# Statistical error bounds for weighted mean and median, with application to robust aggregation of cryptocurrency data

Michaël Allouche, Mnacho Echenim, Emmanuel Gobet, Anne-Claire Maurice

# Statistical error bounds for weighted mean and median, with application to robust aggregation of cryptocurrency data

Michaël ALLOUCHE* Mnacho ECHENIM† Emmanuel GOBET‡
Anne-Claire MAURICE§

December 17, 2024

**Abstract**

We study price aggregation methodologies applied to crypto-currency prices with quotations fragmented on different platforms. An intrinsic difficulty is that the price returns and volumes are heavy-tailed, with many outliers, making averaging and aggregation challenging. While conventional methods rely on Volume-Weighted Average Prices (called VWAPs), or Volume-Weighted Median prices (called VWMs), we develop a new Robust Weighted Median (RWM) estimator that is robust to price and volume outliers. Our study is based on new probabilistic concentration inequalities for weighted means and weighted quantiles under different tail assumptions (heavy tails, sub-gamma tails, sub-Gaussian tails). This justifies that fluctuations of VWAP and VWM are statistically important given the heavy-tailed properties of volumes and/or prices. We show that our RWM estimator overcomes this problem and also satisfies all the desirable properties of a price aggregator. We illustrate the behavior of RWM on synthetic data (within a parametric model close to real data): our estimator achieves a statistical accuracy twice as good as its competitors, and also allows to recover realized volatilities in a very accurate way. Tests on real data are also performed and confirm the good behavior of the estimator on various use cases.

KEYWORDS: robust aggregation; weighted mean and quantile estimation; heavy tails; concentration inequalities; outliers

MSC2020: 62E17; 62G15; 62G35; 62P05
JEL: C13; C18; G15

## 1 Introduction

**Context.** Digital finance has experienced an unprecedented boom in the last decade, with the emergence of digital assets and smart contracts that can be traded in a decentralized (using a Blockchain protocol) or centralized manner, see [Arslanian, 2022] for an overview. As opposed to traditional finance

*Kaiko - Quantitative Data. 2 rue de Choiseul 75002 Paris, France. Email: MICHAEL.ALLOUCHE@KAIKO.COM

†Laboratoire d'Informatique de Grenoble (LIG), CNRS, Grenoble INP, UGA. 700 avenue Centrale, Domaine Universitaire, 38401 Saint Martin d'Hères, France. Email: MNACHO.ECHENIM@UNIV-GRENOBLE-ALPES.FR

‡Centre de Mathématiques Appliquées (CMAP), CNRS, Ecole Polytechnique, Institut Polytechnique de Paris. Route de Saclay, 91128 Palaiseau Cedex, France. Email: EMMANUEL.GOBET@POLYTECHNIQUE.EDU. Corresponding author.

§Kaiko - Quantitative Data. 2 rue de Choiseul 75002 Paris, France. Email: ANNE-CLAIRE.MAURICE@KAIKO.COM

where each given asset is quoted and traded on a centralized stock exchange, information of traded prices/volumes[1] and quotes/full order books is scattered in the case of digital finance. This makes the task of "market consensus" price calculation difficult. Reference prices are essential for the settlement of financial derivatives based on digital assets, or in the net asset value calculation of exchange-traded funds.
Our work tackles the problem of information aggregation to obtain reference prices based on past traded prices/volumes, by developing a new methodology (called `RWM` for Robust Weighted Median) which we prove to be more robust and efficient than existing ones. Theoretical confidence bounds are derived to assess the gain in statistical accuracy, and these results are completed by numerous experiments illustrating the outperformance of RWM.

**State of the art for the methodologies.** Consider the situation where the traded price and volume data $(P_i, V_i)_{i=1}^n$ of a digital asset is collected from several exchanges on a short time window (typically a minute), and for which a "market consensus price" has to be provided at that time. Following the arguments of [Paine and Knottenbelt, 2016], an aggregation rule should enjoy various characteristics such as: Relevance, Timeliness, Manipulation Resistance, Martingale Property, Verifiability, Replicability, Stability, Parsimony. Our `RWM` methodology enjoys all these properties as discussed in Section 2.3.
A first methodology consists in simply averaging prices from different exchanges, yielding the aggregated price

$$\frac{1}{n}\sum_{i=1}^{n} P_i.$$

It is computed this way by, among others, [Vinter, 2021] to obtain prices at the minute scale, while an extra average in time is added to obtain hourly or daily prices (Time Weighted Average Prices, TWAPs in short). Although simple, this volume-independent aggregation is questionable: indeed, it is sensitive to market manipulations that could be performed by trading automata capable of executing a large number of orders with tiny volumes. This explains why aggregation rules that compute weighted averages with higher weights on prices with higher volumes have been investigated. A first well-known methodology is based on the Volume-Weighted Average Price (VWAP), which consists in averaging the traded prices with a weight equal to the traded volume:

$$\widehat{\texttt{VWAP}}_n := \frac{\sum_{i=1}^n V_i P_i}{\sum_{i=1}^n V_i}. \tag{1.1}$$

This methodology is used by, among others, [Nasdaq, 2022] to obtain the Nasdaq Crypto Index (NCI) and by FTSE Russell [FTSE Digital Asset Research, 2022] to compute Digital Asset Reference Prices. However, it is well-known that empirical means are very sensitive to extreme values [Ronchetti and Huber, 2009], and thus, the above approach, as it is, is not robust to outliers. Therefore, [Nasdaq, 2022] and [FTSE Digital Asset Research, 2022] apply an extra level of outlier management: in [Nasdaq, 2022], the weights in the VWAP are modified to account for abnormal prices, abnormal volatilities or abnormal volumes, with several hyper-parameters to set that do not make the method parsimonious ; in [FTSE Digital Asset Research, 2022], data is discarded from the raw data set based on exchange-level and

[1] *i.e.* number of contracts

trade-level outlier detection, with thresholds that once again depend on hyperparameters to tune.
Another well-known methodology is based on the Volume-Weighted Median price (VWM), which is computed as a usual median but with each price data weighted by its volume:

$$\widehat{\texttt{VWM}}_n := \inf\left\{p : \frac{\sum_{i=1}^n V_i \mathbb{1}_{\{P_i \le p\}}}{\sum_{i=1}^n V_i} \ge \frac{1}{2}\right\}. \tag{1.2}$$

To the best of our knowledge, this approach was publicly used for the first time by the Federal Reserve Bank of New York [Federal Reserve Bank of New York, 2015] on interest rates, suggesting to use VWM in order to be more robust to outliers; it is also at work at the Chicago Mercantile Exchange [Paine and Knottenbelt, 2016] [CF Benchmarks, 2022b] [CF Benchmarks, 2022a] and Bloomberg [Gallagher, 2018]. An example is the Bitcoin Reference Rate (BRR) computed every hour by CME: the data set is made of 12 partitions of data over 5-minute timespans, on each partition a VWM is computed, and then the 12 results are averaged. Interestingly, this approach has only a few hyper-parameters to tune.
In Subsection 2.2 which is dedicated to the theoretical statistical analysis of estimators including $\widehat{\texttt{VWAP}}_n$ and $\widehat{\texttt{VWM}}_n$, we will see that both methodologies suffer from a large statistical uncertainty, intrinsically due to the heavy tail characteristics of the data under consideration. This justifies the development of a new price aggregation methodology called Robust Weighted Median (RWM), with high accuracy properties, see Subsection 2.3.

**State of the art for statistical analysis.** Beyond the useful applications in digital finance, our work brings novelties on the theoretical side for the study of weighted empirical means and medians, in particular the convergence in distribution of their re-normalized error (Central Limit Theorem, as $n \to +\infty$) and non-asymptotics error bounds (when $n$ is fixed). Let us discuss this by putting into perspective the existing results from the literature in statistics.
There is a long history on the measure of the centrality of a distribution, and on the debate between mean and median. According to [Bakker and Gravemeijer, 2006, Section 3], around 1755, Boscovich was the first to employ the concept of median in the context of a regression method with L1 norm. Legendre (1753-1833) and Laplace (1749-1825) used the term *milieu de probabilité* to refer to the middle of a distribution. Then, [Farebrother, 2001] reports that in 1883, Galton was the first to employ a median value instead of a mean value, he justified his preference by analogy with a simple majority voting procedure. At the same time, Edgeworth also used the median rather than the mean, as a measure of distribution centrality. During the XXth century, the mean became more important, probably because of its popularity in the society. However, in the seventies, robust statistics were developed to better account for extreme values and outliers in statistical procedures, see the book by Huber and Ronchetti [Ronchetti and Huber, 2009]. Nowadays, statistical procedures accounting for robustness are a must-have.

Without weighting, the theoretical study of empirical mean and empirical median is a standard topics: see [Boucheron et al., 2013] about concentration inequalities around the expectation of the empirical mean computed over $n$ data points (under various distribution assumptions that will be detailed later), or [Reiss, 1989, Chapter 3]-[Boucheron and Thomas, 2012] about concentration inequalities of the empirical median and quantiles. For refined (non-weighted) quantile estimations, see [Hyndman and Fan, 1996].

We should mention that in order to compute an expectation, other robust mean estimation procedures could be used, together with nice concentration inequalities: such methods include median-of-means [Devroye et al., 2016], trimmed means [Lugosi and Mendelson, 2019], or the Catoni estimator [Catoni, 2012]. These estimators have excellent concentration properties even if the data is heavy-tailed; on the other hand, to achieve this, each estimator has to be defined according to a confidence level on which its accuracy is to be guaranteed, without any hint to how the estimator behaves outside this probable event. In addition, the amount of data $n$ required to ensure the validity of the above accuracy confidence bounds is usually well over several hundreds, which is not compatible with the application on cryptocurrency market (with data aggregation at the minute time scale). These computational considerations for Monte Carlo methods are discussed in [Gobet et al., 2022].

When weighting is incorporated into the estimators as it is the case in (1.1)-(1.2), we can refer to [Glynn et al., 1996], which studies the asymptotic normality of empirical quantiles computed using importance sampling methods (a case where $\mathbb{E}[V] = 1$). For a more general case and non-asymptotic results, to the best of our knowledge such a theoretical study for assessing estimator errors has not yet been conducted. Additionally, the theoretical results on the weighted estimators $\widehat{\texttt{VWAP}}_n$ and $\widehat{\texttt{VWM}}_n$ developed in Subsection 2.2 will show that their statistical fluctuations are directly impacted by the tail-heaviness of the distributions of $V \cdot P$ and $V$ respectively. Yet the latter appear to be heavy-tailed in the cryptocurrency market, see Section 5 for details, *i.e.* their quantile functions diverge in such a way that the underlying probability distributions may not have a finite variance. In other words, it is hopeless to expect good statistical properties of $\widehat{\texttt{VWAP}}_n$ and $\widehat{\texttt{VWM}}_n$ on crypto asset data, although the statistical behavior of $\widehat{\texttt{VWM}}_n$ is presumably slightly better than that of $\widehat{\texttt{VWAP}}_n$.

**Contributions.** Our contributions are threefold.

First, we study in a general setting of random weights denoted by $W$ (instead of weights $V$) the properties of weighted averages, weighted medians (as requested by applications) and we also include, for the sake of generality, weighted quantiles. This is achieved on various tail assumptions on $W$ and $W \cdot P$; namely we consider heavy tails, sub-gamma tails, and sub-Gaussian tails. Results on estimation errors include asymptotic normality, as n$\rightarrow +\infty$ (Central Limit Theorem in Theorem 2.2), and non-asymptotic confidence bounds (concentration inequalities in Theorems 2.3, 2.4, 2.5), which fit well in the setting of a small number $n$ of points, similar to the one we can meet in practice (typically $n = 100$ data points over a 1-minute time window). These main theoretical results clearly show the role of the assumptions on the distribution tails, regarding the accuracy of the confidence interval (CI in short); Table 1 summarizes the behavior of the CI width under the various tail assumptions.

Second, in view of previous properties, we design a new aggregated estimator called RWM which enjoys good statistical properties: it fulfills all of the desirable characteristics mentioned in the introduction. In a nutshell, this estimator is based on Weighted Medians but with a weight equal to the log-volume ($W = \log(1 + V)$ up to a scaling defined later). Observe that the estimator RWM behaves like the VWM for small volumes (since $\log(1 + V) \sim V$), but in addition, it has extra robustness properties with respect to extreme volumes and prices. Actually, since $V$ is heavy-tailed,

$W$ has sub-gamma tails and thus, the final estimator RWM exhibits qualitatively smaller statistical fluctuations compared to VWM and VWAP, see Table 1.

Last, we illustrate the aggregation procedure on synthetic and real data (see Sections 4 and 5 respectively), and compare it with VWAP and VWM. In 90 to 95% of the cases, all estimators behave similarly but in the presence of strong outliers, RWM behaves remarkably better, see Figure 1. We also highlight that RWM is the only estimator that permits to accurately estimate the realized volatility (RV), despite the possible large number of outliers in prices and volumes in the data time-series.

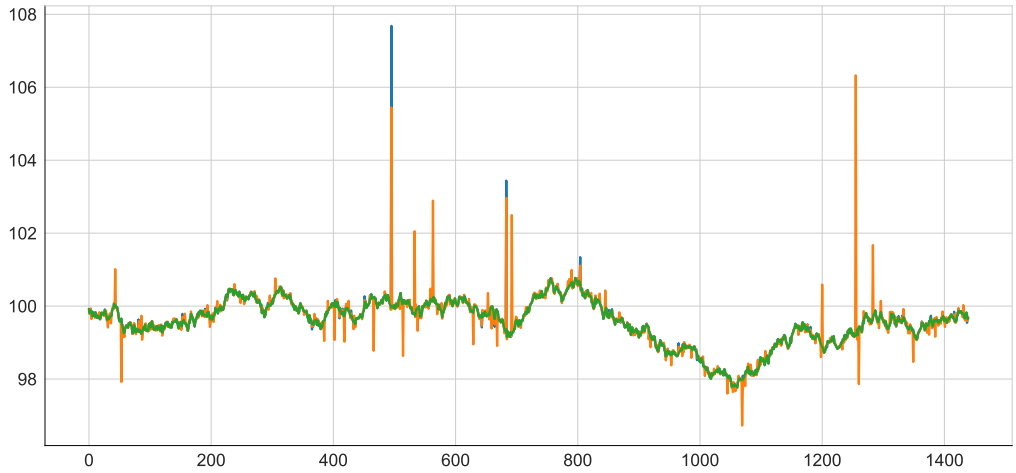



Figure 1: Aggregated volume weighted price estimation on simulated data surrounding an efficient price generated every minute during one day using $\widehat{\mathtt{VWM}}_n$ (blue), $\widehat{\mathtt{VWAP}}_n$ (orange) and $\widehat{\mathtt{RWM}}_n$ (green). See Section 4 for details.

**Organization of the paper.** This paper is organized as follows. The asymptotic and non-asymptotic statistical fluctuations of the weighted average and quantile estimators are studied in Section 2.2. The estimator RWM is presented in Section 2.3 and the real-data are presented in Section 3. The performance of our estimator is illustrated on simulated data (Section 4) and on real cryptocurrency market data (Section 5). Technical proofs are postponed to the Appendix.

## 2 Main theoretical results

This section is mainly devoted to the theoretical study of statistical fluctuations of empirical weighted means and medians, for general weigths $W$ (not necessarily equal to the volume $V$ as in (1.1) and (1.2)) since below, we will choose $W$ as a specific function of $V$. In addition, for the sake of generality of our study, we also investigate the case of weighted quantiles. The purpose of this section is to provide a theoretical foundation for assessing statistical fluctuations of aggregate prices in order to propose an accurate and robust solution.

First in Subsection 2.1, we define our stochastic framework. Then in Subsection 2.2 we state and prove convergence results of weighted means/medians/quantiles: asymptotic fluctuations as the number of data $n \to +\infty$ (Central Limit Theorem) and non-asymptotic deviation arguments. Refined estimators will be defined and studied in the Subsection 2.3.

