Median-of-Means Based Learning Techniques

The Median-of-Means Estimator

Preliminaries

Sample $\mathcal{S}_n = \{Z_1, \ldots, Z_n\} \sim Z$ i.i.d. such that $\mathbb{E}[Z] = \theta$

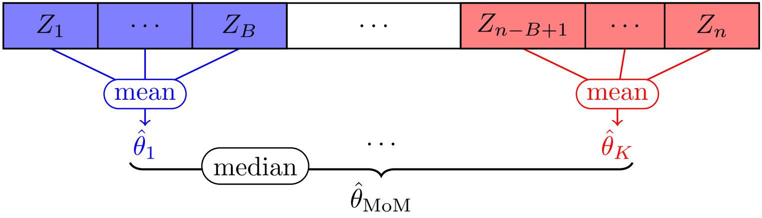
- $\hat{\theta}_{\text{avg}} = \frac{1}{n}\sum_{i=1}^{n} Z_i$
- $\hat{\theta}_{\text{med}} = Z_{\sigma(\frac{n+1}{2})}$, with $Z_{\sigma(1)} \leq \ldots \leq Z_{\sigma(n)}$
- Deviation Probabilities [Catoni, 2012]: $\mathbb{P}\left\{|\hat{\theta} - \theta| > t\right\}$.
- If Z is **bounded** (see Hoeffding's Inequality) or sub-Gaussian:

$$\mathbb{P}\left\{\left|\hat{\theta}_{\text{avg}} - \theta\right| > \sigma\sqrt{\frac{2\ln(2/\delta)}{n}}\right\} \leq \delta.$$

Do estimators exist with same guarantees under weaker assumptions?

How to use them to perform (robust) learning?

The Median-of-Means

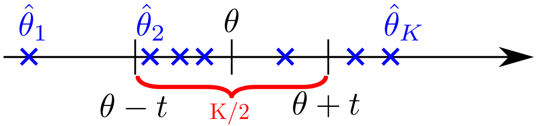


$Z_1, \ldots, Z_n$ i.i.d. realizations of r.v. Z s.t. $\mathbb{E}[Z] = \theta$, $\mathrm{Var}(Z) = \sigma^2$.

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

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

Proof



$$\hat{\theta}_k = \frac{1}{B} \sum_{i \in B_k} Z_i, \qquad \hat{I}_{k,t} = \mathbb{I}\left\{|\hat{\theta}_k - \theta| > t\right\}, \qquad \hat{p}_t = \mathbb{E}[\hat{I}_{1,t}] = \mathbb{P}\left\{|\hat{\theta}_1 - \theta| > t\right\}$$

$$\begin{aligned}
\mathbb{P}\left\{\left|\hat{\theta}_{\mathrm{MoM}} - \theta\right| > t\right\} &\leq \mathbb{P}\left\{\sum_{k=1}^{K} \hat{I}_{k,t} \geq \frac{K}{2}\right\} \leq \mathbb{P}\left\{\frac{1}{K}\sum_{k=1}^{K}(\hat{I}_{k,t} - p_t) \geq \frac{1}{2} - \frac{\sigma^2}{Bt^2}\right\}, \\
&\leq \exp\left(-2K\left(\frac{1}{2} - \frac{\sigma^2}{Bt^2}\right)^2\right), \\
&\leq \delta \text{ for } K = \frac{9\ln(1/\delta)}{2} \text{ and } \frac{\sigma^2}{Bt^2} = \frac{1}{6} \Leftrightarrow t = 3\sqrt{3}\sigma\sqrt{\frac{\ln(1/\delta)}{n}}.
\end{aligned}$$

U-statistics & pairwise learning

Estimator of $\mathbb{E}[h(Z, Z')]$ with minimal variance, defined from an i.i.d. sample $Z_1, \ldots, Z_n$ as:

$$U_n(h) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} h(Z_i, Z_j).$$

Ex: the empirical variance when $h(z, z') = \frac{(z-z')^2}{2}$.

