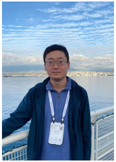
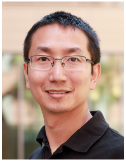
Agnostic Learning of Halfspaces with Gradient Descent via Soft Margins

Spencer Frei* Yuan Cao° Quanquan Gu°

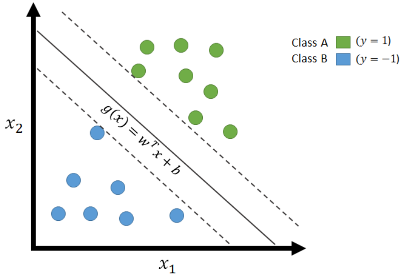
*UCLA Department of Statistics
°UCLA Department of Computer Science

Halfspaces

Halfspaces are classifiers $h : \mathbb{R}^d \to \{\pm 1\}$ where

$$h(x) = \operatorname{sgn}(\langle w, x \rangle - b)$$

for $w \in \mathbb{R}^d, b \in \mathbb{R}$.



$g(x) = w^T x + b = w_1 x_1 + w_2 x_2 + b$

Agnostic (PAC) Learning of Halfspaces

Given distribution $\mathcal{D}$ over $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$.
Consider class of bias-free halfspaces,

$$\mathcal{H} := \{x \mapsto \mathrm{sgn}(\langle w, x\rangle) : w \in \mathbb{R}^d\}.$$

For binary classification, loss of interest is zero-one loss:

$$\ell(y, \widehat{y}) = \mathbb{1}(y \neq \widehat{y}).$$

Denote error of best-fitting halfspace

$$\begin{aligned} \mathsf{OPT} &:= \min_{h \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \mathbb{1}(y \neq h(x)) \\ &= \min_{w \in \mathbb{R}^d} \mathbb{P}_{(x,y)\sim\mathcal{D}}\big(y \neq \mathrm{sgn}(\langle w, x\rangle)\big). \end{aligned}$$

Agnostic Learning of Halfspaces

$$\mathsf{OPT} := \min_{h \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \mathbb{1}(y \neq h(x)) = \min_{w \in \mathbb{R}^d} \mathbb{P}\big(y \neq \mathrm{sgn}(\langle w, x\rangle)\big).$$

- How many samples are necessary to learn a halfspace with error OPT $+ \varepsilon$?
- Are there computationally efficient algorithms for learning a halfspace with error OPT $+ \varepsilon$?
- Do we need assumptions on $\mathcal{D}$ for sample or computational efficiency?

Classical result: sample efficiency of ERM

Given i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$, empirical risk minimizer (ERM) over halfspaces is

$$h_{\mathrm{ERM}}(x) := \mathrm{argmin}_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left(y_i \neq \mathrm{sgn}(\langle w, x_i \rangle)\right).$$

Since VC dimension of halfspaces over $\mathbb{R}^d$ is d,

$\Theta(d/\varepsilon^2)$ samples necessary and sufficient

to achieve $|\mathrm{err}(h_{\mathrm{ERM}}) - \mathsf{OPT}| \leq \varepsilon$.

$\longrightarrow$ no assumptions on $\mathcal{D}$ necessary.

Computational difficulties in finding ERM

$$h_{\mathrm{ERM}}(x) := \mathrm{argmin}_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left(y_i \neq \mathrm{sgn}(\langle w, x_i \rangle)\right).$$

- $\Theta(d/\varepsilon^2)$ samples suffice for learning up to OPT $+ \varepsilon$ error with h_{ERM} (for any $\mathcal{D}$)
- But zero-one loss is <u>non-convex</u>: finding ERM under this loss is highly nontrivial!

