Safe Bayes, Safe Probability

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematical Institute – Leiden University

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

Menu

1. A Problem for Bayes under Misspecification
2. Generalized η-Bayes, Critical Learning Rate η
3. Touch the likelihood!
   - new, simple interpretation of generalized posterior
4. General ‘Safe (Bayesian) Inference’

Bayesian Linear Regression Model

- Model $\mathcal{M}_k = \{p_{\vec{\beta},\sigma^2} \mid \sigma^2 \in \mathbb{R}^+, \vec{\beta} \in \mathbb{R}^{k+1}\}$

  expresses $Y = \sum_{j=0}^{k} \beta_j g_j(X) + \epsilon$

  where ϵ is 0-mean, σ^2 –variance Gaussian random variable, extended to n outcomes by independence:

  $p_{\vec{\beta},\sigma^2}(y^n \mid x^n, \mathcal{M}_k) \propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \sum_{j=0}^{k}\beta_j g_j(x_i))^2}$

  Use standard (Gaussian/Inv. Gamma) priors on β, σ^2

Experiment: Bayes Factor Model Selection for Polynomial Regression

- Model instantiated to $Y = \sum_{j=0}^{k} \beta_j X^j + \epsilon$
- Let's experiment to see what happens if data are sampled from following "true" distribution:

$$X_i \sim \text{Unif.}[-1, 1], \text{i.i.d.}$$

$$Y_i = 0 + \epsilon_i, \epsilon_i \sim \text{Normal}(0, 1), \text{i.i.d.}$$

- Note: model is (for now!) **correct**

Experiment

- Model instantiated to $Y = \sum_{j=0}^{k} \beta_j X^j + \epsilon$
- Let's experiment to see what happens if data are sampled from following "true" distribution:

  $X_i \sim \text{Unif.}[-1,1], \text{i.i.d.}$

  $Y_i = 0 + \epsilon_i, \epsilon_i \sim \text{Normal}(0,1), \text{i.i.d.}$
- Note: model is (for now!) **correct**
- ...and Bayes works perfectly well, selects 0-degree model after just a few outcomes and keeps on doing so for ever

Experiment

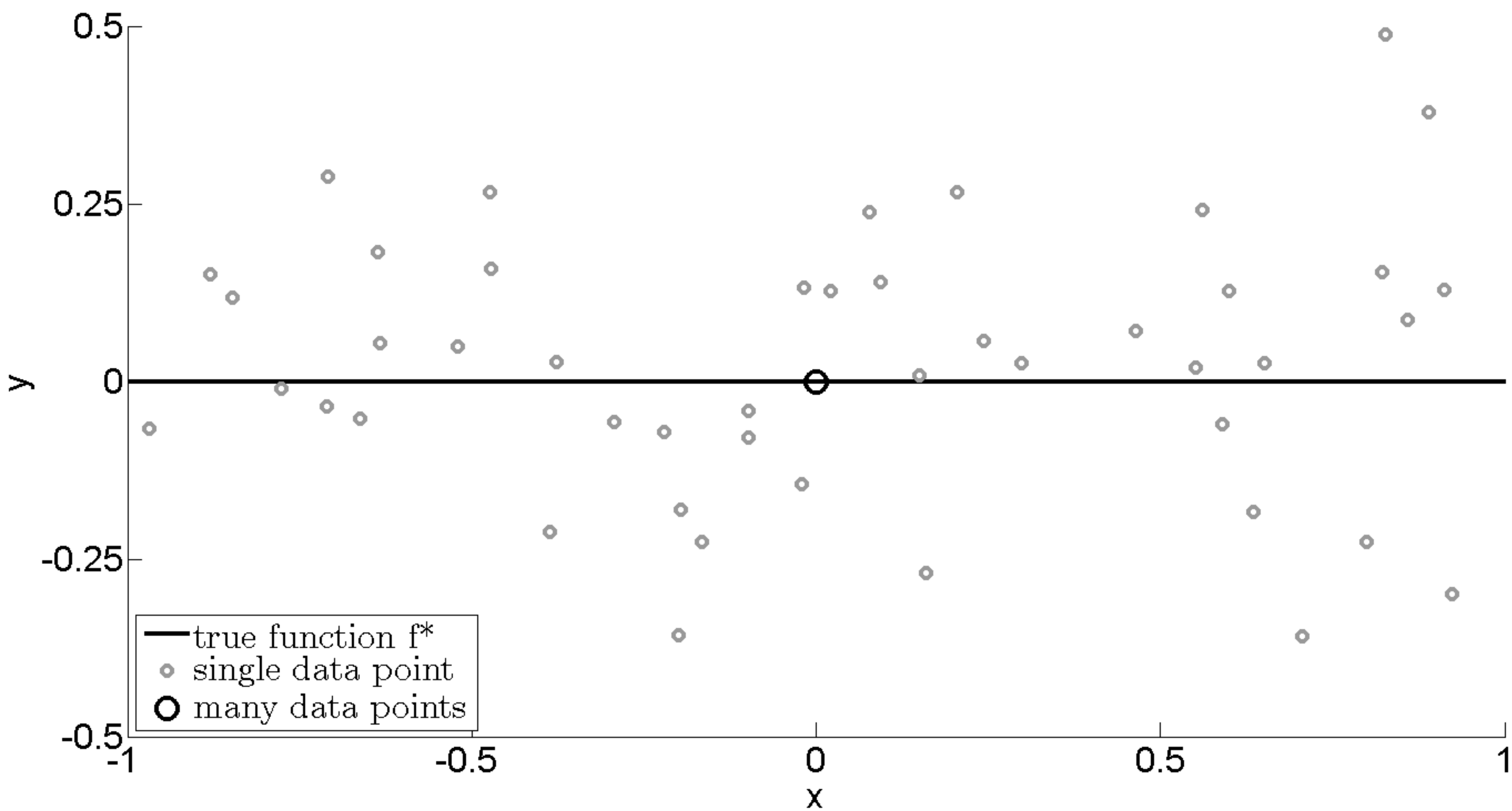
- Model instantiated to $Y = \sum_{j=0}^{k} \beta_j X^j + \epsilon$
- Let's experiment to see what happens if data are sampled from following "true" distribution:
- At each i, we independently toss a fair coin
  - if coin lands heads, as before:

    $X_i \sim \text{Unif.}[-1, 1], \text{i.i.d.}$

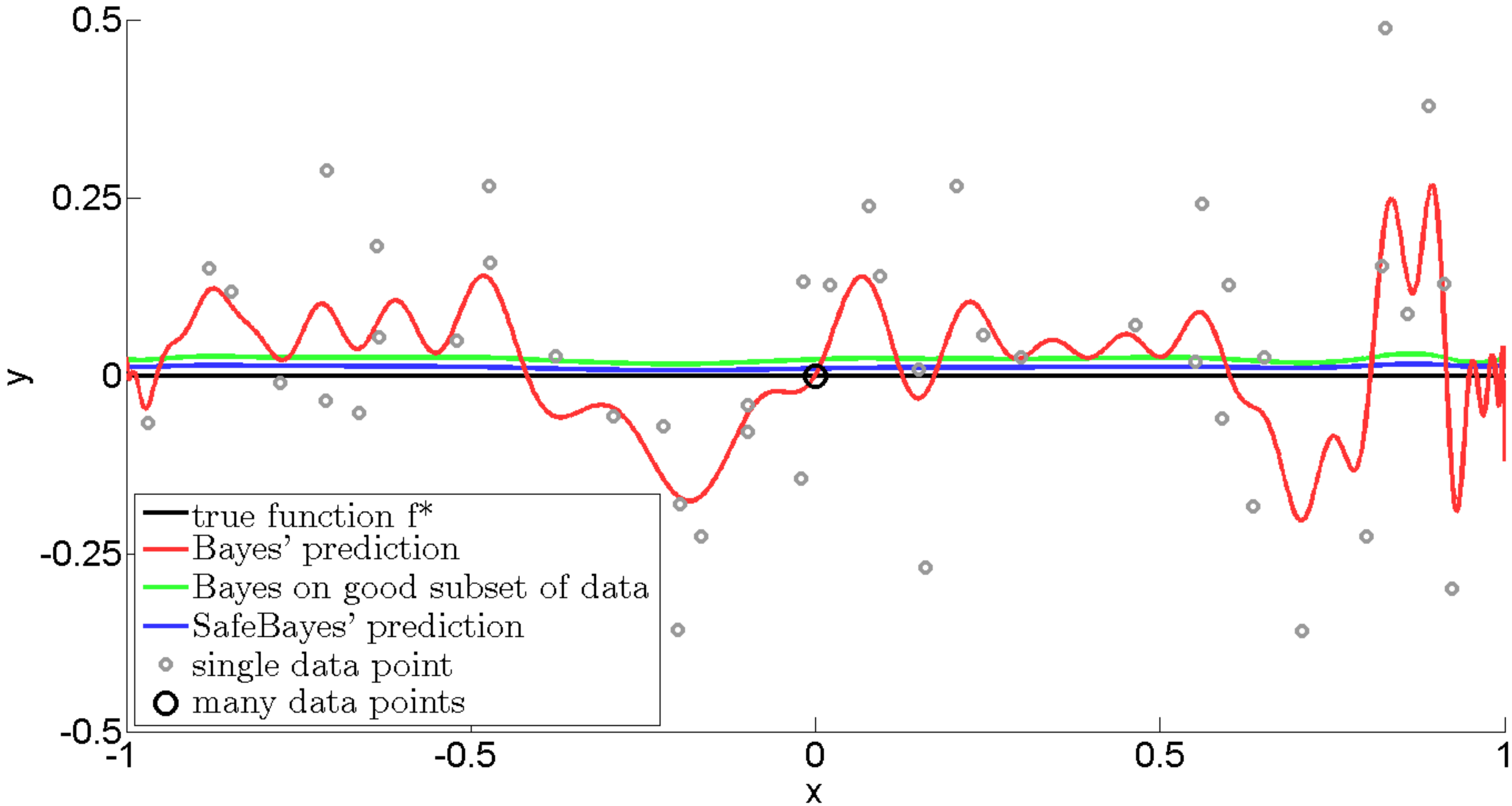
    $Y_i = 0 + \epsilon_i, \epsilon_i \sim \text{Normal}(0, 1), \text{i.i.d.}$

  - if tails, we generate an **easy** example ("**in-lier**")

