Beyond Catoni: Sharper Rates for Heavy-Tailed and Robust Mean Estimation

Shivam Gupta (UT Austin),
Samuel B. Hopkins (MIT), Eric Price (UT Austin)

One-Dimensional Mean Estimation

- Given n samples $x_1, \ldots, x_n$ from a variance σ^2 distribution, would like to produce an estimate of the mean μ

One-Dimensional Mean Estimation

- Given n samples $x_1, \ldots, x_n$ from a variance σ^2 distribution, would like to produce an estimate of the mean μ
- When the distribution is **Gaussian**, the empirical mean is within $\sigma\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$ of the true mean μ with probability $1-\delta$. (Chernoff Bound)

One-Dimensional Mean Estimation

- Given n samples $x_1, \ldots, x_n$ from a variance σ^2 distribution, would like to produce an estimate of the mean μ
- When the distribution is **Gaussian**, the empirical mean is within $\sigma\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$ of the true mean μ with probability $1-\delta$. (Chernoff Bound)
- For the general case:

Estimator	Error
Empirical Mean	$\sigma\sqrt{\frac{1}{n\delta}}$
Median-of-means	$19.2\cdot\sigma\sqrt{\frac{\log\frac{1}{\delta}}{n}}$
Catoni (2012)	$\sigma\sqrt{\frac{2\log\frac{1}{\delta}}{n}}\cdot(1+o(1))$

Table: One-dimensional Estimators

d-Dimensional Heavy-Tailed Estimation

- In d-dimensional estimation, we are given iid samples $x_1, \ldots, x_n \in \mathbb{R}^d$, with $Cov(x_i) \preccurlyeq \sigma^2 I_d$ and want to compute an estimate of the mean μ.

d-Dimensional Heavy-Tailed Estimation

- In d-dimensional estimation, we are given iid samples $x_1, \ldots, x_n \in \mathbb{R}^d$, with $Cov(x_i) \preccurlyeq \sigma^2 I_d$ and want to compute an estimate of the mean μ.
- For simplicity, we will focus on the $\sigma = 1$ case.

d-Dimensional Heavy-Tailed Estimation

- In d-dimensional estimation, we are given iid samples $x_1, \ldots, x_n \in \mathbb{R}^d$, with $Cov(x_i) \preccurlyeq \sigma^2 I_d$ and want to compute an estimate of the mean μ.
- For simplicity, we will focus on the $\sigma = 1$ case.

Estimator	**Error**	**Notes**
Empirical Mean	$\sqrt{\frac{d}{n}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}$	Gaussian/Light-tailed distributions
Catoni (2012) + Net	$\sqrt{\frac{2d}{d+1}} \cdot \left(O\left(\sqrt{\frac{d}{n}}\right) + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right)$	Any distribution
Catoni (2012) + PAC-Bayes	$\sqrt{\frac{2d}{d+1}} \cdot \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right)$	Any distribution
Lee, Valiant (2022)	$\sqrt{\frac{d}{n}}$	Any distribution, when $d \gg \log^2 \frac{1}{\delta}$

Table: Prior Estimators

d-Dimensional Heavy-Tailed Estimation

- In d-dimensional estimation, we are given iid samples $x_1, \ldots, x_n \in \mathbb{R}^d$, with $Cov(x_i) \preccurlyeq \sigma^2 I_d$ and want to compute an estimate of the mean μ.
- For simplicity, we will focus on the $\sigma = 1$ case.

Estimator	**Error**	**Notes**
Empirical Mean	$\sqrt{\frac{d}{n}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}$	Gaussian/Light-tailed distributions
Catoni (2012) + Net	$\sqrt{\frac{2d}{d+1}} \cdot \left(O\left(\sqrt{\frac{d}{n}}\right) + \sqrt{\frac{2\log\frac{1}{\delta}}{n}} \right)$	Any distribution
Catoni (2012) + PAC-Bayes	$\sqrt{\frac{2d}{d+1}} \cdot \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}} \right)$	Any distribution
Lee, Valiant (2022)	$\sqrt{\frac{d}{n}}$	Any distribution, when $d \gg \log^2 \frac{1}{\delta}$

Table: Prior Estimators

- When $\log \frac{1}{\delta} \gg d$, is the $\sqrt{\frac{2d}{d+1}}$-factor loss over the Gaussian rate necessary?

d-Dimensional Heavy-Tailed Estimation

- In d-dimensional estimation, we are given iid samples $x_1, \ldots, x_n \in \mathbb{R}^d$, with $Cov(x_i) \preccurlyeq \sigma^2 I_d$ and want to compute an estimate of the mean μ.
- For simplicity, we will focus on the $\sigma = 1$ case.

Estimator	**Error**	**Notes**
Empirical Mean	$\sqrt{\frac{d}{n}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}$	Gaussian/Light-tailed distributions
Catoni (2012) + Net	$\sqrt{\frac{2d}{d+1}} \cdot \left(O\left(\sqrt{\frac{d}{n}}\right) + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right)$	Any distribution
Catoni (2012) + PAC-Bayes	$\sqrt{\frac{2d}{d+1}} \cdot \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right)$	Any distribution
Lee, Valiant (2022)	$\sqrt{\frac{d}{n}}$	Any distribution, when $d \gg \log^2\frac{1}{\delta}$

Table: Prior Estimators

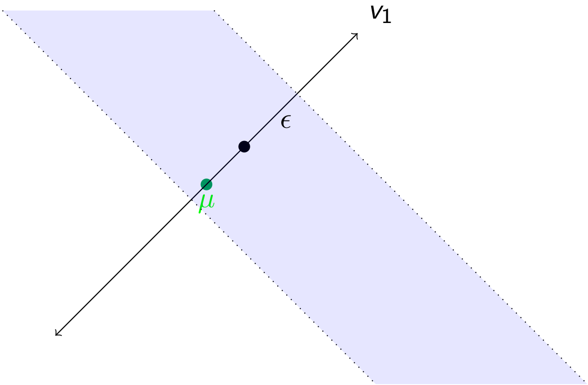
- When $\log\frac{1}{\delta} \gg d$, is the $\sqrt{\frac{2d}{d+1}}$-factor loss over the Gaussian rate necessary?
- We show that the answer is **no** – we show an estimator with error $(1-\tau) \cdot \sqrt{\frac{2d}{d+1}} \cdot \left(\sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right)$ for a small constant $\tau > 0$.

Catoni + Net argument

- Suppose, we have an estimate $\widehat{\mu}_v = \langle \mu, v \rangle \pm \epsilon$ for every unit vector v.

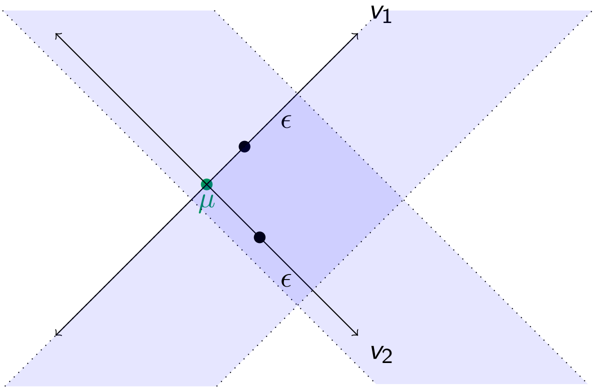
Catoni + Net argument

- Suppose, we have an estimate $\widehat{\mu}_v = \langle \mu, v \rangle \pm \epsilon$ for every unit vector v.



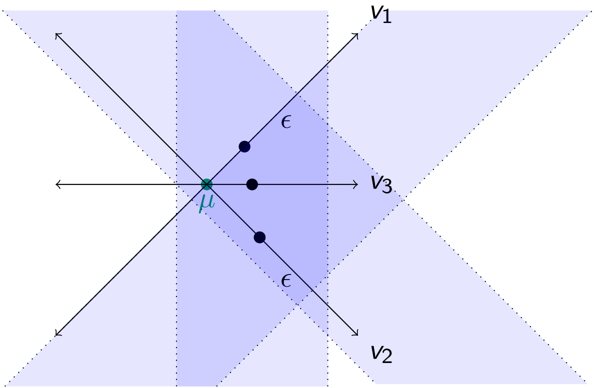
Catoni + Net argument

- Suppose, we have an estimate $\widehat{\mu}_v = \langle \mu, v \rangle \pm \epsilon$ for every unit vector v.



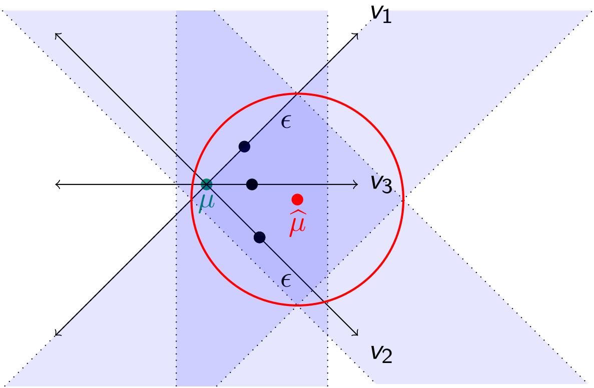
Catoni + Net argument

- Suppose, we have an estimate $\widehat{\mu}_v = \langle \mu, v \rangle \pm \epsilon$ for every unit vector v.



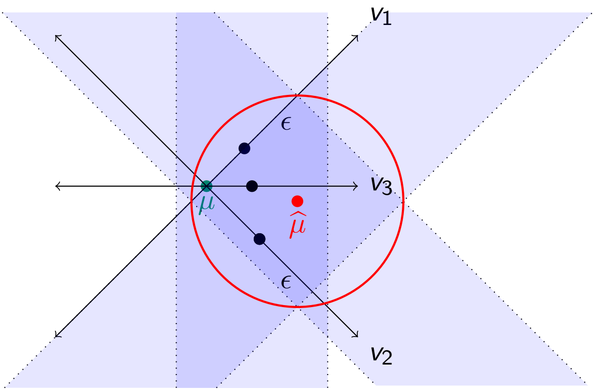
Catoni + Net argument

- Suppose, we have an estimate $\widehat{\mu}_v = \langle \mu, v \rangle \pm \epsilon$ for every unit vector v.



- Taking the center of the enclosing sphere of these confidence regions is guaranteed to be within $\sqrt{\frac{2d}{d+1}} \cdot \epsilon$ of μ in ℓ_2 (Jung's theorem).

d-Dimensional Heavy-Tailed Estimation



- Using a net argument with Catoni's estimate in each direction v, combined with this argument produces $\widehat{\mu}$ with

$$\|\widehat{\mu} - \mu\| \leq \sqrt{\frac{2d}{d+1}} \cdot \sigma \left(O\left(\sqrt{\frac{d}{n}} \right) + \sqrt{\frac{2 \log \frac{1}{\delta}}{n}} \right)$$

Variant of Catoni's One-dimensional Estimator

1. Let μ_0 be estimate via Median-of-means on fraction of the n samples

Variant of Catoni's One-dimensional Estimator

1. Let μ_0 be estimate via Median-of-means on fraction of the n samples
2. With $T = \sqrt{\frac{n}{2 \log \frac{1}{\delta}}}$, refine the estimate:

$$\widehat{\mu} = \mu_0 + \frac{1}{n} \sum_{i=1}^{n} T \psi \left(\frac{x_i - \mu_0}{T} \right)$$

Variant of Catoni's One-dimensional Estimator

1. Let μ_0 be estimate via Median-of-means on fraction of the n samples
2. With $T = \sqrt{\frac{n}{2 \log \frac{1}{\delta}}}$, refine the estimate:

$$\widehat{\mu} = \mu_0 + \frac{1}{n} \sum_{i=1}^{n} T \psi \left(\frac{x_i - \mu_0}{T} \right)$$

- When $\psi(x) = x$, then the estimate $\widehat{\mu}$ is just the empirical mean.

Variant of Catoni's One-dimensional Estimator

1. Let μ_0 be estimate via Median-of-means on fraction of the n samples
2. With $T = \sqrt{\frac{n}{2\log\frac{1}{\delta}}}$, refine the estimate:

$$\widehat{\mu} = \mu_0 + \frac{1}{n}\sum_{i=1}^{n} T\psi\left(\frac{x_i - \mu_0}{T}\right)$$

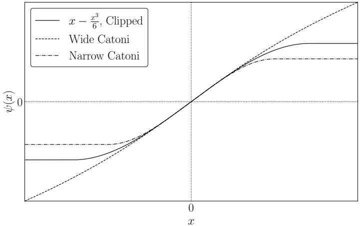
- When $\psi(x) = x$, then the estimate $\widehat{\mu}$ is just the empirical mean.
- Catoni prescribes a specific way of selecting ψ to downweight outliers that achieves the optimal $\sigma\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$ error.

Variant of Catoni's One-dimensional Estimator

1. Let μ_0 be estimate via Median-of-means on fraction of the n samples
2. With $T = \sqrt{\frac{n}{2\log\frac{1}{\delta}}}$, refine the estimate:

$$\widehat{\mu} = \mu_0 + \frac{1}{n}\sum_{i=1}^{n} T\psi\left(\frac{x_i - \mu_0}{T}\right)$$

- When $\psi(x) = x$, then the estimate $\widehat{\mu}$ is just the empirical mean.
- Catoni prescribes a specific way of selecting ψ to downweight outliers that achieves the optimal $\sigma\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$ error.
- ψ satisfies: $-\log\left(1 - x + \frac{x^2}{2}\right) \leq \psi(x) \leq \log\left(1 + x + \frac{x^2}{2}\right)$



Variant of Catoni's One-dimensional Estimator

- Estimate is given by

$$\widehat{\mu} = \mu_0 + r(\mu_0) := \mu_0 + \frac{1}{n}\sum_{i=1}^{n} T\psi\left(\frac{x_i - \mu_0}{T}\right)$$

Variant of Catoni's One-dimensional Estimator

- Estimate is given by

$$\widehat{\mu} = \mu_0 + r(\mu_0) := \mu_0 + \frac{1}{n}\sum_{i=1}^{n} T\psi\left(\frac{x_i - \mu_0}{T}\right)$$

- **Claim:** $r(\mu_0)$ is $(\mu - \mu_0) \pm \sigma\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$ with probability $1 - \delta$.

Variant of Catoni's One-dimensional Estimator

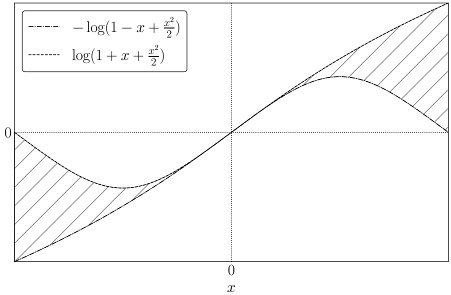
- Estimate is given by

$$\widehat{\mu} = \mu_0 + r(\mu_0) := \mu_0 + \frac{1}{n}\sum_{i=1}^{n} T\psi\left(\frac{x_i - \mu_0}{T}\right)$$

- **Claim:** $r(\mu_0)$ is $(\mu - \mu_0) \pm \sigma\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$ with probability $1-\delta$.
- **Proof:** Its MGF is given by

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{n}{T}r(\mu_0)\right)\right] &= \prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\psi\left(\frac{x_i-\mu_0}{T}\right)\right)\right] \\
&\le \prod_{i=1}^{n}\mathbb{E}\left[\left(1+\frac{x_i-\mu_0}{T}+\frac{(x_i-\mu_0)^2}{2T^2}\right)\right] \text{ since } \psi(x) \le \log(1+x+x^2/2) \\
&\le \exp\left(\frac{n}{T}(\mu-\mu_0) + \frac{n}{2T^2}\cdot\left[\sigma^2(1+o(1))\right]\right)
\end{aligned}$$

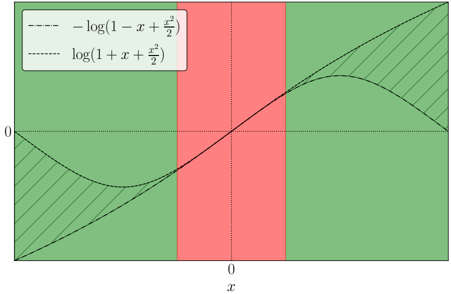
Towards an improved estimator



- Catoni uses

$$-\log\left(1 - x + \frac{x^2}{2}\right) \leq \psi(x) \leq \log\left(1 + x + \frac{x^2}{2}\right)$$

Towards an improved estimator

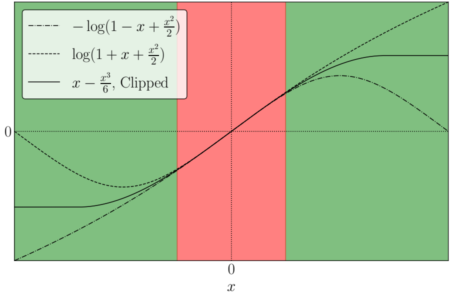


- Catoni uses

$$-\log\left(1 - x + \frac{x^2}{2}\right) \leq \psi(x) \leq \log\left(1 + x + \frac{x^2}{2}\right)$$

- There is slack in the choice for outliers

Towards an improved estimator



- Catoni uses

$$-\log\left(1-x+\frac{x^2}{2}\right) \leq \psi(x) \leq \log\left(1+x+\frac{x^2}{2}\right)$$

- There is slack in the choice for outliers
- Taking advantage of the slack gives smaller error when the distribution has many outliers

Towards an improved estimator

- Recall: $T = \sqrt{\frac{n}{2\log\frac{1}{\delta}}}$ is the *scale* of the outliers
- **Strategy for two-dimensional estimator:** The distribution of x_i is *either* **outlier-heavy**, or **outlier-light**

Towards an improved estimator

- Recall: $T = \sqrt{\frac{n}{2\log\frac{1}{\delta}}}$ is the *scale* of the outliers
- **Strategy for two-dimensional estimator:** The distribution of x_i is *either* **outlier-heavy**, or **outlier-light**
    1. **outlier-heavy**: Elements less than $T/100$ contribute less than 99% of the variance. Then, the Catoni estimate with our improved ψ constraint has a sharper error rate.

Towards an improved estimator

- Recall: $T = \sqrt{\frac{n}{2 \log \frac{1}{\delta}}}$ is the *scale* of the outliers
- **Strategy for two-dimensional estimator:** The distribution of x_i is *either* **outlier-heavy**, or **outlier-light**
  1. **outlier-heavy**: Elements less than $T/100$ contribute less than 99% of the variance. Then, the Catoni estimate with our improved ψ constraint has a sharper error rate.
  2. **Outlier-light**: Elements more than $T/100$ contribute less than 1% of the variance. So, we can *trim* samples past this threshold, and compute an empirical mean.

Towards an improved estimator

- Recall: $T = \sqrt{\frac{n}{2 \log \frac{1}{\delta}}}$ is the *scale* of the outliers
- **Strategy for two-dimensional estimator:** The distribution of x_i is *either* **outlier-heavy**, or **outlier-light**
    1. **outlier-heavy**: Elements less than $T/100$ contribute less than 99% of the variance. Then, the Catoni estimate with our improved ψ constraint has a sharper error rate.
    2. **Outlier-light**: Elements more than $T/100$ contribute less than 1% of the variance. So, we can *trim* samples past this threshold, and compute an empirical mean.
- In both cases, we achieve an improved rate. We can test which case we are in using a small fraction of samples.

Towards an improved estimator

- Recall: $T = \sqrt{\frac{n}{2 \log \frac{1}{\delta}}}$ is the *scale* of the outliers
- **Strategy for two-dimensional estimator:** The distribution of x_i is *either* **outlier-heavy**, or **outlier-light**
    1. **outlier-heavy**: Elements less than $T/100$ contribute less than 99% of the variance. Then, the Catoni estimate with our improved ψ constraint has a sharper error rate.
    2. **Outlier-light**: Elements more than $T/100$ contribute less than 1% of the variance. So, we can *trim* samples past this threshold, and compute an empirical mean.
- In both cases, we achieve an improved rate. We can test which case we are in using a small fraction of samples.
- A generalization of Jung's theorem allows us to lift this estimator to d dimensions with an **improved rate**.

Main Theorems

Heavy-Tailed Estimation, Upper Bound

Suppose $\log \frac{1}{\delta} \gg d$. There is an algorithm that takes n samples in $\mathbb{R}^d$ with covariance of each sample $\Sigma \preceq \sigma^2 I$, and outputs an estimate $\widehat{\mu}$ with

$$\|\widehat{\mu} - \mu\| \leq (1-\tau) \cdot \sqrt{\frac{2d}{d+1}} \cdot \sigma \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$$

with $1-\delta$ probability, for some constant $\tau > 10^{-6}$.

Main Theorems

Heavy-Tailed Estimation, Upper Bound

Suppose $\log \frac{1}{\delta} \gg d$. There is an algorithm that takes n samples in $\mathbb{R}^d$ with covariance of each sample $\Sigma \preceq \sigma^2 I$, and outputs an estimate $\widehat{\mu}$ with

$$\|\widehat{\mu} - \mu\| \leq (1-\tau) \cdot \sqrt{\frac{2d}{d+1}} \cdot \sigma \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$$

with $1-\delta$ probability, for some constant $\tau > 10^{-6}$.

Robust Mean Estimation, Lower Bound

For every $d \geq 1$ and $\varepsilon \leq \frac{1}{2}$, every algorithm for robust estimation of d-dimensional distributions with covariance $\Sigma \preceq \sigma^2 I$ has error rate

$$\mathbb{E}[\|\widehat{\mu} - \mu\|] \geq \sqrt{\frac{2d}{d+1}} \cdot (1 + O(\varepsilon)) \cdot \sqrt{2\sigma^2 \varepsilon}$$

on some input distribution, in the population limit.