State-Space Models for Time Series Forecasting. Application to the Electricity Markets

Joseph De Vilmarest

Sorbonne Université
LPSM

École doctorale **École Doctorale Sciences Mathématiques de Paris Centre**

Unité de recherche **Laboratoire de Probabilités, Statistique et Modélisation**

Thèse présentée par **Joseph de Vilmarest**

En vue de l'obtention du grade de docteur de Sorbonne Université

Discipline **Mathématiques**

Spécialité **Statistiques**

Modèles espace-état pour la prévision de séries temporelles. Application aux marchés électriques.

Thèse dirigée par Olivier Wintenberger directeur
Yannig Goude Encadrant industriel
Thi Thu Huong Hoang Encadrante industrielle

Modèles espace-état pour la prévision de séries temporelles. Application aux marchés électriques.

Résumé

L'électricité étant difficile à stocker, prévoir la demande est un enjeu majeur pour maintenir l'équilibre entre la production et la consommation. L'évolution des usages de l'électricité, le déploiement des énergies renouvelables, et plus récemment la crise du coronavirus, motivent l'étude de modèles qui évoluent au cours du temps, pour tenir compte des changements de comportements. L'objectif de ce travail est de proposer des méthodes adaptatives de prévision, et nous nous sommes intéressés tout spécialement au cadre des modèles espace-état. Dans ce paradigme, on représente l'environnement (ou le contexte) par un état caché. À chaque instant, la demande dépend de cet état que nous cherchons donc à estimer grâce aux observations dont nous disposons, et selon les hypothèses que l'on effectue sur la dynamique du système. L'estimation de l'état nous permet ensuite de prévoir la demande.
Un premier objectif de la thèse est de contribuer au lien entre l'optimisation et l'estimation dans les modèles espace-état. Nous interprétons en effet les méthodes que nous utilisons comme diverses façons de paramétrer un algorithme de descente de gradient de second ordre, et nous avons détaillé ce lien dans un cas particulier. Une seconde contribution de la thèse est de proposer différentes méthodes d'estimation dans les modèles espace-état. Le principal enjeu nous semble être de définir la dynamique avec lequel évolue l'état, et nous proposons deux méthodes dans ce but. Le troisième apport de ce manuscrit est d'appliquer ces méthodes espace-état à la prévision de consommation d'électricité. Nos prévisions s'appuient sur des modèles de prévision existants, par exemple le modèle additif généralisé, que nous cherchons à adapter. Ainsi, nous tirons parti de certaines dépendances complexes capturées par les modèles existants, par exemple la sensibilité de la consommation d'électricité à la température, tout en profitant de la faculté d'adaptation des modèles espace-état.

Mots clés : modèles espace-état, prévision de consommation électrique, séries temporelles

Abstract

Electricity storage capacities are still negligible compared to the demand. Therefore, it is fundamental to maintain the equilibrium between consumption and production, and to that end, we need load forecasting. Numerous patterns motivate the study of time-varying models, including: changes in people's habits, increasing renewable capacities, more recently the coronavirus crisis. This thesis aims to propose adaptive methods for time series forecasting. We focus on state-space models, where the environment (or context) is represented by a hidden state on which the demand depends. Thus, we try to estimate that state based on the observations at our disposal. Based on our estimate, we forecast the load.
The first objective of the thesis is to enrich the link between optimization and state-space estimation. Indeed, we see our methods as second-order stochastic gradient descent algorithms, and we treat a particular case to detail that link. The second contribution concerns variance estimation in state-space models. Indeed, the variances are the parameters on which the models' dynamics crucially relies. The third part of the manuscript is the application of these methods to electricity load forecasting. Our methods build on existing forecasting methods like generalized additive models. The procedure allows to leverage advantages of both. On the one hand, statistical models learn complex relations to explanatory variables like temperature. On the other hand, state-space methods yield model adaptation.

Keywords: electricity load forecasting, state-space models, time series

Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France

Table des matières

Outline of Contributions

Chapters 1 and 2 are respectively the French and English versions of an introduction to the thesis subject, both from an industrial perspective and a formal one. We give an overview of the contributions.

Stochastic Optimization as a Static State-Space Model

In **Chapter** 3, we propose a new convergence proof of the *stochastic gradient descent* algorithm (Robbins and Monro, 1951). We interpret our analysis as a non-asymptotic variant of the Robbins-Siegmund theorem.

Chapter 4 is a study of the extended Kalman filter (Fahrmeir, 1992) in a degenerate setting called static. We compare this algorithm to standard gradient algorithms such as the *online Newton step* of Hazan, Agarwal, and Kale, 2007.

The Choice of the Variances in a State-Space Model

Chapter 5 considers the setting of hyper-parameters in a state-space model where the variances are assumed to be constant. We present the application of the well-known *expectation-maximization* algorithm. However, as the loss that is optimized is non-convex, we propose another heuristic that we call *iterative grid-search*, where we look for the optimum on a grid but in an iterative way to reduce complexity.

In **Chapter** 6, we present an approach that we name *Viking* to estimate the variances of a state-space model adaptively. We augment the state-space model, treating the variances as auxiliary latent variables that we estimate jointly with the state. We rely on *variational Bayes* (Šmídl and Quinn, 2006).

Application to Electricity Load Forecasting

The first application of the thesis was on the confidential data from EDF, and we obtained promising results. In this manuscript, we present our studies on public data sets.

In **Chapter** 7, we apply state-space approaches to adapt forecasting models of the French electricity load. We focus on the break of Spring 2020 caused by the coronavirus crisis, and we test pre-covid, break, and post-covid performances.

In **Chapter** 8, we apply our methods to a city-wide demand of an unknown location. We present the data set of a competition that we won, and we present our strategy. The competition

was held in an online manner. Each of 30 consecutive days, we had to forecast the next day's hourly load, and we received feedback of the load in the past to adjust our models.

In **Chapter** 9, we present a smaller scale forecasting task. We present our predictions on the data set of a competition that we won where the objective was to predict the electricity load of a building. This competition was also held in an online manner. Each of 5 consecutive days, we had to forecast the next day's load with 15-minute intervals, before receiving feedback.

Chapter 10 is a study of adaptive models for probabilistic forecasting. Indeed, while forecasting the mean consumption is important, risk management must have some information on the forecast error distribution. For instance, it is not the same to be sure to have a 50 GW consumption, or to forecast that there is a 90% chance the consumption ranges between 48 GW and 52 GW.

Publications

Chapters 4, 7 and 8 led to the following publications:
— Joseph de Vilmarest and Olivier Wintenberger, Stochastic Online Optimization using Kalman Recursion, *Journal of Machine Learning Research* 22.223, pp. 1-55, 2021.
— David Obst, Joseph de Vilmarest and Yannig Goude, Adaptive Methods for Short-Term Electricity Load Forecasting During COVID-19 Lockdown in France, *IEEE Transactions on Power Systems* 36.5, pp. 4754–4763, 2021.
— Joseph de Vilmarest and Yannig Goude, State-Space Models for Online Post-Covid Electricity Load Forecasting Competition, *IEEE Open Access Journal of Power and Energy*, 2022.
Chapter 6 is based on a submitted article.

Presentations at International Conferences

— Poster presentation at the *time-series workshop* of the International Conference on Machine Learning, 2021.
— Presentation at the panel session *Performance evaluation of artificial intelligence methods for energy consumption forecasting using open data sets* at the IEEE Power & Energy Society General Meeting, 2021.

Competitions

We were awarded first place in the two following competitions:
— *Day-Ahead Electricity Demand Forecasting: Post-COVID Paradigm* hosted by IEEE DataPort (Farrokhabadi, 2020). See Chapter 8.
— *Competition on building energy consumption forecasting.* See Chapter 9.

Implementation

The experiments of the thesis were realized using the R language. The implementation of the forecasting strategy presented in Chapter 8 is available online [1].

1. https://gitlab.com/JosephdeVilmarest/state-space-post-covid-forecasting

Chapitre 1

Introduction (française)

L'objectif de ce chapitre est d'introduire précisément la thèse en français. Nous présentons les motivations industrielles, ainsi que le cadre théorique dans lequel elle s'inscrit. Puis nous décrivons brièvement les contributions apportées par cette thèse.

Sommaire

1.1 Cadre industriel

La prévision de séries temporelles est un enjeu majeur. Il peut s'agir de prévision météorologique ou climatique pour les agriculteurs, de prévision de ventes et donc de stocks pour les commerçants, de prix et de coûts pour les industriels, chacun dépend fondamentalement de prévisions qui l'ont mené à prendre une décision. Dans cette thèse nous étudions le secteur électrique qui présente une singularité en comparaison des applications mentionnées. Comme on ne peut pas stocker l'électricité à grande échelle, il est primordial pour le réseau électrique que l'équilibre

offre-demande soit assuré : à chaque instant, la production d'électricité (offre) doit être égale à la consommation (demande).

Pour obtenir cet équilibre il est nécessaire de prévoir la demande à tous les horizons de temps. À long, voire très long terme (mois, années, décennies), la prévision est utilisée pour déterminer le parc de production adéquat, et ainsi construire de nouvelles unités et décider les dates de maintenance des unités existantes. Les prévisions moyen terme (semaines, mois) permettent d'actualiser les programmes de maintenance, et permettent de décider de l'exploitation ou non d'une unité de production pilotable en fonction de son utilité potentielle future. À court-terme (jour), prévoir la consommation permet de définir un planning de production pour les moyens pilotables dont le changement de puissance demande quelques heures (les centrales nucléaires, dans une moindre mesure les centrales thermiques à flamme). À très court-terme (moins d'un jour) ce planning de production est réactualisé pour que l'erreur résiduelle soit la plus faible possible, et celle-ci est traitée par le gestionnaire de réseau avec des moyens très réactifs, à commencer par les barrages hydrauliques.

Ce bref aperçu de la gestion d'un réseau électrique présuppose que la demande serait une grandeur unique et qu'une entité centralisée prendrait l'ensemble des décisions. Ce n'est pas le cas pour différentes raisons. D'une part, l'ouverture à la concurrence du secteur électrique français réduit la centralisation. Aujourd'hui les fournisseurs d'électricité doivent s'occuper de l'équilibre offre demande pour leurs clients : ils doivent produire l'électricité consommée par leurs clients, ou bien se fournir auprès d'un autre producteur. Chaque fournisseur gère cet équilibre au mieux puis, le réseau étant utilisé par l'ensemble, l'équilibre national est géré par RTE (Réseau de Transport d'Electricité). Pour résorber l'écart résiduel national RTE fait appel à des réserves (des producteurs qui acceptent d'augmenter ou baisser leur production) et inflige des pénalités aux fournisseurs selon les écarts de chacun. D'autre part, s'il est nécessaire de respecter l'équilibre à l'échelle nationale, les contraintes du réseau incitent à fournir la demande à une échelle plus locale. Certains pays présentent une décorrélation spatiale de la production et de la consommation, et l'acheminement de l'électricité produite dans une région vers une autre est coûteuse, non seulement à cause des pertes en ligne mais également parce que cela nécessite une infrastructure plus importante, puisque le réseau est dimensionné en proportion du transport qu'il assume.

EDF a continument amélioré la prévision de la demande au cours des dernières décennies. Classiquement, les méthodes de prévision modélisent le comportement observé dans un certain historique (typiquement cinq ans). Cependant, pour pouvoir extrapoler et prévoir la consommation future, une hypothèse de stabilité est nécessaire.

Ces dernières années, cette hypothèse a été remise en question pour de nombreuses raisons. Suite à l'ouverture à la concurrence EDF prévoit la consommation d'un portefeuille de clients qui évolue au cours du temps. D'autre part, l'électrification des usages et en particulier le développement des véhicules électriques pourrait changer structurellement la consommation. Bien que l'impact soit aujourd'hui négligeable, il est à prévoir que ce ne soit plus le cas dans quelques années. De plus, si la stabilité de la consommation est vérifiée à l'échelle nationale c'est de moins en moins vrai à mesure que l'échelle devient plus locale. Le cas extrême de la prévision d'un logement individuel est évocateur : si le logement devient vacant il est primordial de changer le modèle de prévision. Le récent développement des énergies renouvelables dites fatales (le solaire et l'éolien, sur lesquels on ne peut influer) change fondamentalement la gestion de l'équilibre entre l'offre et la demande. Les moyens pilotables ne sont plus employés pour satisfaire la consommation, mais pour satisfaire la consommation *nette*, une fois soustraite la production non pilotable. La variable d'intérêt devient donc la demande nette, qui présente une plus forte variabilité. Enfin, la crise liée au coronavirus a fortement déstabilisé le réseau électrique. Les confinements mis en place dans de nombreux pays ont fait chuter brutalement la demande, et les évolutions fréquentes sur les restrictions ont impliqué de plus grandes évolutions des comportements que d'ordinaire, ce

qui a donné lieu à une difficulté accrue pour la prévision.

De nouveaux modèles plus réactifs ont été mis au point par EDF pour être plus performants dans des situations instables. Parmi eux, l'agrégation d'experts est une technique générale qui combine plusieurs modèles de prévisions et tire parti de leur diversité. En effet, les modèles ne sont pas tous performants dans les mêmes contextes et l'agrégation a pour objet d'être plus robuste, puisqu'elle permet de profiter de chaque modèle. Cette méthode très étudiée par la communauté de prévision de séries temporelles a été significativement améliorée au cours d'une thèse chez EDF (Gaillard, 2015).

L'objectif de cette thèse est de poursuivre la recherche de modèles plus adaptatifs. Pour ce faire on se concentre sur les représentations espace-état, un cadre qui a déjà été étudié pour la prévision de consommation au cours d'une thèse chez EDF (Dordonnat, 2009). Le changement de comportement des données étudiées toujours plus instables motive de nouveaux travaux sur le sujet. De plus, une différence notable avec les travaux de Dordonnat, 2009 est que nous profitons de modèles statistiques ou de Machine Learning comme le *multi-layer perceptron*, que nous combinons avec des modèles espace-état. En particulier, le modèle additif généralisé (GAM) est aujourd'hui largement utilisé. Certaines méthodes ont été mises au point pour l'adapter au cours du temps, en changeant directement les coefficients du modèle (Ba et al., 2012), ou bien en utilisant un modèle correctif sur les résidus (auto-régressif, par exemple). Notre approche étend la première option, nous ne corrigeons pas les résidus du modèle, mais nous adaptons directement certains coefficients.

1.2 Cadre théorique

Dans cette section nous détaillons le formalisme sur lequel se construit la thèse. Nous notons $y_t \in \mathbb{R}$ la variable à prévoir à l'instant t (par exemple, la consommation électrique). Cet instant est un entier qui commence à 1 par convention, et le pas de temps peut être le jour, l'heure, la demi-heure ... Pour évaluer une méthode fournissant une prévision $\hat{y}_t$, nous nous intéressons à l'erreur commise $y_t - \hat{y}_t$, que l'on cherche à minimiser. Plus généralement nous cherchons à minimiser une perte $\ell(y_t, \hat{y}_t)$, la plus classique étant la perte quadratique $\ell(y_t, \hat{y}_t) = \frac{1}{2}(y_t - \hat{y}_t)^2$.

Pour prévoir y_t nous avons accès à de l'information, et l'on note $x_t \in \mathbb{R}^d$ un vecteur de taille d, dont chaque entrée est une variable explicative (par exemple la température, le jour de la semaine).

1.2.1 Prévision adaptative

Classiquement on paramètre le modèle de prévision par un vecteur : on cherche le meilleur θ de sorte à prévoir $\hat{y}_t = f_\theta(x_t)$, par exemple $f_\theta(x_t) = \theta^\top x_t$. Dans le cas non adaptatif ou *offline*, on cherche à optimiser θ sur un ensemble d'entraînement. On définit ainsi l'*Empirical Risk Minimizer* (ERM) de la façon suivante :

$$\hat{\theta}_N^{(ERM)} \in \arg\min_\theta \frac{1}{N} \sum_{t=1}^{N} \ell(y_t, f_\theta(x_t)) \,,$$

qui peut être transformé en ajoutant des termes de pénalités au problème d'optimisation précédent afin de le rendre plus robuste.

Cependant, le cadre de la thèse est celui des modèles adaptatifs, dans lequel on souhaite prévoir non plus $f_\theta(x_t)$ mais $f_{\theta_t}(x_t)$. On peut utiliser une méthode *offline* pour définir une méthode dite *incremental offline* dans laquelle au lieu d'utiliser $\hat{\theta}_N^{(ERM)}$ on estime $\hat{\theta}_t^{(ERM)}$ à chaque étape.

Cependant cette stratégie donne lieu à deux inconvénients. D'une part, la complexité d'une telle procédure peut être prohibitive (il peut être coûteux de calculer l'ERM, et d'autant plus s'il faut le ré-estimer à chaque étape). D'autre part, cette méthode donne une méthode "faiblement adaptative", dans le sens où elle évolue très lentement, alors que l'on peut souhaiter un algorithme plus réactif.

Le problème de l'optimisation en ligne est donc de trouver une transformation Φ telle que l'on définisse récursivement $\hat{\theta}_{t+1} = \Phi(\hat{\theta}_t, x_t, y_t)$ et la qualité de la prévision $\hat{y}_t = f_{\hat{\theta}_t}(x_t)$ est évaluée par la perte subie $\ell(y_t, f_{\hat{\theta}_t}(x_t))$. L'algorithme d'optimisation en ligne le plus simple est l'*Online Gradient Descent* (OGD) introduit par Zinkevich, 2003, qui consiste à faire un pas dans la direction opposée au gradient de la perte instantanée : à chaque étape on définit

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \gamma_t \frac{\partial \ell(y_t, f_\theta(x_t))}{\partial \theta}\Big|_{\hat{\theta}_t}, \tag{1.1}$$

où $\gamma_t > 0$ est le paramètre de l'algorithme appelé pas de gradient. Les algorithmes développés pendant la thèse sont très proches de l'OGD, le lien est présenté en section 1.2.3.

1.2.2 Modèle espace-état et filtre de Kalman

Au cours de la thèse on se concentre sur des modèles espace-état, dont un exemple est le modèle linéaire gaussien suivant :

$$\begin{aligned}
&\text{État :} && \theta_t = \theta_{t-1} + \eta_t\,, && \eta_t \sim \mathcal{N}(0, Q_t)\,,\\
&\text{Espace :} && y_t = \theta_t^\top x_t + \varepsilon_t\,, && \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)\,.
\end{aligned}$$

L'équation d'état régit la dynamique du modèle : l'état suit ici une marche aléatoire, à chaque étape on lui ajoute un bruit gaussien centré de matrice de covariance $Q_t \in \mathbb{R}^{d\times d}$. L'équation d'espace définit la distribution de l'observation sachant les variables explicatives x_t et l'état du modèle θ_t. Les variances (Q_t et σ_t^2) sont les deux hyper-paramètres du modèle.

Remarquons tout d'abord un cadre dégénéré intéressant, que l'on appelle statique : $Q_t = 0$ et $\sigma_t^2 = \sigma^2$. Dans ce cas on retrouve un modèle non adaptatif, avec $\theta_t = \theta_{t-1}$.

Dans le cadre général, sous l'hypothèse que le modèle espace-état soit vérifié on cherche à estimer l'état à l'instant t avec comme information les observations allant jusqu'à un instant m, et l'on s'intéresse essentiellement à l'espérance et la variance conditionnelles de l'état :

$$\begin{aligned}
\hat{\theta}_{t|m} &= \mathbb{E}[\theta_t \mid x_1, y_1, \cdots, x_m, y_m]\,,\\
P_{t|m} &= \mathbb{E}[(\theta_t - \hat{\theta}_{t|m})(\theta_t - \hat{\theta}_{t|m})^\top \mid x_1, y_1, \cdots, x_m, y_m]\,.
\end{aligned}$$

Une propriété intéressante du modèle espace-état linéaire gaussien est que lorsque la distribution initiale de l'état est gaussienne, la distribution de θ_t sachant les observations jusqu'en m reste gaussienne. Cela justifie l'estimation de $\hat{\theta}_{t|m}$ et $P_{t|m}$ puisque l'ensemble de la distribution est déterminée par son espérance et sa matrice de covariance.

Le problème le plus classique est celui de l'estimation de la distribution de l'état à l'instant t sachant les observations passées (jusqu'en $t-1$). C'est l'objectif principal dans le cadre de la prévision. Cette estimation est réalisée de façon exacte par le filtre de Kalman (Kalman et Bucy, 1961).

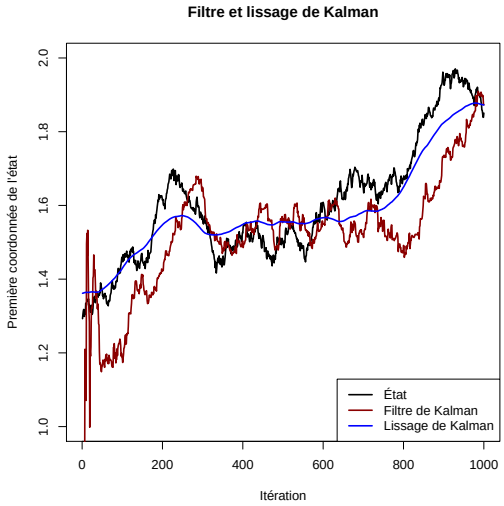


FIGURE 1.1 – Différence entre le filtre et le lissage de Kalman. Les données ont été générées selon le modèle espace-état à variances constantes, et l'on peut donc se comparer à la vraie valeur de θ_t.

Théorème 1.1 (Filtre de Kalman)**.** *Sous réserve que le modèle espace-état soit vérifié pour des variances connues $(\sigma_t^2, Q_t)_t$, les récursions suivantes sont satisfaites :*

$$P_{t|t-1} = P_{t-1|t-1} + Q_t\,, \qquad \hat{\theta}_{t|t-1} = \hat{\theta}_{t-1|t-1}\,,$$
$$P_{t|t} = P_{t|t-1} - \frac{P_{t|t-1}x_t x_t^\top P_{t|t-1}}{x_t^\top P_{t|t-1}x_t + \sigma_t^2}\,, \qquad \hat{\theta}_{t|t} = \hat{\theta}_{t|t-1} - \frac{P_{t|t}}{\sigma_t^2}\left(x_t(\hat{\theta}_{t|t-1}^\top x_t - y_t)\right)\,.$$

Inversement, le lissage de Kalman (*Kalman smoothing*) permet d'obtenir la distribution de l'état sachant les observations futures par une récursion dans le sens inverse :

Théorème 1.2 (Lissage de Kalman)**.** *Sous réserve que le modèle espace-état soit vérifié pour des variances connues, les récursions suivantes sont satisfaites :*

$$\hat{\theta}_{t|n} = \hat{\theta}_{t|t} + P_{t|t}P_{t+1|t}^{-1}(\hat{\theta}_{t+1|n} - \hat{\theta}_{t+1|t})\,,$$
$$P_{t|n} = P_{t|t} + P_{t|t}P_{t+1|t}^{-1}(P_{t+1|n} - P_{t+1|t})P_{t+1|t}^{-1}P_{t|t}\,.$$

Ce second résultat nous permet de lisser l'estimation de l'état sur une trajectoire, c'est ce qu'on illustre en figure 1.1. En effet, le premier passage sur les données donne le meilleur estimateur sachant le passé, mais le bruit sur l'observation (ε_t) trompe le filtre de Kalman. En ayant accès au futur, l'estimateur est plus lisse car la tendance de fond est bien capturée.

Le filtre de Kalman dispose de nombreux avantages. Les mises à jour sont récursives au sens où l'on obtient les estimateurs à partir de leurs valeurs précédentes ainsi que de l'observation de x_t, y_t. Ces mises à jour sont efficaces (le coût est proportionnel à d^2, d étant la dimension de l'état). De plus elles donnent de façon exacte l'espérance et la variance de l'état sachant les observations passées.

La difficulté majeure pour appliquer le filtre de Kalman est que dans la plupart des applica-

tions on ne connait pas les variances telles que le modèle espace-état soit satisfait. Au contraire, dans de nombreuses applications, le modèle est dit mal spécifié, ce qui signifie que l'hypothèse espace-état n'est pas satisfaite, quelles que soient les variances σ_t^2, Q_t. Une vaste littérature porte sur l'estimation de ces variances, les vraies dans le cas bien spécifié, et les "meilleures" dans le cas mal spécifié, "meilleures" étant un terme à définir, par exemple relativement à l'erreur de prévision.

Une première méthode porte sur l'optimisation de variances constantes : on suppose que $\sigma_t^2 = \sigma^2$, $Q_t = Q$, puis on estime les meilleures valeurs de σ^2 et Q en maximisant la vraisemblance (Brockwell et Davis, 2016 ; Durbin et Koopman, 2012 ; Fahrmeir et Tutz, 2013). Nous mettons en pratique ce paradigme au cours de la thèse, mais notons qu'il présente deux inconvénients. Premièrement, le modèle obtenu est moins vaste puisqu'on a restreint les valeurs possibles des variances au cas où elles sont constantes au cours du temps. Deuxièmement, maximiser la vraisemblance est un problème complexe pour lequel on n'obtient pas d'autre garantie que la convergence vers un optimum local, non global.

D'autres méthodes d'optimisation des variances ont donné lieu à de nombreux algorithmes appelés *adaptive Kalman filters*, qui sont donc adaptatifs plus en profondeur, dans lesquels le filtre de Kalman est appliqué avec des variances estimées au cours du temps (Mehra, 1972).

Enfin, nous ne nous sommes pas exclusivement intéressés au modèle espace-état linéaire gaussien, même si c'est le cadre que nous utilisons pour l'application à la prévision de la consommation électrique. Les autres cadres étudiés dans ce manuscrit s'écrivent de la façon suivante :

$$\text{État :} \qquad \theta_t = K\theta_{t-1} + \eta_t \,, \tag{1.2}$$

$$\text{Espace :} \qquad y_t = h(\theta_t^\top x_t) + \varepsilon_t \,, \tag{1.3}$$

où η_t et ε_t sont des bruits centrés non nécessairement gaussiens de variances respectives Q_t et s_t^2, et h est une fonction de lien entre un modèle linéaire et l'espérance de y_t. Des extensions au filtre de Kalman ont été développées dans ce cadre plus général. Nous avons considéré l'utilisation de l'*Extended Kalman Filter* (EKF) (Jazwinski, 1970), bien décrit par Durbin et Koopman, 2012, qui consiste à linéariser l'équation d'espace. Précisément, en notant h' la dérivée de h, on fait l'approximation de premier ordre suivante :

$$y_t \approx h(\hat{\theta}_t^\top x_t) + h'(\hat{\theta}_t^\top x_t) x_t^\top (\theta_t - \hat{\theta}_t) + \varepsilon_t \,.$$

Puis on applique le filtre de Kalman standard, la différence étant que les bruits ne sont pas nécessairement gaussiens et donc on estime l'espérance et la variance de l'état sans avoir la propriété de loi *a posteriori* gaussienne et connue exactement. Partant des estimateurs de la moyenne et de la variance de θ_0, notés $\theta_{0|0}, P_{0|0}$, cet estimateur est donné par les récursions suivantes à tout instant t :

$$P_{t|t-1} = KP_{t-1|t-1}K^\top + Q_t \,, \qquad \hat{\theta}_{t|t-1} = K\hat{\theta}_{t-1|t-1} \,,$$

$$P_{t|t} = P_{t|t-1} - \frac{h'(\hat{\theta}_t^\top x_t)^2 P_{t|t-1} x_t x_t^\top P_{t|t-1}}{h'(\hat{\theta}_t^\top x_t)^2 x_t^\top P_{t|t-1} x_t + s_t^2} \,, \quad \hat{\theta}_{t|t} = \hat{\theta}_{t|t-1} - \frac{P_{t|t}}{s_t^2} \left(h'(\hat{\theta}_t^\top x_t) x_t (h(\hat{\theta}_{t|t-1}^\top x_t) - y_t) \right) .$$

Par ailleurs, mentionnons l'*Unscented Kalman Filter* de Julier et Uhlmann, 1997 comme alternative à l'EKF pour l'estimation de l'état dans le cadre non linéaire.

1.2.3 Le filtre de Kalman, un algorithme de descente de gradient

Revenons à notre problème qui est de prévoir la variable y_t en minimisant une perte $\ell(y_t, \hat{y}_t)$. Dans le cas classique où $\ell(y_t, \hat{y}_t) = \frac{1}{2}(y_t - \hat{y}_t)^2$, on peut observer que le filtre de Kalman (Théorème 1.1) donne une formule de mise à jour de $\hat{\theta}_t$ très proche de celle de l'OGD (Equation 1.1). En effet, en explicitant le gradient dans l'OGD on a :

$$\text{Online Gradient Descent :} \qquad \hat{\theta}_{t+1} = \hat{\theta}_t - \gamma_t \left(x_t(\hat{\theta}_t^\top x_t - y_t) \right) ,$$

$$\text{Filtre de Kalman :} \qquad \hat{\theta}_{t+1|t} = \hat{\theta}_{t|t-1} - \frac{P_{t|t}}{\sigma_t^2} \left(x_t(\hat{\theta}_{t|t-1}^\top x_t - y_t) \right) .$$

La différence entre les deux méthodes se situe donc au niveau du pas de gradient, c'est un scalaire dans le cas de l'OGD, et une matrice dans le cas du filtre de Kalman. Ainsi l'OGD effectue un pas dans la direction opposée à celle du gradient, alors que le filtre de Kalman utilise une matrice de *pré-conditionnement* pour transformer la direction du gradient. Ainsi on peut interpréter le filtre de Kalman comme un algorithme de descente de gradient de second ordre, que l'on peut rapprocher des méthodes de Newton (la matrice P_t sera proche en un certain sens de la hessienne de la perte). Alors que l'OGD nécessite de choisir γ_t, le filtre de Kalman apprend une matrice $P_{t|t}$ à partir des hyper-paramètres σ_t^2 et Q_t, que l'on peut donc voir comme une paramétrisation du gradient. Q_t est la variance du bruit de la marche aléatoire de l'état et peut donc être vue comme la vitesse d'évolution du système. Plus Q_t est grande, plus le système est perturbé au cours du temps. Observons dans le théorème 1.1 que la matrice $P_{t|t}$ est plus grande pour Q_t grand, et le pas de gradient effectué est donc plus grand, ce qui est bien ce que l'on souhaite.

Cette double interprétation du filtre de Kalman comme une méthode bayésienne (estimation de la distribution *a posteriori* de l'état) et comme un algorithme de descente de gradient en ligne peut être généralisée à l'EKF, comme l'a noté Ollivier, 2018.

Nous gardons cette vision tout au long de la thèse. La première partie sur l'analyse de l'EKF dans le cas statique fait le lien avec les algorithmes de descente de gradient dont le pas décroit au cours du temps ($\gamma_t \to 0$). Puis dans la seconde nous considérons le cas où $Q_t \succcurlyeq 0$, et nous voyons le choix des variances comme un problème d'estimation du pas optimal dans une descente de gradient.

1.3 L'optimisation stochastique, un modèle espace-état statique

La première contribution de cette thèse est de renforcer le lien entre les statistiques bayésiennes et l'optimisation en s'appuyant sur le parallèle présenté en section 1.2.3. Nous étudions l'EKF dans le cadre statique, c'est le but de la partie I.

1.3.1 Cadre des modèles linéaires généralisés

Pour notre analyse nous nous restreignons aux fonctions de perte qui s'écrivent comme l'opposé de la log-vraisemblance d'un modèle linéaire généralisé (McCullagh et Nelder, 1989). Formellement, nous supposons que la perte est de la forme $\ell(y, \theta^\top x) = -\log p_\theta(y \mid x)$, et que p_θ appartient à une classe de la famille exponentielle paramétrée ainsi :

$$p_\theta(y \mid x) = c(y) \exp\left(\frac{y\theta^\top x - b(\theta^\top x)}{a} \right) ,$$

où a est une constante et b et c sont des fonctions univariées. Ce modèle inclue notamment la régression gaussienne, la régression logistique (voir l'exemple ci-dessous) et la régression de Poisson. Notre analyse utilise des hypothèses supplémentaires sur la perte ℓ, notamment la convexité.

Exemple (Regression Logistique). On considère la classification binaire de $y \in \{-1, 1\}$, et l'on modélise $\mathcal{L}(y \mid x)$ ainsi :

$$p_\theta(y \mid x) = \frac{1}{1 + e^{-y\theta^\top x}} = \exp\left(\frac{y\theta^\top x - (2\log(1 + e^{\theta^\top x}) - \theta^\top x)}{2}\right) .$$

Alors la perte est $\ell(y, \theta^\top x) = \log(1 + e^{-y\theta^\top x})$.

