Learning General Halfspaces with Adversarial Label Noise via Online Gradient Descent

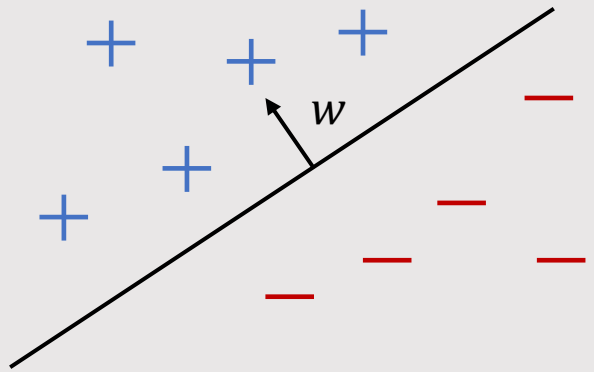
Nikos Zarifis

Joint work with: Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos

University of Wisconsin-Madison

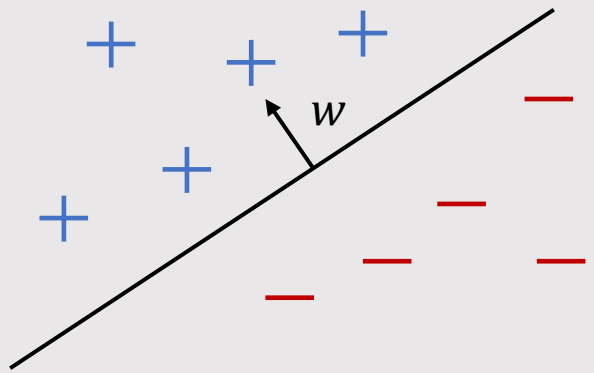
Learning Halfspaces

- General Halfspaces: Functions of the form $\text{sign}(w \cdot x + t)$ for some normal vector $w \in \mathbb{R}^d$ and $t \in \mathbb{R}$.



Learning Halfspaces

- General Halfspaces: Functions of the form $\text{sign}(w \cdot x + t)$ for some normal vector $w \in \mathbb{R}^d$ and $t \in \mathbb{R}$.
- Homogeneous Halfspaces: Halfspaces pass through the origin ($t = 0$).



Learning Halfspaces with Noise

- Noise-Free Setting: samples (x, y) with $y = \text{sign}(w^* \cdot x + t)$ from a distribution D supported on $\mathbb{R}^d \times \mathbb{R}$.

Learning Halfspaces with Noise

- Noise-Free Setting: samples (x, y) with $y = \text{sign}(w^* \cdot x + t)$ from a distribution D supported on $\mathbb{R}^d \times \mathbb{R}$.
- Goal: Find hypothesis h that minimizes the 0-1 loss:

$$\mathbb{P}_{(x,y)\sim D}[h(x) \neq y]$$

Learning Halfspaces with Noise

- Noise-Free Setting: samples (x, y) with $y = \text{sign}(w^* \cdot x + t)$ from a distribution D supported on $\mathbb{R}^d \times \mathbb{R}$.
- Goal: Find hypothesis h that minimizes the 0-1 loss:

$$\mathbb{P}_{(x,y)\sim D}[h(x) \neq y]$$

- Adversarial Noise: samples (x, y) from a distribution D but opt fraction of the labels is corrupted. We want to find a hypothesis h such that

$$\mathbb{P}_{(x,y)\sim D}[h(x) \neq y] \leq \text{opt} + \epsilon$$

Learning Halfspaces with Noise is HARD

- **Gaussian Covariates:**
  - Finding h that gets error opt + ϵ needs super-polynomial time.
    - Klivans Kothari (2014), Diakonikolas Kane Z (2020) , Goel Gollakota Klivans (2021), Diakonikolas Kane Pittas Z (2021)

Learning Halfspaces with Noise is HARD

- **Gaussian Covariates:**
  - Finding h that gets error opt + ϵ needs super-polynomial time.
    - Klivans Kothari (2014), Diakonikolas Kane Z (2020) , Goel Gollakota Klivans (2021), Diakonikolas Kane Pittas Z (2021)

- This Work:
  - We assume that the covariates follow Gaussian distribution.
  - We want to find a hypothesis h that gets error $O(opt) + \epsilon$.