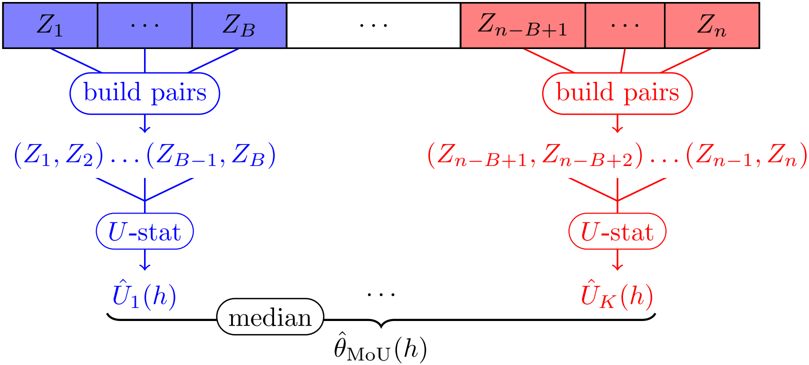
Encountered *e.g.* in **pairwise ranking** and **metric learning**:

$$\widehat{\mathcal{R}}_n(r) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \mathbb{I}\left\{ r(X_i, X_j) \cdot (Y_i - Y_j) \leq 0 \right\}.$$

$$\widehat{\mathcal{R}}_n(d) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \mathbb{I}\left\{ Y_{ij} \cdot (d(X_i, X_j) - \epsilon) > 0 \right\}.$$

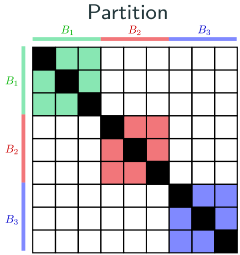
How to extend MoM to U-statistics?

The Median-of-U-statistics



w.p. $1-\delta$, $\quad \left|\hat{\theta}_{\mathrm{MoU}}(h) - \theta(h)\right| \leq C_1(h)\sqrt{\frac{1+\ln(1/\delta)}{n}} + C_2(h)\,\frac{1+\ln(1/\delta)}{n}$

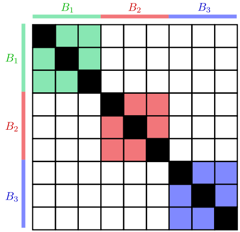
Why randomization?



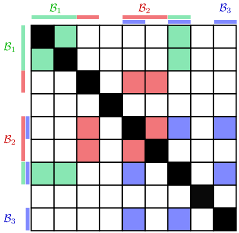
Build all possible blocks
[Joly and Lugosi, 2016]

Why randomization?

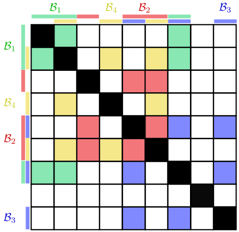
Partition



Random Blocks

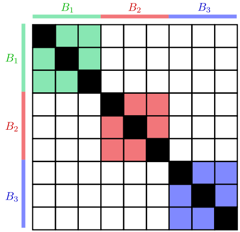


Build all possible blocks
[Joly and Lugosi, 2016]

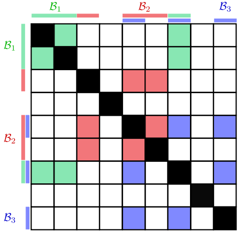


Why randomization?

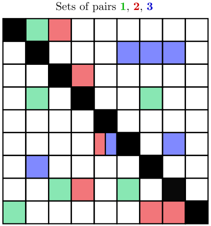
Partition



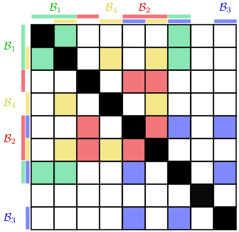
Random Blocks

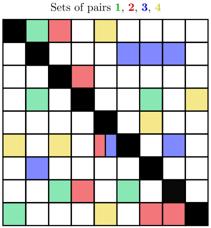


Random Pairs



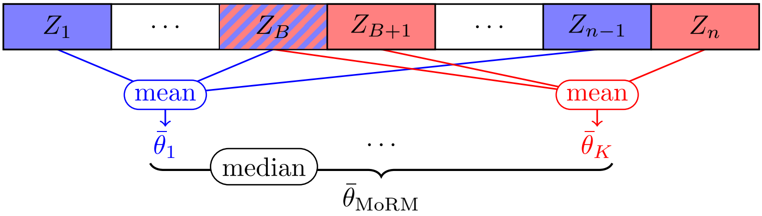
Build all possible blocks
[Joly and Lugosi, 2016]





