Neyman-Pearson classification

under a strict constraint

Philippe Rigollet

Princeton University

with Xin Tong (Princeton)

Binary classification

- $(X, Y) \in \mathbb{R}^d \times \{-1, 1\}$ random couple
- Goal: construct a classifier $h : \mathbb{R}^d \to \{-1, 1\}$.
- Performance is usually measured by classification error:

$$R(h) = \mathbb{P}[h(X) \neq Y] = \mathbb{P}[-h(X)Y \geq 0]$$

- We may be interested in errors for each class $\{-1, +1\}$.

Type I error

$$R^-(h) = \mathbb{P}\left[-h(X)Y \geq 0 | Y = -1\right]$$

Type II error

$$R^+(h) = \mathbb{P}\left[-h(X)Y \geq 0 | Y = +1\right]$$

Neyman-Pearson paradigm

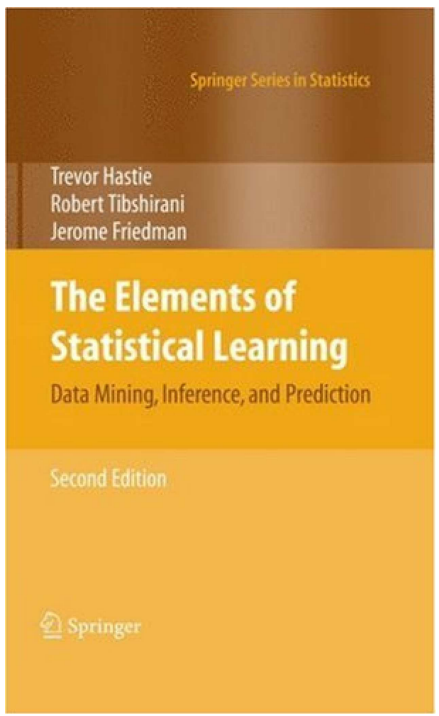
- Two quantities to control simultaneously.
- Classification error is given by

$$R(h) = \mathbb{P}(Y = -1)R^-(h) + \mathbb{P}(Y = 1)R^+(h)$$

  $\rightsquigarrow$ convex combination of the two errors
- The Neyman-Pearson paradigm offers a systematic way to control two types of error.
- Idea: fix one error under a user-specified level $\alpha \in (0, 1)$ and minimize the other error under this constraint.
- Introduces asymmetry in the two types of error.

Asymmetry in the errors

- Why introduce asymmetry in the errors? Classification error is simple and natural...
- But it may not be adapted the a given classification task.
- Let us look at some examples from a best-seller (hereafter FHT):



Some examples of binary classification from FHT

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.

Some examples of binary classification from FHT

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.

  Type I error: predict no second heart attack

Some examples of binary classification from FHT

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.

  Type I error: predict no second heart attack

- Phoneme Recognition: "aa" Vs "ao" based on time series.

Some examples of binary classification from FHT

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.

  Type I error: predict no second heart attack

- Phoneme Recognition: "aa" Vs "ao" based on time series.

  Type I error: who cares?

Some examples of binary classification from FHT

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.

  Type I error: predict no second heart attack

- Phoneme Recognition: "aa" Vs "ao" based on time series.

  Type I error: who cares?

- Use systolic blood pressure to predict coronary heart disease (CHD).

Some examples of binary classification from FHT

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.

  Type I error: predict no second heart attack

- Phoneme Recognition: "aa" Vs "ao" based on time series.

  Type I error: who cares?

- Use systolic blood pressure to predict coronary heart disease (CHD).

  Type I error: predict no CHD

Some examples of binary classification from FHT

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.

  Type I error: predict no second heart attack

- Phoneme Recognition: "aa" Vs "ao" based on time series.

  Type I error: who cares?

- Use systolic blood pressure to predict coronary heart disease (CHD).

  Type I error: predict no CHD

- Use word counts to predict SPAMs

Some examples of binary classification from FHT

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.

  Type I error: predict no second heart attack

- Phoneme Recognition: "aa" Vs "ao" based on time series.

