Statistical Science
2015, Vol. 30, No. 4, 559–581
DOI: 10.1214/15-STS530

Proximal Algorithms in Statistics and Machine Learning

Nicholas G. Polson, James G. Scott and Brandon T. Willard

Abstract. Proximal algorithms are useful for obtaining solutions to difficult optimization problems, especially those involving nonsmooth or composite objective functions. A proximal algorithm is one whose basic iterations involve the *proximal operator* of some function, whose evaluation requires solving a specific optimization problem that is typically easier than the original problem. Many familiar algorithms can be cast in this form, and this "proximal view" turns out to provide a set of broad organizing principles for many algorithms useful in statistics and machine learning. In this paper, we show how a number of recent advances in this area can inform modern statistical practice. We focus on several main themes: (1) variable splitting strategies and the augmented Lagrangian; (2) the broad utility of envelope (or variational) representations of objective functions; (3) proximal algorithms for composite objective functions; and (4) the surprisingly large number of functions for which there are closed-form solutions of proximal operators. We illustrate our methodology with regularized Logistic and Poisson regression incorporating a nonconvex bridge penalty and a fused lasso penalty. We also discuss several related issues, including the convergence of nondescent algorithms, acceleration and optimization for nonconvex functions. Finally, we provide directions for future research in this exciting area at the intersection of statistics and optimization.

Key words and phrases: Bayes MAP, shrinkage, sparsity, splitting, Kurdyka–Łojasiewicz, nonconvex, envelopes, regularization, ADMM, optimization, Divide and Concur.

1. INTRODUCTION

1.1 Proximal Algorithms for Optimization

Optimization problems that involve a trade-off between model fit and model complexity sit at the heart of modern statistical practice. They arise, for example, in sparse regression (Tibshirani, 1996), spatial smoothing (Tibshirani et al., 2005), covariance estimation (Witten, Tobshirani and Hastie, 2009), image processing (Geman and Reynolds, 1992; Geman and Yang, 1995; Rudin, Osher and Faterni, 1992), nonlinear curve fitting (Tibshirani, 2014), Bayesian MAP inference (Polson and Scott, 2012), multiple hypothesis testing (Tansey et al., 2014) and shrinkage/sparsity-inducing prior regularization problems (Green et al., 2015).

The goal of this paper is to introduce researchers in statistics and machine learning to the large body of literature on *proximal algorithms* for solving such optimization problems. By a proximal algorithm, we mean an algorithm whose steps involve evaluating a *proximal operator* related to some term in the objective function. Both of these concepts will be defined precisely in the next section, but the basic idea is simple. Evaluating a proximal operator requires solving a specific

Nicholas G. Polson is Professor of Econometrics and Statistics and Brandon T. Willard is Research Consultant, Booth School of Business, University of Chicago, 5807 South Woodlawn Avenue, Chicago, Illinois 60637, USA (e-mail: ngp@chicagobooth.edu; bwillard@uchicago.edu). James G. Scott is Associate Professor of Statistics, McCombs School of Business and Department of Statistics and Data Sciences, University of Texas at Austin, 2110 Speedway, B6500, Austin, Texas 78712, USA (e-mail: James.Scott@mccombs.utexas.edu).

optimization subproblem that is (one hopes) easier than the original problem of interest. By iteratively solving such subproblems, a proximal algorithm converges on the solution to the original problem. Chrétien and Hero III (2000) provide a general relation between EM and proximal point (PP) algorithms and show that the latter can provide dramatic improvements in rates of convergence.

The early foundational work in this area dates to the study of iterative fixed-point algorithms in Banach spaces (Von Neumann, 1951; Brègman, 1967; Hestenes, 1969; Martinet, 1970; Rockafellar, 1976). As these techniques matured, they became widely used in several different fields. As a result, they have been referred to by a diverse set of names, including proximal gradient, proximal point, alternating direction method of multipliers (ADMM) (Boyd et al., 2011), divide and concur (DC), Frank–Wolfe (FW), Douglas–Rachford splitting, operator splitting and alternating split Bregman (ASB) methods. The field of image processing has developed most of these ideas independently of statistics—for example, in the form of total variation (TV) de-noising and half-quadratic (HQ) optimization (Geman and Yang, 1995; Geman and Reynolds, 1992; Nikolova and Ng, 2005). Many other widely-known methods—including, for example, fast iterative shrinkage thresholding (FISTA), expectation maximization (EM), majorization-minimization (MM) and iteratively reweighed least squares (IRLS)—also fall into the proximal framework.

Recently there has been a spike of interest in proximal algorithms, with a handful of recent broad surveys appearing in the last few years (Cevher, Becker and Schmidt, 2014; Komodakis and Pesquet, 2014; Combettes and Pesquet, 2011; Boyd et al., 2011). Indeed, the use of specific proximal algorithms has become commonplace in statistics and machine learning (e.g., Bien, Taylor and Tibshirani, 2013; Tibshirani, 2014; Tansey et al., 2014). However, there has not been a real focus on the general family of approaches that underly these algorithms, with specific attention to the issues of most direct interest to statisticians. Our review is designed to fill this gap.

The rest of the paper proceeds as follows. Section 1.2 provides notation and basic properties of proximal operators and envelopes. Section 2 describes the proximal operator and Moreau envelope. Section 3 describes the basic proximal algorithms and their extensions. Section 4 describes common algorithms and techniques, such as ADMM and Divide and Concur, that rely on proximal algorithms. Section 5 discusses envelopes and how proximal algorithms can be viewed as envelope gradients. Section 6 considers the general problem of composite operator optimization and shows how to compute the exact proximal operator with a general quadratic envelope and a composite regularization penalty. Section 7 illustrates the methodology with applications to logistic and Poisson regression with fused lasso penalties. A bridge regression penalty illustrates the nonconvex case and we apply our algorithm to the prostate data of Hastie, Tibshirani and Friedman (2009). Finally, Section 8 concludes with directions for future research, while Appendix A discusses convergence results for both convex and nonconvex cases together with Nesterov acceleration.

We also include several useful summaries in table form. Table 1 lists commonly used proximal operators, Table 2 documents several examples of half-quadratic envelopes, and Table 3 provides convergence rates for a variety of algorithms.

1.2 Notation

In this paper we consider optimization problems of the form

$$\text{minimize} \quad F(x) := l(x) + \phi(x), \tag{1}$$

where $l(x)$ is a measure of fit depending implicitly on some observed data y, and $\phi(x)$ is a regularization term that imposes structure or effects a favorable bias-variance trade-off. Often $l(x)$ is a smooth function and $\phi(x)$ is nonsmooth—like a lasso or bridge penalty—so as to induce sparsity. We will assume that l and p are convex and lower semi-continuous except when explicitly stated to be nonconvex.

We will pay particular attention to composite penalties of the form $\phi(Bx)$, where B is a matrix corresponding to some constraint or structural penalty, such as the discrete difference operator in fused lasso or polynomial trend filtering. We use $x = (x_1, \ldots, x_d)$ to denote a d-dimensional parameter of interest, y an n-vector of outcomes, A a fixed $n \times d$ matrix whose rows are covariates (or features) $a_i^\top$, and B a fixed $k \times d$ matrix, b a prior mean or target for shrinkage, and $\gamma > 0$ a regularization parameter that will trace out a solution path. Observations are indexed by i, parameters by j, and iterations of an algorithm by t. Unless stated otherwise, all vectors are column vectors. Putting these together, this paper treats general composite objectives of the form

$$F(x) := \sum_{i=1}^{n} l\bigl(y_i, a_i^\top x\bigr) + \gamma \sum_{j=1}^{k} \phi\bigl([Bx - b]_j\bigr). \tag{2}$$

TABLE 1
Sources: Chaux et al. (2007), Hu, Li and Yang (2015)

Type	$\boldsymbol{\phi(x)}$	$\mathbf{prox}_{\gamma\boldsymbol{\phi}}(y)$
Laplace	$\omega\|x\|$	$\operatorname{sgn}(x)\max(\|x\|-\omega,0)$
Gaussian	$\tau\|x\|^2$	$x/(2\tau+1)$
Group-sparse, ℓ_p	$\kappa\|x\|^p$	$\operatorname{sgn}(x)\rho$, ρ s.t. $\rho+p\kappa\rho^{p-1}=\|x\|$
⋮	$p=4/3$	$x+\frac{4\kappa}{32^{1/3}}((\chi-x)^{1/3}-(\chi+x)^{1/3})$, $\chi=\sqrt{x^2+256\kappa^3/729}$
⋮	$p=3/2$	$x+9\kappa^2\operatorname{sgn}(x)(1-\sqrt{1+16\|x\|/(9\kappa^2)})/8$
⋮	$p=3$	$\operatorname{sgn}(x)(\sqrt{1+12\kappa\|x\|}-1)/(6\kappa)$
⋮	$p=4$	$(\frac{\chi+x}{8\kappa})^{1/3}-(\frac{\chi-x}{8\kappa})^{1/3}$, $\chi=\sqrt{x^2+1/(27\kappa)}$
Gamma, Chi	$-\kappa\ln x+\omega x$	$\frac{1}{2}(x-\omega+\sqrt{(x-\omega)^2+4\kappa})$
Double–Pareto	$\gamma\log(1+\|x\|/a)$	$\frac{\operatorname{sgn}(x)}{2}\{\|x\|-a+\sqrt{(a-\|x\|)^2+4d(x)}\}$, $d(x)=(a\|x\|-\gamma)_+$
Huber dist.	$\begin{cases}\tau x^2, & \|x\|\le\omega/\sqrt{2\tau},\\ \omega\sqrt{2\tau}\|x\|-\omega^2/2, & \text{otherwise}\end{cases}$ $\omega,\tau\in(0,+\infty)$	$\begin{cases}\frac{x}{2\tau+1}, & \|x\|\le\omega(2\tau+1)/\sqrt{2\tau},\\ x-\omega\sqrt{2\tau}\operatorname{sgn}(x), & \|x\|>\omega(2\tau+1)/\sqrt{2\tau}\end{cases}$
Max-entropy dist.	$\omega\|x\|+\tau\|x\|^2+\kappa\|x\|^p$ $2\neq p\in(1,+\infty)$, $\omega,\tau,\kappa\in(0,+\infty)$	$\operatorname{sgn}(x)\operatorname{prox}_{\kappa\|\cdot\|^p/(2\tau+1)}(\frac{1}{2\tau+1}\max(\|x\|-\omega,0))$
Smoothed-laplace dist.	$\omega\|x\|-\ln(1+\omega\|x\|)$	$\operatorname{sgn}(x)\frac{\omega\|x\|-\omega^2-1+\sqrt{\|\omega\|x\|-\omega^2-1\|^2+4\omega\|x\|}}{2\omega}$
Exponential dist.	$\begin{cases}\omega x, & x\ge 0,\\ +\infty, & x<0\end{cases}$	$\begin{cases}x-\omega, & x\ge\omega,\\ 0, & x<\omega\end{cases}$
Uniform dist.	$\begin{cases}-\omega, & x<-\omega,\\ x, & \|x\|\le\omega,\\ \omega, & x>\omega\end{cases}$	$\begin{cases}x-\omega, & x\ge\omega,\\ 0, & x<\omega\end{cases}$
Triangular dist.	$\begin{cases}-\ln(x-\omega)+\ln(-\omega), & x\in(\omega,0),\\ -\ln(\hat\omega-x)+\ln(\hat\omega), & x\in(0,\hat\omega),\\ +\infty, & \text{otherwise}\end{cases}$ $\omega\in(-\infty,0]$, $\hat\omega\in(0,\infty)$	$\begin{cases}\frac{x+\omega+\sqrt{\|x-\omega\|^2+4}}{2}, & x<1/\omega,\\ \frac{x+\hat\omega-\sqrt{\|x-\hat\omega\|^2+4}}{2}, & x>1/\hat\omega\end{cases}$
Weibull dist.	$\begin{cases}-\kappa\ln x+\omega x^p, & x>0,\\ +\infty, & x\le 0\end{cases}$ $p\in(1,+\infty)$ $\omega,\kappa\in(-\infty,0]$	π s.t. $p\omega\pi^p+\pi^2-x\pi=\kappa$
GIG dist.	$\begin{cases}-\kappa\ln x+\omega x+\rho/x, & x>0,\\ +\infty, & x\le 0\end{cases}$ $\omega,\kappa,\rho\in(-\infty,0]$	π s.t. $\pi^3+(\omega-x)\pi^2-\kappa\pi=\rho$

For example, lasso can be viewed as a simple statistical model with the negative log-likelihood from $y=Ax+\varepsilon$, where ε is a standard normal measurement error, corresponding to the norm $l(x)=\|Ax-y\|^2$, and each parameter x_j has independent Laplace priors corresponding to the regularization penalty $\phi(x)=|x|$. To keep the notation light, we overload the symbols l and ϕ: they can refer either to the overall loss and penalty terms [as in equation (1)] or to the individual component-wise terms that are added to produce the overall loss or penalty [as in equation (2)]. We have taken care to ensure that their meaning will always be clear in context.

We also use the following conventions: $\operatorname{sgn}(x)$ is the algebraic sign of x, and $x_+=\max(x,0)$; $\iota_C(x)$ is the set indicator function taking the value 0 if $x\in C$ and ∞ if $x\notin C$; $\mathbb{R}^+=[0,\infty)$, $\mathbb{R}^{++}=(0,\infty)$, and $\overline{\mathbb{R}}$ is the extended real line $\mathbb{R}\cup\{-\infty,\infty\}$.

