Projection onto Minkowski Sums with Application to Constrained Learning

Joong-Ho (Johann) Won[1] Jason Xu[2] Kenneth Lange[3]

[1]Department of Statistics, Seoul National University
[2]Department of Statistical Science, Duke University
[3]Departments of Biomathematics, Human Genetics, and Statistics, UCLA

June 11, 2019

International Conference on Machine Learning

Outline

- Minkowski sum and projection
- Why are Minkowski sums useful for constrained learning?
- Constrained learning via projection onto Minkowski sums
- Minkowski projection algorithm
- Applications to constrained learning
- Conclusion

Minkowski sum of sets

$$A + B \triangleq \{\boldsymbol{a} + \boldsymbol{b} : \boldsymbol{a} \in A,\, \boldsymbol{b} \in B\}, \quad A, B \subset \mathbb{R}^d$$

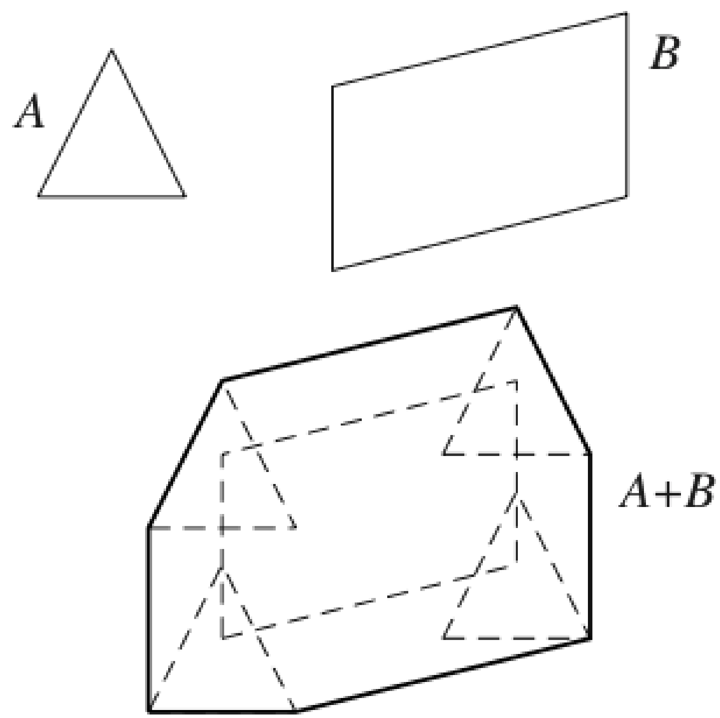


Image source: Christophe Weibel

https://sites.google.com/site/christopheweibel/research/minkowski-sums

Projection onto Minkowski sums

$$P_{A+B}(\boldsymbol{x}) = \underset{\boldsymbol{u}\in A+B}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{u}-\boldsymbol{x}\|_2^2, \quad \boldsymbol{x} \notin A+B \tag{P}$$

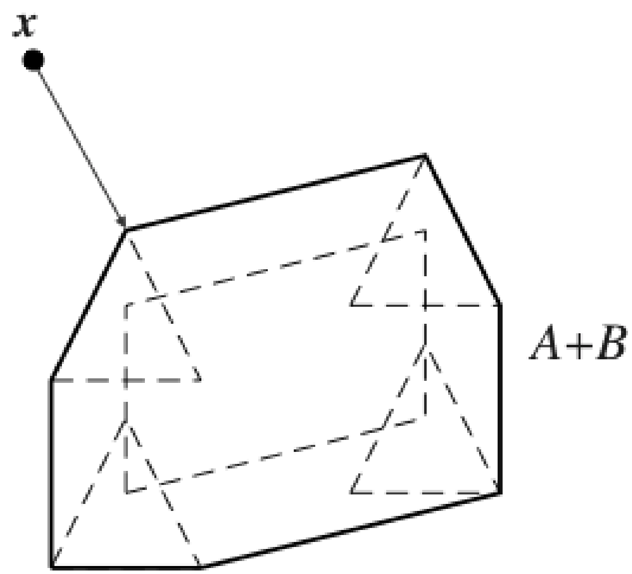


Image source: Christophe Weibel

https://sites.google.com/site/christopheweibel/research/minkowski-sums

Why are Minkowski sums useful for constrained learning?

