Robust Algorithms and Other Contributions to Machine Learning

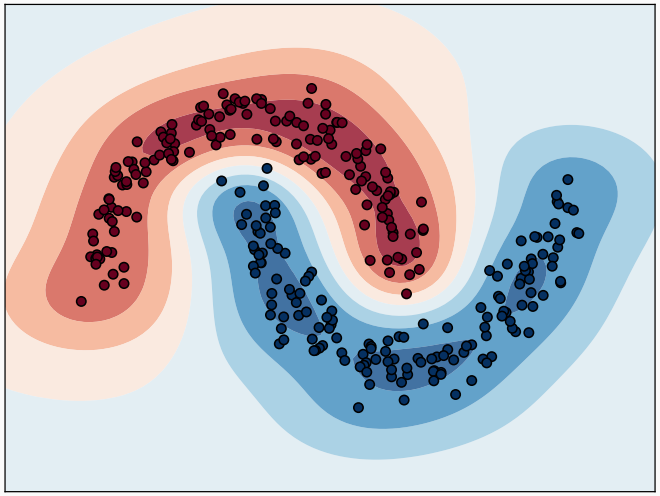
Ibrahim Merad

LPSM Université Paris Cité

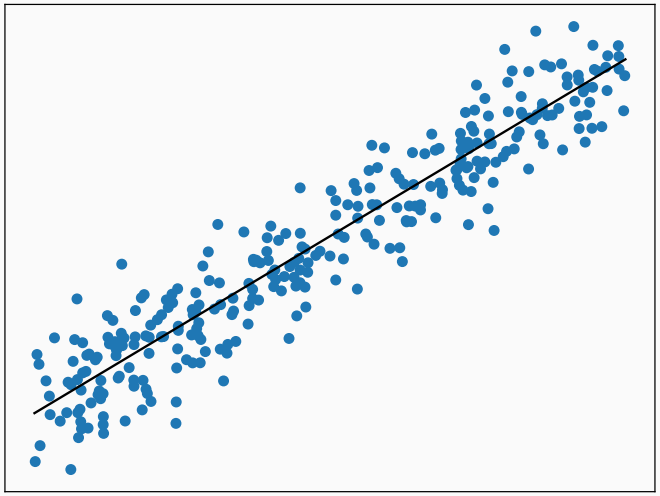
Machine learning algorithms

Learn a **task** from repeated exposure to data instances

Supervised Learning problems



Classification



Regression

Find predictor ϕ minimizing the **Risk** over sample, label pairs X, Y

$$\min_{\phi \in \Phi} \mathcal{R}(\phi) := \mathbb{E}_{X,Y}[\ell(\phi(X), Y)] \tag{1}$$

Predictor ϕ chosen among a **class of functions** Φ.

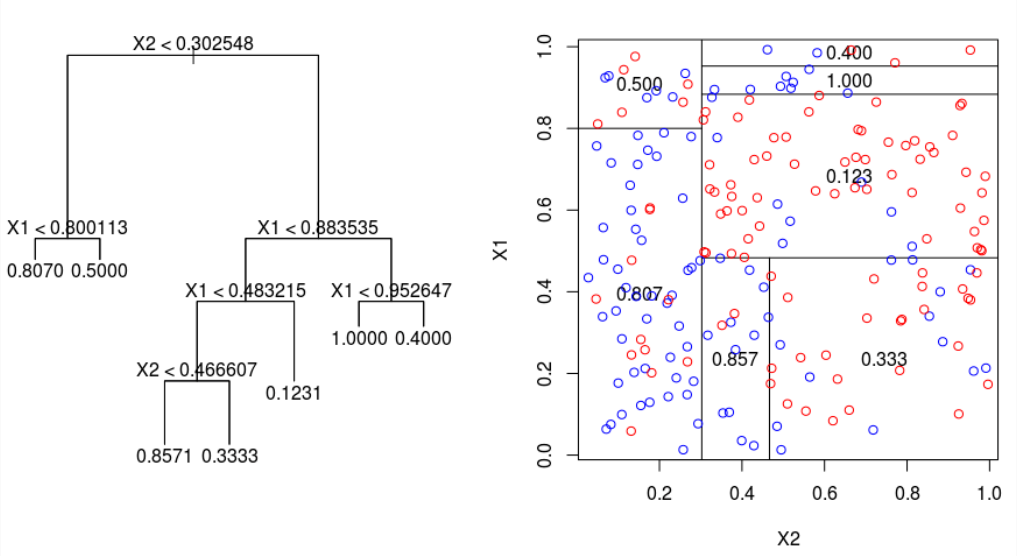
Table of contents

Wildwood : improved random forests via aggregation

Random forests

Introduced by Breiman et al. (1984).

Decision trees : set $\Phi =$ piecewise constant functions on space partition with **axis-aligned splits**, greedily trained to fit data



Random forest : tree ensemble via Bagging and feature subsampling

Predict using average over trees $\frac{1}{M}\sum_{j=1}^{M}\phi_{\mathcal{T}_j}(\cdot)$.

Bagging $\implies$ In-the-bag/Out-of-bag samples for each tree $(\mathcal{T}_j)_{j=1}^{M}$

Wildwood

Replace each tree predictor $\phi_{\mathcal{T}}(\cdot)$ by exponential average over its prunings

$$\widehat{\phi}_{\mathcal{T}}(x) = \frac{\sum_{T \subset \mathcal{T}} \pi(T) e^{-\eta L_T} \phi_T(x)}{\sum_{T \subset \mathcal{T}} \pi(T) e^{-\eta L_T}}, \quad \eta > 0 \text{ temperature.} \tag{2}$$

with $L_T = \sum_{i \in I_{oob}} \ell(\phi_T(X_i), Y_i)$ and $\pi(T) = 2^{-\|T\|}$

$\hookrightarrow$ regularization via validation score and model complexity

$\hookrightarrow$ Efficient implementation by adapting the **Context Tree Weighting** algorithm (Catoni, 2004)

Oracle inequality

$$\frac{1}{n_{oob}} \sum_{i \in I_{oob}} \ell(\widehat{\phi}_{\mathcal{T}}(X_i), Y_i) \leqslant \inf_{T \subset \mathcal{T}} \left\{ \frac{1}{n_{oob}} \sum_{i \in I_{oob}} \ell(\phi_T(X_i), Y_i) + \frac{\log 2}{\eta} \frac{\|T\|}{n_{oob} + 1} \right\}, \tag{3}$$

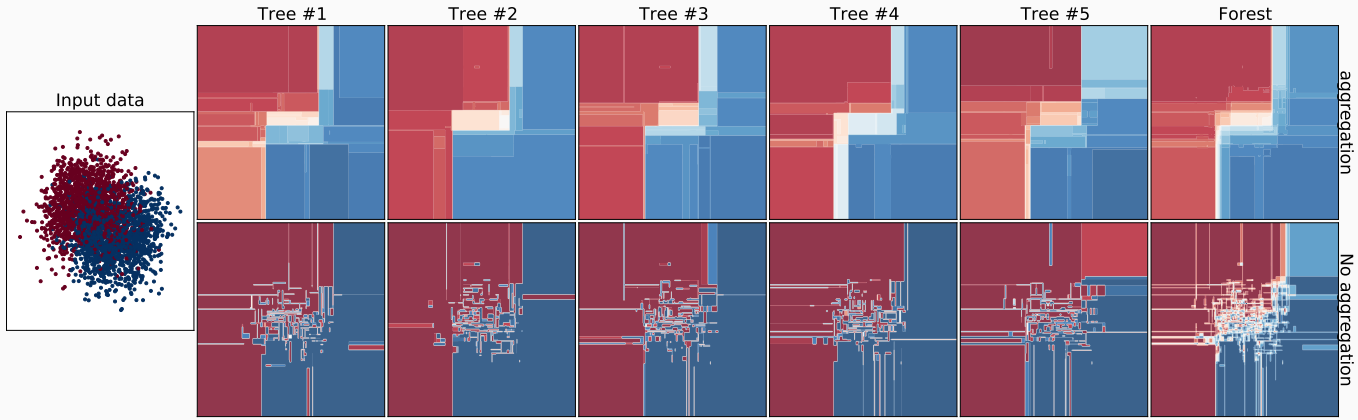
Aggregated tree performance nearly matches the best pruning

Joint work with Stéphane Gaïffas and Yiyang Yu

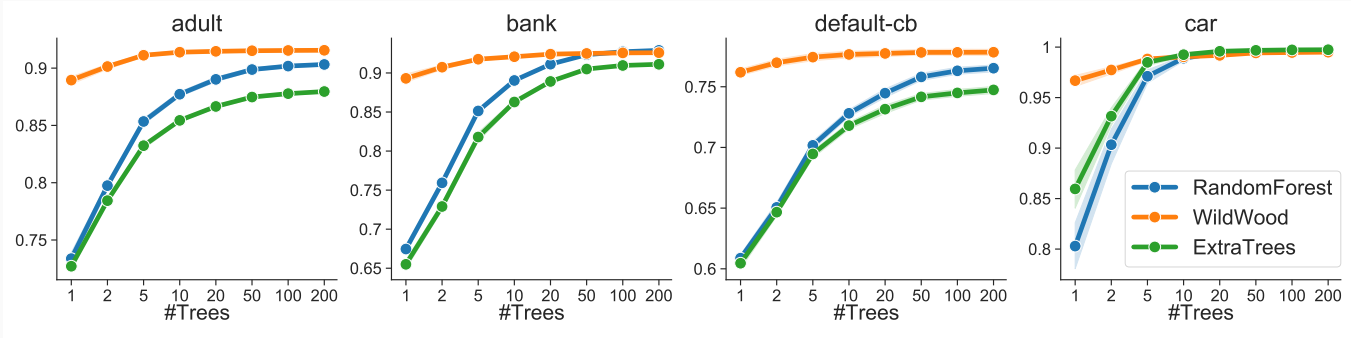
Implementation and experiments

Implemented in the Python library ***wildwood*** available on GitHub

Smoother decision boundaries for classification



Improved performance with fewer trees



Faster training, lighter models, competitive with Boosting libraries.

