Appendix to "Online ℓ_1-Dictionary Learning with Application to Novel Document Detection"

Shiva Prasad Kasiviswanathan **Huahua Wang** **Arindam Banerjee**

Prem Melville

A Background about ADMM

In this section, we give a brief review of the general framework of ADMM. ADMM has recently gathered significant attention in the machine learning community due to its wide applicability to a range of learning problems with complex objective functions [1, 2].

Let $p(\mathbf{x}) : \mathbb{R}^a \to \mathbb{R}$ and $q(\mathbf{y}) : \mathbb{R}^b \to \mathbb{R}$ be convex functions, $F \in \mathbb{R}^{c\times a}$, $G \in \mathbb{R}^{c\times b}$, and $\mathbf{z} \in \mathbb{R}^c$. Consider the following optimization problem

$$\min_{\mathbf{x},\mathbf{y}} \; p(\mathbf{x}) + q(\mathbf{y}) \text{ s.t. } F\mathbf{x} + G\mathbf{y} = \mathbf{z}, \tag{1}$$

where the variable vectors $\mathbf{x}$ and $\mathbf{y}$ are separate in the objective, and coupled only in the constraint. The augmented Lagrangian for the above problem is given by

$$\mathcal{L}(\mathbf{x}, \mathbf{y}, \rho) = p(\mathbf{x}) + q(\mathbf{y}) + \rho^\top(\mathbf{z} - F\mathbf{x} - G\mathbf{y}) + \frac{\varphi}{2}\|\mathbf{z} - F\mathbf{x} - G\mathbf{y}\|_2^2 \ ,$$

where $\rho \in \mathbb{R}^c$ is the Lagrangian multiplier and $\varphi > 0$ is a penalty parameter. ADMM utilizes the separability form of (1) and replaces the joint minimization over $\mathbf{x}$ and $\mathbf{y}$ with two simpler problems. The ADMM first minimizes $\mathcal{L}$ over $\mathbf{x}$, then over $\mathbf{y}$, and then applies a proximal minimization step with respect to the Lagrange multiplier ρ. The entire ADMM procedure is summarized in Algorithm 1. The $\gamma > 0$ is a constant. The subscript (i) denotes the ith iteration of the ADMM procedure. The ADMM procedure has been proved to converge to the global optimal solution under quite broad conditions [2].

Algorithm 1 : ADMM Update Equations for Solving (1)

Iterate until *convergence*

$$\begin{cases} \mathbf{x}_{(i+1)} \leftarrow \text{argmin}_{\mathbf{x}} \; \mathcal{L}(\mathbf{x}, \mathbf{y}_{(i)}, \rho_{(i)}), \\ \mathbf{y}_{(i+1)} \leftarrow \text{argmin}_{\mathbf{y}} \; \mathcal{L}(\mathbf{x}_{(i+1)}, \mathbf{y}, \rho_{(i)}), \\ \rho_{(i+1)} \leftarrow \rho_{(i)} + \gamma\varphi(\mathbf{z} - F\mathbf{x}_{(i+1)} - G\mathbf{y}_{(i+1)}). \end{cases}$$

A.1 ADMM Equations for updating X and A's

Consider the ℓ_1-dictionary learning problem

$$\min_{A\in\mathcal{A}, X\geq\mathbf{0}} \; \|P - AX\|_1 + \lambda\|X\|_1,$$

where $\mathcal{A}$ is defined in Section 3.1. We use the following algorithm from [4] to solve this problem. It is quite easy to adapt the ADMM updates outlined in Algorithm 1 to update X's and A's, when the other variable is fixed (see e.g., [4]).

ADMM for updating X, given fixed A. Here we are given matrices $P \in \mathbb{R}^{m\times n}$ and $A \in \mathbb{R}^{m\times k}$, and we want to solve the following optimization problem

$$\min_{X\geq\mathbf{0}} \; \|P - AX\|_1 + \lambda\|X\|_1.$$

Algorithm 2 shows the ADMM update steps for solving this problem. The entire derivation is presented in [4] and we are reproducing them here for completeness. In our experiments, we set

$\varphi = 5$, $\kappa = 1/\Psi_{\max}(A)$, and $\gamma = 1.89$. These parameters are chosen based on the ADMM convergence results presented in [4, 6].

