# Large Scale Optimization for Machine Learning

Meisam Razaviyayn

Lecture 21

razaviya@usc.edu

# Announcements:

- Midterm exams
  - Return it next Tuesday

# Non-smooth Objective Function

- Sub-gradient
  - Typically slow and no good termination criteria (other than cross validation)

- Proximal Gradient
  - Fast assuming each iteration is easy

- Block Coordinate Descent
  - Also helpful for exploiting multi-block structure

- Alternating Direction Method of Multipliers (ADMM)
  - Will be covered later

# Multi-Block Structure and BCD Method

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}_1, \ldots, \mathbf{x}_m) \\ \text{s.t.} \quad & \mathbf{x}_i \in \mathcal{X}_i, \ \forall i \end{aligned}$$

**Block Coordinate Descent (BCD) Method:**

At iteration *r*, choose an index *i* and

$$\mathbf{x}_i^{r+1} = \arg\min_{\mathbf{x}_i \in \mathcal{X}_i} \quad f(\mathbf{x}_1^r, \ldots, \mathbf{x}_{i-1}^r, \mathbf{x}_i, \mathbf{x}_{i+1}^r, \ldots, \mathbf{x}_m^r)$$

$$\mathbf{x}_k^{r+1} = \mathbf{x}_k^r, \quad \forall k \neq i$$

**Choice of index *i*:** Cyclic, randomized, Greedy

Simple and scalable: Lasso example

Very different than previous incremental GD, SGD,…

**Geometric Interpretation**

# Convergence of BCD Method

**Proposition 2.7.1: (Convergence of Block Coordinate Descent)** Suppose that $f$ is continuously differentiable over the set $X$ of Eq. (2.111). Furthermore, suppose that for each $x = (x_1, \ldots, x_m) \in X$ and $i$,

$$f(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_m)$$

viewed as a function of $\xi$, attains a unique minimum $\bar{\xi}$ over $X_i$, and is monotonically nonincreasing in the interval from $x_i$ to $\bar{\xi}$. Let $\{x^k\}$ be the sequence generated by the block coordinate descent method (2.112). Then, every limit point of $\{x^k\}$ is a stationary point.

"Nonlinear programming", D.P. Bertsekas **for cyclic update rule**

**Assumptions:**

- Separable Constraints
- Differentiable/smooth objective
- Unique minimizer at each step

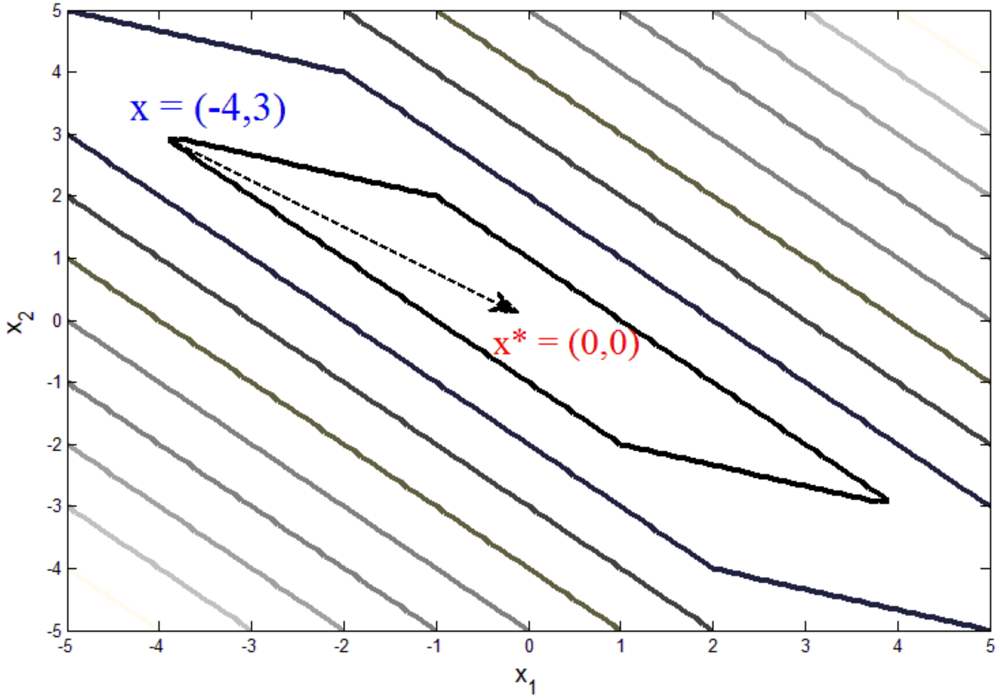
**Proof?**

**Necessary assumptions?**

# Necessity of Smoothness Assumption

$$f(\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|_1, \quad \mathbf{A} = [3 \;\; 4; 2 \;\; 1];$$

