Bayesian Supervised Learning using Full and Limited Support Priors

Doctoral Qualifying Examination

Paul Rademacher

The George Washington University
Department of Electrical and Computer Engineering

December 6, 2019

Table of Contents

Table of Contents

- The debate as to whether or not Bayesian approaches are suitable for applications of statistics such as detection and estimation has a long history[1]
- The distinction between deterministic and Bayesian methods in classical inference has been inherited by the widely-popular field of Machine Learning
- Specific focus on *parametric learning* is a consequence of the constraints of real-world implementation - efficient data representations are needed for practical learning solutions

[1] George E.P. Box et al. *Bayesian Inference in Statistical Analysis*. Wiley, 1973.

From Classical Inference to Machine Learning

Part II

- Much of the attention on machine learning today can be attributed to the resurgence of the multilayer perceptron and the success of deep neural networks (DNN) on classification benchmark challenges
  - Ex: Speech recognition on the TIMIT database[2], ILSVRC 2010[3]
- Many researchers attribute these advances to increased computing power, enabling the training of large numbers of parameters on voluminous collections of high-dimensional data
- Like other historically popular supervised learning algorithms (support vector machines, decision trees, etc.), these deep learning algorithms do not derive from a Bayesian viewpoint.

---

[2] Abdel-rahman Mohamed et al. "Deep Belief Networks for phone recognition". In: *Science* (2010).

[3] Alex Krizhevsky et al. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems*. 2012.

The Bayesian Perspective

Part I

- Should the unknown parameters $\theta \in \Theta$ that statistically model the data $\mathrm{D} \in \mathcal{D}$ be treated as random?
  - **Classical perspective**: there are often no environmental factors that suggest the model is randomly generated
  - **Bayesian perspective**: prior knowledge reflects the user's confidence in different data-generating models before data is observed[4]
- The success or failure of Bayesian learning methods hinges on how well the prior distribution selected by the designer matches reality:

Non-Informative

- Weights the models without preference, lets data "speak for itself"
- Robust solution for all models, but may be insufficient if data volume is limited

Informative

- Accurate localized priors enable low risk learning, even with limited training data
- Poor prior design leads to worst-case performance

[4] George E.P. Box et al. *Bayesian Inference in Statistical Analysis*. Wiley, 1973.

The Bayesian Perspective

Part II

- With high-dimensionality data, designing a sensible prior distribution for Bayesian learning is challenging and often prohibitive. However, this complexity is also reflective of the wide variety of learning approaches afforded
- Many learning methods based on a deterministic treatment of the data-generating model have equivalents in Bayesian learning
  - Classical estimation via Maximum Likelihood is identical to Bayesian Maximum *a posteriori* estimation with a uniform prior[5]:

$$\hat{\theta}_{\mathrm{ML}}(D) = \arg\max_{\theta} \mathrm{P}_{\mathrm{D}\,|\,\theta}(D|\theta) \equiv \arg\max_{\theta} \mathrm{P}_{\theta|\,\mathrm{D}}(\theta|D) = \hat{\theta}_{\mathrm{MAP}}(D)$$

  - Empirical squared-error minimization with L_2 regularization is equivalent to Bayesian MAP with Gaussian likelihood and prior functions[6]

Many non-Bayesian learning methods implicitly express a lack of model preference

[5] Steven M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, 1993.

[6] Sergios Theodoridis. *Machine Learning: A Bayesian and Optimization Perspective*. Elsevier, 2015.

Table of Contents

Learning via Empirical Risk Minimization

Part I

- Design of a parametric learning function can be decomposed into two topics:
  - **Parameter Training**: Definition of how the training data $\mathrm{D} \in \mathcal{D}$ maps to the finite-dimensional parameter space Θ
  - **Parameter Mapping**: Specification of how the parameters $\theta \in \Theta$ map to the higher-dimensional space of decision functions
- The parameter training operators used by many modern non-Bayesian methods, specifically empirical risk minimization, are not notably different from operators used historically

What about the Parameter Mapping of popular non-Bayesian methods makes them effective for certain applications?

Learning via Empirical Risk Minimization

Part II

- The failures of non-Bayesian methods are frequently attributed to a phenomenon termed overfitting, where the decision function achieves low empirical risk but performs poorly on novel data not used during training
- Even simple learning algorithms such as Nearest-Neighbor[7] can achieve minimal empirical risk

  $\Rightarrow$ The empirical risk is not actually the objective function that the designer wants to minimize

- Parametric learning approaches implicitly disallow the use of the complete function space. The selection of this function subspace should be thought of as the imposition of *prior knowledge*

[7] Luc Devroye et al. *A Probabilisitic Theory of Pattern Recognition*. Springer-Verlag, 1996.

The Necessity of Prior Knowledge

- This new perspective enables a Bayesian interpretation of all parametric learning methods, even classical methods
  - Essentially all uses of maximum likelihood search a finite-dimensional subspace of the set of probability functions[8] - this is equivalent to Bayesian MAP with a degenerate prior
- To define a learning approach that truly gives no preference to any data-generating model, the full set of probability distributions should be used
  - The maximum likelihood estimate is the empirical distribution generated using the training data[9], dictating the use of the empirical risk metric

The use of prior knowledge is mandatory for effective machine learning on most data-limited applications

[8] Athanasios Papoulis et al. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 2002.

[9] C. Radhakrishna Rao. "Maximum Likelihood Estimation for the Multinomial Distribution". In: *Sankhyā: The Indian Journal of Statistics* (1957).

Table of Contents

Research Goals

THEORY:

Analyze how prior knowledge of varying subjectivity affects parametric learning for different volumes of training data

APPLICATION:

Use our prior knowledge for regression and classification problems that mimic human prediction/recognition tasks

Data Model

Observable random element: $\mathrm{x} \in \mathcal{X}$

Unobservable random element: $\mathrm{y} \in \mathcal{Y}$

Observable training data: $\mathrm{D} \in \mathcal{D} = \{\mathcal{Y} \times \mathcal{X}\}^N$

Independently, identically distributed according to an unknown probability mass function (PMF)

$$\theta \in \Theta = \left\{ \theta \in \mathbb{R}_{\geq 0}{}^{\mathcal{Y} \times \mathcal{X}} : \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \theta(y, x) = 1 \right\} ,$$

such that $\mathrm{P}_{\mathrm{y},\mathrm{x}\,|\,\theta}(y, x|\theta) = \mathrm{P}_{\mathrm{D}_n\,|\,\theta}(y, x|\theta) = \theta(y, x)$.

