Proximal Algorithms in Statistics and Machine Learning

Nicholas G. Polson
Booth School of Business
University of Chicago

James G. Scott
McCombs School of Business
University of Texas at Austin

Brandon T. Willard
Booth School of Business
University of Chicago

First Draft: October 2014
This Draft: March 2015

Abstract

In this paper we develop proximal methods for statistical learning. Proximal point algorithms are useful in statistics and machine learning for obtaining optimization solutions for composite functions. Our approach exploits closed-form solutions of proximal operators and envelope representations based on the Moreau, Forward-Backward, Douglas-Rachford and Half-Quadratic envelopes. Envelope representations lead to novel proximal algorithms for statistical optimisation of composite objective functions which include both non-smooth and non-convex objectives. We illustrate our methodology with regularized Logistic and Poisson regression and non-convex bridge penalties with a fused lasso norm. We provide a discussion of convergence of non-descent algorithms with acceleration and for non-convex functions. Finally, we provide directions for future research.

Keywords: Bayes MAP; shrinkage; sparsity; splitting; fused lasso; Kurdyka-
Lojasiewicz; non-convex optimisation; proximal operators; envelopes; regularization; ADMM; optimization; Divide and Concur.

*Polson is Professor of Econometrics and Statistics at the Chicago Booth School of Business. email: ngp@chicagobooth.edu. Scott is Associate Professor of Statistics at the McCombs School of Business, University of Texas at Austin. email: James.Scott@mccombs.utexas.edu. Brandon T. Willard. email: bwillard@uchicago.edu. We thank the participants at the 2014 ASA meetings for their comments.
1 Introduction

Our goal is to introduce statisticians to the large body of literature on \textit{proximal algorithms} for solving statistical regularization problems. By a proximal algorithm, we mean an algorithm whose steps involve evaluating the \textit{proximal operator} of some term in the objective function. Both of these concepts will be defined precisely in the next section. The canonical optimization problem of minimising a measure of fit, together with a regularization penalty, sits at the heart of modern statistical practice and it arises, for example, in sparse regression \cite{Tibshirani, 1996}, spatial smoothing \cite{Tibshirani et al., 2005}, covariance estimation \cite{Witten et al., 2009}, image processing \cite{Geman and Reynolds, 1992, Geman and Yang, 1995, Rudin et al., 1992}, nonlinear curve fitting \cite{Tibshirani, 2014}, Bayesian MAP inference \cite{Polson and Scott, 2012}, multiple hypothesis testing \cite{Tansey et al., 2014} and shrinkage/sparsity-inducing prior regularisation problems \cite{Green et al., 2015}.

The techniques we employ here are often referred to as Proximal Gradient, Proximal Point, Alternating Direction Method of Multipliers (ADMM) \cite{Boyd et al., 2011}, Divide and Concur (DC), Frank-Wolfe (FW), Douglas-Ratchford (DR) splitting or alternating split Bregman (ASB) methods. The field of image processing has developed many of these ideas in the form of Total Variation (TV) de-noising and half-quadratic (HQ) optimization \cite{Geman and Yang, 1995, Geman and Reynolds, 1992, Nikolova and Ng, 2005}. Other methods such as fast iterative shrinkage thresholding algorithm (FISTA), expectation maximization (EM), majorisation-minimisation (MM) and iteratively reweighed least squares (IRLS) fall into our proximal framework. Although such approaches are commonplace in statistics and machine learning \cite{Bien et al., 2013}, there hasn’t been a real focus on the general family of approaches that underly these algorithms. Early work on iterative proximal fixed point algorithms in Banach spaces is due to \cite{Von Neumann, 1951, Bregman, 1967, Hestenes, 1969, Martinet, 1970, Rockafellar, 1976}.

A useful feature of proximal algorithms are acceleration techniques \cite{Nesterov, 1983} which lead to non-descent algorithms that can provide an order-of-magnitude increase in efficiency. When both functions are convex, and one has a smooth Lipschitz continuous gradient, a simple convergence result based on the reverse Pythagoras inequality is available. Convergence rates of the associated gradient descent algorithms can vary and typically each analysis has to be dealt with on a case-by-case basis. We illustrate acceleration for a sparse logistic regression with a fused lasso penalty.

The rest of the paper proceeds as follows. Section 1.1 provides notation and basic properties of proximal operators and envelopes. Section 2 describes the proximal operator and Moreau envelope. Section 3 describes the proximal gradient algorithm and its extensions. Section 4 discusses general envelopes and how proximal algorithms can be viewed as envelope gradients. Section 5 considers the default problem of composite operator optimisation. We show how to compute the exact proximal operator with a general quadratic envelope and a composite regularisation penalty. Section 6 illustrates our methodology with applications to logistic and Poisson regression with fused lasso penalties. A bridge regression penalty illustrates the non-convex case and we apply our algorithm to the prostate data of Hastie et al. \cite{Hastie et al., 2009}.
Table 1 provides commonly used proximal operators, Table 2 documents examples of half-quadratic envelopes and Table 3 lists convergence rates for a variety of algorithms. Appendix A discusses convergence results for both convex and non-convex cases together with Nesterov acceleration. Finally, Section 7 concludes with directions for future research.

1.1 Preliminaries

Many statistical regularisation settings are faced with the task of solving the following optimization problem

$$\arg\min_{x \in X} F(x) = l(x) + \phi(x)$$

where $l(x)$ is a measure of fit depending implicitly on some observed data $y$, $\phi(x)$ is a regularization term that imposes structure or effects a favorable bias-variance trade-off. Typically, $l(x)$ is a smooth function and $\phi(x)$ is non-smooth–like a lasso or bridge penalty–so as to induce sparsity. 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) $a_i^T$, and $B$ a fixed $d \times k$ matrix to encode some structural penalty on the parameter (as in the group lasso or fused lasso), $b$ are prior loadings and centerings and $\gamma > 0$ is a regularisation parameter that will trace out a solution path. All together, we have a composite objective of the form

$$F(x) = \sum_{i=1}^{n} l(y_i, a_i^T x) + \gamma \sum_{j=1}^{d} \phi([Bx - b]_j)$$

A simple statistical model that can take this form is in its negative log likelihood is $y = Ax + \epsilon$, where $\epsilon$ is a standard normal measurement error corresponding to the norm $l(x) = \|Ax - y\|^2$ and each $x_j$ has independent Laplace priors corresponding to $\phi(x) = \gamma \sum_{j=1}^{d} |x_j|$. Observations are indexed by $i$, parameters by $j$, and iterations of an algorithm by $t$. Unless stated otherwise, all functions are lower semi-continuous, and all vectors are column vectors. We will pay particular attention to the composite penalty $\phi(x) = \gamma \phi(Bx)$ where $B$ corresponds to some constraint space, such as the discrete difference operator in fused Lasso.

Splitting is a key tool that exploits an equivalence between the unconstrained optimisation problem and a constrained one that includes a latent–or slack–variable, $z$, where we write

$$\min_x \{l(x) + \phi(Bx)\} \equiv \min_{x, z} \{l(x) + \phi(z)\} \text{ subject to } z = Bx$$

To solve the latter problem, we can use augmented Lagrangian methods (ALM) (a.k.a. the alternating direction method of multipliers, ADMM).

Envelopes are another way of introducing latent variables. For example, we will assume that the objective $l(x)$ takes the form of an envelope;

1. a linear envelope $l(x) = \sup_y \{xy - l^*(y)\}$ where $l^*$ denotes the convex dual.
2. a quadratic envelope $l(x) = \inf_y \left\{ \frac{1}{2} x^T \Lambda(y)x - \eta^T(y)x + \psi(y) \right\}$ for some $\Lambda, \xi, \psi$.

The convex conjugate of $l(x)$, $l^*(z)$, 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 the following dual relationship holds between $l$ and its conjugate:

$$l(x) = \sup_{\lambda} \{ \lambda^T x - l^*(\lambda) \} \text{ where } l^*(\lambda) = \sup_x \{ \lambda^T x - l(x) \}.$$  

If $l(x)$ is differentiable, the maximizing value of $\lambda$ is $\hat{\lambda}(x) = \nabla l(x)$.

In sum, latent variables allow us to view the problem of $\min_x F(x)$ as one of a joint minimisation of $\min_x F(x, z)$ where the augmented $F(\cdot, c)$ can be easily minimisation in a conditional fashion. Such alternating minimisation or iterated conditional mode (ICM) [Besag, 1986], [Csisz et al., 1984] algorithms have a long history in statistics. The additional insight is that proximal operators allow the researcher to perform the alternating minimisation step for the non-smooth penalty, $\phi$, in an elegant closed-form fashion. Moreover, divide and concur (DC) methods allow difficult high dimensional problems to be broken down into a collection of smaller tractable subproblems with the global solution being retrieved from the solutions to the subproblems.

The following definitions will be useful. 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 $f(x)$. A $\rho$-strong convex function satisfies

$$f(x) \geq f(y) + u^T(x - y) + \frac{\rho}{2} \|x - y\|_2^2, \text{ where } u \in \delta f(y)$$

where $\delta$ denotes the sub-differential. A $\rho$-smooth function satisfies

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

We also use the following conventions: $\text{sgn}(x)$ 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 $\infty$ if $x \notin C$; $\mathbb{R}^+ = [0, \infty)$, $\mathbb{R}^{++} = (0, \infty)$, and $\mathbb{R}$ is the extended real line $\mathbb{R} \cup \{-\infty, \infty\}$.

