Nonlinear Acceleration of Primal-Dual Algorithms

Raghu Bollapragada
Northwestern University

Damien Scieur
Princeton University

Alexandre d'Aspremont
CNRS & École Normale Supérieure

Abstract

We study convergence acceleration schemes for multi-step optimization algorithms. The extrapolated solution is written as a nonlinear average of the iterates produced by the original optimization algorithm. Our analysis of Regularized Nonlinear Acceleration, aka Anderson acceleration, does not need the underlying fixed-point operator to be symmetric, hence handles e.g. algorithms with momentum terms such as Nesterov's accelerated method, or primal-dual methods such as Chambolle-Pock. The weights are computed via a simple linear system and we analyze performance in both online and offline modes. We use Crouzeix's conjecture to show that acceleration is controlled by the solution of a Chebyshev problem on the numerical range of a non-symmetric operator modelling the behavior of iterates near the optimum. Numerical experiments are detailed on image processing and logistic regression problems.

1 Introduction

Extrapolation techniques, such as Aitken's Δ^2 or Wynn's ε-algorithm, provide an improved estimate of the limit of a sequence using its last few iterates, and we refer the reader to (Brezinski and Zaglia, 2013) for a complete survey. These methods have been extended to vector sequences, where they are known as e.g. Anderson acceleration (Walker and Ni, 2011), minimal polynomial extrapolation (Cabay and Jackson, 1976) or reduced rank extrapolation (Eddy, 1979).

Classical optimization algorithms typically retain only the last iterate or the average (Polyak and Juditsky, 1992) of iterates as their best estimate of the optimum, throwing away all the information contained in the converging sequence. This is highly wasteful from a statistical perspective and extrapolation schemes estimate instead the optimum using a weighted average of the last iterates produced by the underlying algorithm, where the weights depend on the iterates (i.e. a *nonlinear* average). Overall, computing those weights means solving a small linear system, so nonlinear acceleration has marginal computational complexity.

Recent results by (Scieur et al., 2016) adapted classical extrapolation techniques related to Aitken's Δ^2 and minimal polynomial extrapolation to design extrapolation schemes for accelerating the convergence of basic optimization methods such as gradient descent. They showed that by using only iterates from fixed-step gradient descent, these extrapolation algorithms achieve the optimal convergence rate of (Nesterov, 2013) *without any modification to the original algorithm.* However, these results are only applicable to iterates produced by single-step algorithms such as gradient descent, where the underlying operator is symmetric, thus excluding much faster momentum-based methods such as SGD with momentum or Nesterov's algorithm.

Our results here seek to extend those of (Scieur et al., 2016) to multi-step methods, i.e. to accelerate accelerated methods. We use Crouzeix's recent results (Crouzeix, 2007; Crouzeix and Palencia, 2017; Greenbaum et al., 2017) to show that, in the general non-symmetric case, acceleration performance is controlled by the solution of a Chebyshev problem on the numerical range of the linear, non-symmetric operator modelling the behavior of iterates near the optimum. We characterize the shape of this numerical range for various classical multi-step algorithms such as Nesterov's method (Nesterov, 1983), and Chambolle-Pock's algorithm (Chambolle and Pock, 2011).

We then study the performance of our techniques on several classical applications, e.g. image processing problems using extrapolation on Chambolle-Pock's algorithm.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

2 Nonlinear Acceleration

Consider the following optimization problem

$$\min_{x\in\mathbb{R}^n} f(x) \tag{1}$$

in the variable $x \in \mathbb{R}^n$, where $f(x)$ is strongly convex with parameter μ with respect to the Euclidean norm, and has a Lipschitz continuous gradient with parameter L with respect to the same norm. Assume we solve this problem using an iterative algorithm of the form

$$x_i = g(x_{i-1}) \quad \text{for } i = 1, ..., k, \tag{2}$$

where $x_i \in \mathbb{R}^n$, k is the number of iterations. As in (Scieur et al., 2016) we will focus on improving our estimates of the solution to problem (1) by tracking only the sequence of iterates x_i produced by an optimization algorithm, without any further oracle calls to $g(x)$. As in (Scieur et al., 2016), we will first focus here on the case where g is a linear mapping

$$g(x) = A(x - x^*) + x^*. \tag{3}$$

However, the main difference with (Scieur et al., 2016) is that we allow the mapping A to be *non-symmetric*, which is typically the case for methods with momentum terms, or primal-dual algorithms (where x is a concatenation of several iterates x_i, x_{i-1}, ..., of the primal and dual iterates).

We now briefly recall the key ideas driving nonlinear acceleration schemes. Nonlinear acceleration aims to find an approximation of the fixed point x^* (assumed to be unique) of g, i.e.

$$x^* = g(x^*)$$

using a linear combination of previous iterates x_i with coefficients c_i. The optimal coefficients c^* to approximate the fixed point using $\sum_{i=1}^k c_i x_i$ are found by minimizing the residual of the linear combination,

$$c^* = \arg\min_c \left\| g\left(\sum_{i=1}^k c_i x_i\right) - \sum_{i=1}^k c_i x_i \right\|,$$

in the variable $c \in \mathbb{R}^k$. Of course, this subproblem can be hard to solve for nonlinear functions g, so we will solve instead

$$\begin{aligned} c^* &= \arg\min_{c:c^T\mathbf{1}=1} \left\| \sum_{i=1}^k c_i g(x_i) - \sum_{i=1}^k c_i x_i \right\| \\ &= \arg\min_{c:c^T\mathbf{1}=1} \left\| \sum_{i=1}^k c_i r_i \right\|, \end{aligned}$$

