Online Dictionary Learning with Group Structure Inducing Norms

Zoltán Szabó[1], Barnabás Póczos[2], András Lőrincz[1]

[1]Eötvös Loránd University, Budapest, Hungary
[2]Carnegie Mellon University, Pittsburgh, USA

ICML, Structured Sparsity
July 2, 2011

Contents

- Sparse coding, structured sparsity,
- Structured dictionary learning:
  - Our requirements,
  - Cost function,
  - Special cases,
  - Optimization.
- Numerical examples.

Sparse coding

- Observation ($\mathbf{x}$) = linear combination of a *few* vectors (α) from a fixed dictionary ($\mathbf{D}$).
- ℓ_0-norm solution: NP-hard.
- Popular relaxations: ℓ_p ($0 < p \leq 1$) norm.
- Special case: ℓ_1, Lasso problem, efficient algorithms,

$$\min_{\alpha} \left[\frac{1}{2} \|\mathbf{x} - \mathbf{D}\alpha\|_2^2 + \kappa \|\alpha\|_1 \right]. \tag{1}$$

- Disadvantage: prior knowledge on the structure of the hidden code is not taken into account.

Structured sparsity

Different kind of structures (e.g., disjunct groups, trees) on the sparse codes $\Rightarrow$ increased performances in several applications:

- robust CS with substantially fewer observations,
- multi-task learning problems,
- structure learning in graphical models,
- natural language processing,
- fMRI analysis,
- face expression discrimination/recognition.

Structured dictionary learning

- Both dictionary learning
  - (sparse) principal component analysis,
  - (sparse) non-negative matrix factorization (NMF),
  - independent component analysis,
  - independent subspace analysis,

  and structured sparse coding are very popular.
- However, very few works have focused on the *combination* of these two tasks.

Structured dictionary learning: wanted properties

Interested in algorithms with the following four properties:

- handle general, overlapping group structures,
- online: fast, memory efficient, adaptive,
- non-convex sparsity inducing regularization:
  - fewer measurements,
  - weaker conditions on the dictionary,
  - robust (w.r.t. noise, compressibility).
- can deal with missing information.

Current approaches: handle ≤ 2.

Cost function

- Notation: $\boldsymbol{\alpha}$ hidden representation, $\mathbf{x}$ observation, $\mathbf{D}$ dictionary, $\mathcal{G}$ group structure (set system) $\subseteq 2^{\{1,\dots,d_\alpha\}}$.
- Group structure inducing on the hidden representation:

$$\Omega(\boldsymbol{\alpha}) = \left\| \left(\|\boldsymbol{\alpha}_G\|_2 \right)_{G\in\mathcal{G}} \right\|_\eta, \tag{2}$$

$$\Omega(\boldsymbol{\alpha}) = \left\| \left(\|\mathbf{d}^G \circ \boldsymbol{\alpha}\|_2 \right)_{G\in\mathcal{G}} \right\|_\eta, \tag{3}$$

$$\Omega(\boldsymbol{\alpha}) = \left\| \left(\|\mathbf{A}^G \boldsymbol{\alpha}\|_2 \right)_{G\in\mathcal{G}} \right\|_\eta, \quad \eta \in (0,2). \tag{4}$$

- Approximate on the observed coordinates ($\mathbf{x}_O$):

$$\frac{1}{2} \|\mathbf{x}_O - \mathbf{D}_O \boldsymbol{\alpha}\|_2^2 . \tag{5}$$

Cost function – continued

- Loss for a fixed observation ($\kappa > 0$):

$$l(\mathbf{x}_O, \mathbf{D}_O) = \min_{\boldsymbol{\alpha}} \left[\frac{1}{2} \|\mathbf{x}_O - \mathbf{D}_O \boldsymbol{\alpha}\|_2^2 + \kappa \Omega(\boldsymbol{\alpha}) \right]. \qquad (6)$$

- Goal (OSDL): minimize the average loss of the dictionary

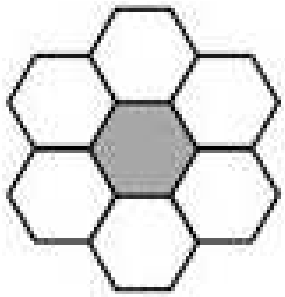
$$\min_{\mathbf{D}} f_t(\mathbf{D}) := \frac{1}{t} \sum_{i=1}^{t} l(\mathbf{x}_{O_i}, \mathbf{D}_{O_i}). \qquad (7)$$

- Possible dictionary/representation constraints:
  - $\mathbf{D} \in \mathcal{D} = \times_{i=1}^{d_\alpha} \mathcal{D}_i \subseteq \mathbb{R}^{d_x \times d_\alpha}$: closed, convex, and bounded.
  - $\boldsymbol{\alpha} \in \mathcal{A} \subseteq \mathbb{R}^{d_\alpha}$: convex, closed.