Previous Work

- **Homogeneous Halfpsaces:**
  - Awasthi Balcan Long (2013): First algorithm for learning halfspaces with $O(\text{opt})$ error.
  - Diakonikolas Kontonis Tzamos Z (2020): Non-Convex SGD for learning halfspaces with $O(\text{opt})$ error.

Previous Work

- **Homogeneous Halfpsaces:**
  - Awasthi Balcan Long (2013): First algorithm for learning halfspaces with $O(\text{opt})$ error.
  - Diakonikolas Kontonis Tzamos Z (2020): Non-Convex SGD for learning halfspaces with $O(\text{opt})$ error.
- **General Halfspaces:**
  - Diakonikolas Kane Stewart (2018): First algorithm with O(opt) error.
    - Complicated and non-practical.
    - $(d/\epsilon)^{O(1)}$ sample complexity (sub-optimal).

Homogeneous vs General Halfspaces

- Homogeneous Halfspaces: $\text{sign}(w^* \cdot x)$
- Adapt Homogeneous to General:
  - Add extra coordinate: $x \to (x, 1)$
- Works: In the distribution-free setting.

Homogeneous vs General Halfspaces

- Homogeneous Halfspaces: $\text{sign}(w^* \cdot x)$
- Adapt Homogeneous to General:
  - Add extra coordinate: $x \to (x, 1)$
- Works: In the distribution-free setting.
- Does not work in distribution specific setting.
  - Provably the previous algorithms for homogeneous halfspaces does not work with this transformation.
  - The $\mathbf{x}$-marginal of the transformed instance is no longer Gaussian.

Is there a simple **iterative method** to learn general halfspaces with adversarial label noise?

Is there a simple **iterative method** to learn general halfspaces with adversarial label noise?

This work: Positive answer with an efficient algorithm.

Our Result – Online Gradient Descent to the Rescue

Theorem:
Performing online gradient descent on sequence of *non-convex* losses $\mathcal{L}_k(w,t)$, for k = poly(log(1/ϵ)) iterations finds w, t:

$$\mathbb{P}_{(x,y)\sim D}[\text{sign}(w \cdot x + t) \neq y] \leq O(\text{opt}) + \epsilon$$

- Near-optimal sample complexity: $\tilde{O}(\text{d}/\epsilon^2)$.

Technical Vignette

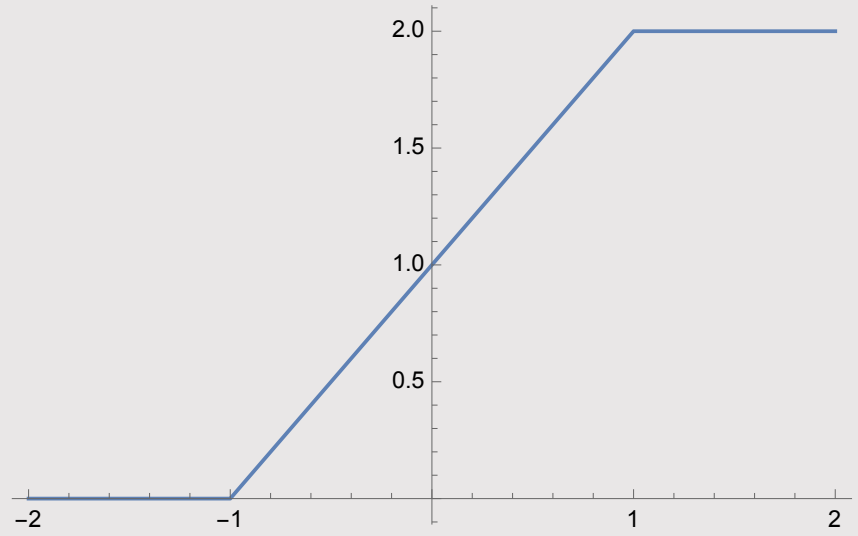
Gradient Descent for Homogeneous Halfspaces

- [DKTZ20] Optimizing a smooth approximation of the loss:

$$\mathcal{L}(w) = -\mathbb{E}_{(x,y)\sim D}\left[r\left(\frac{w\cdot x}{\|w\|_2}\frac{1}{\epsilon}\right)y\right]$$

suffices to get $O(\text{opt}) + \epsilon$ error.