## 2.1 Probabilistic setting

To prepare the estimators study, we need to define a few probabilistic notions. Let us consider a standard probability space $(\Omega, \mathscr{F}, \mathbb{P})$ on which we are given two positive random variables $P$ and $W$, which stand respectively for the price of a given asset and the weight based on which we want to compute weighted estimators. We assume the integrability conditions $\mathbb{E}[W \cdot P] < +\infty$, $\mathbb{E}[W] < +\infty$ and to avoid degenerate situations, we also assume $\mathbb{P}(W > 0) = 1$ and that $P$ is not constant.

**Model-based definitions.** Let us define the cumulative distribution function (c.d.f.) of the weighted price as

$$F_{\mathtt{W}}(x) := \frac{\mathbb{E}\left[W \cdot \mathbb{1}_{\{P \le x\}}\right]}{\mathbb{E}[W]}, \qquad x \ge 0. \tag{2.1}$$

As we will focus on specific weights $W$, we index our quantities by $W$ to highlight this dependence. Although the notation $F_{\mathtt{W}}$ may suggest that this quantity refers to the c.d.f. of $W$, we insist that it rather refers to the distribution of $P$ with weights $W$.

**Remark 2.1.** The above is analogue to a change of measure (with weights proportional to $W$) also known as *importance sampling* [Asmussen and Glynn, 2007, Section V.1 ]. The difference with importance sampling is that the expectation of $W$ is not necessarily equal to 1. The above reads more as a Feynman-Kac formula occurring in mean field simulation [Del Moral, 2013]. Despite the similarity, the literature on these topics is not very helpful to obtain the following results.

The expectation (resp. probability) w.r.t. the probability measure induced by $F_{\mathtt{W}}$ is denoted by $\mathbb{E}_{\mathtt{W}}[.]$ (resp. $\mathbb{P}_{\mathtt{W}}(.)$). For any measurable function $g : \mathbb{R}^+ \mapsto \mathbb{R}^+$, we have

$$\mathbb{E}_{\mathtt{W}}\left[g(P)\right] = \frac{\mathbb{E}\left[W \cdot g(P)\right]}{\mathbb{E}[W]}. \tag{2.2}$$

The $W$-weighted average price (WAP) corresponds to the choice $g(x) = x$ and writes

$$\mathtt{WAP} := \frac{\mathbb{E}[W \cdot P]}{\mathbb{E}[W]}. \tag{2.3}$$

It coincides with the VWAP when $W = V$. Observe that the $\mathtt{WAP}$ is also the (unique) minimizer of the quadratic problem

$$\mathtt{WAP} := \arg\inf_{p} \mathbb{E}_{\mathtt{W}}\left[|P - p|^2\right].$$

The $W$-weighted median is defined by considering a similar minimization problem, but with the abso-

lute value loss (instead of the square):

$$\mathtt{WM} := \arg\inf_p \mathbb{E}_{\mathtt{W}}\left[|P-p|\right]. \tag{2.4}$$

There may be several minimizers, all characterized by the equations

$$\frac{1}{2} \geq \mathbb{P}_{\mathtt{W}}\left(P \geq \mathtt{WM}\right) = \frac{\mathbb{E}\left[W\mathbb{1}_{\{P\geq \mathtt{WM})\}}\right]}{\mathbb{E}[W]}, \qquad \frac{1}{2} \leq \mathbb{P}_{\mathtt{W}}\left(P \leq \mathtt{WM}\right) = \frac{\mathbb{E}\left[W\mathbb{1}_{\{P\leq \mathtt{WM})\}}\right]}{\mathbb{E}[W]}.$$

For the discussion of this Subsection and Subsection 2.2, we take the convention of considering the smallest minimizer. Last, we define the $W$-weighted quantile at level $\alpha \in (0,1)$ by

$$q_{\mathtt{W}}(\alpha) := \inf\left\{p : \frac{\mathbb{E}\left[W\mathbb{1}_{\{P\leq p\}}\right]}{\mathbb{E}[W]} \geq \alpha\right\}.$$

Observe that the weighted median, as the smallest minimizer of (2.4), is obtained via $q_{\mathtt{W}}(\frac{1}{2})$.

**Data-based definitions.** We now turn to the empirical versions of the previous definitions. Given $n$ data points $(P_i, W_i)_{i=1}^n$ sampled independently and supposedly from the distribution of $(P,W)$, the empirical counterpart of $F_{\mathtt{W}}$ is

$$\widehat{F_{\mathtt{W},\mathtt{n}}}(x) := \frac{\sum_{i=1}^n W_i \mathbb{1}_{\{P_i\leq x\}}}{\sum_{i=1}^n W_i}. \tag{2.5}$$

This is related to the empirical measure $\sum_{i=1}^n W_i \delta_{P_i}(\mathrm{d}x)/\sum_{i=1}^n W_i$. Applying the previous definitions with the above empirical distribution, we obtain

$$\widehat{\mathtt{WAP}}_n := \frac{\sum_{i=1}^n W_i P_i}{\sum_{i=1}^n W_i}, \tag{2.6}$$

$$\widehat{q_{\mathtt{W},n}}(\alpha) := \inf\left\{p : \frac{\sum_{i=1}^n W_i \mathbb{1}_{\{P_i\leq p\}}}{\sum_{i=1}^n W_i} \geq \alpha\right\}, \tag{2.7}$$

$$\widehat{\mathtt{WM}}_n := \widehat{q_{\mathtt{W},n}}\left(\frac{1}{2}\right). \tag{2.8}$$

**Comments about the i.i.d. assumption on $(P_i, W_i)_{i=1}^n$.** From a practical point of view, the *stationarity* assumption can be easily accepted since in our application, we aggregate information over short time-intervals (typically a second or a minute). The assumption of strict independence seems to be harder to argue: nevertheless, the crypto market is fragmented and so requires to aggregate data coming from dozens of platforms (Centralized Exchanges like *Binance*, *Coinbase*, or from Decentralized Exchanges like *Uniswap* Automated Market Makers, or *Curve* AMMs) that can be roughly considered as independent actors. Anyway, even in the case of non-independent data but satisfying mixing assumptions, it is known that concentration-of-measures inequalities can be transferred (with minor adjustments on the final results) from the i.i.d. case to the non independent i.d. case using decoupling results (like Berbee's lemma), see [Lerasle, 2009] in the context of density estimation, see [Ren and Mojirsheibani, 2010, Kuznetzov and Mohri, 2017] for regression problems under various mixing assumptions: we guess that our subsequent results are valid also in the non i.i.d. case under mixing conditions, upon small adjustments on the error bounds.

We leave the proof of these technicalities to the reader, and refer for instance to the previous references [Lerasle, 2009, Ren and Mojirsheibani, 2010, Kuznetzov and Mohri, 2017] for the strategy of analysis.

## 2.2 Statistical fluctuations, error bounds

The convergence study of $\widehat{\texttt{WAP}}_n, \widehat{q_{\texttt{W},n}}(\alpha)$ to $\texttt{WAP}, q_{\texttt{W}}(\alpha)$ will be conducted under various assumptions on the tails of the distributions of $W$, $W \cdot P$ and on the c.d.f. $F_{\texttt{W}}$. We recall that $\alpha \in (0,1)$. Below, $X$ is a generic scalar non-negative random variable.

**(H0):** The c.d.f. $F_{\texttt{W}}$ in (2.1) has a density $f_{\texttt{W}}$:

$$F_{\texttt{W}}(x) = \int_0^x f_{\texttt{W}}(x')\,\mathrm{d}x'.$$

**($\mathbf{H}_{\mathbf{X}}^{\kappa}$):** $X$ has a finite moment of order $\kappa$: $\mathbb{E}[X^{\kappa}] < +\infty$.

**($\mathbf{H}_{\mathbf{X}}^{\Gamma}$):** $X$ has a sub-gamma distribution, i.e. $\mathbb{E}\left[e^{cX}\right] < +\infty$ for some $c > 0$.

**($\mathbf{H}_{\mathbf{X}}^{\mathbf{G}}$):** $X$ has a sub-Gaussian distribution, i.e. $\mathbb{E}\left[e^{cX^2}\right] < +\infty$ for some $c > 0$.

In ($\mathbf{H}_{\mathbf{X}}^{\kappa}$), the parameter $\kappa$ will be taken equal to 2 or greater than 2, according to theorem statements. These assumptions on distribution tails range from the mildest to the strongest, in order to clearly distinguish the behavior of the fluctuations depending on the tail hypotheses. The 3 types of distribution tails are quite standard, but they could be refined if necessary: we leave to the reader the generalization of the following results to the $\eta$-exponential tails ($\mathbb{E}\left[e^{c|X|^{\eta}}\right] < +\infty$ for some $c > 0$, see [Chamakh et al., 2020]) or to the $\eta$-heavy tails ($\mathbb{E}\left[e^{\log(|X/c|+1)^{\eta}}\right] < +\infty$ for some $c > 0$, see [Chamakh et al., 2022]).
For the current application to price aggregation and in view of the characteristics of cryptomarket data (see Section 5), we will see that

- ($\mathbf{H}_{\mathbf{X}}^{\kappa}$) with $X = V$ and $X = V \cdot P$ often hold with $\kappa > 2$, and allows to analyse confidence bounds for VWAP and VWM (see Theorem 2.3),
- ($\mathbf{H}_{\mathbf{X}}^{\Gamma}$) with $X = \log(1+V)$ holds, and allows to analyse our new RWM estimator (see Theorem 2.4).

However, to make our results more widely applicable (beyond the current context), we add to the analysis the convergence properties under the stronger assumption of sub-Gaussian tails (Theorem 2.5).
We first consider the asymptotic normality of estimators (Central Limit Theorem, CLT in short), as $n \to +\infty$. We denote the weak convergence by $\Rightarrow$.

**Theorem 2.2** (Asymptotic fluctuations on WM and WAP)**.** *Assume (**H0**) with a continuous and non-vanishing density $f_{\texttt{W}}$ at $q_{\texttt{W}}(\alpha)$, and assume ($\mathbf{H}_{\mathbf{X}}^{\kappa}$) for $X = W$ and $\kappa = 2$. Then*

$$\sqrt{n}\left(\widehat{q_{\texttt{W},n}}(\alpha) - q_{\texttt{W}}(\alpha)\right) \Rightarrow \mathcal{N}\left(0, \frac{\mathbb{E}\left[W^2\left(\alpha - \mathbb{1}_{\{P \le q(\alpha)\}}\right)^2\right]}{\left(\mathbb{E}[W]\, f_{\texttt{W}}\left(q_{\texttt{W}}(\alpha)\right)\right)^2}\right). \tag{2.9}$$

*Assume ($\mathbf{H}_{\mathbf{X}}^{\kappa}$) for $X = W$ and $X = W \cdot P$ and $\kappa = 2$, then*

$$\sqrt{n}\left(\widehat{\texttt{WAP}}_n - \texttt{WAP}\right) \Rightarrow \mathcal{N}\left(0, \frac{\mathbb{E}\left[W^2(\texttt{WAP} - P)^2\right]}{(\mathbb{E}[W])^2}\right). \tag{2.10}$$

As for usual CLTs, the random variables under consideration are required only to have second finite moments ($(\mathbf{H}_{\mathbf{X}}^{\kappa})$ with $\kappa = 2$). However, for finite size samples, the estimators may be far from being Gaussian because of the possible heavy tails: we recall that in our applications, the sample size $n$ is around 100, thus assessing error bounds in a non-asymptotic way is really important. This is the purpose of the next three results. Namely, we are interested in deriving sharp bounds for left and right deviations: let $\alpha \in (0,1)$ and $x > 0$, define

$$\mathscr{Q}_n^{>}(\alpha,x) := \mathbb{P}\left(\widehat{q_{\mathtt{W},n}}(\alpha) - q_{\mathtt{W}}(\alpha) > \frac{x}{\sqrt{n}}\right), \qquad \mathscr{Q}_n^{\leq}(\alpha,x) := \mathbb{P}\left(\widehat{q_{\mathtt{W},n}}(\alpha) - q_{\mathtt{W}}(\alpha) \leq -\frac{x}{\sqrt{n}}\right), \tag{2.11}$$

$$\mathscr{W}_n^{\geq}(x) := \mathbb{P}\left(\widehat{\mathtt{WAP}}_n - \mathtt{WAP} \geq \frac{x}{\sqrt{n}}\right), \qquad \mathscr{W}_n^{\leq}(x) := \mathbb{P}\left(\widehat{\mathtt{WAP}}_n - \mathtt{WAP} \leq -\frac{x}{\sqrt{n}}\right). \tag{2.12}$$

**Theorem 2.3** (Deviation bounds under Heavy Tail assumptions)**.** *Assume* $(\mathbf{H0})$ *and* $(\mathbf{H}_{\mathbf{X}}^{\kappa})$ *for* $X = W$ *and some* $\kappa > 2$*. Then*

$$\mathscr{Q}_n^{>}(\alpha,x) \leq \left(\frac{\kappa+2}{\kappa}\right)^{\kappa} \frac{(2[\alpha \vee (1-\alpha)])^{\kappa}\mathbb{E}[W^{\kappa}]}{n^{\frac{\kappa}{2}-1}\left[x\mathbb{E}[W] f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha), \frac{x}{\sqrt{n}}])\right]^{\kappa}} + \exp\left(-\frac{2\left[x\mathbb{E}[W] f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha), \frac{x}{\sqrt{n}}])\right]^2}{(\kappa+2)^2 e^{\kappa}[\alpha \vee (1-\alpha)]^2 \mathbb{E}\left[W^2\right]}\right), \tag{2.13}$$

$$\mathscr{Q}_n^{\leq}(\alpha,x) \leq \left(\frac{\kappa+2}{\kappa}\right)^{\kappa} \frac{(2[\alpha \vee (1-\alpha)])^{\kappa}\mathbb{E}[W^{\kappa}]}{n^{\frac{\kappa}{2}-1}\left[x\mathbb{E}[W] f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha), \frac{-x}{\sqrt{n}}])\right]^{\kappa}} + \exp\left(-\frac{2\left[x\mathbb{E}[W] f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha), \frac{-x}{\sqrt{n}}])\right]^2}{(\kappa+2)^2 e^{\kappa}[\alpha \vee (1-\alpha)]^2 \mathbb{E}\left[W^2\right]}\right). \tag{2.14}$$

*Assume* $(\mathbf{H}_{\mathbf{X}}^{\kappa})$ *for* $X = W$ *and* $X = W \cdot P$*, and for some* $\kappa > 2$*. Then*

$$\mathscr{W}_n^{\geq}(x) \leq \left(\frac{\kappa+2}{\kappa}\right)^{\kappa} \frac{2^{\kappa}\mathbb{E}\left[W^{\kappa}|P - \mathtt{WAP} - \frac{x}{\sqrt{n}}|^{\kappa}\right]}{n^{\frac{\kappa}{2}-1}[x\mathbb{E}[W]]^{\kappa}} + \exp\left(-\frac{2[x\mathbb{E}[W]]^2}{(\kappa+2)^2 e^{\kappa}\mathbb{E}\left[W^2(P - \mathtt{WAP} - \frac{x}{\sqrt{n}})^2\right]}\right), \tag{2.15}$$

$$\mathscr{W}_n^{\leq}(x) \leq \left(\frac{\kappa+2}{\kappa}\right)^{\kappa} \frac{2^{\kappa}\mathbb{E}\left[W^{\kappa}|P - \mathtt{WAP} + \frac{x}{\sqrt{n}}|^{\kappa}\right]}{n^{\frac{\kappa}{2}-1}[x\mathbb{E}[W]]^{\kappa}} + \exp\left(-\frac{2[x\mathbb{E}[W]]^2}{(\kappa+2)^2 e^{\kappa}\mathbb{E}\left[W^2(P - \mathtt{WAP} + \frac{x}{\sqrt{n}})^2\right]}\right). \tag{2.16}$$

In the above result, we restrict ourselves to the case where $\kappa > 2$. The extension to $\kappa \in (1,2]$ could be handled using the Bahr-Esseen inequality [Bahr and Esseen, 1965, Section 5], which would lead to fluctuations in (2.11)-(2.12) of order $1/n^{1-1/\kappa}$ instead of $1/\sqrt{n}$; we leave this case aside because its statistical fluctuation properties are bad.

