

論文 / 著書情報
Article / Book Information

| | |
|-----------------|--|
| 論題 | 二値重み空間でのBinary Neural Networksの学習 |
| Title | |
| 著者 | 澁谷 辰吉, 井上中順, 川上 玲, 佐藤育郎 |
| Author(s) | Tatsukichi Shibuya, Nakamasa Inoue, Rei Kawakami, Ikuro Sato |
| 出典 | MIRU2023 Extended Abstract集, , , |
| Citation | , , , |
| 発行日 / Pub. date | 2023, 7 |

背景と本研究の貢献

背景 Binary Neural Networksは推論時のメモリ削減に成功
 → 一方, 学習時は**実数重み**を持つため**メモリ消費大**
 → 学習時メモリ削減はエッジAIのローカル学習に重要

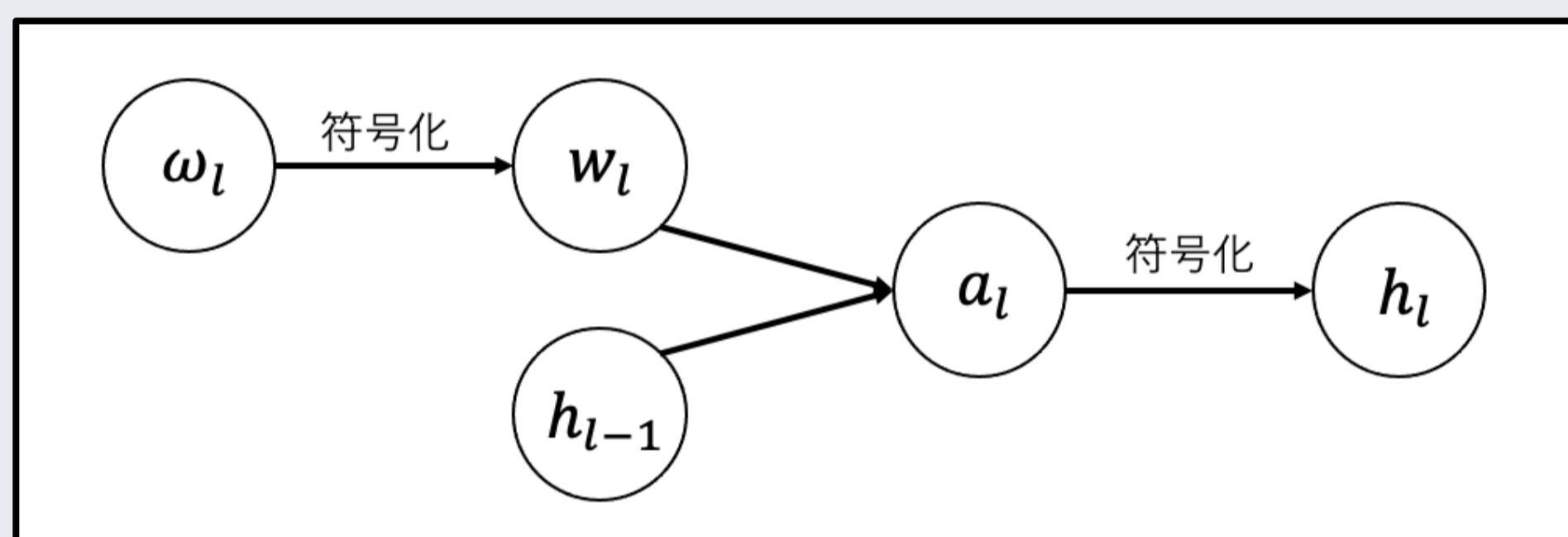
貢献 BNNsの二値重みを直接更新する学習手法を提案
 • MNISTでの実数重みなしBNNsの学習に初めて成功
 • 学習時のメモリ消費量は解析的に最大33倍削減
 • 従来の学習手法からの性能劣化は2.0%以内に留まる

従来手法 (実数重み空間での学習)

