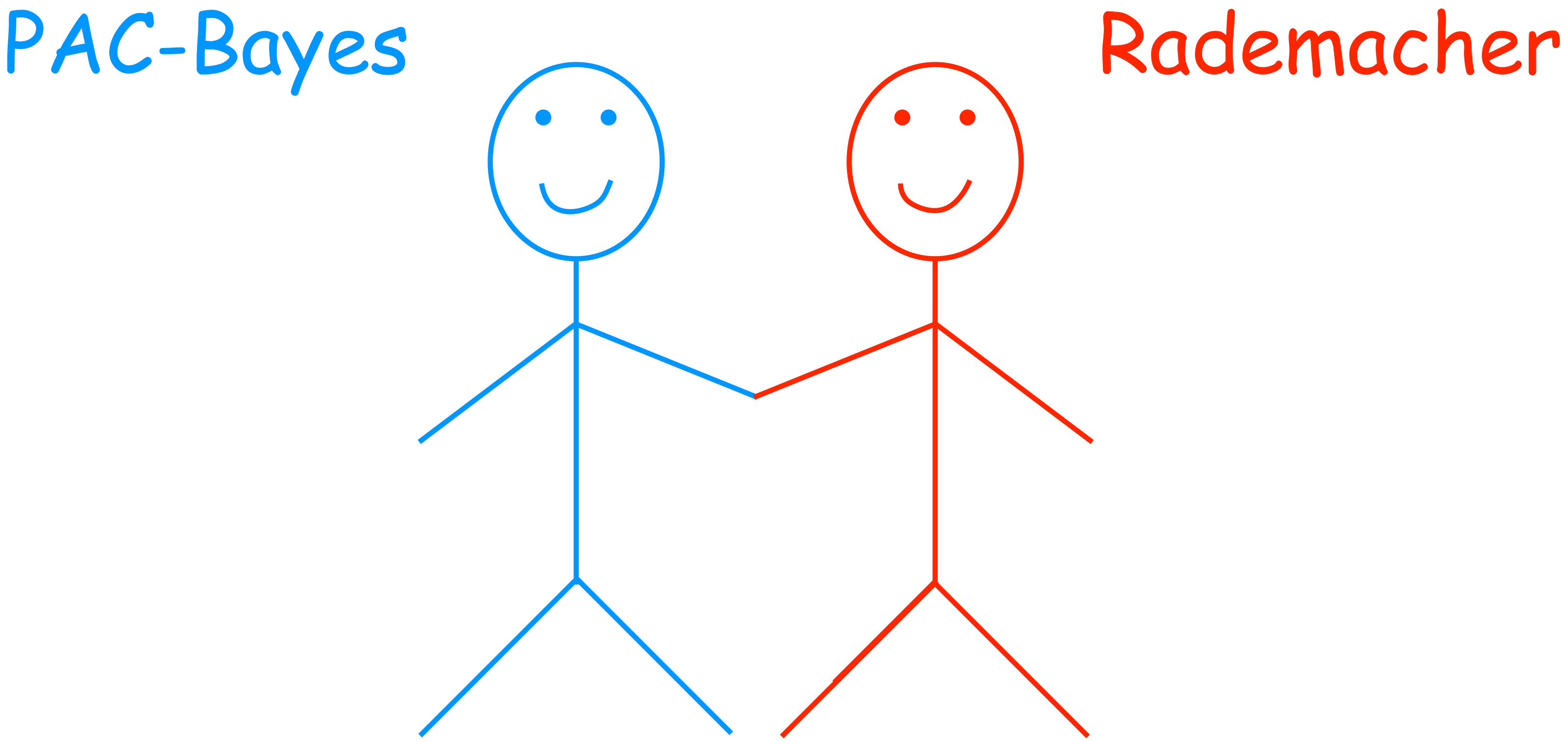
A Tight Excess Risk Bound via a Unified PAC-Bayesian–Rademacher–Shtarkov–MDL Complexity



Peter Grünwald

CWI

Universiteit Leiden

Nishant Mehta

University of Victoria

Overview

Grand goal:

Recover empirical process-style (Rademacher) bounds and PAC-Bayesian bounds using a single theoretical framework and from a **single type of complexity**

(excess risk bounds in i.i.d. statistical learning)

By-products:

New results for individual sequence prediction with large models!

Overview

Grand goal:

Recover empirical process-style (Rademacher) bounds and PAC-Bayesian bounds using a single theoretical framework and from a **single type of complexity**

(excess risk bounds in i.i.d. statistical learning)

By-products:

New results for individual sequence prediction with large models!

