Discrete-Continuous ADMM for Transductive Inference in Higher-Order MRFs –Supplementary Material–

Emanuel Laude [1] Jan-Hendrik Lange [2] Jonas Schüpfer [1] Csaba Domokos [1]
Laura Leal-Taixé [1] Frank R. Schmidt [1 3] Bjoern Andres [2 3 4] Daniel Cremers [1]

[1] Technical University of Munich [2] Max Planck Institute for Informatics, Saarbrücken
[3] Bosch Center for Artificial Intelligence [4] University of Tübingen

1. Theoretical Results

In the remainder of this section we make use of the following properties of L-smooth functions (known as the descent-lemma) and m-semiconvexity [1, 8], which are standard results and therefore stated without proof:

Lemma. *Let $f : \mathbb{R}^{k\times n} \to \mathbb{R}$ be continuously differentiable and let $x, y \in \mathbb{R}^{k\times n}$.*

- *If f is L-smooth (meaning that ∇f is Lipschitz continuous with modulus L), then*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x - y\|_F^2. \quad (1)$$

- *If f is m-semiconvex (meaning that $f + \frac{m}{2}\|\cdot\|_F^2$ is convex), then*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle - \frac{m}{2}\|x - y\|_F^2. \quad (2)$$

For showing convergence we make the following assumptions on our problem:

- The function f is L-smooth, m-semiconvex and lower-bounded.
- For all $y_i \in \mathcal{L}$, $\ell(y_i; \cdot)$ is lower-bounded.
- The kernel matrix $K \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is surjective, i.e. the smallest eigenvalue $\sigma_{\min}(K^\top K) > 0$ is positive.
- After finitely many iterations the penalty parameter ρ is sufficiently large and kept fixed such that

$$\frac{L^2}{\rho\sigma_{\min}(K^\top K)} + \frac{m - \rho\sigma_{\min}(K^\top K)}{2} < 0. \quad (3)$$

1.1. Proof of Lemma 1

In [7, 6], to show convergence of nonconvex ADMM, a monotonic decrease of the augmented Lagrangian is guaranteed. Following a similar line of argument, we show that the "discrete-continuous" augmented Lagrangian

$$\begin{aligned}\mathfrak{L}_\rho(\alpha, \beta, \lambda, y) := &\sum_{i\in\mathcal{V}} \ell(y_i; \beta_i) + f(\alpha) \\ &+ \sum_{C\in\mathcal{C}} E_C(y) + \langle \lambda, K\alpha - \beta\rangle + \frac{\rho}{2}\|K\alpha - \beta\|_F^2.\end{aligned} \quad (4)$$

monotonically decreases with the iterates. Whereas its value decreases with the primal and discrete variable updates, the dual update yields a positive contribution to the overall estimate. Yet, for $\rho > 0$ chosen large enough, K surjective and f being L-smooth, this ascent can be dominated by a sufficiently large descent in the primal block α, updated last.

We need the following notation. Let $B^{t+1}_{:,y^t}$ denote the matrix whose i-th row is given by B^{t+1}_{i,y_i^t}. In particular, by definition of the β update, this means $\beta^{t+1} = B^{t+1}_{:,y^{t+1}}$.

Lemma. *Let $K \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ be surjective and $\delta \geq 0$. For ρ meeting condition (3) we have that*

1. *The discrete-continuous augmented Lagrangian (4) decreases monotonically with the iterates $(\alpha^t, \beta^t, \lambda^t, y^t)$:*

$$\begin{aligned}&\mathfrak{L}_\rho(\alpha^{t+1}, \beta^{t+1}, \lambda^{t+1}, y^{t+1}) - \mathfrak{L}_\rho(\alpha^t, \beta^t, \lambda^t, y^t) \\ &\leq \left(\tfrac{L^2}{\rho\sigma_{\min}(K^\top K)} + \tfrac{m-\rho\sigma_{\min}(K^\top K)}{2}\right)\|\alpha^{t+1} - \alpha^t\|_F^2 \\ &\qquad - \delta[\![y^{t+1} \neq y^t]\!], \quad (5)\end{aligned}$$

   where $[\![\cdot]\!]$ denotes the Iverson bracket.