Algorithm 1 Regularized Nonlinear Acceleration (Complexity: $\mathcal{O}(k^2 d)$ if $k \ll d$)

Input: Sequences of k iterates x_i generated by (2), regularization parameter λ.
Compute matrix of residues $R = [x_1 - x_0, \ldots, x_k - x_{k-1}]$.
Solve the linear system $(R^T R + \lambda I)z = \mathbf{1}$.
Normalize $c = z/(\mathbf{1}^T z)$.
Output: The extrapolated point $\sum_{i=1}^k c_i x_{i-1}$.

with $x_i = g(x_{i-1})$ and $r_i = x_i - x_{i-1}$. Minimizing on the residues may be unstable, so we add a Tychonov regularization term, which leads to the Regularized Nonlinear Acceleration (RNA) algorithm in (Scieur et al., 2016).

As in (Scieur et al., 2016), we can link the accuracy of the extrapolation $\sum_{i=1}^k c_i x_i$ with the norm of a matrix polynomial $p(A)$ where A is the linear fixed point operator in (3). In fact, (Scieur et al., 2016, Prop. 2.2) shows

$$\begin{aligned} \min_{c:c^T\mathbf{1}=1} \left\| \sum_{i=1}^k c_i r_i \right\| &= \min_{p\in\mathcal{P}_{k-1}:p(1)=1} \|p(A) r_1\| \quad (4) \\ &\leq \|r_1\| \min_{p\in\mathcal{P}_{k-1}:p(1)=1} \|p(A)\| \quad (5) \end{aligned}$$

where $\mathcal{P}_k$ is the linear space of polynomials of degree at most k. This then directly yields a convergence bound on $\| \sum_{i=0}^k c_i x_{i-1} - x^* \|$, since

$$\begin{aligned} \left\| \sum_{i=1}^k c_i x_{i-1} - x^* \right\| &= \left\| (A-I)^{-1} \sum_{i=1}^k c_i (x_i - x_{i-1}) \right\| \\ &\leq \left\| (A-I)^{-1} \right\| \left\| \sum_{i=1}^k c_i r_i \right\|. \end{aligned}$$

Thus, if the coefficients c^* in (4) are used, and if we assume $A - I$ invertible, then,

$$\left\| \sum_{i=1}^k c_i^* x_{i-1} - x^* \right\| \leq \left\| (A-I)^{-1} \right\| \|r_1\| \min_{\substack{p\in\mathcal{P}_{k-1} \\ p(1)=1}} \|p(A)\|. \tag{6}$$

Of course, these results hold only if x_k is generated with a linear mapping $g(x)$. We show that these results can be extended to nonlinear and stochastic iterations in section 6.3 of the supplementary material.

In the next section, we will see how to control the convergence bound using the numerical range of A.

3 Crouzeix's Conjecture & Chebyshev Polynomials on the Numerical Range

We have seen that the convergence rate of nonlinear acceleration is bounded by the norm of a matrix polynomial,

$$\min_{\substack{p\in\mathcal{P}_k\\ p(1)=1}} \|p(A)\|.$$

Here and in the rest of this paper, $\|\cdot\|$ is the ℓ_2 norm. The results in (Scieur et al., 2016) recalled above handle the case where the operator A is *symmetric*. Bounding $\|p(A)\|_2$ when A is non-symmetric is not as direct. Fortunately, Crouzeix's conjecture (Crouzeix, 2004) allows us to bound $\|p(A)\|_2$ by solving a Chebyshev problem on the numerical range of A, in the complex plane.

Theorem 3.1 (Crouzeix (2004)) *Let $A \in \mathbb{C}^{n\times n}$, and $p(x) \in \mathbb{C}[x]$, we have*

$$\|p(A)\|_2 \leq c \max_{z\in W(A)} |p(z)|$$

for some absolute constant $c \geq 2$.

Here $W(A) \subset \mathbb{C}$ is the numerical range of the matrix $A \in \mathbb{R}^{n\times n}$, i.e. the range of the Rayleigh quotient

$$W(A) \triangleq \{x^*Ax : \|x\|_2 = 1, x \in \mathbb{C}^n\}. \tag{7}$$

(Crouzeix, 2007) shows $c \leq 11.08$ and Crouzeix's conjecture states that this can be further improved to $c = 2$, which is tight. A more recent bound in (Crouzeix and Palencia, 2017) yields $c = 1 + \sqrt{2}$ and there is significant numerical evidence in support of the $c = 2$ conjecture (Greenbaum et al., 2017). This conjecture has played a vital role in providing convergence results for e.g. the GMRES method (Saad and Schultz, 1986) (see (Choi and Greenbaum, 2015)).

Crouzeix's result allows us to turn the problem of finding uniform bounds for the norm of the matrix polynomial $\|p(A)\|_2$ to that of bounding $p(z)$ over the numerical range of A in the complex plane, an arguably much simpler two-dimensional Chebyshev problem.