Alternate Notation: $\theta \Leftrightarrow \left(\theta', \tilde{\theta}\right)$

- Marginal model $\theta' \equiv \sum_{y \in \mathcal{Y}} \theta(y, \cdot) = \mathrm{P}_{\mathrm{x}\,|\,\theta}$ over the set $\mathcal{X}$
- Conditional models $\tilde{\theta}(\mathrm{x}) \equiv \theta(\cdot, \mathrm{x})/\theta'(x) = \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$ over the set $\mathcal{Y}$

Objective

Design Metric

Decisions: $h \in \mathcal{H}$
Loss function: $\mathcal{L} : \mathcal{H} \times \mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$

Regression: the squared-error loss
$\mathcal{L}(h, y) = (h - y)^2$

Classification: the 0–1 loss
$\mathcal{L}(h, y) = 1 - \delta[h, y]$

Design Task

Create a decision function $f : \mathcal{D} \mapsto \mathcal{H}^{\mathcal{X}}$ that minimizes the conditional expected loss, or conditional "risk",

$$\begin{aligned}\mathcal{R}_{\Theta}(f; \theta) &= \mathrm{E}_{\mathrm{y,x,D}\,|\theta}\left[\mathcal{L}\left(f(\mathrm{x}; \mathrm{D}), \mathrm{y}\right)\right] \\ &= \mathrm{E}_{\mathrm{D}\,|\theta}\left[\mathrm{E}_{\mathrm{x}\,|\theta}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\mathcal{L}\left(f(\mathrm{x}; \mathrm{D}), \mathrm{y}\right)\right]\right]\right].\end{aligned}$$

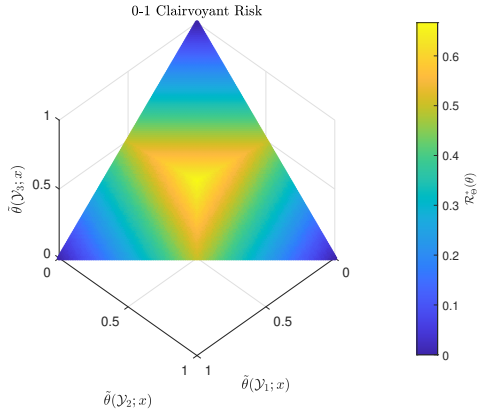
Clairvoyant Decision and Risk

The "clairvoyant"[10] decision function $f_\theta : \Theta \mapsto \mathcal{H}^{\mathcal{X}}$ is

$$
\begin{aligned}
f_\Theta(\mathrm{x};\theta) &= \arg\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\,\mathcal{L}(h,\mathrm{y})\right] \\
&\equiv \arg\min_{h\in\mathcal{H}} \sum_{y\in\mathcal{Y}} \tilde{\theta}(y;\mathrm{x})\,\mathcal{L}(h,y)
\end{aligned}
$$

$$
\Downarrow \quad \Downarrow
$$

$$
\begin{aligned}
\mathcal{R}^*_\Theta(\theta) &\equiv \mathrm{E}_{\mathrm{x}\,|\theta}\left[\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\,\mathcal{L}(h,\mathrm{y})\right]\right] \\
&\leq \mathcal{R}_\Theta(f;\theta)
\end{aligned}
$$



Clairvoyant Risk = Lower Bound for Conditional Risk

[10] Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*.

Human Recognition Tasks

Part I

- Our recognition abilities can be partially attributed to the immeasurable volume of training data that we consume, but we also have a type of prior knowledge

- Our sensory inputs are highly structured due to the phenomenology of the physical world

- Non-invertible feature extraction occurs before memorization - we possess a mechanism for "dimensionality reduction"

Figure: Randomly generated RGB image

Human Recognition Tasks

Part II

- **Perspective**: the labels we use are extrinsic to our observations

$$\Downarrow$$

- We defined the classes ourselves, ensuring joint probability distributions that lead to low clairvoyant risk

$$\Downarrow$$

- With accurate prior knowledge, parametric decision functions that achieve minimal risk are possible

Figure: Handwritten digit samples from the MNIST Database [12]

[12] Lecun et al., "Gradient-based learning applied to document recognition".

Table of Contents

Statistical Learning Work

- Fundamental results in classical statistics include the existence of asymptotically consistent estimators[13] and the development of error bounds for empirical risk minimization[14]
- Practical methods for prediction in supervised Bayesian learning will be relevant for certain priors
  - Approximate inference methods such as variational approximation[15] enable posterior evaluation for intractable problems
  - Non-deterministic approaches including Markov chain Monte Carlo (MCMC) methods[16] provide estimates of the predictive distribution
- Dirichlet processes have been considered in non-parametric learning, but typically only for unsupervised learning with infinite numbers of clusters[17]

[13] Charles J. Stone. "Consistent Nonparametric Regression". In: *The Annals of Statistics* 5.4 (1977).

[14] V. Vapnik et al. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities". In: *Theory of Probability and its Applications*. Vol. 16. 1971.

[15] Matthew J. Beal. "Variational algorithms for approximate Bayesian inference". PhD thesis. University College London, 2003.

[16] W. K. Hastings. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57.1 (1970).

[17] Samuel J. Gershman et al. "A tutorial on Bayesian nonparametric models". In: *Journal of Mathematical Psychology* 56 (2012).

Feature Extraction for Speech/Image Recognition

- Our sensory processing systems decompose information at a variety of scales - multiresolution signal decomposition with wavelet representations[18] is popular for these learning applications
- For tasks such as optical character recognition (OCR), the translation dependency of a feature can significantly raise the probability of error; reconciling our desire for features that describe spatial locality with the need for translation-invariant representation is challenging
  - Even popular features such as those based on multiresolution wavelet transformations[19] do not provide both
  - Focused research on shift-invariant features for 2-dimensional data[20] aims to ensure this property

[18] S. G. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. 3rd. Academic Press, Inc., 2008.

[19] S. G. Mallat. "A theory for multiresolution signal decomposition: the wavelet representation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11.7 (1989).

[20] S. Del Marco et al. "An M-band, 2-dimensional translation-invariant wavelet transform and applications". In: *1995 International Conference on Acoustics, Speech, and Signal Processing*. Vol. 2. 1995.

Table of Contents

Generalized Bayesian Framework

- Having motivated the perspective that *all* machine learning algorithms either implicitly or explicitly use prior knowledge, this thesis will focus on a strictly Bayesian approach to parametric supervised learning
- Existing treatments of parametric Bayesian learning effect significant data dimensionality reduction and yet fail to sufficiently consider the implications of how the mapping between the parameters and the probability model is selected
- Designs using both full support and limited support prior distributions will be analyzed and subsequently applied to human recognition tasks

$\Downarrow \quad \Downarrow$ **Model Unknown. Select Prior** $\mathbf{p_\theta}$ $\quad \Downarrow \quad \Downarrow$

$$\mathcal{R}(f) = \mathrm{E}_\theta \left[\mathcal{R}_\Theta(f;\theta) \right] = \mathrm{E}_\mathrm{D} \left[\mathrm{E}_{\mathrm{x}\,|\,\mathrm{D}} \left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} \left[\mathcal{L}\left(f(\mathrm{x};\mathrm{D}), \mathrm{y} \right) \right] \right] \right]$$

Optimal Decision:

$$f^*(\mathrm{x};\mathrm{D}) = \arg\min_{h \in \mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} \left[\mathcal{L}(h, \mathrm{y}) \right]$$

Minimum Bayes Risk:

$$\mathcal{R}^* \equiv \mathrm{E}_\mathrm{D} \left[\mathrm{E}_{\mathrm{x}\,|\,\mathrm{D}} \left[\min_{h \in \mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} \left[\mathcal{L}(h, \mathrm{y}) \right] \right] \right]$$

Predictive Distributions

Bayesian PMF is the conditional expectation of the true PMF $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} \equiv \tilde{\theta}(\mathrm{x})$:

$$\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} = \mathrm{E}_{\theta|\,\mathrm{x},\mathrm{D}} \left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} \right] \equiv \mu_{\tilde{\theta}(\mathrm{x})|\,\mathrm{x},\mathrm{D}}$$

