Frank-Wolfe Splitting via Augmented Lagrangian Method

Gauthier Gidel gauthier.gidel@umontreal.ca
MILA & DIRO Université de Montréal
Fabian Pedregosa fabian.pedregosa@inria.fr
INRIA Paris - Sierra team
Simon Lacoste-Julien simon.lacoste-julien@umontreal.ca
MILA & DIRO Université de Montréal

Abstract

Minimizing a function over an intersection of convex sets is an important task in optimization, but it is often much more challenging than minimizing over each individual constraint set. While traditional methods such as Frank-Wolfe (FW) or proximal gradient descent assume access to a linear or quadratic oracle on the intersection, splitting techniques, developed in the context of proximal methods, take advantage of the structure of each sets, and only require to perform the projection step over each individual constraint. In this work we develop and analyze the Frank-Wolfe Augmented Lagrangian (FW-AL), a method for minimizing a smooth function over an intersection of constraints that only requires access to a linear minimization oracle over the individual constraints. This method is based on the Augmented Lagrangian Method (ALM), a.k.a Method of Multipliers, but only requires access to linear (instead of quadratic) minimization oracles. We use recent advances in the analysis of FW and the alternating direction method of multipliers algorithms to prove a linear convergence rate of our algorithm over an intersection of polytopes and a sublinear convergence rate over general convex compact sets.

1 Introduction

The Frank-Wolfe (FW) or conditional gradient algorithm has seen an impressive revival in recent years, notably due to its very favorable properties for the optimization of sparse problems (Jaggi, 2013). This algorithm assumes knowledge of a linear minimization oracle (LMO) over the set of constraints. This oracle is inexpensive to compute for sets such as the ℓ_1 or trace norm ball. However, inducing complex priors often requires to consider *multiple* constraints, leading to a constraint set formed by the intersection of the original constraints. Unfortunately, evaluating the LMO over this intersection can be very challenging even if the LMO on the individual sets are inexpensive.

The problem of minimizing over an intersection of convex constraints is pervasive in machine learning and signal processing. For example, one can seek for a matrix that is both sparse and low rank by constraining the solution to have *both* small ℓ_1 and trace norm (Richard et al., 2012) or find a set of brain maps which are both sparse and piecewise constant by constraining both the ℓ_1 and total variation pseudonorm (Gramfort et al., 2013).

The objective of this paper is to describe and analyze FW-AL, an optimization method that can solve convex optimization problems subject to multiple constraints, assuming we have access to a LMO on each of the set of constraints.

Previous work. One of the most popular algorithm to solve optimization problems over an intersection of constraints is the alternating direction method of multipliers (ADMM), proposed by Glowinski and Marroco (1975), studied by Gabay and Mercier (1976), and revisited many times; see for instance (Boyd et al., 2011; Yan and Yin, 2016). On some cases, such as constraints on the trace norm (Cai et al., 2010) or the latent group lasso (Obozinski et al., 2011), the projection step can be a time-consuming operation, while the Frank-Wolfe LMO is much cheaper in both cases. Moreover, for some highly structured polytopes such as those appearing in alignment constraints (Alayrac et al., 2016) or Structured SVM (Lacoste-Julien et al., 2013), there exists a fast and elegant dynamic programming algorithm to compute the LMO, while there is no known tractable algorithm to compute the projection. Recently, Yen et al. (2016a) proposed a FW variant for objectives with a linear loss function over an intersection of polytopes named Greedy Direction Method of Multipliers

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

(GDMM). A similar version of GDMM is also used in (Yen et al., 2016b; Huang et al., 2017) to optimize a function over a Cartesian product of spaces related to each other by a linear consistency constraint. The constraints are incorporated through the augmented Lagrangian method and its convergence analysis crucially uses recent progress in the analysis of ADMM by Hong and Luo (2017).

Contributions. Our main contribution is the development of a novel variant of FW for the optimization of a function over an intersection of sets and its rigorous analysis. We name this method Frank-Wolfe via Augmented Lagrangian method (FW-AL). With respect to Yen et al. (2016a,b); Huang et al. (2017), our framework generalizes GDMM by providing a method to optimize a general class of functions over an intersection of an arbitrary number of compact sets, which are *not* restricted to be polytopes. We show that FW-AL converges for any smooth objective function. We prove that a standard gap measure converges linearly (i.e., with a geometric rate) when the constraint sets are polytopes, and sublinearly for general compact convex sets. We also show that when the function is strongly convex, this gap measure gives a bound on the distance to the set of optimal solutions. This is of key practical importance since the applications that we consider (e.g., minimization with trace norm constraints) verify these assumptions.

