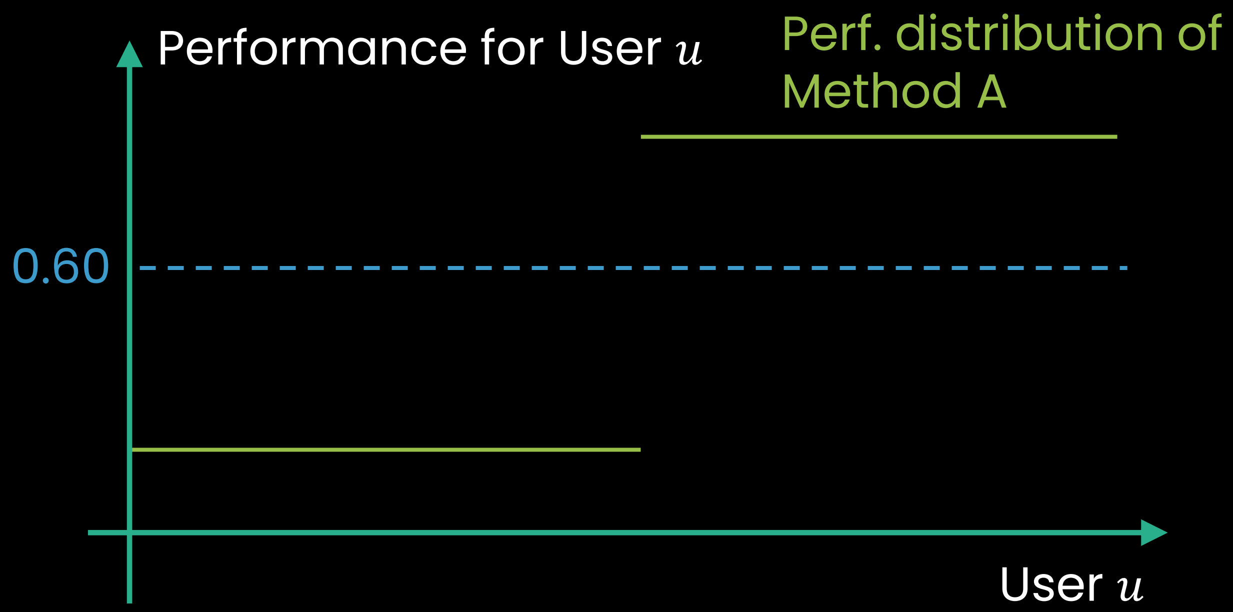
Safe Collaborative Filtering

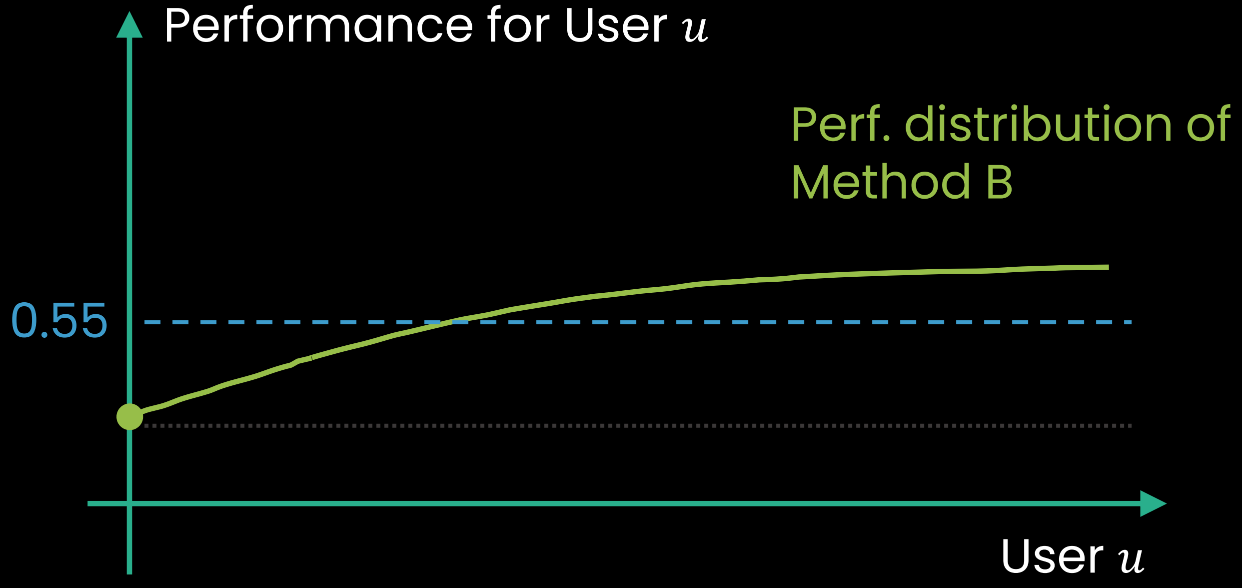
Riku Togashi*, Tatsushi Oka†, Naoto Ohsaka*, Tetsuro Morimura*
*CyberAgent †Department of Economics, Keio University

Which Variant Is Better?