  Type I error: who cares?

- Use systolic blood pressure to predict coronary heart disease (CHD).

  Type I error: predict no CHD

- Use word counts to predict SPAMs

  Type I error: depends on user

Classification error may be dangerous

- Classification error is given by

$$R(h) = \mathbb{P}(Y = -1)R^{-}(h) + \mathbb{P}(Y = 1)R^{+}(h)$$

- Clearly if one class is under-represented (e.g. $\mathbb{P}(Y = -1)$ is very small) then type I error gets a lower weight and may be overlooked in the classification task.
- This is typically the case in anomaly detection where the observations consist mostly ($> 95\%$) of normal observations.

Observations

- To construct our estimator we have observations.
- Two sampling schemes can be considered:
  1. $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d copies of (X, Y) and $\mathbb{P}(Y = +1) = p$
  2. Two i.i.d. samples $X_1^-, \ldots, X_{n^-}^-$, and $X_1^+, \ldots, X_{n^+}^+$, where the sample sizes n^- and n^+ are deterministic.
- Note that in case 2. the two samples need not be mutually independent.
- We focus on case 1. and case 2. is even simpler.
- We denote by $X_1^-, \ldots, X_{N^-}^-$ and $X_1^+, \ldots, X_{N^+}^+$, the (random) partitioning of the sample of features according to their label.

Previous work

- Introducing asymmetry in classification errors is not new.
- Some relevant papers:
  1. Learning with the Neyman-Pearson and min-max criteria. (2002). Cannon *et al.*
  2. A Neyman-Pearson approach to statistical learning. (2005). Scott and Nowak.
  3. Comparison and design of Neyman-Pearson classifiers. (2005). Scott.
  4. Performance measures for Neyman-Pearson classification. (2007). Scott.

  Note that all of them mention Neyman-Pearson but none implements it strictly speaking.

Idea of Cannon *et al.*: relaxed constraint

- Fix a constant $\varepsilon_0 > 0$
- $\mathcal{H}$: given set of classifiers with finite VC dimension.
- Solve

$$\min_{\substack{h \in \mathcal{H} \\ \hat{R}^-(h) \leq \alpha + \varepsilon_0}} \hat{R}^+(h),$$

where

$$\hat{R}^-(h) = \frac{1}{N^-} \sum_{i=1}^{N^-} \mathbb{1}(h(X_i^-) \geq 0)\,,$$

$$\hat{R}^+(h) = \frac{1}{N^+} \sum_{i=1}^{N^+} \mathbb{1}(h(X_i^+) \leq 0)$$

denote the empirical type I and type II errors respectively.

Idea of Cannon *et al.*: relaxed constraint

- Fix a constant $\varepsilon_0 > 0$
- $\mathcal{H}$: given set of classifiers with finite VC dimension.
- Solve

$$\min_{\substack{h \in \mathcal{H} \\ \hat{R}^-(h) \leq \alpha + \varepsilon_0}} \hat{R}^+(h),$$

Ensures that solution has type I error $\leq \alpha$ whp

where

$$\hat{R}^-(h) = \frac{1}{N^-} \sum_{i=1}^{N^-} \mathbb{1}(h(X_i^-) \geq 0)\,,$$

$$\hat{R}^+(h) = \frac{1}{N^+} \sum_{i=1}^{N^+} \mathbb{1}(h(X_i^+) \leq 0)$$

denote the empirical type I and type II errors respectively.

Comments

- The relaxed constraint $\hat{R}^{-}(h) \leq \alpha + \varepsilon_0$ potentially yields classifiers with true type I error strictly larger than α.
- This goes against the Neyman-Pearson paradigm.
- We actually want to strengthen the constraint:

$$\hat{R}^{-}(h) \leq \alpha - \varepsilon_0$$

Empirical risk

Two motivations for not using empirical risk:

1. Non smooth, non convex optimization problem both in objective and constraint.
2. Cannot control type II error in a distribution free way (proved later).

Instead, we use a convex surrogate.

