DINFK

Tight bounds for maximum ℓ_1-margin classifiers

February 2024, Algorithmic Learning Theory

Stefan Stojanovic*, **Konstantin Donhauser***, Fanny Yang

Statistical Machine Learning group, CS department, ETH Zurich

ETH zürich

Linear classification setting

Data model for n samples (standard model in 1-bit compressive sensing):

- **Gaussian covariates** $x_i \sim N(0, I_d)$ of dimension d and **binary labels** $y_i = \mathrm{sgn}(\langle w^*, x_i \rangle)\xi_i$
- Noise model $\xi_i | x_i \sim \mathbb{P}_\sigma\,(.\,; \langle x_i, w^* \rangle)$ only depends on x_i in the **direction** of w^*
  - Special cases: random label flips, logistic regression model, random noise before quantization
- **Sparse high-dimensional** regime where $d \gg n$ and $\|w^*\|_0 = s \ll n$

Performance measure:

$$\mathbf{R}(\widehat{w}) = \mathbb{E}_{x \sim N(0,I)}\, 1[\mathrm{sgn}(\langle w^*, x \rangle) \neq \mathrm{sgn}(\langle \widehat{w}, x \rangle)] \approx \frac{1}{\pi} \left\| \frac{\widehat{w}}{\left|\left|\widehat{w}\right|\right|_2} - w^* \right\|$$

Maximum ℓ_1-margin classifier

We study the **maximum ℓ_1-margin classifier**, given by

Interpolation constraint

$$\widehat{w} = \underset{w}{\operatorname{argmin}} \ \|w\|_1 \qquad \text{s.t.} \ \ \forall i: \ \boxed{y_i \langle w, x_i \rangle \geq 1}$$

$$\widehat{w} \propto \underset{w}{\operatorname{argmax}} \min_i y_i \langle w, x_i \rangle \qquad \text{s.t.} \quad \|w\|_1 \leq 1 \, .$$

- Classifier may not be unique. Our results **hold simultaneously for all solutions!**

Maximum ℓ_1-margin classifier

We study the **maximum ℓ_1-margin classifier**, given by

Interpolation constraint

$$\widehat{w} = \underset{w}{\operatorname{argmin}} \ \|w\|_1 \qquad \text{s.t.} \ \ \forall i: \ \boxed{y_i \langle w, x_i \rangle \geq 1}$$

$$\widehat{w} \propto \underset{w}{\operatorname{argmax}} \min_i y_i \langle w, x_i \rangle \qquad \text{s.t.} \quad \|w\|_1 \leq 1 \, .$$

- Classifier may not be unique. Our results **hold simultaneously for all solutions!**

Classifier arises naturally as limiting solution of

- ℓ_1-penalized logistic regression as $\lambda \to 0$
- coordinate descent algorithms such as Adaboost (implicit bias) [T13]

Problem 1: Behavior in the high noise regime

- Practically speaking, one should **use regularization**!
- Goal, understand why and to what extend interpolating classifiers can ‘‘absorb the noise”

Problem 1: Behavior in the high noise regime

- Practically speaking, one should **use regularization**!
- Goal, understand why and to what extend interpolating classifiers can ‘‘absorb the noise”

Related work

- [CKLvdG22],[W10]: existing upper bounds of **order $\boldsymbol{O(1)}$** for **general (adversarial) corruptions.**

Problem 1: Behavior in the high noise regime

- Practically speaking, one should **use regularization**!
- Goal, understand why and to what extend interpolating classifiers can ‘‘absorb the noise”

Related work

- [CKLvdG22],[W10]: existing upper bounds of **order $\boldsymbol{O(1)}$** for **general (adversarial) corruptions.**

- [CLvdG21],[WDY21]: In the related sparse linear regression setting, rates for **adversarial noise** are of order $\boldsymbol{O(\sigma^2)}$ while rates for **randomized noise** are of order $\boldsymbol{O\left(\frac{\sigma^2}{\log\left(\frac{d}{n}\right)}\right)}$.

Problem 1: Behavior in the high noise regime

- Practically speaking, one should **use regularization**!
- Goal, understand why and to what extend interpolating classifiers can ‘‘absorb the noise”

Related work

- [CKLvdG22],[W10]: existing upper bounds of **order $O(1)$** for **general (adversarial) corruptions.**

- [WDY21]: In the related sparse linear regression setting, rates for **adversarial noise** are of order $O(\sigma^2)$ while rates for **randomized noise** are of order $O\left(\frac{\sigma^2}{\log\left(\frac{d}{n}\right)}\right)$.

