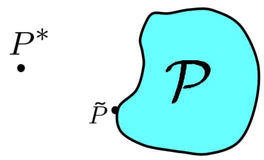
Bayesian Inconsistency under **Misspecification**

Peter Grünwald
CWI, Amsterdam



Extension of joint work with John Langford, TTI Chicago (COLT 2004)

The Setting

- 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 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 $\mathcal{P}$
- If $P^*_{Y|X} \notin \mathcal{P}$, then Bayes is **consistent** under ~~very mild~~ conditions on Π and $\mathcal{P}$

not quite so mild!

Misspecification Inconsistency

- We exhibit:
  1. a model $\mathcal{P}$,
  2. a prior Π
  3. a "true" P^*

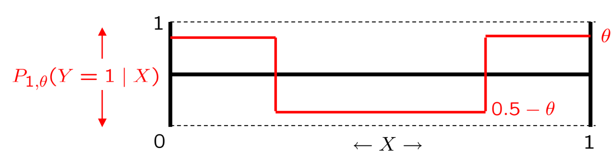
  such that
  - $\mathcal{P}$ contains a good approximation $\tilde{P}_{Y|X}$ to $P^*_{Y|X}$
  - Yet $P^*_{Y|X} \notin \mathcal{P}$
  - and Bayes will be inconsistent in various ways

The Model (Class) $\mathcal{P}$

- Let $\mathcal{P} := \mathcal{P}_1 \cup \mathcal{P}_2 \cup \ldots$
- Each $\mathcal{P}_k$ is a 1-dimensional parametric model,

$$\mathcal{P}_k := \left\{ P_{k,\theta} \mid \tfrac{1}{2} < \theta \leq 1 \right\}$$

- Example: $\mathcal{P}_1$

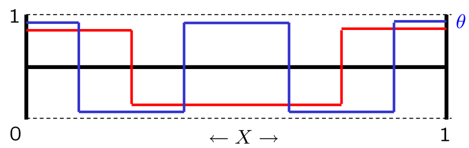


The Model (Class) $\mathcal{P}$

- Let $\mathcal{P} := \mathcal{P}_1 \cup \mathcal{P}_2 \cup \ldots$
- Each $\mathcal{P}_k$ is a 1-dimensional parametric model,

$$\mathcal{P}_k := \left\{ P_{k,\theta} \mid \tfrac{1}{2} < \theta \leq 1 \right\}$$

- Example: $\mathcal{P}_1$, $\mathcal{P}_2$

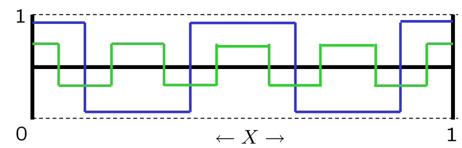


The Model (Class) $\mathcal{P}$

- Let $\mathcal{P} := \mathcal{P}_1 \cup \mathcal{P}_2 \cup \ldots$
- Each $\mathcal{P}_k$ is a 1-dimensional parametric model,

$$\mathcal{P}_k := \left\{ P_{k,\theta} \mid \tfrac{1}{2} < \theta \leq 1 \right\}$$

- Example: $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$



The Prior Π

- Prior mass Π on the model index k must be such that, for some small $\alpha > 0$, for all large k ,

$$\Pi(k) > \frac{1}{k^{1+\alpha}}$$

- Prior density $\pi(\theta \mid k) = \pi(\theta)$ of $P_{k,\theta}$ given k must be
  - continuous
  - not depend on k
  - bounded away from 0, i.e. $\inf_{\theta \in (0.5,1]} \pi(\theta) > 0$

The Prior Π

- Prior mass Π on the model index k must be such that, for some small $\alpha > 0$, for all large k ,

$$\Pi(k) > \frac{1}{k^{1+\alpha}}$$

$\boxed{\text{e.g. } \Pi(k) = \frac{1}{k(k+1)}}$

- Prior density $\pi(\theta \mid k) = \pi(\theta)$ of $P_{k,\theta}$ given k must be
  - continuous
  - not depend on k
  - bounded away from 0, i.e. $\inf_{\theta \in (0.5,1]} \pi(\theta) > 0$

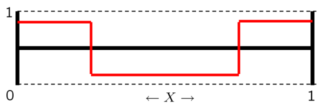
$\boxed{\text{e.g. } \pi(\theta) \text{ uniform}}$

The "True" P^* - benign version

- marginal P^*_X is uniform on $[0,1]$
- conditional $P^*_{Y|X}$ is given by

$$P^*(Y=1 \mid X=x) = \begin{cases} \frac{1}{3} & \text{if } \frac{1}{4} < x < \frac{3}{4} \\ \frac{2}{3} & \text{if } x \in \left[0,\frac{1}{4}\right] \cup \left[\frac{3}{4},1\right] \end{cases}$$