Training Data Sufficient Statistic

Likelihood Function:

$$\mathrm{P}_{\mathrm{D}\,|\theta}(D|\theta) = \prod_{y\in\mathcal{Y}}\prod_{x\in\mathcal{X}} \theta(y,x)^{\bar{N}(y,x;D)}$$

Transform: $\bar{N} : \mathcal{D} \mapsto \bar{\mathcal{N}}$
$\bar{N}(y,x;D) = \sum_{n=1}^{N} \delta\big[(y,x), D_n\big]$

$$\bar{\mathcal{N}} = \left\{ \bar{n} \in \mathbb{Z}_{\geq 0}{}^{\mathcal{Y}\times\mathcal{X}} : \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} \bar{n}(y,x) = N \right\}$$

- Empirical count $\bar{N}(\mathrm{D})$ is a sufficient statistic[21] for the model θ
- $|\bar{\mathcal{N}}| = \binom{N+|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1}{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \leq |\,\mathcal{D}\,|$

$$\Downarrow \quad \Downarrow$$

Efficient Representation

Express distributions using new random process $\bar{\mathrm{n}} \equiv \bar{N}(\mathrm{D}) \in \bar{\mathcal{N}}$

- "Marginal" process: $\mathrm{n}' \equiv \sum_{y\in\mathcal{Y}} \bar{\mathrm{n}}(y,\cdot)$

[21] Bernardo et al., *Bayesian Theory*.

Full Support Priors

- Full support priors are necessary to ensure asymptotically consistent estimation of any model θ and thus optimal decisions in the limit of training data volume N
- Often, priors are termed non-informative as long as they are approximately uniform over their limited support - to be *truly* non-informative, the prior must have full support
- The Dirichlet distribution has full support over the space of data-generating distributions and can be parameterized in different ways to achieve both informative and non-informative priors
  - Additionally, it is a *conjugate prior* for likelihood functions in the exponential family[22], enabling tractable posterior evaluation for I.I.D. training data

[22] Sergios Theodoridis. *Machine Learning: A Bayesian and Optimization Perspective*. Elsevier, 2015.

- The model probability density function (PDF) is Dirichlet[23] with parameters $\alpha : \mathcal{Y} \times \mathcal{X} \mapsto \mathbb{R}^+$:

Joint Model:

$$\begin{aligned} \mathrm{p}_\theta(\theta) &= \mathrm{Dir}\left(\theta; \alpha\right) \\ &= \beta(\alpha)^{-1} \prod_{y \in \mathcal{Y}} \prod_{x \in \mathcal{X}} \theta(y,x)^{\alpha(y,x)-1} \end{aligned}$$

$\Rightarrow$

$\Rightarrow$

Conditional Model:

$$\begin{aligned} \mathrm{p}_{\tilde{\theta}|\theta'}\left(\tilde{\theta}|\theta'\right) &= \mathrm{p}_{\tilde{\theta}}\left(\tilde{\theta}\right) \\ &= \prod_{x \in \mathcal{X}} \mathrm{Dir}\left(\tilde{\theta}(x); \alpha(\cdot, x)\right) \end{aligned}$$

Concentration: $\alpha_0 \equiv \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \alpha(y,x)$

Concentration: $\alpha'(x) \equiv \sum_{y \in \mathcal{Y}} \alpha(y,x)$

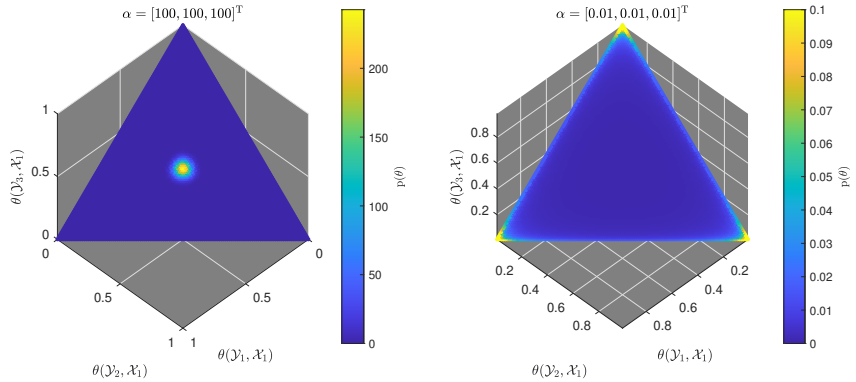
True predictive distribution $\tilde{\theta}$ is independent of the marginal model θ' and inherits a Dirichlet PDF

[23] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Dirichlet Prior Distribution

Part II: Concentration Trends

- Non-informative: $\alpha = 1$ creates the uniform Dirichlet prior, $\mathrm{p}_\theta = \big(|\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1\big)!$
- Maximal ($\alpha_0 \to \infty$) and minimal ($\alpha_0 \to 0$) localization around mean $\mu_\theta = \frac{\alpha}{\alpha_0}$



Limited Support Priors

- The most informative priors have limited support, specifically, support of limited dimensionality, such that $\dim(\Theta) < |\mathcal{Y}||\mathcal{X}| - 1$
- The performance trade-offs dependent on the support dimensionality will be a primary focus
  - Potential risk reduction relative to full support priors for data-limited problems
  - Failure to meet the clairvoyant risk in the limit of increasing training data volume
- Special attention will be given to the design of low-dimensionality support sets that are sensible for human recognition tasks

Fewer degrees of freedom needed to characterize the support

$\Downarrow$

Minimal risk achievable by simpler parametric functions

Informative Priors for Human Recognition Tasks

Part I: Data Dimensionality

- Data processed during human recognition is highly structured

$$\Downarrow$$

- Relatively low *intrinsic dimensionality*, limited support data model θ'

$$\Downarrow$$

$$\left\|\theta'\right\|_0 = \sum_{x \in \mathcal{X}} \Big(1 - \delta[\theta'(x), 0]\Big) \leq M_{\mathcal{X}} < |\,\mathcal{X}\,|$$

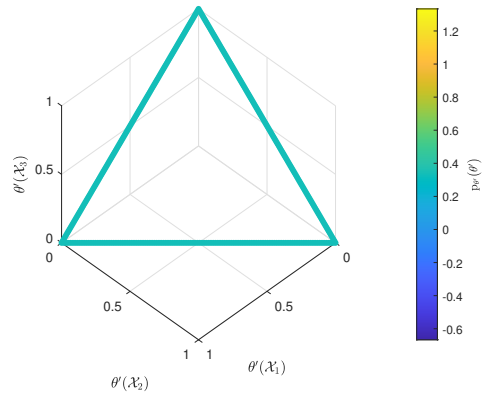


Figure: Marginal model prior for $\left\|\theta'\right\|_0 \leq 2$

Informative Priors for Human Recognition Tasks

Part II: Sufficient Statistics

- Our internal "pre-processing" greatly reduces the complexity of our sensory input

$$\Downarrow$$

- Parametric learning functions should be able to perform feature extraction and effect significant dimensionality reduction without any loss of performance

$$\Downarrow$$

- The prior distribution's support should be limited to a subspace that guarantees a low-dimensional sufficient statistic

