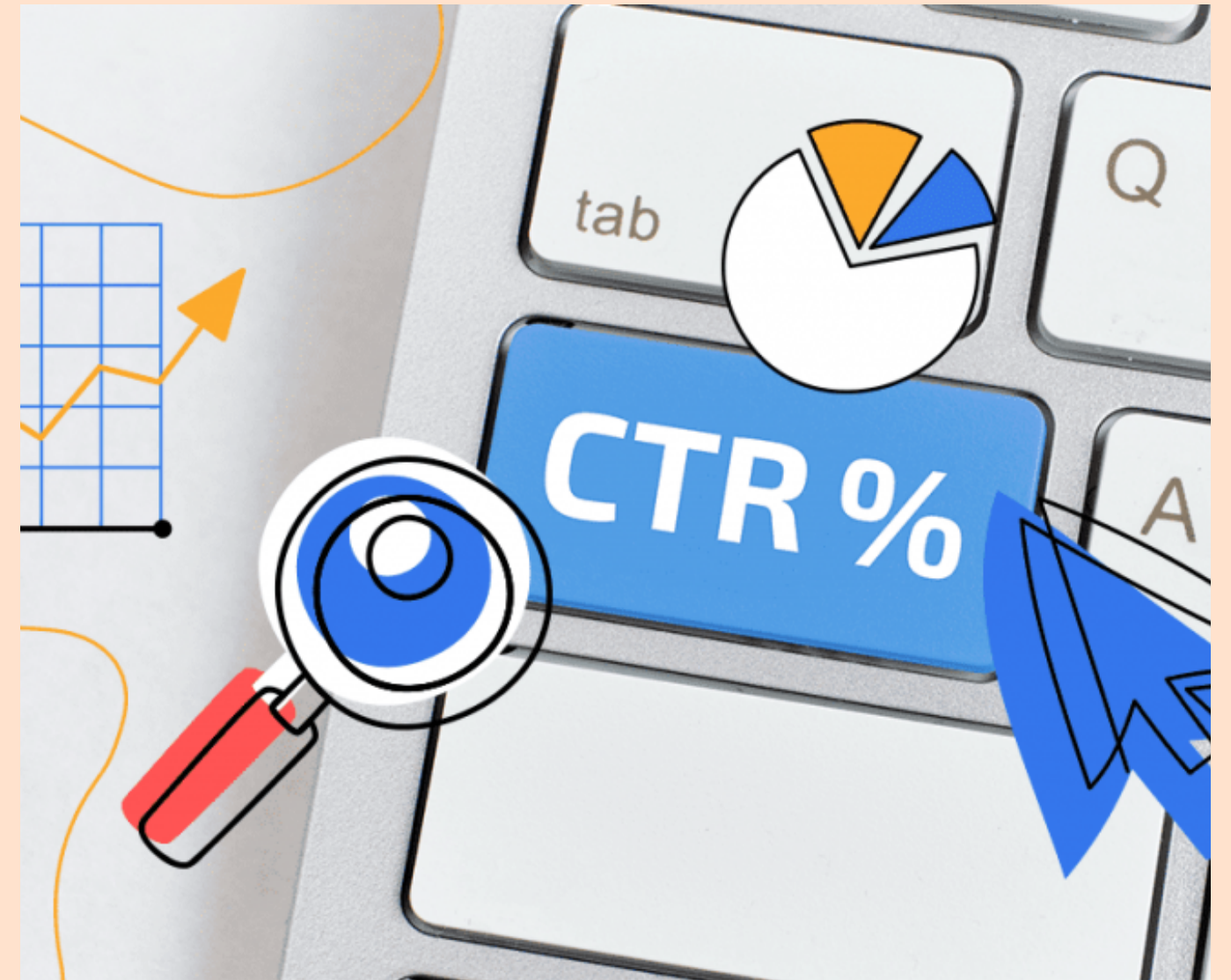


FIELD-AWARE FACTORIZATION MACHINES IN A REAL-WORLD ONLINE ADVERTISING SYSTEM



Elise Chin

Mathilde Da Cruz

Vincent Duchauffour

INTRODUCTION : PRÉSENTATION DE L'ARTICLE

1

- a. Les auteurs
- b. Définitions
- c. Contributions
- d. Contexte

LES AUTEURS

Papier publié en 2017 et présenté à la
World Wide Web Conférence (IW3C2)

Olivier Chapelle
Google

Yuchin Juan
Criteo Research

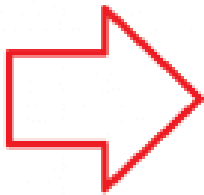
Damien Lefortier
Facebook

DÉFINITIONS

- Computational advertising (CA)
 - ➔ Publicité la plus pertinente selon un contexte
 - ➔ information retrieval, statistical modeling, machine learning, optimization large scale search, text analysis
 - Sponsored search
 - Contextual advertising
 - Display advertising

DÉFINITIONS

Sponsored search



Google

advertising agency

All Images Maps Videos News More Settings Tools

About 32,800,000 results (0.43 seconds)

Digital Marketing Agency - Get visualized online - josewalls.co.uk
(Ad) www.josewalls.co.uk/ ▼
Get optimized your website and campaigns to increase your sales

Full Service Public Relations - Marketing Agency in the US
(Ad) www.frankadvertisingus.com/ ▼
Full service agency with expertise with K&B products manufactured for the US.
Contact Us · Client Portfolio

Talent For Good - createathon.org
(Ad) www.createathon.org/ ▼ +1 804-398-4857
Get started with CreateAthon and make a difference
Inspired Results · Professional Development
Highlights: More Than 100 Partner Organizations, Professional Development, Inspired Results...
About CreateAthon · For Non Profits · CreateAthon Blog · Become A Partner