**Not "Regular"** $\longrightarrow$




Function $f(\cdot)$ is **regular** at point $\mathbf{z}$ if

$$f'(\mathbf{z}; (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{d}_k, \mathbf{0}, \ldots, \mathbf{0})) \geq 0, \; \forall \; k, \; \forall \; \mathbf{d}_k \quad \Rightarrow \quad f'(\mathbf{z}; \mathbf{d}) \geq 0, \text{ for all } \mathbf{d} = (\mathbf{d}_1, \ldots, \mathbf{d}_m)$$

$$f(\mathbf{x}) = \overset{\text{smooth}}{g(\mathbf{x})} + \sum_k h_k(\mathbf{x}_k) \Rightarrow \quad f \text{ is regular}$$

Examples: Lasso

# BCD and Non-smooth Objective

**Theorem [Tseng 2001]**

Assume

1) Feasible set is compact.
2) The uniqueness of minimizer at each step.
3) Separable constraint
4) **Regular objective function**

⇒ Every limit point of the iterates is a stationary point

True for cyclic/randomized/greedy rule

Definition of stationarity for nonsmooth?

Rate of convergence of BCD:

- Similar to GD: sublinear for general convex and linear for strongly convex
- Same results can be shown in most of the non-smooth popular objectives

# Uniqueness of the Minimizer

**[Michael J. D. Powell 1973]**

$$(x-c)_+^2 = \begin{cases} 0, & \text{if } x \le c \\ (x-c)^2, & \text{if } x \ge c \end{cases}$$

$$f(x,y,z) = -xy - yz - xz + (x-1)_+^2 + (-x-1)_+^2 + (y-1)_+^2 + (-y-1)_+^2 + (z-1)_+^2 + (-z-1)_+^2$$

$$\begin{array}{ccc}
(-1-\epsilon, 1+\epsilon/2, -1-\epsilon/4) & \longrightarrow & (1+\epsilon/8, 1+\epsilon/2, -1-\epsilon/4) \\
 & & \downarrow \\
 & & (1+\epsilon/8, -1-\epsilon/16, -1-\epsilon/4) \\
 & & \downarrow \\
(-1-\epsilon/64, 1+\epsilon/128, -1-\epsilon/256) & & (1+\epsilon/8, -1-\epsilon/16, 1+\epsilon/32) \\
\uparrow & & \downarrow \\
(-1-\epsilon/64, 1+\epsilon/128, 1+\epsilon/32) & \longleftarrow & (-1-\epsilon/64, -1-\epsilon/16, 1+\epsilon/32)
\end{array}$$

# Uniqueness of the Minimizer

**Tensor PARAFAC Decomposition**

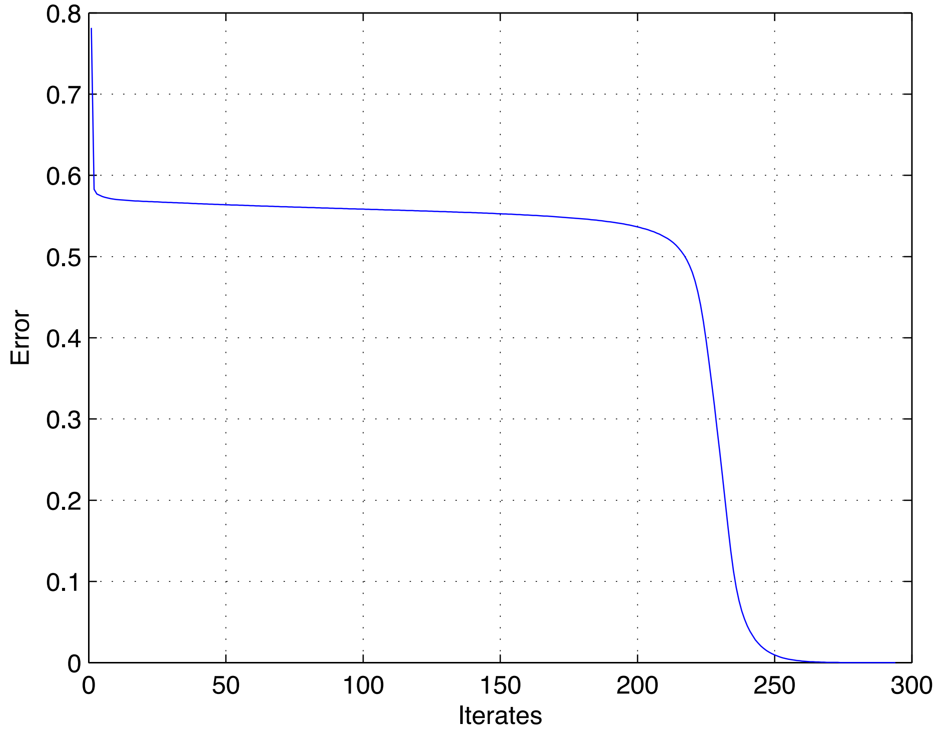
$$\mathfrak{X} \in \mathbb{R}^{I \times J \times K}$$