- Suppose the following results in an A/B test

	Average of User CTR
Variant A	0.60
Variant B	0.55

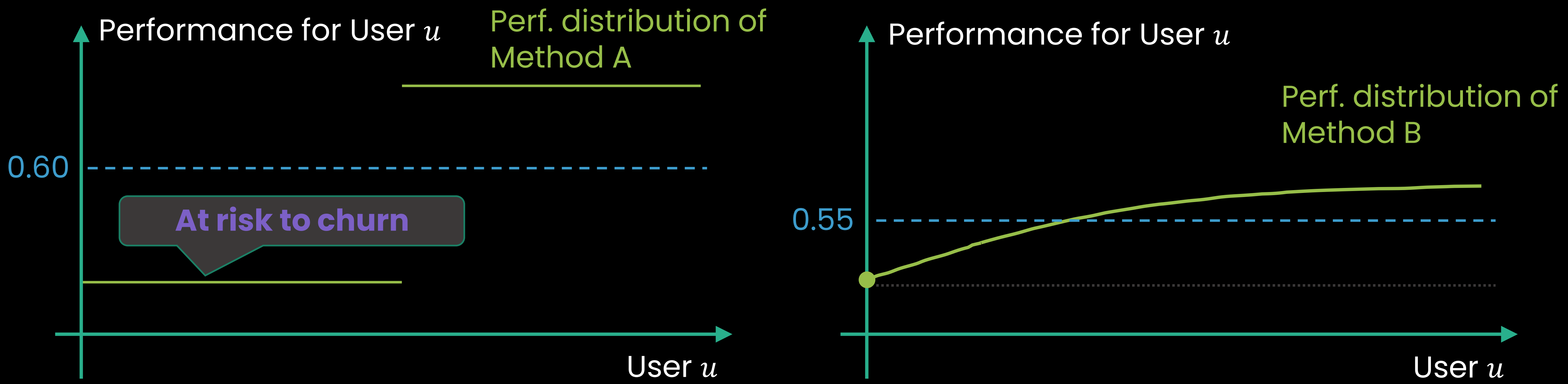




User-Oriented Safety

- We often want ***to avoid the churn of less-satisfied users***
  - Monetization relies on *user retention/growth*
  - Subscription-based services: e.g., video/music streaming platforms

