Agnostic Learnability of Halfspaces via Logistic Loss

Ziwei Ji, Kwangjun Ahn, Pranjal Awasthi, Satyen Kale, Stefani Karp

Problem setting

Comparison of prior results and our results

Details

Problem setting

Comparison of prior results and our results

Details

Problem setting

- Binary classification: unknown distribution P over $\mathbb{R}^d \times \{-1, +1\}$;
  we have i.i.d. samples from P.

Problem setting

- Binary classification: unknown distribution P over $\mathbb{R}^d \times \{-1, +1\}$;
  we have i.i.d. samples from P.
- Goal: compete with the optimal linear classifier $\bar{u}$ with zero-one/misclassification risk $\mathrm{OPT} > 0$ over P, i.e.,

$$\mathcal{R}_{0-1}(\bar{u}) := \Pr_{(x,y)\sim P}\left(\mathrm{sign}(\langle \bar{u}, x\rangle) \neq y\right) = \mathrm{OPT}.$$

Problem setting

Comparison of prior results and our results

Details

Logistic regression

A natural heuristic is logistic regression.
Notation: let $\ell_{\log}(z) := \ln(1 + e^{-z})$, and let

$$\mathcal{R}_{\log}(w) := \mathbb{E}_{(x,y)\sim P}\left[\ell_{\log}\left(y\langle w, x\rangle\right)\right]$$

denote the population logistic risk of w over P.

We can sample a training set and minimize the empirical risk,
or have a sequence of samples and run stochastic optimization.

Prior lower and upper bounds for logistic regression

Known upper and lower bounds don't match:

- With no assumption on P, logistic regression may attain zero-one risk as bad as $1 - \mathrm{OPT}$ (Ben-David et al., 2012).

Prior lower and upper bounds for logistic regression

Known upper and lower bounds don't match:

- With no assumption on P, logistic regression may attain zero-one risk as bad as $1 - \mathrm{OPT}$ (Ben-David et al., 2012).
- With isotropic log-concave distributions, $\widetilde{\Omega}(\mathrm{OPT})$ lower bound can be shown (Diakonikolas et al., 2020).

Prior lower and upper bounds for logistic regression

Known upper and lower bounds don't match:

- With no assumption on P, logistic regression may attain zero-one risk as bad as $1 - \mathrm{OPT}$ (Ben-David et al., 2012).
- With isotropic log-concave distributions, $\widetilde{\Omega}(\mathrm{OPT})$ lower bound can be shown (Diakonikolas et al., 2020).
- For "well-behaved" and sub-exponential distributions, SGD attains zero-one risk $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ (Frei et al., 2021).

Prior lower and upper bounds for logistic regression

Known upper and lower bounds don't match:

- With no assumption on P, logistic regression may attain zero-one risk as bad as $1 - \mathrm{OPT}$ (Ben-David et al., 2012).
- With isotropic log-concave distributions, $\widetilde{\Omega}(\mathrm{OPT})$ lower bound can be shown (Diakonikolas et al., 2020).
- For "well-behaved" and sub-exponential distributions, SGD attains zero-one risk $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ (Frei et al., 2021).

Here "well-behaved" conditions:

- standard concentration and anti-concentration conditions;
- a mixture of log-concave distributions (e.g., a Gaussian mixture) is a nice example.

Prior lower and upper bounds for logistic regression

Known upper and lower bounds don't match:

- With no assumption on P, logistic regression may attain zero-one risk as bad as $1 - \mathrm{OPT}$ (Ben-David et al., 2012).
- With isotropic log-concave distributions, $\widetilde{\Omega}(\mathrm{OPT})$ lower bound can be shown (Diakonikolas et al., 2020).
- For "well-behaved" and sub-exponential distributions, SGD attains zero-one risk $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ (Frei et al., 2021).

Here "well-behaved" conditions:

- standard concentration and anti-concentration conditions;
- a mixture of log-concave distributions (e.g., a Gaussian mixture) is a nice example.

Q. Can we close these gaps?

Prior lower and upper bounds for logistic regression

Known upper and lower bounds don't match:

- With no assumption on P, logistic regression may attain zero-one risk as bad as $1 - \mathrm{OPT}$ (Ben-David et al., 2012).
- With isotropic log-concave distributions, $\widetilde{\Omega}(\mathrm{OPT})$ lower bound can be shown (Diakonikolas et al., 2020).
- For "well-behaved" and sub-exponential distributions, SGD attains zero-one risk $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ (Frei et al., 2021).

