A Pseudocode for gradient descent implementation

For concreteness, we write down in Algorithm 1 the gradient descent algorithm for solving

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2\lambda} \|X\beta - y\|^2_2 + \|\beta\|_1,$$

where we recall that $X \in \mathbb{R}^{m \times n}$. The choice of $\lambda = 0$ corresponds to the Basis-pursuit setting. Note that $\nabla f(\beta_t) = g_t$ is computed either as in line 5 or line 9 of the algorithm and one can use these computations for any gradient based algorithm (e.g. BFGS). Note also that this is simply gradient descent on a smooth function, and one can apply typical methods to choosing the stepsize $\gamma_k$, such as the Barzilai-Borwein stepsize [Barzilai and Borwein, 1988].

Algorithm 1: Gradient descent implementation of Ncvx-Pro for solving Lasso.

1. initialization $v_0 \in \mathbb{R}^n$ (with no zero entries), stepsize $\gamma_t > 0$;
2. Result: $\beta_t$
3. while not converged do
4. if $n \leq m$ and $\lambda > 0$ then
5. $u_t = - (\text{diag}(v_t)X^TX \text{diag}(v_t) + \lambda \text{Id})^{-1} (v_t \odot X^T y);$  
6. $g_t = v_t \odot v_t + \frac{1}{2} u_t \odot X^T (Xu_t \odot v_t - y);$  
7. $\beta_t = u_t \odot v_t;$
8. else
9. $\alpha_t = - (X \text{diag}(v_t \odot v_t)X^T + \lambda \text{Id})^{-1} y;$  
10. $g_t = v_t \odot v_t - v_t \odot |X^T \alpha_t|^2;$  
11. $\beta_t = -v_t \odot v_t \odot X^T \alpha_t;$
12. end
13. $v_{t+1} = v_t - \gamma_t g_t$
14. end

B Proofs and additional results for Section 2

Proof to Theorem 1 To show that i) implies ii), recall that a convex, proper and lower semicontinuous function $-\varphi$ can be written in terms of its convex conjugate which has domain $\mathbb{R}^d$. By writing $\beta^2 \triangleq \beta \circ \beta$, using the definition of $R$, we have

$$-R(\beta^2) = -\varphi(\beta^2) = \sup_{v \leq 0 \atop u \geq 0} \beta^2(v, u) = -\inf_{u \geq 0} (\beta^2, u) + (\varphi)^*(\beta^2).$$

which is ii) with $\psi(u) \triangleq (\varphi)^*(\beta^2)$ as required.

*University of Bath, Bath BA2 7AY, UK cmshp20@bath.ac.uk  †CNRS and DMA, Ecole Normale Supérieure, 45 rue d’Ulm, F-75230 PARIS cedex 05, FRANCE, gabriel.peyre@ens.fr

Conversely, if $R$ is of the form in ii), then
\[
R(\beta) = \inf_{u \in \mathbb{R}_+^2} \langle u, \beta^2 \rangle + \psi(u) = - \sup_{u \in \mathbb{R}_+^2} -\langle u, \beta^2 \rangle - \psi(u),
\]
so $R(\beta) = -\psi^*(-\beta \odot \beta)$ and $-\psi^*(-\cdot)$ is clearly a proper, upper semicontinuous, concave function.

For the expression of $\psi$ when $R$ is a norm, from the above, we know that $\psi = (-\varphi)^*(-z)$, and recall that for any norm, $R(\beta) = \max_{R^*(w) \leq 1} (w, \beta)$. So,
\[
\psi(z) = \max_{u \geq 0} (-u, z) + \varphi(u)
= \max_{\beta} (-\beta^2, z) + \varphi(\beta^2) = \max_{\beta} (-\beta^2, z) + R(\beta)
= \max(-z, \beta^2) + \max_{R^*(w) \leq 1} (\beta, w) = \max \left\{ 1 \frac{u^2}{w_i}, \right\}
\]
where in the line, we swapped the two maximums and used the optimality condition over $\beta$ which is $2\beta \odot z = u$. That is, $h(\eta) = 2\psi(-\frac{\lambda}{2\eta}) = \max_{R^*(w) \leq 1} \sum_i w_i^2 \eta_i$.