Advertising agency - Wikipedia
https://en.wikipedia.org/wiki/Advertising_agency ▼
An advertising agency, often referred to as a creative agency, is a business dedicated to creating, planning, and handling advertising and sometimes other forms of promotion and marketing for its clients. An ad agency is generally independent from the client; it may be an internal department or agency that provides an ...
History · Global advertising agency · Client relationships · Advertising effects

List of advertising agencies - Wikipedia
https://en.wikipedia.org/wiki/List_of_advertising_agencies ▼
Largest agencies. The five largest agencies, with their estimated worldwide revenues in 2014: WPP Group, London \$19.0 billion; Omnicom Group, New York City \$15.3 billion; Publicis Groupe, Paris \$9.6 billion; Interpublic Group, New York City \$7.5 billion; Dentsu, Tokyo \$6.0 billion ...

Advertising agency

An advertising agency, often referred to as a creative agency, is a business dedicated to creating, planning, and handling advertising and sometimes other forms of promotion and marketing for its clients.
Wikipedia

People also search for

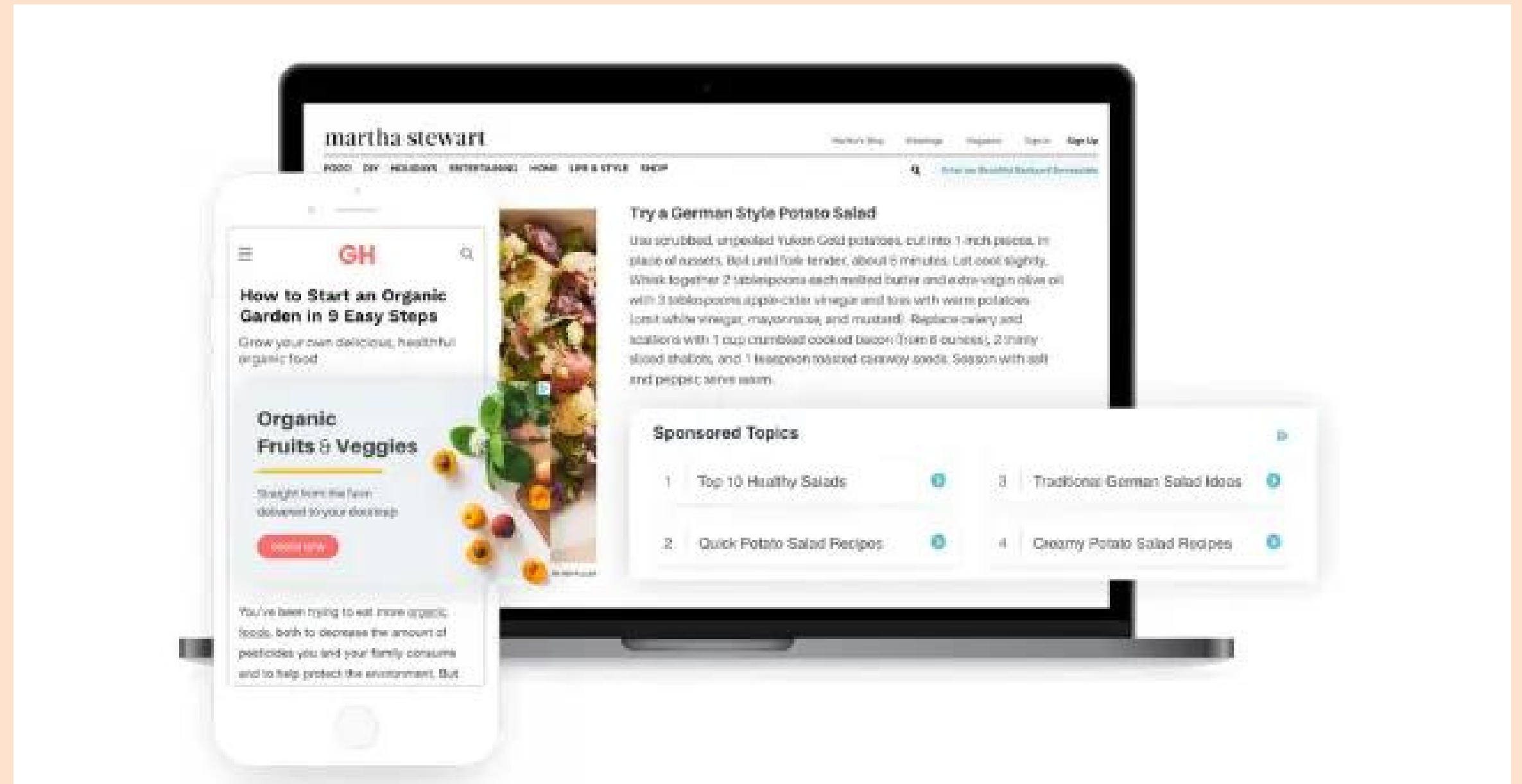
Advertising Marketing Web design Public Relations Graphic design

View 15+ more

Feedback

DÉFINITIONS

Contextual Advertising



DÉFINITIONS

Display Advertising

The screenshot displays the homepage of The Weather Channel website, which is an IBM Business. The header includes a search bar for "Search City or Zip Code", a location indicator for "62° Summerville, SC", and navigation tabs for "Today", "Hourly", "10 Day", "Weekend", "Monthly", "Radar", "Covid-19", and "More Forecasts".

