Regret, Games, and Boosting

a Reverie of Gambles and Bounds

Nicolò Cesa-Bianchi

Università degli Studi di Milano
Politecnico di Milano

Joint work with. . .

Marco Bressan
UNIMI

Nataly Brukhim
IAS Princeton

Emmanuel Esposito
UNIMI & IIT

Yishay Mansour
Tel Aviv & Google

Shay Moran
Technion & Google

Max Thiessen
TU Wien

Learning a binary classifier

- Finite sample $\mathcal{X}$ of datapoints with binary labels $f : \mathcal{X} \to \{-1, 1\}$

Learning a binary classifier

- Finite sample $\mathcal{X}$ of datapoints with binary labels $f : \mathcal{X} \to \{-1, 1\}$
- We know that *simple* explanations of $(\mathcal{X}, f)$ have good predictive power

Learning a binary classifier

- Finite sample $\mathcal{X}$ of datapoints with binary labels $f : \mathcal{X} \to \{-1, 1\}$
- We know that *simple* explanations of $(\mathcal{X}, f)$ have good predictive power
- Fix a simple (e.g., low VC-dimension) class of $\{-1, 1\}$-valued functions

Learning a binary classifier

- Finite sample $\mathcal{X}$ of datapoints with binary labels $f : \mathcal{X} \to \{-1, 1\}$
- We know that *simple* explanations of $(\mathcal{X}, f)$ have good predictive power
- Fix a simple (e.g., low VC-dimension) class of $\{-1, 1\}$-valued functions
- Simple explanations: (convex) combination of functions in the class whose sign correlates well with f on the sample $\mathcal{X}$

Weak Learning

- Let $\mathcal{H}$ be the projection on $\mathcal{X}$ of the functions in our VC class
- Each $h \in \mathcal{H}$ has the form $h : \mathcal{X} \to \{-1, 1\}$

Weak Learning

- Let $\mathcal{H}$ be the projection on $\mathcal{X}$ of the functions in our VC class
- Each $h \in \mathcal{H}$ has the form $h : \mathcal{X} \to \{-1, 1\}$
- WL (weak learning) assumption: There exists $\gamma > 0$ such that

$$\forall\, \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \quad \mathbb{P}_{X \sim \boldsymbol{q}}\big(h(X) \neq f(X)\big) \le \frac{1}{2} - \gamma$$

Weak Learning

- Let $\mathcal{H}$ be the projection on $\mathcal{X}$ of the functions in our VC class
- Each $h \in \mathcal{H}$ has the form $h : \mathcal{X} \to \{-1, 1\}$
- WL (weak learning) assumption: There exists $\gamma > 0$ such that

$$\forall \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \quad \mathbb{P}_{X \sim \boldsymbol{q}}\big(h(X) \neq f(X)\big) < \frac{1}{2}$$

Connections with minimax

- Let M be the $|\mathcal{H}| \times |\mathcal{X}|$ boolean matrix of elements $M(h,x) = \mathbb{I}\{h(x) \neq f(x)\}$

Connections with minimax

- Let M be the $|\mathcal{H}| \times |\mathcal{X}|$ boolean matrix of elements $M(h, x) = \mathbb{I}\{h(x) \neq f(x)\}$
- Rewrite WL assumption using M:

$$\forall\, \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \quad \mathbb{P}_{X \sim \boldsymbol{q}}\big(h(X) \neq f(X)\big) < \frac{1}{2} \qquad \Longleftrightarrow \qquad \max_{\boldsymbol{q} \in \Delta_{\mathcal{X}}} \min_{h \in \mathcal{H}} M(h, \boldsymbol{q}) < \frac{1}{2}$$

Connections with minimax

- Let M be the $|\mathcal{H}| \times |\mathcal{X}|$ boolean matrix of elements $M(h,x) = \mathbb{I}\{h(x) \neq f(x)\}$
- Rewrite WL assumption using M:

$$\forall\, \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \quad \mathbb{P}_{X \sim \boldsymbol{q}}\big(h(X) \neq f(X)\big) < \frac{1}{2} \qquad \Longleftrightarrow \qquad \max_{\boldsymbol{q} \in \Delta_{\mathcal{X}}} \min_{h \in \mathcal{H}} M(h, \boldsymbol{q}) < \frac{1}{2}$$

- By the minimax theorem, we know that

$$\max_{\boldsymbol{q} \in \Delta_{\mathcal{X}}} \min_{h \in \mathcal{H}} M(h, \boldsymbol{q}) = \min_{\boldsymbol{p} \in \Delta_{\mathcal{H}}} \max_{x \in \mathcal{X}} M(\boldsymbol{p}, x)$$

Connections with minimax

- Let M be the $|\mathcal{H}| \times |\mathcal{X}|$ boolean matrix of elements $M(h,x) = \mathbb{I}\{h(x) \neq f(x)\}$
- Rewrite WL assumption using M:

$$\forall\, \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \quad \mathbb{P}_{X \sim \boldsymbol{q}}\big(h(X) \neq f(X)\big) < \frac{1}{2} \qquad \Longleftrightarrow \qquad \max_{\boldsymbol{q} \in \Delta_{\mathcal{X}}} \min_{h \in \mathcal{H}} M(h, \boldsymbol{q}) < \frac{1}{2}$$

- By the minimax theorem, we know that

$$\max_{\boldsymbol{q} \in \Delta_{\mathcal{X}}} \min_{h \in \mathcal{H}} M(h, \boldsymbol{q}) = \min_{\boldsymbol{p} \in \Delta_{\mathcal{H}}} \max_{x \in \mathcal{X}} M(\boldsymbol{p}, x)$$

- Therefore there exists a distribution p^* over $\mathcal{H}$ such that

$$\max_{x \in \mathcal{X}} M(\boldsymbol{p}^*, x) < \frac{1}{2} \qquad \Longleftrightarrow \qquad \mathbb{P}_{H \sim \boldsymbol{p}^*}\big(H(x) \neq f(x)\big) < \frac{1}{2} \quad \text{for all } x \in \mathcal{X}$$

