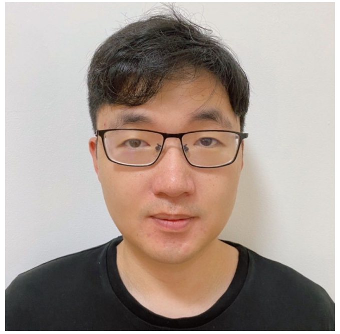
Omnipredicting Single-Index Models with Multi-Index Models

Kevin Tian (UT Austin)

Simons Institute IFML/MPG Symposium

Based on joint work with:



Lunjia Hu (Harvard → Northeastern), Chutong Yang (UT Austin)

Roadmap

- Overview
  - Omniprediction
  - SIMs
  - Our results
- Isotron
  - Realizable setting
  - Agnostic setting
- Efficient omniprediction
  - Sample complexity
  - Runtime complexity

Loss-based supervised learning

$$\boxed{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(p(\mathbf{x}),y)\right] \leq \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon}$$

…traditional paradigm in supervised learning…

Loss-based supervised learning

$$\boxed{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(p(\mathbf{x}),y)\right] \leq \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon}$$

The main characters:

- Distribution $\mathcal{D}$ over $\{\mathbf{x} \in \mathbb{R}^d \,|\, \|\mathbf{x}\| \leq 1\}$ x $\{0,1\}$

Loss-based supervised learning

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(p(\mathbf{x}),y)\right] \leq \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon$$

The main characters:

- Distribution $\mathcal{D}$ over $\{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\| \leq 1\}$ x $\{0,1\}$
- Comparator class $\mathcal{C}$ (e.g., linear functions)

Loss-based supervised learning

$$\boxed{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(p(\mathbf{x}),y)\right] \leq \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon}$$

The main characters:

- Distribution $\mathcal{D}$ over $\{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\| \leq 1\}$ x $\{0,1\}$
- Comparator class $\mathcal{C}$ (e.g., linear functions)
- Loss function ℓ (e.g., squared loss, cross entropy, …)

Loss-based supervised learning

$$\boxed{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(p(\mathbf{x}),y)\right] \le \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon}$$

The main characters:

- Distribution $\mathcal{D}$ over $\{\mathbf{x} \in \mathbb{R}^d \,|\, \|\mathbf{x}\| \le 1\}$ x $\{0,1\}$
- Comparator class $\mathcal{C}$ (e.g., linear functions)
- Loss function ℓ (e.g., squared loss, cross entropy, …)
- Predictor p : can be *proper* *(*$p \in \mathcal{C}$*)* or *improper*

Loss-based supervised learning

$$\boxed{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(p(\mathbf{x}),y)\right] \leq \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon}$$

What if loss not known in advance?

- Depends on parameters unknown at training

Loss-based supervised learning

$$\boxed{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(p(\mathbf{x}),y)\right] \leq \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon}$$

What if loss not known in advance?

- Depends on parameters unknown at training
- Multiple tasks (e.g., weights of false pos/neg)

Loss-based supervised learning

$$\boxed{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(p(\mathbf{x}),y)\right] \le \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon}$$

What if loss not known in advance?

- Depends on parameters unknown at training
- Multiple tasks (e.g., weights of false pos/neg)
- "Fundamental truth" of $\mathcal{D}$ independent of loss
  - Drives us closer to ground truth $p^\star(\mathbf{x}) \coloneqq \mathbb{E}[y|\mathbf{x}]$

Omniprediction

$$\boxed{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(k_\ell(p(\mathbf{x})),y)\right] \le \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon \qquad \forall \ell \in \mathcal{L}}$$

[Gopalan-Kalai-Reingold-Sharan-Wieder '22]

Omniprediction

$$\boxed{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(k_\ell(p(\mathbf{x})),y)\right] \leq \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon \qquad \forall \ell \in \boldsymbol{\mathcal{L}}}$$

The additional characters:

- Loss function *family* $\mathcal{L}$ (e.g., proper losses)

Omniprediction

$$\boxed{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(k_\ell(p(\mathbf{x})),y)\right] \leq \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon \qquad \forall \ell \in \mathcal{L}}$$

The additional characters:

- Loss function *family* $\mathcal{L}$ (e.g., proper losses)
- Loss-specific post-processings $\{k_\ell\}_{\ell\in\mathcal{L}}$
  - Distribution-independent
  - Role of p: “supervised sufficient statistics” for $\mathcal{D}$

Omniprediction

$$\boxed{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(k_\ell(p(\mathbf{x})),y)\right] \leq \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon \qquad \forall \ell \in \mathcal{L}}$$

The additional characters:

- Loss function *family* $\mathcal{L}$ (e.g., proper losses)
- Loss-specific post-processings $\{k_\ell\}_{\ell\in\mathcal{L}}$
  - Distribution-independent
  - Role of p: "supervised sufficient statistics" for $\mathcal{D}$
- Fundamentally an *agnostic learning* guarantee!

