NESTEROV'S METHOD FOR CONVEX OPTIMIZATION

NOEL J. WALKINGTON*

Abstract. While Nesterov's algorithm for computing the minimum of a convex function is now over forty years old, it is rarely presented in texts for a first course in optimization. This is unfortunate since for many problems this algorithm is superior to the ubiquitous steepest descent algorithm, and equally simple to implement. This article presents an elementary analysis of Nesterov's algorithm that parallels that of steepest descent. It is envisioned that this presentation of Nesterov's algorithm could easily be covered in a few lectures following the introductory material on convex functions and steepest descent included in every course on optimization.

Key words. Convex Optimization, Nesterov's Algorithm, Steepest Descent.

AMS subject classifications. 65K10, 60C46, 60C25

1. Introduction. Given a closed convex subset $U \subset X$ of a Hilbert[1] space X and a convex function $f : X \to \mathbb{R}$, this article considers algorithms for solving the problem,

$$x_* \in U, \qquad f(x_*) \le f(x), \qquad x \in U. \tag{1.1}$$

The unconstrained problem, $U = X$, is first considered, and the latter sections consider the situation where

$$U = \{x \in X \mid \phi_i(x) \le 0, 1 \le i \le m\},$$

where $\phi_i : X \to \mathbb{R}$ is convex, $1 \le i \le m$.

This article considers *first order* algorithms for the solution of (1.1); that is, algorithms which only require $f(x)$ and the gradient $\nabla f(x)$ to be evaluated. This class of algorithm is useful when evaluation of higher order derivatives is not tractable; for example, if the dimension of X is large storing the matrix of second derivatives may exceed the available memory. First order methods are robust in the presence of degeneracy; for example, when $D^2 f(x)$ is singular and when solutions of the optimization problem are not unique. In this context the most direct approach to finding a minimum of f is to repeatably step in the direction of steepest descent; that is, given $x_0 \in X$, let

$$x_{i+1} = x_i - \tau_i \nabla f(x_i), \qquad i = 0, 1, 2, \ldots,$$

where $\tau_i > 0$ are step sizes. Specification of the step sizes, the Armijo rule in particular, is discussed below in Section 3. For unconstrained optimization, $U = X$, this rule becomes [2]

$$f(x_{i+1}) \le f(x_i) - (\tau_i/2)\|\nabla f(x_i)\|^2.$$

The following modification of the steepest descent algorithm was introduced by Nesterov in 1983; given $x_0 = y_0 \in X$, let

$$x_{i+1} = y_i - \tau_i \nabla f(y_i), \qquad y_{i+1} = x_{i+1} + \frac{\lambda_i - 1}{\lambda_{i+1}}(x_{i+1} - x_i), \qquad i = 0, 1, 2, \ldots, \tag{1.2}$$

*Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA 15213. Supported in part by National Science Foundation Grants DMS–2012259 This work was also supported by the NSF through the Center for Nonlinear Analysis.

[1]Prototypically $X = \mathbb{R}^n$; however, finite dimensionality is never required.

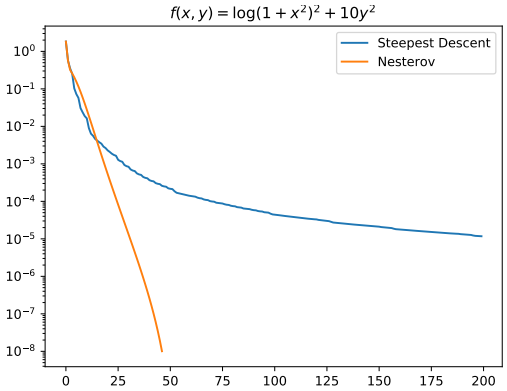

FIG. 1.1. *Function values for steepest descent and Nesterov algorithms for Example 1.2.*

where $\tau_i > 0$ is the step size and $\{\lambda_i\}_{i=0}^{\infty}$ is the sequence, with

$$\lambda_0 = 0, \qquad \text{and} \qquad \lambda_{i+1} = \frac{1 + \sqrt{1 + 4\lambda_i^2}}{2}. \qquad i = 0, 1, 2, \ldots \tag{1.3}$$

EXAMPLE 1.1. *When $X = \mathbb{R}^n$ and $f(x) = (1/2)x^\top Ax - b.x$ with $A \in \mathbb{R}^{n\times n}$ a symmetric positive definite (SPD) matrix and $b \in \mathbb{R}^n$, the method of steepest decent*

$$x_{i+1} = x_i - \tau(Ax_i - b),$$

is Richardson iteration [17, 24] for the solution of the linear system $Ax = b$ with step sizes $\tau_i = \tau$. This algorithm will converge if (and only if) $\tau < 1/\lambda_{max}$ where λ_{max} is the maximum eigenvalue of A.

Nesterov algorithm becomes

$$x_{i+1} = y_i - \tau(Ay_i - b), \qquad y_{i+1} = x_{i+1} + \frac{\lambda_i - 1}{\lambda_{i+1}}(x_{i+1} - x_i).$$

This scheme is similar in spirit to the "successively over relaxed" (SOR) variant of Richardson iteration, the difference being that SOR has y_i instead of x_i in the update formula for y_{i+1}.

When f is smooth, the sequence generated by the method of steepest descent will satisfy $f(x_n) - f(x_*) \leq O(1/n)$, and in his groundbreaking paper [16] Nesterov proved that $f(x_n) - f(x_*) \leq O(1/n^2)$ for his algorithm. In situations where these rates are optimal, this is a dramatic improvement over steepest descent.

EXAMPLE 1.2. *The function $f(x, y) = \log(1 + x^2)^2 + 10y^2$ is convex on $[-2.9, 2.9] \times \mathbb{R}$ and has a unique minimum $0 = f(0, 0)$ at which both the gradient and determinant of the Hessian vanish. Figure 1.1 plots the function values generated by the steepest descent and Nesterov algorithms with initial data $(x, y) = (1, 1)$ and the step sizes determined with the Armijo rule given in Lemma 3.2. The stopping criterion $|f(\mathbf{x}_n) - f(\mathbf{x}_{n-1})| < 10^{-8}$ results in the following approximations of the minimum.*

Scheme	*Iterations*	*Min value*	*Norm of gradient*
Steepest Descent	384	$3.529730e-06$	$3.380372e-04$
Nesterov	47	$1.006851e-08$	$2.334551e-06$

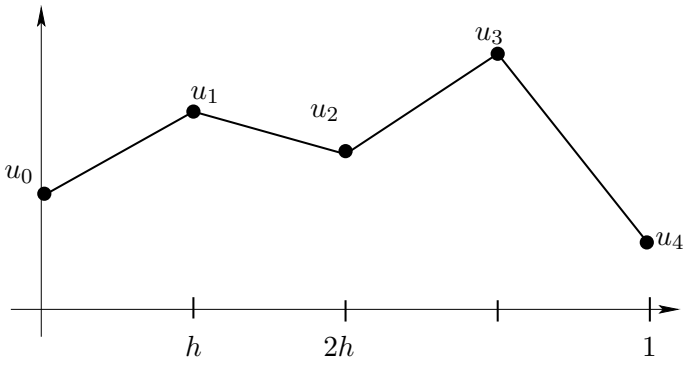

FIG. 1.2. *Discrete approximation of* $u(x)$

Figure 1.1 shows that initially the function values for method of steepest descent decreased more rapidly than those for Nesterov's algorithm. This is due to the fact that the Nesterov step sizes are required to satisfy $\tau_i \leq \tau_{i-1}$ *in addition to the Armijo rule, while for the method of steepest descent step sizes may increase.*

In the absence of degeneracy steepest descent will give a linear rate of convergence, $f(x_n) - f(x_*) \leq (1 - 1/\kappa)(f(x_{n-1}) - f(x_*))$ where $\kappa > 1$ is the condition number ($\kappa = L/\alpha$ below), so may perform better than the original Nesterov algorithm if κ is modest. If an estimate of κ is available, Nesterov's algorithm with fixed parameter $\lambda_n = (\sqrt{\kappa}+1)/2$ will converge linearly with $f(x_n) - f(x_*) \leq (1 - 1/\sqrt{\kappa})(f(x_{n-1}) - f(x_*))$; the proof of this is given in Section 3.3.

EXAMPLE 1.3. *Given noisy signal or image,* $\hat{u} : [0, 1] \to \mathbb{R}$*, variational techniques to recover the underlying signal construct minima of the function*

$$f(u) = \int_0^1 \left((1/2)(u - \hat{u})^2 + (\alpha/\beta)|u'|^\beta \right),$$

where $\beta \in [1, \infty)$ *and* $\alpha > 0$ *are parameters of the method [1, 20]. When* $\beta = 2$ f *is quadratic so the minima satisfies a linear equation. Discrete approximations construct a uniform partition* $x_i = ih$ *of* $[0, 1]$*, with* $i = 0, 1, \ldots, N$ *and* $h = 1/N$*. Writing* $\hat{u}_i = \hat{u}(ih)$ *and letting* u_i *be an approximation of* $u(ih)$ *and* $\mathbf{u} = \{u_i\}_{i=0}^N$*, the natural approximation of* f *is (see Figure 1.2)*

$$f(\mathbf{u}) = (h/4)\left((u_0 - \hat{u}_0)^2 + (u_N - \hat{u}_N)^2\right) + (h/2)\sum_{i=1}^{N-1}(u_i - \hat{u}_i)^2 + (\alpha/\beta)\sum_{i=1}^{N} h^{1-\beta}|u_i - u_{i-1}|^\beta.$$

Figure 1.3 plots $f(\mathbf{u}_n)$ *for the steepest descent, Nesterov, Nesterov with fixed parameter, and conjugate gradient, methods with parameters* $\alpha = 0.001$ *and* $\beta = 2$*, and* $N = 50$ *and* $N = 100$*, when* $\hat{u}$ *is a random perturbation of the "ground truth" signal*

$$\tilde{u}(x) = \begin{cases} \sqrt{x} & 0 \leq x < 1/2 \\ (1/2) + 2|x - 3/4| & 1/2 \leq x \leq 1. \end{cases}$$

The step sizes were computed using the Armijo rule for the steepest descent and Nesterov schemes with the additional restriction $\tau_i \leq \tau_{i-1}$ *for the Nesterov scheme, and stopping criteria* $\|\nabla f(\mathbf{u}_n)\| < 10^{-4}$*.*

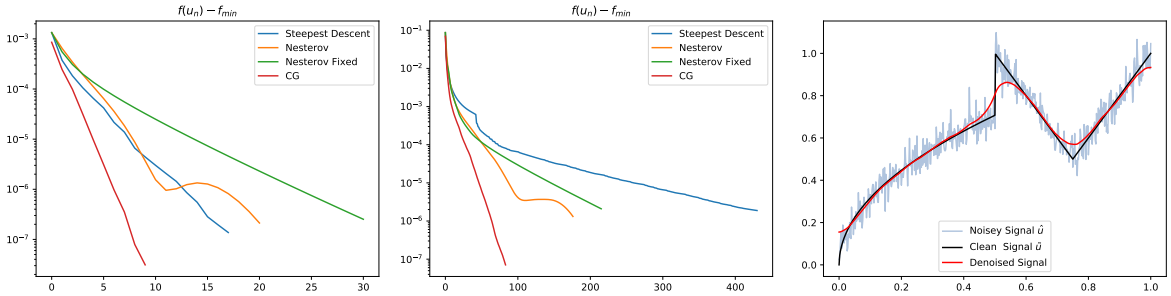

FIG. 1.3. *Function values for $N = 50$, $N = 500$ and solution for Example 1.2.*

	$N = 50$		$N = 500$	
Scheme	*Iterations*	*Norm of gradient*	*Iterations*	*Norm of gradient*
Steepest Descent	18	$8.832316e-05$	431	$9.913039e-05$
Nesterov	21	$8.527445e-05$	177	$9.970484e-05$
Nesterov Fixed	31	$9.700755e-05$	216	$9.951276e-05$
CG	10	$5.826981e-05$	84	$9.852721e-05$

This variational problem is well studied [17, 22] and it is known that the condition number $\kappa = \lambda_{max}/\lambda_{min} = O(\alpha N^2)$ and this estimate was used for the Nesterov scheme with fixed parameter. The following trends are well illustrated by this example.