$$\mathfrak{X} = \sum_{\ell=1}^{L} \mathbf{a}_\ell \circ \mathbf{b}_\ell \circ \mathbf{c}_\ell$$

NP-hard [Hastad 1990]

[Carroll 1970], [Harshman1970]: Alternating Least Squares

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \quad \|\mathfrak{X} - \sum_{\ell=1}^{L} \mathbf{a}_\ell \circ \mathbf{b}_\ell \circ \mathbf{c}_\ell\|^2$$




“Swamp” effect

# BCD Limitations

$$\min_{\mathbf{x}} \quad f(\mathbf{x}_1, \ldots, \mathbf{x}_m)$$
$$\text{s.t.} \quad \mathbf{x}_i \in \mathcal{X}_i, \ \forall i$$

- Uniqueness of minimizer
- Each sub-problem needs to be easily solvable

# BCD Limitations

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}_1, \ldots, \mathbf{x}_m) \\ \text{s.t.} \quad & \mathbf{x}_i \in \mathcal{X}_i, \ \forall i \end{aligned}$$

- Uniqueness of minimizer
- Each sub-problem needs to be easily solvable

**Popular Solution: Inexact BCD**

At iteration *r*, choose an index *i* and

$$\mathbf{x}_i^{r+1} = \arg \min_{\mathbf{x}_i \in \mathcal{X}_i} \quad u_i(\mathbf{x}_i, \mathbf{x}^r)$$

$$\mathbf{x}_k^{r+1} = \mathbf{x}_k^r, \ \ \forall k \neq i$$

# BCD Limitations

$$\min_{\mathbf{x}} \quad f(\mathbf{x}_1, \ldots, \mathbf{x}_m)$$
$$\text{s.t.} \quad \mathbf{x}_i \in \mathcal{X}_i, \; \forall i$$

- Uniqueness of minimizer
- Each sub-problem needs to be easily solvable

**Popular Solution: Inexact BCD**

At iteration *r*, choose an index *i* and

$$\mathbf{x}_i^{r+1} = \arg\min_{\mathbf{x}_i \in \mathcal{X}_i} \; u_i(\mathbf{x}_i, \mathbf{x}^r)$$
$$\mathbf{x}_k^{r+1} = \mathbf{x}_k^r, \quad \forall k \neq i$$

