Efficient PAC Learning for Realizable-Statistic Models via Convex Surrogates

Shivani Agarwal

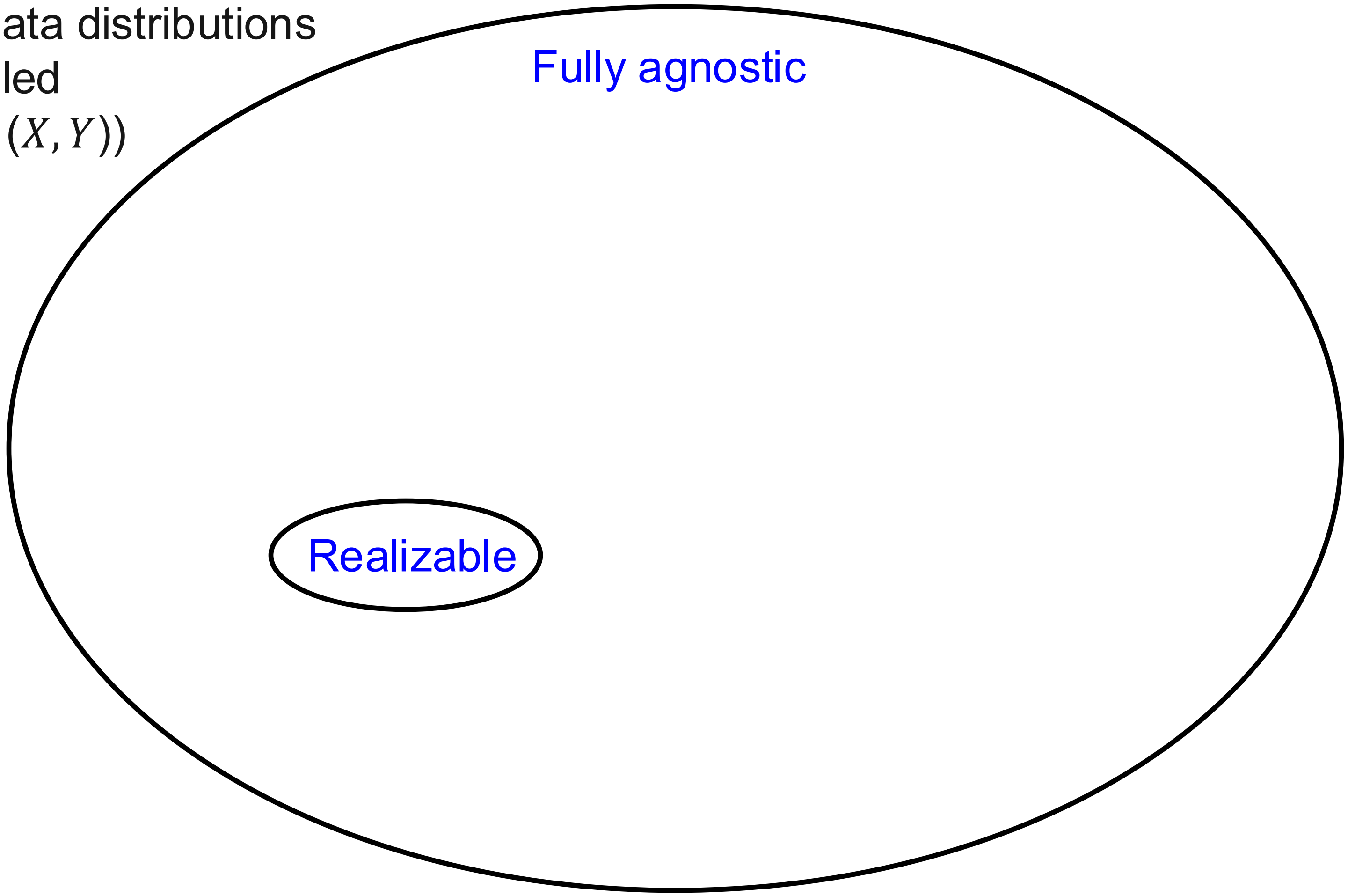
University of Pennsylvania

Probably Approximately Correct (PAC) Learning Model: Common Settings

- Realizable PAC learning [Valiant, 1984]
- (Fully) Agnostic PAC learning [Haussler, 1992; Kearns et al., 1994]

Probably Approximately Correct (PAC) Learning Model: Common Settings

Class of data distributions (over labeled examples (X, Y))

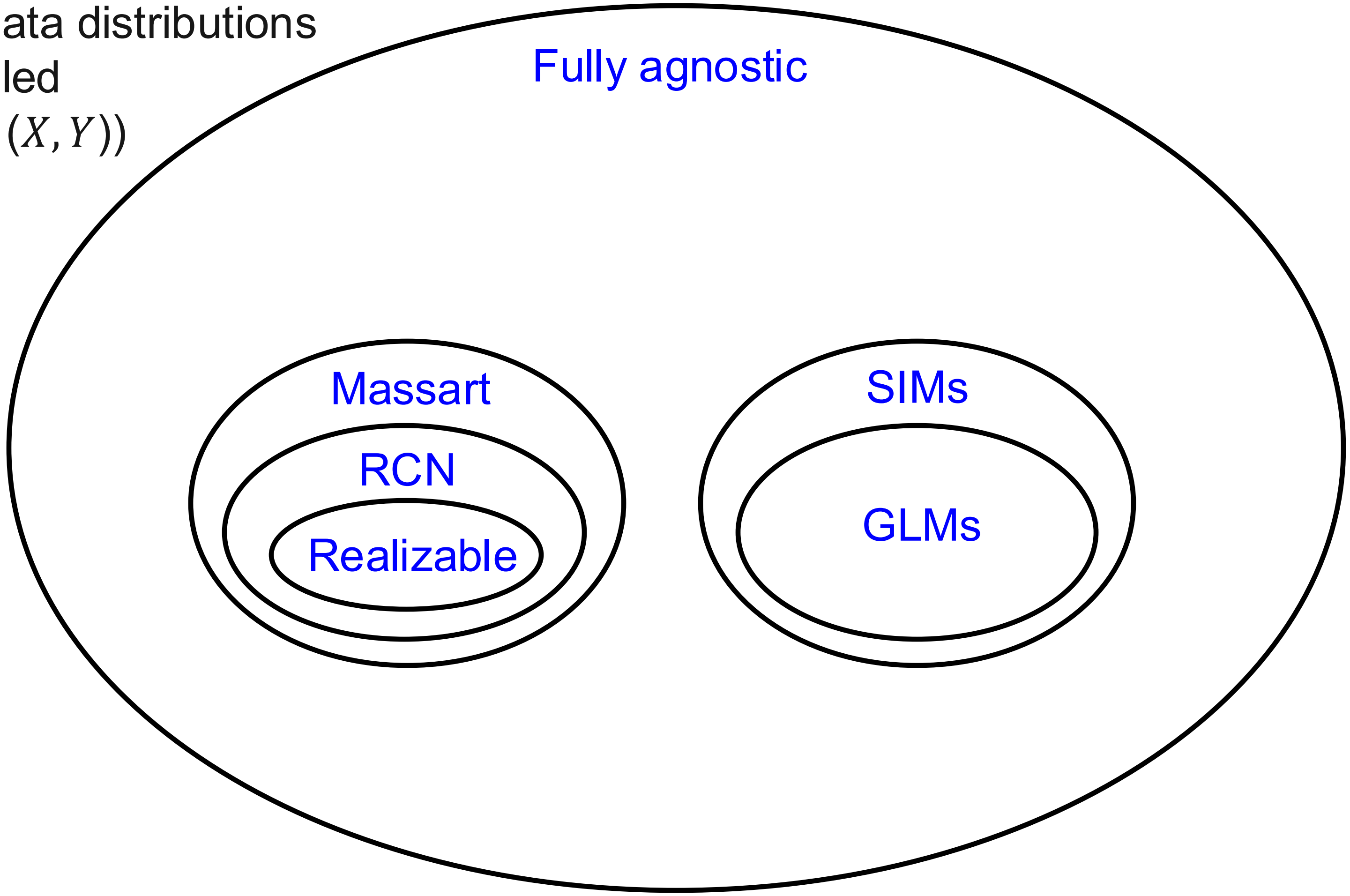


Intermediate PAC Learning Models

- Random classification noise (RCN) [Angluin & Laird, 1988; Bylander, 1994; Blum et al., 1998; Kearns, 1998; Long & Servedio, 2010]
- Probabilistic concepts [Kearns & Schapire, 1994]
- Massart noise [Sloan, 1988; Massart & Nédélec, 2006; Awasthi et al., 2015; 2016; Zhang et al., 2017; Diakonikolas et al., 2019; Chen et al., 2020]
- Generalized linear models (GLMs) and single index models (SIMs) [Kalai & Sastry, 2009; Kakade et al., 2011]