TABLE 2

Minimizers for the multiplicative form are $\sigma(t) = \begin{cases} \phi''(0^+), & \text{if } t=0, \\ \phi'(t)/t, & \text{if } t \neq 0 \end{cases}$ *and for additive form* $\sigma(t) = ct - \phi'(t)$. *See* Nikolova and Ng (2005)

Penalty	**Minimizer**	
$\boldsymbol{\phi(t) = \min_s\{Q(t,s) + \psi(s)\}}$	$\boldsymbol{Q(t,s) = \frac{1}{2}t^2 s}$	$\boldsymbol{Q(t,s) = (t-s)^2}$
$\lvert t\rvert^\alpha$, $\alpha \in (1,2]$	$\alpha \lvert t\rvert^{\alpha-2}$	
$\sqrt{\alpha + t^2}$	$\frac{1}{\sqrt{\alpha+t^2}}$	$ct - \frac{t}{\sqrt{\alpha+t^2}}$
$\frac{\lvert t\rvert}{\alpha} - \log(1 + \frac{\lvert t\rvert}{\alpha})$	$\frac{1}{\alpha(\alpha+\lvert t\rvert)}$	$ct - \frac{t}{\alpha(\alpha+\lvert t\rvert)}$
$\begin{cases} \frac{t^2}{2}, & \lvert t\rvert \le \alpha, \\ \alpha\lvert t\rvert - \frac{\alpha^2}{2}, & \lvert t\rvert > \alpha \end{cases}$	$\begin{cases} 1, & \lvert t\rvert \le \alpha, \\ \frac{\alpha}{\lvert t\rvert}, & \lvert t\rvert > \alpha \end{cases}$	$\begin{cases} (c-1)t, & \lvert t\rvert \le \alpha, \\ ct - \alpha\,\mathrm{sgn}(t), & \lvert t\rvert > \alpha \end{cases}$
$\log(\cosh(\alpha t))$	$\alpha \frac{\tanh(\alpha t)}{t}$	$ct - \alpha\tanh(\alpha t)$
$-\frac{1}{1+\lvert \mathbf{x}\rvert}$	$\begin{cases} -2, & \text{for } t = 0, \\ \frac{\mathrm{sgn}(t)}{t(\lvert t\rvert+1)^2}, & \text{otherwise} \end{cases}$	$ct - \frac{\mathrm{sgn}(t)}{(\lvert t\rvert+1)^2}$
$-\frac{1}{1+\sqrt{\mathbf{x}}}$	$\begin{cases} -\infty, & \text{for } t = 0, \\ \frac{1}{2t^{3/2}(\sqrt{t}+1)^2}, & \text{otherwise} \end{cases}$	$ct - \frac{1}{2\sqrt{t}(\sqrt{t}+1)^2}$

Further preliminaries. We now briefly introduce several useful concepts and definitions to be described further in subsequent sections. First, *splitting* is a key tool that exploits an equivalence between an unconstrained optimization problem and a constrained one that includes a latent or slack variable z. For example, suppose that the original problem is

$$\underset{x}{\text{minimize}} \quad l(x) + \phi(Bx).$$

To apply splitting to this problem, we formulate the equivalent problem

$$\underset{x,z}{\text{minimize}} \quad l(x) + \phi(z)$$
$$\text{subject to} \quad Bx = z,$$

so that the objective is split into two terms involving separate sets of primal variables.

The *convex conjugate* of $l(x)$, $l^\star(z)$, is defined as

$$l^\star(\lambda) = \sup_x \{\lambda^\top x - l(x)\}.$$

The conjugate function $l^\star(\lambda)$ is the point-wise supremum of a family of affine (and therefore convex) functions in z; it is convex even when $l(x)$ is not. But if $l(x)$ is convex (and closed and proper), then we also have that $l(x) = \sup_\lambda\{\lambda^\top x - l^\star(\lambda)\}$, so that l and $l^\star$ are dual to one another. If $l(x)$ is differentiable, the maximizing value of λ is $\hat{\lambda}(x) = \nabla l(x)$.

The convex conjugate is our first example of an *envelope*, which is a way of representing functions in terms of a pointwise extremum of a family of functions indexed by a latent variable. Another example is

TABLE 3

See Duckworth (2014)

Algorithm	**Error rate**		**Per-iteration cost**
	Convex	**Strongly convex**	
Accelerated gradient descent	$O(1/\sqrt{\varepsilon})$	$O(\log(1/\varepsilon))$	$O(n)$
Proximal gradient descent	$O(1/\varepsilon)$	$O(\log(1/\varepsilon))$	$O(n)$
Accelerated proximal gradient descent	$O(1/\sqrt{\varepsilon})$	$O(\log(1/\varepsilon))$	$O(n)$
ADMM	$O(1/\varepsilon)$	$O(\log(1/\varepsilon))$	$O(n)$
Frank-wolfe/conditional gradient algorithm	$O(1/\varepsilon)$	$O(1/\sqrt{\varepsilon})$	$O(n)$
Newton's method		$O(\log\log(1/\varepsilon))$	$O(n^3)$
Conjugate gradient descent		$O(n)$	$O(n^2)$
L-BFGS		Between $O(\log(1/\varepsilon))$ and $O(\log\log(1/\varepsilon))$	$O(n^2)$

a quadratic envelope, where we represent l as

$$l(x) = \inf_z \left\{ \frac{1}{2} x^\top \Lambda(z) x - \eta(z)^\top x + \psi(z) \right\}$$

for some Λ, η, ψ. We will draw heavily on the use of envelope (or variational) representations of functions.

A function $g(x)$ is said to *majorize* another function $f(x)$ at x_0 if $g(x_0) = f(x_0)$ and $g(x) \geq f(x)$ for all $x \neq x_0$. If the same relation holds with the inequality sign flipped, $g(x)$ is said to be a *minorizing* function for $f(x)$.

The *subdifferential* of a function f at the point x is defined as the set

$$\partial f(x) = \{v : f(z) \geq f(x) + v^\top (z - x), \quad \forall z, x \in \mathrm{dom}(f)\}.$$

Any such element is called a subgradient. If the function is differentiable, then the subdifferential is a singleton set comprising the ordinary gradient from differential calculus.

Finally, a ρ-strong convex function satisfies

$$f(x) \geq f(z) + u^\top (x - z) + \frac{\rho}{2} \|x - z\|_2^2 \quad \text{where } u \in \partial f(z),$$

while a ρ-smooth function satisfies

$$f(x) \leq f(z) + \nabla f(z)^\top (x - z) + \frac{\rho}{2} \|x - z\|_2^2 \quad \forall x, z.$$

2. PROXIMAL OPERATORS AND MOREAU ENVELOPES

2.1 Basic Properties

Our perspective throughout this paper will be to view a proximal algorithm as taking a gradient-descent step for a suitably defined envelope function. By constructing different envelopes, one can develop new optimization algorithms. We build up to this perspective by first discussing the basic properties of the proximal operator and its relationship to the gradient of the standard Moreau envelope.

Let $f(x)$ be a lower semi-continuous function, and let $\gamma > 0$ be a scalar. The Moreau envelope $f^\gamma(x)$ and proximal operator $\mathrm{prox}_{\gamma f}(x)$ with parameter γ are defined as

$$f^\gamma(x) = \inf_z \left\{ f(z) + \frac{1}{2\gamma} \|z - x\|_2^2 \right\} \leq f(x),$$
$$\mathrm{prox}_{\gamma f}(x) = \operatorname*{argmin}_z \left\{ f(z) + \frac{1}{2\gamma} \|z - x\|_2^2 \right\}. \tag{3}$$

Intuitively, the Moreau envelope is a regularized version of f. It approximates f from below and has the same set of minimizing values (Rockafellar and Wets, 1998, Chapter 1G). The proximal operator specifies the value that solves the minimization problem defined by the Moreau envelope. It balances the two goals of minimizing f and staying near x, with γ controlling the trade-off. Table 1 provides an extensive list of closed-form solutions.

Parikh and Boyd (2013) provide several interesting interpretations of the proximal operator. Each one provides some intuition about why proximal operators might be useful in optimization. We highlight three of these interpretations here.

First, the proximal operator behaves similarly to a gradient-descent step for the function f. There are many ways of motivating this connection, but one simple way is to consider the Moreau envelope $f^\gamma(x)$. Observe that the Moreau derivative is

$$\partial f^\gamma(x) = \partial \inf_z \left\{ f(z) + \frac{1}{2\gamma} \|z - x\|_2^2 \right\} = \frac{1}{\gamma} [x - \hat{z}(x)],$$

where $\hat{z}(x) = \mathrm{prox}_{\gamma f}(x)$ is the value that achieves the minimum. Hence,

$$\mathrm{prox}_{\gamma f}(x) = x - \gamma \partial f^\gamma(x).$$

Thus, evaluating the proximal operator can be viewed as a gradient-descent step for the Moreau envelope, with γ as a step-size parameter.

Second, the proximal operator generalizes the notion of the Euclidean projection. To see this, consider the special case where $f(x) = \iota_C(x)$ is the set indicator function of some convex set C. Then $\mathrm{prox}_f(x) = \mathrm{argmin}_{z \in C} \|x - z\|_2^2$ is the ordinary Euclidean projection of x onto C. This suggests that, for other functions, the proximal operator can be thought of as a generalized projection. A constrained optimization problem $\min_{x \in C} f(x)$ has an equivalent solution as an unconstrained proximal operator problem. Proximal approaches are, therefore, directly related to convex relaxation and quadratic majorization, through the addition of terms like $\frac{\rho}{2} \|x - v\|^2$ to an objective function, where ρ might be a constant that bounds an operator or the Hessian of a function. We can choose where these quadratic terms are introduced, which variables the terms can involve, and the order in which optimization steps are taken. The envelope framework highlights such choices, leading to many distinct and familiar algorithms.

Finally, there is a close connection between proximal operators and fixed-point theory, in that $\mathrm{prox}_{\gamma f}(x^\star) =$

$x^\star$ if and only if $x^\star$ is a minimizing value of $f(x)$. To see this informally, consider the *proximal minimization* algorithm, in which we start from some point x_0 and repeatedly apply the proximal operator:

$$x^{t+1} = \operatorname*{prox}_{\gamma f}(x^t) = x^t - \gamma \nabla f^\gamma(x^t).$$

At convergence, we reach a minimum point $x^\star$ of the Moreau envelope, and thus a minimum of the original function. At this minimizing value, we have $\nabla f^\gamma(x^\star) = 0$, and thus $\operatorname{prox}_{\gamma f}(x^\star) = x^\star$.

Another key property of proximal operators is the Moreau decomposition for the proximal operator of $f^\star$, the dual of f:

$$\begin{aligned} x &= \operatorname*{prox}_{\lambda f}(x) + \lambda \operatorname*{prox}_{f^\star/\lambda}(\lambda x), \\ \Big(I - \operatorname*{prox}_{\lambda f}\Big)(x) &= \lambda \operatorname*{prox}_{f^\star/\lambda}(\lambda x). \end{aligned} \tag{4}$$

The Moreau identity allows one to easily alter steps within a proximal algorithm so that some computations are performed in the dual (or primal) space. Applications of this identity can also succinctly explain the relationship between a number of different optimization algorithms, as described in Section 6.

All three of these ideas—taking gradient-descent steps, projecting points onto constraint regions, and finding fixed points of suitably defined operators—arise routinely in many classical optimization algorithms. It is therefore easy to imagine that the proximal operator, which relates to all these ideas, could also prove useful.

2.2 Simple Examples of Proximal Operators

Many intermediate steps in statistical optimization problems can be written very concisely in terms of proximal operators of log-likelihoods or penalty functions. However, this conciseness is practically useful only if the proximal operator can be evaluated in closed form or at modest computational cost. Here are two simple examples where this holds.

First, Figure 1 shows a simple proximal operator and Moreau envelope. The solid black line shows the function $f(x) = |x|$, and the dotted line shows the corresponding Moreau envelope $f^1(x)$ with parameter $\gamma = 1$. The grey line shows the function $|x| + (1/2)(x - x_0)^2$ for $x_0 = 1.5$, whose minimum (shown as a red cross) defines the Moreau envelope and proximal operator. This point has ordinate $\operatorname{prox}_f(x_0) = 0.5$ and abscissa $f^1(x_0) = 1$, and is closer than x_0 to the overall minimum at $x = 0$. The blue circle shows the point $(x_0, f^1(x_0))$, emphasizing the point-wise construction of the Moreau envelope in terms of a simple optimization problem.

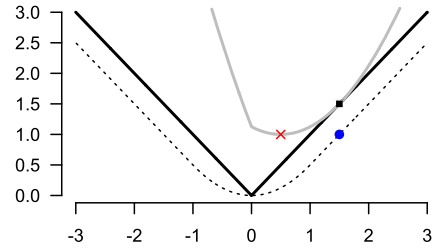

FIG. 1. *A simple example of the proximal operator and Moreau envelope. The solid black line shows the function $f(x) = |x|$, and the dotted line shows the corresponding Moreau envelope with parameter $\gamma = 1$. The grey line shows the function $|x| + (1/2)(x - x_0)^2$ for $x_0 = 1.5$, whose minimum (shown as a red cross) defines the Moreau envelope and proximal operator.*

Let $\phi(x) = \lambda\|x\|_1$ and consider the proximal operator $\operatorname{prox}_{\gamma\phi}(x)$. In this case the proximal operator is clearly separable in the components of x, and the problem that must be solved for each component is

$$\operatorname*{minimize}_{z\in\mathbb{R}} \left\{ \lambda|z| + \frac{\gamma}{2}(z - x)^2 \right\}.$$

This problem has solution

$$\begin{aligned} \hat{z} &= \operatorname*{prox}_{\lambda|x|/\gamma}(x) = \operatorname{sgn}(x)\big(|x| - \lambda/\gamma\big)_+ \\ &= S_{\lambda/\gamma}(x), \end{aligned} \tag{5}$$

the soft-thresholding operator with parameter λ/γ.

As a second example, quadratic terms of the form

$$l(x) = \tfrac{1}{2}x^\top P x + q^\top x + r \tag{6}$$

are very common in statistics. They correspond to conditionally Gaussian sampling models and arise in weighted least squares problems, in ridge regression and in EM algorithms based on scale-mixtures of normals. For example, if we assume that we observe data $(y|x) \sim \mathrm{N}(Ax, \Omega^{-1})$, then $l(x) = (y - Ax)^\top \Omega (y - Ax)/2$ or

$$P = A^\top \Omega A, \quad q = -A^\top \Omega y, \quad r = y^\top \Omega y/2$$

in the general form given above (6). If $l(x)$ takes this form, its proximal operator (with parameter $1/\gamma$) may be directly computed as

$$\operatorname*{prox}_{l/\gamma}(x) = (P + \gamma I)^{-1}\big(\gamma x - q\big),$$

assuming the relevant inverse exists.

3. PROXIMAL ALGORITHMS: SIMPLE EXAMPLES

3.1 The Proximal Gradient Method

We note by way of Introduction that starting at a point x_0 and iteratively applying the proximal operator of some function f is the most basic proximal algorithm for finding the minimum of that function. It is usually called the proximal point method or, simply, proximal iteration. It is not widely useful, since taking proximal points steps is typically no easier than simply minimizing f directly.