- *When the condition number is modest, $N = 50$, steepest descent exhibits a linear rate of convergence and can be competitive with the Nesterov scheme.*
- *When the condition number is large, $N = 500$, the linear rate of convergence for the method of steepest descent becomes negligible and the algebraic rate $O(1/n)$ is observed. In this situation the Nesterov scheme with rate of $O(1/n^2)$ is superior.*
- *The sequence $\{f(\mathbf{u}_n)\}_{n=0}^{\infty}$ is not monotone decreasing for the Nesterov scheme.*
- *The Nesterov scheme with judiciously chosen fixed parameter does exhibit a linear rate of convergence with constant smaller than steepest descent. However, it is quite sensitive to the choice of parameter.*
- *When $\beta = 2$ the minima are smooth and well approximate the original signal where it is smooth; however, the approximation is poor elsewhere. Since jumps in contrast are essential features of images, $\beta = 2$ is a poor choice for debluring images. In Section 4.1 this example is considered with $\beta = 1$ which is known to capture such features with better fidelity [7, 20].*

1.1. Outline. The next section reviews the elementary properties of linear spaces and convex functions, and the notation, used in the sequel. Most of this material will be covered in any course on optimization, and may be skipped upon first reading. Section 3 considers the unconstrained optimization of smooth objective functions, and is the nucleus of the manuscript.

- This section begins with a discussion elucidating the interplay between convexity and descent criteria, and the Armijo rule in particular. The two lemmas, Lemma 3.1 and Lemma 3.2, establish the essential properties of first order algorithms used in the convergence proofs.
- Section 3.1 gives a proof of convergence for the method of steepest descent. The properties of the descent step established in the two lemmas provide a very succinct proof which

provides a "blue print" for the convergence proofs of the Nesterov schemes.

- Section 3.2 establishes the convergence properties of Nesterov's method. The proof closely follows that in Nesterov's original manuscript [16].
- Section 3.3 considers Nesterov's scheme with a fixed, but judiciously chosen, parameter. The convergence proof follows the same line of argument used for the steepest descent and original Nesterov algorithm and, granted the identity introduced in Exercise 3.8, could easily be posed as an exercise. The proof given here closely follows that in [6], and is substantially simpler than those appearing in [18, 5, 15].

Section 4 considers problems where the objective function $f : X \to \mathbb{R} \cup \{\infty\}$ my take extended values and/or may not be smooth. This provides the mathematical setting for problems involving L^1 minimization and/or constraints. This section starts with the analog of Lemma 3.1 which elucidates the interplay between convexity and descent criteria in the non–smooth setting. Upon substituting "Lemma 4.3" for "Lemma 3.1", proofs of convergence for the steepest descent, Nesterov, and fixed parameter Nesterov algorithms are then verbatim copies of the proofs for the unconstrained/smooth case, so are omitted. It is envisioned that this economy of presentation will accord discussion of structure over technical detail.

Section 4 finishes with two applications of the theory. Section 4.1 considers to the signal recovery problem of Example 1.3 with L^1 regularization ($\beta = 1$). Upon posing the problem in a "mixed formulation", and passing to the dual, convergence of the three algorithms then follow for this problem. Section 4.2 reviews Uzawa's algorithm for computing minima on convex sets of the form $U = \{x \in X \mid \phi_i(x) \leq 0,\ 1 \leq i \leq m\}$ for which the projection $P_U : X \to \mathbb{R}$ may not be computable.

The first order methods considered in this manuscript can be interpreted as explicit time stepping schemes to approximating the solution of related differential equations. This connection is briefly presented in the Appendix, and the descent properties of the discrete algorithms and solutions of the differential equations compared.

1.2. Pedagogy. Optimization is a core component of essentially every discipline, and for this reason courses with this title range from pure application to pure analysis. Due to this breadth, and the relative infancy of [16], an otherwise ideal text for a course may not cover Nesterov's algorithm; in fact, very few texts include it. For example, the classical texts by Dennis and Schnabel [10] and Ciarlet [8, 9] were written before 1983, and many more recent texts still do not discuss Nesterov's algorithm [4, 12, 14]. This note provides a concise introduction to Nesterov's algorithm that could supplement such texts.

In the following an effort has been made to stage the level of technical detail. For example, essentially no background beyond calculus and linear algebra is required for Section 3. If a discussion of non–smooth functions is beyond the scope of a course, proofs of convergence for the projected algorithms are obtained by replacing the operators $(I + \tau \partial f^0)^{-1}$ in Section 4 with the projections onto convex sets and sub–gradients with gradients. Similarly, the discussion of L^1 optimization in Section 4.1 and Uzawa's algorithm in Section 4.2 are the only places where duality is used, and the sub–gradient calculus is not required to follow this material.

2. Convex Optimization. The material in this section reviews the essential properties of convex functions that enter into the analysis of algorithms to compute their minima. Throughout this manuscript X will denote a Hilbert space and $f : X \to \mathbb{R}$ will be a convex function to be

minimized; minimizers will be denoted as $x_* \in X$.

2.1. Convex Functions. A function $f : X \to \mathbb{R}$ defined on (a convex subset of) a linear space is convex if

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y), \qquad x, y \in X, \quad \lambda \in [0,1].$$

First order methods for finding minima of f assume that the derivative is readily computable; that is, for each $x \in X$ there exists $f'(x) \in X'$ (the dual of X) for which

$$f(y) = f(x) + f'(x)(y-x) + o(\|y-x\|).$$

Below X will always be a Hilbert space with inner product $(.,.)$, in which case there exists $\nabla f(x) \in X$ (the gradient) such that $f'(x)(y) = (\nabla f(x), y)$. When $X = \mathbb{R}^n$ it is traditional to let $(.,.)$ be the canonical dot product; however, in many applications it is advantageous to use a different inner product (preconditioning). If $(x,y) = x^\top A y$ with $A \in \mathbb{R}^{n\times n}$ SPD, computing $\nabla f(x)$ involves solving a system the linear equations $A\nabla f(x) = f'(x)$, in which case a Cholesky decomposition would be pre–computed for efficiency.

LEMMA 2.1. *If $f : X \to \mathbb{R}$ is differentiable then the following are equivalent. For all x, $y \in X$,*

1. $f((1-\lambda)x + \lambda y) \le (1-\lambda)f(x) + \lambda f(y)$ *for all* $\lambda \in [0,1]$.
2. $(\nabla f(x), y - x) \le f(y) - f(x)$.
3. $(\nabla f(y) - \nabla f(x), y - x) \ge 0$.

All of the convergence results will require the gradient of f to be Lipschitz; that is, there exists $L \ge 0$ such that

$$\|\nabla f(y) - \nabla f(x)\| \le L\|y - x\|, \qquad x, y \in X.$$

If f is twice differentiable this is equivalent to[2] $D^2 f(x) \le LI$, and if f is convex the first order Taylor expansion is bounded on both sides,

$$f(x) + (\nabla f(x), y-x) \le f(y) \le f(x) + (\nabla f(x), y-x) + (L/2)\|y-x\|^2.$$

A differentiable function f is strongly convex with parameter $\alpha \ge 0$ if the function $x \mapsto f(x) - (\alpha/2)\|x\|^2$ is convex. If f is twice differentiable this is equivalent to $\alpha I \le D^2 f(x)$. Strong convexity strengthens the convexity inequalities of Lemma 2.1.

LEMMA 2.2. *If f is differentiable and $\alpha \ge 0$ the following are equivalent. For all x, $y \in X$,*

- $f\big(\lambda x + (1-\lambda)y\big) \le \lambda f(x) + (1-\lambda)f(y) - (\alpha/2)\lambda(1-\lambda)\|y-x\|^2$ *for all* $\lambda \in [0,1]$.
- $\big(\nabla f(x), y - x\big) \le f(y) - f(x) - (\alpha/2)\|y-x\|^2$.
- $\big(\nabla f(y) - \nabla f(x), y - x\big) \ge \alpha\|y-x\|^2$.

A strongly convex function f with Lipschitz gradient is bounded above and below by quadratic functions,

$$f(x) + (\nabla f(x), y-x) + (\alpha/2)\|y-x\|^2 \le f(y) \le f(x) + (\nabla f(x), y-x) + (L/2)\|y-x\|^2;$$

in particular, these functions have unique minima.

[2] If A, $B \in \mathbb{R}^{n\times n}$ are symmetric, we write $A \le B$ if $x^\top A x \le x^\top B x$ for all $x \in \mathbb{R}^n$

Functions such as $f(x) = |x|^\beta$ for $1 < \beta < \infty$ will not simultaneously have Lipschitz gradients and be strongly convex (unless $\beta = 2$). However, function values $\{f(x_n)\}_{m=0}^\infty$ computed with the steepest descent and Nesterov algorithms with the Armijo rule will always be bounded. In this situation assumptions on Lipschitz continuity of the derivative or strong convexity of f appearing in the statements of the theorems need only hold on sub–level sets of f; that is sets of the form $\{x \in X \mid f(x) \leq C\}$.

Convex functions lacking gradients, such as the absolute value $f(x) = |x|$, and functions $f : X \to \mathbb{R} \cup \{\infty\}$ taking extended values, such as the indicator function of a convex set $U \subset X$,

$$I_U(x) = \begin{cases} 0 & x \in U \\ \infty & x \notin U \end{cases},$$

arise frequently in applications. The domain of a function $f : X \to \mathbb{R} \cup \{\infty\}$ is $D(f) = \{x \in X \mid f(x) < \infty\}$, and f is *proper* if $D(f) \neq \emptyset$. Many of the usual results from calculus are available for these functions when sub–differentials

$$\partial f(x) = \{z \in X \mid (z, y - x) \leq f(y) - f(x),\ y \in X\} \subset X,$$

are substituted for gradients. If f is differentiable at x then the sub–gradient is the singleton set containing $\nabla f(x)$. The calculus for sub–differentials of convex functions is presented in [11, 19, 21]. The following lemma will be used in Section 4 below.