Intermediate PAC Learning Models

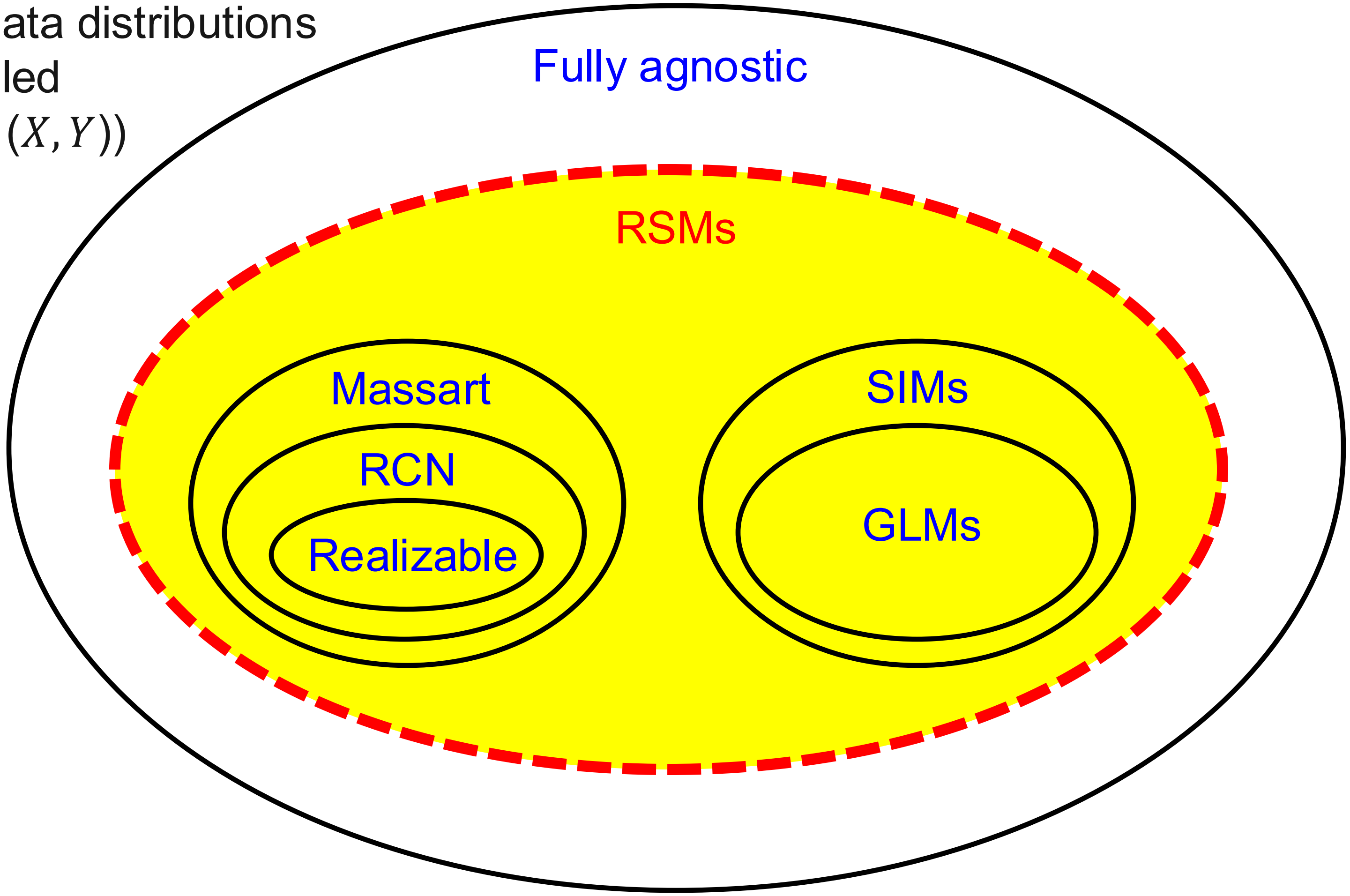
Class of data distributions (over labeled examples (X, Y))



This Work:
Realizable-Statistic Models (RSMs)

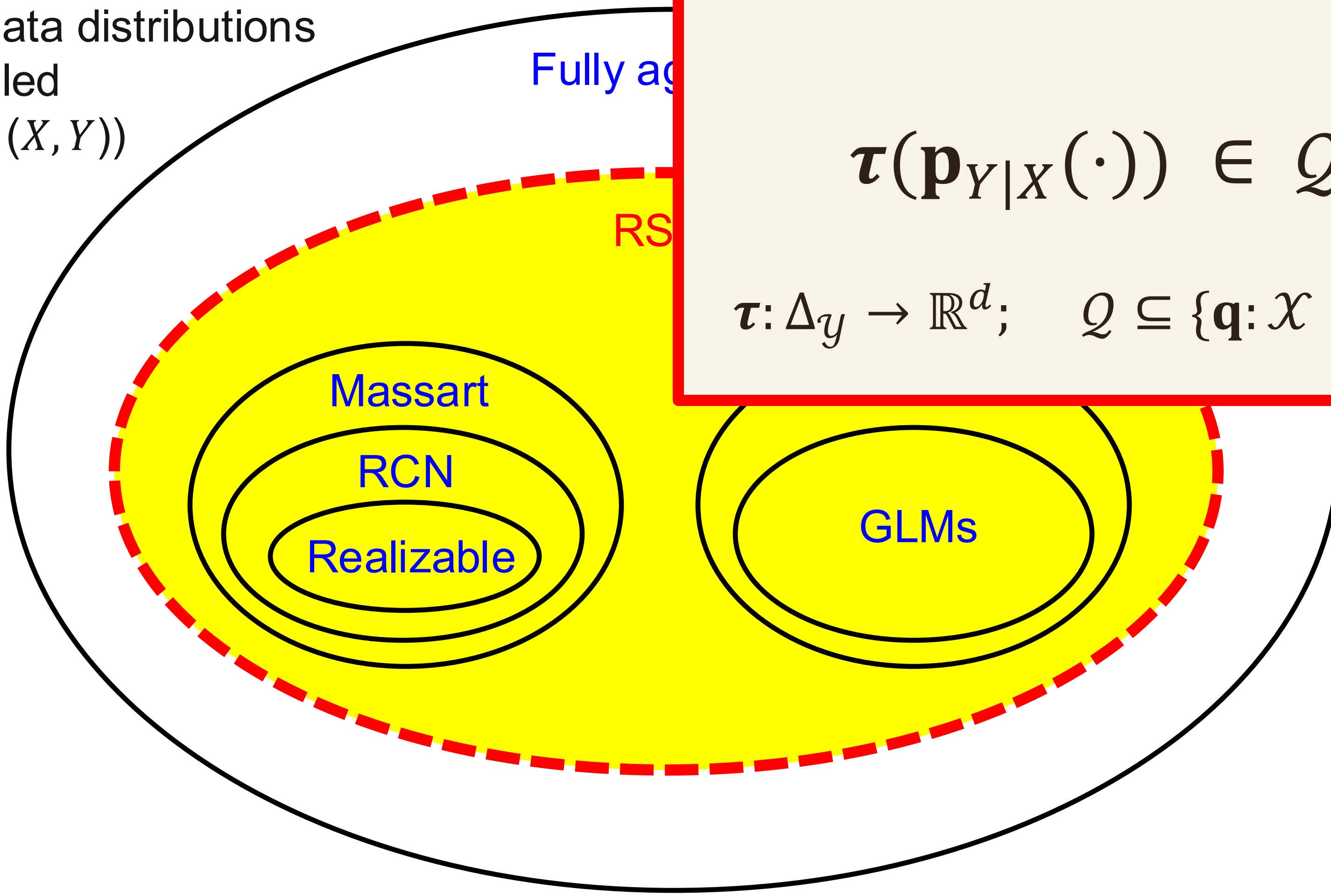
Realizable-Statistic Models (RSMs)

Class of data distributions (over labeled examples (X, Y))



Realizable-Statistic Models (RSMs)

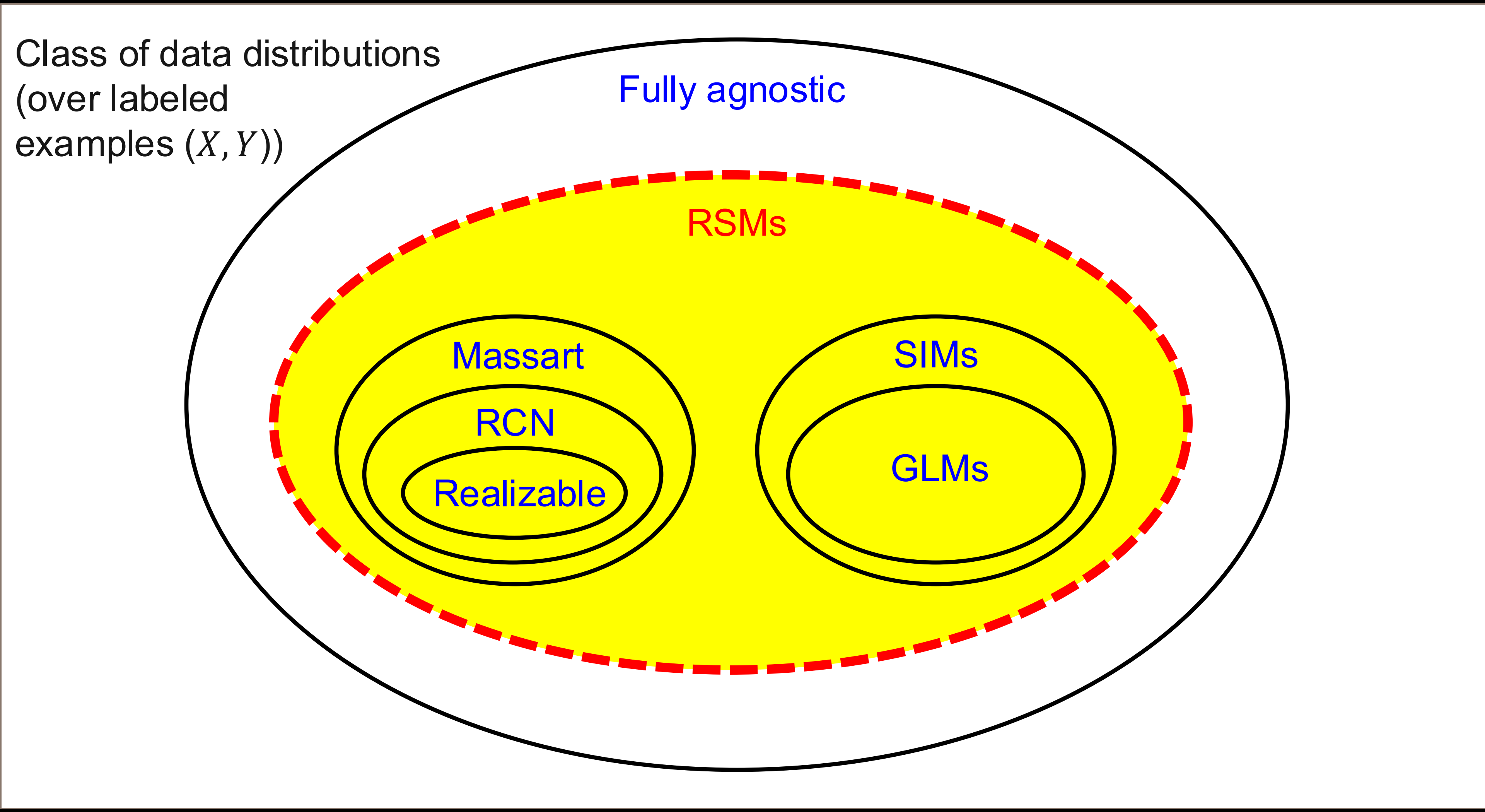
Class of data distributions (over labeled examples (X, Y))



