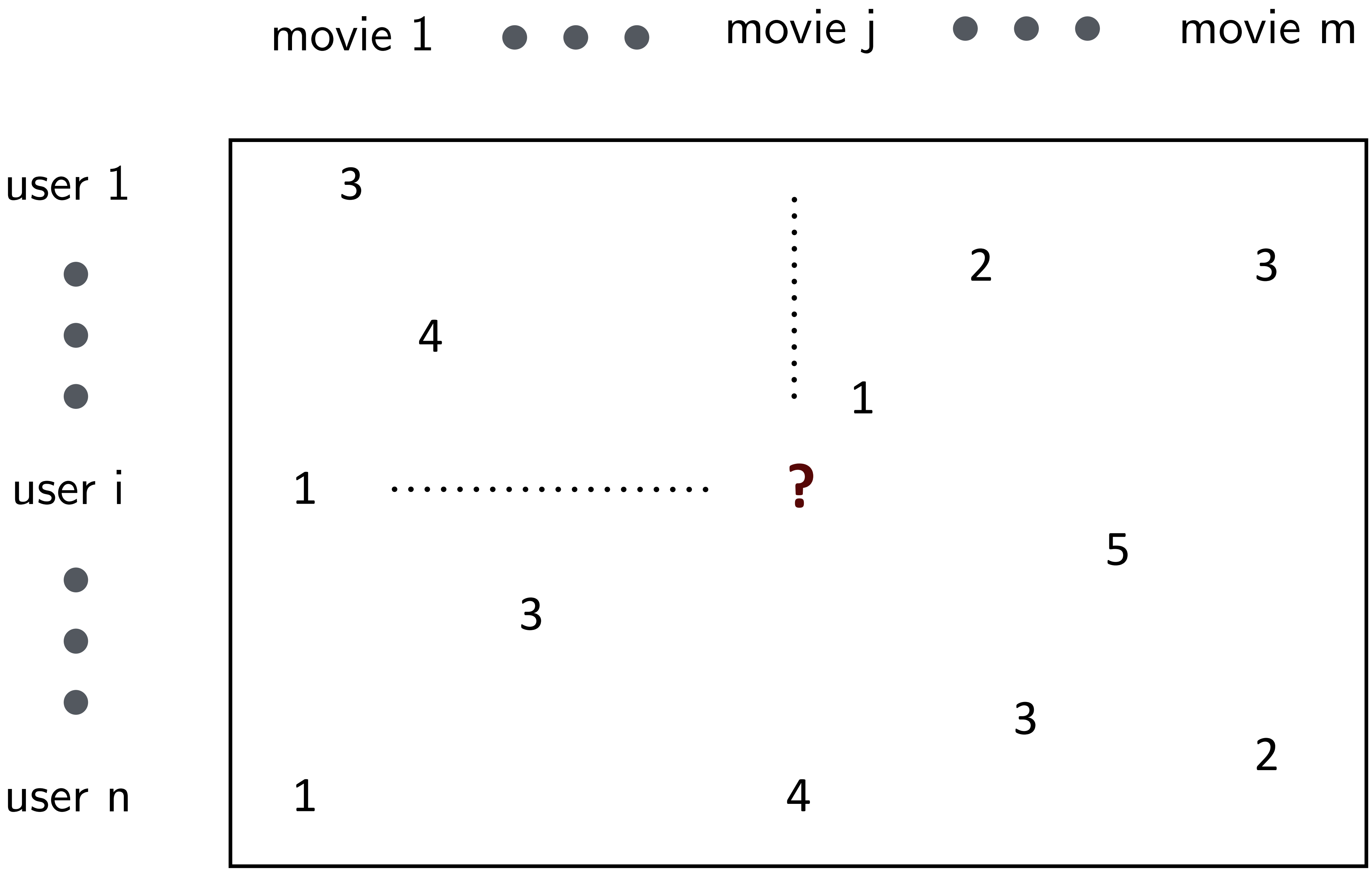
Iterative Collaborative Filtering for Sparse Matrix Estimation

Christina Lee Yu
Cornell University
cleeyu@cornell.edu

Joint work with Christian Borgs (MSR), Jennifer Chayes (MSR), Devavrat Shah (MIT), Dogyoon Song (MIT)

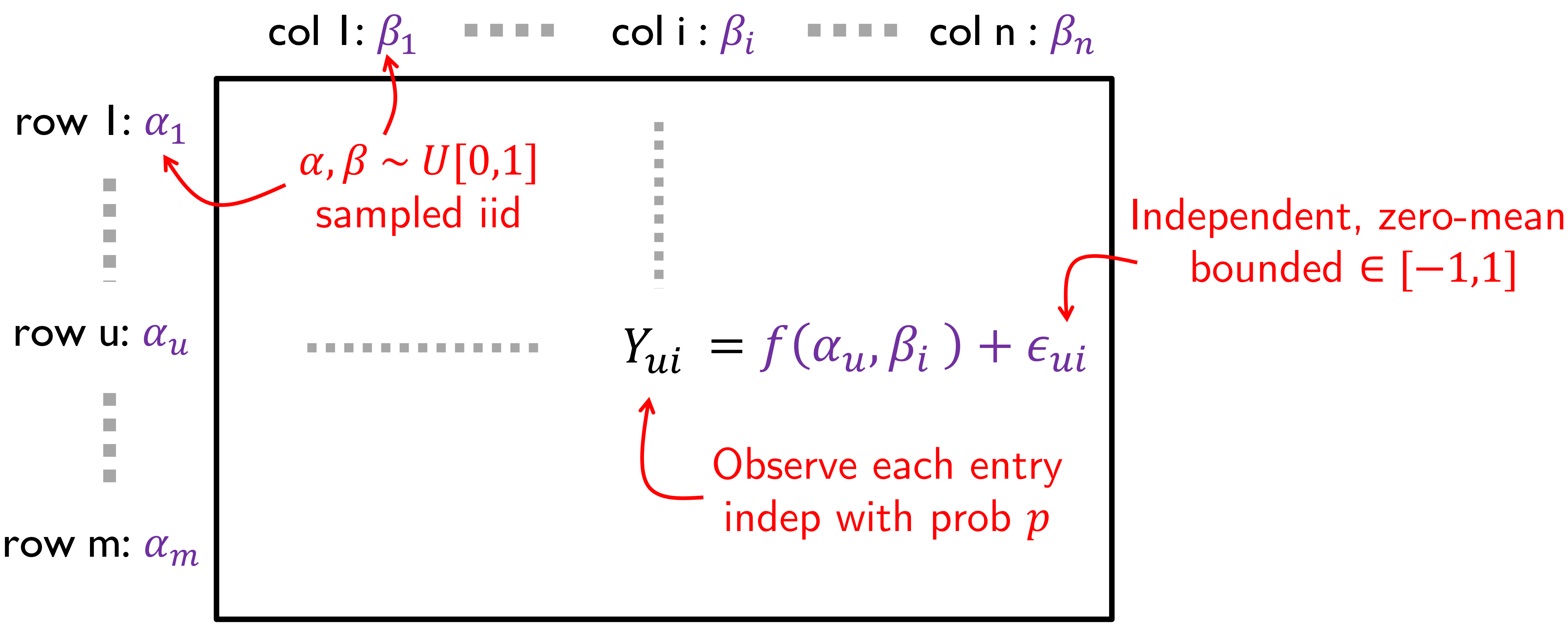
Example of Matrix Estimation: Recommendation Systems



Rating Matrix

Non-parametric latent variable model

- Row/column latent variables α, β sampled iid
- Each entry Y_{ui} observed independently with probability p
- Lipschitz latent function f such that $\mathbb{E}[Y_{ui}] = f(\alpha_u, \beta_i)$



- Low rank corresponds to function f being the inner product
- Model motivated as canonical representation of exchangeable model

Matrix Estimation

- Observations
  - Noisy observations for subset of entries: $\{Y_{ui}\}_{(u,i)\in\Omega}$
  - Subject to some `noise' model: $\mathbb{E}[Y_{ui}] = f(\alpha_u, \beta_i)$
- Goal
  - Produce an estimate $\hat{F} \in \mathbb{R}^{n\times m}$
  - So that prediction error is *small*

$$\mathrm{MSE} = \frac{1}{nm}\sum_{ui}(\hat{F}_{ui} - f(\alpha_u, \beta_i))^2$$

Algorithm for Matrix Estimation: Overview

- Matrix Factorization a la Singular Value Thresholding [Keshavan Montenari Oh, Chatterjee, ...]

