Active & Online Learning

Frequency-aware Truncated methods for Sparse Online Learning

Hidekazu Oiwa, Shin Matsushima, Hiroshi Nakagawa
University of Tokyo

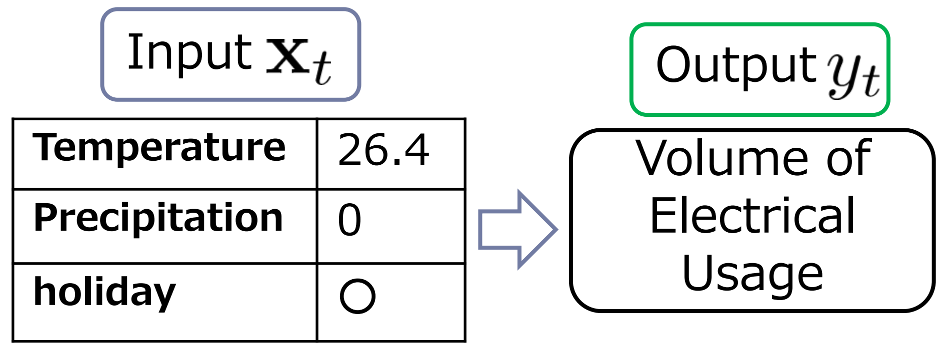
$$f(\mathbf{x}_t) = y_t$$

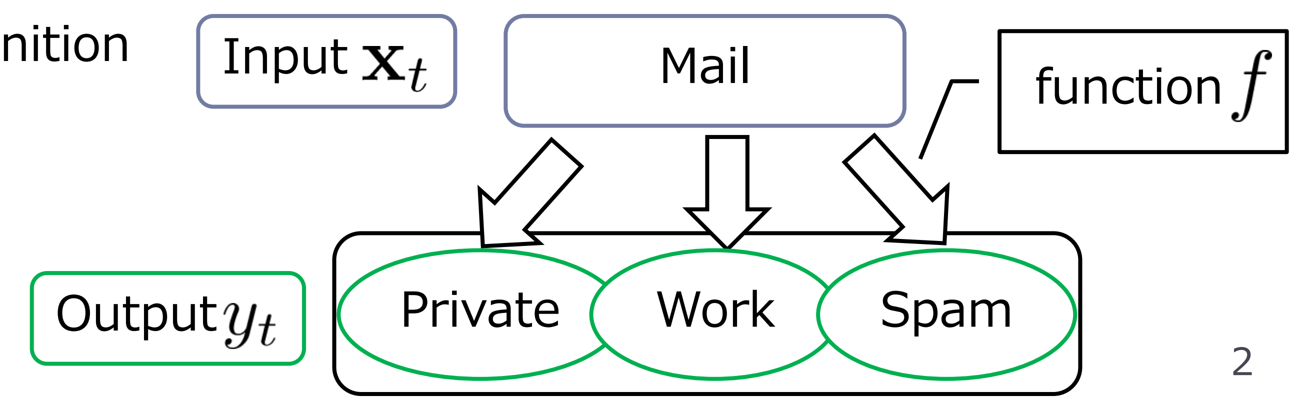
Problem Setting

- Supervised Learning

- Predict output y_t from a given input $\mathbf{x}_t$
  - Learn appropriate function $f(\cdot)$ by using dataset $(\mathbf{x}_t, y_t)$

- Examples of Application
  - Regression
    - Precipitation rate Prediction
    - Electrical Usage Prediction
  - Classification
    - Mail Filtering
    - News Categorization
    - Image Recognition





Notation

- Input $\mathbf{x} \in \mathbf{X} \subset \Re^n$
  - Feature : each component of $\mathbf{x} = \{\, 0 \,,\, 1 \,,\, 0 \,, \ldots ,\, 0 \,,\, 1 \,\}$ (Coach, Curling)
- Output $y \in \mathbf{Y} \subset \Re$

$$y = \begin{cases} 1 & \text{sports article} \\ -1 & \text{Not sports article} \end{cases}$$

- Weight vector $\mathbf{w} \in \mathbf{W} \subset \Re^n$
  - Linear prediction
    - Predict the $\hat{y}$ using value of inner product $\mathbf{w}, \mathbf{x}$
  - Truncation : Component of $\mathbf{w}$ becomes zero

Predicted value

$$\hat{y} = \langle \mathbf{w}, \mathbf{x} \rangle > 0 \quad \text{sports article}$$

$$\hat{y} = \langle \mathbf{w}, \mathbf{x} \rangle < 0 \quad \text{Not sports article}$$

Optimization Problem

Minimize sum of loss function and regularized term

$$\mathbf{w} = \arg\min_{\mathbf{w}} \sum_t \{\ell_t(\mathbf{w}) + r_t(\mathbf{w})\}$$

t : data id

Loss function

$$\ell_t(\mathbf{w}) : \mathbf{W} \rightarrow \Re_+$$

Evaluate performance of data fitting

Regularized term

$$r_t(\mathbf{w}) : \mathbf{W} \rightarrow \Re_+$$

Evaluate complexity of weight vector

Derive optimal function $f(\cdot)$

⇩

Derive optimal $\mathbf{w}$

Online Learning

Be able to run even when
the only part of data is observable

- Update $\mathbf{w}_t$ on one piece of data at each round

$\mathbf{w}_{t-1}$

Datum

$$\mathbf{x}_{t-1} = \{0, 1, 0, 0, 2\}$$
$$y_{t-1} = 1$$

$\mathbf{w}_t$

Datum

$$\mathbf{x}_t = \{2, 0, 0, 1, 0\}$$
$$y_t = -1$$

$\mathbf{w}_{t+1}$

Datum

$$\mathbf{x}_{t+1} = \{0, 0, 5, 2, 4\}$$
$$y_{t+1} = 1$$

$\vdots$

<u>Change Optimization Problem</u>

$$\min_{\mathbf{w}} \sum_t \{\ell_t(\mathbf{w}) + r_t(\mathbf{w})\}$$

$$\Downarrow$$

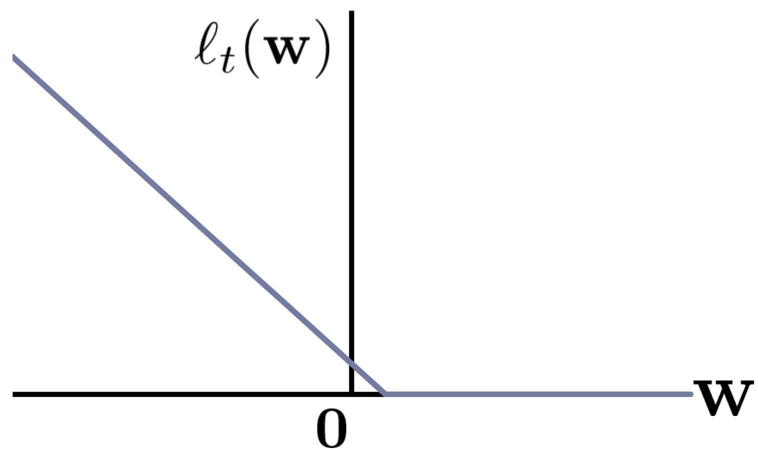
$$\min_{\mathbf{w}_t} \sum_t \{\ell_t(\mathbf{w}_t) + r_t(\mathbf{w}_t)\}$$

Loss function $\ell_t(\mathbf{w}) : \mathbf{W} \rightarrow \Re_+$

- Evaluate prediction accuracy of $\mathbf{w}$
  - Loss function's gradient is proportional to $\mathbf{x}$

<u>Hinge-Loss</u>

$$\ell_t(\mathbf{w}) = [1 - y_t \langle \mathbf{w}, \mathbf{x}_t \rangle]_+$$



In addition,
<u>Squared-Loss</u> etc..

Difference between $\hat{y}$ and y is large

⇩

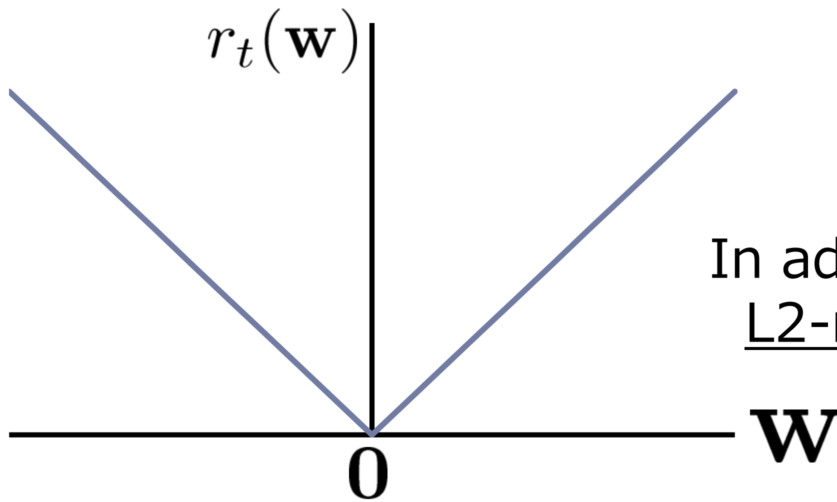
Value of loss function is large

Regularized term $r_t(\mathbf{w}) : \mathbf{W} \to \Re_+$

- Prevent over-fitting of $\mathbf{w}$

L1-regularization (Lasso)

$$r_t(\mathbf{w}) = \lambda \|\mathbf{w}\|_1$$



In addition,
L2-regularization etc..

