

Adaptive Regularization of Weight Vectors

Koby Crammer, Alex Kulesza, Mark Dredze(NIPS 2009)

機械学習勉強会 2010/07/01

東大中川研M1 大岩秀和

Introduction

- ▶ NLPなどの特徴次元数が多い2値分類問題で高い性能を示すOnline学習手法として、**Confidence Weighted Linear Classification** (以下、CW)と呼ばれるアルゴリズムが提案されている
- ▶ しかし、CWはLabel noiseに対して脆弱
 - ▶ Label noise:分類が間違っている教師データ
- ▶ 今回紹介する論文では、訓練例を正しく分類することを最重要としていたCWの問題点を改良したアルゴリズムを提案
 - ▶ Adaptive regularization of Weight Vectors(AROW)
- ▶ Label noiseによる訓練例の急な変化にも頑健
- ▶ state-of-the-artなAlgorithmと遜色ない性能
 - ▶ 特にLabel noiseが存在する場合には、既存手法より非常に高い性能を示した



目次

- ▶ Confidence Weighted Linear Classification(CW)
 - ▶ 既存手法の問題点
 - ▶ CW algorithm
- ▶ Adaptive Regularization of Weight Vectors(AROW)
 - ▶ CWの問題点
 - ▶ Algorithm
 - ▶ Analysis
 - ▶ Experiment
 - ▶ Summary



目次

- ▶ Confidence Weighted Linear Classification(CW)
 - ▶ 既存手法の問題点
 - ▶ CW algorithm
- ▶ Adaptive Regularization of Weight Vectors(AROW)
 - ▶ CWの問題点
 - ▶ Algorithm
 - ▶ Analysis
 - ▶ Experiment
 - ▶ Summary



Online Learning (オンライン学習)

▶ Batch Learning (バッチ学習)

- ▶ 全ての訓練例を受け取ってから、weighted vectorを更新
- ▶ 結果の精度は高いが、収束するまで時間がかかり、実装も複雑

▶ Online Learning (オンライン学習)

- ▶ 訓練例を1つ観測した時点で、weighted vectorを更新
- ▶ 収束が早く、実装も単純、メモリもあまり必要としない

▶ Online Learningの既存手法

- ▶ Perceptron (Rosenblatt[1958])
- ▶ Passive-aggressive (Crammer et al.[2006])



既存手法の問題点

- ▶ NLPなどの分類問題では、特徴次元が非常に大きくなる
- ▶ さらに、多くの特徴が極一部の訓練例にしか出現しない
 - ▶ このような特徴が分類問題において、大きな役割を果たすことも多い
- ▶ 既存手法では、訓練例に特徴Aが出現した時に、特徴Aに対応する重みパラメータが更新される
 - ▶ 頻繁に出現する特徴は、対応するパラメータも頻繁に更新される
 - ▶ 一方、余り出現しない特徴はあまり更新されない
- ▶ しかし、多くの既存手法では、パラメータ更新の際に頻出する単語と稀にしか現れない単語を区別していない



既存手法の問題点 (Ex : Passive Aggressive)

INPUT: aggressiveness parameter $C > 0$

INITIALIZE: $\mathbf{w}_1 = (0, \dots, 0)$

For $t = 1, 2, \dots$

- receive instance: $\mathbf{x}_t \in \mathbb{R}^n$
- predict: $\hat{y}_t = \text{sign}(\mathbf{w}_t \cdot \mathbf{x}_t)$
- receive correct label: $y_t \in \{-1, +1\}$
- suffer loss: $\ell_t = \max\{0, 1 - y_t(\mathbf{w}_t \cdot \mathbf{x}_t)\}$
- update:

スカラー

1. set:

$$\tau_t = \frac{\ell_t}{\|\mathbf{x}_t\|^2} \quad (\text{PA})$$

$$\tau_t = \min \left\{ C, \frac{\ell_t}{\|\mathbf{x}_t\|^2} \right\} \quad (\text{PA-I})$$

$$\tau_t = \frac{\ell_t}{\|\mathbf{x}_t\|^2 + \frac{1}{2C}} \quad (\text{PA-II})$$

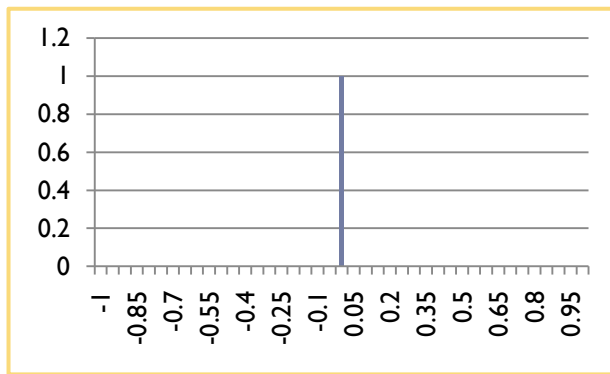
スカラー

2. update: $\mathbf{w}_{t+1} = \mathbf{w}_t + \tau_t y_t \mathbf{x}_t$

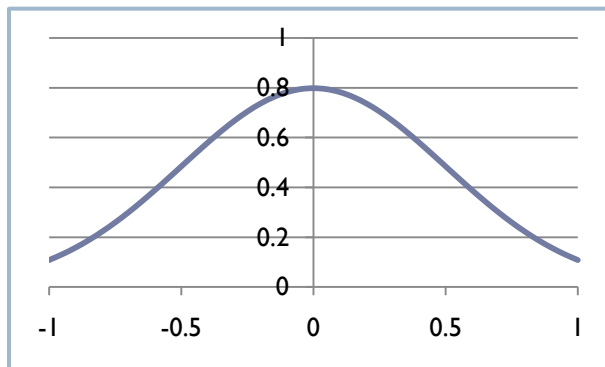
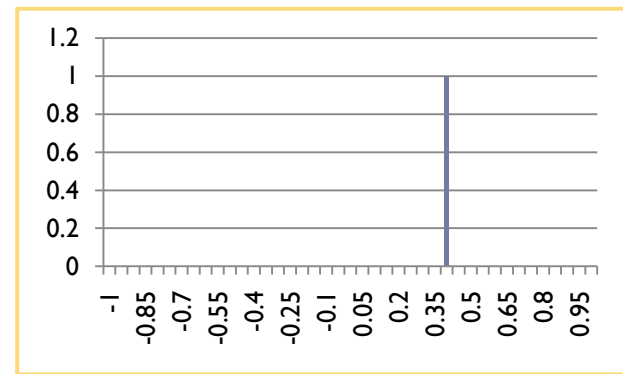
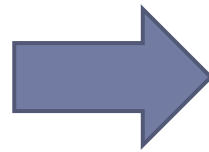
\mathbf{x}_i の出現頻度とは関係なく \mathbf{w}_i が更新される