Omniprediction recipes

$$\boxed{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(k_\ell(p(\mathbf{x})),y)\right] \leq \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon \qquad \forall \ell \in \mathcal{L}}$$

Idea 1: multicalibration suffices [GKRSW '22]

Omniprediction recipes

$$\boxed{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(k_\ell(p(\mathbf{x})),y)\right] \leq \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon \qquad \forall \ell \in \mathcal{L}}$$

Idea 1: multicalibration suffices [GKRSW '22]

- Powerful property: agrees with ground truth on parameterized *conditional dists*

Omniprediction recipes

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(k_\ell(p(\mathbf{x})), y)\right] \le \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x}, y))\right] + \epsilon \qquad \forall \ell \in \mathcal{L}$$

Idea 1: multicalibration suffices [GKRSW '22]

- Powerful property: agrees with ground truth on parameterized *conditional dists*
- Reduce from *agnostically learning $\mathcal{C}$* via iterative boosting [Hébert-Johnson-Kim-Reingold-Rothblum '18]

Omniprediction recipes

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(k_\ell(p(\mathbf{x})),y)\right] \le \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon \qquad \forall \ell \in \mathcal{L}$$

Idea 1: multicalibration suffices [GKRSW '22]

- Powerful property: agrees with ground truth on parameterized *conditional dists*
- Reduce from *agnostically learning* $\mathcal{C}$ via iterative boosting [Hébert-Johnson-Kim-Reingold-Rothblum '18]
- Computationally-intractable in many settings… (e.g. halfspaces [Guruswami-Raghavendra '06, …])

Omniprediction recipes

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(k_\ell(p(\mathbf{x})),y)\right] \leq \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon \qquad \forall \ell \in \mathcal{L}$$

Idea 2: weaker conditions suffice

[Gopalan-Hu-Kim-Reingold-Wieder '23]

- "Statistical tests" parameterized by $\mathcal{C}\times\mathcal{L}$

Omniprediction recipes

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(k_\ell(p(\mathbf{x})),y)\right] \le \min_{c\in\mathcal{C}}\left[\ell(c(\mathbf{x},y))\right] + \epsilon \qquad \forall \ell \in \mathcal{L}$$

Idea 2: weaker conditions suffice

[Gopalan-Hu-Kim-Reingold-Wieder '23]

- "Statistical tests" parameterized by $\mathcal{C}\times\mathcal{L}$
- Calibration + "multi-accuracy" suffice: improved quantitative bounds for explicit families

Single-index models

$$\boxed{\mathbb{E}\left[y \mid \mathbf{x}\right] \approx \sigma(\mathbf{w} \cdot \mathbf{x})}$$

Single-index models

$$\boxed{\mathbb{E}\left[y \mid \mathbf{x}\right] \approx \sigma\left(\mathbf{w} \cdot \mathbf{x}\right)}$$

"Semi-parametric" model family

- Parametric: *linear predictor* $\mathbf{w} \in \mathcal{W} := \{\mathbf{w} \in \mathbb{R}^d \mid ||\mathbf{w}|| \leq 1\}$

Single-index models

$$\boxed{\mathbb{E}\left[y \mid \mathbf{x}\right] \approx \sigma(\mathbf{w} \cdot \mathbf{x})}$$

"Semi-parametric" model family

- Parametric: *linear predictor* $\mathbf{w} \in \mathcal{W} := \{\mathbf{w} \in \mathbb{R}^d \,|\, \|\mathbf{w}\| \leq 1\}$
- Non-parametric: *link function* $\sigma \in \mathcal{S} :=$ β-Lipschitz, monotone functions $\sigma: [-1,1] \rightarrow [0,1]$
  - Known link: generalized linear model

Single-index models

Every link σ has…

…induced *matching loss*

$$\boxed{\ell_{\mathsf{m},\sigma}(t, y) := \int_0^t (\sigma(\tau) - y)\mathrm{d}\tau}$$

Single-index models

Every link σ has…

…induced *matching loss*

$$\boxed{\ell_{\mathsf{m},\sigma}(t, y) := \int_0^t (\sigma(\tau) - y)\mathrm{d}\tau}$$

$$\frac{\partial}{\partial t}\ell_{\mathsf{m},\sigma}(t, y) = \sigma(t) - y$$

(convex)

Single-index models

Every link σ has…

…induced *matching loss*

$$\boxed{\ell_{\mathsf{m},\sigma}(t,y) := \int_0^t (\sigma(\tau) - y)\mathrm{d}\tau}$$

$$\sigma^{-1}(\mathbb{E}[y]) \in \mathrm{argmin}_t \left\{\ell_{\mathsf{m},\sigma}(t,y)\right\} \quad \Longleftarrow \quad \frac{\partial}{\partial t}\ell_{\mathsf{m},\sigma}(t,y) = \sigma(t) - y$$

(convex)

Single-index models

Every link σ has…

…induced *matching loss*

$$\ell_{\mathsf{m},\sigma}(t, y) := \int_0^t (\sigma(\tau) - y)\mathrm{d}\tau$$

…induced *proper loss*

$$\ell_{\mathsf{p},\sigma}(v, y) := \ell_{\mathsf{m},\sigma}(\sigma^{-1}(v), y)$$

(if *y* ~ Bernoulli, minimized by ground truth)

Learning SIMs in squared loss

$$\sigma(\mathbf{w} \cdot \mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$$

Isotron learns SIMs!

[Kalai-Sastry '09, Kakade-Kalai-Kanade-Shamir '11]

realizable

Learning SIMs in squared loss

$$\sigma(\mathbf{w} \cdot \mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$$

Isotron learns SIMs!

[Kalai-Sastry '09, Kakade-Kalai-Kanade-Shamir '11]

- Very simple algo (gradient descent + isotonic regression)
- Proper hypotheses

realizable

Learning SIMs in squared loss

$$\sigma(\mathbf{w} \cdot \mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$$

Isotron learns SIMs!

[Kalai-Sastry ‘09, Kakade-Kalai-Kanade-Shamir ‘11]

- Very simple algo (gradient descent + isotonic regression)
- Proper hypotheses

realizable

$$\sigma(\mathbf{w} \cdot \mathbf{x}) \approx^{?} \mathbb{E}[y \mid \mathbf{x}]$$

“Constant-factor” learners

[Gollakota-Gopalan-Klivans-Stavropoulos ‘23, Zarifis-Wang-Diakonikolasx2 ‘24]

agnostic

Learning SIMs in squared loss

$$\sigma(\mathbf{w} \cdot \mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$$

Isotron learns SIMs!