λ : parameter between loss minimization and regularization

$r_t(\cdot)$ is proportional to complexity of $\mathbf{w}$

⇩

Prevent over-fitting to previous data

Additional Property of Lasso

- Lasso can truncate parameters of $\mathbf{w}$

<u>Update formula</u>

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \left\{ \|\mathbf{w} - \mathbf{w}_t\|_2^2 + \|\mathbf{w}\|_1 \right\}$$

$$\mathbf{w}_t = \{ \quad 3 \quad , \quad 2 \quad , \quad 1 \quad , \quad 0.5 \quad , \quad 0.1 \quad \}$$

⇩

$$\mathbf{w}_{t+1} = \{ \quad 2.5 \quad , \quad 1.5 \quad , \quad 0.5 \quad , \quad 0 \quad , \quad 0 \quad \}$$

Make $\mathbf{w}$ sparse and so can learn faster

Previous Work

- Online Learning+Lasso

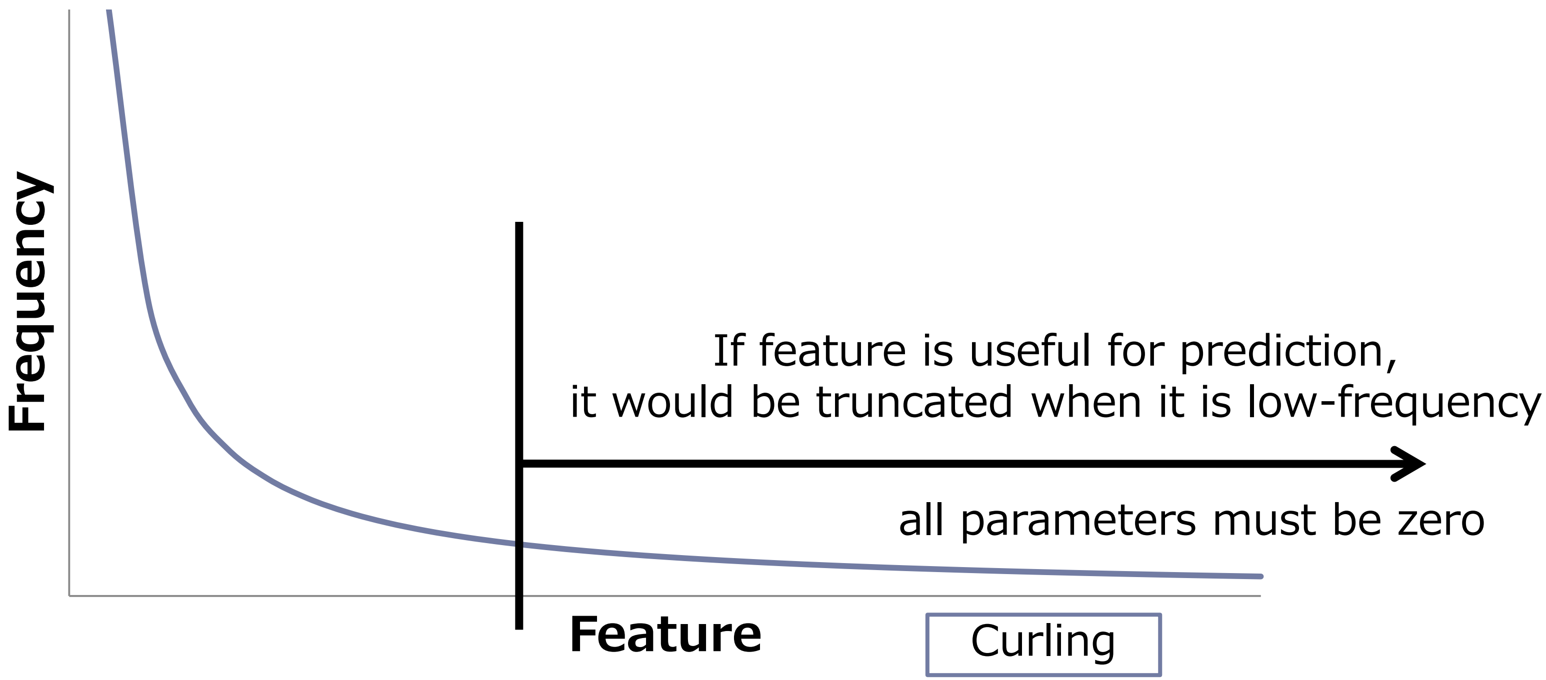
- Forward Backward Splitting (FOBOS) [Duchi et al., 2009]
  - Combine online Learning with Lasso
  - Perform two-step update at each round
  - [Langford et al., 2009] proposed similar method

- Regularized Dual-Averaging methods (RDA) [Xiao, 2009]
  - Dual-Averaging(DA) is optimization method for sequential data [Nesterov, 2009]
  - RDA introduce Lasso into DA

In our research,
we propose the extensional method of FOBOS

Disadvantage of Previous Work

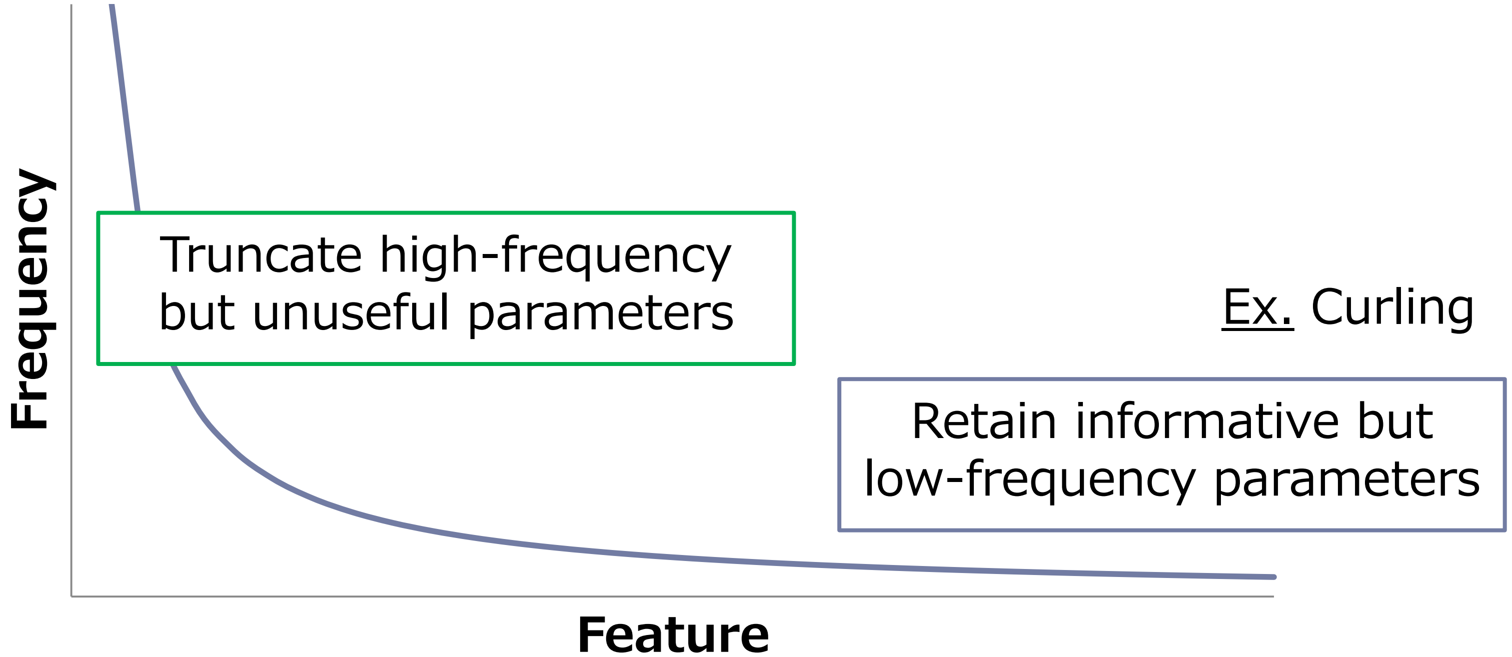
- Low-frequency features tend to be truncated
  - Difficult to use these features for prediction



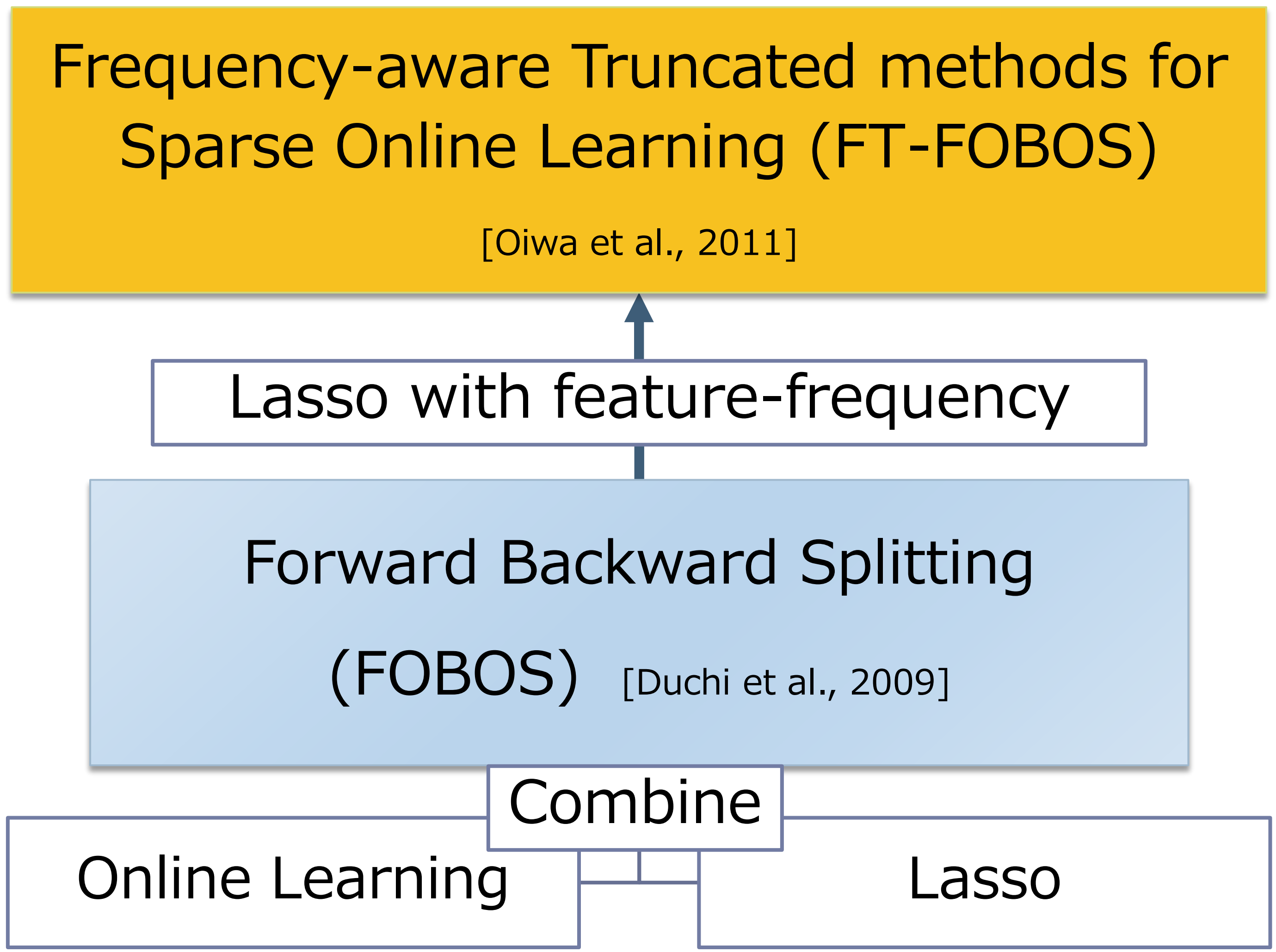
- Cannot achieve objective of Lasso
  - Useful but low-frequency feature would be missed