- Define transform $T : \mathcal{X} \mapsto \mathcal{T}$, such that $\mathcal{X}_{\mathrm{s}}(t) = \{x \in \mathcal{X} : T(x) = t\}$
- **Data reduction**: $|\mathcal{T}| < |\mathcal{X}|$

Sufficiency

Statistic $\mathrm{t} \equiv T(\mathrm{x})$ must satisfy[24]

$$\mathrm{P}_{\mathrm{x}|\theta',\mathrm{t}}(x|\theta', t) = \frac{\theta'(x)}{\sum_{x' \in \mathcal{X}_{\mathrm{s}}(t)} \theta'(x')} = g(x; t)$$

$\forall\, x \in \mathcal{X}_s(t),\ t \in \mathcal{T}$, with no additional dependency on the model

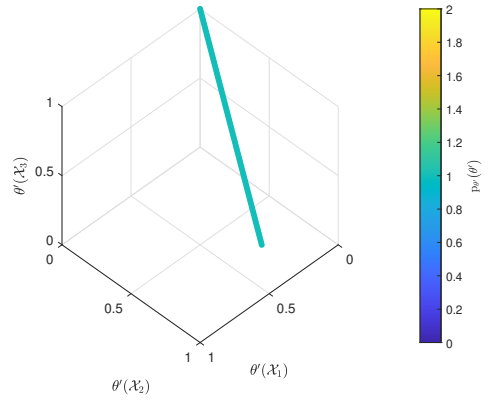


Figure: A marginal model prior for $|\mathcal{T}| = 2$

[24] Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*.

- Humans effectively perform *unsupervised learning* on sensory observations, creating a joint distribution θ that leads to low clairvoyant risk $\mathcal{R}^*_\Theta(\theta)$

$$\Downarrow \quad \Downarrow$$

- The true predictive models $\tilde{\theta}(x)$ should be highly definitive, having low entropy or demonstrating sparsity

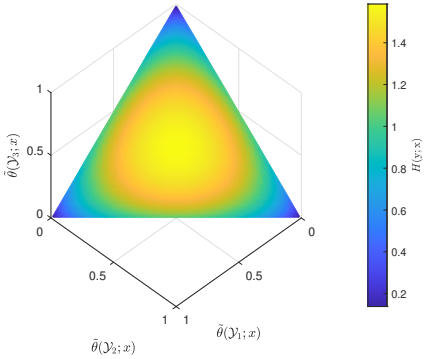


Figure: Conditional entropy for $\tilde{\theta}(x)$

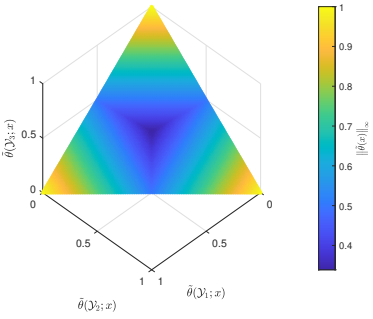
Informative Priors for Human Recognition Tasks

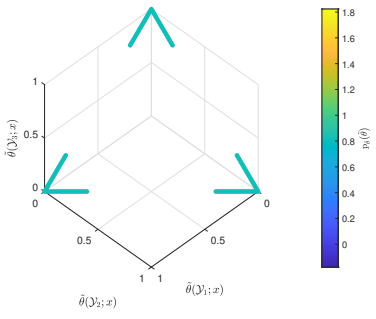
Part III: Low Risk Prediction

L^p-norm restrictions:

$$\left\|\tilde{\theta}(x)\right\|_\infty = \max_{y\in\mathcal{Y}} |\tilde{\theta}(y;x)| \geq \rho > 1/|\,\mathcal{Y}\,|$$

$$\left\|\tilde{\theta}(x)\right\|_0 \leq M_{\mathcal{Y}} < |\,\mathcal{Y}\,|$$





Condition $\left\|\tilde{\theta}(x)\right\|_\infty = 1$ restricts the support of the predictive model to a countable set, limiting computational complexity of implementation

Table of Contents

Publications

- Initial work has focused on optimal decision functions derived from Dirichlet prior distributions for finite sets $\mathcal{Y}$ and $\mathcal{X}$
- Risk analysis of resultant Bayesian estimates and hypotheses for regression and classification, respectively, has been performed
- Publications to date:
  - *Predictive Distribution Estimation for Bayesian Machine Learning using a Dirichlet Prior*, presented at the 53rd Annual Asilomar Conference on Signals, Systems, and Computers
  - *Bayesian Learning for Classification using a Uniform Dirichlet Prior*, presented at the 7th IEEE Global Conference on Signal and Information Processing

Training Data Characterization

Empirical Statistic Likelihood:

$$\begin{aligned} \mathrm{P}_{\bar{\mathrm{n}}\,|\theta}(\bar{n}|\theta) &= \mathrm{Multi}(\bar{n}; N, \theta) \\ &= \mathcal{M}(\bar{n}) \prod_{y \in \mathcal{Y}} \prod_{x \in \mathcal{X}} \theta(y,x)^{\bar{n}(y,x)} \end{aligned}$$

$\Rightarrow$

Statistic "Evidence":

$$\begin{aligned} \mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) &= \mathrm{DM}(\bar{n}; N, \alpha) \\ &= \mathcal{M}(\bar{n}) \beta(\alpha)^{-1} \beta(\alpha + \bar{n}) \end{aligned}$$

Marginal Statistic: $\mathrm{n}'\,|\theta \sim \mathrm{Multi}(N, \theta')$

Conditional Statistic:
$\bar{\mathrm{n}}(\cdot, x)|\,\mathrm{n}'(x), \theta \sim \mathrm{Multi}\left(\mathrm{n}'(x), \tilde{\theta}(x)\right)$

$\Rightarrow$

Marginal Statistic:
$\mathrm{n}' \sim \mathrm{DM}(N, \alpha')$

Conditional Statistic:
$\bar{\mathrm{n}}(\cdot, x)|\,\mathrm{n}'(x) \sim \mathrm{DM}\left(\mathrm{n}'(x), \alpha(\cdot, x)\right)$

Dirichlet Prior leads to tractable Dirichlet-Multinomial[25] Evidence function

[25] Norman L. Johnson et al. *Discrete Multivariate Distributions*. John Wiley & Sons, 1997.