Randomization allows for a better exploration

The Median-of-Randomized-Means [Laforgue et al., 2019]



With blocks formed by SWoR, $\forall\ \tau \in]0, 1/2[,\ \forall\ \delta \in [2e^{-\frac{8\tau^2 n}{9}}, 1[$, set

$K := \left\lceil \frac{\ln(2/\delta)}{2(1/2-\tau)^2} \right\rceil$, and $B := \left\lfloor \frac{8\tau^2 n}{9 \ln(2/\delta)} \right\rfloor$, it holds:

$$\mathbb{P}\left\{ \left| \bar{\theta}_{\text{MoRM}} - \theta \right| > \frac{3\sqrt{3}\ \sigma}{2\ \tau^{3/2}} \sqrt{\frac{\ln(2/\delta)}{n}} \right\} \leq \delta.$$

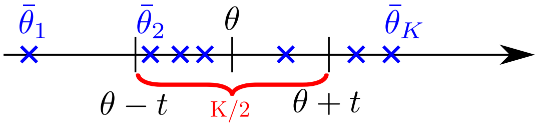
Proof

Random block $\mathcal{B}_k$ characterized by random vector $\boldsymbol{\epsilon}_k = (\epsilon_{k,1}, \ldots, \epsilon_{k,n}) \in \{0,1\}^n$ i.i.d. uniformly over $\Lambda_{n,B} = \left\{\boldsymbol{\epsilon} \in \{0,1\}^n : \mathbf{1}^\top \boldsymbol{\epsilon} = B\right\}$, of cardinality $\binom{n}{B}$.

$$\bar{\theta}_k = \frac{1}{B} \sum_{i \in \mathcal{B}_k} Z_i, \qquad \bar{I}_{\boldsymbol{\epsilon}_k,t} = \mathbb{I}\{|\bar{\theta}_k - \theta| > t\}, \qquad \bar{p}_t = \mathbb{E}[\bar{I}_{\boldsymbol{\epsilon}_k,t}] = \mathbb{P}\left\{|\bar{\theta}_1 - \theta| > t\right\}$$

$$\bar{U}_{n,t} = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\frac{1}{K}\sum_{k=1}^{K} \bar{I}_{\boldsymbol{\epsilon}_k,t} \middle| \mathcal{S}_n\right] = \frac{1}{\binom{n}{B}} \sum_{\boldsymbol{\epsilon} \in \Lambda(n,B)} \bar{I}_{\boldsymbol{\epsilon},t} = \frac{1}{\binom{n}{B}} \sum_{I} \mathbb{I}\left\{\left|\frac{1}{B}\sum_{j=1}^{B} X_{I_j} - \theta\right| > t\right\}$$

$$\mathbb{P}\left\{\left|\bar{\theta}_{\text{MoRM}} - \theta\right| > t\right\} \leq \mathbb{P}\left\{\frac{1}{K}\sum_{k=1}^{K} \bar{I}_{\boldsymbol{\epsilon}_k,t} \qquad\qquad \geq \frac{1}{2} \qquad\qquad \right\},$$



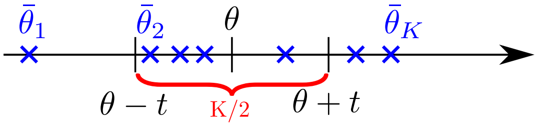
Proof

Random block $\mathcal{B}_k$ characterized by random vector $\boldsymbol{\epsilon}_k = (\epsilon_{k,1}, \ldots, \epsilon_{k,n}) \in \{0,1\}^n$ i.i.d. uniformly over $\Lambda_{n,B} = \left\{\boldsymbol{\epsilon} \in \{0,1\}^n : \mathbf{1}^\top \boldsymbol{\epsilon} = B\right\}$, of cardinality $\binom{n}{B}$.