Many penalized or constrained learning problems are of the form

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}) + \sum_{i=1}^{k} \sigma_{C_i}(\boldsymbol{x})$$

- $\sigma_C(\boldsymbol{x}) = \sup_{\boldsymbol{y} \in C} \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is the support function of convex set C.
- Example: elastic net $\min_{\boldsymbol{x}} f(\boldsymbol{x}) + \lambda_1 \|\boldsymbol{x}\|_1 + \lambda_2 \|\boldsymbol{x}\|_2$,

  $C_1 = \{\boldsymbol{x} : \|\boldsymbol{x}\|_\infty \leq \lambda_1\}$, $C_2 = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq \lambda_2\}$ (dual norm balls)

Why are Minkowski sums useful for constrained learning?

Many penalized or constrained learning problems are of the form

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) + \sum_{i=1}^{k} \sigma_{C_i}(\boldsymbol{x}) = \boxed{\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) + \sigma_{C_1+\cdots+C_k}(\boldsymbol{x})} \tag{1}$$

- Support functions are additive over Minkowski sums (Hiriart-Urruty and Lemaréchal 2012).
- New perspective on LHS: minimizing sum of *two* (convex) functions instead of $k+1$ functions.

Multiple/overlapping norm penalties

$\ell_{1,p}$ group lasso/multitask learning (Yuan and Lin 2006) with overlaps allowed:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) + \lambda \sum_{i=1}^{k} \|\boldsymbol{x}_{i1}\|_p, \quad p \geq 1$$

where $\boldsymbol{x}_{i1}$=subvector of $\boldsymbol{x}$; $i1 \subset \{1,\ldots,d\}$=group index.

- Involved sets: ℓ_q-norm disks.

$$\begin{aligned} C_i &= \{\boldsymbol{y} = (\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}) : \|\boldsymbol{y}_{i1}\|_q \leq \lambda,\ \boldsymbol{y}_{i2} = \boldsymbol{0}\}, \\ \frac{1}{p} &+ \frac{1}{q} = 1, \quad i2 = \{1,\ldots,d\} \setminus i1. \end{aligned} \tag{2}$$

- No distinction between overlapping vs. non-overlapping groups!

Conic constraints

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}) \text{ subject to } \boldsymbol{x} \in K_1^* \cap K_2^* \cap \cdots \cap K_k^*$$

where $K_i^* = \{\boldsymbol{y} : \langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq 0, \forall \boldsymbol{x} \in K_i\}$ is the polar cone of closed convex cone K_i.

- Use the fact $\iota_{K_i^*}(\boldsymbol{x}) = \sigma_{K_i}(\boldsymbol{x})$ to express it as

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}) + \sum_{i=1}^{k} \iota_{K_i^*}(\boldsymbol{x}) = \min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}) + \sum_{i=1}^{k} \sigma_{K_i}(\boldsymbol{x}).$$

- $\iota_S = 0/\infty$ indicator of set S

Constrained lasso: mix-and-match

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|_1 \text{ subject to } \boldsymbol{B}\boldsymbol{x} = \boldsymbol{0},\ \boldsymbol{C}\boldsymbol{x} \leq \boldsymbol{0},$$

which subsumes the generalized lasso (Tibshirani and Taylor 2011) as a special case (James, Paulson, and Rusmevichientong 2013; Gaines, Kim, and Zhou 2018).

- Involved sets: cone, subspace, and ℓ_∞-norm ball

$$\begin{aligned} C_1 &= \{\boldsymbol{x} : \boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}\}^* = \{\boldsymbol{x} : \boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}\}^\perp, \\ C_2 &= \{\boldsymbol{x} : \boldsymbol{C}\boldsymbol{x} \leq \boldsymbol{0}\}^*, \quad C_3 = \{\boldsymbol{x} : \|\boldsymbol{x}\|_\infty \leq \lambda\} \end{aligned} \tag{3}$$

Constrained learning via projection onto Minkowski sums

Contemporary methods for solving problem (1) (e.g., proximal gradient) requires computing the proximity operator of $\sigma_{C_1+\cdots+C_k}$:

$$\mathrm{prox}_{\gamma\sigma_{C_1+\cdots+C_k}}(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^d} \sigma_{C_1+\cdots+C_k}(\boldsymbol{u}) + \frac{1}{2\gamma}\|\boldsymbol{u}-\boldsymbol{x}\|_2^2$$

- Proximal gradient:

$$\boldsymbol{x}^{(t+1)} = \mathrm{prox}_{\gamma_t\sigma_{C_1+\cdots+C_k}}\left(\boldsymbol{x}^{(t)} - \gamma_t^{-1}\nabla f(\boldsymbol{x}^{(t)})\right)$$

- Can be computed via Minkowski projection

- Duality:

$$\sigma^*_{C_1+\cdots+C_k}(\boldsymbol{y}) = \iota_{C_1+\cdots+C_k}(\boldsymbol{y}), \quad (\iota_S(\boldsymbol{u}) = 0 \text{ if } \boldsymbol{u} \in S,\ \infty \text{ otherwise})$$

  if $C_1 + \cdots + C_k$ is closed convex; $g^*(\boldsymbol{y}) = \sup_{\boldsymbol{x}} \langle \boldsymbol{x}, \boldsymbol{y} \rangle - g(\boldsymbol{x})$ is the Fenchel conjugate of g.

