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

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematical Institute – Leiden University

Joint work with Nishant Mehta
University of Victoria

Overview – Problem 1

- PAC-Bayesian Excess Risk bounds give clean data-dependent KL-bounds for randomized predictions with, e.g., generalized Bayes/Gibbs posteriors...

....but also have weaknesses:

  - for ‘large’ classes, rates obtained not minimax optimal (analysis hard to combine with chaining)
  - relation to (more commonly used!) Rademacher complexity bounds is unclear; relation to VC bounds is difficult!

Overview – Problem 2

- The (nonstochastic) minimax cumulative log-loss sequential prediction (codelength) regret, also known as NML or Shtarkov complexity ‘looks’ similar to Rademacher complexity...is there a connection?

Overview

- ...we solve both seemingly entirely different problems in one fell swoop:
- bounding excess risk in terms of a new data-dependent complexity that specializes to PAC-Bayesian KL complexity and/or NML complexity depending on choice of **luckiness function** (generalization of prior) and estimator
- We further bound NML complexity in terms of Rademacher* complexity, and show that this leads to optimal rates even for large classes (Rademacher complexity is amenable to chaining technique...)

...where I come from

- tend to visit both Bayesian and ML conferences
- have been obsessed with Bayes under misspecification, generalized Bayes & Gibbs posteriors, learning rates and the like since around 2011...

Generalized Bayes posteriors

- $\{ p_f : f \in \mathcal{F} \}$ set of densities

$$\pi^B_{n,\eta}(f) := \pi(f \mid Z^n, \eta) \propto \prod_{i=1}^{n} p_f(Z_i)^\eta \cdot \pi_0(f)$$

Generalized and Gibbs posteriors

- $\{ p_f : f \in \mathcal{F} \}$ set of densities

$$\pi^B_{n,\eta}(f) := \pi(f \mid Z^n, \eta) \propto \prod_{i=1}^{n} p_f(Z_i)^\eta \cdot \pi_0(f)$$

- $\mathcal{F}$ set of predictors
- ℓ_f: $\mathcal{Z} \to \mathbb{R}$ loss function for predictor f
- e.g. squared error loss,
- $Z_i = (X_i, Y_i)$; $\ell_f((x,y)) = \big(y - f(x)\big)^2$

$$\pi^B_{n,\eta}(f) := \pi(f \mid Z^n, \eta) \propto \prod_{i=1}^{n} e^{-\eta \ell_f(Z_i)} \cdot \pi_0(f)$$

...where I come from

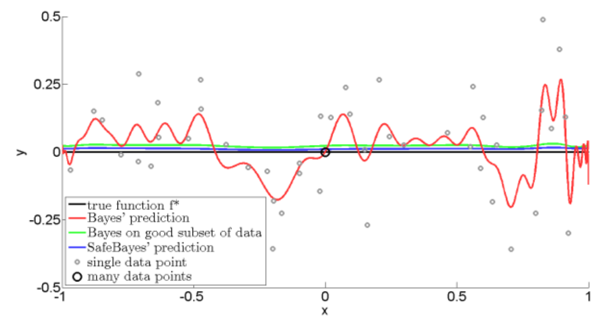
- tend to visit both Bayesian and ML conferences
- Have been obsessed with Bayes under misspecification, generalized η-Bayes & η-Gibbs posteriors, learning rate η since around 2011...
- My earlier work in this direction:
  - The Safe Bayesian (COLT 2011, ALT 2012): learning the learning rate automatically

My earlier work in this direction

- The Safe Bayesian (COLT 2011, ALT 2012): learning the learning rate automatically so that you are consistent under misspecification and...
- $\boldsymbol{\eta = 1}$ (standard Bayes) can be disastrous under misspecification (model wrong but useful)

  G. and van Ommen. Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing it . *Bayesian Analysis*, 2017

  (transplanting G. and Langford '04 to a realistic model)



My earlier work in this direction

- The Safe Bayesian (COLT 2011, ALT 2012): learning the learning rate automatically so that you are consistent under misspecification and...
- $\boldsymbol{\eta = 1}$ (standard Bayes) can be disastrous under misspecification (model wrong but useful)

  G. and van Ommen. *BA* ‘17

  also contains novel interpretation of learning rate

My earlier work in this direction

- The Safe Bayesian (COLT 2011, ALT 2012): learning the learning rate automatically so that you are consistent under misspecification and...
- $\boldsymbol{\eta = 1}$ (standard Bayes) can be disastrous under misspecification (G. and van Ommen, BA 2017)
- Learning η in individual sequence, nonstochastic online setting (various papers with De Rooij, Van Erven, Koolen - JMLR ’14, NIPS ’15 ’16)

This Year

- G. and Mehta, arXiv (2016 and 2017b).
  Fast Rates for General Unbounded Loss Functions: from ERM to Generalized Bayes
- G. and Mehta, arXiv (2017a)
- **A Tight Excess Risk Bound via a Unified PAC-Bayesian-Rademacher-Shtarkov-MDL Complexity**

Both works extend a prevous PAC-Bayes-style excess risk bound due to Tong Zhang (2006a,b) [closely related, partially more general works by Catoni (e.g.'03,'07) & Audibert!]

Zhang's (2004,2006) PAC-Bayes Excess Risk Bound

For every learning algorithm $\hat{\Pi}_n := \hat{\Pi}|Z^n$ that outputs a distribution on model $\mathcal{F}$, every 'prior' Π_0 every $\eta > 0$:

$$\mathbf{E}_{f \sim \hat{\Pi}_n} \mathbf{E}^{\mathrm{ann},\eta}_{Z \sim P} \left[r_f(Z) \right] \trianglelefteq_{\eta n} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[\frac{1}{n} \sum_{i=1}^{n} r_f(Z_i) \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n}$$

- holds for general distribution-output estimators (including deterministic estimators)
- distribution can be, but need not be, a generalized posterior/Gibbs distribution

Zhang's Excess Risk Bound

For every learning algorithm $\widehat{\Pi}_n := \widehat{\Pi}|Z^n$ that outputs a distribution on model $\mathcal{F}$, every 'prior' Π_0 every $\eta > 0$:

$$\mathbf{E}_{f \sim \hat{\Pi}_n} \mathbf{E}^{\mathrm{ann},\eta}_{Z \sim P} \left[r_f(Z) \right] \trianglelefteq_{\eta n} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[\frac{1}{n} \sum_{i=1}^{n} r_f(Z_i) \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n}$$

- G. & Mehta 2016,2017b mostly about extending the left-hand side
- **TODAY: G. & Mehta 2017a ; mostly about the right-hand side**

Zhang's Excess Risk Bound

For every learning algorithm $\widehat{\Pi}_n := \widehat{\Pi}|\mathrm{Z}^{\mathrm{n}}$ that outputs a distribution on model $\mathcal{F}$, every 'prior' Π_0 every $\eta > 0$:

$$\mathbf{E}_{f\sim\hat{\Pi}_n}\ \mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P}\ [r_f(Z)] \trianglelefteq_{\eta n} \mathbf{E}_{f\sim\hat{\Pi}_n}\left[\frac{1}{n}\sum_{i=1}^{n} r_f(Z_i)\right] + \frac{\mathrm{KL}(\hat{\Pi}_n\|\Pi_0)}{\eta\cdot n}$$