Key player: **NML complexity**

Bears similarity to Rademacher complexity

Enjoys tight connection to PAC-Bayesian bounds

The standard log-loss game

Protocol:

For rounds $t = 1, 2, \ldots, n$

1) Learner plays probability distribution Q_t , conditional on $y_1, \ldots, y_{t-1}$
2) Nature plays y_t
3) Learner suffers loss $-\log Q_t(y_t)$

Equivalently, Learner plays joint distribution at very start:

$$Q(y^n) = \prod_{t=1}^{n} Q_t(y_t \mid y_1, \ldots, y_{t-1})$$

The standard log-loss game

Learner plays joint distribution at very start:

$$Q(y^n) = \prod_{t=1}^{n} Q_t(y_t \mid y_1, \ldots, y_{t-1})$$

Learner seeks low worst-case regret relative to set of strategies $\{P_\theta : \theta \in \Theta\}$

$$\sup_{y^n} \left\{ -\log Q(y^n) - \underbrace{\inf_{\theta \in \Theta} \{-\log P_\theta(y^n)\}}_{-\log P_{\hat{\theta}_{|y^n}}(y^n)} \right\}$$

maximum likelihood estimator!

Normalized Maximum Likelihood distribution

Minimax optimal distribution?

$$P_{\mathrm{NML}}(y^n) = \frac{P_{\hat{\theta}_{|y^n}}(y^n)}{\int p_{\hat{\theta}_{|x^n}}(x^n) d\nu(x^n)}$$

Normalized Maximum Likelihood distribution

Minimax optimal distribution?

$$P_{\text{NML}}(y^n) = \frac{P_{\hat{\theta}_{|y^n}}(y^n)}{\int p_{\hat{\theta}_{|x^n}}(x^n) d\nu(x^n)}$$

Normalized Maximum Likelihood (NML) distribution

AKA Shtarkov distribution

(**Shtarkov, 1988**; Rissanen, 1996; Grünwald, 2007)

NML is minimax optimal

Against *every* sequence y^n, NML distribution obtains regret

$$\log \underbrace{\int p_{\hat{\theta}|x^n}(x^n)\,d\nu(x^n)}_{\text{Shtarkov integral}}$$

P_{NML} is an equalizer strategy!

NML is minimax optimal

Against *every* sequence y^n, NML distribution obtains regret

$$\log \underbrace{\int p_{\hat{\theta}|x^n}(x^n)\,d\nu(x^n)}_{\text{Shtarkov integral}}$$

“NML complexity“

P_{NML} is an equalizer strategy!

First Main Result:
Bounding minimax regret for large classes

Let $\mathcal{F}$ be class of probability densities with:

- $\log \mathcal{N}(\mathcal{F}, \varepsilon, L_2(P_n)) \leq \left(\frac{A}{\varepsilon}\right)^{2\rho}$ (polynomial empirical L_2 entropy)
- all densities uniformly lower bounded by $c > 0$

Then minimax individual sequence regret is at most

$$O\left(n^{\frac{\rho}{1+\rho}}\right)$$

Previous results in this setting required bounded L_∞ entropy!

(Opper & Haussler, 1999)

(Cesa-Bianchi & Lugosi, 2001; Rakhlin & Sridharan, 2015)

Example: Monotone densities

Let $\mathcal{P}$ be class of monotone probability densities on $[0, 1]$ with all densities uniformly lower bounded by $c > 0$

Then empirical L_2 entropy is $O\left(\frac{1}{\varepsilon}\right)$

$$\Downarrow$$

Minimax individual sequence regret is $O(n^{1/3})$

Example: Monotone densities

Let $\mathcal{P}$ be class of monotone probability densities on $[0,1]$ with all densities uniformly lower bounded by $c > 0$

Then empirical L_2 entropy is $O\left(\frac{1}{\varepsilon}\right)$

$$\Downarrow$$

Minimax individual sequence regret is $O(n^{1/3})$

L_∞ entropy of $\mathcal{P}$ is unbounded!

Previous results based on L_∞ entropy, cannot bound regret

But how can we use this for statistical learning?

And how is this useful for general classes of hypotheses
and general loss functions?

