Matrix Multiplicative Weights and Nearly-Linear Time Robust Statistics

Kevin Tian (UT Austin)

New Frontiers in Robust Statistics Workshop

TTIC, 2024

with many thanks to…

…and more

Talk outline

- A gentle introduction to MMW
  - Regret minimization
  - Matrix analysis
  - Implementation
  - Relatives of MMW
- Robust statistics primitives via MMW
  - Mean estimation
  - A tour of applications

References:

- MMW intro
  - Course notes (Continuous Algorithms, Spring ‘24)
  - Lectures 3, 5-7
  - Email me for pointers
- Robust statistics applications
  - …Continuous Algorithms, Spring ‘25?
  - Email me for pointers

Regret minimization

$$\mathcal{X} \subset \mathbb{R}^d$$

convex, compact
action space

(e.g. norm ball)

Regret minimization

Goal: predict the future!

$$\sum_{t\in[T]} \ell_t(x_t) - \min_{x^\star\in\mathcal{X}} \sum_{t\in[T]} \ell_t(x^\star)$$

Importantly, x_t played before ℓ_t known
(Also, ℓ_t can be adaptive)

$$\mathcal{X} \subset \mathbb{R}^d$$

convex, compact
action space

(e.g. norm ball)

Regret minimization

Goal: *sublinear* regret

$$\sum_{t\in[T]} \ell_t(x_t) - \min_{x^\star\in\mathcal{X}} \sum_{t\in[T]} \ell_t(x^\star) = o(T)$$

Importantly, x_t played before ℓ_t known
(Also, ℓ_t can be adaptive)

$$\mathcal{X} \subset \mathbb{R}^d$$

convex, compact
action space

(e.g. norm ball)

Linear regret minimization

Goal: *sublinear* regret

$$\sum_{t\in[T]} \langle g_t, x_t \rangle - \min_{x^\star \in \mathcal{X}} \sum_{t\in[T]} \langle g_t, x^\star \rangle = o(T)$$

Importantly, x_t played before g_t known
(Also, g_t can be adaptive)

$$\mathcal{X} \subset \mathbb{R}^d$$

convex, compact
action space

(e.g. norm ball)

Linear regret minimization

Goal: *sublinear* regret

$$\sum_{t \in [T]} \langle g_t, x_t \rangle - \min_{x^\star \in \mathcal{X}} \sum_{t \in [T]} \langle g_t, x^\star \rangle = o(T)$$

Application 1: convex optimization

$$g_t = \nabla f(x_t)$$

adaptive!

Linear regret minimization

Goal: *sublinear* regret

$$\underbrace{\sum_{t\in[T]} \langle g_t, x_t\rangle - \min_{x^\star\in\mathcal{X}} \sum_{t\in[T]} \langle g_t, x^\star\rangle}_{\mathrm{Regret}_T} = o(T)$$

Application 1: convex optimization

$$g_t = \nabla f(x_t)$$

adaptive!

$$\mathrm{Regret}_T \geq \sum_{t\in[T]} f(x_t) - f(x^\star) \geq T(f(\bar{x}) - f(x^\star))$$

Linear regret minimization

Goal: *sublinear* regret

$$\underbrace{\sum_{t\in[T]}\langle g_t, x_t\rangle - \min_{x^\star\in\mathcal{X}}\sum_{t\in[T]}\langle g_t, x^\star\rangle}_{\mathrm{Regret}_T} = o(T)$$

Application 1: convex optimization

$$g_t = \nabla f(x_t)$$

adaptive!

$$\mathrm{Regret}_T \geq \sum_{t\in[T]} f(x_t) - f(x^\star) \geq T(f(\bar{x}) - f(x^\star))$$

vanishing suboptimality gap!

\+ works for cvx-ccv saddle point

Linear regret minimization

Goal: *sublinear* regret

$$\underbrace{\sum_{t\in[T]} \langle g_t, x_t \rangle - \min_{x^\star \in \mathcal{X}} \sum_{t\in[T]} \langle g_t, x^\star \rangle}_{\text{Regret}_T} = o(T)$$

Application 2: dual certificates

$$\mathcal{X} := \{x \mid \|x\| \leq 1\}$$

Linear regret minimization

Goal: *sublinear* regret

$$\underbrace{\sum_{t\in[T]} \langle g_t, x_t\rangle - \min_{x^\star\in\mathcal{X}} \sum_{t\in[T]} \langle g_t, x^\star\rangle}_{\mathrm{Regret}_T} = o(T)$$

Application 2: dual certificates

$$\mathcal{X} := \{x \mid \|x\| \le 1\}$$

$$-\min_{x\in\mathcal{X}} \sum_{t\in[T]} \langle -g_t, x^\star\rangle = \max_{x\in\mathcal{X}} \left\langle \sum_{t\in[T]} g_t, x^\star \right\rangle = \left\| \sum_{t\in[T]} g_t \right\|_*$$

..if we choose g_t, regret minimization algos certify bounds

Linear regret minimization

Examples:

Action set

$$\max_{x^\star \in \mathcal{X}} \langle s, x^\star \rangle$$

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid \|x\|_q \leq 1 \right\} \qquad \|s\|_p, \; \frac{1}{p} + \frac{1}{q} = 1$$

q-norm ball

Application 2: dual certificates

$$\mathcal{X} := \{x \mid \|x\| \leq 1\}$$

Linear regret minimization

Examples:

Action set

$$\max_{x^\star \in \mathcal{X}} \langle s, x^\star \rangle$$

$$\mathcal{X} = \left\{x \in \mathbb{R}^d \mid \|x\|_q \leq 1\right\}$$

$$\|s\|_p, \ \frac{1}{p} + \frac{1}{q} = 1$$

q-norm ball

$$\mathcal{X} = \left\{x \in \mathbb{R}^d_{\geq 0} \mid \|x\|_1 = 1\right\}$$

$$\max_{i \in [d]} s_i$$

(probability) simplex

Application 2: dual certificates

$$\mathcal{X} := \{x \mid \|x\| \leq 1\}$$

Linear regret minimization

Examples:

Action set

$$\max_{\mathbf{X}^\star \in \mathcal{X}} \langle \mathbf{S}, \mathbf{X}^\star \rangle$$

$$\mathcal{X} = \left\{ \mathbf{X} \in \mathrm{Sym}^{d\times d} \mid \|\mathbf{X}\|_q \leq 1 \right\} \quad \|\mathbf{S}\|_p \,, \ \frac{1}{p} + \frac{1}{q} = 1$$

Schatten *q*-norm ball

$$\mathcal{X} = \{ \mathbf{X} \in \mathrm{PSD}^{d\times d} \mid \mathrm{Tr}(\mathbf{X}) = 1 \} \qquad \lambda_{\max}(\mathbf{S})$$

Spectraplex

Application 2: dual certificates

$$\mathcal{X} := \{ x \mid \|x\| \leq 1 \}$$

Linear regret minimization

Examples:

Action set

$$\max_{\mathbf{X}^\star \in \mathcal{X}} \langle \mathbf{S}, \mathbf{X}^\star \rangle$$

$$\mathcal{X} = \left\{ \mathbf{X} \in \mathrm{Sym}^{d \times d} \mid \|\mathbf{X}\|_q \leq 1 \right\} \quad \|\mathbf{S}\|_p, \; \frac{1}{p} + \frac{1}{q} = 1$$

Schatten *q*-norm ball

$$\mathcal{X} = \{ \mathbf{X} \in \mathrm{PSD}^{d \times d} \mid \mathrm{Tr}(\mathbf{X}) = 1 \} \qquad \lambda_{\max}(\mathbf{S})$$

