Eclectic Lectures Part III: Safe Bayes, Statistical Learning

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematical Institute – Leiden University

Joint work with Nishant Mehta, Thijs van Ommen, Rianne de Heide

for all $P \in \mathcal{H}_0$:

$$\mathbf{E}_{S \sim P}[S] \leq 1$$

Invariably, *S* nonnegative

Rough Plan of Lectures

1. Safe Testing (Statistics/AB Testing)
2. Safe Testing (Information Theory)
3. Safe and Generalized Bayes
   – Zhang-G.-Mehta Thm density estimation
4. Fast Rate Conditions in Statistical (stochastic) and Online (nonstochastic) Learning
   – Zhang-G-Mehta Thm general loss fns
5. Safety and Luckiness

Generalized Posterior

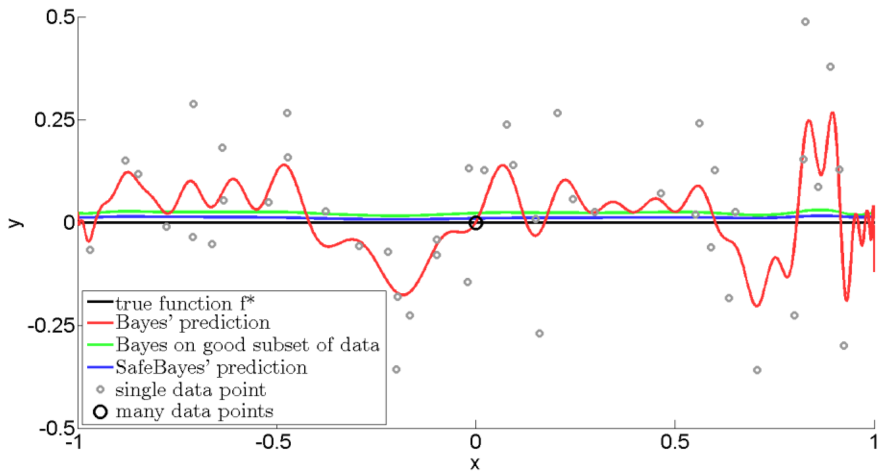
- Let $\{ p_f : f \in \mathcal{F} \}$ be a model, i.e. a set of densities
- We define the η-generalized posterior to be

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

cf. Vovk (1990), Walker & Hjort (2001), Zhang (2006), G. (2011, 2012)

$$\pi(f \mid X^n, Y^n, \eta) \propto \prod_{i=1}^{n} p_f(Y_i \mid X_i)^\eta \cdot \pi(f)$$

$\eta = 1$ (standard Bayes) behaves badly under misspecification; problem goes away with $\eta < 0.4$



- See G. and Van Ommen. Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing it . *Bayesian Analysis*, December 2017 (also ISBA 2016). Also R. de Heide, Master's Thesis, Leiden 2016 (real-world data)

The Critical $\bar{\eta}$

Let $Z_1, Z_2, \ldots \sim$ i.i.d. P

Let f^* be element of $\mathcal{F}$ minimizing KL divergence to P

Let $\bar{\eta}$ be largest $\eta > 0$ such that for all $f \in \mathcal{F}$,

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

(assume both f^* and $\bar{\eta}$ exist for now)

η-Bayes "works" for any $\eta < \bar{\eta}$

What is critical $\bar{\eta}$?

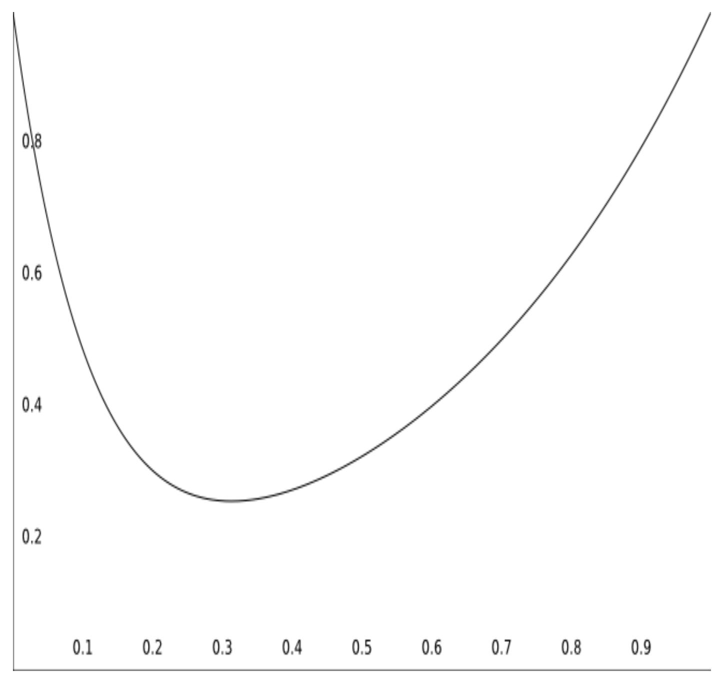
- Define $A(\eta) = \mathbf{E}_{Z \sim P} \left(\frac{p_f}{p_{f^*}} \right)^{\eta}$

- 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



Misspecified Case

- If model $\mathcal{F}$ is convex, then (Li '99) for all $f \in \mathcal{F}$

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

so again, η-Bayes with any $\eta \leq 1$ will work...

This is just the
Reverse Information Projection Theorem!

Misspecified Case

- If model $\mathcal{F}$ is convex, then (Li '99) for all $f \in \mathcal{F}$

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

so again, η-Bayes with any $\eta \leq 1$ will work...

- We require set of *densities* to be convex; most statistical models are *not* convex in this sense. e.g. linear regression with convex set of regression functions is not.

