

Better Tools for Proving Convergence for MM Algorithms?

Eric Chi

Department of Statistics, North Carolina State University

March 21, 2019

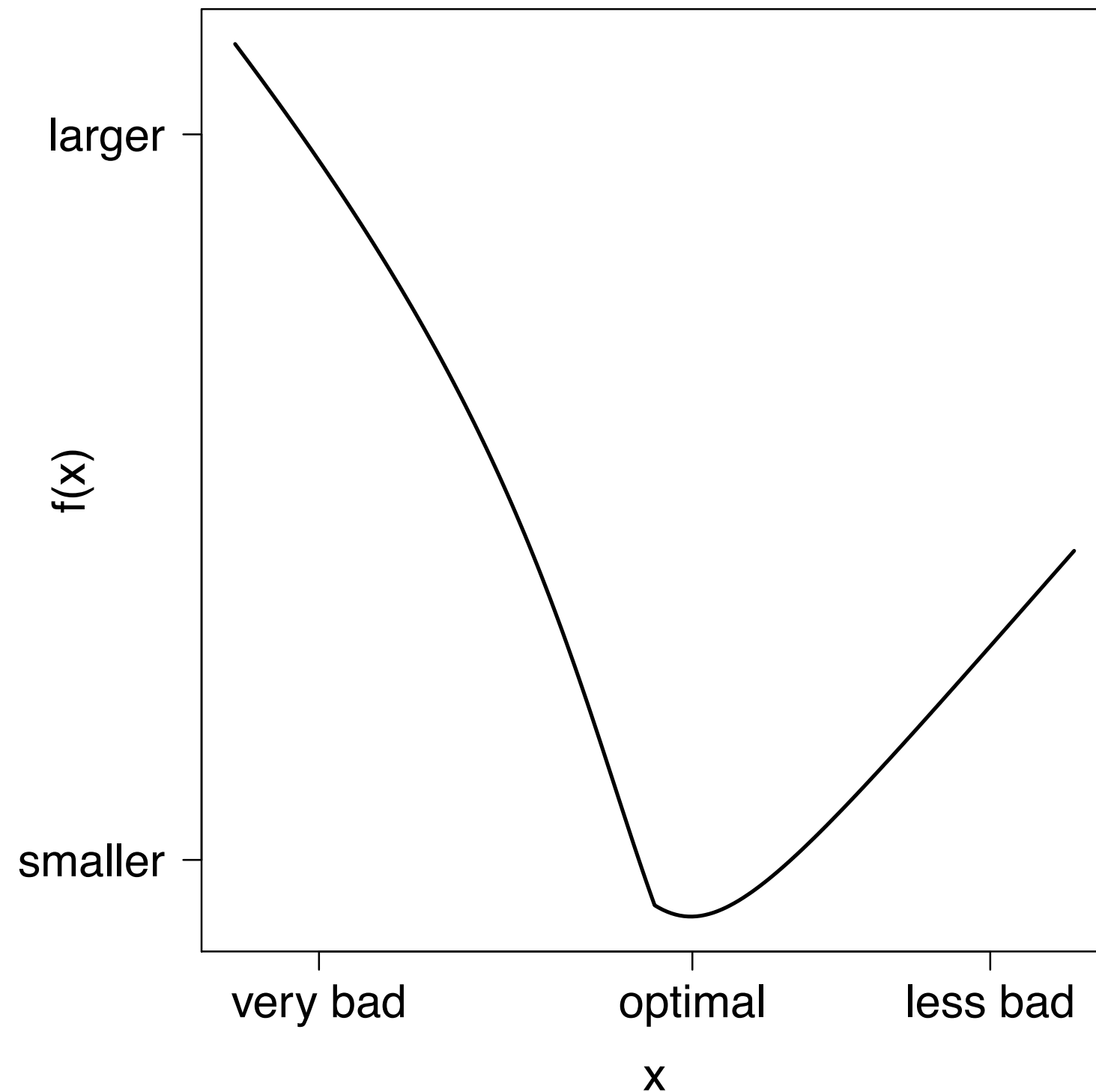
Definition

A function $g(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ is said to majorize the function $f(\boldsymbol{\theta})$ at $\tilde{\boldsymbol{\theta}}$ if

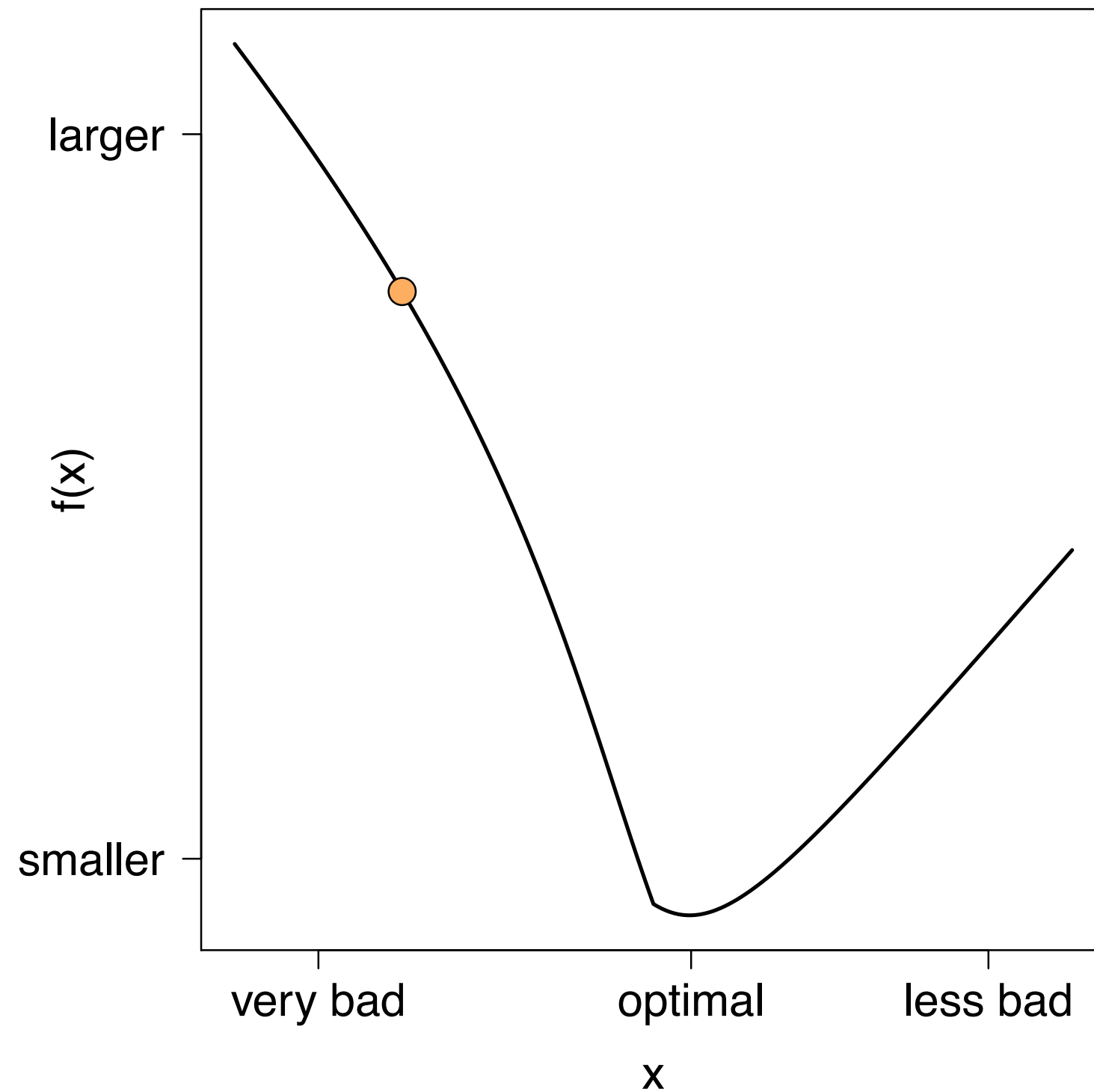
1. $g(\tilde{\boldsymbol{\theta}} \mid \tilde{\boldsymbol{\theta}}) = f(\tilde{\boldsymbol{\theta}})$ and
2. $g(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}) \geq f(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$.

- MM = Majorization-Minimization (or Minorization-Maximization)
- The EM algorithm is an example of the latter