## 2 Proximal operators and Moreau envelopes

The key tools we employ are proximal operators and Moreau envelopes. 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 $\text{prox}_\gamma f(x)$ with parameter $\gamma$ are defined as

$$f^\gamma(x) = \inf_z \left\{ f(z) + \frac{1}{2\gamma} \|z - x\|_2^2 \right\} \leq f(x) \tag{3}$$

$$\text{prox}_\gamma f(x) = \arg \min_z \left\{ f(z) + \frac{1}{2\gamma} \|z - x\|_2^2 \right\}.$$
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 $\gamma$ controlling the trade-off. Table 1 provides an extensive list of closed-form solutions.

## 2.1 Properties of Proximal Operators

Our perspective throughout this paper will be to view proximal fixed point algorithm as the gradient of a suitably defined envelope function. By constructing different envelopes one can develop new optimisation 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. For further information, see Parikh and Boyd [2013] who provide 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 that relate to the envelope perspective.

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 $f^\gamma(x)$, which approximates $f$ from below. 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 $\hat{Z}(x) = \text{prox}_{\gamma f}(x)$ is the value that achieves the minimum. Hence,

$$
\text{prox}_{\gamma f}(x) = x - \gamma \partial f^\gamma(x),
$$

where $\partial h(x)$ is understood to be the sub-differential, defined by

$$
\partial h(x) = \{ z : h(y) \geq h(x) + z^T(y - x), \forall y, x \in \text{dom}(h) \},
$$

when $h(x)$ is non-differentiable. Thus, evaluating the proximal operator can be viewed as a gradient-descent step for a regularized version of the original function, with $\gamma$ 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 $\text{prox}_f(x) = \arg \min_{z \in C} \|x - z\|_2^2$ 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 $\frac{\rho}{2} \|x - v\|_2^2$ to an objective function—where $\rho$ 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.

There is a close connection between proximal operators and fixed-point theory,
including that $\text{prox}_{\gamma f}(x^*) = x^*$ if and only if $x^*$ 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:

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

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

Finally, another key property of proximal operators is the Moreau decomposition
for the proximal operator of $f^*$, the dual of $f$:

$$x = \text{prox}_{\lambda f}(x) + \lambda \text{prox}_{\lambda f^*}(\lambda x)$$

$$I - \text{prox}_{\lambda f}(x) = \lambda \text{prox}_{\lambda f^*}(\lambda x) \quad (4)$$

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 5.

All three of these ideas—projecting points onto constraint regions, taking gradient-
descent steps, 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
compactly in terms of proximal operators of log likelihoods or penalty functions.
Here are two examples.

Figure 1 provides a graphical depiction of these two concepts for the simple case $f(x) = |x|$. In general the proximal operator may be set-valued, but it is scalar-valued
in the special case where $f(x)$ is a proper convex function.

**Example 1.** 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 corre-
sponding 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 $\text{prox}_f(x_0) = 0.5$
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.

Let \(\phi(x) = \lambda\|x\|_1\) and consider the proximal operator \(\text{prox}_{\gamma\phi}(x)\). 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

\[
\min_{z \in \mathbb{R}} \left\{ \lambda|z| + \frac{\gamma}{2}(z - x)^2 \right\}.
\]

This problem has solution

\[
\hat{z} = \text{prox}_{\lambda\|x\|_1/\gamma}(x) = \sgn(x)(|x| - \lambda/\gamma)_+ = S_{\lambda/\gamma}(x),
\]

the soft-thresholding operator with parameter \(\lambda/\gamma\).

**Example 2.** Quadratic terms of the form

\[
l(x) = \frac{1}{2} x^T P x + q^T x + r,
\]

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 \((y|x) \sim N(Ax, \Omega^{-1})\), then \(l(x) = (y - Ax)^T \Omega (y - Ax)/2\), or

\[
P = A^T \Omega A, \quad q = -A^T \Omega y, \quad r = y^T \Omega y/2
\]

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

\[
\text{prox}_{l/\gamma}(x) = (P + \gamma I)^{-1}(\gamma A^T x - q),
\]

assuming the relevant inverse exists.

General lesson: the proximal operator provides concise description of many iterative algorithms. Practically useful only if the proximal operator can be evaluated in closed form or at modest computational cost.
3 The proximal gradient method

The goal here is to further describe one of the simplest proximal algorithms, the proximal-gradient method. We do so both because the proximal-gradient method is broadly useful in its own right, and because it is 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 non-differentiable 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 non-differentiable term \( \phi(x) \) at the intermediate point \( v^t \):

\[
x^{t+1} = \text{prox}_{\gamma \phi} (v^t) = \text{prox}_{\gamma \phi} \{ x^t - \gamma \nabla l(x^t) \} .
\] (7)

This can be motivated in at least two ways.

**As an MM algorithm.** Suppose that \( l(x) \) has a Lipschitz-continuous gradient with modulus \( L \). This allows us to construct a majorizing function: whenever \( \gamma \in (0, 1/L] \), we have the majorization

\[
l(x) + \phi(x) \leq l(x_0) + (x - x_0)^T \nabla l(x_0) + \frac{1}{2\gamma} \| x - x_0 \|^2 + \phi(x) ,
\]

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

\[
\hat{x} = \arg \min_x \left\{ \phi(x) + \frac{1}{2\gamma} \| x - u \|^2 \right\} , \quad \text{where} \quad u = x_0 - \gamma \nabla l(x_0) .
\]

This is nothing but the proximal operator of \( \phi \), evaluated at an intermediate gradient-descent step for \( l(x) \).

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 \( L \), i.e.

\[
\| \nabla l(x) - \nabla l(y) \|_2 \leq L \| x - y \|_2 \quad \forall x, y .
\]
3. \( \phi \) is closed and convex, ensuring that \( \text{prox}_{\gamma \phi} \) makes sense.
4. the optimal value is finite and obtained at \( x^* \).

If these conditions are met, than the proximal gradient method converges at rate \( 1/t \) with fixed step size \( \gamma^t = 1/L \).
As the fixed point of a “forward-backward” operator. 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. This has connections (not pursued here) with the forward-backward method for solving partial differentiable equations. Let $\partial$ be the subdifferential operator. A necessary and sufficient condition that $x^*$ minimizes $f(x)$ is that

$$0 \in \partial \{l(x) + \phi(x)\} = \nabla l(x) + \partial \phi(x), \quad (8)$$

the sum of a point and a set. We will use this fact to characterize $x^*$ as the fixed point of the following operator:

$$x^* = \prox_{\gamma \phi} \{x^* - \lambda \nabla l(x^*)\}. \quad (9)$$

To see this, let $I$ be the identity operator. Observe that the optimality condition (8) is equivalent to

$$0 \in \gamma \nabla l(x^*) - x^* + x^* + \partial \phi(x^*)$$

$$x^* - \gamma \nabla l(x^*) \in x^* + \gamma \partial \phi(x^*)$$

$$(I - \gamma \nabla f)x^* \in (I + \gamma \partial \phi)x^*$$

$$x^* = (I + \gamma \partial \phi)^{-1}(I - \gamma \nabla l)x^*$$

$$= \prox_{\gamma \partial \phi}(I - \gamma \nabla l)x^*, $$

the composition of two operators. The final line appeals to the fact (see below) that the proximal operator is the resolvent of the subdifferential operator: $\prox_{\gamma \partial \phi}(x) = (I + \gamma \partial \phi)^{-1}(x)$. Thus to find the solution, we repeatedly apply the operator having $x^*$ as a fixed point:

$$x^{t+1} = \prox_{\gamma^t \phi} \{x^t - \gamma^t \nabla l(x^t)\}. $$

This is precisely the proximal gradient method.

We now show that the proximal operator is the resolvent of the subdifferential operator. By definition, if $z \in (I + \gamma \partial l)^{-1}x$, then

$$x \in (I + \gamma \partial l)z$$

$$x \in z + \gamma \partial l(z)$$

$$0 \in \frac{1}{\gamma}(z - x) + \partial l(x)$$

$$0 \in \partial z \left\{ \frac{1}{2\gamma} \|z - x\|_2^2 + l(x) \right\}. $$

But for 0 to be in the subdifferential (with respect to $z$) of the function on the right-hand side is a necessary and sufficient condition for $z$ to satisfy

$$z = \arg \min_u \left\{ \frac{1}{2\gamma} \|u - x\|_2^2 + l(u) \right\} = \prox_{\gamma l}(x). $$
Therefore \( z = \text{prox}_{\gamma l}(x) \) if and only if \( z \in (I + \gamma \partial l)^{-1}x \). It is interesting that \((I + \gamma \partial l)^{-1}\) is single-valued and therefore a function, even though \( \partial l \) is set-valued.

Our proximal framework also applies to non-convex regularisation penalties, e.g. \( \ell_q \) for \( 0 \leq q \leq 1 \), for which we provide an example in Section 6.4.