The main content area features several display advertisements:

- Postclick:** A large ad with a woman's profile and the text "We turn your ad clicks into more conversions." with a "See how" button.
- Weather Forecast:** A map of the United States showing "Severe Storms" and "JET STREAM" patterns.
- Why Easter Weekend Forecast Has Us Worried:** A headline with a subtext "A potentially dangerous severe outbreak is possible".
- Here's Who Will See a White Easter:** A map of the United States showing the distribution of white Easter eggs.
- Stimulus Payments Set to Go Out:** A small ad with a map of the United States.
- Emgality:** An advertisement for migraine treatment featuring a woman and child, with text: "Even if you have just one migraine headache day a week, ask your doctor if Emgality is right for you." and a "Discover Emgality" button.

DÉFINITIONS

- Computational advertising (CA)
 - ➔ Publicité la plus pertinente selon un contexte
 - ➔ information retrieval, statistical modeling, machine learning, optimization large scale search, text analysis
 - Sponsored search
 - Contextual advertising
 - Display advertising
- Click-through / Click-through Rate (CTR)
 - ➔ nombre de clics / nombre de vues

DÉFINITIONS

- Computational advertising (CA)
 - ➔ Publicité la plus pertinente selon un contexte
 - ➔ information retrieval, statistical modeling, machine learning, optimization large scale search, text analysis
 - Sponsored search
 - Contextual advertising
 - Display advertising
- Click-through / Click-through Rate (CTR)
 - ➔ nombre de clics / nombre de vues
- Conversion / Conversion rate
 - ➔ Action spécifique de l'utilisateur

CONTRIBUTIONS

Utiliser le FFM dans un système de production pour prédire le CTR et CR.

Cette méthode est-elle aussi efficace dans le monde réel que dans des compétitions kaggle ?

1. Application de FFM dans un cadre industriel

2. Entraînement distribué

3. Pré mature Warm Start

CONTEXTE

Research Prediction Competition

Display Advertising Challenge

Predict click-through rates on display ads

\$16,000
Prize Money

717 teams · 8 years ago

Overview

Data

Code

Discussion

Leaderboard

Rules

Late Submission

...

Overview

Description

Evaluation

Prizes

Timeline

About-criteo

Winners

Display advertising is a billion dollar effort and one of the central uses of machine learning on the Internet. However, its data and methods are usually kept under lock and key. In this research competition, CriteoLabs is sharing a week's worth of data for you to develop models predicting ad click-through rate (CTR). Given a user and the page he is visiting, what is the probability that he will click on a given ad?

Prediction
Contest

criteo.

The goal of this challenge is to benchmark the most accurate ML algorithms for CTR estimation. All winning models will be released under an open source license. As a participant, you are given a chance to access the traffic logs from Criteo that include various undisclosed features along with the click labels.

Kaggle compétition : Predict ad CTR
"Given a user and the page he is visiting, what is the probability that he will click on a given ad ?"

7 jours de données : prédire le jour suivant

1. Field-aware Factorization Machines

2/3 ? Logistic Regression with cross-features

CONTEXTE

Système de production = Contraintes et objectifs spécifiques ! Différent d'une compétition académique.



Cas du Netflix Prize (2006)

- ➔ \$1 million à l'équipe qui améliorerait l'accuracy de leur système de recommandation d'au moins 10%
- ➔ Codes gagnants jamais utilisés par Netflix
- ✗ Code écrit pour 100M notes (VS +5Milliards en réalité)
- ✗ Code ne s'adapte pas à de nouvelles notes
- ✗ Trop de "Engineering efforts" pour le gain
- ✗ Shift des recommandations (transition DVD streaming)
= données obsolètes

CONTEXTE

SOTA Kaggle challenge = SOTA real world ?

Différences entre la compétition et le monde réel :

- Prédire si clic ou non
- Données limitées
- Tout prédire avant la fin de la compétition
- etc.

- Prédire Click-Through Rate ET Conversion Rate
- Données continues
- Prédire chaque valeur au bon moment = TEMPS LIMITÉ !
- etc.

MODÈLES

2

- a. Linéaire
- b. Polynomial de degré 2 (Poly2) et Factorization Machine (FM)
- d. Field-aware Factorization Machine (FFM) : SOTA

DATASET

$$\{\mathbf{x}_i, y_i\}_{i=1}^m$$

y_i label

\mathbf{x}_i vecteur de features
de dimension n

Terminologie :

- **Champ** : {Publisher, Advertiser}
- **Feature** : {ESPN, Vogue, NBC, Nike, ...}

		Publisher	Advertiser
+80	−20	ESPN	Nike
+10	−90	ESPN	Gucci
+0	−1	ESPN	Adidas
+15	−85	Vogue	Nike
+90	−10	Vogue	Gucci
+10	−90	Vogue	Adidas
+85	−15	NBC	Nike
+0	−0	NBC	Gucci
+90	−10	NBC	Adidas

Table 1: An artificial CTR data set, where $+$ ($-$) represents the number of clicked (unclicked) impressions.

LINÉAIRE

Formulation :

$$\phi_{LM}(\mathbf{w}, \mathbf{x}) = \mathbf{w}^T \mathbf{x} = \sum_{j \in C_1} w_j x_j$$

où C_1 est l'ensemble des éléments non nuls dans \mathbf{x}

Exemple :

$$\phi_{LM}(\mathbf{w}, \mathbf{x}) = w_{ESPN} x_{ESPN} + w_{Vogue} x_{Vogue} + w_{Gucci} x_{Gucci} + \dots$$

Limite : ne permet pas de représenter l'effet d'une feature sur une autre (e.g. taux de clics plus élevé pour (Vogue, Gucci) que (ESPN, Gucci))

		Publisher	Advertiser
+80	-20	ESPN	Nike
+10	-90	ESPN	Gucci
+0	-1	ESPN	Adidas
+15	-85	Vogue	Nike
+90	-10	Vogue	Gucci
+10	-90	Vogue	Adidas
+85	-15	NBC	Nike
+0	-0	NBC	Gucci
+90	-10	NBC	Adidas

Table 1: An artificial CTR data set, where + (−) represents the number of clicked (unclicked) impressions.