In the next result, we make use of a Young function (convex, continuously increasing, mapping $[0,+\infty)$ to $[0,+\infty)$) parametrized by $L > 0$:

$$\Psi_L(x) := \exp\left(\left(\frac{\sqrt{1+2Lx}-1}{L}\right)^2\right) - 1.$$

It permits to define an Orlicz norm [van de Geer and Lederer, 2013] (called the Bernstein-Orlicz norm) on sub-gamma random variables: for any real random variable $X$, set

$$\|X\|_{\Psi_L} := \inf\left\{c > 0 : \mathbb{E}\left[\Psi_L\left(\frac{|X|}{c}\right)\right] \leq 1\right\}.$$

See [Krasnoselskii and Rutickii, 1961] for a general account on Orlicz norms.

**Theorem 2.4** (Deviation bounds under sub-gamma assumptions)**.** *Assume* (**H0**) *and* ($\mathbf{H}_{\mathbf{X}}^{\Gamma}$) *for* $X = W$. *Then* $W$ *has a finite Bernstein-Orlicz norm* $\|W\|_{\Psi_L}$. *In addition,*

$$\mathscr{Q}_n^{>}(\alpha, x) \leq \exp\left( - \frac{\left( \mathbb{E}[W]\, x f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha), \frac{x}{\sqrt{n}}]) / (2[\alpha \vee (1-\alpha)]\, \|W\|_{\Psi_L}(1+L)) \right)^2}{\left( \sqrt{\frac{\mathbb{E}[W] x f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha), \frac{x}{\sqrt{n}}])}{(2[\alpha\vee(1-\alpha)]\|W\|_{\Psi_L}(1+L))\sqrt{n}}} + \frac{1}{2} + \frac{1}{\sqrt{2}} \right)^2} \right), \tag{2.17}$$

$$\mathscr{Q}_n^{\leq}(\alpha, x) \leq \exp\left( - \frac{\left( \mathbb{E}[W]\, x f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha), \frac{-x}{\sqrt{n}}]) / (2[\alpha \vee (1-\alpha)]\, \|W\|_{\Psi_L}(1+L)) \right)^2}{\left( \sqrt{\frac{\mathbb{E}[W] x f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha), \frac{-x}{\sqrt{n}}])}{(2[\alpha\vee(1-\alpha)]\|W\|_{\Psi_L}(1+L))\sqrt{n}}} + \frac{1}{2} + \frac{1}{\sqrt{2}} \right)^2} \right). \tag{2.18}$$

*Assume* ($\mathbf{H}_{\mathbf{X}}^{\Gamma}$) *for* $X = W$ *and* $X = W \cdot P$. *Then,*

$$\max\left(\mathscr{W}_n^{\geq}(x), \mathscr{W}_n^{\leq}(x)\right) \leq \exp\left( - \frac{\left( x\mathbb{E}[W] / (2(\|W(P - \mathtt{WAP})\|_{\Psi_L} + \|W\|_{\Psi_L} \frac{x}{\sqrt{n}})(1+L)) \right)^2}{\left( \sqrt{\frac{x\mathbb{E}[W]}{(2(\|W(P-\mathtt{WAP})\|_{\Psi_L} + \|W\|_{\Psi_L} \frac{x}{\sqrt{n}})(1+L))\sqrt{n}}} + \frac{1}{2} + \frac{1}{\sqrt{2}} \right)^2} \right). \tag{2.19}$$

The next result is dedicated to sub-Gaussian distribution, for which we consider an appropriate Orlicz norm

$$\|X\|_{\Psi_G} := \inf\left\{ c > 0 : \mathbb{E}\left[ \Psi_G\left( \frac{|X|}{c} \right) \right] \leq 1 \right\},$$

where $\Psi_G(.)$ is the Young function

$$\Psi_G(x) = \exp\left(x^2\right) - 1.$$

**Theorem 2.5** (Deviation bounds under sub-Gaussian assumptions)**.** *Assume* (**H0**) *and* ($\mathbf{H}_{\mathbf{X}}^{\mathbf{G}}$) *for* $X = W$. *Then* $W$ *has a finite Orlicz norm* $\|W\|_{\Psi_G}$. *In addition,*

$$\mathscr{Q}_n^{>}(\alpha, x) \leq \exp\left( - \frac{(\mathbb{E}[W]\, x f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha), \frac{x}{\sqrt{n}}]))^2}{8[\alpha \vee (1-\alpha)]^2 \|W\|_{\Psi_G}^2} \right), \tag{2.20}$$

$$\mathscr{Q}_n^{\leq}(\alpha, x) \leq \exp\left( - \frac{(\mathbb{E}[W]\, x f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha), \frac{-x}{\sqrt{n}}]))^2}{8[\alpha \vee (1-\alpha)]^2 \|W\|_{\Psi_G}^2} \right). \tag{2.21}$$

*Assume* ($\mathbf{H}_{\mathbf{X}}^{\Gamma}$) *for* $X = W$ *and* $X = W \cdot P$. *Then,*

$$\max\left(\mathscr{W}_n^{\geq}(x), \mathscr{W}_n^{\leq}(x)\right) \leq \exp\left( - \frac{(\mathbb{E}[W]\, x)^2}{8(\|W(P - \mathtt{WAP})\|_{\Psi_G} + \|W\|_{\Psi_G} \frac{x}{\sqrt{n}})^2} \right). \tag{2.22}$$

The proofs of Theorems 2.2, 2.3, 2.4 and 2.5 are given in Appendix A.
In the above statements, there is no reason why the constants should be optimal. The above bounds make use of sharp concentration-inequality results (under various distribution tail assumptions) but

it could be that our aggregate bounds are sub-optimal in the dependency in $x$ and $n$. Addressing the optimality of the above bounds is a delicate issue.
To better grasp the essence of the above bounds, let us focus on the dependency in $x$ and $n$ only, and replace all other factors by a generic constant $c$: in that case, left and right deviations are similar and can be summarized in the following table.

| | $\max(\mathcal{Q}_n^{>}(\alpha,x),\mathcal{Q}_n^{<}(\alpha,x))$ | $\max(\mathcal{W}_n^{\geq}(x),\mathcal{W}_n^{\leq}(x))$ |
|---|---|---|
| Heavy-Tail assumptions | $\dfrac{c}{n^{\frac{\kappa}{2}-1}x^{\kappa}}+\exp\left(-cx^2\right)$ | $\dfrac{c\left(1+\frac{x}{\sqrt{n}}\right)^{\kappa}}{n^{\frac{\kappa}{2}-1}x^{\kappa}}+\exp\left(-\dfrac{cx^2}{1+c\frac{x^2}{n}}\right)$ |
| Sub-gamma assumptions | $\exp\left(-\dfrac{cx^2}{1+c\frac{x}{\sqrt{n}}}\right)$ | $\exp\left(-\dfrac{cx^2}{\left(1+c\frac{x}{\sqrt{n}}\right)^3}\right)$ |
| Sub-Gaussian assumptions | $\exp\left(-cx^2\right)$ | $\exp\left(-\dfrac{cx^2}{1+c\frac{x^2}{n}}\right)$ |

Table 1: Simplified bounds for Theorems 2.3, 2.4, 2.5

Unsurprisingly, the sub-Gaussian setting offers the best confidence intervals: for a relatively small value of $x$, the probability of deviation is quite small (exponentially small in $x$), ensuring a good accuracy of both estimators $\widehat{q_{\mathtt{W},n}}(\alpha)$ and $\widehat{\mathtt{WAP}}_n$. Under a sub-gamma hypothesis, the accuracy is comparable and slightly deteriorated by the term $x/\sqrt{n}$: this setting is essential for us because it corresponds to the framework in which we study our RWM estimator for the choice $W=\log(1+V)$ (Equation (2.23)), see details in Subsection 2.3. On the other hand, the accuracy under heavy-tail assumptions is poor because of the term $\frac{c}{n^{\frac{\kappa}{2}-1}x^{\kappa}}$, especially when $n$ is small. Note that $\kappa$ is linked to the heaviness of the tail, which is characterized by the tail-index $\gamma$ (see Section 3). Therefore, even in favorable cases ($\kappa\geq 2 \Leftrightarrow \gamma\leq 0.5$), we have $n^{\frac{\kappa}{2}-1}\approx 1$ which is quite small and forces $x$ to be very large to guarantee a high probability of confidence intervals. In other words, in the heavy-tail setting, statistical fluctuations of the estimators are presumably huge compared to the sub-gamma setting. We recall that the well-know aggregation price estimators $\widehat{\mathtt{VWM}}_n$ and $\widehat{\mathtt{VWAP}}_n$ can be analyzed via Theorem 2.3 (because volumes are heavy tailed, see Section 5), which shows that high statistical accuracy for these estimators cannot be expected.

## 2.3 Final estimator

As highlighted previously, the statistical fluctuations of $\widehat{\mathtt{VWM}}_n$ (*i.e.* $\widehat{q_{\mathtt{W},n}}\left(\frac{1}{2}\right)$ with $W=V$) are directly impacted by the tail-heaviness of $V$, and the latter empirically appeared to be heavy tailed (see Figure 2b). Therefore, in order to benefit from the accuracy bounds discussed above under the sub-gamma setting, we shall consider a weighting variable $W$ with the adequate properties. To this end, for all $(x,y)\in[0,\infty)\times(0,\infty)$, let

$$\varphi(x,y):=\log\left(1+\frac{x}{y}\right),$$

which is similar to the Cauchy/Lorentzian loss function [Black and Anandan, 1996], well-known for the robustness property it entails for statistical learning problems. Thus, define our proposed estimator

$$\widehat{\mathtt{WM}}_n := \widehat{q_{\mathtt{W},n}}\left(\frac{1}{2}\right) \quad \text{with} \quad W := \log\left(1 + \frac{V}{q_V(1/2)}\right) \tag{2.23}$$

and $q_V(1/2)$ the median of $V$. This new estimator behaves like $\widehat{\mathtt{VWM}}_n$ for small volume data points because $\log(1+x) \sim_{x\to 0} x$, and it increasingly weighs prices as volumes increase, which is one of the desired properties (as for $\widehat{\mathtt{VWM}}_n$ and $\widehat{\mathtt{VWAP}}_n$) of an aggregator price. In addition, it is straightforward to observe that that since the survival function of the heavy-tailed random variable $V$ decays by definition (see Paragraph 3) at a rate $v^{-1/\gamma_V}$ as $v \to \infty$ with $\gamma_V > 0$, $W$ satisfies $\mathbf{H}^{\Gamma}_{\mathbf{X}}$ for $X = W$. Therefore, Theorem 2.4 applies and yields small statistical fluctuations of $\widehat{\mathtt{WM}}_n$, even in presence of small number $n$ of data points.

**Practical computation of $\widehat{\mathtt{WM}}_n$.** Starting from the price and weight observations $\{(P_i, W_i)\}_{i=1}^n$, consider a permutation $\pi$ of the price indices $\{1, \cdots, n\}$ such that $P_{\pi(1)} \le P_{\pi(2)} \le \cdots \le P_{\pi(n)}$. Let $h_k^{\uparrow}$ be the probability of the cumulative weight

$$h_k^{\uparrow} := \frac{\sum_{i=1}^k W_{\pi(i)}}{\sum_{i=1}^n W_i} \quad \text{for} \quad k \in \{1, \ldots, n\}, \qquad \text{with the convention that} \quad h_0^{\uparrow} := 0,$$

then define $k^\star$ as the unique value such that $h_{k^\star}^{\uparrow} \le 0.5$ and $h_{k^\star+1}^{\uparrow} > 0.5$. Note that this weighting rule is based on the increasing ordering of $P$. We are now in a position to write the estimator $\widehat{\mathtt{WM}}_n$ usually found in the literature (see [CF Benchmarks, 2022b, p.9]):

$$\widehat{\mathtt{WM}}_n = \begin{cases} \frac{P_{\pi(k^\star)} + P_{\pi(k^\star+1)}}{2}, \text{ if } k^\star \neq 0, \\ P_{\pi(1)}, \text{ otherwise.} \end{cases} \tag{2.24}$$

**Practical computation of $\widehat{\mathtt{RWM}}_n$.** In addition to the new weight transformation (2.23), our proposed RWM estimator introduces two additional features compared to $\widehat{\mathtt{WM}}_n$ in order to estimate a weighted median: a) a quantile interpolation and b) an additional decreasing ordering of prices. First, the half-sum in (2.24) is replaced by a linear interpolation (which allow to have an output continuous with respect to the volume inputs). Second, an additional estimator $\widehat{\mathtt{WM}_n^{\downarrow}}$ is computed using the reversed cumulative weight, i.e. $h_k^{\downarrow} = \frac{\sum_{i=1}^k W_{\pi(n-i+1)}}{\sum_{i=1}^n W_i}$ (equivalent to consider a decreasing ordering of $P$), and afterward, an average of decreasing and increasing ordering estimators is produced.[2] All in all, the RWM estimator is defined as

$$\widehat{\mathtt{RWM}}_n := \widehat{\mathtt{RWM}}_n(P, W) := \frac{\widehat{\mathtt{WM}_n^{\uparrow}}(P, W) + \widehat{\mathtt{WM}_n^{\downarrow}}(P, W)}{2},$$

[2] Usually, researchers and practitioners don't pay much attention to the distinction of decreasing and increasing ordering. However, this makes a difference when weighting is accounted for, as well as for small sample size.

where $\widehat{\mathtt{WM}}_n^{\uparrow}$ and $\widehat{\mathtt{WM}}_n^{\downarrow}$ respectively define the robust weighted median estimator from below and from above, i.e.,

$$\widehat{\mathtt{WM}}_n^{\uparrow}(P,W) := \begin{cases} P_{\pi(k^\star)} + (0.5 - h^{\uparrow}_{k^\star}) \left( \frac{P_{\pi(k^\star+1)} - P_{\pi(k^\star)}}{h^{\uparrow}_{k^\star+1} - h^{\uparrow}_{k^\star}} \right) \text{ if } k^\star \neq 0 \\ P_{\pi(1)} \text{ otherwise,} \end{cases}$$

and

$$\widehat{\mathtt{WM}}_n^{\downarrow}(P,W) := -\widehat{\mathtt{WM}}_n^{\uparrow}(-P,W).$$

We discuss the properties of $\widehat{\mathtt{RWM}}_n$ as a price aggregator, as mentioned in the introduction. We follow the classification of [Paine and Knottenbelt, 2016].

*Relevance.* The estimator $\widehat{\mathtt{RWM}}_n$ is a relevant price aggregator because, since it is built as a median, the estimator reflects a price consensus.

*Timeliness.* In our tests, the estimator is calculated every minute (so as to have around 100 data points to aggregate), as such it is reactive to market movements (on a minute scale). It could certainly aggregate prices at a higher frequency.

*Manipulation Resistance.* The question is how tiny or huge orders affect our estimator $\widehat{\mathtt{RWM}}_n$. On the one hand, micro-orders will multiply the number of trades but without really affecting the estimator because the price data associated with small volumes are not weighted very much (but proportionally to the volumes). On the other hand, it will cost a trader considerable effort to significantly move the price estimator due to the logarithmic function in the volume; note that it is easier to manipulate prices based on $\widehat{\mathtt{VWM}}_n$ and $\widehat{\mathtt{VWAP}}_n$. All in all, our estimator $\widehat{\mathtt{RWM}}_n$ is manipulation resistant.

*Martingale Property.* Since our estimator is localized in time, it is not possible to game the market by predicting the behavior of the price estimator. We will see in our tests (Sections 4 and 5) that $\widehat{\mathtt{RWM}}_n$ allows to recover the realized variance of a noisy price time series, showing that the trajectory properties are not altered by our estimator.

*Verifiability.* By publishing the index calculation methodology, we guarantee the transparency of the method and the possibility to independently verify the calculations.