$$r(t) = \begin{cases} 0 & \text{if} \quad t \in (-\infty, -1] \\ t+1 & \text{if} \quad t \in (-1, 1) \\ 2 & \text{otherwise} \end{cases}$$



Gradient Descent for Homogeneous Halfspaces

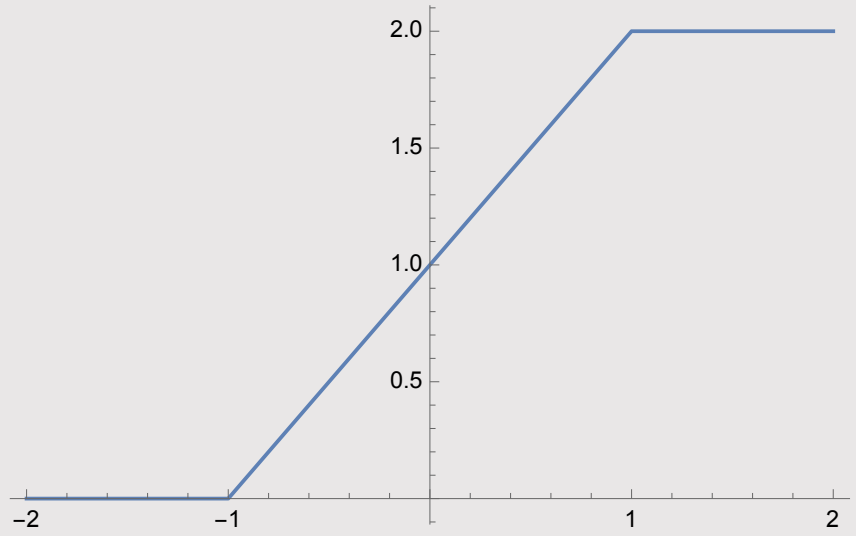
- [DKTZ20] Optimizing a smooth approximation of the loss:

$$\mathcal{L}(w) = -\mathbb{E}_{(x,y)\sim D}\left[r\left(\frac{w\cdot x}{\|w\|_2}\frac{1}{\epsilon}\right)y\right]$$

suffices to get $O(\mathrm{opt}) + \epsilon$ error.

Does not work for general halfspaces!

$$r(t) = \begin{cases} 0 & \text{if} \quad t \in (-\infty, -1] \\ t+1 & \text{if} \quad t \in (-1, 1) \\ 2 & \text{otherwise} \end{cases}$$

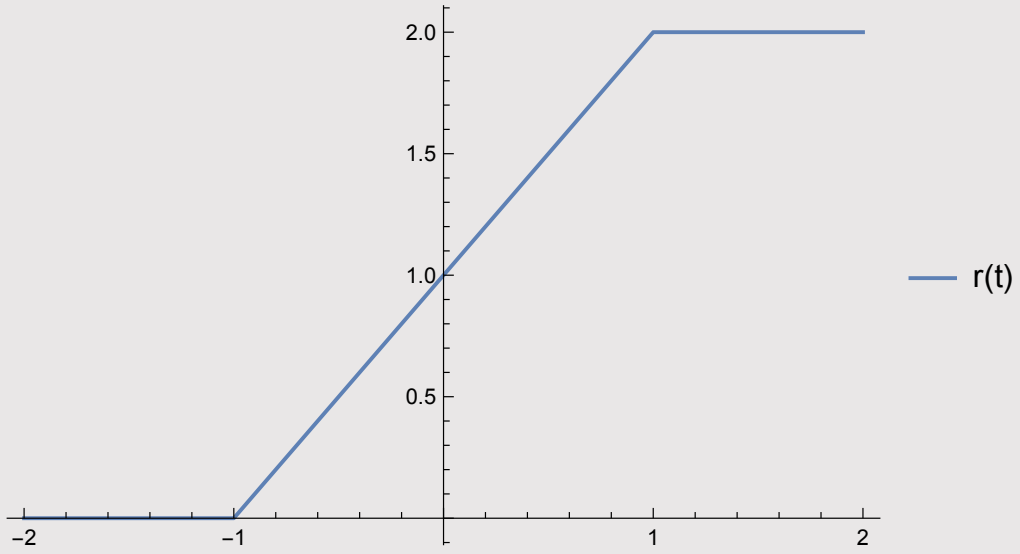


Online Gradient Descent for General Halfspaces

- Sequences of loss functions:

$$\mathcal{L}_k(w) = -\mathbb{E}_{(x,y)\sim D}\left[r\left(\frac{a_k(w\cdot x)}{\|w\|_2} - b_k\right)y\right]$$

$$r(t) = \begin{cases} 0 & \text{if} \quad t \in (-\infty, -1] \\ t+1 & \text{if} \quad t \in (-1, 1) \\ 2 & \text{otherwise} \end{cases}$$