Confidence-Weighted Algorithm(CW)

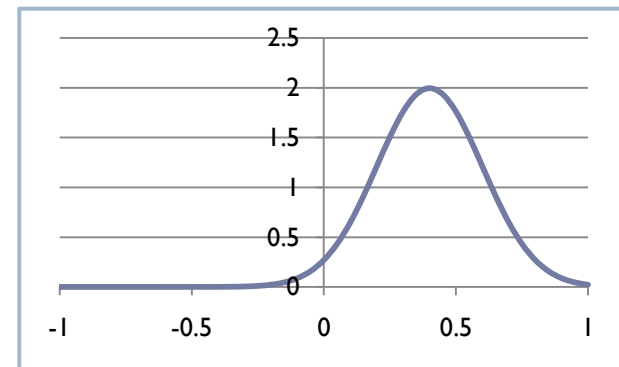
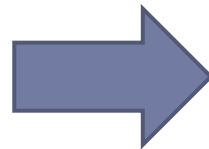
- ▶ 重みベクトル上にガウス分布を導入
- ▶ 訓練例を受け取るごとに、重みベクトルの平均・共分散を更新



既存手法



CW



Confidence-Weighted Algorithm(CW)

- ▶ 分散の大きい(更新回数の少ない)重みパラメータは大きく更新し、分散の小さいパラメータは小さく更新するアルゴリズムを実現
- ▶ Ex.ある本の評価

I liked this author.

- ▶ likedに対応する重みパラメータは上昇

I liked this author, but found the book dull.

- ▶ dullはrareな特徴。likeはfrequentな特徴。
- ▶ dullに対応するパラメータは減少するが、likedに対応するパラメータは減少しない。



Confidence-Weighted Algorithm(CW)

数式の準備

$\mathbf{x} \in \mathbf{R}^d$: 特徴ベクトル

$\boldsymbol{\mu} \in \mathbf{R}^d$: 重みベクトルの平均

$\boldsymbol{\sigma} \in \mathbf{R}^d$: 重みベクトルの分散

$\Sigma \in \mathbf{R}^{d \times d}$: 重みベクトルの共分散

$\mathbf{w} \sim N(\boldsymbol{\mu}, \Sigma)$: 重みベクトル

$y \in \{-1, 1\}$: 正解ラベル

$\eta \in (0.5, 1]$: しきい値



Classification

特徴ベクトル \mathbf{x}_i から重みベクトル \mathbf{w}_i を用いて、 y_i を予測したい

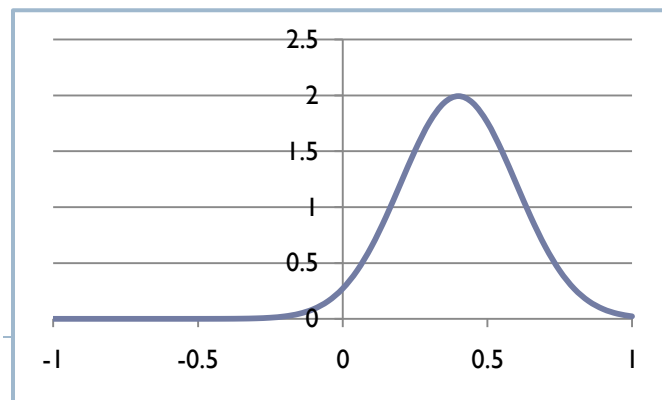
$$\text{sign}(\mathbf{x}_i \cdot \mathbf{w}_i) < 0 \quad y_i = -1$$

$$\text{sign}(\mathbf{x}_i \cdot \mathbf{w}_i) > 0 \quad y_i = 1$$

この時、 \mathbf{x}_i が正しく分類される確率は、

$$M \sim N(y_i(\boldsymbol{\mu}_i \cdot \mathbf{x}_i), \mathbf{x}_i^T \boldsymbol{\Sigma}_i \mathbf{x}_i)$$

$$\Pr_{\mathbf{w} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}[M \geq 0] = \Pr_{\mathbf{w} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}[y_i(\mathbf{w} \cdot \mathbf{x}_i) \geq 0]$$



Update(Constraint function)

以下の式で、最適化を行う

以前の多変量ガウス分布に
最も近いガウス分布を選択する

$$(\boldsymbol{\mu}_{i+1}, \Sigma_{i+1}) = \min D_{KL}(N(\boldsymbol{\mu}, \Sigma) \| N(\boldsymbol{\mu}_i, \Sigma_i))$$

$$s.t. \quad \Pr_{\mathbf{w} \sim N(\boldsymbol{\mu}_i, \Sigma_i)}[y_i(\mathbf{w} \cdot \mathbf{x}_i) \geq 0] \geq \eta$$

誤識別率が $1-\eta$ 以下となるガウス分布の中で

まず、制約式を変形する

$$\Pr[M \leq 0] = \Pr\left[\frac{M - \mu_M}{\sigma_M} \leq \frac{-\mu_M}{\sigma_M}\right] \leq 1 - \eta$$

標準正規分布に従う確率変数

$$\frac{-\mu_M}{\sigma_M} \leq \Phi^{-1}(1 - \eta) = -\Phi^{-1}(\eta)$$

$\Phi(\eta)$: 標準正規分布の累積密度関数

Update(Constraint function)

$$M \sim N(y_i(\boldsymbol{\mu}_i \cdot \mathbf{x}_i), \mathbf{x}_i^T \Sigma_i \mathbf{x}_i) \text{より、}$$

$$\mu_M = y_i(\boldsymbol{\mu}_i \cdot \mathbf{x}_i)$$

$$\sigma_M = \sqrt{\mathbf{x}_i^T \Sigma_i \mathbf{x}_i}$$

$$\frac{-\mu_M}{\sigma_M} \leq -\Phi^{-1}(\eta)$$

$$\Leftrightarrow y_i(\boldsymbol{\mu} \cdot \mathbf{x}_i) \geq \Phi^{-1}(\eta) \sqrt{\mathbf{x}_i^T \Sigma_i \mathbf{x}_i}$$

$$(\boldsymbol{\mu}_{i+1}, \Sigma_{i+1}) = \min D_{KL}(N(\boldsymbol{\mu}, \Sigma) \| N(\boldsymbol{\mu}_i, \Sigma_i))$$

$$s.t. \quad \Pr_{\mathbf{w} \sim N(\boldsymbol{\mu}_i, \Sigma_i)}[y_i(\mathbf{w} \cdot \mathbf{x}_i) \geq 0] \geq \eta$$