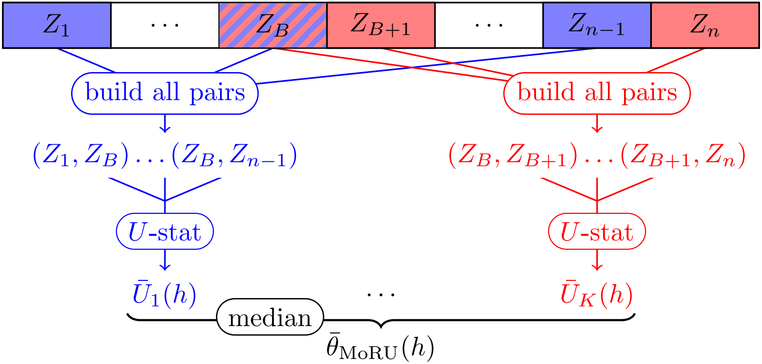
$$\bar\theta_k = \frac{1}{B}\sum_{i\in\mathcal{B}_k} Z_i, \qquad \bar I_{\boldsymbol{\epsilon}_k,t} = \mathbb{I}\{|\bar\theta_k-\theta|>t\}, \qquad \bar p_t = \mathbb{E}[\bar I_{\boldsymbol{\epsilon}_k,t}] = \mathbb{P}\left\{|\bar\theta_1-\theta|>t\right\}$$

$$\bar U_{n,t} = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\frac{1}{K}\sum_{k=1}^K \bar I_{\boldsymbol{\epsilon}_k,t}\middle|\mathcal{S}_n\right] = \frac{1}{\binom{n}{B}}\sum_{\boldsymbol{\epsilon}\in\Lambda(n,B)} \bar I_{\boldsymbol{\epsilon},t} = \frac{1}{\binom{n}{B}}\sum_{I}\mathbb{I}\left\{\left|\frac{1}{B}\sum_{j=1}^B X_{I_j}-\theta\right|>t\right\}$$

$$\begin{aligned}\mathbb{P}\left\{\left|\bar\theta_{\text{MoRM}}-\theta\right|>t\right\} &\leq \mathbb{P}\left\{\frac{1}{K}\sum_{k=1}^K \bar I_{\boldsymbol{\epsilon}_k,t} - \bar U_{n,t} + \bar U_{n,t} - \bar p_t \geq \frac{1}{2} - \bar p_t + \tau - \tau\right\}, \\ &\leq \exp\left(-2K\left(\frac{1}{2}-\tau\right)^2\right) + \exp\left(-2\frac{n}{B}\left(\tau - \frac{\sigma^2}{Bt^2}\right)^2\right).\end{aligned}$$



The Median-of-Randomized-U-statistics [Laforgue et al., 2019]



w.p.a.l. $1-\delta$, $\quad \left|\bar{\theta}_{\mathrm{MoRU}} - \theta(h)\right| \leq C_1(h,\tau)\sqrt{\frac{\ln(2/\delta)}{n}} + C_2(h,\tau)\,\frac{\ln(2/\delta)}{n}$

Robustness to outliers (1/2) [Laforgue et al., 2020]

The sample $\mathcal{D}_n = \{Z_1, \ldots, Z_n\}$ is now composed of $n - n_o > n/2$ realizations of the r.v. Z and n_o outliers with arbitrary distributions.

Let $\tau = n_o/n < 1/2$, define $\alpha(\tau) = 4\tau/(1 + 2\tau)$, and $\beta(\tau) = 4/(1 - 2\tau)$.

Then, for all $\delta \in [\exp(-n/\beta(\tau)), \exp(-n\alpha(\tau)/\beta(\tau))]$, choosing $K = \lceil \beta(\tau) \log(1/\delta) \rceil$, it holds with probability at least $1 - \delta$:

$$\left|\hat{\theta}_{\mathrm{MoM}} - \theta\right| \leq \frac{6\sqrt{2e}\,\sigma}{(1-2\tau)^{3/2}} \sqrt{\frac{1 + \log(1/\delta)}{n}}.$$

Robustness to outliers (2/2) [Laforgue et al., 2020]