3.1 Numerical Range Approximations

There are no tractable methods for computing the exact numerical range of a general operator A. However, efficient numerical methods approximate the numerical range based on its key properties. The Toeplitz-Hausdorff theorem (Hausdorff, 1919; Toeplitz, 1918) in particular states that the numerical range $W(A)$ is a closed convex bounded set. Therefore, it suffices to characterize points on the boundary, the convex hull then yields the numerical range.

Johnson (1978) made the following observations using the properties of the numerical range,

$$\max_{z\in W(A)} Re(z) = \max_{r\in W(H(A))} r = \lambda_{max}(H(A)) \tag{8}$$

$$W(e^{i\theta}A) = e^{i\theta}W(A), \quad \forall\theta \in [0, 2\pi), \tag{9}$$

where $Re(z)$ is the real part of complex number z, $H(A)$ is the Hermitian part of A, i.e. $H(A) = (A + A^*)/2$ and $\lambda_{max}(H(A))$ is the maximum eigenvalue of $H(A)$. The first property implies that the line parallel to the imaginary axis is tangent to $W(A)$ at $\lambda_{max}(H(A))$. The second property can be used to determine other tangents via rotations. Using these observations Johnson (1978) showed that the points on the boundary of the numerical range can be characterized as $p_\theta = \{v_\theta^* A v_\theta : \theta \in [0, 2\pi)\}$ where v_θ is the normalized eigenvector corresponding to the largest eigenvalue of the Hermitian matrix

$$H_\theta = \frac{1}{2}(e^{i\theta}A + e^{-i\theta}A^*) \tag{10}$$

The numerical range can thus be characterized as follows.

Theorem 3.2 *(Johnson, 1978) For any $A \in \mathbb{C}^{n\times n}$, we have*

$$W(A) = Co\{p_\theta : 0 \leq \theta < 2\pi\}$$

where $Co\{Z\}$ is the convex hull of the set Z.

Note that p_θ cannot be uniquely determined as the eigenvectors v_θ may not be unique but the convex hull above is uniquely determined.

3.2 Chebyshev Bounds & Convergence Rate

Crouzeix's result means that bounding the convergence rate of accelerated algorithms can be achieved by bounding the optimum of the Chebyshev problem

$$\min_{\substack{p\in\mathbb{C}[z]\\ p(1)=1}} \max_{z\in W(A)} |p(z)| \tag{11}$$

where $A \in \mathbb{C}^{n\times n}$. This problem has a trivial answer when the numerical range $W(A)$ is spherical, but the convergence rate can be significantly improved when $W(A)$ is less isotropic.

3.2.1 Exact Bounds on Ellipsoids

We can use an outer ellipsoidal approximation of $W(A)$, bounding the optimum value of the Chebyshev problem (11) by

$$\min_{\substack{p(z)\in\mathbb{C}[x]\\ p(1)=1}} \max_{z\in\mathcal{E}_r} |p(z)| \tag{12}$$

where

$$\mathcal{E}_r \triangleq \{z \in \mathbb{C} : |z-1| + |z+1| \leq r + 1/r\}. \qquad (13)$$

This Chebyshev problem has an explicit solution in certain regimes. As in the real case, we will use $C_n(z)$, the Chebyshev polynomial of degree k. Fischer and Freund (1991) derived the optimal solution to problem (12) on ellipsoids, recalled as Theorem 6.2 in the supplementary material.

The optimal polynomial for a general ellipse $\mathcal{E}$ can be obtained by a simple change of variables. That is, the polynomial $C_k(\frac{c-z}{d})/C_k(\frac{c-1}{d})$ is optimal for the problem (12) over any ellipse $\mathcal{E}$ with center c, focal distance d and semi-major axis a. It can be easily seen that the maximum value is achieved at the point a on the real axis. That is the solution to the min max problem is given by $\bar{T}_k(a)$. Figure 1 shows the surface of the optimal polynomial with degree 5 for $a = 0.8, d = 0.76$ and $c = 0$.

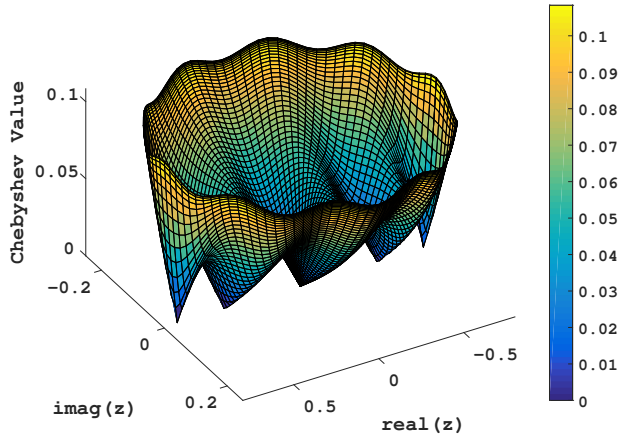

Figure 1: Surface of the optimal polynomial $\bar{T}_n(z)$ with degree 5 for $a = 0.8, d = 0.76$ and $c = 0$.

Figure 2 shows the solutions to the problem (12) with degree 5 for various ellipses with center at origin, various eccentricity values $e = d/a$ and semi-major axis a. Here, zero eccentricity corresponds to a sphere, while an eccentricity of one corresponds to a line.