Contribution of this talk

- A classification procedure based on convex optimization that strictly implements the NP paradigm.
- Theoretical insights on its performance (oracle inequalities)
- Some consequences on chance constrained optimization

Convexification

- $\varphi : [-1, 1] \to \mathbb{R}^+$ is a convex surrogate if $\varphi \nearrow$, continuous, convex and if $\varphi(0) = 1$. Dominates $\mathbb{1}_{\geq 0}(\cdot)$

  - hinge loss $\varphi(x) = (1 + x)_+$
  - logit loss $\varphi(x) = \log_2(1 + e^x)$
  - exponential loss $\varphi(x) = e^x$

- φ-risk:

$$R(h) = \mathbb{E}[\mathbb{1}_{\geq 0}(-Yh(X))] \quad \rightsquigarrow \quad R_\varphi(h) = \mathbb{E}[\varphi(-Yh(X))]$$

- Convex class of classifiers: Given $h_1, \ldots, h_M, M \geq 2$:

$$\mathcal{H}^{\text{conv}} = \{\mathsf{h}_\lambda = \sum_{j=1}^{M} \lambda_j h_j, \lambda \in \Lambda\},$$

where is the flat simplex of $\mathbb{R}^M$
(convex combinations = weighted majority votes).

Convexified NP problem

- Our goal is to "mimic" the solution of

$$\min_{\substack{h \in \mathcal{H}^{\text{conv}} \\ R_\varphi^-(h) \leq \alpha}} R_\varphi^+(h),$$

where

$$R_\varphi^-(h) = \mathbb{E}\left[\varphi(-Yh(X)) | Y = -1\right]$$

and

$$R_\varphi^+(h) = \mathbb{E}\left[\varphi(-Yh(X)) | Y = 1\right].$$

- Find $\hat{h}$ such that with high probability:
  1. $R_\varphi^-(\hat{h}) \leq \alpha$ ... not $\alpha + \varepsilon_0$
  2. $R_\varphi^+(\hat{h}) \approx \min_{\substack{h \in \mathcal{H}^{\text{conv}} \\ R_\varphi^-(h) \leq \alpha}} R_\varphi^+(h)$

Neyman-Pearson classifier

- To ensure that the constraint on type I error is satisfied, we strengthen the constraint
- We propose the following classifier:

$$\tilde{h}^{\tau} = \underset{\substack{h \in \mathcal{H}^{\text{conv}} \\ \hat{R}_{\varphi}^{-}(h) \leq \alpha - \tau/\sqrt{N^{-}}}}{\operatorname{argmin}} \hat{R}_{\varphi}^{+}(h)\,.$$

  where ties are broken arbitrarily.
- We call it NP classifier

Theorem

Fix constants $\delta, \alpha \in (0,1), L > 0$ and let $\varphi : [-1,1] \to \mathbb{R}^+$ be a given L-Lipschitz convex surrogate. Define

$$\tau = 4\sqrt{2}L\sqrt{\log\left(\frac{2M}{\delta}\right)}\,.$$

Then for any classifier $h \in \mathcal{H}^{\text{conv}}$ that satisfies $\hat{R}_\varphi^-(h) \leq \alpha_\tau$, we have

$$R^-(h) \leq R_\varphi^-(h) \leq \alpha\,,$$

with probability at least $(1-\delta)(1-e^{-\frac{n(1-p)^2}{2}})$.

Control of Type I error

An assumption

- To control the type II error, we need to make the following assumption.
  There exists a positive constant $\varepsilon < 1$ such that the set of classifiers

$$\left\{h \in \mathcal{H}^{\mathrm{conv}} \,:\, R_{\varphi}^{-}(h) \leq \varepsilon\alpha\right\}$$

  is nonempty.
  $\rightsquigarrow$ strengthening the constraint makes sense.
- Let n_0 be the smallest integer such that

$$n_0 \geq \left(\frac{26\tau}{(1-\varepsilon)\alpha\sqrt{1-p}}\right)^2 .$$

Theorem

if $n > n_0$, we have with probability $(1-2\delta)(1-e^{-\frac{n(1-p)^2}{2}})$,

$$R^-(\tilde{h}_n^\tau) \leq R_\varphi^-(\tilde{h}_n^\tau) \leq \alpha$$

and

$$R_\varphi^+(\tilde{h}_n^\tau) - \min_{\substack{h \in \mathcal{H}^{\text{conv}} \\ R_\varphi^-(h) \leq \alpha}} R_\varphi^+(h) \leq \frac{4\varphi(1)\tau}{(1-\bar{\varepsilon})\alpha\sqrt{N^-}} + \frac{2\tau}{\sqrt{N^+}}\,.$$