- Maximizing **user-average** performance (e.g., Mean nDCG) is **not** safe



Empirical Risk Minimization

Standard ERM

$$\min_{\theta} \mathbb{E}_{p(x,y)}[\ell(f_\theta(x), y)]$$

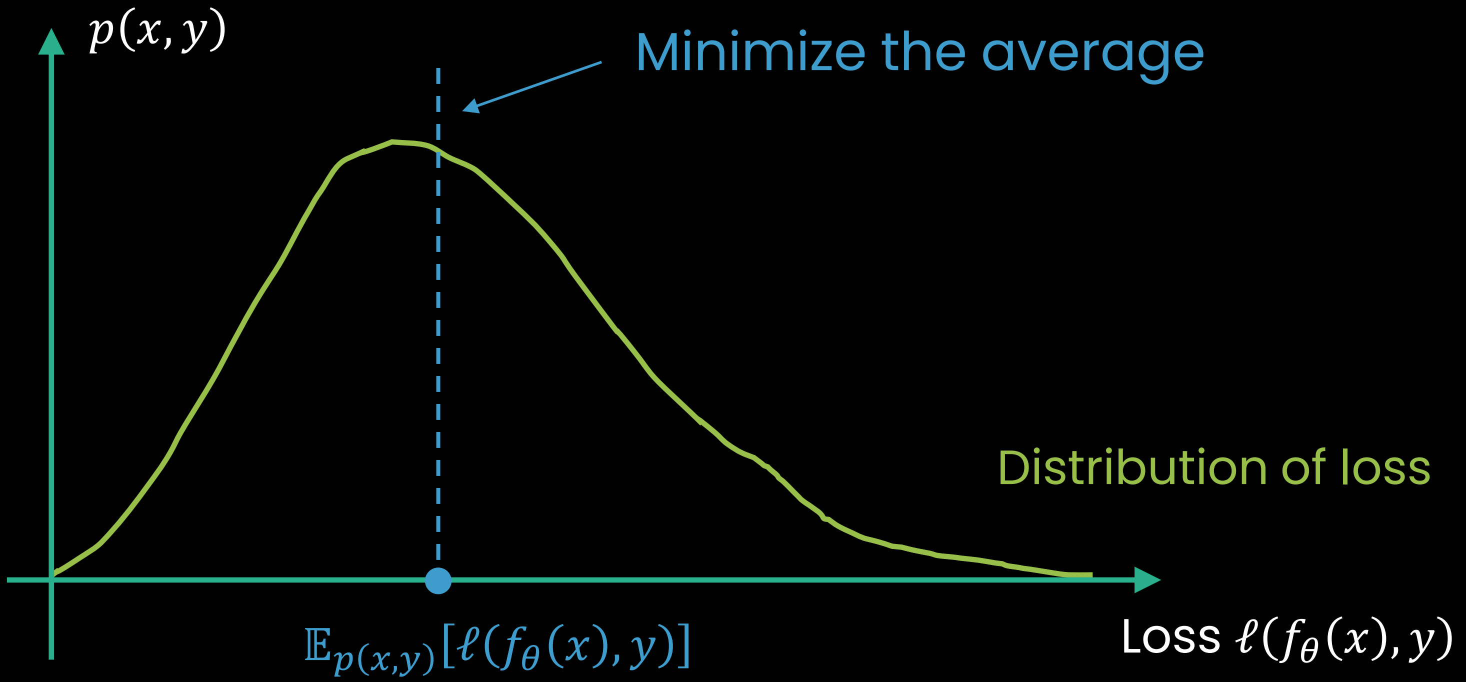
x: Input

y: Label

θ: Model parameter

f_θ: Prediction function

ℓ: loss function

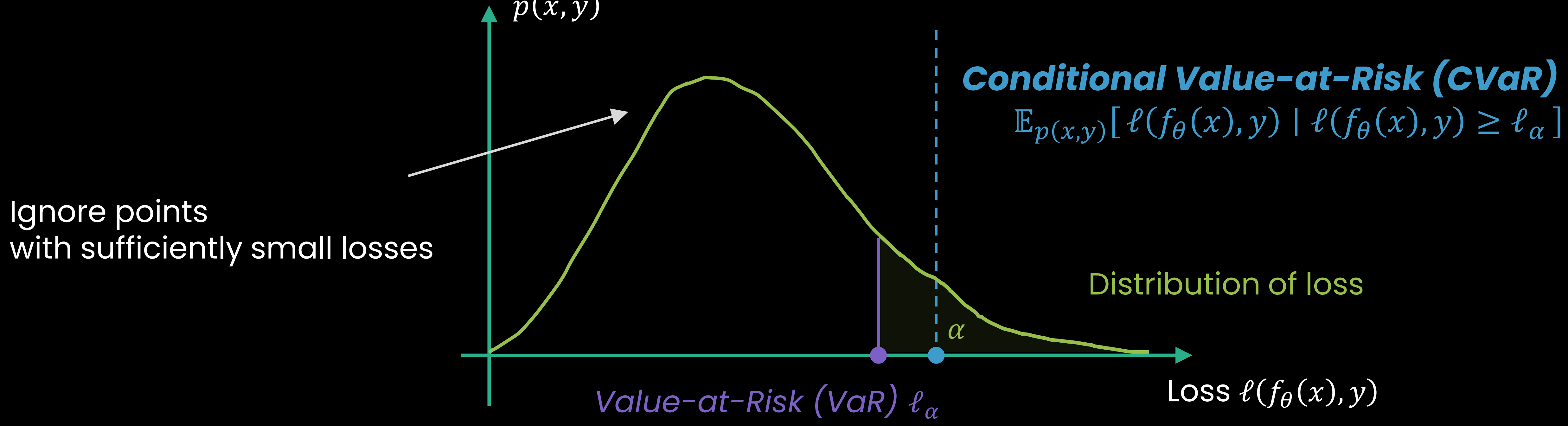


Conditional Value-at-Risk (CVaR)