    $(X_i, Y_i) = (0, 0)$

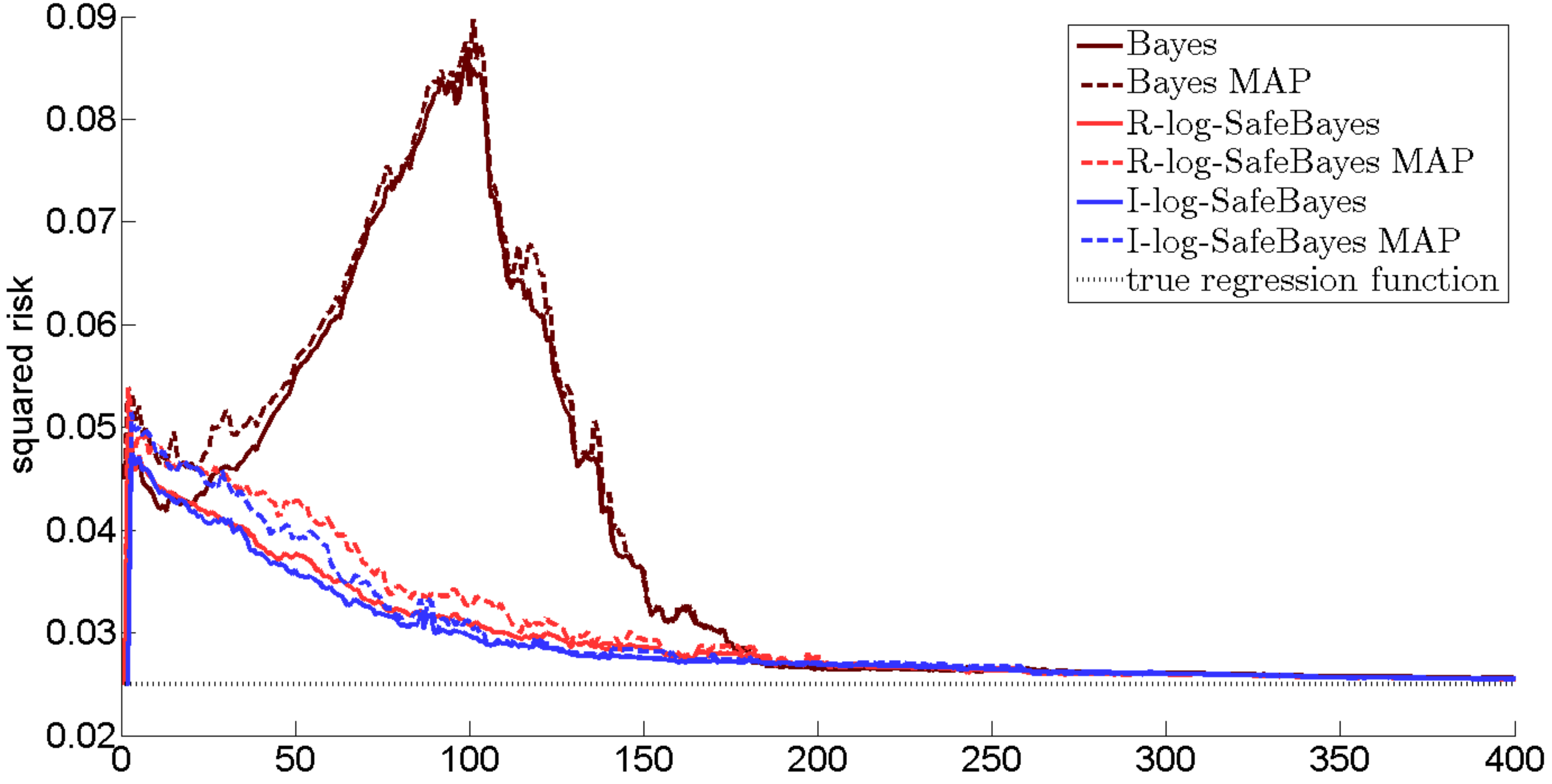


0.5
0.25
0
-0.25
-0.5
y
-1
-0.5
0
0.5
1
x
true function f*
single data point
many data points


0.5
0.25
0
-0.25
-0.5
y
-1
-0.5
0
0.5
1
x
true function f*
Bayes' prediction
Bayes on good subset of data
SafeBayes' prediction
single data point
many data points

$$\sigma^2 = 1/20 \to 1/40 = 0.025$$

Risk Graph



Risk measured in Expected Squared Loss on a new outcome

Important Remark

- If nr of basis functions is finite, then problem does go away *at some point*
- **Real issue**: if we take an infinite nr of basis functions (e.g. polynomials of all degree)
  - Bayes converges straight away if model correct
  - Bayes *never* converges if model contains 50% easy points

Menu

1. A Problem for Bayes under Misspecification
2. Generalized η-Bayes, Critical Learning Rate η
3. Touch the likelihood!
   – new, simple interpretation of generalized posterior
4. General ‘Safe (Bayesian) Inference’

Generalized Posterior

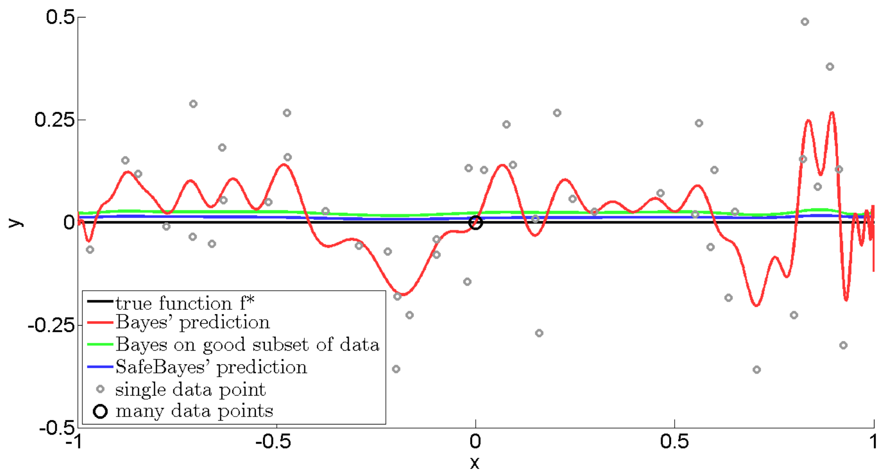
- Let $\{ p_f : f \in \mathcal{F} \}$ be a model, i.e. a set of densities
- We define the η-generalized posterior to be

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

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)