To derive the identity for $R(\beta)^2$, by the Cauchy-Schwarz inequality
\[
R(\beta)^2 = \sup_{R^*(w) \leq 1} (\beta, w)^2 \leq \sup_{R^*(w) \leq 1} \left( \sum_i \beta_i^2 \eta_i \right) \left( \sum_i \frac{u_i^2}{\eta_i^2} \right) = 4\psi(\eta) \sum_i \beta_i^2 \eta_i
\]
for all $\eta > 0$. Therefore,
\[
R(\beta)^2 = \inf_{\eta \geq 0, \psi(\eta) \leq \frac{1}{4}} \sum_i \beta_i^2 \eta_i
= \sup_{\lambda \geq 0} \inf_{\eta \geq 0} \lambda (\psi(\eta) - \frac{1}{4}) + \sum_i \beta_i^2 \eta_i
= \sup_{\lambda > 0} \frac{\lambda}{4} + \lambda R(\beta/\lambda) = \sup_{\lambda > 0} \frac{\lambda}{4} + \sqrt{\lambda} R(\beta) = R(\beta)^2
\]
where we used the identity $\lambda R(\beta/\sqrt{\lambda}) = \sum_i \beta_i^2 \eta_i + \lambda \psi(\eta)$ and the fact that $R$ is one positive homogeneous.

We derive some properties of the function $h$:

**Lemma 1.** Consider the function $\varphi$ and $\psi$ from Theorem [7]. If $\varphi : [0, \infty) \to [L, U]$, where $L > -\infty$ and $U \in \mathbb{R} \cup \{+\infty\}$, then $\psi$ is an decreasing function with domain contained in $[0, \infty)$, taking values in $[L, U]$. If $R$ is coercive, then $\lim_{\|z\| \to \infty} \psi(z) = +\infty$.

**Proof to Lemma [7]** Let $\varphi : [0, \infty) \to [L, U]$, where $L > -\infty$ and $U \in \mathbb{R} \cup \{+\infty\}$. We describe the properties of the function $\psi(z) = (-\varphi)^*(-z) = \sup_{u \geq 0} (z, u) + \varphi(u)$.

(i) $\text{dom}(\psi) \subset [0, \infty)$: since $\varphi$ is bounded below, it is clear that for $z < 0$, $\sup_{u \geq 0} (z, u) + \varphi(u) = +\infty$.

(ii) $\psi(0) = \sup_{u \geq 0} \varphi(u) = U$.

(iii) Suppose $M \triangleq \sup \{ v \setminus v \in \partial \varphi(u), u \geq 0 \} < \infty$. Then, for all $z \geq M$, $-z + \partial \varphi(u) < 0$ for all $u \geq 0$ and hence, $\psi(z) = \varphi(0) \geq L$. Therefore, $\psi$ takes values in $[L, U]$. So, if $M$ is finite, then one can restrict the optimisation over $z$ to values in $[0, M]$.

(iv) $\psi$ is decreasing: By Danskin’s theorem [Bertsekas 1997, Prop. B.25], for $z \in \{ v \setminus v \in \partial \varphi(u), u \geq 0 \}$,
\[
\partial \psi(z) = \{-u \setminus u \in \arg\min_{u \geq 0} (u, -z) + \varphi(u)\} \subset (-\infty, 0).
\]

\[\square\]
B.1 Functions on matrices

We have the following result for matrix valued functions. Let $\mathbb{S}^n_+$ denote the set of symmetric positive semidefinite matrices.

**Theorem 1.** Let $R : \mathbb{R}^{n \times m} \to \mathbb{R}$. The following are equivalent

1. $R(B) = \varphi(BB^T)$ where $\varphi$ is a proper concave upper semi-continuous function with domain $\mathbb{S}^n_+$.

2. There exists a convex function $\psi$ such that $R(B) = \min_{Z \in \mathbb{S}^n_+} \text{tr}(B^T ZB) + \psi(Z)$.

Moreover, we have $\psi(Z) = (-\varphi)^*(-Z)$. If $R$ is a norm, then $\psi$ can be written as

$$\psi(Z) = \max_{R^*(W) \leq 1} \frac{1}{4} \text{tr}(W^T Z^{-1} W).$$

Moreover,

$$R(B)^2 = \inf_{Z \in \mathbb{S}^n_+} \left\{ \text{tr}(B^T ZB) \mid \psi(Z) \leq \frac{1}{4} \right\}.$$  \hspace{1cm} (2)

**Nuclear norm** $R(W) = \text{tr}(\sqrt{WW^T})$ where $\sqrt{\cdot}$ is the matrix square root. On the space of symmetric positive semidefinite matrices, $\varphi(B) = \text{tr}(\sqrt{B})$ is concave and $\psi(D) = \frac{1}{4} \text{tr}(D^{-1})$, where we use $\partial_A \text{tr}(A) = (2\sqrt{A})^{-1}$ for all symmetric positive semidefinite matrices and $\partial_A \text{tr}(AB) = B$.