4 Accelerating Non-symmetric Algorithms

We have seen in the previous section that controlling the convergence rate of the nonlinear acceleration scheme in Algorithm 1 means bounding the optimal value of the Chebyshev optimization problem in (11) over the numerical range of the operator driving iterations. In what follows, we explicitly detail this operator and approximate its numerical range for two classical algorithms, Nesterov's accelerated method (Nesterov, 1983) and Chambolle-Pock's Primal-Dual Algorithm (Chambolle and Pock, 2011).

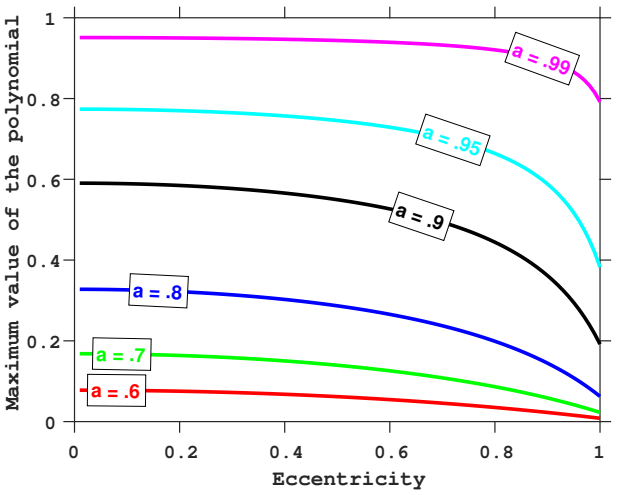

Figure 2: Optimal value of the Chebyshev problem (12) for ellipses with centers at origin. Lower values of the maximum of the Chebyshev problem mean faster convergence. The higher the eccentricity here, the faster the convergence.

4.1 Nesterov's Accelerated Gradient Method

The iterates formed by Nesterov's accelerated gradient descent method for minimizing smooth strongly convex functions with constant stepsize follow

$$\begin{cases} x_k = y_{k-1} - \alpha \nabla f(y_{k-1}) \\ y_k = x_k + \beta(x_k - x_{k-1}) \end{cases} \qquad (14)$$

with $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$, where L is the gradient's Lipschitz continuity constant and μ is the strong convexity parameter. This algorithm is better handled using the results in (Scieur et al., 2018), and we only use it here to better illustrate our results on non-symmetric operators.

4.1.1 Nesterov's Operator in the quadratic case

When minimizing quadratic functions $f(x) = \frac{1}{2}\|Bx - b\|^2$, using constant stepsize $1/L$, these iterations become,

$$\begin{cases} x_k - x^* &= y_{k-1} - x^* - \frac{1}{L}B^T(By_{k-1} - b) \\ y_k - x^* &= x_k - x^* + \beta(x_k - x^* - x_{k-1} + x^*). \end{cases}$$

or again,

$$\begin{bmatrix} x_k - x^* \\ y_k - x^* \end{bmatrix} = \begin{bmatrix} 0 & A \\ -\beta I & (1+\beta)A \end{bmatrix} \begin{bmatrix} x_{k-1} - x^* \\ y_{k-1} - x^* \end{bmatrix}$$

where $A = I - \frac{1}{L}B^TB$. We write O the *non-symmetric* linear operator in these iterations, i.e.

$$O = \begin{bmatrix} 0 & A \\ -\beta I & (1+\beta)A \end{bmatrix} \tag{15}$$

The results in Section 2 show that we can accelerate the sequence $z_k = (x_k, y_k)$ if the solution to the min-max problem (11) defined over the numerical range of the operator O is bounded.

4.1.2 Numerical Range

We can compute the numerical range of the operator O using the techniques described in Section (2). In the particular case of Nesterov's accelerated gradient method, the numerical range is a convex hull of ellipsoids. We show this by considering the 2×2 operators obtained by replacing the symmetric positive matrix A with its eigenvalues, to form

$$O_j = \begin{bmatrix} 0 & \lambda_j \\ -\beta I & (1+\beta)\lambda_j \end{bmatrix} \quad \text{for } j \in \{1, 2, \cdots, n\} \tag{16}$$

where $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n < 1$ are the eigenvalues of the matrix A. We have the following result (proved in the supplementary material).

Theorem 4.1 *The numerical range of operator O is given as the convex hull of the numerical ranges of the operators O_j, i.e. $W(O) = \mathbf{Co}\{W(O_1), W(O_2), \cdots, W(O_n)\}$.*

To minimize the solution of the Chebyshev problem in (11) and control convergence given the normalization constraint $p(1) = 1$, the point $(1, 0)$ should be outside the numerical range. Because the numerical range is convex and symmetric w.r.t. the real axis (the operator O is real), this means checking if the maximum real value of the numerical range is less than 1.

For 2×2 matrices, the boundary of the numerical range is given by an ellipse (Donoghue, 1957), so the numerical range of Nesterov's accelerated gradient method is the convex hull of ellipsoids. The ellipse in (Donoghue, 1957) can be determined directly from the entries of the matrix as in Johnson (1974), with details provided in Theorem 6.1 of the Supplementary Material. This allows us to compute the maximum real value of $W(O)$, as the point of intersection of $W(O)$ with the real line which can be computed explicitly as,

$$\begin{aligned} re(O) &= \max Re(W(O)) = \max_j Re(W(O_j)) \\ &= \frac{1}{2}\left((1+\beta)\lambda_n + \sqrt{\lambda_n^2(1+\beta)^2 + (\lambda_n - \beta)^2}\right) \end{aligned}$$

where $\lambda_n = 1 - \frac{\mu}{L}$.