Convex Luckiness

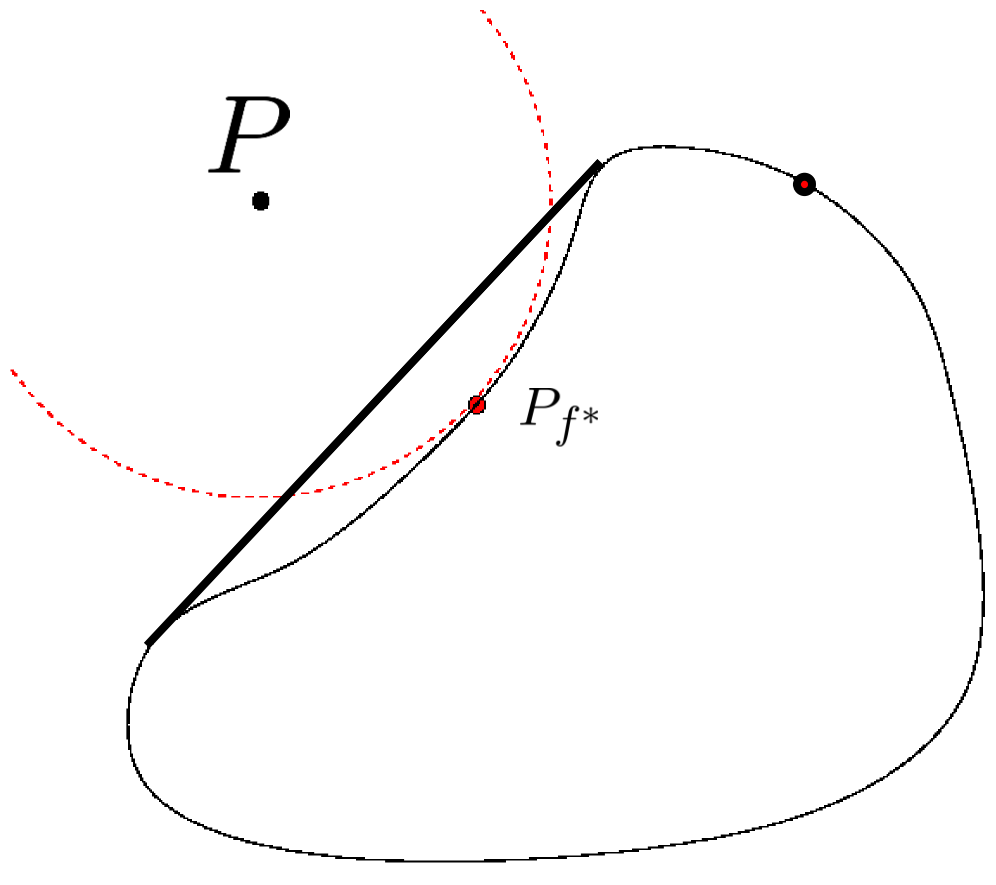
- We say that **convex luckiness** holds if

$$\inf_{f \in \mathcal{F}} D(P \| P_f) = \inf_{f \in \textsc{conv-hull}(\mathcal{F})} D(P \| P_f)$$

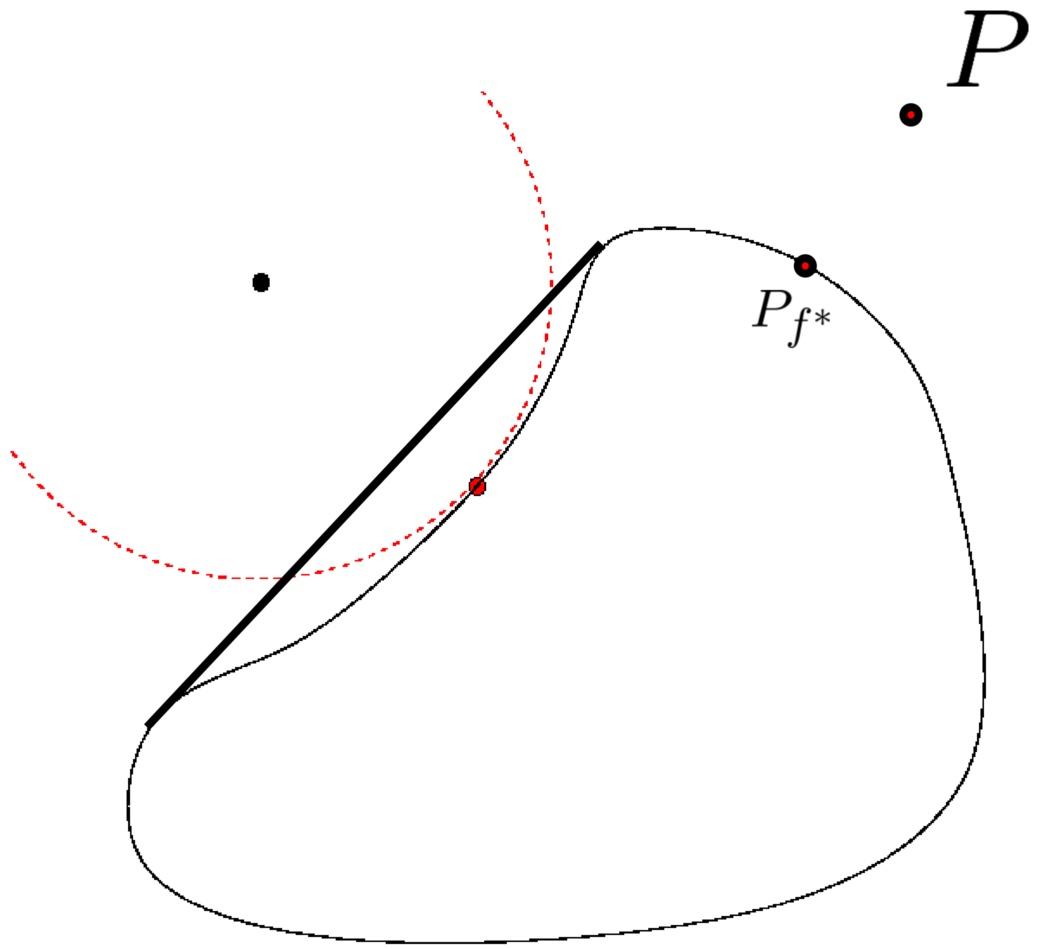
(Van Erven et al. ’15, G & Mehta ’17b)

- Under convex luckiness, we can ‘get away’ with (almost) standard Bayes: η-Bayes with any $\eta < 1$ will “work”...