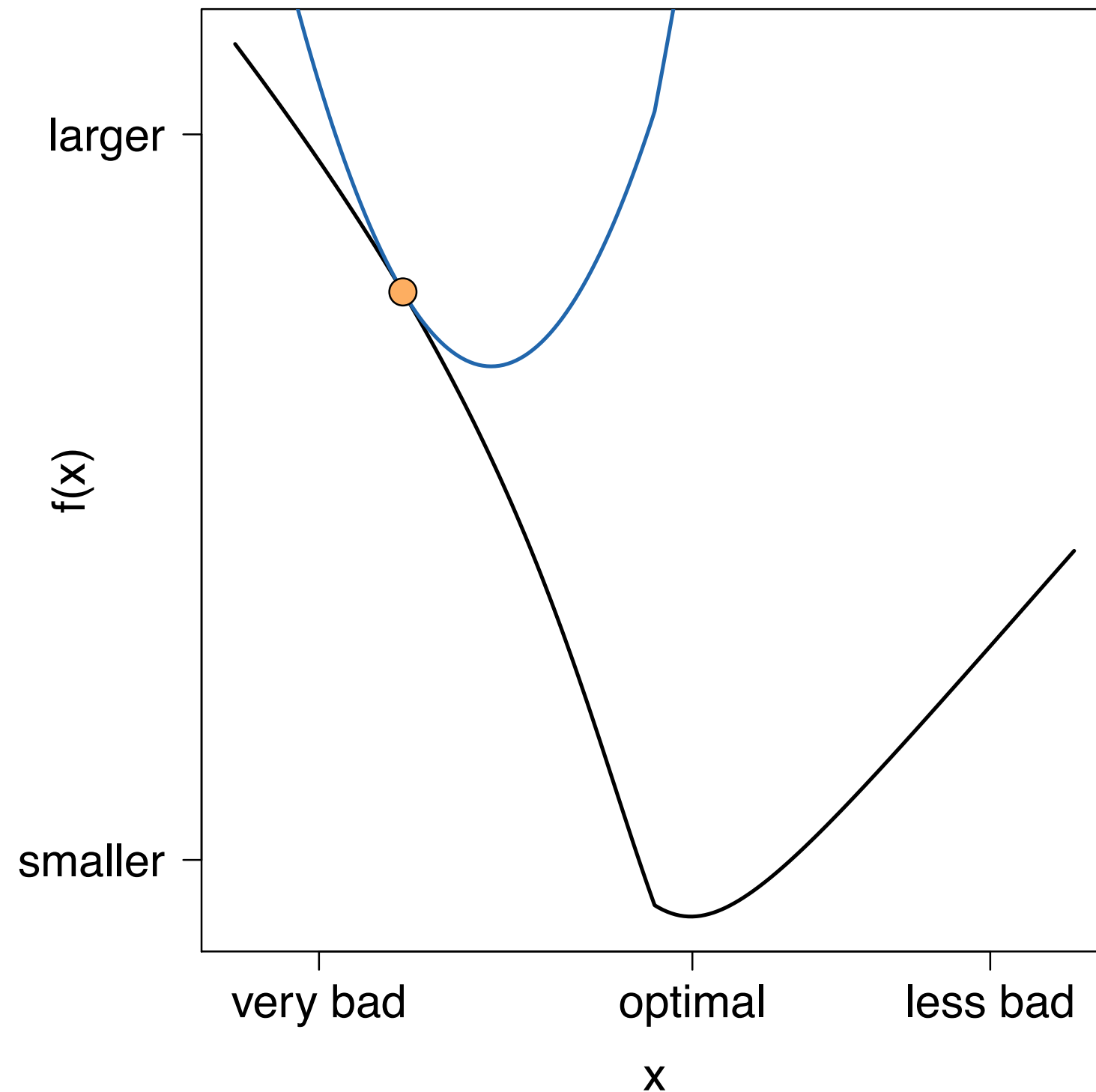
MM Algorithm in Action



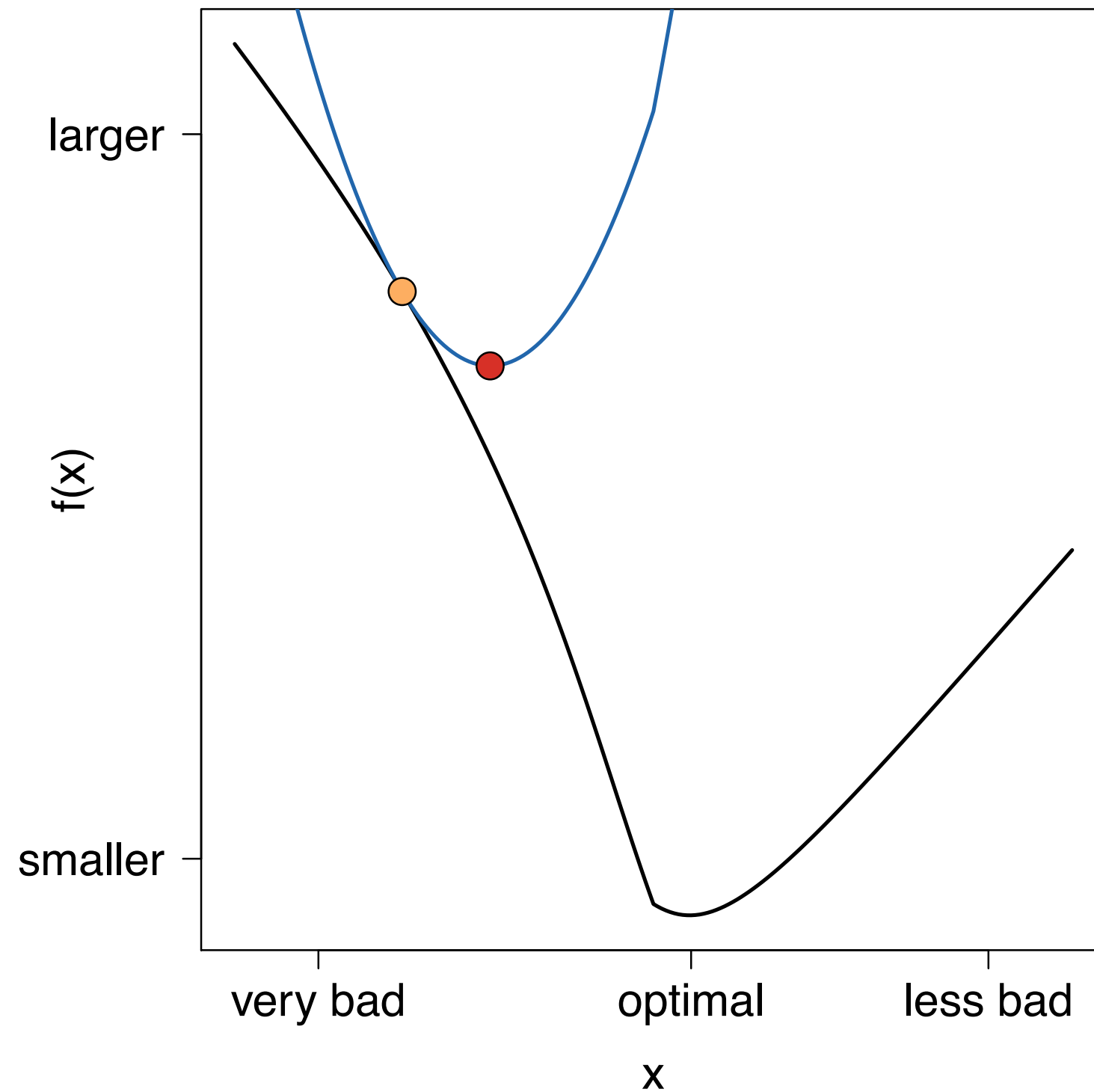
MM Algorithm in Action



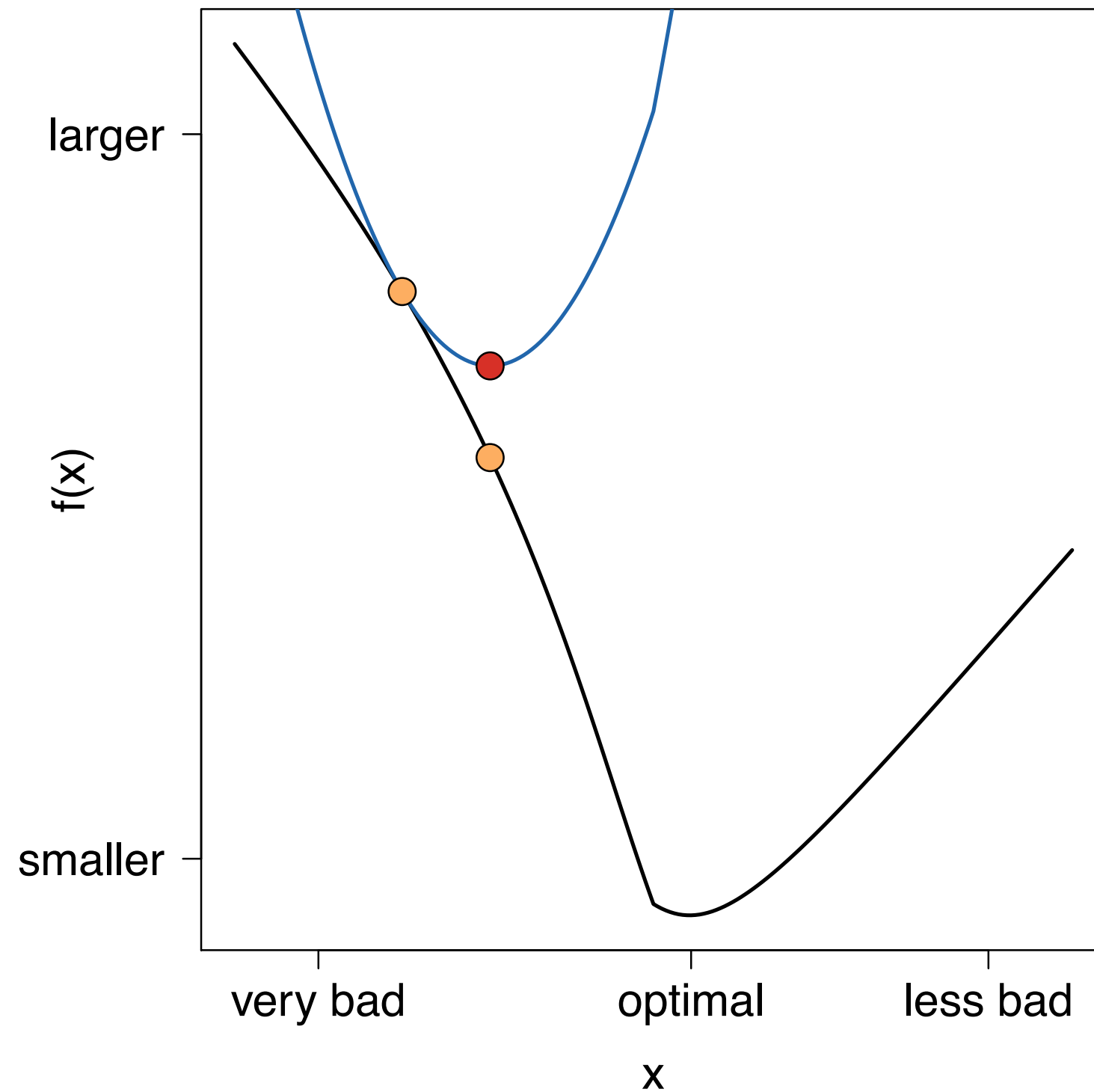
MM Algorithm in Action



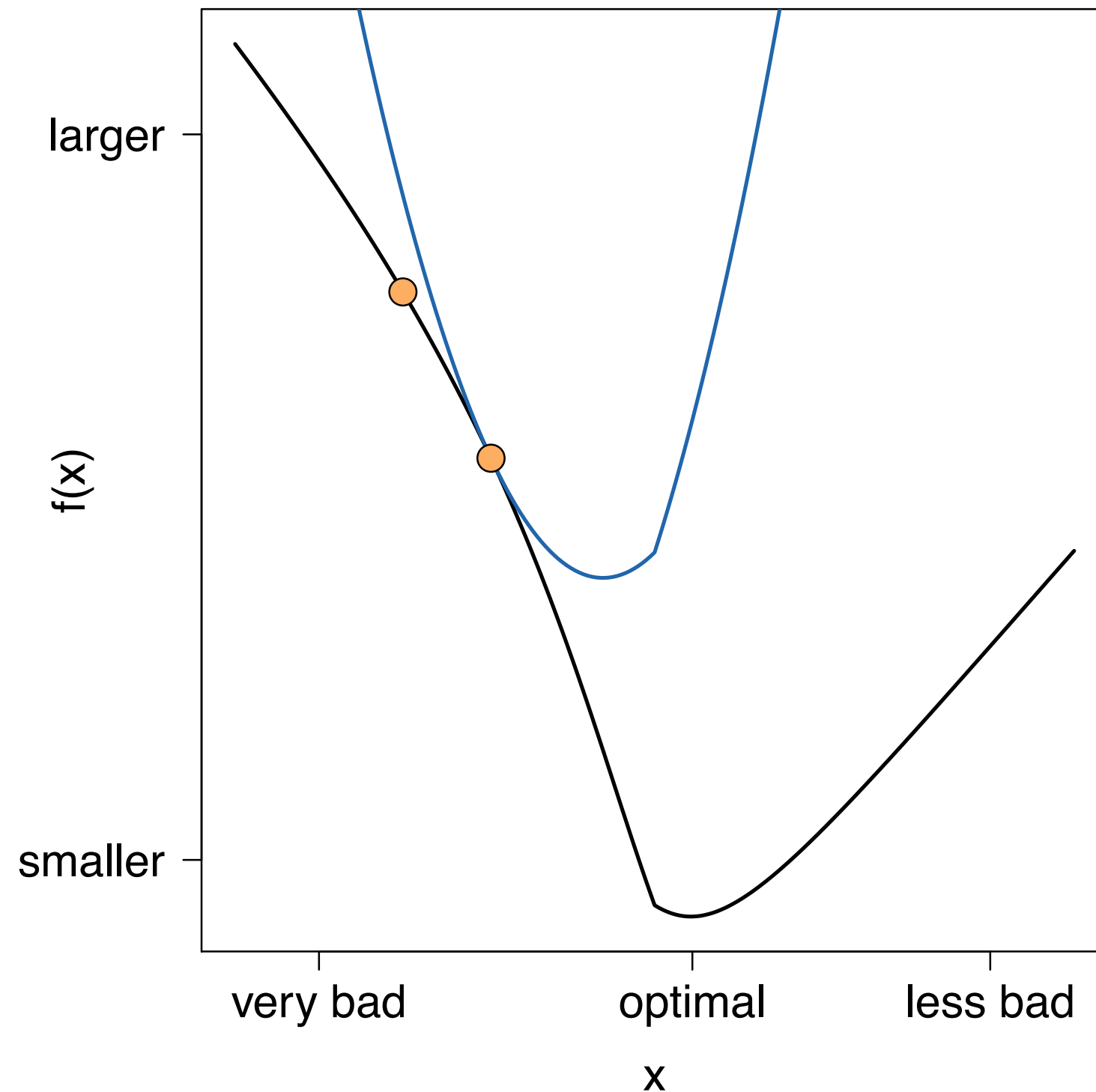
MM Algorithm in Action



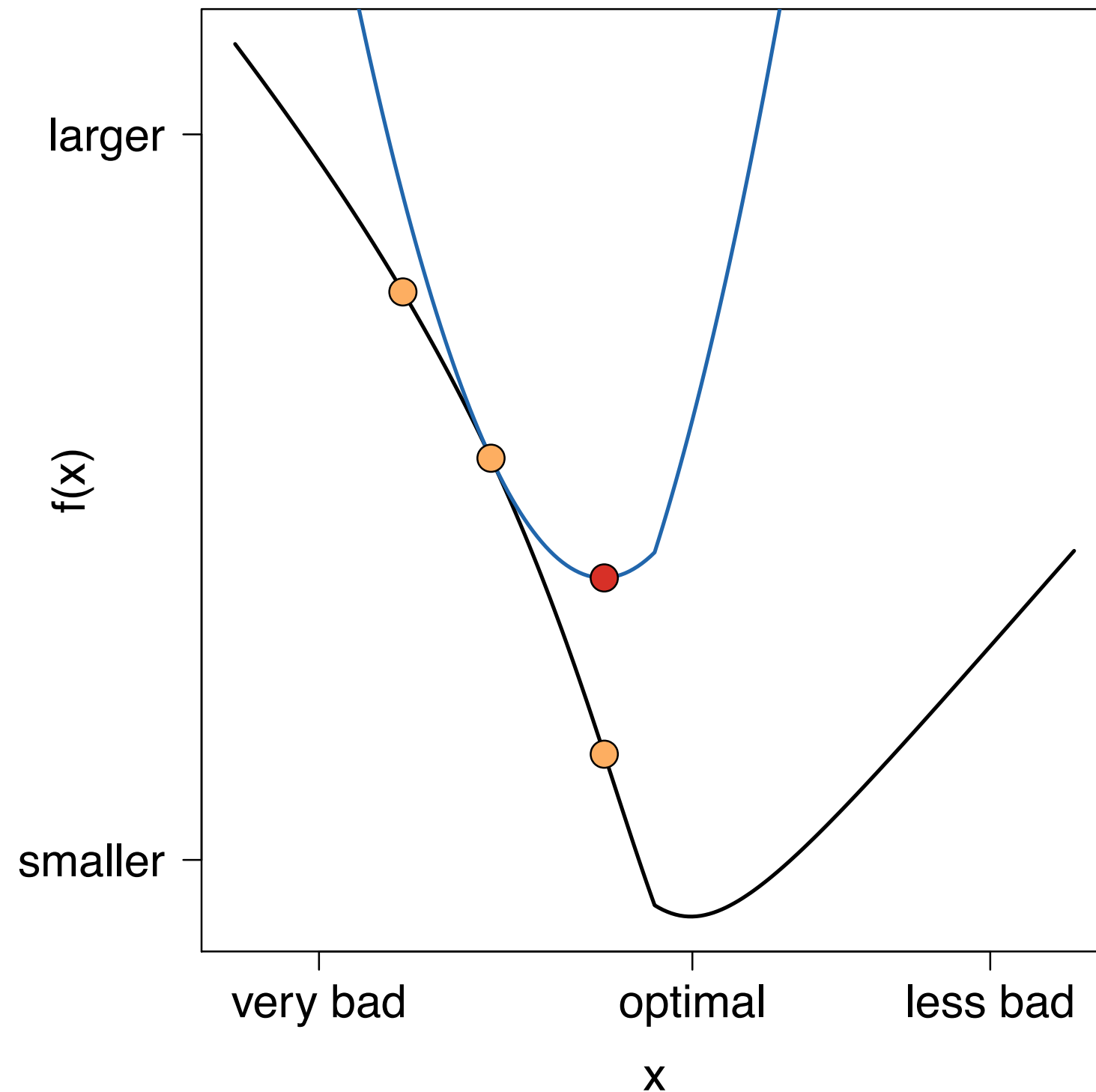
MM Algorithm in Action



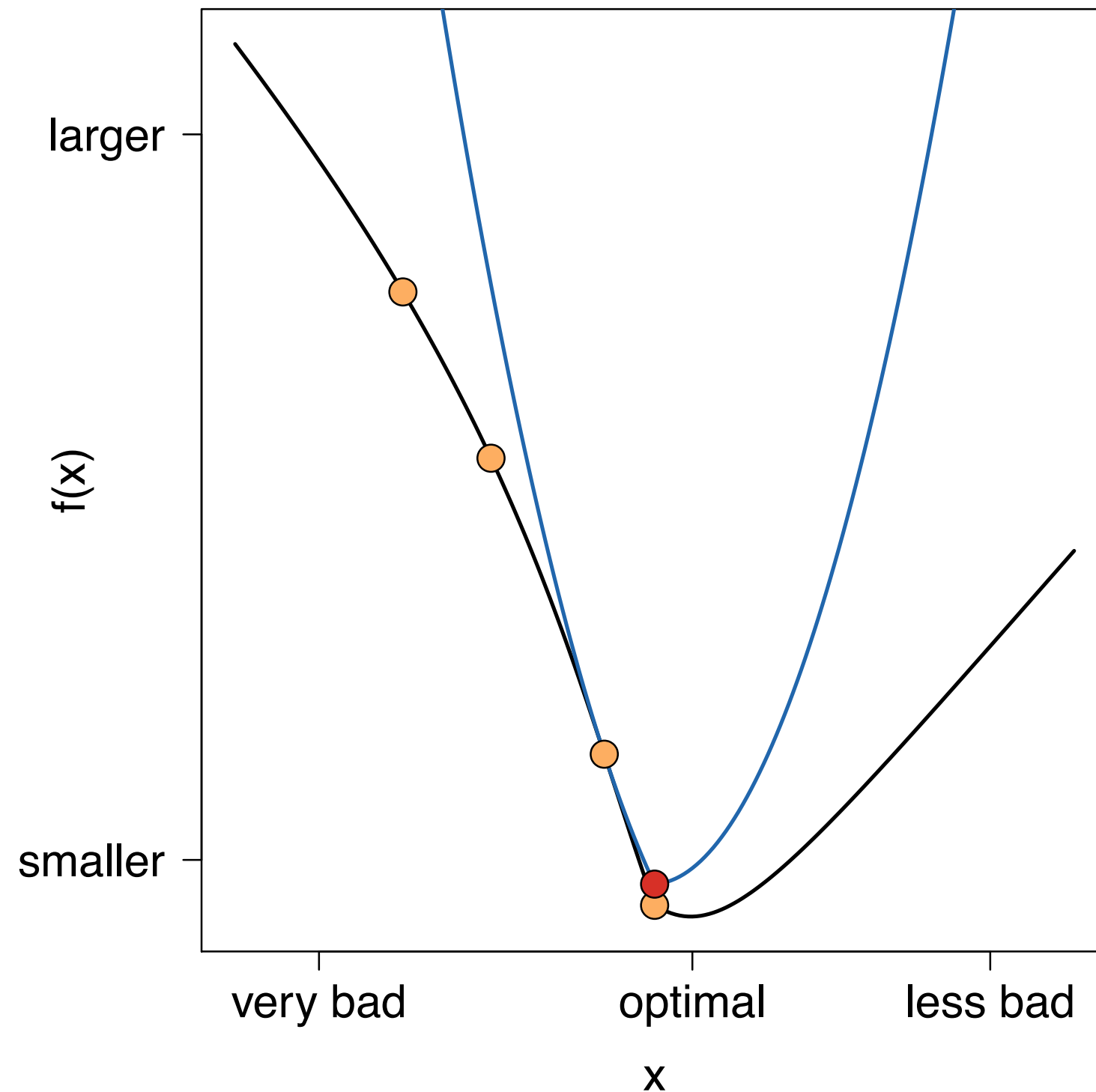
MM Algorithm in Action



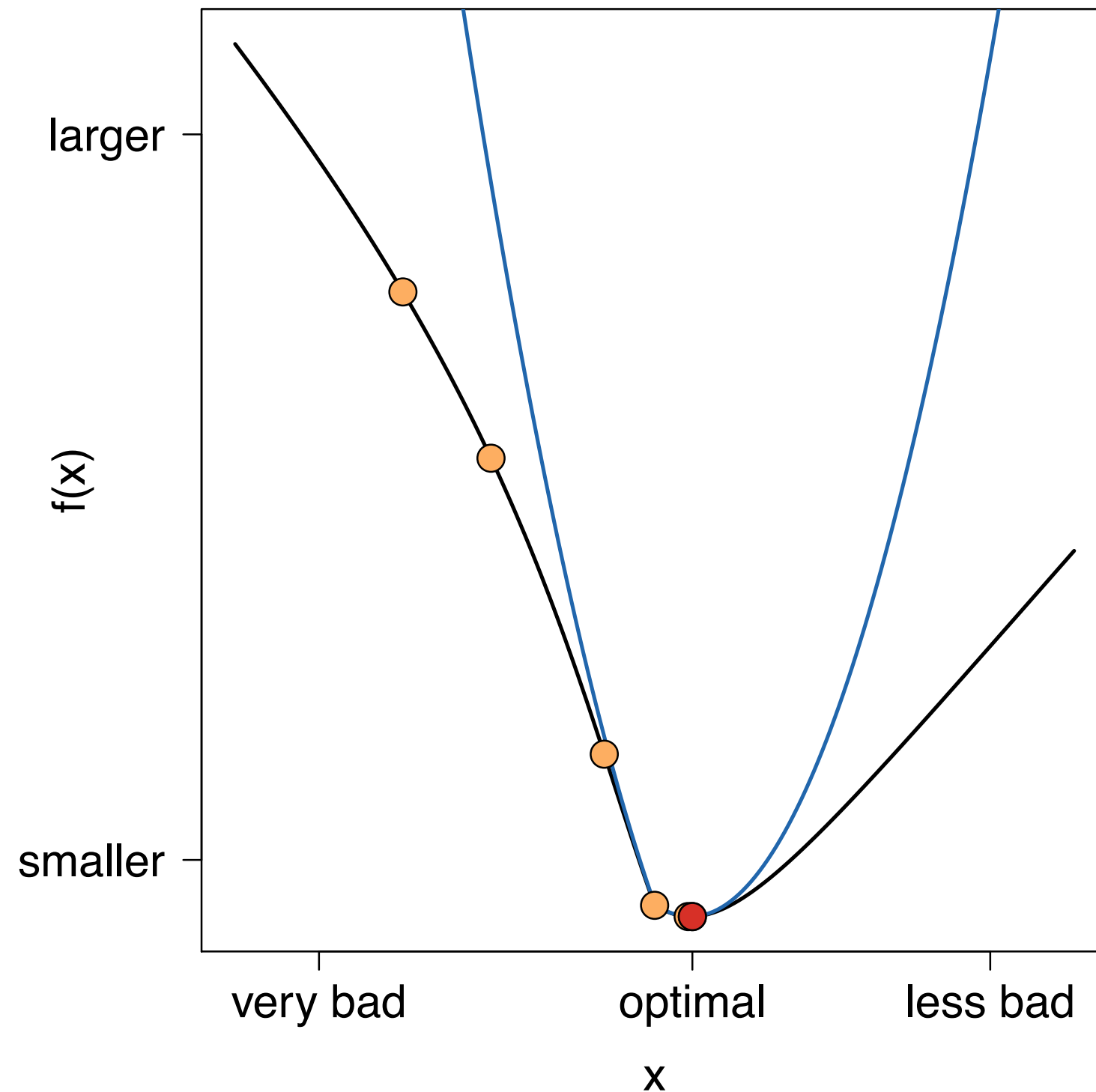
MM Algorithm in Action



MM Algorithm in Action



MM Algorithm in Action



Strategies for Constructing Majorizations

Lipschitz differentiability

- Gradient descent, proximal gradient descent

Jensen's inequality

- EM algorithms
- Nonnegative matrix/tensor factorizations

Supporting hyperplane property of convex functions

- SCAD, MCP

Cauchy-Schwartz inequality

- Multidimensional scaling

Majorizing zero

- Matrix completion

Convergence of MM algorithms

Proposition (Simple Zangwill / Meyer): Let f be a continuous function on a domain $\mathcal{D}$, and let ψ be a continuous iterative map from $\mathcal{D}$ into $\mathcal{D}$ such that $f(\psi(x)) < f(x)$ for all $x \in \mathcal{D}$ with $\psi(x) \neq x$. Suppose there is an x_0 such that the set $\mathcal{L}_f(x_0) \equiv \{x \in \mathcal{D} : f(x) \leq f(x_0)\}$ is compact. Define $x_{n+1} = \psi(x_n)$ for $n \in \mathbb{N}$. Then

1. the sequence of iterates $\{x_n\}$ has at least one limit point and all its limit points are fixed points of ψ .
2. the distance between successive iterates converges to 0, i.e., $\|x_{n+1} - x_n\| \rightarrow 0$.