Robust learning with CGD

Robust Statistics

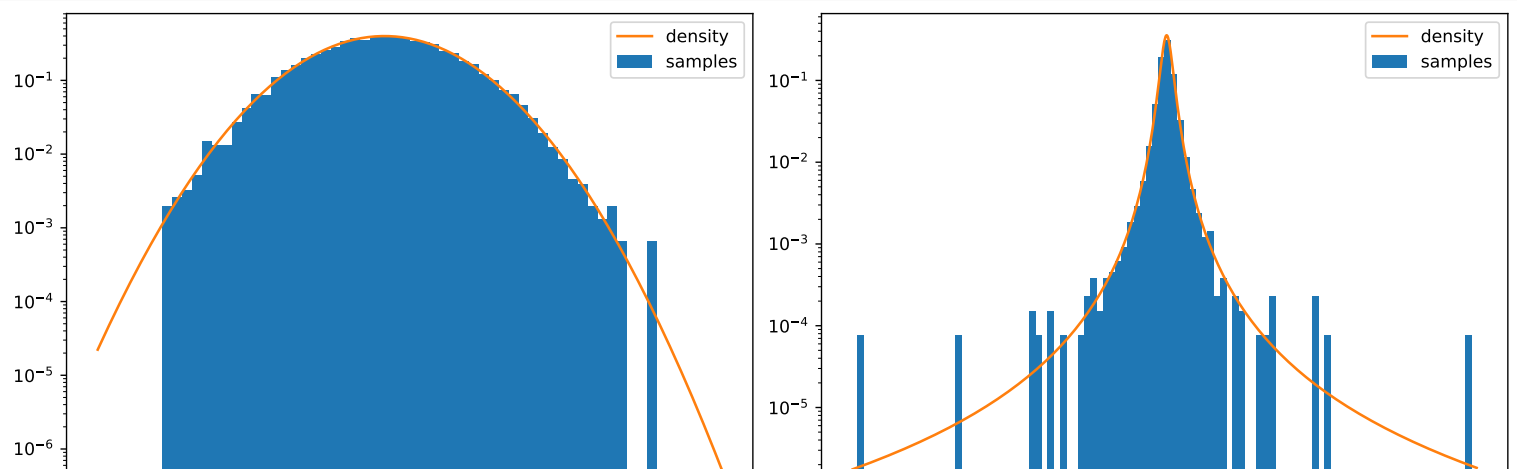
– ML algorithms' performance depends on data's statistical properties

– Classical results assume i.i.d *Gaussian* or even *bounded* data (Vapnik, 1999; Massart and Nédélec, 2006; Geer and van de Geer, 2000)

– In practice, data may be

- **Heavy-tailed** : has infinite moments
- **Corrupted** : has mistaken, irrelevant or adversarial samples.

Degraded performances ↪ Need for Robust methods



Problem

– Consider **linear learning** problems $\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) = \mathbb{E}[\ell(X^\top \theta, Y)]$

– ML objectives $\mathcal{L}(\theta)$ often optimized with **gradient descent**

$$\theta^{(t+1)} = \theta^{(t)} - \beta \nabla \mathcal{L}(\theta^{(t)}) \tag{4}$$

$\hookrightarrow$ need to estimate $\nabla \mathcal{L}(\theta^{(t)}) = \mathbb{E}[\frac{d}{d\theta} \ell(X^\top \theta^{(t)}, Y)] \in \mathbb{R}^d$

– The standard mean over data samples $\frac{1}{N} \sum_{i=1}^N \frac{d}{d\theta} \ell(X_i^\top \theta^{(t)}, Y_i)$ is **not robust**

– Robust multidimensional averages exist but are **computationally heavy** or not robust enough (Diakonikolas et al., 2020; Minsker et al., 2015; Lugosi and Mendelson, 2021)

Robust Coordinate Gradient Descent

– For **scalar** estimation, we have robust and efficient estimators

– **Idea** : Combine a robust scalar estimator with Coordinate Gradient Descent

$$\theta_j^{(t+1)} = \begin{cases} \theta_j^{(t)} - \beta_j \widehat{g}_j(\theta^{(t)}) & \text{if } j = j_t \\ \theta_j^{(t)} & \text{otherwise,} \end{cases} \tag{5}$$

β_j : step sizes, j_t coordinates

$\widehat{g}_j(\theta^{(t)})$ estimator of $g_j(\theta^{(t)}) := \frac{\partial}{\partial \theta_j} \mathcal{L}(\theta^{(t)})$

$\hookrightarrow$ fast and robust learning algorithms

Convergence result under strong convexity

Theorem

Assume $\mathcal{L}(\theta)$ λ-strongly convex. Let $\theta^{(T)}$ be CGD's output, we have w.p $\geqslant 1-\delta$

$$\mathbb{E}\big[\mathcal{L}(\theta^{(T)})\big] - \mathcal{L}^\star \leqslant \big(\mathcal{L}(\theta^{(0)}) - \mathcal{L}^\star\big)\Big(1 - \frac{\lambda}{\sum_{j\in[\![d]\!]} L_j}\Big)^T + \frac{1}{2\lambda}\big\|\epsilon(\delta)\big\|_2^2. \quad (6)$$

$\mathcal{L}^\star = \min_\theta \mathcal{L}(\theta)$, L_j coordinate Lipschitz smooth constants.

$\epsilon(\delta)$ is the **error vector** such that for all $j \in [\![d]\!]$:

$$\mathbb{P}\Big[\sup_\theta \big|\widehat{g}_j(\theta) - g_j(\theta)\big| \leqslant \epsilon_j(\delta)\Big] \geqslant 1-\delta. \quad (7)$$

Depends on estimator choice.

Robust averages

Comparison for n samples with $|\mathcal{O}|$ outliers (corruptions)

	Sub-Gaussian deviation	Robustness to outliers	Numerical complexity	Hyper-parameter
ERM	No	None	$O(n)$	None
MOM	Yes	Yes for $\|\mathcal{O}\| < K/2$	$O(n+K)$	$K \in [\![n]\!]$
CH	Yes	None	$O(n)$	Scale s
TM	Yes	Yes for $\|\mathcal{O}\| < n/8$	$O(n)$	$\epsilon \in [0, 1/2)$

ERM : standard mean, *MOM* : Median-Of-Means, *CH* : Catoni-Holland, *TM* : trimmed mean

$\hookrightarrow$ Best compromise of computational performance and robustness with the **Trimmed mean**.

Some theoretical aspects and implementation

- Bounds for gradients with $1 + \alpha$ **moments** for $\alpha \in (0, 1]$ (sub-Gaussian when variance exists)
- Results for CGD with importance sampling, uniform sampling or deterministic coordinates
- Result for **non strongly convex** objectives
  - slower convergence and statistical rate
  - requires thresholding

Efficiently implemented in the Python library *linlearn*[1] along with concurrent baselines (Prasad et al., 2020; Lecué et al., 2020; Holland and Ikeda, 2019).

