Today

1. Complexity
   - Individual Sequence Prediction with Log-Loss: the NML distribution and Complexity
   - Extending the Right-Hand Side of Zhang’s Bound

2. Safe Probability, Safe Statistics

Three Complexity Notions

- Shtarkov or NML Complexity
  - central notion in nonstochastic log-loss individual sequence prediction.
- PAC-Bayesian Complexity
  - right-hand side in a strong excess risk bound in (stochastic) statistical learning for arbitrary loss fns
  - especially suited for (pseudo-) Bayesian methods but not for very large classes
- Rademacher Complexity
  - right-hand side in stochastic excess risk bound that deals well with large classes but not with log-loss and priors

The Shtarkov/MDL Complexity

- Minimax Cumulative Regret for Individual Sequence Prediction with Log Loss (Shtarkov '88, Rissanen '96), also known as Shtarkov complexity or MDL/stochastic complexity:

$$\mathcal{M} = \{P_\theta : \theta \in \Theta\}$$

$$\mathrm{comp}_n(\mathcal{M}) = \log \sum_{y^n \in \mathcal{Y}^n} p_{\hat{\theta}(y^n)}(y^n)$$

On-Line "Probabilistic" Prediction

- Consider sequence $y_1, y_2, \cdots$, all $y_i \in \mathcal{Y}$
- Goal: sequentially predict y_i given past $y^{i-1} = y_1, \ldots, y_{i-1}$ using a 'probabilistic prediction' P_i (distribution on $\mathcal{Y}$)
- prediction strategy S is function mapping, for all i, 'histories' $y_1, \ldots, y_{i-1}$ to distributions for i -th outcome

$$S : \cup_{n=1}^{\infty} \mathcal{Y}^n \rightarrow \text{set of distributions on } \mathcal{Y}$$

prediction strategy = distribution

- If we think that $Y_1, \ldots, Y_n \sim P$ (not necessarily i.i.d !) then should predict Y_i using conditional distribution

$$P(\cdot \mid y^{i-1}) := P(Y_i = \cdot \mid Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1})$$

- note that then joint probability mass/density is equal to the product of the predictions: $P(y^n) = \prod_{i=1}^{n} P(y_i \mid y^{i-1})$

prediction strategy = distribution

- If we think that $Y_1, \ldots, Y_n \sim P$ (not necessarily i.i.d !) then should predict Y_i using conditional distribution

$$P(\cdot \mid y^{i-1}) := P(Y_i = \cdot \mid Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1})$$

- note that then joint probability mass/density is equal to the product of the predictions: $P(y^n) = \prod_{i=1}^{n} P(y_i \mid y^{i-1})$

Conversely, every prediction strategy S may be thought of as a distribution on $(Y_1, \ldots, Y_n)$, by defining:

$$P(\cdot \mid y^{i-1}) := S(y^{i-1})$$
$$P(y_1, \ldots, y_n) := \prod_{i=1}^{n} P(y_i \mid y^{i-1})$$

Logarithmic Loss

- To compare performance of different prediction strategies, we need a measure of prediction quality
- One popular measure of quality is the log loss:

$$\text{loss}(y, P) := -\log_2 P(y)$$

$$\text{loss}(y_1 \ldots, y_n, S) := \sum_{i=1}^{n} \text{loss}(y_i, S(y_1, \ldots, y_{i-1}))$$

- corresponds to two important practical settings:
  - **data compression**: $\text{loss}(y_1 \ldots, y_n, S)$ is number of bits needed to encode $y_1, \ldots, y_n$ using code S
  - 'Kelly' gambling: loss = log capital growth factor

Log loss & likelihood

- For every "prediction strategy" P, all n,

$$\sum_{i=1}^{n} \mathsf{loss}(y_i, P(\cdot \mid y^{i-1})) = \sum_{i=1}^{n} -\log P(y_i \mid y^{i-1}) = -\log P(y_1, \ldots, y_n)$$

-

$$\sum_{i=1}^{n} -\log P(y_i \mid y^{i-1}) = -\log \prod_{i=1}^{n} P(y_i \mid y^{i-1}) = -\log \prod \frac{P(y_i)}{P(y^{i-1})} = -\log P(y_1, \ldots, y_n)$$

Log loss & likelihood

- For every "prediction strategy" P, all n,

$$\sum_{i=1}^{n} \text{loss}(y_i, P(\cdot \mid y^{i-1})) = \sum_{i=1}^{n} -\log P(y_i \mid y^{i-1}) = -\log P(y_1, \ldots, y_n)$$

- Accumulated log loss = minus log likelihood

Dawid '84, Rissanen '84

Universal Prediction

- Let $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ be a set of predictors (identified with probability distributions on $\mathcal{Y}^\infty$)
  - Simplest example: $\mathcal{M}$ is the Bernoulli model
  - Nonparametric example: $\mathcal{Y}$ is unit interval, $\mathcal{M}$ is set of all monotonically decreasing probability ensities

- GOAL: given $\mathcal{M}$, construct a new predictor predicting data 'almost as well' as any of the $P_\theta \in \mathcal{M}$ *no matter what data arrive* (a nonstochastic setting!)