- Moreau's decomposition

$$\boldsymbol{x} = \operatorname{prox}_{\gamma g}(\boldsymbol{x}) + \gamma \operatorname{prox}_{\gamma^{-1} g^*}(\gamma^{-1}\boldsymbol{x})$$

In terms of Minkowski projection,

$$\begin{aligned} \operatorname{prox}_{\gamma\sigma_{C_1+\cdots+C_k}}(\boldsymbol{x}) &= \boldsymbol{x} - \gamma \operatorname{prox}_{\gamma^{-1}\iota_{C_1+\cdots+C_k}}(\gamma^{-1}\boldsymbol{x}) \\ &= \boxed{\boldsymbol{x} - \gamma P_{C_1+\cdots+C_k}(\gamma^{-1}\boldsymbol{x})} \end{aligned}$$

Minkowski projection algorithm

Goal: to develop an efficient method for computing $P_{C_1+\cdots+C_k}(\boldsymbol{x})$, in case projection onto each set $P_{C_i}(\boldsymbol{x})$ is simple.

MM algorithm:

1: Input: External point $\boldsymbol{x} \notin C_1 + \ldots + C_k$;
Projection operator P_{C_i} onto set C_i, $i = 1, \ldots, k$;
initial value $\boldsymbol{a}_0^i$, $i = 1, \ldots, k$; viscosity parameter $\rho \geq 0$

2: Initialization: $n \leftarrow 0$

3: Repeat

4: For $i = 1, 2, \ldots, k$

5: $\boldsymbol{a}_{n+1}^{(i)} \leftarrow P_{C_i}\Big(\frac{1}{1+\rho}\big(\boldsymbol{x} - \sum_{j=1}^{i-1} \boldsymbol{a}_{n+1}^{(j)} - \sum_{j=i+1}^{k} \boldsymbol{a}_n^{(j)}\big) + \frac{\rho}{1+\rho}\boldsymbol{a}_n^{(i)}\Big)$

6: End For

7: $n \leftarrow n + 1$

8: Until Convergence

9: Return $\sum_{i=1}^{k} \boldsymbol{a}_n^{(i)}$

Properties of the Algorithm

- Assume $k = 2$ for exposition purpose: $A = C_1$, $B = C_2$.

Proposition 1. If both A and B are closed and convex, and $A + B$ is closed, then the Algorithm with $\rho = 0$ generates a sequence converging to $P_{A+B}(\boldsymbol{x})$.

≫ Proof: paracontraction (Elsner, Koltracht, and Neumann 1992; Lange 2013).

Theorem 1. If in addition either A or B is *strongly convex*, then the sequence generated by Algorithm with $\rho = 0$ converges *linearly* to $P_{A+B}(\boldsymbol{x})$.

≫ Set $C \subset \mathbb{R}^d$ is α-strongly convex with respect to norm $\|\cdot\|$ if there is a constant $\alpha > 0$ such that for any $\boldsymbol{a}$ and $\boldsymbol{b}$ in C and any $\gamma \in [0, 1]$, C contains a ball of radius $r = \gamma(1-\gamma)\frac{\alpha}{2}\|\boldsymbol{a} - \boldsymbol{b}\|^2$ centered at $\gamma\boldsymbol{a} + (1-\gamma)\boldsymbol{b}$ (Garber and Hazan 2015).

≫ Ex) ℓ_q-norm ball for $q \in (1, 2]$

Theorem 2. If A and B are closed and subanalytic (possibly non-convex), and at least one of them is bounded, then the sequence generated by the Algorithm with $\rho > 0$ converges to a critical point of (P) regardless of the initial values.

- $\gg$ Proof: Kurdyka-Łojasiewicz inequality (Bolte, Daniilidis, and Lewis 2007).

Theorem 3. If $A + B$ is polyhedral, then the Algorithm with $\rho > 0$ generates a sequence converging *linearly* to $P_{A+B}(\boldsymbol{x})$.