[1] *https://github.com/linlearn/linlearn*

Numerical experiments

Example for classification on real data

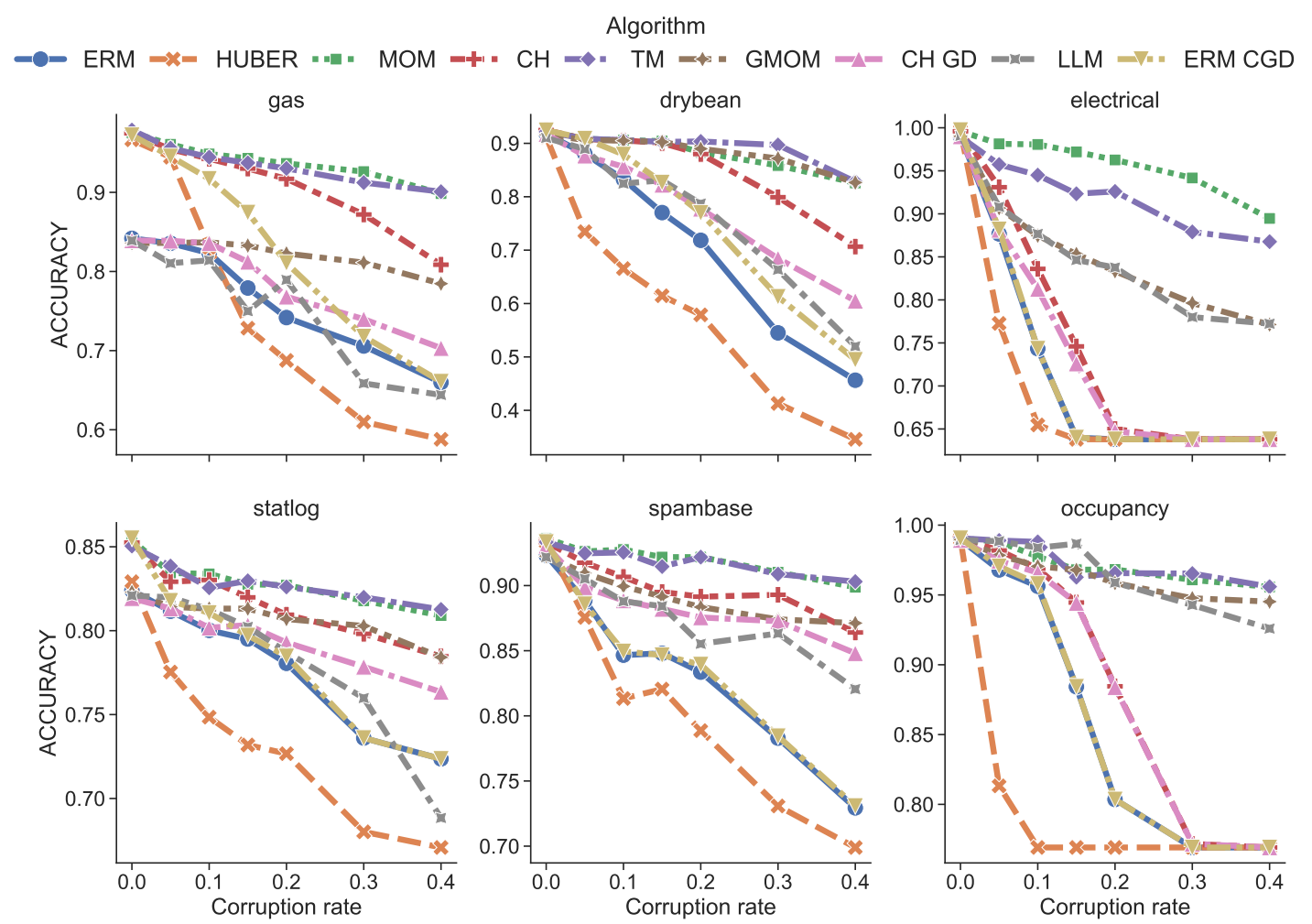


Figure 1: *TM* and *MOM* are the most robust to data corruption

Robust high-dimensional learning

The High-dimensional setting

Linear learning $\mathcal{L}(\theta) = \mathbb{E}[\ell(X^\top \theta, Y)]$ in **high-dimension** : $X, Y \in \mathbb{R}^d \times \mathcal{Y}$ where $d \gg n$.

Assumption

The optimal $\theta^\star$ is s-sparse (in a generic sense)

For high-dimension, we use **non-Euclidean** optimization $\Longrightarrow$ **different metric** on gradient error $\Longrightarrow$ adapted estimators

For smooth objectives, use **Mirror Descent**

$$\theta_{t+1} = \arg\min_{\theta} \langle \beta \widehat{g}(\theta_t), \theta \rangle + V_{\theta_0}(\theta, \theta_t) \tag{8}$$

β : step size, $\widehat{g}$ gradient estimator and V_{θ_0} Bregman divergence with reference point θ_0

For non-smooth objectives, use **Dual Averaging**

AMMD Algorithm

Approximate **M**ultistage **M**irror **D**escent, inspired from Juditsky et al. (2022)

$\hookrightarrow$ Multiple stages of Mirror descent followed by sparsification

Theorem

Let $\theta^\star$ be s-**sparse** and $\mathcal{L}$ be L-**smooth** with **quadratic growth** such that $\mathcal{L}(\theta) - \mathcal{L}^\star \geqslant \kappa \|\theta - \theta^\star\|_2^2$. After K stages of AMMD we have

$$\|\theta^{(K)} - \theta^\star\|_2 \leqslant 2^{-K} R + \frac{40 \bar{\epsilon} \sqrt{\bar{s}}}{\kappa}, \tag{9}$$

with $\theta_0 \in B_{\|\cdot\|}(\theta^\star, R)$ and $\bar{\epsilon} = \max_t \|\widehat{g}(\theta_t) - g(\theta_t)\|_*$ (dual norm).

$\hookrightarrow$ **Generic** for different **forms of sparsity**.

AMDA Algorithm

Approximate **M**ultistage **D**ual **A**veraging

$\hookrightarrow$ Dual averaging stages followed by sparsification

Weaker assumptions : no smoothness and only **pseudo-linear** growth :

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^\star) \geqslant \frac{\kappa \|\theta - \theta^\star\|_2^2}{\lambda + \|\theta - \theta^\star\|_2}, \quad \lambda, \kappa > 0. \tag{10}$$

Theorem

Let $\theta^\star$ be s-**sparse** and $\mathcal{L}$ be Lipschitz with **pseudo-linear growth**. After K stages of AMDA we have

$$\|\theta^{(K)} - \theta^\star\|_2 \leqslant R_K / \sqrt{s}, \tag{11}$$

where $\theta_0 \in B_{\|\cdot\|}(\theta^\star, R_0)$ and R_k converges geometrically to $O\big(\frac{\lambda \bar{s} \bar{\epsilon}}{\kappa}\big)$ and $\bar{\epsilon}$ as before.

Requires *longer* stages leading to *slower convergence*.

Applications

The main application is **Vanilla sparsity** : set $\|\cdot\| = \|\cdot\|_1$ on the parameter space $\implies$ $\|\cdot\|_* = \|\cdot\|_\infty$ on the gradient space

$\hookrightarrow$ coordinatewise trimmed mean estimation leads to a **nearly optimal** rate.

We have with probability $\geqslant 1-\delta$

$$\left\|\widehat{g}^{TM}(\theta) - g(\theta)\right\|_\infty \leqslant 7\sigma_{\max}\sqrt{4\eta + 6\frac{\log(4/\delta) + \log(d)}{n}}. \tag{12}$$

where $\sigma_{\max} = \max_j \mathrm{Var}(g_j(\theta))$ and η : corruption rate.

Plugging into previous theorems, we obtain an **efficient** and **highly robust** algorithms with rate $s\log(d)/n$.

Other applications

- **Group sparsity** : coordinates are partitioned into groups $[\![d]\!] = \cup_j G_j$.
  - Sparsity in terms of **non-zero groups**
  - Take the ℓ_1/ℓ_2 norm as metric $\|\theta\| = \sum_j \|\theta_{G_j}\|_2$ resulting in the ℓ_∞/ℓ_2 norm as dual.
  - Proposed estimator : stability based robust vector mean (Diakonikolas et al., 2020) *groupwise*.
- **Low-rank matrix recovery** : consider a matrix space $\theta \in \mathbb{R}^{p\times q}$.
  - Sparsity in terms of **rank**.
  - Take the nuclear norm $\|\sigma(\theta)\|_1$ as metric ($\sigma(\cdot)$ = singular values) $\implies$ the dual is the operator norm $\|\cdot\|_* = \|\cdot\|_{\mathrm{op}}$.
  - Proposed estimator : MOM version of Minsker (2018) (an extension of Catoni's estimator for matrices)