Update(Constraint function)

$$y_i(\boldsymbol{\mu} \cdot \mathbf{x}_i) \geq \Phi^{-1}(\eta) \sqrt{\mathbf{x}_i^T \boldsymbol{\Sigma} \mathbf{x}_i}$$

上の形のままでは、 $\boldsymbol{\Sigma}$ について凸ではない
根号を取り除く

$$y_i(\boldsymbol{\mu} \cdot \mathbf{x}_i) \geq \Phi^{-1}(\eta)(\mathbf{x}_i^T \boldsymbol{\Sigma} \mathbf{x}_i)$$

一方、Exact-CW(Dredze et al.[2008])では、
根号の形を維持したままで、
凸性を持つ解析的な形で上式が解ける事を示している

$$\begin{aligned} (\boldsymbol{\mu}_{i+1}, \boldsymbol{\Sigma}_{i+1}) &= \min D_{KL}(N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \| N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)) \\ \text{s.t. } &\Pr_{\mathbf{w} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}[y_i(\mathbf{w} \cdot \mathbf{x}_i) \geq 0] \geq \eta \end{aligned}$$

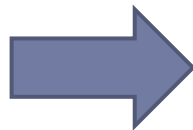
Update(Objective function)

次に、目的関数の変形を行う

$$\begin{aligned}(\boldsymbol{\mu}_{i+1}, \Sigma_{i+1}) &= \min D_{KL}(N(\boldsymbol{\mu}, \Sigma) \| N(\boldsymbol{\mu}_i, \Sigma_i)) \\ &= \min \left\{ \frac{1}{2} \log \left(\frac{\det \Sigma_i}{\det \Sigma} \right) + \frac{1}{2} \text{Tr} \left(\Sigma_i^{-1} \Sigma \right) + \frac{1}{2} (\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \Sigma_i^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}) \right\}\end{aligned}$$

$$(\boldsymbol{\mu}_{i+1}, \Sigma_{i+1}) = \min D_{KL}(N(\boldsymbol{\mu}, \Sigma) \| N(\boldsymbol{\mu}_i, \Sigma_i))$$

$$s.t. \quad \Pr_{\mathbf{w} \sim N(\boldsymbol{\mu}_i, \Sigma_i)} [y_i(\mathbf{w} \cdot \mathbf{x}_i) \geq 0] \geq \eta$$


$$\begin{aligned}& \min \left\{ \frac{1}{2} \log \left(\frac{\det \Sigma_i}{\det \Sigma} \right) + \frac{1}{2} \text{Tr} \left(\Sigma_i^{-1} \Sigma \right) + \frac{1}{2} (\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \Sigma_i^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}) \right\} \\ & s.t. \quad y_i(\boldsymbol{\mu} \cdot \mathbf{x}_i) \geq \Phi^{-1}(\eta)(\mathbf{x}_i^T \Sigma \mathbf{x}_i)\end{aligned}$$

▶ KKT-conditionで解析的に解くことが出来る

Algorithm

Algorithm 1 Variance Algorithm (Approximate)

Input: confidence parameter $\phi = \Phi^{-1}(\eta)$
initial variance parameter $a > 0$

Initialize: $\mu_1 = \mathbf{0}$, $\Sigma_1 = aI$

for $i = 1, 2 \dots$ **do**

Receive $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$

Set the following variables:

α_i as in Lemma 1

$$\mu_{i+1} = \mu_i + \alpha_i y_i \Sigma_i \mathbf{x}_i \quad (11)$$

$$\Sigma_{i+1}^{-1} = \Sigma_i^{-1} + 2\alpha_i \phi \text{diag}(\mathbf{x}_i) \quad (17) \quad \leftarrow \text{対角成分のみ}$$

end for



目次

- ▶ Confidence Weighted Linear Classification(CW)
 - ▶ 既存手法の問題点
 - ▶ CW algorithm
- ▶ Adaptive Regularization of Weight Vectors(AROW)
 - ▶ CWの問題点
 - ▶ Algorithm
 - ▶ Analysis
 - ▶ Experical Evaluation
 - ▶ Summary



CWの問題点

▶ CWのUpdateは非常にAggressive

- ▶ 必ず、 $\Pr_{\mathbf{w} \sim N(\boldsymbol{\mu}_i, \Sigma_i)}[y_i(\mathbf{w} \cdot \mathbf{x}_i) \geq 0]$ となるように更新
- ▶ $\Rightarrow y_i(\boldsymbol{\mu} \cdot \mathbf{x}_i) \geq \Phi^{-1}(\eta)\sqrt{\mathbf{x}_i^T \Sigma \mathbf{x}_i}$ より、 $\boldsymbol{\mu}$ が正しく分類されるように更新
- ▶ したがって、**Label noiseに弱く**、過学習を起こしやすい

▶ CWでは教師データが線形分離可能な状態を仮定

- ▶ CW(Exact-CW)のMistake Boundは、線形分離可能な場合しか保証されない
- ▶ しかし、CWの形を維持したままでは、線形分離不可能な場合にも対応できるように制約条件を緩めるのが困難



AROWの特徴

- ▶ Online学習での各既存手法の特徴をいいとこ取り
 - ▶ Large margin training
 - ▶ Non-mistakeの場合でもUpdate
 - ▶ Confidence weighting
 - ▶ 更新回数の少ない重みベクトルをより大きく更新
 - ▶ Handling non-separable data
 - ▶ 線形分離不可能なデータに対しても、精度があまり落ちない
- ▶ 最適化関数
 - ▶ Label noiseにも頑健
 - ▶ 教師データの線形分離性を仮定せずに、mistake boundを導出



Algorithm(AROW)

$$C(\boldsymbol{\mu}, \Sigma) = D_{KL}(N(\boldsymbol{\mu}, \Sigma) \| N(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})) + \lambda_1 \ell_{h^2}(y_t, \boldsymbol{\mu} \cdot \mathbf{x}_t) + \lambda_2 \mathbf{x}_t^T \Sigma \mathbf{x}_t$$

$$\ell_{h^2}(y_t, \boldsymbol{\mu} \cdot \mathbf{x}_t) = (\max\{0, 1 - y_t(\boldsymbol{\mu} \cdot \mathbf{x}_t)\})^2$$