Spectraplex

Application 2: dual certificates

$$\mathcal{X} := \{ x \mid \|x\| \leq 1 \}$$

von Neumann trace inequality:

$$\max_{\mathbf{V} \text{ unitary}} \langle \mathbf{V} \mathbf{D_2} \mathbf{V}^\top, \mathbf{U} \mathbf{D_1} \mathbf{U}^\top \rangle$$

achieved iff **V** = **U** up to permutation and subspace invariance

Linear regret minimization

Examples:

Action set

$$\max_{\mathbf{X}^\star \in \mathcal{X}} \langle \mathbf{S}, \mathbf{X}^\star \rangle$$

$$\mathcal{X} = \left\{ \mathbf{X} \in \mathrm{Sym}^{d \times d} \mid \|\mathbf{X}\|_q \le 1 \right\} \quad \|\mathbf{S}\|_p \,, \; \frac{1}{p} + \frac{1}{q} = 1$$

Schatten *q*-norm ball

$$\mathcal{X} = \{ \mathbf{X} \in \mathrm{PSD}^{d \times d} \mid \mathrm{Tr}(\mathbf{X}) = 1 \} \qquad \lambda_{\max}(\mathbf{S})$$

Spectraplex

Application 2: dual certificates

$$\mathcal{X} := \{ x \mid \|x\| \le 1 \}$$

Upside:

Most SOTA fast robust stats algos based on this connection to regret minimization!

Linear regret minimization

Examples:

$$\mathcal{X} = \{\mathbf{X} \in \mathrm{PSD}^{d\times d} \mid \mathrm{Tr}(\mathbf{X}) = 1\}$$

Spectraplex

Application 3: SDP feasibility

$$\exists y \in \mathcal{Y} : \sum_{i\in[n]} y_i \mathbf{A}_i \succeq \mathbf{0}_d?$$

Linear regret minimization

Examples:

$$\mathcal{X} = \{\mathbf{X} \in \mathrm{PSD}^{d\times d} \mid \mathrm{Tr}(\mathbf{X}) = 1\}$$

Spectraplex

Application 3: SDP feasibility

$$\exists y \in \mathcal{Y} : \sum_{i\in[n]} y_i \mathbf{A}_i \succeq \mathbf{0}_d?$$

Regret minimization for approximate saddle point of:

$$\min_{\mathbf{X}\in\mathcal{X}} \max_{y\in\mathcal{Y}} \left\langle \mathbf{X}, \sum_{i\in[n]} y_i \mathbf{A}_i \right\rangle$$

Mirror descent

Assumptions

$$\|g_t\|_* \leq G \text{ for all } t$$

$$\varphi : \mathcal{X} \to \mathbb{R} \text{ is 1-s.c. in } \|\cdot\|$$

$$\max_{x \in \mathcal{X}} \varphi(x) - \min_{x \in \mathcal{X}} \varphi(x) \leq \Theta$$

Mirror descent

Assumptions

$$\|g_t\|_* \leq G \text{ for all } t$$

$$\varphi : \mathcal{X} \to \mathbb{R} \text{ is 1-s.c. in } \|\cdot\|$$

$$\max_{x \in \mathcal{X}} \varphi(x) - \min_{x \in \mathcal{X}} \varphi(x) \leq \Theta$$

$$v^\top \nabla^2 \varphi(x) v \geq \|v\|^2$$

Mirror descent

$$\|g_t\|_* \leq G \text{ for all } t$$

$$\varphi : \mathcal{X} \to \mathbb{R} \text{ is 1-s.c. in } \|\cdot\|$$

$$\max_{x \in \mathcal{X}} \varphi(x) - \min_{x \in \mathcal{X}} \varphi(x) \leq \Theta$$

$$\max_{x^\star \in \mathcal{X}} \sum_{t \in [T]} \langle g_t, x_t - x^\star \rangle \lesssim G\sqrt{\Theta T}$$

Mirror descent

$$\|g_t\|_* \leq G \text{ for all } t$$

$$\varphi : \mathcal{X} \to \mathbb{R} \text{ is 1-s.c. in } \|\cdot\|$$

$$\max_{x \in \mathcal{X}} \varphi(x) - \min_{x \in \mathcal{X}} \varphi(x) \leq \Theta$$

Different regularity:

RHS is $\text{poly}\,(G, \Theta)$

$$\max_{x^\star \in \mathcal{X}} \sum_{t \in [T]} \langle g_t, x_t - x^\star \rangle \lesssim G\sqrt{\Theta T}$$

Aside: convex duality

$$\varphi^*(y) := \max_{x \in \mathcal{X}} \langle y, x \rangle - \varphi(x)$$

Conjugate of convex function

Aside: convex duality

$$\varphi^*(y) := \max_{x \in \mathcal{X}} \langle y, x \rangle - \varphi(x)$$

Conjugate of convex function

e.g. $\varphi(x) = \frac{1}{2}\|x\|^2, \ \varphi^*(y) = \frac{1}{2}\|y\|_*^2$

Aside: convex duality

$$\varphi^*(y) := \max_{x \in \mathcal{X}} \langle y, x \rangle - \varphi(x)$$

Conjugate of convex function

$$\nabla \varphi^*(y) = \operatorname{argmax}_{x \in \mathcal{X}} \langle y, x \rangle - \varphi(x)$$

Maximizing argument

Mirror descent

$$x_t \leftarrow \nabla\varphi^* \left(-\eta \sum_{s<t} g_s \right)$$

$$\|g_t\|_* \leq G \text{ for all } t$$

$$\varphi : \mathcal{X} \to \mathbb{R} \text{ is 1-s.c. in } \|\cdot\|$$

$$\max_{x \in \mathcal{X}} \varphi(x) - \min_{x \in \mathcal{X}} \varphi(x) \leq \Theta$$

$$\max_{x^\star \in \mathcal{X}} \sum_{t \in [T]} \langle g_t, x_t - x^\star \rangle \lesssim G\sqrt{\Theta T}$$

Mirror descent

$$\|g_t\|_* \leq G \text{ for all } t$$

$$\varphi : \mathcal{X} \to \mathbb{R} \text{ is 1-s.c. in } \|\cdot\|$$

$$\max_{x \in \mathcal{X}} \varphi(x) - \min_{x \in \mathcal{X}} \varphi(x) \leq \Theta$$

$$x_t \leftarrow \nabla \varphi^* \left(-\eta \sum_{s<t} g_s \right)$$

$$= \operatorname{argmin}_{x \in \mathcal{X}} \left\{ \eta \sum_{s<t} \langle g_s, x \rangle + \varphi(x) \right\}$$

"follow the regularized leader"

$$\max_{x^\star \in \mathcal{X}} \sum_{t \in [T]} \langle g_t, x_t - x^\star \rangle \lesssim G\sqrt{\Theta T}$$

Mirror descent

$$\|g_t\|_* \leq G \text{ for all } t$$

$$\varphi : \mathcal{X} \to \mathbb{R} \text{ is 1-s.c. in } \|\cdot\|$$

$$\max_{x \in \mathcal{X}} \varphi(x) - \min_{x \in \mathcal{X}} \varphi(x) \leq \Theta$$

Mirror descent

$$\|g_t\|_* \leq G \text{ for all } t$$

Gold standard
(for εT regret over norm ball):

$$T = \text{poly}\left(\frac{G\log(d)}{\epsilon}\right)$$

Quantum entropy

$$\mathcal{X} = \{\mathbf{X} \in \mathrm{PSD}^{d \times d} \mid \mathrm{Tr}(\mathbf{X}) = 1\}$$

Basic primitive: spectral bounds

Quantum entropy

$$\mathcal{X} = \{\mathbf{X} \in \mathrm{PSD}^{d\times d} \mid \mathrm{Tr}(\mathbf{X}) = 1\}$$

$$\varphi(\mathbf{X}) = \sum_{i\in[d]} \lambda_i(\mathbf{X}) \log \lambda_i(\mathbf{X})$$

Quantum (von Neumann) entropy

Quantum entropy

$$\mathcal{X} = \{\mathbf{X} \in \mathrm{PSD}^{d\times d} \mid \mathrm{Tr}(\mathbf{X}) = 1\}$$

$$\varphi(\mathbf{X}) = \sum_{i\in[d]} \lambda_i(\mathbf{X}) \log \lambda_i(\mathbf{X})$$

Checklist:

- Strongly convex?
- Bounded?
- Implementable?

