Pénalités minimales pour la sélection de modèle

Olivier Sorba

NNT : 2017SACLS043

THÈSE DE DOCTORAT

de

L'UNIVERSITÉ PARIS-SACLAY

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Université Paris-Sud

Laboratoire d'accueil : Laboratoire de mathématiques d'Orsay, UMR 8628 CNRS

Spécialité de doctorat : Mathématiques appliquées

M. Olivier SORBA

Pénalités minimales pour la sélection de modèle
Minimal penalties for model selection

Thèse présentée et soutenue le 9 février 2017 à Orsay

Après avis des rapporteurs :
M. FABRICE GAMBOA (Institut de Mathématiques de Toulouse)
MME PATRICIA REYNAUD-BOURET (Université de Nice Sophia-Antipolis)

Jury de soutenance :

M. GÉRARD BIAU (Université Pierre et Marie Curie)	Examinateur
M. GILLES CELEUX (INRIA-Futurs)	Codirecteur de thèse
M. FABRICE GAMBOA (Institut de Mathématiques de Toulouse)	Rapporteur
MME CÉLINE LÉVY-LEDUC (AgroParisTech)	Présidente
M. PASCAL MASSART (Université Paris-Sud)	Directeur de thèse
MME PATRICIA REYNAUD-BOURET (Université de Nice Sophia-Antipolis)	Rapporteur

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Pénalités minimales pour la sélection de modèle

Mots Clefs : moindres carrés pénalisés, sélection de modèle, pénalité minimale, segmentation de signal gaussien, estimation de densité, contraste pénalisé, détection de ruptures multiples, CART, arbres de régression

Résumé : Dans le cadre de la sélection de modèle par contraste pénalisé, L. Birgé et P. Massart ont prouvé que le phénomène de pénalité minimale se produit pour la sélection libre parmi des variables gaussiennes indépendantes. Nous étendons certains de leurs résultats à la partition d'un signal gaussien lorsque la famille de partitions envisagée est suffisamment riche, notamment dans le cas des arbres de régression. Nous montrons que le même phénomène se produit dans le cadre de l'estimation de densité. La richesse de la famille de modèles est liée à une forme d'isotropie. De ce point de vue le phénomène de pénalité minimale est intrinsèque. Pour corroborer et illustrer ce point de vue, nous montrons que le même phénomène se produit pour une famille de modèles d'orientation aléatoire uniforme.

Title : Minimal penalties for model selection

Keywords : penalized least-squares, model selection, minimal penalties, Gaussian signal segmentation, density estimation, penalized contrast, multiple changepoints detection, CART, regression trees

Abstract : L. Birgé and P. Massart proved that the minimum penalty phenomenon occurs in Gaussian model selection when the model family arises from complete variable selection among independent variables. We extend some of their results to discrete Gaussian signal segmentation when the model family corresponds to a sufficiently rich family of partitions of the signal's support. This is the case of regression trees. We show that the same phenomenon occurs in the context of density estimation. The richness of the model family can be related to a certain form of isotropy. In this respect the minimum penalty phenomenon is intrinsic. To corroborate this point of view, we show that the minimum penalty phenomenon occurs when the models are chosen randomly under an isotropic law.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

Remerciements

Mes remerciements vont tout d'abord à mon directeur de thèse Pascal Massart qui - malgré ses multiples charges dans la communauté mathématique - a bien voulu prendre le risque d'accueillir un élève atypique puis l'aider dans la durée à se frayer un chemin vers une question intéressante.

C'est un immense privilège scientifique et humain de partager le statut d'élève de Pascal.

Le soutien, l'attention, les conseils avisés et la grande expérience de Gilles Celeux, co-directeur, m'ont également été précieux.

Je dois une reconnaissance particulière aux rapporteurs de cette thèse pour leur lecture attentive et leurs observations, sources de réelles améliorations : Fabrice Gamboa et Patricia Reynaud-Bouret dont les questions ont de plus suscité plusieurs approfondissements.

Céline Lévy-Leduc – comme présidente - et Gérard Biau m'ont également fait l'honneur de siéger au jury de soutenance. Jury dont chaque membre m'a accordé une écoute aussi aiguisée que bienveillante. Cette écoute m'a été précieuse tout comme les suggestions de directions de recherche ultérieures.

J'ai une très grande dette envers François, mieux connu comme Professeur François Labourie, qui m'a le premier suggéré l'idée de revenir sérieusement vers la recherche, et qui m'a inlassablement fait partager sa connaissance scientifique et humaine de la communauté mathématique.

Ce travail m'a permis de rencontrer certains membres du Laboratoire de Mathématiques d'Orsay, que je remercie de leur accueil et de leur amitié. Je pense en particulier à Elisabeth Gassiat et Pierre Pansu. Je voudrais aussi remercier Valérie Lavigne qui m'a facilité de nombreuses démarches administratives, suivie en cela par Florence Rey.

J'ai pu aussi retrouver sur un autre plan de nombreux amis comme Indira Chatterji, Claude Viterbo, Carl Graham ou Pierre Bertrand qui, scientifiques chevronnés, m'ont cependant fait l'honneur de s'intéresser à mes travaux.

Cette thèse n'aurait pas vu le jour sans le soutien du groupe Lagardère avec en particulier l'appui de Thierry Funck-Brentano et l'amitié de mes collègues : parmi eux les membres du réseau innovation Julien Durand, Rémi Vion et Edouard Minc, Pierre Sellier, Anne Solente, Norbert Giaoui, Isabelle Juppé et Sophie Reille (Sophie merci d'avoir longtemps supporté ma distraction !).

Marianne Sorba puis Christophe Geissler et l'équipe d'Advestis – Noureddine Boumlaik, Vincent Margot, Raphaël Minato - ont bien voulu se prêter à des répétitions de soutenance chacune suivie d'une discussion soutenue. C'est aussi Christophe qui a pris le temps de vérifier certains calculs.

De nombreux amis et proches - pardon à tous ceux que j'omets - m'ont toujours accompagné, soutenu et encouragé de bien des manières. Je pense en particulier à André Vig, Dominique D'Hinnin, Christophe Geissler, Bruno Jouhier, François Rémion, Martine Magnan, Jean-Christophe Pettinotti, Geneviève Almouzni («*less is more !*») et Sylvie Laure.

Je pense également à mes parents, qui m'ont donné très tôt la curiosité des choses de la nature et le goût de la science.

Les membres des très secrets cercles de congratulation mutuelle se reconnaîtront : en particulier Pierre, François, Daniel, Marie-Jeanne, Sophie, Florence, Clément, Marie, Vincent, Paul, Grégoire, Séverine, Samuel, Marie-Hélène, Françoise, Mathieu, Julien, Michel, Valérian, Daniel, Christine, Véronique, Amandine et Felicity.

Je pense enfin avec reconnaissance à Claire et nos filles Camille, Laure, Clémentine et Marianne pour tout ce que signifient leur soutien et leur présence joyeuse.

Contents

I Introduction 9

1 Objet de ce travail 10

1.1 Les principes de la sélection de modèle 10

1.2 Estimation par minimum de contraste 14

1.3 Sélection de modèle par pénalisation 15

1.4 Sélection de modèle non asymptotique par pénalisation 17

1.5 Pénalités minimales 18

2 État de l'art 21

2.1 Difficultés propres à la complexité exponentielle 22

2.2 Segmentations et partitions 23

3 Principaux résultats 27

4 Main results 29

II Minimal penalties 30

5 Technical Preliminaries for Gaussian model selection 33

5.1 Gaussian model selection 33

5.2 Birgé-Massart's Gaussian model selection theorem 34

5.3 Variable selection, a minimal penalty theorem of L.Birgé and P. Massart 36

6 A motivating example: estimating a continuous signal under white noise with step-functions by recursive partitioning 39

6.1 Estimation procedure 39

6.2 Assessing the complexity of the model family 40

6.3 Sufficient penalties 41

6.4 A geometrical property 42

6.5 Minimal penalties 44

7 Technical preliminaries on signal segmentation and partitioning 45

7.1 Partitions . . . 45
7.2 Recursive partitioning . . . 47

8 Minimal penalties for the partitioning of a discrete Gaussian signal 48
8.1 Definitions and assumptions on the model family structure . . . 49
8.2 Sufficient penalties . . . 53
8.3 Minimal penalties . . . 53
8.4 Numerical experiments for section 8 . . . 57
8.5 Numerical experiments for section 8: results . . . 62

9 A first extension to histogram selection for density estimation 122
9.1 Technical preleminaries: model selection for density estimation with histograms . . . 122
9.2 Definitions and main assumptions . . . 127
9.3 Sufficient penalties under the null hypothesis . . . 131
9.4 A risk and dimension lower bound . . . 132

10 Remark on the isotropy of the model family 135
10.1 Toy problem . . . 135

11 Proofs for Sections 5, 6 and 8 138
11.1 Proof of Corollary 5.7 . . . 138
11.2 Proof of Proposition 6.1 . . . 139
11.3 Proof of Lemma 6.2 . . . 142
11.4 Proof of proposition 6.4 . . . 143
11.5 Proof of Proposition 8.5 . . . 147
11.6 Proof of Proposition 8.6 . . . 150
11.7 Proof of Proposition 8.7 . . . 151
11.8 Proof of Proposition 8.8 . . . 156

12 Proofs for section: 9 A first extension to histogram selection ... 159
12.1 Proof of Proposition 9.13 . . . 159
12.2 Proof of Proposition 9.14 . . . 160

13 Proofs for section: 10 Remark on the isotropy of the model family 170

13.1 Proof of Proposition 10.1 170

III Deviation Inequalities 173

14 Tail lower bound inequalities 174

14.1 Tail functions 174
14.2 Some known concentration inequalities 176
14.3 Tail lower bound for a binomial distribution 178
14.4 Inequalities based on the properties of the incomplete beta function . . . 178
14.5 Inequalities based on the properties of the incomplete gamma function . 181

15 Tools for sorted chi-square samples 183

15.1 Order statistics 183
15.2 Lower bound on the partial sum of an ordered chi-square sample 183
15.3 Upper bound on the partial sum of an ordered chi-square sample 185

16 Proofs for Section 14 Tail lower bound inequalities 186

16.1 Proof of Lemma 14.2 186
16.2 Proof of Lemma 14.4 186
16.3 Proof of Lemma 14.6 187
16.4 Proof of Lemma 14.10 188
16.5 Proof of Lemma 14.11 189
16.6 Proof of Lemma 14.13 190
16.7 Proof of Lemma 14.14 192
16.8 Proof of Corollary 14.15 194
16.9 Proof of Lemma 14.18 195
16.10Proof of Corollary 14.19 196
16.11Proof of Lemma 14.20 199
16.12Proof of Lemma 14.21 200

17 Proofs for Section 15 Tools for sorted chi-square samples 201

17.1 Proof of Lemma 15.3 201
17.2 Proof of Lemma 15.4 203

IV Appendix 204

A Assessing the complexity of the set of hyperrectangle partitions in a discrete hyperrectangle 205

B Complexity of recursive segmentation families 208

B.1 Introduction 208

B.2 Multiple segmentation of an integer segment 208

B.3 Regression trees 210

B.4 Enumerating the models 211

References 226

Figures and tables 230

Index 231

Première partie

Introduction

Sommaire

1 Objet de ce travail **10**

1.1 Les principes de la sélection de modèle 10

1.2 Estimation par minimum de contraste 14

1.3 Sélection de modèle par pénalisation 15

1.4 Sélection de modèle non asymptotique par pénalisation 17

1.5 Pénalités minimales 18

2 État de l'art **21**

2.1 Difficultés propres à la complexité exponentielle 22

2.2 Segmentations et partitions 23

3 Principaux résultats **27**

4 Main results **29**

1 Objet de ce travail

1.1 Les principes de la sélection de modèle

Une personne intéressée peut se familiariser de façon simple avec les principes de la sélection de modèle en essayant informellement d'améliorer la qualité d'une image bruitée, que ce soit avec les fonctions de traitement d'image d'un téléphone portable, avec un logiciel librement disponible, ou simplement en fermant progressivement ses paupières en regardant l'image. La table 1 offre un exemple, où l'on constate que lisser progressivement l'image bruitée a l'effet désirable d'éroder ce que l'œuil humain perçoit comme le bruit, mais aussi l'effet indésirable d'éroder les détails de l'image originale. Bien que ce soit une question de préférence, beaucoup d'observateurs choisiraient de retenir un certain degré de lissage. Dans cette combinaison particulière d'image originale et de bruit, ceci est dû au fait que le bruit - contrairement à l'image originale - n'a pas de structure de longue portée et répond bien au lissage, qui est une opération de courte portée.

Une observation clef est que l'image sélectionnée n'est pas la plus contrastée, mais celle qui - dans le jugement de l'observateur - se trouve au point où l'opération de lissage commence à éroder les détails de l'image originale autant qu'elle diminue le bruit. En d'autres mots, au point où la perte incrémentale en biais (erreur de structure) commence à l'emporter sur le gain incrémental en bruit (erreur de variance)

La sélection de modèle formalise cette intuition. En pratique une image rectangulaire est enregistrée comme un vecteur de valeurs d'exposition de dimension égale au nombre de pixels. Chaque intensité de lissage est considérée comme un modèle statistique, autrement dit comme un moyen de projeter ou de réduire l'image en un vecteur de moindre dimension effective. Informellement, lisser une image réduit le nombre de paramètres nécessaires pour la décrire, comme dans l'expérience courante d'en réduire la définition. En général, la capacité d'un modèle à capturer indûment le bruit varie directement avec sa dimension effective, si bien qu'un bon modèle doit être parcimonieux en paramètres.

La première étape de la sélection de modèle est de choisir une mesure de contraste pour quantifier la distance entre deux images de mêmes dimensions. Ce peut être par exemple une moyenne des distances entre les valeurs d'exposition des pixels de même position. L'objectif quantifié de l'expérimentateur est d'obtenir le plus petit contraste possible entre l'image lissée et l'image originale. Comme cette dernière est inconnue, on peut la remplacer par l'image bruitée, et quantifier les distances entre l'image bruitée et chacune de ses versions lissées. Malheureusement, le bruit contribue à cette mesure de contraste dite 'empirique', surtout pour les modèles de grande dimension, et cette procédure seule ne permet pas de conclure.

Pour contourner cette difficulté, l'expérimentateur peut utiliser sa connaissance statistique du bruit pour corriger le contraste empirique d'une façon appropriée et pénaliser ainsi les modèles de trop grande dimension. Finalement, le modèle sélectionné sera le minimiseur du contraste empirique pénalisé, qui a donné son nom à cette méthode. La section 8.4 offre plusieurs exemples de courbe de contraste empirique pénalisé en fonction de la dimension du modèle.

Une observation importante est que certaines régions de l'image originale contiennent

moins de détails et que l'expérimentateur pourrait chercher à concentrer le lissage sur elles sans perdre beaucoup de la structure d'origine. La table 2 offre un exemple du résultat d'une telle opération avec la méthode dite de seuillage d'ondelettes. Avec cette flexibilité accrue, l'expérimentateur espère améliorer la qualité générale du résultat final, mais en même temps, il accroît considérablement le nombre de modèles à envisager pour chaque dimension, que l'on nomme la complexité de la famille de modèles considérés. Là non plus on ne connaît pas l'image d'origine, et ce processus adaptatif doit être basé sur l'observation de l'image bruitée, avec un risque accru de s'adapter au bruit plutôt qu'à la structure de l'image originale. En témoignent les nombreux pixels blancs qui restent sur les images de la Table 2, en contrepartie d'une meilleure fidélité à l'image originale. C'est pourquoi la complexité de la famille de modèles envisagés doit être prise en compte pour concevoir un terme de pénalité approprié.

Cette conception d'un terme de pénalité revient à incorporer dans la procédure une certaine hiérarchie préalable au sein de la famille de modèles, ce qui est très proche, en théorie comme en pratique, des méthodes bayésiennes, où cette hiérarchie préalable s'exprime sous la forme d'une loi de probabilité *a priori* sur la famille de modèles envisagés. Les deux méthodes offrent un cadre formalisé pour arbitrer entre d'une part qualité d'ajustement et d'autre part simplicité.

La sélection de modèle par contraste pénalisé ne se limite pas au traitement d'image, mais constitue au contraire un cadre très général qui trouve son application dans de nombreux domaines comme par exemple la génomique, les bio-statistiques, la classification ou le traitement du signal. Deux questions importantes de ce domaine sont de déterminer quelles pénalités sont suffisantes pour garantir un modèle de qualité acceptable, et inversement quel est le niveau minimal de pénalité en deçà duquel le modèle sélectionné adopte un comportement erratique, en général de pair avec une dimension excessive.

L'importance pratique des pénalités minimales provient du fait qu'en général l'intensité du bruit n'est pas connue à l'avance. Dans ce cas, même si l'on connaît la forme générale d'un terme de pénalité acceptable, il manque un facteur général de calibration. Une solution empirique consiste alors à observer quel est le niveau minimal de ce terme de calibration à partir duquel la procédure sélectionne un modèle raisonnable, c'est-à-dire de dimension limitée. Dans certaines situations, on sait empiriquement ou par une preuve théorique que le niveau optimal de pénalité est le double du seuil minimal. En réinjectant ce seuil, la procédure peut donc se mener entièrement sur la base des seules observations.

Cependant, la validité de cette heuristique pour des familles comportant un très grand nombre de modèles reste largement une question ouverte. Un résultat de L. Birgé et P. Massart apporte une réponse positive dans le cas dit de la sélection libre parmi une famille de variables gaussiennes indépendantes. Le but de ce travail est d'étendre ce résultat à des situations plus générales souvent rencontrées en pratique, comme par exemple les arbres de régression.

TABLE 1 – A first experiment in model selection, averaging

Domenico Scarlatti
in 1738 by
Domingo Antonio Velasco,
1267 × 1600 pixels

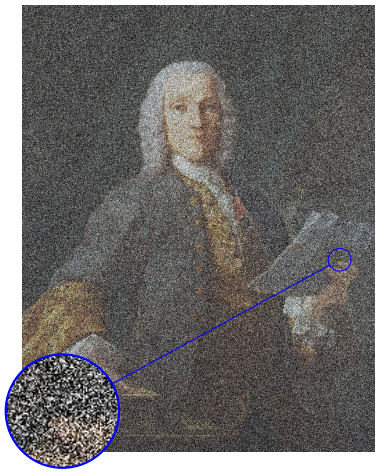
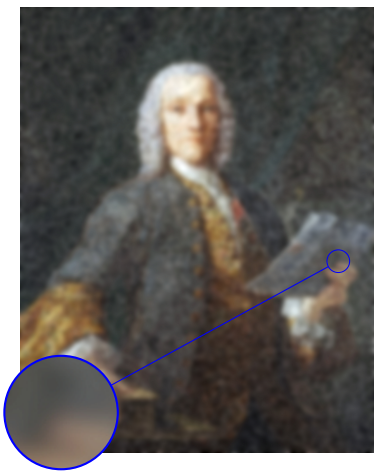
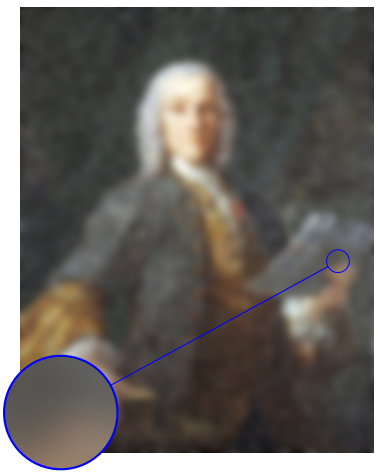
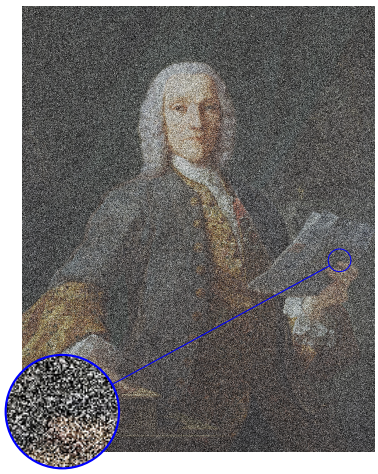
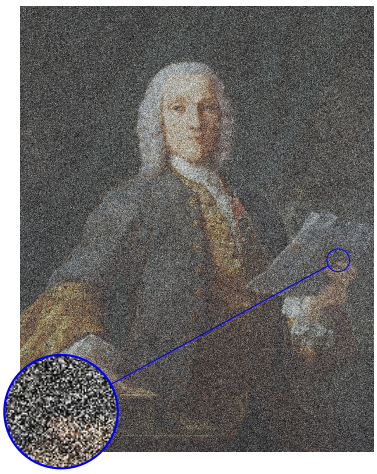
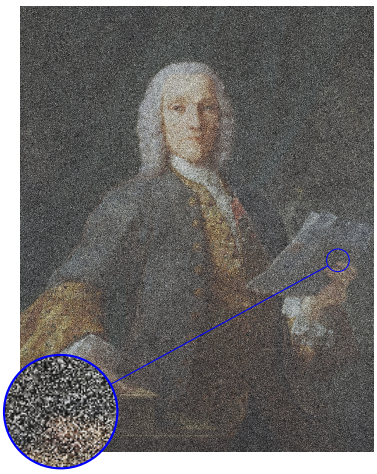
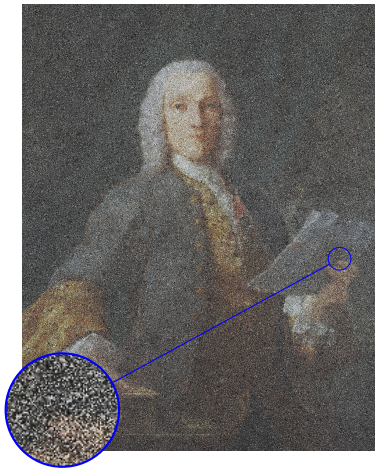
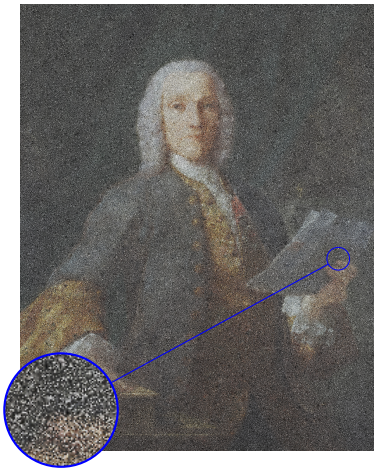
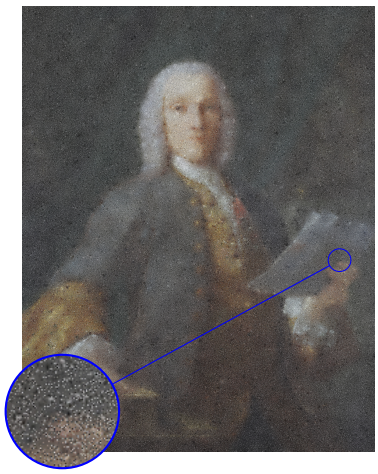
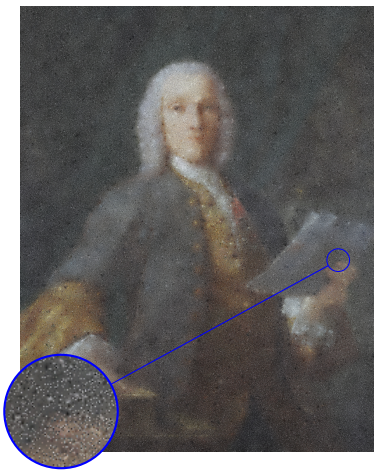
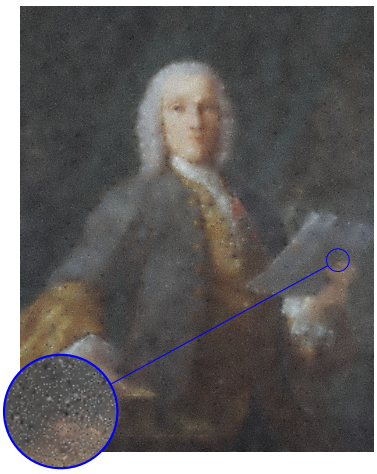
Noisy image	average radius ∼ 5 pixels	average radius ∼ 10 pixels
		
average radius ∼ 20 pixels	average radius ∼ 30 pixels	average radius ∼ 50 pixels
		

TABLE 2 – A first experiment in model selection, wavelet thresholding

Noisy image	Threshold 1	Threshold 2
		
Threshold 3	**Threshold 4**	**Threshold 5**
		
Threshold 8	**Threshold 9**	**Threshold 10**
		

1.2 Estimation par minimum de contraste

Dans cette section, nous décrivons le cadre de la sélection de modèle par contraste pénalisé tel que le proposent L. Birgé et P. Massart dans [Mas07], [BM01] and [BM07].

En sélection de modèle, on observe typiquement une variable aléatoire $\xi^{(n)}$ dont la distribution dépend d'une quantité inconnue $s \in \mathcal{S}$ à estimer. Dans le cadre de la régression linéaire il peut s'agir d'un échantillon de variables explicatives et de réponse (expliquées), la quantité s est la fonction de régression, c'est-à-dire l'espérance de la variable de réponse sachant les valeurs des variables explicatives, et l'ensemble $\mathcal{S}$ est $\mathbb{L}_2(\mu)$ où μ désigne la distribution des variables explicatives.

Pour la régression à effet fixe comme dans l'exemple de traitement d'une image en niveau de gris ci-dessus, on observe $Y_i = s(x_i) + \epsilon_i, 1 \leq i \leq n$ et $\mathcal{S}$ est simplement $\mathbb{R}^n$ où n est dans notre exemple la taille de l'image en pixels.

En estimation de densité s est une densité et $\mathcal{S}$ peut être choisi comme l'ensemble de toutes les densités de probabilité par rapport à une mesure de probabilité donnée μ. On considère un contraste empirique $\gamma_n(\cdot)$, basé sur les observations et tel que la fonction $t \mapsto \mathbb{E}[\gamma_n(t)]$ atteint un minimum à $t = s$. Alors la fonction de perte associée $l(s,t) = \mathbb{E}[\gamma_n(t)] - \mathbb{E}[\gamma_n(s)]$ a la propriété désirable d'être non-négative pour tout $t \in \mathcal{S}$.

Ce qu'on appelle un modèle dans ce contexte est un sous-ensemble donné de S de $\mathcal{S}$, et un estimateur par minimum de contraste est un minimiseur $\hat{s}$ du contraste empirique $\gamma_n(\cdot)$ sur S. L'idée est bien-entendu de substituer $\gamma_n(\cdot)$ à son espérance, qui est inconnue.

Avec le choix d'une fonction de perte γ adéquate, on peut en particulier définir un contraste empirique $\gamma_n(t)$ proportionnel à $\sum_{i=1}^n \gamma(t, \xi_i)$, entrant ainsi dans le cadre de la M-estimation. Des choix courants sont :

- pour la régression, le critère des moindres carrés, avec
  - $\gamma(t,(x,y)) = -2y\,t(x) + t^2(x)$,
  - $\gamma_n(t) = \frac{1}{n}\sum_{i=1}^n \gamma(t, \xi_i)$,
  - auquel cas la fonction de perte est donnée par $l(s,t) = \frac{1}{n}\|s-t\|^2$ où $\|\cdot\|$ désigne la norme quadratique dans $\mathbb{L}_2(\mu)$ ou $\mathbb{R}^n$.
- pour l'estimation de densité le critère de maximum de vraisemblance avec
  - $\gamma(t,\xi) = -\log(t(\xi))$,
  - $\gamma_n(t) = \frac{1}{n}\sum_{i=1}^n \gamma(t, \xi_i)$
  - et la fonction de perte correspondante est donnée par le nombre d'information de Kullback-Leibler $l(s,t) = \mathbf{K}(s,t) = \int s \log\left(\frac{s}{t}\right)$

Le problème central de l'estimation par minimum de contraste est de choisir un modèle S adéquat. Pour illustrer la principale difficulté, considérons une régression linéaire sous un bruit gaussien de loi $\mathcal{N}(0, \epsilon^2 \mathbb{I}_n)$, and choisissons pour modèle S un sous-espace D-dimensionnel de $\mathcal{S}$. Un calcul rapide montre que

$$\mathbb{E}\left[\|s - \hat{s}\|^2\right] = d^2(s,S) + \epsilon^2 D$$

et

$$\mathbb{E}\left[\gamma_n\left(\hat{s}\right)\right] = -\frac{\left\|\bar{s}\right\|^2}{n} - \epsilon^2\frac{D}{n}$$

où $\bar{s}$ est la projection de s sur le sous-espace D-dimensionnel S et $d(\cdot,\cdot)$ la distance euclidienne standard de $\mathbb{R}^n$. Le choix d'un modèle de faible dimension D garantit donc que le terme de variance $\epsilon^2 D$ reste sous contrôle, mais cela peut être en contrepartie d'un biais $d(s,S)$ très important. Inversement, choisir un modèle S très grand peut conduire à un mauvais estimateur même si le paramètre s appartient au modèle S, par le fait que le terme de variance $\epsilon^2 D$ peut être très grand.

En pratique, on considérant une collection dénombrable (et habituellement finie) de modèles , $(S_m)_{m\in\mathcal{M}}$ on doit choisir choisir le 'meilleur' estimateur au sein de la collection $(\hat{s}_m)_{m\in\mathcal{M}}$.

Idéalement on aimerait retenir le minimiseur du risque $\mathbb{E}\left[l(s,\hat{s}_m)\right]$ pour $m \in \mathcal{M}$, que l'on appelle l'oracle, mais ce modèle dépend du paramètre inconnu s. Il vient naturellement à l'esprit de substituer à s le signal observé Y, et de construire un critère basé sur les données en corrigeant le contraste empirique ainsi obtenu par une compensation adéquate de son terme de variance. L'oracle, inconnu, ne peut servir que de référence pour évaluer la performance d'une procédure de sélection de modèle basée sur les données. Il est utile de noter que souvent, même si le vrai paramètre s appartient à l'un des modèles candidats, l'oracle peut être différent de s et en général de moindre dimension.

1.3 Sélection de modèle par pénalisation

Une procédure de sélection de modèle par pénalisation consiste, à partir d'une fonction de pénalité appropriée $\text{pen} : \mathcal{M} \to \mathbb{R}$ à choisir comme modèle $\hat{m}$ et comme estimateur $\hat{s}_{\hat{m}}$ un minimiseur du critère de contraste pénalisé

$$\gamma_n\left(\hat{s}_m\right) + \text{pen}\left(m\right).$$

Les premiers exemples de critères pénalisé ont été proposés au début des années 1970 par Akaike and Mallows, sous forme de fonctions proportionnelles au nombre de paramètre D_m du modèle correspondant S_m.

- Aikaike : $\frac{D_m}{n}$ pour l'estimation de densité [Aka73],
- Mallows's C_p : $2\frac{D_m}{n}\epsilon^2$ pour la régression par moindres carrés [Mal73].

Le C_p de Mallows, tout comme son pendant le critère d'Akaike reposent sur l'heuristique suivante, dite de l'estimation de risque sans biais. Supposons que l'on observe $Y = s + \epsilon W$ où s est un élément de $\mathbb{R}^n$ et W est un vecteur gaussien standard $\mathcal{N}(0,\mathbb{I}_n)$. A chaque modèle linéaire S_m, $m \in \mathcal{M}$ correspond l'estimateur des moindres carrés $\hat{s}_m = \Pi_m(s+\epsilon W)$, où Π_m est le projecteur orthogonal sur S_m.

Le meilleur choix de modèle serait :

$$\hat{m} \in \underset{m\in\mathcal{M}}{\text{argmin}}\, n^{-1}\|\hat{s}_m - s\|^2.$$

Cependant ce critère ne peut pas être utilisé puisqu'il dépend du signal inconnu s. On a donc recours à un critère empirique de la forme :

$$\text{crit}(m) \simeq n^{-1}\|\hat{s}_m - Y\|^2 + \text{pen}\,(m)\,.$$

La forme idéale de la pénalité $\text{pen}\,(m)$ serait

$$\text{pen}_{id}\,(m) = n^{-1}\|\hat{s}_m - s\|^2 - n^{-1}\|\hat{s}_m - Y\|^2,$$

mais de nouveau celle-ci dépend de s et ne peut pas être connue. L'heuristique de Mallows consiste à la remplacer par son estimateur sans biais

$$\mathbb{E}\left[n^{-1}\|\hat{s}_m - s\|^2 - n^{-1}\|\hat{s}_m - Y\|^2\right] = -\epsilon^2\left(1 - 2\frac{\text{Tr}\,[\Pi_m]}{n}\right),$$

ou plus précisément par sa composante variable en fonction du modèle m, à savoir $2\epsilon^2\frac{\text{Tr}[\Pi_m]}{n} = 2\epsilon^2\frac{D_m}{n}$ où D_m est la dimension du modèle m.

La proposition de Mallows se place donc clairement du point de vue de l'optimalité, puisqu'elle vise à minimiser le risque quadratique de l'estimateur sélectionné, et non à maximiser la probabilité de sélectionner un éventuel modèle contenant s, ce qui serait le point de vue de l'efficacité. A l'inverse, une approche visant l'efficacité chercherait non à minimiser le risque mais à maximiser les chances d'identifier correctement le paramètre s au vu des données.

Cette autre approche a donné naissance au *critère d'information bayésien*, ou *BIC*. Dans un contexte gaussien, la pénalité correspondante est de la forme $\epsilon^2 D_m \log(n)$, et le contraste pénalisé correspondant de la forme

$$BIC(m) \simeq n^{-1}\|\hat{s}_m - Y\|^2 + n^{-1}D_m \log(n).$$

Historiquement le critère d'information bayésien est proposé par Gideon Schwarz dans son article [Sch78], et suscite depuis une littérature fournie et de nombreuses applications pratiques. Dans un contexte bayésien, on suppose une loi *a priori* non-informative (uniforme) sur une famille finie $\mathcal{M}$ de modèles en général emboités, dont cherche à retenir le plus vraisemblable au vu des données. La forme $D_m \log(n)$ de la pénalité résulte d'une approximation de la vraisemblance d'un modèle m, en n grand. Cette méthode fournit pour chaque modèle m, outre le critère de contraste pénalisé $BIC(m)$, une probabilité *a posterori*

$$p_m = \frac{e^{-\frac{1}{2}BIC(m)}}{\sum_{m'\in\mathcal{M}} e^{-\frac{1}{2}BIC(m')}}$$

qui, sous certaines conditions, en n grand tend vers 1 pour le "vrai modèle", c'est-à-dire - lorsqu'il existe - le plus petit modèle de la famille $\mathcal{M}$ qui contienne le paramètre à estimer s. Lorsque ce dernier n'appartient à aucun modèle de la famille $\mathcal{M}$, les mêmes considérations s'appliquent au "quasi-vrai" modèle, c'est-à-dire l'élément de $\mathcal{M}$ qui en est le plus proche en un sens à définir [LMH04]. C'est en ce sens que le critère d'information bayésien est considéré comme consistant pour la dimension. Les auteurs de [LMH04] détaillent cet aspect et discutent également de façon détaillée en quoi le critère d'information bayésien et le critère d'Akaike diffèrent par leur nature et par leur intention.

Dans l'article [Yan05], Y.Yang montre que les qualités du critère d'information bayésien d'une part et celles du critère d'Akaike d'autre part sont incompatibles, en ce sens que si le signal s est contenu dans un des modèles candidats, et que la procédure est efficace et trouve de ce fait asymptotiquement ce modèle avec une forte probabilité, cette même procédure ne peut qu'être sous-optimale en terme d'erreur quadratique moyenne. Il appartient donc au praticien de choisir quel critère convient le mieux à l'objectif de l'estimation : efficacité et identification d'une part ou optimalité et risque d'autre part.

Ce travail se place du point de vue de l'optimalité, avec pour objectif le risque quadratique de l'estimateur sélectionné. Dans ce cadre, on juge souvent de la performance d'une procédure non par comparaison avec la performance d'un éventuel vrai modèle contenant s, mais par rapport à celle d'un oracle, c'est-à-dire d'un modèle qui parmi la famille de candidats minimise le risque quadratique moyen. Il n'est donc pas nécessaire de faire l'hypothèse qu'un 'vrai' modèle contenant le paramètre s existe dans la famille $\mathcal{M}$. Le but des procédures de sélection de modèle est alors d'obtenir des inégalités dites "d'oracle" vraies avec une forte probabilité et de la forme :

$$n^{-1}\|\hat{s}_{\hat{m}} - s\|^2 \leq C_n \inf_{m\in\mathcal{M}} \left\{ n^{-1}\|\hat{s}_m - s\|^2 \right\} + R_n$$

où la constante C_n s'approche autant que possible de 1, et où le terme R_n est négligeable devant le risque de l'oracle.

Les propositions de Mallows et Akaike consistent en substance à considérer qu'avec un grand nombre d'observations l'espérance du minimum du contraste pénalisé

$$\mathbb{E}\left[\inf_{m\in\mathcal{M}} \left\{ n^{-1}\|\hat{s}_m - s\|^2 + \mathrm{pen}\,(m) \right\} \right]$$

peut être remplacée sans dommage par le minimum des espérances

$$\inf_{m\in\mathcal{M}} \left\{ \mathbb{E}\left[n^{-1}\|\hat{s}_m - s\|^2 \right] + \mathrm{pen}\,(m) \right\}.$$

Ces propositions conviennent dans un point de vue asymptotique et leur preuves reposent fortement sur les hypothèses que la dimension des modèles considérés reste bornée quand le nombre d'observations n tend vers l'infini, de même que le nombre de ces modèles.

Malheureusement ce point de vue perd sa validité quand ces hypothèses ne sont plus satisfaites. C'est par exemple le cas avec la détection de ruptures multiples où l'on tente de reconstituer un signal séquentiel par des fonctions en escaliers, ou des histogrammes. Dans ce cas un modèle de dimension d correspond à $d-1$ points de rupture, et le nombre de tels modèles est $\binom{n-1}{d-1}$, qui croît de façon polynomiale avec n.

Sur la base de ces observations, les auteurs de [BM97] and [BBM99] ont développé une approche non-asymptotique pour la sélection de modèle par contraste pénalisé.

1.4 Sélection de modèle non asymptotique par pénalisation

Dans le cadre de la sélection de modèle non asymptotique, il est possible de faire varier le nombre et la dimension des modèles envisagés avec le nombre d'observations, et on choisit souvent la famille de modèles en fonction de ses propriétés d'approximation,

comme par exemple des familles fonctions en escalier pour la segmentation de signal, ou bien les espaces d'ondelettes en traitement d'image.

La complexité de la famille de modèles candidats est prise en compte via un ensemble de poids $\{L_m\}_{m\in\mathcal{M}}$ avec la restriction

$$\sum_{m\in\mathcal{M},D_m>0} e^{-L_m D_m} \leq 1,$$

ou dans certains cas simplement

$$\sum_{m\in\mathcal{M}} e^{-L_m D_m} < \infty.$$

Des résultats comme ceux de L. Birgé et P. Massart (Théorème 5.6) pour les moindres carrés ou de G. Castellan (Théorème 9.7) pour l'estimation de densité garantissent la sélection d'un modèle en un certain sens raisonnable avec des pénalités de la forme

$$\left(c_1 + c_2\sqrt{L_m} + c_3 L_m\right)\frac{D_m}{n},$$

où les constantes c_1, c_2 and c_3 ne dépendent pas de n.

Techniquement ces résultats reposent sur des inégalités de concentration pour contrôler uniformément les fluctuations d'un contraste $\gamma_n(t)$ autour de son espérance pour $t \in S_m, m \in \mathcal{M}$. Bien que ce travail fasse appel à certaines inégalités de concentration, nous n'essayons pas de discuter ici ce domaine très large. Des exposés approfondis figurent dans [BLM13] ou [Mas07].

Notons simplement, pour préparer la discussion du rôle de la complexité dans les pénalité minimales, que le contrôle uniforme des variations du contraste empirique pour $m \in \mathcal{M}$ s'opère en quelque sorte au prix du remplacement des seuils $\{D_m\}_{m\in\mathcal{M}}$ par des seuils proches de $\left\{D_m\left(1+2\sqrt{L_m}+2L_m\right)\right\}_{m\in\mathcal{M}}$.

En particulier, au vu des hypothèses des deux théorèmes de sélection de modèles cités ci-dessus, on observe que des pénalités suffisantes pour garantir un modèle raisonnable doivent *a minima* vérifier les conditions suivantes :

- $\text{pen}(m) > \epsilon^2 \frac{D_m}{n}\left(1+2\sqrt{L_m}+2L_m\right)$ dans le cas des moindres carrés gaussiens,
- $\text{pen}(m) \geq \frac{D_m}{n}\left(\sqrt{c_1}+\sqrt{2(1+c_1)L_m}\right)^2$ avec $c_1 > \frac{1}{2}$ dans le cas du théorème de G. Castellan pour l'estimation de densité.

Ceci soulève naturellement la question miroir de savoir si il existe un niveau minimal de pénalités en deçà duquel la procédure de sélection de modèle produit invariablement un modèle de mauvaise qualité. Au delà de son aspect théorique, cette question revêt une grande importance pratique.

1.5 Pénalités minimales

Le concept de pénalités minimale a été introduit par L. Birgé et P. Massart dans [BM07] en 2007 puis étudié plus avant dans [AM09], dans le but de fournir une estimation du terme de variance ϵ^2 directement à partir de la procédure de sélection.

L'heuristique de la pénalité minimale est la suivante : par les relations

$$\begin{aligned} Y &= s + \epsilon W, \\ \hat{s}_m &= \Pi_m(s + \epsilon W), \end{aligned}$$

le risque empirique de chaque modèle $m \in \mathcal{M}$ s'écrit

$$\begin{aligned} \gamma_n(\hat{s}_m) - \gamma_n(s) &= n^{-1}\left[\|\hat{s}_m\|^2 - 2\langle \hat{s}_m, Y\rangle - \|s\|^2 + 2\langle s, Y\rangle\right], \\ &= n^{-1}\left[\|\hat{s}_m - Y\|^2 - \|s - Y\|^2\right], \\ &= n^{-1}\left[\|\hat{s}_m - Y\|^2 - \|\epsilon W\|^2\right], \end{aligned} \tag{1.1}$$

et son espérance :

$$\mathbb{E}\left[\gamma_n(\hat{s}_m) - \gamma_n(s)\right] = n^{-1}\|(\mathbb{I}_n - \Pi_m)s\|^2 - \epsilon^2 \frac{2\operatorname{Tr}\left[\Pi_m\right] - \operatorname{Tr}\left[\Pi_m^T\Pi_m\right]}{n}, \tag{1.2}$$

où le terme positif représente un biais et le terme négatif la contribution de la variance du bruit. Le risque quadratique d'un même modèle m s'écrit

$$\mathbb{E}\left[n^{-1}\|\hat{s}_m - s\|^2\right] = n^{-1}\|(\mathbb{I}_n - \Pi_m)s\|^2 + \epsilon^2 \frac{\operatorname{Tr}\left[\Pi_m^T\Pi_m\right]}{n}. \tag{1.3}$$

Dans le cadre de la régression linéaire les opérateurs Π_m sont des projecteurs et vérifient $\operatorname{Tr}\left[\Pi_m\right] = \operatorname{Tr}\left[\Pi_m^T\Pi_m\right] = D_m$. Cependant les auteurs de [AM09] observent que ce raisonnement s'applique dans le cas beaucoup plus général d'estimateurs linéaires, sous la seule condition $\operatorname{Tr}\left[\Pi_m^T\Pi_m\right] < 2\operatorname{Tr}\left[\Pi_m\right]$. C'est le cas par exemple avec la régression régularisée de Tikhonov (*ridge regression*), avec $\Pi_m = K\left[\lambda_m\mathbb{I}_n + K\right]^{-1}$ où K est la matrice d'un noyau prédéfini et $\lambda_m \geq 0$. La quantité $\operatorname{Tr}\left[\Pi_m\right]$ joue le rôle de dimension effective du modèle m, et s'apparente à un nombre de degrés de liberté.

Supposons de façon heuristique que par le fait d'inégalités de concentration (c'est-à-dire pour une famille de modèles modérément riche), le comportement des quantités ci-dessus soit uniformément assimilable à celui de leurs espérances, comme le suppose l'heuristique de Mallows. Dans ce cas si on choisit une fonction de pénalité de la forme $\operatorname{pen}(m) = \alpha\epsilon^2\frac{2\operatorname{Tr}[\Pi_m] - \operatorname{Tr}\left[\Pi_m^T\Pi_m\right]}{n}$, un minimiseur du risque empirique pénalisé

$$n^{-1}\|\hat{s}_m - Y\|^2 - n^{-1}\|\epsilon W\|^2 + \operatorname{pen}(m)$$

se comportera comme un minimiseur de l'espérance correspondante :

$$\mathbb{E}\left[n^{-1}\|\hat{s}_m - Y\|^2 + n^{-1}\|\epsilon W\|^2 + \operatorname{pen}(m)\right] + \alpha\epsilon^2\frac{2\operatorname{Tr}\left[\Pi_m\right] + \operatorname{Tr}\left[\Pi_m^T\Pi_m\right]}{n}$$

soit par Equation 1.2

$$n^{-1}\|(\mathbb{I}_n - \Pi_m)s\|^2 - (1-\alpha)\epsilon^2\frac{2\operatorname{Tr}\left[\Pi_m\right] - \operatorname{Tr}\left[\Pi_m^T\Pi_m\right]}{n}.$$

Avec des hypothèses modérées sur les qualité d'approximation de la famille $\mathcal{M}$ (par exemple $\sup_{m\in\mathcal{M},\dim(m)=D}\|\Pi_m(s)\|^2$ croissant avec D), on constate l'apparition de deux régimes :

- si $\alpha < 1$, alors la procédure choisit un modèle de très grande dimension, et sur-ajuste (*overfits*)
- si $\alpha > 1$, le critère pénalisé croit avec $\mathrm{Tr}\left[\Pi_m\right]$ à partir d'un certain seuil, et la procédure choisit un modèle de bien moins grande dimension.

En ce sens la fonction $m \mapsto pen_{min}(m) = \epsilon^2 \frac{2\,\mathrm{Tr}[\Pi_m] - \mathrm{Tr}\left[\Pi_m^T \Pi_m\right]}{n}$ a bien les qualités d'une pénalité minimale observable à partir des seules données par une discontinuité de la dimension du modèle sélectionné.

La pénalité de Mallows $m \mapsto \mathrm{pen}\left(m\right) = \epsilon^2 \frac{2\,\mathrm{Tr}[\Pi_m]}{n}$ peut quant à elle être considérée comme optimale dans la mesure où un minimiseur du risque empirique ainsi pénalisé se comporte comme un minimiseur du risque quadratique exprimé par l'Equation 1.3, ce qui est le critère recherché.

Revenant au cas des moindres carrés où les estimateurs $\{\Pi_m\}_{m\in\mathcal{M}}$ sont des projecteurs orthogonaux, les considérations heuristiques ci-dessus peuvent se résumer en :

- pénalité minimale : $\epsilon^2 \frac{\dim(m)}{n}$
- pénalité optimale : $2\epsilon^2 \frac{\dim(m)}{n}$

ce qui a amené les auteurs de [BM07] à formuler la règle empirique

$$\text{“pénalité optimale"} = 2 \times \text{“pénalité minimale"}$$

où “pénalité minimale" s'entend comme le niveau de pénalité à partir duquel la dimension du modèle choisit explose. Les auteurs de [AB09] ont ensuite ré-interprété le facteur 2 de cette relation en $\frac{2\,\mathrm{Tr}[\Pi_m]}{2\,\mathrm{Tr}[\Pi_m] - \mathrm{Tr}[\Pi_m^T \Pi_m]}$, comme il ressort de la discussion ci-dessus, où Π_m n'est plus nécessairement un projecteur.

L'heuristique de pente et la méthode du coude Les auteurs [AM09] traduisent l'heuristique de pente en l'algorithme suivant, dit du saut de dimension, applicable notamment au cas de la régression linéaire :

- intrants : $Y \in \mathbb{R}^n$ et une collection de matrices $\{\Pi_m\}_{m\in\mathcal{M}}$
- $\forall C > 0$, déterminer
  $\hat{m}(C) \in \mathrm{argmin}_{m\in\mathcal{M}} \left\{ \|\Pi_m(Y) - Y\|^2 + C\left(2\,\mathrm{Tr}\left[\Pi_m\right] - \mathrm{Tr}\left[\Pi_m^T \Pi_m\right]\right) \right\}$
- déterminer $\hat{C}$ tel que $\mathrm{Tr}\left[\Pi_{\hat{m}(\hat{C})}\right] \in [n/10, n/3]$
- sélectionner $\hat{m} \in \mathrm{argmin}_{m\in\mathcal{M}} \left\{ \|\Pi_m(Y) - Y\|^2 + 2\hat{C}\,\mathrm{Tr}\left[\Pi_m\right] \right\}$

Une méthode proche, également très utilisée, consiste à tracer le contraste empirique en fonction du nombre de degrés de liberté $\mathrm{Tr}\left[\Pi_m\right]$ et à déterminer $\hat{C}$ à partir du plus grand saut de dimension dans cette séquence. Cette heuristique est à rapprocher de la désormais classique heuristique du coude pour la régularisation de Tikhonov ([RH09] cité dans [AB09]).

Enfin une méthode alternative, non équivalente, consiste à estimer et utiliser la variance inconnue du bruit à partir de la pente de cette même courbe contraste/dimension en grande dimension.

É. Lebarbier [Leb05a] a implémenté et testé avec succès de telles méthodes de calibration pour la détection de rupture multiples dans la moyenne d'un processus gaussien.

Dans [BMM12], J.P. Baudry, C. Maugis et B. Michel donnent une discussion détaillée des aspects théoriques et des difficultés pratiques de ces algorithmes, et proposent une implémentation.

Au vu des apports prometteurs de ces algorithmes, de nombreux travaux théoriques ont visé à déterminer leurs conditions validité.

2 État de l'art

Comme indiqué ci-dessus, l'heuristique de pente apparaît dans [BM01] et les premières preuves théoriques figurent dans [BM07] pour la régression gaussienne en effet fixe sous bruit homoscédastique. Les auteurs distinguent plusieurs régimes de complexité, selon que l'effectif des modèles considérés croît avec la dimension de ces derniers de façon linéaire, polynomiale ou exponentielle. Dans cette dernière situation, qui est la plus ardue, les auteurs se placent dans le cadre de la sélection parmi une famille de variables indépendantes.

Dans [AM09] S. Arlot et P. Massart étendent ces résultats à la régression hétéroscédastique avec effet aléatoire sans faire d'hypothèse gaussienne, sous complexité polynomiale. Les preuves sont limitées au cas des histogrammes (ou régressogrammes) mais les auteurs conjecturent que les raisons en sont purement techniques et que l'heuristique reste valide dans un cadre plus général.

S. Arlot et F. Bach ont ensuite élargi l'approche théorique à un ensemble d'estimateurs linéaires, sortant du cadre strict de la régression où les opérateurs considérés sont des projecteurs [AB09]. Leur approche s'applique notamment à la régression régularisée, au lissage par spline, à la régression locale, ou enfin au choix parmi plusieurs noyaux pour la régression régularisée, sous bruit gaussien homoscédastique. L'hypothèse de complexité est polynomiale. Plus précisément les auteurs considèrent des réunions de familles continues d'opérateurs d'estimation du type $\left\{ A_\lambda = K \left[K + \lambda \mathbb{I}_n \right]^{-1} \right\}_{\lambda \in \mathbb{R}^+}$ où K est une matrice semi-définie positive donnée. Il y a donc une famille non dénombrable de modèles, cependant les auteurs montrent que la relation d'ordre naturelle au sein de chacune de ces familles limite la complexité effective qui en résulte.

Récemment A. Saumard élargit la validation de l'heuristique de pente à une classe générale de contrastes présentant certaines conditions de régularité [Sau10]. Il montre l'optimalité de cette même heuristique pour la régression sous bruit hétéroscédastique et effet aléatoire [Sau13]. Il propose des pénalités non déterministes du type *hold out*.

C. Lacour et P. Massart ont récemment mis en évidence un phénomène de pénalité minimale pour la méthode de A. Goldenshluger et O. Lepski [LM15], dans le cadre d'un modèle de bruit blanc gaussien d'une part, et pour l'estimation de densité par noyau d'autre part.

Pour l'estimation de densité par histogrammes, un théorème de G. Castellan offre une borne inférieure du risque de Hellinger d'un histogramme sélectionné par log-vraisemblance pénalisée, cependant sous l'hypothèse qu'il n'existe qu'un seul histogramme par dimension, qui est définie comme le nombre d'intervalles dans la partition correspondante, moins un (voir le Théorème 9.8 ou bien [Cas99] cité dans [Mas07, Th. 7.10 p.238]). M. Lerasle a ensuite validé l'heuristique de pente pour une perte quadratique et montré qu'elle est en un certain sens optimale, toujours sous l'hypothèse de complexité polynomiale, mais avec des pénalités non nécessairement déterministes, obtenues par ré-échantillonnage [Ler12].

2.1 Difficultés propres à la complexité exponentielle

Comme il ressort de l'exposé précédent, la preuve du phénomène de pénalité minimale présente une difficulté particulière en situation de complexité exponentielle, c'est-à-dire quand le nombre de modèle croît exponentiellement avec leur dimension. Le seul résultat dans ce sens est à notre connaissance le résultat de L. Birgé et P. Massart dans [BM07] (voir Théorème 5.8), pour la sélection libre parmi des variables gaussiennes indépendantes.

La raison technique en est la suivante : nous avons vu plus haut que se donner des poids $\{L_m\}_{m\in\mathcal{M}}$ avec la condition $\sum_{m\in\mathcal{M},D_m>0} e^{-L_m D_m} < 1$ permet définir des pénalités suffisantes d'une forme proche de

$$\mathrm{pen}\,(m) = \epsilon^2 \frac{D_m}{n}\left(1 + 2\sqrt{L_m} + 2L_m\right),$$

le contrôle uniforme sur les risques empiriques correspondants s'effectuant en quelque sorte au prix du remplacement des seuils de Mallows $\{D_m\}_{m\in\mathcal{M}}$ par des seuils proches de $\left\{D_m\left(1 + 2\sqrt{L_m} + 2L_m\right)\right\}_{m\in\mathcal{M}}$.

La condition $\sum_{m\in\mathcal{M},D_m>0} e^{-L_m D_m} < 1$ entraîne notamment pour chaque modèle m de dimension non nulle que

$$\sum_{\substack{m'\in\mathcal{M}\\ D_{m'}=D_m}} e^{-L_{m'} D_m} < 1$$

et par un simple argument de convexité que

$$\bar{L}_m \geq \frac{\log(N_m)}{D_m}$$

où N_m désigne le nombre des modèles de même dimension que m et $\bar{L}_m$ la moyenne des poids associés.

Les poids L_m deviennent donc significatifs voire prédominants dès lors que la complexité de la famille de modèles est exponentielle, contrairement à la situation de complexité polynomiale. En d'autres termes, en situation exponentielle, le contrôle uniforme sur le risque empirique ne s'obtient qu'à des niveaux significativement éloignés de son espérance pour chaque modèle donné, ce qui a trait bien entendu à la forme particulière des queues de distributions de type $\chi_2(D_m)$. Ceci a pour effet de donner un rôle central aux propriétés d'approximation de la famille de modèles $\mathcal{M}$.

En effet, les théorèmes généraux qui valident les formes de pénalités suffisantes indiquées ci-dessus ne font que très peu d'hypothèses sur la famille de modèles candidats. Rien n'interdit par exemple à cette famille d'être redondante, certains modèles figurants plusieurs fois, ou étant très proches les uns des autres (au sens d'une distance entre matrices de projection par exemple). Cependant les pénalités proposées correspondent nécessairement au cas le plus adverse où la famille concernée posséderait les meilleures propriétés d'approximation possibles compte tenu de sa complexité. La situation de référence dans ce domaine est celle de la sélection libre parmi des variables gaussiennes indépendantes, du fait de l'isotropie par échange de coordonnées de la famille de modèles considérée.

L'analyse du phénomène de pénalité minimale demande pour sa part de minorer le supremum de quantités telles que $\|\Pi_m(Y)\|^2 = \|\Pi_m(s+\epsilon W)\|^2$ pour les modèles m de même dimension D recensés dans la famille $\mathcal{M}$, dans l'espoir de montrer que ce supremum s'élève suffisamment au-dessus de la valeur de référence

$$\sup_{m\in\mathcal{M},\dim(S_m)=D} \|\Pi_m(s)\|^2 + \epsilon^2 D$$

pour se rapprocher des valeurs suggérées par les pénalités suffisantes.

Ceci n'est possible qu'avec certaines hypothèses sur les qualités d'approximation de la famille de modèles considérés. De façon heuristique, en situation de complexité polynomiale, l'écart à combler n'est pas très important et la simple hypothèse qu'il existe par exemple un modèle par dimension, ou un modèle d'assez grande dimension suffit à jouer ce rôle, sans préjuger d'autres difficultés techniques. En revanche, par le rôle prédominant qu'y joue la complexité représentée par les poids L_m, la situation exponentielle requiert quant à elle des hypothèses plus fortes sur la famille considérée.

Dans ce travail, nous examinons des familles de modèles de complexité exponentielle couramment employées à base d'arbres de régression, de partition récursive, de segmentation ou enfin de pavage en bloc rectangulaires.

2.2 Segmentations et partitions

Souvent les observations disponibles proviennent de relevés à intervalles réguliers dans le temps : points d'une série chronologique, ou dans l'espace : pixels d'une image. Un but naturel de l'estimation statistique est alors d'identifier des ensembles connexes d'observations dont certaines propriétés statistiques sont homogènes, et les frontières entre ces régions. En segmentation d'un signal séquentiel, ces frontières sont communément appelées ruptures (*breakpoints*) ou points de changement (*changepoints*). La détermination de leur nombre et de leur position est un problème classique en statistiques, objet de recherches très actives qui concerne de nombreux domaines d'application comme par exemple la climatologie [RCW+07], la génomique [PRL+05, PLH+11, PHL+12], la sécurité des systèmes, la détection d'anomalie ou d'intrusion [SG12], la finance [LT05], la géologie [FC03], l'analyse des signaux vitaux [Lav05], l'analyse de la parole et du son [DLRY06], la détection de séquence dans les enregistrements audiovisuels.

On cherche donc souvent à identifier une structure spatiale dans le signal étudié. Une motivation indépendante, du point de vue de l'efficacité de l'estimation d'un signal source caché, peut consister à chercher à tirer profit des bonnes propriétés d'approximation et

de la maniabilité de familles de fonction en escalier convenablement choisies. De ce cas on sait souvent que le signal source n'a pas la simplicité d'une telle une fonction en escalier, et ce n'est pas nécessaire, même si on en cherche une approximation de ce type. Dans ce domaine, les techniques d'arbres de décision et de régression proposées par Breiman, Friedman, Olshen et Stone en 1984 [BFOS84, GN05] ont acquis une grande popularité auprès des praticiens en raison de leur efficacité notamment dans le cadre de méthodes d'ensemble comme les forêts aléatoires [HTF09, Arl14].

Un modèle de segmentation ou de partition se décrit en général de façon hiérarchique : une segmentation définit des régions spatiales ou temporelles, et sur chacune de celles-ci certains paramètres statistiques considérés comme constants, par exemple la moyenne d'un signal gaussien homoscédastique. Le signal observé est considéré comme un reflet bruité de ces paramètres, selon une loi d'émission qui respecte en général l'indépendance entre les régions de la segmentation (voir [CLLLR15] pour une exception à ce principe). Dans l'approche fréquentiste, le problème se présente comme une sélection de modèle, chaque segmentation candidate représentant un modèle différent [Leb05a, AM09]. Dans l'approche bayésienne (par exemple [Fea06, LL01]), la structure de la segmentation et les paramètres individuels de chaque segment peuvent être eux-mêmes probabilisés, selon une loi hiérachique dont on cherche à déterminer les hyper-paramètres ou à échantillonner la distribution *a posteriori*.

Si les approches fréquentiste et bayésienne diffèrent par leurs choix de modélisation initiaux, elles ont de nombreux points communs puisque la sélection de modèle demande de produire, sinon une loi de probabilité, une pondération de somme finie sur les modèles candidats et des pénalités dont il convient de calibrer le niveau. Le choix de référence dans ce domaine est le contraste L_2 pénalisé avec le jeu de pénalités proposé par É. Lebarbier dans [Leb05a, Leb05b]. La figure 1 offre un exemple de selection de modèle basé sur de telles pénalités pour des ruptures dans la moyenne d'un signal séquentiel gaussien. Les segmentations sont obtenues par la méthode dite de la programmation dynamique, sur laquelle nous revenons par la suite. Cette figure fait apparaître également les courbes de contraste pour des pénalités insuffisantes et pour une séquence de modèles 'sur-segmentés' qui font l'objet de la section 1.5.

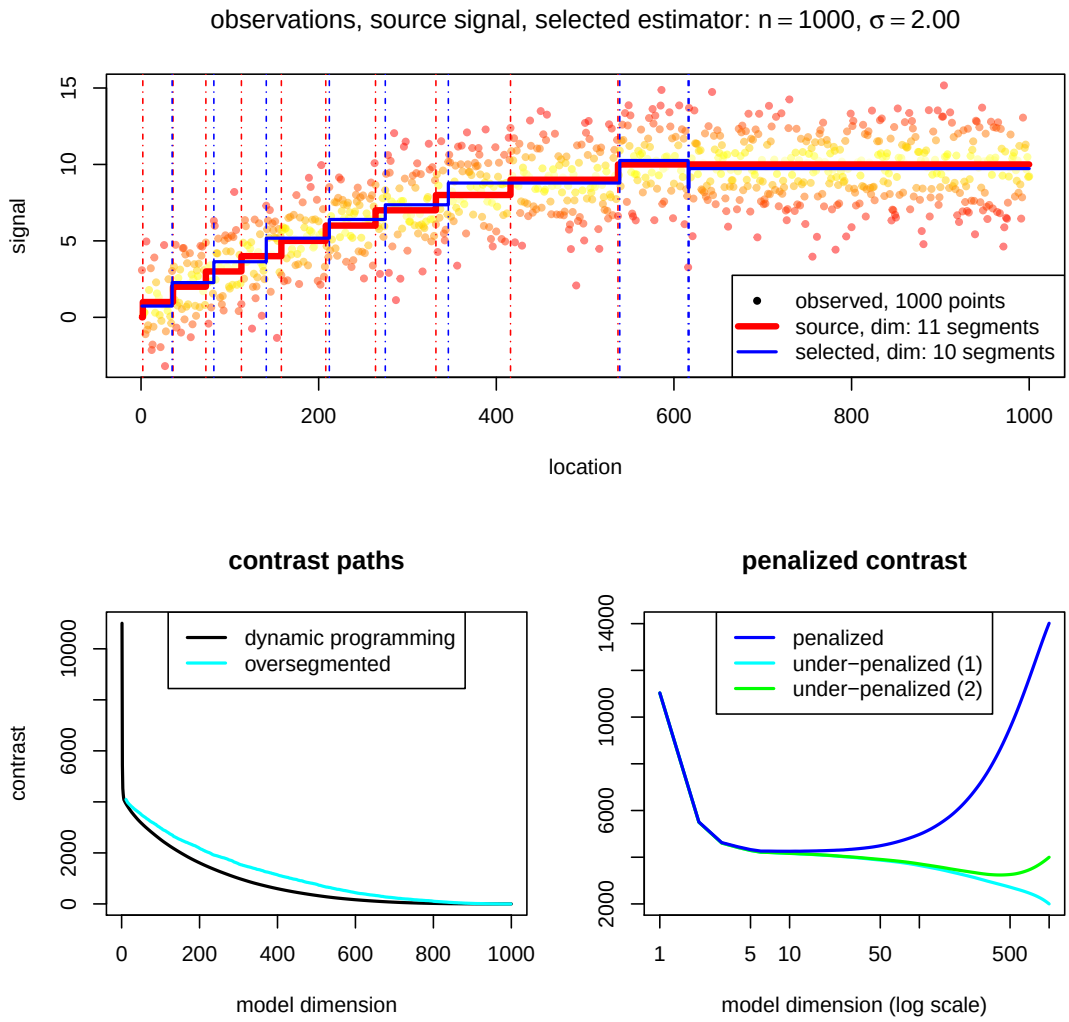


FIGURE 1 – Detection de ruptures dans la moyenne d'un signal gaussien. Segmentations optimales à dimension fixées obtenue par l'algorithme de programmation dynamique.

Une introduction aux approches les plus utilisées peut se trouver dans [Leb05b],[EFK11], [Pic07], [BN93] (pour la détection en ligne), [GM05] pour une courte synthèse des principaux algorithmes, [CG00, CG12]. Citons également le site [KNAE12] qui propose différentes ressources et références.

La combinatoire des segmentations exclut en général de traiter celles-ci par énumération naïve : il y a par exemple plus de 10^{12} manières de segmenter une séquence aussi courte que 100 points en 10 intervalles non vides, et plus de 10^{21} pour une séquence de 1000 points. De ce fait, dans ce domaine de recherche les considérations algorithmiques reçoivent autant d'attention que la modélisation proprement dite.

Sans entrer dans le détail des algorithmes les plus utilisés, nous pourrions les diviser grossièrement en trois classes : les algorithmes dérivés de la programmation dynamique, les méthodes de segmentation récursives et les méthodes du type *Markov Chain Monte Carlo*.

D'un point vue numérique, ces méthodes font en général intervenir un critère de coût à optimiser, qu'il s'agisse de la log-vraisemblance des paramètres dans le cadre

bayésien ou d'un contraste pénalisé dans le cadre de la sélection de modèle. La structure hiérarchique employée lors de la modélisation permet idéalement, pour une segmentation donnée, de calculer ce critère sous la forme d'une somme de termes calculables segment par segment, aisément manipulable lors des calculs. Une propriété désirable est alors que les restrictions d'une segmentation optimale sont elles-mêmes optimales pour le problème restreint.

Les méthodes dérivées de l'algorithme de la programmation dynamique [BD62, AL89] s'appliquent à la segmentation d'un signal séquentiel et consistent à rechercher une segmentation optimale de dimension donnée en élargissant progressivement une zone sur laquelle les segmentations optimales de dimension inférieure sont connues, comme pour la recherche de géodésique dans un graphe. Si l'on écarte l'énumération naïve des partitions, cette famille de méthodes est la seule à identifier exactement la segmentation optimale [CLLLR15, p.2]. L'algorithme de Viterbi [CMR05, p. 125] pour le décodage des chaînes de Markov cachées [CMR05] peut être considéré comme une de ses généralisations. Dans la pratique, cette méthode requiert un temps de calcul quadratique en fonction du nombre d'observations, ce qui est prohibitif pour certaines applications. Des heuristiques sont alors nécessaires pour opérer une sélection sur les points de ruptures envisagés [FL07, KFE12, Rig10].

La méthode de segmentation récursive, issue de la méthode dite 'binaire' dont les origines sont attribuées à [ECS65, AJS74], est certainement la plus utilisée en pratique, notamment pour le calibrage des arbres de régression [BFOS84]. Elle consiste à choisir un unique point rupture prometteur du point de vue du gain de contraste, et à itérer cette opération sur les segments obtenus, sans jamais remettre en cause les choix précédents. On obtient ainsi un arbre de segmentation de profondeur ajustable, que l'on réduit ensuite en éliminant les branches qui contribuent le moins à l'objectif. La Section 8.4 propose une description plus détaillée de l'algorithme CART (*classification and regression trees*) très employé pour les arbres de régression et de décision, qui - bien que s'interdisant d'emblée la plupart des segmentations - produit des résultats souvent satisfaisants en pratique. Dans [GN05], S. Gey and É. Nedelec détaillent cet algorithme et analysent sa performance.

Enfin dans le cadre bayésien les méthodes de type *Markov Chain Monte Carlo* [CMR05, chap. 6] tirent parti de l'algorithme de Metropolis-Hastings [LL01] pour échantillonner l'espace des segmentations selon sa loi *a posteriori*, estimer les probabilités marginales de présence d'un point de rupture, ou encore exhiber la segmentation jugée la plus probable. De façon similaire, dans la perspective fréquentiste de la sélection de modèle, les méthodes d'optimisation stochastiques comme celle du recuit simulé [CMR05, p. 502] permettent d'exhiber dans l'espace des segmentations un optimum local, que l'on espère proche de l'optimum global, sinon égal à ce dernier. Mentionnons une version en 'température nulle' simpliste mais très efficace en pratique qui consiste itérer l'opération suivante à partir d'une segmentation arbitraire : supprimer aléatoirement un point de changement pour le replacer à la position de moindre coût.

De nombreux travaux cherchent à élargir le cadre présenté brièvement ci-dessus, qu'il s'agisse par exemple d'analyser les données au fil d'une seule lecture en ligne [FC03], de s'affranchir de l'hypothèse d'indépendance entre les segments [CLLLR15] ou de repérer des changements dans des objets dont la métrique n'est pas euclidienne comme des graphes ou des histogrammes [Arl14].

3 Principaux résultats

Dans ce travail, nous présentons des résultats pour le phénomène de pénalité minimale pour la sélection de modèle non-asymptotique par contraste pénalisé, pour certaines familles de modèles de complexité exponentielle.

Dans le cadre gaussien, le premier résultat traite de l'estimation d'un signal continu sur un domaine de $\mathbb{R}^d$ sous bruit blanc brownien, quand la famille de modèles provient d'une méthode de partition récursive b-adique, comme il est courant avec les arbres de régression. Nous concluons qu'un niveau de pénalité minimale existe et qu'il est lié au niveau de pénalité suffisante offert par le théorème de sélection de modèle de L. Birgé et P. Massart (5.6). La preuve repose sur une observation combinatoire simple sur ces familles de modèles vues comme des familles d'arbres planaires, plus précisément, sur une borne du nombre minimum de feuilles (nœuds terminaux) d'un tel arbre lorsqu'un ensemble donné de feuilles est prescrit. Nous utilisons cette borne pour montrer comment certains modèles - en un certain sens de dimension modérée - peuvent capter dans le contraste empirique une large part de la variabilité du bruit. Bien que ce premier résultat se place en dimension infinie et soit de ce point de vue intrinsèquement de nature asymptotique, il entre bien dans le cadre de la sélection de modèle non asymptotique.

Le deuxième et le troisième résultat traitent de situations analogues cette fois-ci pour un signal discret sur un domaine fini, d'abord avec une borne inférieure sur le risque du modèle sélectionné, puis - sous l'hypothèse nulle $s = 0$ - avec une borne inférieure sur sa dimension. La borne inférieure de risque ne demande aucune hypothèse sur le paramètre inconnu s. Nous faisons une hypothèse combinatoire simple sur les familles de modèles considéré, hypothèse qui se trouve remplie dans de nombreux cas pratiques où les familles de modèles sont construites par segmentation d'un intervalle ou par partitionnement récursif d'un domaine de $\mathbb{N}^d$, comme, ici aussi, dans les exemples de la segmentation linéaire ou des arbres de régression. Les pénalités minimales obtenues sont plus faible par un facteur $\mathrm{C}_{\mathrm{ext}}$ que celles que prouvent L. Birgé et P. Massart pour la sélection libre parmi des variables gaussiennes indépendantes. Ce facteur $\mathrm{C}_{\mathrm{ext}}$ dépend de la géométrie particulière de la famille de modèles, il est proche de 2 pour la segmentation libre d'une séquence. Nous proposons un ensemble d'expériences numériques basés sur plusieurs méthodes de partitionnement récursif d'images bidimensionnelles. Ce choix permet de visualiser le comportement de la procédure de sélection de modèle en relativement grande dimension.

Le quatrième résultat montre qu'avec des hypothèses analogues sur une famille d'histogrammes, le même phénomène de pénalité minimale apparaît pour l'estimation de densité par log-vraisemblance pénalisée, du moins dans l'hypothèse où la densité inconnue est constante. Le résultat proposé minore conjointement le risque et la dimension du modèle sélectionné.

Enfin nous présentons un cas simpliste gaussien pour illustrer le lien qui existe à notre sens entre le phénomène de pénalité minimale sous forte complexité et les propriétés d'approximation qu'une forme d'isotropie confère à la famille de modèles candidats.

Ces différents résultats font appel pour leur preuve à un ensemble d'inégalités de déviation qui sont regroupées dans une section séparée.

Enfin dans un but pratique nous regroupons en annexe une synthèse de différentes

méthodes pour équiper de pondérations différentes familles de modèles construites par partitionnement récursif.

4 Main results

In this work we present some non-asymptotic results relating to the minimal penalty phenomenon with model families of exponential complexity.

In the Gaussian framework, our first result deals with the estimation of a continuous signal over a domain in $\mathbb{R}^d$ under white noise when the model family arises from a recursive b-adic partitioning process, as in segmentation or with regression trees. Our main conclusion is that a minimum penalty threshold exists relatively close to the level of sufficient penalties. The proof principally relies on simple combinatorial observations on certain families of planar trees, more precisely on an estimate on the minimum number of leaves of a b-tree with a prescribed set of leaves. We use this estimate to show how certain models of in a given sense moderate dimension capture a large part of the noise variability into the empirical contrast.

The second and the third result deal with similar situations for a discrete signal over a finite domain, first with a risk lower bound and then with a dimension lower bound on the selected model. The results are based on combinatorial assumptions on the model family that we believe to be customary in practice, notably in the domain of regression trees or linear signal segmentation. We provide a set of numerical experiments based on various segmentation methods for a two-dimensional image. This choice allows to visualize the behaviour of the model selection procedure in situations of relatively high dimension.

The fourth result shows that with similar assumptions on a family of histograms, also under exponential complexity, the same minimal penalties phenomenon appears in the context of density estimation by penalized log-likelihood.

Last we present a Gaussian toy problem to illustrate how in our sense, the minimal penalty phenomenon relates to a form of isotropy of the model family, as a proxy to its approximation properties.

The results presented call on a set of deviation inequalities, that we gathered in a dedicated section.

Last for reference in annex we offer a summary of different methods to equip with weights different model families arising from recursive partitioning.

Part II

Minimal penalties

Summary

5 Technical Preliminaries for Gaussian model selection 33

5.1 Gaussian model selection . . . 33

5.2 Birgé-Massart's Gaussian model selection theorem . . . 34

5.2.1 Potential difficulties connected with bad penalties choice . . . 36

5.2.2 Risk of the penalized estimator and dimension of the selected model . . . 36

5.2.3 Application to segmentation . . . 36

5.3 Variable selection, a minimal penalty theorem of L.Birgé and P. Massart . . . 36

6 A motivating example: estimating a continuous signal under white noise with step-functions by recursive partitioning 39

6.1 Estimation procedure . . . 39

6.2 Assessing the complexity of the model family . . . 40

6.3 Sufficient penalties . . . 41

6.4 A geometrical property . . . 42

6.5 Minimal penalties . . . 44

7 Technical preliminaries on signal segmentation and partitioning 45

7.1 Partitions . . . 45

7.2 Recursive partitioning . . . 47

8 Minimal penalties for the partitioning of a discrete Gaussian signal 48

8.1 Definitions and assumptions on the model family structure . . . 49

8.1.1 Completion rule . . . 49

8.1.2 Binomial complexity . . . 51

8.1.3 Linking completion rule and binomial complexity . . . 52

8.2 Sufficient penalties . . . 53

8.3 Minimal penalties . . . 53

8.3.1 A risk lower bound . . . 53

8.3.2 A dimension lower bound . . . 55

8.4 Numerical experiments for section 8 . . . 57

8.4.1 Introduction and motivation . . . 57

8.4.2 Description . . . 57

8.4.3 Discussion . . . 60

8.5 Numerical experiments for section 8: results . . . 62
8.5.1 Image squares, free binary tree . . . 62
8.5.2 Image squares, free quad tree . . . 67
8.5.3 Image squares, regular binary tree . . . 72
8.5.4 Image squares, regular quad tree . . . 77
8.5.5 Image lone tree, free binary tree . . . 82
8.5.6 Image lone tree, free binary tree . . . 87
8.5.7 Image lone tree, free binary tree . . . 92
8.5.8 Image lone tree, free binary tree . . . 97
8.5.9 Image face, free binary tree . . . 102
8.5.10 Image face, free binary tree . . . 107
8.5.11 Image face, free binary tree . . . 112
8.5.12 Image face, free binary tree . . . 117

9 A first extension to histogram selection for density estimation 122
9.1 Technical preleminaries: model selection for density estimation with histograms . . . 122
9.1.1 Castellan's histogram selection theorem for density estimation 124
9.1.2 Castellan's minimal penalty theorem for density estimation . 125
9.2 Definitions and main assumptions . . . 127
9.3 Sufficient penalties under the null hypothesis . . . 131
9.4 A risk and dimension lower bound . . . 132
9.4.1 Minimal penalty . . . 133

10 Remark on the isotropy of the model family 135
10.1 Toy problem . . . 135

11 Proofs for Sections 5, 6 and 8 138
11.1 Proof of Corollary 5.7 . . . 138
11.2 Proof of Proposition 6.1 . . . 139
11.3 Proof of Lemma 6.2 . . . 142
11.4 Proof of proposition 6.4 . . . 143
11.5 Proof of Proposition 8.5 . . . 147
11.6 Proof of Proposition 8.6 . . . 150
11.7 Proof of Proposition 8.7 . . . 151
11.7.1 A risk lower bound lemma for nested models . . . 151
11.7.2 Proof of Lemma 11.1 . . . 152
11.7.3 Concluding the proof of Proposition 8.7 . . . 153
11.8 Proof of Proposition 8.8 . . . 156
11.8.1 Controlling the contrast contribution of low dimensional models 156

11.8.2 Concluding the proof of Proposition 8.8 157

12 Proofs for section: 9 A first extension to histogram selection ... 159

12.1 Proof of Proposition 9.13 159

12.2 Proof of Proposition 9.14 160

12.2.1 Deviation count, deviation rate 160

12.2.2 Selected parts and adverse model 160

12.2.3 Expected empirical contrast of the adverse model 161

12.2.4 Risk 163

12.2.5 Variability of the number of selected parts 164

12.2.6 Concluding the proof of Proposition 9.14 166

13 Proofs for section: 10 Remark on the isotropy of the model family 170

13.1 Proof of Proposition 10.1 170

In this part we present some non-asymptotic results relating to the minimal penalty phenomenon with model families of exponential complexity.

5 Technical Preliminaries for Gaussian model selection

In this section, we recall some definitions for Gaussian model selection and state two theorems from L. Birgé and P. Massart.

5.1 Gaussian model selection

Recall that, in some Hilbert space $\mathcal{H}$, with norm $\|\cdot\|$ and scalar product $\langle \cdot, \cdot \rangle$, a linear *isonormal process* indexed by a suitable linear subspace S is a centered and Gaussian linear process with covariance structure $\mathbb{E}\left[W(t)\,W(u)\right] = \langle t, u \rangle$. We denote $d(\cdot, \cdot)$ the associated quadratic distance.

L. Birgé and P. Massart consider the statistical problem of estimating the unknown parameter $s \in \mathcal{H}$ when one observes the Gaussian linear process Y indexed by S defined by

Definition 5.1 (Gaussian linear process)**.**

$$Y(t) := \langle s, t \rangle + \epsilon\, W(t) \text{ for all } t \in S,$$

where W denotes a linear isonormal process and ϵ is a known level of noise.

Let us define a *linear model* as a finite-dimensional (and possibly zero-dimensional) subspace of S. Some countable (possibly finite) collection $\{ S_m, m \in \mathcal{M} \}$ of models, with respective dimensions $\{ D_m, m \in \mathcal{M} \}$ (also denoted $\{ |m|, m \in \mathcal{M} \}$), is available. For a model m the unknown s is not usually an element of S_m, contrary to its projection s_m:

Definition 5.2 (Projections)**.** For any model S_m, define the projection s_m of s on S_m as:

$$s_m := \operatorname*{argmin}_{t \in S_m} \|s - t\|^2,$$

or equivalently

$$s_m := \operatorname*{argmin}_{t \in S_m} \|t\|^2 - 2\langle s, t \rangle .$$

As these projections are unknown, one builds on each model S_m the corresponding least squares estimator, $\hat{s}_m$ defined as the minimizer with respect of $t \in S_m$ of the least squares criterion.