Algorithm 2 : ADMM for Updating X

ADMM procedure for solving $\min_{X \geq 0} \ \|P - AX\|_1 + \lambda \|X\|_1$
Input: $A \in \mathbb{R}^{m \times k}$, $P \in \mathbb{R}^{m \times n}$, $\lambda \geq 0$, $\gamma \geq 0$, $\psi \geq 0$, $\kappa \geq 0$
$X_{(1)} \leftarrow \mathbf{0}_{k \times n}$, $E_{(1)} \leftarrow P$, $\rho_{(1)} \leftarrow \mathbf{0}_{m \times n}$
for $i = 1, 2, \ldots,$ **to** *convergence* **do**
$E_{(i+1)} \leftarrow \text{soft}(P - AX_{(i)} + \rho_{(i)}/\varphi, 1/\varphi)$
$G \leftarrow A^\top (AX_{(i)} + E_{(i+1)} - P - \rho_{(i)}/\varphi)$
$X_{(i+1)} \leftarrow \max\{X_{(i)} - \kappa G - (\lambda\kappa)/\varphi, 0\}$
$\rho_{(i+1)} \leftarrow \rho_{(i)} + \gamma\varphi(P - AX_{(i+1)} - E_{(i+1)})$
Return X at convergence

ADMM for Updating A, given fixed X. Given inputs $P \in \mathbb{R}^{m \times n}$ and $X \in \mathbb{R}^{k \times n}$, consider the following optimization problem

$$\min_{A \in \mathcal{A}} \ \|P - AX\|_1.$$

When repeating this optimization over multiple timesteps, we use warm starts for faster convergence, i.e., instead of initializing $A_{(1)}$ to $\mathbf{0}_{m \times k}$, we initialize $A_{(1)}$ to the dictionary obtained at the end of the previous timestep.

Algorithm 3 : ADMM for Updating A

ADMM procedure for solving $\min_{A \in \mathcal{A}} \ \|P - AX\|_1$
Input: $X \in \mathbb{R}^{k \times n}$, $P \in \mathbb{R}^{m \times n}$, $\gamma \geq 0$, $\psi \geq 0$, $\kappa \geq 0$
$A_{(1)} \leftarrow \mathbf{0}_{m \times k}$, $E_{(1)} \leftarrow P$, $\rho_{(1)} \leftarrow \mathbf{0}_{m \times n}$
for $i = 1, 2, \ldots,$ **to** *convergence* **do**
$E_{(i+1)} \leftarrow \text{soft}(P - A_{(i)}X + \rho_{(i)}/\varphi, 1/\varphi)$
$G \leftarrow (A_{(i)}X + E_{(i+1)} - P - \rho_{(i)}/\varphi)X^\top$
$A_{(i+1)} \leftarrow \Pi_{\mathcal{A}}(\max\{A_{(i)} - \kappa G, 0\})$
$\rho_{(i+1)} \leftarrow \rho_{(i)} + \gamma\varphi(P - A_{(i+1)}X - E_{(i+1)})$
Return A at convergence

B Analysis of OIADMM: Proofs from Section 4

First, lets recap the OIADMM update rules.

$$\Gamma_{t+1} = \operatorname*{argmin}_{\Gamma} \|\Gamma\|_1 + \langle \Delta_t, \widetilde{\Gamma}_t - \Gamma \rangle + \frac{\beta_t}{2} \|\widetilde{\Gamma}_t - \Gamma\|_F^2. \tag{2}$$

$$\hat{A}_{t+1} = \operatorname*{argmin}_{A \in \mathcal{A}} \beta_t (\langle G_{t+1}, A - \hat{A}_t \rangle + \frac{1}{2\tau_t} \|A - \hat{A}_t\|_F^2), \tag{3}$$

$$\Delta_{t+1} = \Delta_t + \beta_t (P_t - \hat{A}_{t+1} X_t - \Gamma_{t+1}). \tag{4}$$

Let A^{opt} be the optimum solution to (the batch problem)

$$\min_{A \in \mathcal{A}} \sum_{t=1}^{T} \|P_t - AX_t\|_1.$$

Let $\widetilde{\Gamma}_t = P_t - \hat{A}_t \hat{X}_t$ and $\widehat{\Gamma}_t = P_t - \hat{A}_{t+1} \hat{X}_t$. For any, $A^\star \in \mathcal{A}$, let $\Gamma_t^\star = P_t - A^\star \hat{X}_t$. The lemmas below hold for any $A^\star \in \mathcal{A}$ so in particular it holds for $A^\star$ set as A^{opt}.

Proof Flow. Although the algorithm is relatively simple, the analysis is somewhat involved. Define, $\overline{\Gamma}_t^{\text{opt}} = P_t - A^{\text{opt}} X_t$. Then the regret of the OIADMM is

$$R(T) = \sum_{t=1}^{T} \|\widetilde{\Gamma}_t\|_1 - \|\overline{\Gamma}_t^{\text{opt}}\|_1.$$

We split the proof into three technical lemmas. We first upper bound $\langle \Delta_t, \widehat{\Gamma}_t - \Gamma_t^\star \rangle$ (Lemma B.2), and use it to bound $\|\Gamma_{t+1}\|_1 - \|\Gamma_t^\star\|_1$ (Lemma B.3). In the proof of Lemma B.4, we bound $\|\widetilde{\Gamma}_t\|_1 - \|\Gamma_{t+1}\|_1$ and this when added to the bound on $\|\Gamma_{t+1}\|_1 - \|\Gamma_t^\star\|_1$ (from Lemma B.3) gives a bound

on $\|\widetilde{\Gamma}_t\|_1 - \|\Gamma_t^\star\|_1$. The proof of the regret bound uses a canceling telescoping sum on the bound on $\|\widetilde{\Gamma}_t\|_1 - \|\Gamma_t^\star\|_1$.