Clean up workflow:

- Show that $\text{Fix-}\psi$ are optima / stationary points of f
- Show that there is only one optimum / finitely many stationary points of f

Case Study 1: The Biclustering Problem

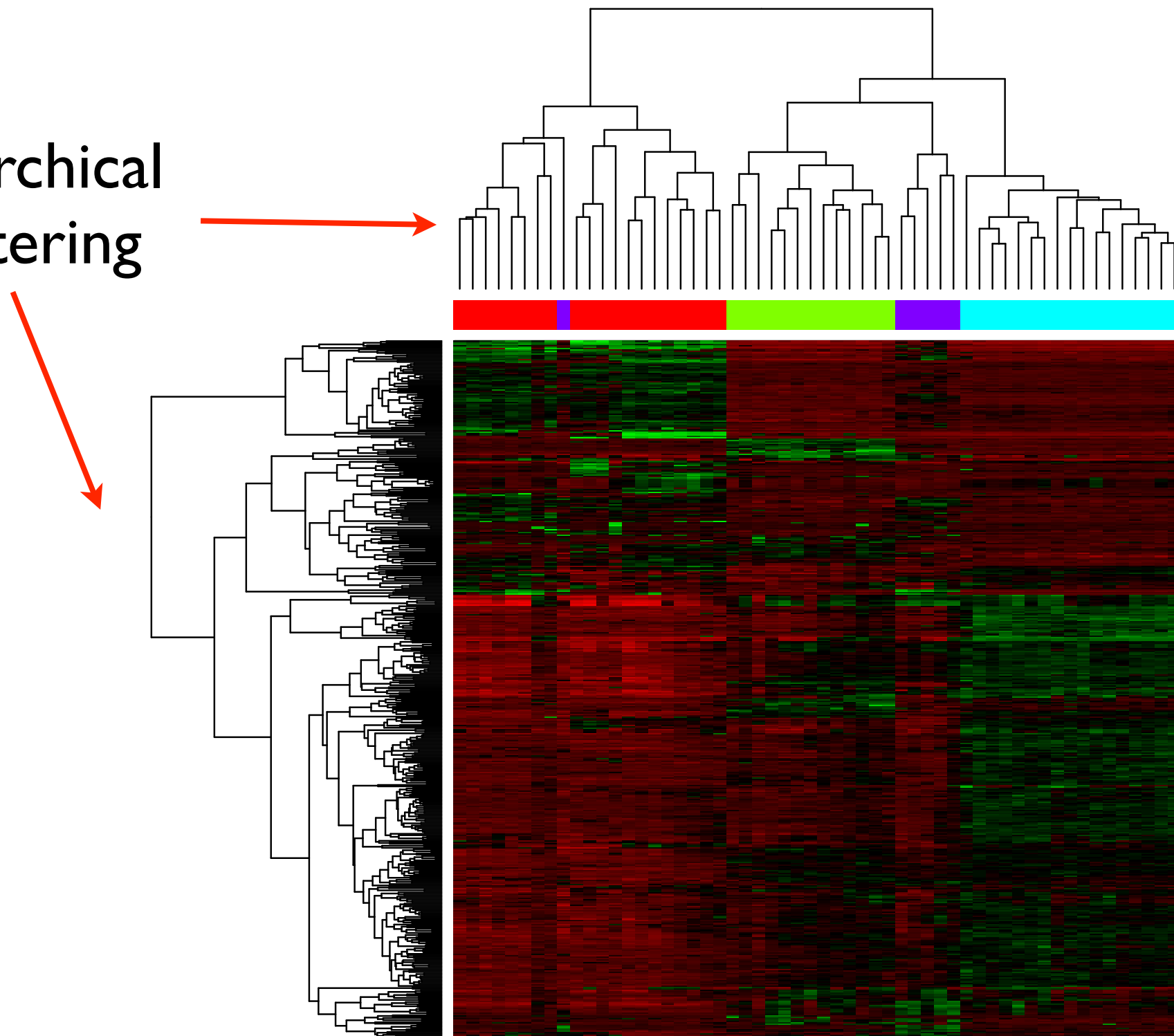
Task

Given a data matrix $\mathbf{X} \in \mathbb{R}^{m \times p}$, find subgroups of rows & columns that go together.

- **Text mining:** similar documents share a small set of highly correlated words.
- **Collaborative filtering:** likeminded customers share similar preferences for a subset of products
- **Cancer genomics:** subtypes of cancerous tumors share similar molecular profiles over a subset of genes

Simple Solution: Cluster Dendrogram

Hierarchical
Clustering



Genes

Tissue Sample

Simple Solution: Cluster Dendrogram

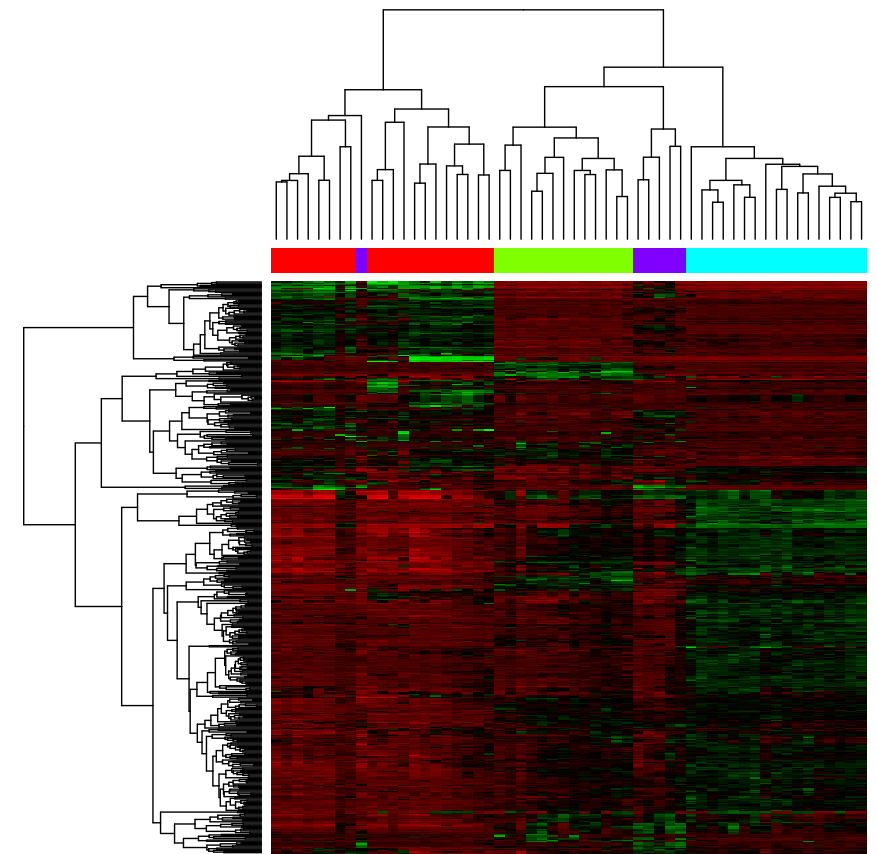
The Good

- Easy to interpret
- Fast computation - greedy algorithm

The Bad: Non-convex optimization problem

- Local Minimizers
- Instability (initialization, tuning parameters, or data)

The Ugly: How to choose number of biclusters?

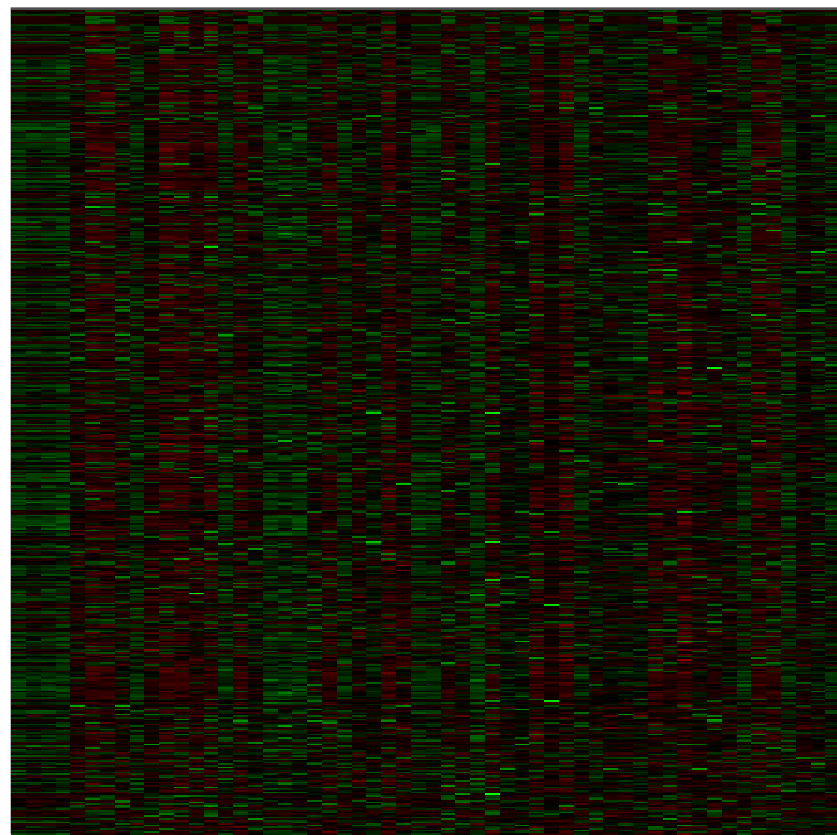


Solution: Convex Relaxation

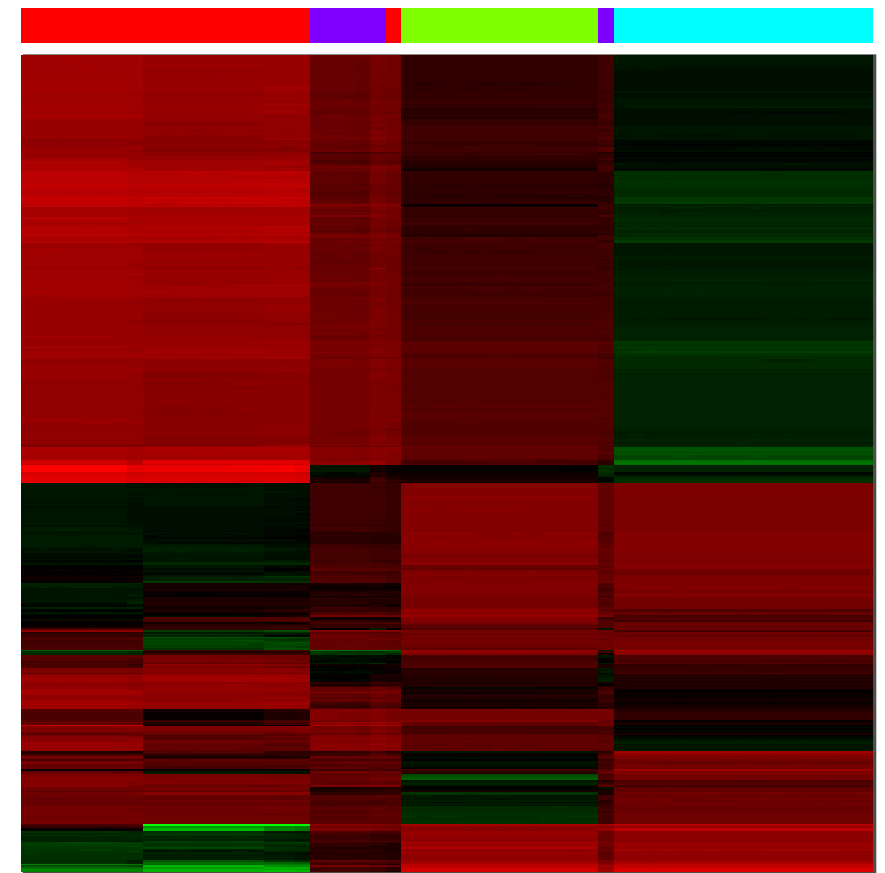
Solve combinatorially hard problem with a convex surrogate.