Definition 5.3 (Empirical contrast)**.** The *least squares criterion* or *empirical contrast* $\gamma_n(t)$ is defined for any $t \in \mathbb{R}^n$ as

$$\gamma_n(t) = \|t\|^2 - 2\,\mathrm{Y}(t).$$

The quality of a model S_m is quantified by the corresponding quadratic risk:

Definition 5.4 (Quadratic risk)**.** The *quadratic risk* $\mathcal{R}_m(s)$ of a model m is defined as:

$$\begin{aligned}\mathcal{R}_m(s) &:= \mathbb{E}\left[\|\hat{s}_m - s\|^2\right], \\ &= \mathrm{d}^2\left(s, S_m\right) + \epsilon^2 D_m.\end{aligned}$$

where $\mathbb{E}[\cdot]$ denotes the expectation of functions of the process $Y(\cdot)$ described by Definition 5.1. An ideal model s is one that minimizes $\mathcal{R}_m(s)$, an inaccessible goal since the bias term $\mathrm{d}^2\left(s, S_m\right)$ in Definition 5.4 is unknown. It is also typically impossible to choose an exactly ideal model from the data, but with the penalization approach one can hope to build some $\hat{m}$ satisfying

$$\mathbb{E}\left[\|\hat{s}_{\hat{m}} - s\|^2\right] \leq \mathcal{C}_n \inf_{m \in \mathcal{M}} \mathbb{E}\left[\|\hat{s}_m - s\|^2\right] + R_n \tag{5.1}$$

where the constant $\mathcal{C}_n$ is as close as possible to 1 and the R_n term is small compared to the risk of the oracle. The penalization approach to model selection consists in selecting the model that minimizes a suitably defined penalized contrast criterion:

Definition 5.5 (Selected model)**.** The *selected model* $\hat{m}$ is defined as

$$\begin{aligned}\hat{m} &= \operatorname*{argmin}_{m \in \mathcal{M}} \left\{\gamma_n\left(\hat{s}_m\right) + \operatorname{pen}(m)\right\}, \\ &= \operatorname*{argmin}_{m \in \mathcal{M}} \left\{-\|\hat{s}_m\|^2 + \operatorname{pen}(m)\right\},\end{aligned}$$

when this quantities are defined, where $\operatorname{pen}(\cdot)$ denotes a suitable non-negative function defined on $\mathcal{M}$.

Note that for all $m \in \mathcal{M}$ the quantities $\gamma_n\left(\hat{s}_m\right)$ and $-\|\hat{s}_m\|^2$ are equal since the estimator $\hat{s}_m$ is the orthogonal projection $\Pi_m(Y)$ of the observed Y on the linear model S_m. So by Definition 5.3,

$$\begin{aligned}\gamma_n\left(\hat{s}_m\right) &= \|\Pi_m(Y)\|^2 - 2Y(\Pi_m(Y)), \\ &= \|\Pi_m(Y)\|^2 - 2\left\langle \Pi_m(Y), Y\right\rangle, \\ &= -\|\Pi_m(Y)\|^2, \\ &= -\|\hat{s}_m\|^2.\end{aligned}$$

5.2 Birgé-Massart's Gaussian model selection theorem

In the framework of Gaussian linear models, the following theorem offers an answer to the main question of model selection: how to choose the penalization function $\operatorname{pen}(\cdot)$ in order to ensure that the minimum penalized contrast selection procedure in Definition 5.5 actually yields an estimator $\hat{s}_{\hat{m}}$, and that $\hat{s}_{\hat{m}}$ is of good quality. L. Birgé and P. Massart give an answer with a non-asymptotic risk bound on the selected estimator $\hat{s}_{\hat{m}}$, in the form of an *oracle inequality*.

Theorem 5.6. *[BM07, Theorem 1] Given the collection of models* $\{S_m\}_{m\in\mathcal{M}}$*, let us consider*

a family of non-negative weights $\{L_m\}_{m\in\mathcal{M}}$ *satisfying*

$$\Sigma = \sum_{m\in\mathcal{M}, D_m>0} \exp\left[-L_m D_m\right] < +\infty,$$

two numbers, $\theta \in (0,1)$ *and* $\kappa > 2-\theta$*,*

and let us assume that there exists a finite (possibly empty) subset $\bar{\mathcal{M}}$ *of* $\mathcal{M}$ *such that the penalty function* $\mathrm{pen}(\cdot)$ *satisfies:*

$$\mathrm{pen}(m) \geq \mathcal{Q}_m \text{ for } m \in \mathcal{M}\setminus\bar{\mathcal{M}}$$

with

$$\mathcal{Q}_m = \epsilon^2 D_m\left(\kappa + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1}L_m\right) \forall m \in \mathcal{M}.$$

Then the corresponding penalized projection estimator $\tilde{s} = \hat{s}_{\hat{m}}$ *with* $\hat{m}$ *given by Definition 5.5 exists a.s. and satisfies:*

$$\begin{aligned}(1-\theta)\,\mathbb{E}\left[\|s-\tilde{s}\|^2\right] \leq \inf_{m\in\mathcal{M}}&\left\{\mathrm{d}^2(s,S_m) + \mathrm{pen}(m) - \epsilon^2 D_m\right\}\\ &+ \sup_{m\in\bar{\mathcal{M}}}\left\{\mathcal{Q}_m - \mathrm{pen}(m)\right\}\\ &+ \epsilon^2\Sigma\left[(2-\theta)^2(\kappa+\theta-2)^{-1} + 2\theta^{-1}\right].\end{aligned}$$

If, in particular we fix $\kappa = 2$ *and* $\mathrm{pen}(m) = \mathcal{Q}_m$ *whatever* $m \in \mathcal{M}$ *then*

$$\begin{aligned}(1-\theta)\,\mathbb{E}\left[\|s-\tilde{s}\|^2\right] \leq \inf_{m\in\mathcal{M}}&\left\{\mathrm{d}^2(s,S_m) + \epsilon^2 D_m\left[1 + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1}L_m\right]\right\}\\ &+ \epsilon^2\Sigma\theta^{-1}\left[(2-\theta)^2 + 2\right].\end{aligned}$$

The following Corollary is a simple specialization of Theorem 5.6 where the penalties are close to the limit form, at which we come back later on:

$$\mathrm{pen}(m) > \epsilon^2 D_m\left(1 + 2\sqrt{L_m} + 2L_m\right).$$

Corollary 5.7. *Consider a model family* $\mathcal{M}$*, a family of non-negative weights* $\{L_m\}_{m\in\mathcal{M}}$ *satisfying*

$$\sum_{m\in\mathcal{M}, D_m>0} \exp\left[-L_m D_m\right] < 1,$$

and a number η *with* $0 < \eta < 1$ *such that the penalty function satisfies for any* $m \in \mathcal{M}$*:*

$$\mathrm{pen}(m) \geq \frac{1+\eta}{1-\eta}\epsilon^2 D_m\left(1 + 2\sqrt{L_m} + 2L_m\right).$$

Then the corresponding penalized projection estimator $\tilde{s} = \hat{s}_{\hat{m}}$ *with* $\hat{m}$ *given by Definition 5.5 exists a.s. and satisfies:*

$$\eta\,\mathbb{E}\left[\|s-\tilde{s}\|^2\right] \leq \inf_{m\in\mathcal{M}}\left\{\mathrm{d}^2(s,S_m) + \mathrm{pen}(m) - \epsilon^2 D_m\right\} + \epsilon^2\frac{(1+3\eta)}{\eta(1-\eta)}.$$

The proof is given in Section 11.1.

5.2.1 Potential difficulties connected with bad penalties choice

It follows from Theorem 5.6 that a typical choice of penalties should be of the form

$$\text{pen}(m) = K\epsilon^2 D_m \left(1 + a\sqrt{L_m} + bL_m\right), \tag{5.2}$$

and the limiting condition

$$\text{pen}(m) > K\epsilon^2 D_m \left(1 + 2\sqrt{L_m} + 2L_m\right) \tag{5.3}$$

is required for the proof given by Birgé and Massart. This raises the following question: does violating the limiting condition above lead to a bad penalized estimator ? This question is usually expressed in terms of existence of a *minimal penalty*. A large literature exists on the issue of minimal penalties.

5.2.2 Risk of the penalized estimator and dimension of the selected model

A strong motivation for studying the minimal penalties phenomenon is that usually the ϵ^2 (or σ^2/n) term in Equation 5.5 is unknown. One then has to resort to data-driven methods, of which a large number consist in monitoring the dimension of the model that the penalized least squares procedure selects for various penalty levels. When gradually increasing the pen(m) penalty function, a minimal penalty level is typically associated to a drop in the dimension of the selected model, drop towards a 'reasonable' dimension.

5.2.3 Application to segmentation

For the multiple segmentation of a Gaussian sequential signal, one typically encounters situations where the number of models of the same dimension than S_m is in the same order of magnitude than $\binom{n-1}{|m|-1}$. Based on a previous version of Theorem 5.6, É. Lebarbier showed in [Leb05a] that penalties of the form

$$\text{pen}(m) = |m|\epsilon^2 \left\{ c_1 \log\left(\frac{n}{|m|}\right) + c_2 \right\} \tag{5.4}$$

lead to satisfying oracle inequalities, numerical calibration pointing towards the choice of $c_1 = 2$ and $c_5 = 5$. A key observation in order to choose L_m weights satisfying the conditions of theorem 5.6 is that $\sum_{j=0}^{D} \binom{n}{j} \leq \left(\frac{en}{D}\right)^D$ for all integers $1 \leq D \leq n$ [Mas07, proposition 2.5].

5.3 Variable selection, a minimal penalty theorem of L.Birgé and P. Massart

A typical example is the problem of variable selection in Gaussian linear regression.

We observe n independent variables $Y_1, \cdots, Y_n$ with $Y_i \sim \mathcal{N}(s_i, \sigma^2)$, which can be written in vector form as

$$Y = s + \sigma\xi, \tag{5.5}$$

with

$$\begin{aligned} Y &= (Y_i)\,, \\ s &= (s_i) \in \mathbb{R}^n, \\ \xi &\sim \mathcal{N}(0, \mathbf{Id}_n)\,. \end{aligned}$$

In this case Y can be identified by duality with a linear operator on the Hilbert space $\mathbb{R}^n$, or equivalently to the Gaussian linear process $Y(\cdot)$ indexed by $\mathbb{R}^n$ and defined by

$$\begin{aligned} Y(t) &= \langle Y,\, t\rangle_n \\ &= \langle s,\, t\rangle_n + \sigma\,\langle \xi,\, t\rangle_n \\ &= \langle s,\, t\rangle_n + \epsilon\, W(t), \end{aligned} \tag{5.6}$$

with $\epsilon = \sigma/\sqrt{n}$ where W is a linear isonormal process indexed by $\mathbb{R}^n$ and $\langle\cdot,\,\cdot\rangle_n$ denotes the scalar product corresponding to the normalized Euclidean norm $\|\cdot\|_n$ on $\mathbb{R}^n$ defined by $\|t\|^2_n = n^{-1}\sum_{i=1}^n t_1^2$.

In order to estimate the unknown parameter s, one typically considers a set of potential variables $\{\phi_\lambda, \lambda \in \Lambda_N\}$, $\Lambda_N = \{1, 2, \cdots, N\}$, with $\phi_\lambda \in \mathbb{R}^n$ and N possibly large. To each subset m of Λ_N corresponds a linear regression model:

$$Y(i) = \sum_{\lambda\in m} \alpha_\lambda \phi_\lambda(i) + \sigma\xi_i \text{ for } 1 \le i \le n \text{ with } \xi_1, \cdots, \xi_n \text{ i.i.d. } \mathcal{N}(0,1)\,.$$

Building a good model amounts to select some influential variables from the set $\{\phi_\lambda, \lambda \in \Lambda_N\}$. To be precise, one would like to select a subset m of Λ_N that minimizes (at least approximately) the risk $\mathbb{E}\left[\|\hat{s}_m - s\|^2_n\right]$ where $\hat{s}_m$ denotes the least squares estimator corresponding to the stochastic model. Using the identification given by Equation 5.6, assuming that σ is known and denoting by S_m the linear span of the vectors ϕ_λ, $\lambda \in m$, a variable selection problem can be viewed as a model selection problem among a subset of the collection $\{S_m\}_{m\subset\Lambda_N}$, as previously defined.

In free selection among N variables, the model family $\mathcal{M}$ is taken from all the subsets of Λ_N, so that $\mathcal{M} = \{S_m\}_{m\subset\Lambda_N}$ and the number of models of dimension D is $\binom{N}{D}$. the sum $\sum_{m\in\mathcal{M}, D_m>0} 2^{-D_m}\binom{N}{D_m}^{-1}$ is less than 1, and by the inequality $\binom{N}{D} \le \left(\frac{en}{D}\right)^D$ for $0 < D \le N$ [Mas07, proposition 2.5], a suitable choice of weights for the application of Theorem 5.6 is

$$L_m = \log\left(\frac{2eN}{D_m}\right)$$

for which sufficient penalties satisfy the limiting condition:

$$\text{pen}\,(m) > \epsilon^2 D_m\left[1 + 2\sqrt{L_m} + 2L_m\right].$$

A proposition of L. Birgé and P. Massart in the same article than Theorem 5.6 shows that for large N this limiting condition is close to a minimal penalty level, in the context of free variable selection as just described, when the variables are independent, or in other words when $\langle\phi_\lambda,\, \phi_{\lambda'}\rangle = 0$ whenever $\lambda \neq \lambda'$:

Proposition 5.8. *[BM07, Proposition 2] Let s be the true unknown function to estimate and set $\Lambda_1 = \{\lambda \in \Lambda_N, \langle s, \phi_\lambda \rangle \neq 0\}$. Assume that there exist numbers δ, α, A and η with:*

$$0 \leq \delta < 1, \quad 0 \leq \alpha < 1, \quad A > 0, \quad \text{and} \quad 0 < \eta < 2(1-\alpha),$$

and some $\bar{m} \in \mathcal{M}$ with

$$|\Lambda_1| \leq \delta\,|\bar{m}|\,, \quad |\bar{m}| \leq AN^\alpha, \quad \text{and} \quad \text{pen}\,(\bar{m}) \leq (2-2\alpha-\eta)(1-\delta)\epsilon^2\,|\bar{m}|\log N.$$

Then one can find positive constants κ and N_0 depending on δ, α, A and η, such that

$$\mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \geq \kappa\epsilon^2\,|\bar{m}|\log N \quad \text{for all } N \geq N_0.$$

On one hand, as the authors point, if following Theorem 5.6, the penalty takes the form

$$\text{pen}\,(m) = (1+\eta)\epsilon^2 D_m\left(1+2\sqrt{L_m}+2L_m\right) \text{ with } \eta > 0,$$

then

$$\mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \leq C(\eta)\inf_{m\in\mathcal{M}}\left\{d^2(s,S_m)+\epsilon^2 D_m\left[1+\log\left(\frac{N}{D_m}\right)\right]\right\}.$$

In particular if the source signal s satisfies the assumptions of the proposition above with $|\Lambda_1| \geq 3$, then the source signal s belongs to the $|\Lambda_1|$-dimensional model $\text{Span}\,\{\phi_\lambda\}_{\lambda\in\Lambda_1}$, and

$$\mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \leq C(\eta)\epsilon^2\,|\Lambda_1|\log N.$$

On the other hand, by the proposition above, under the same assumptions, if

$$\text{pen}\,(\bar{m}) \leq (1-\alpha-\eta/2)(1-\delta)\,|\bar{m}|\,\epsilon^2 2\log N,$$

then

$$\mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \geq \kappa\epsilon^2\,|\bar{m}|\log N$$

when N is large enough. It follows that the estimator associated with some too small penalties has a risk much larger that one would get with larger penalty from Theorem 5.6, the ratio tending to infinity when both $\frac{N}{|\bar{m}|}$ and N tend to infinity. Comparing the thresholds in Theorem 5.6 and Proposition 5.8 with $m = \bar{m}$ and $|\bar{m}| \sim N^\alpha$ shows that

$$\text{pen}\,(m) = \epsilon^2 D_m\left[1+2\log\left(\frac{N}{D_m}\right)\right]$$

is the borderline formula for the penalty, at least when N is very large and D_m is of the order N^α with $0 < \alpha < 1$.

To our knowledge, Proposition 5.8 is the only result available for minimal penalties under high (binomial) complexity. One of its notable consequences is to explain why Mallow's C_p is not suitable in such situations.

The proof of Proposition 5.8 relies heavily on the assumption of independence between the explanatory variables $\{\phi_\lambda\}_{\lambda\in\Lambda_N}$ available for selection. The main goal of our work is to address situations where this assumption is lifted.

6 A motivating example: estimating a continuous signal under white noise with step-functions by recursive partitioning

The following results deals with the idealized situation of estimating a continuous signal under white noise with step-functions by recursive partitioning. It corroborates empirical evidence on minimal penalties. The model family has the relatively simple structure of the set of rooted b-ary trees, which allow us to introduce the mechanisms at hand with minimal apparatus, postponing technical preliminaries on partition models. We first describe the estimation procedure, then assess the complexity of the corresponding family, provide a sufficient penalty proposition and finally a minimal penalty proposition.

6.1 Estimation procedure

Consider a domain U in $\mathbb{R}^n$, that we call the *underlying space*, and μ Lebesgue's measure on U, with $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ the associated norm and scalar product. Assume for simplicity that $\mu(U) = 1$.

Consider a Brownian sheet B over U with the covariance structure:

$$\mathbb{E}\left[\int f \,\mathrm{d}\, B \int g \,\mathrm{d}\, B\right] = \int fg \,\mathrm{d}\,\mu = \langle f,\, g\rangle \;\text{ for } (f,g) \in \mathbb{L}_2(\mu)^2.$$

Assume one wants to recover a function $s \in \mathbb{L}_2(\mu)$ from the observations of the process indexed by $\langle Y,\, f\rangle = \int sf \,\mathrm{d}\,\mu + \epsilon \int f \,\mathrm{d}\, B$ for $f \in \mathbb{L}_2(\mu)$, with the least squares minimum contrast criterion: $\gamma(t) = \int (t^2 - 2tY) \,\mathrm{d}\,\mu = \|t\|^2 - 2\,\langle Y,\, t\rangle$.

Each model m of the family $\mathcal{M}$ consists of step-functions constant over the elements of some finite partition p_m of the underlying space U into parts of positive measure, so that:

$$\begin{aligned} S_m &= \mathrm{Span}\left\{ \mathbb{1}_{\{\tau\}}(\cdot) \right\}_{\tau \in p_m}, \\ U &= \bigsqcup_{\tau \in p_m} \tau \end{aligned}$$

where $\sqcup$ stands for disjoint union. As each linear space in $(S_m)_{m \in \mathcal{M}}$ corresponds to a unique partition p_m, we identify m and p_m and write:

$$\begin{aligned} S_m &= \mathrm{Span}\left\{ \mathbb{1}_{\{\tau\}}(\cdot) \right\}_{\tau \in m}, \\ U &= \bigsqcup_{\tau \in m} \tau. \end{aligned}$$

The least squares estimator $\hat{s}_m$ associated with a model m is the step function obtained by averaging the observed Y over each part of m:

$$\begin{aligned} \hat{s}_m &= \sum_{\tau \in m} \frac{\langle Y,\, \mathbb{1}_{\{\tau\}}\rangle}{\langle \mathbb{1}_{\{\tau\}},\, \mathbb{1}_{\{\tau\}}\rangle}\, \mathbb{1}_{\{\tau\}}, \\ &= \sum_{\tau \in m} \frac{\langle Y,\, \mathbb{1}_{\{\tau\}}\rangle}{\mu(\mathbb{1}_{\{\tau\}})}\, \mathbb{1}_{\{\tau\}}. \end{aligned}$$

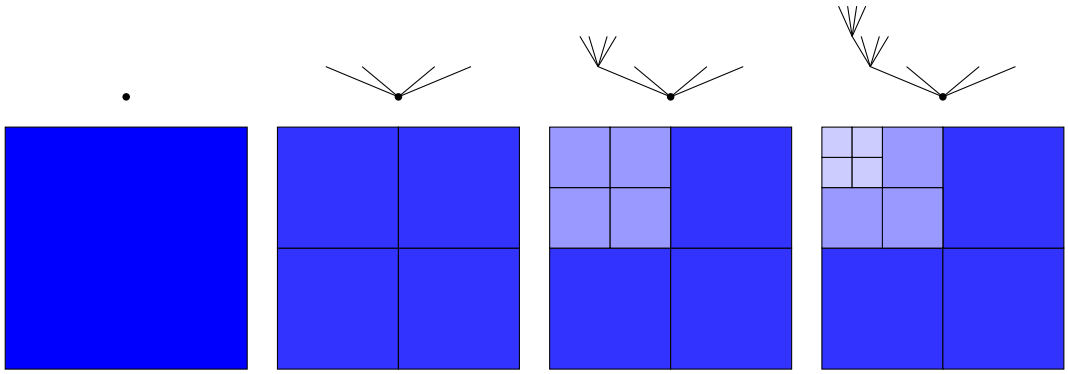


Figure 2 – Recursive partitioning and rooted plane trees

as described in more details in Section 7.1.

The model family is built by recursively partitioning the underlying space U as follows:

- root partition: the singleton partition $\{U\}$ corresponding to constant functions belongs to the family $\mathcal{M}$
- recursive partitioning: there is an integer $b > 0$ so that for any partition/model m and any part $\tau \in m$, there is a unique partition of τ in b subsets $\{\tau_1, \cdots, \tau_b\}$ of equal measures so that the partition $(m \setminus \tau) \cup \{\tau_1, \cdots, \tau_b\}$ indexes a model of $\mathcal{M}$. We say that any part of a partition in $\mathcal{M}$ may be split into b sub-parts of equal measure in a unique way.
- any model of $\mathcal{M}$ is obtained by a finite sequence of such operations starting from the root partition $\{U\}$.

Figure 2 offers an illustration with $U = [0,1]^2$ and $b = 4$. This procedure is described later as 'b-adic partitioning' in Definition 7.7.

Note that the corresponding model family $\mathcal{M}$ is infinite but countable, and is indexed by the set of finite rooted plane b-ary trees (see for example [FS09]).

6.2 Assessing the complexity of the model family

As the dimension of any model m in $\mathcal{M}$ is equal to the number of leaves of the corresponding tree, the number of models N_D of dimension D in $\mathcal{M}$ is the number of rooted b-ary trees with D leaves. This number satisfies [FS09, Example I.14 p. 68]

$$N_D = \begin{cases} \frac{1}{D}\binom{b\frac{D-1}{b-1}}{\frac{D-1}{b-1}} & \text{if } D-1 \equiv 0 \bmod b-1, \\ 0 & \text{otherwise.} \end{cases}$$

and by Stirling's relation $\log(k!) = k\log\left(\frac{k}{e}\right) + \frac{1}{2}\log(2\pi k) + o\left(\frac{1}{k}\right)$, the number N_D has the following behaviour in large D: when $D - 1 \equiv 0 \bmod b - 1$ and tends to ∞,

$$\log N_D = (1 + o(1))(D-1)\left[\frac{\log b}{b-1} + \log\left(\frac{b}{b-1}\right)\right].$$

This confirms that the considered model family is of exponential complexity.

In view of applying the Gaussian model selection Theorem 5.6 note that by Lemma B.2, the set of weights $\{L_m\}_{m\in\mathcal{M}}$ defined by

$$L_m = \frac{\log b}{b-1} + \log\left(\frac{b}{b-1}\right) \quad \forall m \in \mathcal{M},$$

satisfies

$$\sum_{m\in\mathcal{M}} e^{-|m|L_m} < 1.$$

6.3 Sufficient penalties

As expected, choosing a large enough level of penalties based on the complexity assessment above ensures that the model selection procedure behaves reasonably despite the infinite number of models as stated in Proposition 6.1 below which follows from Theorem 5.6. The case of a non constant Lipschitz source function is illustrative in the sense that, as it may not be a step function, such a function does not belong to any model of the family $\mathcal{M}$.

Proposition 6.1. *For some integer b and the model selection procedure described in the present section, choose a number η with $0 < \eta < 1$ and denote L_b the number $\frac{\log b}{b-1} + \log\left(\frac{b}{b-1}\right)$. Assume that*

$$\operatorname{pen}(m) = \epsilon^2 \frac{1+\eta}{1-\eta} |m| \left(1 + 2\sqrt{L_b} + 2L_b\right) \forall m \in \mathcal{M}. \tag{6.1}$$

Then a minimiser $\hat{m}$ of the penalized empirical contrast

$$\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ -\|\hat{s}_m\|^2 + \operatorname{pen}(m) \right\},$$

exists almost surely and the following relation holds:

$$\begin{aligned}\eta\,\mathbb{E}\left[\|s - \hat{s}_{\hat{m}}\|^2\right] \leq \inf_{m\in\mathcal{M}} \left\{ \mathrm{d}^2(s, S_m) + \epsilon^2 |m| \frac{1+\eta}{1-\eta} \left[\frac{2\eta}{1+\eta} + 2\sqrt{L_b} + 2L_b\right] \right\} \\ + \epsilon^2 \frac{1+3\eta}{\eta(1-\eta)}.\end{aligned} \tag{6.2}$$

Case of an exact model *Assume in addition that there is an exact model $\tilde{m} \in \mathcal{M}$ containing the source signal s, then the following holds:*

$$\mathbb{E}\left[\|s - \hat{s}_{\hat{m}}\|^2\right] \leq \epsilon^2 |\tilde{m}| \frac{1+\eta}{\eta(1-\eta)} \left[\frac{8}{\sqrt[4]{b}} + \frac{2\eta}{1+\eta}\right] + \epsilon^2 \frac{1+3\eta}{\eta^2(1-\eta)}. \tag{6.3}$$

Case of a Lipschitz unknown source function *Assume in addition that the unknown source function is L-Lipschitz, that $\mu(U) = 1$ and that there is a number $d \geq 1$ so that the diameter δ_τ of any part τ arising in the recursive partitioning process*

is bounded by $\delta_\tau \leq \delta_U \mu(\tau)^{\frac{1}{d}}$. *Then there is a constant* $C(b,d)$ *depending only on* b *and* d *so that the following relations hold:*

$$C(b,d) \leq 5\left(1+2\frac{b}{d}\right), \tag{6.4}$$

and

$$\eta\,\mathbb{E}\left[\|s-\hat{s}_{\hat{m}}\|^2\right] \leq C(b,d)\left(\frac{L\delta_U}{2}\right)^{2\frac{d}{d+2}}\epsilon^{2\frac{2}{d+2}}\left(\frac{1+\eta}{1-\eta}\right)^{\frac{2}{d+2}}+\epsilon^2\frac{1+3\eta}{\eta(1-\eta)}. \tag{6.5}$$

The proof is given in Section 11.2.

6.4 A geometrical property

After providing a sufficient penalties result for the model selection procedure at hand, we look for a minimal penalty result. This will rely on a seemingly unrelated property of the model family $\mathcal{M}$, presented in the following proposition as a completion rule. Informally, this states the following: consider the set (informally the grid) of the b^h parts of measure b^{-h} that may arise from h partitioning steps, then any subset of this set of parts may be completed into a model element of $\mathcal{M}$, model of - in a certain sense - moderate dimension, without smaller parts, as illustrated by Figure 3 .

Lemma 6.2 (B-ary completion rule)**.** *Consider an integer* $b>1$, *an integer* $h\geq 0$, $T_{b,h}$ *the maximal rooted* b*-ary tree of height* h, *and a subset* I *of the set formed by the* b^h *leaves of* $T_{b,h}$. *Then there is a rooted* b*-ary sub-tree of* $T_{b,h}$

- *in which all the elements of* I *are leaves,*
- *with not more than* $1+(b-1)h\,\mathrm{Card}(I)$ *leaves.*

The proof is given in Section 11.3

This geometrical rule relates with minimal penalties in the following way: consider an arbitrarily large value of the height h and the corresponding partition (grid) of U into b^h parts. The average values of the process B over each of these b^h parts form an i.i.d b^h-dimensional Gaussian sample. In the proof of the minimal penalty result in Proposition 6.4, we define a way to select large values among this sample. In large h, this yields, by the completion rule above, a noise-dependant model of moderate dimension but largely negative enough expected empirical contrast. This allows us in turn to explicit a minimal penalty level.

This method builds on the method of proof employed by the authors of Theorem 5.6 in the context of free selection among independent explanatory variables: among the available feature variables, select a certain number of those presenting the largest projected Gaussian noise to build a noise-driven adverse model. In free selection among independent variables, the linear span of any set of explanatory variables yields a valid model in $\mathcal{M}$, contrary to our case where we need to invoke the completion rule.

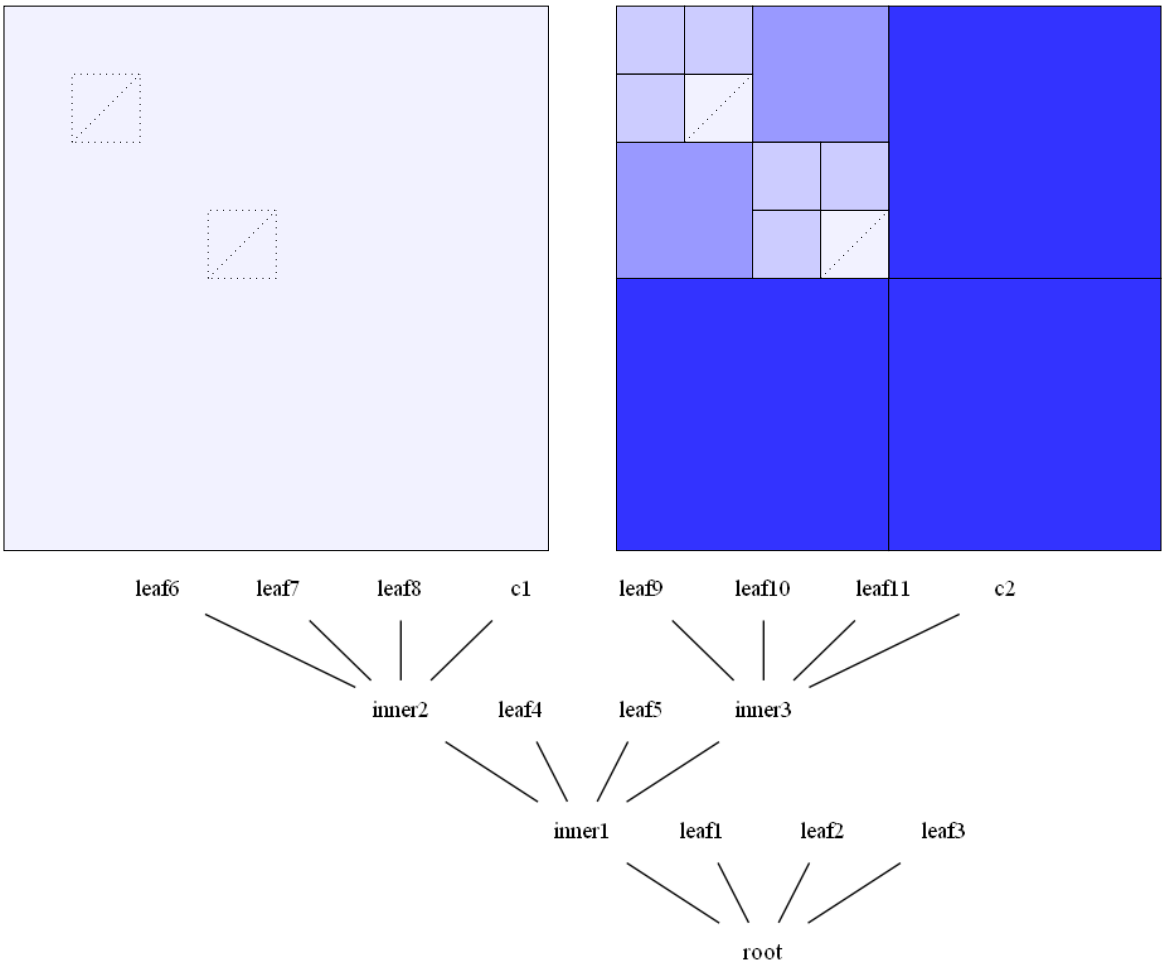


Figure 3 – Extending a rooted 4-ary tree to contain two leaves of height 3

6.5 Minimal penalties

The proposition to follow offers a threshold for minimal penalties. To address the limit case where the penalized empirical contrast $\gamma(\hat{s}_m)$ admits a lower bound over $\mathcal{M}$ but no minimizer, we rely on the following definition of an η-minimizer:

Definition 6.3. For $\eta > 0$ and the model selection procedure described in the present section, we call a model $m_\eta \in \mathcal{M}$ an *η-minimizer* of the penalized empirical contrast if $\forall m' \in \mathcal{M}$,

$$\gamma(\hat{s}_{m_\eta}) + \text{pen}\left(\hat{s}_{m_\eta}\right) \leq \gamma(\hat{s}_{m'}) + \text{pen}\left(\hat{s}_{m'}\right) + \eta.$$

Proposition 6.4. *Considering an integer $b \geq 2$ and the model selection procedure described in the present section with $\epsilon > 0$, assume there is a number $\rho < 1$ so that for any $m \in \mathcal{M}$,*

$$\text{pen}\left(m\right) < \epsilon^2 \left|m\right| \left(1 + 2\rho \frac{\log b}{b-1}\right).$$

Then

$$\mathbb{E}\left[\inf_{m \in \mathcal{M}} \left[\gamma(\hat{s}_m) + \text{pen}\left(m\right)\right]\right] = -\infty,$$

and for any $\eta > 0$, either

- *the penalized empirical contrast does not almost surely admit an η-minimiser,*
- *or such an η-minimizer $\hat{m}_\eta$ is defined almost surely, in which case $\mathbb{E}\left[\left\|\hat{s}_{\hat{m}_\eta} - s\right\|^2\right] = \infty$*

The proof is given in Section 11.4

It results from Proposition 6.4 that for the model selection procedure at hand, a minimal penalty level

$$\text{pen}\left(m\right) = \epsilon^2 \left|m\right| \left(1 + 2\frac{\log b}{b-1}\right)$$

exists above what Mallow's heuristic suggests: $\text{pen}\left(m\right) = \epsilon^2 \left|m\right|$ (see Section 1.5).

It is natural to investigate what happens when the integer b takes large values. In this situation, both the sufficient penalty and the minimal penalty get close to Mallow's penalty, the former with a gap $2\sqrt{\frac{\log(b)}{b-1}}\left[1 + o(1)\right]$, the later with a gap $2\frac{\log(b)}{b-1}\left[1 + o(1)\right]$ when b tends to ∞. The convergence to Mallows's heuristic could be attributed to a reduction in complexity when b grows. For instance, for $k \in \mathbb{N}$, any tree with branching factor b^k can be represented as a tree of branching factor b with only leaves of height $h \in k\mathbb{N}$ (informally "always take k partitioning turns together"). In the limit forms, the minimal penalty is much closer to Mallows's heuristic than to the sufficient penalty, making the situation difficult to interpret in terms of tightness of the result in Proposition 6.4. However, investigating this topic with a different method gives us a strong indication in favor of a minimal penalty level $\text{pen}\left(m\right) = \left|m\right|\left(1 + \Delta(b)\right) \forall m \in \mathcal{M}$, where $\Delta(b) = \left[1 + o(1)\right] 2\sqrt{\frac{\log(b)}{b}}$ when $b \to \infty$. For instance, $\Delta(2) = \frac{2}{\pi}$.

The key elements leading to Proposition 6.4 are on one hand the exponential complexity of the model family, and on the other hand the completion rule in Lemma 6.2.

We will try to apply a similar method to more practical situations with finite but large model families, such as the estimation of a discrete signal on a finite domain. Before doing so we recall some technical aspects of signal partitioning and segmentation.

7 Technical preliminaries on signal segmentation and partitioning

At this point we must introduce some definitions and notations regarding partitions of a feature space. In many estimation procedures, there is a natural geometrical structure on the set of feature variables available. Notably, in image partitioning, signal intensities are often recorded at regularly spaced positions in physical space. In the case of the segmentation of a sequential Gaussian signal, one observes an unknown signal s at times or positions $1, 2, \cdots, n$, $(n \geq 2)$ often with homoscedastic Gaussian errors, assuming that the source signal is exactly or approximately piecewise constant on the (unknown) intervals of some segmentation of $[1, n]$. The same geometrical situation appears when trying to recover a density from empirical densities on regularly spaced bins, as described in Section 9.1. In this case each model is a linear space of step functions based on a common set of breakpoints, and it is customary to index the models by the corresponding sets of breakpoints. When the signal's *support* (its *underlying space* or *feature space*) has a natural multidimensional geometry, like for instance with a 2-D image, it is more convenient to index the model family by partitions of this support. The parts of such a partition play the same role than the segments play in segmentation of a sequential signal.

Often the practical interest of working with segmentations, generally partitions in convex parts, is to ease computations and limit the complexity of the model families used in practice.

In the Gaussian model selection framework, a key assumption is the isonormal property of the noise components in the models to estimate. As a consequence, any orthogonal basis of the corresponding Euclidean space (e.g. $\mathbb{R}^n$) can play the role of the underlying space of the signal. This for instance is put to profit in signal estimation by wavelet component selection [AB96].

In the next section we formalize the relation between partitions of the signal's underlying space and the associated linear models. At the possible expense of readability, we treat the case of a discrete signal in the same framework than the case of a continuous signal.

7.1 Partitions

Consider some set U that we call *underlying space* equiped with a measure μ with $0 < \mu(U) < \infty$. Often U is a domain in $\mathbb{R}^n$ and μ Lebesgue's measure, or U is a finite space, often a rectangular bloc in $\mathbb{N}^d$, and μ the counting measure. To any set $\mathcal{M}$ of finite partitions of U into measurable sets of positive measure, we bijectively associate a set of linear models, describing each model as the linear span of the indicator functions of the *parts* of the associated partition. More precisely:

Notation 7.1. Denote $\mathcal{P}_U$ the set of finite partitions of U into measurable sets of positive measure.

Definition 7.2 (Partition models)**.** Considering a family $\mathcal{M} \subset \mathcal{P}_U$ of finite measurable partitions of U into sets of positive measure, for any partition $m \in \mathcal{M}$, define the linear model S_m as:

$$\begin{aligned} S_m &:= \operatorname{Span}\left(\{\mathbb{1}_\tau(\cdot)\}_{\tau\in m}\right), \\ &= \left\{\sum_{\tau\in m} \alpha_\tau \mathbb{1}_\tau(\cdot), \alpha \in \mathbb{R}^m\right\}. \end{aligned}$$

If for instance U is a set of cardinal n and μ the counting measure, for any $m \in \mathcal{M}$, the linear space S_m is associated to the linear regression model:

$$Y_i = \sum_{\tau\in m} \alpha_\tau \mathbb{1}_\tau(i) + \sigma\xi_i \text{ for } 1 \le i \le n \text{ with } \xi_1, \cdots, \xi_n \text{ i.i.d. } \mathcal{N}(0,1).$$

In what follows, when no ambiguity results, knowing a partition family $\mathcal{M}$ we will abuse the notation in designating the model family $\{S_m, m \in \mathcal{M}\}$ by the symbol $\mathcal{M}$, and in identifying as m either the corresponding partition, or the corresponding linear model S_m.

To express the projections over a model S_m it is convenient to define the means of a signal u over the parts in m:

Definition 7.3 (mean over a part)**.** For any $u \in \mathbb{L}_2(\mu)^2$ and τ a measurable subset of U of positive measure, define the *mean* $\bar{u}_\tau$ *of* u *over* τ as:

$$\bar{u}_\tau := \mu(\tau)^{-1} \int_\tau u \,\mathrm{d}\,\mu.$$

Or equivalently if μ is the counting measure on $\{1, \cdots, n\}$ and $u \in \mathbb{R}^n$, with the shorthand $|\tau| := \mu(\tau)$,

$$\bar{u}_\tau := |\tau|^{-1} \sum_{i\in\tau} u_i.$$

Then the projection s_m of s on S_m may be expressed as a function of the means of s over each part in m:

$$s_m = \sum_{\tau\in m} \bar{s}_\tau \mathbb{1}_\tau.$$

We denote Π_m the associated orthogonal projector of the n-dimensional Euclidean space:

$$\Pi_m : \left| \begin{array}{rcl} \mathbb{R}^n & \longrightarrow & \mathbb{R}^n \\ u & \longmapsto & \Pi_m(u) = \sum\limits_{\tau\in m} \bar{u}_\tau \mathbb{1}_\tau. \end{array} \right.$$

The associated minimum least squares estimator or *regressogram* for m may be written:

$$\hat{s}_m = \sum_{\tau\in m} \bar{Y}_\tau \mathbb{1}_\tau = \Pi_m Y.$$

The orthogonal projector Π_m may also be written:

$$\Pi_m = \sum_{\tau \in m} \frac{1}{|\tau|} \mathbb{1}_\tau \otimes \mathbb{1}_\tau,$$

making apparent the relation $\mathrm{Tr}\left[\Pi_m\right] = \mathrm{Card}(m)$. Finally the minimum least squares contrast is expressed as:

$$\begin{aligned} \gamma_n\left(\hat{s}_m\right) &= \|\hat{s}_m\|^2 - 2\left\langle \hat{s}_m, Y \right\rangle, \\ &= -\sum_{\tau \in m} |\tau| \, \bar{Y}_\tau^2, \\ &= -\|\Pi_m(Y)\|^2. \end{aligned} \tag{7.1}$$

As it will be discussed in the next section, in the context of density estimation when the contrast criterion is based on the log-likelihood, the same linear projections appear, the empirical measure taking the place of the Gaussian observed signal Y.

For the following, unless otherwise stated, the index set $\mathcal{M}$ of the model family is a subset of $\mathcal{P}_U$ for some underlying space U. The following shorthand is introduced for the sake of concision:

Notation 7.4. For any partition $m \in \mathcal{P}_n$ we denote $|m|$ the cardinal of m.

Finally the following definition will play a central role in this work, in connection with large components of the noise vector W:

Definition 7.5 (Isolate)**.** We say that a partition m *isolates* a subset $\tau \subset U$ if the subset τ is a part of the partition m. We say that a partition m *isolates* an element $x \in U$ if the singleton $\{x\}$ is a part of the partition m.

7.2 Recursive partitioning

In the following we provide definitions for recursive partitioning of a feature space U. Our aim is simply to formalise a current procedure with regression tree: building a family of partitions of some space U into parts by recursive subdivision of those parts according to a predefined rule. We describe such partitions with trees of split operations in view of getting enumeration informations, as trees are well studied objects.

Definition 7.6 (recursive partitioning)**.** We say that a set $\mathcal{M}$ of partitions of some set U arises from *recursive partitioning* of U if there is

- a finite or countable family C of measurable subsets of U each of positive measure, that we call the *parts* of $\mathcal{M}$
- for each part $\tau \in C$ a set P_τ of finite ordered partitions of τ into elements of C, partitions that we call the *admissible splits* of τ,

so that:

- $\{U\}$ is an element of $\mathcal{M}$,

- for any part $\tau \in C$ and any admissible split $s = (\nu_1, \cdots) \in P_\tau$, for any partition $m_1 = \{\tau, \tau_1, \cdots\} \in \mathcal{M}$, then $m_2 = \{\nu_1, \cdots\} \cup \{\tau_1, \cdots\} \in \mathcal{M}$. In this case we say that m_2 results from a *direct split* of the part τ in m_1 with the split $s = (\nu_1, \cdots)$.
- all the elements in $\mathcal{M}$ can be obtained from $\{U\}$ by this recursive process: for any m in $\mathcal{M}$, there is a finite sequence $(m_i)_{i=1}^k$ of elements of $\mathcal{M}$ with $m_1 = \{U\}$ and $m_k = m$ so that for any $i \in [1, k-1]$, m_{i+1} results of a direct split of a part τ_i in m_i with a split s_i.
- in this case we say that the rooted plane tree in which:
  - the root is U
  - the nodes are the elements of the set of parts $\cup_{i\in[1,k]} m_i$
  - the set of edges is the set of ordered pairs $\cup_{i\in[1,k-1]} \cup_{\nu\in s_i} (\tau_i, \nu)$
  - for $i \in [1, k-1]$, the list of children of the node τ_i is ordered as the admissible split s_i.

  is associated with the partition m.

It follows from this definition that if a set of partitions $\mathcal{M}$ arises from recursive partitioning, any partition $m \in \mathcal{M}$ is associated with at least one finite rooted plane tree. We call the couple $(C, \{P_\tau\}_{\tau\in C})$ a *partitioning rule* or *partitioning scheme*. If all the admissible splits have the same cardinal b, we call the partitioning scheme *b-ary*.

In many cases, several trees may be associated to a single partition under the same partitioning rule.

In b-ary partitioning, if all splits are composed of parts of equal measure, any part τ of height h has measure $\mu(\tau) = b^{-h}\mu(U)$.

The following provides a definition for the partitioning of a continuous domain in the geometric manner of an infinitely fine dyadic grid (see Section 6):

Definition 7.7 (b-adic partitioning)**.** Consider a set of partitions $\mathcal{M}$ arising from b-ary partitioning of a measurable domain U of $\mathbb{R}^n$ of positive Lesbegues's measure $\mu(U)$. We say that $\mathcal{M}$ arises from *b-adic partitioning* of U if any $\tau \in C$ has one and only one admissible split. In this case we say that a part τ is of *height* h if they is a sequence $\tau_0, \cdots, \tau_h$ of parts so that $\tau_0 = U$ and $\tau_h = \tau$, and so that for $0 \leq i < h$, τ_{i+1} is an element of an admissible split of τ_i.

In the case of b-adic partitioning, the relation between trees and partitions is bijective, and their set is countably infinite.

8 Minimal penalties for the partitioning of a discrete Gaussian signal

In this section we show that the minimal penalty phenomenon occurs in the context of segmentation or partitioning of a discrete Gaussian signal, for a large class of model selection methods by partition of a set of independent variables.

The proof relies on the fact that in this setup, if the partitions family $\mathcal{M}$ used for model selection is large enough in a certain sense, the atoms of the associated partitions family play a role analog to the role of individual feature variables in free independent variables selection. This will be expressed by the completion rule in Assumption 8.1.

Thanks to this assumption, we will need no explicit reference to the recursive partitioning methods introduced in the preceding section, even though they remain our primary target.

If additionally the model family is not too rich in low dimensional model, the odds are that - under too low penalties - the excess risk of the selection procedure will materialize in a high dimension selected model. This is the purpose of the binomial complexity Definition 8.4.

8.1 Definitions and assumptions on the model family structure

8.1.1 Completion rule

We will first state a structure assumption on the model family $\mathcal{M}$. Without mentioning any specific geometry of the signal's support nor any partitioning procedure, it intends to capture the consequences of the statement "to isolate any single point of the signal's support, it takes to create at most $\mathrm{C}_{\mathrm{ext}}$ new parts ". We view the number $\mathrm{C}_{\mathrm{ext}}$ as an inverse measure of the geometric diversity of the model family, in the sense of its proximity with the isotropic situation of free variable selection, where the equivalent of the number $\mathrm{C}_{\mathrm{ext}}$ is one.

Assumption 8.1 (completion rule)**.** *For a set $\mathcal{M}$ of linear models in an Euclidean space, and an orthonormal family $B = \{u_1, \cdots, u_n\}$ in the same Euclidean space, we say that Assumption 8.1 is satisfied if there is a integer $\mathrm{C}_{\mathrm{ext}}$ that we call the extension factor so that for any $m \in \mathcal{M}$ and any subset b of B, there is a model in $\mathcal{M}$ denoted $m \odot b$ satisfying*

$$\begin{aligned} &S_m \oplus \mathrm{Span}(b) \subset S_{m\odot b}, \\ &\quad dim(S_{m\odot b}) \leq dim(S_m) + \mathrm{C}_{\mathrm{ext}}\, Card(b). \end{aligned} \tag{8.1}$$

Remark 8.2. Assumption 8.1 is suited to situations where the models in $\mathcal{M}$ are made of step functions based on partitions of an underlying feature space represented by the basis B in the assumption, and are built by recursively splitting regions of this space in a predefined way, as for instance in a regression tree. This situation is formalized in the preceding section (see Section 7) where the underlying space U is the analog of the basis B in the present section by the identification $B \sim \{\mathbb{1}_x(\cdot)\}_{x\in U}$. The observed signal lies in the Euclidean space $L_2(U) \sim \mathrm{Span}(B)$, so that any two functions with disjoint supports are orthogonal, which is true in particular when these functions are indicator functions of disjoint parts of the underlying space U. A partition model (see Definition 7.2) is the linear span of the indicator functions of the parts of a given partition of U. It is equivalent to state that the indicator function $\mathbb{1}_{\{x\}}(\cdot)$ of a singleton $\{x\}$ belongs to a model m, or to say that the singleton $\{x\}$ is a part in the partition indexing the model m, in other words that m isolates x. If a model m does not isolate a point x, one could

want to modify it along the recursive partitioning procedure to obtain a finer partition indexing a larger model that does isolate the point x. This only requires to modify (split possibly several times) one single part in m, the part where the point to isolate belongs. Note that all segmentation methods do not enjoy this property: in Voronoi segmentation for instance, it is not straightforward to see how to split a part without disturbing its neighbours.

Typically on a d-dimensional underlying space U, with freely located splits across coordinates level sets of U, like for regression trees, the factor $\mathrm{C}_{\mathrm{ext}}$ can be chosen equal to $2d$, since it requires at most 2 splits per coordinate to isolate a point, producing at most one additional sub-part per split. This applies *a fortiori* when the partitioning rule allows any partition of the underlying space by hyperrectangles (hyperrectangle layout).(see Figure 4). The most simple instance of this procedure is free segmentation of an integer segment, where the factor $\mathrm{C}_{\mathrm{ext}}$ can be chosen equal to 2.

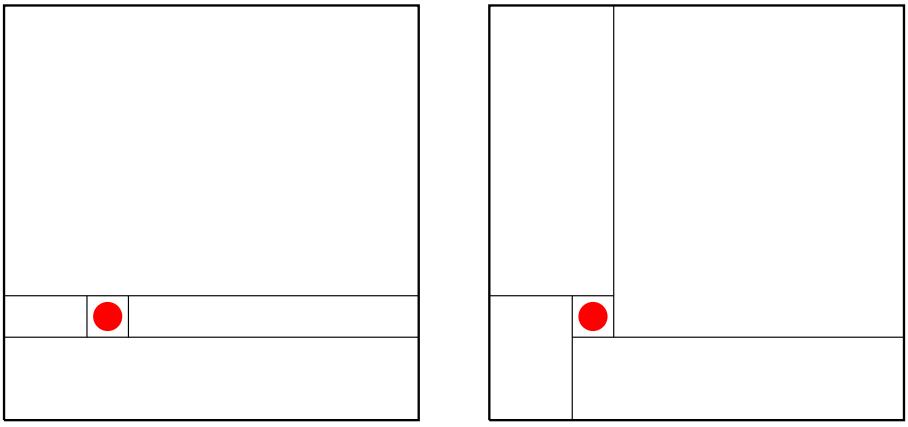


Figure 4 – Splitting method of a regression tree or hyperrectangle layout: $\mathrm{C}_{\mathrm{ext}} \leq 2d$

If the chosen method only allows to split a part into sub-parts of equal measure (up to rounding choices), like for instance with a dyadic grid or similar cases, the number $\mathrm{C}_{\mathrm{ext}}$ may be larger than in the previous examples, by a factor which is logarithmic in the dimensions of the underlying space.

Part 8.4 offers several instances of such model selection procedures, and Part B details their construction.

Remark 8.3. The choice of an integer for the extension factor does not involve any loss of generality. Indeed if the model family $\mathcal{M}$ satisfies a condition similar to Assumption 8.1 with the only difference that a real number c plays the role of the integer factor $\mathrm{C}_{\mathrm{ext}}$, for any $b = \{e_{i_1}, \ldots, e_{i_k}\} \subset B$ and any $m \in M$, as the dimensions are integers, the relation:

$$|m \odot \{e_{i_1}\}| \leq |m| + c$$

implies that

$$|m \odot \{e_{i_1}\}| \leq |m| + \lfloor c \rfloor$$

leading after $k = \operatorname{Card} b$ iterations to:

$$|(\ldots(m \odot \{e_{i_1}\}) \odot \ldots) \odot \{e_{i_k}\}| \leq |m| + \lfloor c \rfloor \operatorname{Card} b.$$

The model above is a valid completion of the model m by the sub-basis b, of dimension not more than $|m| + \lfloor c \rfloor \operatorname{Card} b$. Then the model family $\mathcal{M}$ satisfies Assumption 8.1 with extension factor $\mathrm{C}_{\mathrm{ext}} = \lfloor c \rfloor$

8.1.2 Binomial complexity

In numerous situations, the minimal penalty phenomenon is observable with a change in the dimension of the selected model when one gradually increases the penalty function by a scaling parameter (see Section 1.5). A closer look reveals that this phenomenon requires the model family $\mathcal{M}$ to be relatively poor in low dimensional models, in a certain sense. The following definition offers a quantification of this complexity property.

Definition 8.4 (Binomial complexity rate)**.** Consider a finite set $\mathcal{M}$ of linear models in an n-dimensional Euclidean space. We call *binomial complexity rate* of $\mathcal{M}$ the real number $\mathrm{B}_{\mathcal{M}}$ defined by:

$$\mathrm{B}_{\mathcal{M}} = \inf \left\{ \beta \geq 0, \sum_{m \in \mathcal{M}, |m|>0} 2^{-|m|} \left(\frac{en}{|m|} \right)^{-\beta |m|} \leq 1 \right\}.$$

Denote N_D the number of models of dimension D in $\mathcal{M}$. Choosing $\beta = \sup \left\{ \frac{\log N_D}{D \log\left(\frac{en}{D}\right)} \right\}_{0<D\leq n}$ yields:

$$\sum_{m \in \mathcal{M}, |m|>0} 2^{-|m|} \left(\frac{n}{|m|} \right)^{-\beta |m|} \leq \sum_{D=1}^{n} 2^{-|m|} \leq 1,$$

so that the following bound on the binomial complexity rate holds:

$$\mathrm{B}_{\mathcal{M}} \leq \sup \left\{ \frac{\log N_D}{D \log\left(\frac{en}{D}\right)} \right\}_{0<D\leq n}.$$

Although the definition above is arbitrary in its form, we chose it so that the binomial complexity rate $\mathrm{B}_{\mathcal{M}}$ is less than a few units for many standard practical cases. If the model family $\mathcal{M}$ has not more than one model per dimension, then $\mathrm{B}_{\mathcal{M}} = 0$ by the remark above. The binomial complexity rate is less than one for free variable selection or segmentation of a finite sequence: in the first case the number of models of dimension $D \leq n$ is $\binom{n}{D}$ and in the second case $\binom{n-1}{D-1}$. By the relation $\binom{n}{D} \leq \left(\frac{en}{D}\right)^D$ for $1 \leq D \leq n$ [Mas07, p.20], in both cases the sum in the definition of $\mathrm{B}_{\mathcal{M}}$ above is less than $\sum_{D=1}^{n} 2^{-D} < 1$ for $\beta = 1$. More generally, if the model family arises from a partitioning scheme of a finite hyperrectangle domain in $\mathbb{N}^d$ into hyperrectangles, its binomial complexity rate satisfies $\mathrm{B}_{\mathcal{M}} \leq \frac{\log 2d}{\log 2}$, as a direct consequence of the complexity assessment in Proposition A.1 for hyperrectangle layouts. This applies for instance to segmentation of a sequence as well as regression trees and other types of estimation by step functions constant over (hyper)rectangular blocs .

Of course the definition of the number $\mathrm{B}_{\mathcal{M}}$ intends to ensure that the weights

$$\left\{ L_m = \log 2 + \mathrm{B}_{\mathcal{M}} \log \left(\frac{en}{|m|} \right) \right\}_{m \in \mathcal{M}}$$

satisfy

$$\sum_{m \in \mathcal{M}, |m|>0} e^{-|m| L_m} \leq 1,$$

as required by several model selection results mentioned in this work (Theorems 5.6 and 9.7) and the proof of Proposition 8.8 in what follows.

8.1.3 Linking completion rule and binomial complexity

It is natural to ask what are the consequences of the completion rule in Assumption 8.1 on the complexity of a model family $\mathcal{M}$. A first observation is that the (unrealistic) model family formed with a single model of maximal dimension n obeys the completion rule with any constant $\mathrm{C}_{\mathrm{ext}} \geq 0$, despite its very low complexity. So we need to assume the family $\mathcal{M}$ contains at least one model of moderate dimension d_0, in general 0 or 1. The following shows that in this case, if $d_0 = 0$ or $d_0 \leq \lfloor \sqrt{n} \rfloor$, for a fixed extension factor $\mathrm{C}_{\mathrm{ext}}$ and moderately large n, the model family $\mathcal{M}$ has strictly positive binomial complexity, with a lower bound of the form $\frac{1}{\mathrm{C}_{\mathrm{ext}}} \left[1 - O\left(\frac{1}{\log(n)} \right) \right]$, with equality for certain types of model families.

Proposition 8.5 (Linking completion rule and binomial complexity)**.** *Consider a model family $\mathcal{M}$ in a n-dimensional Euclidean space and denote d_0 the minimum of the dimensions of the elements of $\mathcal{M}$. Assume that $d_0 < n$ and that $\mathcal{M}$ satisfies the completion rule in Assumption 8.1 with extension factor $\mathrm{C}_{\mathrm{ext}}$. Then the extension factor $\mathrm{C}_{\mathrm{ext}}$ satisfies:*

$$\mathrm{C}_{\mathrm{ext}} \geq 1.$$

If $d_0 = 0$ and additionally

$$n > \mathrm{C}_{\mathrm{ext}} 2^{\mathrm{C}_{\mathrm{ext}}},$$

then the binomial complexity rate $\mathrm{B}_{\mathcal{M}}$ introduced in Definition 8.4 satisfies:

$$\mathrm{B}_{\mathcal{M}} \geq \frac{1}{\mathrm{C}_{\mathrm{ext}}} \left[1 - \frac{\mathrm{C}_{\mathrm{ext}} \log(2) + 1}{\log(n) - \log(\mathrm{C}_{\mathrm{ext}}) + 1} \right] > 0.$$

If $d_0 > 0$ and additionally

$$d_0 \leq \lfloor \sqrt{n} \rfloor,$$

and

$$n > \left[e(1 + \mathrm{C}_{\mathrm{ext}}) 2^{1+\mathrm{C}_{\mathrm{ext}}} \right]^2,$$

then the binomial complexity rate $\mathrm{B}_{\mathcal{M}}$ satisfies:

$$\mathrm{B}_{\mathcal{M}} \geq \frac{1}{\mathrm{C}_{\mathrm{ext}} + \frac{d_0}{\lfloor \sqrt{n} \rfloor}} \left[1 - \frac{(1 + \mathrm{C}_{\mathrm{ext}}) \log(2) + 2}{\frac{1}{2} \log(n) - \log(1 + \mathrm{C}_{\mathrm{ext}}) + 1} \right] > 0.$$

Moreover, for any integers $c \geq 1$ and n multiple of c, there exists a model family $\mathcal{M}_{n,c}$ in $\mathbb{R}^n$, satisfying the completion rule with $\mathrm{C}_{\mathrm{ext}} = c$ and of binomial complexity rate satisfying:

$$\mathrm{B}_{\mathcal{M}_{n,c}} \leq \frac{1}{\mathrm{C}_{\mathrm{ext}}}.$$

The proof is given in section 11.5.

8.2 Sufficient penalties

Here again, choosing a large enough level of penalties based on the complexity rate in Definition 8.4 above ensures that the model selection procedure behaves reasonably whatever the number of models, as stated in the following sufficient penalty result, in direct application of Theorem 5.6:

Proposition 8.6. *Let $\mathcal{M}$ be a finite model family of binomial complexity rate $\mathrm{B}_{\mathcal{M}}$. Choose a number $\eta \in (0,1)$ and consider the set of weights $\left\{ L_m = \log 2 + \mathrm{B}_{\mathcal{M}} \log \frac{en}{|m|} \right\}_{m \in \mathcal{M}}$. Assume that*

$$\mathrm{pen}\,(m) = \epsilon^2 \,|m|\, \frac{1+\eta}{1-\eta} \left[1 + 2\sqrt{L_m} + 2L_m \right] \forall m \in \mathcal{M}.$$

Then the minimiser $\hat{m}$ of the penalized empirical contrast

$$\hat{m} = \underset{m \in \mathcal{M}}{\mathrm{argmin}} \left\{ -\|\hat{s}_m\|^2 + \mathrm{pen}\,(m) \right\},$$

satisfies

$$\begin{aligned} \eta\, \mathbb{E}\left[\|s - \tilde{s}\|^2 \right] &\leq \inf_{m \in \mathcal{M}} \left\{ \mathrm{d}^2\,(s, S_m) - \epsilon^2 \,|m|\, \frac{1+\eta}{1-\eta} \left[\frac{2\eta}{1+\eta} + 2\sqrt{L_m} + 2L_m \right] \right\} + \epsilon^2 \frac{(1+3\eta)}{\eta(1-\eta)}, \\ &\leq \inf_{m \in \mathcal{M}} \left\{ \mathrm{d}^2\,(s, S_m) + \epsilon^2 \,|m|\, \frac{1+\eta}{1-\eta} \left(5 + 3\mathrm{B}_{\mathcal{M}} \log \frac{en}{|m|} \right) \right\} + \epsilon^2 \frac{(1+3\eta)}{\eta(1-\eta)}. \end{aligned}$$

If in addition there is an exact model $\tilde{m} \in \mathcal{M}$ containing the source signal s, then the following relation holds:

$$\eta\, \mathbb{E}\left[\|s - \hat{s}_{\hat{m}}\|^2 \right] \leq \epsilon^2 \,|\tilde{m}|\, \frac{1+\eta}{1-\eta} \left(5 + 3\mathrm{B}_{\mathcal{M}} \log \frac{en}{|\tilde{m}|} \right) + \epsilon^2 \frac{(1+3\eta)}{\eta(1-\eta)}.$$

The proof is given in Section 11.6.

8.3 Minimal penalties

The following results extends the minimal penalty result of L. Birgé and P. Massart (Theorem 5.8) to Gaussian model selection in a larger context than free variable selection. Proposition 8.7 provides a risk bound and Proposition 8.8 a bound on the dimension of the selected model. Both rely on a common set of assumptions and address different facets of the same phenomenon.

8.3.1 A risk lower bound

Note that the function $(n, l, \theta) \mapsto r(n, l, \theta)$ is defined in Lemma 15.3 and goes to 0^+ when $\frac{n}{l(1+\theta)}$ goes to ∞.

Proposition 8.7 (Quadratic risk lower bound). *Let $\mathcal{M}$ be a finite model family following the completion rule in Assumption 8.1 for some integer $\mathrm{C}_{\mathrm{ext}}$. Consider a number $t > 0$ and set $\theta = 1 + 2\log(2) + 2t$. Assume there is a number ρ with*

$$\rho \mathrm{C}_{\mathrm{ext}} < 1,$$

an integer l with

$$1 \leq l \leq \frac{e^{-\frac{1}{2}}}{(1+\theta)} n,$$

and a function $f : \mathbb{N} \to \mathbb{R}^+$, so that the penalty function is defined $\forall m \in \mathcal{M}$ by $\mathrm{pen}\,(m) = f(|m|)$ and satisfies both conditions:

$$\frac{f(|m|)}{|m|} \text{ is non increasing with } |m|\,, \tag{8.2}$$

$$0 \leq f(l) \leq \rho\epsilon^2 l\,[1 - r(n,l,\theta)\,]\left[1 + 2\log\frac{n}{(1+\theta)l}\right]. \tag{8.3}$$

Then if $\hat{s}_{\hat{m}}$ is a minimizer of the penalized contrast criterion $-\|\hat{s}_m\|^2 + \mathrm{pen}\,(m)$ over $m \in \mathcal{M}$, the following risk lower bound holds apart from an event of probability less than e^{-t}:

$$\|s - \hat{s}_{\hat{m}}\|^2 \geq \epsilon^2 C(\rho\mathrm{C}_{\mathrm{ext}})\,[1 - r(n,l,\theta)\,]\,l\left(1 + 2\log\frac{n}{(1+\theta)l}\right), \tag{8.4}$$

where

$$C(\rho\mathrm{C}_{\mathrm{ext}}) = \begin{cases} \frac{(1-\rho\mathrm{C}_{\mathrm{ext}})^2}{4} & \text{if } s \neq 0, \\ 1 - \rho\mathrm{C}_{\mathrm{ext}} & \text{if } s = 0. \end{cases}$$

The proof is given in Section 11.7.

On one hand, Proposition 8.6 shows that for any η with $0 < \eta < 1$ a penalty function of the form $\forall m \in \mathcal{M}$:

$$\mathrm{pen}\,(m) = \epsilon^2\,|m|\,\frac{1+\eta}{1-\eta}\left[1 + 2\sqrt{\log(2) + \mathrm{B}_{\mathcal{M}}\log\frac{en}{|m|}} + 2\left(\log(2) + \mathrm{B}_{\mathcal{M}}\log\frac{en}{|m|}\right)\right],$$

warrants a reasonable selected model, where the binomial complexity rate $\mathrm{B}_{\mathcal{M}}$ is a constant of a few units in many practical cases (see Definition 8.4 and its following discussion), with an asymptotic lower bound in $\frac{1}{\mathrm{C}_{\mathrm{ext}}}$ in large n, which is tight for certain classes of model families (see Proposition 8.5). This provides us with the reference asymptotic form in large $\frac{n}{|m|}$ for a lower bound on sufficient penalties obtained by the general Gaussian model selection Theorem 5.6:

$$\mathrm{pen}\,(m) > [1 + o(1)]\,\frac{1+\eta}{1-\eta}\epsilon^2\,|m|\,\mathrm{B}_{\mathcal{M}} 2\log\frac{n}{|m|}, \tag{8.5}$$

$$\text{with } \mathrm{B}_{\mathcal{M}} \gtrsim \frac{1}{\mathrm{C}_{\mathrm{ext}}}.$$

On the other hand, in the point of view of Proposition 8.7 above, assuming ρ and $\mathrm{C}_{\mathrm{ext}}$ satisfy $\rho\mathrm{C}_{\mathrm{ext}} < 1$, consider large values of n and for instance $l \sim n^\alpha$ for some

$0 < \alpha < 1$. If the penalty function is moderate enough to meet the conditions 8.2 and 8.3 of Proposition 8.7, then by Inequality 8.4 for n large enough the risk of the corresponding selection procedure can be arbitrarily bounded away from 0 with strong probability whatever $\|s\|^2$. In this case the resulting estimator is drastically worse than with sufficient penalties. In that sense the limiting condition $\rho \mathrm{C}_{\mathrm{ext}} < 1$ is essential. The reference limit form for insufficient penalties obtained by Proposition 8.7 is

$$\mathrm{pen}\,(l) < \epsilon^2\,|l|\,\frac{1}{\mathrm{C}_{\mathrm{ext}}} 2\log\frac{n}{l} \tag{8.6}$$

for some large l and $\frac{n}{l}$, a form to be compared with Inequality 8.5. Within our understanding of the relation between the extension factor $\mathrm{C}_{\mathrm{ext}}$ and the binomial complexity rate $\mathrm{B}_{\mathcal{M}}$ (see Proposition 8.5), this is a hint that the result in Proposition 8.6 is in a certain sense tight for the class of model families following the completion rule in Assumption 8.1.

In this respect, the constant $\mathrm{C}_{\mathrm{ext}}$ acts as an inverse measure of the approximation ability of the model family $\mathcal{M}$, the most favorable situation (at constant n) being found with free selection among independent variables, which offers complete isotropy in $\mathbb{R}^n$. The extension rule covers many practical situation where the model family is generated by a recursive partitioning scheme of a feature space, often an (hyper)rectangle in $\mathbb{N}^d$, like with regression trees, segmentation of a sequence or many image partitioning schemes. In this case, as formalized in Section 7, each model is indexed by such a partition and consists of step functions constant over the elements this partition. As mentioned, partitioning schemes with freely located (possibly multiple) splits along coordinates lead to values of the factor $\mathrm{C}_{\mathrm{ext}}$ close to a few unit, for instance 2 for segmentation of a sequence. To the contrary, situations where the splits are only placed to central locations (in the manner of a dyadic grid) impose an additional $\log(n)$ factor on the number $\mathrm{C}_{\mathrm{ext}}$, leading to potentially unrealistic numerical values in Proposition 8.7.

Note also that when n is sufficiently large, the term

$$\rho\,[1 - r(n, l, \theta)]\left[1 + 2\log\frac{n}{(1+\theta)l}\right]$$

can be arbitrarily large even for small ρ, so that Proposition 8.7 covers cases where the minimal penalty lies above Mallow's heuristic $\mathrm{pen}\,(m) = \epsilon^2\,|m|$ (see section 1.5).

We believe Proposition 8.7 covers many practical situations. It is inspired by the result of L. Birgé and P. Massart's for independent variables selection in [BM07]. To our knowledge at the time of writing, this extension to more intricate setups in exponential complexity is original.

8.3.2 A dimension lower bound

The following result shows, at least when the source signal is null, that lack of a strong enough penalty term, the selection procedure will very likely choose a model of excessive dimension. In such a situation, under-penalization is apparent, which can be of practical importance when the magnitude of the noise factor ϵ^2 is unknown.

As mentioned, such a phenomenon appears when the model family is sparse enough to likely prevent models of very low dimension to provide a best fit, but rich of enough

approximation ability to overcome the penalty term in high dimension. The former property is quantified by the binomial complexity rate defined above, and the later is again provided by the completion rule defined in Assumption 8.1.

The following proposition keeps the definitions and the assumptions of Proposition 8.7, in the null hypothesis $s = 0$.

Recall that the function $(n, l, \theta) \mapsto r(n, l, \theta)$ is defined in Lemma 15.3 and goes to 0^+ when $\frac{n}{l(1+\theta)}$ goes to ∞.

Proposition 8.8. *Assume the model family $\mathcal{M}$ follows the completion rule in Assumption 8.1 for some integer $\mathrm{C}_{\mathrm{ext}}$, denote $\mathrm{B}_{\mathcal{M}}$ its binomial complexity rate in the sense of Definition 8.4. Assume the source signal s is null.*

Consider an number $t > 0$ and set $\theta = 1 + 2\log(2) + 2t$. Assume there is a number ρ with

$$\rho \mathrm{C}_{\mathrm{ext}} < 1,$$

an integer l with

$$1 \leq l \leq \frac{e^{-\frac{1}{2}}}{(1+\theta)} n,$$

and a function $f : \mathbb{N} \to \mathbb{R}^+$, so that the penalty function is defined $\forall m \in \mathcal{M}$ by $\mathrm{pen}\,(m) = f(|m|)$ and satisfies both conditions:

$$\frac{f(|m|)}{|m|} \text{ is non increasing with } |m|\,, \tag{8.7}$$

$$0 \leq f(l) \leq \rho\epsilon^2 l\,[1 - r(n, l, \theta)] \left[1 + 2\log\frac{n}{(1+\theta)l}\right]. \tag{8.8}$$

Then apart of an event of probability less than $2e^{-t}$, the following bound holds:

$$|\hat{m}| \geq l\frac{2}{3}(1 - \rho\mathrm{C}_{\mathrm{ext}})\,[1 - r(n, l, \theta)]\,\frac{\log\left(\frac{n}{l(1+\theta)}\right)}{\mathrm{B}_{\mathcal{M}}\log(en) + 2} - 3t, \tag{8.9}$$

where $\hat{s}_{\hat{m}}$ is a minimizer of the penalized contrast criterion $-\|\hat{s}_m\|^2 + \mathrm{pen}\,(m)$ over $m \in \mathcal{M}$ and $\mathrm{B}_{\mathcal{M}}$.

The prof is given in section 11.8

As anticipated, Inequality 8.9 in Proposition 8.8 above shows that the dimension jump phenomenon appears as a combination of several driving factors. On the favorable side the existence of a large under-penalized dimension, quantified by the integer l, and the approximation properties of the model family, inversely quantified by $\mathrm{C}_{\mathrm{ext}}$. On the unfavorable side, the strength of the penalization, quantified by ρ, and the family's complexity, quantified by the binomial complexity rate $\mathrm{B}_{\mathcal{M}}$.

Taking for instance $l = \left\lfloor \frac{n^\alpha}{e(1+\theta)} \right\rfloor$ for some α with $0 < \alpha < 1$ and some large fixed θ, then if $\mathrm{B}_{\mathcal{M}} > 0$, which is customary as discussed above (*see* Poposition 8.5), with strong probability the following bound holds when $n \to \infty$

$$|\hat{m}| \geq (1 - \alpha)n^\alpha \frac{2}{3e(1+\theta)} \frac{1 - \rho\mathrm{C}_{\mathrm{ext}}}{\mathrm{B}_{\mathcal{M}}}\,[1 - o(1)]\,. \tag{8.10}$$

As also anticipated, the limit condition $\rho C_{\text{ext}} < 1$ found and discussed with Proposition 8.7 appears again, and the same consideration on the limit form of the penalty function apply.

To the contrary, the binomial complexity rate $B_{\mathcal{M}}$ was not appearing in the risk analysis in Proposition 8.7. As mentioned, in presence of a high binomial complexity rate, low dimensional models could be numerous enough to likely provide an erroneous penalized best fit, leading to a high risk of the procedure despite a low dimensional selected model. Recall that by Proposition 8.5, the number $B_{\mathcal{M}}$ has an asymptotic lower bound $\frac{1}{C_{\text{ext}}}$ in large n for model families with at least one small model, but is not prevented to take very large values. However the results in Appendices A and B show that in many cases like regression tree model families or simply bloc-layout regressograms, the number $B_{\mathcal{M}}$ has moderate values. Notably its value is less than $\frac{\log(2d)}{\log(2)}$ for the family indexed by all bloc layouts of a finite (hyper)rectangle in $\mathbb{N}^d$ (see Proposition A.1), and we guess this value could be improved towards 1.

To our knowledge at the time of writing, the result in Proposition 8.8 is also original.

8.4 Numerical experiments for section 8

8.4.1 Introduction and motivation

This section is devoted to numerical illustrations of the minimal penalty phenomenon. In the example proposed, the model selection task consist in de-noising an image by mean of regression tree techniques. Although we do not specifically advocate this method in practical image processing, working on images provides a straightforward perception of the behavior of the selection procedures even in large dimension. Moreover, regression trees satisfy the completion rule in Assumption 8.1 with moderate values of the parameter C_{ext}, and come with algorithms of minimal complexity.

The purpose of the examples is to observe the behavior and the respective contrast estimates of:

- the minimizer of the empirical contrast at specific model dimensions,
- the sequence of the adverse models produced by isolating the peeks of the noise, starting from an estimate of an oracle model,
- the corresponding sufficient penalties proposed in Massart's Theorem 5.6 and its application in Proposition 8.6,
- the minimal penalties proposed in Propositions 8.7 and 8.8.

8.4.2 Description

The followings describes how the numerical experiments were conducted.

Experience plan Each experiment was conducted as follows:

- Choose a gray-level *source image* represented by a vector in $\mathbb{R}^{n=h\times w}$, playing the role of an unknown source signal to reconstruct.

- Choose a partitioning rule and the corresponding family of linear models $\mathcal{M}$ (see Section 7).

- Choose a set of weights on this family suitable for sufficient penalties in application of Proposition 8.6

- Choose a value for the variance of an i.i.d. Gaussian vector of the same dimension than the source image.

- Determine a sequence of partition models of increasing dimension, each approaching the minimizer of the quadratic risk (as in Definition 5.4) among the models of the same dimension. We refer to this sequence as the *segmentation path* of the source image.

- From this sequence, choose a model approaching the minimizer of the quadratic risk among the family $\mathcal{M}$. We refer to this model as the *near oracle*.

- Draw an i.i.d. Gaussian vector (the noise) of the chosen variance, and add it to the source image to produce a *noisy image* playing the role of the observed signal.

- Determine a sequence of partition models of increasing dimension, each approaching the minimizer of the empirical contrast (as in Definition 5.3) among the models of the same dimension. We refer to this sequence as the *segmentation path* of the noisy image.

- Determine the corresponding sequence of weights and sufficient penalties. We retained slightly smaller values which amount to set $\eta = 0$ in in Proposition 8.6.

- Build a sequence of models starting from the near-oracle by sequentially isolating the largest locations of the noise according to the partitioning rule in Assumption 8.1. Determine the sequence of the corresponding empirical contrasts. We refer this sequence of models as *adverse path* or *oversegmented path*.

- Determine a corresponding sequence of minimal penalties as proposed in Propositions 8.7 and 8.8 and. We retained slightly larger values amounting to set $t = 0$ and $\rho = 1$ in this propositions.

- Check that the empirical contrast penalized with the minimal penalty proposition is decreasing up to a dimension much larger than the dimension of the oracle, indicating a poorly selected model and an insufficient penalization.

- Check the contrast spread between the empirical minimizer sequence and the adverse model sequence, to assert if the construction of the adverse model was a sound choice.

Algorithms We relied on the CART partitioning algorithm. Classification and regression trees (CART) where proposed by Breiman, Friedman Olshen and Stone in [BFOS84]. They are now extensively used in the practice of statistics. In [GN05], S. Gey and É. Nedelec offer, with a description of the algorithms involved, an analysis of their performance. In regression tree, a model families consists in step functions constant over the elements of some partition of the observed signal's underlying space, as in Definition 7.2.

The CART algorithm consists in two phases: growing and pruning. In the growing phase, a partition is recursively refined by choosing for each part the split offering the best contrast gain, taking no account of the contribution of possible subsequent splits. A variety of choices exists to administer the order of the parts chosen and the depth of the tree. Ideally the algorithm will consider a final partition tree of maximal ramification, with only singleton leafs. In practice various stopping rules are employed.

In the pruning phase, the same tree is sequentially reduced by suppressing sub-trees based on an impurity criterion. Finally the procedure yields a finite sequence of nested partitions or equivalently trees, representing a sequence of nested models of increasing dimension. Each model in this sequence is a proxy for the minimizer of the empirical contrast among the models of the same dimension in the family considered. Last, one may apply a penalized contrast procedure to this sequence to retain a final selected model.

By their sequential nature, CART algorithms in large dimension only explore a fraction of the available models, and are as such - at least theoretically - sub-optimal in minimizing the empirical contrast among a model family. However this is not a preoccupation for the present work, as our purpose is, on the contrary, to offer upper bounds on the empirical contrast. In other words, we expect the CART contrast proxy to be somewhat flatter than the true empirical contrast. Then when in an experience the CART proxy overcomes a certain penalty function, the odds are the true empirical contrast will also overcome that same penalty function. Moreover regression tree algorithms offer good performance in many practical situations. Based on this reduced effective complexity, S. Gey and É. Nedelec [GN05] established that sufficient penalties in the CART procedure may be less than the form in Theorem 5.6. As our main focus is the generic minimal penalty phenomenon, and not specifically the CART algorithm, we kept the form of Theorem 5.6 in the examples.

The algorithms for the examples where coded in Python. We tested several methods, ad-hoc or publicly available. We did not notice any significant discrepancy in the results obtained, except for speed of execution and with a single caveat: the well known necessity to allow the growing phase to go deeper by several levels of sub-partition than the largest dimension desired for the analysis.