Universal Prediction

- More concretely: find, for fixed n, the predictor P achieving the minimax cumulative log-loss regret

$$\min_{P} \left\{ \sup_{y^n \in \mathcal{Y}^n} \left(\text{loss}(y^n, P) - [\inf_{\theta \in \Theta} \text{loss}(y^n, P_\theta)] \right) \right\}$$

where $\text{loss}(y^n, Q) = \sum_{i=1}^{n} - \log Q(y_i \mid y^{i-1})$

- Solution was given by Shtarkov in 1988 (!)

Universal Prediction

- More concretely: find, for fixed n, the predictor P achieving the minimax cumulative log-loss regret

$$\min_P \left\{ \sup_{y^n \in \mathcal{Y}^n} \left(\text{loss}(y^n, P) - [\inf_{\theta \in \Theta} \text{loss}(y^n, P_\theta)] \right) \right\}$$

$$= \min_P \left\{ \sup_{y^n \in \mathcal{Y}^n} \left(-\log P(y^n) - [\inf_{\theta \in \Theta} -\log P_\theta(y^n)] \right) \right\}$$

$$= \min_P \left\{ \sup_{y^n \in \mathcal{Y}^n} \left(-\log P(y^n) + \log P_{\hat{\theta}(y^n)}(y^n) \right) \right\}$$

Universal Prediction

$$\min_{P} \left\{ \sup_{y^n \in \mathcal{Y}^n} \left(-\log P(y^n) + \log P_{\hat{\theta}(y^n)}(y^n) \right) \right\}$$

- uniquely achieved* by Shtarkov or NML (Normalized Maximum Likelihood) Distribution, given by

$$P_{\mathsf{nml}}(y^n) = \frac{P_{\hat{\theta}(y^n)}(y^n)}{\sum_{y^n \in \mathcal{Y}^n} P_{\hat{\theta}(y^n)}(y^n)}$$

- ...and its regret satisfies, for all $y^n \in \mathcal{Y}^n$,

$$-\log P_{\mathsf{nml}}(y^n) - [-\log P_{\hat{\theta}(y^n)}(y^n)] = \mathsf{comp}_n(\mathcal{M}) = \log \sum_{y^n \in \mathcal{Y}^n} p_{\hat{\theta} y^n}(y^n)$$

Complexity for Parametric Models

- So $$\mathrm{comp}_n(\mathcal{M}) = \log \sum_{y^n \in \mathcal{Y}^n} p_{\hat{\theta}(y^n)}(y^n)$$

is cumulative minimax regret relative to model $\mathcal{M}$

For d-dimensional exponential families with bounded density ratios (Rissanen '96, G. '07),

$$\mathrm{comp}_n(\mathcal{M}) = \frac{d}{2} \log \frac{n}{2\pi} + \log \int \sqrt{\det I(\theta)} + o(1) = O(\log n)$$

Complexity for Parametric Models

$$\mathsf{comp}_n(\mathcal{M}) = \frac{d}{2} \log \frac{n}{2\pi} + \log \int \sqrt{\det I(\theta)} + o(1) = O(\log n)$$

...whereas the Bayesian marginal likelihood

$$P_{\mathsf{Bayes}}(y^n) = \int P_\theta(y^n) w(\theta) d\theta$$

is known to satisfy*

$$-\log P_{\mathsf{Bayes}}(y^n) - [-\log P_{\hat{\theta}(y^n)}(y^n)] =$$

$$\frac{d}{2} \log \frac{n}{2\pi} - \log w(\theta) + \log \sqrt{\det I(\theta)} + o(1) = O(\log n)$$

Complexity for Parametric Models

$$\mathsf{comp}_n(\mathcal{M}) = \frac{d}{2}\log\frac{n}{2\pi} + \log\int\sqrt{\det I(\theta)} + o(1) = O(\log n)$$

...whereas the Bayesian marginal likelihood

$$P_{\mathsf{Bayes}}(y^n) = \int P_\theta(y^n)w(\theta)d\theta$$

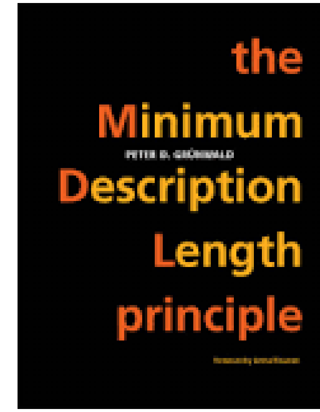
is known to satisfy*

$$-\log P_{\mathsf{Bayes}}(y^n) - [-\log P_{\hat{\theta}(y^n)}(y^n)] =$$

$$\frac{d}{2}\log\frac{n}{2\pi} - \log w(\theta) + \log\sqrt{\det I(\theta)} + o(1) = O(\log n)$$