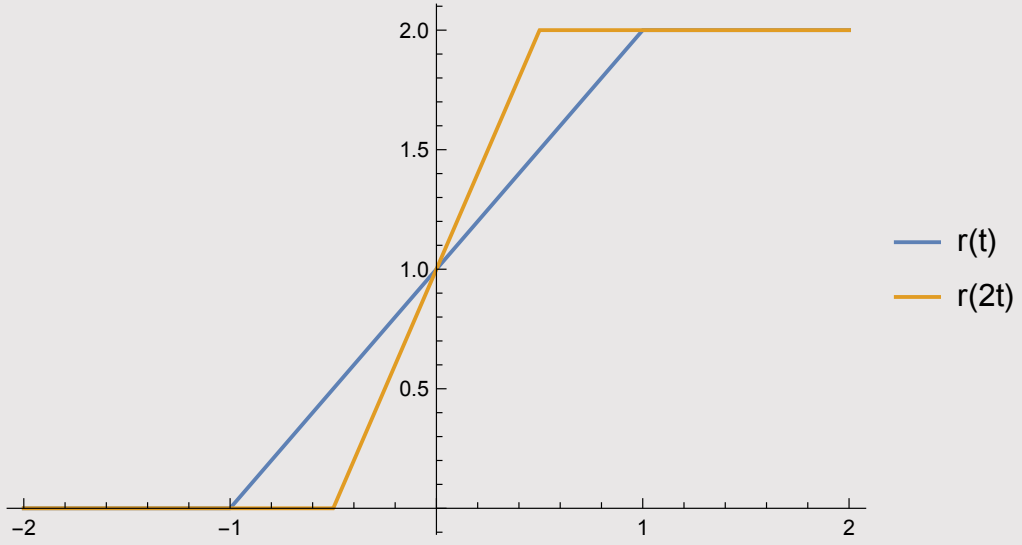


Online Gradient Descent for General Halfspaces

- Sequences of loss functions:

$$\mathcal{L}_k(w) = -\mathbb{E}_{(x,y)\sim D}\left[r\left(\frac{a_k(w\cdot x)}{\|w\|_2} - b_k\right)y\right]$$

$$r(t) = \begin{cases} 0 & \text{if} \quad t \in (-\infty, -1] \\ t+1 & \text{if} \quad t \in (-1, 1) \\ 2 & \text{otherwise} \end{cases}$$

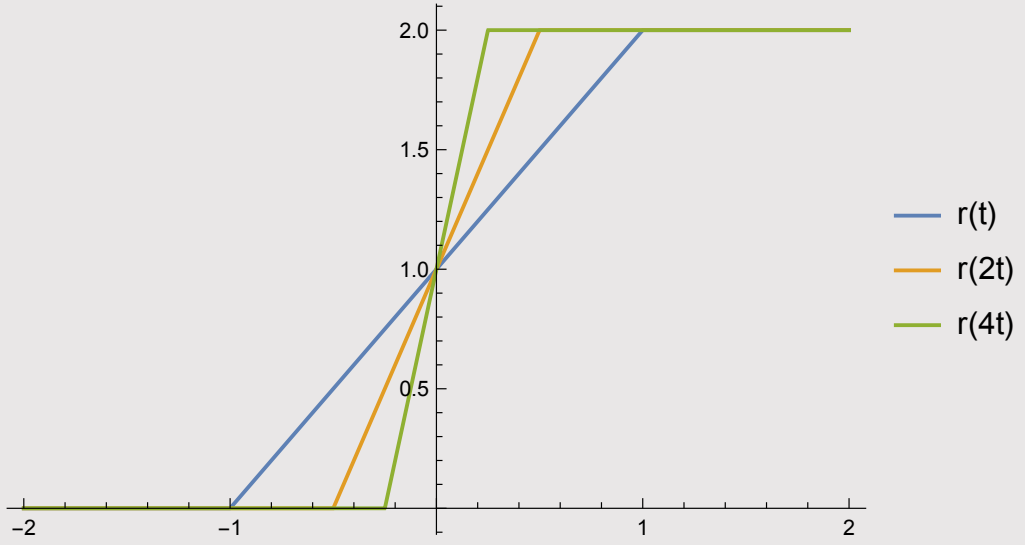


Online Gradient Descent for General Halfspaces

- Sequences of loss functions:

$$\mathcal{L}_k(w) = -\mathbb{E}_{(x,y)\sim D}\left[r\left(\frac{a_k(w\cdot x)}{\|w\|_2} - b_k\right)y\right]$$

$$r(t) = \begin{cases} 0 & \text{if} \quad t \in (-\infty, -1] \\ t+1 & \text{if} \quad t \in (-1, 1) \\ 2 & \text{otherwise} \end{cases}$$

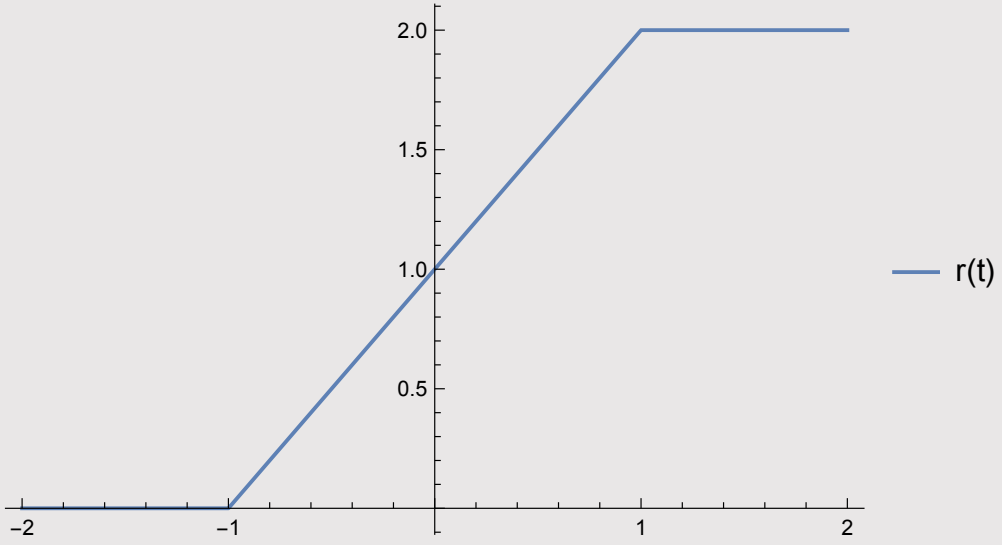


Online Gradient Descent for General Halfspaces

- Sequences of loss functions:

$$\mathcal{L}_k(w) = -\mathbb{E}_{(x,y)\sim D}\left[r\left(\frac{a_k(w\cdot x)}{\|w\|_2} - b_k\right)y\right]$$

$$r(t) = \begin{cases} 0 & \text{if} \quad t \in (-\infty, -1] \\ t+1 & \text{if} \quad t \in (-1, 1) \\ 2 & \text{otherwise} \end{cases}$$