We use the following simple inequality in our proofs.

Lemma B.1. *For matrices $M_1, M_2, M_3, M_4 \in \mathbb{R}^{m\times n}$, we have the following*
$2\langle M_1 - M_2, M_3 - M_4\rangle = \|M_1 - M_4\|_F^2 + \|M_2 - M_3\|_F^2 - \|M_1 - M_3\|_F^2 - \|M_2 - M_4\|_F^2$.

Lemma B.2. *Let $\{\Gamma_t, \hat{A}_t, \Delta_t\}$ be the sequences generated by the* OIADMM *procedure. For any $A^\star \in \mathcal{A}$, we have*

$$\langle \Delta_t, \widehat{\Gamma}_t - \Gamma_t^\star\rangle \le \frac{\beta_t}{2\tau_t}\left(\|A^\star - \hat{A}_t\|_F^2 - \|A^\star - \hat{A}_{t+1}\|_F^2\right)$$
$$+\frac{\beta_t}{2}\left(\|\Gamma_t^\star - \Gamma_{t+1}\|_F^2 - \|\Gamma_{t+1} - \widehat{\Gamma}_t\|_F^2 - \|\Gamma_t^\star - \widetilde{\Gamma}_t\|_F^2\right) - \frac{\beta_t}{2}\left(\frac{1}{\tau_t} - \Psi_{\max}(\hat{X}_t)\right)\|\hat{A}_{t+1} - \hat{A}_t\|_F^2.$$

Proof. For any $A^\star \in \mathcal{A}$, (3) is equivalent to the following variational inequality [5]:

$$\beta_t\langle G_{t+1} + \frac{1}{\tau_t}(\hat{A}_{t+1} - \hat{A}_t), A^\star - \hat{A}_{t+1}\rangle \ge 0. \tag{5}$$

Using $\widehat{\Gamma}_t = P_t - \hat{A}_{t+1}\hat{X}_t$ and substituting for G_{t+1}, we have

$$\begin{aligned}\beta_t\langle G_{t+1}, A^\star - \hat{A}_{t+1}\rangle &= -\beta_t\langle(\Delta_t/\beta_t + \widetilde{\Gamma}_t - \Gamma_{t+1})\hat{X}_t^\top, A^\star - \hat{A}_{t+1}\rangle \\ &= \beta_t\langle \Delta_t/\beta_t + \widetilde{\Gamma}_t - \Gamma_{t+1}, \hat{A}_{t+1}\hat{X}_t - A^\star\hat{X}_t\rangle \\ &= \beta_t\langle \Delta_t/\beta_t + \widetilde{\Gamma}_t - \Gamma_{t+1}, P_t - A^\star\hat{X}_t - (P_t - \hat{A}_{t+1}\hat{X}_t)\rangle \\ &= \langle \Delta_t, \Gamma_t^\star - \widehat{\Gamma}_t\rangle + \beta_t\langle\widetilde{\Gamma}_t - \Gamma_{t+1}, \Gamma_t^\star - \widehat{\Gamma}_t\rangle. \end{aligned} \tag{6}$$

Substituting (6) into (5) and rearranging the terms yield

$$\langle \Delta_t, \widehat{\Gamma}_t - \Gamma_t^\star\rangle \le \beta_t\langle\widetilde{\Gamma}_t - \Gamma_{t+1}, \Gamma_t^\star - \widehat{\Gamma}_t\rangle + \frac{\beta_t}{\tau_t}\langle\hat{A}_{t+1} - \hat{A}_t, A^\star - \hat{A}_{t+1}\rangle. \tag{7}$$

By using Lemma B.1, the first term on the right side can be rewritten as

$$\langle\widetilde{\Gamma}_t - \Gamma_{t+1}, \Gamma_t^\star - \widehat{\Gamma}_t\rangle = \frac{1}{2}\big(\|\widetilde{\Gamma}_t - \widehat{\Gamma}_t\|_F^2 + \|\Gamma_t^\star - \Gamma_{t+1}\|_F^2 - \|\Gamma_{t+1} - \widehat{\Gamma}_t\|_F^2 - \|\Gamma_t^\star - \widetilde{\Gamma}_t\|_F^2\big). \tag{8}$$

Substituting the definitions of $\widehat{\Gamma}_t$ and $\widetilde{\Gamma}_t$, we have

$$\|\widetilde{\Gamma}_t - \widehat{\Gamma}_t\|_F^2 = \|P_t - \hat{A}_t\hat{X}_t - (P_t - \hat{A}_{t+1}\hat{X}_t)\|_F^2 = \|(\hat{A}_{t+1} - \hat{A}_t)\hat{X}_t\|_F^2 \le \Psi_{\max}(\hat{X}_t)\|\hat{A}_{t+1} - \hat{A}_t\|_F^2, \tag{9}$$

