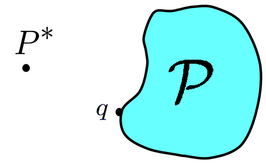
Safe Learning

CWI Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam
Mathematical Institute – Leiden University

Universiteit Leiden

Model Misspecification



Menu

1. Bayesian inconsistency under misspecification
   - G. and Langford, Machine Learning J. 2007
2. Learning Rate - Relation to Convexity, PAC-Bayes
3. Sequential Prediction Detour
   - **paradox:** Bayesian posterior good and bad at same time
4. The Safe Bayesian Algorithm
   - Use optimal learning rate, itself "learned" from data
5. "Unifying" Bayes and PAC-Bayes

Setting of Inconsistency Result

- Let $\mathcal{X} = [0,1], \mathcal{Y} = \{0,1\}$ (classification setting)
- Let $\mathcal{P}$ be a set of conditional distributions $P_{Y|X}$, and let Π be a prior on $\mathcal{P}$

Bayesian Consistency

- Let $\mathcal{X} = [0,1], \mathcal{Y} = \{0,1\}$ (classification setting)
- Let $\mathcal{P}$ be a set of conditional distributions $P_{Y|X}$, and let Π be a prior on $\mathcal{P}$
- Let P^* be a distribution on $\mathcal{X} \times \mathcal{Y}$
- Let $(X_1, Y_1), (X_2, Y_2), \ldots$ i.i.d. $\sim P^*$
- If $P^*_{Y|X} \in \mathcal{P}$, then Bayes is **consistent** under very mild conditions on Π and $\mathcal{P}$

Bayesian Consistency

- Let $\mathcal{X} = [0,1], \mathcal{Y} = \{0,1\}$ (classification setting)
- Let $\mathcal{P}$ be a set of conditional distributions $P_{Y|X}$, and let Π be a prior on $\mathcal{P}$
- Let P^* be a distribution on $\mathcal{X} \times \mathcal{Y}$
- Let $(X_1, Y_1), (X_2, Y_2), \ldots$ i.i.d. $\sim P^*$
- If $P^*_{Y|X} \in \mathcal{P}$, then Bayes is **consistent** under very mild conditions on Π and $\mathcal{P}$
  – "consistency" can be defined in number of ways, e.g. posterior distribution $\Pi(\cdot \mid X^n, Y^n)$ "concentrates" on "neighborhoods" of P^*

Bayesian Consistency

- Let $\mathcal{X} = [0,1], \mathcal{Y} = \{0,1\}$ (classification setting)
- Let $\mathcal{P}$ be a set of conditional distributions $P_{Y|X}$, and let Π be a prior on $\mathcal{P}$
- Let P^* be a distribution on $\mathcal{X} \times \mathcal{Y}$
- Let $(X_1, Y_1), (X_2, Y_2), \ldots$ i.i.d. $\sim P^*$
- If $P^*_{Y|X} \in \mathcal{P}$, then Bayes is **consistent** under very mild conditions on Π and $\mathcal{P}$
- If $P^*_{Y|X} \notin \mathcal{P}$, then Bayes is **consistent** under mild conditions on Π and $\mathcal{P}$

see e.g. Kleijn and Van der Vaart 2006

Bayesian Consistency

- Let $\mathcal{X} = [0,1], \mathcal{Y} = \{0,1\}$ (classification setting)
- Let $\mathcal{P}$ be a set of conditional distributions $P_{Y|X}$, and let Π be a prior on $\mathcal{P}$
- Let P^* be a distribution on $\mathcal{X} \times \mathcal{Y}$
- Let $(X_1, Y_1), (X_2, Y_2), \ldots$ i.i.d. $\sim P^*$
- If $P^*_{Y|X} \in \mathcal{P}$, then Bayes is **consistent** under very mild conditions on Π and $\mathcal{P}$
- If $P^*_{Y|X} \notin \mathcal{P}$, then Bayes is **consistent** under mild conditions on Π and $\mathcal{P}$

but not nearly so mild!

Bayesian Inconsistency

- Let $\mathcal{X} = [0,1], \mathcal{Y} = \{0,1\}$
- Let $\mathcal{P}$ be a set of conditional distributions $P_{Y|X}$, and let Π be a prior on $\mathcal{P}$ such that $\pi(Q) > 0$
- Let $(X_1, Y_1), (X_2, Y_2), \ldots$ i.i.d. $\sim P^*$
- $D(P^* \| Q) = \min_{P \in \mathcal{P}} D(P^* \| P) > 0$
- Here D is (conditional) KL divergence: $D(P^* \| P) = E_{X,Y \sim P^*}\left[-\log \frac{p(Y \mid X)}{p^*(Y \mid X)}\right]$

"Theorem", G. & Langford 2007

- Let $\mathcal{X} = [0,1], \mathcal{Y} = \{0,1\}$
- Let $\mathcal{P}$ be a set of conditional distributions $P_{Y|X}$, and let Π be a prior on $\mathcal{P}$ such that $\pi(Q) > 0$
- Let $(X_1, Y_1), (X_2, Y_2), \ldots$ i.i.d. $\sim P^*$
- $D(P^* \| Q) = \min_{P \in \mathcal{P}} D(P^* \| P) > 0$

For all $K > 0$ there exist $(\mathcal{P}, P^*, Q)$ satisfying these conditions such that P^*-a.s., we have

$$\lim_{n \to \infty} \Pi(\{P : D(P^* \| P) > D(P^* \| Q) + K\} \mid X^n, Y^n) = 1$$

"Theorem", G. & Langford 2007

very different from Diaconis-Freedman!