■ Binary Neural Networks (BNNs) [Courbariaux+, 2016]
 • 活性化値 h_l および実数重み ω_l を符号化関数sgnで二値化

$$a_l = \text{sgn}(\omega_l) h_{l-1}$$

$$h_l = \text{sgn}(a_l)$$



■ Straight Through Estimator (STE) による勾配近似
 • 逆伝播時にsgn関数をHardtanh関数で近似し勾配消失を回避

$$\frac{\partial \mathcal{L}}{\partial a_l} \approx \frac{\partial \mathcal{L}}{\partial h_l} \circ \mathbf{1}_{|a_l| \leq 1}, \quad \frac{\partial \mathcal{L}}{\partial \omega_l} \approx \frac{\partial \mathcal{L}}{\partial w_l}$$

■ 実数重み空間での勾配降下法

• 時刻 $t-1$ での**実数重み** ω_{t-1} を目標値 ω_t^* との内分点で更新

学習率($0 < \eta < 1$) 連続重みに関する勾配

$$\omega_t = (1 - \eta)\omega_{t-1} + \eta\omega_t^*, \quad \omega_t^* = \omega_{t-1} - g_t$$

$$\Leftrightarrow \omega_t \in \text{argmin}_{\omega \in \mathbb{R}^N} (\|\omega_{t-1} - \omega\|_2 + \|\omega - \omega_t^*\|_2)$$

提案：二値重み空間での学習

1. 勾配降下法の二値重み空間への拡張

• 二値重み w_{t-1} を目標値 $w_t^* = \text{sgn}(-g_t)$ との内分点で更新
 ハミング距離 (異なる成分の数)

$$w_t \in \text{argmin}_{w \in \{-1,1\}^N} (d(w_{t-1}, w) + d(w, w_t^*)) \Leftrightarrow$$

$$w_t = (1 - m_t) \odot w_{t-1} + m_t \odot w_t^*, \quad m_t \in \{0,1\}^N$$

マスキングを行うベクトル (ハイパーマスク)

2. ハイパーマスクの確率的サンプリング

• 現在の二値重み w_{t-1} と連続値の勾配 g_t からマスクを確率的にサンプリング → $m_t \sim P(M_t | w_{t-1}, g_t)$