LEMMA 2.3. *Let X be a Hilbert space and $f : X \to \mathbb{R} \cup \{\infty\}$ be convex, proper, and lower semi–continuous. Then $(I + \tau\partial f) : X \to X$ is surjective for all $\tau > 0$, and the inverse is a contraction.*

2.2. Constraints & Duality. If $U \subset X$ is a closed convex subset of the Hilbert space X, the projection $P_U : X \to U$ is the function satisfying

$$P_U(x) \in U, \qquad \|x - P_U(x)\| \leq \|x - y\|, \qquad y \in U,$$

or equivalently

$$P_U(x) \in U, \qquad (P_U(x) - x, y - P_U(x)) \geq 0, \qquad y \in U.$$

The projection and indicator of U are related through

$$x \in (I + \tau\partial I_U)(z) \qquad \Leftrightarrow \qquad z = P_U(x), \qquad \tau > 0.$$

Formulae for the projections onto simple sets such as half spaces, hyperplanes, cubes, and balls are available; however, explicit formula are not available for projections onto sets of the form $U = \{x \in X \mid \phi_i(x) \leq 0,\ 1 \leq i \leq m\}$. Duality theory can be utilized to circumvent this difficulty. Given $f : X \to \mathbb{R}$, the Lagrangian $L : X \times [0, \infty)^m \to \mathbb{R}$ associated with problem (1.1) is[3]

$$L(x, \mu) = f(x) + \mu.\phi(x), \qquad \text{where} \qquad \phi(x) = \big(\phi_1(x), \ldots, \phi_m(x)\big) \in \mathbb{R}^m.$$

A saddle point of L is a point $(x_*, \mu_*) \in X \times [0, \infty)^m$ for which

$$\sup_{\mu \in [0,\infty)^m} \inf_{x \in X} L(x, \mu) = L(x_*, \mu_*) = \inf_{x \in X} \sup_{\mu \in [0,\infty)^m} L(x, \mu).$$

[3]The inner product of a, $b \in \mathbb{R}^m$ is denoted $a.b$.

Under appropriate hypotheses on the functions f and ϕ_i (see Theorem 4.9), saddle points exist and satisfy $\mu_*.\phi(x_*) = 0$,

$$x_* \in U, \qquad f(x_*) \leq f(y) \qquad y \in U, \qquad \text{and} \qquad 0 \in \partial f(x_*) + \sum_{i=1}^{m} \mu_i \nabla \phi_i(x_*).$$

The dual variables μ_i are *generalized Lagrange, or KKT, multipliers*. If $\phi = (a, x) + \alpha$ is affine, then both $\pm\phi$ are convex, so affine equality constraints are can be implemented as $\phi(x) \leq 0$ and $-\phi(x) \leq 0$ in which case the corresponding KKT multipliers combine to form a Lagrange multiplier,

$$\mu_+ \nabla \phi(x) + \mu_- \nabla(-\phi)(x) = (\mu_+ - \mu_-)\nabla\phi(x) \equiv \lambda \nabla \phi(x), \qquad \lambda \in \mathbb{R}.$$

The dual function

$$g(\mu) = \inf_{x \in X} L(x, \mu) = \inf_{x \in X} \left(f(x) + \mu.\phi(x) \right), \qquad g : [0, \infty)^m \to \mathbb{R} \cup \{-\infty\},$$

is concave. Since projections onto $P_+ : \mathbb{R}^m \to [0, \infty)^m$ are easily evaluated, one strategy (Uzawa's algorithm) for solving (1.1) is to find a maximizer μ_* of g, using, for example, projected steepest descent, and to then compute an (unconstrained) minimizer of $x \mapsto f(x) + \mu_*.\phi(x)$.

If $x = \arg\min_x f(x) + \mu.\phi(x)$ and (μ_*, x_*) a saddle point then

$$g(\mu) = f(x) + \mu.\phi(x) \leq g(\mu_*) = f(x_*) + \mu_*.\phi(x_*) = f(x_*),$$

so that $f(x) - f(x_*) = -\mu.\phi(x)$. The right hand side is the *duality gap* and in a computational context is an explicitly computable error indicator.

2.3. First Order Methods & Complexity. We finish this overview with a summary of the complexity of first order schemes. These results highlight the essential role of the smoothness properties introduced above. Proofs of the following theorem can be found in [5, 15].

THEOREM 2.4. *Let X be an infinite dimensional Hilbert space and set $x_0 = 0$.*

- *There exists a convex function $f : X \to \mathbb{R}$ with Lipschitz gradient and minima $f(x_*) > -\infty$ such that for any sequence satisfying*

$$x_{i+1} \in \text{Span}\{\nabla f(x_0), \nabla f(x_1), \ldots \nabla f(x_i)\}, \qquad i = 0, 1, 2, \ldots.$$

  there holds

$$\min_{1 \leq i \leq n} f(x_i) - f(x_*) \geq \frac{3L}{32} \frac{\|x_1 - x_*\|^2}{(n+1)^2},$$

  where L is the Lipschitz constant of the gradient.
- *There exists a strongly convex function $f : X \to \mathbb{R}$ with constant $\alpha > 0$ having Lipschitz gradient and minima $f(x_*) > -\infty$ such that for any sequence satisfying*

$$x_{i+1} \in \text{Span}\{\nabla f(x_0), \nabla f(x_1), \ldots \nabla f(x_i)\}, \qquad i = 0, 1, 2, \ldots.$$

  there holds

$$\min_{1 \leq i \leq n} f(x_i) - f(x_*) \geq \frac{\alpha}{2} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2(n-1)} \|x_1 - x_*\|^2,$$

  where $\kappa = L/\alpha$ and L is the Lipschitz constant of the gradient.

3. Unconstrained Problem. Given a convex function $f : X \to \mathbb{R}$ defined on a Hilbert space X, this section considers the unconstrained minimization problem:

$$x_* \in X, \qquad f(x_*) \le f(x), \qquad x \in X.$$

Both the steepest descent and Nesterov's algorithm involve updates of the form $x = y - \tau\nabla f(y)$ in combination with a descent criteria. The essential estimate in the convergence proofs uses the following amalgamation of these two ingredients.

LEMMA 3.1. *Let X be a Hilbert space and $f : X \to \mathbb{R}$ be differentiable, strongly convex with parameter $\alpha \ge 0$, and let x, $y \in X$, and $\tau > 0$. If*

$$x = y - \tau\nabla f(y) \qquad \textit{and} \qquad f(x) \le f(y) + (\nabla f(y), x - y) + \frac{1}{2\tau}\|y - x\|^2,$$

then

$$2\tau f(x) + \|x - z\|^2 \le 2\tau f(z) + (1 - \alpha\tau)\|y - z\|^2, \qquad z \in X. \tag{3.1}$$

Writing $x - y = -\tau\nabla f(y)$ in the descent condition gives the equivalence,

$$f(x) \le f(y) + (\nabla f(y), x - y) + \frac{1}{2\tau}\|y - x\|^2 \quad \Leftrightarrow \quad (\tau/2)\|\nabla f(y)\|^2 \le f(y) - f(x). \tag{3.2}$$

Proof. Subtracting z from both sides of the equation for x and taking the inner product with $x - z$ gives

$$\|x - z\|^2 = (y - z, x - z) + \tau\left(\nabla f(y), z - x\right),$$

and using the identity $2(a, b) = \|a\|^2 - \|b - a\|^2 + \|b\|^2$ shows

$$\|x - z\|^2 + \|y - x\|^2 = \|y - z\|^2 + 2\tau\left(\nabla f(y), z - x\right).$$

To estimate the last term, write $z - x = (z - y) + (y - x)$ and bound each summand separately,

- Strong convexity of f shows $\big(\nabla f(y), z - y\big) \le f(z) - f(y) - (\alpha/2)\|z - y\|^2$.
- The descent condition gives $\big(\nabla f(y), y - x\big) \le f(y) - f(x) + (1/2\tau)\|y - x\|^2$.

The lemma follows upon substituting these bounds into the above. □

The following lemma shows that step sizes satisfying the descent condition can be computed using bisection (backtracking) with a guaranteed lower bound.

LEMMA 3.2 (Armijo Rule). *Let $f : X \to \mathbb{R}$ have Lipschitz gradient with constant $L > 0$, then the descent condition*

$$f(x) \le f(y) + (\nabla f(y), x - y) + \frac{1}{2\tau}\|y - x\|^2 \tag{3.3}$$

is satisfied whenever $\tau \le 1/L$. In particular, if $\tau = 1/2^m$ where m is the smallest integer for which the descent condition is satisfied then $1/2L \le \tau$.

Proof. When the gradient of f is Lipschitz, the fundamental theorem of calculus gives

$$f(x) \le f(y) + (\nabla f(y), x - y) + (L/2)\|y - x\|^2. \tag{3.4}$$

Then,

- The descent criteria is satisfied if $L \le 1/\tau$; that is, $\tau \le 1/L$.
- If $\tau = 1/2^m$ where m is the smallest integer for which the descent condition holds, then doubling τ violates the descent criteria in which case $L \ge 1/2\tau$; that is, $\tau > 1/2L$.

□

EXERCISE 3.3. *Starting with the calculation,*

$$\begin{aligned} f(x) &= f(y) + \int_0^1 \big(\nabla f((1-t)x + ty), y - x\big)\, dt \\ &= f(y) + \big(\nabla f(y), y - x\big) + \int_0^1 \big(\nabla f((1-t)x + ty) - \nabla f(x), y - x\big)\, dt, \end{aligned}$$

verify that equation (3.4) holds when the gradient of f satisfies $\|\nabla f(z) - \nabla f(x)\| \le L\|z - x\|$.

3.1. Steepest Descent. We provide a proof of convergence of the steepest descent method for finding a minima of f which is both simple and naturally extends to proof of convergence of Nesterov's method.

THEOREM 3.4. *Let X be a Hilbert space and $f : X \to \mathbb{R}$ be convex with Lipschitz gradient with constant $L > 0$ and strongly convex with constant $\alpha \ge 0$. Let $x_0 \in X$ and*

$$x_{i+1} = x_i - \tau_i \nabla f(x_i), \qquad i = 0, 1, 2, \ldots,$$

where the step size $\tau_i > 0$ is determined by the Armijo rule (3.3). Assume there exists $x_* \in X$ for which $f_* \equiv f(x_*) \le f(x)$ for all $x \in X$. Then

$$f(x_n) - f(x_*) \le \frac{L\|x_0 - x_*\|^2}{n} \qquad \text{and} \qquad \|x_n - x_*\|^2 \le (1 - \alpha/2L)^n \|x_0 - x_*\|^2.$$

In addition, $\|\nabla f(x_n)\|^2 \le 4L\big(f(x_n) - f_*\big)$ and $f(x_n) - f_* \le (L/2)\|x_n - x_*\|^2$.

Proof. Setting $(x, y, z) = (x_{i+1}, x_i, x_i)$ in Lemma 3.1 shows $f(x_{i+1}) \le f(x_i)$.