Remember that $\Psi_{\max}(\hat{X}_t)$ is the maximum eigenvalue of $X^\top X$. Using Lemma B.1, we get that the second term in the right hand side of (7) is equivalent to

$$\langle\hat{A}_{t+1} - \hat{A}_t, A^\star - \hat{A}_{t+1}\rangle = \frac{1}{2}\left(\|A^\star - \hat{A}_t\|_F^2 - \|A^\star - \hat{A}_{t+1}\|_F^2 - \|\hat{A}_{t+1} - \hat{A}_t\|_F^2\right). \tag{10}$$

Combining results in (7), (8), (9), and (10), we get the desired bound. □

Lemma B.3. *Let $\{\Gamma_t, \hat{A}_t, \Delta_t\}$ be the sequences generated by the OIADMM procedure. For any $A^\star \in \mathcal{A}$, we have*

$$\|\Gamma_{t+1}\|_1 - \|\Gamma_t^\star\|_1 \le \frac{1}{2\beta_t}\left(\|\Delta_t\|_F^2 - \|\Delta_{t+1}\|_F^2\right) + \frac{\beta_t}{2\tau_t}\left(\|A^\star - \hat{A}_t\|_F^2 - \|A^\star - \hat{A}_{t+1}\|_F^2\right)$$
$$- \frac{\beta_t}{2}\left(\frac{1}{\tau_t} - \Psi_{\max}(\hat{X}_t)\right)\|\hat{A}_{t+1} - \hat{A}_t\|_F^2 - \frac{\beta_t}{2}\|\Gamma_{t+1} - \widetilde{\Gamma}_t\|_F^2.$$

Proof. Let $\partial\|\Gamma_{t+1}\|_1$ denote the subgradient of $\|\Gamma_{t+1}\|_1$. Now Γ_{t+1} is a minimizer of (2). Therefore, $\mathbf{0}_{m\times n} \in \partial\|\Gamma_{t+1}\|_1 - \Delta_t - \beta_t(\widetilde{\Gamma}_t - \Gamma_{t+1})$. Rearranging the terms gives $\Delta_t + \beta_t(\widetilde{\Gamma}_t - \Gamma_{t+1}) \in \partial\|\Gamma_{t+1}\|_1$. Since $\|\Gamma_{t+1}\|_1$ is a convex function, we have

$$\begin{aligned}\|\Gamma_{t+1}\|_1 - \|\Gamma_t^\star\|_1 &\le \langle \Delta_t + \beta_t(\widetilde{\Gamma}_t - \Gamma_{t+1}), \Gamma_{t+1} - \Gamma_t^\star\rangle \\ &\le \langle \Delta_t, \Gamma_{t+1} - \widehat{\Gamma}_t\rangle + \langle \Delta_t, \widehat{\Gamma}_t - \Gamma_t^\star\rangle + \beta_t\langle\widetilde{\Gamma}_t - \Gamma_{t+1}, \Gamma_{t+1} - \Gamma_t^\star\rangle. \end{aligned} \tag{11}$$

Using Lemma B.1, the last term can be rewritten as

$$\beta_t\langle\widetilde{\Gamma}_t - \Gamma_{t+1}, \Gamma_{t+1} - \Gamma_t^\star\rangle = \frac{\beta_t}{2}\big(\|\Gamma_t^\star - \widetilde{\Gamma}_t\|_F^2 - \|\Gamma_t^\star - \Gamma_{t+1}\|_F^2 - \|\Gamma_{t+1} - \widetilde{\Gamma}_t\|_F^2\big) \tag{12}$$

Combining the inequality of Lemma B.2 with (12) gives

$$\langle \Delta_t, \widehat{\Gamma}_t - \Gamma_t^\star \rangle + \beta_t \langle \widetilde{\Gamma}_t - \Gamma_{t+1}, \Gamma_{t+1} - \Gamma_t^\star \rangle \leq \frac{\beta_t}{2\tau_t} \left(\|A^\star - \hat{A}_t\|_F^2 - \|A^\star - \hat{A}_{t+1}\|_F^2 \right) - \frac{\beta_t}{2} \left(\frac{1}{\tau_t} - \Psi_{\max}(\hat{X}_t) \right) \|\hat{A}_{t+1} - \hat{A}_t\|_F^2 - \frac{\beta_t}{2} (\|\Gamma_{t+1} - \widetilde{\Gamma}_t\|_F^2 - \|\Gamma_{t+1} - \widehat{\Gamma}_t\|_F^2). \quad (13)$$

Since $\Gamma_{t+1} - \widehat{\Gamma}_t = (\Delta_t - \Delta_{t+1})/\beta_t$, we have

$$\begin{aligned} \langle \Delta_t, \Gamma_{t+1} - \widehat{\Gamma}_t \rangle - \frac{\beta_t}{2} \|\Gamma_{t+1} - \widehat{\Gamma}_t\|_F^2 &= \frac{1}{2\beta_t} \left(2\langle \Delta_t, \Delta_t - \Delta_{t+1} \rangle - \|\Delta_t - \Delta_{t+1}\|_F^2 \right) \\ &= \frac{1}{2\beta_t} \left(\|\Delta_t\|_F^2 - \|\Delta_{t+1}\|_F^2 \right). \quad (14) \end{aligned}$$