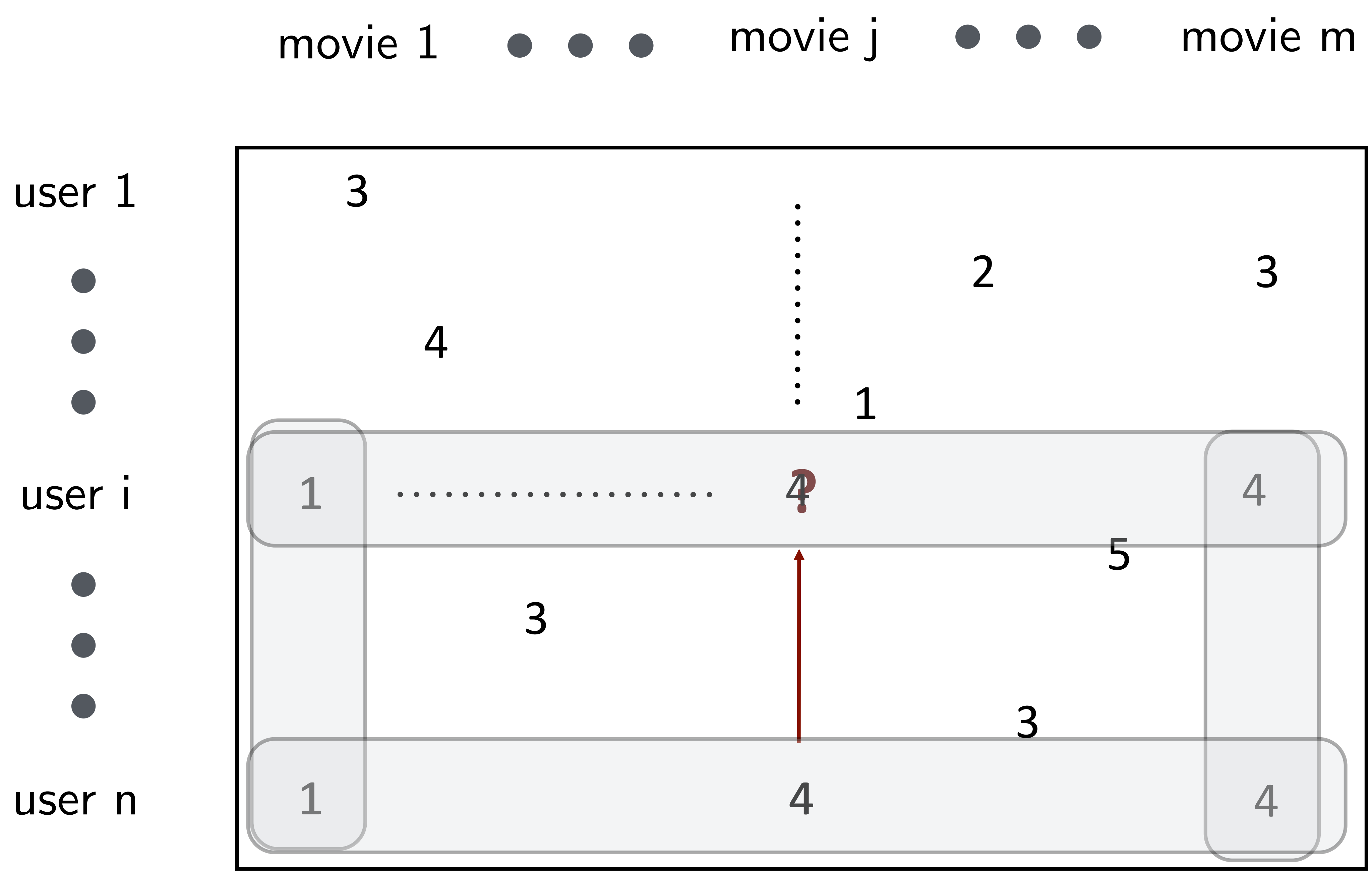
- Optimization or Risk Minimization
  - Convex relaxation via Nuclear Norm Minimization
    - [Candes-Tao, Candes-Retch, Candes-Plan, Negahban-Wainwright, Mazumdar et al, ...]
  - Tacking non-convex objective directly
    - Good Initialization, e.g. via Matrix Factorization
    - Further improvement via local minimization, e.g. via
      - (Projected) Gradient Descent [Keshavan et al][Chen et al], ...
      - Alternative Least Squares [Jain et al][Hardt, Hardt et al], ...
    - All local minima = global minima [Sun-Luo, Ge et al], ...

- Nearest Neighbors
    - Relation to Collaborative Filtering and Non-parametric representation [Goldberg et al 92], [Lee et al, Borgs et al, Zhang et al], ...

Sample Complexity Comparison

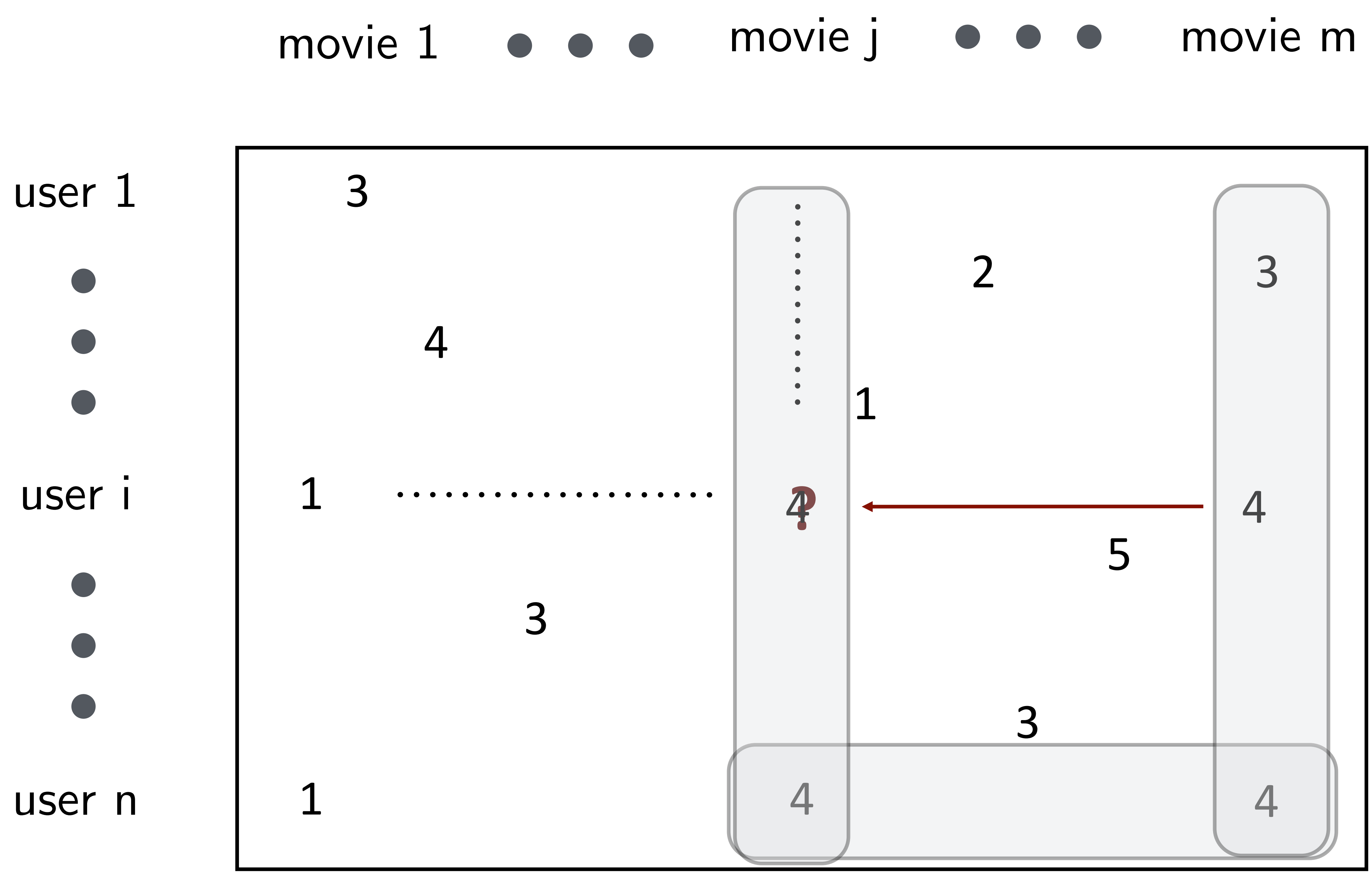
Algorithm	References	Function Class	Noise Model	Guaranteed Recovery	Observations mnp (m=n)
SVT	[Chatterjee]	Lipschitz	Arbitrary	Approx.	$n^{\frac{2r+2}{r+2}} \log^6 n$
SVT	[Chatterjee]	Low-rank	Arbitrary	Approx.	$nr \log^6 n$
Convex	[Recht]	Low-rank	No Noise	Exact	$nr \log^2 n$
Convex	[CandesPlan]	Low-rank	Additive	Approx.	$nr \log^2 n$
Non-Convex	[KeMonOh]	Low-rank	No Noise	Exact	$nr \log n$
Non-Convex	[KeMonOh]	Low-rank	Additive	Approx.	$nr \log n$
Near Nghbr		?			?

Nearest Neighbor and Collaborative Filtering (user-user)



Rating Matrix

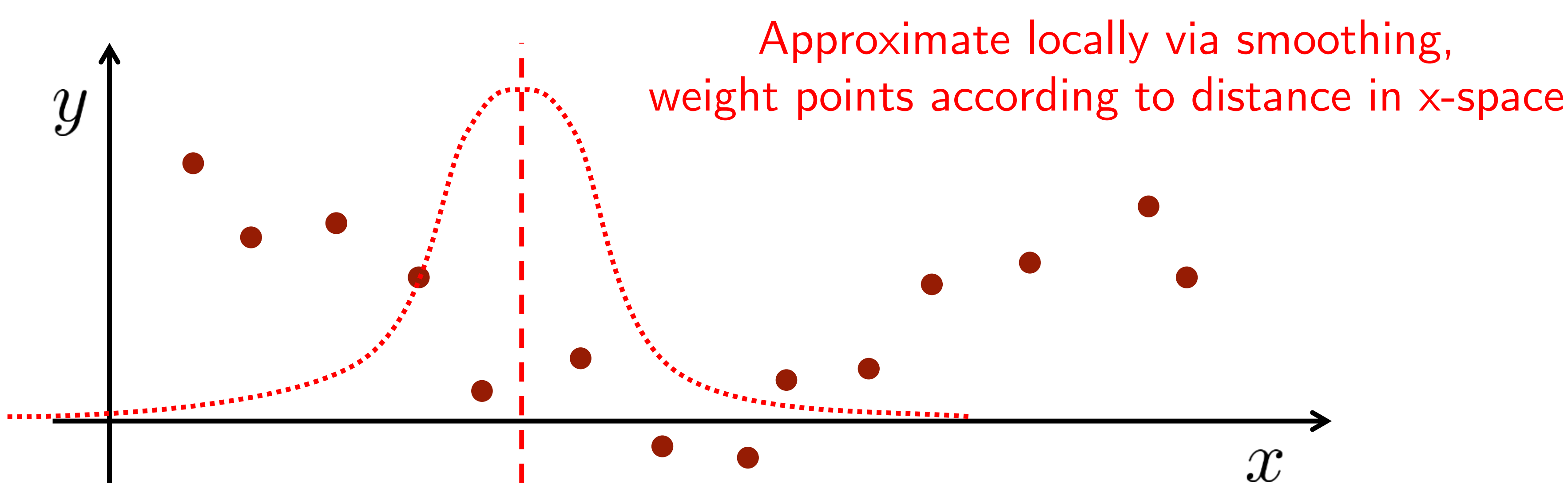
Nearest Neighbor and Collaborative Filtering (item-item)



Rating Matrix

Why study Collaborative Filtering algorithm?