Connections with minimax

- Let M be the $|\mathcal{H}| \times |\mathcal{X}|$ boolean matrix of elements $M(h,x) = \mathbb{I}\{h(x) \neq f(x)\}$
- Rewrite WL assumption using M:

$$\forall\, \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \quad \mathbb{P}_{X \sim \boldsymbol{q}}\big(h(X) \neq f(X)\big) < \frac{1}{2} \qquad \Longleftrightarrow \qquad \max_{\boldsymbol{q} \in \Delta_{\mathcal{X}}} \min_{h \in \mathcal{H}} M(h, \boldsymbol{q}) < \frac{1}{2}$$

- By the minimax theorem, we know that

$$\max_{\boldsymbol{q} \in \Delta_{\mathcal{X}}} \min_{h \in \mathcal{H}} M(h, \boldsymbol{q}) = \min_{\boldsymbol{p} \in \Delta_{\mathcal{H}}} \max_{x \in \mathcal{X}} M(\boldsymbol{p}, x)$$

- Therefore there exists a distribution p^* over $\mathcal{H}$ such that

$$\max_{x \in \mathcal{X}} M(\boldsymbol{p}^*, x) < \frac{1}{2} \qquad \Longleftrightarrow \qquad \mathbb{P}_{H \sim \boldsymbol{p}^*}\big(H(x) \neq f(x)\big) < \frac{1}{2} \quad \text{for all } x \in \mathcal{X}$$

- This $\boldsymbol{p}^*$ *explains* $(f, \mathcal{X})$: for all $x \in \mathcal{X}$ $\quad \mathrm{sgn}\big(\mathbb{E}_{\boldsymbol{p}^*}[H(x)]\big) = f(x)$ $\quad$ where $\mathrm{sgn}(0) = \perp$

Weak learning and simple explanations

- WL assumption is equivalent to $\mathrm{sgn}(\mathbb{E}_{\boldsymbol{p}^*}[H]) = f$, existence of a simple explanation for $(\mathcal{X}, f)$

Weak learning and simple explanations

- WL assumption is equivalent to $\mathrm{sgn}(\mathbb{E}_{\boldsymbol{p}^*}[H]) = f$, existence of a simple explanation for $(\mathcal{X}, f)$
- We can find $\boldsymbol{p}^*$ using LP, but boosting is a very simple and efficient alternative

Weak learning and simple explanations

- WL assumption is equivalent to $\mathrm{sgn}(\mathbb{E}_{\boldsymbol{p}^*}[H]) = f$, existence of a simple explanation for $(\mathcal{X}, f)$
- We can find $\boldsymbol{p}^*$ using LP, but boosting is a very simple and efficient alternative

- WL oracle for $z \in [0,1]$: Given $\boldsymbol{q} \in \Delta_{\mathcal{X}}$ the oracle returns $h \in \mathcal{H}$ such that

$$\mathbb{P}_{X \sim \boldsymbol{q}}\big(h(X) \neq f(X)\big) \leq z$$

Weak learning and simple explanations

- WL assumption is equivalent to $\operatorname{sgn}(\mathbb{E}_{\boldsymbol{p}^*}[H]) = f$, existence of a simple explanation for $(\mathcal{X}, f)$
- We can find $\boldsymbol{p}^*$ using LP, but boosting is a very simple and efficient alternative
- WL oracle for $z \in [0,1]$: Given $\boldsymbol{q} \in \Delta_{\mathcal{X}}$ the oracle returns $h \in \mathcal{H}$ such that

$$\mathbb{P}_{X \sim \boldsymbol{q}}\big(h(X) \neq f(X)\big) \leq z$$

Definition

$z \in [0,1]$ is boostable if we can find $\boldsymbol{p}^*$ such that $\operatorname{sgn}(\mathbb{E}_{\boldsymbol{p}^*}[H]) = f$ using a WL oracle for z

Weak learning and simple explanations

- WL assumption is equivalent to $\mathrm{sgn}(\mathbb{E}_{\boldsymbol{p}^*}[H]) = f$, existence of a simple explanation for $(\mathcal{X}, f)$
- We can find $\boldsymbol{p}^*$ using LP, but boosting is a very simple and efficient alternative

- WL oracle for $z \in [0, 1]$: Given $\boldsymbol{q} \in \Delta_{\mathcal{X}}$ the oracle returns $h \in \mathcal{H}$ such that

$$\mathbb{P}_{X \sim \boldsymbol{q}}\big(h(X) \neq f(X)\big) \leq z$$

Definition

$z \in [0, 1]$ is boostable if we can find $\boldsymbol{p}^*$ such that $\mathrm{sgn}(\mathbb{E}_{\boldsymbol{p}^*}[H]) = f$ using a WL oracle for z

We now show how to boost a WL oracle using online learning

Online learning

Recall $M(h, x) = \mathbb{I}\{h(x) \neq f(x)\}$

The online learning protocol

For each round $t \geq 1$:

1. The learner chooses $\boldsymbol{p}_t \in \Delta_{\mathcal{H}}$
2. The adversary reveals $x_t \in \mathcal{X}$
3. The learner suffers loss $M(\boldsymbol{p}_t, x_t) = \mathbb{P}_{H \sim \boldsymbol{p}_t}\big(H(x_t) \neq f(x_t)\big)$

Online learning

Recall $M(h,x) = \mathbb{I}\{h(x) \neq f(x)\}$

The online learning protocol

For each round $t \geq 1$:

1. The learner chooses $\boldsymbol{p}_t \in \Delta_{\mathcal{H}}$
2. The adversary reveals $x_t \in \mathcal{X}$
3. The learner suffers loss $M(\boldsymbol{p}_t, x_t) = \mathbb{P}_{H \sim \boldsymbol{p}_t}\big(H(x_t) \neq f(x_t)\big)$