Special cases

- $O_i = \{1, \ldots, d_x\}$ $(\forall i)$: fully observed OSDL task.
- Special cases for $\mathcal{G}$:

'Traditional' sparse dictionary	$\mathcal{G} = \{\{1\}, \{2\}, \ldots, \{d_\alpha\}\}$.
Hierarchical dictionary	$\mathcal{G}$ = descendants of the nodes.
Grid adopted dictionary	$\mathcal{G}$ = nearest neighbors of the nodes.
Group Lasso	$\mathcal{G}$ = partition.
Elastic net	$\mathcal{G}$ = singletons and $\{1, \ldots, d_\alpha\}$.
Contiguous code	$\mathcal{G}$ = intervals.



Special cases – continued

Special cases for $\{\mathbf{A}^G\}_{G\in\mathcal{G}}$:

Fused Lasso	$\Omega(\boldsymbol{\alpha}) = \sum_{j=1}^{d_\alpha - 1} \lvert\alpha_{j+1} - \alpha_j\rvert.$
Graph-guided fusion penalty	$\Omega(\boldsymbol{\alpha}) = \sum_{e=(i,j)\in E : i<j} w_{ij}\lvert\alpha_i - v_{ij}\alpha_j\rvert.$
Linear trend/polynomial filtering	$\Omega(\boldsymbol{\alpha}) = \sum_{j=2}^{d_\alpha - 1} \lvert -\alpha_{j-1} + 2\alpha_j - \alpha_{j+1}\rvert.$
Generalized Lasso penalty	$\Omega(\boldsymbol{\alpha}) = \lVert\mathbf{A}\boldsymbol{\alpha}\rVert_1.$
Total variation	$\Omega(\boldsymbol{\alpha}) = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} \left\lVert(\nabla\boldsymbol{\alpha})_{ij}\right\rVert_2.$

Special cases – continued

Special cases for $\mathcal{D}, \mathcal{A}$:

'Traditional' setting	ℓ_2 constrained **D**.
Structured NMF	non-negative **D** and α.
Structured mixture-of-topics	ℓ_1 constained **D**, non-negative **D**, α.
'Hard' representation constraints	group norm/elastic net/ fused Lasso constrained α.
Double structured dictionaries	group norm constraints to α *and* **D**.

OSDL optimization

Online optimization of $\mathbf{D}$ through alternations:

- For fix $\mathbf{D}_{t-1}$ and $\mathbf{x}_{O_t}$, α_t is the solution of

$$\alpha_t = \underset{\alpha \in \mathcal{A}}{\operatorname{argmin}} \left[\frac{1}{2} \left\| \mathbf{x}_{O_t} - (\mathbf{D}_{t-1})_{O_t} \alpha \right\|_2^2 + \kappa \Omega(\alpha) \right]. \quad (8)$$

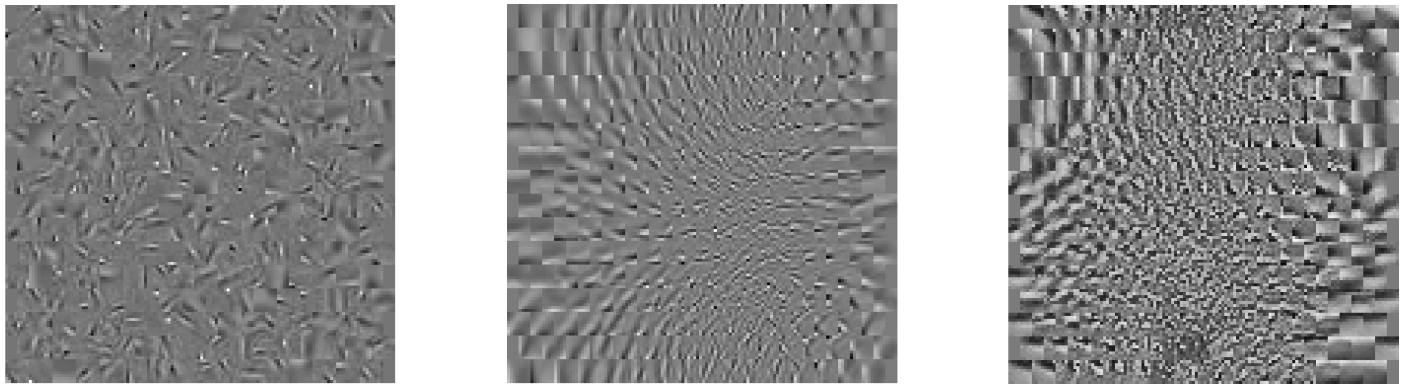
- 'Using' $\{\alpha_i\}_{i=1}^t$, $\mathbf{D}_t$ is updated by means of the quadratic optimization