The paper is organized as follows. In Section 2, we introduce the general setting, provide some motivating applications and present the augmented Lagrangian formulation of our problem. In Section 3, we describe the algorithm FW-AL and provide its analysis.

2 Problem Setting

Let $f : \mathbb{R}^p \to \mathbb{R}$ be a convex smooth function and for $k \in [K]$, $\mathcal{X}_k \subset \mathbb{R}^p$ be convex compact sets. We will consider the following minimization problem,

$$\underset{\boldsymbol{x} \in \bigcap_{k=1}^K \mathcal{X}_k}{\text{minimize}} \; f(\boldsymbol{x}) \,, \quad \text{with only access to} \quad \text{LMO}_k(\boldsymbol{r}) \in \underset{\boldsymbol{s} \in \mathcal{X}_k}{\arg\min} \, \langle \boldsymbol{s}, \boldsymbol{r} \rangle \,, \; k \in [K] \,. \tag{OPT}$$

This framework models several problems that arise in machine learning and signal processing. We denote by $\mathcal{X}^*$ the set of optimal points of the optimization problem (OPT) and we will assume that this problem is feasible, i.e., the set of solutions is non empty. By casting (OPT) into the problem of finding a saddle point of an augmented Lagrangian (Bertsekas, 1996), we can split the constraints in a way in which the linear oracle is computed over the product space $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_K \subset \mathbb{R}^m$. Noting $\boldsymbol{x} := (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(K)})$ and $\bar{f}(\boldsymbol{x}) := \frac{1}{K} \sum_{k=1}^K f(\boldsymbol{x}^{(k)})$ we can consider the augmented Lagrangian formulation of (OPT)

$$\underset{\boldsymbol{x}}{\text{minimize}} \; \max_{\boldsymbol{y} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) \quad \text{s.t.} \quad \boldsymbol{x}^{(k)} \in \mathcal{X}_k, \;\; k \in \{1, \ldots, K\} \,, \tag{OPT2}$$

where $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) := \bar{f}(\boldsymbol{x}) + \langle \boldsymbol{y}, M\boldsymbol{x} \rangle + \frac{\lambda}{2} \|M\boldsymbol{x}\|^2$, and M is such that $M\boldsymbol{x} = 0 \Leftrightarrow \boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k+1)}$, $k \in [K-1]$. This saddle point formulation is the one onto our algorithm FW-AL is performed.

3 FW-AL Algorithm

The augmented Lagrangian method alternates a primal update on $\boldsymbol{x}$ (approximately) minimizing[1] the augmented Lagrangian $\mathcal{L}(\cdot, \boldsymbol{y}_t)$, with a dual update on $\boldsymbol{y}$ by taking a gradient ascent step on $\mathcal{L}(\boldsymbol{x}_{t+1}, \cdot)$. The FW-AL algorithm follows the general iteration of the augmented Lagrangian method, but with the crucial difference that Lagrangian minimization is replaced by one Frank-Wolfe step on $\mathcal{L}(\cdot, \boldsymbol{y}_t)$. The algorithm is thus composed by two loops: an outer loop presented in (2) and an inner loop noted $\mathcal{FW}$ which can be chosen to be one of the FW step variants described in Alg. 1 or 2.

FW steps. In our algorithm, we need to ensure that the $\mathcal{FW}$ inner loop makes sufficient progress. For general sets, we can use one iteration of the classical Frank-Wolfe algorithm with line-search Jaggi (2013) as given in Algorithm 2. When working over polytopes, we can get faster (linear) convergence by taking one *non-drop* step (defined below) of the away-step variant of the FW algorithm (AFW) Lacoste-Julien and Jaggi (2015), as described in Algorithm 1).

[1] An example of approximate minimization is taking one proximal gradient step, as used for example in the Linearized ADMM algorithm (Goldfarb et al., 2013; Yang and Yuan, 2013). We replace this step with a FW one.

FW Augmented Lagrangian method (FW-AL)

At each iteration $t \geq 1$, we first update the primal variable blocks $\boldsymbol{x}_t$ with a Frank-Wolfe step and then update the dual multiplier $\boldsymbol{y}_t$ using the updated primal variables:

$$\begin{cases} \boldsymbol{x}_{t+1} = \mathcal{FW}(\boldsymbol{x}_t; \mathcal{L}(\cdot, \boldsymbol{y}_t)) \ , \\ \boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \eta_t M \boldsymbol{x}_{t+1} \ , \end{cases} \tag{2}$$

where $\eta_t > 0$ is the step size for the dual update and $\mathcal{FW}$ is either Alg. 1 or Alg. 2.