Next, set $(x, y, z) = (x_{i+1}, x_i, x_*)$ to get

$$2\tau_i f(x_{i+1}) + \|x_{i+1} - x_*\|^2 \le 2\tau_i f(x_*) + (1 - \alpha\tau_i)\|x_i - x_*\|^2.$$

The bounds on $f(x_n) - f_*$ and $\|x_n - x_*\|$ follow since Lemma 3.2 guarantees $\tau_i \ge 1/2L$. The bound on the gradient follows from the descent condition (3.2), and the additional bound on $f(x_n) - f_*$ follows from (3.4) (with $y = x_*$) since $\nabla f(x_*) = 0$. □

Commentary: When used in combination with the Armijo rule the method of steepest descent is parameter free in the sense that no knowledge of the strong convexity or Lipschitz constants are required as input. In particular, the algorithm will converge at a linear rate if the f is strongly convex, and algebraically if it is not.

3.2. Nesterov's Algorithm. This section considers Nesterov's scheme (1.2) with step sizes $\tau_i > 0$ determined by the Armijo rule (3.3) with the additional requirement that $\tau_{i+1} \le \tau_i$. Note that this additional condition does not alter the lower bound $1/2L \le \tau_i$ in Lemma 3.2.

If f is convex, but not strongly convex, the essential inequality to be satisfied by the sequence $\{\lambda_i\}_{i=0}^{\infty}$ is $\lambda_i(\lambda_i - 1) \le \lambda_{i-1}^2$. The sequence in equation (1.3) satisfies this relation with equality, so is optimal in the sense that λ_i is as large as possible.

EXERCISE 3.5. *Let $\{\lambda_i\}_{i=0}^{\infty}$ be the sequence in equation (1.3).*

1. *For $i \geq 1$ show that*

$$\lambda_i \geq 1, \qquad \lambda_{i-1}^2 = (\lambda_i - 1)\lambda_i, \qquad \text{and} \qquad i/2 \leq \lambda_{i-1} \leq i - 1.$$

2. *Show that the ratio $(\lambda_i - 1)/\lambda_{i+1}$ is monotone increasing and converges to 1.*
3. *Show that the same properties hold for the sequence $\lambda_0 = 0$ and $\lambda_i = (i+1)/2$ for $i \geq 1$.*

Frequently Nestorov's method is presented with this choice of parameters since $(\lambda_i - 1)/\lambda_{i+1} = (i-1)/(i+2)$ is explicitly computable.

To establish convergence we mimic the proof presented for the steepest descent method.

THEOREM 3.6. *Let X be a Hilbert space and $f : X \to \mathbb{R}$ be convex with Lipschitz gradient with constant $L > 0$. Let $x_0 = y_0 \in X$ and*

$$x_{i+1} = y_i - \tau_i \nabla f(y_i), \qquad y_{i+1} = x_{i+1} + \frac{\lambda_i - 1}{\lambda_{i+1}}(x_{i+1} - x_i), \qquad i = 0, 1, 2, \ldots,$$

where the step size $\tau_i > 0$ is determined by the Armijo rule (3.3) with the additional requirement that $\tau_i \leq \tau_{i-1}$. Assume that the parameters $\{\lambda_i\}_{i=0}^{\infty}$ satisfy

$$\lambda_0 = 0, \qquad \text{and} \qquad \lambda_i \geq 1, \qquad \lambda_i(\lambda_i - 1) \leq \lambda_{i-1}^2, \qquad i = 1, 2, \ldots.$$

and that there exists $x_* \in X$ for which $f_* \equiv f(x_*) \leq f(x)$ for all $x \in X$. Then

$$f(x_n) - f_* \leq L\frac{\|x_0 - x_*\|^2}{\lambda_{n-1}^2}, \qquad \text{and} \qquad \|\nabla f(y_n)\| \leq \frac{4L\|x_0 - x_*\|}{\lambda_n};$$

in addtion, $\|\lambda_n y_n - (\lambda_n - 1)x_n - x_*\| \leq \|x_0 - x_*\|$.

Proof. Set $(x, y, z) = \big(x_{i+1}, y_i, (1 - 1/\lambda_i)x_i + (1/\lambda_i)x_*\big)$ in Lemma 3.1 with $\alpha = 0$, to get

$$2\tau_i f(x_{i+1}) + (1/\lambda_i)^2\|z_{i+1}\|^2 \leq 2\tau_i f\left((1 - 1/\lambda_i)x_i + (1/\lambda_i)x_*\right) + (1/\lambda_i^2)\|z_i\|^2,$$

where $z_i \equiv \lambda_i y_i - (\lambda_i - 1)x_i - x_*$, and the update formula for y_{i+1} was used to get the formula for z_{i+1},

$$z_{i+1} \equiv \lambda_{i+1}y_{i+1} - (\lambda_{i+1} - 1)x_{i+1} - x_* = \lambda_i x_{i+1} - (\lambda_i - 1)x_i - x_*.$$

From the convexity of f it follows that

$$2\tau_i f(x_{i+1}) + (1/\lambda_i)^2\|z_{i+1}\|^2 \leq 2\tau_i\left((1 - 1/\lambda_i)f(x_i) + (1/\lambda_i)f(x_*)\right) + (1/\lambda_i^2)\|z_i\|^2,$$

which can be rearranged to give

$$2\tau_i\lambda_i^2\left(f(x_{i+1}) - f(x_*)\right) + \|z_{i+1}\|^2 \leq 2\tau_i\lambda_i(\lambda_i - 1)\left(f(x_i) - f(x_*)\right) + \|z_i\|^2. \tag{3.5}$$

The hypotheses guarantee that $\tau_i\lambda_i(\lambda_i - 1) \leq \tau_{i-1}\lambda_{i-1}^2$, and $\lambda_0 = 0$, so the recursion telescopes to give

$$2\tau_n\lambda_n^2\left(f(x_{n+1}) - f(x_*)\right) + \|z_{n+1}\|^2 \leq \|z_0\|^2 = \|x_0 - x_*\|^2.$$

The bound upon z_n is direct, and the bound upon $f(x_{n+1}) - f(x_*)$ follows since Lemma 3.2 guarantees $\tau_n \geq 1/2L$. The gradient bound follows from the identity $z_n - z_{n+1} = \lambda_n\tau_n\nabla f(y_n)$. ∎

EXERCISE 3.7. *If $a \geq 2$ show that the sequence with $\lambda_0 = 0$ and $\lambda_i = (i + a - 1)/a$ for $i \geq 1$ satisfies*

$$\lambda_{i-1}^2 - \lambda_i(\lambda_i - 1) \geq \left(\frac{a-2}{a^2}\right) i.$$

If $a > 2$ and this sequence is used for the Nesterov scheme, show $\sum_{i=1}^{\infty} i\left(f(x_i) - f(x_*)\right) < \infty$. Conclude that $\liminf_{n\to\infty} n^2 \log(n)\left(f(x_n) - f(x_*)\right) = 0$.

Commentary: When paired with the Armijo rule Nesterov's algorithm is parameter free, and is optimal in the absence of strong convexity assumptions. It is natural to ask if, like steepest descent, a better rate of convergence is achieved when f is strongly convex. This does not appear to be the case. If f is strongly convex with parameter $\alpha > 0$, the recurrence relation in the proof becomes

$$\begin{aligned} 2\tau_i\lambda_i^2\left(f(x_{i+1}) - f(x_*)\right) + \|z_{i+1}\|^2 \\ \leq 2\tau_i\lambda_i(\lambda_i - 1)\left(f(x_i) - f(x_*)\right) + (1 - \alpha\tau_i)\|z_i\|^2 - \alpha\tau_i(\lambda_i - 1)\|x_i - x_*\|^2. \end{aligned}$$

In general $\lambda_i(\lambda_i - 1) \not\leq (1 - \alpha\tau_i)\lambda_{i-1}^2$, so linear convergence does not follow. Note though, this inequality does hold if $\lambda_{i-1} = \lambda_i \leq 1/\alpha\tau_i$. Thus if an estimate of α is available it may be advantageous to fix λ_i to be constant once it attains this value. The next section shows that this is the case; in fact, a linear rate with constant $(1 - \sqrt{\alpha\tau_i})$ is possible.

3.3. Nesterov's Algorithm with Fixed Parameter. We consider the Nesterov scheme

$$x_{i+1} = y_i - \tau_i\nabla f(y_i), \qquad y_{i+1} = x_{i+1} + \frac{\lambda - 1}{\lambda + 1}(x_{i+1} - x_i), \qquad i = 1, 2, \ldots,$$

with fixed parameter[4]. Letting

$$\hat{z} = (\lambda + 1)y - \lambda x - x_* \qquad \text{and} \qquad z = \lambda y - (\lambda - 1)x - x_*,$$

the convergence proof uses the following identity relating these two quantities,

$$(1 - 1/\lambda^2)\|z\|^2 = (1 - 1/\lambda)\|\hat{z}\|^2 + (1/\lambda^2)(\lambda - 1)\|x - x_*\|^2 - (\lambda - 1/\lambda)\|y - x\|^2. \tag{3.6}$$

EXERCISE 3.8. *Figure 3.1 shows the results of a* MAPLE *calculation. Show that identities involving scalars and the squares of norms in a real Hilbert space can be validated using the corresponding identities for polynomials. In particular, the* MAPLE *calculation proves identity (3.6).*

THEOREM 3.9. *Let X be a Hilbert space and let $f : X \to \mathbb{R}$ be strongly convex with constant $\alpha > 0$ and have Lipschitz gradient with constant $L > 0$ and fix $\lambda \geq \sqrt{2L/\alpha} > 1$. Let $x_0 = y_0 \in X$ and*

$$x_{i+1} = y_i - \tau_i\nabla f(y_i), \qquad y_{i+1} = x_{i+1} + \frac{\lambda - 1}{\lambda + 1}(x_{i+1} - x_i), \qquad i = 0, 1, 2, \ldots,$$

where the step size $\tau_i > 0$ is determined by the Armijo rule (3.3) with the additional requirement that $\tau_i \leq \tau_{i-1}$.