For all T and for all $x_1, \dots, x_T \in \mathcal{X}$, if the learner runs the Hedge algorithm, then

$$\frac{1}{T}\sum_{t=1}^{T} M(\boldsymbol{p}_t, x_t) \leq \min_{h \in \mathcal{H}} \frac{1}{T}\sum_{t=1}^{T} M(h, x_t) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$

Boosting via online learning

We run Hedge over the dual game $M' = \mathbf{1}\mathbf{1}^\top - M^\top$ against a WL oracle for z as adversary

Boosting algorithm

For each round $t \geq 1$:

1. Hedge chooses $\boldsymbol{p}_t \in \Delta_{\mathcal{X}}$ (Hedge learns distributions over $\mathcal{X}$)
2. The WL oracle returns $h_t \in \mathcal{H}$ satisfying $M(h_t, \boldsymbol{p}_t) \leq z$
3. Hedge gets loss $M'(\boldsymbol{p}_t, h_t) = 1 - M(h_t, \boldsymbol{p}_t)$

Analysis

- The WL oracle for z satisfies $\quad M'(\boldsymbol{p}_t, h_t) = 1 - M(h_t, \boldsymbol{p}_t) \geq 1 - z \qquad t = 1, 2, \ldots$

Analysis

- The WL oracle for z satisfies $\quad M'(\boldsymbol{p}_t, h_t) = 1 - M(h_t, \boldsymbol{p}_t) \geq 1 - z \qquad t = 1, 2, \dots$
- Therefore, by the properties of Hedge

$$1 - z \leq \frac{1}{T} \sum_{t=1}^{T} M'(\boldsymbol{p}_t, h_t) \leq \min_{x \in \mathcal{X}} \frac{1}{T} \sum_{t=1}^{T} M'(x, h_t) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$

Analysis

- The WL oracle for z satisfies $\quad M'(\boldsymbol{p}_t, h_t) = 1 - M(h_t, \boldsymbol{p}_t) \ge 1 - z \qquad t = 1, 2, \dots$
- Therefore, by the properties of Hedge

$$1 - z \le \frac{1}{T} \sum_{t=1}^{T} M'(\boldsymbol{p}_t, h_t) \le \min_{x \in \mathcal{X}} \frac{1}{T} \sum_{t=1}^{T} M'(x, h_t) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$

- Hence, if $z < \frac{1}{2}$ and T is large enough,

$$\min_{x \in \mathcal{X}} \frac{1}{T} \sum_{t=1}^{T} M'(x, h_t) > \frac{1}{2}$$

Analysis

- The WL oracle for z satisfies $\quad M'(\boldsymbol{p}_t, h_t) = 1 - M(h_t, \boldsymbol{p}_t) \geq 1 - z \qquad t = 1, 2, \dots$
- Therefore, by the properties of Hedge

$$1 - z \leq \frac{1}{T} \sum_{t=1}^{T} M'(\boldsymbol{p}_t, h_t) \leq \min_{x \in \mathcal{X}} \frac{1}{T} \sum_{t=1}^{T} M'(x, h_t) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$

- Hence, if $z < \frac{1}{2}$ and T is large enough,

$$\min_{x \in \mathcal{X}} \frac{1}{T} \sum_{t=1}^{T} M'(x, h_t) > \frac{1}{2}$$

- So, $h_t(x) = f(x)$ for more than half of the h_t on each $x \in \mathcal{X}$, which implies

$$\operatorname{sgn}\left(\frac{1}{T} \sum_{t=1}^{T} h_t\right) = f$$

A game-theoretic view

- Game matrix $B = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ for binary classification

A game-theoretic view

- Game matrix $B = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ for binary classification
- $\displaystyle \frac{1}{2} = \min_{\boldsymbol{u} \in \Delta_{\mathcal{Y}}} \max_{\boldsymbol{v} \in \Delta_{\mathcal{Y}}} B(\boldsymbol{u}, \boldsymbol{v}) = V(B)$ is the value of game B $\qquad$ ($\mathcal{Y} = \{-1, 1\}$)

A game-theoretic view

- Game matrix $B = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ for binary classification
- $\dfrac{1}{2} = \min_{\boldsymbol{u}\in\Delta_{\mathcal{Y}}} \max_{\boldsymbol{v}\in\Delta_{\mathcal{Y}}} B(\boldsymbol{u},\boldsymbol{v}) = V(B)$ is the value of game B $\qquad (\mathcal{Y} = \{-1,1\})$

Theorem (Boosting Theorem)

Any $z < V(B)$ is boostable

A game-theoretic view

- Game matrix $B = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ for binary classification
- $\displaystyle \frac{1}{2} = \min_{\boldsymbol{u}\in\Delta_{\mathcal{Y}}} \max_{\boldsymbol{v}\in\Delta_{\mathcal{Y}}} B(\boldsymbol{u},\boldsymbol{v}) = V(B)$ is the value of game B $\qquad (\mathcal{Y} = \{-1,1\})$

Theorem (Boosting Theorem)

Any $z < V(B)$ is boostable

$$V(B) = \max_{\boldsymbol{v}\in\Delta_{\mathcal{Y}}} \min_{\boldsymbol{u}\in\Delta_{\mathcal{Y}}} B(\boldsymbol{u},\boldsymbol{v})$$

is the smallest error attainable with a biased coin $\boldsymbol{u}$ against the worst-possible distribution $\boldsymbol{v}$

A game-theoretic view

- Game matrix $B = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ for binary classification
- $\displaystyle \frac{1}{2} = \min_{\boldsymbol{u} \in \Delta_{\mathcal{Y}}} \max_{\boldsymbol{v} \in \Delta_{\mathcal{Y}}} B(\boldsymbol{u}, \boldsymbol{v}) = V(B)$ is the value of game B $\quad (\mathcal{Y} = \{-1, 1\})$

Theorem (Boosting Theorem)