$$\boldsymbol{\tau}(\mathbf{p}_{Y|X}(\cdot)) \in Q$$

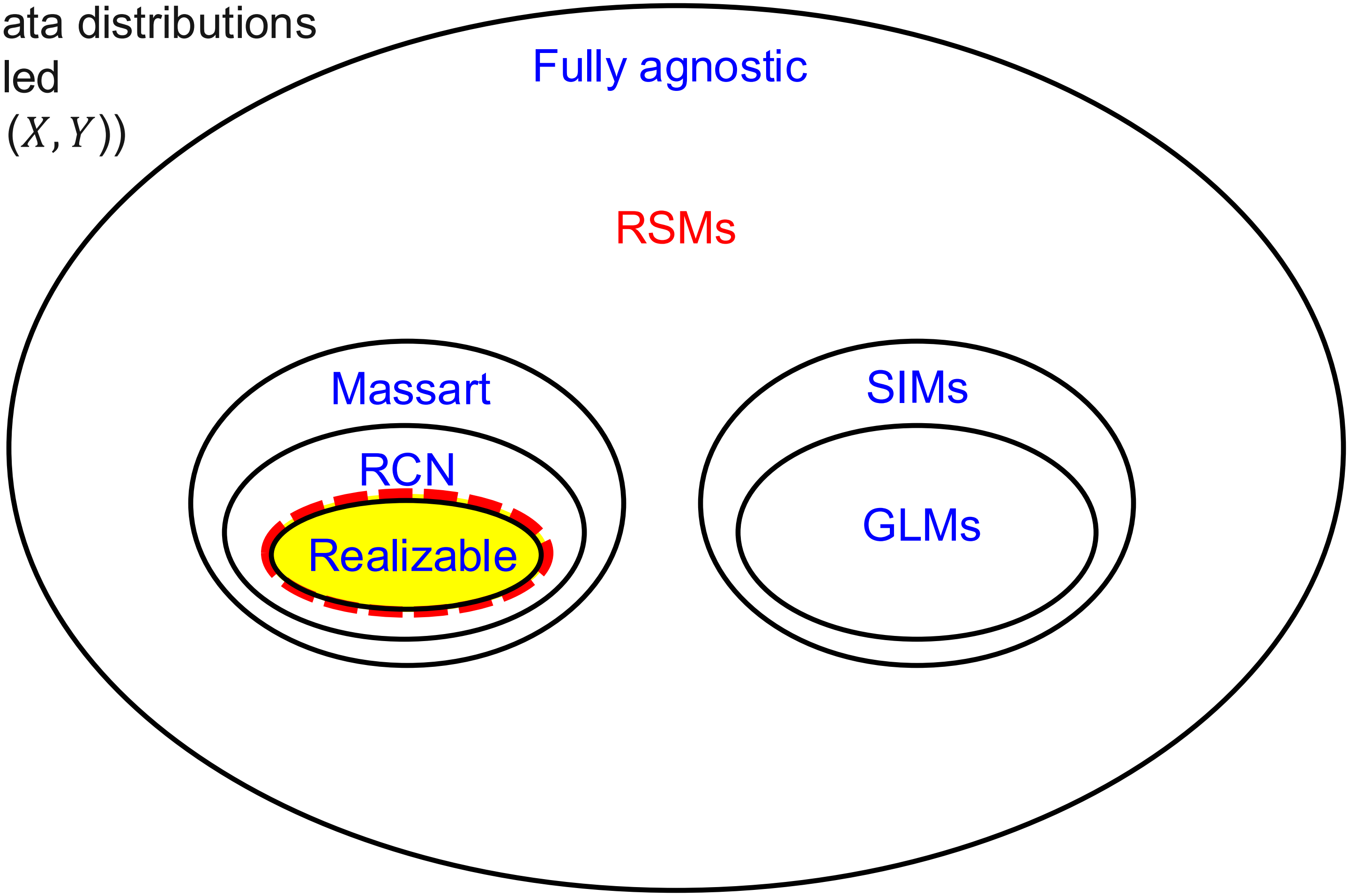
$$\boldsymbol{\tau}: \Delta_y \to \mathbb{R}^d; \quad Q \subseteq \{\mathbf{q}: \mathcal{X} \to \mathbb{R}^d\}$$

Realizable-Statistic Models (RSMs)



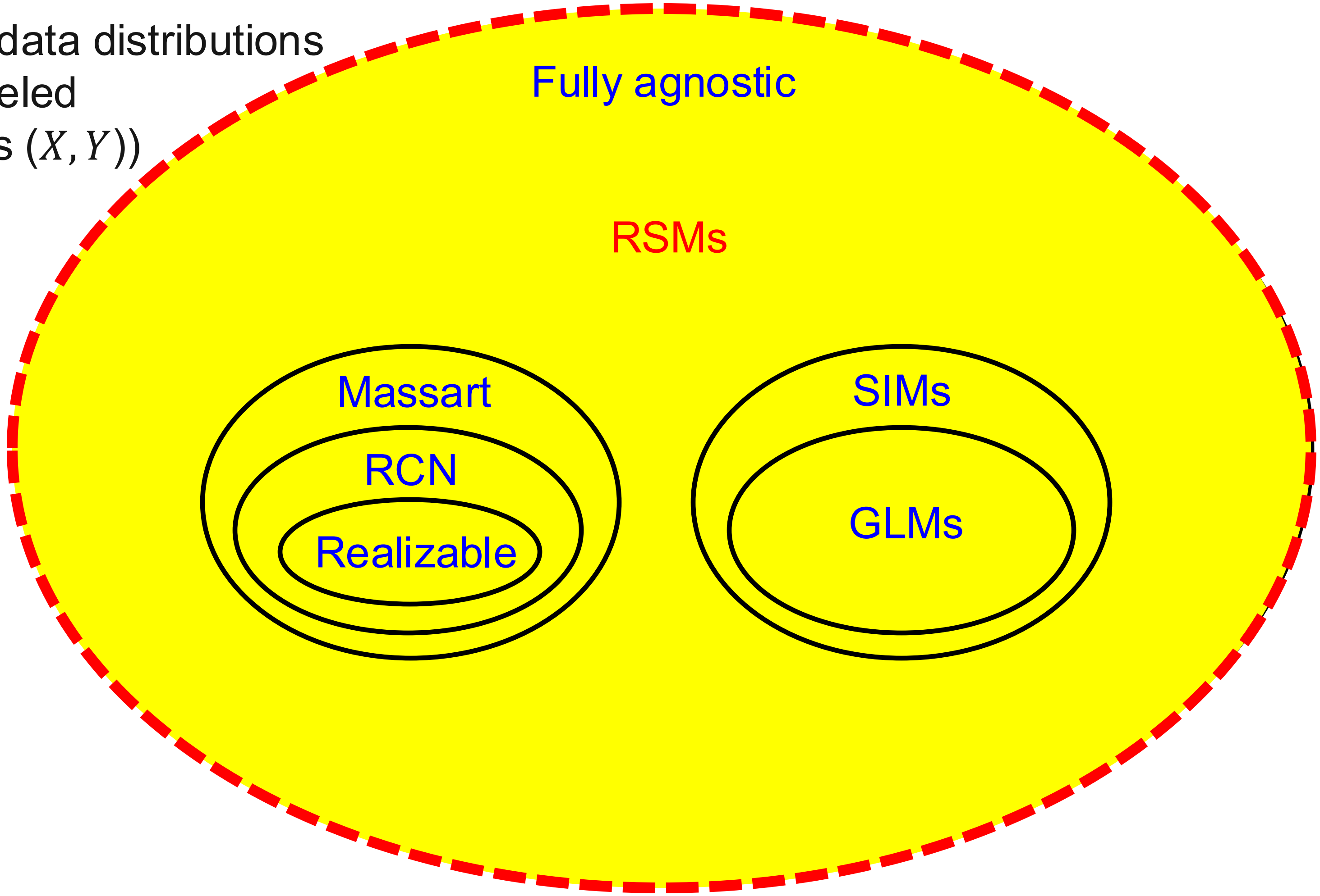
Realizable-Statistic Models (RSMs)

Class of data distributions (over labeled examples (X, Y))