[4]Setting $\lambda = (\tilde{\lambda} + 1)/2$ in equation (1.3) gives $(\lambda - 1)/\lambda = (\tilde{\lambda} - 1)/(\tilde{\lambda} + 1)$. This change of variable simplifies some of the formula in the convergence proof.

```
> restart:
  zhat := (lambda+1)*y - lambda*x - xstar;
  zz := lambda*y - (lambda-1)*x - xstar;
  factor((1-1/lambda^2)*zz^2 - (1-1/lambda)*(zhat^2+(1/lambda)*(x-xstar)^2));
```

$$zhat := (\lambda+1)\,y-\lambda\,x-xstar$$

$$zz := \lambda\,y-(\lambda-1)\,x-xstar$$

$$-\frac{(x-y)^2\,(\lambda-1)\,(\lambda+1)}{\lambda}$$

FIG. 3.1. *Proof of the identity (3.6).*

Let $x_* \in X$ satisfy $f_* \equiv f(x_*) \le f(x)$ for all $x \in X$, then

$$(\lambda^2/L)(f(x_n)-f_*)+\|(\lambda+1)y_n-\lambda x_n-x_*\|^2 \le (1-1/\lambda)^n\left(2\tau_0\lambda^2(f(x_0)-f_*)+\|x_0-x_*\|^2\right).$$

In addition,

$$\|x_n-x_*\|^2 \le (2/\alpha)(f(x_n)-f_*), \qquad \text{and} \qquad \|\nabla f(y_n)\|^2 \le (8L^2/\alpha)(f(x_{n+1})-f_*).$$

Proof. Set $(x,y,z)=\left(x_{i+1},y_i,(1-1/\lambda)x_i+(1/\lambda)x_*\right)$ in Lemma 3.1, to get

$$2\tau_i f(x_{i+1})+(1/\lambda)^2\|\hat{z}_{i+1}\|^2 \le 2\tau_i f\left((1-1/\lambda)x_i+(1/\lambda)x_*\right)+(1/\lambda^2)(1-\alpha\tau_i)\|z_i\|^2,$$

where

$$\hat{z}_i \equiv (\lambda+1)y_i-\lambda x_i-x_* \qquad \text{and} \qquad z_i \equiv \lambda y_i-(\lambda-1)x_i-x_*.$$

The update formula for y_{i+1} was used to get the alternative formula for $\hat{z}_{i+1}$,

$$\hat{z}_{i+1} \equiv (\lambda+1)y_{i+1}-\lambda x_{i+1}-x_* = \lambda x_{i+1}-(\lambda-1)x_i-x_*.$$

From the strong convexity of f it follows that

$$2\tau_i\lambda^2 f(x_{i+1})+\|\hat{z}_{i+1}\|^2 \le 2\tau_i\lambda(\lambda-1)f(x_i)+2\tau_i\lambda f(x_*)-\alpha\tau_i(\lambda-1)\|x_i-x_*\|^2+(1-\alpha\tau_i)\|z_i\|^2,$$

which can be rearranged to give

$$2\tau_i\lambda^2\left(f(x_{i+1})-f(x_*)\right)+\|\hat{z}_{i+1}\|^2 \le 2\tau_i\lambda(\lambda-1)\left(f(x_i)-f(x_*)\right)-\alpha\tau_i(\lambda-1)\|x_i-x_*\|^2+(1-\alpha\tau_i)\|z_i\|^2.$$

Writing the last term as $(1/\lambda^2-\alpha\tau_i)\|z_i\|^2+(1-1/\lambda^2)\|z_i\|^2$ and using identity (3.6) to bound the second summand shows

$$\begin{aligned} 2\tau_i\lambda^2\left(f(x_{i+1})-f(x_*)\right)+\|\hat{z}_{i+1}\|^2 \;\le\; & 2\tau_i\lambda(\lambda-1)\left(f(x_i)-f(x_*)\right)+(1-1/\lambda)\|\hat{z}_i\|^2-(\lambda-1/\lambda)\|y_i-x_i\|^2 \\ & +(1/\lambda^2-\alpha\tau_i)(\lambda-1)\left(\|x_i-x_*\|^2+\|z_i\|^2\right). \end{aligned} \tag{3.7}$$

By hypothesis $1/\lambda^2 \le \alpha/2L \le \alpha\tau_i$, the latter following from Lemma 3.2, and $\tau_i \le \tau_{i-1}$, so

$$2\tau_i\lambda^2\left(f(x_{i+1})-f(x_*)\right)+\|\hat{z}_{i+1}\|^2+(\lambda-1/\lambda)\|y_i-x_i\|^2 \le (1-1/\lambda)\left(2\tau_{i-1}\lambda^2\left(f(x_i)-f(x_*)\right)+\|\hat{z}_i\|^2\right).$$

The bound on $\|x_n-x_*\|^2$ follow from strong convexity,

$$0=\left(\nabla f(x_*),x_n-x_*\right) \le f(x_n)-f(x_*)-(\alpha/2)\|x_n-x_*\|^2, \quad \text{so} \quad \|x_n-x_*\|^2 \le (2/\alpha)\left(f(x_n)-f(x_*)\right).$$

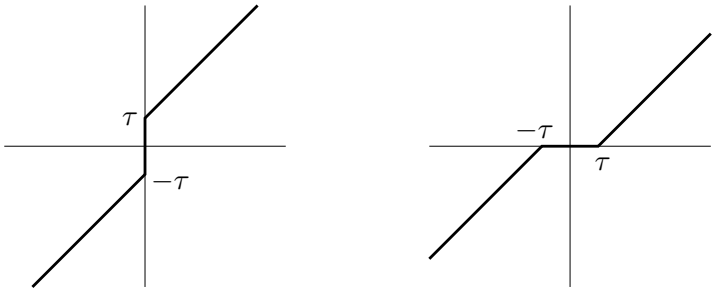

FIG. 4.1. *$I + \tau\partial|.|$ and its inverse.*

Next, monotonicity of the gradient shows

$$0 \le \big(\nabla f(y_n) - \nabla f(x_*), y_n - x_*\big) \quad \text{so that} \quad \big(\nabla f(y_n), y_n - x_{n+1}\big) \le \big(\nabla f(y_n), x_* - x_{n+1}\big)\,.$$

The update formula, $\tau_n \nabla f(y_n) = y_n - x_{n+1}$, and Cauchy Schwarz inequality give

$$\tau_n \|\nabla f(y_n)\| \le \|x_{n+1} - x_*\| \le \big((2/\alpha)(f(x_{n+1}) - f(x_*))\big)^{1/2}\,,$$

and the gradient bound follows since $\tau_n \ge 1/2L$. □

Commentary: Computational examples show that the hypothesis $\lambda \ge \sqrt{2L/\alpha}$ is essential. The development of parameter free modifications of Nesterov's algorithm exhibiting linear convergence when $\alpha > 0$, and rate $O(1/n^2)$ otherwise, is still an active area of investigation [13].

4. Constrained and Non–Smooth Optimization. The objective function $f : X \to \mathbb{R} \cup \{\infty\}$ is often the sum $f = f^0 + f^1$ of a smooth convex function f^1 and non–smooth convex part f^0 for which the inverse of $(I + \tau\partial f^0) : X \to X$ can be readily evaluated. It is then possible to treat the non–smooth part implicitly; that is, replacing the descent step

$$x = y - \tau\nabla f(y) \simeq y - \tau\left(\nabla f^0(y) + \nabla f^1(y)\right) \qquad \text{with} \qquad x \in y - \tau\left(\partial f^0(x) + \nabla f^1(y)\right).$$

This may be alternatively expressed as

$$(I + \tau\partial f^0)(x) \ni y - \tau\nabla f^1(y), \qquad \text{or} \qquad x = (I + \tau\partial f^0)^{-1}\left(y - \tau\nabla f^1(y)\right),$$

or equivalently,

$$x = \operatorname*{arg\,min}_{x\in X}\left\{\tau f^0(x) + \tau\left(\nabla f^1(y), x\right) + (1/2)\|x - y\|^2\right\}.$$

EXERCISE 4.1. *(Constrained Optimization) If $U \subset X$ is a closed convex subset of the Hilbert space X, let $I_U : X \to \mathbb{R}\cup\{\infty\}$ denote the indicator function of U. Show that $(I+\tau\partial I_U)^{-1}(z) = P_U(z)$ where $P_U : X \to U$ is the projection.*

EXERCISE 4.2. *(L^1 Minimization)*

1. *If $X = \mathbb{R}$ and $f(x) = |x|$ and $\tau > 0$, show that (see Figure 4.1)*

$$(I + \tau\partial f)(x) = \begin{cases} x + \tau & 0 < x \\ [-\tau, \tau] & x = 0 \\ x - \tau & x < 0 \end{cases} \qquad \textit{and} \qquad (I + \tau\partial f)^{-1}(y) = \begin{cases} y - \tau & 0 \le y \\ 0 & -\tau \le y \le \tau \\ y + \tau & y \le 0 \end{cases}$$

2. *If $X = \mathbb{R}^n$ and $f(x) = |x|_{\ell^1} \equiv \sum_{i=1}^n |x_i|$, show that $(I + \tau\partial f)^{-1}(y)_i = (I + \tau\partial|.|)^{-1}(y_i)$; that is, the inverse of $I + \tau\partial f$ can be computed compoentwise.*
3. *If $X = L^2(\Omega)$ and $f(x) = \|x\|_{L^1(\Omega)}$ if $x \in L^2(\Omega) \cap L^1(\Omega)$ and infinity otherwise, show that $(I + \tau\partial f^0)^{-1}(y)(\omega) = (I + \tau\partial|.|)^{-1}(y(\omega))$ for $\omega \in \Omega$.*

The convergence proofs of the steepest descent and Nesterov algorithms in the previous section each started by invoking an instance of Lemma 3.1, and the remainder of the proof involved routine convexity arguments. Using the following analog of Lemma 3.1, proofs of convergence in the current context are identical to those in the previous section.

LEMMA 4.3. *Let X be a Hilbert space and $f : X \to \mathbb{R} \cup \{\infty\}$ be the sum of two convex functions $f = f^0 + f^1$ with f^0 proper, convex, and lower semi–continuous, and f^1 differentiable and strongly convex with constant $\alpha \geq 0$. For $y \in X$ and $\tau > 0$ set*

$$x = (I+\tau\partial f^0)^{-1}\left(y - \tau\nabla f^1(y)\right), \qquad \text{and suppose} \qquad f^1(x) \leq f^1(y) + \big(\nabla f^1(y), x - y\big) + 1/(2\tau)\|x - y\|^2,$$

then

$$2\tau f(x) + \|x - z\|^2 \leq 2\tau f(z) + (1 - \alpha\tau)\|y - z\|^2, \qquad z \in X.$$

Proof. The function $\Phi(z) = \tau f^0(z) + \tau\left(\nabla f^1(y), z\right) + (1/2)\|z - y\|^2$ is strongly convex with unit parameter, and since x minimizes Φ it follows that

$$\Phi(x) + (1/2)\|x - z\|^2 \leq \Phi(z), \qquad z \in X.$$

Expanding this shows

$$\tau f^0(x) + \tau\left(\nabla f^1(y), x\right) + (1/2)\|x - y\|^2 + (1/2)\|x - z\|^2 \leq \tau f^0(z) + \tau\left(\nabla f^1(y), z\right) + (1/2)\|y - z\|^2,$$

and rearranging this gives

$$2\tau f^0(x) + \|x - y\|^2 + \|x - z\|^2 \leq 2\tau f^0(z) + 2\tau\left(\nabla f^1(y), z - x\right) + \|y - z\|^2.$$

Writing $z - x = (z - y) + (y - x)$ the corresponding terms in the inner product are bounded via

- Strong convexity of f^1:

$$\big(\nabla f^1(y), z - y\big) \leq f^1(z) - f^1(y) - (\alpha/2)\|y - z\|^2.$$

- The descent criteria:

$$\big(\nabla f^1(y), y - x\big) \leq f^1(y) - f^1(x) + (1/2\tau)\|x - y\|^2.$$

Combining the above completes the proof. □

Using this lemma in place of Lemma 3.1, convergence proofs for the steepest descent and Nesterov algorithms in this context follow *mutatis mutandis* as in the smooth case. Note though that the equivalence (3.2) no longer holds, so bounds upon $\nabla f(y)$ (which need not exist) no longer follow from the descent condition (3.3); putting $z = y$ in the Lemma shows $\|y - x\|^2 \leq 2\tau\left(f(y) - f(x)\right)$ instead.