2. *$\{\mathfrak{L}_\rho(\alpha^{t+1}, \beta^{t+1}, \lambda^{t+1}, y^{t+1})\}_{t\in\mathbb{N}}$ is lower bounded.*

3. $\{\mathfrak{L}_\rho(\alpha^{t+1},\beta^{t+1},\lambda^{t+1},y^{t+1})\}_{t\in\mathbb{N}}$ *converges.*

Proof. We rewrite the difference of two consecutive "discrete-continuous" augmented Lagrangians as

$$\begin{aligned}
&\mathfrak{L}_\rho(\alpha^{t+1},\beta^{t+1},\lambda^{t+1},y^{t+1}) - \mathfrak{L}_\rho(\alpha^t, B^t_{:,y^t},\lambda^t,y^t)\\
&= \mathfrak{L}_\rho(\alpha^t, B^{t+1}_{:,y^t},\lambda^t,y^t) - \mathfrak{L}_\rho(\alpha^t, B^t_{:,y^t},\lambda^t,y^t)\\
&\quad + \mathfrak{L}_\rho(\alpha^t,\beta^{t+1},\lambda^t,y^{t+1}) - \mathfrak{L}_\rho(\alpha^t, B^{t+1}_{:,y^t},\lambda^t,y^t)\\
&\quad + \mathfrak{L}_\rho(\alpha^{t+1},\beta^{t+1},\lambda^t,y^{t+1}) - \mathfrak{L}_\rho(\alpha^t,\beta^{t+1},\lambda^t,y^{t+1})\\
&\quad + \mathfrak{L}_\rho(\alpha^{t+1},\beta^{t+1},\lambda^{t+1},y^{t+1})\\
&\quad - \mathfrak{L}_\rho(\alpha^{t+1},\beta^{t+1},\lambda^t,y^{t+1})
\end{aligned}$$

We now bound each of the four differences separately:

Since the augmented Lagrangian is separable in β and we solve for any y_i a minimization problem in β_{y_i} globally optimal we have that

$$\mathfrak{L}_\rho(\alpha^t, B^{t+1}_{:,y^t},\lambda^t,y^t) - \mathfrak{L}_\rho(\alpha^t, B^t_{:,y^t},\lambda^t,y^t) \le 0. \tag{6}$$

A similar estimate holds for the the discrete variable y^{t+1} due to the update in the algorithm:

$$\begin{aligned}
&\mathfrak{L}_\rho(\alpha^t,\beta^{t+1},\lambda^t,y^{t+1}) - \mathfrak{L}_\rho(\alpha^t, B^{t+1}_{:,y^t},\lambda^t,y^t)\\
&\qquad \le -\delta[\![y^{t+1}\neq y^t]\!].
\end{aligned} \tag{7}$$

Now we devise a bound for the third term given by

$$\begin{aligned}
&\mathfrak{L}_\rho(\alpha^{t+1},\beta^{t+1},\lambda^t,y^{t+1}) - \mathfrak{L}_\rho(\alpha^t,\beta^{t+1},\lambda^t,y^{t+1})\\
&= f(\alpha^{t+1}) - f(\alpha^t) + \langle K\alpha^{t+1} - K\alpha^t, \lambda^t\rangle\\
&\qquad + \frac{\rho}{2}\|K\alpha^{t+1}-\beta^{t+1}\|_F^2 - \frac{\rho}{2}\|K\alpha^t-\beta^{t+1}\|_F^2.
\end{aligned}$$