Proposed method : Intention

- Lasso with feature-frequency
  - Capture low-frequency but informative feature
  - Proposed several work in batch-learning field
    - Ex. TF-IDF (Natural Language Processing)
    - However, these methods cannot be applied in online learning framework



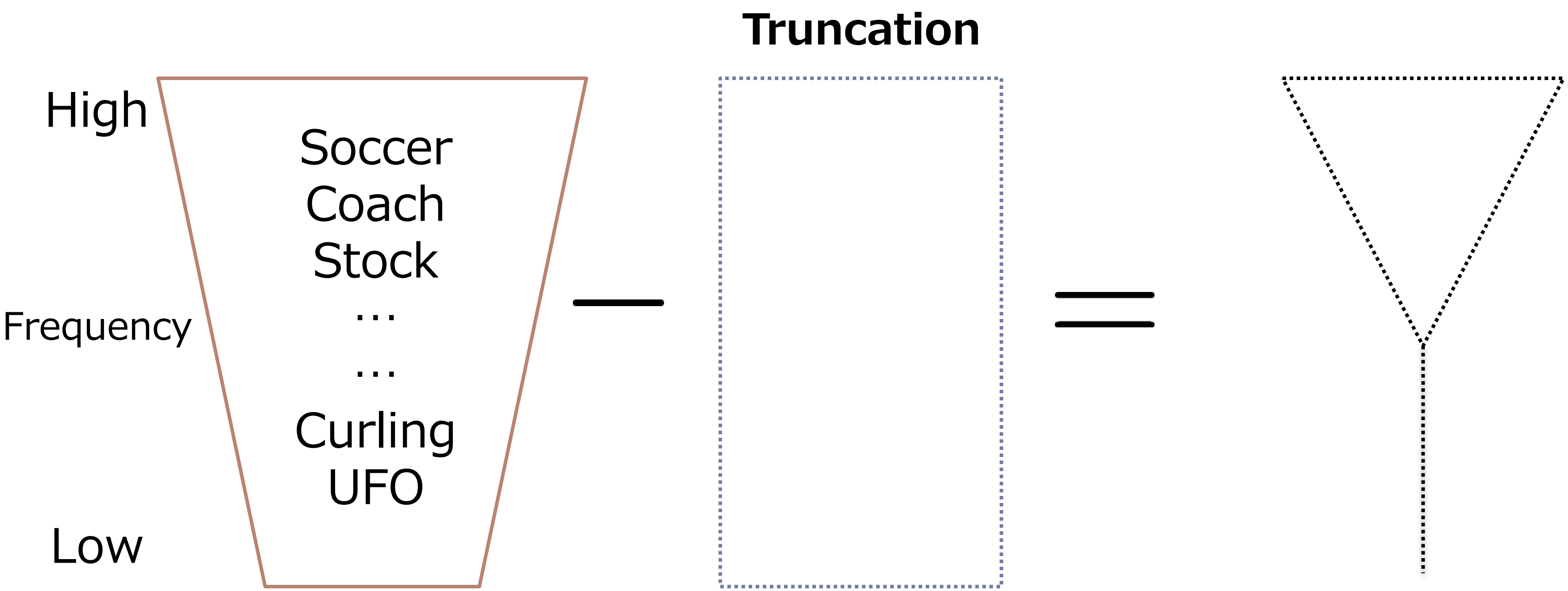
Proposed method (FT-FOBOS)



Proposed method [1/2]

Frequency-aware Truncated FOBOS (FT-FOBOS)

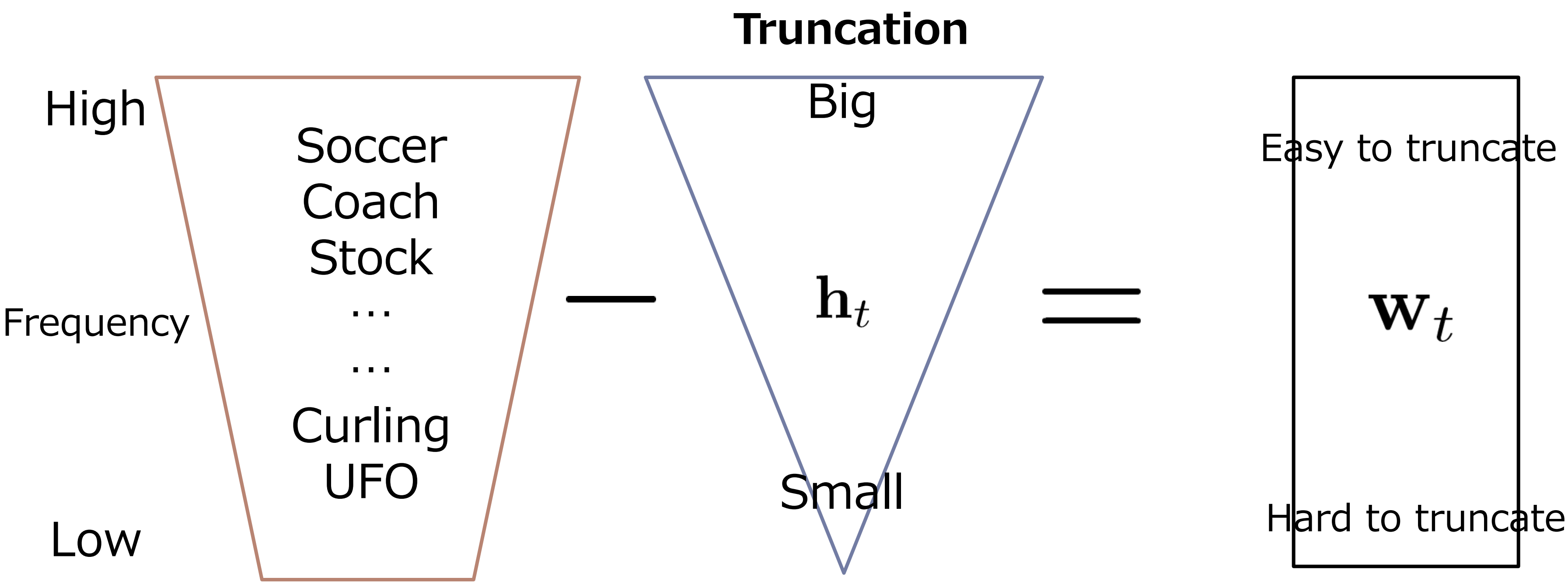
- Introduce $\mathbf{h}_t$ which has correlation with feature-frequency into Lasso



Proposed method [1/2]

Frequency-aware Truncated FOBOS (FT-FOBOS)

- Introduce $\mathbf{h}_t$ which has correlation with feature-frequency into Lasso

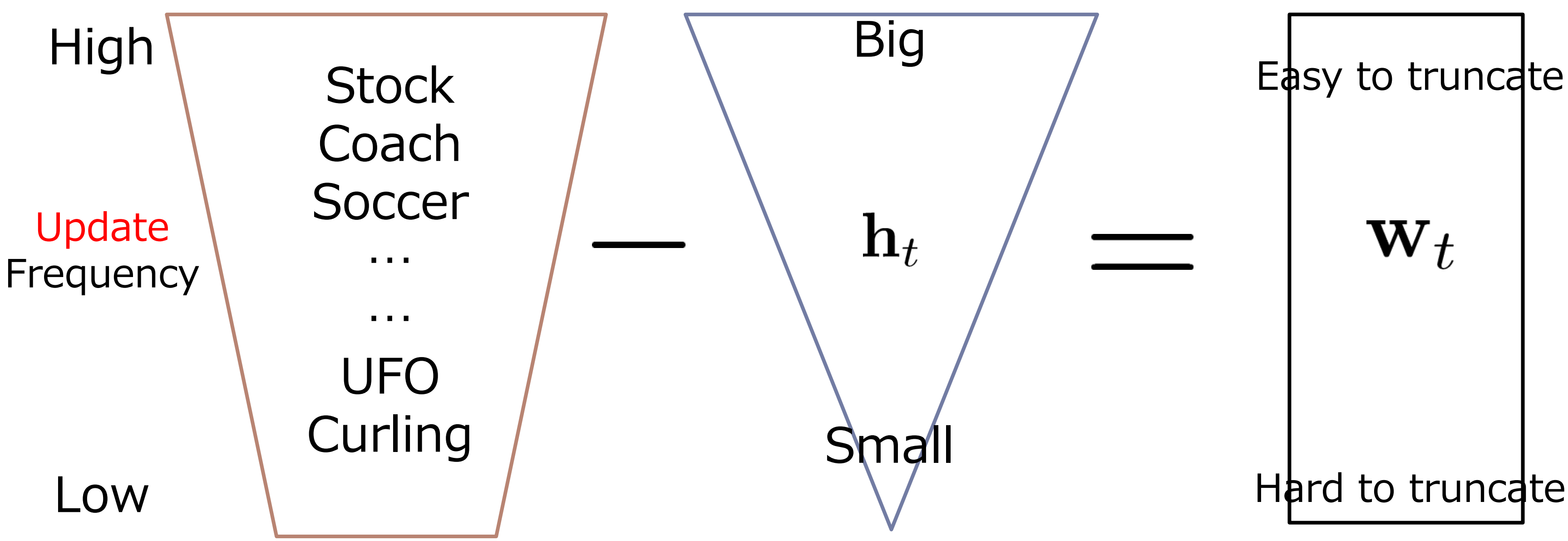


Proposed method [2/2]

