Quantum Entropy Scoring for Fast Robust Mean Estimation and Improved Outlier Detection

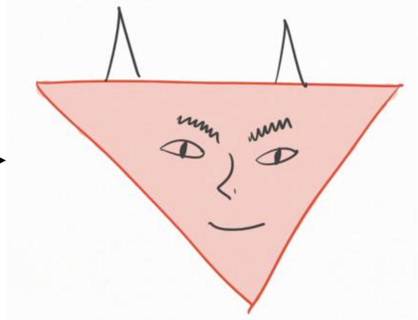
Yihe Dong
Microsoft

Sam Hopkins
UC Berkeley

Jerry Li
Microsoft

Outliers in High Dimensions

Robust mean estimation (theory): estimate mean of D on R^d from ϵ-corrupted samples

$X_1, \ldots, X_n$ →  → ϵ-corrupted samples

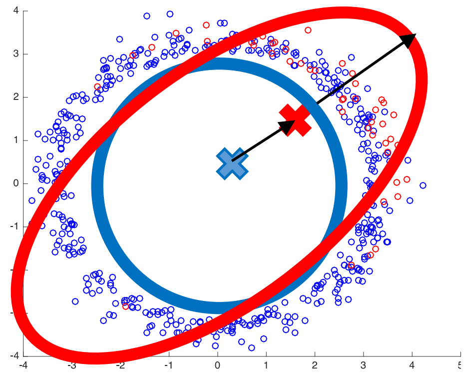
Outlier scoring (experiments): given $X_1, \ldots, X_n$, assign scores $\tau(X_i) \geq 0$ so that outliers receive higher scores

[Anscome '60, Tukey '60, Huber '64, Tukey '75]

Outliers in High Dimensions

2010s robust stats: **PCA** is a good way to find outliers

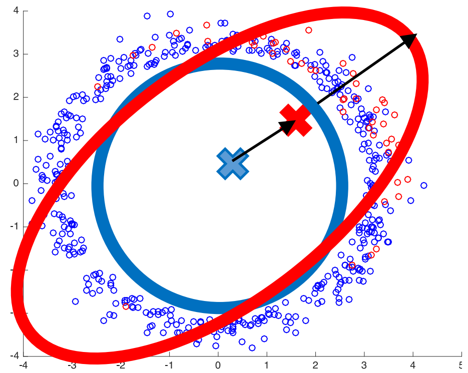
$$\tau(X) = \langle \mathrm{X}, \text{principle component} \rangle^2$$



[Diakonikolas-Kamath-Kane-Li-Moitra-Stewart ’16, Lai-Rao-Vempala ’16]

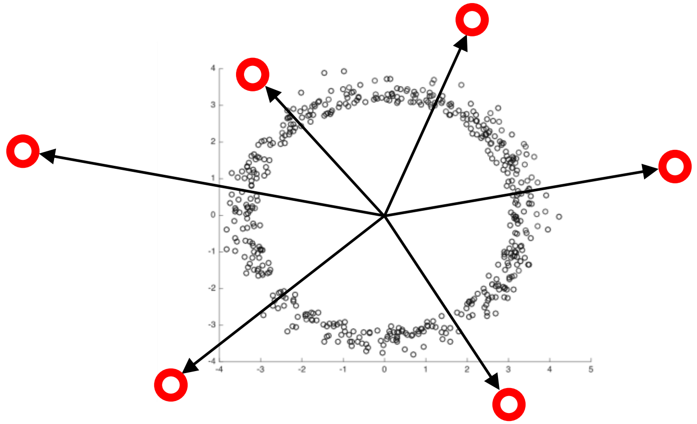
Outliers in High Dimensions

2010s robust stats: **PCA** is a good way to find outliers



$$\tau(X) = \langle \mathrm{X}, \text{principle component} \rangle^2$$

Drawback: only in one direction at a time



Outliers in $\Omega(d)$ directions → $\Omega\left(d^2\right)$ running times

[Diakonikolas-Kamath-Kane-Li-Moitra-Stewart '16, Lai-Rao-Vempala '16]

Quantum Entropy Scoring (QUE)

Given: $X_1, \ldots, X_n \in R^d$, empirical covariance Σ

1. **Regularized** PCA problem

$$W = \arg \max_{\substack{Tr\ U=1 \\ U \succcurlyeq 0}} \alpha \cdot \langle U, \Sigma \rangle + S(U)$$

