Increasing Iterate Averaging for Solving Saddle-Point Problems

Yuan Gao, Christian Kroer, Donald Goldfarb
Columbia University, Department of Industrial Engineering and Operations Research
gao.yuan@columbia.edu, christian.kroer@columbia.edu, goldfarb@columbia.edu

Abstract

Many problems in machine learning and game theory can be formulated as saddle-point problems, for which various first-order methods have been developed and proven efficient in practice. Under the general convex-concave assumption, most first-order methods only guarantee an ergodic convergence rate, that is, the uniform averages of the iterates converge at a $O(1/T)$ rate in terms of the saddle-point residual. However, numerically, the iterates themselves can often converge much faster than the uniform averages. This observation motivates increasing averaging schemes that put more weight on later iterates, in contrast to the usual uniform averaging. We show that such increasing averaging schemes, applied to various first-order methods, are able to preserve the $O(1/T)$ convergence rate with no additional assumptions or computational overhead. Extensive numerical experiments on zero-sum game solving, market equilibrium computation and image denoising demonstrate the effectiveness of the proposed schemes. In particular, the increasing averages consistently outperform the uniform averages in all test problems by orders of magnitude. When solving matrix and extensive-form games, increasing averages consistently outperform the last iterates as well. For matrix games, a first-order method equipped with increasing averaging outperforms the highly competitive CFR$^+$ algorithm.

Introduction

Consider saddle point problems (SPP) of the form

$$\min_{x\in\mathbb{X}} \max_{y\in\mathbb{Y}} \mathcal{L}(x,y) \tag{1}$$

where $\mathcal{L}$ is a general convex-concave function and $\mathbb{X}$, $\mathbb{Y}$ are Euclidean spaces. For any $(x,y)\in\mathbb{X}\times\mathbb{Y}$, denote its *saddle-point residual* (SPR) as

$$\epsilon_{\text{sad}}(x,y) = \max_{y'\in\mathbb{Y}} \mathcal{L}(x,y') - \min_{x'\in\mathbb{X}} \mathcal{L}(x',y).$$

Many problems in machine learning (Juditsky, Nemirovski et al. 2011; Chambolle and Pock 2016), image processing (Chambolle and Pock 2011, 2016) and game theory (Koller, Megiddo, and Von Stengel 1996; Kroer et al. 2018) can be formulated as (1). Many primal-dual first-order methods (FOMs) are suitable for these problems and have been proven efficient in practice. For example, Chambolle and Pock (2011) gives an algorithm for solving saddle-point problems involving bilinear and separable, nonsmooth terms and demonstrate its effectiveness in image denoising. Kroer, Farina, and Sandholm (2018) uses the Excessive Gap Technique (EGT) (Nesterov 2005), with a specific distance-generating function, to solve saddle-point formulation of zero-sum extensive-form games (EFG). Given the general convex-concave structure without strong convexity and smoothness assumptions, these algorithms only guarantee *ergodic* convergence rate, that is, $\epsilon_{\text{sad}}\left(\frac{1}{T}\sum_{t=1}^T x^t, \frac{1}{T}\sum_{t=1}^T y^t\right) = O(1/T)$. Meanwhile, numerically, (x^t, y^t) (the "last iterates") often converge much more rapidly (see, e.g., (Chambolle and Pock 2016, §7.2.2)). This observation motivates new averaging schemes that put more weight on later iterates rather than uniformly across all of them. Let (x^t, y^t), $t = 1, 2, \dots$ denote the iterates generated by a first-order method (more specifically, the iterates used in forming the convergent uniform averages; for certain algorithms, they are not necessarily denoted as (x^t, y^t); see, e.g., Theorem 2). Let w_t be positive, nondecreasing weights. We consider averages of the form

$$\bar{x}^T = \frac{1}{S_T}\sum_{t=1}^T w_t x^t, \quad \bar{y}^T = \frac{1}{S_T}\sum_{t=1}^T w_t y^t, \tag{2}$$

where $S_T = \sum_{t=1}^T w_t$. For example, $w_t = 1, t, t^2$, and t^3 give *uniform*, *linear*, *quadratic* and *cubic* averages, respectively. We refer to such choices of positive, nondecreasing w_t as *increasing iterate averaging schemes* (IIAS) and $\bar{x}^T, \bar{y}^T$ as *increasing averages*. In fact, in solving extensive-form games, the highly successful CFR$^+$ algorithm uses a form of linear averaging (Tammelin et al. 2015). Similar averaging techniques have also been used in other scenarios, such as algorithms for solving large-scale sequential games that achieve superhuman performance in poker (Bowling et al. 2015; Moravčík et al. 2017; Brown and Sandholm 2018), and efficient large-scale GAN training (Yazici et al. 2018). On the theory side, (Golowich et al. 2020) shows that, for unconstrained smooth saddle-point problems, the last iterates of Extragradient, a primal-dual FOM, converges slower than the averages in a strict sense. Through a unified analysis, (Davis and Yin 2016) shows $O(1/\sqrt{T})$ ergordic and last-iterate convergence rates of general splitting

schemes, which apply to the special case of the general standard form (4) with $f = 0$. In the context of (stochastic) convex minimization, theoretical guarantees and practical effectiveness of similar increasing averaging schemes have been studied (Lacoste-Julien, Schmidt, and Bach 2012; Shamir and Zhang 2013; Nesterov 2018). To the best of our knowledge, in solving saddle-point problems, no theoretical justification has been given for the practical speedup from last iterates as compared to uniform averages, as well as the potentially harder problem of explaining the speedup from increasing iterate averaging.

In this work, we show that for a wide class of FOMs, IIAS produces averages that converges at a rate $O(1/T)$ in terms of SPR. Algorithms compatible with IIAS include the (vanilla) primal-dual algorithm (PDA) (Chambolle and Pock 2011), its relaxed version (RPDA), inertial version (IPDA) (Chambolle and Pock 2016), and linesearch version (PDAL) (Malitsky and Pock 2018), as well as Mirror Descent (MD) (Nemirovski and Yudin 1983; Beck and Teboulle 2003; Ben-Tal and Nemirovski 2019) and Mirror Prox (MP) (Nemirovski 2004; Ben-Tal and Nemirovski 2019). For most of the algorithms, in order to preserve the convergence of $(\bar{x}^T, \bar{y}^T)$, it suffices to choose $w_t = t^q$ for some weight exponent $q \geq 0$, completely independent of the problem instance and the algorithm. For algorithms with inertial terms or linesearch subroutine, in order to ensure the theoretical convergence rate, w_t needs to satisfy additional inequalities, which makes w_t depend on previous iterations. Still, simple formulas suffice (e.g., Theorem 3 and 4). Finally, we emphasize that for all first-order methods considered here, IIAS does *not* alter the execution of the original algorithm. In other words, the performance boost is achieved without any extra computation or memory - we simply replace the uniform averages by increasingly weighted ones. Meanwhile, the averaging weights w_t, the sum of weights $\sum_{t=1}^{T} w_t$ and the averages $(\bar{x}^T, \bar{y}^T)$ can all be updated incrementally along the way.