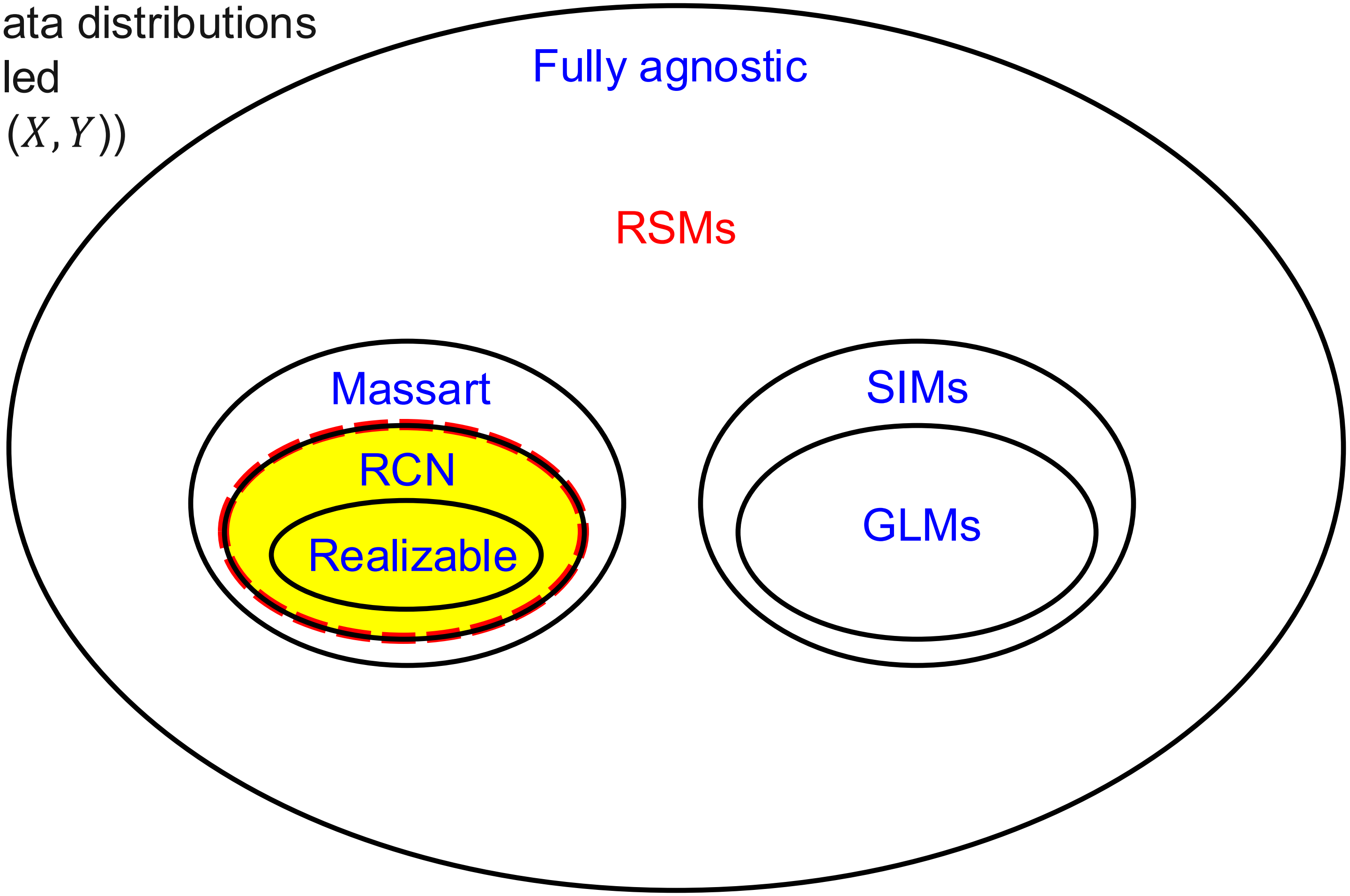
Realizable-Statistic Models (RSMs)

Class of data distributions
(over labeled
examples (X, Y))



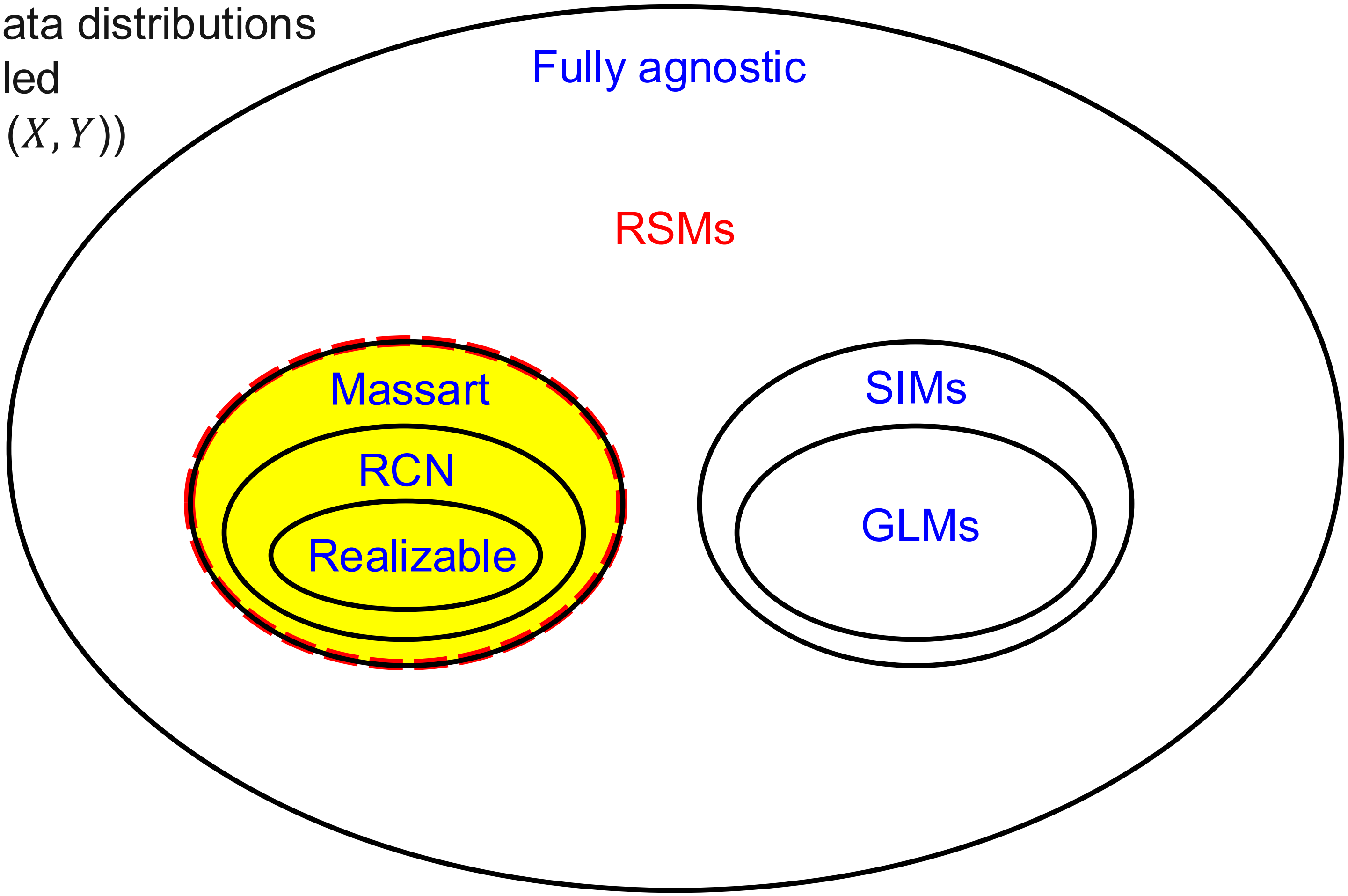
Realizable-Statistic Models (RSMs)

Class of data distributions (over labeled examples (X, Y))