**Nonconvex spectral regularisation** Given a symmetric psd matrix $Z = U \text{diag}(\sigma_i) U^T$ and $\alpha > 0$, let $Z^\alpha = U \text{diag}(\sigma_i^\alpha) U^T$. For $\alpha \in (0, 1)$, consider $R(W) = \text{tr}((WW^T)^{\alpha/2}) = \sum_i \sigma_i^\alpha$ where $\sigma_i$ are the singular values of $W$. Then, given a symmetric psd matrix, $\varphi(Z) = \text{tr}(Z^{\alpha/2})$ which is concave [Bhatia 2009 Thm 4.2.3] and

$$\psi(Z) = \min_{V \in \mathbb{S}^n_+} \text{tr}(VZ) + \varphi(V) = \min_{U \in \mathbb{S}^n_+} \text{tr}(\text{diag}(\sigma) U Z U^T) + \sum_i \sigma_i^\alpha/2$$

$$= \min_{U \in \mathbb{S}^n_+, \sigma \geq 0} -\sum_i \hat{Z}_{ii} \sigma_i + \sum_i \sigma_i^\alpha/2 \quad \text{where} \quad \hat{Z} = U Z U^T$$

$$= C_\alpha \min_{U \in \mathbb{S}^n_+, \sigma \geq 0} -\sum_i \hat{Z}_{ii} \sigma_i + \sum_i \sigma_i^\alpha/2 \quad \text{where} \quad \hat{Z} = U Z U^T \quad \text{and} \quad U \in \mathbb{S}^n_+$$

$$= C_\alpha \text{tr}(Z^{\alpha/2})$$

Therefore,

$$R(B) = \inf_{Z \in \mathbb{S}^n_+} \text{tr}(B^T ZB) + C_\alpha \text{tr}(Z^{\alpha/2}).$$

**Proof of Theorem 1.** To derive (1),

$$\psi(Z) = \max_{U \in \mathbb{S}^n_+} -(U, Z) + \varphi(U) = \max_{V \in \mathbb{S}^n_+} -(VV^T, Z) + \varphi(VV^T)$$

$$= \max_{V \in \mathbb{S}^n_+} -(VV^T, Z) + R(V)$$

Then, (I) follows, since by convex duality and definition of $R^*$,

$$\max_{R^*(W) \leq 1} \frac{1}{4} \text{tr}(W^T Z^{-1} W) = \max_{R^*(W) \leq 1} \max_{V \in \mathbb{S}^n_+} -\langle -Z, VV^T \rangle + \langle V, W \rangle = \max_{V \in \mathbb{S}^n_+} \text{tr}(W^T Z^{-1} W) \text{tr}(B^T ZB)$$

Finally, by the submultiplicative property of the Frobenius norms, for all $Z \in \mathbb{S}^n_+$ with $Z \succ 0$,

$$R(B)^2 = \sup_{R^*(W) \leq 1} \text{tr}(W^T Z^{-1} W) \text{tr}(B^T ZB)$$

It follows that just as in the proof of Theorem 1 that

$$R(B)^2 \leq \inf_{Z \in \mathbb{S}^n_+} \text{tr}(B^T ZB) \quad \text{where} \quad \psi(Z) \leq \frac{1}{4}.$$  \hspace{1cm} □
C  Proof of Section 3

Proof of Proposition 3. Let
\[ G(u, v) \triangleq \frac{1}{2} |u|^2 + \frac{1}{2} |v|^2 + \frac{1}{2\lambda} |X(v \odot_G u) - y|^2. \]