- Let $\mathcal{X} = [0,1], \mathcal{Y} = \{0,1\}$
- Let $\mathcal{P}$ be a set of conditional distributions $P_{Y|X}$, and let Π be a prior on $\mathcal{P}$ such that $\pi(Q) > 0$
- Let $(X_1, Y_1), (X_2, Y_2), \ldots$ i.i.d. $\sim P^*$
- $D(P^* \| Q) = \min_{P \in \mathcal{P}} D(P^* \| P) > 0$

For all $K > 0$ there exist $(\mathcal{P}, P^*, Q)$ satisfying these conditions such that P^*-a.s., we have

$$\lim_{n \to \infty} \Pi(\{P : D(P^* \| P) > D(P^* \| Q) + K\} \mid X^n, Y^n) = 1$$

Root of the Problem

- Consider very simple case $\mathcal{P} = \{P_{\text{bad}}, P_{\text{good}}\}$
- Let $Y_1, Y_2, \ldots$ i.i.d. $\sim P^*$
- Then (Markov's inequality)

$$P^*\left(\frac{p_{\text{bad}}(Y^n)}{p_{\text{good}}(Y^n)} > K\right) \leq \inf_{\lambda > 0} \frac{1}{K^\lambda} \left(E_{P^*}\left(\frac{p_{\text{bad}}(Y_1)}{p_{\text{good}}(Y_1)}\right)^\lambda \right)^n$$

- If $p^* = p_{\text{good}}$ already get interesting bound for $\lambda = 1$ since then generalized Hellinger affinity

$$A(\lambda) := E_{P^*}\left(\frac{p_{\text{bad}}(Y_1)}{p_{\text{good}}(Y_1)}\right)^\lambda = 1 \text{ at } \lambda = 1, \text{ and strictly increasing}$$

Root of the Problem

- Consider very simple case $\mathcal{P} = \{P_{\text{bad}}, P_{\text{good}}\}$
- Let $Y_1, Y_2, \ldots$ i.i.d. $\sim P^*$
- Then (Markov's inequality)

$$P^*\left(\frac{p_{\text{bad}}(Y^n)}{p_{\text{good}}(Y^n)} > K\right) \leq \inf_{\lambda>0} \frac{1}{K^\lambda}\left(E_{P^*}\left(\frac{p_{\text{bad}}(Y_1)}{p_{\text{good}}(Y_1)}\right)^\lambda\right)^n$$

- If $p^* = p_{\text{good}}$ already get interesting bound for $\lambda = 1$ since then generalized Hellinger affinity

$$A(\lambda) := E_{P^*}\left(\frac{p_{\text{bad}}(Y_1)}{p_{\text{good}}(Y_1)}\right)^\lambda = 1 \text{at } \lambda = 1, \text{ and strictly increasing}$$

- Yet if $D(P^* \| P_{\text{bad}}) > D(P^* \| P_{\text{good}}) > 0$ then may have $A(1) > 1$

Bound becomes worthless (exp. large) for all but very small λ

Root of the Problem

- If $\mathcal{P}$ finite or "regular parametric", then for large n get consistency anyway by uniform law of large numbers

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \frac{1}{n} \log \frac{p(Y^n)}{q(Y^n)} \rightarrow E_{P^*}\left[\log \frac{p(Y)}{q(Y)}\right]$$

- G. & Langford '07 give countably infinite $\mathcal{P}$ with
  - no uniform convergence; **in**consistency
  - **relevant** since in practice we often do apply Bayes in nonparametric situations without uniform convergence/optimal convergence rate depends on underlying degree of "smoothness"

Possible Solutions

- Let q achieve $\inf_{P\in\langle\mathcal{P}\rangle} D(P^* \| P)$
- It turns out that, for convex $\langle\mathcal{P}\rangle$ for all $P \in \mathcal{P}$

$$A(\lambda, p) := E_{P^*}\left(\frac{p(Y}{q(Y)}\right)^\lambda \leq 1 \text{ at } \lambda = 1, \text{ and strictly increasing}$$

- ...so indeed o.k. if we restrict to convex models (Barron & Li '99, Kleijn and v.d. Vaart '06)
- But we often *want* to use nonconvex models (e.g. regression)!

What to do for nonconvex models?