Quantum entropy

$$\mathcal{X} = \{\mathbf{X} \in \mathrm{PSD}^{d \times d} \mid \mathrm{Tr}(\mathbf{X}) = 1\}$$

$$\varphi(\mathbf{X}) = \sum_{i \in [d]} \lambda_i(\mathbf{X}) \log \lambda_i(\mathbf{X})$$

Spoilers:

- Strongly convex in the trace norm
- Range: $\tilde{O}(1)$
- Matvec to $\nabla \varphi^*(\mathbf{Y})$ in near-linear time

Quantum entropy

$$\mathcal{X} = \{\mathbf{X} \in \mathrm{PSD}^{d \times d} \mid \mathrm{Tr}(\mathbf{X}) = 1\}$$

$$\varphi(\mathbf{X}) = \sum_{i \in [d]} \lambda_i(\mathbf{X}) \log \lambda_i(\mathbf{X})$$

Matrix multiplicative weights: mirror descent w.r.t. quantum entropy

Spoilers:

- Strongly convex in the trace norm
- Range: $\tilde{O}(1)$
- Matvec to $\nabla \varphi^*(\mathbf{Y})$ in near-linear time

Gradients

$$\varphi^*(\mathbf{Y}) = \max_{\mathbf{X}\in\mathcal{X}} \langle \mathbf{Y}, \mathbf{X} \rangle - \sum_{i\in[d]} \lambda_i(\mathbf{X}) \log \lambda_i(\mathbf{X})$$

$$= \log \mathrm{Tr} \exp\left(\mathbf{Y}\right)$$

Gradients

$$\varphi^*(\mathbf{Y}) = \max_{\mathbf{X} \in \mathcal{X}} \langle \mathbf{Y}, \mathbf{X} \rangle - \sum_{i \in [d]} \lambda_i(\mathbf{X}) \log \lambda_i(\mathbf{X})$$

$$= \log \mathrm{Tr} \exp (\mathbf{Y})$$

Useful fact: $$f_{\mathrm{mat}}(\mathbf{M}) = f_{\mathrm{vec}}(\lambda(\mathbf{M}))$$ "spectral function"

$$\mathbf{M} = \mathbf{U} \mathrm{diag}(\lambda(\mathbf{M})) \mathbf{U}^\top$$

Gradients

$$\varphi^*(\mathbf{Y}) = \max_{\mathbf{X}\in\mathcal{X}} \langle \mathbf{Y}, \mathbf{X} \rangle - \sum_{i\in[d]} \lambda_i(\mathbf{X}) \log \lambda_i(\mathbf{X})$$

$$= \log \operatorname{Tr} \exp\left(\mathbf{Y}\right)$$

Useful fact: $$f_{\text{mat}}(\mathbf{M}) = f_{\text{vec}}(\lambda(\mathbf{M}))$$ *convex*, spectral

$$\nabla f_{\text{mat}}(\mathbf{X}) = \mathbf{U}\text{diag}(\nabla f_{\text{vec}}(\lambda(\mathbf{M})))\mathbf{U}^\top$$

Gradients

$$\varphi^*(\mathbf{Y}) = \max_{\mathbf{X}\in\mathcal{X}} \langle \mathbf{Y}, \mathbf{X} \rangle - \sum_{i\in[d]} \lambda_i(\mathbf{X}) \log \lambda_i(\mathbf{X})$$

$$= \log \operatorname{Tr} \exp\left(\mathbf{Y}\right)$$

Useful fact:

$$f_{\text{mat}}(\mathbf{M}) = f_{\text{vec}}(\lambda(\mathbf{M}))$$

$$\nabla f_{\text{mat}}(\mathbf{X}) = \mathbf{U}\text{diag}(\nabla f_{\text{vec}}(\lambda(\mathbf{M})))\mathbf{U}^\top$$

Why?
…von Neumann!

Gradients

$$\varphi^*(\mathbf{Y}) = \max_{\mathbf{X}\in\mathcal{X}} \langle \mathbf{Y}, \mathbf{X} \rangle - \sum_{i\in[d]} \lambda_i(\mathbf{X}) \log \lambda_i(\mathbf{X})$$

$$= \log \operatorname{Tr} \exp(\mathbf{Y})$$

Example: $f_{\text{vec}}(\lambda) = \log\left(\sum_{i\in[d]} \exp(\lambda_i)\right)$ $\nabla f_{\text{vec}}(\lambda) = \frac{\exp(\lambda)}{\|\exp(\lambda)\|_1}$

Gradients

$$\varphi^*(\mathbf{Y}) = \max_{\mathbf{X} \in \mathcal{X}} \langle \mathbf{Y}, \mathbf{X} \rangle - \sum_{i \in [d]} \lambda_i(\mathbf{X}) \log \lambda_i(\mathbf{X})$$

$$= \log \operatorname{Tr} \exp(\mathbf{Y})$$

Example:

$$\nabla \varphi^*(\mathbf{Y}) = \frac{\exp(\mathbf{Y})}{\operatorname{Tr} \exp(\mathbf{Y})}$$

Strong convexity

$$\varphi^*(\mathbf{Y}) = \log \operatorname{Tr} \exp(\mathbf{Y})$$

Useful fact: $$v^\top \nabla^2 \varphi^*(y) v \le \|v\|_*^2$$

$$\iff v^\top \nabla^2 \varphi(x) v \ge \|v\|^2$$

Strong convexity

$$\varphi^*(\mathbf{Y}) = \log \operatorname{Tr} \exp(\mathbf{Y})$$

Useful fact:

$$v^\top \nabla^2 \varphi^*(y) v \le \|v\|_*^2$$

$$\iff v^\top \nabla^2 \varphi(x) v \ge \|v\|^2$$

"Smoothness-strong convexity duality"

Strong convexity

$$\varphi^*(\mathbf{Y}) = \log \operatorname{Tr} \exp(\mathbf{Y})$$

Useful fact:

$$v^\top \nabla^2 \varphi^*(y) v \leq \|v\|_*^2$$

$$\iff v^\top \nabla^2 \varphi(x) v \geq \|v\|^2$$

"Smoothness-strong convexity duality"

Why?
Taylor expansion +

$$\varphi(x) = \frac{1}{2}\|x\|^2, \ \varphi^*(y) = \frac{1}{2}\|y\|_*^2$$

Aside: "disentangling" lemma

$$\mathrm{Tr}\left(\mathbf{M}^{\alpha}\mathbf{N}\mathbf{M}^{1-\alpha}\mathbf{N}\right) \leq \mathrm{Tr}(\mathbf{M}\mathbf{N}^2)$$