The Critical $\bar{\eta}$

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

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

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

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

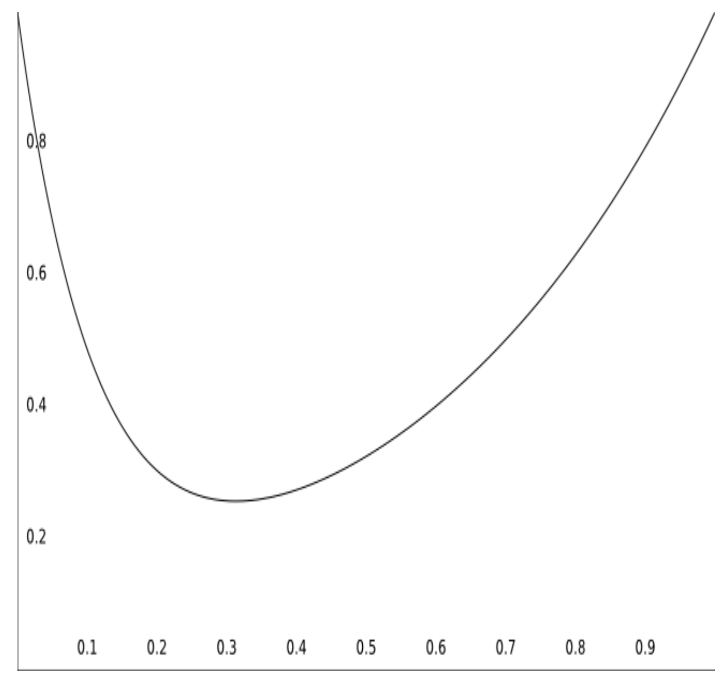
- If model correct, $\bar{\eta}$= 1, since

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

$$\int p_{f^*} \frac{p_f}{p_{f^*}} = 1$$

...and $A(0) = 1$ and $A(\eta)$

is (strictly) convex



First (Frequentist) Reason for $\bar{\eta}$

Let $Z_1, Z_2, \ldots \sim$ i.i.d. P . Let $f^* = \arg\min_{f \in \mathcal{F}} D(P \| P_f)$

- "Theorem" For any $0 < \eta < \bar{\eta}$, η-generalized Bayes tends to concentrate around f^* at minimax rate up to log factors (parametric and nonparametric settings)
- Reason, abstractly put:

For $\eta \leq \bar{\eta}$, $(p_f / p_{f^*})^{\eta}$ defines a supermartingale

For $\eta < \bar{\eta}$, it defines a strictly-super-martingale

First (Frequentist) Reason for $\bar{\eta}$

Let $Z_1, Z_2, \ldots \sim$ i.i.d. P . Let $f^* = \arg\min_{f \in \mathcal{F}} D(P \| P_f)$

- "Theorem" For any $0 < \eta < \bar{\eta}$, η-generalized Bayes tends to concentrate around f^* at minimax rate up to log factors (parametric and nonparametric settings)
- Reason, abstractly put:

  For $\eta \leq \bar{\eta}$, $(p_f / p_{f^*})^\eta$ defines a supermartingale

  For $\eta < \bar{\eta}$, it defines a strictly-super-martingale

indeed can extend notion to non-iid settings:

$$\mathbf{E}_{Z_n}\left[\left(\frac{p_f(Z_1,\ldots,Z_{n-1},Z_n)}{p_{f^*}(Z_1,\ldots,Z_{n-1},Z_n)}\right)^\eta \mid \mathcal{F}_{n-1}\right] \leq \left(\frac{p_f(Z_1,\ldots,Z_{n-1})}{p_{f^*}(Z_1,\ldots,Z_{n-1})}\right)$$

First Reason for $\overline{\eta}$

- Posterior Concentration Theorem
- Follows because, abstractly put, $(p_f/p_{f^*})^\eta$ defines supermartingale
- Less abstractly put:

Markov's inequality with union bound

e.g. for countable $\mathcal{F}$

$$P\left(\exists f \in \mathcal{F}: \ \pi(f) \cdot \left(\frac{p_f(Z^n)}{p_{f^*}(Z^n)}\right)^\eta > K\right) \leq \frac{1}{K}\left(\ \mathbf{E}\left(\frac{p_f(Z)}{p_{f^*}(Z)}\right)^\eta \ \right)^n$$

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 !

Well-Specified Case

Theorem thus says that if model is correct, then generalized Bayes with any $\eta < 1$ has posterior convergence property solely under the prior-KL-property

- Previous nonparametric posterior concentration results invariably
  - either have additional (more complicated) conditions (GGV: entropy nr condition ; Barron/Schervish/Wasserman/Zhang condition)
  - or also require $\eta < 1$ ...

(Walker, Hjort '01; Zhang '06; Barron & Cover, '91 (!))