One of the simplest nontrivial proximal algorithms is the proximal-gradient method, which provides an important starting point for the more advanced techniques we describe in subsequent sections. Suppose as in (2) that the objective function is $F(x) = l(x) + \phi(x)$, where $l(x)$ is differentiable but $\phi(x)$ is not. An archetypal case is that of a generalized linear model with a nondifferentiable penalty designed to encourage sparsity. The proximal gradient method is well suited for such problems. It has only two basic steps which are iterated until convergence:

(1) *Gradient step*. Define an intermediate point v^t by taking a gradient step with respect to the differentiable term $l(x)$:

$$v^t = x^t - \gamma \nabla l(x^t).$$

(2) *Proximal operator step*. Evaluate the proximal operator of the nondifferentiable term $\phi(x)$ at the intermediate point v^t:

$$x^{t+1} = \operatorname*{prox}_{\gamma\phi}(v^t) = \operatorname*{prox}_{\gamma\phi}\{x^t - \gamma \nabla l(x^t)\}. \tag{7}$$

This can be motivated in several ways. We outline what is perhaps the most transparent motivation for statisticians by showing that the proximal gradient is an MM (majorize/minimize) algorithm.

Suppose that $l(x)$ has a Lipschitz-continuous gradient with modulus γ_l. This allows us to construct a majorizing function for $l(x)$, and therefore for the whole objective. Whenever $\gamma \in (0, 1/\gamma_l]$, we have the majorization

$$l(x) + \phi(x) \le l(x_0) + (x - x_0)^\top \nabla l(x_0) + \frac{1}{2\gamma}\|x - x_0\|_2^2 + \phi(x),$$

with equality at $x = x_0$. Simple algebra shows that the optimum value of the right-hand side is

$$\hat{x} = \operatorname*{argmin}_x \left\{ \phi(x) + \frac{1}{2\gamma}\|x - u\|_2^2 \right\},$$

where

$$u = x_0 - \gamma \nabla l(x_0).$$

Thus, to find the minimum of the majorizing function, we perform precisely the two steps prescribed by the proximal gradient-method: (1) form the intermediate point u taking the gradient-descent step for $l(x)$ from x_0, and (2) evaluate the proximal operator of ϕ at this point u.

The fact that we may write this method as an MM algorithm leads to the following basic convergence result. Suppose that:

1. $l(x)$ is convex with domain $\mathbb{R}^n$.
2. $\nabla l(x)$ is Lipschitz continuous with modulus γ_l, that is,
$$\|\nabla l(x) - \nabla l(z)\|_2 \le \gamma_l \|x - z\|_2 \quad \forall x, z.$$
3. ϕ is closed and convex, ensuring that $\operatorname{prox}_{\gamma\phi}$ makes sense.
4. the optimal value is finite and obtained at $x^\star$.

If these conditions are met, then the proximal gradient method converges at rate $1/t$ with fixed step size $\gamma = 1/\gamma_l$ (Beck and Teboulle, 2010).

The proximal gradient method can also be interpreted as a means for finding the fixed point of a "forward–backward" operator derived from the standard optimality conditions from subdifferential calculus. For this reason the method is sometimes referred to as forward–backward splitting. This has connections (not pursued here) with the forward–backward method for solving partial differential equations. We refer the reader to Parikh and Boyd (2013) for details.

3.2 Iterative Shrinkage Thresholding

Consider the proximal gradient method applied to a quadratic-form log-likelihood (6), as in a weighted least squares problem, with a penalty function $\phi(x)$. Then $\nabla l(x) = A^\top \Omega A x - A^\top \Omega y$, and the proximal gradient method becomes

$$x^{t+1} = \operatorname*{prox}_{\gamma^t\phi}\{x^t - \gamma^t A^\top \Omega(Ax^t - y)\}.$$

This algorithm has been widely studied under the name of IST, or iterative shrinkage thresholding (Figueiredo and Nowak, 2003). Its primary computational costs at each iteration are as follows: (1) multiplying the current iterate x^t by A, and (2) multiplying the residual $Ax^t - y$ by $A^\top \Omega$. Typically, the proximal operator for ϕ will be simple to compute, as in the case of a quadratic or ℓ^1-norm/Lasso penalty discussed in the previous section. Thus, the evaluation of the proximal operator will contribute a negligible amount to the overall complexity of the algorithm.

3.3 Proximal Newton

As we have described, the proximal gradient method is a generalization of classical gradient approaches. It uses only first-order information about the smooth term $l(x)$. However, one can naturally use higher-order expansions to construct different envelopes that take into account second-order information about l, leading to improvements analogous to the manner in which Newton's method improves upon gradient descent. Consider a family of functions of the form

$$F_H(x,z) = l(z) + \nabla l(z)^\top (x-z) + \tfrac{1}{2}(x-z)^\top H_z (x-z),$$

and use this to define an envelope in the manner of (3). Then we can calculate the generalized proximity operator

$$\operatorname*{prox}_{F_H}(z) = z - \left(\gamma^{-1} I + H_z\right)^{-1} \nabla l(z). \tag{8}$$

Instead of directly using the Hessian, $H_z = \nabla^2 l(z)$, approximations can be employed, leading to quasi-Newton-style approaches. As we will soon describe, the second-order bound, and approximations to the Hessian, are one way to interpret the half-quadratic (HQ) approach, as well as introduce quasi-Newton methods into the proximal framework. Proximal Newton methods are even possible for some nonconvex problems, as in Chouzenoux, Pesquet and Repetti (2014) and Appendix D.

3.4 Nesterov Acceleration

A useful feature of proximal algorithms is the ability to use acceleration techniques (Nesterov, 1983), often referred to as Nesterov acceleration. Acceleration leads to nondescent algorithms that can provide substantial increases in efficiency versus their nonaccelerated counterparts.

The idea of acceleration is to add an intermediate "momentum" variable z, prior to evaluating the forward and backward steps:

$$\begin{aligned} z^{t+1} &= x^t + \theta_{t+1}\left(\theta_t^{-1} - 1\right)\left(x^t - x^{t-1}\right), \\ x^{t+1} &= \operatorname*{prox}_{\gamma^{-1}\phi}\left(z^{t+1} - \gamma^{-1}\nabla l\left(z^{t+1}\right)\right), \end{aligned}$$

where standard choices are $\theta_t = 2/(t+1)$ and $\theta_{t+1}(\theta_t^{-1} - 1) = (t-1)/(t+2)$.

When ϕ is convex, the proximal problem is strongly convex, and advanced acceleration techniques can be used (Zhang, Saha and Vishwanathan, 2010; Meng and Chen, 2011).

4. REDUNDANCY, SPLITTING AND THE AUGMENTED LAGRANGIAN

4.1 Overview

In this section we show how the splitting technique described in the Introduction leads to many well-known proximal algorithms. As a running example, consider the problem of minimizing $l(x)+\phi(x)$, where we apply the splitting strategy to formulate the equivalent problem

$$\begin{aligned} &\operatorname*{minimize}_{x,z} \quad l(x) + \phi(z) \\ &\text{subject to} \quad x - z = 0. \end{aligned} \tag{9}$$

The advantage of such a variable-splitting approach is that now the fit and penalty terms are decoupled in the objective function of the primal problem. A standard tactic for exploiting this fact is to write down and solve the dual problem corresponding to the original (primal) constrained problem. This is sometimes referred to as *dualization*. Many well-known references exist on this topic (e.g., Bertsekas, 2011). For this reason we focus on problem formulation and algorithms for solving (9), avoiding standard material on duality or optimality conditions.

4.2 Dual Ascent, the Augmented Lagrangian and Scaled Form

Consider first the ordinary Lagrangian of problem (9):

$$L(x,z,\lambda) = l(x) + \phi(z) + \lambda^\top (x-z),$$

with Lagrange multiplier λ. The dual function is $g(\lambda) = \inf_{x,z} L(x,z,\lambda)$, and the dual problem is to maximize $g(\lambda)$.

Let $p^\star$ and $d^\star$ be the optimal values of the primal and dual problems, respectively. Assuming that strong duality holds, the optimal values of the primal and dual problems are the same. Moreover, we may recover a primal-optimal point $(x^\star, z^\star)$ from a dual-optimal point $\lambda^\star$ using the fact that

$$\begin{aligned} \left(x^\star, z^\star\right) &= \operatorname*{argmin}_{x,z} L\left(x,z,\lambda^\star\right) \\ &\iff \quad 0 \in \partial_{x,z} L\left(x^\star, z^\star, \lambda^\star\right). \end{aligned}$$

The idea of dual ascent is to solve the dual problem using gradient ascent, exploiting the fact that

$$\nabla g(\lambda) = \nabla_\lambda L(\hat{x}_\lambda, \hat{z}_\lambda, \lambda),$$

where

$$(\hat{x}_\lambda, \hat{z}_\lambda) = \operatorname*{argmin}_{x,z} L(x,z,\lambda).$$

Thus, the required gradient is simply the residual for the primal constraint: $\nabla_\lambda L(x,z,\lambda) = x - z$. Therefore, dual ascent involves iterating two steps:

$$(x^{t+1}, z^{t+1}) = \operatorname*{argmin}_{x,z} L(x, z, \lambda^t),$$
$$\lambda^{t+1} = \lambda^t + \alpha_t(x^{t+1} - z^{t+1})$$

for appropriate step size α_t.

An obvious issue with dual ascent for problem (9) is that the update in x and z must be done jointly, rather than one at a time. This is rarely practical for problems of this form. But a discussion of dual ascent is an important starting point for building up to more realistic algorithms.

We also note that in the case where g is not differentiable, it is possible to replace the gradient with the negative of a subgradient of $-g$, leading to dual subgradient ascent; see Shor (1985).

Augmented Lagrangian and the method of multipliers. Take problem (9) as before, with Lagrangian $L(x,z,\lambda) = l(x) + \phi(z) + \lambda^\top(x - z)$. The augmented-Lagrangian approach (also known as the method of multipliers) seeks to stabilize the intermediate steps of dual ascent by adding a ridge-like term to the Lagrangian:

$$L_\gamma(x,z,\lambda) = l(x) + \phi(z) + \lambda^\top(x - z) + \frac{\gamma}{2}\|x - z\|_2^2,$$

where γ is a scale or step-size parameter. One way of viewing this augmented Lagrangian is as the standard Lagrangian for the equivalent problem

$$\begin{aligned} \operatorname*{minimize}_{x,z} \quad & l(x) + \phi(z) + \frac{\gamma}{2}\|x - z\|_2^2 \\ \text{subject to} \quad & x - z = 0. \end{aligned}$$

We can see that this is equivalent to the original because, for any primal-feasible x and z, the new objective takes the same value as the original objective, and thus has the same minimum. The dual function corresponding to this augmented Lagrangian is $g_\gamma(\lambda) = \inf_{x,z} L_\gamma(x,z,\lambda)$, which is differentiable and strongly convex under mild conditions. (The ordinary dual function need not be either of these things, which is a key advantage of using the augmented Lagrangian.)

The method of multipliers is to use dual ascent for the modified problem, iterating

$$(x^{t+1}, z^{t+1}) = \operatorname*{argmin}_{x,z} L_\gamma(x, z, \lambda^t),$$
$$\lambda^{t+1} = \lambda^t + \gamma(x^{t+1} - z^{t+1}).$$

Thus, the dual-variable update does not change compared to standard dual ascent. But the joint (x,z) update has a regularization term added to it, whose magnitude depends upon the tuning parameter γ. Notice that the step size γ is used in the dual-update step.

Scaled form. Many proximal algorithms have more concise updates when the dual variable λ is expressed in scaled form. Specifically, rescale the dual variable as $u = \gamma^{-1}\lambda$. We can rewrite the augmented Lagrangian in terms of u as

$$\begin{aligned} &L_\gamma(x,z,u) \\ &\quad = l(x) + \phi(z) + \gamma u^\top(x - z) + \frac{\gamma}{2}\|x - z\|_2^2 \\ &\quad = l(x) + \phi(z) + \frac{\gamma}{2}\|r + u\|_2^2 - \frac{\gamma}{2}\|u\|_2^2, \end{aligned}$$

where $r = x - z$ is the primal residual. This leads to the following dual-update formulas:

$$\begin{aligned} &(x^{t+1}, z^{t+1}) \\ &\quad = \operatorname*{argmin}_{x,z}\left\{ l(x) + \phi(z) + \frac{\gamma}{2}\|x - z + u^t\|_2^2 \right\}, \\ &u^{t+1} = u^t + (x^{t+1} - z^{t+1}). \end{aligned}$$

Bregman iteration. The augmented Lagrangian method for solving ℓ^1-norm problems is called "Bregman iteration" in the compressed-sensing literature. Here the goal is to solve the exact-recovery problem via basis pursuit:

$$\begin{aligned} \operatorname*{minimize}_{x} \quad & \|x\|_1 \\ \text{subject to} \quad & Ax = y, \end{aligned}$$

where y is measured, x is the unknown signal, and A is a known "short and fat" sensing matrix (meaning more coordinates of x than there are observations).

The scaled-form augmented Lagrangian corresponding to this problem is

$$L_\gamma(x,u) = \|x\|_1 + \frac{\gamma}{2}\|Ax - y + u\|_2^2 - \frac{\gamma}{2}\|u\|_2^2,$$

with steps

$$x^{t+1} = \operatorname*{argmin}_x \left\{ \|x\|_1 + \frac{\gamma}{2}\|Ax - z^t\|_2^2 \right\},$$
$$z^{t+1} = y + z^t - Ax^{t+1},$$

where we have redefined $z^t = y - u^t$ compared to the usual form of the dual update. Thus, each intermediate step of the Bregman iteration is like a lasso regression problem. (This algorithm also has an alternate derivation in terms of Bregman divergences, hence its name.)