We apply the identity $\|a+c\|_F^2 - \|b+c\|_F^2 = -\|b-a\|_F^2 + 2\langle a+c, a-b\rangle$ with $a := K\alpha^{t+1}$, $b := K\alpha^t$ and $c = -\beta^{t+1}$ and obtain

$$\begin{aligned}
&f(\alpha^{t+1}) - f(\alpha^t) - \frac{\rho}{2}\|K\alpha^{t+1}-K\alpha^t\|_F^2\\
&\qquad + \langle K\alpha^{t+1}-K\alpha^t, \lambda^t + \rho(K\alpha^{t+1}-\beta^{t+1})\rangle.
\end{aligned}$$

The optimality condition for the update of the variable α is given as

$$0 = \nabla f(\alpha^{t+1}) + K^\top(\rho(K\alpha^{t+1}-\beta^{t+1}) + \lambda^t). \tag{8}$$

We replace the term $\langle K\alpha^{t+1}-K\alpha^t, \lambda^t + \rho(K\alpha^{t+1}-\beta^{t+1})\rangle = \langle \alpha^{t+1}-\alpha^t, K^\top(\lambda^t + \rho(K\alpha^{t+1}-\beta^{t+1}))\rangle$ and obtain from the optimality condition of the α update that

$$\begin{aligned}
&f(\alpha^{t+1}) - f(\alpha^t) - \frac{\rho}{2}\|K\alpha^{t+1}-K\alpha^t\|_F^2\\
&\qquad + \langle \alpha^{t+1}-\alpha^t, -\nabla f(\alpha^{t+1})\rangle\\
&\le f(\alpha^{t+1}) - f(\alpha^t) - \frac{\rho\sigma_{\min}(K^\top K)}{2}\|\alpha^{t+1}-\alpha^t\|_F^2\\
&\qquad + \langle \alpha^t-\alpha^{t+1}, \nabla f(\alpha^{t+1})\rangle.
\end{aligned}$$

Moreover, due to the m-semiconvexity of the f we know that

$$\begin{aligned}
&f(\alpha^t) + \frac{m}{2}\|\alpha^{t+1}-\alpha^t\|_F^2\\
&\qquad \ge f(\alpha^{t+1}) + \langle \nabla f(\alpha^{t+1}), \alpha^t - \alpha^{t+1}\rangle.
\end{aligned}$$

Overall we can bound

$$\begin{aligned}
&\mathfrak{L}_\rho(\alpha^{t+1},\beta^{t+1},\lambda^t,y^{t+1}) - \mathfrak{L}_\rho(\alpha^t,\beta^{t+1},\lambda^t,y^{t+1})\\
&\qquad \le \frac{m-\rho\sigma_{\min}(K^\top K)}{2}\|\alpha^{t+1}-\alpha^t\|_F^2.
\end{aligned} \tag{9}$$

Since by assumption K is surjective, the smallest eigenvalue of $K^\top K$ is greater than zero: $\sigma_{\min}(K^\top K) > 0$. This means there exists some $\rho > 0$ large enough so that $\frac{m-\rho\sigma_{\min}(K^\top K)}{2} < 0$.

Finally, we estimate the last term:

$$\begin{aligned}
&\mathfrak{L}_\rho(\alpha^{t+1},\beta^{t+1},\lambda^{t+1},y^{t+1}) - \mathfrak{L}_\rho(\alpha^{t+1},\beta^{t+1},\lambda^t,y^{t+1})\\
&\qquad = \langle K\alpha^{t+1}-\beta^{t+1}, \lambda^{t+1}-\lambda^t\rangle = \frac{1}{\rho}\|\lambda^{t+1}-\lambda^t\|_F^2.
\end{aligned}$$

From the update of the dual variable and the optimality condition for the α update (8) it follows that

$$-\nabla f(\alpha^{t+1}) = K^\top\lambda^{t+1}. \tag{10}$$

Further, since f is L-smooth we know that

$$\|\nabla f(\alpha^{t+1}) - \nabla f(\alpha^t)\|_F^2 \le L^2\|\alpha^{t+1}-\alpha^t\|_F^2. \tag{11}$$