- ▶ 第一項-- $D_{KL}(N(\boldsymbol{\mu}, \Sigma) \| N(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1}))$
 - ▶ 以前のパラメータから大きく更新しない
- ▶ 第二項-- $\ell_{h^2}(y_t, \boldsymbol{\mu} \cdot \mathbf{x}_t)$
 - ▶ 損失関数を最小にする
 - ▶ Hinge-loss以外の損失関数でも良い
- ▶ 第三項-- $\mathbf{x}_t^T \Sigma \mathbf{x}_t$
 - ▶ 学習するにつれ、 Σ を小さくする

$\mathbf{x} \in \mathbf{R}^d$: 特徴ベクトル

$\boldsymbol{\mu} \in \mathbf{R}^d$: 重みベクトルの平均

$\boldsymbol{\sigma} \in \mathbf{R}^d$: 重みベクトルの分散

$\Sigma \in \mathbf{R}^{d \times d}$: 重みベクトルの共分散

$\mathbf{w} \sim N(\boldsymbol{\mu}, \Sigma)$: 重みベクトル

$y \in \{-1, 1\}$: 正解ラベル

$\eta \in (0.5, 1]$: しきい値

λ_1, λ_2 : hyperparameters

Update

第一項のKL-divergenceを分解

$$C(\boldsymbol{\mu}, \Sigma) = \frac{1}{2} \log \left(\frac{\det \Sigma_{t-1}}{\det \Sigma} \right) + \frac{1}{2} \text{Tr}(\Sigma_{t-1}^{-1} \Sigma) + \frac{1}{2} (\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu})^T \Sigma_{t-1}^{-1} (\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}) - \frac{d}{2} \\ + \frac{1}{2r} \ell_{h^2}(y_t, \boldsymbol{\mu} \cdot \mathbf{x}_t) + \frac{1}{2r} \mathbf{x}_t^T \Sigma \mathbf{x}_t \quad \leftarrow \quad \boldsymbol{\mu}, \Sigma \text{ は分離可能} \quad (\lambda_1 = \lambda_2 = \frac{1}{2r})$$

$\boldsymbol{\mu}, \Sigma$ それぞれ独立に、argminを求めればよい

1. $\boldsymbol{\mu}_t$ を更新

$$\boldsymbol{\mu}_t = \arg \min_{\boldsymbol{\mu}} C(\boldsymbol{\mu}, \Sigma)$$

2. $\boldsymbol{\mu}_t \neq \boldsymbol{\mu}_{t-1}$ のとき、 Σ を更新

$$\Sigma_t = \arg \min_{\Sigma} C(\boldsymbol{\mu}, \Sigma)$$

Update($\boldsymbol{\mu}$)

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}} C(\boldsymbol{\mu}, \Sigma) &= -\Sigma_{t-1}^{-1} (\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}) + \frac{1}{2r} \left[\frac{\partial}{\partial z} \ell_{h^2}(y_t, z) \Big|_{z=\boldsymbol{\mu} \cdot \mathbf{x}_t} \right] \mathbf{x}_t = 0 \\ \Leftrightarrow \begin{cases} \boldsymbol{\mu} = \boldsymbol{\mu}_{t-1} + \frac{y_t}{2r} (1 - y_t (\boldsymbol{\mu} \cdot \mathbf{x}_t)) \Sigma_{t-1} \mathbf{x}_t & (1 - y_t (\boldsymbol{\mu} \cdot \mathbf{x}_t) \geq 0) \\ \boldsymbol{\mu} = \boldsymbol{\mu}_{t-1} & (\text{otherwise}) \end{cases} \end{aligned}$$

両辺で \mathbf{x}_t との内積を取り、 $\boldsymbol{\mu} \cdot \mathbf{x}_t$ を求めて、上式に代入

$$\boldsymbol{\mu} = \boldsymbol{\mu}_{t-1} + \frac{\max(0, 1 - y_t (\mathbf{x}_t^T \boldsymbol{\mu}_{t-1}))}{\mathbf{x}_t^T \Sigma_{t-1} \mathbf{x}_t + r} \Sigma_{t-1} y_t \mathbf{x}_t$$



Update(Σ)

$$\begin{aligned}\frac{\partial}{\partial \Sigma} C(\boldsymbol{\mu}, \Sigma) &= -\frac{1}{2} \Sigma^{-1} + \frac{1}{2} \Sigma_{t-1}^{-1} + \frac{1}{2r} \mathbf{x}_t \mathbf{x}_t^T \\ &\Leftrightarrow \Sigma^{-1} = \Sigma_{t-1}^{-1} + \frac{\mathbf{x}_t \mathbf{x}_t^T}{r}\end{aligned}$$

Woodbury identityを適用する

$$\begin{aligned}(\Sigma^{-1})^{-1} &= \Sigma_{t-1} - \Sigma_{t-1} \mathbf{x}_t (r + \mathbf{x}_t^T \Sigma_{t-1} \mathbf{x}_t)^{-1} \mathbf{x}_t^T \Sigma_{t-1} \\ &= \Sigma_{t-1} - \frac{\Sigma_{t-1} \mathbf{x}_t \mathbf{x}_t^T \Sigma_{t-1}}{r + \mathbf{x}_t^T \Sigma_{t-1} \mathbf{x}_t}\end{aligned}$$

対角項に着目すると、
 Σ の固有値は単調減少していることが分かる



Algorithm

Input parameters r

Initialize $\mu_0 = 0$, $\Sigma_0 = I$,

For $t = 1, \dots, T$

- Receive a training example $x_t \in \mathbb{R}^d$
- Compute margin and confidence $m_t = \mu_{t-1}^\top x_t$ $v_t = x_t^\top \Sigma_{t-1} x_t$
- Receive true label y_t , and suffer loss $\ell_t = 1$ if $\text{sign}(m_t) \neq y_t$
- If $m_t y_t < 1$, update using eqs. (7) & (9):

$$\mu_t = \mu_{t-1} + \alpha_t \Sigma_{t-1} y_t x_t$$

$$\Sigma_t = \Sigma_{t-1} - \beta_t \Sigma_{t-1} x_t x_t^\top \Sigma_{t-1}$$

$$\beta_t = \frac{1}{x_t^\top \Sigma_{t-1} x_t + r}$$

$$\alpha_t = \max\left(0, 1 - y_t x_t^\top \mu_{t-1}\right) \beta_t$$

Output: Weight vector μ_T and confidence Σ_T .