3.1 Proximal Newton

Proximal gradient, or forward-backward splitting, is a generalisation of the classical gradient approaches. They only require first-order information and their speed can be improved by using second order information, where the resulting algorithms mimic quasi-Newton procedures. To do this, notice that the quadratic bound used in (7) implements a linear approximation of \( l(x) \); however, one can, naturally, use higher order expansions to construct envelopes. If we let

\[
F_H(x, y) = l(y) + \nabla l(y)^T (x - y) + \frac{1}{2} (x - y)^T H_y(x - y)
\]

Then we can calculate the proximal operators,

\[
\text{prox}_{F_H}(y) = y - (\gamma^{-1} I + H_y)^{-1} \nabla l(y)
\]

Instead of directly using the Hessian, \( H_y = \nabla^2 l(y) \), approximations can be employed leading to quasi-Newton approaches. 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.

Another advantage of proximal Newton methods, when second-derivative information is available, is that convergence results can be extended to non-convex problems; see, for example, [Chouzenoux et al., 2014] and Appendix D.

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 \( \nabla l(x) = A^T \Omega Ax - A^T \Omega y \), and the proximal gradient method becomes

\[
l(x) = \frac{1}{2} (y - Ax)^T \Omega (y - Ax)
\]

\[
x^{t+1} = \text{prox}_{\gamma \phi} \{ x^t - \gamma^t A^T \Omega (Ax^t - y) \}.
\]

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: (1) multiplying the current iterate \( x^t \) by \( A \), and (2) multiplying the residual \( Ax^t - y \) by \( A^T \Omega \). Typically the proximal operator for \( \phi \) will be simple to compute, as in the case of a quadratic or \( \ell_1 \) penalty, and will contribute a negligible amount to the overall complexity of the algorithm.
3.3 Acceleration

One advantage of proximal algorithms is that we can accelerate the sequences within algorithms like (7) by introducing an intermediate step that adds a momentum term to the slack variable, $y$, before evaluating the forward and backwards steps,

$$y = x^t + \theta_{t+1} (\theta_{t}^{-1} - 1) (x^t - x^{t-1})$$

$$x^{t+1} = \text{prox}_{L^{-1} \phi} \left( y - L^{-1} \nabla l(y) \right)$$

with $\theta_t = 2/(t + 1)$ and $\theta_{t+1} (\theta_t^{-1} - 1) = (t - 1)/(t + 2)$.

When $\phi$ is convex the proximal problem is strongly convex, and even more advanced acceleration techniques can be used [Zhang et al., 2010, Meng and Chen, 2011].

4 Envelope Methods

In this section we introduce different types of envelopes: the forward-backward envelope (FBE), Douglas-Rachford envelope (DRE), and the half-quadratic (HQ) envelope, and Bregman divergence envelopes. Within this framework, new algorithms are generated as a gradient step of an envelope, for instance, ADMM methods will be viewed as the gradient step of the dual FBE envelope. Section 5 extends these envelopes to the case of composite functions and describes their relationship to Lagrangian formulations.

4.1 Forward-Backward Envelope

Suppose that we have to minimise $F = l + \phi$ where $l$ is strongly convex and possesses a continuous gradient with Lipschitz constant $L_l$ so that $|\nabla^2 l(x)| \leq L_l$. The penalty $\phi$ is only assumed to be proper lower semi-continuous and convex. If we don’t have an “exact” quadratic envelope (see the discussion in 5.1), then we can argue as follows.

First, we define the FBE, $F_\gamma(x)$, which will possess some desirable properties (see Patrinos and Bemporad [2013]).

$$F_\gamma(x) := \min_v \left\{ l(x) + \nabla l(x)^T (v - x) + \phi(v) + \frac{1}{2\gamma} \|v - x\|^2 \right\}$$

$$= l(x) - \frac{\gamma}{2} \|\nabla l(x)\|^2 + \phi^* (x - \gamma \nabla l(x))$$

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

$$\nabla F_\gamma(x) = (I - \gamma \nabla^2 l(x)) G_\gamma(x)$$

where $G_\gamma(x) = \gamma^{-1} (x - P_\gamma(x))$.

$$P_\gamma(x) = \text{prox}_{\gamma \phi} (x - \gamma \nabla l(x)).$$
With these definitions, we can establish the descent property for the FBE
\[ F_\gamma(x) \leq F(x) - \frac{\gamma}{2} \|G_\gamma(x)\|^2 \]
\[ F(P_\gamma(x)) \leq F_\gamma(x) - \frac{\gamma}{2} (1 - \gamma L_l) \|G_\gamma(x)\|^2 . \]
Hence for \( \gamma \in (0, L_l^{-1}) \) 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.

For example, in the common quadratic case \( l(x) = \frac{1}{2} x^T A x + \eta^T x \) we have strong convexity of the envelope \( F_\gamma(x) \) with
\[ \mu_{F_\gamma} = \min \left( (1 - \gamma \mu_l) \mu_l, (1 - \gamma L_l) L_l \right) \text{ and } L_{F_\gamma} = 2(1 - \gamma \mu_l) \gamma . \]
We can then pick \( \gamma = 1/(\mu_l + L_l) \) where \( \mu_l = \sigma_{\min}(A) \) and \( L_l = \sigma_{\max}(A) \) are the usual singular values.

4.2 Douglas-Rachford Envelope

Mimicking the forward-backward approach, Patrinos et al. [2014] derive the Douglas-Rachford envelope (DRE)
\[ F^{\text{DR}}_\gamma(x) = f^{\gamma}(x) - \frac{\gamma}{2} \|\nabla f^{\gamma}(x)\|^2 + g^{\gamma}(x - 2\gamma \nabla f^{\gamma}(x)) \]
\[ = \min_z \left\{ f \left( \text{prox}_{\gamma f}(x) \right) + \nabla^T (z - \text{prox}_{\gamma f}(x)) + g(z) + \frac{1}{2\gamma} \|z - \text{prox}_{\gamma f}(x)\|^2 \right\} . \]
where \( f^{\gamma} \) is, again, the Moreau envelope of the function \( f \). This can be interpreted as a backward-backward envelope and is a special case of a FBE evaluated at the proximal operator of \( \gamma f \), namely
\[ F^{\text{DR}}_\gamma(x) = F^{\text{FB}}_{\gamma f} \left( \text{prox}_{\gamma f}(x) \right) . \]
Again the gradient of this envelope produces the following proximal algorithm which converges to the solution to \( \min_x \{ f(x) + g(x) \} \) given by the iterations
\[ y^{t+1} = \text{prox}_{\gamma f}(x^t) \]
\[ z^{t+1} = \text{prox}_{\gamma g}(2y^t - x^t) \]
\[ x^{t+1} = x^t + (z^t - y^t) \]
There are many ways to re-arrange the DR algorithm. For example, with an intermediate variable, \( w = y - x \), we could equal well iterate
\[ y^{t+1} = \text{prox}_{\gamma f}(x^t - w^t) , \quad x^{t+1} = \text{prox}_{\gamma g}(y^t + w^t) , \quad w^{t+1} = w^t + (y^t - x^t) \].

We now turn to Half-Quadratic Envelopes (HQE) which are useful for non-convex problems.
4.3 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 $l(x) = \|Ax - y\|^2$. See Nikolova and Ng [2005] for convergence rates and comparisons of the different algorithms.

The half-quadratic envelope (HQE) is defined by

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

where $Q(x, v) = vAx^2$ or $(v - x)^2$

and 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 other function $g(x)$.

**Example 3.** (Half-Quadratic $f + \ell^2$ where $f$ is an envelope).

Suppose that we wish to minimise the functional

$$J(x) = \frac{1}{2} \|Ax - y\|^2 + \gamma f(x) \quad \text{where} \quad f(x) = \sum_{i=1}^{d} f((B^T x - b)_i)$$

Then we need to solve the joint criterion

$$J(x, \lambda) = \frac{1}{2} \|Ax - y\|^2 + \gamma \sum_{i=1}^{d} Q(\delta_i, \lambda_i) + \gamma \sum_{i=1}^{d} \psi(\lambda_i) .$$

where $\delta_i = (B^T x - b)_i$. There is an equivalence between gradient linearisation and quasi-Newton. These algorithms give the iterative mappings:

$$x^{t+1} = L(\hat{\lambda}(x^t))^{-1} A^T y \quad \text{and} \quad x^{t+1} = x^t - L(x^t)^{-1}\nabla_x J(x^t),$$

respectively. They are identical, as the derivative information is

$$\nabla_x J(x) = A^T Ax - A^T y + \gamma \sum_{i=1}^{d} B_i \frac{f'(||\delta_i||)}{||\delta_i||} B_i^T x$$

$$= (A^T A + \gamma B\Lambda(x)B^T)x - A^T y \quad \text{where} \quad \Lambda(x) = \text{diag}(\hat{\lambda}(||\delta_i||_{i=1}^d))$$

$$= L(\hat{\lambda}(x))x - A^T y \quad \text{where} \quad L(\hat{\lambda}(x)) = A^T A + \lambda B\Lambda(x)B^T .$$

Here $\hat{\lambda}(x) = f'(x)/2x$ for Geman-Yang (GY) and $\hat{\lambda}(x) = x - f'(x)$ for Geman-Reynolds (GR).