Summary of contributions. First, we provide easily implementable IIAS for a variety of FOMs and establish their respective convergence properties. The high-level idea of the analysis can be summarized as follows. For each of the first-order methods, in the proof of the $O(1/T)$ rate of convergence, we identify the *critical inequality* being summed across all time steps to derive the final rate. We then take a weighted sum instead where the weights are the increasing averaging weights w_t. Then, through telescoping the summation, we bound the right hand side by $O\left(w_T/S_T\right)$, which is $O(1/T)$ (with an extra constant $(q+1)$ compared to the original ones under uniform averaging) as long as w_t grows *polynomially*, that is, $w_t > 0$, nondecreasing, and $\frac{w_{t+1}}{w_t} \leq \frac{(t+1)^q}{t^q}$ for all t, for some $q \geq 0$. Second, we perform extensive numerical experiments on various first-order methods and saddle-point point problems to demonstrate the consistent, strong performance gain of IIAS. Test problems include matrix games of different sizes and generative distributions, extensive-form games, Fisher market equilibrium computation and the TV-ℓ_1 image denoising model. As the results demonstrate, increasing averages consistently outperform uniform averages in all experiments by orders of magnitude. When solving matrix and extensive-form games, increasing averages consistently outperform the last iterates as well. For matrix games, PDA and RPDA equipped with IIAS also outperforms the highly competitive CFR$^+$ algorithm. For EFGs, RPDA under *static, theoretically safe* hyperparameters equipped with quadratic averaging outperforms EGT with unsafe, sophisticated, adaptive stepsizing.

Organization. We first present IIAS for the primal-dual algorithm (PDA) of (Chambolle and Pock 2016) and its analysis in detail. Then, we propose and analyze IIAS, with additional constraints on the weights, for relaxed, inertial and linesearch variants of PDA. Next, we discuss IIAS for Mirror Prox and Mirror Descent algorithms. Then, we present and discuss numerical experiment results.

Proofs of technical results and additional experiments can be found in an extended manuscript of this paper available at https://arxiv.org/abs/1903.10646.

The Primal-Dual Algorithm

Setup and notation. We follow the setup in (Chambolle and Pock 2016). Let $\mathbb{X}$ and $\mathbb{Y}$ be real Euclidean spaces with norms $\|\cdot\|_{\mathbb{X}}$ and $\|\cdot\|_{\mathbb{Y}}$, respectively. Denote the dual space of $\mathbb{X}$ as $\mathbb{X}^*$. Its corresponding dual norm, for any $x^* \in \mathbb{X}^*$, is defined as $\|x^*\|_{\mathbb{X},*} = \sup_{\|x\|=1} \langle x^*, x\rangle$. Define $\mathbb{Y}^*$ and $\|y^*\|_{\mathbb{Y},*}$ similarly. The subscripts on the norms are dropped when there is no ambiguity. Let $K : \mathbb{X} \to \mathbb{Y}^*$ be a bounded linear operator. and $K^* : \mathbb{Y} \to \mathbb{X}^*$ be its adjoint operator. The (operator) norm of K is defined as $\|K\| = \sup_{\|x\|\leq 1,\, \|y\|\leq 1} \langle Kx, y\rangle$. Let $\psi_{\mathbb{X}}$ and $\psi_{\mathbb{Y}}$ be 1-strongly convex (w.r.t. to their respective norms) smooth functions (known as distance-generating functions, DGF). Let $D_{\mathbb{X}}$ and $D_{\mathbb{Y}}$ be their respective Bregman divergence functions, that is, for $\mathbb{V} = \mathbb{X}, \mathbb{Y}$, $v, v' \in \mathbb{V}$,

$$D_{\mathbb{V}}(v', v) := \psi_{\mathbb{V}}(v') - \psi_{\mathbb{V}}(v) - \langle \nabla \psi_{\mathbb{V}}(v), v' - v\rangle. \quad (3)$$

Let f be a proper lower-semicontinuous (l.s.c.) convex function whose gradient ∇f is L_f-Lipschitz continuous w.r.t. $\|\cdot\|_{\mathbb{X}}$. Let g, h be proper l.s.c. convex functions whose proximal maps $\mathrm{Prox}_{\tau g}(x) = \arg\min_u \{\tau g(u) + D_{\mathbb{X}}(x, u)\}$ and $\mathrm{Prox}_{\sigma h^*}(y) = \arg\min_v \{\sigma h^*(v) + D_{\mathbb{Y}}(y, v)\}$, $\tau, \sigma > 0$ can be easily computed. In addition, assume $\mathrm{dom}\, g \subseteq \mathrm{dom}\, \psi_{\mathbb{X}}$ and $\mathrm{dom}\, h^* \subseteq \mathrm{dom}\, \psi_{\mathbb{Y}}$. Define the matrix $M_{\tau,\sigma} = \begin{bmatrix} \frac{1}{\tau} I & -K^* \\ -K & \frac{1}{\sigma} I \end{bmatrix}$, which is positive definite (semidefinite) as long as $\tau, \sigma > 0$ and $\tau\sigma L^2 < 1$ (≤ 1). With the above setup, consider the following SPP:

$$\min_{x\in\mathbb{X}} \max_{y\in\mathbb{Y}} \mathcal{L}(x, y) := \langle Kx, y\rangle + f(x) + g(x) - h^*(y). \quad (4)$$

For $(\bar{x}, \bar{y}) \in \mathbb{X}\times\mathbb{Y}$, $(\tilde{x}, \tilde{y}) \in \mathbb{X}\times\mathbb{Y}$, the generic primal-dual iteration $(\hat{x}, \hat{y}) = PD_{\tau,\sigma}(\bar{x}, \bar{y}, \tilde{x}, \tilde{y})$ is

$$\hat{x} = \arg\min_x \left\{ \begin{matrix} f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + g(x) \\ +\langle Kx, \tilde{y}\rangle + \frac{1}{\tau} D_{\mathbb{X}}(x, \bar{x}) \end{matrix} \right\},$$

$$\hat{y} = \arg\min_y \left\{ h^*(y) - \langle K\tilde{x}, y\rangle + \frac{1}{\sigma} D_{\mathbb{Y}}(y, \bar{y}) \right\}.$$

Here, the updates are *asymmetric*: in PD, y^{t+1} depends on x^{t+1}, which then depends on (x^t, y^t). Chambolle and

Algorithm 1 Nonlinear primal-dual algorithm (PDA)

Input: Initial iterate $(x^0, y^0) \in \mathbb{X} \times \mathbb{Y}$, stepsizes $\tau, \sigma > 0$, Bregman divergences $D_{\mathbb{X}}$ and $D_{\mathbb{Y}}$.
Iterations: For $t = 0, 1, 2, \ldots$, compute $(x^{t+1}, y^{t+1}) = PD_{\tau,\sigma}(x^t, y^t, 2x^{t+1} - x^t, y^t)$.

Pock (2016) proposes a primal-dual algorithm (PDA), which is listed here as Algorithm 1. Theorem 1 in their paper shows that the uniform averages converge at $O(1/T)$ in SPR. The proof relies on a lemma regarding the generic iteration $PD_{\tau,\sigma}$, which is restated below.