THEOREM 4.4. *Let X be a Hilbert space and $f : X \to \mathbb{R} \cup \{\infty\}$ be the sum of two convex functions $f = f^0 + f^1$ with f^0 proper, convex, and lower semi–continuous, and f^1 having Lipschitz gradient with constant $L > 0$ and strongly convex with constant $\alpha \geq 0$ on $D(f^0)$. Let $x_0 \in X$ and*

$$x_{i+1} + \tau_i\partial f^0(x_{i+1}) \ni x_i - \tau_i\nabla f^1(x_i), \qquad i = 0, 1, 2, \ldots.$$

with step size $\tau_i = 1/2^m$ *where* m *is the smallest integer for which*

$$f^1(x_{i+1}) \le f^1(x_i) + \big(\nabla f^1(x_i), x_{i+1} - x_i\big) + 1/(2\tau_i)\|x_{i+1} - x_i\|^2.$$

Assume there exists $x_* \in X$ *for which* $f_* \equiv f(x_*) \le f(x)$ *for all* $x \in X$. *Then*

$$f(x_n) - f(x_*) \le \frac{L\|x_0 - x_*\|^2}{n}, \quad \text{and} \quad \|x_n - x_*\|^2 \le (1 - \alpha/2L)^n \|x_0 - x_*\|^2.$$

In addition, $\|x_{n+1} - x_n\|^2 \le 2\tau_n \big(f(x_{n+1}) - f(x_*)\big)$.

One difference between the method of steepest descent and the Nesterov algorithm is that in the former the gradient of f^1 is always evaluated at points $x_i \in D(f^0)$; however, the extrapolation step of the Nesterov algorithm may produce points $y_i \notin D(f^0)$, so we assume $D(f^1) = X$.

THEOREM 4.5. *Let* X *be a Hilbert space and* $f : X \to \mathbb{R} \cup \{\infty\}$ *be the sum of two convex functions* $f = f^0 + f^1$ *with* f^0 *proper, convex, and lower semi–continuous, and* $f^1 : X \to \mathbb{R}$ *having Lipschitz gradient with constant* $L > 0$. *Let* $x_0 = y_0 \in X$ *and*

$$x_{i+1} + \tau_i \partial f^0(x_{i+1}) \ni y_i - \tau_i \nabla f^1(y_i), \quad y_{i+1} = x_{i+1} + \frac{\lambda_i - 1}{\lambda_{i+1}}(x_{i+1} - x_i), \qquad i = 0, 1, 2, \ldots,$$

where $\tau_i = \min(\tau_{i-1}, 1/2^m)$ *with* m *the smallest integer for which*

$$f^1(x_{i+1}) \le f^1(y_i) + \big(\nabla f^1(y_i), x_{i+1} - y_i\big) + 1/(2\tau_i)\|x_{i+1} - y_i\|^2,$$

and $\{\lambda_i\}_{i=0}^{\infty}$ *is a sequence satisfying*

$$\lambda_0 = 0, \qquad \lambda_1 = 1, \qquad \lambda_i(\lambda_i - 1) \le \lambda_{i-1}^2, \qquad i = 1, 2, \ldots$$

Assume there exists $x_* \in X$ *for which* $f_* \equiv f(x_*) \le f(x)$ *for all* $x \in X$. *Then*

$$f(x_{n+1}) - f(x_*) \le \frac{L\|x_0 - x_*\|^2}{\lambda_n^2}, \qquad \text{and} \qquad \|\lambda_n y_n - (\lambda_n - 1)x_n - x_*\| \le \|x_0 - x_*\|.$$

THEOREM 4.6. *Let* X *be a Hilbert space and* $f : X \to \mathbb{R} \cup \{\infty\}$ *be the sum of two convex functions* $f = f^0 + f^1$ *with* f^0 *proper, convex, and lower semi–continuous, and* f^1, *having Lipschitz gradient with constant* $L > 0$ *and strongly convex with constant* $\alpha \ge 0$, *and fix* $\lambda \ge \sqrt{2L/\alpha} > 1$. *Let* $x_0 = y_0 \in X$ *and*

$$x_{i+1} + \tau_i \partial f^0(x_{i+1}) \ni y_i - \tau_i \nabla f^1(y_i), \qquad y_{i+1} = x_{i+1} + \frac{\lambda - 1}{\lambda + 1}(x_{i+1} - x_i), \qquad i = 0, 1, 2, \ldots,$$

where $\tau_i = \min(\tau_{i-1}, 1/2^m)$ *with* m *the smallest integer for which*

$$f^1(x_{i+1}) \le f^1(y_i) + \big(\nabla f^1(y_i), x_{i+1} - y_i\big) + 1/(2\tau_i)\|x_{i+1} - y_i\|^2.$$

Let $x_* \in X$ *satisfy* $f_* \equiv f(x_*) \le f(x)$ *for all* $x \in X$, *then*

$$(\lambda^2/L)(f(x_n) - f_*) + \|(\lambda + 1)y_n - \lambda x_n - x_*\|^2 \le (1 - 1/\lambda)^n \left(2\tau_0 \lambda^2 (f_0 - f_*) + \|x_0 - x_*\|^2\right).$$

In addition, $\|x_n - x_*\|^2 \le (2/\alpha)(f(x_n) - f_*)$.

4.1. L^1 Minimization. To illustrate the interplay between constraints and duality we consider the signal reconstruction problem from Example 1.3 with parameter $\beta = 1$,

$$f(u) = \int_0^1 \left((1/2)(u - \hat{u})^2 + \alpha|u'|\right).$$

The discrete approximation this problem is

$$f(\mathbf{u}) = (h/4)\left((u_0 - \hat{u}_0)^2 + (u_N - \hat{u}_N)^2\right) + \sum_{i=1}^{N-1}(h/2)(u_i - \hat{u}_i)^2 + \alpha\sum_{i=1}^{N}|u_i - u_{i-1}|.$$

While the sub-gradient of the last term is readily computable, it's inverse is not, which motivates the following "mixed" formulation. Set $p_i = u_i - u_{i-1}$ and $U = \{(\mathbf{u}, \mathbf{p}) \mid p_i - u_i + u_{i-1} = 0\} \subset \mathbb{R}^{2N+1}$, and find

$$(\mathbf{u}, \mathbf{p}) \in U \quad \text{such that} \quad f^1(\mathbf{u}) + f^0(\mathbf{p}) \leq f^1(\mathbf{v}) + f^0(\mathbf{q}), \qquad (\mathbf{v}, \mathbf{q}) \in U,$$

where

$$f^1(\mathbf{u}) = (1/2)(\mathbf{u} - \hat{\mathbf{u}})^\top D(\mathbf{u} - \hat{\mathbf{u}}), \qquad \text{and} \qquad f^0(\mathbf{p}) = \alpha\sum_{i=1}^{N}|p_i|,$$

with $D = \operatorname{diag}(h/2, h, h, \ldots, h, h/2) \in \mathbb{R}^{(N+1)\times(N+1)}$.

Writing the constraint as $\mathbf{p} = C\mathbf{u}$ with $C \in \mathbb{R}^{N\times(N+1)}$, the Lagrangian for the mixed formulation is

$$L\left((\mathbf{u}, \mathbf{p}), \boldsymbol{\lambda}\right) = f^1(\mathbf{u}) + f^0(\mathbf{p}) + \boldsymbol{\lambda}.\left(\mathbf{p} - C\mathbf{u}\right).$$

For this problem the dual $g(\boldsymbol{\lambda}) = \inf_{(\mathbf{u},\mathbf{p})} L\left((\mathbf{u}, \mathbf{p}), \boldsymbol{\lambda}\right)$ can be computed explicitly,

$$g(\boldsymbol{\lambda}) = -(1/2)\boldsymbol{\lambda}^\top(CD^{-1}C^\top)\boldsymbol{\lambda} - C\hat{\mathbf{u}}.\boldsymbol{\lambda} - I_{[-\alpha,\alpha]^N}(\boldsymbol{\lambda}) \equiv g^1(\boldsymbol{\lambda}) + g^0(\boldsymbol{\lambda}),$$

where $I_{[-\alpha,\alpha]^N} : \mathbb{R}^N \to \mathbb{R} \cup \{\infty\}$ is the indicator for the cube. To verify this note that

- $\inf_{p\in\mathbb{R}}(\alpha|p| + \lambda p) = -\infty$ if $\lambda \notin [-\alpha, \alpha]$ and is zero otherwise.
- $\frac{\partial}{\partial \mathbf{u}} L\left((\mathbf{u}, \mathbf{p}), \boldsymbol{\lambda}\right) = D(\mathbf{u} - \hat{\mathbf{u}}) - C^\top\boldsymbol{\lambda}$. Equating this to zero gives $\mathbf{u} = \hat{\mathbf{u}} + D^{-1}C^\top\boldsymbol{\lambda}$, and evaluating the Lagrangian at this minima gives the formula for $g^1(\boldsymbol{\lambda})$.

Since the projection $P_{[-\alpha,\alpha]^N} : \mathbb{R}^N \to [-\alpha, \alpha]^N$ is trivial to compute, the projected steepest descent or projected Nesterov algorithms may can be used to find $\boldsymbol{\lambda} = \arg\max g(\boldsymbol{\lambda})$. The update step for these algorithms is

$$\boldsymbol{\lambda} \mapsto P_{[-\alpha,\alpha]^N}\left(\boldsymbol{\lambda} - \tau C(D^{-1}C^\top + \hat{u})\boldsymbol{\lambda}\right),$$

where $\tau > 0$ is the step size. The solution of the primal problem is then $\mathbf{u} = \hat{\mathbf{u}} + D^{-1}C^\top\boldsymbol{\lambda}$. For this example $\operatorname{rank}(C) = N$, so $g^1(\boldsymbol{\lambda})$ is strongly convex. It follows that the Lagrange multiplier is unique, and if an estimate of the condition number is available Nesterov's algorithm with fixed parameter can be utilized.

EXERCISE 4.7. *Show that* $(N+1)^{-2}|\boldsymbol{\lambda}|^2 \leq |C^\top\boldsymbol{\lambda}^2| \leq 2|\boldsymbol{\lambda}|^2$, *and use this to estimate the condition number of* $CD^{-1}C^\top$.