Bad and Good Misspecification



Bad and Good Misspecification



Misspecified Case, Example

- Standard Linear Regression Model with Fixed Variance $\tilde{\sigma}^2$, i.e. $\mathcal{F}$ is set of functions $\mathcal{X} \rightarrow \mathcal{Y} = \mathbb{R}$

$$p_f(y|x) \propto e^{-\frac{(y-f(x))^2}{2\tilde{\sigma}^2}}$$

- Suppose "true" $P(Y|X)$ has exponentially small tails*, and for some $f^* \in \mathcal{F}$ $\quad \mathbf{E}_P[Y \mid X] = f^*(X)$

and variance $\sigma_x^2 := \mathbf{E}_P[Y - f^*(X))^2 \mid X = x]$

(signal well-specified, noise misspecified)

- ...then

$$\bar{\eta} \geq \frac{\tilde{\sigma}^2}{\sup_x \sigma_x^2}$$

Simple Example - Critical $\bar{\eta} < 1$

- Let $X_1, X_2, \ldots$ be i.i.d. Bernoulli(p^*)
- Model is $p \in \{0.2, 0.8\}$,
- Prior is $w(0.2) = w(0.8) = 1/2$.
- "True" $p^* =$½ (in practice: close to ½)
- By CLT: $w(p \mid X^n) = O(e^{-\eta\sqrt{n}})$ for either $p = 0.2$ or $p = 0.8$
- Bayes is very convinced that one of the two hypotheses is true, even though they're equally false
- If we set $\eta = 1/\sqrt{n}$, this will not happen. Indeed this is 'optimal' value in this case.

Critical $\bar{\boldsymbol{\eta}} > 1$: borderline case

- Model is $p \in [0.2,0.8]$.
- "True" $p^* = 1$ (hence we see 1,1,1,1....)
- $\tilde{p} = 0.8$ is closest to p^* in KL divergence.
- Now data are more informative for learning $\tilde{p}$ than you would expect them to be if $\tilde{p}$ where true...
- ...hence it makes sense to learn faster than usual: set $\eta \gg 1$ ($\bar{\eta} = \infty \rightarrow$ Bayes puts all mass on ML estimator $\hat{p} = 0.8$)
- In realistic cases $\bar{\eta}$ not so high but might still be > 1

Reasons why using $\eta < \bar{\eta}$ does work

1. Union Bound/Zhang-G.M. Convergence Theorem
2. "No Hypercompression" Theorem
3. η-generalized Bayes becomes standard Bayes for modified model!

Posterior Concentration Theorem

G. & Mehta, 2017b

For all $0 < \eta < \bar{\eta}$, under no further conditions

$$\mathbf{E}_{Z^n \sim P} \mathbf{E}_{f \sim \Pi | Z^n} \left[d^2_{\text{GEN. HELLINGER},\eta}(f^* \| f) \right]$$
$$\leq C_\eta \cdot \inf_{\epsilon \geq 0} \left\{ \epsilon + \frac{-\log \Pi_0(B_{D_P}(f^*, \epsilon))}{\eta \cdot n} \right\}$$

$f^* = \arg\min_{f \in \mathcal{F}} D(P \| P_f)$ represents KL-optimal density

$D_P(P_{f^*} \| P_f) = \mathbf{E}_{Z \sim P} \left[\log \frac{p_{f^*}(Z)}{p_f(Z)} \right]$ is generalized KL div.

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

Retrieve Ghosal, Gosh, VDVaart (2000), under weaker conditions !

So why $\eta < \bar{\eta}$ rather than $\eta = \bar{\eta}$?

- If we take $\eta = \bar{\eta}$ then this is sufficient to prove consistency/convergence (at right rate) of Bayes posterior predictive distribution

i.e.

$$\bar{p}_\eta(z_i \mid z^{i-1}) := \int_{\mathcal{F}} p'_{f,\eta}(z_i)\, d\Pi(f \mid z^{i-1})$$

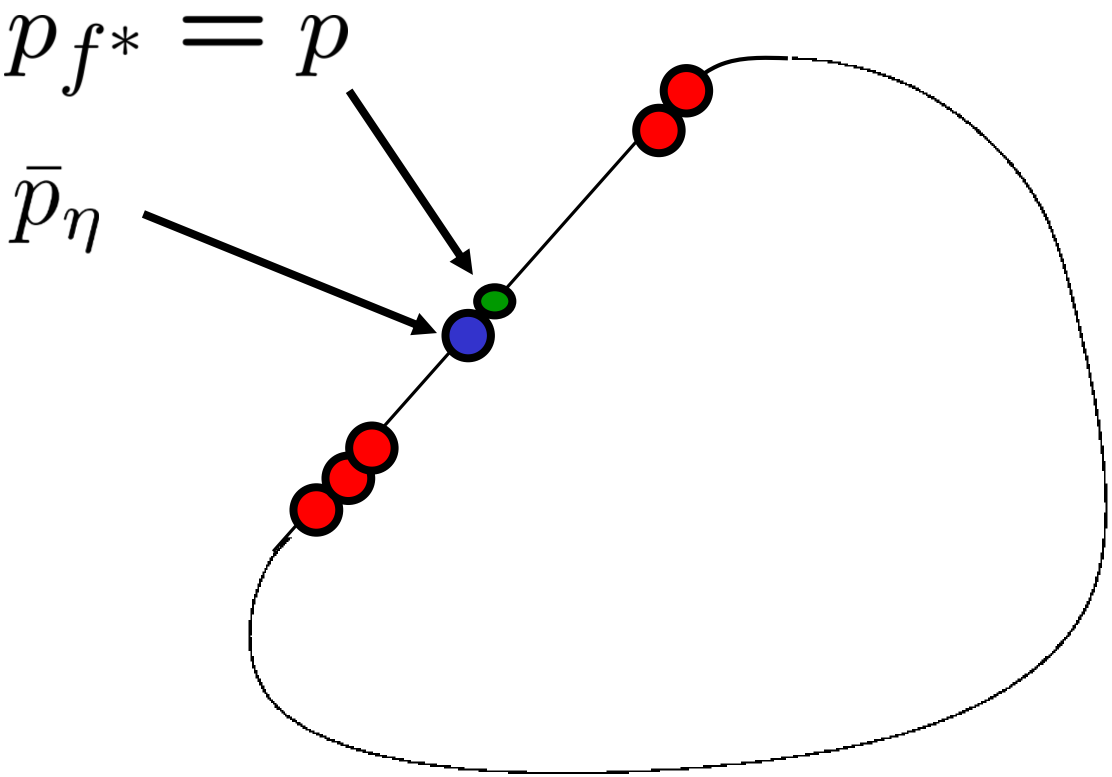
$$\bar{p}_\eta(Z_i = \cdot \mid Z^{i-1}) \to p_{f^*,\eta}$$

where the convergence is ‘in mean sum’ (Barron ISBA ‘98, Grünwald ‘07)

So why $\boldsymbol{\eta} < \bar{\boldsymbol{\eta}}$ rather than $\boldsymbol{\eta} = \bar{\boldsymbol{\eta}}$?

