STATS 606 Report:

A proximal alternating direction method for regularized multi-task regression

SIMON FONTAINE [1,*] , JINMING LI [1,**] and YANG LI [1,†]

[1] *University of Michigan, Department of Statistics. West Hall, 1085 South University, Ann Arbor, MI, U.S.A., 48109.*
E-mail: [*] simfont@umich.edu; [**] lijinmin@umich.edu; [†] yangly@umich.edu

Abstract. We propose an alternating direction of multipliers method (ADMM) based on the consensus scheme for solving regularized multi-task problems where the loss function is separable across tasks. We discuss the convergence of the algorithm and survey its statistical properties in the sparse group Lasso case. Through numerical experiments, we find that our proposed method can be particularly efficient in the number of proximal evaluations required to achieve convergence when compared to a more general method such as FISTA.

1. Introduction

Suppose we have a collection $\mathcal{D} = \cup_{k=1}^{K} \mathcal{D}_k$ of K datasets $\mathcal{D}_k = \{(X_i^{(k)}, Y_i^{(k)})\}_{i=1}^{n_k} \subset \mathbb{R}^p \times \mathbb{R}$ and wish to model the relationship between the features $X_i^{(k)}$ and the response $Y_i^{(k)}$ by minimizing the following objective function,

$$\mathcal{F}(\boldsymbol{\beta} \mid \mathcal{D}) = \mathcal{L}(\boldsymbol{\beta} \mid \mathcal{D}) + \lambda \mathcal{P}(\boldsymbol{\beta}) \tag{1.1}$$

where $\mathcal{L}$ is a loss function of the parameters $\boldsymbol{\beta} \in \mathbb{R}^{p \times K}$ that is L-strongly smooth, in the sense $\nabla^2 \mathcal{L}(\boldsymbol{\beta} \mid \mathcal{D}) \preceq L I_{Kp}$ uniformly in $\boldsymbol{\beta}$, $\lambda \geqslant 0$ is a tuning parameter, and $\mathcal{P} \geqslant 0$ is a convex regularization term.

The structure of the problem assumes that all K datasets share the same number of features p: in particular, this *multi-task* setting is often used when the K datasets share the same p features. In that case, we model the relationship of these p features to K different types of outcomes. We identify two common cases where such structure exists: multiple outcomes are measured on the same observational unit (the *multivariate* setting) and a single outcome is measured for each unit, but the same features are available.

When the K responses are related, we expect the parameter of interest $\boldsymbol{\beta}$ so exhibit some interesting structure. A common choice of regularization $\mathcal{P}$ inducing sparsity at the feature level as well as at the within-feature level is the *sparse group Lasso* penalty (Simon et al., 2013), $\mathcal{P}_{\alpha,q}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \alpha \|\beta_j\|_1 + (1-\alpha)\|\beta_j\|_q$, where $\alpha \in [0,1]$ is a tuning parameter controlling the mixing of the group and individual sparsity, and where q is often chosen to be 2 of ∞.

Now, most methods of fitting such models do not fully exploit the separability structure of $\mathcal{L}$ or of $\mathcal{P}$ and optimize naively the original problem (1.1). However, in most cases, such separability exists so we can apply an *alternating direction of multipliers method* (ADMM) to take advantage of it. For example, the loss function may be separable across tasks or may depend only on linear predictors $\eta^{(k)} = \mathbf{X}^{(k)}\,\beta^{(k)}$; the penalty is often separable across features (e.g., the sparse group Lasso).

2. Algorithm

Consensus ADMM scheme Suppose the loss function $\mathcal{L}$ is separable across the K tasks, in the sense that it can be split in a sum of losses depending only on $\beta^{(k)}$, $k \in [K]$, that is, $\mathcal{L}(\boldsymbol{\beta} \mid \mathcal{D}) = \sum_{k=1}^{K} \mathcal{L}^{(k)}(\beta^{(k)} \mid \mathcal{D})$. In that case, we consider the following augmented problem, reminiscent of the *consensus* ADMM (Boyd et al., 2011, Section 7.1.1), where we introduce copies of $\beta^{(k)}$ for each loss,

$$\begin{aligned} \underset{\boldsymbol{\beta},\mathbf{B}}{\text{minimize}} \quad & \sum_{k=1}^{K} \mathcal{L}^{(k)}(\mathbf{B}^{(k)} \mid \mathcal{D}) + \lambda\,\mathcal{P}\,(\boldsymbol{\beta}) \\ \text{subject to} \quad & \mathbf{B} - \boldsymbol{\beta} = 0, \end{aligned}$$

where $\mathbf{B} \in \mathbb{R}^{p\times K}$. The corresponding (scaled) augmented Lagrangian is then given by

$$L\,(\boldsymbol{\beta}, \mathbf{B}, \mathbf{D}) = \sum_{k=1}^{K} \mathcal{L}^{(k)}(\mathbf{B}^{(k)} \mid \mathcal{D}) + \lambda\,\mathcal{P}\,(\boldsymbol{\beta}) + \rho\,\langle \mathbf{D}, \mathbf{R}\rangle + \frac{\rho}{2}\,\langle \mathbf{R}, \mathbf{R}\rangle\,,$$

where $\mathbf{D} \in \mathbb{R}^{p\times K}$ is a Lagrange multiplier, $\mathbf{R} = \mathbf{B} - \boldsymbol{\beta}$ and $\rho > 0$ is the ADMM regularization parameter. The sequential updates are given by

$$\mathbf{B}^{(k)}[t+1] = \underset{\mathbf{B}^{(k)}}{\arg\min}\, \mathcal{L}^{(k)}(\mathbf{B}^{(k)} \mid \mathcal{D}) + \frac{\rho}{2}\left\|\mathbf{B}^{(k)} - (\boldsymbol{\beta}^{(k)}[t] - \mathbf{D}^{(k)}[t])\right\|_2^2, \quad k \in [K], \tag{2.1}$$