Lemma 1. *Given $\tau, \sigma > 0$, let $(\hat{x}, \hat{y}) = PD_{\tau,\sigma}(\bar{x}, \bar{y}, \tilde{x}, \tilde{y})$. For any $(x, y) \in \mathbb{X} \times \mathbb{Y}$, one has*
$\mathcal{L}(\hat{x}, y) - \mathcal{L}(x, \hat{y}) \le \frac{1}{\tau}\left(D_{\mathbb{X}}(x, \bar{x}) - D_{\mathbb{X}}(x, \hat{x}) - D_{\mathbb{X}}(\hat{x}, \bar{x})\right)$
$+ \frac{L_f}{2}\|\hat{x} - \bar{x}\|^2 + \frac{1}{\sigma}\left(D_{\mathbb{Y}}(y, \bar{y}) - D_{\mathbb{Y}}(y, \tilde{y}) - D_{\mathbb{Y}}(\hat{y}, \bar{y})\right)$
$+ \langle K(x - \hat{x}), \tilde{y} - \hat{y}\rangle - \langle K(\tilde{x} - \hat{x}), y - \hat{y}\rangle.$

Based on Lemma 1, we obtain the following extension of Theorem 1 in Chambolle and Pock (2016) regarding the convergence of IIAS for PDA.

Theorem 1. *For $t = 0, 1, 2, \ldots$, let $w_t = t^q$ for some $q \ge 0$ and (x^t, y^t), $t = 1, 2, \ldots$ generated by* PDA*, where stepsizes τ, σ are chosen such that, for all $x, x' \in \operatorname{dom} g$ and $y, y' \in \operatorname{dom} h^*$, it holds that*

$$\left(\frac{1}{\tau} - L_f\right) D_{\mathbb{X}}(x, x') + \frac{1}{\sigma} D_{\mathbb{Y}}(y, y') - \langle K(x - x'), y - y'\rangle \ge 0. \tag{5}$$

Let the averages $\bar{x}^T$, $\bar{y}^T$ be as in (2)*. Denote $A(x, y, x', y') = \frac{1}{\tau}D_{\mathbb{X}}(x, x') + \frac{1}{\sigma}D_{\mathbb{Y}}(y, y') - \langle K(x - x'), y - y'\rangle$. Let $\Omega = \sup_{x,x' \in \operatorname{dom} g,\, y,y' \in \operatorname{dom} h^*} A(x, y, x', y')$. Then, for any $T \ge 1$ and $(x, y) \in \mathbb{X} \times \mathbb{Y}$, one has*

$$\mathcal{L}(\bar{x}^T, y) - \mathcal{L}(x, \bar{y}^T) \le \frac{(q+1)\Omega}{T}.$$

Proof. As in the proof of Theorem 1 in (Chambolle and Pock 2016), Lemma 1 and PDA imply the following *critical inequality* regarding the iterates (x^t, y^t) and (x^{t+1}, y^{t+1}):

$$\begin{aligned}
&\mathcal{L}(x^{t+1}, y) - \mathcal{L}(x, y^{t+1}) \\
&\le A(x, y, x^t, y^t) - A(x, y, x^{t+1}, y^{t+1}) \\
&- \left(A(x^{t+1}, y^{t+1}, x^t, y^t) - \frac{L_f}{2}\|x^{t+1} - x^t\|^2\right) \\
&\le A(x, y, x^t, y^t) - A(x, y, x^{t+1}, y^{t+1})
\end{aligned} \tag{6}$$

where the last inequality is by (5). Multiplying (6) by w_{t+1} and summing up over $t = 0, 1, \ldots, T-1$ yield

$$\begin{aligned}
&\sum_{t=1}^{T} w_t \left(\mathcal{L}(x^t, y) - \mathcal{L}(x, y^t)\right) \\
&\le \sum_{t=1}^{T} w_t \left(A(x, y, x^{t-1}, y^{t-1}) - A(x, y, x^t, y^t)\right) \\
&= \sum_{t=1}^{T} (w_t - w_{t-1}) A(x, y, x^{t-1}, y^{t-1}) \le \Omega w_T.
\end{aligned}$$

The convex-concave structure of $\mathcal{L}$ implies

$$\mathcal{L}(\bar{x}^T, y) - \mathcal{L}(x, \bar{y}^T) \le \frac{1}{S_T} \sum_{t=1}^{T} w_t \left(\mathcal{L}(x^t, y) - \mathcal{L}(x, y^t)\right).$$

Furthermore, $S_T \ge \int_0^T x^q \, dx = \frac{T^{q+1}}{q+1}$. Combining the above inequalities yields the claim. □

The key proof idea is to take a weighted sum of the critical inequalities at all t and bound the right hand side via telescoping summation. This is also used in subsequent analysis. Here, the constant in the bound increases with q. Nonetheless, numerical experiments show that a nonzero, small value of q always yields significant speedup. We also remark that our result is no more restrictive than (Chambolle and Pock 2016) in terms of the domain boundedness assumption. In fact, Ω often takes on a small, finite value in many realistic scenarios. For example, for a two-person zero-sum game, g and h^* are indicator functions of the strategy spaces X, Y (which are simplexes for a matrix game), which are bounded polytopes with small diameters. The linear map K corresponds to the payoffs, which can be normalized to $\|K\| = 1$ w.l.o.g. Finally, note that Theorem 1 and subsequent theorems present *point-wise* inequalities, that is, a uniform bound on $\mathcal{L}(\bar{x}^T, y) - \mathcal{L}(x, \bar{y}^T)$ independent of (x, y). A bound on the saddle-point residual $\epsilon_{\mathrm{sad}}(\bar{x}^T, \bar{y}^T)$ can be easily obtained by taking $\min_{x \in \mathbb{X}} \max_{y \in \mathbb{Y}}$ on both sides.

Extensions of the Primal-Dual Algorithm

Similar IIAS can be applied to the relaxed and inertial versions of PDA, as described in (Chambolle and Pock 2016), as well as a nontrivial extension with linesearch (PDAL) (Malitsky and Pock 2018). We inherit the notation of the previous section and further assume $\|\cdot\|_{\mathbb{X}}$, $\|\cdot\|_{\mathbb{Y}}$ are Euclidean 2-norms and $D_{\mathbb{X}}(x, x') = \frac{1}{2}\|x - x'\|_2^2$, $D_{\mathbb{Y}}(y, y') = \frac{1}{2}\|y - y'\|_2^2$.

Relaxed primal dual algorithm. The relaxed primal-dual algorithm (RPDA) in (Chambolle and Pock 2016) is listed here as Algorithm 2. Similar to Theorem 1, we have the following convergence guarantee for RPDA. Here, denote $\Omega = \sup_{z,z' \in \mathbb{X} \times \mathbb{Y}} \frac{1}{2}\|z - z'\|_{M_{\tau,\sigma}}$, where $M_{\tau,\sigma}$ is the positive semidefinite matrix defined in the previous section.