We denote by $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ the iterates computed by FW-AL after t steps and by $\mathcal{A}_t$ the set of atoms previously given by the FW oracle (including the initialization point). If the constraint set is the convex hull of a set of atoms $\mathcal{A}$, the iterate $\boldsymbol{x}_t$ has a sparse representation as a convex combination of the first iterate and the atoms previously given by the FW oracle. The set of atoms which appear in this expansion with non-zero weight is called the *active set* $\mathcal{S}_t$. Similarly, $\boldsymbol{y}_t$ is by construction in the cone generated by $\{M\boldsymbol{x}_s\}_{s \leq t}$, that is, they both have the sparse expansion:

$$\boldsymbol{x}_t = \sum_{\boldsymbol{v} \in \mathcal{S}_t} \alpha_{\boldsymbol{v}}^{(t)} \boldsymbol{v}, \quad \text{and} \quad \boldsymbol{y}_t = \sum_{\boldsymbol{v} \in \mathcal{A}_t} \xi_{\boldsymbol{v}}^{(t)} M\boldsymbol{v} \ , \tag{1}$$

Algorithm 1 AFW: Lacoste-Julien and Jaggi (2015)

input: $(\boldsymbol{x}_t, \mathcal{S}_t, f)$
while $\gamma_{\max} < 1$ (away step) **do**
  $\boldsymbol{s}_t := \textsc{Lmo}\left(\nabla f(\boldsymbol{x}_t)\right)$
  $\boldsymbol{v}_t \in \arg\max_{\boldsymbol{v} \in \mathcal{S}_t} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v} \rangle$
  if $\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{v}_t + \boldsymbol{s}_t - 2\boldsymbol{x}_t \rangle \geq 0$ **then**
    $\boldsymbol{d}_t := \boldsymbol{s}_t - \boldsymbol{x}_t$ and $\gamma_{\max} := 1$
  else
    $\boldsymbol{d}_t := \boldsymbol{x}_t - \boldsymbol{v}_t$ and $\gamma_{\max} := \alpha_{\boldsymbol{v}_t}^{(t)} / (1 - \alpha_{\boldsymbol{v}_t}^{(t)})$
  end if
  Compute $\gamma_t \in \arg\min_{\gamma \in [0, \gamma_{\max}]} f\left(\boldsymbol{x}_t + \gamma \boldsymbol{d}_t\right)$
  Update $\boldsymbol{x}_{t+1} := \boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t$
  Update $\alpha_{\boldsymbol{v}}^{(t+1)}$ according to (1)
  Update $\mathcal{S}_{t+1} := \{\boldsymbol{v} \in \mathcal{A} \text{ s.t. } \alpha_{\boldsymbol{v}}^{(t+1)} > 0\}$
  Update $t \leftarrow t + 1$
end while
return: $(\boldsymbol{x}_{t+1}, \mathcal{S}_{t+1})$

When we choose to use the AFW Alg. 1 as inner loop algorithm, it can choose an *away* direction to remove mass from "bad" atoms in the active set, i.e. to reduce $\alpha_{\boldsymbol{v}}^{(t)}$ for some $\boldsymbol{v}$ (see L8 of Alg. 1). On the other hand, the maximal step size for an *away* step can be quite small ($\gamma_{\max} = \alpha_{\boldsymbol{v}}^{(t)} / 1 - \alpha_{\boldsymbol{v}}^{(t)}$, where $\alpha_{\boldsymbol{v}}^{(t)}$ is the weight of the away vertex in (1)), yielding to arbitrary small suboptimality progress when the line-search is truncated to such small step-sizes. A step removing an atom from the active set is called a *drop step* , Alg. 1 loops until it performs one.One of the issue of ALM is that it is a non feasible method and consequently the usual suboptimality convergence criterion is no longer a satisfying one (since it can be negative). In the following section we wonder what could be the quantities to look at in order to get a sufficient condition of convergence.