$$\boldsymbol{\beta}[t+1] = \mathbf{prox}_{\lambda\,\mathcal{P}\,/\rho}\,(\mathbf{B}[t+1] + \mathbf{D}[t])\,, \tag{2.2}$$

$$\mathbf{D}[t+1] = \mathbf{D}[t] + \mathbf{R}[t+1]. \tag{2.3}$$

We note that the $\mathbf{B}^{(k)}$ updates are ℓ_2-regularized problems and the $\boldsymbol{\beta}$ update is a simple proximal operator. This scheme can be particularly efficient in the case that $\mathcal{L}^{(k)}$ depends only on $\mathcal{D}^{(k)}$, i.e., $\mathcal{L}^{(k)}(\mathbf{B}^{(k)} \mid \mathcal{D}) = \mathcal{L}^{(k)}(\mathbf{B}^{(k)} \mid \mathcal{D}^{(k)})$: indeed, the optimization can be easily parallelized and distributed. If the loss function is not separable across $k \in [K]$, the same procedure can be performed, but the update (2.1) can no longer be split into K optimization problems.

Alternative ADMM schemes We also consider algorithms using the *sharing* ADMM scheme where, instead of copying the matrix of coefficients $\boldsymbol{\beta}$ into $\mathbf{B}$, we clone linear

predictors. The alternating steps therefore correspond to minimizing the loss w.r.t. linear predictors and finding the coefficients best approximating the linear predictors. In particular, three types of sharing scheme can be defined: the first, following Boyd et al. (2011, Section 7.3), introduces partial linear predictors $\eta_j^{(k)} = X_j^{(k)}\beta_j^{(k)}$, $j \in [p]$, $k \in [K]$; the second acts on the linear predictors $\eta^{(k)} = X^{(k)}\beta^{(k)}$, $k \in [K]$; the third, following Xiao, Wu and He (2012), introduces residuals $R^{(k)} = Y^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)}$ and therefore only applies to losses depending on residuals only (e.g., least squares). Details of these algorithms are given in Appendices A.2 to A.4, respectively. We find that the sharing schemes do not perform as well as the consensus scheme for the regularized multi-task problem and we thus only consider the consensus method in the main text.

Solving the subproblems In our proposed schemes, we identify two types of subproblem we need to solve. First, update (2.2) is an explicit proximal operator and can be solved exactly for many choices of regularizations (see, e.g., Combettes and Pesquet, 2011). For sparse group Lasso penalty, we have analytical solutions for $q = 2$ as well as $q = \infty$ (Jenatton et al., 2011); solutions for general $q \geqslant 1$ can be found in Liu and Ye (2010). Second, updates (2.1) corresponds to a ridge-regularized differentiable loss and can be solved using first-order methods such as Nesterov's (Nesterov, 1983).

3. Theoretical Properties

Convergence We explore the literature and prove that our optimization algorithm will indeed converge to the optimal. Eckstein and Bertsekas (1992, Theorem 8), whose details are given in Appendix A.5, has discussed convergence properties of inexact ADMM. As a consequence, we only need to make sure each sub-problem has been well optimized.

In the case of consensus ADMM, the algorithm is guaranteed to reach optimal value so long as

$$\left\| \mathbf{B}[t+1] - \arg\min_{\mathbf{B}} L\left(\boldsymbol{\beta}[t], \mathbf{B}, \mathbf{D}[t]\right) \right\|_2 \leq \mu_t,$$

for a summable sequence of thresholds $\sum \mu_t < \infty$. Indeed, the updates of $\boldsymbol{\beta}$ in (2.1) are an explicit proximal operator and an exact solution is generally available for most regularization $\mathcal{P}$. Now, the ridge problem (2.1) is solved using an accelerated first-order method which guarantees convergence provided the loss function $\mathcal{L}$ is L-strongly smooth. Furthermore, the quadratic term introduced by the ADMM renders the objective of (2.1) to be ρ-strongly convex, which generally helps convergence.

Statistical Properties Here we discuss some statistical properties of sparse group lasso problem with literature in regularized M-estimators. Negahban et al. (2012) and Lee, Sun and Taylor (2013) introduced unified analysis framework for sparse models with decomposable regularizers. Decomposability of penalty w.r.t. model sparsity structure plays an important role here, and many problems fit into such framework: sparse

linear model with ℓ_1 penalty; or group-structured sparse model with $\ell_{1,q}$ penalty. Both consistency and convergence rates are given in Negahban et al. (2012).

Sparse group lasso penalty, however, does not directly fit into the framework aforementioned due to its corresponding complex sparsity structure. It is decomposable under hierarchical structure: namely, $\beta_j^{(k)} = 0$ if $||\beta^{(k)}||_1 = 0$. To this end, sparse group lasso penalty is decomposable w.r.t. group sparsity structure, just like group lasso. In Chatterjee et al. (2012) the authors followed decomposable framework and showed consistency results w.r.t group sparsity structure. But this is not satisfactory, since sparse group lasso is originally proposed to reveal in-group sparsity structure.

Another line of works try to overcome this shortage by assuming additive structure of parameters with different types of sparsity structures. For example, recently authors of Yang and Lozano (2017) considered

$$\underset{\Theta}{\text{minimize}}\, \mathcal{L}(\Theta) + ||\Theta||_\lambda,$$

where $||\Theta||_\lambda = \inf_{\alpha+\beta=\Theta} (\lambda_1||\alpha||_1 + \lambda_2||\beta||_{1,q})$. This is closely related to sparse group lasso problem, but still not the same.