- $\gg$ Proof: Luo-Tseng error bound (Karimi, Nutini, and Schmidt 2018).
- $\gg$ Ex) $\ell_{1,\infty}$ overlapping group penalty/multitask learning; polyhedra are not strongly convex

Applications to constrained learning

Overlapping group penalties/multitask learning

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) + \lambda \sum_{i=1}^{k} \|\boldsymbol{x}_{i1}\|_p,$$

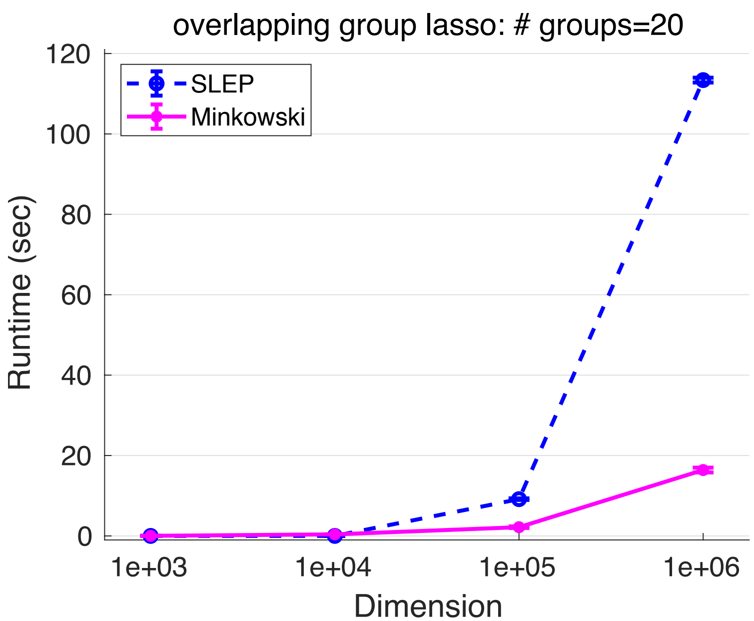
$$C_i = \{\boldsymbol{y} = (\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}) : \|\boldsymbol{y}_{i1}\|_q \leq \lambda,\ \boldsymbol{y}_{i2} = \mathbf{0}\}$$

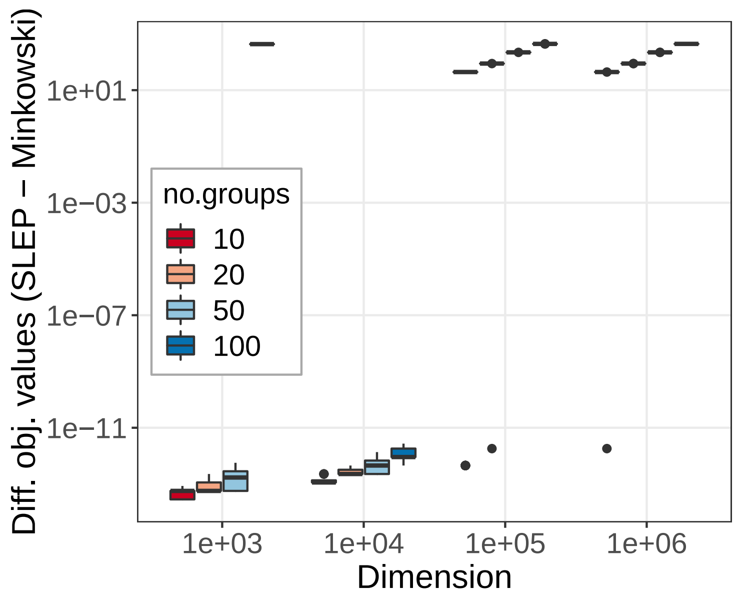
- Overlaps automatically handled with Minkowski projection.

- If $p \in [2, \infty)$, dual ℓ_q-norm disks are strongly convex; if $p = \infty$, polyhedral (linear convergence)

- Fast and reliable algorithm for projection onto ℓ_q-norm disks available (Liu and Ye 2010).