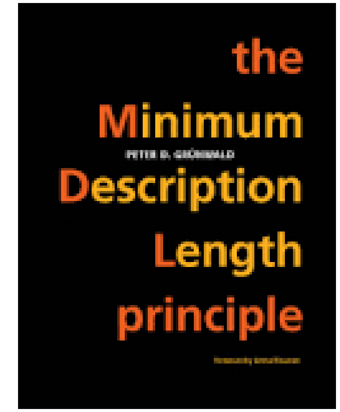
for Jeffreys' prior, $w(\theta) \propto \sqrt{\det I(\theta)}$ asymptotically same!

Aside



- In its simplest form, the MDL Principle (Rissanen, ’89) states that to compare 2 statistical models $\mathcal{M}_0, \mathcal{M}_1$ for the same data, one should associate them both with a lossless universal code (i.e. a code that gives small codelengths whenever ‘the model fits the data well’ ...)
- ... and then pick the model which allows for the shortest codelength of the data

- A lossless code is just a sequential log-loss prediction strategy... It is a good universal code if it has small regret

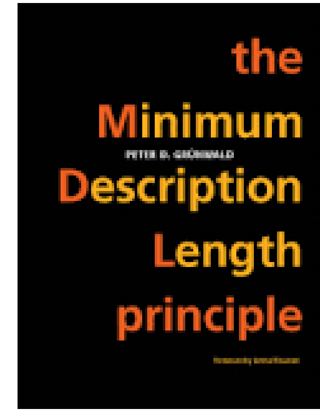


Aside

- pick the model $\mathcal{M}_j$ which allows for shortest codelength of data if encoded with good universal code
- A lossless code is just a sequential log-loss prediction strategy... it is a good universal code if it has small regret

- i.e. MDL tells you to pick $\mathcal{M}_1$ with 'confidence' $K > 0$ iff

$$-\log P_{\mathsf{nml}}(y^n \mid \mathcal{M}_1) - (-\log P_{\mathsf{nml}}(y^n \mid \mathcal{M}_0)) \leq -K$$



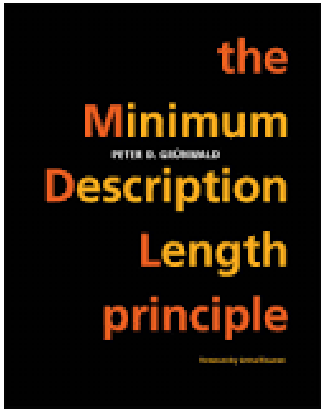
Aside

- pick the model $\mathcal{M}_j$ which allows for shortest codelength of data if encoded with good universal code
- A lossless code is just a sequential log-loss prediction strategy... it is a good universal code if it has small regret

- i.e. MDL tells you to pick $\mathcal{M}_1$ with 'confidence' $K > 0$ iff

$$-\log P_{\mathrm{nml}}(y^n \mid \mathcal{M}_1) - (-\log P_{\mathrm{nml}}(y^n \mid \mathcal{M}_0)) \leq -K$$

i.e. $$\frac{P_{\mathrm{nml}}(y^n \mid \mathcal{M}_1)}{P_{\mathrm{nml}}(y^n \mid \mathcal{M}_0)} \geq 2^K$$



Aside

- pick $\mathcal{M}_1$ with 'confidence' $K > 0$ iff

$$S = \frac{P_{\mathsf{nml}}(y^n \mid \mathcal{M}_1)}{P_{\mathsf{nml}}(y^n \mid \mathcal{M}_0)} \geq 2^K$$

- If null model is simple, then S is an S-value ($\mathbf{E}[S] \leq 1$)
- ... More generally, one also allows ratios of other P's that correspond to codes with small regret, such as Bayesian, 'prequential', 'switch'
- Ryabko & Monarev:

$$S = \frac{P_{\mathsf{gzip}}(y^n)}{P_0(y^n)}$$

Complexity for Parametric Models

$$\text{comp}_n(\mathcal{M}) = \frac{d}{2}\log\frac{n}{2\pi} + \log\int\sqrt{\det I(\theta)} + o(1) = O(\log n)$$

...whereas the Bayesian marginal likelihood

$$P_{\text{Bayes}}(y^n) = \int P_\theta(y^n)w(\theta)d\theta$$

is known to satisfy*

$$-\log P_{\text{Bayes}}(y^n) - [-\log P_{\hat{\theta}(y^n)}(y^n)] =$$

$$\frac{d}{2}\log\frac{n}{2\pi} - \log w(\theta) + \log\sqrt{\det I(\theta)} + o(1) = O(\log n)$$

for Jeffreys' prior, $w(\theta) \propto \sqrt{\det I(\theta)}$ asymptotically same!