Theorem 2. *Let $\tau, \sigma > 0$ and $0 \le \rho_t \le \rho_{t+1} \le \rho$, where $\rho \in (0, 2)$ satisfy*[1] *$\left(\frac{1}{\tau} - \frac{L_f}{2-\rho}\right)\frac{1}{\sigma} \ge \|K\|_2^2$. Let (ξ^t, η^t), $t = 1, 2, \ldots$ be generated by* RPDA *and $\bar{x}^T = \frac{1}{S_T}\sum_{t=1}^{T} w_t \xi^t$, $\bar{y}^T = \frac{1}{S_T}\sum_{t=1}^{T} w_t \eta^t$. For any $z = (x, y) \in \mathbb{X} \times \mathbb{Y}$, one has*

$$\mathcal{L}(\bar{x}^T, y) - \mathcal{L}(x, \bar{y}^T) \le \frac{(q+1)\Omega}{\rho_0 T}.$$

[1] Theorem 2 in (Chambolle and Pock 2016) requires strict inequality. In fact, a non-strict one suffices. The strict inequality is required for sequence convergence of the last iterates (Chambolle and Pock 2016, Remark 3). The same holds for Theorem 3. However, there we assume strict inequality as in (Chambolle and Pock 2016, Theorem 3), as it conveniently ensures eventual polynomial growth of the weights.

Algorithm 2 Relaxed primal-dual algorithm (RPDA)

Input: Initial iterates $z^0 = (x^0, y^0) \in \mathbb{X} \times \mathbb{Y}$, $\tau, \sigma > 0$, relaxation parameters ρ_t, Euclidean $D_{\mathbb{X}}$ and $D_{\mathbb{Y}}$.
Set: $z^t = (x^t, y^t)$ and $\zeta^t = (\xi^t, \eta^t)$.
Iterations: For $t = 0, 1, 2, \ldots$, compute
$(\xi^{t+1}, \eta^{t+1}) = PD_{\tau,\sigma}(x^t, y^t, 2\xi^{t+1} - x^t, y^t)$,
$z^{t+1} = (1 - \rho_n) z^t + \rho_n \zeta^{t+1}$.

Algorithm 3 Inertial primal-dual algorithm (IPDA)

Input: $(x^{-1}, y^{-1}) = (x^0, y^0) \in \mathbb{X} \times \mathbb{Y}$, $\tau, \sigma > 0$, inertial parameters α_t, Euclidean $D_{\mathbb{X}}$ and $D_{\mathbb{Y}}$.
Set: $z^t = (x^t, y^t)$ and $\zeta^t = (\xi^t, \eta^t)$.
Iterations: For $t = 0, 1, 2, \ldots$, compute
$\zeta^t = z^t + \alpha_t (z^t - z^{t-1})$,
$z^{t+1} = PD_{\tau,\sigma}(\xi^t, \eta^t, 2x^{t+1} - \xi^t, \eta^t)$.

Inertial primal-dual algorithm. Another useful variant of PDA is the inertial primal-dual algorithm (IPDA) (Chambolle and Pock 2016), listed here as Algorithm 3. In contrast to PDA and RPDA, in order to preserve the rate of convergence of IPDA, w_t needs to satisfy a few additional inequalities arising from the analysis of the telescoping sum. The choice of w_t and the convergence guarantee are summarized below. The proof is based on the same principles but is much more involved. Note that the strict inequality involving $\tau, \sigma, \|K\|$ makes $b_t \geq b^* > 1$, which ensures eventual polynomial growth of w_t (and an asymptotical $O(1/T)$ rate).

Theorem 3. *Let $\alpha > 0$ be such that $\left(\frac{1}{\tau} - \frac{(1+\alpha)^2}{1-3\alpha} L_f\right) \frac{1}{\sigma} > \|K\|_2^2$ and $0 \leq \alpha_t \leq \alpha_{t+1} \leq \alpha < \frac{1}{3}$ for all t. Let (x^t, y^t) be generated by* IPDA. *Let w_t be chosen as follows:*

$$w_0 = 0,\ w_1 = 1,\ w_{t+1} = w_t \cdot \min\left\{b_t, \tfrac{(t+1)^q}{t^q}\right\},\ t \geq 2,$$

where

$$b_t = \min\left\{\frac{1-\alpha_{t-1}}{\alpha_t}, \frac{r(1-\alpha_{t-1}) - (1+\alpha_{t-1})}{\alpha_t(1+2r+\alpha_t)}\right\},$$
$$r = \frac{\frac{1}{\tau} - \sigma\|K\|^2}{L_f}.$$

Then, it holds that $w_t \leq w_{t+1}$ and $b_t \geq b^*$ for some $b^* > 1$ for all t. Furthermore, let S_T, $\bar{x}^T$ and $\bar{y}^T$ be as in (2)*. For any $z = (x, y) \in \mathbb{X} \times \mathbb{Y}$, one has*

$$\begin{aligned}
&\mathcal{L}(x, \bar{y}^T) - \mathcal{L}(\bar{x}^T, y) \\
&\leq \frac{(1-\alpha_0) w_1 A_0 + \Omega\left[(1-\alpha_{T-1}) w_T + \alpha_1 w_2 - w_1\right]}{S_T} \\
&\leq \frac{(q+1)(2-\alpha_0)\Omega}{T},
\end{aligned}$$

where $A_0 := \|z - z^0\|^2_{M_{\tau,\sigma}}$ and Ω is as in Theorem 2.

Primal-dual algorithm with linesearch. Recently, Malitsky and Pock (2018) proposed a primal-dual algorithm with linesearch (PDAL), which is listed as Algorithm 4 here. To align the primal and dual iterates in the increasing averaging version of PDAL, y^t here correspond to y^{t+1} in the original paper, $t = 0, 1, 2, \ldots$ Assume $\|\cdot\|_{\mathbb{X}}$, $\|\cdot\|_{\mathbb{Y}}$, $D_{\mathbb{X}}$, $D_{\mathbb{Y}}$ are all Euclidean as in Theorem 2 and $f = 0$ in (4). Theorem 3.5 in (Malitsky and Pock 2018) establishes the ergodic convergence rate of PDAL. We show the following theorem on IIAS for PDAL. Note that the averaging involves not only the weights w_t but also the stepsizes τ_t. Similar is true for the subsequent Mirror-type algorithms. Here, denote $\Omega_{\mathbb{X}} = \sup_{x,x' \in \operatorname{dom} g} \frac{1}{2}\|x - x'\|^2$, $\Omega_{\mathbb{Y}} = \sup_{y,y' \in \operatorname{dom} h^*} \frac{1}{2}\|y - y'\|^2$.

Algorithm 4 PDAL: Primal-dual algorithm with linesearch

Input: $(x^0, y^0) \in \mathbb{X} \times \mathbb{Y}$, initial stepsize $\tau_0 > 0$, backtracking discount factor μ, backtracking break tolerance δ, primal-dual ratio $\beta > 0$, Euclidean $D_{\mathbb{X}}$ and $D_{\mathbb{Y}}$.
Set: stepsize growth factor $\theta_0 = 1$.
Iterations: For $t = 0, 1, 2, \ldots$, compute
$x^{t+1} = \operatorname{Prox}_{\tau_t g}(x^t - \tau_t K^* y^t)$.