Model Posterior Distribution

Part I: Closed-Form

- Independence of $\tilde{\theta}$ from θ' implies conditional independence of $\tilde{\theta}$ from x given $\bar{n}$
  - No "unsupervised" inference is possible
- Since the likelihood $P_{\bar{n}|\theta}$ has exponential form, the Dirichlet PDF is a conjugate prior

$$\begin{aligned} p_{\tilde{\theta}|\bar{n},x}\left(\tilde{\theta}|\bar{n},x\right) = p_{\tilde{\theta}|\bar{n}}\left(\tilde{\theta}|\bar{n}\right) &= \prod_{x' \in \mathcal{X}} p_{\tilde{\theta}(x')|\bar{n}(\cdot,x')}\left(\tilde{\theta}(x')|\bar{n}(\cdot,x')\right) \\ &= \prod_{x' \in \mathcal{X}} \mathrm{Dir}\left(\tilde{\theta}(x'); \alpha(\cdot,x') + \bar{n}(\cdot,x')\right) \end{aligned}$$

Posterior for model $\tilde{\theta}(x)$ is Dirichlet and dependent solely on the sufficient statistic elements $\bar{n}(\cdot, x)$

Model Posterior Distribution

Part II: Asymptotic Trends

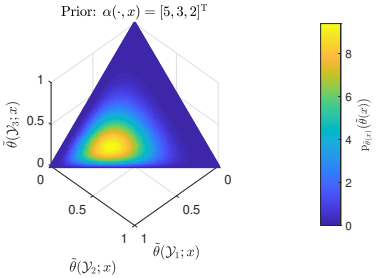
- Covariance of Dirichlet $\mathrm{p}_{\tilde{\theta}(x)|\,\bar{n}(\cdot,x)}$ decreases monotonically with concentration $\alpha'(x) + n'(x)$

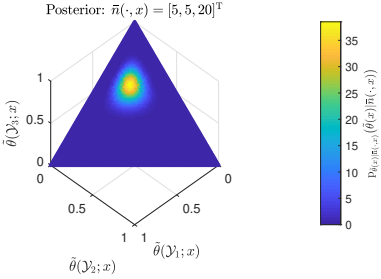
$$\Downarrow \quad \Downarrow$$

- As $n'(x) \to \infty$, the posteriors tend to

$$\mathrm{p}_{\tilde{\theta}(x)|\,\bar{n}(\cdot,x)}\left(\tilde{\theta}(x)|\bar{n}(\cdot,x)\right) \to \delta\left(\tilde{\theta}(x) - \mu_{\tilde{\theta}(x)|\,\bar{n}(\cdot,x)}\right)$$

Asymptotically consistent estimation of $\tilde{\theta}$ due to full support of prior





Bayesian Predictive Distribution

The Bayesian predictive distribution is a convex combination of two conditional PMF's:

$$\begin{aligned} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(\cdot|x,\bar{n}) &= \mu_{\tilde{\theta}(x)|\,\bar{\mathrm{n}}(\cdot,x)}\big(\bar{n}(\cdot,x)\big) \\ &\equiv \left(\frac{\alpha'(x)}{\alpha'(x)+n'(x)}\right)\frac{\alpha(\cdot,x)}{\alpha'(x)} + \left(\frac{n'(x)}{\alpha'(x)+n'(x)}\right)\frac{\bar{n}(\cdot,x)}{n'(x)} \end{aligned}$$

Prior Mean Uniform PMF $\mu_{\tilde{\theta}(x)} = \alpha(\cdot,x)/\alpha'(x)$ dependent only on the Dirichlet parameterization

Conditional Empirical PMF $\bar{n}(\cdot,x)/n'(x)$ dependent only on training data

As the data volume $n'(x)$ increases relative to the concentration $\alpha'(x)$, the predictive distribution tends toward the empirical PMF

Assess estimation of θ using a difference process: $\Delta(\mathrm{x}, \bar{\mathrm{n}}, \theta) \equiv \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} - \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$

$$\begin{aligned}\mathrm{Bias}(\mathrm{x}, \mathrm{n}', \theta) &= \mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\Delta(\mathrm{x}, \bar{\mathrm{n}}, \theta)\right] \\ &= \frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})}\left(\frac{\alpha(\cdot, \mathrm{x})}{\alpha'(\mathrm{x})} - \tilde{\theta}(\mathrm{x})\right)\end{aligned}$$

$\alpha'(x)$	Bias
$\to \infty$	$\frac{\alpha(\cdot,\mathrm{x})}{\alpha'(\mathrm{x})} - \tilde{\theta}(\mathrm{x})$
$\to 0$	0

$$\begin{aligned}\mathrm{Cov}(y, y'; \mathrm{x}, \mathrm{n}', \theta) &= \mathrm{C}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(\cdot\,|\,\mathrm{x}, \bar{\mathrm{n}})\right](y, y') \\ &= \frac{\mathrm{n}'(\mathrm{x})}{\left(\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})\right)^2}\left(\tilde{\theta}(y; \mathrm{x})\delta[y, y'] - \tilde{\theta}(y; \mathrm{x})\tilde{\theta}(y'; \mathrm{x})\right)\end{aligned}$$

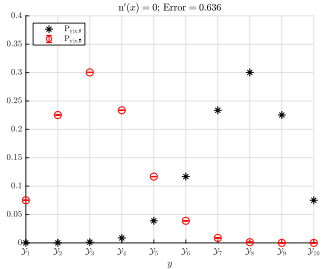
$\alpha'(x)$	Covariance
$\to \infty$	0
$\to 0$	$\frac{\Sigma_{\bar{\mathrm{n}}(\cdot,\mathrm{x})\,|\,\mathrm{n}'(\mathrm{x}),\tilde{\theta}(\mathrm{x})}(y,y')}{\mathrm{n}'(\mathrm{x})^2}$

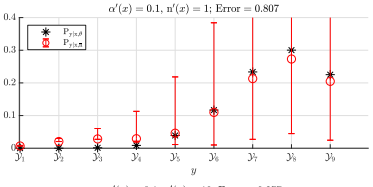
$$\mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\Delta(y; \mathrm{x}, \bar{\mathrm{n}}, \theta)\Delta(y'; \mathrm{x}, \bar{\mathrm{n}}, \theta)\right] = \mathrm{Bias}(y; \mathrm{x}, \mathrm{n}', \theta)\mathrm{Bias}(y'; \mathrm{x}, \mathrm{n}', \theta) + \mathrm{Cov}(y, y'; \mathrm{x}, \mathrm{n}', \theta)$$

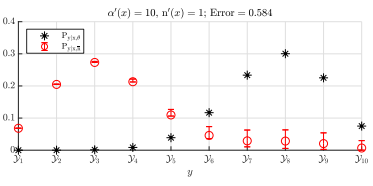
Model Estimation Perspective

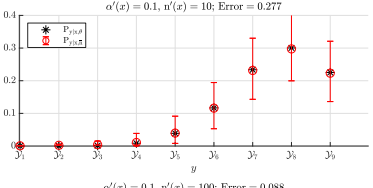
Part II

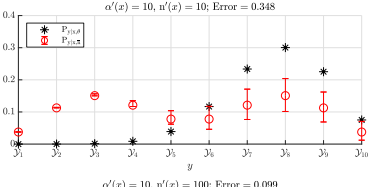
Concentration parameter controls a Bias-Variance trade-off