*Replicability.* It is obviously very easy to reproduce the estimated values from the input data.

*Stability.* By construction, the estimator is robust to outliers or extreme values in prices and volumes, as a median is performed on prices and a logarithmic weighting on volumes is performed. Moreover, Theorem 2.4 shows excellent theoretical properties related to confidence intervals, which will be confirmed in subsequent tests. Its statistical stability, comparable to a Gaussian estimator (see Table 1), shows that it can work with scarce data (and thus is resistant to missing data).

*Parsimony.* Our estimator has no parameters to set, and thus, it offers optimal parsimony properties.

This section provided theoretical guidelines in order to 1) explain the poor performance of the commonly used $\widehat{\mathtt{VWM}}_n$ and $\widehat{\mathtt{VWAP}}_n$ estimators, and 2) justify the appropriate heavy-tailed volume transformation for

defining the weights in order to benefit from better statistical stability. Both points will be illustrated in the next sections through experiments on both synthetic and real data.

## 3 Data description

The $\widehat{\mathtt{RWM}}_n$ estimator is tested on both simulated and real cryptocurrency data. In order to parameterize the former dataset, we first study the statistical behavior of the latter. Using Kaiko's Instrument Explorer[3] and Trades product [4], we selected all the trades made on all the available exchanges for 67 pairs during 4 different days, including both calm (2022-06-28, 2022-12-09) and stressed (2022-06-12, 2022-05-05) periods in the market. In total, we collected 268 use cases (one pair for one period) for a total of 80,448,919 trades with their associated volume, price and computed ordinary returns. More details on the data statistics are reported in Table 2. In the next paragraph we study the tail behavior of the variables of interest which, as previously proved in Section 2.2, plays a crucial role in the deviation bounds of the different estimators.

| | Number of trades | | Returns | | |
|---|---|---|---|---|---|
| Period | Mean | Std. | Mean | Std. | Ann. Vol. |
| 2022-05-05 | 96 | 156 | $-0.44\cdot10^{-4}$ | $8\cdot10^{-4}$ | 55% |
| 2022-06-12 | 188 | 311 | $-0.19\cdot10^{-4}$ | $15\cdot10^{-4}$ | 104% |
| 2022-06-28 | 94 | 149 | $-0.24\cdot10^{-4}$ | $8\cdot10^{-4}$ | 56% |
| 2022-12-09 | 42 | 74 | $-0.04\cdot10^{-4}$ | $5\cdot10^{-4}$ | 36% |

Table 2: Data statistics (number of trades, returns) per minute by discarding 5% of the lowest and the highest data in order to obtain robust metrics that will be used for generating synthetic data. The annualized volatility associated with the standard deviation of the returns is added.

**Heavy-tail modelling.** Extreme value theory provides a statistical framework on tail behaviors and it clusters distributions according to the decay rate of their survival functions driven by a real parameter $\gamma$, called the tail-index (a.k.a. extreme-value index). Heavy-tailed distributions (*e.g.* Pareto, Student, Cauchy) are related to $\gamma > 0$ where the associated survival function decays at a rate $x^{-1/\gamma}$ as $x \to \infty$, which implies ([de Haan and Ferreira, 2006, Theorem 1.2.1] and [de Haan and Ferreira, 2006, Proposition B.1.9.9]) that the quantile diverges at a rate $(1-u)^{-\gamma}$ as $u \to 1$. Heavy-tailed distributions have been revealed useful to describe the tail structure of actuarial and financial data [Embrechts et al., 1997, p.8], [Resnick, 2007, Section 1.3.2].

**Estimating the tail-index.** The most well-known tail-index estimator is the Hill estimator [Hill, 1975], which is known to be biased [de Haan and Ferreira, 2006, Theorem 3.2.5, p. 74]. Since then, several works have proposed bias reduction techniques in order to improve the original Hill estimator. Such techniques mainly rely on the second-order condition [de Haan and Ferreira, 2006, Section 2.3] introducing a second-order parameter $\rho < 0$ which is the main driver of the bias in most of the tail-index estimators. Using the second-order estimator implemented in the `R` package `evt0` [Manjunath et al., 2013] on all

[3] https://instruments.kaiko.com/#/instruments
[4] https://docs.kaiko.com/#historical-trades

the use cases, it appears that $\hat{\rho} \approx -0.7$ which is close to what was estimated on the traditional financial market during a stressed period [Allouche et al., 2022b]. Such a high value of $\hat{\rho}$ indicates a relatively large bias situation, and so it supports the use of a refined method for the tail-index estimator. In this context, we consider the one proposed recently in [Allouche et al., 2022a], which is illustrated in Figure 2a for one use case. The advantage of this method is that it searches for anchor points in a much smaller area than the usual methods, which makes it easier to find them, especially in an automated way when data sets may have quite different sizes. Estimating the tail-index for all the use cases, it appears in Figure 2b that the distributions of the volumes, of the price returns and of the product of the two are heavy-tailed. Note that the tail of the volume ($\hat{\gamma}_V \approx 0.6$) is very heavy (no finite variance) and is significantly heavier than the one on the price returns ($\hat{\gamma}_P \approx 0.4$), while the tail of the product is heavier than the max of the two but lower than their sum. This phenomenon indicates the presence of tail dependence.

**Tail dependence.** On the one hand, it is known [McNeil et al., 2005, Theorem 7.35 p. 295], [Breiman, 1965] that the survival function of the product of two independent heavy-tailed random variables behaves as the one with the heaviest tail. On the other hand it is easy to check that the tail-index of the product of two fully dependent and comonotone heavy-tailed random variables is equal to the sum of the two tail-indices. In this context, this requires the study of the empirical joint dependence and more precisely in the right tail, which is related to the delicate notion of upper tail dependence (see [Joe, 1997, Section 2.1] among others). For a random vector $(X_1, X_2)$, the latter is defined as

$$\lambda = \lim_{u \to 1^-} \mathbb{P}\left(X_1 > F_{X_1}^{-1}(u) | X_2 > F_{X_2}^{-1}(u)\right)$$

if the limit exists. Several non-parametric estimators have been developed (see [Garcin and Nicolas, 2021] for an overview), but here we use the one proposed in [Frahm et al., 2005], illustrated in Figure 2c for one use case. Estimating the upper tail dependence for all of them, it appears in Figure 2d that our assumption on the presence of tail dependence is confirmed with $\hat{\lambda} \approx 0.1$.

# 4 Validation of $\widehat{\texttt{RWM}}_n$ on synthetic data

## 4.1 Experimental design

The objective is to assess the statistical behavior of our aggregated estimator on simulated prices and volumes, using a model having similar statistical properties than those described in Section 3. To mimic the situation of data that is fragmented over different platforms, we consider a single price path $\{S_t\}_{t=1}^T$ which is supposedly not observed directly and which plays the role of an efficient price to be recovered from observations; at each time $t$ we assume the observations of $n = 100$ noisy prices $\widetilde{S}_t^j = S_t(1 + r_t^j)$, perturbed by either positive or negative noises $r_t^j$ for all $j \in \{1, \dots, n\}$. The efficient price path is simulated arbitrarily with a Geometric Brownian Motion (GBM) parameterized by a drift $\mu$ and a volatility $\sigma$:

$$S_{t+1} = S_t \exp\left((\mu - \sigma^2/2)d_t + \sigma\varepsilon\sqrt{d_t}\right),$$

with the time variable $t = 0, \dots, T-1$ for $T = 1440$ minutes (1 day) and with a time step $d_t = 1/(365 \times 24 \times 60)$. We set $\mu = 0$ and $\sigma = 0.5$ according to Table 2 on real data statistics.
First, the noisy returns $r_t^j$ are drawn from a mixture model with two components, each distributed

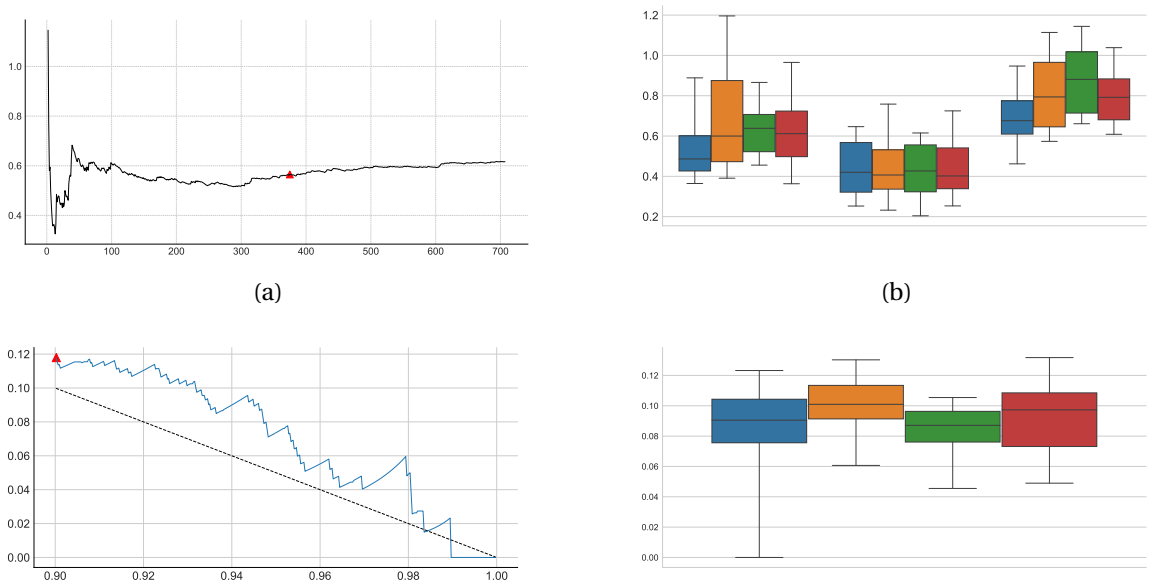



Figure 2: Illustration of extreme statistics on real cryptocurrency pairs aggregated per second. (a): Hill estimators as functions of anchor points in a restricted adapted range for bias reduction purposes [Allouche et al., 2022a] (black) regarding the price returns of the pair btc/usdc on 2022-05-05. The estimated tail-index selected by [Allouche et al., 2022a, Algorithm 1] is emphasized by a triangle. (b): Box plot of the estimated tail-index on the volumes (left), on the price returns (middle) and on the product of the two (right) for all considered pairs on 2022-05-05 (blue), 2022-06-12 (orange), 2022-06-28 (green) and 2022-12-09 (red). Outliers are discarded from the box plot. (c): Upper tail dependence estimator between price returns and volumes (blue), and the independence case (black dashed line) $u \mapsto 1-u$ as a function of $u \in (0.9, 1)$ for the pair btc/usdc on 2022-05-05. The estimated upper tail dependence selected by [Frahm et al., 2005, Section 4.4] is emphasized by a triangle. (d): Box plot of the estimated upper tail dependence for all considered pairs and periods. Both the color legend and the outliers display are similar to 2b. The whiskers, in the box plots 2b, 2d and in the following ones, represent respectively the percentiles 5% and 95%.

respectively as light tailed (small returns) and heavy tailed (large returns), with a mixture parameter $\omega \in [0,1]$ controlling the proportion of the former regime. Incorporating a mixture proportion parameter $\omega$ will allow us to simulate an increase in the number of data points with heavy tails on the returns, so as to mimic the effect of outliers. Thus, we will be able to test how robust the price aggregator RWM is to outliers, with an increasing proportion of outliers. In [CF Benchmarks, 2022b] [CF Benchmarks, 2022a], the percentage of erroneous data and outliers is estimated to 10% for normal situations, which corresponds to $\omega = 0.9$ in our framework. Since the efficient prices have Gaussian log-returns, it is natural to suppose that the small returns are centered Gaussian perturbations with a variance of order $(8 \cdot 10^{-4})^2$ related to the one observed in Table 2. The large returns are simulated from a Burr$(\gamma, \rho)$ distribution parameterized by the first two estimated order parameters $\hat{\gamma}_P = 0.4$ and $\hat{\rho}_P = -0.7$ according to Figure 2b,

with a quantile function defined as

$$q^{\texttt{Burr}}_{\gamma,\rho}(u) = \left((1-u)^{\rho}-1\right)^{-\frac{\gamma}{\rho}}, \qquad \forall\, u \in (0,1).$$

Second, we consider two scenarios for simulated volumes that are either light or heavy-tailed. In the former case we consider a standard Normal distribution with absolute value, while in the latter one we consider a Burr distribution with appropriate parameters $\hat{\gamma}_V = 0.6$ and $\hat{\rho}_V = -0.7$ according to Figure 2b. Regarding the modelling of the dependency between returns and volumes, and based on the empirically observed dependency property, the data simulation is performed using copula (see [Nelsen, 2006] for an overview), which allows to model separately the dependence structure and the margins with parameters coming from real data statistics in Section 3. The simulation of the dependent variables $P$ and $V$ is done through an Archimedean copula defined in a bivariate case for all $(u_1, u_2) \in [0,1]^2$ as

$$C_\theta(u_1,u_2) = \psi_\theta(\psi_\theta^{-1}(u_1) + \psi_\theta^{-1}(u_2)),$$

where $\psi_\theta$ is a decreasing and convex function on $[0,1] \to [0,\infty]$ parameterized by $\theta$ with $\psi_\theta(1) = 0$ and $\psi_\theta(0) = \infty$. See [McNeil and Nešlehová, 2009] for all listed properties on Archimedean copulas and [Allouche et al., 2022b, Appendix A p.25] for the sampling schema in a bivariate case. Here we focus on the Gumbel copula where the associated generating function is $\psi_\theta(t) = \exp(-t^{1/\theta})$ defined for all $\theta \geq 1$ and $t \geq 0$. It is well known [Nelsen, 2006, 5.4 p.215] that $\lambda = 2 - 2^{1/\theta}$, so we consider $\theta = \log 2/\log 1.9$ based on $\hat{\lambda} = 0.1$ estimated in Figure 2d.