Implementation and experiments

Previous algorithms and estimators implemented in *linlearn*

Baselines : LASSO variants and iterative thresholding gradient descent

SDP-based baselines dropped due to **heavy computational cost**

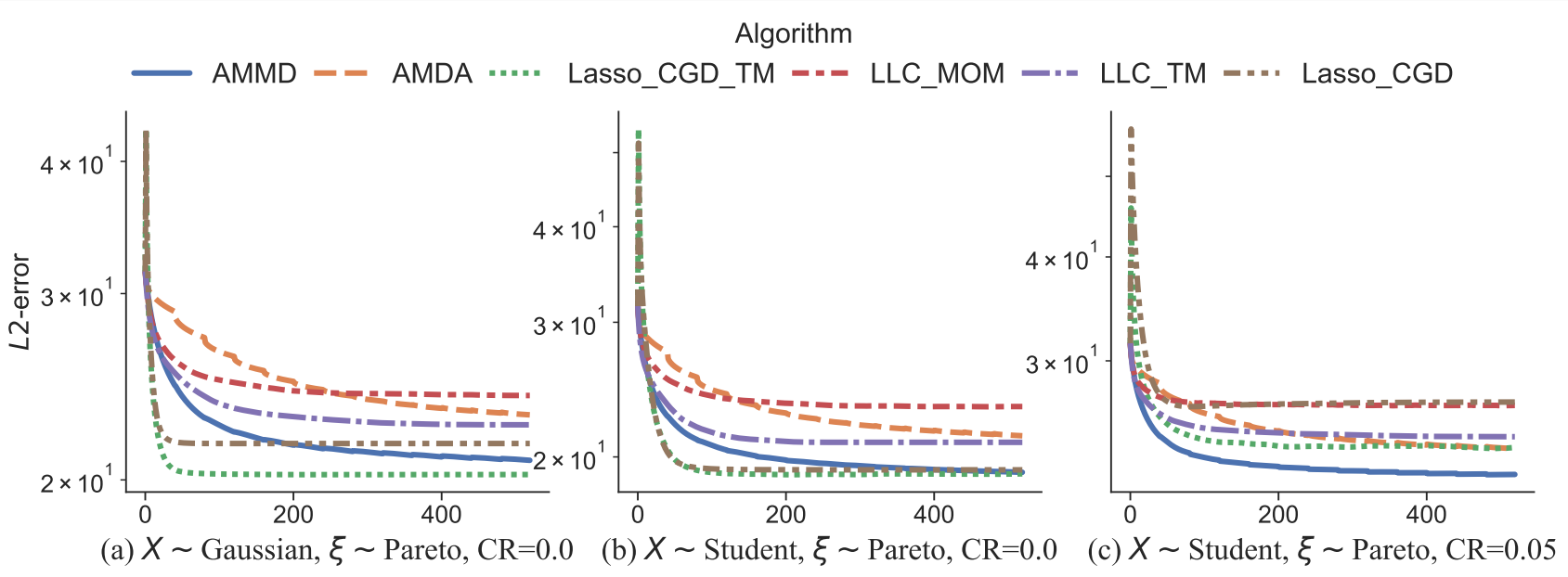
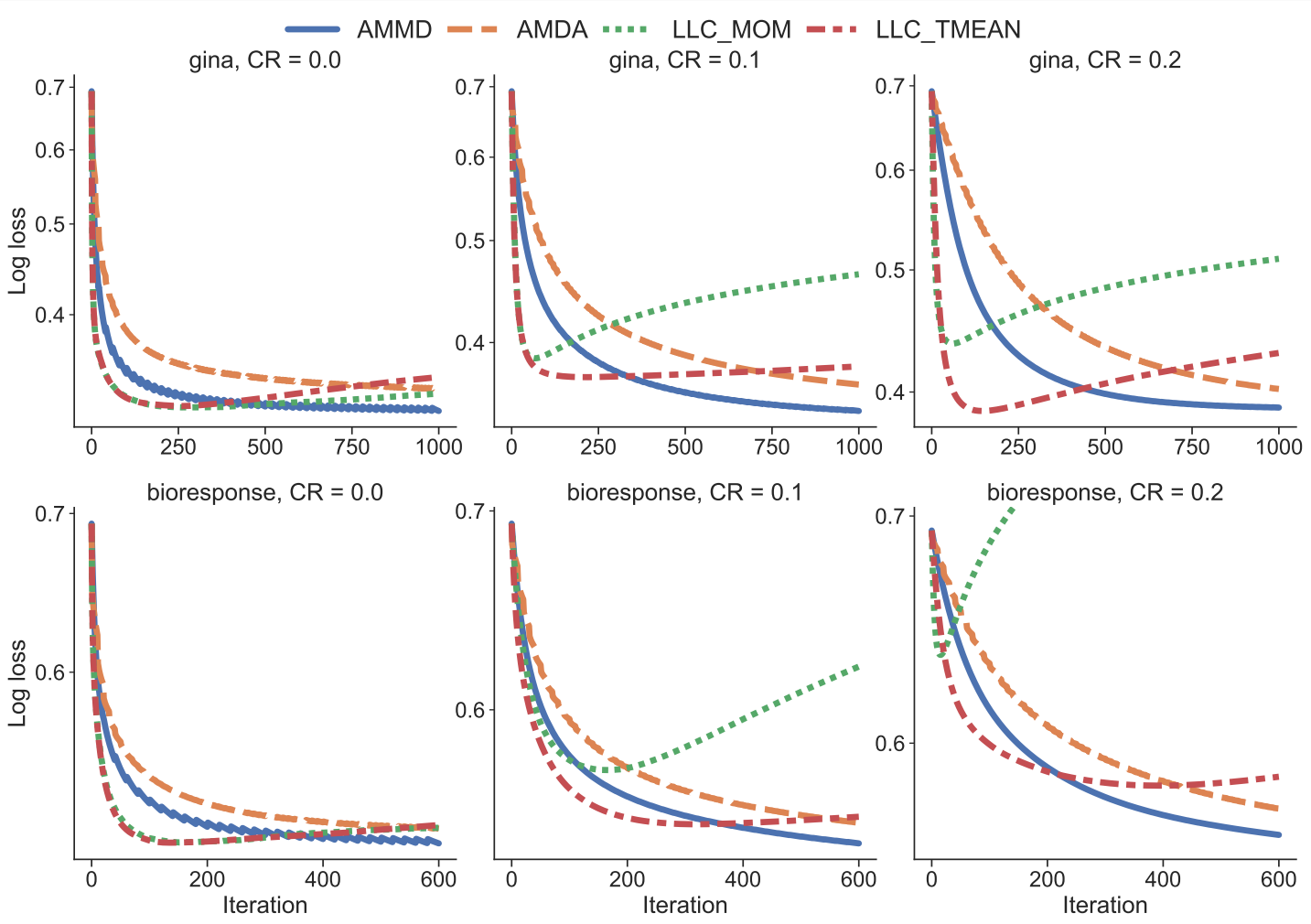


Figure 2: High dimensional regression on synthetic data

Classification on real data



– AMMD and AMDA are more robust to corruption

– Slower convergence for AMDA

Robust stochastic optimisation

Problem and Relevant Literature

– **Stochastic optimisation** of an objective $\mathcal{L}(\theta)$ (back to "*moderate*" dimension)

– Gradient samples received **individually** and **sequentially**

Published papers on heavy-tail robustness with **clipped SGD**

- Gorbunov et al. (2020) : constant step size, uniformly bounded variance assumption, requires minibatching
- Tsai et al. (2022) : decreasing step-size, loose variance bound, no minibatching

Can we handle **infinite variance** $\implies$ Yes (Nguyen et al., 2023; Sadiev et al., 2023)

Can we handle **corruption** ? $\implies$ Yes : via Quantile clipping

Our Setting

– No batch data + $O(d)$ memory constraint $\implies$ **cannot** use robust estimators.

– Corrupted gradients $G_1, \ldots, G_T \in \mathbb{R}^d$ w.p $\eta < 1/2$ independently

$$G(\theta_t, \zeta_t) = G_t = U_t \underbrace{\check{G}_t}_{\text{corrupted}} + (1 - U_t) \underbrace{\widetilde{G}_t}_{\text{true}} \quad \text{with} \quad U_t \sim^{iid} \mathcal{B}(\eta), \quad (13)$$

where $\mathbb{E}[\widetilde{G}_t] = \nabla \mathcal{L}(\theta_t)$.

– Heavy-tails i.e we have the **Moment assumption**

$$\mathbb{E}\left[\|\widetilde{G}_t - \nabla \mathcal{L}(\theta_t)\|^q | \theta_t\right]^{1/q} \leqslant A_q \|\theta_t - \theta^\star\| + B_q, \quad (14)$$

for some $q > 1$.

Quantile Clipped SGD

$$\theta_{t+1} = \theta_t - \alpha_{\theta_t} \beta G_t \quad \text{with} \quad \alpha_{\theta_t} = \min\left(1, \frac{Q_p(\|\widetilde{G}_t\|)}{\|G_t\|}\right) \tag{15}$$

$Q_p(\cdot)$: quantile of index p

Idea :

- No need to *identify* corruptions.

$\hookrightarrow$ Cap the contribution of each sample and the majority (true samples) prevails.

Parameters : step-size β and quantile index p.

Define $P_{\beta,p}$ the **Markov transition kernel**

$\Longrightarrow$ For $\theta_t \sim \nu$, successor distribution is $\theta_{t+1} \sim \nu P_{\beta,p}$

Convergence

Define $V(\theta) = 1 + \|\theta - \theta^\star\|^2$, we show the drift property

$$\mathbb{E}[V(\theta_{t+1})|\theta_t] \leqslant \lambda V(\theta_t) + b \cdot \mathbf{1}_{\theta_t \in \mathcal{C}} \tag{16}$$

$\lambda < 1, b > 0, \mathcal{C} \subset \mathbb{R}^d$ bounded.

Theorem [geometric ergodicity] (Informal)

Assume $\mathcal{L}$ **smooth** and **strongly convex** and gradients G_t as before. For well chosen β, p, there is a unique invariant measure $\pi_{\beta,p}$ and $\rho < 1$ and $M < \infty$ s.t.

$$\left\|\mathcal{D}(\theta_t) - \pi_{\beta,p}\right\|_{\mathrm{TV}} = \left\|\delta_{\theta_0} P_{\beta,p}^t - \pi_{\beta,p}\right\|_{\mathrm{TV}} \leqslant M\rho^t\left(1 + \|\theta_0 - \theta^\star\|^2\right). \tag{17}$$

$\mathcal{D}(\theta_t)$: distribution of θ_t

Based on continuous state Markov Chain theory (Meyn and Tweedie, 1993)

Invariant Distribution properties

Using invariance property of $\pi_{\beta,p}$, for $p = 1 - \eta$, we have

$$\mathbb{E}_{\theta \sim \pi_{\beta,p}}\left[\|\theta - \theta^\star\|^2\right] \leqslant 20\left(\frac{\eta^{1-1/q} B_q}{\kappa}\right)^2. \tag{18}$$

- 👍 Robust to corruption and heavy-tails (even infinite variance)
- 👍 Optimal rate in η
- 👎 Optimal convergence speed is hard to prove
- 👎 B_q may be of order $\sqrt{d} \implies$ sub-optimal dimension dependence

By restricting the optimization to a **bounded** set

- For $\eta = 0$ (no corruption), $\pi_{\beta,p}$ is sub-Gaussian
- For $\eta > 0$ (with corruption) $\pi_{\beta,p}$ is sub-exponential

Non convex objectives

If $\mathcal{L}$ is only smooth and positive, assume the set

$$\left\{\theta \in \mathbb{R}^d : \left\|\nabla \mathcal{L}(\theta)\right\|^2 \leqslant O\left(\eta^{2-\frac{2}{q}} B_q^2\right)\right\} \quad \text{is } \textbf{bounded}. \tag{19}$$

Show the **drift property**

$$\mathbb{E}[V(\theta_{t+1})|\theta_t] - V(\theta_t) \leqslant -1 + b \cdot \mathbf{1}_{\theta_t \in \mathcal{C}} \tag{20}$$

$\implies$ $(\theta_t)_{t \geqslant 0}$ is ergodic with invariant distribution $\pi_{\beta,p}$

$$\left\|\mathcal{D}(\theta_t) - \pi_{\beta,p}\right\|_{\mathsf{TV}} = \left\|\delta_{\theta_0} P_{\beta,p}^t - \pi_{\beta,p}\right\|_{\mathsf{TV}} \leqslant M/t, \qquad M > 0 \tag{21}$$

$\hookrightarrow$ **Slower** convergence.

Characterization of $\pi_{\beta,p}$

$$\mathbb{E}_{\theta \sim \pi_{\beta,p}}\left[\left\|\nabla \mathcal{L}(\theta)\right\|^2\right] \leqslant O\left(\eta^{2-\frac{2}{q}} B_q^2\right). \tag{22}$$

$\hookrightarrow$ A low gradient neighborhood is reached.