We know from Theorem 2 that \( f \) is differentiable with
\[ \nabla f(v) = \partial_u G(u, v) = v + \lambda^{-1} u \odot X^\top (X v \odot_G u - y) \]
where \( u = \arg\min_u G(u, v) \). In particular,
\[ 0 = \partial_u G(u, v) = u + \lambda^{-1} X^\top (X v \odot_G u - y). \]
Since \( \partial_{uu} G = \lambda^{-1} (v g X^\top g X_h v_h)_{g, h} + \text{Id} \) is invertible, by the implicit function theorem \( u \) is a smooth function of \( v \) with \( \partial_u u = [\partial_{uu} G]^{-1} \partial_{uv} G \). In particular,
\[ \nabla^2 f(v) = \partial_{vv} G(u, v) + \partial_{uv} G(u, v) \partial_v u. \]

So, the Hessian of \( f \) is the Schur complement of the Hessian of \( G \) (as also observed in Ruhe and Wedin [1980], van Leeuwen and Aravkin [2016]). We write \( \nabla^2 G = \begin{pmatrix} A & B \\ B^\top & D \end{pmatrix} \) where
\[
A \triangleq \partial_{vv} G = \lambda^{-1} (u g X^\top g X_h v_h)_{g, h} + \text{Id}
B \triangleq \partial_{uv} G = \lambda^{-1} ((u g X^\top g X_h v_h)_{g, h} + \text{diag}(\xi_g))
D \triangleq \partial_{uu} G = \lambda^{-1} (v g X^\top g X_h v_h)_{g, h} + \text{Id}
\]
where \( \xi = \frac{1}{\lambda} X^\top (X (u \odot v) - y) \). Then, \( \nabla^2 f(t) = A - B D^{-1} B^\top \). Note that in fact, \( u \) is infinitely differentiable by the implicit function theorem, and so, \( f \) is also infinitely differentiable.

We now derive a formula for the Hessian of \( f \). By permuting the rows and columns of \( \nabla^2 G \), we can assume that, letting \( J \) denote the support of \( v \),
\[
A \triangleq \begin{pmatrix}
\lambda^{-1} (u g X^\top g X_h v_h)_{g, h} + \text{Id}_J & 0 \\
0 & \text{Id}_{J^c}
\end{pmatrix}
B \triangleq \lambda^{-1} ((u g X^\top g X_h v_h)_{g, h} + \text{diag}(\xi_g))
D \triangleq \begin{pmatrix}
\lambda^{-1} (v g X^\top g X_h v_h)_{g, h} + \text{Id}_J & 0 \\
0 & \text{Id}_{J^c}
\end{pmatrix}
\]
Note that \( A \) and \( D \) is positive definite. So, \( \nabla^2 G \) is positive semidefinite if and only if \( A - B D^{-1} B^\top \) is positive semidefinite. Note that \( A - B D^{-1} B^\top \) is a block diagonal matrix, with the bottom right block as \( \text{Id}_{J^c} - \text{diag}(\{\xi_g|_{J^c}\}) \).

To work out the expression for the top left block of \( A - B D^{-1} B^\top \), let us first examine the top left block of the matrix \( B \): Note that by definition of \( u \),
\[ \lambda^{-1} v g X^\top g (X (u \odot v) - y) + u_g = 0 \implies \forall g \in \text{Supp}(v), \xi_g = -u_g/v_g. \]
Define the block diagonal matrix \( U_{u/v} = \text{diag}(u_g/v_g)_{g \in J} \), then \( U_{u/v}^\top U_{u,v} = \text{diag}(u_g^2/v_g^2)_{g \in J} \).
The top left block of \( B \) is
\[
= u_{u/v}^\top \begin{pmatrix}
\lambda^{-1} (v g X^\top g X_h v_h)_{g, h} + \text{diag}(\xi_g) & 0 \\
0 & \text{Id}_{J^c}
\end{pmatrix}
- 2 U_{u/v}^\top W - 2 U_{u/v}^\top.
\]
Therefore, the top left block of \( D^{-1} \) is \( H^{-1} \). So, the top left block of \( B D^{-1} B^\top \) is
\[ (U_{u/v}^\top H - 2 U_{u/v}^\top)(U_{u/v} - 2 H^{-1} U_{u/v}) = U_{u/v}^\top H U_{u/v} - 4 U_{u/v}^\top U_{u/v} + 4 U_{u/v}^\top H^{-1} U_{u/v}. \]
where \( \varepsilon \)