Here $\trianglelefteq_{\eta n}$ means inequality holds both in expectation and with very high probability over $Z^n = (Z_1, \ldots, Z_n) = \big((X_1, Y_1,), \ldots, (X_n, Y_n)\big) \sim$ i.i.d. P

$$X \trianglelefteq_\gamma Y \quad \Leftrightarrow \quad \mathbf{E}\left[e^{\gamma(X-Y)}\right] \leq 1$$

Zhang's Excess Risk Bound

For every learning algorithm $\widehat{\Pi}_n := \widehat{\Pi}|\mathrm{Z}^{\mathrm{n}}$ that outputs a distribution on model $\mathcal{F}$, every 'prior' Π_0 every $\eta > 0$:

$$\mathbf{E}_{f \sim \hat{\Pi}_n} \mathbf{E}_{Z \sim P}^{\mathrm{ann},\eta} \left[r_f(Z) \right] \trianglelefteq_{\eta n} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[\frac{1}{n} \sum_{i=1}^{n} r_f(Z_i) \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n}$$

Here $\trianglelefteq_{\eta n}$ means inequality holds both in expectation and with very high probability over
$Z^n = (Z_1, \ldots, Z_n) = \big((X_1, Y_1,), \ldots, (X_n, Y_n)\big) \sim$ i.i.d. P

$$X \trianglelefteq_{\gamma} Y \quad \Leftrightarrow \quad \mathbf{E}\left[e^{\gamma(X-Y)}\right] \leq 1 \Rightarrow \mathbf{E}[X] \leq \mathbf{E}[Y]$$

$$\mathbf{E}\left[e^{\gamma(X-Y)}\right] \leq 1 \Rightarrow P(X \geq Y + a) \leq e^{-\gamma a}$$

Zhang's Excess Risk Bound

For every learning algorithm $\widehat{\Pi}_n := \widehat{\Pi}|\mathrm{Z}^{\mathrm{n}}$ that outputs a distribution on model $\mathcal{F}$, every 'prior' Π_0 every $\eta > 0$:

$$\mathbf{E}_{f\sim\hat{\Pi}_n}\ \mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P}\ [r_f(Z)] \trianglelefteq_{\eta n} \mathbf{E}_{f\sim\hat{\Pi}_n}\left[\frac{1}{n}\sum_{i=1}^{n} r_f(Z_i)\right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n}$$

$r_f(Z) := \ell_f\ (Z) - \ell_{f^*}(Z)$ is excess loss on Z

Zhang's Excess Risk Bound

For every learning algorithm $\widehat{\Pi}_n := \widehat{\Pi}|Z^n$ that outputs a distribution on model $\mathcal{F}$, every 'prior' Π_0 every $\eta > 0$:

$$\mathbf{E}_{f\sim\hat{\Pi}_n} \mathbf{E}_{Z\sim P}^{\text{ann},\eta} \left[r_f(Z)\right] \trianglelefteq_{\eta n} \mathbf{E}_{f\sim\hat{\Pi}_n}\left[\frac{1}{n}\sum_{i=1}^{n} r_f(Z_i)\right] + \frac{\text{KL}(\hat{\Pi}_n\|\Pi_0)}{\eta\cdot n}$$

$r_f(Z) := \ell_f(Z) - \ell_{f^*}(Z)$ is excess loss on Z

ℓ can be any loss function

e.g. $Z = (X,Y),\ \ell_f((X,Y)) = |Y - f(X)|$ (0/1-loss)

$Z = (X,Y),\ \ell_f((X,Y)) = \left(Y - f(X)\right)^2$ (sq. Err. loss)

$\ell_f(Z) = -\log p_f(Z)$ (log loss)

Zhang's Excess Risk Bound

For every learning algorithm $\widehat{\Pi}_n := \widehat{\Pi}|Z^n$ that outputs a distribution on model $\mathcal{F}$, every 'prior' Π_0 every $\eta > 0$:

$$\mathbf{E}_{f\sim\hat{\Pi}_n} \mathbf{E}^{\text{ann},\eta}_{Z\sim P} \left[r_f(Z)\right] \trianglelefteq_{\eta n} \mathbf{E}_{f\sim\hat{\Pi}_n}\left[\frac{1}{n}\sum_{i=1}^{n} r_f(Z_i)\right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n}$$

$r_f(Z) := \ell_f\,(Z) - \ell_{f^*}(Z)$ is excess loss on Z

ℓ can be any loss function (0/1, square, log-loss, ...)

f^* is risk minimizer in $\mathcal{F}$:

$$f^* := \arg\min_{f\in\mathcal{F}} \mathbf{E}_{Z\sim P}[\ell_f(Z)]$$

Zhang's Excess Risk Bound

For every learning algorithm $\widehat{\Pi}_n := \widehat{\Pi}|Z^n$ that outputs a distribution on model $\mathcal{F}$, every 'prior' Π_0 every $\eta > 0$:

$$\mathbf{E}_{f \sim \hat{\Pi}_n} \mathbf{E}^{\mathrm{ann},\eta}_{Z \sim P} \left[r_f(Z) \right] \trianglelefteq_{\eta n} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[\frac{1}{n} \sum_{i=1}^{n} r_f(Z_i) \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n}$$

Special Case of deterministic $\hat{f}$

For every learning algorithm $\hat{f}$ that upon observing Z^n outputs predictor $\hat{f}_{|Z^n}$ in countable subset $\ddot{\mathcal{F}} \subseteq \mathcal{F}$,every 'prior' mass fn π_0 every $\eta > 0$:

$$\cancel{\mathbf{E}_{f\sim\hat{\Pi}_n}}\ \mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P}\left[r_f(Z)\right] \trianglelefteq_{\eta n} \cancel{\mathbf{E}_{f\sim\hat{\Pi}_n}}\left[\frac{1}{n}\sum_{i=1}^{n} r_f(Z_i)\right] + \frac{\mathrm{KL}(\hat{\Pi}_n\|\Pi_0)}{\eta\cdot n}$$

Special Case of deterministic $\hat{f}$

For every learning algorithm $\hat{f}$ that upon observing Z^n outputs predictor $\hat{f}_{|Z^n}$ in countable subset $\ddot{\mathcal{F}} \subseteq \mathcal{F}$,every 'prior' mass fn π_0 every $\eta > 0$:

$$\mathbf{E}^{\text{ann},\eta}_{Z\sim P}\left[r_{\hat{f}_{|Z^n}}(Z)\right] \trianglelefteq_{\eta n} \left[\frac{1}{n}\sum_{i=1}^{n} r_{\hat{f}_{|Z^n}}(Z_i)\right] + \frac{-\log \pi_0(\hat{f}_{|Z^n})}{\eta \cdot n}$$

Here $\trianglelefteq_{\eta n}$ means inequality holds both in expectation and with very high probability over $Z^n \sim$ i.i.d. P

$r_f(Z) \coloneqq \ell_f\,(Z) - \ell_{f^*}(Z)$ is excess loss on Z

Zhang's Excess Risk Bound

$$\underbrace{\mathbf{E}_{f\sim\hat{\Pi}_n}\ \mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P}\ [r_f(Z)]}_{\text{annealed excess risk of draw of } f \text{ according to 'posterior'}} \trianglelefteq_{\eta n} \mathbf{E}_{f\sim\hat{\Pi}_n}\underbrace{\left[\frac{1}{n}\sum_{i=1}^{n} r_f(Z_i)\right]}_{f\text{'s empirical excess risk}} + \underbrace{\frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta\cdot n}}_{\text{data-dependent complexity term}}$$