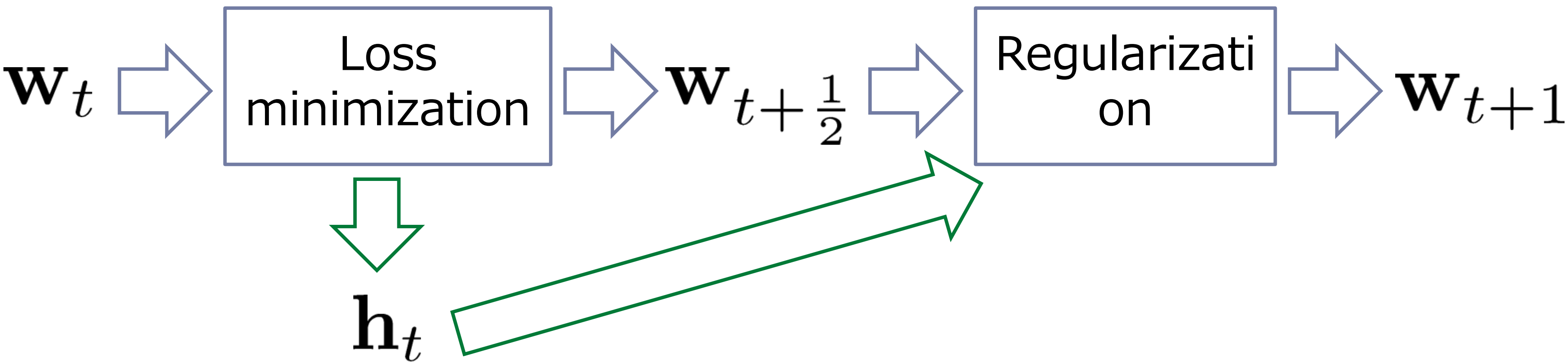
Frequency-aware Truncated FOBOS (FT-FOBOS)

- Bad when simply $\mathbf{h}_t$ is proportional to frequency
  - Value ranges of $\mathbf{w}_t$ depend more on update-frequency than feature-frequency

- Make $\mathbf{h}_t$ as a correlation with update-frequency



Algorithm of FT-FOBOS [1/3]



<u>Loss minimization step</u>

Update $\mathbf{w}_t$ into reverse direction of subgradient $\ell_t(\mathbf{w}_t)$

$$\mathbf{w}_{t+\frac{1}{2}} = \mathbf{w}_t - \eta_t \mathbf{g}_t^{\ell}$$

$\eta_t > 0$: Step size

$\mathbf{g}_t^{\ell} \in \partial \ell_t(\mathbf{w}_t)$: Subgradient of loss

Algorithm of FT-FOBOS [2/3]

Define $\mathbf{h}_t$ using constant $p > 0$

$$h_t^{(i)} = h_{t,p}^{(i)} = \sqrt[p]{\sum_{\tau=1}^{t} \left| \eta_\tau g_\tau^{\ell,(i)} \right|^p}$$

$\left| \eta_\tau g_\tau^{\ell,(i)} \right|$: Step size in loss minimization

weight of i component rarely update ⇨ only small number of $g_\tau^{\ell,(i)}$ is non-zero ⇨ value of $h_t^{(i)}$ becomes small

$\mathbf{h}_t$ can be calculated in $O(\mathbf{g}_t^\ell$'s nonzero number)

Algorithm of FT-FOBOS [3/3]

<u>Lasso step</u>

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2} \|\mathbf{w} - \mathbf{w}_{t+\frac{1}{2}}\| + \eta_{t+\frac{1}{2}} \lambda \|\mathbf{M}_t \mathbf{w}\|_1 \right\}$$

$$s.t. \quad \mathbf{M}_t = \begin{pmatrix} h_t^{(1)} & 0 & \dots & 0 \\ 0 & h_t^{(2)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & h_t^{(n)} \end{pmatrix}$$

<u>Update formula</u>

$$w_{t+1}^{(i)} = sign(w_t^{(i)} - \eta_t g_t^{\ell,(i)}) \left[|w_t^{(i)} - \eta_t g_t^{\ell,(i)}| - \eta_{t+\frac{1}{2}} \lambda h_t^{(i)} \right]_+$$

$\mathbf{w}_{t+1}$ can be calculated in $O(\mathbf{g}_t^{\ell}$ ‘s nonzero number)

<u>Same order as FOBOS</u>

Theoretical Evaluation

- Proof with "Regret"
  - Regret's Definition

$$R_{\ell+r}(T) = \sum_{t=1}^{T} \{\ell_t(\mathbf{w}_t) + r_t(\mathbf{w}_t)\} - \inf_{\mathbf{w}} \sum_{t=1}^{T} \{(\ell_t(\mathbf{w}) + r_t(\mathbf{w}))\}$$

Cumulative Loss and regularization

Minimal Loss and regularization ex post

- Prove convergence to optimal solution
  - Regret's Upper Bound is smaller than $O(T)$
  - Regret per datum converges 0 as data increase
  - Weight vector converges to optimal solution

$$R_{\ell+r}(T) < O(T) \Leftrightarrow \lim_{T\to\infty} \frac{R_{\ell+r}(T)}{T} = 0$$

FT-FOBOS's Regret

$p \leq 2$ の時は，定数 V で上限を定める

Let $\mathbf{h}_t$ be

$$h_t^{(i)} = \begin{cases} \min\left(h_{t,p}^{(i)}, V\right) & p \leq 2 \\ h_{t,p}^{(i)} & p > 2 \end{cases} \quad s.t.\ h_{t,p}^{(i)} = \sqrt[p]{\sum_{\tau=1}^{t} \left|\eta_\tau g_\tau^{\ell,(i)}\right|^p}$$

both loss function and regularized term are convex functions, and they satisfy

$$\forall \mathbf{w}_t \quad \|\mathbf{w}_t - \mathbf{w}^*\| \leq D, \|\partial \ell_t(\mathbf{w}_t)\| \leq G, \|\partial r_t(\mathbf{w}_t)\| \leq G$$

where scalars D, G .

In this case, we can prove

<u>Same order as FOBOS</u>

$$R_{\ell+r}(T) \leq 2GD + \left(D^2/2c + 8G^2c\right)\sqrt{T} = O(\sqrt{T})$$

where we set a scalar $c > 0$ and stepsize $\eta_t = \eta_{t+\frac{1}{2}} = \frac{c}{\sqrt{t}}$

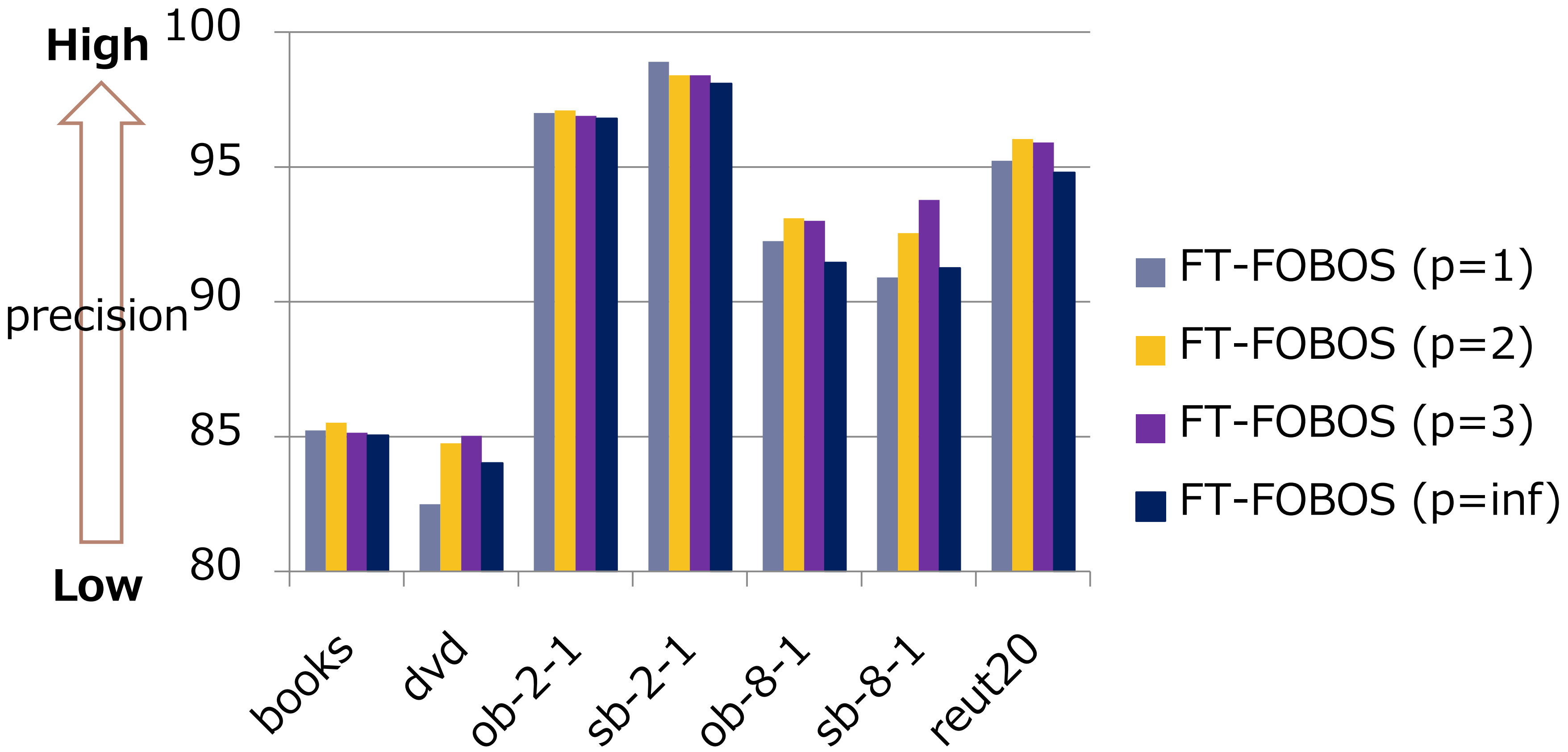
Experimental Evaluations

- Seven real dataset experiments
  - Loss function : Hinge-Loss
  - 10-fold cross validation for adjusting λ
  - 20 iterations
  - Algorithms : FOBOS, RDA, FT-FOBOS $p = 1, 2, 3, \infty$
  - Step size : $\eta_t = \eta_{t+\frac{1}{2}} = \frac{1}{\sqrt{t}}$