- All local minima are global minima
- Algorithms converge to global minimizer regardless of initialization

Solve a convex optimization problem to go from A to B



A



B

Convex Biclustering

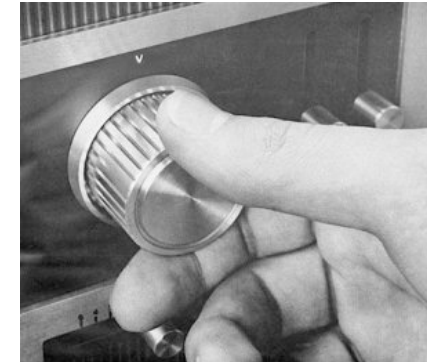
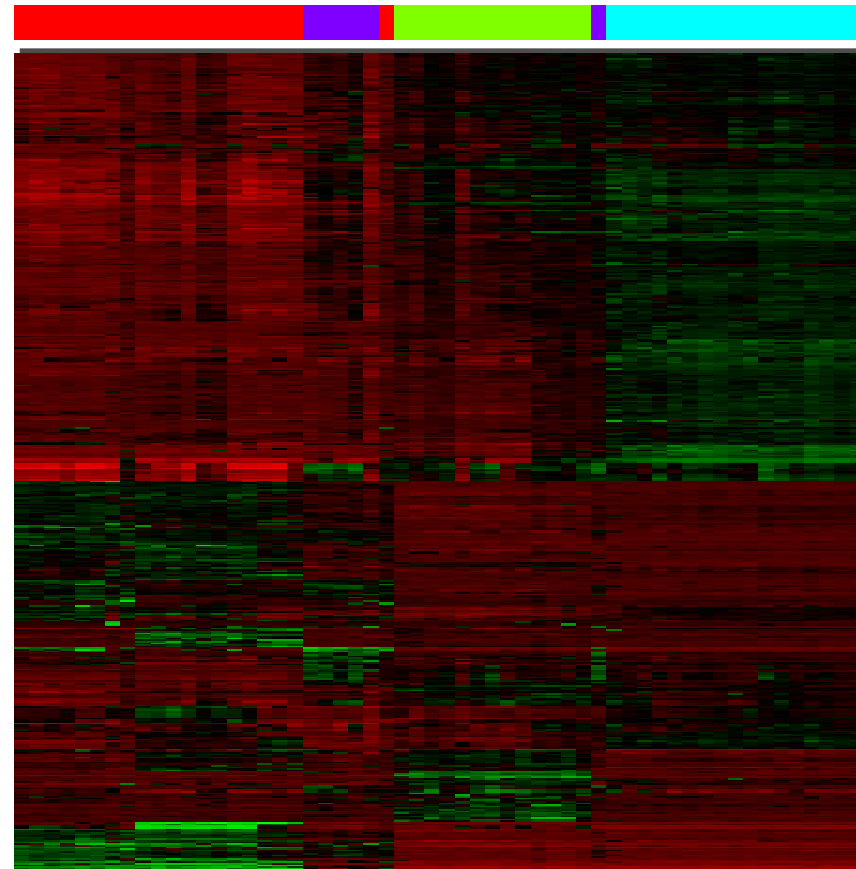
- Apply fusion penalties to **columns** and **rows**

$$\underset{\mathbf{U}}{\text{minimize}} F_{\gamma}(\mathbf{U}) = \frac{1}{2} \|\mathbf{X} - \mathbf{U}\|_F^2 + \gamma J(\mathbf{U})$$

$$J(\mathbf{U}) = \sum_{i < j} w_{ij} \|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \sum_{k < l} \tilde{w}_{kl} \|\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}\|_2$$

- Leads to “checkerboard” biclustering - every (i,j)th entry is assigned to one bicluster
 - ▶ *i*th and *j*th **columns** belong to same **column** cluster iff $\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j} = \mathbf{0}$
 - ▶ *k*th and *l*th **rows** belong to same **row** cluster iff $\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot} = \mathbf{0}$
- Need to make sure row and column penalties are on the same scale

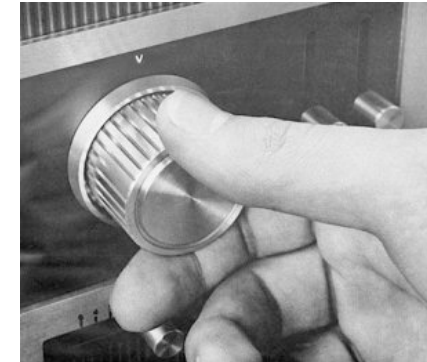
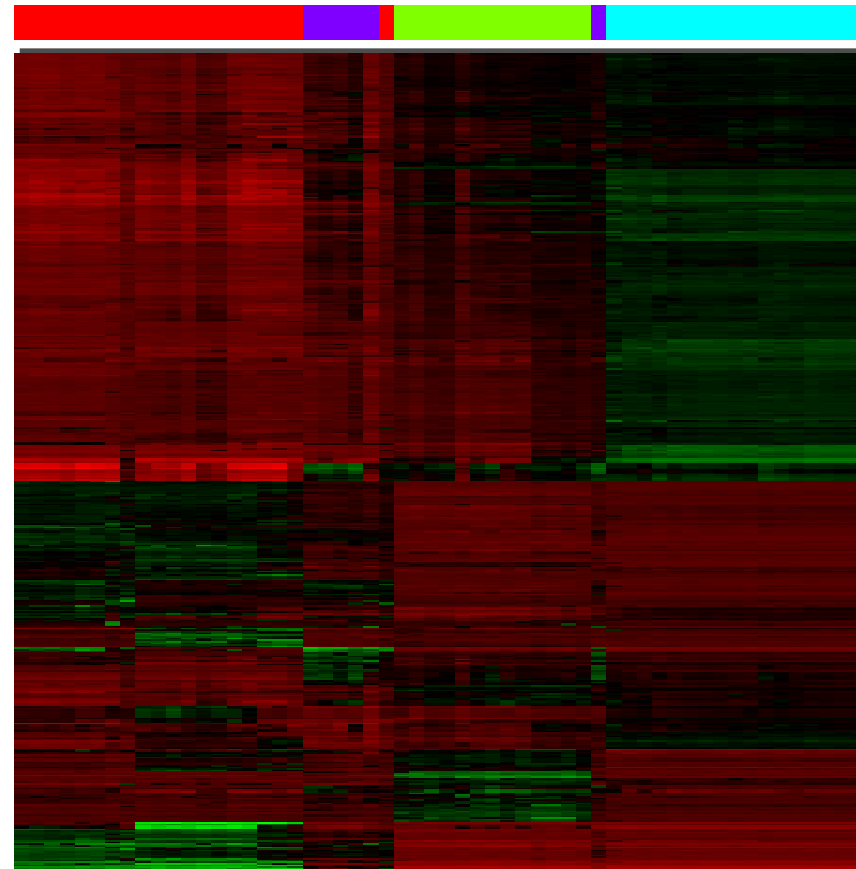
Biclustering path



γ

$$\text{minimize } \frac{1}{2} \|\mathbf{X} - \mathbf{U}\|_F^2 + \gamma \left[\sum_{i < j} w_{ij} \|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \sum_{k < l} \tilde{w}_{kl} \|\mathbf{U}_{k \cdot} - \mathbf{U}_{l \cdot}\|_2 \right]$$

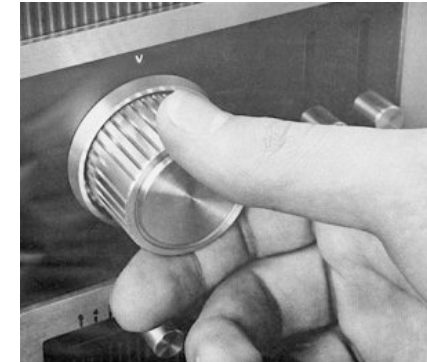
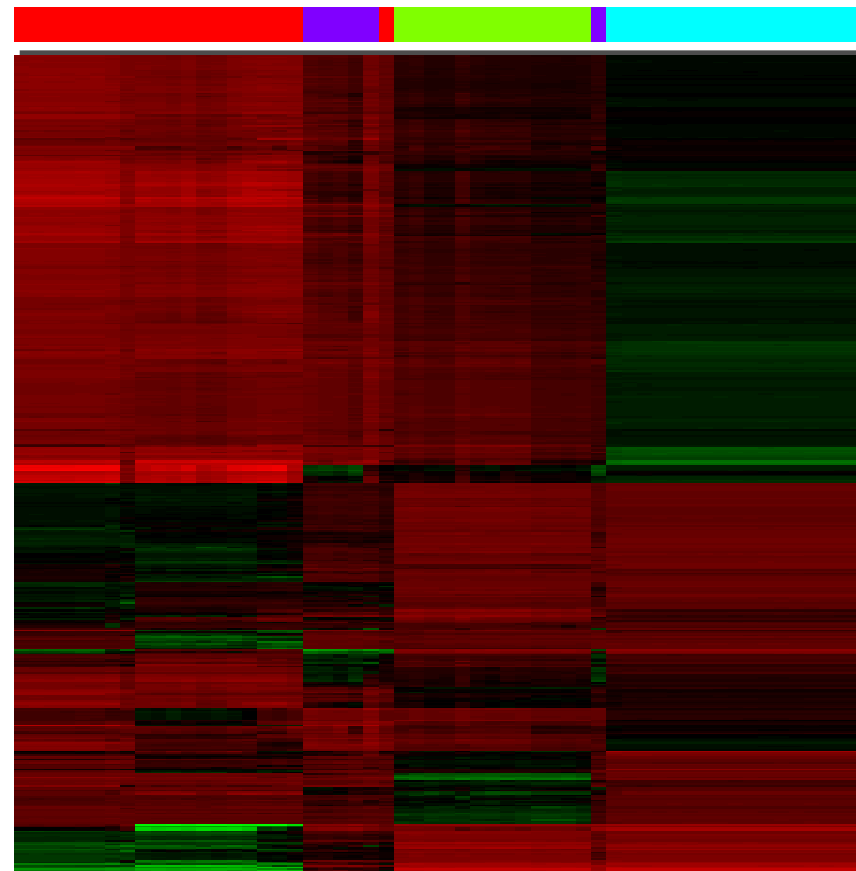
Biclustering path



γ

$$\text{minimize } \frac{1}{2} \|\mathbf{X} - \mathbf{U}\|_F^2 + \gamma \left[\sum_{i < j} w_{ij} \|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \sum_{k < l} \tilde{w}_{kl} \|\mathbf{U}_{k \cdot} - \mathbf{U}_{l \cdot}\|_2 \right]$$

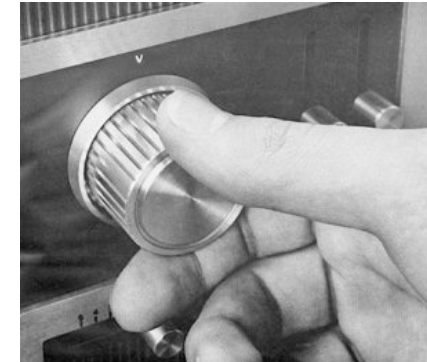
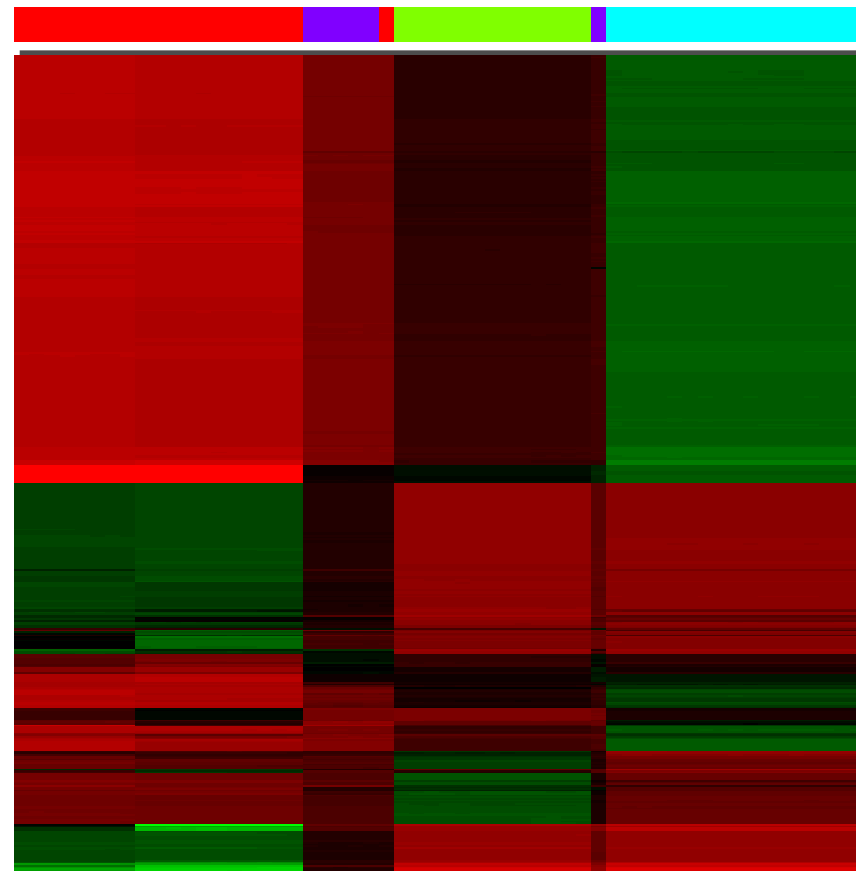
Biclustering path