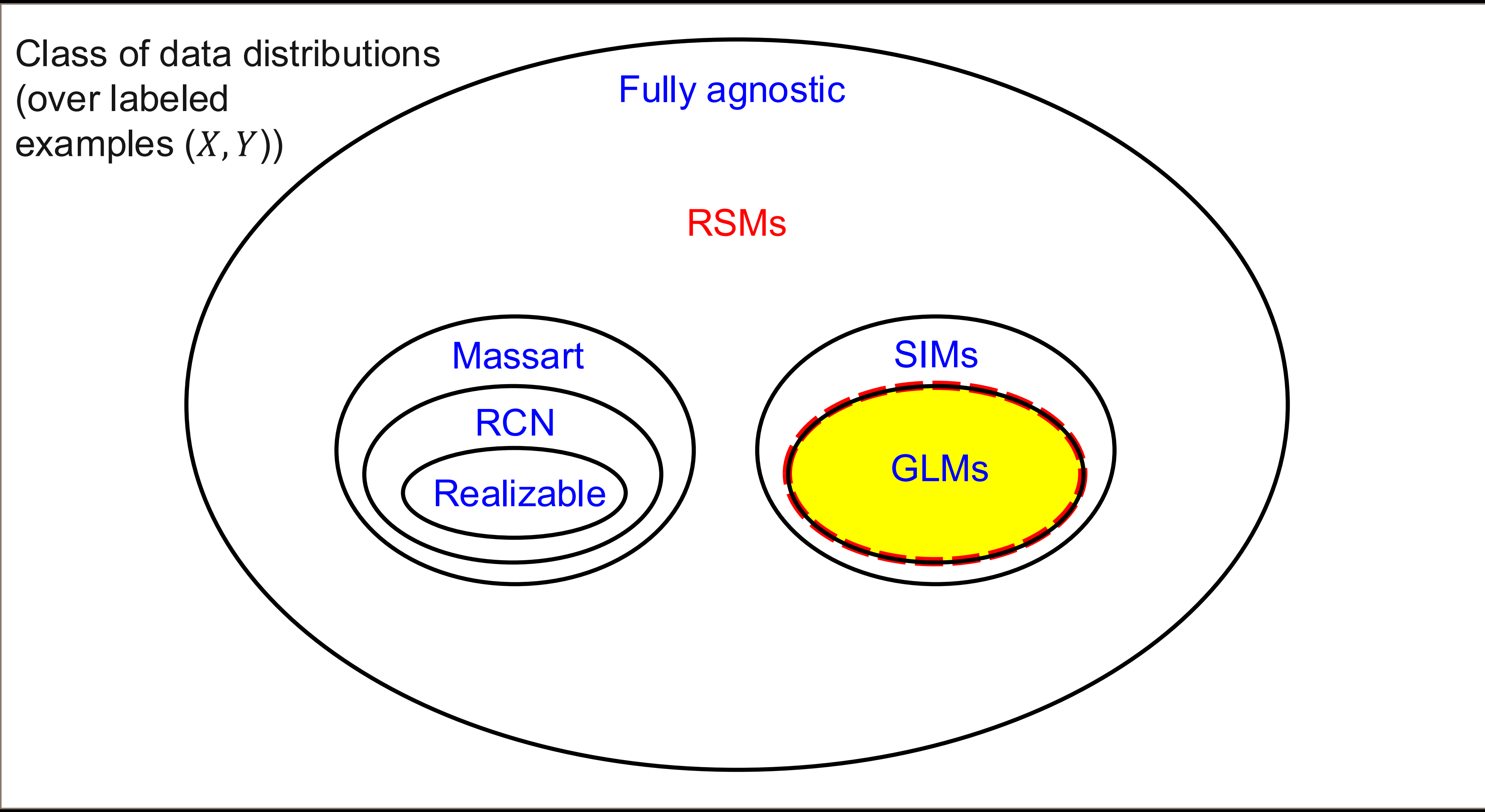


Realizable-Statistic Models (RSMs)

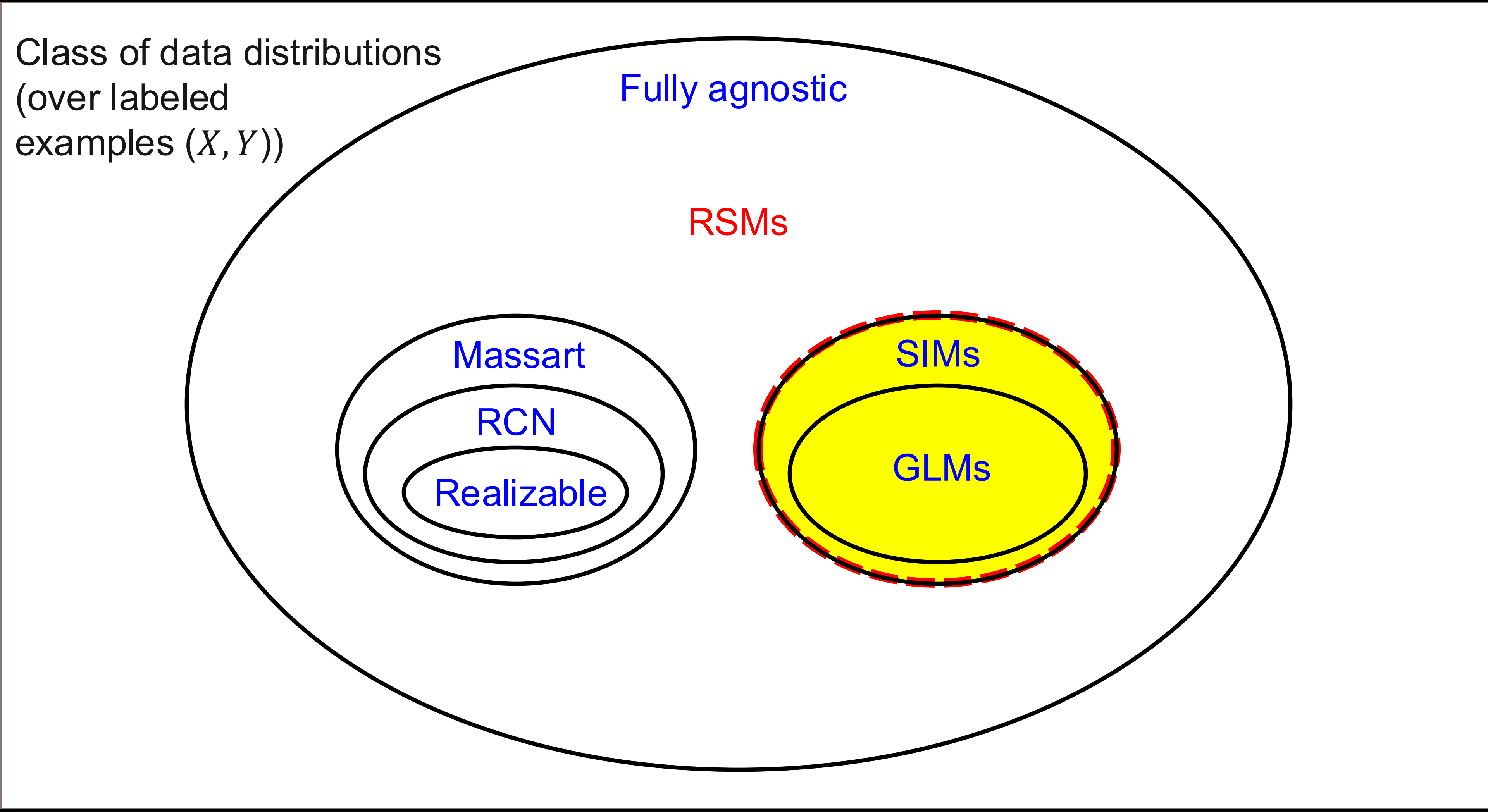
Class of data distributions (over labeled examples (X, Y))



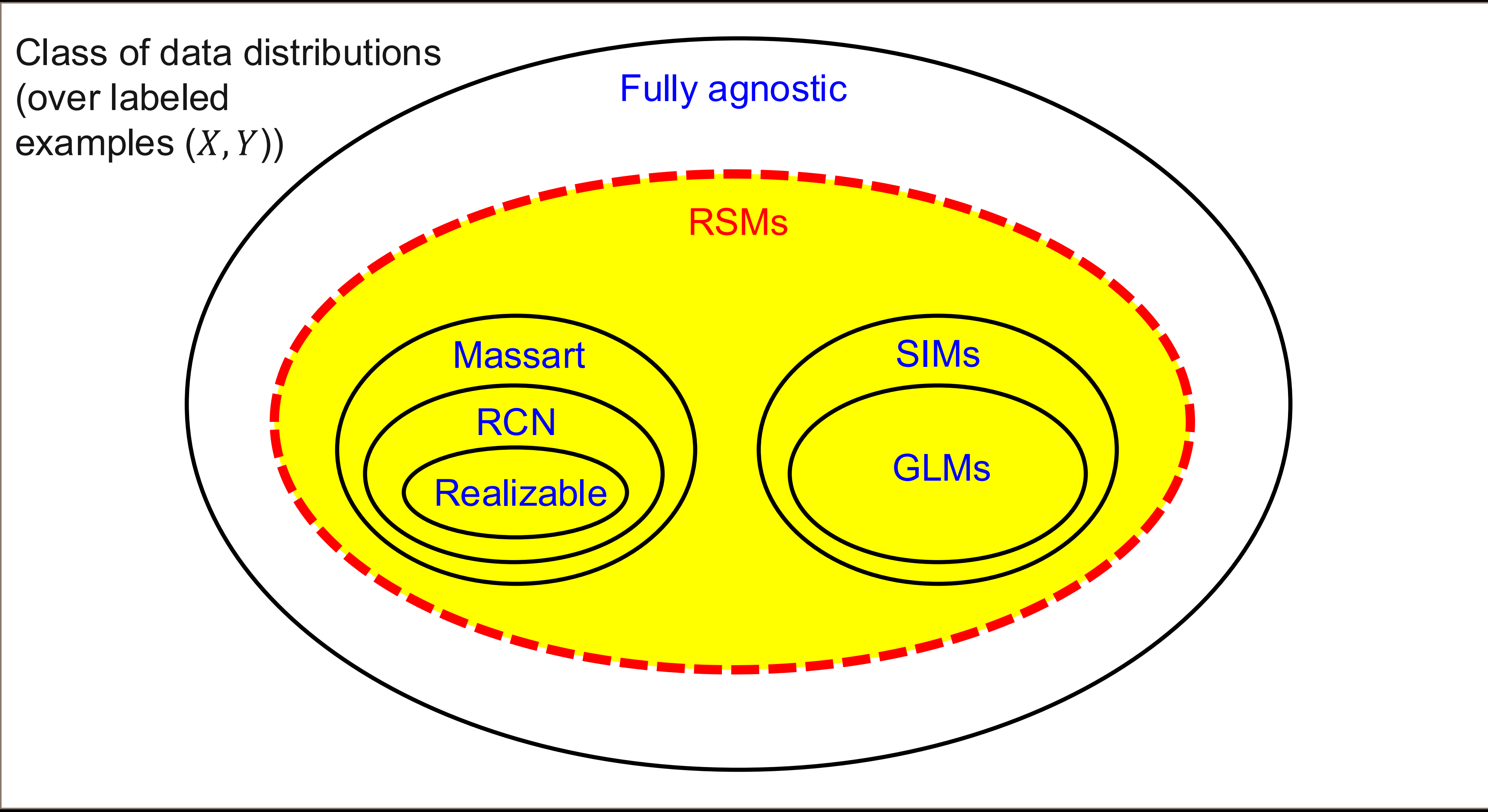
Realizable-Statistic Models (RSMs)



Realizable-Statistic Models (RSMs)



Realizable-Statistic Models (RSMs)