Plugging (13) and (14) into (11) yields the result. □

Lemma B.4. *Let $\{\Gamma_t, \hat{A}_t, \Delta_t\}$ be the sequences generated by the OIADMM procedure. If τ_t satisfies $\frac{1}{\tau_t} \geq 2\Psi_{\max}(\hat{X}_t)$. Then*

$$\|\widetilde{\Gamma}_t\|_1 - \|\Gamma_t^\star\|_1 \leq \frac{1}{2\beta_t} \|\Lambda_t\|_F^2 + \frac{1}{2\beta_t} \left(\|\Delta_t\|_F^2 - \|\Delta_{t+1}\|_F^2 \right) + \frac{\beta_t}{2\tau_t} \left(\|A^\star - \hat{A}_t\|_F^2 - \|A^\star - \hat{A}_{t+1}\|_F^2 \right),$$

where $\Lambda_t \in \partial \|\widetilde{\Gamma}_t\|_1$.

Proof. Let $\Lambda_t \in \partial \|\widetilde{\Gamma}_t\|_1$. Therefore, $\|\widetilde{\Gamma}_t\|_1 - \|\Gamma_{t+1}\|_1 \leq \langle \Lambda_t, \widetilde{\Gamma}_t - \Gamma_{t+1} \rangle$. Now,

$$\langle \Lambda_t, \widetilde{\Gamma}_t - \Gamma_{t+1} \rangle = \langle \Lambda_t / \sqrt{\beta_t}, \sqrt{\beta_t} (\widetilde{\Gamma}_t - \Gamma_{t+1}) \rangle \leq \frac{1}{2\beta_t} \|\Lambda_t\|_F^2 + \frac{\beta_t}{2} \|\widetilde{\Gamma}_t - \Gamma_{t+1}\|_F^2$$

Therefore,

$$\|\widetilde{\Gamma}_t\|_1 - \|\Gamma_{t+1}\|_1 \leq \frac{1}{2\beta_t} \|\Lambda_t\|_F^2 + \frac{\beta_t}{2} \|\widetilde{\Gamma}_t - \Gamma_{t+1}\|_F^2. \quad (15)$$

Adding (15) and the inequality of Lemma B.3 together we get

$$\begin{aligned} \|\widetilde{\Gamma}_t\|_1 - \|\Gamma_t^\star\|_1 \leq \frac{1}{2\beta_t} \|\Lambda_t\|_F^2 + \frac{1}{2\beta_t} (\|\Delta_t\|_F^2 - \|\Delta_{t+1}\|_F^2) + \frac{\beta_t}{2\tau_t} \left(\|A^\star - \hat{A}_t\|_F^2 - \|A^\star - \hat{A}_{t+1}\|_F^2 \right) \\ - \frac{\beta_t}{2} \left(\frac{1}{\tau_t} - \Psi_{\max}(\hat{X}_t) \right) \|\hat{A}_{t+1} - \hat{A}_t\|_F^2. \end{aligned}$$

Setting $1/\tau_t \geq 2\Psi_{\max}(\hat{X}_t)$ means that $(-\beta_t/2)(\frac{1}{\tau_t} - \Psi_{\max}(\hat{X}_t)) \|\hat{A}_{t+1} - \hat{A}_t\|_F^2 \leq 0$, Therefore,

$$\|\widetilde{\Gamma}_t\|_1 - \|\Gamma_t^\star\|_1 \leq \frac{1}{2\beta_t} \|\Lambda_t\|_F^2 + \frac{1}{2\beta_t} \left(\|\Delta_t\|_F^2 - \|\Delta_{t+1}\|_F^2 \right) + \frac{\beta_t}{2\tau_t} \left(\|A^\star - \hat{A}_t\|_F^2 - \|A^\star - \hat{A}_{t+1}\|_F^2 \right),$$

□

Theorem B.5 (Theorem 4.2 Restated)**.** *Let $\{\Gamma_t, \hat{A}_t, \Delta_t\}$ be the sequences generated by the OIADMM procedure and $R(T)$ be defined as above. Assume the following conditions hold: (i) the Frobenius norm of $\partial \|\Gamma_t\|_1$ is upper bounded by Φ, (ii) $\hat{A}_0 = \mathbf{0}_{m \times k}$, $\|A^{\text{opt}}\|_F \leq D$, (iii) $\Delta_0 = \mathbf{0}_{m \times n}$, and (iv) $1/\tau_t \geq 2\Psi_{\max}(\hat{X}_t)$. Setting $\beta_t = (\Phi/D)\sqrt{\tau_t T}$, we have*

$$R(T) \leq \frac{\Phi D \sqrt{T}}{(2\sqrt{\tau_t})} + \sum_{t=1}^{T} \|A^{\text{opt}} E_t\|_1.$$