Moreover, with probability $1 - 2\delta - e^{-\frac{n(1-p)^2}{2}} - e^{-\frac{np^2}{2}}$, we have also

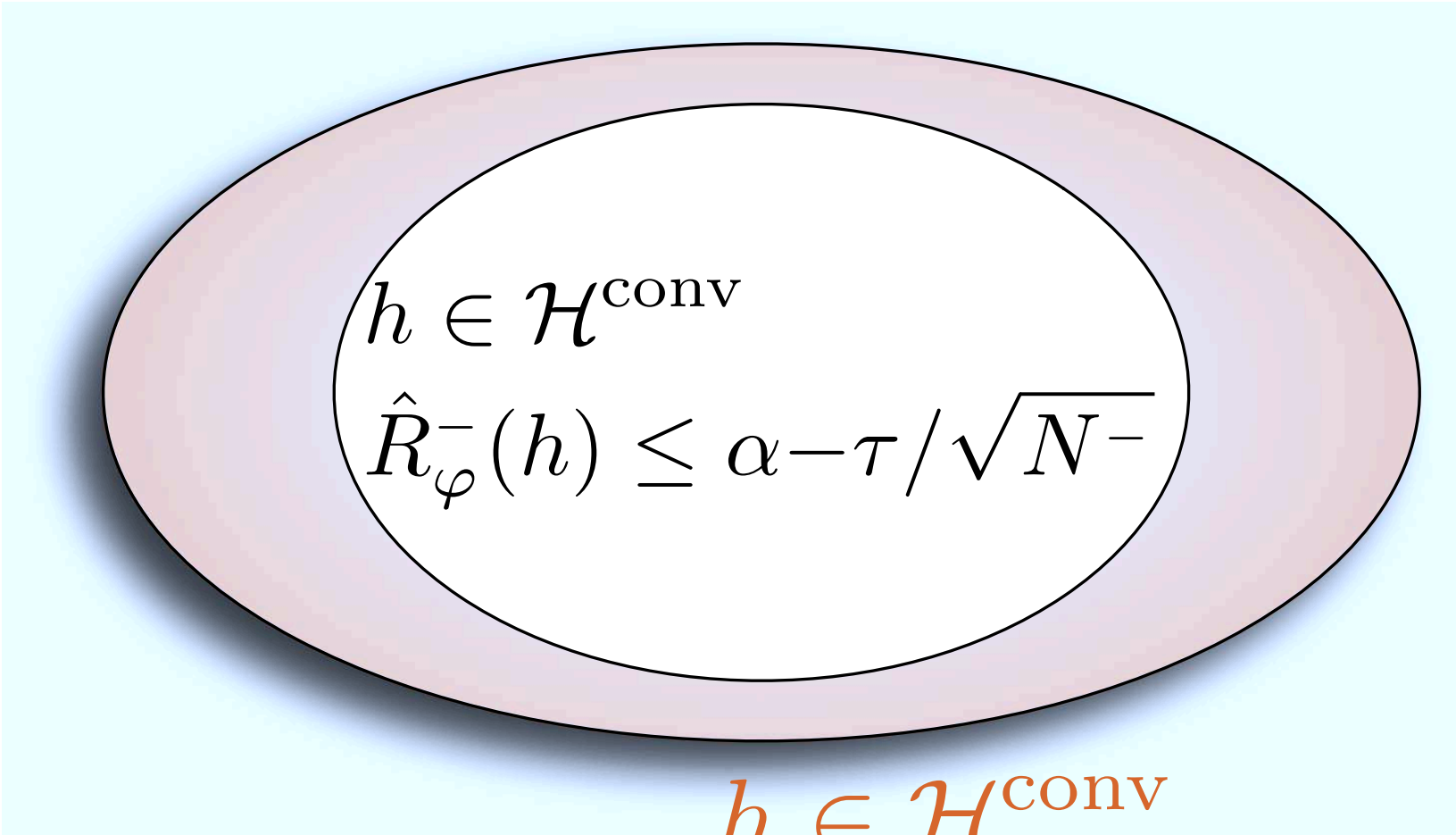
$$R_\varphi^+(\tilde{h}_n^\tau) - \min_{\substack{h \in \mathcal{H}^{\text{conv}} \\ R_\varphi^-(h) \leq \alpha}} R_\varphi^+(h) \leq \frac{4\sqrt{2}\varphi(1)\tau}{(1-\varepsilon)\alpha\sqrt{n(1-p)}} + \frac{2\sqrt{2}\tau}{\sqrt{np}}\,.$$