Types of regression trees The numerical experiments are based on four different partitioning rules: free binary tree (regression tree), regular binary tree, free quad tree, regular quad tree, described in detail in part B. We chose them only in view of the possibility to illustrate the minimal penalty phenomenon in different configurations, and not particularly to advocate their use in practice.

theme	image credit	tag	partitioning rule	near oracle dimension($\simeq$)	section
squares		sqs	free binary	300	8.5.1
squares		sqs	free quad	300	8.5.2
squares		sqs	regular binary	300	8.5.3
squares		sqs	regular quad	300	8.5.4
face	[Ima]	face	free binary	300	8.5.5
face		face	free binary	1000	8.5.6
face		face	free binary	3000	8.5.7
face		face	free binary	5000	8.5.8
tree	[Pho13]	lontr	free binary	300	8.5.9
tree		lontr	free binary	1000	8.5.10
tree		lontr	free binary	3000	8.5.11
tree		lontr	free binary	5000	8.5.12

Table 3 – List of numerical experiments for Section 8

Choice of the source images The source images proposed for the examples are of size amenable to computation on a laptop computer, but large enough to be deemed high dimensional. Proposition 8.8 assumes a null source signal, which correspond to a uniformly grey source image. This is not the case in the examples. The conclusions still appear robust to this liberty. As regression trees tend to favor horizontal and vertical boundaries, we try to present examples with a dominance of such lines, and some others with features in many different orientations.

Choice of the weights for the sufficient penalties The sufficient penalties are taken from Proposition 8.6 and the binomial complexity rate assessments in Appendix B (see Table 16). This choice is mostly a matter of taste, as the sufficient penalties are presented only for convenience.

Signal over noise ratio We tried to present versions of the same examples with different signal over noise ratio, leading to different dimensions of the near oracle.

Experiments presented For reading convenience the numerical results are set out separately in Section 8.5. Table 3 lists the numerical experiment results presented.

8.4.3 Discussion

The experiment yield the anticipated result, in the sense that all cases show, on a large dimension range, a negative slope of the empirical contrast penalized with minimal penalties based on Propositions 8.7 and 8.8, and hence a selected model much larger in dimension than the oracle. We conducted a number of other experiments with various source images and noise levels, confirming the same phenomenon.

As mentioned, Proposition 8.8 makes the assumption that the source signal is null, which is not the case in the experiments. Empirically its conclusion seem to remain valid

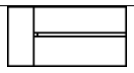
beyond this assumption, for source images of various level of complexity and different noise levels.

The proposed minimal penalty covers a significant part of the contrast slope gap between the noisy image and the source image with only two partition methods: free binary tree and free quad-tree. With the two other methods tested: regular binary tree and regular quad-tree, the minimal penalty estimate of Propositions 8.7 and 8.8 largely understates the suggestion from the experiments. We attribute this to the following: the proof of these propositions requires an estimate ($\mathrm{C_{ext}}$) of the largest dimension increase needed to sub-partition a model enough to isolate an individual point (pixel) of the source image. The theoretical minimal penalty estimate is inversely proportional to the extension factor $\mathrm{C_{ext}}$. This factor is fairly low for the free binary and free quad methods, where the boundaries can be freely located, in the manner of a classical regression tree. The two other methods work in the manner of a dyadic grid, with only centrally located splits. This inflates the number $\mathrm{C_{ext}}$ by an additional logarithmic factor in the image dimension. In practice, this factor depends on the dimension of the part one chooses to split, and not on the dimension of the entire image. Moreover in the proofs as in the experiments, the construction of the sequence of adverse (oversegmented) models relies on the sorted squared noise components. This gives the character of a random draw without replacement to the distribution of the sequence of points to isolate. In this context, for a model dimension D of the adverse model sequence, the effective logarithmic factor in $\mathrm{C_{ext}}$ is in some sense closer to the average depth $\sim \log\left(\frac{n}{D}\right)$ than to the initial depth $\sim \log(n)$. Taking this effect into account could plausibly allow to get the theoretical minimal penalty closer to the observed one, within the same general method of proof. Another plausible direction would be to quantify the model family's complexity with an exponential growth rate like in Section 6 instead of a binomial complexity rate.

8.5 Numerical experiments for section 8: results

8.5.1 Image squares, free binary tree

dimensions				std. dev.	
size	height	width	near oracle	source	noise
$4,194,304$	2048	2048	149	0.28	0.35

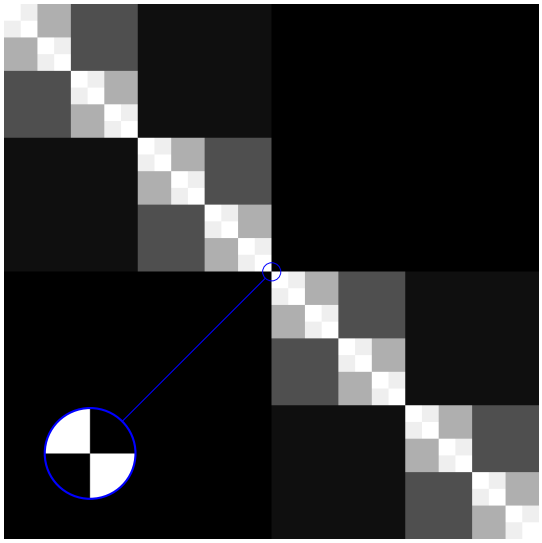
splitting		maximum dimension increase	
mode	max branching factor	to isolate a point (in this image size)	per split
free binary tree 	$b = 2$	$C_{ext} = 4$	$b - 1 = 1$

algorithm		
growing method	contrast criterion	max. dim.
impurity	l2	10000

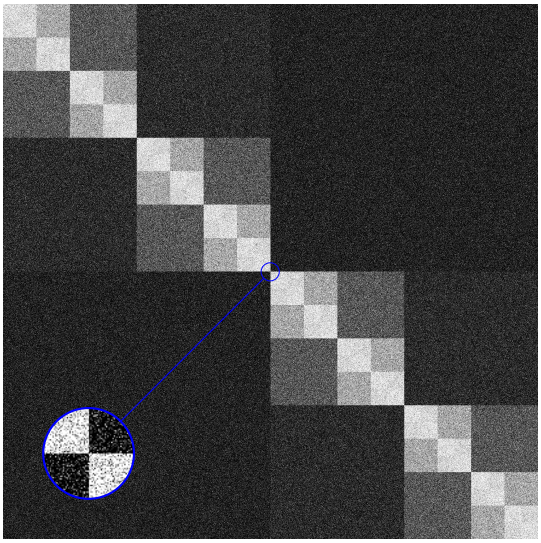
sufficient penalties				
type	expression	L_m	$B_{\mathcal{M}}$	θ
binom	$\epsilon^2 \lvert m\rvert \left[1 + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1} L_m\right]$	$\log(2) + B_{\mathcal{M}} \log(\frac{en}{\lvert m\rvert})$	2	1.00

minimal penalties (limit)
expression
$\epsilon^2 \frac{\lvert m\rvert}{C_{ext}} \left[2\log(\frac{0.48n}{\lvert m\rvert}) - \log(2\log(\frac{0.48n}{\lvert m\rvert})) - \frac{\log(\lvert m\rvert)+2}{\lvert m\rvert}\right]$

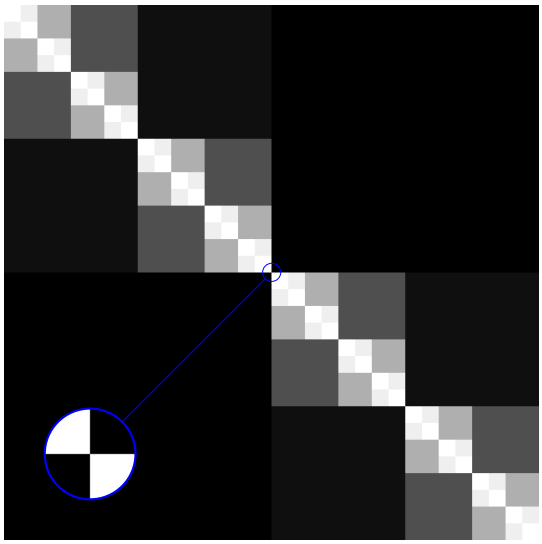
Table 4 – Data for image squares example squares.2fold.(2048, 2048)



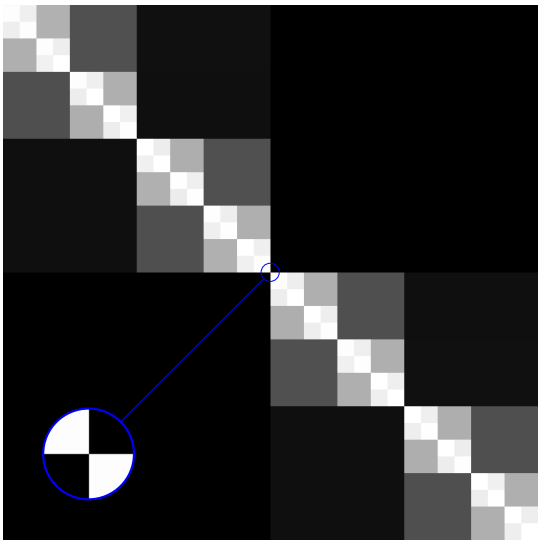
(a) Source image



(b) Noisy image



(c) Source image projected on near oracle of dimension 149



(d) Noisy image projected on near oracle of dimension 149

Figure 5 – Image sqs300binom2fisourceimage.png

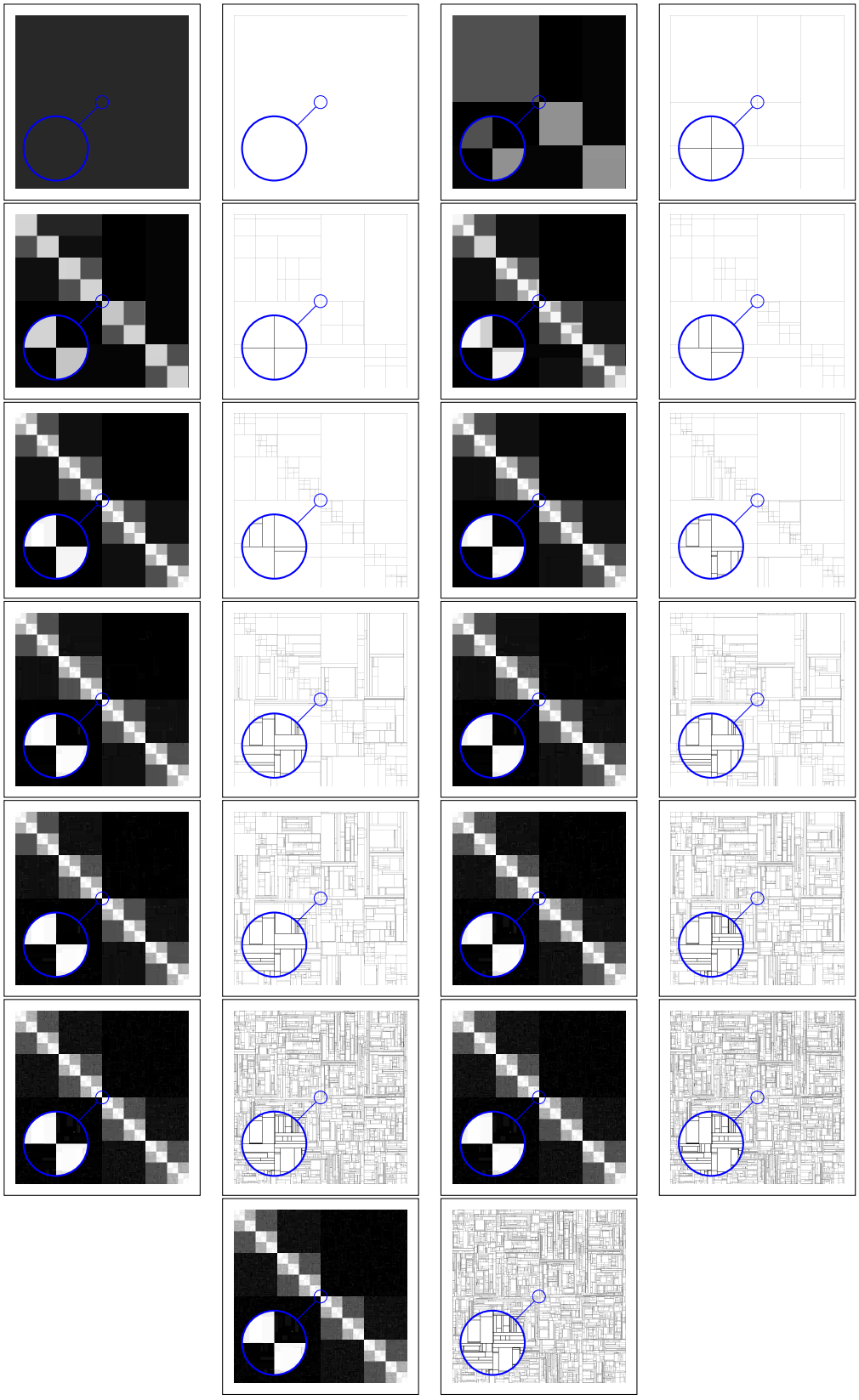


Figure 6 – Segmentation path of noisy image, at dimensions [1, 10, 32, 64, 128, 251, 508, 1008, 2048, 4071, 8192, 9985, 9993](estimates and borders)

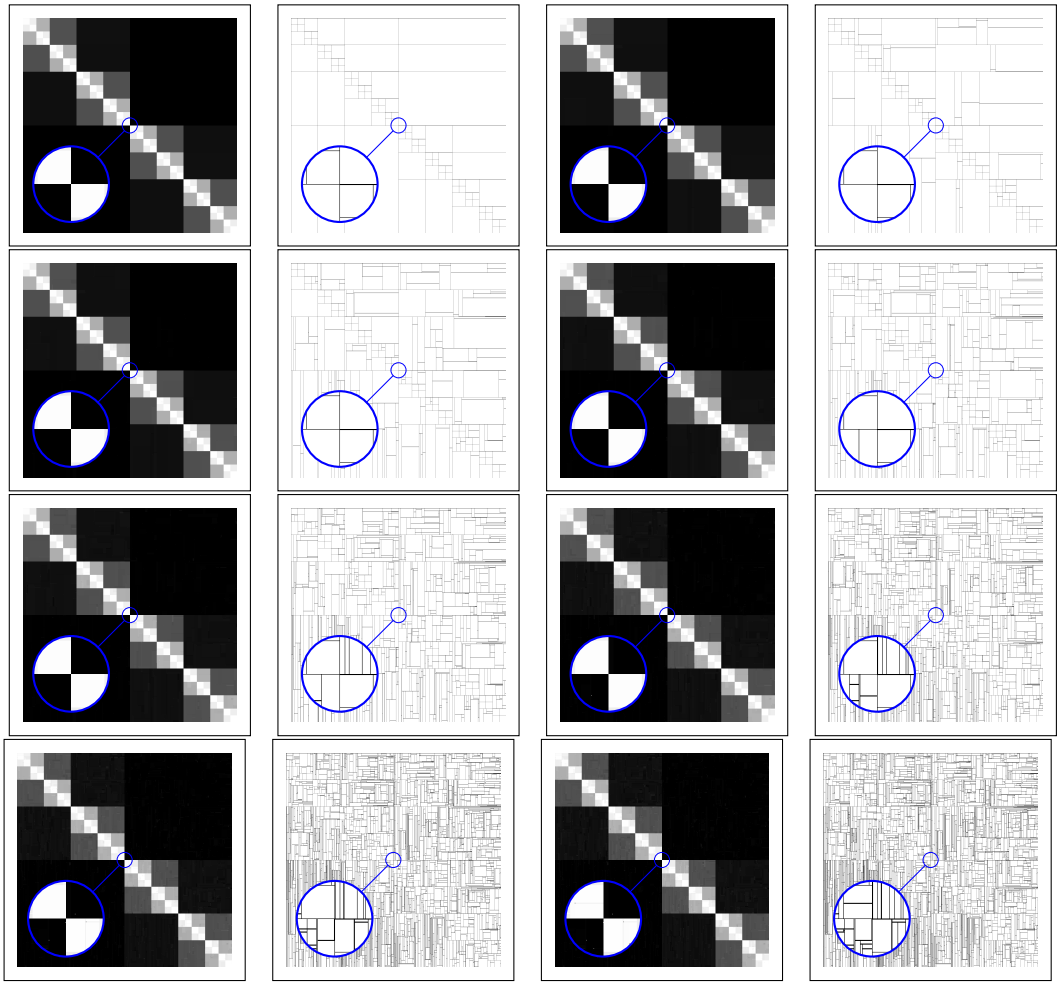


Figure 7 – Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [149, 257, 515, 1027, 2050, 4098, 8193, 10002](estimates and borders)

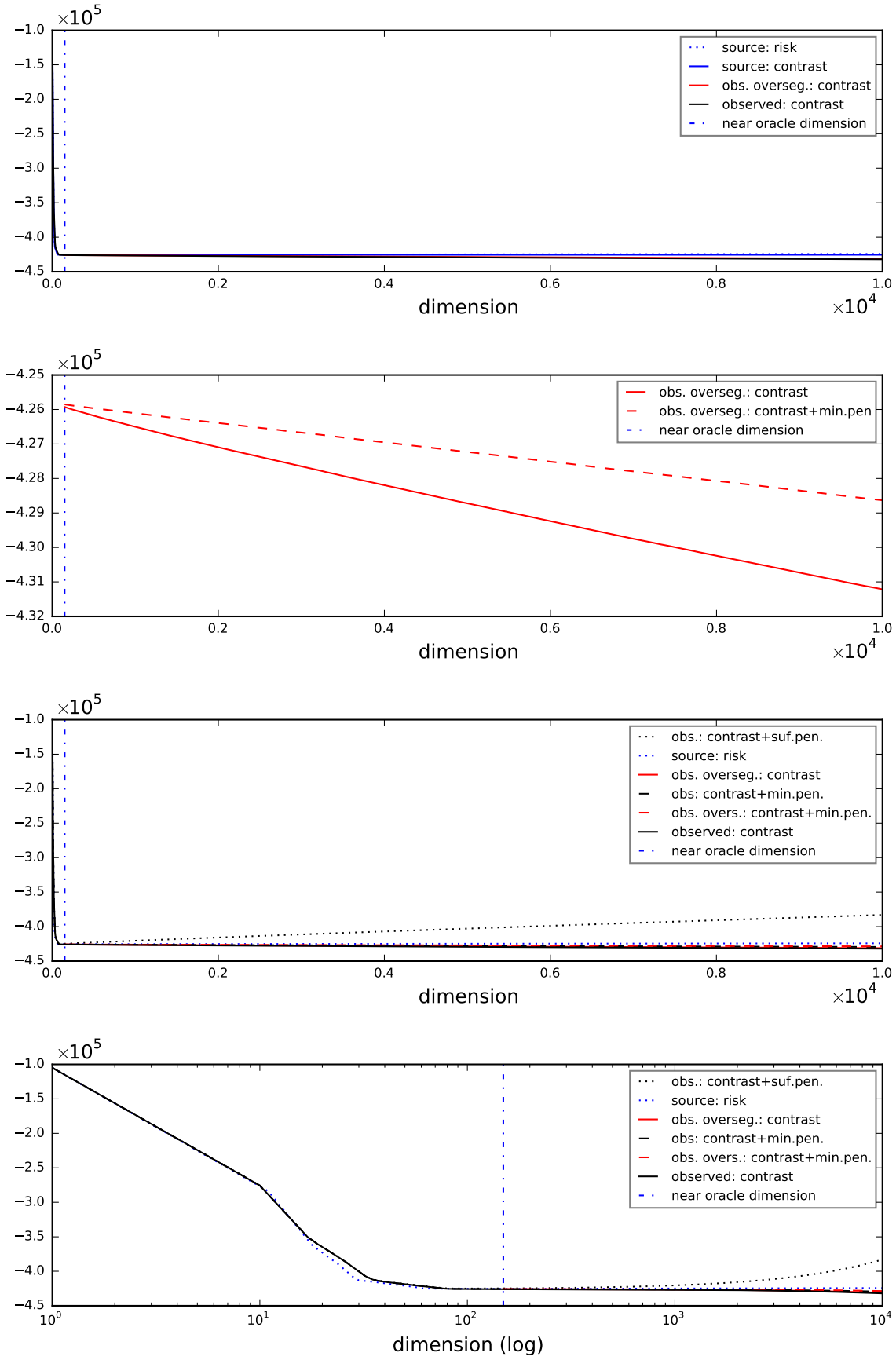
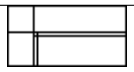


Figure 8 – contrasts and penalized contrast: segmentation paths of source and observed (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties.

8.5.2 Image squares, free quad tree

dimensions				std. dev.	
size	height	width	near oracle	source	noise
$4,194,304$	2048	2048	94	0.28	2.31

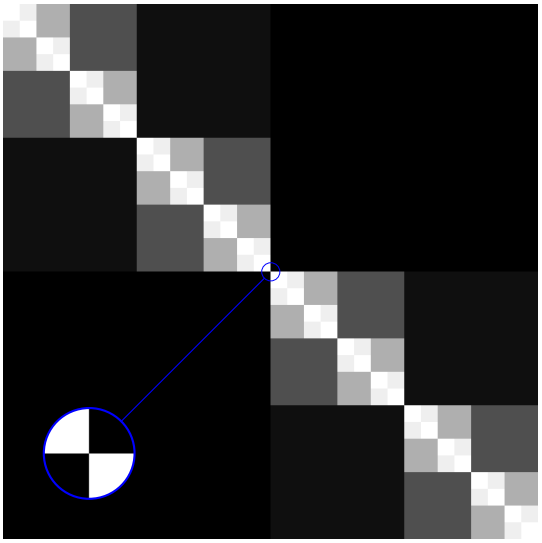
splitting		maximum dimension increase	
mode	max branching factor	to isolate a point (in this image size)	per split
free quad tree 	$b = 4$	$C_{ext} = 6$	$b - 1 = 3$

algorithm		
growing method	contrast criterion	max. dim.
impurity	l2	10000

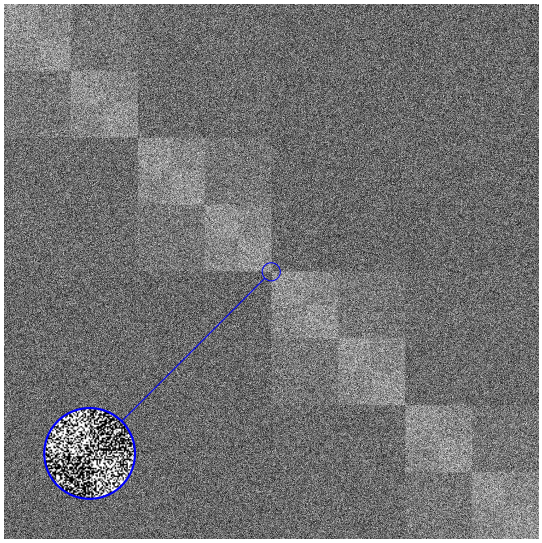
sufficient penalties				
type	expression	L_m	$B_{\mathcal{M}}$	θ
binom	$\epsilon^2 \lvert m\rvert \left[1 + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1} L_m\right]$	$\log(2) + B_{\mathcal{M}} \log(\frac{en}{\lvert m\rvert})$	2	1.00

minimal penalties (limit)
expression
$\epsilon^2 \frac{\lvert m\rvert}{C_{ext}} \left[2\log(\frac{0.48n}{\lvert m\rvert}) - \log(2\log(\frac{0.48n}{\lvert m\rvert})) - \frac{\log(\lvert m\rvert)+2}{\lvert m\rvert}\right]$

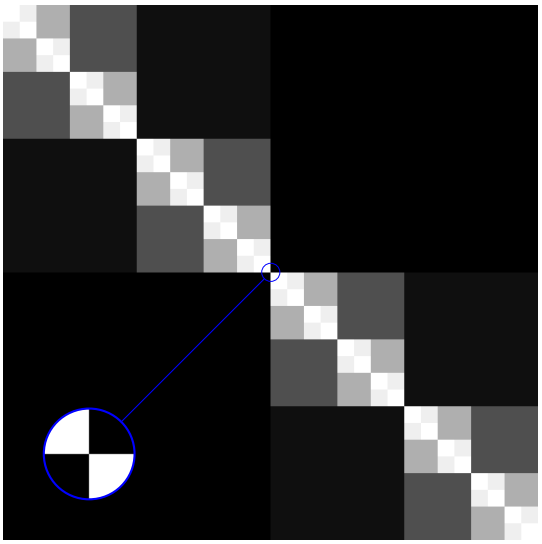
Table 5 – Data for image squares example squares.4fold.(2048, 2048)



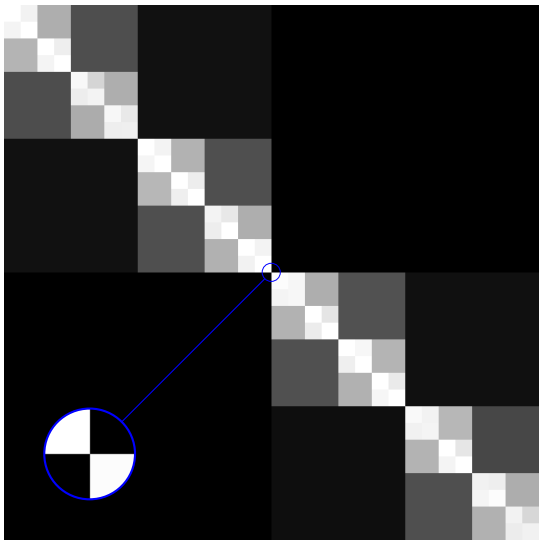
(a) Source image



(b) Noisy image



(c) Source image projected on near oracle of dimension 94



(d) Noisy image projected on near oracle of dimension 94

Figure 9 – Image sqs300binom4fisourceimage.png

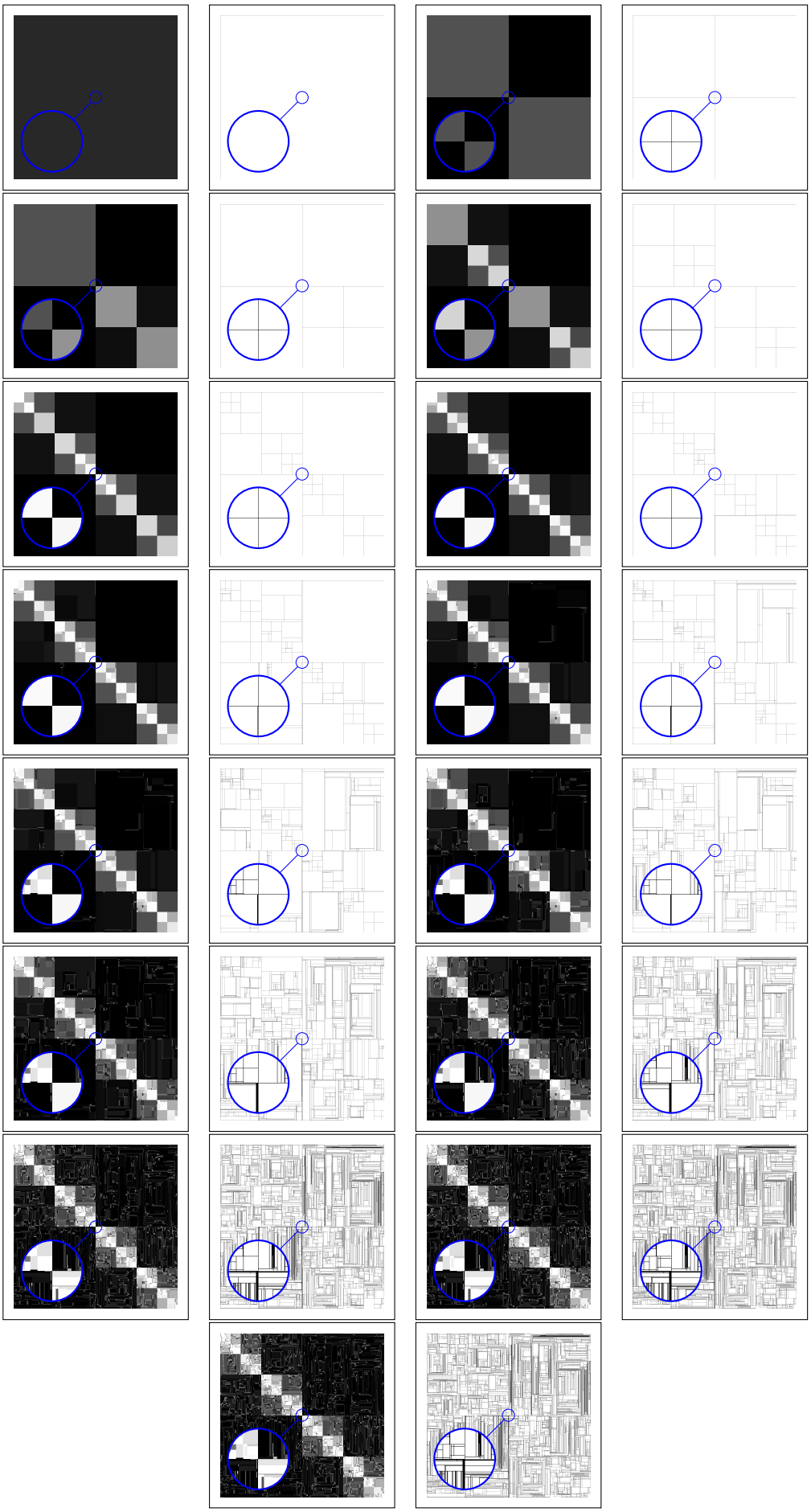


Figure 10 – Segmentation path of noisy image, at dimensions [1, 4, 7, 16, 31, 64, 127, 255, 498, 1023, 2046, 4088, 8188, 9976, 9979](estimates and borders)

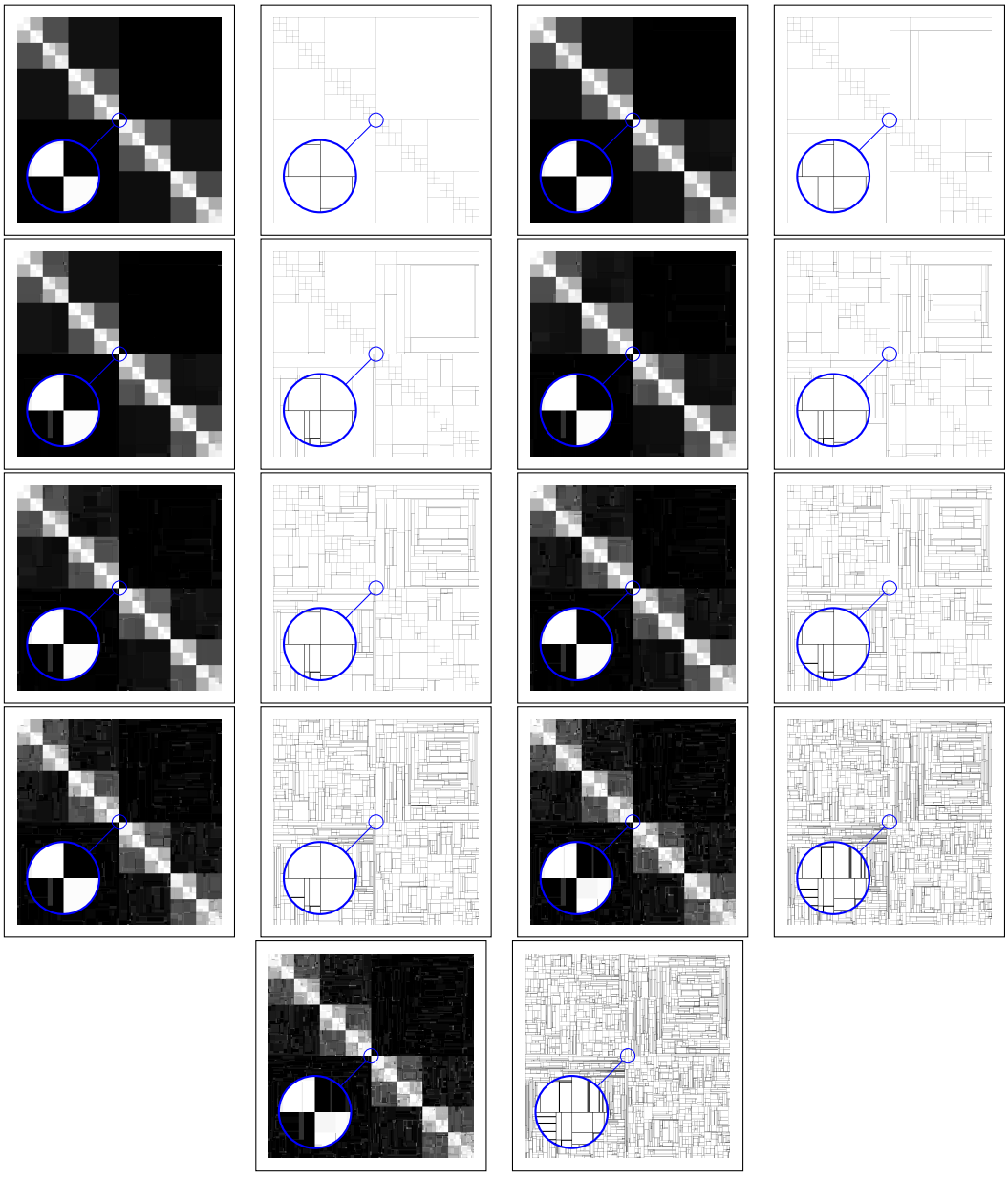


Figure 11 – Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [94, 130, 256, 512, 1028, 2050, 4098, 8195, 10001](estimates and borders)

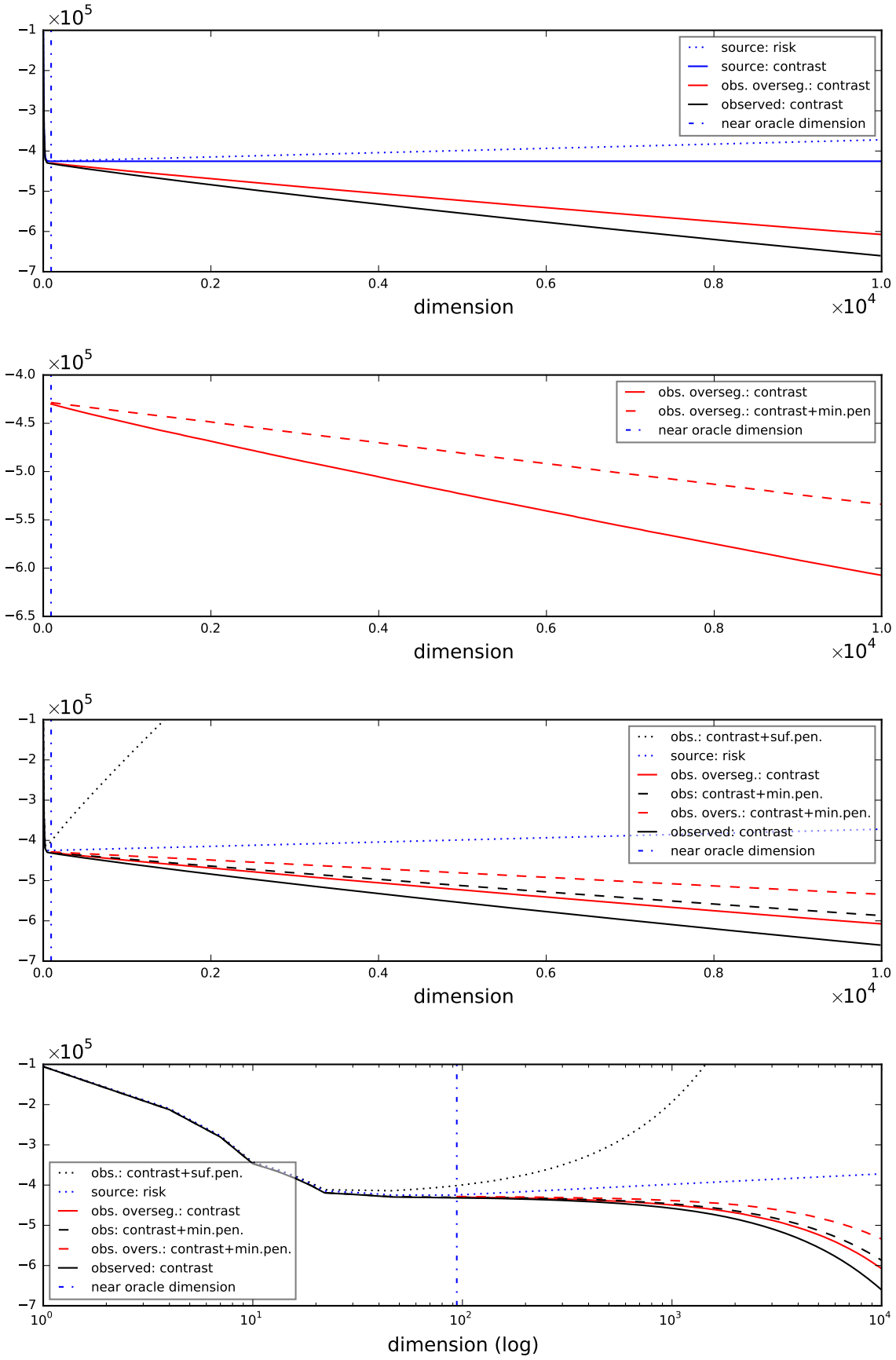
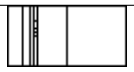


Figure 12 – contrasts and penalized contrast: segmentation paths of source and observed (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties.

8.5.3 Image squares, regular binary tree

dimensions				std. dev.	
size	height	width	near oracle	source	noise
$4,194,304$	2048	2048	94	0.28	2.31

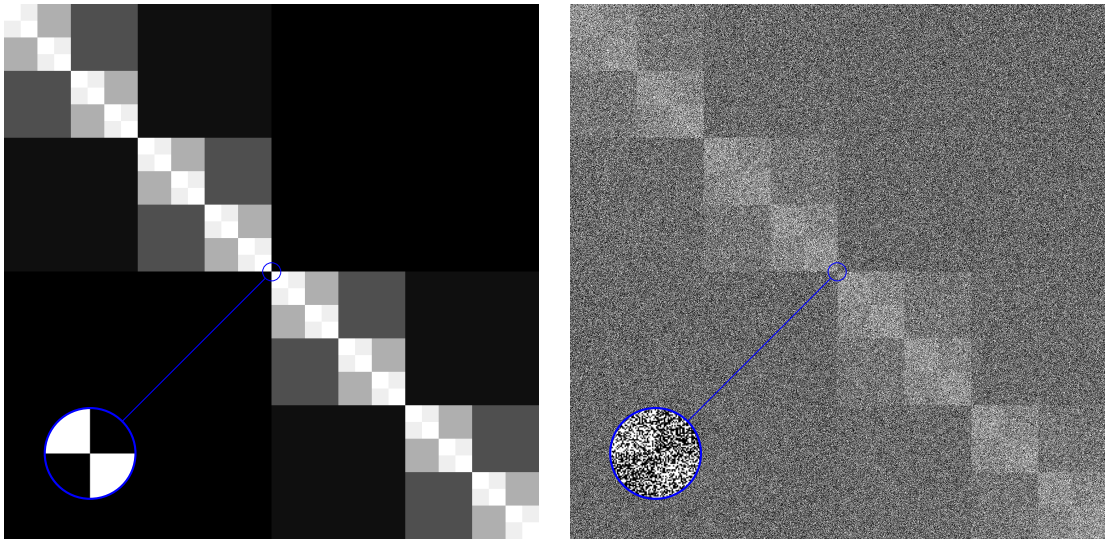
splitting		maximum dimension increase	
mode	max branching factor	to isolate a point (in this image size)	per split
regular binary tree 	$b = 2$	$C_{ext} = 22.0$	$b - 1 = 1$

algorithm		
growing method	contrast criterion	max. dim.
impurity	l2	10000

sufficient penalties				
type	expression	L_m	$B_{\mathcal{M}}$	θ
binom	$\epsilon^2 \lvert m\rvert \left[1 + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1} L_m\right]$	$\log(2) + B_{\mathcal{M}} \log(\frac{en}{\lvert m\rvert})$	1.4	1.00

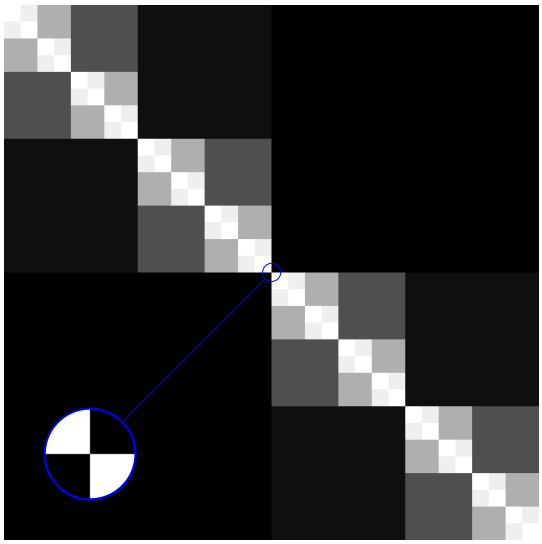
minimal penalties (limit)
expression
$\epsilon^2 \frac{\lvert m\rvert}{C_{ext}} \left[2\log(\frac{0.48n}{\lvert m\rvert}) - \log(2\log(\frac{0.48n}{\lvert m\rvert})) - \frac{\log(\lvert m\rvert)+2}{\lvert m\rvert}\right]$

Table 6 – Data for image squares example squares.2foldequal.(2048, 2048)

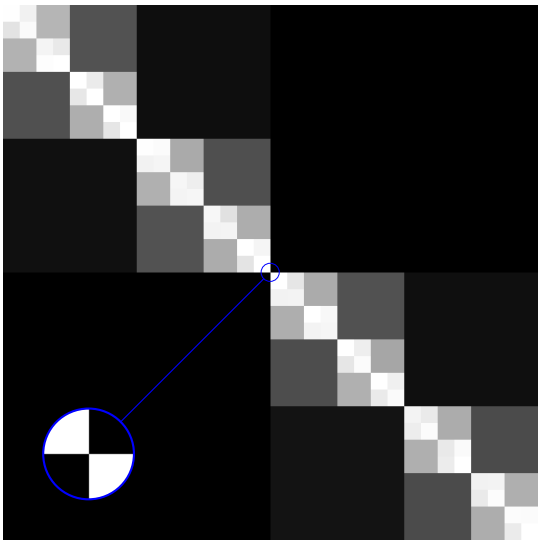


(a) Source image

(b) Noisy image



(c) Source image projected on near oracle of dimension 94



(d) Noisy image projected on near oracle of dimension 94

Figure 13 – Image sqs300binom2risourceimage.png

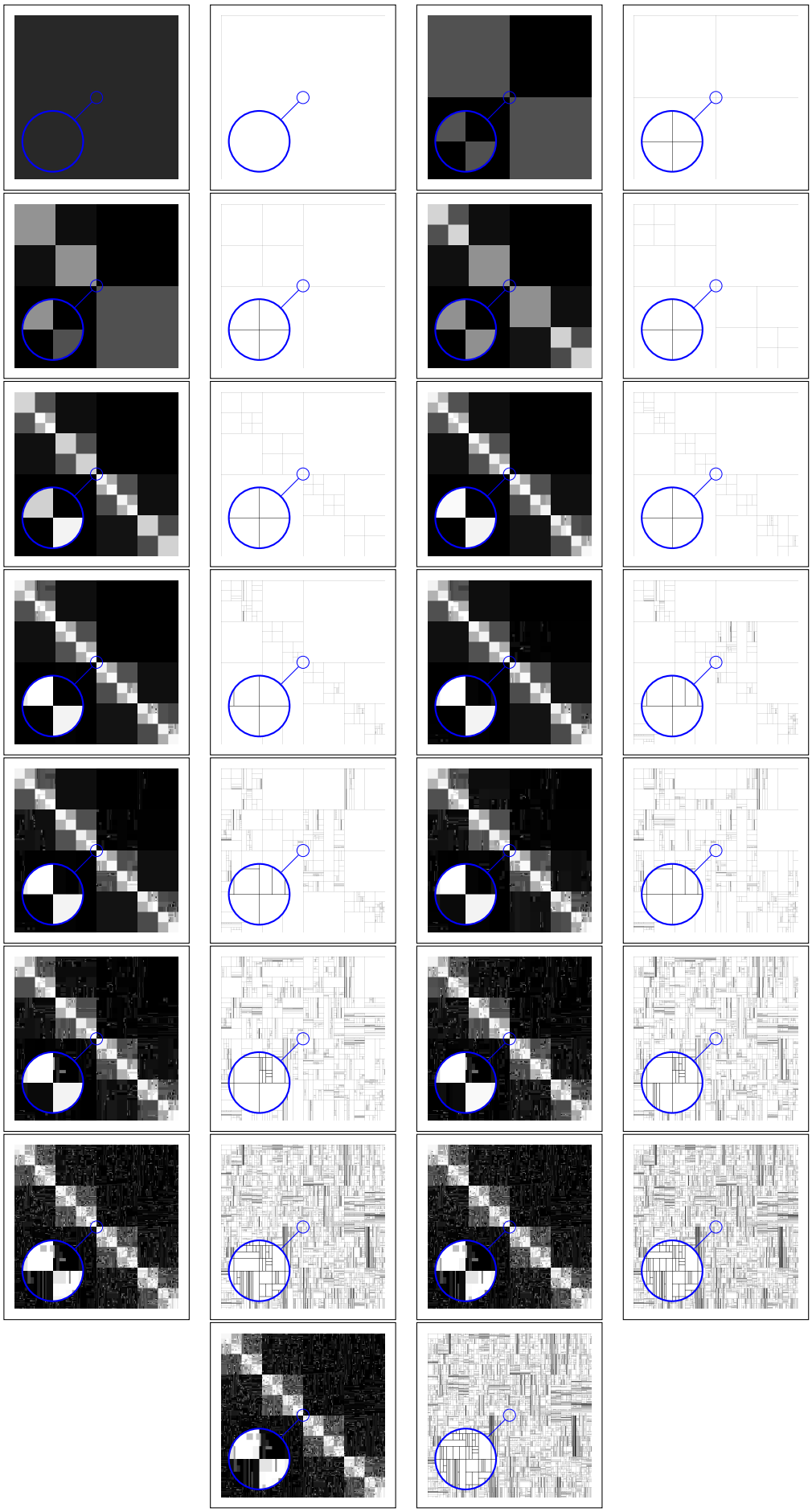


Figure 14 – Segmentation path of noisy image, at dimensions [1, 4, 7, 16, 31, 63, 128, 251, 503, 1024, 2045, 4096, 8191, 9992, 9993](estimates and borders)

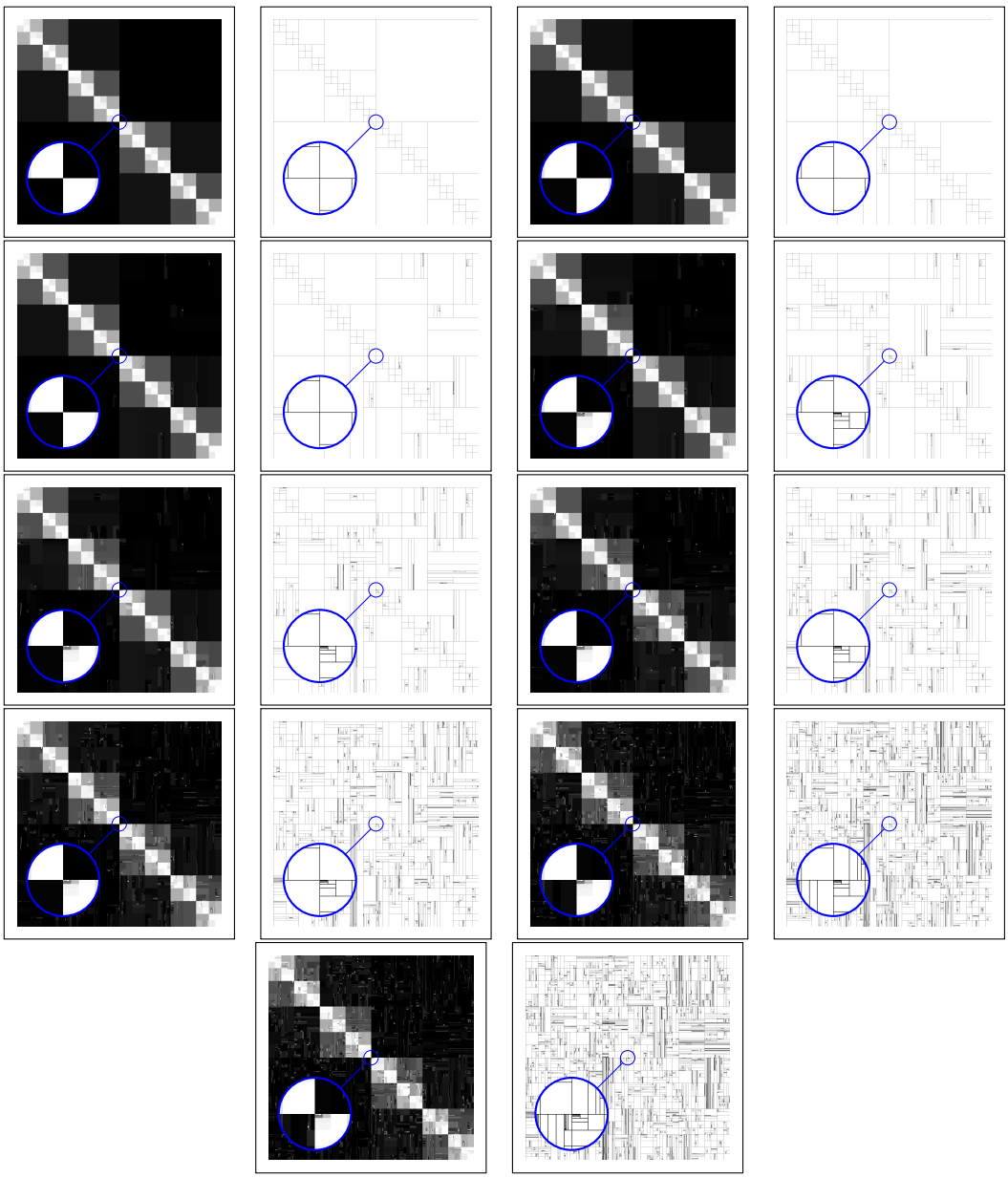


Figure 15 – Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [94, 132, 257, 513, 1024, 2053, 4108, 8198, 10006](estimates and borders)

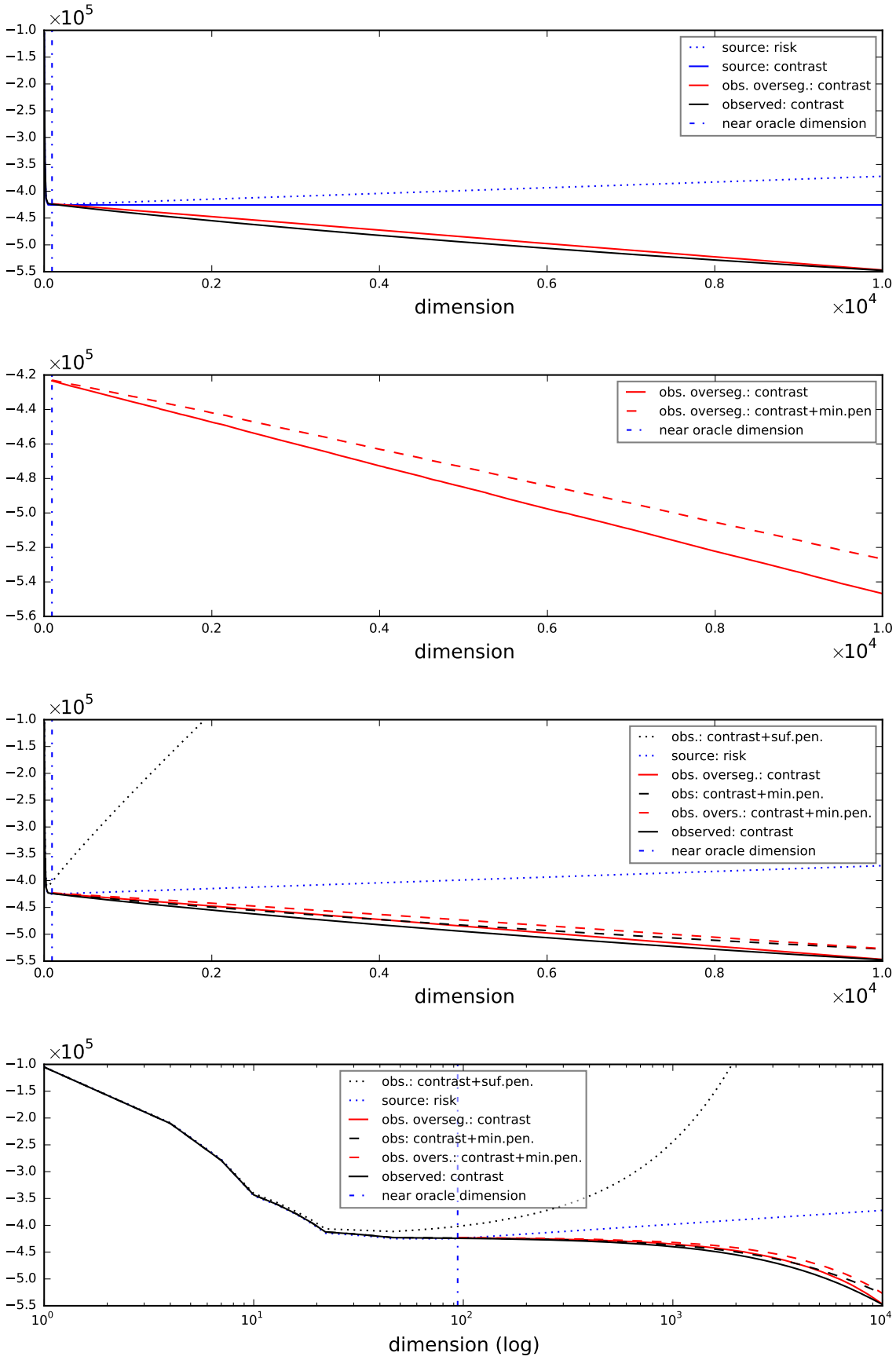


Figure 16 – contrasts and penalized contrast: segmentation paths of source and observed (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties.

8.5.4 Image squares, regular quad tree

dimensions				std. dev.	
size	height	width	near oracle	source	noise
$4,194,304$	2048	2048	94	0.28	2.31

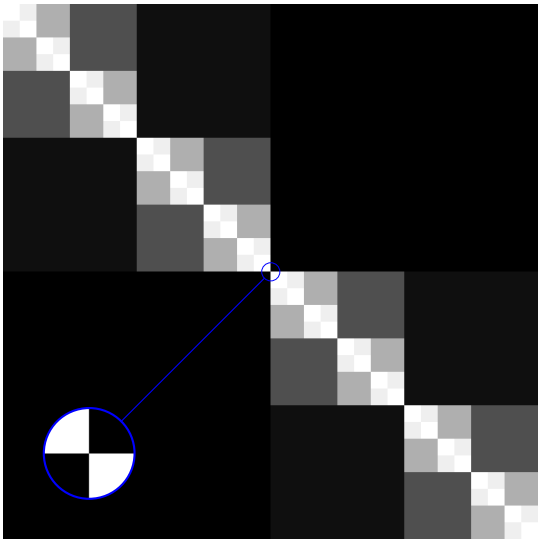
splitting		maximum dimension increase	
mode	max branching factor	to isolate a point (in this image size)	per split
regular quad tree 	$b = 4$	$C_{ext} = 33.0$	$b - 1 = 3$

algorithm		
growing method	contrast criterion	max. dim.
impurity	l2	10000

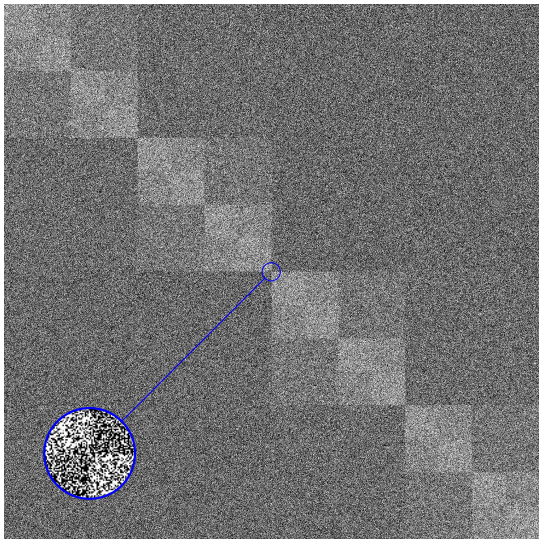
sufficient penalties				
type	expression	L_m	$B_{\mathcal{M}}$	θ
binom	$\epsilon^2 \lvert m\rvert \left[1 + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1} L_m\right]$	$\log(2) + B_{\mathcal{M}} \log(\frac{en}{\lvert m\rvert})$	0.057	1.00

minimal penalties (limit)
expression
$\epsilon^2 \frac{\lvert m\rvert}{C_{ext}} \left[2\log(\frac{0.48n}{\lvert m\rvert}) - \log(2\log(\frac{0.48n}{\lvert m\rvert})) - \frac{\log(\lvert m\rvert)+2}{\lvert m\rvert}\right]$

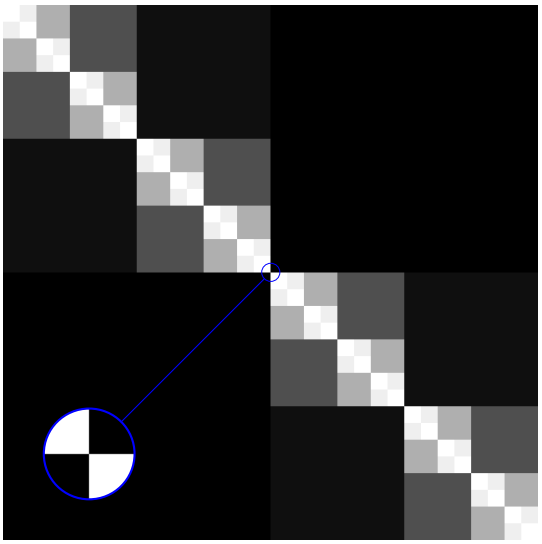
Table 7 – Data for image squares example squares.4foldequal.(2048, 2048)



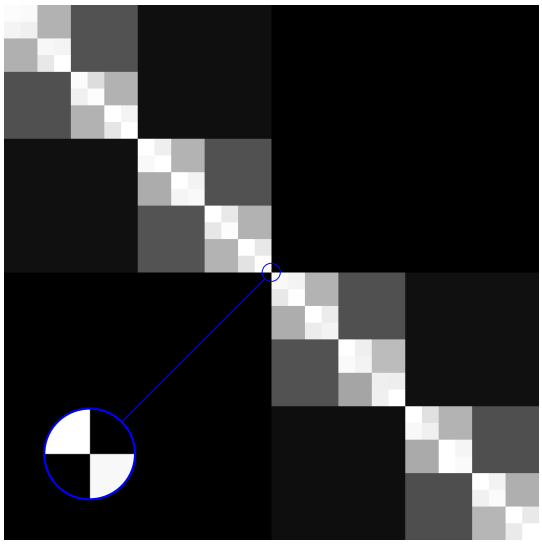
(a) Source image



(b) Noisy image



(c) Source image projected on near oracle of dimension 94



(d) Noisy image projected on near oracle of dimension 94

Figure 17 – Image sqs300binom4risourceimage.png

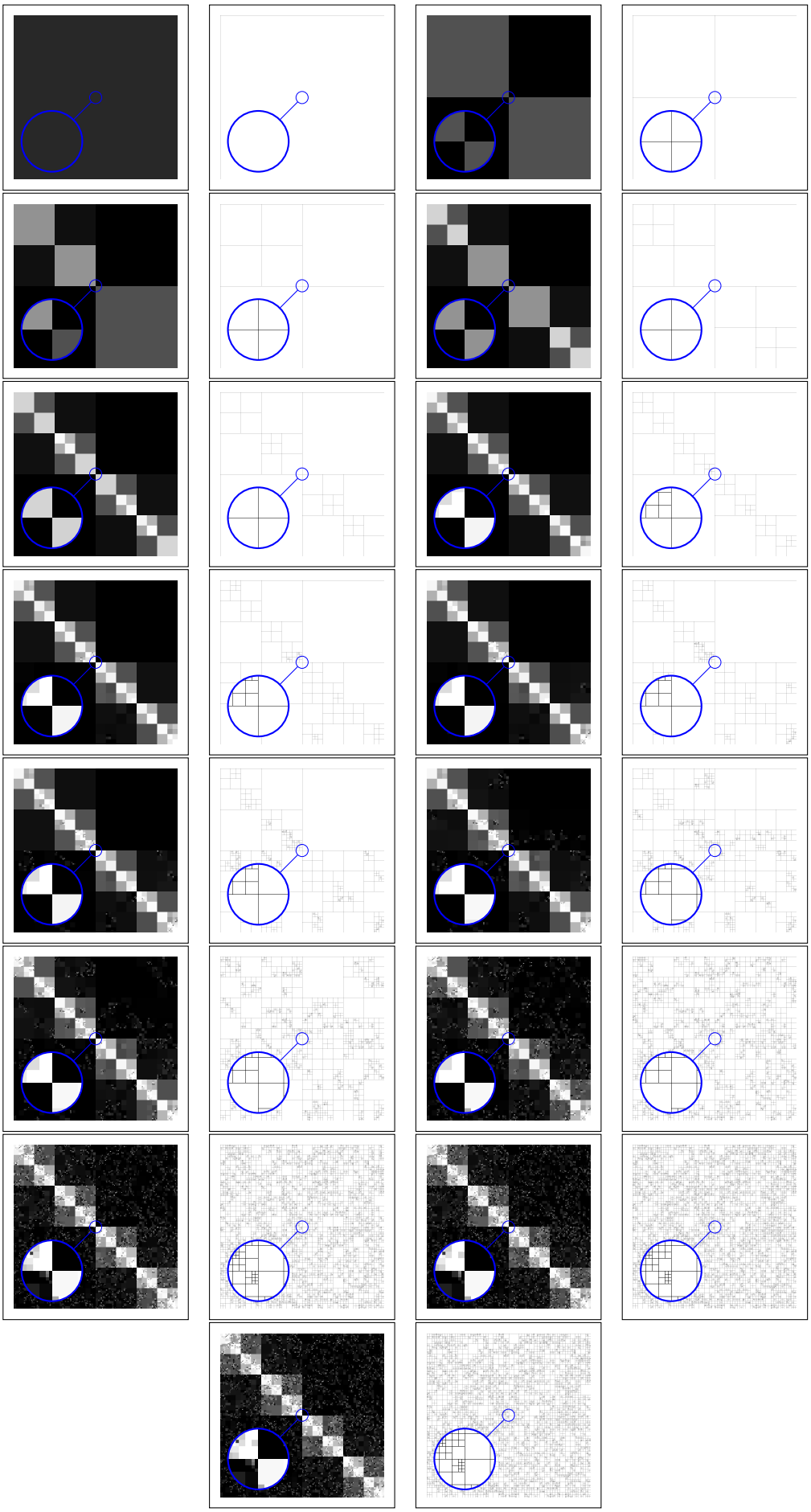


Figure 18 – Segmentation path of noisy image, at dimensions [1, 4, 7, 16, 31, 58, 121, 256, 511, 970, 2038, 4093, 8185, 9985, 9988](estimates and borders)

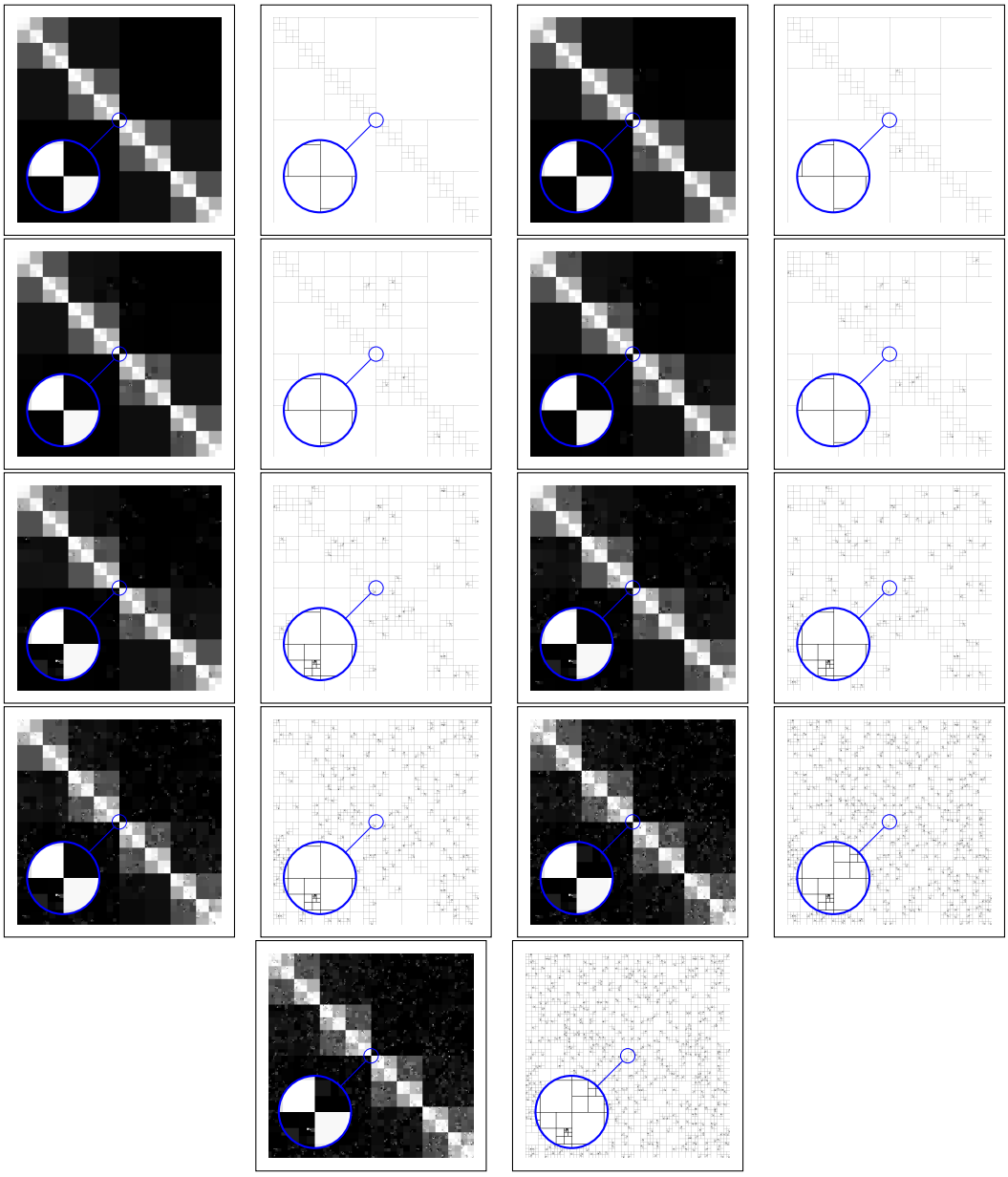


Figure 19 – Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [94, 148, 280, 517, 1036, 2068, 4105, 8203, 10012](estimates and borders)

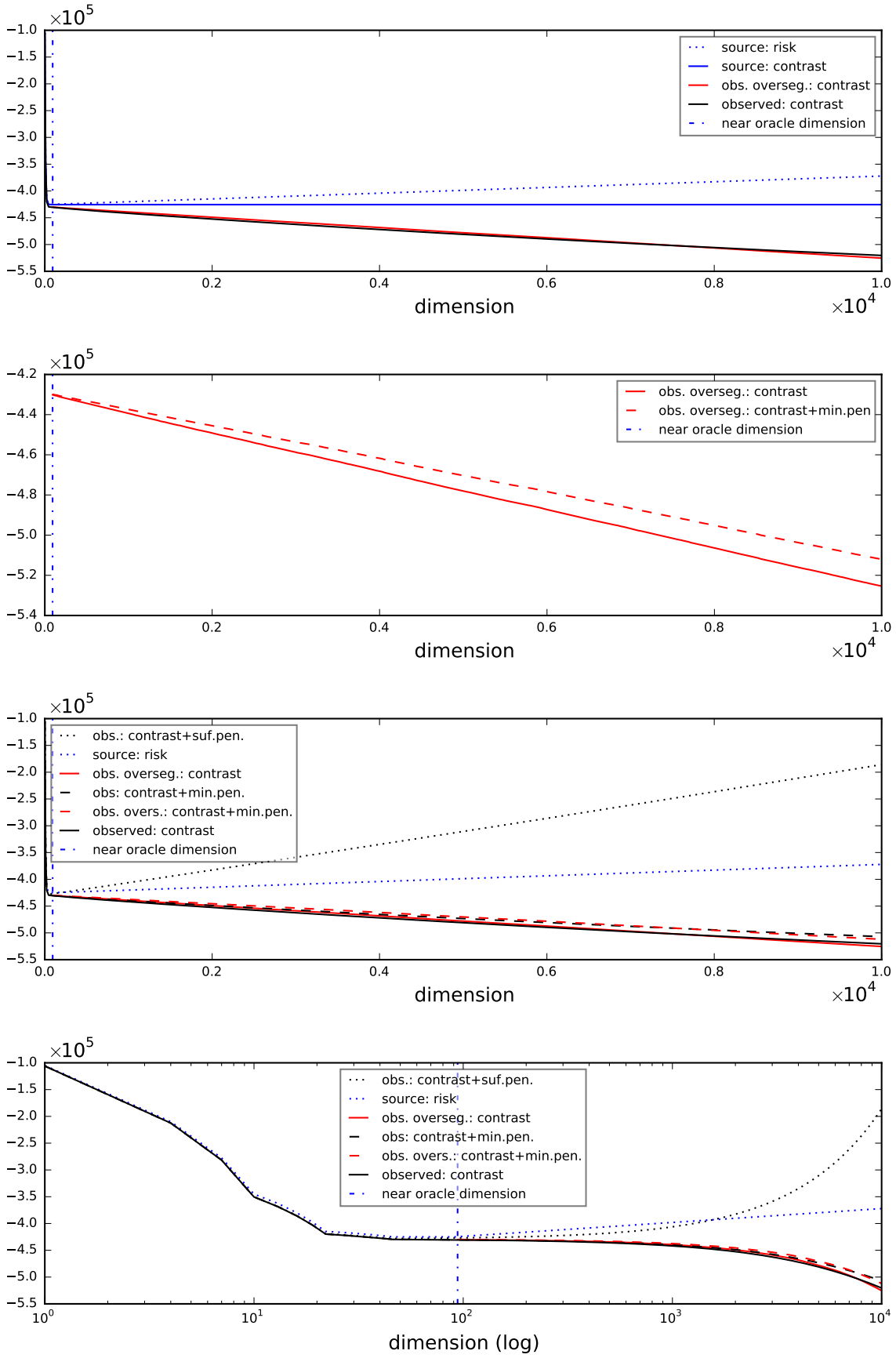


Figure 20 – contrasts and penalized contrast: segmentation paths of source and observed (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties.

8.5.5 Image lone tree, free binary tree

dimensions				std. dev.	
size	height	width	near oracle	source	noise
$1,835,856$	1098	1672	301	0.27	3.83

splitting		maximum dimension increase	
mode	max branching factor	to isolate a point (in this image size)	per split
free binary tree 	$b = 2$	$C_{ext} = 4$	$b - 1 = 1$

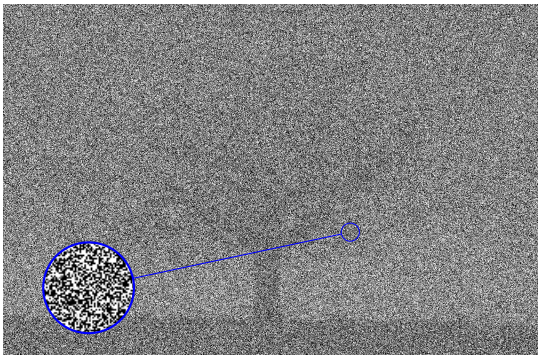
algorithm		
growing method	contrast criterion	max. dim.
impurity	l2	10000

sufficient penalties				
type	expression	L_m	$B_{\mathcal{M}}$	θ
binom	$\epsilon^2 \lvert m\rvert \left[1 + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1} L_m\right]$	$\log(2) + B_{\mathcal{M}} \log(\frac{en}{\lvert m\rvert})$	2	1.00

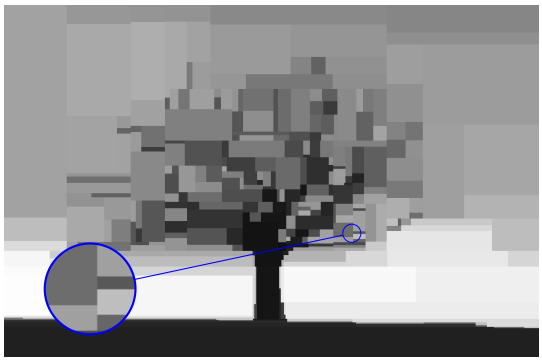
minimal penalties (limit)
expression
$\epsilon^2 \frac{\lvert m\rvert}{C_{ext}} \left[2\log(\frac{0.48n}{\lvert m\rvert}) - \log(2\log(\frac{0.48n}{\lvert m\rvert})) - \frac{\log(\lvert m\rvert)+2}{\lvert m\rvert}\right]$

Table 8 – Data for image lone tree example lone tree.2fold.(1098, 1672)

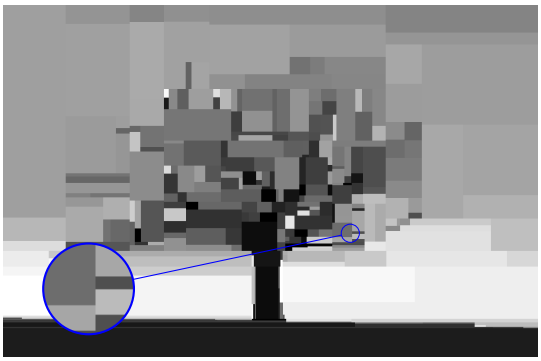
(a) Source image



(b) Noisy image



(c) Source image projected on near oracle of dimension 301



(d) Noisy image projected on near oracle of dimension 301

Figure 21 – Image lontr300binom2fisourceimage.png

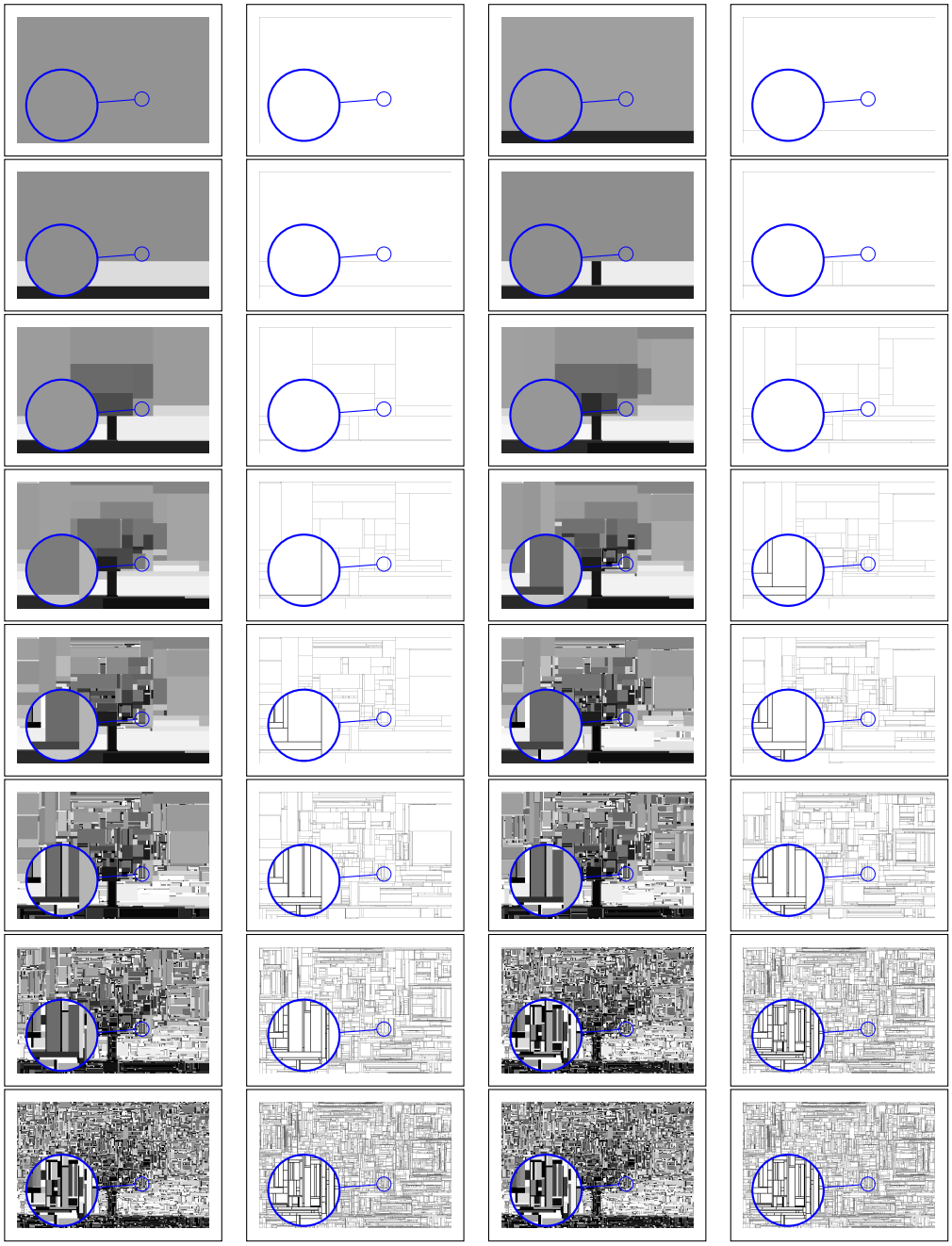


Figure 22 – Segmentation path of noisy image, at dimensions [1, 2, 3, 6, 15, 32, 64, 128, 238, 512, 1000, 2043, 4086, 8192, 9997, 9999](estimates and borders)

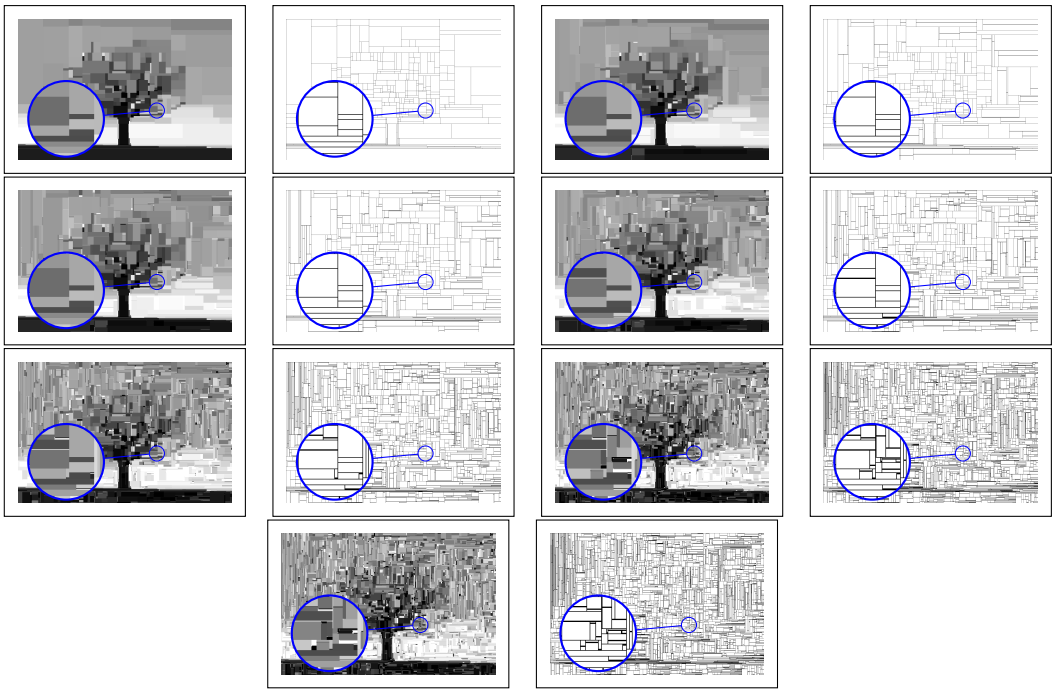


Figure 23 – Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [301, 514, 1027, 2048, 4098, 8194, 10000](estimates and borders)

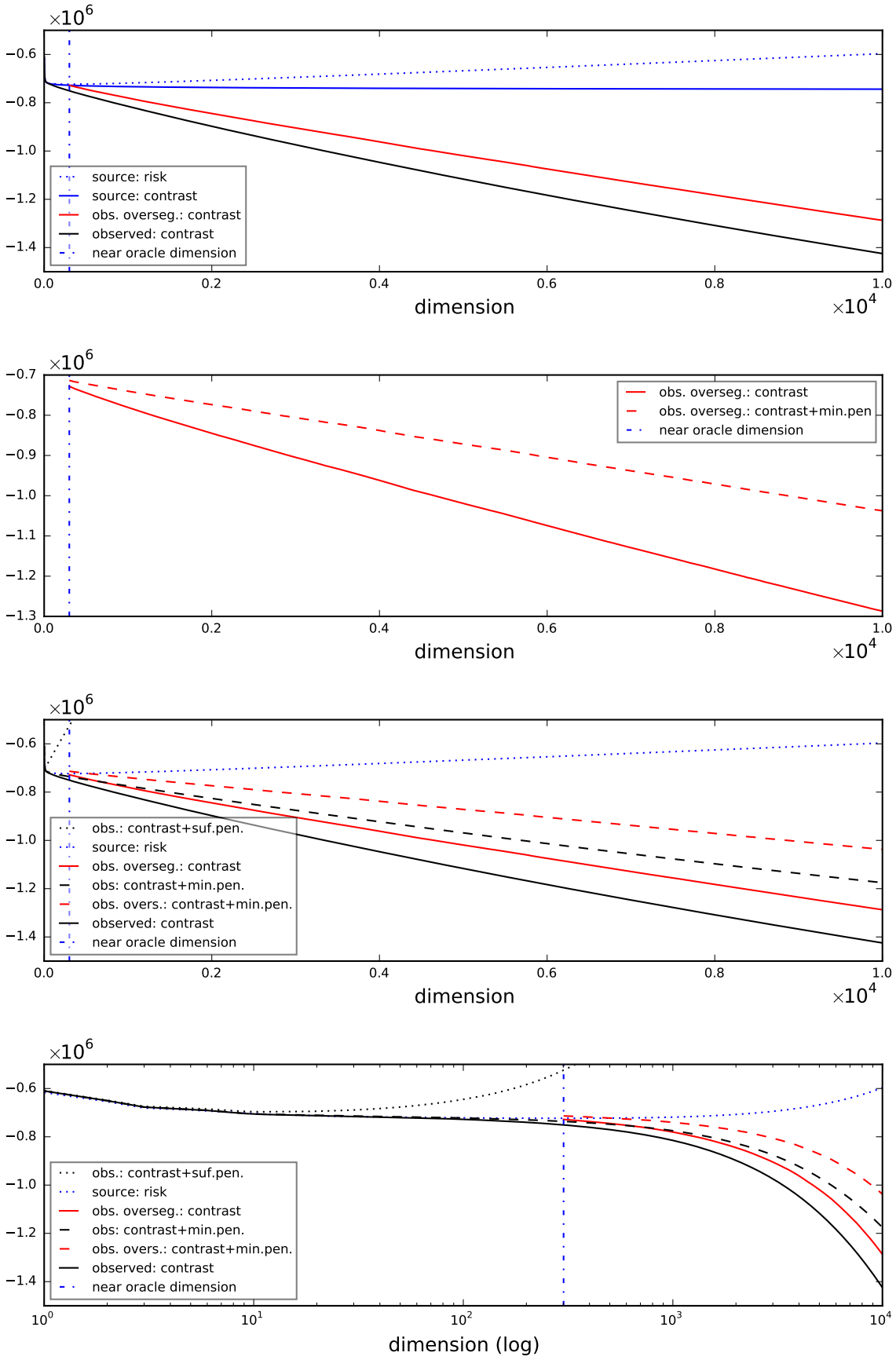


Figure 24 – contrasts and penalized contrast: segmentation paths of source and observed (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties.