Any $z < V(B)$ is boostable

$$V(B) = \max_{\boldsymbol{v} \in \Delta_{\mathcal{Y}}} \min_{\boldsymbol{u} \in \Delta_{\mathcal{Y}}} B(\boldsymbol{u}, \boldsymbol{v})$$

is the smallest error attainable with a biased coin $\boldsymbol{u}$ against the worst-possible distribution $\boldsymbol{v}$

Any $z \in [0, 1]$ is either boostable or coin attainable

Cost-sensitive binary classification

- Game matrix $C = \begin{pmatrix} 0 & c^+ \\ c^- & 0 \end{pmatrix}$ for cost-sensitive classification
  where $0 < c^+, c^- \leq 1$ are costs for false positive and false negative mistakes

Cost-sensitive binary classification

- Game matrix $C = \left(\begin{smallmatrix} 0 & c^+ \\ c^- & 0 \end{smallmatrix}\right)$ for cost-sensitive classification where $0 < c^+, c^- \leq 1$ are costs for false positive and false negative mistakes
- $\displaystyle \min_{\boldsymbol{u} \in \Delta_{\mathcal{Y}}} \max_{\boldsymbol{v} \in \Delta_{\mathcal{Y}}} C(\boldsymbol{u}, \boldsymbol{v}) = \frac{c^- c^+}{c^- + c^+} = V(C)$ is the value of game C

Cost-sensitive binary classification

- Game matrix $C = \begin{pmatrix} 0 & c^+ \\ c^- & 0 \end{pmatrix}$ for cost-sensitive classification where $0 < c^+, c^- \le 1$ are costs for false positive and false negative mistakes
- $\displaystyle \min_{\boldsymbol{u}\in\Delta_{\mathcal{Y}}} \max_{\boldsymbol{v}\in\Delta_{\mathcal{Y}}} C(\boldsymbol{u},\boldsymbol{v}) = \frac{c^- c^+}{c^- + c^+} = V(C)$ is the value of game C
- $(\mathcal{H} \times \mathcal{X})$-matrix:

$$M_C(h,x) = c^+ \underbrace{\mathbb{I}\{h(x) = 1 \wedge f(x) = -1\}}_{\text{false positive}} + c^- \underbrace{\mathbb{I}\{h(x) = -1 \wedge f(x) = 1\}}_{\text{false negative}}$$

Cost-sensitive binary classification

- Game matrix $C = \begin{pmatrix} 0 & c^+ \\ c^- & 0 \end{pmatrix}$ for cost-sensitive classification
  where $0 < c^+, c^- \leq 1$ are costs for false positive and false negative mistakes
- $\min_{\boldsymbol{u}\in\Delta_\mathcal{Y}} \max_{\boldsymbol{v}\in\Delta_\mathcal{Y}} C(\boldsymbol{u},\boldsymbol{v}) = \dfrac{c^- c^+}{c^- + c^+} = V(C)$ is the value of game C
- $(\mathcal{H} \times \mathcal{X})$-matrix:

$$M_C(h,x) = c^+ \underbrace{\mathbb{I}\{h(x) = 1 \wedge f(x) = -1\}}_{\text{false positive}} + c^- \underbrace{\mathbb{I}\{h(x) = -1 \wedge f(x) = 1\}}_{\text{false negative}}$$

- WL oracle for z: $\max_{\boldsymbol{q}\in\Delta_\mathcal{X}} \min_{h\in\mathcal{H}} M_C(h,\boldsymbol{q}) \leq z$

Cost-sensitive binary classification

- Game matrix $C = \begin{pmatrix} 0 & c^+ \\ c^- & 0 \end{pmatrix}$ for cost-sensitive classification where $0 < c^+, c^- \leq 1$ are costs for false positive and false negative mistakes
- $\displaystyle\min_{\boldsymbol{u}\in\Delta_{\mathcal{Y}}} \max_{\boldsymbol{v}\in\Delta_{\mathcal{Y}}} C(\boldsymbol{u},\boldsymbol{v}) = \frac{c^- c^+}{c^- + c^+} = V(C)$ is the value of game C
- $(\mathcal{H} \times \mathcal{X})$-matrix:

$$M_C(h,x) = c^+ \underbrace{\mathbb{I}\{h(x) = 1 \wedge f(x) = -1\}}_{\text{false positive}} + c^- \underbrace{\mathbb{I}\{h(x) = -1 \wedge f(x) = 1\}}_{\text{false negative}}$$

- WL oracle for z: $\displaystyle\max_{\boldsymbol{q}\in\Delta_{\mathcal{X}}} \min_{h\in\mathcal{H}} M_C(h,\boldsymbol{q}) \leq z$
- Which values of z are boostable?

Cost-sensitive binary classification

- Game matrix $C = \begin{pmatrix} 0 & c^+ \\ c^- & 0 \end{pmatrix}$ for cost-sensitive classification where $0 < c^+, c^- \le 1$ are costs for false positive and false negative mistakes
- $\displaystyle \min_{\boldsymbol{u}\in\Delta_{\mathcal{Y}}} \max_{\boldsymbol{v}\in\Delta_{\mathcal{Y}}} C(\boldsymbol{u},\boldsymbol{v}) = \frac{c^- c^+}{c^- + c^+} = V(C)$ is the value of game C
- $(\mathcal{H} \times \mathcal{X})$-matrix:

$$M_C(h,x) = c^+ \underbrace{\mathbb{I}\{h(x) = 1 \wedge f(x) = -1\}}_{\text{false positive}} + c^- \underbrace{\mathbb{I}\{h(x) = -1 \wedge f(x) = 1\}}_{\text{false negative}}$$

- WL oracle for z: $\displaystyle \max_{\boldsymbol{q}\in\Delta_{\mathcal{X}}} \min_{h\in\mathcal{H}} M_C(h,\boldsymbol{q}) \le z$
- Which values of z are boostable?