[Kalai-Sastry '09, Kakade-Kalai-Kanade-Shamir '11]

- Very simple algo (gradient descent + isotonic regression)
- Proper hypotheses

realizable

$$\sigma(\mathbf{w} \cdot \mathbf{x}) \approx^{?} \mathbb{E}[y \mid \mathbf{x}]$$

"Constant-factor" learners

[Gollakota-Gopalan-Klivans-Stavropoulos '23, Zarifis-Wang-Diakonikolasx2 '24]

- More distributional assumptions, structure (e.g. bi-Lipschitz, anti-conc.)
- Very large overheads (e.g., [ZWDD24] needs $d\kappa^{44}$ samples)

agnostic

Omnipredicting SIMs

Our goal: for all $(\sigma, \mathbf{w}) \in \mathcal{S} \times \mathcal{W} \ldots$

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_{\mathsf{m},\sigma}\left(k_\sigma(p(\mathbf{x})), y\right)\right] \leq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_{\mathsf{m},\sigma}\left(\mathbf{w}\cdot\mathbf{x}, y\right)\right] + \epsilon$$

Omnipredicting SIMs

Our goal: for all $(\sigma, \mathbf{w}) \in \mathcal{S} \times \mathcal{W}$…

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_{\mathsf{m},\sigma}\left(k_\sigma(p(\mathbf{x})), y\right)\right] \leq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_{\mathsf{m},\sigma}\left(\mathbf{w}\cdot\mathbf{x}, y\right)\right] + \epsilon$$

…another view when *p* scalar…

$$\mathbb{E}_{(\mathbf{x},y)}\left[\ell_{\mathsf{p},\sigma}(p(\mathbf{x}), y)\right] \leq \mathbb{E}_{(\mathbf{x},y)}\left[\ell_{\mathsf{p},\sigma}(\sigma(\mathbf{w}\cdot\mathbf{x}), y)\right]$$

"competing with all (proper) SIMs"

Omnipredicting SIMs

Our goal: for all $(\sigma, \mathbf{w}) \in \mathcal{S} \times \mathcal{W} \ldots$

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_{\mathsf{m},\sigma}\left(k_\sigma(p(\mathbf{x})), y\right)\right] \leq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_{\mathsf{m},\sigma}\left(\mathbf{w}\cdot\mathbf{x}, y\right)\right] + \epsilon$$

Existing construction [GHKRW '23]:
$\approx \epsilon^{-10}$ samples

Iterate until MA and CAL:

- Calibrated residual
  - Bucket + estimate quantiles
- Multiaccurate residual
  - Repeated truncation + boosting

Omnipredicting SIMs

Our goal: for all $(\sigma, \mathbf{w}) \in \mathcal{S} \times \mathcal{W}$…

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_{\mathsf{m},\sigma}\left(k_\sigma(p(\mathbf{x})), y\right)\right] \le \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_{\mathsf{m},\sigma}\left(\mathbf{w}\cdot\mathbf{x}, y\right)\right] + \epsilon$$

Existing construction [GHKRW '23]:
$\approx \epsilon^{-10}$ samples

- Complex algo / hypothesis
  - Highly-improper (interpretability?)
  - Large sequential depth

Iterate until MA and CAL:

- Calibrated residual
  - Bucket + estimate quantiles
- Multiaccurate residual
  - Repeated truncation + boosting

Omnipredicting SIMs

Our goal: for all $(\sigma, \mathbf{w}) \in \mathcal{S} \times \mathcal{W} \ldots$

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_{\mathsf{m},\sigma}\left(k_\sigma(p(\mathbf{x})), y\right)\right] \le \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_{\mathsf{m},\sigma}\left(\mathbf{w}\cdot\mathbf{x}, y\right)\right] + \epsilon$$

Existing construction [GHKRW '23]:
$\approx \epsilon^{-10}$ samples

- Complex algo / hypothesis
  - Highly-improper (interpretability?)
  - Large sequential depth
- Loose sample / runtime complexity?

Iterate until MA and CAL:

- Calibrated residual
  - Bucket + estimate quantiles
- Multiaccurate residual
  - Repeated truncation + boosting

Philosophy

"TCS"-style results

- Challenging setup (agnostic, heterogeneous, nonconvex, …)
- Provable guarantees!!!
- Polynomial time / samples!!!

Philosophy

…okay, but what polynomial?

“TCS”-style results

- Challenging setup (agnostic, heterogeneous, nonconvex, …)
- Provable guarantees!!!
- Polynomial time / samples!!!

Philosophy

…okay, but what polynomial?

“TCS”-style results

- Challenging setup (agnostic, heterogeneous, nonconvex, …)
- Provable guarantees!!!
- Polynomial time / samples!!!

Strive for “right” algorithms, analyses in tractable settings to make impact on applications.

Our results

Theorem [HTY '24]: There is an omnipredictor for SIMs using…

- $\frac{\beta^2}{\epsilon^4}$ samples (for β-Lipschitz, monotone links)
- $\frac{\beta^2}{\alpha^2\epsilon^2}$ samples (for (α, β)-bi-Lipschitz links)

…in nearly-linear time $\tilde{O}(nd \cdot \epsilon^{-2})$

Our results

Theorem [HTY '24]: There is an omnipredictor for SIMs using…

- $\frac{\beta^2}{\epsilon^4}$ samples (for β-Lipschitz, monotone links)
- $\frac{\beta^2}{\alpha^2\epsilon^2}$ samples (for (α, β)-bi-Lipschitz links)

…with the "multi-index model" form

$$p(\mathbf{x}) = \{\sigma_t(\mathbf{w}_t \cdot \mathbf{x})\}_{t \in [O(\epsilon^{-2})]}$$

Our results

Theorem [HTY '24]: There is an omnipredictor for SIMs using…

- $\frac{\beta^2}{\epsilon^4}$ samples (for β-Lipschitz, monotone links)
- $\frac{\beta^2}{\alpha^2\epsilon^2}$ samples (for (α, β)-bi-Lipschitz links)

The algo is Isotron with custom iso-reg solver + post-processing.