Une propriété notable des distributions de la famille exponentielle est la forme explicite de leur espérance et de leur variance. Avec nos notations nous avons $\mathbb{E}[y \mid \theta^\top x] = b'(\theta^\top x)$ et $Var[y \mid \theta^\top x] = ab''(\theta^\top x)$, où b' et b'' sont les deux premières dérivées de la fonction b. Ainsi nous considérons le modèle espace-état

$$\begin{aligned} &\text{Etat :} && \theta_t = \theta_{t-1} , \\ &\text{Espace :} && y_t = b'(\theta_t^\top x_t) + \varepsilon_t , \end{aligned}$$

pour lequel les formules de mise à jour deviennent

$$\begin{aligned} &P_{t|t-1} = P_{t-1|t-1} , && \hat\theta_{t|t-1} = \hat\theta_{t-1|t-1} , \\ &P_{t|t} = P_{t|t-1} - \frac{b''(\hat\theta_{t|t-1}^\top x_t) P_{t|t-1} x_t x_t^\top P_{t|t-1}}{b''(\hat\theta_{t|t-1}^\top x_t) x_t^\top P_{t|t-1} x_t + a} , && \hat\theta_{t|t} = \hat\theta_{t|t-1} - \frac{P_{t|t}}{a}\left(x_t(b'(\hat\theta_{t|t-1}^\top x_t) - y_t)\right) . \end{aligned}$$

Puis, la formule de Sherman et Morrison donne

$$P_{t|t}^{-1} = P_{t-1|t-1}^{-1} + \frac{b''(\hat\theta_{t|t-1}^\top x_t) x_t x_t^\top}{a} .$$

Enfin, en notant ℓ' et ℓ'' les deux premières dérivées de ℓ par rapport à la seconde variable, puis définissant $P_t = P_{t|t-1}$ et $\hat\theta_t = \hat\theta_{t|t-1}$, nous obtenons

$$P_{t+1}^{-1} = P_t^{-1} + \ell''(y_t, \hat\theta_t^\top x_t) x_t x_t^\top , \qquad \hat\theta_{t+1} = \hat\theta_t - P_{t+1}\left(x_t \ell'(y_t, \hat\theta_t^\top x_t)\right) . \tag{1.4}$$

Cette dernière récursion donne l'écriture suivante pour P_{t+1} :

$$P_{t+1} = \left(P_1^{-1} + \sum_{s=1}^{t} \ell''(y_s, \hat\theta_s^\top x_s) x_s x_s^\top\right)^{-1} .$$

Nous avons intuitivement une décroissance de P_t en $1/t$, et l'EKF est donc proche d'une descente de gradient avec un pas de gradient proportionnel à $1/t$. Cependant nous avons une matrice de pré-conditionnement au lieu d'un pas de gradient scalaire. Comme $\ell''(y_s, \hat\theta_s^\top x_s) x_s x_s^\top$ est la hessienne de la perte à l'instant s, sous réserve que $\hat\theta_t$ converge, P_{t+1} devrait être proche de H^{-1}/t, où H est l'espérance de la hessienne de la perte à la limite de $\hat\theta_t$.

1.3.2 Description des résultats

Nous catégorisons les garanties sur des algorithmes d'optimisation en deux types. Dans le cadre *adversarial*, aucune hypothèse n'est faite sur le processus de génération des données et (x_t, y_t) peut être défini par un adversaire, autrement dit l'objectif est une analyse dans le pire des cas. L'objectif est de borner le regret $\sum_{t=1}^n \ell(y, \hat{\theta}_t^\top x_t) - \ell(y, \theta^{\star\top} x_t)$, différence entre la perte subie et la perte subie par l'oracle constant.

Inversement, dans le cadre *stochastique*, les données (x_t, y_t) sont supposées indépendantes et identiquement distribuées, puis l'on définit le risque $L(\theta) = \mathbb{E}[\ell(y, \theta^\top x)]$. L'objectif est alors de minimiser ce risque.

Nous nous plaçons dans un cadre intermédiaire et nous obtenons des garanties sur le risque cumulé $\sum_{t=1}^n L(\hat{\theta}_t) - L(\theta^\star)$ où $\theta^\star$ minimise le risque.

Dans un premier temps, nous obtenons un résultat sous une hypothèse très forte de convergence de l'EKF, définie ci-dessous.

Hypothèse (Localisée)**.** *Définissons* $\tau(\zeta) = \min\{k \in \mathbb{N} \mid \forall t > k, \|\hat{\theta}_t - \theta^\star\| \leq \zeta\}$ *pour tout* $\zeta > 0$. *Pour tout* $\delta, \zeta > 0$*, il existe* $T(\zeta, \delta) \in \mathbb{N}$ *tel que* $\mathbb{P}\big(\tau(\zeta) \leq T(\zeta, \delta)\big) \geq 1 - \delta$.

L'hypothèse consiste à supposer que, à partir d'un certain rang, avec grande probabilité, l'estimateur de l'EKF est piégé dans une boule de rayon ζ arbitrairement petit autour de $\theta^\star$. Nous prouvons cette propriété ensuite dans les cadres quadratique et logistique.

Théorème 1.3. *En partant de* $\hat{\theta}_1 \in \mathbb{R}^d$ *et* $P_1 \succ 0$*, sous certaines hypothèses dont l'hypothèse localisée, pour tout* $\delta > 0$*, nous avons simultanément pour* $n \geq 1$

$$\sum_{t=T(\zeta,\delta)+1}^{T(\zeta,\delta)+n} L(\hat{\theta}_t) - L(\theta^\star) \leq C(\log n + \log \delta^{-1}) \,,$$

avec probabilité au moins $1 - 3\delta$.

Structure de la preuve. Nous décomposons la preuve en trois étapes.

1. Le point de départ est une borne dans le cadre *adversarial* sur le développement du regret à l'ordre 2 :

$$\sum_{t=1}^n \left(\Big(\ell'(y_t, \hat{\theta}_t^\top x_t) x_t\Big)^\top (\hat{\theta}_t - \theta^\star) - \frac{1}{2} (\hat{\theta}_t - \theta^\star)^\top \Big(\ell''(y_t, \hat{\theta}_t^\top x_t) x_t x_t^\top\Big) (\hat{\theta}_t - \theta^\star) \right) = O(\log n) \,. \tag{1.5}$$

   Cette borne présentée dans le lemme 4.2 est obtenue directement depuis les formules récursives (1.4) et tient donc sans hypothèse sur (x_t, y_t). Puis notre travail consiste à faire le lien entre cette borne et une borne sur le risque cumulé. Ce sont les étapes 2 et 3.

2. Le problème de la borne précédente est qu'elle concerne le développement d'ordre 2 de la perte, mais on ne sait pas borner la perte par cette expression. Nous passons alors au risque L (espérance de la perte ℓ), et nous montrons que nous pouvons borner le risque par un développement d'ordre 2. Précisément, nous comparons les termes d'ordre 1 et 2 (proposition 4.1). Pour tout $\rho < 1$, il existe un voisinage V_ρ de $\theta^\star$ tel que pour tout $\theta \in V_\rho$,

$$\frac{\partial L}{\partial \theta}\Big|_\theta^\top (\theta - \theta^\star) \geq \rho (\theta - \theta^\star)^\top \frac{\partial^2 L}{\partial \theta^2}\Big|_\theta^\top (\theta - \theta^\star) \,.$$

En tirant parti de cette propriété et en utilisant la convexité de la perte (et donc du risque) nous obtenons la borne d'ordre 2 suivante sur le risque (proposition 4.2) : pour tout $\theta \in V_\rho$ et $0 < c < \rho$,

$$L(\theta) - L(\theta^\star) \leq \frac{\rho}{\rho - c} \left(\frac{\partial L}{\partial \theta}\Big|_\theta^\top (\theta - \theta^\star) - c(\theta - \theta^\star)^\top \frac{\partial^2 L}{\partial \theta^2}\Big|_\theta (\theta - \theta^\star) \right) . \tag{1.6}$$

3. Enfin nous relions les équations (1.5) et (1.6). Pour ce faire nous étudions la différence entre les termes d'ordre 1 et 2 du développement de Taylor de la perte et ceux du développement du risque. Nous étudions donc dans le lemme 4.1 la martingale définie ainsi :

$$\Delta M_t = \sum_{t=1}^{n} \left(\frac{\partial L}{\partial \theta}\Big|_\theta - \ell'(y_t, \hat{\theta}_t^\top x_t) x_t \right)^\top (\hat{\theta}_t - \theta^\star) .$$

□

Ce résultat nous permet d'avoir une borne optimale sur le risque cumulé, mais sous l'hypothèse de localisation. Nous prouvons cette hypothèse dans deux cas. Pour la perte quadratique nous appliquons les résultats de Hsu, Kakade et Zhang, 2012. Dans le cas logistique, nous prouvons la propriété de convergence pour un EKF statique légèrement modifié à la manière de Bercu, Godichon et Portier, 2020 :

Proposition 1.1. *Rappelons la perte logistique* $\ell(y, \theta^\top x) = \log(1 + e^{-y\theta^\top x})$. *Soit* $0 < \beta < \frac{1}{2}$. *Définissons l'algorithme suivant :*

$$P_{t+1}^{-1} = P_t^{-1} + \max\left(\ell''(y_t, \hat{\theta}_t^\top x_t), \frac{1}{t^\beta}\right) x_t x_t^\top , \qquad \hat{\theta}_{t+1} = \hat{\theta}_t - P_{t+1}\left(x_t \ell'(y_t, \hat{\theta}_t^\top x_t)\right) ,$$

avec les mêmes notations que l'EKF statique. Alors cet algorithme que l'on appelle tronqué vérifie la propriété localisée, et de plus sa récursion coïncide avec celle de l'EKF statique à partir d'un certain rang.

Formellement, en gardant la notation pour $\tau(\zeta)$ *dans le cadre de cet algorithme tronqué, nous avons pour tout* $\delta, \zeta > 0$, *l'existence d'un* $T(\zeta, \delta)$ *défini explicitement tel qu'avec probabilité au moins* $1 - \delta$,

$$\tau(\zeta) \leq T(\zeta, \delta) ,$$
$$\forall t \geq T(\zeta, \delta),\ \ell''(y_t, \hat{\theta}_t^\top x_t) \geq \frac{1}{t^\beta} .$$

La coïncidence avec l'EKF statique est cruciale car elle permet d'appliquer l'analyse locale à partir de $T(\zeta, \delta)$. Ainsi, les $T(\zeta, \delta)$ premiers termes sont traités à part, puis le risque cumulé est borné grâce au Théorème 1.3.

Structure de la preuve. Nous décomposons la preuve en trois étapes.

1. Le seuil $\frac{1}{t^\beta}$ est introduit dans un objectif clair, celui de bien contrôler P_t. En effet, on peut facilement borner P_t inférieurement ($P_t \succcurlyeq cI/t$). À l'inverse, on ne peut pas obtenir de borne supérieure du type $P_t \preccurlyeq cI/t$ lorsque $\ell''(y_t, \hat{\theta}_t^\top x_t)$ peut être arbitrairement petit, ce qui est le cas de la régression logistique. Ainsi, le seuil nous permet d'obtenir le contrôle suivant dans la proposition 4.4 : pour tout $\delta > 0$, il existe $T_1(\delta)$ tel qu'avec probabilité au

moins $1-\delta$,

$$\forall t > T_1(\delta), \qquad P_t \preccurlyeq \frac{4}{\Lambda_{\min} t^{1-\beta}} I\,,$$

avec $\Lambda_{\min}$ la plus petite valeur propre de $\mathbb{E}[xx^\top]$. Une hypothèse est donc que cette matrice soit inversible.

2. En repartant de l'équation de mise à jour de $\hat{\theta}_t$ et en utilisant le fait que $\ell'' \leq \frac{1}{4}$, nous obtenons la récursion suivante sur le risque :

$$L(\hat{\theta}_{t+1}) \leq L(\hat{\theta}_t) - \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^\top P_t\Big(\ell'(y_t, \hat{\theta}_t^\top x_t)x_t\Big) + 2D_X^4 \lambda_{\max}(P_t)^2\,. \tag{1.7}$$

De cette récursion, Bercu, Godichon et Portier, 2020 obtiennent la convergence presque sure de $\hat{\theta}_t$ vers $\theta^\star$ en appliquant le théorème de Robbins-Siegmund. En effet, ce résultat est intuitif car grâce au contrôle obtenu sur P_t nous avons

$$\sum_t (2D_X^4 \lambda_{\max}(P_t)^2) < \infty\,,$$

et le terme du milieu permet une décroissance en espérance du risque que l'on peut borner inférieurement par $\frac{1}{t}\|\frac{\partial L}{\partial \theta}|_{\hat{\theta}_t}\|^2$.

3. Pour obtenir notre borne non asymptotique, nous estimons la probabilité d'avoir un risque loin du risque optimal (à une distance supérieure à un certain $\eta > 0$). Pour ce faire, nous utilisons le fait que les variations de l'algorithme sont lentes, et nous regardons la dernière itération (si elle existe) telle que le risque soit proche du risque optimum (à une distance inférieure à $\eta/2$).

   Formellement, nous définissons $B_{k,t}$ l'événement $(\forall k < s < t, L(\hat{\theta}_s) - L(\theta^\star) > \eta/2)$, et nous pouvons utiliser la loi des probabilités totales :

$$\begin{aligned}\mathbb{P}(L(\hat{\theta}_t) - L(\theta^\star) > \eta) = \mathbb{P}\Big((L(\hat{\theta}_t) - L(\theta^\star) > \eta) \cap B_{0,t}\Big) \\ + \sum_{k=1}^{t-1} \mathbb{P}\Big((L(\hat{\theta}_t) - L(\theta^\star) > \eta) \cap \big(L(\hat{\theta}_k) - L(\theta^\star) \leq \frac{\eta}{2}\big) \cap B_{k,t}\Big)\,.\end{aligned}$$

   Afin d'estimer ces probabilités, nous itérons l'équation (1.7) :

$$L(\hat{\theta}_t) - L(\hat{\theta}_k) \leq \sum_{s=k}^{t-1} \left(\Delta M_s - \lambda_{\min}(P_s)\left\|\frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_s}\right\|^2 + 2D_X^4 \lambda_{\max}(P_s)^2\right)\,,$$

   où (ΔM_t) est une différence de martingale. Alors nous séparons les différents k en deux, ce que nous illustrons par la figure 1.2. Pour k suffisamment petit par rapport à t, la décroissance en espérance du risque rend improbable de rester longtemps loin de l'optimum. Pour k plus proche de t, le contrôle de P_t permet de borner la probabilité que l'algorithme se soit éloigné de l'optimum en $t-k$ étapes.

□

Dans le chapitre 3 nous présentons une version plus simple de cette preuve de convergence pour l'algorithme *stochastic gradient descent* à pas décroissant. Cet algorithme nous permet de

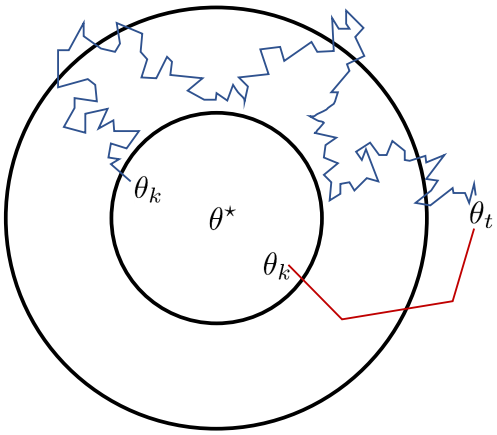


FIGURE 1.2 – Illustration de la preuve de convergence de l'EKF ou de SGD. La trajectoire bleue est peu probable car l'algorithme est loin de l'optimum pendant longtemps. La trajectoire rouge est peu probable car l'algorithme s'éloigne de l'optimum rapidement.

traiter une classe de fonctions L plus large. Comme le pas de gradient utilisé est un scalaire au lieu d'une matrice de pré-conditionnement, nous n'avons pas besoin de l'étape 1 de la preuve. De même que pour l'EKF statique, nous établissons une version non-asymptotique de la preuve de convergence établie par ROBBINS et MONRO, 1951.

1.4 Choix des variances dans un modèle espace-état

Nous n'avons pas introduit le modèle espace-état pour ne considérer que le cas statique, mais plutôt pour le cadre dynamique. Rappelons le modèle espace-état linéaire gaussien qui nous intéresse particulièrement :

$$\text{Etat :} \qquad \theta_t = \theta_{t-1} + \eta_t\,, \qquad \eta_t \sim \mathcal{N}(0, Q_t)\,, \tag{1.8}$$

$$\text{Espace :} \qquad y_t = \theta_t^\top x_t + \varepsilon_t\,, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)\,. \tag{1.9}$$

Le filtre de Kalman donne une estimation exacte et récursive de l'état, à variances connues Q_t, et σ_t^2, voir la section 1.2.2. Ces variances sont les *hyper-paramètres* du modèle espace-état, et ne sont pas connues dans la plupart des applications. Il n'y a pas de consensus quant à leur choix.

Nous proposons dans la partie II différentes approches, que nous divisons en deux paradigmes : ou bien nous considérons que les variances sont constantes au cours du temps et nous les estimons sur un historique d'entraînement, ou bien nous ne les supposons pas constantes et nous les estimons dynamiquement. Pour reprendre le parallèle du filtre de Kalman avec un algorithme de gradient (1.2.3), nous pouvons voir le cas statique comme un pas de gradient convergeant vers 0, le cas dynamique à variances constantes est proche d'un pas de gradient adaptatif tel Adam avec pas constant (KINGMA et BA, 2014) et le cas dynamique à variances dynamiques est une étape supplémentaire d'adaptation.

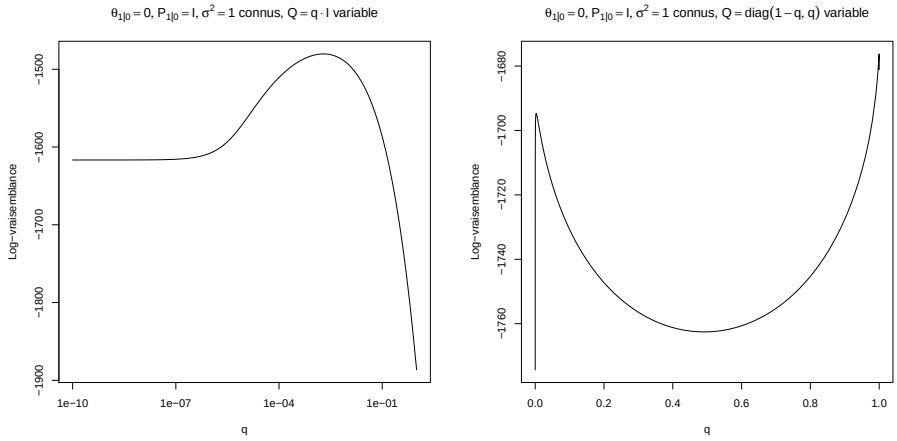


FIGURE 1.3 – Log-vraisemblance dans un cadre bien spécifié en dimension 2. Nous affichons la log-vraisemblance en fonction de Q, en fixant les autres hyper-paramètres. À gauche : $Q = qI_2$. À droite : $Q = diag(1-q, q)$.

1.4.1 Variances constantes

Le choix plus courant dans la littérature que nous avons considérée consiste à supposer que les variances sont constantes au cours du temps (BROCKWELL et DAVIS, 2016 ; DURBIN et KOOPMAN, 2012 ; FAHRMEIR et TUTZ, 2013). Formellement, les variances des équations (1.8) et (1.9) sont alors $Q_t = Q$ et $\sigma_t^2 = \sigma^2$.

Dans ce cadre, l'objectif qui fait consensus est de maximiser la vraisemblance sur un certain jeu de données $(x_t, y_t)_{1 \leq t \leq n}$. Pour cela, le principal algorithme est l'*Expectation-Maximization* (EM), dans lequel on alterne entre deux étapes :

1. **Expectation** : à variances fixées, on estime les *paramètres* $(\hat{\theta}_{t|n}, P_{t|n})_t$ par le filtre de Kalman (théorème 1.1) et le lissage de Kalman (théorème 1.2). Puis on en déduit l'espérance de la log-vraisemblance complète comme une fonction de Q et σ^2.
2. **Maximization** : à paramètres fixés, on estime les *hyper-paramètres* Q et σ^2 en maximisant l'espérance de la log-vraisemblance complète.

Dans le cadre du modèle linéaire gaussien, ces deux étapes admettent des formes closes : à hyper-paramètres fixés, l'estimation des paramètres est exacte, et à paramètres fixés, il en est de même de l'estimation des hyper-paramètres qui maximisent l'espérance de la log-vraisemblance complète. De plus, cette procédure itérative présente une garantie attirante : à chaque étape la vraisemblance croît.

L'algorithme EM présente cependant deux inconvénients qui ne sont pas négligeables. D'une part, c'est un algorithme coûteux qui converge lentement. D'autre part, s'il garantit la convergence vers un maximum local de la vraisemblance, il ne converge pas vers un maximum global. En effet, la log-vraisemblance n'est pas nécessairement une fonction concave, voir la figure 1.3.

Une alternative que nous proposons consiste en une maximisation de la vraisemblance par une *recherche itérative sur une grille* sur Q dans laquelle nous nous restreignons aux matrices diagonales. Cette restriction revient à supposer que les coefficients de θ_t évoluent indépendamment les uns des autres et nous parait une restriction raisonnable sur le modèle. Notons que cette hypothèse d'évolution indépendante des coefficients ne se traduit pas en une évolution indépendante des coefficients de l'estimateur $\hat{\theta}_t$. Comme son nom l'indique, nous optimisons les

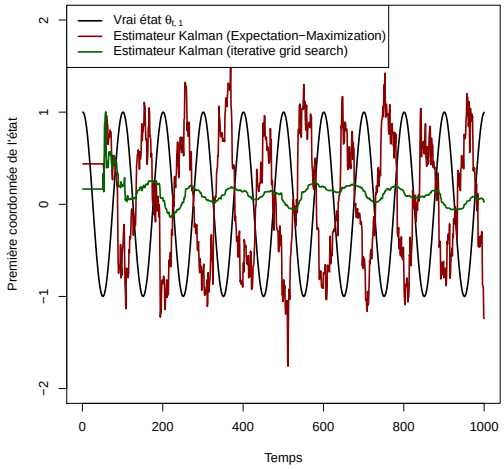


FIGURE 1.4 – Données synthétiques dans un cadre mal spécifié. Nous utilisons une dimension $d = 2$, un état déterministe $\theta_t = \cos(\frac{2\pi t}{100}) \cdot (1, 1)^\top$, puis $x_t \sim \mathcal{N}(0, I)$ et $y_t - \theta_t^\top x_t \sim \mathcal{N}(0, 1)$. Le délai est $k = 50$, ce qui signifie que nous sommes en opposition de phase. Le mieux est de varier très peu.

coefficients diagonaux de Q sur une grille, et nous cherchons de manière itérative donc non exhaustive. À chaque étape nous calculons la vraisemblance d'un certain nombre de matrices Q ayant un seul coefficient différent de celle de l'itération précédente, et nous gardons celle qui obtient la vraisemblance la plus grande. Nous n'obtenons pas non plus de garantie de convergence vers le maximum global, mais nous avons constaté de meilleurs résultats en pratique.

Il nous semble que ces meilleurs résultats proviennent d'une plus grande robustesse à deux phénomènes provenant du monde réel. Premièrement, en général, le modèle espace-état linéaire gaussien à variances constantes est mal spécifié. Cela signifie que les données ne suivent pas réellement le modèle, que les variances ne sont pas constantes par exemple. Cela ne remet pas en cause le modèle qui approche bien les données, l'interprétation comme un algorithme de gradient en section 1.2.3 justifie que l'algorithme soit robuste, mais cela justifie qu'une méthode plus empirique puisse mieux fonctionner.

Deuxièmement, la recherche itérative sur une grille permet d'introduire un délai de disponibilité des données, qui existe dans la plupart des applications. En effet, dans le cas du réseau électrique nous ne savons pas instantanément quelle est la consommation, elle est estimée petit à petit. Nous disposons d'estimations fiables au bout de quelques jours, et les valeurs finales consolidées ne sont publiées par RTE que des mois plus tard. Ce délai signifie formellement que pour prévoir y_t, nous disposons des variables x_t et des observations $x_1, y_1, \ldots, x_{t-k}, y_{t-k}$ où k est le délai. Notre approche permet d'optimiser une fonction proche de la vraisemblance qui en tient compte, et nous évitons ainsi une forme de sur-apprentissage. Ce n'est pas le cas de l'EM, comme nous le verrons en section 5.4, et nous illustrons notre propos en figure 1.4 avec un exemple jouet.

Enfin, la simplicité de la méthode introduite implique sa généralité. L'algorithme peut être appliqué à n'importe quel modèle espace-état et n'importe quelle variante du filtre de Kalman.

1.4.2 Variances dynamiques

Un second paradigme consiste à estimer les variances d'un modèle espace-état au cours du temps, c'est ce qui a été fréquemment appelé *adaptive Kalman filtering* (Mehra, 1972). Dans le chapitre 6, nous développons une nouvelle méthode d'estimation conjointe de l'état et des variances que nous appelons Viking.

Formellement, notons $\mathcal{F}_t = \sigma(x_1, y_1, \dots, x_t, y_t)$ la filtration naturelle. Nous cherchons à appliquer une méthode bayésienne. Nous partons d'un prior $p(\theta_0, \sigma_0^2, Q_0 \mid \mathcal{F}_0)$ et nous supposons un modèle $p(\theta_t, \sigma_t^2, Q_t \mid \theta_{t-1}, \sigma_{t-1}^2, Q_{t-1})$ sur la dynamique de ces trois variables. À l'instant t, nous appliquons une étape de prévision (suivant la dynamique du modèle), et une étape de filtrage (loi de Bayes) :

$$\begin{aligned} &\text{Prévision :} && p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_{t-1})\,, \\ &\text{Filtrage :} && p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_t)\,. \end{aligned}$$

Cependant, il n'y a pas de distribution paramétrique naturelle sur la loi jointe telle que la distribution *a posteriori* reste dans la classe de distributions considérée. Nous appliquons alors l'approche *Variational Bayes* (VB) (Šmídl et Quinn, 2006). Cela consiste à approcher la distribution jointe par une distribution produit dont chaque loi marginale a une forme simple.

Nous utilisons une loi gaussienne pour l'état afin de coïncider avec la distribution exacte *a posteriori* dans le cas particulier où les variances sont connues. Nous introduisons alors des distributions paramétriques sur σ_t^2 et Q_t de la forme $P_{\Phi_{t|t}}$ et $P_{\Psi_{t|t}}$, de densités $p_{\Phi_{t|t}}$ et $p_{\Psi_{t|t}}$, où $\Phi_{t|t}$ et $\Psi_{t|t}$ sont les paramètres. Nous estimons alors $\hat{\theta}_{t|t}, P_{t|t}, \Phi_{t|t}, \Psi_{t|t}$ tels que la distribution produit $\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times P_{\Phi_{t|t}} \times P_{\Psi_{t|t}}$ soit la "meilleure" approximation de la distribution *a posteriori* notée $P_{\mathcal{F}_t}$. Formellement, nous minimisons la divergence de Kullback-Leibler :

$$KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times P_{\Phi_{t|t}} \times P_{\Psi_{t|t}} \mid\mid P_{\mathcal{F}_t}\Big),$$

où $KL(P \mid\mid Q) = \int_x \log(p(x)/q(x))p(x)dx$ pour toutes distributions P et Q de densités p et q. Dans l'approche VB nous avons donc à résoudre à chaque étape un problème d'optimisation en trois distributions.

L'étape de prévision est déterminée par la dynamique du modèle supposé. Notons $\mathcal{N}(x \mid \mu, \Sigma)$ la densité de la distribution gaussienne $\mathcal{N}(\mu, \Sigma)$ au point x. Avec les hypothèses adéquates, partant d'un prior

$$p(\theta_{t-1}, \sigma_{t-1}^2, Q_{t-1} \mid \mathcal{F}_{t-1}) = \mathcal{N}(\theta_{t-1} \mid \hat{\theta}_{t-1|t-1}, P_{t-1|t-1}) p_{\Phi_{t-1|t-1}}(\sigma_{t-1}^2) p_{\Psi_{t-1|t-1}}(Q_{t-1})\,,$$

nous obtenons naturellement la densité ci-dessous :

$$p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_{t-1}) = \mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^\top + Q_t) p_{\Phi_{t|t-1}}(\sigma_t^2) p_{\Psi_{t|t-1}}(Q_t)\,,$$

où nous avons une forme simple pour $\Phi_{t|t-1}$ et $\Psi_{t|t-1}$. Cette étape de prévision est introduite comme un prior dans l'étape de filtrage, qui donne la densité suivante pour la distribution posterior :

$$\begin{aligned} p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_t) = \frac{p(x_t, \mathcal{F}_{t-1})}{p(\mathcal{F}_t)} &\mathcal{N}(y_t \mid \theta_t^\top x_t, \sigma_t^2) \\ &\mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^\top + Q_t) p_{\Phi_{t|t-1}}(\sigma_t^2) p_{\Psi_{t|t-1}}(Q_t)\,. \end{aligned}$$

Notons que quelles que soient les distributions introduites pour σ_t^2 et Q_t, la distribution jointe *a posteriori* de l'état et des variances ne peut être factorisée. Il est donc naturel d'appliquer l'approche VB pour estimer la distribution *a posteriori* avec une forme simple. De plus, le terme croisé entre θ_t et Q_t dans la densité précédente nous empêche d'appliquer la méthode proposée par Tzikas, Likas et Galatsanos, 2008 pour obtenir une forme analytique de minimisation de la divergence KL. Nous obtenons des bornes supérieures de la divergence KL, dont la minimisation admet des formes closes. Minimiser ces bornes supérieures ne garantit pas de minimiser la divergence KL mais de la réduire à chaque étape, c'est la règle de l'*evidence-lower bound* (ELBO).

Dans ce bref aperçu de l'approche VB, nous avons traité des distributions paramétriques générales pour σ_t^2 et Q_t. Détaillons enfin notre choix de ces distributions, qui donne lieu à l'algorithme Viking. L'estimation récursive de la distribution *a posteriori* ne motive pas de distributions naturelles qui simplifient l'optimisation de la divergence KL. Nous choisissons alors de représenter les variances par des variables gaussiennes. Un avantage notable d'une variable latente gaussienne est d'introduire naturellement une dynamique sous la forme d'une marche aléatoire. Cependant, comme les variances doivent être positives, nous transformons ces variables latentes. Précisément, nous utilisons $\sigma_t^2 = \exp(a_t)$ (loi log-normale) et $Q_t = f(b_t)$, où a_t, b_t suivent des distributions gaussiennes. La fonction f est un paramètre de l'algorithme dont nous proposons différents choix possibles. Grâce à la représentation gaussienne des différentes variables latentes, l'approche est résumée par le modèle ci-dessous :

$$
\begin{aligned}
&\theta_0 \sim \mathcal{N}(\hat{\theta}_0, P_0)\,, \quad a_0 \sim \mathcal{N}(\hat{a}_0, s_0)\,, \quad b_0 \sim \mathcal{N}(\hat{b}_0, \Sigma_0)\,,\\
&a_t - a_{t-1} \sim \mathcal{N}(0, \rho_a)\,, \quad b_t - b_{t-1} \sim \mathcal{N}(0, \rho_b I)\,,\\
&\theta_t - K\theta_{t-1} \sim \mathcal{N}(0, f(b_t))\,,\\
&y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \exp(a_t))\,,
\end{aligned}
$$

dans lequel nous introduisons les paramètres (positifs ou nuls) ρ_a et ρ_b qui régissent la dynamique des variables a_t et b_t, représentant les variances σ_t^2 et Q_t.

1.5 Application à la prévision de consommation électrique

Dans la partie III, nous appliquons des modèles espace-état pour la prévision de consommation d'électricité dans différents pays et à différentes échelles. Dans un cadre il s'agit de prévision de consommation nette, différence entre la consommation et la production non pilotable. Nous introduisons en section 1.5.1 une méthode générique pour se rapporter à un modèle linéaire gaussien. En section 1.5.2 nous présentons la prévision *en moyenne* (ou en médiane), au sens où nous mesurons une erreur absolue et nous attribuons la même performance à une erreur positive qu'à une erreur négative. Dans le chapitre 10 introduit en section 1.5.3, nous estimons les différents quantiles de la consommation, c'est la prévision *probabiliste*.

1.5.1 Utilisation du modèle linéaire gaussien

Nous résumons la méthode par un schéma en figure 1.5. Plus précisément, dans chaque problème de prévision de série temporelle, nous avons des variables explicatives et nous cherchons à prévoir une variable d'intérêt. Notre approche se synthétise en 4 étapes, la troisième étant le coeur de la thèse :

1. **Pré-traitement**. Nous nettoyons les données en sélectionnant les variables explicatives intéressantes. De plus, nous pré-calculons des quantités d'intérêt comme des lissages exponentiels de la température, nous corrigeons des prévisions météorologiques ...

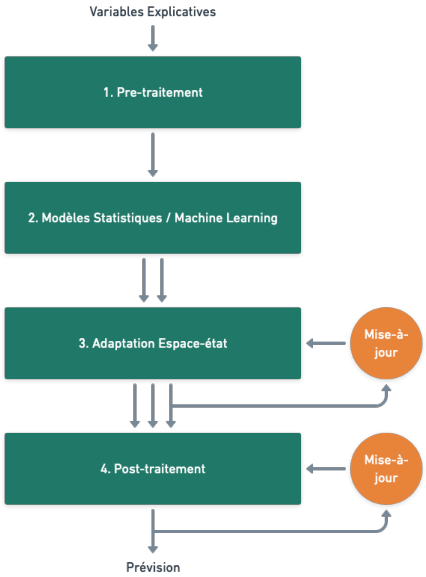


FIGURE 1.5 – Schéma de la méthode employée pour la prévision de consommation électrique. La démultiplication des flèches indique une augmentation du nombre de modèles due à une étape, et tous ces modèles sont fusionnés finalement dans la dernière étape pour ne donner qu'une unique prévision.

2. **Modèles Statistiques / Machine Learning**. Nous utilisons des modèles classiques pour la prévision de séries temporelles : l'auto-régressif, la régression linéaire, le modèle additif généralisé (GAM) et le *multi-layer perceptron* (MLP). Dans la plupart de nos applications ces modèles sont calibrés par heure de la journée, s'il s'agit de prévision horaire nous avons donc 24 modèles différents, chacun étant adapté à une heure précise de la journée. Cette étape nous donne déjà potentiellement plusieurs choix possibles.
3. **Adaptation Espace-état**. Nous adaptons les différents modèles obtenus à l'étape 2 par le modèle espace-état linéaire gaussien, avec les différentes estimations des variances introduites en partie II. Pour ce faire nous linéarisons les modèles, car le modèle espace-état n'a pas de sens dans le cas, par exemple, de l'adaptation d'un réseau de neurones. Pour le GAM, nous figeons les effets non linéaires et nous adaptons une combinaison linéaire de ces effets, linéaires et non linéaires. Pour le MLP, nous figeons les couches les plus profondes et nous n'adaptons que la dernière. Autrement dit, nous utilisons l'étape 2 pour apprendre de nouveaux *features* que nous utilisons comme variables explicatives dans un modèle espace-état linéaire.