Control of Type II error

Sketch of the proof

- Control of type I error is classical empirical process theory: symmetrization, contraction, geometry of the simplex
- Control of type II is more tricky. Involves sensitivity of optimal value of a convex program.
- Surprisingly it does not imply the Lipschitz constant but only convexity of φ.

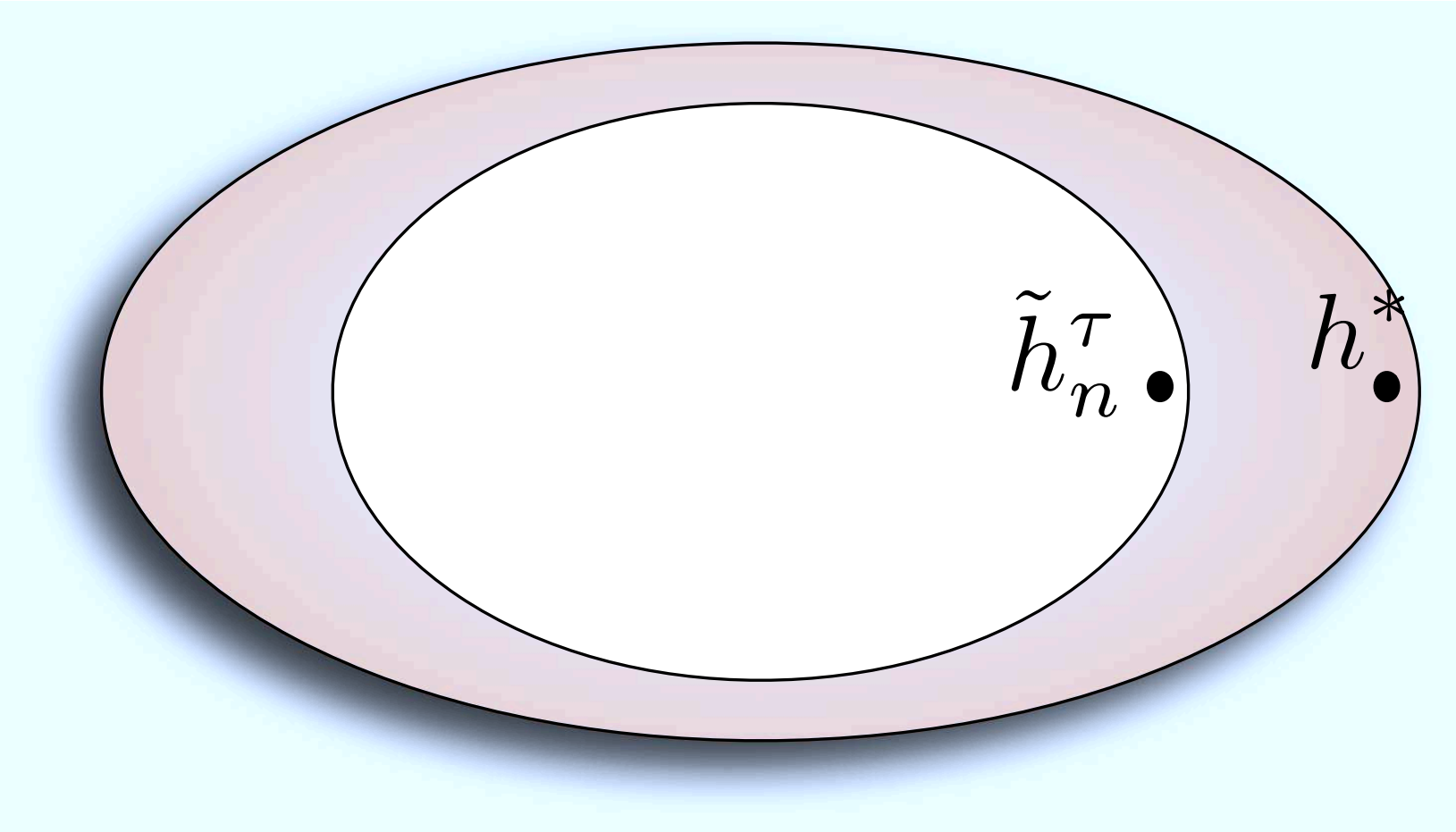
Some intuition



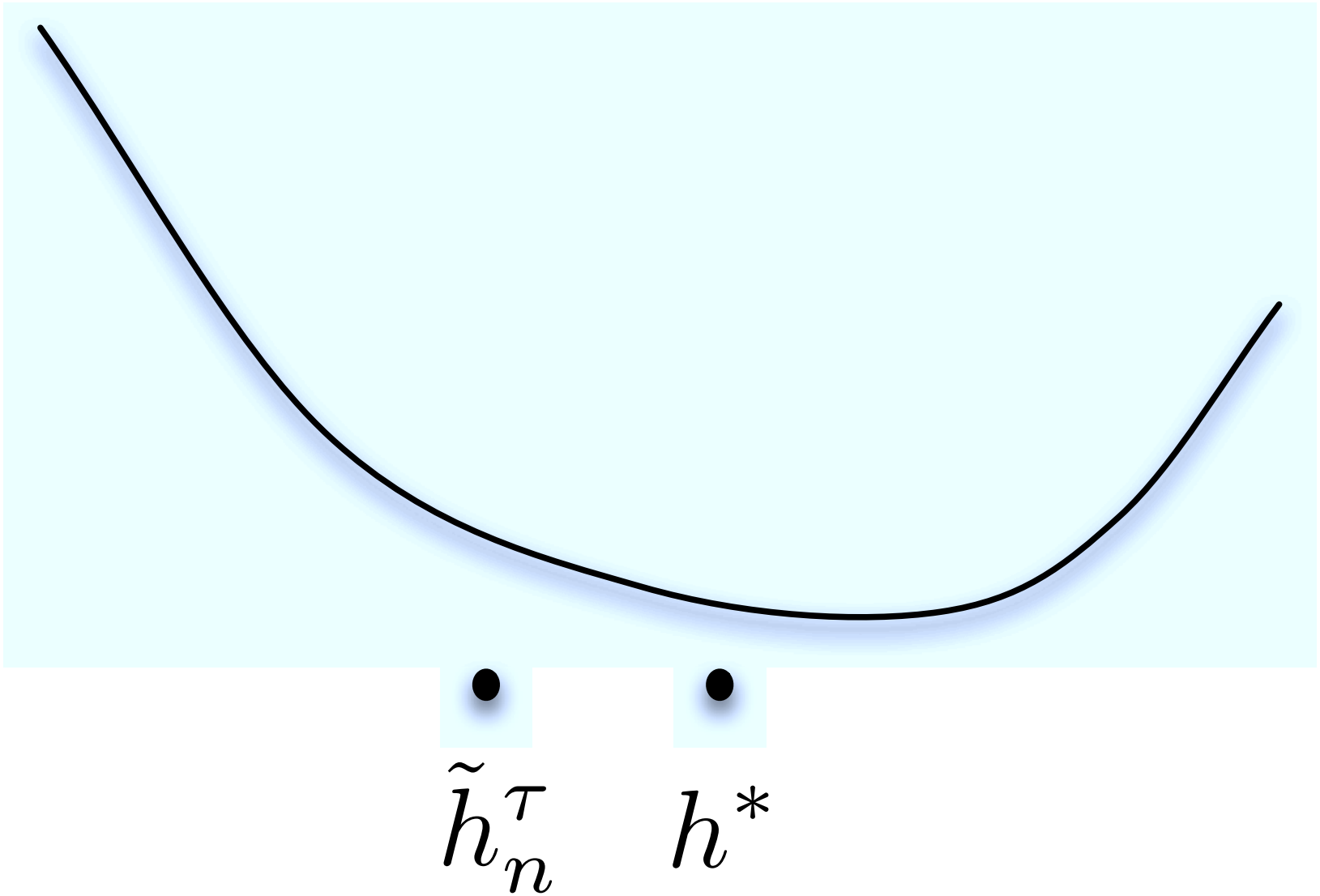
$$h \in \mathcal{H}^{\text{conv}}$$

$$R_\varphi^-(h) \leq \alpha$$

Some intuition



Some intuition



Proposition

Assume that there exists $\nu_0 > 0$ such that the set of classifiers $\left\{h \in \mathcal{H}^{\text{conv}} \,:\, R_\varphi^-(h) \leq \alpha - \nu_0\right\}$ is nonempty. Then, for any $\nu \in (0, \nu_0)$,

$$\min_{\substack{h \in \mathcal{H}^{\text{conv}} \\ R_\varphi^-(h) \leq \alpha - \nu}} R_\varphi^+(h) - \min_{\substack{h \in \mathcal{H}^{\text{conv}} \\ R_\varphi^-(h) \leq \alpha}} R_\varphi^+(h) \leq \varphi(1) \frac{\nu}{\nu_0 - \nu}\,.$$