We call it the *Omnitron*.

Our results

“ERM” omniprediction

Better time / sample complexities (second moment bound)

Same results existentially hold for population-level omniprediction

Omnipredicting SIMs in $\mathbb{R}$

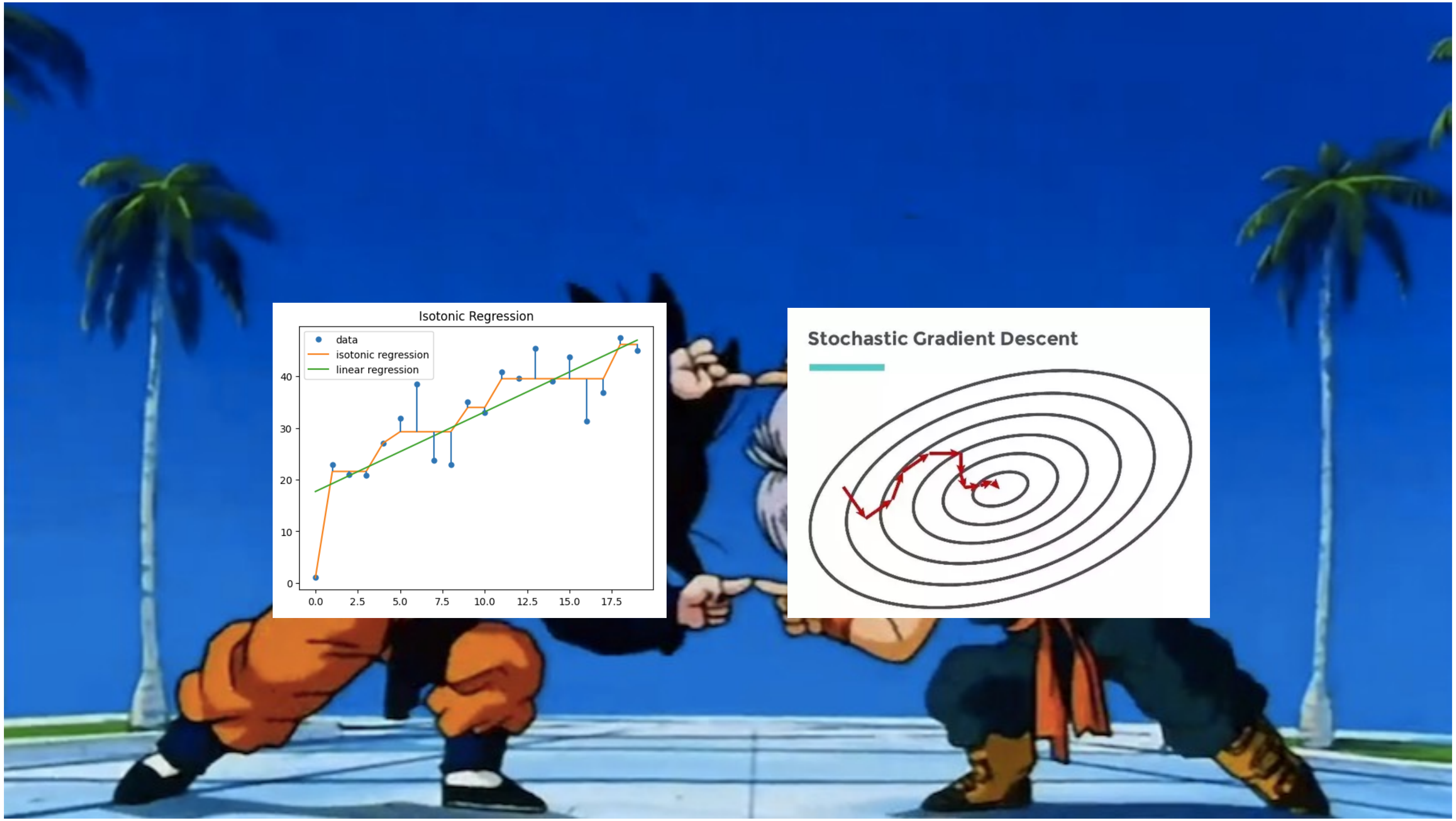
Theorem [HTY ‘24]: Two SIMs suffice (“double-index model”)

Open Q: is there a proper omnipredictor, even in 1-d?

Roadmap

- Overview
  - Omniprediction
  - SIMs
  - Our results
- Isotron
  - Realizable setting
  - Agnostic setting
- Efficient omniprediction
  - Sample complexity
  - Runtime complexity

Isotron



Isotonic Regression
data
isotonic regression
linear regression
0
10
20
30
40
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Stochastic Gradient Descent

Isotron

Algorithm 1: $\mathsf{Isotron}(\mathcal{D}, T, \eta)$

Input: Distribution $\mathcal{D}$ from Model 2, iteration count $T \in \mathbb{N}$, step size $\eta > 0$

$\mathbf{w}_0 \leftarrow \mathbf{0}_d$

for $0 \le t < T$ **do**

  $\sigma_t \leftarrow \arg\min_{\sigma \in \mathcal{S}_{0,\beta}} \{\ell_{\mathsf{sq}}(\sigma, \mathbf{w}_t; \mathcal{D})\}$