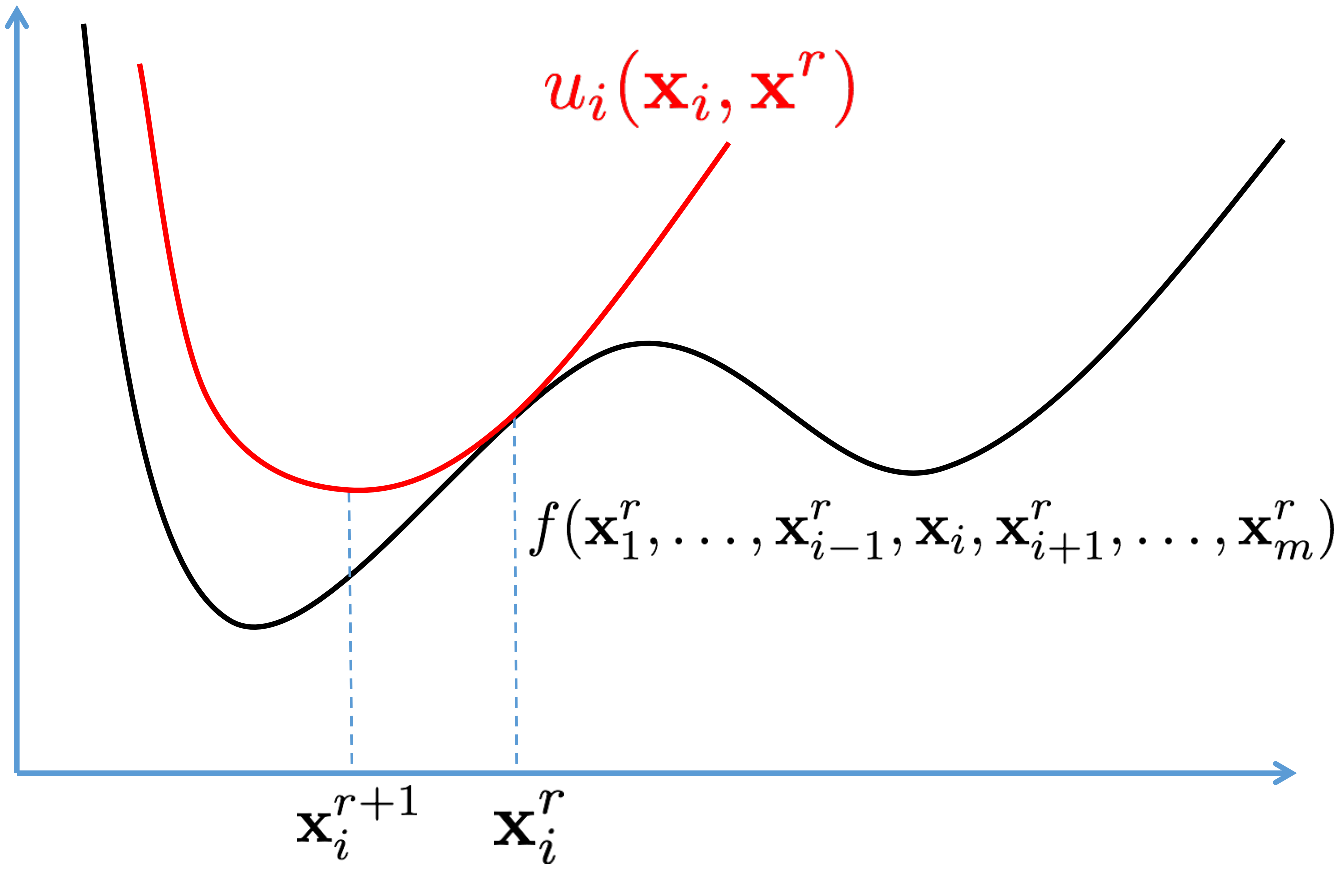
Local approximation of the objective function

Block successive upper-bound minimization, block successive convex approximation, convex-concave procedure, majorization minimization, dc-programming, BCGD,…

# Idea of Block Successive Upper-bound Minimization

Global upper-bound:

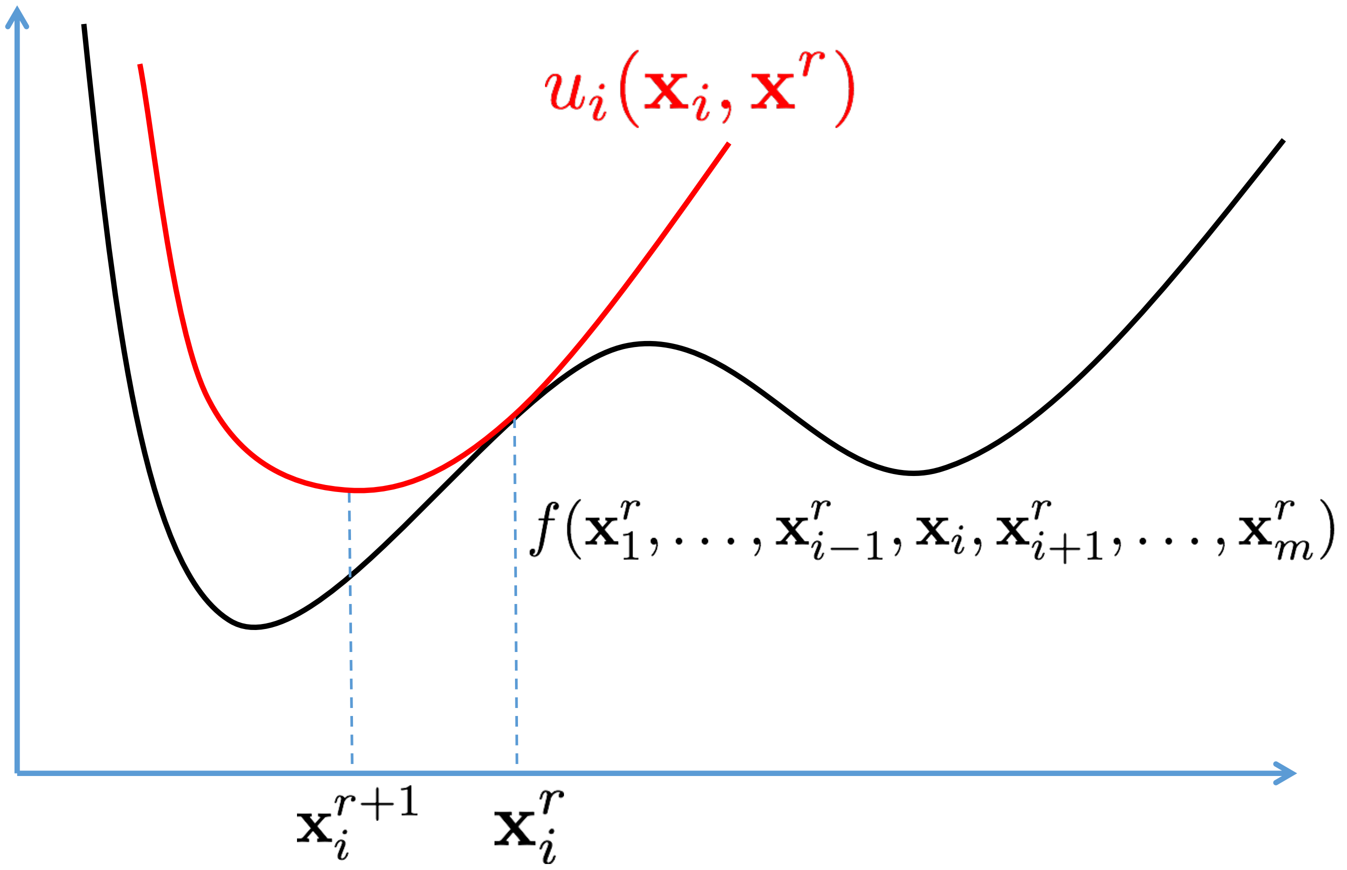
$u_i(\mathbf{x}_i, \mathbf{x}^r) \geq f(\mathbf{x}_1^r, \ldots, \mathbf{x}_{i-1}^r, \mathbf{x}_i, \mathbf{x}_{i+1}^r, \ldots, \mathbf{x}_m^r)$



# Idea of Block Successive Upper-bound Minimization

Global upper-bound:

$u_i(\mathbf{x}_i, \mathbf{x}^r) \geq f(\mathbf{x}_1^r, \ldots, \mathbf{x}_{i-1}^r, \mathbf{x}_i, \mathbf{x}_{i+1}^r, \ldots, \mathbf{x}_m^r)$




Locally tight:

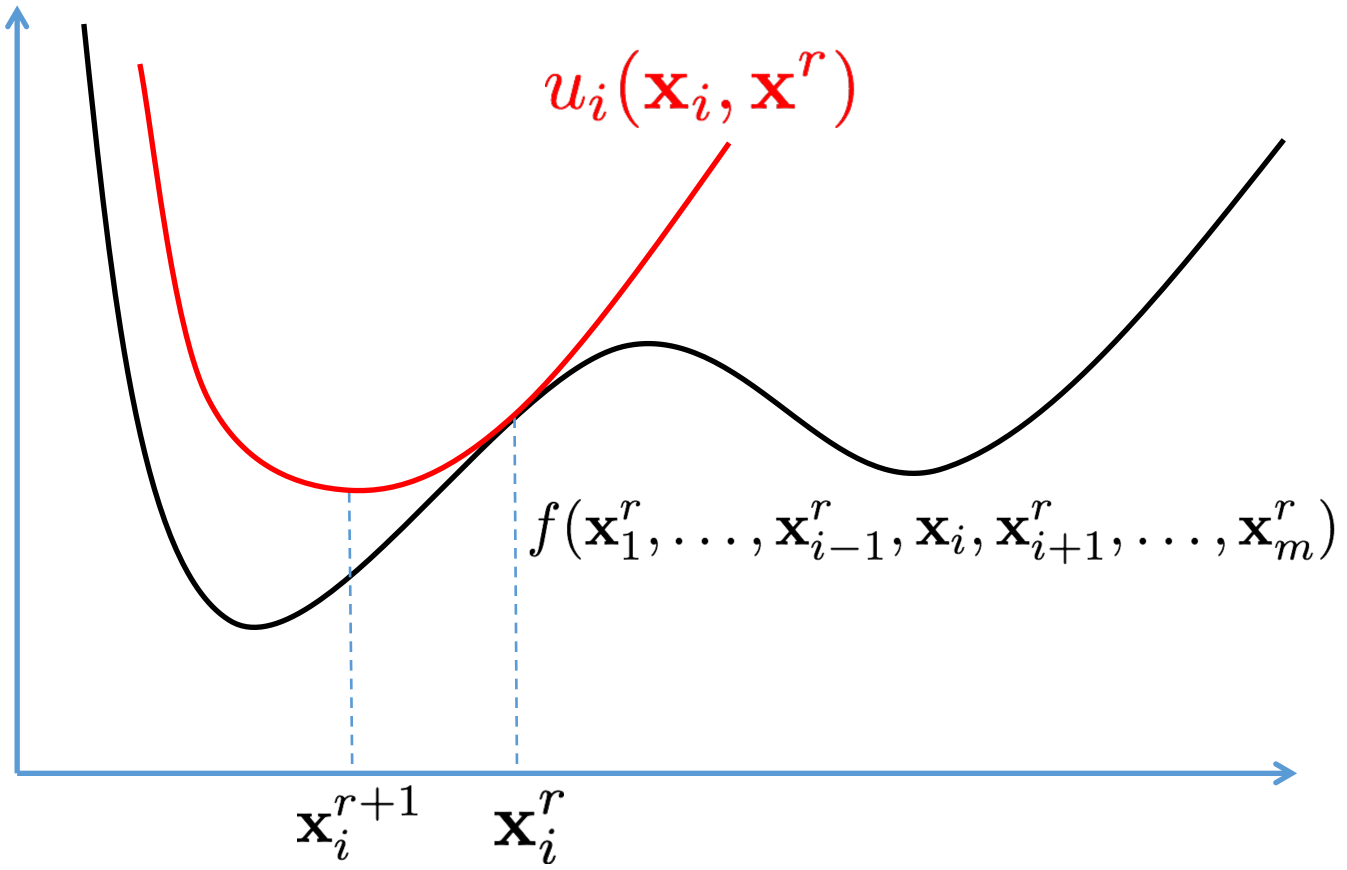
$$u_i(\mathbf{x}_i^r, \mathbf{x}^r) = f(\mathbf{x}^r)$$

$$u'(\mathbf{x}_i, \mathbf{x}^r; \mathbf{d}_i)\Big|_{\mathbf{x}_i = \mathbf{x}_i^r} = f'(\mathbf{x}^r; \mathbf{d}), \ \ \forall \mathbf{d} = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{d}_i, \mathbf{0}, \ldots, \mathbf{0})$$

# Idea of Block Successive Upper-bound Minimization

Global upper-bound:

$u_i(\mathbf{x}_i, \mathbf{x}^r) \geq f(\mathbf{x}_1^r, \ldots, \mathbf{x}_{i-1}^r, \mathbf{x}_i, \mathbf{x}_{i+1}^r, \ldots, \mathbf{x}_m^r)$




Locally tight:

$$u_i(\mathbf{x}_i^r, \mathbf{x}^r) = f(\mathbf{x}^r)$$

$$u'(\mathbf{x}_i, \mathbf{x}^r; \mathbf{d}_i)\Big|_{\mathbf{x}_i = \mathbf{x}_i^r} = f'(\mathbf{x}^r; \mathbf{d}), \;\; \forall \mathbf{d} = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{d}_i, \mathbf{0}, \ldots, \mathbf{0})$$

Monotone Algorithm

Every limit point is a stationary point

# Example 1: Block Coordinate (Proximal) Gradient Descent

Smooth Scenario: $\mathbf{x}_i^{r+1} = \mathbf{x}_i^r - \alpha^r \nabla_{\mathbf{x}_i} f(\mathbf{x}^r)$

# Example 1: Block Coordinate (Proximal) Gradient Descent

Smooth Scenario: $\mathbf{x}_i^{r+1} = \mathbf{x}_i^r - \alpha^r \nabla_{\mathbf{x}_i} f(\mathbf{x}^r)$

Non-smooth Scenario: $\mathbf{x}_i^{r+1} = \mathbf{x}_i^r - \alpha^r \widetilde{\nabla}_{\mathbf{x}_i} f(\mathbf{x}^r; \alpha^r)$