Towards more general results

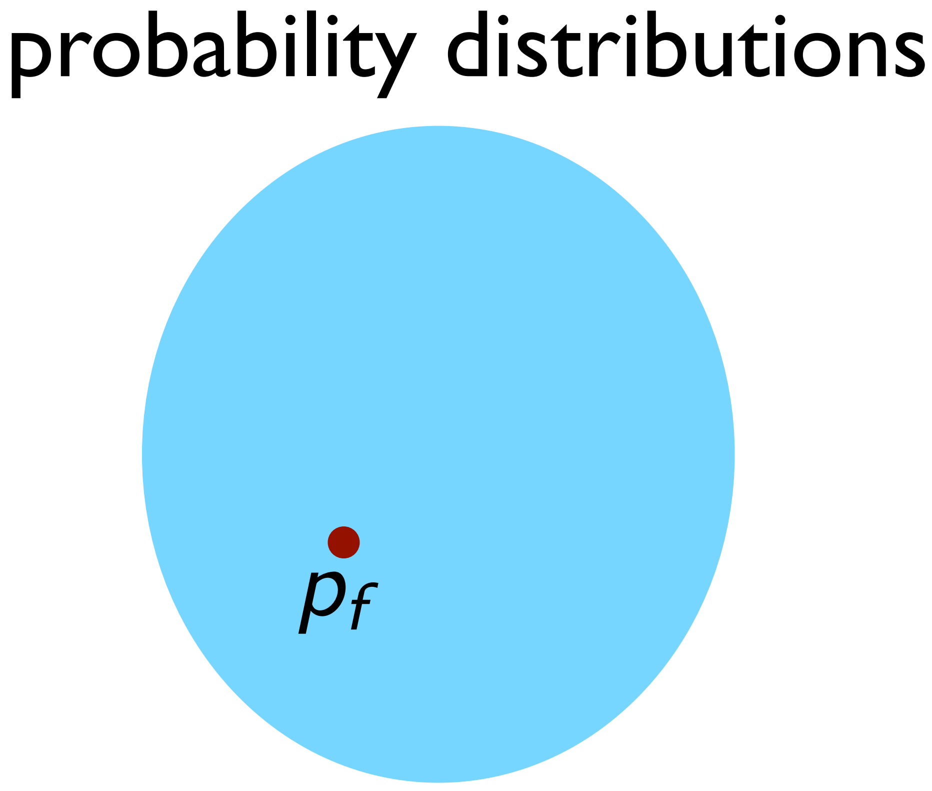
So far, results for density estimation with log loss

Can all this be made to work more generally?

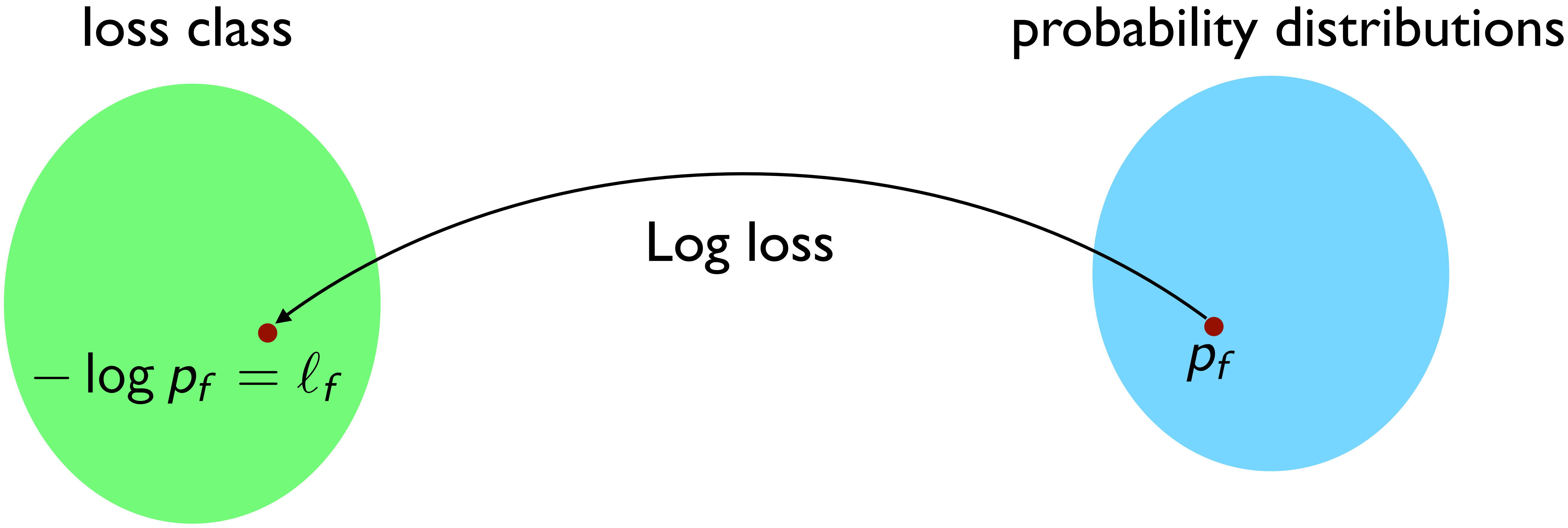
Skeptic: Seems unlikely. The derivation was fundamentally linked to log loss and the Shtarkov integral.