We can speed up the GY algorithm by rescaling $f_\gamma = \gamma f$ with $\gamma = 1/L$. The normal equation matrix is $B^\gamma_{GY} = A^T A + \gamma^{-1} BB^T$ and we perform iterations with an over-relaxation sequence, $\theta_k$, given by $x^{t+1} = x^t - \theta_k (B^\gamma_{GY})^{-1}\nabla J(x^t)$. For GR, the normal equation matrix is $B_{GR} = A^T A + B\Lambda(x)B^T$. The algorithm iterates

$$\hat{\lambda}_i^{t+1} = f'(\delta_i^t)/\delta_i^t$$

$$x^{t+1} = (B_{GR})^{-1} \left( A^T y + B\hat{\lambda}^{t+1} w \right)$$
4.4 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 minimisation of an objective function of the form $D(x, v) + \phi(x) + \psi(v)$. To minimise over $(x, v)$ we can use an alternating Bregman projection method. To perform the minimisation 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, $D(x, v) \geq 0$ and attains equality at $x = v$. The Bregman divergence has a three-point law of cosines triangle inequality, which helps to establish descent in proximal algorithms (see the note Appendix A). Many commonly used EM and MM algorithms in statistics and variational Bayes models use envelopes of this type.

The key insight is that the proximal operator generated by the $D$-Moreau envelope allows one to add non-smooth regularisation penalties to traditional exponential family models. In our applications, we illustrate this with logistic and Poisson regression both of which can be interpreted as Bregman divergence measures of fit in the objective function.

We now turn to the general case of a quadratic envelope with a composite regularisation penalty.

5 Proximal operators of composite functions

The most common situation in statistical learning is a general composite problem, $l + \phi \circ B$, where we have an objective function of the form

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

One can view much of the optimization landscape for such problems in terms of the following (re)formulations:

- **primal** $F(x) = l(x) + \phi(Bx)$
- **primal-dual** $F_{PD}(x, z) = l(x) + z^T(Bx) - \phi^*(z)$
- **split primal** $F_{SP}(x, y, z) = l(x) + \phi(x) + z^T(Bx - y)$
- **split dual** $F_{SD}(x, y, z) = l^*(y) + \phi^*(z) + x^T(-B^T z - y)$

where $\phi^*$ denotes the dual of $\phi$. The motivation for the primal-dual and the split problems (see Esser et al. [2010]) lies in how they decouple $\phi$ from $B$ without affecting its solution to the primal problem $F(x)$. We refer to this class of problems as joint objective problems. Essentially, we are calculating a marginal mode as a joint mode.

The split problems are Lagrangian formulations that each arise separately from the definition of the convex conjugate or Fenchel dual, and relate to each other, in
the general case, by the Max-Min inequality [Boyd and Vandenberghe, 2009]
\[
\sup_q \inf_v F(q, v) \leq \inf_v \sup_q F(q, v)
\]

In the special case of closed proper convex functions, we have the following
\[
\min_x F(x) = \min_x \sup_z F_{PD}(x, z) = \max_z \min_y F_{SP}(x, y, z) = \max_z \min_y F_{SD}(x, y, z),
\]
made possible for the dual problems by noting the equality in
\[
\phi(Bx) = \sup_z \{z^T Bx - \phi^*(z)\},
\]
when \(\phi\) is convex. In this case,
\[
\min_{y \geq 0} F_{SP}(x, y, z) = \min_{y \geq 0} \{\phi(y) + l(x) + z^T (Bx - y)\}
\]
\[
= l(x) + z^T Bx + \min_{y \geq 0} \{\phi(y) - z^T y\}
\]
\[
= l(x) + z^T Bx - \phi^*(z)
\]
\[
= F_{PD}(x, z)
\]
The solution values \(x^*, y^*, z^*\) that tie all of the objective problems together are the saddle point values, as expected.

Given an objective problem, we must next specify the exact steps to solve the sub-problems within it, i.e. the problems in \(y\) and/or \(z\). In some cases, one might not have closed forms for the dual functions, but instead exact solutions to related problems that share the same critical points, as in the case of proximal algorithm solutions. In fact, the proximal envelopes in Section 4 can be derived as specific sequences of proximal solutions for the variables in an objective problem.

Especially in cases where multiple majorization steps are taken (to solve for–say–\(y\) and \(z\) in a \(F_{SP}\) problem) the use of proximal operators, their properties, and the associated fixed-point theory can simplify otherwise lengthy constructions and convergence arguments. As well, using the proximal operator’s properties, such as the Moreau identity, one can move easily between the different objective problems and, thus, primal and dual spaces. It is also worth mentioning that the efficacy of certain acceleration techniques can depend on the objective problem (see Beck and Teboulle [2014]) and, similarly, the proximal steps taken.

A useful interpretation of the additional squared term introduced by a proximal step is as an augmented Lagrangian for a linear constraint. Specifically, the addition of a squared term in the \(F_{SP}\) problem leads to the ADMM estimation technique in which one iterates through conditional solutions to \(x\) and \(z\) at each step, with solutions given by proximal points. Both Parikh and Boyd [2013] and Chen and Teboulle [1994] observe that, for the splitting/composite problem, the augmented Lagrangian for ADMM is
\[
\phi(y) + l(x) + z^T (Bx - y) + \frac{\rho}{2} \|Ax - z\|^2
\]
\[ F_{SP}(x, y, z) + \frac{\rho}{2} \|Ax - z\|^2 \]

The implied proximal operator for an optimization step in \( x \) would still involve the composite argument \( Ax \), so when the solution to the composite operator isn’t available one can consider linearizing \( \frac{\rho}{2} \|Ax - z\|^2 \) with \( \frac{\rho}{2\lambda_A} \|x - z\|^2 \), where \( \sigma_{\text{max}}(A^TA) \leq \lambda_A \), yielding
\[
\frac{\rho}{2} \|Ax - z\|^2 \leq \frac{\rho}{2\lambda_A} \|x - z\|^2
\]

Further details are provided for quadratic \( l(x) \) in Section 5.1.

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] details such splitting methods in terms of augmented-Lagrangians for non-convex objectives.

We now give an example of an alternative, and sometimes faster approach, which can arrive at similar, or even identical, results by using basic properties of proximal operators.

**Example 4.** For proper, convex \( l(x), \phi(x) \) we start with the split-dual problem
\[
\max \inf \{ l(x) + z^T(Bx) - \phi^*(z) \}
\]
and notice that the argmin for the part in \( x \), \( l(x) + z^T(Bx) \), is given by the fixed point
\[
x^* = \text{prox}_{\lambda(l(x) + z^T Bx)}(x^*),
\]
when an appropriate bounding term \( \lambda_l > 0 \) exists. In other words, when we can find a quadratic majorizer for \( l(x) + z^T Bx \). By a property of proximal operators, namely
\[
\text{prox}_{g(q^* + u^T z)}(q) = \text{prox}_{g(q - u)}(q), \tag{11}
\]
which is obtained by completing the square in the definition of the operator, we have
\[
x^* = \text{prox}_{\lambda_l(l + z^TB)(z)}(x^*) = \text{prox}_{\lambda_l}(x^* - \lambda_l B^T z).
\]
As an optimization only in \( z \),
\[
\max_z \{ l(x^*) + z^T(Bx^*) - \phi^*(z) \} = -\min_z \{ \phi^*(z) - z^T(Bx^*) - l(x^*) \}
\]
Next, we can take yet another proximal/majorization step, for the minimization problem, \( \phi^*(z) - z^T(Bx^*) \), in \( z \) with constant \( \lambda_\phi \). Using (11) and (4), we find that the argmin satisfies
\[
z^* = \text{prox}_{\lambda_\phi \phi^*}(z^* + \lambda_\phi Bx^*)
\]
with the latter given by (4). Hence, we have the following implied iterative algorithm:

\[
\begin{align*}
\lambda_t &= 1 \\
\phi &= 1
\end{align*}
\]

\[
\begin{align*}
x^* &= \text{prox}(x^* - \lambda_B^T z^*) \\
\phi^* &= \text{prox}(\phi^* + Bx^*)
\end{align*}
\]

If we further separate the last step in (12) into two steps—and simplify by setting \(\lambda_t = \lambda_\phi = 1\)—we arrive at

\[
\begin{align*}
x^* &= \text{prox}(x^* - B^T u^*) \\
w^* &= \text{prox}(u^* + Bx^*) \\
u^* &= u^* - (w^* - Bx^*)
\end{align*}
\]

This has the basic form of techniques like alternating split Bregman, ADMM, split inexact Uzawa, etc. The differences often involve assumptions on \(l\) and the exact order of steps (see Chen et al. [2013] for a detailed description).

### 5.1 General Quadratic Composition

Consider, now, the most general form of a quadratic objective

\[
\begin{align*}
\arg\min_x \inf_y \left\{ F(x, y) = \frac{1}{2} x^T \Lambda(y) x - \eta^T y + \phi(Bx) \right\}
\end{align*}
\]

where \(\Lambda(y) \succ 0\). Again, such forms can arise when one majorizes with a second-order approximation of \(l(x)\) around \(y\). This also makes (13) the Moreau envelope defined in (3). The general quadratic case, in which \(\Lambda(y)\) is not necessarily diagonal, can be addressed with splitting techniques.