Overall, we obtain

$$\begin{aligned}
\sigma_{\min}(K^\top K)\|\lambda^{t+1}-\lambda^t\|_F^2 &\le \|K^\top\lambda^{t+1} - K^\top\lambda^t\|_F^2\\
&\le L^2\|\alpha^{t+1}-\alpha^t\|_F^2.
\end{aligned}$$

This gives the bound for the last term:

$$\begin{aligned}
&\mathfrak{L}_\rho(\alpha^{t+1},\beta^{t+1},\lambda^{t+1},y^{t+1}) - \mathfrak{L}_\rho(\alpha^{t+1},\beta^{t+1},\lambda^t,y^{t+1})\\
&\qquad \le \frac{L^2}{\rho\sigma_{\min}(K^\top K)}\|\alpha^{t+1}-\alpha^t\|_F^2.
\end{aligned}$$

Then, by merging the four estimates we obtain the desired result.

We proceed showing the lower boundedness of $\{\mathfrak{L}_\rho(\alpha^{t+1},\beta^{t+1},\lambda^{t+1},y^{t+1})\}_{t\in\mathbb{N}}$. Since K is surjective, there exists α' such that $K\alpha' = \beta^{t+1}$ and it holds that

$$-\frac{L}{2}\|\alpha^{t+1}-\alpha'\|_F^2 \ge -\frac{L}{2\sigma_{\min}(K^\top K)}\|K\alpha^{t+1}-K\alpha'\|_F^2.$$

Let $\rho > \frac{L}{\sigma_{\min}(K^\top K)}$. Then, since f is L-smooth, we have

$$\begin{aligned}
&f(\alpha^{t+1}) + \langle \lambda^{t+1}, K\alpha^{t+1} - \beta^{t+1}\rangle \\
&\quad + \frac{\rho}{2}\|K\alpha^{t+1} - \beta^{t+1}\|_F^2 \\
&= f(\alpha^{t+1}) + \langle K^\top\lambda^{t+1}, \alpha^{t+1} - \alpha'\rangle \\
&\quad + \frac{\rho}{2}\|K\alpha^{t+1} - \beta^{t+1}\|_F^2 \\
&= f(\alpha^{t+1}) + \langle \nabla f(\alpha^{t+1}), \alpha' - \alpha^{t+1}\rangle \\
&\quad + \frac{\rho}{2}\|K\alpha^{t+1} - \beta^{t+1}\|_F^2 \\
&\geq f(\alpha') - \frac{L}{2}\|\alpha^{t+1} - \alpha'\|_F^2 \\
&\quad + \frac{\rho}{2}\|K\alpha^{t+1} - \beta^{t+1}\|_F^2 \\
&\geq f(\alpha') + \frac{\rho\sigma_{\min}(K^\top K) - L}{2\sigma_{\min}(K^\top K)}\|K\alpha^{t+1} - \beta^{t+1}\|_F^2 \geq f(\alpha').
\end{aligned}$$

Overall, since by assumption f and $\ell(y_i;\cdot)$ are bounded from below (for all $y_i \in \mathcal{L}$), this means

$$\{\mathfrak{L}_\rho(\alpha^{t+1}, \beta^{t+1}, \lambda^{t+1}, y^{t+1})\}_{t\in\mathbb{N}}$$

is bounded from below.

Since $\{\mathfrak{L}_\rho(\alpha^{t+1}, \beta^{t+1}, \lambda^{t+1}, y^{t+1})\}_{t\in\mathbb{N}}$ is monotonically decreasing and bounded from below, $\{\mathfrak{L}_\rho(\alpha^{t+1}, \beta^{t+1}, \lambda^{t+1}, y^{t+1})\}_{t\in\mathbb{N}}$ converges. This completes the proof. □