(a) Choose $\tau_{t+1} \in [\tau_t, \tau_t\sqrt{1+\theta_t}]$, compute $\theta_{t+1} = \frac{\tau_{t+1}}{\tau_t}$, $\tilde{x}^{t+1} = x^{t+1} + \theta_{t+1}(x^{t+1} - x^t)$ and $y^{t+1} = \operatorname{Prox}_{\beta\tau_{t+1} h^*}(y^t + \beta\tau_{t+1} K \tilde{x}^{t+1})$.
(b) Break if $\sqrt{\beta}\tau_{t+1}\|K^* y^{t+1} - K^* y^t\| \leq \delta\|y^{t+1} - y^t\|$. Otherwise, set $\tau_{t+1} \leftarrow \tau_{t+1}\mu$ and go to ((a)).

Theorem 4. *Let (x^t, y^t) be generated by* PDAL *on problem* (4) *with $f = 0$. Let w_t be as follows:*

$$w_0 = 0,\ w_1 = 1,\ w_{t+1} = w_t \cdot \min\left\{\tfrac{1+\theta_t}{\theta_{t+1}}, \tfrac{(t+1)^q}{t^q}\right\},\ t \geq 1.$$

Let $S_T = \sum_{t=1}^T w_t \tau_t$ and

$$\bar{x}^T = \frac{w_1\theta_1\tau_1 x_0 + \sum_{t=1}^N w_t\tau_t\tilde{x}^t}{w_t\tau_1\theta_1 + S_T},\ \bar{y}^T = \frac{\sum_{t=1}^T w_t\tau_t y^t}{S_T}.$$

For any $(x, y) \in \mathbb{X} \times \mathbb{Y}$, $T \geq 2$, we have

$$\begin{aligned}
\mathcal{L}(\bar{x}^T, y) - \mathcal{L}(x, \bar{y}^T) &\leq \frac{w_T\left(\Omega_{\mathbb{X}} + \frac{1}{\beta}\Omega_{\mathbb{Y}}\right) + w_1\tau_1\theta_1 P_0}{S_T} \\
&\leq \frac{(q+1)\left(\Omega_{\mathbb{X}} + \frac{1}{\beta}\Omega_{\mathbb{Y}} + \tau_1\theta_1 P_0\right)}{T},
\end{aligned}$$

where $P_0 = g(x^0) - g(x) + \langle K(x^0 - x), y\rangle$.

Mirror-Type Algorithms

Next, we consider Mirror Descent (MD) (Nemirovski and Yudin 1983; Beck and Teboulle 2003) and Mirror Prox (MP) (Nemirovski 2004), which require a setup different from that of the previous sections. Specifically, assume that $\mathbb{X}$ and $\mathbb{Y}$ are Euclidean spaces with norms $\|\cdot\|_{\mathbb{X}}$ and $\|\cdot\|_{\mathbb{Y}}$. Let $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$ and $\|\cdot\|$ be be any norm (with dual norm $\|\cdot\|_*$) on $\mathbb{Z}$. Let $\mathcal{X} \subseteq \mathbb{X}$ and $\mathcal{Y} \subseteq \mathbb{Y}$ be closed convex sets. Let $\phi : \mathcal{Z} = \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a convex-concave cost function, that is, $\phi(\cdot, y)$ is convex for any $y \in \mathcal{Y}$ and $\phi(x, \cdot)$ is concave for any $x \in \mathcal{X}$. The saddle-point problem that Mirror-type algorithms solve is

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

Algorithm 5 Mirror Descent (MD) and Mirror Prox (MP)

Input: initial iterate $z^0 = (x^0, y^0) \in \mathcal{Z}$, stepsizes τ_t, Bregman function $D_{\mathbb{Z}}$.

Iterations: For $t = 0, 1, 2, \ldots$, compute

$$\text{MD: } z^{t+1} = \text{Prox}^{\mathcal{Z}}_{\langle \tau_t F(z^t), \cdot \rangle}(z^t),$$

$$\text{MP: } \begin{cases} \tilde{z}^t = \text{Prox}^{\mathcal{Z}}_{\langle \tau_t F(z^t), \cdot \rangle}(z^t), \\ z^{t+1} = \text{Prox}^{\mathcal{Z}}_{\langle \tau_t F(\tilde{z}^t), \cdot \rangle}(z^t). \end{cases}$$

Let $\psi_{\mathbb{Z}} : \mathbb{Z} \to \mathbb{R}$ be a 1-strongly convex (w.r.t. $\|\cdot\|_{\mathbb{Z}}$) smooth function, that is, a DGF on $\mathbb{Z}$. For example, given DGF $\psi_{\mathbb{X}}(x)$ and $\psi_{\mathbb{Y}}$ for $\mathbb{X}$ and $\mathbb{Y}$, the DGF $\psi_{\mathbb{Z}}(z) := \psi_{\mathbb{X}}(x) + \psi_{\mathbb{Y}}(y)$, $z = (x, y)$, is 1-strongly convex w.r.t. the norm $\|z\| := \sqrt{\|x\|^2_{\mathbb{X}} + \|y\|^2_{\mathbb{Y}}}$. Let the Bregman divergence function $D_{\mathbb{Z}}$ be defined as (3) with $\mathbb{V} = \mathbb{Z}$. Let $\Omega = \sup_{z \in \mathcal{Z}} D_{\mathbb{Z}}(z)$.[2] The "gradient vector field" associated with ϕ is $F(z) = \left(\frac{\partial}{\partial_x}\phi(z), -\frac{\partial}{\partial_y}\phi(z) \right)$. We assume F is bounded and L-Lipschitz continuous on Z, that is, $M_F = \sup_{z \in \mathcal{Z}} \|F(z)\|_* < \infty$ and $\|F(x) - F(z')\|_* \leq L\|z - z'\|$ for any $z, z' \in \mathcal{Z}$. For $z, \xi \in \mathbb{Z}$, define the (constrained) proximal mapping (of a linear function) as $\text{Prox}^{\mathcal{Z}}_{\langle \xi, \cdot \rangle}(z) := \arg\min_{w \in \mathcal{Z}} \{\langle \xi, w \rangle + D_{\mathbb{Z}}(w, z)\}$. The algorithms are listed together here as Algorithm 5. Similarly, we show that IIAS applied to MD and MP preserves their respective convergence rate, where the averaging weights involve the stepsizes τ_t.[3] In addition, (Ben-Tal and Nemirovski 2019, Theorem 5.3.5) show that a similar bound holds for MD with increasing averaging.