   Formellement, nous cherchons à prévoir une quantité y_t à l'aide de variables explicatives x_t. L'étape 2 nous permet d'obtenir de nouvelles covariables qui définissent un vecteur $f(x_t)$, et nous obtenons le modèle espace-état suivant :

$$\begin{aligned} &\text{Etat :} && \theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t)\,, \\ &\text{Espace :} && y_t - \theta_t^\top f(x_t) \sim \mathcal{N}(0, \sigma_t^2)\,. \end{aligned}$$

   Nous testons alors les différentes méthodes d'estimation de σ_t^2 et Q_t. Nous utilisons le

cas dégénéré statique ($Q_t = 0$), le paradigme où les variances sont constantes, et le cadre de variances dynamiques. Les données incluant la période de crise du coronavirus présentaient une telle rupture que nous avons introduit un cadre incluant cette rupture. Nous la modélisons comme une matrice de covariance Q_T grande à l'instant T de la rupture, en comparaison avec les valeurs de Q_t en dehors de cette rupture.

4. **Post-traitement**. Nous devons transformer les prévisions de l'étape 3 pour obtenir une unique prévision, et pour ce faire nous avons souvent eu recours à l'agrégation d'experts, détaillée dans la thèse de Gaillard, 2015. Nous avons des prévisions (experts) $\hat{y}_{t,1}, \dots \hat{y}_{t,K}$ qui proviennent de l'étape 3 et notre prévision finale s'écrit comme une moyenne de ces prévisions, $\sum_k p_{t,k}\hat{y}_{t,k}$. Nous utilisons une moyenne convexe, *i.e.* $\sum_k p_{t,k} = 1$. Les poids $p_{t,1}, \dots, p_{t,K}$ sont estimés dynamiquement, ils évoluent au cours du temps pour prendre en compte le fait que les performances d'un expert peuvent évoluer au cours du temps.

   Préalablement à l'agrégation, nous utilisons pour l'un des jeux de données une étape de correction infra-journalière, qui permet de prendre en compte la corrélation entre les différentes heures de la journée.

1.5.2 Prévision en moyenne

Nous avons appliqué cette méthode sur les jeux de données détaillés ci-dessous et présentés graphiquement en figure 1.6. Nous évaluons la performances des modèles par différentes fonctions de l'erreur. Les plus classiques sont l'erreur moyenne absolue (*mean absolute error*, MAE) et la racine de l'erreur quadratique moyenne (*root-mean-square error*, RMSE), définie pour des observations $(y_1, \dots y_n)$ et leurs prévisions associées $(\hat{y}_1, \dots \hat{y}_n)$ par

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t| \,, \qquad RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \,.$$

Chapitre 7 : rupture du printemps 2020 en France

L'application première de la thèse est la consommation électrique française. Prévoir la consommation française est en effet un enjeu important pour EDF. Nous étudions donc l'intérêt des méthodes présentées dans la thèse pour ce jeu de données. Nous évaluons sur une période "normale" préalable à Mars 2020, puis nous étudions tout particulièrement la période de crise du coronavirus dont l'impact fut très fort, spécialement lors du premier confinement de Mars à Mai 2020.

Par rapport à un modèle additif généralisé non adaptatif, auquel nous appliquons une correction ARIMA, nous obtenons une réduction de la RMSE de 7% sur une période pré-covid. Puis, pendant le premier mois de confinement, le filtre de Kalman réduit la RMSE de 11%, boosté à 28% en modélisant la rupture. Durant les 2 mois suivants (une période plus stabilisée mais toujours chahutée), nous réduisons la RMSE de 42%. L'agrégation d'experts permet d'augmenter significativement le gain.

Chapitre 8 : compétition sur la prévision post-covid d'une ville

Puis nous avons participé à la compétition *Day-Ahead Electricity Demand Forecasting : Post-COVID Paradigm* (Farrokhabadi, 2020). L'objectif des organisateurs était de mettre au point de nouvelles techniques de prévision robustes à une rupture telle que celle du coronavirus, et le jeu de données était une ville dont la localisation n'était pas donnée. Alors que dans le chapitre 7 nous nous intéressons avant tout à la période de confinement (le coeur de la rupture), dans cette

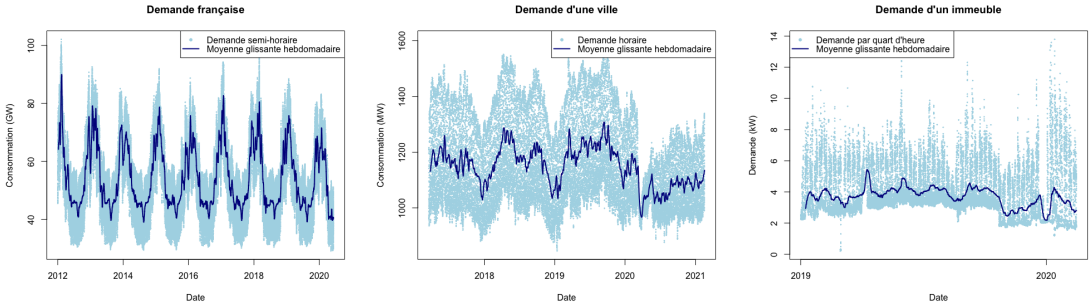


FIGURE 1.6 – Jeux de données utilisés pour la prévision de consommation électrique à différentes échelles : la France (gauche), une ville inconnue (milieu) ou un immeuble de localisation inconnue aussi (droite).

compétition, la période d'évaluation est en Janvier-Février 2021, un an environ après la rupture. Ainsi, alors que la rupture d'un confinement est un événement extrême, la compétition visait une période moins chahutée, et avait pour objectif de définir des modèles de prévision adaptés à l'avenir plutôt que focalisés sur une période passée.

Nous avons remporté cette compétition, améliorant de plus de 30% en MAE la méthode naïve donnée par la compétition (persistance).

Chapitre 9 : compétition sur la prévision d'un immeuble

Nous avons participé à une seconde compétition, appelée *Competition on Building Energy Consumption Forecasting*, dont l'objectif était de prévoir la consommation d'électricité d'un immeuble. Cette compétition était l'occasion pour nous d'appliquer la méthode à une échelle beaucoup plus locale. Il nous semble que les méthodes adaptatives sont nécessaires dans ce cadre car le changement de comportement d'un individu a un impact non marginal, ce qui n'est pas le cas lorsque les données sont agrégées au niveau d'un pays ou d'une grande ville.

Nous avons également remporté cette compétition, en améliorant de plus de 30% en MAE une méthode naïve (persistance).

1.5.3 Prévision probabiliste

Dans le chapitre 10 nous nous intéressons à la prévision probabiliste. Plutôt que de chercher à minimiser l'erreur absolue de notre prévision avec la variable à prévoir, nous cherchons à prévoir ses quantiles. Rappelons que $y_t \in \mathbb{R}$ est la variable à prévoir. Nous cherchons à prévoir $\hat{y}_{t,q} \in \mathbb{R}$ tel que $\mathbb{P}(y_t < \hat{y}_{t,q}) = q$. Nous testons différentes approches.

— Remarquons tout d'abord que le filtre de Kalman fournit une prévision probabiliste par essence. En effet, nous estimons la distribution *a posteriori* de l'état θ_t comme la loi gaussienne $\mathcal{N}(\hat{\theta}_{t|t-1}, P_{t|t-1})$. Or l'état est défini tel que $y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma_t^2)$. Ainsi, nous pouvons prévoir y_t comme une loi gaussienne. Son q-quantile s'écrit alors

$$\hat{y}_{t,q} = \hat{\theta}_{t|t-1}^\top x_t + U_q \sqrt{\sigma_t^2 + x_t^\top P_{t|t-1} x_t}\,,$$

où U_q est le q-quantile de la loi normale centrée réduite.

— Cependant, la propriété de distribution *a posteriori* gaussienne de l'état repose fortement sur le modèle espace-état, et gardons à l'esprit que le monde réel ne vérifie pas l'hypothèse

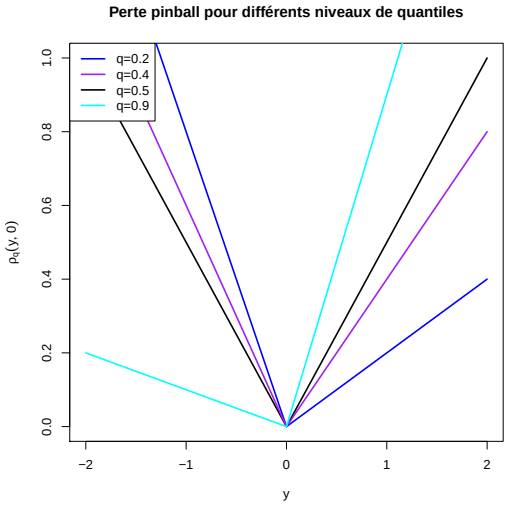


FIGURE 1.7 – Variation de la perte pinball selon le niveau de quantile souhaité pour $\hat{y}_q = 0$ et y variable. Lorsque $q \leq \frac{1}{2}$ la perte est plus grande lorsque y prend des valeurs négatives, et inversement.

de modèle linéaire espace-état de variances connues. Ainsi, il est préférable en pratique d'utiliser les modèles espace-état pour prévoir la consommation moyenne, puis nous modélisons la distribution des résidus (différence $y_t - \hat{y}_t$, où $\hat{y}_t$ est la prévision en moyenne). Plutôt que de chercher à minimiser l'erreur absolue de notre prévision avec la variable à prévoir, nous utilisons une erreur asymétrique qui attribue une perte plus grande à une erreur négative (ou positive). Cette perte est appelée la perte *pinball*, définie comme $\rho_q(y, \hat{y}_q) = (\mathbb{1}_{y<\hat{y}_q} - q)\ (\hat{y}_q - y)$ et représentée en figure 1.7. L'optimisation de cette perte résulte en des prévisions probabilistes, ce qui est justifié par la proposition suivante :

Proposition 1.2. *Soit Y une variable aléatoire à valeurs réelles. Pour tout $0 < q < 1$, notons Y_q le quantile de Y de niveau q. Alors nous avons $Y_q \in \arg\min \mathbb{E}[\rho_q(Y, Y_q)]$.*

Intuitivement, lorsque $q \leq \frac{1}{2}$, la perte pinball prenant des valeurs plus grandes lorsque l'erreur est négative, nous aurons tendance à sous-estimer y : nous aurons une prévision pour le quantile q, qui inférieur à la médiane.

Nous considérons deux jeux de données, représentés en figure 1.8.

Données régionales de consommation nette en Grande-Bretagne

Nous considérons le jeu de données introduit par BROWELL et FASIOLO, 2021. Il consiste en la consommation nette (demande réduite de la production solaire et éolienne), en Grande-Bretagne. La Grande-Bretagne est décomposée en 14 régions, et nous reprenons les modèles de BROWELL et FASIOLO, 2021, calibrés région par région. En effet, un aspect peu abordé dans cette thèse est qu'il est primordial non seulement d'équilibrer le réseau au niveau national, mais également au niveau local.

Nous évaluons sur une période normale pré-covid, puis nous regardons l'évolution des différents modèles en 2020 et 2021.

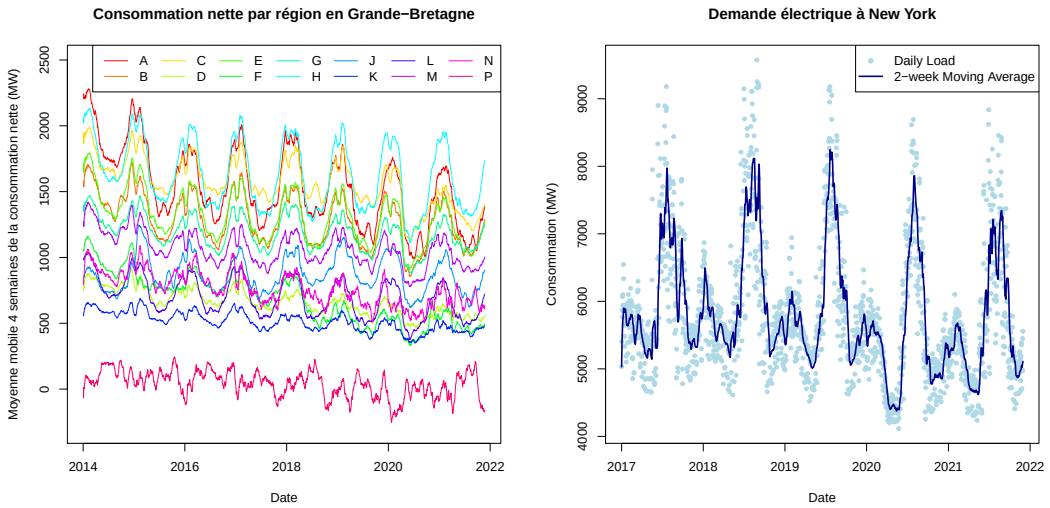


FIGURE 1.8 – Jeux de données utilisés pour la prévision probabiliste : consommation nette en Grande-Bretagne décomposée en 14 régions (à gauche), consommation à New-York (à droite).

Consommation de New York

Le dernier jeu de données utilisé au cours de ce travail porte sur des données américaines, introduites par RUAN et al., 2020. Nous étudions l'évolution de la consommation à New York. Nous nous intéressons à la consommation quotidienne, alors que les jeux de données précédents considéraient une consommation à un pas de temps plus fin (15 min, 30 min ou une heure).

Chapter 2

Introduction (English)

In this chapter, we introduce the thesis. We first present the industrial context motivating this work. Then we detail the theoretical framework that will be followed throughout the manuscript. The rest of the chapter is devoted to the description of our contributions.

Contents

2.1 Industrial Context

Time series forecasting is a fundamental issue. A few non-exhaustive examples are weather and climate forecasting for farmers, sales and inventory forecasting in retail, price and cost forecasting for industrials. Every decision we make crucially depends on forecasts. In this thesis, we study the electricity sector, presenting a specificity compared to the applications mentioned. As electricity cannot be stored on a large scale, it is essential for the electrical network that the supply-demand balance is ensured. At any time, electricity production (supply) must equal the consumption (demand).

To achieve this equilibrium, it is necessary to forecast the demand at various time horizons. In the long or very long term (months, years, or even decades ahead), the forecast is used to determine the adequate production units. The forecast allows to decide whether new units must be built and the maintenance dates of existing units. Medium-term forecasts (weeks, months ahead) are used to update the maintenance calendar and decide whether to operate a controllable production unit according to future needs. In the short term (day-ahead), forecasting consumption allows defining a production schedule for controllable means whose power change requires a few hours (nuclear power plants and, to a lesser extent, fossil fuel power plants). This production schedule is updated frequently to reduce the residual error in the very short term (less than a day ahead). This final error is handled by the transmission system operator with very reactive production means such as hydroelectricity.

This brief overview of power system management assumes that demand is a single quantity and that a central entity makes decisions. This is not the case for several reasons. On the one hand, the opening of the French electricity sector to competition reduces centralization. Nowadays, electricity suppliers have to take care of the supply-demand balance for their customers. They have to produce the electricity their customers consume or buy it from another producer. Each supplier manages this balance as well as possible. Since all the suppliers share the network, the global equilibrium is managed by RTE (Réseau de Transport d'Electricité), the French transmission system operator. To reduce the residual national imbalance, RTE calls on reserves (producers who agree to increase or decrease their production) and imposes penalties on suppliers according to their imbalances. On the other hand, while it is necessary to respect the balance at the national level, it is also preferable to satisfy the supply-demand equilibrium at a more local level. Some countries have a spatial decorrelation between generation and consumption. The transmission of electricity from one region to another is costly because of line losses and because it requires more extensive infrastructure. Indeed, the network is sized in proportion to the transmission.

EDF has significantly improved demand forecasting over the past decades. Classical forecasting methods are based on the observed behavior over typically five years. However, to extrapolate and forecast future consumption, an assumption of stability is necessary.

This assumption has been reassessed in recent years for many reasons. Following the opening to competition, EDF forecasts the consumption of an evolving customer portfolio. Moreover, changes in consumption and in particular the development of electric vehicles could structurally change load patterns. Although the impact of electric cars is still negligible today, it is expected that this will no longer be the case in a few years. Furthermore, while the national load is relatively stable, this is less and less true as the forecasting scale is more local. The extreme case of the forecast of an individual home is interesting: if the dwelling becomes vacant, the forecasting model needs to be changed. The recent development of renewable energies that are not controllable (solar and wind energy) fundamentally changes the management of the supply-demand equilibrium. Controllable means are no longer used to achieve the consumption but the *net* consumption (reduced of solar and wind production). The variable of interest thus becomes net demand, which is more unstable and has more volatility. Finally, the coronavirus crisis has been a significant source of instability for the power system. The lockdowns in many countries caused a sudden drop in demand. Frequent evolutions on restrictions have implied more time-variability than before. Load forecasting difficulty was increased.

EDF has designed new strategies to be more efficient in an unstable context. Expert aggregation is a model-agnostic approach that combines several forecasting models and leverages their diversity. Indeed, models do not perform well in the same contexts, and the aggregation aims to be more robust since it takes advantage of each model's specificities. This method has been widely studied by the time series community and has been significantly improved during a thesis

at EDF (Gaillard, 2015).

This thesis aims to pursue the research of more adaptive models. We focus on state-space representations, a framework that has already been studied during a thesis at EDF (Dordonnat, 2009). Continuing the investigation is motivated by the data itself as the load becomes more unstable. A notable difference with the work of Dordonnat, 2009 is that we take advantage of statistical or Machine Learning models such as the *multi-layer perceptron*, and we combine it with space-state models. In particular, the generalized additive model (GAM) is widely used today at EDF. A few methods have already been developed to adapt it, either by directly changing the coefficients of the model (Ba et al., 2012) or by using a corrective model on the residuals (auto-regressive, for instance). Our approach extends the first option, we do not correct the model's residuals, but we directly adapt some model coefficients.

2.2 Theoretical Framework

In this section, we introduce the theoretical setting of the thesis. We denote by $y_t \in \mathbb{R}$ the variable of interest at time t (for instance, the electricity consumption). By convention, the time t starts at 1, and the frequency may be daily, hourly, half-hourly ... We build a forecast $\hat{y}_t$, and we evaluate it through the error $y_t - \hat{y}_t$ that we try to minimize. More formally, we minimize a loss $\ell(y_t, \hat{y}_t)$, for instance the quadratic loss $\ell(y_t, \hat{y}_t) = \frac{1}{2}(y_t - \hat{y}_t)^2$.

To forecast y_t, we have access to some explanatory variables represented by a vector $x_t \in \mathbb{R}^d$. Each coordinate of x_t is an explanatory variable such as the temperature, the day of the week ...

2.2.1 Adaptive Forecasting

It is usual to parametrize the forecasting model by a vector: we look for the best possible θ in order to forecast $\hat{y}_t = f_\theta(x_t)$, for instance $f_\theta(x_t) = \theta^\top x_t$. In the nonadaptive setting, often denoted by the term *offline* in this manuscript, we find θ performing best during a training set. That is the definition of the empirical risk minimizer (ERM):

$$\hat{\theta}_N^{(ERM)} \in \arg\min_\theta \frac{1}{N} \sum_{t=1}^{N} \ell(y_t, f_\theta(x_t)) \,.$$

We can add penalty terms to the previous optimization problem to make it more robust.

However, the objective of the thesis is to design adaptive models, in which we wish to predict $f_{\theta_t}(x_t)$ instead of $f_\theta(x_t)$. From the *offline* method it is natural to define the *incremental offline* method, where instead of $\hat{\theta}_N^{(ERM)}$ we estimate $\hat{\theta}_t^{(ERM)}$ at each time step t. However, this strategy has several drawbacks. On the one hand, the complexity of such a procedure can be prohibitive (in general, it is expensive to estimate the ERM at each step). On the other hand, this method yields a "weakly adaptive" method because it evolves very slowly, whereas one might wish for a more reactive algorithm.

The goal of online optimization is thus to define an efficient transformation Φ such that the model is updated recursively by $\hat{\theta}_{t+1} = \Phi(\hat{\theta}_t, x_t, y_t)$. The quality of the forecast $\hat{y}_t = f_{\hat{\theta}_t}(x_t)$ is evaluated by the loss $\ell(y_t, f_{\hat{\theta}_t}(x_t))$. A simple online optimization algorithm is the online gradient descent (OGD) introduced by Zinkevich, 2003, which consists in applying a step in the direction opposite to the gradient of the instantaneous loss: at each step we define

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \gamma_t \frac{\partial \ell(y_t, f_\theta(x_t))}{\partial \theta}\Big|_{\hat{\theta}_t} \,, \tag{2.1}$$

where $\gamma_t > 0$ is the algorithm parameter called gradient step size. The methods applied in the thesis are linked to the OGD, see Section 2.2.3.

2.2.2 State-Space Model and Kalman Filter

We focus on state-space models, an example of which is the following linear Gaussian model:

$$\begin{aligned} &\text{State:} && \theta_t = \theta_{t-1} + \eta_t\,, && \eta_t \sim \mathcal{N}(0, Q_t)\,,\\ &\text{Space:} && y_t = \theta_t^\top x_t + \varepsilon_t\,, && \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)\,. \end{aligned}$$

The state equation governs the dynamics of the model. The state follows a random walk, that is, at each step, we add a centered Gaussian noise of covariance matrix $Q_t \in \mathbb{R}^{d\times d}$. The space equation defines the distribution of the observation given the explanatory variables x_t and the state θ_t. The variances (Q_t and σ_t^2) are the two *hyper-parameters* of the model.

Let us first notice an interesting degenerate setting called static: $Q_t = 0$ and $\sigma_t^2 = \sigma^2$. In this case we find a time-invariant model $\theta_t = \theta_{t-1}$.

In the general framework, under the hypothesis that the space-state model is verified, we try to estimate the state at time t with information from observations up to time m, and we are essentially interested in the conditional expectation and covariance matrix of the state:

$$\begin{aligned} \hat{\theta}_{t|m} &= \mathbb{E}[\theta_t \mid x_1, y_1, \cdots, x_m, y_m]\,,\\ P_{t|m} &= \mathbb{E}[(\theta_t - \hat{\theta}_{t|m})(\theta_t - \hat{\theta}_{t|m})^\top \mid x_1, y_1, \cdots, x_m, y_m]\,. \end{aligned}$$

An attractive property of the linear Gaussian state-space model is that provided that the initial state distribution is Gaussian, the distribution of the state θ_t given observations up to time m is also Gaussian. This motivates the estimation of $\hat{\theta}_{t|m}$ and $P_{t|m}$, as the distribution of a Gaussian distribution is entirely determined by its mean and covariance matrix.

The main objective in the context of forecasting is estimating the state's distribution at time t, knowing the past observations (up to $t-1$). This estimation is done exactly by the *Kalman filter* (Kalman and Bucy, 1961):

Theorem 2.1 (Kalman Filter)**.** *Provided that the data-generating process is the state-space model with known variances $(\sigma_t^2, Q_t)_t$, the following recursions are satisfied:*

$$\begin{aligned} P_{t|t-1} &= P_{t-1|t-1} + Q_t\,, && \hat{\theta}_{t|t-1} = \hat{\theta}_{t-1|t-1}\,,\\ P_{t|t} &= P_{t|t-1} - \frac{P_{t|t-1}x_t x_t^\top P_{t|t-1}}{x_t^\top P_{t|t-1} x_t + \sigma_t^2}\,, && \hat{\theta}_{t|t} = \hat{\theta}_{t|t-1} - \frac{P_{t|t}}{\sigma_t^2}\left(x_t(\hat{\theta}_{t|t-1}^\top x_t - y_t)\right)\,. \end{aligned}$$

Conversely, *Kalman smoothing* yields the distribution of the state given future observations. This is done with a backward recursion:

Theorem 2.2 (Kalman Smoothing)**.** *Provided that the data-generating process is the state-space model, the following recursions are satisfied:*

$$\begin{aligned} \hat{\theta}_{t|n} &= \hat{\theta}_{t|t} + P_{t|t}P_{t+1|t}^{-1}(\hat{\theta}_{t+1|n} - \hat{\theta}_{t+1|t})\,,\\ P_{t|n} &= P_{t|t} + P_{t|t}P_{t+1|t}^{-1}(P_{t+1|n} - P_{t+1|t})P_{t+1|t}^{-1}P_{t|t}\,. \end{aligned}$$

This result yields a smoothing of the state estimation trajectory. Indeed, the forward recursion gives the best estimator knowing the past, but the noise on the observation (ε_t) means the

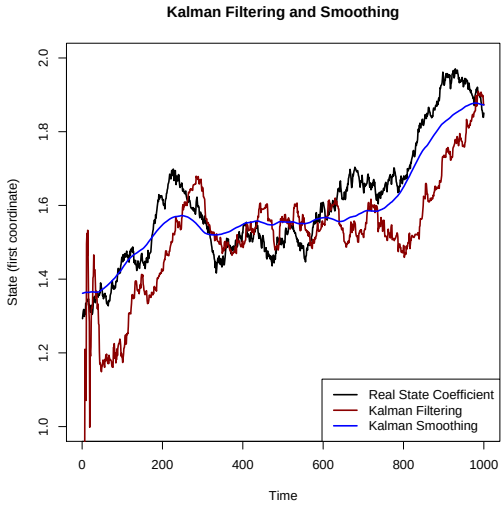


Figure 2.1 – Kalman Filtering and Smoothing. Data is generated under the state-space model with constant variances, and we can therefore compare the estimates with the true state value θ_t.

trajectory of the Kalman filter is a little chaotic. The backward recursion is smoother because the trend is well captured, having access to the future. We illustrate that in Figure 2.1.

The Kalman filter has many advantages. The updates are recursive in the sense the estimators are obtained from their previous values and the new observation x_t, y_t. These updates are efficient (the cost is proportional to d^2, d being the dimension of the state). Moreover, they give the exact expected value and covariance matrix of the state given the past observations.

The primary issue in applying the Kalman filter is that in most applications, one does not know the variances such that the state-space model is satisfied. On the contrary, in many applications, the model is called misspecified. This means the state-space assumption is not satisfied, whatever the variances σ_t^2 and Q_t. A vast literature is devoted to estimating these variances, the true ones in the well-specified case and the "best" ones in the misspecified case, where "best" may be defined, for instance, with respect to the model likelihood or the forecasting error.

A first paradigm consists in assuming that the variances are constant: we assume that $\sigma_t^2 = \sigma^2, Q_t = Q$ and we estimate the best values of σ^2 and Q by maximizing the likelihood (Brockwell and Davis, 2016; Durbin and Koopman, 2012; Fahrmeir and Tutz, 2013). We apply this approach in the course of the thesis, but we observe that it has two drawbacks. First, the resulting model is less extensive since we have restricted the possible values of the variances to the case where they are time-invariant. Second, maximizing the likelihood is a complex problem for which we can only guarantee convergence to a local but not a global optimum.

Other variance choices have led to many algorithms called *adaptive Kalman filters*, which are thus more deeply adaptive. In these approaches, the variances are estimated over time, and the Kalman filter is applied with these variance estimators (Mehra, 1972).

Finally, we are not exclusively interested in the linear Gaussian state-space model, although this is the framework we use for the application to power consumption forecasting. The other

frameworks developed in this manuscript are written as follows:

$$\text{État:} \qquad \theta_t = K\theta_{t-1} + \eta_t \,, \tag{2.2}$$

$$\text{Espace:} \qquad y_t = h(\theta_t^\top x_t) + \varepsilon_t \,, \tag{2.3}$$

where η_t and ε_t are centered noises of respective variances Q_t et s_t^2 and that are not necessarily Gaussian. In the space equation, h is a link function between a linear model and the expected value of y_t. Extensions of the Kalman filter have been developed in this more general setting. We have considered the *extended Kalman filter* (EKF), where the space equation is linearized (Jazwinski, 1970; Durbin and Koopman, 2012). Precisely, denoting by h' the first-order derivative of h, we use the following first-order approximation:

$$y_t \approx h(\hat{\theta}_t^\top x_t) + h'(\hat{\theta}_t^\top x_t) x_t^\top (\theta_t - \hat{\theta}_t) + \varepsilon_t \,.$$

Then we can apply the standard Kalman filter. The difference lies in the fact the noises are not necessarily Gaussian, and therefore the *a posteriori* distribution is not necessarily Gaussian. Moreover, the estimation of the expectation and variance is not exact anymore. Starting from initial estimators of the mean and covariance matrix of θ_0, denoted by $\theta_{0|0}, P_{0|0}$, the EKF is defined by the following recursions at time t:

$$P_{t|t-1} = KP_{t-1|t-1}K^\top + Q_t \,, \qquad \hat{\theta}_{t|t-1} = K\hat{\theta}_{t-1|t-1} \,,$$

$$P_{t|t} = P_{t|t-1} - \frac{h'(\hat{\theta}_t^\top x_t)^2 P_{t|t-1} x_t x_t^\top P_{t|t-1}}{h'(\hat{\theta}_t^\top x_t)^2 x_t^\top P_{t|t-1} x_t + s_t^2} \,, \quad \hat{\theta}_{t|t} = \hat{\theta}_{t|t-1} - \frac{P_{t|t}}{s_t^2}\Big(h'(\hat{\theta}_t^\top x_t) x_t (h(\hat{\theta}_{t|t-1}^\top x_t) - y_t)\Big) \,.$$

Finally, let us mention the *Unscented Kalman Filter* (Julier and Uhlmann, 1997) as an alternative to the EKF for state estimation in the nonlinear framework.

2.2.3 Kalman Filter as a Gradient Descent Algorithm

Let us keep in mind that our objective is to predict a variable y_t by minimizing a loss $\ell(y_t, \hat{y}_t)$. When the loss is the classical quadratic loss $\ell(y_t, \hat{y}_t) = \frac{1}{2}(y_t - \hat{y}_t)^2$, we remark that the Kalman filter (Theorem 2.1) is very similar to the OGD (Equation 2.1). Indeed, if we write explicitly the gradient we have:

$$\text{Online Gradient Descent:} \qquad \hat{\theta}_{t+1} = \hat{\theta}_t - \gamma_t \Big(x_t(\hat{\theta}_t^\top x_t - y_t)\Big) \,,$$

$$\text{Kalman Filter:} \qquad \hat{\theta}_{t+1|t} = \hat{\theta}_{t|t-1} - \frac{P_{t|t}}{\sigma_t^2}\Big(x_t(\hat{\theta}_{t|t-1}^\top x_t - y_t)\Big) \,.$$

Therefore, the difference between both methods lies in the gradient step. It is a scalar in the OGD and a matrix in the Kalman filter. Thus the OGD performs a step in the opposite direction of the gradient, while the Kalman filter uses a *preconditioning matrix* to transform the direction of the gradient. Thus we can interpret the Kalman filter as a second-order gradient descent algorithm, which can be compared to Newton's methods (the P_t matrix is close in some sense to the Hessian of the loss). While the OGD depends on the choice of the gradient step size γ_t, the Kalman filter estimates the matrix $P_{t|t}$ from the hyper-parameters σ_t^2 et Q_t. These variances may thus be interpreted as parameters of a gradient descent. Q_t is the state noise covariance matrix and may be interpreted as the speed of system evolution. The larger Q_t is, the more the system is disturbed over time. We remark in Theorem 2.1 that the larger Q_t is, the larger $P_{t|t}$

is, and the larger the gradient step on $\hat{\theta}_{t|t-1}$. That is what we expect.

This dual interpretation of the Kalman filter as a Bayesian method (estimating the *a posteriori* distribution of the state) and as an online gradient descent algorithm can be generalized to the EKF, as described by Ollivier, 2018.

We develop this interpretation throughout the thesis. The first part, on the analysis of the EKF in the static case, links the EKF to gradient descent algorithms whose step decreases with time ($\gamma_t \to 0$). In the second part, we consider the case where $Q_t \succcurlyeq 0$, and we present the choice of variances as a problem of estimating the optimal gradient step of a gradient descent.

2.3 Stochastic Optimization as a Static State-Space Model

The first contribution of this thesis is to enrich the link between Bayesian statistics and stochastic optimization, building on the parallel presented in section 2.2.3. We study the EKF in the static setting; this is the goal of Part I.

2.3.1 Generalized Linear Models

In our analysis presented in Chapter 4, we focus on loss functions that may be written as the negative log-likelihood of a generalized linear model (McCullagh and Nelder, 1989). Precisely, we assume that the loss function is of the form $\ell(y, \theta^\top x) = -\log p_\theta(y \mid x)$, where p_θ belongs to a subclass of the exponential family parametrized as follows:

$$p_\theta(y \mid x) = c(y) \exp\left(\frac{y\theta^\top x - b(\theta^\top x)}{a}\right),$$

where a is a constant and b and c are univariate functions. This includes Gaussian linear regression, logistic regression (see the example below), and Poisson regression. Our analysis requires further assumptions about the loss ℓ such as convexity.

Example (Logistic Regression). We consider binary classification. We predict $y \in \{-1, 1\}$ and we model $\mathcal{L}(y \mid x)$ by the following distribution parametrized by θ:

$$p_\theta(y \mid x) = \frac{1}{1 + e^{-y\theta^\top x}} = \exp\left(\frac{y\theta^\top x - (2\log(1 + e^{\theta^\top x}) - \theta^\top x)}{2}\right).$$

The loss function becomes $\ell(y, \theta^\top x) = \log(1 + e^{-y\theta^\top x})$.