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

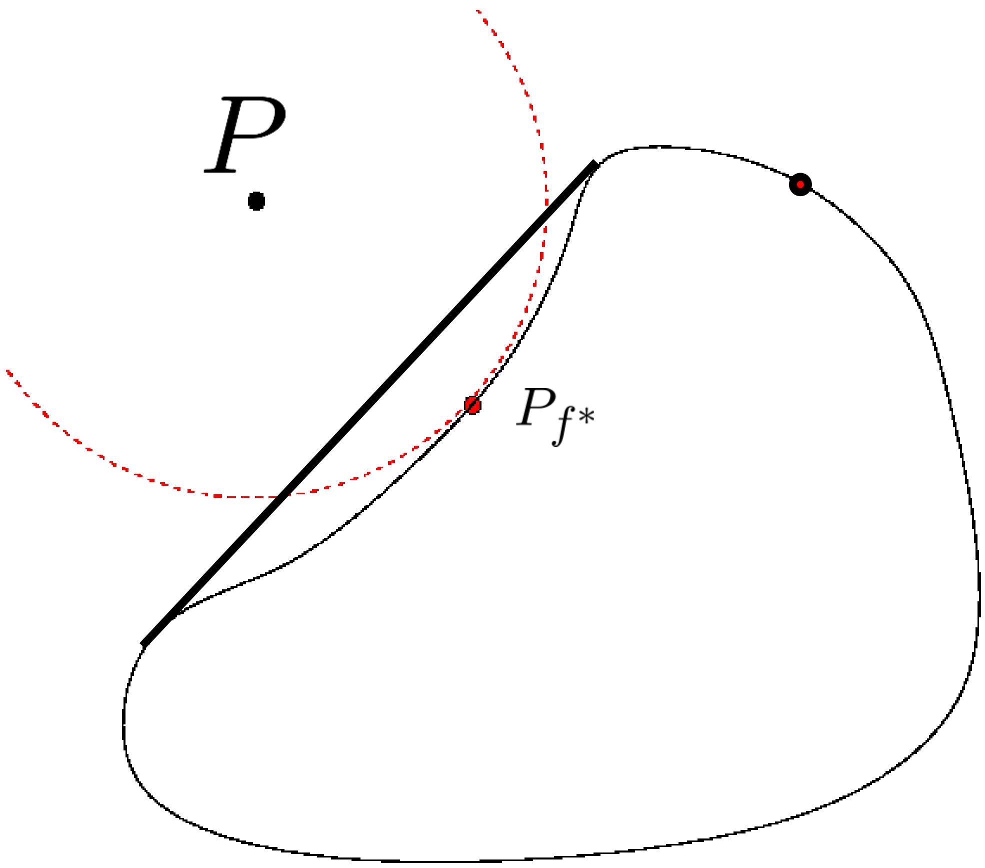
- We say that **convex luckiness** holds if

$$\inf_{f \in \mathcal{F}} D(P \| P_f) = \inf_{f \in \text{CONV-HULL}} D(P \| P_f)$$

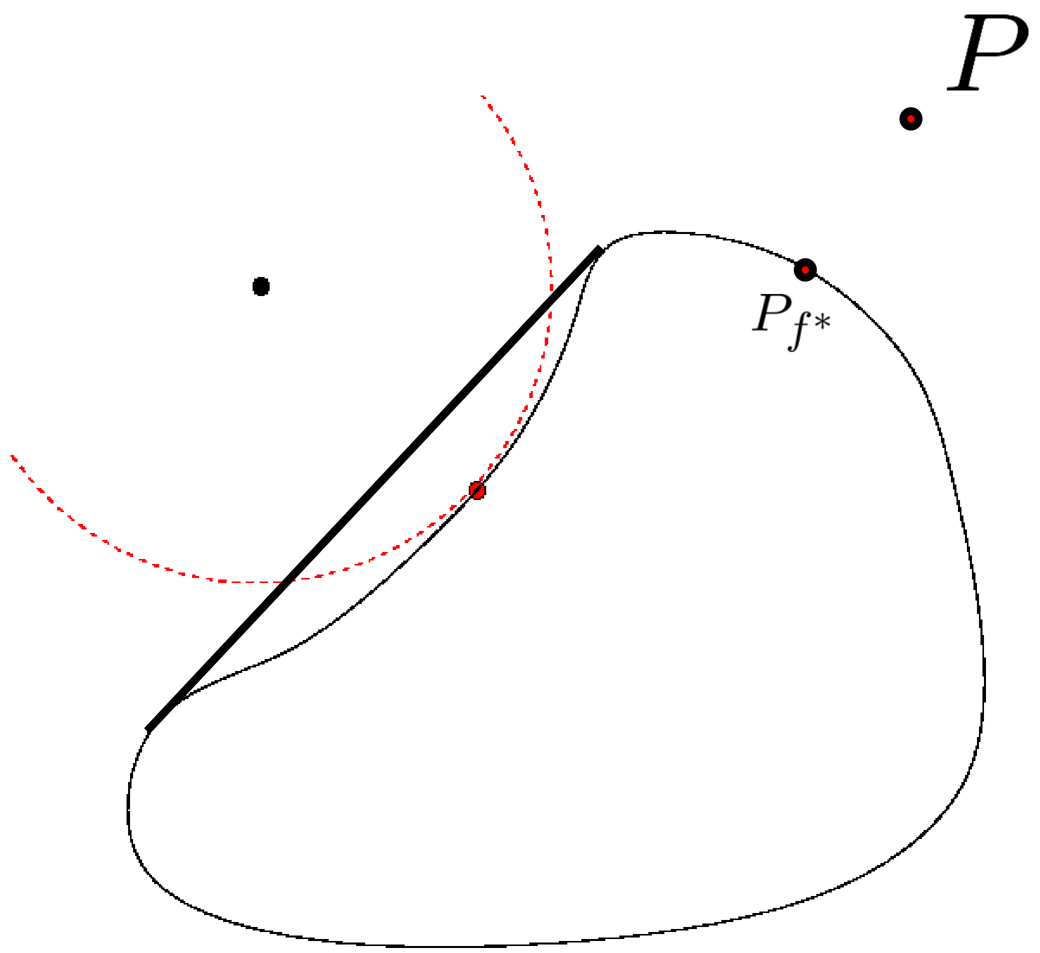
(Van Erven et al. '15, G & Mehta '17b)

- 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 $f^* \in \mathcal{F}$ $\mathbf{E}_P[Y \mid X] = f^*(X)$

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

(signal well-specified, noise misspecified)

