An optimization approach to post-processing for classification with system constraints

Evgeny CHZHEN

SCOR
FOUNDATION FOR SCIENCE

Post-processing for multi-class classification

with system constraints

joint works with G. Taturyan and M. Hebiri

Workshop SCOR Foundation
Paris, France

Evgenii Chzhen
CNRS, Univ. Paris-Saclay

What this talk is about?

A post-processing algorithm for classification under various constraints

Plan

1. A quick intro into randomized classifiers
2. A glimpse into different constraints that one might want to incorporate
3. General framework and proposed methodology
4. Some numerical experiments

Multi-class classification: randomized classifiers

Data: $(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times [K]$ $\qquad p_k(\boldsymbol{X}) := \mathbb{P}(Y = k \mid \boldsymbol{X})$

Prediction kernel: $\pi : \mathcal{X} \to \Delta([K])$
Prediction: $\widehat{Y}_\pi \mid \boldsymbol{X} \sim \pi(\cdot \mid \boldsymbol{X})$ and $(\widehat{Y}_\pi \perp\!\!\!\perp Y) \mid \boldsymbol{X}$

Classical setup: perform as well as Bayes classifier

$$\pi^\star \in \arg\min_\pi \left\{ \mathcal{R}(\pi) = \mathbb{P}(Y \neq \widehat{Y}_\pi) \right\}$$

an LP in π

Folklore result: $\widehat{Y}_{\pi^*} = \arg\max_{k \in [K]} p_k(\boldsymbol{X})$.

Sometimes we might want more: rejection

Observations: $(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times [K]$ $\qquad p_k(\boldsymbol{X}) = \mathbb{P}(Y = k \mid \boldsymbol{X})$

Prediction kernel: $\pi : \mathcal{X} \to \Delta([K] \cup \{r\})$

Prediction: $\widehat{A}_\pi \mid \boldsymbol{X} \sim \pi(\cdot \mid \boldsymbol{X})$ and $(\widehat{A}_\pi \perp\!\!\!\perp Y) \mid \boldsymbol{X}$

Possible goal: perform as well as

$$\pi^\star \in \arg\min_\pi \left\{ \mathcal{R}(\pi) = \mathbb{P}(Y \neq \widehat{A}_\pi, \widehat{A}_\pi \neq r) \ : \ C(\pi) = \mathbb{P}(\widehat{A}_\pi = r) \leq \alpha \right\}$$

an LP in π

Optimal classifier: under some continuity assumptions

$$\widehat{A}_{\pi^*} = \begin{cases} r & \max_{k \in [K]} p_k(\boldsymbol{X}) \leq \lambda(\alpha, \mathbb{P}) \\ \arg\max_{k \in [K]} p_k(\boldsymbol{X}) & \text{otherwise} \end{cases} .$$

(Chow, 1970; Denis and Hebiri, 2015)

Sometimes we might want more: fairness

Observations: $(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times [L] \times [K]$

Prediction kernel: $\pi : \mathcal{X} \to \Delta([K])$
Prediction: $\widehat{A}_\pi \mid \boldsymbol{X} \sim \pi(\cdot \mid \boldsymbol{X})$ and $(\widehat{A}_\pi \perp\!\!\!\perp Y) \mid \boldsymbol{X}$

Possible goal: perform as well as

$$\pi^\star \in \arg\min_{\pi} \left\{ \mathcal{R}(\pi) = \mathbb{P}(Y \neq \widehat{A}_\pi) \ : \ C_{s,y}(\pi) \leq \alpha_{s,y} \quad \forall (s,y) \right\}$$

with $C_{s,y}(\pi) = |\mathbb{P}(\widehat{A}_\pi = y \mid S = s) - \mathbb{P}(\widehat{A}_\pi = y)|$ an LP in π

Under continuity assumptions, the form of $\pi^\star$ is again deterministic

proven and re-proven in many papers

Sometimes we might want more: set-valued

Data: $(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times [K]$ $\qquad p_k(\boldsymbol{X}) = \mathbb{P}(Y = k \mid \boldsymbol{X})$