## 4.2 Performance assessment

The numerical tests allow to evaluate for different mixture scenarios

$$\omega \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$$

(the larger amount of heavy tails – or outliers – on returns corresponding to $\omega = 0.9$ assuming we have to include unreliable data sources), the capacity of our model to estimate: a) the efficient price $S_t$ at a time $t$, and b) the annualized Realized Volatility (RV) associated with the efficient price

$$\sigma^{\texttt{eff}} = \sqrt{365}\sqrt{\sum_{t=1}^{T-1} \log\left(\frac{S_{t+1}}{S_t}\right)^2}, \tag{4.1}$$

which is a key tool in risk management. The experiments are replicated $M = 1000$ times in order to compare the Relative Mean Square Error (RMSE) on the estimated quantities $\widehat{S}_{t,i}$ and $\widehat{\sigma}_i$ between the $\widehat{\texttt{RWM}}_n$ and the other competitors ($\widehat{\texttt{VWM}}_n$ and $\widehat{\texttt{VWAP}}_n$) for all $i \in \{1,\dots,M\}$ where

$$\text{RMSE}_t^{\texttt{price}} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{\widehat{S}_{t,i}}{S_{t,i}} - 1\right)^2 \quad \text{and} \quad \text{RMSE}^{\texttt{RV}} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{\widehat{\sigma}_i}{\sigma_i^{\texttt{eff}}} - 1\right)^2. \tag{4.2}$$

### 4.3 Simulation process

For each $t \in \{1,\dots,T\}$ and $j \in \{1,\dots,n\}$:

- Sample a random pair $U_1, U_2$ from the bivariate copula $C_\theta$.
- Simulate noisy prices $\widetilde{S}_t^j = S_t(1 + r_t^j)$ according to the simulated returns

$$r_t^j := \nu r_{t-1}^j + (1-\nu) R_t^j Z_t^j,$$

defined through an auto-regressive structure with a fixed $\nu = 0.5$, a Rademacher random variable $R_t^j$ and a mixture random variable

$$Z_t^j = \xi_t^j Y_t^j + (1-\xi_t^j)\frac{X_t^j - \mathbb{E}[X]}{q^{\texttt{Burr}}_{\hat{\gamma}_P,\hat{\rho}_P}(1-1/(nT))},$$

composed by

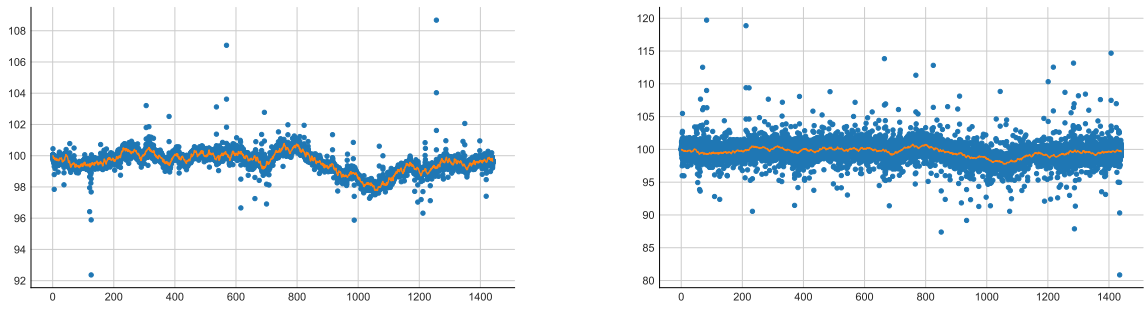
$$\begin{aligned} Y_t^j &= q^{\texttt{Normal}}_{\mu,\sigma^2}(U_1), \\ X_t^j &= q^{\texttt{Burr}}_{\hat{\gamma}_P,\hat{\rho}_P}(U_1), \quad \mathbb{E}[X] = \frac{\Gamma(-1/\hat{\rho}_P + \hat{\gamma}_P/\hat{\rho}_P)\Gamma(1-\hat{\gamma}_P/\hat{\rho}_P)}{\Gamma(-1/\hat{\rho}_P)}, \\ \xi_i^j &\sim \texttt{Bern}(1-\omega). \end{aligned}$$

Note that if a simulated price is negative, which should not statistically be that often thanks to the normalization $q^{\texttt{Burr}}_{\hat{\gamma}_P,\hat{\rho}_P}(1-1/(nT))$ in order to restrict $\left|r_t^j\right| < 1$, we clip it to $10^{-8}$.

- Simulate volumes either from a light-tailed $\widetilde{V}_t^j = |q^{\texttt{Normal}}_{0,1}(U_2)|$, or an heavy-tailed $\widetilde{V}_t^j = q^{\texttt{Burr}}_{\hat{\gamma}_V,\hat{\rho}_V}(U_2)$ distribution.

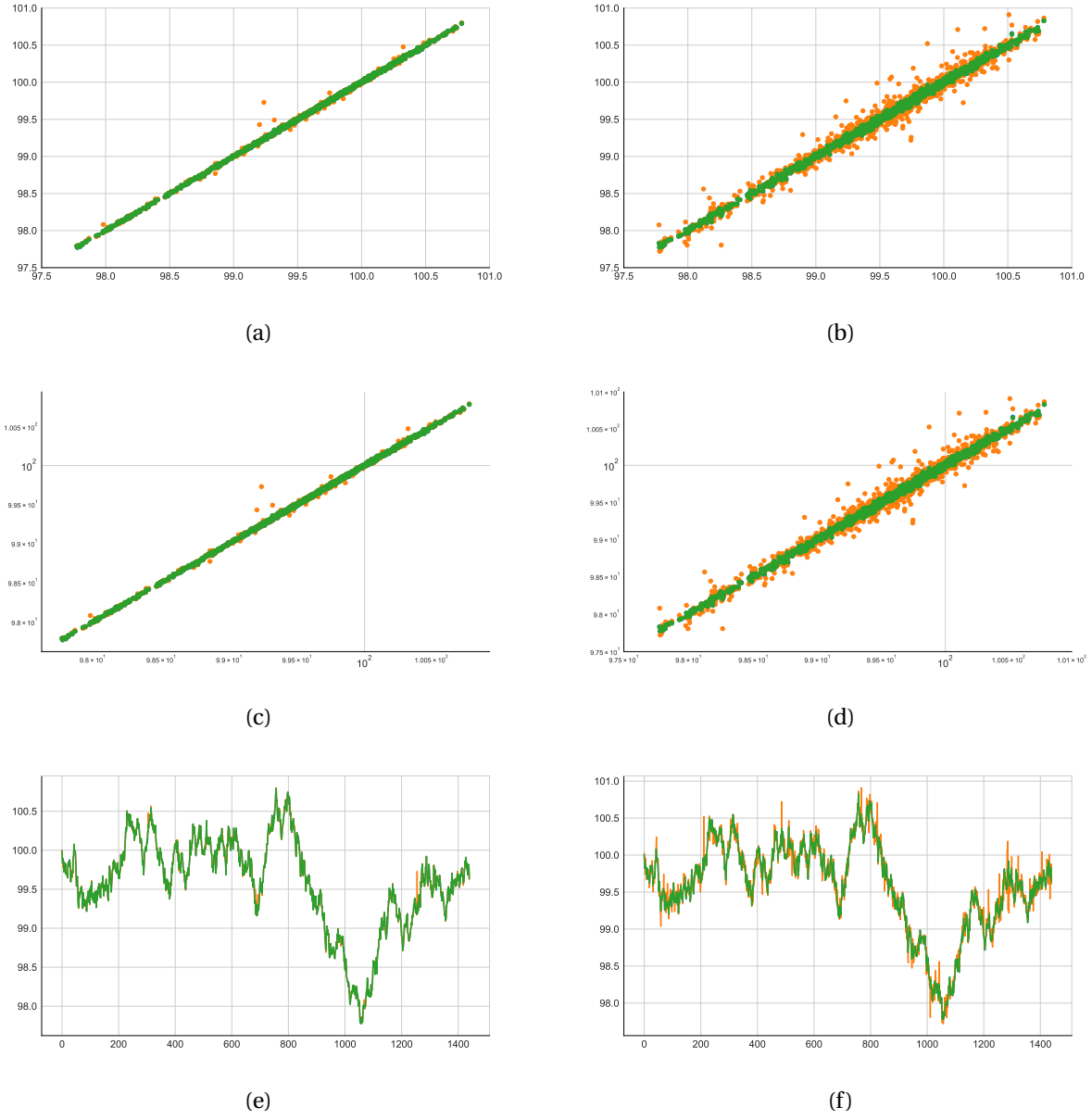
### 4.4 Results

We observe respectively in Figures 4 and 5 the light and heavy-tailed volume aggregated prices computed by the three estimators, based on the noisy prices simulated around efficient prices using either $\omega = 0.01$ (Figure 3a) or $\omega = 0.3$ ( Figure 3b). It appears that when $\omega = 0.01$ (left sides in Figures 4 and 5), all the estimators globally behave in the same way, with few spikes in the presence of heavy-tailed volumes. Conversely, when $\omega = 0.3$ (right sides in Figures 4 and 5) when large values occur on prices – coherently with heavy-tail empirical distributions –, and on volumes – coherently with either light or heavy-tail empirical distributions – both $\widehat{\texttt{RWM}}_n$ and $\widehat{\texttt{VWM}}_n$ or just $\widehat{\texttt{RWM}}_n$ are the only estimators to be resilient. Performance results (4.2) on the price estimation and on the realized volatility computed on each price aggregator estimations are presented in Tables 3 and 4 for each volume scenario considered. It appears that $\widehat{\texttt{RWM}}_n$ outperforms both $\widehat{\texttt{VWM}}_n$ and $\widehat{\texttt{VWAP}}_n$ on both volume assumptions, on both metrics and for all considered configurations of $\omega$. Although all the estimators are robust (RMSE is less than one) in the price estimation at $t = 1$, the estimators $\widehat{\texttt{VWM}}_n$ and $\widehat{\texttt{VWAP}}_n$ are robust in the annualized RV on light-tailed volumes until $\omega = 0.8$ and $\omega = 0.1$ respectively but none of them on heavy-tailed volumes, while $\widehat{\texttt{RWM}}_n$ is robust for all values of $\omega$ (except for $\omega = 0.9$ in the heavy-tailed case). This phenomenon highlights the ability of our estimator to compute key statistics in any situation, while the other methods fail.




(a) (b)

Figure 3: Efficient price time-series $(S_t)_{t=1}^{1440}$ (orange line) simulated with $\sigma = 0.5$ ($\sigma^{\texttt{eff}} = 0.51$), and its associated noisy prices $(\widetilde{S}_t^j, j \in \{1,\dots,100\})_{t=1}^{1440}$ (blue dots) with $\omega = 0.01$ (a) and $\omega = 0.3$ (b).



(a) (b)

(c) (d)

(e) (f)

Figure 4: Aggregation with light-tailed volumes using $\omega = 0.01$ (left side) and $\omega = 0.3$ (right side): (a)-(b): Scatter plots of the associated estimated prices $(\widehat{S}_t)_{t=1}^{1440}$ ($\widehat{\mathtt{VWM}}_n$: blue, $\widehat{\mathtt{VWAP}}_n$: orange, $\widehat{\mathtt{RWM}}_n$: green) with respect to the reference price. The best estimator is illustrated by the black dashed regression line $x \mapsto y = x$. Both horizontal and vertical axes are limited to 97.5 and 101 for a better display. (c)-(d): Scatter plots of the reference price estimation with log-scale axes. (e)-(f): Estimated price $(\widehat{S}_t)_{t=1}^{1440}$.

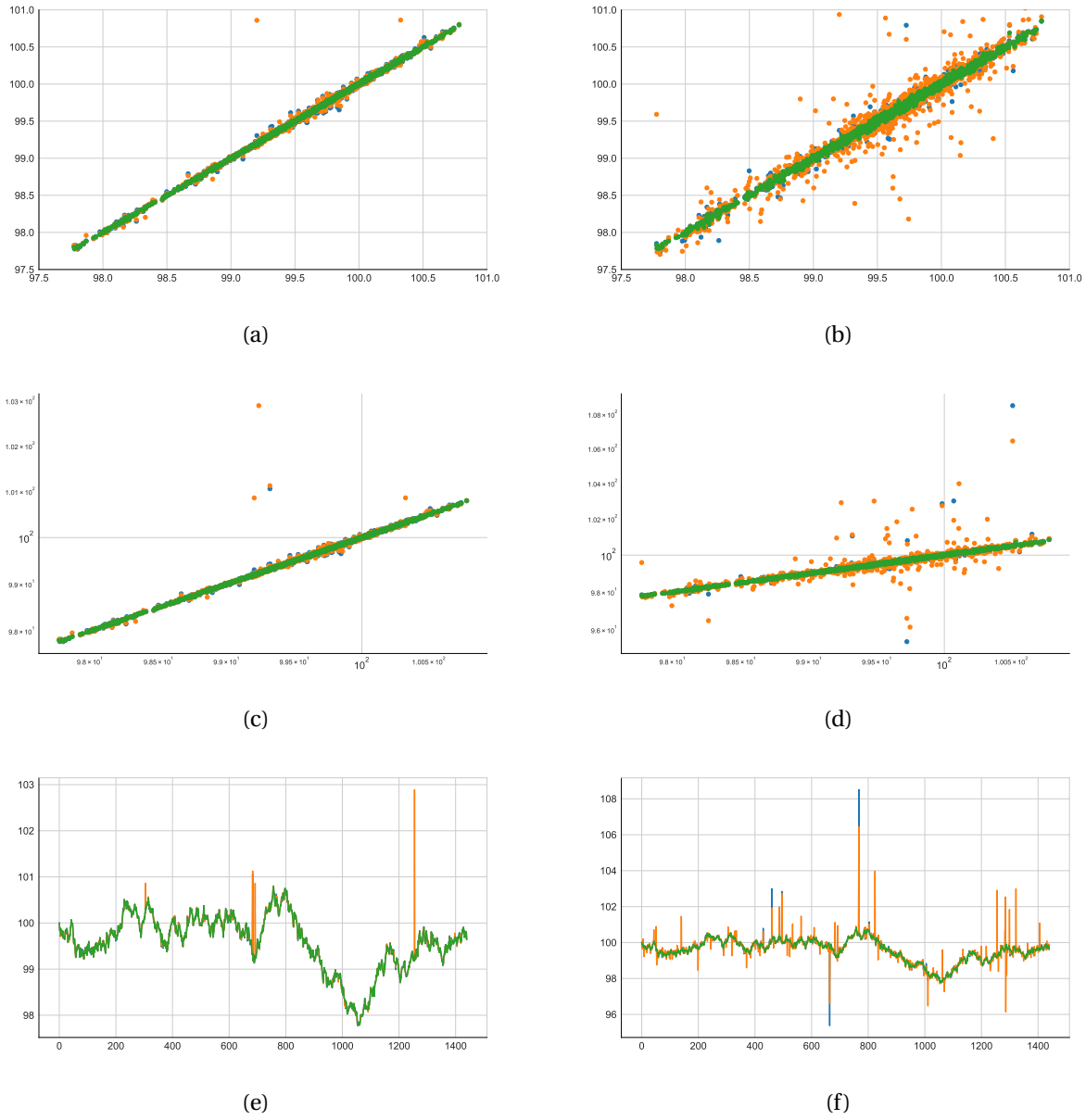


Figure 5: Aggregation with heavy-tailed volumes using $\omega = 0.01$ (left side) and $\omega = 0.3$ (right side): (a)-(b): Scatter plots of the associated estimated prices $(\widehat{S}_t)_{t=1}^{1440}$ ($\widehat{\mathtt{VWM}}_n$: blue, $\widehat{\mathtt{VWAP}}_n$: orange, $\widehat{\mathtt{RWM}}_n$: green) with respect to the reference price. The best estimator is illustrated by the black dashed regression line $x \mapsto y = x$. Both horizontal and vertical axes are limited to 97.5 and 101 for a better display. (c)-(d): Scatter plots of the reference price estimation with log-scale axes. (e)-(f): Estimated price $(\widehat{S}_t)_{t=1}^{1440}$.

| | $\text{RMSE}_1^{\texttt{price}}$ | | | $\text{RMSE}^{\texttt{RV}}$ | | |
|---|---|---|---|---|---|---|
| $\omega$ | $\widehat{\texttt{VWM}}_n$ | $\widehat{\texttt{VWAP}}_n$ | $\widehat{\texttt{RWM}}_n$ | $\widehat{\texttt{VWM}}_n$ | $\widehat{\texttt{VWAP}}_n$ | $\widehat{\texttt{RWM}}_n$ |
| 0.01 | 0.011 | 0.027 | **0.009** | 0.05 | 9.82 | **0.04** |
| 0.05 | 0.014 | 0.352 | **0.013** | 0.09 | 37.77 | **0.08** |
| 0.10 | 0.020 | 0.447 | **0.018** | 0.18 | 68.82 | **0.15** |
| 0.20 | 0.037 | 0.688 | **0.033** | 0.64 | – | **0.54** |
| 0.30 | 0.069 | 1.007 | **0.063** | 1.95 | – | **1.63** |
| 0.40 | 0.114 | 1.261 | **0.104** | 5.27 | – | **4.41** |
| 0.50 | 0.202 | 1.273 | **0.183** | 12.4 | – | **10.38** |
| 0.60 | 0.307 | 1.598 | **0.273** | 25.05 | – | **21.10** |
| 0.70 | 0.438 | 1.849 | **0.392** | 44.32 | – | **37.54** |
| 0.80 | 0.579 | 2.183 | **0.520** | 70.63 | – | **60.01** |
| 0.90 | 0.705 | 2.381 | **0.624** | – | – | **87.54** |

Table 3: Comparison between $\widehat{\texttt{VWM}}_n$, $\widehat{\texttt{VWAP}}_n$ and $\widehat{\texttt{RWM}}_n$ results for different simulation scenarios with light-tailed volumes and $\omega \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ using two performance criteria. Left: RMSE on price estimation at time $t = 1$, all results are scaled by $10^6$. Right: RMSE on realized volatility, all results are scaled by $10^2$. RMSEs larger than 1 are not reported. Best results are emphasized in bold.

