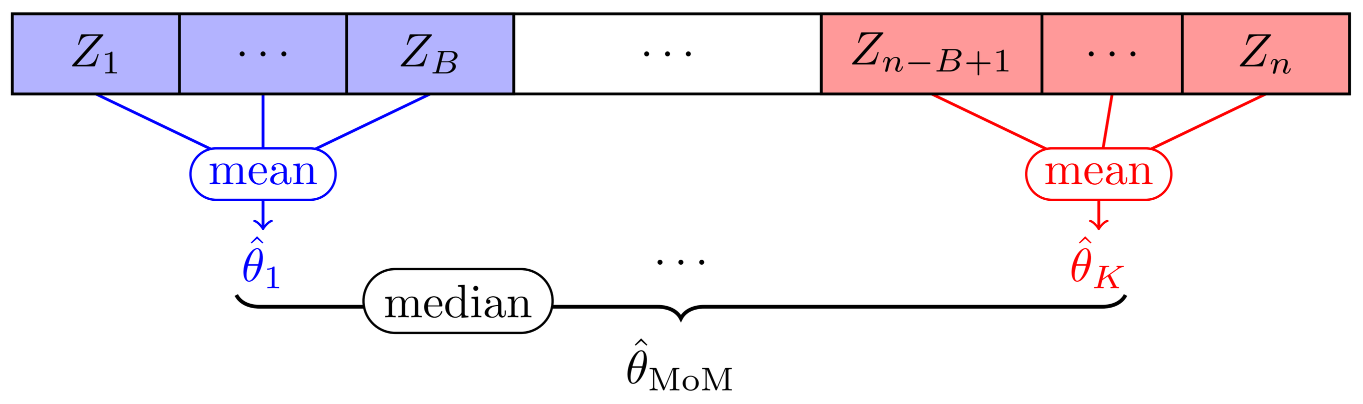
Generalization Bounds in the Presence of Outliers: a Median-of-Means Study

P. Laforgue[1], G. Staerman[2], S. Clémençon[2]

[1] University of Milan, Italy [2] LTCI, Télécom ParisTech, Institut Polytechnique de Paris, France

MoM in the presence of outliers



Assumption 1. *The sample $\mathcal{S}_n = \{Z_1, \ldots, Z_n\}$ contains $n - n_{\mathcal{O}}$ inliers drawn i.i.d. from P, and $n_{\mathcal{O}}$ outliers, upon which no assumption is made. Let $\varepsilon = n_{\mathcal{O}}/n$ be the fraction of outliers among sample $\mathcal{S}_n$.*

Assumption 2. *Let $\alpha\colon [0, 1/2] \to [0, 1]$ be such that: $\forall \varepsilon \in (0, 1/2)$, $2\varepsilon < \alpha(\varepsilon) < 1$. From mapping α, we define the following functions:*

$$\beta\colon \varepsilon \mapsto \frac{2\alpha(\varepsilon)}{\alpha(\varepsilon) - 2\varepsilon}, \qquad \gamma\colon \varepsilon \mapsto \frac{\sqrt{\alpha(\varepsilon)}(\alpha(\varepsilon) - \varepsilon)}{(\alpha(\varepsilon) - 2\varepsilon)^{\frac{3}{2}}},$$

$$\Gamma\colon \varepsilon \mapsto \sqrt{\frac{\alpha(\varepsilon)}{\alpha(\varepsilon) - 2\varepsilon}}, \qquad \Delta\colon \varepsilon \mapsto \sqrt{\frac{\alpha(\varepsilon)}{\varepsilon}}.$$

Proposition 1. *Let $\mathcal{S}_n$ and $\alpha, \beta, \gamma, \Gamma, \Delta$ satisfying Assumptions 1 and 2 respectively. Then, for any $\delta \in [e^{-n/\beta(\varepsilon)}, e^{-n\alpha(\varepsilon)/\beta(\varepsilon)}]$, choosing $K = \lceil \beta(\varepsilon) \log(1/\delta) \rceil$, it holds w.p.a.l. $1 - \delta$:*

$$\left|\hat{\theta}_{\mathrm{MoM}} - \theta\right| \;\le\; 4\sqrt{e}\sigma\ \gamma(\varepsilon)\ \sqrt{\frac{1 + \log(1/\delta)}{n}}.$$