- Since $P^*_{Y|X} = P_{1,\tilde\theta}$ for $\tilde\theta = \frac{2}{3}$, we have $P^*_{Y|X} \in \mathcal{P}$
- Bayes is consistent



Two Natural "Distance" Measures

- Kullback-Leibler (KL) divergence

$$D(P^*\|P) := E_{X,Y\sim P^*}[-\log P(Y|X) + \log P^*(Y|X)]$$

- Classification performance (measured in 0/1-risk)

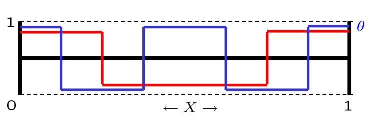
$$\delta_P(x) := \begin{cases} 1 & \text{if } P(Y=1 \mid X=x) > \frac{1}{2} \\ 0 & \text{if } P(Y=1 \mid X=x) < \frac{1}{2}. \end{cases}$$

$$\text{risk}_{01}(P^*,P) := E_{X,Y\sim P^*}[\,|Y-\delta_P(X)|\,]$$

KL and 0/1-behaviour of $\mathcal{P}_1$

- $\min_{P\in\mathcal{P}_1} D(P^*\|P) = 0$
- For all $k>1$: $\min_{P\in\mathcal{P}_k} D(P^*\|P) = 0.09..$
- For all $P \in \mathcal{P}_1$: $\text{risk}_{01}(P^*,P) = \frac{1}{3}$
- For all $P \in \mathcal{P}_k$, $k>1$: $\text{risk}_{01}(P^*,P) = \frac{1}{2}$

$\mathcal{P}_1$ is true and, of course, optimal for classification

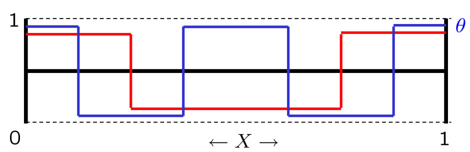


The "True" P^* - evil version

- First throw a coin $Z \in \{0,1\}$ with bias $P^*(Z=1) = 0.4$
- If $Z=1$, sample (X,Y) from P^* as before
- If $Z=0$, set $X=0, Y=1$

The “True” P^* - evil version

- First throw a coin $Z \in \{0,1\}$ with bias $P^*(Z=1)=0.4$
- If $Z=1$, sample (X,Y) from P^* as before
- If $Z=0$, set $X=0, Y=1$
- All $P \in \mathcal{P}$ satisfy $P(Y=1 \mid X=1) = \theta_P > \frac{1}{2}$ so these are **easy** examples



$\mathcal{P}_1$ is still optimal

- For all $P \in \mathcal{P}_1$, $\text{risk}_{01}(P^*, P) = 0.2$
- For all $P \in \mathcal{P}_k$, $k > 1$: $\text{risk}_{01}(P^*, P) = 0.3$

$\mathcal{P}_1$ is still optimal for classification

- $\min_{P \in \mathcal{P}} D(P^* \| P) = \min_{P \in \mathcal{P}_1} D(P^* \| P) = 0.35$
- For all $k > 1$, $\min_{P \in \mathcal{P}_k} D(P^* \| P) = 0.51$

$\mathcal{P}_1$ is still optimal in KL divergence

Inconsistency

- Both from a KL and from a classification perspective, we’d hope that the posterior converges on $\mathcal{P}_1$
- But this does not happen!

Inconsistency

- Theorem: With P^* - probability 1:
- $\frac{\max_k \Pi(\mathcal{P}_k \mid X^n, Y^n)}{\Pi(\mathcal{P}_1 \mid X^n, Y^n)} \geq e^{c\sqrt{n}}$ for some $c > 0$, all large n
- $\arg\max_k \Pi(\mathcal{P}_k \mid X^n, Y^n) \rightarrow \infty$

Posterior puts almost all mass on ever larger k, none of which are optimal

- $\Pi(\theta \in [a,1] \mid X^n, Y^n) \rightarrow 1$ for all $a < 1$

1. Kullback-Leibler Inconsistency

- Bad News: With P^* - probability 1, for all $B > 0$,
  $\Pi(\{P_{k,\theta} \ : \ D(P^* \| P_{k,\theta}) > B\} \mid X^n, Y^n) \rightarrow 1$
- Posterior "concentrates" on very bad distributions
  Note: if we restrict all $\mathcal{P}_k$ to $\theta \in (0.5, a]$ for some $a < 1$
  then this only holds for all B smaller than some $B_{\max}$

1. Kullback-Leibler Inconsistency

- Bad News: With P^* - probability 1, for all $B > 0$,
  $\Pi(\{P_{k,\theta} \ : \ D(P^* \| P_{k,\theta}) > B\} \mid X^n, Y^n) \rightarrow 1$
- Posterior "concentrates" on very bad distributions
- Good News: With P^* - probability 1, for all large n ,
  $\bar{D}(P^* \| P_{\text{Bayes}}(Y_{n+1} \mid X_{n+1}, X^n, Y^n))$
  $\leq \min_{P \in \mathcal{P}} D(P^* \| P) - 0.32$
- Predictive distribution does perform well!