- If we take $\eta = \bar{\eta}$ then this is sufficient to prove consistency/convergence (at right rate) of Bayes posterior predictive distribution
- But if we want concentration of the posterior, then something weird can (and sometimes does) happen...
  - Barron (ISBA '99), Cziszar & Shields (inconsistency of Bayes model selection for Markov models) and Zhang ('06)...
  - Very different from Diaconis-Freedman Bayes inconsistency!

Bad Posterior, Good Predictive



Zhang, G. and Mehta

- Posterior Concentration Theorem is slight extension of Zhang's (2006) bound, itself related to Catoni/Audibert's bounds
- Zhang's bound also applies to general loss fns rather than log likelihood)
- In recent work, G&M (2016, 2017) tremendously generalized Zhang's bound
- Plan:

1. Posterior concentration version of Zhang's bound
2. Extension to general loss fns
3. Our Extensions

Zhang's ('06) Bound, Special Case

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 $0 < \eta < \bar{\eta}$:

$$\mathbf{E}_{Z^n \sim P} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[d^2_{\bar{\eta}}(f^* \| f) \right] \leq$$

$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[\mathbf{E}_{f \sim \hat{\Pi}_n} \left[-\frac{1}{n} \cdot \log \frac{p_f(Z^n)}{p_{f^*}(Z^n)} \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n} \right]$$

Zhang's ('06) 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 $0 < \eta < \bar{\eta}$:

$$\mathbf{E}_{Z^n \sim P} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[d^2_{\bar{\eta}}(f^* \| f) \right] \leq$$

$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[\mathbf{E}_{f \sim \hat{\Pi}_n} \left[-\frac{1}{n} \cdot \log \frac{p_f(Z^n)}{p_{f^*}(Z^n)} \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n} \right]$$

generalized Hellinger distance: under $\bar{\eta} = 1$ and well-specification, this becomes squared standard Hellinger distance:

$$d_1^2(f^* \| f) = \int \left(\sqrt{p_{f^*}(z)} - \sqrt{p_f(z)} \right)^2 dz$$

Zhang's ('06) 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 $0 < \eta < \bar{\eta}$:

$$\mathbf{E}_{Z^n \sim P} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[d^2_{\bar{\eta}}(f^* \| f) \right] \leq$$

$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[\mathbf{E}_{f \sim \hat{\Pi}_n} \left[-\frac{1}{n} \cdot \log \frac{p_f(Z^n)}{p_{f^*}(Z^n)} \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n} \right]$$

Zhang's ('06) 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 $0 < \eta < \bar{\eta}$:

$$\mathbf{E}_{Z^n \sim P} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[d^2_{\bar{\eta}}(f^* \| f) \right] \leq$$
$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[\mathbf{E}_{f \sim \hat{\Pi}_n} \left[-\frac{1}{n} \cdot \log \frac{p_f(Z^n)}{p_{f^*}(Z^n)} \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n} \right]$$

Example: $\mathcal{F}$ finite, $\widehat{\Pi}_n$ implements ML, i.e. puts probability 1 on ML estimator $\hat{f}$:

$$\mathbf{E}_{Z^n \sim P} \left[d^2_{\bar{\eta}}(f^* \| \hat{f}_{|Z^n}) \right] \leq$$
$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[-\frac{1}{n} \cdot \log \frac{p_{\hat{f}_{Z^n}}(Z^n)}{p_{f^*}(Z^n)} + \frac{-\log \pi_0(\hat{f}_{|Z^n})}{\eta \cdot n} \right]$$

Zhang's ('06) 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 $0 < \eta < \bar{\eta}$:

$$\mathbf{E}_{Z^n \sim P} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[d^2_{\bar{\eta}}(f^* \| f) \right] \leq$$

$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[\mathbf{E}_{f \sim \hat{\Pi}_n} \left[-\frac{1}{n} \cdot \log \frac{p_f(Z^n)}{p_{f^*}(Z^n)} \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n} \right]$$

Example: $\mathcal{F}$ finite, $\widehat{\Pi}_n$ implements ML, i.e. puts probability 1 on ML estimator $\hat{f}$:

$$\mathbf{E}_{Z^n \sim P} \left[d^2_{\bar{\eta}}(f^* \| \hat{f}_{|Z^n}) \right] \leq$$

$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[\cancel{-\frac{1}{n} \cdot \log \frac{p_{\hat{f}_{|Z^n}}(Z^n)}{p_{f^*}(Z^n)}} + \frac{-\log \pi_0(\hat{f}_{|Z^n})}{\eta \cdot n} \right]$$

Zhang's ('06) 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 $0 < \eta < \bar{\eta}$:

$$\mathbf{E}_{Z^n \sim P} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[d^2_{\bar{\eta}}(f^* \| f) \right] \leq$$

$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[\mathbf{E}_{f \sim \hat{\Pi}_n} \left[-\frac{1}{n} \cdot \log \frac{p_f(Z^n)}{p_{f^*}(Z^n)} \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n} \right]$$

Example: $\mathcal{F}$ finite, $\widehat{\Pi}_n$ implements ML, i.e. puts probability 1 on ML estimator $\hat{f}$:

$$\mathbf{E}_{Z^n \sim P} \left[d^2_{\bar{\eta}}(f^* \| \hat{f}_{|Z^n}) \right] \leq$$

$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[\qquad \frac{-\log \pi_0(\hat{f}_{|Z^n})}{\eta \cdot n} \right]$$

Zhang's ('06) 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 $0 < \eta < \bar{\eta}$:

$$\mathbf{E}_{Z^n \sim P} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[d^2_{\bar{\eta}}(f^* \| f) \right] \leq$$

$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[\mathbf{E}_{f \sim \hat{\Pi}_n} \left[-\frac{1}{n} \cdot \log \frac{p_f(Z^n)}{p_{f^*}(Z^n)} \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n} \right]$$

