Preamble: Aggregating Least Squares Regressions

- Full Model: Data $Y \sim \text{Normal}(\mu, \sigma^2 I)$ with μ in $\mathbb{R}^n$
- Linear Models: μ is in the span of a design matrix
- M models indexed by m, dimensions d_m
- Let $\hat{\mu}_m$ be the least squares projection of Y for model m
- Individual risk function $r_m = \mathbb{E}\|\hat{\mu}_m - \mu\|^2$.
- Stein's unbiased estimate of risk

$$\hat{r}_m = \|Y - \hat{\mu}_m\|^2 + \sigma^2(2d_m - n)$$

- Model Selection: $\hat{m} = \mathrm{argmin}_m \hat{r}_m$
- Model Aggregation: $\hat{\mu} = \sum_m w_m \hat{\mu}_m$
- Advocate weights w_m proportional to $\exp[-(\beta/\sigma^2)\hat{r}_m]$

Stein Estimate of Risk of Aggregated Least Squares

- Model Selection: $\hat{m} = \text{argmin}_m \hat{r}_m$
- Model Aggregation: $\hat{\mu} = \sum_m w_m \hat{\mu}_m$
- Risk $r = \mathbb{E}\|\hat{\mu} - \mu\|^2$
- What is the Stein unbiased estimate of this risk?
- Studied in Leung and Barron (2006)
- Advocate weights w_m proportional to $\exp[-(\beta/\sigma^2)\hat{r}_m]$
- $\beta = 1/2$ for posterior weights; $\beta = 1/4$ for risk simplification
- The Stein estimate of risk of $\hat{\mu}$ simplifies to $\hat{r} = \sum_m w_m \hat{r}_m$ which may be expressed as

$$\hat{r} = \hat{r}_{\hat{m}} + 4\sigma^2[H(w) + \log w_{\hat{m}}]$$

so that

$$\hat{r} \leq \min_m \hat{r}_{\hat{m}} + 4\sigma^2 \log M$$

Accordingly the risk of the aggregated $\hat{\mu}$ satisfies

$$r \leq \min_m \mathbb{E}\|\hat{\mu}_m - \mu\|^2 + 4\sigma^2 \log M$$

Information Theory and Statistical Learning: Foundations and a Modern Perspective

ANDREW R BARRON

Department of Statistics and Data Science
Yale University

Robert Bohrer Workshop in Statistics
University of Illinois at Urbana Champaign, April 22, 2023

Models and Likelihood

- **Likelihood:** Early statistical foundations
  Bayes, Laplace, Gauss shared a Bayesian perspective.
  R. A. Fisher championed likelihood.
- **Model:** Input X, output Y with center $f(X, \theta)$, parameters θ.
  For instance, a linear model or a modern deep network.
- **Probability Model:** for finite precision X, Y.
  Design distribution $p(x)$, output condit. distrib. $p(y|x, \theta)$.
- **Data:** For *training* and for *future evaluation*
  $\text{data} = (X_i, Y_i)_{i=1}^n \quad \text{data}' = (X_i', Y_i')_{i=1}^n$
- **LIKELIHOOD:** $p(\text{data}|\theta)$
  Independent observations case: $\prod_i p(x_i)p(y_i|x_i, \theta)$.
- **Likelihood Criterion:** Prefer θ with small

$$\log 1/p(\text{data}|\theta)$$

- **Information Theory Viewpoint:** Shannon, Cover, Rissanen
  Prefer shorter codelength.

Maximum Likelihood Estimation

What's good about the maximum likelihood estimate $\hat{\theta}_n$?

