Accelerated refresh of Generalized Linear Models in production

Matthieu KIRCHMEYER (presenter), Syrine KRICHENE, Eugene KHARITONOV

Criteo

Motivation: refreshing GLMs in production at reduced costs

- Production models at Criteo (GLMs) are refreshed around 4 times a day
- A single GLM learning takes 5 - 6 hours
- The training data from consecutive learnings (21 days, 3bln records, 3TB) overlap

Reducing learning time

How could we reuse the information from learning t in learning $t+1$?

Our approach a.k.a "model update" (MU)

As GLMs are **smooth** and **convex**, MU can be viewed as sequential

- Bayesian updates with a **Laplace** approximation of the weight posterior [1]
- batch learnings with a **Taylor** expansion of the loss ("optimization" view)

Practically we propose to:

- carry over the **Hessian** $H(\theta_t^{opt})$ with the optimal θ_t^{opt} at t for learning $t+1$
- to limit the training data to **non-overlapping sets** at each iteration:
  - the 1st learning reads 21 days ("seed model")
  - the following learnings take varying training windows from 3h to 3 days

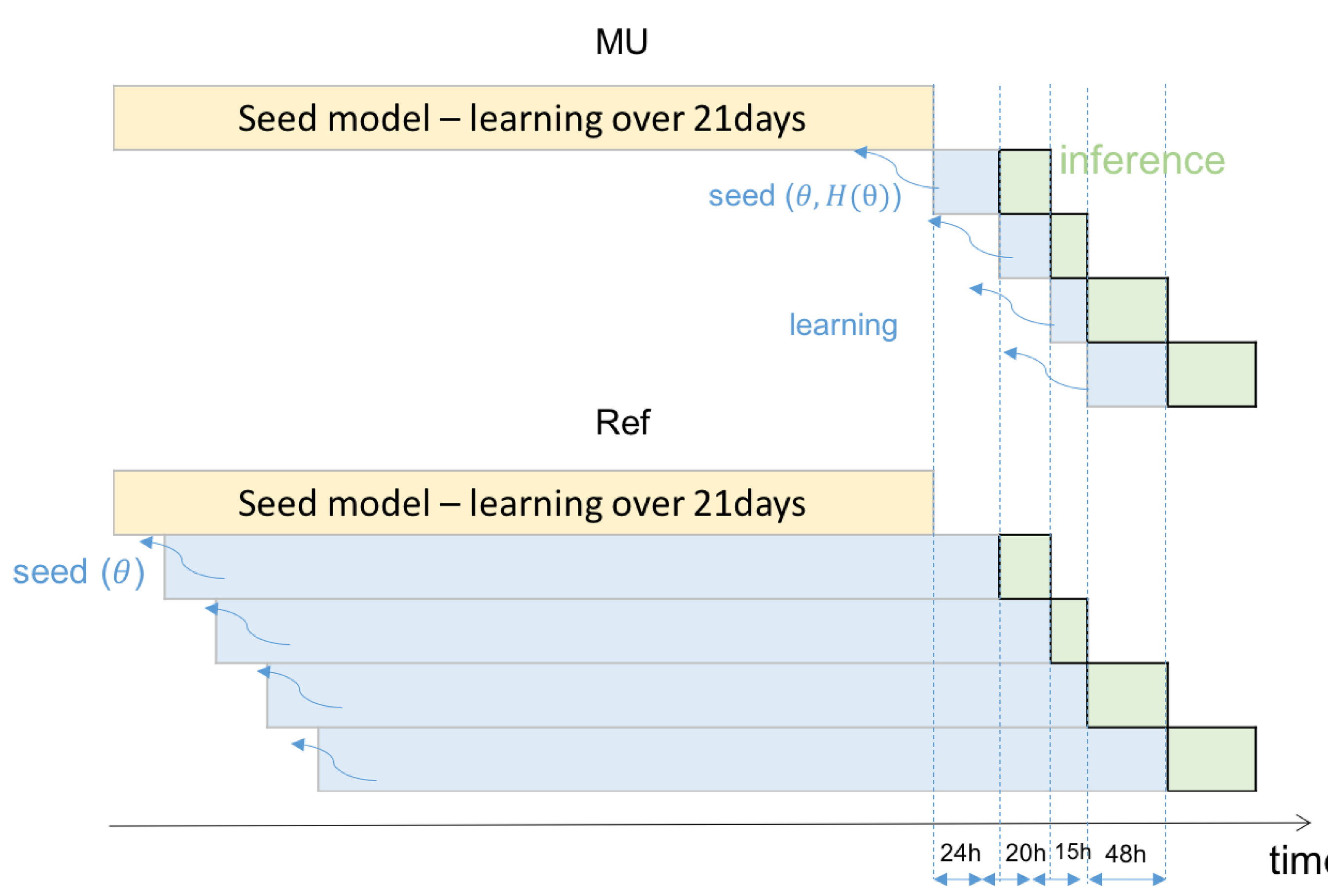

Figure 1: A typical production learning sequence for the reference and MU setting

Mathematical derivation of MU (optimization view)

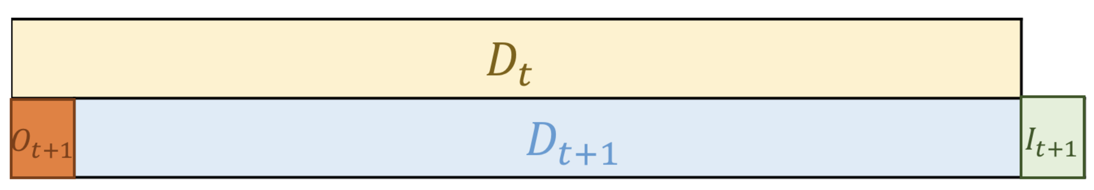

Figure 2: Training data for 2 consecutive learnings t and $t+1$

We assume that:

$$\mathrm{Loss}(\mathrm{D}_{t+1};\theta) = \mathrm{Loss}(I_{t+1};\theta) + \mathrm{Loss}(\mathrm{D}_t;\theta)$$