POLY2

Formulation :

$$\phi_{Poly2}(\mathbf{w}, \mathbf{x}) = \sum_{j_1, j_2 \in C_2} w_{j_1, j_2} x_{j_1} x_{j_2}$$

où C_2 est l'ensemble des couples d'éléments non nuls dans \mathbf{x}

FACTORIZATION MACHINE (FM)

Formulation :

$$\phi_{FM}(\mathbf{w}, \mathbf{x}) = \sum_{j_1, j_2 \in C_2} \langle \mathbf{w}_{j_1}, \mathbf{w}_{j_2} \rangle x_{j_1} x_{j_2}$$

où :

- \mathbf{w}_{j_1} et \mathbf{w}_{j_2} sont deux vecteurs de dimension k , k étant un hyperparamètre
- C_2 est l'ensemble des couples d'éléments non nuls dans \mathbf{x}

		Publisher	Advertiser
+80	-20	ESPN	Nike
+10	-90	ESPN	Gucci
+0	-1	ESPN	Adidas
+15	-85	Vogue	Nike
+90	-10	Vogue	Gucci
+10	-90	Vogue	Adidas
+85	-15	NBC	Nike
+0	-0	NBC	Gucci
+90	-10	NBC	Adidas

Table 1: An artificial CTR data set, where + (-) represents the number of clicked (unclicked) impressions.

Exemple :

$$\phi_{Poly2}(\mathbf{w}, \mathbf{x}) = w_{ESPN, Gucci} \times x_{ESPN} x_{Gucci} + w_{Vogue, Gucci} \times x_{Vogue} x_{Gucci} + \dots$$

$$\phi_{FM}(\mathbf{w}, \mathbf{x}) = \langle \mathbf{w}_{ESPN}, \mathbf{w}_{Gucci} \rangle \times x_{ESPN} x_{Gucci} + \langle \mathbf{w}_{Vogue}, \mathbf{w}_{Gucci} \rangle \times x_{Vogue} x_{Gucci} + \dots$$

FIELD-AWARE FACTORIZATION MACHINE (FFM)

Formulation :

$$\phi_{FFM}(\mathbf{w}, \mathbf{x}) = \sum_{j_1, j_2 \in C_2} \langle \mathbf{w}_{j_1, f_2}, \mathbf{w}_{j_2, f_1} \rangle x_{j_1} x_{j_2}$$

où :

- f_1 et f_2 sont respectivement les champs des features j_1 et j_2
- w_{j_1, f_2} et w_{j_2, f_1} sont deux vecteurs pour les features j_1 et j_2 dans le contexte des champs f_2 et f_1 respectivement
- C_2 est l'ensemble des couples d'éléments non nuls dans \mathbf{x}

Exemple :

Clicked	Publisher (P)	Advertiser (A)	Gender (G)
Yes	ESPN	Nike	Male

FM

$$\mathbf{w}_{\text{ESPN}} \cdot \mathbf{w}_{\text{Nike}} + \mathbf{w}_{\text{ESPN}} \cdot \mathbf{w}_{\text{Male}} + \mathbf{w}_{\text{Nike}} \cdot \mathbf{w}_{\text{Male}}$$

FFM

$$\mathbf{w}_{\text{ESPN}, A} \cdot \mathbf{w}_{\text{Nike}, P} + \mathbf{w}_{\text{ESPN}, G} \cdot \mathbf{w}_{\text{Male}, P} + \mathbf{w}_{\text{Nike}, G} \cdot \mathbf{w}_{\text{Male}, A}$$

FFM APPLIQUÉ DANS UN CADRE INDUSTRIEL

3

- a. Approches Offline vs Online
- b. Résultats

APPROCHES OFFLINE VS ONLINE

Offline learning :

Apprentissage fondé sur un dataset complet

Online learning :

La complétude des données d'entraînement n'est pas assurée. Complétion des données effectué selon un plan programmé

Méthode générale :

Comparaison du FFM avec une régression logistique comme baseline (L-BFGS avec warn start).

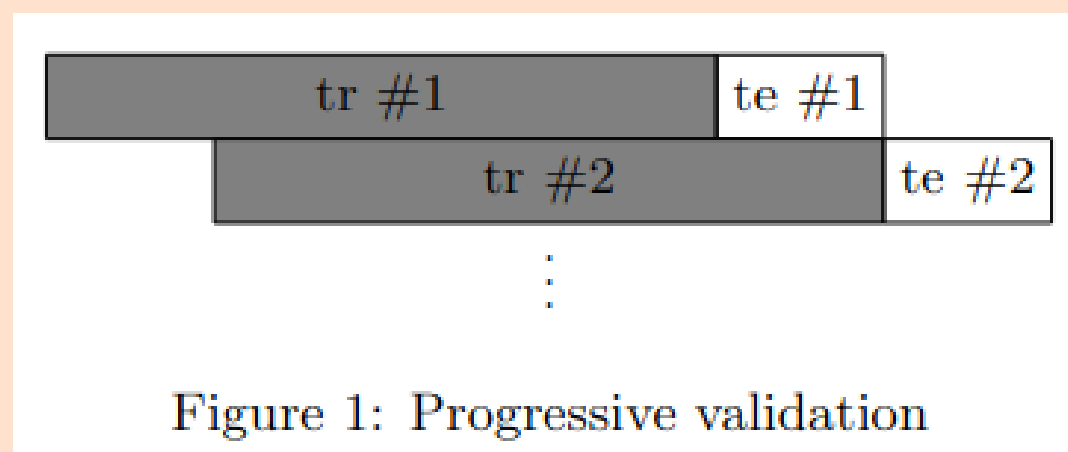
Utilisation du hashing trick pour réduire la dimension des données

Les variables prédites par les modèles sont les CTR et CR.

