Appendix

1 Proof of Proposition 2.2

Here we show how efficient computation can be done with the proposed models and prove Proposition 2.2.

PROPOSITION 2.2. *Time complexity of making a prediction with SH$\mathcal{A}^2$ is $\mathcal{O}(kd)$.*

Proof. Following (3.5) and dropping the bias term, we derive efficient computation for SHFMs and SH$\mathcal{A}^2$ with:

(1.1)

$$\sum_{f=1}^{k} \beta_f \mathcal{A}^2(\boldsymbol{V}'_{:,f}, \boldsymbol{x}') = \sum_{f=1}^{k} \beta_f \sum_{i=0}^{d} \sum_{j=i+1}^{d} V_{i,f} x_i V_{j,f} x_j$$

$$= \frac{1}{2} \sum_{f=1}^{k} \beta_f \Big(\sum_{i=0}^{d} \sum_{j=0}^{d} V_{i,f} V_{j,f} x_i x_j - \sum_{i=0}^{d} V_{i,f} V_{i,f} x_i x_i\Big)$$

$$= \frac{1}{2} \sum_{f=1}^{k} \beta_f \Big[\Big(\sum_{i=0}^{d} V_{i,f} x_i\Big)^2 - \sum_{i=0}^{d} (V_{i,f} x_i)^2\Big]$$

Therefore, the computation is linear in terms of both k and $d+1$, so the time complexity is $\mathcal{O}(k(d+1)) = \mathcal{O}(kd)$.

2 FTRL-Proximal for SH$\mathcal{A}^2$

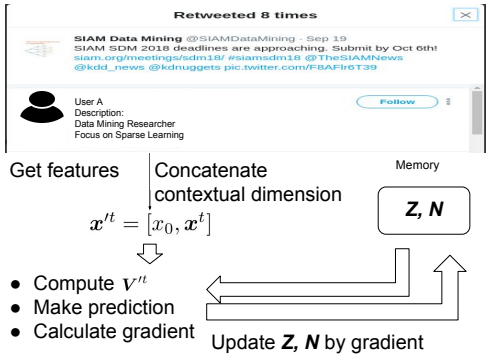

Figure 1: Visualization of Algorithm 1 with our running example.

In Figure 1, we show how Algorithm 1 updates parameters when a retweet is observed with interaction of the two features: *data mining researcher* in the user profile and *SDM* in the tweet text.

3 Experimental Setup Details

3.1 When Do You Retweet Dataset In table 1, we briefly describe the WDYR dataset. Then we describe the details for one-hot encoding as below:

- Text: we concatenate user description and tweet text together and use Bag of Words (BoW) to model them with vocabulary of 10,000 plus 1 dimension for the other infrequent words.
- Numerical: one-hot encoding is applied to bins of numerical attributes which are grouped into bins by their deciles.
- Time stamp (hours): we divide 24 hours of a day into 4 groups: 12am - 5:59am, 6am - 11:59am, 12pm - 5:59pm, 6pm - 11:59pm.

Table 1: One-hot encoding of features for WDYR

Type	Field	Dimension
user id	user id	10,157
tweet id	original tweet id	3,771
text	user description	10,001
	tweet text	
numerical	tweet count etc.	10
time stamp: year	create time, t_0	10
time stamp: month		12
time stamp: weekday		7
time stamp: hour		4

3.2 Grid Search Table 2 shows the domain of grid search for hyper-parameters (λ_1,λ_2,α and k).

Table 2: Grid search for hyper-parameters.

	WDYR	E2006
λ_1	$\{10^{-2}, 10^{-3}, ..., 10^{-5}\}$	$\{100.0, 10.0, ..., 10^{-3}\}$
λ_2	$\{0.1, 10^{-2}, 10^{-3}10^{-4}\}$	
α	$\{1.0, 0.1, ..., 10^{-3}\}$	$\{1, 0.5, 0.2, ..., 0.02, 10^{-2}\}$
k	$\{5, 10, 20, 50\}$	