Theorem 5. *Let $\tilde{z}^t = (\tilde{x}^t, \tilde{y}^t)$ be generated by* MP *and $\delta_t := \tau_t \langle F(\tilde{z}^t), \tilde{z}_t - z_{t+1} \rangle - D_{\mathbb{Z}}(z_{t+1}, z_t)$. Let $w_t \geq 0$ be nondecreasing weights and*

$$S_T = \sum_{t=1}^{T} w_t \tau_t, \ \bar{z}^T = (\bar{x}^T, \bar{y}^T) = \frac{\sum_{t=1}^{T} w_t \tau_t \tilde{z}^t}{S_T}.$$

Then, for any $T \geq 1$ and any $(x, y) \in \mathbb{X} \times \mathbb{Y}$, it holds that

$$\phi(\bar{x}^T, y) - \phi(x, \bar{y}^T) \leq \frac{w_T \Omega + \sum_{t=1}^{T} w_t \delta_t}{S_T}.$$

In particular, constant stepsizes $\tau_t = \frac{1}{L}$ and $w_t = t^q$, $q \geq 0$ ensure $\delta_t \leq 0$ and

$$\phi(\bar{x}^T, y) - \phi(x, \bar{y}^T) \leq \frac{(q+1)L\Omega}{T}.$$

[2]The constant Ω here corresponds to Θ in (Ben-Tal and Nemirovski 2019), defined on page 356.

[3]During the course of this research, we note that IIAS for MP has been analyzed in the growing lecture notes of Ben-Tal and Nemirovski (2019). Specifically, Theorem 5.6.2 in (Ben-Tal and Nemirovski 2019) presents the weighted averaging version of MP, although without no explicit weight formulaes.

Numerical Experiments

We demonstrate the numerical effectiveness of IIAS in solving zero-sum games, computing Fisher market equilibria and image-denoising. For matrix games and EFG, our best FOMs with IIAS are compared against state-of-the-art methods for equilibrium computation.

Matrix games. Here, we briefly describe the experiment setup. A matrix game can be formulated as a bilinear SPP

$$\min_{x \in \Delta^{n_1}} \max_{y \in \Delta^{n_2}} \langle x, Ay \rangle,$$

where $\Delta^d := \{x \in \mathbb{R}^d \mid x^\top \mathbf{e} = 1, \, x \geq 0\}$ is the unit simplex in $\mathbb{R}^d$. To use a PDA-type algorithm to solve a matrix game, let g and h^* be the indicator functions of Δ^{n_1} and Δ^{n_2}, respectively, and $K = A^\top$. For MP, we take $\mathcal{X} = \Delta^{n_1}$ and $\mathcal{Y} = \Delta^{n_2}$ and $\phi(x, y) = \langle x, Ay \rangle$. We also try a linesearch variant of MP, referred to as MPL, that performs linesearch as described in (Ben-Tal and Nemirovski 2019, pp. 443). For all algorithms, we use their "default" stepsizes and Euclidean DGF. We generate matrix games of different dimensions with i.i.d. entries and solve them using all six algorithms. For each matrix game and each algorithm, we perform $T = 2000$ iterations, take increasing averages of the iterates and compute their saddle-point residuals. Residuals are normalized to make them commensurable in magnitude. The above is repeated 50 times. We plot the averages and standard errors (which have small magnitudes and are nearly invisible) of the normalized residuals along sample paths of each setup-algorithm-averaging combination. Figure 1 displays the plots. As can be seen, IIAS leads to significant performance improvement for all algorithms across all experiments, both against uniform averages as expected, but, perhaps surprisingly, also against the last iterates. The plots for IPDA and PDAL can be found in the full arXiv version. Next, we compare PDA, RPDA, and CFR$^+$ on the same set of random matrix games as well as on a 2×2 matrix game with payoff matrix $A = \begin{bmatrix} 5 & -1 \\ 0 & 1 \end{bmatrix}$ (Farina, Kroer, and Sandholm 2019). Figure 2 displays the results for these experiments: the upper plot is for the random matrix game experiments and displays averaged, normalized residuals and standard deviations of the 3 settings, similar to Figure 1; the lower plot is for the 2×2 matrix game and displays the (unnormalized) residual values. Clearly, PDA and RPDA outperform CFR$^+$ in both settings. Moreover, for the 2×2 matrix game, the last iterates converge rapidly, suggesting the use of a large weight exponent q. As the lower subplot in Figure 2 displays, using $q = 10$ can even outperform the last iterates. We stress that we test $q = 10$ mostly for experimental curiosity. In general, using $q = 1, 2$ yields significant speedup. Furthermore, even a large q does not lead to numerical issues in any experiment when the averages are incrementally updated.

Extensive-form games. An EFG can be written as a bilinear saddle-point problem (BLSP) $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \langle x, Ay \rangle$ where $\mathcal{X} \subseteq \mathbb{R}^{n_1}$ and $\mathcal{Y} \subseteq \mathbb{R}^{n_2}$ are polytopes encoding players' strategy spaces, known as *treeplexes* (Hoda et al. 2010). We perform IIAS with uniform, linear, quadratic and cubic averaging for all first-order methods on two classic

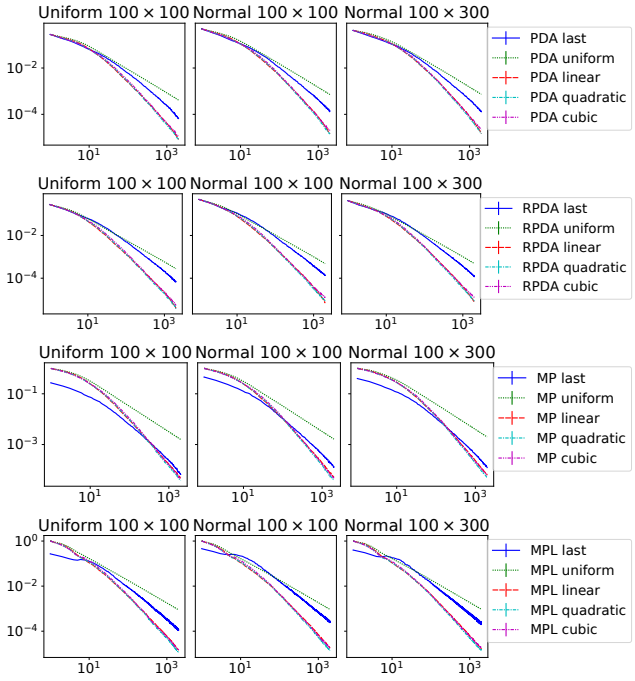

Figure 1: FOM with IIAS, matrix games, normalized SPR v.s. number of gradient computations

EFG benchmark instances Kuhn and Leduc poker (see, e.g., (Kroer et al. 2018)). The choices of algorithm hyperparameters are completely analogous to those in solving matrix games. As Figure 3 shows, increasing averages outperform uniform averages for all algorithms in both games. For all algorithms except MP, increasing averages also outperform the last iterates. For both games, we also compare RPDA with quadratic and $q = 10$ averaging, CFR$^+$ (Tammelin et al. 2015) and EGT with the dilated entropy DGF (Kroer et al. 2018). We plot the saddle-point residuals against the number of *gradient computations* $x \mapsto A^\top x$ and $y \mapsto Ay$, since EGT uses linesearch and may require more than one gradient computations in each iteration. Here, since we are interested in the regime where gradient computations dominate the overall computation cost, we assume computing the proximal mappings under different DGFs takes much less time in comparison. As Figure 4 shows, RPDA with quadratic and $q = 10$ averaging significantly outperforms CFR$^+$ and EGT on Kuhn but are not as fast as CFR$^+$ on Leduc. In addition, even using theoretically safe, highly conservative stepsizes, RPDA with quadratic averaging outperforms the EGT implementation, which employs sophisticated, adaptive stepsizing heuristics.