Proof. Substituting, $\Gamma_t^{\text{opt}} = P_t - A^{\text{opt}}\hat{X}_t$ for $\Gamma_t^\star$ and A^{opt} for $A^\star$ in Lemma B.4 and summing the inequality of over t from 1 to T, we get the following canceling telescoping sum

$$\begin{aligned}
&\sum_{t=1}^{T} \|\widetilde{\Gamma}_t\|_1 - \|\Gamma_t^{\text{opt}}\|_1 \\
&\leq \frac{1}{2\beta_t}\sum_{t=1}^{T} \|\Lambda_t\|_F^2 + \frac{1}{2\beta_t}(\|\Delta_0\|_F^2 - \|\Delta_{T+1}\|_F^2) + \frac{\beta_t}{2\tau_t}(\|A^{\text{opt}} - \hat{A}_0\|_F^2 - \|A^{\text{opt}} - \hat{A}_{T+1}\|_F^2) \\
&\leq \frac{1}{2\beta_t}\sum_{t=1}^{T} \|\Lambda_t\|_F^2 + \frac{\beta_t}{2\tau_t}\|A^{\text{opt}}\|_F^2 - \frac{1}{2\beta_t}\|\Delta_{T+1}\|_F^2 - \frac{\beta_t}{2\tau_t}\|A^{\text{opt}} - \hat{A}_{T+1}\|_F^2 \\
&\leq \frac{\Phi^2 T}{2\beta_t} + \frac{D^2\beta_t}{2\tau_t}.
\end{aligned}$$

Since

$$\overline{\Gamma}_t^{\text{opt}} = P_t - A^{\text{opt}}X_t = P_t - A^{\text{opt}}(\hat{X}_t + E_t) = \Gamma_t^{\text{opt}} - A^{\text{opt}}E_t,$$

we have then $\|\overline{\Gamma}_t^{\text{opt}}\|_1 \geq \|\Gamma_t^{\text{opt}}\|_1 - \|A^{\text{opt}}E_t\|_1$. The regret is bounded as follows:

$$R(T) = \sum_{t=1}^{T} \|\widetilde{\Gamma}_t\|_1 - \|\overline{\Gamma}_t^{\text{opt}}\|_1 \leq \frac{\Phi^2 T}{2\beta_t} + \frac{D^2\beta_t}{2\tau_t} + \sum_{t=1}^{T} \|A^{\text{opt}}E_t\|_1 .$$

Setting $\beta_t = \frac{\Phi}{D}\sqrt{\tau_t T}$ yields desired bound. □

As mentioned in Section 4, OIADMM can violate the equality constraint at each t (i.e., $P_t - \hat{A}_{t+1}\hat{X}_t \neq \Gamma_{t+1}$). However, we show in Theorem B.6 that the accumulated loss caused by the violation of equality constraint is sublinear in T, i.e., the equality constraint is satisfied on average in the long run.

Theorem B.6. *Let $\{\Gamma_t, \hat{A}_t, \Delta_t\}$ be the sequences generated by the OIADMM procedure and $R(T)$ be defined as above. Assume the following conditions hold: (i) the Frobenius norm of $\partial\|\Gamma_t\|_1$ is upper bounded by Φ, (ii) $\hat{A}_0 = \mathbf{0}_{m\times k}$, $\|A^{\text{opt}}\|_F \leq D$, (iii) $\Delta_0 = \mathbf{0}_{m\times n}$, (iv) $1/\tau_t \geq 2\Psi_{\max}(\hat{X}_t)$, and (v) $\|\Gamma_t^{\text{opt}}\|_1 \leq \Upsilon$. Setting $\beta_t = (\Phi/D)\sqrt{\tau_t T}$, we have*

$$\sum_{t=1}^{T} \|\Gamma_{t+1} - \widehat{\Gamma}_t\|_2^2 \leq \frac{2D^2}{\tau_t} + \frac{4\Upsilon D}{\Phi\sqrt{\tau_t}}\sqrt{T}.$$

Proof. Lets look at $\|\Gamma_{t+1} - \widehat{\Gamma}_t\|_F^2$.

$$\begin{aligned}
\|\Gamma_{t+1} - \widehat{\Gamma}_t\|_F^2 = \|\Gamma_{t+1} - \widetilde{\Gamma}_t + \widetilde{\Gamma}_t - \widehat{\Gamma}_t\|_F^2 &\leq 2\left(\|\Gamma_{t+1} - \widetilde{\Gamma}_t\|_F^2 + \|\widetilde{\Gamma}_t - \widehat{\Gamma}_t\|_F^2\right) \\
&\leq 2\left(\|\Gamma_{t+1} - \widetilde{\Gamma}_t\|_F^2 + \Psi_{\max}(\hat{X}_t)\|\hat{A}_{t+1} - \hat{A}_t\|_F^2\right). \qquad (16)
\end{aligned}$$

