Appendix: Future Gradient Descent for Adapting the Temporal Shifting Data Distribution in Online Recommendation Systems

Mao Ye, Ruichen Jiang, Haoxiang Wang, Dhruv Choudhary, Xiaocong Du, Bhargav Bhushanam, Aryan Mokhtari, Arun Kejariwal, Qiang Liu

1The University of Texas at Austin.
2The University of Illinois at Urbana-Champaign.
3Meta.

Extra Notation We introduce several new notations for the appendix. We use $\langle \cdot, \cdot \rangle$ to denote the inner product between two vectors and use $\circ$ to denote the entrywise product.

1 PROOF OF THEOREM 1

Proof. We start with a simple decomposition using the triangle inequality:

$$\|u_{w,t}(\theta_t)\| \leq \|u_{w,t}(\theta_t) - \bar{m}(\theta_t; t)\| + \|\bar{m}(\theta_t; t)\|.$$  

By the termination condition of Algorithm 2, we have $\|\bar{m}(\theta_t; t)\| \leq \delta$. Furthermore, it follows from (5) that

$$\|u_{w,t}(\theta_t) - \bar{m}(\theta_t; t)\| = \frac{1}{w} \|\nabla r_t(\theta_t) - m(\theta_t; t)\|.$$  

Hence, we obtain

$$\|u_{w,t}(\theta_t)\|^2 \leq \left( \delta + \frac{1}{w} \|\nabla r_t(\theta_t) - m(\theta_t; t)\| \right)^2 \leq 2\delta^2 + \frac{2}{w^2} \|\nabla r_t(\theta_t) - m(\theta_t; t)\|^2.$$  

This further implies that

$$R_w(T) = \frac{1}{T} \sum_{t=1}^{T} \|u_{w,t}(\theta_t)\|^2 \leq \frac{2}{w^2 T} \sum_{t=1}^{T} \|\nabla r_t(\theta_t) - m(\theta_t; t)\|^2 + 2\delta^2,$$  

and the main result follows from the fact that $\|\nabla r_t(\theta_t) - m(\theta_t; t)\|^2 \leq \sup_{\theta} \|\nabla r_t(\theta) - m(\theta; t)\|^2$ for all $t \in [T]$.

Furthermore, under the boundedness assumption, we have for all $t \in [T]$

$$\|\nabla r_t(\theta_t) - m(\theta_t; t)\|^2 \leq (\|\nabla r_t(\theta_t)\| + \|m(\theta_t; t)\|)^2 \leq 4M^2.$$  

Hence, (3) also implies $R_w(T) \leq 8M^2/w^2 + 2\delta^2$, which leads to $R_w(T) = O(1/w^2)$ when $\delta = 1/w$.  

2 DETAILS OF THE RESULT IN SECTION 4.4

Algorithm. Given $\theta_t$, define $h_t(\phi) = \|\nabla r_t(\theta_t) - m(\theta_t; \phi, t)\|^2$ as a function of $\phi$, where we view $\theta_t$ as a constant. Thus, if follows from that (4) that

$$R_w(T) \leq \frac{2}{w^2 T} \sum_{t=1}^{T} h_t(\phi_t) + 2\delta^2.$$  

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).
where we used Cauchy-Schwarz inequality in (6), the triangle inequality in (7) and the boundedness of the gradients in (8).

### Algorithm 1

**Input:** The learning rate $\eta, \eta_\phi$ for updating the model parameter $\theta$ and $\phi$.

Initialize $\phi_1 = [1/b, \ldots, 1/b]$.

for $t \in [T]$ do

- Deploy the prediction model $f_{\theta_t}$ with the parameter $\theta_t$ and collect the new dataset $D_t$.
- Construct the function $h_t(\phi) = \|\nabla r_t(\theta_t) - m(\theta_t; \phi, t)\|^2$.
- $\phi_{t+1} = \frac{\phi_t \exp(-\eta_\phi \nabla h_t(\phi_t))}{\|\phi_t \exp(-\eta_\phi \nabla h_t(\phi_t))\|_1}$. $\triangleright$ One step of Exponentiated gradient descent from $\phi_t$.
- Initialize the model parameter $\theta_{t+1}$.
- while $\|\tilde{m}(\theta_{t+1}; \phi_{t+1}, t + 1)\| \geq \delta$ do
  - $\theta_{t+1} = \theta_{t+1} - \eta \tilde{m}(\theta_{t+1}; \phi_{t+1}, t + 1)$.
- end while

end for

Thus, our goal is to minimize $\sum_{t=1}^T h_t(\phi_t)$ in an online manner, since we can only access $h_t(\phi_t)$ after $\phi_t$ is chosen. To achieve this, we use the classic exponentiated gradient method to update $\phi_t$. Specifically, for any $\phi = [a_1, \ldots, a_b] \in S_b$, define the negative potential function $\psi(\phi) = \sum_{i=1}^b a_i \log a_i$ and its Bregman divergence

$$B_\psi(\phi; \phi') = \psi(\phi) - \psi(\phi') - \langle \nabla \psi(\phi'), \phi - \phi' \rangle = \sum_{i=1}^b a_i \log \frac{a_i}{\tilde{a}_i}.$$  