If in addition the r.v. Z is sub-Gaussian with parameter $\rho > 0$, then for all $\delta \in]0, \exp(-4n\alpha(\tau))]$, with $K = \lceil \alpha(\tau) n \rceil$, it holds w.p.a.l. $1 - \delta$:

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

If finally it also holds $n_o \leq C_o^2 n^{\alpha_o}$, with $\alpha_o < 1$, we have:

$$\mathbb{E}\left|\hat{\theta}_{\mathrm{MoM}} - \theta\right| \;\leq\; \frac{2\sqrt{2}\,\rho}{\sqrt{1-2\tau}}\,\left(\frac{8\,C_o}{n^{(1-\alpha_o)/2}} + \sqrt{\frac{\pi}{n}}\right).$$

References (estimator concentration)

- First mentions [Nemirovsky and Yudin, 1983, Jerrum et al., 1986, Alon et al., 1999]
- Deviation study [Catoni, 2012], concentration [Devroye et al., 2016]
- Multi-D [Minsker et al., 2015, Hsu and Sabato, 2016, Lugosi and Mendelson, 2017]
- *U*-stats and randomized blocks [Joly and Lugosi, 2016, Laforgue et al., 2019]
- Robustness to outliers [Depersin and Lecué, 2019, Laforgue et al., 2020]

Learning from MoM's principle

Direct applications (1/2)

- **Robust bandits** [Bubeck et al., 2013]: $B_{k,s,t} = \hat{\mu}_{k,s,t} + \sqrt{\frac{2vc \log t}{s}}$.

Direct applications (1/2)

- **Robust bandits** [Bubeck et al., 2013]: $B_{k,s,t} = \hat{\mu}_{k,s,t} + \sqrt{\frac{2vc \log t}{s}}$.

- **Robust mean embedding** [Lerasle et al., 2019]:

$$\mu_{\mathbb{P}} = \int k(\cdot, x) d\mathbb{P}(x) \qquad = \operatorname*{argmin}_{h \in \mathcal{H}_k} \int \|h - k(\cdot, x)\|^2_{\mathcal{H}_k} d\mathbb{P}(x),$$

$$\hat{\mu}_{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, x_i), \qquad \bar{\mu}_{\mathbb{P}} = \operatorname*{argmin}_{h \in \mathcal{H}_k} \sup_{h' \in \mathcal{H}_k} \text{MoM}\left\{ \|h - k(\cdot, x)\|^2_{\mathcal{H}_k} - \|h' - k(\cdot, x)\|^2_{\mathcal{H}_k} \right\}.$$

$$\text{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k} = \sup_{h \in \mathcal{B}_k} \langle h, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k} = \sup_{h \in \mathcal{B}_k} \int \langle h, k(\cdot, x) - k(\cdot, y) \rangle_{\mathcal{H}_k} d\mathbb{P}(x) d\mathbb{Q}(y),$$

$$\widehat{\text{MMD}}(\mathbb{P}, \mathbb{Q}) = \sup_{h \in \mathcal{B}_k} \langle h, \hat{\mu}_{\mathbb{P}} - \hat{\mu}_{\mathbb{Q}} \rangle_{\mathcal{H}_k} = \sup_{h \in \mathcal{B}_k} \frac{1}{n} \sum_{i=1}^{n} h(x_i) - h(y_i),$$

$$\overline{\text{MMD}}(\mathbb{P}, \mathbb{Q}) = \sup_{h \in \mathcal{B}_k} \text{MoM}\left\{ \langle h, k(\cdot, x) - k(\cdot, y) \rangle_{\mathcal{H}_k} \right\} = \sup_{h \in \mathcal{B}_k} \text{med}\left(\langle h, \hat{\mu}_{\mathbb{P},k} - \hat{\mu}_{\mathbb{Q},k} \rangle_{\mathcal{H}_k},\ k \le K \right).$$

Direct Applications (2/2)