APPROCHES OFFLINE

Protocole :

- validation progressive
- réduction de la taille de l'espace de hashing afin de garantir le même nombre de paramètres pour la LR et le FFM



$$\mathbf{LL}(p) = - \sum_{i=1}^N y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \quad (2)$$

$$\mathbf{NLL}(p) = \frac{\mathbf{LL}(\bar{p}) - \mathbf{LL}(p)}{\mathbf{LL}(\bar{p})} \quad (3)$$

$$\mathbf{Utility} = \sum_i \int_0^{p(\mathbf{x}_i)v_i} (y_i \cdot v_i - \tilde{c}) \Pr(\tilde{c} | c_i) d\tilde{c} \quad (4)$$

APPROCHES ONLINE

Protocole :

- A/B test
- entraînement sur données live (~5B de display)
- rafraîchissement synchronisé des données pour LR et FFM
- utilisation d'un ROI comme métrique de comparaison avec la méthode offline

RÉSULTATS

Prediction model with FFM	NLL on all advertisers	NLL on small advertisers
CTR	+3.71%▲	+5.9%▲
CTR + CR	+1.21%▲	+6.2%▲

Prediction model with FFM	Utility $_{\beta=10}$ on all advertisers	Utility $_{\beta=10}$ on small advertisers	Utility $_{\beta=1000}$ on all advertisers	Utility $_{\beta=1000}$ on small advertisers
CTR	+6.29%▲	+9.70%▲	+2.22%▲	+4.39%▲
CTR + CR	+11.42%▲	+38.44%▲	+5.43%▲	+18.34%▲

Observation :
Meilleurs résultats sur les
petits annonceurs

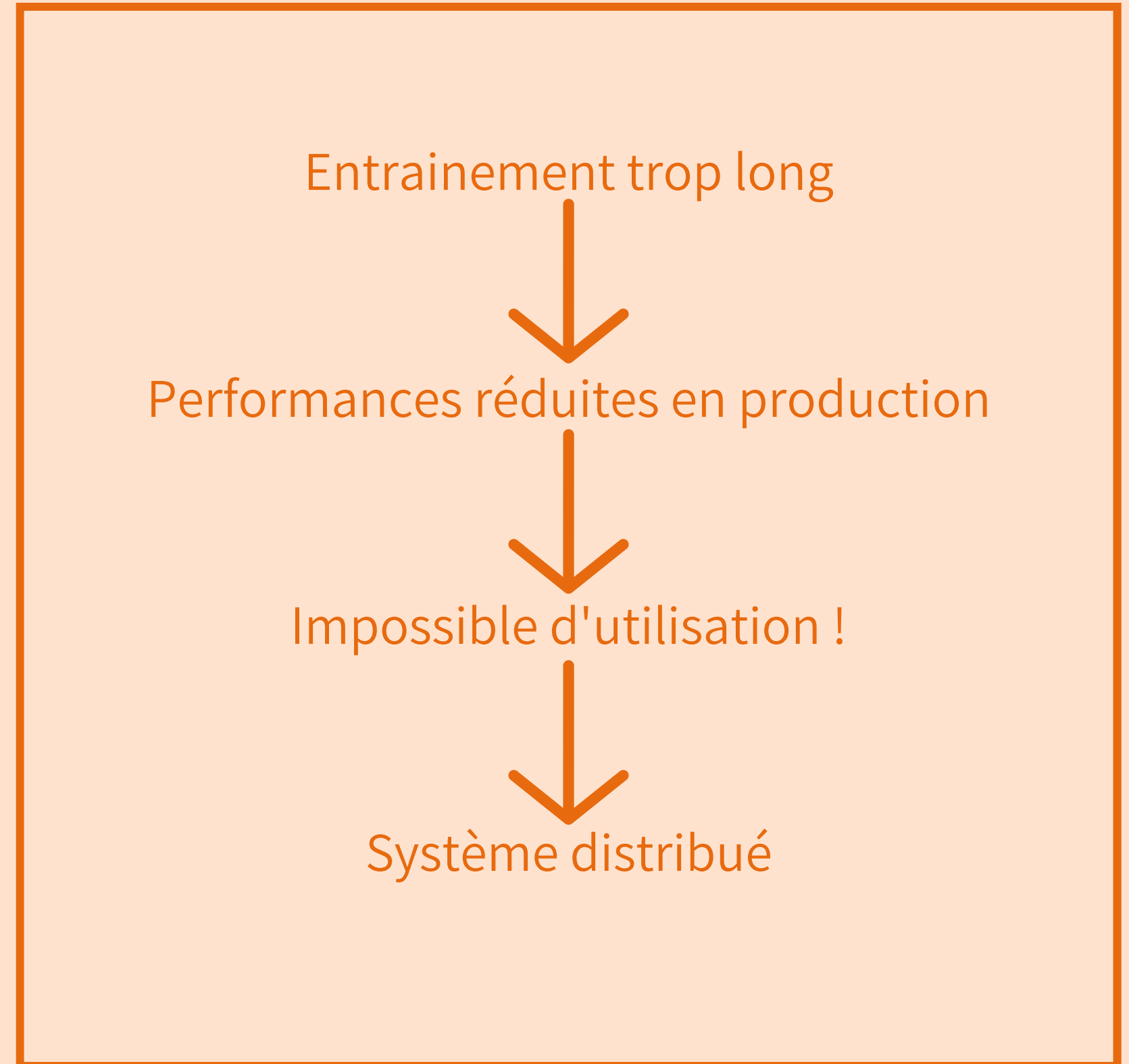
Prediction model with FFM	ROI on all advertisers	ROI on small advertisers
CTR + CR	+0.97%▲	+2.61%▲