$\mathbf{M}, \mathbf{N} \in \mathrm{PSD}^{d \times d}$

$\alpha \in [0, 1]$

Aside: "disentangling" lemma

$$\mathrm{Tr}\left(\mathbf{M}^{\alpha}\mathbf{N}\mathbf{M}^{1-\alpha}\mathbf{N}\right) \leq \mathrm{Tr}(\mathbf{M}\mathbf{N}^{2})$$

$$\mathbf{M}, \mathbf{N} \in \mathrm{PSD}^{d \times d}$$

$$\alpha \in [0, 1]$$

Proof sketch:

$$\begin{pmatrix} \mathbf{N} & -\mathbf{N}^{\frac{1}{2}}\mathbf{M}^{\alpha}\mathbf{N}^{\frac{1}{2}} \\ -\mathbf{N}^{\frac{1}{2}}\mathbf{M}^{\alpha}\mathbf{N}^{\frac{1}{2}} & \mathbf{N}^{\frac{1}{2}}\mathbf{M}^{2\alpha}\mathbf{N}^{\frac{1}{2}} \end{pmatrix} \in \mathrm{PSD}^{2d \times 2d}$$

$f(\alpha) = \mathrm{Tr}(\mathbf{M}^{\alpha}\mathbf{N}\mathbf{M}^{1-\alpha}\mathbf{N})$ is convex

Dual smoothness

$$\varphi^*(\mathbf{Y}) = \log \operatorname{Tr} \exp(\mathbf{Y})$$

$$\nabla^2 \varphi^*(\mathbf{Y})[\mathbf{M}, \mathbf{M}] \leq \frac{1}{\operatorname{Tr} \exp(\mathbf{Y})} \langle \mathbf{M}, \nabla(\langle \mathbf{M}, \exp(\mathbf{Y}) \rangle) \rangle$$

Dual smoothness

$$\varphi^*(\mathbf{Y}) = \log \operatorname{Tr} \exp(\mathbf{Y})$$

$$\nabla^2 \varphi^*(\mathbf{Y})[\mathbf{M}, \mathbf{M}] \leq \frac{1}{\operatorname{Tr} \exp(\mathbf{Y})} \langle \mathbf{M}, \nabla(\langle \mathbf{M}, \exp(\mathbf{Y}) \rangle) \rangle$$

$$\langle \mathbf{M}, \nabla \left(\langle \mathbf{M}, \exp(\mathbf{Y}) \rangle \right) \rangle = \sum_{k=1}^{\infty} \sum_{i=0}^{k-1} \frac{1}{k!} \langle \mathbf{M}, \mathbf{Y}^i \mathbf{M} \mathbf{Y}^{k-1-i} \rangle$$

Dual smoothness

$$\varphi^*(\mathbf{Y}) = \log \operatorname{Tr} \exp(\mathbf{Y})$$

$$\nabla^2 \varphi^*(\mathbf{Y})[\mathbf{M}, \mathbf{M}] \leq \frac{1}{\operatorname{Tr} \exp(\mathbf{Y})} \langle \mathbf{M}, \nabla(\langle \mathbf{M}, \exp(\mathbf{Y}) \rangle) \rangle$$

$$\langle \mathbf{M}, \nabla \left(\langle \mathbf{M}, \exp(\mathbf{Y}) \rangle \right) \rangle = \sum_{k=1}^{\infty} \sum_{i=0}^{k-1} \frac{1}{k!} \langle \mathbf{M}, \mathbf{Y}^i \mathbf{M} \mathbf{Y}^{k-1-i} \rangle$$

$$\leq \sum_{k=0}^{\infty} \frac{1}{k!} \langle \mathbf{M}^2, \mathbf{Y}^k \rangle \qquad \text{(disentangling)}$$

Dual smoothness

$$\varphi^*(\mathbf{Y}) = \log \operatorname{Tr} \exp(\mathbf{Y})$$

$$\nabla^2 \varphi^*(\mathbf{Y})[\mathbf{M}, \mathbf{M}] \leq \frac{1}{\operatorname{Tr} \exp(\mathbf{Y})} \langle \mathbf{M}, \nabla(\langle \mathbf{M}, \exp(\mathbf{Y}) \rangle) \rangle$$

$$\langle \mathbf{M}, \nabla\left(\langle \mathbf{M}, \exp(\mathbf{Y}) \rangle\right) \rangle = \sum_{k=1}^{\infty} \sum_{i=0}^{k-1} \frac{1}{k!} \langle \mathbf{M}, \mathbf{Y}^i \mathbf{M} \mathbf{Y}^{k-1-i} \rangle$$

$$\leq \sum_{k=0}^{\infty} \frac{1}{k!} \langle \mathbf{M}^2, \mathbf{Y}^k \rangle = \langle \mathbf{M}^2, \exp(\mathbf{Y}) \rangle$$

Dual smoothness

$$\varphi^*(\mathbf{Y}) = \log \operatorname{Tr} \exp(\mathbf{Y})$$

$$\nabla^2 \varphi^*(\mathbf{Y})[\mathbf{M}, \mathbf{M}] \leq \left\langle \mathbf{M}^2, \frac{\exp(\mathbf{Y})}{\operatorname{Tr} \exp(\mathbf{Y})} \right\rangle$$

Dual smoothness

$$\varphi^*(\mathbf{Y}) = \log \operatorname{Tr} \exp(\mathbf{Y})$$

$$\begin{aligned}\nabla^2 \varphi^*(\mathbf{Y})[\mathbf{M}, \mathbf{M}] &\leq \left\langle \mathbf{M}^2, \frac{\exp(\mathbf{Y})}{\operatorname{Tr}\exp(\mathbf{Y})} \right\rangle \\ &\leq \lambda_{\max}(\mathbf{M}^2) = \|\mathbf{M}\|_\infty^2\end{aligned}$$

Dual smoothness

$$\varphi^*(\mathbf{Y}) = \log \operatorname{Tr} \exp(\mathbf{Y})$$

$$\begin{aligned}\nabla^2 \varphi^*(\mathbf{Y})[\mathbf{M}, \mathbf{M}] &\leq \left\langle \mathbf{M}^2, \frac{\exp(\mathbf{Y})}{\operatorname{Tr} \exp(\mathbf{Y})} \right\rangle \\ &\leq \lambda_{\max}(\mathbf{M}^2) = \|\mathbf{M}\|_\infty^2\end{aligned}$$

…so, entropy is strongly convex in Schatten-1!

Dual smoothness

$$\varphi^*(\mathbf{Y}) = \log \operatorname{Tr} \exp(\mathbf{Y})$$

$$\nabla^2 \varphi^*(\mathbf{Y})[\mathbf{M}, \mathbf{M}] \leq \left\langle \mathbf{M}^2, \frac{\exp(\mathbf{Y})}{\operatorname{Tr} \exp(\mathbf{Y})} \right\rangle$$

$$\leq \lambda_{\max}(\mathbf{M}^2) = \|\mathbf{M}\|_\infty^2$$

"local norms"
smoothness bound

Implementation

$$\mathbf{Y} = -\eta \sum_{s<t} \mathbf{G}_s$$

$$\nabla \varphi^*(\mathbf{Y}) = \frac{\exp \mathbf{Y}}{\operatorname{Tr} \exp(\mathbf{Y})}$$

Implementation

$$\mathbf{Y} = -\eta \sum_{s<t} \mathbf{G}_s \qquad \|\mathbf{Y}\|_{\text{op}} = \text{poly}\left(\frac{G\log(d)}{\epsilon}\right)$$

$$\nabla \varphi^*(\mathbf{Y}) = \frac{\exp \mathbf{Y}}{\operatorname{Tr}\exp(\mathbf{Y})}$$