4.3 ADMM

The alternating-direction method of multipliers (or ADMM) is a proximal algorithm that combines three ideas for solving problems like (9): splitting, the augmented Lagrangian, and alternating-direction updates. Recall that the scaled-form augmented Lagrangian for this problem is

$$L_\gamma(x,z,u) = l(x)+\phi(z)+\frac{\gamma}{2}\|x-z+u\|_2^2-\frac{\gamma}{2}\|u\|_2^2.$$

ADMM is similar to dual ascent for this problem, except that we optimize the Lagrangian in x and z individually, rather than jointly, in each pass. (Hence, "alternating direction.") For our problem, the updates become

$$\begin{aligned} x^{t+1} &= \operatorname*{argmin}_x\left\{l(x)+\frac{\gamma}{2}\|x-z^t+u^t\|_2^2\right\} \\ &= \operatorname*{prox}_{l/\gamma}(z^t-u^t), \\ z^{t+1} &= \operatorname*{argmin}_z\left\{\phi(z)+\frac{\gamma}{2}\|x^{t+1}-z+u^t\|_2^2\right\} \\ &= \operatorname*{prox}_{\phi/\gamma}(x^{t+1}+u^t), \\ u^{t+1} &= u^t+x^{t+1}-z^{t+1}. \end{aligned}$$

The first two steps are to evaluate the proximal operators of l and ϕ, respectively.

4.4 Divide and Concur

Divide and Concur (e.g., Gravel and Elser, 2008) is another type of splitting strategy that provides a general approach to statistical models that require optimization of a sum of $J+1$ composite functions of the form

$$\operatorname*{minimize}_x \quad \sum_{j=1}^{J} l_j(A_j x)+\phi(x).$$

In Divide and Concur, we add slack variables z_j for $j\in\{1,\dots,J+1\}$ to divide the problem together, with equality constraints so that the solutions concur. Specifically, we form the equivalent constrained optimization problem

$$\begin{aligned} \operatorname*{minimize}_{x,z} \quad & \sum_{j=1}^{J+1} l_j(z_j) \\ \text{subject to} \quad & z_j = A_j x, \end{aligned}$$

where $l_{J+1}=\phi$ and $A_{J+1}=I$. This can be solved using an iterative proximal splitting algorithm (e.g., multiple ADMM, split Bregman). For example, under ADMM (Parikh and Boyd, 2013) the updates are

$$\begin{aligned} x_j^{t+1} &= \operatorname*{prox}_{\lambda l_j\circ A_j}(\bar{x}^t-u_j^k), \\ u_j^{t+1} &= u_j^t+x_j^{t+1}-\bar{x}^{t+1}, \end{aligned}$$

where $\bar{x}^t=\frac{1}{J+1}\sum_{j=1}^{J+1}x_j^t$.

Divide and Concur methods provide a natural approach to hierarchical models or to very large problems—for example, where each l_j corresponds to a negative log-likelihood for a subset of the data stored on one machine. In this case, DC allows the overall problem to be broken into many tractable, independently computable subproblems via splitting. Only the intermediate solutions to these subproblems, rather every subset of the actual data, need to be broadcast between machines.

4.5 Other Forms of Redundancy

Other redundant parameterizations are certainly possible, beyond the basic splitting strategy considered here. For example, consider the case of an exponential-family model for outcome y with cumulant-generating function $\psi(z)$ and with natural parameter z:

$$p(y)=p_0(y)\exp\{yz-\psi(z)\}.$$

There is a unique Bregman divergence associated with every exponential family. It corresponds precisely to the relationship between the natural parameterization and the mean-value parameterization. There is a corresponding class of Bregman proximal point algorithms.

In a generalized linear model, the natural parameter for outcome y_i is a linear regression on covariates, $z_i=a_i^\top x$. In this case $l(x)$ may be written as

$$l(x)=\sum_{i=1}^{N} l_i(x) \quad \text{where } l_i(x)=\psi(a_i^\top x)-y_i(a_i^\top x),$$

up to an additive constant not depending on x. Now introduce slack variables $z_i=a_i^\top x$. This leads to the equivalent primal problem

$$\begin{aligned} \min_{x,z} \quad & \sum_{i=1}^{N}\{\psi(z_i)-y_i z_i\}+\phi(x) \\ \text{subject to} \quad & Ax-z=0. \end{aligned}$$

For example, in a Poisson model $(y_i|\mu_i)\sim\text{Pois}(\mu_i)$, $\mu_i=\exp(\theta_i)$ with natural parameter $\theta_i=a_i^\top x$. The cumulant generating function is $b(\theta)=\exp(\theta)$, and thus

$d(\mu) = \mu \log \mu - \mu$. After simplification, the divergence $D_d(y, \mu) = \mu - y \log \mu + (\mu - y)$. The optimization problem can then be split as

$$\min_{x,z} \quad \sum_{i=1}^{N} (z_i - y_i \log z_i) + \phi(x)$$
$$\text{subject to} \quad a_i^\top x = \log z_i.$$

These same optimization problems arise when one considers scale mixtures, or convex variational forms (Palmer et al., 2005, Polson and Scott, 2015). The connection is made explicit by the dual function for a density and its relationship with scale-mixture decompositions. For instance, one can obtain the following equality for appropriate densities $p(x), q(z)$ and constants μ, κ:

$$\begin{aligned} -\log p(x) &= -\sup_{z>0} \log\bigl(p_N\bigl(x; \mu + \kappa/z, z^{-1}\bigr) q(z)\bigr) \\ &= \inf_{z>0} \left\{ \frac{z}{2}(x - \mu - \kappa/z)^2 - \log\bigl(\sqrt{z} q(z)\bigr) \right\}, \end{aligned}$$

where $p_N(x; \mu, \sigma^2)$ is the density function for a normal distribution with mean μ and variance σ^2. The form resulting from this normal scale-mixture envelope is similar to the half-quadratic envelopes described in Section 5. Polson and Scott (2015) describe these relationships in further detail.

5. ENVELOPE METHODS

In this section we describe several types of envelopes: the *forward–backward* (FB) envelope, the *Douglas–Rachford* (DR) envelope, the *half-quadratic* (HQ) envelope, and the *Bregman divergence* envelopes. These all build upon the idea of a Moreau envelope and lead to analogous proximal algorithms. Within this framework, various algorithms may be generated in terms of gradient steps for the corresponding envelope. (For instance, ADMM methods will be viewed as the gradient step of the dual FB envelope.) Section 6 dissects these envelopes in further detail, shows their relationship to Lagrangian approaches, and provides a framework within which they can be derived and extended.

5.1 Forward–Backward Envelope

Suppose as in (9) that we have to minimize $F = l + \phi$, under the assumptions that l is strongly convex and possesses a continuous gradient with Lipschitz constant γ_l, so that $|\nabla^2 l(x)| \le \gamma_l$, and that ϕ is proper lower semi-continuous and convex.

First, we define the FB envelope, $F_\gamma^{\mathrm{FB}}(x)$, which will possess some desirable properties (see Patrinos and Bemporad, 2013):

$$\begin{aligned} F_\gamma^{\mathrm{FB}}(x) &:= \min_v \left\{ l(x) + \nabla l(x)^\top (v - x) + \phi(v) \right. \\ &\qquad \left. + \frac{1}{2\gamma} \|v - x\|^2 \right\} \\ &= l(x) - \frac{\gamma}{2} \|\nabla l(x)\|^2 + \phi^\gamma\bigl(x - \gamma \nabla l(x)\bigr). \end{aligned}$$

If we pick $\gamma \in (0, \gamma_l^{-1})$, the matrix $I - \gamma \nabla^2 l(x)$ is symmetric and positive definite. The stationary points of the envelope $F_\gamma^{\mathrm{FB}}(x)$ are the solutions $x^\star$ of the original problem which satisfy $x = \mathrm{prox}_{\gamma\phi}(x - \gamma \nabla l(x))$. This follows from the derivative information

$$\nabla F_\gamma^{\mathrm{FB}}(x) = \bigl(I - \gamma \nabla^2 l(x)\bigr) G_\gamma(x),$$

where $G_\gamma(x) = \gamma^{-1}(x - P_\gamma(x))$ and $P_\gamma(x) = \mathrm{prox}_{\gamma\phi}(x - \gamma \nabla l(x))$.

With these definitions, we can establish the following descent property for gradient steps based on the FB envelope:

$$\begin{aligned} F_\gamma^{\mathrm{FB}}(x) &\le F(x) - \frac{\gamma}{2} \|G_\gamma(x)\|^2, \\ F\bigl(P_\gamma(x)\bigr) &\le F_\gamma^{\mathrm{FB}}(x) - \frac{\gamma}{2}(1 - \gamma\gamma_l) \|G_\gamma(x)\|^2. \end{aligned}$$

Hence, for $\gamma \in (0, \gamma_l^{-1})$, the envelope value always decreases on application of the proximal operator of $\gamma\phi$, and we can determine the stationary points. See Appendix A for further details.

5.2 Douglas–Rachford Envelope

Mimicking the forward–backward approach, Patrinos, Lorenzo and Alberto (2014) define the Douglas–Rachford (DR) envelope as

$$\begin{aligned} F_\gamma^{\mathrm{DR}}(x) &= l^\gamma(x) - \frac{\gamma}{2} \|\nabla l^\gamma(x)\|_2^2 + \phi^\gamma\bigl(x - 2\gamma \nabla l^\gamma(x)\bigr) \\ &= \min_z \left\{ l(x^\star) + \nabla l(x^\star)^\top (z - x^\star) + \phi(z) \right. \\ &\qquad \left. + \frac{1}{2\gamma} \|z - x^\star\|^2 \right\}, \end{aligned}$$

where we recall that l^γ is the Moreau envelope of the function l and $x^\star = \mathrm{prox}_{\gamma l}(x)$.

This can be interpreted as a backward–backward envelope. It is a special case of a FB envelope evaluated at the proximal operator of γl, namely,

$$F_\gamma^{\mathrm{DR}}(x) = F_\gamma^{\mathrm{FB}}\Bigl(\mathop{\mathrm{prox}}_{\gamma l}(x)\Bigr).$$

Again, the gradient of this envelope produces a proximal algorithm (see Patrinos, Lorenzo and Alberto, 2014) which converges to the minimum of $\{l(x) + \phi(x)\}$. The iterations are

$$w^{t+1} = \underset{\gamma l}{\text{prox}}(x^t),$$
$$z^{t+1} = \underset{\gamma \phi}{\text{prox}}(2w^t - x^t),$$
$$x^{t+1} = x^t + (z^t - w^t).$$

There are many ways to rearrange the basic DR algorithm. For example, with an intermediate variable, $v = w - x$, we could equally well iterate

$$w^{t+1} = \underset{\gamma l}{\text{prox}}(x^t - v^t),$$
$$x^{t+1} = \underset{\gamma \phi}{\text{prox}}(w^t + v^t), \quad v^{t+1} = v^t + (w^t - x^t).$$

5.3 Half-Quadratic Envelopes

We now provide an illustration of a quasi-Newton algorithm within the class of Half-Quadratic (HQ) optimization problems (Geman and Yang, 1995; Geman and Reynolds, 1992). This envelope applies to the commonly used L^2-norm where $l(x) = \|Ax - y\|^2$, and can be used in conjunction with some nonconvex ϕ. See Nikolova and Ng (2005) for convergence rates and comparisons of the different algorithms.

The half-quadratic (HQ) envelope is defined by

$$F^{\text{HQ}}(x) = \inf_v \{Q(x, v) + \psi(v)\},$$

where

$$Q(x, v) = vx^2 \quad \text{or} \quad (v - x)^2,$$

that is, the function $Q(x, v)$ is "half-quadratic" in the variable v. In the HQ framework, the term $\psi(v)$ is usually understood to be the convex conjugate of some function, for example, $\psi(v) = \phi^\star(x)$.

As an initial example, suppose that we wish to minimize the function

$$F(x) = \tfrac{1}{2}\|Ax - y\|^2 + \gamma \Phi(x),$$

where

$$\Phi(x) = \sum_{i=1}^{d} \phi\big((B^\top x - b)_i\big),$$

and that the penalty is specified in terms of the representation $\phi(x) = F^{\text{HQ}}(x)$. Then we need to minimize the higher-dimensional function

$$F(x, v) = \frac{1}{2}\|Ax - y\|^2 + \gamma \sum_{i=1}^{d} Q(\delta_i, v_i) + \gamma \sum_{i=1}^{d} \psi(v_i),$$

where $\delta_i = (B^\top x - b)_i$.

This establishes an equivalence between gradient linearization and quasi-Newton. These algorithms give the iterative mappings

$$x^{t+1} = L\big(\hat{v}(x^t)\big)^{-1} A^\top y$$

and

$$x^{t+1} = x^t - L\big(x^t\big)^{-1} \nabla_x F\big(x^t\big),$$

respectively, where $L(x^t)$ is a step size function. They turn out to be identical, with derivative information

$$\begin{aligned}\nabla_x F(x) &= A^\top Ax - A^\top y + \gamma \sum_{i=1}^{d} B_i \frac{\phi'(\|\delta_i\|)}{\|\delta_i\|} B_i^\top x \\ &= \big(A^\top A + \gamma B\, \mathrm{V}(x) B^\top\big)x - A^\top y \\ &= L\big(\hat{v}(x)\big)x - A^\top y\end{aligned}$$

for $\mathrm{V}(x) = \text{diag}(\hat{v}(\|\delta\|_{i=1}^d))$ and $L(\hat{v}(x)) = A^\top A + vB\,\mathrm{V}(x)B^\top$.

See Polson and Scott (2015) for further explanation of the half-quadratic class of penalties.

5.4 Bregman Divergence Envelopes

Many statistical models, such as those generated by an exponential family distribution, can be written in terms of a Bregman divergence. One is then faced with the joint minimization of an objective function of the form $F(x, v) = D(x, v) + \phi(x) + \psi(v)$. To minimize over (x, v), we can use an alternating Bregman projection method. To perform the minimization of v given x, we can make use of the D-Moreau envelope, which is defined by

$$\phi^D(x) = \inf_v \{D(x, v) + \phi(v)\},$$

where $D(x, v)$ is a Bregman divergence. A key feature here is that a Bregman divergence satisfies a three-point law of cosines triangle inequality, which helps to establish the descent property for proximal algorithms derived from these envelopes (see Appendix A). Many commonly used EM, MM and variational EM algorithms in statistics implicitly use envelopes of this type.