1.2. Proof of Lemma 2

Lemma. *Let $\{(\alpha^t, \beta^t, \lambda^t, y^t)\}_{t\in\mathbb{N}}$ be the iterates produced by Algorithm 1. Then $\{(\alpha^t, \beta^t, \lambda^t, y^t)\}_{t\in\mathbb{N}}$ is a bounded sequence. Furthermore, for $t \to \infty$ the distance of two consecutive continuous iterates vanishes, and feasibility is achieved in the limit:*

$$\|\alpha^{t+1} - \alpha^t\|_F \to 0, \tag{12}$$
$$\|\beta^{t+1} - \beta^t\|_F \to 0, \tag{13}$$
$$\|\lambda^{t+1} - \lambda^t\|_F \to 0, \tag{14}$$
$$\|K\alpha^{t+1} - \beta^{t+1}\|_F \to 0. \tag{15}$$

Finally, if $\delta > 0$ is chosen strictly positive, then there exists some $T \in \mathbb{N}$ such $y^{t+1} = y^t$ for all $t > T$.

Proof. We sum over the estimate (5) which yields

$$\begin{aligned}
-\infty &< \lim_{t\to\infty} \mathfrak{L}_\rho(\alpha^t, \beta^t, \lambda^t, y^t) - \mathfrak{L}_\rho(\alpha^1, \beta^1, \lambda^1, y^1) \\
&\leq \sum_{t=1}^{\infty} \left(\tfrac{L^2}{\rho\sigma_{\min}(K^\top K)} + \tfrac{m - \rho\sigma_{\min}(K^\top K)}{2}\right)\|\alpha^{t+1} - \alpha^t\|_F^2 \\
&\quad - \sum_{t=1}^{\infty} \delta[\![y^{t+1} \neq y^t]\!]
\end{aligned}$$

Due to the lowerboundedness, the infinite sums have to converge. This yields that $\|\alpha^{t+1} - \alpha^t\|_F \to 0$. Since $0 \leq \sigma_{\min}(K^\top K)\|\lambda^{t+1} - \lambda^t\|_F^2 \leq \frac{L^2}{\rho\sigma_{\min}(K^\top K)}\|\alpha^{t+1} - \alpha^t\|_F^2$ and $\sigma_{\min}(K^\top K) > 0$ also $\|\lambda^{t+1} - \lambda^t\|_F \to 0$. Since due to the dual update $\lambda^{t+1} - \lambda^t = \rho(K\alpha^{t+1} - \beta^{t+1})$, also $\|K\alpha^{t+1} - \beta^{t+1}\|_F \to 0$. Moreover, it holds that

$$\begin{aligned}
\|\beta^{t+1} - \beta^t\|_F &\leq \|\beta^{t+1} - K\alpha^{t+1}\|_F + \|K\alpha^{t+1} - K\alpha^t\|_F \\
&\quad + \|K\alpha^t - \beta^t\|_F \\
&\leq \|K\alpha^{t+1} - \beta^{t+1}\|_F \\
&\quad + \|K\|\|\alpha^{t+1} - \alpha^t\|_F \\
&\quad + \|K\alpha^t - \beta^t\|_F \to 0
\end{aligned}$$

for $t \to \infty$.

Finally, suppose that there exists an infinite subsequence $\{t_j\}_{j=1}^\infty \subset \{t\}_{t=1}^\infty$ so that $y^{t_j+1} \neq y^{t_j}$. The last sum rewrites as,

$$\sum_{t=1}^{\infty} \delta[\![y^{t+1} \neq y^t]\!] = \sum_{j=1}^{\infty} \delta$$

which diverges for $\delta > 0$ positive. This however contradicts the lower boundedness of $\mathfrak{L}_\rho(\alpha^t, \beta^t, \lambda^t, y^t)$. □

1.3. Proof of Proposition 1

Definition ("Discrete-continuous" critical point)**.** *We call $(\alpha^*, \beta^*, \lambda^*, y^*)$ a "discrete-continuous" critical point of the discrete-continuous augmented Lagrangian (8) if it satisfies*

$$0 \in \partial(\ell(y_i^*;\cdot))(\beta_i^*) - \lambda_i^*, \qquad \forall\, i \tag{16}$$
$$0 = \nabla f(\alpha^*) + K^\top\lambda^* \tag{17}$$
$$K\alpha^* = \beta^*, \tag{18}$$