Our action: need to "access" efficiently

Implementation

$$a \approx v^\top \exp(\mathbf{Y}) v$$

Warmup: single quadratic form

Implementation

$$a \approx v^\top \exp(\mathbf{Y})v$$

Key idea: polynomial approximation

$$\text{degree-}\Delta\ p$$

$$\implies \mathcal{T}_{\mathrm{mv}}(p(\mathbf{Y})) = O(\mathcal{T}_{\mathrm{mv}}(\mathbf{Y}) \cdot \Delta)$$

Implementation

$$a \approx v^\top \exp(\mathbf{Y})v$$

$$p(x) \approx \exp(x) \text{ for } x \in [\lambda_{\min}(\mathbf{Y}), \lambda_{\max}(\mathbf{Y})]$$

Implementation

$$a \approx v^\top \exp(\mathbf{Y}) v$$

$$p(x) \approx \exp(x) \text{ for } x \in [\lambda_{\min}(\mathbf{Y}), \lambda_{\max}(\mathbf{Y})]$$

$$\lambda_{\max}(\mathbf{M}) - \lambda_{\min}(\mathbf{M}) \leq R$$

Degree $\approx R$ Taylor approximation is high-accuracy

Implementation

$$a = v^\top p(\mathbf{Y}) v$$

Computable in "nearly-linear time":

$$\mathcal{T}_{\mathrm{mv}}(\mathbf{Y}) \cdot \mathrm{poly}\left(\frac{G \log(d)}{\epsilon}\right)$$

Implementation

$$\left\{a_i \approx v_i^\top \exp\left(\mathbf{Y}\right) v_i\right\}_{i \in [n]}$$

…what good is one quadratic form?

Implementation

$$\left\{a_i \approx v_i^\top \exp\left(\mathbf{Y}\right) v_i\right\}_{i \in [n]}$$

$$v^\top \exp\left(\mathbf{Y}\right) v \approx v^\top \mathbf{Q} \exp\left(\mathbf{Y}\right) \mathbf{Q}^\top v$$

(**Q** is any JL matrix)

Key idea: reuse multiplies via sketching

(warning: independence)

Implementation

$$\left\{a_i \approx v_i^\top \exp\left(\mathbf{Y}\right) v_i\right\}_{i\in[n]}$$

$$\begin{pmatrix} \tilde{q}_1^\top \\ \vdots \\ \tilde{q}_k^\top \end{pmatrix} = p\left(\frac{1}{2}\mathbf{Y}\right) \begin{pmatrix} q_1^\top \\ \vdots \\ q_k^\top \end{pmatrix}$$

Implementation

$$\left\{a_i \approx v_i^\top \exp\left(\mathbf{Y}\right) v_i\right\}_{i\in[n]}$$

$$\begin{pmatrix} \tilde{q}_1^\top \\ \vdots \\ \tilde{q}_k^\top \end{pmatrix} = p\left(\frac{1}{2}\mathbf{Y}\right) \begin{pmatrix} q_1^\top \\ \vdots \\ q_k^\top \end{pmatrix}$$

runtime: $\mathcal{T}_{\mathrm{mv}}\left(\mathbf{Y}\right) \cdot \mathrm{poly}\left(\frac{G\log(d)}{\epsilon}\right)$

$$k = \mathrm{poly}\left(\frac{G\log(d)}{\epsilon}\right)$$

Implementation

$$\left\{a_i \approx v_i^\top \exp\left(\mathbf{Y}\right) v_i\right\}_{i\in[n]}$$

…at "test time"…

$$a_i = \left\| \begin{pmatrix} \tilde{q}_1^\top \\ \vdots \\ \tilde{q}_k^\top \end{pmatrix} v_i \right\|_2^2$$

runtime:

$$d \cdot \mathrm{poly}\left(\frac{G \log(d)}{\epsilon}\right)$$

$$k = \mathrm{poly}\left(\frac{G \log(d)}{\epsilon}\right)$$

MMW summary

$$\max_{\substack{\mathbf{X}^\star \in \mathrm{PSD}^{d\times d} \\ \mathrm{Tr}(\mathbf{X}^\star)=1}} \sum_{t\in[T]} \langle \mathbf{G}_t, \mathbf{X}^\star - \mathbf{X}_t \rangle \lesssim G\sqrt{T}$$

mirror descent bound

$$\|\mathbf{G}_t\|_\infty \leq G \text{ for all } t \in [T]$$

MMW summary

$$\max_{\substack{\mathbf{X}^\star \in \mathrm{PSD}^{d\times d} \\ \mathrm{Tr}(\mathbf{X}^\star)=1}} \frac{1}{T} \sum_{t\in[T]} \langle \mathbf{G}_t, \mathbf{X}^\star - \mathbf{X}_t \rangle \le \epsilon$$

$$T = \mathrm{poly}\left(\frac{G \log(d)}{\epsilon} \right)$$

iteration
count

MMW summary

$$\max_{\substack{\mathbf{X}^\star \in \mathrm{PSD}^{d\times d} \\ \mathrm{Tr}(\mathbf{X}^\star)=1}} \frac{1}{T} \sum_{t\in[T]} \langle \mathbf{G}_t, \mathbf{X}^\star - \mathbf{X}_t \rangle \leq \epsilon$$

$$T = \mathrm{poly}\left(\frac{G\log(d)}{\epsilon}\right) \qquad \left(\sum_{t\in[T]} \mathcal{T}_{\mathrm{mv}}\left(\mathbf{G}_t\right)\right) \cdot \mathrm{poly}\left(\frac{G\log(d)}{\epsilon}\right)$$

iteration
count

cost for "implementing"
each iteration

Improvement: runtime

$$\max_{\substack{\mathbf{X}^\star \in \mathrm{PSD}^{d\times d} \\ \mathrm{Tr}(\mathbf{X}^\star)=1}} \frac{1}{T} \sum_{t\in[T]} \langle \mathbf{G}_t, \mathbf{X}^\star - \mathbf{X}_t \rangle \leq \epsilon$$

[CDST19], see also [BBN13] for different SOTA tradeoff

$$\left(\frac{G\log(d)}{\epsilon}\right)^2$$

iteration count

$$\left(\frac{G\log(d)}{\epsilon}\right)^{0.5}$$

cost for "implementing" each iteration

Improvement: local norms

"prediction error"
per iteration,
controlled by s.c.

$$\max_{\substack{\mathbf{X}^\star \in \mathrm{PSD}^{d \times d} \\ \mathrm{Tr}(\mathbf{X}^\star)=1}} \sum_{t \in [T]} \langle \mathbf{G}_t, \mathbf{X}^\star - \mathbf{X}_t \rangle \lesssim \frac{1}{\eta} + \eta G^2 T$$

size of
regularizer

Improvement: local norms

$$\max_{\substack{\mathbf{X}^\star \in \mathrm{PSD}^{d\times d} \\ \mathrm{Tr}(\mathbf{X}^\star)=1}} \sum_{t\in[T]} \langle \mathbf{G}_t, \mathbf{X}^\star - \mathbf{X}_t \rangle \lesssim \frac{1}{\eta} + \eta G^2 T$$

$$\max_{\substack{\mathbf{X}^\star \in \mathrm{PSD}^{d\times d} \\ \mathrm{Tr}(\mathbf{X}^\star)=1}} \sum_{t\in[T]} \langle \mathbf{G}_t, \mathbf{X}^\star - \mathbf{X}_t \rangle \lesssim \frac{1}{\eta} + \eta G \sum_{t\in[T]} \langle \mathbf{G}_t, \mathbf{X}_t \rangle$$

can drastically improve
if $\mathbf{G}_t$ reacts to $\mathbf{X}_t$

Extension: Schatten-norm setups

$$\varphi(\mathbf{X}) = \frac{1}{2(q-1)}\|\mathbf{X}\|_q^2$$

globally 1-s.c. in Schatten-*q* norm

$$\varphi(\mathbf{X}) = \frac{1}{2q(q-1)}\|\mathbf{X}\|_q^q$$

1-s.c. in Schatten-*q* norm on unit ball

Extension: Schatten-norm setups

$$\varphi(\mathbf{X}) = \frac{1}{2(q-1)} \|\mathbf{X}\|_q^2$$

globally 1-s.c. in Schatten-*q* norm

$$\varphi(\mathbf{X}) = \frac{1}{2q(q-1)} \|\mathbf{X}\|_q^q$$

1-s.c. in Schatten-*q* norm on unit ball

Who cares?