6. PROXIMAL ALGORITHMS FOR COMPOSITE FUNCTIONS

6.1 Overview

Building off the general objective in (1), we now consider the optimization of a general composite objective of the form

$$F(x) := l(x) + \phi(Bx)$$

or, in split form,

$$\begin{aligned} &\underset{x,z}{\text{minimize}} && l(x)+\phi(z) \\ &\text{subject to} && Bx=z. \end{aligned} \tag{10}$$

Composite penalties arise in statistical models that account for structural constraints or spatiotemporal correlations (e.g., Tibshirani and Taylor, 2011; Tibshirani, 2014; Tansey et al., 2014). The most famous examples of problems in this class are total-variation denoising (Rudin, Osher and Faterni, 1992) and the fused lasso (Tibshirani et al., 2005).

We start by noting that many approaches for solving this problem, including the ones in Section 4, can be characterized in terms of one of the four general forms of the objective functions/Lagrangians that result from appealing to splitting and conjugate functions:

$$\begin{aligned} \text{primal} \qquad & F(x)=l(x)+\phi(Bx), \\ \text{split–primal} \qquad & F_{\mathrm{SP}}(x,z,\lambda)=l(x)+\phi(z) \\ & \qquad +\lambda^{\top}(Bx-w), \\ \text{primal–dual} \qquad & F_{\mathrm{PD}}(x,\lambda)=l(x)+\lambda^{\top}(Bx) \\ & \qquad -\phi^{\star}(\lambda), \\ \text{split–dual} \qquad & F_{\mathrm{SD}}(x,z,\lambda)=l^{\star}(z)+\phi^{\star}(\lambda) \\ & \qquad +x^{\top}\bigl(-B^{\top}\lambda-z\bigr). \end{aligned}$$

From a statistical perspective, it is natural to think of z and λ as latent variables, and of each of these splitting/duality strategies as defining a higher-dimensional objective function. Such ideas are familiar in statistics, where alternating minimization, iterated conditional mode (ICM), EM and MM algorithms have a long history (e.g., Dempster, Laird and Rubin, 1977; Csiszár and Tusnády, 1984; Besag, 1986). Indeed, Polson and Scott (2015) show how many such algorithms that appeal to convex conjugacy have a natural EM-like interpretation in terms of missing data.

For problem (10), the motivation for using the primal–dual and the split forms (see Esser, Zhang and Chan, 2010) lies in how they decouple ϕ from the linear mapping B; it is precisely the composition of these functions that poses the difficulty for problems like TV denoising and the fused lasso. Note that the primal–dual formulation follows from profiling the slack variable z out of the split–primal objective:

$$\begin{aligned} \inf_{z} L(x,z,\lambda) &= \inf_{z}\bigl\{l(x)+\phi(z)+\lambda^{\top}(Bx-z)\bigr\} \\ &= l(x)+\lambda^{\top}Bx-\phi^{\star}(\lambda), \end{aligned}$$

and the split–dual by a similar argument. These two formulations are related via the Max–Min inequality (Boyd and Vandenberghe, 2004):

$$\sup_{q}\inf_{v} F(q,v)\le\inf_{v}\sup_{q} F(q,v).$$

In the special case of closed proper convex functions, we have

$$\begin{aligned} \min_{x} F(x) &= \min_{x}\sup_{z} F_{\mathrm{PD}}(x,\lambda) \\ &= \max_{\lambda}\min_{x,z} F_{\mathrm{SP}}(x,z,\lambda) \\ &= \max_{x}\min_{\lambda,z} F_{\mathrm{SD}}(x,z,\lambda), \end{aligned}$$

where we exploit the fact that

$$\phi(Bx)=\sup_{z}\bigl\{z^{\top}Bx-\phi^{\star}(z)\bigr\}$$

whenever ϕ is convex. $F_{\mathrm{SP}}(x,z,\lambda)$ and $F_{\mathrm{PD}}(x,\lambda)$ are also related by

$$\begin{aligned} \min_{z\ge 0} F_{\mathrm{SP}}(x,z,\lambda) &= \min_{z\ge 0}\bigl\{\phi(z)+l(x)+\lambda^{\top}(Bx-z)\bigr\} \\ &= l(x)+\lambda^{\top}Bx+\min_{z\ge 0}\bigl\{\phi(z)-\lambda^{\top}z\bigr\} \\ &= l(x)+\lambda^{\top}Bx-\phi^{\star}(\lambda) \\ &= F_{\mathrm{PD}}(x,\lambda). \end{aligned}$$

6.2 Proximal Solutions

In most statistical problems of form (10), it is typically the case that closed-form expressions for one or more of $l(x)$, $l^{\star}(z)$, $\phi(z)$ or $\phi^{\star}(\lambda)$ will be unavailable or inefficient to compute. However, exact solutions to related problems that share the same critical points may be easily accessible. We now step through several such approaches for solving (10), explaining how they relate to the ideas introduced thus far. We highlight whenever proximal operators enter the analysis. Because proximal operators are so well understood, their presence in an algorithm is convenient: the properties of proximal operators and the associated fixed-point theory can simplify otherwise lengthy constructions and convergence arguments. Moreover, by exploiting the proximal operator's known properties, like the Moreau identity, one can move easily between the different formulations above, and thus between the primal and dual spaces. It is also worth mentioning that the efficacy of certain acceleration techniques can depend on which formulation is used, and therefore implicitly on the specific proximal steps taken. We refer the reader to Beck and Teboulle (2014) for further discussion.

First, proximal operators arise naturally whenever we augment the Lagrangian for problem (10), which entails adding a ridge term to the split–primal objective:

$$\begin{aligned} L_\rho(x,z,\lambda) &= l(x)+\phi(z)+\lambda^\top(Bx-z) \\ &\quad +\frac{\rho}{2}\|Bx-z\|^2 \\ &= F_{\mathrm{SP}}(x,z,\lambda)+\frac{\rho}{2}\|Bx-z\|^2. \end{aligned} \tag{11}$$

As already detailed, this leads naturally to an ADMM algorithm whose intermediate iterates involve proximal operators.

Second, we also are not restricted to using the proximal operators directly implied by one of these four problem formulations, such as those that appear when l, $l^\star$, ϕ and/or $\phi^\star$ contain quadratic terms. We can also apply a surrogate or approximation (e.g., an envelope or majorizer) to certain terms. For example, when exact solutions to the composite proximal operator are not available, one can consider "linearizing" $\frac{\rho}{2}\|Bx-z\|^2$ with $\frac{\rho}{2\lambda_B}\|x-z\|^2$, where $\sigma_{\max}(B^\top B)<\lambda_B$, yielding

$$\begin{aligned} &F_{\mathrm{SP}}(x,w,z)+\frac{\rho}{2}\|B_x-z\|^2 \\ &\quad\le F_{\mathrm{SP}}(x,w,z)+\frac{\rho}{2\lambda_B}\|x-z\|^2. \end{aligned}$$

This approach can be seen as a simple majorization and, when combined with the proximal solution for z, as a forward–backward envelope for the subproblem. Implementations of this approach include the linearized ADMM technique or the split inexact Uzawa method, and are described in the context of Lagrangians by Chen and Teboulle (1994) and primal–dual algorithms in Chambolle and Pock (2011). Magnússon et al. (2014) detail splitting methods in terms of augmented-Lagrangians for nonconvex objectives.

Finally, one can represent one of the terms in the objective using one of the envelopes described in Section 5, in which case the iterates of the resulting algorithm will involve proximal operators. In fact, the envelope representation can itself be seen as a way to encode the iterates in each of a problem's latent/slack/splitting terms as proximal operators.

An example: *The primal–dual*. To demonstrate these ideas, we give an example of how proximal operators and their properties can be used to derive an algorithm starting from the primal–dual formulation

$$\max_\lambda \min_x \{l(x)+\lambda^\top(Bx)-\phi^\star(\lambda)\}.$$

First, notice that the argmin for the subproblem in x, $l(x)+\lambda^\top(Bx)$, can be characterized in terms of the following fixed point whenever $\gamma_l>0$:

$$x^\star = \operatorname*{prox}_{\gamma_l(l(x)+\lambda^\top Bx)}(x^\star).$$

We now use the fact that

$$\operatorname*{prox}_{g(z)+u^\top z}(q)=\operatorname*{prox}_g(q-u), \tag{12}$$

for a generic function $g(z)$ and variables q, z and u; this is obtained by completing the square in the definition of the operator. Appealing to (12) gives

$$x^\star = \operatorname*{prox}_{\gamma_l(l(x)+\lambda^\top Bx)}(x^\star)=\operatorname*{prox}_{\gamma_l l}(x^\star-\gamma_l B^\top\lambda).$$

Now we're left with only the subproblem in λ:

$$\begin{aligned} &\max_\lambda\{l(x^\star)+\lambda^\top(Bx^\star)-\phi^\star(\lambda)\} \\ &\quad=-\min_\lambda\{\phi^\star(\lambda)-\lambda^\top(Bx^\star)-l(x^\star)\}. \end{aligned}$$

We can take yet another proximal step, for the minimization of $\phi^\star(\lambda)-\lambda^\top(Bx^\star)$, in λ with step size γ_ϕ. Using (12) and (4), we find that the argmin satisfies

$$\lambda^\star=\operatorname*{prox}_{\gamma_\phi\phi^\star}(\lambda^\star+\gamma_\phi Bx^\star).$$

Using the Moreau decomposition in (4), we can derive yet another strategy. Note that

$$\begin{aligned} &\operatorname*{prox}_{\gamma_\phi\phi^\star}(\lambda^\star+\gamma_\phi Bx^\star) \\ &\quad=\frac{1}{\gamma_\phi}\Big(I-\operatorname*{prox}_{\phi/\gamma_\phi}\Big)\circ(\gamma_\phi(\lambda^\star+Bx^\star)). \end{aligned}$$

Hence, we can characterize the solution to the primal–dual problem in terms of fixed points of the following two operators:

$$\begin{aligned} x^\star &= \operatorname*{prox}_{\gamma_l l}(x^\star-\gamma_l B^\top\lambda^\star), \\ \lambda^\star &= \frac{1}{\gamma_\phi}\Big(I-\operatorname*{prox}_{\phi/\gamma_\phi}\Big)\circ(\gamma_\phi(\lambda^\star+Bx^\star)). \end{aligned} \tag{13}$$

If we separate the last step implied by (13) into two steps and simplify by setting $\gamma_l=\gamma_\phi=1$, we arrive at

$$\begin{aligned} x^\star &= \operatorname*{prox}_l(x^\star-B^\top u^\star), \\ w^\star &= \operatorname*{prox}_\phi(u^\star+Bx^\star), \\ u^\star &= u^\star-(w^\star-Bx^\star). \end{aligned}$$

This has the same basic form of techniques like ADMM, alternating split Bregman, split inexact Uzawa and so forth. See Chen, Huang and Zhang (2013) for more details.

6.3 Composition in General Quadratic Envelopes

Consider now the most general form of a quadratic envelope involving a composite penalty function:

$$F_\Lambda(x) = \inf_z \left\{ \frac{1}{2} x^\top \Lambda(z) x - \eta^\top(z) x + \phi(Bx) \right\}, \tag{14}$$

where $\Lambda(z)$ is symmetric positive definite. Such forms can arise when one majorizes $l(x)$ using a second-order approximation of around z. This general quadratic case in which $\Lambda(z)$ is not necessarily diagonal encompasses the approaches of Geman and Yang (1995), Geman and Reynolds (1992), and can be addressed with splitting techniques.

If $B^\top B$ is positive definite, a proximal point solution can be obtained by setting $l(x) = x^\top \Lambda(z) x - \eta^\top x$ in (13). The general solution to a quadratic-form proximal operator (6), together with the split–dual formulation, implies a proximal point algorithm that exploits the fact that the optimal values satisfy

$$\begin{aligned} x^\star &= \operatorname*{prox}_{\gamma_l l(x)}\left(x^\star - \gamma_l B^\top z^\star\right) \\ &= \left(I + \gamma_l \Lambda(z^\star)\right)^{-1}\left(x^\star - \gamma_l B^\top z^\star + \gamma_l \eta\right), \\ z^\star &= \frac{1}{\gamma_\phi}\left(I - \operatorname*{prox}_{\phi/\gamma_\phi}\right) \circ \left(\gamma_\phi\left(z^\star + Bx^\star\right)\right). \end{aligned}$$

This formulation introduces the subproblem of solving a system of linear equations. Using the exact solution to this system would reflect methods that involve Levenberg–Marquardt steps, quasi-Newton methods and Tikhonov regularization, and is related to the use of second-order Taylor approximations to an objective function. Naturally, the efficiency of computing exact solutions depends very much on the properties of $I + \gamma_l \Lambda(z)$, since the system defined by this term will need to be solved on each iteration of a fixed-point algorithm. When $\Lambda(z)$ is constant, a decomposition can be performed at the start and reused, so that solutions are computed quickly at each step. For some matrices, this can mean only $O(n)$ operations per iteration. In general, however, the post-startup iteration cost is $O(n^2)$.

Other approaches, like those in Chen, Huang and Zhang (2013) and Argyriou et al. (2011), do not attempt to directly solve the aforementioned system of equations. Instead they use a forward–backward algorithm on the dual objective, F_{PD}. In particular, we call attention to the approach of Argyriou et al. (2011). They show how to evaluate the proximal operator of $\phi(Bx)$ directly, by finding the fixed point of the operator

$$H_k = \kappa I + (1 - \kappa) H,$$

for $\kappa \in (0, 1)$, where

$$H(v) := \left(I - \operatorname*{prox}_{\gamma^{-1}\phi}\right)\left(BA^{-1}\eta + \left(I - \gamma BA^{-1}B^\top\right)v\right) \qquad \forall v \in \mathbb{R}^p.$$

Here $0 < \gamma < 2/\sigma_{\max}(BA^{-1}B^\top)$ and $A = \Lambda(z)$. The operator H is understood to be nonexpansive, so, by Opial's theorem, one is guaranteed convergence; when H is a contraction, this convergence is linear. After finding the fixed point $v^\star$, one sets $x^\star = A^{-1}(\eta - xB^\top v^\star)$.

7. APPLICATIONS

7.1 Logit Loss Plus Lasso Penalty

To illustrate our approach, we simulate observations from the model

$$\begin{aligned} (y_i \mid p_i) &\sim \mathrm{Binom}(m_i, p_i), \\ p_i &= \mathrm{logit}^{-1}(a_i^\top x), \end{aligned}$$

where $i = 1, \ldots, 100$, $a_i^\top$ is a row vector of $A \in \mathbb{R}^{100\times 300}$, $x \in \mathbb{R}^{300}$. The A matrix is simulated from $\mathrm{N}(0, 1)$ variates and normalized column-wise. The signal x is also simulated from $\mathrm{N}(0, 1)$ variates, but with only 10% of entries being nonzero.