- Comparison to the dual projected gradient method used in SLEP (Yuan, Liu, and Ye 2011; Liu, Ji, and Ye 2011; Zhou, Zhang, and So 2015):





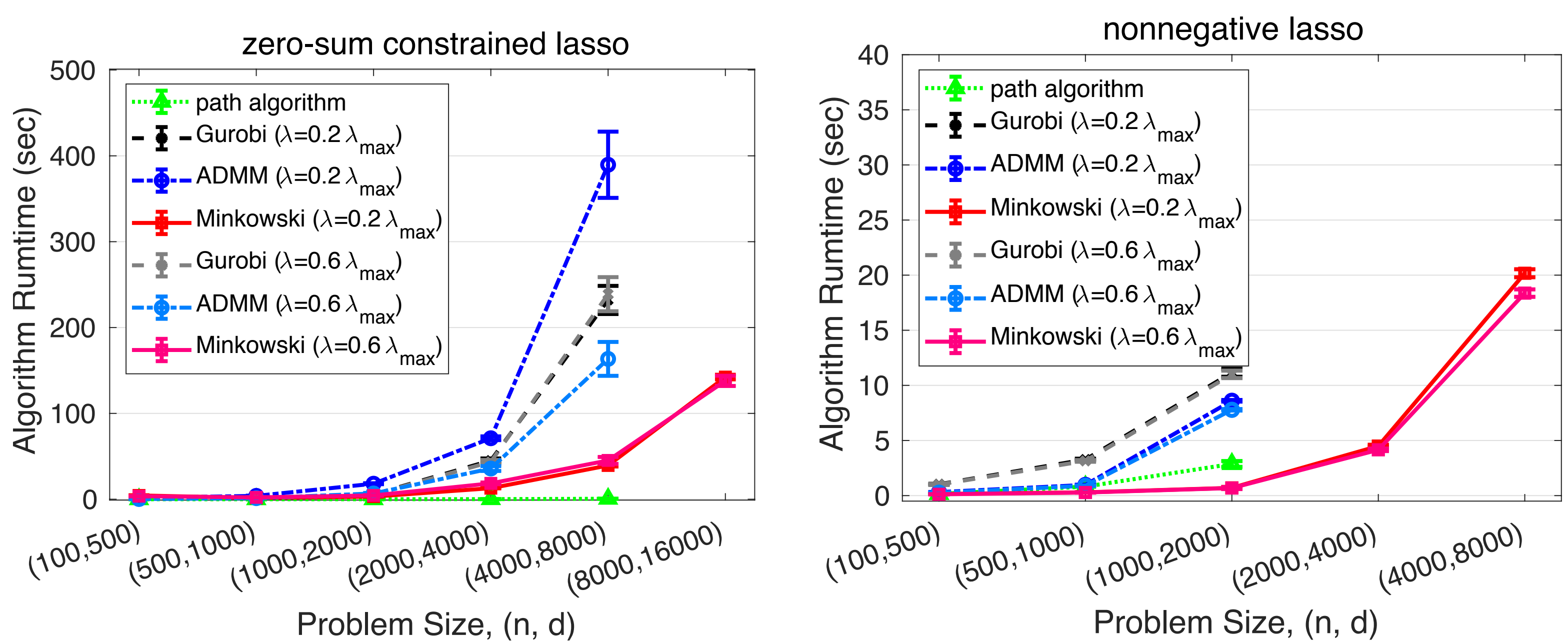
Constrained lasso

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|_1 \text{ subject to } \boldsymbol{B}\boldsymbol{x} = \boldsymbol{0},\ \boldsymbol{C}\boldsymbol{x} \le \boldsymbol{0},$$

- Zero-sum constrained lasso (Lin et al. 2014; Altenbuchinger et al. 2017): $C_1 = \{\boldsymbol{x} : \sum_{j=1}^d x_j = 0\}^\perp$, $C_2 = \{\boldsymbol{0}\}$, $C_3 = \{\boldsymbol{x} : \|\boldsymbol{x}\|_\infty \le \lambda\}$ ($\boldsymbol{B} = \boldsymbol{1}^T$, $\boldsymbol{C} = \boldsymbol{0}$).

- Nonnegative lasso (Efron et al. 2004; El-Arini et al. 2013): $C_1 = \{\boldsymbol{0}\}$, $C_2 = \{\boldsymbol{x} : -\boldsymbol{x} \le \boldsymbol{0}\}^*$, $C_3 = \{\boldsymbol{x} : \|\boldsymbol{x}\|_\infty \le \lambda\}$ ($\boldsymbol{B} = \boldsymbol{0}$, $\boldsymbol{C} = -\boldsymbol{I}$).

- Comparison to generic methods by Gaines, Kim, and Zhou (2018), including path algorithm, ADMM, and commercial solver Gurobi:



Conclusion

- Reconsider constrained learning problems:

  $\gg$ structural complexities such as non-separability can be handled gracefully via formulations involving Minkowski sums.