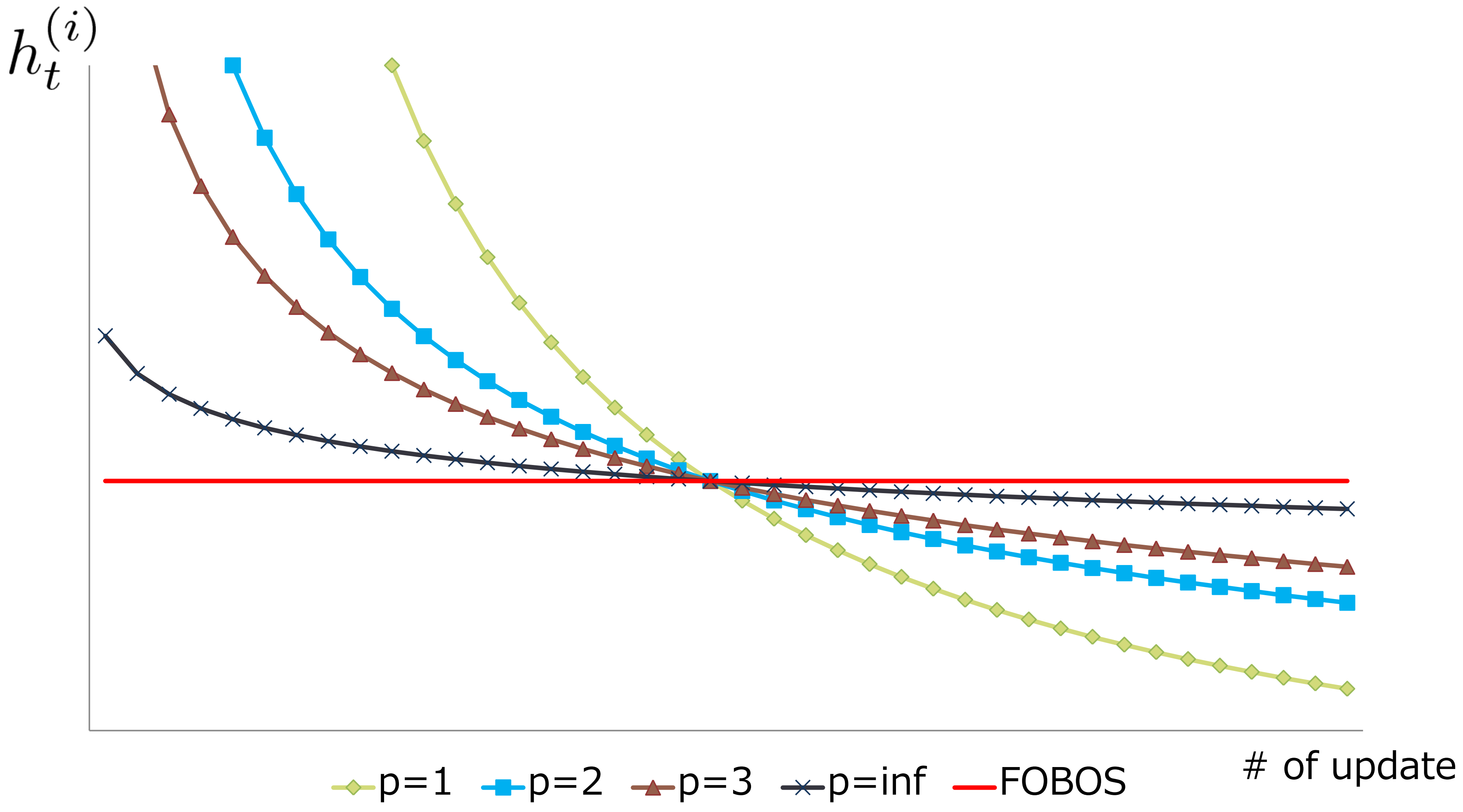
	# of data	# of feature	# of class
books	4,465	332,441	2
dvd	3,586	282,901	2
ob-2-1	1,000	5,942	2
sb-2-1	1,000	6,276	2
ob-8-1	4,000	13,890	8
sb-8-1	4,000	16,282	8
reut20	7,800	34,488	20