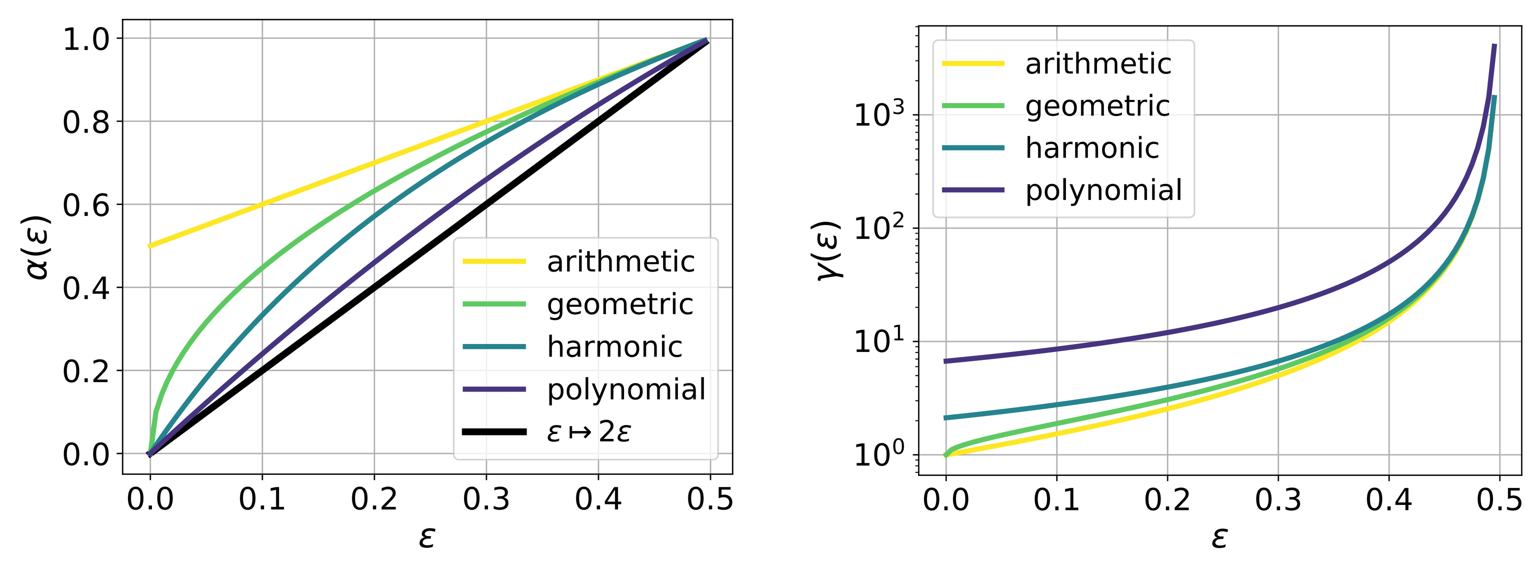
If in addition distribution P is ρ sub-Gaussian, then for all $\delta \in (0, e^{-4n\alpha(\varepsilon)}]$, with $K = \lceil \alpha(\varepsilon) n \rceil$, it holds w.p.a.l. $1 - \delta$:

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

If furthermore $n_{\mathcal{O}} \le C_{\mathcal{O}} n^{\alpha_{\mathcal{O}}}$, the same K gives:

$$\mathbb{E}\left[\left|\hat{\theta}_{\mathrm{MoM}} - \theta\right|\right] \;\le\; 2\rho\ \Gamma(\varepsilon) \left(4C_{\mathcal{O}} \frac{\Delta(\varepsilon)}{n^{(1-\alpha_{\mathcal{O}})/2}} + \sqrt{\frac{\pi}{n}}\right).$$