γ

$$\text{minimize } \frac{1}{2} \|\mathbf{X} - \mathbf{U}\|_F^2 + \gamma \left[\sum_{i < j} w_{ij} \|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \sum_{k < l} \tilde{w}_{kl} \|\mathbf{U}_{k \cdot} - \mathbf{U}_{l \cdot}\|_2 \right]$$

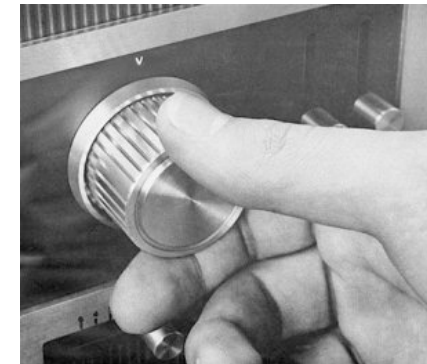
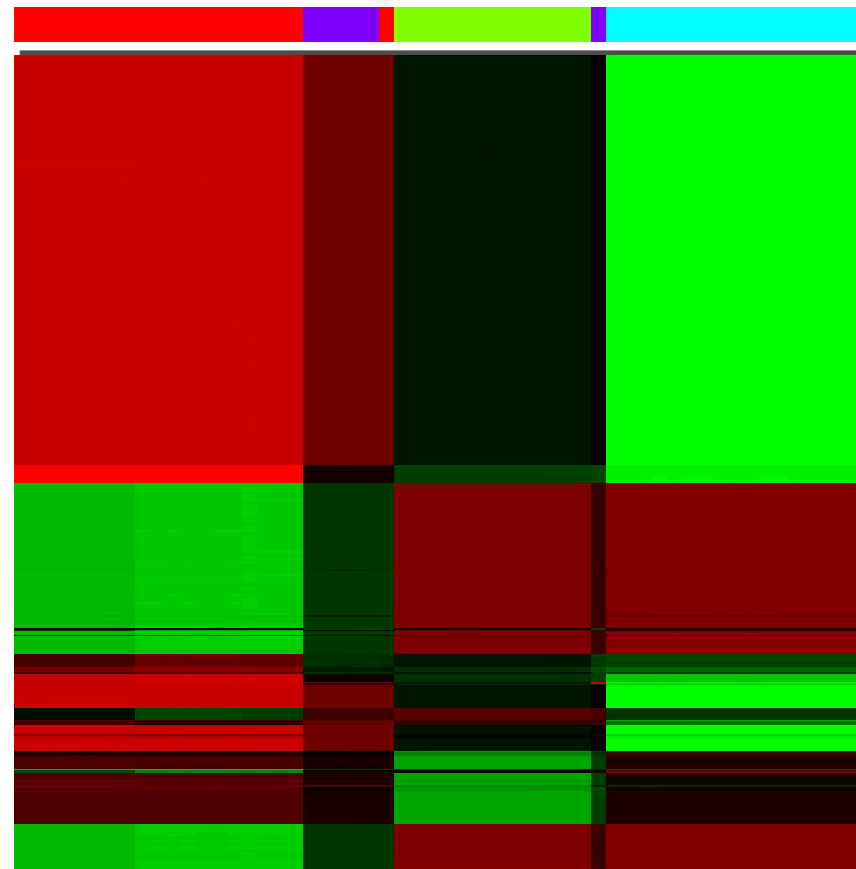
Biclustering path



γ

$$\text{minimize } \frac{1}{2} \|\mathbf{X} - \mathbf{U}\|_F^2 + \gamma \left[\sum_{i < j} w_{ij} \|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \sum_{k < l} \tilde{w}_{kl} \|\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}\|_2 \right]$$

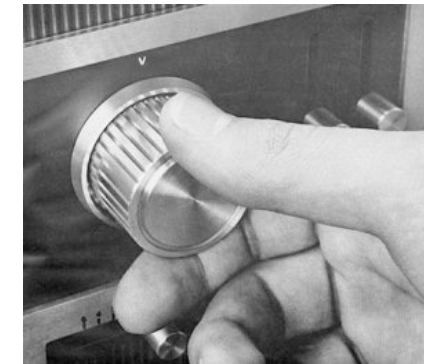
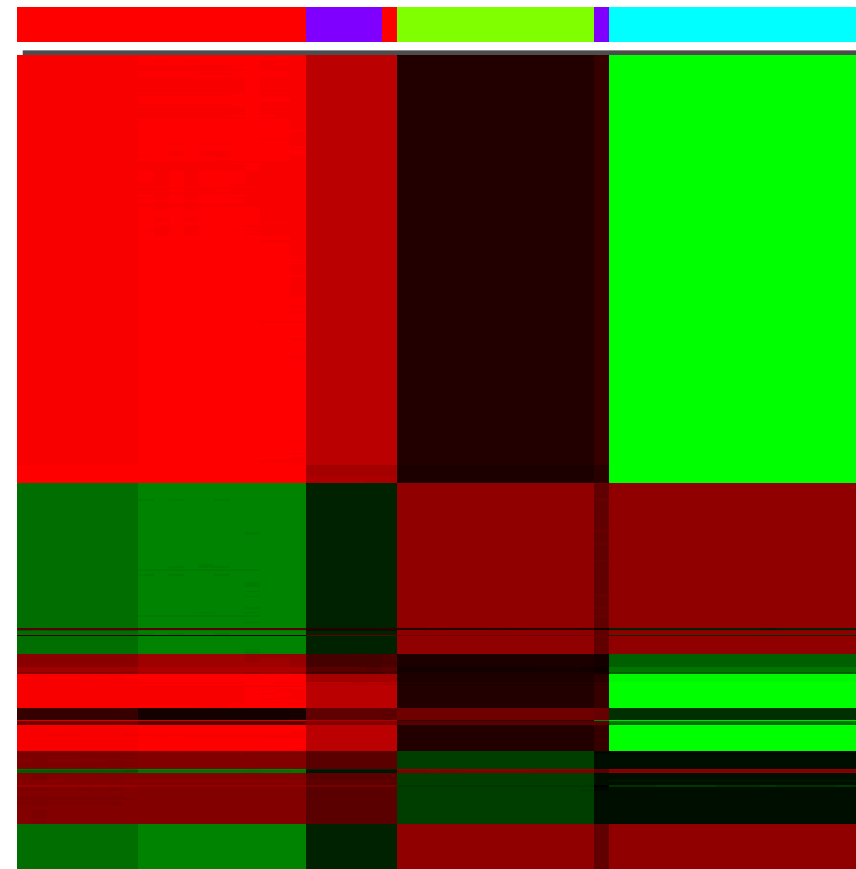
Biclustering path



γ

$$\text{minimize } \frac{1}{2} \|\mathbf{X} - \mathbf{U}\|_F^2 + \gamma \left[\sum_{i < j} w_{ij} \|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \sum_{k < l} \tilde{w}_{kl} \|\mathbf{U}_{k \cdot} - \mathbf{U}_{l \cdot}\|_2 \right]$$

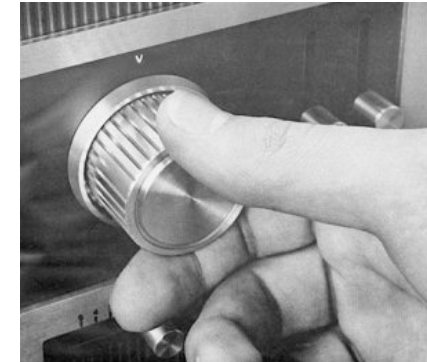
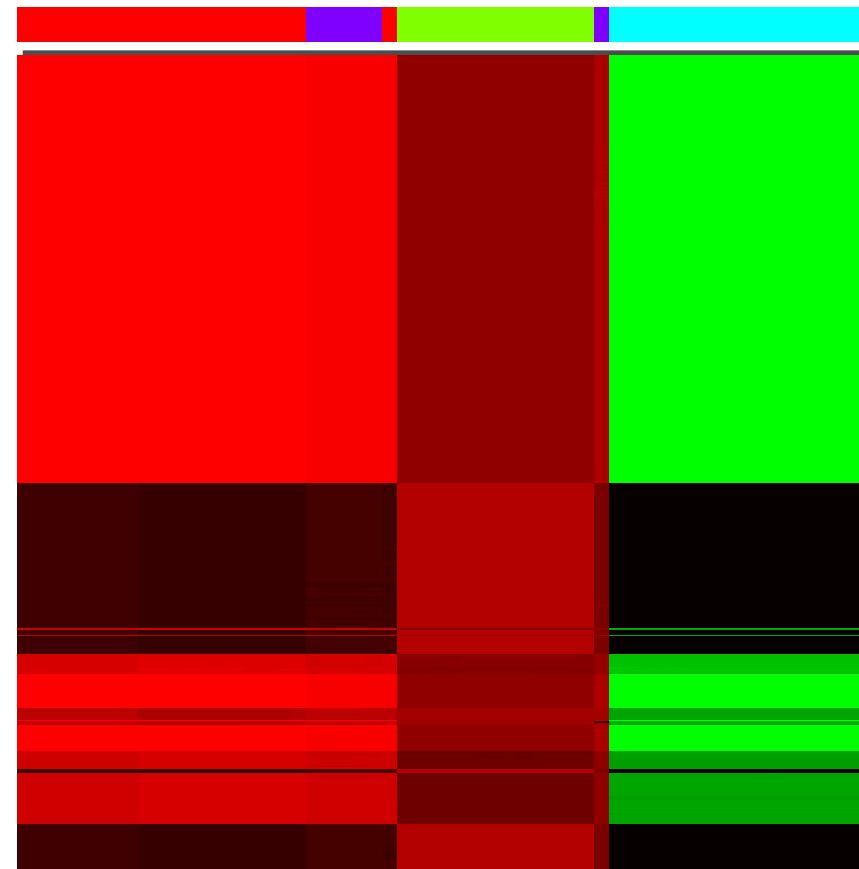
Biclustering path



γ

$$\text{minimize } \frac{1}{2} \|\mathbf{X} - \mathbf{U}\|_F^2 + \gamma \left[\sum_{i < j} w_{ij} \|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \sum_{k < l} \tilde{w}_{kl} \|\mathbf{U}_{k \cdot} - \mathbf{U}_{l \cdot}\|_2 \right]$$

