1 Intro

In this document, we will be discussing a stateful version of mirror descent used to encourage periodicity in a sequence of parameter estimates and/or data transformations. This method is intended for scenarios in which a signal is known to be approximately periodic with some drift across periods. Currently, I am applying this method to a model that is everywhere differentiable, so we will assume the existence of a gradient at all points.

2 Notation

The parameter estimate for time step t within period n will be denoted $\mathbf{\Phi}_t^{(n)} \in \mathbb{R}^{p\times k}$. The minibatch of data received at time step t of period n will be denoted $\mathbf{X}_t^{(n)} \in \mathbb{R}^{m\times p}$ where m is the batch size and p is the ambient dimension of the data.

3 Stateful Link/Bregman/Proximal Functions for Periodicity

Our proximal function at time step $t+1$ during the n-th period will be the sum of a strongly convex penalty for stability and an additional convex penalty to encourage periodicity. The obvious choice for the strongly convex penalty is the typical squared Frobenius proximal function $\left\|\mathbf{\Phi} - \mathbf{\Phi}_t^{(n)}\right\|_F^2$.

There are two immediately clear options for the second term depending on whether we want filtered data values or the parameters themselves to be approximately periodic. To encourage periodicity of the filtered data, we will add a term $\left\|\mathbf{X}_{t+1}^{(n)}\mathbf{\Phi} - \mathbf{X}_{t+1}^{(n-1)}\mathbf{\Phi}_{t+1}^{(n-1)}\right\|_F^2$. To encourage periodicity in the parameter estimates themselves, we replace that term with a similar term sans the data, $\left\|\mathbf{\Phi} - \mathbf{\Phi}_{t+1}^{(n-1)}\right\|_F^2$. Our two potential proximal functions are then

$$\Psi_{t+1}^{(n)}(\mathbf{\Phi}) = \left\|\mathbf{\Phi} - \mathbf{\Phi}_t^{(n)}\right\|_F^2 + c_1\left\|\mathbf{X}_{t+1}^{(n)}\mathbf{\Phi} - \mathbf{X}_{t+1}^{(n-1)}\mathbf{\Phi}_{t+1}^{(n-1)}\right\|_F^2 \quad (1)$$

$$\Psi_{t+1}^{(n)}(\mathbf{\Phi}) = \left\|\mathbf{\Phi} - \mathbf{\Phi}_t^{(n)}\right\|_F^2 + c_1\left\|\mathbf{\Phi} - \mathbf{\Phi}_{t+1}^{(n-1)}\right\|_F^2, \quad (2)$$

where $c_1 > 0$ is a constant chosen to weight the importance of one term against the other. It is indexed with 1 because we will later introduce other constants in the full proximal operator expression.

Our unregularized proximal operators with no dual averaging then take the form

$$\mathbf{\Phi}_{t+1}^{(n)} = \arg\min_{\mathbf{\Phi}} \left\langle \mathbf{\Phi}, \nabla f_t^{(n)}\left(\mathbf{\Phi}_t^{(n)}\right)\right\rangle + \frac{c_2}{2}\left(\left\|\mathbf{\Phi} - \mathbf{\Phi}_t^{(n)}\right\|_F^2 + c_1\left\|\mathbf{X}_{t+1}^{(n)}\mathbf{\Phi} - \mathbf{X}_{t+1}^{(n-1)}\mathbf{\Phi}_{t+1}^{(n-1)}\right\|_F^2\right) \quad (3)$$

$$\mathbf{\Phi}_{t+1}^{(n)} = \arg\min_{\mathbf{\Phi}} \left\langle \mathbf{\Phi}, \nabla f_t^{(n)}\left(\mathbf{\Phi}_t^{(n)}\right)\right\rangle + \frac{c_2}{2}\left(\left\|\mathbf{\Phi} - \mathbf{\Phi}_t^{(n)}\right\|_F^2 + c_1\left\|\mathbf{\Phi} - \mathbf{\Phi}_{t+1}^{(n-1)}\right\|_F^2\right), \quad (4)$$