# Example 1: Block Coordinate (Proximal) Gradient Descent

Smooth Scenario: $\mathbf{x}_i^{r+1} = \mathbf{x}_i^r - \alpha^r \nabla_{\mathbf{x}_i} f(\mathbf{x}^r)$

Non-smooth Scenario: $\mathbf{x}_i^{r+1} = \mathbf{x}_i^r - \alpha^r \widetilde{\nabla}_{\mathbf{x}_i} f(\mathbf{x}^r; \alpha^r)$

Using Bregman divergence

# Example 1: Block Coordinate (Proximal) Gradient Descent

Smooth Scenario: $\mathbf{x}_i^{r+1} = \mathbf{x}_i^r - \alpha^r \nabla_{\mathbf{x}_i} f(\mathbf{x}^r)$

Non-smooth Scenario: $\mathbf{x}_i^{r+1} = \mathbf{x}_i^r - \alpha^r \widetilde{\nabla}_{\mathbf{x}_i} f(\mathbf{x}^r; \alpha^r)$

Using Bregman divergence

Alternating Proximal Minimization:

$$\mathbf{x}_i^{r+1} = \arg\min_{\mathbf{x}_i} \; f(\mathbf{x}_1^r, \ldots, \mathbf{x}_{i-1}^r, \mathbf{x}_i, \mathbf{x}_{i+1}^r, \ldots, \mathbf{x}_m^r) + \frac{\mu}{2}\|\mathbf{x}_i - \mathbf{x}_i^r\|_2^2$$

# Example 2: Expectation Maximization Algorithm

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \; \ln p(\mathbf{w}|\boldsymbol{\theta})$$

# Example 2: Expectation Maximization Algorithm

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \ \ln p(\mathbf{w}|\boldsymbol{\theta})$$

- E-Step: Calculate $g(\boldsymbol{\theta}, \boldsymbol{\theta}^r) \triangleq \mathbb{E}_{\mathbf{z}|\mathbf{w},\boldsymbol{\theta}^r}\{\ln p(\mathbf{w}, \mathbf{z}|\boldsymbol{\theta})\}$
- M-Step: $\boldsymbol{\theta}^{r+1} = \arg\max_{\boldsymbol{\theta}} \ g(\boldsymbol{\theta}, \boldsymbol{\theta}^r)$

# Example 2: Expectation Maximization Algorithm

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \ \ln p(\mathbf{w}|\boldsymbol{\theta})$$

- E-Step: Calculate $g(\boldsymbol{\theta}, \boldsymbol{\theta}^r) \triangleq \mathbb{E}_{\mathbf{z}|\mathbf{w},\boldsymbol{\theta}^r}\{\ln p(\mathbf{w}, \mathbf{z}|\boldsymbol{\theta})\}$
- M-Step: $\boldsymbol{\theta}^{r+1} = \arg\max_{\boldsymbol{\theta}} \ g(\boldsymbol{\theta}, \boldsymbol{\theta}^r)$

$$\begin{aligned}
-\ln p(\mathbf{w}|\boldsymbol{\theta}) &= -\ln \ \mathbb{E}_{\mathbf{z}|\boldsymbol{\theta}} \ p(\mathbf{w}|\mathbf{z}, \boldsymbol{\theta}) \\
&= -\ln \ \mathbb{E}_{\mathbf{z}|\boldsymbol{\theta}} \left[\frac{p(\mathbf{z}|\mathbf{w}, \boldsymbol{\theta}^r)p(\mathbf{w}|\mathbf{z}, \boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{w}, \boldsymbol{\theta}^r)}\right] \\
&= -\ln \ \mathbb{E}_{\mathbf{z}|\mathbf{w},\boldsymbol{\theta}^r} \left[\frac{p(\mathbf{z}|\boldsymbol{\theta})p(\mathbf{w}|\mathbf{z}, \boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{w}, \boldsymbol{\theta}^r)}\right] \\
&\leq -\mathbb{E}_{\mathbf{z}|\mathbf{w},\boldsymbol{\theta}^r} \ln \left[\frac{p(\mathbf{z}|\boldsymbol{\theta})p(\mathbf{w}|\mathbf{z}, \boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{w}, \boldsymbol{\theta}^r)}\right] \qquad \text{Jensen's inequality} \\
&= -\mathbb{E}_{\mathbf{z}|\mathbf{w},\boldsymbol{\theta}^r} \ln p(\mathbf{w}, \mathbf{z}|\boldsymbol{\theta}) + \mathbb{E}_{\mathbf{z}|\mathbf{w},\boldsymbol{\theta}^r} \ln p(\mathbf{z}|\mathbf{w}, \boldsymbol{\theta}^r) \triangleq u(\boldsymbol{\theta}, \boldsymbol{\theta}^r)
\end{aligned}$$