- The average loss of $100\alpha\%$ worse-off samples

$$\mathbb{E}_{p(x,y)}[\,\ell(f_\theta(x),y) \mid \ell(f_\theta(x),y) \geq \ell_\alpha\,]$$

ℓ_α: $1-\alpha$-quantile (*Value-at-Risk; VaR*)



CVaR Minimization

CVaR

$$\mathbb{E}_{p(x,y)}[\, \ell(f_\theta(x), y) \mid \ell(f_\theta(x), y) \geq \ell_\alpha \,]$$

Dual of CVaR

$$\min_{\ell_\alpha} \left\{ \ell_\alpha + \frac{1}{\alpha} \mathbb{E}_{p(x,y)}[\max(0, \ell(f_\theta(x), y) - \ell_\alpha)] \right\}$$

- Minimizing its empirical approx. using i.i.d. samples $(x_1, y_1), \ldots, (x_n, y_n)$

$$\min_{\theta} \min_{\ell_\alpha} \left\{ \ell_\alpha + \frac{1}{\alpha n} \sum_{i=1}^{n} \max(0, \ell(f_\theta(x_i), y_i) - \ell_\alpha) \right\}$$

Inefficiency issue in CVaR+RecSys

Matrix factorization + CVaR dual

$$\min_{\mathbf{U},\mathbf{V}} \min_{\xi} \left\{ \xi + \frac{1}{\alpha|\mathcal{U}|} \sum_{i=1}^{|\mathcal{U}|} \max(0, \ell(\mathbf{V}\mathbf{u}_i, \mathcal{V}_i) - \xi) + \Omega(\mathbf{U}, \mathbf{V}) \right\}$$

where $\mathcal{V}_i$ is the set of i's clicked items

Quadratic loss function

$$\ell(\mathbf{V}\mathbf{u}_i, \mathcal{V}_i) = \frac{1}{|\mathcal{V}_i|} \sum_{j \in \mathcal{V}_i} \frac{1}{2} \left(\mathbf{u}_i^\top \mathbf{v}_j - 1 \right)^2 + \frac{\beta}{2} \|\mathbf{V}\mathbf{u}_i\|_2^2$$

L2 regularization (with Tikhonov weight matrices)

$$\Omega(\mathbf{U}, \mathbf{V}) = \frac{1}{2} \left\| \mathbf{\Lambda}_U^{1/2} \mathbf{U} \right\|_F^2 + \frac{1}{2} \left\| \mathbf{\Lambda}_V^{1/2} \mathbf{V} \right\|_F^2$$

Remark

Non-linear $\max(0,\cdot)$ breaks separability w.r.t. the rows of $\mathbf{V}$ and smoothness
→ the objective is **not scalable for many items** and **difficult to exploit second-order information**

Inefficiency issue in CVaR+RecSys

Matrix factorization + CVaR dual

$$\min_{\mathbf{U},\mathbf{V}} \min_{\xi} \left\{ \xi + \frac{1}{\alpha|\mathcal{U}|} \sum_{i=1}^{|\mathcal{U}|} \max(0, \ell(\mathbf{V}\mathbf{u}_i, \mathcal{V}_i) - \xi) + \Omega(\mathbf{U}, \mathbf{V}) \right\}$$

where $\mathcal{V}_i$ is the set of i's clicked items

Quadratic loss function

$$\ell(\mathbf{V}\mathbf{u}_i, \mathcal{V}_i) = \frac{1}{|\mathcal{V}_i|} \sum_{j\in\mathcal{V}_i} \frac{1}{2}\left(\mathbf{u}_i^\top \mathbf{v}_j - 1\right)^2 + \frac{\beta}{2}\|\mathbf{V}\mathbf{u}_i\|_2^2$$

Separable upper bound of ranking loss

L2 regularization (with Tikhonov weight matrices)

$$\Omega(\mathbf{U}, \mathbf{V}) = \frac{1}{2}\left\|\mathbf{\Lambda}_U^{1/2}\mathbf{U}\right\|_F^2 + \frac{1}{2}\left\|\mathbf{\Lambda}_V^{1/2}\mathbf{V}\right\|_F^2$$

Remark

Non-linear $\max(0,\cdot)$ breaks separability w.r.t. the rows of $\mathbf{V}$ and smoothness

→ the objective is **not scalable for many items** and **difficult to exploit second-order information**