This form, when \(\Lambda(y)\) is symmetric positive definite, encompasses the approaches of Geman and Yang [1995], Geman and Reynolds [1992]. Assuming \(B\) is positive definite, a proximal point solution can be obtained by setting \(l(x) = x^T \Lambda(y) x - \eta^T x\) in (12). The general solution to a quadratic-form proximal operator—like (6)—is, again, given by

\[
\text{prox}_{\lambda l(x)}(q) = (I + \lambda l \Lambda(y))^{-1} (q + \lambda l \eta)
\]

which, together with the split-dual formulation, implies a proximal point algorithm of the form

\[
x^* = \text{prox}_{\lambda l(x)}(x^* - \lambda_B^T z^*)
\]
Then, with $\lambda f$

Again, we can use a forward-backward proximal solution to the above problem, where

Solving for $x$ with the split-dual objective for $A$ be symmetric positive definite, and $\|z\|$ bound inspired by the 2-norm inequality

This would still arrive at the same solution by a first-order quadratic forward-backward algorithm on the dual objective, $F_{PD}$. For simplicity, let $\Lambda(y) = A$ be symmetric positive definite, and $A = R^TR$ its Cholesky decomposition. Starting with the split-dual objective for $f(x) = \frac{1}{2}x^TAx - \eta^Tz$,

\[
\min_{x} \max_{z} \left\{ \frac{1}{2}x^TAx - \eta^Tz + z^TBx - \phi^*(z) \right\} = \min_{x} \max_{z} \left\{ \frac{1}{2}\|Rx - R^{-1}(\eta - B^Tz)\|^2 - \frac{1}{2}\|R^{-1}(\eta - B^Tz)\|^2 - \phi^*(z) \right\}
\]

Solving for $x$ is straight-forward from the first line and $x^* = A^{-1}(\eta - B^Tz)$, but from the second line we could still arrive at the same solution by a first-order quadratic bound inspired by the 2-norm inequality $\|Mv\| \leq \|M\|\|v\|$. That is

\[
\|Rx - R^{-1}(\eta - B^Tz)\|^2 \leq \|R\|^2\|x - A^{-1}(\eta - B^Tz)\|^2 \\
\leq \sigma_{\max}(A)\|x - A^{-1}(\eta - B^Tz)\|^2
\]

Now, at $x = x^*$ we have the following problem in $z$:

\[
\max_{z} \left\{ -\frac{1}{2}\|R^{-1}(\eta - B^Tz)\|^2 - \phi^*(z) \right\} = \min_{q} \left\{ \frac{1}{2}\|R^{-1}B^Tz - R^{-1}\eta\|^2 + \phi^*(z) \right\}
\]

Again, we can use a forward-backward proximal solution to the above problem, where $f(z) = \frac{1}{2}\|R^{-1}B^Tz - R^{-1}\eta\|^2$, so that

$\nabla f(z) = BR^{-T}(R^{-1}B^Tz - R^{-1}\eta) = \lambda_2(BA^{-1}B^Tz - BA^{-1}\eta)$,

Then, with $\lambda_2 \geq \sigma_{\max}(BA^{-1}B^T)/2$, we can obtain $z^*$ as the proximal solution

\[
z^* = \text{prox}(z - \lambda_2 \nabla l(z))
\]
\[ z^* = \text{prox}(z - \lambda_2(BA^{-1}B^T z + BA^{-1} \eta)) \]
\[ = \left( I - \text{prox}_{\lambda_2^{-1}\phi} \right) \circ \left( (I - \lambda_2 BA^{-1}B^T) z + BA^{-1} \eta \right) \]  

(14)

In sum, we have an implied proximal point algorithm similar to (12) that is, instead, based on a first-order forward-backward method.

**Example 5.** A related example of this variety of split forward-backward algorithm is used by Argyriou et al. [2011], who apply Picard-Opial iterations given by

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

for \( \kappa \in (0, 1) \), to find a fixed point, \( v^* \), of the operator

\[ H(v) := \left( I - \text{prox}_{\gamma^{-1}\phi} \right) (BA^{-1} \eta + (I - \gamma BA^{-1}B^T)v), \forall v \in \mathbb{R}^p \]

where \( 0 < \gamma < 2/\sigma_{\text{max}}(BA^{-1}B^T) \). The operator \( H \) is understood to be non-expansive, so, by Opial’s theorem, one is guaranteed convergence, and, when \( H \) is a contraction, this convergence is linear. After finding \( v^* \), one sets \( x^* = A^{-1}(\eta - xB^Tv^*) \).

Noting the similarities with (14), we see that \( v \) here can be interpreted as the dual variable \( z \). What distinguishes this approach from others is that there are fewer upfront restrictions on the matrix operator \( B \). Chen et al. [2013] discuss the number of iterations, \( k \), in the process of finding the fixed point \( v^* \) and detail a one-step algorithm with similar scope.

### 5.2 Divide and Concur

In the most general setting we have a sum of \( J \) composite functions and an optimization problem of the form

\[ \max_{x \in X} \sum_{j=1}^{J+1} l_j(A_jx) + \phi(Bx) \]

The approach is to add slack variables, \( z_j \) for \( j \in [1, \ldots, J+1] \), to “divide” the problem together with equality constraints so that the solutions “concur”. We have the equivalent constrained optimization problem

\[ \max_{x,z} \sum_{j=1}^{J+1} l_j(z_j) \text{ under constraints } z_j = A_jx, z_{J+1} = Bx. \]

where \( l_{J+1} = \phi, A_{J+1} = B \). This can be solved using an iterative proximal splitting algorithm (e.g. multiple ADMM, split Bregman). Specifically, under ADMM one finds (see Parikh and Boyd [2013])

\[ x_j^{t+1} = \text{prox}_{\lambda_j \circ A_j}(\bar{x}^t - u_j^k) \]
\[ u^t_j + 1 = u^t_j + x^t_j + 1 - \bar{x}^t + 1. \]

where \( \bar{x}^t = \frac{1}{J + 1} \sum_{j=1}^{J+1} x^t_j. \) Divide and Concur [Gravel and Elser, 2008] methods are the natural approach to big data problems as they break a hard high-dimensional problem into tractable sub-problems using the divide variables and then they use a concur step to find the global solution from the solutions to each sub-problem.

6 Applications

6.1 Logit + \( \ell^2 \) Simulation

To illustrate our approach, we simulate observations from the model

\[
(y_i | p_i) \sim \text{Binom}(J, p_i) \\
p_i = \text{logit}^{-1}(a^T_i x)
\]

where \( i = 1, \ldots, 100, a^T_i \) is a row vector of \( A \in \mathbb{R}^{100 \times 300}, x \in \mathbb{R}^{300} \) and \( J = 2 \). The \( A \) matrix is simulated from \( \text{N}(0, 1) \) variates and normalized column-wise. The signal \( x \) is also simulated from \( \text{N}(0, 1) \) variates, but with only 10% of entries being non-zero.

Here \( m_i \) are the number of trials, \( y_i \) the number of successes and \( m = \sum_{i=1}^n m_i \) the total number of trials in the classification problem. The composite objective function for sparse logistic regression is then given by

\[
\arg\min_x \sum_{i=1}^n \left\{ m_i \log(1 + e^{a^T_i x}) - y_i a^T_i x \right\} + \lambda \sum_{j=1}^p |x_j|
\]

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 \( L \) given by \( \| A^T A \|_2 / 4 = \sigma_{\max}(A) / 4 \), and a penalty coefficient \( \lambda \) 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 non-descent nature of the algorithm and the clear advantage of adding acceleration.

6.2 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

\[
\Lambda(v) = 2 \sum_{i=1}^n m_i \lambda(a^T_i v) a_i a^T_i = 2 A^T \text{diag}(m \cdot \lambda(Av)) A
\]

\[
\kappa^T = 2 \sum_{i=1}^n (y_i - m_i / 2) a_i^T.
\]
Figure 2: (Adjusted) objective values for iterations of the proximal gradient method, with and without acceleration, applied to a logistic regression problem with an $\ell_1$ penalty.

Now we compute $x_t$, conditional on $y$, for the envelope

$$\sum_{i=1}^{n} \left\{ m_i \log(1 + e^{a_i^T x}) - y_i a_i^T x \right\} + \| D^{(1)} x \|_1 = \min_y \left\{ \frac{1}{2} x^T \Lambda(y) x - \kappa^T x + c(y) + \gamma \| D^{(1)} x \|_1 \right\}$$

To do this, we employ the Picard-Opial composite method of Argyriou et al. [2011]. Simulations were performed in a similar fashion as Section 6.2 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 pre-conditioning, we can perform the following decompositions: $A = U \Sigma V^T$, the singular value decomposition (SVD), $\Lambda^{-1}(v) = \frac{1}{2} A^{-1} D^{-1} A^T$, where $D = \text{diag}(m \cdot \lambda(Av))$. This implies that one SVD of $A$, or generalized inverse, is required to compute all future $\Lambda^{-1}(v)$ and thus providing computational savings.