Prediction kernel: $\pi_k : \mathcal{X} \to [0, 1]$ and $\pi = (\pi_1, \ldots, \pi_K)$

Flip a coin for each class: $\widehat{Y}_{\pi,k} \mid \boldsymbol{X} \sim \pi_k(\cdot \mid \boldsymbol{X})$ and $(\widehat{Y}_{\pi,k} \perp\!\!\!\perp Y) \mid \boldsymbol{X}$

Set-valued prediction: $\Gamma_\pi = \{k \in [K] \,:\, \widehat{Y}_{\pi,k} = 1\}$

Possible goal: perform as well as

$$\pi^\star \in \arg\min_{\pi} \{\mathcal{R}(\pi) = \mathbb{P}(Y \notin \Gamma_\pi) \,:\, \mathbb{E}|\Gamma_\pi| \leq B\}$$

an LP in π

Under continuity assumptions

$$\Gamma_{\pi^\star} = \{k \in [K] \,:\, p_k(\boldsymbol{X}) \geq G^{-1}(B)\} \quad \text{where} \quad G(t) = K - \sum_{k=1}^{K} \mathbb{P}(p_k(\boldsymbol{X}) \leq t)$$

adapted from (Denis and Hebiri, 2017; Sadinle, Lei, and Wasserman, 2018)

Depending on the goals one might want to consider an arbitrary combination of constraints

General setup

Observations: $(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times [K]$

Prediction kernel: $\pi : \mathcal{X} \to \Delta(\mathcal{A})$
Prediction: $\widehat{A}_\pi \mid \boldsymbol{X} \sim \pi(\cdot \mid \boldsymbol{X})$ and $(\widehat{A}_\pi \perp\!\!\!\perp Y) \mid \boldsymbol{X}$

Some risk: $\pi \mapsto \mathcal{R}(\pi)$

$$\mathcal{R}(\pi) = \mathbb{E}[r(\boldsymbol{X}, \widehat{A}_\pi)]$$

Some constraints: $\pi \mapsto C_j(\pi)$

$$C_j(\pi) = \mathbb{E}[c_j(\boldsymbol{X}, \widehat{A}_\pi)] \qquad \text{for } j = 1, \dots, m$$

Goal: "mimic" the optimal classifier

$$\pi^\star \in \arg\min_\pi \{\mathcal{R}(\pi) \,:\, C_j(\pi) \leq 0 \quad \forall j \in [m]\}$$

an LP in π

How to treat the general case?

$$\mathcal{R}(\pi) = \mathbb{E}[r(\boldsymbol{X}, \widehat{A}_\pi)] \qquad C_j(\pi) = \mathbb{E}[c_j(\boldsymbol{X}, \widehat{A}_\pi)]$$

Goal: "mimic" the optimal classifier

$$\pi^\star \in \arg\min_\pi \{\mathcal{R}(\pi) \,:\, C_j(\pi) \leq 0 \quad \forall j \in [m]\}$$

Plan

1. Assuming that r and c_j are known but $\mathbb{P}_{\boldsymbol{X}}$ is not, how to solve the above problem in distribution-free manner and what can we guarantee?

2. Assuming estimators $\hat{r}$ and $\hat{c}_j$ are available, what can we do?

3. Some numerical results on fairness example

Generalized Neyman-Pearson lemma

NP is a statement about LP

Setup: for now $\mathbb{E}$ unknown, r, c_j known
Primal problem

$$\min_{\pi} \{\mathcal{R}(\pi) \,:\, C_j(\pi) \leq 0 \quad \forall j \in [m]\}$$

Dual problem

$$\max_{\boldsymbol{\lambda} \geq 0} \left\{ \mathbb{E} \min_{a \in \mathcal{A}} \left(r(\boldsymbol{X}, a) + \sum_{j=1}^{m} \lambda_j c_j(\boldsymbol{X}, a) \right) := F(\boldsymbol{\lambda}) \right\}$$