Therefore, the smallest eigenvalue of \( A \) is \( \max \langle \nabla^2 f(v_0), v_0 \rangle \). Note that \( \max_{g \in G} \| \varepsilon_g \|^2 \). We have a global Lipschitz bound on the gradient of \( f \) if \( \| \varepsilon_g \| \leq L \) for some \( L \), which is true because for each \( v, u \) minimises

\[
\min_u \frac{1}{2} \| u \|^2 + \frac{1}{2\lambda} \| X(v \odot g) u - y \|^2 \leq \frac{\| y \|^2}{2\lambda}
\]

So, \( \max_{g \in G} \| \varepsilon_g \|^2 \leq \min_{g \in G} \| X_g \|^2 / \lambda^2 \).

At stationary points, we also have

\[
u_g^T X_g^T (X(v \odot g) u - y) + \lambda v_g = 0 \implies \forall g \in \text{Supp}(v), \ u_g^T \varepsilon_g = -v_g.
\]

Together, this means that at stationary points, \( \| u_g \|^2 = v_g^2 \) and \( U_{u/v} U_{u/v} = \text{Id}_J \). Therefore, the top left block of \( A - BD^{-1}B^T \) becomes

\[
4\text{Id}_J - 4U_{u/v} \left( \lambda^{-1}(v_g X_g^T X_h v_h)_{g,h \in J} + \text{Id}_J \right)^{-1} U_{u/v} \succeq 0
\]

since \( \lambda^{-1}(v_g X_g^T X_h v_h)_{g,h \in J} + \text{Id}_J \succeq (1 + \mu)\text{Id} \), where \( \mu = \min \text{Eig} \left( \lambda^{-1}(v_g X_g^T X_h v_h)_{g,h \in J} \right) \). Therefore, the smallest eigenvalue of \( A - BD^{-1}B^T \) is at least

\[
\lambda \min \left( 4\mu / (1 + \mu), \min_{g \notin J} \left( 1 - \| \varepsilon_g \|^2 \right) \right) \geq \min_{g \notin J} \left( 1 - \| \varepsilon_g \|^2 \right)
\]

Moreover, if \( A - BD^{-1}B^T \succeq 0 \), then \( \min_{g \notin J} \left( 1 - \| \varepsilon_g \|^2 \right) \geq 0 \), which implies that \( (u, v) \) defines a minimiser to the original group Lasso problem, hence, \( (u, v) \) defines a global minimum. Therefore, every stationary point is either a global minimum or a strict saddle point.

\[\square\]

Remarks on the comparison with ISTA in the introduction To explain the observed behavior, note that gradient descent for \( f \) with stepsize \( \gamma \) reads \( v_{k+1} = v_k - \gamma \nabla f(v_k) = v_k - \gamma (1 - \| \varepsilon_k \|^2) \) where \( \varepsilon_k \triangleq \frac{1}{\lambda} X^T (X v_k \odot u_k - y) \) (see Proposition 3). Note that if \( \beta_\ast \) is a minimiser, then \( \varepsilon_\ast \triangleq \frac{1}{\lambda} X^T (X \beta_\ast - y) \) satisfies \( \| \varepsilon_\ast \|_\infty \leq 1 \) and the set \( \{ i \mid \| (\varepsilon_\ast)_i \| = 1 \} \) is often called the extended support and contains the support of \( \beta_\ast \). It is clear that we can expect coefficients outside the extended support to (eventually) decay to 0 geometrically. Since \( \varepsilon_k \) is uniformly bounded (see Proposition 2), for \( \gamma \) sufficiently small, \( v_k \) never changes sign and any sign change in the iterate \( \beta_k \triangleq v_k \odot u_k \) is due to \( u_k \). In contrast, the ISTA dynamics is \( \beta_{k+1} = \text{sign}(\beta_k - \gamma \varepsilon_k) \max \left( \| \beta_k - \gamma \varepsilon_k \| - \gamma, 0 \right) \). Due to the thresholding operation, a coefficient of \( \beta_k \) is initialised with the wrong sign will spend some iterations as 0 before correcting its sign.