A remarkable property of distributions in the exponential family is the explicit form of their expectation and variance. In our notations we have $\mathbb{E}[y \mid \theta^\top x] = b'(\theta^\top x)$ and $Var[y \mid \theta^\top x] = ab''(\theta^\top x)$, where b' and b'' are the first two derivatives of the function b. We consider the following static state-space model:

$$\begin{aligned} &\text{State:} && \theta_t = \theta_{t-1}\,, \\ &\text{Space:} && y_t = b'(\theta_t^\top x_t) + \varepsilon_t\,. \end{aligned}$$

The recursive updates become:

$$P_{t|t-1} = P_{t-1|t-1}\,, \qquad \hat{\theta}_{t|t-1} = \hat{\theta}_{t-1|t-1}\,,$$
$$P_{t|t} = P_{t|t-1} - \frac{b''(\hat{\theta}_{t|t-1}^\top x_t) P_{t|t-1} x_t x_t^\top P_{t|t-1}}{b''(\hat{\theta}_{t|t-1}^\top x_t) x_t^\top P_{t|t-1} x_t + a}\,, \quad \hat{\theta}_{t|t} = \hat{\theta}_{t|t-1} - \frac{P_{t|t}}{a}\Big(x_t(b'(\hat{\theta}_{t|t-1}^\top x_t) - y_t)\Big)\,.$$

The Sherman-Morrison formula yields

$$P_{t|t}^{-1} = P_{t-1|t-1}^{-1} + \frac{b''(\hat{\theta}_{t|t-1}^\top x_t) x_t x_t^\top}{a}\,.$$

We denote by ℓ' and ℓ'' the first two derivatives of ℓ with respect to the second variable. We define $P_t = P_{t|t-1}, \hat{\theta}_t = \hat{\theta}_{t|t-1}$ and we obtain:

$$P_{t+1}^{-1} = P_t^{-1} + \ell''(y_t, \hat{\theta}_t^\top x_t) x_t x_t^\top\,, \qquad \hat{\theta}_{t+1} = \hat{\theta}_t - P_{t+1}\Big(x_t \ell'(y_t, \hat{\theta}_t^\top x_t)\Big)\,. \tag{2.4}$$

This update rule yields the following expression:

$$P_{t+1} = \Big(P_1^{-1} + \sum_{s=1}^{t} \ell''(y_s, \hat{\theta}_s^\top x_s) x_s x_s^\top\Big)^{-1}\,.$$

Intuitively, P_{t+1} decays at the rate $1/t$. Therefore, the EKF should be close to a gradient descent where the gradient step size is proportional to $1/t$. However, instead of a scalar gradient step size we have a *preconditioning matrix*. As $\ell''(y_s, \hat{\theta}_s^\top x_s) x_s x_s^\top$ is the Hessian of the instantaneous loss at time step s, provided that $\hat{\theta}_t$ converges, P_{t+1} should be similar to H^{-1}/t where H is the Hessian of the expected loss taken at the limit of $\hat{\theta}_t$.

2.3.2 Results

We categorize guarantees on optimization algorithms into two types. In the *adversarial* framework, no assumptions are made on the generation of the data and (x_t, y_t) can be defined by an adversary, *id est*, the objective is a worst-case analysis. The only assumption is that the data are bounded, and the objective is to bound the regret $\sum_{t=1}^n \ell(y, \hat{\theta}_t^\top x_t) - \ell(y, \theta^{\star\top} x_t)$, the difference between the loss incurred and the loss of a constant oracle.

Conversely, in the *stochastic* framework, (x_t, y_t) are assumed to be independent and identically distributed. It allows the definition of the risk $L(\theta) = \mathbb{E}[\ell(y, \theta^\top x)]$. The objective is to minimize this risk.

We consider an intermediate setting and we obtain bounds on the cumulative risk, defined as $\sum_{t=1}^n L(\hat{\theta}_t) - L(\theta^\star)$ where $\theta^\star$ minimizes the risk.

First, we obtain an upper bound under a strong assumption of convergence of the EKF, defined as follows.

Assumption (Localized)**.** *We define* $\tau(\zeta) = \min\{k \in \mathbb{N} \mid \forall t > k, \|\hat{\theta}_t - \theta^\star\| \leq \zeta\}$ *for any* $\zeta > 0$. *For any* $\delta, \zeta > 0$, *we have* $T(\zeta, \delta) \in \mathbb{N}$ *such that* $\mathbb{P}\big(\tau(\zeta) \leq T(\zeta, \delta)\big) \geq 1 - \delta$.

The assumption states that, from a certain time, with high probability, the EKF estimator is trapped in a ball of radius ζ arbitrarily small around $\theta^\star$. We prove this property next in the quadratic and logistic settings.

Theorem 2.3. *Starting from $\hat{\theta}_1 \in \mathbb{R}^d, P_1 \succ 0$, under some assumptions including the localized assumption, for any $\delta > 0$, it holds simultaneously for $n \geq 1$:*

$$\sum_{t=T(\zeta,\delta)+1}^{T(\zeta,\delta)+n} L(\hat{\theta}_t) - L(\theta^\star) \leq C(\log n + \log \delta^{-1}) \,,$$

with probability at least $1 - 3\delta$.

Sketch of Proof. We decompose the proof into three steps.

1. We start from an *adversarial* bound on the second-order Taylor expansion of the regret:

$$\sum_{t=1}^{n} \left(\left(\ell'(y_t, \hat{\theta}_t^\top x_t) x_t \right)^\top (\hat{\theta}_t - \theta^\star) - \frac{1}{2} (\hat{\theta}_t - \theta^\star)^\top \left(\ell''(y_t, \hat{\theta}_t^\top x_t) x_t x_t^\top \right) (\hat{\theta}_t - \theta^\star) \right) = O(\log n) \,. \tag{2.5}$$

   This bound is presented in Lemma 4.2. We obtain it directly from the recursive updates (2.4). It holds without any assumption on (x_t, y_t).
   Then we transform this bound into a guarantee for the cumulative risk; that is what we do in Steps 2 and 3.

2. The issue with the previous logarithmic bound is that we cannot control the loss with its second-order expansion. We consider the risk L (expected value of the loss ℓ), and we prove a control of the risk by a second-order expansion (proposition 4.1). For any $\rho < 1$, there exists a neighbourhood of $\theta^\star$ denoted by V_ρ such that for any $\theta \in V_\rho$,

$$\frac{\partial L}{\partial \theta}\Big|_\theta^\top (\theta - \theta^\star) \geq \rho (\theta - \theta^\star)^\top \frac{\partial^2 L}{\partial \theta^2}\Big|_\theta^\top (\theta - \theta^\star) \,.$$

   From this property, and using the convexity of the loss (hence of the risk) we obtain the following second-order bound on the risk (proposition 4.2): for any $\theta \in V_\rho$ and $0 < c < \rho$,

$$L(\theta) - L(\theta^\star) \leq \frac{\rho}{\rho - c} \left(\frac{\partial L}{\partial \theta}\Big|_\theta^\top (\theta - \theta^\star) - c (\theta - \theta^\star)^\top \frac{\partial^2 L}{\partial \theta^2}\Big|_\theta (\theta - \theta^\star) \right) . \tag{2.6}$$

3. Finally, we combine Equations (2.5) and (2.6). To that end, we estimate the difference between the first two terms of the Taylor expansion of the loss and those of the second-order expansion of the risk. In particular, we study in Lemma 4.1 the following martingale

$$\Delta M_t = \sum_{t=1}^{n} \left(\frac{\partial L}{\partial \theta}\Big|_\theta - \ell'(y_t, \hat{\theta}_t^\top x_t) x_t \right)^\top (\hat{\theta}_t - \theta^\star) \,.$$

□

This result yields an optimal bound on the cumulative risk; however, we need the strong localized assumption. We prove this hypothesis in two settings. For the quadratic loss, we apply the results of Hsu, Kakade, and Zhang, 2012. In the logistic setting, we prove the localized assumption for a slightly modified variant of the static EKF, in the manner of Bercu, Godichon, and Portier, 2020:

Proposition 2.1. *We recall the logistic loss $\ell(y, \theta^\top x) = \log(1 + e^{-y\theta^\top x})$. Let $0 < \beta < \frac{1}{2}$. We*

define the following algorithm:

$$P_{t+1}^{-1} = P_t^{-1} + \max\left(\ell''(y_t, \hat{\theta}_t^\top x_t), \frac{1}{t^\beta}\right) x_t x_t^\top \,, \qquad \hat{\theta}_{t+1} = \hat{\theta}_t - P_{t+1}\left(x_t \ell'(y_t, \hat{\theta}_t^\top x_t)\right) \,,$$

where we keep the notations of the static EKF with some abuse. This so-called truncated algorithm satisfies the localized assumption, and its recursion coincides with the one of the static EKF after some time.

Formally, keeping the notation $\tau(\zeta)$ for the truncated algorithm, for any $\delta, \zeta > 0$, we can define explicitly $T(\zeta, \delta)$ such that with probability at least $1 - \delta$, it holds:

$$\tau(\zeta) \leq T(\zeta, \delta) \,,$$
$$\forall t \geq T(\zeta, \delta), \ \ell''(y_t, \hat{\theta}_t^\top x_t) \geq \frac{1}{t^\beta} \,.$$

It is crucial that the recursions coincide with these of the static EKF because it allows us to apply the local analysis from $T(\zeta, \delta)$. We treat the first $T(\zeta, \delta)$ terms independently, and the cumulative risk is bounded based on Theorem 2.3.

Sketch of Proof. We decompose the proof into three steps.

1. The objective of the threshold $\frac{1}{t^\beta}$ is to control P_t. Indeed, it is easy to lower bound P_t by $P_t \succcurlyeq cI/t$. However, we don't have upper bounds of the form $P_t \preccurlyeq cI/t$ when $\ell''(y_t, \hat{\theta}_t^\top x_t)$ may be arbitrarily small, which is true for logistic regression. The threshold yields the following control in Proposition 4.4: for any $\delta > 0$, we have $T_1(\delta)$ such that with probability at least $1 - \delta$, it holds:

$$\forall t > T_1(\delta), \qquad P_t \preccurlyeq \frac{4}{\Lambda_{\min} t^{1-\beta}} I \,,$$

   where $\Lambda_{\min}$ is the smallest eigenvalue of $\mathbb{E}[xx^\top]$. A necessary assumption is thus that the latter matrix is invertible.

2. From the recursive updates (2.4), and using $\ell'' \leq \frac{1}{4}$, we obtain the following recursion on the risk:

$$L(\hat{\theta}_{t+1}) \leq L(\hat{\theta}_t) - \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^\top P_t\Big(\ell'(y_t, \hat{\theta}_t^\top x_t) x_t\Big) + 2D_X^4 \lambda_{\max}(P_t)^2 \,. \tag{2.7}$$

   From this recursion Bercu, Godichon, and Portier, 2020 obtain the almost sure convergence of $\hat{\theta}_t$ to $\theta^\star$ applying the Robbins-Siegmund theorem. This result is intuitive because the control on P_t yields

$$\sum_t (2D_X^4 \lambda_{\max}(P_t)^2) < \infty \,,$$

   and the middle term of Equation (2.7) yields a decrease of the risk in expectation. The expected value of the middle term is lower bounded by $\frac{1}{t}\|\frac{\partial L}{\partial \theta}|_{\hat{\theta}_t}\|^2$.

3. We obtain non-asymptotic convergence by estimating the probability of the risk being far from the optimal risk (at a distance greater than $\eta > 0$). To that end, we use the fact that the variations of the algorithm are slow, and we look at the last iteration (if it exists) such that the risk is close to the optimum risk (at a distance at most $\eta/2$).

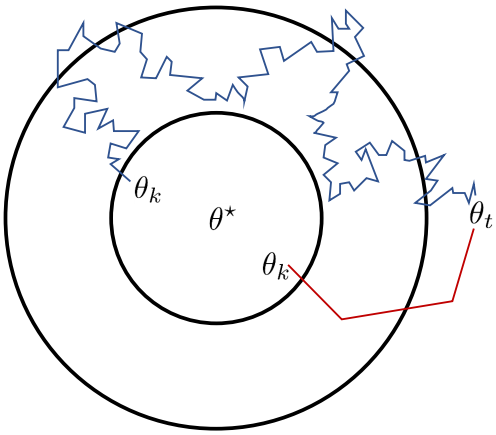


Figure 2.2 – Illustration of the proof of convergence of EKF or SGD. The blue trajectory is unlikely because the iterates are far from the optimum during a long period. The red trajectory is unlikely because the algorithm jumps from the inner ball to the outer ball in a few steps.

Formally we denote by $B_{k,t}$ the event $(\forall k < s < t, L(\hat{\theta}_s) - L(\theta^\star) > \eta/2)$. We use the law of total probability:

$$\mathbb{P}(L(\hat{\theta}_t) - L(\theta^\star) > \eta) = \mathbb{P}\left((L(\hat{\theta}_t) - L(\theta^\star) > \eta) \cap B_{0,t}\right) + \sum_{k=1}^{t-1} \mathbb{P}\left((L(\hat{\theta}_t) - L(\theta^\star) > \eta) \cap \left(L(\hat{\theta}_k) - L(\theta^\star) \leq \frac{\eta}{2}\right) \cap B_{k,t}\right) .$$

To estimate each probability we iterate Equation (2.7):

$$L(\hat{\theta}_t) - L(\hat{\theta}_k) \leq \sum_{s=k}^{t-1} \left(\Delta M_s - \lambda_{\min}(P_s) \left\| \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_s} \right\|^2 + 2D_X^4 \lambda_{\max}(P_s)^2 \right) ,$$

where (ΔM_t) is a martingale difference. We divide the various k into two groups, and we illustrate in Figure 2.2. For k small enough compared to t, the decrease of the risk in expectation makes it unlikely that the estimate stays far from the optimum for a long period. For k closer to t, the control of P_t yields a control on the probability that the algorithm moves fast, and it makes it unlikely that the estimate moves away from the optimum in $t - k$ steps.

□

In Chapter 3 we present a more straightforward convergence proof for the *stochastic gradient descent* algorithm with annealing step size. This algorithm allows us to consider a broader class of functions L. As the gradient step is a scalar instead of a preconditioning matrix, we don't have to apply Step 1. Similarly as for the static EKF, we establish a non-asymptotic version of the almost sure convergence proof Robbins and Monro, 1951.

2.4 The Choice of the Variances in a State-Space Model

Obviously, we have not introduced the state-space model to consider the degenerate static setting only. Our objective is to study the dynamic setting. We recall the linear Gaussian state-space model, in which we are especially interested in this thesis:

$$\text{State:} \qquad \theta_t = \theta_{t-1} + \eta_t\,, \qquad \eta_t \sim \mathcal{N}(0, Q_t)\,, \tag{2.8}$$

$$\text{Space:} \qquad y_t = \theta_t^\top x_t + \varepsilon_t\,, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)\,. \tag{2.9}$$

As presented in Section 2.2.2, the Kalman filter provides the exact estimation of the state recursively for known variances Q_t and σ_t^2. In most applications, these variances are not known. There is no consensus in the literature regarding the choice of these variances in a state-space model.

In Part II we propose several approaches that we segment into two different paradigms. Either we assume the variances are time-invariant and we estimate them on a training period, or we don't assume time-invariance and we estimate the variances adaptively. Let us prolong the parallel between the Kalman filter and a gradient algorithm (2.2.3). The static setting is similar to a gradient step converging to 0. The dynamic setting with constant variances is close to an adaptive step size such as Adam with constant step sizes (Kingma and Ba, 2014). The dynamic setting where variances are estimated online is a deeper adaptation level.

2.4.1 Constant Variances

The most usual choice in the literature we considered is to assume that the variances are time-invariant (Brockwell and Davis, 2016; Durbin and Koopman, 2012; Fahrmeir and Tutz, 2013). Formally, the variances of Equations (2.8) and (2.9) are $Q_t = Q$ and $\sigma_t^2 = \sigma^2$.

In that paradigm, the consensual objective is maximum likelihood on a training data set $(x_t, y_t)_{1\leq t\leq n}$. The most widely used method is the *expectation-maximization* (EM), consisting of two alternating steps:

1. **Expectation**: for fixed variances, we estimate $(\hat{\theta}_{t|n}, P_{t|n})_t$ using Kalman filtering (Theorem 2.1) and Smoothing (Theorem 2.2). Then we deduce the expectation of the complete log-likelihood as a function of Q, σ^2.
2. **Maximization**: for fixed $(\hat{\theta}_{t|n}, P_{t|n})_t$ we estimate the *hyper-parameters* Q, σ^2 maximizing the expectation of the complete log-likelihood.

In the linear Gaussian state-space model, these two steps admit closed-form solutions. Fixing the variances, Kalman filtering and smoothing is exact. Conversely, fixing the Kalman estimates, the maximum of the expectation of the complete log-likelihood has a closed-form expression. Furthermore, this iterative procedure yields an appealing guarantee: at each step, the likelihood increases.

However, the EM algorithm has two significant drawbacks. On the one hand, it is a costly algorithm that converges slowly. On the other hand, while it guarantees the convergence towards a local maximum of the likelihood, it does not converge towards a global maximum. Indeed, the log-likelihood is not necessarily a concave function; see Figure 2.3.

We propose an alternative procedure. We estimate maximum-likelihood by an *iterative grid search* on Q, and we restrict ourselves to diagonal covariance matrices. This restriction means we assume that the coefficients of θ_t evolve independently from each other, and we believe it is a reasonable restriction on the model. Remark that this assumption of independent evolution of the coefficients does not yield independent evolution of the coefficients of the estimator $\hat{\theta}_t$.

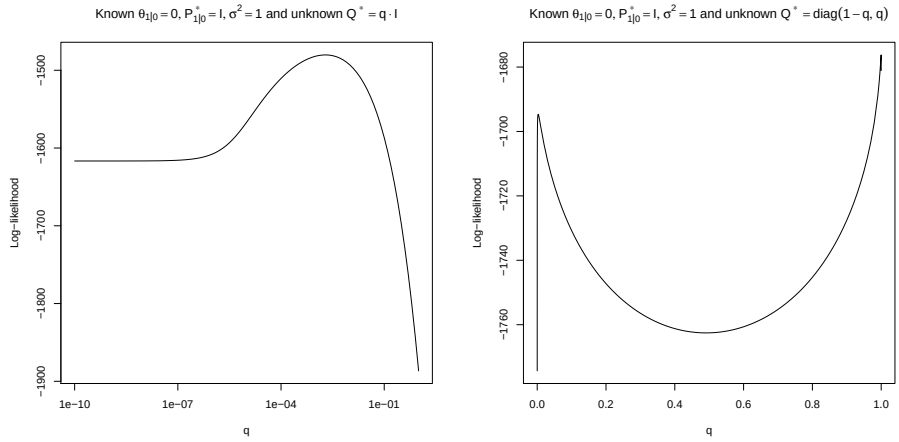


Figure 2.3 – Log-likelihood for varying Q in a well-specified setting in dimension $d = 2$. We fix $\sigma^2, \hat{\theta}_{1|0}, P_{1|0}$. On the left: $Q = qI_2$. On the right: $Q = diag(1 - q, q)$.

As the name *iterative grid search* suggests, we search in a grid for the diagonal coefficients of Q; however, we do it iteratively and not exhaustively. At each step, we compute the likelihood of the matrices Q having only one coefficient different from the previous iteration, and we keep the one achieving the highest likelihood. Similarly as the EM algorithm, *iterative grid search* does not guarantee the convergence to the global maximum, but we have seen better results in practice.

We believe these better results come from greater robustness to two real-world phenomena. First, the linear Gaussian state-space model with time-invariant variances is usually misspecified. This means that the data are not generated by Equations (2.8) and (2.9) whatever the constant variances $Q_t = Q$ and $\sigma_t^2 = \sigma^2$. This does not undermine the model in practice. The interpretation as a gradient algorithm in Section 2.2.3 justifies that the algorithm is robust. However, the misspecification may explain that a more empirical method may work better.

The second phenomenon exists in a lot of applications. It is the data availability delay. For instance, in the case of the electric network, the consumption is not perfectly known in real-time; instead, it is incrementally estimated. We have reliable estimation after a few days, and the final consolidated load is published by RTE only months afterward. Formally, the delay means that in order to forecast y_t, we have at our disposal x_t and observations $x_1, y_1, \ldots, x_{t-k}, y_{t-k}$ where k is the delay. In the *iterative grid search*, we can optimize a variant of the likelihood taking the delay into account. In a sense, we avoid overfitting. That is not the case of the EM algorithm, as we show in Section 5.4. We illustrate that phenomenon in Figure 2.4 with a toy example.

Finally, an advantage of *iterative grid search* is its simplicity. It may be applied in any state-space model with any variant of the Kalman filter.

2.4.2 Dynamical Variances

A second paradigm consists in estimating the variances of a state-space model over time. This has frequently been called *adaptive Kalman filtering* (Mehra, 1972). In Chapter 6, we develop a new method called Viking, estimating the state and the variances jointly.

Formally, we denote by $\mathcal{F}_t = \sigma(x_1, y_1, \ldots, x_t, y_t)$ the natural filtration. We seek to apply a Bayesian method. We start with a prior $p(\theta_0, \sigma_0^2, Q_0 \mid \mathcal{F}_0)$ and we assume a model on the

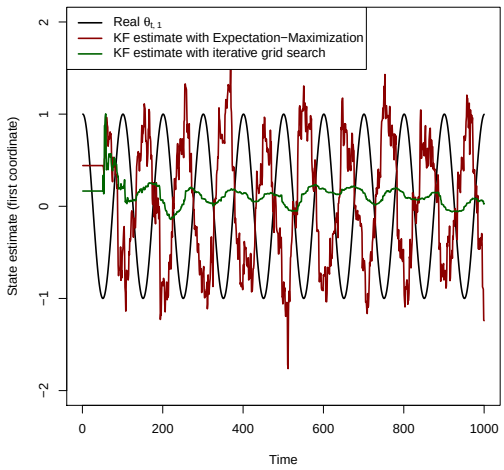


Figure 2.4 – Synthetic data in a misspecified setting. We use the dimension $d = 2$. The state is defined as $\theta_t = \cos(\frac{2\pi t}{100}) \cdot (1,1)^\top$, then $x_t \sim \mathcal{N}(0, I)$ and $y_t - \theta_t^\top x_t \sim \mathcal{N}(0,1)$. The delay is $k = 50$, which is the worst-case scenario in terms of phase offset. The best is to have a small matrix Q and not move too fast.

dynamics of these three variables $p(\theta_t, \sigma_t^2, Q_t \mid \theta_{t-1}, \sigma_{t-1}^2, Q_{t-1})$. At each iteration t, we apply a prediction step (following the dynamics of the model), and a filtering step (Bayes' rule):

$$\begin{aligned} &\text{Prediction:} && p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_{t-1})\,, \\ &\text{Filtering:} && p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_t)\,. \end{aligned}$$

However, there is no natural parametric class of distributions on the joint distribution such that the posterior remains in the class considered. We then apply the *Variational Bayes* (VB) approach (Šmídl and Quinn, 2006). This consists in approximating the joint distribution with a product distribution of which each marginal has a simple form.

We keep a Gaussian marginal for θ_t in order to coincide with the exact posterior in the degenerate setting where the variances are known. We introduce parametric distributions on σ_t^2 and Q_t of the form $P_{\Phi_{t|t}}$ and $P_{\Psi_{t|t}}$ and of densities $p_{\Phi_{t|t}}$ and $p_{\Psi_{t|t}}$, where $\Phi_{t|t}$ and $\Psi_{t|t}$ are the parameters. We then estimate $\hat{\theta}_{t|t}, P_{t|t}, \Phi_{t|t}, \Psi_{t|t}$ such that the product $\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times P_{\Phi_{t|t}} \times P_{\Psi_{t|t}}$ is the "best" approximation of the posterior distribution denoted by $P_{\mathcal{F}_t}$. Formally, we minimize the Kullback-Leibler (KL) divergence:

$$KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times P_{\Phi_{t|t}} \times P_{\Psi_{t|t}} \mid\mid P_{\mathcal{F}_t}\Big),$$

where $KL(P \mid\mid Q) = \int_x \log(p(x)/q(x))p(x)dx$ for any distributions P and Q of densities p and q. At each step, the VB approach yields a coupled optimization problem in three distributions.

The prediction step is determined by the dynamics we propose in the model. We denote by $\mathcal{N}(x \mid \mu, \Sigma)$ the density of the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ at point x. Starting from the prior

$$p(\theta_{t-1}, \sigma_{t-1}^2, Q_{t-1} \mid \mathcal{F}_{t-1}) = \mathcal{N}(\theta_{t-1} \mid \hat{\theta}_{t-1|t-1}, P_{t-1|t-1}) p_{\Phi_{t-1|t-1}}(\sigma_{t-1}^2) p_{\Psi_{t-1|t-1}}(Q_{t-1})\,,$$

we naturally obtain the following density with suitable assumptions:

$$p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_{t-1}) = \mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^\top + Q_t) p_{\Phi_{t|t-1}}(\sigma_t^2) p_{\Psi_{t|t-1}}(Q_t)\,,$$

where $\Phi_{t|t-1}$ and $\Psi_{t|t-1}$ have simple forms. This prediction step is introduced as a prior in the filtering step, yielding the following posterior distribution:

$$\begin{aligned} p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_t) = \frac{p(x_t, \mathcal{F}_{t-1})}{p(\mathcal{F}_t)} &\mathcal{N}(y_t \mid \theta_t^\top x_t, \sigma_t^2) \\ &\mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^\top + Q_t) p_{\Phi_{t|t-1}}(\sigma_t^2) p_{\Psi_{t|t-1}}(Q_t)\,. \end{aligned}$$

We remark that whatever the parametric distributions on σ_t^2 and Q_t, the joint posterior distribution of the state and variances cannot be factorized. Therefore, it is natural to apply the VB approach in order to estimate the posterior distribution with a simple factorized one. Furthermore, the crossed factor involving θ_t and Q_t in the preceding density prevents us from applying the method introduced by Tzikas, Likas, and Galatsanos, 2008 to obtain an analytical form for the KL minimum. We obtain upper bounds on the KL divergence, whose minimization yields closed forms. Minimizing these upper bounds does not yield the KL minimum, but we reduce the KL divergence at each step; this is the *evidence-lower bound* (ELBO) rule.

This brief overview of the VB approach was written for general parametric distributions for σ_t^2 and Q_t. Let us detail how we propose to define them in order to derive the algorithm Viking. The recursive estimation of the posterior distribution does not suggest natural distributions simplifying the KL divergence minimization. Therefore, we choose to represent the variances using Gaussian variables. A significant advantage of a Gaussian latent variable is that dynamics is naturally introduced in the form of a random walk. However, as variances must be nonnegative, we transform these Gaussian variables. Specifically, we use $\sigma_t^2 = \exp(a_t)$ (log-normal distribution) and $Q_t = f(b_t)$, where a_t, b_t follow Gaussian distributions. The function f is a parameter of the algorithm, for which we propose different possible choices. Thanks to this full Gaussian representation, the approach may be summarized as follows:

$$\begin{aligned} &\theta_0 \sim \mathcal{N}(\hat{\theta}_0, P_0)\,, \quad a_0 \sim \mathcal{N}(\hat{a}_0, s_0)\,, \quad b_0 \sim \mathcal{N}(\hat{b}_0, \Sigma_0)\,, \\ &a_t - a_{t-1} \sim \mathcal{N}(0, \rho_a)\,, \quad b_t - b_{t-1} \sim \mathcal{N}(0, \rho_b I)\,, \\ &\theta_t - K\theta_{t-1} \sim \mathcal{N}(0, f(b_t))\,, \\ &y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \exp(a_t))\,, \end{aligned}$$

where we introduce the (nonnegative) parameters ρ_a and ρ_b. These parameters govern the dynamics of the latent variables a_t and b_t, representing the variances σ_t^2 and Q_t.

2.5 Application to Electricity Load Forecasting

In Part III, we apply state-space models to electricity load forecasting in various countries and at various scales. One data set considers electricity net-load, defined as the difference between the consumption and the embedded solar and wind productions, that we cannot control. We introduce in Section 2.5.1 the generic framework on which we build our forecasting strategies. In Section 2.5.2 we present *mean* forecasting (or median forecasting): we measure absolute errors, and we attribute similar performances to positive and negative errors. In Chapter 10 introduced in Section 2.5.3 we estimate consumption quantiles, that is *probabilistic* forecasing.

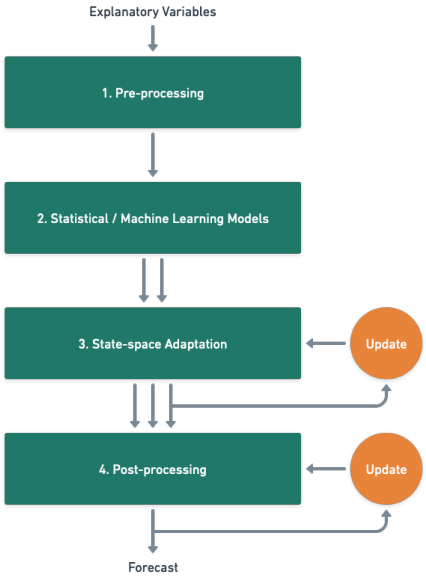


Figure 2.5 – Flowchart of the method applied to forecast the electricity load. The increasing number of arrows means an increase of the model numbers at a step. The different forecasting models are aggregated or merged in the last step to yield a final forecast.

2.5.1 Application of the Linear Gaussian State-Space Model

We sum up the method in the flowchart of Figure 2.5. Specifically, for each time series forecasting problem encountered, we have explanatory variables, and we seek to forecast a variable of interest. Our approach can be summarized in 4 steps, and we highlight the third as it is the core of the thesis:

1. **Pre-processing**. We clean the data, and we select interesting explanatory variables. We compute hand-designed features like exponential smoothing of the temperature. We correct meteorological forecasts ...

2. **Statistical / Machine Learning Models**. We use classical time series forecasting models. We focus on autoregressive, linear regression, *generalized additive model* (GAM) and *multi-layer perceptron* (MLP). In most of our applications, these models are estimated independently for each time of day. For instance, if we forecast the hourly consumption, we have 24 different models. This step yields several possible models.

3. **State-space Adaptation**. We adapt the different models of Step 2 using the linear Gaussian state-space model. We use the different variance estimation introduced in Part II. Before applying the state-space model, we linearize the models that we adapt. For the GAM, we freeze the nonlinear effects, and we adapt a linear combination of the (linear and nonlinear) effects. For the MLP, we freeze the deepest layers, and we adapt the last one only. In other words, we use Step 2 to learn new features, and we use these features as explanatory variables in a linear Gaussian state-space model.

   Formally, we forecast a quantity y_t given explanatory variables x_t. From Step 2 we obtain

a new covariate vector $f(x_t)$, and we consider the following state-space model:

$$\begin{aligned} &\text{State:} && \theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t)\,, \\ &\text{Space:} && y_t - \theta_t^\top f(x_t) \sim \mathcal{N}(0, \sigma_t^2)\,. \end{aligned}$$

We test the different estimation methods for σ_t^2, Q_t. We use the degenerate static setting ($Q_t = 0$), the time-invariant variances paradigm, and the dynamical variances framework. The coronavirus crisis yielded a big break in the data. Therefore, in the data sets including this crisis period, we introduce a break in the model itself at a specified time (mid-March 2020). To take the break into account, we define a big covariance matrix Q_T at the specified break time T, and this matrix is set large compared to the usual values of Q_t.

4. **Post-processing**. Finally, we transform the forecasts of Step 3 to obtain a final forecast. To that end, we often relied on expert aggregation, detailed in the thesis of Gaillard, 2015. Formally, we have different forecasts (experts) $\hat{y}_{t,1}, \dots \hat{y}_{t,K}$ provided by Step 3. Our final prediction is a weighted average of these experts, $\sum_k p_{t,k}\hat{y}_{t,k}$. In the linear combination we use $\sum_k p_{t,k} = 1$. The weights $p_{t,1}, \dots p_{t,K}$ are estimated dynamically, that is, they are time-varying to take into account the evolution of each expert's recent performances.
Before this aggregation step, we use for one of the data sets an intraday correction in order to take into account the correlation between the consumption at the different times of the day.

2.5.2 Mean Forecast

We have applied this method on the data sets detailed below and displayed in Figure 2.6. We evaluate through various evaluation metrics. The most classical are the mean absolute error (MAE) and the root-mean-square error (RMSE), defined below for observations $(y_1, \dots y_n)$ and their associated forecasts $(\hat{y}_1, \dots \hat{y}_n)$:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t|\,, \qquad RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}\,.$$

Chapter 7: Spring 2020 in France

The first application of the thesis is the French electricity consumption. Indeed, forecasting French consumption is an essential issue for EDF. We therefore study in particular the interest of the approach presented in the thesis for the French data set. We evaluate on a "normal" period before March 2020. Then we study the special period of the coronavirus crisis, whose impact was very strong, especially during the first lockdown from March to May 2020.

Compared to a non-adaptive generalized additive model, to which we apply an ARIMA correction, we obtain a 7% reduction in RMSE on a pre-covid period. During the first month of lockdown, the Kalman filter reduces the RMSE by 11%, boosted to 28% when we model the break. During the following two months (a more stabilized but still chaotic period), we reduce the RMSE by 42%. The aggregation of experts increases the gain significantly.