where $c_2 > 0$ is a constant chosen to weigh the importance of the proximal function, and $f_t^{(n)}$ is the loss function at time step t in the n-th period. To find a minimizer, we take the gradient of the expressions inside each $\arg\min_{\mathbf{\Phi}}$ and set to 0, then solve for $\mathbf{\Phi}$. For the first proximal operator, the gradient of hte proximal expression is equal to

$$0 = \nabla f_t^{(n)}\left(\mathbf{\Phi}_t^{(n)}\right) + c_2\left(\mathbf{\Phi} - \mathbf{\Phi}_t^{(n)} + c_1\left(\mathbf{X}_{t+1}^{(n)}\right)^\top\left(\mathbf{X}_{t+1}^{(n)}\mathbf{\Phi} - \mathbf{X}_{t+1}^{(n-1)}\mathbf{\Phi}_{t+1}^{(n-1)}\right)\right). \quad (5)$$

Sending all of the terms that include $\mathbf{\Phi}$ to one side, we have

$$\left(c_2\mathbf{I}_p + c_2\cdot c_1\left(\mathbf{X}_{t+1}^{(n)}\right)^\top\left(\mathbf{X}_{t+1}^{(n)}\right)\right)\mathbf{\Phi} = c_2\mathbf{\Phi}_t^{(n)} + c_2\cdot c_1\left(\mathbf{X}_{t+1}^{(n)}\right)^\top\mathbf{X}_{t+1}^{(n-1)}\mathbf{\Phi}_{t+1}^{(n-1)} - \nabla f_t^{(n)}\left(\mathbf{\Phi}_t^{(n)}\right). \quad (6)$$

Finally, we get the following expression for the optimal $\mathbf{\Phi}_{t+1}^{(n)}$.

$$\mathbf{\Phi}_{t+1}^{(n)} \;=\; \left(c_2\mathbf{I}_p + c_2 \cdot c_1 \left(\mathbf{X}_{t+1}^{(n)}\right)^\top \left(\mathbf{X}_{t+1}^{(n)}\right)\right)^{-1} \left(c_2\mathbf{\Phi}_t^{(n)} + c_2 \cdot c_1 \left(\mathbf{X}_{t+1}^{(n)}\right)^\top \mathbf{X}_{t+1}^{(n-1)}\mathbf{\Phi}_{t+1}^{(n-1)} - \nabla f_t^{(n)} \left(\mathbf{\Phi}_t^{(n)}\right)\right). \tag{7}$$

For the second proximal expression, we follow similar steps to get, first, the gradient set to zero.

$$0 \;=\; \nabla f_t^{(n)} \left(\mathbf{\Phi}_t^{(n)}\right) + c_2 \left(\mathbf{\Phi} - \mathbf{\Phi}_t^{(n)} + c_1 \left(\mathbf{\Phi} - \mathbf{\Phi}_{t+1}^{(n-1)}\right)\right). \tag{8}$$

Moving every term with $\mathbf{\Phi}$ to one side, we get

$$\left(\left(c_2 + c_2 \cdot c_1\right)\mathbf{I}_p\right)\mathbf{\Phi} \;=\; c_2\mathbf{\Phi}_t^{(n)} + c_2 \cdot c_1\mathbf{\Phi}_{t+1}^{(n-1)} - \nabla f_t^{(n)} \left(\mathbf{\Phi}_t^{(n)}\right). \tag{9}$$

Finally, we have the following expression for the optimal $\mathbf{\Phi}_{t+1}^{(n)}$.

$$\mathbf{\Phi}_{t+1}^{(n)} \;=\; \left(\left(c_2 + c_2 \cdot c_1\right)\mathbf{I}_p\right)^{-1} \left(c_2\mathbf{\Phi}_t^{(n)} + c_2 \cdot c_1\mathbf{\Phi}_{t+1}^{(n-1)} - \nabla f_t^{(n)} \left(\mathbf{\Phi}_t^{(n)}\right)\right). \tag{10}$$