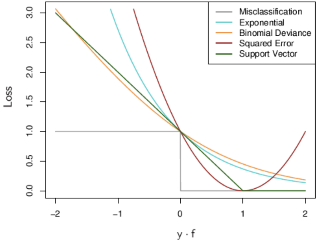
Computational difficulties in learning halfspaces

- If OPT $= 0$, linear programming methods are efficient.
- If OPT > 0, more complicated.
  - There exist $\mathcal{D}_x$ s.t. learning up to $O(\text{OPT}) + \varepsilon$ requires superpoly runtime. [Daniely, 2016]
  - If $\mathcal{D}_x = N(0, I)$, learning up to OPT $+ \varepsilon$ requires $d^{\text{poly}(1/\varepsilon)}$ runtime for SQ algorithms [Diakonikolas et al. 2020, Goel et al. 2020]
  - Efficient algorithms known to learn up to $O(\text{OPT}) + \varepsilon$ under assumptions on $\mathcal{D}_x$ [Awasthi et al. 2014, Diakonikolas et al. 2020]

Black-box optimization for classification

Standard approach for learning linear classifiers: gradient descent on convex surrogates (efficient).

$$w_{t+1} = w_t - \eta \nabla \widehat{L}(w_t)$$
$$= w_t - \eta \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(y_i \cdot \langle w_t, x_i \rangle),$$



with

$$\ell(z) \in \{\log(1+\exp(-z)),\ \max(1-z, 0),\ \exp(-z),\ \cdots\}.$$

When OPT $= 0$ this works. But when OPT > 0, unknown!

Agnostic learning of halfspaces with G.D.

Theorem

Suppose ℓ is convex, Lipschitz, decreasing. Assume $\mathcal{D}_x$ is sub-exponential and satisfies 'anti-concentration': $\exists U > 0$, such that p.d.f. $p_{\langle w,\cdot\rangle}(z) \leq U$ along 1D projections $\langle w, x\rangle$. Then G.D. on ℓ learns halfspaces with classification error at most $C \cdot \sqrt{\mathsf{OPT}}$ in poly time/sample complexity.

- Covers log-concave isotropic $\mathcal{D}_x$ (Gaussian, uniform, ...)
- Although learning up to OPT is hard, black-box optimization learns up to $C\sqrt{\mathsf{OPT}}$ efficiently.
- First positive result showing standard G.D. learns halfspaces with noise.

Proof idea: compare minimizers of surrogates for 0-1 vs. for 0-1 itself

If $L^\ell(w) = \mathbb{E}\ell(y\langle w, x\rangle)$, $L^{01}(w) = \min \mathbb{E}\,\mathbb{1}(y\langle w, x\rangle < 0)$,

$$w_\ell^* := \min_{\|w\| \le R} L^\ell(w) \quad \text{(finding minima is easy)},$$

vs.

$$w_{01}^* := \min_{w} L^{01}(w) \quad \text{(finding minima is hard)}.$$

Proof idea: compare minimizers of surrogates for 0-1 vs. for 0-1 itself

For $v \in \mathbb{R}^d$, $\|v\| = 1$,

soft margin function at $v := \phi_v(\gamma) = \mathbb{P}_{x \sim \mathcal{D}_x}(|\langle v, x \rangle| \leq \gamma)$.

For convex, 1-Lipschitz, decreasing ℓ with $\ell(0) = 1$ (for $\|x\| \leq 1$), want to compare

$\mathbb{E}\ell(y\langle w, x \rangle)$ vs. $\mathbb{E}\,\mathbb{1}(y\langle w, x \rangle < 0)$

Consider *normalized margin* $y\langle w/\|w\|, x\rangle$. Three cases:

1. Correct, large margin: $y\langle w/\|w\|, x\rangle \geq \gamma > 0$.
2. Correct, small ['soft'!] margin: $y\langle w/\|w\|, x\rangle \in [0, \gamma)$
3. Incorrect: $y\langle w/\|w\|, x\rangle < 0$.

Proof idea: compare minimizers of surrogates for 0-1 vs. for 0-1 itself

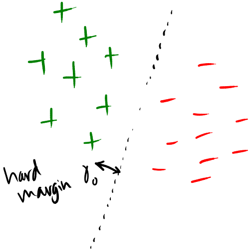
For $v \in \mathbb{R}^d$, $\|v\| = 1$,

soft margin function at $v := \phi_v(\gamma) = \mathbb{P}_{x \sim \mathcal{D}_x}(|\langle v, x\rangle| \leq \gamma)$.