- Let $\eta_{\text{crit}} > 0$ be largest η such that $\sup_{P\in\mathcal{P}} E_{P^*}\left(\frac{p(Y)}{q(Y)}\right)^\eta \leq 1$
- "scale down" model
  ...by defining "generalized posterior"
  (Vovk, Zhang, Hjort, Walker, Barron, G.)

$$\pi(p \mid Y^i, \eta) := \frac{\pi(p)p^\eta(Y^i)}{\sum_{p\in\mathcal{P}} \pi(p)p^\eta(Y^i)}$$

- and do Bayesian inference for $\eta < \eta_{\text{crit}}$
- This works, but of course we don't know η_{crit} ....

Interpretation of Generalized Posterior

- In case of regression, decreasing η simply means increasing the variance of the model

$$\pi(p \mid X^n, Y^n, \eta) := \frac{\pi(p)p^\eta(Y^n \mid X^n)}{\sum_{p\in\mathcal{P}} \pi(p)p^\eta(Y^n \mid X^n)}$$

$$\pi(h \mid X^n, Y^n, \eta) = \frac{\pi(dh)e^{-\eta\sum_{i=1}^n (Y_i - h(X_i))^2}}{\int_{h'\in\mathcal{H}} \pi(dh')e^{-\eta\sum_{i=1}^n (Y_i, h'(X_i))^2}}$$

Interpretation of Generalized Posterior

- In case of regression, decreasing η simply means increasing the variance of the model

$$\pi(p \mid X^n, Y^n, \eta) := \frac{\pi(p)p^\eta(Y^n \mid X^n)}{\sum_{p\in\mathcal{P}} \pi(p)p^\eta(Y^n \mid X^n)}$$

- In general though interpretation not so easy
  - can get super- and sub-probabilities
  - for exponential families

$$\eta \leq \eta_{\text{crit}} \Rightarrow \text{COV}_{P^*}[X] \preceq \text{COV}_{Q|\eta}[X]$$

But in general converse only holds 'locally'

Interpretation of Generalized Posterior

- In case of regression, decreasing η simply means increasing the variance of the model

$$\pi(p \mid X^n, Y^n, \eta) := \frac{\pi(p)p^{\eta}(Y^n \mid X^n)}{\sum_{p \in \mathcal{P}} \pi(p)p^{\eta}(Y^n \mid X^n)}$$

- In general though interpretation not so easy
- What does hold in general: the smaller η the larger the weight of the prior/regularization term in MAP

$$\hat{p}_{\text{map}} = \arg\min_{p \in \mathcal{P}} \left(\frac{1}{\eta} \cdot (-\log \pi(p)) - \log p(Y^n \mid X^n) \right)$$

PAC-Bayes: beyond Log-Loss

- Let loss : $\mathcal{Y} \times \mathcal{A} \rightarrow [0, \infty]$ be arbitrary loss fn.
- Define generalized posterior on set of predictors $\mathcal{H}$ as

$$\pi(h \mid Z^n, \eta) = \frac{\pi(dh) e^{-\eta \sum_{i=1}^n \text{loss}(Y_i, h(X_i))}}{\int_{h' \in \mathcal{H}} \pi(dh') e^{-\eta \sum_{i=1}^n \text{loss}(Y_i, h'(X_i))}}$$

McAllester '02, Seeger 02, Audibert '04, Zhang '06, Catoni '07

- With log-loss this reduces to original generalized posterior ; most often used for 0/1-loss

PAC-Bayes: beyond Log-Loss

- Define generalized posterior on set of predictors $\mathcal{H}$

$$\pi(h \mid Z^n, \eta) = \frac{\pi(dh) e^{-\eta \sum_{i=1}^n \text{loss}(Y_i, h(X_i))}}{\int_{h' \in \mathcal{H}} \pi(dh') e^{-\eta \sum_{i=1}^n \text{loss}(Y_i, h'(X_i))}}$$

- For 0/1-loss, this "procedure" "works" (posterior concentrates around best $\tilde{h}$) if $\eta < \eta'_{\text{crit}}$

$$\eta'_{\text{crit}} = \sup \left\{ \eta : \sup_{P \in \mathcal{P}} E_{P^*} \left(\frac{\mathrm{p}(Y)}{\mathrm{q}(Y)} \right)^{\eta} \leq 1 + \frac{1}{n} \right\}$$

  - optimal contraction rate determined by η'_{crit}

PAC-Bayes: beyond Log-Loss

- Define generalized posterior on set of predictors $\mathcal{H}$

$$\pi(h \mid Z^n, \eta) = \frac{\pi(dh) e^{-\eta \sum_{i=1}^n \text{loss}(Y_i, h(X_i))}}{\int_{h' \in \mathcal{H}} \pi(dh') e^{-\eta \sum_{i=1}^n \text{loss}(Y_i, h'(X_i))}}$$