- **Short codelength** interpretation provides motivation.
- **Consistency:** Wald(1948) iid case. Target θ^* is limit of $\hat{\theta}$.
  Proof idea: Maximizing likelihood is same as minimizing
  $$\frac{1}{n}\sum_{i=1}^{n} \log p(\text{data}_i|\theta^*)/p(\text{data}_i|\theta),$$
  which (akin to the AEP) is asymptotically close to its expectation
  $$\mathbb{E}\left[\, \log p(\text{data}_1|\theta^*)/p(\text{data}_1|\theta)\,\right],$$
  uniformly so with Wald's finite expected infimum condition, so the empirical minimizer approaches the minimizer of the expectation.
- **Expected Favorability**: Wald(1948), credited to Doob, showed that this expectation, later called Kullback divergence, is indeed positive (also known as the Gibbs, Shannon inequality).
- **Empirical Risk Min**: Gauss, Vapnik least squares, other settings
- **Accuracy:** The finite sample risk is controlled by the best trade-off of Kullback approximation error and metric entropy relative to sample size, as discussed later here.

Maximum Likelihood Estimation

What can go wrong with likelihood maximization?

- **Lack of Parsimony:** For nested models, it prefers larger, more complex, models.
- **Non-adaptive:** Accuracy (or lack thereof) dictated by the largest size, in metric entropy, of the models considered.
- **Over-fit:** Suppose the family includes the target, then $\log 1/p(\text{data}|\hat{\theta})$ will be smaller than $\log 1/p(\text{data}|\theta^*)$.

  Such over-fit is traditionally regarded as problematic.
  We will come back to that.

Penalized Likelihood

Penalized Log Likelihood

$$\log 1/p(\text{data}|\theta) + \text{pen}_n(\theta)$$

Aims of Penalized Log Likelihood

- Overcome limitations of maximum likelihood
- Allow adaptivity
- Overcome problematic over-fit

Penalized Likelihood

Forms of penalized log-likelihood:

- **Bayes**: Prior provides a penalty. Posterior favors smallest
  $$\log 1/p(\text{data}|\theta) \; + \; \log 1/\text{prior}(\theta)$$
- **Minimum Description Length** (MDL):
  Codelength $L_n(\theta)$ for θ, plus codelength for data given θ
  $$\log 1/p(\text{data}|\theta) + L_n(\theta)$$
- **Parameter Dimension Penalty**:
  $\frac{\dim}{2} \log n$ Schwartz BIC, Rissanen MDL.
- **Fisher Information Penalty**:
  $\frac{\dim}{2} \log \frac{n}{2\pi} \; + \; \log\left(|I(\theta)|^{1/2}/w(\theta)\right)$ Barron, Clarke, Rissanen.
- ℓ_1 **Norm Penalty**: prop. to $|\theta|_1 = \sum_{k=1}^{\dim} |\theta_j|$ in linear models.
- ℓ_1 **Norm of Path Weights**: In deep ReLU networks. (e.g. Klusowski, Barron 2020).
- **Roughness Penalty**: e.g. Tapia, Thompson (1978).
- **Structural Minimization**: Vapnik.

Information-Theoretic Unification of Pen Likelihood

Information-Theoretically Valid Penalty: Codelength valid if the Shannon, Kraft inequality $\sum_{\cdot} 2^{-L(\cdot)} \leq 1$ holds for the criterion

$$L(\text{data}) = \min_{\theta \in \Theta} \{\log 1/p(\text{data}|\theta) + \text{pen}_n(\theta)\}$$

Description length interpretation that remains valid for continuous θ.

Mechanisms to Establish Information-Theoretic Validity

- Compare $L(\text{data})$ to the Bayes Mixture Codelength:

  $$\log 1/\int p(\text{data}|\theta) w(\theta) d\theta$$

  Laplace approx. shows Fisher Info penalty is codelength valid
- Compare $L(\text{data})$ to a Discrete Two-Stage MDL:

  $$\min_{\tilde{\theta} \in \tilde{\Theta}} \{\log 1/p(\text{data}|\tilde{\theta}) + L_n(\tilde{\theta})\}$$

  where $\tilde{\Theta}$ is a discrete set and $L_n(\tilde{\theta})$ satisfies the Kraft inequality.
- The ℓ_1 norm penalty $\text{pen}_n(\theta) = \lambda_n |\theta|_1$ is codelength valid for $\lambda_n \geq \sqrt{n \log \text{dim}}$ (Barron, Huang, Li, Liu 2008)

Information-Theoretic Unification of Pen Likelihood

Penalty doubling produces statistical generalization benefits.