D Supplementary to Section 4

D.1 Remarks on numerical experiments

Initialisation points We generated random initialisation point from the normal distribution. In our experiments, methods which are not reparameterized (e.g. the proximal methods), are given the same random initial point, while reparameterized methods have their own random initialisation, since some of these require positive starting points and some need double the number of variables. We find that the comparisons are not much affected by the choice of initial points.
Inversion of linear systems As mentioned in Corollary (1), for the Lasso, when computing the gradient of $f$, one can either invert a $n \times n$ linear system or an $m \times m$ linear system. The same applies to Quad-variational, since the solution to the inner maximisation problem is, by the Woodbury identity,

$$\alpha = (\lambda \text{Id}_m + X_\eta X_\eta^T)^{-1} y = \frac{1}{\lambda} y - \frac{1}{\lambda} X_\eta (\lambda \text{Id}_n + X_\eta^T X_\eta)^{-1} (X_\eta^T y)$$

where $X_\eta = X \text{ diag}(\sqrt{\eta})$, with the correspondence that $\beta = \eta \odot X^T \alpha$. Throughout, we simply use backslash in MATLAB for the matrix inversion.

Implementation details All numerics are done in Matlab with the exception of CELER which is in Python:

- CELER are conducted in Python and we used the code [https://mathurinm.github.io/celer/](https://mathurinm.github.io/celer/) provided by the original paper [Massias et al. 2018]
- 0-mem SR1, FISTA w/ BB and SPG/SpaRSA use the Matlab code from [https://github.com/stephenbeckr/zeroSR1](https://github.com/stephenbeckr/zeroSR1) of the paper [Becker et al. 2019].
- Interior point method uses the Matlab code [https://web.stanford.edu/~boyd/l1_ls/](https://web.stanford.edu/~boyd/l1_ls/) of [Koh et al. 2007].
- We had our own implementation of Non-cvx-Alternating-min and IRLS.
- Quad-variational, Non-cvx-LBFGS and Noncvx-Pro are written in Matlab using the L-BFGS-B solver from [https://github.com/stephenbeckr/L-BFGS-B-C](https://github.com/stephenbeckr/L-BFGS-B-C) which is a Matlab wrapper for C code converted from the well known Fortran implementation of [Byrd et al. 1995].

D.2 Additional examples

Lasso In Figure 2 we show additional numerics for the Lasso, testing against datasets from the Libsvm repository. The regularisation parameter $\lambda$ associated to each plots is found by cross validation on the mean squared error.

Group Lasso In Figure 3 we show additional numerics for the multitask Lasso setup described in Section 4.2. We test on two synthetic datasets of size $(m, n, q) = (300, 1000, 100)$ with 5 relevant features and $(m, n, q) = (50, 1200, 20)$ with 10 relevant features. The data matrix $X$ has entries drawn from a normal distribution. We also test on a MEG/EEG dataset with $(m, n, q) = (305, 22494, 85)$ from the MNE repository [https://mne.tools/0.11/manual/datasets_index.html](https://mne.tools/0.11/manual/datasets_index.html). We display convergence plots for different regularisation parameters.

Trace norm In Figure 1 we show additional numerics for the multifeature learning setup described in Section 4.3. The data matrices $X_t$ has entries drawn uniformly at random from $[0, 1]$. We consider different number of tasks $T$ tasks, $n = 30$ features and $m = 10,000$ samples in total (the samples are split at random across the different tasks).
Figure 2: Comparisons of Lasso with different regularisation parameters on datasets from Libsvm. The first column shows the optimal regularisation parameter $\lambda_*$ found by cross validation. The second, third and fourth columns correspond to different fractions of $\lambda_{\text{max}} = \|X^\top y\|_{\infty}$ which is the parameter for which the Lasso solution is identically zero. The smaller this fraction, the less sparse the solution.
Figure 3: Comparisons for multitask Lasso at different regularisation strengths. The problem sizes \((m, n, q)\) are displayed on the left of each row. The top two rows are synthetic datasets generated by random Gaussian variables with 5 and 10 active features respectively. The last row corresponds to a MEG/EEG dataset from the MNE website.

E  Douglas-Rachford and Primal-Dual Algorithms

We consider the resolution of a constrained group Lasso problem

\[
\min_{X\beta = y} \|\beta\|_{1,2} = \sum_g \|\beta_g\|_2
\]

which we write as the minimization of either \(F(\beta) + G(\beta)\) (for DR) or \(F(\beta) + G_0(X\beta)\) where \(F = \|\cdot\|_{1,2}, G = \iota_C\) where the constraint set is \(C = \{\beta \mid X\beta = y\}\) and \(G_0 = \iota_{\{y\}}\). Here \(\iota_C\) is the convex indicator function of a closed convex set \(C\).