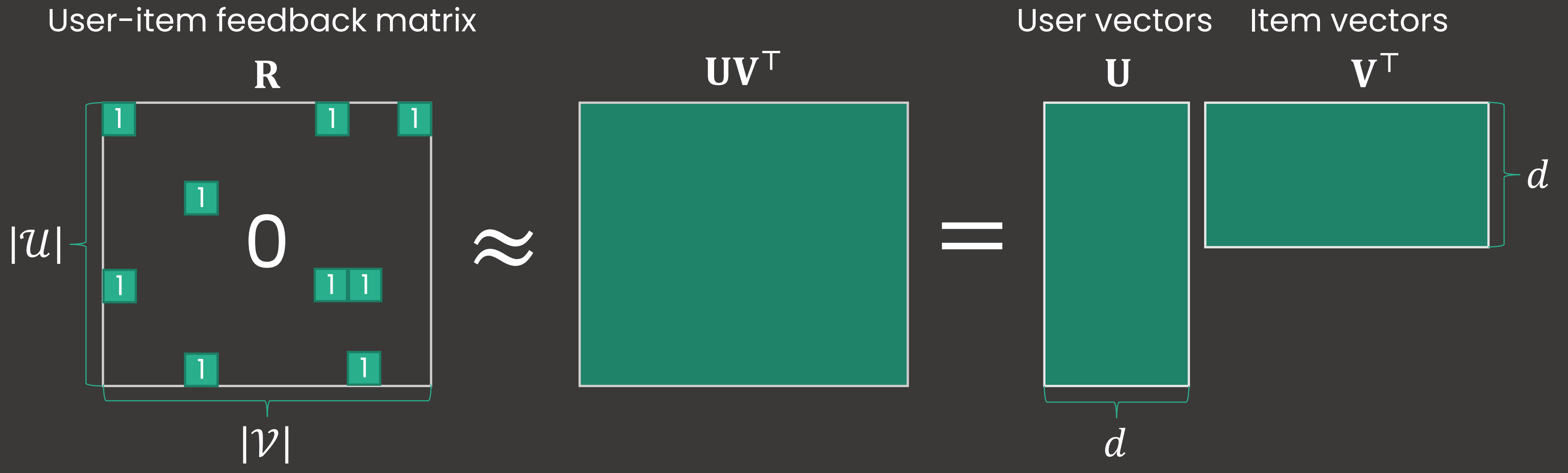
Inefficiency issue in CVaR+RecSys

Matrix factorization + CVaR dual

$$\min_{\mathbf{U},\mathbf{V}} \min_{\xi} \left\{ \xi + \frac{1}{\alpha|\mathcal{U}|} \sum_{i=1}^{|\mathcal{U}|} \max(0, \ell(\mathbf{V}\mathbf{u}_i, \mathcal{V}_i) - \xi) + \Omega(\mathbf{U}, \mathbf{V}) \right\}$$