  $\mathbf{w}_{t+1} \leftarrow \mathbf{\Pi}_{\mathcal{W}}(\mathbf{w}_t - \eta \nabla_{\mathbf{w}} \ell_{\mathsf{m},\sigma_t}(\mathbf{w}_t; \mathcal{D}))$

end

return $\{\sigma_t\}_{0 \le t \le T-1}, \{\mathbf{w}_t\}_{0 \le t \le T}$

Isotron

Algorithm 1: $\mathsf{Isotron}(\mathcal{D}, T, \eta)$

Input: Distribution $\mathcal{D}$ from Model 2, iteration count $T \in \mathbb{N}$, step size $\eta > 0$

$\mathbf{w}_0 \leftarrow \mathbf{0}_d$

for $0 \leq t < T$ **do**

  $\sigma_t \leftarrow \arg\min_{\sigma \in \mathcal{S}_{0,\beta}} \{\ell_{\mathsf{sq}}(\sigma, \mathbf{w}_t; \mathcal{D})\}$

  $\mathbf{w}_{t+1} \leftarrow \mathbf{\Pi}_{\mathcal{W}}(\mathbf{w}_t - \eta \nabla_{\mathbf{w}} \ell_{\mathsf{m},\sigma_t}(\mathbf{w}_t; \mathcal{D}))$

end

return $\{\sigma_t\}_{0 \leq t \leq T-1}, \{\mathbf{w}_t\}_{0 \leq t \leq T}$

$$\ell_{\mathsf{sq}}(\sigma, \mathbf{w}; \mathcal{D}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2\right]$$

$$\nabla_{\mathbf{w}} \ell_{\mathsf{m},\sigma}(\mathbf{w}; \mathcal{D}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\cdot\mathbf{x}\right]$$

Isotron analysis (realizable setting)

Idea 1: regret minimization

$$\frac{1}{T}\sum_{0\leq t<T}\langle \nabla_{\mathbf{w}}\ell_{\mathsf{m},\sigma_t}(\mathbf{w}_t;\mathcal{D}),\mathbf{w}_t-\mathbf{w}\rangle\leq\epsilon$$

Isotron analysis (realizable setting)

Idea 2: optimality of iso-reg

$$\langle \nabla_{\mathbf{w}} \ell_{\mathsf{m},\sigma_t}(\mathbf{w}_t; \mathcal{D}), \mathbf{w}_t - \mathbf{w} \rangle = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\mathbf{w}_t \cdot \mathbf{x} - \mathbf{w} \cdot \mathbf{x}) \right]$$

Isotron analysis (realizable setting)

Idea 2: optimality of iso-reg

$$
\begin{aligned}
\langle \nabla_{\mathbf{w}} \ell_{\mathsf{m},\sigma_t}(\mathbf{w}_t;\mathcal{D}), \mathbf{w}_t - \mathbf{w}\rangle &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma_t(\mathbf{w}_t\cdot\mathbf{x}) - y)(\mathbf{w}_t\cdot\mathbf{x} - \mathbf{w}\cdot\mathbf{x})\right] \\
&= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma_t(\mathbf{w}_t\cdot\mathbf{x}) - y)(\mathbf{w}_t\cdot\mathbf{x} - \sigma^{-1}(\mathbf{w}_t\cdot\mathbf{x}))\right] \\
&+ \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma_t(\mathbf{w}_t\cdot\mathbf{x}) - y)(\sigma^{-1}(\mathbf{w}_t\cdot\mathbf{x}) - \mathbf{w}\cdot\mathbf{x})\right]
\end{aligned}
$$

Isotron analysis (realizable setting)

Idea 2: optimality of iso-reg

$$\begin{aligned}
\langle \nabla_{\mathbf{w}} \ell_{\mathrm{m},\sigma_t}(\mathbf{w}_t; \mathcal{D}), \mathbf{w}_t - \mathbf{w} \rangle &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\mathbf{w}_t \cdot \mathbf{x} - \mathbf{w} \cdot \mathbf{x}) \right] \\
&= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\mathbf{w}_t \cdot \mathbf{x} - \sigma^{-1}(\mathbf{w}_t \cdot \mathbf{x})) \right] \\
&\quad \boxed{+\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\sigma^{-1}(\mathbf{w}_t \cdot \mathbf{x}) - \mathbf{w} \cdot \mathbf{x}) \right]}
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\sigma^{-1}(\mathbf{w}_t \cdot \mathbf{x}) - \mathbf{w} \cdot \mathbf{x}) \right] \\
&\geq \mathbb{E}_{\mathbf{x}} \left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - \sigma(\mathbf{w} \cdot \mathbf{x}))^2 \right]
\end{aligned}$$

(excess squared loss)

Isotron analysis (realizable setting)

Idea 2: optimality of iso-reg

$$\langle \nabla_{\mathbf{w}} \ell_{\mathrm{m},\sigma_t}(\mathbf{w}_t; \mathcal{D}), \mathbf{w}_t - \mathbf{w} \rangle = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\mathbf{w}_t \cdot \mathbf{x} - \mathbf{w} \cdot \mathbf{x}) \right]$$

$$\boxed{= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\mathbf{w}_t \cdot \mathbf{x} - \sigma^{-1}(\mathbf{w}_t \cdot \mathbf{x})) \right]}$$

$$+ \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\sigma^{-1}(\mathbf{w}_t \cdot \mathbf{x}) - \mathbf{w} \cdot \mathbf{x}) \right]$$

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\mathbf{w}_t \cdot \mathbf{x} - \sigma^{-1}(\mathbf{w}_t \cdot \mathbf{x})) \right] = 0$$

(iso-reg solution calibrated)

Isotron analysis (realizable setting)