- For 0/1-loss, this "procedure" "works" (posterior concentrates around best $\tilde{h}$) if $\eta < \eta'_{\text{crit}}$

$$\eta'_{\text{crit}} = \sup \left\{ \eta : \sup_{P \in \mathcal{P}} E_{P^*} \left(\frac{\mathrm{p}(Y)}{\mathrm{q}(Y)} \right)^{\eta} \leq 1 + \frac{1}{n} \right\}$$

  - optimal contraction rate determined by η'_{crit}

$$E_{P^*} \left[\frac{e^{-\eta \text{loss}(Y, h(X))}}{e^{-\eta \text{loss}(Y, \tilde{h}(X))}} \right]$$

PAC-Bayes: beyond Log-Loss

- Define generalized posterior on set of predictors $\mathcal{H}$

$$\pi(h \mid Z^n, \eta) = \frac{\pi(dh) e^{-\eta \sum_{i=1}^n \text{loss}(Y_i, h(X_i))}}{\int_{h' \in \mathcal{H}} \pi(dh') e^{-\eta \sum_{i=1}^n \text{loss}(Y_i, h'(X_i))}}$$

- For 0/1-loss, this "procedure" "works" (posterior concentrates around best $\tilde{h}$) if $\eta < \eta'_{\text{crit}}$

$$\eta'_{\text{crit}} = \sup \left\{ \eta : \sup_{P \in \mathcal{P}} E_{P^*} \left(\frac{\mathrm{p}(Y)}{\mathrm{q}(Y)} \right)^{\eta} \leq 1 + \frac{1}{n} \right\}$$

  - optimal contraction rate determined by η'_{crit}
  - we know $\eta'_{\text{crit}} > 1/\sqrt{n}$
  - if $(P^*, \mathcal{H})$ satisfies Tsybakov-Mammen condition then $\eta'_{\text{crit}} \asymp n^{-\alpha}$, for some $\alpha \in [0, 1/2]$

Bayesian and PAC-Bayesian Motivation

- Standard Bayesian inference uses $\eta = 1$
  - This may not converge at all if model is wrong. Want to use smaller η , but how to find it?
- Standard PAC-Bayesian inference uses $\eta = 1/\sqrt{n}$
  - This converges (but slowly). If situation is "nice", we can converge faster by using larger η , but how to find it?

Bayesian and PAC-Bayesian Motivation

- Standard Bayesian inference uses $\eta = 1$
  - This may not converge at all if model is wrong. Want to use smaller η , but how to find it?
- Standard PAC-Bayesian inference uses $\eta = 1/\sqrt{n}$
  - This converges (but slowly). If situation is "nice", we can converge faster by using larger η , but how to find it?

in fact, in both cases:
if $\eta > \eta'_{\text{crit}}$ then we may not converge at all,
if $\eta \ll \eta'_{\text{crit}}$ we may converge too slowly

Part 2: Towards a solution via a paradox

- So again: **How to learn the learning rate?**
  - "learning" learning rate η by empirical Bayes can give disastrous results (GL '07)
  - "hierarchical Bayes" (integrating out η) can give disastrous results! (GL '07)

Part 2: Towards a solution via a paradox

- So again: **How to learn the learning rate?**
  - "learning" learning rate η by empirical Bayes can give disastrous results (GL '07)
  - "hierarchical Bayes" (integrating out η) can give disastrous results! (GL '07)
- I recently "solved" issue (after 10 year long search...)
- **Paradox:** Bayesian predictive distribution behaves well in terms of cumulative KL risk even when model is completely wrong
- Understanding the paradox leads to a solution

Menu

1. Bayesian inconsistency under misspecification
   - G. and Langford, Machine Learning J. 2007
2. Learning Rate - Relation to Convexity, PAC-Bayes
3. **Sequential Prediction Detour**
   - **paradox:** Bayesian posterior good and bad at same time
4. The Safe Bayesian Algorithm
   - Use optimal learning rate, itself "learned" from data
5. "Unifying" Bayes and PAC-Bayes

Barron's Theorem (baby version)

- Let P^* be arbitrary distribution on Y, extended to n outcomes by independence. We have

$$E_{Y^n \sim P^*}\left[\sum_{i=1}^{n} D_i^*\right] \leq -\log \pi(Q)$$

where $D_i^* = D(P^* \| P_{\text{Bayes}}(\cdot \mid Y^{i-1})) - D(P^* \| Q)$ **(KL risk)**

$$p_{\text{Bayes}}(y \mid Y^{i-1}) = \sum_{p \in \mathcal{P}} p(y)\pi(p \mid Y^{i-1})$$

- Bayes predictive distribution has small cumulative KL risk even if model wrong

Barron's Theorem

$$E_{Y^n \sim P^*}\left[\sum_{i=1}^{n} D_i^*\right] \leq -\log \pi(Q) \quad \text{where } D_i^* = D(P^* \| P_{\text{Bayes}}(\cdot \mid Y^{i-1})) - D(P^* \| Q)$$

- Can easily extend this to uncountable $\mathcal{P}$, RHS then determined by discretization
- If model correct, then Bayes cumulative KL risk is usually minimax optimal by suitable choice of priors (Barron '98)
- If model wrong – paradox??