Then $\phi_{t+1}$ is given by

$$\phi_{t+1} = \arg \min_{\phi \in S_b} \left( \nabla h_t, \phi + \frac{1}{\eta_\phi} B_\psi(\phi; \phi_t) \right) = \frac{\phi_t \exp(-\eta_\phi \nabla h_t(\phi_t))}{\|\phi_t \exp(-\eta_\phi \nabla h_t(\phi_t))\|_1},$$

where $\eta_\phi$ is the learning rate. See Section 6.6 in [Orabona 2019] for the derivation of the last equality. Intuitively, $\frac{1}{\eta_\phi} B_\psi(\phi; \phi_t)$ stabilizes the algorithm by ensuring that $\phi_{t+1}$ remains close to $\phi_t$.

This simplified version of FGD is summarized in Algorithm 1. Note that when updating $\phi$, we only use the last recommendation model $\theta_t$.

**Lemma 1.** Suppose that we have $\|\nabla r_1(\theta)\| \leq M$ for all $\theta \in \Theta$ and $t$. Then $\|\nabla h_t(\phi)\|_\infty \leq 8M^2$ for all $\phi \in S_b$.

**Proof.** By definition, we have

$$h_t(\phi) = \|\nabla r_t(\theta_t) - \sum_{i=1}^b a_i \nabla r_{t-1}(\theta_t)\|^2 = \|\sum_{i=1}^b a_i (\nabla r_t(\theta_t) - \nabla r_{t-1}(\theta_t))\|^2,$$

where we used the fact that $\sum_{i=1}^b a_i = 1$. Direct computation shows that

$$\frac{\partial h_t}{\partial a_i}(\phi) = 2\left(\langle \nabla r_t(\theta_t) - \nabla r_{t-1}(\theta_t), \sum_{j=1}^b a_j (\nabla r_t(\theta_t) - \nabla r_{t-1}(\theta_t)) \rangle\right) = 2\|\nabla r_t(\theta_t) - \nabla r_{t-1}(\theta_t)\| \left\| \sum_{j=1}^b a_j (\nabla r_t(\theta_t) - \nabla r_{t-1}(\theta_t)) \right\| \leq 2\|\nabla r_t(\theta_t)\| + \|\nabla r_{t-1}(\theta_t)\| \left(\sum_{j=1}^b a_j \|\nabla r_t(\theta_t)\| + \|\nabla r_{t-1}(\theta_t)\|\right) \leq 8M^2,$$

where we used Cauchy-Schwarz inequality in (6), the triangle inequality in (7) and the boundedness of the gradients in (8).

Hence, we conclude that $\|\nabla h_t(\phi)\|_\infty \leq 8M^2$. \qed
Proof of Theorem 2. Now we proceed to the proof of Theorem 2. This is a standard result in the online learning literature (see, e.g., Orabona [2019]). For completeness, we present the proof below.

**Proof.** As $\psi$ is $\lambda$-strongly convex with $\lambda = 1$, we have

$$\mathcal{B}_\psi(\phi; \phi') \geq \frac{1}{2} \|\phi - \phi'\|_1^2.$$  \hspace{1cm} (9)

Throughout the proof, we slightly abuse the notation by writing $\eta_\phi = \eta$ and $\nabla h_t = \nabla h_t(\phi_t)$ for simplicity. Notice that by our update rule $\phi_{t+1}$ is given by

$$\phi_{t+1} = \arg \min_{\phi \in S_b} (\eta(\nabla h_t, \phi) + \mathcal{B}_\psi(\phi; \phi_t)).$$

From the first-order optimality condition, we get for any $\phi \in S_b$,

$$\langle \eta \nabla h_t + \nabla \psi(\phi_{t+1}) - \nabla \psi(\phi_t), \phi_{t+1} - \phi \rangle \leq 0$$

$$\Leftrightarrow \eta \langle \nabla h_t, \phi_t - \phi \rangle \leq \eta \langle \nabla h_t, \phi_t - \phi_{t+1} \rangle + \langle \nabla \psi(\phi_{t+1}) - \nabla \psi(\phi_t), \phi - \phi_{t+1} \rangle$$

$$\Leftrightarrow \eta \langle \nabla h_t, \phi_t - \phi \rangle \leq \eta \langle \nabla h_t, \phi_t - \phi_{t+1} \rangle - \mathcal{B}_\psi(\phi; \phi_{t+1}) + \mathcal{B}_\psi(\phi; \phi_t) - \mathcal{B}_\psi(\phi_{t+1}; \phi_t),$$

where we used the three-point equality [Chen and Teboulle, 1993] in the last inequality. Furthermore,