Main Results

- For many RSM learning problems, minimizing a suitable convex ‘strongly proper composite’ surrogate loss yields a computationally efficient learning algorithm with finite sample complexity bounds

- Applications to binary classification, multiclass classification, multi-label prediction, subset ranking

Step 1: Strongly Proper Composite Surrogate Losses

Definition 3 (Strongly proper composite surrogate losses for a statistic $\boldsymbol{\tau}$). *Let* $d \in \mathbb{Z}_+$ *and* $\mathcal{C} \subseteq \mathbb{R}^d$, *and let* $\boldsymbol{\tau} : \Delta_{\mathcal{Y}} \to \mathcal{C}$ *be any statistic of interest. Let* $d' \in \mathbb{Z}_+$, *and let* $\mathcal{C}' \subseteq \mathbb{R}^{d'}$ *be such that* $\mathcal{C}$ *is in one-to-one correspondence with a subset of* $\mathcal{C}'$. *If* $\mathcal{C}$ *is in one-to-one correspondence with* $\mathcal{C}'$ *itself, then let* $\boldsymbol{\lambda} : \mathcal{C} \to \mathcal{C}'$ *be an invertible mapping with inverse* $\boldsymbol{\lambda}^{-1} : \mathcal{C}' \to \mathcal{C}$; *otherwise, let* $\boldsymbol{\lambda} : \mathcal{C} \to \mathcal{C}'$ *be a one-to-one mapping and let* $\mathcal{S} = \{\mathcal{S}_{\mathbf{q}} : \mathbf{q} \in \mathcal{C}\}$ *be a partition of* $\mathcal{C}'$ *such that* $\boldsymbol{\lambda}(\mathbf{q}) \in \mathcal{S}_{\mathbf{q}}\ \forall \mathbf{q} \in \mathcal{C}$, *and let* $\boldsymbol{\lambda}^{-1} : \mathcal{C}' \to \mathcal{C}$ *denote an 'extended' inverse that assigns* $\boldsymbol{\lambda}^{-1}(\mathbf{u}) = \mathbf{q}\ \forall \mathbf{u} \in \mathcal{S}_{\mathbf{q}}$. *Let* $\gamma > 0$. *A surrogate loss* $\psi : \mathcal{Y} \times \mathcal{C}' \to \mathbb{R}_+$ *acting on* $\mathcal{C}'$ *is* γ-strongly proper composite for statistic $\boldsymbol{\tau}$ with link function $\boldsymbol{\lambda}$ *if* $\mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \mathbf{u}) - \psi(Y, \boldsymbol{\lambda}(\boldsymbol{\tau}(\mathbf{p})))] \geq \frac{\gamma}{2} \|\boldsymbol{\lambda}^{-1}(\mathbf{u}) - \boldsymbol{\tau}(\mathbf{p})\|_2^2\ \forall \mathbf{p} \in \Delta_{\mathcal{Y}},\ \mathbf{u} \in \mathcal{C}'$.

Step 1: Strongly Proper Composite Surrogate Losses

Definition 3 (Strongly proper composite surrogate losses for a statistic $\boldsymbol{\tau}$). *Let $d \in \mathbb{Z}_+$ and*

$\mathcal{C} \subseteq$

is i

itse

be c

and

sur

fun

$$\mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,\mathbf{u}) - \psi(Y,\boldsymbol{\lambda}(\boldsymbol{\tau}(\mathbf{p})))]$$

$$\geq \frac{\gamma}{2}\|\boldsymbol{\lambda}^{-1}(\mathbf{u}) - \boldsymbol{\tau}(\mathbf{p})\|_2^2$$

$$\forall \mathbf{p} \in \Delta_{\mathcal{Y}|},\ \mathbf{u} \in \mathcal{C}'$$

Step 2: Surrogate Regret Transfer Bound for (a Broad Class of) RSMs

Theorem 1 (Surrogate regret transfer bound for RSMs that admit strongly proper composite surrogate losses). *Let $\mathcal{X}$ be any instance space and $\mathcal{Y}, \widehat{\mathcal{Y}}$ be any label and prediction spaces, respectively. Let $\mathbf{L} \in \mathbb{R}_+^{\mathcal{Y}\times\widehat{\mathcal{Y}}}$ be a loss matrix. Let $d \in \mathbb{Z}_+$ and $\mathcal{C} \subseteq \mathbb{R}^d$. Let $\boldsymbol{\tau} : \Delta_{\mathcal{Y}} \to \mathcal{C}$ and* $\text{pred} : \mathcal{C} \to \widehat{\mathcal{Y}}$ *be such that $(\boldsymbol{\tau}, \text{pred})$ is an $\mathbf{L}$-calibrated statistic-mapping pair, and suppose $\exists \kappa > 0$ s.t.*

$$\mathbf{E}_{Y\sim\mathbf{p}}[L_{Y,\text{pred}(\mathbf{q})}] - \min_{\widehat{y}\in\mathcal{Y}} \mathbf{E}_{Y\sim\mathbf{p}}[L_{Y,\widehat{y}}] \;\le\; \kappa\|\mathbf{q} - \boldsymbol{\tau}(\mathbf{p})\|_2 \quad \forall \mathbf{p} \in \Delta_{\mathcal{Y}}, \mathbf{q} \in \mathcal{C}\,.$$

Let $\mathcal{Q} \subseteq \{\mathbf{q} : \mathcal{X}\to\mathcal{C}\}$ be a class of 'statistic' functions, and let $\psi : \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}_+$ be a γ-strongly proper composite surrogate loss for $\boldsymbol{\tau}$ with link function $\boldsymbol{\lambda} : \mathcal{C} \to \mathbb{R}^d$.[4] *Let $\mathcal{H} \subseteq \{h : \mathcal{X}\to\widehat{\mathcal{Y}}\}$ be defined as $\mathcal{H} := \text{pred} \circ \mathcal{Q} = \{h : \mathcal{X}\to\widehat{\mathcal{Y}} \mid \exists \mathbf{q} \in \mathcal{Q} \text{ s.t. } h(x) = \text{pred}(\mathbf{q}(x))\ \forall x \in \mathcal{X}\}$, let $\mathcal{F} \subseteq \{f : \mathcal{X}\to\mathbb{R}^d\}$ be defined as $\mathcal{F} := \boldsymbol{\lambda} \circ \mathcal{Q} = \{\mathbf{f} : \mathcal{X}\to\mathbb{R}^d \mid \exists \mathbf{q} \in \mathcal{Q} \text{ s.t. } \mathbf{f}(x) = \boldsymbol{\lambda}(\mathbf{q}(x))\ \forall x \in \mathcal{X}\}$, and define* $\text{decode} : \mathbb{R}^d \to \widehat{\mathcal{Y}}$ *as* $\text{decode} := \text{pred} \circ \boldsymbol{\lambda}^{-1}$. *Suppose that $\psi(y, \mathbf{f}(x)) \in [0, B]\ \forall x \in \mathcal{X}, y \in \mathcal{Y}, \mathbf{f} \in \mathcal{F}$ for some $B > 0$. Then for any $\mathbf{f} \in \mathcal{F}$ and any $D \in \mathcal{D}_{(\boldsymbol{\tau},\mathcal{Q})\text{-RSM}}$,*

$$\text{er}_D^{\mathbf{L}}[\underbrace{\text{decode} \circ \mathbf{f}}_{h}] - \text{er}_D^{\mathbf{L}}[\mathcal{H}] \;\le\; \kappa \cdot \sqrt{\mathbf{E}_X[\|\boldsymbol{\lambda}^{-1}(\mathbf{f}(X)) - \boldsymbol{\tau}(\mathbf{p}(X))\|_2^2]} \;\le\; \kappa \cdot \sqrt{\tfrac{2}{\gamma}(\text{er}_D^{\psi}[\mathbf{f}] - \text{er}_D^{\psi}[\mathcal{F}])}\,.$$

Step 2: Surrogate Regret Transfer Bound for (a Broad Class of) RSMs

Theorem 1 (Surrogate regret transfer bound for RSMs that admit strongly proper composite surrogate losses). *Let $\mathcal{X}$ be any instance space and $\mathcal{Y}, \widehat{\mathcal{Y}}$ be any label and prediction spaces, respectively. Let* $\mathbf{L} \in \mathbb{R}_+^{\mathcal{Y} \times \widehat{\mathcal{Y}}}$ *be a loss matrix. Let* $d \in \mathbb{Z}_+$ *and* $\mathcal{C} \subseteq \mathbb{R}^d$. *Let* $\boldsymbol{\tau} : \Delta_{\mathcal{Y}} \rightarrow \mathcal{C}$ *and*

pred

s.t.