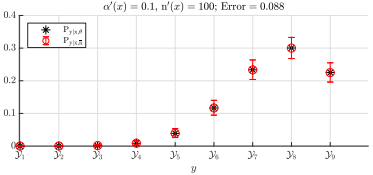


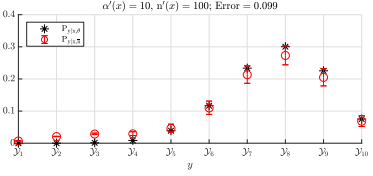












Application to Common Loss Functions

Regularized Empirical Risk

- Recall that the Bayesian decision is $f^*(\mathrm{x}; \bar{\mathrm{n}}) = \arg\min_{h \in \mathcal{H}} \mathrm{E}_{\mathrm{y} \mid \mathrm{x}, \bar{\mathrm{n}}} \left[\mathcal{L}(h, \mathrm{y}) \right]$. The Bayesian conditional risk objective for a Dirichlet prior is:

$$\mathrm{E}_{\mathrm{y} \mid \mathrm{x}, \bar{\mathrm{n}}} \left[\mathcal{L}(h, \mathrm{y}) \right]$$
$$\equiv \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})} \right) \sum_{y \in \mathcal{Y}} \frac{\alpha(y, \mathrm{x})}{\alpha'(\mathrm{x})} \mathcal{L}(h, y) + \left(\frac{\mathrm{n}'(\mathrm{x})}{\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})} \right) \sum_{y \in \mathcal{Y}} \frac{\bar{\mathrm{n}}(y, \mathrm{x})}{\mathrm{n}'(\mathrm{x})} \mathcal{L}(h, y)$$

$$\Downarrow \quad \Downarrow$$

- *Convex combination* of the expected risk with respect to two distributions: the prior conditional mean and the conditional empirical distribution. Prior parameters provide a regularizing term for the empirical loss
- Relative weight $\alpha'(x)/n'(x)$ for regularizing term tends to zero with training data volume, dictating empirical risk minimization

Optimal Estimate: the *Bayes predictive mean* ($\mathcal{H} = \mathbb{R} \supset \mathcal{Y}$)

$$f^*(\mathrm{x}; \bar{\mathrm{n}}) = \mu_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} = \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})}\right) \sum_{y \in \mathcal{Y}} y \frac{\alpha(y, \mathrm{x})}{\alpha'(\mathrm{x})} + \left(\frac{\mathrm{n}'(\mathrm{x})}{\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})}\right) \sum_{y \in \mathcal{Y}} y \frac{\bar{\mathrm{n}}(y, \mathrm{x})}{\mathrm{n}'(\mathrm{x})}$$

* Convex combination of prior estimate $\mu_{\mathrm{y}\,|\,\mathrm{x}} \equiv \sum_{y \in \mathcal{Y}} y \frac{\alpha(y,\mathrm{x})}{\alpha'(\mathrm{x})}$ and empirical mean

---

Minimum Bayes Squared-Error: the expected Bayesian predictive variance

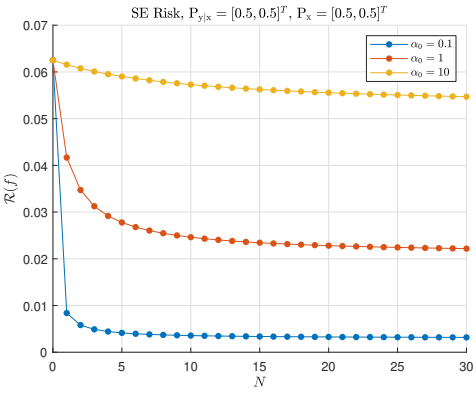
$$\mathcal{R}^* = \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right] = \mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{P}_{\mathrm{x}}(\mathrm{x}) + (\alpha_0 + N)^{-1}}{\mathrm{P}_{\mathrm{x}}(\mathrm{x}) + \alpha_0^{-1}} \Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]$$

* *Note*: $\mathrm{P}_{\mathrm{x}} \equiv \alpha'/\alpha_0$ and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} \equiv \alpha(\cdot, \mathrm{x})/\alpha'(\mathrm{x})$

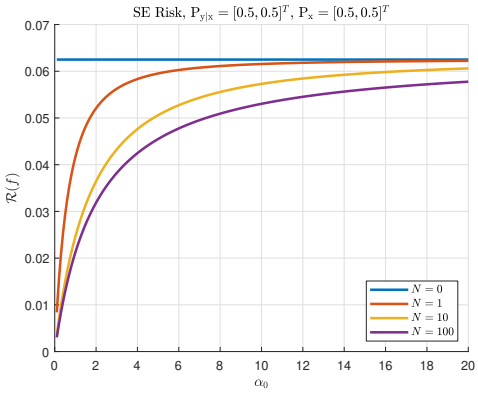
Regression: the Squared-Error Loss

Part II: Minimum Bayes Risk Trends

- Unit interval: $\mathcal{Y} = \{i/M_{\mathrm{y}} : i = 0, \ldots, M_{\mathrm{y}} - 1\}$, $\mathcal{X} = \{i/M_{\mathrm{x}} : i = 0, \ldots, M_{\mathrm{x}} - 1\}$



$$\lim_{N \to \infty} \mathcal{R}^* \equiv \mathrm{E}_{\mathrm{x}} \left[\frac{\mathrm{P}_{\mathrm{x}}(\mathrm{x})}{\mathrm{P}_{\mathrm{x}}(\mathrm{x}) + \alpha_0^{-1}} \Sigma_{\mathrm{y}\,|\,\mathrm{x}} \right]$$



$$\lim_{\alpha_0 \to \infty} \mathcal{R}^* \equiv \mathrm{E}_{\mathrm{x}} \left[\Sigma_{\mathrm{y}\,|\,\mathrm{x}} \right]$$

Regression: the Squared-Error Loss

Part III: Conditional Risk

General squared-error $\mathcal{R}_\Theta(f;\theta)$ is the sum of the clairvoyant risk $\mathcal{R}^*_\Theta(\theta) = \mathrm{E}_{\mathrm{x}\,|\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]$ and the expected squared-bias between the clairvoyant estimate $f_\Theta(\mathrm{x};\theta) = \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}$ and the Bayes estimate $\mu_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$

$$\begin{aligned}\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) &= \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}\,|\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right] \\ &= \mathrm{E}_{\mathrm{x}\,|\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\,\mathrm{E}_{\mathrm{n}'(\mathrm{x})|\theta'(\mathrm{x})}\left[\frac{\mathrm{n}'(\mathrm{x})}{\left(\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})\right)^2}\right]\right] \\ &\quad + \mathrm{E}_{\mathrm{x}\,|\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2 \mathrm{E}_{\mathrm{n}'(\mathrm{x})|\theta'(\mathrm{x})}\left[\left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})}\right)^2\right]\right]\end{aligned}$$

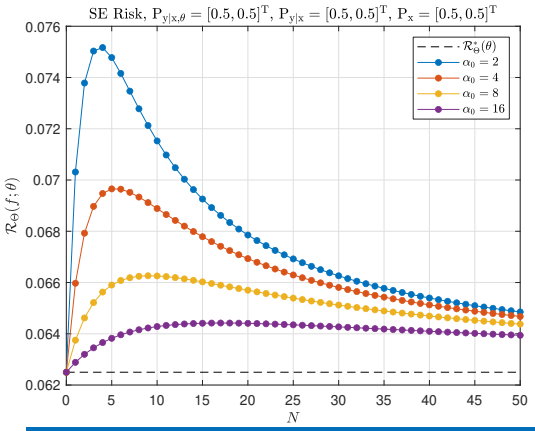
Note: Scaling via expectations of $\mathrm{n}'(x)|\theta'(x) \sim \mathrm{Bi}\left(N,\theta'(x)\right)$

Excess squared-error combines an excess variance term and a squared-bias term for the data-independent estimate $\mu_{\mathrm{y}\,|\,\mathrm{x}}$