Quantum Entropy Scoring (QUE)

Given: $X_1, \dots, X_n \in R^d$, empirical covariance Σ

1. **Regularized** PCA problem

$$W = \arg\max_{\substack{Tr\,U=1 \\ U \succcurlyeq 0}} \alpha \cdot \langle U, \Sigma \rangle + S(U)$$

$= \mathrm{Tr}\, U \log U$ ("quantum entropy")

Quantum Entropy Scoring (QUE)

Given: $X_1, \ldots, X_n \in R^d$, empirical covariance Σ

1. **Regularized** PCA problem

$$W = \arg \max_{\substack{Tr\, U = 1 \\ U \succcurlyeq 0}} \alpha \cdot \langle U, \Sigma \rangle + S(U)$$

$= \mathrm{Tr}\, U \log U$ ("quantum entropy")

2. Score $\tau(X) = \langle X X^\top, W \rangle$

Quantum Entropy Scoring (QUE)

Given: $X_1, \dots, X_n \in R^d$, empirical covariance Σ

1. **Regularized** PCA problem

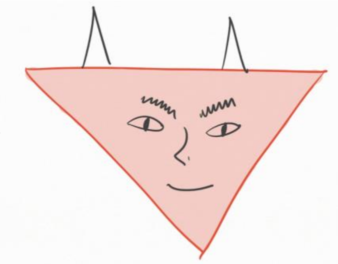
$$W = \arg\max_{\substack{Tr\ U=1 \\ U \succcurlyeq 0}} \alpha \cdot \langle U, \Sigma \rangle + S(U)$$

$= \mathrm{Tr}\, U \log U$ ("quantum entropy")

2. Score $\tau(X) = \langle XX^\top, W \rangle$

Alternatively, $\tau(X) \propto \sum_{i \le d} e^{\alpha \lambda_i} \cdot \langle X, v_i \rangle^2$

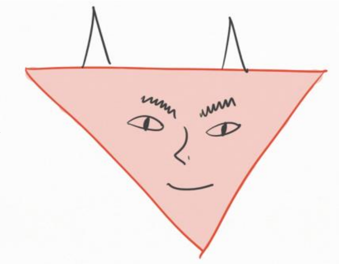
Fast Robust Mean Estimation

$X_1, \ldots, X_n$ →  → ϵ-corrupted samples

Standard Assumption: underlying dist. has bounded covariance

[Diakonikolas-Kamath-Kane-Li-Moitra-Stewart ’17, Cheng-Diakonikolas-Ge ’19]

Fast Robust Mean Estimation

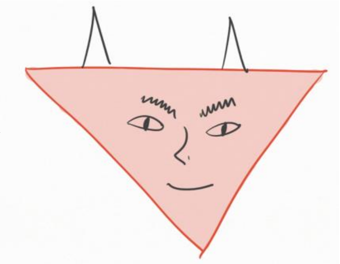
$X_1, \dots, X_n$ →  → ϵ-corrupted samples

Standard Assumption: underlying dist. has bounded covariance

Theorem: can solve with optimal error in nearly-linear time $\tilde{O}(nd)$

[Diakonikolas-Kamath-Kane-Li-Moitra-Stewart ’17, Cheng-Diakonikolas-Ge ’19]

Fast Robust Mean Estimation

$X_1, \dots, X_n$ →  → ϵ-corrupted samples

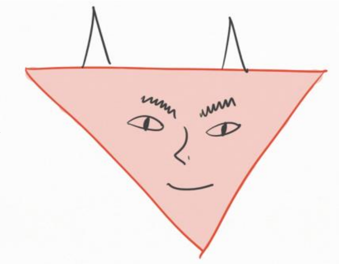
Standard Assumption: underlying dist. has bounded covariance

Theorem: can solve with optimal error in nearly-linear time $\tilde{O}(nd)$

Best previous: $\tilde{O}(nd^2)$, $\tilde{O}(nd/\epsilon^6)$

[Diakonikolas-Kamath-Kane-Li-Moitra-Stewart ’17, Cheng-Diakonikolas-Ge ’19]