- ...then

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

Generalized Linear Models

- Similar result holds for GLMs. Suppose that:

  1. for some $\lambda > 0$, $\sup_x \mathbf{E}_{Z\sim P}\left[e^{\lambda|Y|} \mid X = x\right] < \infty$
  2. $\{p_f : f \in \mathcal{F}\}$ contains true conditional mean, i.e. there exists $f^* \in \mathcal{F}$ with $\mathbf{E}_{P_{f^*}}[Y \mid X] = \mathbf{E}_P[Y \mid X]$
  3. boring technical stuff about link function

  ...then $\bar{\eta} > 0$ and moreover $\bar{\eta}$ converges to

$$\bar{\eta}\left(\mathcal{F} \cap B_D^{\mathrm{KL}}(f^*, \epsilon)\right) \to \frac{\tilde{\sigma}^2_{f^*}}{\sup_x \sigma_x^2}$$

as $\epsilon \to 0$, i.e. we “shrink” model to f^* (G. & Mehta,’17b)

Menu

1. A Problem for Bayes under Misspecification
2. Generalized η-Bayes, Critical Learning Rate η
3. Touch the likelihood!
   – new, simple interpretation of generalized posterior
4. General 'Safe (Bayesian) Inference'

"Don't Touch the Likelihood!"

Even though...

- even for well-specified models, anomalies can occur with $\eta = 1$, i.e. standard Bayes (Barron '99, Zhang '06, Csiszar & Shields, '00)
- Under misspecification, $\eta = 1$ can yield disastrous results and $\eta \ll 1$ works fine
- Posterior concentration can be proven under much weaker conditions once $\eta < 1$ ..

...Bayesians are hesitant to use generalized Bayes....even **frequentist Bayesians** are...

“Don’t Touch the Likelihood!”

- In G. and van Ommen (2017, Section 4.1), we give a novel interpretation of generalized Bayes that, we hope, will help convince people...

Entropification

- Following G. ('98), Li ('99), Van Erven et al. ('15), define reweighted measures

$$p'_{f,\eta}(z) := p(z) \cdot \left(\frac{p_f(z)}{p_{f^*}(z)} \right)^{\eta}$$

- For $\eta \leq \bar{\eta}$, we have for all $f \in \mathcal{F}$: $\int p'_{f,\eta}(z)\, d\mu(z) \leq 1$

Entropification

- Following G. ('98), Li ('99), Van Erven et al. ('15), define reweighted measures

$$p'_{f,\eta}(z) := p(z) \cdot \left(\frac{p_f(z)}{p_{f^*}(z)} \right)^{\eta}$$

- For $\eta \leq \bar{\eta}$, we have for all $f \in \mathcal{F}$: $\int p'_{f,\eta}(z)\, d\mu(z) \leq 1$
- Let $\mathcal{Z}' = \mathcal{Z} \cup \{\circ\}$, where $\circ$ is a fake outcome that will never actually occur. Extend $p'_{f,\eta}$ to $\mathcal{Z}'$ by setting

$$P'_{f,\eta}(Z = \circ) := 1 - \int_{z \in \mathcal{Z}} p'_{f,\eta}(z) d\mu(z)$$

- Now $\{p'_{f,\eta} : f \in \mathcal{F}\}$ is a probability model

INSIGHT 1: $\{p'_{f,\eta} : f \in \mathcal{F}\}$, even though it contains many silly distributions that waste some of their mass on things that will never happen, is a well-specified model for every $\eta > 0$!

$$p'_{f^*,\eta}(z) := p(z) \cdot \left(\frac{p_{f^*}(z)}{p_{f^*}(z)}\right)^{\eta} = p(z)$$

INSIGHT 1: $\{p'_{f,\eta} : f \in \mathcal{F}\}$ is well-specified model!

INSIGHT 2: The **standard** Bayesian posterior for this new model **coincides** with the η-Bayesian posterior for model $\{ p_f : f \in \mathcal{F}\}$

G. and van Ommen, BA 2017, Section 4.1.

(this is new insight, not to be found in earlier arxiv version and ISBA 2016 presentation!)

Touch the Likelihood!

INSIGHT 1: $\{p'_{f,\eta} : f \in \mathcal{F}\}$ is well-specified model!

INSIGHT 2: The **standard** Bayesian posterior for this new model **coincides** with the η-Bayesian posterior for model $\{ p_f : f \in \mathcal{F}\}$

- Thus, under misspecification, generalized Bayes with right η actually has interpretation as applying Bayes' theorem to a well-specified model; **standard Bayes does not!**
- So once you accept misspecification it's more Bayesian to touch the likelihood than to not touch it!

We should perhaps embrace η – Bayes more fully!

- We often use pseudo-likelihoods to simplify computations etc.
  - variational Bayes, substitution likelihood, rank-based likelihood....
- Our interpretation suggests that in all such cases, it might be better to use appropriate $\eta \neq 1$ since then our posterior is still interpretable as applying Bayes rule to a correct model!

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

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

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

i.e.