Experimental Results among FT-FOBOS

- Compare <u>precision</u>
  - $p = 2$ achieves the best performance

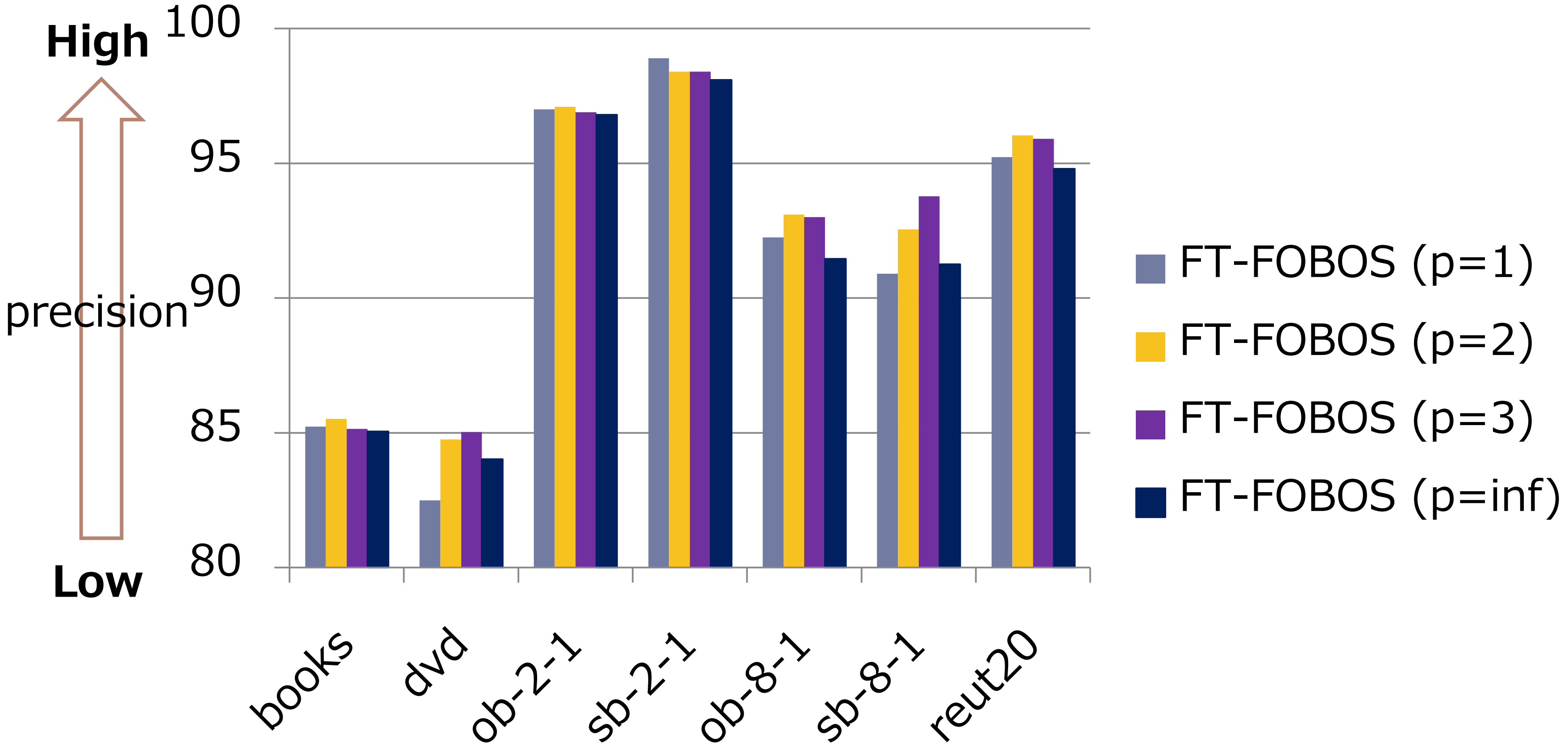


$h_t^{(i)}$'s disparity when change p



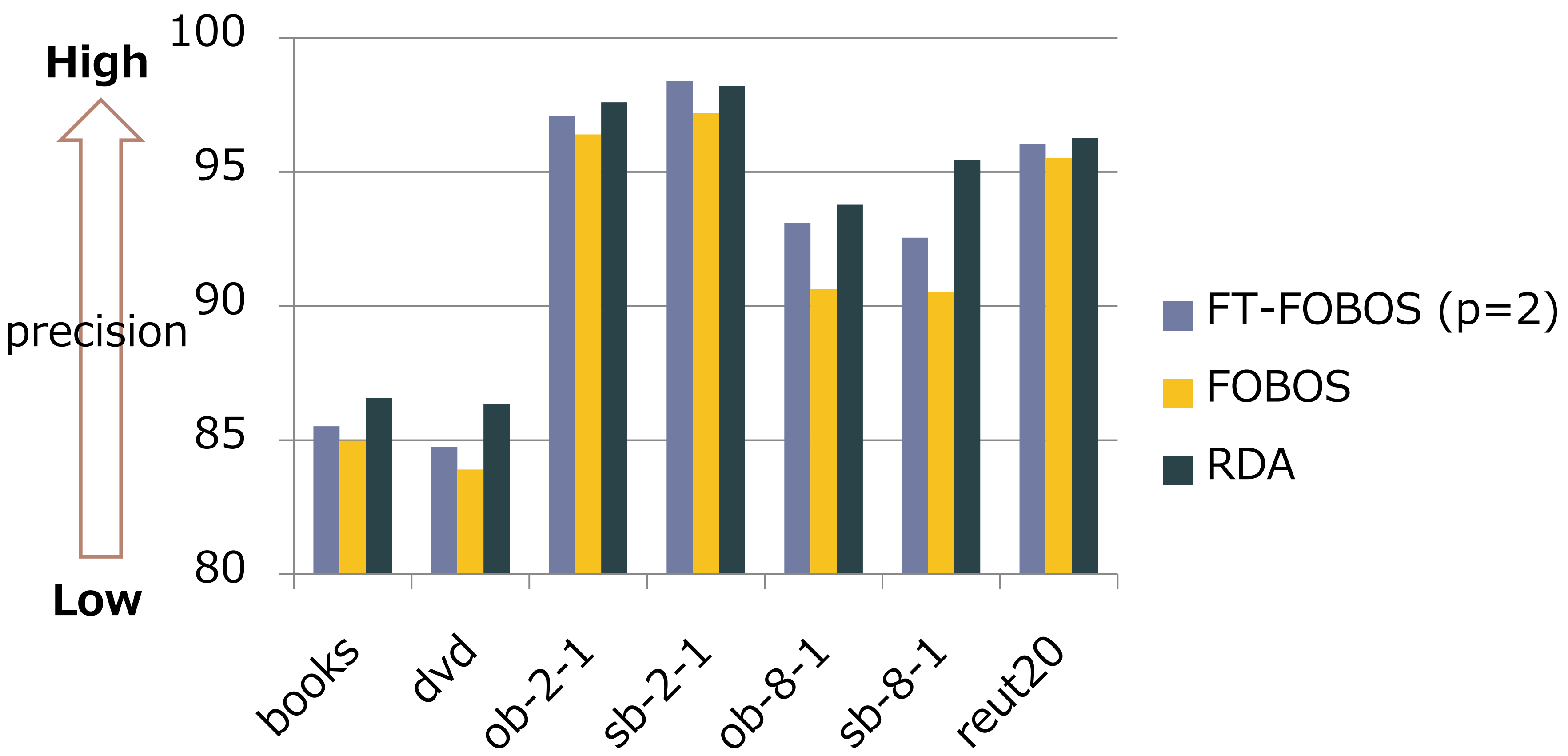
Experimental Results among FT-FOBOS

- Compare <u>precision</u>
  - $p = 2$ achieves the best performance



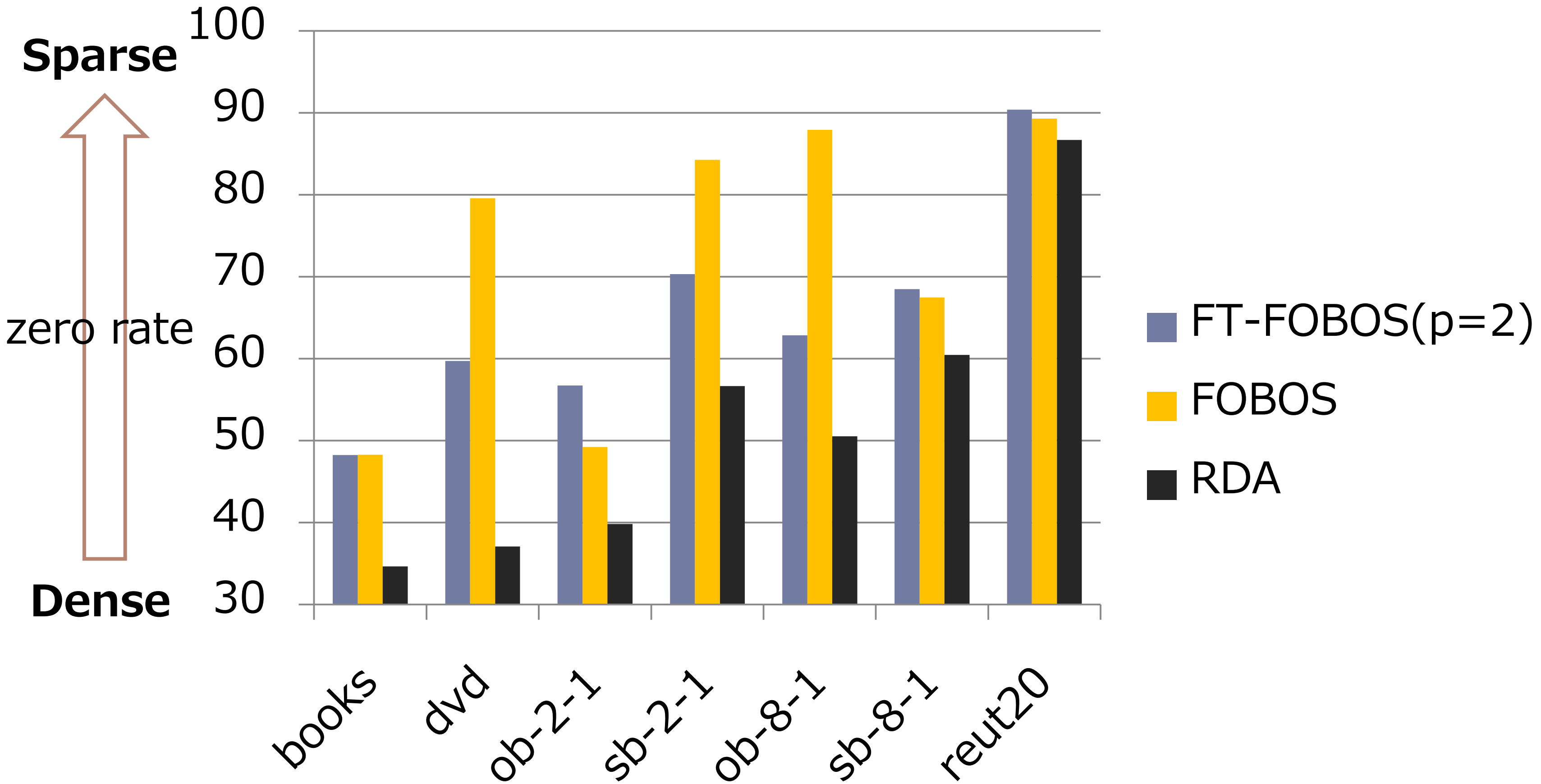
Experimental Results of all algorithms

- Compare precision
  - FT-FOBOS outperforms FOBOS in all datasets
  - RDA is better than FOBOS and FT-FOBOS



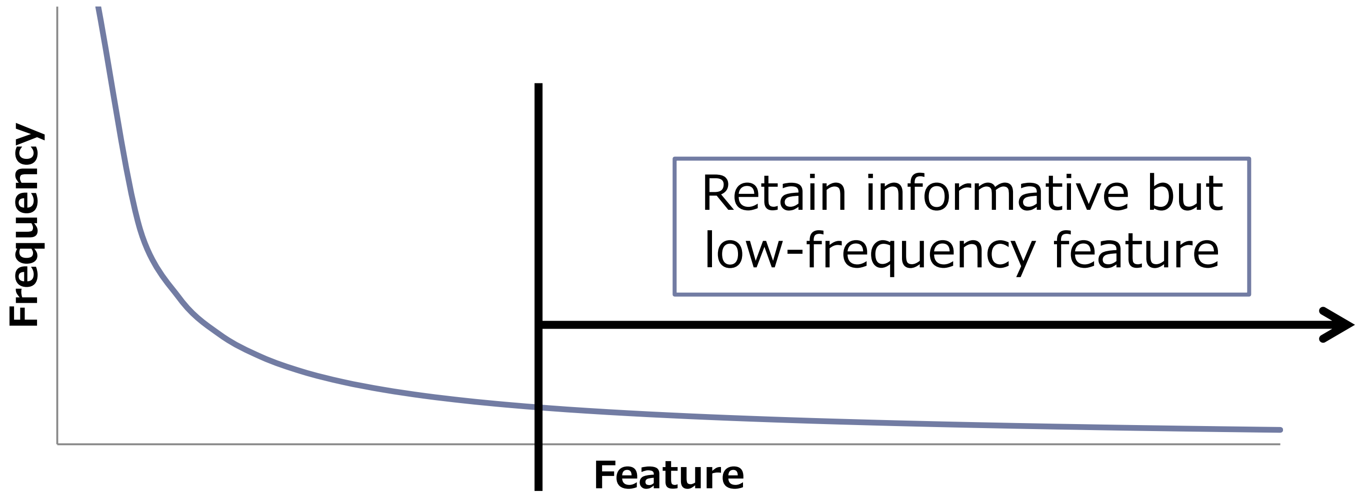
Experimental Results of all algorithms

- Compare <u>sparseness of weight vector</u>
  - FT-FOBOS improve accuracy while obtaining almost same sparseness



Summary

- Lasso with Feature Frequency

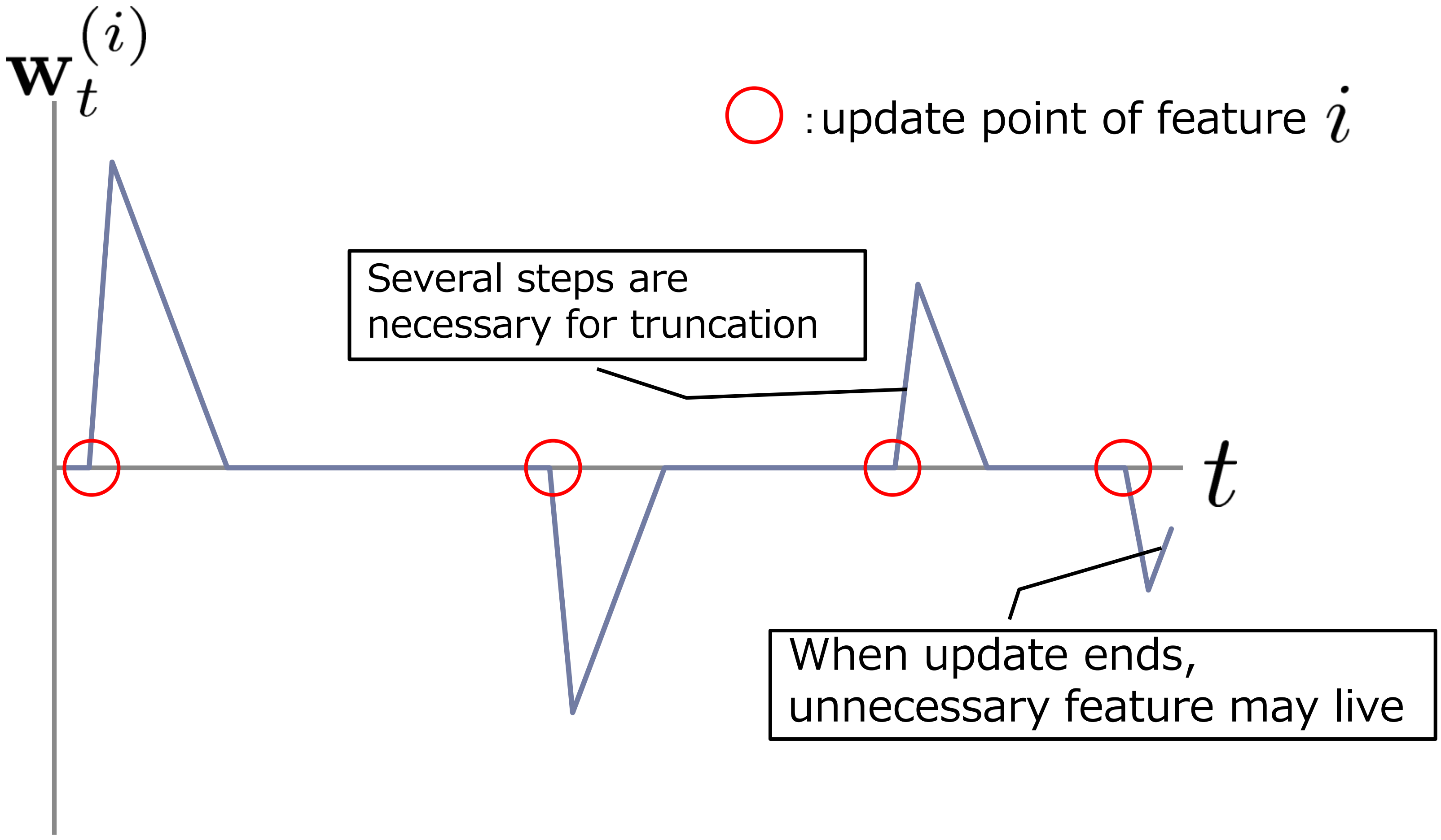


  - Prove regret upper bound $O(\sqrt{T})$
  - Propose FT-FOBOS with Cumulative Penalty
  - Outperform FOBOS in all datasets