Analysis (Representer Theorem)

Lemma 1 (Representer Theorem) Assume that $\Sigma_0 = I$ and $\mu_0 = \mathbf{0}$. The mean parameters μ_t and confidence parameters Σ_t produced by updating via (7) and (9) can be written as linear combinations of the input vectors (resp. outer products of the input vectors with themselves) with coefficients depending only on inner-products of input vectors.

- ▶ μ, Σ は、それぞれ入力ベクトルの線形和・入力ベクトル同士の外積の線形和で表現することが出来る
- ▶ 線形和の係数は、入力ベクトル同士の内積にのみ依存する
- ▶ Proof sketch

$$\mu_t = \sum_p^{i-1} v_p^{(i)} \mathbf{x}_p \quad \Sigma_t = \sum_{p,q=1}^{i-1} \pi_{p,q}^{(i)} \mathbf{x}_p \mathbf{x}_q^T + aI$$

とにおいて、帰納法により確認することが出来る

Theorem 2(Mistake Bound)

► Mistake bound

Theorem 2 For any reference weight vector $\mathbf{u} \in \mathbb{R}^d$, the number of mistakes made by AROW (Fig. 1) is upper bounded by

$$M \leq \sqrt{r \|\mathbf{u}\|^2 + \mathbf{u}^\top \mathbf{X}_A \mathbf{u}} \sqrt{\log \left(\det \left(I + \frac{1}{r} \mathbf{X}_A \right) \right)} + U + \sum_{t \in M \cup U} g_t - U, \quad (10)$$

where $g_t = \max(0, 1 - y_t \mathbf{u}^\top \mathbf{x}_t)$.

$$\mathbf{X}_A = \mathbf{X}_M + \mathbf{X}_U$$

$$\mathbf{X}_M : \sum_{i \in M} \mathbf{x}_i \mathbf{x}_i^T \quad [y_t(\boldsymbol{\mu}_{t-1} \cdot \mathbf{x}_t) \leq 0] \quad \text{Mistake}$$

$$\mathbf{X}_U : \sum_{i \in U} \mathbf{x}_i \mathbf{x}_i^T \quad [0 < y_t(\boldsymbol{\mu}_{t-1} \cdot \mathbf{x}_t) \leq 1] \quad \text{Not mistake but update}$$

$$M = |M| \quad U = |U|$$



Theorem 2(Mistake Bound)

$$M \leq \sqrt{r \|u\|^2 + u^\top \mathbf{X}_A u} \sqrt{\log \left(\det \left(I + \frac{1}{r} \mathbf{X}_A \right) \right)} + U + \sum_{t \in M \cup U} g_t - U$$

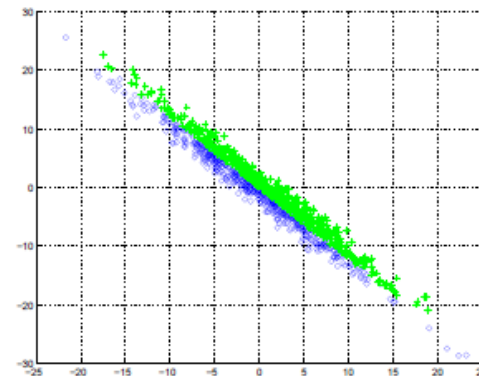
- ▶ r は最適なboundの調節に用いることが考えられる
 - ▶ r は左の根号式を単調増加、右の根号式を単調減少させる
 - ▶ ただし、 \mathbf{X}_A などの項も r に影響を受けるため直接評価することは不可能
- ▶ $U = \phi$ の場合、second-order perceptronと同じboundで押さえることが可能



Experical Evaluation

▶ 人工データによる実験

- ▶ 20次元の実数ベクトル
- ▶ ある2つの次元は、 45° 傾けた分散1のガウス分布
- ▶ ラベルは、上の楕円の長軸より上か下かで決める
- ▶ 他の18次元は、それぞれ平均0,分散2の独立したガウス分布
- ▶ データにLabel noiseを加える



▶ 実データによる実験

- ▶ Amazon(category classification)
- ▶ 20 Newsgroups(newsgroups classification)
- ▶ Reuters(news category classification)
- ▶ Sentiment(positive/negative)
- ▶ ECML/PKDD Challenge(spam/ham)
- ▶ OCR[MNIST/USPS](Digit recognition)
- ▶ 全てのデータセットに対してLabel noiseを加える

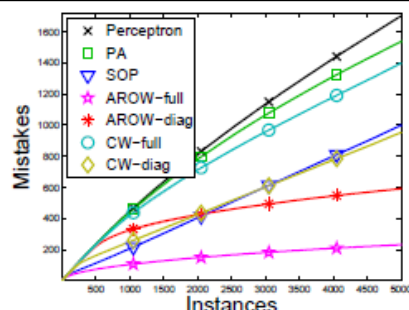
Experical Evaluation

実験結果

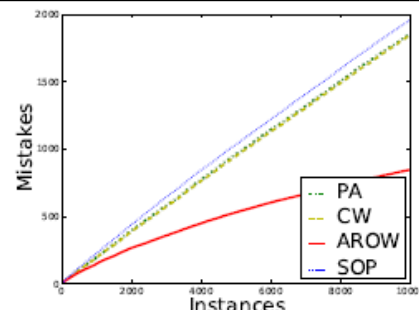
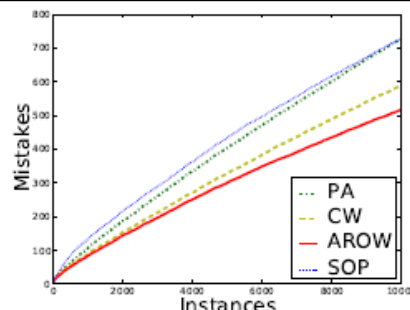
Noise level別の平均順位

Algorithm	Noise level					
	0.0	0.05	0.1	0.15	0.2	0.3
<i>AROW</i>	1.51	1.44	1.38	1.42	1.25	1.25
<i>CW</i>	1.63	1.87	1.95	2.08	2.42	2.76
<i>PA</i>	2.95	2.83	2.78	2.61	2.33	2.08
<i>SOP</i>	3.91	3.87	3.89	3.89	4.00	3.91

Table 1: Mean rank (out of 4, over all datasets) at different noise levels. A rank of 1 indicates that an algorithm outperformed all the others.