Chapter 8: Post-covid City-wide Forecasting Competition

We participated in the competition *Day-Ahead Electricity Demand Forecasting: Post-COVID Paradigm* (Farrokhabadi, 2020). The organizers' objective was to develop new forecasting strategies robust to disruption, such as the coronavirus. The data set was a city whose location was

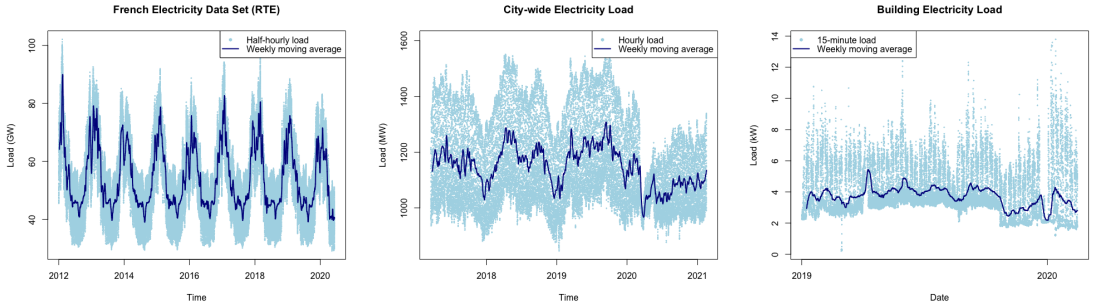


Figure 2.6 – Data sets used for electricity load forecasting at different scales: France (left), an unknown city (middle) or a building of unknown location (right).

undisclosed. While in Chapter 7 we are primarily interested in the lockdown period, in this competition, the evaluation period was January-February 2021, about one year after the break. Thus, while the lockdown is an extreme event, the competition's setting was a less turbulent period and aimed at defining prediction models adapted to the future rather than focused on a past period.

We won the competition, improving the naive method given by the competition (persistence) by over 30% in MAE.

Chapter 9: Building Forecasting Competition

We participated in a second competition, called *Competition on Building Energy Consumption Forecasting*, whose objective was to forecast the electricity consumption of a building. This competition allowed us to apply the method at a much smaller scale. It seems that adaptive methods are necessary in this framework because the behavior change of one person has a non-marginal impact, whereas this is not the case when the data are aggregated at the level of a country or a large city.

We also won this competition, improving a naive method (persistence) by more than 30% in MAE.

2.5.3 Probabilistic Forecast

In Chapter 10, we consider probabilistic forecasting. Instead of trying to minimize the absolute error of our forecast with the variable of interest we forecast its quantiles. Precisely, let us recall that y_t is the variable to forecast, we seek to predict $\hat{y}_{t,q} \in \mathbb{R}$ such that $\mathbb{P}(y_t < \hat{y}_{t,q}) = q$. We test different approaches.

— Let us first notice that the Kalman filter provides a probabilistic prediction by essence. Indeed, we estimate the *a posteriori* distribution of the state θ_t as the Gaussian distribution $\mathcal{N}(\hat{\theta}_{t|t-1}, P_{t|t-1})$. Furthermore, the state is defined such that $y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma_t^2)$. Thus, we can predict y_t as a Gaussian distribution. Its q-quantile is then written as

$$\hat{y}_{t,q} = \hat{\theta}_{t|t-1}^\top x_t + U_q \sqrt{\sigma_t^2 + x_t^\top P_{t|t-1} x_t}\,,$$

where U_q is the q-quantile of the standard normal distribution.

— However, the Gaussian *a posteriori* distribution property of the state relies heavily on the state-space model. Let us keep in mind that the linear state-space model of known

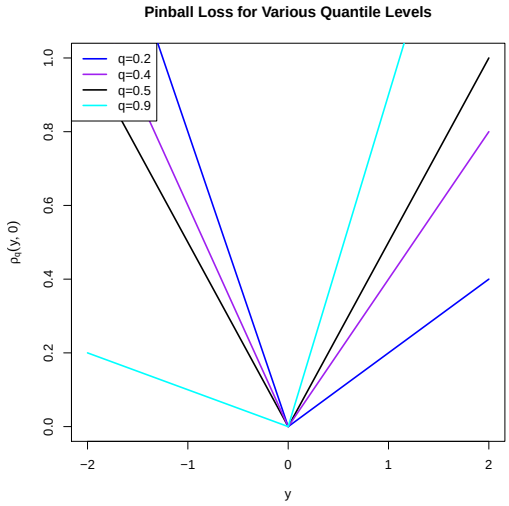


Figure 2.7 – Value of the pinball loss for different quantile levels, for $\hat{y}_q = 0$ and y variable. When $q \leq \frac{1}{2}$, the loss is more significant when y takes negative values and vice versa.

variances is not satisfied in the real world. Thus, in practice, it is best to use the state-space models to forecast average consumption. Then we build a model on the distribution of residuals (difference $y_t - \hat{y}_t$, where $\hat{y}_t$ is the forecast average).
Rather than seeking to minimise the absolute error of our forecast with the variable to be forecast, we use an asymmetric error that assigns a larger loss to a negative (or positive) error. This loss is called the pinball loss, defined as $\rho_q(y, \hat{y}_q) = (\mathbb{1}_{y<\hat{y}_q} - q)\ (\hat{y}_q - y)$ and displayed in Figure 2.7. The following proposition justifies that the optimization of this loss results in probabilistic forecasts:
Proposition 2.2. *Let Y be a real-valued random variable. For any $0 < q < 1$, we denote by Y_q the q-quantile of Y. Then we have $Y_q \in \arg\min \mathbb{E}[\rho_q(Y, Y_q)]$.*
Intuitively, when $q \leq \frac{1}{2}$, the pinball loss takes larger values when the error is negative. We thus tend to underestimate y: we will have a forecast for the quantile q, which is lower than the median.

We consider two data sets displayed in Figure 2.8.

Regional Net-load in Great Britain

We use the data set introduced by Browell and Fasiolo, 2021. It consists of the net-load (demand reduced by solar and wind production) in Great Britain. Great Britain is decomposed into 14 regions, and we use the models of Browell and Fasiolo, 2021, calibrated region by region. Indeed, an aspect not much discussed in this thesis is that it is essential to balance the network at more local levels, besides national equilibrium.

We evaluate on a normal pre-covid period and then look at the evolution of the different models in 2020 and 2021.

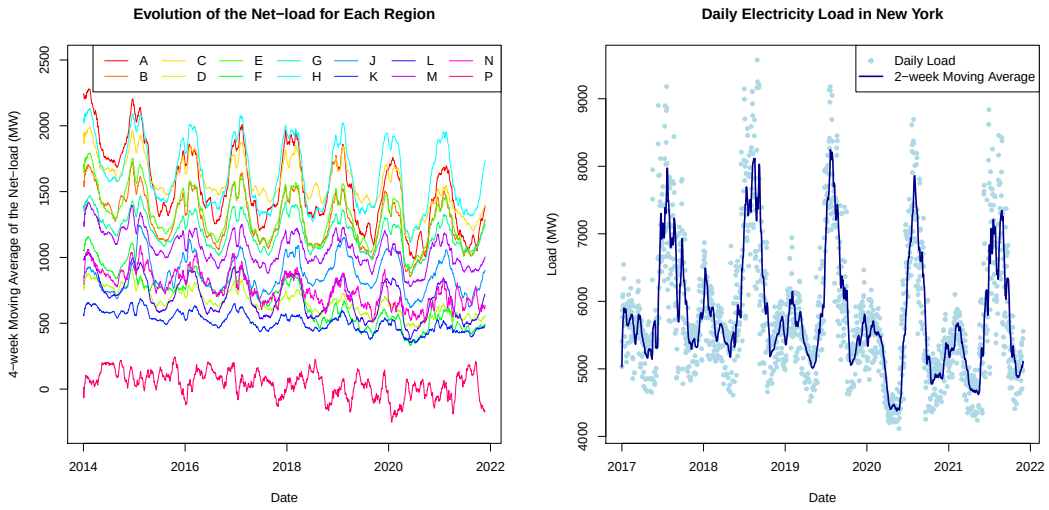


Figure 2.8 – Data sets considered for probabilistic forecasting: net-load in Great Britain decomposed in 14 regions (left), load in New-York (right).

New York Load

The last data set used in this thesis is US data introduced by Ruan et al., 2020. We study the evolution of consumption in New York. We are interested in daily consumption, whereas previous data sets considered consumption at a finer time granularity (15 min, 30 min, or one hour).

Part I

Stochastic Optimization as a Static State-Space Model

Chapter 3

Non-asymptotic Robbins-Monro

In this chapter, we present a convergence analysis for a simple gradient descent algorithm, relying on Azuma-Hoeffding inequality. This yields, with high probability, the convergence of unconstrained stochastic gradient algorithms to a *localized* phase where the iterates are trapped around the optimum. It may be helpful in order to extend guarantees obtained on constrained algorithms to their equivalent on the unconstrained version. This technique is applied in Chapter 4.

Contents

3.1 Introduction

We consider the classical problem of minimizing a convex function, having access to unbiased estimates of its gradient. This problem is well-understood, and numerous guarantees have been obtained on various algorithms.

The most widely used algorithm remains the simple stochastic gradient descent (SGD) of Robbins and Monro, 1951, as well as its averaged variant (Polyak and Juditsky, 1992; Bach, 2014). More complex methods have been designed relying on second-order information. The online Newton step (Hazan, Agarwal, and Kale, 2007) has been designed in the online setting where the objective is to compete with an adversary, and guarantees have been generalized to the stochastic setting where the gradients are independent and identically distributed (Mahdavi, Zhang, and Jin, 2015). More simple methods have been designed to reduce the computational cost of algorithms implying the estimation of a matrix of size the squared dimension of the problem, still leveraging second-order information. That is the goal of AdaGrad (Duchi, Hazan, and Singer, 2011) and Adam (Kingma and Ba, 2014).

The guarantees obtained on all these algorithms are not necessarily comparable because they split into two interesting settings in the optimization community, online and stochastic. In the

online setting, we have access to a series of loss functions $\ell_1, \ell_2 \dots$ or their gradients. The objective is to minimize the regret $\sum_{t=1}^{n}(\ell_t(\theta) - \ell_t(\theta^\star))$, where $\theta^\star$ is the minimum of the cumulative loss, depending on the horizon n and the loss functions $\ell_1, \dots \ell_n$ that may be designed by an adversary. We focus here on the stochastic setting where the goal is to minimize a convex function L that is fixed. We assume that there exists a minimizer for L that we denote by $\theta^\star$, and we have access to $\ell_1, \ell_2 \dots$ such that $L(\theta) = \mathbb{E}[\ell_t(\theta)]$, or to $\nabla\ell_1, \nabla\ell_2 \dots$ such that $\nabla L(\theta) = \mathbb{E}[\nabla\ell_t(\theta)]$, where ∇L is the gradient of L. Then a stochastic optimization algorithm defines an estimate θ_n having observed $\ell_1, \dots \ell_n$ (or their gradient equivalents), and the objective is to upper-bound $L(\theta_n) - L(\theta^\star)$.

We refer to Section 4.1.1 for a detailed presentation of the various guarantees existing in the stochastic optimization literature. We don't claim in this chapter to provide optimal bounds. Instead, our objective is to provide a more elementary proof of convergence for SGD. We interpret our result as a non-asymptotic variant of Robbins-Monro analysis. The latter yields the almost sure convergence of the estimate to the optimum. We prove, with high probability, the convergence of the algorithm to a *localized* phase where the estimates of the algorithm are trapped in a small region around the optimum. This is a general method that can be applied to any stochastic gradient algorithm, as is done in Chapter 4. Applying our approach leads to deriving a two-step analysis for unconstrained algorithms. The first phase is a convergence phase where the algorithm is hard to control. We control the duration of that first phase. In the second (*localized*) phase, we can apply tight analyses that have been designed for constrained algorithms.

In what follows, we consider the following recursion:

$$\begin{aligned}\theta_1 &\in \mathbb{R}^d\,,\\ \theta_{t+1} &= \theta_t - \gamma_t \nabla\ell_t(\theta_t)\,,\end{aligned}$$

where γ_t is the gradient step size and $\nabla\ell_t$ is the unbiased estimate of the gradient of L, which we observe at time t. Our analysis considers $\gamma_t = 1/t^\beta$ for some β.

We focus on SGD as it yields proofs that are more reader-friendly than algorithms relying on a *preconditioning* matrix before the gradient. Instead of applying a scalar step size before the gradient, a matrix is used to transform the gradient direction in these latter algorithms. This matrix is generally designed to take second-order information into account.

3.1.1 Related Work: Application of Robbins-Siegmund Theorem

We present in this section a standard convergence analysis of SGD. We assume the gradients of the losses are bounded, as well as the Hessian of the risk:

Assumption 3.1. *There exists constants $g, h > 0$ such that for any $\theta \in \mathbb{R}^d$, it holds:*
- *$\|\nabla\ell_t(\theta)\| \le g$ for any $t \ge 1$.*
- *The Hessian of L denoted by $\nabla^2 L(\theta)$ satisfies $0 \preccurlyeq \nabla^2 L(\theta) \preccurlyeq hI$.*

Lemma 3.1. *Let $t \ge 1$. Under Assumption 3.1, we have:*

$$L(\theta_{t+1}) \le L(\theta_t) - \gamma_t \|\nabla L(\theta_t)\|^2 + \gamma_t^2 g^2 h + \Delta M_t\,,$$

where $\Delta M_t = -\gamma_t \nabla L(\theta_t)^\top \Big(\nabla\ell_t(\theta_t) - \nabla L(\theta_t)\Big)$.

Proof. We first apply a second-order Taylor expansion on L: there exists $0 \le \alpha_t \le 1$ such that

$$L(\theta_{t+1}) = L(\theta_t) + \nabla L(\theta_t)^\top(\theta_{t+1} - \theta_t) + \frac{1}{2}(\theta_{t+1} - \theta_t)^\top \nabla^2 L(\theta_t + \alpha_t(\theta_{t+1} - \theta_t))(\theta_{t+1} - \theta_t)\,.$$

Using the update formula on θ, as well as the upper bound on the Hessian of L, we obtain

$$L(\theta_{t+1}) \leq L(\theta_t) - \gamma_t \nabla L(\theta_t)^\top \nabla \ell_t(\theta_t) + \gamma_t^2 g^2 h \,,$$

and the result follows. □

Lemma 3.1 allows a recursive control of $L(\theta_t)$. Indeed, we have $\mathbb{E}[\Delta M_t] = 0$. Therefore, the recursive evolution of $L(\theta_t)$ in expectation depends on the sign of $-\gamma_t \|\nabla L(\theta_t)\|^2 + \gamma_t^2 g^2 h$. The classical convergence proof relies on the following theorem (Robbins and Siegmund, 1971):

Theorem 3.1 (Robbins-Siegmund Theorem)**.** *Let $(z_t), (\beta_t), (\xi_t), (\zeta_t)$ be non-negative random variables adapted to a filtration $(\mathcal{F}_t)$, and satisfying*

$$\mathbb{E}[z_{t+1} \mid \mathcal{F}_t] \leq z_t(1 + \beta_t) + \xi_t - \zeta_t, \qquad t \geq 1.$$

Then on $\{\sum_{t=1}^\infty \beta_t < \infty, \sum_{t=1}^\infty \xi_t < \infty\}$, we have almost surely the convergence of (z_t) to a finite limit, and almost surely $\sum_{t=1}^\infty \zeta_t < \infty$.

This theorem, along with Lemma 3.1, yields the convergence of SGD under a few hypotheses. Standard assumptions in the stochastic setting include independence and identical distribution (i.i.d.), as well as the existence of a minimizer:

Assumption 3.2. *$\nabla \ell_1, \nabla \ell_2, \ldots$ are i.i.d. copies and $\nabla L = \mathbb{E}[\nabla \ell_1]$.*

Assumption 3.3. *There exists $\theta^\star \in \mathbb{R}^d$ such that $L(\theta^\star) = \min_{\theta \in \mathbb{R}^d} L(\theta)$.*

To apply the Robbins-Siegmund theorem and deduce the almost sure convergence, we will see that the following assumption is natural.

Assumption 3.4. *For any $\eta > 0$, there exists D_η such that for any $\theta \in \mathbb{R}^d$,*

$$L(\theta) - L(\theta^\star) > \frac{\eta}{2} \implies \|\nabla L(\theta)\| > D_\eta \,.$$

These definitions yield the following convergence result.

Theorem 3.2. *If Assumptions 3.1, 3.2, 3.3 and 3.4 are satisfied, and if the gradient steps satisfy $\sum_{t=1}^\infty \gamma_t = +\infty$ and $\sum_{t=1}^\infty \gamma_t^2 < \infty$, then almost surely $L(\theta_t)$ converges to $L(\theta^\star)$.*

Proof. Let $(\mathcal{F}_t)$ be the natural filtration $(\sigma(\nabla \ell_1, \ldots, \nabla \ell_t))_t$. Then from Lemma 3.1 we obtain

$$\mathbb{E}[L(\theta_{t+1}) \mid \mathcal{F}_t] \leq L(\theta_t) - \gamma_t \|\nabla L(\theta_t)\|^2 + \gamma_t^2 g^2 h \,.$$

We apply Theorem 3.1 with $\sum_{t=1}^\infty \gamma_t^2 g^2 h < \infty$. Almost surely $L(\theta_t)$ converges to a finite limit, and also almost surely $\sum_{t=1}^\infty \gamma_t \|\nabla L(\theta_t)\|^2 < \infty$.

We note that $\sum_{t=1}^\infty \gamma_t \|\nabla L(\theta_t)\|^2 < \infty$ and $\sum_{t=1}^\infty \gamma_t = \infty$ alone do not imply the convergence of $\|\nabla L(\theta_t)\|$ to 0. Furthermore, as we do not assume that $\theta^\star$ is the unique minimum of L, the convergence of $\|\nabla L(\theta_t)\|$ would not in turn imply the convergence of $L(\theta_t)$ to $L(\theta^\star)$. Assumption 3.4 is the natural sufficient condition such that $\sum_{t=1}^\infty \gamma_t \|\nabla L(\theta_t)\|^2 < \infty$ is incompatible with the convergence of $L(\theta_t)$ to a different limit than $L(\theta^\star)$. Indeed, for any $\eta > 0$, we have:

$$\begin{aligned} L(\theta_t) \to L(\theta^\star) + \eta &\implies \max\{t \in \mathbb{N} \mid \|\nabla L(\theta_t)\| \leq D_\eta\} < \infty \\ &\implies \sum_{t=1}^\infty \gamma_t \|\nabla L(\theta_t)\|^2 = \infty \,. \end{aligned}$$

□

Assumption 3.4 is related to the local strong convexity that we define below. We show in Proposition 3.1 that the local strong convexity is stronger than Assumption 3.4.

Assumption 3.5 (Local Strong Convexity)**.** *For any $\theta_1, \theta_2 \in \mathbb{R}^d$ such that $\|\theta_1 - \theta^\star\| \leq \varepsilon$ and $\|\theta_2 - \theta^\star\| \leq \varepsilon$, it holds*

$$L(\theta_2) \geq L(\theta_1) + \nabla L(\theta_1)^\top (\theta_2 - \theta_1) + \frac{\mu_\varepsilon}{2} \|\theta_2 - \theta_1\|^2 .$$

Proposition 3.1. *If Assumption 3.5 is satisfied for given ε and μ_ε, then Assumption 3.4 holds for $D_\eta = \sqrt{\min(\mu_\varepsilon^2 \varepsilon^2, \mu_\varepsilon \eta)}$.*

Proposition 3.1 is proved in Appendix A.1. However, we state in the following example that Assumptions 3.4 and 3.5 are not equivalent. Our interpretation is that Assumption 3.4 is not sensitive to null eigenvalues in the Hessian, while the local strong convexity constant is the minimal eigenvalue of the Hessian at $\theta^\star$ when ε is arbitrarily small.

Example. Let $d = 2$ and $L((\theta_1 \ \theta_2)^\top) = \frac{1}{2}\theta_1^2$. L is not locally strongly convex but satisfies Assumption 3.4 with $D_\eta = \sqrt{\eta}$.

3.2 From Robbins-Siegmund to Azuma-Hoeffding

In this section, we rely on Azuma-Hoeffding inequality to prove the convergence of $L(\theta_t)$ to $L(\theta^\star)$ with a non-asymptotic rate on SGD with the gradient step size $\gamma_t = 1/t^\beta$. We restrict to the case $1/2 < \beta < 1$, for which Robbins-Monro conditions stated in Theorem 3.2 are satisfied.

Our convergence proof crucially relies on the recursive upper bound on the risk provided by Lemma 3.1. In this control, we need to estimate the decrease in expectation, that is, we need to lower bound $\|\nabla L(\theta_t)\|^2$. As in the proof of Theorem 3.2 we rely on Assumption 3.4. These similar condition on Theorem 3.2 and 3.3 highlight the link between both results and motivate our expression of *non-asymptotic Robbins-Monro*.

Theorem 3.3. *Under Assumptions 3.1, 3.2, 3.3 and 3.4, for any $\eta > 0$ and*

$$t \geq \max\Big(\Big(\frac{2(1-\beta)}{D_\eta^2 (1 - (1/2)^{1-\beta})} \Big(\frac{2g^2 h}{2\beta - 1} + L(\theta_1) - L(\theta^\star) \Big) \Big)^{\frac{1}{1-\beta}}, \quad 2 + 2 \Big(\frac{4g^2 h}{\eta} \Big)^{\frac{1}{2\beta - 1}} \Big) ,$$

it holds:

$$\begin{aligned} \mathbb{P}\left(L(\theta_t) - L(\theta^\star) > \eta \right) \leq (1 + t/2) \exp\Big(- D_\eta^4 t^{2(1-\beta)} \Big(\frac{1 - (1/2)^{1-\beta}}{2(1-\beta)} \Big)^2 \frac{(2\beta - 1)}{16 g^4} \Big) \\ + (1 + t/2) \exp\Big(- \eta^2 (t/2 - 2)^{2\beta - 1} \frac{(2\beta - 1)}{128 g^4} \Big) . \end{aligned}$$

The two terms on the right-hand side outline the main idea of the proof. Indeed, similarly as in the Robbins-Siegmund theorem, we rely on Lemma 3.1, and we consider a compromise between two phenomena. On the one hand, the recursive evolution in expectation with $\sum_{t=1}^\infty \gamma_t = \infty$ yields an estimate of the probability that the algorithm stays far from the optimum during a long period. On the other hand, the decrease of the gradient step sizes with $\sum_{t=1}^\infty \gamma_t^2 < \infty$ yields a bound on the probability that the algorithm moves fast. That is illustrated in Figure 3.1.

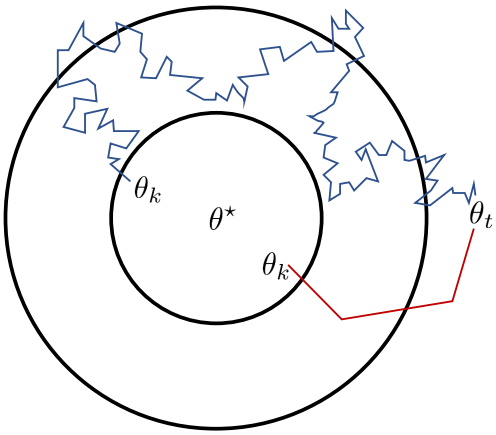


Figure 3.1 – Illustration of the proof of Theorem 3.3. We estimate the probability that the algorithm moves from a small ball around $\theta^\star$ to a larger one. We build on the compromise between the decrease in expectation and the slowness of the algorithm. The blue trajectory is unlikely because the iterates are far from the optimum during a long period. The red trajectory is unlikely because the algorithm jumps from the inner ball to the outer ball in a few steps. We note that Theorem 3.3 is a result on the convergence of $L(\theta_t)$ to $L(\theta^\star)$, not of θ_t to $\theta^\star$ (in particular, $\theta^\star$ may not be unique). This graph is nonetheless more intuitive than a unidimensional graph on $L(\theta_t)$.

Theorem 3.3 implies that for any $\eta > 0$, $\mathbb{P}\left(L(\theta_t) - L(\theta^\star) > \eta\right)$ converges to 0. We translate this into a rate of convergence. To that end, we need an explicit definition of D_η, otherwise we cannot relate both exponential terms of the theorem to one another. Motivated by Proposition 3.1, we consider the case $D_\eta^2 = \eta\mu$, implied by local strong convexity (but not equivalent). Setting $\beta = 3/4$ is optimal.

Corollary 3.1. *We set $\beta = 3/4$. We assume 3.1, 3.2, 3.3 are satisfied, as well as 3.4 for $D_\eta^2 = \eta\mu$ with $0 < \mu \leq 1/\sqrt{2}$. For $t \geq 8$ and $0 < \delta \leq 1$, it holds:*

$$L(\theta_t) - L(\theta^\star) \leq \frac{18g^2(\sqrt{\ln \delta^{-1}} + \sqrt{\ln(t+2)})}{t^{1/4}} \max\left(\mu^{-1}, \left(2h + \frac{L(\theta_1) - L(\theta^\star)}{4g^2}\right)^{1/3}, \left(\frac{h}{10g^2}\right)^{1/2}\right),$$

with probability at least $1 - \delta$.

Corollary 3.1 is proved in Appendix A.2. We now prove the convergence result.

Proof of Theorem 3.3. We start from Lemma 3.1: for any t,

$$L(\theta_{t+1}) \leq L(\theta_t) - \frac{1}{t^\beta}\|\nabla L(\theta_t)\|^2 + \frac{1}{t^{2\beta}}g^2 h + \Delta M_t\,,$$

where $\Delta M_t = -\frac{1}{t^\beta}\nabla L(\theta_t)^\top\Big(\nabla \ell_t(\theta_t) - \nabla L(\theta_t)\Big)$. It yields, for any $k < t$,

$$L(\theta_t) - L(\theta_k) \leq \sum_{s=k}^{t-1}\left(\Delta M_s - \frac{1}{s^\beta}\left\|\nabla L(\theta_s)\right\|^2 + \frac{g^2 h}{s^{2\beta}}\right)\,. \tag{3.1}$$

We are then interested in $\mathbb{P}(L(\theta_t) - L(\theta^\star) > \eta)$ for some $\eta > 0$. For $0 \leq k \leq t$, we define $B_{k,t}$ the event $(\forall k < s < t, L(\theta_s) - L(\theta^\star) > \eta/2)$. Then we use the law of total probability:

$$\begin{aligned}\mathbb{P}(L(\theta_t) - L(\theta^\star) > \eta) &= \mathbb{P}\left((L(\theta_t) - L(\theta^\star) > \eta) \cap B_{0,t}\right) \\ &\quad + \sum_{k=1}^{t-1} \mathbb{P}\left((L(\theta_t) - L(\theta^\star) > \eta) \cap \left(L(\theta_k) - L(\theta^\star) \leq \frac{\eta}{2}\right) \cap B_{k,t}\right) \\ &\leq \mathbb{P}\left((L(\theta_t) - L(\theta^\star) > \eta) \cap B_{0,t}\right) + \sum_{k=1}^{t-1} \mathbb{P}\left(\left(L(\theta_t) - L(\theta_k) > \frac{\eta}{2}\right) \cap B_{k,t}\right).\end{aligned}$$

We get from Equation (3.1) and Assumption 3.4 that for any $1 \leq k < t$,

$$\begin{aligned}\mathbb{P}\left((L(\theta_t) - L(\theta_k) > \eta/2) \cap B_{k,t}\right) &\leq \mathbb{P}\left(\left(\sum_{s=k}^{t-1} \Delta M_s > f(k,t)\right) \cap B_{k,t}\right) \\ &\leq \mathbb{P}\left(\sum_{s=k}^{t-1} \Delta M_s > f(k,t)\right),\end{aligned}$$

where $f(k,t) = \frac{\eta}{2} + D_\eta^2 \sum_{s=k+1}^{t-1} \frac{1}{s^\beta} - g^2 h \sum_{s=k}^{t-1} \frac{1}{s^{2\beta}}$ for any $1 \leq k < t$. Similarly, we get

$$\mathbb{P}\left((L(\theta_t) - L(\theta^\star) > \eta) \cap B_{0,t}\right) \leq \mathbb{P}\left(\sum_{s=1}^{t-1} \Delta M_s > f_0(t)\right),$$

with $f_0(t) = \eta - (L(\theta_1) - L(\theta^\star)) + D_\eta^2 \sum_{s=1}^{t-1} \frac{1}{s^\beta} - g^2 h \sum_{s=1}^{t-1} \frac{1}{s^{2\beta}}$ for any $t \geq 1$.

We have $\mathbb{E}[\Delta M_s \mid x_1, y_1, ..., x_{s-1}, y_{s-1}] = 0$, and almost surely $|\Delta M_s| \leq \frac{2g^2}{s^\beta}$. We can therefore apply Azuma-Hoeffding inequality: for t, k such that $f(k,t) > 0$,

$$\mathbb{P}\left(\sum_{s=k}^{t-1} \Delta M_s > f(k,t)\right) \leq \exp\left(-f(k,t)^2 \frac{(2\beta-1)\max\left(1/2, (k-1)^{2\beta-1}\right)}{8g^4}\right)$$

because $\sum_{s=k}^{t-1} \frac{1}{s^{2\beta}} < \sum_{s=k}^{+\infty} \frac{1}{s^{2\beta}} \leq \frac{1}{(2\beta-1)\max(1/2,(k-1)^{2\beta-1})}$. Similarly, for t such that $f_0(t) > 0$,

$$\mathbb{P}\left(\sum_{s=1}^{t-1} \Delta M_s > f_0(t)\right) \leq \exp\left(-f_0(t)^2 \frac{2\beta-1}{16g^4}\right).$$

We use the following controls on $f(k,t), f_0(t)$:

— If $1 \leq k \leq t/2 - 1$, as $\sum_{s=k+1}^{t-1} \frac{1}{s^\beta} \geq \frac{t^{1-\beta} - (k+1)^{1-\beta}}{1-\beta}$ and $\sum_{s=k}^{t-1} \frac{1}{s^{2\beta}} \leq \frac{2}{(2\beta-1)}$ we obtain:

$$f(k,t) \geq \frac{\eta}{2} - g^2 h \frac{2}{(2\beta-1)} + D_\eta^2 \frac{1 - (1/2)^{1-\beta}}{1-\beta} t^{1-\beta},$$

and for $t \geq \left(\frac{2(1-\beta)}{D_\eta^2(1-(1/2)^{1-\beta})}\left(\frac{2g^2h}{2\beta-1} + L(\theta_1) - L(\theta^\star)\right)\right)^{\frac{1}{1-\beta}}$ we have

$$f(k,t) \geq D_\eta^2 \frac{1-(1/2)^{1-\beta}}{2(1-\beta)} t^{1-\beta}\,,$$
$$f_0(t) \geq D_\eta^2 \frac{1-(1/2)^{1-\beta}}{2(1-\beta)} t^{1-\beta}\,.$$

— If $k > t/2 - 1 \geq 2$, as $\sum_{s=k}^{t-1} \frac{1}{s^{2\beta}} \leq \frac{1}{(2\beta-1)(k-1)^{2\beta-1}}$ we obtain

$$f(k,t) \geq \frac{\eta}{2} - g^2 h \frac{1}{(2\beta-1)(t/2-2)^{2\beta-1}}\,,$$

and for $t \geq 2 + 2\left(\frac{4g^2h}{\eta}\right)^{\frac{1}{2\beta-1}}$ we have $f(k,t) \geq \frac{\eta}{4}$.

This yields the result. □

3.3 Application to Averaged Stochastic Gradient Descent

In this section, we apply our result to averaged SGD. Our sub-optimal convergence rates are thus motivated by a two-step procedure relying first on the high probability convergence bound and then accelerated to obtain optimal bounds. Averaging the estimates of SGD is known to accelerate the convergence (Polyak and Juditsky, 1992). For any n, we define the averaged iterate:

$$\overline{\theta}_n = \frac{1}{n}\sum_{t=1}^{n} \theta_t\,.$$

We provide an upper bound on the excess risk $L(\overline{\theta}_n) - L(\theta^\star)$. To that end, we use Jensen's inequality:

$$L(\overline{\theta}_n) - L(\theta^\star) \leq \frac{1}{n}\sum_{t=1}^{n} (L(\theta_t) - L(\theta^\star))\,, \tag{3.2}$$

and we estimate the cumulative excess risk.

We start from the definition of SGD update. In the adversarial setting, we derive the following bound on the first-order Taylor expansion of the excess risk. It holds without the need for i.i.d. assumption or any form of convexity assumption.

Lemma 3.2. *For any $1 \leq k \leq n$ and $\theta \in \mathbb{R}^d$, the following inequality holds in the adversarial setting:*

$$\sum_{t=k}^{n} \nabla L(\theta_t)^\top (\theta_t - \theta) \leq \sum_{t=k}^{n} \Delta N_t + \frac{1}{2}\sum_{t=k}^{n} \|\theta_t - \theta\|^2 \frac{\beta}{t^{1-\beta}} + \frac{1}{2}\sum_{t=k}^{n} \frac{\|\nabla \ell_t(\theta_t)\|^2}{t^\beta}\,,$$

where $\Delta N_t = (\nabla L(\theta_t) - \nabla \ell_t(\theta_t))^\top (\theta_t - \theta)$.

From this result, we could obtain sub-optimal rates under Assumption 3.4. Indeed, from Theorem 3.3 we obtain the convergence of θ_t to $\theta^\star$. Therefore, the first term of the right-hand

side of Lemma 3.2 is of order $O(\sqrt{n})$, the second term of order $O(n^\beta)$ and the third of order $O(n^{1-\beta})$. As $\beta > 1/2$, the largest is $O(n^\beta)$ asymptotically. Using Equation (3.2) and the first-order convexity bound, we would bound the excess risk in $O(1/n^{1-\beta})$.

We show in the following that under local strong convexity, we accelerate to the optimal $O(1/n)$ excess risk, with high probability.

In the vein of Corollary 3.1 we obtain the following corollary. For a locally strongly convex risk, for any ball arbitrarily small around the optimum, we define explicitly a convergence time such that the iterates of the SGD are trapped in the local ball after that convergence time with high probability.

Corollary 3.2. *We set $\beta = 3/4$ and we assume 3.1, 3.2, 3.3, and 3.5. For any $\varepsilon, \delta > 0$, we define*

$$k = \left(\frac{2560 g^4}{\mu_\varepsilon^4 \varepsilon^4} \Big(\ln \delta^{-1} + \ln(g^4/(\mu_\varepsilon^4 \varepsilon^4)) + 9 \Big) \right)^2 .$$

Then it holds $\mathbb{P}\left(\bigcap_{t=k+1}^{\infty} (\|\theta_t - \theta^\star\| \leq \varepsilon) \right) \geq 1 - \delta$.

Motivated by this convergence property, we propose a two-step analysis for the averaged SGD, and we decompose the cumulative risk into two parts. For a given ε and δ, we have a first convergence phase that we bound using Corollary 3.2. The second phase is localized: the iterates of SGD are trapped in the ball of radius ε around the optimum with high probability. In that localized phase, we use local strong convexity in order to obtain the optimal rate.

We consider a final assumption which is that the gradients are Lipschitz. We already assumed the gradient of the risk is Lipschitz, due to the second point of Assumption 3.1, but the assumption on the losses is a little stronger.

Assumption 3.6. *For any $\theta \in \mathbb{R}^d$ such that $\|\theta - \theta^\star\| \leq \varepsilon$ it holds $\|\nabla \ell_t(\theta)\| \leq C_{\mathrm{Lip}} \|\theta - \theta^\star\|$ almost surely for any t.*

We can now state our result for the averaged SGD.

Theorem 3.4. *We assume 3.1, 3.2, 3.3, 3.5 and 3.6 are satisfied. We assume that we have an integer $k \geq 1$ satisfying $k \geq (4 \max(1, C_{\mathrm{Lip}}^2)/\mu_\varepsilon)^{1/\beta}$ and $\mathbb{P}(\cap_{t=k+1}^{\infty}(\|\theta_t - \theta^\star\| \leq \varepsilon)) \geq 1 - \delta$. Then for any $n \geq 1$ and $\delta > 0$, it holds*

$$L(\overline{\theta}_n) - L(\theta^\star) \leq \frac{16 g^2 \ln \delta^{-1}}{\mu_\varepsilon n} + \frac{1}{n} \sum_{t=1}^{k} (L(\theta_t) - L(\theta^\star)) ,$$

with probability at least $1 - 2\delta$.

The first term of the bound $O(g^2 \ln \delta^{-1}/(\mu_\varepsilon n))$ is optimal. However, we remark that if we set $\beta = 3/4$ and if we use the value of k provided by Corollary 3.2, then the leading term in Theorem 3.4 is the second term. Indeed, applying Corollary 3.1, we should be able to bound this initializing cumulative excess risk (from 1 to k) as $O(g^2 \mu_\varepsilon^{-1}(\sqrt{\ln \delta} + \sqrt{\ln k}) k^{3/4})$. Removing logarithmic terms and applying Corollary 3.2, we would then obtain

$$\sum_{t=1}^{k} (L(\theta_t) - L(\theta^\star)) = O\Big(\frac{g^8}{\mu_\varepsilon^7 \varepsilon^6} (\ln \delta^{-1})^2 \Big) . \tag{3.3}$$

We obtain the fast $1/n$ rate but the leading constant is far from the optimal $g^2 \ln \delta^{-1}/\mu_\varepsilon$. There is room for improvement:

— The first way to improve Theorem 3.4 is to consider a gradient step $\gamma_t = \gamma/t^\beta$ for a well-chosen γ depending on the horizon n as in Bach, 2014. Indeed, if we set $1/2 < \beta < 3/4$ and $\gamma = 1/n^{(3-4\beta)/2}$, the proofs of Theorem 3.3 and Corollary 3.1 are almost unchanged and we claim it yields an excess risk with high probability of $O(g^2\mu^{-1}\sqrt{\ln \delta^{-1}}/n^{1-\beta})$ for iterate n (of SGD, not the averaged version). This yields an improvement on Corollary 3.2, which in turn propagates to Theorem 3.4.
This strategy would reduce the exponents of Equation (3.3) but the second term of Theorem 3.4 will still be greater than $O(g^2\mu_\varepsilon^{-1} \ln \delta^{-1}/n)$.

— A second lead would be to relax the convergence property. Indeed, the assumption of Theorem 3.4 (that we prove) is that the algorithm stays trapped in a small region around the optimum. The property that with high probability, the algorithm does not leave this small region is substantial. However, it could be possible to relax that property and to split the cumulative excess risk into two sums. On the one hand, the sum over the time steps when the algorithm is inside the ball around the optimum would be estimated similarly with $O(g^2\mu_\varepsilon^{-1} \ln \delta^{-1}/n)$. On the other hand, the second sum would be smaller than the one of Theorem 3.4 by definition of k.

— Finally, a more structural change would be to estimate k and average only after the first k steps. Algorithmically, it may be possible to compute the average only on the last steps included in a small ball around the last iterate.

3.4 Conclusion

In this chapter, we proposed a new analysis of stochastic gradient descent. Relying on Azuma-Hoeffding inequality, we proved a strong convergence property. With high probability, the estimates stay trapped in a region arbitrarily small around the optimum. This motivates a two-phase analysis, using that convergence and tighter bounds in the local phase. We applied that idea in order to obtain a high probability bound on the excess risk of averaged stochastic gradient descent.

We apply a very similar analysis in Chapter 4, where we analyze the static setting of the extended Kalman filter.

Chapter 4

Stochastic Online Optimization using Kalman Recursion

We study the Extended Kalman Filter in constant dynamics, offering a bayesian perspective of stochastic optimization. For generalized linear models, we obtain high probability bounds on the cumulative excess risk in an unconstrained setting, under the assumption that the algorithm reaches a local phase. In order to avoid any projection step we propose a two-phase analysis. First, for linear and logistic regressions, we prove that the algorithm enters a local phase where the estimate stays in a small region around the optimum. We provide explicit bounds with high probability on this convergence time, slightly modifying the Extended Kalman Filter in the logistic setting. Second, for generalized linear regressions, we provide a martingale analysis of the excess risk in the local phase, improving existing ones in bounded stochastic optimization. The algorithm appears as a parameter-free online procedure that optimally solves some unconstrained optimization problems.

This chapter is based on a joint work with Olivier Wintenberger published in Journal of Machine Learning Research.

Contents

4.1 Introduction

The optimization of convex functions is a long-standing problem with many applications. In supervised machine learning it frequently arises in the form of the prediction of an observation $y_t \in \mathbb{R}$ given explanatory variables $X_t \in \mathbb{R}^d$. The aim is to minimize a cost depending on the prediction and the observation. We focus in this article on linear predictors, hence the loss function is of the form $\ell(y_t, \theta^\top X_t)$.

Two important settings have emerged in order to analyse learning algorithms. In the online setting (X_t, y_t) may be set by an adversary. The assumption required is boundedness and the goal is to upper estimate the regret (cumulative excess loss compared to the optimum). In the stochastic setting with independent identically distributed (i.i.d.) (X_t, y_t), we define the risk $L(\theta) = \mathbb{E}[\ell(y, \theta^\top X)]$. We focus on the cumulative excess risk $\sum_{t=1}^n L(\hat{\theta}_t) - L(\theta^\star)$ where $\theta^\star$ minimizes the risk. We obtain bounds holding with high probability simultaneously for any horizon, that is, we control the whole trajectory of the risk. Furthermore, our bounds on the cumulative risk all lead to similar ones on the excess risk at any step for the averaged version of the algorithm.

Due to its low computational cost the Stochastic Gradient Descent (SGD) of Robbins and Monro, 1951 has been widely used, along with its equivalent in the online setting, the online gradient descent (Zinkevich, 2003) and a simple variant where the iterates are averaged (Ruppert, 1988; Polyak and Juditsky, 1992). More recently Bach and Moulines, 2013 provided a sharp bound in expectation on the excess risk for a two-step procedure that has been extended to the average of SGD with a constant step size (Bach, 2014). Second-order methods based on stochastic versions of Newton-Raphson algorithm have been developed in order to converge faster in iterations, although with a bigger computational cost per iteration (Hazan, Agarwal, and Kale, 2007).

In order to obtain a parameter-free second-order algorithm we apply a bayesian perspective, seeing the loss as a negative log-likelihood and approximating the maximum-likelihood estimator at each step. We get a state-space model interpretation of the optimization problem: in a well-specified setting the space equation is $y_t \sim p_{\theta_t}(\cdot \mid X_t) \propto \exp(-\ell(\cdot, \theta_t^\top X_t))$ with $\theta_t \in \mathbb{R}^d$ and the state equation defines the dynamics of the state θ_t. The stochastic convex optimization setting corresponds to a degenerate constant state-space model $\theta_t = \theta_{t-1}$ called static. As usual in State-Space models, the optimization is realized with the Kalman recursion (Kalman and Bucy, 1961) for the quadratic loss and the Extended Kalman Filter (EKF) (Fahrmeir, 1992) in a more general case. A correspondence has recently been made by Ollivier, 2018 between the static EKF and the online natural gradient (Amari, 1998). This motivates a risk analysis in order to enrich the link between Kalman filtering and the optimization community. We may see the static EKF as the online approximation of bayesian model averaging, and similarly to its analysis derived by Kakade and Ng, 2005 our analysis is robust to misspecification, that is we don't assume the data to be generated by the probabilistic model.

The static EKF is very close to the Online Newton Step (ONS) introduced by Hazan, Agarwal, and Kale, 2007 as both are second-order online algorithms and our results are of the same flavor as those obtained on the ONS (Mahdavi, Zhang, and Jin, 2015). However the ONS requires the knowledge of the region in which the optimization is realized. It is involved in the choice of the gradient step size and a projection step is done to ensure that the search stays in the chosen

region. On the other hand the static EKF yields two advantages at the cost of being less generic.

First, there is no costly projection step and each recursive update runs in $O(d^2)$ operations, where d is the dimension of the features X_t. Therefore, our comparison of the static EKF with the ONS provides a lead to the open question of Koren, 2013. Indeed, the problem of the ONS pointed out by Koren, 2013 is to control the cost of the projection step and the question is whether it is possible to perform better than the ONS in the stochastic exp-concave setting. We don't answer the open question in the general setting. However, we suggest a general way to get rid of the projection by dividing the analysis between a convergence proof of the algorithm to the optimum and a second phase where the estimate stays in a small region around the optimum where no projection is required.

Second, the algorithm is (nearly) parameter-free. We believe that bayesian statistics is the reasonable approach in order to obtain parameter-free online algorithms in the unconstrained setting. Parameter-free is not exactly correct as there are initialization parameters, which we see as a smoothed version of the hard constraint imposed by bounded algorithm, but they have no impact on the leading terms of our bounds. Static Kalman filter coincides with the ridge forecaster and similarly the static EKF may be seen as the online approximation of a regularized empirical risk minimizer.

4.1.1 Related Work

Theoretical guarantees for online and stochastic algorithms are multi-criteria and of various natures. The comparison of upper-bounds or computational complexity highly depends on the values of d the dimension of the explanatory vectors and n the time horizon, leading to different views on whether the dependence on d or n is the most important. The nature of the guarantee obviously depends on the objective pursued.

In the advesarial setting, the learner suffers a loss $\ell_t(\hat{\theta}_t)$ depending on its estimate $\hat{\theta}_t$ at each time step t. It is natural to minimize the cumulative loss, or equivalently the regret

$$\sum_{t=1}^{n} \ell_t(\hat{\theta}_t) - \sum_{t=1}^{n} \ell_t(\theta^\star)\,,$$

where $\theta^\star$ reaches the minimum value of the cumulative loss and thus highly depends on $(\ell_t)_{1\leq t\leq n}$. Under an assumption of bounded gradients, Zinkevich, 2003 proved that a first-order online gradient descent yields a regret bound in $O(\sqrt{n})$. The Online Newton Step (ONS) is a second-order online gradient descent that has been designed to obtain a regret bound in $O(\ln n)$ (Hazan, Agarwal, and Kale, 2007) under the assumption that the losses are exp-concave. The improved guarantee comes at a cost of $O(d^2)$ operations per step instead of $O(d)$, along with a projection at each step whose cost depends on the data.

In the stochastic setting where the losses (ℓ_t) are assumed i.i.d., the aim is to minimize the risk $L(\theta) = \mathbb{E}[\ell(\theta)]$. A natural candidate is the Empirical Risk Minimizer (ERM), whose asymptotics are well understood (see for example Murata and Amari, 1999). Assuming the existence of $\theta^\star$ minimizing the risk and that the Hessian matrix $H^\star = \frac{\partial^2}{\partial\theta^2} L(\theta^\star)$ is positive definite, the ERM $\hat{\theta}_n$ satisfies

$$\mathbb{E}[L(\hat{\theta}_n)] - L(\theta^\star) = \frac{\operatorname{tr}(G^\star H^{\star -1})}{2n} + o(1/n)\,, \qquad G^\star = \mathbb{E}\Big[\frac{\partial}{\partial\theta}\ell(y, \theta^{\star\top} X)\frac{\partial}{\partial\theta}\ell(y, \theta^{\star\top} X)^\top\Big]\,.$$

Although in the well-specified setting the identity $\operatorname{tr}(G^\star H^{\star -1}) = d$ holds, in the misspecified case there is no general estimate for $\operatorname{tr}(G^\star H^{\star -1})$. Recently a non-asymptotic bound

	ERM	Averaged SGD	ONS	**This article**
Regret			$O(\ln n)$	
Excess risk in expectation	$O(\frac{1}{n})$	$O(\frac{1}{n})$		
Excess risk w.h.p.	$O(\frac{\ln \delta^{-1}}{n})$	$O(\frac{\ln \delta^{-1}}{\sqrt{n}})$		
Cumulative excess risk w.h.p.			$O(\ln n + \ln \delta^{-1})$	$O(\ln n + \ln \delta^{-1} + S(\delta))$
Cost per iteration	Implicit	$O(d)$	$O(d^2) + T_{\text{proj}}$	$O(d^2)$

Table 4.1 – Summary of existing results along with the static EKF for which we prove the bound for the cumulative excess risk. We focus on the dependence on n, and δ for the bounds holding with probability $1-\delta$ (w.h.p.). $S(\delta)$ is the cumulative excess risk of the convergence phase. In Chapter 3, we prove that averaged SGD can achieve an excess risk $O((\ln \delta^{-1})^2/n)$ with probability $1-\delta$, but with sub-optimal constants (Theorem 3.4).

$L(\hat{\theta}_n) - L(\theta^\star) = O(\mathrm{tr}(G^\star H^{\star -1}) \ln \delta^{-1}/n)$ holding with probability $1-\delta$ has been shown by Ostrovskii and Bach, 2021 on the ERM. However the ERM is defined only implicitly and may have important computational cost, hence recursive algorithms based on gradient descent have been studied under different sets of assumptions to bound $\mathrm{tr}(G^\star H^{\star -1})$.

The most simple is Stochastic Gradient Descent (SGD), where each step is in the opposite direction of the gradient. This algorithm has been widely used with various step sizes. Sharp results have been obtained by Bach, 2014 for a constant gradient step size $C/\sqrt{n}$ with fixed horizon n. Under the assumption that gradients are bounded by R we have $\mathrm{tr}(G^\star H^{\star -1}) \leq R^2/\mu$ where μ is the minimal eigenvalue of $H^\star$. The fast rate $\mathbb{E}[L(\overline{\theta}_n)] - L(\theta^\star) = O(R^2/(\mu n))$ is obtained by Bach, 2014 for the averaged estimate $\overline{\theta}_n$ of SGD. In the same article the author also derives a bound with high probability but with a slower rate: it degrades into $L(\overline{\theta}_n) - L(\theta^\star) = O(\log \delta^{-1}/\sqrt{n})$ with probability $1-\delta$. Finally, in the quadratic setting a fast rate $L(\overline{\theta}_n) - L(\theta^\star) = O(1/(n\delta^\alpha))$ is achieved with probability $1-\delta$ for a defined $\alpha > 0$ (Bach and Moulines, 2013).

To obtain fast rates with high probability beyond the quadratic setting, it seems necessary to use second-order information as in the ONS (Mahdavi, Zhang, and Jin, 2015). Under the assumption that the loss is α-exp-concave, $\mathrm{tr}(G^\star H^{\star -1}) \leq d/\alpha$ and for the averaged version of the ONS the rate $L(\overline{\theta}_n) - L(\theta^\star) = O(d(\ln n + \ln \delta^{-1})/(\alpha n))$ with probability $1-\delta$ is obtained. From our perspective, the result is stronger than what is claimed by Mahdavi, Zhang, and Jin, 2015: the bound obtained is

$$\sum_{t=1}^{n} L(\hat{\theta}_n) - L(\theta^\star) = O(\ln n + \ln \delta^{-1})\,, \tag{4.1}$$

holding simultaneously for any n with probability $1-\delta$. Note that although averaging this bound with Jensen's inequality leads to a sub-optimal bound on the excess risk of the last averaged estimate, it is conversely not possible to obtain Equation (4.1) from

$$L(\hat{\theta}_n) - L(\theta^\star) = O(\ln \delta^{-1}/n)\,,$$

holding with probability $1-\delta$.

4.1.2 Contributions

Our first contribution is a local analysis of the static EKF under assumptions defined in Section 4.2, and provided that consecutive steps stay in a small ball around the optimum $\theta^\star$. We

derive local bounds on the cumulative risk with high probability from a martingale analysis. Our analysis of Section 4.3 is similar to the one of Mahdavi, Zhang, and Jin, 2015 and we slightly refine their constants as a by-product.

We then show that the convergence property crucial in our analysis is reachable. To that end we focus on linear regression and logistic regression as these two well-known problems are challenging in the unconstrained setting. In linear regression, the gradient of the loss is not bounded globally. In logistic regression, the loss is strictly convex, but neither strongly convex nor exp-concave in the unconstrained setting. In Section 4.4, we develop a global bound in the logistic setting on a slight modification of the algorithm introduced by Bercu, Godichon, and Portier, 2020. We prove that this modified algorithm converges and stays into a local region around $\theta^\star$ after a finite number of steps. Moreover we show that it coincides with the static EKF and thus our local analysis applies. In Section 4.5, we apply our local analysis to the quadratic setting. We rely on Hsu, Kakade, and Zhang, 2012 to obtain the convergence of the algorithm after exhibiting the correspondence between Kalman filter in constant dynamics and the ridge forecaster, and we therefore obtain similarly a global bound.

Finally, we demonstrate numerically the competitiveness of the static EKF for logistic regression in Section 4.6.

4.2 Definitions and Assumptions

We consider loss functions that may be written as the negative log-likelihood of a generalized linear model (McCullagh and Nelder, 1989). Formally, the loss is defined as $\ell(y, \theta^\top X) = -\log p_\theta(y \mid X)$ where $\theta \in \mathbb{R}^d$, $(X, y) \in \mathcal{X} \times \mathcal{Y}$ for some $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ and p_θ is of the form

$$p_\theta(y \mid X) = h(y) \exp\left(\frac{y\,\theta^\top X - b(\theta^\top X)}{a}\right), \tag{4.2}$$

where a is a constant and h and b are one-dimensional functions on which a few assumptions are required (Assumption 4.3). This includes logistic and quadratic regressions, see Sections 4.4 and 4.5. We display the static EKF in Algorithm 1 in this setting (see Appendix B.1 for a derivation relying on Durbin and Koopman, 2012). In the quadratic setting, noting that the EKF estimate $\hat{\theta}_t$ does not depend on the (unknown) variance σ^2, we consider the quadratic loss $\ell(y, \hat{y}) = (y - \hat{y})^2/2$ by convention.

Algorithm 1 : Static Extended Kalman Filter for Generalized Linear Model

1. *Initialization*: P_1 is any positive definite matrix, $\hat{\theta}_1$ is any initial parameter in $\mathbb{R}^d$.
2. *Iteration*: at each time step $t = 1, 2, \ldots$
   (a) Update $P_{t+1} = P_t - \frac{P_t X_t X_t^\top P_t}{1 + X_t^\top P_t X_t \alpha_t}\alpha_t$ with $\alpha_t = \frac{b''(\hat{\theta}_t^\top X_t)}{a}$.
   (b) Update $\hat{\theta}_{t+1} = \hat{\theta}_t + P_{t+1}\frac{(y_t - b'(\hat{\theta}_t^\top X_t))X_t}{a}$.

Due to matrix-vector and vector-vector multiplications, Algorithm 1 has a running-time complexity of $O(d^2)$ at each iteration and thus $O(nd^2)$ for n iterations.

Note that although we need the loss function to be derived from a likelihood of the form (4.2), we do not need the data to be generated under this process. We need two standard hypotheses on the data. The first one is the i.i.d. assumption and bounded random design (all along the

paper $\|.\|$ is the Euclidean norm):

Assumption 4.1. *The observations $(X_t, y_t)_t$ are i.i.d. copies of the pair $(X, y) \in \mathcal{X} \times \mathcal{Y}$ and there exists D_X such that $\|X_t\| \leq D_X$ almost surely (a.s.).*

Working under Assumption 4.1, we define the risk function $L(\theta) = \mathbb{E}\left[\ell(y, \theta^\top X)\right]$. Note that in Section 4.3 we don't need $\mathbb{E}[XX^\top]$ invertible, but we will make such an assumption in Sections 4.4 and 4.5 to prove the convergence of the algorithm in the logistic and quadratic settings, respectively. In order to work on a well-defined optimization problem we assume there exists a minimum:

Assumption 4.2. *There exists $\theta^\star \in \mathbb{R}^d$ such that $L(\theta^\star) = \inf_{\theta \in \mathbb{R}^d} L(\theta)$.*

We treat two different settings requiring different assumptions, summarized in Assumption 4.3 and 4.4 respectively. First, motivated by logistic regression we define:

Assumption 4.3. *There exists $(\kappa_\varepsilon)_{\varepsilon>0}, (h_\varepsilon)_{\varepsilon>0}$ and $\rho_\varepsilon \xrightarrow[\varepsilon \to 0]{} 1$ such that for any $\varepsilon > 0$ and any $\theta, \theta_0 \in \mathbb{R}^d$ satisfying $max(\|\theta - \theta^\star\|, \|\theta_0 - \theta^\star\|) \leq \varepsilon$, we have*
- *$\ell'(y, \theta^\top X)^2 \leq \kappa_\varepsilon \ell''(y, \theta^\top X)$ a.s.*
- *$\ell''(y, \theta^\top X) \leq h_\varepsilon$ a.s.*
- *$\ell''(y, \theta^\top X) \geq \rho_\varepsilon \ell''(y, \theta_0^\top X)$ a.s.*

Here ℓ' and ℓ'' are the first and second derivatives of ℓ with respect to the second variable.

Assumption 4.3 requires local exp-concavity (around $\theta^\star$) along with some regularity on ℓ'' (ℓ'' continuous and $\ell''(y, \theta^{\star\top} X) \geq \gamma > 0$ a.s. is sufficient). That setting implies $\mathcal{Y}$ bounded, because ℓ' depends on y whereas ℓ'' doesn't. In logistic regression, $\mathcal{Y} = \{-1, +1\}$ and Assumption 4.3 is satisfied for $\kappa_\varepsilon = e^{D_X(\|\theta^\star\| + \varepsilon)}, h_\varepsilon = \frac{1}{4}, \rho_\varepsilon = e^{-\varepsilon D_X}$.

Second, we consider the quadratic loss, corresponding to a Gaussian model. In order to include the well-specified setting and to bound $G^\star = \mathbb{E}[(y - \theta^{\star\top} X)^2 XX^\top]$, we assume y sub-gaussian conditionally to X and not too far away from the model:

Assumption 4.4. *The distribution of $(X, y) \in \mathcal{X} \times \mathcal{Y}$ satisfies*
- *There exists $\sigma^2 > 0$ such that for any $s \in \mathbb{R}$, $\mathbb{E}\left[e^{s(y - \mathbb{E}[y|X])} \mid X\right] \leq e^{\frac{\sigma^2 s^2}{2}}$ a.s.,*
- *There exists $D_{\text{app}} \geq 0$ such that $|\mathbb{E}[y \mid X] - \theta^{\star\top} X| \leq D_{\text{app}}$ a.s.*

Both conditions of Assumption 4.4 hold with $\mathcal{Y} = \mathbb{R}$ and $D_{\text{app}} = 0$ for the well-specified Gaussian linear model with random bounded design. The second condition of Assumption 4.4 is satisfied for $D_{\text{app}} > 0$ in misspecified sub-gaussian linear model with a.s. bounded approximation error.

4.3 The Algorithm Around the Optimum

In this section, we analyse the cumulative risk under a strong convergence assumption. Precisely we define

$$\tau(\varepsilon) = \min\{k \in \mathbb{N} \mid \forall t > k, \|\hat{\theta}_t - \theta^\star\| \leq \varepsilon\},$$

where $(\hat{\theta}_t)_t$ are the estimates of the static EKF, and with the convention $\min \emptyset = +\infty$. We assume a bound on $\tau(\varepsilon)$ holding with high probability:

Assumption 4.5. *For any $\delta, \varepsilon > 0$, there exists $T(\varepsilon, \delta) \in \mathbb{N}$ such that*

$$\mathbb{P}\big(\tau(\varepsilon) \leq T(\varepsilon, \delta)\big) \geq 1 - \delta\,.$$

Assumption 4.5 states that with high probability there exists a convergence time after which the algorithm stays trapped in a local region around the optimum. Sections 4.4 and 4.5 are devoted to define explicitly such a convergence time for a modified EKF in the logistic setting and for the true EKF in the quadratic setting.

We present our result in the bounded and sub-gaussian settings. The results and their proofs are very similar, but two crucial steps are different. First, Assumption 4.3 yields a bound on the gradient holding almost surely. We relax the boundedness condition for the quadratic loss with a sub-gaussian hypothesis, requiring a specific analysis. Second, our analysis is based on a second-order expansion. The quadratic loss is equal to its second-order Taylor expansion but we need Assumption 4.5 along with the third point of Assumption 4.3 otherwise.

The following theorem is our result in the bounded setting.

Theorem 4.1. *Starting the static EKF from any $\hat{\theta}_1 \in \mathbb{R}^d, P_1 \succ 0$, if Assumptions 4.1, 4.2, 4.3, 4.5 are satisfied and if $\rho_\varepsilon > 0.95$, for any $\delta > 0$, it holds for any $n \geq 1$ simultaneously*

$$\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} L(\hat{\theta}_t) - L(\theta^\star) \leq \frac{5}{2} d\kappa_\varepsilon \ln\left(1 + n\frac{h_\varepsilon \lambda_{\max}(P_1) D_X^2}{d}\right) + 5\lambda_{\max}\left(P_{T(\varepsilon,\delta)+1}^{-1}\right)\varepsilon^2 + 30\left(2\kappa_\varepsilon + h_\varepsilon \varepsilon^2 D_X^2\right)\ln \delta^{-1}\,,$$

with probability at least $1 - 3\delta$.

The constant 0.95 may be chosen arbitrarily close to 0.5 with growing constants in the bound on the cumulative risk. There is a hidden trade-off in ε: on the one hand, the smaller ε the better our upper-bound, but on the other hand $T(\varepsilon, \delta)$ increases when ε decreases, and thus our bound applies after a bigger convergence time.

For the quadratic loss, we obtain the following result under the sub-gaussian hypothesis.

Theorem 4.2. *In the quadratic setting, starting the static EKF from any $\hat{\theta}_1 \in \mathbb{R}^d, P_1 \succ 0$, if Assumptions 4.1, 4.2, 4.4 and 4.5 are satisfied, for any $\delta > 0$ and any $\varepsilon > 0$, it holds for any $n \geq 1$ simultaneously*

$$\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} L(\hat{\theta}_t) - L(\theta^\star) \leq \frac{15}{2} d\left(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2\right)\ln\left(1 + n\frac{\lambda_{\max}(P_1) D_X^2}{d}\right) + 5\lambda_{\max}\left(P_{T(\varepsilon,\delta)+1}^{-1}\right)\varepsilon^2 + 115\left(\sigma^2\left(4 + \frac{\lambda_{\max}(P_1) D_X^2}{4}\right) + D_{\text{app}}^2 + 2\varepsilon^2 D_X^2\right)\ln \delta^{-1}\,,$$

with probability at least $1 - 5\delta$.

We observe a similar trade-off in ε as in Theorem 4.1. Up to numerical constants, the tight constant $d(\sigma^2 + D_{\text{app}}^2)$ (see for instance Hsu, Kakade, and Zhang, 2012) is achieved by choosing ε arbitrarily small, at the cost of a loose control of the $T(\varepsilon, \delta)$ first steps.

Both results follow from a regret analysis close to the one on the ONS (see Section 4.3.1), and on a control on the martingales stated below:

Lemma 4.1. *Let $k \geq 0$ and $(\Delta N_t)_{t>k}$ be any martingale difference adapted to the filtration $(\mathcal{F}_t)_{t\geq k}$ such that for any $t > k$, $\mathbb{E}[\Delta N_t^2 \mid \mathcal{F}_{t-1}] < \infty$. For any $\delta, \lambda > 0$, we have the simultaneous*

Algorithm 2 : Recursive updates of the ONS and the static EKF

Online Newton Step	**Static Extended Kalman Filter**
$P_{t+1}^{-1} = P_t^{-1} + \ell'(y_t, w_t^\top X_t)^2 X_t X_t^\top \,,$	$P_{t+1}^{-1} = P_t^{-1} + \ell''(y_t, \hat{\theta}_t^\top X_t) X_t X_t^\top \,,$
$\nabla_t = \ell'(y_t, w_t^\top X_t) X_t \,,$	$\nabla_t = \ell'(y_t, \hat{\theta}_t^\top X_t) X_t \,,$
$w_{t+1} = \prod_{\mathcal{K}}^{P_{t+1}^{-1}} \left(w_t - \frac{1}{\gamma} P_{t+1} \nabla_t \right) \,,$	$\hat{\theta}_{t+1} = \hat{\theta}_t - P_{t+1} \nabla_t \,,$

where $\prod_{\mathcal{K}}^{P_{t+1}^{-1}}$ is the projection on $\mathcal{K}$ for the norm $\|.\|_{P_{t+1}^{-1}}$.

property

$$\sum_{t=k+1}^{k+n} \left(\Delta N_t - \frac{\lambda}{2} ((\Delta N_t)^2 + \mathbb{E}[(\Delta N_t)^2 \mid \mathcal{F}_{t-1}]) \right) \leq \frac{\ln \delta^{-1}}{\lambda}, \qquad n \geq 1 \,,$$

with probability at least $1 - \delta$.

This result proved in Appendix B.2.1 is a corollary of a martingale inequality from Bercu and Touati, 2008 and a stopping time construction (Freedman, 1975).

We detail the key ideas of the proofs in the rest of the Section, and we defer to Appendix B.2 the proof of the intermediate results along with the detailed proof of Theorems 4.1 and 4.2. Specifically, we display in Section 4.3.1 the parallel with the ONS, where we compare with the existing result on the cumulative risk, and a similar analysis yields an adversarial bound on a second-order expansion of the loss. In Section 4.3.2 we compare the excess risk with its second-order expansion thanks to Assumption 4.5, and we use a martingale analysis to obtain a bound on the cumulative excess risk.

4.3.1 Comparison with Online Newton Step and Adversarial Analysis

We display the parallel between the ONS and the static EKF in Algorithm 2 through their recursive updates. We observe that the square of the first derivative of ℓ is replaced with the second derivative. Thus tP_t^{-1} in the static EKF is an estimate of the Hessian $H^\star$ which is the optimal preconditioning matrix as shown in Corollary 3 of Murata and Amari, 1999. Then the recursion on the parameter (w_t and $\hat{\theta}_t$) has two differences: there is a gradient step size $1/\gamma$ in the ONS absent in the static EKF, and after the gradient step the ONS applies a projection. Lemma 4.1 yields the following refinement on the bound of Mahdavi, Zhang, and Jin, 2015 obtained on the cumulative excess risk of the ONS:

Corollary 4.1. *Assume the search region $\mathcal{K}$ has diameter D and the gradients are bounded by R. Let $(w_t)_t$ be the ONS estimates starting from $w_1 \in \mathcal{K}, P_1 = \lambda I$ and using a step-size $\gamma = \frac{1}{2}\min(\frac{1}{4RD}, \alpha)$ with α the exp-concavity constant of ℓ on $\mathcal{K}$. Then for any $\delta > 0$, it holds for any $n \geq 1$ simultaneously*

$$\sum_{t=1}^{n} L(w_t) - L(\theta^\star) \leq \frac{3}{2\gamma} d \ln \left(1 + \frac{nR^2}{\lambda d} \right) + \frac{\lambda \gamma}{6} D^2 + \left(\frac{12}{\gamma} + \frac{4\gamma R^2 D^2}{3} \right) \ln \delta^{-1} \,,$$

with probability at least $1 - 2\delta$.

For the sake of consistency, we display Corollary 4.1 as a bound on the cumulative excess risk, whereas Theorem 3 of Mahdavi, Zhang, and Jin, 2015 is a bound on the excess risk of the averaged ONS. The latter follows directly from an application of Jensen's inequality. The proof of Corollary 4.1 consists in replacing Theorem 4 of Mahdavi, Zhang, and Jin, 2015 with Lemma 4.1. We obtain similar constants in Theorem 4.1 and in Corollary 4.1, as κ_ε is the inverse of the exp-concavity constant α. The use of second-order methods with well-tuned preconditioning is crucial in order to replace the leading constant R^2/μ obtained for first-order methods by d/α (μ is the minimum eigenvalue of the hessian $H^\star$).

Our results on the static EKF are less general than the ones obtained on the ONS as a control of the convergence time $\tau(\varepsilon) \leq T(\varepsilon, \delta)$ is required with high probability. On the other hand the results obtained on the ONS require the knowledge of the exp-concavity constant α whereas the static EKF is parameter-free. That is why we argue that the static EKF provides an optimal way to tune the step size and the preconditioning matrix. Indeed, as ε is a parameter of the EKF analysis but not of the algorithm, we can improve the leading constant κ_ε on a local region arbitrarily small around $\theta^\star$, at a cost of a loose control of the $T(\varepsilon, \delta)$ first steps. In the ONS the choice of a diameter $D > \|\theta^\star\|$ makes the gradient step-size sub-optimal and impacts the leading constant.

Once the parallel between the ONS and the static EKF has been displayed (Algorithm 2), it is natural to adopt an approach similar to the one in Hazan, Agarwal, and Kale, 2007. The cornerstone of our local analysis is the derivation of an adversarial bound on the second-order Taylor expansion of ℓ, from the recursive update formulae.

Lemma 4.2. *For any sequence $(X_t, y_t)_t$, starting from any $\hat{\theta}_1 \in \mathbb{R}^d, P_1 \succ 0$, it holds for any $\theta^\star \in \mathbb{R}^d$ and $n \in \mathbb{N}$ that*

$$\sum_{t=1}^{n} \left(\left(\ell'(y_t, \hat{\theta}_t^\top X_t) X_t \right)^\top (\hat{\theta}_t - \theta^\star) - \frac{1}{2} (\hat{\theta}_t - \theta^\star)^\top \left(\ell''(y_t, \hat{\theta}_t^\top X_t) X_t X_t^\top \right) (\hat{\theta}_t - \theta^\star) \right)$$
$$\leq \frac{1}{2} \sum_{t=1}^{n} X_t^\top P_{t+1} X_t \ell'(y_t, \hat{\theta}_t^\top X_t)^2 + \frac{\|\hat{\theta}_1 - \theta^\star\|^2}{\lambda_{\min}(P_1)} \,.$$

We cannot compare the excess loss with the second-order Taylor expansion in general, and it is natural to use a step size parameter. In Hazan, Agarwal, and Kale, 2007, the regret analysis of the ONS is based on a very similar bound on

$$\left(\ell'(y_t, w_t^\top X_t) X_t \right)^\top (w_t - \theta^\star) - \frac{\gamma}{2} (w_t - \theta^\star)^\top \left(\ell'(y_t, w_t^\top X_t)^2 X_t X_t^\top \right) (w_t - \theta^\star) \,,$$

where γ is a step size parameter. Then the regret bound follows from the exp-concavity property, bounding the excess loss $\ell(y_t, w_t^\top X_t) - \ell(y_t, \theta^{\star\top} X_t)$ with the previous quantity for a specific γ. The dependence of γ on the exp-concavity constant and the bound on the gradients require that the algorithm stays in a bounded region around the optimum $\theta^\star$, and a projection on this region is used, potentially at each step.

We follow a very different approach, to stay parameter-free, unconstrained and to avoid any additional cost in the leading constant. In the stochastic setting, we observe that we can upper-bound the excess risk with a second-order expansion, up to a multiplicative factor.

4.3.2 From Adversarial to Stochastic: the Cumulative Risk

In order to compare the excess risk with a second-order expansion, we compare the first-order term with the second-order one.

Proposition 4.1. *If Assumptions 4.1, 4.2 and 4.3 are satisfied, for any $\theta \in \mathbb{R}^d$, it holds*

$$\frac{\partial L}{\partial \theta}\Big|_\theta^\top (\theta - \theta^\star) \geq \rho_{\|\theta-\theta^\star\|}(\theta - \theta^\star)^\top \frac{\partial^2 L}{\partial \theta^2}\Big|_\theta (\theta - \theta^\star)\,.$$

This result leads immediately to the following proposition, using the first-order convexity property of L.

Proposition 4.2. *If Assumptions 4.1, 4.2 and 4.3 are satisfied, for any $\theta \in \mathbb{R}^d$, $0 < c < \rho_{\|\theta-\theta^\star\|}$, it holds*

$$L(\theta) - L(\theta^\star) \leq \frac{\rho_{\|\theta-\theta^\star\|}}{\rho_{\|\theta-\theta^\star\|} - c}\left(\frac{\partial L}{\partial \theta}\Big|_\theta^\top (\theta - \theta^\star) - c(\theta - \theta^\star)^\top \frac{\partial^2 L}{\partial \theta^2}\Big|_\theta (\theta - \theta^\star)\right).$$

Lemma 4.2 motivates the use of $c > \frac{1}{2}$, thus we need at least $\rho_{\|\theta-\theta^\star\|} > \frac{1}{2}$. In the quadratic setting, it holds as an equality with $\rho = 1$ because the second derivative of the quadratic loss is constant. In the bounded setting we need to control the second derivative in a small range, and we can achieve that only locally, therefore we separate the condition $\rho_{\|\theta-\theta^\star\|} > \frac{1}{2}$ between the third point of Assumption 4.3 and Assumption 4.5.

Then we are left to obtain a bound on the cumulative risk from Lemma 4.2. In order to compare the derivatives of the risk and the losses, we need to control the martingale difference adapted to the natural filtration $(\mathcal{F}_t)$ and defined as

$$\Delta M_t = \left(\frac{\partial L}{\partial \theta}\Big|_{\hat\theta_t} - \nabla_t\right)^\top (\hat\theta_t - \theta^\star), \qquad \text{where } \nabla_t = \ell'(y_t, \hat\theta_t^\top X_t) X_t\,.$$

We thus apply Lemma 4.1 to this martingale difference.

Lemma 4.3. *Starting the static EKF from any $\hat\theta_1 \in \mathbb{R}^d, P_1 \succ 0$, if Assumptions 4.1 and 4.2 are satisfied, for any $k \geq 0$ and $\delta, \lambda > 0$, it holds simultaneously*

$$\sum_{t=k+1}^{k+n} \left(\Delta M_t - \lambda(\hat\theta_t - \theta^\star)^\top \left(\nabla_t \nabla_t^\top + \frac{3}{2}\mathbb{E}\big[\nabla_t \nabla_t^\top \mid \mathcal{F}_{t-1}\big]\right)(\hat\theta_t - \theta^\star)\right) \leq \frac{\ln \delta^{-1}}{\lambda}, \qquad n \geq 1\,,$$

with probability at least $1 - \delta$.

The proof of Theorems 4.1 and 4.2 deferred to Appendix B.2 builds on the above results. Summing Lemma 4.2 and 4.3, we obtain for any $\delta, \lambda > 0$ the simultaneous bound

$$\begin{aligned}\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} &\left(\frac{\partial L}{\partial \theta}\Big|_{\hat\theta_t}^\top (\hat\theta_t - \theta^\star) - (\hat\theta_t - \theta^\star)^\top \left(\frac{1}{2}\nabla_t^{(2)} + \lambda \nabla_t \nabla_t^\top + \frac{3}{2}\lambda \mathbb{E}\big[\nabla_t \nabla_t^\top \mid \mathcal{F}_{t-1}\big]\right)(\hat\theta_t - \theta^\star)\right) \\ &\leq \frac{1}{2}\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} X_t^\top P_{t+1} X_t \ell'(y_t, \hat\theta_t^\top X_t)^2 + \frac{\|\hat\theta_1 - \theta^\star\|^2}{\lambda_{\min}(P_{T(\varepsilon,\delta)+1})} + \frac{\ln \delta^{-1}}{\lambda}, \qquad n \geq 1\,,\end{aligned}$$

with probability at least $1 - \delta$, where we define $\nabla_t^{(2)} = \ell''(y_t, \hat\theta_t^\top X_t) X_t X_t^\top$ for any t. In the last equation, we control (see Appendix B.2.4 and B.2.5) the quadratic term in $\hat\theta_t - \theta^\star$ on the left hand-side in terms of $(\hat\theta_t - \theta^\star)^\top \frac{\partial^2 L}{\partial \theta^2}|_{\hat\theta_t}(\hat\theta_t - \theta^\star)$ in order to lower-bound the left expression proportionally to the cumulative excess risk using Proposition 4.2 for well chosen λ.

Algorithm 3 : Truncated Extended Kalman Filter for Logistic Regression

1. *Initialization*: P_1 is any positive definite matrix, $\hat{\theta}_1$ is any initial parameter in $\mathbb{R}^d$.
2. *Iteration*: at each time step $t = 1, 2, \dots$
   (a) Update $P_{t+1} = P_t - \frac{P_t X_t X_t^\top P_t}{1+X_t^\top P_t X_t \alpha_t}\alpha_t$, with $\alpha_t = \max\left(\frac{1}{t^\beta}, \frac{1}{(1+e^{\hat{\theta}_t^\top X_t})(1+e^{-\hat{\theta}_t^\top X_t})}\right)$.
   (b) Update $\hat{\theta}_{t+1} = \hat{\theta}_t + P_{t+1}\frac{y_t X_t}{1+e^{y_t \hat{\theta}_t^\top X_t}}$.

4.4 Logistic Setting

Logistic regression is a widely used statistical model in classification. The prediction of a binary random variable $y \in \mathcal{Y} = \{-1, 1\}$ consists in modelling $\mathcal{L}(y \mid X)$ with

$$p_\theta(y \mid X) = \frac{1}{1 + e^{-y\theta^\top X}} = \exp\left(\frac{y\theta^\top X - (2\ln(1 + e^{\theta^\top X}) - \theta^\top X)}{2}\right).$$

In the GLM notations, it yields $a = 2$ and $b(\theta^\top X) = 2\ln(1 + e^{\theta^\top X}) - \theta^\top X$.

4.4.1 Results for the Truncated Algorithm

In order to prove the convergence of the algorithm needed in the local phase, we follow a trick introduced by Bercu, Godichon, and Portier, 2020 consisting in changing slightly the update on P_t. Indeed, when the authors tried to prove the asymptotic convergence of the static EKF (which they named stochastic Newton step) using Robbins-Siegmund Theorem, they needed the convergence of $\sum_t \lambda_{\max}(P_t)^2$. This seems very likely to hold as we have intuitively $P_t \propto 1/t$. However, in order to obtain $\lambda_{\max}(P_t) = O(1/t)$, one needs to lower-bound α_t, that is, to upper-bound $|\hat{\theta}_t^\top X_t|$, and that is impossible in the global logistic setting. Therefore, the idea is to force a lower-bound on α_t in its definition. We thus define, for some $0 < \beta < 1/2$,

$$\alpha_t = \max\left(\frac{1}{t^\beta}, \frac{1}{(1 + e^{\hat{\theta}_t^\top X_t})(1 + e^{-\hat{\theta}_t^\top X_t})}\right), \qquad t \geq 1\,.$$

This modification yields Algorithm 3, where we keep the notations $\hat{\theta}_t, P_t, \tau(\varepsilon)$ with some abuse in the rest of this section. We impose a decreasing threshold on α_t ($\beta > 0$) and we prove that the recursion coincides with Algorithm 1 after some steps. The sensitivity of the algorithm to β is discussed at the end of Section 4.4.2. Also, note that the threshold could be c/t^β, $c > 0$, as in Bercu, Godichon, and Portier, 2020. We consider $1/t^\beta$ for clarity. We control the convergence time $\tau(\varepsilon)$ of this version of the EKF:

Proposition 4.3. *Starting Algorithm 3 from $\hat{\theta}_1 = 0$ and any $P_1 \succ 0$, if Assumptions 4.1 and 4.2 are satisfied and $\mathbb{E}[XX^\top]$ is invertible, for any $\varepsilon, \delta > 0$, it holds $\tau(\varepsilon) \leq T(\varepsilon, \delta)$ along with*

$$\forall t > T(\varepsilon, \delta), \qquad \alpha_t = \frac{1}{(1 + e^{\hat{\theta}_t^\top X_t})(1 + e^{-\hat{\theta}_t^\top X_t})}\,,$$

with probability at least $1 - \delta$, where $T(\varepsilon, \delta) \in \mathbb{N}$ is defined in Corollary 4.2.

Besides the convergence of the truncated EKF, the proposition states that the truncated recursions coincide with the static EKF ones after the first $T(\varepsilon, \delta)$ steps. Thus we can apply our analysis of Section 4.3. We state the global result for $\varepsilon = 1/(20D_X)$:

Theorem 4.3. *Under the assumptions of Proposition 4.3, for any $\delta > 0$, it holds for any $n \geq 1$ simultaneously*

$$\sum_{t=1}^{n} L(\hat{\theta}_t) - L(\theta^\star) \leq 3de^{D_X\|\theta^\star\|} \ln\left(1 + n\frac{\lambda_{\max}(P_1)D_X^2}{4d}\right) + \frac{\lambda_{\max}(P_1^{-1})}{75D_X^2} + 64e^{D_X\|\theta^\star\|}\ln\delta^{-1} + T\Big(\frac{1}{20D_X}, \delta\Big)\Big(\frac{1}{300} + D_X\|\hat{\theta}_1 - \theta^\star\|\Big) + T\Big(\frac{1}{20D_X}, \delta\Big)^2 \frac{\lambda_{\max}(P_1)D_X^2}{2},$$

with probability at least $1 - 4\delta$, where $T(1/(20D_X), \delta)$ is defined in Corollary 4.2.

4.4.2 Explicit Definition of $T(\varepsilon, \delta)$ in Proposition 4.3

It is proved that $\|\hat{\theta}_n - \theta^\star\|^2 = O(\ln n/n)$ almost surely (Bercu, Godichon, and Portier, 2020, Theorem 4.2). We don't obtain a non-asymptotic version of this rate of convergence, but the aim of this paragraph is to prove Proposition 4.3 for an explicit value of $T(\varepsilon, \delta)$ for any $\delta, \varepsilon > 0$.

The objective of the truncation introduced in the algorithm is to improve the control on P_t. We state that fact formally with a concentration result relying on Tropp, 2012. We define $\Lambda_{\min}$ the smallest eigenvalue of $\mathbb{E}[XX^\top]$.

Proposition 4.4. *Under the assumptions of Proposition 4.3, for any $\delta > 0$, it holds simultaneously*

$$\forall t > \left(\frac{20D_X^4}{\Lambda_{\min}^2}\ln\left(\frac{625dD_X^8}{\Lambda_{\min}^4\delta}\right)\right)^{1/(1-\beta)}, \qquad \lambda_{\max}(P_t) \leq \frac{4}{\Lambda_{\min}t^{1-\beta}},$$

with probability at least $1 - \delta$.

This proposition justifies the choice $\beta < 1/2$ in the introduction of the truncated algorithm to satisfy the condition $\sum_t \lambda_{\max}(P_t)^2 < +\infty$ with high probability. Motivated by Proposition 4.4, we define, for $C > 0$, the event

$$A_C := \bigcap_{t=1}^{\infty}\left(\lambda_{\max}(P_t) \leq \frac{C}{t^{1-\beta}}\right).$$

To obtain a control on P_t holding for any t, we use the relation $\lambda_{\max}(P_t) \leq \lambda_{\max}(P_1)$ holding almost surely. We thus define

$$C_\delta = \max\left(\frac{4}{\Lambda_{\min}}, \lambda_{\max}(P_1)\left(\frac{20D_X^4}{\Lambda_{\min}^2}\ln\Big(\frac{625dD_X^8}{\Lambda_{\min}^4\delta}\Big)\right)\right),$$

and we obtain $\mathbb{P}(A_{C_\delta}) \geq 1 - \delta$. We obtain the following theorem under that condition.

Theorem 4.4. *Under the assumptions of Proposition 4.3, we have for any $\delta, \varepsilon > 0$ and $t \geq$*

$\exp\left(\frac{2^8 D_X^8 C_\delta^2 (1+e^{D_X(\|\theta^\star\|+\varepsilon)})^3}{\Lambda_{\min}^3 (1-2\beta)^{3/2}\varepsilon^2}\right)$,

$$\mathbb{P}(\|\hat{\theta}_t - \theta^\star\| > \varepsilon \mid A_{C_\delta}) \leq (\sqrt{t}+1)\exp\left(-\frac{\Lambda_{\min}^6(1-2\beta)\varepsilon^4}{2^{16}D_X^{12}C_\delta^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^6}\ln(t)^2\right) + t\exp\left(-\frac{\Lambda_{\min}^2(1-2\beta)\varepsilon^4}{2^{11}D_X^4 C_\delta^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^2}(\sqrt{t}-1)^{1-2\beta}\right).$$

The beginning of our convergence proof starts similarly as the analysis of Bercu, Godichon, and Portier, 2020: we obtain a recursive inequality ensuring that $(L(\hat{\theta}_t))_t$ is decreasing in expectation. However, in order to obtain a non-asymptotic result we cannot apply Robbins-Siegmund Theorem. Instead we use the fact that the variations of the algorithm $\hat{\theta}_t$ are slow thanks to the control on P_t. Thus, if the algorithm was far from the optimum, the last estimates were far too which contradicts the decrease in expectation of the risk. Consequently, we look at the last $k \leq t$ such that $\|\hat{\theta}_k - \theta^\star\| < \varepsilon/2$, if it exists. We decompose the probability of being outside the local region in two scenarii, yielding the two terms in Theorem 4.4. If $k < \sqrt{t}$, the recursive decrease in expectation makes it unlikely that the estimate stays far from the optimum for a long period. If $k > \sqrt{t}$, the control on P_t allows a control on the probability that the algorithm moves fast, in $t-k$ steps, away from the optimum.

The following corollary explicitly defines a guarantee for the convergence time.

Corollary 4.2. *Proposition 4.3 holds with for any $\varepsilon, \delta > 0$*

$$T(\varepsilon,\delta) = \max\left(\left(2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})\right)^{1/\beta}, \exp\left(\frac{3\cdot 2^{15}D_X^{12}C_{\delta/2}^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^6}{\Lambda_{\min}^6(1-2\beta)^{3/2}\varepsilon^4}\right), 6\delta^{-1}\right).$$

This definition of $T(\varepsilon,\delta)$ allows a discussion on the dependence of the bound Theorem 4.3 on the different parameters. Note that the choice $\varepsilon = 1/(20D_X)$ in Theorem 4.3 is artificially made for simplifying constants since the bound actually holds for any $\varepsilon > 0$ simultaneously. The truncation has introduced an extra parameter $0 < \beta < 1/2$ that does not impact the leading term in Theorem 4.3. However, it impacts the first step control in an intricate way. On the one hand, when β is close to 0, the algorithm is slow to coincide with the true EKF as $T(\varepsilon,\delta) = e^{O(1)/\beta}$. On the other hand, the larger β, the larger our control on $\lambda_{\max}(P_t)$ and thus we get $T(\varepsilon,\delta) = e^{O(1)/(1-2\beta)^{3/2}}$. Practical considerations show that the truncation is artificial and can even deteriorate the performence of the EKF, see Section 4.6. Thus Bercu, Godichon, and Portier, 2020 suggest to choose $\beta = 0.49$.

The dependence on δ is even more complex. The third constraint on $T(\varepsilon,\delta)$ is $O(\delta^{-1})$ which should not be sharp. To improve this lousy dependence in the bound, one needs a better control of P_t. It would follow from a specific analysis of the $O(\ln \delta^{-1})$ first recursions in order to "initialize" the control on P_t. However the objective of Corollary 4.2 was to prove Proposition 4.3 and not to get an optimal value of $T(\varepsilon,\delta)$. A refinement of our convergence analysis following from a tighter control on P_t of the EKF than the one provided by Tropp, 2012 is a very important and challenging open question.

4.5 Quadratic Setting

We obtain a global result for the quadratic loss where Algorithm 1 becomes the standard Kalman filter (recall that we take $\sigma^2 = 1$, that is $\ell(y,\hat{y}) = (y-\hat{y})^2/2$ and $a = 1, b'(\hat{\theta}_t^\top X_t) = \hat{\theta}_t^\top X_t, \alpha_t = 1$).

The parallel with the ridge forecaster was evoked by Diderrich, 1985, and it is crucial that the static Kalman filter is the ridge regression estimator for a decaying regularization parameter. It highlights that the static EKF may be seen as an approximation of the regularized ERM:

Proposition 4.5. *In the quadratic setting, for any sequence (X_t, y_t), starting from any $\hat{\theta}_1 \in \mathbb{R}^d$ and $P_1 \succ 0$, the static EKF satisfies the optimisation problem*

$$\hat{\theta}_t = \arg\min_{\theta\in\mathbb{R}^d} \left(\frac{1}{2}\sum_{s=1}^{t-1}(y_s - \theta^\top X_s)^2 + \frac{1}{2}(\theta - \hat{\theta}_1)^\top P_1^{-1}(\theta - \hat{\theta}_1) \right), \qquad t \geq 1\,.$$

Note that the static Kalman filter provides automatically a right choice of the ridge regularization parameter. This equivalence yields a logarithmic regret bound for the Kalman filter (Theorem 11.7 of Cesa-Bianchi and Lugosi, 2006). It follows from Lemma 4.2 as the quadratic loss coincides with its second-order Taylor expansion. The leading term of the bound is $d \ln n \max_t (y_t - \hat{\theta}_t^\top X_t)^2$, thus $y_t - \hat{\theta}_t^\top X_t$ needs to be bounded.

As the static Kalman filter estimator is exactly the ridge forecaster, we can also use the regularized empirical risk minimization properties to control $T(\varepsilon, \delta)$. In particular, we apply the ridge analysis of Hsu, Kakade, and Zhang, 2012, and we check Assumption 4.5:

Proposition 4.6. *Starting from any $\hat{\theta}_1 \in \mathbb{R}^d$ and $P_1 \succ 0$, if Assumptions 4.1, 4.2 and 4.4 hold and if $\mathbb{E}[XX^\top]$ is invertible then Assumption 4.5 holds for $T(\varepsilon, \delta)$ defined explicitly in Appendix B.4, Corollary B.1.*

Up to universal constants, defining $\Lambda_{\min}$ as the smallest eigenvalue of $\mathbb{E}[XX^\top]$, we get

$$T(\varepsilon,\delta) \lesssim h\Bigg(\frac{\varepsilon^{-1}}{\Lambda_{\min}} \Bigg(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{p_1} + \frac{D_X^2}{\Lambda_{\min}}(1 + D_{\text{app}}^2)\sqrt{\ln \delta^{-1}} + \sigma^2 d \\ + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\text{app}} + D_X\|\theta^\star\|) + \frac{\|\hat{\theta}_1 - \theta^\star\|}{\sqrt{p_1}} + \sigma^2 \right) \ln \delta^{-1} \Bigg) \Bigg)\,,$$

with $h(x) = x \ln x$. We obtain a much less dramatic dependence on ε than in the logistic setting. However we could not avoid a $\Lambda_{\min}^{-1}$ factor in the definition of $T(\varepsilon, \delta)$. It is not surprising since the convergence phase relies deeply on the behavior of P_t.

As for the logistic setting, we split the cumulative risk into two sums. The sum of the first terms is roughly bounded by a worst case analysis, and the sum of the last terms is estimated thanks to our local analysis (Theorem 4.2). However, as the loss and its gradient are not bounded we cannot obtain a similar almost sure upper-bound on the convergence phase. The sub-gaussian assumption provides a high probability bound instead.

Theorem 4.5. *Under the assumptions of Proposition 4.6, for any $\varepsilon, \delta > 0$, it holds simultaneously*

$$\sum_{t=1}^{n} L(\hat{\theta}_t) - L(\theta^\star) \leq \frac{15}{2} d \left(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2\right) \ln\left(1 + n\frac{\lambda_{\max}(P_1) D_X^2}{d}\right) + 5\lambda_{\max}(P_1^{-1})\varepsilon^2 \\ + 115\left(\sigma^2(4 + \frac{\lambda_{\max}(P_1)D_X^2}{4}) + D_{\text{app}}^2 + 2\varepsilon^2 D_X^2\right)\ln\delta^{-1} \\ + D_X^2\left(5\varepsilon^2 + 2(\|\hat{\theta}_1 - \theta^\star\|^2 + 3\lambda_{\max}(P_1) D_X \sigma \ln\delta^{-1})^2\right) T(\varepsilon,\delta) \\ + \frac{2\lambda_{\max}(P_1)^2 D_X^4 (3\sigma + D_{\text{app}})^2}{3} T(\varepsilon,\delta)^3, \qquad n \geq 1\,,$$

with probability at least $1 - 6\delta$.

Note that the dependence of the cumulative excess risk of the convergence phase on δ is $O(\log(\delta^{-1})^3)$.

4.6 Experiments

We experiment the static EKF for logistic regression. Precisely, we compare the following sequential algorithms that we all initialize at 0:

— The static EKF and the truncated version (Algorithm 3). We take the default value $P_1 = I_d$ along with the value $\beta = 0.49$ suggested by Bercu, Godichon, and Portier, 2020. Note that a threshold $10^{-10}/t^{0.49}$ as recommended by Bercu, Godichon, and Portier, 2020 would coincide with the static EKF.

— The ONS and the averaged version. The convex region of search is a ball centered in 0 and of radius $D_\theta = 1.1\|\theta^\star\|$, a setting where we have good knowledge of $\theta^\star$. We consider two choices of the exp-concavity constant on which the ONS crucially relies to define the gradient step size. First, we use the only available bound $e^{-D_\theta D_X}$. Second, in the settings where the step size is so small that the ONS doesn't move, we use the exp-concavity constant κ_0 at $\theta^\star$. This yields a bigger step size, though the exp-concavity is not satisfied on the region of search.

— Two Averaged Stochastic Gradient Descent as described by Bach, 2014. First we test the choice of the gradient step size $\gamma = 1/(2D_X^2\sqrt{N})$ denoted by ASGD and a second version with $\gamma = \|\theta^\star\|/(D_X\sqrt{N})$ denoted by ASGD oracle. Note that these algorithms are with fixed horizon, thus at each step t, we have to re-run the whole procedure.

4.6.1 Synthetic Data

We first consider well-specified data generated by the process of Bercu, Godichon, and Portier, 2020. The explanatory variables $X = (1, Z^\top)^\top$ are of dimension $d = 11$ where Z is a random vector composed of 10 independent components uniformly generated in $[0, 1]$, thus $D_X = \sqrt{d}$. With this distribution for X we define three synthetic settings that we evaluate:

— **Well-specified 1**: we define $\theta^\star = (-9, 0, 3, -9, 4, -9, 15, 0, -7, 1, 0)^\top$, and at each iteration t, the variable $y_t \in \{-1, 1\}$ is a Bernoulli variable of parameter $(1 + e^{-\theta^{\star\top} X_t})^{-1}$.

— **Well-specified 2**: in the first well-specified setting the Bernoulli parameter is mostly distributed around 0 and 1 (see Figure 4.1), thus we try a less discriminated setting with $\theta^\star = \frac{1}{10}(-9, 0, 3, -9, 4, -9, 15, 0, -7, 1, 0)^\top$.

— **Misspecified**: In order to demonstrate the robustness of the EKF we test the algorithms in a misspecified setting switching randomly between two well-specified logistic processes. We define $\theta_1 = \frac{1}{10}(-9, 0, 3, -9, 4, -9, 15, 0, -7, 1, 0)^\top$ and θ_2 where we have only changed the first coefficient from $-9/10$ to $15/10$. Then y_t is a Bernoulli random variable whose parameter is either $(1 + e^{-\theta_1^\top X_t})^{-1}$ or $(1 + e^{-\theta_2^\top X_t})^{-1}$ uniformly at random. We checked that Assumption 4.2 is still satisfied.

We evaluate the different algorithms with the mean squared error $\mathbb{E}[\|\hat{\theta}_t - \theta^\star\|^2]$ that we approximate by its empirical version on 100 samples. We display the results in Figure 4.2.

4.6.2 Real Data Sets

To illustrate better the robustness to misspecification, we run the same procedures on real data sets:

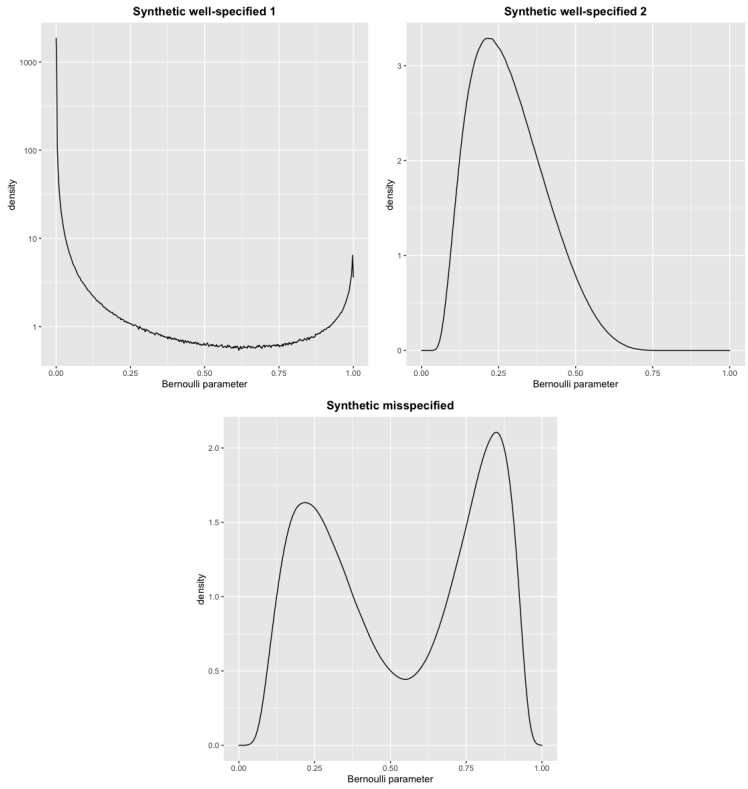


Figure 4.1 – Density of the Bernoulli parameter on 10^7 samples: on the left and on the middle density of $(1+e^{-\theta^{\star\top}X})^{-1}$ for the two well-specified settings (left, the ordinate is in log scale), and on the right density of $(1+e^{-\theta_j^\top X_t})^{-1}$ with $j \in \{1,2\}$ uniformly at random for the misspecified setting. On the right we observe the two modes $\mathbb{E}[(1+e^{-\theta_1^\top X_t})^{-1}] \approx 0.28$ and $\mathbb{E}[(1+e^{-\theta_2^\top X_t})^{-1}] \approx 0.79$.

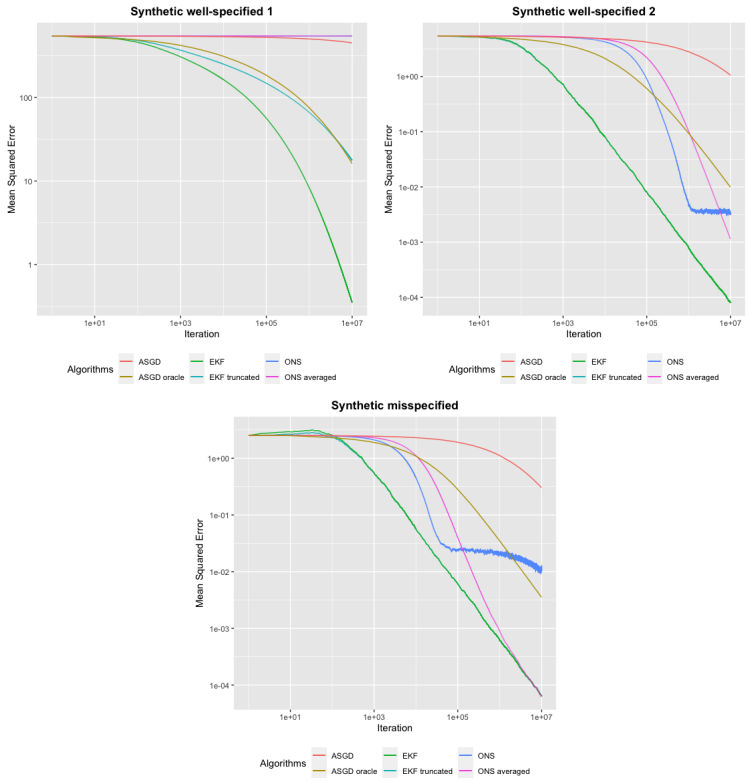


Figure 4.2 – Mean squared error in log-log scale for the three synthetic settings. For the first well-specified setting (left) the ONS is applied using the exp-concavity constant $\kappa_0 \approx 1.7 \cdot 10^{-15}$ instead of $e^{-D_\theta \sqrt{d}}$ to accelerate the algorithm, and both the ONS and its averaged version still don't move. In the other two (middle and right) we use $e^{-D_\theta \sqrt{d}}$ for the ONS. We observe that the EKF and the truncated version coincide in the two last settings.

Setting	d	$\lambda_{\max}(H^\star)/\mu$	$\mathrm{tr}(G^\star H^{\star -1})$	R^2/μ	$de^{D_\theta D_X}$	$d\kappa_0$
Synthetic well-specified 1	11	$6.9 \cdot 10^2$	$1.7 \cdot 10^2$	$1.0 \cdot 10^5$	$9.2 \cdot 10^{37}$	$6.4 \cdot 10^{15}$
Synthetic well-specified 2	11	$1.5 \cdot 10^2$	$7.1 \cdot 10^1$	$2.5 \cdot 10^3$	$5.4 \cdot 10^4$	$3.3 \cdot 10^2$
Synthetic misspecified 3	11	$1.5 \cdot 10^2$	$7.1 \cdot 10^1$	$2.0 \cdot 10^3$	$3.6 \cdot 10^3$	$7.4 \cdot 10^1$
Forest cover type	54	∞	∞	∞	$9.2 \cdot 10^{32}$	$3.8 \cdot 10^4$
Adult income	98	$2.5 \cdot 10^7$	$7.2 \cdot 10^5$	$5.3 \cdot 10^8$	$1.8 \cdot 10^{62}$	$1.8 \cdot 10^7$

Table 4.2 – For the different experimental settings we display the dimension d and the condition number of the Hessian at $\theta^\star$ ($\lambda_{\max}(H^\star)$ and μ are the maximal and minimal eigenvalues of $H^\star$). We present the value of $\mathrm{tr}(G^\star H^{\star -1})$ which is bounded either by R^2/μ, or by $de^{D_\theta D_X}$ because $e^{-D_\theta D_X}$ bounds the exp-concavity constant on the centered ball of radius D_θ. We add to the table $d\kappa_0 \leq de^{D_\theta D_X}$ where κ_0 is the inverse of the exp-concavity constant of the loss at $\theta^\star$.