Paradox?

$$E_{Y^n \sim P^*}\left[\sum_{i=1}^{n} D_i^*\right] \le -\log \pi(Q) \quad \text{where } D_i^* = D(P^* \| P_{\text{Bayes}}(\cdot \mid Y^{i-1})) - D(P^* \| Q)$$

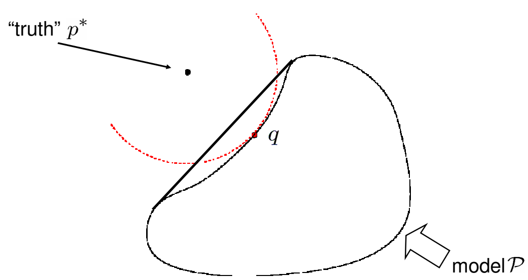
- D_i^* must be very small at most i
- If model correct, $P^* = Q$, good behaviour of Bayes' predictive distribution implies posterior concentration:
  D_i^* small ⟹ **all** P with substantial posterior weight must have $D(P^* \| P)$ close to $D(P^* \| Q) = 0$

Paradox?

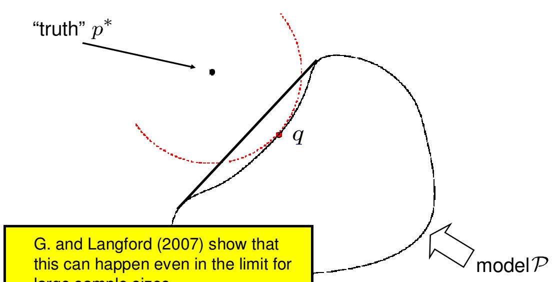
$$E_{Y^n \sim P^*}\left[\sum_{i=1}^{n} D_i^*\right] \le -\log \pi(Q) \quad \text{where } D_i^* = D(P^* \| P_{\text{Bayes}}(\cdot \mid Y^{i-1})) - D(P^* \| Q)$$

- D_i^* must be very small at most i
- If model correct, $P^* = Q$, good behaviour of Bayes' predictive distribution implies posterior concentration:
  D_i^* small ⟹ **all** P with substantial posterior weight must have $D(P^* \| P)$ close to $D(P^* \| Q) = 0$
- Yet if model misspecified, we can have good behaviour of Bayes predictive without concentration!

Bad and Good Misspecification



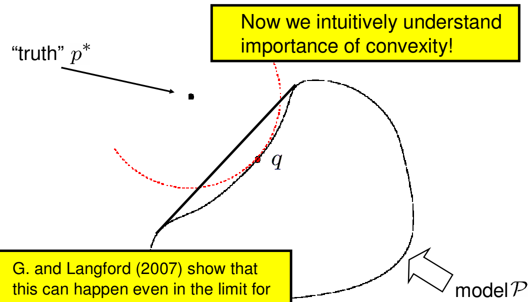
Bad and Good Misspecification



G. and Langford (2007) show that this can happen even in the limit for large sample sizes

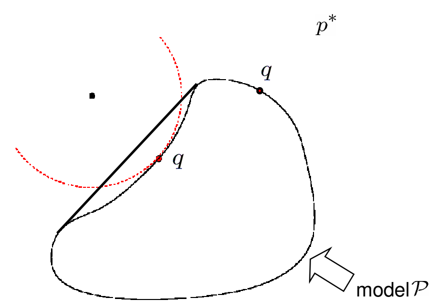
Bad and Good Misspecification

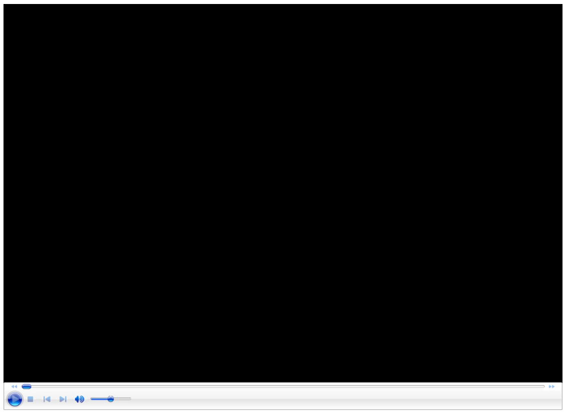
Now we intuitively understand importance of convexity!