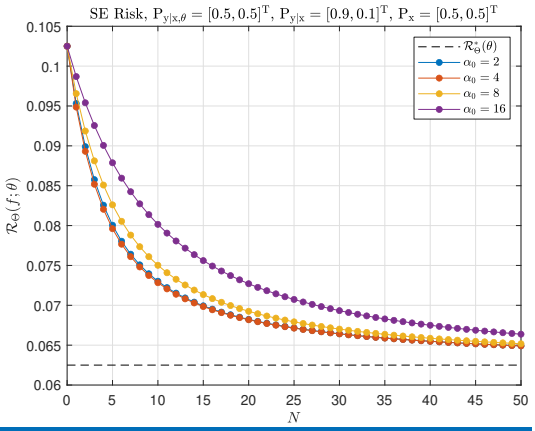
Regression: the Squared-Error Loss

Part IV: Conditional Risk Trends

Unbiased: $\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta} = 0$



Biased: $\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta} \neq 0$



Full prior support guarantees $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to 0$ as $N \to \infty$

Conditional concentration $\alpha'(x)$ controls the Bias-Variance risk trade-off

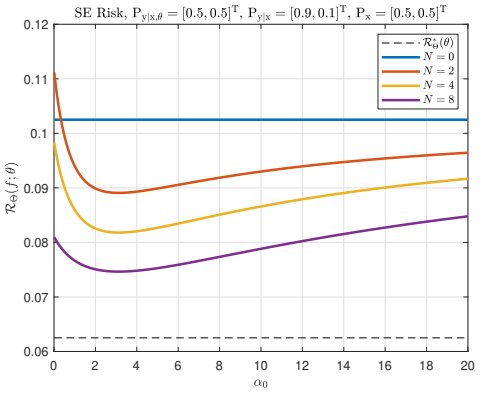
Excess Risk $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta)$

$\underline{\alpha'(x) \to 0}$:

$$\approx \mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\sum_{n=1}^{N}\binom{N}{n}\theta'(\mathrm{x})^n\big(1-\theta'(\mathrm{x})\big)^{N-n}\frac{1}{n}\right]$$

$\underline{\alpha'(x) \to \infty}$:

$$= \mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\big(\mu_{\mathrm{y}\,|\,\mathrm{x}}-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\big)^2\right]$$



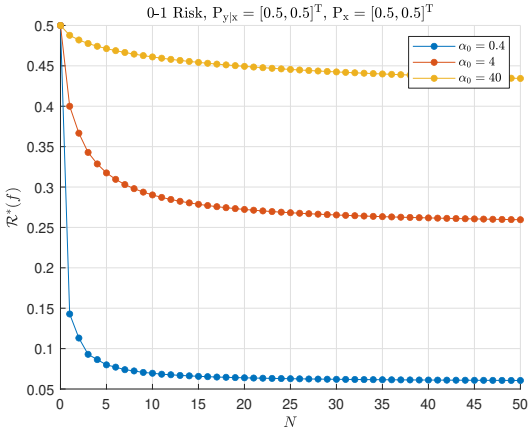
Concentration $\alpha'(x) \equiv \dfrac{\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}}{\big(\mu_{\mathrm{y}\,|\,\mathrm{x}}-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\big)^2}$ minimizes squared-error given $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}$

Optimal Hypothesis:
Weighted conditional majority decision

$$f^*(x;\bar{n}) = \arg\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(y|x,\bar{n}) = \arg\max_{y\in\mathcal{Y}} \left(\alpha(y,x) + \bar{n}(y,x)\right)$$

Minimum Bayes Probability of Error:

$$\mathcal{R}^* = 1 - \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(y|\,\mathrm{x},\bar{\mathrm{n}})\right] = 1 - \sum_{x\in\mathcal{X}} \frac{\mathrm{E}_{\bar{\mathrm{n}}}\left[\max_{y\in\mathcal{Y}}\left(\alpha(y,x) + \bar{\mathrm{n}}(y,x)\right)\right]}{\alpha_0 + N}$$



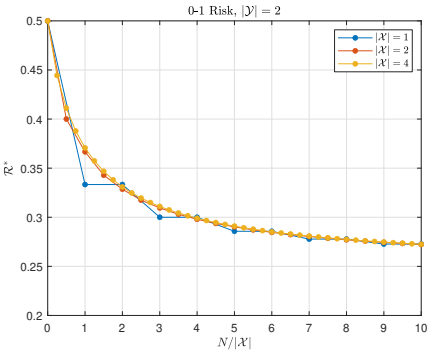
Classification: the 0–1 Loss

Part II: Conditional Majority Decision

With a non-informative ($\alpha = 1$) prior, the majority decision minimizes the empirical risk

$$f^*(x; \bar{n}) = \arg\max_{y \in \mathcal{Y}} \bar{n}(y, x)$$

$$\mathcal{R}^* = 1 - \frac{\sum_{m=1}^{|\mathcal{Y}|} \binom{|\mathcal{Y}|}{m} (-1)^{m-1} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \prod_{l=1}^{|\mathcal{Y}||\mathcal{X}|-1} \left(1 - \frac{mn}{N+l}\right)}{|\mathcal{Y}| + N/|\mathcal{X}|}$$



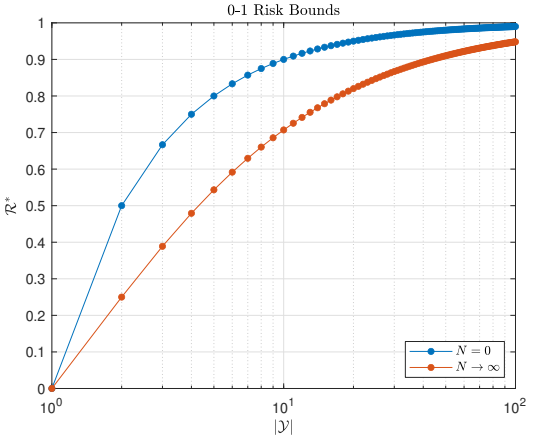
- Computationally efficient formula derived using *Inclusion-Exclusion principle*
- Minimal risk can be approximated by a function dependent only on $N/|\mathcal{X}|$

Classification: the 0–1 Loss

Part II: Conditional Majority Decision

- For binary classification, infinite training data reduces the expected probability of error only from 0.5 to 0.25
- As $|\mathcal{Y}|$ increases, the probability of error tends to unity and any improvement due to training data becomes negligible

N	$\mathcal{R}^*$
0	$1 - \|\mathcal{Y}\|^{-1}$
$\to \infty$	$1 - \|\mathcal{Y}\|^{-1} \sum_{m=1}^{\|\mathcal{Y}\|} m^{-1}$