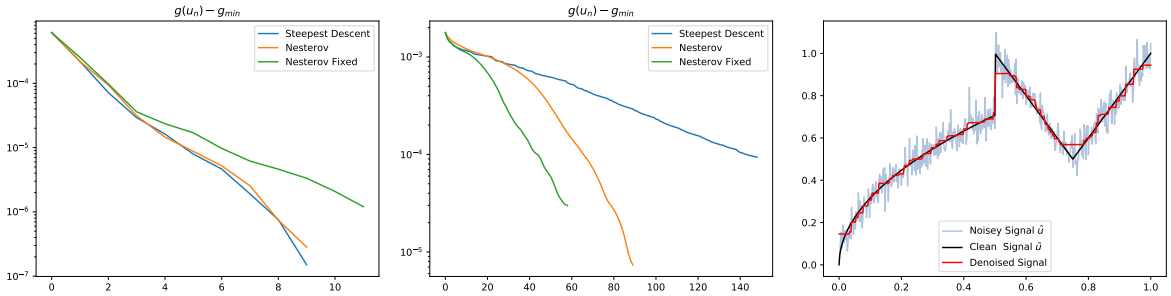

FIG. 4.2. *Dual function values with $N = 50$, $N = 500$ and solution for Example 4.8.*

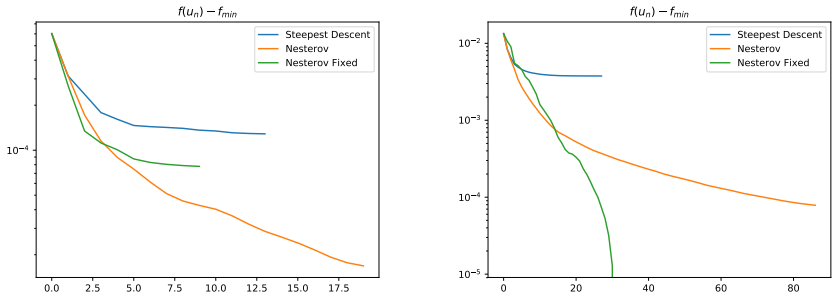

FIG. 4.3. *Function values for $N = 50$, $N = 500$ computed using the primal formulation of Example 4.8.*

EXAMPLE 4.8. *Figure 4.2 plots the values of the dual function $g_{max} - g(\lambda_n)$ computed with the steepest descent, Nesterov, and fixed parameter Nesterov algorithms with for $N = 50$ and $N = 500$, and the solution for the latter. The Armijo rule was used to compute the step sizes, and the estimate for the square root of the condition number for the fixed parameter Nesterov scheme was taken to be $N/4$. The stopping criteria was $|g(\boldsymbol{\lambda}_{n+1}) - g(\boldsymbol{\lambda}_n)| \leq 10^{-6}$, and the estimate of g_{max} was found by setting the tolerance to 10^{-12}. It is clear that the solution with L^1 regularization of the gradient captures the discontinuities much better than the L^2 regularization in Example 1.3, which just smooths the whole signal.*

	$N = 50$		$N = 500$	
Scheme	*Iterations*	*Duality Gap*	*Iterations*	*Duality Gap*
Steepest Descent	10	$1.1800e-05$	149	$2.4068e-04$
Nesterov	10	$2.4549e-05$	90	$5.0773e-04$
Nesterov Fixed	12	$3.8914e-05$	59	$1.1047e-03$

The figures and table show that the method of steepest descent is competitive with Nesterov's algorithm when the condition number is small but when the condition number is large the contraction factor for the linear rate of convergence of steepest descent becomes negligible.

For comparison, Figure 4.3 shows the results of the three algorithms for the primal problem with selection $0 \in \partial|.|(0)$ used for the gradient calculation. The stopping criteria was $|f(x_{n+1} - f_n| \leq 10^{-6}$, and the square root of the condition number for the fixed parameter Nesterov scheme was taken to be $\sqrt{\alpha}N$.

4.2. Uzawa's Algorithm. Frequently the constraints take the form $\phi_i(x) \le 0$ for $1 \le i \le m$, where $\phi_i : X \to \mathbb{R}$ are convex functions, so that the feasible set takes the form

$$U = \{x \in X \mid \phi_i(x) \le 0,\ 1 \le i \le m\}.$$

In this situation the projection $P_U : X \to U$ is not readily computable; however, the dual variable takes values in $[0,\infty)^m \subset \mathbb{R}^m$, and the projection onto this set is easily computed component wise as $\max(0, \mu_i)$. The following theorem summarizes the duality theory needed to utilize this property.

THEOREM 4.9. *Let X be a Hilbert space and $f : X \to \mathbb{R}$ be strongly convex with parameter $\alpha > 0$. Let $\phi : X \to \mathbb{R}^m$ be convex, differentiable, and Lipschitz, and assume that there exists a point $\hat{x} \in X$ where*

$$\phi_i(\hat{x}) \le 0,\ \ 1 \le i \le m, \qquad \text{and} \qquad \phi_i(\hat{x}) < 0, \quad \text{if } \phi_i \text{ is not affine.}$$

Define the Lagrangian $L : X \times [0,\infty)^m \to \mathbb{R}$ to be $L(x,\mu) = f(x) + \mu.\phi(x)$, and dual $g : [0,\infty)^m \to \mathbb{R}$ to be $g(\mu) = \inf_{x\in X} L(x,\mu)$. Then

- *$g : [0,\infty)^m \to \mathbb{R}$ is concave, and $\nabla g(\mu) = \phi(x_\mu)$ where $x_\mu = \arg\min_x L(x,\mu)$.*
- *∇g is Lipschitz on $[0,\infty)^m$ with constant L_ϕ^2/α, where L_ϕ is the Lipschitz constant of ϕ.*
- *Saddle points $(x_*, \mu_*) \in X \times [0,\infty)^m$ exist and satisfy*
  - *$g(\mu_*) \ge g(\mu)$ for all $\mu \in [0,\infty)^m$.*
  - *$\mu_*.\phi(x_*) = 0$.*
  - *The set $U \equiv \{x \in X \mid \phi_i(x) \le 0,\ 1 \le i \le m\}$ is non–empty,*

$$x_* \in U, \quad f(x_*) \le f(x) \quad x \in U, \qquad \text{and} \qquad 0 \in \partial f(x_*) + \sum_{i=1}^{m} \mu_{*i} \nabla \phi_i(\mu_*).$$

Uzawa's algorithm is the projected gradient method applied to the dual variable, thereby reducing an intractable constrained optimization problem to a sequence of tractable unconstrained problems. Let $\mu_0 \in [0,\infty)^m$ and

$$x_n = \arg\min_{y\in X} \left(f(y) + \mu_n.\phi(y)\right), \qquad \mu_{n+1} = P_+\left(\mu_n + \tau_n \phi(x_n)\right), \qquad n = 0, 1, 2, \ldots, \tag{4.1}$$

where $P_+ : \mathbb{R}^m \to [0,\infty)^m$ is the projection, and $\tau_n > 0$.

- In general, the function $y \mapsto f(y) + \mu.\phi(y)$ is convex provided $\mu \in [0,\infty)^m$, so it is necessary to use the projected steepest descent algorithm to guarantee $\mu_n \in [0,\infty)^m$. Nesterov's algorithm can then be used to compute x_n.
  However, if constraint functions $\phi_i : X \to \mathbb{R}$ are affine, then $y \mapsto f(y) + \nu.\phi(y)$ is convex for $\nu \in \mathbb{R}^m$, so the projected Nesterov scheme can be used for the dual variable

$$\mu_{n+1} = P_+\left(\nu_n + \phi(\nu_n)\right), \qquad \nu_{n+1} = \mu_{n+1} + \frac{\lambda_i - 1}{\lambda_{i+1}}\left(\mu_{n+1} - \mu_n\right).$$

- The Armijo rule can be used to compute the step sizes; this may involve multiple evaluations of $g(\mu) = \min_y f(y) + \mu.\phi(y)$. Theorem 4.4 then shows $g(\mu_*) - g(\mu_n) = O(1/n)$; in addition, the duality gap is explicitly computable, $f(x_n) - f(x_*) \le -\mu_n.\phi(x_n)$.

- In general, g may not be strongly concave, so $\{\mu_n\}_{n=1}^{\infty}$ may not converge. However, strong convexity of f guarantees uniqueness of the primal minima x_*, and the next theorem shows that if the step sizes satisfy $\tau_n \leq \alpha/L_\phi^2$, then $x_n \to x_*$.

THEOREM 4.10. *Adopting the notation and hypotheses of Theorem 4.9, assume that the step size in Uzawa's algorithm (4.1) satisfies* $0 < \tau \leq \alpha/L_\phi^2$. *Then*

$$g(\mu_*) - g(\mu_n) \leq \frac{\|\mu_0 - \mu_*\|^2}{2n\tau}, \qquad \textit{and} \qquad \|x_n - x_*\|^2 \leq \frac{2\|\mu_0 - \mu_*\|^2}{\alpha\tau\sqrt{n}}.$$

Proof. Since $\tau \leq \alpha/L_\phi^2 \leq 1/|\nabla g|_{Lip}$ it follows from Lemma 3.2 (Amiljo rule) that the descent condition is always satisfied. Theorem 4.4 then gives

$$g(\mu_*) - g(\mu_n) \leq \frac{\|\mu_0 - \mu_*\|^2}{2n\tau} \qquad \text{and} \qquad \|\mu_{n+1} - \mu_n\| \leq \frac{\|\mu_0 - \mu_*\|}{\sqrt{n}}.$$

- Consistency of the algorithm follows from the property that saddle points (x_*, μ_*) satisfy $P_+(\mu_* + \tau\phi(x_*)) = \mu_*$ for all $\tau \geq 0$.
  To verify this, let $\mu \in [0,\infty)^m$ and compute

  $$\|\mu_* + \tau\phi(x_*) - \mu\|^2 = \|\mu_* - \mu\|^2 + 2\tau(\phi(x_*), \mu_* - \mu) + \tau^2\|\phi(x_*)\|^2 \geq \|\mu_* - \mu\|^2.$$

  The inequality holds since $\mu_*.\phi(x_*) = 0$ and $\phi_i(x_*) \leq 0$. Then μ_* is the closest point in $[0,\infty)^m$ to $\mu_* + \tau\phi(x_*)$, so is the projection.
- We next show that

  $$\alpha\|x_n - x_*\|^2 \leq (\mu_* - \mu_n).(\phi(x_n) - \phi(x_*)). \tag{4.2}$$

  The necessary conditions for optimality of the primal problems are

  $$0 \in \partial f(x_n) + \sum_{i=1}^{m} \mu_{ni}\nabla\phi_i(x_n), \qquad \text{and} \qquad 0 \in \partial f(x_*) + \sum_{i=1}^{m} \mu_{*i}\nabla\phi_i(x_*).$$

  Subtracting and and using the convexity of ϕ gives (a selection of the sub-gradients for which)

  $$\begin{aligned}\big(\partial f(x_n) - \partial f(x_*), x_n - x_*\big) &= \sum_{i=1}^{m} \mu_{*i}\left(\nabla\phi_i(x_*), x_n - x_*\right) + \mu_{ni}\left(\nabla\phi_i(x_n), x_* - x_n\right) \\ &\leq (\mu_* - \mu_n).(\phi(x_n) - \phi(x_*)),\end{aligned}$$

  the inequality following since the components of μ_n and μ_* are non–negative. Equation (4.2) then follows from strong convexity of f.
- Convergence of the primal variable will follow from the estimate

  $$\tau(2\alpha - \tau L_\phi^2)\|x_n - x_*\|^2 \leq \|\mu_n - \mu_*\|^2 - \|\mu_{n+1} - \mu_*\|^2. \tag{4.3}$$

  Granted this, the the difference of the squares can be factored to give

  $$\tau(2\alpha - \tau L_\phi^2)\|x_n - x_*\|^2 \leq 2\|\mu_0 - \mu_*\|\|\mu_{n+1} - \mu_n\| \leq 2\|\mu_0 - \mu_*\|^2/\sqrt{n}.$$

The theorem then follows from since $0 < \tau \leq \alpha/L_\phi^2$.
Equation (4.3) results upon writing the update step of the dual variable as

$$\mu_{n+1} - \mu_* = P_+(\mu_n + \tau\phi(x_n)) - \mu_* = P_+(\mu_n + \tau\phi(x_n)) - P_+(\mu_* + \tau\phi(x_*)).$$

Taking the norm of both sides and recalling that $P_+ : X \to [0,\infty)^m$ is a contraction gives

$$\begin{aligned}
\|\mu_{n+1} - \mu_*\|^2 &= \|P_+(\mu_n + \tau\phi(x_n)) - P_+(\mu_* + \tau\phi(x_*))\|^2 \\
&\leq \|(\mu_n + \tau\phi(x_n)) - (\mu_* + \tau\phi(x_*))\|^2 \\
&= \|\mu_n - \mu_*\|^2 + 2\tau(\mu_n - \mu_*).(\phi(x_n) - \phi(x_*)) + \tau^2\|\phi(x_n) - \phi(x_*)\|^2 \\
&\leq \|\mu_n - \mu_*\|^2 + 2\tau(\mu_n - \mu_*).(\phi(x_n) - \phi(x_*)) + \tau^2 L_\phi^2\|x_n - x_*\|^2,
\end{aligned}$$

and (4.3) follows upon using equation (4.2) to bound the cross term.