Fisher market equilibrium. In a *Fisher market* of n buyers, each buyer i has a *valuation* $v_i \in \mathbb{R}^m_+$ over m goods. An *allocation* $x_i \in \mathbb{R}^m_+$ gives a utility of $v_i^\top x_i$ to buyer i. Each buyer i has budget $B_i > 0$ and each good j has supply $s_j > 0$. A *market equilibrium* is a set of prices $p \in \mathbb{R}^m_+$ for the goods and an (aggregate) *allocation* $x = [x_1, \ldots, x_n]$ such that $x_i \in \arg\max \{v_i^\top x_i \mid p^\top x_i \le B_i\}$ for all buyers (i.e.,

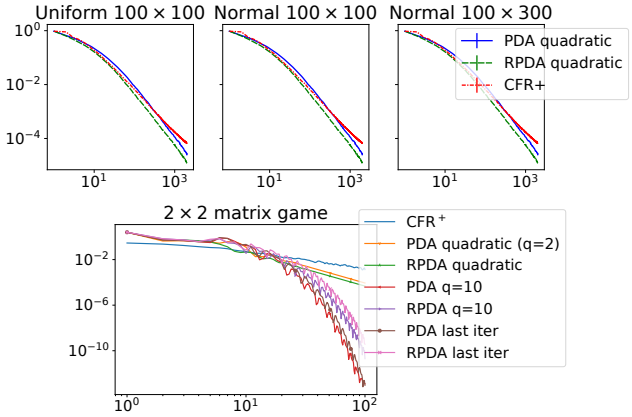

Figure 2: FOM with IIAS v.s. CFR$^+$, matrix games, normalized SPR v.s. number of gradient computations

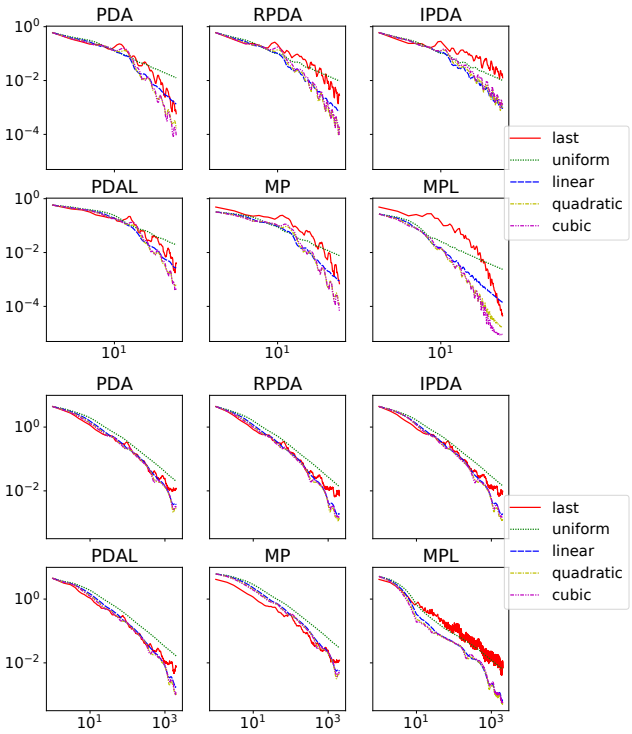

Figure 3: FOM with IIAS on Kuhn (upper 6 plots) and Leduc (lower 6 plots), SPR v.s. number of gradient computations

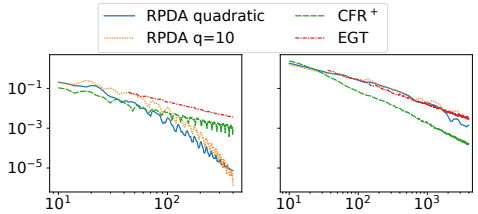

Figure 4: RPDA v.s. EGT v.s. CFR$^+$ on Kuhn (left) and Leduc (right), SPR v.s. number of gradient computations

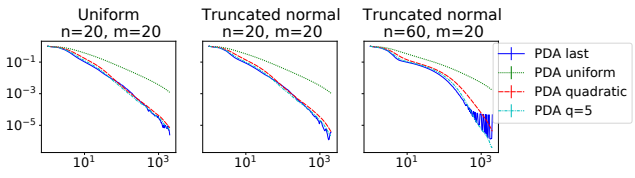

Figure 5: PDA with IIAS for Fisher market equilibria, normalized SPR v.s. number of iterations

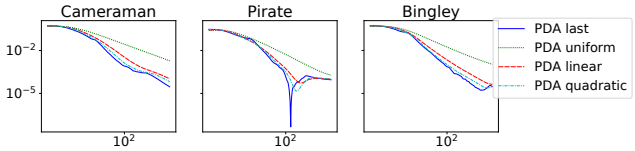

Figure 6: PDA with IIAS for TV-ℓ_1 minimization, normalized loss v.s. number of iterations

buyer are optimal given their budgets) i and $\sum_i x_{ij} = s_j$ for all items j with $p_j > 0$ (i.e., all valuable items are sold). It is well known that market equilibria are captured by the solutions of the *Eisenberg-Gale convex program* (Jain and Vazirani 2010, Eq. (1)). For more details, see (Eisenberg and Gale 1959; Eisenberg 1961; Jain and Vazirani 2010). This convex program can be formulated as a saddle-point problem (see, e.g., (Kroer et al. 2019)):

$$\min_{x \geq 0} \max_{p:\, 0 \leq p_i \leq \|B\|_1} \left[-\sum_i B_i \log v_i^\top x_i + \sum_i p^\top x_i - p^\top s \right].$$

We generate random instances of different sizes, solve them using PDA and compute the saddle-point residuals of the last iterates and various averages. Repeat each experiment 50 times, normalize residuals and compute mean and standard deviations similar to the procedures in matrix games. Figure 5 displays the normalized residuals. The standard deviations have small magnitudes and thus become invisible. As the plots shows, increasing averages can converge as fast as, and sometimes even more rapidly ($q = 5$ averaging in the rightmost subplot) than the last iterates.