Theorem (Cost-sensitive Boosting Theorem)

Any $z < V(C)$ is boostable *(boostable vs. coin attainable dichotomy)*

Relationship to Bayes optimal prediction

- Bayes optimal prediction minimizes the conditional risk:

$$f^*(x) = \operatorname*{argmin}_{\widehat{y} \in \mathcal{Y}} \mathbb{E}\big[\ell(\widehat{y}, Y) \mid X = x\big]$$

Relationship to Bayes optimal prediction

- Bayes optimal prediction minimizes the conditional risk:

$$f^*(x) = \operatorname*{argmin}_{\widehat{y} \in \mathcal{Y}} \mathbb{E}\big[\ell(\widehat{y}, Y) \mid X = x\big]$$

- Worst-case conditional Bayes risk:

$$\max_{p(Y|X=x)} \min_{\widehat{y} \in \mathcal{Y}} \mathbb{E}\Big[\ell(\widehat{y}, Y) \mid X = x\Big]$$

Relationship to Bayes optimal prediction

- Bayes optimal prediction minimizes the conditional risk:

$$f^*(x) = \operatorname*{argmin}_{\widehat{y}\in\mathcal{Y}} \mathbb{E}\big[\ell(\widehat{y},Y) \mid X = x\big]$$

- Worst-case conditional Bayes risk:

$$\max_{p(Y|X=x)} \min_{\widehat{y}\in\mathcal{Y}} \mathbb{E}\Big[\ell(\widehat{y},Y) \mid X = x\Big]$$

- Cost-sensitive binary classification

$$\max_{p(Y|X=x)} \min_{\widehat{y}\in\mathcal{Y}} \mathbb{E}\Big[C(\widehat{y},Y) \mid X = x\Big] = \min_{\boldsymbol{u}\in\Delta_{\mathcal{Y}}} \max_{\boldsymbol{v}\in\Delta_{\mathcal{Y}}} C(\boldsymbol{u},\boldsymbol{v}) = V(C)$$

Relationship to Bayes optimal prediction

- Bayes optimal prediction minimizes the conditional risk:

$$f^*(x) = \operatorname*{argmin}_{\widehat{y} \in \mathcal{Y}} \mathbb{E}\big[\ell(\widehat{y}, Y) \mid X = x\big]$$

- Worst-case conditional Bayes risk:

$$\max_{p(Y|X=x)} \min_{\widehat{y} \in \mathcal{Y}} \mathbb{E}\Big[\ell(\widehat{y}, Y) \mid X = x\Big]$$

- Cost-sensitive binary classification

$$\max_{p(Y|X=x)} \min_{\widehat{y} \in \mathcal{Y}} \mathbb{E}\Big[C(\widehat{y}, Y) \mid X = x\Big] = \min_{\boldsymbol{u} \in \Delta_{\mathcal{Y}}} \max_{\boldsymbol{v} \in \Delta_{\mathcal{Y}}} C(\boldsymbol{u}, \boldsymbol{v}) = V(C)$$

- The worst-case conditional Bayes risk for a binary prediction game defines the threshold between boostability and coin attainability

Simultaneous false positive and false negative guarantees

- The FP game $W^+ = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ and the FN game $W^- = \left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right)$

Simultaneous false positive and false negative guarantees

- The FP game $W^+ = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ and the FN game $W^- = \left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right)$
- The associated $(\mathcal{H} \times \mathcal{X})$-matrices

$$M^+(h,x) = \mathbb{I}\{h(x) = 1 \wedge f(x) = -1\} \qquad M^-(h,x) = \mathbb{I}\{h(x) = -1 \wedge f(x) = 1\}$$

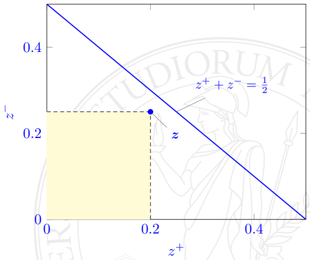
Simultaneous false positive and false negative guarantees

- The FP game $W^+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and the FN game $W^- = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$
- The associated $(\mathcal{H} \times \mathcal{X})$-matrices

$$M^+(h,x) = \mathbb{I}\{h(x) = 1 \wedge f(x) = -1\} \qquad M^-(h,x) = \mathbb{I}\{h(x) = -1 \wedge f(x) = 1\}$$

- WL oracle for $\boldsymbol{z} = (z^+, z^-)$
(simultaneous guarantees for FP and FN mistakes):

$$\forall\, \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \qquad M^+(h,\boldsymbol{q}) < z^+ \wedge M^-(h,\boldsymbol{q}) < z^-$$



Simultaneous false positive and false negative guarantees

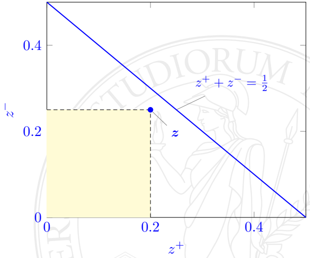
- The FP game $W^+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and the FN game $W^- = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$
- The associated $(\mathcal{H} \times \mathcal{X})$-matrices

$$M^+(h,x) = \mathbb{I}\{h(x) = 1 \wedge f(x) = -1\} \qquad M^-(h,x) = \mathbb{I}\{h(x) = -1 \wedge f(x) = 1\}$$

- WL oracle for $\boldsymbol{z} = (z^+, z^-)$
(simultaneous guarantees for FP and FN mistakes):

$$\forall\, \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \qquad M^+(h,\boldsymbol{q}) < z^+ \wedge M^-(h,\boldsymbol{q}) < z^-$$

- Which values of $\boldsymbol{z} = (z^+, z^-)$ are boostable?