We observe that $re(O)$ is a function of the condition number of the problem and takes the values in the interval $[0, 2]$. Therefore, RNA will only work on Nesterov's accelerated gradient method when $re(O) < 1$ holds, which implies that the condition number of the problem $\kappa = \frac{L}{\mu}$ should be less than around 2.5 which is highly restrictive.

An alternative approach is to use RNA on a sequence of iterates sampled every few iterations, which is equivalent to using powers of the operator O. We expect the numerical radius of some power of operator O to be less than 1 for any conditioning of the problem. This is because the iterates are converging at an $R-$linear rate and so the norm of the power of the operator is decreasing at an $R-$linear rate with the powers. Therefore, using the property that the numerical radius is bounded by the norm of the operator we have,

$$re(O^p) = \max Re(W(O^p)) \leq r_{O^p} \leq \|O^p\| \leq C_p \rho^p$$

where r_{O^p} is the numerical radius of O^p. Figure 3 shows the numerical range of the powers of the operator O for a random matrix B^TB with dimension $d = 50$. We observe that after some threshold value for the power p, $(1, 0)$ lies outside the field values corresponding to O^p thus guaranteeing that the acceleration scheme will work. We also observe that the boundaries of the field values are almost circular for higher powers p, which is consistent with results on optimal matrices in (Lewis and Overton, 2018). When the numerical range is circular, the solution of the Chebyshev problem is trivially equal to z^p so RNA simply picks the last iterate and does not accelerate convergence.

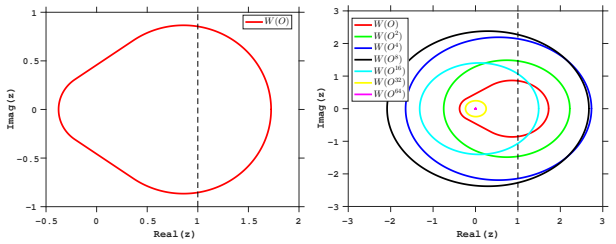

Figure 3: Numerical range for the linear operator in Nesterov's method, on a random quadratic problem with dimension 50. Left: Operator O. Right: Various operator powers O^p. The RNA scheme will improve convergence whenever the point $(1, 0)$ lies outside of the numerical range of the operator.

The difficulty in performing RNA on Nesterov's accelerated gradient method arises due to the fact that the iterates can be non-monotonic. The restriction that 1 should be outside the numerical range is necessary for both non-symmetric and symmetric operators. In symmetric operators, the numerical range is a line segment on the real axis and the numerical radius and

spectral radius are equal, so this restriction is equivalent to having spectral radius less than 1, i.e. having monotonically converging iterates.

4.2 Chambolle-Pock's Primal-Dual Algorithm

Chambolle-Pock is a first-order primal-dual algorithm used for minimizing composite functions of the form

$$\min_x h_p(x) := f(Ax) + g(x) \tag{17}$$

where f and g are convex functions and A is a continuous linear map. Optimization problems of this form arise in e.g. imaging applications like total variation minimization (see Chambolle and Pock (2016)). The Fenchel dual of this problem is given by

$$\max_y h_d(y) := -f^*(-y) - g^*(A^*y) \tag{18}$$

where f^*, g^* are the convex conjugate functions of f, g respectively. These problems are primal dual formulations of the general saddle point problem,

$$\min_x \max_y < Ax, y > +g(x) - f^*(y), \tag{19}$$

where f^*, g are closed proper functions. Chambolle and Pock (2011) designed a first-order primal-dual algorithm for solving these problems, where primal-dual iterates are given by

$$\begin{cases} y_{k+1} = \mathbf{Prox}^{\sigma}_{f^*}(y_k + \sigma A\bar{x}_k) \\ x_{k+1} = \mathbf{Prox}^{\tau}_{g}(x_k - \tau A^* y_{k+1}) \\ \bar{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k) \end{cases} \tag{20}$$

where σ, τ are the step length parameters, $\theta \in [0, 1]$ is the momentum parameter and the proximal mapping of a function f is defined as

$$\mathbf{Prox}^{\tau}_f(y) = \arg\min_x \left\{ \|y - x\|^2/(2\tau) + f(x) \right\}$$

Note that if the proximal mapping of a function is available then the proximal mapping of the conjugate of the function can be easily computed using Moreau's identity, with

$$\mathbf{Prox}^{\tau}_f(y) + \mathbf{Prox}^{1/\tau}_{f^*}(y/\tau) = y$$

The optimal strategy for choosing the step length parameters σ, τ and the momentum parameter θ depend on the smoothness and strong convexity parameters of the problem. When f^* and g are strongly convex with strong convexity parameters δ and γ respectively then these parameters are chosen to be constant values given as

$$\sigma = \frac{1}{\|A\|}\sqrt{\frac{\gamma}{\delta}} \quad \tau = \frac{1}{\|A\|}\sqrt{\frac{\delta}{\gamma}} \qquad \theta = \left(1 + \frac{2\sqrt{\gamma\delta}}{\|A\|}\right)^{-1} \tag{21}$$

to yield the optimal linear rate of convergence. When only one of f^* or g is strongly convex with strong convexity parameter γ, then these parameters are chosen adaptively at each iteration as

$$\theta_k = (1 + 2\gamma\tau_k)^{-1/2} \quad \sigma_{k+1} = \sigma_k/\theta_k \quad \tau_{k+1} = \tau_k\theta_k \tag{22}$$

to yield the optimal sublinear rate of convergence.