Hinges on the fact that the function $\gamma(\alpha) = \inf_{\substack{h \in \mathcal{H}^{\text{conv}} \\ R_\varphi^-(h) \leq \alpha}} R_\varphi^+(h)$, is a non-increasing convex function on $[0, 1]$. (See Lehmann).

Chance constrained optimization

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad F(\lambda) \leq 0,$$

Chance constrained optimization

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad \mathbb{P}\{F(\lambda, \xi) \leq 0\} \geq 1 - \alpha,$$

where $\xi \in \Xi$ is a random vector, $\Lambda \subset \mathbb{R}^M$ is convex, $\alpha > 0$ and f is convex and deterministic.

- Relaxation of robust optimization ($\alpha = 0$). See Ben-Tal *et al.* (2009).
- Not a convex problem in general. Nemirovski and Shapiro (2006) proposed to use the following conservative approximation to this problem:

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad \mathbb{E}\{\varphi[F(\lambda, \xi)]\} \leq \alpha,$$

where φ is a convex surrogate.

Chance constrained optimization

- We provide an answer to this problem in the following case. Let $g_j, j \in \{1, \dots, M\}$ be functions in $[-1, 1]$, $F(\lambda, \xi) = \sum_{j=1}^{N} \lambda_j g_j(\xi)$ and Λ is the simplex.
- Sample $(\xi_1, \dots, \xi_n)$ independent copies of ξ. We propose to solve

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad \sum_{i=1}^{n} \varphi(F(\lambda, \xi_i)) \leq n\alpha - \tau\sqrt{n}\,,$$

Denote by $\tilde{\lambda}$ any solution to this problem.
Denote by λ^* any solution to the original problem.

1. Does $\tilde{\lambda}$ satisfy the chance constraint?
2. $f(\tilde{\lambda}) - f(\lambda^*) \leq ?$

Corollary

Fix constants $\delta, \alpha \in (0,1), L > 0$ and let $\varphi : [-1,1] \to \mathbb{R}^+$ be a given L-Lipschitz convex surrogate. Define

$$\tau = 4\sqrt{2}L\sqrt{\log\left(\frac{2M}{\delta}\right)}\,.$$

Then, with probability at least $1 - 2\delta$

(i) $\mathbb{E}\{\varphi[F(\tilde{\lambda}, \xi)]\} \leq \alpha$.