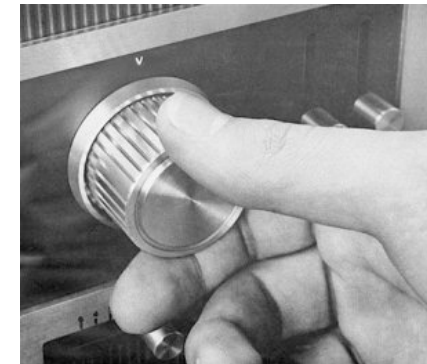
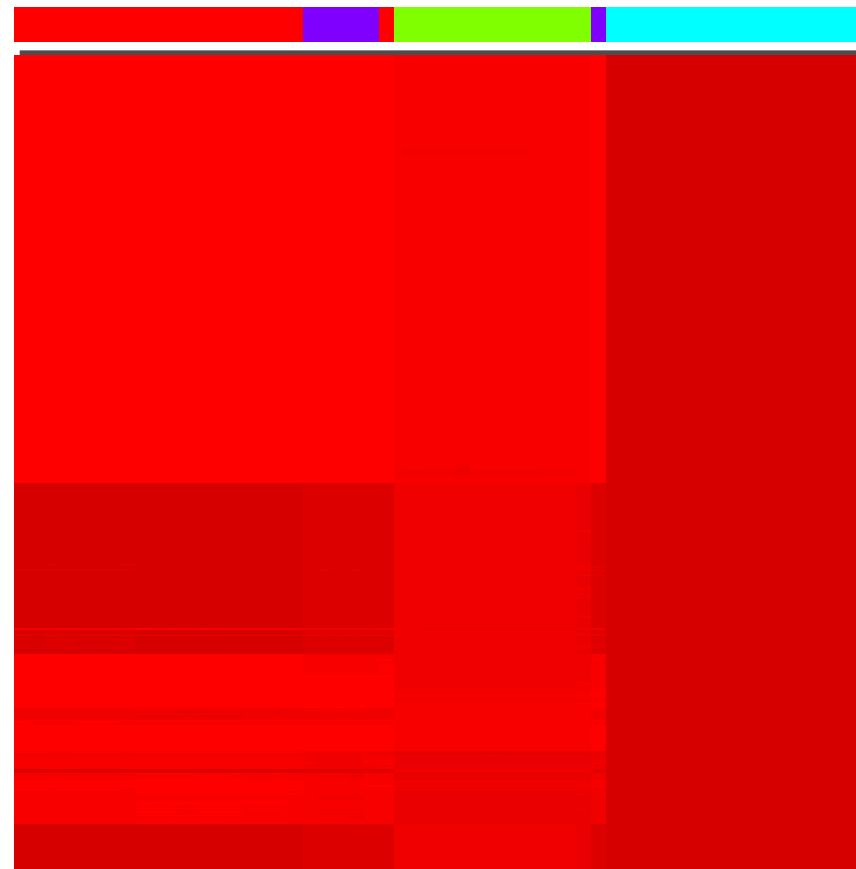
Biclustering path



γ

$$\text{minimize } \frac{1}{2} \|\mathbf{X} - \mathbf{U}\|_F^2 + \gamma \left[\sum_{i < j} w_{ij} \|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \sum_{k < l} \tilde{w}_{kl} \|\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}\|_2 \right]$$

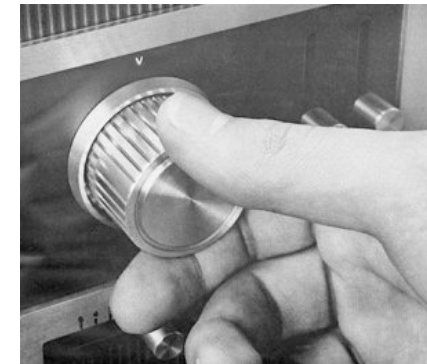
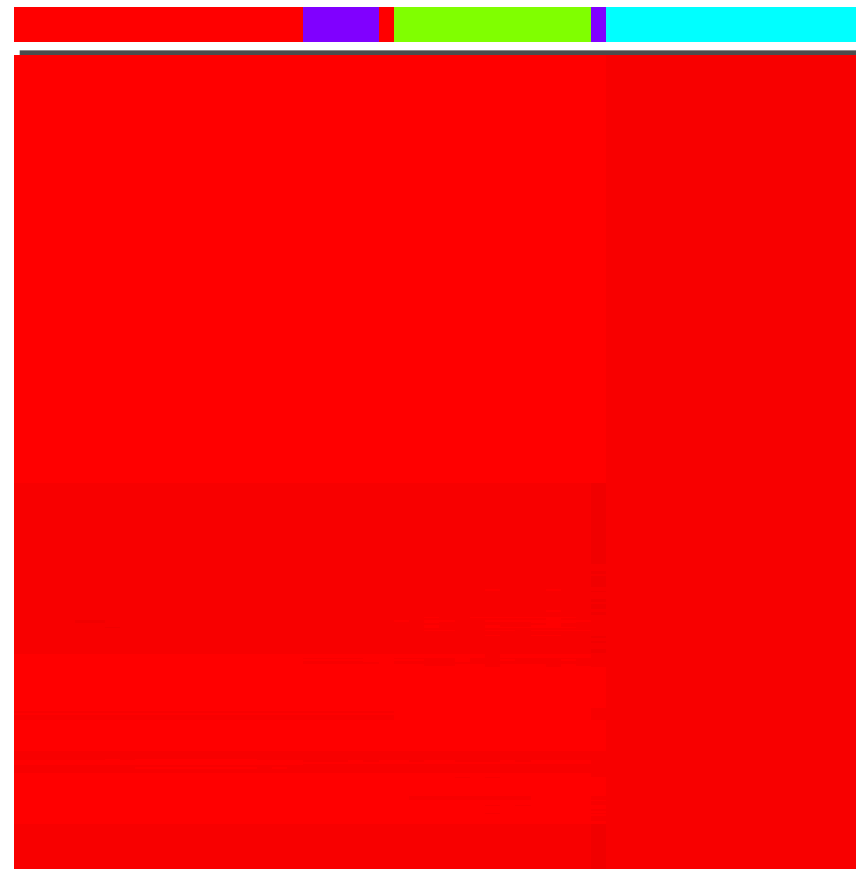
Biclustering path



γ

$$\text{minimize } \frac{1}{2} \|\mathbf{X} - \mathbf{U}\|_F^2 + \gamma \left[\sum_{i < j} w_{ij} \|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \sum_{k < l} \tilde{w}_{kl} \|\mathbf{U}_{k \cdot} - \mathbf{U}_{l \cdot}\|_2 \right]$$

Biclustering path



γ

$$\text{minimize } \frac{1}{2} \|\mathbf{X} - \mathbf{U}\|_F^2 + \gamma \left[\sum_{i < j} w_{ij} \|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \sum_{k < l} \tilde{w}_{kl} \|\mathbf{U}_{k \cdot} - \mathbf{U}_{l \cdot}\|_2 \right]$$

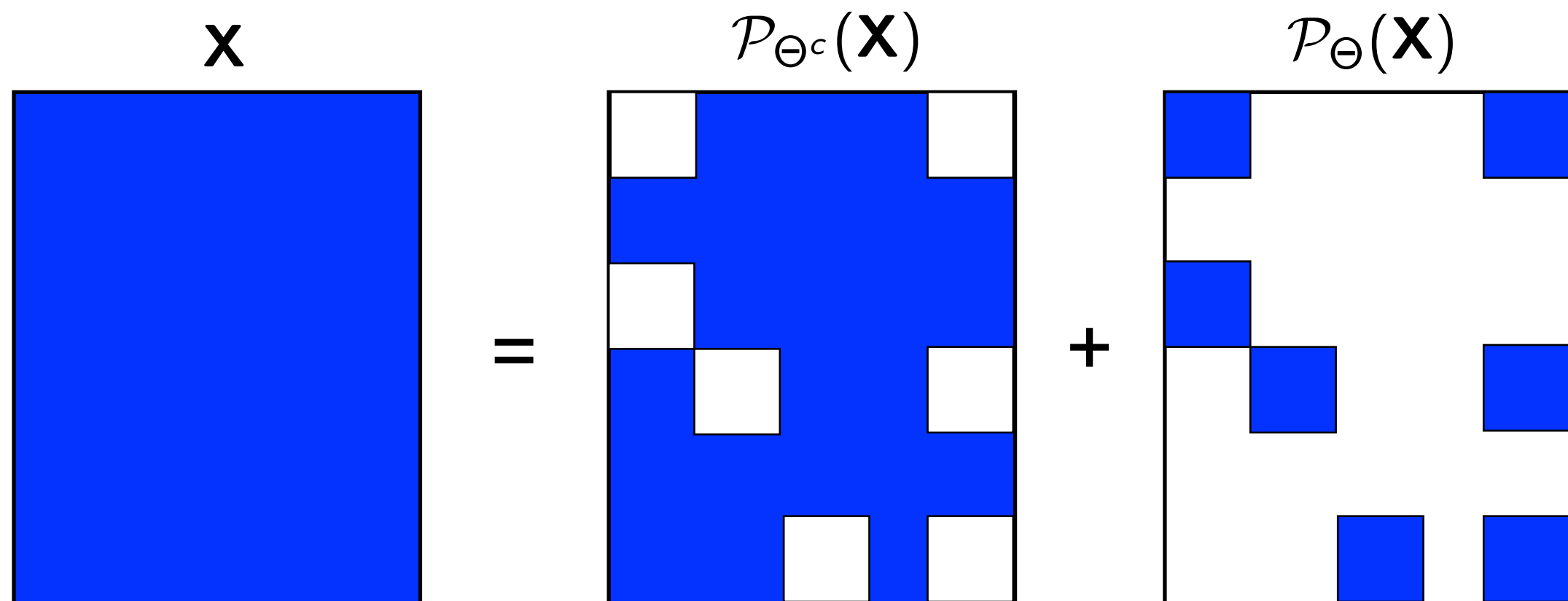
Model Selection

Validation Scheme (Wold, 1978)

- Randomly select a hold out set ($\sim 10\%$ of all entries)

$$\Theta \subset \{1, \dots, p\} \times \{1, \dots, n\}$$

$$\square = 0$$

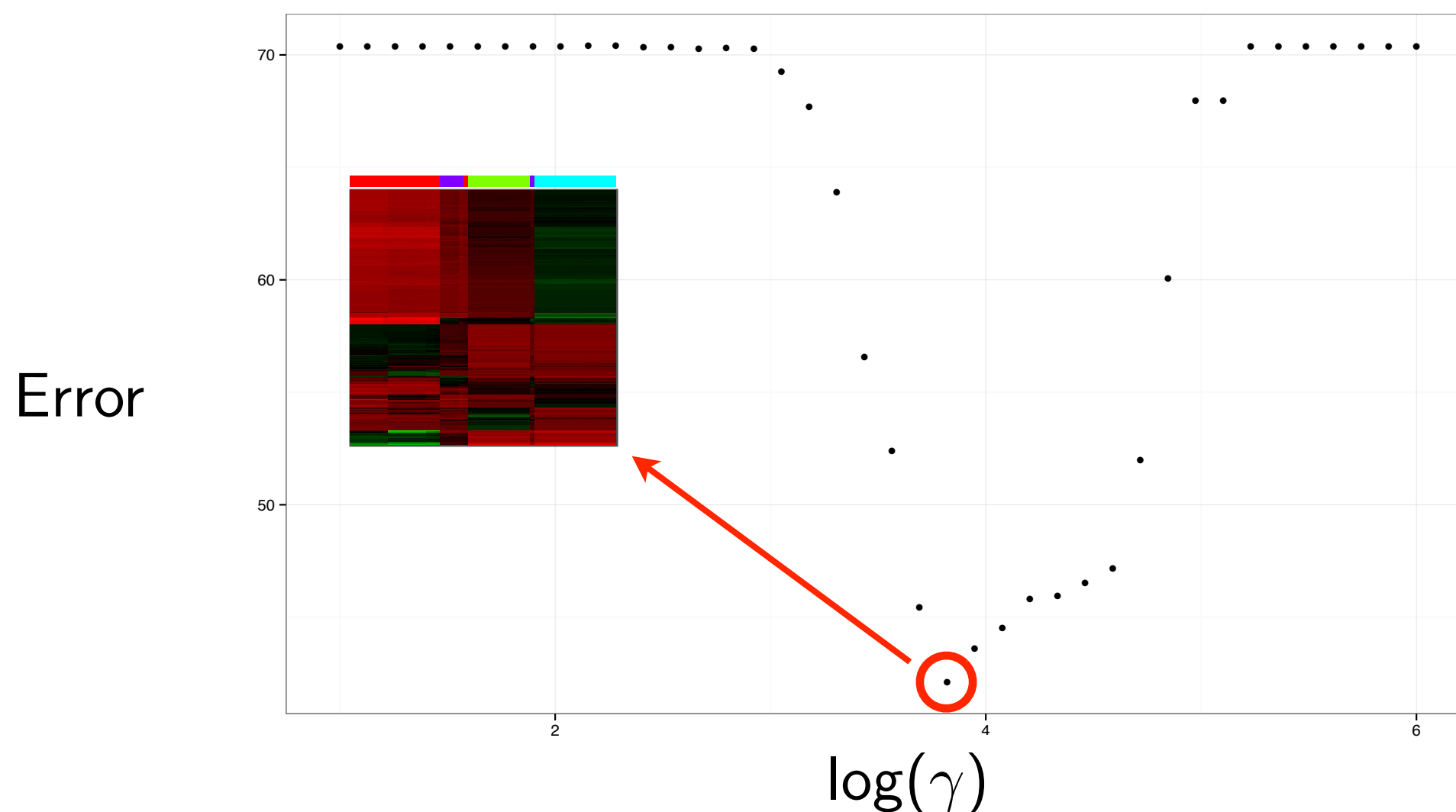


Model Selection

- Solve the matrix completion problem for a sequence of candidate γ_m

$$\mathbf{U}_{\gamma_m}^* = \arg \min_{\mathbf{U}} \frac{1}{2} \|\mathcal{P}_{\Theta^c}(\mathbf{X}) - \mathcal{P}_{\Theta^c}(\mathbf{U})\|_F^2 + \gamma_m J(\mathbf{U})$$