Here "well-behaved" conditions:

- standard concentration and anti-concentration conditions;
- a mixture of log-concave distributions (e.g., a Gaussian mixture) is a nice example.

Q. Can we close these gaps? $\rightarrow$ precise scope of this work!

Our lower and upper bounds for logistic regression

- $\Omega\left(\sqrt{\mathrm{OPT}}\right)$ lower bound for "well-behaved" sub-exponential distributions;
  matching $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ upper bound from (Frei et al., 2021).

Our lower and upper bounds for logistic regression

- $\Omega\left(\sqrt{\mathrm{OPT}}\right)$ lower bound for "well-behaved" sub-exponential distributions;
  matching $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ upper bound from (Frei et al., 2021).
- $\widetilde{O}(\mathrm{OPT})$ upper bound with additional "radial Lipschitzness."

Upper bounds beyond logistic regression

- Diakonikolas et al. (2020) designed a nonconvex SGD method that achieves $O(\mathrm{OPT}) + \epsilon$ risk using $\widetilde{O}(d/\epsilon^4)$ samples. They can also handle heavy-tailed distributions.

Upper bounds beyond logistic regression

- Diakonikolas et al. (2020) designed a nonconvex SGD method that achieves $O(\mathrm{OPT}) + \epsilon$ risk using $\widetilde{O}(d/\epsilon^4)$ samples. They can also handle heavy-tailed distributions.
- Other prior algorithms achieving $O(\mathrm{OPT}) + \epsilon$ risk involve solving multiple rounds of convex program (Awasthi et al., 2014; Daniely, 2015).

Upper bounds beyond logistic regression

- Diakonikolas et al. (2020) designed a nonconvex SGD method that achieves $O(\mathrm{OPT}) + \epsilon$ risk using $\widetilde{O}(d/\epsilon^4)$ samples. They can also handle heavy-tailed distributions.
- Other prior algorithms achieving $O(\mathrm{OPT}) + \epsilon$ risk involve solving multiple rounds of convex program (Awasthi et al., 2014; Daniely, 2015).
- We design a simple two-phase convex program (logistic regression followed by Perceptron) that achieves $O(\mathrm{OPT} \ln(1/\mathrm{OPT})) + \epsilon$ risk using $\widetilde{O}(d/\epsilon^2)$ samples.

Problem setting

Comparison of prior results and our results

Details

Our $\Omega\left(\sqrt{\mathrm{OPT}}\right)$ lower bound

Theorem

There exists a distribution on $\mathbb{R}^2 \times \{-1, +1\}$, such that:

- *the feature distribution is isotropic and a mixture of log-concave distributions;*
- *the population logistic risk $\mathcal{R}_{\log}$ has a global minimizer w^* with*

$$\mathcal{R}_{0-1}(w^*) = \Omega\left(\sqrt{\mathrm{OPT}}\right).$$

Our $\Omega\left(\sqrt{\mathrm{OPT}}\right)$ lower bound

Theorem

There exists a distribution on $\mathbb{R}^2 \times \{-1, +1\}$, such that:

- *the feature distribution is isotropic and a mixture of log-concave distributions;*
- *the population logistic risk $\mathcal{R}_{\log}$ has a global minimizer w^* with*

$$\mathcal{R}_{0-1}(w^*) = \Omega\left(\sqrt{\mathrm{OPT}}\right).$$

- Matches $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ upper bound from (Frei et al., 2021).

Our $\widetilde{O}(\mathrm{OPT})$ upper bound under radial Lipschitzness

Assumption

There exists a measurable function $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ such that for any two-dimensional subspace V, letting p_V denote the density of the projection of feature distribution onto V, then

$$\left| p_V(r,\theta) - p_V(r,\theta') \right| \leq \kappa(r)|\theta - \theta'|.$$

Our $\widetilde{O}(\mathrm{OPT})$ upper bound under radial Lipschitzness

Assumption

There exists a measurable function $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ such that for any two-dimensional subspace V, letting p_V denote the density of the projection of feature distribution onto V, then

$$\left|p_V(r,\theta) - p_V(r,\theta')\right| \le \kappa(r)|\theta - \theta'|.$$

- Holds if p_V is Lipschitz continuous (e.g., Gaussian mixtures).
- Does not hold for general log-concave distributions.