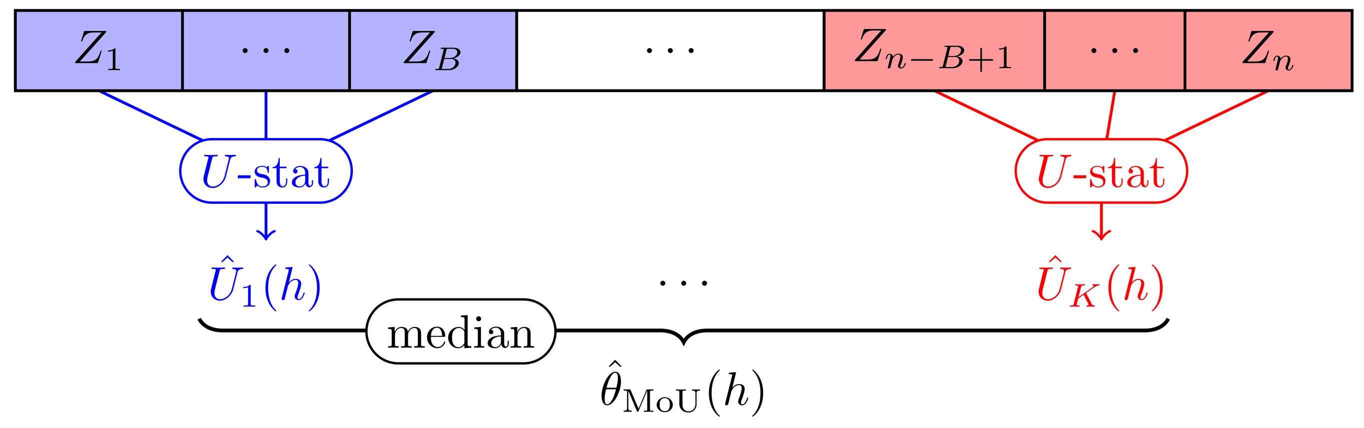
For most α, we have $\gamma(\varepsilon) \le 3\sqrt{5}/(1-2\varepsilon)^{3/2}$ and $\Gamma(\varepsilon) \le \sqrt{5}/\sqrt{1 - 2\varepsilon}$.



MoU in the presence of outliers

A U-statistic is the MVU estimator of $\mathbb{E}\left[h(Z_1, \ldots, Z_d)\right]$. It is given by

$$\hat{U}_n(h) = \frac{1}{\binom{n}{d}} \sum_{1 \le i_1 < \ldots < i_d \le n} h(Z_{i_1}, \ldots, Z_{i_d}).$$

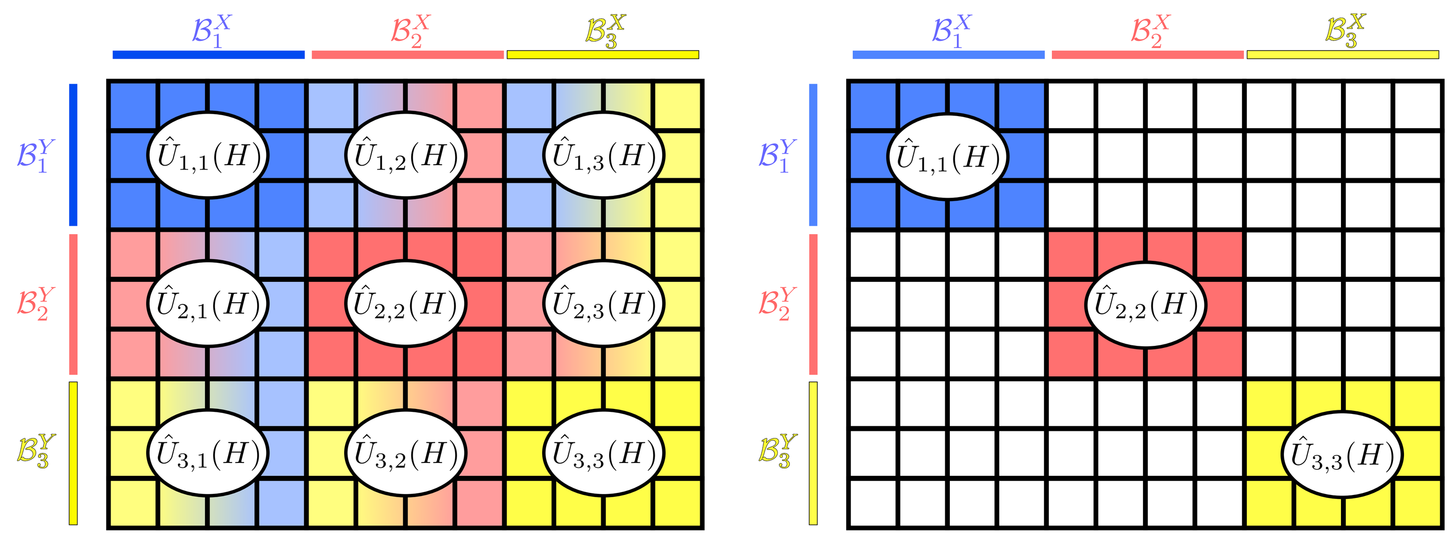


Proposition 2. *Let $\Sigma^2(h) = d! \sum_{c=1}^{d} \binom{d}{c} \zeta_c(h)$, and $M = \|h(Z_1, \ldots, Z_d)\|_\infty$. Then, under the assumptions of Proposition 1, with the same choices of K (and the same ranges of δ), we have wp.a.l. $1 - \delta$:*