Zhang's Excess Risk Bound

$$\mathbf{E}_{f\sim\hat{\Pi}_n} \underbrace{\mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P} \; [r_f(Z)]} \trianglelefteq_{\eta n} \mathbf{E}_{f\sim\hat{\Pi}_n}\left[\frac{1}{n}\sum_{i=1}^{n} r_f(Z_i)\right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta\cdot n}$$

annealed excess risk $\mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P}[r_f] \;:= -\frac{1}{\eta}\log \mathbf{E}_{Z\sim P}\left[e^{-\eta r_f(Z)}\right]$

Zhang's Excess Risk Bound

But we are really interested in the **actual** excess risk $\mathbf{E}[r_f]$!

$$\mathbf{E}_{f \sim \hat{\Pi}_n} \ \mathbf{E}^{\text{ann},\eta}_{Z \sim P} \ [r_f(Z)] \trianglelefteq_{\eta n} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[\frac{1}{n} \sum_{i=1}^{n} r_f(Z_i) \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n}$$

annealed excess risk $\mathbf{E}^{\text{ann},\eta}_{Z \sim P}[r_f] := -\frac{1}{\eta} \log \mathbf{E}_{Z \sim P} \left[e^{-\eta r_f(Z)} \right]$

Zhang's Excess Risk Bound

But we are really interested in the **actual** excess risk $\mathbf{E}[r_f]$!

$$\mathbf{E}_{f \sim \hat{\Pi}_n} \underbrace{\mathbf{E}^{\mathrm{ann},\eta}_{Z \sim P} \; [r_f(Z)]}_{} \trianglelefteq_{\eta n} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[\frac{1}{n} \sum_{i=1}^{n} r_f(Z_i) \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n}$$

annealed excess risk $\mathbf{E}^{\mathrm{ann},\eta}_{Z \sim P}[r_f] \; := -\frac{1}{\eta} \log \mathbf{E}_{Z \sim P} \left[e^{-\eta r_f(Z)} \right]$

Annealed excess risk is lower bound on actual excess risk

(can even be negative!)

Indeed with annealed risk result holds completely generally, no further conditions!

Zhang's Excess Risk Bound

But we are really interested in the **actual** excess risk $\mathbf{E}[r_f]$!

$$\mathbf{E}_{f \sim \hat{\Pi}_n} \underbrace{\mathbf{E}^{\mathrm{ann},\eta}_{Z \sim P} \; [r_f(Z)]} \trianglelefteq_{\eta n} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[\frac{1}{n} \sum_{i=1}^{n} r_f(Z_i) \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n}$$

annealed excess risk $\mathbf{E}^{\mathrm{ann},\eta}_{Z \sim P}[r_f] \; := -\frac{1}{\eta} \log \mathbf{E}_{Z \sim P} \left[e^{-\eta r_f(Z)} \right]$

annealed excess risk is lower bound on actual excess risk

but for right choice of η also upper bounds actual excess risk

or Hellinger^2 distance (density estimation) up to constant factor

From Annealed Risk to Hellinger:

- log-loss with well-specified probability model: for any $\eta < 1$ annealed risk larger than constant times Hellinger distance^2 (Zhang '06)
- One retrieves celebrated standard thms on posterior concentration for Bayesian inference by Gosh, Ghosal and Van der Vaart (2000; many follow-up papers) under substantially weaker conditions *as soon as* one uses generalized Bayes with $\eta < 1$

Zhang's Excess Risk Bound

For every learning algorithm $\hat{\Pi}_n := \hat{\Pi}|Z^n$ that outputs a distribution on model $\mathcal{F}$, every 'prior' Π_0 every $\eta > 0$:

$$\mathbf{E}_{f\sim\hat{\Pi}_n}\ \mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P}\ [r_f(Z)] \trianglelefteq_{\eta n} \quad \cdot \left(\mathbf{E}_{f\sim\hat{\Pi}_n}\left[\frac{1}{n}\sum_{i=1}^{n} r_f(Z_i)\right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n} \right)$$

Zhang's Excess Risk Bound

For every 'prior' Π_0 , every $0 < \eta < 1$, for the generalized η-Bayesian posterior, every well-specified probability model $\{ p_f : f \in \mathcal{F} \}$

$$\mathbf{E}_{f \sim \hat{\Pi}_n} \cancel{\mathbf{E}^{\mathrm{ann},\eta}_{Z \sim P} [r_f(Z)]} \trianglelefteq_{\eta n} C_\eta \cdot \left(\mathbf{E}_{f \sim \hat{\Pi}_n} \cancel{\left[\frac{1}{n} \sum_{i=1}^{n} r_f(Z_i) \right]} + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n} \right)$$

$$d_{\mathrm{H}}^2(f^*, f) \qquad\qquad -\frac{1}{n} \cdot \log \frac{p_f(Z^n)}{p_{f^*}(Z^n)}$$

Zhang's Excess Risk Bound

For every 'prior' Π_0 , every $0 < \eta < 1$, for the generalized η-Bayesian posterior, every well-specified probability model $\{ p_f : f \in \mathcal{F} \}$

$$\mathbf{E}_{f\sim\hat{\Pi}_n}\ \cancel{\mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P}\left[{\color{red} r_f(Z)} \right]} \trianglelefteq_{\eta n} {\color{blue} C_\eta} \cdot \left(\mathbf{E}_{f\sim\hat{\Pi}_n} \cancel{\left[\frac{1}{n}\sum_{i=1}^{n} {\color{red} r_f(Z_i)} \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta\cdot n}} \right)$$

$$d^2_{\mathrm{H}}(f^*, f)$$

$$-\frac{1}{\eta\cdot n}\cdot\log\frac{\int_{\mathcal{F}} (\, p_f(Z^n)\,)^{\eta}\, d\Pi_0(f)}{(\, p_{f^*}(Z^n)\,)^{\eta}}$$

Zhang's Excess Risk Bound

For every 'prior' Π_0 , every $0 < \eta < 1$, for the generalized η-Bayesian posterior, every well-specified probability model $\{ p_f : f \in \mathcal{F}\}$

$$\mathbf{E}_{f\sim\hat{\Pi}_n}\ \cancel{\mathbf{E}^{\text{ann},\eta}_{Z\sim P}\left[r_f(Z)\right] \trianglelefteq_{\eta n}}\ C_\eta \cdot \left(\mathbf{E}_{f\sim\hat{\Pi}_n}\left[\cancel{\frac{1}{n}\sum_{i=1}^{n} r_f(Z_i)} \right] + \cancel{\frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta\cdot n}} \right)$$

$$d^2_{\mathrm{H}}(f^*, f) \qquad\qquad -\frac{1}{\eta\cdot n}\cdot \log \frac{\int_{\mathcal{F}} (\, p_f(Z^n)\,)^{\eta}\, d\Pi_0(f)}{(\, p_{f^*}(Z^n)\,)^{\eta}}$$

Retrieve Ghosal, Gosh, VDVaart!

$$\leq^* \inf_{\epsilon \geq 0} \left\{ \epsilon \ + \ \frac{-\log \Pi_0(B_{D_P}(f^*, \epsilon))}{\eta \cdot n} \right\}$$

$$B_{D_P}(f^*, \epsilon) = \{ f \in \mathcal{F} : D_P(f^* \| f) \leq \epsilon \}$$