G. and Langford (2007) show that this can happen even in the limit for large sample sizes

Bad and Good Misspecification





Three Observations

1. If posterior "concentrated" then predicting by randomizing using posterior not much worse than standard Bayes prediction (which mixes by posterior)

$$E_{Y_{i+1}\sim P^*}\left[E_{p\sim\Pi|Y^i}\left[-\log\frac{p(Y_{i+1})}{q(Y_{i+1})}\right]\right] \leq C\cdot E_{Y_{i+1}\sim P^*}\left[-\log E_{p\sim\Pi|Y^i}\left[\frac{p(Y_{i+1})}{q(Y_{i+1})}\right]\right]$$

2. If GL phenomenon takes place, then randomized predictions much worse than mixed predictions
   - right-hand side may even be negative

Three Observations

1. If posterior "concentrated" then predicting by randomizing using posterior not much worse than standard Bayes prediction (which mixes by posterior)

$$E_{Y_{i+1}\sim P^*}\left[E_{p\sim\Pi|Y^i}\left[-\log\frac{p(Y_{i+1})}{q(Y_{i+1})}\right]\right] \leq C\cdot E_{Y_{i+1}\sim P^*}\left[-\log E_{p\sim\Pi|Y^i}\left[\frac{p(Y_{i+1})}{q(Y_{i+1})}\right]\right]$$

2. If GL phenomenon takes place, then randomized predictions much worse than mixed predictions
3. If we use generalized posterior with $\eta < \eta_{\text{crit}}$ then posterior **will** tend to concentrate! (Thm.!)

Three Observations

1. If posterior "concentrated" then predicting by randomizing using posterior not much worse than standard Bayes prediction (which mixes by posterior)

$$E_{Y_{i+1}\sim P^*}\left[E_{p\sim\Pi|Y^i}\left[-\log\frac{p(Y_{i+1})}{q(Y_{i+1})}\right]\right] \leq C\cdot E_{Y_{i+1}\sim P^*}\left[-\log E_{p\sim\Pi|Y^i}\left[\frac{p(Y_{i+1})}{q(Y_{i+1})}\right]\right]$$

2. If GL phenomenon takes place, then randomized predictions much worse than mixed predictions
3. If we use generalized posterior with $\eta < \eta_{\text{crit}}$ then posterior **will** tend to concentrate! (Thm.!)

Idea: determine η that optimizes the fit of a randomizing rather than a mixing Bayesian!

The Safe Bayesian Algorithm

- First idea (which does not yet work): find η maximizing

$$\begin{aligned}\log p_{\text{Bayes}}(Y^n \mid \eta) &= \sum_{i=1}^{n} \log p_{\text{Bayes}}(Y_i \mid Y^{i-1}, \eta)\\ &= \sum_{i=1}^{n} \log \sum_{p} p(Y_i)\pi(p \mid Y^{i-1}, \eta)\\ &= \sum_{i=1}^{n} \log \mathbf{E}_{p\sim\Pi|Y^{i-1},\eta}\, p(Y_i)\end{aligned}$$

- This is like empirical Bayes. Similarly might try to put a prior on η and integrate it out (as a real Bayesian would do). That also doesn't work...

The Safe Bayesian Algorithm

- First idea (which does not yet work): find η maximizing

$$\begin{aligned}\log p_{\text{Bayes}}(Y^n \mid \eta) &= \sum_{i=1}^{n} \log p_{\text{Bayes}}(Y_i \mid Y^{i-1}, \eta)\\ &= \sum_{i=1}^{n} \log \sum_{p} p(Y_i)\pi(p \mid Y^{i-1}, \eta)\\ &= \sum_{i=1}^{n} \log \mathbf{E}_{p\sim\Pi|Y^{i-1},\eta}\, p(Y_i)\end{aligned}$$

- Instead we maximize

$$\sum_{i=1}^{n} \mathbf{E}_{p\sim\Pi|Y^{i-1},\eta} \log p(Y_i)$$

The Safe Bayesian Algorithm

- Want to do Bayesian inference for $\eta \approx \eta_{\text{crit}}$
- But of course we don't know η_{crit} ....
- Instead we pick $\hat{\eta}(Y^n) \in [1/\sqrt{n}, 1]$ which maximizes posterior-expected log-likelihood according to **sequentially randomized Bayes predictive distr.**
  - (cf. Freund & Shapire's "Hedge" algorithm!)
- We then use the corresponding randomized predictive distribution as a (randomized) "estimator" /predictor of P^*
- This (almost) works!

Preparing Main Result

- Let $\mathcal{P}$ be a set of conditional distributions $P_{Y|X}$, and let Π be a prior on $\mathcal{P}$
- Let P^* be a distribution on $\mathcal{X} \times \mathcal{Y}$
- Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be i.i.d. $\sim P^*$
- Let Q achieve $\inf_{P \in \langle \mathcal{P} \rangle} D(P^* \| P)$
- Let
$$\eta_{\text{crit}} = \sup \left\{ \eta : \sup_{p \in \mathcal{P}} \mathbf{E}_{P^*} \left(\frac{p(Y \mid X)}{q(Y \mid X)} \right)^{\eta} \leq 1 + \frac{1}{n} \right\}$$
- Proposition:
  - if model correct ($Q \in \langle \mathcal{P} \rangle$) or convex then $\eta_{\text{crit}} \geq 1$
  - if $\text{ess}\sup_{P, P' \in \mathcal{P}} \frac{p(Y \mid X)}{p'(Y \mid X)} \leq V$ then $\eta_{\text{crit}} \geq \frac{C}{\log V \cdot \sqrt{n}}$