8.5.6 Image lone tree, free binary tree

dimensions				std. dev.	
size	height	width	near oracle	source	noise
$1,835,856$	1098	1672	1002	0.27	2.24

splitting		maximum dimension increase	
mode	max branching factor	to isolate a point (in this image size)	per split
free binary tree 	$b = 2$	$C_{ext} = 4$	$b - 1 = 1$

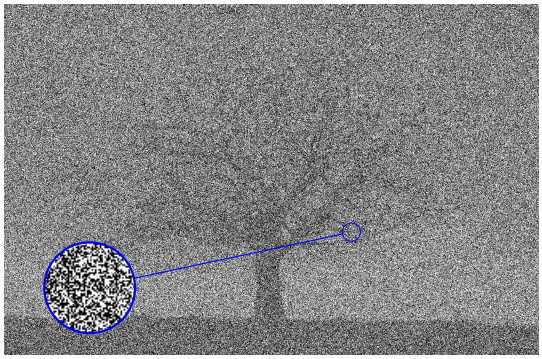
algorithm		
growing method	contrast criterion	max. dim.
impurity	l2	10000

sufficient penalties				
type	expression	L_m	$B_{\mathcal{M}}$	θ
binom	$\epsilon^2 \lvert m\rvert \left[1 + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1} L_m\right]$	$\log(2) + B_{\mathcal{M}} \log(\frac{en}{\lvert m\rvert})$	2	1.00

minimal penalties (limit)
expression
$\epsilon^2 \frac{\lvert m\rvert}{C_{ext}} \left[2\log(\frac{0.48n}{\lvert m\rvert}) - \log(2\log(\frac{0.48n}{\lvert m\rvert})) - \frac{\log(\lvert m\rvert)+2}{\lvert m\rvert}\right]$

Table 9 – Data for image lone tree example lone tree.2fold.(1098, 1672)

(a) Source image



(b) Noisy image

(c) Source image projected on near oracle of dimension 1002

(d) Noisy image projected on near oracle of dimension 1002

Figure 25 – Image lontr1000binom2fisourceimage.png

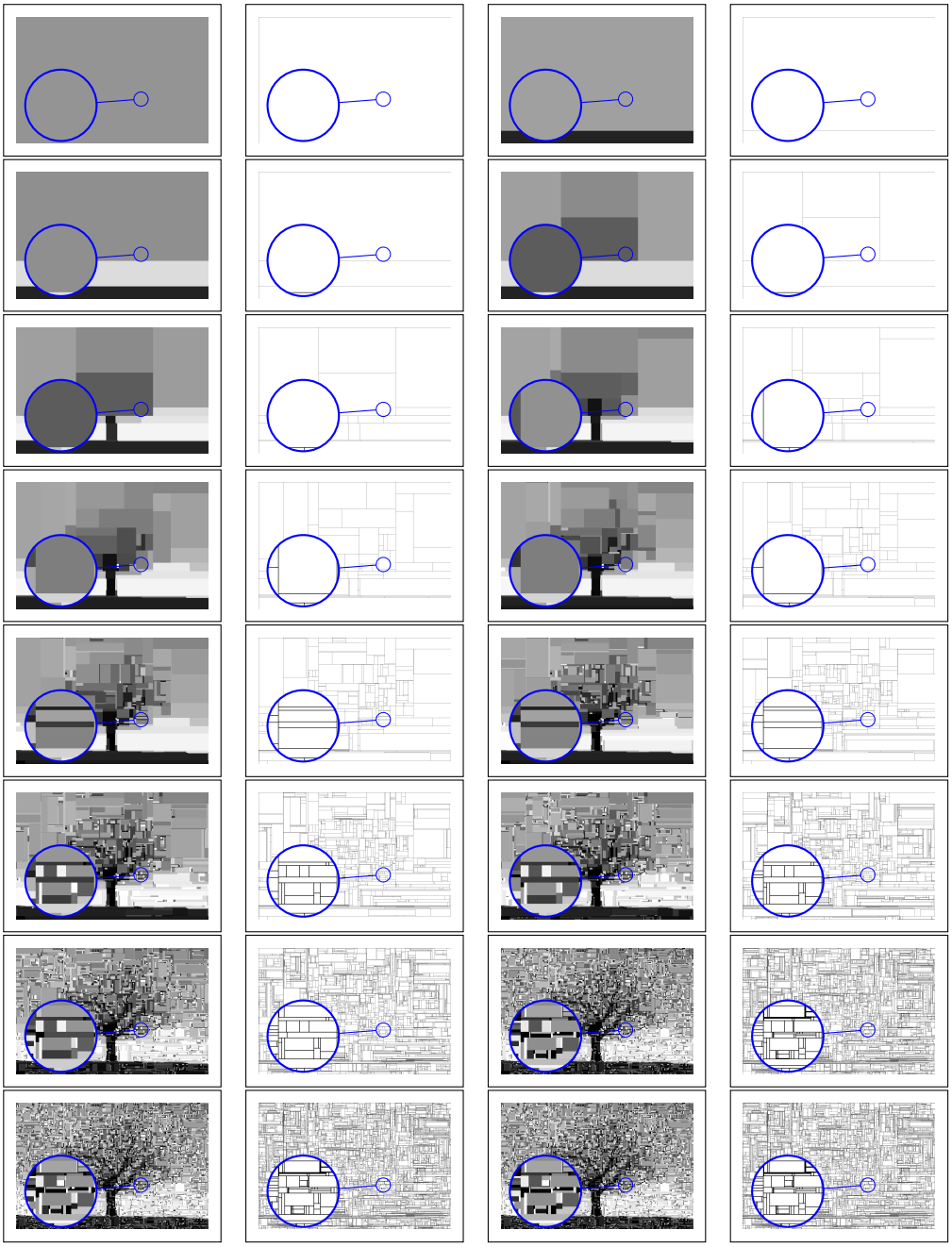


Figure 26 – Segmentation path of noisy image, at dimensions [1, 2, 3, 6, 16, 30, 63, 127, 255, 511, 1023, 2045, 4091, 8192, 9995, 9999](estimates and borders)

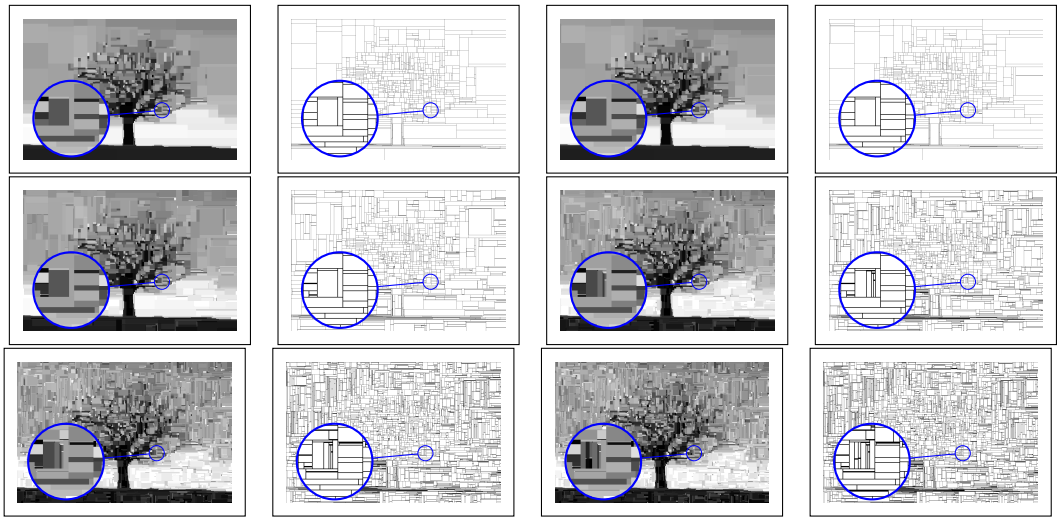


Figure 27 – Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [1002, 1026, 2048, 4096, 8195, 10001](estimates and borders)

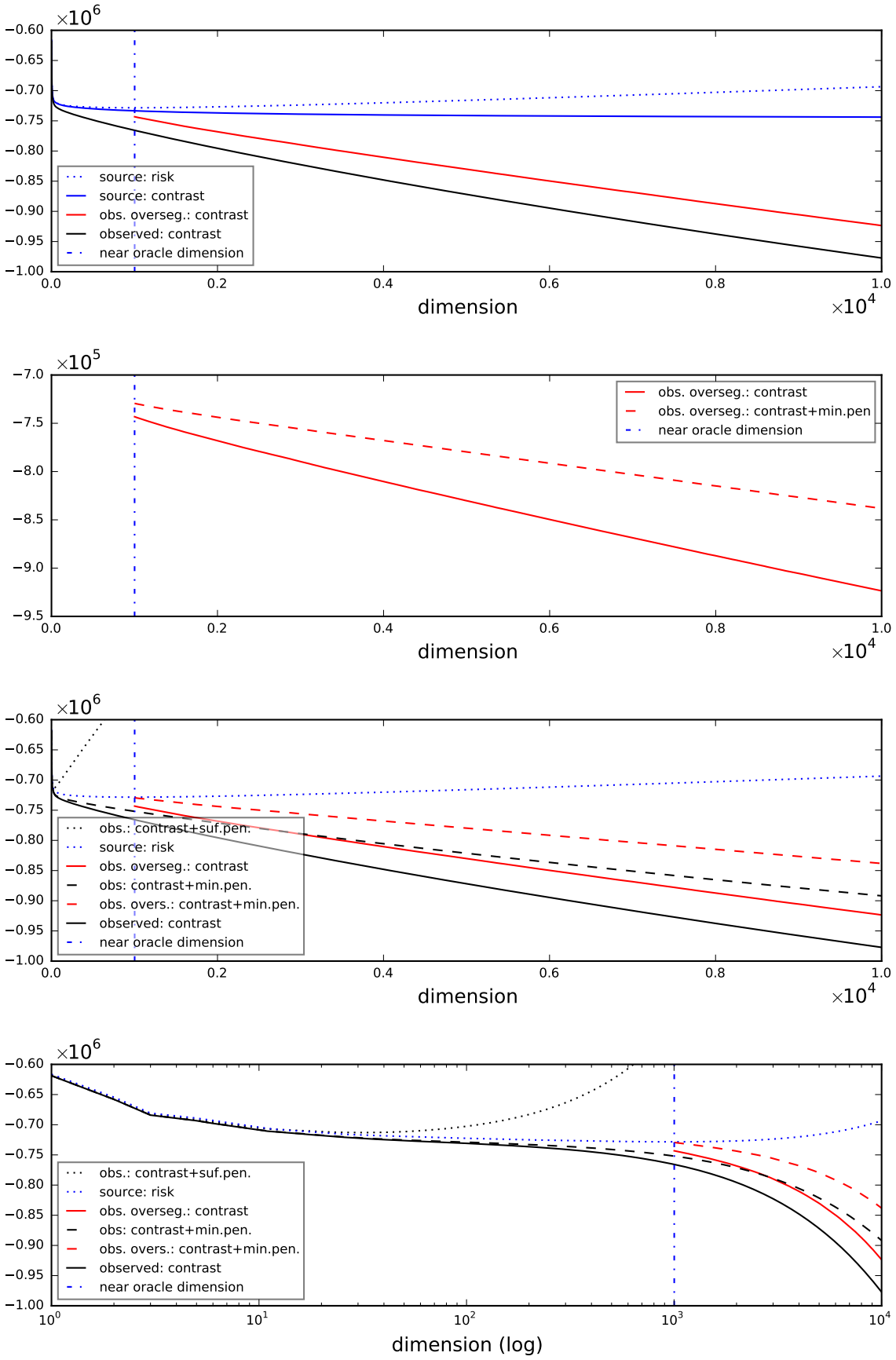


Figure 28 – contrasts and penalized contrast: segmentation paths of source and observed (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties.

8.5.7 Image lone tree, free binary tree

dimensions				std. dev.	
size	height	width	near oracle	source	noise
$1,835,856$	1098	1672	3001	0.27	1.28

splitting		maximum dimension increase	
mode	max branching factor	to isolate a point (in this image size)	per split
free binary tree 	$b = 2$	$C_{ext} = 4$	$b - 1 = 1$

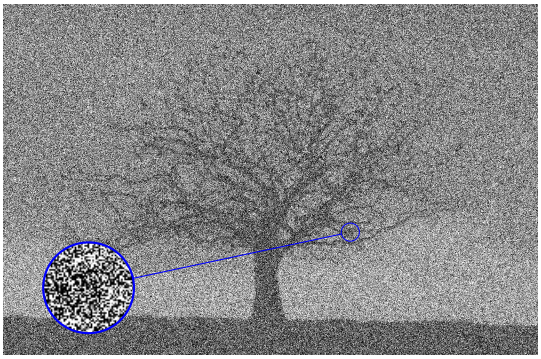
algorithm		
growing method	contrast criterion	max. dim.
impurity	l2	10000

sufficient penalties				
type	expression	L_m	$B_{\mathcal{M}}$	θ
binom	$\epsilon^2 \lvert m \rvert \left[1 + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1} L_m\right]$	$\log(2) + B_{\mathcal{M}} \log(\frac{en}{\lvert m \rvert})$	2	1.00

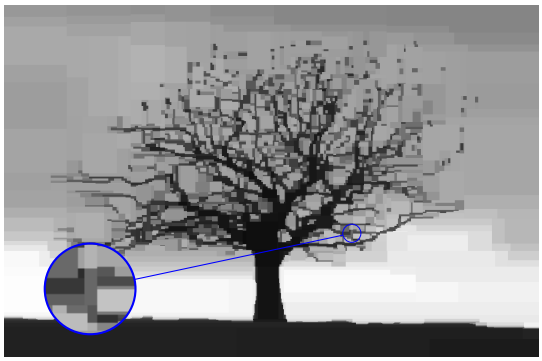
minimal penalties (limit)
expression
$\epsilon^2 \frac{\lvert m \rvert}{C_{ext}} \left[2\log(\frac{0.48n}{\lvert m \rvert}) - \log(2\log(\frac{0.48n}{\lvert m \rvert})) - \frac{\log(\lvert m \rvert)+2}{\lvert m \rvert}\right]$

Table 10 – Data for image lone tree example lone tree.2fold.(1098, 1672)

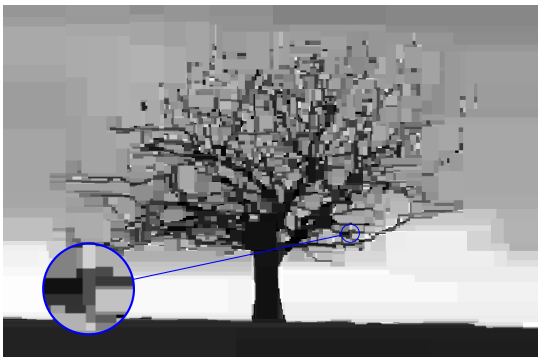
(a) Source image



(b) Noisy image



(c) Source image projected on near oracle of dimension 3001



(d) Noisy image projected on near oracle of dimension 3001

Figure 29 – Image lontr3000binom2fisourceimage.png

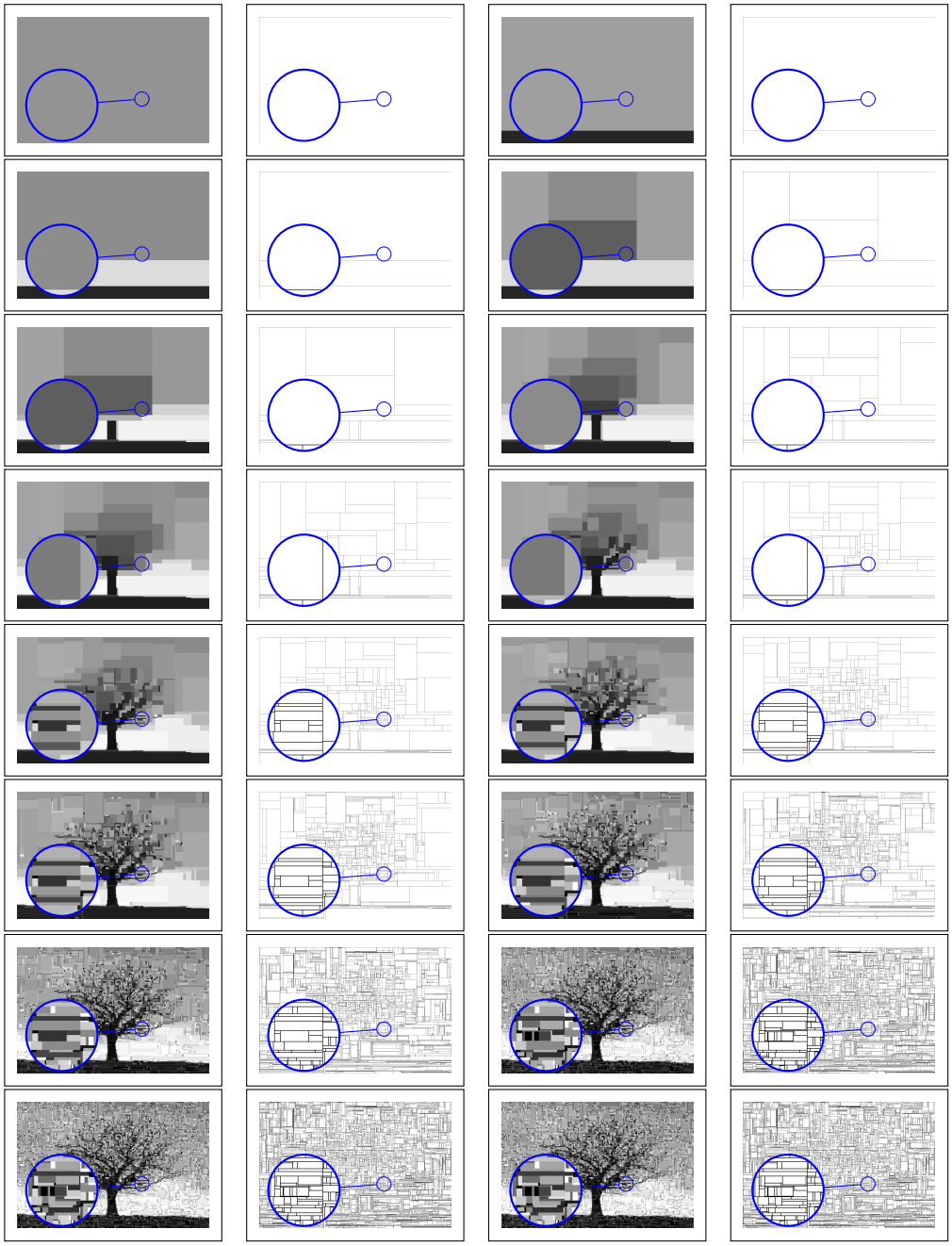


Figure 30 – Segmentation path of noisy image, at dimensions [1, 2, 3, 6, 16, 31, 63, 127, 250, 508, 1024, 2033, 4094, 8186, 9992, 9998](estimates and borders)

Figure 31 – Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [3001, 4096, 8195, 10001](estimates and borders)

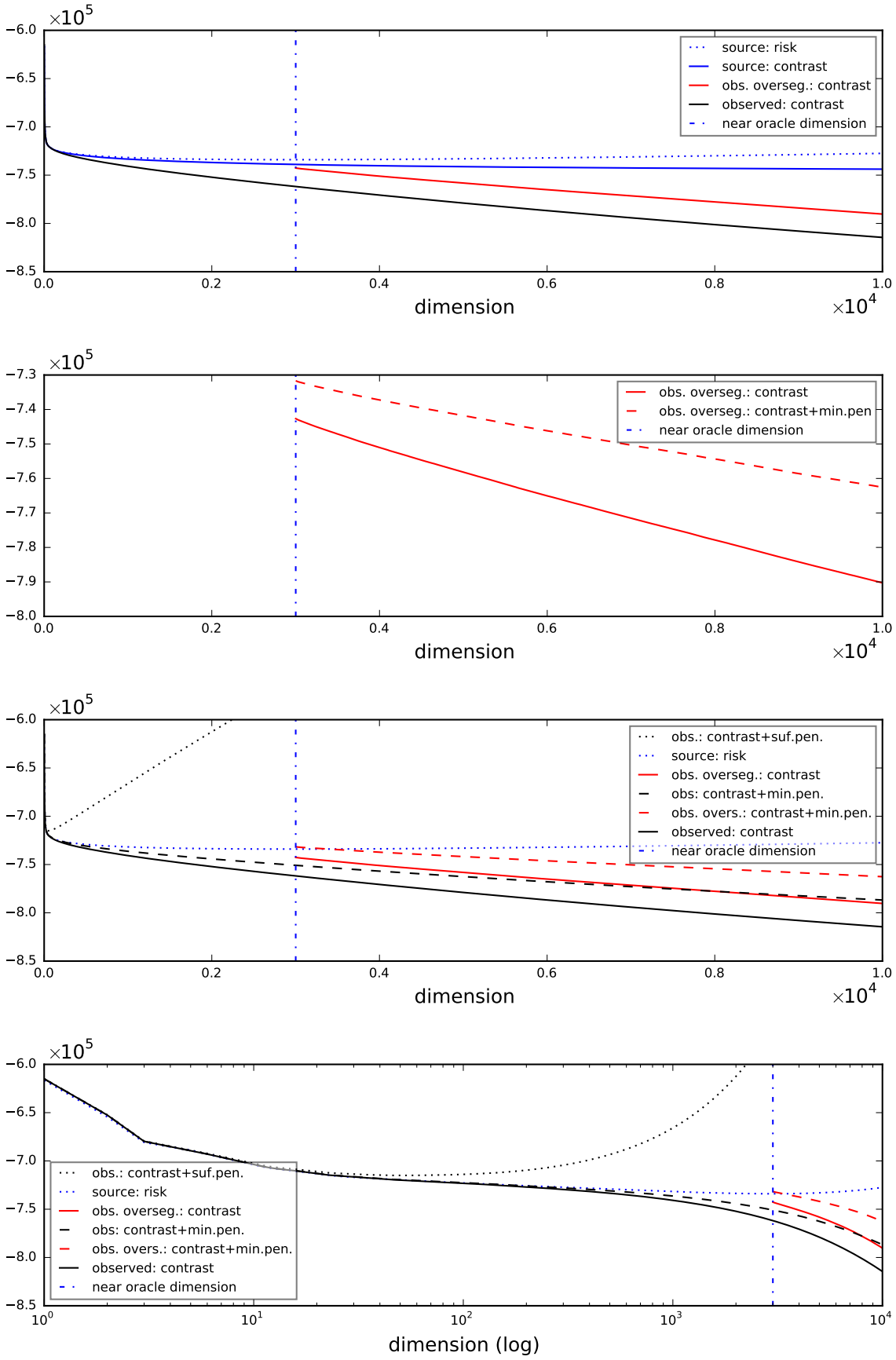


Figure 32 – contrasts and penalized contrast: segmentation paths of source and observed (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties.

8.5.8 Image lone tree, free binary tree

dimensions				std. dev.	
size	height	width	near oracle	source	noise
$1,835,856$	1098	1672	5001	0.27	0.91

splitting		maximum dimension increase	
mode	max branching factor	to isolate a point (in this image size)	per split
free binary tree 	$b = 2$	$C_{ext} = 4$	$b - 1 = 1$

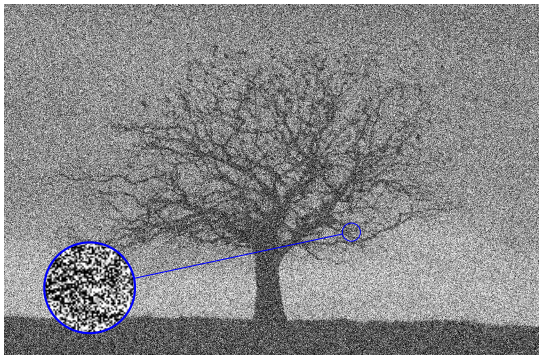
algorithm		
growing method	contrast criterion	max. dim.
impurity	l2	10000

sufficient penalties				
type	expression	L_m	$B_{\mathcal{M}}$	θ
binom	$\epsilon^2 \lvert m\rvert \left[1 + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1} L_m\right]$	$\log(2) + B_{\mathcal{M}} \log(\frac{en}{\lvert m\rvert})$	2	1.00

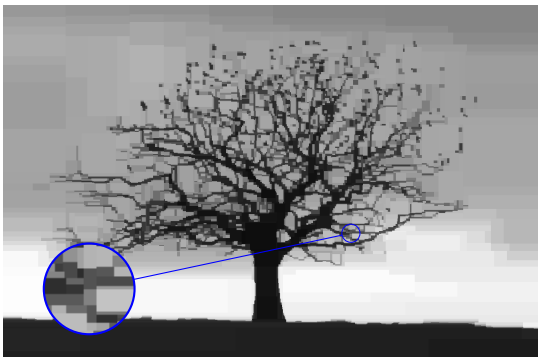
minimal penalties (limit)
expression
$\epsilon^2 \frac{\lvert m\rvert}{C_{ext}} \left[2\log(\frac{0.48n}{\lvert m\rvert}) - \log(2\log(\frac{0.48n}{\lvert m\rvert})) - \frac{\log(\lvert m\rvert)+2}{\lvert m\rvert}\right]$

Table 11 – Data for image lone tree example lone tree.2fold.(1098, 1672)

(a) Source image



(b) Noisy image



(c) Source image projected on near oracle of dimension 5001

(d) Noisy image projected on near oracle of dimension 5001

Figure 33 – Image lontr5000binom2fisourceimage.png

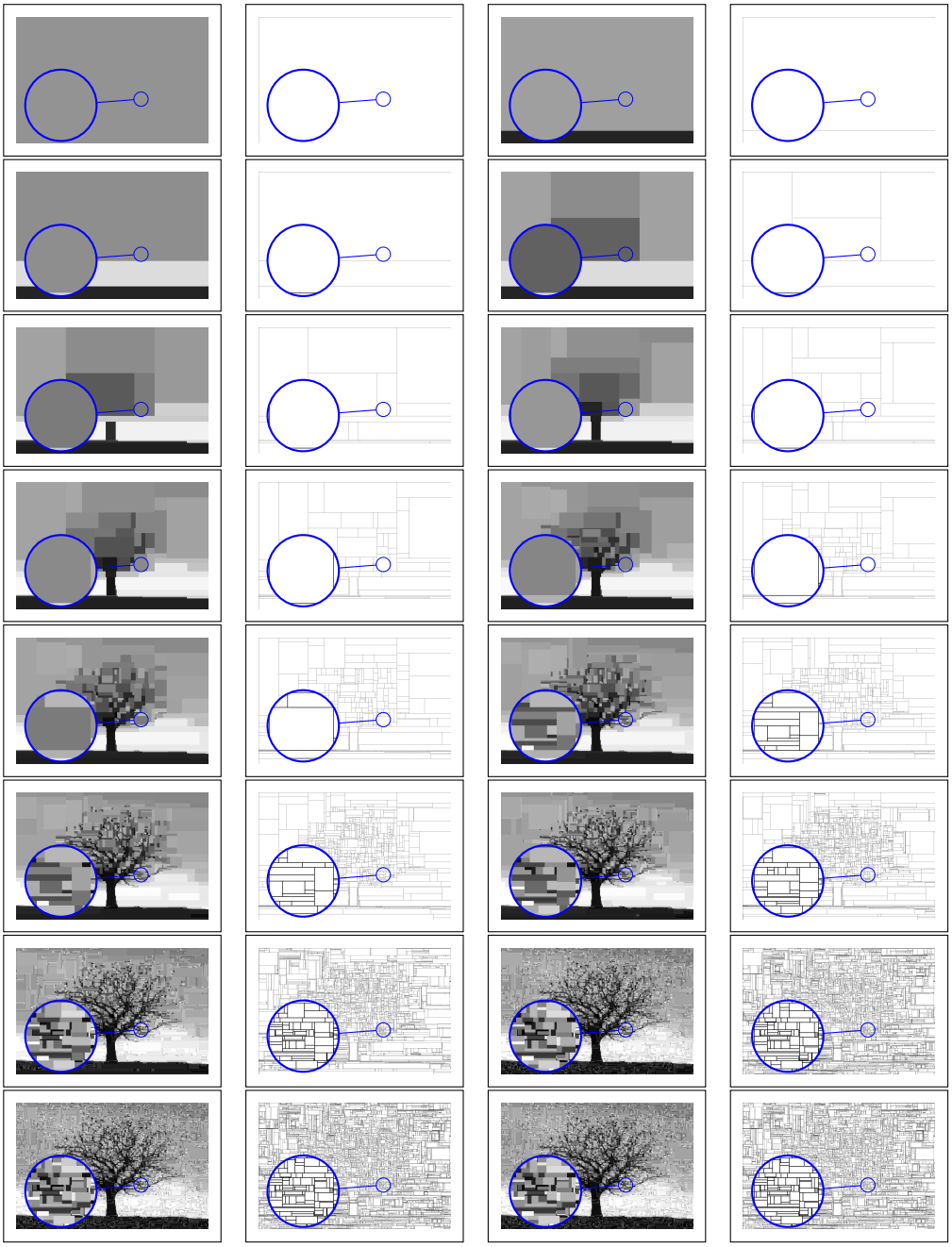


Figure 34 – Segmentation path of noisy image, at dimensions [1, 2, 3, 6, 15, 32, 63, 128, 256, 510, 1024, 2040, 4095, 8192, 9995, 9996](estimates and borders)

Figure 35 – Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [5001, 8192, 10001](estimates and borders)

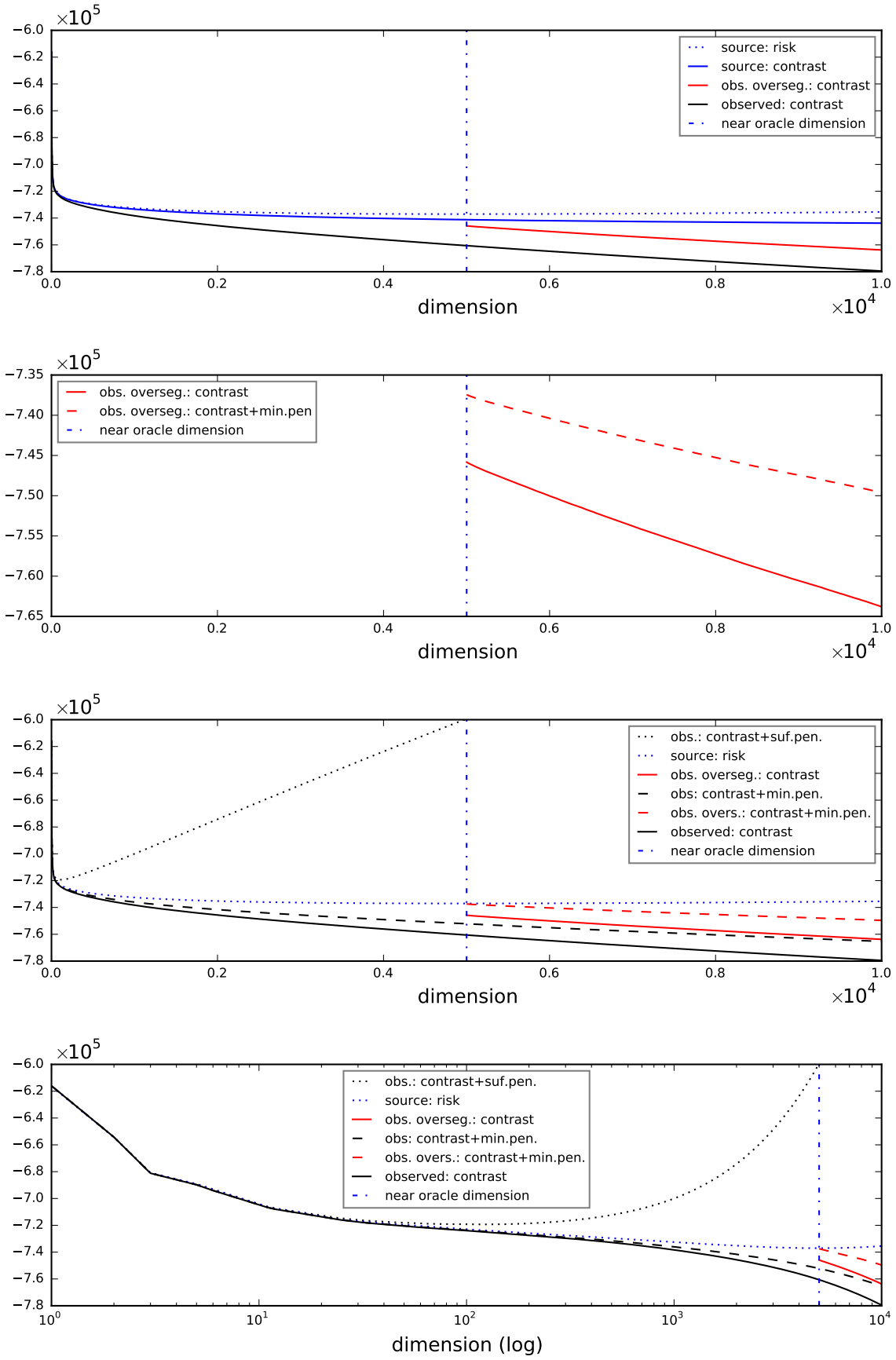


Figure 36 – contrasts and penalized contrast: segmentation paths of source and observed (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties.

8.5.9 Image face, free binary tree

dimensions				std. dev.	
size	height	width	near oracle	source	noise
262,144	512	512	301	0.21	1.82

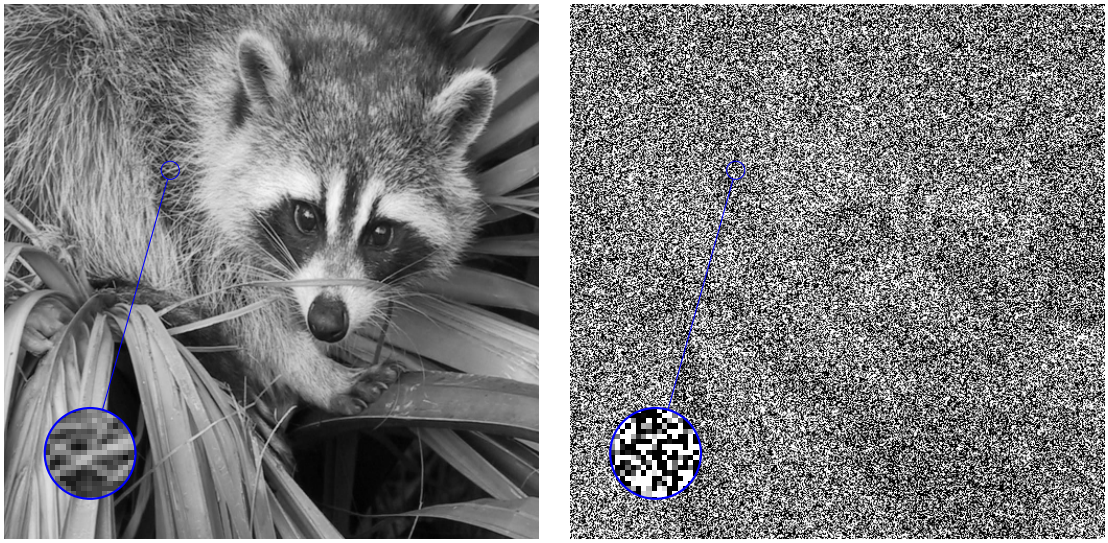
splitting		maximum dimension increase	
mode	max branching factor	to isolate a point (in this image size)	per split
free binary tree 	$b = 2$	$C_{ext} = 4$	$b - 1 = 1$

algorithm		
growing method	contrast criterion	max. dim.
impurity	l2	10000

sufficient penalties				
type	expression	L_m	$B_{\mathcal{M}}$	θ
binom	$\epsilon^2 \lvert m\rvert \left[1 + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1} L_m\right]$	$\log(2) + B_{\mathcal{M}} \log(\frac{en}{\lvert m\rvert})$	2	1.00

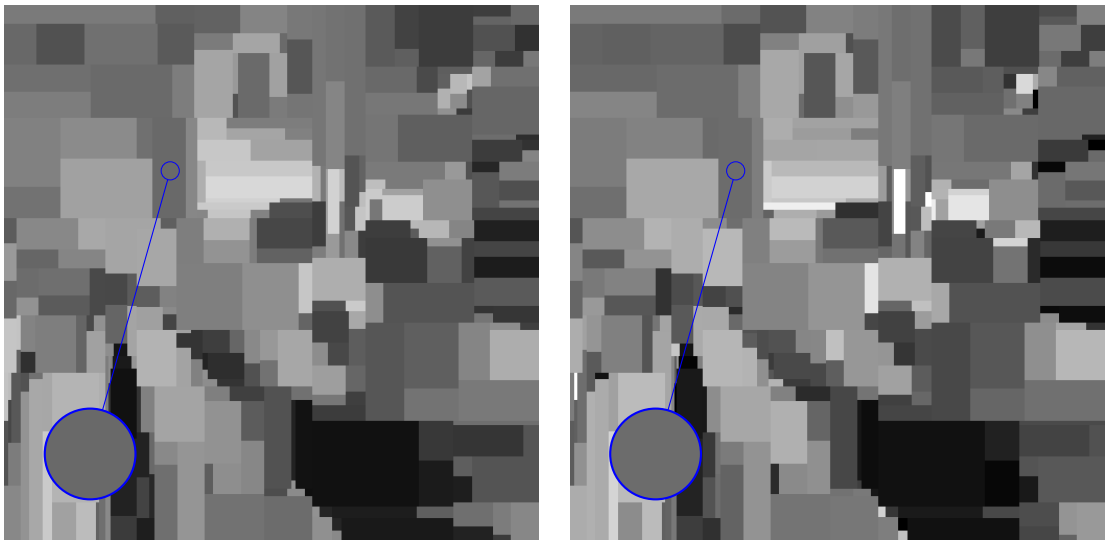
minimal penalties (limit)
expression
$\epsilon^2 \frac{\lvert m\rvert}{C_{ext}} \left[2\log(\frac{0.48n}{\lvert m\rvert}) - \log(2\log(\frac{0.48n}{\lvert m\rvert})) - \frac{\log(\lvert m\rvert)+2}{\lvert m\rvert}\right]$

Table 12 – Data for image face example face.2fold.(512, 512)



(a) Source image

(b) Noisy image



(c) Source image projected on near oracle of dimension 301

(d) Noisy image projected on near oracle of dimension 301

Figure 37 – Image face300binom2fisourceimage.png

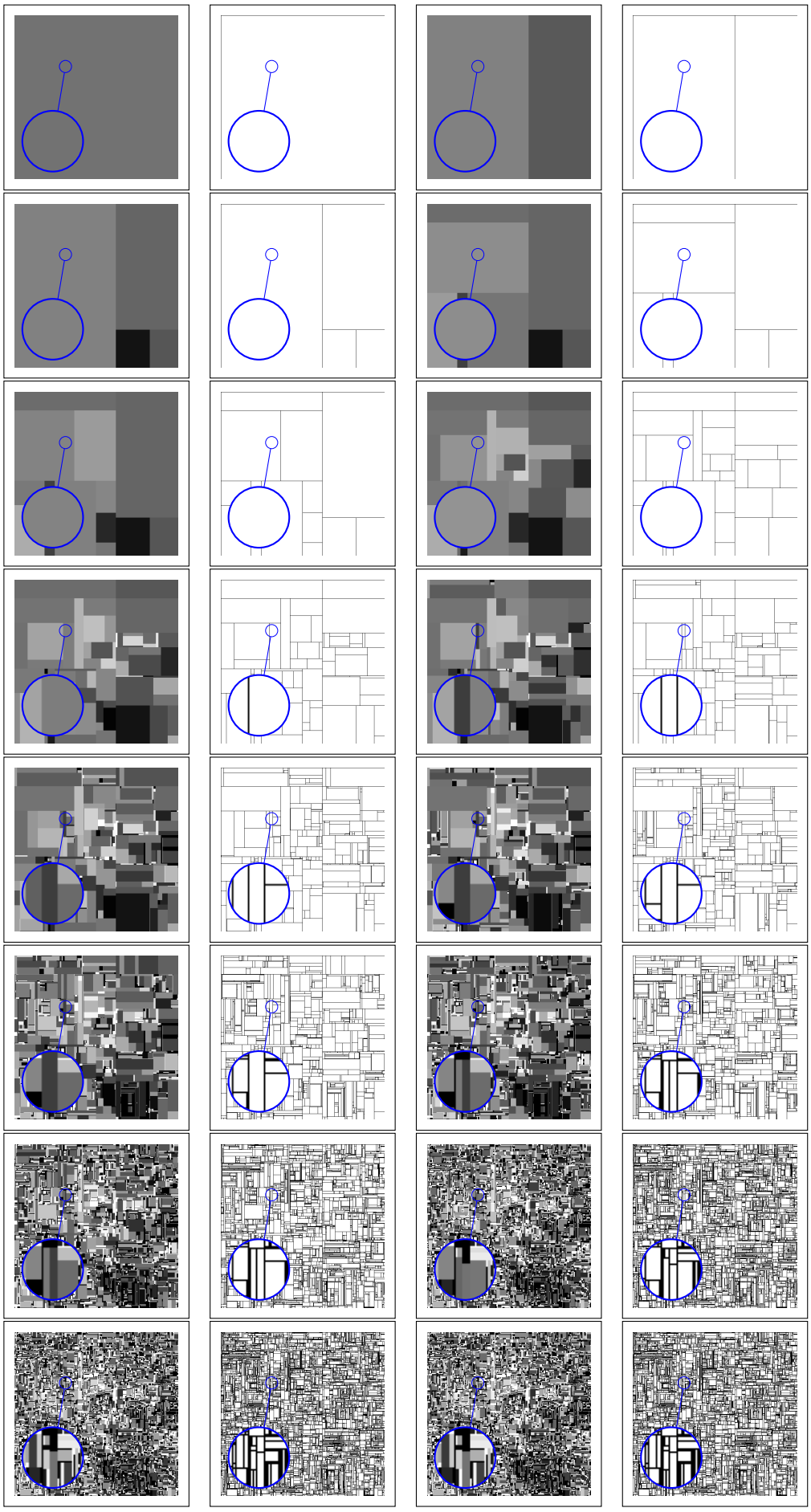


Figure 38 – Segmentation path of noisy image, at dimensions [1, 2, 4, 8, 13, 32, 64, 124, 254, 512, 1022, 2047, 4090, 8189, 9997, 9998](estimates and borders)

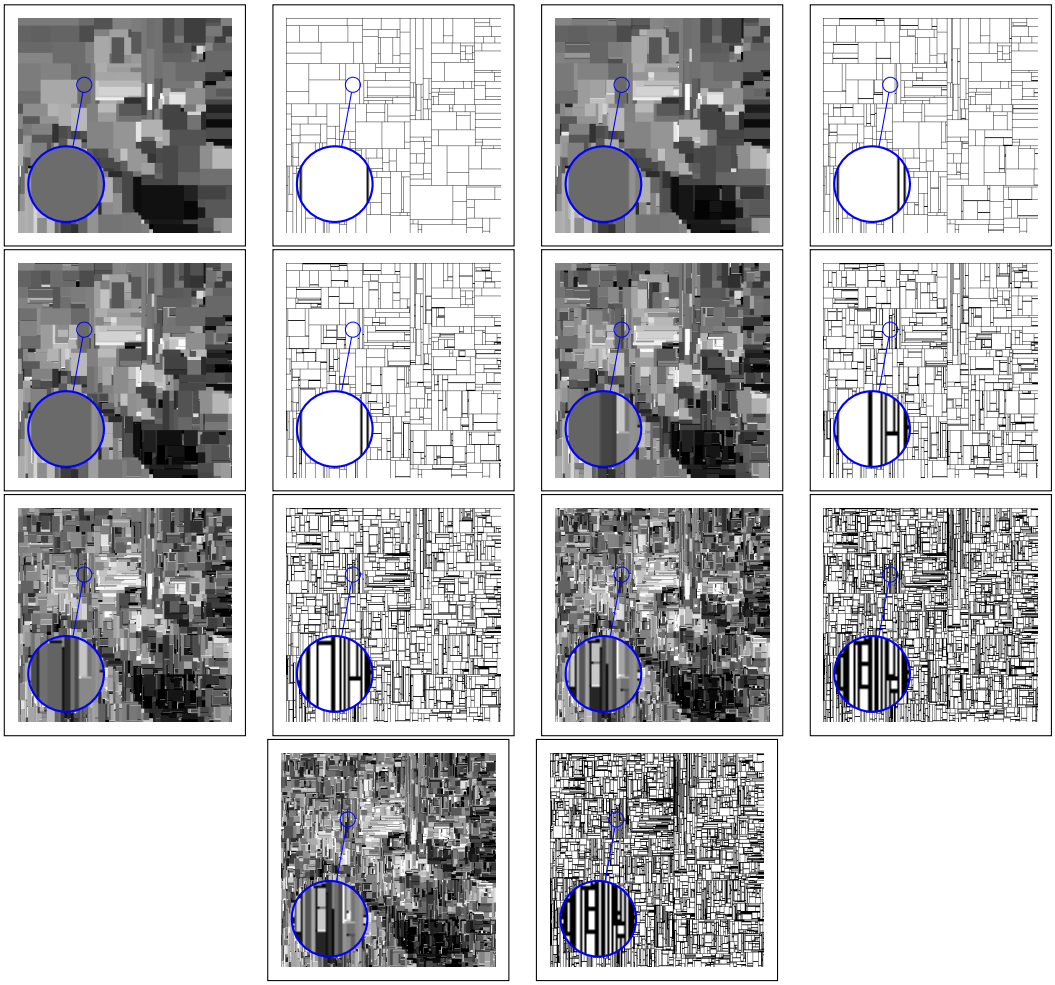


Figure 39 – Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [301, 513, 1024, 2049, 4098, 8195, 10000](estimates and borders)

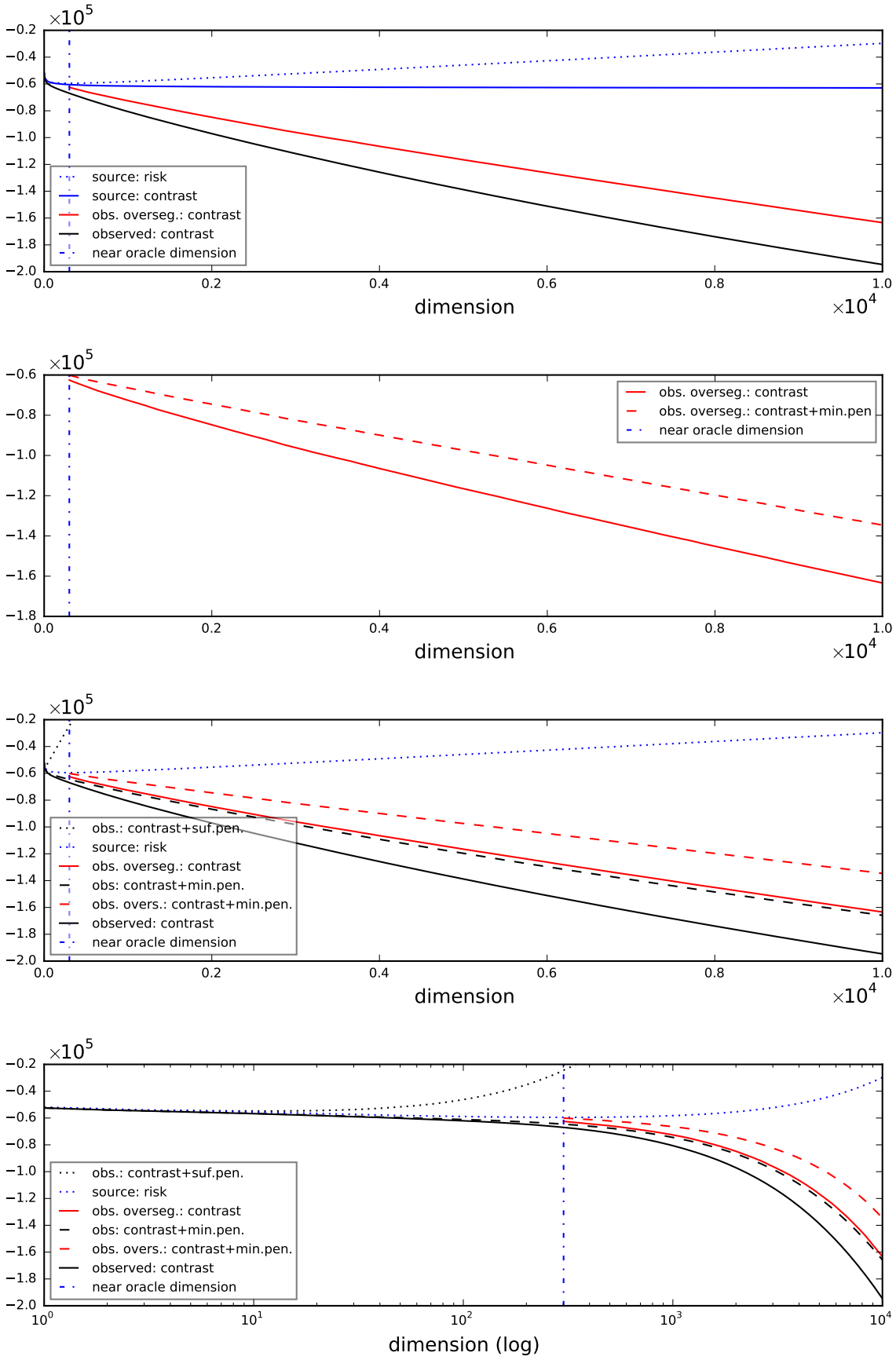


Figure 40 – contrasts and penalized contrast: segmentation paths of source and observed (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties.

8.5.10 Image face, free binary tree

dimensions				std. dev.	
size	height	width	near oracle	source	noise
262, 144	512	512	1002	0.21	0.86

splitting		maximum dimension increase	
mode	max branching factor	to isolate a point (in this image size)	per split
free binary tree 	$b = 2$	$C_{ext} = 4$	$b - 1 = 1$

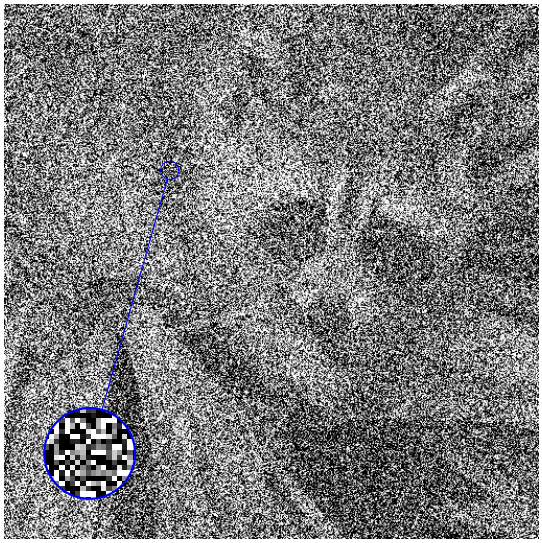
algorithm		
growing method	contrast criterion	max. dim.
impurity	l2	10000

sufficient penalties				
type	expression	L_m	$B_{\mathcal{M}}$	θ
binom	$\epsilon^2 \lvert m\rvert \left[1 + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1} L_m\right]$	$\log(2) + B_{\mathcal{M}} \log(\frac{en}{\lvert m\rvert})$	2	1.00

minimal penalties (limit)
expression
$\epsilon^2 \frac{\lvert m\rvert}{C_{ext}} \left[2\log(\frac{0.48n}{\lvert m\rvert}) - \log(2\log(\frac{0.48n}{\lvert m\rvert})) - \frac{\log(\lvert m\rvert)+2}{\lvert m\rvert}\right]$

Table 13 – Data for image face example face.2fold.(512, 512)

(a) Source image



(b) Noisy image

(c) Source image projected on near oracle of dimension 1002

(d) Noisy image projected on near oracle of dimension 1002

Figure 41 – Image face1000binom2fisourceimage.png

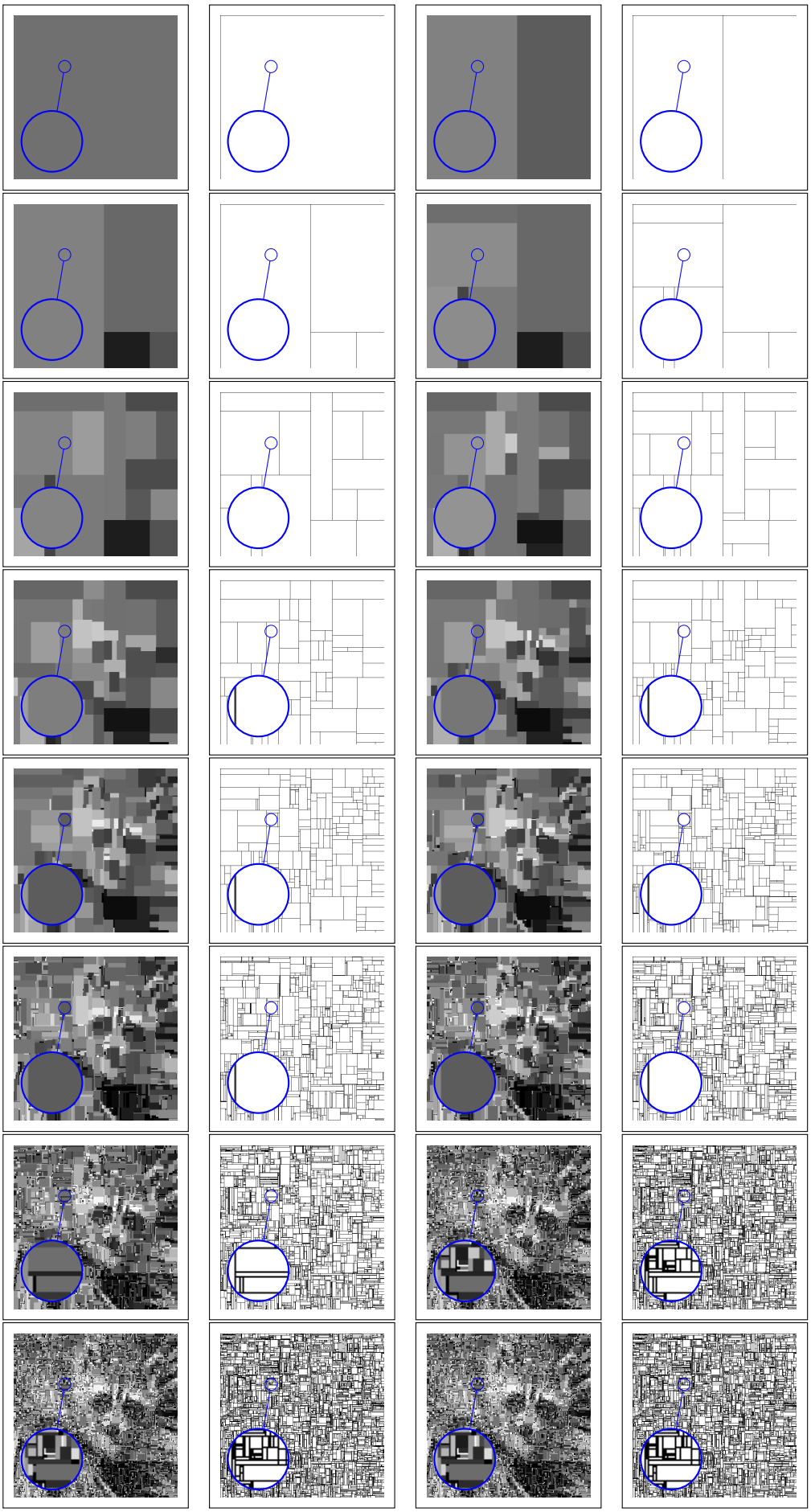


Figure 42 – Segmentation path of noisy image, at dimensions [1, 2, 4, 8, 16, 32, 64, 127, 252, 504, 1013, 2037, 4093, 8192, 9996, 9997](estimates and borders)

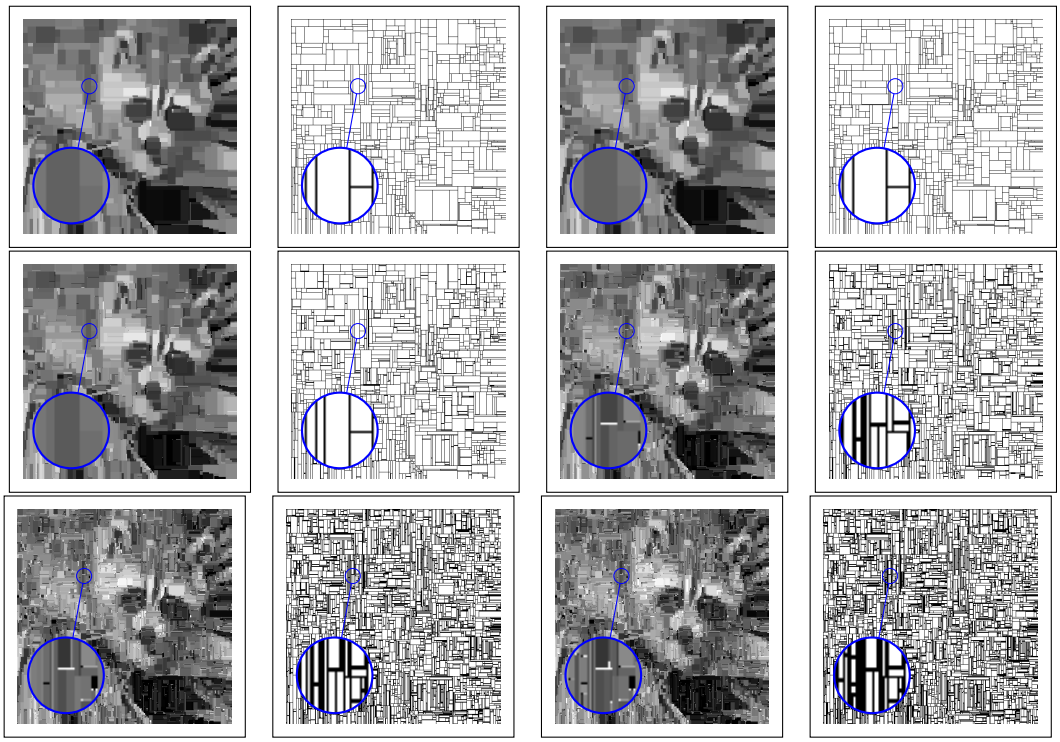


Figure 43 – Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [1002, 1024, 2050, 4096, 8192, 10000](estimates and borders)

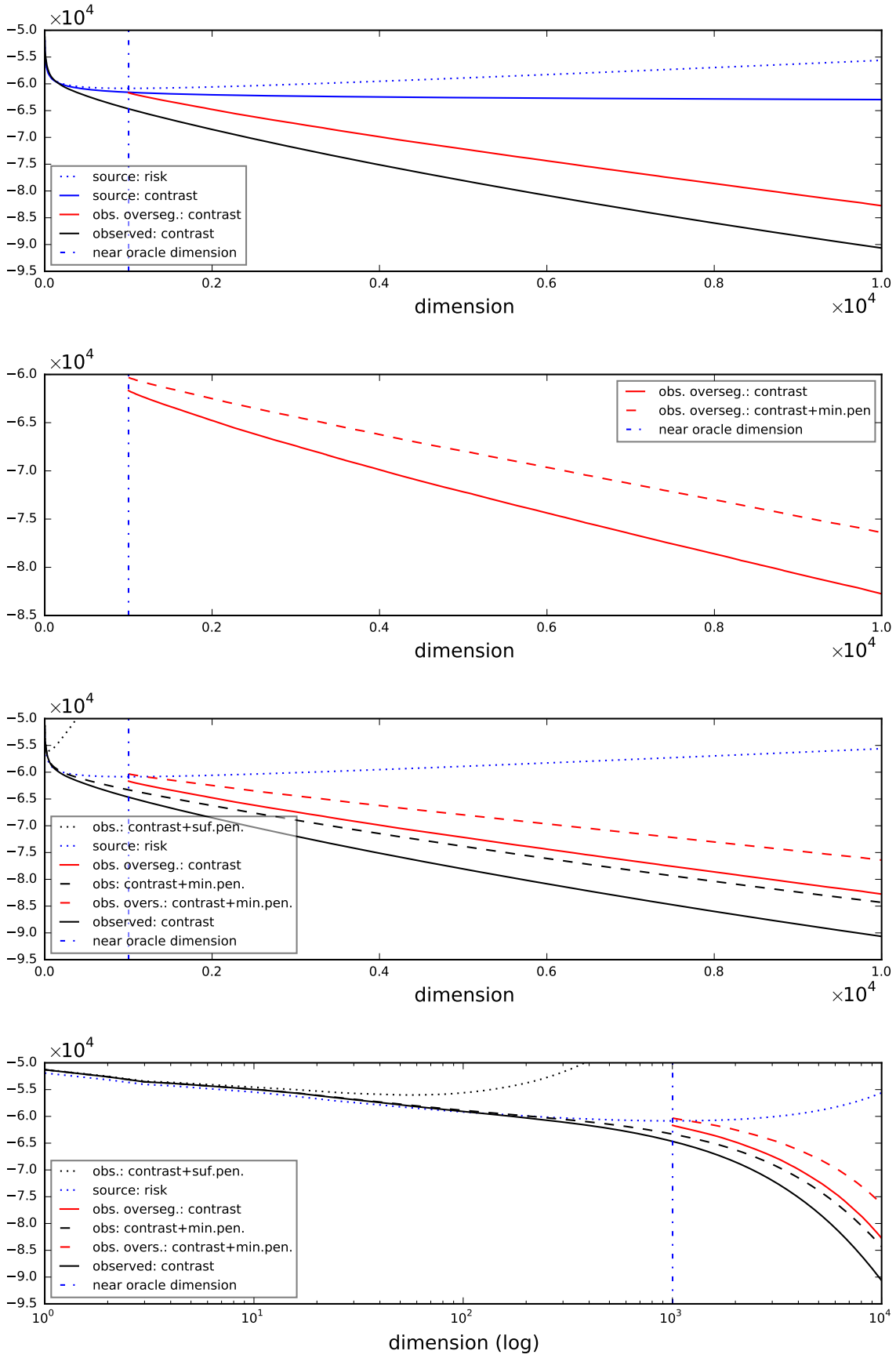


Figure 44 – contrasts and penalized contrast: segmentation paths of source and observed (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties.

8.5.11 Image face, free binary tree

dimensions				std. dev.	
size	height	width	near oracle	source	noise
262, 144	512	512	3003	0.21	0.44

splitting		maximum dimension increase	
mode	max branching factor	to isolate a point (in this image size)	per split
free binary tree 	$b = 2$	$C_{ext} = 4$	$b - 1 = 1$

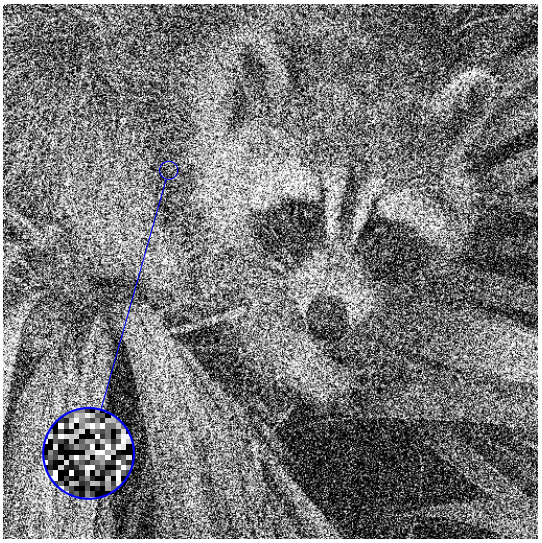
algorithm		
growing method	contrast criterion	max. dim.
impurity	l2	10000

sufficient penalties				
type	expression	L_m	$B_{\mathcal{M}}$	θ
binom	$\epsilon^2 \lvert m \rvert \left[1 + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1} L_m\right]$	$\log(2) + B_{\mathcal{M}} \log(\frac{en}{\lvert m \rvert})$	2	1.00

minimal penalties (limit)
expression
$\epsilon^2 \frac{\lvert m \rvert}{C_{ext}} \left[2\log(\frac{0.48n}{\lvert m \rvert}) - \log(2\log(\frac{0.48n}{\lvert m \rvert})) - \frac{\log(\lvert m \rvert)+2}{\lvert m \rvert}\right]$

Table 14 – Data for image face example face.2fold.(512, 512)

(a) Source image



(b) Noisy image

(c) Source image projected on near oracle of dimension 3003

(d) Noisy image projected on near oracle of dimension 3003

Figure 45 – Image face3000binom2fisourceimage.png

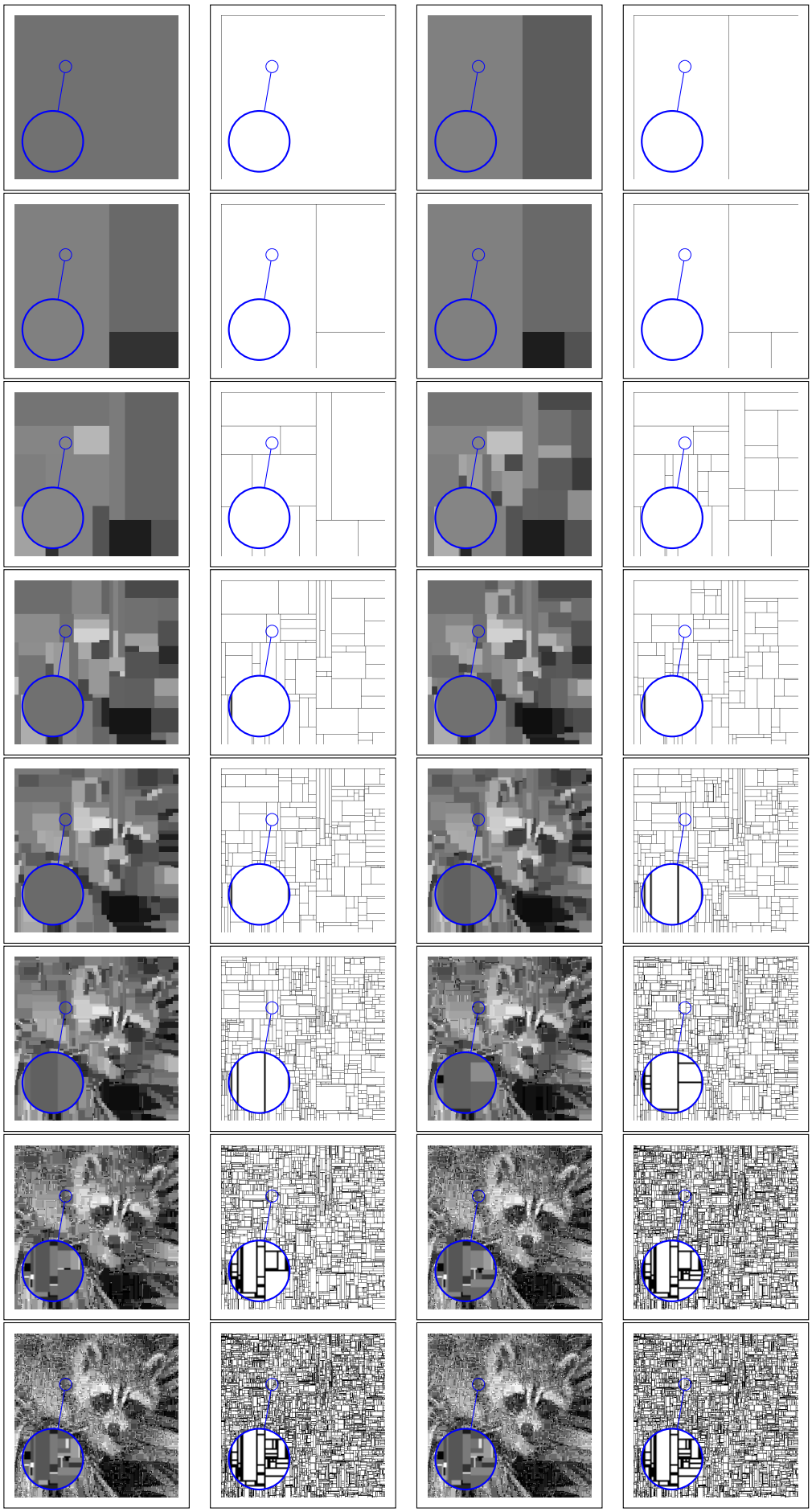


Figure 46 – Segmentation path of noisy image, at dimensions [1, 2, 3, 4, 13, 32, 64, 128, 256, 512, 1021, 2045, 4096, 8192, 9998, 9999](estimates and borders)

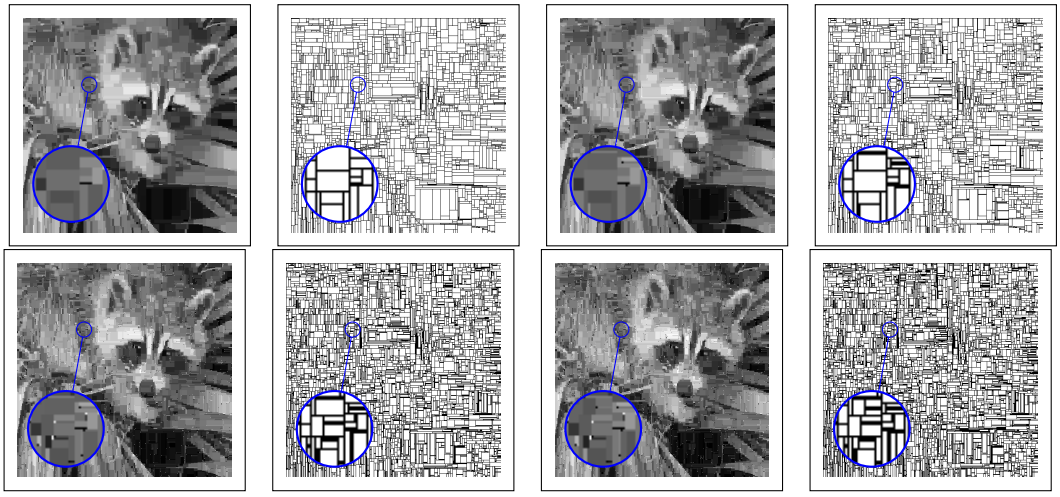


Figure 47 – Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [3003, 4098, 8194, 10000](estimates and borders)

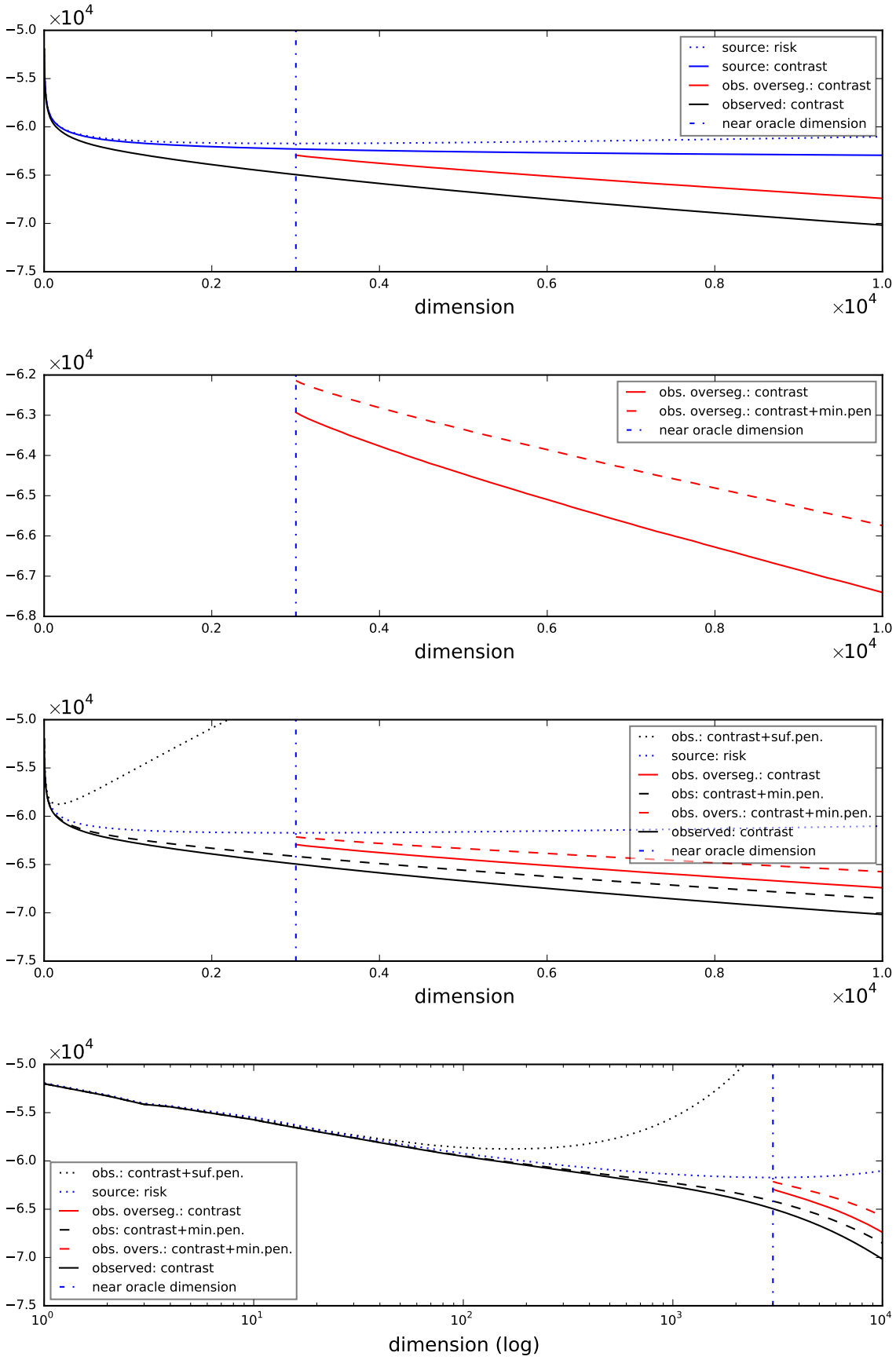


Figure 48 – contrasts and penalized contrast: segmentation paths of source and observed (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties.