Idea 2: optimality of iso-reg

$$
\begin{aligned}
\langle \nabla_{\mathbf{w}} \ell_{\mathsf{m},\sigma_t}(\mathbf{w}_t; \mathcal{D}), \mathbf{w}_t - \mathbf{w} \rangle &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\mathbf{w}_t \cdot \mathbf{x} - \mathbf{w} \cdot \mathbf{x}) \right] \\
&= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\mathbf{w}_t \cdot \mathbf{x} - \sigma^{-1}(\mathbf{w}_t \cdot \mathbf{x})) \right] \\
&+ \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\sigma^{-1}(\mathbf{w}_t \cdot \mathbf{x}) - \mathbf{w} \cdot \mathbf{x}) \right] \\
&\geq \ell_{\mathsf{sq}}(\sigma_t, \mathbf{w}_t; \mathcal{D}) - \ell_{\mathsf{sq}}(\sigma, \mathbf{w}; \mathcal{D})
\end{aligned}
$$

…some iterate is good, i.e., proper learner

Isotron analysis (agnostic setting)

$$\langle \nabla_{\mathbf{w}} \ell_{\mathrm{m},\sigma_t}(\mathbf{w}_t; \mathcal{D}), \mathbf{w}_t - \mathbf{w} \rangle = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\mathbf{w}_t \cdot \mathbf{x} - \mathbf{w} \cdot \mathbf{x})\right]$$

$$\boxed{= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\mathbf{w}_t \cdot \mathbf{x} - \sigma^{-1}(\mathbf{w}_t \cdot \mathbf{x}))\right]}$$

$$+\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\sigma^{-1}(\mathbf{w}_t \cdot \mathbf{x}) - \mathbf{w} \cdot \mathbf{x})\right]$$

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma_t(\mathbf{w}_t \cdot \mathbf{x}) - y)(\sigma^{-1}(\mathbf{w}_t \cdot \mathbf{x}) - \mathbf{w} \cdot \mathbf{x})\right] \geq 0$$

…still OK by KKT conditions of *Lipschitz* iso-reg! [Lemma 1, KKKS '11]

Isotron analysis (agnostic setting)

$$\begin{aligned}\langle \nabla_{\mathbf{w}} \ell_{\mathrm{m},\sigma_t}(\mathbf{w}_t;\mathcal{D}), \mathbf{w}_t - \mathbf{w}\rangle &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma_t(\mathbf{w}_t\cdot\mathbf{x}) - y)(\mathbf{w}_t\cdot\mathbf{x} - \mathbf{w}\cdot\mathbf{x})\right] \\ &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma_t(\mathbf{w}_t\cdot\mathbf{x}) - y)(\mathbf{w}_t\cdot\mathbf{x} - \sigma^{-1}(\mathbf{w}_t\cdot\mathbf{x}))\right] \\ &\boxed{+\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma_t(\mathbf{w}_t\cdot\mathbf{x}) - y)(\sigma^{-1}(\mathbf{w}_t\cdot\mathbf{x}) - \mathbf{w}\cdot\mathbf{x})\right]}\end{aligned}$$

What about…

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma_t(\mathbf{w}_t\cdot\mathbf{x}) - y)(\mathbf{w}_t\cdot\mathbf{x} - \sigma^{-1}(\mathbf{w}_t\cdot\mathbf{x}))\right]$$

Omnigap

$$\mathsf{OG}(p) := \sup_{(\mathbf{w},\sigma)\in\mathcal{W}\times\mathcal{S}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(p(\mathbf{x})-y)(\sigma^{-1}(p(\mathbf{x}))-\mathbf{w}\cdot\mathbf{x})\right]$$

Omnigap

$$\mathsf{OG}(p) := \sup_{(\mathbf{w},\sigma)\in\mathcal{W}\times\mathcal{S}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(p(\mathbf{x})-y)(\sigma^{-1}(p(\mathbf{x}))-\mathbf{w}\cdot\mathbf{x})\right]$$

Ground truth p has zero omnigap!

Interpretation: small omnigap → passing many "indistinguishability" statistical tests

Omnigap

"calibration"

$$\mathrm{OG}(p) := \sup_{(\mathbf{w},\sigma)\in\mathcal{W}\times\mathcal{S}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(p(\mathbf{x})-y)(\sigma^{-1}(p(\mathbf{x}))-\mathbf{w}\cdot\mathbf{x})\right]$$

"multi-accuracy"

…turns out to be a one-sided variant of

"loss outcome indistinguishability" [GHKRW '23]

Omnigap

$$\mathsf{OG}(p) := \sup_{(\mathbf{w},\sigma)\in\mathcal{W}\times\mathcal{S}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(p(\mathbf{x})-y)(\sigma^{-1}(p(\mathbf{x}))-\mathbf{w}\cdot\mathbf{x})\right]$$

Theorem [implicit, GHKRW '23]:

$$\mathsf{OG}(p) \le \epsilon \implies p \text{ is an } \epsilon\text{-omnipredictor for SIMs}$$

Omnigap

Proof. Let $\widehat{\mathcal{D}}$ be the distribution on $\mathcal{X} \times \{0,1\}$ which draws $\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}$ and then $y \mid \mathbf{x} \sim \mathrm{Bern}(p(\mathbf{x}))$. We have that the following hold, because the integral part of Definition 1 cancels in each line:

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_{\mathsf{p},\sigma}(p(\mathbf{x}),y)\right] - \mathbb{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}\left[\ell_{\mathsf{p},\sigma}(p(\mathbf{x}),y)\right] = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(p(\mathbf{x})-y)\,\sigma^{-1}\left(p(\mathbf{x})\right)\right],$$
$$\mathbb{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}\left[\ell_{\mathsf{m},\sigma}(\mathbf{w}\cdot\mathbf{x},y)\right] - \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_{\mathsf{m},\sigma}(\mathbf{w}\cdot\mathbf{x},y)\right] = -\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(p(\mathbf{x})-y)\,(\mathbf{w}\cdot\mathbf{x})\right].$$