For the first inequality, we used the simple fact that for any two matrices M_1 and M_2 $\|M_1 - M_2\|_F^2 \leq 2(\|M_1\|_F^2 + \|M_2\|_F^2)$. The second inequality is because of (9). Firstly, since $\|\Gamma_{t+1}\|_1 \geq 0$

$$\|\Gamma_{t+1}\|_1 - \|\Gamma_t^{\text{opt}}\|_1 \geq -\|\Gamma_t^{\text{opt}}\|_1 \geq -\Upsilon.$$

Using this and rearranging terms in the inequality of Lemma B.3 (with A^{opt} instead of $A^\star$) gives

$$\begin{aligned}
\|\Gamma_{t+1} - \widetilde{\Gamma}_t\|_F^2 \leq \frac{1}{\beta_t^2}\left(\|\Delta_t\|_F^2 - \|\Delta_{t+1}\|_F^2\right) + \frac{1}{\tau_t}\left(\|A^{\text{opt}} - \hat{A}_t\|_F^2 - \|A^{\text{opt}} - \hat{A}_{t+1}\|_F^2\right) \\
- \left(\frac{1}{\tau_t} - \Psi_{\max}(\hat{X}_t)\right)\|\hat{A}_{t+1} - \hat{A}_t\|_F^2 + \frac{2\Upsilon}{\beta_t},
\end{aligned}$$

Plugging this into (16) yields

$$\begin{aligned}
\|\Gamma_{t+1} - \widehat{\Gamma}_t\|_F^2 \leq \frac{2}{\beta_t^2}\left(\|\Delta_t\|_F^2 - \|\Delta_{t+1}\|_F^2\right) + \frac{2}{\tau_t}\left(\|A^{\text{opt}} - \hat{A}_t\|_F^2 - \|A^{\text{opt}} - \hat{A}_{t+1}\|_F^2\right) \\
- 2\left(\frac{1}{\tau_t} - 2\Psi_{\max}(\hat{X}_t)\right)\|\hat{A}_{t+1} - \hat{A}_t\|_F^2 + \frac{4\Upsilon}{\beta_t}.
\end{aligned}$$

Letting $1/\tau_t \geq 2\Psi_{\max}(\hat{X}_t)$ and summing over t from 1 to T, we have

$$\sum_{t=1}^{T} \|\Gamma_{t+1} - \widehat{\Gamma}_t\|_F^2 \leq \frac{2}{\beta_t^2}\left(\|\Delta_0\|_F^2 - \|\Delta_{T+1}\|_F^2\right) + \frac{2}{\tau_t}\left(\|A^{\text{opt}} - \hat{A}_0\|_F^2 - \|A^{\text{opt}} - \hat{A}_{T+1}\|_F^2\right) + \frac{4\Upsilon T}{\beta_t}$$

$$\leq \frac{2D^2}{\tau_t} + \frac{4\Upsilon T}{\beta_t}.$$

Setting $\beta_t = \frac{\Phi}{D}\sqrt{\tau_t T}$ yields the desired bound. □

C Pseudo-Codes from Section 5

Let us start by extending the definition of $\mathcal{A}$, define
$\mathcal{A}_{k_t} = \{A \in \mathbb{R}^{m \times k_t} \ : \ A \geq \mathbf{0}_{m \times k_t} \ \forall j = 1, \ldots, k_t \ , \|A_j\|_1 \leq 1\}$, where A_j is the jth column in A.
We use $\Pi_{\mathcal{A}_{k_t}}$ to denote the projection onto the nearest point in the convex set $\mathcal{A}_{k_t}$.

Algorithm 4 : BATCH-IMPL

Input: $P_{[t-1]} \in \mathbb{R}^{m \times N_{t-1}}$, $X_{[t-1]} \in \mathbb{R}^{k_t \times N_{t-1}}$, $P_t = [\mathbf{p}_1, \ldots, \mathbf{p}_{n_t}] \in \mathbb{R}^{m \times n_t}$, $A_t \in \mathbb{R}^{m \times k_t}$, $\lambda, \zeta, \eta \geq 0$
Novel Document Detection Step:
for $j = 1$ **to** n_t **do**
  Solve: $\mathbf{x}_j = \text{argmin}_{\mathbf{x} \geq 0} \ \|\mathbf{p}_j - A_t\mathbf{x}\|_1 + \lambda\|\mathbf{x}\|_1$ (solved using Algorithm 2)
  if $\|\mathbf{p}_j - A_t\mathbf{x}_j\|_1 + \lambda\|\mathbf{x}_j\|_1 > \zeta$
    Mark $\mathbf{p}_j$ as novel
Batch Dictionary Learning Step:
Set $k_{t+1} \leftarrow k_t + \eta$
Set $Z_{[t]} \leftarrow [X_{[t-1]} \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_{n_t}]$
Set $X_{[t]} \leftarrow \begin{bmatrix} Z_{[t]} \\ \hline \mathbf{0}_{\eta \times N_t} \end{bmatrix}$
Set $P_{[t]} \leftarrow [P_{[t-1]} \,|\, \mathbf{p}_1, \ldots, \mathbf{p}_{n_t}]$
for $i = 1$ **to** *convergence* **do**
  Solve: $A_{t+1} = \text{argmin}_{A \in \mathcal{A}_{k_{t+1}}} \ \|P_{[t]} - AX_{[t]}\|_1$ (solved using Algorithm 3 with warm starts)
  Solve: $X_{[t]} = \text{argmin}_{X \geq \mathbf{0}} \ \|P_{[t]} - A_{t+1}X\|_1 + \lambda\|X\|_1$ (solved using Algorithm 2)