Implementation and experiments

- True quantile $Q_p(\|\widetilde{G}_t\|)$ not available in practice

$\hookrightarrow$ Use a rolling quantile estimate $\widehat{Q}_p = \lfloor pS \rfloor$-th **order statistic** of a buffer $(\|G(\theta_{t-j}, \zeta_{t-j})\|)_{1 \leqslant j \leqslant S}$ of size $S \in \mathbb{N}^*$

$\Longrightarrow$ $O(S)$ memory and time complexity

- Experiments on vector mean estimation and linear/logistic regression
  Synthetic data with heavy-tails and corruption.

Results for linear/logistic regression

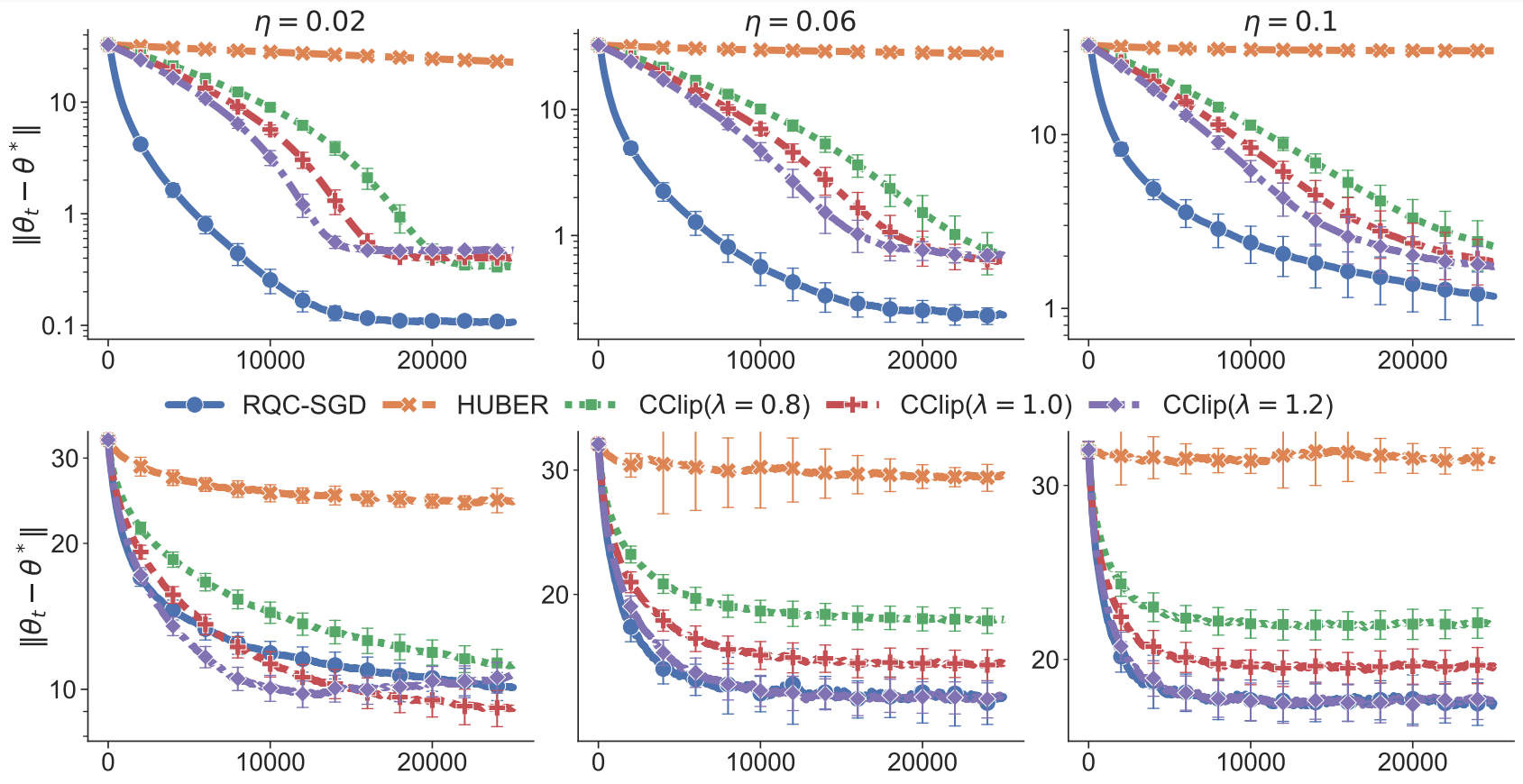


Figure 3: Linear regression (top row) and Logistic regression (bottom row)

SGD Convergence and Concentration

SGD limit distribution concentration

Non-robust setting (no heavy-tails/corruption).

Smooth and μ-**strongly convex** objective $\mathcal{L}(\theta) = \mathbb{E}_\zeta[\ell(\theta, \zeta)]$ optimized with constant step SGD

$$\theta_{t+1} = \theta_t - \beta G(\theta_t, \zeta_t) \quad \text{with} \quad \mathbb{E}[G(\theta_t, \zeta_t)] = \nabla \mathcal{L}(\theta_t) \tag{23}$$

$\hookrightarrow$ Markov chain approach, convergence established as previously.

Proposition (Concentration)

Assume $\mathcal{L}$ as above and $G(\theta, \zeta) - \nabla \mathcal{L}(\theta)$ K-sub-Gaussian (resp. K-sub-exponential) uniformly in θ.

Then SGD **geometrically** converges in TV to an invariant distribution π_β which is $4K\sqrt{\beta/\mu}$-sub-Gaussian (resp. sub-exponential).

Key proof ingredient : unbiased gradient samples

$\hookrightarrow$ small β + geometric ergodicity $\Longrightarrow$ high confidence estimate of $\theta^\star$

Wasserstein convergence

Besides TV convergence we have

Proposition (Wasserstein contraction)

Let $\mathcal{L}$ as before and assume gradient errors $\varepsilon_\zeta(\theta) = G(\theta, \zeta) - \nabla\mathcal{L}(\theta)$ satisfy the Wasserstein Lipschitz condition

$$\mathcal{W}_2^2\big(\mathcal{D}(\varepsilon_\zeta(\theta)), \mathcal{D}(\varepsilon_\zeta(\theta'))\big) \leqslant L_{\mathcal{W}} \|\theta - \theta'\|_2^2 \quad \text{for all} \quad \theta, \theta' \tag{24}$$

with $\mathcal{D}(\varepsilon_\zeta(\theta)) =$ distribution of $\varepsilon_\zeta(\theta)$. Then we have the contraction

$$\mathcal{W}_2^2(\nu_1 P_\beta, \nu_2 P_\beta) \leqslant \big(\underbrace{(1-\beta\mu)^2 + \beta^2 L_{\mathcal{W}}}_{<1 \text{ for small } \beta} \big) \mathcal{W}_2^2(\nu_1, \nu_2) := \alpha_{\mathcal{W}}^2 \mathcal{W}_2^2(\nu_1, \nu_2). \tag{25}$$

$\hookrightarrow$ SGD iteration is a contraction in Wasserstein distance.

$\Longrightarrow$ convergence to a unique stationary $\pi_\beta \in \mathcal{P}_2(\mathbb{R}^d)$ in $\mathcal{W}_2$ distance.

Confidence bound for Polyak-Ruppert average

Proposition

Let $\mathcal{L}$ as before with **linear** gradient and $(\theta_t)_{t\geqslant 0}$ be started from $\theta_0 \sim \nu$ with β small enough. Then, there exist $\rho < 1$ and $M < +\infty$ s.t. for $\delta > 0$ and $n, n_0 > 0$:

$$\left\| \frac{1}{n} \sum_{t=n_0+1}^{n_0+n} \theta_t - \theta^\star \right\| \leqslant \sqrt{\frac{2}{n} \frac{1+\alpha}{1-\alpha} \left(\alpha_{\mathcal{W}}^{n_0} \mathcal{W}_2^2(\nu, \pi_\beta) + \frac{\beta\sigma^2}{\mu} \right)} + \frac{4K\sqrt{\beta/\mu}}{1-\alpha_{\mathcal{W}}} \sqrt{\frac{\log(1/\delta)}{n}}$$

w.p. $\geqslant 1 - \Upsilon(\nu, n_0)\delta$, where $\Upsilon(\nu, n_0) = 1 + M\rho^{n_0} \left\| \frac{d\nu}{d\pi_\beta} \right\|_\infty$ and $\alpha = 1 - \beta\mu$.

$\hookrightarrow$ β should have **constant** order.

Conclusion

Conclusion

In this presentation

- A new random forest algorithm with built-in regularization mechanism to prevent overfitting
- Computationally efficient algorithms for robust learning
- Important role of optimisation methods for robustness
- Tailored algorithms for different problems and error metrics
- Robustness for streaming data with an appropriate clipping strategy
- Convergence and concentration properties of standard SGD and Polyak-Ruppert averaging

Thank you !