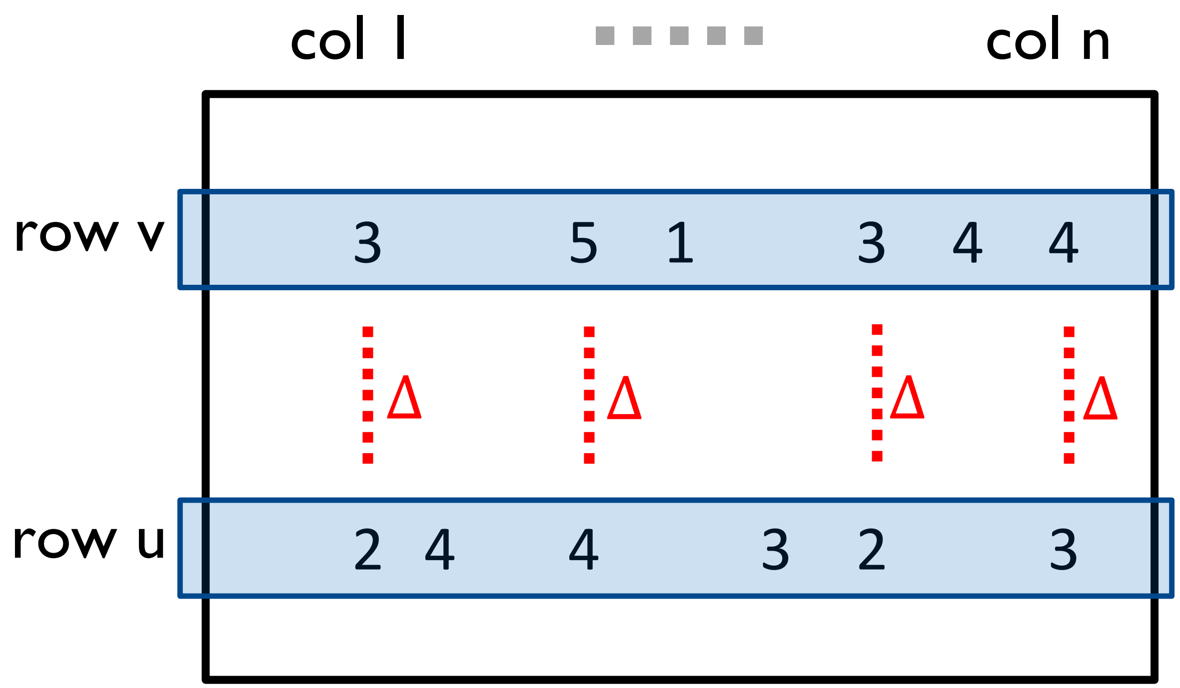
- Extensively used in practice
- Scalable implementation via approximate nearest neighbor
- Incremental (and hence robust)
- Interpretable:
  - watch *Captain America* because you liked *Iron Man* and those who liked *Iron Man* also liked *Captain America*
- Conceptual relationship to Kernel regression / nearest neighbor / nonparametric approach



- Don't know distance in covariate space, thus approximate from data

Simplified Collaborative Filtering Algorithm

- Approx distance with mean squared difference of common observations

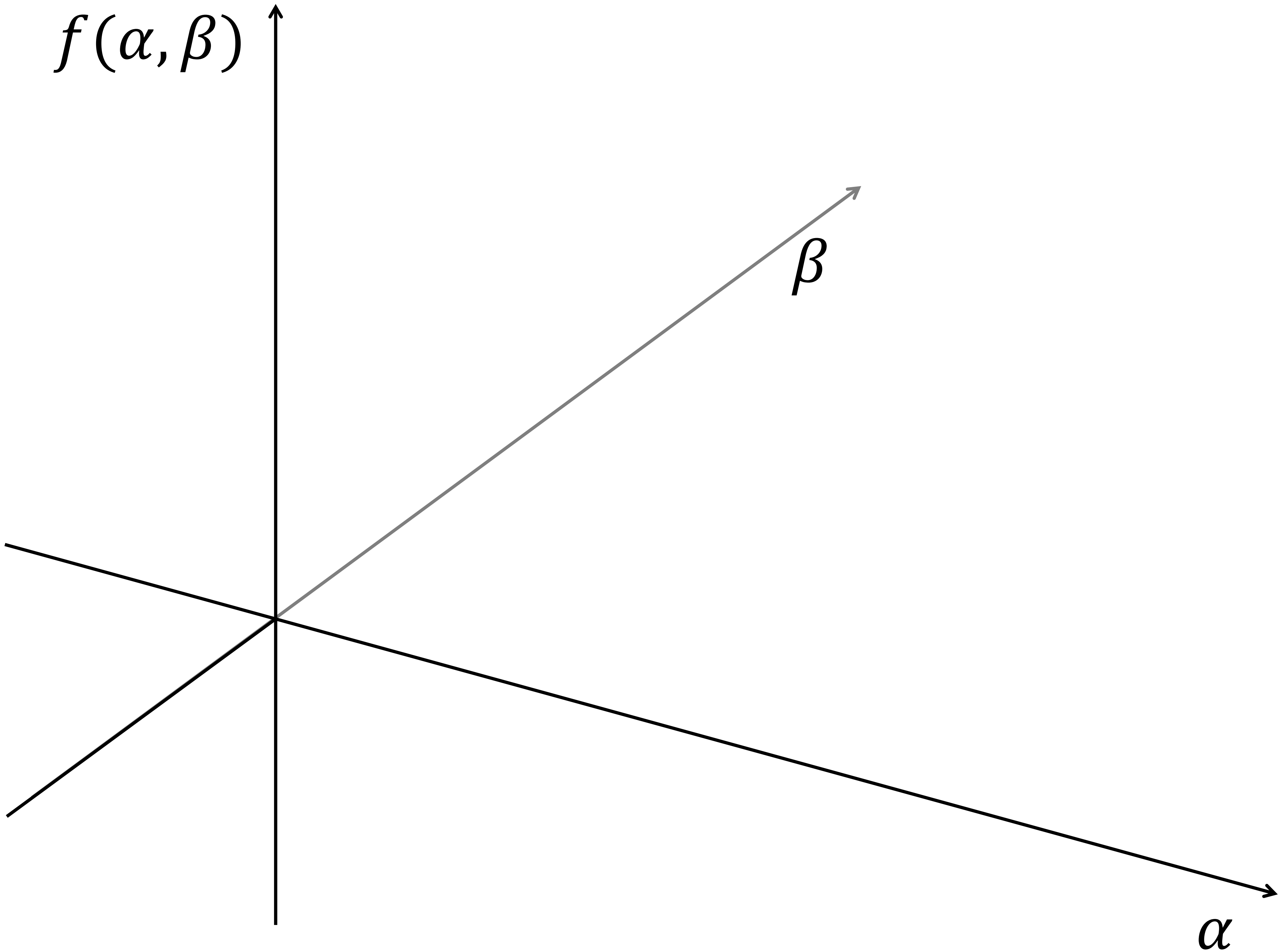


$$d_{uv} = \frac{1}{|N_u \cap N_v| - 1} \sum_{i \in N_u \cap N_v} (Y_{ui} - Y_{vi})^2 - 2\sigma^2$$

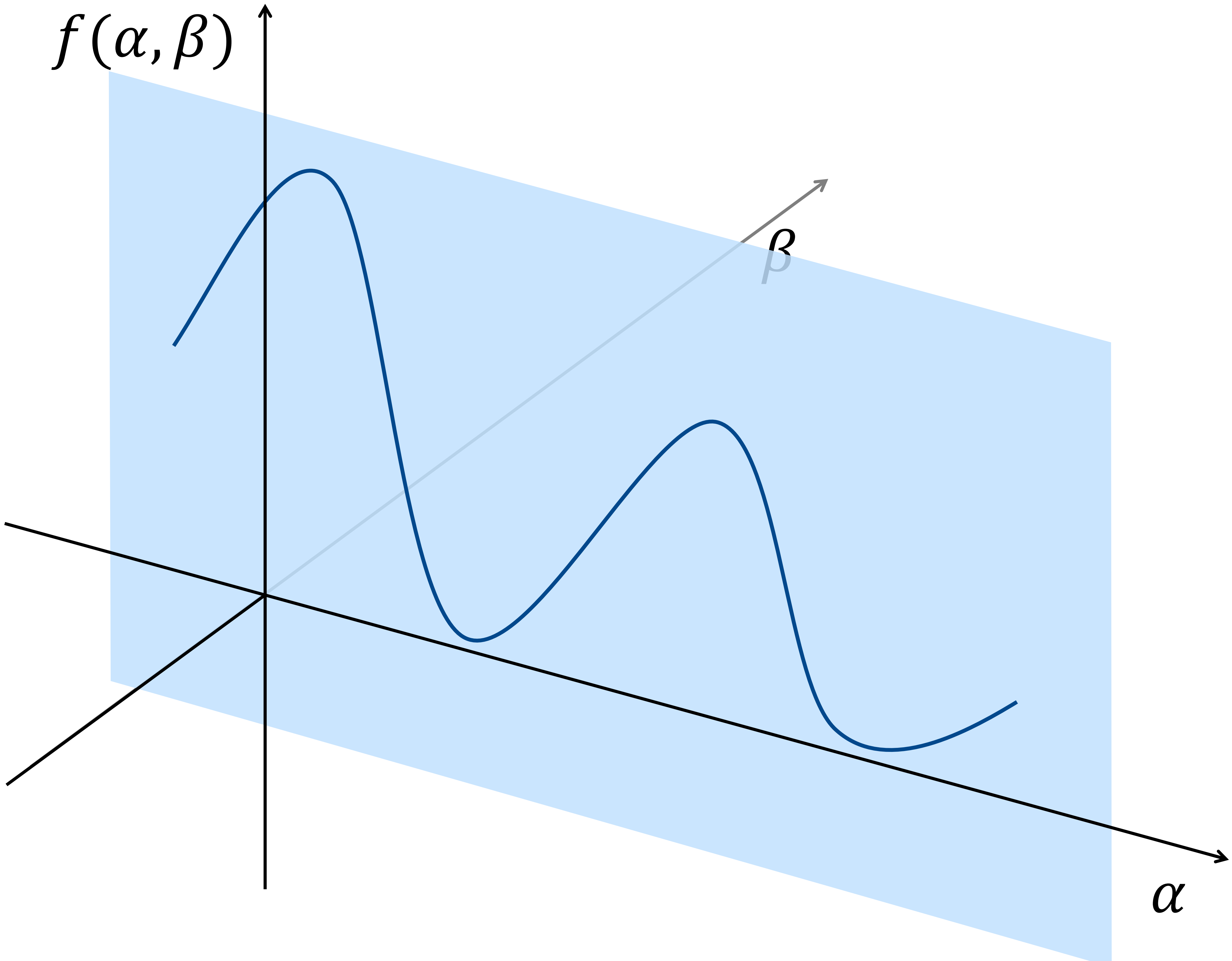
- Final estimate averages datapoints from "nearby" rows

$$\hat{F}_{ui} = \frac{1}{|\{v : d_{uv} \leq \eta\}|} \sum_{v : d_{uv} \leq \eta} Y_{vi}$$