Define $\mathbb{A}_{k_t}$ as
$\mathbb{A}_{k_t} = \{A \in \mathbb{R}^{m \times k_t} \ : \ A \geq \mathbf{0}_{m \times k_t} \ \forall j = 1, \ldots, k_t \ , \|A_j\|_2 \leq 1\}$, where A_j is the jth column in A.
We use $\Pi_{\mathbb{A}_{k_t}}$ to denote the projection onto the nearest point in the convex set $\mathbb{A}_{k_t}$.

Algorithm 5 : L2-BATCH

Input: $P_{[t-1]} \in \mathbb{R}^{m \times N_{t-1}}$, $P_t = [\mathbf{p}_1, \ldots, \mathbf{p}_{n_t}] \in \mathbb{R}^{m \times n_t}$, $A_t \in \mathbb{R}^{m \times k_t}$, $\lambda \geq 0, \zeta \geq 0, \eta \geq 0$
Novel Document Detection Step:
for $j = 1$ **to** n_t **do**
  Solve: $\mathbf{x}_j = \text{argmin}_{\mathbf{x} \geq 0} \ \|\mathbf{p}_j - A_t\mathbf{x}\|_2 + \lambda\|\mathbf{x}\|_1$ (solved using the LARS method [3])
  if $\|\mathbf{p}_j - A_t\mathbf{x}_j\|_2 + \lambda\|\mathbf{x}_j\|_1 > \zeta$
    Mark $\mathbf{p}_j$ as novel
ℓ_2-batch Dictionary Learning Step:
Set $k_{t+1} \leftarrow k_t + \eta$
Set $P_{[t]} \leftarrow [P_{[t-1]} \,|\, \mathbf{p}_1, \ldots, \mathbf{p}_{n_t}]$
$[A_{t+1}, X_{[t]}] = \text{argmin}_{A \in \mathbb{A}_{k_{t+1}}, X \geq \mathbf{0}} \ \|P_{[t]} - AX\|_2 + \lambda\|X\|_1$ (non-negative sparse coding problem)

D Additional Experimental Evaluation

In Figure 1, we show the effect of the size of the dictionary on the performance of Algorithm ONLINE. The average AUC is computed as in Table 1. Not surprisingly, as the size of dictionary (k) increases the average AUC also increases, but correspondingly the running time of the algorithm also increases. The plot suggests that there is a diminishing return on AUC with increase in the size of the dictionary, and this increase in AUC comes at the cost of higher running times.

Post-processing done to Generate Table 2. In each timestep, instead of thresholding by ζ, we take the top 10% of tweets measured in terms of the sparse coding objective value and run a dictionary-based clustering, described in [4], on it. Further post-processing is done to discard clusters without much support and to pick a representative tweet for each cluster.

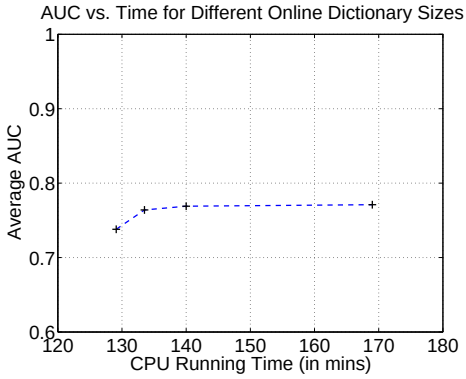

Figure 1: TDT2 dataset: Average AUC vs. running time for different values of dictionary sizes (k) in Algorithm ONLINE. The points plotted (from left to right) are for $k = 50, 100, 150$, and 200.

References

[1] S. 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*, 2011.

[2] P. Combettes and J. Pesquet. Proximal Splitting Methods in Signal Processing. arXiv:0912.3522, 2009.

[3] J. Friedman, T. Hastie, H. Hfling, and R. Tibshirani. Pathwise Coordinate Optimization. *The Annals of Applied Statistics*, 1(2), 2007.

[4] S. P. Kasiviswanathan, P. Melville, A. Banerjee, and V. Sindhwani. Emerging Topic Detection using Dictionary Learning. In *CIKM*, pages 745–754, 2011.

[5] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer-Verlag, 2004.

[6] J. Yang and Y. Zhang. Alternating Direction Algorithms for L1-Problems in Compressive Sensing. *SIAM Journal of Scientific Computing*, 33(1):250–278, 2011.