Wildwood training speed

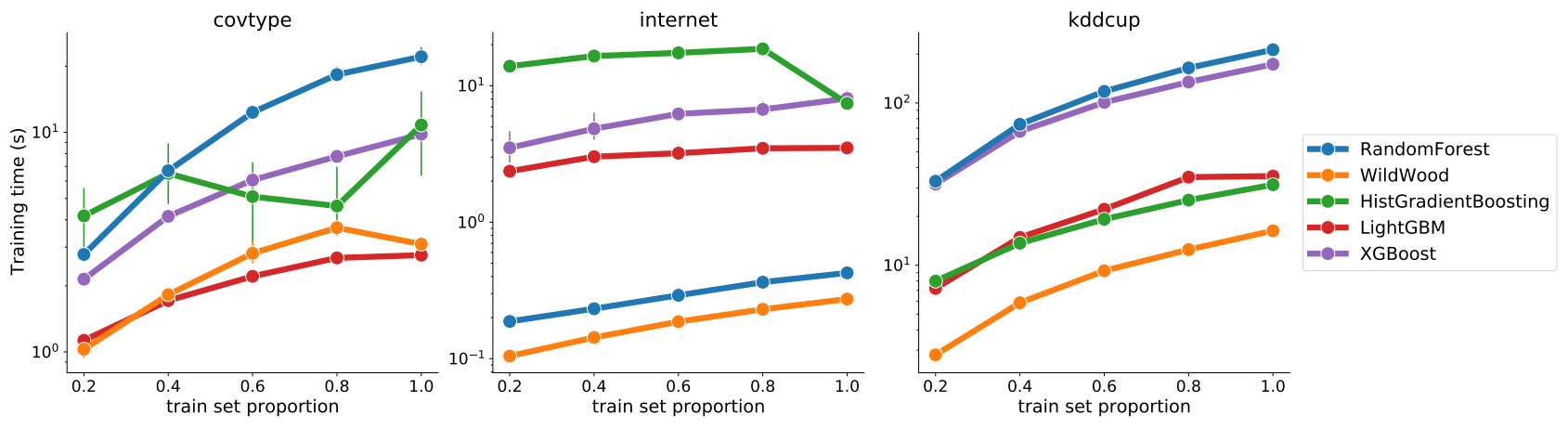


Figure 4: Training times on increasing fractions of a few large datasets for WildWood and a few baselines

Wildwood trains fast compared to other algorithms.

Wildwood model size

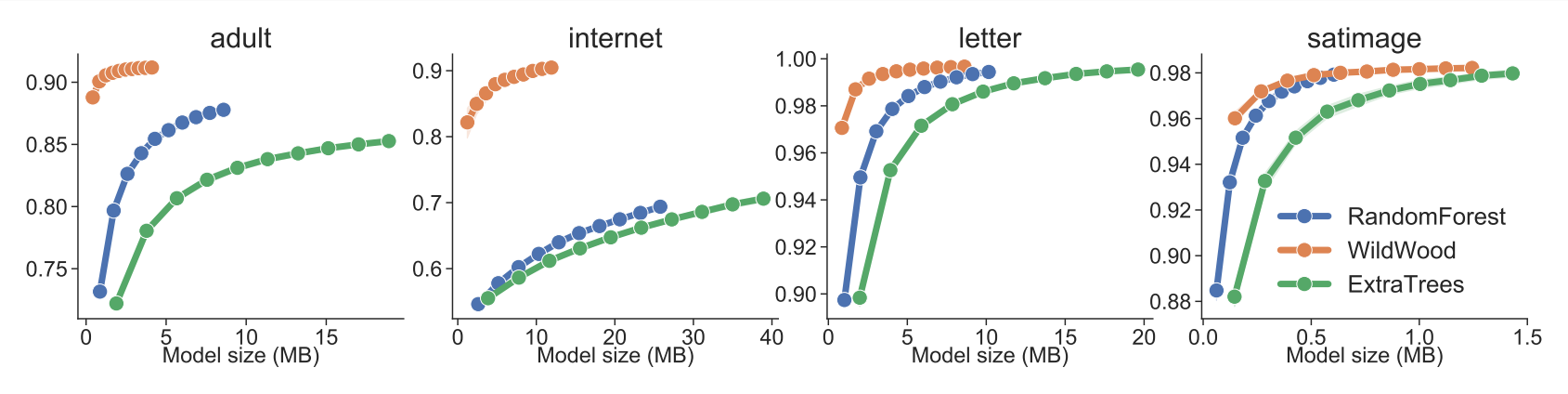


Figure 5: Classification performance measured through AUC as a function of model size in megabytes.

Wildwood reaches better performance using smaller models compared to baselines.

Sub-Gaussian and sub-exponential distributions

– If X is a real random variable, we say that X is K-sub-Gaussian for $K > 0$ if

$$\mathbb{E}\exp(\lambda^2 X^2) \leqslant e^{\lambda^2 K^2} \quad \text{for} \quad |\lambda| \leqslant 1/K. \tag{26}$$

$\Longrightarrow$ with probability at least $1 - \delta$, we have

$$|X| \leqslant K\sqrt{\log(e/\delta)}. \tag{27}$$

– We say that X is K-sub-exponential for $K > 0$ if

$$\mathbb{E}\exp(\lambda|X|) \leqslant \exp(\lambda K) \quad \text{for all} \quad 0 \leqslant \lambda \leqslant 1/K. \tag{28}$$

$\Longrightarrow$ with probability at least $1 - \delta$, we have

$$|X| \leqslant 2eK\log(2/\delta). \tag{29}$$

– For random vectors $X \in \mathbb{R}^d$, the properties are defined in terms of $\langle X, u\rangle$ for all $\|u\| = 1$.

Median-Of-Means estimator

Given samples $X_1, \dots, X_n$ and number of blocks K

Let $\{1, \dots, n\} = B_1 \cup \cdots \cup B_K$ disjoint with $|B_k| = n/K$, compute :

$$\mu^{(k)} = \frac{1}{|B_k|} \sum_{i \in B_k} X_i \quad \text{for} \quad k = 1, \dots, K \tag{30}$$

$\hookrightarrow$ Return $\widehat{\mu}_n^{\mathrm{MOM}} = \mathsf{median}\left(\mu^{(1)}, \dots, \mu^{(K)}\right)$

Setting $K = \lceil 18 \log(1/\delta) \rceil$, if $\mu = \mathbb{E}[X]$, $\|X\|_{L^{1+\alpha}} \leqslant M_\alpha$ and corrupted set $|\mathcal{O}| \leqslant K/12$, we have

$$\left|\widehat{\mu}_n^{\mathrm{MOM}} - \mu\right| \leqslant c_\alpha M_\alpha \left(\frac{\log(1/\delta)}{n}\right)^{\alpha/(1+\alpha)}$$

with probability $\geqslant 1 - \delta$, where $c_\alpha := 2^{(3+\alpha)/(1+\alpha)} 3^{(1+2\alpha)/(1+\alpha)}$.

Trimmed mean estimator

Given samples $X_1, \dots, X_n$ and a quantile $0 < \epsilon < 1/2$

Divide $\{1, \dots, n\} = \{1, \dots, n/2\} \cup \{n/2 + 1, \dots, n\}$ and compute :

$$q_\epsilon = X^{([\epsilon n/2])} \quad \text{and} \quad q_{1-\epsilon} = X^{([(1-\epsilon)n/2])} \tag{31}$$

where $X^{(1)}, \dots, X^{(n/2)}$ are the order statistics of $(X_i)_{i \leqslant n/2}$

$\hookrightarrow$ Return $\widehat{\mu}_n^{\mathrm{TM}} = \frac{2}{n} \sum_{i=n/2+1}^{n} q_\epsilon \vee X_i \wedge q_{1-\epsilon}$

If $\mu = \mathbb{E}[X]$, $\|X\|_{L^{1+\alpha}} \leqslant M_\alpha$ and corrupted set $|\mathcal{O}| \leqslant \eta n$, setting $\epsilon = 8\eta + 12 \log(4/\delta)/n$, we have

$$\left|\widehat{\mu}_n^{\mathrm{TM}} - \mu\right| \leqslant 7 M_\alpha \left(4\eta + \frac{6 \log(4/\delta)}{n}\right)^{\alpha/(1+\alpha)}$$

with probability $\geqslant 1 - \delta$.

Catoni's estimator

Define the "influence" function

$$\psi(x) = \begin{cases} \log(1 + x + x^2/2) & \text{if } x \geqslant 0 \\ -\log(1 - x + x^2/2) & \text{otherwise} \end{cases} \tag{32}$$

Given a sample $X_1, \ldots, X_n$ and *scale parameter* s.

Define $\widehat{\mu}_n^{\mathrm{CH}}$ as the solution in ζ to

$$\sum_{i=1}^{n} \psi\Big(\frac{X_i - \zeta}{s}\Big) = 0 \tag{33}$$

Can be computed iteratively.

If $\mathbb{E}[X] = \mu$ and $\mathbb{E}(X - \mu)^2 \leqslant \sigma^2$ and $s = \sigma\sqrt{\frac{n}{2\log(4/\delta)}}$ and $\mathcal{O} = \emptyset$

$$\big|\widehat{\mu}_n^{\mathrm{CH}} - \mu\big| \leqslant \sigma\sqrt{\frac{8\log(4/\delta)}{n}} \tag{34}$$

with probability at least $1 - \delta$.

CGD convergence with deterministic coordinates

$$\begin{cases} \theta_j^{(t+1)} \leftarrow \theta_j^{(t)} - \beta_j \widehat{g}_j(\theta^{(t)}) & \text{if } j = j_t \\ \theta_j^{(t+1)} \leftarrow \theta_j^{(t)} & \text{otherwise} \end{cases} \tag{35}$$

Theorem

Assume $\mathcal{L}(\theta)$ is λ-strongly convex with coordinate Lipschitz smooth constants L_j. Let $\theta^{(T)}$ be CGD's output with deterministic coordinates j_t s.t. $\forall t$

$$\{j_{td+1}, j_{td+2}, \ldots, j_{(t+1)d-1}\} = [\![d]\!] \tag{36}$$

Then, we have w. p. $\geqslant 1 - \delta$,

$$\mathcal{L}(\theta^{(T)}) - \mathcal{L}^\star \leqslant \left(\mathcal{L}(\theta^{(0)}) - \mathcal{L}^\star\right)\left(1 - 2\lambda\kappa\right)^T + \frac{3}{8\lambda\kappa L_{\min}} \left\|\epsilon(\delta)\right\|_2^2 \tag{37}$$

where $\kappa = \frac{1}{8L_{\max}(1+d(L_{\max}/L_{\min}))}$ and $(L_{\min}, L_{\max}) = (\min_j L_j, \max_j L_j)$.