Example: $\mathcal{F}$ finite, $\widehat{\Pi}_n$ implements ML, i.e. puts probability 1 on ML estimator $\hat{f}$, **Π_0 uniform**:

$$\mathbf{E}_{Z^n \sim P} \left[d^2_{\bar{\eta}}(f^* \| \hat{f}_{|Z^n}) \right] \leq$$

$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[\qquad \frac{\overset{\log |\mathcal{F}|}{\cancel{-\log \pi_0(\hat{f}_{|Z^n})}}}{\eta \cdot n} \right]$$

Zhang's ('06) 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 $0 < \eta < \bar{\eta}$:

$$\mathbf{E}_{Z^n \sim P} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[d^2_{\bar{\eta}}(f^* \| f) \right] \leq$$

$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[\mathbf{E}_{f \sim \hat{\Pi}_n} \left[-\frac{1}{n} \cdot \log \frac{p_f(Z^n)}{p_{f^*}(Z^n)} \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n} \right]$$

Zhang's ('06) bound

For ***η*-generalized Bayes posterior** $\widehat{\Pi}_n := \widehat{\Pi}|\mathrm{Z}^{\mathrm{n}}$ based on arbitrary 'prior' Π_0 every $0 < \eta < \bar{\eta}$:

$$\mathbf{E}_{Z^n \sim P} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[d^2_{\bar{\eta}}(f^* \| f) \right] \leq$$

$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[\mathbf{E}_{f \sim \hat{\Pi}_n} \left[-\frac{1}{n} \cdot \log \frac{p_f(Z^n)}{p_{f^*}(Z^n)} \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n} \right]$$

Zhang's ('06) bound

For ***η*-generalized Bayes posterior** $\widehat{\Pi}_n := \widehat{\Pi}|\mathrm{Z}^{\mathrm{n}}$ based on arbitrary 'prior' Π_0 every $0 < \eta < \bar{\eta}$:

$$\mathbf{E}_{Z^n \sim P} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[d^2_{\bar{\eta}}(f^* \| f) \right] \leq$$

$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[\mathbf{E}_{f \sim \hat{\Pi}_n} \left[-\frac{1}{n} \cdot \log \frac{p_f(Z^n)}{p_{f^*}(Z^n)} \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n} \right]$$

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

Zhang's ('06) bound

For ***η*-generalized Bayes posterior** $\widehat{\Pi}_n := \widehat{\Pi}|Z^n$ based on arbitrary 'prior' Π_0 every $0 < \eta < \bar{\eta}$:

$$\mathbf{E}_{Z^n \sim P} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[d^2_{\bar{\eta}}(f^* \| f) \right] \leq$$

$$C_\eta \cdot \mathbf{E}_{Z^n \sim P} \left[\mathbf{E}_{f \sim \hat{\Pi}_n} \left[-\frac{1}{n} \cdot \log \frac{p_f(Z^n)}{p_{f^*}(Z^n)} \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n} \right]$$

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

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

$$D_P(P_{f^*} \| P_f) = \mathbf{E}_{Z \sim P} \left[\log \frac{p_{f^*}(Z)}{p_f(Z)} \right] \quad B_{D_P}(f^*, \epsilon) = \{ f \in \mathcal{F} : D_P(f^* \| f) \leq \epsilon \}$$

$$D_P(P_{f^*} \| P_f) = \mathbf{E}_{Z \sim P}\left[\log \frac{p_{f^*}(Z)}{p_f(Z)}\right]$$

Zhang's ('06) bound

For **η-generalized Bayes posterior** $\widehat{\Pi}_n := \widehat{\Pi} | Z^n$ based on arbitrary 'prior' Π_0 every $0 < \eta < \bar{\eta}$:

$$\mathbf{E}_{Z^n \sim P} \mathbf{E}_{f \sim \hat{\Pi}_n} \left[d^2_{\bar{\eta}}(f^* \| f) \right] \leq$$

$$C_\eta \cdot \mathbf{E}_{Z^n \sim P}\left[\mathbf{E}_{f \sim \hat{\Pi}_n}\left[-\frac{1}{n} \cdot \log \frac{p_f(Z^n)}{p_{f^*}(Z^n)} \right] + \frac{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}{\eta \cdot n} \right]$$

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

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

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

Retrieve Ghosal, Gosh, VDVaart!

Zhang, G. and Mehta

- Posterior Concentration Theorem is slight extension of Zhang's (2006) bound, itself related to Catoni/Audibert's bounds
- Zhang's bound also applies to general loss fns rather than log likelihood
- In recent work, G&M (2016, 2017) tremendously generalized Zhang's bound
- Plan:
  1. Posterior concentration version of Zhang's bound
  2. Extension to general loss fns
  3. Our Extensions

First Extension: ESI notation

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 $0 < \eta < \bar{\eta}$:

$$\mathbf{E}_{f\sim\hat{\Pi}_n}\left[d^2_{\bar{\eta}}(f^*\|f)\right] \trianglelefteq_{\eta n} C_\eta \cdot \left(\mathbf{E}_{f\sim\hat{\Pi}_n}\left[-\frac{1}{n}\cdot\log\frac{p_f(Z^n)}{p_{f^*}(Z^n)}\right] + \frac{\mathrm{KL}(\hat{\Pi}_n\|\Pi_0)}{\eta\cdot n}\right)$$

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

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

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

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
- $\ell_f \colon \mathcal{Z} \to \mathbb{R}$ loss function for predictor f

e.g. squared error loss,

$$Z_i = (X_i, Y_i) \; ; \; \ell_f((x,y)) = \big(y - f(x)\big)^2$$

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

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
- $\ell_f \colon \mathcal{Z} \to \mathbb{R}$ loss function for predictor f

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

- Works for arbitrary loss functions; for log-loss, $\ell_f(Z) = -\log p_f(Z)$, Gibbs posterior reduces to generalized posterior

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 $\eta > 0$:

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

holds for general distribution-output estimators (including deterministic estimators, e.g. ERM)

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 $\eta > 0$:

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

G. & Mehta 2016 mostly about extending the left-hand side

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 $\eta > 0$:

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

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

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

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 $Z^n = (Z_1, \dots, Z_n) = \big((X_1, Y_1,), \dots, (X_n, Y_n)\big) \sim$ i.i.d. P

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

$$\Rightarrow \quad P(X \geq Y + a) \leq e^{-\gamma a}$$

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 $\eta > 0$:

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

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

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 $\eta > 0$:

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

$r_f(Z) := \ell_f\ (Z) - \ell_{f^*}(Z)$ 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 $\widehat{\Pi}_n := \widehat{\Pi}|\mathrm{Z}^{\mathrm{n}}$ 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

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

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

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 $\eta > 0$:

$$\cancel{\mathbf{E}_{f \sim \hat{\Pi}_n}} \; \mathbf{E}^{\mathrm{ann},\eta}_{Z \sim P} \left[r_f(Z) \right] \trianglelefteq_{\eta n} \cancel{\mathbf{E}_{f \sim \hat{\Pi}_n}} \left[\frac{1}{n} \sum_{i=1}^{n} r_f(Z_i) \right] + \frac{\cancelto{-\log \pi_0(\hat{f}_{|Z^n})}{\mathrm{KL}(\hat{\Pi}_n \| \Pi_0)}}{\eta \cdot n}$$

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 $\eta > 0$:

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

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

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 $\eta > 0$:

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

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

Left-hand side: 'annealed' excess risk.
Can under some conditions be replaced by actual excess risk for sufficiently small η

Let us assume that we can do this for now!

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 $\eta > 0$:

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

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

Left-hand side: 'annealed' excess risk.
Can under some conditions be replaced by actual excess risk for sufficiently small η

Let us assume that we can do this for now!

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 , under appropriate conds. on (P, ℓ_f, η)

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

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

Example: ERM (empirical risk minimization)

(think of e.g. least squares)

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 , under appropriate conds. on (P, ℓ_f, η)

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

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

Example: ERM (empirical risk minimization)

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 , under appropriate conds. on (P, ℓ_f, η)

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

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

Example: ERM (empirical risk minimization)

...with uniform prior and finite $\mathcal{F}$ ...

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 , under appropriate conds. on (P, ℓ_f, η)

$$\mathbf{E}_{Z \sim P} \left[r_{\hat{f}_{|Z^n}}(Z) \right] \trianglelefteq_{\eta n} C \cdot \frac{\log |\mathcal{F}|}{\eta n}$$

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

Example: ERM (empirical risk minimization)

...with uniform prior and finite $\mathcal{F}$ ...

get O(1/n) convergence rate!

Log-Loss

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 , under appropriate conds. on (P, ℓ_f, η)

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

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

For log-loss, left-hand side is generalized KL divergence and right-hand side is log-likelihood ratio!

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

KL vs Hellinger

- Apparently, if the ‘special conditions’ hold that allow us to replace annealed excess risk by actual excess risk and we consider log-loss, we get original version of Zhang’s theorem back but with a KL instead of a Hellinger on the left!

  (works also with probabilistic estimator)
- Both stronger and conceptually nicer!

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

Zhang's Excess Risk Bound

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

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

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 $\mathbf{E}[r_f]$!

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

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

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! (that's why we state it like this)

Zhang's Excess Risk Bound

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

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

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

annealed excess risk is lower bound on actual excess risk

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

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)
- log-loss with misspecified model: for any $\eta < \bar{\eta}$ annealed risk larger than constant times generalized Hellinger distance^2 (G&M '17a)
- But from now on we are only interested in excess risk on the left
  - For log-loss & well-specified this is nicer
  - For other loss fns / misspecified this is essential! (otherwise noninterpretable)

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 $\eta > 0$:

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

U-Central Condition

(Van Erven et al. 2015)

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.

Probability that any fixed f performs much better than optimal-in-expectation f^* is exponentially small

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.

Eqv. to: $\forall 0 < \eta \leq u(\epsilon) : \mathbf{E}\left[e^{\eta(\ell_{f^*} - \ell_f)}\right] \leq e^{\eta\epsilon}$

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.

eqv. to: $\forall 0 < \eta \leq u(\epsilon) : \mathbf{E}\left[e^{\eta(\ell_{f^*} - \ell_f)}\right] \leq e^{\eta\epsilon}$

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

Our main equation is back!

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.

eqv. to: $\forall 0 < \eta \leq u(\epsilon) : \mathbf{E}\left[e^{\eta(\ell_{f^*} - \ell_f)}\right] \leq e^{\eta\epsilon}$

For general loss fns, we say that **strong central** holds if u-central holds for constant $u(0) = u(\epsilon) = \bar{\eta}$ (best case!)

If it only holds for u with $\lim_{\epsilon \downarrow 0} u(\epsilon) = 0$, then we say that **weak central** holds

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

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

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 $f \in \mathcal{F}, \epsilon > 0$:

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

...so now annealed risk on left replaced by actual risk (symmetric result)

$$\frac{1}{n}\sum_{i=1}^{n} r_f(Z_i)$$

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 $f \in \mathcal{F}, \epsilon > 0$:

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

...so now annealed risk on left replaced by actual risk (symmetric result)

Proof: simply plug previous result into Zhang!

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 $f \in \mathcal{F}, \epsilon > 0$:

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

...best case for strong central/critical $\bar{\eta}$: $O(KL/n)$ bounds

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 $f \in \mathcal{F}, \epsilon > 0$:

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

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

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

Fast vs. Slow Excess Risk Rates

- Convergence Rate of order $\sqrt{\text{COMP}/n}$ called **slow rate** in machine learning theory
- Convergence Rate of order COMP/n called **fast rate** in machine learning theory
- G-Mehta-Zhang Thm implies that slow rate can always be achieved for bounded losses
- Fast rate can be achieved under strong central
- Intermediate rates $(\text{COMP}/n)^{1/(1+\beta)}$ can be achieved under u $-$central with $u(\epsilon) = \epsilon^{\beta}$, $0 < \beta < 1$

The Fast Rate

- Fast Rate thus achieveable for log-loss, for well-specified ($\bar{\eta} = 1$) and convex models ($\bar{\eta} \geq 1$) and more generally ($\bar{\eta} > 0$) for misspecified models with ‘exponentially small loss tails’
- Strong central also holds, and fast rate therefore achieveable, for every mixable loss function as long as convex luckiness holds

Van Erven et al., 2015

The Fast Rate

- Strong central also holds, and fast rate therefore achieveable, for every mixable loss function as long as convex luckiness holds
  - log-loss, bounded range is mixable
  - every strongly convex loss is exp-concave. Every exp-concave loss is mixable
  - e.g. squared loss, bounded range is mixable; logistic loss (classification) is mixable