G&M 2016,2017b

For every ‘prior’ Π_0 , every $0 < \eta < \bar{\eta}$ for the generalized η-Bayesian posterior, where $\bar{\eta}$ is **critical learning rate** for (possibly misspecified) probability model $\{ p_f : f \in \mathcal{F} \}$

$$\mathbf{E}_{f \sim \hat{\Pi}_n} \cancel{\mathbf{E}^{\text{ann},\eta}_{Z \sim P} [r_f(Z)]} \trianglelefteq_{\eta n} C_\eta \cdot \left(\mathbf{E}_{f \sim \hat{\Pi}_n} \left[\cancel{\frac{1}{n} \sum_{i=1}^{n} r_f(Z_i)} \right] \cancel{+ \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n}} \right)$$

$$d^2_{\text{H, generalized}}(f^*, f)$$

$$- \frac{1}{\eta \cdot n} \cdot \log \frac{\int_{\mathcal{F}} (\, p_f(Z^n) \,)^\eta \, d\Pi_0(f)}{(\, p_{f^*}(Z^n) \,)^\eta}$$

$$\leq^* \inf_{\epsilon \geq 0} \left\{ \epsilon \;+\; \frac{- \log \Pi_0(B_{D_P}(f^*, \epsilon))}{\eta \cdot n} \right\}$$

$$B_{D_P}(f^*, \epsilon) = \{ f \in \mathcal{F} : D_P(f^* \| f) \leq \epsilon \}$$

From Annealed to Actual Excess Risk: G&M 2016, 2017b

- log-loss/density estimation: for any $\eta < \bar{\eta}$ annealed risk larger than constant times Hellinger distance^2
- general loss functions: ‘Hellinger’ not too meaningful. Want actual risk on the left

U-Central Condition

Suppose there exists an increasing function $u : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ such that :

$$\forall f \in \mathcal{F}, \epsilon \geq 0 : \quad \ell_{f^*} - \ell_f \trianglelefteq_{u(\epsilon)} \epsilon$$

then we say that the ***u*-central condition** holds.

(Van Erven et al. 2015)

log-loss: if there is a fixed critical $\bar{\eta}$ then u-central holds for the special case with $u \equiv \bar{\eta}$ constant!

Theorem for general u-central

Suppose loss bounded and u-central holds, i.e.

$$\forall f \in \mathcal{F}, \epsilon > 0 : \quad \ell_{f^*} - \ell_f \trianglelefteq_{u(\epsilon)} \epsilon$$

Then (G. & Mehta 2016) there is $C > 0$ such that for every $f \in \mathcal{F}, \epsilon > 0$

$$\mathbf{E}_{Z \sim P}[r_f] \leq C \cdot \left(\mathbf{E}^{\mathrm{ann}, u(\epsilon)}[r_f] + \epsilon \right)$$

C is linear in loss range

Theorem for general u-central

Suppose loss bounded and u-central holds*, i.e.

$$\forall f \in \mathcal{F}, \epsilon > 0 : \quad \ell_{f^*} - \ell_f \trianglelefteq_{u(\epsilon)} \epsilon$$

Then there is $C > 0$ such that for every distribution-output learning algorithm Π_n , every prior Π_0 every $f \in \mathcal{F}, \epsilon > 0$:

$$\mathbf{E}_{f \sim \Pi_n} \mathbf{E}_{Z \sim P} [r_f] \trianglelefteq_{n \cdot u(\epsilon)} C \cdot \left(\mathbf{E}_{f \sim \Pi_n} [r_f(Z^n)] + \frac{\mathrm{KL}(\Pi_n \| \Pi_0)}{u(\epsilon) \cdot n} + \epsilon \right)$$

Proof: simply plug previous result into Zhang!

For bounded loss, u-central with linear u always holds:

Can get $O\left(\sqrt{\mathrm{KL}/n}\right)$ rate

Bernstein, Central

- Bounded losses: for $\beta \in [0,1]$:
- $u(x) \asymp x^{\beta}$ – central equivalent to $(1-\beta)$-**Bernstein** condition (Van Erven et al., 2015):

$$\mathbf{E}_{Z\sim P}[(r_f)^2] \leq C \cdot (\ \mathbf{E}_{Z\sim P}[r_f]\ )^{\beta}$$

- Bernstein condition, a generalization of the **Tsybakov noise condition**, is *the* condition studied in statistical learning theory that allows for fast rates of ERM, Gibbs and related methods (cf. Tsybakov '04, Audibert '04, Bartlett and Mendelson, '06)

Theorem (G. & Mehta, 2017b)

Suppose loss potentially **unbounded** and u-central holds

$$\forall f \in \mathcal{F}, \epsilon \geq 0 : \quad \ell_{f^*} - \ell_f \lhd_{u(\epsilon)} \epsilon$$

and **????**

Then there is $C > 0$ such that for every $f \in \mathcal{F}, \epsilon > 0$:

$$\mathbf{E}_{Z \sim P} [r_f] \leq \mathbf{E}^{\mathrm{ann}, u(\epsilon)} [r_f] + \epsilon$$

Left vs Right Zhang

- G & Mehta, 2017b is about extending left-hand side of Zhang’s Theorem

- Remainder of talk is about G& Mehta, 2017a, which extends the right-hand side. Substantially more novel!

G & M Excess Risk Bound

For every $\widehat{\Pi}_n = \widehat{\Pi} \mid Z^n$, every prior Π_0, every $\eta > 0$:

$$\mathbf{E}_{f \sim \hat{\Pi}_n} \mathbf{E}^{\mathrm{ann},\eta}_{Z \sim P} \left[r_f(Z) \right] \trianglelefteq_{\eta n} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[\frac{1}{n} \sum_{i=1}^{n} r_f(Z_i) \right] + \frac{\mathrm{KL}(\color{red}{\hat{\Pi}_n} \| \color{blue}{\Pi_0})}{\eta \cdot n}$$

G & M Excess Risk Bound

For every $\widehat{\Pi}_n = \widehat{\Pi} \mid Z^n$, ~~every prior Π_0~~, every $\eta > 0$:

$$\mathbf{E}_{f \sim \hat{\Pi}_n} \ \mathbf{E}^{\text{ann},\eta}_{Z \sim P} \ [r_f(Z)] \trianglelefteq_{\eta n}$$

$$\mathbf{E}_{f \sim \hat{\Pi}_n} \left[\frac{1}{n} \sum_{i=1}^{n} r_f(Z_i) \right] + \cancel{\frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n}}$$

G & M Excess Risk Bound

For every $\widehat{\Pi}_n = \widehat{\Pi} \mid Z^n$, every **luckiness function** w, every $\eta > 0$:

$$\mathbf{E}_{f\sim\hat{\Pi}_n} \mathbf{E}^{\text{ann},\eta}_{Z\sim P} \left[r_f(Z)\right] \trianglelefteq_{\eta n}$$

$$\mathbf{E}_{f\sim\hat{\Pi}_n}\left[\frac{1}{n}\sum r_f(Z_i)\right] + \text{COMP}_\eta(\mathcal{F},\hat{\Pi},w,Z^n)$$

$$\text{COMP}_\eta(\mathcal{F},\hat{\Pi}_n,w,Z^n) = \frac{1}{\eta}\cdot\left(\mathbf{E}_{f\sim\hat{\Pi}_n}\left[-\log w(z^n,f)\right] + \log S(\mathcal{F},\hat{\Pi},w)\right)$$

data-dependent part

data-independent part

Bounding the novel complexity

- By different choices of w, $\mathrm{COMP}_\eta(\mathcal{F}, \hat{\Pi}, w, Z^n)$ can be further bounded so as to become a
  - KL divergence between prior and posterior (recovering and improving Zhang's bound)
  - Normalized Maximum Likelihood (NML) or Shtarkov Integral

    which can be further bounded in terms of ***Rademacher complexity****, VC dim, entropy nrs (right rates for polynomial entropy classes)*
  - Luckiness NML

    (useful for penalized estimators e.g. Lasso)

Bounding COMP for ERM/ML $\hat{f}$

- Let us take $\widehat{\Pi} \equiv \hat{f}$ to be ERM (note that for the log loss, this is just maximum likelihood)
- and let us take $w(z^n, f) \equiv 1$ *constant*

  Assume bounded losses here!