A special case of the primal-dual algorithm with no momentum term, i.e., $\theta = 0$ in (20) is also known as the Arrow-Hurwicz method (Mizoguchi (1960)). Although theoretical complexity bounds for this algorithm are worse compared to methods including a momentum term, it is observed experimentally that the performance is either on par or sometimes better, when step length parameters are chosen as above.

We first consider algorithms with no momentum term and apply RNA to the primal-dual sequence $z_k = (y_k, x_k)$. We note that, as observed in the Nesterov's case, RNA can only be applied on non-symmetric operators for which the normalization constant 1 is outside their numerical range. Therefore, the step length parameters τ, σ should be suitably chosen such that this condition is satisfied.

4.2.1 Chambolle-Pock's Operator in the Quadratic Case

When minimizing smooth strongly convex quadratic functions where $f(Ax) = \frac{1}{2}\|Ax - b\|^2$ and $g(x) = \frac{\mu}{2}\|x\|^2$, the proximal operators have closed form solutions. That is

$$\mathbf{Prox}^{\sigma}_{f^*}(y) = \frac{y - \sigma b}{1 + \sigma} \quad \text{and} \quad \mathbf{Prox}^{\tau}_{g}(x) = \frac{1}{1 + \tau\mu}.$$

Iterates of the primal-dual algorithm with no momentum term can be written as,

$$y_{k+1} = \frac{y_k + \sigma A x_k - \sigma b}{1 + \sigma}, \quad x_{k+1} = \frac{x_k - \tau A^T y_{k+1}}{1 + \tau\mu}$$

Note that the optimal primal and dual solutions satisfy $y^* = Ax^* - b$ and $x^* = \frac{-1}{\mu}A^T y$. This yields the following operator for iterations

$$O = \begin{bmatrix} \frac{I}{1+\sigma} & \frac{\sigma A}{1+\sigma} \\ \frac{\tau A^T}{(1+\sigma)(1+\tau\mu)} & \frac{I}{1+\tau\mu} - \frac{\tau\sigma A^T A}{(1+\sigma)(1+\tau\mu)} \end{bmatrix}. \tag{23}$$

Note that O is a non-symmetric operator except when $\sigma = \frac{\tau}{1+\tau\mu}$, in which case the numerical range is a line segment on the real axis and the spectral radius is equal to the numerical radius.

4.2.2 Numerical Range

The numerical range of the operator can be computed using the techniques described in Section 2. As men-

tioned earlier, the point 1 should be outside the numerical range for the Chebyshev polynomial to be bounded. Therefore, using (8), we have, $re(O) = \max Re(W(O)) = \lambda_{max}\left(\frac{O+O^*}{2}\right)$ The step length parameters σ, τ should be chosen such that the above condition is satisfied. We observe empirically that there exists a range of values for the step length parameters such that $re(O) < 1$. Figure 6 in the supplementary material shows the numerical range of operator O for $\sigma = 4, \tau = 1/\|A^T A\|$ with two different regularization constants and Figure 7 in the supplementary material shows the regions for which $re(O^p) \leq 1$ (converging) for different values of σ and τ.

We also consider non-smooth problems in addition to the smooth strongly convex problems in the numerical experiments section. While our scheme does not explicitly handle nonsmoothness but we report some preliminary empirical results which show the benefits of RNA.

5 Numerical Results

We now study the performance of our techniques on several classical applications, e.g. image processing problems using extrapolation on Chambolle-Pock's algorithm. We consider two different classes of problems: smooth strongly convex problems and non-smooth convex problems.

5.1 Smooth Problems

We consider ridge regression and l_2 regularized logistic regression problems which are of the form $h(x) := f(Ax) + g(x)$ where $f(Ax) = \frac{1}{2}\|Ax - b\|^2$ for ridge regression and $f(Ax) = \sum \log(1 + \exp(-a_i^T x b_i))$ for logistic regression, and $g(x) = \frac{\mu}{2}\|x\|^2$. The following methods are tested in this experiment.

- **GD.** Gradient descent, with $x_{k+1} = x_k - \frac{1}{L}\nabla h(x_k)$, where L is the gradient's Lipschitz constant.
- **Nesterov.** Nesterov's accelerated gradient method

  $$x_{k+1} = y_k - \frac{1}{L}\nabla h(y_k),\ y_{k+1} = y_k + \beta(y_k - y_{k-1})$$

  where $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$, L is the gradient's Lipschitz constant.
- **LBFGS.** The LBFGS method (Liu and Nocedal, 1989) $x_{k+1} = x_k - \alpha_k H_k \nabla h(x_k)$ where the steplength parameter α_k is chosen via Armijo backtracking line search and the memory parameter is chosen to be 10.
- **PDGM.** The primal-dual gradient method (Chambolle and Pock, 2011; Mizoguchi, 1960)

  $$y_{k+1} = Prox^{\sigma}_{f^*}(y_k + \sigma A x_k)$$
  $$x_{k+1} = Prox^{\tau}_{g}(x_k - \tau A^* y_{k+1})$$

