Supplementary Material for An Efficient Algorithm for Large Scale Compressive Feature Learning

Hristo Paskov

1 Minimizations for Linearized ADMM

We show how to minimize the augmented Lagrangian

$$\mathcal{L}_\rho(w,\gamma,y) = h(w) + g(\gamma) + y^T(Aw-\gamma) + \frac{\rho}{2}\|Aw-\gamma\|_2^2 \tag{1}$$

Note that $\gamma = \begin{bmatrix} z \\ \theta \end{bmatrix}$ and hence $y = \begin{bmatrix} y^{(z)} \\ y^{(\theta)} \end{bmatrix}$ are partitioned variables. We use the following definitions for h, g, and A:

$$\begin{aligned} h(w) &= w^T d + \sum_{s\in\mathcal{S}} c(s)\|w_{J(s)}\|_\infty + I\{w \geq 0\} \\ g(z,\theta) &= I\{z \geq 1\} + I\left\{\left\|\theta_{T(i)}\right\|_\infty \geq \zeta \ \forall i = 1,\dots,n\right\} \\ A &= \begin{bmatrix} X \\ I_m \end{bmatrix} \end{aligned} \tag{2}$$

We use the notation $I\{\bullet\}$ as an indicator function that is 0 if the condition insides the braces is met and is ∞ otherwise. Starting with w, recall that we use the linearized form of the Lagrangian. The relevant parts of the optimization problem are

$$w^T d + \sum_{s\in\mathcal{S}} c(s)\left\|w_{J(s)}\right\|_\infty + y^{(z)T}Xw + y^{(\theta)T}w + \rho(X\bar{w}-z)^T Xw + \frac{\rho}{2}\|w-\theta\|_2^2 + \frac{\mu}{2}\|w-\bar{w}\|_2^2 \tag{3}$$

$$\text{subject to } w \geq 0$$

We collect all linear terms and complete the square to obtain the equivalent formulation

$$\sum_{s\in\mathcal{S}} c(s)\left\|w_{J(s)}\right\|_\infty + \frac{\mu+\rho}{2}\|w-q\|_2^2 \tag{4}$$

$$q = -\frac{1}{\mu+\rho}\left(d + y^{(z)T}X + y^{(\theta)} + \rho X^T(X\bar{w}-z) - \mu\bar{w} - \rho\theta\right) \tag{5}$$

Next, the relevant terms for γ are

$$-\begin{bmatrix} z^T & \theta^T \end{bmatrix}\begin{bmatrix} y^{(z)} \\ y^{(\theta)} \end{bmatrix} + \frac{\rho}{2}\left\|\begin{bmatrix} Xw \\ w \end{bmatrix} - \begin{bmatrix} z \\ \theta \end{bmatrix}\right\|_2^2 \tag{6}$$

$$\text{subject to } z \geq 1$$

$$\left\|\theta_{T(i)}\right\|_\infty \geq \zeta \;\forall i = 1,\ldots,n \tag{7}$$

This function is clearly separable in z and θ. It simply projects $Xw + \rho^{-1}y^{(z)}$ to have all entries ≥ 1 and hence $z = \max(Xw + \rho^{-1}y^{(z)}, 1)$. The solution for θ is discussed in the paper.

2 Structure of X

We assume that a collection of N documents of sizes $n_1,\ldots,n_N$ is being compressed and define $n = \sum_{i=1}^N n_i$. Assuming that we allow all K-grams that respect document boundaries as potential pointers, X has special structure. Note that there are $m_j = \sum_{i=1}^K (n_j - i + 1)$ potential pointers for document j and that $X \in \{0,1\}^{n\times m}$ where $m = \sum_{i=1}^N m_i$. Each column in this matrix corresponds to a particular potential pointer in $\mathcal{P}$ and we are free to select how to order the pointers and hence columns of X. An efficient way to do this is to let the first m_1 columns correspond to the pointers for document 1, the next m_2 columns to the pointers for document 2, and so on. X then becomes a block diagonal matrix

$$X = \begin{bmatrix} X^{(1)} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & X^{(N)} \end{bmatrix} \tag{8}$$

(where $X^{(i)} \in \{0,1\}^{n_i \times m_i}$ corresponds to D_i) because the pointers for document i cannot be used to reconstruct any other documents (because they respect document boundaries). It is easy to see that with this order,

$$XX^T = \begin{bmatrix} X^{(1)}X^{(1)T} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & X^{(N)}X^{(N)T} \end{bmatrix} \tag{9}$$

is a block diagonal positive semidefinite matrix.

We further order the columns within each $X^{(i)}$ by ordering the pointers according to size first and then starting location. Thus, column j for $1 \leq j \leq n_i$ corresponds to the pointer for the unigram at position j. Column j for $n_i + 1 \leq j \leq 2n_i - 1$ corresponds to the pointer for the bi-gram that starts at position $j - n_i$, and so on. When $K = 2$, for example, $X^{(i)}$ looks like

$$X^{(i)} = \begin{bmatrix} I_{n_i} & A^{(i2)} \end{bmatrix} \tag{10}$$

where I_{n_i} is the $n_i \times n_i$ identity and $A^{(i2)} \in \{0,1\}^{n_i \times (n_i - 1)}$ with column k a vector of all 0's except for two 1's in positions k and $k+1$. If we continue this line of reasoning, we see that $X^{(i)}$ can be expressed as K block matrices