Here m_i are the number of trials and y_i the number of successes. The composite objective function for sparse logistic regression is then given by

$$\operatorname*{argmin}_x \sum_{i=1}^n \left\{ m_i \log\left(1 + e^{a_i^T x}\right) - y_i a_i^\top x \right\} + \lambda \sum_{j=1}^p |x_j|.$$

To specify a proximal gradient algorithm, all we need is an envelope such as those commonly used in Variational Bayes. In this example, we use the simple quadratic majorizer with Lipschitz constant given by $\|A^\top A\|_2/4 = \sigma_{\max}(A)/4$, and a penalty coefficient λ set to $0.1\sigma_{\max}(A)$.

Figure 2 shows the (adjusted) objective values per iteration with and without Nesterov acceleration. We can see the nondescent nature of the algorithm and the clear advantage of adding acceleration.

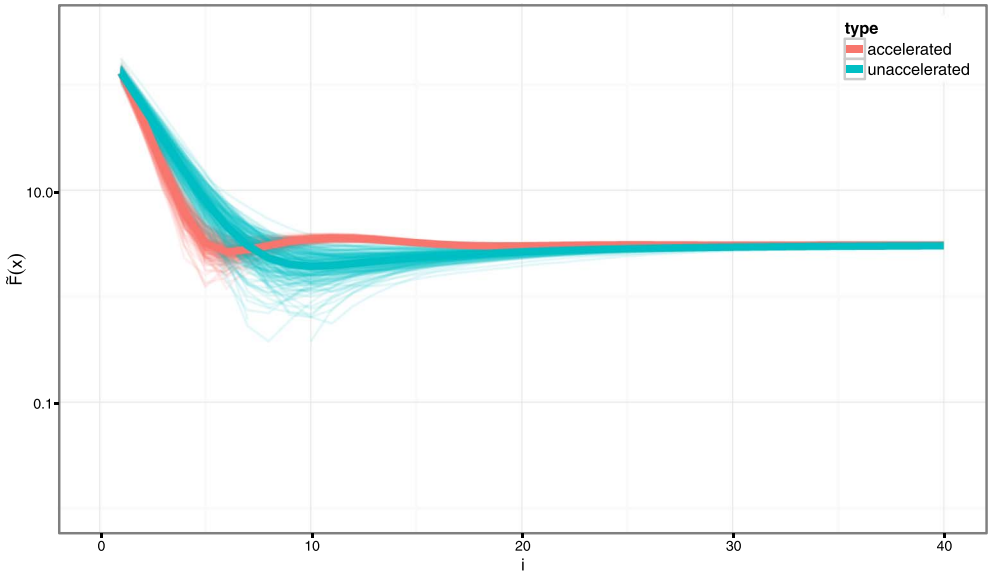

FIG. 2. *(Adjusted) objective values for iterations of the proximal gradient method, with and without acceleration, applied to a logistic regression problem with an ℓ^1-norm penalty.*

7.2 Logit Fused Lasso

To illustrate a logit fused lasso problem, we compare a Geman–Reynolds inspired quadratic envelope for the multinomial logit loss and a fused lasso penalty with the standard Lipschitz-bounded gradient step. We define the following quantities:

$$\begin{aligned}\Lambda(v) &= 2\sum_{i=1}^{n} m_i \lambda(a_i^\top v) a_i a_i^\top \\ &= 2A^\top \operatorname{diag}\bigl(m \cdot \lambda(Av)\bigr) A, \\ \eta^\top &= 2\sum_{i=1}^{n} (y_i - m_i/2) a_i^\top,\end{aligned}$$

where $\lambda(v) = \frac{1}{2v}(\frac{1}{1+e^{-v}} - \frac{1}{2})$. Now we compute x_t, conditional on w, for the envelope

$$\begin{aligned}&\sum_{i=1}^{n} \{m_i \log(1 + e^{a_i^T x}) - y_i a_i^\top x\} + \bigl\| D^{(1)} x \bigr\|_1 \\ &\quad = \min_w \left\{ \frac{1}{2} x^\top \Lambda(w) x - \eta^\top x + c(w) + \gamma \bigl\| D^{(1)} x \bigr\|_1 \right\}.\end{aligned}$$

To do this, we employ the Picard–Opial composite method of Argyriou et al. (2011).

Simulations were performed in a similar fashion as Section 7.2 but with $N = 100$, $M = 400$, $m = 2$ and where $D^{(1)}x$ has a fused lasso construction consisting of first-order differences of x. Figure 3 show the objective values for iterations of each formulation. With the use of second-order information, we have extremely fast convergence to the solution.

For data preconditioning, we perform the following decompositions: $A = U\Sigma V^\top$, the singular value decomposition (SVD), $\Lambda^{-1}(v) = \frac{1}{2}A^{-1}D^{-1}A^{-\top}$, where $D = \operatorname{diag}(\mathbf{m} \cdot \lambda(Av))$. This implies that one SVD of A, or generalized inverse, is required to compute all future $\Lambda^{-1}(v)$, thus providing computational savings.

7.3 Poisson Fused Lasso

To illustrate an objective that is not Lipschitz but still convex, we use a Poisson regression example with a fused lasso penalty. We simulated a signal given from the model

$$\begin{aligned}(y|x) &\sim \operatorname{Pois}\bigl(\exp(Ax)\bigr), \\ \phi(x) &= \bigl\| D^{(1)} x \bigr\|_1 = \sum_{j=1}^{p} |x_j - x_{j-1}|.\end{aligned}$$

In our simulation, the true sparse parameter vector x has 10% nonzero signals from $\mathrm{N}(0, 1)$. The design matrix $A \in \mathbb{R}^{100\times 300}$ is also generated from $\mathrm{N}(0, 1)$, then column normalized.

In sum, we have a negative log-likelihood and regularization penalty of the composite form

$$\begin{aligned}F(x) &= \sum_{i=1}^{n} \exp(a_i^\top x) - y_i a_i^\top x + \sum_{j=1}^{p} |x_j - x_{j-1}| \\ &= \sum_{i=1}^{n} \exp(a_i^\top x) - y_i a_i^\top x + \bigl\| D^{(1)} x \bigr\|_1,\end{aligned}$$

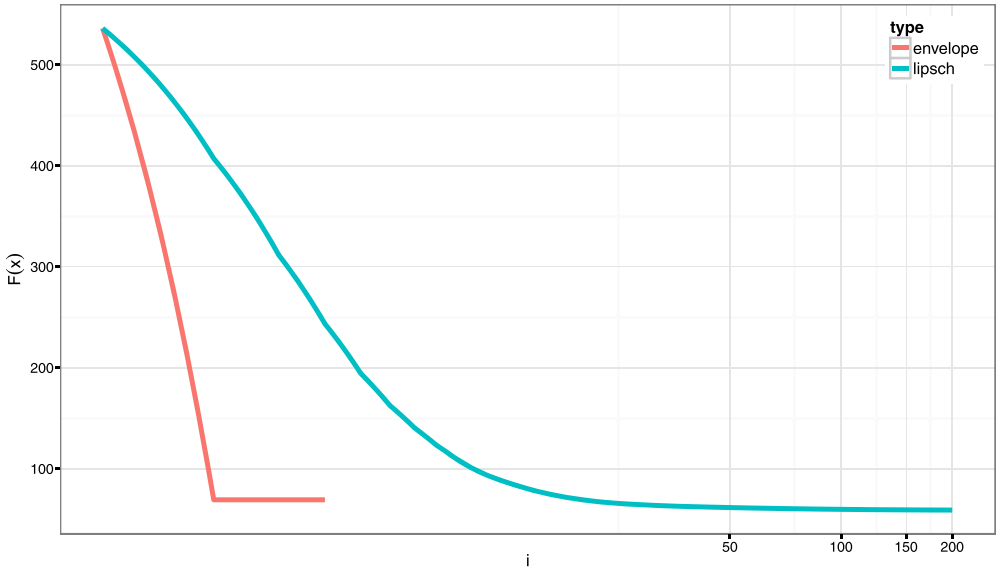

FIG. 3. *Objective values for iterations of two proximal composite formulations applied to a multinomial logistic regression problem with a composite ℓ^1-norm penalty. Both are run until the same numeric precision is reached.*

where a_i are the column vectors of A and $D^{(1)}x$ is the matrix operator of first-order differences in x. Since the Poisson loss function is not Lipschitz but still convex, we replace the constant gradient step with a back-tracking line search. This can be accomplished with a back-tracking line search step.

Figure 4 shows the objective value results for each method, with and without acceleration. An alternative approach is given by Green (1990), who describes an implementation of an EM algorithm for penalized likelihood estimation.

7.4 L^2-Norm Loss Plus L^q-Norm Penalty for $0 < q < 1$

A common nonconvex penalty is the L^q-norm for $0 < q < 1$. There are a number of ways of developing a proximal algorithm to solve such problems. The proximal operator of L^q-norm has a closed-form, multi-

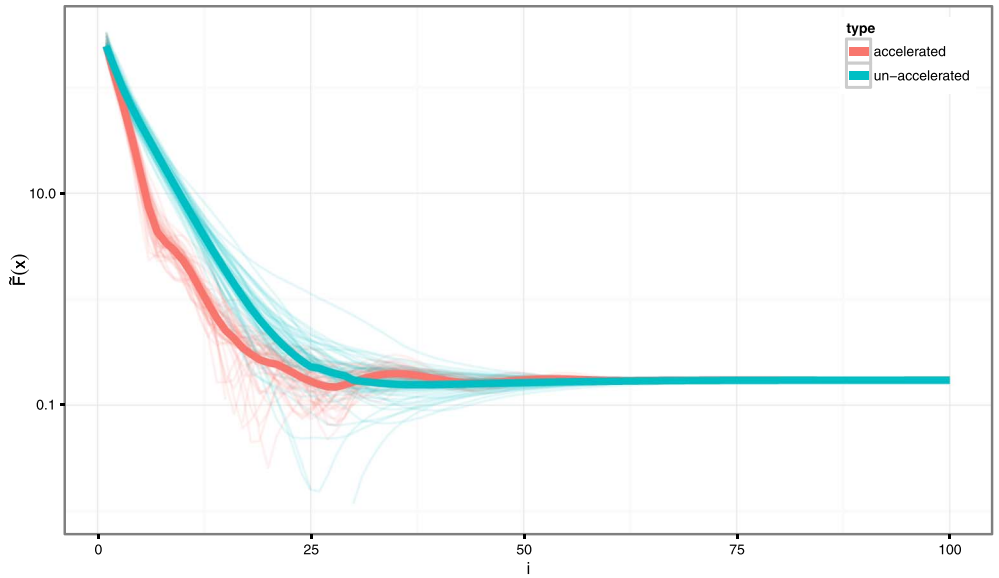

FIG. 4. *(Adjusted) objective values for iterations of the proximal gradient method, with and without acceleration, applied to a Poisson regression problem with a fused ℓ^1-norm penalty.*

valued solution, and convergence results are available for proximal methods in Marjanovic and Solo (2013) and Attouch, Bolte and Svaiter (2013). For this example, we choose the former approach.

The regularization problem involves finding the minimizer of an L^2-norm loss with an L^q-norm penalty for $0 < q < 1$, so that

$$\hat{x}_\lambda^q := \operatorname*{argmin}_x \left\{ \frac{1}{2}\|y - Ax\|^2 + \lambda \sum_{j=1}^{p} |x_i|^q \right\}.$$

The component-wise, set-valued proximal L^q-norm operator is given by

$$\operatorname*{prox}_{\lambda\phi_q}(v) = \begin{cases} 0, & \text{if } |v| < h_\lambda, \\ \{0, \operatorname{sgn}(v)x_\lambda\}, & \text{if } |v| = h_\lambda, \\ \operatorname{sgn}(v)\hat{x}, & \text{if } |v| > h_\lambda, \end{cases}$$

where

$$b_{\lambda,q} = \big(2\lambda(1-q)\big)^{1/(2-q)},$$

$$h_{\lambda,q} = b_{\lambda,q} + \lambda q b_{\lambda,q}^{q-1},$$

$$\hat{x} + \lambda q \hat{x}^{q-1} = |v|, \quad \hat{x} \in (b_{\lambda,q}, |x|).$$

Attouch, Bolte and Svaiter (2013) describe how the objective for this problem is a Kurdyka–Łojasiewicz (KL) function, which provides convergence results for an inexact (multi-valued proximal operator) forward–backward algorithm given by

$$x^{t+1} \in \operatorname*{prox}_{\lambda\gamma_t \|\cdot\|_p} \big(x^t - \gamma_t(A^\top A x^t - A^\top b)\big).$$

Interestingly, the KL convergence results for forward–backward splitting on appropriate nonconvex continuous functions bounded below imply that the solution choice for multi-valued proximal maps—as in the L^q-norm case—does not affect the convergence properties. See Appendix D for more information.

An alternative approach is the variational representation of the L^q-norm; however, this does not satisfy the convergence conditions of Allain, Idier and Goussard (2006) within the half-quadratic framework.

Marjanovic and Solo (2013) detail how cyclic descent can be used to apply the proximal operator in a per-coordinate fashion under a squared-error loss. The cyclic descent method is derived from the following algebra. First, a single solution to the squared-error loss minimization problem can be given for a component i of x, by

$$\begin{aligned} 0 &= \nabla_i l(x) = A_i^\top (Ax - y) \\ &= A_i^\top (A_i x_i + A_{-i} x_{-i} - y), \end{aligned}$$

where A_i is column i of A, and A_{-i}, x_{-i} have column/element i removed. Applied to a quadratic majorization scheme, we find that at iteration t

$$x_i^{t+1} = \frac{A_i^\top (y - A_{-i} x_i^{t+1})}{A_i^\top A_i} = \frac{A_i^\top r^t}{\|A_i\|^2} + x_i^t$$

with $y - Ax^t = r^t$. In a similar fashion to gradient descent, this involves $O(n)$ operations for updates of $A_i^\top r^t$, so one cycle is $O(np)$.