### 6.3 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

$$(y|x) \sim \text{Pois}(\exp(Ax))$$
In our simulation, the true sparse parameter vector $x$ has 10% non-zero signals from $\mathcal{N}(0,1)$. The design matrix $A \in \mathbb{R}^{100 \times 300}$ is also generated from $\mathcal{N}(0,1)$, then column normalized.

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

$$F(x) = \sum_{i=1}^{n} \exp(a_i^T x) - y_i a_i^T x + \sum_{j=1}^{p} |x_j - x_{j-1}| = \sum_{i=1}^{n} \exp(a_i^T x) - y_i a_i^T x + \|D^{(1)}x\|_1 .$$

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 penalised likelihood estimation.
6.4 Non-Convex $\| \cdot \|^2 + \ell^q, 0 < q < 1$.

A common non-convex penalty is the bridge norm, $\ell^q$ for $0 < q < 1$. There are a number of ways of developing a proximal algorithm to solve such problems. The proximal operator of $\ell^q$ has a closed-form, multi-valued solution and convergence results are available for proximal methods in Marjanovic and Solo [2013] and Attouch et al. [2013]. For this example, we choose the former approach.

The regularization problem is to find the $\| \cdot \|^2 + \ell^q$ minimizer for $0 < q < 1$ of the regression problem

$$
\hat{x}^q_\lambda := \arg\min_{x} \left\{ \frac{1}{2} \|y - Ax\|^2 + \lambda \sum_{j=1}^{p} |x_j|^q \right\},
$$

The component-wise, set-valued proximal $\ell^q$-operator is given by

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

where

$$
b_{\lambda, q} = (2\lambda(1 - q))^{\frac{1}{q - 1}} \\
h_{\lambda, q} = b_{\lambda, q} + \lambda q b_{\lambda, q}^{q-1}
$$
\[ \hat{x} + \lambda q \hat{x}^{q-1} = |y|, \hat{x} \in (b_{\lambda,q}, |x|) \] 

Attouch et al. [2013] describe how the objective for this problem is a Kurdyka-Lojasiewicz (KL) function, which provides convergence results for an inexact (multi-valued proximal operator) forward-backward algorithm given by

\[ x^{t+1} \in \text{prox}_{\lambda_t \gamma} \left( x^t - \gamma_t (A^T Ax^t - A^T b) \right) . \]

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

An alternative approach is the variational representation of the \( \ell^q \)-norm; however, this doesn’t satisfy the convergence conditions of Allain et al. [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

\[ 0 = \nabla_i f(x) = A_i^T (Ax - y) = A_i^T (A_i x_i + A_{-i} x_{-i} - y) \]

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

\[ x_i^{t+1} = \frac{A_i^T (y - A_{-i} x_i^{t+1})}{A_i^T A_i} = \frac{A_i^T r_t}{\|A_i\|^2} + x_i^t \]

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

We simulate a data vector \( y \in \mathcal{R}^n \) from a regression model

\[ y = Ax + \sigma \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, 1) \]

with an underlying sparse parameter value \( x \in \mathcal{R}^d \) with \( n = 100, d = 256 \), in which the true sparse \( x \) has 5% non-zero signals generated from \( \mathcal{N}(0, 1) \). The design matrix \( X \in \mathcal{R}^{100 \times 256} \) is also generated from \( \mathcal{N}(0, 1) \) 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-regularisation penalty and the power in the \( \ell^q \) penalty. Essentially, this consists of contours of \( \log_{10}(\text{MSE}(\hat{\beta})) \) 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 \( \hat{\beta}_q^\lambda \) can jump to sparsity as \( 0 < q < 1 \), and this will be illustrated in a regularized path for the next example.
6.5 Prostate Data

As a practical example of our methodology, we consider the prostate cancer dataset, 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 regularisation path using

$$
\hat{\beta}_\lambda^q := \arg\min_{\beta} \left\{ \frac{1}{2} ||y - X\beta||^2 + \lambda \sum_{j=1}^{p} |\beta_i|^q \right\}.
$$

We can use our exact proximal operator for $\ell^q$ and solve the harder non-convex problem. Figure 6 shows the regularisation path. The major difference is, again, in the jumps to a sparse solution.

7 Discussion

Proximal algorithms provide an extension of classical gradient descent methods and have properties that can be used to arrive at many different algorithmic implementations. They are iterative shrinkage methods that extend traditional EM and MM
algorithms—which are presently commonplace in statistics. Beck and Sabach [2013] provide a historical perspective on iterative shrinkage algorithms by mainly focusing on Weiszfeld [1937] algorithm. The split Lagrangian methods described here were originally developed by Hestenes [1969] and Rockafellar [1973]. 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-Lojasiewicz inequality (see Attouch et al. [2013]).

The purpose of our approach was to develop proximal algorithms for composite functions which are a sum of a linear or quadratic envelope together with a function that has a closed-form proximal operator and is easy to evaluate. Numerous studies exist that demonstrate the efficacy and breadth of application of this approach. Micchelli et al. [2013, 2011] study proximal operators for composite operators for $\ell^2$ and $\ell^1/TV$ denoising models. Argyriou et al. [2011] describe numerical advantages of the proximal operator approach versus traditional fused lasso implementations. Chen et al. [2013] provides a further class of fixed point algorithms that don’t rely on fully solving the Picard-Opial iterates for the nonlinear operator, $H$. Their algorithm proceeds by single updates before other gradient steps.

Many MM block descent algorithms converge very slowly and there are a number of tools available to speed convergence. 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 non-smooth composite problems. Attouch and Bolte [2009], Noll [2014] provide further convergence properties for non-smooth functions. O’Donoghue and Candes [2012] use adaptive restart to improve the convergence rate of accelerated gradient schemes. Giselsson and Boyd [2014] show how preconditioning can help with convergence for ill-conditioned problems. 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 (backwards) steps with dual (backwards) progress. By linearly coupling these two steps they improve convergence. Giselsson and Boyd [2014] 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 mixed exponential family models, and the relationship between proximal splitting and variational Bayes methods in graphical models.

References

Marc Allain, Jérôme Idier, and Yves Goussard. On global and local convergence of half-quadratic algorithms. *Image Processing, IEEE Transactions on*, 15(5):1130–1142, 2006.

Zeyuan Allen-Zhu and Lorenzo Orecchia. A novel, simple interpretation of nesterov’s
accelerated method as a combination of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

Andreas Argyriou, Charles A Micchelli, Massimiliano Pontil, Lixin Shen, and Yuesheng Xu. Efficient first order methods for linear composite regularizers. *arXiv preprint arXiv:1104.1436*, 2011.

Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116 (1-2):5–16, 2009.

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the kurdyka-lojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137 (1-2):91–129, 2013.

Amir Beck and Shoham Sabach. Weiszfeld’s method: Old and new results. *Journal of Optimization Theory and Applications*, pages 1–40, 2013.

Amir Beck and Marc Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Mathematical Methods of Operations Research*, 59(2):235–247, 2004.

Amir Beck and Marc Teboulle. A fast dual proximal gradient algorithm for convex minimization and applications. *Operations Research Letters*, 42(1):1–6, 2014.

Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 259–302, 1986.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141, 2013.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2009.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.
Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

Caroline Chaux, Patrick L Combettes, Jean-Christophe Pesquet, and Valérie R Wajs. A variational formulation for frame-based inverse problems. *Inverse Problems*, 23(4):1495, 2007.

Gong Chen and Marc Teboulle. A proximal-based decomposition method for convex minimization problems. *Mathematical Programming*, 64(1-3):81–101, 1994.

Peijun Chen, Jianguo Huang, and Xiaoqun Zhang. A primal–dual fixed point algorithm for convex separable minimization with applications to image restoration. *Inverse Problems*, 29(2):025011, 2013.

Emilie Chouzenoux, Jean-Christophe Pesquet, and Audrey Repetti. Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. *Journal of Optimization Theory and Applications*, 162(1):107–132, 2014.

I Csisz, Gábor Tusnády, et al. Information geometry and alternating minimization procedures. *Statistics and decisions*, 1984.

Daniel Duckworth. The big table of convergence rates. [https://github.com/duckworthd/duckworthd.github.com/blob/master/blog/big-table-of-convergence-rates.html](https://github.com/duckworthd/duckworthd.github.com/blob/master/blog/big-table-of-convergence-rates.html), 2014.

Ernie Esser, Xiaoqun Zhang, and Tony F Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010.

M.A.T. Figueiredo and R.D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12:906–16, 2003.

Pierre Frankel, Guillaume Garrigos, and Juan Peypouquet. Splitting methods with variable metric for kl functions and general convergence rates. 2014.

Donald Geman and George Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Transactions on pattern analysis and machine intelligence*, 14(3):367–383, 1992.

Donald Geman and Chengda Yang. Nonlinear image recovery with half-quadratic regularization. *Image Processing, IEEE Transactions on*, 4(7):932–946, 1995.

Pontus Giselsson and Stephen Boyd. Preconditioning in fast dual gradient methods. In *Proceedings of the 53rd Conference on Decision and Control*, 2014.