Open Question: Does the prediction error for the maximum ℓ_1-margin classifier **yield vanishing rates** when labels are **randomly** corrupted?

First main result: Yes!

Theorem 2 – Noisy classification

Suppose $\|w^*\|_0 \lesssim \frac{n}{\log\left(\frac{d}{n}\right)^5}$. For any $n \geq \kappa_1$, and $\kappa_1 n \leq d \leq \exp(\kappa_3 n^{1/5})$,

$$\mathbf{R}(\widehat{w}) = \frac{\kappa_\sigma}{\sqrt{\log(d/n)}} + O\left(\frac{1}{\log^{3/4}(d/n)}\right)$$

w.p. $\geq 1 - c_1 \exp\left(-c_2 \frac{n}{\log\left(\frac{d}{n}\right)^5}\right) - c_3 \exp\left(-c_4 \frac{n}{\log(n)\log\left(\frac{d}{n}\right)^{\frac{3}{2}}}\right)$ and κ_σ only depends $\mathbb{P}_\sigma$

First main result: Yes!

Theorem 2 – Noisy classification

Suppose $\|w^*\|_0 \lesssim \frac{n}{\log\left(\frac{d}{n}\right)^5}$. For any $n \geq \kappa_1$, and $\kappa_1 n \leq d \leq \exp(\kappa_3 n^{1/5})$,

$$\mathbf{R}(\widehat{w}) = \frac{\kappa_\sigma}{\sqrt{\log(d/n)}} + O\left(\frac{1}{\log^{3/4}(d/n)}\right)$$

w.p. $\geq 1 - c_1 \exp\left(-c_2 \frac{n}{\log\left(\frac{d}{n}\right)^5}\right) - c_3 \exp\left(-c_4 \frac{n}{\log(n)\log\left(\frac{d}{n}\right)^{\frac{3}{2}}}\right)$ and κ_σ only depends $\mathbb{P}_\sigma$

Exact non-asymptotic error bound (including constants). **Vanishing error** as $n, d \to \infty$ and $d \gg n$

First main result: Yes!

Theorem 2 – Noisy classification

Suppose $\|w^*\|_0 \lesssim \frac{n}{\log\left(\frac{d}{n}\right)^5}$. For any $n \geq \kappa_1$, and $\kappa_1 n \leq d \leq \exp(\kappa_3 n^{1/5})$,

$$\mathbf{R}(\widehat{w}) = \frac{\kappa_\sigma}{\sqrt{\log(d/n)}} + O\left(\frac{1}{\log^{3/4}(d/n)}\right)$$

w.p. $\geq 1 - c_1 \exp\left(-c_2 \frac{n}{\log\left(\frac{d}{n}\right)^5}\right) - c_3 \exp\left(-c_4 \frac{n}{\log(n)\log\left(\frac{d}{n}\right)^{\frac{3}{2}}}\right)$ and κ_σ only depends $\mathbb{P}_\sigma$

Sparsity can go up to almost n. Open question whether we can replace the constraint with $\|w^*\|_1 \lesssim \frac{n}{\log\left(\frac{d}{n}\right)^5}$?

Problem 2: Behavior in the noiseless regime ($\xi = 1$)

- Maximum-margin classifiers are natural choices in the noiseless (low-noise) regimes

Problem 2: Behavior in the noiseless regime ($\xi = 1$)

- Maximum-margin classifiers are natural choices in the noiseless (low-noise) regimes

Related work

- [CKLvdG22],[W10] Show an upper bound of order $\tilde{O}\left(\frac{\|w^*\|_1^2}{n}\right)^{1/3}$ **for general (non-sparse) vectors**
- Adaptivity to sparsity: many sparse classifiers/regressors show a gap in rates between hard-sparse (bounded ℓ_0-norm) and soft-sparse (bounded ℓ_1-norm) ground truth w^*

Problem 2: Behavior in the noiseless regime ($\xi = 1$)

- Maximum-margin classifiers are natural choices in the noiseless (low-noise) regimes

Related work

- [CKLvdG22],[W10] Show an upper bound of order $\tilde{O}\left(\frac{\|w^*\|_1^2}{n}\right)^{1/3}$ **for general (non-sparse) vectors**
- Adaptivity to sparsity: many sparse classifiers/regressors show a gap in rates between hard-sparse (bounded ℓ_0-norm) and soft-sparse (bounded ℓ_1-norm) ground truth w^*

Open question: Can we improve the rates in [CKLvdG22] when w^* is (hard) sparse?