Optimist: We can generalize the concept of the Shtarkov integral using an idea called **entropification**.

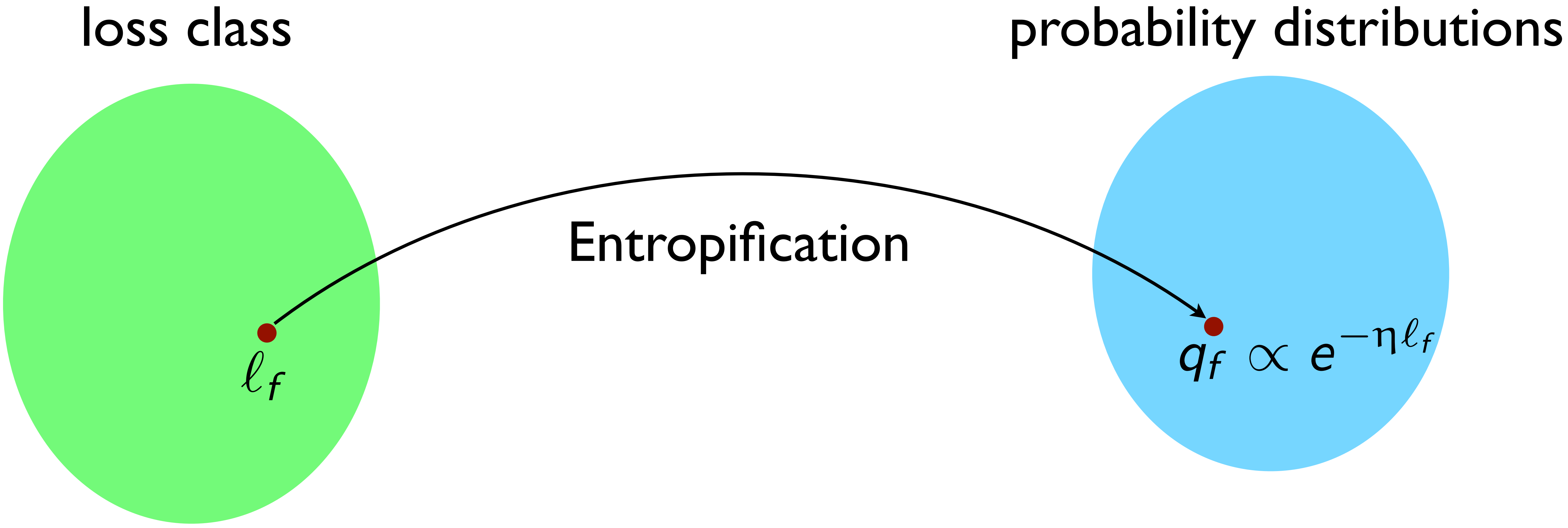
The key transformation: Entropification



The key transformation: Entropification



The key transformation: Entropification



Normalized Maximum Likelihood

	well-specified density estimation with log loss
loss	$-\log p_\theta(z^n)$
excess loss	$-\log \dfrac{p_\theta(z^n)}{p(z^n)}$
probability density	$p_\theta(z^n)$
Shtarkov integral	$\int p_{\hat{\theta}_{\mid z^n}}(z^n)\,d\nu(z^n)$

Generalized Normalized Maximum Likelihood

	well-specified density estimation with log loss	general statistical learning
loss	$-\log p_\theta(z^n)$	$\ell_f(z^n)$
excess loss	$-\log \frac{p_\theta(z^n)}{p(z^n)}$	$R_f(z^n) = \ell_f(z^n) - \ell_{f^*}(z^n)$
probability density	$p_\theta(z^n)$	$q_f(z^n) = \frac{p(z^n) \cdot e^{-\eta R_f(z^n)}}{\mathrm{E}_{\bar{Z}^n \sim P}\left[e^{-\eta R_f(\bar{Z}^n)}\right]}$
Shtarkov integral	$\int p_{\hat{\theta}_{\lvert z^n}}(z^n) d\nu(z^n)$	$\int q_{\hat{f}_{\lvert z^n}}(z^n) d\nu(z^n)$