We simulate a data vector $y \in \mathbb{R}^n$ from a regression model

$$y = Ax + \sigma\varepsilon \quad \text{where } \varepsilon \sim \mathrm{N}(0, 1)$$

with an underlying sparse parameter value $x \in \mathbb{R}^d$ with $n = 100, d = 256$, in which the true sparse x has 5% nonzero signals generated from $\mathrm{N}(0, 1)$. The design matrix $A \in \mathbb{R}^{100\times 256}$ is also generated from $\mathrm{N}(0, 1)$, then column normalized. We set the signal-to-noise ratio at 16.5 to match the simulated example from Marjanovic and Solo (2013), which gives $\sigma = 0.0369$.

Figure 5 plots the mean squared error (MSE) versus the log-regularization penalty and the power in the L^q-norm penalty. Essentially, this consists of contours of $\log_{10}(\mathrm{MSE}(\hat{x}))$ on a plot of $0 < q < 1$ versus the amount of regularization $\log_{10}(\lambda)$. One interesting feature of this model is that the estimated regression coefficients $\hat{x}_\lambda^q$ can jump to sparsity as $0 < q < 1$, and this will be illustrated in a regularized path for the next example.

7.5 Prostate Data

As a practical example of our methodology, we consider the prostate cancer data set, which examines the relationship between the level of a prostate specific antigen and a number of clinical factors. The variables are log cancer volume (`lcavol`), log prostate weight (`lweight`), age (`age`), log of the amount of benign prostatic hyperplasia (`lbph`), seminal vesicle invasion (`svi`), log of capsular penetration (`lcp`), Gleason score (`gleason`) and percent of Gleason scores 4 or 5 (`pgg45`).

A common regularized approach is to use lasso and elastic net; see Tibshirani (1996) and in Zou and Hastie (2005), respectively. Alternatively, we fit the regularization path using

$$\hat{x}_\lambda^q := \operatorname*{argmin}_x \left\{ \frac{1}{2}\|y - Ax\|^2 + \lambda \sum_{j=1}^{p} |x_i|^q \right\}.$$

We can use the exact proximal operator for the L^q-norm and solve the harder nonconvex problem. Figure 6 shows the regularization path. The major difference is, again, in the jumps to a sparse solution.

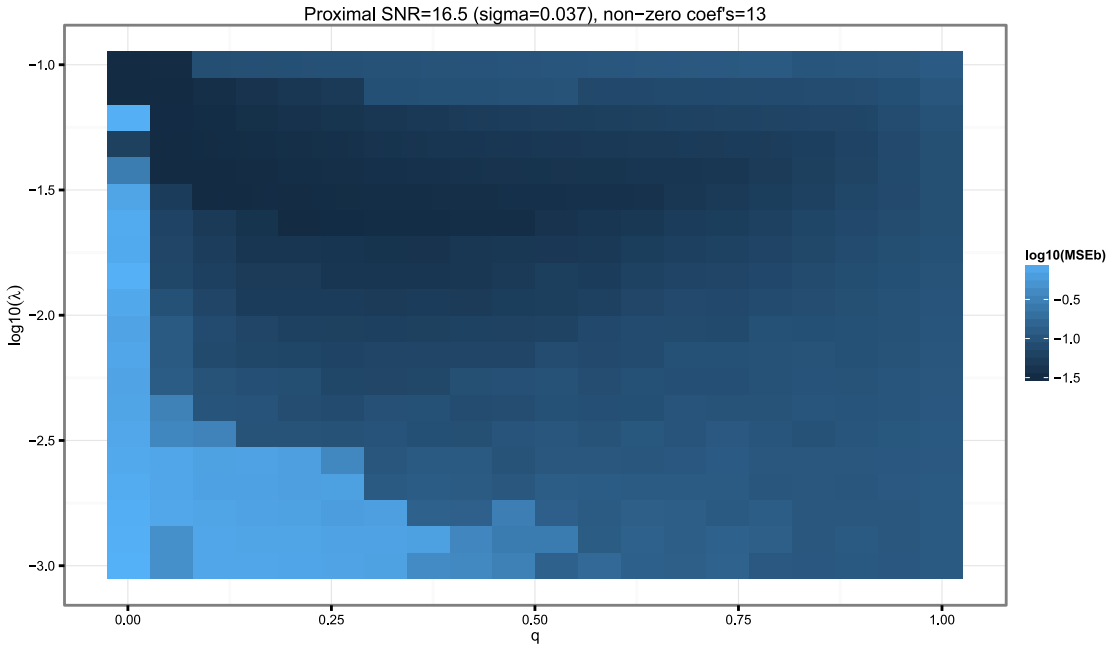

FIG. 5. *Penalty weight, λ, vs. MSE and q for an L^2-norm error with an L^q-norm penalty, $0 < q < 1$, estimated via cyclic descent and proximal solutions.*

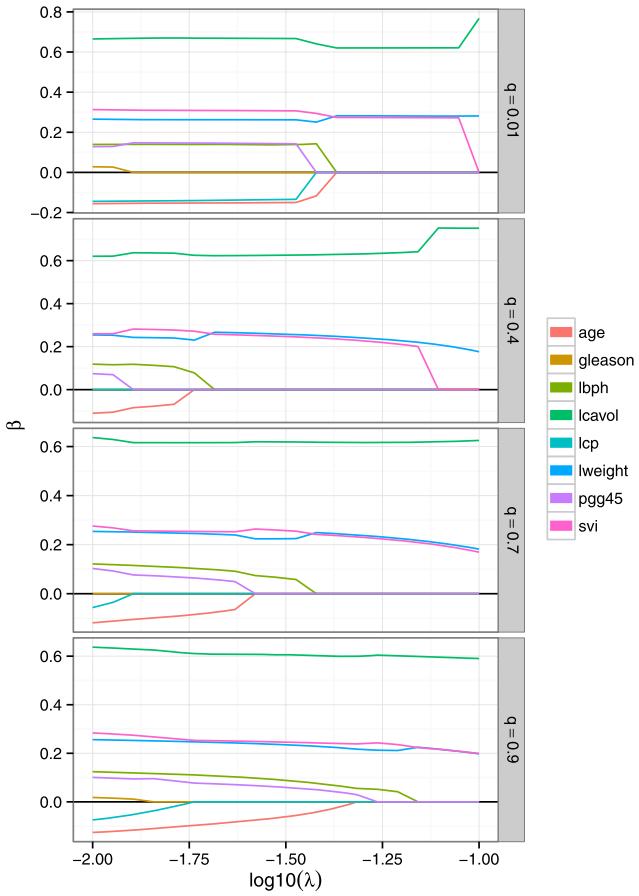

FIG. 6. *Proximal results for the prostate data example under the L^q-norm penalty.*

8. DISCUSSION

Proximal algorithms are a widely used approach for solving optimization problems. They provide an elegant extension of classical gradient descent method and have properties that—much like EM or MM algorithms—can be used to derive many different approaches for solving a given problem.

For readers interest in further historical details, we recommend Beck and Sabach (2015), who provide a historical perspective on iterative shrinkage algorithms by focusing mainly on the Weiszfeld algorithm (Weiszfeld, 1937) for computing an ℓ^1 median. The split Lagrangian methods described here were originally developed by Hestenes (1969) and Rockafellar (1974). More recently, there is work being done to extend the range of applicability of these methods outside of the class of convex functions to the broader class of functions satisfying the Kurdyka–Łojasiewicz inequality (Attouch, Bolte and Svaiter, 2013).

The purpose of our review has been to describe and apply proximal algorithms to some archetypical optimization problems that arise in statistics. These problems often involve composite functions that are representable by a sum of a linear or quadratic envelope, together with a function that has a closed-form proximal operator that is easy to evaluate. Many papers demonstrate the efficacy and breadth of application of this approach: for example, Micchelli et al.

(2013) and Micchelli, Shen and Xu (2011) study proximal operators for composite operators for L^2-norm and ℓ^1-norm/TV denoising models; Argyriou et al. (2011) describe numerical advantages of the proximal operator approach versus traditional fused lasso implementations; and Chen, Huang and Zhang (2013) provide a further class of fixed-point algorithms that advance the proximal approach in the composite setting.

Another nice property of proximal algorithms is the ease with which acceleration techniques can be applied. The most common approach involves Nesterov acceleration; see Nesterov (1983) and Beck and Teboulle (2004), who introduce a momentum term for gradient-descent algorithms applied to nonsmooth composite problems. Attouch and Bolte (2009), Noll (2014) provide further convergence properties for nonsmooth functions. O'Donoghue and Candes (2015) use adaptive restart to improve the convergence rate of accelerated gradient schemes. Meng and Chen (2011) modify Nesterov's gradient method for strongly convex functions with Lipschitz continuous gradients. Allen-Zhu and Orecchia (2014) provide a simple interpretation of Nesterov's scheme as a two-step algorithm with gradient-descent steps which yield proximal (forward) progress coupled with mirror-descent (backward) steps with dual (backward) progress. By linearly coupling these two steps they improve convergence. Giselsson and Boyd (2014) also show how preconditioning can help with convergence for ill-conditioned problems.

There are a number of directions for future research on proximal methods in statistics, for example, exploring the use of Divide and Concur methods for exponential-family mixed models and studying the relationship between proximal splitting and variational Bayes methods in graphical models. Another interesting area of research involves combining proximal steps with MCMC algorithms (Pereyra, 2013). Of course, the proximal methods developed here are not designed to provide standard errors and the advantage of MCMC methods is the ability to assess uncertainty through the full posterior distribution.

APPENDIX A: PROXIMAL GRADIENT CONVERGENCE

We now outline convergence results for the proximal gradient solution, given by (4), to the fixed point problem

$$x^\star = \operatorname*{prox}_{\phi/\lambda}\bigl(x - \nabla l(x)/\lambda\bigr),$$

when l and ϕ are convex, lower semi-continuous and ∇l is Lipschitz continuous. We also assume that $\operatorname{prox}_{\phi/\lambda}$ is nonempty and can be evaluated independently in each component.

Recalling the translation property of proximal operators stated in (12), we can say

$$\begin{aligned} x^\star &= \operatorname*{prox}_{\phi/\lambda}\bigl(x - \nabla l(x)/\lambda\bigr) = \operatorname*{prox}_{(\phi(z)+\lambda\nabla l(x)^\top z)/\lambda}(x) \\ &= \operatorname*{argmin}_z\left\{\phi(z) + \nabla l(x)^\top(z-x) + \frac{\lambda}{2}\|x-z\|^2\right\}. \end{aligned}$$

By the proximal operator's minimizing properties, its solution $x^\star$ satisfies

$$\phi\bigl(x^\star\bigr) + \nabla l(x)^\top\bigl(x^\star - x\bigr) + \frac{\lambda}{2}\bigl\|x - x^\star\bigr\|^2 \le \phi(x),$$

providing a quadratic minorizer for $F(w)$ in the form of

$$\begin{aligned} &l(w) + \phi\bigl(x^\star\bigr) + \nabla l(w)^\top\bigl(x^\star - w\bigr) + \frac{\lambda}{2}\bigl\|w - x^\star\bigr\|^2 \\ &\quad\le l(w) + \phi(w) \equiv F(w). \end{aligned}$$

The Lipschitz continuity of $\nabla l(x)$, that is,

$$l(x) \le l(w) + \nabla l(w)^\top(x-w) + \frac{\gamma}{2}\|x-w\|^2,$$

also gives us a quadratic majorizer

$$\begin{aligned} F(x) &\equiv l(x) + \phi(x) \\ &\le l(w) + \phi(x) + \nabla l(w)^\top(x-w) \\ &\quad + \frac{\gamma}{2}\|x-w\|^2, \end{aligned}$$

which, when evaluated at $x = x^\star$ and combined with our minorizer, yields

$$(\lambda - \gamma)\tfrac{1}{2}\bigl\|x^\star - w\bigr\|^2 \le F(w) - F\bigl(x^\star\bigr).$$

Thus, if we want to ensure that the objective value will decrease in this procedure, we need to fix $\lambda \ge \gamma$. Furthermore, functional characteristics of l and ϕ, such as strong convexity, can improve the bounds in the steps above and guarantee good- or optimal-decreases in $F(w) - F(x^\star)$.

Finally, when we compound up the errors we obtain a $O(1/k)$ convergence bound. This can be improved by adding a momentum term that includes the first derivative information.

These arguments can be extended to Bregman divergences by way of the general law of cosines inequality:

$$\begin{aligned} D_\phi(x,z) &= D_\phi(x,w) + D_\phi(w,z) \\ &\quad - \bigl(\nabla\phi(z) - \nabla\phi(w)\bigr)^\top(x-w), \end{aligned}$$

so that $D_\phi(x,z) \ge D_\phi(x,w) + D_\phi(w,z)$ where $w = \operatorname{argmin}_v D_\phi(v,z)$.

APPENDIX B: NESTEROV ACCELERATION

A powerful addition is Nesterov acceleration. Consider a convex combination, with parameter θ, of upper bounds for the proximal operator inequality $z = x$ and $z = x^\star$. We are free to choose variables $z = \theta x + (1-\theta)x^+$ and w. If ϕ is convex, $\phi(\theta x + (1-\theta)x^+) \le \theta\phi(x) + (1-\theta)\phi(x^+)$, then we have

$$\begin{aligned}
&F(x^+) - F^\star - (1-\theta)\big(F(x) - F^\star\big)\\
&\quad = F(x^+) - \theta F^\star - (1-\theta)F(x)\\
&\quad \le \lambda(x^+ - w)^\top(\theta x^\star + (1-\theta)x - x^+)\\
&\qquad + \frac{\lambda}{2}\|x^+ - w\|^2\\
&\quad = \frac{\lambda}{2}\big(\|w - (1-\theta)x - \theta x^\star\|^2\\
&\qquad - \|x^+ - (1-\theta)x - \theta x^\star\|^2\big)\\
&\quad = \frac{\theta^2\lambda}{2}\big(\|u - x^\star\|^2 - \|u^+ - x^\star\|^2\big),
\end{aligned}$$

where w is given in terms of the intermediate steps

$$\begin{aligned}
\theta u &= w - (1-\theta)x,\\
\theta u^+ &= x^+ - (1-\theta)x,
\end{aligned}$$

introducing a sequence θ_t with iteration subscript, t. The second identity, $\theta u = x - (1-\theta)x^-$, then yields an update for w as the current state x plus a momentum term, depending on the direction $(x - x^-)$, namely,

$$w = (1-\theta_t)x + \theta_t u = x - \theta_{t-1}(1-\theta_t)\big(x - x^-\big).$$

APPENDIX C: QUASI-CONVEX CONVERGENCE

Consider an optimization problem $\min_{x\in\mathcal{X}} l(x)$ where l is quasi-convex, continuous and has a nonempty set of finite global minima. Let x^t be generated by the proximal point algorithm

$$x^t \in \operatorname{argmin}\left\{l(x) + \frac{\lambda_t}{2}\|x - x^t\|^2\right\}.$$

Papa Quiroz and Oliveira (2009) show that these iterates converge to the global minima, although the proximal operator at each step may be set-valued, due to the nonconvexity of l. A function l is quasi-convex when

$$l\big(\theta x + (1-\theta)z\big) \le \max\big(l(x), l(z)\big),$$

which accounts for a number of nonconvex functions like $|x|^q$, when $0 < q < 1$, and functions involving appropriate ranges of $\log(x)$ and $\tanh(x)$. In this setting, using the level-sets generated by the sequence, that is, $U = \{x \in \operatorname{dom}(l) : l(x) \le \inf_t l(x^t)\}$, one finds that U is a nonempty closed convex set and that x^t is a Fejér sequence of finite length, $\sum_t \|x^{t+1} - x^t\| < \infty$, and that it converges to a critical point of l as long as $\min\{l(x) : x \in \mathbb{R}^d\}$ is nonempty.