Coin attainable guarantees

- Recall WL oracle for $\boldsymbol{z} = (z^+, z^-)$:

$$\forall\, \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \qquad M^+(h, \boldsymbol{q}) < z^+ \;\wedge\; M^-(h, \boldsymbol{q}) < z^-$$

Coin attainable guarantees

- Recall WL oracle for $\boldsymbol{z} = (z^+, z^-)$:

$$\forall\, \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \qquad M^+(h, \boldsymbol{q}) < z^+ \;\wedge\; M^-(h, \boldsymbol{q}) < z^-$$

- This oracle returns $h : \mathcal{X} \to \mathcal{Y}$ attaining the desired guarantees against a given distribution over $\mathcal{X}$

Coin attainable guarantees

- Recall WL oracle for $\boldsymbol{z} = (z^+, z^-)$:
$$\forall\, \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \qquad M^+(h, \boldsymbol{q}) < z^+ \;\wedge\; M^-(h, \boldsymbol{q}) < z^-$$
- This oracle returns $h : \mathcal{X} \to \mathcal{Y}$ attaining the desired guarantees against a given distribution over $\mathcal{X}$
- Coin-attainable region: guarantees that are attainable with a biased coin
$$\mathcal{K} \equiv \left\{ (z^+, z^-) \,:\, \forall \boldsymbol{v} \in \Delta_{\mathcal{Y}} \quad \exists \boldsymbol{u} \in \Delta_{\mathcal{Y}} \quad W^+(\boldsymbol{u}, \boldsymbol{v}) \le z^+ \;\wedge\; W^-(\boldsymbol{u}, \boldsymbol{v}) \le z^- \right\}$$

Coin attainable guarantees

- Recall WL oracle for $\boldsymbol{z} = (z^+, z^-)$:

$$\forall\, \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \qquad M^+(h, \boldsymbol{q}) < z^+ \;\wedge\; M^-(h, \boldsymbol{q}) < z^-$$

- This oracle returns $h : \mathcal{X} \to \mathcal{Y}$ attaining the desired guarantees against a given distribution over $\mathcal{X}$
- Coin-attainable region: guarantees that are attainable with a biased coin

$$\mathcal{K} \equiv \Big\{ (z^+, z^-) \,:\, \forall \boldsymbol{v} \in \Delta_{\mathcal{Y}} \quad \exists \boldsymbol{u} \in \Delta_{\mathcal{Y}} \quad W^+(\boldsymbol{u}, \boldsymbol{v}) \le z^+ \;\wedge\; W^-(\boldsymbol{u}, \boldsymbol{v}) \le z^- \Big\}$$

- Can an oracle providing guarantees that are attainable with a coin be used for boosting?

Coin attainable guarantees

- Recall WL oracle for $\boldsymbol{z} = (z^+, z^-)$:

$$\forall\, \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \qquad M^+(h, \boldsymbol{q}) < z^+ \;\wedge\; M^-(h, \boldsymbol{q}) < z^-$$

- This oracle returns $h : \mathcal{X} \to \mathcal{Y}$ attaining the desired guarantees against a given distribution over $\mathcal{X}$
- Coin-attainable region: guarantees that are attainable with a biased coin

$$\mathcal{K} \equiv \left\{ (z^+, z^-) \,:\, \forall \boldsymbol{v} \in \Delta_{\mathcal{Y}} \quad \exists \boldsymbol{u} \in \Delta_{\mathcal{Y}} \quad W^+(\boldsymbol{u}, \boldsymbol{v}) \le z^+ \;\wedge\; W^-(\boldsymbol{u}, \boldsymbol{v}) \le z^- \right\}$$

- Can an oracle providing guarantees that are attainable with a coin be used for boosting?
- Scalar case: any $z \ge V(C)$ is coin attainable by definition and therefore not boostable by the cost-sensitive Boosting theorem

Coin attainable guarantees

- Recall WL oracle for $\boldsymbol{z} = (z^+, z^-)$:
$$\forall\, \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \qquad M^+(h, \boldsymbol{q}) < z^+ \;\wedge\; M^-(h, \boldsymbol{q}) < z^-$$
- This oracle returns $h : \mathcal{X} \to \mathcal{Y}$ attaining the desired guarantees against a given distribution over $\mathcal{X}$
- Coin-attainable region: guarantees that are attainable with a biased coin
$$\mathcal{K} \equiv \left\{ (z^+, z^-) \,:\, \forall \boldsymbol{v} \in \Delta_{\mathcal{Y}} \quad \exists \boldsymbol{u} \in \Delta_{\mathcal{Y}} \quad W^+(\boldsymbol{u}, \boldsymbol{v}) \le z^+ \;\wedge\; W^-(\boldsymbol{u}, \boldsymbol{v}) \le z^- \right\}$$
- Can an oracle providing guarantees that are attainable with a coin be used for boosting?
- Scalar case: any $z \ge V(C)$ is coin attainable by definition and therefore not boostable by the cost-sensitive Boosting theorem

Theorem (Multi-Objective Boosting Theorem)

Any $\boldsymbol{z} \notin \mathcal{K}$ is boostable

Coin attainable guarantees

- Recall WL oracle for $\boldsymbol{z} = (z^+, z^-)$:

$$\forall\, \boldsymbol{q} \in \Delta_{\mathcal{X}} \quad \exists\, h \in \mathcal{H} \qquad M^+(h, \boldsymbol{q}) < z^+ \;\wedge\; M^-(h, \boldsymbol{q}) < z^-$$

- This oracle returns $h : \mathcal{X} \to \mathcal{Y}$ attaining the desired guarantees against a given distribution over $\mathcal{X}$
- Coin-attainable region: guarantees that are attainable with a biased coin

$$\mathcal{K} \equiv \left\{ (z^+, z^-) \,:\, \forall \boldsymbol{v} \in \Delta_{\mathcal{Y}} \quad \exists \boldsymbol{u} \in \Delta_{\mathcal{Y}} \quad W^+(\boldsymbol{u}, \boldsymbol{v}) \le z^+ \;\wedge\; W^-(\boldsymbol{u}, \boldsymbol{v}) \le z^- \right\}$$

- Can an oracle providing guarantees that are attainable with a coin be used for boosting?
- Scalar case: any $z \ge V(C)$ is coin attainable by definition and therefore not boostable by the cost-sensitive Boosting theorem

Theorem (Multi-Objective Boosting Theorem)

Any $\boldsymbol{z} \notin \mathcal{K}$ is boostable

What is the geometry of $\mathcal{K}$?