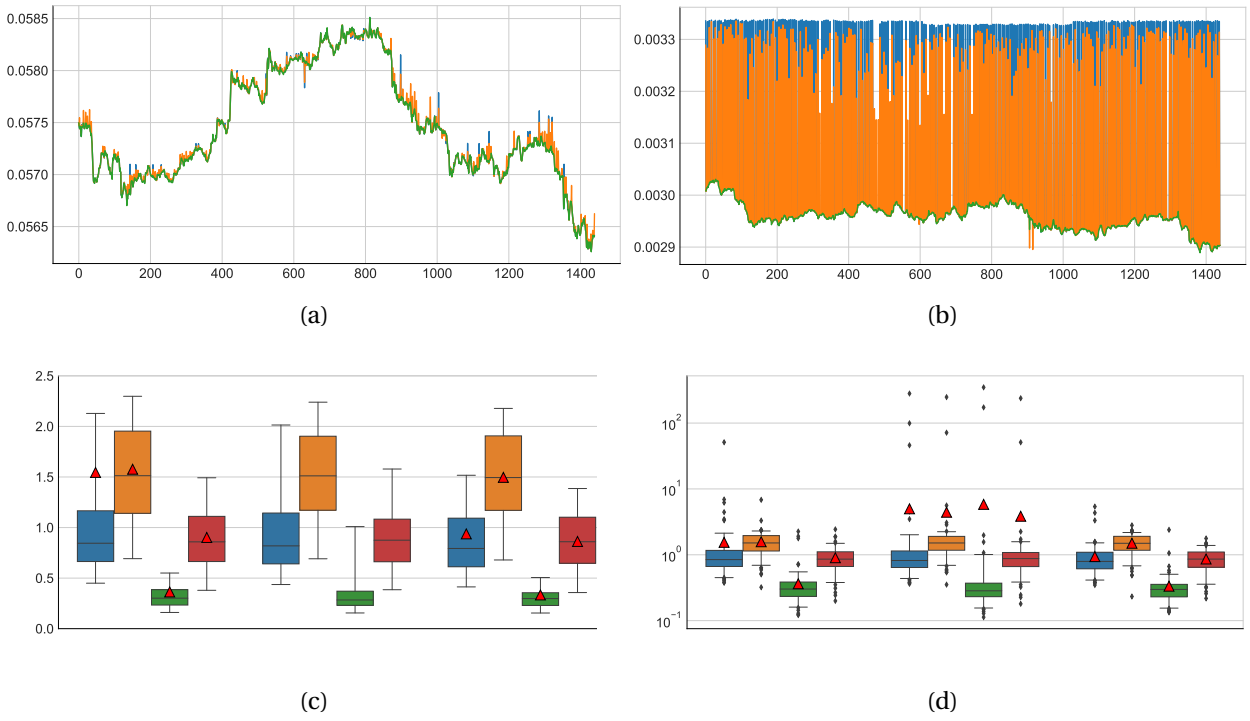
| | $\text{RMSE}_1^{\texttt{price}}$ | | | $\text{RMSE}^{\texttt{RV}}$ | | |
|---|---|---|---|---|---|---|
| $\omega$ | $\widehat{\texttt{VWM}}_n$ | $\widehat{\texttt{VWAP}}_n$ | $\widehat{\texttt{RWM}}_n$ | $\widehat{\texttt{VWM}}_n$ | $\widehat{\texttt{VWAP}}_n$ | $\widehat{\texttt{RWM}}_n$ |
| 0.01 | 0.061 | 0.054 | **0.010** | - | - | **0.06** |
| 0.05 | 0.070 | 3.534 | **0.014** | - | - | **0.10** |
| 0.10 | 0.377 | 4.425 | **0.02** | - | - | **0.20** |
| 0.20 | 2.591 | 6.156 | **0.037** | - | - | **0.68** |
| 0.30 | 6.281 | 8.849 | **0.066** | - | - | **2.05** |
| 0.40 | 6.838 | 12.259 | **0.114** | - | - | **5.50** |
| 0.50 | 26.302 | 34.546 | **0.207** | - | - | **12.91** |
| 0.60 | 37.159 | 42.815 | **0.301** | - | - | **26.00** |
| 0.70 | 37.987 | 45.717 | **0.433** | - | - | **46.09** |
| 0.80 | 41.842 | 47.865 | **0.566** | - | - | **73.23** |
| 0.90 | 44.582 | 49.523 | **0.701** | - | - | - |

Table 4: Comparison between $\widehat{\texttt{VWM}}_n$, $\widehat{\texttt{VWAP}}_n$ and $\widehat{\texttt{RWM}}_n$ results for different simulation scenarios with heavy-tailed volumes and $\omega \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ using two performance criteria. Left: RMSE on price estimation at time $t = 1$, all results are scaled by $10^6$. Right: RMSE on realized volatility, all results are scaled by $10^2$. RMSEs larger than 1 are not reported. Best results are emphasized in bold.

## 5 Illustration of $\widehat{\texttt{RWM}}_n$ on real data

Our goal here is to compare the different aggregated price estimators on real trades described in Section 3 based on the annualized RV criteria (4.1). Two use cases illustrated in Figure 6a and in Figure 6b

highlight real applications of price estimation on common and liquid pairs. While the outputs of the $\widehat{\texttt{RWM}}_n$ estimator are smooth, those of both $\widehat{\texttt{VWM}}_n$ and $\widehat{\texttt{VWAP}}_n$ estimators suffer from small (Figure 6a) and very large (Figure 6b) peaks due to outliers. Note that, similarly to the other estimators, $\widehat{\texttt{RWM}}_n$ does not suffer from delayed jump reactions (see for example minutes 50, 425 and 500 in Figure 6a). Computing the annualized RV for all the use cases, it appears in Figure 6c that the median is very close for all considered periods between the three estimators, which confirms that most of the time they behave similarly. Nevertheless, the lower average for the $\widehat{\texttt{RWM}}_n$ shows that it is less impacted by outliers compared to its competitors.




(a) (b)

(c) (d)

Figure 6: (a)-(b): Comparison between $\widehat{\texttt{VWM}}_n$ (blue), $\widehat{\texttt{VWAP}}_n$ (orange) and $\widehat{\texttt{RWM}}_n$ (green) on real data aggregation per minute of the pair eth/btc (a) and zec/btc (b) on 2022-06-28. (c): Box plot of the annualized RV (over all pairs) from $\widehat{\texttt{VWM}}_n$ (left), $\widehat{\texttt{VWAP}}_n$ (middle) and $\widehat{\texttt{RWM}}_n$ (right) on 2022-06-28 (blue), on 2022-06-12 (orange), on 2022-09-12 (green) and on 2022-05-05 (red). The mean is emphasized by a red triangle and outliers are discarded. d)Box plot of the annualized RV taking into account outliers with the y-axis in log-scale.

## 6 Conclusion

Our study has shown that standard price aggregators such as Volume Weighted Average (VWAP) and Volume Weighted Median (VWM) prices suffer from instability and lack of robustness when applied to cryptocurrency data that are heavy-tailed. Based on a deep statistical study on the real data, we have built an estimator, called Robust Weighted Median (RWM), that is robust to large statistical fluctuations caused by the heavy-tailed behavior of both the volumes and the price returns, while satisfying all the desired properties of a price aggregator. Our analysis is supported by new theoretical results in statistics: in particular, we have derived, to the best of our knowledge, the first probabilistic concentration

inequalities for weighted means and weighted quantiles under different tail assumptions (heavy tails, sub-gamma tails, sub-Gaussian tails). The sub-gamma case turns out to be the relevant setting for analyzing RWM, while the accuracy of usual price aggregation estimators are related to the concentration inequalities for heavy tails.
From a practical point of view, the estimator RWM outperforms both WVAP and VWM in terms of robustness and regularity during stressed scenarios (about 5 to 10% of all scenarios); while all the estimators behave more or less similarly most of the remaining time. The accurate estimation of risk management metrics such as the realized volatility in presence of contaminated data highlights the ability of RWM to compute key statistics in any situation, while the other methods fail. The excellent performance of the RWM price aggregator shows that it can be a reliable tool in the perspective of automatic price calculation, limiting to the maximum the post-processing by hand.

## References

[Allouche et al., 2022a] Allouche, M., El Methni, J., and Girard, S. (2022a). A refined Weissman estimator for extreme quantiles. *Extremes*. doi:10.1007/s10687-022-00452-8

[Allouche et al., 2022b] Allouche, M., Girard, S., and Gobet, E. (2022b). EV-GAN: Simulation of extreme events with ReLU neural networks. *Journal of Machine Learning Research*, 23(150):1–39.

[Arslanian, 2022] Arslanian, H. (2022). *The Book of Crypto*. Palgrave Macmillan Cham. The Complete Guide to Understanding Bitcoin, Cryptocurrencies and Digital Assets.

[Asmussen and Glynn, 2007] Asmussen, S. and Glynn, P. (2007). *Stochastic simulation: Algorithms and analysis*. Stochastic Modelling and Applied Probability 57. New York, NY: Springer.

[Bahr and Esseen, 1965] Bahr, B. and Esseen, C.-G. (1965). Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Annals of Mathematical Statistics*, 36(1):299–303.

[Bakker and Gravemeijer, 2006] Bakker, A. and Gravemeijer, K. An historical phenomenology of mean and median. *Educational Studies In Mathematics*. 62 pp. 149-168 (2006)

[Bhattacharya and Rao, 2010] Bhattacharya, R. and Rao, R. (2010). *Normal Approximation and Asymptotic Expansions*, volume 64. SIAM.

[Black and Anandan, 1996] Black, M. J. and Anandan, P. (1996). The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer vision and image understanding*, 63(1):75–104.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities. A nonasymptotic theory of independence*. Clarendon Press, Oxford. doi:10.1093/acprof:oso/9780199535255.001.0001

[Boucheron and Thomas, 2012] Boucheron, S. and Thomas, M. (2012). Concentration inequalities for order statistics. *Electronic Communications in Probability*, 17:1–12. doi:10.1214/ECP.v17-2210

[Breiman, 1965] Breiman, L. (1965). On some limit theorems similar to the arc-sin law. *Theory of Probability & Its Applications*, 10(2):323–331.

[Catoni, 2012] Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'IHP Probabilités et statistiques*, 48(4):1148–1185. doi:10.1214/11-AIHP454

[CF Benchmarks, 2022a] CF Benchmarks (2022a). CME CF cryptocurrency real time indices, version 15.1. https://docs.cfbenchmarks.com/CME%20CF%20Real%20Time%20Indices%20Methodology.pdf.

[CF Benchmarks, 2022b] CF Benchmarks (2022b). CME CF cryptocurrency reference rates, version 15.1. https://docs.cfbenchmarks.com/CME%20CF%20Reference%20Rates%20Methodology.pdf.

[Chamakh et al., 2022] Chamakh, L., Gobet, E., and Liu, W. (2022). Orlicz norms and concentration inequalities for $\beta$-heavy tailed random variables. *In minor revision for Bernoulli*.

[Chamakh et al., 2020] Chamakh, L., Gobet, E., and Szabo, Z. (2020). Orlicz random Fourier feature. *Journal of Machine Learning Research*, 21:1–37.

[de Haan and Ferreira, 2006] de Haan, L. and Ferreira, A. (2006). *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York. doi:10.1007/0-387-34471-3

[Del Moral, 2013] Del Moral, P. (2013). *Mean field simulation for Monte-Carlo integration*, volume 126 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL.

[Devroye et al., 2016] Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016). Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725. doi:10.1214/16-AOS1440

[Embrechts and Hofert, 2013] Embrechts, P. and Hofert, M. (2013). A note on generalized inverses. *Mathematical Methods of Operations Research*, 77(3):423–432.

[Embrechts et al., 1997] Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling extremal events*, volume 33 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin. For insurance and finance. doi:10.1007/978-3-642-33483-2

[Farebrother, 2001] Farebrother, R. W. (2001). Francis galton. In Heyde, C. C., Seneta, E., Crépel, P., Fienberg, S. E., and Gani, J., editors, *Statisticians of the Centuries*, pages 181–184. Springer New York, New York, NY.

[Federal Reserve Bank of New York, 2015] Federal Reserve Bank of New York (2015). Technical note concerning the methodology for calculating the effective federal funds rate. https://www.newyorkfed.org/medialibrary/media/markets/EFFR-technical-note-070815.pdf.

[Frahm et al., 2005] Frahm, G., Junker, M., and Schmidt, R. (2005). Estimating the tail-dependence coefficient: properties and pitfalls. *Insurance: mathematics and Economics*, 37(1):80–100. doi:10.1016/j.insmatheco.2005.05.008

[FTSE Digital Asset Research, 2022] FTSE Digital Asset Research (2022). Guide to the calculation of the FTSE DAR Digital Asset Prices and FTSE DAR Reference Prices. https://research.ftserussell.com/products/downloads/Guide_to_the_Calculation_of_FTSE_DAR_Digital_Asset_Prices_and_Reference_Prices_Fixes.pdf.

[Gallagher, 2018] Gallagher, C. (2018). CFIX methodology, Bloomberg cryptocurrency solutions. https://data.bloomberglp.com/professional/sites/10/CFIX-Methodology.pdf.

[Garcin and Nicolas, 2021] Garcin, M. and Nicolas, M. L. (2021). Nonparametric estimator of the tail dependence coefficient: balancing bias and variance. *arXiv preprint arXiv:2111.11128*.

[Glynn et al., 1996] Glynn, P. W. et al. (1996). Importance sampling for monte carlo estimation of quantiles. In *Mathematical methods in stochastic simulation and experimental design: Proceedings of the 2nd st. petersburg workshop on simulation*, pages 180–185. Citeseer.

[Gobet et al., 2022] Gobet, E., Lerasle, M., and Métivier, D. (2022). Mean estimation for randomized Quasi Monte Carlo method. *HAL preprint, https://hal.archives-ouvertes.fr/hal-03631879/document*.

[Hall and Heyde, 2014] Hall, P. and Heyde, C. C. (2014). *Martingale limit theory and its application*. Academic press.

[Hill, 1975] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

[Hyndman and Fan, 1996] Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365.

[Joe, 1997] Joe, H. (1997). *Multivariate models and dependence concepts*, volume 73 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London. doi:10.1201/b13150

[Krasnoselskii and Rutickii, 1961] Krasnoselskii, M. and Rutickii, Y. (1961). *Convex functions and Orlicz spaces*. Translated from the first Russian edition by Leo F. Boron. P. Noordhoff Ltd., Groningen.

[Kuznetzov and Mohri, 2017] Kuznetzov, V. and Mohri, M. (2017). Generalization bounds for non-stationary mixing processes. *Machine Learning*, 06(1):93–117.

[Lerasle, 2009] Lerasle, M. (2009). Adaptive density estimation of stationary $\beta$-mixing and $\tau$-mixing processes. *Mathematical Methods of statistics*, 18(1):59–83.

[Lugosi and Mendelson, 2019] Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190.

[Manjunath et al., 2013] Manjunath, B., Caeiro, F., Gomes, M., and Alves, M. (2013). evt0: Mean of order p, peaks over random threshold hill and high quantile estimates. *R package version*, pages 1–1.

[McNeil and Nešlehová, 2009] McNeil, A. and Nešlehová, J. (2009). Multivariate Archimedean copulas, $d$-monotone functions and $l_1$-norm symmetric distributions. *The Annals of Statistics*, 37(5B):3059–3097.

[McNeil et al., 2005] McNeil, A. J., Frey, R., and Embrechts, P. (2005). *Quantitative risk management*. Princeton Series in Finance. Princeton University Press, Princeton, NJ. Concepts, techniques and tools.

[Nasdaq, 2022] Nasdaq (2022). Nasdaq Crypto Index: index methodology. https://indexes.nasdaqomx.com/docs/methodology_NCI.pdf.