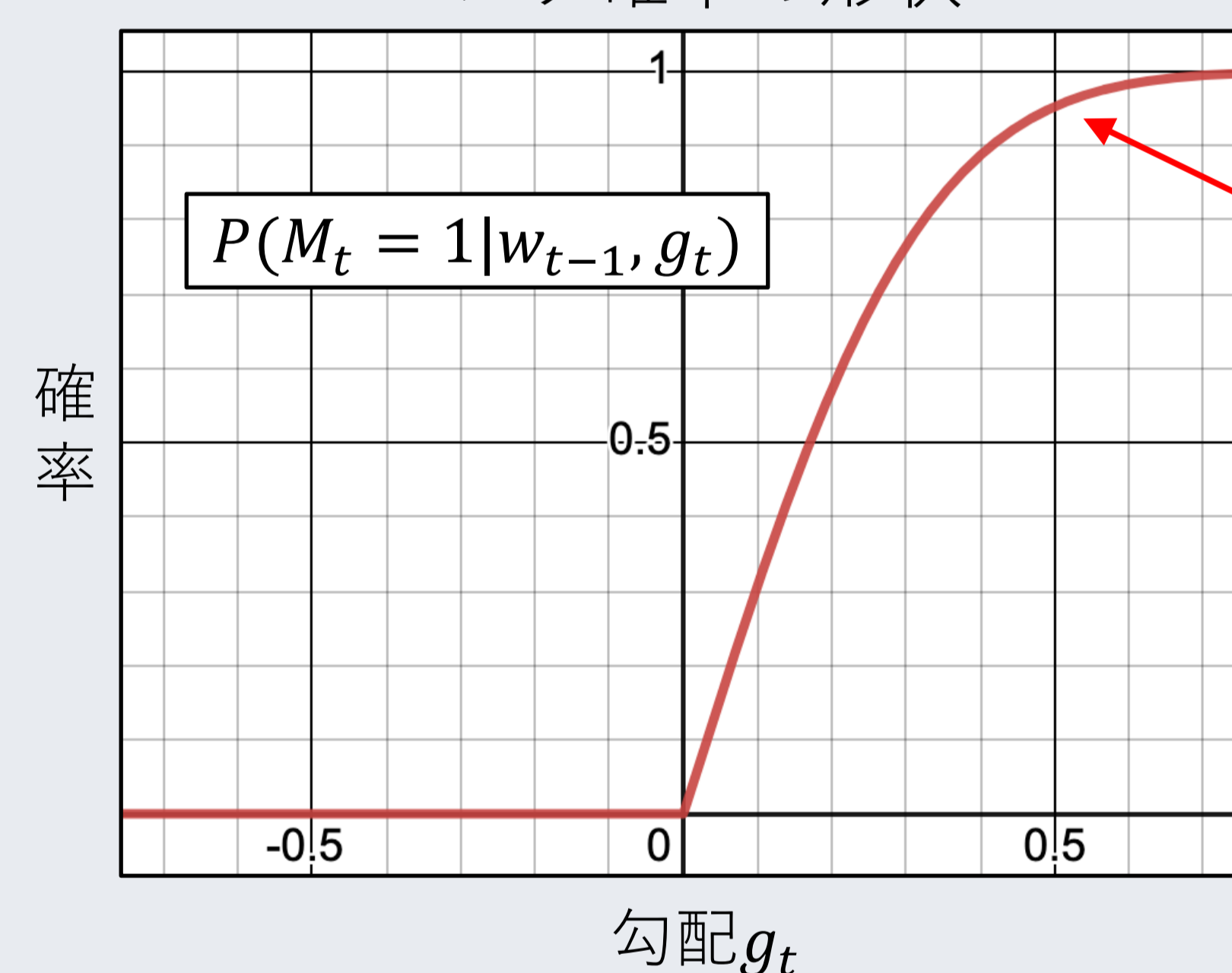
3. ハイパーマスクの生成分布の定義

$$P(M_t = 1 | w_{t-1}, g_t) = \begin{cases} \text{erf}(\alpha_{t-1} g_t) - \text{erf}(\min(\alpha_{t-1} g_t, 0)) & (w_{t-1} = 1) \\ \text{erf}(\max(\alpha_{t-1} g_t, 0)) - \text{erf}(\alpha_{t-1} g_t) & (w_{t-1} = -1) \end{cases}$$

$$\alpha_t = \frac{\eta}{\sqrt{2(\sigma_0^2 + \eta^2 \sum_{k=1}^t \sigma_{g_k}^2)}}, \quad \eta: \text{学習率}, \sigma_{g_k}^2: \text{勾配} g_k \text{の標準偏差}$$

この分布による二値重みの更新は, 勾配 g_t と実数重み w_{t-1} の独立性の仮定のもとで**実数重み空間での更新と期待値的に一致**

マスク確率の形状



勾配 g_t が大きいかほど
 M_t が1になる確率が高くなる
↓
勾配 g_t が大きいかほど
二値重みが更新されやすい

評価実験

実験1. 汎化性能

モデル 4層の全結合型ニューラルネットワーク
 中間層: 256次元 (Small) or 8192次元 (Large)