$$\hat{f}_t(\mathbf{D}_t) = \min_{\mathbf{D} \in \mathcal{D}} f_t(\mathbf{D}, \{\alpha_i\}_{i=1}^t). \quad (9)$$

Solution idea: variational property of $\|\cdot\|_\eta$; BCD + 3 different $\hat{f}_t$ statistics + matrix recursions.

Numerical examples – inpainting of natural images

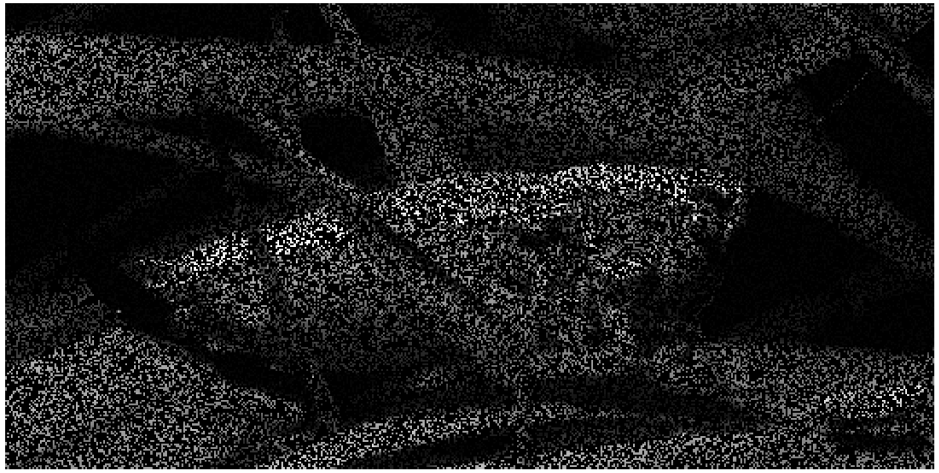
- Structured (toroid) vs. unstructured dictionary: $13-19\%$ improvement.
- Efficiency in case of missing observations: MSE grows slowly, $p_{tr}=0.9$ (training incompleteness: 90%) is still OK.



Left: unstructured; center: structured; right: structured, incomplete observations.