4. Implementation

We implement both proposed ADMM schemes, along with an accelerated proximal gradient method (FISTA), for solving arbitrary regularized losses in a `Python` package. [1] The only requirements are (1) that the loss and its gradient w.r.t. $\boldsymbol{\beta}$ are readily available, (2) the loss has a known and computable Hessian upper bound and (3) the regularization is convex and has a computable proximal operator.

The least squares and multi-label logistic regression losses are implemented as well as the sparse group Lasso regularization. Furthermore, the solution path (along a sequence of λ values) is implemented and it is possible to initiate it automatically to the smallest λ value yielding a null model (by studying the KKT conditions with $\boldsymbol{\beta} = \mathbf{0}$). Finally, we use the warm-start trick to speed-up the optimization.

5. Numerical experiments

In order to assess the performance of the consensus algorithm, we procede with two numerical experiments comparing the results to optimization using Consensus ADMM to FISTA. To put both algorithms on equal footings, we opt to compare convergence w.r.t. the number of gradients computed as well as to the number of proximal operations performed—the two computational bottlenecks in both algorithms. Indeed, the raw number of iterations would put ADMM at an advantage since more work is done at each step. Also, we refrain from commenting on computing time since it would depend on

[1] Available at `https://github.com/fontaine618/MTSGL`.

the implementation. We note that the K ridge problems solved by ADMM can be easily parallelized so the number of gradients computed—as well as computing time—should be divided across workers; yet, we report the serial statistics to produce a fair comparison.

Multi-task Least Squares Regression In this first experiment, we consider a multi-task least squares problem with $n = 3000$ observations split into $K = 5$ tasks each having $p = 100$ features; the true $\boldsymbol{\beta}$ is generated to be sparse across features as well as within features (across tasks). The full solution path is computed starting from the smallest λ value such that $\boldsymbol{\beta} = \mathbf{0}$ and ending at 1/100th of that value. Similar convergence criterion are used for both algorithm. Furthermore, we proceed at optimizing the problem for $\lambda = 0.3$ starting from $\boldsymbol{\beta} = \mathbf{0}$.

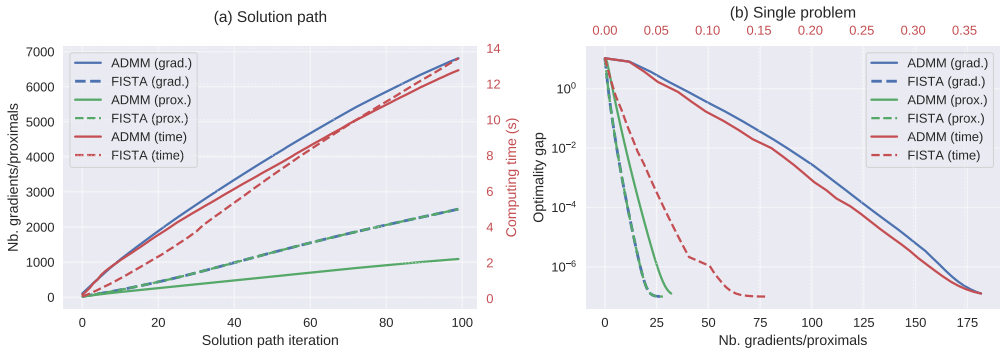

Figure 1: *Multi-task least squares regression experiment results. Number of gradient computed (blue) and number of proximal operations (green) and computing time (red) by the consensus ADMM (solid) and the FISTA (dashed) algorithms.*

Results for the solution path, contained in Figure 1(a)[2], show that consensus ADMM requires significantly less proximals to achieve the same convergence as FISTA summed over all solution path iterations. Results for a single problem, in Figure 1(b), indicate that FISTA achieves faster convergence in terms of gradients and similar convergence in terms of proximals when the problem is initialized at $\boldsymbol{\beta} = \mathbf{0}$.

Multi-label Classification In this second experiment, we consider a multi-label classification problem with $n = 1000$ observations split into $K = 50$ tasks each having $p = 50$ features. This setting is therefore of the high-dimensional case $p > n$ since there are $Kp = 2500$ parameters, but only a fraction of those are truly non-zero. The solution path again starts from the smallest λ with $\boldsymbol{\beta} = \mathbf{0}$ until 1/20th of that value. Furthermore, we proceed at optimizing the problem for $\lambda = 0.02$ starting from $\boldsymbol{\beta} = \mathbf{0}$.

Results for the solution path, contained in Figure 2(a), show that consensus ADMM again requires much fewer proximal evaluations than FISTA to yield similar convergence

[2]Code for all results can be found at `https://github.com/fontaine618/MTSGL` under `experiments`.

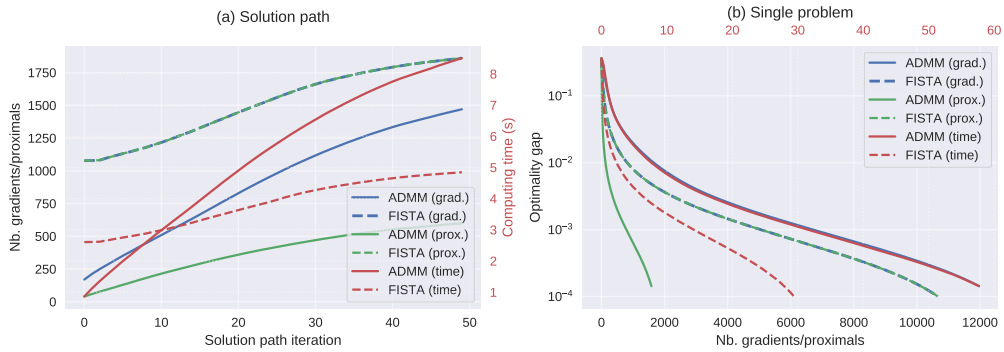

Figure 2: *Multi-label classification experiment results. Number of gradient computed (blue) and number of proximal operations (green) and computing time (red) by the consensus ADMM (solid) and the FISTA (dashed) algorithms.*

when the whole path is considered. The improvement seems to emerge mostly in the first few iterations. Results for a single problem, in Figure 2(b), show that consensus ADMM achieves faster convergence in terms of proximals, but similar to FISTA in terms of gradients.