[Nelsen, 2006] Nelsen, R. (2006). *An introduction to copulas*. Springer Series in Statistics. Springer, New York, second edition. doi:10.1007/s11229-005-3715-x

[Olver et al., 2010] Olver, F. W. J., Lozier, D. W., Boisvert, R. F., and Clark, C. W., editors (2010). *NIST handbook of mathematical functions*. U.S. Department of Commerce, National Institute of Standards and Technology, Washington, DC; Cambridge University Press, Cambridge.

[Paine and Knottenbelt, 2016] Paine, A. and Knottenbelt, W. J. (2016). Analysis of the CME CF Bitcoin Reference Rate and Real Time Index. https://www.cmegroup.com/trading/files/bitcoin-white-paper.pdf.

[Reiss, 1989] Reiss, R.-D. (1989). *Approximate distributions of order statistics*. Springer Series in Statistics. Springer-Verlag, New York. With applications to nonparametric statistics. doi:10.1007/978-1-4613-9620-8

[Ren and Mojirsheibani, 2010] Ren, Q. and Mojirsheibani, M. (2010). A note on nonparametric regression with $\beta$-mixing sequences. *Communications in Statistics-Theory and Methods*, 39(12):2280–2287.

[Resnick, 2007] Resnick, S. I. (2007). *Heavy-tail phenomena*. Springer Series in Operations Research and Financial Engineering. Springer, New York. Probabilistic and statistical modeling.

[Rio, 2017] Rio, E. (2017). About the constants in the Fuk-Nagaev inequalities. *Electronic Communications in Probability*, 22:1–12.

[Ronchetti and Huber, 2009] Ronchetti, E. M. and Huber, P. J. (2009). *Robust statistics*. John Wiley & Sons.

[van de Geer and Lederer, 2013] van de Geer, S. and Lederer, J. (2013). The Bernstein-Orlicz norm and deviation inequalities. *Probability Theory and Related Fields*, 157(1-2):225–250.

[Vinter, 2021] Vinter (2021). Crypto reference rates for single assets, version 2.1. https://methodology.vinter.co/vinter/reference-rates.

# A Proofs of main theoretical results

## A.1 General identities on weighted quantiles and weighted mean

By using the properties of generalized inverse [Embrechts and Hofert, 2013, Proposition 2.3 - item (5)] applied to (2.5)-(2.7)-(2.8), and because each $W_i$ is positive, for any $y$ we have

$$\begin{aligned}\left\{\widehat{q_{\mathtt{W},n}}(\alpha) \le y\right\} &= \left\{\alpha \le \widehat{F_{\mathtt{W},\mathtt{n}}}(y)\right\} \\ &= \left\{\sum_{i=1}^{n} \alpha W_i \le \sum_{i=1}^{n} W_i \mathbb{1}_{\{P_i \le y\}}\right\} \\ &= \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(W_i(\alpha - \mathbb{1}_{\{P_i\le y\}}) - \mathbb{E}\left[W(\alpha - \mathbb{1}_{\{P\le y\}})\right]\right) \le -\sqrt{n}\mathbb{E}\left[W(\alpha - \mathbb{1}_{\{P\le y\}})\right]\right\}. \qquad \text{(A.1)}\end{aligned}$$

Taking the complement gives

$$\{\widehat{q_{\mathtt{W},n}}(\alpha) > y\} = \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( W_i(\alpha - \mathbb{1}_{\{P_i \le y\}}) - \mathbb{E}\left[W(\alpha - \mathbb{1}_{\{P \le y\}})\right] \right) > -\sqrt{n}\mathbb{E}\left[W(\alpha - \mathbb{1}_{\{P \le y\}})\right] \right\}. \tag{A.2}$$

The following lemma gives an estimate of the term $\sqrt{n}\mathbb{E}\left[W(\alpha - \mathbb{1}_{\{P \le y\}})\right]$ for $y$ close to the weighted quantile $q_{\mathtt{W}}(\alpha)$.

**Lemma A.1.** *Assume (**H0**). Set $y = q_{\mathtt{W}}(\alpha) + \frac{x}{\sqrt{n}}$ for $x \in \mathbb{R}$. Then*

$$-\sqrt{n}\mathbb{E}\left[W(\alpha - \mathbb{1}_{\{P \le y\}})\right] = x\,\mathbb{E}[W]\, f_{\mathtt{W}}\left(\left[q_{\mathtt{W}}(\alpha), \frac{x}{\sqrt{n}}\right]\right),$$
$$\textit{with} \qquad f_{\mathtt{W}}\left(\left[q_{\mathtt{W}}(\alpha), \frac{x}{\sqrt{n}}\right]\right) := \int_0^1 f_{\mathtt{W}}\left(q_{\mathtt{W}}(\alpha) + u\frac{x}{\sqrt{n}}\right) \mathrm{d}u.$$

*Proof.* First, since $F_{\mathtt{W}}$ is assumed to be continuous (density assumption (**H0**)), we have $\alpha = F_{\mathtt{W}}(q_{\mathtt{W}}(\alpha))$. Then, using the density $f_{\mathtt{W}}$ for writing expectations as in (2.2), we derive

$$\begin{aligned}\mathbb{E}\left[W(\alpha - \mathbb{1}_{\{P \le y\}})\right] &= \mathbb{E}[W] \int_0^{q_{\mathtt{W}}(\alpha)} f_{\mathtt{W}}(x')\,\mathrm{d}x' - \mathbb{E}[W] \int_0^{q_{\mathtt{W}}(\alpha)+\frac{x}{\sqrt{n}}} f_{\mathtt{W}}(x')\,\mathrm{d}x' \\ &= -\mathbb{E}[W]\,\frac{x}{\sqrt{n}} \int_0^1 f_{\mathtt{W}}\left(q_{\mathtt{W}}(\alpha) + u\frac{x}{\sqrt{n}}\right) \mathrm{d}u.\end{aligned}$$

□

Similarly to the weighted quantile, for (2.6) we write

$$\begin{aligned}\{\widehat{\mathtt{WAP}}_n \le y\} &= \left\{ \sum_{i=1}^{n} W_i P_i \le y \sum_{i=1}^{n} W_i \right\} \\ &= \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( W_i(y - P_i) - \mathbb{E}\left[W(y-P)\right] \right) \ge -\sqrt{n}\mathbb{E}\left[W(y-P)\right] \right\}.\end{aligned} \tag{A.3}$$

Taking $y = \mathtt{WAP} - \frac{x}{\sqrt{n}}$ where $\mathtt{WAP} = \frac{\mathbb{E}[W\cdot P]}{\mathbb{E}[W]}$ (see (2.3)), we get

$$-\sqrt{n}\mathbb{E}\left[W(y-P)\right] = x\mathbb{E}[W], \tag{A.4}$$

and using (2.12), we have

$$\mathscr{W}_n^{\le}(x) = \mathbb{P}\left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( W_i(\mathtt{WAP} - \frac{x}{\sqrt{n}} - P_i) - \mathbb{E}\left[ W(\mathtt{WAP} - \frac{x}{\sqrt{n}} - P) \right] \right) \ge x\mathbb{E}[W] \right). \tag{A.5}$$

We can proceed similarly with $\left\{\widehat{\mathtt{WAP}}_n \ge \mathtt{WAP} + \frac{x}{\sqrt{n}}\right\}$, and we get

$$\mathscr{W}_n^{\ge}(x) = \mathbb{P}\left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( W_i(P_i - \mathtt{WAP} - \frac{x}{\sqrt{n}}) - \mathbb{E}\left[ W(P - \mathtt{WAP} - \frac{x}{\sqrt{n}}) \right] \right) \ge x\mathbb{E}[W] \right). \tag{A.6}$$

## A.2 Proof of Theorem 2.2

**Proof of (2.9).** Starting from (A.1) with $y = q_{\mathtt{W}}(\alpha) + \frac{x}{\sqrt{n}}$ for $x \in \mathbb{R}$, we get

$$\left\{\sqrt{n}\left(\widehat{q_{\mathtt{W},n}}(\alpha) - q_{\mathtt{W}}(\alpha)\right) \le x\right\} = \left\{\frac{1}{\sigma_n\sqrt{n}}\sum_{i=1}^{n}\left(Z_i^n - \mathbb{E}\left[Z^n\right]\right) \le -\frac{\sqrt{n}}{\sigma_n}\mathbb{E}\left[Z^n\right]\right\}, \tag{A.7}$$

where

$$Z_i^n := W_i(\alpha - \mathbb{1}_{\{P_i \le q_{\mathtt{W}}(\alpha) + \frac{x}{\sqrt{n}}\}}), \qquad \sigma_n := \sqrt{\mathbb{V}\mathrm{ar}\left[Z^n\right]}, \qquad Z^n := W(\alpha - \mathbb{1}_{\{P \le q_{\mathtt{W}}(\alpha) + \frac{x}{\sqrt{n}}\}}).$$

By Lemma A.1 and using the continuity of $f_{\mathtt{W}}(\cdot)$ at $q_{\mathtt{W}}(\alpha)$, we have

$$-\sqrt{n}\mathbb{E}\left[Z^n\right] = x\mathbb{E}\left[W\right] f_{\mathtt{W}}(q_{\mathtt{W}}(\alpha)) + o(1)$$

as $n \to +\infty$. In addition, the dominated convergence theorem yields

$$\mathbb{E}\left[\left(Z^n\right)^2\right] \to \mathbb{E}\left[W^2\left(\alpha - \mathbb{1}_{\{P \le q_{\mathtt{W}}(\alpha)\}}\right)^2\right].$$

Since $\mathbb{E}\left[Z^n\right] \to 0$, we deduce $\sigma_n \to \sigma := \sqrt{\mathbb{E}\left[W^2\left(\alpha - \mathbb{1}_{\{P \le q_{\mathtt{W}}(\alpha)\}}\right)^2\right]}$ as $n \to \infty$. Observe that $\sigma \neq 0$. We now claim that the renormalised sum in (A.7) converges weakly to a standard Gaussian random variable $\mathcal{N}(0,1)$: indeed, this is a normalized sum of independent, centered and unit variance random variables, and since $|Z^n| \le W$, we obtain the Lindeberg condition, i.e. for all $\varepsilon > 0$

$$\mathbb{E}\left[\left|Z^n - \mathbb{E}\left[Z^n\right]\right|^2 \mathbb{1}_{\{|Z^n - \mathbb{E}[Z^n]| > \varepsilon\sqrt{n}\}}\right] \le \mathbb{E}\left[2\left(W^2 + \mathbb{E}\left[W\right]^2\right)\mathbb{1}_{\{W + \mathbb{E}[W] > \varepsilon\sqrt{n}\}}\right] \underset{n\to+\infty}{\to} 0,$$

which is a consequence of the dominated convergence theorem and of $(\mathbf{H}_{\mathbf{X}}^{\kappa})$ with $X = W$ and $\kappa = 2$. Thus the Central Limit Theorem for triangular arrays of random variables under the Lindeberg condition applies ([Bhattacharya and Rao, 2010, Corollary 18.2]). All in all, we have proved

$$\mathbb{P}\left(\sqrt{n}\left(\widehat{q_{\mathtt{W},n}}(\alpha) - q_{\mathtt{W}}(\alpha)\right) \le x\right) \to \mathbb{P}\left(\mathcal{N}(0,1) \le \frac{x\mathbb{E}\left[W\right] f_{\mathtt{W}}(q_{\mathtt{W}}(\alpha))}{\sigma}\right),$$

which leads to the announced result. □

**Proof of (2.10).** Starting from (A.3) with $y = \mathtt{WAP} + \frac{x}{\sqrt{n}}$ and using (A.4), we get

$$\mathbb{P}\left(\sqrt{n}(\widehat{\mathtt{WAP}}_n - \mathtt{WAP}) \le x\right) = \mathbb{P}\left(\frac{1}{\sigma_n\sqrt{n}}\sum_{i=1}^{n}\left(Z_i^n - \mathbb{E}\left[Z^n\right]\right) \ge -\frac{x\mathbb{E}\left[W\right]}{\sigma_n}\right),$$

where, here, we set

$$Z_i^n := W_i(\mathtt{WAP} + \frac{x}{\sqrt{n}} - P_i), \qquad \sigma_n := \sqrt{\mathbb{V}\mathrm{ar}\left[Z^n\right]}, \qquad Z^n := W(\mathtt{WAP} + \frac{x}{\sqrt{n}} - P).$$

As for the previous CLT proof, it is easy to check that, under the standing assumptions ($\mathbf{H}^{\kappa}_{\mathbf{X}}$) with $X = W \cdot P$ and $X = W$ and $\kappa = 2$),

$$\sigma_n \to \sigma := \sqrt{\mathbb{E}\left[W^2(\mathtt{WAP} - P)^2\right]},$$
$$|Z^n| \le W(\mathtt{WAP} + x + P) =: \overline{Z} \in L_2,$$
$$\mathbb{E}\left[\left|Z^n - \mathbb{E}\left[Z^n\right]\right|^2 \mathbb{1}_{\{|Z^n - \mathbb{E}[Z^n]| > \varepsilon\sqrt{n}\}}\right] \le \mathbb{E}\left[2\left(\overline{Z}^2 + \mathbb{E}\left[\overline{Z}\right]^2\right)\mathbb{1}_{\{\overline{Z} + \mathbb{E}\left[\overline{Z}\right] > \varepsilon\sqrt{n}\}}\right] \underset{n\to+\infty}{\to} 0,$$

for all $\varepsilon > 0$. This leads to

$$\mathbb{P}\left(\sqrt{n}(\widehat{\mathtt{WAP}}_n - \mathtt{WAP}) \le x\right) \to \mathbb{P}\left(\mathcal{N}(0,1) \ge -\frac{x\mathbb{E}[W]}{\sigma}\right),$$

and the convergence (2.10) readily follows. □

## A.3 Proof of Theorem 2.3

**Proof of (2.13) and (2.14).** We start with (2.13). We apply the Fuk-Nagaev inequality [Rio, 2017], which states that for any set of i.i.d. centered random variables $X_i$'s, we have

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \ge \lambda\right) \le \left(\frac{\kappa+2}{\kappa}\right)^{\kappa} \frac{\mathbb{E}\left[X_+^{\kappa}\right]}{n^{\frac{\kappa}{2}-1}\lambda^{\kappa}} + \exp\left(-\frac{2\lambda^2}{(\kappa+2)^2 e^{\kappa}\mathbb{V}\mathrm{ar}[X]}\right), \qquad \lambda > 0, \tag{A.8}$$

provided that $\kappa > 2$ and[5] $\mathbb{E}\left[X_+^{\kappa}\right] < +\infty$. We recall, for the sake of completeness, that the above inequality is more accurate than the Bahr-Esseen inequality [Bahr and Esseen, 1965, Section 5] or the Markov inequality combined with the Burkholder inequality [Hall and Heyde, 2014, Th. 2.10].
Setting $X = W(\alpha - \mathbb{1}_{\{P \le y\}}) - \mathbb{E}\left[W(\alpha - \mathbb{1}_{\{P \le y\}})\right]$ and applying (A.8) to it. First using elementary computations, we obtain the following crude upper bounds

$$\mathbb{V}\mathrm{ar}[X] \le [\alpha \vee (1-\alpha)]^2 \mathbb{E}\left[W^2\right] < \infty \qquad \text{and} \qquad \mathbb{E}\left[X_+^{\kappa}\right] \le (2[\alpha \vee (1-\alpha)])^{\kappa}\mathbb{E}\left[W^{\kappa}\right] < \infty,$$

assuming ($\mathbf{H}^{\kappa}_{\mathbf{W}}$). Combining (A.2) (where $y = q_{\mathtt{W}}(\alpha) + \frac{x}{\sqrt{n}}$, $x > 0$) and Lemma A.1, observe now that $\mathcal{Q}_n^{>}(\alpha, x)$ in (2.11) is bounded by the right-hand side of (A.8) with $\lambda = x\mathbb{E}[W]\, f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha), \frac{x}{\sqrt{n}}])$ and the above bounds on $\mathbb{V}\mathrm{ar}[X]$ and $\mathbb{E}\left[X_+^{\kappa}\right]$: indeed, this follows from (A.2) (where $y = q_{\mathtt{W}}(\alpha) + \frac{x}{\sqrt{n}}$, $x > 0$) and Lemma A.1. Thus, we obtain the claimed inequality (2.13).
The bound $\mathcal{Q}_n^{\le}(\alpha, x) = \mathbb{P}\left(\widehat{q_{\mathtt{W},n}}(\alpha) - q_{\mathtt{W}}(\alpha) \le -\frac{x}{\sqrt{n}}\right)$ works similarly, using $y = q_{\mathtt{W}}(\alpha) - \frac{x}{\sqrt{n}}$, we skip details. □

**Proof of (2.15) and (2.16).** In view of (A.6), (2.15) is a direct application of the Fuk-Nagaev inequality (A.8) with

$$X_i = W_i(P_i - \mathtt{WAP} - \frac{x}{\sqrt{n}}) - \mathbb{E}\left[W(P - \mathtt{WAP} - \frac{x}{\sqrt{n}})\right],$$

$\lambda = x\mathbb{E}[W]$, and using $\mathbb{V}\mathrm{ar}[X] \le \mathbb{E}\left[W^2(P - \mathtt{WAP} - \frac{x}{\sqrt{n}})^2\right] < \infty$, $\mathbb{E}\left[X_+^{\kappa}\right] \le 2^{\kappa}\mathbb{E}\left[W^{\kappa}|P - \mathtt{WAP} - \frac{x}{\sqrt{n}}|^{\kappa}\right] < \infty$ assuming ($\mathbf{H}^{\kappa}_{\mathbf{W}}$) and ($\mathbf{H}^{\kappa}_{\mathbf{W}\cdot\mathbf{P}}$).