…better captures multiplicative (vs. additive)

…offers different tradeoffs (e.g. lower moment bounds)

Extension: Positive SDP

$$\min_{\mathbf{X}\in\mathcal{X}} \max_{y\in\mathcal{Y}} \left\langle \mathbf{X}, \sum_{i\in[n]} y_i \mathbf{A}_i \right\rangle$$

Canonical application:
feasibility SDP via saddle points

Extension: Positive SDP

$$\min_{\mathbf{X}\in\mathcal{X}} \max_{y\in\mathcal{Y}} \left\langle \mathbf{X}, \sum_{i\in[n]} y_i \mathbf{A}_i \right\rangle$$

Canonical application:
feasibility SDP via saddle points

If all $\mathbf{A}_i$ are PSD or NSD, can get *multiplicative* error guarantees with no dependence on "width"

$$G := \max_{i\in[n]} \lambda_{\max}\left(\mathbf{A}_i\right)$$

Extension: Positive SDP

$$\min_{\mathbf{X} \in \mathcal{X}} \max_{y \in \mathcal{Y}} \left\langle \mathbf{X}, \sum_{i \in [n]} y_i \mathbf{A}_i \right\rangle$$

Canonical application:
feasibility SDP via saddle points

If all $\mathbf{A}_i$ are PSD or NSD, can get *multiplicative* error guarantees with no dependence on "width"

- Works at every scale
- Often the case in robust statistics! (Sample covariances)

Talk outline

- A gentle introduction to MMW
  - Regret minimization
  - Matrix analysis
  - Implementation
  - Relatives of MMW
- Robust statistics primitives via MMW
  - Mean estimation
  - A tour of applications

Robust mean estimation

$$\{X_i^\star\}_{i\in[n]} \sim_{\text{i.i.d.}} \mathcal{D}$$

We observe: $\{X_i = X_i^\star\}_{i\in G}$

$$\{X_i\}_{i\in B}\,,\ |B| \approx \epsilon n$$

Setting

Robust mean estimation

$$\{X_i^\star\}_{i\in[n]} \sim_{\text{i.i.d.}} \mathcal{D}$$

We observe: $\{X_i = X_i^\star\}_{i\in G}$

$\{X_i\}_{i\in B}$, $|B| \approx \epsilon n$

Goal: estimate "true" mean $\mu(\mathcal{D})$

Setting

Robust mean estimation

$$\{X_i^\star\}_{i\in[n]} \sim_{\text{i.i.d.}} \mathcal{D}$$

$$\{X_i = X_i^\star\}_{i\in G}$$

$$\{X_i\}_{i\in B}\,,\ |B| \approx \epsilon n$$

Goal: estimate $\mu^\star = \mu(\mathcal{D})$

Setting

$$\frac{1}{|G|}\sum_{i\in G}(X_i - \mu^\star)(X_i - \mu^\star)^\top \preceq \mathbf{I}_d$$

Meta-algo

Robust mean estimation

$$\{X_i^\star\}_{i\in[n]} \sim_{\text{i.i.d.}} \mathcal{D}$$

$$\{X_i = X_i^\star\}_{i\in G}$$

$$\{X_i\}_{i\in B}\,,\ |B| \approx \epsilon n$$

Goal: estimate $\mu^\star = \mu(\mathcal{D})$

Setting

$$\frac{1}{|G|}\sum_{i\in G}(X_i - \mu^\star)(X_i - \mu^\star)^\top \preceq \mathbf{I}_d$$

Return empirical mean μ_w if:

$$\sum_{i\in[n]} w_i(X_i - \mu_w)(X_i - \mu_w)^\top \preceq O(1)\mathbf{I}_d$$

$$\sum_{i\in[n]} w_i \leq 1,\ \sum_{i\in G} w_i \geq 1 - O(\epsilon)$$

Meta-algo

Robust mean estimation

$$\{X_i^\star\}_{i\in[n]} \sim_{\text{i.i.d.}} \mathcal{D}$$

$$\{X_i = X_i^\star\}_{i\in G}$$

$$\{X_i\}_{i\in B}\,,\ |B| \approx \epsilon n$$

Goal: estimate $\mu^\star = \mu(\mathcal{D})$

Setting

$$\frac{1}{|G|}\sum_{i\in G}(X_i - \mu^\star)(X_i - \mu^\star)^\top \preceq \mathbf{I}_d$$

Return empirical mean μ_w if:

$$\sum_{i\in[n]} w_i(X_i - \mu_w)(X_i - \mu_w)^\top \preceq O(1)\mathbf{I}_d$$

$$\sum_{i\in[n]} w_i \le 1,\ \sum_{i\in G} w_i \ge 1 - O(\epsilon)$$

Meta-algo

Invariant:
"saturation"

Robust mean estimation

$$\{X_i^\star\}_{i\in[n]} \sim_{\text{i.i.d.}} \mathcal{D}$$

$$\{X_i = X_i^\star\}_{i\in G}$$

$$\{X_i\}_{i\in B}\,,\ |B| \approx \epsilon n$$

Goal: estimate $\mu^\star = \mu(\mathcal{D})$

Setting

$$\frac{1}{|G|}\sum_{i\in G}(X_i - \mu^\star)(X_i - \mu^\star)^\top \preceq \mathbf{I}_d$$

Else:

$$\exists \mathbf{X} \in \mathrm{PSD}^{d\times d} : \mathrm{Tr}\mathbf{X} = 1$$

$$\mathbb{E}_{i\sim w}\left\langle (X_i - \mu_w)(X_i - \mu_w)^\top, \mathbf{X}\right\rangle \gg 1$$

Meta-algo

Robust mean estimation

$$\{X_i^\star\}_{i\in[n]} \sim_{\text{i.i.d.}} \mathcal{D}$$

$$\{X_i = X_i^\star\}_{i\in G}$$

$$\{X_i\}_{i\in B}\,,\ |B| \approx \epsilon n$$

Goal: estimate $\mu^\star = \mu(\mathcal{D})$

Setting

$$\frac{1}{|G|}\sum_{i\in G}(X_i - \mu^\star)(X_i - \mu^\star)^\top \preceq \mathbf{I}_d$$

Else:

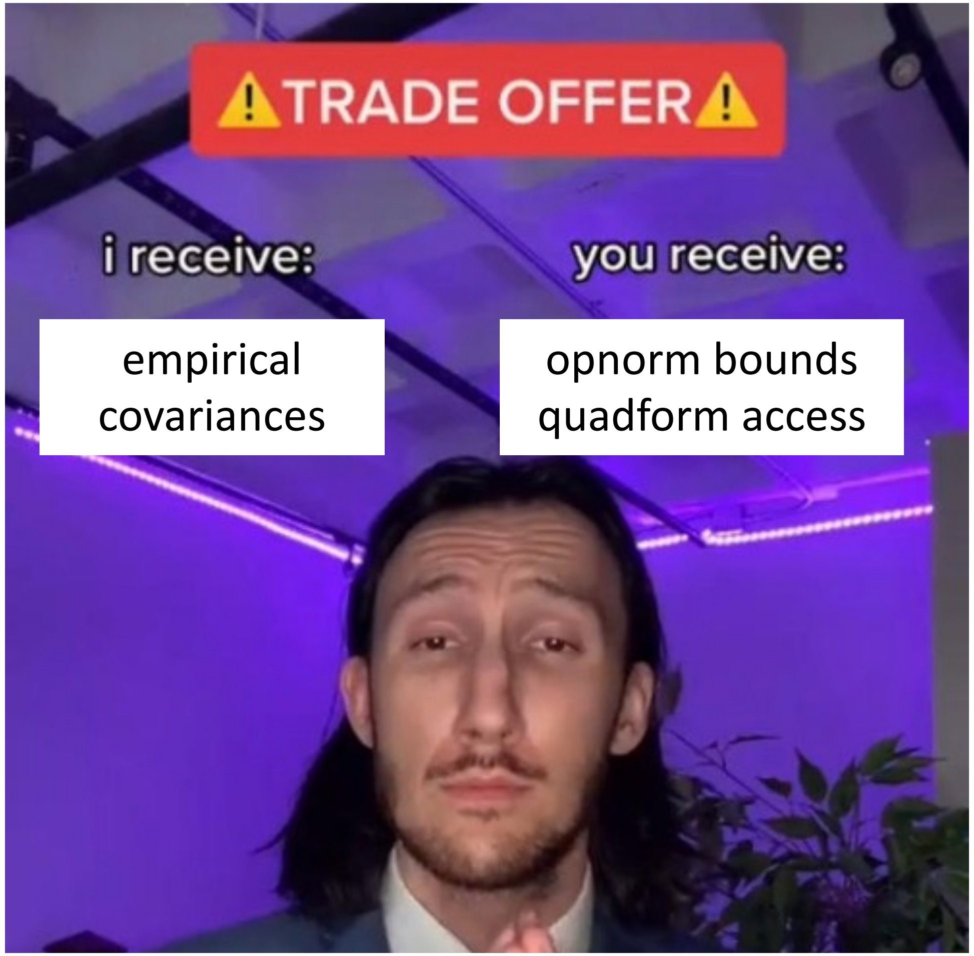
$$\exists \mathbf{X} \in \mathrm{PSD}^{d\times d} : \mathrm{Tr}\mathbf{X} = 1$$

$$\mathbb{E}_{i\sim w}\left\langle (X_i - \mu_w)(X_i - \mu_w)^\top, \mathbf{X}\right\rangle \gg 1$$

Many fast ways of preserving saturation

Meta-algo

Robust mean estimation



$$\frac{1}{|G|}\sum_{i\in G}(X_i-\mu^\star)(X_i-\mu^\star)^\top \preceq \mathbf{I}_d$$

Else:

$$\exists \mathbf{X} \in \mathrm{PSD}^{d\times d} : \mathrm{Tr}\mathbf{X} = 1$$

$$\mathbb{E}_{i\sim w}\left\langle (X_i-\mu_w)(X_i-\mu_w)^\top, \mathbf{X}\right\rangle \gg 1$$

Many fast ways of preserving saturation

Meta-algo

Robust mean estimation

$$\max_{\substack{\mathbf{X}^\star \in \mathrm{PSD}^{d\times d} \\ \mathrm{Tr}(\mathbf{X}^\star)=1}} \sum_{t\in[T]} \langle \mathbf{G}_t, \mathbf{X}^\star - \mathbf{X}_t \rangle \lesssim \frac{1}{\eta} + \eta G \sum_{t\in[T]} \langle \mathbf{G}_t, \mathbf{X}_t \rangle$$

Robust mean estimation

$$\max_{\substack{\mathbf{X}^\star \in \mathrm{PSD}^{d\times d} \\ \mathrm{Tr}(\mathbf{X}^\star)=1}} \sum_{t\in[T]} \langle \mathbf{G}_t, \mathbf{X}^\star - \mathbf{X}_t \rangle \lesssim \frac{1}{\eta} + \eta G \sum_{t\in[T]} \langle \mathbf{G}_t, \mathbf{X}_t \rangle$$

$$\left\| \sum_{t\in[T]} \mathbf{G}_t \right\|_{\mathrm{op}} \lesssim G + 2 \sum_{t\in[T]} \langle \mathbf{G}_t, \mathbf{X}_t \rangle$$

Robust mean estimation

$$\mathbf{G}_t = \sum_{i \in [n]} [w_t]_i \left(X_i - \mu_{w_t}\right)\left(X_i - \mu_{w_t}\right)^\top \qquad \mathbf{G}_0 \preceq G_0 \mathbf{I}_d$$

$$G_0 \gg 1$$

$$\left\| \sum_{t \in [T]} \mathbf{G}_t \right\|_{\mathrm{op}} \lesssim G_0 + 2 \sum_{t \in [T]} \langle \mathbf{G}_t, \mathbf{X}_t \rangle$$

Robust mean estimation

$$\mathbf{G}_t = \sum_{i \in [n]} [w_t]_i \left(X_i - \mu_{w_t}\right)\left(X_i - \mu_{w_t}\right)^\top$$

$$\mathbf{G}_0 \preceq G_0 \mathbf{I}_d$$

$$G_0 \gg 1$$

$$\left\|\sum_{t \in [T]} \mathbf{G}_t\right\|_{\mathrm{op}} \lesssim G_0 + 2\sum_{t \in [T]} \langle \mathbf{G}_t, \mathbf{X}_t \rangle$$

$$\lesssim G_0 + O(T)$$

(filter in each iteration)

Robust mean estimation

$$\mathbf{G}_t = \sum_{i\in[n]} [w_t]_i \left(X_i - \mu_{w_t}\right)\left(X_i - \mu_{w_t}\right)^\top \qquad \mathbf{G}_0 \preceq G_0 \mathbf{I}_d$$

$$G_0 \gg 1$$

$$T\,\|\mathbf{G}_T\|_{\mathrm{op}} \lesssim \left\|\sum_{t\in[T]} \mathbf{G}_t\right\|_{\mathrm{op}} \lesssim G_0 + 2\sum_{t\in[T]} \langle \mathbf{G}_t, \mathbf{X}_t\rangle$$

monotone
feedbacks

$$\lesssim G_0 + O(T)$$

Robust mean estimation

$$\mathbf{G}_t = \sum_{i \in [n]} [w_t]_i \left(X_i - \mu_{w_t}\right)\left(X_i - \mu_{w_t}\right)^\top \qquad \mathbf{G}_0 \preceq G_0 \mathbf{I}_d$$

$$T \left\|\mathbf{G}_T\right\|_{\mathrm{op}} \lesssim \left\| \sum_{t \in [T]} \mathbf{G}_t \right\|_{\mathrm{op}} \lesssim G_0 + 2 \sum_{t \in [T]} \langle \mathbf{G}_t, \mathbf{X}_t \rangle$$

$$\lesssim G_0 + O(T)$$

monotone
feedbacks

$$T \lesssim 1$$

$$\left\|\mathbf{G}_T\right\|_{\mathrm{op}} \leq \frac{G_0}{2}$$

Punchline

Robust mean estimation

$$\mathbf{G}_t = \sum_{i \in [n]} [w_t]_i \left(X_i - \mu_{w_t}\right) \left(X_i - \mu_{w_t}\right)^\top \qquad \mathbf{G}_0 \preceq G_0 \mathbf{I}_d$$