□

Appendix A. Discrete and Differential Equations. Convergence properties of the method of steepest descent are frequently motivated by identifying it with a discrete approximation of the differential equation,

$$x'(t) + \nabla f(x(t)) = 0, \qquad x(0) = x_0.$$

If f is strongly convex with parameter $\alpha \geq 0$, a calculation shows

$$\frac{d}{dt}\Big(t(f(t) - f(x_*)) + (1/2)\|x(t) - x_*\|^2\Big) \leq -(\alpha/2)\|x(t) - x_*\|^2.$$

It follows that

$$f(x(t)) - f(x_*) \leq \frac{\|x_0 - x_*\|^2}{2t}, \qquad \text{and} \qquad \|x(t) - x_*\|^2 \leq \|x_0 - x_*\|^2 \exp(-\alpha t).$$

A discrete version of this appears in the convergence proof of steepest descent in Section 3.1. When the gradient of f is Lipschitz with constant L,

$$f(x(t)) - f(x_*) \leq (L/2)\|x(t) - x_*\|^2 \leq (L/2)\|x_0 - x_*\|^2 \exp(-\alpha t),$$

which is an asymptotically better than $O(1/t)$ when $\alpha > 0$.

Nesterov's original manuscript was succinct; the scheme was not introduced as either an extrapolation (SOR variant) of steepest descent, or as a discrete approximation of a differential equation. However, writing the Nesterov scheme as the three term recurrence,

$$\lambda_{i+1}(y_{i+1} - 2y_i + y_{i-1}) + (\lambda_{i+1} - \lambda_i + 1)(y_i - y_{i-1}) + (\lambda_{i+1} + \lambda_i - 1)\tau_i\nabla f(y_i) - (\lambda_i - 1)\tau_{i-1}\nabla f(y_{i-1}) = 0,$$

motivates formal connections with second order differential equations.

In [23] the authors substituted the asymptotic approximation $\lambda_i \simeq (i-1)/2$ and terminal step size $\tau = \tau_i = \tau_{i-1} = s^2$ into the three term recurrence to get

$$i(y_{i+1} - 2y_i + y_{i-1}) + 3(y_i - y_{i-1}) + s^2(2i - 3)\nabla f(y_i) - s^2(i - 3)\nabla f(y_{i-1}) \simeq 0.$$

Rewriting this as

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{s^2} + \frac{3}{is}\frac{y_i - y_{i-1}}{s} + (2 - 3/i)\nabla f(y_i) - (1 - 3/i)\nabla f(y_{i-1}) \simeq 0,$$

```
> restart:
  ee := t^2*(f(y(t)) - fstar) + 2*(y(t) - xstar + (t/2)*diff(y(t),t))^2;
  d2ydt := -D(f)(y(t)) - (3/t)*diff(y(t),t);
  simplify( subs(diff(y(t),t,t)=d2ydt, diff(ee,t)) );
```

$$ee := t^2\,(f(y(t)) - fstar) + 2\left(y(t) - xstar + \frac{t\left(\frac{\mathrm{d}}{\mathrm{d}t}\,y(t)\right)}{2}\right)^2$$

$$d2xdt := -\mathrm{D}(f)\,(y(t)) - \frac{3\left(\frac{\mathrm{d}}{\mathrm{d}t}\,y(t)\right)}{t}$$

$$-2\,((-xstar + y(t))\,\mathrm{D}(f)\,(y(t)) + fstar - f(y(t)))\,t \qquad \textbf{(1)}$$

FIG. A.1. *Derivation of the dissipation relation (A.1).*

and identifying s as a time step motivates the differential equation

$$y''(t) + (3/t)y'(t) + \nabla f(y(t)) = 0, \qquad y(0) = y_0,\ y'(0) = 0.$$

Solutions of this equation satisfy

$$\frac{d}{dt}\Big(t^2\left(f(y(t)) - f(x_*)\right) + 2\|y(t) - x_* + (t/2)y'(t)\|^2\Big) \le 0. \tag{A.1}$$

The analog of (3.5) is then immediate

$$t^2\left(f(y(t)) - f(x_*)\right) + 2\|y(t) - x_* + (t/2)y'(t)\|^2 \le 2\|y_0 - x_*\|^2.$$

EXERCISE A.1. *Figure A.1 shows the results of a* MAPLE *calculation. If f is strongly convex with parameter $\alpha \ge 0$ show that a sharper statement of equation (A.1) is*

$$\frac{d}{dt}\Big(t^2\left(f(y(t)) - f(x_*)\right) + 2\|y(t) - x_* + (t/2)y'(t)\|^2\Big) \le -t\alpha\|y(t) - x_*\|^2.$$

In [18] the authors considered the three term recurrence for Nesterov's algorithm with λ fixed and terminal step size $\tau = \tau_i = \tau_{i-1} = s^2$,

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{s^2} + \frac{1}{\lambda s}\frac{y_i - y_{i-1}}{s} + (2 - 1/\lambda)\nabla f(y_i) - (1 - 1/\lambda)\nabla f(y_{i-1}) = 0.$$

Setting $\lambda = \sqrt{L/\alpha} \equiv \sqrt{\kappa}$ and step size $\tau = s^2 = 1/L$ gives

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{s^2} + \sqrt{\alpha}\,\frac{y_i - y_{i-1}}{s} + (2 - s\sqrt{\alpha})\nabla f(y_i) - (1 - s\sqrt{\alpha})\nabla f(y_{i-1}) = 0,$$

which motivates the differential equation

$$y''(t) + \sqrt{\alpha}\,y'(t) + \nabla f(y(t)) = 0, \qquad y(0) = y_0,\ y'(0) = 0.$$

When f is strongly convex with parameter $\alpha \ge 0$, solutions of this differential equation satisfy

$$\frac{d}{dt}\exp(\sqrt{\alpha t})\Big(f(y(t)) - f(x_*) + (1/2)\|\sqrt{\alpha}(y(t) - x_*) + y'(t)\|^2\Big) \le 0.$$

Integrating this gives the analog of the descent rate in Theorem 3.9,

$$f(y(t)) - f(x_*) + (1/2)\|\sqrt{\alpha}(y(t) - x_*) + y'(t)\|^2 \leq \exp(-\sqrt{\alpha}t)\big(f(y_0) - f(x_*) + (\alpha/2)\|y_0 - x_*\|^2\big).$$

These differential equations provide some insight into the descent properties of the associated scheme. While differential equations with better descent properties are known, [3], the development of explicit time stepping schemes to approximate their solutions which inherit their descent (i.e. stability) properties is difficult. Note that Lipschitz continuity of ∇f is not required to establish descent of solutions to the differential equations, this is a required in the discrete setting to establish stability of explicit schemes.

REFERENCES

[1] L. Alvarez, F. Guichard, P.-L. Lions, and J.-M. Morel, *Axioms and fundamental equations of image processing*, Arch. Rational Mech. Anal., 123 (1993), pp. 199–257.

[2] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[3] H. Attouch, Z. Chbani, and H. Riahi, *Fast proximal methods via time scaling of damped inertial dynamics*, SIAM J. Optim., 29 (2019), pp. 2227–2256.

[4] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, Cambridge, 2004.

[5] S. Bubeck, *Convex optimization: Algorithms and complexity*, Foundations and Trends in Machine Learning, 8 (2015), pp. 231–357.

[6] A. Chambolle and C. Dossal, *On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm"*, J. Optim. Theory Appl., 166 (2015), pp. 968–982.

[7] T. F. Chan, J. Shen, and L. Vese, *Variational PDE models in image processing*, Notices Amer. Math. Soc., 50 (2003), pp. 14–26.

[8] P. G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree], Masson, Paris, 1982.

[9] P. G. Ciarlet, *Introduction to Numerical Linear Algebra and Optimisation*, Cambridge, 1988.

[10] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*, vol. 16 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. Corrected reprint of the 1983 original.

[11] I. Ekeland and R. Témam, *Convex analysis and variational problems*, vol. 28 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, english ed., 1999. Translated from the French.

[12] W. Forst and D. Hoffmann, *Optimization—theory and practice*, Springer Undergraduate Texts in Mathematics and Technology, Springer, New York, 2010.

[13] S. V. Guminov, Y. E. Nesterov, P. E. Dvurechensky, and A. V. Gasnikov, *Accelerated primal-dual gradient descent with linesearch for convex, nonconvex, and nonsmooth optimization problems*, Doklady Mathematics, 99 (2019), pp. 125–128.

[14] C. T. Kelley, *Iterative methods for optimization*, vol. 18 of Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.

[15] Y. Nesterov, *Introductory lectures on convex programming*, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.693.855, (1998).

[16] Y. E. Nesterov, *A method for solving the convex programming problem with convergence rate $O(1/k^2)$*, Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543–547.

[17] J. M. Ortega, *Numerical analysis. A second course*, Academic Press, New York-London, 1972. Computer Science and Applied Mathematics.

[18] J.-H. Park, A. J. Salgado, and S. M. Wise, *Preconditioned accelerated gradient descent methods for locally Lipschitz smooth objectives with applications to the solution of nonlinear PDEs*, J. Sci. Comput., 89 (2021), pp. Paper No. 17, 37.

[19] R. T. Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, N.J., 1970.

[20] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268. Experimental mathematics: computational issues in nonlinear science (Los Alamos, NM, 1991).

[21] R. E. Showalter, *Monotone operators in Banach space and nonlinear partial differential equations*, vol. 49 of Mathematical Surveys and Monographs, American Mathematical Society, Providence, RI, 1997.

[22] G. Strang, *Linear algebra and its applications*, Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1976.

[23] W. Su, S. Boyd, and E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: theory and insights*, J. Mach. Learn. Res., 17 (2016), pp. Paper No. 153, 43.

[24] D. M. Young, *Iterative solution of large linear systems*, Academic Press, New York-London, 1971.