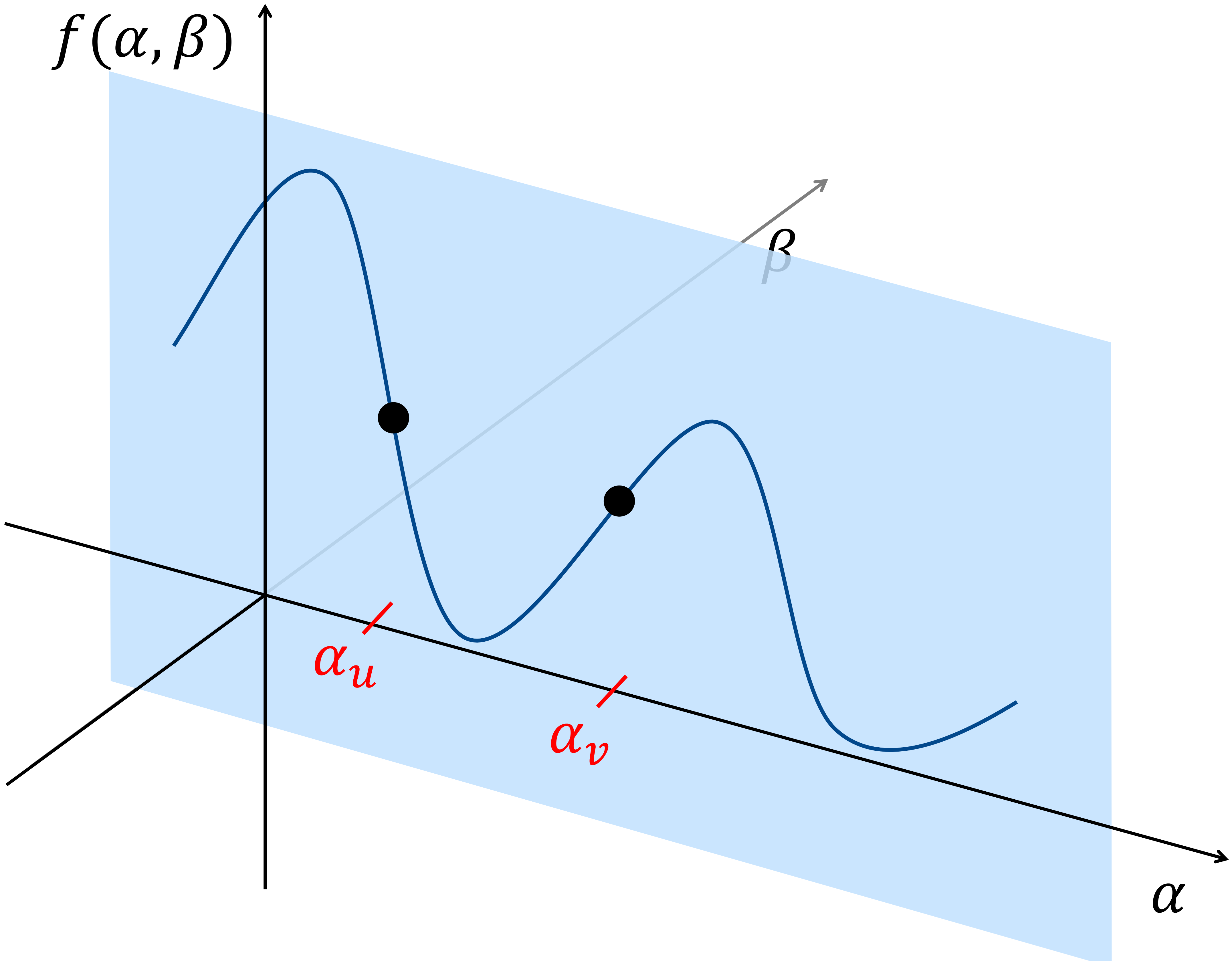
- Similar concept to kernel regression, "average over nearby datapoints"