Simon Gravel and Veit Elser. Divide and concur: A general approach to constraint satisfaction. *Physical Review E*, 78(3):036706, 2008.
P. J. Green, K. Łatuszyński, M. Pereyra, and C. P. Robert. Bayesian computation: a perspective on the current state, and sampling backwards and forwards. *ArXiv e-prints*, February 2015.

Peter J Green. On use of the em for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 443–452, 1990.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

Magnus R Hestenes. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320, 1969.

YH Hu, C Li, and XQ Yang. Proximal gradient algorithm for group sparse optimization.

Sindri Magnússon, Pradeep Chathuranga Weeraddana, Michael G Rabbat, and Carlo Fischione. On the convergence of alternating direction lagrangian methods for non-convex structured optimization problems. *arXiv preprint arXiv:1409.8033*, 2014.

Goran Marjanovic and Victor Solo. On exact ℓq denoising. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6068–6072. IEEE, 2013.

Bernard Martinet. Brève communication. régularisation d’inéquations variationnelles par approximations successives. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 4(R3):154–158, 1970.

Xiangrui Meng and Hao Chen. Accelerating nesterov’s method for strongly convex functions with lipschitz gradient. *arXiv preprint arXiv:1109.6058*, 2011.

Charles A Micchelli, Lixin Shen, and Yuesheng Xu. Proximity algorithms for image models: denoising. *Inverse Problems*, 27(4):045009, 2011.

Charles A Micchelli, Lixin Shen, Yuesheng Xu, and Xueying Zeng. Proximity algorithms for the ℓ1/tv image denoising model. *Advances in Computational Mathematics*, 38(2):401–426, 2013.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate o(1/k²). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

Mila Nikolova and Michael K Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal on Scientific computing*, 27(3):937–966, 2005.

Dominikus Noll. Convergence of non-smooth descent methods using the kurdyka–lojasiewicz inequality. *Journal of Optimization Theory and Applications*, 160(2):553–572, 2014.
Brendan O’Donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, pages 1–18, 2012.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

Panagiotis Patrinos and Alberto Bemporad. Proximal newton methods for convex composite optimization. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 2358–2363. IEEE, 2013.

Panagiotis Patrinos, Lorenzo Stella, and Alberto Bemporad. Douglas-rachford splitting: complexity estimates and accelerated variants. *arXiv preprint arXiv:1407.6723*, 2014.

Nicholas G. Polson and James G. Scott. Local shrinkage rules, Lévy processes, and regularized regression. *Journal of the Royal Statistical Society (Series B)*, 74(2):287–311, 2012.

EA Papa Quiroz and P Roberto Oliveira. Proximal point methods for quasiconvex and convex functions with bregman distances on hadamard manifolds. *J. Convex Anal*, 16(1):46–69, 2009.

R. Tyrell Rockafellar and R. J-B Wets. *Variational Analysis*. Springer, 1998.

R Tyrell Rockafellar. Conjugate duality and optimization. Technical report, DTIC Document, 1973.

R Tyrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(259–68), 1992.

Wesley Tansey, Oluwasanmi Koyejo, Russell A. Poldrack, and James G. Scott. False discovery rate smoothing. Technical report, University of Texas at Austin, 2014.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society (Series B)*, 67:91–108, 2005.

R.J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1):285–323, 2014.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

John Von Neumann. *Functional operators: The geometry of orthogonal spaces*. Princeton University Press, 1951.
Endre Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Math. J.*, 43(355-386):2, 1937.

Daniella M. Witten, Robert Tobshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–34, 2009.

Xinhua Zhang, Ankan Saha, and SVN Vishwanathan. Regularized risk minimization by nesterov’s accelerated gradient methods: Algorithmic extensions and empirical studies. *arXiv preprint arXiv:1011.0472*, 2010.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320, 2005.

A Convergence

We now establish convergence results for the forward-backward proximal solution to (13) given in (9)

\[ x^* = \text{prox}_{\phi/\lambda} \{ x - \nabla l(x)/\lambda \} , \]

when \( l \) and \( \phi \) are lower semi-continuous and \( \nabla l \) is Lipschitz continuous. We also assume that \( \text{prox}_{\phi/\lambda} \) is non-empty and can be evaluated independently in each component of \( y \).

Recalling the translation property of proximal operators stated in 11, we can say

\[ x^* = \text{prox}_{\phi/\lambda} \{ x - \nabla l(x)/\lambda \} = \text{prox}_{(\phi(z) + \lambda \nabla l(z)^T z)/\lambda} (x) \]

\[ = \arg\min_z \left\{ \phi(z) + \nabla l(z)^T (z - x) + \frac{\lambda}{2} \| x - z \|^2 \right\} \]

By the proximal operator’s minimizing properties, its solution \( x^* \) satisfies

\[ \phi(x^*) + \nabla l(x^*)^T (x^* - x) + \frac{\lambda}{2} \| x - x^* \|^2 \leq \phi(x) \]

providing a sort of quadratic minorizer for \( F(y) \) in the form of

\[ l(y) + \phi(x^*) + \nabla l(x^*)^T (x^* - y) + \frac{\lambda}{2} \| y - x^* \|^2 \leq l(y) + \phi(y) \equiv F(y) \]

The Lipschitz continuity of \( \nabla l(x) \), i.e.

\[ l(x) \leq l(y) + \nabla l(y)^T (x - y) + \frac{\gamma}{2} \| x - y \|^2 , \]

also gives us a quadratic majorizer

\[ F(x) \equiv l(x) + \phi(x) \leq l(y) + \nabla l(y)^T (y - x) + \frac{\gamma}{2} \| x - x^* \|^2 \]
which, when evaluated at \( x = x^* \) and combined with our minorizer yields

\[
(\lambda - \gamma) \frac{1}{2} \| x^* - y \|^2 \leq F(y) - F(x^*)
\]

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

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

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

\[
D(x, y) = D(x, z) + D(y, z) + (\nabla l(z) - \nabla l(y))^T(x - y),
\]

so that \( D(x, y) \geq D(x, P(y)) + D(P(y), y) \) where \( P(y) = \arg\min_v D(v, y) \).

**B Nesterov Acceleration**

A powerful addition is Nesterov acceleration. Consider a convex combination, with parameter \( \theta \), of upper bounds for the proximal operator inequality \( z = x \) and \( z = x^* \). We are free to choose \( z = \theta x + (1 - \theta)x^+ \) and \( y \).

If \( \phi \) is convex, \( \phi(\theta x + (1 - \theta)x^+) \leq \theta \phi(x) + (1 - \theta)\phi(x^+) \), then we have

\[
F(x^+) - F^* - (1 - \theta)(F(x) - F^*)
= F(x^+) - \theta F^* - (1 - \theta)F(x)
\leq L(x^+ - y)^T(\theta x^* + (1 - \theta)x - x^+) + \frac{L}{2} \| x^+ - y \|^2
\]

\[
= \frac{L}{2} \left( \| y - (1 - \theta)x - \theta x^* \|^2 - \| x^+ - (1 - \theta)x - \theta x^* \|^2 \right)
= \frac{\theta^2 L}{2} \left( \| u - x^* \|^2 - \| u^+ - x^* \|^2 \right)
\]

Where \( y \) is given in terms of the intermediate steps

\[
\theta u = y - (1 - \theta)x
\]

\[
\theta u^+ = x^+ - (1 - \theta)x
\]

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

\[
y = (1 - \theta_t)x + \theta_t u = x - \theta_{t-1}(1 - \theta_t)(x - x^-)
\]
C  Quasi-convex Convergence

Consider an optimisation problem \( \min_{x \in X} f(x) \) where \( f \) is quasi-convex, continuous and has non-empty set of finite global minima. Let \( x^t \) be generated by the proximal point algorithm

\[
x^t \in \arg\min \left\{ f(x) + \frac{\lambda}{2} \|x - x^t\|^2 \right\}.
\]

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 non-convexity of \( f \). A function \( f \) is quasi-convex when

\[
f(\theta x + (1 - \theta)y) \leq \max\{f(x), f(y)\},
\]

which accounts for a number of non-convex 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, i.e.

\[
U = \{x \in \text{dom}(f) : f(x) \leq \inf_t f(x^t)\},
\]

one finds that \( U \) is a non-empty 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 \( f \) as long as \( \min\{f(x) : x \in \mathbb{R}^d\} \) is nonempty.

D  Non-convex: Kurdyka-Łojasiewicz (KL)

A locally Lipschitz function \( f : \mathbb{R}^d \to \mathbb{R} \) satisfies KL at \( x^* \in \mathbb{R}^d \) if and only if \( \exists \eta \in (0,\infty) \) and a neighbourhood \( U \) of \( x^* \) 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 \( f(x^*) < f(x) < f(x^*) + \eta \) we have

\[
k' \{f(x) - f(x^*)\} \text{dist}(0, \partial f(x)) \geq 1
\]

where \( \text{dist}(0, A) \equiv \sup_{x \in A} \|x\|^2 \).