[5] here, $x_+$ stands for the positive part of $x$.

The derivation of (2.16) from (A.5) is similar. Details are left to the reader. □

## A.4 Proof of Theorem 2.4

**Proof of (2.17) and (2.18).** It is an easy exercice to check that $(\mathbf{H}_X^{\Gamma})$ for $X = W$ implies that $\|W\|_{\Psi_L}$ is finite for any $L > 0$. The strategy of proof consists in deriving concentration inequalities under sub-gamma assumptions for the renormalized sum

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( W_i(\alpha - \mathbb{1}_{\{P_i \le y\}}) - \mathbb{E}\left[ W(\alpha - \mathbb{1}_{\{P \le y\}}) \right] \right),$$

in order to quantity the probability of the events (A.1) and (A.2).
Observe that we can not directly use results of [van de Geer and Lederer, 2013] because they are stated under the assumption that

$$Z := W(\alpha - \mathbb{1}_{\{P \le y\}}) - \mathbb{E}\left[ W(\alpha - \mathbb{1}_{\{P \le y\}}) \right]$$

has a finite Bernstein-Orlicz norm and concentration bounds are expressed in terms of $\|Z\|_{\Psi_L}$. On the one hand, the latter quantity depends on $P$ and $y$. On the other hand, our goal is to express results using $\|W\|_{\Psi_L}$ only. Obviously,

$$\left\| W(\alpha - \mathbb{1}_{\{P \le y\}}) \right\|_{\Psi_L} \le [\alpha \vee (1-\alpha)] \, \|W\|_{\Psi_L} \tag{A.9}$$

just using that $|\alpha - \mathbb{1}_{\{P \le y\}}| \le [\alpha \vee (1-\alpha)]$, which means that bounding works well with Orlicz norm while keeping the same bounds on norms (up to constants). But, the same bounding does not work well when dealing with the centering term $\mathbb{E}\left[ W(\alpha - \mathbb{1}_{\{P \le y\}}) \right]$. This explains why our way of proof follows slightly different arguments. The objective of the proof is to derive the Bernstein inequality (for non-centered random variable) to

$$W_{\alpha,y,P} := W(\alpha - \mathbb{1}_{\{P \le y\}}), \tag{A.10}$$

by first showing that $W_{\alpha,y,P}$ satisfies the required conditions on $\mathbb{E}\left[ |W_{\alpha,y,P}|^m \right]$, for $m \in \{2,3,\dots\}$ (see [Boucheron et al., 2013, Theorem 2.10, p.37]). In view of (A.9), set $\tau := [\alpha \vee (1-\alpha)] \, \|W\|_{\Psi_L} \ge \left\| W_{\alpha,y,P} \right\|_{\Psi_L}$. Invoking [van de Geer and Lederer, 2013, Lemma 1], we get

$$\mathbb{P}\left( |W_{\alpha,y,P}| > \tau \left( \sqrt{t} + \frac{Lt}{2} \right) \right) \le 2e^{-t} \qquad \forall\, t \ge 0. \tag{A.11}$$

We claim that

$$\mathbb{E}\left[ |W_{\alpha,y,P}|^m \right] \le m \tau^m 2^{m-1} \left( \Gamma(m/2) + L^m \Gamma(m) \right), \qquad \forall\, m \ge 1, \tag{A.12}$$

where $\Gamma(.)$ is the usual gamma function. The proof of (A.12) is given at the end. This shows that $W_{\alpha,y,P}$ satisfies conditions of the Bernstein inequality: for any $m \in \{2,3,\dots\}$,

$$\begin{aligned} \mathbb{E}\left[ |W_{\alpha,y,P}|^m \right] &\le m \tau^m 2^{m-1} \Gamma(m) \left( 1 + L^m \right) \\ &\le \frac{1}{2} m! \left( 2\tau(1+L) \right)^m, \end{aligned}$$

using that $\Gamma(.)$ is increasing [6] on $[3/2,+\infty)$, satisfies $\Gamma(m)=(m-1)!$ for $m\in\mathbb{N}^*$. Therefore, for any $t\geq 0$, applying Bernstein's inequality we get

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(W_i(\alpha-\mathbb{1}_{\{P_i\leq y\}})-\mathbb{E}\left[W(\alpha-\mathbb{1}_{\{P\leq y\}})\right]\right)\geq 2\tau(1+L)\left(\sqrt{2t}+\frac{t}{\sqrt{n}}\right)\right)\leq e^{-t}, \tag{A.13}$$

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(W_i(\alpha-\mathbb{1}_{\{P_i\leq y\}})-\mathbb{E}\left[W(\alpha-\mathbb{1}_{\{P\leq y\}})\right]\right)\leq -2\tau(1+L)\left(\sqrt{2t}+\frac{t}{\sqrt{n}}\right)\right)\leq e^{-t}. \tag{A.14}$$

It is easy to check the following relation between non-negative $t$ and $\zeta$:

$$\zeta=c\left(\sqrt{2t}+\frac{t}{\sqrt{n}}\right)\Longleftrightarrow t=\frac{(\zeta/c)^2}{\left(\sqrt{\frac{\zeta}{c\sqrt{n}}+\frac{1}{2}}+\frac{1}{\sqrt{2}}\right)^2}. \tag{A.15}$$

Invoking (A.2) with $y=q_{\mathtt{W}}(\alpha)+\frac{x}{\sqrt{n}}$ (for $x>0$), setting $\zeta:=-\sqrt{n}\mathbb{E}\left[W(\alpha-\mathbb{1}_{\{P\leq y\}})\right]=x\mathbb{E}[W]\,f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha),\frac{x}{\sqrt{n}}])$ (Lemma A.1) and combining this with (A.13) and (A.15) (using $c=2\tau(1+L)$ and $\tau=[\alpha\vee(1-\alpha)]\,\|W\|_{\Psi_L}$) readily leads to the announced bound (2.17).
Proving (2.18) is similar by using (A.14) and (A.1) for $y=q_{\mathtt{W}}(\alpha)-\frac{x}{\sqrt{n}}$ (for $x>0$). □

**Proof of** (A.12)**.** Take $m\geq 1$:

$$\mathbb{E}\left[|W_{\alpha,y,P}|^m\right]=m\int_0^{+\infty}x^{m-1}\mathbb{P}\left(|W_{\alpha,y,P}|>x\right)\mathrm{d}x$$

(set $x=\tau(\sqrt{t}+\frac{L}{2}t)$ and use (A.11))

$$\leq m\int_0^{+\infty}\tau^m(\sqrt{t}+\frac{L}{2}t)^{m-1}\left(\frac{1}{2\sqrt{t}}+\frac{L}{2}\right)2e^{-t}\,\mathrm{d}t$$
$$\leq m\tau^m\int_0^{+\infty}(\sqrt{t}+Lt)^m\frac{e^{-t}}{t}\,\mathrm{d}t$$

(use $(a+b)^m\leq 2^{m-1}(a^m+b^m)$ and integrate using the $\Gamma(.)$-definition)

$$\leq m\tau^m 2^{m-1}\left(\Gamma(m/2)+L^m\Gamma(m)\right).$$

□

**Proof of** (2.19)**.** Let us begin with $\mathscr{W}_n^{\geq}(x)$. In view of (A.6), set $X_i=W_i(P_i-\mathtt{WAP}-\frac{x}{\sqrt{n}})$: the triangle inequality on Bernstein-Orlicz norm yields $\|X_i\|_{\Psi_L}\leq\|W(P-\mathtt{WAP})\|_{\Psi_L}+\|W\|_{\Psi_L}\frac{x}{\sqrt{n}}$. Then, we proceed as for the proof of (2.17)-(2.18), by replacing

$$\tau=[\alpha\vee(1-\alpha)]\,\|W\|_{\Psi_L}\quad\text{ by }\quad\tau=\|W(P-\mathtt{WAP})\|_{\Psi_L}+\|W\|_{\Psi_L}\frac{x}{\sqrt{n}}$$

[6] indeed, its logarithmic derivative, the digamma function $\psi(.)$, is increasing and such that $\psi(3/2)=-\gamma-2\ln(2)+2\geq 0$ where $\gamma$ is the Euler constant, see [Olver et al., 2010, Chapter 5].

(as a norm upper bound for random variables at hand), and by replacing

$$\zeta = x\,\mathbb{E}[W]\, f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha), \frac{x}{\sqrt{n}}]) \quad \text{by} \quad \zeta = x\mathbb{E}[W]\,.$$

This leads to the bound (2.19) for $\mathscr{W}_n^{\geq}(x)$.
The proof of the bound for $\mathscr{W}_n^{\leq}(x)$ is analogous, by starting from (A.5) and setting $X_i = W_i(\mathtt{WAP} - \frac{x}{\sqrt{n}} - P_i)$ for which we still have $\|X_i\|_{\Psi_L} \leq \|W(P-\mathtt{WAP})\|_{\Psi_L} + \|W\|_{\Psi_L}\frac{x}{\sqrt{n}}$. Thus, the final bound for $\mathscr{W}_n^{\leq}(x)$ is the same as for that of $\mathscr{W}_n^{\geq}(x)$. □

## A.5 Proof of Theorem 2.5

**Proof of (2.20) and (2.21).** Observe that under the assumption ($\mathbf{H}_{\mathbf{X}}^{\mathbf{G}}$) with $X = W$, $\|W\|_{\Psi_G}$ is finite. In addition,

$$\left\| W(\alpha - \mathbb{1}_{\{P\leq y\}}) \right\|_{\Psi_G} \leq [\alpha \vee (1-\alpha)]\, \|W\|_{\Psi_G} =: \tau.$$

Using the notation (A.10) for $W_{\alpha,y,P}$, it implies that

$$1 \geq \mathbb{E}\left[e^{(W_{\alpha,y,P}/\tau)^2}\right] - 1 = \sum_{m\geq 1} \mathbb{E}\left[(W_{\alpha,y,P}/\tau)^{2m}\right]/m!,$$

hence

$$\mathbb{E}\left[W_{\alpha,y,P}^{2m}\right] \leq \tau^{2m} m!, \qquad \forall\, m \in \mathbb{N}^*.$$

The objective is 1) to derive an upper bound on the Laplace transform of $(W_{\alpha,y,P} - \mathbb{E}\left[W_{\alpha,y,P}\right])$ and 2) to apply Chernoff's inequality in order to get the desired deviation bound. Define $W'_{\alpha,y,P}$ an independent copy of $W_{\alpha,y,P}$: by Jensen inequality, we have

$$\mathbb{E}\left[e^{\lambda(W_{\alpha,y,P} - \mathbb{E}[W_{\alpha,y,P}])}\right] \leq \mathbb{E}\left[e^{\lambda(W_{\alpha,y,P} - W'_{\alpha,y,P})}\right], \qquad \forall \lambda \in \mathbb{R}.$$

The distribution of $W_{\alpha,y,P} - W'_{\alpha,y,P}$ is symmetric, thus all odd polynomial moments are zero. For even moments, use $(a+b)^m \leq 2^{m-1}(a^m + b^m)$ to get

$$\mathbb{E}\left[(W_{\alpha,y,P} - W'_{\alpha,y,P})^{2m}\right] \leq 2^{2m}\mathbb{E}\left[(W_{\alpha,y,P})^{2m}\right] \leq (2\tau)^{2m} m!.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(W_{\alpha,y,P} - \mathbb{E}[W_{\alpha,y,P}])}\right] &\leq 1 + \sum_{m\geq 1} \frac{\lambda^{2m}\mathbb{E}\left[(W_{\alpha,y,P} - W'_{\alpha,y,P})^{2m}\right]}{(2m)!} \\ &\leq 1 + \sum_{m\geq 1} \frac{(2\tau\lambda)^{2m} m!}{(2m)!}.\end{aligned}$$

Note that $\frac{(2m)!}{m!} = \prod_{i=1}^m (m+i) \geq \prod_{i=1}^m (2i) = 2^m m!$, thus

$$\mathbb{E}\left[e^{\lambda(W_{\alpha,y,P} - \mathbb{E}[W_{\alpha,y,P}])}\right] \leq \sum_{m\geq 0} \frac{(2\tau^2\lambda^2)^m}{m!} = e^{2\tau^2\lambda^2}.$$

Therefore, the Chernoff inequality [Boucheron et al., 2013, Section 2.3] gives, for any $\zeta \geq 0$,

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(W_i(\alpha - \mathbb{1}_{\{P_i \leq y\}}) - \mathbb{E}\left[W(\alpha - \mathbb{1}_{\{P \leq y\}})\right]\right) \geq \zeta\right) \leq e^{-\frac{\zeta^2}{8\tau^2}},$$
$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(W_i(\alpha - \mathbb{1}_{\{P_i \leq y\}}) - \mathbb{E}\left[W(\alpha - \mathbb{1}_{\{P \leq y\}})\right]\right) \leq -\zeta\right) \leq e^{-\frac{\zeta^2}{8\tau^2}}.$$

Recalling $\tau = [\alpha \vee (1-\alpha)]\,\|W\|_{\Psi_G}$, using (A.2) with $y = q_{\mathtt{W}}(\alpha) + \frac{x}{\sqrt{n}}$ (for $x > 0$) and setting $\zeta := -\sqrt{n}\mathbb{E}\left[W(\alpha - \mathbb{1}_{\{P \leq y\}})\right] = \mathbb{E}[W]\, x f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha), \frac{x}{\sqrt{n}}])$ (Lemma A.1) readily leads to the announced bound (2.20).
Proving (2.21) is similar by using (A.1) with $y = q_{\mathtt{W}}(\alpha) - \frac{x}{\sqrt{n}}$ (for $x > 0$). □

**Proof of (2.22).** We follow the arguments used before for establishing (2.20) and (2.21). Consider first the bound on $\mathscr{W}_n^{\geq}(x)$, which is expressed as (A.6). A quick inspection of the previous proof based on Chernoff inequality shows that it suffices to replace $\tau = [\alpha \vee (1-\alpha)]\,\|W\|_{\Psi_G}$ by $\tau = \|W(P - \mathtt{WAP})\|_{\Psi_G} + \|W\|_{\Psi_G}\frac{x}{\sqrt{n}}$, and $\zeta = \mathbb{E}[W]\, x f_{\mathtt{W}}([q_{\mathtt{W}}(\alpha), \frac{x}{\sqrt{n}}])$ by $\zeta = x\mathbb{E}[W]$. The bound for $\mathscr{W}_n^{\geq}(x)$ readily follows. The analysis for $\mathscr{W}_n^{\leq}(x)$ is similar and we skip it. □