where $\mathcal{V}_i$ is the set of i's clicked

Quadratic loss function



L2 regularization (

Remark

Non-linear $\max(0,\cdot)$

→ the objective is **no**

Inefficiency issue in CVaR+RecSys

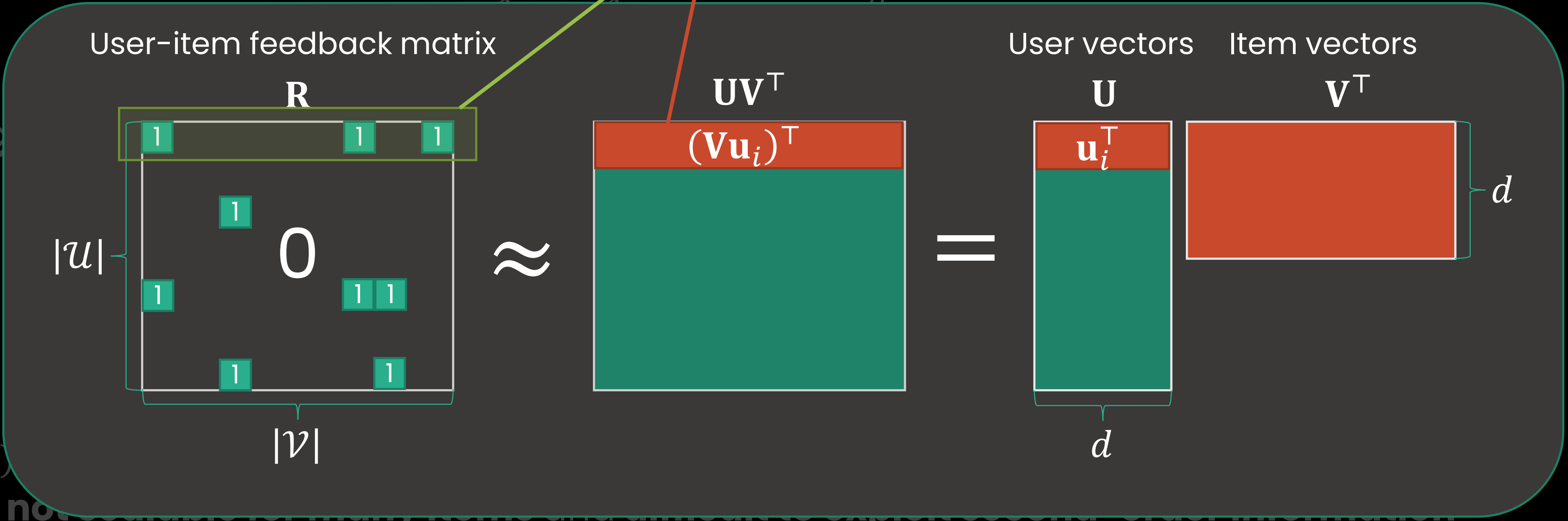
Matrix factorization + CVaR dual

$$\min_{\mathbf{U},\mathbf{V}} \min_{\xi} \left\{ \xi + \frac{1}{\alpha|\mathcal{U}|} \sum_{i=1}^{|\mathcal{U}|} \max(0, \ell(\mathbf{V}\mathbf{u}_i, \mathcal{V}_i) - \xi) + \Omega(\mathbf{U},\mathbf{V}) \right\}$$

where $\mathcal{V}_i$ is the set of i's clicked items

Quadratic loss function

L2 regularizatio



Remark

Non-linear $\max(0,\cdot)$

→ the objective is **no**

Inefficiency issue in CVaR+RecSys

Matrix factorization + CVaR dual

$$\min_{\mathbf{U},\mathbf{V}} \min_{\xi} \left\{ \xi + \frac{1}{\alpha|\mathcal{U}|} \sum_{i=1}^{|\mathcal{U}|} \max(0, \ell(\mathbf{V}\mathbf{u}_i, \mathcal{V}_i) - \xi) + \Omega(\mathbf{U}, \mathbf{V}) \right\}$$

where $\mathcal{V}_i$ is the set of i's clicked items

Quadratic loss function

L2 regularization (w

Non-separable w.r.t. items!

$$\max(0, f(\mathbf{V})) \neq \sum_{j=1}^{|\mathcal{V}|} g_j(\mathbf{v}_j)$$

Remark

Non-linear $\max(0,\cdot)$ breaks separability w.r.t. the rows of $\mathbf{V}$ and smoothness
$\rightarrow$ the objective is **not scalable for many items** and **difficult to exploit second-order information**

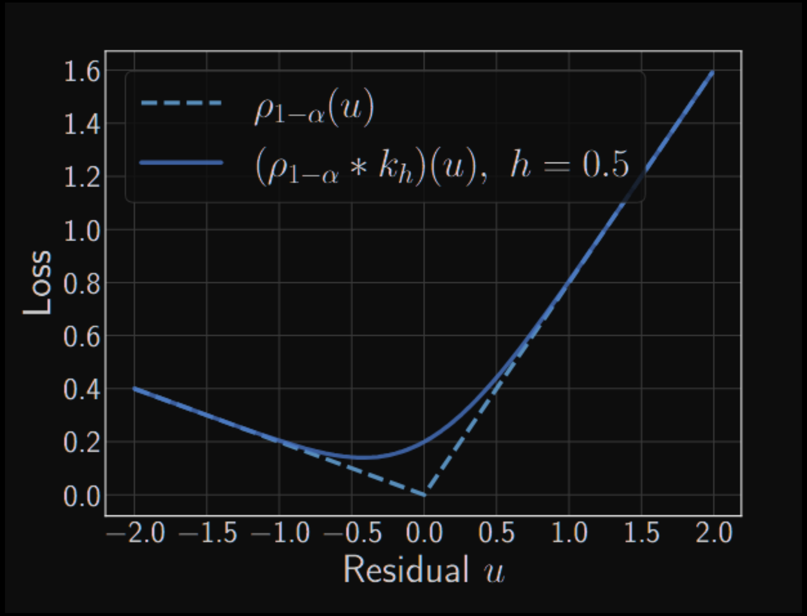
Convolution-type Smoothing

Convolution-type smoothing

Consider the convolution between $\rho_1(u) = \max(0, u)$ and some proper kernel $k_h(\cdot)$,