6. Discussion

We proposed an ADMM scheme for solving regularized multi-task regression problems relying on the separability of the loss function across K task. Alternating between K ridge problems and a single proximal operator, our algorithm is easily parallelizable, as the data is only used within each K ridge updates, with minimal communication (only $\boldsymbol{\beta}^{(k)}$, $\mathbf{B}^{(k)}$ and $\mathbf{D}^{(k)}$ have to be exchanged).

Simulation studies show that this consensus ADMM algorithm is particularly economical in the number of proximal operations performed compared to a general method such as FISTA, especially in the high-dimensional case; this observation can inform on the applicability of the method. Indeed, problems with large K or with expensive proximal may greatly benefit from such a scheme. For example, the sparse group Lasso with $q = \infty$ requires a ℓ_1-projection which is typically done through sorting in $\mathcal{O}(K \log K)$[3] and an algorithm performing as few proximal evaluation as possible should be preferred.

Our proposed method appears greedier than FISTA in terms of gradients computed particularly when optimizing a single problem. Improved tuning of the threshold sequence and of the quadratic penalty parameter ρ can potentially fix these issues. When a whole solution path is considered, we observe more encouraging results: the warm-start trick seems really beneficial for fast convergence of the consensus ADMM.

[3] Other methods (e.g., Duchi et al., 2008) have *expected* complexity $\mathcal{O}(K)$ but worst-case $\mathcal{O}(K^2)$.

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning* **3** 1–122.

Chatterjee, S., Steinhaeuser, K., Banerjee, A., Chatterjee, S. and Ganguly, A. (2012). Sparse group lasso: Consistency and climate applications. In *Proceedings of the 2012 SIAM International Conference on Data Mining* 47–58. SIAM.

Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering. Springer Optim. Appl.* **49** 185–212. Springer, New York. MR2858838

Duchi, J., Shalev-Shwartz, S., Singer, Y. and Chandra, T. (2008). Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning* 272–279.

Eckstein, J. and Bertsekas, D. P. (1992). On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Programming* **55** 293–318. MR1168183

Jenatton, R., Mairal, J., Obozinski, G. and Bach, F. (2011). Proximal methods for hierarchical sparse coding. *J. Mach. Learn. Res.* **12** 2297–2334. MR2825428

Lee, J. D., Sun, Y. and Taylor, J. E. (2013). On model selection consistency of penalized m-estimators: a geometric theory. In *Advances in Neural Information Processing Systems* 342–350.

Liu, J. and Ye, J. (2010). Efficient l1/lq norm regularization. *arXiv preprint arXiv:1009.4766*.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133

Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence O (1/kˆ 2). In *Doklady an ussr* **269** 543–547.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. MR3173712

Xiao, Y., Wu, S.-Y. and He, B.-S. (2012). A proximal alternating direction method for $\ell_{2,1}$-norm least squares problem in multi-task feature learning. *J. Ind. Manag. Optim.* **8** 1057–1069. MR2981693

Yang, E. and Lozano, A. C. (2017). Sparse+ group-sparse dirty models: Statistical guarantees without unreasonable conditions and a case for non-convexity. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* 3911–3920. JMLR. org.

Appendix A: Additional ADMM Schemes and Further Details

A.1. Consensus ADMM

Let $r^{(k)} = \mathbf{B}^{(k)} - \boldsymbol{\beta}^{(k)}$ denote the primal residuals and $s^{(k)} = \rho r^{(k)}$ denote the dual residuals. Then, following Boyd et al. (2011, Section 3.3), a stopping criteria can be defined using

$$\left\| r^{(k)} \right\|_2 \leqslant \varepsilon^{\text{primal}} \qquad \left\| s^{(k)} \right\|_2 \leqslant \varepsilon^{\text{dual}},$$

where,

$$\begin{aligned} \varepsilon^{\text{primal}} &= \sqrt{Kp}\,\varepsilon^{\text{abs}} + \varepsilon^{\text{rel}} \left[\|\boldsymbol{\beta}\|_2 \vee \|B\|_2 \right] \\ \varepsilon^{\text{dual}} &= \sqrt{Kp}\,\varepsilon^{\text{abs}} + \varepsilon^{\text{rel}}\, \rho \|\mathbf{D}\|_2, \end{aligned}$$

for some fixed $\varepsilon^{\text{abs}} > 0$ given by the problem context and ε^{rel} typically of the order of 10^{-3} or 10^{-4}.

A.2. Sharing ADMM

This is the approach in Boyd et al. (2011, Section 8.3), where we need only the separability of the linear predictor and of the penalty across j; in particular, this extends to objective functions that is not separable across k.

Introduce the partial linear predictors $\eta_j^{(k)} \in \mathbb{R}^{n_k}$, $k \in [K]$, $j \in [p]$. Then, we can considered the following augmented problem:

$$\begin{aligned} \underset{\boldsymbol{\beta}, l}{\text{minimize}} \quad & \sum_{k=1}^{K} \mathcal{L}^{(k)} \left(Y^{(k)}, \sum_{j=1}^{p} \eta^{(k)} \right) + \lambda \mathcal{P}(\boldsymbol{\beta}) \\ \text{subject to} \quad & \eta_j^{(k)} - \mathbf{X}_j^{(k)} \boldsymbol{\beta}_j^{(k)} = 0, \quad k \in [K], j \in [p]. \end{aligned}$$