Main Result

Let $\eta < \eta_{\text{crit}}$. We (almost!) have

$$\mathbf{E}_{Z^n \sim P^*} [\text{D}(Q \| \text{ random draw from posterior at } \hat{\eta}(Z^n))] \leq$$

$$\frac{1}{n} \cdot \frac{1}{\eta} (\text{complexity term, sublinear in } n)$$

where in case model is correct, the complexity term is within a constant factor of the minimax optimal rates that can be obtained in such cases

Main Result

Let $\eta < \eta_{\text{crit}}$ We have

$$\Pi_{\text{Ces}} \mid Z^n, \eta := n^{-1} \sum_{i=1}^{n} \Pi \mid Z^{i-1}, \eta$$

$$\mathbf{E}_{Z^n \sim P^*} \mathbf{E}_{P \sim \Pi_{\text{Ces}} | Z^n, \hat{\eta}(Z^n)} [D^*(Q \| P)] \leq$$

$$\frac{1}{n} \cdot \frac{1}{\eta} (\text{complexity term, sublinear in } n)$$

where in case model is correct, the complexity term is constant if model is countable, $O(\log n)$ if model parametric, and $O(n^{\gamma})$ for general nonparametric models

Note: D behaves like *square* of most common distances

Main Result (Oracle Bound)

Let $\eta < \eta_{\text{crit}}$. We have

$$\mathbf{E}_{Z^n \sim P^*} \mathbf{E}_{P \sim \Pi_{\text{Ces}} | Z^n, \hat{\eta}(Z^n)} [D^*(Q \| P)] \leq$$

$$\frac{C_\eta}{n} \mathbf{E}_{Z^n \sim P^*} \left[-\log \frac{p_{\text{Bayes}}(Y^n \mid X^n, \eta)}{q(Y^n \mid X^n)} + O\left(\frac{\log \log n}{\eta} \right) \right]$$

where C_η decreasing in η and $C_{\eta_{\text{crit}}/2} \leq 2 + \eta_{\text{crit}} \log V$

Main Result

Let $\eta < \eta_{\text{crit}}$. We have

$$\mathbf{E}_{Z^n \sim P^*} \mathbf{E}_{P \sim \Pi_{\text{Ces}} | Z^n, \hat{\eta}(Z^n)} [D^*(Q \| P)] \leq$$

$$\underbrace{\frac{C_\eta}{n} \mathbf{E}_{Z^n \sim P^*} \left[-\log \frac{p_{\text{Bayes}}(Y^n \mid X^n, \eta)}{q(Y^n \mid X^n)}\right.}_{\boxed{\leq C_\eta \cdot \inf_{\epsilon \geq 0} \left(\epsilon + \frac{-\log \Pi(p \ : \ D^*(q \| p) \leq \epsilon)}{n\eta} \right) \leq C_\eta \cdot \frac{-\log \pi(q)}{n\eta}}} \left. + O\left(\frac{\log \log n}{\eta} \right) \right]$$

"resolvability"

Main Result

Let $\eta < \eta_{\text{crit}}$. We have

$$\mathbf{E}_{Z^n \sim P^*} \mathbf{E}_{P \sim \Pi_{\text{Ces}} | Z^n, \hat{\eta}(Z^n)} [D^*(Q \| P)] \leq$$

$$\frac{C_\eta}{n} \mathbf{E}_{Z^n \sim P^*} \left[-\log \frac{p_{\text{Bayes}}(Y^n \mid X^n, \eta)}{q(Y^n \mid X^n)} + O\left(\frac{\log \log n}{\eta}\right) \right]$$

where C_η decreasing in η and $C_{\eta_{\text{crit}}/2} \leq 2 + \eta_{\text{crit}} \log V$

- If model correct and C_η finite, this is as good as bounds for standard Bayes up to constant factor – leading to optimal rates by suitable choice of prior (see Barron '98)
- If model incorrect, we still have "consistency", and we get optimal rates in classification under Tsybakov conditions

PAC-Bayes: beyond Log-Loss

- Let loss : $\mathcal{Y} \times \mathcal{A} \to [0, \infty]$ be arbitrary loss fn.
- Define generalized posterior on set of predictors $\mathcal{H}$ as

$$\pi(h \mid Z^n, \eta) = \frac{\pi(dh) e^{-\eta \sum_{i=1}^n \text{loss}(Y_i, h(X_i))}}{\int_{h' \in \mathcal{H}} \pi(dh') e^{-\eta \sum_{i=1}^n \text{loss}(Y_i, h'(X_i))}}$$

McAllester '02, Audibert '04, Zhang '06, Catoni '07

- With log-loss this reduces to original generalized posterior

Main Result, general loss fns.

