}_{z_k}(z_\star)\right] \le \frac{1}{2^k} V^{\omega}_{z_0}(z_\star).$$

Proof By applying the optimality conditions on $z^{\star}_{k+1}$, strong convexity-concavity of F, and (18), and letting Φ^F be the gradient operator of F,

$$\begin{aligned}
\left\langle \Phi^F(z^{\star}_{k+1}), z^{\star}_{k+1} - z_\star \right\rangle &\le \frac{\mu^{\mathsf{x}}}{4} \left\langle z^{\mathsf{x}}_k - [z^{\star}_{k+1}]^{\mathsf{x}}, [z^{\star}_{k+1}]^{\mathsf{x}} - z^{\mathsf{x}}_\star \right\rangle \\
&\quad + \frac{\mu^{\mathsf{y}}}{4} \left\langle z^{\mathsf{y}}_k - [z^{\star}_{k+1}]^{\mathsf{y}}, [z^{\star}_{k+1}]^{\mathsf{y}} - z^{\mathsf{y}}_\star \right\rangle \\
\implies V^{\omega}_{z^{\star}_{k+1}}(z_\star) &\le \left\langle \Phi^F(z^{\star}_{k+1}), z^{\star}_{k+1} - z_\star \right\rangle \\
&\le \frac{1}{4} V^{\omega}_{z_k}(z_\star) - \frac{1}{4} V^{\omega}_{z^{\star}_{k+1}}(z_\star) - \frac{1}{4} V_{z_k}(z^{\star}_{k+1}).
\end{aligned}$$

Further by the triangle inequality and $(a+b)^2 \le 2a^2 + 2b^2$, we have

$$V^{\omega}_{z_{k+1}}(z_\star) \le 2V^{\omega}_{z_{k+1}}(z^{\star}_{k+1}) + 2V^{\omega}_{z^{\star}_{k+1}}(z_\star).$$

Hence, combining these pieces,

$$\begin{aligned}
\mathbb{E}V^{\omega}_{z_{k+1}}(z_\star) &\le 2V^{\omega}_{z^{\star}_{k+1}}(z_\star) + 2\mathbb{E}V^{\omega}_{z_{k+1}}(z^{\star}_{k+1}) \\
&\le 2V^{\omega}_{z^{\star}_{k+1}}(z_\star) + \frac{1}{2} V^{\omega}_{z_k}(z^{\star}_{k+1}) \\
&\le \frac{1}{2} V^{\omega}_{z_k}(z_\star) - \frac{1}{2} V^{\omega}_{z^{\star}_{k+1}}(z_\star) \le \frac{1}{2} V^{\omega}_{z_k}(z_\star).
\end{aligned}$$

■

We apply this reduction in order to prove Corollary 41, for minimax finite sum optimization problems with the set of relaxed conditions in Assumption 4.

Corollary 41 *Suppose the summands $\{f_i, g_i, h_i\}_{i\in[n]}$ in (36) satisfy Assumption 4, and F_{mmfs} is μ^{x}-strongly convex in x, μ^{y}-strongly convex in y, with saddle point $(x_\star, y_\star)$. Further, suppose we have $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$ such that $\text{Gap}_{F_{\text{mmfs}}}(x_0, y_0) \le \epsilon_0$. Algorithm 6 using Algorithm 7 and 8 to implement steps returns $(x, y) \in \mathcal{X} \times \mathcal{Y}$ with $\mathbb{E}\text{Gap}(x, y) \le \epsilon$ in N_{tot} iterations, using a total of $O(N_{\text{tot}})$ gradient calls each to some f_i, g_i, or h_i for $i \in [n]$, where*

$$N_{\text{tot}} = O\left(\kappa_{\text{mmfs}} \log(\kappa_{\text{mmfs}}) \log\left(\frac{\kappa_{\text{mmfs}}\epsilon_0}{\epsilon}\right)\right),$$

$$\textit{for } \kappa_{\text{mmfs}} := n + \frac{1}{\sqrt{n}} \sum_{i\in[n]} \left(\sqrt{\frac{L_i^{\mathsf{x}}}{\mu^{\mathsf{x}}}} + \sqrt{\frac{L_i^{\mathsf{y}}}{\mu^{\mathsf{y}}}} + \frac{\Lambda_i^{\mathsf{xx}}}{\mu^{\mathsf{x}}} + \frac{\Lambda_i^{\mathsf{xy}}}{\sqrt{\mu^{\mathsf{x}}\mu^{\mathsf{y}}}} + \frac{\Lambda_i^{\mathsf{yy}}}{\mu^{\mathsf{y}}}\right).$$

Proof The overhead K is asymptotically the same here as the logarithmic term in the parameter T in Theorem 39, by analogous smoothness and strong convexity arguments. Moreover, we use Theorem 39 with μ^{x}, μ^{y} rescaled by constants to solve each subproblem required by Algorithm 9; in particular, the subproblem is equivalent to approximately finding a saddle point to $F_{\text{fs}}(z) + \frac{\mu^{\mathsf{x}}}{8}\|z^{\mathsf{x}}\|^2 - \frac{\mu^{\mathsf{y}}}{8}\|z^{\mathsf{y}}\|^2$, up to a linear shift which does not affect any smoothness bounds. We note that we will initialize the subproblem solver in iteration k with z_k. We hence can set $T = O(\gamma)$, yielding the desired iteration bound. ■