G & M Excess Risk Bound

For every $\widehat{\Pi}_n = \widehat{\Pi} \mid Z^n$, every luckiness fn w , every $\eta > 0$:

$$\mathbf{E}_{f\sim\hat{\Pi}_n}\ \mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P}\ [r_f(Z)] \trianglelefteq_{\eta n}$$

$$\mathbf{E}_{f\sim\hat{\Pi}_n}\left[\frac{1}{n}\sum r_f(Z_i)\right] + \mathrm{COMP}_\eta(\mathcal{F},\hat{\Pi},w,Z^n)$$

$$\mathrm{COMP}_\eta(\mathcal{F},\hat{\Pi}_n,w,Z^n) = \frac{1}{\eta}\cdot\left(\mathbf{E}_{f\sim\hat{\Pi}_n}\left[-\log w(z^n,f)\right] + \log S(\mathcal{F},\hat{\Pi},w)\right)$$

G & M Excess Risk Bound

For every deterministic $\hat{f}$, every luckiness fn w , $\eta > 0$:

$$\cancel{\mathbf{E}_{f \sim \hat{\Pi}_n}} \; \mathbf{E}^{\mathrm{ann},\eta}_{Z \sim P} \left[r_{\hat{f}_{|Z^n}}(Z) \right] \trianglelefteq_{\eta n} \cancel{\mathbf{E}_{f \sim \hat{\Pi}_n}} \left[\frac{1}{n} \sum r_{\hat{f}_{|Z^n}}(Z_i) \right] + \mathrm{COMP}_\eta(\mathcal{F}, \hat{\Pi}, w, Z^n)$$

$$\mathrm{COMP}_\eta(\mathcal{F}, \hat{\Pi}_n, w, Z^n) = \frac{1}{\eta} \cdot \left(\mathbf{E}_{f \sim \hat{\Pi}_n} \left[-\log w(z^n, f) \right] + \log S(\mathcal{F}, \hat{\Pi}, w) \right)$$

G & M Excess Risk Bound

For every deterministic $\hat{f}$, constant $w \equiv 1$, $\eta > 0$:

$$\cancel{\mathbf{E}_{f\sim\hat{\Pi}_n}}\ \mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P}\left[r_{\hat{f}_{|Z^n}}(Z)\right] \trianglelefteq_{\eta n}$$

$$\cancel{\mathbf{E}_{f\sim\hat{\Pi}_n}}\left[\frac{1}{n}\sum r_{\hat{f}_{|Z^n}}(Z_i)\right] + \mathrm{COMP}_\eta(\mathcal{F},\hat{\Pi},w,Z^n)$$

$$\mathrm{COMP}_\eta(\mathcal{F},\hat{\Pi}_n,w,Z^n) = \frac{1}{\eta}\cdot\left(\mathbf{E}_{f\sim\hat{\Pi}_n}\cancel{\left[-\log w(z^n,f)\right]} + \log S(\mathcal{F},\hat{\Pi},w)\right)$$

data-dependent part disappears

G & M Excess Risk Bound

For **ERM** $\hat{f}$, constant $w \equiv 1$, $\eta > 0$:

$$\cancel{\mathbf{E}_{f \sim \hat{\Pi}_n}} \; \mathbf{E}^{\mathrm{ann},\eta}_{Z \sim P} \left[r_{\hat{f}_{|Z^n}}(Z) \right] \trianglelefteq_{\eta n}$$

$$\cancel{\mathbf{E}_{f \sim \hat{\Pi}_n}} \left[\cancel{\frac{1}{n} \sum r_{\hat{f}_{|Z^n}}} (Z_i) \right] + \mathrm{COMP}_\eta(\mathcal{F}, \hat{\Pi}, w, Z^n)$$

$$\mathrm{COMP}_\eta(\mathcal{F}, \hat{\Pi}_n, w, Z^n) = \frac{1}{\eta} \cdot \left(\mathbf{E}_{f \sim \hat{\Pi}_n} \cancel{[-\log w(z^n, f)]} + \log S(\mathcal{F}, \hat{\Pi}, w) \right)$$

data-dependent part disappears

Excess Risk Bound for ERM

$$\mathbf{E}^{\text{ann},\eta}_{Z\sim P}\left[r_{\hat{f}_{|Z^n}}(Z)\right] \trianglelefteq_{\eta n} \eta^{-1} \cdot \log S(\mathcal{F}, \hat{f}, w_{\text{uniform}})$$

Excess Risk Bound for ERM

$$\mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P}\left[r_{\hat{f}_{|Z^n}}(Z)\right] \trianglelefteq_{\eta n} \eta^{-1}\cdot\log S(\mathcal{F},\hat{f},w_{\mathrm{uniform}})$$

$$S(\mathcal{F};\hat{f},w_{\mathrm{uniform}}) := \mathbf{E}_{Z^n\sim P}\left[\frac{e^{-\eta r_{\hat{f}_{|Z^n}}(Z^n)}}{C(\hat{f}_{|Z^n})}\right]$$

$$C(f) := \mathbf{E}_{Z^n\sim P}\left[e^{-\eta r_f(Z^n)}\right]$$

Excess Risk Bound for ERM

$$\mathbf{E}^{\text{ann},\eta}_{Z\sim P}\left[r_{\hat{f}_{|Z^n}}(Z)\right] \trianglelefteq_{\eta n} \eta^{-1}\cdot\log S(\mathcal{F},\hat{f},w_{\text{uniform}})$$

$$S(\mathcal{F};\hat{f},w_{\text{uniform}}) := \mathbf{E}_{Z^n\sim P}\left[\frac{e^{-\eta r_{\hat{f}_{|Z^n}}(Z^n)}}{C(\hat{f}_{|Z^n})}\right]$$

$$C(f) := \mathbf{E}_{Z^n\sim P}\left[e^{-\eta r_f(Z^n)}\right]$$

...to interpret this, define probability density fns q_f as

$$q_f(z) := p(z)\cdot\frac{e^{-\eta r_f(z)}}{\int p(z)e^{-\eta r_f(z)}d\nu(z)}$$

...and note that

$$S(\mathcal{F};\hat{f},w_{\text{uniform}}) = \int q_{\hat{f}_{|z^n}}(z^n)d\nu(z^n) \le \int q_{\hat{f}_{\mathbf{ML}|z^n}}(z^n)d\nu(z^n)$$

Excess Risk Bound for ERM

$$\mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P}\left[r_{\hat{f}_{|Z^n}}(Z)\right] \trianglelefteq_{\eta n} \eta^{-1}\cdot \log S(\mathcal{F},\hat{f},w_{\mathrm{uniform}})$$