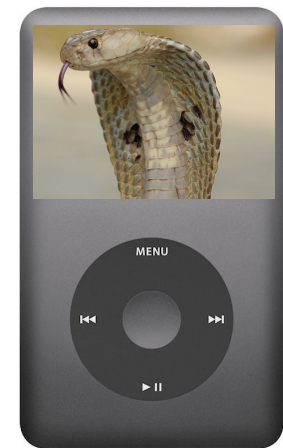
- Pick γ_m that minimizes prediction error on Θ



Matrix Completion with COBRA-POD

COBRA with **P**artially **O**bserved **D**ata

- Fill in missing entries using last iterate
- Use COBRA to get next iterate
- Repeat



<http://pitchfork.com/news/56629-apple-discontinues-ipod-classic>

Algorithm 1 COBRA-POD

- 1: Initialize $\mathbf{U}^{(0)}$.
 - 2: **repeat**
 - 3: $\mathbf{M} \leftarrow \mathcal{P}_{\Theta^c}(\mathbf{X}) + \mathcal{P}_{\Theta}(\mathbf{U}^{(n)})$
 - 4: $\mathbf{U}^{(n+1)} \leftarrow \text{COBRA}(\mathbf{M})$
 - 5: **until** convergence
-

- This is a majorization-minimization (MM) algorithm
- Same majorization used in Soft-Impute: Mazumder et al. (2010)

Convergence of COBRA-POD

Proposition (Simple Zangwill / Meyer): Let f be a continuous function on a domain $\mathcal{D}$, and let ψ be a continuous iterative map from $\mathcal{D}$ into $\mathcal{D}$ such that $f(\psi(x)) < f(x)$ for all $x \in \mathcal{D}$ with $\psi(x) \neq x$. Suppose there is an x_0 such that the set $\mathcal{L}_f(x_0) \equiv \{x \in \mathcal{D} : f(x) \leq f(x_0)\}$ is compact. Define $x_{n+1} = \psi(x_n)$ for $n \in \mathbb{N}$. Then

1. the sequence of iterates $\{x_n\}$ has at least one limit point and all its limit points are fixed points of ψ .
2. the distance between successive iterates converges to 0, i.e., $\|x_{n+1} - x_n\| \rightarrow 0$.

$$f(\mathbf{U}) = \frac{1}{2} \|\mathcal{P}_{\Theta^c}(\mathbf{X}) - \mathcal{P}_{\Theta^c}(\mathbf{U})\|_{\mathbb{F}}^2 + \gamma_m J(\mathbf{U})$$

Clean up workflow:

- Fix- ψ are optima of f (modified objective)
- Show that there is only one optimum?

Convergence of COBRA-POD

Theorem 5.14 (Bauschke & Combettes) Let $\psi : \mathcal{H} \mapsto \mathcal{H}$ be a nonexpansive operator such that $\text{Fix-}\psi$ is not empty. Let $x_0 \in \mathcal{H}$ and $(\forall n \in \mathbb{N}) x_{n+1} = \psi(x_n)$. Suppose that $\psi(x_n) - x_n \rightarrow 0$. Then $(x_n)_{n \in \mathbb{N}}$ converges to a point in $\text{Fix-}\psi$.

Proposition (Simple Zangwill / Meyer): Let f be a continuous function on a domain $\mathcal{D}$, and let ψ be a continuous iterative map from $\mathcal{D}$ into $\mathcal{D}$ such that $f(\psi(x)) < f(x)$ for all $x \in \mathcal{D}$ with $\psi(x) \neq x$. Suppose there is an x_0 such that the set $\mathcal{L}_f(x_0) \equiv \{x \in \mathcal{D} : f(x) \leq f(x_0)\}$ is compact. Define $x_{n+1} = \psi(x_n)$ for $n \in \mathbb{N}$. Then

1. the sequence of iterates $\{x_n\}$ has at least one limit point and all its limit points are fixed points of ψ .
2. the distance between successive iterates converges to 0, i.e., $\|x_{n+1} - x_n\| \rightarrow 0$.

upshot: If MM algorithm map is nonexpansive, MM iterates converge to fixed point.

Convergence of COBRA-POD

Algorithm 2 COBRA-POD

- 1: Initialize $\mathbf{U}^{(0)}$.
 - 2: **repeat**
 - 3: $\mathbf{M} \leftarrow \mathcal{P}_{\Theta^c}(\mathbf{X}) + \mathcal{P}_{\Theta}(\mathbf{U}^{(n)})$
 - 4: $\mathbf{U}^{(n+1)} \leftarrow \text{COBRA}(\mathbf{M})$
 - 5: **until** convergence
-

$$\text{minimize } \frac{1}{2} \|\mathbf{X} - \mathbf{U}\|_{\text{F}}^2 + \gamma \left[\sum_{i < j} w_{ij} \|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \sum_{k < l} \tilde{w}_{kl} \|\mathbf{U}_{k \cdot} - \mathbf{U}_{l \cdot}\|_2 \right]$$

- COBRA is a proximal map
- Imputation step is also non-expansive

$$\|\mathcal{P}_{\Theta^c}(\mathbf{X}) + \mathcal{P}_{\Theta}(\mathbf{U}) - \mathcal{P}_{\Theta^c}(\mathbf{X}) - \mathcal{P}_{\Theta}(\mathbf{V})\| \leq \|\mathcal{P}_{\Theta}(\mathbf{U}) - \mathcal{P}_{\Theta}(\mathbf{V})\| \leq \|\mathbf{U} - \mathbf{V}\|$$

Case Study 2: Robust Low-Rank Matrix Approximation

$$\text{minimize } \sum_{ij} \rho(x_{ij} - u_{ij}) + \gamma \|\mathbf{U}\|_*$$

where

$$\rho(t) = -\exp\left[-\frac{\alpha t^2}{2}\right]$$

- ρ is Lipschitz differentiable

$$g(\mathbf{U} \mid \tilde{\mathbf{U}}) = \frac{1}{2} \|\varphi(\tilde{\mathbf{U}}) - \mathbf{U}\|_F^2 + \gamma \|\mathbf{U}\|_*$$

$$[\varphi(\tilde{\mathbf{U}})]_{ij} = \varphi_{ij}(\tilde{\mathbf{U}}) = w_{ij}(\tilde{\mathbf{U}})x_{ij} + (1 - w_{ij}(\tilde{\mathbf{U}}))\tilde{u}_{ij},$$

where

$$w_{ij}(\tilde{\mathbf{U}}) = \exp\left[-\frac{\alpha}{2}(x_{ij} - \tilde{u}_{ij})^2\right].$$

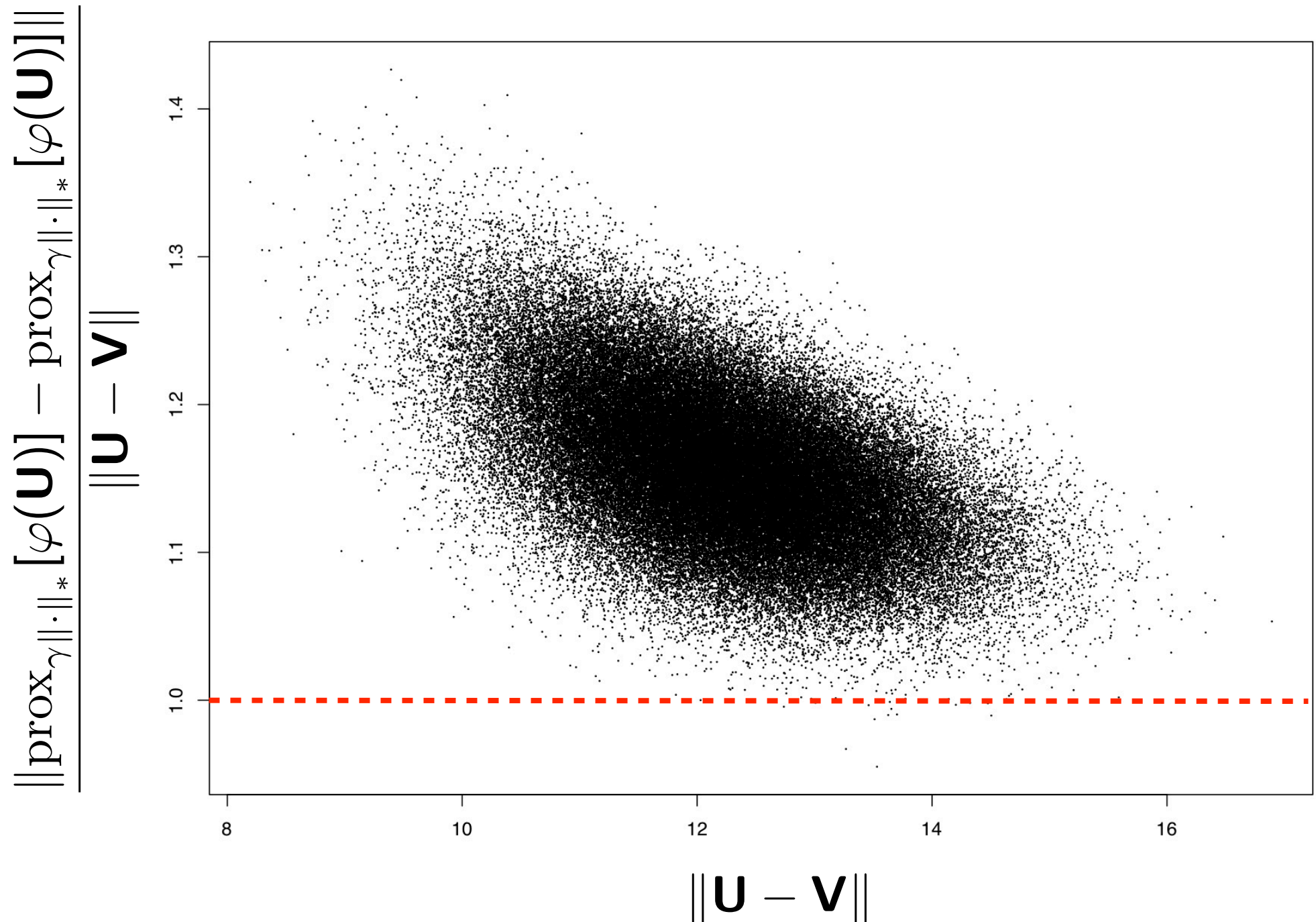
Convergence Robust Low-Rank Approximation

Algorithm 3 Robust Low-Rank

```
1: Initialize  $\mathbf{U}^{(0)}$ .  
2: repeat  
3:    $\mathbf{M} \leftarrow \varphi(\mathbf{U}^{(n)})$   
4:    $\mathbf{U}^{(n+1)} \leftarrow \text{prox}_{\gamma\|\cdot\|_*}(\mathbf{M})$   
5: until convergence
```

- Can apply Zangwill / Meyer in this case; fixed points are stationary points - directional derivative is non-negative in all directions.
- But can't apply Theorem 5.12
- φ is Lipschitz continuous with constant $1 + 2e^{-3/2} \approx 1.4462603$.

Convergence Robust Low-Rank Approximation



Open Questions

1. Is there a way to modify the MM algorithm map to enforce non-expansivity that won't alter the fixed point set?
2. Are there additional regularity conditions (Lipschitz smoothness of MM algorithm map?) that could ensure convergence?