8.5.12 Image face, free binary tree

dimensions				std. dev.	
size	height	width	near oracle	source	noise
262, 144	512	512	5001	0.21	0.33

splitting		maximum dimension increase	
mode	max branching factor	to isolate a point (in this image size)	per split
free binary tree 	$b = 2$	$C_{ext} = 4$	$b - 1 = 1$

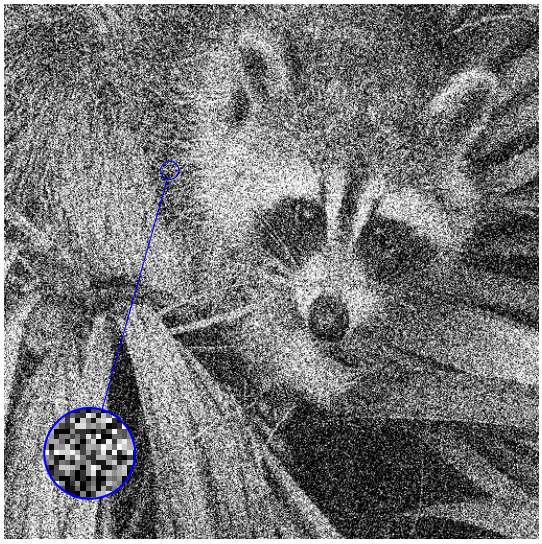
algorithm		
growing method	contrast criterion	max. dim.
impurity	l2	10000

sufficient penalties				
type	expression	L_m	$B_{\mathcal{M}}$	θ
binom	$\epsilon^2 \lvert m\rvert \left[1 + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1} L_m\right]$	$\log(2) + B_{\mathcal{M}} \log(\frac{en}{\lvert m\rvert})$	2	1.00

minimal penalties (limit)
expression
$\epsilon^2 \frac{\lvert m\rvert}{C_{ext}} \left[2\log(\frac{0.48n}{\lvert m\rvert}) - \log(2\log(\frac{0.48n}{\lvert m\rvert})) - \frac{\log(\lvert m\rvert)+2}{\lvert m\rvert}\right]$

Table 15 – Data for image face example face.2fold.(512, 512)

(a) Source image



(b) Noisy image

(c) Source image projected on near oracle of dimension 5001

(d) Noisy image projected on near oracle of dimension 5001

Figure 49 – Image face5000binom2fisourceimage.png

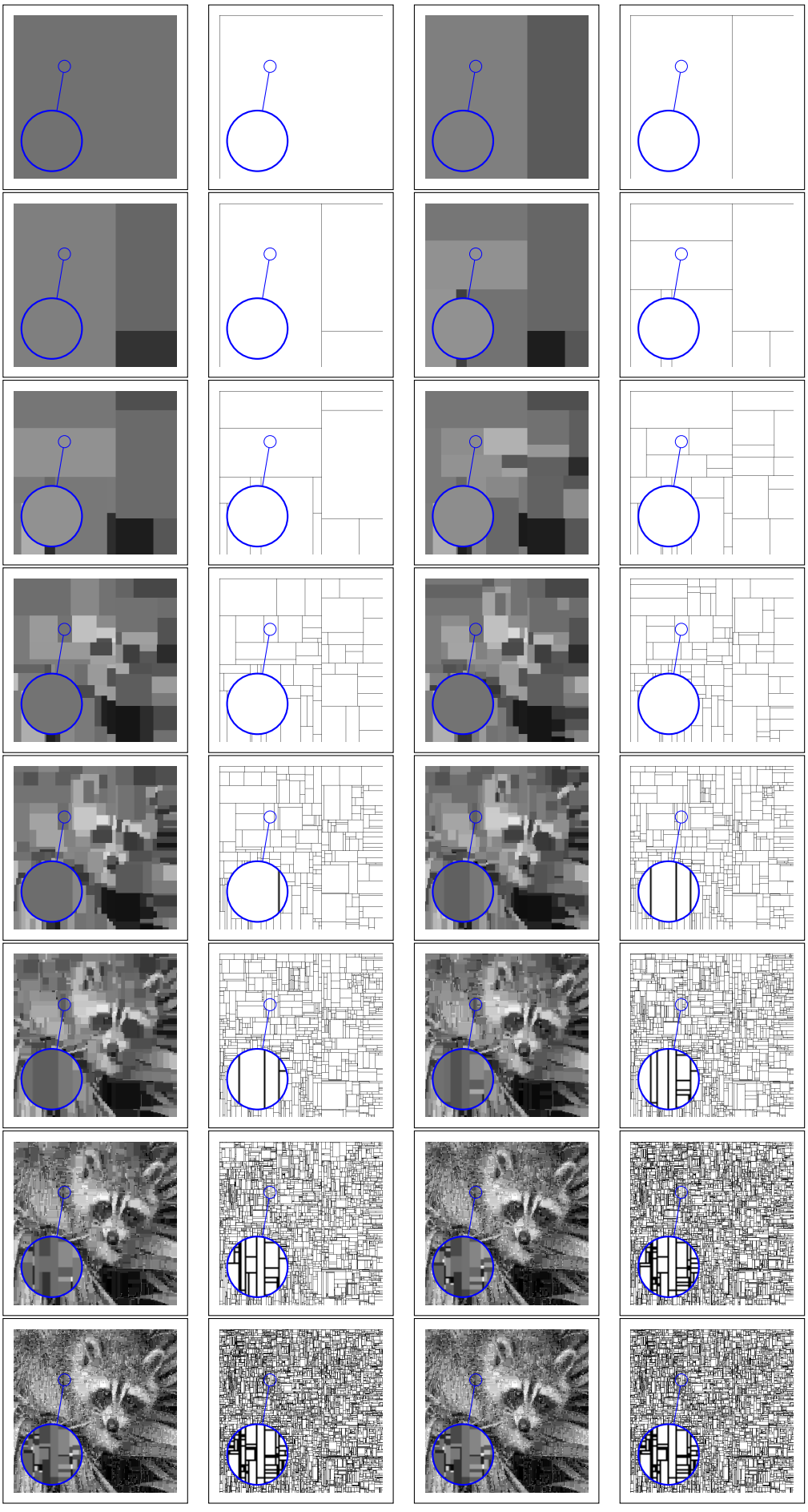


Figure 50 – Segmentation path of noisy image, at dimensions [1, 2, 3, 8, 14, 32, 64, 127, 256, 511, 1022, 2047, 4095, 8188, 9993, 9994](estimates and borders)

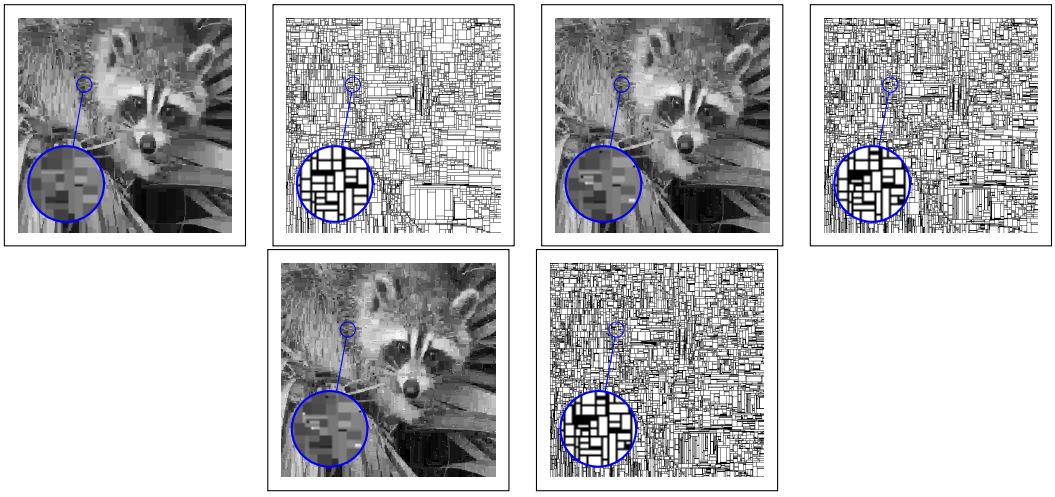


Figure 51 – Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [5001, 8194, 10001](estimates and borders)

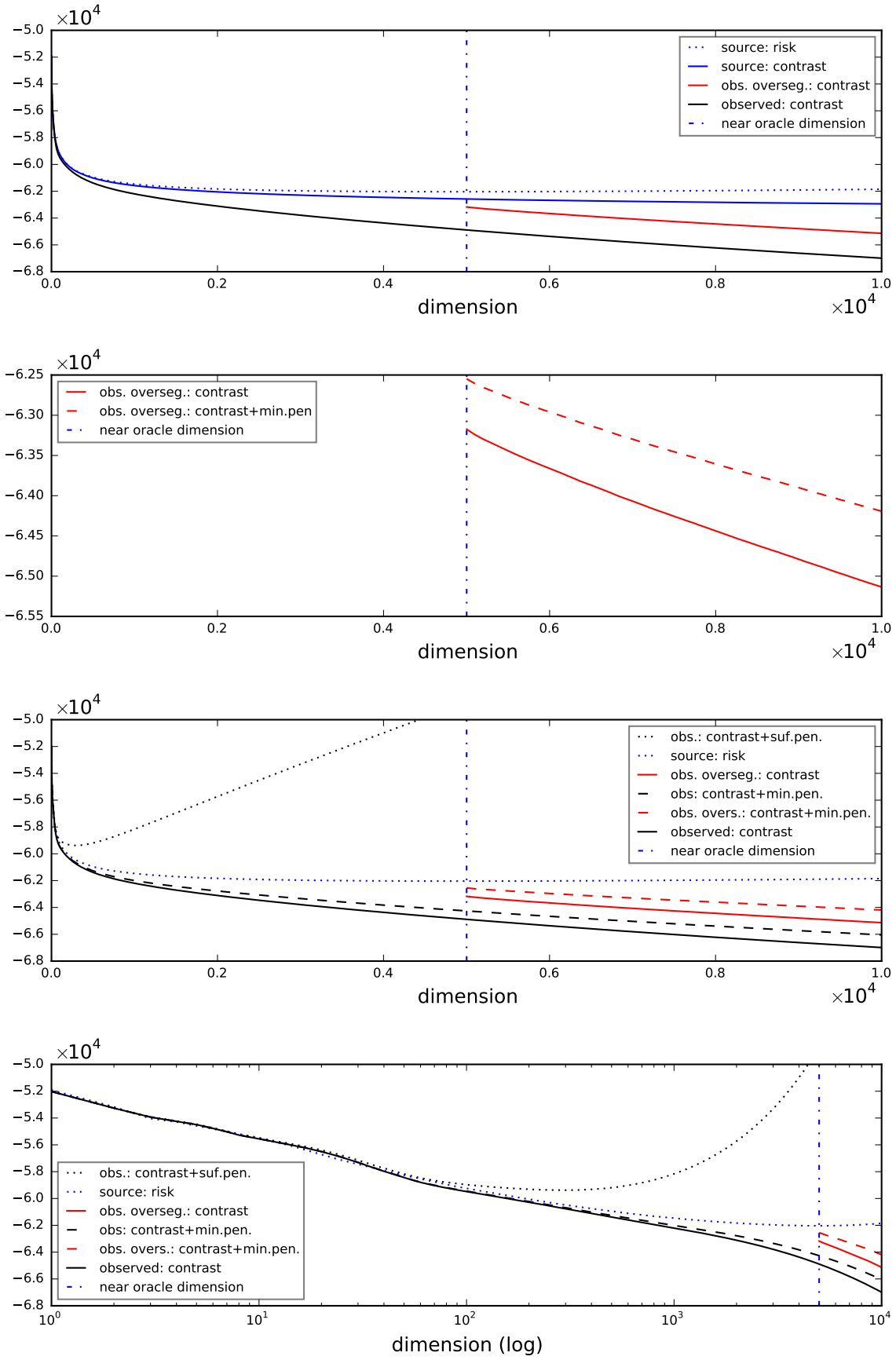


Figure 52 – contrasts and penalized contrast: segmentation paths of source and observed (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties.

9 A first extension to histogram selection for density estimation

In the previous sections we developed a method to prove the occurrence of the minimal penalty phenomenon in certain situations of Gaussian signal partitioning and segmentation.

Oracle inequalities are also available for model selection in the case of density estimation along a real interval. A theorem of G. Castellan for histogram selection shows that sufficient penalties must satisfy:

$$pen(m) \geq \frac{D_m}{n}\left(\sqrt{c_1} + \sqrt{2(1+c_1)L_m}\right)^2, \text{ for all } m \in \mathcal{M},$$

where c_1 denotes any constant larger than $\frac{1}{2}$, $D_m + 1$ denotes the number of elements of the partition m, and $\{L_m\}_{m\in\mathcal{M}}$ denotes a sequence of weights with $\sum_{m\in\mathcal{M}} e^{-D_m L_m} < \infty$ (see Theorem 9.7 or [Mas07, Th.7.9 p.232] or [Cas99]).

G. Castellan on one hand proves that the choice of $c_1 = 1$ optimizes the risk bound, and on the other hand shows that choosing $c_1 < \frac{1}{2}$ can lead to trouble, with a lower bound for the Kullbak-Leibler risk in the case of irregular histogram selection when $c_1 < \frac{1}{2}$, although under the assumption that there is only one model per dimension (see Theorem 9.8 or [Mas07, Th.7.10 p.238]).

The purpose of this section is to show, at least in the null hypothesis, that a similar phenomenon takes place with a model family of exponential complexity under certain customary geometrical assumptions.

9.1 Technical preleminaries: model selection for density estimation with histograms

In the framework of density estimation, one considers a distribution $P = s\mu$ where μ denotes a known normalized measure on some set U, the *underlying space*. The unknown density s plays the role of the source signal to reconstruct. Under the null hypothesis, s is equal to 1 almost everywhere.

Often the underlying space U is a product of intervals in an Euclidean space, or simply the real interval $[0, 1]$, and μ is the normalized Lebesgue measure on U.

One observes a sample $\xi_1, \ldots, \xi_n$ of n independent random variable with common distribution $s\mu$, defining an *empirical measure* on the underlying space by: $P_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(\xi_i)$ for any real function f on U.

We consider the task of estimating a density by histograms (step functions) based on a specified finite collection of models identified to a family $\mathcal{M}$ of partitions of the underlying space U in measurable subsets of positive measure (i.e. parts, bins, or regions). We keep the same definitions and notation regarding partitions than in Section 7.1 above. The following definition sets out how a partition relates to an estimator:

Definition 9.1. *(Histogram estimator)*

For any $m \in \mathcal{M}$, define $|m|$ as the number of parts of m, namely its cardinal.

Also define for any $m \in \mathcal{M}$, respectively the *histogram estimator* and the *histogram projection* of s, based on m:

$$\hat{s}_m = \sum_{\tau \in m} \frac{\mathbb{P}_{\mathrm{n}}(\tau)}{\mu(\tau)} \mathbb{1}_\tau$$
$$s_m = \sum_{\tau \in m} \frac{\mathrm{P}(\tau)}{\mu(\tau)} \mathbb{1}_\tau$$

As with Gaussian model selection, we need a contrast criterion:

Definition 9.2. *(log-likelihood criterion)* For any probability density t define its empirical contrast $\gamma_n(t)$ as

$$\begin{aligned}\gamma_n(t) &:= \mathbb{P}_{\mathrm{n}}(-\log(t)) \\ &:= -\int \hat{s}_m \log(t)\, \mathrm{d}\,\mu,\end{aligned}$$

Consider some penalty function pen : $\mathcal{M} \longrightarrow \mathbb{R}^+$,

Let the selected model $\hat{m}$ be, when it exists, the minimizer of the penalized log-likelihood criterion over $m \in \mathcal{M}$:

$$\begin{aligned}\mathrm{crit}(\hat{s}_m) &:= \gamma_n(\hat{s}_m) + \mathrm{pen}(m), \\ &:= -\int \hat{s}_m \log \hat{s}_m \,\mathrm{d}\,\mu + \mathrm{pen}(m).\end{aligned}$$

Remark 9.3. Although we kept the same definition of $|m|$ than in discrete signal partition, the effective dimension of the corresponding model in now $|m| - 1$ as opposed to $|m|$ previously, by the normalising constraint on the estimated density.

Remark 9.4. Since

$$\int \hat{s}_m \,\mathrm{d}\,\mu = 1,$$

the criterion above is equivalent to

$$\mathrm{crit}(\hat{s}_m) = -\int h(\hat{s}_m - 1)\,\mathrm{d}\,\mu + \mathrm{pen}(m)$$

where (see Definition 14.1) for any x with $-1 < x$:

$$h(x) := (1+x)\log(1+x) - x$$

In view of what follows, we recall the definitions of the *Kullback-Leibler divergence* and of the *Hellinger distance*.

Definition 9.5 (Kullback-Leibler information number)**.** For two probability densities s and t with respect to a measure μ, define the Kullback-Leibler information number $\mathbf{K}(s,t)$ between the densities s and t as

$$\mathbf{K}(s,t) = \int s \log\left(\frac{s}{t}\right) \mathrm{d}\,\mu$$

if $s\mu$ is absolutely continuous with respect to $t\mu$ and $s\log\left(\frac{s}{t}\right) \in L_1(\mu)$, and

$$\mathbf{K}(s,t) = \infty$$

otherwise.

Definition 9.6 (Hellinger distance)**.** For two probability densities s and t with respect to a measure μ, define the Hellinger distance $\mathbf{h}(s,t)$ between the densities s and t by

$$\mathbf{h}^2(s,t) = \frac{1}{2}\int \left(\sqrt{s} - \sqrt{t}\right)^2 \mathrm{d}\,\mu$$

9.1.1 Castellan's histogram selection theorem for density estimation

The following theorem is due to G. Castellan (see [Cas99] or [Mas07, Theorem 7.9 p.232]). It offers a threshold for sufficient penalties in the situation of density estimation by histogram selection, in a manner analog to the Gaussian model selection Theorem 5.6 when specified to signal segmentation. The penalized contrast is based on the log-likelihood. A set of weights x_m on the histogram family $\mathcal{M}$ is introduced with the condition that there exists Σ with $\sum_{m\in\mathcal{M}} e^{-x_m} \leq \Sigma < \infty$. and a limiting condition for sufficient penalties is the following, with $c_1 > \frac{1}{2}$:

$$\mathrm{pen}\,(m) \geq \frac{c_1}{n}\left(\sqrt{D_m} + \sqrt{2(1+c_1^{-1})x_m}\right)^2, \quad \forall m \in \mathcal{M}.$$

Expressing the weights as $x_m = L_m D_m, \forall m \in \mathcal{M}$ leads to the borderline formula mentioned in introduction:

$$\mathrm{pen}\,(m) \geq c_1 \frac{D_m}{n}\left(1 + \sqrt{2(1+c_1^{-1})L_m}\right)^2, \quad \forall m \in \mathcal{M}.$$

Castellan's theorem offers a control over the Hellinger risk of the selected estimator $\mathbf{h}^2(s,\tilde{s})$, which is very close to its Kullback-Leibler counterpart $\mathbf{K}(s,\tilde{s})$.

Theorem 9.7. *[Mas07, Theorem 7.9 p.232] Let $\xi_1, \cdots, \xi_n$ be some independent $[0,1]$-valued random variables with common distribution $P = s\mu$, where μ denotes the Lebesgue measure. Consider a finite family $\mathcal{M}$ of partitions satisfying:*

- *there is an integer N such that $N \leq n\log(n)^{-2}$ and m_N a partition of $[0,1]$ the elements of which are intervals with equal length $(N+1)^{-1}$,*
- *every element of any partition m belonging to $\mathcal{M}$ is the union of pieces of m_N.*

Let for every partition m in $\mathcal{M}$

$$\hat{s}_m = \sum_{I\in m} \frac{P_n(I)}{\mu(I)}\mathbb{1}_I \quad \textit{and} \quad s_m = \sum_{I\in m} \frac{P(I)}{\mu(I)}\mathbb{1}_I$$

be respectively the histogram indicator and the histogram projection of s, based on m. Consider some absolute constant Σ and some family of non-negative weights $\{x_m\}_{m\in\mathcal{M}}$ such that

$$\sum_{m\in\mathcal{M}} e^{-x_m} \leq \Sigma$$

Let $c_1 > \frac{1}{2}$, $c_2 = 2(1+c_1^{-1})$ and consider some penalty function $\mathrm{pen} : \mathcal{M} \to \mathbb{R}^+$ such that

$$\mathrm{pen}\,(m) \geq \frac{c_1}{n}\left(\sqrt{D_m} + \sqrt{c_2 x_m}\right)^2, \forall m \in \mathcal{M},$$

where $D_m + 1$ denotes the number of elements of a partition m. Let $\hat{m}$ be the minimizer over $m \in \mathcal{M}$ of the penalized log-likelihood criterion

$$-\int \hat{s}_m \log(\hat{s}_m)\,\mathrm{d}\,\mu + \mathrm{pen}\,(m)$$

and define the penalized maximum likelihood estimator by $\tilde{s} = s_{\hat{m}}$. If for some positive real number ρ, $s \geq \rho$ almost everywhere and $\int s(\log s)^2\,\mathrm{d}\,\mu \leq L < \infty$, then for some constant $C(c_1, \rho, L, \Sigma)$,

$$\frac{1}{2}\mathbb{E}\left[\mathbf{h}^2(s,\tilde{s})\right] \leq \frac{(2c_1)^{1/5}}{(2c_1)^{1/5}-1}\inf_{m\in\mathcal{M}}\left\{\mathbf{K}(s,s_m) + \mathrm{pen}\,(m)\right\} + \frac{C(c_1,\rho,L,\Sigma)}{n}.$$

If for instance the source density s belongs to some model $\bar{m} \in \mathcal{M}$, then $\mathbf{K}(s, s_m) = 0$ and under the assumption of Theorem 9.7,

$$\frac{1}{2}\mathbb{E}\left[\mathbf{h}^2(s,\tilde{s})\right] \leq \frac{(2c_1)^{1/5}}{(2c_1)^{1/5}-1}\frac{c_1}{n}\left(\sqrt{D_{\bar{m}}} + \sqrt{c_2 x_{\bar{m}}}\right)^2 + \frac{C(c_1,\rho,L,\Sigma)}{n}.$$

so that $\mathbb{E}\left[\mathbf{h}^2(s,\tilde{s})\right] \leq \frac{O(1)}{n}$, which the risk behaviour of a predefined model m_0 containing the source density s, since for instance within the assumption on Theorem, when $n \to \infty$,

$$\begin{aligned}\mathbb{E}\left[\mathbf{h}^2(s,\hat{s}_{m_0})\right] &= \mathbf{h}^2(s,\hat{s}_{m_0}) + \frac{D_{m_0}}{8n}\left[1+o(1)\right],\\ &= \frac{D_{m_0}}{8n}\left[1+o(1)\right] \qquad (9.1)\end{aligned}$$

(see [BR06, Theorem 1] quoted in [Cas99, 2.2.3 p 8]).

9.1.2 Castellan's minimal penalty theorem for density estimation

The following result by G. Castellan complements Theorem 9.7 by showing that the choice of $c_1 < \frac{1}{2}$ can lead to very bad results.

Theorem 9.8. *[Mas07, Theorem 7.10 p.238] Let $\xi_1, \cdots, \xi_n$ be some independent $[0,1]$-valued random variables with common distribution $P = s\mu$ with $s = \mathbb{1}_{[0,1]}$. Consider some finite family of partitions $\mathcal{M}$ such that for each integer D, there exists only one partition m such that $D_m = D$. Moreover assume that $\mu(I) \geq \log(n)^2/n$ for every $I \in m$ and $m \in \mathcal{M}$.*

Assume that for some partition $m_N \in \mathcal{M}$ with $N+1$ pieces one has

$$\mathrm{pen}\,(m_N) = c\frac{N}{n}$$

with $c < \frac{1}{2}$. Let $\hat{m}$ be the minimizer over $\mathcal{M}$ of the penalized log-likelihood criterion

$$-\int \hat{s}_m \log(\hat{s}_m)\,\mathrm{d}\,\mu + \mathrm{pen}\,(m)\,.$$

Then, whatever the values of $\mathrm{pen}\,(m)$ *for $m \neq m_n$, there exists positive numbers N_0, L and $C(c)$, depending only on c, such that, for all $N \geq N_0$,*

$$\mathbb{P}\left[D_{\hat{m}} \geq \frac{1-4c^2}{4}N\right] \geq 1-\beta(c),$$

where

$$\beta(c) = \Sigma(L)\exp\left[-\frac{L}{2}((1-4c^2)N)^{\frac{1}{2}}\right] + \frac{C(c)}{n} \text{ with } \Sigma(L) = \sum_{D\geq 1} e^{-L\sqrt{D}}.$$

Moreover, if $\tilde{s} = \hat{s}_{\hat{m}}$,

$$\mathbb{E}\left[\mathbf{K}(s,\tilde{s})\right] \geq \delta(c)\left[1-\beta(c)\right]\frac{N}{n}$$

where $\delta(c) = (1-2c)(1+2c)^2/16$.

As a result taking $\mathrm{pen}\,(m) = c\frac{D_m}{n}$ for some $c < \frac{1}{2}$ leads to a disaster in the sense that, if the true signal s is uniform, the model selection procedure will choose high dimensional models with high probability, and the estimator's normalized risk $n\,\mathbb{E}\left[\mathbf{K}(s,\tilde{s})\right]$ will be bounded away from 0 when n goes to infinity. Taking for instance $N = \lceil\sqrt{n}\rceil$ ensures that simultaneously

$$\mathbb{P}\left[D_{\hat{m}} \geq \frac{1-4c^2}{4}\sqrt{n}\right] = 1 - O\left(\frac{1}{n}\right),$$

and

$$n\,\mathbb{E}\left[\mathbf{K}(s,\tilde{s})\right] \geq \left[1-O\left(\frac{1}{n}\right)\right]\frac{1-4c^2}{16}\sqrt{n}.$$

Theorem 9.7 relies on the low complexity assumption that there is only a single model per dimension in the model family $\mathcal{M}$. We will try in Section 9 to address an analog situation under high (binomial) complexity.

9.2 Definitions and main assumptions

The following assumption specifies the type of partition family used for the model selection procedure, it is taken from Castellan's Theorem 9.7:

Assumption 9.9. *For what follows, we assume specified a positive integer N and m_N a partition of $[0,1]$ in $N+1$ parts of equal measure $\frac{1}{N+1}$.*

We assume that every partition element of the family $\mathcal{M}$ has for parts unions of the parts of m_N.

The most immediate example of such a partition is a segmentation of $[0,1]$ in $N+1$ segments of equal length. Under the null hypothesis, for any part $\tau \in m_N$ of measure $\frac{1}{N+1}$, the (random) *occupancy count* $N_\tau := n\,\mathbb{P}_{\mathrm{n}}(\tau) = n\int_\tau \hat{s}_m \,\mathrm{d}\,\mu$ follows a binomial law $\mathcal{B}(n, \frac{1}{N+1})$. We denote $\bar{k}$ the *expected occupancy count* $\bar{k} := \frac{n}{N+1}$. In order to facilitate approximations of the binomial tail, we make a light additional assumption on the number N. This assumption is part of Assumption H_0 in Castellan's histogram selection Theorem 9.7.

Assumptions on the number of parts For what follows, we will make the following assumption, expressing that a large enough number of sample points is expected in each of the $N+1$ parts:

Assumption 9.10. *Assume that*

$$N \leq \frac{n}{\log(n)^2}$$

Structure assumption on the model family As in the context of Gaussian signal partitioning (see Section 8), we make a structure assumption on the model family $\mathcal{M}$, with the following completion rule, stating that the null model can be extended by elementary parts with a limited increase in dimension:

Assumption 9.11 (completion rule)**.** *Assume there is a number $\mathrm{C}_{\mathrm{ext}} \geq 1$ so that for any subset b of the partition m_N, there is a model denoted m_b in the family $\mathcal{M}$ so that*

- *m_b extends b, in the sense that any part in the part subset b is also a part in the partition m_b,*
- *and $|m_b| \leq 1 + \mathrm{C}_{\mathrm{ext}}\,\mathrm{Card}(b)$.*

We think this geometrical situation is very customary, and the observations made in Section 8 for Gaussian segmentation apply in the same way.

Instances of histogram families over $[0,1]$ satisfying Assumption 9.11 It is customary that the partition m_N is formed of $N+1$ regularly spaced segments of $[0,1]$ of length $\frac{1}{N+1}$.

For multiple change detection in the parameter of a sequencially observed frequency, one often relies on free segmentation, where any partition element of $\mathcal{M}$ is formed of

segments which are unions of consecutive elements of m_N. In this case the number $\mathrm{C}_{\mathrm{ext}}$ is equal to 2, as discussed in a similar context in Section 8.1.1 and illustrated in Figure 53.

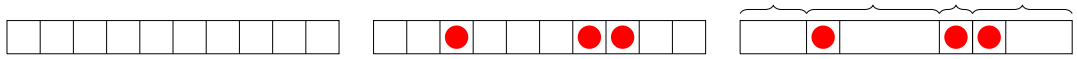


Figure 53 – Free line segmentation for histograms: instances of m_N, selected parts in m_N and completed segmentation. $\mathrm{C}_{\mathrm{ext}} = 2$

In the case of dyadic partitioning, the number of elementary parts $N+1$ introduced Assumption 9.9 satisfies $N+1 = 2^h$ for some $h \in \mathbb{N}$. It follows from Lemma 6.2 with $b = 2$ and the same value of the parameter h that in this case $\mathrm{C}_{\mathrm{ext}} \leq h$, as illustrated in Figure 54.

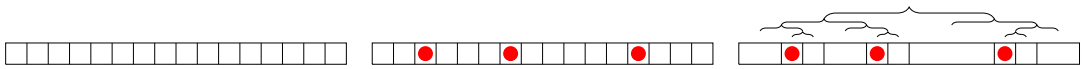


Figure 54 – Dyadic line segmentation for histograms with $h = 4$ and $N+1 = 2^4$: instances of m_N, selected parts in m_N and completed segmentation. $\mathrm{C}_{\mathrm{ext}} \leq (2-1)h = 4$

Instances of histogram families over $[0,1]^d$ satisfying Assumption 9.11 Notice that the essential ingredient in Assumption 9.9 is to ensure that all the considered partitions are made of unions of $N+1$ elementary parts of $[0,1]$ of equal measure $\frac{1}{N+1}$, regardless of their actual geometry. This allows to extend the results of this section to families of histograms over $[0,1]^d$ with $d > 1$ by only replacing $[0,1]$ by $[0,1]^d$ in the set of assumptions. In this case it is customary to take for m_N a set of homothetic regularly spaced rectangular blocs (see Figure 55), in other words a regular grid.

A common instance of partitionning method is the method of regression trees, described in more details in Section 7.1. In such a case, the extension factor satisfies $\mathrm{C}_{\mathrm{ext}} \leq 2d$, as illustrated in Figure 55.

The model family obtained by partionning into arbitrary rectangular blocs (bloc layouts) is a superset of the preceding one (regression trees), so that also in this case the extension factor satisfies $\mathrm{C}_{\mathrm{ext}} \leq 2d$, as illustrated in Figure 56.

Last a well known way to partition a domain in $\mathbb{R}^d$ is Voronoi's method. The following is an adaptation intended to respect Assumption 9.9. The element of the model family $\mathcal{M}$ are indexed by subsets of the set (grid) m_N. Informally, the partition m_I associated to such a subset $I \subset m_N$ is formed by grouping any element $\tau \in m_N$ with an element $\tau' \in I$ realising the minimum of $d(\tau, \tau')$, for some distance $d(.,.)$. Figure 57 offers an illustration. For each part $\tau \in m_N$, there is a smallest set of parts $I_\tau \subset m_N$ so that the part τ is isolated in the partition m_{I_τ}. In the exemple of Figure 57, the set of parts I_τ is formed of τ and its left, right, top and bottom immediate neihbours. Given a subset $b \in m_N$, the subset $I_b = \bigcup_{\tau \in b} I_\tau$ indexes an element of $\mathcal{M}$ associated with a partition where all the parts element of b are isolated, so that the model m_{I_b} extends b in the sense of Assumption 9.11. It follows that the number $\mathrm{C}_{\mathrm{ext}}$ is not larger than the quantity $\sup_{\tau \in m_N} \mathrm{Card}\, I_\tau$, which is $2d+1 = 5$ in the instance of Figure 57.

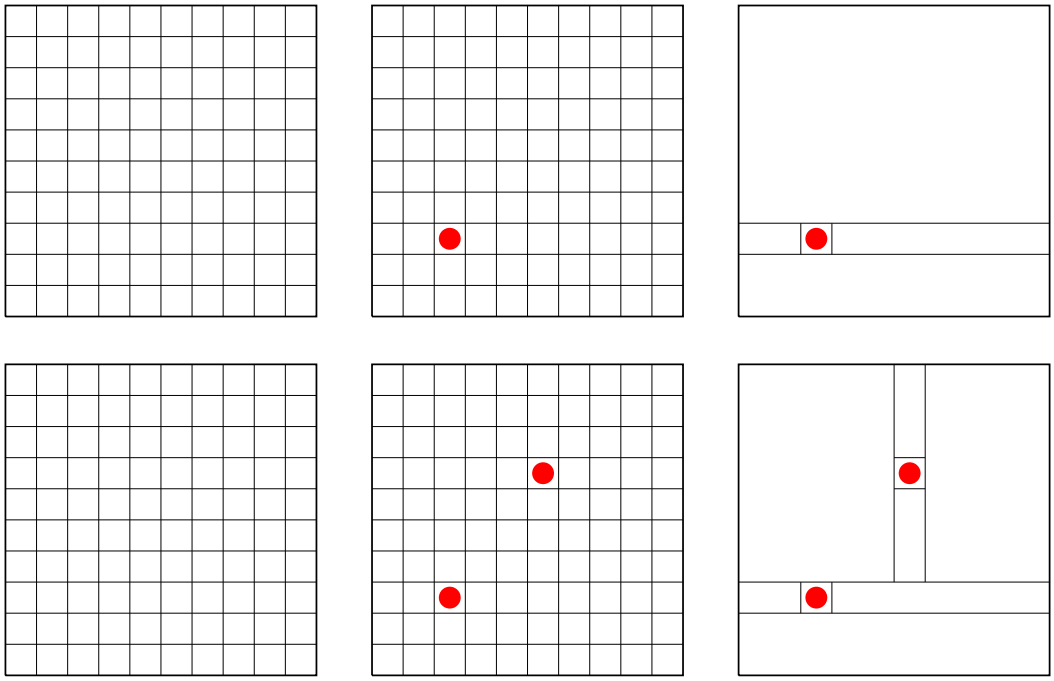


Figure 55 – Regression tree for histograms over $[0,1]^d$: m_N, selected parts in m_N, completed segmentations. $\mathrm{C}_{\mathrm{ext}} \leq 2d$

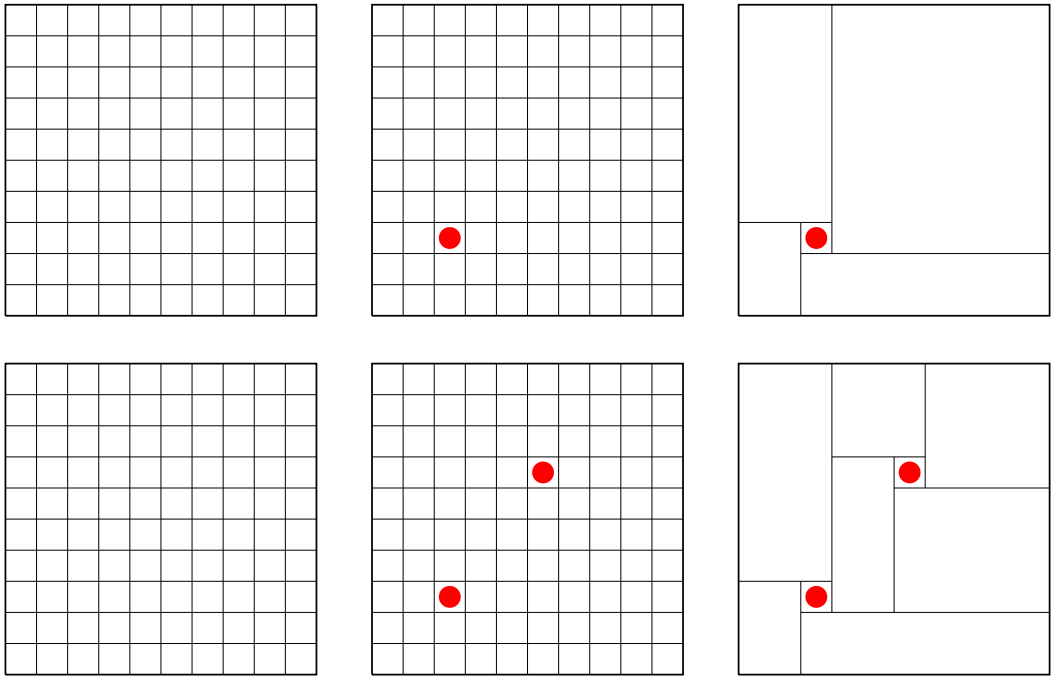


Figure 56 – Bloc layouts for histograms over $[0,1]^d$: m_N, selected parts in m_N, completed segmentations. $\mathrm{C}_{\mathrm{ext}} \leq 2d$

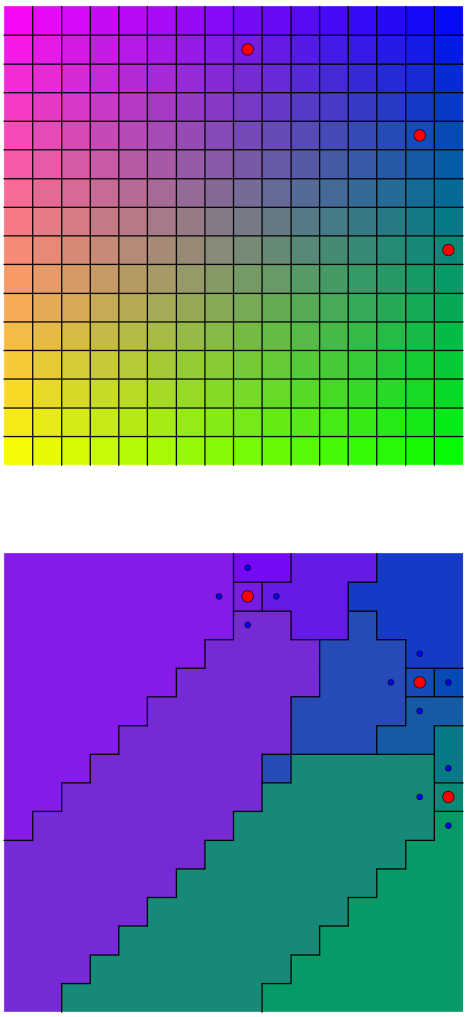


Figure 57 – Voronoi segmentation for histograms over $[0,1]^d$: m_N and selected parts, completed segmentations.

Complexity rate on the model family As in the context of Gaussian signal segmentation, we measure the complexity of a family of histogram models by its *binomial complexity rate*:

Definition 9.12 (binomial complexity rate for a histogram family)**.** Assume specified, as in Assumption 9.9 a positive integer N and m_N a partition of $[0,1]$ in $N+1$ parts of equal measure $\frac{1}{N+1}$ so that that every partition element of the family $\mathcal{M}$ has for parts unions of the parts of m_N. Recall that $|m|$ stands for the number of part of a partition m. We call *binomial complexity rate* of the model family $\mathcal{M}$ the number

$$\mathrm{B}_{\mathcal{M}} = \inf\left\{ \beta \geq 0, \sum_{m\in\mathcal{M},|m|>1} 2^{-(|m|-1)} \left(\frac{eN}{|m|-1}\right)^{-\beta(|m|-1)} \leq 1 \right\}.$$

Except for the absence of an empty partition, this definition is very close to its equivalent in the Gaussian context (see Definition 8.4), and similar considerations apply. Notably, in the case of segmentation of a real interval $[0,1]$, the binomial complexity rate satisfies $\mathrm{B}_{\mathcal{M}} \leq 1$, as the number of models with $|m|$ parts is $\binom{N}{|m|-1} \leq \left(\frac{eN}{|m|-1}\right)^{|m|-1}$.

9.3 Sufficient penalties under the null hypothesis

The following proposition will provide us with a comparison benchmark for the minimal penalty result in the next section. For a histogram model $m \in \mathcal{M}$ of , the histogram estimator $\hat{s}_m$ corresponds to Definition 9.1.

Proposition 9.13 (Risk lower bound for density estimation)**.** *Let $\xi_1, \cdots, \xi_n$ be some independent $[0,1]$-valued random variables with common distribution $P = s\mu$, where μ denotes the Lebesgue measure. Consider a finite family $\mathcal{M}$ of partitions of $[0,1]$ satisfying to Assumption 9.9 for some integer N and with binomial complexity rate $\mathrm{B}_{\mathcal{M}}$. Assume Assumption 9.10 holds so that $\frac{n}{N} \geq \log(n)^2$.*
Assume the unknown density s is equal to 1 almost everywhere.
Let $c_1 > \frac{1}{2}$ and consider a penalty function $\mathrm{pen} : \mathcal{M} \to \mathbb{R}^+$ such that $\forall m \in \mathcal{M}$,

$$\mathrm{pen}(m) = \begin{cases} (|m|-1)\frac{c_1}{n}\left(1+\sqrt{2(1+c_1^{-1})\left[\log(2)+\mathrm{B}_{\mathcal{M}}\log\left(\frac{eN}{|m|-1}\right)\right]}\right)^2 & \text{if } |m|>1, \\ 0 & \text{if } |m|=1. \end{cases}$$

Let $\hat{m}$ be the minimizer over $m \in \mathcal{M}$ of the penalized log-likelihood criterion

$$-\int \hat{s}_m \log(\hat{s}_m)\,\mathrm{d}\,\mu + \mathrm{pen}(m).$$

Then for some constant $C(c_1)$,

$$\frac{1}{2}\,\mathbb{E}\left[\mathbf{h}^2(s,\hat{s}_{\hat{m}})\right] \leq \frac{C(c_1)}{n}$$

The proof, based on Castellan's Theorem 9.7, is given in Section 12.1.

9.4 A risk and dimension lower bound

In a different context than Castellan's Theorem 9.8, the following result shows that, under the null hypothesis and the completion rule in Assumption 9.11, lack of a minimal penalty, the model selection procedure will retain a model of excessive risk and dimension.

Recall that by Assumption 9.9 the bins of the histograms in the model family $\mathcal{M}$ are unions of parts of a partition of $[0,1]$ in $N+1$ parts of individual measure $\frac{1}{N+1}$. The expected occupancy count of an individual part is $\bar{k} = \frac{n}{N+1} = np$. The Poisson tail function $h(\cdot)$ (see Definition 14.1) is defined for $-1 < x$ by $h(x) := (1+x)\log(1+x) - x$ and the binomial tail function $h_p(\cdot)$ for $0 < x < 1$ and $0 < p < 1$ by $h_p(x) = x \log \frac{x}{p} + (1-x) \log \frac{1-x}{1-p}$.

Proposition 9.14 (Risk and dimension lower bound for density estimation)**.** *Assume the model selection procedure is based on a family of histograms satisfying Assumption 9.9 for some number N and some partition m_N. Set $p = \frac{1}{N+1}$ and $\bar{k} = \frac{n}{N+1}$.*

Assume the unknown density s to estimate is equal to 1 almost everywhere, that $n \geq 9$, that Assumption 9.10 holds ($\frac{n}{N} \geq \log(n)^2$) and that the completion rule in Assumption 9.11 applies for $\mathcal{M}$ and some number $\mathrm{C}_{\mathrm{ext}} \geq 1$.

Consider an integer k with

$$np + \sqrt{2np} + \frac{2}{3} < k < n - \sqrt{np(1-p)} - 1. \tag{9.2}$$

Then the number $\bar{k} h\left(\frac{k-\bar{k}}{\bar{k}}\right)$ is larger than one.

Assume the penalty function $\mathrm{pen}(\cdot)$ is bounded for any $m \in \mathcal{M}$ by:

$$0 \leq \mathrm{pen}(m) \leq \frac{|m| - 1}{n\, \mathrm{C}_{\mathrm{ext}}} \left[\bar{k} h\left(\frac{k - \bar{k}}{\bar{k}} \right) - 1 \right]. \tag{9.3}$$

Let $\hat{m}$ be the minimizer over $\mathcal{M}$ of the penalized log-likelihood criterion

$$\mathrm{crit}_n(\hat{s}_{m^\star}) = \gamma_n(\hat{s}_m) + \mathrm{pen}(m) = -\int \hat{s}_m \log(\hat{s}_m)\, \mathrm{d}\mu + \mathrm{pen}(m).$$

Set

$$L = n h_p\left(\frac{k + \sqrt{np(1-p)} + \frac{1-p}{2}}{n} \right). \tag{9.4}$$

Then there is a random set of parts $b^\star{}_k \subset m_N$ (as later defined in Definition 12.3) satisfying the following relations:

$$\mathbf{K}(\hat{s}_{\hat{m}}, s) = \int \hat{s}_{\hat{m}} \log \hat{s}_{\hat{m}}\, \mathrm{d}\mu \geq \frac{|b^\star{}_k|}{n} \ \ a.s. \ , \tag{9.5}$$

$$\mathbb{E}\left[|b^\star{}_k|\right] \geq \frac{N+1}{3} e^{-L}. \tag{9.6}$$

Moreover the following bounds hold, where E_k *denotes the expectation* $\mathbb{E}\left[|b^\star{}_k|\right]$ *:*

$$\mathbb{V}\left[|b^\star{}_k|\right] \leq E_k\left(1 - \frac{E_k}{N+1}\right). \tag{9.7}$$

and for any t *with* $0 \leq t < E_k$*,*

$$\mathbb{P}\left[|b^\star{}_k| - E_k \leq -t\right] \leq \exp\left[-E_k h\left(-\frac{t}{E_k}\right)\right]. \tag{9.8}$$

If in addition the following conditions are satisfied:

$$N \geq 3, \tag{9.9}$$

$$k \leq np + \sqrt{np(1-p)}\left(\sqrt{\log(N)} - 1\right) - \frac{1-p}{2}, \tag{9.10}$$

then the following bound holds:

$$\bar{k}h\left(\frac{k-\bar{k}}{\bar{k}}\right) \geq \left[1 + \frac{1}{2}\log(n)^{-\frac{1}{2}}\right]^{-1} \frac{N}{N+1}\left[\left(\sqrt{L} - \frac{1}{\sqrt{2}} - \log(n)^{-1}\right)^+\right]^2. \tag{9.11}$$

The proof is given in section 12.2. Before discussing the results, remark that the additional assumptions in Inequalities 9.9 and 9.10 are cheap, since the upper bound on the number N is $N \leq \frac{n}{\log(n)^2}$ by Assumption 9.9 and the lower bound on the number k is $k > np + \sqrt{2np} + \frac{2}{3}k$ in Inequality 9.2.

9.4.1 Minimal penalty

This section explicits the relation of Proposition 9.14 to minimal penalties. On one hand Proposition 9.13 shows that under cheap assumptions on the family of histograms, under the null assumption $s = 1$ a.s. and with penalties of the form

$$\operatorname{pen}(m) = \begin{cases} \frac{|m|-1}{n}\left(\sqrt{c_1} + \sqrt{2(c_1+1)\left[\log(2) + \mathrm{B}_{\mathcal{M}}\log\left(\frac{eN}{|m|-1}\right)\right]}\right)^2 & \text{if } |m| > 1, \\ 0 & \text{if } |m| = 1. \end{cases} \tag{9.12}$$

with $c_1 > 2$, the minimizer $\hat{m}$ of the penalized log-likelihood criterion has a Hellinger risk bounded by $\frac{C_1(c_1)}{n}$, which is typically the risk of a fixed model of dimension of the order of magnitude of $C_1(c_1)$. As a result, the normalized Hellinger risk $n\frac{1}{2}\mathbb{E}\left[\mathbf{h}^2(s, \hat{s}_{\hat{m}})\right]$ of the selected model never diverges away from the null risk of the true model represented by the singleton partition $\{[0,1]\}$, or from the normalized risk of a fixed predefined model (see Inequality 9.1). In large N, as $c_1 > \frac{1}{2}$ the dominant term in sufficient penalties is at least

$$\frac{|m|-1}{n} 3\mathrm{B}_{\mathcal{M}} \log\left(\frac{eN}{|m|-1}\right).$$

As a reference at the other extreme, with a constant source density s, the worst model is the most ramified one, represented by the partition m_N of cardinal $N+1$.

Recall that N_τ represents the occupancy count of a part τ by sample points. If τ_0 is any elementary part of measure $\frac{1}{N+1}$, the corresponding Hellinger risk is $\frac{1}{2}\mathbb{E}\left[(1-\sqrt{\frac{N_{\tau_0}}{k}})^2\right]$, which is close to the variance term $\frac{1}{8}\mathbb{V}\left[\frac{N_{\tau_0}}{k}\right] = \frac{N}{8n}$ for $n >> N$.

Turning now to minimal penalties, in the point of view of proposition 9.14, assume in large N that $L = \alpha \log N$ with $\alpha < 1$. Under penalties where the dominant term is

$$(1+o(1))\frac{|m|-1}{n\mathrm{C}_{\mathrm{ext}}}\frac{N}{N+1}\alpha\log(N),$$

the selected estimator $\hat{s}_{\hat{m}}$ satisfies the risk relation:

$$\mathbb{E}\left[\mathbf{K}(\hat{s}_{\hat{m}}, s)\right] \geq \frac{N+1}{3n}e^{-L} \geq \frac{N^{1-\alpha}}{3n}.$$

(See Inequalities 9.3 and 9.7 for the minimal penalty level, and Inequalities 9.5 and 9.6 for the risk lower bound).
As a result, for large N, despite the null hypothesis, the selected model will be of poor quality if the penalty term $\mathrm{pen}(m)$ has an asymptotic equivalent under $[1+o(1)]\,\alpha\frac{|m|-1}{n}\frac{\log N}{\mathrm{C}_{\mathrm{ext}}}$, so that an indicative limiting condition for a reasonable penalty term is

$$\mathrm{pen}(m) \geq [1+o(1)]\,\frac{|m|-1}{n}\frac{\log N}{\mathrm{C}_{\mathrm{ext}}}.$$

Comparing with the sufficient penalty level proposed in Equation 9.12 in introduction of this section, it appears that asymptotically in partition and sample size, there is only a discrepancy by a ratio $3\mathrm{B}_{\mathcal{M}}\mathrm{C}_{\mathrm{ext}}$ between on one hand the minimal penalty term in Proposition 9.14 and on the other hand the sufficient penalty term from Castellan's model selection Theorem 9.8. For instance in the case of free segmentation of a sequence, the factor $\mathrm{C}_{\mathrm{ext}}$ is equal to 2 (see Figure 53) and the rate $\mathrm{B}_{\mathcal{M}}$ close to 1 (see Definition 9.12). In a different context but with very similar definitions, Proposition 8.5 shows that model families with at least one small model satisfy for large N the asymptotic bound $\mathrm{B}_{\mathcal{M}}\mathrm{C}_{\mathrm{ext}} \geq 1 + o(1)$, with equality for certain classes of model families. This offers an intuition on the tightness of the minimal penalty estimate in 9.14.

As with Gaussian signal partition, the term $\mathrm{C}_{\mathrm{ext}}$ relates to the specific geometry of the chosen partition family $\mathcal{M}$, which is more constraint that in the reference situation of free variable selection, despite the completion rule in Assumption. 9.11. Informally, the number $\mathrm{B}_{\mathcal{M}}\mathrm{C}_{\mathrm{ext}}$ informs on the degree of redundancy of the model family, and reflects on the gap between minimal and sufficient penalties, at least in their asymptotic form.

10 Remark on the isotropy of the model family

The purpose of this section is to illustrate the role of isotropy in the minimal penalty phenomenon. The result of L. Birgé and P. Massart in Proposition 5.8 takes place in the framework of free selection among independent feature variables, which act as an orthogonal family in their ambiant Euclidean space. Any permutation among those variables leaves the model family $\mathcal{M}$ invariant. The types of model families studied in this work often do not present this kind of isotropy by coordinate exchange, but the completion rule in Assumption 8.1 ensures a weakened form of this property with the existence of models of moderate dimension containing any subset of a given orthonormal basis which is analog to the set of independent variables just mentioned.

To corroborate the intuition that the degree of isotropy of the model family is a driving factor in the minimal penalty phenomenon, one would like to investigate other types of nearly isotropic model families. A natural way of producing such a family is to draw a sample from a uniformly distributed subspace of the Euclidean $\mathbb{R}^n$. This is the purpose of the following section.

10.1 Toy problem

Consider the following model selection toy problem: in the Gaussian situation of Definition 5.1, the observed signal belongs to the n-dimensional Euclidean space E_n and is of the form $Y = s + \epsilon W$, where s is the fixed unknown source signal to recover, and W a standard Gaussian random vector. The model family $\mathcal{M}$ is formed by the null model $\{0\}$ and a uniform random draw of N hyperspaces of common dimension d, as formalized in Definition 14.16, independent from the noise process W. We denote Π the corresponding uniform random d-dimensional projector, so that the model family is

$$\begin{aligned}\mathcal{M} &:= \{\{0\}, m_1, \cdots, m_N\}, \\ &= \{\{0\}, \Pi_1 E_n, \cdots, \Pi_N E_n\},\end{aligned}$$

where $(\Pi_1, \cdots, \Pi_N)$ is an i.i.d. sample of Π. By the rotation invariance of the distributions involved, this can be realized by choosing the explanatory variables of each non-null model as an independent d-sample of an uniformly oriented random vector.

For simplicity we assume that $s = 0$ and $\epsilon^2 = 1$. The non null models are obviously worse than the null one in any respect, in the sense that the risk of the null model is

$$\mathbb{E}_{\mathcal{M},W}\left[\|s - 0\|^2\right] = 0,$$

when the risk of any individual non null model m_i for $1 \leq i \leq N$ is

$$\mathbb{E}_{\mathcal{M},W}\left[\|s - \Pi_i(s + \epsilon W)\|^2\right] = \mathbb{E}_{\mathcal{M}}\left[\mathbb{E}_W\left[\|\Pi(W)\|^2\right]\right] = \mathbb{E}_{\mathcal{M}}[d] = d.$$

Denote $\tilde{\Pi}$ a model element of $\mathcal{M}$ realising the minimum

$$\inf_{\substack{m\in\mathcal{M},\\ |m|>0}} -\|\Pi_m(s + \epsilon W)\|^2 = \inf_{\substack{m\in\mathcal{M},\\ |m|>0}} -\|\Pi_m(W)\|^2.$$

In the absence of penalty, the minimum of the least square criterion is formally

$$0 \wedge -\left\|\tilde{\Pi}(s + W)\right\|^2 = 0 \wedge -\left\|\tilde{\Pi}(W)\right\|^2,$$

and the model selection procedure almost surely retains the non-null model $\tilde{\Pi}$. The risk of the entire procedure is then

$$\mathbb{E}_{\mathcal{M},W}\left[\|s-\hat{s}_{\hat{m}}\|^2\right]=\mathbb{E}_{\mathcal{M},W}\left[\left\|\tilde{\Pi}(W)\right\|^2\right].$$

which is by definition not less than $\mathbb{E}_{\Pi,W}\left[\|\Pi(W)\|^2\right]=d$, and larger if $N>1$. As anticipated, the selection procedure among several random models only worsen the poor expected risk of an individual one.

Choose a penalty term based on the model dimension,

$$\mathrm{pen}\,(m)=\begin{cases}0 & \text{if } |m|=0,\\ \mathrm{pen}_d>0 & \text{if } |m|=d.\end{cases}$$

The selected model is now the maximizer of

$$\mathrm{pen}_d\vee\left\|\tilde{\Pi}(W)\right\|^2,$$

so that to contain the risk of the selection procedure, the penalty pen_d should be set to some high quantile of the random variable $\left\|\tilde{\Pi}(W)\right\|^2$. This is the purpose of the following proposition:

Proposition 10.1 (Penalty level for toy problem)**.** *Consider the model selection procedure described in the present section, assume that* $4\leq d\leq\frac{n}{2}$ *and consider a number* t *with* $0<t\leq\frac{N}{10}$*. Assume that for some number* pen_d *the penalty function is defined* $\forall m\in\mathcal{M}$ *by :*

$$\mathrm{pen}\,(m)=\begin{cases}0 & \textit{if } |m|=0,\\ \mathrm{pen}_d & \textit{if } |m|=d.\end{cases}$$

Denote $\hat{s}_{\hat{m}}$ *a minimizer of the penalized contrast criterion* $-\|\hat{s}_m\|^2+\mathrm{pen}\,(m)$ *over* $m\in\mathcal{M}$*.*

Sufficient penalty *If for some* $\eta\in(0,1)$ *the number* pen_d *satisfies:*

$$\mathrm{pen}_d=\frac{1+\eta}{1-\eta}\left(d+2\sqrt{d\log(N)}+2\log(N)\right). \tag{10.1}$$

then the following bound holds:

$$\mathbb{E}_{\mathcal{M},W}\left[\|s-\hat{s}_{\hat{m}}\|^2\right]=\mathbb{E}_{\mathcal{M},W}\left[\|\hat{s}_{\hat{m}}\|^2\right]\leq\frac{(1+3\eta)}{\eta^2(1-\eta)}. \tag{10.2}$$

Minimal penalty *There is a real positive function defined for* $4\leq d\leq\frac{n}{2}$ *and* $0<t\leq\frac{N}{10}$*:*

$$(n,d,N,t)\mapsto d_t^\star(n,d,N,t)$$

so that if the penalty function satisfies:

$$\mathrm{pen}_d\leq d_t^\star(n,d,N,t),$$

then the following bounds hold:

$$\mathbb{P}_{\mathcal{M},W}\left[|\hat{m}| = d\right] \geq (1 - e^{-t})^2, \tag{10.3}$$

and

$$\mathbb{E}_{\mathcal{M},W}\left[\left\|s - \hat{s}_{\hat{m}}\right\|^2\right] = \mathbb{E}_{\mathcal{M},W}\left[\left\|\hat{s}_{\hat{m}}\right\|^2\right] \geq (1 - e^{-t})^2 d_t^\star(n, d, N, t). \tag{10.4}$$

Moreover, when n, d and N tend to ∞ while $d \leq \frac{n}{2}$, for any fixed value of t,

$$d_t^\star(n, d, N, t) = [1 + o(1)] \left[d + \frac{2\sqrt{d\left(1 - \frac{d}{n}\right)\log(N)}}{1 + \frac{2}{n-d}\sqrt{d\left(1 - \frac{d}{n}\right)\log(N)}} \right]. \tag{10.5}$$

The proof is given in Section 13.1

A first conclusion is that the limit form of the minimal penalty estimate is always above Mallows's heuristic ($\text{pen}_d = d$ in this case). This effect becomes significant when the quantity $\frac{\log N}{d}$ exceeds 1, which is customary: for instance with free variable selection $\log(N) \sim d \log\left(\frac{en}{d}\right)$. Setting $d = o(n)$ and for some $r > 0$

$$N = [1 + o(1)]\, e^{rd},$$

yields for the sufficient penalty the estimate:

$$[1 + o(1)]\, d\left(1 + 2\sqrt{r} + 2r\right),$$

and for the minimal penalty:

$$[1 + o(1)]\, d\left(1 + 2\sqrt{r}\right).$$

For small values of the rate r, these two estimates can be as close to each other than desired. In this situation both estimates get near to Mallows's heuristic, however it may be worth mentioning that the two estimates are always closer to each other (by $2r$) than to Mallows's heuristic (by $2\sqrt{r}$).

A point of interest is the status of the denominator of the right term in 10.5. When $\log N$ is large compared to $\frac{n(n-d)}{d}$, the effect of the isotropy is dominant and the minimal penalty level gets close to its absolute threshold n, which is the expected squared norm of the noise W. In this case the sufficient penalty estimate becomes excessively larger than n, showing we are using it outside its natural domain of interest.

Altogether this toy problem corroborates our initial intuition that isotropy is a driving factor in the minimal penalty phenomenon, plausibly by ensuring a form of non redundancy of the model family.

11 Proofs for Sections 5, 6 and 8

11.1 Proof of Corollary 5.7

Proof. Let us apply the general Gaussian model selection Theorem 5.6 with

$$\begin{aligned}\bar{\mathcal{M}} &= \emptyset,\\ \theta &= 1-\eta,\\ \kappa &= 1+2\eta.\end{aligned}$$

Then as required, $\theta \in (0,1)$ and $\kappa > 2-\theta = 1+\eta$. Moreover, since $1+2\eta < \frac{1+\eta}{1-\eta}$,

$$\begin{aligned}\mathcal{Q}_m &= \epsilon^2 D_m \left(\kappa + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1}L_m\right),\\ &= \epsilon^2 D_m \left(1+2\eta+2(1+\eta)\sqrt{L_m} + 2\frac{1}{1-\eta}L_m\right),\\ &\leq \frac{1+\eta}{1-\eta}\epsilon^2 D_m \left(1+2\sqrt{L_m}+2L_m\right).\end{aligned}$$

Then the relation

$$\forall m \in \mathcal{M},\ \mathrm{pen}\,(m) \geq \frac{1+\eta}{1-\eta}\epsilon^2 D_m \left(1+2\sqrt{L_m}+2L_m\right),$$

is enough for the penalty function to satisfy the condition

$$\forall m \in \mathcal{M},\ \mathrm{pen}\,(m) \geq \mathcal{Q}_m.$$

It follows by Theorem 5.6 that the corresponding penalized projection estimator $\tilde{s} = \hat{s}_{\hat{m}}$ with $\hat{m}$ given by Definition 5.5 exists a.s. and satisfies:

$$\begin{aligned}(1-\theta)\,\mathbb{E}\left[\|s-\tilde{s}\|^2\right] \leq \inf_{m\in\mathcal{M}} \left\{\mathrm{d}^2\,(s,S_m) + \mathrm{pen}\,(m) - \epsilon^2 D_m\right\}\\ + \epsilon^2\left[(2-\theta)^2(\kappa+\theta-2)^{-1} + 2\theta^{-1}\right].\end{aligned} \tag{11.1}$$

After substitution, since

$$\begin{aligned}(2-\theta)^2(\kappa+\theta-2)^{-1} + 2\theta^{-1} &= \frac{(1+\eta)^2}{\eta} + 2\frac{1}{1-\eta},\\ &= \frac{1}{\eta(1-\eta)}\left((1+\eta)(1-\eta^2)+2\eta\right),\\ &\leq \frac{(1+3\eta)}{\eta(1-\eta)},\end{aligned}$$

Inequality 11.1 implies:

$$\begin{aligned}\eta\,\mathbb{E}\left[\|s-\tilde{s}\|^2\right] \leq \inf_{m\in\mathcal{M}} \left\{\mathrm{d}^2\,(s,S_m) + \mathrm{pen}\,(m) - \epsilon^2 D_m\right\}\\ + \epsilon^2\frac{(1+3\eta)}{\eta(1-\eta)}.\end{aligned}$$

□

11.2 Proof of Proposition 6.1

Proof. As mentioned in introduction of the proposition, by Lemma B.2, the set of weights defined by

$$\forall m \in \mathcal{M}, L_m = L_b \text{ with } L_b := \frac{\log b}{b-1} + \log\left(\frac{b}{b-1}\right)$$

satisfies

$$\sum_{m \in \mathcal{M}} e^{-|m| L_m} < 1.$$

A direct application of the model selection result in Corollary 5.7 with the constant set of weights above and

$$\forall m \in \mathcal{M},\ \mathrm{pen}\,(m) = \epsilon^2\,|m|\,\frac{1+\eta}{1-\eta}\left[1 + 2\sqrt{L_b} + 2L_b\right],$$

ensures that $\hat{s}_{\hat{m}}$ exists a.s., and yields Inequality 6.2 in the proposition:

$$\begin{aligned}
\eta\,\mathbb{E}\left[\|s - \hat{s}_{\hat{m}}\|^2\right] &\leq \inf_{m \in \mathcal{M}} \left\{ \mathrm{d}^2\,(s, S_m) + \epsilon^2\,|m|\,\frac{1+\eta}{1-\eta}\left[1 + 2\sqrt{L_b} + 2L_b\right] - \epsilon^2\,|m| \right\} \\
&\quad + \epsilon^2 \frac{1+3\eta}{\eta(1-\eta)}, \\
&\leq \inf_{m \in \mathcal{M}} \left\{ \mathrm{d}^2\,(s, S_m) + \epsilon^2\,|m|\,\frac{1+\eta}{1-\eta}\left[\frac{2\eta}{1+\eta} + 2\sqrt{L_b} + 2L_b\right] \right\} \\
&\quad + \epsilon^2 \frac{1+3\eta}{\eta(1-\eta)}.
\end{aligned} \tag{11.2}$$

If $s \in \tilde{m}$ then $d(s, S_{\tilde{m}}) = 0$ so that:

$$\eta\,\mathbb{E}\left[\|s - \hat{s}_{\hat{m}}\|^2\right] \leq \epsilon^2\,|\tilde{m}|\,\frac{1+\eta}{1-\eta}\left[\frac{2\eta}{1+\eta} + 2\sqrt{L_b} + 2L_b\right] + \epsilon^2 \frac{1+3\eta}{\eta(1-\eta)}.$$

or equivalently the announced relation 6.3:

$$\mathbb{E}\left[\|s - \hat{s}_{\hat{m}}\|^2\right] \leq \epsilon^2\,|\tilde{m}|\,\frac{1+\eta}{\eta(1-\eta)}\left[\frac{2\eta}{1+\eta} + 2\sqrt{L_b} + 2L_b\right] + \epsilon^2 \frac{1+3\eta}{\eta^2(1-\eta)}.$$

The function $x \mapsto f(x) = \frac{\log x}{x-1} + \log \frac{x}{x-1}$ is decreasing for $x \geq 2$, and by the relation $\log(1+x) \leq x$ for $x > -1$,

$$\begin{aligned} L_b &= \frac{\log b}{b-1} + \log \frac{b}{b-1}, \\ &= \frac{2\log\sqrt{b}}{b-1} + \log\left(1 + \frac{1}{b-1}\right), \\ &\leq \frac{2\sqrt{b}-1}{b-1}, \\ &= \frac{1}{\sqrt{b}} \frac{2b-\sqrt{b}}{b-1}, \\ &\leq \frac{1}{\sqrt{b}} \frac{2b-\sqrt{2}}{b-1}, \text{ since } b \geq 2 \\ &= \frac{1}{\sqrt{b}} \left[2 + \frac{2-\sqrt{2}}{b-1}\right], \\ &\leq \frac{4-\sqrt{2}}{\sqrt{b}}, \qquad (11.3) \\ &\lesssim \frac{2.59}{\sqrt{b}}. \end{aligned}$$

moreover as the weight L_b decreases with b and $L_2 = 2\log(2)$,

$$\begin{aligned} 2\sqrt{L_b} + 2L_b &\leq 2\sqrt{L_b}\left(1 + \sqrt{L_2}\right), \qquad (11.4) \\ &= 2\sqrt{L_b}\left(1 + \sqrt{2\log(2)}\right), \qquad (11.5) \end{aligned}$$

so that altogether by 11.3 and 11.5 the following bound holds whatever $b \geq 2$:

$$\begin{aligned} 2L_b + 2\sqrt{L_b} &\leq 2\sqrt{\frac{4-\sqrt{2}}{\sqrt{b}}}\left(1 + \sqrt{L_2}\right), \\ &= \frac{1}{\sqrt[4]{b}} 2\sqrt{4-\sqrt{2}}\left(1 + \sqrt{2\log(2)}\right), \\ &\lesssim \frac{7.003}{\sqrt[4]{b}}, \\ &\leq \frac{8}{\sqrt[4]{b}}. \qquad (11.6) \end{aligned}$$

Finally, Inequality 11.2 implies the annouced relation 6.3 :

$$\mathbb{E}\left[\|s - \hat{s}_{\hat{m}}\|^2\right] \leq \epsilon^2 |\tilde{m}| \frac{1+\eta}{\eta(1-\eta)} \left[\frac{8}{\sqrt[4]{b}} + \frac{2\eta}{1+\eta}\right] + \epsilon^2 \frac{1+3\eta}{\eta^2(1-\eta)}.$$

Case of a Lipschitz source function Assume that the unknown source function is L-Lipschitz, that $\mu(U) = 1$ and that there is a number $d_U \leq 1$ so that the diameter of any part τ of height h (and then of measure b^{-h}) is bounded by

$$\begin{aligned} \delta_\tau &\leq \delta_U \mu(\tau)^{\frac{1}{d_U}}, \\ &= \delta_U b^{-\frac{h}{d_U}}. \end{aligned}$$

Notice that for instance in the 2-dimensional example of Figure 2, the number d_U is 2. Consider for each integer $h \leq 0$ the model m_h represented by the maximal b-ary tree of height h. Then as s_{m_h} is the projection of the source function s on the model m_h,

$$\begin{aligned} d(s, S_{m_h}) &= \|s - s_{m_h}\|^2, \\ &\leq \sum_{\tau \in m_h} \int_\tau \left(f - \frac{\sup_\tau f + \inf_\tau f}{2} \right)^2 \mathrm{d}\,\mu, \\ &\leq \sum_{\tau \in m_h} \left(\frac{\sup_\tau f - \inf_\tau f}{2} \right)^2 \mu(\tau), \\ &\leq \sum_{\tau \in m_h} \frac{L^2 \delta_\tau^2}{4} \mu(\tau), \\ &\leq \sum_{\tau \in m_h} \frac{L^2 \delta_U^2}{4} b^{-2\frac{h}{d}} \mu(\tau), \\ &\leq \frac{L^2 \delta_U^2}{4} b^{-2\frac{h}{d}}. \end{aligned}$$

Moreover the dimension of the model m_h is b^h. Consider a number $\Lambda > 0$. Then

$$\begin{aligned} \inf_{m \in \mathcal{M}} \left\{ d(s, S_{m_h}) + \Lambda \left| m \right| \right\} &\leq \inf_{h \geq 0} \left\{ d(s, S_{m_h}) + \Lambda \left| m \right| \right\}, \\ &\leq \inf_{h \geq 0} \left\{ \frac{L^2 \delta_U^2}{4} b^{-2\frac{h}{d}} + \Lambda b^h \right\}. \end{aligned} \tag{11.7}$$

The minimum of this bound for h over the real line is obtained for

$$b^{h^\star} = \left(\frac{L^2 \delta_U^2}{2 \Lambda d} \right)^{\frac{d}{d+2}}.$$

The nearest larger integer $\lceil h^\star \rceil$ satisfies

$$\begin{aligned} b^{\lceil h^\star \rceil} &\leq b^{h^\star + 1}, \\ &= b \left(\frac{L^2 \delta_U^2}{2 \Lambda d} \right)^{\frac{d}{d+2}}. \end{aligned}$$

Taking the bound in Inequality 11.7 at $h = \lceil h^\star \rceil$ yields,

$$\begin{aligned} \inf_{m \in \mathcal{M}} d(s, S_{m_h}) + \Lambda \left| m \right| &\leq \frac{L^2 \delta_U^2}{4} \left(\frac{L^2 \delta_U^2}{2 \Lambda d} \right)^{\frac{-2}{d+2}} + \Lambda b \left(\frac{L^2 \delta_U^2}{2 \Lambda d} \right)^{\frac{d}{d+2}}, \\ &= \left(\frac{L^2 \delta_U^2}{4} \right)^{\frac{d}{d+2}} \left(\frac{\Lambda d}{2} \right)^{\frac{2}{d+2}} \left(1 + 2\frac{b}{d} \right). \end{aligned}$$

Combining with the oracle-like Inequality 11.2 after setting the parameter Λ to:

$$\begin{aligned} \Lambda &= \epsilon^2 \frac{1+\eta}{1-\eta} \left[\frac{2\eta}{1+\eta} + 2\sqrt{L_b} + 2L_b \right], \\ &\leq \epsilon^2 2 \frac{1+\eta}{1-\eta} \left[\frac{1}{2} + \sqrt{L_b} + L_b \right], \text{ since } \eta < 1, \end{aligned}$$

yields the announced relation 6.5:

$$\begin{aligned}\eta\,\mathbb{E}\left[\|s-\hat{s}_{\hat{m}}\|^2\right] &\leq \left(\frac{L^2\delta_U^2}{4}\right)^{\frac{d}{d+2}}\left(\epsilon^2\frac{1+\eta}{1-\eta}\right)^{\frac{2}{d+2}} d^{\frac{2}{d+2}}\left[\frac{1}{2}+\sqrt{L_b}+L_b\right]^{\frac{2}{d+2}}\left(1+2\frac{b}{d}\right)\\ &\quad+\epsilon^2\frac{1+3\eta}{\eta(1-\eta)},\\ &= C(b,d)\left(\frac{L\delta_U}{2}\right)^{2\frac{d}{d+2}}\epsilon^{2\frac{2}{d+2}}\left(\frac{1+\eta}{1-\eta}\right)^{\frac{2}{d+2}}+\epsilon^2\frac{1+3\eta}{\eta(1-\eta)},\end{aligned}$$

where

$$C(b,d)=d^{\frac{2}{d+2}}\left[\frac{1}{2}+\sqrt{L_b}+L_b\right]^{\frac{2}{d+2}}\left(1+2\frac{b}{d}\right).$$

For $d>1$, the term $d^{\frac{2}{d+2}}$ can be bounded by 2 as follows, with the shorthand $z:=\log(d)$,

$$\begin{aligned}d^{\frac{2}{d+2}} &= \exp\left[\frac{2z}{e^z+2}\right],\\ &\leq \exp\left[\frac{2z}{3+z+\frac{1}{2}z^2}\right],\\ &= \exp\left[\frac{2}{1+\frac{3}{z}+\frac{z}{2}}\right],\\ &\leq \exp\left[\frac{2}{1+\sqrt{6}}\right], \text{ by the relation } x+y\geq 2\sqrt{xy} \text{ for } x>0 \text{ and } y>0\\ &\simeq 1.79,\\ &< 2.\end{aligned}\tag{11.8}$$

Finally, as the weight L_b decreases with b, and $d\geq 1$, the bound announced in 6.4 follows:

$$\begin{aligned}C(b,d) &\leq 2\left[\frac{1}{2}+\sqrt{L_2}+L_2\right]^{\frac{2}{3}}\left(1+2\frac{b}{d}\right),\\ &\leq 2\left[\frac{1}{2}+\sqrt{2\log 2}+2\log 2\right]^{\frac{2}{3}}\left(1+2\frac{b}{d}\right),\\ &\lesssim 4.22\left(1+2\frac{b}{d}\right),\\ &\leq 5\left(1+2\frac{b}{d}\right).\end{aligned}$$

□

11.3 Proof of Lemma 6.2

Proof. For each part $c\in I$, consider the minimal sub-tree T_c of $T_{b,h}$ containing c. The sub-tree T_c has for inner nodes all the antecedents of c, down to the root node. Its leaves are the siblings of c and its antecedents. See for instance Figure 58. The sub-tree T_c

has c as leaf, and no inner node of height above $h-1$. Then the combined rooted tree $T_I = \cup\{T_c,\, c \in I\}$ has all the elements of I as leaves, and no leaf of height above h.

Moreover since any part $c \in I$ is of height h, it has exactly h antecedents and the minimal tree T_c has exactly h inner nodes. Then the combined tree T_I has not more than $|I|\,h$ inner nodes. As a b-ary tree with n inner nodes has $1+(b-1)n$ leaves, the tree T_I has not more than $1+(b-1)\,|I|\,h$ leaves. See for instance Figure 3.

Altogether the combined tree T_I fulfills the requirements for the tree announced in Lemma 6.2. □

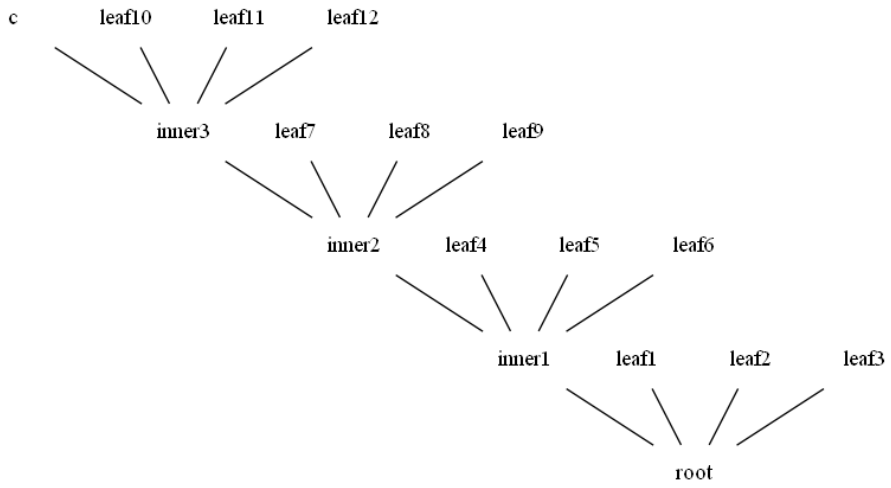


Figure 58 – Minimal rooted b-ary tree with a prescribed part c as leaf

11.4 Proof of proposition 6.4

Proof. Consider a positive integer height h, that we will later make tend to infinity. The b^h parts of $\mathcal{M}$ of height h form a partition of the underlying space U. Denote them $(\tau_{h,i})_{i=1}^{b^h}$. By the covariance structure of the Brownian sheet B, the sequence $\left(X_{h,i} = b^{\frac{h}{2}} \int_{\tau_{h,i}} \mathrm{d}\, B\right)_{i=1}^{b^h}$ is an i.i.d sequence of standard Gaussian random variables.

Consider a real number $x_h > 0$, to be determined later. Denote α_h the probability $\alpha_h = \mathbb{P}\left[|X| \geq x_h\right]$ where X is a standard Gaussian variable, and define the quantities:

$$\mathfrak{w}_h^+ := \mathbb{E}\left[X^2 - 1 \,\middle|\, |X| \geq x_h\right],$$

and

$$\mathfrak{w}_h^- := \mathbb{E}\left[1 - X^2 \,\middle|\, |X| < x_h\right].$$

The balance condition $\alpha_h(1+\mathfrak{w}_h^+) + (1-\alpha_h)(1-\mathfrak{w}_h^-) = \mathbb{E}\left[X^2\right] = 1$ implies that

$$\mathfrak{w}_h^- := \mathbb{E}\left[1 - X^2 \,\middle|\, |X| < x_h\right] = \mathfrak{w}_h^+ \frac{\alpha_h}{1-\alpha_h}. \tag{11.9}$$

Consider the random set of indexes

$$I_h^\star := \left\{ i \in \left[1, b^h\right], |X_{h,i}| \geq x_h \right\}. \tag{11.10}$$

The random set $I_h^\star$ only depends on the modules of the i.i.d. symmetrical random variables $(X_{h,i})_{i=1}^{b^h}$ and is independent of their signs. As a result, conditioned to a realisation of $I_h^\star$, both the sequences $(X_{h,i})_{i\in I_h^\star}$ and $(X_{h,i})_{i\in\left[1,b^h\right]\setminus I_h^\star}$, when not empty, are i.i.d. sequences of sign symmetrical random variables. In particular, the expectations of both sequences are null, and their common covariance matrix is diagonal. We also note that the respective marginal distributions are the normalized restrictions of the Gaussian distribution to $[-x_h, x_h]$ and $\mathbb{R}\setminus[-x_h, x_h]$ whatever $I_h^\star$.

By the completion rule in Lemma 6.2 for any set of indexes $I \subset \left[1, b^h\right]$ there is a model $m_h(I) \in \mathcal{M}$ of dimension $|m_h(I)| \leq 1+(b-1)h\,\mathrm{Card}(I)$ so that the corresponding partition contains each of the part sequence $(\tau_{h,i})_{i\in I}$ and has no part of height larger than h. (Figure 3 provides a simple instance).