Nonparametric Models

- Opper & Haussler ('96), Cesa-Bianchi & Lugosi ('01) and more recently Rakhlin and Sridharan ('15) gave bounds using chaining based on L_∞-covering nrs:

$$\mathsf{comp}_n(\mathcal{M}) \leq \inf_{\epsilon > 0} \log N_\infty(\mathcal{M}, \epsilon) + 24 \int_0^\epsilon \sqrt{\log N_\infty(\mathcal{M}, \delta)}\, d\delta$$

- If the model is i.i.d., then $N_\infty(\mathcal{M}, \epsilon)$ is ϵ-covering nr under metric $d(P, Q) = \sup_{y \in \mathcal{Y}} |-\log P(Y) + \log Q(Y)|$

Nonparametric Models

- Opper & Haussler ('96), Cesa-Bianchi & Lugosi ('01) and more recently Rakhlin and Sridharan ('15) gave bounds using chaining based on L_∞-covering nrs:

$$\mathrm{comp}_n(\mathcal{M}) \leq \inf_{\epsilon > 0} \log N_\infty(\mathcal{M}, \epsilon) + 24 \int_0^\epsilon \sqrt{\log N_\infty(\mathcal{M}, \delta)} d\delta$$

- If the model is i.i.d., then $N_\infty(\mathcal{M}, \epsilon)$ is ϵ-covering nr under metric $d(P,Q) = \sup_{y \in \mathcal{Y}} |-\log P(Y) + \log Q(Y)|$

- With this bound they obtained for variety of nonparametric models $\mathrm{comp}_n(\mathcal{M}) = O(n^\gamma)$

Two Observations

$$\mathrm{comp}_n(\mathcal{M}) \leq \inf_{\epsilon > 0} \log N_\infty(\mathcal{M}, \epsilon) + 24 \int_0^\epsilon \sqrt{\log N_\infty(\mathcal{M}, \delta)}\, d\delta$$

- Bound is often **better** than best regret bound that can be given for prediction by Bayes marginal likelihood (n^γ vs. n^β for $\beta > \gamma$)
  - ...and for some models it is indeed known that Bayesian prediction has larger worst-case regret
- ...yet bound is **void** if $N_\infty(\mathcal{M}, \epsilon) = \infty$

Two Observations

$$\mathrm{comp}_n(\mathcal{M}) \leq \inf_{\epsilon > 0} \log N_\infty(\mathcal{M}, \epsilon) + 24 \int_0^\epsilon \sqrt{\log N_\infty(\mathcal{M}, \delta)} d\delta$$

1. Bound is often **better** than best regret bound that can be given for prediction by Bayes marginal likelihood (n^γ vs. n^β for $\beta > \gamma$)
   - ...and for some $\mathcal{M}$ it is indeed known that Bayesian prediction has larger worst-case regret
2. ...yet bound is **void** if $N_\infty(\mathcal{M}, \epsilon) = \infty$
   - Take e.g. $\mathcal{M}$ to be all i.i.d. extensions of monotonically decreasing densities (bounded away from 0 and ∞) on unit interval

Two Complexity Notions, Two Results

- Shtarkov or NML Complexity
  - central notion in log-loss individual sequence prediction. Existing bounds are in terms of L_∞-entropy nrs; we have bound in terms of $L_{1/2.}(P)$ nrs.
- PAC-**Bayesian** Complexity
  - right-hand side in a strong excess risk bound in (stochastic) statistical learning for arbitrary loss fns; not suited for very large classes. We will unify with Shtarkov Complexity and thus make bound suitable for large classes.

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 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 $\eta > 0$:

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

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

Zhang's Excess Risk Bound

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

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

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

ℓ can be any loss function

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

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

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 $\eta > 0$:

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

$$-\frac{1}{\eta \cdot n} \cdot \log \frac{p'_{f,\eta}(Z^n)}{p'_{f^*,\eta}(Z^n)}$$

where $p'_{f,\eta}(z) = p(z) \cdot e^{-\eta r_f(z)} = p(z) \cdot e^{-\eta(\ell_f(z) - \ell^*_f(z))}$ are the 'entropified' probabilities we discussed earlier

Zhang's Excess Risk Bound

For every 'prior' Π_0 , every $0 < \eta$, for the generalized η-Bayesian posterior, every 'prior' Π_0 every $\eta > 0$:

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

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

Zhang's Excess Risk Bound

For every 'prior' Π_0 , every $0 < \eta$, for the generalized η-Bayesian posterior, every 'prior' Π_0 every $\eta > 0$:

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

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

Insight: excess risk bound in terms of the cumulative log-loss of a Bayesian prediction strategy

Two Observations

$$\mathrm{comp}_n(\mathcal{M}) \leq \inf_{\epsilon>0} \log N_\infty(\mathcal{M},\epsilon)+24\int_0^\epsilon \sqrt{\log N_\infty(\mathcal{M},\delta)}d\delta$$