Properties of Online Learning

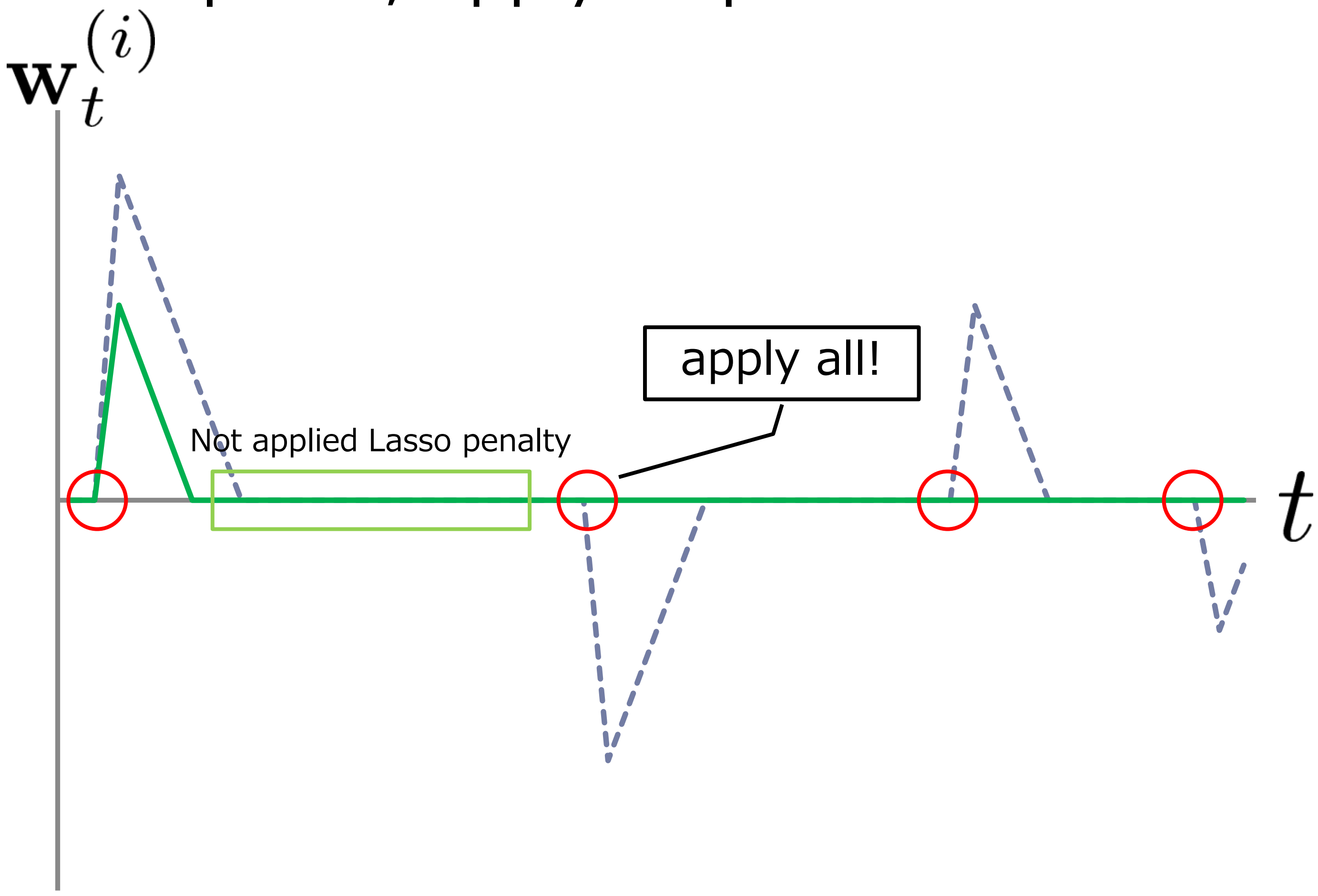
- Be able to run in the condition that only see part of data at a time
  - be able to learn from streaming data
  - don't have to put all data into memory at a time
- Easy to re-learning
  - previously used data are not necessary to re-learn

Additional Problem of FOBOS



Cumulative Penalty [Tsuruoka et al. 2009]

- When update, apply all previous truncation



Frequency-aware Truncated methods with Cumulative Penalty

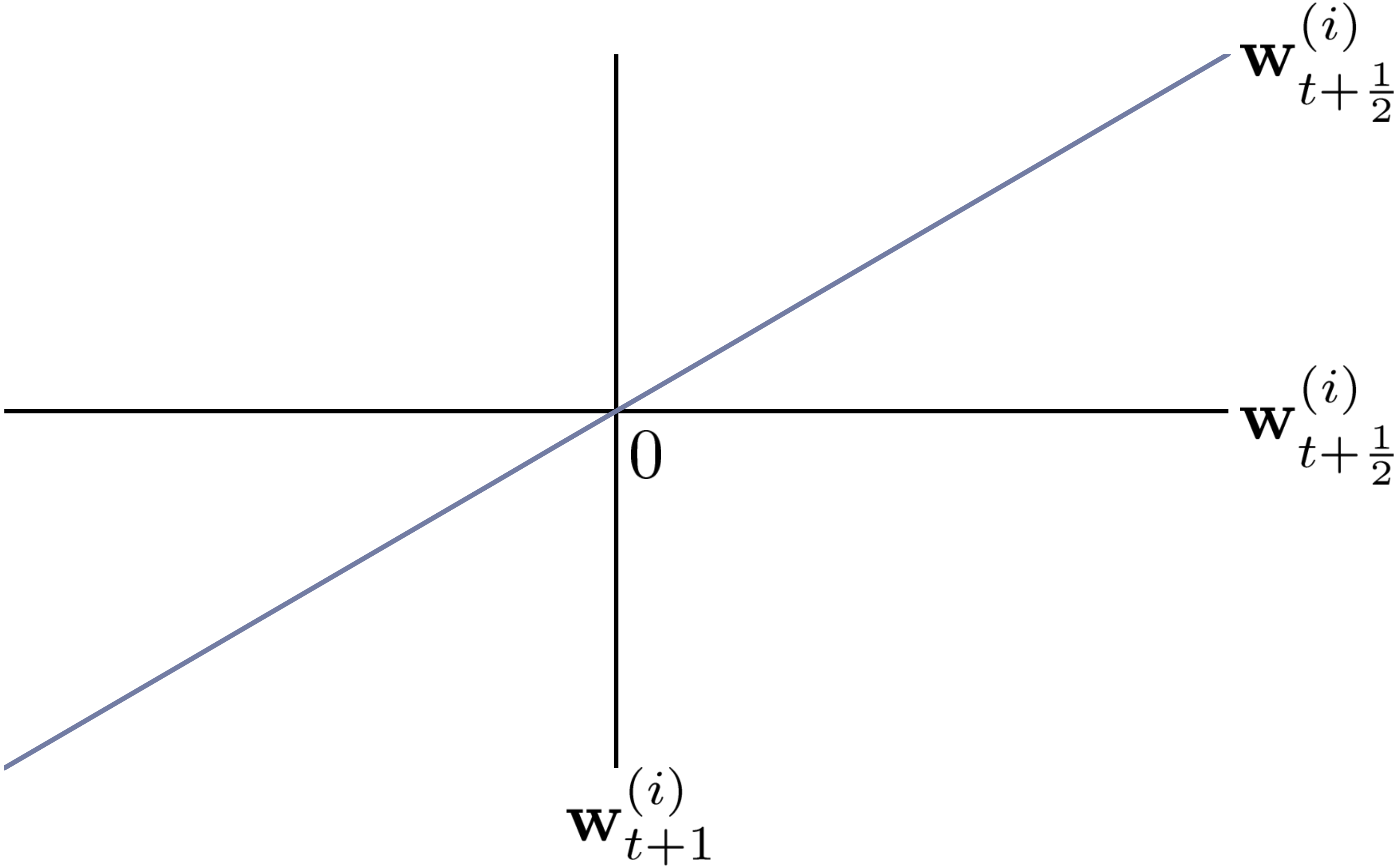
- Combine cumulative penalty framework into FT-FOBOS
  - However, experimental results were not good.