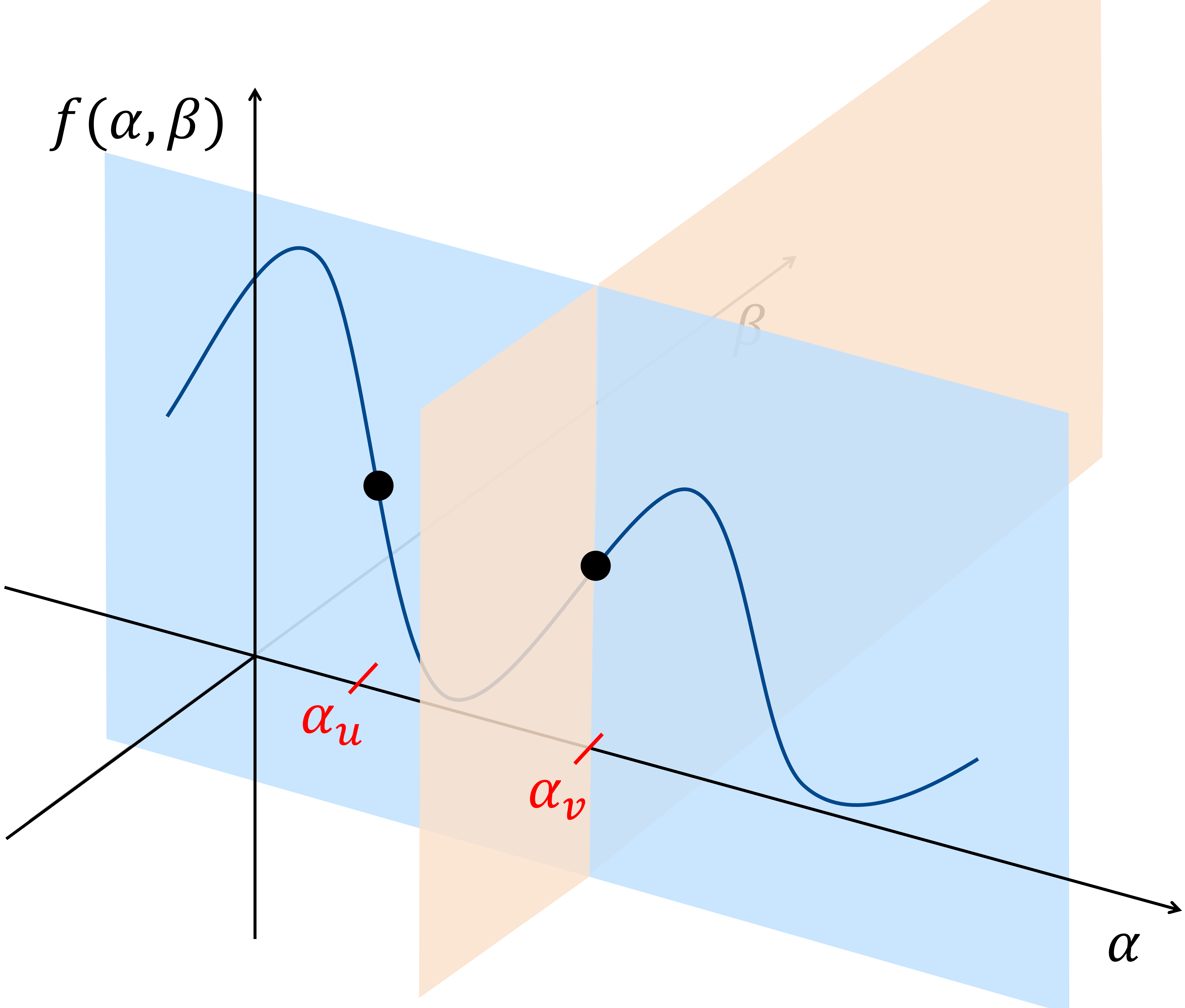
$f(\alpha,\beta)$
β
α



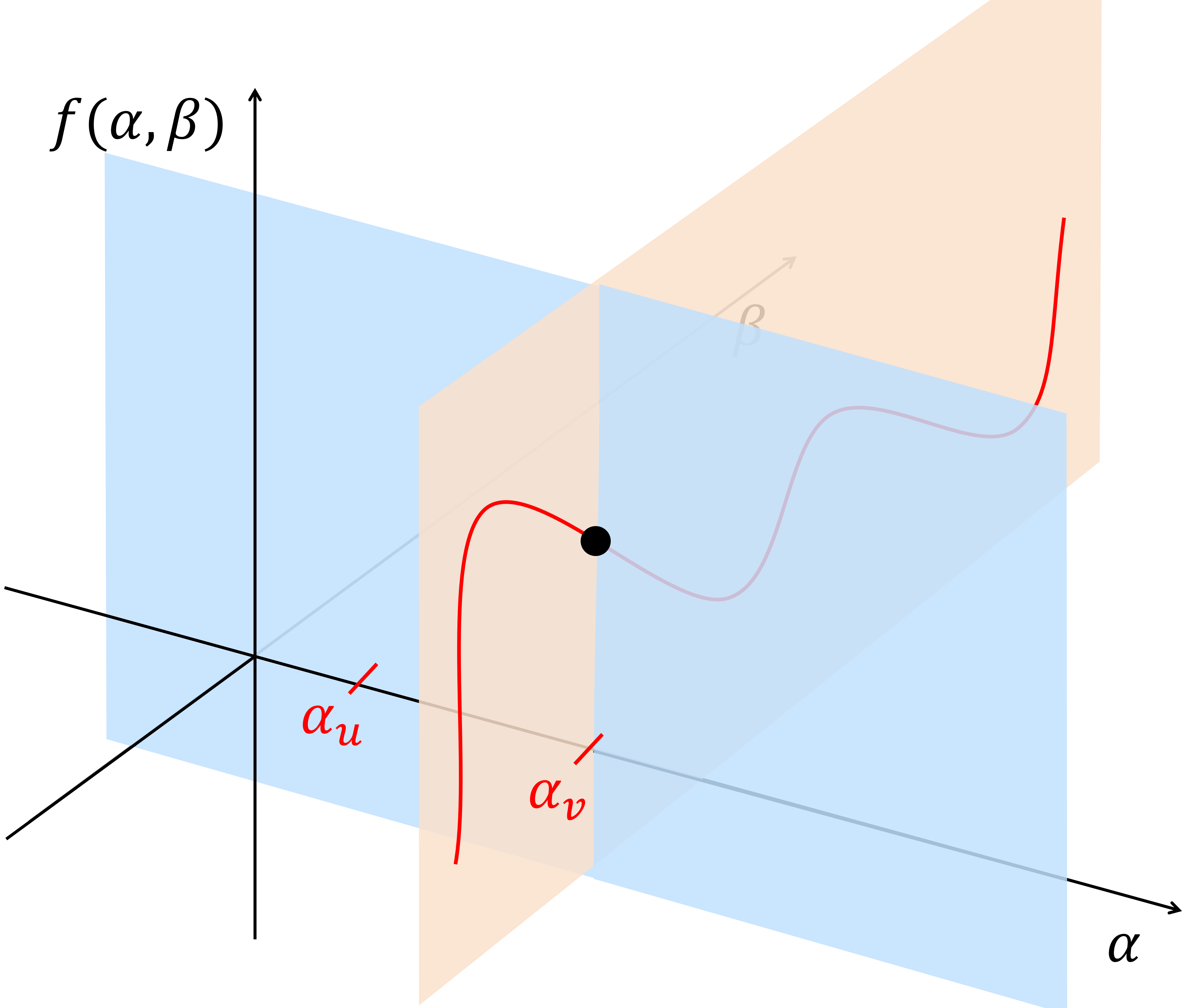
$f(\alpha, \beta)$
β
α



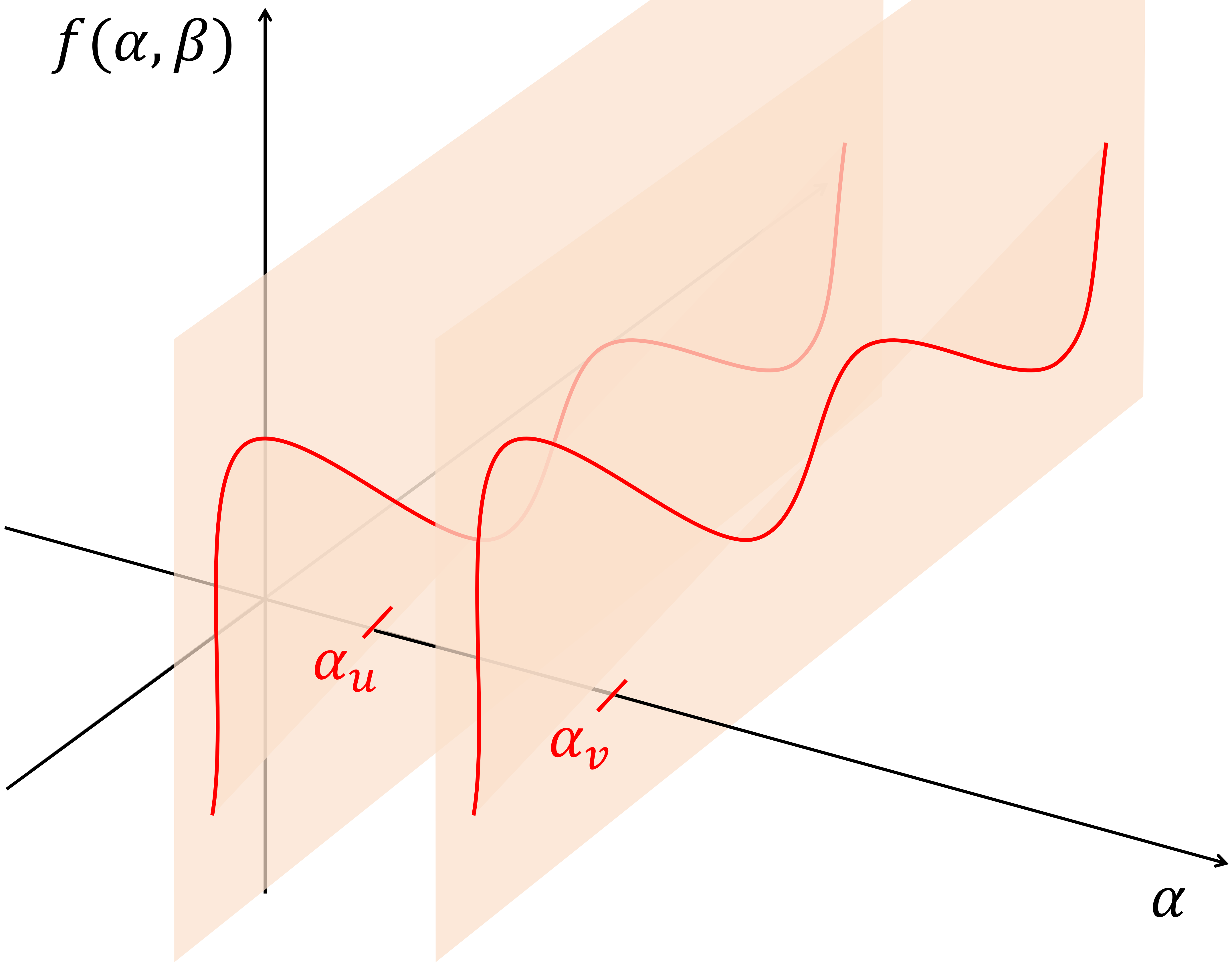
$f(\alpha, \beta)$
β
α_u
α_v
α

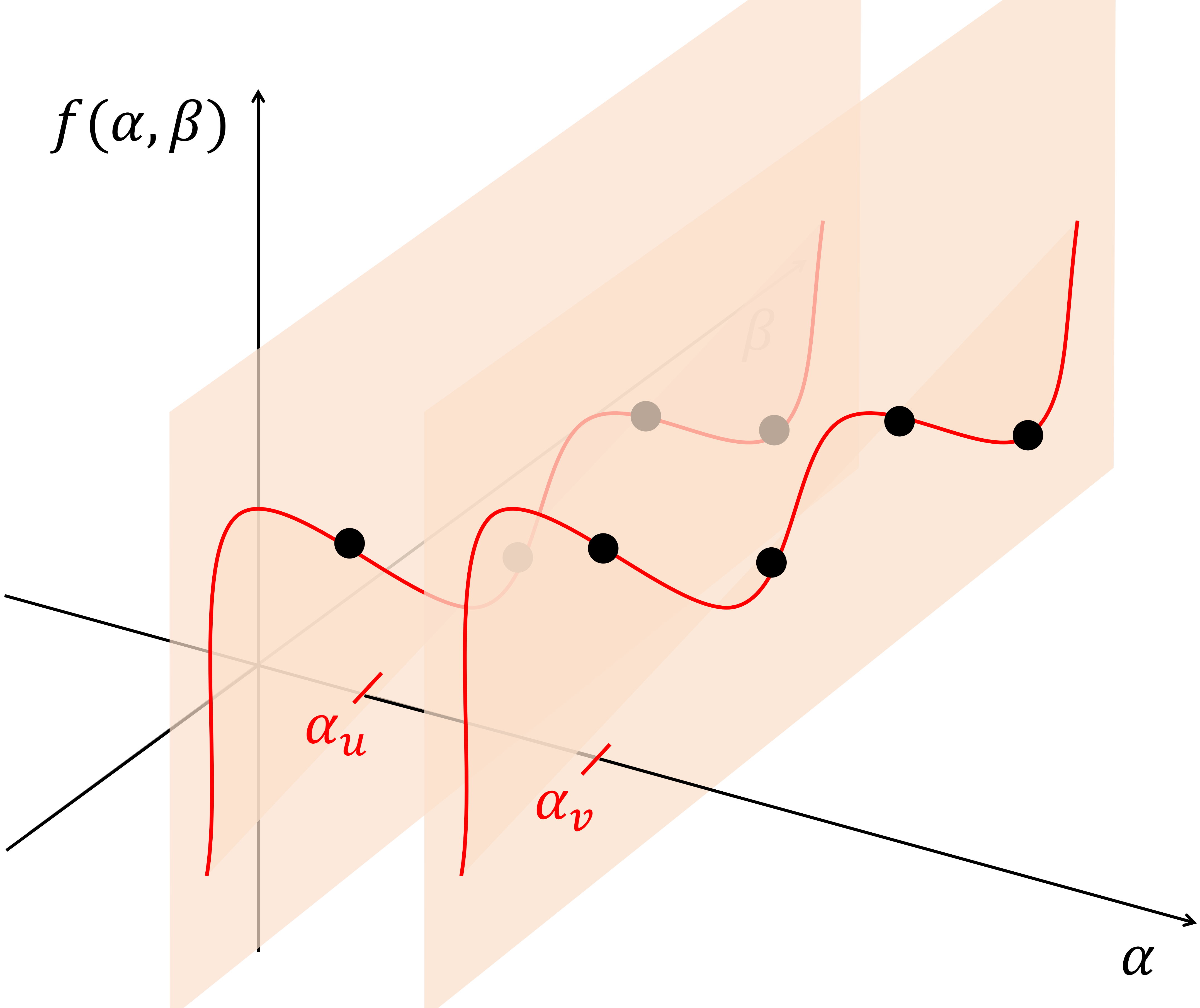


$f(\alpha, \beta)$
β
α_u
α_v
α



$f(\alpha, \beta)$
β
α_u
α_v
α


$f(\alpha, \beta)$
α_u
α_v
α



Instead of learning true latent distance, use an estimated proxy which reflects L_2 distance between associated latent functions

Theorem [Lee-Li-Shah-Song]