Our $\widetilde{O}(\mathrm{OPT})$ upper bound under radial Lipschitzness

Theorem

If the distribution is well-behaved, sub-exponential and radially-Lipschitz, then with learning rate $\widetilde{\Theta}(1/d)$, using poly$(d, 1/\epsilon, \ln(1/\delta))$ *samples and iterations, with probability $1-\delta$, projected gradient descent outputs w_t with*

$$\mathcal{R}_{0-1}(w_t) = \widetilde{O}(\mathrm{OPT}) + \epsilon.$$

Why radial Lipschitzness?

Lemma

If the distribution is well-behaved, sub-exponential and radially-Lipschitz, and suppose $\hat{w}$ satisfies $\mathcal{R}_{\log}(\hat{w}) \leq \mathcal{R}_{\log}(\|\hat{w}\|\bar{u}) + \epsilon'$, then

$$\mathcal{R}_{0-1}(\hat{w}) = \widetilde{O}\left(\max\left\{\mathrm{OPT}, \sqrt{\tfrac{\epsilon'}{\|\hat{w}\|}}, \tfrac{C_\kappa}{\|\hat{w}\|^2}\right\}\right).$$

Why radial Lipschitzness?

Lemma

If the distribution is well-behaved, sub-exponential and radially-Lipschitz, and suppose $\hat{w}$ satisfies $\mathcal{R}_{\log}(\hat{w}) \leq \mathcal{R}_{\log}(\|\hat{w}\|\bar{u}) + \epsilon'$, then

$$\mathcal{R}_{0-1}(\hat{w}) = \widetilde{O}\left(\max\left\{\mathrm{OPT}, \sqrt{\frac{\epsilon'}{\|\hat{w}\|}}, \frac{C_\kappa}{\|\hat{w}\|^2}\right\}\right).$$

- $C_\kappa = O(\ln(1/\mathrm{OPT})^2)$ for Lipschitz continuous density.

Why radial Lipschitzness?

Lemma

If the distribution is well-behaved, sub-exponential and radially-Lipschitz, and suppose $\hat{w}$ satisfies $\mathcal{R}_{\log}(\hat{w}) \leq \mathcal{R}_{\log}(\|\hat{w}\|\bar{u}) + \epsilon'$, then

$$\mathcal{R}_{0-1}(\hat{w}) = \widetilde{O}\left(\max\left\{\mathrm{OPT}, \sqrt{\frac{\epsilon'}{\|\hat{w}\|}}, \frac{C_\kappa}{\|\hat{w}\|^2}\right\}\right).$$

- $C_\kappa = O(\ln(1/\mathrm{OPT})^2)$ for Lipschitz continuous density.
- We can find $\hat{w}$ with small ϵ' with PGD; $\|\hat{w}\| = \widetilde{\Omega}\left(1/\sqrt{\mathrm{OPT}}\right)$.

Our $\widetilde{O}(\mathrm{OPT})$ upper bound: two-phase algorithm

Key observation: the lemma holds for the hinge loss $\ell_h(z) := \max\{-z, 0\}$ **without** radial Lipschitzness!

Lemma

For **hinge loss**, *if the distribution is well-behaved and sub-exponential, and suppose $\hat{w}$ satisfies $\mathcal{R}_h(\hat{w}) \leq \mathcal{R}_h(\|\hat{w}\|\bar{u}) + \epsilon'$, then*

$$\mathcal{R}_{0-1}(\hat{w}) = \widetilde{O}\left(\max\left\{\mathrm{OPT}, \sqrt{\frac{\epsilon'}{\|\hat{w}\|}}\right\}\right).$$

Our $\widetilde{O}(\mathrm{OPT})$ upper bound: two-phase algorithm

Key observation: the lemma holds for the hinge loss $\ell_h(z) := \max\{-z, 0\}$ **without** radial Lipschitzness!