Numerical examples – inpainting, *full* unseen image

Learning: $p_{tr} = 0.5$. Inpainting: $p_{test}^{val} = 0.7$

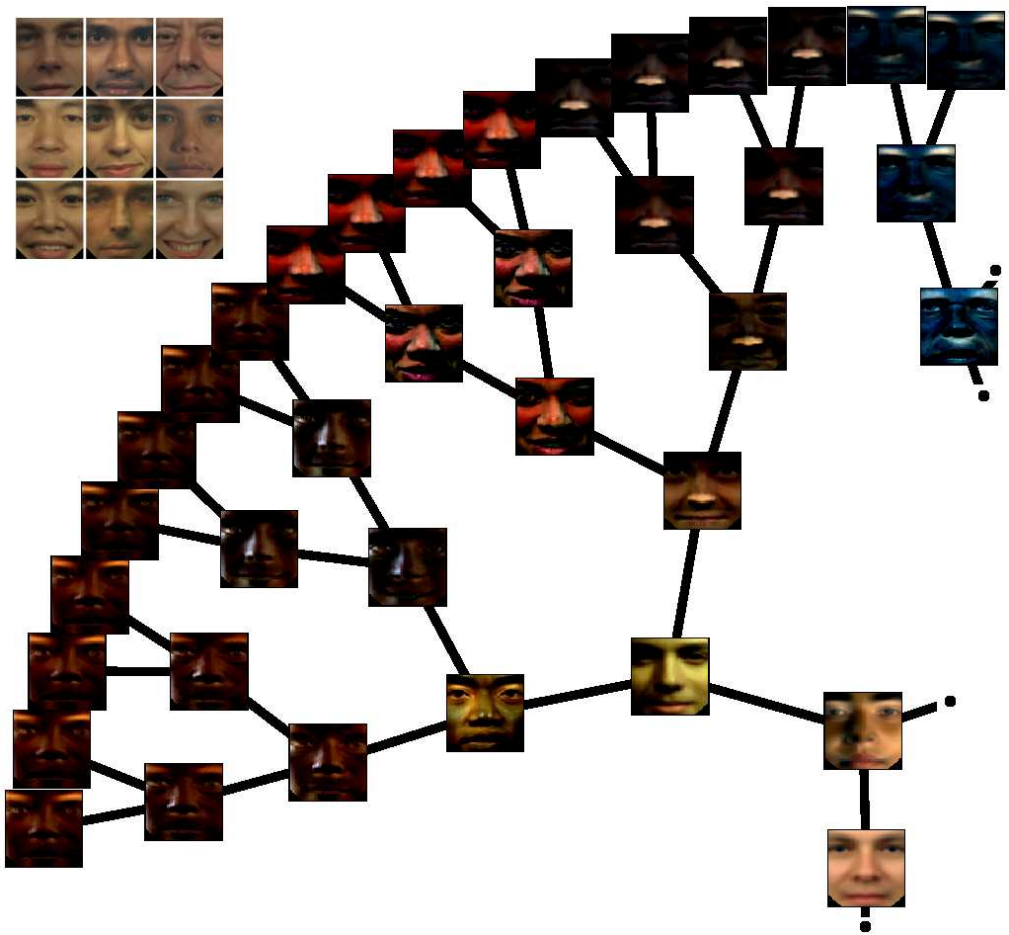


Numerical examples – inpainting, *full* unseen image

Learning: $p_{tr} = 0.5$. Inpainting: $p_{test}^{val} = 0.7$ (PSNR = 29 dB):

Numerical examples – online structured NMF on faces

- Online, $\mathcal{G}$-NMF: special case of OSDL.
- Illustration: color FERET, 140×120 sized facial dataset.
- $\mathcal{G}$: complete, 8-level binary tree ($d_\alpha = 255$).



Numerical examples – collaborative filtering

- Joke recommendation (Jester): 100 jokes $\times$ $73,421$ users.
- Observation: $\mathbf{x}_{O_t}$ = ratings of the t^{th} user.
- Baseline: best known RMSEs
  - 4.1123 (item neighbor),
  - 4.1229 (unstructured dictionary, $d_\alpha = 100$).
- Result:
  - toroid $\mathcal{G}$ ($d_\alpha = 100$): RMSE = 4.0774,
  - hierarchical $\mathcal{G}$ ($d_\alpha = 15$): RMSE = 4.1220.

Conclusions

- We developed a dictionary learning method, which
  - enables general overlapping group structures,
  - is online,
  - applies non-convex sparsity inducing regularization,
  - can deal with missing information.
- $\Rightarrow$ Dictionary learning for several actively studied structured sparse coding problems.
- Numerical examples: inpainting of natural images, structured NMF, collaborative filtering.

Acknowledgments

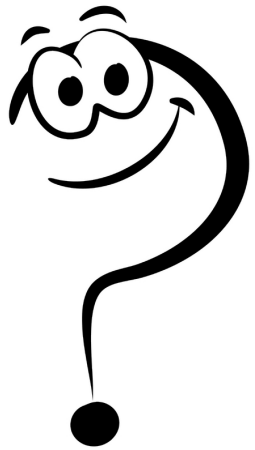
The research was partly supported by the Department of Energy (grant number DESC0002607).

Nemzeti Fejlesztési Ügynökség
www.ujszechenyiterv.gov.hu
06 40 638 638

MAGYARORSZÁG MEGÚJUL

The Project is supported by the European Union and co-financed by the European Social Fund (grant agreements no. TÁMOP 4.2.1/B-09/1/KMR-2010-0003 and KMOP-1.1.2-08/1-2008-0002).

Thank you for the attention!



Representation optimization (α)

- Structured sparse coding task:

$$\frac{1}{2}\left\|\mathbf{x}_{O_t} - (\mathbf{D}_{t-1})_{O_t}\boldsymbol{\alpha}\right\|_2^2 + \kappa\Omega(\boldsymbol{\alpha}) \to \min_{\boldsymbol{\alpha}\in\mathcal{A}}. \tag{10}$$

- Solution: let us use the

$$\|\mathbf{y}\|_\eta = \min_{\mathbf{z}\in\mathbb{R}_+^d}\left[\frac{1}{2}\sum_{j=1}^{d}\frac{y_j^2}{z_j} + \frac{1}{2}\|\mathbf{z}\|_\beta\right], \tag{11}$$

variational property of $\|\cdot\|_\eta$, where

- $\mathbf{y}\in\mathbb{R}^d$, $\beta = \frac{\eta}{2-\eta}$, and
- the minimum value is attained at $z_i^* = |y_i|^{2-\eta}\|\mathbf{y}\|_\eta^{\eta-1}$.