Around θ_t^{opt} at learning t we form a 2nd order approximation of $\mathrm{Loss}(\mathrm{D}_t;\theta)$ as

$$\mathrm{Loss}(\mathrm{D}_t;\theta) = C + \frac{1}{2}(\theta-\theta_t^{opt})^\intercal H(\theta_t^{opt})(\theta-\theta_t^{opt})$$

The final optimization problem for $t+1$ (solved through L-BFGS) is:

$$\mathrm{argmin}_{\theta\in\mathbb{R}^p}\,\mathrm{Loss}(I_{t+1};\theta) + \frac{1}{2}(\theta-\theta_t^{opt})^\intercal H(\theta_t^{opt})(\theta-\theta_t^{opt})$$

- The Hessian at $t+1$ is updated from the one at t
- Regularization is included by fixing the Hessian for the seed model to $H_0 = \lambda I$

The diagonal approximation of the Hessian

Our input space is of size 2^{26} requiring a storage-efficient representation of the Hessian.

We approximate the Hessian using a **diagonal** matrix.

Input: I_{t+1}; θ_t^{opt}; $H(\theta_t^{opt})$
Output: θ_{t+1}^{opt}; $H(\theta_{t+1}^{opt})$

$$\theta_{t+1}^{opt} = \mathrm{argmin}_{\theta\in\mathbb{R}^p}\,\mathrm{Loss}(I_{t+1};\theta) + \frac{1}{2}(\theta-\theta_t^{opt})^\intercal H(\theta_t^{opt})(\theta-\theta_t^{opt})$$

$$H(\theta_{t+1}^{opt}) = H(\theta_t^{opt}) + \mathrm{diag}\left(\frac{\partial^2 \mathrm{Loss}(I_{t+1};\theta)}{\partial\theta^2}|_{\theta=\theta_{t+1}^{opt}}\right)$$

Algorithm 1: Model update diagonal approximation recursion

Online behavior of MU on a Click prediction model

The model that was AB tested is a logistic loss predicting pClick given a display.