Assuming that

- Latent function f is L-Lipschitz and bounded in [0,1]
- Latent features α, β sampled iid $\sim U[0,1]$
- Each entry observed independently w/prob $p = \omega(\max(m^{-1}, n^{-\frac{1}{2}}))$
- Additive Gaussian noise $\sim N(0, \sigma^2)$

the algorithm achieves

$$MSE = O\left(L^{10/7}\sigma^{8/7}(mp)^{-\frac{4}{7}}, L^2\sigma^{4/3}(np^2)^{-\frac{1}{3}+\delta}\right).$$

Proof Sketch

- Analysis involves first showing distance proxy is "good" because empirical squared difference converges to L_2 functional distance as long as sufficiently many common observations between rows

$$\mathbb{P}\left(\cap_{v\in[m]}\{|N(u)\cap N(v)| \geq c_0 np^2\}\right) \geq 1 - m\exp\left(-\frac{(1-c_0)^2 np^2}{2}\right)$$

- Concentration of squared difference

$$\begin{aligned} d_{uv} &= \frac{1}{|N_u \cap N_v|}\sum_{i\in N_u\cap N_v}(f(\alpha_u,\beta_i)+\epsilon_{ui}-f(\alpha_v,\beta_i)+\epsilon_{vi})^2 - 2\sigma^2 \\ &\approx \frac{1}{|N_u \cap N_v|}\sum_{i\in N_u\cap N_v}(f(\alpha_u,\beta_i)-f(\alpha_v,\beta_i))^2 \\ &\approx \int (f(\alpha_u,\beta)-f(\alpha_v,\beta))^2 d\beta := \|f_u - f_v\|_2^2 \end{aligned}$$

$$\mathbb{P}\left(|d_{uv} - \|f_u - f_v\|_2^2| > \delta\right) \leq 4\exp\left(-\frac{c_1 np^2\delta^2}{\sigma^4}\right)$$

- If f is L-Lipschitz, then

$$\|f_u - f_v\|_\infty := \sup_\beta |f(\alpha_u,\beta) - f(\alpha_v,\beta)| \leq (6L\|f_u - f_v\|_2^2)^{1/3}$$

Proof Sketch

- Recall final estimate is $\hat{F}_{ui} = \frac{1}{|\{v : d_{uv} \le \eta\}|} \sum_{v : d_{uv} \le \eta} Y_{vi}$
- Use standard bias variance tradeoff calculations as in kernel regression
- Sufficiently many nearby neighbors

$$\mathbb{P}(\|f_u - f_v\|_2^2 \le \eta) \ge \frac{\sqrt{\eta}}{L}$$

$$\mathbb{P}\left(|\{v : d_{uv} \le \eta, (v,i) \in \Omega\}| \le (1 - c_2)\frac{mp\sqrt{\eta - \delta}}{L}\right) \le \exp\left(-\frac{c_2^2 mp\sqrt{\eta-\delta}}{L}\right)$$

- Choose η to tradeoff between bias and variance of $\hat{F}_{ui}$

$$MSE \le \frac{\sigma^2 L}{(1-\delta)mp\sqrt{\eta-\delta}} + \left((6L(\eta+\delta))^{1/3} + 4\exp\left(-\frac{c_1 np^2\delta^2}{2\sigma^4}\right)\right)^2$$

- Therefore, algorithm is provably convergent for Lipschitz functions
- .... But very expensive sample complexity!!

Sample Complexity Bottleneck

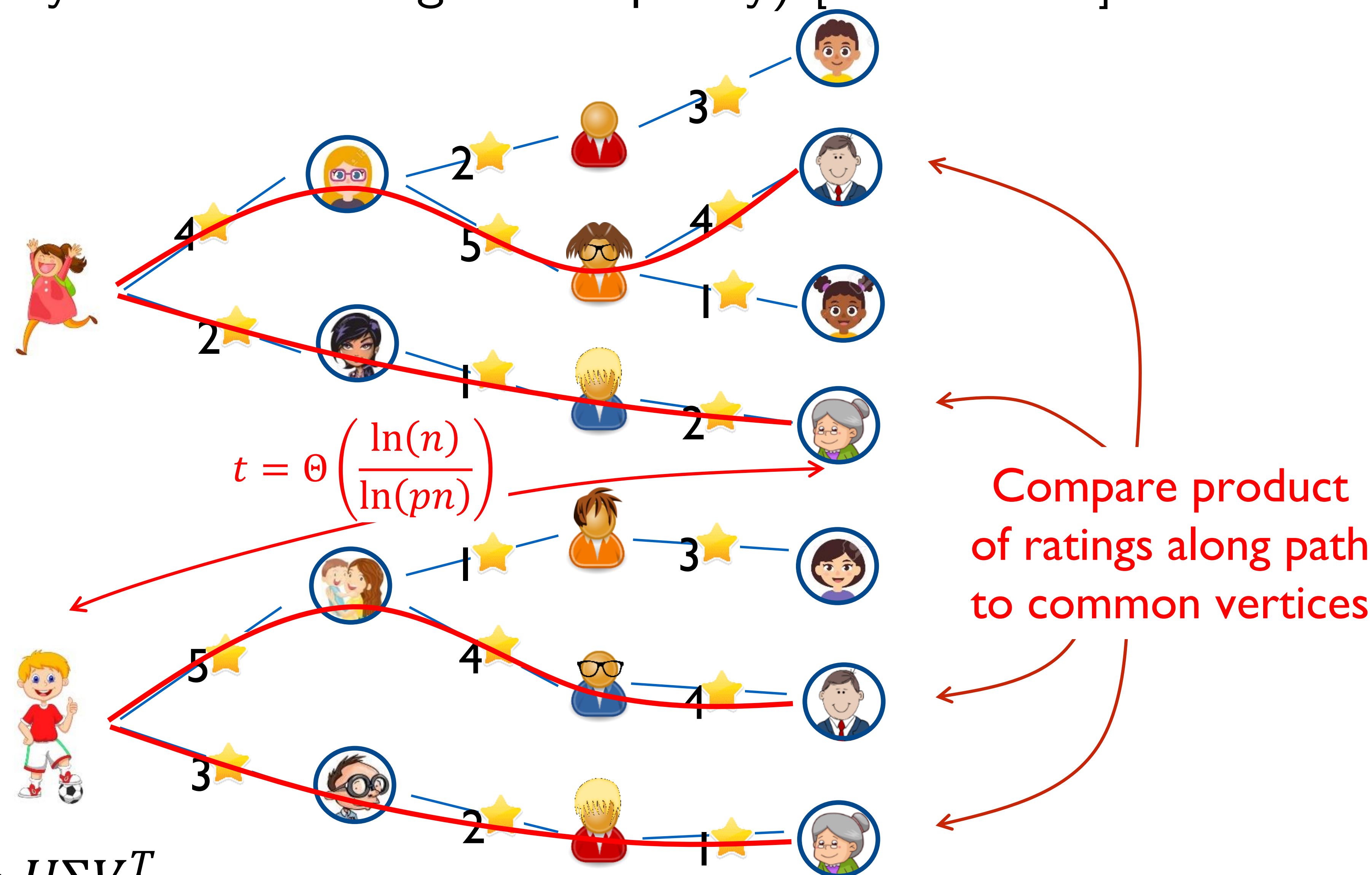
Computing similarity requires common observations

Birthday Paradox requires sample complexity $\widetilde{\Omega}(n^{3/2})$

What if observations are too sparse? $p = \omega(n^{-\frac{1}{2}})$ is quite expensive

Expanding Neighborhood of Data

- Idea: use higher order data, consider graph representation of data (consider symmetric setting for simplicity) [Abbe-Sandon] for SBM



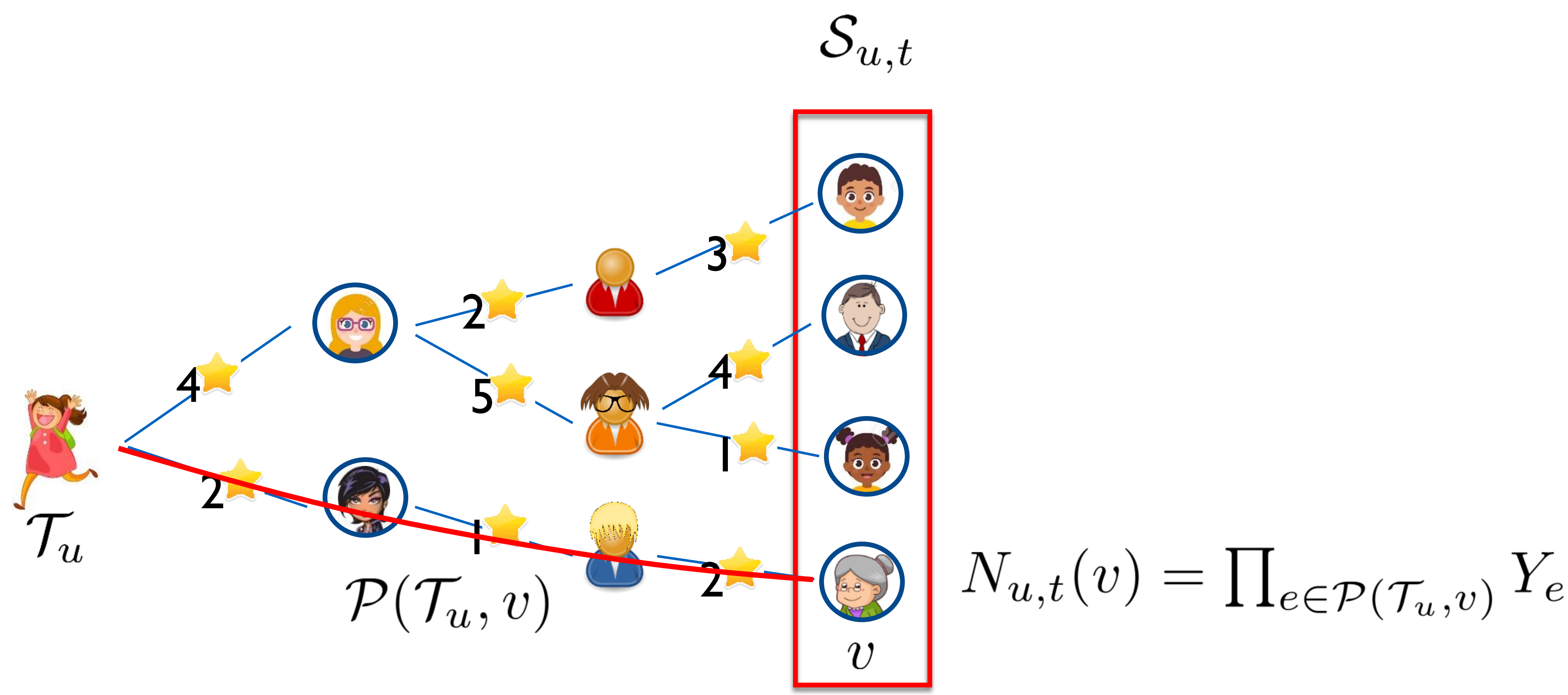
$$\mathbb{E}[Y] = U\Sigma V^T$$

$$Y \quad \bullet \quad Y^T \quad \bullet \quad Y \quad \approx U\Sigma^3 V^T$$

- Compare neighbors $\|(U_{} - U_{})\Sigma\|_2^2$ vs. t-boundaries $\|(U_{} - U_{})\Sigma^t\|_2^2$
- Need additional steps when t grows with n