$$(\rho_1 * k_h)(u) = \int_{-\infty}^{\infty} \rho_1(v) k_h(v-u) dv$$
$$= \int_{0}^{\infty} v \cdot k_h(v-u) dv$$
$$= \int_{0}^{\infty} \{1 - K_h(v-u)\} dv,$$

where $K_h(u) = \int_{-\infty}^{u} k_h(v) dv$ is the kernel CDF



Remark

The first/second derivatives of $(\rho_1 * k_h)$ have tractable forms:

$$\nabla_u (\rho_1 * k_h)(u) = 1 - K_h(-u)$$
$$\nabla_u^2 (\rho_1 * k_h)(u) = k_h(-u)$$

SAFER_2

SAFER_2 (Smoothing Approach for Efficient Risk-averse Recommendation)

$$\min_{\mathbf{U},\mathbf{V},\xi}\left\{\xi + \frac{1}{\alpha|\mathcal{U}|}\sum_{i=1}^{|\mathcal{U}|}(\rho_1 * k_h)(\ell(\mathbf{V}\mathbf{u}_i,\mathcal{V}_i)-\xi) + \Omega(\mathbf{U},\mathbf{V})\right\}$$

smoothed $\max(0,\cdot)$

Efficient block-coordinate algorithm

Alternating optimization

$$\xi^{(k+1)} = \operatorname*{argmin}_{\xi}\left\{\xi + \frac{1}{\alpha|\mathcal{U}|}\sum_{i=1}^{|\mathcal{U}|}(\rho_1 * k_h)\left(\ell\left(\mathbf{V}^{(k)}\mathbf{u}_i^{(k)},\mathcal{V}_i\right)-\xi\right)\right\}$$

$$(\mathbf{U}^{(k+1)},\mathbf{V}^{(k+1)}) = \operatorname*{argmin}_{\mathbf{U},\mathbf{V}}\left\{\frac{1}{\alpha|\mathcal{U}|}\sum_{i=1}^{|\mathcal{U}|}(\rho_1 * k_h)\left(\ell(\mathbf{V}\mathbf{u}_i,\mathcal{V}_i)-\xi^{(k+1)}\right) + \Omega(\mathbf{U},\mathbf{V})\right\}$$

$$= \operatorname*{argmin}_{\mathbf{U},\mathbf{V}}\max_{\mathbf{z}}\left\{\frac{1}{\alpha|\mathcal{U}|}\sum_{i=1}^{|\mathcal{U}|}\left[z_i\cdot\left(\ell(\mathbf{V}\mathbf{u}_i,\mathcal{V}_i)-\xi^{(k+1)}\right)-(\rho_1 * k_h)^*(z_i)\right] + \Omega(\mathbf{U},\mathbf{V})\right\}$$