Representation optimization (α) – continued

Our problem is equivalent to the solution of

$$J(\alpha, \mathbf{z}) = \frac{1}{2}\left\|\mathbf{x}_{O_t} - (\mathbf{D}_{t-1})_{O_t}\alpha\right\|_2^2 + \kappa\frac{1}{2}\left(\alpha^T\mathbf{H}\alpha + \|\mathbf{z}\|_\beta\right) \to \min_{\alpha\in\mathcal{A}, \mathbf{z}\in\mathbb{R}_+^{|\mathcal{G}|}},$$

where

$$\mathbf{H} = \mathbf{H}(\mathbf{z}) = \sum_{G\in\mathcal{G}} (\mathbf{A}^G)^T\mathbf{A}^G/z^G. \tag{12}$$

One can optimize $J(\alpha, \mathbf{z})$ by iterative alternating steps:

- For given α: explicit formula for the optimal $\mathbf{z} = (z^G)_{G\in\mathcal{G}}$

$$z^G = \|\mathbf{A}^G\alpha\|_2^{2-\eta}\|(\|\mathbf{A}^G\alpha\|_2)_{G\in\mathcal{G}}\|_\eta^{\eta-1}. \tag{13}$$

- For given α: quadratic cost on the convex set $\mathcal{A}$.

Dictionary optimization ($\mathbf{D}$)

- Cost function (ρ: non-negative forgetting factor):

$$\hat{f}_t(\mathbf{D}) = \frac{1}{\sum_{j=1}^{t}(j/t)^{\rho}} \sum_{i=1}^{t} \left(\frac{i}{t}\right)^{\rho} \left[\frac{1}{2}\left\|\mathbf{x}_{O_i} - \mathbf{D}_{O_i}\boldsymbol{\alpha}_i\right\|_2^2 + \kappa\Omega(\boldsymbol{\alpha}_i)\right] \to \min_{\mathbf{D}\in\mathfrak{D}}.$$

- Optimization (BCD):
  - optimize in $\mathbf{d}_j$, while the other columns ($\mathbf{d}_i$, $i \neq j$) are fixed.
  - $\hat{f}_t$ is quadratic in $\mathbf{d}_j$:
    1. Solve the equation:

$$\frac{\partial \hat{f}_t}{\partial \mathbf{d}_j}(\mathbf{u}_j) = \mathbf{0}. \tag{14}$$

    2. Project the solution to the constraint set $\mathfrak{D}_j$:

$$\mathbf{d}_j = \Pi_{\mathfrak{D}_j}(\mathbf{u}_j). \tag{15}$$

Computation of $\mathbf{u}_j$

- Task:

$$\frac{\partial \hat{f}_t}{\partial \mathbf{d}_j}(\mathbf{u}_j) = \mathbf{0}. \tag{16}$$

- Solution: $\mathbf{u}_j$ satisfies the linear equation

$$\mathbf{C}_{j,t}\mathbf{u}_j = \mathbf{b}_{j,t} - \mathbf{e}_{j,t} + \mathbf{C}_{j,t}\mathbf{d}_j, \tag{17}$$

where for the $\{\{\mathbf{C}_{j,t}\}_{j=1}^{d_\alpha}, \mathbf{B}_t, \{\mathbf{e}_{j,t}\}_{j=1}^{d_\alpha}\}$ statistics

Computation of $\mathbf{u}_j$ – continued

$$\mathbf{C}_{j,t} = \sum_{i=1}^{t} \left(\frac{i}{t}\right)^{\rho} \boldsymbol{\Delta}_i \alpha_{i,j}^2 \in \mathbb{R}^{d_x \times d_x} \quad (j = 1, \ldots, d_\alpha), \tag{18}$$

$$\mathbf{B}_t = \sum_{i=1}^{t} \left(\frac{i}{t}\right)^{\rho} \boldsymbol{\Delta}_i \mathbf{x}_i \boldsymbol{\alpha}_i^T = [\mathbf{b}_{1,t}, \ldots, \mathbf{b}_{d_\alpha,t}] \in \mathbb{R}^{d_x \times d_\alpha}, \tag{19}$$