RÉDUCTION DU TEMPS D'ENTRAÎNEMENT

4

- a. Entraînement distribué
- b. Pre-mature warm-start

ENTRAÎNEMENT DISTRIBUÉ



ENTRAÎNEMENT DISTRIBUÉ

2 méthodes pour distribuer
l'algorithme de descente du gradient

Distribution synchrone :

Agrégation des modèles après une
certaines quantités de données
traités (en général 1 epoch)

➡ Iterative Parameter Mixing (IPM)

Distribution asynchrone :

Des machines sont dédié au stockage
du modèle tandis que d'autres met à
jour le modèle avec leurs copies
locales

➡ Parameter Server Approach

ENTRAÎNEMENT DISTRIBUÉ

$$\text{speed-up} = \# \text{machines} \times \frac{\# \text{epochs with multiple machines}}{\# \text{epochs with one machine}}$$

Le choix se porte sur la distribution synchrone du gradient car :

- implémentation technique simple
- suppose que le temps de calcul est équitablement distribué entre les machines
- le coût de communication entre les machines est négligeable

Cependant :

- la convergence réduit à mesure que l'on augmente le nombre de machines
- il faut donc également augmenter la valeur du learning rate

η	#epochs	log loss
0.2	157	0.44585
0.5	70	0.44569
1.0	37	0.44590
2.0	26	0.44622
3.0	21	0.44654
4.0	19	0.44688
5.0	18	0.44721

ENTRAÎNEMENT DISTRIBUÉ

En modifiant l'algorithme IPM de telle façon à ce que le gradient soit le résultat d'une agrégation, on peut avoir un learning rate élevé et une loss faible.

On peut ainsi accélérer l'entraînement de :
32 * (8/22) soit 12 fois

Algorithm 2 Improved IPM for AdaGrad

```

1: Spread  $m$  data points into  $k$  machines
2: Initialize  $\mathbf{w}$ 
3: Initialize  $G \leftarrow I$ 
4: for  $t \in \{1, \dots, T\}$  do ▷  $T$ : number of epochs
5:   Let  $\mathbf{w}_i \leftarrow \mathbf{w} \quad \forall i \in \{1, \dots, k\}$ 
6:   Let  $G_i \leftarrow G \quad \forall i \in \{1, \dots, k\}$ 
7:   for  $i \in \{1, \dots, k\}$  parallel do
8:     for each data point do
9:       Calculate the gradient  $\mathbf{g}$ 
10:      Update  $G_i$ :  $G_i \leftarrow G_i + \text{diag}(\mathbf{g}\mathbf{g}^T)$ 
11:      Update  $\mathbf{w}_i$ :  $\mathbf{w}_i \leftarrow \mathbf{w}_i - \eta G_i^{-1/2} \mathbf{g}$ 
12:    $\mathbf{w} \leftarrow \sum_{i=1}^k \mathbf{w}_i / k$ 
13:    $G \leftarrow \sum_{i=1}^k G_i$ 

```

PRE-MATURE WARM-START

Warm-start :

initialisation d'un modèle avec les poids d'un autre modèle

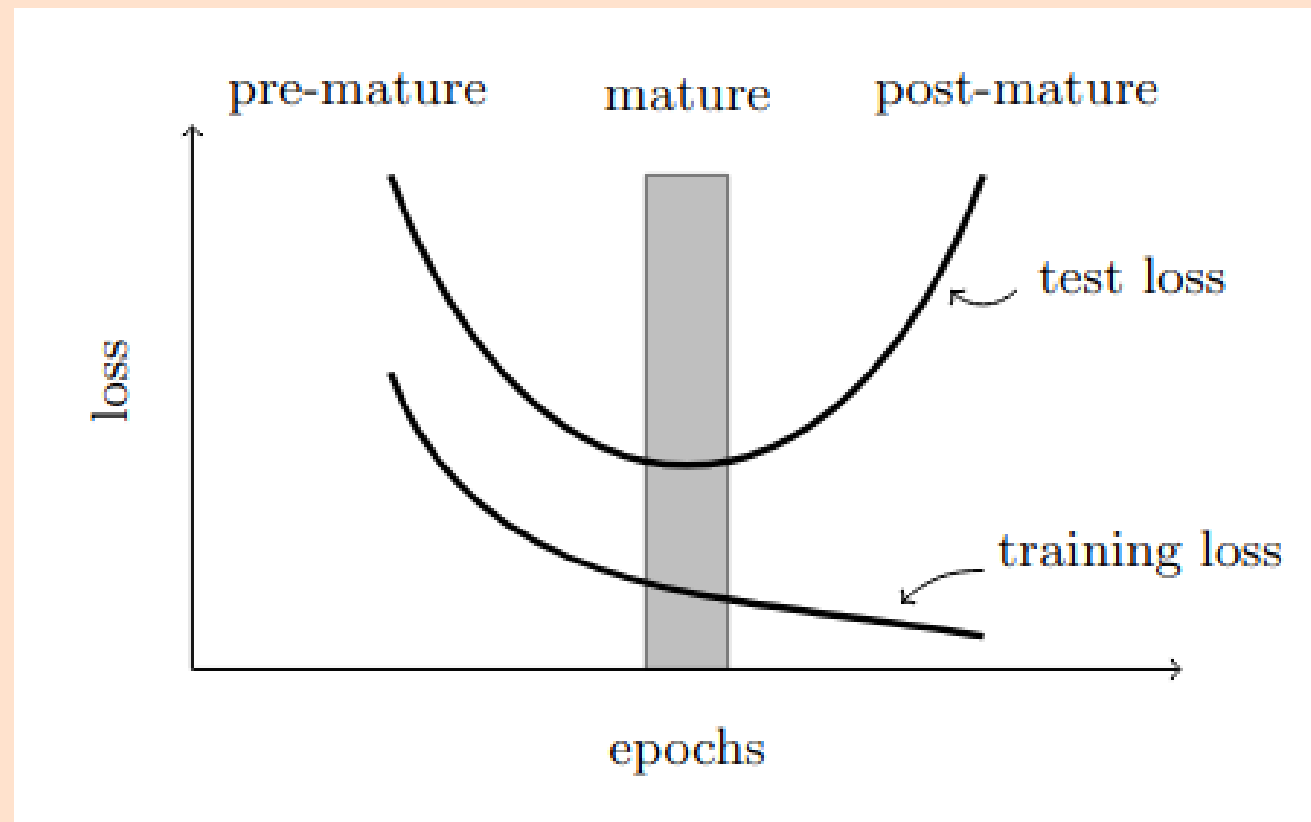
Early-stopping :