Primal-dual link (under continuity assumptions)

$$\widehat{A}_{\pi^\star_{\boldsymbol{\lambda}}} = \arg\min_{a \in \mathcal{A}} \left(r(\boldsymbol{X}, a) + \sum_{j=1}^{m} \lambda_j c_j(\boldsymbol{X}, a) \right)$$

(Dantzig and Wald, 1951)

Generalized Neyman-Pearson lemma: algorithmic implications

$$\max_{\boldsymbol{\lambda}\geq 0}\left\{\mathbb{E}\min_{a\in\mathcal{A}}\left(r(\boldsymbol{X},a)+\sum_{j=1}^{m}\lambda_j c_j(\boldsymbol{X},a)\right) := F(\boldsymbol{\lambda})\right\}$$

Primal-dual link (under continuity assumptions)

$$\widehat{A}_{\pi^\star_{\boldsymbol{\lambda}}} = \arg\min_{a\in\mathcal{A}}\left(r(\boldsymbol{X},a)+\sum_{j=1}^{m}\lambda_j c_j(\boldsymbol{X},a)\right)$$

Perform simple ERM-like algorithm:

$$\widehat{\boldsymbol{\lambda}} \in \arg\max_{\boldsymbol{\lambda}\geq 0}\left\{\widehat{\mathbb{E}}\min_{a\in\mathcal{A}}\left(r(\boldsymbol{X},a)+\sum_{j=1}^{m}\lambda_j c_j(\boldsymbol{X},a)\right) := \hat{F}(\boldsymbol{\lambda})\right\}$$

set

$$\widehat{A}_{\pi^\star_{\widehat{\boldsymbol{\lambda}}}} = \arg\min_{a\in\mathcal{A}}\left(r(\boldsymbol{X},a)+\sum_{j=1}^{m}\hat{\lambda}_j c_j(\boldsymbol{X},a)\right)$$

Issues of this approach

$$\widehat{\boldsymbol{\lambda}} \in \arg\max_{\boldsymbol{\lambda} \geq 0} \left\{ \widehat{\mathbb{E}} \min_{a \in \mathcal{A}} \left(r(\boldsymbol{X}, a) + \sum_{j=1}^{m} \lambda_j c_j(\boldsymbol{X}, a) \right) := \hat{F}(\boldsymbol{\lambda}) \right\}$$

1. This approach relies on some kind of continuity assumption; in the absence of it one has to do some strange tricks (e.g., add small perturbations)

2. $\boldsymbol{\lambda} \mapsto \hat{F}(\boldsymbol{\lambda})$ is concave but non-smooth (issue comes from the min)

3. Statistical analysis typically goes through $0 \in \partial \hat{F}(\widehat{\boldsymbol{\lambda}})$

4. The condition $0 \in \partial \hat{F}(\widehat{\boldsymbol{\lambda}})$ is nearly impossible to guarantee for non-smooth problems (think about $x \mapsto |x|$)

5. Combining with unknown r and c_j is possible, but not very elegant

Proposition: Nesterov's smoothing (Nesterov, 2005)

Assumptions: for now $\mathbb{E}$ unknown, r, c_j known

Primal problem

$$\min_{\pi} \{\mathcal{R}(\pi) \,:\, C_j(\pi) \leq 0 \quad \forall j \in [m]\}$$

Smoothed dual

$$\max_{\boldsymbol{\lambda} \geq 0} \left\{ \mathbb{E}\,\mathrm{SoftMin}_\beta \left(r(\boldsymbol{X}, a) + \sum_{j=1}^{m} \lambda_j c_j(\boldsymbol{X}, a) \right) := F_\beta(\boldsymbol{\lambda}) \right\}$$

Smoothed-primal-dual link

$$\pi^\beta_{\boldsymbol{\lambda}}(a \mid \boldsymbol{x}) \propto \exp\left(-\beta \left\{ r(\boldsymbol{x}, a) + \sum_{j=1}^{m} \lambda_j c_j(\boldsymbol{x}, a) \right\} \right)$$

Observations

1. F_β is still concave
2. F_β has β-Lipschitz gradient!
3. We are loosing determinism, $\pi^\beta_{\boldsymbol{\lambda}}$ is defuse

What can we guarantee?