$$\mathbf{e}_{j,t} = \sum_{i=1}^{t} \left(\frac{i}{t}\right)^{\rho} \boldsymbol{\Delta}_i \mathbf{D} \boldsymbol{\alpha}_i \alpha_{i,j} \in \mathbb{R}^{d_x} \quad (j = 1, \ldots, d_\alpha), \tag{20}$$

where $\mathbf{C}_{j,t}$ and $\boldsymbol{\Delta}_i$s are diagonal; $\boldsymbol{\Delta}_i$ matrix $\Leftrightarrow$ O_i (element j in the diagonal is 1 if $j \in O_i$, and 0 otherwise).

Matrix recursion lemma

Let

- $\mathbf{N}_t \in \mathbb{R}^{L_1 \times L_2}$ $(t = 1, 2, \ldots)$ be a given matrix series,
- $\gamma_t = \left(1 - \frac{1}{t}\right)^{\rho}$, $\rho \geq 0$,
- the $\mathbf{M}_t$ and $\mathbf{M}'_t$ matrix series be defined as

$$\mathbf{M}_t = \gamma_t \mathbf{M}_{t-1} + \mathbf{N}_t \in \mathbb{R}^{L_1 \times L_2} \quad (t = 1, 2, \ldots), \tag{21}$$

$$\mathbf{M}'_t = \sum_{i=1}^{t} \left(\frac{i}{t}\right)^{\rho} \mathbf{N}_i \in \mathbb{R}^{L_1 \times L_2} \quad (t = 1, 2, \ldots). \tag{22}$$

If $\rho = 0$, then $\mathbf{M}_t = \mathbf{M}_0 + \mathbf{M}'_t$ $(\forall t \geq 1)$. When $\rho > 0$, then $\mathbf{M}_t = \mathbf{M}'_t$ $(\forall t \geq 1)$.

Computation of $\mathbf{u}_j$ – continued

- Matrix recursion lemma $\Rightarrow$ one can update $\mathbf{C}_{j,t}$ and $\mathbf{B}_t$ as

$$\mathbf{C}_{j,t} = \gamma_t \mathbf{C}_{j,t-1} + \boldsymbol{\Delta}_t \alpha_{tj}^2, \tag{23}$$

$$\mathbf{B}_t = \gamma_t \mathbf{B}_{t-1} + \boldsymbol{\Delta}_t \mathbf{x}_t \boldsymbol{\alpha}_t^T, \tag{24}$$

  with $\mathbf{C}_{j,0} = \mathbf{0}$, $\mathbf{B}_0 = \mathbf{0}$ ($\rho = 0$), or arbitrary initialization ($\rho > 0$).

- Numerical experiences $\Rightarrow$ efficient online approximation for $\mathbf{e}_{j,t}$:

$$\mathbf{e}_{j,t} = \gamma_t \mathbf{e}_{j,t-1} + \boldsymbol{\Delta}_t \mathbf{D} \boldsymbol{\alpha}_t \alpha_{t,j}, \tag{25}$$

  with the actual estimation $\mathbf{D}$ and initialization $\mathbf{e}_{j,0} = \mathbf{0}$.

Special, fully observable case

In this case ($\mathbf{\Delta}_i = \mathbf{I}, \forall i$):

$$\mathbf{C}_{j,t} = \mathbf{I} \sum_{i=1}^{t} \left(\frac{i}{t}\right)^{\rho} \alpha_{i,j}^2, \quad \mathbf{B}_t = \sum_{i=1}^{t} \left(\frac{i}{t}\right)^{\rho} \mathbf{x}_i \boldsymbol{\alpha}_i^T, \tag{26}$$

$$\mathbf{e}_{j,t} = \sum_{i=1}^{t} \left(\frac{i}{t}\right)^{\rho} \mathbf{D}\boldsymbol{\alpha}_i \alpha_{i,j} = \mathbf{D} \sum_{i=1}^{t} \left(\frac{i}{t}\right)^{\rho} \boldsymbol{\alpha}_i \alpha_{i,j}, \tag{27}$$

that is

- $\mathbf{D}$ can be pulled out from $\mathbf{e}_{j,t}$s, and
- it is sufficient to maintain 2 statistics, $\mathbf{B}_t$ and

$$\mathbf{A}_t = \sum_{i=1}^{t} \left(\frac{i}{t}\right)^{\rho} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^T \in \mathbb{R}^{d_\alpha \times d_\alpha}. \tag{28}$$