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 minimisation algorithms for problems like

\[
\min_{x,y} L(x,y) := f(x) + Q(x,y) + g(y),
\]

where \( \nabla Q \) is Lipschitz continuous (see Attouch et al. [2010, 2013]). A typical application involves solving \( \min_{z \in \mathbb{R}^d} \{f(z) + g(z)\} \) via the augmented Lagrangian

\[
L(x,y) = f(x) + g(y) + \lambda^\top (x - y) + \frac{\rho}{2} \|x - y\|^2
\]

where \( \rho \) is a relaxation parameter.

A useful class of functions that satisfy KL as ones that possess uniform convexity

\[
f(y) \geq f(x) + u^\top (y - x) + K\|y - x\|^p, \text{ where } p \geq 1, \forall u \in \partial f(x).
\]

Then \( f \) satisfies KL on \( \text{dom}(f) \) for \( \kappa(s) = pK^{-\frac{1}{p}}s^\frac{1}{p} \).

For explicit convergence rates in the KL setting, see [Frankel et al., 2014].
Figure 6: Proximal results for the prostate data example under the $\ell^q$ penalty.
| Type                  | \(\phi(x)\)                                                                 | \(\text{prox}_{\gamma\phi}(y)\)                                      |
|----------------------|------------------------------------------------------------------------------|---------------------------------------------------------------------|
| Laplace              | \(\omega \|x\|\)                                                         | \(\text{sgn}(x) \max(|x| - \omega, 0)\)                             |
| Gaussian             | \(\tau \|x\|^2\)                                                         | \(x / (2\tau + 1)\)                                               |
| Group-sparse, \(\ell_p\) | \(\kappa \|x\|^p\)       | \(\text{sgn}(x)\rho,\rho \text{ s.t. } \rho + p\kappa\rho^{p-1} = \|x\|\) |
| :                   | \(p = 4/3\)                                                                 | \(x + \frac{4\kappa}{2\tau + 1} ((\chi - x)^{1/3} - (\chi + x)^{1/3})\) |
| :                   | \(p = 3/2\)                                                                 | \(x + 9\kappa^2 \text{sgn}(x) (1 - \sqrt{1 + 16|x|/(9\kappa^2)}) / 8\) |
| :                   | \(p = 3\)                                                                  | \(\text{sgn}(x) \left(\sqrt{1 + 12\kappa|x| - 1}\right) / (6\kappa)\) |
| :                   | \(p = 4\)                                                                  | \(\left(\frac{\chi + x}{\kappa}\right)^{1/3} - \left(\frac{\chi - x}{\kappa}\right)^{1/3}\) |
| Gamma, Chi           | \(-\kappa \log x + \omega x\)                                             | \(\frac{1}{2} \left( x - \omega + \sqrt{(x - \omega)^2 + 4\kappa} \right)\) |
| Double-Pareto        | \(\gamma \log (1 + |x|/a)\)                                              | \(\frac{\text{sgn}(x)}{2} \left\{|x| - a + \sqrt{(a - |x|)^2 + 4d(x)}\right\},\) |
| Huber dist.          | \begin{cases} \tau x^2 & \text{if } |x| \leq \omega / \sqrt{2\tau} \\ \omega \sqrt{2\tau} |x| - \omega^2 / 2 & \text{if } \omega, \tau \in (0, +\infty) \end{cases} | \begin{cases} x/2\tau + 1 & \text{if } |x| \leq \omega (2\tau + 1) / \sqrt{2\tau} \\ x - \omega \sqrt{2\tau} \text{sgn}(x) & \text{if } |x| > \omega (2\tau + 1) / \sqrt{2\tau} \end{cases} |
| Max-entropy dist.    | \(\omega |x| + \tau |x|^2 + \kappa |x|^p\) \(\omega, \tau, \kappa \in (0, +\infty)\), \(2 \neq p \in (1, +\infty)\) | \(\text{sgn}(x) \text{ prox}_{\kappa|\cdot|^p / (2\tau + 1)} \left(\frac{1}{2\tau + 1} \max(|x| - \omega, 0)\right)\) |
| Smoothed-laplace dist. | \(\omega |x| - \ln(1 + \omega |x|)\)                                           | \(\text{sgn}(x) \omega |x| - \omega^2 - 1 + \sqrt{\omega |x| - \omega^2 - 1}^2 + 4\omega |x| \) \(\frac{1}{4\omega}\) |
| Exponential dist.    | \begin{cases} \omega x & x \geq 0 \\ +\infty & x < 0 \end{cases} | \(\begin{cases} x - \omega & x \geq \omega \\ 0 & x < \omega \end{cases}\) |
| Uniform dist.        | \begin{cases} -\omega & x < -\omega \\ x & |x| \leq \omega \\ \omega & x > \omega \end{cases} | \(\begin{cases} x - \omega & x \geq \omega \\ 0 & x < \omega \end{cases}\) |
| 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} | \begin{cases} x + \omega + \sqrt{|x - \omega|^2 + 1} & x < 1/\omega \\ x + \hat{\omega} - \sqrt{|x - \omega|^2 + 1} / 2 & x > 1/\hat{\omega} \end{cases}\) |
| Weibull dist.        | \begin{cases} -\kappa \ln x + \omega x^p & x > 0 \\ +\infty & x \leq 0 \end{cases} \(p \in (1, +\infty), \omega, \kappa \in (-\infty, 0)\) | \(\pi \text{ s.t. } p\omega \pi^p + \pi^2 - x\pi = \kappa\) |
| GIG dist.            | \begin{cases} -\kappa \ln x + \omega x + \rho / x & x > 0 \\ +\infty & x \leq 0 \end{cases} \(\omega, \kappa, \rho \in (-\infty, 0)\) | \(\pi \text{ s.t. } \pi^3 + (\omega - x)\pi^2 - \kappa\pi = \rho\) |

Table 1: Sources: [Chaux et al., 2007] [Hu et al.]
Penalty

$$\phi(t) = \min_s \{ Q(t, s) + \psi(s) \}$$

Minimizer

$$Q(t, s) = \frac{1}{2} t^2 s$$

$$Q(t, s) = (t - s)^2$$

$$|t|^\alpha, \alpha \in (1, 2]$$

$$\sqrt{\alpha + t^2}$$

$$\sqrt{\frac{t^2}{\alpha} - \log \left( 1 + \frac{|t|}{\alpha} \right)}$$

$$\begin{cases} \frac{t^2}{2} \quad & |t| \leq \alpha \\ \alpha |t| - \frac{\alpha^2}{2} \quad & |t| > \alpha \end{cases}$$

$$\log(\cosh(\alpha t))$$

$$-\frac{1}{t + |x|}$$

$$-\frac{1}{1 + \sqrt{\alpha}}$$

$$\alpha |t|^{\alpha - 2}$$

$$\frac{1}{\sqrt{\alpha + t^2}}$$

$$\frac{1}{\alpha (\alpha + |t|)}$$

$$\frac{1}{|t|}$$

$$\frac{1}{\tanh(\alpha t)}$$

$$-2$$

$$\frac{\text{sgn}(t)}{t(|t|+1)^2}$$

$$\frac{\text{sgn}(t)}{t(|t|+1)^2}$$

$$\frac{-\infty}{t(|t|+1)^2}$$

$$\frac{-\infty}{t(|t|+1)^2}$$

$$ct - \frac{t}{\sqrt{\alpha + t^2}}$$

$$ct - \frac{t}{\alpha (\alpha + |t|)}$$

$$\begin{cases} (c - 1)t \quad & |t| \leq \alpha \\ ct - \alpha \text{sgn}(t) \quad & |t| > \alpha \end{cases}$$

$$ct - \alpha \tanh(\alpha t)$$

$$ct - \frac{\text{sgn}(t)}{(|t|+1)^2}$$

$$ct - \frac{1}{2\sqrt{t(|t|+1)^2}}$$

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].

| Algorithm                              | Error Rate                  |
|----------------------------------------|-----------------------------|
| **Algorithm**                          | **Convex**                  | **Strongly Convex** | **Per-Iteration Cost** |
| Accelerated Gradient Descent           | $$O(1/\sqrt{\epsilon})$$   | $$O(\log(1/\epsilon))$$ | $$O(n)$$               |
| Proximal Gradient Descent              | $$O(1/\epsilon)$$           | $$O(\log(1/\epsilon))$$ | $$O(n)$$               |
| Accelerated Proximal Gradient Descent  | $$O(1/\sqrt{\epsilon})$$   | $$O(\log(1/\epsilon))$$ | $$O(n)$$               |
| ADMM                                   | $$O(1/\epsilon)$$           | $$O(\log(1/\epsilon))$$ | $$O(n)$$               |
| Frank-Wolfe / Conditional Gradient Algorithm | $$O(1/\epsilon)$$     | $$O(1/\sqrt{\epsilon})$$ | $$O(n)$$               |
| Newton’s Method                        | $$O(\log(1/\epsilon))$$   | $$O(n^2)$$            |                         |
| Conjugate Gradient Descent            | $$O(n)$$                    |                         | $$O(n^2)$$             |
| L-BFGS                                 | Between $$O(\log(1/\epsilon))$$ and $$O(\log \log(1/\epsilon))$$ | $$O(n^2)$$             |                         |

Table 3: See [Duckworth, 2014].

36