2. 0/1-Risk Inconsistency

- Bad News: With P^* - probability 1,
  $\Pi(\{P_{k,\theta} \ : \ \text{risk}_{01}(P^*, P_{k,\theta}) = 0.3\} \mid X^n, Y^n) \rightarrow 1$
- Posterior "concentrates" on bad distributions
- More bad news: With P^* - probability 1,
  $\text{risk}_{01}(P^*, P_{\text{Bayes}}(Y_{n+1} \mid X_{n+1}, X^n, Y^n)\ ) \rightarrow 0.3.$
- Now predictive distribution is no good either!

Theorem 1: worse 0/1 news

- Prior $\Pi(\mathcal{P}_k) \approx \frac{1}{k^{1+\alpha}}$ depends on parameter α
- True distribution P^* depends on two parameters β, γ
  - With probability β , generate "easy" example
  - With probability $1 - \beta$, generate example according to $P_{1,\gamma}$

Theorem 1: worse 0/1 news

- Theorem 1: for each desired $\mu \in (0, 1/2)$, we can set α, β, γ such that

  $\min_{P \in \mathcal{P}} \text{risk}_{01}(P^*, P) = \mu$

  yet with P^*-probability 1, as $n \to \infty$

  $\text{risk}_{01}(P^*, P_{\text{map}(X^n,Y^n)}(Y_{n+1} \mid X_{n+1})\ ) \to \frac{1}{2} H(\mu)$

  $\text{risk}_{01}(P^*, P_{\text{Bayes}}(Y_{n+1} \mid X_{n+1}, X^n, Y^n)\ ) \to H(\mu)$
- Here $H(\mu)$ is the binary entropy

  $H(\mu) = -\mu \log_2 \mu - (1-\mu) \log_2 (1-\mu)$

Theorem 1, graphically

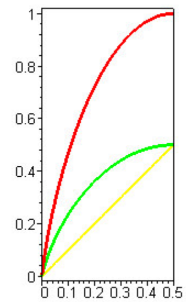
- X-axis: $\text{risk}_{01}(P^*, \tilde{P}) = \mu$
- —— = maximum 0/1 risk

  Bayes MAP $\frac{1}{2} H(\mu)$

achieved with probability 1, for all large *n*

- —— = maximum 0/1 risk

  full Bayes $H(\mu)$
- Maximum difference 0.5 at $\mu = 0.5$



Theorem 1, graphically

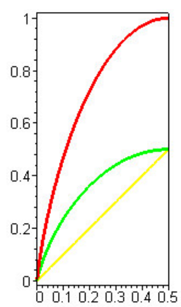
- X-axis: $\text{risk}_{01}(P^*, \tilde{P}) = \mu$
- —— = maximum 0/1 risk

  Bayes MAP

Bayes can get worse than random guessing!

- —— = maximum 0/1 risk

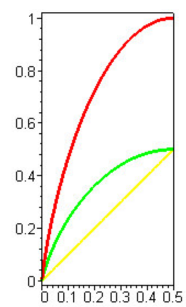
  full Bayes
- Maximum difference 0.5 at $\mu = 0.5$



Thm 2: full Bayes result is tight

- Let $\mathcal{P}$ be an arbitrary countable set of conditional distributions
- Let Π be an arbitrary prior on $\mathcal{P}$ with full support
- Let data be i.i.d. according to an arbitrary P^* on $\mathcal{X} \times \mathcal{Y}$, and let

  $\mu = \arg \min_{P \in \mathcal{P}} \text{risk}_{0/1}(P^*, P)$
- Then the 0/1-risk of Bayes predictive distribution is bounded by $H(\mu)$ (red line)



KL and Classification

- It is trivial to construct a model $\mathcal{P}'$ such that, for some P^*

$$\arg\min_{P \in \mathcal{P}} D(P^* \| P) \neq \arg\min_{P \in \mathcal{P}} \text{risk}_{01}(P^*, P)$$