Update formula

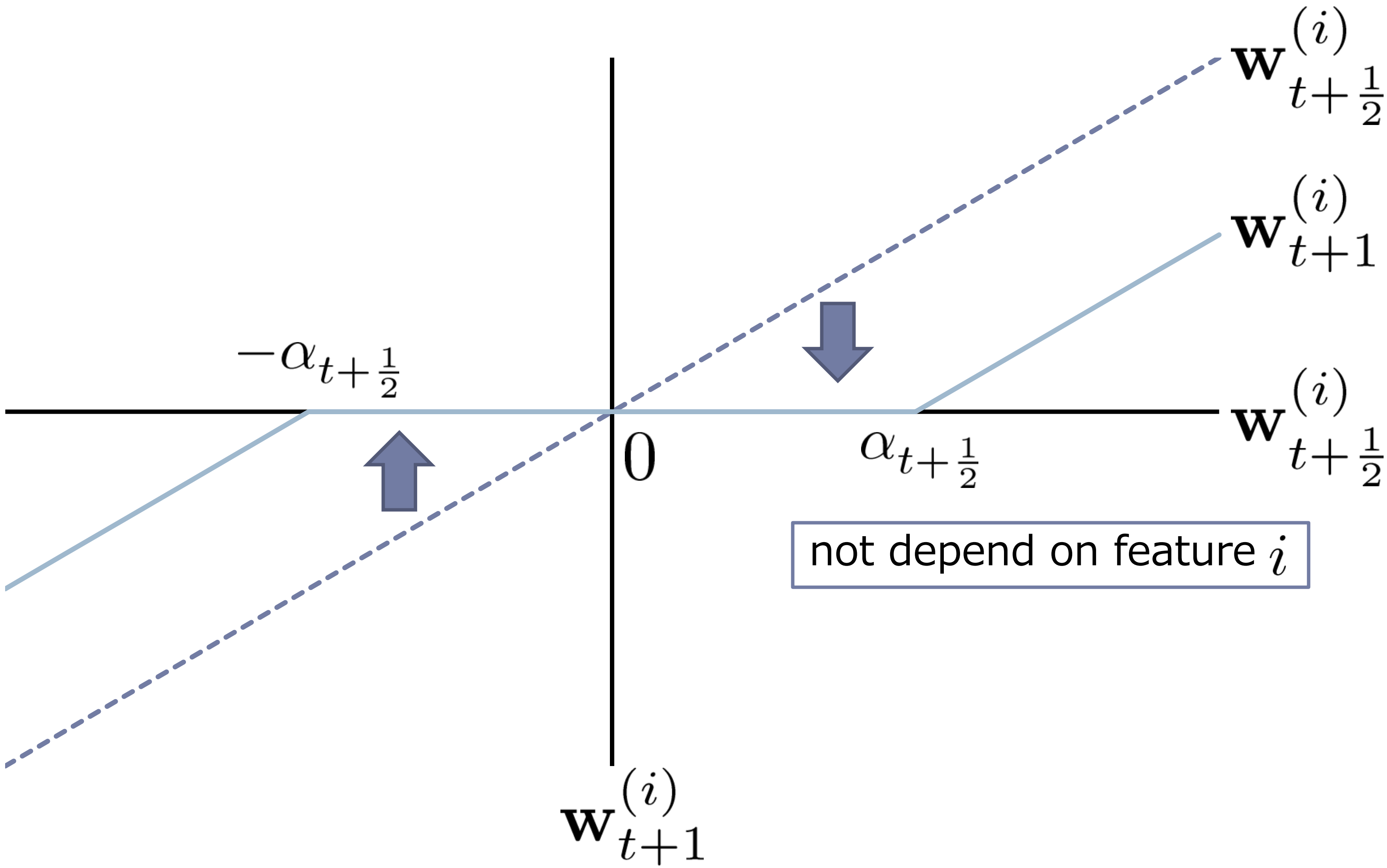
$$w_{t+1}^{(j)} = \begin{cases} \max\left(0, w_{t+1/2}^{(j)} - (h_{t,p}^{(j)} u_t + q_t^{(j)})\right) & w_{t+1/2}^{(j)} \geq 0 \\ \min\left(0, w_{t+1/2}^{(j)} + (h_{t,p}^{(j)} u_t - q_t^{(j)})\right) & w_{t+1/2}^{(j)} < 0 \end{cases}$$

$$q_t^{(j)} = q_{t-1}^{(j)} + (w_{t+1}^{(j)} - w_{t+1/2}^{(j)}) \qquad u_n = \lambda \sum_{t=1}^{n} \eta_{t+1/2}$$

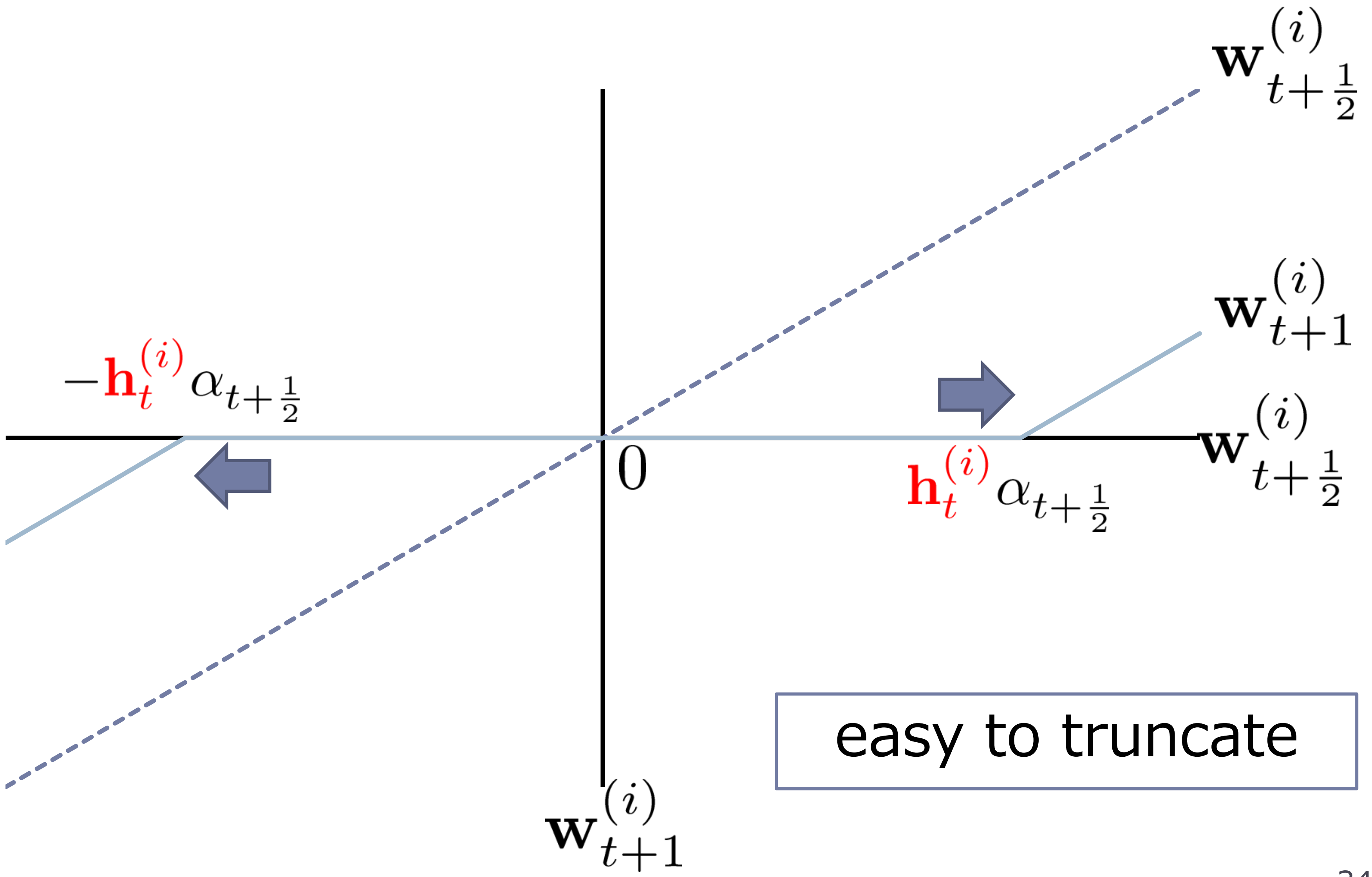
Lasso in FOBOS



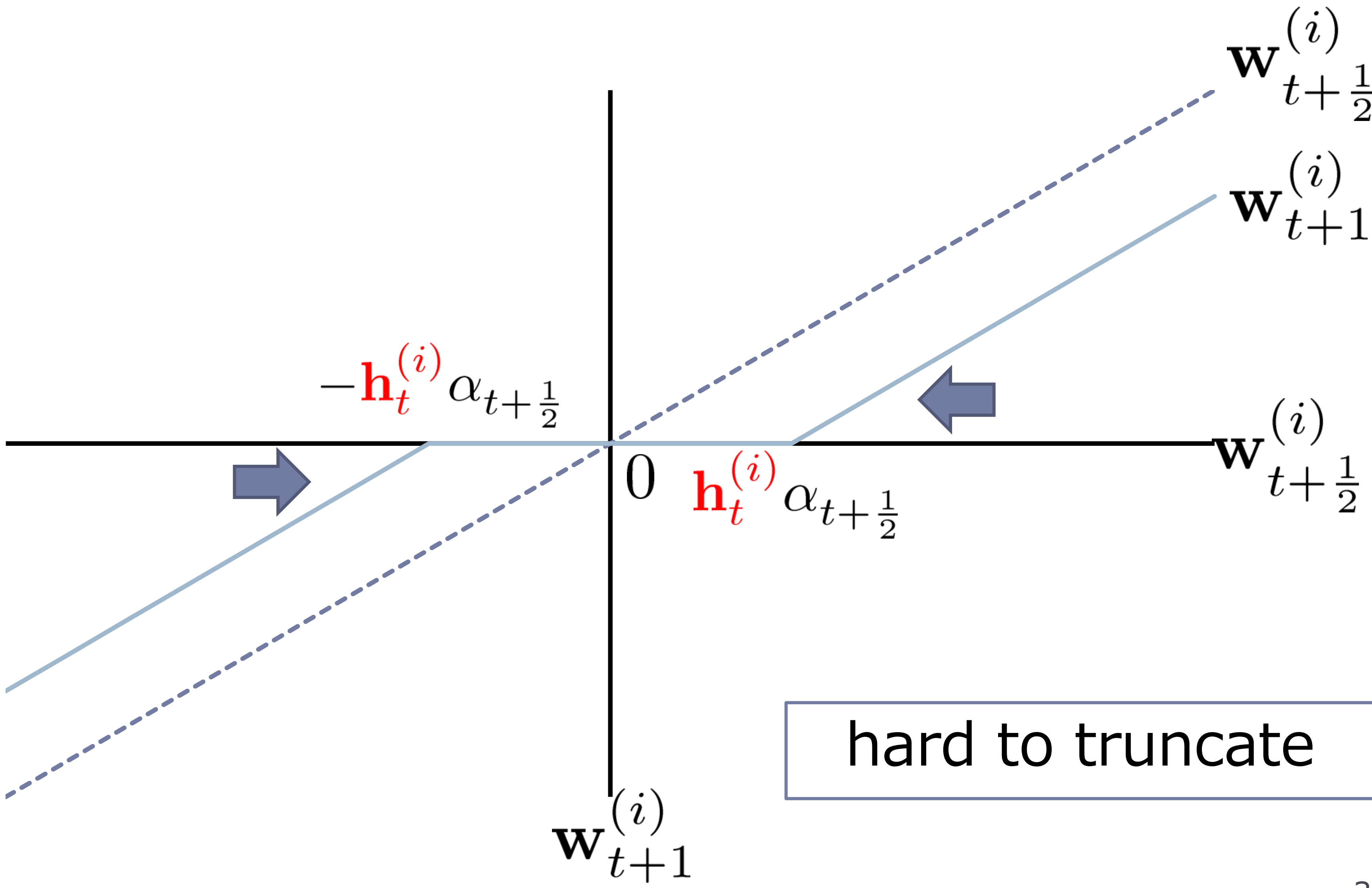
Lasso in FOBOS



Lasso in FT-FOBOS ($\mathbf{h}_t^{(i)}$ is big)



Lasso in FT-FOBOS ($\mathbf{h}_t^{(i)}$ is small)



hard to truncate