Fast Robust Mean Estimation

$X_1, \ldots, X_n$ →  → ϵ-corrupted samples

Standard Assumption: underlying dist. has bounded covariance

Theorem: can solve with optimal error in nearly-linear time $\tilde{O}(nd)$

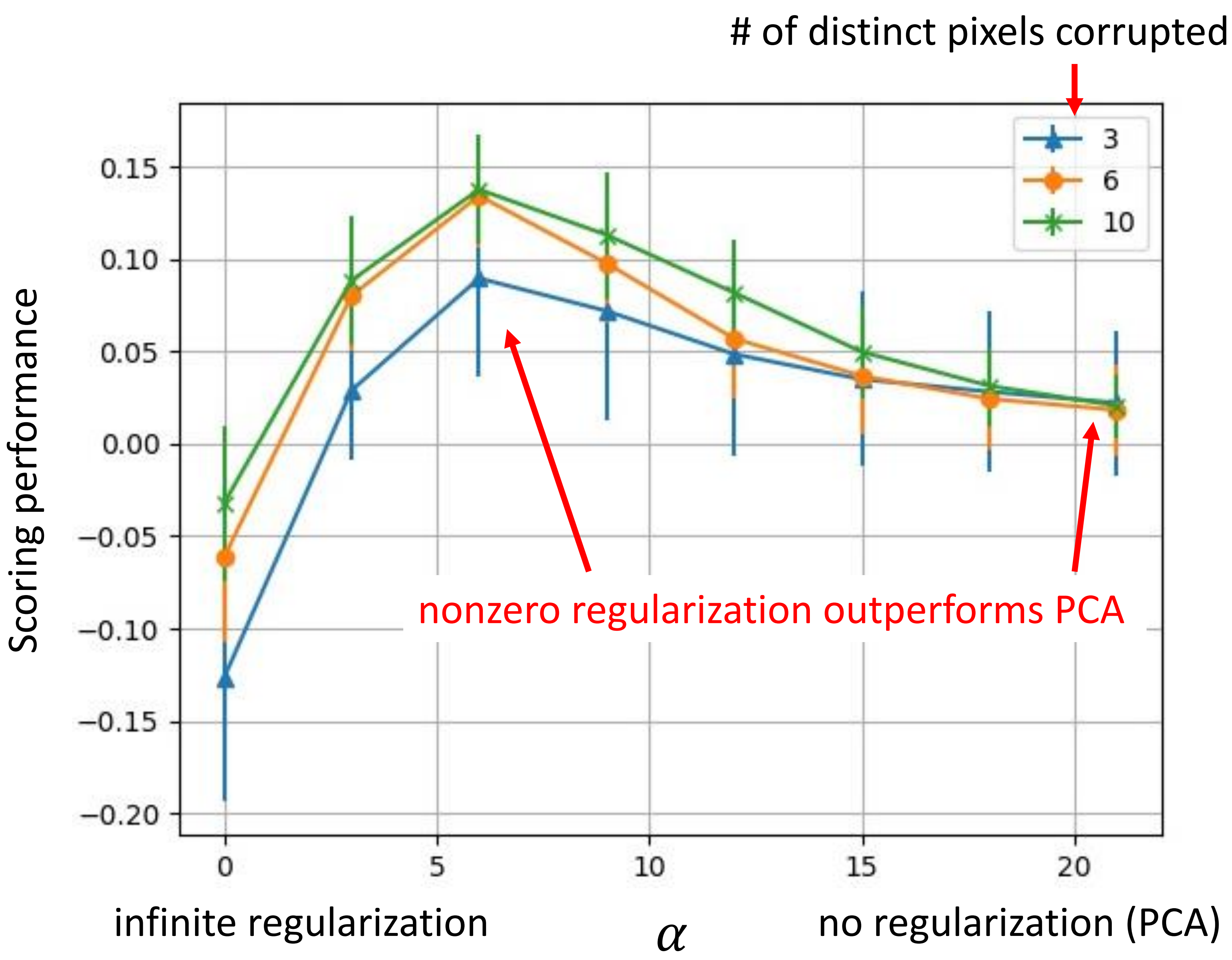
Best previous: $\tilde{O}(nd^2)$, $\tilde{O}(nd/\epsilon^6)$

Algorithm: QUE + matrix multiplicative weights (a.k.a. mirror descent)

[Diakonikolas-Kamath-Kane-Li-Moitra-Stewart ’17, Cheng-Diakonikolas-Ge ’19]