(ii) If there exists $\varepsilon \in (0,1)$ such that the constraint $\mathbb{E}[\varphi(F(\lambda, \xi))] \leq \varepsilon\alpha$ is feasible for some $\lambda \in \Lambda$, then for $n \geq \left(\frac{4\tau}{(1-\varepsilon)\alpha}\right)^2$, we have

$$f(\tilde{\lambda}) - f(\lambda^*) \leq \frac{4\varphi(1)\tau}{(1-\varepsilon)\alpha\sqrt{n}}\,.$$

About the classification error

- Theorems are only about convexified classification errors. What about the real ones?
- We clearly have

$$R^-(\tilde{h}_n^\tau) \leq R_\varphi^-(\tilde{h}_n^\tau) \leq \alpha$$

  since $\varphi(\cdot) \geq \mathbb{1}(\cdot)$.
- Very likely, we have $R^-(\tilde{h}_n^\tau) < \alpha$. This is the price to pay to enforce the constraint strictly.
- Under a strict constraint, the following negative result holds.

Proposition

For any $\varepsilon > 0$, there exist base classifiers h_1, h_2 and a probability distribution for (X, Y) for which any pseudo-classifier $h_{\tilde{\lambda}} = \tilde{\lambda} h_1 + (1 - \tilde{\lambda}) h_2$, $0 \leq \tilde{\lambda} \leq 1$, such that $R^-(h_{\tilde{\lambda}}) = \alpha - \varepsilon$ and

$$R^+(h_{\tilde{\lambda}}) - \min_{\substack{R^-(h_\lambda) \leq \alpha \\ \lambda \in [0,1]}} R^+(h_\lambda) \geq \alpha \,.$$

About the classification error

- It is not hard to see that

$$\min_{R^-(h)\leq\alpha} R^+(h) = \inf_{R^-(h)\leq\alpha} R^+_\varphi(h)$$

- Therefore, analogous to Zhang's lemma for classification error we find (without extra conditions) that for any classifier $\tilde{h}$:

$$R^+(\tilde{h}) - \min_{R^-(h)\leq\alpha} R^+(h) \leq R^+_\varphi(\tilde{h}) - \inf_{R^-(h)\leq\alpha} R^+_\varphi(h)\,.$$

About the classification error

- It is not hard to see that

$$\min_{R^-(h)\leq\alpha} R^+(h) = \inf_{R^-(h)\leq\alpha} R^+_\varphi(h)$$

- Therefore, analogous to Zhang's lemma for classification error we find (without extra conditions) that for any classifier $\tilde{h}$:

$$R^+(\tilde{h}) - \min_{R^-(h)\leq\alpha} R^+(h) \leq R^+_\varphi(\tilde{h}) - \inf_{R^-(h)\leq\alpha} R^+_\varphi(h)\,.$$

- But the benchmark is too strong...

About the classification error

We can split the excess of type II error as follows

$$R^+(\tilde{h}^\kappa) - \min_{R^-(h)\leq\alpha} R^+(h) \leq T_1 + T_2\,,$$

where

$$T_1 = R^+_\varphi(\tilde{h}^\kappa) - \min_{\substack{h\in\mathcal{H}^{\mathrm{conv}}\\ R^-(h)\leq\alpha}} R^+_\varphi(h)\,,$$

$$T_2 = \min_{\substack{h\in\mathcal{H}^{\mathrm{conv}}\\ R^-(h)\leq\alpha}} R^+_\varphi(h) - \inf_{R^-(h)\leq\alpha} R^+_\varphi(h)\,.$$

- T_1 is the stochastic term of order $\sqrt{\dfrac{\log M}{n}}$.
- T_2 is the systematic error: the price to pay to use a restricted class of classifiers.