β-Dual problem

$$\max_{\boldsymbol{\lambda} \geq 0} \left\{ \mathbb{E}\, \mathrm{SoftMin}_\beta \left(r(\boldsymbol{X}, a) + \sum_{j=1}^{m} \lambda_j c_j(\boldsymbol{X}, a) \right) := F_\beta(\boldsymbol{\lambda}) \right\}$$

β-Primal-dual link $\pi^\beta_{\boldsymbol{\lambda}}(a \mid \boldsymbol{x}) \propto \exp\left(-\beta \left\{ r(\boldsymbol{x}, a) + \sum_{j=1}^{m} \lambda_j c_j(\boldsymbol{x}, a) \right\} \right)$

Proposition (with a slight lie)

For any $\boldsymbol{\lambda} \geq 0$

$$\sum_{j=1}^{M} \max\left(C_j(\pi^\beta_{\boldsymbol{\lambda}}),\, 0 \right)^2 \leq \|\nabla F_\beta(\boldsymbol{\lambda})\|^2$$

$$\mathcal{R}(\pi^\beta_{\boldsymbol{\lambda}}) \leq \mathcal{R}(\pi^\star) + \|\boldsymbol{\lambda}\| \cdot \|\nabla F_\beta(\boldsymbol{\lambda})\| + \frac{\log(\mathcal{A})}{\beta}$$

Question: How to find $\boldsymbol{\lambda}$ with small $\|\nabla F_\beta(\boldsymbol{\lambda})\|$?

Stochastic convex optimization with stationarity guarantees

The problem we want to solve with unknown $\mathbb{E}$

$$\max_{\boldsymbol{\lambda} \geq 0} \left\{ \mathbb{E}\, \mathrm{SoftMin}_\beta \left(r(\boldsymbol{X}, a) + \sum_{j=1}^{m} \lambda_j c_j(\boldsymbol{X}, a) \right) := F_\beta(\boldsymbol{\lambda}) \right\}$$

Observation: Sample one $\boldsymbol{X}$ from $\mathbb{P}_{\boldsymbol{X}}$

$$\nabla_{\boldsymbol{\lambda}}\, \mathrm{SoftMin}_\beta \left(r(\boldsymbol{X}, a) + \sum_{j=1}^{m} \lambda_j c_j(\boldsymbol{X}, a) \right)$$

is an unbiased gradient estimator of $\nabla F_\beta(\boldsymbol{\lambda})$

Goal: Using N i.i.d. samples from $\mathbb{P}_{\boldsymbol{X}}$ can we find $\widehat{\boldsymbol{\lambda}}$ such that

$$\mathbf{E}\|\nabla F_\beta(\widehat{\boldsymbol{\lambda}})\|^2 \quad \text{is small?}$$

N.B. A more typical quantity to control in the case of convex problems is functionality gap

$$\mathbf{E}\left[F_\beta^\star - F_\beta(\widehat{\boldsymbol{\lambda}}) \right]$$

Enforcing constraint in unsupervised fashion

SGD3 and its' variants are tailored exactly for this setup (Allen-Zhu, 2021; Foster et al., 2019)

Theorem (case of known r and c_j)

Using a fancy version of SGD and having access to N i.i.d. samples from $\mathbb{P}_{\boldsymbol{X}}$ with $\beta = N$, we can return $\widehat{\boldsymbol{\lambda}} \geq 0$ such that

$$\mathbf{E}\left[\sum_{j=1}^{M} \max\left(C_j(\pi^{\beta}_{\widehat{\boldsymbol{\lambda}}}), 0\right)^2\right] \leq O(1/N)$$

$$\mathbf{E}[\mathcal{R}(\pi^{\beta}_{\widehat{\boldsymbol{\lambda}}})] \leq \mathcal{R}(\pi^{\star}) + O(1/\sqrt{N})$$