Separable reformulation

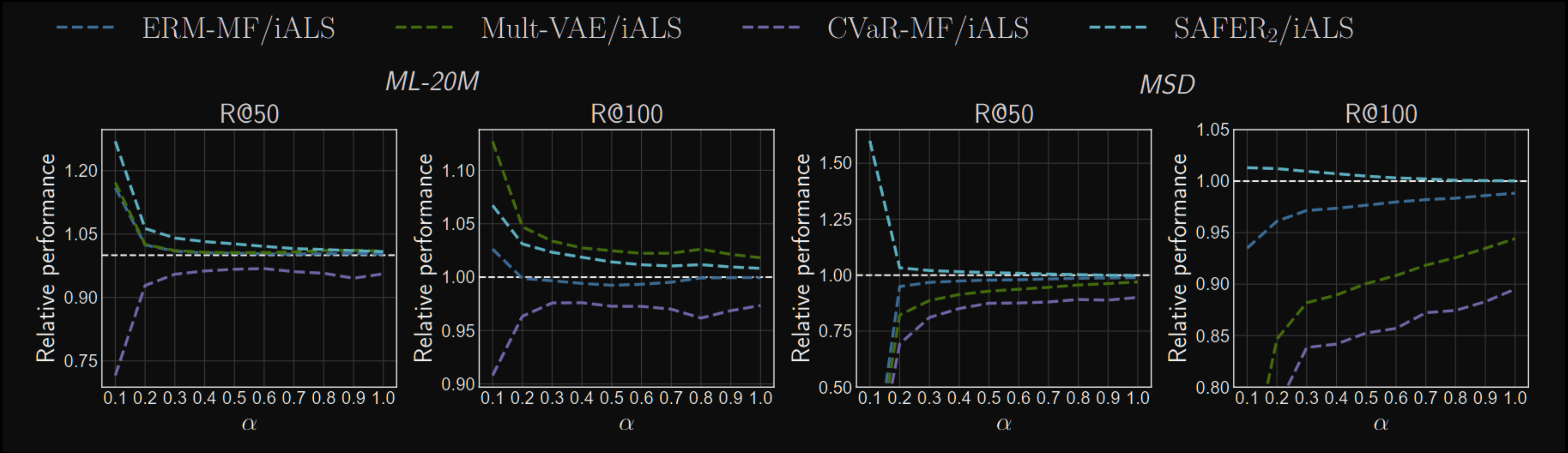
Numerical Results

Safety

- SAFER_2 shows stable performance for the tail users (small α)

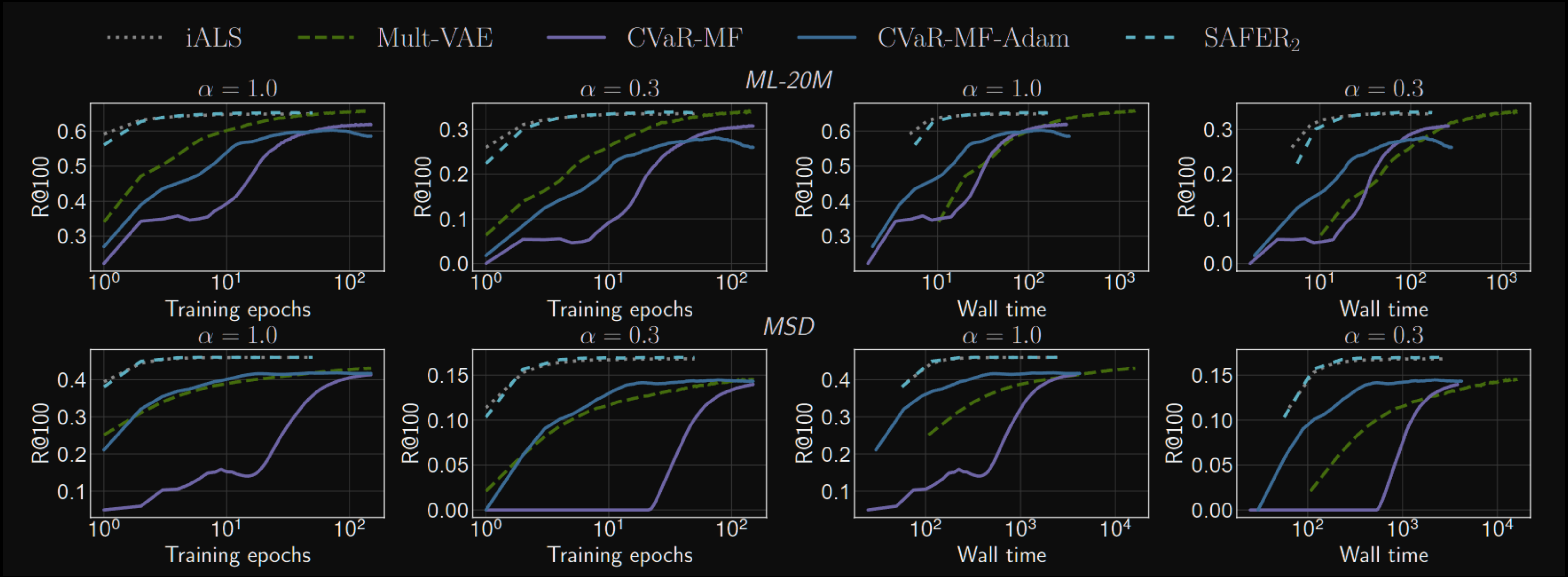
Quality

- SAFER_2 preserves competitive average performance ($\alpha = 1.0$)



Convergence Speed

SAFER_2 (- - -) achieves competitive training speed compared to the fastest method (iALS).



Summary

- We proposed **safety-aware recommendation** via CVaR minimization beyond ERM

- We develop **a safe and scalable method, SAFER_2**, which
  - **overcomes the non-parallelizable property** of CVaR formulation
  - enables an ALS-type optimization with **fast training convergence**

- Further technical details can be found in the paper
  - Discussions on CVaR + convolution-type smoothing
  - Customized Tikhonov regularization for SAFER_2
  - Various extensions of SAFER_2
    - Stochastic quantile/VaR estimation based on sub-sampled users
    - Subspace-based BCD for large embedding sizes