The corresponding augmented Lagrangian is given by

$$\begin{aligned} L(\boldsymbol{\beta}, \eta, m) = \sum_{k=1}^{K} \mathcal{L}^{(k)} \left(Y^{(k)}, \sum_{j=1}^{p} \eta^{(k)} \right) + \lambda \mathcal{P}(\boldsymbol{\beta}) + \sum_{k=1}^{K} \sum_{j=1}^{p} \left\langle m_j^{(k)}, \eta^{(k)} - \mathbf{X}_j^{(k)} \boldsymbol{\beta}_j^{(k)} \right\rangle + \\ \frac{\rho}{2} \sum_{k=1}^{K} \sum_{j=1}^{p} \left\langle \eta^{(k)} - \mathbf{X}_j^{(k)} \boldsymbol{\beta}_j^{(k)}, \eta^{(k)} - \mathbf{X}_j^{(k)} \boldsymbol{\beta}_j^{(k)} \right\rangle, \end{aligned}$$

where $m_j^{(k)} \in \mathbb{R}^{n_k}$, $k \in [K]$, $j \in [p]$, is a Lagrange multiplier and $\rho > 0$ is the ADMM parameter. The update of β is given by

$$\boldsymbol{\beta}_j[t+1] = \arg\min_{\boldsymbol{\beta}_j} \lambda \mathcal{P}\left(\boldsymbol{\beta}_j\right) + \frac{\rho}{2} \sum_{k=1}^{K} \left\| \mathbf{X}_j^{(k)} \boldsymbol{\beta}_j^{(k)} - \left(\eta_j^{(k)}[t] + \frac{1}{\rho} m_j^{(k)}[t] \right) \right\|_2^2.$$

Now we consider the update steps for η. Since we are mostly interested in the sum of the partial linear predictors, we can use ideas from sharing (Boyd et al., 2011, Section 7.3). Consider

$$a_j^{(k)} := \mathbf{X}_j^{(k)} \boldsymbol{\beta}_j^{(k)}[t+1] - \frac{1}{\rho} m_j^{(k)}[t], \qquad \bar{a}^{(k)} := \frac{1}{p} \sum_{j=1}^{p} a_j^{(k)},$$

We can write the optimization for the update of η as

$$\begin{aligned} \underset{\eta, \bar{\eta}}{\text{minimize}} \quad & \sum_{k=1}^{K} \mathcal{L}^{(k)}\left(Y^{(k)}, p\bar{\eta}^{(k)}\right) + \frac{\rho}{2} \sum_{k=1}^{K} \sum_{j=1}^{p} \left\| \eta_j^{(k)} - a_j^{(k)} \right\|_2^2 \\ \text{subject to} \quad & \bar{\eta}^{(k)} - \frac{1}{p} \sum_{j=1}^{p} \eta_j^{(k)} = 0, \quad k \in [K]. \end{aligned}$$

An optimal solution will be such that

$$\eta_j^{(k)} - a_j^{(k)} = \bar{\eta}^{(k)} - \bar{a}^{(k)},$$

so that we can first solve for the mean linear predictor $\bar{\eta}^{(k)}$ and then compute the partial linear predict updates. Thus, the updates are given by

$$\begin{aligned} \bar{\eta}[t+1] &= \arg\min_{\bar{\eta}} \sum_{k=1}^{K} \mathcal{L}^{(k)}\left(Y^{(k)}, p\bar{\eta}^{(k)}\right) + \frac{\rho}{2} \sum_{k=1}^{K} \sum_{j=1}^{p} \left\| \bar{\eta}^{(k)} - \bar{a}^{(k)} \right\|_2^2 \\ \eta_j^{(k)}[t+1] &= a_j^{(k)} + \bar{\eta}^{(k)}[t+1] - \bar{a}^{(k)}, \\ m_j^{(k)}[t+1] &= \bar{m}^{(k)}[t] + \rho \left(\bar{\eta}^{(k)}[t+1] - \frac{1}{p} \mathbf{X}^{(k)} \boldsymbol{\beta}^{(k)}[t+1] \right) \end{aligned}$$

A.3. Sharing ADMM using Linear Predictors

Inspection of the scheme presented in Appendix A.2 indicates that we optimize mostly w.r.t. the linear predictors $\eta^{(k)} \in \mathbb{R}^{n_k}$, $k \in [K]$. This observation suggests to consider the following augmented problem:

$$\begin{aligned} \underset{\boldsymbol{\beta}, l}{\text{minimize}} \quad & \sum_{k=1}^{K} \mathcal{L}^{(k)}\left(Y^{(k)}, \eta^{(k)}\right) + \lambda \mathcal{P}_{q,\alpha}\left(\boldsymbol{\beta}\right) \\ \text{subject to} \quad & \eta^{(k)} - \mathbf{X}^{(k)} \boldsymbol{\beta}^{(k)} = 0, \quad k \in [K]. \end{aligned}$$

The corresponding augmented (scaled) Lagrangian is given by

$$L\left(\boldsymbol{\beta}, \eta, z\right) = \sum_{k=1}^{K} \mathcal{L}^{(k)}\left(Y^{(k)}, \eta^{(k)}\right) + \lambda\, \mathcal{P}\left(\boldsymbol{\beta}\right) + \frac{\rho}{2} \sum_{k=1}^{K} \left\| \eta^{(k)} - \mathbf{X}^{(k)}\, \boldsymbol{\beta}^{(k)} + z^{(k)} \right\|_2^2$$