- **Robust optimal transport** [Staerman et al., 2020]:

$$\mathcal{W}_1(\mu,\nu) = \sup_{\phi \in B_L} \mathbb{E}_\mu[\phi(X)] - \mathbb{E}_\nu[\phi(Y)] = \sup_{\phi \in B_L} \mathbb{E}_{\mu \otimes \nu}[\phi(X) - \phi(Y)]$$

$$\widehat{\mathcal{W}}(\mu,\nu) = \sup_{\phi \in B_L} \frac{1}{n}\sum_{i=1}^{n} \phi(X_i) - \frac{1}{m}\sum_{j=1}^{m} \phi(Y_j)$$

$$\overline{\mathcal{W}}_{\mathrm{MoM}}(\mu,\nu) = \sup_{\phi \in B_L} \mathrm{MoM}_{\mathcal{S}_X}[\phi(X)] - \mathrm{MoM}_{\mathcal{S}_Y}[\phi(Y)]$$

$$\overline{\mathcal{W}}_{\mathrm{MoU}}(\mu,\nu) = \sup_{\phi \in B_L} \mathrm{MoU}_{\mathcal{S}_{XY}}[\phi(X) - \phi(Y)]$$

$$\mathbb{E}\left|\overline{\mathcal{W}}_{\mathrm{MoM}}(\mu,\nu) - \mathcal{W}(\mu,\nu)\right| \leq \frac{C}{n^{1/(d+2)}}$$

MoM minimization [Lecué et al., 2018]

$$\textbf{ERM:} \quad \min_{h \text{ measurable}} \mathbb{E}\left[\ell(h, Z)\right] \quad \rightarrow \quad \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h, z_i)$$

$$\textbf{MoM min:} \quad \min_{h \in \mathcal{H}} \text{MoM}\Big\{\ell(h, Z)\Big\} = \text{med}\left(\frac{1}{|B_k|} \sum_{i \in B_k} \ell(h, z_i),\ k \leq K\right)$$

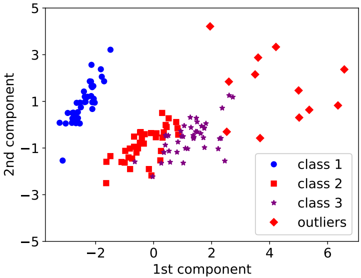
How to adapt Gradient Descent?

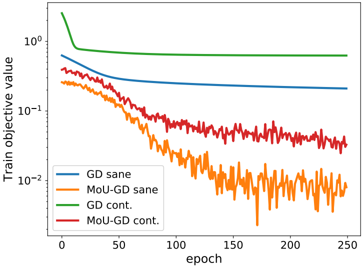
- compute empirical risk on each block
- perform batch GD on block with median risk
- needs for reshuffling the partition at each iteration

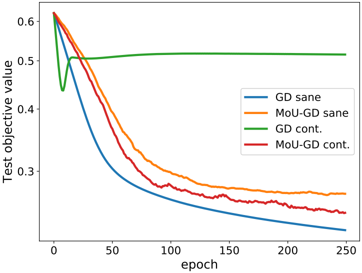
MoU Gradient Descent for metric learning

We want to minimize for $M \in S_q^+(\mathbb{R})$:

$$\frac{2}{n(n-1)} \sum_{i<j} \max\Big(0, 1+y_{ij}(d_M^2(x_i, x_j)-2)\Big)$$







The tournament procedure [Lugosi and Mendelson, 2016]

We want $g^* \in \operatorname*{argmin}_{g \in \mathcal{G}} \mathcal{R}(g) = \mathbb{E}[(g(X) - Y)^2]$. For any pair $(g, g') \in \mathcal{G}^2$:

1) Compute the MoM estimate of $\|g - g'\|_{L_1}$

$$\Phi_{\mathcal{S}}(g, g') = \text{median}\left(\hat{\mathbb{E}}_1 |g - g'|, \ldots, \hat{\mathbb{E}}_K |g - g'|\right).$$

2) If it is *large enough*, compute the *match*

$$\Psi_{\mathcal{S}'}(g, g') = \text{median}\Big(\hat{\mathbb{E}}_1[(g(X) - Y)^2 - (g'(X) - Y)^2], \ldots, \hat{\mathbb{E}}_K[(g(X) - Y)^2 - (g'(X) - Y)^2]\Big).$$

3) $\hat{g}$ winning all its matches verifies w.p.a.l. $1 - \exp(c_0 n \min\{1, r^2\})$