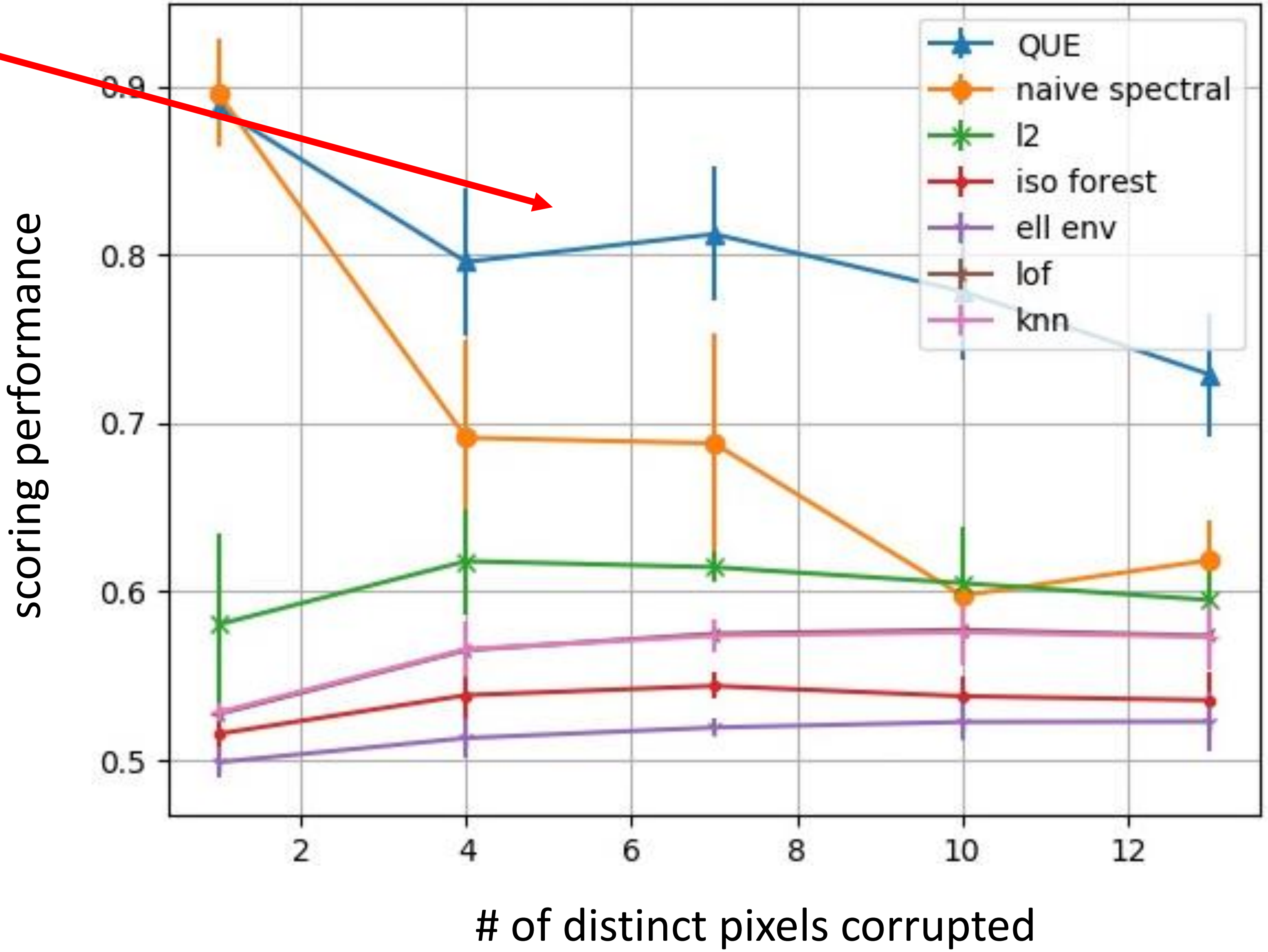
Outlier Scoring: Image Dataset

QUE Scoring: Results



QUE Scoring: Results

QUE outperforms nearest-neighbor methods



[Campos-Zimek-Sander-Campello-Micenková-Schubert-Assent-Houle ’17]

Quantum Entropy Scoring for Fast Robust Mean Estimation and Improved Outlier Detection

Poster No. 228