- There are problem dependent constants related to matrix norm of constraints
- The parameter $\beta = N$ is very high, so for large N the classifier $\pi^{\beta}_{\widehat{\boldsymbol{\lambda}}}$ is very close to deterministic

The case of unknown r and c_j: unlabeled post-processing

Assume that estimators $\hat{r}$ and $\hat{c}_j$ are available (e..g, via a labeled dataset)

Theorem

Using a fancy version of `SGD` and having access to N i.i.d. samples from $\mathbb{P}_{\boldsymbol{X}}$ with $\beta = N$, we can return $\widehat{\boldsymbol{\lambda}} \geq 0$ such that

$$\mathbf{E}\left[\sum_{j=1}^{M} \max\left(C_j(\pi^{\beta}_{\widehat{\boldsymbol{\lambda}}}), 0\right)^2\right] \leq O(1/N) + \sum_j \mathbf{E}\|c_j - \hat{c}_j\|^2$$

$$\mathbf{E}[\mathcal{R}(\pi^{\beta}_{\widehat{\boldsymbol{\lambda}}})] \leq \mathcal{R}(\pi^{\star}) + O(1/\sqrt{N}) + \sum_j \mathbf{E}\|c_j - \hat{c}_j\| + \mathbf{E}\|r - \hat{r}\|$$

Unlabeled post-processing interpretation:

- Estimators $\hat{r}, \hat{c}$ typically depend only on estimators of $\mathbb{P}(Y = k \mid \boldsymbol{X})$
- Given initial estimator of $\mathbb{P}(Y = k \mid \boldsymbol{X})$ above algorithm transforms it into a randomized classifier "satisfying" constraints
- Quality of post-processing depends separately on: quality of your initial estimation; number of unlabeled data