Image denoising via TV-ℓ_1 minimization. The Total Variation (TV)-ℓ_1 model is a means for image denoising through convex optimization (Chambolle and Pock 2011). We use the saddle-point formulation of the convex optimization problem (Chambolle and Pock 2011, pp. 132). Let $\mathbb{X} = \mathbb{R}^{m \times n}$ be the image domain. Let div denote the divergence operator, that is, is the negative adjoint of the gradient operator $\nabla : \mathbb{X} \to \mathbb{X}$. Let $\mathbb{Y} = X \times X = \mathbb{R}^{m,n,2}$ be the set of discrete finite differences and

$$P = \left\{ p \in \mathbb{Y} \mid (p_{ij}^1)^2 + (p_{ij}^2)^2 \leq 1 \right\}$$

be the point-wise unit ℓ_2-ball. The saddle-point formulation of the TV-ℓ_1 model is

$$\min_{u \in \mathcal{X}} \max_{p \in \mathcal{Y}} -\langle u, \operatorname{div} p \rangle + \lambda \|u - g\|_1 - \delta_P(p),$$

where $\lambda > 0$ is the regularization strength hyperparameter. Following Chambolle and Pock (2011), to align it with

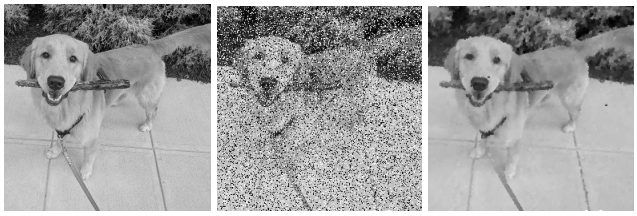

Figure 7: Original (left), corrupted (middle) and reconstructed images (right) of Bingley

(4), choose $f = 0$, $g(u) = \lambda \|u - g\|_1$ with $\lambda = 1.5$ and $h^*(p) = \delta_P(p)$; in this way, the proximal mappings yield closed-form formulas. See (Chambolle and Pock 2011, pp. 135-156) for more details on the saddle-point problem setup. We add salt-and-pepper noise to three 256×256 gray-scale images to obtain corrupted inputs and use the TV-ℓ_1 minimization procedure for reconstruction. To solve the resulting saddle-point problems, we use PDA with default, static hyperparameters used in (Chambolle and Pock 2011) and run for $T = 1000$ iterations. We compute the values of the original primal TV-ℓ_1 loss values of the last iterates and the various increasing averages. The loss values are normalized similarly and are displayed in Figure 6. Here, the last iterates converge fast, while linear and quadratic averages still perform nearly as well, and are far ahead of uniform averages. Figure 7 displays the original, corrupted, and reconstructed images of Bingley. The reconstruction is based on the quadratic averages of the PDA iterates at $T = 1000$.

Conclusion

We proposed increasing iterate averaging schemes for various first-order methods, provided simple, implementable choices of averaging weights and established convergence properties of the new iterate averages. Extensive numerical experiments on various saddle-point problems demonstrated their ability to accelerate numerical convergence by orders of magnitude without modifying the original algorithm or incurring extra computation. We reiterate that the algorithms are run unaltered with untuned, theoretically safe hyperparameters, while the averaging weights are chosen simply based on their respective convergence theorems. Even so, IIAS can bring the algorithms close to, and sometimes make them beat, other carefully engineered and tuned approaches.

Ethics Statement

Since our work is primarily algorithmic and works for a broad class of convex optimization problems, we do not see any direct ethical impacts of our work. Our work leads to substantial practical speed-up on all classes of problems we have tried it on so far. Thus, impact could arise indirectly by enabling larger scale on certain problems. For example, an increasing iterate averaging scheme may help efficient computation of Nash equilibria of zero-sum games, as shown in this work. Methods for training generative adversarial networks can potentially benefit from these schemes as well.

However, the ethical and societal implications of individual applications are beyond the scope of our paper.

References

Beck, A.; and Teboulle, M. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters* 31(3): 167–175.

Ben-Tal, A.; and Nemirovski, A. 2019. Lectures on modern convex optimization. *Online version: http://www2. isye. gatech. edu/˜ nemirovs/Lect_ ModConvOpt* .

Bowling, M.; Burch, N.; Johanson, M.; and Tammelin, O. 2015. Heads-up limit hold'em poker is solved. *Science* 347(6218): 145–149.

Brown, N.; and Sandholm, T. 2018. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science* 359(6374): 418–424.

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

Chambolle, A.; and Pock, T. 2016. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming* 159(1-2): 253–287.

Davis, D.; and Yin, W. 2016. Convergence rate analysis of several splitting schemes. In *Splitting methods in communication, imaging, science, and engineering*, 115–163. Springer.

Eisenberg, E. 1961. Aggregation of utility functions. *Management Science* 7(4): 337–350.

Eisenberg, E.; and Gale, D. 1959. Consensus of subjective probabilities: The pari-mutuel method. *The Annals of Mathematical Statistics* 30(1): 165–168.

Farina, G.; Kroer, C.; and Sandholm, T. 2019. Optimistic Regret Minimization for Extensive-Form Games via Dilated Distance-Generating Functions. In *Advances in Neural Information Processing Systems*, 5222–5232.

Golowich, N.; Pattathil, S.; Daskalakis, C.; and Ozdaglar, A. 2020. Last Iterate is Slower than Averaged Iterate in Smooth Convex-Concave Saddle Point Problems. *arXiv preprint arXiv:2002.00057* .

Hoda, S.; Gilpin, A.; Pena, J.; and Sandholm, T. 2010. Smoothing techniques for computing Nash equilibria of sequential games. *Mathematics of Operations Research* 35(2): 494–512.

Jain, K.; and Vazirani, V. V. 2010. Eisenberg–Gale markets: Algorithms and game-theoretic properties. *Games and Economic Behavior* 70(1): 84–106.

Juditsky, A.; Nemirovski, A.; et al. 2011. First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. *Optimization for Machine Learning* 121–148.

Koller, D.; Megiddo, N.; and Von Stengel, B. 1996. Efficient computation of equilibria for extensive two-person games. *Games and economic behavior* 14(2): 247–259.

Kroer, C.; Farina, G.; and Sandholm, T. 2018. Solving large sequential games with the excessive gap technique. In *Advances in Neural Information Processing Systems*, 864–874.

Kroer, C.; Peysakhovich, A.; Sodomka, E.; and Stier-Moses, N. E. 2019. Computing large market equilibria using abstractions. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, 745–746.

Kroer, C.; Waugh, K.; Kılınç-Karzan, F.; and Sandholm, T. 2018. Faster algorithms for extensive-form game solving via improved smoothing functions. *Mathematical Programming* 1–33.

Lacoste-Julien, S.; Schmidt, M.; and Bach, F. 2012. A simpler approach to obtaining an O (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002* .

Malitsky, Y.; and Pock, T. 2018. A first-order primal-dual algorithm with linesearch. *SIAM Journal on Optimization* 28(1): 411–432.

Moravčík, M.; Schmid, M.; Burch, N.; Lisỳ, V.; Morrill, D.; Bard, N.; Davis, T.; Waugh, K.; Johanson, M.; and Bowling, M. 2017. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* 356(6337): 508–513.

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

Nemirovski, A.; and Yudin, D. B. 1983. Problem complexity and method efficiency in optimization. .

Nesterov, Y. 2005. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization* 16(1): 235–249.

Nesterov, Y. 2018. Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming* 171(1-2): 311–330.

Shamir, O.; and Zhang, T. 2013. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, 71–79.

Tammelin, O.; Burch, N.; Johanson, M.; and Bowling, M. 2015. Solving heads-up limit Texas Hold'em. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Yazici, Y.; Foo, C.-S.; Winkler, S.; Yap, K.-H.; Piliouras, G.; and Chandrasekhar, V. 2018. The unusual effectiveness of averaging in GAN training. *arXiv preprint arXiv:1806.04498* .