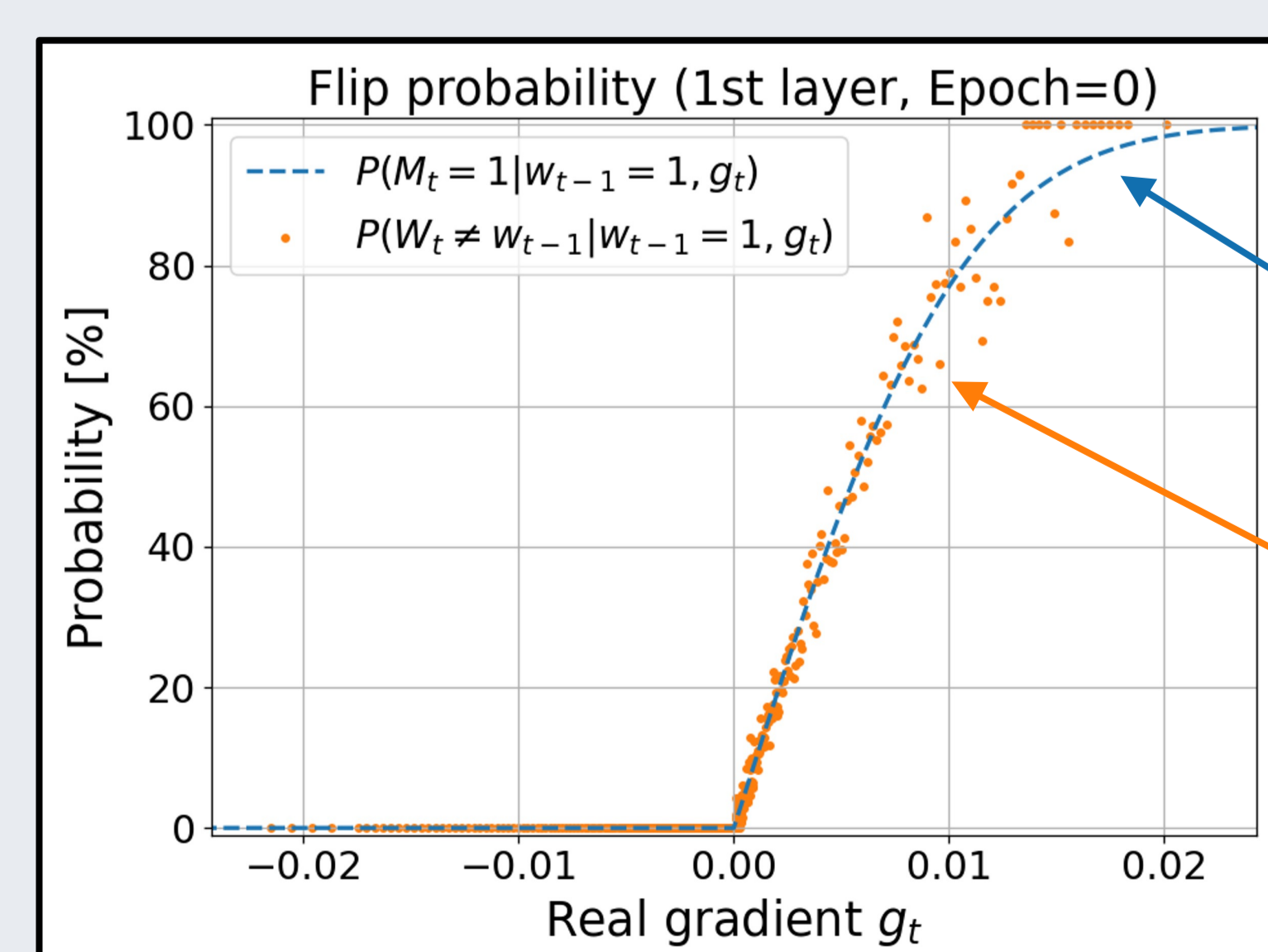
結果 • Small: 実数空間での学習から20%弱の性能劣化
 • Large: 性能劣化は約2%に留まる

| | 学習時の重み空間 | 学習アルゴリズム | MNIST | |
|------|----------------|-----------|------------------|-----------------|
| | | | Small | Large |
| NNs | \mathbb{R}^N | SGD | 1.79 \pm 0.10 | 2.32 \pm 0.04 |
| BNNs | \mathbb{R}^N | SGD + STE | 4.79 \pm 0.42 | 3.63 \pm 0.17 |
| | \mathbb{B}^N | 提案手法 | 23.78 \pm 1.04 | 5.57 \pm 0.22 |

実験2. ハイパーマスクの解析

設定 勾配 g_t に対応する要素のマスク確率 $P(M_t = 1 | w_t, g_t)$ と実空間学習での離散重みの反転確率 $P(W_t \neq w_t)$ を比較

結果 マスク確率と反転確率はほぼ一致している
 → 実数重みと関係なく, 勾配大の二値重みを更新で学習可



二値空間学習でのマスク確率

実数空間学習での二値重みの反転確率