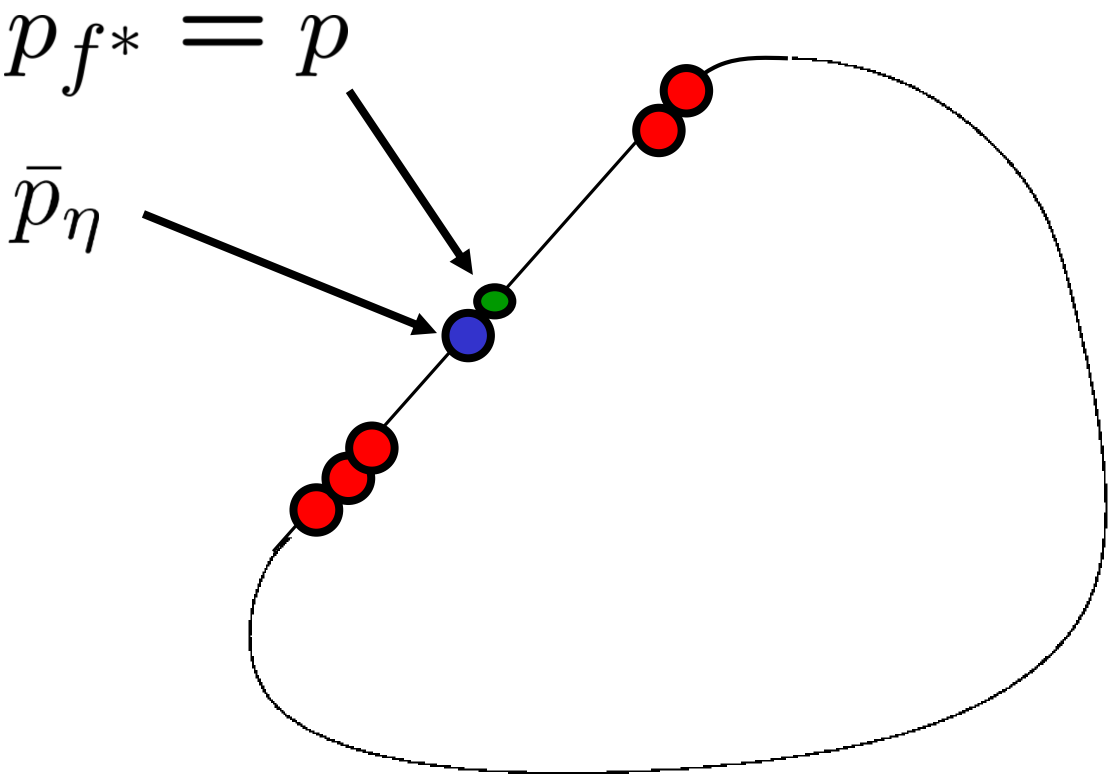
$$\bar{p}_\eta(Z_i = \cdot \mid Z^{i-1}) \rightarrow p_{f^*,\eta}$$

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

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

- 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)...

Bad Posterior, Good Predictive



Posterior concentration

Posterior concentration guaranteed if we take η strictly (but slightly) smaller than $\bar{\eta}$, since

(a) model remains correct, i.e. $p'_{f^*,\eta}$ remains true distribution, wasting 0 mass to fake outcomes

(b) convergence/consistency thm remains valid (although convergence will be slightly slower)

(c) ...

Posterior concentration

Posterior concentration guaranteed if we take η strictly (but slightly) smaller than $\bar{\eta}$, since

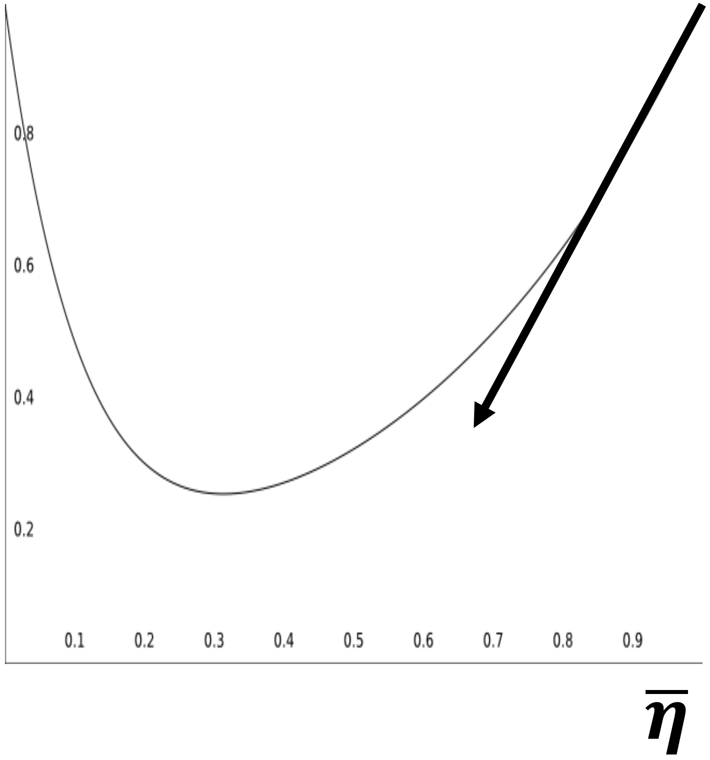
(a) model remains correct, i.e. $p'_{f^*,\eta}$ remains true distribution, wasting 0 mass to fake outcomes

(b) convergence/consistency thm remains valid (although convergence will be slightly slower)

(c) *all other* $p'_{f,\eta}$ now assign strictly positive probability to fake outcomes... hence so do their mixtures, so these mixtures can never become competitive with $p'_{f^*,\eta}$

- ...hence convergence of the predictive now implies concentration of the posterior the worse f , the more mass it will start wasting:

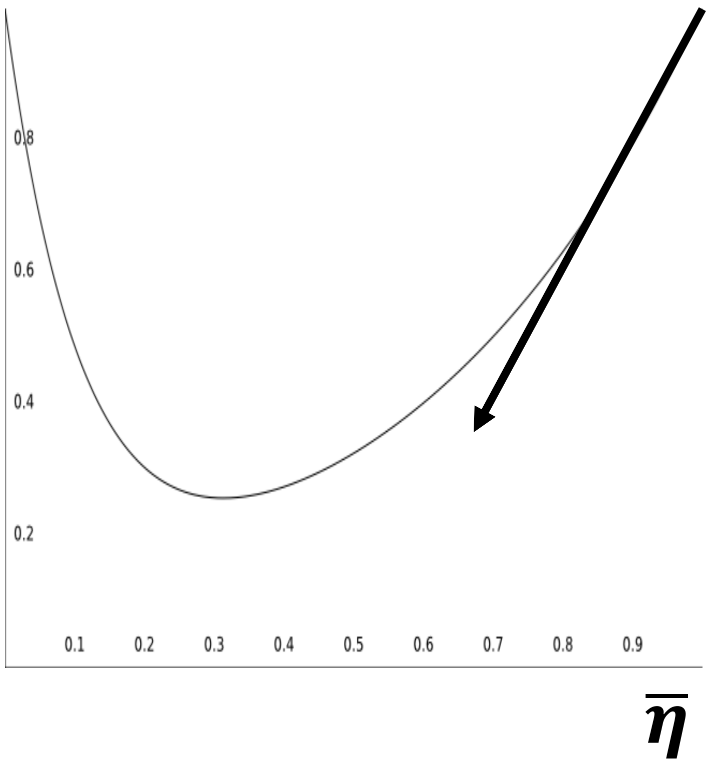
derivative of $A(\eta) = E\left(\frac{p_f}{p_{f*}}\right)^{\eta}$ at $\eta = \bar{\eta}$ “proportional” to $D_P(f^*||f)$



Prediction easier than identification!

the worse f , the more mass it will start wasting:

derivative of $A(\eta) = E\left(\frac{p_f}{p_{f*}}\right)^{\eta}$ at $\eta = \bar{\eta}$ “proportional” to $D_P(f^*||f)$



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 probabilities should not be taken fully seriously; their application should be restricted to **safe** uses

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)
     - R Package on CRAN for regression using η-generalized Bayes and SafeBayes (De Heide, ’16)
     - Provably finds ‘right $\bar{\eta}$’ for bounded likelihood ratios
     - In practice significantly outperforms Bayesian Lasso (De Heide, ‘16)
     - I am not wed to this algorithm however!
     - I am wed to claim that ‘$\bar{\eta} < \eta$’ is ‘right value to use’!

     (INVITE: write R packages for other models than regression)

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 probabilities should not be taken fully seriously; their application should be restricted to **safe** uses

Safe Probability

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

- **misspecification:** If your model is incorrect, then you might still converge (with the right η) to a distribution that estimates the conditional mean correctly, but perhaps not the conditional median; or that would give a very bad idea about the noise distribution (cf Watson & Holmes '16, contextuality of misspecification)
- **priors...even if model correct**

Safe Misspecified Bayes

- KL-associated prediction tasks: those on which you can give guarantees, as long as you use right η so that you converge to the KL-optimal distribution in your model
- For linear regression model, 2 KL-associated tasks:
  - **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 Bayes, Safe Probability

Even if your model is correct, in Bayesian practice you often cannot assume that your prior *really* captures your beliefs

Safe Priors

Even if your model is correct, in Bayesian practice you often cannot assume that your prior *really* captures your beliefs

In that case, you should *restrict* the applicability of your prior: state what it can be used for and not.

Safe Priors

Even if your model is correct, in Bayesian practice you often cannot assume that your prior *really* captures your beliefs

Example 1: Bernoulli with Jeffreys' prior. If you really believe the prior, you would be willing to play the following game: 10000 outcomes will be generated ; then:

- if empirical average is between 0.45-0.55, you pay 9$
- If between 0 and 0.05 you get 1$
- Otherwise nothing happens

Who in this room would actually want to play this game!?

Safe Priors

Even if your model is correct, in *objective Bayes* approaches (that's what we're here for!) you cannot assume that your prior *really* captures your beliefs

In that case, you should *restrict* the applicability of your prior: state what it can be used for and not.

A Vision: Safe Probability

A principled way to state what your model/prior should and should not be used for. For example, if you do a Bayesian regression analysis, you could, depending on how sure you are of model/prior, state that

– inference is safe for learning the optimal squared error predictor within your model

– inference is safe for learning the true regression function (i.e. you have to be right conditional on X)

– inference is safe for making probability rather than in-expectation statements of Y (noise process correct)

A Vision: Safe Probability

In hypothesis testing, you could state for example:

- my priors are safe for a given sampling plan
- my priors are safe under optional continuation
- my priors are safe under optional stopping
- my priors are safe for gambling
    - you really believe them in the sense that you would be willing to pay 1$ for a bet that pays out 2$ if θ lies in a set of prior prob > ½)

If we would all adopt such a stance, it would lead to (yes!) safer statistics.

A first, theoretical stab in this direction is made by G. 2017, Safe Probability, *Journ.Stat. Planning & Inference*

Thank you
for your attention!

Further Reading and Doing:

- G. and Van Ommen, *Bayesian Analysis, Dec. 2017*
- G. and Mehta, Fast Rates for Unbounded Losses, *arXiv* (2016, 2017b – first part is about Bayesian consistency and convergence under misspecification)
- G. **Safe Probability**. *Journal of Statistical Planning and Inference, 2017*

- R-Package SafeBayes for regression

Additional Material

Part II: 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

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

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"

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

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!

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

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 $\mathbf{E}[Y|X]$?

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!

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

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 $\hat{\beta}_n^T X$ give a good estimate of $\mathbf{E}[Y|X]$?

B: Of course!

Q: Does $\hat{\beta}_n^T X$ 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

- In reality one is often 'somewhere 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^* which will also satisfy
  - **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 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

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 ((penalized) conditional likelihood maximization of logistic model) 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 $\tilde{P}$

Optional stopping!

Safe Probability

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