Scalarization

For any $0 \le \alpha \le 1$

$$W_\alpha = \alpha W^+ + (1-\alpha)W^- = \left(\begin{smallmatrix} 0 & \alpha \\ 1-\alpha & 0 \end{smallmatrix}\right)$$

$$M_\alpha = \alpha \underbrace{\mathbb{I}\{h(x) = 1 \wedge f(x) = -1\}}_{\text{FP}} + (1-\alpha) \underbrace{\mathbb{I}\{h(x) = -1 \wedge f(x) = 1\}}_{\text{FN}}$$

Scalarization

For any $0 \le \alpha \le 1$

$$W_\alpha = \alpha W^+ + (1-\alpha)W^- = \left(\begin{smallmatrix} 0 & \alpha \\ 1-\alpha & 0 \end{smallmatrix}\right)$$

$$M_\alpha = \alpha \, \underbrace{\mathbb{I}\{h(x) = 1 \wedge f(x) = -1\}}_{\text{FP}} + (1-\alpha) \, \underbrace{\mathbb{I}\{h(x) = -1 \wedge f(x) = 1\}}_{\text{FN}}$$

- This is equivalent to convex costs $\alpha, 1-\alpha$ for false positive and false negative mistakes

Scalarization

For any $0 \le \alpha \le 1$

$$W_\alpha = \alpha W^+ + (1-\alpha)W^- = \begin{pmatrix} 0 & \alpha \\ 1-\alpha & 0 \end{pmatrix}$$

$$M_\alpha = \alpha \underbrace{\mathbb{I}\{h(x) = 1 \wedge f(x) = -1\}}_{\text{FP}} + (1-\alpha) \underbrace{\mathbb{I}\{h(x) = -1 \wedge f(x) = 1\}}_{\text{FN}}$$

- This is equivalent to convex costs $\alpha, 1-\alpha$ for false positive and false negative mistakes
- In the scalar case $W_\alpha = \begin{pmatrix} 0 & \alpha \\ 1-\alpha & 0 \end{pmatrix}$, any $z \ge V(W_\alpha)$ is coin attainable

Scalarization

For any $0 \le \alpha \le 1$

$$W_\alpha = \alpha W^+ + (1-\alpha) W^- = \left(\begin{smallmatrix} 0 & \alpha \\ 1-\alpha & 0 \end{smallmatrix}\right)$$

$$M_\alpha = \alpha \underbrace{\mathbb{I}\{h(x) = 1 \wedge f(x) = -1\}}_{\text{FP}} + (1-\alpha) \underbrace{\mathbb{I}\{h(x) = -1 \wedge f(x) = 1\}}_{\text{FN}}$$

- This is equivalent to convex costs $\alpha, 1-\alpha$ for false positive and false negative mistakes
- In the scalar case $W_\alpha = \left(\begin{smallmatrix} 0 & \alpha \\ 1-\alpha & 0 \end{smallmatrix}\right)$, any $z \ge V(W_\alpha)$ is coin attainable

Theorem (Duality)

A multi-objective guarantee $\boldsymbol{z}$ is coin-attainable iff it has no boostable scalarizations:

$$\mathcal{K} \equiv \left\{ \boldsymbol{z} = (z^+, z^-) \,:\, \forall \boldsymbol{\alpha} = (\alpha, 1-\alpha) \quad \langle \boldsymbol{\alpha}, \boldsymbol{z} \rangle \ge V(W_{\boldsymbol{\alpha}}) \right\}$$

Scalarization

For any $0 \le \alpha \le 1$

$$W_\alpha = \alpha W^+ + (1-\alpha) W^- = \left(\begin{smallmatrix} 0 & \alpha \\ 1-\alpha & 0 \end{smallmatrix}\right)$$

$$M_\alpha = \alpha \underbrace{\mathbb{I}\{h(x) = 1 \wedge f(x) = -1\}}_{\mathrm{FP}} + (1-\alpha) \underbrace{\mathbb{I}\{h(x) = -1 \wedge f(x) = 1\}}_{\mathrm{FN}}$$

- This is equivalent to convex costs $\alpha, 1-\alpha$ for false positive and false negative mistakes
- In the scalar case $W_\alpha = \left(\begin{smallmatrix} 0 & \alpha \\ 1-\alpha & 0 \end{smallmatrix}\right)$, any $z \ge V(W_\alpha)$ is coin attainable

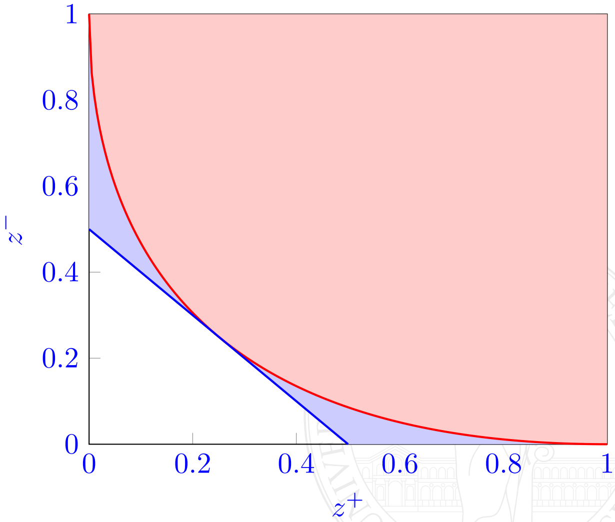
Theorem (Duality)