Let $\mathcal{Q}$

prope

defin

$\mathcal{X} \rightarrow \mathbb{I}$

define

$\mathcal{Y}, \mathbf{f} \in$

$\mathrm{er}_D^{\mathbf{L}}[\mathrm{d}$

h

$$\mathrm{er}_D^{\mathbf{L}}[\underbrace{\mathrm{decode} \circ \mathbf{f}}_{h}] - \mathrm{er}_D^{\mathbf{L}}[\mathcal{H}]$$

$$\leq \kappa \cdot \sqrt{\mathbf{E}_X[\|\boldsymbol{\lambda}^{-1}(\mathbf{f}(X)) - \boldsymbol{\tau}(\mathbf{p}(X))\|_2^2]}$$

$$\leq \kappa \cdot \sqrt{\frac{2}{\gamma}(\mathrm{er}_D^{\psi}[\mathbf{f}] - \mathrm{er}_D^{\psi}[\mathcal{F}])}$$

Step 3: RSM Learning Bounds for Surrogate Risk Minimizers

Theorem 2 (RSM learning bounds for surrogate risk minimizers via d_1 covering numbers). *Under the conditions of Theorem 1, suppose the surrogate loss ψ is ρ_1-Lipschitz in the second argument with respect to the L^1 metric, so that $\psi(y, \mathbf{u}_1) - \psi(y, \mathbf{u}_2) \le \rho_1 \|\mathbf{u}_1 - \mathbf{u}_2\|_1 \; \forall y, \mathbf{u}_1, \mathbf{u}_2$, and suppose that the function classes $\mathcal{F}^j = \{f_j : \mathcal{X} \to \mathbb{R} \mid \exists \mathbf{f} \in \mathcal{F} \text{ s.t. } f_j(x) = (\mathbf{f}(x))_j \; \forall x\}$, $j \in [d]$ each have bounded d_1 covering numbers $\mathcal{N}_1(\epsilon, \mathcal{F}^j, m)$ (polynomial in m and $1/\epsilon$). Then a surrogate risk minimization algorithm $\mathcal{A}$ which, given a training sample S of size m, finds an $(16B/\sqrt{m})$-approximate minimizer $\widehat{\mathbf{f}}_S \in \mathcal{F}$ of the empirical surrogate risk $\frac{1}{m}\sum_{i=1}^m \psi(y_i, \mathbf{f}(x_i))$ over $\mathcal{F}$, and produces a $\boldsymbol{\tau}$-statistic estimate $\widehat{\mathbf{q}}_S(x) = \boldsymbol{\lambda}^{-1}(\widehat{\mathbf{f}}_S(x))$ and a prediction model $\widehat{h}_S \in \mathcal{H}$ given by $\widehat{h}_S(x) = \mathrm{decode}(\widehat{\mathbf{f}}_S(x))$ (or equivalently, $\widehat{h}_S(x) = \mathrm{pred}(\widehat{\mathbf{q}}_S(x))$), is a PAC learning algorithm for the RSM learning problem $(\mathbf{L}, \mathcal{H}, \mathcal{D}_{(\boldsymbol{\tau},\mathcal{Q})\text{-RSM}})$ with squared $\boldsymbol{\tau}$-estimation error sample complexity $m^{\boldsymbol{\tau}}_{\mathcal{A}}(\epsilon, \delta) \le \min\big\{m_0 \in \mathbb{Z}_+ : m \ge m_0 \implies m \ge \frac{1152B^2}{\gamma^2\epsilon^2}\big(\sum_{j=1}^d \ln\big(\mathcal{N}_1\big(\frac{\gamma\epsilon}{48\rho_1 d}, \mathcal{F}^j, 2m\big)\big) + \ln\big(\frac{4}{\delta}\big)\big)\big\}$, and with target loss sample complexity $m^{\mathbf{L}}_{\mathcal{A}}(\epsilon, \delta) \le \min\big\{m \in \mathbb{Z}_+ : m \ge m_0 \implies m \ge \frac{1152\kappa^4B^2}{\gamma^2\epsilon^4}\big(\sum_{j=1}^d \ln\big(\mathcal{N}_1\big(\frac{\gamma\epsilon^2}{48\kappa^2\rho_1 d}, \mathcal{F}^j, 2m\big)\big) + \ln\big(\frac{4}{\delta}\big)\big)\big\}$. In particular, if the d_1 covering numbers of the function classes $\mathcal{F}^j$ have upper bounds of the form $\mathcal{N}_1(\epsilon, \mathcal{F}^j, m) \le \phi(\epsilon, \mathcal{F}^j)$ (i.e., bounds independent of sample size m), then $m^{\boldsymbol{\tau}}_{\mathcal{A}}(\epsilon, \delta) \le \frac{1152B^2}{\gamma^2\epsilon^2}\big(\sum_{j=1}^d \ln\big(\phi\big(\frac{\gamma\epsilon}{48\rho_1 d}, \mathcal{F}^j\big)\big) + \ln\big(\frac{4}{\delta}\big)\big)$, and $m^{\mathbf{L}}_{\mathcal{A}}(\epsilon, \delta) \le \frac{1152\kappa^4B^2}{\gamma^2\epsilon^4}\big(\sum_{j=1}^d \ln\big(\phi\big(\frac{\gamma\epsilon^2}{48\kappa^2\rho_1 d}, \mathcal{F}^j\big)\big) + \ln\big(\frac{4}{\delta}\big)\big)$.*

Step 3: RSM Learning Bounds for Surrogate Risk Minimizers

Theorem 2 (RSM learning bounds for surrogate risk minimizers via d_1 covering numbers). *Under the conditions of Theorem 1, suppose the surrogate loss ψ is ρ_1-Lipschitz in the second argument with respect to the L^1 metric, so that $\psi(y, \mathbf{u}_1) - \psi(y, \mathbf{u}_2) \leq \rho_1 \|\mathbf{u}_1 - \mathbf{u}_2\|_1 \ \forall y, \mathbf{u}_1, \mathbf{u}_2$, and suppose that the function classes $\mathcal{F}^j = \{f_j : \mathcal{X} \to \mathbb{R} \mid \exists \mathbf{f} \in \mathcal{F} \text{ s.t. } f_j(x) = (\mathbf{f}(x))_j \ \forall x\}$, $j \in [d]$*

For $\mathcal{N}_1(\epsilon, \mathcal{F}^j, m) \leq \phi(\epsilon, \mathcal{F}^j)$:

$$m^{\boldsymbol{\tau}}_{\mathcal{A}}(\epsilon, \delta) \leq \frac{1152 B^2}{\gamma^2 \epsilon^2} \Big(\sum_{j=1}^{d} \ln \Big(\phi\Big(\frac{\gamma \epsilon}{48 \rho_1 d}, \mathcal{F}^j \Big) \Big) + \ln \Big(\frac{4}{\delta} \Big) \Big)$$

$$m^{\mathbf{L}}_{\mathcal{A}}(\epsilon, \delta) \leq \frac{1152 \kappa^4 B^2}{\gamma^2 \epsilon^4} \Big(\sum_{j=1}^{d} \ln \Big(\phi\Big(\frac{\gamma \epsilon^2}{48 \kappa^2 \rho_1 d}, \mathcal{F}^j \Big) \Big) + \ln \Big(\frac{4}{\delta} \Big) \Big)$$

$m^{\mathbf{L}}_{\mathcal{A}}(\epsilon, \delta) \leq \frac{1152 \kappa^4 B^2}{\gamma^2 \epsilon^4} \big(\sum_{j=1}^{d} \ln \big(\phi\big(\frac{\gamma \epsilon^2}{48 \kappa^2 \rho_1 d}, \mathcal{F}^j \big) \big) + \ln \big(\frac{4}{\delta} \big) \big)$.

Step 3: RSM Learning Bounds for Surrogate Risk Minimizers

Theorem 3 (RSM learning bounds for surrogate risk minimizers via Rademacher complexities). *Under the conditions of Theorem 1, suppose the surrogate loss ψ is ρ_2-Lipschitz in the second argument with respect to the Euclidean metric, so that $\psi(y, \mathbf{u}_1) - \psi(y, \mathbf{u}_2) \le \rho_2 \|\mathbf{u}_1 - \mathbf{u}_2\|_2 \; \forall y, \mathbf{u}_1, \mathbf{u}_2$, and suppose that the function classes $\mathcal{F}^j = \{f_j : \mathcal{X} \to \mathbb{R} \,|\, \exists \mathbf{f} \in \mathcal{F} \text{ s.t. } f_j(x) = (\mathbf{f}(x))_j \; \forall x\}$, $j \in [d]$ each have non-negative, decreasing Rademacher complexities $\mathcal{R}_m(\mathcal{F}^j)$ (decreasing in m). Then a surrogate risk minimization algorithm $\mathcal{A}$ which, given a training sample S of size m, finds an $(B/(2\sqrt{m}))$-approximate minimizer $\widehat{\mathbf{f}}_S \in \mathcal{F}$ of the empirical surrogate risk $\frac{1}{m}\sum_{i=1}^m \psi(y_i, \mathbf{f}(x_i))$ over $\mathcal{F}$, and produces a $\boldsymbol{\tau}$-statistic estimate $\widehat{\mathbf{q}}_S(x) = \boldsymbol{\lambda}^{-1}(\widehat{\mathbf{f}}_S(x))$ and a prediction model $\widehat{h}_S \in \mathcal{H}$ given by $\widehat{h}_S(x) = \mathrm{decode}(\widehat{\mathbf{f}}_S(x))$ (or equivalently, $\widehat{h}_S(x) = \mathrm{pred}(\widehat{\mathbf{q}}_S(x))$), is a PAC learning algorithm for the RSM learning problem $(\mathbf{L}, \mathcal{H}, \mathcal{D}_{(\boldsymbol{\tau},\mathcal{Q})\text{-RSM}})$ with squared $\boldsymbol{\tau}$-estimation error sample complexity $m^{\boldsymbol{\tau}}_{\mathcal{A}}(\epsilon,\delta) \le \min\{m_0 \in \mathbb{Z}_+ : m \ge m_0 \implies 3\Big(2\sqrt{2}\rho_2 \cdot \sum_{j=1}^d \mathcal{R}_m(\mathcal{F}^j) + B\sqrt{\frac{\ln(2/\delta)}{m}}\Big) \le \frac{\gamma\epsilon}{2}\}$, and with target loss sample complexity $m^{\mathbf{L}}_{\mathcal{A}}(\epsilon,\delta) \le \min\{m \in \mathbb{Z}_+ : m \ge m_0 \implies 3\Big(2\sqrt{2}\rho_2 \cdot \sum_{j=1}^d \mathcal{R}_m(\mathcal{F}^j) + B\sqrt{\frac{\ln(2/\delta)}{m}}\Big) \le \frac{\gamma\epsilon^2}{2\kappa^2}\}$. In particular, if $\exists C > 0$ such that the Rademacher complexities of the function classes $\mathcal{F}^j$ have upper bounds of the form $\mathcal{R}_m(\mathcal{F}^j) \le C/\sqrt{m} \; \forall j \in [d]$, then $m^{\boldsymbol{\tau}}_{\mathcal{A}}(\epsilon,\delta) \le \frac{36}{\gamma^2\epsilon^2}\big(2\sqrt{2}\rho_2 C d + B\sqrt{\ln(2/\delta)}\,\big)^2$, and $m^{\mathbf{L}}_{\mathcal{A}}(\epsilon,\delta) \le \frac{36\kappa^4}{\gamma^2\epsilon^4}\big(2\sqrt{2}\rho_2 C d + B\sqrt{\ln(2/\delta)}\,\big)^2$.*