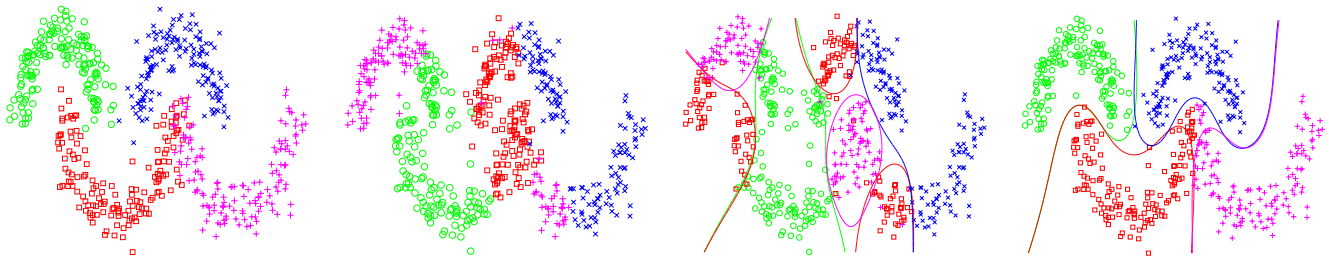
for y^* with $E_C(y^*) < \infty$ for all $C \in \mathcal{C}$. Here, $\partial f(x)$ denotes the "limiting" subdifferential [9, Definition 8.3] of the function f at x with $f(x) < \infty$.

Proposition. *Let $\delta \geq 0$. Then any limit point $(\alpha^*, \beta^*, \lambda^*, y^*)$ of the sequence $\{(\alpha^t, \beta^t, \lambda^t, y^t)\}_{t\in\mathbb{N}}$ is a "discrete-continuous" critical point.*

Proof. Let $(\alpha^*, \beta^*, \lambda^*, y^*)$ be a limit point of $\{(\alpha^t, \beta^t, \lambda^t, y^t)\}_{t\in\mathbb{N}}$, and let $\{t_j\}_{j=1}^\infty \subset \{t\}_{t=1}^\infty$ be the corresponding subsequence of indices. The optimality conditions for the update of the variables β_i (for any i) and α are given as:

$$0 \in \partial\ell(y_i^{t_j}; \beta_i^{t_j}) - \rho(K_i\alpha^{t_j-1} - \beta_i^{t_j} + {}^1\!/\!_\rho\lambda_i^{t_j-1}) \tag{19}$$
$$0 = \nabla f(\alpha^{t_j}) + \rho K^\top(K\alpha^{t_j} - \beta^{t_j} + {}^1\!/\!_\rho\lambda^{t_j-1}). \tag{20}$$

Passing the limit $j \to \infty$ and applying Lemma 2 we arrive at conditions (16)–(18). This completes the proof. □



(a) (b) (c) (d)

Figure 1: Form left to right: Ground-truth, RBF-kernel-k-means, coordinate descent, proposed method. The label inference errors are 66.6% for constrained RBF-kernel k-means, 68.5% for coordinate descent and 2.5% for our method.