A multi-objective guarantee $\boldsymbol{z}$ is coin-attainable iff it has no boostable scalarizations:

$$\begin{aligned} \mathcal{K} &\equiv \left\{ \boldsymbol{z} = (z^+, z^-) \,:\, \forall \boldsymbol{\alpha} = (\alpha, 1-\alpha) \quad \langle \boldsymbol{\alpha}, \boldsymbol{z} \rangle \ge V(W_{\boldsymbol{\alpha}}) \right\} \\ &\equiv \left\{ (z^+, z^-) \in [0,1]^2 \,:\, \sqrt{z^+} + \sqrt{z^-} \ge 1 \right\} \end{aligned}$$

Geometry of $\mathcal{K}$

- Red: Coin attainable region
- Blue: Boostable only using oracle with simultaneous guarantees
- White: Boostable by both oracles



Extensions to the multiclass case

- Game matrix D for labels $\{1, \ldots, n\}$ where $D(i,j)$ is cost of predicting i when true label is j

Extensions to the multiclass case

- Game matrix D for labels $\{1,\dots,n\}$ where $D(i,j)$ is cost of predicting i when true label is j
- For any subset $J \subseteq \mathcal{Y}$ of labels: $V_J(D) = \min_{\boldsymbol{u}\in\Delta_{\mathcal{Y}}} \max_{\boldsymbol{v}\in\Delta_J} D(\boldsymbol{u},\boldsymbol{v})$

Extensions to the multiclass case

- Game matrix D for labels $\{1, \ldots, n\}$ where $D(i, j)$ is cost of predicting i when true label is j
- For any subset $J \subseteq \mathcal{Y}$ of labels: $V_J(D) = \min\limits_{\boldsymbol{u} \in \Delta_{\mathcal{Y}}} \max\limits_{\boldsymbol{v} \in \Delta_J} D(\boldsymbol{u}, \boldsymbol{v})$
- This is the smallest loss one can ensure by predicting with a die knowing that the correct label is in J.

Extensions to the multiclass case

- Game matrix D for labels $\{1,\dots,n\}$ where $D(i,j)$ is cost of predicting i when true label is j
- For any subset $J \subseteq \mathcal{Y}$ of labels: $V_J(D) = \min_{\boldsymbol{u}\in\Delta_{\mathcal{Y}}} \max_{\boldsymbol{v}\in\Delta_J} D(\boldsymbol{u},\boldsymbol{v})$
- This is the smallest loss one can ensure by predicting with a die knowing that the correct label is in J.
- Boosting is only possible within any two consecutive game values

$$0 = v_1 < v_2 < \cdots < \boxed{v_j < v_{j+1}} < \cdots < v_N = V_{\mathcal{Y}}(D)$$

Extensions to the multiclass case

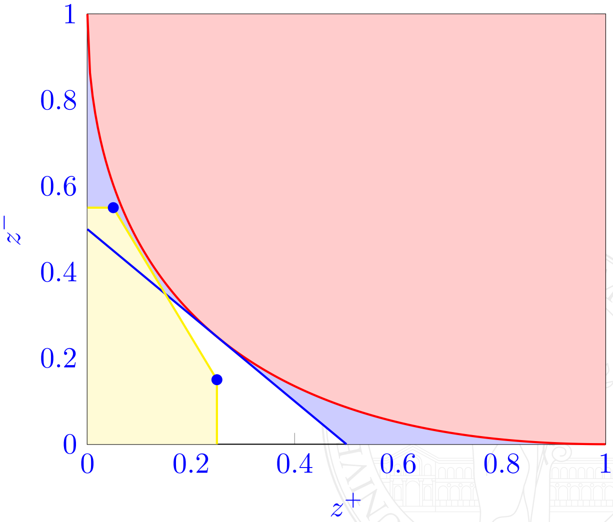
- Game matrix D for labels $\{1, \ldots, n\}$ where $D(i, j)$ is cost of predicting i when true label is j
- For any subset $J \subseteq \mathcal{Y}$ of labels: $V_J(D) = \min_{\boldsymbol{u} \in \Delta_{\mathcal{Y}}} \max_{\boldsymbol{v} \in \Delta_J} D(\boldsymbol{u}, \boldsymbol{v})$
- This is the smallest loss one can ensure by predicting with a die knowing that the correct label is in J.
- Boosting is only possible within any two consecutive game values

$$0 = v_1 < v_2 < \cdots < \boxed{v_j < v_{j+1}} < \cdots < v_N = V_{\mathcal{Y}}(D)$$

- Extension to multiobjective multiclass

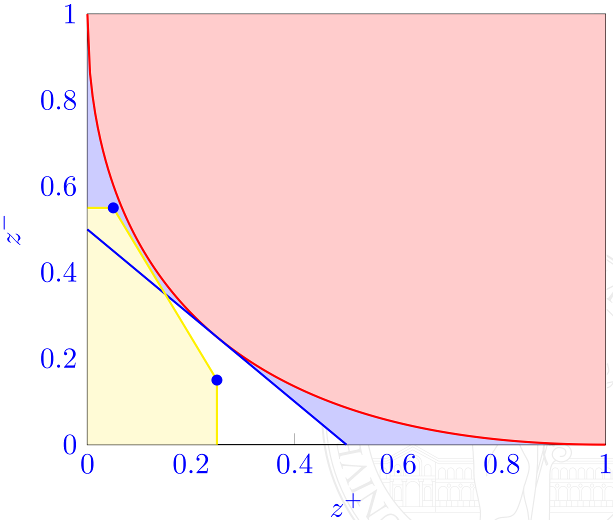
Open problems

- Suppose that the WL oracle provides guarantees in some convex region



Open problems

- Suppose that the WL oracle provides guarantees in some convex region
- Can we boost as long as the oracle region does not intersect the coin attainable region?



Open problems

- Suppose that the WL oracle provides guarantees in some convex region
- Can we boost as long as the oracle region does not intersect the coin attainable region?
- Can we do it adaptively?