Step 3: RSM Learning Bounds for Surrogate Risk Minimizers

Theorem 3 (RSM learning bounds for surrogate risk minimizers via Rademacher complexities). *Under the conditions of Theorem 1, suppose the surrogate loss* ψ *is* ρ_2*-Lipschitz in the second argument with respect to the Euclidean metric, so that* $\psi(y, \mathbf{u}_1) - \psi(y, \mathbf{u}_2) \leq \rho_2 \left\| \mathbf{u}_1 - \mathbf{u}_2 \right\|_2 \; \forall y, \mathbf{u}_1, \mathbf{u}_2$*, and suppose that the function classes* $\mathcal{F}^j = \{ f_j : \mathcal{X} \to \mathbb{R} \mid \exists \mathbf{f} \in \mathcal{F} \text{ s.t. } f_j(x) = (\mathbf{f}(x))_j \; \forall x \}$ ...

For $\mathcal{R}_m(\mathcal{F}^j) \leq C/\sqrt{m}$:

$$m^{\boldsymbol{\tau}}_{\mathcal{A}}(\epsilon, \delta) \leq \frac{36}{\gamma^2 \epsilon^2} \left(2\sqrt{2}\rho_2 C d + B\sqrt{\ln(2/\delta)} \right)^2,$$

$$m^{\mathbf{L}}_{\mathcal{A}}(\epsilon, \delta) \leq \frac{36\kappa^4}{\gamma^2 \epsilon^4} \left(2\sqrt{2}\rho_2 C d + B\sqrt{\ln(2/\delta)} \right)^2.$$

bounds of the form $\mathcal{R}_m(\mathcal{F}^j) \leq C/\sqrt{m} \; \forall j \in [d]$*, then* $m^{\boldsymbol{\tau}}_{\mathcal{A}}(\epsilon, \delta) \leq \frac{36}{\gamma^2 \epsilon^2} \left(2\sqrt{2}\rho_2 C d + B\sqrt{\ln(2/\delta)} \right)^2$*, and* $m^{\mathbf{L}}_{\mathcal{A}}(\epsilon, \delta) \leq \frac{36\kappa^4}{\gamma^2 \epsilon^4} \left(2\sqrt{2}\rho_2 C d + B\sqrt{\ln(2/\delta)} \right)^2$.

Applications

- Binary classification (0-1 loss)
- Multiclass classification (0-1 loss)
- Multi-label prediction (Hamming loss)
- Subset ranking (DCG metric)

Applications

Assumption on conditional label distribution $\mathbf{P}(Y \mid X = x)$	Learning target	Sample complexity (for squared estimation error $\le \epsilon$)	Sample complexity (for target loss based regret $\le \epsilon$)	Computational complexity (m = sample complexity from column 3 or 4)
Binary classification with 0-1 loss $[\mathcal{X} \subseteq \mathbb{R}^p, \mathcal{Y} = \widehat{\mathcal{Y}} = \{\pm 1\}]$				
Noisy LTF: RCN [10, 17, 21]	Best LTF		$\text{poly}(p, 1/\epsilon)$	$\text{poly}(p, 1/\epsilon)$
Noisy LTF: Massart noise [15]	Upper bound η on Massart noise		$\widetilde{O}(\text{poly}(p)/\epsilon^3)$	$\text{poly}(p, 1/\epsilon)$
GLM [25] (Kakade et al., 2011)	Best LTF	$\widetilde{O}(1/\epsilon^2)$		$\widetilde{O}(m^{3/2}p)$
SIM [25]	Best LTF	*(i)* $\widetilde{O}(p/\epsilon^3)$ *(ii)* $\widetilde{O}(1/\epsilon^4)$		*(i)* $\widetilde{O}(m^{4/3}p)$ *(ii)* $\widetilde{O}(m^{5/4}p)$
Sigmoid-of-linear [as special case of RSMs]	Best LTF	$\widetilde{O}(1/\epsilon^2)$	$\widetilde{O}(1/\epsilon^4)$	$\widetilde{O}(m^{5/4}p)$
Multiclass classification with 0-1 loss (n classes) $[\mathcal{X} \subseteq \mathbb{R}^p, \mathcal{Y} = \mathcal{Y} = [n]]$				
Softmax-of-multilinear [as special case of RSMs]	Best multilinear multiclass classifier	*(i)* $\widetilde{O}(np/\epsilon^2)$ *(ii)* $\widetilde{O}(n^2/\epsilon^2)$	*(i)* $\widetilde{O}(np/\epsilon^4)$ *(ii)* $\widetilde{O}(n^2/\epsilon^4)$	*(i)* $\widetilde{O}(m^{5/4}np)$ *(ii)* $\widetilde{O}(m^{5/4}np)$
Multi-label prediction with Hamming loss (s tags) $[\mathcal{X} \subseteq \mathbb{R}^p, \mathcal{Y} = \widehat{\mathcal{Y}} = \{0,1\}^s]$				
Sigmoid-of-linear marginals [as special case of RSMs]	Best multilinear multi-label prediction model	$\widetilde{O}(s^3/\epsilon^2)$	$\widetilde{O}(s^5/\epsilon^4)$	$\widetilde{O}(m^{5/4}sp)$
Subset ranking with DCG metric (s items, r rating levels) $[\mathcal{X} \subseteq \mathbb{R}^p, \mathcal{Y} = \{0,1,\ldots,r\}^s, \widehat{\mathcal{Y}} = \Pi_s]$				
Sigmoid-of-linear scaled marginal expectations [as special case of RSMs]	Best multilinear subset ranking model	$\widetilde{O}(s^3/\epsilon^2)$	$\widetilde{O}(r^4 s^5/\epsilon^4)$	$\widetilde{O}(m^{5/4}sp)$

Applications

Assumption on conditional label distribution $\mathbf{P}(Y\|X=x)$	Learning target	Sample complexity (for squared estimation error $\leq \epsilon$)	Sample complexity (for target loss based regret $\leq \epsilon$)	Computational complexity (m = sample complexity from column 3 or 4)
Binary classification with 0-1 loss $[\mathcal{X} \subseteq \mathbb{R}^p, \mathcal{Y} = \widehat{\mathcal{Y}} = \{\pm 1\}]$				
Noisy LTF: RCN [10, 17, 21]	Best LTF		$\text{poly}(p, 1/\epsilon)$	$\text{poly}(p, 1/\epsilon)$
Noisy LTF: Massart noise [15]	Upper bound η on Massart noise		$\widetilde{O}(\text{poly}(p)/\epsilon^3)$	$\text{poly}(p, 1/\epsilon)$
GLM [25] (Kakade et al., 2011)	Best LTF	$\widetilde{O}(1/\epsilon^2)$		$\widetilde{O}(m^{3/2}p)$
SIM [25]	Best LTF	*(i)* $\widetilde{O}(p/\epsilon^3)$ *(ii)* $\widetilde{O}(1/\epsilon^4)$		*(i)* $\widetilde{O}(m^{4/3}p)$ *(ii)* $\widetilde{O}(m^{5/4}p)$
Sigmoid-of-linear [as special case of RSMs]	Best LTF	$\widetilde{O}(1/\epsilon^2)$	$\widetilde{O}(1/\epsilon^4)$	$\widetilde{O}(m^{5/4}p)$
Multiclass classification with 0-1 loss (n classes) $[\mathcal{X} \subseteq \mathbb{R}^p, \mathcal{Y} = \widehat{\mathcal{Y}} = [n]]$				
Softmax-of-multilinear [as special case of RSMs]	Best multilinear multiclass classifier	*(i)* $\widetilde{O}(np/\epsilon^2)$ *(ii)* $\widetilde{O}(n^2/\epsilon^2)$	*(i)* $\widetilde{O}(np/\epsilon^4)$ *(ii)* $\widetilde{O}(n^2/\epsilon^4)$	*(i)* $\widetilde{O}(m^{5/4}np)$ *(ii)* $\widetilde{O}(m^{5/4}np)$
Multi-label prediction with Hamming loss (s tags) $[\mathcal{X} \subseteq \mathbb{R}^p, \mathcal{Y} = \widehat{\mathcal{Y}} = \{0,1\}^s]$				
Sigmoid-of-linear marginals [as special case of RSMs]	Best multilinear multi-label prediction model	$\widetilde{O}(s^3/\epsilon^2)$	$\widetilde{O}(s^5/\epsilon^4)$	$\widetilde{O}(m^{5/4}sp)$
Subset ranking with DCG metric (s items, r rating levels) $[\mathcal{X} \subseteq \mathbb{R}^p, \mathcal{Y} = \{0,1,\ldots,r\}^s, \widehat{\mathcal{Y}} = \Pi_s]$				
Sigmoid-of-linear scaled marginal expectations [as special case of RSMs]	Best multilinear subset ranking model	$\widetilde{O}(s^3/\epsilon^2)$	$\widetilde{O}(r^4 s^5/\epsilon^4)$	$\widetilde{O}(m^{5/4}sp)$

Realizable-Statistic Models (RSMs)

Class of data distributions (over labeled examples (X, Y))