Interpretation:

MMW as a multi-directional filter

$$T \lesssim 1$$

$$\|\mathbf{G}_T\|_{\text{op}} \leq \frac{G_0}{2}$$

Punchline

Robust mean estimation

$$\mathbf{G}_t = \sum_{i \in [n]} [w_t]_i \left(X_i - \mu_{w_t}\right)\left(X_i - \mu_{w_t}\right)^\top \qquad \mathbf{G}_0 \preceq G_0 \mathbf{I}_d$$

Interpretation:

MMW as a multi-directional filter

…[DHL '19] Robust mean estimation in time $\tilde{O}(nd)$

$$T \lesssim 1$$

$$\|\mathbf{G}_T\|_{\mathrm{op}} \leq \frac{G_0}{2}$$

Punchline

Relatives of MMW

$$\mathrm{Tr}\left(\mathbf{Y}^p\right) \propto \max_{\substack{\mathbf{X}\in \mathrm{Sym}^{d\times d} \\ \|\mathbf{X}\|_q \leq 1}} \langle \mathbf{X}, \mathbf{Y} \rangle$$

as a potential

Relatives of MMW

$$\mathrm{Tr}\left(\mathbf{Y}^p\right) \propto \max_{\substack{\mathbf{X}\in\mathrm{Sym}^{d\times d}\\ \|\mathbf{X}\|_q\le 1}} \langle \mathbf{X}, \mathbf{Y}\rangle$$

as a potential

Upshot:

- Single-iteration progress (MMW non-monotone)
  - Multifilter [DKKLT '22], list-decoding

Relatives of MMW

$$\mathrm{Tr}\left(\mathbf{Y}^p\right) \propto \max_{\substack{\mathbf{X}\in \mathrm{Sym}^{d\times d} \\ \|\mathbf{X}\|_q \le 1}} \langle \mathbf{X}, \mathbf{Y} \rangle$$

as a potential

Upshot:

- Single-iteration progress (MMW non-monotone)
  - Multifilter [DKKLT '22], list-decoding
- More natural interpretation?
  - Power method [DKKP '23], PCA

Relatives of MMW

$$\mathrm{Tr}\left(\mathbf{Y}^p\right) \propto \max_{\substack{\mathbf{X}\in\mathrm{Sym}^{d\times d} \\ \|\mathbf{X}\|_q\leq 1}} \langle \mathbf{X}, \mathbf{Y} \rangle$$

as a potential

Downside(?)

- Less obvious connection to regret minimization
- (Does not apply to Daniel Kane)

Relatives of MMW

$$\mathrm{Tr}\left(\mathbf{Y}^p\right) \propto \max_{\substack{\mathbf{X}\in\mathrm{Sym}^{d\times d} \\ \|\mathbf{X}\|_q \le 1}} \langle \mathbf{X}, \mathbf{Y} \rangle$$

as a potential

Downside(?)

- Less obvious connection to regret minimization
- Suggest: mirror descent as a catch-all
- Smarter filters for specific problem

Relatives of MMW

$$\min_{\substack{w \in \mathbb{R}^n_{\geq 0} \\ \|w\|_1 \leq 1}} \left\| \sum_{i \in [n]} w_i \left(X_i - \mu_w\right) \left(X_i - \mu_w\right)^\top \right\|_{\text{op}}$$

Packing SDP

Relatives of MMW

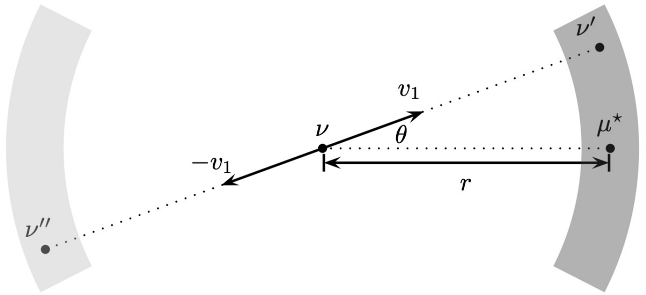
$$\min_{\substack{w \in \mathbb{R}^n_{\geq 0} \\ \|w\|_1 \leq 1}} \left\| \sum_{i \in [n]} w_i \left(X_i - \mu_w\right)\left(X_i - \mu_w\right)^\top \right\|_{\mathrm{op}}$$

Use case: local reweightings (e.g. gradient descent)

Iterative methods: *O*(1) approx. is OK [PSBR '18, CDG '19, …]

Relatives of MMW

$$\min_{\substack{w \in \mathbb{R}^n_{\geq 0} \\ \|w\|_1 \leq 1}} \left\| \sum_{i \in [n]} w_i \left(X_i - \mu_w\right)\left(X_i - \mu_w\right)^\top \right\|_{\mathrm{op}}$$



value = step size
dual = descent direction

Very general strategy for stochastic optimization problems…

Relatives of MMW

$$\mathcal{X} = \{\mathbf{X} \in \mathrm{PSD}^{d \times d} \mid \|\mathbf{X}\|_{\mathrm{op}} \le 1, \mathrm{TrX} \le k\}$$

“Fantope” = cvx hull of projection matrices

Relatives of MMW

$$\mathcal{X} = \left\{ \mathbf{X} \in \mathrm{PSD}^{d \times d} \mid \|\mathbf{X}\|_{\mathrm{op}} \leq 1, \mathrm{Tr}\mathrm{X} \leq k \right\}$$

“Fantope” = cvx hull of projection matrices

Multi-direction filters:
list-decoding [DKKLT ‘21], optimal Huber contamination [DKPP ‘23]

Relatives of MMW

$$\mathbf{G} = \frac{1}{\kappa} \sum_{i \in [n]} w_i \mathbf{A}_i$$

$$\sum_{i \in [n]} w_i \mathbf{A}_i \preceq \mathbf{I}_d$$

e.g. solution to a packing SDP

Relatives of MMW

$$\mathbf{G} = \frac{1}{\kappa} \sum_{i \in [n]} w_i \mathbf{A}_i$$

$$\frac{O(1)}{\kappa} \mathbf{I}_d \preceq \sum_{i \in [n]} \bar{w}_i \mathbf{A}_i \preceq \mathbf{I}_d$$

$$\sum_{i \in [n]} w_i \mathbf{A}_i \preceq \mathbf{I}_d$$

Regret minimization: two-sided constraints

Relatives of MMW

$$\mathbf{G} = \frac{1}{\kappa} \sum_{i \in [n]} w_i \mathbf{A}_i$$

$$\frac{O(1)}{\kappa} \mathbf{I}_d \preceq \sum_{i \in [n]} \bar{w}_i \mathbf{A}_i \preceq \mathbf{I}_d$$

$$\sum_{i \in [n]} w_i \mathbf{A}_i \preceq \mathbf{I}_d$$

Regret minimization: two-sided constraints

e.g. planted well-conditioning, semi-random linear models [JLMSST '23]

Thank you!

Contact
kjtian.github.io
kjtian@cs.utexas.edu

$$\varphi(\mathbf{X}) = \sum_{i \in [d]} \lambda_i(\mathbf{X}) \log \lambda_i(\mathbf{X})$$