1. Bound is often **better** than best regret bound that can be given for prediction by Bayes marginal likelihood (n^γ vs. n^β for $\beta > \gamma$)
   - ...and for some $\mathcal{M}$ it is indeed known that Bayesian prediction has larger worst-case regret

Recall: Two Complexity Notions

- Shtarkov or NML Complexity
  - central notion in log-loss individual sequence prediction
- PAC-**Bayesian** Complexity
  - right-hand side in a strong excess risk bound in (stochastic) statistical learning for arbitrary loss fns; not suited for very large classes. We will unify with Shtarkov Complexity and thus make bound suitable for large classes.

G & M Excess Risk Bound (Thm)

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

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

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

G & M Excess Risk Bound

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

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

G & M Excess Risk Bound

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

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

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

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

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 $\eta > 0$:

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

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

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

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 $w \equiv 1$, $\eta > 0$:

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

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

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

data-dependent part disappears

Excess Risk Bound for ERM

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

Excess Risk Bound for ERM

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

...to define S, define probability density fns q_f as

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

[note that with log-loss and $\eta = 1$ and a correctly specified model, $q_f(z) = p_f(z)$!]

Then

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

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

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

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

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

G & M Excess Risk Bound

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

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

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

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

G & M Excess Risk Bound

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

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

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

G & M Excess Risk Bound

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

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

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

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

G & M Excess Risk Bound

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

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

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

...and now

$$S(\mathcal{F}, \hat{f}, w) := \int q_{\hat{f}_{|z^n}}(z^n) w(z^n) d\nu(z^n)$$

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 :

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

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

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)

Now we reduce to Zhang...

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

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

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

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

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

Excess Risk $\leq$ Codelength Diff.

- If we estimate by 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
- If we take deterministic $\hat{f}$ and nonconstant w then RHS has a 'luckiness NML' (Bartlett et al. 2013) codelength interpretation

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

General insight: **right-hand side of bound always has a codelength interpretation**, different w's corresponding to different codes

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

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

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

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

Two Observations

$$\mathrm{comp}_n(\mathcal{M}) \leq \inf_{\epsilon>0} \log N_\infty(\mathcal{M},\epsilon) + 24 \int_0^\epsilon \sqrt{\log N_\infty(\mathcal{M},\delta)}\, d\delta$$

1. Bound is often **better** than best regret bound that can be given for prediction by Bayes marginal likelihood (n^γ vs. n^β for $\beta > \gamma$)
   - ...and for some $\mathcal{M}$ it is indeed known that Bayesian prediction has larger worst-case regret
2. ...yet bound is **void** if $N_\infty(\mathcal{M},\epsilon) = \infty$
   - Take e.g. $\mathcal{M}$ to be all i.i.d. extensions of monotonically decreasing densities (bounded away from 0 and ∞) on unit interval

Two Complexity Notions, Two Results

- Shtarkov or NML Complexity
  - central notion in log-loss individual sequence prediction. Existing bounds are in terms of L_∞-entropy nrs; we have comparable bound in terms of $L_{1/2.}(P)$ nrs. (but haven't shown you)
- PAC-**Bayesian** Complexity
  - right-hand side in a strong excess risk bound in (stochastic) statistical learning for arbitrary loss fns with Bayesian codelength interpretation; not suited for very large classes. We have unified with Shtarkov Complexity (smaller codelengths) and thus made bound suitable for large classes.

Three Complexity Notions

- Shtarkov or NML Complexity
  - central notion in nonstochastic log-loss individual sequence prediction.
- PAC-Bayesian Complexity
  - right-hand side in a strong excess risk bound in (stochastic) statistical learning for arbitrary loss fns
  - especially suited for (pseudo-) Bayesian methods but not for very large classes
- Rademacher Complexity
  - right-hand side in stochastic excess risk bound that deals well with large classes but not with log-loss and priors

Thm 2: Shtarkov bounded by Rademacher Complexity

- 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 $\hat{f}$,

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

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

where ϵ is diameter of $\mathcal{F}$ in $L_2(P)$-pseudometric

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

Bounding excess risk, minimax regret in terms of L_2 entropy nrs

- Recall Lugosi/Cesa-Bianchi log-loss result:

$$\mathrm{COMP}_1(\mathcal{F}, \hat{f}, w_{\mathrm{UNIFORM}}) \leq$$

$$\inf_{\epsilon>0} \log N_\infty(\mathcal{F}, \epsilon) + 24 \int_0^\epsilon \sqrt{\log N_\infty(\mathcal{F}, \delta)}\, d\delta$$