DR and PD are generic algorithm to solve minimization of function of the form \(F + G\) and \(F + G_0 \circ X\) when one is able to compute efficiently the so-called proximal operator of the involved functionals, where the proximal operator of some convex function \(H\) and some step size \(\tau \geq 0\) is

\[
\text{Prox}_\tau H(\beta) \triangleq \arg\min_{\beta'} \frac{1}{2}\|\beta - \beta'\|_2^2 + H(\beta').
\]

In our special case, one has

\[
\text{Prox}_\tau F(\beta) = \left(\max(\|\beta_g\| - \tau, 0) \frac{\beta_g}{\|\beta_g\|_g}\right)_g, \quad \text{Prox}_\tau G(\beta) = \beta + X^T(XX^T)^{-1}(y - X\beta),
\]

and \(\text{Prox}_{\tau G_0}(\beta) = y\).

**DR algorithm.** We denoted the reflected proximal map as \(r\text{Prox}_\tau H(\beta) = 2 \text{Prox}_\tau H(\beta) - \beta\). For some step size \(\mu > 0\) and weight \(0 < \gamma < 2\) (which is set to \(\gamma = 1\) in our experiments), the iterates \((\beta_k)_k\) of DR are \(\beta_k \triangleq \text{Prox}_\mu G(z_k)\) where \(z_k\) satisfies

\[
z_{k+1} = (1 - \frac{\gamma}{2})z_k + \frac{\gamma}{2} r\text{Prox}_\mu F(r\text{Prox}_\mu G(z_k)).
\]

**PD algorithm.** Denoting \(G_0^*(u) = \sup_{\beta}(\beta, u) - G_0(\beta)\) the Legendre transform of \(G_0\), the PD iterations read

\[
w_{k+1} = \text{Prox}_{\sigma G_0^*}(w_k + \sigma X(\tilde{\beta}_k))
\]

\[
\beta_{k+1} = \text{Prox}_F(\beta_k - \tau K^T(w_{k+1}))
\]

\[
\tilde{\beta}_{k+1} = \beta_{k+1} + \theta(\beta_{k+1} - \beta_k).
\]
In our case, one has $G_G^\theta(u) = \langle u, y \rangle$ so that $\text{Prox}_{\tau G_G^\theta}(u) = u - \tau y$. Convergence of the PD algorithm is ensure as long as $\tau \sigma |X|^2 < 1$ where $|X|$ is the operator norm, and $0 < \theta \leq 1$ (we use $\theta = 1$ in the numerical simulation). In our numerical simulation, we set $\tau \sigma |X|^2 = 0.9$ and tuned the value of the parameter $\sigma$.

F Non-convex optimisation with $\ell_q$ quasi-norms

As mentioned, for $q \in (0, 2)$, $R(\beta) \triangleq \|\beta\|_q^q = \sum_j |\beta_j|^q$ has a quadratic variational form. In the case where $q > 2/3$, we have the following bilevel smooth formulation:

**Corollary 1.** When $q > 2/3$, $(P_X)$ is equivalent to

$$
\inf_{v \in \mathbb{R}^n} f(v) \triangleq \inf_{u \in \mathbb{R}^n} \frac{1}{2} |u|^2 + C_q \sum_{j=1}^n |\beta_j|^q + \frac{1}{\lambda} L(X(u \circ v), y) \tag{3}
$$

where $C_q = (2 - q)q^{q/(2-q)}$. The function $f$ is differentiable function provided that $q > 2/3$. Its gradient can be computed as in Theorem 2.

**Remark 1 (Existing approaches).** Existing approaches to $\ell_q$ minimisation are typically iterative thresholding/proximal algorithms [Bredies et al., 2015, IRLS Chartrand and Staneva, 2008, Daubechies et al., 2004] or iterative reweighted $\ell_1$ algorithms Foucart and Lai [2009]. Iterative thresholding algorithms are applicable only for the case where the loss function is differentiable, and hence not applicable for Basis pursuit problems which we describe below. Moreover, computation of the proximal operation requires solving a nonlinear equation. For iterative reweighted algorithms, they require gradually decreasing an additional regularisation parameter $\varepsilon > 0$. This can be problematic in practice and for finite $\varepsilon$, one does not solve the original optimisation problem.