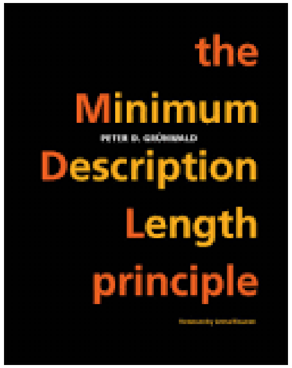
...where

$$S(\mathcal{F};\hat{f},w_{\mathrm{uniform}}) \leq S(\mathcal{F};\hat{f}_{\mathrm{ML}},w_{\mathrm{uniform}}) = \int q_{\hat{f}_{\mathrm{ML}|z^n}}(z^n)d\nu(z^n)$$

$\log S$ is cumulative minimax individual sequence regret for log-loss prediction relative to the set of densities $\{q_f : f \in \mathcal{F}\}$

- ...a.k.a. as Shtarkov or NML (normalized ML) complexity (Shtarkov 1988, Rissanen 1996, G. ‘07)

Excess Risk Bound for ERM



$$\mathbf{E}^{\text{ann},\eta}_{Z\sim P}\left[r_{\hat{f}_{|Z^n}}(Z)\right] \trianglelefteq_{\eta n} \eta^{-1}\cdot \log S(\mathcal{F},\hat{f}, w_{\text{uniform}})$$

...where

$$S(\mathcal{F};\hat{f}, w_{\text{uniform}}) \le S(\mathcal{F};\hat{f}_{\text{ML}}, w_{\text{uniform}}) = \int q_{\hat{f}_{\text{ML}|z^n}}(z^n)\,d\nu(z^n)$$

$\log S$ is cumulative minimax individual sequence regret for log-loss prediction relative to the set of densities $\{q_f : f \in \mathcal{F}\}$

...a.k.a. as Shtarkov or NML (normalized ML) complexity

...both intriguing and highly useful!

G & M Excess Risk Bound

For every $\widehat{\Pi}_n = \widehat{\Pi} \mid Z^n$, every luckiness fn w , every $\eta > 0$:

$$\mathbf{E}_{f \sim \hat{\Pi}_n} \; \mathbf{E}^{\text{ann},\eta}_{Z \sim P} \; [r_f(Z)] \trianglelefteq_{\eta n}$$

$$\mathbf{E}_{f \sim \hat{\Pi}_n} \left[\frac{1}{n} \sum r_f(Z_i) \right] + \text{COMP}_\eta(\mathcal{F}, \hat{\Pi}, w, Z^n)$$

$$\text{COMP}_\eta(\mathcal{F}, \hat{\Pi}_n, w, Z^n) = \frac{1}{\eta} \cdot \left(\mathbf{E}_{f \sim \hat{\Pi}_n} \left[-\log w(z^n, f) \right] + \log S(\mathcal{F}, \hat{\Pi}, w) \right)$$

G & M Excess Risk Bound

For every deterministic $\hat{f}$, every luckiness fn w , $\eta > 0$:

$$\cancel{\mathbf{E}_{f \sim \hat{\Pi}_n}} \; \mathbf{E}^{\mathrm{ann},\eta}_{Z \sim P} \left[r_{\hat{f}_{|Z^n}}(Z) \right] \trianglelefteq_{\eta n} \cancel{\mathbf{E}_{f \sim \hat{\Pi}_n}} \left[\frac{1}{n} \sum r_{\hat{f}_{|Z^n}}(Z_i) \right] + \mathrm{COMP}_{\eta}(\mathcal{F}, \hat{f}, w, Z^n)$$

$$\mathrm{COMP}_{\eta}(\mathcal{F}, \hat{f}, w, z^n) = \frac{1}{\eta} \cdot \left(- \log w(z^n, \hat{f}_{|z^n}) + \log S(\mathcal{F}, \hat{f}, w) \right)$$

G & M Excess Risk Bound

For every deterministic $\hat{f}$, every simple luckiness fn w :

$$\cancel{\mathbf{E}_{f\sim\hat{\Pi}_n}} \; \mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P}\left[r_{\hat{f}_{|Z^n}}(Z)\right] \trianglelefteq_{\eta n}$$

$$\cancel{\mathbf{E}_{f\sim\hat{\Pi}_n}}\left[\frac{1}{n}\sum r_{\hat{f}_{|Z^n}}(Z_i)\right] + \mathrm{COMP}_{\eta}(\mathcal{F}, \hat{f}, w, Z^n)$$

$$\mathrm{COMP}_{\eta}(\mathcal{F}, \hat{f}, w, z^n) = \frac{1}{\eta} \cdot \left(-\log w(z^n, \cancel{\hat{f}_{z^n}}) + \log S(\mathcal{F}, \hat{f}, w)\right)$$

G & M Excess Risk Bound

$$\cancel{\mathbf{E}_{f\sim\hat{\Pi}_n}} \; \mathbf{E}^{\text{ann},\eta}_{Z\sim P}\left[r_{\hat{f}_{|Z^n}}(Z)\right] \trianglelefteq_{\eta n}$$

$$\cancel{\mathbf{E}_{f\sim\hat{\Pi}_n}}\left[\frac{1}{n}\sum r_{\hat{f}_{|Z^n}}(Z_i)\right] + \text{COMP}_\eta(\mathcal{F},\hat{f},w,Z^n)$$

$$\text{COMP}_\eta(\mathcal{F},\hat{f},w,z^n) = \frac{1}{\eta}\cdot\left(-\log w(z^n,\cancel{\hat{f}_{|z^n}}) + \log S(\mathcal{F},\hat{f},w)\right)$$

...and now

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

Bounds for Penalized ERM

For every deterministic $\hat{f}$, every simple luckiness fn w :

$$\mathbf{E}^{\text{ann},\eta}_{Z\sim P}\left[r_{\hat{f}_{|Z^n}}(Z)\right] \trianglelefteq_{\eta n} \frac{1}{n}\sum r_{\hat{f}_{|Z^n}}(Z_i) + \text{COMP}_\eta(\mathcal{F},\hat{f},w,Z^n)$$

$$\text{COMP}_\eta(\mathcal{F},\hat{f},w,z^n) = \frac{1}{\eta}\cdot\left(-\log w(z^n) + \log S(\mathcal{F},\hat{f},w)\right)$$

Taking $w(z^n) = \exp(-\text{PEN}(\hat{f}_{|z^n}))$ for a penalization function PEN the bound is optimized if we take

$$\hat{f}_{|z^n} := \arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} \ell_f(z_i) + \eta^{-1}\text{PEN}(f)$$

Bounds for Penalized ERM

For every deterministic $\hat{f}$, every simple luckiness fn w :

$$\mathbf{E}_{Z\sim P}^{\mathrm{ann},\eta}\left[r_{\hat{f}_{|Z^n}}(Z)\right] \trianglelefteq_{\eta n} \frac{1}{n}\sum r_{\hat{f}_{|Z^n}}(Z_i) + \mathrm{COMP}_\eta(\mathcal{F},\hat{f},w,Z^n)$$

$$\mathrm{COMP}_\eta(\mathcal{F},\hat{f},w,z^n) = \frac{1}{\eta}\cdot\left(-\log w(z^n) + \log S(\mathcal{F},\hat{f},w)\right)$$

Taking $w(z^n) = \exp(-\mathrm{PEN}(\hat{f}_{|z^n}))$ for a penalization function PEN the bound is optimized if we take

$$\hat{f}_{|z^n} := \arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} \ell_f(z_i) + \eta^{-1}\mathrm{PEN}(f)$$