# Example 3: Transcript Abundance Estimation

$$
\begin{aligned}
Pr\left(R_1, \ldots, R_N; \rho_1, \ldots, \rho_M\right) &= \prod_{n=1}^{N} Pr\left(R_n; \rho_1 \ldots \rho_M\right) \\
&= \prod_{n=1}^{N} \left( \sum_{m=1}^{M} Pr\left(R_n \mid \text{read } R_n \text{ from sequence } s_m\right) Pr(s_m) \right) \\
&= \prod_{n=1}^{N} \left( \sum_{m=1}^{M} \alpha_{nm} \rho_m \right),
\end{aligned}
$$

# Example 3: Transcript Abundance Estimation

$$
\begin{aligned}
Pr\left(R_1, \ldots, R_N; \rho_1, \ldots, \rho_M\right) &= \prod_{n=1}^{N} Pr\left(R_n; \rho_1 \ldots \rho_M\right) \\
&= \prod_{n=1}^{N} \left( \sum_{m=1}^{M} Pr\left(R_n \mid \text{read } R_n \text{ from sequence } s_m\right) Pr(s_m) \right) \\
&= \prod_{n=1}^{N} \left( \sum_{m=1}^{M} \alpha_{nm} \rho_m \right),
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\rho}_{ML} = \arg\min_{\rho} \quad & -\sum_{n=1}^{N} \log \left( \sum_{m=1}^{M} \alpha_{nm} \rho_m \right) \\
\text{s.t.} \quad & \sum_{m=1}^{M} \rho_m = 1, \quad \text{and} \quad \rho_m \geq 0,\ \forall m = 1, \ldots, M.
\end{aligned}
$$

# Example 3: Transcript Abundance Estimation

$$\widehat{\rho}_{ML} = \arg\min_{\rho} \quad -\sum_{n=1}^{N} \log\left(\sum_{m=1}^{M} \alpha_{nm}\rho_m\right)$$

$$\text{s.t.} \quad \sum_{m=1}^{M} \rho_m = 1, \quad \text{and} \quad \rho_m \geq 0,\ \forall m = 1, \ldots, M.$$

$$\rho^{r+1} = \arg\min_{\rho} \quad -\sum_{n=1}^{N}\left(\sum_{m=1}^{M}\left(\frac{\alpha_{nm}\rho_m^r}{\sum_{m'=1}^{M}\alpha_{nm'}\rho_{m'}^r}\log\left(\frac{\rho_m}{\rho_m^r}\right)\right) + \log\left(\sum_{m=1}^{M}\alpha_{nm}\rho_m^r\right)\right)$$

$$\text{s.t.} \quad \sum_{m=1}^{M} \rho_m = 1, \quad \text{and} \quad \rho_m \geq 0,\ \forall m = 1, \ldots, M.$$

# Example 3: Transcript Abundance Estimation

$$\widehat{\rho}_{ML} = \arg\min_{\rho} \quad -\sum_{n=1}^{N} \log\left(\sum_{m=1}^{M} \alpha_{nm}\rho_m\right)$$

$$\text{s.t.} \quad \sum_{m=1}^{M} \rho_m = 1, \quad \text{and} \quad \rho_m \geq 0,\ \forall m = 1, \ldots, M.$$

$$\rho^{r+1} = \arg\min_{\rho} \quad -\sum_{n=1}^{N}\left(\sum_{m=1}^{M}\left(\frac{\alpha_{nm}\rho_m^r}{\sum_{m'=1}^{M}\alpha_{nm'}\rho_{m'}^r}\log\left(\frac{\rho_m}{\rho_m^r}\right)\right) + \log\left(\sum_{m=1}^{M}\alpha_{nm}\rho_m^r\right)\right)$$

$$\text{s.t.} \quad \sum_{m=1}^{M} \rho_m = 1, \quad \text{and} \quad \rho_m \geq 0,\ \forall m = 1, \ldots, M.$$

$$\rho_m^{r+1} = \frac{1}{N}\sum_{n=1}^{N}\frac{\alpha_{nm}\rho_m^r}{\sum_{m'=1}^{M}\alpha_{nm'}\rho_{m'}^r}, \quad \forall m = 1, \ldots, M,$$

**Closed form update!**