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g^*) \leq cr.$$

Can be extended to pairwise learning thanks to MoU

References (applications)

- Robust bandit strategies [Bubeck et al., 2013]
- Robust mean embedding [Lerasle et al., 2019]
- Robust optimal transport [Staerman et al., 2020]

- Le Cam's [Lecué and Lerasle, 2019], tournament [Lugosi and Mendelson, 2019]
- MoM minimization [Lecué et al., 2018], MoM min-max [Lecué and Lerasle, 2017]

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \left\{ \sup_{h' \in \mathcal{H}} \mathbb{E}\left[\ell(h, Z) - \ell(h', Z) \right] \right\}$$

Take home messages

Conclusion

- **MoM has sub-Gaussian behavior with finite variance only**
- **MoM is robust to outliers**
- **MoM can replace any empirical mean in algorithms**
(bandits, MMD, OT)
- **MoM provides alternatives to ERM**
(MoM-minimization, MoM-GD, MoM tournament, MoM-minimax)

References I

Alon, N., Matias, Y., and Szegedy, M. (1999).
The space complexity of approximating the frequency moments.
Journal of Computer and system sciences, 58(1):137–147.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013).
Bandits with heavy tail.
IEEE Transactions on Information Theory, 59(11):7711–7717.

Catoni, O. (2012).
Challenging the empirical mean and empirical variance: a deviation study.
In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185. Institut Henri Poincaré.

References II

Depersin, J. and Lecué, G. (2019).
Robust subgaussian estimation of a mean vector in nearly linear time.
arXiv preprint arXiv:1906.03058.

Devroye, L., Lerasle, M., Lugosi, G., Oliveira, R. I., et al. (2016).
Sub-gaussian mean estimators.
The Annals of Statistics, 44(6):2695–2725.

Hsu, D. and Sabato, S. (2016).
Loss minimization and parameter estimation with heavy tails.
The Journal of Machine Learning Research, 17(1):543–582.

Jerrum, M. R., Valiant, L. G., and Vazirani, V. V. (1986).
Random generation of combinatorial structures from a uniform distribution.
Theoretical Computer Science, 43:169–188.

References III

Joly, E. and Lugosi, G. (2016).
Robust estimation of u-statistics.
Stochastic Processes and their Applications, 126(12):3760–3773.

Laforgue, P., Clémençon, S., and Bertail, P. (2019).
On medians of (randomized) pairwise means.
In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*.

Laforgue, P., Staerman, G., and Clémençon, S. (2020).
How robust is the median-of-means? concentration bounds in presence of outliers.
arXiv preprint arXiv:2006.05240.

References IV

Lecué, G. and Lerasle, M. (2017).
Robust machine learning by median-of-means: theory and practice.
arXiv preprint arXiv:1711.10306.

Lecué, G. and Lerasle, M. (2019).
Learning from mom's principles: Le cam's approach.
Stochastic Processes and their applications, 129(11):4385–4410.

Lecué, G., Lerasle, M., and Mathieu, T. (2018).
Robust classification via mom minimization.
arXiv preprint arXiv:1808.03106.

References V

Lerasle, M., Szabo, Z., Mathieu, T., and Lecué, G. (2019).
Monk – outlier-robust mean embedding estimation by median-of-means.
In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*.

Lugosi, G. and Mendelson, S. (2016).
Risk minimization by median-of-means tournaments.
arXiv preprint arXiv:1608.00757.

Lugosi, G. and Mendelson, S. (2017).
Sub-gaussian estimators of the mean of a random vector.
arXiv preprint arXiv:1702.00482.

Lugosi, G. and Mendelson, S. (2019).
Risk minimization by median-of-means tournaments.
Journal of the European Mathematical Society.

References VI

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

Nemirovsky, A. S. and Yudin, D. B. (1983).
***Problem Complexity and Method Efficiency in Optimization*.**
John Wiley & Sons Ltd.

Staerman, G., Laforgue, P., Mozharovskyi, P., and d'Alché Buc, F. (2020).
When ot meets mom: Robust estimation of wasserstein distance.
arXiv preprint arXiv:2006.10325.