Information-Theoretically Valid Penalty: Codelength valid if the Shannon, Kraft inequality $\sum_{\cdot} 2^{-L(\cdot)} \leq 1$ holds for the criterion

$$L(\text{data}) = \min_{\theta \in \Theta} \{\log 1/p(\text{data}|\theta) + \text{pen}_n(\theta)\}$$

Description length interpretation that remains valid for continuous θ.

Mechanisms to Establish Information-Theoretic Validity

- Compare L(data) to the Bayes Mixture Codelength:

  $$\log 1/\int p(\text{data}|\theta) w(\theta) d\theta$$

  Laplace approx. shows Fisher Info penalty is codelength valid
- Compare L(data) to a Discrete Two-Stage MDL:

  $$\min_{\tilde{\theta} \in \tilde{\Theta}} \{\log 1/p(\text{data}|\tilde{\theta}) + 2L_n(\tilde{\theta})\}$$

  where $\tilde{\Theta}$ is a discrete set and $L_n(\tilde{\theta})$ satisfies the Kraft inequality.
- The ℓ_1 norm penalty $\text{pen}_n(\theta) = \lambda_n |\theta|_1$ is codelength valid for $\lambda_n \geq 2\sqrt{n \log \text{dim}}$ (Barron, Huang, Li, Liu 2008)

Statistical Aim

- From training data $\underline{X}, \underline{Y}$ obtain an estimator $\hat{p} = p_{\hat{\theta}}$
- Generalize to subsequent data$' = \underline{X}', \underline{Y}'$
- Want $\log 1/\hat{p}(\text{data}')$ to compare favorably to $\log 1/p(\text{data}')$
- For targets p which are close to or even inside the families
- With data$'$ expectation, loss becomes Kullback divergence
- Bhattacharyya, Hellinger loss also relevant

Loss

- Kullback Information-divergence:

$$D_n(\theta^*||\theta) = \mathbb{E}\big[\ \log p(\text{data}|\theta^*)/p(\text{data}|\theta)\ \big]$$

- Bhattacharyya, Hellinger divergence:

$$d_n(\theta^*||\theta) = 2\log 1/\mathbb{E}[p(\text{data}|\theta)/p(\text{data}|\theta^*)]^{1/2}$$

- Indep. ident. distrib. case: data $= (\text{data}_1, \ldots, \text{data}_n)$

$$D_n(\theta^*\|\theta) = n\, D(\theta^*\|\theta)$$

$$d_n(\theta^*, \theta) = n\, d(\theta^*, \theta)$$

- Relationship: $d \leq D \leq (2 + B)\, d$ if the log density ratio $\leq B$.

Index of Resolvability

The empirical criterion

$$\min_{\theta \in \Theta} \left\{ \log\left[1/p(\text{data}|\theta)\right] + \text{pen}_n(\theta) \right\}$$

equivalently

$$\min_{\theta \in \Theta} \left\{ \log\left[p(\text{data}|\theta^*)/p(\text{data}|\theta)\right] + \text{pen}_n(\theta) \right\}$$

has the population counterpart

$$\min_{\theta \in \Theta} \left\{ D_n(\theta^* || \theta) + \text{pen}_n(\theta) \right\}$$

The minimizing parameter θ_n^* best resolves the target.