Algorithm

- Define the following quantities
  - $\mathcal{T}_u$ is a breadth first tree rooted at u
  - $\mathcal{P}(\mathcal{T}_u, v)$ is the path from u to v in $\mathcal{T}_u$
  - $\mathcal{S}_{u,t} = \{v \in [n] \ s.t. \ |\mathcal{P}(\mathcal{T}_u, v)| = t\}$
  - $N_{u,t}(v) = \mathbb{I}_{\{|\mathcal{P}(\mathcal{T}_u,v)|=t\}} \prod_{e \in \mathcal{P}(\mathcal{T}_u,v)} Y_e$



Algorithm

- Define the following quantities
  - $\mathcal{T}_u$ is a breadth first tree rooted at u
  - $\mathcal{P}(\mathcal{T}_u, v)$ is the path from u to v in $\mathcal{T}_u$
  - $\mathcal{S}_{u,t} = \{v \in [n] \ s.t. \ |\mathcal{P}(\mathcal{T}_u, v)| = t\}$
  - $N_{u,t}(v) = \mathbb{I}_{\{|\mathcal{P}(\mathcal{T}_u,v)|=t\}} \prod_{e\in\mathcal{P}(\mathcal{T}_u,v)} Y_e$

- Compute distance according to

$$d_{uv} = \frac{1}{p}\left(\frac{N_{u,t}}{|\mathcal{S}_{u,t}|} - \frac{N_{v,t}}{|\mathcal{S}_{v,t}|}\right)^T Y \left(\frac{N_{u,t+1}}{|\mathcal{S}_{u,t+1}|} - \frac{N_{v,t+1}}{|\mathcal{S}_{v,t+1}|}\right)$$

- Compute the final estimate by averaging over nearby points

$$\hat{F}_{ui} = \frac{1}{|\mathcal{B}_{ui}|}\sum_{(v,j)\in\mathcal{B}_{ui}} Y_{vj}$$

  where $\mathcal{B}_{ui} = \{v : d_{uv} \leq \eta, j : d_{ij} \leq \eta\}$

*omitted sample splitting for different algorithm steps, used in analysis

Theorem [Borgs-Chayes-Lee-Shah]

Assuming that

- Latent function f is L-Lipschitz, bounded in $[0,1]$, symmetric

  $f(\alpha_u, \alpha_v) = \sum_{k=1}^{r} \lambda_k q_k(\alpha_u) q_k(\alpha_v)$

- Latent features α sampled iid $\sim U[0,1]$
- Each entry observed independently w/prob $p = \omega(r^5 n^{-1})$
- Independent bounded noise $\in [-1,1]$

the algorithm with $t = \Theta\left(\frac{\ln(1/p)}{\ln(pn)}\right)$ achieves

$$\mathrm{MSE} = O\left(\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{2t} \frac{r^2 \lambda_{\max}}{(pn)^{\frac{1}{2}-\theta}}\right).$$

Intuition

- Let F denote Hilbert-Schmidt integral operator assoc to kernel f

$$(Fg)(x) = \int f(x,\alpha)g(\alpha)d\alpha$$

- Consider spectral decomposition of F

$$f(\alpha_u, \alpha_v) = \sum_{k=1}^{r} \lambda_k q_k(\alpha_u) q_k(\alpha_v)$$

for $\int q_k(\alpha)q_h(\alpha)d\alpha = \begin{cases} 0 & \text{if } k \neq h \\ 1 & \text{if } k = h \end{cases}$

- Analyze neighborhood growth via spectrum of F, e.g. the expected product of weights along path of length three from α_0 to α_3

$$\int \int f(\alpha_0, \alpha_1) f(\alpha_1, \alpha_2) f(\alpha_2, \alpha_3) d\alpha_2 d\alpha_1 = \sum_k \lambda_k^3 q_k(\alpha_0) q_k(\alpha_3)$$

Proof Sketch

- Ideally, we would like our distance estimates to approximate

$$\|f_u - f_v\|_2^2 := \int (f(\alpha_u, \alpha) - f(\alpha_v, \alpha))^2 d\alpha = \sum_k \lambda_k^2 (q_k(\alpha_u) - q_k(\alpha_v))^2$$

- We can show that with high probability,

$$\boxed{\begin{aligned} d_{uv} &= \frac{1}{p}\left(\frac{N_{u,t}}{\mathcal{S}_{u,t}} - \frac{N_{v,t}}{\mathcal{S}_{v,t}}\right)^T Y \left(\frac{N_{u,t+1}}{\mathcal{S}_{u,t+1}} - \frac{N_{v,t+1}}{\mathcal{S}_{v,t+1}}\right) \\ &\approx \sum_k \lambda_k^{2(t+1)} (q_k(\alpha_u) - q_k(\alpha_v))^2 \end{aligned}}$$

- Therefore, with high probability (and using Lipschitz assumption)

$$\begin{aligned} d_{uv} \le \eta \implies & \|f_u - f_v\|_2^2 \le \eta \lambda_{\max}^{-2t} \\ \implies & \|f_u - f_v\|_\infty \le \left(6L\eta\lambda_{\max}^{-2t}\right)^{1/3} \end{aligned}$$

- Choose final parameter η to tradeoff between bias and variance

Proof Sketch

- Want to show

$$(*)\quad \boxed{\frac{1}{p}\left(\frac{N_{u,t}}{\mathcal{S}_{u,t}} - \frac{N_{v,t}}{\mathcal{S}_{v,t}}\right)^T Y \left(\frac{N_{u,t+1}}{\mathcal{S}_{u,t+1}} - \frac{N_{v,t+1}}{\mathcal{S}_{v,t+1}}\right) \approx \sum_k \lambda_k^{2(t+1)} (q_k(\alpha_u) - q_k(\alpha_v))^2}$$

- Lemma 1: With high probability, for all k,

$$(**)\quad \sum_i q_k(\alpha_i) \frac{N_{u,t}(i)}{|\mathcal{S}_{u,t}|} \approx \lambda_k^t q_k(\alpha_u)$$

  - Conditioned on s-radius neighborhood, $\sum_i q_k(\alpha_i) \frac{N_{u,s+1}(i)}{|\mathcal{S}_{u,s+1}|}$ is a sum of iid r.v. with mean $\lambda_k \sum_i q_k(\alpha_i) \frac{N_{u,s}(i)}{|\mathcal{S}_{u,s}|}$

- Lemma 2: Conditioned on (**), then (*) holds with high probability
  - Assuming Y is fresh sample, then LHS of (*) is a sum of iid r.v.

Proof Sketch of Lemma 1

- We want to show that $\boxed{\sum_i q_k(\alpha_i)\frac{N_{u,r}(i)}{|\mathcal{S}_{u,r}|} \approx \lambda_k^r q_k(\alpha_u)}$
- Recall that $N_{u,r}(i) = \mathbb{I}_{\{|\mathcal{P}(\mathcal{T}_u,i)|=r\}} \prod_{e\in\mathcal{P}(\mathcal{T}_u,i)} Y_e$
- By Chernoff's bound, for all $s \in \{0,1,2,\dots r-1\}$, $|\mathcal{S}_{u,s}| \approx (pn)^s$
- For all $s \in \{0,1,2,\dots r-1\}$, conditioning on the s-radius ball around u (i.e. vertices and edges within distance s of u), for $w \in [n] \setminus S_{u,s}$

$$\mathbb{P}(w \notin \mathcal{S}_{u,s+1}) = (1-p)^{|\mathcal{S}_{u,s}|}$$

$$\sum_i q_k(\alpha_i)\frac{N_{u,s+1}(i)}{|\mathcal{S}_{u,s+1}|} = \frac{1}{|\mathcal{S}_{u,s+1}|}\sum_i \underbrace{q_k(\alpha_i)\sum_{b\in\mathcal{S}_{u,s}} \mathbb{I}((b,i)\in\Omega)Y_{bi}N_{u,s}(b)}_{X_i \text{ iid}}$$

Proof Sketch of Lemma 1

- Conditioned on the s-radius ball around u and the set $S_{u,s+1}$

$$\mathbb{E}[X_i] = \sum_{b \in \mathcal{S}_{u,s}} \mathbb{E}[\mathbb{I}((b,i) \in \Omega)]\mathbb{E}[Y_{bi} q_k(\alpha_i)] N_{u,s}(b)$$

edges independent from data value

Proof Sketch of Lemma 1

- Conditioned on the s-radius ball around u and the set $S_{u,s+1}$

$$\mathbb{E}[X_i] = \sum_{b \in \mathcal{S}_{u,s}} \mathbb{E}[\mathbb{I}((b,i) \in \Omega)]\mathbb{E}[Y_{bi} q_k(\alpha_i)] N_{u,s}(b)$$

parent equally likely to be any $b \in S_{u,r}$

$$= \sum_{b \in \mathcal{S}_{u,s}} \frac{N_{u,s}(b)}{|\mathcal{S}_{u,s}|} \mathbb{E}[(f(\alpha_b, \alpha_i) + \epsilon_{bi}) q_k(\alpha_i)]$$

$$= \sum_{b \in \mathcal{S}_{u,s}} \frac{N_{u,s}(b)}{|\mathcal{S}_{u,s}|} \lambda_k q_k(\alpha_b)$$

By orthogonality of $\{q_k\}$

$$\mathrm{Var}[X_i] \le (\sigma^2 + 1) \sum_{b \in \mathcal{S}_{u,s}} \frac{N_{u,s}^2(b)}{|\mathcal{S}_{u,s}|} \le (\sigma^2+1) \frac{\|N_{u,s}\|_1}{|\mathcal{S}_{u,s}|} = O(r \lambda_{\max}^s)$$