Moreover, by the definition of $\widehat{\mathcal{D}}$ (i.e., labels are $\sim \mathrm{Bern}(p(\mathbf{x}))$), because $\ell_{\mathsf{p},\sigma}$ is a proper loss,

$$E_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}\left[\ell_{\mathsf{p},\sigma}(p(\mathbf{x}),y)\right] - \mathbb{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}\left[\ell_{\mathsf{m},\sigma}(\mathbf{w}\cdot\mathbf{x},y)\right] \le 0.$$

Summing up the above displays, we obtain for any $(\sigma,\mathbf{w}) \in \mathcal{S}\times\mathcal{W}$,

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_{\mathsf{p},\sigma}(p(\mathbf{x}),y)\right] - \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell_{\mathsf{m},\sigma}(\mathbf{w}\cdot\mathbf{x},y)\right] &\le \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(p(\mathbf{x})-y)(\sigma^{-1}(p(\mathbf{x}))-\mathbf{w}\cdot\mathbf{x})] \\ &= \mathsf{OG}(p;\sigma,\mathbf{w}).\end{aligned} \tag{13}$$

Because $(\sigma,\mathbf{w})$ were arbitrary, by using $\mathsf{OG}(p)\le\varepsilon$, we have the claim. $\square$

Omnigap

$$\mathrm{OG}(p) := \sup_{(\mathbf{w},\sigma)\in\mathcal{W}\times\mathcal{S}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(p(\mathbf{x})-y)(\sigma^{-1}(p(\mathbf{x}))-\mathbf{w}\cdot\mathbf{x})\right]$$

Yields a new, simpler proof of PAV optimality in 1-d

(key ingredient in our 1-d omnipredictor construction)

Omnitron

Algorithm 3: Omnitron($\{(\mathbf{x}_t, y_t)\}_{0 \le t < T}, T, \eta, \mathcal{O}, \varepsilon$)

Input: $\{(\mathbf{x}_t, y_t)\}_{0 \le t < T} \sim_{\text{i.i.d.}} \mathcal{D}$ for a distribution $\mathcal{D}$ from Model 2, iteration count $T \in \mathbb{N}$, step size $\eta > 0$, ε-approximate BIR oracle $\mathcal{O}$ (Definition 6)

$\mathbf{w}_0 \leftarrow \mathbf{0}_d$

for $0 \le t < T$ **do**

  $\sigma_t \leftarrow \mathcal{O}(\mathbf{w}_t)$

  $\tilde{\mathbf{g}}_t \leftarrow (\sigma_t(\mathbf{w}_t \cdot \mathbf{x}_t) - y_t)\mathbf{x}_t$

  $\mathbf{w}_{t+1} \leftarrow \mathbf{\Pi}_{\mathcal{W}}(\mathbf{w}_t - \eta\tilde{\mathbf{g}}_t)$

end

return $p : \mathbf{x} \to \{\sigma_t(\mathbf{w}_t \cdot \mathbf{x})\}_{0 \le t \le T-1},\ k_\sigma : \{p_t\}_{0 \le t < T} \to \frac{1}{T}\sum_{0 \le t < T} \sigma^{-1}(p_t)$

(rest of talk)

(“adaptive” stochastic optimization)

Roadmap

- Overview
  - Omniprediction
  - SIMs
  - Our results
- Isotron
  - Realizable setting
  - Agnostic setting
- Efficient omniprediction
  - Sample complexity
  - Runtime complexity

Robust omniprediction

For input $\widehat{\mathbf{w}} \in \mathcal{W}$ return $\widehat{\sigma} \in \mathcal{S}$ s.t. $\text{for all } \sigma \in \mathcal{S}$

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\widehat{\sigma}(\widehat{\mathbf{w}} \cdot \mathbf{x}) - y)(\mathbf{w} \cdot \mathbf{x} - \sigma^{-1}(\widehat{\sigma}(\mathbf{w} \cdot \mathbf{x})))\right] \geq -\epsilon$$

Robust omniprediction

For input $\widehat{\mathbf{w}} \in \mathcal{W}$ return $\widehat{\sigma} \in \mathcal{S}$ s.t. $\text{for all } \sigma \in \mathcal{S}$

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\widehat{\sigma}(\widehat{\mathbf{w}} \cdot \mathbf{x}) - y)(\mathbf{w} \cdot \mathbf{x} - \sigma^{-1}(\widehat{\sigma}(\mathbf{w} \cdot \mathbf{x})))\right] \geq -\epsilon$$

(population-level iso-reg suffices but intractable)

Robust omniprediction

For input $\widehat{\mathbf{w}} \in \mathcal{W}$ return $\widehat{\sigma} \in \mathcal{S}$ s.t. $\text{for all } \sigma \in \mathcal{S}$

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\widehat{\sigma}(\widehat{\mathbf{w}}\cdot\mathbf{x}) - y)(\mathbf{w}\cdot\mathbf{x} - \sigma^{-1}(\widehat{\sigma}(\mathbf{w}\cdot\mathbf{x})))\right] \geq -\epsilon$$

(population-level iso-reg suffices but intractable)

$$\approx \mathbb{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}_n}\left[(\widehat{\sigma}(\widehat{\mathbf{w}}\cdot\mathbf{x}) - y)(\mathbf{w}\cdot\mathbf{x} - \sigma^{-1}(\widehat{\sigma}(\mathbf{w}\cdot\mathbf{x})))\right]$$

Key ideas for uniform convergence:

- Smoothing (slightly anti-Lipschitz w.l.o.g.)
- Dudley's generic chaining

Robust omniprediction in nearly-linear time

$$\min_{\{v_i\}_{i\in[n]}\subseteq\mathbb{R}} \sum_{i\in[n]} (v_i - y_i)^2$$

s.t. $\quad a_i \le v_{i+1} - v_i \le b_i \text{ for all } i \in [n-1]$

(empirical "bounded" iso-reg)