  where $\sigma = \frac{1}{\|A\|}\sqrt{\frac{\mu}{\delta}}$, $\tau = \frac{1}{\|A\|}\sqrt{\frac{\delta}{\mu}}$, δ is the strong convexity parameters of f^*.
- **PDGM + Momentum.** The primal-dual gradient method with momentum (Chambolle and Pock, 2011)

  $$\begin{aligned} y_{k+1} &= Prox^{\sigma}_{f^*}(y_k + \sigma A \bar{x}_k \\ x_{k+1} &= Prox^{\tau}_{g}(x_k - \tau A^* y_{k+1}) \\ \bar{x}_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k) \end{aligned}$$

  where $\sigma = \frac{1}{\|A\|}\sqrt{\frac{\mu}{\delta}}$, $\tau = \frac{1}{\|A\|}\sqrt{\frac{\delta}{\mu}}$, $\theta = \frac{1}{1+\frac{2\sqrt{\mu\delta}}{\|A\|}}$, δ is the strong convexity parameters of f^*.

The Lipschitz constant L is $\|A\|^2 + \mu$ for ridge regression and is $\frac{\|A\|^2}{4} + \mu$ for logistic regression. The strong convexity parameter δ of the dual function f^* is 1 for ridge regression and is 4 for logistic regression. The proximal operators used in the primal - dual algorithms have closed form solutions for ridge regression. That is, $Prox^{\tau}_{g}(x) = \frac{1}{1+\tau\mu}$ and $Prox^{\sigma}_{f^*}(y) = \frac{y-\sigma b}{1+\sigma}$. In logistic regression, the approximate proximal operator of f^* is obtained by running Newton's method till some tolerance on the accuracy is achieved or a maximum of 100 iterations is reached. Note that the dominant cost in computing the gradients or proximal operators is the cost of computing the matrix vector products Ax and A^*y which are of the order $O(Nd)$ and the cost of performing Newton's method to obtain the proximal operator is of order N times the maximum number of iterations t. Therefore, when $t < d$ one can ignore the additional cost of performing Newton's method.

We use online RNA (described in section 6.2 of supplementary material) on GD, Nesterov and PDGM with a fixed window size $m = 10$ and set $\lambda = 10^{-8}\|R^T R\|_2$. As discussed in Section 4, RNA can be applied only with specific choices of the step-length parameters in the case of primal-dual methods. In the case of smooth problems, we observe that the choice $\tau = \frac{1}{\|A\|}$ and $\sigma = \frac{1}{\|A\|}$ yields stability for applying RNA on PDGM. We note this choice is not an optimal choice and one can improve the results by suitably tuning these parameters.

Figure 4 in the supplementary material shows the performance of different variants of the primal-dual algorithms on ridge regression problems for two different regularization constants. We observe that there

is no significant difference in the performance of the method with the momentum term (θ) as compared to the one with no momentum term. We also observe that although the choice of the steplength parameters mentioned above have consistent performance across different problems, the improvements obtained with RNA are not very significant. However, choosing $\sigma = \tau = 1/\|A\|$ and applying RNA to the PDGM has consistently outperformed all other variants. This is in consistent with theoretical observations made in Section 4 that one can find optimal steplength parameters for which RNA is stable and obtains the optimal performance. Additional results corresponding to different algorithms on logistic and ridge regression problems are given in section 6.5 of the supplementary material.

We also compare the performance of offline, restart (Scieur et al., 2016) and online versions of RNA on primal-dual gradient methods in Figure 5 of the supplementary material. We observe that the improvement in the performance is more pronounced in the online version (where the algorithm is restarted at each iteration) of RNA as compared to the offline version.

5.2 Non-Smooth Problems

We consider denoising an image that is degraded by Gaussian noise using total variation. We refer the reader to Chambolle and Pock (2016) for details about the total variation models. The optimization problem is given as,

$$\min_x \|\nabla x\|_1 + \mu\|x - b\|^2/2$$

where, $\|\nabla x\|_1 = \sum_{i,j} \sqrt{((\nabla x)^1_{i,j})^2 + ((\nabla x)^2_{i,j})^2}$ and b is a 256 by 256 noisy input image. This optimization problem is in the form (17) with $f(\nabla x) = \|\nabla x\|_1$ and $g(x) = \frac{\mu}{2}\|x - b\|^2$. The gradient operator ∇x is discretized by forward differencing (see Chambolle and Pock (2011)). The convex conjugate of f is an indicator function of the convex set P where,

$$P = \{p : \|p\|_\infty \leq 1\}, \quad |p\|_\infty = \max_{i,j} \sqrt{(p^1_{i,j})^2 + (p^2_{i,j})^2}$$

and so the proximal operator is a point wise projection on to this set, so $Prox^\sigma_{f^*}(p)_{i,j} = p_{i,j}/\max(1, |p_{i,j}|)$.