$$X^{(i)} = \begin{bmatrix} A^{(i1)} & \dots & A^{(iK)} \end{bmatrix} \tag{11}$$

where $A^{(ik)} \in \{0,1\}^{n_i \times (n_i - k + 1)}$. Column j of $A^{(ik)}$ is all 0's except for a contiguous string of k 1's starting at position j. We will use the notation $e^{(jk)}$ to denote column j in $A^{(ik)}$ (we have dropped reference to the document number for brevity). An immediate consequence of this representation is that vector multiplication by $X^{(i)}$ and $X^{(i)T}$ is very fast. For instance, the product

$$X^{(i)T} w = \sum_{k=1}^{K} A^{(ik)T} w \tag{12}$$

and multiplying by $A^{(ik)T}$ amounts to convolving w with a signal of k 1's and is easily performed in $\Theta(n_i)$ operations. The overall running time is therefore $O(Kn_i)$ for matrix-vector multiplication by $X^{(i)T}$ and a similar $O(Kn_i)$ algorithm can be obtained for multiplication by $X^{(i)}$.

We can now express the product $B^{(i)} = X^{(i)} X^{(i)T}$ as

$$B^{(i)} = \sum_{k=1}^{K} \sum_{i=1}^{n_i - k + 1} e^{(jk)} e^{(jk)T} \tag{13}$$

or, equivalently, as sum of squares of 1's of side lengths $1, \dots, K$ whose upper left corner is positioned along the main diagonal. We will call these squares k-squares.

To start, $B^{(i)}$ is symmetric because it is a covariance matrix so we only consider its upper triangle. We start with the "middle" entries, assuming that $n_i > 2k - 2$. Then $B^{(i)}_{st}$ for $t \geq s$ and $s \geq k$ can be expressed as an appropriate sum of 1's. Note that if $t = s + 1$, a 1-square cannot contribute to the entry. Extending this reasoning to the general case, we see that if $z = t - s$, then only $z + 1, \dots, K$ squares can contribute to $B^{(i)}_{st}$. This implies that $B^{(i)}_{st} = 0$ if $t \geq s + K$, i.e. each $B^{(i)}$ and hence XX^T is $K - 1$ banded and symmetric.

Next, assuming $k \in z + 1, \dots, K$, a k-square whose upper left corner is in row j can only contribute if it is non-zero at position (s, t). This happens when $s - j + 1 \leq k$ and $t - j + 1 \leq k$, i.e. $j \geq s - k + 1$ and $j \geq t - k + 1$. Since $t \geq s$, we only need to check the second inequality. Finally, we also know that $j \leq s$, and so our entry can be expressed as

$$B_{st}^{(i)} = \sum_{k=z+1}^{K} \sum_{i=t-k+1}^{s} 1 = \sum_{k=z+1}^{K} (k-z) = \sum_{k=1}^{K-z} (k+z-z) = \frac{(K-z)(K-z+1)}{2} \tag{14}$$

Next, suppose that $t \geq s$ and $1 \leq s < k$. The outer summation stays the same, but the inner one must account for when $t-k+1 < 1$. In those cases, the inner summation contributes only s instead of $k-z$. This situation happens when $k > t$, so we divide the summation into

$$B_{st}^{(i)} = \sum_{k=z+1}^{t'} (k-z) + \sum_{k=t'+1}^{K} s = \sum_{k=1}^{t'-z} k + s(K-t') = s(K-t') = \frac{(t'-z)(t'-z+1)}{2} \tag{15}$$

where $t' = \min(K, t)$. Finally, our matrix is not only symmetric but also symmetric with respect to its minor diagonal. This can be seen from redoing all of our formulas using the bottom right of each square rather than the top left.

It is easy to see now that $B^{(i)}$ is nearly Toeplitz. Indeed, if we chop off the top and bottom $K-1$ rows, this is the case. The sum of each row of this Toeplitz matrix can be expressed as

$$\frac{K(K+1)}{2} + 2\sum_{z=1}^{K-1} \frac{(K-z)(K-z+1)}{2} = \frac{K^3}{3} + \frac{K^2}{2} + \frac{K}{6} \tag{16}$$

In addition, it is easy to see that each of the top and bottom rows we removed must sum to an integer less than $\frac{K^3}{3} + \frac{K^2}{2} + \frac{K}{6}$ since each entry in these rows has fewer k-squares added to it than the rows in the middle.

A sample 12×12 matrix $B^{(i)}$ with $K = 5$ is shown below:

$$\begin{matrix}
5 & 4 & 3 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 9 & 7 & 5 & 3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 7 & 12 & 9 & 6 & 3 & 1 & 0 & 0 & 0 & 0 & 0 \\
2 & 5 & 9 & 14 & 10 & 6 & 3 & 1 & 0 & 0 & 0 & 0 \\
1 & 3 & 6 & 10 & 15 & 10 & 6 & 3 & 1 & 0 & 0 & 0 \\
0 & 1 & 3 & 6 & 10 & 15 & 10 & 6 & 3 & 1 & 0 & 0 \\
0 & 0 & 1 & 3 & 6 & 10 & 15 & 10 & 6 & 3 & 1 & 0 \\
0 & 0 & 0 & 1 & 3 & 6 & 10 & 15 & 10 & 6 & 3 & 1 \\
0 & 0 & 0 & 0 & 1 & 3 & 6 & 10 & 14 & 9 & 5 & 2 \\
0 & 0 & 0 & 0 & 0 & 1 & 3 & 6 & 9 & 12 & 7 & 3 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 3 & 5 & 7 & 9 & 4 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 4 & 5
\end{matrix} \tag{17}$$