For convex, 1-Lipschitz, decreasing ℓ with $\ell(0) = 1$ (for $\|x\| \leq 1$)

$$\begin{aligned}
\mathbb{E}\ell(y\langle w, x\rangle) = \mathbb{E}\ell(y\langle w, x\rangle)\Big[&\mathbb{1}(y\langle w/\|w\|, x\rangle \geq \gamma) \\
&+ \mathbb{1}(y\langle w/\|w\|, x\rangle \in [0,\gamma)) + + \mathbb{1}(y\langle w/\|w\|, x\rangle < 0)\Big] \\
&\leq \ell(\gamma\,\|w\|) + \phi_{w/\|w\|}(\gamma) \\
&\quad + (1 + \|w\|)\mathbb{P}(y \neq \operatorname{sgn}(\langle w, x\rangle))
\end{aligned}$$

Soft margins connect minimizers of 0-1 and surrogates



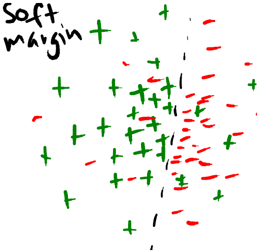
Recall: $\phi_{v/\|v\|}(\gamma) = \mathbb{P}_{x\sim\mathcal{D}_x}(|\langle v/\|v\|, x\rangle| \le \gamma)$.

$$\mathbb{E}\ell(y\langle w, x\rangle) \le (1 + \|w\|)\mathrm{err}(w) + \phi_{w/\|w\|}(\gamma) + \ell(\|w\|\gamma).$$

Assume w^*, $\|w^*\| = 1$ is s.t. $\mathrm{err}(w^*) = \mathsf{OPT}$.

- If 'hard margin' of γ_0, $\phi_{w^*}(\gamma) = 0$ for $\gamma \le \gamma_0$, so for $\rho > 0$, $\mathbb{E}\ell(y\rho\gamma_0^{-1}\langle w^*, x\rangle) \le (1 + \rho\gamma_0^{-1})\mathsf{OPT} + \ell(\rho) = O(\gamma_0^{-1}\mathsf{OPT})$.
- Matches lower bound of Ben-David et al., 2012

Soft margins connect minimizers of 0-1 and surrogates



Recall: $\phi_{v/\|v\|}(\gamma) = \mathbb{P}_{x\sim\mathcal{D}_x}(|\langle v/\|v\|, x\rangle| \le \gamma)$.

$$\mathbb{E}\ell(y\langle w, x\rangle) \le (1 + \|w\|)\mathrm{err}(w) + \phi_{w/\|w\|}(\gamma) + \ell(\|w\|\gamma).$$

Assume w^*, $\|w^*\| = 1$ is s.t. $\mathrm{err}(w^*) = \mathsf{OPT}$.

- If anti-concentration, $\phi_{w^*}(\gamma) = O(\gamma)$, so for $\rho > 0$,
  $\mathbb{E}\ell(y\rho\gamma^{-1}\langle w^*, x\rangle) \le (1 + \rho\gamma^{-1})\mathsf{OPT} + C\cdot\gamma + \ell(\rho)$.
- Take $\gamma = \mathsf{OPT}^{1/2}$ gives $O(\mathsf{OPT}^{1/2})$.

Summary

- Understanding G.D. for minimizing classification error requires understanding minimizers of surrogate vs 0-1
- Soft margin (& benign distrib. assumptions) connect the minimizers of surrogate to 0-1.
- G.D. is efficient, somewhat noise-robust, but not optimally so
- Soft margin idea useful in other contexts
  - Adversarial robustness + adversarial training (Zou*, F.*, Gu, ICML 2021)
  - Learning with neural networks trained by G.D. (F., Cao, Gu, ICML 2021)