**Remark 2.** Since we have a differentiable unconstrained problem, the problem (3) can be handled using descent algorithms and convergence analysis is standard. For example, since $f$ is coercive, for any descent algorithm applied to $f$, we can assume that the generated sequence $v_k$ is uniformly bounded and $\nabla f(v_k)$ is also uniformly bounded. So, by applying standard results [Bertsekas, 1997, Proposition 1.2.1], we can conclude that all limit points of sequences $v_k$ generated by descent methods under line search on the stepsize are stationary points. In fact, since we have an unconstrained minimisation problem with a continuously differentiable $f$ which is also semialgebraic (for rational $q$) and hence satisfy the KL inequality [Attouch et al., 2013], convergence of the full sequence by descent methods with line search can be guaranteed [Noll and Rondepierre, 2013].

F.1 Basis pursuit

In this section, we focus on the basis pursuit problem with $q \in (2/3, 1)$,

$$
\min_{\beta} \|\beta\|^q \quad \text{where} \quad X \beta = y.
$$

The set of local minimums are all $\beta$ for which $X \beta = y$ and there exists $\alpha$ such that $(X^T \alpha)_i = q|\beta_i|^{q-1} \text{sign}(\beta_i)$ on the support of $\beta$. When $q > 2/3$ and $f$ is differentiable with

$$
\nabla f(v) = q^{2/q} |\gamma|^{-1} \text{sign}(v) - v \circ |X^T \alpha|^2, \quad \text{where} \quad \gamma \triangleq 2q/(2-q) > 1,
$$

At a stationary point $v$, letting $\beta = -v^2 \circ X^T \alpha$, we have $X \beta = y$ and $\nabla f(v) = 0$ implies that on the support of $v$, $q|v|^2 = |\beta|^{2-q}$ and so,

$$
X^T \alpha = -v^{-2} \beta = -q \text{sign}(\beta) \circ |\beta|^{q-1},
$$

which is precisely the optimality condition of the original problem.

**Illustrations for Basis pursuit** In Figure 4, we show that gradient descent dynamics for $f$ in the case of the indicator function $L(\cdot, y) = \ell_{(y)}$ and a random Gaussian matrix $X \in \mathbb{R}^{10 \times 20}$, that is

$$
v_{k+1} = v_k - \tau \nabla f(v_k) = v_k - \tau \left( q^{2/q} |v_k|^{2q-2} \circ \text{sign}(v) - v_k \circ |X^T \alpha_k|^2 \right)
$$

where

$$
X \text{diag}(v_k \circ v_k)X^T \alpha_k = -y.
$$
Figure 4: Evolution of 20 coefficients for Basis pursuit with $\ell_q$ regularisation. The same stepsize $\tau$ is used for all plots. Top row show the evolution of $\beta_k$ and the bottom row show the evolution of $v_k$.

Figure 5: Number of successful recovery by $\ell_q$ minimisation.

Observe that as $q \to 2/3$, the evolution paths of $v_k$ becomes increasingly linear.

In Figure 5 we follow the experiment setup of Chartrand and Staneva [2008] and generate 100 problem instances $(\bar{X}, \bar{y}, \bar{\beta})$. Each problem instance consist of a matrix $X \in \mathbb{R}^{m \times n}$ with $m = 140$ rows and $n = 256$ columns whose entries are identical independent distributed Gaussian random variable with mean 0 and variance, a vector $\beta$ of size $n$ with $K = 40$ entries uniformly distributed on $\{1, \ldots, n\}$ and whose nonzero entries are iid Gaussian with mean 0 and variance 1 and $\bar{y} \triangleq \bar{X}\bar{\beta}$. For each problem, we carry out the following procedure. For each $m \in \{60, \ldots, 140\} \cap 2\mathbb{N}$, we let $X$ be the matrix from the first $m$ rows of $\bar{X}$, and $y$ be the first $m$ entries of $\bar{y}$. We then compute $\beta$ by minimising $f$ for this $X$ and $y$ using BFGS with 10 randomly generated starting points and declare “success” if $\|\beta - \bar{\beta}\|_2 \leq 10^{-3}$ for one of these starting points.

References