KL and Classification

- It is trivial to construct a model $\mathcal{P}'$ such that, for some P^*

$$\arg\min_{P \in \mathcal{P}} D(P^* \| P) \neq \arg\min_{P \in \mathcal{P}} \text{risk}_{01}(P^*, P)$$

- However, here we constructed $\mathcal{P}$ such that, for arbitrary P^* on $\mathcal{X} \times \mathcal{Y}$,

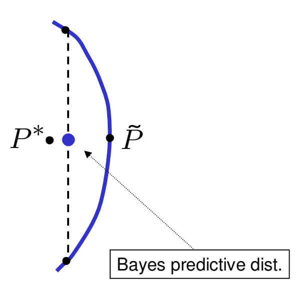
$$\arg\min_{P \in \mathcal{P}} D(P^* \| P)$$

achieved for $\tilde{P} = P_{\tilde{k},\tilde{\theta}}$ that also achieves

$$\min_{P \in \mathcal{P}} \text{risk}_{01}(P^*, P) = \tilde{\theta}$$

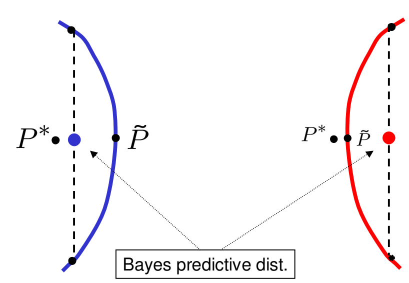
Therefore, the bad classification performance of Bayes is really surprising!

KL…



KL geometry

KL… and classification



KL geometry 0/1 geometry

What's new?

- There exist various infamous theorems showing that Bayesian inference can be inconsistent even if $P^* \in \mathcal{P}$
  - Diaconis and Freedman (1986), Barron (Valencia 6, 1998)
- So why is result interesting?
- Because we can choose $\mathcal{P}$ to be countable

Bayesian Consistency Results

- Doob (1949), Blackwell & Dubins (1962), Barron(1985)
  Suppose $\mathcal{P}$
  - **Countable**
  - Contains 'true' conditional distribution $P^*_{Y|X}$

  Then with P^*-probability 1, as $n \to \infty$,

$$\Pi(\{P^*_{Y|X}\} \mid X^n, Y^n) \to 1$$

Countability and Consistency

- Thus: if model well-specified and countable, Bayes must be consistent, and previous inconsistency results do not apply.
- We show that in misspecified case, can even get inconsistency if model countable.
  - Previous results based on priors with 'very small' mass on neighborhoods of true P^*
  - In our case, prior can be arbitrarily close to 1 on $\tilde{P}$ achieving $\min_{P \in \mathcal{P}} D(P^* \| P)$

Discussion

1. "Result not surprising because Bayesian inference was never designed for misspecification"
   - I agree it's not too surprising, but it *is* disturbing, because in practice, Bayes is used with misspecified models **all the time**

Discussion

1. “Result not surprising because Bayesian inference was never designed for misspecification”
   - I agree it’s not too surprising, but it *is* disturbing, because in practice, Bayes is used with misspecified models **all the time**
2. “Result irrelevant for true (subjective) Bayesian because a ‘true’ distribution does not exist anyway”
   - I agree true distributions don’t exist, but can rephrase result so that it refers only to realized *patterns* in *data* (not distributions)

Discussion

1. “Result not surprising because Bayesian inference was never designed for misspecification”
   - I agree it’s not too surprising, but it *is* disturbing, because in practice, Bayes is used with misspecified models **all the time**
2. “Result irrelevant for true (subjective) Bayesian because a ‘true’ distribution does not exist anyway”
   - I agree true distributions don’t exist, but can rephrase result so that it refers only to realized *patterns* in *data* (not distributions)
3. “Result irrelevant if you use a nonparametric model class, containing *all* distributions on $\mathcal{X} \times \mathcal{Y}$”
   - For small samples, your prior then severely restricts “effective” model size (there are versions of our result for small samples)

Discussion - II

- One objection remains: scenario is very unrealistic!
  - Goal was to discover the **worst possible scenario**
- Note however that
  - Variation of result still holds for $\mathcal{P}$ containing distributions with differentiable densities
  - Variation of result still holds if precision of X-data is finite
  - Priors are not so strange
  - Other methods such as **McAllester’s PAC-Bayes** do perform better on this type of problem.
  - They are guaranteed to be consistent under misspecification but often need much more data than Bayesian procedures

see also Clarke (2003), Suzuki (2005)

Conclusion

- Conclusion should *not* be

  “Bayes is bad under misspecification”,

- but rather

  “more work needed to find out what types of misspecification are problematic for Bayes”

Thank you, and let's Party!