Arrêt automatique d'un entraînement lorsqu'une métrique ne présente aucune amélioration

État pre-mature : le modèle est entraîné avec trop peu d'epoch

État mature : le modèle est entraîné avec un nombre suffisant d'epoch

État post-mature : le modèle est entraîné avec un trop grand nombre d'epoch



PRE-MATURE WARM-START

Naive warm-start :

- Initialisation : modèle "mature"
- Après entraînement : obtention d'un modèle "post-mature", donc overfitting sur les anciennes données
- Observation : dégradation des performances au fur et à mesure des apprentissages.

Pre-mature warm-start (solution) :

- Initialisation : modèle "pre-mature"
- Après entraînement : obtention d'un modèle "mature", donc apprentissage des nouvelles données sans sur-apprentissage sur les anciennes données.
- Nouveau modèle mature pour la prédiction, nouveau modèle pre-mature pour l'initialisation du prochain modèle

Algorithm 4 Our proposed "pre-mature" warm-start

Require: an initial model \mathbf{w}_{-1}

$\mathbf{w} \leftarrow \mathbf{w}_0 \leftarrow \mathbf{w}_{-1}$

calculate the validation loss L_0

for $t \in \{1, \dots, T\}$ **do**

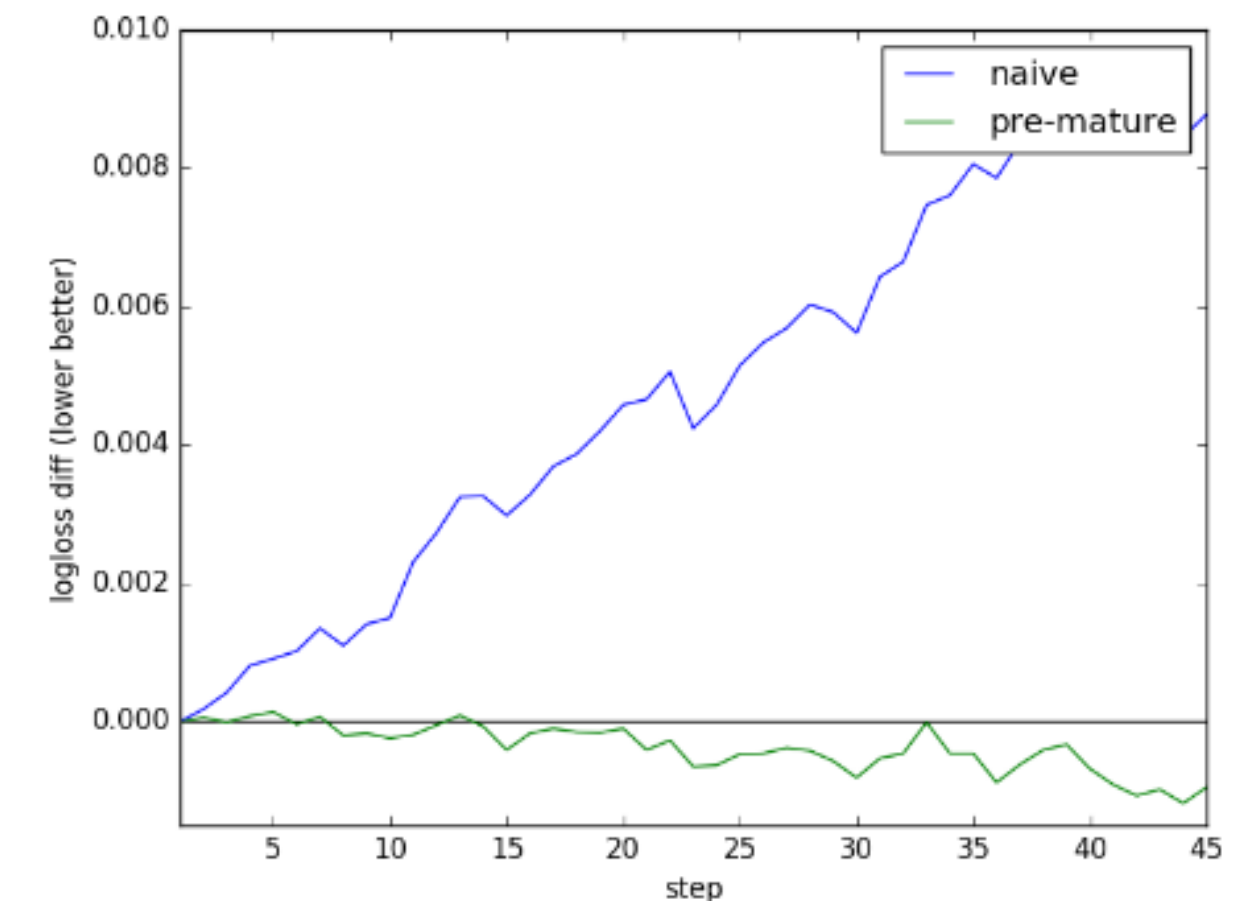
 update \mathbf{w}

$\mathbf{w}_t \leftarrow \mathbf{w}$

 calculate the validation loss L_t

if $L_t > L_{t-1}$ **then**

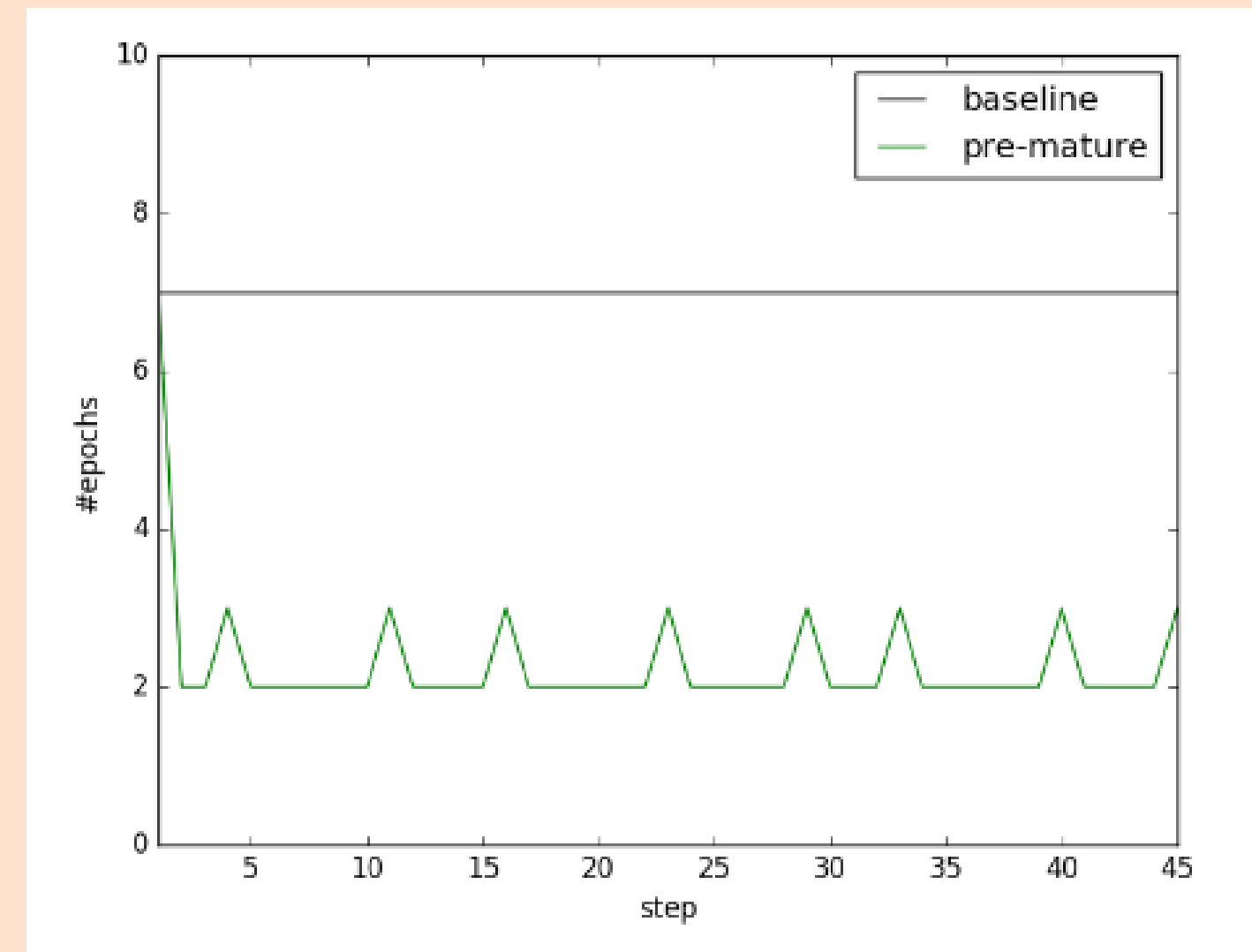
return $(\mathbf{w}_{t-1}, \mathbf{w}_{t-2})$



PRE-MATURE WARM-START

Le pre-mature warm start permet ainsi :

- de réduire significativement le nombre d'epoch nécessaire (et donc le temps d'entraînement)
- de conserver des performances équivalentes



CONCLUSION



Field-aware Factorization Machines
peut être déployé avec succès dans
des systèmes de publicité en ligne à
grande échelle !

- Amélioration des "business metrics"
- Meilleure généralisation que la regression logistique (petits annonceurs)
- L'entrainement a été accéléré efficacement
 - distributed learning
 - warm start