- Very simple and efficient algorithm for projecting points onto Minkowski sums of sets:

  $\gg$ Linear rate of convergence whenever at least one summand is strongly convex or the Luo-Tseng error bound condition is satisfied.

- Our algorithm can serve as an inner loop in, e.g, proximal gradient:

  $\gg$ Competitive performance
  $\gg$ Fast (inner loop) convergence is crucial.

*References

Altenbuchinger, Michael, Thorsten Rehberg, HU Zacharias, Frank Stämmler, Katja Dettmer, Daniela Weber, Andreas Hiergeist, Andre Gessner, Ernst Holler, Peter J Oefner, et al. 2017. "Reference point insensitive molecular data analysis." *Bioinformatics* 33 (2): 219–226.

El-Arini, Khalid, Min Xu, Emily B Fox, and Carlos Guestrin. 2013. "Representing documents through their readers." In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 14–22. ACM.

Bolte, Jérôme, Aris Daniilidis, and Adrian Lewis. 2007. "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems." *SIAM Journal on Optimization* 17 (4): 1205–1223.

Boyd, Stephen, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. 2010. "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers." *Foundations and Trends in Machine Learning* 3 (1): 1–122.

Davis, Damek, and Wotao Yin. 2017. "A three-operator splitting scheme and its optimization applications." *Set-valued and Variational Analysis* 25 (4): 829–858.

Efron, Bradley, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. 2004. "Least angle regression." *Annals of Statistics* 32 (2): 407–499.

Elsner, Ludwig, Israel Koltracht, and Michael Neumann. 1992. "Convergence of sequential and asynchronous nonlinear paracontractions." *Numerische Mathematik* 62 (1): 305–319.

Gaines, Brian R., Juhyun Kim, and Hua Zhou. 2018. "Algorithms for Fitting the Constrained Lasso." *Journal of Computational and Graphical Statistics* 27 (4): 861–871.

Garber, Dan, and Elad Hazan. 2015. "Faster rates for the Frank-Wolfe method over strongly-convex sets." In *Proceedings of the 32nd International Conference on Machine Learning,* 37:541–549.

Hiriart-Urruty, Jean-Baptiste, and Claude Lemaréchal. 2012. *Fundamentals of Convex Analysis.* Springer Science & Business Media.

James, Gareth M, Courtney Paulson, and Paat Rusmevichientong. 2013. "Penalized and constrained regression." *Unpublished Manuscript, available at* `http://www-bcf.usc.edu/~gareth/research/Research.html`*.*

Karimi, Hamed, Julie Nutini, and Mark Schmidt. 2018. "Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition." *arXiv preprint arXiv:1608.04636 v3.*

Lange, Kenneth. 2013. *Optimization.* 2nd. Springer.

Lin, Wei, Pixu Shi, Rui Feng, and Hongzhe Li. 2014. “Variable selection in regression with compositional covariates.” *Biometrika* 101 (4): 785–797.

Liu, Jun, Shuiwang Ji, and Jieping Ye. 2011. *SLEP: Sparse learning with efficient projections.* Technical report. Arizona State University. `https://github.com/jiayuzhou/SLEP`.

Liu, Jun, and Jieping Ye. 2010. “Efficient ℓ_1/ℓ_q norm regularization.” *arXiv preprint arXiv:1009.4766.*

Tibshirani, Ryan J., and Jonathan Taylor. 2011. “The solution path of the generalized lasso.” *Annals of Statistics* 39 (3): 1335–1371.

Yuan, Lei, Jun Liu, and Jieping Ye. 2011. “Efficient methods for overlapping group lasso.” In *Advances in Neural Information Processing Systems,* 352–360.

Yuan, Ming, and Yi Lin. 2006. "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (1): 49–67.

Zhou, Zirui, Qi Zhang, and Anthony Man-Cho So. 2015. "$\ell_{1,p}$-Norm Regularization: Error Bounds and Convergence Rate Analysis of First-Order Methods." In *Proceedings of the 32nd International Conference on Machine Learning,* 37:1501–1510.

Comparison to other algorithms

- Splitting methods: ADMM (Boyd et al. 2010), Davis-Yin three-operator splitting (Davis and Yin 2017)
- Do not produce descent algorithms, and introduce additional variables as well as intermediate steps.
- We do not know whether these methods can achieve a linear convergence rate under, e.g., strong convexity of a summand set.
- Sublinear rates for non-strongly convex sets can be achieved with our algorithm with $\rho > 0$ as well.