The Fast Rate

- Strong central also holds, and fast rate therefore achieveable, for every mixable loss function as long as convex luckiness holds

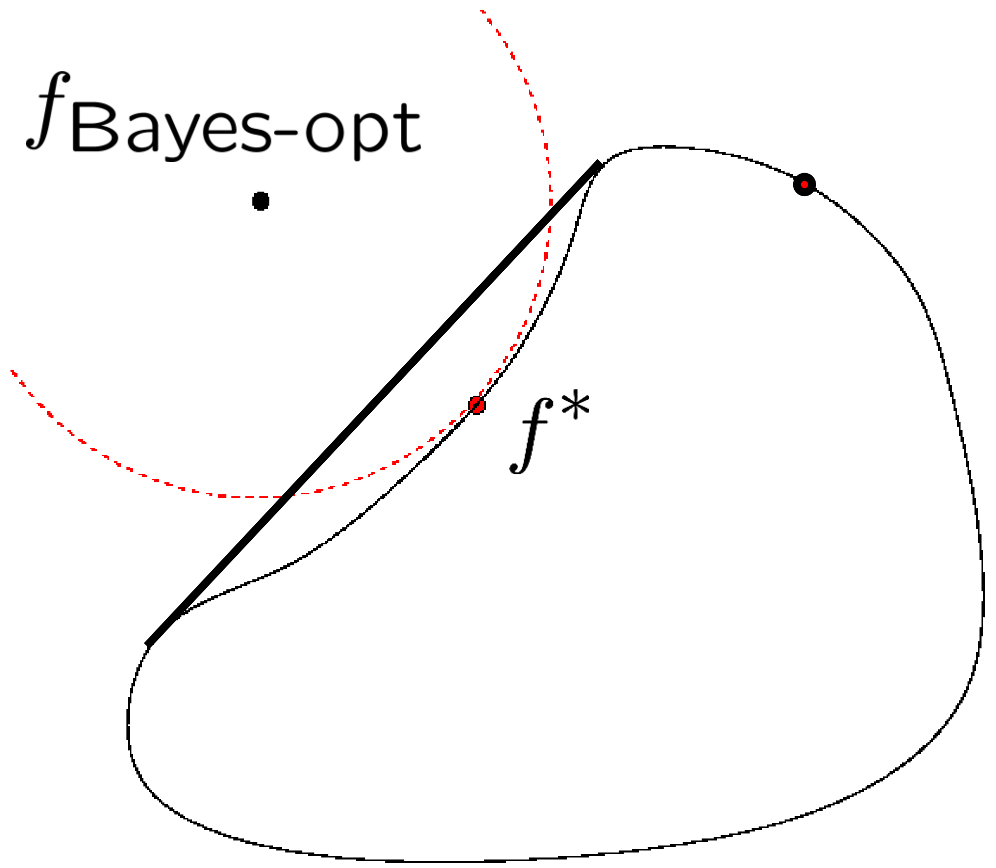
Convex Luckiness

- We say that **convex luckiness** holds if

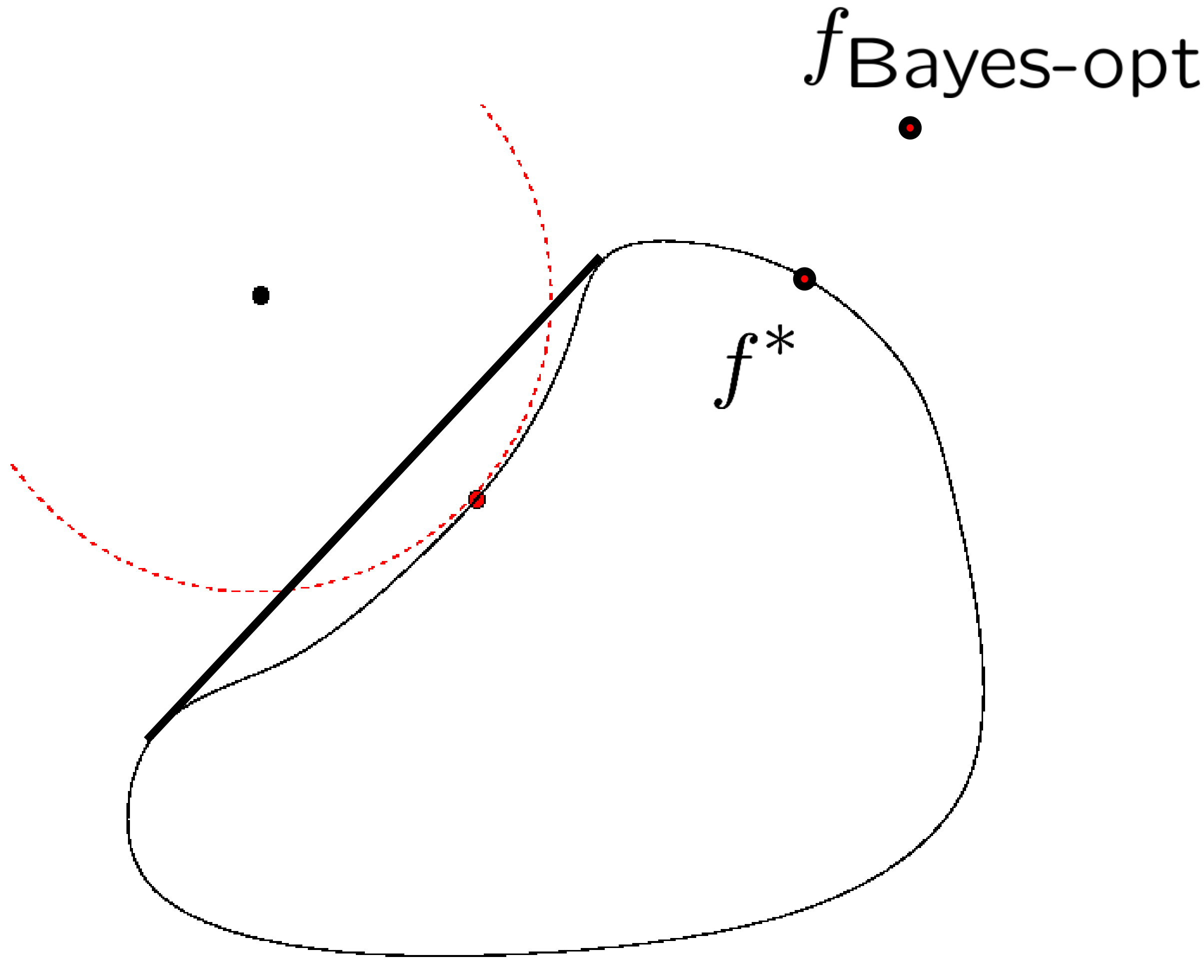
$$\inf_{f \in \mathcal{F}} \mathbf{E}_{Z \sim P}[\ell_f(Z)] = \inf_{f \in \text{CONV-HULL}(\mathcal{F})} \mathbf{E}_{Z \sim P}[\ell_f(Z)]$$