- Via existing bounds on Rademacher using chaining we get

$$\mathrm{COMP}_\eta(\mathcal{F}, \hat{f}, w_{\mathrm{UNIFORM}}) \leq$$

$$\inf_{\epsilon>0} \log N_{L_2(P)}(\mathcal{F}, \epsilon) + 24 \int_0^\epsilon \sqrt{\log N_{L_2(P)}(\mathcal{F}, \delta)}\, d\delta + Cn\eta\epsilon^2$$

For class of monotone decreasing densities, now get $O(n^{1/3})$ which is tight; previous bound was void

Today

1. Complexity
   - Individual Sequence Prediction with Log-Loss: the NML distribution and Complexity
   - Extending the Right-Hand Side of Zhang's Bound

2. Safe Inference

Safe Bayes, Safe Probability

- In previous work, I used phrase ‘safe Bayes’ in two senses:
  1. Specific algorithm for learning η from the data (‘G. ‘12, The Safe Bayesian; G. and vOmmen ‘17)

  2. General idea that in practice probabilities should not be taken fully seriously; their application should be restricted to **safe** uses

     (G., Safe Probability, JSPI ‘18)

Two Extreme Views on Learning – yet using almost same methods

- **Vapnik's ML Theory**
  ('statistical learning theory', 50000 citations)

Can only do one single thing with the function learned from data

- **Bayesian Inference (at least De Finetti brand)**

Every single inference task that can be formulated in terms of measurable fns on my domain can be answered by my posterior

Two Extremist Views on Learning – yet using almost same methods

- **Vapnik's ML Theory**
  ('statistical learning theory', 50000 citations)

Can only do one single thing with the function I learned from data

- **Bayesian Inference (at least De Finetti brand)**

Every single inference task that can be formulated in terms of measurable fns on my domain can be answered by my posterior

Example: Ridge/Lasso Regression

$$\widehat{\beta}_n := \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \lambda \|\beta\|_2^2$$

- V: assume X_i, Y_i i.i.d.$\sim P$ .For large enough n, 'right' λ, we have

$$\mathbf{E}_{(X,Y)\sim P}(Y - \widehat{\beta}_n^T X)^2 \approx \min_{\beta \in \mathbb{R}^k} \mathbf{E}_{(X,Y)\sim P}(Y - \beta^T X)^2$$

- "Hence I can get small squared error when predicting a new Y based on a new X from the same distribution"

$$\hat{\beta}_n := \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \lambda \|\beta\|_2^2$$

- V: assume X_i, Y_i i.i.d.$\sim P$ .For large enough n, 'right' λ, we have

$$\mathbf{E}_{(X,Y)\sim P}(Y - \hat{\beta}_n^T X)^2 \approx \min_{\beta \in \mathbb{R}^k} \mathbf{E}_{(X,Y)\sim P}(Y - \beta^T X)^2$$

- "Hence I can get small squared error when predicting a new Y based on a new X from the same distribution"
- Q: What if new X drawn from different distribution?
- V: You can't say anything!

$$\widehat{\beta}_n := \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \lambda \|\beta\|_2^2$$

- V: assume X_i, Y_i i.i.d. $\sim P$ .For large enough n, 'right' λ, we have

$$\mathbf{E}_{(X,Y)\sim P}(Y - \widehat{\beta}_n^T X)^2 \approx \min_{\beta \in \mathbb{R}^k} \mathbf{E}_{(X,Y)\sim P}(Y - \beta^T X)^2$$

- "Hence I can get small squared error when predicting a new Y based on a new X from the same distribution"
- Q: What if new X drawn from different distribution?
- V: You can't say anything!
- Q: Does $\widehat{\beta}_n^T X$ give a good estimate of $\mathbf{E}[Y|X]$?
- V: Can't say!

$$\widehat{\beta}_n := \arg\min_{\beta \in \mathbb{R}^k} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \frac{\lambda}{\sigma^2} \|\beta\|_2^2$$

- B: $\hat{\beta}_n$ is also posterior mean (even with prior on σ^2)
- So I agree that I can get small squared error when predicting a new Y based on a new X from same distr.
- Q: What if new X drawn from different distribution?
- B: You'll still be o.k.!
- Q: Does $\widehat{\beta}_n^T X$ give a good estimate of $\mathbf{E}[Y|X]$?
- B: Of course!

$$\widehat{\beta}_n := \arg\min_{\beta \in \mathbb{R}^k} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \frac{\lambda}{\sigma^2} \|\beta\|_2^2$$

- B: $\hat{\beta}_n$ is also posterior mean (even with prior on σ^2)
- So I agree that I can get small squared error when predicting a new Y based on a new X from same distr.
- Q: What if new X drawn from different distribution?
- B: You'll still be o.k.!
- Q: Does $\widehat{\beta}_n^T X$ give a good estimate of $\mathbf{E}[Y \mid X]$?
- B: Of course!
- Q: Does $\widehat{\beta}_n^T X$ give good estimate of median of Y given X?
- B: Of course!
- Q: Is $P(Y|X)$ unimodal? B: Of course! etc etc