Let $\eta < \eta_{\text{crit}}$. We have

$$\boxed{R(\tilde{h}) = \inf_{h \in \mathcal{H}} R(h)}$$

$$\mathbf{E}_{Z^n \sim P^*} \mathbf{E}_{h \sim \Pi_{\text{Ces}} | Z^n, \hat{\eta}(Z^n)} \left[D^*(\tilde{h} \| h) \right] \leq C_\eta \cdot$$

$$\left(\frac{1}{n} \mathbf{E}_{Z^n \sim P^*} \left[-\log p_{\text{Bayes}}(Y^n \mid X^n, \eta) \right] - R(\tilde{h}) + O\left(\frac{\log \log n}{n \cdot \eta}\right) \right)$$

$$\boxed{\begin{aligned} &= R(h) - R(\tilde{h}) = \\ &\quad \mathbf{E}_{P^*}[L(Y, h(X)) - L(Y, \tilde{h}(X))] \end{aligned}}$$

Main Result, general loss fns.

Let $\eta < \eta_{\text{crit}}$. We have

$$\mathbf{E}_{Z^n \sim P^*} \mathbf{E}_{h \sim \Pi_{\text{Ces}} | Z^n, \hat{\eta}(Z^n)} \left[D^*(\tilde{h} \| h) \right] \leq C_\eta \cdot$$

$$\left(\underbrace{\frac{1}{n} \mathbf{E}_{Z^n \sim P^*} \left[-\log p_{\text{Bayes}}(Y^n \mid X^n, \eta) \right] - R(\tilde{h})}_{\boxed{\leq \inf_{\epsilon \geq 0} \left(\epsilon + \frac{-\log \Pi(h \,:\, D^*(\tilde{h} \| h) \leq \epsilon)}{n\eta} \right)}} + O\left(\frac{\log \log n}{n \cdot \eta}\right) \right)$$

$$\boxed{= R(h) - R(\tilde{h})}$$

Main Result - Oracle Bound

Let $\eta < \eta_{\text{crit}}$. We have

$$\mathbf{E}_{Z^n \sim P^*} \mathbf{E}_{h \sim \Pi_{\text{Ces}} | Z^n, \hat{\eta}(Z^n)} \left[D^*(\tilde{h} \| h) \right] \leq C_\eta \cdot$$

$$\left(\inf_{\epsilon \geq 0} \left(\epsilon + \frac{-\log \Pi(h \,:\, D^*(\tilde{h} \| h) \leq \epsilon)}{n\eta} \right) + O\left(\frac{\log \log n}{n \cdot \eta}\right) \right)$$

- RHS corresponds to best rates obtainable if η_{crit} **known**, at least in many cases (Zhang 06a,06b)
- Thus result implies convergence of 'randomized safe Bayesian estimator' at optimal rates in such cases

Main Result - Oracle Bound

Let $\eta < \eta_{\text{crit}}$. We have

$$\mathbf{E}_{Z^n \sim P^*} \mathbf{E}_{h \sim \Pi_{\text{Ces}} | Z^n, \hat{\eta}(Z^n)} \left[D^*(\tilde{h} \| h) \right] \leq C_\eta \cdot$$

$$\left(\inf_{\epsilon \geq 0} \left(\epsilon + \frac{-\log \Pi(h \,:\, D^*(\tilde{h} \| h) \leq \epsilon)}{n\eta} \right) + O\left(\frac{\log \log n}{n \cdot \eta}\right) \right)$$

- If loss fn. is η- mixable, then $\eta_{\text{crit}} \geq \eta$ **(!)** and for 'simple' $\mathcal{H}$ we get rates up to $O(1/n)$ [Van Erven, G. et at., subm.]
- For 0/1-loss, if $(P^*, \mathcal{H})$ satisfies a (generalized) Tsybakov margin condition with parameter $\kappa \in [1, \infty]$, then we get rates up to $O(n^{-\kappa/(2\kappa-1)})$- which are the minimax rates

note that we can do "model aggregation"

The Bayesian Belief in Concentration

- Under very weak conditions on prior, a Bayesian will believe that her posterior will concentrate, i.e. prediction by randomization not much worse than prediction by mixing:

$$\Pi\left\{E_{p\sim\Pi|Y^i}\left[-\log\frac{p(Y_{i+1})}{q(Y_{i+1})}\right]\rightarrow C\times\left(-\log E_{p\sim\Pi|Y^i}\left[\frac{p(Y_{i+1})}{q(Y_{i+1})}\right]\right)\right\}=1$$

- Can view our work as a test **(posterior predictive check!?!?)** of Bayesian assumption. If test fails, we modify our model (not to make it true – that would be too ambitious – but to make Bayes predict well!)

Thank you for your attention!

- Preliminary version of work appears in ALT 2012
- Related work in worst-case setting:
  Van Erven, G., De Rooij, Koolen: *Adaptive Hedge,* NIPS '11
- See also Larry Wasserman's blog "normal deviate" under "self-repairing Bayesian inference"

> "If a subjective distribution P attaches probability zero to a non-ignorable event, and if this event happens, then P must be treated with suspicion, and **modified** or replaced"
> - A. P. Dawid in *The Well-Calibrated Bayesian*, JASA 1982