Generalized NML Complexity

Generalized Shtarkov integral:

$$S(\mathcal{F}, \hat{f}) = \int q_{\hat{f}_{|z^n}}(z^n) d\nu(z^n) = \mathrm{E}_{Z^n \sim P}\left[\frac{e^{-\eta R_{\hat{f}_{|Z^n}}(Z^n)}}{C(\hat{f}_{|Z^n})} \right]$$

normalizer

Generalized NML Complexity

Generalized Shtarkov integral:

$$\mathsf{S}(\mathcal{F}, \hat{f}) = \int q_{\hat{f}_{|z^n}}(z^n)\, d\nu(z^n) = \mathrm{E}_{Z^n \sim P}\left[\frac{e^{-\eta R_{\hat{f}_{|Z^n}}(Z^n)}}{C(\hat{f}_{|Z^n})}\right]$$

normalizer

Generalized NML complexity:

$$\textsc{comp}(\mathcal{F}, \hat{f}) = \frac{1}{\eta} \log \mathsf{S}(\mathcal{F}, \hat{f})$$

Generalized NML Complexity

Generalized Shtarkov integral:

$$\mathsf{S}(\mathcal{F}, \hat{f}) = \int q_{\hat{f}_{|z^n}}(z^n) d\nu(z^n) = \mathsf{E}_{Z^n \sim P}\left[\frac{e^{-\eta R_{\hat{f}_{|Z^n}}(Z^n)}}{C(\hat{f}_{|Z^n})} \right]$$

normalizer

Generalized NML complexity:

$$\textsc{comp}(\mathcal{F}, \hat{f}) = \frac{1}{\eta} \log \mathsf{S}(\mathcal{F}, \hat{f})$$

Later: extended to data-dependent complexity based on "luckiness" function

First risk bound - ERM case

Under the central condition, with probability at least $1 - \delta$

$$\mathsf{E}_{Z \sim P}\left[R_{\hat{f}}(Z)\right] \lesssim \frac{1}{n}\mathrm{COMP}_{\eta/2}(\mathcal{F}, \hat{f}) + \frac{\log\frac{1}{\delta}}{\eta n}$$

First risk bound - ERM case

Under the central condition, with probability at least $1-\delta$

$$\mathrm{E}_{Z\sim P}\left[R_{\hat{f}}(Z)\right] \lesssim \frac{1}{n}\mathrm{COMP}_{\eta/2}(\mathcal{F},\hat{f}) + \frac{\log\frac{1}{\delta}}{\eta n}$$