V&B use almost same method but draw very weak vs very strong conclusions!

$$\widehat{\beta}_n := \arg\min_{\beta \in \mathbb{R}^k} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \frac{\lambda}{\sigma^2} \|\beta\|_2^2$$

- B: $\hat{\beta}_n$ is also posterior mean (even with prior on σ^2)
- So I agree that I can get small squared error when predicting a new Y based on a new X from same distr.
- Q: What if new X drawn from different distribution?
- B: You'll still be o.k.!
- Q: Does $\widehat{\beta}_n^T X$ give a good estimate of $\mathbf{E}[Y|X]$?
- B: Of course!
- Q: Does $\widehat{\beta}_n^T X$ give good estimate of median of Y given X?
- B: Of course!
- Q: Is $P(Y|X)$ unimodal? B: Of course! Etc etc
-

Safe Statistics: Go Inbetween

- If I do η –Bayesian linear regression with normal prior on β , standard prior on variance σ^2 and $\eta < \bar{\eta}$, then if data i.i.d. I can guarantee convergence to KL optimal $f^*(x) = \beta^{*T} x$ and σ^* s.t.:
  - **Optimality** of squared error predictions of p_{f^*}

$$\mathbf{E}_{(X,Y)\sim P}\left[(Y - f^*(X))^2\right] = \min_{f\in\mathcal{F}} \mathbf{E}_{(X,Y)\sim P}\left[(Y - f(X))^2\right]$$

  - **Safety** of your error assessment thereof

$$\mathbf{E}_{Y\sim p_{f^*}}\left[(Y - f^*(X))^2 \mid X\right] = \sigma_2^* = \mathbf{E}_{(X,Y)\sim P}\left[(Y - f^*(X))^2\right]$$

Safe Statistics: Go Inbetween

- If I assume data i.i.d. I can guarantee
- **Optimality** of squared error predictions of p_{f^*}

- **Safety** of error assessment thereof

- If(f) I am further willing to assume that $\mathcal{F}$ contains Bayes-optimal decision rule...

$$\arg\min_{f:\mathcal{X}\to\mathbb{R}} \mathbf{E}_{(X,Y)\sim P}(Y - f(X))^2$$

- ....then I can guarantee that $f^*(X) = \mathbf{E}[Y \mid X]$
- If on top I want to assume that $P(Y|X)$ is symmetric then I can guarantee that $f^*(X)$ is median of $P(Y \mid X)$

I have a Dream

- Imagine a world in which statisticians/data analysts would, as a matter of principle, be asked to express what their probability model can be used for and what not.
- Then indeed we would have a safer statistics
- ...in the paper 'Safe Probability' I make a first attempt to develop a formal language for specifying this

Hypothesis Testing

- Suppose you test between two models using a Bayes factor
- If you choose $\bar{p}_0(y^n) = \int p_\theta(y^n) w_0(\theta) d\theta$ because your prior w_0 really expresses prior knowledge, and $\bar{p}_0(y^n) \gg \bar{p}_1(y^n)$, then you might be willing to use the Bayes posterior $w_0(\theta|y^n)$ for making actual predictions: you might claim it is safe for all bounded loss fns.
- But if you choose $\bar{p}_0$ because it is the RIPr of $\bar{p}_1$, then you definitely cannot trust the poster and you do not want to make such claims!

New Mathematical Questions/Concepts

- Optimality: If I assume <X>, for what inference/prediction tasks am I (sufficiently) optimal?

- Some scattered nontrivial results exist in machine learning theory literature.

New Mathematical Questions/Concepts

- Optimality: If I assume <X>, for what inference/prediction tasks am I (sufficiently) optimal?

- Some scattered nontrivial results exist in machine learning theory literature. For example:

  if you do logistic regression and you are really interested in classification, then your KL optimal parameters (to which you'll converge) also give you the smallest expected 0/1-loss when used for classification *if* your model contains the Bayes optimal classifier (Bartlett, Jordan, McAullife '06)

New Mathematical Questions/Concepts

- Optimality: If I assume <X>, for what inference/prediction tasks am I (sufficiently) optimal?
- Safety: central concept of G. 2018.

A distribution $\tilde{P}$ is safe for predicting against loss function L with 'true' distribution P if it holds that

$$\mathbf{E}_{Z\sim P}\left[L(Z,\delta_{\tilde{P}})\right] = \mathbf{E}_{Z\sim \tilde{P}}[L(Z,\delta_{\tilde{P}})]$$

where $\delta_{\tilde{P}}$ is the Bayes act according to $\tilde{P}$

Safe Probability

- Safety: Simplest form:

A distribution $\tilde{P}$ is safe for predicting against loss function L with 'true' distribution P if it holds that

$$\mathbf{E}_{Z\sim P}\left[L(Z,\delta_{\tilde{P}})\right] = \mathbf{E}_{Z\sim \tilde{P}}[L(Z,\delta_{\tilde{P}})]$$

where $\delta_{\tilde{P}}$ is the Bayes act according to $\tilde{P}$

If you act as your model prescribes, the world behaves as your model predicts, even though your model may be wrong and there may be better predictions!