(a) synthetic data



(b) MNIST data

Figure 2: Learning curves for AROW (full/diagonal) and baseline methods. (a) 5k synthetic training examples and 10k test examples (10% noise, 100 runs). (b) MNIST 3 vs. 5 binary classification task for different amounts of label noise (left: 0 noise, right: 10%).

Experimental Evaluation

► CW vs AROW

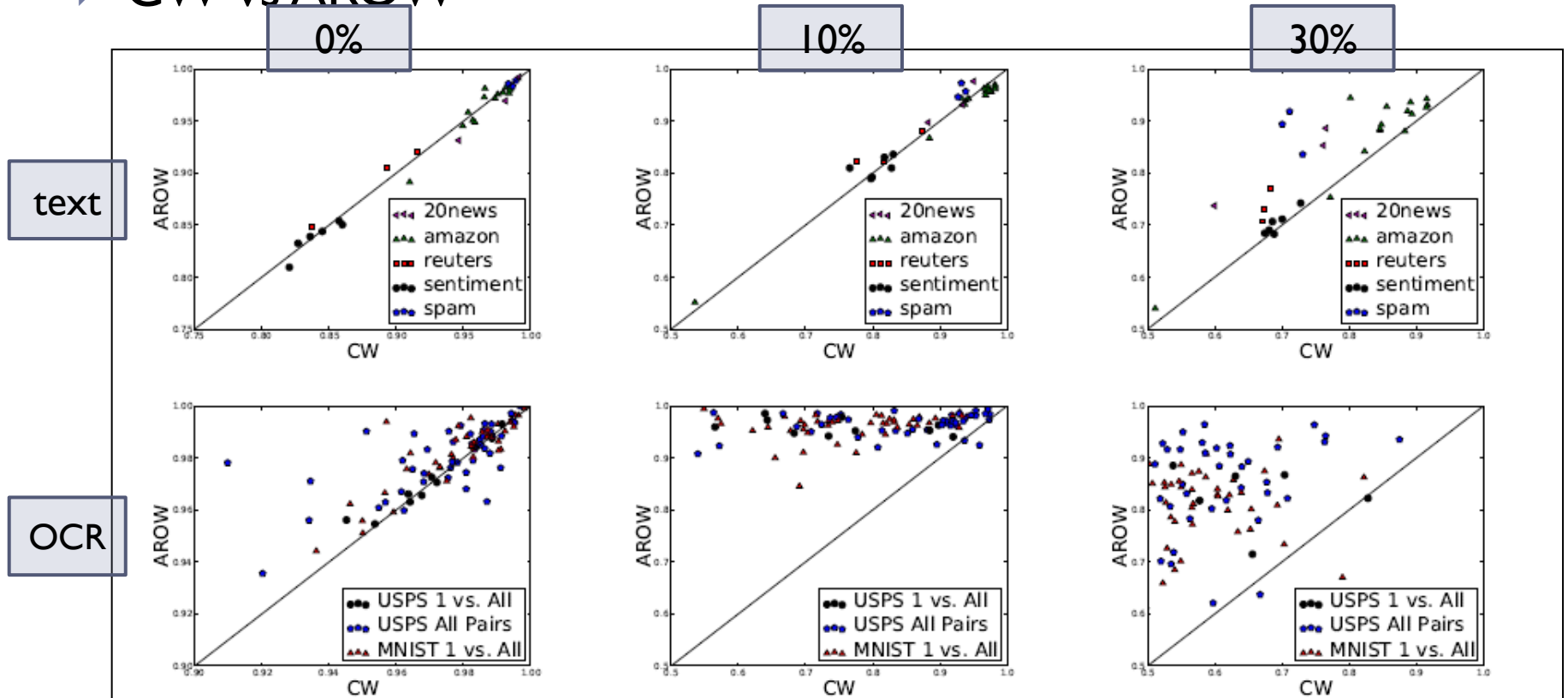


Figure 3: Accuracy on text (top) and OCR (bottom) binary classification. Plots compare performance between AROW and CW, the best performing baseline (Table 1). Markers above the line indicate superior AROW performance and below the line superior CW performance. Label noise increases from left to right: 0%, 10% and 30%. AROW improves relative to CW as noise increases.

Experical Evaluation

► Algorithm property review

Algorithm	Noise level					
	0.0	0.05	0.1	0.15	0.2	0.3
<i>AROW</i>	1.51	1.44	1.38	1.42	1.25	1.25
<i>CW</i>	1.63	1.87	1.95	2.08	2.42	2.76
<i>PA</i>	2.95	2.83	2.78	2.61	2.33	2.08
<i>SOP</i>	3.91	3.87	3.89	3.89	4.00	3.91

Table 1: Mean rank (out of 4, over all datasets) at different noise levels. A rank of 1 indicates that an algorithm outperformed all the others.

<i>Algorithm</i>	<i>Large Margin</i>	<i>Confidence</i>	<i>Non-Separable</i>	<i>Adaptive Margin</i>
<i>PA</i>	Yes	No	Yes	No
<i>SOP</i>	No	Yes	Yes	No
<i>CW</i>	Yes	Yes	No	Yes
<i>AROW</i>	Yes	Yes	Yes	No

Table 2: Online algorithm properties overview.

Summary

- ▶ AROWは、SOPやCWの長所を利用したアルゴリズム
 - ▶ SOPは、mistakeの場合しか更新を行わない
 - ▶ CWの制約条件を緩和(non-separable)
- ▶ Hazan[2008]でも、勾配降下法に対して、Confidenceに近い概念を利用して対数リグレットを示している
 - ▶ AROWでは、勾配降下法などをせずに直接最適化問題を解いている
 - ▶ AROWでは、Loss Boundを直接求めている
 - ▶ Hazan[2008]では、重みベクトルの更新をした後に、重みベクトルの正規化を行っている

ONLINE GRADIENT DESCENT.

Inputs: convex set $\mathcal{P} \subset \mathbb{R}^n$, step sizes $\eta_1, \eta_2, \dots \geq 0$, initial $\mathbf{x}_1 \in \mathcal{P}$.

- In iteration 1, use point $\mathbf{x}_1 \in \mathcal{P}$.
- In iteration $t > 1$: use point

$$\mathbf{x}_t = \Pi_{\mathcal{P}}(\mathbf{x}_{t-1} - \eta_t \nabla f_{t-1}(\mathbf{x}_{t-1}))$$

Here, $\Pi_{\mathcal{P}}$ denotes the *projection* onto nearest point in \mathcal{P} , $\Pi_{\mathcal{P}}(\mathbf{y}) = \arg \min_{\mathbf{x} \in \mathcal{P}} \|\mathbf{x} - \mathbf{y}\|_2$.