Robust omniprediction in nearly-linear time

$$\min_{\{v_i\}_{i\in[n]}\subseteq\mathbb{R}} \sum_{i\in[n]} (v_i - y_i)^2$$

s.t. $$a_i \le v_{i+1} - v_i \le b_i \text{ for all } i \in [n-1]$$

anti-Lipschitz / monotonicity constraints

Lipschitz constraints

(also in [KKKS '11, ZWDD '24])

Robust omniprediction in nearly-linear time

$$\min_{\{v_i\}_{i\in[n]}\subseteq\mathbb{R}} \sum_{i\in[n]} (v_i - y_i)^2$$

s.t. $\quad a_i \le v_{i+1} - v_i \le b_i \text{ for all } i \in [n-1]$

Prev. solver: inexact, $\Omega(n^2)$ time

[HTY '24]: exact, $O(n(\log(n))^2)$

Fast DP on piecewise quadratic

Robust omniprediction in nearly-linear time

$$\min_{\{v_i\}_{i\in[n]}\subseteq\mathbb{R}} \sum_{i\in[n]} (v_i - y_i)^2$$

s.t. $\quad a_i \le v_{i+1} - v_i \le b_i$ for all $i \in [n-1]$

Prev. solver: inexact, $\Omega(n^2)$ time

[HTY '24]: exact, $O(n(\log(n))^2)$

Fast DP on piecewise quadratic

Testing Calibration in Nearly-Linear Time

[Hu-Jambulapati-Tian-Yang '24]

See Chutong's poster!

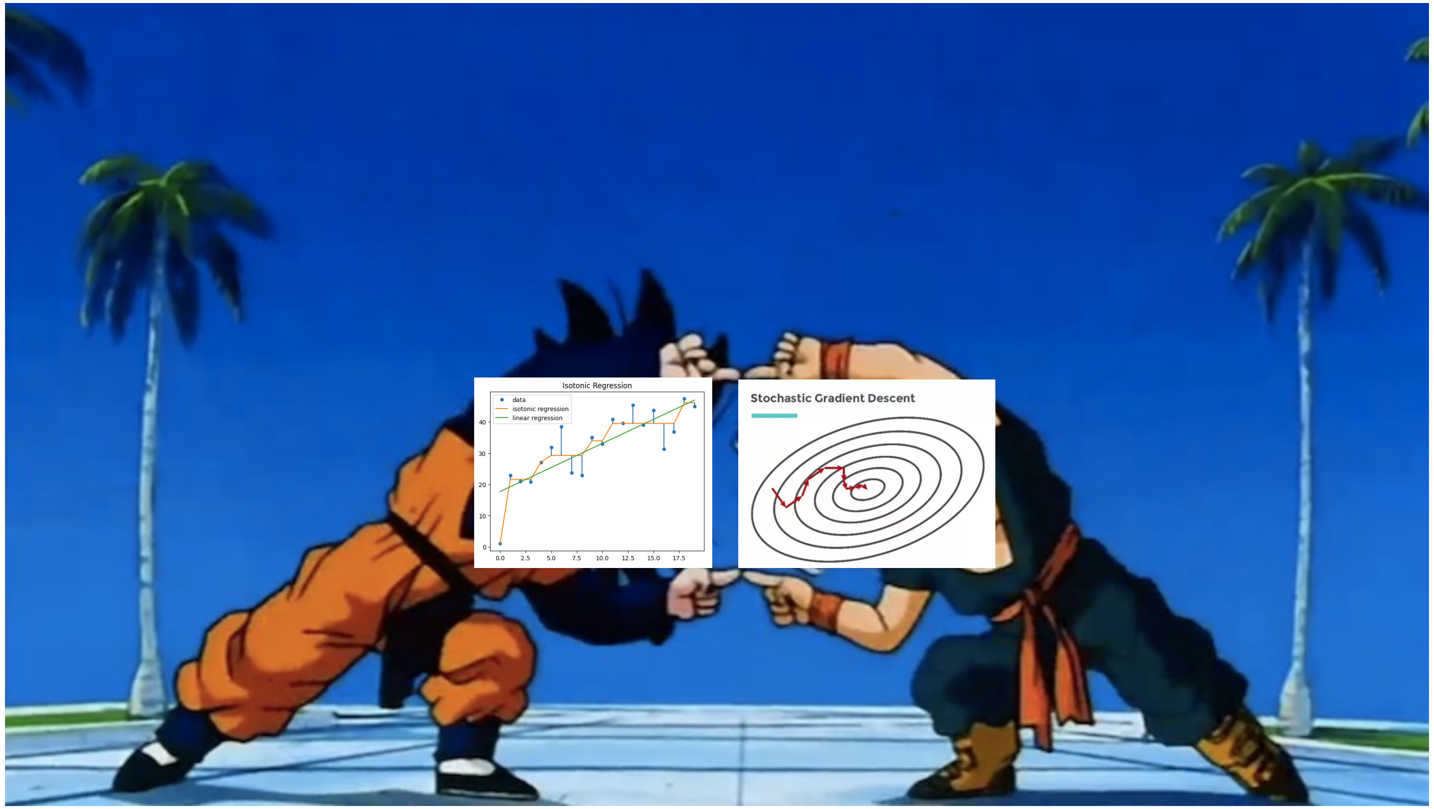
What else?

1. Omnipredicting structured families?
   - Regression setting?
   - Multi-class classification?
   - Multi-objective optimization? (Thanks Han ☺)
2. Proper omnipredictors?
   - Do they exist?
   - Some partial characterizations in our paper…
3. Practical implications?
   - Multigroup fairness, e.g., for fine-tuning

Thank you!

Contact

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



arXiv… soon™