(Van Erven et al. ’15, G & Mehta ’17b)

Convexily Unlucky



Convexily **Lucky**



The Fast Rate

- Strong central also holds, and fast rate therefore achieveable, for every mixable loss function as long as convex luckiness holds
- Every convex loss satisfies convex luckiness as long as either
  - the set of predictors is convex, or
  - the Bayes predictor against P is contained in $\mathcal{F}$.

The Fast Rate

- Strong central also holds, and fast rate therefore achieveable, for every mixable loss function as long as convex luckiness holds
- Every convex loss satisfies convex luckiness as long as either
  - the set of predictors is convex, or
  - the Bayes predictor against P is contained in $\mathcal{F}$.

very strong condition for density estimation,

not so strong for some other losses

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)

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/Tsybakov often hold in realistic situations even if loss fn not convex! (classification loss)

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 $f \in \mathcal{F}, \epsilon > 0$:

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

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 $\Pi_{\mathbf{n}}$, every prior Π_0 every $f \in \mathcal{F}, \epsilon > 0$:

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

Can we also replace annealed excess risk on left by true excess risk in some unbounded cases?

Theorem (G. & Mehta, 2016)

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 $f \in \mathcal{F}, \epsilon > 0$:

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

Theorem (G. & Mehta, 2016)

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

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

and **Witness-of-Badness Condition** holds

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

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

Theorem (G. & Mehta, 2016)

Suppose risk (not loss) bounded and u-central holds

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

and **Witness-of-Badness Condition** holds

Then

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

Witness-of-Badness

There is $A, c > 0$ such that:

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

- automatically holds for bounded loss
- there should be no f that is extremely bad with extremely small probability
- Condition requires that we witness f's badness in the training set!
  - If we don't, learning does seem impossible...

Witness-of-Badness

There is $A, c > 0$ such that:

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

- automatically holds for bounded loss
- condition surprisingly weak: hold e.g. for squared error loss with convex $\mathcal{F}$ as long as **E** $[|Y|^3|X]$ bounded

Unbounded Loss: One-Sided Conditions

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** holds: there is $A, c > 0$ such that:

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

much weaker sort of tail-control of r_f

Then ... ...

Zhang-G-M with Witness/Central

Suppose **witness condition** 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 $f \in \mathcal{F}, \epsilon > 0$:

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

...result holds with annealed excess risk (generalized Hellinger) replaced by real excess risk (KL divergence)

Left vs Right Zhang

- G & Mehta, 2016 is about extending left-hand side of Zhang's Theorem
  - central, witness, fast rate conditions etc.
- G & Mehta, 2017a is about extending the right-hand side!
  - Relation to data compression, "really complex" models, etc.

Some History

- The oldest precursor of the Zhang-G-M bound is probably Barron & Cover (1991), *Minimum Complexity Density Estimation:* log-loss, Hellinger/Rényi on left, countable $\mathcal{F}$, in-probability
- Barron & Yang (‘98), Birgé & Massart (‘98) give tight bounds between KL and Hellinger/Rényi divergence if ratio of probability densities is bounded
- Wong & Shen (‘95) give condition under which ratio KL/Hellinger is bounded by log-factor for some unbounded cases
- **Witness** Condition/Theorem generalizes all these results to misspecification, general loss functions

extended to in-expectation by Barron (2000)

Some History

- The oldest precursor of the Zhang-G-M bound is probably Barron & Cover (1991), *Minimum Complexity Density Estimation:* log-loss, Hellinger/Rényi on left, countable $\mathcal{F}$, in-probability
- Barron & Yang ('98), Birgé & Massart ('98) give tight bounds between KL and Hellinger/Rényi divergence if ratio of probability densities is bounded
- Wong & Shen ('95) give condition under which ratio KL/Hellinger is bounded by log-factor for some unbounded cases
- **Witness** Condition/Theorem generalizes all these results to misspecification, general loss functions

Some History

- McAllester (1998) gives first PAC-Bayesian generalization bound with KL on the right (avoiding need for countable $\mathcal{F}$): with prob. at least $1-\delta$,

$$\mathbf{E}_{f\sim\Pi_n}\mathbf{E}_{Z\sim P}\left[r_f\right] \leq C_\eta \cdot \left(\mathbf{E}_{f\sim\Pi_n}\left[\frac{1}{n}\sum \ell_f(Z_i)\right] + \frac{\mathrm{KL}(\Pi_n \| \Pi_0) + C \cdot \log(1/\delta)}{\eta \cdot n} \right)$$

- Catoni ('03), Audibert ('04) give various extensions of this bound focusing on excess risk instead of generalization bounds
- Zhang (Ann. Stats' 06, IEEE Tr. Inf. Th. '06) is first to connect both strands of work into a single bound
- G&M add witness and u-central on the left, and also extensions on right

Rough Plan of Lectures

1. Safe Testing (Statistics/AB Testing)
2. Safe Testing (Information Theory!)
3. Safe and Generalized Bayes
   – Zhang-G.-Mehta Thm density estimation
4. Fast Rate Conditions in Statistical (stochastic) and Online (nonstochastic) Learning
   – Zhang-G-Mehta Thm general loss fns
5. Safety and Luckiness

Thank you for your attention!

Further Reading:

- Van Erven, G., Mehta, Reed, Williamson. Fast Rates in Statistical and Online Learning. *Journal of Machine Learning Research, 2015*
- G. and Van Ommen, *Bayesian Analysis, Dec. 2017*
- G. and Mehta, Fast Rates for Unbounded Losses, arXiv (2016)
- G. and Mehta. *A Tight Excess Risk bound in terms of a Unified PAC-Bayesian-Rademacher-MDL Complexity*, arXiv (2017)