Lemma

For **hinge loss**, *if the distribution is well-behaved and sub-exponential, and suppose $\hat{w}$ satisfies $\mathcal{R}_h(\hat{w}) \leq \mathcal{R}_h(\|\hat{w}\|\bar{u}) + \epsilon'$, then*

$$\mathcal{R}_{0-1}(\hat{w}) = \widetilde{O}\left(\max\left\{\mathrm{OPT}, \sqrt{\frac{\epsilon'}{\|\hat{w}\|}}\right\}\right).$$

But, we are not quite done since the global minimizer of $\mathcal{R}_h$ is 0...

Our $\widetilde{O}(\mathrm{OPT})$ upper bound: two-phase algorithm

Lemma

If the distribution is well-behaved and sub-exponential, and suppose $\hat{w}$ satisfies $\mathcal{R}_h(\hat{w}) \leq \mathcal{R}_h(\|\hat{w}\|\bar{u}) + \epsilon'$, then

$$\mathcal{R}_{0-1}(\hat{w}) = \widetilde{O}\left(\max\left\{\mathrm{OPT}, \sqrt{\frac{\epsilon'}{\|\hat{w}\|}}\right\}\right).$$

Ideas:

- first find a unit v that is $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ away from $\bar{u}$;
- then minimize $\mathcal{R}_h$ over $\mathcal{D} := \left\{w \middle| \langle w, v\rangle \geq 1\right\}$.

Our $\widetilde{O}(\mathrm{OPT})$ upper bound: two-phase algorithm

Lemma

If the distribution is well-behaved and sub-exponential, and suppose $\hat{w}$ satisfies $\mathcal{R}_h(\hat{w}) \le \mathcal{R}_h(\|\hat{w}\|\bar{u}) + \epsilon'$, then

$$\mathcal{R}_{0-1}(\hat{w}) = \widetilde{O}\left(\max\left\{\mathrm{OPT}, \sqrt{\frac{\epsilon'}{\|\hat{w}\|}}\right\}\right).$$

Ideas:

- first find a unit v that is $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ away from $\bar{u}$;
- then minimize $\mathcal{R}_h$ over $\mathcal{D} := \left\{w \middle| \langle w, v\rangle \ge 1\right\}$.
  - $\forall w \in \mathcal{D}$, $\|w\| \ge 1$.

Our $\widetilde{O}(\mathrm{OPT})$ upper bound: two-phase algorithm

Lemma

If the distribution is well-behaved and sub-exponential, and suppose $\hat{w}$ satisfies $\mathcal{R}_h(\hat{w}) \leq \mathcal{R}_h(\|\hat{w}\|\bar{u}) + \epsilon'$, then

$$\mathcal{R}_{0-1}(\hat{w}) = \widetilde{O}\left(\max\left\{\mathrm{OPT}, \sqrt{\frac{\epsilon'}{\|\hat{w}\|}}\right\}\right).$$

Ideas:

- first find a unit v that is $\widetilde{O}\left(\sqrt{\mathrm{OPT}}\right)$ away from $\bar{u}$;
- then minimize $\mathcal{R}_h$ over $\mathcal{D} := \left\{w \middle| \langle w, v\rangle \geq 1\right\}$.
  - $\forall w \in \mathcal{D}$, $\|w\| \geq 1$.
  - $\|\hat{w}\|\bar{u}$ may not in $\mathcal{D}$, but $\left(1 + \widetilde{O}(\mathrm{OPT})\right)\|\hat{w}\|\bar{u} \in \mathcal{D}$!
    Since we choose v close to $\bar{u}$.

Our $\widetilde{O}(\mathrm{OPT})$ upper bound: two-phase algorithm

Another ingredient: when minimizing hinge loss, we use SGD (instead of GD) for sample efficiency;

Our $\widetilde{O}(\mathrm{OPT})$ upper bound: two-phase algorithm

Another ingredient: when minimizing hinge loss, we use SGD (instead of GD) for sample efficiency;
basically it's **Perceptron** with a restricted domain and warm start given by v.

Theorem

If the distribution is well-behaved and sub-exponential, using $\widetilde{O}(d/\epsilon^2)$ samples, SGD can achieve zero-one risk $O(\mathrm{OPT} \ln(1/\mathrm{OPT})) + \epsilon$.

Thanks, please come to our poster!