▶ 今後の課題

- ▶ AROWのmulti-classへの適用
- ▶ 回帰問題への適用(RLSと同じアルゴリズムとなる)
- ▶ 第3項 $\mathbf{x}_t^T \sum \mathbf{x}$ の変形



補足(多変量ガウス分布のKL-divergence)

$$\begin{aligned}
 KLD &= \int_{-\infty}^{\infty} \frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \cdot \exp \left[-\frac{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}{2} \right] \cdot \log \left(\frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \cdot \exp \left[-\frac{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}{2} \right] \right) \\
 &\quad - \frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \cdot \exp \left[-\frac{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}{2} \right] \cdot \log \left(\frac{|\Sigma_i|^{-\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \cdot \exp \left[-\frac{(\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i)}{2} \right] \right) \cdot d\mathbf{x} \\
 &= \frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \log \left(\frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \right) \int_{-\infty}^{\infty} \exp \left[-\frac{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}{2} \right] \cdot d\mathbf{x} \\
 &\quad - \frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \log \left(\frac{|\Sigma_i|^{-\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \right) \int_{-\infty}^{\infty} \exp \left[-\frac{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}{2} \right] \cdot d\mathbf{x} \\
 &\quad + \frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \int_{-\infty}^{\infty} -\frac{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}{2} \exp \left[-\frac{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}{2} \right] \cdot d\mathbf{x} \\
 &\quad - \frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \int_{-\infty}^{\infty} -\frac{(\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i)}{2} \exp \left[-\frac{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}{2} \right] \cdot d\mathbf{x} \\
 &= \frac{1}{2} \log \frac{|\Sigma_i|}{|\Sigma|} + \frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \text{Tr} \left(\Sigma^{-1} \int_{-\infty}^{\infty} -\frac{(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T}{2} \exp \left[-\frac{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}{2} \right] \cdot d\mathbf{x} \right) \\
 &\quad - \frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \text{Tr} \left(\Sigma_i^{-1} \int_{-\infty}^{\infty} -\frac{(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T}{2} \exp \left[-\frac{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}{2} \right] \cdot d\mathbf{x} \right) \\
 &\quad + \frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \int_{-\infty}^{\infty} -\frac{(\mu_i - \mu)^T \Sigma_i^{-1} (\mu_i - \mu)}{2} \exp \left[-\frac{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}{2} \right] \cdot d\mathbf{x} + (\text{定数項}) \\
 &= \frac{1}{2} \log \frac{|\Sigma_i|}{|\Sigma|} + \frac{1}{2} \text{tr}(\Sigma_i^{-1} \Sigma) + \frac{1}{2} (\mu_i - \mu)^T \Sigma_i^{-1} (\mu_i - \mu) + (\text{定数項})
 \end{aligned}$$

補足(Lemma 1 Proof)

The proof proceeds by induction. The initial parameters $\mu_1 = 0$ and $\Sigma_1 = aI$ can be trivially written in the desired form. For the induction step we first substitute (16) in (8) and get,

$$\mu_{i+1} = \mu_i + \alpha_i y_i \Sigma_i x_i = \sum_{p=1}^{i-1} \left(\nu_p^{(i)} + \alpha_i y_i \sum_{q=1}^{i-1} \pi_{p,q}^{(i)} x_q^\top x_i \right) x_p + a x_i ,$$

which is of the desired form with

$$\nu_i^{(i+1)} = a \quad \text{and} \quad \nu_p^{(i+1)} = \nu_p^{(i)} + \alpha_i y_i \sum_{q=1}^{i-1} \pi_{p,q}^{(i)} x_q^\top x_i \quad \text{for } p < i . \quad (20)$$

A similar elementary calculation can be done for the covariance to obtain

$$\pi_{p,q}^{(i+1)} = -\beta_i \sum_{r,s} \pi_{p,r}^{(i)} \pi_{s,q}^{(i)} x_r^\top x_s + \pi_{p,q}^{(i)} , \quad \pi_{p,i}^{(i+1)} = \pi_{i,p}^{(i+1)} = -\beta_i a \sum_{p,r=1}^{i-1} \pi_{p,r}^{(i)} (x_r^\top x_i) , \quad \pi_{i,i}^{(i+1)} = -\beta_i a^2 , \quad (21)$$

for $p = 1 \dots i-1$, where

$$\beta_i = (\alpha_i \phi) / \left(\sqrt{x_i^\top \Sigma_{i+1} x_i} + (x_i^\top \Sigma_i x_i) \alpha_i \phi \right) = (\alpha_i \phi) / (\sqrt{u_i} + v_i \alpha_i \phi) . \quad (22)$$

Finally, we show that the coefficients $\{\nu_p^{(i)}\}$ and $\{\pi_{p,q}^{(i)}\}$ depend on the data only through inner products. From (11) we have that both m_i and v_i can be written only using inner products. From (14), α_i can also be written as a function of inner products, which in turn, together with (12) implies that u_i can be written that way. Therefore, β_i can also be written as a function of inner products. Finally, using (20) and (21) we conclude that $\{\nu_p^{(i)}\}$ and $\{\pi_{p,q}^{(i)}\}$ depend on the data only through inner products.

補足(Theorem2 Proof)

Lemma 3 Let $\ell_t = \max(0, 1 - y_t \mu_{t-1}^\top x_t)$ and $\chi_t = x_t^\top \Sigma_{t-1} x_t$. Then, for every $t \in \mathcal{M} \cup \mathcal{U}$,

$$\begin{aligned} u^\top \Sigma_t^{-1} \mu_t &= u^\top \Sigma_{t-1}^{-1} \mu_{t-1} + \frac{y_t u^\top x_t}{r} \\ \mu_t^\top \Sigma_t^{-1} \mu_t &= \mu_{t-1}^\top \Sigma_{t-1}^{-1} \mu_{t-1} + \frac{\chi_t + r - \ell_t^2 r}{r(\chi_t + r)} \end{aligned}$$

Lemma 4 Let T be the number of rounds. Then

$$\sum_t \frac{\chi_t r}{r(\chi_t + r)} \leq \log(\det(\Sigma_{T+1}^{-1})) .$$

Proof: We compute the following quantity:

$$x_t^\top \Sigma_t x_t = x_t^\top (\Sigma_{t-1} - \beta_t \Sigma_{t-1} x_t x_t^\top \Sigma_{t-1}) x_t = \chi_t - \frac{\chi_t^2}{\chi_t + r} = \frac{\chi_t r}{\chi_t + r} .$$

