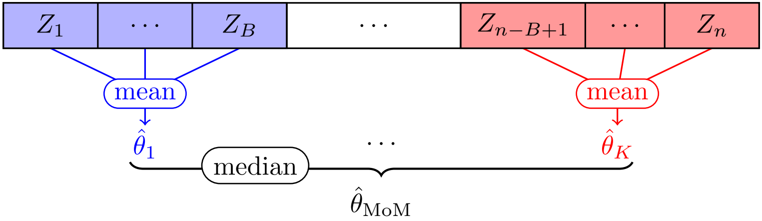
Generalization Bounds in the Presence of Outliers: a Median-of-Means Study

Pierre Laforgue[1], Guillaume Staerman[2], Stephan Clémençon[2]

[1] Università degli Studi di Milano, Italy

[2] LTCI, Télécom Paris, Institut Polytechnique de Paris, France

The Median-of-Means (MoM) for heavy-tailed data



$Z_1, \ldots, Z_n$ i.i.d. realizations of r.v. Z s.t. $\mathbb{E}[Z] = \theta$, $\mathrm{Var}(Z) = \sigma^2$.

$\forall \delta \in [e^{1-\frac{2n}{9}}, 1[$, for $K = \left\lceil \frac{9}{2} \ln(1/\delta) \right\rceil$ it holds [Devroye et al. 2016]:

$$\mathbb{P}\left\{ \left| \hat{\theta}_{\mathrm{MoM}} - \theta \right| > 3\sqrt{6}\sigma \sqrt{\frac{1 + \ln(1/\delta)}{n}} \right\} \leq \delta.$$

The Median-of-Means (MoM) for outliers

$\{Z_1, \ldots, Z_n\}$ contains $n - n_O$ *inliers* drawn i.i.d. from P, and n_O *outliers*. We denote $\varepsilon = n_O/n$. Choosing $K = \lceil \beta(\varepsilon) \log(1/\delta) \rceil$, we have w.p.a.l. $1 - \delta$:

$$\left|\hat{\theta}_{\mathrm{MoM}} - \theta\right| \leq \frac{12\sqrt{5e}\sigma}{(1-2\varepsilon)^{3/2}} \sqrt{\frac{1 + \log(1/\delta)}{n}}.$$

If in addition P is ρ sub-Gaussian, with $K = \lceil \alpha(\varepsilon) n \rceil$, we have w.p.a.l. $1 - \delta$:

$$\left|\hat{\theta}_{\mathrm{MoM}} - \theta\right| \leq \frac{4\sqrt{5}\rho}{\sqrt{1-2\varepsilon}} \sqrt{\frac{\log(1/\delta)}{n}}.$$

If furthermore $n_O \leq C_O n^{\alpha_O}$, with the same K we have:

$$\mathbb{E}\left[\left|\hat{\theta}_{\mathrm{MoM}} - \theta\right|\right] \leq \frac{2\sqrt{5}\rho}{\sqrt{1-2\varepsilon}} \left(4 C_O \frac{\Delta(\varepsilon)}{n^{(1-\alpha_O)/2}} + \sqrt{\frac{\pi}{n}}\right).$$

Similar guarantees for U-statistics, with application to Integral Probability Metrics [Staerman et al. 2021]

MoU minimization (adaptation fom [Lecué et al. 2018]):

$$\hat{g}_{MoU} = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \ \operatorname{median}\left(\sum_{i<j \in \mathcal{B}_1} \ell(g, Z_i, Z_j), \ldots, \sum_{i<j \in \mathcal{B}_K} \ell(g, Z_i, Z_j) \right).$$

Algorithm 1 MoU Gradient Descent (MoU-GD)

input: $\mathcal{S}_n$, K, $T \in \mathbb{N}^*$, $(\gamma_t)_{t \le T} \in \mathbb{R}_+^T$, $u_0 \in \mathbb{R}^p$

for *epoch from* 1 *to* T **do**

```
// Randomly partition the data
```
Choose a random permutation π of $\{1, \ldots, n\}$

Build a partition $B_1, \ldots, B_k$ of $\{\pi(1), \ldots, \pi(n)\}$

```
// Select block with median risk
```
for $k \le K$ **do**

$\hat{U}_{B_k} = \sum_{i<j \in B_k^2} \ell(g_{u_t}, Z_i, Z_j)$

Set B_{med} s.t. $\hat{U}_{B_{\text{med}}} = \text{median}(\hat{U}_{B_k}, \ldots \hat{U}_{B_K})$

```
// Gradient step
```
$u_{t+1} = u_t - \gamma_t \sum_{i<j \in B_k^2} \nabla_{u_t} \ell(g_{u_t}, Z_i, Z_j)$

return u_T

With guarantees in the presence of outliers:

$$\mathcal{R}(\hat{g}_{\text{alg}}) - \mathcal{R}(g^*) \le \frac{8\sqrt{10}M}{\sqrt{1-2\varepsilon}} \sqrt{\frac{\mathrm{VC}_{\text{dim}}(\mathcal{G})(1+\log(n)) + \log(1/\delta)}{n}}.$$

Numerical experiments

Application to metric learning on the *iris* dataset: