# Safe Learning

*On Bayes, PAC-Bayes, MDL, Tsybakov*

*and* ***Empirical Convexity***

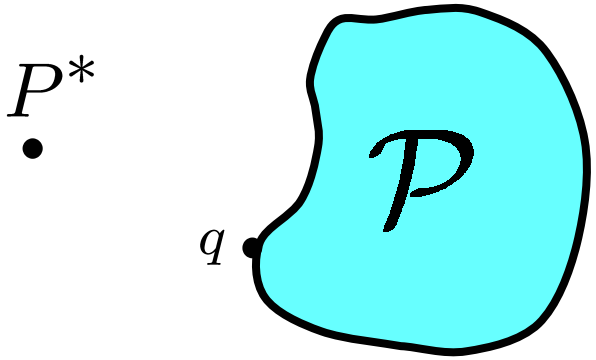

CWI

Peter Grünwald

Universiteit Leiden

Centrum Wiskunde & Informatica – Amsterdam

Mathematical Institute – Leiden University

# Two Seemingly Different Problems

1. Penalization-Based Classification in Learning Theory:
   - Empirical risk penalized by $\sqrt{-\log w(h)/n}$ in worst-case
   - Empirical and Oracle Risk Bounds involve terms of order $\sqrt{1/n}$
   - **Q**: In what situations can we **root out the square root**?

# Two Seemingly Different Problems

1. Penalization-Based Classification in Learning Theory:
   - Empirical risk penalized by $\sqrt{-\log w(h)/n}$ in worst-case
   - Empirical and Oracle Risk Bounds involve terms of order $\sqrt{1/n}$

   **Q**: In what situations can we **root out the square root**?
2. Bayesian inference: if model is "correct", then Bayes (and MDL) converge at fast rates. If model is incorrect, they can **fail** (G & Langford, '04, '07).

   **Q**: Can we design a generalization of MDL/Bayes that is **safe**, i.e. also performs well if model is incorrect?

# Two Seemingly Different Problems

**Q1**: In what situation can we **root out the square root**?

**Q2**: Can we design a generalization of MDL/Bayes that is **safe**, i.e. that also performs well if model is incorrect?

We find such a generalization of MDL/Bayes. When applied to classification problems it achieves Tsybakov-optimal rates and new empirical bounds.

Main Idea: models can be wrong in good and bad ways, we can tell from the data by a "convexity" test whether we are in the 'bad' case, and if we are, we can adjust priors/codelengths

# Basic 2-part code MDL

- Data: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ i.i.d. $\sim P^*$
- Model $\mathcal{P}$ : countable set of conditional densities
- Each $p \in \mathcal{P}$ of form $p(y \mid x)$ , $p(y^n \mid x^n) := \prod_{i=1}^{n} p(y_i \mid x_i)$
- $w$ : prior probability mass function on $\mathcal{P}$
  satisfying $w(p) > 0$, for all $p \in \mathcal{P}$

# Basic 2-part code MDL

- Data: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ i.i.d. $\sim P^*$
- Model $\mathcal{P}$ : countable set of conditional densities
- Each $p \in \mathcal{P}$ of form $p(y \mid x)$ , $p(y^n \mid x^n) := \prod_{i=1}^{n} p(y_i \mid x_i)$
- $w$ : prior probability mass function on $\mathcal{P}$ satisfying $w(p) > 0$, for all $p \in \mathcal{P}$
- Define the 2-MDL estimator $\ddot{p}$ as

$$\ddot{p} := \arg\min_{p \in \mathcal{P}} -2 \log w(p) - \log p(Y^n \mid X^n)$$

# Convergence of 2-MDL

- Let $p^*(y \mid x)$ be conditional density of $P^*$
- KL divergence: $D(P^* \| p) := E_{X,Y \sim P^*} \left[ \log \frac{p^*(Y \mid X)}{p(Y \mid X)} \right]$
- Let $q = \arg \min_{p \in \langle \mathcal{P} \rangle} D(P^* \| p)$

# Convergence of 2-MDL

- Let $p^*(y \mid x)$ be conditional density of $P^*$
- KL divergence: $D(P^* \| p) := E_{X,Y \sim P^*}\left[\log \frac{p^*(Y \mid X)}{p(Y \mid X)}\right]$
- Let $q = \arg\min_{p \in \langle \mathcal{P} \rangle} D(P^* \| p)$

- "Theorem" (Barron & Cover '91, Zhang '06): if $q = p^*$ ("model correct"!)
  then with $P^*$- probability 1, as $n \to \infty$ ,

$$D(P^* \| \dot{\dot{p}}) \longrightarrow D(P^* \| q) = 0$$

# Convergence of 2-MDL

- Let $p^*(y \mid x)$ be conditional density of $P^*$
- KL divergence: $D(P^* \| p) := E_{X,Y \sim P^*}\left[\log \frac{p^*(Y \mid X)}{p(Y \mid X)}\right]$
- Let $q = \arg\min_{p \in \langle \mathcal{P} \rangle} D(P^* \| p)$

- "Theorem" (Barron & Cover '91, Zhang '06): if $q = p^*$ **OR** **"model "convex"** (Li'99) then with $P^*$- probability 1, as $n \to \infty$ ,

$$D(P^* \| \ddot{p}) \longrightarrow D(P^* \| q) \geq 0$$

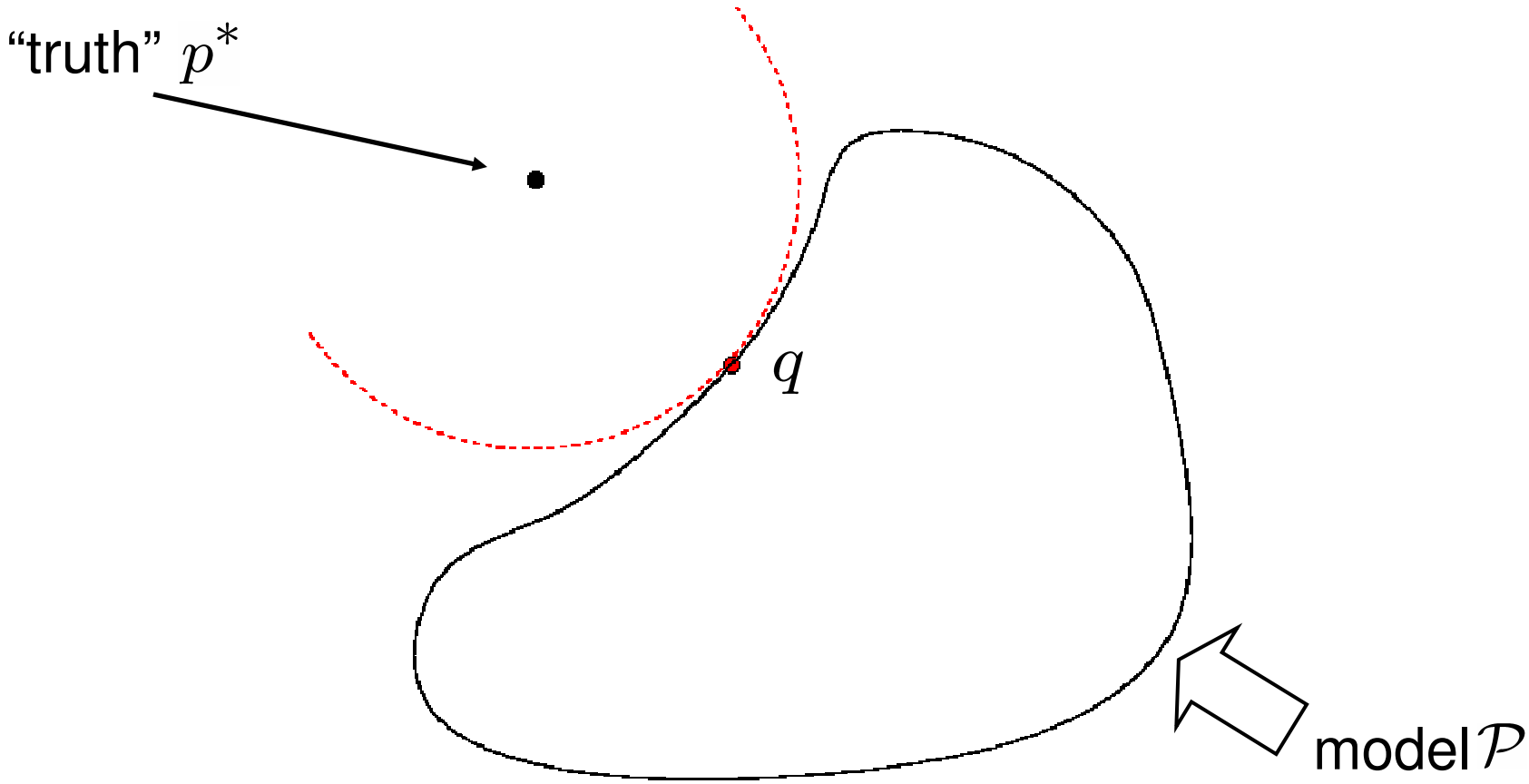
# First Insight

- We know: if model is wrong, then 2-MDL still converges if model is “convex”
- Then 2-MDL estimator must also converge if we have

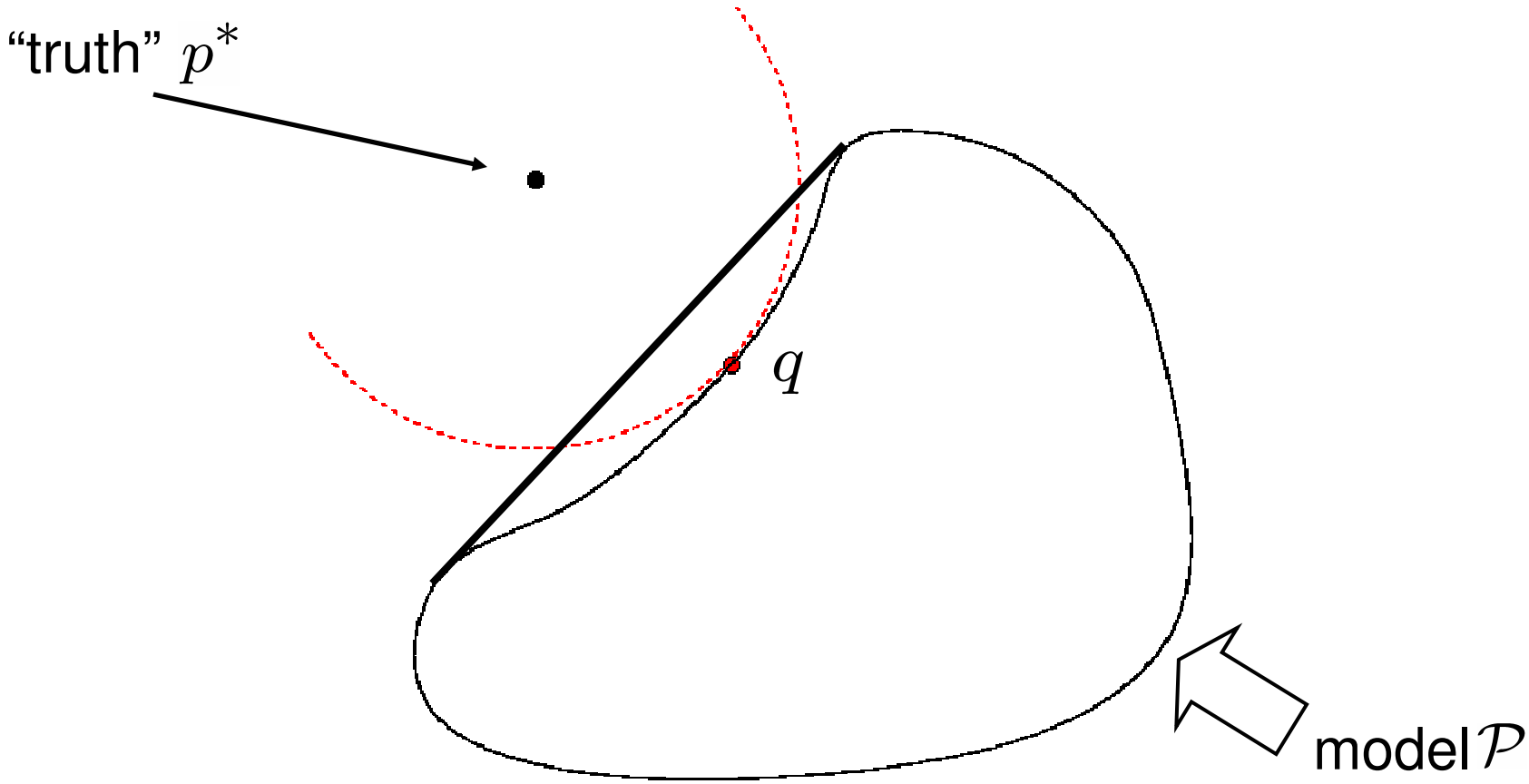
$$\inf_{q \in \mathcal{P}} D(P^* \| q) = \inf_{q \in \text{conv-hull}(\mathcal{P})} D(P^* \| q)$$

i.e. if model cannot be improved by taking its convex hull

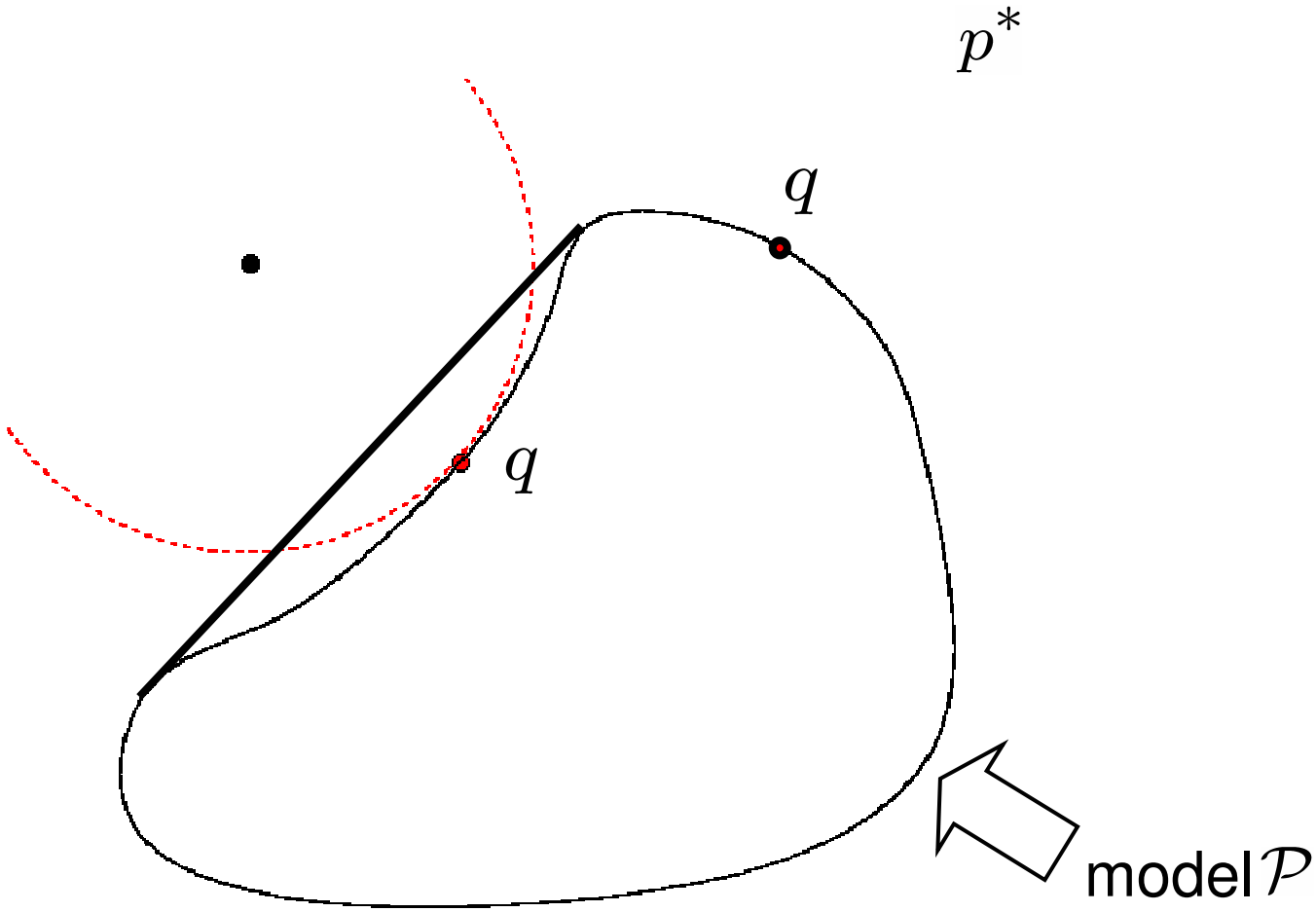
# **Bad** and Good Misspecification



# **Bad** and Good Misspecification



# Bad and Good Misspecification



# Can we test (tell from the data) wether we are in the bad situation?

First idea: see whether an element of convex hull of $\mathcal{P}$ beter compresses the data, i.e. is

$$\min_{p \in \mathrm{conv}(\mathcal{P})} -\log p(Y^n \mid X^n) - 2\log w'(p)$$
$$\ll -\log \ddot{p}(y^n \mid x^n) - 2\log w(\ddot{p})$$

for some prior $w$' on convex hull of $\mathcal{P}$ ?

This may be o.k. but it turns out to be overkill

# YES: we can test wether it's bad!

Second idea: see whether a special two-component mixture beter compresses the data, i.e. is

$$\min_{p \in \text{mini-conv}(\mathcal{P}, X^n, Y^n)} -\log p(Y^n \mid X^n) - 2\log w'(p) \ll -\log \ddot{p}(y^n \mid x^n) - 2\log w(\ddot{p})$$

where $\text{mini-conv}(\mathcal{P}, X^n, Y^n)$ is set of all distributions of form

$$p(Y \mid X) = \alpha \ddot{p}(Y \mid X) + (1-\alpha) p^\circ(Y \mid X)$$

with

$$\alpha \in \{0, 1/n, 2/n, \ldots, 1\}, p^\circ \in \mathcal{P}, w(p^\circ) \geq w(\ddot{p})$$

# Can we adjust model or priors to "repair" situation?

- Define $\mathcal{P}^{(\eta)} := \left\{ \frac{1}{Z(\eta)} p^{\eta} \mid p \in \mathcal{P} \right\}$
- Proposition: there exists $\eta_{\text{crit}} > 0$ s.t. for $0 \leq \eta \leq \eta_{\text{crit}}$ :

$$\inf_{q \in \mathcal{P}^{(\eta)}} D(P^* \| q) = \inf_{q \in \text{conv}(\mathcal{P}^{(\eta)})} D(P^* \| q)$$

**The situation becomes good for** $\eta \leq \eta_{\text{crit}}$

# Can we adjust model or priors to "repair" situation?

- Define $\mathcal{P}^{(\eta)} := \left\{ \frac{1}{Z(\eta)} p^\eta \mid p \in \mathcal{P} \right\}$
- Proposition: there exists $\eta_{\text{crit}} > 0$ s.t. for $0 \leq \eta \leq \eta_{\text{crit}}$ :

$$\inf_{q \in \mathcal{P}^{(\eta)}} D(P^* \| q) = \inf_{q \in \text{conv}(\mathcal{P}^{(\eta)})} D(P^* \| q)$$

- The $2k$-MDL estimator for $\mathcal{P}$ is defined as

$$\arg\min_{p \in \mathcal{P}} -2k \log w(p) - \log p(Y^n \mid X^n)$$

- Proposition: $p$ is the $2/\eta$-MDL estimator for $\mathcal{P}$ iff $\frac{1}{Z(\eta)} p^\eta$ is the 2-MDL estimator for $\mathcal{P}^{(\eta)}$

# YES: we can adjust models/priors to bad situation!

For all $0 < \eta < 1$ see whether special two-component mixture beter compresses the data, i.e. is

$$\min_{p \in \text{mini-conv}(\mathcal{P}^{(\eta)}, X^n, Y^n)} -\log p(Y^n \mid X^n) - 2\log w'(p)$$
$$\ll \min_{p \in \mathcal{P}^{(\eta)}} -\log p(Y^n \mid X^n) - 2\log w(p)$$

where $\text{mini-conv}(\mathcal{P}, X^n, Y^n)$ is as before

safe estimator $\approx (2/\eta)$-MDL estimator for largest $\eta$ for which inequality becomes equality!

# Safe Estimation

- "Theorem 0" (trivial extension of BC '91, Z '06, G' 07):
  Suppose model correct or convex.
  Uniformly for all $0 < \eta \leq 1$, all $K$, with $P^*$ -probability at least $1 - e^{-K}$:

  $$D^*(q \| \ddot{p}_{2/\eta}) \leq \frac{C}{n} \text{ (redundancy} + K)$$

  here $D^*(q\|p) := E_{X,Y\sim P^*}\left[-\log \frac{p(Y \mid X)}{q(Y \mid X)}\right]$

  $$\text{redundancy} = -\frac{2}{\eta}\log w(\ddot{p}_{2/\eta}) - \log \ddot{p}_{2/\eta}(Y^n \mid X^n) - [-\log q(Y^n \mid X^n)]$$

- "Theorem 0"

  Suppose model correct or convex.

  Uniformly for all $0 < \eta \leq 1$, with high probability:

  $$D^*(q\|\ddot{p}_{2/\eta}) \leq \frac{C}{n} \text{ (redundancy} + K)$$

  here $D^*(q\|p) := E_{X,Y\sim P^*}\left[-\log \frac{p(Y \mid X)}{q(Y \mid X)}\right]$

  $$\text{redundancy} = -\frac{2}{\eta}\log w(\ddot{p}_{2/\eta}) - \log \ddot{p}_{2/\eta}(Y^n \mid X^n) - [-\log q(Y^n \mid X^n)]$$

- "Theorem 0"

  Suppose model correct or convex.

  Uniformly for all $0 < \eta \leq 1$, with high probability:

$$D^*(q \| \ddot{p}_{2/\eta}) \leq \frac{C}{n} \text{ (redundancy} + K)$$

  here $D^*(q\|p) := E_{X,Y\sim P^*}\left[-\log \frac{p(Y \mid X)}{q(Y \mid X)}\right]$

$$\text{redundancy} = -\frac{2}{\eta}\log w(\ddot{p}_{2/\eta}) - \log \ddot{p}_{2/\eta}(Y^n \mid X^n) - [-\log q(Y^n \mid X^n)]$$

  nr of extra bits needed to encode data compared to best $q \in \mathcal{P}$

By choosing prior $w$ cleverly can make this logarithmic (parametric case) or $n^\gamma, \gamma \ll 1$ (nonparametric case) – minimax optimal convergence rates can be achieved!

- "Theorem 0"

  Suppose model correct or convex.

  Uniformly for all $0 < \eta \leq 1$, with high probability:

$$D^*(q \| \ddot{p}_{2/\eta}) \leq \frac{C}{n} \text{ (redundancy} + K)$$

  here $D^*(q\|p) := E_{X,Y\sim P^*}\left[-\log \frac{p(Y \mid X)}{q(Y \mid X)}\right]$

$$\text{redundancy} = -\frac{2}{\eta}\log w(\ddot{p}_{2/\eta}) - \log \ddot{p}_{2/\eta}(Y^n \mid X^n) - [-\log q(Y^n \mid X^n)]$$

nr of extra bits needed to encode data compared to best $q \in \mathcal{P}$

# Main Result

- "Theorem 1"

~~Suppose model correct or convex.~~

Uniformly for all $0 < \eta \leq 1$, with high probability:

$$D^*(q \| \ddot{p}_{2/\eta}) \leq \frac{C}{n} \text{ (redundancy + conv-lack + } K)$$

here

$$\text{conv-lack} = \frac{C'}{\eta}\left( \min_{p \in \mathcal{P}^{(\eta)}} -2 \log w(p) - \log p(Y^n \mid X^n) - \min_{p \in \text{mini-conv}(\mathcal{P}^{(\eta)})} [-2 \log w'(p) - \log p(Y^n \mid X^n)] \right)$$

# Main Result

- "Theorem 1"

~~Suppose model correct or convex.~~

Uniformly for all $0 < \eta \leq 1$, with high probability:

$$D^*(q \| \ddot{p}_{2/\eta}) \leq \frac{C}{n} \text{ (redundancy + conv-lack + } K)$$

here

$$\text{conv-lack} = \frac{C'}{\eta}\left( \min_{p \in \mathcal{P}^{(\eta)}} -2 \log w(p) - \log p(Y^n \mid X^n) - \min_{p \in \text{mini-conv}(\mathcal{P}^{(\eta)})} [-2 \log w'(p) - \log p(Y^n \mid X^n)] \right)$$

decreasing in $\eta$

increasing in $\eta$

# Main Result

- "Theorem **1**"

~~Suppose model correct or convex.~~

Uniformly for all $0 < \eta \leq 1$, with high probability:

$$D^*(q \| \ddot{p}_{2/\eta}) \leq \frac{C}{n} \left(\text{redundancy} + \text{conv-lack} + K\right) \quad \textbf{(*)}$$

here

$$\text{conv-lack} = \frac{C'}{\eta}\left( \min_{p \in \mathcal{P}^{(\eta)}} -2 \log w(p) - \log p(Y^n \mid X^n) - \min_{p \in \text{mini-conv}(\mathcal{P}^{(\eta)})} [-2 \log w'(p) - \log p(Y^n \mid X^n)] \right)$$

**Safe** estimator defined as:

$(2/\eta)$ -MDL estimator, with $\eta$ chosen to minimize (*)

# Second Result: What Actually Happens

- "Theorem 1"

  Uniformly for all $0 < \eta \leq 1$, with high probability:

  $$D^*(q \| \ddot{p}_{2/\eta}) \leq \frac{C}{n} \text{ (redundancy + conv-lack + } K)$$

- "Theorem 3": with high probability, for all $\eta \leq \eta_{\text{crit}}$ :

  $$\text{conv-lack} \leq C'' \text{redundancy} + K$$

This implies that if model is correct or convex, ( $\eta_{\text{crit}} = 1$) you **converge at same fast rate** as with standard MDL.

# Classification!

- "Theorem 1" : WHP $D^*(q\|\ddot{p}_{2/\eta}) \leq \frac{C}{n}$ (redundancy + conv-lack + $K$)

- "Theorem 3": WHP, for $\eta \leq \eta_{\text{crit}}$ : conv-lack $\leq C''$redundancy

- Now assume you do classification:
- Countable set of classifiers $\mathcal{H} : \mathcal{X} \to \{0, 1\}$
- Fix some $\beta$ and map $\mathcal{H}$ to set of distr. $\mathcal{P} := \{P_{h,\beta} \mid h \in \mathcal{H}\}$ by standard transformation

$$p_{h,\beta}(y \mid x) := \frac{1}{Z(\beta)} e^{-\beta \text{loss}_{01}(y,h(x))}$$

# Classification!

- "Theorem 1" : WHP $D^*(q\|\ddot{p}_{2/\eta}) \leq \frac{C}{n}$ (redundancy + conv-lack + $K$)

$\beta E_{P^*}[\text{loss}_{01}(Y, \ddot{h}_{2/\eta}(X)) - \text{loss}_{01}(Y, h_{\text{opt}}(X))] \leq \frac{C}{n}$ (redundancy + conv-lack + $K$)

- "Theorem 3": WHP, for $\eta \leq \eta_{\text{crit}}$ : conv-lack $\leq C''$redundancy

- Now assume you do classification:
- Countable set of classifiers $\mathcal{H} : \mathcal{X} \to \{0, 1\}$
- Fix some $\beta$ and map $\mathcal{H}$ to set of distr. $\mathcal{P} := \{P_{h,\beta} \mid h \in \mathcal{H}\}$ by standard transformation

$$p_{h,\beta}(y \mid x) := \frac{1}{Z(\beta)} e^{-\beta \text{loss}_{01}(y, h(x))}$$

$\mathcal{P}$ may be very wrong but we don't care!

# Classification!

- "Theorem 1" : WHP

$$\beta E_{P^*}[\text{loss}_{01}(Y, \ddot{h}_{2/\eta}(X)) - \text{loss}_{01}(Y, h_{\text{opt}}(X))] \leq \frac{C}{n} \text{ (redundancy + conv-lack + } K)$$

- "Theorem 3": WHP, for $\eta \leq \eta_{\text{crit}}$ : conv-lack $\leq C''$redundancy

- Can show: for *all* choices of $P^*, \mathcal{H}$ : $\eta_{\text{crit}} \geq 1/\sqrt{n}$
- With $\eta = 1/\sqrt{n}$ we have RHS in Thm 1 bounded by

$$O\left(\frac{-\log w(\ddot{h})}{\sqrt{n}}\right)$$

optimal!

By choosing $\eta$ slightly larger it becomes $O\left(\sqrt{\frac{-\log w(\ddot{h})}{n}}\right)$

# Final Remarks

- We addressed a number of issues in one fell swoop:
  1 MDL and Bayes when model is wrong
  2 connection between Tsybakov and convexity
  3 A Misspecification Test for (MAP) Bayesian Inference (see Gelman & Shalizi '10, Dawid 82)
  4 Validation of Rissanen's "Universal Yardstick"
  5 Theorem 2: new variation of McAllester-type empirical PAC-Bayes bound