$$\eta \langle \nabla h_t, \phi_t - \phi_{t+1} \rangle - \mathcal{B}_\psi(\phi; \phi_{t+1}) \leq \eta \|\nabla h_t\|_\infty \|\phi_t - \phi_{t+1}\|_1 - \frac{1}{2} \|\phi_t - \phi_{t+1}\|_1^2$$

$$\leq \frac{\eta^2}{2} \|\nabla h_t\|_\infty^2 + \frac{1}{2} \|\phi_t - \phi_{t+1}\|_2^2 - \frac{1}{2} \|\phi_t - \phi_{t+1}\|_1^2$$

$$= \frac{\eta^2}{2} \|\nabla h_t\|_\infty^2.$$

Combining these two bounds, we have

$$\eta \langle \nabla h_t, \phi_t - \phi \rangle \leq \mathcal{B}_\psi(\phi; \phi_t) - \mathcal{B}_\psi(\phi; \phi_{t+1}) + \frac{\eta^2}{2} \|\nabla h_t\|_\infty^2.$$  

Since $h_t(\phi)$ is convex in $\phi$, we have $h_t(\phi_t) - h_t(\phi) \leq \langle \nabla h_t, \phi_t - \phi \rangle$ for any $\phi \in S_b$. By telescoping, we obtain

$$\sum_{t=1}^T (h_t(\phi_t) - h_t(\phi)) \leq \sum_{t=1}^T \langle \nabla h_t, \phi_t - \phi \rangle$$

$$\leq \frac{1}{\eta} \sum_{t=1}^T \left[ \mathcal{B}_\psi(\phi; \phi_t) - \mathcal{B}_\psi(\phi; \phi_{t+1}) + \frac{\eta^2}{2} \|\nabla h_t\|_\infty^2 \right]$$

$$= \frac{1}{\eta} (\mathcal{B}_\psi(\phi; \phi_1) - \mathcal{B}_\psi(\phi; \phi_{T+1})) + \frac{\eta}{2} \sum_{t=1}^T \|\nabla h_t\|_\infty^2$$

$$\leq \frac{1}{\eta} \log b + 32\eta M^4 T,$$

where we used Lemma 8 $\mathcal{B}_\psi(\phi; \phi_{T+1}) \geq 0$ and $\mathcal{B}_\psi(\phi; \phi_t) = \psi(\phi) + \log b \leq \log b$ in the last inequality. Choosing $\eta = c \sqrt{\log b / (TM^4)}$ with some constant $c > 0$ leads to

$$\sum_{t=1}^T (h_t(\phi_t) - h_t(\phi)) \leq O(M^2 \sqrt{T \log b}).$$  \hspace{1cm} (10)
Algorithm 2 Generalized Future Gradient Descent for Smoothed Loss

**Input:** The learning rate $\eta, \eta_\phi$ for updating the model parameter $\theta$ and $\phi$. The initial trajectory buffer $B$.

for $t \in [T]$ do
  Deploy the prediction model $f_{\theta_t}$ with parameter $\theta_t$. Then collect the new dataset $D_t$.
  Initialize the parameter of MFGG $\phi_{t+1}$.
  for Inner loop iteration $k \in K$ do
    $\phi_{t+1} \leftarrow \phi_{t+1} - \eta_\phi \sum_{\theta \in B} \nabla_\phi \|m(\theta; \phi_{t+1}, t) - \nabla r_t(\theta)\|^2$. \▷ Update the meta network.
  end for
  Initialize the trajectory buffer $B = \emptyset$ and model parameter $\theta_{t+1}$.
  end for

end for

Note that (10) holds for any $\phi \in S_b$. In particular, we can set $\phi = \phi^*$ defined by $\phi^* = \arg\min_{\phi \in S_b} \sum_{t=1}^T h_t(\phi)$. Therefore,

$$
\sum_{t=1}^T h_t(\phi_t) \leq \sum_{t=1}^T h_t(\phi^*) + O(M^2 \sqrt{T \log b})
$$

$$
= \min_{\phi \in S_b} \sum_{t=1}^T \|\nabla r_t(\phi_t) - m(\theta_t; \phi, t)\|^2 + O(M^2 \sqrt{T \log b})
$$

$$
\leq \min_{\phi \in S_b} \sum_{t=1}^T \sup_{\theta} \|\nabla r_t(\theta) - m(\theta; \phi, t)\|^2 + O(M^2 \sqrt{T \log b}) = \min_{m \in M} Q[T; m] + O(M^2 \sqrt{T \log b}).
$$

We thus conclude from (4) that

$$
\mathcal{R}_w(T) \leq \frac{2}{w^2 T} (\min_{m \in M} Q[T; m] + O(M^2 \sqrt{T \log b})) + 2\delta^2.
$$

3 A PRACTICAL GENERALIZED FGD ALGORITHM.

Compared with FGD in Algorithm[2], we use a smoothed version of MFGG $\tilde{m}$ for training, which is due to the consideration of minimizing a smoothed loss in (2). For completeness, we also summarize the practical algorithm of the generalized version of FGD in Algorithm[2].

References