Second main result: No!

Theorem 3 – Noiseless classification

Suppose $\|w^*\|_0 \lesssim n^{\frac{2}{3}}\log(d)^{-14/3}$. For any $n \geq \kappa_1$, and $\kappa_1 m_n \leq d \leq \exp(\kappa_3 n^{1/12})$,

$$\mathbf{R}(\widehat{w}) = \left(\frac{8\|w^*\|_1^2}{\sqrt{3\pi^5}\, n \log\left(\frac{\mathrm{d}}{\mathrm{m_n}}\right)^{\frac{1}{2}}}\right)^{1/3} + O\left(\frac{\|w^*\|_1^2}{n \log\left(\frac{\mathrm{d}}{\mathrm{m_n}}\right)}\right)^{1/3}$$

w.p. $\geq 1 - c_1 d^{-1} - c_2 \exp\left(-c_3 \frac{n^{1/3}}{\log\left(\frac{d}{m_n}\right)^4}\right)$ and κ_0 some constant and $m_n = \widetilde{\Theta}(n\|w^*\|_1)^{2/3}$

Second main result: No!

Theorem 3 – Noiseless classification

Suppose $\|w^*\|_0 \lesssim n^{\frac{2}{3}} \log(d)^{-14/3}$. For any $n \geq \kappa_1$, and $\kappa_1 m_n \leq d \leq \exp(\kappa_3 n^{1/12})$,

$$\mathbf{R}(\widehat{w}) = \left(\frac{8\|w^*\|_1^2}{\sqrt{3\pi^5}\, n \log\left(\frac{\mathrm{d}}{\mathrm{m_n}}\right)^{\frac{1}{2}}} \right)^{1/3} + O\left(\frac{\|w^*\|_1^2}{n \log\left(\frac{\mathrm{d}}{\mathrm{m_n}}\right)} \right)^{1/3}$$

w.p. $\geq 1 - c_1 d^{-1} - c_2 \exp\left(-c_3 \frac{n^{1/3}}{\log\left(\frac{d}{m_n}\right)^4}\right)$ and κ_0 some constant and $m_n = \widetilde{\Theta}(n\|w^*\|_1)^{2/3}$

Exact non-asymptotic error bound (including constants) **of order** $\mathbf{R}(\widehat{w}) = \widetilde{O}\left(\frac{\|w^*\|_1^2}{n}\right)^{1/3}$ **!**

Second main result: No!

Theorem 3 – Noiseless classification

Suppose $\|w^*\|_0 \lesssim n^{\frac{2}{3}}\log(d)^{-14/3}$. For any $n \geq \kappa_1$, and $\kappa_1 m_n \leq d \leq \exp(\kappa_3 n^{1/12})$,

$$\mathbf{R}(\widehat{w}) = \left(\frac{8\|w^*\|_1^2}{\sqrt{3\pi^5}\, n \log\left(\frac{\mathrm{d}}{\mathrm{m_n}}\right)^{\frac{1}{2}}}\right)^{1/3} + O\left(\frac{\|w^*\|_1^2}{n \log\left(\frac{\mathrm{d}}{\mathrm{m_n}}\right)}\right)^{1/3}$$

w.p. $\geq 1 - c_1 d^{-1} - c_2 \exp\left(-c_3 \frac{n^{1/3}}{\log\left(\frac{d}{m_n}\right)^4}\right)$ and κ_0 some constant and $m_n = \widetilde{\Theta}(n\|w^*\|_1)^{2/3}$

Maximum ℓ_1-margin classifier is **not adaptive to sparsity**

Second main result: No!

Theorem 3 – Noiseless classification

Suppose $\|w^*\|_0 \lesssim n^{\frac{2}{3}}\log(d)^{-14/3}$. For any $n \geq \kappa_1$, and $\kappa_1 m_n \leq d \leq \exp(\kappa_3 n^{1/12})$,

$$\mathbf{R}(\widehat{w}) = \left(\frac{8\|w^*\|_1^2}{\sqrt{3\pi^5}\, n \log\left(\frac{\mathrm{d}}{\mathrm{m_n}}\right)^{\frac{1}{2}}}\right)^{1/3} + O\left(\frac{\|w^*\|_1^2}{n \log\left(\frac{\mathrm{d}}{\mathrm{m_n}}\right)}\right)^{1/3}$$

w.p. $\geq 1 - c_1 d^{-1} - c_2 \exp\left(-c_3 \frac{n^{1/3}}{\log\left(\frac{d}{m_n}\right)^4}\right)$ and κ_0 some constant and $m_n = \widetilde{\Theta}(n\|w^*\|_1)^{2/3}$