Define the random model $m_h^\star = m_h(I_h^\star)$. By construction any part element of $\{\tau_i\}_{i\in I_h^\star}$ is an element of $m_h^\star$, or in other words a leaf in the corresponding tree (see the crossed parts in Figure 3). Also by construction, as the tree $m_h^\star$ as no part of height above h, each part in the 'complement partition' $m_h^\star \setminus \{\tau_i\}_{i\in I_h^\star}$ is a union of elements of the set $\{\tau_i\}_{i\in\left[1,b^h\right]\setminus I_h^\star}$ (see the non crossed parts in Figure 3). From the expression Π_m of the orthogonal projector on a model S_m:

$$\Pi_m = \sum_{\tau\in m} \frac{1}{\mu(\tau)}\, \mathbb{1}_\tau \otimes \mathbb{1}_\tau,$$

we draw that given $I \subset \left[1, b^h\right]$, the following holds:

$$\begin{aligned}
\mathbb{E}\left[\gamma(\hat{s}_{m_h^\star}) \,\middle|\, I^\star = I\right] &= \mathbb{E}\left[-\sum_{\tau\in m_h^\star} \mu(\tau)^{-1} \Big(\int_\tau s\,\mathrm{d}\,\mu + \epsilon \int_\tau \mathrm{d}\,B\Big)^2 \,\middle|\, I^\star = I \right], \\
&= \mathbb{E}\left[-\sum_{\tau\in m_h^\star} \mu(\tau)^{-1} \left(\int_\tau s\,\mathrm{d}\,\mu + \epsilon \sum_{\substack{1\leq i\leq b^h \\ \tau_i\subset\tau}} \mu(\tau_i)^{\frac{1}{2}} X_{h,i}\right)^2 \,\middle|\, I^\star = I \right],
\end{aligned}$$

and by the observations above on the moments and the marginal distributions of the sequences $(X_{h,i})_{i\in I_h^\star}$ and $(X_{h,i})_{i\in\left[1,b^h\right]\setminus I_h^\star}$ conditioned to a realisation of $I_h^\star$,

$$\begin{aligned}
\mathbb{E}\left[\gamma(\hat{s}_{m_h^\star}) \,\middle|\, I^\star = I\right] = &-\mathbb{E}\left[\sum_{\tau\in m_h^\star} \mu(\tau)^{-1} \left(\int_\tau s\,\mathrm{d}\,\mu\right)^2 \,\middle|\, I^\star = I \right] \\
&- \epsilon^2\, \mathbb{E}\left[\sum_{\tau\in m_h^\star} \sum_{\substack{1\leq i\leq b^h \\ \tau_i\subset\tau}} \mu(\tau)^{-1}\mu(\tau_i) X_{h,i}^2 \,\middle|\, I^\star = I \right].
\end{aligned}$$

We also noted that conditioned to $I_h^\star = I$, the random sequences $(X_{h,i})_{i\in I}$ and $(X_{h,i})_{i\in\left[1,b^h\right]\setminus I_h^\star}$ are both exchangeable. It follow that:

$$\begin{aligned}\mathbb{E}\left[\gamma(\hat{s}_{m_h^\star}) \,\middle|\, I^\star = I\right] = &-\mathbb{E}\left[\sum_{\tau\in m_h^\star} \mu(\tau)^{-1}\left(\int_\tau s\,\mathrm{d}\,\mu\right)^2 \,\middle|\, I^\star = I\right] \\ &- \epsilon^2\left[\sum_{\tau\in I_h^\star} 1\right]\mathbb{E}\left[X_{h,i}^2 \,\middle|\, i\in I^\star\right] \\ &- \epsilon^2\left[\sum_{\tau\in {m^\star}_h\setminus I_h^\star} 1\right]\mathbb{E}\left[X_{h,i}^2 \,\middle|\, i\in\left[1,b^h\right]\setminus I_h^\star\right].\end{aligned}$$

Recalling that the random model ${m^\star}_h$ is defined as $m_h(I_h^\star)$ and that the random set $I_h^\star$ is defined in Equation 11.10 as $\left\{i\in\left[1,b^h\right], |X_{h,i}|\geq x_h\right\}$, and ignoring the contribution of the source function s,

$$\begin{aligned}\mathbb{E}\left[\gamma(\hat{s}_{m_h^\star}) \,\middle|\, I^\star = I\right] &\leq -\epsilon^2\,\mathrm{Card}(I)\,\mathbb{E}\left[X^2 \,\middle|\, |X|\geq x_h\right] \\ &\quad - \epsilon^2(|m_h(I)| - \mathrm{Card}(I))\,\mathbb{E}\left[X^2 \,\middle|\, |X|< x_h\right], \\ &= -\epsilon^2\left[\mathrm{Card}(I)(1+\mathfrak{w}_h^+) + (|m_h(I)| - \mathrm{Card}(I))(1-\mathfrak{w}_h^-)\right].\end{aligned}$$

By the balance Equation 11.9, $\mathfrak{w}_h^- = \mathfrak{w}_h^+\frac{\alpha_h}{1-\alpha_h}$, so that

$$\mathbb{E}\left[\gamma(\hat{s}_{m_h^\star}) \,\middle|\, I^\star = I\right] \leq -\epsilon^2\left[|m_h(I)| + \frac{\mathrm{Card}(I) - \alpha_h\,|m_h(I)|}{1-\alpha_h}\mathfrak{w}_h^+\right].$$

Releasing the conditioning on I and recalling the assumption that

$$\mathrm{pen}\,(m) < \epsilon^2\,|m|\left(1 + 2\rho\frac{\log b}{b-1}\right)\ \forall m\in\mathcal{M},$$

leads to:

$$\mathbb{E}\left[\gamma(\hat{s}_{m_h^\star}) + \mathrm{pen}\,(m_h^\star)\right] \leq \epsilon^2\,\mathbb{E}\left[|m_h^\star|\,2\rho\frac{\log b}{b-1} - \frac{\mathrm{Card}(I_h^\star) - \alpha_h\,|m_h^\star|}{1-\alpha_h}\mathfrak{w}_h^+\right],$$

By the completion rule in Lemma 6.2, the relation $|m_h^\star| \leq 1+(b-1)h\,\mathrm{Card}(I_h^\star)$ holds a.s.. After restating it into $\mathrm{Card}(I_h^\star) \geq \frac{|m_h^\star|-1}{(b-1)h}$ and combining it with the contrast inequality above:

$$\begin{aligned}\mathbb{E}\left[\gamma(\hat{s}_{m_h^\star}) + \mathrm{pen}\,(m_h^\star)\right] &\leq \epsilon^2\frac{\mathfrak{w}_h^+}{(b-1)h(1-\alpha_h)} \\ &\quad + \epsilon^2\,\mathbb{E}\left[|m_h^\star|\left[2\rho\frac{\log b}{b-1} - \frac{1-(b-1)h\alpha_h}{1-\alpha_h}\frac{\mathfrak{w}_h^+}{(b-1)h}\right]\right], \\ &= \epsilon^2\frac{\mathfrak{w}_h^+}{(b-1)h(1-\alpha_h)} \\ &\quad + \epsilon^2\,\mathbb{E}\left[|m_h^\star|\right]\left[2\rho\frac{\log b}{b-1} - \frac{1-(b-1)h\alpha_h}{1-\alpha_h}\frac{\mathfrak{w}_h^+}{(b-1)h}\right].\end{aligned}$$

To show that the expectation above can be arbitrarily negative away from 0, we set

$$\alpha_h = b^{-\frac{1+\rho}{2}h},$$

in view of making the height h tend to infinity. Knowing from Lemma 14.20 that $e^{-\frac{1}{2}(x_h+1)^2} \leq \alpha_h$ and classically that $\alpha_h \leq e^{-\frac{1}{2}x_h^2}$, it follows that when $h \to \infty$, the following asymptotic relation holds:

$$\begin{aligned} x_h^2 &= (1+o(1))\, 2\log(\alpha_h^{-1}), \\ &= (1+o(1))\,(1+\rho)h\log b, \\ &= (1+\rho+o(1))\, h\log(b). \end{aligned}$$

Let us show that the expected excess $\mathfrak{w}_h^+$ is asymptotically equivalent to x_h^2, by invoking l'Hospital's rule twice:

$$\begin{aligned} \lim_{x_h\to\infty}\left(\mathfrak{w}_h^+ - x_h^2\right) &= \lim_{x_h\to\infty}\left[\mathbb{E}\left[X^2-1 \,\middle|\, |X|\geq x_h\right] - x_h^2\right], \\ &= \lim_{x_h\to\infty} \frac{\int_{x_h}^\infty (x^2-x_h^2)e^{-\frac{x^2}{2}}\,\mathrm{d}\,x}{\int_{x_h}^\infty e^{-\frac{x^2}{2}}\,\mathrm{d}\,x} - 1, \\ &= \lim_{x_h\to\infty} \frac{-2x_h\int_{x_h}^\infty e^{-\frac{x^2}{2}}\,\mathrm{d}\,x}{-e^{-\frac{x_h^2}{2}}} - 1, \text{ derivative of each term} \\ &= \lim_{x_h\to\infty} 2\frac{\int_{x_h}^\infty e^{-\frac{x^2}{2}}\,\mathrm{d}\,x}{x_h^{-1}e^{-\frac{x_h^2}{2}}} - 1, \\ &= \lim_{x_h\to\infty} 2\frac{-e^{-\frac{x_h^2}{2}}}{-(1+x_h^{-2})e^{-\frac{x_h^2}{2}}} - 1, \text{ derivative of each term} \\ &= 1. \end{aligned}$$

Consequently, when $h \to \infty$, the following asymptotic relations hold:

- $\mathfrak{w}_h^+ := \mathbb{E}\left[X^2-1 \,\middle|\, |X|\geq x_h\right] = (1+\rho+o(1))\, h\log b$
- $\alpha_h(b-1)h = o(1)$
- and $\frac{\mathfrak{w}_h^+}{(b-1)h(1-\alpha_h)} = (1+\rho+o(1))\,\frac{\log b}{b-1}$

Altogether, there are, depending only on ρ and b, a constant $c(\rho,b)$ and a function $h \mapsto r_{\rho,\beta}(h)$ going to 0 when $h \to \infty$ so that,

$$\begin{aligned} \mathbb{E}\left[\gamma(\hat{s}_{m_h^\star}) + \mathrm{pen}\,(m_h^\star)\right] &\leq \epsilon^2 c(\rho,b) + \epsilon^2\,\mathbb{E}\left[|m_h^\star|\right]\left[2\rho\frac{\log b}{b-1} - (1+\rho+r_{\rho,\beta}(h))\,\frac{\log b}{b-1}\right], \\ &= \epsilon^2 c(\rho,b) - \epsilon^2\,\mathbb{E}\left[|m_h^\star|\right](1-\rho+r_{\rho,\beta}(h))\,\frac{\log b}{b-1}. \end{aligned}$$

Since $\rho < 1$ by assumption, there is an integer $\bar{h}(\rho,b)$ depending only on ρ and b so that for any $h > \bar{h}(\rho,b)$ the quantity $1-\rho+r_{\rho,\beta}(h)$ is positive. On the other hand by construction

$$\begin{aligned} \mathbb{E}\left[|m_h^\star|\right] &\geq \mathbb{E}\left[\mathrm{Card}(I_h^\star)\right], \\ &= \alpha_h b^h, \\ &= b^{\frac{1-\rho}{2}h}, \end{aligned}$$

so that for $h > \bar{h}(\rho, b)$ the following relation holds:

$$\mathbb{E}\left[\gamma(\hat{s}_{m_h^\star}) + \mathrm{pen}\,(m_h^\star)\right] \leq \epsilon^2 c(\rho, b) - \epsilon^2 b^{\frac{1-\rho}{2}h}\,(1 - \rho + r_{\rho,\beta}(h))\,\frac{\log b}{b-1}.$$

As a result, $\lim_{h\to\infty} \mathbb{E}\left[\gamma(\hat{s}_{m_h^\star}) + \mathrm{pen}\,(m_h^\star)\right] = -\infty$, and by Fatou's lemma on the countable family $\mathcal{M}$,

$$\mathbb{E}\left[\liminf_{m\in\mathcal{M}} \left[\gamma(\hat{s}_m) + \mathrm{pen}\,(m)\right]\right] = -\infty.$$

Assume that an η-minimiser $\hat{m}_\eta$ of the penalized empirical contrast is still defined almost surely for some $\eta > 0$. Then

$$\begin{aligned}
\mathbb{E}\left[-\left\|\hat{s}_{\hat{m}_\eta}\right\|^2\right] &\leq \mathbb{E}\left[-\left\|\hat{s}_{\hat{m}_\eta}\right\|^2 + \mathrm{pen}\,(\hat{m}_\eta)\right], \\
&\leq \mathbb{E}\left[\liminf_{m\in\mathcal{M}} \left[\gamma(\hat{s}_m) + \mathrm{pen}\,(m)\right]\right] + \eta, \\
&= -\infty.
\end{aligned}$$

Since for instance the following holds a.s.:

$$\left\|\hat{s}_{\hat{m}_\eta}\right\|^2 \leq 2\left\|\hat{s}_{\hat{m}_\eta} - s\right\|^2 + 2\|s\|^2,$$

it follows that

$$\mathbb{E}\left[\left\|\hat{s}_{\hat{m}_\eta} - s\right\|^2\right] = \infty.$$

□

11.5 Proof of Proposition 8.5

Proof. By assumption there is a model $m_0 \in \mathcal{M}$ realizing the minimum $|m_0| = d_0 < n$.

The completion rule in Assumption 8.1 provides with an orthonormal basis $B = (e_1, \cdots, e_n)$ and ensures that for any $m \in \mathcal{M}$ and any sub-basis $b \subset \{e_1, \cdots, e_n\}$ there is a model $m \odot b \in \mathcal{M}$ with $|m \odot b| \leq |m| + \mathrm{C_{ext}}\,\mathrm{Card}(b)$ and $m \oplus \mathrm{Span}(b) \subset m \odot b$.

Since $|m_0| = d_0 < n$, there is a basis vector $e_i \in B$ with $e_i \notin m_0$ so that

$$|m_0 \odot \{e_i\}| \geq |m_0 \oplus \mathrm{Span}\,(\{e_i\})| = |m_0| + 1.$$

As $|m \odot \{e_i\}| \leq |m_0| + \mathrm{C_{ext}} \times 1$, the first assertion in the proposition follows:

$$\mathrm{C_{ext}} \geq |m_0 \odot \{e_i\}| - |m_0| \geq 1.$$

Consider an integer $l \geq 1$ to be chosen later and assume that $d_0 + \mathrm{C_{ext}} l \leq n$. Consider S_l the set of all l-sub-bases of the basis B:

$$S_l = \{b \in \mathcal{P}\,(\{e_1, \ldots, e_n\}), \mathrm{Card}(b) = l\}.$$

Consider the function defined for $b \in S_l$ by $f(b) = m_0 \odot b$. Then,

$$S_l = \bigcup_{m\in f(S_l)} f^{-1}(m)$$

and as $\mathrm{Card}(S_l) = \binom{n}{l}$,

$$\binom{n}{l} = \sum_{m \in f(S_l)} \mathrm{Card}\left(f^{-1}(m)\right). \tag{11.11}$$

For any model $m \in f(S_l)$ and sub-basis $b \in f^{-1}(m)$, the relation $m_0 \odot b = m$ holds so that $\mathrm{Span}(b) \subset m$, in other words

$$b \subset B \cap m.$$

Since $|m| \leq d_0 + \mathrm{C}_{\mathrm{ext}} l$, the model m cannot contain more that $d_0 + \mathrm{C}_{\mathrm{ext}} l$ basis vectors, in other words

$$\mathrm{Card}\left(B \cap m\right) \leq d_0 + \mathrm{C}_{\mathrm{ext}} l.$$

It follows that

$$\begin{aligned}
\mathrm{Card}\left(f^{-1}(m)\right) &= \mathrm{Card}\left(\{b \in S_l, m_0 \odot b = m\}\right), \\
&\leq \mathrm{Card}\left(\{b \in S_l, b \subset B \cap m\}\right), \\
&= \binom{\mathrm{Card}\left(B \cap m\right)}{l}, \\
&\leq \binom{d_0 + \mathrm{C}_{\mathrm{ext}} l}{l},
\end{aligned}$$

and by inequality 11.11 that

$$\binom{n}{l} \leq \mathrm{Card}\left(f(S_l)\right) \binom{d_0 + \mathrm{C}_{\mathrm{ext}} l}{l},$$

and

$$\begin{aligned}
\mathrm{Card}\left(f(S_l)\right) &\geq \binom{n}{l} \binom{d_0 + \mathrm{C}_{\mathrm{ext}} l}{l}^{-1}, \\
&= \frac{n!}{(n-l)!} \frac{(d_0 + (\mathrm{C}_{\mathrm{ext}} - 1)l)!}{(d_0 + \mathrm{C}_{\mathrm{ext}} l)!}, \\
&\geq \left(\frac{n - l + 1}{d_0 + \mathrm{C}_{\mathrm{ext}} l}\right)^l.
\end{aligned} \tag{11.12}$$

For $\beta \geq 0$, consider the sum

$$S_\beta = \sum_{m \in \mathcal{M}, |m| > 0} 2^{-|m|} \left(\frac{en}{|m|}\right)^{-\beta |m|}.$$

This sum appears in Definition 8.4, as the binomial complexity rate satisfies:

$$\mathrm{B}_{\mathcal{M}} = \inf\left\{\beta \geq 0, S_\beta \leq 1\right\}. \tag{11.13}$$

In particular

$$S_\beta \geq \sum_{m \in f(S_l)} 2^{-|m|} \left(\frac{en}{|m|}\right)^{-\beta |m|}.$$

By assumption $1 \leq |m| \leq d_0 + \mathrm{C}_{\mathrm{ext}} l \leq n$ for any $m \in f(S_l)$. Moreover the quantity $2^{-|m|}\left(\frac{en}{|m|}\right)^{-\beta|m|}$ is decreasing with $|m|$ for $1 \leq |m| \leq n$. Combined with the enumeration relation in Inequality 11.12, it follows that

$$S_\beta \geq \left(\frac{n-l+1}{d_0+\mathrm{C}_{\mathrm{ext}} l}\right)^{l} 2^{-d_0-\mathrm{C}_{\mathrm{ext}} l}\left(\frac{en}{d_0+\mathrm{C}_{\mathrm{ext}} l}\right)^{-\beta(d_0+\mathrm{C}_{\mathrm{ext}} l)}. \tag{11.14}$$

Transforming Inequality 11.14 shows that the sum S_β is larger than 1 for $\beta < \beta^\star$, where

$$\begin{aligned}\beta^\star &= \frac{1}{d_0+\mathrm{C}_{\mathrm{ext}} l}\frac{l\log\left(\frac{n-l+1}{d_0+\mathrm{C}_{\mathrm{ext}} l}\right)-(d_0+\mathrm{C}_{\mathrm{ext}} l)\log(2)}{\log\left(\frac{n}{d_0+\mathrm{C}_{\mathrm{ext}} l}\right)+1},\\ &= \frac{1}{\frac{d_0}{l}+\mathrm{C}_{\mathrm{ext}}}\left[1-\frac{\log\left(\frac{n}{n-l+1}\right)+\left(\frac{d_0}{l}+\mathrm{C}_{\mathrm{ext}}\right)\log(2)+1}{\log\left(\frac{n}{l}\right)-\log\left(\frac{d_0}{l}+\mathrm{C}_{\mathrm{ext}}\right)+1}\right].\end{aligned} \tag{11.15}$$

By Equation 11.13 $\mathrm{B}_{\mathcal{M}} \geq \beta^\star$.

Case $d_0 = 0$ If $d_0 = 0$, we may choose $l = 1$, so that

$$\beta^\star = \frac{1}{\mathrm{C}_{\mathrm{ext}}}\left[1-\frac{1+\mathrm{C}_{\mathrm{ext}}\log(2)}{1+\log(n)-\log(\mathrm{C}_{\mathrm{ext}})}\right],$$

and the number $\beta^\star$ is positive under the condition

$$n > \mathrm{C}_{\mathrm{ext}} 2^{\mathrm{C}_{\mathrm{ext}}}.$$

Note that this condition implies that $n \geq \mathrm{C}_{\mathrm{ext}} = d_0 + \mathrm{C}_{\mathrm{ext}} l$, as assumed when introducing the integer l.

Case $d_0 > 0$ In the case where $d_0 > 0$, stronger assumptions will help us exhibit the desired asymptotic lower bound under $\mathrm{B}_{\mathcal{M}}$. Assume for instance that $d_0 \leq \lfloor\sqrt{n}\rfloor$ and choose $l = \lfloor\sqrt{n}\rfloor$, under the temporary assumption that $d_0 + \mathrm{C}_{\mathrm{ext}} l < n$.

Noting in Equation 11.15 that $l = \lfloor\sqrt{n}\rfloor$, that $\frac{d_0}{l} \leq 1$ and that:

$$\begin{aligned}\log\left(\frac{n}{n-l+1}\right) &= \log\left(1+\frac{l-1}{n-l+1}\right),\\ &= \log\left(1+\frac{\lfloor\sqrt{n}\rfloor-1}{n-\lfloor\sqrt{n}\rfloor+1}\right),\\ &\leq \frac{\lfloor\sqrt{n}\rfloor-1}{n-\lfloor\sqrt{n}\rfloor+1},\\ &\leq \frac{1}{\lfloor\sqrt{n}\rfloor},\\ &\leq 1,\end{aligned}$$

leads to:

$$\beta^\star \geq \frac{1}{\frac{d_0}{\lfloor\sqrt{n}\rfloor} + \mathrm{C_{ext}}} \left[1 - \frac{(1+\mathrm{C_{ext}})\log(2) + 2}{\frac{1}{2}\log(n) - \log(1+\mathrm{C_{ext}}) + 1}\right],$$

which is positive if

$$n > \left[e(1+\mathrm{C_{ext}})2^{1+\mathrm{C_{ext}}}\right]^2.$$

Note that this condition implies that $1+\mathrm{C_{ext}} < \sqrt{n}$ so that $d_0+\mathrm{C_{ext}}l \leq \sqrt{n}+\mathrm{C_{ext}}\sqrt{n} \leq n$ as assumed when introducing the integer l.

Tightness To prove the tightness of the bounds above, assume that n is a multiple of an integer c and consider an orthonormal basis B of $\mathbb{R}^n$, a partition P_c of B in $\frac{n}{c}$ parts each of cardinal c and $\mathcal{B}_c$ the set of sub-bases formed by unions of parts present in the partition P_c. Consider the model family 'by blocs':

$$\mathcal{M} = \{\mathrm{Span}(p)\}_{p\in\mathcal{B}_c}.$$

This model family satisfies the completion rule with $\mathrm{C_{ext}} = c$, since to extend a model with a given basis vector, it is enough to add the part (bloc) where this vector belongs in the partition P_c. Moreover, for $\beta > 0$, for any integer k with $kc \leq n$, there are exactly $\binom{\frac{n}{c}}{k} \leq \left(\frac{en}{ck}\right)^k$ models of dimension kc. It follows that:

$$\begin{aligned} S_\beta &= \sum_{m\in\mathcal{M},|m|>0} 2^{-|m|}\left(\frac{en}{|m|}\right)^{-\beta|m|}, \\ &\leq \sum_{1\leq ck\leq n}\left(\frac{en}{ck}\right)^k\left(\frac{en}{ck}\right)^{-\beta ck}2^{-ck}, \\ &\leq \sum_{1\leq ck\leq n}\left(\frac{en}{ck}\right)^{(1-\beta c)k}2^{-ck}, \end{aligned}$$

so that S_β is less than one for $\beta = \frac{1}{c} = \frac{1}{\mathrm{C_{ext}}}$. and $\mathrm{B}_\mathcal{M} \leq \frac{1}{\mathrm{C_{ext}}}$.

□

11.6 Proof of Proposition 8.6

Proof. As the model family is finite, a minimizer of the penalized empirical contrast exists almost surely whatever the penalty function. By Definition 8.4 and its following remarks, the set of weights $\left\{L_m = \log 2 + \mathrm{B}_\mathcal{M}\log\frac{en}{|m|}\right\}_{m\in\mathcal{M}}$ satisfies $\sum_{m\in\mathcal{M},|m|>0} e^{-|m|L_m} < 1$. This allows to directly apply Corollary 5.7 with parameter η and the set of weights above, ensuring that if:

$$\mathrm{pen}(m) = \frac{1+\eta}{1-\eta}\epsilon^2|m|\left(1+2\sqrt{L_m}+2L_m\right)\ \forall m\in\mathcal{M}.$$

then the following relation holds:

$$\eta\,\mathbb{E}\left[\|s-\tilde{s}\|^2\right] \leq \inf_{m\in\mathcal{M}}\left\{\mathrm{d}^2\left(s,S_m\right)+\mathrm{pen}\left(m\right)-\epsilon^2\left|m\right|\right\}+\epsilon^2\frac{(1+3\eta)}{\eta(1-\eta)},$$
$$= \inf_{m\in\mathcal{M}}\left\{\mathrm{d}^2\left(s,S_m\right)+\epsilon^2\left|m\right|\frac{1+\eta}{1-\eta}\left(\frac{2\eta}{1+\eta}+2\sqrt{L_m}+2L_m\right)\right\}+\epsilon^2\frac{(1+3\eta)}{\eta(1-\eta)},$$

The second inequality in the proposition follows by the relation $1+x\geq 2\sqrt{x}$ for $x>0$ applied to $x=L_m$, and the bound $\frac{2\eta}{1+\eta}\leq 1$ which amounts to neglect the contribution of the term $-\epsilon^2\left|m\right|$:

$$\eta\,\mathbb{E}\left[\|s-\tilde{s}\|^2\right] \leq \inf_{m\in\mathcal{M}}\left\{\mathrm{d}^2\left(s,S_m\right)+\epsilon^2\left|m\right|\frac{1+\eta}{1-\eta}\left(2+3L_m\right)\right\}+\epsilon^2\frac{(1+3\eta)}{\eta(1-\eta)},$$
$$= \inf_{m\in\mathcal{M}}\left\{\mathrm{d}^2\left(s,S_m\right)+\epsilon^2\left|m\right|\frac{1+\eta}{1-\eta}\left(2+3\log(2)+3\mathrm{B}_{\mathcal{M}}\log\frac{en}{|m|}\right)\right\}+\epsilon^2\frac{(1+3\eta)}{\eta(1-\eta)},$$
$$\leq \inf_{m\in\mathcal{M}}\left\{\mathrm{d}^2\left(s,S_m\right)+\epsilon^2\left|m\right|\frac{1+\eta}{1-\eta}\left(5+3\mathrm{B}_{\mathcal{M}}\log\frac{en}{|m|}\right)\right\}+\epsilon^2\frac{(1+3\eta)}{\eta(1-\eta)}.$$

If the source signal s belongs to an exact model $\tilde{m}$ then $d(s,S_{\tilde{m}})=0$ and the third relation announced in the proposition follows:

$$\mathbb{E}\left[\|s-\hat{s}_{\hat{m}}\|^2\right]\leq\epsilon^2\left|\tilde{m}\right|\frac{1+\eta}{1-\eta}\left(5+3\mathrm{B}_{\mathcal{M}}\log\frac{en}{|\tilde{m}|}\right)+\epsilon^2\frac{(1+3\eta)}{\eta(1-\eta)}.$$

□

11.7 Proof of Proposition 8.7

The proof of Proposition 8.7 relies on Lemma 11.1 that we establish before concluding in Section 11.7.3.

11.7.1 A risk lower bound lemma for nested models

The following lemma will help us to link the risk of the selected model $\hat{m}$ to the risk of any random extension of the same $\hat{m}$. Subsequently, we will rely on Assumption 8.1 to build such an extension. Recall that in the setup of the model selection procedure described in Section 5.1, one observes $Y=s+\epsilon W$ where s is an unknown fixed parameter, W an orthonormal Gaussian process and ϵ a noise parameter. One tries to select a model within a family $\mathcal{M}$ by finding a minimizer $\hat{m}$ over $m\in\mathcal{M}$ of the penalized empirical contrast criterion $-\|\hat{s}_m\|^2+\mathrm{pen}\left(m\right)=-\|\Pi_m(s+\epsilon W)\|^2+\mathrm{pen}\left(m\right)$, where $\Pi_m(\cdot)$ denotes the orthogonal projector on the Euclidean sub-space S_m.

Lemma 11.1 (Risk lower bound for nested models)**.** *In the setup of the model selection procedure described in Section 5.1, assume a minimizer $\hat{m}$ of the penalized least squares criterion $-\|\hat{s}_m\|^2+\mathrm{pen}\left(m\right)$ over $m\in\mathcal{M}$ exists almost surely. Assume specified m' a random model element of $\mathcal{M}$ satisfying $S_{\hat{m}}\subset S_{m'}$ a.s..*
Then for any $\eta\geq 1$ the following relation holds a.s.:

$$(\eta-1)\|s-\hat{s}_{\hat{m}}\|^2\geq\left(1-\frac{1}{\eta}\right)\epsilon^2\|\Pi_{m'}(W)\|^2-\mathrm{pen}\left(m'\right)+\mathrm{pen}\left(\hat{m}\right).$$

Moreover, if the source function s is null, the following relation holds a.s.:

$$\|\hat{s}_{\hat{m}}\|^2 \geq \epsilon^2 \|\Pi_{m'}(W)\|^2 - \text{pen}\,(m') + \text{pen}\,(\hat{m})\,.$$

The proof is given in Section 11.7.2.

11.7.2 Proof of Lemma 11.1

Proof. By the definition of $\hat{m}$ and $\hat{s}_{\hat{m}}$, the following relations hold a.s.:

$$\begin{aligned} \|s - \hat{s}_{\hat{m}}\|^2 &= \|s - \Pi_{\hat{m}}(s + \epsilon W)\|^2, \\ &= \|(\mathbb{I}_n - \Pi_{\hat{m}})(s) - \Pi_{\hat{m}}(\epsilon W)\|^2, \\ &= \|s - \Pi_{\hat{m}}(s)\|^2 + \epsilon^2 \|\Pi_{\hat{m}}(W)\|^2, \end{aligned} \tag{11.16}$$

and since $\hat{m}$ is a minimizer of the penalized empirical contrast,

$$\|\Pi_{\hat{m}}(s + \epsilon W)\|^2 - \text{pen}\,(\hat{m}) \geq \|\Pi_{m'}(s + \epsilon W)\|^2 - \text{pen}\,(m')\,. \tag{11.17}$$

By assumption the model $S_{m'}$ extends the model $S_{\hat{m}}$, so that $\Pi_{m'} - \Pi_{\hat{m}}$ is an orthogonal projector, and the projector $\Pi_{m'}$ admits the decomposition

$$\Pi_{m'} = \Pi_{\hat{m}} + \Pi_{m'}(\mathbb{I}_n - \Pi_{\hat{m}}) \text{ a.s.}, \tag{11.18}$$

with

$$\Pi_{\hat{m}}(\Pi_{m'} - \Pi_{\hat{m}}) = 0 \text{ a.s.}. \tag{11.19}$$

Combining with the contrast relation in Inequality 11.17 above shows that almost surely:

$$\text{pen}\,(m') - \text{pen}\,(\hat{m}) \geq \|(\Pi_{m'} - \Pi_{\hat{m}})(s + \epsilon W)\|^2,$$

and by the inequality $2\langle x, y\rangle \leq \eta\|x\|^2 + \frac{1}{\eta}\|y\|^2$, for $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$

$$\begin{aligned} \text{pen}\,(m') - \text{pen}\,(\hat{m}) &\geq -(\eta - 1)\|(\Pi_{m'} - \Pi_{\hat{m}})(s)\|^2 + \left(1 - \frac{1}{\eta}\right)\epsilon^2 \|(\Pi_{m'} - \Pi_{\hat{m}})(W)\|^2, \\ &= -(\eta - 1)\|(\Pi_{m'} - \Pi_{\hat{m}})(s)\|^2 + \left(1 - \frac{1}{\eta}\right)\epsilon^2 \|\Pi_{m'}(W)\|^2 \\ &\quad - \left(1 - \frac{1}{\eta}\right)\epsilon^2 \|\Pi_{\hat{m}}(W)\|^2, \end{aligned}$$

and after reorganizing the terms:

$$\begin{aligned} (\eta - 1)\|\Pi_{m'}(s) - \Pi_{\hat{m}}(s)\|^2 + \left(1 - \frac{1}{\eta}\right)\epsilon^2 \|\Pi_{\hat{m}}(W)\|^2 \geq \left(1 - \frac{1}{\eta}\right)\epsilon^2 \|\Pi_{m'}(W)\|^2 \\ - (\text{pen}\,(m') - \text{pen}\,(\hat{m})). \end{aligned}$$

Since $\eta \geq 1$ it follows that:

$$\begin{aligned} (\eta - 1)\left[\|\Pi_{m'}(s) - \Pi_{\hat{m}}(s)\|^2 + \epsilon^2 \|\Pi_{\hat{m}}(W)\|^2\right] \geq \left(1 - \frac{1}{\eta}\right)\epsilon^2 \|\Pi_{m'}(W)\|^2 \\ - (\text{pen}\,(m') - \text{pen}\,(\hat{m})). \end{aligned} \tag{11.20}$$

Since the projector $\Pi_{m'}$ extends $\Pi_{\hat{m}}$, the following relation holds a.s. (see Equations 11.18 and 11.19):

$$\|s - \Pi_{\hat{m}}(s)\|^2 = \|s - \Pi_{m'}(s)\|^2 + \|(\Pi_{m'} - \Pi_{\hat{m}})(s)\|^2,$$

so that Inequality 11.20 implies:

$$(\eta - 1)\left[\|s - \Pi_{\hat{m}}(s)\|^2 + \epsilon^2 \|\Pi_{\hat{m}}(W)\|^2\right] \geq \left(1 - \frac{1}{\eta}\right) \epsilon^2 \|\Pi_{m'}(W)\|^2 - (\mathrm{pen}\,(m') - \mathrm{pen}\,(\hat{m}))\ \text{a.s.}.$$

which is the announced statement since by Equation 11.16:

$$\|s - \Pi_{\hat{m}}(s)\|^2 + \epsilon^2 \|\Pi_{\hat{m}}(W)\|^2 = \|s - \hat{s}_{\hat{m}}\|^2.$$

Null source function s If the source function s is null, then the second relation in the Lemma is simply Equation 11.17. □

11.7.3 Concluding the proof of Proposition 8.7

Proof. Relying on the extension rule in Assumption 8.1, we proceed by constructing an extension m_e of the selected model $\hat{m}$ by the basis-vectors that support the largest components of the noise. The risk relation for nested models in Lemma 11.1 provides a link between the respective quadratic risks of these two random models.

Consider the orthonormal basis $B = (u_1, \cdots, u_n)$ provided by assumption 8.1, and the standard Gaussian sequence $\xi_1, \cdots, \xi_n = \langle u_1, W\rangle, \cdots, \langle u_n, W\rangle$. As in Definition 15.1 $\sigma_n(\cdot)$ is a random permutation satisfying

$$\xi^2_{\sigma_n(1)} \leq \cdots \leq \xi^2_{\sigma_n(n)}\ \text{a.s.}.$$

We denote $b^\star \subset B$ the random orthonormal sub-basis

$$b^\star := \left\{u_{\sigma_n(n-l+1)}, \cdots, u_{\sigma_n(n)}\right\},$$

and $\Pi_{b^\star}$ the associated random orthogonal projector

$$\Pi_{b^\star} := \sum_{u \in b^\star} u \otimes u.$$

Informally, $b^\star$ represents the location of the largest components of the noise vector W in the basis B which represents the feature space. Relying on the completion rule in Assumption 8.1, we define the (random) extended model

$$S_{m_e} = S_{\hat{m} \odot b^\star}.$$

Its dimension, by the same completion rule, as $\mathrm{Card}\, b^\star = l$, is bound by

$$|\hat{m}| \vee l \leq |m_e| \leq |\hat{m}| + \mathrm{C}_{\mathrm{ext}} l, \text{a.s.}.$$

Set

$$\begin{aligned}\theta &= 1 + 2\log(2) + 2t, \\ &\simeq 2.39 + 2t.\end{aligned} \tag{11.21}$$

For future reference, this ensures that:

$$\begin{aligned}\left[\exp\left(\frac{\theta^2}{2(1+\theta)}\right) - 1\right]^{-1} &\leq \left[\exp\left(\frac{\theta^2-1}{2(1+\theta)}\right) - 1\right]^{-1}, \\ &= \left[\exp\left(\frac{\theta-1}{2}\right) - 1\right]^{-1}, \\ &= \left[2e^t - 1\right]^{-1}, \\ &\leq e^{-t}.\end{aligned} \tag{11.22}$$

The projection $\|\Pi_{b^\star}(W)\|^2$ has the same distribution than the sum of the l largest terms of a sample of the chi-square distribution. Moreover the numbers n, l and θ satisfy the conditions for Lemma 15.3 which are $\theta > 2.06$ and $l \leq \frac{n}{e^{\frac{1}{2}(1+\theta)}}$, so that the following inequality holds:

$$\begin{aligned}\mathbb{P}\left[\|\Pi_{b^\star}(W)\|^2 < [1 - r(n,l,\theta)]\, l\left(1 + 2\log\frac{n}{(1+\theta)l}\right)\right] &\leq \left[\exp\left(\frac{\theta^2}{2(1+\theta)}\right) - 1\right]^{-1}, \\ &< e^{-t},\end{aligned} \tag{11.23}$$

where the function $r(\cdot,\cdot,\cdot)$ is defined in Lemma 15.3. Recall that by the same lemma, this function goes to 0 when $\frac{n}{l(1+\theta)}$ goes to infinity. Since by the completion rule the extended model S_{m_e} contains the subspace $\operatorname{Span} b^\star$, the inequality above applies *a fortiori* to $\|\Pi_{m_e}(W)\|^2$, so that

$$\mathbb{P}\left[\|\Pi_{m_e}(W)\|^2 < [1 - r(n,l,\theta)]\, l\left(1 + 2\log\frac{n}{(1+\theta)l}\right)\right] < e^{-t}, \tag{11.24}$$

Moreover, since $\eta > 1$ and $S_{\hat{m}} \subset S_{m_e}$ a.s., the first risk relation in Lemma 11.1 ensures that:

$$(\eta - 1)\|s - \hat{s}_{\hat{m}}\|^2 \geq \left(1 - \frac{1}{\eta}\right)\epsilon^2\|\Pi_{m_e}(W)\|^2 - \operatorname{pen}(m_e) + \operatorname{pen}(\hat{m}) \text{ a.s..} \tag{11.25}$$

Combining Inequalities 11.24 and 11.25 above ensures that the relation:

$$\begin{aligned}(\eta - 1)\|s - \hat{s}_{\hat{m}}\|^2 \geq \left(1 - \frac{1}{\eta}\right) l\,[1 - r(n,l,\theta)]\left(1 + 2\log\frac{n}{(1+\theta)l}\right) \\ - \operatorname{pen}(m_e) + \operatorname{pen}(\hat{m}),\end{aligned} \tag{11.26}$$

holds apart from an event Ω of probability less than e^{-t}.

If additionally $s = 0$, by the second risk relation in Lemma or simply by the definition of the minimizer $\hat{m}$, 11.1

$$\|\hat{s}_{\hat{m}}\|^2 \geq \epsilon^2\|\Pi_{m_e}(W)\|^2 - \operatorname{pen}(m_e) + \operatorname{pen}(\hat{m}) \text{ a.s.}, \tag{11.27}$$

and by 11.24 and 11.27, on the same event, if $s = 0$,

$$\begin{aligned}\|\hat{s}_{\hat{m}}\|^2 \geq l\,[1 - r(n,l,\theta)]\left(1 + 2\log\frac{n}{(1+\theta)l}\right)\\ - \operatorname{pen}(m_e) + \operatorname{pen}(\hat{m}),\end{aligned} \tag{11.28}$$

By assumption in the lemma,

- $\operatorname{pen}(m) = f(|m|) \quad \forall m \in \mathcal{M}$,
- $\frac{f(|m|)}{|m|}$ decreases with $|m|$,
- $0 \leq f(l) \leq \rho\epsilon^2 l\left[1 + 2\log\frac{n}{(1+\theta)l}\right]$.

Moreover by their construction based on Assumption 8.1, the models $\hat{m}$ and m_e satisfy a.s. the four relations:

$$\begin{aligned} S_{\hat{m}} &\subset S_{m_e},\\ |\hat{m}| &\leq |m_e|,\\ l &\leq |m_e|,\\ |m_e| - |\hat{m}| &\leq \mathrm{C}_{\mathrm{ext}} l.\end{aligned}$$

It follows that

$$\begin{aligned}\operatorname{pen}(m_e) - \operatorname{pen}(\hat{m}) &= f(|m_e|) - f(|\hat{m}|),\\ &\leq (|m_e| - |\hat{m}|)\frac{f(|m_e|)}{|m_e|},\\ &\leq (|m_e| - |\hat{m}|)\frac{f(l)}{l},\\ &\leq \mathrm{C}_{\mathrm{ext}} l\frac{f(l)}{l},\\ &\leq (|m_e| - |\hat{m}|)\epsilon^2\rho\,[1 - r(n,l,\theta)]\left[1 + 2\log\frac{n}{(1+\theta)l}\right].\end{aligned} \tag{11.29}$$

Combining with Inequality 11.28 shows that the following relation holds apart of the event Ω:

$$(\eta - 1)\|s - \hat{s}_{\hat{m}}\|^2 \geq \epsilon^2\,[1 - r(n,l,\theta)]\left[1 - \frac{1}{\eta} - \rho\mathrm{C}_{\mathrm{ext}}\right] l\left(1 + 2\log\frac{n}{(1+\theta)l}\right), \tag{11.30}$$

In Inequality 11.30 the number η can be chosen arbitrarily larger than 1. Choosing for instance

$$\eta = \frac{2}{1 - \rho\mathrm{C}_{\mathrm{ext}}},$$

ensures that

$$\begin{aligned}\frac{1 - \frac{1}{\eta} - \rho\mathrm{C}_{\mathrm{ext}}}{\eta - 1} &= \frac{(1 - \rho\mathrm{C}_{\mathrm{ext}})^2}{2(1 + \rho\mathrm{C}_{\mathrm{ext}})},\\ &\geq \frac{(1 - \rho\mathrm{C}_{\mathrm{ext}})^2}{4},\end{aligned}$$

so that the following relation holds on the event Ω:

$$\|s - \hat{s}_{\hat{m}}\|^2 \geq \epsilon^2 \frac{1}{2} \frac{(1 - \rho \mathrm{C}_{\mathrm{ext}})^2}{1 + \rho \mathrm{C}_{\mathrm{ext}}} l\epsilon^2 \left[1 - r(n, l, \theta)\right] (1 + 2 \log \frac{n}{(1+\theta)l}),$$

and since $\theta = 1 + 2\log(2) + 2t$, the case $s \neq 0$ of Inequality 8.4 announced in the lemma follows:

$$\|s - \hat{s}_{\hat{m}}\|^2 \geq \epsilon^2 \frac{(1 - \rho \mathrm{C}_{\mathrm{ext}})^2}{4} l \left[1 - r(n, l, \theta)\right] \left(1 + 2 \log \frac{n}{(1+\theta)l}\right).$$

In the the case $s = 0$, combining Inequalities 11.28 and 11.29 leads to Inequality 8.4 announced in the lemma:

$$\|\hat{s}_{\hat{m}}\|^2 \geq \epsilon^2 \left[1 - r(n, l, \theta)\right] \left[1 - \rho \mathrm{C}_{\mathrm{ext}}\right] l \left(1 + 2 \log \frac{n}{(1+\theta)l}\right). \tag{11.31}$$

□

11.8 Proof of Proposition 8.8

The proof of Proposition 8.8 relies on Proposition 8.7 and on Lemma 11.2 that we establish before concluding in Section 11.8.2.

11.8.1 Controlling the contrast contribution of low dimensional models

The following lemma will allow us to control the contrast contribution of low dimensional models, in a classical way. Recall that Definition 8.4 introduces the binomial complexity rate $\mathrm{B}_{\mathcal{M}}$ for any finite model family $\mathcal{M}$ so that the weights $\left\{ L_m = \log 2 + \mathrm{B}_{\mathcal{M}} \log \left(\frac{en}{|m|} \right) \right\}_{m \in \mathcal{M}}$ satisfy $\sum_{m \in \mathcal{M}} e^{-|m| L_m} \leq 1$.

Lemma 11.2 (Low dimensional models)**.** *For any model $m \in \mathcal{M}$, denote Π_m the orthogonal projector on S_m. Consider some set of weights $\{L_m\}_{m \in \mathcal{M}}$ satisfying*

$$\sum_{m \in \mathcal{M}, |m| > 0} e^{-|m| L_m} \leq 1.$$

Then for any $t > 0$ and any $\eta > 0$ the following inequalities hold uniformly over the family $\mathcal{M}$:

$$\|\Pi_m(W)\|^2 \leq |m| \left(1 + 2\sqrt{L_m} + 2L_m\right) + 2\sqrt{|m|\, t} + 2t,$$

and

$$\|\Pi_m(W)\|^2 \leq |m| \left(1 + \eta + 2\sqrt{L_m} + 2L_m\right) + (2 + \eta^{-1})t,$$

apart from an event of probability less than e^{-t}.
The preceding statement holds with $L_m = \log 2 + \mathrm{B}_{\mathcal{M}} \log \left(\frac{en}{|m|} \right)$ $\forall m \in \mathcal{M}$.

Proof of Lemma 11.2. We know from Lemma 14.10 that if a random variable V has for distribution $\chi^2(D)$, then for any positive t the following probability bound holds:

$$\mathbb{P}\left[V \geq D + 2\sqrt{Dt} + 2t\right] \leq e^{-t}.$$

It follows that for any positive t, the following inequality holds for each $m \in \mathcal{M}$:

$$\mathbb{P}\left[\|\Pi_m(W)\|^2 \geq |m| + 2\sqrt{|m|\, t} + 2t\right] \leq e^{-t}.$$

Additionally, since $t > 0$, if a null model of zero dimension is element of $\mathcal{M}$, it fulfills the relation above with probability bound 0. After choosing for each model $t_m = |m|\, L_m + t$ and recalling that the family $(L_m)_{m\in\mathcal{M}}$ satisfies $\sum_{m\in\mathcal{M},|m|>0} e^{-|m|L_m} \leq 1$, a union bound ensures that uniformly over the models in the family $\mathcal{M}$,

$$\|\Pi_m(W)\|^2 \leq \left[|m| + 2\sqrt{|m|\,(|m|\, L_m + t)} + 2(|m|\, L_m + t)\right]$$

apart from an event of probability less than $\sum_{m\in\mathcal{M},|m|>0} e^{-|m|L_m - t} \leq e^{-t}$.

Relying on the inequalities $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $2\sqrt{ab} \leq \alpha^{-1}a + \alpha b$ for positive a, b and α, we note that

$$\begin{aligned} 2\sqrt{|m|\,(|m|\, L_m + t)} &\leq |m|\, 2\sqrt{L_m} + 2\sqrt{|m|\, t}, \\ &\leq |m|\left(2\sqrt{L_m} + \eta\right) + \eta^{-1}t. \end{aligned}$$

so that for any $t > 0$ the following inequalities holds uniformly over the family $\mathcal{M}$ apart from an event of probability less than e^{-t}:

$$\|\Pi_m(W)\|^2 \leq |m|\left(1 + 2L_m + 2\sqrt{L_m}\right) + 2\sqrt{|m|\, t} + 2t,$$

and

$$\|\Pi_m(W)\|^2 \leq |m|\left(1 + \eta + 2L_m + 2\sqrt{L_m}\right) + (2 + \eta^{-1})t.$$

Finally, the preceding statement applies to the particular choice of weights $\left\{L_m = \log 2 + \mathrm{B}_{\mathcal{M}} \log\left(\frac{en}{|m|}\right)\right\}_{m\in\mathcal{M}}$, as Definition 8.4 ensures that $\sum_{m\in\mathcal{M}} e^{-|m|L_m} \leq 1$ in this case. □

11.8.2 Concluding the proof of Proposition 8.8

Proof. Set

$$\theta = 1 + 2\log(2) + 2t.$$

Under the assumptions of Proposition 8.7, the risk relation in the same proposition for the case $s = 0$ holds apart of an event of probability less than e^{-t}:

$$\|\Pi_{\hat{m}}(W)\|^2 \geq l(1 - \rho \mathrm{C}_{\mathrm{ext}})\left[1 - r(n, l, \theta)\right]\left(1 + 2\log\frac{n}{(1+\theta)l}\right). \tag{11.32}$$

The uniform bound in Lemma 11.2 applies in particular to the selected model $\hat{m}$, so that with the parameter η set to $\eta = 1$, apart from an event of probability less than e^{-t},

$$\|\Pi_{\hat{m}}(W)\|^2 \leq |\hat{m}| \left(2 + 2\sqrt{L_{\hat{m}}} + 2L_{\hat{m}}\right) + 3t,$$

with for $|\hat{m}| \geq 1$, $L_{\hat{m}} = \log 2 + \mathrm{B}_{\mathcal{M}} \log\left(\frac{en}{|\hat{m}|}\right)$. Hence on the same event, by the relation $2\sqrt{L_{\hat{m}}} \leq 1 + L_{\hat{m}}$:

$$\|\Pi_{\hat{m}}(W)\|^2 \leq |\hat{m}|\, 3\,(1 + \log 2 + \mathrm{B}_{\mathcal{M}} \log en) + 3t. \tag{11.33}$$

By a union bound between Inequalities 11.32 and 11.33 above, it follows that apart from an event of probability less than $2e^{-t}$, the following relation holds:

$$|\hat{m}|\, 3(1 + \log 2 + \mathrm{B}_{\mathcal{M}} \log en) + 3t \geq l(1 - \rho \mathrm{C}_{\mathrm{ext}})\,[1 - r(n, l, \theta)] \left(1 + 2 \log \frac{n}{(1+\theta)l}\right) \tag{11.34}$$

so that apart from the same event, Inequality 8.9 announced in the proposition in proof follows:

$$\begin{aligned}
|\hat{m}| &\geq l(1 - \rho \mathrm{C}_{\mathrm{ext}})\,[1 - r(n, l, \theta)]\, \frac{1}{3} \frac{1 + 2 \log \frac{n}{(1+\theta)l}}{1 + \log(2) + \mathrm{B}_{\mathcal{M}} \log(en)} - 3t, \\
&\geq l(1 - \rho \mathrm{C}_{\mathrm{ext}})\,[1 - r(n, l, \theta)]\, \frac{1}{3} \frac{2 \log \frac{n}{(1+\theta)l}}{\mathrm{B}_{\mathcal{M}} \log(en) + 2} - 3t, \\
&\geq l \frac{2}{3} (1 - \rho \mathrm{C}_{\mathrm{ext}})\,[1 - r(n, l, \theta)]\, \frac{\log \frac{n}{(1+\theta)l}}{\mathrm{B}_{\mathcal{M}} \log(en) + 2} - 3t,
\end{aligned} \tag{11.35}$$

Assume that $l = \left\lfloor \frac{n^\alpha}{e(1+\theta)} \right\rfloor$ for some α with $0 < \alpha < 1$, Then

$$\begin{aligned}
\frac{\log \frac{n}{l(1+\theta)}}{\mathrm{B}_{\mathcal{M}} \log(en) + 2} &\geq \frac{\log en^{1-\alpha}}{\mathrm{B}_{\mathcal{M}} \log(en) + 2}, \\
&\geq \frac{1 - \alpha}{\mathrm{B}_{\mathcal{M}} + \frac{2}{\log(en)}},
\end{aligned}$$

and on the same event, Inequality 11.35 implies Inequality 8.9 in the proposition:

$$|\hat{m}| \geq (1 - \alpha) \left\lfloor \frac{n^\alpha}{e(1+\theta)} \right\rfloor \frac{2}{3} \frac{1 - \rho \mathrm{C}_{\mathrm{ext}}}{\mathrm{B}_{\mathcal{M}} + \frac{2}{\log(en)}}\,[1 - r(n, l, \theta)] - 3t. \tag{11.36}$$

□

12 Proofs for section: 9 A first extension to histogram selection ...

12.1 Proof of Proposition 9.13

Proof. Proposition 9.13 derives from Castellan's model selection Theorem 9.7. Assumption 9.9 ensures that the model family has the required structure of histograms formed by unions of the elements of a predefined partition of $[0,1]$ in $(N+1)$ parts of same measure $\frac{1}{N+1}$. The assumption that $\frac{n}{N} \geq \log(n)^2$ is also identical.

By the definition of the binomial complexity rate in Definition 9.12, the following inequality holds:

$$\sum_{m\in\mathcal{M},|m|>1} 2^{-(|m|-1)} \left(\frac{eN}{|m|-1}\right)^{-\mathrm{B}_{\mathcal{M}}(|m|-1)} \leq 1$$

so that the weights

$$x_m = \begin{cases} 0 & \text{if } |m| = 1, \\ (|m|-1)\left[\log(2) + \mathrm{B}_{\mathcal{M}} \log\left(\frac{eN}{|m|-1}\right)\right] & \text{otherwise,} \end{cases}$$

satisfy the relations:

$$\sum_{m\in\mathcal{M},|m|>1} e^{-x_m} \leq 1,$$
$$\sum_{m\in\mathcal{M}} e^{-x_m} \leq 2.$$

Moreover the source density s is equal to 1 almost everywhere, so that for the application of Theorem 9.7 the number ρ can be chosen equal to 1. If on one hand the numbers c_1 and c_2 satisfy $c_1 > \frac{1}{2}$ and $c_2 = 2(1+c_1^{-1})$ and on the other hand the penalty function satisfies $\forall m \in \mathcal{M}$:

$$\begin{aligned} \mathrm{pen}\,(m) &= \frac{c_1}{n}\left(\sqrt{D_m} + \sqrt{c_2 x_m}\right)^2, \\ &= \frac{c_1}{n}(|m|-1)\left(1 + \sqrt{2(1+c_1^{-1})\left[\log(2) + \mathrm{B}_{\mathcal{M}} \log\left(\frac{eN}{|m|-1}\right)\right]}\right)^2 \end{aligned}$$

then by Theorem 9.7- for some constant $C_1(c_1) = C(c_1, \rho = 1, L = 0, \Sigma = 2)$, the following holds:

$$\frac{1}{2}\,\mathbb{E}\left[\mathbf{h}^2(s, \hat{s}_{\hat{m}})\right] \leq \frac{(2c_1)^{1/5}}{(2c_1)^{1/5}-1} \inf_{m\in\mathcal{M}} \left\{\mathbf{K}(s, s_m) + \mathrm{pen}\,(m)\right\} + \frac{C_1(c_1)}{n}.$$

Denote m_0 the model associated with the singleton partition $\{[0,1]\}$. As $s = s_{m_0} = 1$ almost everywhere and $\mathrm{pen}\,(m_0) = 0$, then $\mathbf{K}(s, s_{m_0}) + \mathrm{pen}\,(m_0) = 0$, so that $\inf_{m\in\mathcal{M}} \left\{\mathbf{K}(s, s_m) + \mathrm{pen}\,(m)\right\} = 0$. Plugging in the contrast relation above yields:

$$\frac{1}{2}\,\mathbb{E}\left[\mathbf{h}^2(s, \hat{s}_{\hat{m}})\right] \leq \frac{C_1(c_1)}{n}.$$

□

12.2 Proof of Proposition 9.14

To prepare the proof of Proposition 9.14 we present some definitions and technical lemmas, before concluding the proof itself in section 12.2.6.

12.2.1 Deviation count, deviation rate

Let us first introduce events to represent the occupancy of one specific partition interval by more than k sample elements. Recall that in all the histograms in the model family $\mathcal{M}$, the bins are unions of elementary parts taken in a m_N, a partition of $[0,1]$ into of $N+1$ elementary segments (parts) of common measure $\frac{1}{N+1}$.

Definition 12.1 (k-deviation)**.** Assuming predefined a positive integer k with $k > \bar{k} := \frac{n}{N+1}$, for any part $\tau \in m_N$ define the random integer *occupancy counts* N_τ as follows:

$$N_\tau := n\, \mathbb{P}_{\mathrm{n}}(\tau)\,.$$

We say that a *k-deviation* occurs in τ if $N_\tau \geq k$

Notation 12.2 (Deviation rate, expected deviation rate)**.** Assuming predefined a positive integer k with $k > \bar{k} := \frac{n}{N+1}$, call *deviation rate* the ratio

$$r^\star_k := \frac{1}{N+1} \sum_{\tau \in m_N} \mathbb{1}_{\{N_\tau \geq k\}}\,.$$

Denote p^+_k the common *expected deviation rate*: for any part $\tau \in m_N$

$$p^+_k := \mathbb{P}\left[N_\tau \geq k\right].$$

As the source density s is assumed uniform, the expected deviation rate is independent of τ.

12.2.2 Selected parts and adverse model

In the following, we define a sample-dependent adverse model in the hope that it will show a strong enough empirical contrast gain over the null model. To do so we select elementary parts with large occupancy counts, and produce the announced adverse model by applying the completion rule in Assumption 9.11.

A slight difference with the method in the previous section (Gaussian signal partition) is that we base the selection of these parts on an occupancy count threshold, as opposed to applying a predefined threshold on the rank by occupancy count order.

Definition 12.3 (Selected parts and adverse model)**.** Assuming predefined a positive integer k not less than $\bar{k} := \frac{n}{N+1}$, we define the random set of *selected parts* $b^\star{}_k$ as the set formed by the parts of m_N where a k-deviation occurs.

$$b^\star{}_k := \{\tau \in m_N, N_\tau \geq k\}\,.$$

The random *adverse model* $m^\star{}_k$ then is chosen as any model in the family $\mathcal{M}$ that

- isolates every selected location.
- and satisfies $|m| \leq 1 + \mathrm{C}_{\mathrm{ext}} |b^\star{}_k|$.

By Assumption 9.11, we know that $m^\star{}_k$ exists whatever k and the sample empirical distribution.

12.2.3 Expected empirical contrast of the adverse model

The following lemma provides a bound on the expected empirical contrast of the adverse model $m^\star{}_k$. Recall that for $x > -1$,

$$h(x) = (1+x)\log(1+x) - x,$$

and that for $0 < x < 1$ and $0 < p < 1$,

$$\begin{aligned} h_p(x) &= x \log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right), \\ &= ph\left(\frac{x-p}{p}\right) + (1-p)h\left(-\frac{x-p}{1-p}\right). \end{aligned}$$

We take the convention $h_0(0) := 0$.

Lemma 12.4. *Under the assumptions of Proposition 9.14, let $m^\star$ be the random adverse model introduced in Definition 12.3 for a predefined integer k. Let $b^\star$ be the corresponding set of selected parts, $r_k^\star = \frac{|b^\star{}_k|}{N+1}$ the corresponding deviation rate and p_k^+ the expected deviation rate. If $k > \bar{k} := \frac{n}{N+1}$, the following bound is well defined and holds almost surely:*

$$\gamma_n\left(\hat{s}_{m^\star{}_k}\right) \leq -h_{r_k^\star}\left(r_k^\star \frac{k}{\bar{k}}\right) \leq -r_k^\star h\left(\frac{k-\bar{k}}{\bar{k}}\right).$$

In addition, the following bound holds:

$$\mathbb{E}\left[\gamma_n\left(\hat{s}_{m^\star{}_k}\right)\right] \leq -h_{p_k^+}\left(p_k^+ \frac{k}{\bar{k}}\right) \leq -p_k^+ h\left(\frac{k-\bar{k}}{\bar{k}}\right).$$

Proof. By the definitions of the set of selected parts $b^\star{}_k$ and of the adverse model $m^\star{}_k$, any elementary part $\tau \in m_N$ with $N_\tau \geq k$ is an element of $m^\star{}_k$ a.s. In other words, any such part is isolated in $m^\star{}_k$ a.s.. The other parts in $m^\star{}_k$ are unions of one or several elementary parts. Recall that any elementary part is of measure $\frac{1}{N+1}$. By the balance relation $\sum_{\tau \in m_N} N_\tau = (N+1)\bar{k} \geq k\,|b^\star{}_k|$, the ratio $r_k^\star = \frac{|b^\star{}_k|}{N+1}$ is always less than $\frac{\bar{k}}{k} < 1$. In other words, since $k > \bar{k}$, there is always a part occupied by less than k sample elements.

By definition of the empirical contrast,

$$\begin{aligned}\gamma_n(\hat{s}_{m^\star{}_k}) &= -\int \hat{s}_{m^\star{}_k} \log(\hat{s}_{m^\star{}_k})\,\mathrm{d}\,\mu,\\ &= -\int [\hat{s}_{m^\star{}_k} \log(\hat{s}_{m^\star{}_k}) - \hat{s}_{m^\star{}_k} + 1]\,\mathrm{d}\,\mu,\\ &= -\int h(\hat{s}_{m^\star{}_k} - 1)\,\mathrm{d}\,\mu,\\ &= -\sum_{\tau\in m^\star{}_k} h\left(\frac{N_\tau}{n\mu(\tau)} - 1\right)\mu(\tau) - \sum_{\tau\in m^\star{}_k\setminus b^\star{}_k} h\left(\frac{N_\tau}{n\mu(\tau)} - 1\right)\mu(\tau).\end{aligned}$$

The measure of the set $\cup\{\tau \in b^\star{}_k\}$ is $r^\star_k = \frac{|b^\star{}_k|}{N+1}$ and the measure of its complement $\cup\{\tau \in b^\star{}_k\}$ is $1 - r^\star_k$. The $h(\cdot)$ function is convex and invoking Jensen's inequality separately on each of these sets yields whenever $|b^\star{}_k| > 0$:

$$\gamma_n(\hat{s}_{m^\star{}_k}) \le -\frac{|b^\star{}_k|}{N+1} h\left(\frac{\sum_{\tau\in b^\star{}_k} N_\tau}{\bar{k}\,|b^\star{}_k|} - 1\right) - \left(1 - \frac{|b^\star{}_k|}{N+1}\right) h\left(\frac{\sum_{\tau\in m^\star{}_k\setminus b^\star{}_k} N_\tau}{\bar{k}(N+1-|b^\star{}_k|)} - 1\right). \tag{12.1}$$

For any elementary segment $\tau \in m_N$, the number N_τ satisfies $N_\tau \ge k$ if $\tau \in b^\star{}_k$, and $N_\tau < k$ otherwise. As a result, the count $\sum_{\tau\in m^\star{}_k\setminus b^\star{}_k} N_\tau$ satisfies:

$$\begin{aligned}\sum_{\tau\in m^\star{}_k\setminus b^\star{}_k} N_\tau &\le n - |b^\star{}_k|\,k,\\ &= (N+1-|b^\star{}_k|)\bar{k} - |b^\star{}_k|\,(k-\bar{k}),\\ \frac{\sum_{\tau\in m^\star{}_k\setminus b^\star{}_k} N_\tau}{(N+1-|b^\star{}_k|)} &\le \bar{k} - \frac{r^\star_k}{1-r^\star_k}(k-\bar{k}).\end{aligned}$$

The $h(\cdot)$ function is convex with a minimum in 0, so it is decreasing on $\mathbb{R}^-$ and increasing on $\mathbb{R}^+$. It follows from Inequality 12.1 combined with the bounds above that whenever $|b^\star{}_k| > 0$, the following relation holds:

$$\begin{aligned}\gamma_n(\hat{s}_{m^\star{}_k}) \le &-r^\star_k h\left(\frac{k-\bar{k}}{\bar{k}}\right)\\ &-(1-r^\star_k)\,h\left(-\frac{r^\star_k}{1-r^\star_k}\frac{k-\bar{k}}{\bar{k}}\right),\end{aligned}$$

that we may express as:

$$\begin{aligned}\gamma_n(\hat{s}_{m^\star{}_k}) &\le -r^\star_k h\left(\frac{\frac{r^\star_k k}{\bar{k}} - r^\star_k}{r^\star_k}\right) - (1-r^\star_k)h\left(-\frac{\frac{r^\star_k k}{\bar{k}} - r^\star_k}{1-r^\star_k}\right),\\ &= -h_{r^\star_k}\left(r^\star_k\frac{k}{\bar{k}}\right).\end{aligned}$$

As the relation above holds also trivially when $|b^\star{}_k| = 0$ with the convention $h_0(0) = 0$, it holds almost surely, which is the first affirmation in the lemma. The second

derivative of the function $x \mapsto h_x(xu) = x\log(u) + (1-xu)\log\left(\frac{1-xu}{1-x}\right)$ is $\frac{(u-1)^2}{(1-x)^2(1-xu)}$, so this function is convex for $xu < 1$ and $0 < x < 1$, which is the case with $x = r_k^\star$ and $u = \frac{k}{\bar{k}}$. Recall that $\mathbb{E}[r_k^\star] = \mathbb{E}\left[\frac{|b^\star{}_k|}{N+1}\right] = p_k^+$. Taking the expectation on both sides of the bound above yields by Jensen's inequality:

$$\mathbb{E}[\gamma_n(\hat{s}_{m^\star{}_k})] \leq -h_{p_k^+}\left(p_k^+\frac{k}{\bar{k}}\right) \leq -p_k^+ h\left(\frac{k-\bar{k}}{\bar{k}}\right),$$

which is well defined, as the expected deviation rate p_k^+ satisfies $p_k^+ \leq \frac{\bar{k}}{k} < 1$. This is the second affirmation in the lemma. □

12.2.4 Risk

The following explicits a straightforward link between the risk of the model selection procedure and the expected empirical contrast of the adverse model. We still denote $r_k^\star$ the ratio $\frac{|b^\star{}_k|}{N+1}$ and use the convention $h_0(0) = 0$

Lemma 12.5. *Under the assumptions of Proposition 9.14, denote* $\mathrm{C_{ext}}$ *the extension factor for the completion rule in Assumption 9.11. Define the adverse model* $m^\star{}_k$ *as in Definition 12.3 with a predefined threshold* $k > \bar{k} = \frac{n}{N+1}$ *and let* $b^\star$ *be the corresponding random set of selected parts,* $r_k^\star = \frac{|b^\star{}_k|}{N+1}$ *the corresponding deviation rate and* p_k^+ *the expected deviation rate.*
Assume additionally there is a number P *so that the penalty function is bounded by:*

$$0 \leq \mathrm{pen}(m) \leq \frac{|m|-1}{n}P,\ \forall m \in \mathcal{M}$$

Then the minimizer $\hat{m}$ *of the penalized empirical contrast*

$$\mathrm{crit}(m) = -\int \hat{s}_m \log \hat{s}_m \,\mathrm{d}\,\mu + \mathrm{pen}(m)$$

satisfies the following relations:

$$\begin{aligned}\int \hat{s}_{\hat{m}} \log \hat{s}_{\hat{m}} \,\mathrm{d}\,\mu &\geq h_{r_k^\star}\left(r_k^\star\frac{k}{\bar{k}}\right) - \frac{|b^\star{}_k|}{n}\mathrm{C_{ext}}P,\\ &\geq \frac{N+1}{n}r_k^\star\left[\bar{k}h\left(\frac{k-\bar{k}}{\bar{k}}\right) - \mathrm{C_{ext}}P\right] \quad a.s.,\\ \mathbb{E}\left[\int \hat{s}_{\hat{m}} \log \hat{s}_{\hat{m}} \,\mathrm{d}\,\mu\right] &\geq \frac{N+1}{n}\,p_k^+\left[\bar{k}h\left(\frac{k-\bar{k}}{\bar{k}}\right) - \mathrm{C_{ext}}P\right],\end{aligned}$$

Proof. By the completion rule, the adverse model $m^\star{}_k$ has a linear dimension less than $1 + \mathrm{C_{ext}}|b^\star{}_k|$, so that by assumption its penalty term $\mathrm{pen}(m^\star{}_k)$ is less than $\mathrm{C_{ext}}\frac{|b^\star{}_k|}{n}P$ always. In addition, Lemma 12.4 states that

$$\gamma_n(\hat{s}_{m^\star{}_k}) \leq -h_{r_k^\star}\left(r_k^\star\frac{k}{\bar{k}}\right) \leq -r_k^\star h\left(\frac{k-\bar{k}}{\bar{k}}\right),$$

and

$$\mathbb{E}[\gamma_n(\hat{s}_{m^\star{}_k})] \leq -h_{p_k^+}\left(p_k^+\frac{k}{\bar{k}}\right) \leq -p_k^+ h\left(\frac{k-\bar{k}}{\bar{k}}\right).$$

It follows that almost surely:

$$\begin{aligned}
\mathrm{crit}_n(\hat{s}_{m^\star{}_k}) &= \gamma_n\left(\hat{s}_{m^\star{}_k}\right) + \mathrm{pen}\left(m^\star{}_k\right), \\
&\leq -h_{r^\star_k}\left(r^\star_k \frac{k}{\bar{k}}\right) + \frac{|b^\star{}_k|}{n}\mathrm{C}_{\mathrm{ext}}P, \\
&\leq -r^\star_k h\left(\frac{k-\bar{k}}{\bar{k}}\right) + \frac{|b^\star{}_k|}{n}\mathrm{C}_{\mathrm{ext}}P, \qquad (12.2) \\
&= -\frac{N+1}{n} r^\star_k \left[\bar{k} h\left(\frac{k-\bar{k}}{\bar{k}}\right) - \mathrm{C}_{\mathrm{ext}}P\right]. \qquad (12.3)
\end{aligned}$$

By definition the selected model satisfies $\mathrm{crit}_n(\hat{s}_{\hat{m}}) = \min_{m\in\mathcal{M}} \mathrm{crit}_n(\hat{s}_m)$ a.s., so that:

$$\begin{aligned}
\int \hat{s}_{\hat{m}} \log \hat{s}_{\hat{m}} \,\mathrm{d}\,\mu &= -\,\mathrm{crit}_n(\hat{s}_{\hat{m}}) + \mathrm{pen}\left(\hat{m}\right), \\
&\geq -\mathrm{crit}_n(\hat{s}_{\hat{m}}), \\
&\geq -\mathrm{crit}_n(\hat{s}_{m^\star{}_k}), \\
&\geq h_{r^\star_k}\left(r^\star_k \frac{k}{\bar{k}}\right) - \frac{|b^\star{}_k|}{n}\mathrm{C}_{\mathrm{ext}}P \text{ by Ineq. 12.2}, \\
&\geq \frac{|b^\star{}_k|}{n}\left[\bar{k} h\left(\frac{k-\bar{k}}{\bar{k}}\right) - \mathrm{C}_{\mathrm{ext}}P\right] \text{ by Ineq. 12.3}.
\end{aligned}$$

which yields the inequalities in the lemma, one almost surely and one in expectation. □

12.2.5 Variability of the number of selected parts

The following arises from the negative association property of 'competing urns' (Mallows, [Mal68], quoted in [Pem04]). Dubashi and Ranjan study this situation in detail in [DR98]. An important aspect is that for the purpose of stochastic bounds on sums, one can treat the occupancy counts as if they were independent.

Lemma 12.6. *Under the assumptions of Proposition 9.14, let $b^\star{}_k$ be the random set of parts introduced in definition 12.3 for a predefined integer $k \leq n$. Denote its expectation $E_k : \mathbb{E}\left[|b^\star{}_k|\right]$. Then the following bound holds:*

$$\mathbb{V}\left[|b^\star{}_k|\right] \leq E_k\left(1 - \frac{E_k}{N+1}\right).$$

Morevover for any t with $0 \leq t < E_k$, the following bound holds:

$$\mathbb{P}\left[|b^\star{}_k| - E_k \leq -t\right] \leq \exp\left[-E_k h\left(-\frac{t}{E_k}\right)\right].$$

Proof. The sequence of occupancy counts $(N_\tau)_{\tau\in m_N}$ follows a multinomial law of parameter $(n, p, \cdots, p)$ with $p = \frac{1}{N+1}$. This is the situation of competing multinomial bins studied in detail by Dubashi and Ranjan. In particular Theorem 13 and the proof of Proposition 5 in [DR98] state that the components of both the vectors $(N_\tau)_{\tau\in m_N}$ and $(n - N_\tau)_{\tau\in m_N}$ are negatively associated in the sense that for any partition of m_N in two

set of parts I and J and any functions $f_1 : \mathbb{R}^{|I|} \to \mathbb{R}$ and $f_2 : \mathbb{R}^{|J|} \to \mathbb{R}$ that are both non-increasing or non-decreasing, the following holds:

$$\mathbb{E}\left[f_1\left((N_i)_{i\in I}\right) f_2\left((N_j)_{j\in J}\right)\right] \leq \mathbb{E}\left[f_1\left((N_i)_{i\in I}\right)\right]\mathbb{E}\left[f_2\left((N_j)_{j\in J}\right)\right]. \tag{12.4}$$

As a direct consequence, for any non-increasing or non-decreasing function $f : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}\left[\prod_{\tau\in m_N} f(N_\tau)\right] \leq \prod_{\tau\in m_N} \mathbb{E}\left[f(N_\tau)\right]. \tag{12.5}$$

It follows the "negative right ortant dependence" [DR98, proposition 4]: for each pair of disjoint parts (τ_1, τ_2), the corresponding occupancy counts N_1 and N_2 satisfy for any n_1 and n_2 with $0 \leq n_1 \leq n$ and $0 \leq n_2 \leq n$ the marginal probability bound:

$$\mathbb{P}\left[N_2 \geq n_2 \text{ and } N_1 \geq n_1\right] \leq \mathbb{P}\left[N_1 \geq n_1\right]\mathbb{P}\left[N_2 \geq n_2\right],$$

and the relation:

$$\mathbb{P}\left[N_2 \geq n_2 \,|\, N_1 \geq n_1\right] \leq \mathbb{P}\left[N_2 \geq n_2\right].$$

As the number $|b^\star{}_k|$ is the count of k-deviations over the $N+1$ disjoint and identical parts in m_N, it satisfies:

$$\mathbb{E}\left[|b^\star{}_k|^2\right] \leq N(N+1)\,\mathbb{P}\left[N_1 \geq k\right]^2 + (N+1)\,\mathbb{P}\left[N_1 \geq k\right],$$

$$= \left(1 - \frac{1}{N+1}\right)\mathbb{E}\left[|b^\star{}_k|\right]^2 + \mathbb{E}\left[|b^\star{}_k|\right],$$

so that as announced

$$\mathbb{V}\left[|b^\star{}_k|\right] \leq \left[1 - \frac{\mathbb{E}\left[|b^\star{}_k|\right]}{N+1}\right]\mathbb{E}\left[|b^\star{}_k|\right].$$

Denote E_k the quantity $\mathbb{E}\left[|b^\star{}_k|\right]$. Following the observation in [DR98] that Chernoff-Hoeffding's bounds apply in such a situation, note that for any $\lambda > 0$ the functions $N_\tau \mapsto \exp\left(-\lambda\,\mathbb{1}_{N_\tau \geq k}\right)$ are non-increasing so that Inequality 12.5 implies that for any $\lambda > 0$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left[\lambda\left(E_k - |b^\star{}_k|\right)\right]\right] &= e^{\lambda E_k}\,\mathbb{E}\left[\exp\left[-\lambda\,|b^\star{}_k|\right]\right], \\
&= e^{\lambda E_k}\,\mathbb{E}\left[\exp\left[-\lambda \sum_{\tau\in m_N} \mathbb{1}_{N_\tau \geq k}\right]\right], \\
&= e^{\lambda E_k}\,\mathbb{E}\left[\prod_{\tau\in m_N} \exp\left[-\lambda\,\mathbb{1}_{N_\tau \geq k}\right]\right], \\
&= e^{\lambda E_k}\,\mathbb{E}\left[\prod_{\tau\in m_N} \left[1 - \left(1 - e^{-\lambda}\right)\mathbb{1}_{N_\tau \geq k}\right]\right], \\
&\leq e^{\lambda E_k}\,\mathbb{E}\left[\left[1 - \left(1 - e^{-\lambda}\right)\mathbb{1}_{N_\tau \geq k}\right]\right]^{N+1}, \\
&= e^{\lambda E_k}\left[1 - \left(1 - e^{-\lambda}\right)\frac{E_k}{N+1}\right]^{N+1}, \\
&\leq \exp\left[\left(e^{-\lambda} + \lambda - 1\right)E_k\right].
\end{aligned}$$

It follows by Markov's inequality that for $0 < t < E_k$,

$$\mathbb{P}\left[|b^\star{}_k| \leq E_k - t\right] \leq \exp\left[\left(e^{-\lambda} + \lambda - 1\right) E_k - \lambda t\right],$$

and after optimizing with $\lambda = -\log\left(1 - \frac{t}{E_k}\right)$,

$$\begin{aligned}\mathbb{P}\left[|b^\star{}_k| \leq E_k - t\right] &\leq \exp\left[-(E_k - t)\log\left(1 - \frac{t}{E_k}\right) - t\right], \\ &= \exp\left[-E_k h\left(-\frac{t}{E_k}\right)\right].\end{aligned}$$

□

12.2.6 Concluding the proof of Proposition 9.14

Proof. As in the proposition in proof, choose $p = \frac{1}{N+1}$ and assume known a number k and a number L so that:

$$np + \sqrt{2np} + \frac{2}{3} \leq k \leq n - \sqrt{np(1-p)} - 1. \tag{12.6}$$

and

$$L = nh_p\left(\frac{k + \sqrt{np(1-p)} + \frac{1-p}{2}}{n}\right), \tag{12.7}$$

Lemma 14.2 states that the Poisson tail function satisfies $u \leq \sqrt{2h(u)} + \frac{2}{3}h(u)$ for any $u > 0$. Recall that $\bar{k} = \frac{n}{N+1} = np$ and that the function $u \mapsto h(u)$ is increasing for $u > 0$. As $k \geq np + \sqrt{2np} + \frac{2}{3}$, then:

$$\begin{aligned}\frac{k - \bar{k}}{\bar{k}} &\geq \sqrt{\frac{2}{\bar{k}}} + \frac{2}{3\bar{k}}, \\ &\geq h^{-1}\left(\frac{1}{\bar{k}}\right),\end{aligned}$$

so that the first statement in the proposition holds:

$$\bar{k}h\left(\frac{k - \bar{k}}{\bar{k}}\right) \geq 1.$$

The tail lower bound in Corollary 14.13 states that if $B_{n,p}$ is a binomial random variable with parameters n and $p > 0$ and k an integer with $np \leq k \leq n - \sqrt{np(1-p)} - 1$, then the following bound holds:

$$\mathbb{P}\left[B_{n,p} \geq k\right] \geq \frac{1}{3}\exp\left[-nh_p\left(\frac{k + \sqrt{np(1-p)} + \frac{1-p}{2}}{n}\right)\right].$$

Then by the definition of the number L in Equation 12.7 the following holds:

$$\begin{aligned} p_k^+ &= \mathbb{P}\left[B_{n,p} \geq k\right], \\ &\geq \frac{1}{3}\exp\left[-nh_p\left(\frac{k+\sqrt{np(1-p)}+\frac{1-p}{2}}{n}\right)\right], \\ &= \frac{1}{3}\exp\left[-L\right] \end{aligned}$$

so that the third statement in the Proposition holds:

$$\mathbb{E}\left[|b^\star{}_k|\right] = (N+1)p_k^+ \geq \frac{N+1}{3}e^{-L}. \tag{12.8}$$

Finally Lemma 12.5 states that if there is a number P such that for any model $m \in \mathcal{M}$,

$$0 \leq \mathrm{pen}\left(m\right) \leq \frac{|m|-1}{n}P,$$

then the following bounds hold:

$$\int \hat{s}_{\hat{m}} \log \hat{s}_{\hat{m}} \,\mathrm{d}\,\mu \geq \frac{N+1}{n} r_k^\star \left[\bar{k}h\left(\frac{k-\bar{k}}{\bar{k}}\right) - \mathrm{C_{ext}}P\right] \text{ a.s.}, \tag{12.9}$$

and

$$\mathbb{E}\left[\int \hat{s}_{\hat{m}} \log \hat{s}_{\hat{m}} \,\mathrm{d}\,\mu\right] \geq \frac{N+1}{n} p_k^+ \left[\bar{k}h\left(\frac{k-\bar{k}}{\bar{k}}\right) - \mathrm{C_{ext}}P\right]. \tag{12.10}$$

On the other hand the proposition makes the assumption that the penalty function is bounded by

$$\mathrm{pen}\left(m\right) \leq \frac{|m|-1}{n\mathrm{C_{ext}}}\left[\bar{k}h\left(\frac{k-\bar{k}}{k}\right) - 1\right] \;\forall m \in \mathcal{M}.$$

Taking $P = \frac{1}{\mathrm{C_{ext}}}\left[\bar{k}h\left(\frac{k-\bar{k}}{k}\right) - 1\right]$ in Inequalities 12.9 and 12.10 yields:

$$\begin{aligned} \int \hat{s}_{\hat{m}} \log \hat{s}_{\hat{m}} \,\mathrm{d}\,\mu &\geq \frac{N+1}{n} r_k^\star \times 1 \text{ a.s.}, \\ &= \frac{|b^\star|}{n} \text{ a.s.}, \end{aligned}$$

and by the lower bound in Inequality 12.8 above on $p_k^+ = \mathbb{E}\left[\frac{|b^\star|}{N+1}\right]$:

$$\mathbb{E}\left[\int \hat{s}_{\hat{m}} \log \hat{s}_{\hat{m}} \,\mathrm{d}\,\mu\right] \geq \frac{N+1}{n}\frac{1}{3}e^{-L}.$$

Assessing the minimal penalty level Lemma 14.6 states that the binomial tail function admits the following bound for $0 < p < p+u < 1$:

$$u \geq \frac{\sqrt{2p(1-p)h_p(p+u)}}{1+\sqrt{\frac{p}{2(1-p)}h_p(p+u)}}.$$

As $k \geq np$ by Equation 12.6, combining with the definition of the number L in Equation 12.7 ensures that:

$$\frac{k + \sqrt{np(1-p)} + \frac{1-p}{2}}{n} - p \geq \frac{\sqrt{2p(1-p)\frac{L}{n}}}{1 + \sqrt{\frac{p}{2(1-p)}\frac{L}{n}}},$$

and after rearranging the terms:

$$\frac{k - np}{\sqrt{np}} \geq \sqrt{1-p}\left[\frac{\sqrt{2L}}{1 + \sqrt{\frac{p}{2(1-p)}\frac{L}{n}}} - 1 - \frac{1}{2}\sqrt{\frac{1-p}{np}}\right]. \tag{12.11}$$

By the relation $\frac{1}{1+x} \geq 1 - x$ for $0 < x$ equation 12.11 above implies:

$$\frac{k - np}{\sqrt{np}} \geq \sqrt{(1-p)}\left[\sqrt{2L} - L\sqrt{\frac{p}{(1-p)n}} - 1 - \frac{1}{2}\sqrt{\frac{1-p}{np}}\right]. \tag{12.12}$$

To facilitate the calculations, assume additionally as in Inequalities 9.9 and 9.10 in the proposition that $N \geq 3$ (so that $\log(N) > 1$) and that the following bound holds:

$$k \leq np + \sqrt{np(1-p)}\left(\sqrt{\log(N)} - 1\right) - \frac{1-p}{2}.$$

It's purpose is to provide the bound

$$L \leq \log(N),$$

by the following arguments: Lemma 14.6 asserts that $h_p(p+u) \leq \frac{u^2}{p(1-p)}$ for $0 < p < p + u < 1$. Then by the definition of the number L:

$$\begin{aligned}
L &:= nh_p\left(\frac{k - np + \sqrt{np(1-p)} + \frac{1-p}{2}}{n}\right), \\
&\leq nh\left(p + \frac{\sqrt{\log(N)}\sqrt{np(1-p)} - \frac{1-p}{2}}{n}\right), \\
&\leq nh_p\left(p + \frac{\sqrt{\log(N)}\sqrt{np(1-p)}}{n}\right), \text{since by assumption also } k \geq np. \\
&\leq \log(N), \text{ by the bound above on the } h_p(\cdot) \text{ function.}
\end{aligned}$$

So under these additional assumptions the number L is less than $\log(N)$. By the assumptions that $\frac{n}{N} \geq \log(n)^2$ and that $n \geq 9$ and by the relation $\frac{\log(N)}{N} \leq e^{-1}$, the following bound follows:

$$L\sqrt{\frac{p}{n(1-p)}} = L\sqrt{\frac{1}{nN}} \leq \frac{\log(N)}{N}\sqrt{\frac{N}{n}} \leq \frac{1}{e}\sqrt{\frac{N}{n}} \leq \frac{1}{e\,\log(n)},$$

and

$$\frac{1}{2}\sqrt{\frac{1-p}{np}} = \frac{1}{2}\sqrt{\frac{N}{n}} \leq \frac{1}{2}\frac{1}{\log(n)}$$

Plugging in Inequality 12.12 yields, since $\frac{1}{e} + \frac{1}{2} \leq 1$,

$$\frac{k - np}{\sqrt{np}} \geq \sqrt{1-p}\left[\sqrt{2L} - 1 - \log(n)^{-1}\right]. \tag{12.13}$$

We know from Lemma 14.2 that for $u > 0$, the Poisson tail function satisfies $h(u) \geq \frac{u^2}{2\left(1+\frac{u}{3}\right)}$. Combining with Inequality 12.13 yields:

$$np\, h\left(\frac{k-np}{np}\right) \geq \frac{1-p}{2} \frac{\left[\left(\sqrt{2L} - 1 - \log(n)^{-1}\right)^+\right]^2}{1 + \frac{1}{3}\sqrt{\frac{2(1-p)L}{np}}},$$

Again by the assumptions that $L \leq \log N$ and $\frac{n}{N} \geq \log(n)^2$, the following holds:

$$\frac{1}{3}\sqrt{\frac{2(1-p)L}{np}} = \sqrt{\frac{2NL}{9n}} \leq \frac{1}{2}\sqrt{\frac{\log(N)}{\log(n)^2}} \leq \frac{1}{2}\log(n)^{-\frac{1}{2}},$$

so that

$$nph\left(\frac{k-np}{np}\right) \geq (1-p)\frac{\left[\left(\sqrt{L} - \frac{1}{\sqrt{2}} - \log(n)^{-1}\right)^+\right]^2}{1 + \frac{1}{2}\log(n)^{-\frac{1}{2}}},$$

which is he last statement in Proposition 9.14, after replacing np by its shorthand $\bar{k}$ and $1-p$ by its value $\frac{N}{N+1}$:

$$\bar{k}h\left(\frac{k-\bar{k}}{\bar{k}}\right) \geq \frac{N}{N+1}\frac{\left[\left(\sqrt{L} - \frac{1}{\sqrt{2}} - \log(n)^{-1}\right)^+\right]^2}{1 + \frac{1}{2}\log(n)^{-\frac{1}{2}}}, \tag{12.14}$$

Variability of the number of selected parts The last two assertions in Proposition 9.14 result directly from Lemma 12.6 with the same value of k. □

13 Proofs for section: 10 Remark on the isotropy of the model family

13.1 Proof of Proposition 10.1

Proof. We treat the minimal penalty threshold by a deviation argument, and then the sufficient penalty threshold by an application of Corollary 5.7, which derives from the general model selection Theorem 5.6.