Robust CGD without strong convexity

$$\begin{cases} \theta_j^{(t+1)} \leftarrow \mathsf{proj}_{\Theta_j}\left(\theta_j^{(t)} - \beta_j \tau_{\epsilon_j}\left(\widehat{g}_j(\theta^{(t)})\right)\right) & \text{if } j = j_t \\ \theta_j^{(t+1)} \leftarrow \theta_j^{(t)} & \text{otherwise,} \end{cases} \tag{38}$$

with soft-thresholding $\tau_\epsilon(x) = \mathrm{sign}(x)(|x| - \epsilon)_+$ with $(x)_+ = \mathsf{max}(x, 0)$.

Theorem

Assume $\mathcal{L}$ is L_j-smooth along coordinates j, we have w. p. $\geqslant 1 - \delta$

$$\mathbb{E}\left[\mathcal{L}(\theta^{(T)})\right] - \mathcal{L}^\star \leqslant \frac{d}{T+1}\left(\sum_{j \in [\![d]\!]} \frac{L_j}{2}\left(\theta_j^{(0)} - \theta_j^\star\right)^2 + \mathcal{L}(\theta^{(0)})\right)$$
$$+ \frac{2\|\epsilon(\delta)\|_2}{T+1} \sum_{t=0}^{T} \|\theta^{(t)} - \theta^\star\|_2,$$

Moreover, we have $\left\|\theta^{(t)} - \theta^\star\right\|_2 \leqslant \|\theta^{(t-1)} - \theta^\star\|_2$

$\epsilon(\delta)$: **error vector** as before s.t. $\mathbb{P}\left[\sup_\theta \left|\widehat{g}_j(\theta) - g_j(\theta)\right| \leqslant \epsilon_j(\delta)\right] \geqslant 1 - \delta$.

Mirror descent iteration

Let $\omega(\theta)$ distance generating function.

For reference point $\theta_0 \in \Theta$, let $\omega_{\theta_0}(\theta) = \omega(\theta - \theta_0)$ and define the Bregman divergence :

$$V_{\theta_0}(\theta, \theta') = \omega_{\theta_0}(\theta) - \omega_{\theta_0}(\theta') - \langle \nabla \omega_{\theta_0}(\theta'), \theta - \theta' \rangle. \tag{39}$$

For step-size β, proximal step is :

$$\widehat{\text{prox}}_\beta(u, \theta; \theta_0, \Theta) = \underset{\theta' \in \Theta}{\arg\min} \{ \langle \beta u, \theta' \rangle + \beta \bar{\epsilon} \|\theta' - \theta\| + V_{\theta_0}(\theta', \theta) \}. \tag{40}$$

u : gradient at θ, $\bar{\epsilon}$: gradient error.

Approximate Multistage Mirror Descent (AMMD) algorithm

- *Initialization*: $\theta^{(0)}$ and $R > 0$ s.t. $\theta^\star \in \Theta := B_{\|\cdot\|}(\theta_0, R)$. Number of stages $K > 0$. Step size $\beta \leqslant 1/L$. Constant κ. High probability upperbound $\bar{\epsilon}$ on $\|\widehat{g}(\theta) - g(\theta)\|_*$. Upperbound $\bar{s}$ on the sparsity s.
- Set $R_0 = R$.
- Loop over stages $k = 1 \dots K$:
  - Set $\theta_0^{(k)} = \theta^{(k-1)}$ and $\Theta_k = B_{\|\cdot\|}(\theta_0^{(k)}, R_{k-1})$.
  - Run iteration

$$\theta_{t+1}^{(k)} = \widehat{\mathsf{prox}}_\beta\big(\widehat{g}(\theta_t^{(k)}), \theta_t^{(k)}; \theta_0^{(k)}, \Theta_k\big), \tag{41}$$

    for T_k steps with $T_k = \left\lceil \frac{\nu R_{k-1}}{\beta \bar{\epsilon}} \right\rceil$.
  - Set $\theta^{(k)} = \mathsf{sparse}_{\bar{s}}(\widetilde{\theta}^{(k)})$ where $\widetilde{\theta}^{(k)} = \theta_{T_k}^{(k)}$.
  - Set $R_k = \frac{1}{2}\big(R_{k-1} + \frac{40\bar{s}\bar{\epsilon}}{\kappa}\big)$.
- *Output*: The final stage estimate $\theta^{(K)}$.

Dual averaging iteration

For initial $\theta_0 \in \Theta$, step-sizes $(a_i)_{i \geqslant 0}$ and non-decreasing scaling factors $(\gamma_i)_{i \geqslant 0}$, Dual Averaging is defined by :

$$s_t = \frac{1}{A_t} \sum_{i=0}^{t} a_i \widehat{g}_i \quad \text{with} \quad \widehat{g}_i = \widehat{g}(\theta_i) \quad \forall i = 0, \dots, T.$$

$$A_t = \sum_{i=0}^{t} a_i \quad \text{and} \quad \theta_t^+ = \operatorname*{arg\,min}_{\theta \in \Theta} A_t \langle s_t, \theta \rangle + \gamma_t \omega(\theta).$$

$$\theta_{t+1} = (1 - \tau_t)\theta_t + \tau_t \theta_t^+ \qquad \text{where} \qquad \tau_t = \frac{a_{t+1}}{A_{t+1}}.$$

$\widehat{g}_i$: gradient estimator at θ_i.

Approximate Multistage Dual Averaging (AMDA) algorithm

- *Initialization*: θ_0 and $R > 0$ s.t. $\theta^\star \in \Theta := B_{\|\cdot\|}(\theta_0, R)$.
  Pseudolinear minorization constants κ, λ.
  High probability upperbound $\bar{\epsilon}$ on $\|\widehat{g}(\theta) - g(\theta)\|_*$.
  Upperbound $\bar{s}$ on the sparsity s.
- Set $R_0 = R$ and $\tau = \frac{10\sqrt{8\bar{s}}\bar{\epsilon}}{\kappa}$ and $R^\star = \frac{80\lambda\bar{s}\bar{\epsilon}}{\kappa}$.
- Set $k = 0$ and the per stage number of iterations $T' = \left\lceil \left(\frac{\nu + M^2}{\bar{\epsilon}}\right)^2 \right\rceil$
- For $k = 1, \ldots, K$:
  - Set $\theta_0^{(k)} = \theta^{(k-1)}$ and $\Theta_k = B_{\|\cdot\|}(\theta_0^{(k)}, R_{k-1})$.
  - Run Dual averaging with prox-function $\omega_{\theta^{(k-1)}}$, $a_i = R_{k-1}$ and $\gamma_i = \sqrt{i+1}$ for T' iterations.
  - Set $\theta^{(k)} = \mathbf{sparse}_{\bar{s}}(\widetilde{\theta}^{(k)})$ where $\widetilde{\theta}^{(k)} := \theta_{T'}^{(k)}$.
  - Set $R_k = \mathbf{max}\left(\tau R_{k-1}, \frac{1}{2}(R_{k-1} + R^\star)\right)$.
- *Output*: The final stage estimate $\theta^{(K)}$.

Quantile clipped SGD additional details (1)

$\|\nu\|_{\mathsf{TV}} := \sup_{A \in \mathcal{B}(\mathbb{R}^d)} \nu(A) - \inf_{A \in \mathcal{B}(\mathbb{R}^d)} \nu(A)$ **total variation** of signed measure $\nu \rightsquigarrow$ TV distance $\|\nu_1 - \nu_2\|_{\mathsf{TV}}$

Conditions for geometric ergodicity of QC-SGD :

Assume $\mathcal{L}$ μ-strongly convex and L-smooth, corrupted gradients $G(\theta_t, \zeta_t) = G_t = U_t \breve{G}_t + (1 - U_t)\widetilde{G}_t$, $U_t \sim^{iid} \mathcal{B}(\eta)$ with moment bound

$$\mathbb{E}\left[\left\|\widetilde{G}_t - \nabla \mathcal{L}(\theta_t)\right\|^q \mid \theta_t\right]^{1/q} \leqslant A_q \|\theta_t - \theta^\star\| + B_q \tag{42}$$

Assume there is $p \in [\eta, 1 - \eta]$ such that

$$\kappa := (1 - \eta)p\mu - \eta L - (1 - p)^{-\frac{1}{q}} A_q (1 - p(1 - \eta)) > 0. \tag{43}$$

Then, for step-size $\beta < \frac{1}{4} \frac{\kappa}{\mu^2 + 2L^2 + 4\eta^{-\frac{2}{q}} A_q^2}$, for $\theta_0 \in \mathbb{R}^d$, there is $\rho < 1$ and $M < \infty$ such that after T iterations

$$\left\|\delta_{\theta_0} P_{\beta,p}^T - \pi_{\beta,p}\right\|_{\mathsf{TV}} \leqslant M \rho^T \left(1 + \|\theta_0 - \theta^\star\|^2\right). \tag{44}$$

Quantile clipped SGD additional details (2)

Conditions for ergodicity for non-convex objectives :

Assume $\mathcal{L}$ is L-smooth, corrupted gradient and moment condition as before. For step size β and quantile $p \in [\eta, 1-\eta]$. Assume that p and β are such that $3p(1-\eta)/4 > L\beta + \eta$ and that the set

$$\left\{\theta \in \mathbb{R}^d : \frac{1}{2}\left\|\nabla \mathcal{L}(\theta)\right\|^2 \leqslant \frac{B_q^2\left((1-p)^{-\frac{2}{q}}(L\beta + 2\eta^2) + 2\eta^{2-\frac{2}{q}}\right)}{p(1-\eta)(3p(1-\eta)/4 - L\beta - \eta)}\right\} \tag{45}$$

is bounded. Then, for any initial $\theta_0 \in \mathbb{R}^d$, there exists $M < +\infty$ such that after T iterations

$$\left\|\delta_{\theta_0} P_{\beta,p}^T - \pi_{\beta,p}\right\|_{\mathsf{TV}} \leqslant \frac{M}{T}, \tag{46}$$

where δ_{θ_0} is the Dirac at θ_0 and $\pi_{\beta,p}$ is s.t. for $p = 1-\eta$ and $\beta \leqslant \eta^2/L$ we have

$$\mathbb{E}\left\|\nabla \mathcal{L}(\theta)\right\|^2 \leqslant \frac{5\eta^{2-\frac{2}{q}} B_q^2}{p(1-\eta)\big(3p(1-\eta)/4 - L\beta - \eta\big)}. \tag{47}$$

Key lemma for QC-SGD proofs

Under previous assumptions, fix $\theta \in \mathbb{R}^d$ and let $G(\theta) = \widetilde{G}(\theta)$ be non corrupted.

