安全な AI こそ効率的：ロバスト学習による汎化性能向上の研究

マシュー・ホーランド（大阪大学）

「任せて安心！機械学習の性能保証」

本研究の目的

関心の対象は**汎用的学習アルゴリズム**.

(1) 統計的な性能保証の高度化
(2) 安易な実装
(3) 限りなく小さなオーバーヘッド

↓（同時に満たせば）

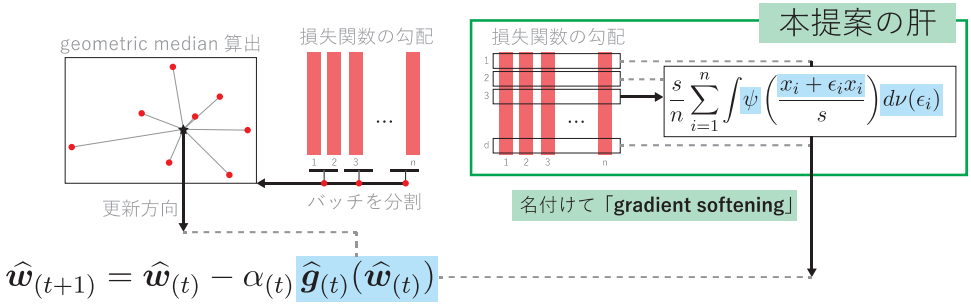
試行錯誤が少なく，
信頼保証つきの AI 基盤技術の刷新

（関連業績）
- AISTATS 2019 、採録決定 x 2 件
- ICML 2019 投稿中
- ECML 2019 投稿中

既存手法とコア技法

拙速型：速いが，性能保証がきわめて弱い（脆弱性）
（例）従来の SGD 系からの Adam，Adagrad，Dropout や Clipping 等のワザも.

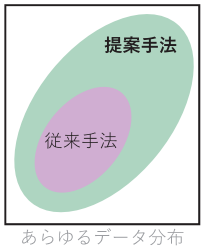
机上の空論型：理論上は担保されるが，コストが高すぎる
（例）geometric median による勾配統合，次元ごとの M 推定量の導入など.



$$\widehat{\boldsymbol{w}}_{(t+1)} = \widehat{\boldsymbol{w}}_{(t)} - \alpha_{(t)}\,\widehat{\boldsymbol{g}}_{(t)}\left(\widehat{\boldsymbol{w}}_{(t)}\right)$$

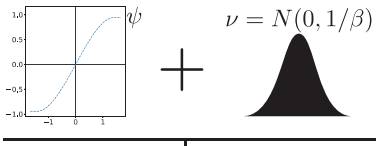
頑健性と計算効率の両立について

巧みな truncation 関数と
ノイズ分布の設計がポイント.



あらゆるデータ分布

高い確率（$1-\delta$）で，
高水準の誤差を担保する
領域を左図に明示



$$\|\widehat{\boldsymbol{g}}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{w})\| \le \frac{\widetilde{\varepsilon}}{\sqrt{n}} \quad \text{（一様上界）}$$

$$R(\widehat{\boldsymbol{w}}_{(T)}) - R^* \le (1-\alpha)^T \lambda \|\widehat{\boldsymbol{w}}_{(0)} - \boldsymbol{w}^*\|^2 + \frac{4\lambda\widetilde{\varepsilon}}{\kappa^2 n}$$

$$= O\left((1-\alpha)^T\right) + O\left(\frac{d(\log(d\delta^{-1}) + d\log(\Delta n))}{n}\right)$$

この保証は，2 次モーメントの有限性だけで十分.
従来手法と比べると，大幅な仮定の緩和.

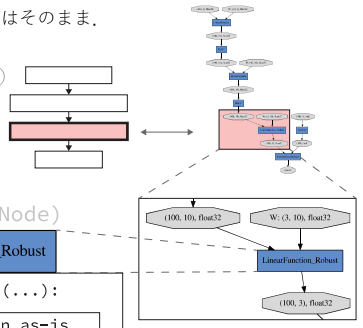
期待値を解析的に求める

$$\mathbf{E}_\nu\,\psi\left(a + b\sqrt{\beta}\epsilon\right) = a\left(a - \frac{b^2}{2}\right) - \frac{a^3}{6} + C(a,b)$$

深層学習向けの実装

逆走部分を書き換えるだけ済む.
既成 backprop はそのまま.

（Chainer での実装例）

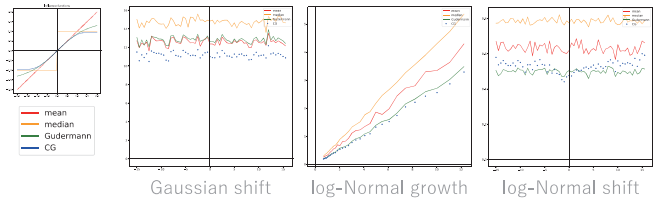


```
(FunctionNode)
LinearFunction_Robust
def forward(...):
    # use fn as-is
def backward(...):
    # core grad step
    ...
```

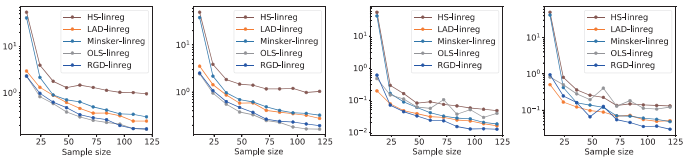
ソース＋デモ公開中
GitHub --> feedbackward --> **softgrad**

種々の実験結果

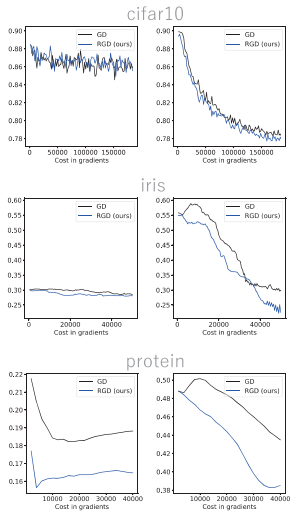
シミュレーション上の振る舞いは理論通り



微調整なしでも多様なノイズに頑健



深層学習への拡張も



今後に向けて

冒頭の (1) ～ (3) を満たす新手法.
取り口は新しく，有効性も入念な検証を経て確認できている.

次なる課題：CNN や RNN への拡張，検証，および実用化.

【問】アーキテクチャーとデータ分布に応じて，いつ，どこをロバスト化すべきか？

【問】モデルとデータに見合ったノイズ分布の設計を盛り込む利点とコストとは？