The updates are then given by

$$\begin{aligned}
\boldsymbol{\beta}[t+1] &= \arg\min_{\boldsymbol{\beta}} \frac{\rho}{2} \sum_{k=1}^{K} \left\| \mathbf{X}^{(k)}\, \boldsymbol{\beta}^{(k)} - \left(\eta^{(k)}[t] + z^{(k)}[t]\right) \right\|_2^2 + \lambda\, \mathcal{P}_{q,\alpha}\left(\boldsymbol{\beta}\right) \\
\eta[t+1] &= \arg\min_{\eta} \sum_{k=1}^{K} \mathcal{L}^{(k)}\left(Y^{(k)}, \eta^{(k)}\right) + \frac{\rho}{2} \sum_{k=1}^{K} \left\| \eta^{(k)} - \left(\mathbf{X}^{(k)}\, \boldsymbol{\beta}^{(k)}[t+1] - z^{(k)}[t]\right) \right\|_2^2 \\
z^{(k)}[t+1] &= z^{(k)}[t] + \left(\eta^{(k)}[t+1] - \mathbf{X}^{(k)}\, \boldsymbol{\beta}^{(k)}[t+1]\right)
\end{aligned}$$

The β update is a simple $\mathcal{P}$-regularized least squares problem with working responses $\eta^{(k)}[t] + z^{(k)}[t]$; the η update is a proximal operator on L. Since L is assumed to be smooth, this also corresponds to a Ridge-regularized problem. Note that if L is separable in $k \in [K]$ this optimization problem can be split into K ridge problems efficiently.

Let's define the primal residuals $r^{(k)} = \eta^{(k)} - \mathbf{X}^{(k)}\, \boldsymbol{\beta}^{(k)} \in \mathbb{R}^{n_k}$, $k \in [K]$. Then, the $z^{(k)}$ updates take the form

$$z^{(k)}[t+1] = z^{(k)}[t] + r^{(k)}[t+1],$$

and therefore corresponds to a running sum of residuals between $\eta^{(k)}$ and the linear predictors $\mathbf{X}^{(k)}\, \boldsymbol{\beta}^{(k)}$. The dual residuals are given by

$$\begin{aligned}
s^{(k)} &= -\,\mathbf{X}^{(k)\top}\left(m^{(k)}[t] - m^{(k)}[t+1]\right) \\
&= \rho\, \mathbf{X}^{(k)\top}\, r^{(k)} \in \mathbb{R}^{p}, \quad k \in [K].
\end{aligned}$$

A stopping criteria (Boyd et al., 2011, Section 3.3) can be defined through

$$\left\| r^{(k)} \right\|_2 \leqslant \varepsilon^{\text{primal}}, \qquad\qquad \left\| s^{(k)} \right\|_2 \leqslant \varepsilon^{\text{dual}},$$

where, using some abuse of notation,

$$\begin{aligned}
\varepsilon^{\text{primal}} &= \sqrt{n}\, \varepsilon^{\text{abs}} + \varepsilon^{\text{rel}}\, [\|\mathbf{X}\, \boldsymbol{\beta}\|_2 \vee \|\eta\|_2] \\
\varepsilon^{\text{dual}} &= \sqrt{Kp}\, \varepsilon^{\text{abs}} + \varepsilon^{\text{rel}}\, \rho \left\| \mathbf{X}^{\top}\, z \right\|_2,
\end{aligned}$$

for some fixed $\varepsilon^{\text{abs}} > 0$ given by the problem context and ε^{rel} typically of the order of 10^{-3} or 10^{-4}.

A.4. Sharing ADMM using Residuals

For a special case of the scheme in Appendix A.3 (Xiao, Wu and He, 2012), we need to assume that the loss function only depends on the residuals $r^{(k)} := \mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)} - Y^{(k)}$, that is,

$$\mathcal{L}^{(k)}\left(Y^{(k)}, \mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)}\right) = \mathcal{L}^{(k)}\left(\mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)} - Y^{(k)}\right) = \mathcal{L}^{(k)}\left(r^{(k)}\right).$$

Then, we can consider the following augmented problem:

$$\begin{aligned} \underset{\boldsymbol{\beta}, r}{\text{minimize}} \quad & \sum_{k=1}^{K} \mathcal{L}^{(k)}\left(r^{(k)}\right) + \lambda\,\mathcal{P}\left(\boldsymbol{\beta}\right) \\ \text{subject to} \quad & r^{(k)} - \left(\mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)} - Y^{(k)}\right) = 0, \quad k \in [K]. \end{aligned}$$

The corresponding augmented (scaled) Lagrangian is given by

$$L\left(\boldsymbol{\beta}, r, m\right) = \sum_{k=1}^{K} \mathcal{L}^{(k)}\left(r^{(k)}\right) + \lambda\,\mathcal{P}_{q,\alpha}\left(\boldsymbol{\beta}\right) + \rho \sum_{k=1}^{K} \left\langle m^{(k)}, r^{(k)} - \left(\mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)} - Y^{(k)}\right)\right\rangle + \frac{\rho}{2} \sum_{k=1}^{K} \left\langle r^{(k)} - \left(\mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)} - Y^{(k)}\right), r^{(k)} - \left(\mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)} - Y^{(k)}\right)\right\rangle,$$

where $m^{(k)} \in \mathbb{R}^{n_k}$, $k \in [K]$, is a Lagrange multiplier and $\rho > 0$ is the ADMM parameter. The updates are given by

$$\begin{aligned} \boldsymbol{\beta}[t+1] &= \arg\min_{\boldsymbol{\beta}} L\left(\boldsymbol{\beta}, r[t], m[t]\right), \\ r[t+1] &= \arg\min_{r} L\left(\boldsymbol{\beta}[t+1], r, m[t]\right), \\ m^{(k)}[t+1] &= m^{(k)}[t] + \left[r^{(k)}[t+1] - \left(\mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)}[t+1] - Y^{(k)}\right)\right] \end{aligned}$$

We find

$$\boldsymbol{\beta}[t+1] = \arg\min_{\boldsymbol{\beta}} \lambda\,\mathcal{P}\left(\boldsymbol{\beta}\right) + \frac{\rho}{2} \sum_{k=1}^{K} \left\| \mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)} - \left(r^{(k)}[t] + Y^{(k)} + m^{(k)}[t]\right)\right\|_2^2,$$

which is one big sparse group Lasso problem. For the second update, we get

$$r[t+1] = \arg\min_{r} \sum_{k=1}^{K} \mathcal{L}^{(k)}\left(r^{(k)}\right) + \frac{\rho}{2} \sum_{k=1}^{K} \left\| r^{(k)} - \left(\mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)}[t+1] - Y^{(k)} - m^{(k)}[t]\right)\right\|_2^2,$$

which has the form of a simple Ridge regression problem.