Example

- If I do η –Bayesian linear regression with normal prior on β , standard prior on variance σ^2 and $\eta < \bar{\eta}$, then if data i.i.d. I can guarantee convergence to KL optimal $f^*(x) = \beta^{*T}x$ and σ^* s.t.:
  - **Optimality** of squared error predictions of p_{f^*}

$$\mathbf{E}_{(X,Y)\sim P}\left[(Y - f^*(X))^2\right] = \min_{f\in\mathcal{F}} \mathbf{E}_{(X,Y)\sim P}\left[(Y - f(X))^2\right]$$

  - **Safety** of your error assessment thereof

$$\mathbf{E}_{Y\sim p_{f^*}}\left[(Y - f^*(X))^2 \mid X\right] = \sigma_2^* = \mathbf{E}_{(X,Y)\sim P}\left[(Y - f^*(X))^2\right]$$

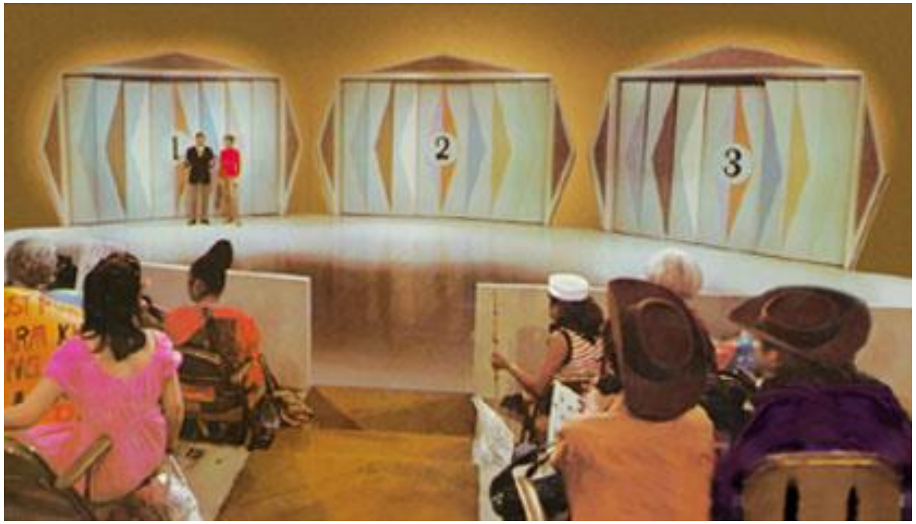
Example 2

- The Weather Forecaster!

Monty Hall (3-door) Problem

Monty Hall 1970

Monty Hall



- There are three doors in the TV studio. Behind one door is a car, behind both other doors a goat. You choose one of the doors. Monty Hall opens one of the other two doors, and shows that there is a goat behind it. You are now allowed to switch to the other door that is still closed. Is it smart to switch?

Monty Hall: The Wikipedia Wars

- I am interested in understanding the Wikipedia Wars (Gill 11, Mlodinow 08) on Monty Hall
  - Both sides agree that switching is smart and increases your chances of winning from 1/3 to 2/3!
  - The “war” is about how to *prove* this:
    - “strictly Bayesian”: via conditioning (in the right space, with additional assumption that Monty chooses by tossing a fair coin) (“MaxEnt-style assumption”)
    - Without additional assumptions, via decision-theoretic argument

Safe Probability applied to Monty Hall

- Under a **symmetric** loss function as in the original formulation of the problem, assuming that Monty flips a fair coin if he has a choice and then conditioning is ***safe*** and ***minimax optimal***
  - 'asymmetric' means e.g. that if the car is behind door B, it is a Ferrari; if it is behind door C, it is a Fiat Panda
- Still holds if same candidate plays each week and can reinvest his prize, hedging over several doors (**horse race losses**), even if prizes asymmetric
- But *not* if loss functions are asymmetric and reinvestment impossible

Thank you!

Further Reading:

- G. and T. van Ommen, Inconsistency of Bayesian Inference for Misspecied Linear Models, and a Proposal for Repairing It. *Bayesian Analysis, Dec. 2017*
- G. and N. Mehta, Fast Rates for Unbounded Losses, arXiv (2016)
- G. and N. Mehta. A Tight Excess Risk bound in terms of a Unified PAC-Bayesian-Rademacher-MDL Complexity, arXiv (2017)
- G. Safe Probability, *Journal of Stat. Planning and Inference,* 2018
- T. van Ommen, W. Koolen and G. *Robust Probability Updating, Intern. Journ. of Approx. Reasoning*, 2016