This model, trained separately over 3 platforms (US, AS, EU), is used to predict:

- CTR (1)
- sales volume given a display (combination with other models) (2)
- sales amount given a display (combination with other models) (3)

Learning time and refresh frequency over all platforms

- Refresh: $1.6\times$ more learnings for the MU configuration per day
- Cluster consumption: $9\times$ speedup during learning for the MU configuration

Online metrics (AB test)

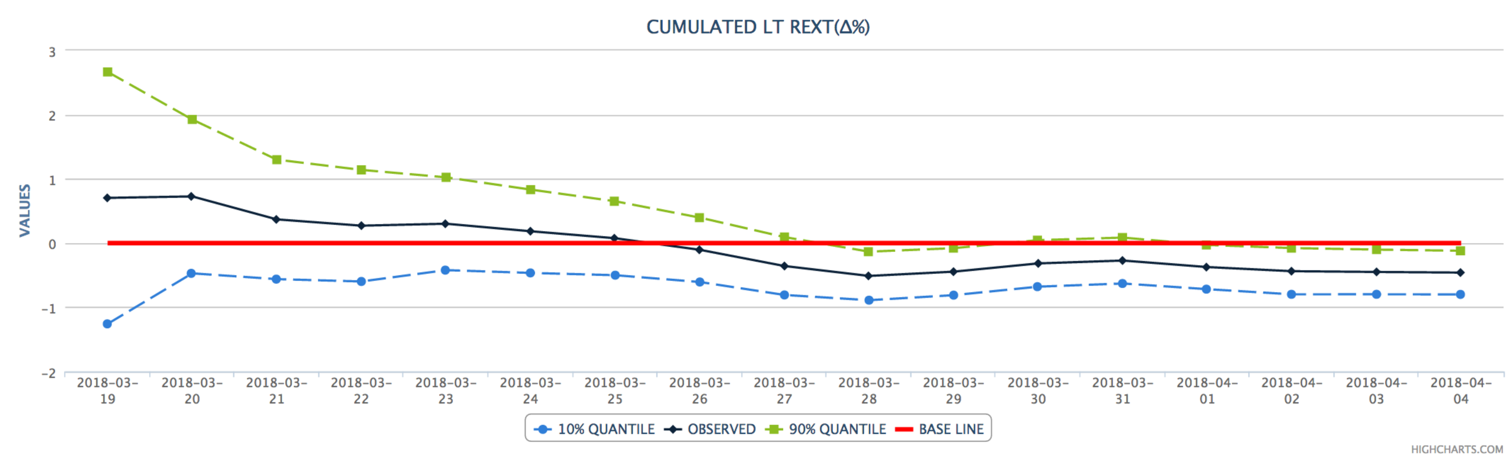

Figure 3: Aggregated online metric over Europe, Asia, USA platforms (2 weeks)

- Neutral long term revenue difference aggregated over (1), (2), (3) and 3 platforms
- Comparable long term behavior for (1), (2), (3) per platform
- **Drift** MU vs. ref on metrics for (1) on US and AS but neutral impact for EU (also observed offline on validation LLH)

Futher investigation will be done into per platform behavior over a longer period.

Offline applications of MU for GLMs

We ran MU on **logistic** and **geometric** losses used at Criteo in an **offline** setting.

We measured the effect on the metrics over consecutive iterations for the Criteo

- bidding,
- recommendation,
- look and feel (L&F) products

MU on various Criteo models

Model	Metric	Diff (%)	Speedup
Logistic Clicks	LLH	0.28	9
Geometric NbSales	RMSE	-0.030	5
Bidding NbSales pipeline (2)	Utility [2]	0.43	7
Recommendation pipeline	Discounted Cumulative Gain	0.023	5
L&F ProductView	Inverse Propensity Score	0.096	3.6

Table 1: Relative difference on the mean over 7 consecutive Model Update validations

Conclusion

Online: Careful monitoring of drift is necessary

- Big win on cluster consumption and refresh frequency
- The diagonal approximation holds well in the first weeks but degrades over time
- Drift on CTR and long term metric per platform will be further investigated

Maintaining MU online would require scheduling automatic resets of the models.

Offline: MU can be applied to all GLM losses used at Criteo

The gains of MU offline are the biggest in the short term.

- Average speedup of $\times 5$
- The effect of drift is negligeable for a typical test configuration

Bibliography

[1] O. Chapelle, E. Manavoglu, and R. Rosales. Simple and scalable response prediction for display advertising. *Transactions on Intelligent Systems and Technology*, 2014

[2] O.Chapelle. Offline Evaluation of Response Prediction in Online Advertising Auctions. *WWW*, 2015

m.kirchmeyer@criteo.com