🤔

First risk bound - ERM case

$$\mathsf{E}\left[e^{-\eta R_f(Z)}\right] \leq 1$$

$$\mathsf{E}\left[R_f^2(Z)\right] \leq C\,\mathsf{E}[R_f(Z)]$$

Bernstein condition

↕

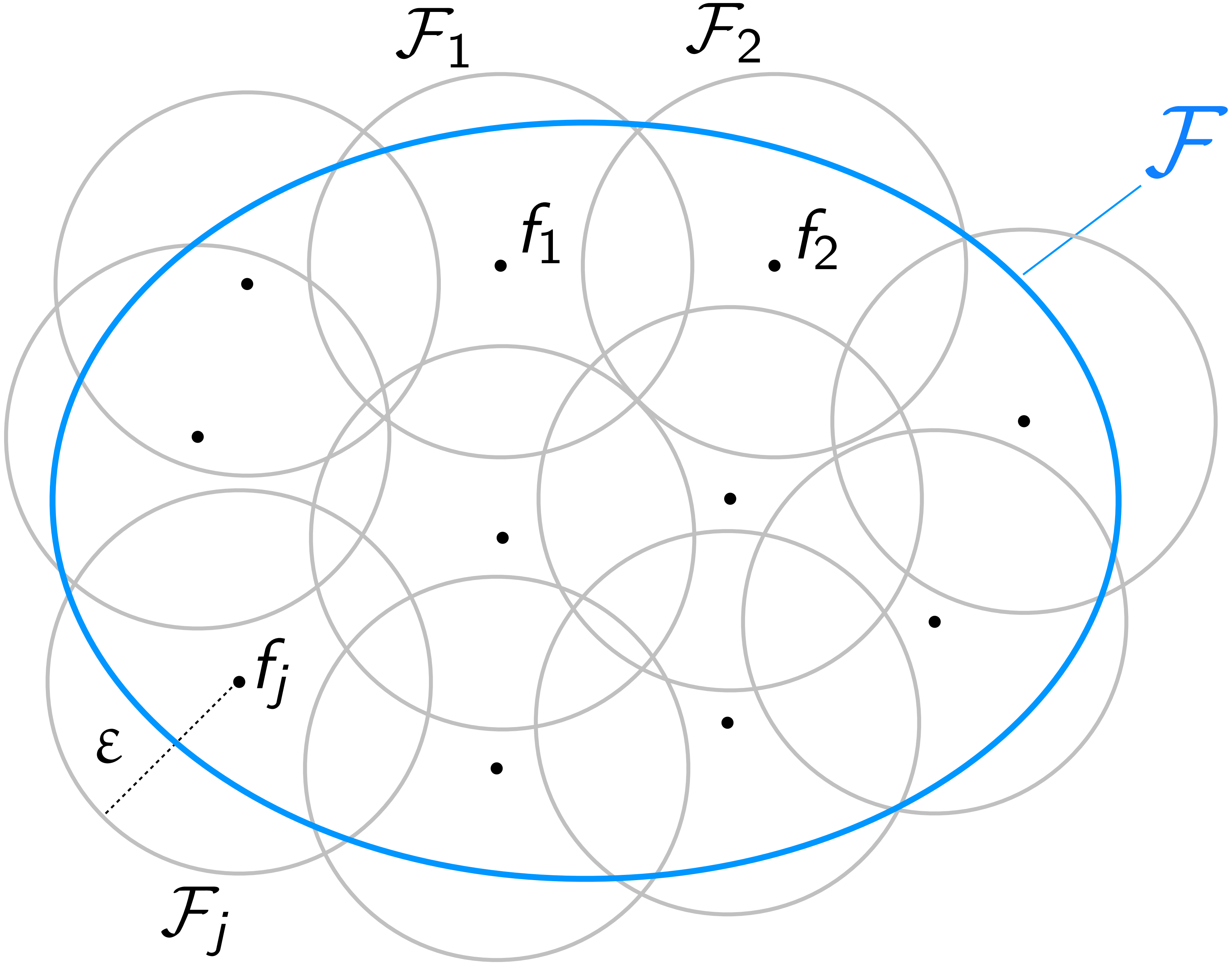
Under the central condition, with probability at least $1-\delta$

$$\mathsf{E}_{Z\sim P}\left[R_{\hat{f}}(Z)\right] \lesssim \frac{1}{n}\mathrm{COMP}_{\eta/2}(\mathcal{F},\hat{f}) + \frac{\log\frac{1}{\delta}}{\eta n}$$