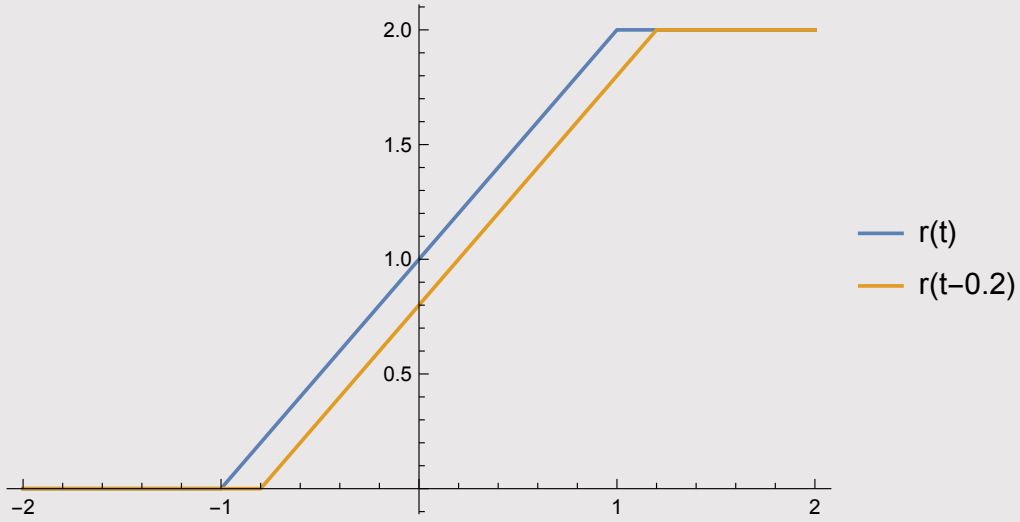


Online Gradient Descent for General Halfspaces

- Sequences of loss functions:

$$\mathcal{L}_k(w) = -\mathbb{E}_{(x,y)\sim D}\left[r\left(\frac{a_k(w\cdot x)}{\|w\|_2} - b_k\right)y\right]$$

$$r(t) = \begin{cases} 0 & \text{if} \quad t \in (-\infty, -1] \\ t+1 & \text{if} \quad t \in (-1, 1) \\ 2 & \text{otherwise} \end{cases}$$



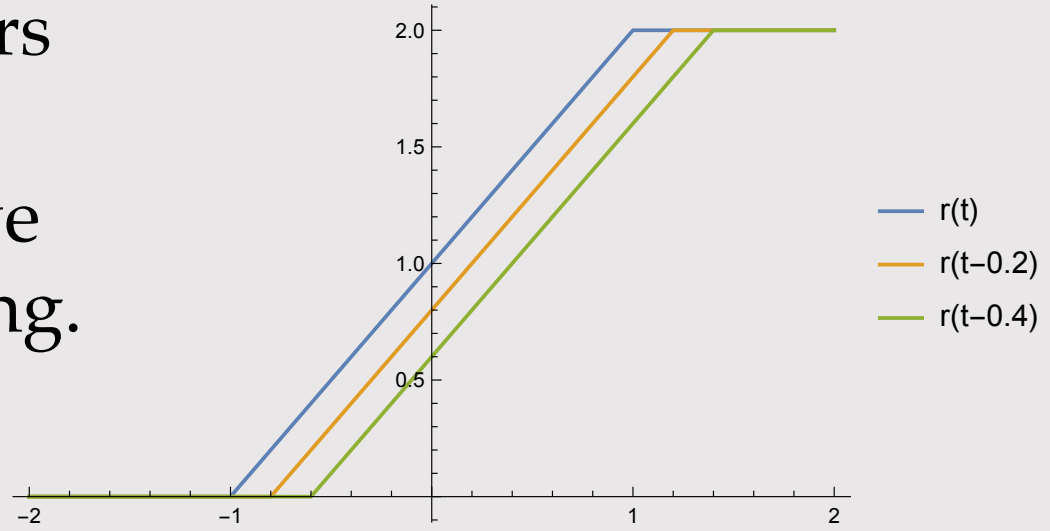
Online Gradient Descent for General Halfspaces

$$r(t) = \begin{cases} 0 & \text{if} \quad t \in (-\infty, -1] \\ t+1 & \text{if} \quad t \in (-1, 1) \\ 2 & \text{otherwise} \end{cases}$$

- Sequences of loss functions:

$$\mathcal{L}_k(w) = -\mathbb{E}_{(x,y)\sim D}\left[r\left(\frac{a_k(w \cdot x)}{\|w\|_2} - b_k\right) y\right]$$

- In each round, we tune the parameters a_k , b_k.
- Structural Result: In each iteration, we guarantee that: $\theta\left(w^k, w^*\right)$ is decreasing.



Conclusion

- We provide an algorithm for learning general halfspaces with adversarial noise achieving best-possible error guarantees and sample complexity.

Conclusion

- We provide an algorithm for learning general halfspaces with adversarial noise achieving best-possible error guarantees and sample complexity.
- Our algorithm is practical as it relies on a simple **online gradient descent iteration.**

Conclusion

- We provide an algorithm for learning general halfspaces with adversarial noise achieving best-possible error guarantees and sample complexity.
- Our algorithm is practical as it relies on a simple **online gradient descent iteration.**

Questions?