Convergence Measures. Variants of ALM update at each iteration both the primal variable and the dual variable. For the purpose of analyzing the popular ADMM algorithm, Hong and Luo (2017) introduced two positives quantities which they called primal and dual gaps that we re-use in the analysis of our algorithm. Let $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ be the current primal and dual variables after t iterations of the FW-AL algorithm (2), the primal and dual gaps are respectively defined as

Algorithm 2 FW: Frank and Wolfe (1956)

input: $(\boldsymbol{x}_t, f)$
Compute $\boldsymbol{r}_t = \nabla f(\boldsymbol{x}_t)$
Compute $\boldsymbol{s}_t := \arg\min_{\boldsymbol{s} \in \mathcal{X}} \langle \boldsymbol{s}, \boldsymbol{r}_t \rangle$
$\gamma_t \in \arg\min_{\gamma \in [0,1]} f(\boldsymbol{x}_t + \gamma(\boldsymbol{s}_t - \boldsymbol{x}_t))$
Update $\boldsymbol{x}_{t+1} := (1 - \gamma)\boldsymbol{x}_t + \gamma \boldsymbol{s}_t$
return: $\boldsymbol{x}_{t+1}$

$$\Delta_t^{(d)} := \max_{\boldsymbol{y} \in \mathbb{R}^d} d(\boldsymbol{y}) - d(\boldsymbol{y}_t) \ \text{ where } \ d(\boldsymbol{y}_t) := \min_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_t) \ \text{ and } \ \Delta_t^{(p)} := \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_t) - d(\boldsymbol{y}_t) \ . \tag{3}$$

Notice that $\Delta_t^{(p)}$ is *not* the suboptimality associated with the primal function $p(\cdot) := \max_{\boldsymbol{y} \in \mathbb{R}^d} \mathcal{L}(\cdot, \boldsymbol{y})$ (which is infinite for every $\boldsymbol{x}$ non feasible). In this paper, we also define $\Delta_t := \Delta_t^{(p)} + \Delta_t^{(d)}$. It is important to realize that since ALM is a non-feasible method, the standard convergence convex minimization certificates could become meaningless. In particular, the quantity $f(\boldsymbol{x}_t) - f^*$ might be negative since $\boldsymbol{x}_t$ does not necessarily belong to the constraint set of (OPT).

Convergence over general convex sets. The GDMM algorithm of (Yen et al., 2016a,b; Huang et al., 2017) relied on the assumption of $\mathcal{X}$ being polytope, hence we obtain from this sublinear decrease a completely new result on ALM with FW. This result covers the case of the simultaneously sparse and low rank matrices where the trace norm ball is not a polytope.

Theorem 1 (Rate of FW-AL with Alg. 2)**.** *If $\mathcal{X}$ is a compact convex set, f is a L-smooth convex function and if the affine hull of $\cap_{k=1}^{K}\mathcal{X}_k$ is equal to the intersection of the affine hull of $\mathcal{X}_k$ then using Alg.2 in* FW-AL *and $\eta_t := \min\left\{\frac{2}{\lambda}, \frac{\alpha^2}{2\delta}\right\}\frac{2}{t+2}$ implies that there exists a bounded $t_0 \geq 0$ such that,*

$$\Delta_t \leq \frac{4\delta(t_0+2)}{t+2}, \qquad \min_{t_0 \leq s-1 \leq t} \|M\boldsymbol{x}_s\|^2 \leq \frac{O(1)}{t-t_0+1}, \quad \forall t \geq t_0 \tag{4}$$

where $D := \max_{\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}} \|\boldsymbol{x}-\boldsymbol{x}'\|$ is the diameter of $\mathcal{X}$, $L_\lambda := L/K + 2\lambda$ the Lipschitz constant of $\mathcal{L}$ the AL function, $\delta := L_\lambda D^2$ and α is a positive constant depending on $\mathcal{X}$. Moreover if f is μ-strongly convex, the optimal set of (OPT) *is reduced to a point, $\mathcal{X}^* = \{\boldsymbol{x}^*\}$, and,*

$$\min_{t_0 \leq s \leq t} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 \leq \frac{4K}{\mu}\frac{O(1)}{t-t_0+1}, \forall t \geq t_0, \tag{5}$$

Some bounds on t_0 can be explicitly given with the constants introduced in this theorem.

Convergence over Polytopes. On the other hand, if $\mathcal{X}$ is a polytope, recent advances on FW proposed global linear convergence rates for a generalized strongly convex objective using FW with away steps (Lacoste-Julien and Jaggi, 2015; Garber and Meshi, 2016). More precisely, we will use the fact that Algorithm 1 performs *geometric decrease* (Lacoste-Julien and Jaggi, 2015, Theorem 1): for $\boldsymbol{x}^+ := \mathcal{FW}(\boldsymbol{x}; \mathcal{L}(\cdot, \boldsymbol{y}))$, there exists $\rho_A < 1$ such that for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathbb{R}^d$,

$$\mathcal{L}(\boldsymbol{x}^+, \boldsymbol{y}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) \leq \rho_A\big[\min_{\boldsymbol{x}' \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}', \boldsymbol{y}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})\big]. \tag{6}$$

The constant ρ_A (Lacoste-Julien and Jaggi, 2015) depends on the smoothness, the generalized strong convexity of $\mathcal{L}(\cdot, \boldsymbol{y})$ (does not depend on $\boldsymbol{y}$, but depends on M) and the condition number of the set $\mathcal{X}$ depending on its geometry.