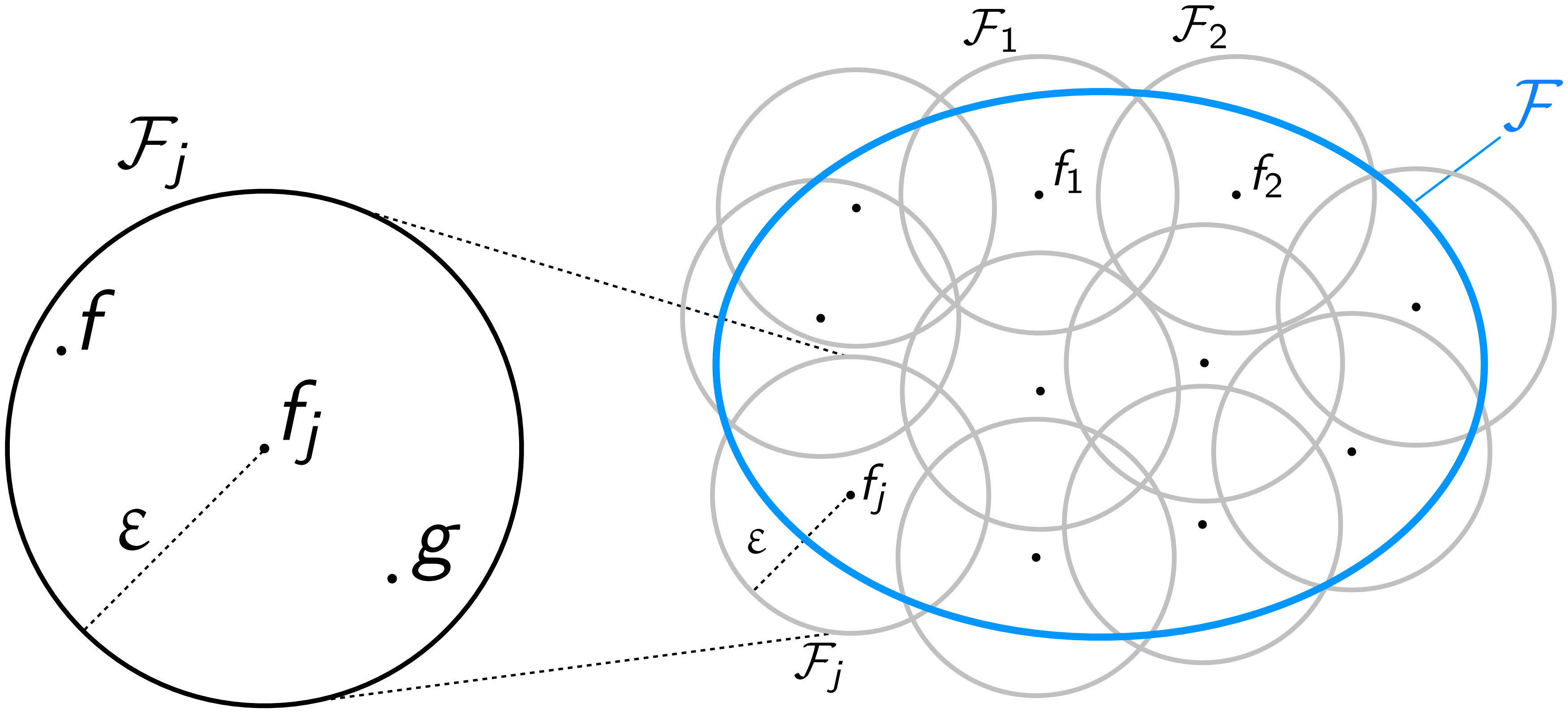
🤔

Decomposing COMP into little COMPs

ε-cover for $\mathcal{F}$ in $L_2(P)$ norm: $\{f_1, f_2, \ldots, f_{N_\varepsilon}\}$



Decomposing COMP into little COMPs



$$\|f - g\|_{L_2(P)} = \mathrm{E}_{X \sim P}\left[(f(X) - g(X))^2\right]^{1/2} \leq \varepsilon$$

Decomposing COMP into little COMPs

$$\textsc{comp}(\mathcal{F}, \hat{f}) \leq \frac{\log N_\varepsilon}{\eta} + \max_{k=1,\ldots,N_\varepsilon} \textsc{comp}(\mathcal{F}_k)$$

(Essentially due to Opper and Haussler (1999))

Bounding COMP

$$\frac{1}{\eta n}\overbrace{\log \mathsf{S}(\mathcal{F}_k)}^{\substack{\mathrm{COMP}_\eta(\mathcal{F}_k)\\ \Big\updownarrow}} \leq \frac{1}{\eta n}\log \mathsf{E}_{Z^n\sim Q_{f_k}}\left[e^{\eta T_n}\right]$$

where T_n is

$$\sup_{f\in\mathcal{F}_k}\underbrace{\left\{\sum_{j=1}^{n}\left(\ell_{f_k}(Z_j)-\ell_f(Z_j)\right)-\mathsf{E}_{Z^n\sim Q_{f_k}}\left[\sum_{j=1}^{n}\left(\ell_{f_k}(Z_j)-\ell_f(Z_j)\right)\right]\right\}}_{\text{centered empirical process}}$$

Bounding COMP

$$\frac{1}{\eta n}\overbrace{\log \mathsf{S}(\mathcal{F}_k)}^{\substack{\mathrm{COMP}_\eta(\mathcal{F}_k)\\ \Updownarrow}} \leq \frac{1}{\eta n}\log \mathsf{E}_{Z^n \sim Q_{f_k}}\left[e^{\eta T_n}\right]$$

$$\lesssim \mathcal{R}_n(\{\ell_{f_K} - \ell_f : f \in \mathcal{F}_k\}) + \eta\varepsilon^2$$

Rademacher complexity!

squared $L_2(P)$ diameter of $\mathcal{F}_k$

Bounding COMP

$$\frac{1}{\eta n}\overbrace{\log \mathsf{S}(\mathcal{F}_k)}^{\mathrm{COMP}_\eta(\mathcal{F}_k)} \leq \frac{1}{\eta n}\log \mathsf{E}_{Z^n \sim Q_{f_k}}\left[e^{\eta T_n}\right]$$

$$\lesssim \mathcal{R}_n(\{\ell_{f_K} - \ell_f : f \in \mathcal{F}_k\}) + \eta\varepsilon^2$$