....we get (sharp!) bounds for Lasso and friends. We see that **multiplier in Lasso is 'just like' learning rate in Bayes**

Bounds for 'Posteriors' including generalized Bayes

For every $\widehat{\Pi}_n = \widehat{\Pi} \mid Z^n$, every luckiness fn w , every $\eta > 0$:

$$\mathbf{E}_{f\sim\hat{\Pi}_n} \ \mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P} \ [r_f(Z)] \trianglelefteq_{\eta n}$$

$$\mathbf{E}_{f\sim\hat{\Pi}_n} \left[\frac{1}{n} \sum r_f(Z_i) \right] + \mathrm{COMP}_\eta(\mathcal{F}, \hat{\Pi}, w, Z^n)$$

$$\mathrm{COMP}_\eta(\mathcal{F}, \hat{\Pi}_n, w, Z^n) = \frac{1}{\eta} \cdot \left(\mathbf{E}_{f\sim\hat{\Pi}_n} [-\log w(z^n, f)] + \log S(\mathcal{F}, \hat{\Pi}, w) \right)$$

$$S(\mathcal{F}, \hat{\Pi}, w) := \mathbf{E}_{Z^n\sim P} \left[\exp\left(-\mathbf{E}_{f\sim\hat{\Pi}|Z^n} [\eta r_f(Z^n) + \log C(f) - \log w(Z^n, f)] \right) \right]$$

Proposition

- Take arbitrary estimator $\widehat{\Pi}$ that outputs distribution over $\mathcal{F}$ and arbitrary prior Π_0. If we take $w(z^n, f) := \frac{\pi_0(f)}{\pi(f|z^n)}$ then we have

$$S(\mathcal{F}, \hat{\Pi}, w) \leq 1$$

(Proof is just Jensen)

- inequality is strict (gap accounts for 'localized' PAC-Bayes bounds; Catoni '03)

Now we reduce to Zhang...

For every $\widehat{\Pi}_n = \widehat{\Pi} \mid Z^n$, luckiness fn $w(z^n, f) := \frac{\pi_0(f)}{\pi(f|z^n)}$

$$\mathbf{E}_{f\sim\hat{\Pi}_n} \mathbf{E}^{\mathrm{ann},\eta}_{Z\sim P}\left[r_f(Z)\right] \trianglelefteq_{\eta n} \mathbf{E}_{f\sim\hat{\Pi}_n}\left[\frac{1}{n}\sum r_f(Z_i)\right] + \mathrm{COMP}_\eta(\mathcal{F},\hat{\Pi}, w, Z^n)$$

$$\mathrm{COMP}_\eta(\mathcal{F},\hat{\Pi}_n,w,Z^n) = \frac{1}{\eta}\cdot\left(\mathbf{E}_{f\sim\hat{\Pi}_n}\left[-\log w(z^n,f)\right] + \cancel{\log S(\mathcal{F},\hat{\Pi},w)}\right)$$

$$\mathbf{E}_{f\sim\hat{\Pi}_n}\left[-\log\frac{\pi_0(f)}{\hat{\pi}(f|z^n)}\right] = \mathrm{KL}(\hat{\Pi}_n\|\Pi_0)$$

More Remarks on Bound

- MDL relation:
  - If we take the generalized Bayesian posterior, RHS has a log-Bayesian marginal likelihood interpretation = codelength under Bayesian code
  - If we take deterministic $\hat{f}$ and constant w then RHS has a NML codelength interpretation

... Bayes and NML are two most important ‘universal coding strategies’ for data compression (G. 07)

... What’s going on here?

More Remarks on Bound

- It turns ot that every luckiness function (up to multiplicative constant) gives...
  - – different (incomparable) bound
  - – different way to code data using code that is ‘universal’ for constructed probability model $\{q_f : f \in \mathcal{F}\}$

....so there may be useful bounds here which nobody has explored yet

More Remarks on Bound

Bound is sharp! Why?

- It says $\mathrm{LHS} \trianglelefteq_{\eta n} \mathrm{RHS}$

  i.e. $\mathbf{E}\left[e^{\eta \cdot (\mathrm{LHS}-\mathrm{RHS})}\right] \leq 1$

...but the proof (which is straightforward rewriting!) actually gives that

$$\mathbf{E}\left[e^{\eta \cdot (\mathrm{LHS}-\mathrm{RHS})}\right] = 1$$

$$\mathrm{LHS} = \mathbf{E}_{f \sim \hat{\Pi}_n} \mathbf{E}^{\mathrm{ann},\eta}_{Z \sim P} \left[r_f(Z) \right]$$

$$\mathrm{RHS} = \mathbf{E}_{f \sim \hat{\Pi}_n} \left[\frac{1}{n} \sum r_f(Z_i) \right] + \mathrm{COMP}_\eta(\mathcal{F}, \hat{\Pi}, w, Z^n)$$

Thm 2: log Shtarkov bounded by Rademacher

- Let $\mathcal{F}$ have radius ε in the $L_2(P)$-pseudometric.
- Fix arbitrary $f^\circ \in \mathcal{F}$ and define $\mathcal{G} = \{\ell_f - \ell_{f^\circ} : f \in \mathcal{F}\}$
- For arbitrary deterministic estimators $\hat{f}$,

$$\mathrm{COMP}_\eta(\mathcal{F}, \hat{f}, w_{\mathrm{UNIFORM}}) \le \frac{1}{\eta} \cdot \log \int q_{\hat{f}_{\mathrm{ml}|z^n}}(z^n)\, dz^n \le$$

$$6n \cdot \mathbf{E}_{Z^n \sim q_{f^\circ}} \left[\mathrm{R\scriptstyle AD}_n(\mathcal{G} \mid Z^n)\right] + n \cdot \eta \cdot C \cdot \varepsilon^2$$

...where $\mathrm{R\scriptstyle AD}_n(\mathcal{G} \mid Z^n) := \mathbf{E}_{\epsilon_1,\ldots,\epsilon_n}\left[\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i g(Z_i) \right| \right]$

More Precisely

- Fix arbitrary $f^\circ \in \mathcal{F}$ and define $\mathcal{G} = \{\ell_f - \ell_{f^\circ} : f \in \mathcal{F}\}$
- Define centered empirical process

$$T_n := \sup_{f \in \mathcal{F}} \left\{ \sum_{j=1}^{n} (\ell_{f^\circ}(Z_j) - \ell_f(Z_j)) - \mathbf{E}_{Z^n \sim Q_{f^\circ}} \left[\sum_{j=1}^{n} (\ell_{f^\circ}(Z_j) - \ell_f(Z_j)) \right] \right\}.$$

- For arbitrary deterministic estimators $\hat{f}$,

$$\begin{aligned} \mathrm{COMP}_\eta(\mathcal{F}, \hat{f}, w_{\mathrm{UNIFORM}}) \le\ & 3 \cdot \mathbf{E}_{Z^n \sim q_{f^\circ}} [T_n] + n \cdot \eta \cdot C \cdot \varepsilon^2 \le \\ & 6n \cdot \mathbf{E}_{Z^n \sim q_{f^\circ}} [\mathrm{RAD}_n(\mathcal{G} \mid Z^n)] + n \cdot \eta \cdot C \cdot \varepsilon^2 \end{aligned}$$