$$\left|\hat{\theta}_{\mathrm{MoU}}(h) - \theta(h)\right| \;\le\; 4\sqrt{e}\ \Sigma(h)\ \gamma(\varepsilon)\ \sqrt{\frac{1 + \log(1/\delta)}{n}},$$

$$\left|\hat{\theta}_{\mathrm{MoU}}(h) - \theta(h)\right| \;\le\; 4\sqrt{d}\ M\ \Gamma(\varepsilon)\ \sqrt{\frac{\log(1/\delta)}{n}},$$

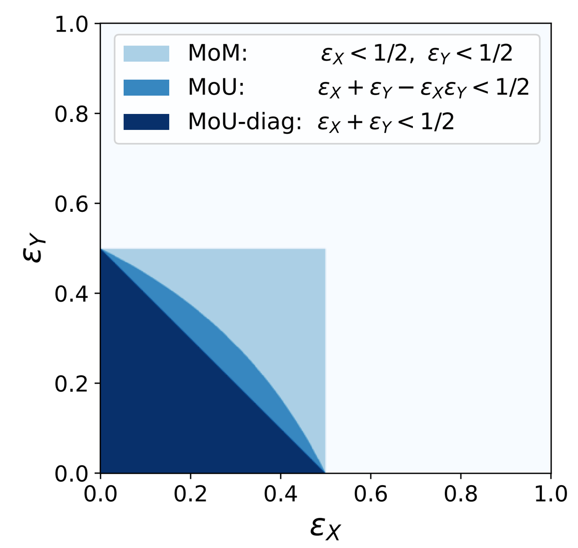
$$\mathbb{E}\left[\left|\hat{\theta}_{\mathrm{MoU}}(h) - \theta(h)\right|\right] \le 2\sqrt{d}\ M\ \Gamma(\varepsilon) \left(4C_{\mathcal{O}} \frac{\Delta(\varepsilon)}{n^{(1-\alpha_{\mathcal{O}})/2}} + \sqrt{\frac{\pi}{n}}\right).$$



The 2-sample U-statistic of degrees $(1, 1)$ is defined as $\hat{U}_{n,m}(H) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} H(X_i, Y_j)$. For IPMs, $h(X, Y) = \sup_\phi\ \phi(X) - \phi(Y)$.

Results are similar to the ones in Propositions 1 and 2, but with the following constants, and the notation $\tilde{\varepsilon} = \varepsilon_X + \varepsilon_Y - \varepsilon_X \varepsilon_Y$.

Asm.	Complete	Diag.
None	$\gamma(\tilde{\varepsilon})$	$\gamma(\varepsilon_X + \varepsilon_Y)$
$\|H\|_\infty < +\infty$	$\emptyset$	$\Gamma(\varepsilon_X + \varepsilon_Y)$
$n_{\mathcal{O}} = \mathcal{O}(n^{\alpha_{\mathcal{O}}})$	$\emptyset$	$\Delta(\varepsilon_X + \varepsilon_Y)$



Outlier robust pairwise learning

Goal: $g^* = \operatorname{argmin}_{g \in \mathcal{G}} \left\{ \mathcal{R}(g) = \mathbb{E}\left[\ell_g(Z, Z')\right] \right\}$ (ranking, metric learning)

Approach: $\hat{g}_{\mathrm{MoU}} = \operatorname{argmin}_{g \in \mathcal{G}}\ \operatorname{median}\Big(\sum_{i<j \in \mathcal{B}_k} \ell_g(Z_i, Z_j),\ \text{for } k \le K \Big)$

Algorithm: adapted from Lecué et al. 2018

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. $\hat{U}_{B_{\mathrm{med}}} = \operatorname{median}(\hat{U}_{B_k}, \ldots \hat{U}_{B_K})$
  // 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

Theorem 1. *Suppose that $\ell_g(Z, Z') < M$, and that $\mathcal{G}$ has finite VC-dimension. Under some technical assumptions taken from Lecué et al. 2018, the output of Algorithm 1 run on a corrupted sample $\mathcal{S}_n$ converges almost surely towards $\hat{g}_{\mathrm{alg}}$, that satisfies w.p.a.l. $1 - \delta$:*

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