A.5. Inexact ADMM

Eckstein and Bertsekas (1992, Theorem 8) show that we don't need to solve each sub-problem exactly, only that we need to solve them with a certain threshold and that the threshold should satisfy some decay bounds. For the purpose of self-contained, we include their theorem as followed:

Consider the minimization of $f(x) + g(Mx)$ where M has full column rank. The augmented Lagrangian

$$L(x, w, p) = f(x) + g(w) + \langle p, Mx - w\rangle + \frac{1}{2}\rho\|Mx - w\|_2^2.$$

The updates are given by

$$\begin{aligned} x[t+1] &= \arg\min_x f(x) + \langle p[t], Mx\rangle + \frac{1}{2}\rho\|Mx - w[t]\|_2^2 \\ w[t+1] &= \arg\min_w g(w) - \langle p[t], w\rangle + \frac{1}{2}\rho\|Mx[t+1] - w\|_2^2 \\ p[t+1] &= p[t] + \rho\left[Mx[t+1] - w[t+1]\right] \end{aligned}$$

Then, Eckstein and Bertsekas (1992, Theorem 8) implies we get convergence provided that

$$\begin{aligned} \left\| x[t+1] - \arg\min_x L(x, w[t], p[t]) \right\|_2 &\leqslant \mu_t \\ \left\| w[t+1] - \arg\min_w L(x[t+1], w, p[t]) \right\|_2 &\leqslant \nu_t \end{aligned}$$

with $\mu_t, \nu_t \geqslant 0$, $\sum \mu_t, \sum \nu_t < \infty$ and where $\rho > 0$.

A.6. KKT Conditions and Solution Path

For the sparse group Lasso penalty, the sub-gradient equations are given by

$$\mathbf{0}_{p\times K} \in \sum_{k=1}^{K} \nabla \mathcal{L}^{(k)}\left(Y^{(k)}, \mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)}\right) + \lambda \partial \mathcal{P}_{q,\alpha}(\boldsymbol{\beta})$$

For $\mathcal{P}_{q,\alpha}(\boldsymbol{\beta}_j) = \omega_j[\alpha\|\boldsymbol{\beta}_j\|_1 + (1-\alpha)\|\boldsymbol{\beta}_j\|_q]$, we find

$$\partial \mathcal{P}_{q,\alpha}(\boldsymbol{\beta}_j) = \omega_j\left[\alpha\partial\|\boldsymbol{\beta}_j\|_1 + (1-\alpha)\partial\|\boldsymbol{\beta}_j\|_q\right],$$

where + denotes set addition and where

$$[\partial\|\mathbf{u}\|_1]^{(k)} = \begin{cases} [-1,1], & \mathbf{u}^{(k)} = 0, \\ \{1\}, & \mathbf{u}^{(k)} > 0, \\ \{-1\}, & \mathbf{u}^{(k)} < 0, \end{cases} \qquad k \in [K];$$

$$\partial\|\mathbf{u}\|_q = \begin{cases} \{g \mid \|g\|_{q^*} \leqslant 1\}, & \mathbf{u} = \mathbf{0}_K, \\ \{\mathbf{u}/\|\mathbf{u}\|_{q^*}\}, & \mathbf{u} \neq \mathbf{0}_K, \end{cases}, \qquad 1 < q < \infty,\ q^* = \frac{q}{q-1};$$

$$\partial\|\mathbf{u}\|_\infty = \mathbf{conv}\left\{\mathbf{w} \mid \|\mathbf{w}\|_1 \leqslant 1, \mathbf{u}^\top \mathbf{w} = \|u\|_\infty\right\}.$$

To find the smallest λ such that all features are excluded, we use these KKT conditions. If $\boldsymbol{\beta} = \mathbf{0}_{p\times K}$, we find

$$\sum_{k=1}^{K} \nabla \mathcal{L}\left(Y^{(k)}, \mathbf{0}_{n^{(k)}}\right) \in \lambda \partial \mathcal{P}_{q,\alpha}\left(\mathbf{0}_{p\times K}\right).$$

Thus, the equation in $\boldsymbol{\beta}_j$ corresponds to

$$\frac{1}{\omega_j \lambda} \sum_{k=1}^{K} \nabla_j \mathcal{L}\left(Y^{(k)}, \mathbf{0}_{n^{(k)}}\right) \in \alpha \partial \|\mathbf{0}_K\|_1 + (1-\alpha)\partial\|\mathbf{0}_K\|_q,$$

that is, there exists $\mathbf{u} \in \partial\|\mathbf{0}_K\|_1$ and $\mathbf{v} \in \partial\|\mathbf{0}_K\|_q$ such that

$$\frac{1}{\omega_j \lambda} \sum_{k=1}^{K} \nabla_j \mathcal{L}\left(Y^{(k)}, \mathbf{0}_{n^{(k)}}\right) = \alpha\mathbf{u} + (1-\alpha)\mathbf{v}.$$