Minimal penalty Denote δ the fraction $\delta = \frac{d-2}{n-4}$. We know from Corollary 14.19 that for any t with t, provided that $4 \leq d \leq \frac{n}{2}$ and $0 < t \leq \frac{N}{10}$, there is a number $x_t^\star$ with

$$x_t^\star \geq \delta + \sqrt{\frac{2\delta(1-\delta)}{n-2}} \left[\frac{\sqrt{2\log\left(\frac{N}{3t}\right)}}{1 + \sqrt{2\log\left(\frac{N}{3t}\right)} \frac{2\delta}{(n-2)(1-\delta)}} - 1 \right]. \tag{13.1}$$

so that

$$\mathbb{P}\left[\left\| \tilde{\Pi}(W) \right\|^2 \geq d_t^\star \right] \geq \left(1 - e^{-t}\right)^2, \tag{13.2}$$

where

$$d_t^\star := n \left(1 - \sqrt{\frac{4t}{n}} \right) x_t^\star. \tag{13.3}$$

Assume that

$$\mathrm{pen}_d \leq d_t^\star. \tag{13.4}$$

Then on an event of probability not less than $(1-e^{-t})^2$, the relation 13.2 translates into

$$\begin{aligned} -\left\| \tilde{\Pi}(W) \right\|^2 + \mathrm{pen}\left(\tilde{\Pi} E_n \right) &= -\left\| \tilde{\Pi}(W) \right\|^2 + d_t^\star, \\ &< 0, \\ &= -\|0 \times W\|^2 + \mathrm{pen}\left(\{0\}\right), \end{aligned}$$

(the case of equality is negligible in a Gaussian context). On this event the non null model $\tilde{\Pi}$ shows a lower penalized contrast than the null one, and the selected the model $\hat{m}$ is $\tilde{\Pi} E_n$. As a consequence, under the condition 13.4, Inequality 10.3 announced in the proposition holds:

$$\mathbb{P}_{\mathcal{M},W}\left[|\hat{m}| = d \right] \geq \left(1 - e^{-t}\right)^2.$$

Moreover, still on the same event, the squared norm $\left\| \tilde{\Pi}(W) \right\|^2$ of the selected estimator is larger than $d_t^\star$, so that Inequality 10.4 announced in the proposition follows:

$$\mathbb{E}_{\mathcal{M},W}\left[\|\hat{s}_{\hat{m}}\|^2 \right] \geq \left(1 - e^{-t}\right)^2 d_t^\star.$$

Consider now $d_t^\star$ as a function $(n, d, N, t) \mapsto d_t^\star(n, d, N, t)$ defined by 13.1 and 13.3 (for simplicity we use the same symbol for the parameter and the function). Then when n, d and N tend to ∞ while $d \leq \frac{n}{2}$, for any fixed value of t, the following equivalents hold:

$$\begin{aligned}
\log \frac{N}{3t} &= [1 + o(1)] \log N, \\
\delta &= [1 + o(1)] \frac{d}{n}, \\
\sqrt{\frac{2\delta(1-\delta)}{n-2}} &= [1 + o(1)] \sqrt{\frac{2d(n-d)}{n^2}} = o\left(\frac{d}{n}\right) = O(1), \\
1 - \sqrt{\frac{4t}{n}} &= 1 + o(1).
\end{aligned}$$

Then Inequality 10.5 announced in the proposition follows from the expression of $d_t^\star(n, d, N, t)$ in 13.1 and 13.3 by:

$$\begin{aligned}
d_t^\star(n, d, N, t) &= n \left(1 - \sqrt{\frac{4t}{n}}\right) \left[\delta + \sqrt{\frac{2\delta(1-\delta)}{n-2}} \left[\frac{\sqrt{2\log\left(\frac{N}{3t}\right)}}{1 + \sqrt{2\log\left(\frac{N}{3t}\right)} \frac{2\delta}{(n-2)(1-\delta)}} - 1\right]\right], \\
&= n\,[1 + o(1)] \left[[1 + o(1)] \frac{d}{n} + \frac{[1 + o(1)]\, 2\sqrt{\frac{d(n-d)}{n^3} \log(N)}}{[1 + o(1)] \left[1 + 2\sqrt{\frac{d}{n(n-d)} \log(N)}\right]} - o\left(\frac{d}{n}\right)\right], \\
&= [1 + o(1)] \left[d + \frac{2\sqrt{d\left(1 - \frac{d}{n}\right)\log(N)}}{1 + \frac{2}{n-d}\sqrt{d\left(1 - \frac{d}{n}\right)\log(N)}}\right] \text{ as the terms are positive.}
\end{aligned}$$

Sufficient penalty The set of weights $\left\{L_m = \frac{\log(N)}{d}\right\}_{m \in \mathcal{M}}$ satisfies

$$\begin{aligned}
\sum_{m \in \mathcal{M}, |m| > 0} e^{-|m| L_m} &= N e^{-d \frac{\log(N)}{d}}, \\
&= 1,
\end{aligned}$$

so that by Corollary 5.7, penalties of the form proposed in Equality 10.1 of the proposition:

$$\begin{aligned}
\text{pen}_0 &= 0, \\
\text{pen}_d &= \frac{1+\eta}{1-\eta} \epsilon^2 d \left(1 + 2\sqrt{\frac{\log(N)}{d}} + 2\frac{\log(N)}{d}\right), \\
&= \frac{1+\eta}{1-\eta} \epsilon^2 \left(d + 2\sqrt{d \log(N)} + 2\log(N)\right).
\end{aligned}$$

warrant that given any realization of the random model family $\mathcal{M}$,

$$\eta\, \mathbb{E}_W \left[\|s - \tilde{s}\|^2\right] \leq \inf_{m \in \mathcal{M}} \left\{ \mathrm{d}^2(s, S_m) + \text{pen}(m) - \epsilon^2 |m| \right\} + \epsilon^2 \frac{(1+3\eta)}{\eta(1-\eta)},$$

and as here $s = 0$ and $\epsilon^2 = 1$, taking the value in $m = \{0\}$ yields:

$$\mathbb{E}_W \left[\|s - \tilde{s}\|^2 \right] \leq \frac{(1+3\eta)}{\eta^2(1-\eta)}.$$

Inequality 10.2 in the proposition follows:

$$\mathbb{E}_{\mathcal{M},W} \left[\|s - \tilde{s}\|^2 \right] \leq \frac{(1+3\eta)}{\eta^2(1-\eta)}.$$

□

Part III

Deviation Inequalities

Summary

14 Tail lower bound inequalities **174**
14.1 Tail functions . . . 174
14.1.1 Poisson tail function . . . 174
14.1.2 Gamma tail function . . . 175
14.1.3 Binomial tail function . . . 175
14.2 Some known concentration inequalities . . . 176
14.3 Tail lower bound for a binomial distribution . . . 178
14.4 Inequalities based on the properties of the incomplete beta function . 178
14.4.1 Tail lower bound for a beta distribution . . . 178
14.4.2 Projection of a Gaussian random vector on randomly oriented d-dimensional subspaces . . . 180
14.5 Inequalities based on the properties of the incomplete gamma function 181

15 Tools for sorted chi-square samples **183**
15.1 Order statistics . . . 183
15.2 Lower bound on the partial sum of an ordered chi-square sample . . 183
15.3 Upper bound on the partial sum of an ordered chi-square sample . . 185

16 Proofs for Section 14 Tail lower bound inequalities **186**
16.1 Proof of Lemma 14.2 . . . 186
16.2 Proof of Lemma 14.4 . . . 186
16.3 Proof of Lemma 14.6 . . . 187
16.4 Proof of Lemma 14.10 . . . 188
16.5 Proof of Lemma 14.11 . . . 189
16.6 Proof of Lemma 14.13 . . . 190
16.7 Proof of Lemma 14.14 . . . 192
16.8 Proof of Corollary 14.15 . . . 194
16.9 Proof of Lemma 14.18 . . . 195
16.10 Proof of Corollary 14.19 . . . 196
16.11 Proof of Lemma 14.20 . . . 199
16.12 Proof of Lemma 14.21 . . . 200

17 Proofs for Section 15 Tools for sorted chi-square samples **201**
17.1 Proof of Lemma 15.3 . . . 201
17.2 Proof of Lemma 15.4 . . . 203

14 Tail lower bound inequalities

This section is dedicated to several instances of tail lower bound inequalities for probability laws encountered in the context of model selection. These inequalities are intended to help proving the occurrence of the minimal penalty phenomenon in various situations. Penalized contrast methods rely heavily on the concentration measure phenomenon, and our general method will be to look for the tightness of measure concentration inequalities by means of lower bounds on the corresponding probability tails.

For reading convenience most of the proofs are given in Section 16.

14.1 Tail functions

The following functions appear frequently in deviation inequalities. Some useful general definitions and remarks are listed bellow without further introduction.

14.1.1 Poisson tail function

Definition 14.1 (Poisson tail function h)**.** Denote h the function defined on $[-1, \infty]$ by:

$$h(u) = (1+u)\log(1+u) - u \text{ for } u > -1$$

and

$$h(-1) = 1$$

Lemma 14.2. *The following inequalities hold:*

for $u \geq 0$

$$\frac{u^2}{2\left(1+\frac{u}{3}\right)} \leq h(u) \leq \frac{u^2}{2}, \tag{14.1}$$

for $0 \leq u \leq 1$

$$\frac{u^2}{2\left(1-\frac{u}{3}\right)} \leq h(-u) \leq \frac{u^2}{2\left(1-\frac{u}{2}\right)}. \tag{14.2}$$

For any $u > 0$*, the following inequalities hold:*

$$\sqrt{2h(u)} \leq u \leq \sqrt{2h(u)} + \frac{2}{3}h(u). \tag{14.3}$$

Proof is given in Section 16.1.

14.1.2 Gamma tail function

Definition 14.3 (Gamma tail function g)**.** Denote g the function defined on $(-1, \infty)$ by:

$$g(u) = u - \log(1+u) \text{ for } u > -1.$$

Lemma 14.4. *The following inequalities hold:*

$$\left(\sqrt{1+u} - 1\right)^2 \leq g(u) \text{ for } u > -1, \tag{14.4}$$

$$\frac{u^2}{2\left(1 + \frac{2u}{3}\right)} \leq g(u) \leq \frac{u^2}{2\left(1 + \frac{u}{2}\right)} \text{ for } u > 0, \tag{14.5}$$

and

$$\frac{u^2}{2\left(1 - \frac{2u}{3}\right)} \leq g(-u) \leq \frac{u^2}{2(1-u)} \text{ for } u < 1. \tag{14.6}$$

The proof is given in Section 16.2.

14.1.3 Binomial tail function

Definition 14.5 (Binomial tail function h_p)**.** For any real p with in $0 < p < 1$, denote h_p the function defined (by continuity) on $[0, 1]$ by:

$$h_p(v) = v \log\left(\frac{v}{p}\right) + (1-v)\log\left(\frac{1-v}{1-p}\right)$$

Lemma 14.6. *The following inequalities hold:*
for u with $p < p + u < 1$

$$h_p(p+u) \geq \frac{u^2}{2\left(p + \frac{u}{3}\right)\left(1 - p - \frac{u}{3}\right)}, \tag{14.7}$$

$$h_p(p+u) \leq \frac{u^2}{2p\left(1 - p - \frac{u}{2}\right)}, \tag{14.8}$$

$$h_p(p+u) \leq \frac{u^2}{p(1-p)}, \tag{14.9}$$

and

$$u \geq \frac{\sqrt{2p(1-p)h_p(p+u)}}{1 + \sqrt{\frac{p}{2(1-p)}h_p(p+u)}}. \tag{14.10}$$

for u with $0 < p - u < p$:

$$h_p(p-u) \leq \frac{u^2}{2\left(p - \frac{u}{2}\right)(1-p)}, \tag{14.11}$$

$$u \geq \frac{\sqrt{2p(1-p)h_p(p-u)}}{1+\sqrt{\frac{1-p}{2p}h_p(p-u)}}, \tag{14.12}$$

and for $0 < x \leq p$*:*

$$p \geq x + \frac{\sqrt{2x(1-x)h_p(x)}}{1+\sqrt{\frac{2xh_p(x)}{1-x}}}, \tag{14.13}$$

$$= x + (1-x)\frac{\sqrt{2xh_p(x)}}{\sqrt{1-x}+\sqrt{2xh_p(x)}}. \tag{14.14}$$

The proof is given in Section 16.3.

14.2 Some known concentration inequalities

The following are instances of known concentration inequalities. The functions $h(\cdot)$, $h_p(\cdot)$ and $g(\cdot)$ are introduced in Definitions 14.1,14.5 and 14.3 in the preceding section. From [Mas07, p. 20]:

Lemma 14.7 (Concentration of a binomial random variable)**.** *If* $B_{n,p}$ *is a binomial random variable of parameters* (n,p)*, for* $k \geq np$*:*

$$\mathbb{P}\left[B_{n,p} \geq k\right] \leq e^{-np\,h\left(\frac{k-np}{np}\right)-n(1-p)h\left(-\frac{k-np}{n(1-p)}\right)}, \tag{14.15}$$

$$= e^{-n\,h_p\left(\frac{k}{n}\right)}. \tag{14.16}$$

$$\mathbb{P}\left[B_{n,p} \leq k\right] \leq e^{-np\,h\left(-\frac{k-np}{np}\right)-n(1-p)h\left(\frac{k-np}{n(1-p)}\right)}, \tag{14.17}$$

$$= e^{-n\,h_{1-p}\left(\frac{n-k}{n}\right)}. \tag{14.18}$$

From [Mas07, p. 19]:

Lemma 14.8 (Concentration of a Poisson random variable)**.** *If* P_λ *is a Poisson random variable of parameter* λ *, for* $t \geq \lambda$*:*

$$\mathbb{P}\left[P_\lambda \geq t\right] \leq e^{-\lambda\,h\left(\frac{t-\lambda}{\lambda}\right)}. \tag{14.19}$$

The following results from the link between the binomial distribution and the order statistics of an uniform n-sample. From [Mas90, equ. 81]:

Lemma 14.9 (Concentration of order statistics)**.** *If* $U_{(1)} \leq U_{(1)} \leq \cdot \leq U_{(n)}$ *is an ordered sample of size* n *of the uniform distribution,* j *a positive integer and* θ *a positive number with* $j(1+\theta) \leq n$*, then:*

$$\mathbb{P}\left[U_{(n+1-j)} \geq 1 - \frac{j(1+\theta)}{n}\right] \leq \exp\left[-\frac{j\theta^2}{2(1+\theta)}\right]. \tag{14.20}$$

Lemma 14.10 (Concentration of a non central χ_2 variable)**.** *Let E_n denote an n-dimensional Euclidean space.*

Let s be an element of E_n, let W be a standard Gaussian random element of E_n: $W \sim \mathcal{N}(0, I_n)$, and σ a real number.

For any $t > 0$, the following inequalities holds:

$$\mathbb{P}\left[\|s + \sigma W\|^2 \leq \|s\|^2 + n\sigma^2 - \sqrt{2t\left(4\|s\|^2\sigma^2 + 2n\sigma^4\right)}\right] \leq e^{-t}, \tag{14.21}$$

and

$$\mathbb{P}\left[\|s + \sigma W\|^2 \geq \|s\|^2 + n\sigma^2 + \sqrt{2t\left(4\|s\|^2\sigma^2 + 2n\sigma^4\right)} + 2\sigma^2 t\right] \leq e^{-t}. \tag{14.22}$$

The proof is given in Section 16.4

Lemma 14.11 (Concentration of a gamma random variable)**.** *If G_n is a gamma random variable of parameter n so that for any $g \geq 0$*

$$\mathbb{P}\left[G_n < g\right] = \frac{\int_0^g e^{-x}x^{n-1}\mathrm{dx}}{\int_0^\infty e^{-x}x^{n-1}\mathrm{dx}},$$

then the following relations hold, for $t > 0$:

$$\mathbb{P}\left[G_n \geq n\left(1 + \sqrt{\frac{t}{n}}\right)^2\right] \leq e^{-t},$$

and for $0 < t < n$:

$$\mathbb{P}\left[G_n \leq n\left(1 - \sqrt{\frac{t}{n}}\right)^2\right] \leq e^{-t},$$
$$\mathbb{P}\left[G_n \leq n - 2\sqrt{nt}\right] \leq e^{-t}.$$

The proof is given in Section 16.5

The following [DG03] leads to a lemma of W.B. Johnson and J. Lindenstrauss well known in compressed sensing:

Lemma 14.12 (Norm concentration of a projected random unit vector, beta random variable)**.** *In the n-dimensional Euclidean space, consider u_{n-1} a random unit vector uniformly distributed on the $(n-1)$-dimensional sphere. Denote Π_d the orthogonal projection onto the first d coordinates .*

Then for any ϵ with $0 < \epsilon < 1$,

$$\mathbb{P}\left[\|\Pi_d u_{n-1}\|^2 \leq \frac{d}{n}(1-\epsilon)\right] \leq \exp\left[\frac{d}{2}\left(\epsilon + \log(1-\epsilon)\right)\right],$$
$$\leq \exp\left[-\frac{d}{4}\epsilon^2\right],$$

For any $\epsilon > 0$,

$$\mathbb{P}\left[\|\Pi_d u_{n-1}\|^2 \geq \frac{d}{n}(1+\epsilon)\right] \leq \exp\left[\frac{d}{2}\left(-\epsilon + \log(1+\epsilon)\right)\right],$$
$$\leq \exp\left[-\frac{d\epsilon^2}{4\left(1+\frac{2}{3}\epsilon\right)}\right].^{1}$$

14.3 Tail lower bound for a binomial distribution

The following lemma offers a lower bound for the tail of a binomial distribution:

Lemma 14.13 (Tail lower bound for a binomial distribution)**.** *Let* $B_{n,p}$ *be a binomial random variable with parameters* n *and* $p > 0$.

Let k *be an integer with* $np \leq k \leq n - \sqrt{np(1-p)} - \frac{1-p}{2}$. *Then the following inequalities hold:*

$$\mathbb{P}\left[B_{n,p} \geq k\right] \geq \frac{1}{3}\exp\left[-\left(n+\frac{1}{2}\right)h_p\left(\frac{k+\sqrt{np(1-p)}+\frac{1}{2}}{n+\frac{1}{2}}\right)\right],$$

and

$$\mathbb{P}\left[B_{n,p} \geq k\right] \geq \frac{1}{3}\exp\left[-nh_p\left(\frac{k+\sqrt{np(1-p)}+\frac{1-p}{2}}{n}\right)\right].$$

The proof is given in Section 16.6

14.4 Inequalities based on the properties of the incomplete beta function

14.4.1 Tail lower bound for a beta distribution

The following lemma provides a non asymptotic probability lower bound for large deviations of a beta random variable:

Lemma 14.14 (Tail lower bound for a beta distribution)**.** *Let* $Z_{\alpha,\beta}$ *be a random variable on* $[0,1]$ *of Lebesgue probability density equal to the regularized beta function of parameters* α *and* β, *namely for* $0 \leq x \leq 1$:

$$\mathbb{P}\left[Z_{\alpha,\beta} \leq x\right] := I_x(\alpha,\beta),$$
$$:= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\int_0^x t^{\alpha-1}\,(1-t)^{\beta-1}\,\mathrm{d}t.$$

[1] by Lemma 14.4

Assume that $\alpha \geq 2$ *and* $\beta \geq 2$ *and set*

$$\begin{aligned} \nu &:= \alpha + \beta, \\ t_0 &:= \frac{\alpha - 1}{\alpha + \beta - 2}, \\ \text{and } \sigma &:= \frac{1}{\alpha + \beta - 2}\sqrt{\frac{(\alpha-1)(\beta-1)}{\alpha+\beta-1}} = \sqrt{\frac{t_0(1-t_0)}{\nu-1}}. \end{aligned}$$

Then for any t *with* $\sigma \leq t < 1$,

$$\mathbb{P}\left[Z_{\alpha,\beta} \leq t\right] \geq \frac{1}{3}\exp\left[-(\nu-2)h_{t-\sigma}(t_0)\right], \tag{14.23}$$

and for any t *with* $0 < t \leq 1 - \sigma$,

$$\mathbb{P}\left[Z_{\alpha,\beta} \geq t\right] \geq \frac{1}{3}\exp\left[-(\nu-2)h_{t+\sigma}(t_0)\right]. \tag{14.24}$$

For any η *with*

$$\eta \geq \frac{1}{2}\frac{1}{1-\sqrt{\frac{t_0}{(1-t_0)(\nu-1)}}},$$

the following inequality holds:

$$\mathbb{P}\left[Z_{\alpha,\beta} \geq t_0 + \sigma\left[\frac{\sqrt{2\eta}}{1+\frac{\sigma}{1-t_0}\sqrt{2\eta}} - 1\right]\right] \geq \frac{1}{3}e^{-\eta}. \tag{14.25}$$

Proof is given in Section 16.7.

Corollary 14.15 (Projection of a random unit vector)**.** *Consider, in the* n*-dimensional Euclidean space, a uniformly distributed random unit vector* $\mathbf{u_{n-1}}$ *and a fixed orthogonal projector* Π_d *of rank* d*. Assume that* $4 \leq d \leq n-4$*. Then for any number* η *with*

$$2\eta \geq \frac{1}{1-\sqrt{\frac{2(d-2)}{(n-d-2)(n-2)}}}$$

the number

$$x_\eta = \frac{d-2}{n-4} + \frac{1}{n-4}\sqrt{\frac{2(d-2)(n-d-2)}{n-2}}\left[\frac{\sqrt{2\eta}}{1+\sqrt{2\eta}\sqrt{2\frac{d-2}{(n-2)(n-d-2)}}} - 1\right]$$

satisfies the relation:

$$\mathbb{P}\left[\left\|\Pi_d(\mathbf{u_{n-1}})\right\|^2 \geq x_\eta\right] \geq \frac{1}{3}e^{-\eta}.$$

The proof is given in Section 16.8

14.4.2 Projection of a Gaussian random vector on randomly oriented d-dimensional subspaces

The following results will help us illustrating the role of isotropy in the minimum penalty phenomenon, in Section 10. To introduce it, we first need to define a uniform measure on the set of r-dimensional subspaces of an n-dimensional Euclidean space.

Definition 14.16 (Uniform measure on a Euclidean Grassmannian variety)**.** Let E_n denote an n-dimensional Euclidean space, $O(n)$ the associated orthogonal group and μ_n its normalized Haar measure. Define the uniform probability measure $\mu_{n,r}$ on the set $G(n,r)$ of r-dimensional Euclidean space of E_n as:

for any measurable $A \subset G(n,r)$,

$$\mu_{n,r}\{A\} := \mu_n\{g \in O(n) \ : \ gV \in A\}$$

where V is an arbitrary r-dimensional subspace of E_n.

Remark 14.17. The definition of $\mu_{n,r}$ ensures that it is a probability measure invariant under the action of $O(n)$. If $\mathbf{g}$ is a random isometry of E_n of law μ_n, then the random subspace $\mathbf{g}V$ is distributed according to $\mu_{n,r}$. The corresponding random orthogonal projector is $\mathbf{g}\Pi_V\mathbf{g}^{-1}$ where Π_V stands for the orthogonal projector on V.

Lemma 14.18 (randomly oriented orthogonal projector, expectations)**.** *Consider a random subspace $\mathbf{V}$ of dimension d of an Euclidean space E_n of dimension n with $d \leq n$, distributed according to the uniform law on $G(n,d)$ introduced in Definition 14.16, and $\Pi_{\mathbf{V}}$ the corresponding random orthogonal projector. Then*

$$\mathbb{E}_V[\Pi_{\mathbf{V}}] = \frac{d}{n}\mathbb{I}_n.$$

In particular, for any linear endomorphism M of E_n,

$$\mathbb{E}_V[\mathrm{Tr}[\Pi_{\mathbf{V}} M \Pi_{\mathbf{V}}]] = \frac{d}{n}\mathrm{Tr}[M],$$

and

$$\mathbb{E}_V[\mathrm{Tr}[(\mathbb{I}_n - \Pi_{\mathbf{V}}) M (\mathbb{I}_n - \Pi_{\mathbf{V}})]] = \left(1 - \frac{d}{n}\right)\mathrm{Tr}[M],$$

and for any vector $u \in E_n$,

$$\mathbb{E}_V\left[\|\Pi_{\mathbf{V}}(u)\|^2\right] = \frac{d}{n}\|u\|^2,$$

and

$$\mathbb{E}_V\left[\|u - \Pi_{\mathbf{V}}(u)\|^2\right] = \left(1 - \frac{d}{n}\right)\|u\|^2.$$

Corollary 14.19 (Projection of a Gaussian random vector on uniformely oriented d–dimensional subspaces)**.** *Let E_n denote an n-dimensional Euclidean space.*

Let w be a standard Gaussian random element of E_n: $W \sim \mathcal{N}(0, \mathbb{I}_n)$. Denote ν_n the associated probability measure on E_n.

Consider an i.i.d. sample $(V_i)_{i=1}^N$ of N d-dimensional subspaces of E_n, drawn along the uniform probability measure $\mu_{n,d}^{\otimes N}$, and $(\Pi_i)_{i=1}^N$ the associated family of orthogonal projectors.

Denote $\mathbb{P}[\cdot]$ the corresponding probability measure $\nu_n \otimes \mu_{n,d}^{\otimes N}$ on $E_n \times G(n,d)^N$.

Denote $\tilde{\Pi}$ a random projector defined so that $\left\|\tilde{\Pi}(W)\right\|^2 = \sup_{1 \le i \le N} \|\Pi_i(W)\|^2$.

Assume that $4 \le d \le \frac{n}{2}$.

Then, for any t with $t \le \frac{N}{10}$ there is a number $x_t^\star$ with

$$x_t^\star \ge \frac{d-2}{n-4} + \frac{1}{n-4}\sqrt{\frac{2(d-2)(n-d-2)}{n-2}}\left[\frac{\sqrt{2\log\left(\frac{N}{3t}\right)}}{1+\sqrt{2\log\left(\frac{N}{3t}\right)\frac{2(d-2)}{(n-2)(n-d-2)}}} - 1\right].$$

so that

$$\mathbb{P}\left[\left\|\tilde{\Pi}(W)\right\|^2 \ge \left(1-\sqrt{\frac{4t}{n}}\right) n x_t^\star\right] \ge \left(1-e^{-t}\right)^2, \tag{14.26}$$

Additionally:

$$\mathbb{E}\left[\left\|\tilde{\Pi}(W)\right\|^2\right] \ge \left(1-e^{-t}\right) n x_t^\star.$$

Proof is given in Section 16.10

14.5 Inequalities based on the properties of the incomplete gamma function

The following lower bound is well known:

Lemma 14.20 (Gaussian tail lower bound, error function)**.** *Denote $\Phi(\cdot)$ the standard normal cumulative distribution function. For any $x > 0$ and $y < \frac{1}{2}e^{-\frac{1}{2}} \approx 0.303$, the following relations hold:*

$$\exp\left(-\frac{1}{2}(x+1)^2\right) \le \frac{\exp\left(-\frac{1+x^2}{2}\right)}{\sqrt{1+x^2}} \le 2\left(1-\Phi(x)\right),$$

and

$$\left[\Phi^{-1}(1-y)\right]^2 \ge -2\log\left(2e^{\frac{1}{2}}y\right) - \log\left(1-2\log\left(2e^{\frac{1}{2}}y\right)\right).$$

The proof is given in Section 16.11

The following will helps us to get a lower bound in probability on the maximum of several independent $\chi^2(m)$ random variables.

Lemma 14.21 (Tail lower bound for a gamma distribution)**.** *Let Z_a be a random variable distributed according to a gamma distribution of parameter $(a, 1)$, namely for $z \geq 0$:*

$$\mathbb{P}\left[Z_a \leq z\right] = \frac{\int_0^z x^{a-1} e^{-x} \mathrm{dx}}{\int_0^\infty x^{a-1} e^{-x} \mathrm{dx}}.$$

Assume that $a > 1$.

Then for any $z \geq a - 1$ the following bound holds:

$$\mathbb{P}\left[Z_a \geq z\right] \geq \frac{1}{e} \exp\left[-(a-1) g\left(\frac{z + \sqrt{a-1} - (a-1)}{a-1}\right)\right].$$

If the random variable X_b has the distribution of a $\chi^2(b)$ random variable, and $b > 2$, then the following bound holds for any $z > b - 2$:

$$\mathbb{P}\left[X_b \geq z\right] \geq \frac{1}{e} \exp\left[-\left(\frac{b}{2} - 1\right) g\left(\frac{z + \sqrt{2b-4} - b - 2)}{b-2}\right)\right].$$

The proof is given in Section 16.12

15 Tools for sorted chi-square samples

15.1 Order statistics

In this section, we define the ordering permutation for a squared i.i.d. standard Gaussian sequence. Consider an i.i.d standard where W is an isonormal Gaussian process. Denote $\xi_1, \cdots, \xi_n$ the random sequence $W(u_1), \cdots, W(u_n)$ where $(u_1, \cdots, u_n)$ is the orthonormal family defined in Assumption 8.1. The sequence $\xi_1, \cdots, \xi_n$ forms a standard Gaussian vector.

In order to define suitable noise-driven models we introduce some definitions and facts regarding the order statistics of the square-noise sequence $\xi_1^2, \xi_2^2, \ldots, \xi_n^2$. We denote ξ^2 this sequence.

Definition 15.1 (Order statistics)**.** Consider an i.i.d. standard Gaussian sequence $\xi_1, \cdots, \xi_n$. We denote $\sigma_n(\cdot)$ the random re-ordering permutation of $[1, n]$ defined almost surely by:

$$\xi^2_{\sigma_n(1)} \leq \xi^2_{\sigma_n(2)} \cdots \leq \xi^2_{\sigma_n(n)}.$$

Denoting F the cumulative distribution function of the absolute value of a normal variable, we define $U_{\sigma_n(1)} \leq U_{\sigma_n(2)} \leq \ldots \leq U_{\sigma_n(2)}$ as the ordered sequence of uniform random variables

$$F(|\xi|_{\sigma_n(1)}) \leq F(|\xi|_{\sigma_n(2)}) \leq \ldots \leq F(|\xi|_{\sigma_n(n)}).$$

L. Birgé and P. Massart studied the behaviour of partial sums of such ordered samples of the chi-square distribution. We recall their result in Section 15.2.

Lemma 15.3 asserts that under reasonable assumptions, the partial sum $\sum_{i=1}^{l} \xi^2_{\sigma_n(n-i+1)}$ behaves like $[1 + o(1)]\, 2 \log \frac{\sqrt{e}n}{l(1+\theta)}$ where θ is a number controlling the approximation in probability.

15.2 Lower bound on the partial sum of an ordered chi-square sample

The following lemma provides a lower bound in probability on the l highest terms of a squared Gaussian sequence.

Lemma 15.2 (Birgé-Massart [Mas07])**.** *Let $\xi^2_{\sigma_n(1)} \leq \cdots \leq \xi^2_{\sigma_n(n)}$ be an ordered sample of size n from the chi-square distribution with one degree of freedom with $\sigma_n(\cdot)$ the corresponding reordering random permutation, j be a positive integer, θ a positive number such that $j(1+\theta) \leq n$ and Φ the standard normal cumulative distribution function. Then*

$$\mathbb{P}\left[\xi^2_{\sigma_n(n+1-j)} \leq \left[\Phi^{-1}\left(1 - \frac{j(1+\theta)}{2n}\right)\right]^2\right] \leq \exp\left[-\frac{j\theta^2}{2(1+\theta)}\right],$$

and consequently if $\theta \geq 2.06$

$$\xi^2_{\sigma_n(n+1-j)} \geq \left[\Phi^{-1}\left(1 - \frac{j(1+\theta)}{2n}\right)\right]^2 \text{ for } 1 \leq j \leq \frac{n}{1+\theta},$$

apart from a set of probability bounded by

$$\left[\exp\left(\frac{\theta^2}{2(1+\theta)}\right)-1\right]^{-1} < 1.$$

Moreover, uniformly for $0 < y < x$,

$$\left[\Phi^{-1}(1-y)\right]^2 = -2\log(y)\left[1+o(1)\right] \text{ when } x \to 0.$$

In the same article, L. Birgé and P. Massart provide a thorough asymptotic analysis of the quantities mentioned in the lemma above. The following lemma illustrates, in a sub-optimal way, that the constants involved are reasonable for practical matters.

Lemma 15.3. *Let* $\xi^2_{\sigma_n(1)} \leq \cdots \leq \xi^2_{\sigma_n(n)}$ *be an ordered i.i.d. sample of size* n *from the chi-square distribution with one degree of freedom,* σ_n *the associated random reordering permutation of* $[1,n]$, l *a positive integer and* θ *a positive number larger than* 2.06 *and such that* $\frac{l(1+\theta)}{n} < e^{-\frac{1}{2}} \approx 0.606$.

Define the random variable $Z_{n,l} = \sum_{i=1}^{l} \xi^2_{\sigma_n(n-i+1)}$.

Then $Z_{n,l}$ *is independent from the random reordering permutation* σ_n *and the function*

$$(n,l,\theta) \mapsto r(n,l,\theta) := \frac{\log\left[2\log\frac{e\,n}{l(1+\theta)}\right] + \frac{\log l + 2}{l}}{1 + 2\log\frac{n}{l(1+\theta)}},$$

is so that:

- $r(n,l,\theta)$ *is positive whenever* $\frac{n}{l(1+\theta)} \geq 1$ *and* $o(1)$ *when* $\frac{n}{(1+\theta)l}$ *tends to infinity,*
- *the following inequality holds:*

$$\mathbb{P}\left[Z_{n,l} < l\left[1-r(n,l,\theta)\right]\left[1+2\log\frac{n}{(1+\theta)l}\right]\right] \leq \left[\exp\left(\frac{\theta^2}{2(1+\theta)}\right)-1\right]^{-1}, \tag{15.1}$$
$$< 1.$$

If $\theta \geq 1 + 2\log 2 + 2t$ *with* $t > 0$, *then the following relation holds:*

$$\left[\exp\left(\frac{\theta^2}{2(1+\theta)}\right)-1\right]^{-1} \leq e^{-t}. \tag{15.2}$$

Additionally the following inequality holds for $n \geq l \geq 1$, $\frac{l(1+\theta)}{n} \leq e^{-\frac{1}{2}}$ *and* $\theta > 2.06$*:*

$$r(n,l,\theta) \leq \frac{2 + 1.5 l^{-\frac{1}{2}}}{1 + \sqrt{2\log\frac{en}{l(1+\theta)}}}. \tag{15.3}$$

The proof is given in Section 17.1.

15.3 Upper bound on the partial sum of an ordered chi-square sample

Mirroring the preceding section, this section is also about the sum of the $n-l$ first squared components of a random Gaussian vector. The following lemma help showing how close it is from the expectation $n-l$ of its 'simply Gaussian' equivalent.

Lemma 15.4. *Let $\xi_1, \cdots, \xi_n$ be a standard Gaussian vector of dimension n, $\xi^2_{\sigma_n(1)} \leq \cdots \leq \xi^2_{\sigma_n(n)}$ be the associated sorted sample from the chi-square distribution with one degree of freedom, $\sigma_n(\cdot)$ the associated reordering permutation and l a integer with $0 < l < n$.*
For any positive real number t the following inequality holds:

$$\mathbb{P}\left[\xi_1^2 \geq 2\log\left(\frac{2n}{l}\right) + 2t \,\middle|\, \xi_1^2 \geq \xi^2_{\sigma_n(n-l+1)}\right] \leq e^{-t}.$$

The following inequalities holds:

$$\mathbb{E}\left[\xi_1^2 \,\middle|\, \xi_1^2 \geq \xi^2_{\sigma_n(n-l+1)}\right] \leq 2\log\left(\frac{n}{l}\right) + 2 + 2\log 2,$$

and

$$\mathbb{E}\left[\xi_1^2 \,\middle|\, \xi_1^2 \leq \xi^2_{\sigma_n(n-l)}\right] \geq 1 - \frac{l}{n-l}\left(2\log\left(\frac{n}{l}\right) + 1 + 2\log 2\right).$$

The proof is given in Section 17.2.

16 Proofs for Section 14 Tail lower bound inequalities

16.1 Proof of Lemma 14.2

Proof. As well known the h function is convex with a null minimum in $u = 0$.

The upper bound in 14.1 for positive u derives from the fact that h has second derivative $h''(u) = \frac{1}{1+u}$.

The lower bound for any u in 14.1 and 14.2 appears in [Mas07, p 24] and can be derived by noticing that the second derivative of $\left(1 + \frac{u}{3}\right) h(u)$ is greater than 1, as shows its expression: $1 - \frac{2}{3}\frac{u}{1+u} + \frac{2}{3}\log(1+u)$ combined with the fact that $\log(1+u) = -\log\left(1 - \frac{u}{1+u}\right) \geq \frac{u}{1+u}$ for $u > -1$.

To show the upper bound in 14.2 for $0 < u < 1$, we observe that $i \leq 2^{i-1}$ for any positive integer i, so that:

$$\begin{aligned}(1-u)\log(1-u) + u &= u - \sum_{i=1}^{\infty}(1-u)\frac{u^i}{i}, \\ &\leq u - \sum_{i=1}^{\infty}(1-u)\frac{u^i}{2^{i-1}}, \\ &= \frac{u^2}{2\left(1 - \frac{u}{2}\right)}.\end{aligned}$$

By algebraic inversion the upper bound in 14.1 for $u > 0$ implies the lower bound in 14.3. The lower bound in 14.1 implies the upper bound in 14.3 as follows:

$$u^2 - 2\frac{h(u)}{3}u - 2h(u) \leq 0,$$

implies

$$u \leq \frac{h(u)}{3} + \sqrt{2h(u) + \frac{h(u)^2}{9}},$$

and by the relation $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x > 0$ and $y > 0$,

$$u \leq \sqrt{2h(u)} + \frac{2}{3}h(u).$$

□

16.2 Proof of Lemma 14.4

Proof. The lower bound in 14.4 results from the fact that for $u > -1$, $\log(1+u) \leq u$ so that $u - \log(1+u) = u - 2\log\left(\sqrt{1+u}\right) \geq u + 2 - 2\sqrt{1+u} = \left(\sqrt{1+u} - 1\right)^2$.

The lower bounds in 14.5 and 14.6 for all $u > -1$ are shown in [Mas90], and results from the fact that the second derivative of the function $u \mapsto \left(1 + \frac{2u}{3}\right)(u - \log(1+u))$ is $1 + \frac{u^2}{3(1+u)^2}$ which is more than 1 for any $u > -1$.

The upper bound in 14.5 results from the observation that the second derivative of the function $u \mapsto \left(1+\frac{u}{2}\right)(u-\log(1+u))$ is $1-\frac{u}{2(1+u)^2}$ which is less than 1 for any $u>0$.

The upper bound in 14.6 is based on the observation that the second derivative of the function $u \mapsto (1-u)(-u-\log(1-u))$ is $1-\frac{u}{1-u}$ which is less than 1 for any $u>0$. □

16.3 Proof of Lemma 14.6

Proof. The lower bound in 14.7 arises from the lower bound of the $h(\cdot)$ function in Lemma 14.2, and is shown in [Mas90, Lemma 1].

The upper bound in Inequalities 14.8 and 14.9 result from the upper bounds on the $h(\cdot)$ function in 14.1 and 14.2 in Lemma 14.2:

$$\begin{aligned}
h_p(p+u) &= ph\left(\frac{u}{p}\right)+(1-p)h\left(-\frac{u}{1-p}\right), \\
&\leq \frac{u^2}{2p}+\frac{u^2}{2\left(1-p-\frac{u}{2}\right)}, \\
&= \frac{u^2}{2p}\frac{1-\frac{u}{2}}{\left(1-p-\frac{u}{2}\right)}, \\
&\leq \frac{u^2}{2p\left(1-p-\frac{u}{2}\right)}, \\
&\leq \frac{u^2}{p(1-p)}, \text{ since } u \leq 1-p.
\end{aligned}$$

With the shorthand $h := h_p(p+u)$, algebraic inversion leads to the upper bound in 14.10:

$$\begin{aligned}
0 &\leq u^2+hpu-2p(1-p)h, \\
u &\geq \sqrt{2p(1-p)h+\frac{h^2p^2}{2}}-\frac{hp}{2}, \\
&= \frac{2p(1-p)h}{\sqrt{2p(1-p)h+\frac{h^2p^2}{2}}+\frac{hp}{2}}, \\
&\geq \frac{2p(1-p)h}{\sqrt{2p(1-p)h}+hp}, \\
&= \frac{\sqrt{2p(1-p)h}}{1+\sqrt{\frac{p}{2(1-p)}h}}.
\end{aligned}$$

The bound for the lower deviations in 14.11 and 14.12 follow by the symmetry $h_p(x) = h_{1-p}(1-x)$.
To prove the inverted form in 14.13 and 14.14, we start from the relation 14.11 for $0 \leq u < p < 1$

$$h_p(p-u) \leq \frac{u^2}{2\left(p-\frac{u}{2}\right)(1-p)},$$

and for $0 < x < x + u < 1$, set $p = x + u$ so that

$$h_{x+u}(x) \leq \frac{u^2}{2\left(x + \frac{u}{2}\right)(1 - x - u)},$$

leading to the the weaker but more tractable:

$$h_{x+u}(x) \leq \frac{u^2}{2x(1 - x - u)} \text{ for } 0 < x < x + u < 1.$$

With the shorthand $h := h_{x+u}(x)$, algebraic inversion leads to:

$$\begin{aligned}
0 &\geq u^2 + 2xhu - 2x(1-x)h, \\
u &\geq \sqrt{2x(1-x)h + x^2h^2} - xh, \\
&= \frac{2x(1-x)h}{\sqrt{2x(1-x)h + x^2h^2} + xh}, \\
&\geq \frac{2x(1-x)h}{\sqrt{2x(1-x)h} + 2xh}, \\
&= \frac{\sqrt{2x(1-x)h}}{1 + \sqrt{\frac{2xh}{1-x}}}, \\
&= (1-x)\frac{\sqrt{2xh}}{\sqrt{1-x} + \sqrt{2xh}}.
\end{aligned}$$

which is the announced relation when choosing $u = p - x$ □

16.4 Proof of Lemma 14.10

Proof. A quick calculation shows that for $\lambda \geq -\frac{1}{2\sigma^2}$,

$$\log \mathbb{E}\left[e^{-\lambda\left(\|s+\sigma W\|^2 - \|s\|^2 - n\sigma^2\right)}\right] = \|s\|^2 \frac{2\lambda^2\sigma^2}{1 + 2\lambda\sigma^2} + n\lambda\sigma^2 - \frac{n}{2}\log\left(1 + 2\lambda\sigma^2\right)$$

and by the relation $u - \log(1+u) \leq \frac{u^2}{2\left(1+\frac{u}{2}\right)}$ for $u > 0$ (see Inequality 14.5 in Lemma 14.4)

$$\log \mathbb{E}\left[e^{-\lambda\left(\|s+\sigma W\|^2 - \|s\|^2 - n\sigma^2\right)}\right] \leq \left(4\|s\|^2\sigma^2 + 2n\sigma^4\right)\frac{\lambda^2}{2}.$$

Denote $\psi(\lambda)$ the quantity on the left of the relation above. Set for $0 \leq t$

$$z_t = \inf\left\{\frac{\psi(\lambda) + t}{\lambda},\ \lambda > 0\right\},$$

and denote $\lambda^\star$ the value of λ realizing the minimum. Then

$$z_t \leq \sqrt{2t\left(4\|s\|^2\sigma^2 + 2n\sigma^4\right)},$$

and Markov's inequality:

$$\begin{aligned}\mathbb{P}\left[\|s+\sigma W\|^2 \leq \|s\|^2 + n\sigma^2 - z_t\right] &\leq e^{\psi(\lambda^\star)-\lambda^\star z_t},\\ &\leq e^{-t},\end{aligned}$$

which is the relation announced in Inequality 14.21, for any $t \geq 0$:

$$\mathbb{P}\left[\|s+\sigma W\|^2 \leq \|s\|^2 + n\sigma^2 - \sqrt{2t\left(4\|s\|^2\sigma^2 + 2n\sigma^4\right)}\right] \leq e^{-t}.$$

Turning now to the upper variations, for $\lambda < \frac{1}{2\sigma^2}$:

$$\begin{aligned}\psi(\lambda) &= \log \mathbb{E}\left[e^{\lambda\left(\|s+\sigma W\|^2 - \|s\|^2 - n\sigma^2\right)}\right]\\ &= \|s\|^2 \frac{2\lambda^2\sigma^2}{1-2\lambda\sigma^2} - n\lambda\sigma^2 - \frac{n}{2}\log\left(1-2\lambda\sigma^2\right),\end{aligned}$$

and by the relation $-\log(1-u) - u \leq \frac{u^2}{2(1-u)}$ for $0 < u < 1$ (see Lemma 14.4)

$$\begin{aligned}\psi(\lambda) &\leq \|s\|^2 \frac{2\lambda^2\sigma^2}{1-2\lambda\sigma^2} + \frac{n}{2}\frac{4\lambda^2\sigma^2}{2\left(1-2\lambda\sigma^2\right)},\\ &= \left(4\|s\|^2\sigma^2 + 2n\sigma^2\right)\frac{\lambda^2}{2\left(1-2\lambda\sigma^2\right)}.\end{aligned}$$

so that

$$\begin{aligned}z_t &= \inf\left\{\frac{\psi(\lambda)+t}{\lambda},\ \lambda > 0\right\},\\ &\leq \sqrt{2t\left(4\|s\|^2\sigma^2 + 2n\sigma^4\right)} + 2\sigma^2 t,\end{aligned}$$

and the relation announced in Inequality 14.22 holds:

$$\mathbb{P}\left[\|s+\sigma W\|^2 \geq \|s\|^2 + n\sigma^2 + \sqrt{2t\left(4\|s\|^2\sigma^2 + 2n\sigma^4\right)} + 2\sigma^2 t\right] \leq e^{-t}.$$

□

16.5 Proof of Lemma 14.11

Proof. For any λ with $0 \leq \lambda < 1$ and $z > 0$, by Markov's inequality

$$\begin{aligned}\mathbb{P}\left[G_n > z\right] &\leq e^{-\lambda z}\,\mathbb{E}\left[e^{\lambda G_n}\right],\\ &= e^{-\lambda z}(1-\lambda)^{-n},\\ &= e^{-n\log(1-\lambda)-\lambda z},\end{aligned}$$

and with $\lambda = 1 - \frac{n}{z}$, by Lemma 14.4,

$$\begin{aligned}\mathbb{P}\left[G_n > z\right] &\leq e^{n\log\left(\frac{z}{n}\right)-z+n},\\ &= e^{-ng\left(\frac{z}{n}-1\right)},\\ &\leq e^{-n\left(\sqrt{\frac{z}{n}}-1\right)^2},\end{aligned}$$

and for $t \geq 0$ setting $z = n\left(1+\sqrt{\frac{t}{n}}\right)^2$,

$$\mathbb{P}\left[G_n > n\left(1+\sqrt{\frac{t}{n}}\right)^2\right] \leq e^{-t}.$$

The lower tail inequality is proved in the same manner:

$$\begin{aligned}\mathbb{P}\left[G_n < z\right] &\leq e^{\lambda z}\,\mathbb{E}\left[e^{-\lambda G_n}\right], \\ &= e^{\lambda z}(1+\lambda)^{-n}, \\ &= e^{-n\log(1+\lambda)+\lambda z},\end{aligned}$$

and with $\lambda = \frac{n}{z} - 1$,

$$\begin{aligned}\mathbb{P}\left[G_n > z\right] &\leq e^{n\log\left(\frac{z}{n}\right)-z+n}, \\ &\leq e^{-n\left(1-\sqrt{\frac{z}{n}}\right)^2},\end{aligned}$$

and for $0 \leq t \leq n$

$$\begin{aligned}\mathbb{P}\left[G_n < n\left(1-\sqrt{\frac{t}{n}}\right)^2\right] &\leq e^{-t}, \\ \mathbb{P}\left[G_n < n - 2\sqrt{nt}\right] &\leq e^{-t}.\end{aligned}$$

□

16.6 Proof of Lemma 14.13

Proof. starting with the relation

$$\mathbb{P}\left[B_{n,p} = l\right] = p^l(1-p)^{n-l}\frac{n!}{l!(n-l)} \text{ for } 0 \leq l \leq n,$$

the Stirling inequality [Rob55]

$$\sqrt{2\pi l}\left(\frac{l}{e}\right)^l \leq l! \leq \sqrt{e^2 l}\left(\frac{l}{e}\right)^l,$$

and Definition 14.5:

$$h_p(x) := x\log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right) \text{ for } 0 < x < 1 \text{ and } 0 < p < 1,$$

we observe that

$$\begin{aligned}\mathbb{P}\left[B_{n,p} = l\right] &\geq \sqrt{2\pi}e^{-2}\sqrt{\frac{n}{l(n-l)}}\left(\frac{np}{l}\right)^l\left(\frac{n(1-p)}{n-l}\right)^{n-l}, \\ &\geq \frac{\sqrt{2\pi}e^{-2}}{\sqrt{np(1-p)}}\left(\frac{np}{l}\right)^{l+\frac{1}{2}}\left(\frac{n(1-p)}{n-l}\right)^{n-l}, \\ &= \frac{\sqrt{2\pi}e^{-2}}{\sqrt{np(1-p)}}\left[\left(\frac{n}{n+\frac{1}{2}}\right)^{n+\frac{1}{2}}\left(\frac{l+\frac{1}{2}}{l}\right)^{l+\frac{1}{2}}\right]\exp\left[-\left(n+\frac{1}{2}\right)h_p\left(\frac{l+\frac{1}{2}}{n+\frac{1}{2}}\right)\right].\end{aligned}$$

The function $x \mapsto \left(x+\frac{1}{2}\right)\log\left(\frac{x+\frac{1}{2}}{x}\right)$ for $x > 0$ has for derivative $\log\left(1+\frac{1}{2x}\right) - \frac{1}{2x}$, which is negative, so that the bracketed term in the inequality above is larger than one since $l \leq n$. This leads to:

$$\mathbb{P}\left[B_{n,p} = l\right] \geq \frac{\sqrt{2\pi}e^{-2}}{\sqrt{np(1-p)}} \exp\left[-\left(n+\frac{1}{2}\right) h_p\left(\frac{l+\frac{1}{2}}{n+\frac{1}{2}}\right)\right].$$

By assumption $np \leq k$, so that $p \leq \frac{k+\frac{1}{2}}{n+\frac{1}{2}}$, and the h_p function is convex with a minimum in p, so that the upper bound above is decreasing with respect to l for $l \geq k$.

By the assumption that $k \leq n - \sqrt{np(1-p)} - \frac{1-p}{2}$, the relation above applies for $k \leq l \leq k + \left\lfloor\sqrt{np(1-p)}\right\rfloor$, which represents $\left\lfloor\sqrt{np(1-p)}\right\rfloor + 1$ values of the integer l. Taking the sum yields,

$$\mathbb{P}\left[B_{n,p} \geq k\right] \geq \sqrt{2\pi}e^{-2} \exp\left[-\left(n+\frac{1}{2}\right) h_p\left(\frac{k+\left\lfloor\sqrt{np(1-p)}\right\rfloor+\frac{1}{2}}{n+\frac{1}{2}}\right)\right],$$

(invoking the log-concavity of the bound would provide us with a slightly better but less tractable expression.) As mentionned since $np \leq k$ then $p \leq \frac{k+\frac{1}{2}}{n+\frac{1}{2}}$, and the argument of the h_p function in the relation above lies in the region where this function is increasing. Then we may slightly degrade the bound above by replacing the integer part $\left\lfloor\sqrt{np(1-p)}\right\rfloor$ by $\sqrt{np(1-p)}$, with:

$$\mathbb{P}\left[B_{n,p} \geq k\right] \geq \sqrt{2\pi}e^{-2} \exp\left[-\left(n+\frac{1}{2}\right) h_p\left(\frac{k+\sqrt{np(1-p)}+\frac{1}{2}}{n+\frac{1}{2}}\right)\right].$$

Since $\sqrt{2\pi}e^{-2} \sim 0.339$ exceeds $\frac{1}{3}$, this yields the first relation announced in the lemma.

The h function is convex, with a null minimum in $x = p$. Then the following relation is defined since by assumption $k \leq n - \sqrt{np(1-p)} - \frac{1-p}{2}$, and holds:

$$\begin{aligned}\left(n+\frac{1}{2}\right) h_p\left(\frac{k+\sqrt{np(1-p)}+\frac{1}{2}}{n+\frac{1}{2}}\right) &\leq n h_p\left(\frac{k+\sqrt{np(1-p)}+\frac{1}{2}-\frac{p}{2}}{n}\right) + \frac{1}{2}h_p(p), \\ &= n h_p\left(\frac{k+\sqrt{np(1-p)}+\frac{1-p}{2}}{n}\right),\end{aligned}$$

yielding the second relation announced:

$$\mathbb{P}\left[B_{n,p} \geq k\right] \geq \frac{1}{3}\exp\left[-n h_p\left(\frac{k+\sqrt{np(1-p)}+\frac{1-p}{2}}{n}\right)\right].$$

□

16.7 Proof of Lemma 14.14

Proof. Set

$$\begin{aligned}\nu &:= \alpha+\beta,\\ t_0 &:= \frac{\alpha-1}{\alpha+\beta-2},\\ \text{and } \sigma &:= \frac{1}{\alpha+\beta-2}\sqrt{\frac{(\alpha-1)(\beta-1)}{\alpha+\beta-1}} = \sqrt{\frac{t_0(1-t_0)}{\nu-1}}.\end{aligned}$$

For future reference, we first check that the number σ is less than t_0 and $1-t_0$. Indeed, since $\min(\alpha-1,\beta-1)\geq 1$:

$$\begin{aligned}\sigma &= \frac{1}{\alpha+\beta-2}\sqrt{\frac{(\alpha-1)(\beta-1)}{\alpha+\beta-1}},\\ &= \min(t_0,1-t_0)\sqrt{\frac{\max(\alpha-1,\beta-1)}{\min(\alpha-1,\beta-1)(\alpha+\beta-1)}},\\ &\leq \min(t_0,1-t_0).\end{aligned}$$

We know from Sonin's formula [Kup] that for $x>0$, there is a number $\theta(x)$ with $0<\theta(x)<\frac{1}{2}$ so that

$$\Gamma(1+x) = \sqrt{2\pi}\,x^{x+\frac{1}{2}}\,e^{-x+\frac{1}{12(x+\theta(x))}}.$$

Since $\alpha-1\geq 1$ and $\beta-1\geq 1$, this leads to the following relations:

$$\begin{aligned}\Gamma(\alpha) &\leq \sqrt{2\pi(\alpha-1)}\left(\frac{\alpha-1}{e}\right)^{\alpha-1} e^{\frac{1}{12(\alpha-1)}},\\ \Gamma(\alpha) &\leq \sqrt{2\pi(\beta-1)}\left(\frac{\beta-1}{e}\right)^{\beta-1} e^{\frac{1}{12(\beta-1)}},\\ \Gamma(\alpha+\beta) &\geq \sqrt{2\pi(\alpha+\beta-1)}\left(\frac{\alpha+\beta-1}{e}\right)^{\alpha+\beta-1},\\ \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} &\leq e^{1+\frac{1}{6}}\sqrt{2\pi}\frac{(\alpha-1)^{\alpha-1+\frac{1}{2}}\,(\beta-1)^{\beta-1+\frac{1}{2}}}{(\alpha+\beta-1)^{\alpha+\beta-1+\frac{1}{2}}},\\ &= e^{1+\frac{1}{6}}\sqrt{2\pi}\frac{1}{\alpha+\beta-2}\sqrt{\frac{(\alpha-1)(\beta-1)}{\alpha+\beta-1}}\left(\frac{\alpha+\beta-2}{\alpha+\beta-1}\right)^{\alpha+\beta-1}\frac{(\alpha-1)^{\alpha-1}\,(\beta-1)^{\beta-1}}{(\alpha+\beta-2)^{\alpha+\beta-2}}.\end{aligned}$$

By the inequality $\left(\frac{\alpha+\beta-2}{\alpha+\beta-1}\right)^{\alpha+\beta-1} = \exp\left[(\alpha+\beta-1)\log\left(1-\frac{1}{\alpha+\beta-1}\right)\right] \leq e^{-1}$, this leads in turn to:

$$\begin{aligned}\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} &\leq e^{\frac{1}{6}}\sqrt{2\pi}\frac{1}{\alpha+\beta-2}\sqrt{\frac{(\alpha-1)(\beta-1)}{\alpha+\beta-1}}\frac{(\alpha-1)^{\alpha-1}\,(\beta-1)^{\beta-1}}{(\alpha+\beta-2)^{\alpha+\beta-2}},\\ &\leq 3\sigma t_0^{\alpha-1}(1-t_0)^{\beta-1}.\end{aligned}$$

Combining with the expression of the incomplete beta function ensures that:

$$\begin{aligned}\mathbb{P}\left[Z_{\alpha,\beta} \leq t\right] &\geq \frac{1}{3\sigma}\int_0^t \exp\left[(\nu-2)\left(t_0 \log\left(\frac{u}{t_0}\right) + (1-t_0)\log\left(\frac{1-u}{1-t_0}\right)\right)\right] \mathrm{du}, \\ &= \frac{1}{3\sigma}\int_0^t \exp\left[-(\nu-2)h_u(t_0)\right] \mathrm{du},\end{aligned}$$

so that with the assumption that $\sigma < t \leq t_0$, since the function $u \mapsto h_u(t_0)$ is decreasing for $u < t_0$ and increasing for $u > t_0$,

$$\mathbb{P}\left[Z_{\alpha,\beta} \leq t\right] \geq \frac{1}{3}\exp\left[-(\nu-2)h_{t-\sigma}(t_0)\right].$$

In the same manner, for $t_0 \leq t < 1-\sigma$

$$\mathbb{P}\left[Z_{\alpha,\beta} \geq t\right] \geq \frac{1}{3}\exp\left[-(\nu-2)h_{t+\sigma}(t_0)\right]. \tag{16.1}$$

To get an inverted form or the relation above, recall that Lemma 14.6 states that for $0 < x < p < 1$, the following bound holds:

$$p \geq x + \frac{\sqrt{2x(1-x)h_p(x)}}{1+\sqrt{\frac{2xh_p(x)}{1-x}}}. \tag{16.2}$$

Assume that

$$\eta \geq (\nu-2)h_{t_0+\sigma}(t_0). \tag{16.3}$$

The function $z \mapsto h_z(t_0)$ goes from 0 to ∞ when z goes from t_0 to ∞. Then there is a threshold $z^\star \geq t_0$ such that the three following relations hold:

$$\begin{aligned}(\nu-2)h_{z^\star+\sigma}(t_0) &= \eta, \\ z^\star &\geq t_0 - \sigma + \frac{\sqrt{2t_0(1-t_0)\eta}}{\sqrt{\nu-2}+\sqrt{\frac{2t_0\eta}{1-t_0}}} \text{ (by Ineq. 16.2)}, \qquad (16.4) \\ \mathbb{P}\left[Z_{\alpha,\beta} \geq z^\star\right] &\geq \frac{1}{3}e^{-\eta} \text{ (by Ineq. 16.1)}. \qquad (16.5)\end{aligned}$$

Lets now determine a range for the parameter η consistent with the condition 16.3. Lemma 14.6 also states that for $0 < p-u < p < 1$,

$$h_p(p-u) \leq \frac{u^2}{2\left(p-\frac{u}{2}\right)(1-p)},$$

so setting $u = \sigma$ and $p = t_0 + \sigma$ ensures that:

$$\begin{aligned}(\nu-2)h_{t_0+\sigma}(t_0) &\leq (\nu-2)\frac{\sigma^2}{2\left(t_0+\frac{\sigma}{2}\right)(1-t_0-\sigma)}, \\ &\leq \frac{\nu-2}{\nu-1}\frac{t_0(1-t_0)}{2t_0\left(1-t_0-\sqrt{\frac{t_0(1-t_0)}{\nu-1}}\right)}, \\ &\leq \frac{1}{2\left(1-\sqrt{\frac{t_0}{(1-t_0)(\nu-1)}}\right)},\end{aligned}$$

and the condition

$$\eta \geq \frac{1}{2}\frac{1}{1-\sqrt{\frac{t_0}{(1-t_0)(\nu-1)}}}, \tag{16.6}$$

is sufficient to ensure that $(\nu-1)h_{t_0+\sigma}(t_0) \leq \eta$ and $z^\star \geq t_0$. Under this condition, inequalities 16.4 and 16.5 imply that

$$\mathbb{P}\left[Z_{\alpha,\beta} \geq t_0 - \sigma + \frac{\sqrt{2t_0(1-t_0)\eta}}{\sqrt{\nu-2}+\sqrt{\frac{2t_0\eta}{1-t_0}}}\right] \geq \frac{1}{3}e^{-\eta}. \tag{16.7}$$

and for slightly better readability:

$$\begin{aligned}
&\mathbb{P}\left[Z_{\alpha,\beta} \geq t_0 - \sigma + \frac{\sqrt{2t_0(1-t_0)\eta}}{\sqrt{\nu-1}+\sqrt{\frac{2t_0\eta}{1-t_0}}}\right] \geq \frac{1}{3}e^{-\eta}, \\
&\mathbb{P}\left[Z_{\alpha,\beta} \geq t_0 + \sigma\left[\frac{\sqrt{2\eta}}{1+\frac{\sigma}{1-t_0}\sqrt{2\eta}} - 1\right]\right] \geq \frac{1}{3}e^{-\eta}.
\end{aligned} \tag{16.8}$$

□

16.8 Proof of Corollary 14.15

Proof. Recall that $\mathbf{u_{n-1}}$ is a uniformly distributed unit vector on the unit sphere S_{n-1} in the n-dimensional Euclidean space. Let start by re-stating the well known fact that the random variable $\|\Pi_d(\mathbf{u_{n-1}})\|^2$ follows a Beta law of parameter $\left(\frac{d}{2}, \frac{n-d}{2}\right)$. To do so consider a $\chi^2(n)$ random variable Z. Then by its isotropy and the distribution of its norm, the vector $Z\mathbf{u_{n-1}}$ has the distribution of a standard Gaussian vector in $\mathbb{R}^n$. It follows that $Z\|\Pi_d(\mathbf{u_{n-1}})\|^2$ and $Z\|(\mathbb{I}_n - \Pi_d)(\mathbf{u_{n-1}})\|^2$ are together distributed like an independent pair of a $\chi^2(d)$ variable X and a $\chi^2(n-d)$ variable Y. Then $\|\Pi_d(\mathbf{u_{n-1}})\|^2$ has the distribution of $\frac{X}{X+Y}$, which is a Beta law of parameter $\left(\frac{d}{2}, \frac{n-d}{2}\right)$.

Then the assumption that $4 \leq d \leq n-4$ allows to invoke the tail lower bound for a beta distribution in Lemma 14.14 with

$$\begin{aligned}
\alpha &= \frac{d}{2} \geq 2, \\
\beta &= \frac{n-d}{2} \geq 2, \\
\nu &= \frac{a+b}{2} = \frac{n}{2}, \\
t_0 &= \frac{\alpha-1}{\nu-2} = \frac{d-2}{n-4}, \\
\sigma &= \sqrt{\frac{t_0(1-t_0)}{\nu-1}} = \frac{1}{n-4}\sqrt{\frac{2(d-2)(n-d-2)}{n-2}}.
\end{aligned}$$

It follows that the number

$$\begin{aligned} x_\eta &:= t_0 + \sigma \left[\frac{\sqrt{2\eta}}{1 + \frac{\sigma}{1-t_0}\sqrt{2\eta}} - 1 \right], \\ &= \frac{d-2}{n-4} + \frac{1}{n-4}\sqrt{\frac{2(d-2)(n-d-2)}{n-2}} \left[\frac{\sqrt{2\eta}}{1 + \sqrt{4\eta \frac{d-2}{(n-2)(n-d-2)}}} - 1 \right] \end{aligned}$$

satisfies for any η with

$$\begin{aligned} 2\eta &\geq \frac{1}{1 - \sqrt{\frac{t_0}{(1-t_0)(\nu-1)}}}, \\ &= \frac{1}{1 - \sqrt{\frac{2(d-2)}{(n-d-2)(n-2)}}} \end{aligned}$$

the relation:

$$\mathbb{P}\left[\|\Pi_d(\mathbf{u_{n-1}})\|^2 \geq x_\eta \right] \geq \frac{1}{3} e^{-\eta}.$$

□

16.9 Proof of Lemma 14.18

Proof. As all the random quantities involved are bounded, all the expectations mentioned in the lemma are finite. If $n = 1$, there is only one projector of rank d and the statements in the lemma are trivial.

If $n > 1$, consider an isometry g of E_n. By construction (see Definition 14.16), the distribution of $\mathbf{V}$ is invariant under g, so that the random projector $g\Pi_{\mathbf{V}}g^{-1}$ follows the same law than $\Pi_{\mathbf{V}}$. It follows that

$$\begin{aligned} \mathbb{E}_V\left[\Pi_{\mathbf{V}}\right] &= \mathbb{E}_V\left[g\Pi_{\mathbf{V}}g^{-1}\right], \\ &= g\,\mathbb{E}_V\left[\Pi_{\mathbf{V}}\right]g^{-1}, \end{aligned}$$

which shows that the operator $\mathbb{E}_V\left[\Pi_{\mathbf{V}}\right]$ commutes with any element of $O(n)$. Denote $(m_{i,j})_{(i,j)\in[1,n]^2}$ the matrix of $\mathbb{E}_V\left[\Pi_{\mathbf{V}}\right]$ in an arbitrary orthogonal basis $(e_1, \cdots, e_n)$ of E_n. In particular for any indexes (i,j) with $i \neq j$, the operator $\mathbb{E}_V\left[\Pi_{\mathbf{V}}\right]$ commutes with the reflexion exchanging e_i with $-e_i$, and with the reflexion exchanging e_i and e_j, respectively

$$\mathbb{I}_n - 2e_i \otimes e_i$$

and

$$\mathbb{I}_n - 2\frac{e_i - e_j}{\sqrt{2}} \otimes \frac{e_i - e_j}{\sqrt{2}} = \mathbb{I}_n - e_i \otimes e_i - e_j \otimes e_j + e_i \otimes e_j + e_j \otimes e_i.$$

A direct calculation shows that for $1 \leq i < j \leq n$

$$m_{i,i} = m_{j,j},$$

and

$$m_{i,j} = -m_{i,j},$$

so that $\mathbb{E}_V[\Pi_{\mathbf{V}}]$ is diagonal and proportional to the identity operator $\mathbb{I}_n$. As

$$\mathrm{Tr}[\Pi_{\mathbf{V}}] = d \text{ a.s.},$$

and

$$\mathrm{Tr}[\mathbb{I}_n] = n,$$

it follows that

$$\mathbb{E}_V[\Pi_{\mathbf{V}}] = \frac{d}{n}\mathbb{I}_n.$$

As $\Pi_{\mathbf{V}}$ is an orthogonal projector,

$$\begin{aligned}\mathbb{E}_V[\mathrm{Tr}[\Pi_{\mathbf{V}} M \Pi_{\mathbf{V}}]] &= \mathbb{E}_V\left[\mathrm{Tr}\left[\Pi_{\mathbf{V}}^2 M\right]\right], \\ &= \mathbb{E}_V[\mathrm{Tr}[\Pi_{\mathbf{V}} M]], \\ &= \mathrm{Tr}[\mathbb{E}_V[\Pi_{\mathbf{V}}] M], \\ &= \frac{d}{n}\mathrm{Tr}[M].\end{aligned}$$

The operator $\mathbb{I}_n - \Pi_{\mathbf{V}}$ is also an orthogonal projector and

$$\begin{aligned}\mathbb{E}_V[\mathrm{Tr}[(\mathbb{I}_n - \Pi_{\mathbf{V}}) M (\mathbb{I}_n - \Pi_{\mathbf{V}})]] &= \mathbb{E}_V[\mathrm{Tr}[(\mathbb{I}_n - \Pi_{\mathbf{V}}) M]], \\ &= \left(1 - \frac{d}{n}\right)\mathrm{Tr}[M].\end{aligned}$$

The last equalities in the lemma follow by bilinearity with:

$$\begin{aligned}\mathbb{E}_V\left[\|\Pi_{\mathbf{V}} u\|^2\right] &= \mathbb{E}_V[\langle u, \Pi_{\mathbf{V}} u\rangle], \\ &= \left\langle u, \frac{d}{n}\mathbb{I}_n u\right\rangle, \\ &= \frac{d}{n}\|u\|^2,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_V\left[\|u - \Pi_{\mathbf{V}} u\|^2\right] &= \|u\|^2 - \mathbb{E}_V[\langle u, \Pi_{\mathbf{V}} u\rangle], \\ &= \left(1 - \frac{d}{n}\right)\|u\|^2.\end{aligned}$$

□

16.10 Proof of Corollary 14.19

Proof. Denote $\mathbf{u}$ a random unit vector satisfying

$$\frac{W}{\|W\|} = \mathbf{u} \text{ a.s.}.$$

For $1 \leq i \leq N$, denote by Π_i the orthogonal projector on V_i, and $\mathbf{u_i}$ and $\mathbf{w_i}$ the images of $\mathbf{u}$, and W by Π_i. Then

$$\mathbf{u_i} = \Pi_i(\mathbf{u}) = \frac{\mathbf{w_i}}{\|W\|} \text{ a.s..}$$

The present proof relies on the i.i.d property of the $(\|\mathbf{u_i}\|)_{i=1}^n$ sample, on the tail lower bound for a beta distribution stated in Lemma 14.14 and on the strong concentration of $\|W\|^2$ around its expectation n.

I.i.d property of the $(\|\mathbf{u_i}\|)_{i=1}^n$ sample To check the i.i.d property of the $(\|\mathbf{u_i}\|)_{i=1}^n$ sample, we choose:

Π_0 an arbitrary d-dimensional projector of the Euclidean space $\mathbb{R}^n$,
R an uniform random isometry satisfying $\Pi = R^{-1}\Pi_0 R$ a.s. as in the remark following Definition 14.16,
u_0 an arbitrary fixed unit vector.

Then $\|\Pi(\mathbf{u})\|$ is equal to $\|\Pi_0 R(\mathbf{u})\|$ a.s. and $R(\mathbf{u})$ is a uniform random unit vector, as such independent of $\mathbf{u}$, so that the sample $(\|\mathbf{u_i}\|)_{i=1}^n$ is distributed as an i.i.d. sample of $\|\Pi_0 R(u_0)\|$.

Projections of the normalized signal Our next step is to apply the tail lower bound for the projection of a uniform random unit vector in Corollary 14.15. This requires that $4 \leq d \leq n-4$, which is satisfied by the assumptions that $4 \leq d \leq \frac{n}{2}$. It follows that for any number η with

$$\eta \geq \frac{1}{2} \frac{1}{1 - \sqrt{\frac{2(d-2)}{(n-d-2)(n-2)}}}$$

the number

$$x_\eta = \frac{d-2}{n-4} + \frac{1}{n-4}\sqrt{\frac{2(d-2)(n-d-2)}{n-2}}\left[\frac{\sqrt{2\eta}}{1+\sqrt{2\eta}\sqrt{\frac{2(d-2)}{(n-2)(n-d-2)}}} - 1\right] \tag{16.9}$$

satisfies the relation for all $i \in [1, N]$:

$$\mathbb{P}\left[\|\mathbf{u_i}\|^2 \geq x_\eta\right] \geq \frac{1}{3}e^{-\eta}. \tag{16.10}$$

By the assumption that $4 \leq d \leq \frac{n}{2}$, the threshold on the number η above satisfies:

$$\begin{aligned}
\frac{1}{2} \frac{1}{1 - \sqrt{\frac{2(d-2)}{(n-d-2)(n-2)}}} &\leq \frac{1}{2} \frac{1}{1 - \sqrt{\frac{2}{(n-2)}}}, \\
&\leq \frac{1}{2} \frac{1}{1 - \sqrt{\frac{1}{3}}}, \\
&\simeq 1.183.
\end{aligned}$$

By the i.i.d. property of the $(\|\mathbf{u_i}\|^2)_{i=1}^n$ sample, for any number η satisfying the bound above,

$$\mathbb{P}\left[\sup_{i\in[1,N]}\|\mathbf{u_i}\|^2 \geq x_\eta\right] \geq 1-\left(1-\frac{1}{3}e^{-\eta}\right)^N, \\ \geq 1-e^{-\frac{N}{3}e^{-\eta}}.$$

Finally, as

$$\frac{1}{3}\exp\left[-\frac{1}{2}\frac{1}{1-\sqrt{\frac{1}{3}}}\right] \simeq 0.102 \geq \frac{1}{10},$$

whenever the number t, as assumed, satisfies

$$t \leq \frac{N}{10},$$

choosing $\eta = \log\left(\frac{N}{3t}\right)$ in 16.9 ensures that

$$\eta \geq \log\frac{10}{3} \geq \frac{1}{2}\frac{1}{1-\sqrt{\frac{1}{3}}},$$

and the following relation holds by Inequality 16.10:

$$\mathbb{P}\left[\sup_{i\in[1,N]}\|\mathbf{u_i}\|^2 \geq x_t^\star\right] \geq 1-e^{-t}, \tag{16.11}$$

with

$$x_t^\star := \frac{d-2}{n-4} + \frac{1}{n-4}\sqrt{\frac{2(d-2)(n-d-2)}{n-2}}\left[\frac{\sqrt{2\log\left(\frac{N}{3t}\right)}}{1+\sqrt{2\log\left(\frac{N}{3t}\right)}\frac{d-2}{2(n-2)(n-d-2)}}-1\right]. \tag{16.12}$$

Concentration of $\|W\|^2$ We get from Lemma 14.10 that the χ_2 variable $\|W\|^2$ satisfies for $t>0$:

$$\mathbb{P}\left[\|W\|^2 \leq n\left(1-\sqrt{\frac{4t}{n}}\right)\right] \leq e^{-t}, \tag{16.13}$$

so that the quantity $\|W\|^2$ is strongly concentrated around its expectation n.