Setting the clipping level as $\tau_\theta = Q_p(\|\widetilde{G}(\theta)\|)$ for $p \in (0,1)$ and denoting

$$\alpha_\theta = \min\left(1, \frac{\tau_\theta}{\|G(\theta)\|}\right) \quad \text{and} \quad \overline{\alpha}_\theta = \mathbb{E}\big[\alpha_\theta | \theta, G(\theta) = \widetilde{G}(\theta)\big], \tag{48}$$

the clipping factor and its average. We have:

$$\big\|\mathbb{E}[\alpha_\theta \widetilde{G}(\theta)] - \overline{\alpha}_\theta \nabla \mathcal{L}(\theta)\big\| \leqslant (1-p)^{1-1/q}\big(A_q\|\theta - \theta^\star\| + B_q\big), \tag{49}$$

and

$$\tau_\theta \leqslant \big\|\nabla \mathcal{L}(\theta)\big\| + (1-p)^{-1/q}\big(A_q\|\theta - \theta^\star\| + B_q\big).$$

QC-SGD results for vector mean estimation

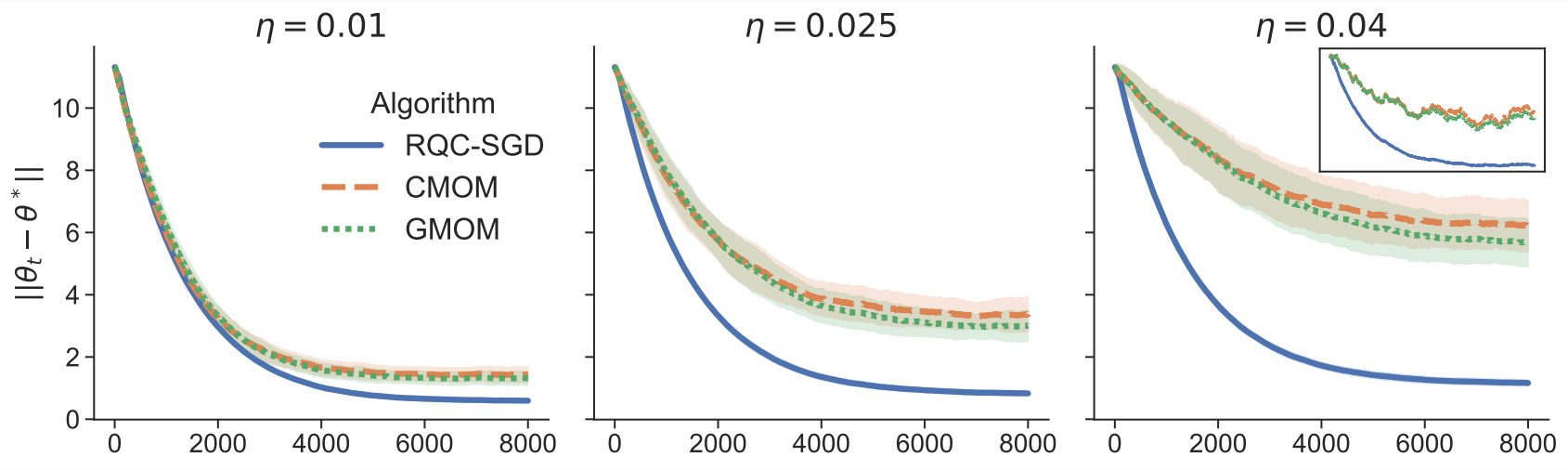


Figure 6: Comparison of QC-SGD with CMOM and GMOM for vector mean estimation via stochastic optimisation. Data are heavy-tailed and η-corrupted.

The geometric and coordinatewise Median-Of-Means estimators (GMOM and CMOM) optimize the following objectives respectively:

$$\mathbb{E}\left\|\theta - \bar{X}_{N_b}\right\|_2 \quad \text{and} \quad \mathbb{E}\left\|\theta - \bar{X}_{N_b}\right\|_1, \tag{50}$$

where $\bar{X}_{N_b}$ is the average of N_b independent copies of X.

Wasserstein contraction for linear regression

For linear regression with samples $\zeta = (X, \xi) \in \mathbb{R}^d \times \mathbb{R}$, the (random) gradient is

$$G(\theta, \zeta) = XX^\top \theta - XY \quad \text{with} \quad Y = X^\top \theta^\star + \xi, \tag{51}$$

and the Wasserstein Lipschitz condition holds as long as X is in L^4.

We have $\varepsilon_\zeta(\theta) = G(\theta, \zeta) - \nabla \mathcal{L}(\theta) = (XX^\top - \mathbb{E}XX^\top)(\theta - \theta^\star) - X\xi$.

We couple $\varepsilon_\zeta(\theta')$ and $\varepsilon_{\zeta'}(\theta')$ by using the same variables $\zeta = \zeta' = (X, \xi)$ so that

$$\begin{aligned} \mathcal{W}_2^2\big(\mathcal{D}(\varepsilon_\zeta(\theta)), \mathcal{D}(\varepsilon_{\zeta'}(\theta'))\big) &\leqslant \mathbb{E}\|\varepsilon_\zeta(\theta) - \varepsilon_{\zeta'}(\theta')\|^2 \\ &\leqslant \mathbb{E}\|XX^\top - \mathbb{E}XX^\top\|_2^2 \|\theta - \theta'\|^2. \end{aligned}$$

A more general theorem for SGD concentration

Let $\mathcal{L}$ be smooth and strongly convex and the gradient errors be K-sub-Gaussian. Assume the Wasserstein Lipschitz condition holds.

Let $f : \Theta^n \to \mathbb{R}$ be a 1-Lipschitz w.r.t. each parameter. Let $\vec{\theta} := (\theta_1, \ldots, \theta_n)$ stationary SGD iterates i.e. such that $\theta_1 \sim \pi_\beta$.

Then $f(\vec{\theta}) - \mathbb{E}f(\vec{\theta})$ is sub-Gaussian with constant $KC_{\mathcal{W}}\sqrt{\beta/\mu + (n-1)\beta^2}$ where $C_{\mathcal{W}} = \left(1 - \sqrt{(1-\beta\mu)^2 + \beta^2 L_{\mathcal{W}}}\right)^{-1}$.

If the gradient errors are Ksub-exponential then $f(\vec{\theta}) - \mathbb{E}f(\vec{\theta})$ is sub-exponential with constant $KC_{\mathcal{W}}\sqrt{\beta/\mu + (n-1)\beta^2}$.

Key Lemma for Polyak-Ruppert bound

Assume $\mathcal{L}$ is smooth and strongly convex, the gradient is linear and the Wasserstein Lipschitz condition holds.

The Markov chain $(\theta_t)_{t\geqslant 0}$ started from $\theta_0 \sim \nu$, satisfies for $0 \leqslant i, j \leqslant n$:

$$\mathbb{E}\langle \theta_i - \theta^\star, \theta_j - \theta^\star \rangle \leqslant 2(1-\beta\mu)^{|i-j|}\left(\left((1-\beta\mu)+\beta^2 L_{\mathcal{W}}\right)^i \mathcal{W}_2^2(\nu,\pi) + \mathsf{Var}_{\pi_\beta}(\theta)\right). \tag{52}$$

where π_β is the invariant limit distribution.

$\hookrightarrow$ covariance decays geometrically between iterates

References

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

Olivier Catoni. *Statistical Learning Theory and Stochastic Optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag Berlin Heidelberg, 2004. ISBN 9783540445074. URL *https://books.google.fr/books?id=3Ih8CwAAQBAJ*.

Ilias Diakonikolas, Daniel M Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. *Advances in Neural Information Processing Systems*, 33:1830–1840, 2020.

References ii

Sara A. Geer and Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

Matthew Holland and Kazushi Ikeda. Better generalization with less data using robust gradient descent. In *International Conference on Machine Learning*, pages 2761–2770. PMLR, 2019.

Anatoli Juditsky, Andrei Kulunchakov, and Hlib Tsyntseus. Sparse recovery by reduced variance stochastic approximation. *Information and Inference: A Journal of the IMA*, 12(2):851–896, 11 2022. ISSN 2049-8772. doi: 10.1093/imaiai/iaac028. URL *https://doi.org/10.1093/imaiai/iaac028*.

References iii

Guillaume Lecué, Matthieu Lerasle, and Timlothée Mathieu. Robust classification via MOM minimization. *Machine Learning*, 109(8): 1635–1665, 2020.

Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

Sean P Meyn and Richard L Tweedie. *Markov Chains and Stochastic Stability*. Springer London, 1993.

Stanislav Minsker. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A): 2871–2903, 2018.

References iv

Stanislav Minsker et al. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Le Nguyen. High probability convergence of clipped-SGD under heavy-tailed noise. *arXiv preprint arXiv:2302.05437*, 2023.

Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):601–627, 2020.

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. *arXiv preprint arXiv:2302.00999*, 2023.

References v

Che-Ping Tsai, Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. Heavy-tailed streaming statistical estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 1251–1282. PMLR, 2022.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer New York, NY, 1999.