Using Lemma D.1 from [2] we have that

$$\frac{1}{r} x_t^\top \Sigma_t x_t = 1 - \frac{\det(\Sigma_{t-1}^{-1})}{\det(\Sigma_t^{-1})} . \quad (11)$$

Combining, we get

$$\sum_t \frac{\chi_t r}{r(\chi_t + r)} = \sum_t \left(1 - \frac{\det(\Sigma_{t-1}^{-1})}{\det(\Sigma_t^{-1})} \right) \leq - \sum_t \log \left(\frac{\det(\Sigma_{t-1}^{-1})}{\det(\Sigma_t^{-1})} \right) \leq \log(\det(\Sigma_{T+1}^{-1})) .$$

補足(Theorem2 Proof)

Proof: We iterate the first equality of Lemma 3 to get

$$u^\top \Sigma_T^{-1} \mu_T = \sum_{t \in \mathcal{M} \cup \mathcal{U}} \frac{y_t u^\top x_t}{r} \geq \sum_{t \in \mathcal{M} \cup \mathcal{U}} \frac{1 - g_t}{r} = \frac{M + U}{r} - \frac{1}{r} \sum_{t \in \mathcal{M} \cup \mathcal{U}} g_t. \quad (12)$$

We iterate the second equality to get

$$\mu_T^\top \Sigma_T^{-1} \mu_T = \sum_{t \in \mathcal{M} \cup \mathcal{U}} \frac{\chi_t + r - \ell_t^2 r}{r(\chi_t + r)} = \sum_{t \in \mathcal{M} \cup \mathcal{U}} \frac{\chi_t}{r(\chi_t + r)} + \sum_{t \in \mathcal{M} \cup \mathcal{U}} \frac{1 - \ell_t^2}{\chi_t + r}. \quad (13)$$

Using Lemma 4 we have that the first term of (13) is upper bounded by $\frac{1}{r} \log(\det(\Sigma_T^{-1}))$. For the second term in (13) we consider two cases. First, if a mistake occurred on example t , then we have that $y_t(x_t \cdot \mu_{t-1}) \leq 0$ and $\ell_t \geq 1$, so $1 - \ell_t^2 \leq 0$. Second, if an the algorithm made an update (but no mistake) on example t , then $0 < y_t(x_t \cdot \mu_{t-1}) \leq 1$ and $\ell_t \geq 0$, thus $1 - \ell_t^2 \leq 1$. We therefore have

$$\sum_{t \in \mathcal{M} \cup \mathcal{U}} \frac{1 - \ell_t^2}{\chi_t + r} \leq \sum_{t \in \mathcal{M}} \frac{0}{\chi_t + r} + \sum_{t \in \mathcal{U}} \frac{1}{\chi_t + r} = \sum_{t \in \mathcal{U}} \frac{1}{\chi_t + r}. \quad (14)$$

Combining and plugging into the Cauchy-Schwarz inequality

$$u^\top \Sigma_T^{-1} \mu_T \leq \sqrt{u^\top \Sigma_T^{-1} u} \sqrt{\mu_T^\top \Sigma_T^{-1} \mu_T},$$

we get

$$\frac{M + U}{r} - \frac{1}{r} \sum_{t \in \mathcal{M} \cup \mathcal{U}} g_t \leq \sqrt{u^\top \Sigma_T^{-1} u} \sqrt{\frac{1}{r} \log(\det(\Sigma_T^{-1})) + \sum_{t \in \mathcal{U}} \frac{1}{\chi_t + r}}. \quad (15)$$

補足(Theorem2 Proof)

Rearranging the terms and using the fact that $\chi_t \geq 0$ yields

$$M \leq \sqrt{r} \sqrt{u^\top \Sigma_T^{-1} u} \sqrt{\log(\det(\Sigma_T^{-1}))} + U + \sum_{t \in \mathcal{M} \cup \mathcal{U}} g_t - U .$$

By definition,

$$\Sigma_T^{-1} = I + \frac{1}{r} \sum_{t \in \mathcal{M} \cup \mathcal{U}} x_t x_t^\top = I + \frac{1}{r} \mathbf{X}_{\mathcal{A}} ,$$

so substituting and simplifying completes the proof:

$$\begin{aligned} M &\leq \sqrt{r} \sqrt{u^\top \left(I + \frac{1}{r} \mathbf{X}_{\mathcal{A}} \right) u} \sqrt{\log \left(\det \left(I + \frac{1}{r} \mathbf{X}_{\mathcal{A}} \right) \right)} + U + \sum_{t \in \mathcal{M} \cup \mathcal{U}} g_t - U \\ &= \sqrt{r \|u\|^2 + u^\top \mathbf{X}_{\mathcal{A}} u} \sqrt{\log \left(\det \left(I + \frac{1}{r} \mathbf{X}_{\mathcal{A}} \right) \right)} + U + \sum_{t \in \mathcal{M} \cup \mathcal{U}} g_t - U . \end{aligned}$$



補足(Second-Order perceptron)

- ▶ Perceptronの拡張(Nicol' o Cesa-Bianchi et al. [2005])
 - ▶ 与えられた入力ベクトルのみではなく、以前の入力ベクトルと与えられた入力ベクトルの外積も考慮して、重みベクトルを更新する

Parameter: $a > 0$.

Initialization: $X_0 = \emptyset$; $\mathbf{v}_0 = \mathbf{0}$; $k = 1$.

Repeat for $t = 1, 2, \dots$:

1. get instance $\mathbf{x}_t \in \mathbb{R}^n$;
2. set $S_t = [X_{k-1} \ \mathbf{x}_t]$;
3. predict $\hat{y}_t = \text{SGN}(\mathbf{w}_t^\top \mathbf{x}_t) \in \{-1, +1\}$,
where $\mathbf{w}_t = (aI_n + S_t S_t^\top)^{-1} \mathbf{v}_{k-1}$;
4. get label $y_t \in \{-1, +1\}$;
5. if $\hat{y}_t \neq y_t$, then:

$$\mathbf{v}_k = \mathbf{v}_{k-1} + y_t \mathbf{x}_t,$$

$$X_k = S_t,$$

$$k \leftarrow k + 1.$$

FIG. 3.1. The second-order Perceptron algorithm with parameter $a > 0$.