For any $1 \leqslant q \leqslant \infty$, we must have $\|\mathbf{u}\|_\infty \leqslant 1$ and $\|\mathbf{v}\|_{q^*} \leqslant 1$. Denote

$$a_j^{(k)} = \nabla_j^{(k)} \mathcal{L}\left(Y^{(k)}, \mathbf{0}_{n^{(k)}}\right),$$

so we need to solve $\frac{1}{\lambda}\mathbf{a}_j = \alpha\mathbf{u}_j + (1-\alpha)\mathbf{v}_j$ where $\|\mathbf{u}_j\|_\infty \leqslant 1$ and $\|\mathbf{v}_j\|_{q^*} \leqslant 1$. By Hölder's inequality, we have $\|\mathbf{v}_j\|_{q^*} \leqslant K^{\frac{1}{q^*}-\frac{1}{\infty}}\|\mathbf{v}_j\|_\infty$. Then, we find the bound

$$\begin{aligned} \frac{1}{\lambda}\|\mathbf{a}_j\|_{q^*} &= \|\alpha\mathbf{u}_j + (1-\alpha)\mathbf{v}_j\|_{q^*} \\ &\leqslant \alpha\|\mathbf{u}_j\|_{q^*} + (1-\alpha)\|\mathbf{v}_j\|_{q^*} \\ &\leqslant \alpha K^{1/q^*}\|\mathbf{u}_j\|_\infty + (1-\alpha)\|\mathbf{v}_j\|_{q^*} \\ &\leqslant \alpha K^{1/q^*} + (1-\alpha). \end{aligned}$$

This allows us to have a heuristic starting value,

$$\lambda_{\max} = \frac{\max_{j\in[p]} \frac{1}{\omega_j}\left\|\sum_{k=1}^{K} \nabla_j \mathcal{L}\left(Y^{(k)}, \mathbf{0}_{n^{(k)}}\right)\right\|_{q^*}}{\alpha K^{1/q^*} + (1-\alpha)}.$$

While this may not be the smallest such lambda, it will be reasonably close to the true value at the cost of possibly performing a few regularization iterations for nothing at first. Note that with $q = 1$, we recover the exact bound since $q^* = \infty$ implies $K^{1/q^*} = 1$. If $\alpha = 0$ then we find the group Lasso regularization which has an exact bound

$$\lambda_{\max} = \max_{j \in [p]} \frac{1}{\omega_j} \left\| \sum_{k=1}^{K} \nabla_j \mathcal{L}\left(Y^{(k)}, \mathbf{0}_{n^{(k)}}\right) \right\|_{q^*},$$

which agrees with the above bound.

Appendix B: Solving the Subproblems

B.1. Regularized Least Squares

We consider the following problem:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \qquad \frac{\rho}{2} \sum_{k=1}^{K} \left\| \mathbf{X}^{(k)} \boldsymbol{\beta}^{(k)} - V^{(k)} \right\|_2^2 + \lambda \mathcal{P}\left(\boldsymbol{\beta}\right)$$

for some vector $V^{(k)} \in \mathbb{R}^{n_k}$, where $\lambda > 0$, $\rho > 0$. We recognize the previous problem to be of the form smooth convex function of $\boldsymbol{\beta}$ + convex function of $\boldsymbol{\beta}$ which can be solve using proximal gradient descent. Let

$$Q(\boldsymbol{\beta}) := \frac{\rho}{2} \sum_{k=1}^{K} \left\| \mathbf{X}^{(k)} \boldsymbol{\beta}^{(k)} - V^{(k)} \right\|_2^2,$$

with

$$\nabla_{\boldsymbol{\beta}^{(k)}} Q(\boldsymbol{\beta}^{(k)}) = \rho \, \mathbf{X}^{(k)\top} \left(\mathbf{X}^{(k)} \boldsymbol{\beta}^{(k)} - V^{(k)} \right).$$

The proximal gradient descent updates are given by

$$\boldsymbol{\beta}[t+1] = \mathbf{prox}_{\eta[t] \lambda \mathcal{P}_{q,\alpha}} \left(\boldsymbol{\beta}[t] - \eta[t] \nabla Q(\boldsymbol{\beta}[t]) \right),$$

where $\eta[t] > 0$ is a step-size. By the separability of $\mathcal{P}_{q,\alpha}$ across $j \in [p]$, these updates can be performed independently:

$$\boldsymbol{\beta}_j[t+1] = \mathbf{prox}_{\eta[t] \lambda \mathcal{P}_{q,\alpha}} \left(\boldsymbol{\beta}_j[t] - \eta[t] \nabla_j Q(\boldsymbol{\beta}_j[t]) \right),$$

Note that we have a constant Hessian:

$$\nabla^2_{\boldsymbol{\beta}^{(k)}, \boldsymbol{\beta}^{(k)}} Q(\boldsymbol{\beta}^{(k)}) = \rho \, \mathbf{X}^{(k)\top} \mathbf{X}^{(k)},$$

$$\nabla^2 Q(\boldsymbol{\beta}) = \rho \, \mathrm{diag} \left(\mathbf{X}^{(k)\top} \mathbf{X}^{(k)}, k \in [K] \right)$$

so we may use this information to find $\eta[t]$. Indeed, we can find an upper bound that will remain constant throughout the ADMM so we can fix the stepsize.

B.2. Ridge Regularization

In a unified form, we identify the following problem:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \qquad F(x) := f(x) + \frac{1}{2\tau}\|x - v\|_2^2,$$

which implies that the optimal solution is given by $\mathbf{prox}_{\tau f}(v)$. When f is assumed to be sufficiently smooth, we can compute its proximal operator using a first-order method. The quadratic penalty term will then help convergence.

The gradient of the objective function is given by

$$\nabla F(x) = \nabla f(x) + \frac{1}{\tau}(x - v),$$

and its Hessian by

$$\nabla^2 F(x) = \nabla^2 f(x) + \frac{1}{\tau} I_d.$$

Thus, if we have an hessian upper bound $\nabla^2 f \preceq L I_p$, we find $\nabla^2 F \preceq (L + \tau^{-1}) I_p$ which can be used to construct a fixed step-size algorithm.