Theorem 2 (Rate of FW-AL with inner loop Alg. 1)**.** *If $\mathcal{X}$ is a compact polytope, f is a L-smooth convex function and if the affine hull of $\cap_{k=1}^{K}\mathcal{X}_k$ is equal to the intersection of the affine hull of $\mathcal{X}_k$ then using Alg 1 as inner loop and a constant step size $\eta_t = \frac{\lambda\rho_A}{4}$, there exists $t_0 \in \mathbb{N}$ such that the quantity Δ_t decreases at least by a uniform amount for finite number of steps t_0 as,*

$$\Delta_{t+1} - \Delta_t \leq -\frac{\lambda\alpha^2\rho_A}{8}, \tag{7}$$

until $\Delta_{t_0} \leq L_\lambda D^2$. Then we have that the gap and the feasibility violation decrease linearly as,

$$\Delta_t \leq \frac{\Delta_{t_0}}{(1+\kappa)^{t-t_0}}, \quad \|M\boldsymbol{x}_{t+1}\|^2 \leq \frac{16}{\lambda \cdot \rho_A}\frac{\Delta_{t_0}}{(1+\kappa)^{t-t_0}}, \ \forall t \geq t_0,$$

where $\kappa := \frac{\rho_A}{2}\min\left\{1, \frac{\lambda\alpha^2}{4L_\lambda D^2}\right\}$, $L_\lambda := L/K + 2\lambda$ and α a constant depending on $\mathcal{X}$. Morever if f is μ-strongly convex, the optimal set of (OPT) *is reduced to a point, $\mathcal{X}^* = \{\boldsymbol{x}^*\}$, and,*

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leq \frac{2K\Delta_{t_0}(\sqrt{2}+1)}{\mu(\sqrt{1+\kappa})^{t-t_0}}, \ \forall t \geq t_0. \tag{8}$$

For an intersection of sets, the theorem above give stronger results than Yen et al. (2016b); Huang et al. (2017) since we prove that the distance to the optimal point as well as the feasibility condition vanish linearly.

References

J.-B. Alayrac, P. Bojanowski, N. Agrawal, I. Laptev, J. Sivic, and S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *CVPR*, 2016.

D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Athena Scientific, 1996.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 2011.

J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 2010.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics*, 1956.

D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 1976.

D. Garber and O. Meshi. Linear-memory and decomposition-invariant linearly convergent conditional gradient algorithm for structured polytopes. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *NIPS*, 2016.

R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis*, 1975.

D. Goldfarb, S. Ma, and K. Scheinberg. Fast alternating linearization methods for minimizing the sum of two convex functions. *Mathematical Programming*, 2013.

A. Gramfort, B. Thirion, and G. Varoquaux. Identifying predictive regions from fMRI with TV-L1 prior. In *International Workshop on Pattern Recognition in Neuroimaging*. IEEE, 2013.

M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction method of multipliers. *Math. Program.*, 2017.

X. Huang, I. E.-H. Yen, R. Zhang, Q. Huang, P. Ravikumar, and I. Dhillon. Greedy direction method of multiplier for MAP inference of large output domain. In *AISTATS*, 2017.

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, 2013.

S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. 2015.

S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-Coordinate Frank-Wolfe Optimization for Structural SVMs. In *ICML*, 2013.

G. Obozinski, L. Jacob, and J.-P. Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv:1110.0413*, 2011.

E. Richard, P.-A. Savalle, and N. Vayatis. Estimation of simultaneously sparse and low rank matrices. In *ICML*, 2012.

M. Yan and W. Yin. Self equivalence of the alternating direction method of multipliers. In *Splitting Methods in Communication, Imaging, Science, and Engineering*. Springer, 2016.

J. Yang and X. Yuan. Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of computation*, 2013.

I. Yen, X. Huang, K. Zhong, R. Zhang, P. Ravikumar, and I. Dhillon. Dual decomposed learning with factorwise oracle for structural SVM with large output domain. In *NIPS*, 2016b.

I. E.-H. Yen, X. Lin, J. Zhang, P. Ravikumar, and I. Dhillon. A convex atomic-norm approach to multiple sequence alignment and motif discovery. In *ICML*, 2016a.