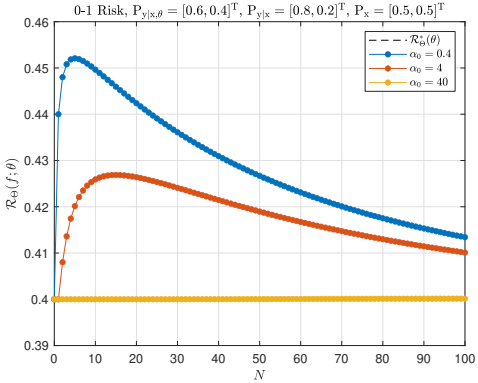


Classification: the 0–1 Loss

Part III: Conditional Risk Trends

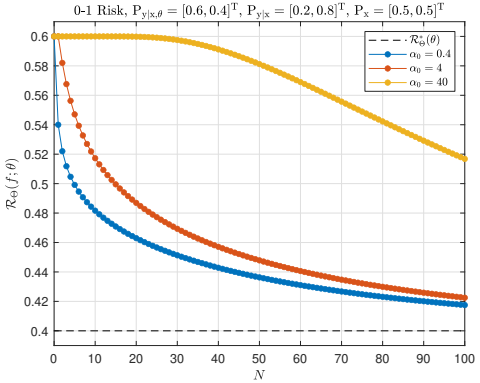
Accurate Prior:

$$\arg\max_{y\in\mathcal{Y}} \alpha(y,x) = \arg\max_{y\in\mathcal{Y}} \tilde{\theta}(y;x)$$



Inaccurate Prior:

$$\arg\max_{y\in\mathcal{Y}} \alpha(y,x) \neq \arg\max_{y\in\mathcal{Y}} \tilde{\theta}(y;x)$$

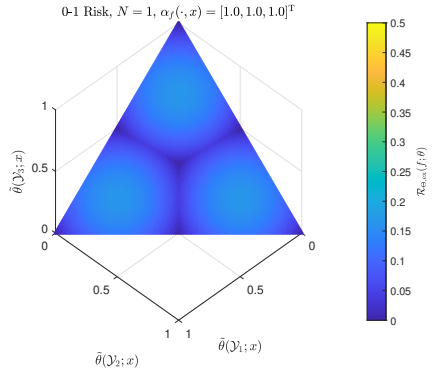


Full prior support guarantees $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to 0$ as $N \to \infty$

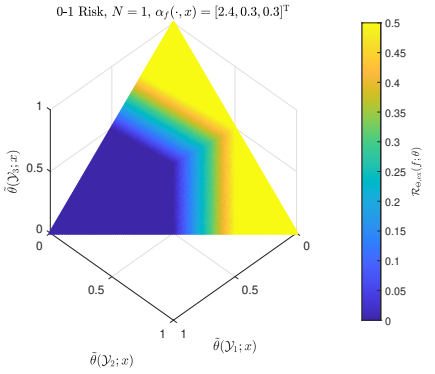
Classification: the 0–1 Loss

Part III: Conditional Risk Trends

Non-informative Prior:



Informative Prior:



Trade-Off

Uniform prior provides a robust classifier for all unknown models θ, limiting the space of high-error models. However, fewer models achieve the clairvoyant risk

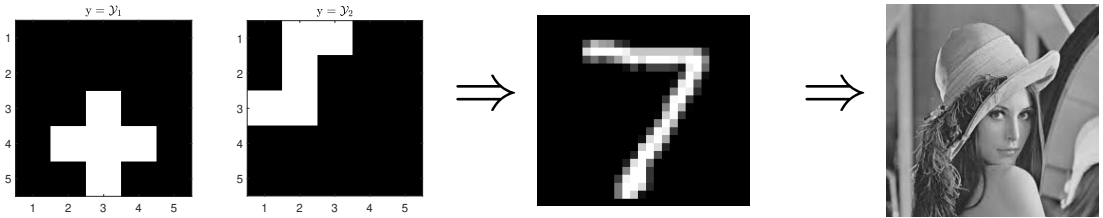
Table of Contents

Learning with Low-Dimensionality Support Priors

- Additional theory needs to be developed for Bayesian learning using priors with limited dimensionality support
  - Quantify risk $\mathcal{R}_\Theta(f;\theta)$ for limited volume N, compare to full-support and non-informative prior results
  - Assess additional risk beyond clairvoyant $\mathcal{R}^*_\Theta(\theta)$ incurred in the limit $N \to \infty$
  - Determine relationship between $\dim(\Theta) < |\mathcal{Y}||\mathcal{X}| - 1$ and computational complexity of implementation. Can data be sufficiently represented after mapping to a set with cardinality $< |\bar{\mathcal{N}}|$?
- Specific research for class of mixture distributions $\theta \equiv \sum_{m=1}^{M} \phi_m h_m$, where $h_m \in \Theta$
  - Hyperparameters ϕ are convex coefficients $\Rightarrow$ characterize with Dirichlet distribution
  - If mixture distributions have disjoint support $h_i(y,x) \cdot h_j(y,x) = 0$, likelihood $\mathrm{P}_{\mathrm{D}|\phi}$ has exponential form

Apply to Human Recognition Tasks

- Bayesian learners will be applied first to simulated data $\mathrm{x} \in \mathcal{X} = \{0,1\}^K$ and subsequently to more complex recognition benchmark data
- Simulated data below (random translations/rotations) demonstrates properties that motivate the use of low-dimensionality support priors
  - Only $9 + 2(9) = 27$ images x are ever observed, compared to the $2^{25} = 33554432$ possible binary images $\Rightarrow$ Prior should be restricted by $\left\|\theta'\right\|_0 \ll |\,\mathcal{X}\,|$
  - Each data sample x is unique to a single class, such that $\left\|\tilde{\theta}(x)\right\|_\infty = 1$



Generalize Theory for Uncountably Infinite Sets

- Initial work assumes finite sets $\mathcal{Y}$ and $\mathcal{X}$, since digital machine learning functions unavoidably use finite data representations. However, numerical data with $|\mathcal{X}| \gg 1$ can often be approximated with real numbers, motivating an analysis for continuous data
- Extending to Euclidean sets (starting with $\mathcal{Y} = \mathcal{X} = \mathbb{R}$), the model θ is treated as a random process
  - Dirichlet process $\theta \sim \mathrm{DP}(\alpha)$ inherits the desirable properties of the PDF[26]
  - Empirical process $\bar{\mathrm{n}} \equiv \sum_{n=1}^{N} \delta(\cdot - \mathrm{D}_n)$ characterized by new concepts: the Multinomial Process and the Dirichlet-Multinomial Process
- Analysis of learning with limited-dimensionality priors will generalize naturally to continuous data sets
  - *Example*: Model θ as a finite mixture of continuous distributions h_m

[26] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Expand Framework for Joint Decisions and Semi-Supervised Learning

- With a Dirichlet prior, the model posterior $\mathrm{p}_{\tilde{\theta}|\,\mathrm{x},\mathrm{D}}$ has no dependency on the novel observation x; using general priors, the datum x refines our statistical understanding of the model, effecting semi-supervised learning
- Generalization to L joint decisions $(h_1, \ldots, h_L) \in \mathcal{H}^L$ based on observations $(\mathrm{x}_1, \ldots, \mathrm{x}_L) \in \mathcal{X}^L$, each corresponding to a distinct unobserved random element $(\mathrm{y}_1, \ldots, \mathrm{y}_L) \in \mathcal{Y}^L$
  - Sample risk comparison with popular non-Bayesian learners using training data D and test data $\big((\mathrm{y}_1, \mathrm{x}_1), \ldots, (\mathrm{y}_L, \mathrm{x}_L)\big)$
  - Dirichlet prior dictates independent decisions due to conditional independence of the model $\tilde{\theta}$ from the novel observations $(\mathrm{x}_1, \ldots, \mathrm{x}_L)$
- Semi-supervised risk trends for $L \to \infty$ will be a primary research focus