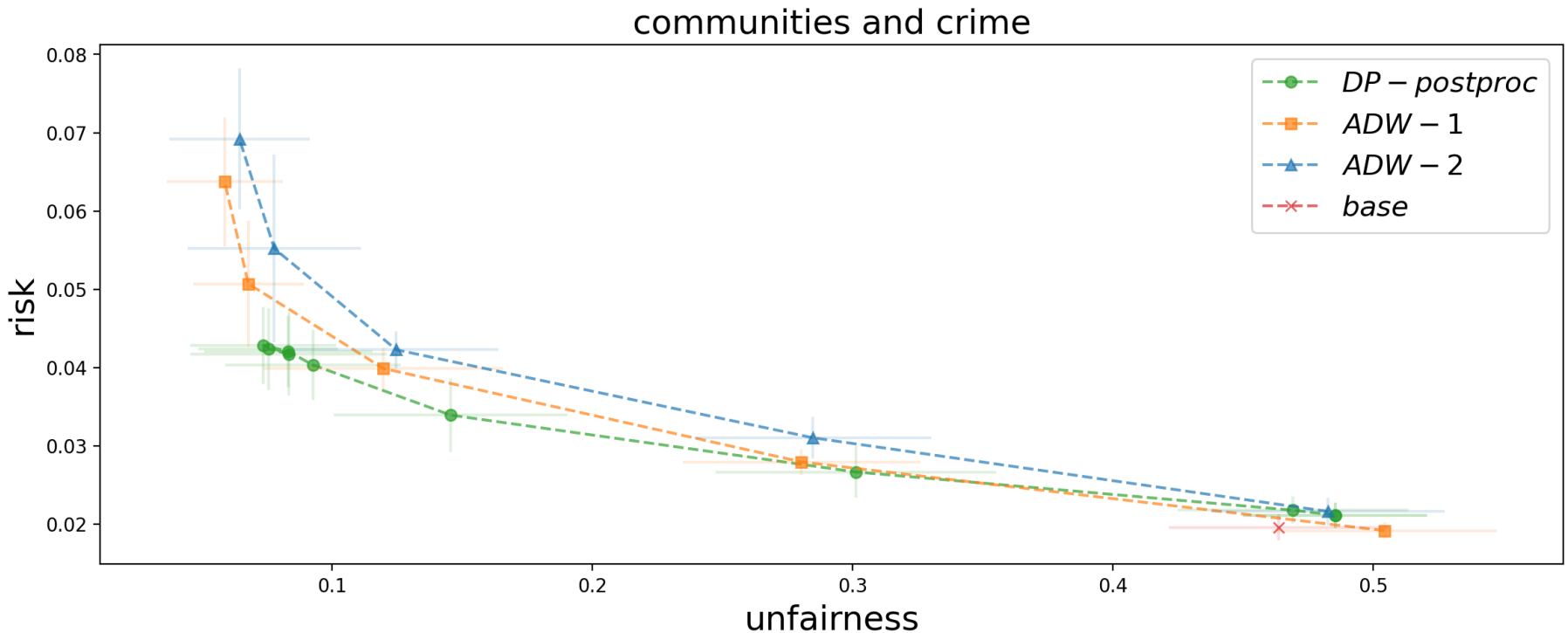
Some numerics

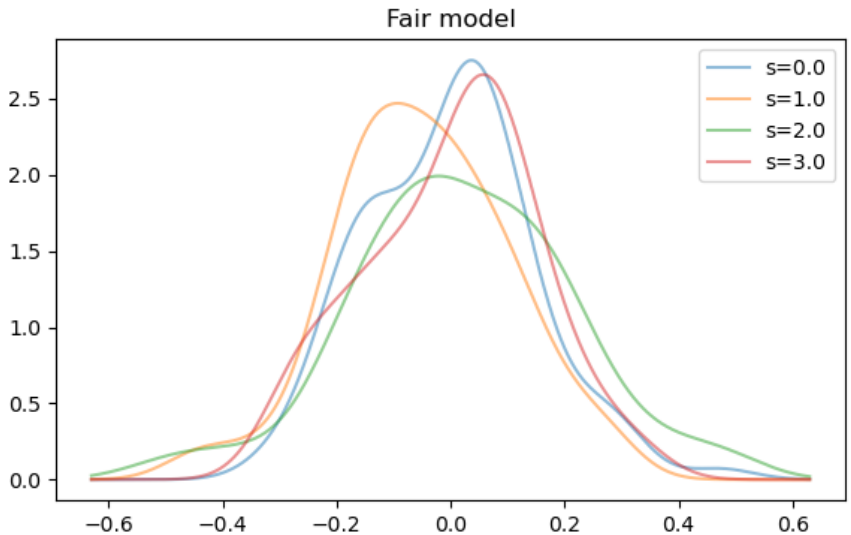
Observations: $(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}\underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times [4] \times [K]$

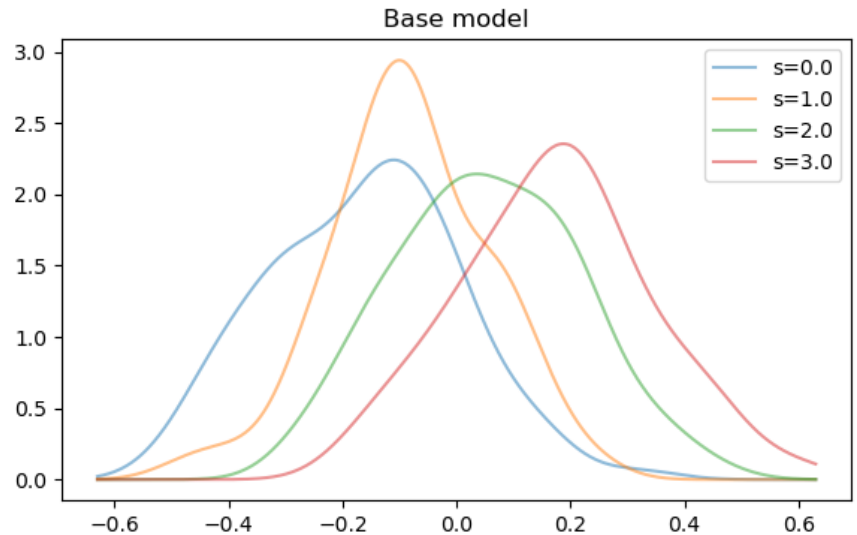
Risk: $\mathbb{E}(f(\boldsymbol{X}) - Y)^2$