APPENDIX D: NONCONVEX: KURDYKA–ŁOJASIEWICZ (KL)

A locally Lipschitz function $l : \mathbb{R}^d \to \mathbb{R}$ satisfies KL at $x^\star \in \mathbb{R}^d$ if and only if $\exists\eta \in (0,\infty)$ and a neighborhood U of $x^\star$ and a concave $\kappa : [0,\eta] \to [0,\infty)$ with $\kappa(0) = 0$, $\kappa \in C^1$, $\kappa' > 0$ on $(0,\eta)$ and for every $x \in U$ with $l(x^\star) < l(x) < l(x^\star) + \eta$ we have

$$\kappa'\big\{l(x) - l\big(x^\star\big)\big\}\operatorname{dist}\big(0, \partial l(x)\big) \ge 1,$$

where $\operatorname{dist}(0, A) := \sup_{x\in A}\|x\|^2$.

The KL condition guarantees summability and therefore a finite length of the discrete subgradient trajectory. Using the KL properties of a function, one can show convergence for alternating minimization algorithms for problems like

$$\min_{x,z} L(x,z) := l(x) + Q(x,z) + \phi(z),$$

where ∇Q is Lipschitz continuous (see Attouch et al., 2010, Attouch, Bolte and Svaiter, 2013). A typical application involves solving $\min_{x\in\mathbb{R}^d}\{l(x) + \phi(x)\}$ via the augmented Lagrangian

$$L(x,z) = l(x) + \phi(z) + \lambda^\top(x - z) + \frac{\rho}{2}\|x - z\|^2,$$

where ρ is a relaxation parameter.

A useful class of functions that satisfy KL is one that possesses uniform convexity

$$l(z) \ge l(x) + u^\top(z - x) + K\|z - x\|^p,$$

where

$$p \ge 1 \quad \forall u \in \partial l(x).$$

Then l satisfies KL on $\operatorname{dom}(l)$ for $\kappa(s) = pK^{-1/p}s^{1/p}$.

For explicit convergence rates in the KL setting, see Frankel, Garrigos and Peypouquet (2015).

ACKNOWLEDGMENTS

We thank the participants at the 2014 ASA meetings for their comments. We also thank the Editor, Associate Editor and two anonymous referees for their help in improving the paper.

REFERENCES

ALLAIN, M., IDIER, J. and GOUSSARD, Y. (2006). On global and local convergence of half-quadratic algorithms. *IEEE Trans. Image Process.* **15** 1130–1142.

ALLEN-ZHU, Z. and ORECCHIA, L. (2014). A novel, simple interpretation of Nesterov's accelerated method as a combination of gradient and mirror descent. Preprint. Available at arXiv:1407.1537.

ARGYRIOU, A., MICCHELLI, C. A., PONTIL, M., SHEN, L. and XU, Y. (2011). Efficient first order methods for linear composite regularizers. Preprint. Available at arXiv:1104.1436.

ATTOUCH, H. and BOLTE, J. (2009). On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.* **116** 5–16. MR2421270

ATTOUCH, H., BOLTE, J. and SVAITER, B. F. (2013). Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss–Seidel methods. *Math. Program.* **137** 91–129. MR3010421

ATTOUCH, H., BOLTE, J., REDONT, P. and SOUBEYRAN, A. (2010). Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka–Łojasiewicz inequality. *Math. Oper. Res.* **35** 438–457. MR2674728

BECK, A. and SABACH, S. (2015). Weiszfeld's method: Old and new results. *J. Optim. Theory Appl.* **164** 1–40. MR3296283

BECK, A. and TEBOULLE, M. (2004). A conditional gradient method with linear rate of convergence for solving convex linear systems. *Math. Methods Oper. Res.* **59** 235–247. MR2063242

BECK, A. and TEBOULLE, M. (2010). Gradient-based algorithms with applications to signal recovery problems. In *Convex Optimization in Signal Processing and Communications* (D. P. Palomar and Y. C. Eldar, eds.) 42–88. Cambridge Univ. Press, Cambridge. MR2840594

BECK, A. and TEBOULLE, M. (2014). A fast dual proximal gradient algorithm for convex minimization and applications. *Oper. Res. Lett.* **42** 1–6. MR3159144

BERTSEKAS, D. P. (2011). Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning* **2010** 1–38.

BESAG, J. (1986). On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B* **48** 259–302. MR0876840

BIEN, J., TAYLOR, J. and TIBSHIRANI, R. (2013). A LASSO for hierarchical interactions. *Ann. Statist.* **41** 1111–1141. MR3113805

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. now Publishers, Hanover, MA.

BRÈGMAN, L. M. (1967). A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **7** 200–217.

CEVHER, V., BECKER, S. and SCHMIDT, M. (2014). Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics. *IEEE Signal Process. Mag.* **31** 32–43.

CHAMBOLLE, A. and POCK, T. (2011). A first-order primal–dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision* **40** 120–145. MR2782122

CHAUX, C., COMBETTES, P. L., PESQUET, J.-C. and WAJS, V. R. (2007). A variational formulation for frame-based inverse problems. *Inverse Probl.* **23** 1495–1518. MR2348078

CHEN, P., HUANG, J. and ZHANG, X. (2013). A primal–dual fixed point algorithm for convex separable minimization with applications to image restoration. *Inverse Probl.* **29** 025011, 33. MR3020432

CHEN, G. and TEBOULLE, M. (1994). A proximal-based decomposition method for convex minimization problems. *Math. Program.* **64** 81–101. MR1274173

CHOUZENOUX, E., PESQUET, J.-C. and REPETTI, A. (2014). Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. *J. Optim. Theory Appl.* **162** 107–132. MR3228518

CHRÉTIEN, S. and HERO, A. O. III (2000). Kullback proximal algorithms for maximum-likelihood estimation. *IEEE Trans. Inform. Theory* **46** 1800–1810. MR1790321

COMBETTES, P. L. and PESQUET, J.-C. (2011). Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering* 185–212. Springer, New York. MR2858838

CSISZÁR, I. and TUSNÁDY, G. (1984). Information geometry and alternating minimization procedures. *Statist. Decisions* **1 (supplement issue)** 205–237. MR0785210

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DUCKWORTH, D. (2014). The big table of convergence rates. Available at https://github.com/duckworthd/duckworthd.github.com/blob/master/blog/big-table-of-convergence-rates.html.

ESSER, E., ZHANG, X. and CHAN, T. F. (2010). A general framework for a class of first order primal–dual algorithms for convex optimization in imaging science. *SIAM J. Imaging Sci.* **3** 1015–1046. MR2763706

FIGUEIREDO, M. A. T. and NOWAK, R. D. (2003). An EM algorithm for wavelet-based image restoration. *IEEE Trans. Image Process.* **12** 906–916. MR2008658

FRANKEL, P., GARRIGOS, G. and PEYPOUQUET, J. (2015). Splitting methods with variable metric for Kurdyka–Lojasiewicz functions and general convergence rates. *J. Optim. Theory Appl.* **165** 874–900.

GEMAN, D. and REYNOLDS, G. (1992). Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.* **14** 367–383.

GEMAN, D. and YANG, C. (1995). Nonlinear image recovery with half-quadratic regularization. *IEEE Trans. Image Process.* **4** 932–946.

GISELSSON, P. and BOYD, S. (2014). Preconditioning in fast dual gradient methods. In *Proceedings of the* 53*rd Conference on Decision and Control*. 5040–5045. Los Angeles, CA.

GRAVEL, S. and ELSER, V. (2008). Divide and concur: A general approach to constraint satisfaction. *Phys. Rev. E* **78** 036706.

GREEN, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **52** 443–452. MR1086796

GREEN, P. J., ŁATUSZYŃSKI, K., PEREYRA, M. and ROBERT, C. P. (2015). Bayesian computation: A perspective on the current state, and sampling backwards and forwards. Preprint. Available at arXiv:1502.01148.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

HESTENES, M. R. (1969). Multiplier and gradient methods. *J. Optim. Theory Appl.* **4** 303–320. MR0271809

HU, Y. H., LI, C. and YANG, X. Q. (2015). Proximal gradient algorithm for group sparse optimization.

KOMODAKIS, N. and PESQUET, J.-C. (2014). Playing with duality: An overview of recent primal–dual approaches for solving large-scale optimization problems. Preprint. Available at arXiv:1406.5429.

MAGNÚSSON, S., WEERADDANA, P. C., RABBAT, M. G. and FISCHIONE, C. (2014). On the convergence of alternating direction lagrangian methods for nonconvex structured optimization problems. Preprint. Available at arXiv:1409.8033.

MARJANOVIC, G. and SOLO, V. (2013). On exact ℓ^q denoising. In *Acoustics*, *Speech and Signal Processing* (*ICASSP*), 2013 *IEEE International Conference on* 6068–6072. IEEE, New York.

MARTINET, B. (1970). Brève communication. Regularisation d'inéquations variationnelles par approximations successives. *ESAIM Math. Modell. Numer. Anal.* **4** 154–158.

MENG, X. and CHEN, H. (2011). Accelerating Nesterov's method for strongly convex functions with Lipschitz gradient. Preprint. Available at arXiv:1109.6058.

MICCHELLI, C. A., SHEN, L. and XU, Y. (2011). Proximity algorithms for image models: Denoising. *Inverse Probl.* **27** 045009, 30. MR2781033

MICCHELLI, C. A., SHEN, L., XU, Y. and ZENG, X. (2013). Proximity algorithms for the L1/TV image denoising model. *Adv. Comput. Math.* **38** 401–426. MR3019155

NESTEROV, YU. E. (1983). A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Sov. Math., Dokl.* **27** 372–376.

NIKOLOVA, M. and NG, M. K. (2005). Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM J. Sci. Comput.* **27** 937–966 (electronic). MR2199915

NOLL, D. (2014). Convergence of non-smooth descent methods using the Kurdyka–Łojasiewicz inequality. *J. Optim. Theory Appl.* **160** 553–572. MR3180983

O'DONOGHUE, B. and CANDES, E. (2015). Adaptive restart for accelerated gradient schemes. *Found. Comput. Math.* **15** 715–732.

PALMER, J., KREUTZ-DELGADO, K., RAO, B. D. and WIPF, D. P. (2005). Variational EM algorithms for non-Gaussian latent variable models. In *Advances in Neural Information Processing Systems* 18 1059–1066. Vancouver, BC, Canada.

PAPA QUIROZ, E. A. and OLIVEIRA, P. R. (2009). Proximal point methods for quasiconvex and convex functions with Bregman distances on Hadamard manifolds. *J. Convex Anal.* **16** 49–69. MR2531192

PARIKH, N. and BOYD, S. (2013). Proximal algorithms. *Foundations and Trends in Optimization* **1** 123–231.

PATRINOS, P. and BEMPORAD, A. (2013). Proximal Newton methods for convex composite optimization. In *Decision and Control* (*CDC*), 2013 *IEEE* 52*nd Annual Conference on* 2358–2363. IEEE, New York.

PATRINOS, P., LORENZO, S. and ALBERTO, B. (2014). Douglas-rachford splitting: Complexity estimates and accelerated variants. Preprint. Available at arXiv:1407.6723.

PEREYRA, M. (2013). Proximal Markov chain Monte Carlo algorithms. Preprint. Available at arXiv:1306.0187.

POLSON, N. G. and SCOTT, J. G. (2012). Local shrinkage rules, Lévy processes and regularized regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 287–311. MR2899864

POLSON, N. G. and SCOTT, J. G. (2015). Mixtures, envelopes, and hierarchical duality. *J. Roy. Statist. Soc. Ser. B*. To appear. Available at arXiv:1406.0177.

ROCKAFELLAR, R. T. (1974). Conjugate duality and optimization. Technical report, DTIC Document, 1973.

ROCKAFELLAR, R. T. (1976). Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.* **14** 877–898. MR0410483

ROCKAFELLAR, R. T. and WETS, R. J.-B. (1998). *Variational Analysis*. Springer, Berlin. MR1491362

RUDIN, L., OSHER, S. and FATERNI, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D* **60** 259–268.

SHOR, N. Z. (1985). *Minimization Methods for Nondifferentiable Functions*. Springer, Berlin. MR0775136

TANSEY, W., KOYEJO, O., POLDRACK, R. A. and SCOTT, J. G. (2014). False discovery rate smoothing. Technical report, Univ. Texas at Austin.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** 285–323. MR3189487

TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641

VON NEUMANN, J. (1951). *Functional Operators*: *The Geometry of Orthogonal Spaces*. Princeton Univ. Press, Princeton, NJ.

WEISZFELD, E. (1937). Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Math. J.* **43** 355–386.

WITTEN, D. M., TOBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

ZHANG, X., SAHA, A. and VISHWANATHAN, S. V. N. (2010). Regularized risk minimization by Nesterov's accelerated gradient methods: Algorithmic extensions and empirical studies. Preprint. Available at arXiv:1011.0472.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic Net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327