Rademacher complexity!

squared $L_2(P)$ diameter of $\mathcal{F}_k$

Key techniques:

Talagrand's inequality as a sort of "Reverse Jensen"

Standard use of symmetrization

Back to "big" COMP

Recall that $\mathrm{COMP}(\mathcal{F}, \hat{f}) \leq \frac{\log N_\varepsilon}{\eta} + \max_{k=1,\dots,N_\varepsilon} \mathrm{COMP}(\mathcal{F}_k)$

$$\frac{1}{n}\mathrm{COMP}_\eta(\mathcal{F}) \lesssim \frac{\log N_\varepsilon}{\eta n} + \max_{k=1,\dots,N_\varepsilon} \mathcal{R}_n(\mathcal{G}_k) + \eta\varepsilon^2$$

$\mathcal{G}_k = \{\ell_{f_k} - \ell_f : f \in \mathcal{F}_k\}$

Risk bounds in the best case

$$\log \mathcal{N}(\mathcal{F}, \varepsilon, L_2(P_n)) \leq \left(\frac{A}{\varepsilon}\right)^{2\rho}$$

Large classes, ERM

$$\mathrm{E}_{Z\sim P}\left[R_{\hat{f}}(Z)\right] = O\left(n^{-\frac{1}{1+\rho}} + \frac{\log\frac{1}{\delta}}{n}\right)$$

Rademacher complexity bounded using (Koltchinskii, 2011)

(results for VC classes in the paper)

Intermediate rates in non-best cases

Intermediate Bernstein condition (or Tsybakov margin condition)

$$\mathrm{E}\left[R_f^2(Z)\right] \leq C\,\mathrm{E}[R_f(Z)]^{1/\kappa} \quad \text{for} \quad \kappa \geq 1$$

Large classes, ERM

$$\mathrm{E}_{Z\sim P}\left[R_{\hat{f}}(Z)\right] = O\left(n^{-\frac{\kappa}{2\kappa-1+\rho}} + \frac{\log\frac{1}{\delta}}{n^{\frac{\kappa+\rho}{2\kappa-1+\rho}}}\right)$$

Intermediate rates in non-best cases

Intermediate Bernstein condition (or Tsybakov margin condition)

$$\mathrm{E}\left[R_f^2(Z)\right] \leq C\,\mathrm{E}[R_f(Z)]^{1/\kappa} \quad \text{for} \quad \kappa \geq 1$$

Large classes, ERM

$$\mathrm{E}_{Z\sim P}\left[R_{\hat{f}}(Z)\right] = O\left(n^{-\frac{\kappa}{2\kappa-1+\rho}} + \frac{\log\frac{1}{\delta}}{n^{\frac{\kappa+\rho}{2\kappa-1+\rho}}}\right)$$

THESE ARE THE OPTIMAL RATES
NOT AVAILABLE WITH PAC-BAYESIAN TYPE COMPLEXITY!

(results for VC classes in the paper, again with optimal rates)

Risk bound - General case

For every deterministic estimator $\hat{f}$
and every luckiness function $w \colon \mathcal{Z}^n \to \mathbb{R}_{\geq 0}$

With probability at least $1 - \delta$:

$$\mathrm{E}_{Z \sim P}\left[R_{\hat{f}}(Z)\right] \lesssim \frac{1}{\eta n}\Big(-\log w(Z^n) + \log \mathrm{S}(\mathcal{F}, \hat{f}, w)\Big)$$

data-dependent penalty

w-weighted version of Shtarkov integral
(See paper for details)

Information Complexity bound

NML complexity can be generalized further to handle randomized estimators $\widehat{\Pi}$

Once again, excess risk bounded by generalized complexity with high probability

Generalized complexity can be bounded by information complexity

↕

$$\mathrm{E}_{f \sim \widehat{\Pi}_{|z^n}} \left[\frac{1}{n} \sum_{j=1}^{n} \left(\ell_f(Z_j) - \ell_{f^*}(Z_j) \right) \right] + \frac{\mathrm{KL}(\widehat{\Pi} \,\|\, \Pi)}{\eta \cdot n}$$

The present and future

Thus far:

New bounds on minimax regret for individual sequence prediction with log loss (for large classes)

Single framework that recovers empirical process theory-style bounds and PAC-Bayesian bounds

Luckiness function can allow for interpolations, like bounds for two-part MDL estimators (can talk offline about this)

The future:

Still much more room for sophisticated interpolations via clever choices of luckiness function

Different analyses of generalized log Shtarkov integral when luckiness function is involved