More Precisely

- Fix arbitrary $f^\circ \in \mathcal{F}$ and define $\mathcal{G} = \{\ell_f - \ell_{f^\circ} : f \in \mathcal{F}\}$
- Define centered empirical process

$$T_n := \sup_{f \in \mathcal{F}} \left\{ \sum_{j=1}^{n} (\ell_{f^\circ}(Z_j) - \ell_f(Z_j)) - \mathbf{E}_{Z^n \sim Q_{f^\circ}} \left[\sum_{j=1}^{n} (\ell_{f^\circ}(Z_j) - \ell_f(Z_j)) \right] \right\}.$$

- For arbitrary deterministic estimators $\hat{f}$,

$$\begin{aligned}\mathrm{COMP}_\eta(\mathcal{F}, \hat{f}, w_{\mathrm{UNIFORM}}) \leq\ & 3 \cdot \mathbf{E}_{Z^n \sim q_{f^\circ}} [T_n] + n \cdot \eta \cdot C \cdot \varepsilon^2 \leq \\ & 6n \cdot \mathbf{E}_{Z^n \sim q_{f^\circ}} [\mathrm{RAD}_n(\mathcal{G} \mid Z^n)] + n \cdot \eta \cdot C \cdot \varepsilon^2\end{aligned}$$

Proof: recycle Opper&Haussler '99 who bound Shtarkov in terms of L_∞ entropy nrs; replace Yurinskii's inequality by Bousquet-Talagrand (see also Cesa-Bianchi&Lugosi '01)

Optimal Rates for Large Classes

$$\mathrm{COMP}_\eta(\mathcal{F}, \hat{f}, w_{\text{UNIFORM}}) \leq 3 \cdot \mathbf{E}_{Z^n \sim q_{f^\circ}} [T_n] + n \cdot \eta \cdot C \cdot \varepsilon^2$$

- Remainder term ε^2 is small enough so as to get optimal rates for really large classes in classification a la Tsybakov '04 under a β – Bernstein condition

$$\mathrm{COMP}_\eta(\mathcal{F}, \hat{f}, w_{\text{UNIFORM}}) \leq 3 \cdot \mathbf{E}_{Z^n \sim q_{f^\circ}}[T_n] + n \cdot \eta \cdot C \cdot \varepsilon^2$$

- We say that a class $\mathcal{F}$ of functions $\mathcal{X} \to \mathcal{Y}$ has polynomial bracketing $L_1(P)$ entropy if for some $A > 0$, $0 < \rho < 1$, all $\varepsilon > 0$, $\log N_{[\cdot]}(\mathcal{F}, L_1(P), \varepsilon) \leq \left(A^2/\varepsilon\right)^\rho$

$$\mathrm{COMP}_\eta(\mathcal{F}, \hat{f}, w_{\mathrm{UNIFORM}}) \leq 3 \cdot \mathbf{E}_{Z^n \sim q_{f^\circ}} [T_n] + n \cdot \eta \cdot C \cdot \varepsilon^2$$

- We say that a class $\mathcal{F}$ of functions $\mathcal{X} \rightarrow \mathcal{Y}$ has polynomial bracketing $L_1(P)$ entropy if for some $A > 0$, $0 < \rho < 1$, all $\varepsilon > 0$, $\log N_{[\cdot]}(\mathcal{F}, L_1(P), \varepsilon) \leq \left(A^2/\varepsilon\right)^\rho$

- Following Massart and Nédélec (2006), we can bound T_n in terms of A and ρ using **chaining** such that bound above becomes

$$\frac{\mathrm{COMP}_\eta}{n} \lesssim (A \cdot C)^{\frac{2\rho}{1+\rho}} \cdot n^{-\frac{1}{1+\rho}} \cdot \eta^{-\frac{1-\rho}{1+\rho}}$$

- Under a $\beta = \frac{1}{k}$ - Bernstein condition, optimizing over η in our excess risk bound then gives the minimax optimal rate for ERM: $n^{-\frac{\kappa}{2\kappa - 1 + \rho}}$

Additional Niceties

- Bounding T_n in terms of $L_1(P)$ bracketing entropy nrs gives optimal rates for large classes
- Opper, Haussler (‘99), Cesa-Bianchi, Lugosi (’01) bounded NML/minimax log-loss regret in terms of L_∞ entropy nrs; by bounding it in terms of T_n/Rademacher complexity which we can further bound in terms of $L_1(P), L_{2(P)}$ and $L_2(P_n)$ entropy nrs, we obtain strictly better bounds!
- No (difficult!) localized Rademacher complexities needed

Thank you for your attention!

Further Reading:

- G. and Van Ommen, *Bayesian Analysis, Dec. 2017*
- G. and Mehta, Fast Rates for Unbounded Losses, arXiv (2016, 2017b)
- G. and Mehta. *A Tight Excess Risk bound in terms of a Unified PAC-Bayesian-Rademacher-MDL Complexity*, arXiv (2017a)

The Critical $\bar{\eta}$

- $\bar{\eta}$ is defined as largest $\eta > 0$ such that for all $f \in \mathcal{F}$,

$$A(\eta) := \mathbf{E}_{Z \sim P} \left(\frac{p_f(Z)}{p_{f^*}(Z)} \right)^{\eta} \leq 1$$

The Critical $\bar{\eta}$

- $\bar{\eta}$ is defined as largest $\eta > 0$ such that for all $f \in \mathcal{F}$,

$$A(\eta) := \mathbf{E}_{Z \sim P} \left(\frac{p_f(Z)}{p_{f^*}(Z)} \right)^{\eta} \leq 1$$

...if model **correct**, $\bar{\eta}$= 1, since

$$A(1) = \mathbf{E}_{Z \sim P_{f^*}} \left(\frac{p_f}{p_{f^*}} \right)^1 = \int p_{f^*} \frac{p_f}{p_{f^*}} = 1$$

- ...and $A(0) = 1$ and $A(\eta)$ is (strictly) convex
- if model **convex**, also $\bar{\eta} \leq 1$ (Barron & Li, ‘99)

...otherwise still often $\bar{\eta} > 0$ but smaller... (G&M ‘17)

Relation to Log-Loss Prediction/Data Compression

- $\log S\,(\mathcal{F}, \hat{f}_{\mathrm{ml}}, w)$ has interpretation as minimax individual sequence regret for the q_f densities with uniform w

- Similarly $\log S\,(\mathcal{F}, \hat{f}_{\mathrm{map}}, w)$ it has interpretation as minimax individual sequence luckiness regret (G. '07, Bartlett et al. '13) for general w, with the corresponding **MAP estimator**

$$\hat{f}_{\mathrm{map}|z^n} := \arg\max_{f \in \mathcal{F}} q_f(z^n) w(z^n)$$

Main Insight of G&M, 2017b: One-Sided unbounded loss conds.

Suppose risk bounded and **u-central holds**

$$\forall f \in \mathcal{F}, \epsilon \geq 0: \quad \ell_{f^*} - \ell_f \trianglelefteq_{u(\epsilon)} \epsilon \text{ i.e. } -r_f \trianglelefteq_{u(\epsilon)} \epsilon$$

exponential tail-control of $-r_f$

and **witness-of-badness** holds: there is $A, c > 0$ s.t.:

$$\forall f \in \mathcal{F}: \ \mathbf{E}_{Z \sim P}\left[r_f \cdot \mathbf{1}_{r_f > A}\right] \leq c \cdot \mathbf{E}_{Z \sim P}\left[r_f \cdot \mathbf{1}_{r_f \leq A}\right]$$

much weaker sort of tail-control of r_f

Then ... ...

Excess risk bounded both **with very high probability** and expectation [Version for Polynomial Tails As Well]