Open Problem: Show that **early stopped coordinate descent** yields **faster (adaptive) rates** and thus early stopping may be helpful **even if there is no noise!**

Proof sketch

Uniform convergence based proof (similar to [KZSS21] etc)

- Step 1 (Localization): Upper bound with high prob.

$$\phi_N = [\min_w \|w\|_1 \quad \text{s.t.} \quad \forall i: \quad y_i \langle w, x_i \rangle \geq 1] \quad \leq M$$

- Step 2 (Uniform Convergence): Upper (resp. lower) bound with high prob.

$$\phi_- / \phi_+ = \min_w / \max_w \left\| \frac{w}{\|w\|_2} - w^* \right\|_2 \quad \text{s.t.} \quad \forall i: \quad y_i \langle w, x_i \rangle \geq 1 \text{ and } \|w\|_1 \leq M$$

Proof sketch

Uniform convergence based proof (similar to [KZSS21] etc)

- Step 1 (Localization): Upper bound with high prob.

$$\phi_N = [\min_w \|w\|_1 \quad \text{s.t.} \quad \forall i: \quad y_i\langle w, x_i\rangle \geq 1] \quad \leq M$$

- Step 2 (Uniform Convergence): Upper (resp. lower) bound with high prob.

$$\phi_- / \phi_+ = \min_w / \max_w \left\| \frac{w}{\|w\|_2} - w^* \right\|_2 \quad \text{s.t.} \quad \forall i: \quad y_i\langle w, x_i\rangle \geq 1 \text{ and } \|w\|_1 \leq M$$

Key difficulty:

- To obtain tight bounds we require **a very sharp analysis (including the constants)**
- For instance, replacing M with $2M$ would only yields loose (already known) bounds

⇨ **Gaussianity is key!**

Main Tool: Convex Gaussian Minimax Theorem

Compare the Gaussian processes

$$\phi_N = \min_w \|w\|_1 \quad \text{s.t.} \quad \forall i: \qquad y_i \langle w, x_i \rangle \geq 1$$

Main Tool: Convex Gaussian Minimax Theorem

Compare the Gaussian processes

$$\phi_N = \min_w \|w\|_1 \quad \text{s.t.} \quad \forall i: \qquad y_i \langle w, x_i \rangle \geq 1$$

$$\Phi_N = \min_w \|w\|_1 \quad \text{s.t.} \quad \langle w_\perp, H \rangle \geq f_n(\|w_\perp\|_2, w_\parallel)$$

- f_n is some (random) function and H an iid Gaussian random vector

Main Tool: Convex Gaussian Minimax Theorem

Compare the Gaussian processes

$$\phi_N = \min_w \|w\|_1 \quad \text{s.t.} \quad \forall i: \quad y_i\langle w, x_i\rangle \geq 1$$

$$\Phi_N = \min_w \|w\|_1 \quad \text{s.t.} \quad \langle w_\perp, H\rangle \geq f_n(\|w_\perp\|_2, w_\parallel)$$

- f_n is some (random) function and H an iid Gaussian random vector
- (C)GMT by [TOH14] : a high prob. upper bound for Φ_N yields a high prob. upper bound for ϕ_N

Main technical contribution is a careful analysis of Φ_N, Φ_+ and Φ_-

Thanks for listening!

- [CKLvdG22]: Chinot, G., Kuchelmeister, F., Löffler, M., & van de Geer, S. (2022). AdaBoost and robust one-bit compressed sensing. *Mathematical Statistics and Learning*, *5*(1), 117-158.
- [KZSS21]: Koehler, F., Zhou, L., Sutherland, D. J., & Srebro, N. (2021). Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. *Advances in Neural Information Processing Systems*, *34*, 20657-20668.
- [T13]: Telgarsky, M. (2013, May). Margins, shrinkage, and boosting. In *International Conference on Machine Learning* (pp. 307-315). PMLR.
- [TOH14]: Thrampoulidis, C., Oymak, S., & Hassibi, B. (2014). The Gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*.
- [WDY22]: Wang, G., Donhauser, K., & Yang, F. (2022, May). Tight bounds for minimum $\ell_1 $-norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics* (pp. 10572-10602). PMLR.
- [W10]: Wojtaszczyk, P. (2010). Stability and instance optimality for Gaussian measurements in compressed sensing. *Foundations of Computational Mathematics*, *10*, 1-13.