Dividing by n yields a statistical rate, the index of resolvability

$$R_n(\theta^*) = \frac{1}{n} \min_{\theta \in \Theta} \left\{ D_n(\theta^* || \theta) + \text{pen}_n(\theta) \right\}$$

For instance, in the i.i.d. case

$$R_n(\theta^*) = \min_{\theta \in \Theta} \left\{ D(\theta^* || \theta) + \frac{\text{pen}_n(\theta)}{n} \right\}$$

Conservative bound

$$R_n(\theta^*) \leq \frac{\text{pen}_n(\theta^*)}{n}$$

One-sided empirical analysis reveals generalization

Idea: **empirical process error may be complexity dependent**

- log likelihood-ratio discrepancy for training and future data

$$\left[\log \frac{p(\text{data}|\theta^*)}{p(\text{data}|\theta)} - d_n(\theta^*, \theta)\right]$$

- Instead, we examine the *penalized discrepancy*

$$\min_{\theta \in \Theta} \left\{ \left[\log \frac{p(\text{data}|\theta^*)}{p(\text{data}|\theta)} - d_n(\theta^*, \theta)\right] + pen_n(\theta) \right\}$$

- **Risk validity condition**: Penalized discrepancy is at least

$$\min_{\tilde{\theta} \in \tilde{\Theta}} \left\{ \left[\log \frac{p(\text{data}|\theta^*)}{p(\text{data}|\tilde{\theta})} - d_n(\theta^*, \tilde{\theta})\right] + 2L_n(\tilde{\theta}) \right\}$$

  where $\tilde{\Theta}$ is a discrete set and $L_n(\tilde{\theta})$ satisfies the Kraft inequality.
- **Key to statistical analysis**:
  With risk valid penalty, the *penalized discrepancy*
  - has expectation greater than or equal to zero and
  - is stochastically greater than minus an exponential(1) r.v.

Li, Barron 1998; extended in Barron, Huang, Li, Luo 2008.

Risk Bounds and Confidence Bounds

For any risk valid $\mathrm{pen}_n(\theta)$, the *penalized discrepancy*

$$\min_{\theta \in \Theta} \left\{ \left[\log \frac{p(\text{data}|\theta^*)}{p(\text{data}|\theta)} - d_n(\theta^*, \theta) \right] + pen_n(\theta) \right\}$$

- has expectation greater than or equal to zero and
- is stochastically greater than minus an exponential(1) r.v.

Risk bound: Apply the expectation inequality at the penalized log likelihood optimizer $\hat{\theta}$ to get the risk bound (from Li, Barron 1998, Grunwald 2007, with extension in Barron, Huang, Li, Liu 2008)

$$\mathbb{E}[d(\theta^*, \hat{\theta})] \leq \frac{1}{n} \mathbb{E} \min_{\theta \in \Theta} \left\{ \log \frac{p(\text{data}|\theta^*)}{p(\text{data}|\theta)} + pen_n(\theta) \right\}.$$

Hence, since the expected min is less than the min of expectations,

$$\mathbb{E}[d(\theta^*, \hat{\theta})] \leq R_n(\theta^*).$$

Thus the population resolvability controls the estimation risk.
Analogous conclusion holds for general (non-iid) models.

Risk Bounds and Confidence Bounds

For any risk valid $pen_n(\theta)$, the *penalized discrepancy*

$$\min_{\theta \in \Theta} \left\{ \left[\log \frac{p(\text{data}|\theta^*)}{p(\text{data}|\theta)} - d_n(\theta^*, \theta) \right] + pen_n(\theta) \right\}$$

- has expectation greater than or equal to zero and
- is stochastically greater than minus an exponential(1) r.v.