Conclusion By definition

$$\left\|\tilde{\Pi}(W)\right\|^2 = \|W\|^2 \sup_{i\in[1,N]}\|\mathbf{u_i}\|^2 \text{ a.s..} \tag{16.14}$$

As the sequence $\{\mathbf{u_i}\}_{i=1}^N$ is independent from the norm $\|W\|^2$, combining Inequalities 16.11 and 16.13 yields the the first statement in the corollary:

$$\mathbb{P}\left[\left\|\tilde{\Pi}(W)\right\|^2 \geq n\left(1-\sqrt{\frac{4t}{n}}\right)x_t^\star\right] \geq (1-e^{-t})^2, \tag{16.15}$$

valid under the conditions $4 \leq d \leq \frac{n}{2}$ and $t \leq \frac{N}{10}$.

Under the same conditions, with the same argument of independence, we draw from 16.14:

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{\Pi}(W)\right\|^2\right] &= \mathbb{E}\left[\|W\|^2\right]\mathbb{E}\left[\sup_{i\in[1,N]}\|\mathbf{u_i}\|^2\right], \\ &= \mathbb{E}\left[\sup_{i\in[1,N]}\|\mathbf{u_i}\|^2\right],\end{aligned}$$

leading by 16.11 to the second statement in the corollary:

$$\mathbb{E}\left[\left\|\tilde{\Pi}(W)\right\|^2\right] \geq \left(1-e^{-t}\right)nx_t^\star. \tag{16.16}$$

□

16.11 Proof of Lemma 14.20

Proof. The first statement in the lemma follows from a short calculation: for any $x > 0$

$$\begin{aligned}1-\Phi(x) &= \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{u^2}{2}}\,\mathrm{d}\,u, \\ &= e^{-\frac{x^2}{2}}\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-\frac{2xh+h^2}{2}}\,\mathrm{d}\,h, \\ &\geq e^{-\frac{x^2}{2}}\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-\frac{1+(1+x^2)h^2}{2}}\,\mathrm{d}\,h, \text{ by the inequality } 2xh \leq 1+x^2h^2 \\ &= \frac{e^{-\frac{1+x^2}{2}}}{2\sqrt{1+x^2}}.\end{aligned}$$

which is the first statement announced. This bound may be simplified by invoking the relation $\log\sqrt{1+x^2} \leq \log(1+x) \leq x$, so that:

$$\frac{e^{-\frac{1+x^2}{2}}}{2\sqrt{1+x^2}} \geq \frac{1}{2}e^{-\frac{1}{2}(x+1)^2}.$$

Conversely if $1-\Phi(x) = y$ with $y \leq \frac{1}{2}e^{-\frac{1}{2}}$, then $x \geq x_1$ where x_1 is the solution of

$$y = \frac{e^{-\frac{1+x_1^2}{2}}}{2\sqrt{1+x_1^2}},$$

which reformulates as

$$-2\log(2y) = (1+x_1^2) + \log\left(1+x_1^2\right),$$

leading to

$$1 + x_1^2 \leq -2\log(2y),$$

and

$$1 + x_1^2 \geq -2\log(2y) - \log\left(-2\log(2y)\right),$$

so that

$$\begin{aligned} x^2 &\geq -1 - 2\log(2y) - \log\left(-2\log(2y)\right), \\ &= -2\log\left(2e^{\frac{1}{2}}y\right) - \log\left(1 - 2\log\left(2e^{\frac{1}{2}}y\right)\right). \end{aligned}$$

□

16.12 Proof of Lemma 14.21

Proof. The gamma density is decreasing for any $z > a-1$. Then by its expression,

$$\begin{aligned} \mathbb{P}\left[Z_a \geq z\right] &= \frac{1}{\Gamma(a)} \int_z^{\infty} x^{a-1} e^{-x} \mathrm{d}x, \\ &\geq \frac{1}{\Gamma(a)} \int_z^{z+\sqrt{a-1}} x^{a-1} e^{-x} \mathrm{d}x, \\ &\geq \sqrt{a-1} \frac{\left(z+\sqrt{a-1}\right)^{a-1} e^{-z-\sqrt{a-1}}}{\Gamma(a)}, \\ &\geq \sqrt{a-1} \frac{\exp\left[-z-\sqrt{a-1} + (a-1)\log\left(z+\sqrt{a-1}\right)\right]}{\Gamma(a)}. \end{aligned}$$

We know by Stirling's formula that:

$$\Gamma(a) \leq e\sqrt{a-1}\left(\frac{a-1}{e}\right)^{a-1}.$$

so that combining the last two inequalities above yields:

$$\begin{aligned} \log\left(\mathbb{P}\left[Z_a \geq z\right]\right) &\geq -1 - z - \sqrt{a-1} + (a-1) + (a-1)\log\left(\frac{z+\sqrt{a-1}}{a-1}\right), \\ &= -1 - (a-1)g\left(\frac{z+\sqrt{a-1}-(a-1)}{a-1}\right). \end{aligned}$$

and

$$\mathbb{P}\left[Z_a \geq z\right] \geq \frac{1}{e}\exp\left[-(a-1)g\left(\frac{z+\sqrt{a-1}-(a-1)}{a-1}\right)\right].$$

which is the first announced statement. The second statement is a direct transposition, as a $\chi^2(b)$ random variable is distributed as $2Z_{\frac{b}{2}}$. □

17 Proofs for Section 15 Tools for sorted chi-square samples

17.1 Proof of Lemma 15.3

Proof. The independence of the random variable $Z_{n,l}$ and the random reordering permutation $\sigma_n(\cdot)$ results from the exchangeability of the unordered sample $\xi_1^2, \cdots, \xi_n^2$.

To show the next statement in the lemma we note that since by assumption $\frac{l(1+\theta)}{n} < e^{-\frac{1}{2}} < 1$ and $\theta > 2.06$, the lower bound in Lemma 15.2 ensures that apart from a set of probability bounded by $\left[\exp\left(\frac{\theta^2}{2(1+\theta)}\right) - 1\right]^{-1} < 1$,

$$\xi^2_{\sigma_n(n+1-j)} \geq \left[\Phi^{-1}\left(1 - \frac{j(1+\theta)}{2n}\right)\right]^2 \text{ for } 1 \leq j \leq l. \tag{17.1}$$

Lemma 14.20 offers the following bound on the inverse normal cumulative distribution function $\Phi^{-1}(\cdot)$: for any $y < \frac{1}{2}e^{-\frac{1}{2}} \approx 0.303$:

$$\left[\Phi^{-1}(1-y)\right]^2 \geq -2\log\left(2e^{\frac{1}{2}}y\right) - \log\left(1 - 2\log\left(2e^{\frac{1}{2}}y\right)\right). \tag{17.2}$$

By assumption , the integers l and n satisfy $2\log\frac{n}{l(1+\theta)} > 1$ so that Inequality 17.2 above applies to value of y set to $\frac{j(1+\theta))}{2n}$ for $1 \leq j \leq l$. Consequently, the following relation holds for $1 \leq j \leq n$:

$$\left[\Phi^{-1}\left(1 - \frac{j(1+\theta)}{2n}\right)\right]^2 \geq 2\log\left(\frac{e^{-\frac{1}{2}}n}{j(1+\theta)}\right) - \log\left(1 + 2\log\frac{e^{-\frac{1}{2}}n}{j(1+\theta)}\right).$$

and on on a set Ω of probability at least $1 - \left[\exp\left(\frac{\theta^2}{2(1+\theta)}\right) - 1\right]^{-1} > 0$, by combining with Inequality 17.1:

$$\sum_{j=1}^{l} \xi^2_{\sigma_n(n-l+j)} \geq \sum_{i=1}^{l}\left[2\log\left(\frac{e^{-\frac{1}{2}}n}{j(1+\theta)}\right) - \log\left(1 + 2\log\frac{e^{-\frac{1}{2}}n}{j(1+\theta)}\right)\right]$$

Note that the function defined for $x > 1$ by $x \mapsto f(x) = 2x - \log(1+2x)$ is increasing and convex. It follows that the following relation holds on the set Ω:

$$\sum_{j=1}^{l} \xi^2_{\sigma_n(n-l+j)} \geq l f(z),$$

where z denotes the mean

$$z = \frac{1}{l}\sum_{1=1}^{l} \log\frac{e^{-\frac{1}{2}}n}{j(1+\theta)} = \log\frac{e^{-\frac{1}{2}}n}{1+\theta} - \frac{1}{l}\log(l!).$$

Then plugging the classical bound [Rob55]:

$$\log l! \leq \left(l + \frac{1}{2}\right)\log l - l + 1 \text{ for } l \geq 1,$$

leads to

$$z \geq \log \frac{e^{\frac{1}{2}} n}{l(1+\theta)} - \frac{\log l + 2}{2l},$$

and on the set Ω,

$$\sum_{j=1}^{l} \xi^2_{\sigma_n(n-l+j)} \geq 2l \log \frac{e^{\frac{1}{2}} n}{l(1+\theta)} - \log l - 2 \\ - l \log \left[1 + 2 \log \frac{e^{\frac{1}{2}} n}{l(1+\theta)} - \frac{\log l + 2}{l} \right].$$

Reordering the terms leads to:

$$\sum_{j=1}^{l} \xi^2_{\sigma_n(n-l+j)} \geq l \left[1 + 2 \log \frac{n}{l(1+\theta)} \right] \\ - l \log \left[2 \log \frac{ne}{l(1+\theta)} - \frac{1}{l} \log l - \frac{2}{l} \right] \\ - \log l - 2.$$

and after neglecting the negative terms $-\frac{1}{l} \log l - \frac{2}{l}$,

$$\begin{aligned} \sum_{j=1}^{l} \xi^2_{\sigma_n(n-l+j)} &\geq l \left[1 + 2 \log \frac{n}{l(1+\theta)} \right] - l \log \left[2 \log \frac{e\, n}{l(1+\theta)} \right] - \log l - 2, \\ &= \left[1 - \frac{\log \left[2 \log \frac{e\, n}{l(1+\theta)} \right] + \frac{\log l + 2}{l}}{1 + 2 \log \frac{n}{l(1+\theta)}} \right] l \left[1 + 2 \log \frac{n}{l(1+\theta)} \right], \\ &= \left[1 - r(n,l,\theta) \right] l \left(1 + 2 \log \frac{n}{l(1+\theta)} \right) \end{aligned}$$

with

$$r(n,l,\theta) := \frac{\log \left[2 \log \frac{e\, n}{l(1+\theta)} \right] + \frac{\log l + 2}{l}}{1 + 2 \log \frac{n}{l(1+\theta)}},$$

which is the first result in the Lemma.

The fact that $r(n,l,\theta) = o(1)$ when $\frac{l(1+\theta)}{n} \to 0$ results from its expression, which also shows that the $r(\cdot)$ function is positive whenever $\frac{n}{l(1+\theta)} \geq e^{-\frac{1}{2}}$ and *a fortiori* within our assumptions.

Reformulating the expression of the $r(\cdot,\cdot,\cdot)$ function as:

$$r(n,l,\theta) := \frac{\log \left[1 + 2 \log \frac{e^{\frac{1}{2}} n}{l(1+\theta)} \right] + \frac{\log l + 2}{l}}{2 \log \frac{e^{\frac{1}{2}} n}{l(1+\theta)}}, \tag{17.3}$$

and using twice the relation for $x > 0$: $\log(1+x) \leq 2\left(\sqrt{1+x} - 1\right) = \frac{2x}{1+\sqrt{1+x}}$ leads to:

$$r(n,l,\theta) \leq \frac{2}{1 + \sqrt{1 + 2 \log \frac{e^{\frac{1}{2}} n}{l(1+\theta)}}} + \frac{2 l^{-\frac{1}{2}}}{2 \log \frac{e^{\frac{1}{2}} n}{l(1+\theta)}}. \tag{17.4}$$

By assumption $2\log\frac{n}{l(1+\theta)} \geq 1$ so that $2\log\frac{ne^{\frac{1}{2}}}{l(1+\theta)} \geq 2$, and it is straightforward to check that $1+\sqrt{1+x} \leq \frac{1+\sqrt{3}}{2}x \leq 1.5x$ for $x > 2$ so that finally we obtain the slightly more tractable inequality:

$$r(n,l,\theta) \leq \frac{2+1.5l^{-\frac{1}{2}}}{1+\sqrt{2\log\frac{en}{l(1+\theta)}}}.$$

□

17.2 Proof of Lemma 15.4

Proof. We simply use a version of the well known Gaussian tail upper bound by Cramér–Chernoff's method. Consider a real number λ to be chosen later and a standard Gaussian variable G. First recall that

$$\begin{aligned}\mathbb{E}\left[e^{\lambda|G|}\right] &\leq 2\,\mathbb{E}\left[e^{\lambda G}\right],\\ &= 2e^{\frac{\lambda^2}{2}}.\end{aligned}$$

Next, a simple enumeration show that

$$\begin{aligned}\mathbb{E}\left[e^{\lambda|\xi_1|}\,\middle|\,\xi_1^2 \geq \xi^2_{\sigma_n(n-l+1)}\right] &= \frac{1}{l}\sum_{i=n-l+1}^{n}\mathbb{E}\left[e^{\lambda\left|\xi_{\sigma_n(i)}\right|}\right],\\ &\leq \frac{1}{l}\sum_{i=1}^{n}\mathbb{E}\left[e^{\lambda\left|\xi_{\sigma_n(i)}\right|}\right],\\ &\leq \frac{2n}{l}e^{\frac{\lambda^2}{2}}.\end{aligned}$$

The first result in the lemma follows by Markov's inequality

$$\begin{aligned}\mathbb{P}\left[\xi_1^2 \geq x^2\,\middle|\,\xi_1^2 \geq \xi^2_{\sigma_n(n-l+1)}\right] &\leq e^{-\lambda|x|}\mathbb{E}\left[e^{\lambda|\xi_1|}\,\middle|\,\xi_1^2 \geq \xi^2_{\sigma_n(n-l+1)}\right]\\ &\leq \frac{2n}{l}e^{-\lambda|x|+\frac{\lambda^2}{2}},\end{aligned}$$

and after optimizing with λ set to $|x|$, for $t > 0$:

$$\mathbb{P}\left[\xi_1^2 \geq 2\log\left(\frac{2n}{l}\right)+2t\,\middle|\,\xi_1^2 \geq \xi^2_{\sigma_n(n-l+1)}\right] \leq e^{-t}$$

The inequality in expectation follows by integration with:

$$\begin{aligned}\mathbb{E}\left[\xi_1^2\,\middle|\,\xi_1^2 \geq \xi^2_{\sigma_n(n-l+1)}\right] &\leq 2\log\left(\frac{2n}{l}\right)+2\int_0^\infty e^{-t}\,\mathrm{d}\,t\\ &= 2\log\left(\frac{2n}{l}\right)+2.\end{aligned}$$

The last statement in the lemma simply follows from the fact that the expectation of the sum $\sum_{i=1}^{n}\xi^2_{\sigma_n(i)}$ is equal to n. □

Part IV

Appendix

Summary

A Assessing the complexity of the set of hyperrectangle partitions in a discrete hyperrectangle **205**

B Complexity of recursive segmentation families **208**

B.1 Introduction 208

B.2 Multiple segmentation of an integer segment 208

B.3 Regression trees 210

B.3.1 Segmentation rules for the examples in the present work . . . 210

B.4 Enumerating the models 211

B.4.1 Tree structure 211

B.4.2 Generating function 212

B.4.3 Enumerating tree structures 213

B.4.4 Enumeration results for segmentation of a rectangle in $\mathbb{N}^2$ . . 216

B.4.5 Enumerating regular binary segmentation trees 217

B.4.6 Enumerating regular segmentation quad-trees 217

B.4.7 Enumerating free binary segmentation trees 218

B.4.8 Enumerating free segmentation quad-trees 218

B.4.9 Summary 219

A Assessing the complexity of the set of hyperrectangle partitions in a discrete hyperrectangle

Most of the model families encountered in this work are based on segmentation schemes of a particular underlying space (see Section 7.1), and as such indexed by a set of partitions the same underlying space. Each linear model is composed of step functions constant over the elements of the corresponding partition. In most cases these partitions are composed of hyper rectangles such as segments in line segmentation, rectangles in $2d$-image segmentation or higher dimensional hyperrectangles for regression trees. In this context, the richest model family for partitioning a particular hyperrectangle underlying space in $\mathbb{N}^d$ corresponds to the set of all unconstrained hyperrectangle partitions, also called hyperrectangle layouts or bloc layouts in the literature. The question of the complexity of this particular model family arises naturally.

The combinatorics of rectangle layouts has recently attracted attention, in particular for the optimisation of micro-electronic circuitry, but also for their intrinsic interest. For instance in [ABP04] Ackerman and al. show that the number of floor plan structures with D non-intersecting segments is equal to the number of partitions of a specific type (Baxter partitions) . In [CGHK78] Chung and al. show that this number is $\frac{32\,2^{3D}}{D^4\pi\sqrt{3}}\left[1-\frac{22}{3D}+O(D^{-2})\right]$. Making the guess that relaxing the condition "with no intersecting segments" would not significantly inflate this estimate, this suggests that in dimension 2, the number of partitions of a discrete rectangle into smaller ones may have an upper bound of the form $c8^d\binom{n-1}{d-1}$ where c is a constant, based on the guess that a such a partition is mostly determined by its structure (the floor plan) and the sequence of the cardinals of its elements.

The following provides a rough complexity assessment when the underlying space is a discrete hyperrectangle in $\mathbb{N}^d$. As its proof relies on a very basic arithmetic argument, we guess it could be largely improved towards a value of the binomial complexity rate close to one in any dimension.

For the present work we denote *hyperrectangle* any product of intervals in $\mathbb{N}^d$, in other words any non-empty set of the form $(i_1,\cdots,j_1)\times\cdots\times(i_d,\cdots,j_d)$ in $\mathbb{N}^d$ with $i_1\leq j_1,\cdots,i_d\leq j_d$.

Recall that Definition 8.4 introduces the binomial complexity rate of a model family $\mathcal{M}$ as the smallest positive number $\mathrm{B}_{\mathcal{M}}$ with:

$$\sum_{m\in\mathcal{M},|m|>0} 2^{-|m|}\left(\frac{en}{|m|}\right)^{-\mathrm{B}_{\mathcal{M}}|m|}\leq 1.$$

Proposition A.1 (Bloc layouts)**.** *Consider an hyperrectangle $U\subset\mathbb{N}^d$ of cardinal n. For any integer D with $1\leq D\leq n$, denote N_D the number of partitions of U into D hyperrectangles. If $D=1$ then $N_1=1$ and if $D\geq 2$,*

$$N_D\leq\binom{n-1}{D-1}\left(\frac{n-1}{D-1}\right)^{\frac{\log d}{\log 2}(D-1)},$$

and

$$\log N_D\leq\frac{\log 2d}{\log 2}(D-1)\log\left(e\frac{n-1}{D-1}\right).$$

The binomial complexity rate of the set of the family of all partitions of U into hyperrectangle is less than $\frac{\log(2d)}{\log 2}$.

Proof. We call shape of a given hyperrectangle $(i_1, \cdots, j_1) \times \cdots \times (i_d, j_d) \subset \mathbb{N}^d$ the sequence of positive integers $j_1 - i_1 + 1, \cdots, j_d - i_d + 1$, and we call origin its first element in the lexicographic order of $\mathbb{N}^d$, namely $(i_1, \cdot, i_d)$.

We first establish that any partition of U in D hyperrectangles is entirely described by the sequence of the shapes of its $D-1$ first elements in lexicographic order of the origins. This statement is visually obvious in one dimension but if needed we detail it in the general case. It may informally be stated as the following "building blocks rule": "always place the origin of the next block at the first empty position in lexicographic order ". To show that only the sequence of shapes up to $D-1$ is needed to identify the entire partition, and that the positions of the origins are not needed, we consider such an ordered partition $U = r_1 \cup r_2 \cup \cdots \cup r_D$, suppose known the corresponding sequence of shapes and proceed by recurrence:

- The origin of the first hyperrectangle r_1 is the first point in U in the lexicographic order, namely the origin of U. Knowing its shape and its origin, r_1 is entirely identified.

- Assume the locations of $r_1, \cdots, r_k$ are known. Consider c the first point of $U \setminus r_1 \setminus \cdots \setminus r_k$ in the lexicographic order, and i the integer in $[k+1, D]$ such that $c \in r_i$. By definition, c is the origin of the set $r_{k+1} \cup \cdots \cup r_D$, and since c belongs to r_i, it is also the origin of r_i and the least of the origins of $r_{k+1} \cdots r_D$ in the lexicographic order. As the sequence $r_{k+1} \cdots r_D$ is ordered by lexicographic order of the origins, this means that $r_i = r_{k+1}$ and that c is the origin of r_{k+1}. Knowing the origin of r_{k+1} and its shape is enough to entirely identify r_{k+1}.

- finally, $r_D = U \setminus r_1 \setminus \cdots \setminus r_{D-1}$.

For any integer m, denote $H_d(m)$ the number of hyperrectangles of cardinal m with origin $(0, \cdots, 0)$ in $\mathbb{N}^d$. The number $H_d(m)$ is also the number of possible shapes for an hyperrectangle of cardinal m. Enumerating all the possible sequences formed of $D-1$ shapes whose cardinals sum to less than $n-1$ leads to:

$$N_D \leq \sum_{\substack{m_1+\cdots+m_D=n \\ m_1>0,\cdots,m_D>0}} \prod_{i=1}^{D-1} H_d(m_i).$$

Assume there is a number $\alpha_d > 0$ to be specified later so that $H_d(m) \leq m^{\alpha_d}$ for any integer $0 < m \leq n$. If $D = 1$ then simply $N_D = N_1 = 1$, for the singleton sequence

$(1, \cdots, 1)$. Otherwise, by concavity of the logarithm,

$$\begin{aligned}
N_D &\leq \sum_{\substack{m_1+\cdots+m_D=n \\ m_1>0,\cdots,m_D>0}} \prod_{i=1}^{D-1} H_d(m_i), \\
&\leq \sum_{\substack{m_1+\cdots+m_D=n \\ m_1>0,\cdots,m_D>0}} \exp\left[\sum_{i=1}^{D-1} \alpha_d \log m_i\right], \\
&\leq \sum_{\substack{m_1+\cdots+m_D=n \\ m_1>0,\cdots,m_D>0}} \exp\left[(D-1)\alpha_d \log \sum_{i=1}^{D-1} \frac{m_i}{D-1}\right], \\
&= \sum_{\substack{m_1+\cdots+m_D=n \\ m_1>0,\cdots,m_D>0}} \exp\left[(D-1)\alpha_d \log \frac{n-m_D}{D-1}\right], \\
&\leq \exp\left[(D-1)\alpha_d \log \frac{n-1}{D-1}\right] \sum_{\substack{m_1+\cdots+m_D=n \\ m_1>0,\cdots,m_D>0}} 1, \\
&= \left(\frac{n-1}{D-1}\right)^{\alpha_d(D-1)} \binom{n-1}{D-1}. \qquad \text{(A.1)}
\end{aligned}$$

To conclude we need to identify a suitable number α_d satisfying for any integer m

$$H_d(m) \leq m^{\alpha_d}. \qquad \text{(A.2)}$$

Assume the decomposition of a given integer m in distinct prime factors is $m = \prod_{i=1}^{k} p_i^{r_i}$. Denote r the number $r = \sum_{i=1}^{k} r_i$. Then the number of decompositions of m in a product of d factors is less than d^r, with equality when m has only simple prime factors. In addition,

$$m = \prod_{i=1}^{k} p_i^{r_i} \geq \prod_{i=1}^{k} 2^{r_i} = 2^r,$$

so that the following simple bound is enough for our need:

$$\begin{aligned}
H_d(m) &\leq d^r, \\
&\leq d^{\frac{\log m}{\log 2}}, \\
&= m^{\frac{\log d}{\log 2}}, \qquad \text{(A.3)}
\end{aligned}$$

and Inequality A.2 holds with $\alpha_d = \frac{\log d}{\log 2}$ so that the first statement in the lemma follows from Inequalities A.1 and A.3:

$$N_D \leq \binom{n-1}{D-1} \left(\frac{n-1}{D-1}\right)^{\frac{\log d}{\log 2}(D-1)},$$

By the inequality (see [Mas07, Proposition 2.5 p. 20])

$$\log \binom{n-1}{D-1} \leq (D-1) \log \left(e \frac{n-1}{D-1}\right),$$

it also follows that

$$\begin{aligned}\log N_D &\leq (D-1)\log\left(e\frac{n-1}{D-1}\right)+\frac{\log d}{\log 2}(D-1)\log\left(\frac{n-1}{D-1}\right),\\ &\leq \frac{\log 2d}{\log 2}(D-1)\log\left(e\frac{n-1}{D-1}\right),\end{aligned}$$

which is the second statement in the lemma. Since $\binom{n-1}{D-1} \leq \binom{n}{D}$ and $d \geq 1$, it follows also that for $D > 0$:

$$\log N_D \leq \frac{\log 2d}{\log 2} D \log\left(e\frac{n}{D}\right).$$

By the inequality $\mathrm{B}_{\mathcal{M}} \leq \sup\left\{\frac{\log N_D}{D\log\left(e\frac{n}{D}\right)}\right\}_{D\leq 1\leq n}$ (see Definition 8.4 and its following remark),

$$\mathrm{B}_{\mathcal{M}} \leq \frac{\log 2d}{\log 2}.$$

□

B Complexity of recursive segmentation families

B.1 Introduction

In the following we explore different choices for the model family $\mathcal{M}$, when it arises from recursive segmentation of a linear or rectangular discrete domain. Our aim is to construct families $(L_m)_{m\in\mathcal{M}}$ with $\sum_{m\in\mathcal{M}} e^{-|m|L_m} < \Sigma < +\infty$, as required by the assumptions of theorem 5.6, or with the same goal, to assess the binomial complexity rate $\mathrm{B}_{\mathcal{M}}$ as introduced in Definition 8.4.

Although we claim no novelty in the facts exposed, we believe gathering them has an interest *per se*, and for the description of numerical experiments shown in part 8.4. As each of the segmentation families we study can be indexed by families of trees, we will rely on known elementary methods to enumerate tree structures.

As indicated in Part 8.4, S. Gey and É. Nedelec offer a thorough analysis of the performance of CART regression trees in [GN05]. A extensive account of analytic combinatorics may be found in [FS09], and precise asymptotic results on the number of binary trees of prescribed height in [FGOR92].

B.2 Multiple segmentation of an integer segment

Consider a positive integer n and an integer interval $[1, n]$, the *underlying space*. In the setup of segmentation of a sequence, the model family $\mathcal{M}$ is indexed by the partitions of the underlying space in segments. This setup was studied by É. Lebarbier [LN07].

Lebarbier's proposal Following [Mas07, p. 93-94] we noted in the remarks following the introduction of the binomial complexity rate $\mathrm{B}_{\mathcal{M}}$ in Definition 8.4 that choosing a segmentation of $[1, n]$ amounts to choose a subset of $[2, n]$ to mark the breakpoints of the segmentation. It follows that for $1 \leq d \leq n$, the segmentations of the underlying space into d segments are in number $\binom{n-1}{d-1} \leq \binom{n}{d}$. By the classical inequality $\binom{n}{d} \leq \left(\frac{en}{d}\right)^d$ (see [Mas07, p. 20]) for $1 \leq d \leq n$, it follows that the weights

$$L_m := \log 2 + \log\left(\frac{en}{|m|}\right),\ m \in \mathcal{M}$$

satisfy:

$$\sum_{m \in \mathcal{M}} e^{-|m|L_m} \leq \sum_{d=1}^{n} 2^{-d} < 1,$$

so that in this case the binomial complexity rate satisfies $\mathrm{B}_{\mathcal{M}} \leq 1$.

Linear weights We may observe that given a positive number d_0, the binomial relation leads to:

$$\begin{aligned}\sum_{m \in \mathcal{M}} \left(\frac{d_0}{n-1}\right)^{|m|} &= \sum_{d=1}^{n} \binom{n-1}{d-1} \left(\frac{d_0}{n-1}\right)^d, \\ &= \frac{d_0}{n-1} 1 + \frac{d_0}{n-1}^{(n-1)}, \\ &\leq \frac{d_0}{n-1} e^{d_0},\end{aligned}$$

so that the weight $L := \log \frac{n-1}{d_0}$ satisfies

$$\sum_{m \in \mathcal{M}} e^{-|m|L} \leq \frac{d_0}{n-1} e^{d_0},$$

with a straightforward comparison with the choice in the paragraph above. The right side of the bound above is less than 1 if $d_0 \leq \log\left(\frac{n-1}{\log(n-1)}\right)$. Though this weight structure is considered less refined than the one described in the preceding section (see [Mas07, p.92]), it may be suited in situations where a selected model of large dimension is acceptable, as the weight profile is comparatively skewed in favor of large dimensions.

Relation with a renewal process To illustrate the analogy between the set of weights above and a tree structure, we may observe that this set of weights borrows its structure from a renewal process of constant survival rate $1 - p = \frac{n-1}{n+d_0-1}$ (so that $\frac{p}{1-p} = e^{-L} = \frac{d_0}{n-1}$) on the underlying space $[1, n]$. Defining for any integer segment τ

$$w_\tau := \begin{cases} (1-p)^{(|\tau|-1)} & \text{if } 1 \in \tau \\ p(1-p)^{(|\tau|-1)} & \text{otherwise} \end{cases}$$

we observe classically that

$$\begin{aligned}\sum_{m \in \mathcal{M}} \prod_{\tau \in m} w_\tau &= \sum_{d=1}^{n} \binom{n-1}{|m|-1} p^{|m|-1} (1-p)^{n-|m|} \\ &= 1\end{aligned}$$

The only difference with the situation above is the normalisation of the sum. Any other renewal process on the integer segment $[1, n]$ would induce a set of weights by the same procedure. From an algorithmic point a view, combined with a linearization of the weight in theorem 5.6 by the relation $a + 2b\sqrt{L_m} + cL_m \leq a + b\eta^{-1} + (c + b\eta)L_m$ for positive a, b, c and η, this presents the advantage of expressing a variety of possible weights as sums on the segments of the models, allowing manipulations at the segment level regardless of the global structure of the model. In a similar context, Fearnhead and Liu took advantage of this relation to build fast online Bayesian methods for multiple changepoints problems [FL07].

B.3 Regression trees

We turn now to the situation of multi-dimensional regression trees, where the underlying space (feature space) is a rectangular domain of $\mathbb{N}^d$, in general $\mathbb{N}^2$, and the family of models is generated by an iterative segmentation rule. The underlying space is sequentially subdivided in rectangular parts, e.g. by splits across one coordinate, resulting in a family of partitions of the underlying space in rectangular parts. The linear model corresponding to such a partition is the linear span of the indicator functions of the parts (elements) of the partition. In other words the estimate is a step function constant over each part of the partition, and its value over each part is the mean of the observed signal over the same part (see Section 7 for regressogram estimators and Section 9.1 for histogram estimators).

The complexity of the model family only depends on the dimensions of the underlying space and on the variety of subdivisions allowed by the segmentation rule.

B.3.1 Segmentation rules for the examples in the present work

Instances of the following appear in the present work: *regular binary tree*, *free binary tree*, *free quad-tree* and *regular quad-tree*, described as follows:

Free binary tree: a part may be split in two non-empty rectangular parts by a split across one coordinate direction, with free choice of the split location. This is the segmentation rule of usual regression trees.

Regular binary tree a part may be split in two non-empty rectangular parts by a split along either one of the coordinate, at a predefined (near)median location . As an example the region $\begin{array}{cccc} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{array}$ may only be split as $\begin{array}{cccc} 1 & 2 & 3 & 4 \\ \hline 5 & 6 & 7 & 8 \end{array}$ or as $\begin{array}{cc|cc} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{array}$.

A small complication arises when the dimensions of the underlying space are not powers of 2 (as customary in wavelets analysis): a convention is needed to divide parts of odd height or width. We choose to allow split immediately below the median, so that a side of length $2n + 1$ with $n \geq 1$ may be split into n and $n + 1$, in this order.

Quad-tree: a part of width larger than 1 and height larger than 1 may be divided by a cross-split at any interior point. A part of width one or height one (a segment) may be split into any pair of contiguous segments. As an example, the region $\begin{array}{ccc}1&2&3\\4&5&6\end{array}$ may only be split as $\begin{array}{c|cc}1&2&3\\\hline4&5&6\end{array}$ or $\begin{array}{cc|c}1&2&3\\\hline4&5&6\end{array}$.

The region $\begin{array}{cccc}1&2&3&4\end{array}$ may be split as $\begin{array}{c|ccc}1&2&3&4\end{array}$ or $\begin{array}{cc|cc}1&2&3&4\end{array}$ or $\begin{array}{ccc|c}1&2&3&4\end{array}$.

A slight complication arises from the fact that in general the branching factor of the trees considered is not constant, due to the limitations on splitting thin or singleton parts. In this respect the rule for splitting segment parts may seem artificial, we retained it order to preserve the ability of the segmentation scheme to isolate individual points, which is at the center of the minimal penalty results presented in this work.

Regular quad-tree: a part may be divided in a single manner by a cross-split at a predefined central position. As an example, the region $\begin{array}{cc}1&2\\3&4\end{array}$ may only be split as $\begin{array}{c|c}1&2\\\hline3&4\end{array}$. The rule for splitting segment parts is the same than for regular binary trees

This procedure is parent to wavelet procedures (see [DJ94]) with informally the additional geometric constraint that if a wavelet is retained in a model, all of its parent and sibling wavelets must also be present, to preserve the tree structure of the associated partition. As with free quad-trees, a slight complication arises from the fact that the splits in thin parts may be two-fold as opposed to four-fold, preventing the branching factor to be constant.

B.4 Enumerating the models

The following is a combination of known facts. We refer to the book by Philippe Flajolet and Robert Sedgewick for combinatorial methods and results on trees [FS09, I.5 p 65].

In this section the underlying space is always a hyper-rectangle in $\mathbb{N}^d$ and when $d = 2$ we denote by W and H its width and its height. Accordingly the number of points (observations) in the underlying space satisfies $n = WH$.

B.4.1 Tree structure

We begin with a simple remark stating that any of the segmentation trees discussed in Section B.3.1 can be described by at least one triplet composed of:

1. a rooted plane tree representing the inclusion relation between parts
2. the list of the orientations of the corresponding splits
3. the list of cardinals of the corresponding parts in $\mathbb{N}^d$

In certain cases, some elements of this description may be redundant: for instance the orientations of the splits for quad-trees in $\mathbb{N}^2$, or the size of the parts for regular binary segmentation trees. We call the first two element of this triplet a *tree structure* for the corresponding regression tree. The following table lists the types of tree-structures and the segmentation rules associated with the four model classes in discussion (w and h stand for the width and height of a given split part):

Model family	branching factor (thin part)	split orientation	number of allowed splits	Description
regular binary	2	horizontal or vertical	2 or 0	e.g. central
free binary	2	horizontal or vertical	$w+h-2$	split starts inside top or left frontier
regular quad	4 (2)	cross split	1 or 0	single choice, e.g. central
free quad	4 (2)	cross split	$(w-1)(h-1)$ if $w>1$ and $h>1$, $w+h-2$ otherwise	inner points

B.4.2 Generating function

For any model family $\mathcal{M}$, we denote by $G_{\mathcal{M}}(t)$ the *generating function* $G_{\mathcal{M}}(t) := \sum_{m\in\mathcal{M}} t^{|m|}$. As we only deal with finite non-empty underlying spaces, $G_{\mathcal{M}}(t)$ is always a convex polynomial in t, with $G_{\mathcal{M}}(0)=0$ and $G'_{\mathcal{M}}(0)=1$. We denote $\left[t^d\right] G_{\mathcal{M}}(t)$ the coefficient of t^d in $G_{\mathcal{M}}(t)$ for any integer $d\geq 0$, which is also the number of models of dimension d. Recall that the dimension $|m|$ of a model m is also the number of leaves of the corresponding segmentation tree. We denote by $G(T)$ the corresponding formal generating series.

The following product relation allows to combine segmentations on disjoint parts of an underlying space (see [FS09, p. 27], Cartesian product):

Lemma B.1 (Cartesian product of model families)**.** *If a model set $\mathcal{M}_1$ can be decomposed as $\mathcal{M}_1 = \left\{ S_{m_2} \oplus^{\perp} S_{m_3} \right\}_{(m_2,m3)\in\mathcal{M}_2\times\mathcal{M}_3}$, then the following relation holds:*

$$G_{\mathcal{M}_1}(t) = G_{\mathcal{M}_2}(t) G_{\mathcal{M}_3}(t). \tag{B.1}$$

Proof. For any pair of models $(m_2, m3) \in \mathcal{M}_2 \times \mathcal{M}_3$, the dimension of $S_{m_2} \oplus^{\perp} S_{m_3}$ is $|m_2| + |m_3|$. The result follows by the relation

$$\begin{aligned} G_{\mathcal{M}_1}(t) &= \sum_{(m_2,m3)\in\mathcal{M}_2\times\mathcal{M}_3} t^{|m_2|+|m_3|}, \\ &= \left[\sum_{m_2\in\mathcal{M}_2} t^{|m_2|}\right]\left[\sum_{m_3\in\mathcal{M}_3} t^{|m_3|}\right], \\ &= G_{\mathcal{M}_2}(t) G_{\mathcal{M}_3}(t). \end{aligned}$$

□

B.4.3 Enumerating tree structures

The following lemmas we help us in assessing the complexity of various tree families encountered with recursive segmentation schemes. Our goal is to enumerate tree structures. Recall that we call tree structure of a segmentation tree in $\mathbb{N}^d$ the associated rooted planar tree with the record of the orientation of the splits, independently of the actual size of the parts and locations of the splits. A segmentation tree is entirely described by at least one tree structure and the cardinals of its parts.

b-ary tree structures

Lemma B.2 (Counting rooted plane b-ary tree structures with tags)**.** *Consider a real polynomial function defined by*

$$t \mapsto G(t) = \sum_{D \in \mathbb{N}} g_D t^D$$

with non negative coefficients $\{g_D\}_{D\in\mathbb{N}}$*. Assume that* $G(0) = g_0 = 0$ *and that there are numbers* $a \geq 1$ *and* $b > 1$ *so that for any* $t > 0$ *the following inequality holds:*

$$G(t) - aG(t)^b \leq t.$$

Then the number

$$\tilde{t} = \left(\frac{1}{ab}\right)^{\frac{1}{b-1}} \left(1 - \frac{1}{b}\right)$$

satisfies:

$$G(\tilde{t}) \leq \left(\frac{1}{ab}\right)^{\frac{1}{b-1}} < 1.$$

For any integer D*, the coefficient* g_D *satisfies:*

$$g_D \leq \left(\frac{1}{ab}\right)^{\frac{1}{b-1}} \left[(ab)^{\frac{1}{b-1}} \frac{b}{b-1}\right]^D.$$

In particular, the preceding bound applies to the number of rooted plane b*-ary tree structures with tags in* $[1, \ldots, a]$ *on the inner nodes.*

If the function $G(\cdot)$ *is the generating function of a model family* $\mathcal{M}$ *in an Euclidean space of dimension* n*, then the binomial complexity rate of this model family satisfies:*

$$\mathrm{B}_{\mathcal{M}} \leq \log\left(\frac{1}{2\tilde{t}}\right)^+ = \left[\frac{\log(ab)}{b-1} + \log\left(\frac{b}{b-1}\right) - \log(2)\right]^+.$$

Proof. Our main reference is [FS09, p. 67-68]. If all the coefficients of the function $G(\cdot)$ are null, the lemma is trivial. Otherwise the polynomial function $t \mapsto G(t)$ is positive, continuous and increasing for $t \geq 0$, going to infinity with t. Denote by $t^\star$ the unique positive real number satisfying $G_{\mathcal{M}}(t)^{b-1} = \frac{1}{ab}$. The inequality in assumption implies

that

$$\begin{aligned} t^\star &\geq G_{\mathcal{M}}(t^\star) - aG_{\mathcal{M}}(t^\star)^b, \\ &= \left(\frac{1}{ab}\right)^{\frac{1}{b-1}} - a\left(\frac{1}{ab}\right)^{\frac{b}{b-1}}, \\ &= \left(\frac{1}{ab}\right)^{\frac{1}{b-1}}\left(1-\frac{1}{b}\right), \\ &= \tilde{t}. \end{aligned}$$

As the function $G(t)$ is increasing for $t \geq 0$, if follows that $G(\tilde{t}) \leq G(t^\star) = \left(\frac{1}{ab}\right)^{\frac{1}{b-1}}$. Since $a \geq 1$ and $b > 1$, the number $G(\tilde{t})$ is less than 1. This is the first announced result. Replacing the value $G(\tilde{t})$ by its development yields

$$G(\tilde{t}) = \sum_{D\in\mathbb{N}} g_D \tilde{t}^D.$$

As all the coefficients are non-negative, for any $D \in \mathbb{N}$,

$$g_D \leq \tilde{t}^{-D} G(\tilde{t}),$$

leading to the second result in the lemma.

If the function $G(\cdot)$ is the generating function of a model family $\mathcal{M}$ in an Euclidean space of dimension n, there are only models of dimension less than n so that

$$G(t) = \sum_{1\leq D\leq n} g_D t^D.$$

Since $G(\tilde{t}) = (ab)^{-\frac{1}{b-1}} \leq 1$, it follows that

$$\begin{aligned} 1 &\geq \sum_{1\leq D\leq n} g_D \tilde{t}^D, \\ &= \sum_{1\leq D\leq n} g_D 2^{-D} e^{-D\log\left(\frac{1}{2\tilde{t}}\right)}, \\ &\geq \sum_{1\leq D\leq n} g_D 2^{-D} e^{-D\log\left(\frac{1}{2\tilde{t}}\right)^+}, \\ &\geq \sum_{1\leq D\leq n} g_D 2^{-D} \left[\frac{en}{D}\right]^{-D\log\left(\frac{1}{2\tilde{t}}\right)^+}, \end{aligned}$$

so that in accordance with Definition 8.4 (binomial complexity rate),

$$\mathrm{B}_{\mathcal{M}} \leq \log\left(\frac{1}{2\tilde{t}}\right)^+ = \left[\frac{\log(ab)}{b-1} + \log\left(\frac{b}{b-1}\right) - \log(2)\right]^+.$$

To apply what precedes to rooted plane b-ary tree structures with tags valued in $[1,\ldots,a]$ on inner nodes, observe that such a tree with D leaves is either the root node alone without tag, or a composition of b trees of the same family but with less leaves, attached to the root node which is now an inner node and carries a tag with a possible values. It follows that the corresponding generating function satisfies the relation:

$$G(t) \leq t + aG(t)^b, \forall t > 0$$

so that the preceding results apply. □

Regression tree structures and phylogenetic trees The segmentation rule of standard regression trees proceeds by recursive bi-partition along coordinate levels. It it straightforward to describe the corresponding tree structures by rooted planar binary trees, with a tag on each inner node representing the coordinate used to define the split. However this representation tends to be redundant, as there are many ways to represent a sequence of parallel splits by by-partition. A slightly more efficient representation is obtained by treating successive split accross the same coordinate as one single multiple split, for instance:

$$Cell_a : (Cell_{aaa}, Cell_{aab}, Cell_{aba}, Cell_{abb})$$

instead of

$$Cell_a : (Cell_{aa} : (Cell_{aaa}, Cell_{aab}), Cell_{ab} : (Cell_{aba}, Cell_{abb})).$$

In the resulting tree structure, the degree of a node can be any integer not less than 2. This tree structure is known in the literature as *phylogenetic* (see [FS09, II.V p 69]). To apply it to regression trees, we make the additional assumption than if a node is split across a given coordinate, then its direct descendants may not be split across the same coordinate. So with d coordinates (in $\mathbb{N}^d$), the root node accepts d possible split orientations but any other node only $d-1$.

Lemma B.3 (Counting regression trees structure)**.** *Consider $t \mapsto G(t)$ the generating function of the regression tree structures with $d \geq 2$ variables. Then for the number*

$$t^\star = \frac{\left(\sqrt{d} - \sqrt{d-1}\right)^2}{d-1} = (d-1)^{-1}\left(\sqrt{d} + \sqrt{d-1}\right)^{-2},$$

the following holds:

$$G(t^\star) < 1.$$

Proof of Lemma B.3. If the formal generating series of a class of objects $\mathcal{C}$ is $t \mapsto H_{\mathcal{C}}(T)$, then by the 'product Lemma' B.1 the generating function of the set of sequences of two or more disjoint copies of these objects is (see [FS09, p.27]):

$$\begin{aligned} G_{seq(\mathcal{C})}(T) &= H^2_{\mathcal{C}}(T) + H^3_{\mathcal{C}}(T) + \cdots, \\ &= \frac{H^2_{\mathcal{C}}(T)}{1 - H_{\mathcal{C}}(T)}. \end{aligned} \tag{B.2}$$

Consider $G(T)$ the formal generating function of the set of regression tree structures with d variables, in $\mathbb{N}^d$. The tree structure representing the un-split root node contributes to the generating function by the term T, as the corresponding dimension is one. In the other tree structures, the first sequence of splits exists, and in each of the resulting parts the subsequent splits are represented by a subtree rooted at this same part. This subtrees may have the same structure than the root tree, but the orientation of their initial splits must differ from the root one, so that only that $d-1$ possible orientation for the first sequence of splits are allowed. As a result the generating function of each non-root subtree is

$$G_1(T) = T + \frac{d-1}{d}\left(G(T) - T\right) = \frac{T + (d-1)G(T)}{d}.$$

By the 'sequence rule' in Equation B.2 the generating function for one whole sequence of sub-trees of the root node is then

$$\frac{G_1(T)^2}{1-G_1(T)}.$$

As the initial split has d posible orientations, $G(T)$ satisfies:

$$\begin{aligned} G(T) &= T + d\frac{G_1(T)^2}{1-G_1(T)}, && \text{(B.3)}\\ &= T + d\frac{\left[\frac{T}{d} + \frac{d-1}{d}G(T)\right]^2}{1-\frac{T}{d}-\frac{d-1}{d}G(T)}, \\ &= T + \frac{[T+(d-1)G(T)]^2}{d-T-(d-1)G(T)}. && \text{(B.4)} \end{aligned}$$

If $d=1$, then $G(T)=\frac{T}{1-T}$, which is the formal generating function of simple non-empty sequences. If $d>1$, as the generating function has no negative coefficients, only one root of Equation B.4 may represent it, and still in the sense or formal series,

$$G(T) = \frac{1-T-\sqrt{1-2(2d-1)T+T^2}}{2(d-1)}. \qquad \text{(B.5)}$$

Note that whith $d=2$, in any of the tree structures at hand, the sequence of the orientations of the splits along any path is alternating, and carries no information expect for the first one which has two possible orientations. We recoup the generating function for planar rooted trees where the degree of any inner node is not less than two (phylogenetic trees), counted by their number of leaves: $T+\frac{1}{2}(G(T)-T) = \frac{1+T-\sqrt{1-6t+T^2}}{4}$ [FS09, II.V p 69].

The lowest root of the polynomial expression under the radical in Equation B.5 is the radius of convergence of the formal series $G(T)$:

$$\begin{aligned} t^\star &= \frac{2d-1-2\sqrt{d(d-1)}}{d-1}, && \text{(B.6)}\\ &= (d-1)^{-1}\left(\sqrt{d}+\sqrt{d-1}\right)^{-2}. \end{aligned}$$

We recoup the value known for phylogenetic trees: $(\sqrt{2}+1)^{-2} = 3-2\sqrt{2} \simeq 0.172 \simeq \frac{1}{5.83}$ (as infered from [FS09, p 69 and 475]). This compares favorably with the value obtained by representing the same regression trees by recursive bi-partition, as in Lemma B.2: $\left(\frac{1}{ab}\right)^{\frac{1}{b-1}}\left(1-\frac{1}{b}\right) = \frac{1}{8}$ with $a=2$ and $b=2$.

In view of Equations B.5 and B.6, the function $t \mapsto G(t)$ is defined and finite for $0 \le t \le t^\star$, and $G(t^\star) \le \frac{1}{2(d-1)} < 1$ holds. □

B.4.4 Enumeration results for segmentation of a rectangle in $\mathbb{N}^2$

The following lemmas offer bounds on the complexity of the model families considered above.

B.4.5 Enumerating regular binary segmentation trees

Lemma B.4 (Enumerating regular binary segmentation trees)**.** *With the segmentation rule set out in Section B.3.1 for regular binary trees in* $\mathbb{N}^2$*, for any integer* $D > 0$*, the number of models of dimension* D *is less than* 2^{3D-2}*. The corresponding binomial complexity rate is less than* $2\log 2 \leq 1.39$*.*

Proof. As discussed in SectionB.4.1, the model family $\mathcal{M}$ is described by a subset of the set of rooted plane trees binary with inner nodes labeled by an element of the two-set $\{h, v\}$. Such a labelled binary tree structure effectively represents a segmentation model if it does not require to split any segment part along its lowest dimension. As a result the generating function is increasing with respect to the inclusion relation of the underlying spaces.

From the preceding observations we infer that the generating function satisfies:

$$G_{\mathcal{M}}(t) \leq t + 2G_{\mathcal{M}}(t)^2, \tag{B.7}$$

expressing the fact that:

- there is only one segmentation of dimension 1, represented by the trivial partition, accounting for the term t in the relation above.
- any segmentation of dimension not less than 2 can be seen as the conjunction of two segmentations of two parts of the underlying space, which accounts for the exponent 2 in the relation above, by the Cartesian product relation in Lemma B.1.
- there is at most two pairs of such parts, arising either from a vertical split or from a horizontal one. This accounts for the factor 2 in the relation above.

The result follows by a direct application of Lemma B.2 with $a = 2$ and $b = 2$ yielding the exponent $\frac{\log(ab)}{b-1} + \log\left(\frac{b}{b-1}\right) = 3\log 2$ and the binomial complexity rate bound $\left[\frac{\log(ab)}{b-1} + \log\left(\frac{b}{b-1}\right) - \log(2)\right]^+ = 2\log 2 \leq 1.39$. □

B.4.6 Enumerating regular segmentation quad-trees

Lemma B.5 (Enumerating regular quad-trees)**.** *With the segmentation rule set out in Section B.3.1 for regular quad-trees in* $\mathbb{N}^2$*, if the underlying space admits at least one split in four parts, for any integer* $D > 0$ *the number of models of dimension* D *is not more than* $4^{-\frac{1}{3}}\left[\frac{4^{\frac{4}{3}}}{3}\right]^D$*. The corresponding binomial complexity rate is less than* $\frac{5}{3}\log 2 - \log 3 \leq 0.057$

Proof. The same arguments apply than for regular binary segmentation, with the difference that there is at most one possible split position per part, so that the tags on split parts may be chosen constant. Here again the cardinal of the model family increases along the inclusion relation for underlying spaces. With the assumption that the

underlying space admits at least a split in four parts, the generating function $G_{\mathcal{M}}(t)$ satisfies:

$$G_{\mathcal{M}}(t) \leq t + G_{\mathcal{M}}(t)^4$$

and the result follows from Lemma B.2 with $a = 1$ and $b = 4$ yielding the exponent $\frac{\log(ab)}{b-1} + \log\left(\frac{b}{b-1}\right) = \frac{8}{3}\log(2) - \log(3) \leq 0.75$ and the binomial complexity rate bound $\left[\frac{\log(ab)}{b-1} + \log\left(\frac{b}{b-1}\right) - \log(2)\right]^+ = \frac{5}{3}\log 2 - \log 3 \leq 0.057$. □

B.4.7 Enumerating free binary segmentation trees

In [FGOR92], Flajolet and al. give a precise but asymptotic account of the number of rooted binary trees of prescribed height. Sticking to elementary methods yields the following:

Lemma B.6 (Enumerating free binary segmentation trees)**.** *With the segmentation rule set out in B.3.1 for free binary regression trees in $\mathbb{N}^2$, for any integer $D > 1$, the number of models of dimension D is not more than $\left[\left(3 + 2\sqrt{2}\right)\frac{en}{D}\right]^D$. The binomial complexity rate of the corresponding model family is less than* 2

Proof. We rely on the observation that knowing the binary tree structure of its segmentation and the orientations of its splits, an element $|m|$ of the present model family is uniquely determined by the sequence of the cardinals of its parts. The number of such sequences is $\binom{n-1}{|m|-1}$. Lemma B.3 provides a bound on the number of regression tree structures for an underlying space of dimension d, of value $(d-1)\left(\sqrt{d} + \sqrt{d-1}\right)^2$, so that with $d = 2$ there are not more than $\left(\sqrt{2}+1\right)^{2D} = \left(3 + 2\sqrt{2}\right)^D \leq 5.83^D$ such structures. By the classical bound $\log\binom{n-1}{D-1} \leq \log\binom{n}{D} \leq D\log\frac{en}{D}$ [Mas07][p. 20] we infer that for $D > 0$, the number N_D of models with dimension D satisfies:

$$\begin{aligned} N_D &\leq \left(3 + 2\sqrt{2}\right)^D \binom{n}{D}, \\ &\leq \left[\left(3 + 2\sqrt{2}\right)\frac{en}{D}\right]^D. \end{aligned}$$

By the remark following definition 8.4, the corresponding binomial complexity rate is less than $\sup\left\{\frac{\log N_D}{D\log\frac{en}{D}}\right\}_{0<D\leq n} = 1 + \log\left(3 + 2\sqrt{2}\right) \leq 2.77$. However the bound $\mathrm{B}_{\mathcal{M}} \leq \frac{\log(2d)}{\log 2}$ from Proposition A.1 on block layouts in $\mathbb{N}^d$ is better with $d = 2$:

$$\mathrm{B}_{\mathcal{M}} \leq \frac{\log(2d)}{\log 2} = 2.$$

□

B.4.8 Enumerating free segmentation quad-trees

Lemma B.7 (Enumerating free segmentation quad-trees)**.** *With the segmentation rule set out in B.3.1 for free quad-trees in $\mathbb{N}^2$, for any integer $D > 1$, the number of models of dimension D is not more than $\left[\frac{4en}{D}\right]^D$. The binomial complexity of the corresponding model family is less than* 2*.*

Proof. Due to the particularities of the segmentation rule, the situation with free quad-trees is more intricate than the previous ones, as the allowed splits may be four-fold for non-segment parts, but only two-fold for segment (thin) parts. As a result it may not be straightforward to associate a free quad-tree segmentation with a rooted plane 4-ary tree structure of same dimension, as the branching factor is not constant. A first observation is that a quad-tree as defined in B.3.1 is also an instance of binary tree in which the orientation of the splits are prescribed (alternating in non-segment parts, and crossing the largest dimension in segment parts). By Lemma B.2 with $a = 1$ and $b = 2$ there is not more than

$$\left(\frac{1}{ab}\right)^{\frac{1}{b-1}}\left[(ab)^{\frac{1}{b-1}}\frac{b}{b-1}\right]^D = \frac{1}{2}4^D$$

such rooted plane binary tree structures with D leaves. We know from what precedes that a quad-tree segmentation is determined by its rooted plane binary tree structure and the sequence of the cardinal of its parts, and that there are not more than $\binom{n-1}{D-1} \leq \binom{n-1}{D-1} \leq \left[e\frac{n}{D}\right]^D$ such sequences. This lead to the first result announced:

$$N_D \leq 4^D\left[e\frac{n}{D}\right]^D.$$

By the remark following definition 8.4, the corresponding binomial complexity rate is less than $\sup\left\{\frac{\log N_D}{D\log\frac{en}{D}}\right\}_{0<D\leq n} = 1 + 2\log 2 \leq 2.39$. However the bound $\mathrm{B}_{\mathcal{M}} \leq \frac{\log(2d)}{\log 2}$ from Proposition A.1 on block layouts in $\mathbb{N}^d$ is better with $d = 2$:

$$\mathrm{B}_{\mathcal{M}} \leq \frac{\log(2d)}{\log 2} = 2.$$

□

B.4.9 Summary

Table 16 summarizes the different situations met and the corresponding complexity results.

Type of tree	bound on number of segmentations with D leaves	binomial complexity rate $B_{\mathcal{M}}$	Reference
bloc layouts, in $\mathbb{N}^d$	$\left(e\frac{n-1}{D-1}\right)^{\frac{\log 2d}{\log 2}(D-1)}$	$\frac{\log(2d)}{\log 2}$	Pro. A.1
abstract regular b-ary, $b>1$ tags in $\mathbb{Z}_a$ on inner nodes	$\left[(ab)^{\frac{1}{b-1}}\frac{b}{b-1}\right]^D$	$\left[\frac{\log(ab)}{b-1}+\log\left(\frac{b}{b-1}\right)-\log(2)\right]^+$	Lem. B.2
abstract regression tree in $\mathbb{N}^d$ ($\sim$phylogenetic)	$\left[(d-1)\left(\sqrt{d}+\sqrt{d-1}\right)^2\right]^D$		Lem. B.3
free line segmentation	$\left[e\frac{n-1}{D-1}\right]^{D-1}$	1	Sec. B.2
regular binary in $\mathbb{N}^2$	8^D	$2\log(2)\leq 1.39$	Lem. B.4
free binary in $\mathbb{N}^2$	$\left[\left(3+2\sqrt{2}\right)e\frac{n}{D}\right]^D$	2	Lem. B.6
regular quad in $\mathbb{N}^2$	$\left[\frac{4\frac{4}{3}}{3}\right]^D$	$\frac{5}{3}\log(2)-\log(3)\leq 0.057$	Lem. B.5
free quad in $\mathbb{N}^2$	$\left[4e\frac{n}{D}\right]^D$	2	Lem. B.7

Table 16 – Complexity of some types of recursive segmentation trees

References

[AB96] F. Abramovich and Y. Benjamini. Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis*, 22:351–361, 1996.

[AB09] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. In *NIPS 2009 - Advances in Neural Information Processing Systems*, volume 22, pages 46–54, Vancouver, Canada, December 2009.

[ABP04] E. Ackerman, G. Barequet, and R. Y. Pinter. On the number of rectangular partitions. In J. Ian Munro, editor, *SODA*, pages 736–745. SIAM, 2004.

[AJS74] M. Knott A. J. Scott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507–512, 1974.

[Aka73] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akadémiai Kiado.

[AL89] I. E. Auger and C. E. Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of mathematical biology*, 51(1):39–54, 1989.

[AM09] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245, 2009.

[Arl14] S. Arlot. *Contributions to statistical learning theory: estimator selection and change-point detection*. Habilitation à diriger des recherches, Université Paris Diderot, December 2014. Rapporteurs: Boucheron, S., Arnak Dalalyan, Lugosi, G.

[BBM99] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413, 1999.

[BD62] R. Bellman and S. Dreyfus. *Applied dynamic programming*. Princeton University Press., 1962.

[BFOS84] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: a Nonasymptotic Theory of Independence*. Oxford university press, Oxford, 2013.

[BM97] L. Birgé and P. Massart. From model selection to adaptive estimation. In David Pollard, Erik Torgersen, and GraceL. Yang, editors, *Festschrift for Lucien Le Cam*, pages 55–87. Springer New York, 1997.

[BM01] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3:203–268, 2001.

[BM07] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33–73, May 2007.

[BMM12] J. P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

[BN93] M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Application (Prentice Hall information and system sciences series)*. Prentice Hall, May 1993.

[BR06] L. Birgé and Y. Rozenholc. How many bins should be put in a regular histogram. *ESAIM: PS*, 10:24–45, 2006.

[Cas99] G. Castellan. Modified Akaike's criterion for histogram density estimation. Technical report, 1999.

[CG00] J. Chen and A. K. Gupta. *Parametric Statistical Change point Analysis*. Birkhauser, 2000.

[CG12] J. Chen and A. K. Gupta. *Parametric Statistical Change Point Analysis*. Birkhäuser/Springer, New York, second edition, 2012. With applications to genetics, medicine, and finance.

[CGHK78] F. R. K. Chung, R. L. Graham, V. E. Hoggatt, and M. Kleiman. The number of Baxter permutations. *Journal of Combinatorial Theory, Series A*, 24(3):382–394, 1978.

[CLLLR15] S. Chakar, É. Lebarbier, C. Lévy-Leduc, and S. Robin. A robust approach for estimating change-points in the mean of an AR(1) process. 2015.

[CMR05] O. Cappé, É. Moulines, and T. Ryden. *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[DG03] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. In *Random Structures and Algorithms*, volume 22, pages 60–65, 2003.

[DJ94] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[DLRY06] R. A. Davis, T. C. M. Lee, and G. A. Rodriguez-Yam. Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223–239, 2006.

[DR98] D. Dubhashi and D. Ranjan. Balls and bins: A study in negative dependence. *Random Structures and Algorithms*, 13(2):99–124, 1998.

[ECS65] A. W. F. Edwards and L. Cavalli-Sforza. A method for cluster analysis. *Biometrics*, 21(2):362–375, 1965.

[EFK11] I. A. Eckley, P. Fearnhead, and R. Killick. Analysis of changepoint models. In *Bayesian Time Series Models*. Cambridge University Press, 2011.

[FC03] P. Fearnhead and P. Clifford. On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):887–899, 2003.

[Fea06] P. Fearnhead. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213, 2006.

[FGOR92] P. Flajolet, Z. Gao, A. M. Odlyzko, and B. Richmond. The distribution of heights of binary trees and other simple trees. Research Report RR-1749, INRIA, 1992.

[FL07] P. Fearnhead and Z. Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605, 2007.

[FS09] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, New York, NY, USA, 1 edition, 2009.

[GM05] A. Gionis and H. Mannila. Segmentation algorithms for time series and sequence data. *SIAM International Conference on Data Mining, Newport Beach, CA*, 2005.

[GN05] S. Gey and É. Nédèlec. Model selection for CART regression trees. *IEEE Transactions on Information Theory*, 51(2):658–670, Feb 2005.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

[Ima] Public Domain Images. Raccoon, procyon, lotor. http://www.public-domain-image.com/free-images/fauna-animals/raccoons/raccoon-procyon-lotor. in Scipy library.

[KFE12] R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[KNAE12] R. Killick, C. F. H. Nam, J. A. D. Aston, and I. A. Eckley. changepoint.info: The changepoint repository, 2012.

[Kup] L. P. Kuptsov. Encyclopedia of Mathematics. http://www.encyclopediaofmath.org/index.php?title=Gamma-function&oldid=29082.

[Lav05] M. Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501–1510, 2005.

[Leb05a] É. Lebarbier. Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85(4):717–736, 2005.

[Leb05b] É. Lebarbier. *Quelques Approches pour la Détection de Rupture à Horizon Fini*. PhD thesis, Université Paris XI Orsay, 2005.

[Ler12] M. Lerasle. Optimal model selection in density estimation. *Annales de l'IHP - Probabilités et Statistiques*, 48(3):884–908, June 2012.

[LL01] M. Lavielle and É. Lebarbier. An application of MCMC methods for the multiple change-points problem. *Signal Processing*, 81(1):39–53, 2001.

[LM15] C. Lacour and P. Massart. Minimal penalty for Goldenshluger-Lepski method. *ArXiv e-prints*, March 2015.

[LMH04] É. Lebarbier and T. Mary-Huard. Le critère BIC : Fondements Théoriques et Interprétation. Research Report RR-5315, INRIA, 2004.

[LN07] É. Lebarbier and É. Nédèlec. Change-points detection for discrete sequences via model selection. Technical report, Statistics for Systems Biology Group, Jouy-en-Josas/Paris/Evry, France, April 2007.

[LT05] M. Lavielle and G. Teyssière. Adaptive detection of multiple change–points in asset price volatility. *Long Memory in Economics. Springer Verlag, Berlin*, pages 129–156, 2005.

[Mal68] C. Mallows. An inequality involving multinomial probabilities. *Biometrika*, 55:422–424, 1968.

[Mal73] C. L. Mallows. Some comments on Cp. *Technometrics*, 15:661–675, 1973.

[Mas90] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3), 1990.

[Mas07] P. Massart. *Concentration Inequalities and Model Selection, Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*. Springer, 2007.

[Pem04] R. Pemantle. Towards a theory of negative dependence. *ArXiv Mathematics e-prints*, April 2004.

[PHL+12] F. Picard, M. Hoebeke, É. Lebarbier, Miele V., Rigaill G., and Robin S. *cghseg: Segmentation Methods for Array CGH Analysis*, 2012. R package version 1.0.1.

[Pho13] SR Photies. Lone tree birds. https://flic.kr/p/jXuQCp, 10 2013.

[Pic07] F. Picard. An introduction to process segmentation. Technical report, Statistics for Systems Biology Group, Jouy-en-Josas/Paris/Evry, France http://genome.jouy.inra.fr/ssb/, March 2007.

[PLH+11] F. Picard, É. Lebarbier, M. Hoebeke, G. Rigaill, B. Thiam, and S. Robin. Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics*, 12(3):413–428, 2011.

[PRL+05] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J. J. Daudin. A statistical approach for array cgh data analysis. *BMC Bioinformatics*, 6, 2005.

[RCW+07] J. Reeves, J. Chen, X. L. Wang, R. Lund, and L. QiQi. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6):900–915, 2007.

[RH09] M. Rezghi and S. M. Hosseini. A new variant of L-curve for Tikhonov regularization. *Journal of Computational and Applied Mathematics*, 231(2):914 – 924, 2009.

[Rig10] G. Rigaill. Pruned dynamic programming for optimal multiple change-point detection. *ArXiv e-prints*, 2010. Provided by the SAO/NASA Astrophysics Data System.

[Rob55] H. Robbins. A remark on Stirling's formula. *Amer. Math. Monthly*, 62:26–29, 1955.

[Sau10] A. Saumard. *Regular Contrast Estimation and the Slope Heuristics*. Thèses, Université Rennes 1, October 2010.

[Sau13] A. Saumard. Optimal model selection in heteroscedastic regression using piecewise polynomial functions. *Electron. J. Statist.*, 7:1184–1223, 2013.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978.

[SG12] J. A. Shine and J. E. Gentle. Alarm activation, pattern discovery, and anomaly detection in sensor networks. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(6):565–570, 2012.

[Yan05] Y. Yang. Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

List of Figures

1 Detection de ruptures dans la moyenne d'un signal gaussien. Segmentations optimales à dimension fixées obtenue par l'algorithme de programmation dynamique. . . . 25

2 Recursive partitioning and rooted plane trees . . . 40

3 Extending a rooted 4-ary tree to contain two leaves of height 3 . . . 43

4 Splitting method of a regression tree or hyperrectangle layout: $C_{ext} \leq 2d$ 50

5 Image sqs300binom2fisourceimage.png . . . 63

6 Segmentation path of noisy image, at dimensions [1, 10, 32, 64, 128, 251, 508, 1008, 2048, 4071, 8192, 9985, 9993](estimates and borders) . . . 64

7 Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [149, 257, 515, 1027, 2050, 4098, 8193, 10002](estimates and borders) . . . 65

8 contrasts and penalized contrast: segmentation paths of source and observerved (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties. . . . 66

9 Image sqs300binom4fisourceimage.png . . . 68

10 Segmentation path of noisy image, at dimensions [1, 4, 7, 16, 31, 64, 127, 255, 498, 1023, 2046, 4088, 8188, 9976, 9979](estimates and borders) . . . 69

11 Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [94, 130, 256, 512, 1028, 2050, 4098, 8195, 10001](estimates and borders) . . . 70

12 contrasts and penalized contrast: segmentation paths of source and observerved (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties. . . . 71

13 Image sqs300binom2risourceimage.png . . . 73

14 Segmentation path of noisy image, at dimensions [1, 4, 7, 16, 31, 63, 128, 251, 503, 1024, 2045, 4096, 8191, 9992, 9993](estimates and borders) . . . 74

15 Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [94, 132, 257, 513, 1024, 2053, 4108, 8198, 10006](estimates and borders) . . . 75

16 contrasts and penalized contrast: segmentation paths of source and observerved (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties. . . . 76

17 Image sqs300binom4risourceimage.png . . . 78

18 Segmentation path of noisy image, at dimensions [1, 4, 7, 16, 31, 58, 121, 256, 511, 970, 2038, 4093, 8185, 9985, 9988](estimates and borders) . . . 79

19 Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [94, 148, 280, 517, 1036, 2068, 4105, 8203, 10012](estimates and borders) . . . 80

20 contrasts and penalized contrast: segmentation paths of source and observerved (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties. . . . 81

21 Image lontr300binom2fisourceimage.png . . . 83

22 Segmentation path of noisy image, at dimensions [1, 2, 3, 6, 15, 32, 64, 128, 238, 512, 1000, 2043, 4086, 8192, 9997, 9999](estimates and borders) 84

23 Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [301, 514, 1027, 2048, 4098, 8194, 10000](estimates and borders) . . . 85

24 contrasts and penalized contrast: segmentation paths of source and observerved (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties. . . . 86

25 Image lontr1000binom2fisourceimage.png . . . 88

26 Segmentation path of noisy image, at dimensions [1, 2, 3, 6, 16, 30, 63, 127, 255, 511, 1023, 2045, 4091, 8192, 9995, 9999](estimates and borders) 89

27 Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [1002, 1026, 2048, 4096, 8195, 10001](estimates and borders) . . . 90

28 contrasts and penalized contrast: segmentation paths of source and observerved (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties. . . . 91

29 Image lontr3000binom2fisourceimage.png . . . 93

30 Segmentation path of noisy image, at dimensions [1, 2, 3, 6, 16, 31, 63, 127, 250, 508, 1024, 2033, 4094, 8186, 9992, 9998](estimates and borders) 94

31 Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [3001, 4096, 8195, 10001](estimates and borders) . . 95

32 contrasts and penalized contrast: segmentation paths of source and observerved (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties. . . . 96

33 Image lontr5000binom2fisourceimage.png . . . 98

34 Segmentation path of noisy image, at dimensions [1, 2, 3, 6, 15, 32, 63, 128, 256, 510, 1024, 2040, 4095, 8192, 9995, 9996](estimates and borders) 99

35 Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [5001, 8192, 10001](estimates and borders) . . . 100

36 contrasts and penalized contrast: segmentation paths of source and observerved (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties. . . . 101

37 Image face300binom2fisourceimage.png 103

38 Segmentation path of noisy image, at dimensions [1, 2, 4, 8, 13, 32, 64, 124, 254, 512, 1022, 2047, 4090, 8189, 9997, 9998](estimates and borders) 104

39 Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [301, 513, 1024, 2049, 4098, 8195, 10000](estimates and borders) 105

40 contrasts and penalized contrast: segmentation paths of source and observerved (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties. 106

41 Image face1000binom2fisourceimage.png 108

42 Segmentation path of noisy image, at dimensions [1, 2, 4, 8, 16, 32, 64, 127, 252, 504, 1013, 2037, 4093, 8192, 9996, 9997](estimates and borders) 109

43 Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [1002, 1024, 2050, 4096, 8192, 10000](estimates and borders) 110

44 contrasts and penalized contrast: segmentation paths of source and observerved (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties. 111

45 Image face3000binom2fisourceimage.png 113

46 Segmentation path of noisy image, at dimensions [1, 2, 3, 4, 13, 32, 64, 128, 256, 512, 1021, 2045, 4096, 8192, 9998, 9999](estimates and borders) 114

47 Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [3003, 4098, 8194, 10000](estimates and borders) . . 115

48 contrasts and penalized contrast: segmentation paths of source and observerved (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties. 116

49 Image face5000binom2fisourceimage.png 118

50 Segmentation path of noisy image, at dimensions [1, 2, 3, 8, 14, 32, 64, 127, 256, 511, 1022, 2047, 4095, 8188, 9993, 9994](estimates and borders) 119

51 Segmentation path of noisy image along noise peeks starting from near oracle, at dimensions [5001, 8194, 10001](estimates and borders) 120

52 contrasts and penalized contrast: segmentation paths of source and observerved (noisy) images, oversegmentation path of observed image, sufficient and minimal penalties. 121

53 Free line segmentation for histograms: instances of m_N, selected parts in m_N and completed segmentation. $C_{ext} = 2$ 128

54 Dyadic line segmentation for histograms with $h = 4$ and $N + 1 = 2^4$: instances of m_N, selected parts in m_N and completed segmentation. $C_{ext} \leq (2-1)h = 4$ 128

55 Regression tree for histograms over $[0,1]^d$: m_N, selected parts in m_N, completed segmentations. $C_{ext} \leq 2d$ 129

56 Bloc layouts for histograms over $[0,1]^d$: m_N, selected parts in m_N, completed segmentations. $C_{ext} \leq 2d$ 129

57 Voronoi segmentation for histograms over $[0,1]^d$: m_N and selected parts, completed segmentations. 130

58 Minimal rooted b-ary tree with a prescribed part c as leaf 143

List of Tables

1 A first experiment in model selection, averaging 12

2 A first experiment in model selection, wavelet thresholding 13

3 List of numerical experiments for Section 8 60

4 Data for image squares example squares.2fold.(2048, 2048) 62

5 Data for image squares example squares.4fold.(2048, 2048) 67

6 Data for image squares example squares.2foldequal.(2048, 2048) 72

7 Data for image squares example squares.4foldequal.(2048, 2048) 77

8 Data for image lone tree example lone tree.2fold.(1098, 1672) 82

9 Data for image lone tree example lone tree.2fold.(1098, 1672) 87

10 Data for image lone tree example lone tree.2fold.(1098, 1672) 92

11 Data for image lone tree example lone tree.2fold.(1098, 1672) 97

12 Data for image face example face.2fold.(512, 512) 102

13 Data for image face example face.2fold.(512, 512) 107

14 Data for image face example face.2fold.(512, 512) 112

15 Data for image face example face.2fold.(512, 512) 117

16 Complexity of some types of recursive segmentation trees 220

Index

C_p, *see* Mallows,C_p
D_m, *see* model,dimension
$G_{\mathcal{M}}(t)$, *see* function,generating function
L_m, *see* weights
N_τ, *see* part,occupancy count,random
P_n , *see* measure,empirical
W, *see* process,isonormal
$Y(t)$, *see* signal,observed,Gaussian
$\mathcal{M}$, *see* model,family
$|\tau|$, *see* measure, of a part
$|m|$, *see* model,dimension, *see* partition,cardinal, *see* partition,cardinal
$\bar{u}_\tau$, mean over part46
$b^\star{}_k$, *see* part,selected parts
ϵ, *see* noise,level
$\mathrm{C}_{\mathrm{ext}}$, extension,factor49
$\gamma_n(\cdot)$, empirical contrast33
$g(\cdot)$, *see* function,Gamma tail
$\mathbf{h}^2(\cdot,\cdot)$, Hellinger distance124
$h(\cdot)$, *see* function,Poisson tail
$h_p(\cdot)$, *see* function,binomial tail
$\bar{k}$, *see* part,occupancy count,expected
$\mathbf{K}(\cdot,\cdot)$, *see* information number,Kullback-Leibler
$\mathcal{R}_m(s)$, *see* risk,quadratic
$\hat{m}$, *see* model,selected
$\odot$, *see* model,extension
$\hat{s}$, *see* estimator,minimum contrast
$\sqcup$, disjoint union39
$\mathbb{V}[\cdot]$, variance134
$d(\cdot,\cdot)$, *see* quadratic distance
p_k^+ , *see* deviation,rate,expected
$r_k^\star$, *see* deviation,rate
s, *see* signal,unknown
s_m, *see* model,projection on
$m^\star{}_k$, *see* model,adverse

Akaike, *see* criterion,Akaike
algorithm
  CART, 59

BIC, *see* criterion,Bayesian information criterion
bin, *see* partition,bin
binomial complexity rate, 51, 131

CART, *see* algorithm,CART
Castellan, 124
changepoint
  multiple changepoint detection, 17
completion rule, *see* model,extension
contrast
  empirical, 14, 33
criterion
  Akaike, 15
  Bayesian information criterion, 16
  critère d'information bayésien, 16
  least squares, 33
  penalized contrast criterion, 15

deviation
  k-deviation, 160
  rate, 160
    expected, 160
distance
  Hellinger, 123
  quadratic, 33
divergence
  Kullback-Leibler, *see* information number

efficacité, 16
estimator
  histogram, 123, 210
  minimum contrast, 14
  projection, 123
  regressogram, 46, 210
extension, *see* model,extension

function
  beta function, 178
  binomial tail function, 175
  Gamma tail function, 175
  generating function, 212
  penalty function, 15
  Poisson tail function, 174

height, *see* part,height
Hellinger , *see* distance, Hellinger
histogram, *see* estimator,histogram
hyperrectangle, 205

image

noisy, 58
source, 58
information number
Kullback-Leibler, 123
isolate
a subset, 47
an element, 47
isonormal, *see* process,isonormal
isotropy, 135, 180

Kullback-Leibler
nombre d'information, 14
Kullback-Leibler , *see* information number

M-estimation, 14
Mallows, 15
C_p, 15
mean $\bar{u}_\tau$ of u over τ, 46
measure
counting, 46
empirical, 122
invariant, 180
of a part, 46
minimal penalty, 36
η-minimizer, 44
model
adverse, 58, 160
dimension, 33
extension
completion rule, 49
factor, 49
symbol $\odot$, 49
family, 15
linear, 33
oversegmented, 58
partition models, 46
projection on, 33, 46
selected, 34
model selection via penalisation, 15

noise
level, 33

occupancy, *see* part,occupancy count
optimality, 16
oracle
near oracle, 58
oracle inequality, 34
overfits, 20

part, *see* partition,part, *see* partition,part
height, 48
occupancy count
assumption on, 127
expected, 127
random, 127, 160
selected parts, 160
partition, 47
bin, 122
cardinal, 47
family, 122
part, 45, 122
parts, 45
region, 122
set, 45
partitioning
b-adic, 48
b-ary, 48
recursive, 47
parts, 47
rule, 48
scheme, 48
process
Gaussian linear, 33
isonormal, 33
renewal, 209
projection, *see* estimator,projection
projector
randomly oriented, 180

region, *see* partition,region
regression
ridge, 19
tree, *see* tree,regression tree
regressogram, *see* estimator,regressogram
risk
Hellinger, 124, 133
quadratic, 34

segmentation
path, 58
segmentation of a sequence, 208
segmentation path, 58
signal
observed, 58
Gaussian, *see* process,Gaussian linear
unknown, 33
space

feature space, 45
underlying, 45
split
admissible, 47
direct, 48
support, *see* space,underlying

Tikhonov, *see* regression,ridge
tree
binary
free, 210
regular, 210
phylogenetic, 215, 216
quad-tree
free, 210
regular, 210
regression tree, 47, 49, 59
bi-dimensional, 210
multi-dimensional, 210
rooted plane, 48
structure, 212
structures
enumerating, 213
tree structure, 213

underlying space, 39, *see* space,underlying, 122, 208

weights, 18

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Pénalités minimales pour la sélection de modèle

Mots Clefs : moindres carrés pénalisés, sélection de modèle, pénalité minimale, segmentation de signal gaussien, estimation de densité, contraste pénalisé, détection de ruptures multiples, CART, arbres de régression

Résumé : Dans le cadre de la sélection de modèle par contraste pénalisé, L. Birgé et P. Massart ont prouvé que le phénomène de pénalité minimale se produit pour la sélection libre parmi des variables gaussiennes indépendantes. Nous étendons certains de leurs résultats à la partition d'un signal gaussien lorsque la famille de partitions envisagée est suffisamment riche, notamment dans le cas des arbres de régression. Nous montrons que le même phénomène se produit dans le cadre de l'estimation de densité. La richesse de la famille de modèles est liée à une forme d'isotropie. De ce point de vue le phénomène de pénalité minimale est intrinsèque. Pour corroborer et illustrer ce point de vue, nous montrons que le même phénomène se produit pour une famille de modèles d'orientation aléatoire uniforme.

Title : Minimal penalties for model selection

Keywords : penalized least-squares, model selection, minimal penalties, Gaussian signal segmentation, density estimation, penalized contrast, multiple changepoints detection, CART, regression trees

Abstract : L. Birgé and P. Massart proved that the minimum penalty phenomenon occurs in Gaussian model selection when the model family arises from complete variable selection among independent variables. We extend some of their results to discrete Gaussian signal segmentation when the model family corresponds to a sufficiently rich family of partitions of the signal's support. This is the case of regression trees. We show that the same phenomenon occurs in the context of density estimation. The richness of the model family can be related to a certain form of isotropy. In this respect the minimum penalty phenomenon is intrinsic. To corroborate this point of view, we show that the minimum penalty phenomenon occurs when the models are chosen randomly under an isotropic law.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France