- Use Bernstein's inequality to show that

$$\approx (pn)^{s+1}$$

$$\mathbb{P}\left(\left| \sum_i q_k(\alpha_i) \frac{N_{u,s+1}(i)}{|\mathcal{S}_{u,s+1}|} - \lambda_k \sum_i q_k(\alpha_i) \frac{N_{u,s}(i)}{|\mathcal{S}_{u,s}|} \right| \ge t_{sk} \right) \le \exp\left(-\frac{c_0 |\mathcal{S}_{u,s+1}| t_{sk}^2}{r \lambda_{\max}^s} \right)$$

Proof Sketch of Lemma 1

- Choose $t_{sk} = (pn)^{-\frac{1}{2}+\theta}\left(\frac{\lambda_k}{2}\right)^s$, then with prob $1 - \exp\left(-\frac{c_1 (pn)^{2\theta}}{d}\right)$ for all $s \in \{0,1,2 \ldots r-1\}$,

$$\left|\sum_i q_k(\alpha_i)\left(\frac{N_{u,s+1}}{|\mathcal{S}_{u,s+1}|} - \lambda_k \frac{N_{u,s}}{|\mathcal{S}_{u,s}|}\right)\right| \geq (pn)^{-\frac{1}{2}+\theta}\left(\frac{\lambda_k}{2}\right)^s$$

- Conditioned on above good event,

$$\begin{aligned}
&\sum_i q_k(\alpha_i)\frac{N_{u,r}(i)}{|\mathcal{S}_{u,r}|} - \lambda_k^r q_k(\alpha_u) \\
&= \sum_{t=1}^r \lambda_k^{t-1} \sum_i q_k(\alpha_i)\left(\frac{N_{u,r-t+1}(i)}{|\mathcal{S}_{u,r-t+1}|} - \lambda_k \frac{N_{u,r-t}(i)}{|\mathcal{S}_{u,r-t}|}\right) \\
&\leq \sum_{t=1}^r \lambda_k^{t-1} (pn)^{-\frac{1}{2}+\theta}\left(\frac{\lambda_k}{2}\right)^{r-t} \\
&\leq \lambda_k^{r-1} (pn)^{-\frac{1}{2}+\theta} \sum_{t=1}^r \left(\frac{\lambda_k}{2\lambda_k}\right)^{r-t} \\
&\leq 2\lambda_k^{r-1} (pn)^{-\frac{1}{2}+\theta}
\end{aligned}$$

- Thus with high probability $\boxed{\sum_i q_k(\alpha_i)\frac{N_{u,r}(i)}{|\mathcal{S}_{u,r}|} \approx \lambda_k^r q_k(\alpha_u)}$

Proof Sketch of Lemma 2

We want to show $\boxed{d_{uv} \approx \sum_k \lambda_k^{2(t+1)} (q_k(\alpha_u) - q_k(\alpha_v))^2}$

$$\mathbb{E}[d_{uv} \mid N_{u,t}, \mathcal{S}_{u,t}, N_{v,t}, \mathcal{S}_{v,t}, N_{u,t+1}, \mathcal{S}_{u,t+1}, N_{v,t+1}, \mathcal{S}_{v,t+1}]$$

$$= \frac{1}{p} \left(\frac{N_{u,t}}{|\mathcal{S}_{u,t}|} - \frac{N_{v,t}}{|\mathcal{S}_{v,t}|} \right)^T \mathbb{E}[Y] \left(\frac{N_{u,t+1}}{|\mathcal{S}_{u,t+1}|} - \frac{N_{v,t+1}}{|\mathcal{S}_{v,t+1}|} \right)$$

Conditioned on expanded t-radius neighborhood, and assuming Y is an independent fresh sample

Proof Sketch of Lemma 2

We want to show $\boxed{d_{uv} \approx \sum_k \lambda_k^{2(t+1)} (q_k(\alpha_u) - q_k(\alpha_v))^2}$

$$\mathbb{E}[d_{uv} \mid N_{u,t}, \mathcal{S}_{u,t}, N_{v,t}, \mathcal{S}_{v,t}, N_{u,t+1}, \mathcal{S}_{u,t+1}, N_{v,t+1}, \mathcal{S}_{v,t+1}]$$

$$= \frac{1}{p} \left(\frac{N_{u,t}}{|\mathcal{S}_{u,t}|} - \frac{N_{v,t}}{|\mathcal{S}_{v,t}|} \right)^T \mathbb{E}[Y] \left(\frac{N_{u,t+1}}{|\mathcal{S}_{u,t+1}|} - \frac{N_{v,t+1}}{|\mathcal{S}_{v,t+1}|} \right)$$

$$= \sum_k \lambda_k \sum_i q_k(\alpha_i) \left(\frac{N_{u,t}(i)}{|\mathcal{S}_{u,t}|} - \frac{N_{v,t}(i)}{|\mathcal{S}_{v,t}|} \right) \sum_j q_k(\alpha_j) \left(\frac{N_{u,t+1}(j)}{|\mathcal{S}_{u,t+1}|} - \frac{N_{v,t+1}(j)}{|\mathcal{S}_{v,t+1}|} \right)$$

By assumption $\mathbb{E}[Y_{ij}] = pf(\alpha_i, \alpha_j) = p \sum_k \lambda_k q_k(\alpha_i) q_k(\alpha_j)$,
in fact can write expression d_{uv} as sum of iid random variables

Proof Sketch of Lemma 2

We want to show $\boxed{d_{uv} \approx \sum_k \lambda_k^{2(t+1)} (q_k(\alpha_u) - q_k(\alpha_v))^2}$

$$\mathbb{E}[d_{uv} \mid N_{u,t}, \mathcal{S}_{u,t}, N_{v,t}, \mathcal{S}_{v,t}, N_{u,t+1}, \mathcal{S}_{u,t+1}, N_{v,t+1}, \mathcal{S}_{v,t+1}]$$

$$= \frac{1}{p} \left(\frac{N_{u,t}}{|\mathcal{S}_{u,t}|} - \frac{N_{v,t}}{|\mathcal{S}_{v,t}|} \right)^T \mathbb{E}[Y] \left(\frac{N_{u,t+1}}{|\mathcal{S}_{u,t+1}|} - \frac{N_{v,t+1}}{|\mathcal{S}_{v,t+1}|} \right)$$

$$= \sum_k \lambda_k \sum_i q_k(\alpha_i) \left(\frac{N_{u,t}(i)}{|\mathcal{S}_{u,t}|} - \frac{N_{v,t}(i)}{|\mathcal{S}_{v,t}|} \right) \sum_j q_k(\alpha_j) \left(\frac{N_{u,t+1}(j)}{|\mathcal{S}_{u,t+1}|} - \frac{N_{v,t+1}(j)}{|\mathcal{S}_{v,t+1}|} \right)$$

$$\approx \sum_k \lambda_k^{2t+2} (q_k(\alpha_u) - q_k(\alpha_v))^2$$

Conditioned on event that for all k and $s \in \{t, t+1\}$

$$\sum_i q_k(\alpha_i) \frac{N_{u,s}(i)}{|\mathcal{S}_{u,s}|} \approx \lambda_k^s q_k(\alpha_u) \text{ and } \sum_i q_k(\alpha_i) \frac{N_{v,s}(i)}{|\mathcal{S}_{v,s}|} \approx \lambda_k^s q_k(\alpha_v)$$

by Lemma 1, this holds with high probability

Proof Sketch of Lemma 2

We want to show $\boxed{d_{uv} \approx \sum_k \lambda_k^{2(t+1)} (q_k(\alpha_u) - q_k(\alpha_v))^2}$

$$\mathbb{E}[d_{uv} \mid N_{u,t}, \mathcal{S}_{u,t}, N_{v,t}, \mathcal{S}_{v,t}, N_{u,t+1}, \mathcal{S}_{u,t+1}, N_{v,t+1}, \mathcal{S}_{v,t+1}]$$
$$= \frac{1}{p} \left(\frac{N_{u,t}}{|\mathcal{S}_{u,t}|} - \frac{N_{v,t}}{|\mathcal{S}_{v,t}|} \right)^T \mathbb{E}[Y] \left(\frac{N_{u,t+1}}{|\mathcal{S}_{u,t+1}|} - \frac{N_{v,t+1}}{|\mathcal{S}_{v,t+1}|} \right)$$
$$= \sum_k \lambda_k \sum_i q_k(\alpha_i) \left(\frac{N_{u,t}(i)}{|\mathcal{S}_{u,t}|} - \frac{N_{v,t}(i)}{|\mathcal{S}_{v,t}|} \right) \sum_j q_k(\alpha_j) \left(\frac{N_{u,t+1}(j)}{|\mathcal{S}_{u,t+1}|} - \frac{N_{v,t+1}(j)}{|\mathcal{S}_{v,t+1}|} \right)$$
$$\approx \sum_k \lambda_k^{2t+2} (q_k(\alpha_u) - q_k(\alpha_v))^2$$

- Concentration follows via Bernstein's inequality and variance bounds

Sample Complexity Comparison

Algorithm	References	Function Class	Noise Model	Guaranteed Recovery	Observations mnp (m=n)
SVT	[Chatterjee]	Lipschitz	Arbitrary	Approx.	$n^{\frac{2r+2}{r+2}} \log^6 n$
SVT	[Chatterjee]	Low-rank	Arbitrary	Approx.	$nr \log^6 n$
Convex	[Recht]	Low-rank	No Noise	Exact	$nr \log^2 n$
Convex	[CandesPlan]	Low-rank	Additive	Approx.	$nr \log^2 n$
Non-Convex	[KeMonOh]	Low-rank	No Noise	Exact	$nr \log n$
Non-Convex	[KeMonOh]	Low-rank	Additive	Approx.	$nr \log n$
Near Nghbr	[LeeLiSoSh]	Lipschitz	Additive	Approx.	$n^{\frac{3}{2}}$ polylog$n$
Near Nghbr	[BoChLeeSh]	Low-rank	Arbitrary	Approx.	$nr^5 \omega(1)$

Discussion

- Connections to acyclic belief propagation
- Connections to nonbacktracking operator
- Is finite spectrum necessary?