Confidence region: Apply the stochastic inequality to any estimate $\hat{\theta}$ to get the following confidence statement. In an event of probability at least $1 - \delta$

$$d(\theta^*, \hat{\theta}) \leq \frac{1}{n} \log \frac{p(\text{data}|\theta^*)}{p(\text{data}|\hat{\theta})} + \frac{pen_n(\hat{\theta})}{n} + \frac{\log 1/\delta}{n}$$

In particular, for any over-fit estimate $\hat{\theta}$, with the same prob,

$$d(\theta^*, \hat{\theta}) \leq \frac{pen_n(\hat{\theta})}{n} + \frac{\log 1/\delta}{n}$$

Risk Bounds and Confidence Bounds

- **Confidence region**: In an event of probability at least $1-\delta$

$$d(\theta^*,\hat{\theta}) \;\le\; \frac{1}{n}\log\frac{p(\text{data}|\theta^*)}{p(\text{data}|\hat{\theta})} \;+\; \frac{pen_n(\hat{\theta})}{n} \;+\; \frac{\log 1/\delta}{n}.$$

  In particular, for any over-fit estimate $\hat{\theta}$, with the same prob,

$$d(\theta^*,\hat{\theta}) \;\le\; \frac{pen_n(\hat{\theta})}{n} \;+\; \frac{\log 1/\delta}{n}$$

- **Implication for linear models and for deep ReLU nets**: for any over-fit estimate $\hat{\theta}$, with prob at least $1-\delta$,

$$d(\theta^*,\hat{\theta}) \le 2|\hat{\theta}|_1\sqrt{\frac{\log \dim}{n}} + \frac{Const}{n} + \frac{\log 1/\delta}{n}$$

- **A fitted over-parameterized deep net** with small ℓ_1 path norm compared to $\sqrt{n/\log \dim}$ yields appropriately confident in the indicated accuracy of generalization.
- Provides understanding of sometimes *benign* over-fitting.

Summary

Statistics and information theory are fundamentally intertwined.

General one-sided penalized empirical proc. analysis provides:

- Risk bound by the index of resolvability.
- Confidence bound from observed penalty, log-likelihood
- Fundamental connection between empirically valid penalties and information -theoretically valid penalties.
- Surprisingly valid penalties.
- Explanation for benign over-fitting.

Extra: Better Risk Bounds for Bayes Estimation

- From prior $\pi(\theta)$ and data get posterior $\pi(\theta|\text{data}^n)$
- Suppose $(\text{data}_1, \ldots, \text{data}_N, \text{data}')$ are i.i.d. $p_{\theta^*}(\cdot) = p(\cdot|\theta^*)$
- Bayes predictive distribution provides a density estimate

$$\hat{p}_n(\text{data}') = p(\text{data}'|\text{data}^n) = \int p(\text{data}'|\theta)\pi(\theta|\text{data}^n)d\theta$$

- Time average Kullback risk $\bar{r}_N(\theta^*) = \frac{1}{N+1}\sum_{n=0}^{N} \mathbb{E}\, D(p_{\theta^*}\|\hat{p}_n)$
- Resolvability bound (Barron 1986,1998)

$$\bar{r}_N(\theta^*) \le \min_B \left\{ \max_{\theta\in B} D(p_{\theta^*}\|p_\theta) + \frac{1}{N+1}\log\frac{1}{\pi(B)} \right\}$$

- Example: Discrete parameter and singleton sets $B = \{\theta\}$

$$\bar{r}_N(\theta^*) \le \min_\theta \left\{ D(p_{\theta^*}\|p_\theta) + \frac{1}{N+1}\log\frac{1}{\pi(\theta)} \right\}$$

and in particular

$$\bar{r}_N(\theta^*) \le \frac{1}{N+1}\log\frac{1}{\pi(\theta^*)}$$

Extra: Better Risk Bounds for Bayes Estimation

- Consequence using convexity of Kullback divergence
- Time average estimate

$$\hat{\hat{p}}(data') = \frac{1}{N+1} \sum_{n=0}^{N} p(data' | data^n)$$

where data^n may use the n most recent observations.

- Kullback risk

$$\mathbb{E}\, D(p_{\theta*} \| \hat{\hat{p}}) \leq \bar{r}_N$$

- Thus have estimator with risk at least as good as the time average risk of Bayes predictive estimators
- As we saw, this risk is controlled by the resolvability