We compare the performance of the two variants of primal-dual methods with RNA for two different noise levels ζ with two different regularization constants μ. The step-length parameters are chosen adaptively at each iteration as follows:

- **PDGM**

  $\hat{\theta}_k = (1+2\gamma\tau_k)^{-1/2} \quad \sigma_{k+1} = \sigma_k/\hat{\theta}_k \quad \tau_{k+1} = \tau_k\hat{\theta}_k$

  with $\gamma = 0.2\mu$, $\theta = 0$, $\tau_0 = 0.02$, $\sigma_0 = \frac{4}{\tau_0\|\nabla\|^2}$

- **PDGM + Momentum** as above with $\gamma = 0.7\mu$ and $\sigma_0 = \tau_0 = 1/\|\nabla\|$

with $\|\nabla\|^2 = 8$. These adaptive choices are the standard choices used in the literature and yield the optimal theoretical convergence rates for the momentum variants. We note that these parameters are not carefully fine-tuned to give the best performance for each variant but are chosen based on some simple observations. We used the offline RNA instead of online RNA as we consistently observed that the offline RNA is more robust in the high accuracy regime and the online variants needed some stability inducing techniques like line searches. Moreover, for the online RNA, the improvement in the performance on these non-smooth problems is small and so the additional cost of solving the linear system is not well justified. Results showing the number of iterations required to achieve different accuracy levels on image denoising problem are given in Table 1 of the supplementary material. We observe that the offline RNA variant of PDGM method consistently outperformed PDGM and its momentum variant.

Acknowledgements

The authors are very grateful to Lorenzo Stella for fruitful discussions on acceleration and the Chabolle-Pock method. AA is at CNRS & département d'informatique, École normale supérieure, UMR CNRS 8548, 45 rue d'Ulm 75005 Paris, France, INRIA and PSL Research University. The authors would like to acknowledge support from the *ML & Optimisation* joint research initiative with the *fonds AXA pour la recherche* and Kamet Ventures, as well as a Google focused award. DS was supported by a European Union Seventh Framework Programme (FP7-PEOPLE-2013-ITN) under grant agreement n.607290 SpaRTaN. RB was supported by Department of Energy grant DE-FG02-87ER25047 and DARPA grant 650-4736000-60049398.

References

Claude Brezinski and M Redivo Zaglia. *Extrapolation methods: theory and practice*, volume 2. Elsevier, 2013.

Stan Cabay and LW Jackson. A polynomial extrapolation method for finding limits and antilimits of vector sequences. *SIAM Journal on Numerical Analysis*, 13(5): 734–752, 1976.

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 and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

Daeshik Choi and Anne Greenbaum. Roots of matrices in the study of gmres convergence and crouzeix's conjecture. *SIAM Journal on Matrix Analysis and Applications*, 36(1):289–301, 2015.

Michel Crouzeix. Bounds for analytical functions of matrices. *Integral Equations and Operator Theory*, 48(4): 461–477, 2004.

Michel Crouzeix. Numerical range and functional calculus in hilbert space. *Journal of Functional Analysis*, 244(2): 668–690, 2007.

Michel Crouzeix and César Palencia. The numerical range as a spectral set. *arXiv preprint arXiv:1702.00668*, 2017.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

William F. Donoghue. On the numerical range of a bounded operator. *Michigan Math. J.*, 4(3):261–263, 1957.

RP Eddy. Extrapolating to the limit of a vector sequence. In *Information linkage between applied mathematics and industry*, pages 387–396. Elsevier, 1979.

Bernd Fischer and Roland Freund. Chebyshev polynomials are not always optimal. *Journal of Approximation Theory*, 65(3):261–272, 1991.

R Paul Gorman and Terrence J Sejnowski. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1:75, 1988.

Anne Greenbaum, Adrian S Lewis, and Michael L Overton. Variational analysis of the crouzeix ratio. *Mathematical Programming*, 164(1-2):229–243, 2017.

Isabelle Guyon. Design of experiments of the nips 2003 variable selection benchmark, 2003.

Felix Hausdorff. Der wertvorrat einer bilinearform. *Mathematische Zeitschrift*, 3(1):314–316, 1919.

Charles R Johnson. Computation of the field of values of a 2× 2 matrix. *J. Res. Nat. Bur. Standards Sect. B*, 78: 105, 1974.

Charles R Johnson. Numerical determination of the field of values of a general complex matrix. *SIAM Journal on Numerical Analysis*, 15(3):595–602, 1978.

A. Lewis and M. Overton. Partial smoothness of the numerical radius at matrices whose fields of values are disks. *Working paper (mimeo)*, 2018.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Toshiyuki Mizoguchi. K.j. arrow, l. hurwicz and h. uzawa, studies in linear and non-linear programming. *Economic Review*, 11(3):349–351, 1960.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Youcef Saad and Martin H Schultz. Gmres: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on scientific and statistical computing*, 7(3):856–869, 1986.

Damien Scieur, Alexandre d'Aspremont, and Francis Bach. Regularized nonlinear acceleration. In *Advances In Neural Information Processing Systems*, pages 712–720, 2016.

Damien Scieur, Francis Bach, and Alexandre d'Aspremont. Nonlinear acceleration of stochastic algorithms. In *Advances in Neural Information Processing Systems*, pages 3985–3994, 2017.

Damien Scieur, Edouard Oyallon, Alexandre d'Aspremont, and Francis Bach. Nonlinear acceleration of deep neural networks. *arXiv preprint arXiv:1805.09639*, 2018.

Otto Toeplitz. Das algebraische analogon zu einem satze von fejér. *Mathematische Zeitschrift*, 2(1-2):187–197, 1918.

Homer F Walker and Peng Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011.