Constraints: $f(\boldsymbol{X}) \perp\!\!\!\perp S$

Strategy: discretize the output domain of f and reduce to multi-class classification




Fair model
s=0.0
s=1.0
s=2.0
s=3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6


Base model
s=0.0
s=1.0
s=2.0
s=3.0
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

Thank you

1. Many constraint learning problems are LPs in the space of randomized classifiers

2. One can build estimators using Generalized Neyman-Pearson lemma, but this approach is not very operational

3. We propose a general purpose post-processing algorithm using Nesterov's smoothing

4. The quantity is norm of the gradient even though the considered function is convex

Seems to work

Bibliography I

Allen-Zhu, Z. (2021). *How To Make the Gradients Small Stochastically: Even Faster Convex and Nonconvex SGD*. arXiv: 1801.02982 [cs.LG].

Chow, C.-K. (1970). "On optimum error and reject trade-off". In: *IEEE Transactions on Information Theory* 16, pp. 41–46.

Dantzig, George B and Abraham Wald (1951). "On the fundamental lemma of Neyman and Pearson". In: *The Annals of Mathematical Statistics* 22.1, pp. 87–93.

Denis, C. and M. Hebiri (2015). "Confidence Sets for Classification". In: *International Symposium on Statistical Learning and Data Sciences*. Springer, pp. 301–312.

– (2017). "Confidence sets with expected sizes for Multiclass Classification". In: *JMLR* 18.1, pp. 3571–3598.

Foster, D. et al. (2019). "The complexity of making the gradient small in stochastic convex optimization". In: *Conference on Learning Theory*. PMLR, pp. 1319–1345.

Nesterov, Y. (2005). "Smooth minimization of non-smooth functions". In: *Mathematical programming* 103.3, pp. 127–152.

Bibliography II

Sadinle, M., J. Lei, and L. Wasserman (2018). "Least Ambiguous Set-Valued Classifiers With Bounded Error Levels". In: *Journal of the American Statistical Association*, pp. 1–12.

Allen-Zhu, Z. (2021). *How To Make the Gradients Small Stochastically: Even Faster Convex and Nonconvex SGD*. arXiv: 1801.02982 [cs.LG].

Chow, C.-K. (1970). "On optimum error and reject trade-off". In: *IEEE Transactions on Information Theory* 16, pp. 41–46.

Dantzig, George B and Abraham Wald (1951). "On the fundamental lemma of Neyman and Pearson". In: *The Annals of Mathematical Statistics* 22.1, pp. 87–93.

Denis, C. and M. Hebiri (2015). "Confidence Sets for Classification". In: *International Symposium on Statistical Learning and Data Sciences*. Springer, pp. 301–312.

– (2017). "Confidence sets with expected sizes for Multiclass Classification". In: *JMLR* 18.1, pp. 3571–3598.

Foster, D. et al. (2019). "The complexity of making the gradient small in stochastic convex optimization". In: *Conference on Learning Theory*. PMLR, pp. 1319–1345.

Bibliography III

Nesterov, Y. (2005). "Smooth minimization of non-smooth functions". In: *Mathematical programming* 103.3, pp. 127–152.

Sadinle, M., J. Lei, and L. Wasserman (2018). "Least Ambiguous Set-Valued Classifiers With Bounded Error Levels". In: *Journal of the American Statistical Association*, pp. 1–12.