1.4. Proof of Proposition 2

Proposition. *Let $\ell(y_i;\cdot)$ and f be proper, convex and lower-semicontinuous and let $\delta > 0$. Then the sequence $\{(\alpha^t, \beta^t, \lambda^t, y^t)\}_{t\in\mathbb{N}}$ produced by Algorithm 1 converges to a "discrete-continuous" critical point of (8) and α^* solves the problem (27) to global optimality.*

Proof. Let $\delta > 0$. Then, due to Lemma 2 the discrete variable converges, i.e. there is $T > 0$ so that for all $t > T$

$$y^{t+1} = y^t. \tag{21}$$

Then, since f and $\ell(y_i;\cdot)$ are convex proper and lsc., after finitely many iterations our scheme Alg. 1 reduces to convex ADMM and the global convergence is a direct consequence of [5, 4, 3]. This completes the proof. □

2. Additional Experimental Results

2.1. Proof of Concept

As a proof of concept we conduct a synthetic experiment with data sampled from 2D moon-shape distributions (600 samples, 4 classes, 150 per class). We sample 25 (possibly overlapping) cliques $C \subset \mathcal{V}$ of cardinality 25 from the set of examples. The synthetic labeling prior in this experiment is given in terms of constraints, that balance the label assignment within each clique. More precisely, it restricts the maximal deviation of the determined labeling from the true labeling to a given bound within each clique $C \in \mathcal{C}$. Mathematically, the higher order energies E_C in the MRF are defined so that $E_C(y_C) = 0$ if $L_C^j \leq |\{i \in C : y_C^i = j\}| \leq U_C^j$, and ∞ otherwise. The bounds L_C^j and U_C^j are fixed and chosen a-priori, such that the number of samples $i \in C$ assigned to class j deviates by at most 3 from the true number within clique C. This means that we do not provide any exact labels to the algorithm.

The overall task is to infer the correct labels from both, the distribution of the examples in the feature space, and the combinatorial prior encoded within the higher order energies. Within the algorithm, we solve the LP-relaxation of the higher order MRF-subproblem (14) with the dual-simplex method and threshold the solution. On this task, we compare our method to constrained kernel k-means and plain discrete-continuous coordinate descent on (5) with an RBF kernel and an SVM-loss (see Figure 1). Like [10, 11, 12, 2], we apply k-means in the RBF-kernel space and solve the E-step w.r.t. to (14). It can be seen that both, coordinate descent on the SVM-based-model (Figure 1c) and constrained kernel k-means (Figure 1b) get stuck in poor local minima. In contrast, our method is able to infer the correct labels of most examples and finds a reasonable classifier, even for a trivial initialization of the parameters, cf. Figure 1d. The label errors are 66.6% for constrained RBF-kernel k-means, 68.5% for coordinate descent and 2.5% for our method.

References

[1] M. Artina, M. Fornasier, and F. Solombrino. Linearly constrained nonsmooth and nonconvex minimization. *SIAM Journal on Optimization*, 23(3):1904–1937, 2013. 1

[2] S. Basu, I. Davidson, and K. Wagstaff. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC, 1 edition, 2008. 4

[3] S. P. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 4

[4] J. Eckstein and D. P. Bertsekas. On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992. 4

[5] D. Gabay. Applications of the method of multipliers to variational inequalities. *Studies in Mathematics and Its Applications*, 15:299–331, 1983. 4

[6] M. Hong, Z.-Q. Luo, and M. Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1):337–364, 2016. 1

[7] G. Li and T. K. Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015. 1

[8] T. Möllenhoff, E. Strekalovskiy, M. Möller, and D. Cremers. The primal-dual hybrid gradient method for semi-convex splittings. *SIAM Journal on Imaging Sciences*, 8(2):827–857, 2015. 1

[9] R. Rockafellar and R.-B. Wets. *Variational Analysis*. Springer, 1998. 3

[10] M. Tang, I. B. Ayed, D. Marin, and Y. Boykov. Secrets of grabcut and kernel k-means. In *IEEE International Conference on Computer Vision, ICCV*, pages 1555–1563, 2015. 4

[11] M. Tang, D. Marin, I. B. Ayed, and Y. Boykov. Normalized cut meets MRF. In *European Conference on Computer Vision, ECCV*, pages 748–765, 2016. 4

[12] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In C. E. Brodley and A. P. Danyluk, editors, *Proceedings of the 18th International Conference on Machine Learning, ICML*, pages 577–584. Morgan Kaufmann, 2001. 4