— **Forest cover-type** (Blackard and Dean, 1999): the feature vector is of dimension $d = 54$, and as it is a multi-class task (7 classes) we focus on classifying 2 versus all others. There are $n = 581012$ instances and we randomly split in two halves for training and testing.
— **Adult income** (Kohavi, 1996): the objective is to predict whether a person's annual income is smaller or bigger than 50K. There are 14 explanatory variables, and we obtain $d = 98$ once categorical variables are transformed into binary variables. We use the canonical split between training (32561 instances) and testing (16281 instances).

For each data set, we standardize X such that each feature ranges from 0 to 1. At each step we sample within the training set (with replacement). We evaluate through an empirical version of $\mathbb{E}[L(\hat{\theta}_n)] - L(\theta^\star)$ estimated on 100 samples and where L is estimated on the test set, see Figure 4.3.

4.6.3 Summary

Our experiments show the superiority of the EKF for logistic regression compared to the ONS or to averaged SGD in all the settings we tested. We display in Table 4.2 a few indicators of the data sets. In particular, it is interesting that the static EKF works well even in a setting where the Hessian matrix $H^\star$ is singular.

It appears clear that low exp-concavity constants are responsible of the poor performances of the ONS. One may tune the gradient step size at the cost of losing the exp-concavity property and thus the regret guarantee of (Hazan, Agarwal, and Kale, 2007) or its analogous for the cumulative risk (Mahdavi, Zhang, and Jin, 2015). Averaging is crucial for the ONS, whereas it is useless for the static EKF. Indeed we chose not to plot the averaged version of the EKF for clarity, but the EKF performs better than its averaged version.

It is important to note that in the first synthetic setting the truncation deteriorates the performance of the EKF, as well as in the adult income data set to a lesser extent, whereas the results are the same in the other settings. Bercu, Godichon, and Portier, 2020 argue that the truncation is artificially introduced for the convergence property, thus they use the threshold $10^{-10}/t^{0.49}$ instead of $1/t^{0.49}$ and the truncated version almost coincides with the true EKF. We confirm here that the truncation may be damaging if the threshold is set too high and we recommend to use the EKF in practice, or equivalently the truncated version with the low threshold suggested by Bercu, Godichon, and Portier, 2020.

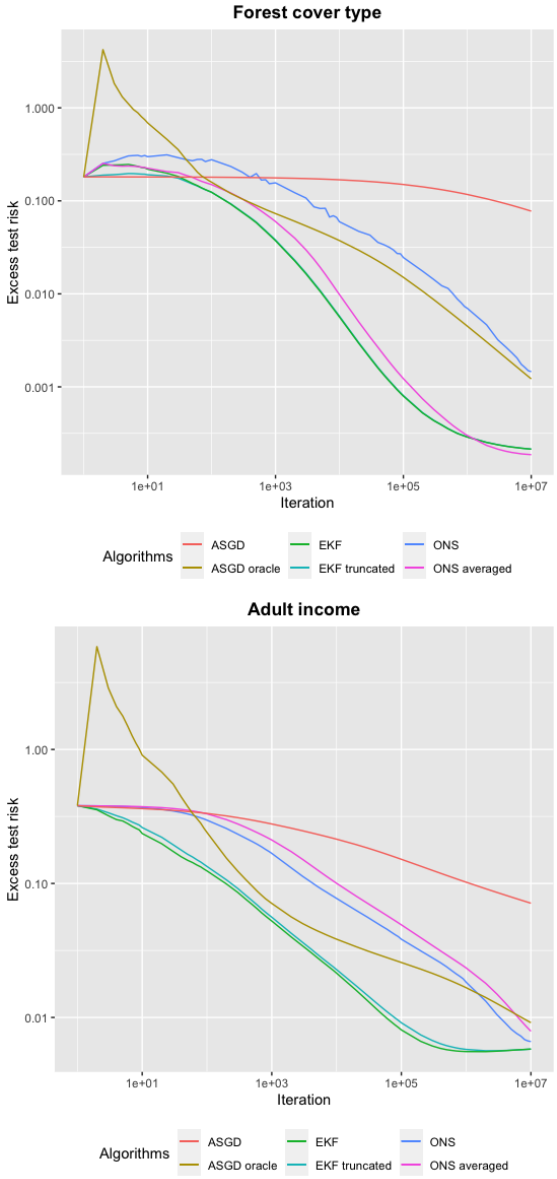


Figure 4.3 – Excess test risk for forest cover type (left) and adult income (right). As the ONS doesn't move when applied with the exp-concavity constant $e^{-D_\theta D_X}$ we use instead the exp-concavity constant at $\theta^\star$: $\kappa_0 \approx 1.4 \cdot 10^{-3}$ for forest cover type and $\kappa_0 \approx 5.5 \cdot 10^{-6}$ for adult income. The EKF and the truncated version almost coincide for both data sets.

4.7 Conclusion

We studied an efficient way to tackle some unconstrained optimization problems, in which we get rid of the projection step of bounded algorithm such as the ONS. We presented a bayesian approach where we transformed the loss into a negative log-likelihood. We used the Kalman recursion to provide a parameter free approximation of the maximum-likelihood estimator. We demonstrated the optimality of the local phase for locally exp-concave losses which can be expressed as GLM log-likelihoods. We proved the finiteness of the convergence phase in logistic and quadratic regressions. We illustrated our theoretical results with numerical experiments for logistic regression. It would be interesting to generalize our results to a larger class of optimization problems.

Finally, this article aimed at strengthening the bridge between Kalman recursion and the optimization community. Therefore we made the i.i.d. assumption, standard in the stochastic optimization literature and we focus on the static EKF. It may lead the way to a risk analysis of the general EKF in non i.i.d. state-space models.

Part II

The Choice of the Variances in a State-Space Model

Chapter 5

Constant Variances in a Kalman Filter with Delayed Observations

In this chapter, we discuss the choice of the hyper-parameters in a linear Gaussian state-space model. We compare two heuristics. First, we derive the widely used expectation-maximization algorithm. Second, building on the fact the log-likelihood is not necessarily a convex function, we propose a more empirical approach. We provide a toy example on a time series with delayed observations, and we claim that the proposed algorithm takes that practical delay into account, while that is not the case of the EM algorithm.

Contents

5.1 Introduction

We consider the linear Gaussian state-space model with time-invariant variances:

$$y_t = \theta_t^\top x_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2)\,, \tag{5.1}$$
$$\theta_{t+1} = \theta_t + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, Q)\,. \tag{5.2}$$

For any integers t and m we define

$$\hat{\theta}_{t|m} = \mathbb{E}\left[\theta_t \mid x_{1:m}, y_{1:m}\right], \qquad P_{t|m} = \mathbb{E}\left[(\theta_t - \hat{\theta}_{t|m})(\theta_t - \hat{\theta}_{t|m})^\top \mid x_{1:m}, y_{1:m}\right],$$

where $x_{1:m}$ denotes the set $(x_1, \dots x_m)$ and similarly for y. The context of this thesis is time series forecasting, where $x_t \in \mathbb{R}^d$ is the covariate vector and $y_t \in \mathbb{R}$ is the variable to predict. Therefore we are mostly interested in $\hat{\theta}_{t|t-1}, P_{t|t-1}$. These values are exactly obtained by the Kalman filter (Kalman and Bucy, 1961):

$$P_{t|t} = P_{t|t-1} - \frac{P_{t|t-1}x_t x_t^\top P_{t|t-1}}{x_t^\top P_{t|t-1} x_t + \sigma^2}, \qquad \hat{\theta}_{t|t} = \hat{\theta}_{t|t-1}, \tag{5.3}$$

$$P_{t+1|t} = P_{t|t} + Q, \qquad \hat{\theta}_{t+1|t} = \hat{\theta}_{t|t-1} + P_{t|t}x_t(y_t - \hat{\theta}_{t|t-1}^\top x_t). \tag{5.4}$$

Then, at each time step, we know that the distribution of y_t given $x_{1:(t-1)}, y_{1:(t-1)}, x_t$ is the Gaussian distribution of mean $\hat{\theta}_{t|t-1}^\top x_t$ and variance $x_t^\top P_{t|t-1} x_t + \sigma^2$.

The problem at hand is the choice of the *hyper-parameters* of the state-space model: the model variances σ^2, Q, along with the prior $\hat{\theta}_{1|0}, P_{1|0}$. We define the set of hyper-parameters $\Theta = (\sigma^2, Q, \hat{\theta}_{1|0}, P_{1|0})$.

It is classical to maximize the likelihood in order to choose hyper-parameters. In this chapter we present two heuristics. We first derive the expectation-maximization (EM) algorithm (see, for instance, Watson and Engle, 1983), where the likelihood is expressed as the expectation of the complete likelihood:

$$p(x_{1:n}, y_{1:n} \mid \Theta) = \int_{\theta_{1:n}} p(x_{1:n}, y_{1:n}, \theta_{1:n} \mid \Theta).$$

Then we present a new heuristics maximizing directly the likelihood with an iterative algorithm close to a grid search. It relies on the following expression of the log-likelihood:

$$\begin{aligned}\ln p(x_{1:n}, y_{1:n} \mid \Theta) &= \sum_{t=1}^{n} \ln p(y_t \mid x_{1:(t-1)}, y_{1:(t-1)}, x_t, \Theta) + c, \\ &= -\frac{1}{2}\sum_{t=1}^{n}\left(\ln(2\pi) + \ln(\sigma^2 + x_t^\top P_{t|t-1}x_t) + \frac{(y_t - \hat{\theta}_{t|t-1}^\top x_t)^2}{\sigma^2 + x_t^\top P_{t|t-1}x_t}\right) + c, \end{aligned} \tag{5.5}$$

where $c = \sum_{t=1}^{n} \ln p(x_t \mid x_{1:(t-1)}, y_{1:(t-1)}, \Theta)$ is assumed independent of Θ in the model. In what follows, we remove the constant c. For clarity, we don't write $\hat{\theta}_{t|t-1}, P_{t|t-1}$ as functions of Θ.

5.2 The Complete Likelihood and EM Algorithm

The most standard approach to optimize the log-likelihood is the EM algorithm. This algorithm is an iterative algorithm increasing the likelihood at each step.

5.2.1 Principle

Building on the complete likelihood, we observe that for any probability density function q, it holds:

$$\begin{aligned}
\ln p(x_{1:n}, y_{1:n} \mid \Theta) &= \ln \int_{\theta_{1:n}} p(x_{1:n}, y_{1:n}, \theta_{1:n} \mid \Theta) d\theta_{1:n} \\
&= \ln \int_{\theta_{1:n}} \frac{p(x_{1:n}, y_{1:n}, \theta_{1:n} \mid \Theta)}{q(\theta_{1:n})} q(\theta_{1:n}) d\theta_{1:n} \\
&\geq \int_{\theta_{1:n}} \ln \left(\frac{p(x_{1:n}, y_{1:n}, \theta_{1:n} \mid \Theta)}{q(\theta_{1:n})} \right) q(\theta_{1:n}) d\theta_{1:n} \\
&= \mathbb{E}_q \left[\ln p(x_{1:n}, y_{1:n}, \theta_{1:n} \mid \Theta) \right] + H(q) := \mathcal{L}(q, \Theta) \,,
\end{aligned}$$

where $H(q) = -\int q \ln q$ is the entropy of q. Third line is the application of Jensen inequality. This inequality is an equality if $q(\theta_{1:n}) = p(\theta_{1:n} \mid x_{1:n}, y_{1:n}, \Theta)$. Thus we get for any Θ,

$$\ln p(x_{1:n}, y_{1:n} \mid \Theta) = \mathcal{L}(p(. \mid x_{1:n}, y_{1:n}, \Theta), \Theta) = \max_q \mathcal{L}(q, \Theta) \,.$$

Furthermore, if $\Theta^\star$ is in $\arg\max_{\Theta'} \mathcal{L}(p(. \mid x_{1:n}, y_{1:n}, \Theta), \Theta')$ then

$$\ln p(x_{1:n}, y_{1:n} \mid \Theta^\star) \geq \mathcal{L}(p(. \mid x_{1:n}, y_{1:n}, \Theta), \Theta^\star) \geq \mathcal{L}(p(. \mid x_{1:n}, y_{1:n}, \Theta), \Theta) = \ln p(x_{1:n}, y_{1:n} \mid \Theta) \,.$$

We have thus an iterative way to increase the log-likelihood at each step, called the expectation-maximization algorithm. At each iteration k we have $\Theta^{(k)}$ in hand. The E-step consists in estimating $p(. \mid x_{1:n}, y_{1:n}, \Theta^{(k)})$. The M-step consists in maximizing $\mathcal{L}(p(. \mid x_{1:n}, y_{1:n}, \Theta^{(k)}), \Theta)$ in Θ. The guarantee of this algorithm (proved above) is that the likelihood is increased at each step. However, there is no global guarantee.

We consider a deterministic prior $\theta_1 = \hat{\theta}_{1|0}$, that is $P_{1|0} = 0$. This is not too strong because the algorithm provides $Q \succ 0$ and thus $(P_{t|t-1})$ converges to an ergodic stationary process (Bougerol, 1992), that is, the initial matrix $P_{1|0}$ is vanishing. Then we have

$$\begin{aligned}
\ln p(x_{1:n}, y_{1:n}, \theta_{1:n} \mid \hat{\theta}_{1|0}, Q, \sigma^2) = & -\frac{1}{2} \sum_{t=1}^{n-1} (\theta_{t+1} - \theta_t)^\top Q^{-1} (\theta_{t+1} - \theta_t) - \frac{n-1}{2} \ln \det Q \\
& - \frac{1}{2} \sum_{t=1}^{n} \frac{(y_t - \theta_t^\top x_t)^2}{\sigma^2} - \frac{n}{2} \ln(\sigma^2) - \frac{2n-1}{2} \ln 2\pi \,. \qquad (5.6)
\end{aligned}$$

As the previous expression is independent of $\hat{\theta}_{1|0}$, we incorporate in the E-step a maximization of the log-likelihood with respect to $\hat{\theta}_{1|0}$. The EM algorithm thus consists of the following two steps at each iteration:

— **E-step**: estimate $\hat{\theta}_{1|0}^{(k+1)}$ maximizing the log-likelihood with fixed $Q^{(k)}, \sigma^{2(k)}$, then compute $\hat{\theta}_{t|n}, P_{t|n}$ for given $(\hat{\theta}_{1|0}^{(k+1)}, Q^{(k)}, \sigma^{2(k)})$.

— **M-step**: maximize $\mathbb{E}\left[\ln p(x_{1:n}, y_{1:n}, \theta_{1:n} \mid \hat{\theta}_{1|0}^{(k+1)}, Q^{(k)}, \sigma^{2(k)}) \mid x_{1:n}, y_{1:n}, (\hat{\theta}_{t|n}, P_{t|n})_t \right]$ with respect to $Q^{(k)}$ and $\sigma^{2(k)}$.

The increase of the log-likelihood still holds even with this optimization in $\hat{\theta}_{1|0}$ in the E-step.

Algorithm 4 : Expectation-Maximization

— **Input**: $x_{1:n}, y_{1:n}, \hat{\theta}_1^{(0)}$ (default 0), $Q^{(0)}$ (default I), $\sigma^{2(0)}$ (default 1), N.
— **Iteration**: for k in $0 : (N-1)$:

E-step:

1. Kalman Filter. Estimate $\hat{\theta}^*_{t|t-1}, P_{t|t-1}$ as well as $C_{t|t-1}$ using Equations (5.3) and (5.4), as well as (5.7) starting from $\hat{\theta}^*_{1|0} = 0, P_{1|0} = 0$ and using the variances $Q^{(k)}, \sigma^{2(k)}$.
2. Set $\hat{\theta}_1^{(k+1)}$ using Equation (5.8). Then set $\hat{\theta}_{t|t-1} = \hat{\theta}^*_{t|t-1} + C_{t|t-1}\hat{\theta}_1^{(k+1)}$.
3. Kalman Smoothing. Compute $(\hat{\theta}_{t|n}, P_{t|n})_t$ with Equations (5.9) and (5.10).

M-step:

4. Compute $\sigma^{2(k+1)}$ using Equation (5.11).
   Compute $Q^{(k+1)}$ using Equation (5.12).

— **Outputs**: $\hat{\theta}_1^{(N)}, P_1 = 0, \sigma^{2(N)}, Q^{(N)}$.

Indeed, the update of $\hat{\theta}^{(k)}_{1|0}$ increases the log-likelihood, and so does the M-step as explained at the beginning of the section.

5.2.2 The EM Algorithm

We derive precisely the step-by-step procedure, and we summarize in Algorithm 4.

1. We first remark from Equations (5.3) and (5.4) that the estimate of the Kalman filter can be decomposed as $\hat{\theta}_{t|t-1} = \hat{\theta}^*_{t|t-1} + C_{t|t-1}\hat{\theta}_{1|0}$, where $\hat{\theta}^*_{t|t-1}$ is the estimate obtained with $\hat{\theta}^*_{1|0} = 0$, and

$$C_{t+1|t} = \left(I - P^*_{t|t} x_t x_t^\top\right) C_{t|t-1}, \qquad t \geq 1. \tag{5.7}$$

2. Therefore the expression of the log-likelihood given in Equation (5.5) may be rewritten as:

$$-\frac{1}{2}\sum_{t=1}^{n}\left(\ln(\sigma^2 + x_t^\top P_{t|t-1}x_t) + \frac{(y_t - (\hat{\theta}^*_{t|t-1} + C_{t|t-1}\hat{\theta}_{1|0})^\top x_t)^2}{\sigma^2 + x_t^\top P_{t|t-1}x_t}\right),$$

up to a constant. Therefore, we see that for fixed $Q^{(k)}, \sigma^{2(k)}$ and corresponding values of $(\hat{\theta}_{t|t-1}, P_{t|t-1}, C_{t|t-1})_t$, the maximum likelihood with respect to $\hat{\theta}_{1|0}$ is obtained for

$$\hat{\theta}^{(k+1)}_{1|0} = \left(\sum_{t=1}^{n}\frac{C^\top_{t|t-1}x_t x_t^\top C_{t|t-1}}{1 + x_t^\top P_{t|t-1}x_t}\right)^{-1}\sum_{t=1}^{n}\frac{(y_t - \hat{\theta}^{*\top}_{t|t-1}x_t)C^\top_{t|t-1}x_t}{1 + x_t^\top P_{t|t-1}x_t}. \tag{5.8}$$

3. We then need Kalman smoothing in addition to Kalman filtering: this is a downward

recursion:

$$\hat{\theta}_{t|n} = \hat{\theta}_{t|t} + P_{t|t}P_{t+1|t}^{-1}(\hat{\theta}_{t+1|n} - \hat{\theta}_{t+1|t})\,, \tag{5.9}$$

$$P_{t|n} = P_{t|t} + P_{t|t}P_{t+1|t}^{-1}(P_{t+1|n} - P_{t+1|t})P_{t+1|t}^{-1}P_{t|t}\,. \tag{5.10}$$

4. We obtain also a closed-form solution for the M-step. The expression of the expected complete log-likelihood (5.6) is easily maximized with respect to σ^2:

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[(y_t - \theta_t^\top x_t)^2 \mid x_{1:n}, y_{1:n}, \theta_t \sim \mathcal{N}(\hat{\theta}_{t|n}, P_{t|n})] \\ &= \frac{1}{n}\sum_{t=1}^{n}\Big((y_t - \hat{\theta}_{t|n}^\top x_t)^2 + x_t^\top P_{t|n} x_t\Big)\,. \end{aligned} \tag{5.11}$$

Similarly, the maximum with respect to Q is obtained with:

$$\begin{aligned}Q &= \frac{1}{n-1}\sum_{t=1}^{n-1}\mathbb{E}\left[(\theta_{t+1} - \theta_t)(\theta_{t+1} - \theta_t)^\top \mid \theta_t \sim \mathcal{N}(\hat{\theta}_{t|n}, P_{t|n})\right] \\ &= \frac{1}{n-1}\sum_{t=1}^{n-1}\Big(P_{t+1|n} + (\hat{\theta}_{t+1|n} - \hat{\theta}_{t|n})(\hat{\theta}_{t+1|n} - \hat{\theta}_{t|n})^\top + P_{t|n} \qquad (5.12)\\ &\qquad\qquad - P_{t+1|n}P_{t+1|t}^{-1}P_{t|t} - P_{t|t}P_{t+1|t}^{-1}P_{t+1|n}\Big)\,. \qquad (5.13)\end{aligned}$$

5.3 Likelihood Optimization by Grid Search

This section is much inspired by Section 8.5 of Brockwell and Davis, 2016. In Section 5.3.1 we follow the same ideas. However no method was proposed to optimize the reduced likelihood (Brockwell and Davis, 2016, Equation (8.5.12)), and we show in Section 5.3.2 that the obtained optimization problem is not convex. We provide a heuristic in Section 5.3.3.

5.3.1 Derivation of the log-likelihood

We start from Equation (5.5):

$$\ln p(x_{1:n}, y_{1:n} \mid \Theta) = -\frac{1}{2}\sum_{t=1}^{n}\left(\ln(2\pi) + \ln(\sigma^2 + x_t^\top P_{t|t-1}x_t) + \frac{(y_t - \hat{\theta}_{t|t-1}^\top x_t)^2}{\sigma^2 + x_t^\top P_{t|t-1}x_t}\right)\,.$$

We define $P^*_{t|t-1} = P_{t|t-1}/\sigma^2$ and $Q^* = Q/\sigma^2$. Indeed, a remarkable property is that $\hat{\theta}_{t|t-1}, P^*_{t|t-1}$ only depend on $\hat{\theta}_{1|0}, P^*_{1|0}, Q^*$. In terms of forecasting, this means that the best mean forecast fo y_t, which is $\hat{\theta}_{t|t-1}^\top x_t$, depends only on $\hat{\theta}_{1|0}, P^*_{1|0}, Q^*$. For probabilistic forecasting it is however necessary to estimate σ^2 and $P_{t|t-1}$, because the conditional variance of y_t is $x_t^\top P_{t|t-1}x_t + \sigma^2$. Therefore,

$$\ln p(x_{1:n}, y_{1:n} \mid \Theta) = -\frac{n}{2}\big(\ln(2\pi) + \ln(\sigma^2)\big) - \frac{1}{2}\sum_{t=1}^{n}\Big(\ln(1 + x_t^\top P^*_{t|t-1}x_t) + \frac{1}{\sigma^2}\frac{(y_t - \hat{\theta}_{t|t-1}^\top x_t)^2}{1 + x_t^\top P^*_{t|t-1}x_t}\Big)\,.$$

For fixed $\hat{\theta}_{1|0}, P^*_{1|0}, Q^*$ and corresponding $(\hat{\theta}_{t|t-1}, P^*_{t|t-1})$, the maximum likelihood with respect to σ^2 is obtained for

$$\sigma^2 = \frac{1}{n}\sum_{t=1}^{n} \frac{(y_t - \hat{\theta}_{t|t-1}^\top x_t)^2}{1 + x_t^\top P^*_{t|t-1} x_t}\,. \tag{5.14}$$

It allows reducing the dimension of the optimization problem. Removing constants, we minimize in $\hat{\theta}_{1|0}, P^*_{1|0}, Q^*$ the following quantity:

$$S(\hat{\theta}_{1|0}, P^*_{1|0}, Q^*) = \frac{1}{2}\sum_{t=1}^{n} \ln(1 + x_t^\top P^*_{t|t-1} x_t) + \frac{n}{2}\ln\left(\frac{1}{n}\sum_{t=1}^{n} \frac{(y_t - \hat{\theta}_{t|t-1}^\top x_t)^2}{1 + x_t^\top P^*_{t|t-1} x_t}\right).$$

As in the last section, we observe that if $\hat{\theta}^*_{t|t-1}$ is the Kalman estimate obtained with $\hat{\theta}^*_{1|0} = 0$, we have $\hat{\theta}_{t|t-1} = \hat{\theta}^*_{t|t-1} + C_{t|t-1}\hat{\theta}_1$ where $C_{t|t-1}$ is defined recursively (5.7). It yields:

$$S(\hat{\theta}_{1|0}, P^*_{1|0}, Q^*) = \frac{1}{2}\sum_{t=1}^{n} \ln(1 + x_t^\top P^*_{t|t-1} x_t) + \frac{n}{2}\ln\left(\frac{1}{n}\sum_{t=1}^{n} \frac{(y_t - \hat{\theta}^{*\top}_{t|t-1} x_t - \hat{\theta}_{1|0}^\top C_{t|t-1}^\top x_t)^2}{1 + x_t^\top P^*_{t|t-1} x_t}\right).$$

For given $P^*_{1|0}, Q^*$, the minimum is reached for

$$\hat{\theta}_{1|0} = \left(\sum_{t=1}^{n} \frac{C_{t|t-1}^\top x_t x_t^\top C_{t|t-1}}{1 + x_t^\top P^*_{t|t-1} x_t}\right)^{-1} \sum_{t=1}^{n} \frac{(y_t - \hat{\theta}^{*\top}_{t|t-1} x_t) C_{t|t-1}^\top x_t}{1 + x_t^\top P^*_{t|t-1} x_t}\,. \tag{5.15}$$

Eventually, the maximum likelihood is reduced to the minimization of

$$\ell(P^*_{1|0}, Q^*) = \frac{1}{2}\sum_{t=1}^{n} \ln(1 + x_t^\top P^*_{t|t-1} x_t) + \frac{n}{2}\ln\left(\frac{1}{n}\sum_{t=1}^{n} \frac{(y_t - \hat{\theta}^{*\top}_{t|t-1} x_t - \hat{\theta}_{1|0}^\top C_{t|t-1}^\top x_t)^2}{1 + x_t^\top P^*_{t|t-1} x_t}\right). \tag{5.16}$$

5.3.2 A Non-convex Log-likelihood

We relied on the ideas of Brockwell and Davis, 2016 to reduce the maximum likelihood problem to the minimization of $\ell(P^*_{1|0}, Q^*)$. However, we claim that optimizing this function is a complex problem because it is not convex. In particular, there can be multiple local optima.

We consider the smallest possible setting. We use 2-dimensional $x_t \sim \mathcal{N}(0, I_2)$, and then y_t is generated by the well-specified state-space model: $\theta_1 \sim \mathcal{N}(0, I_2)$ and

$$\begin{aligned} y_t - \theta_t^\top x_t &\sim \mathcal{N}(0, 1)\,, \\ \theta_{t+1} - \theta_t &\sim \mathcal{N}(0, 10^{-3} I_2)\,. \end{aligned}$$

For this data we plot the log-likelihood and its reduced version for varying Q^* in Figure 5.1. We observe that the log-likelihood has a nice shape for matrices proportional to I_2, and the maximum of the likelihood is reached for $q \approx 4 \cdot 10^{-4}$. However, on the segment between $diag(1, 0)$ and $diag(0, 1)$, the log-likelihood admits two local optima. It proves that the negative log-likelihood is not necessarily convex or even quasiconvex. It explains why maximum likelihood estimation has been a long-standing problem, and is likely to stay an interesting topic with no consensus but various heuristics. The second conclusion of Figure 5.1 is that *a priori* knowledge on Q may be very useful: maximizing the likelihood is simple when we know that Q is proportional to I_2.

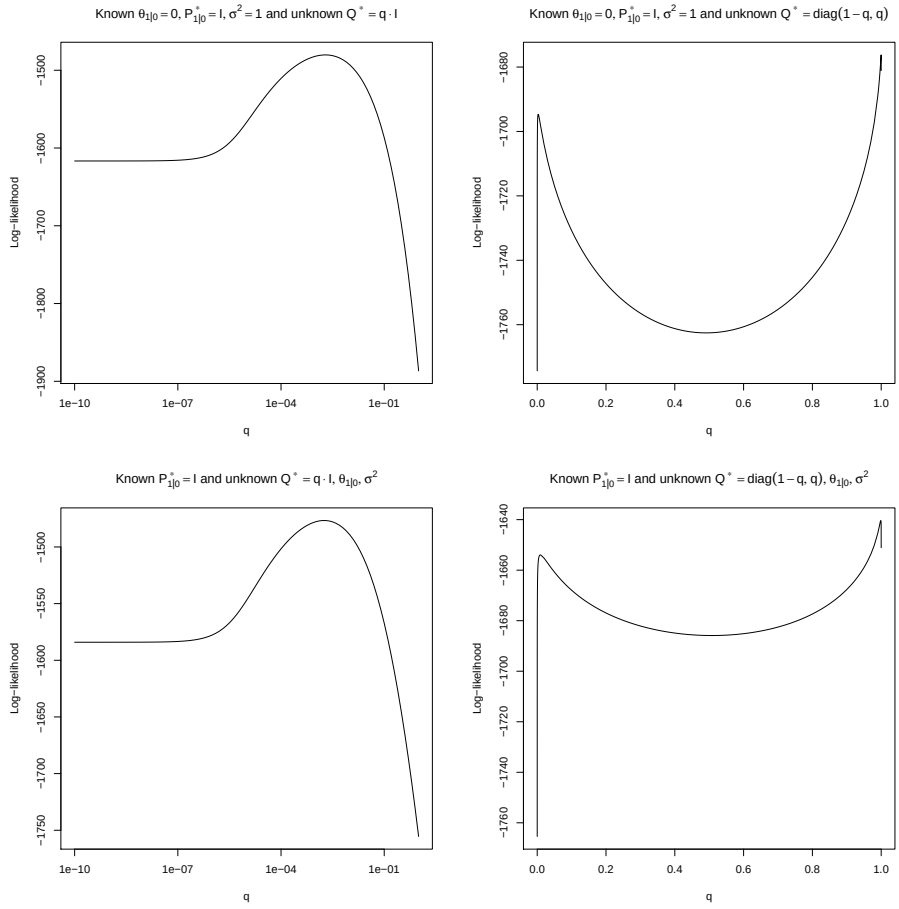


Figure 5.1 – We fix $P^*_{1|0} = I$ and we plot the log-likelihood for different Q^* with $n = 1000$. On the top the likelihood is the one of (5.5) for known $\hat{\theta}_{1|0}, P^*_{1|0}, \sigma^2$. On the bottom we display the reduced likelihood obtained for varying Q^* and the associated values of σ^2 and $\hat{\theta}_{1|0}$ obtained by Equations (5.14) and (5.15). We add constants to obtain comparable functions. On the left: $Q^* = qI_2$. On the right: $Q^* = diag(1-q, q)$.

Algorithm 5 : Iterative Grid Search

— **Input**: $x_{1:n}, y_{1:n}$, $P^*_{1|0} \succcurlyeq 0$ (default I_d), *qlist* (default $(0, 2^{-31}, 2^{-30}, \dots 1)$).
— **Initialization**: $Q^{*(0)} = 0, k = 0$.
— **Iteration**: while $Q^{*(k)} \neq Q^{*(k-1)}$:
  — For any $i \in \{1, \dots d\}$, for each q in *qlist*, define $Q^*_{i,q} = Q^{*(k)}$ and set the coefficient i of the diagonal of $Q^*_{i,q}$ to q, then:
    1. Compute $\hat{\theta}^*_{t|t-1}, P^*_{t|t-1}$ using Kalman filter starting from $\hat{\theta}_{1|0} = 0, P_{1|0} = P^*_{1|0}$ and with variances $\sigma^2 = 1, Q = Q^*_{i,q}$. Compute also C_t from Equation (5.7).
    2. Compute $\hat{\theta}_{1|0}$ from Equation (5.15).
    3. Compute $\ell(P^*_{1|0}, Q^*_{i,q})$ from Equation (5.16).
  — Define $Q^{*(k+1)}$ as the one maximizing $\ell(P^*_{1|0}, Q^*_{i,q})$ among the matrices $Q^*_{i,q}$. Increment $k \to k + 1$.
— **Final steps**: Set $Q^* = Q^{*(k)}$. As in each iteration step $\hat{\theta}_1$ is computed given $P^*_{1|0}, Q^*$. Then we also compute σ^2 from Equation (5.14).
— **Outputs**: $\hat{\theta}_{1|0}, P_1 = \sigma^2 P^*_{1|0}, \sigma^2, Q = \sigma^2 Q^*$.

In what follows, we derive a new algorithm to choose hyper-parameters. It consists in constraining the matrix Q to be diagonal, then selecting the best by an iterative grid search to minimize the expression of Equation (5.16). Similarly as for the EM algorithm, the guarantee is an increase of the likelihood at each step.

5.3.3 Grid Search

As the function $\ell(P^*_{1|0}, Q^*)$ is not convex, we don't apply gradient descent, but we would rather do some grid search. Precisely, our procedure consists in fixing $P^*_{1|0}$ to some value (by default I_d), and then try different diagonal matrices Q^* where the coefficients are in a defined list (by default, 0 or $2^k, k \in \{-30, \dots, 0\}$).

A grid search would yield a complexity $O(L^d)$ where L is the size of the list, and that could be very big even for moderate dimensions ($L = 31$ and $d = 10$ yields $L^d \approx 8 \cdot 10^{14}$). Instead, we optimize it in a greedy fashion first described in Obst, Vilmarest, and Goude, 2021: we begin with $Q^{*(0)} = 0$, and at each step of the procedure, we change one coefficient increasing the most the likelihood, see Algorithm 5. We stop when there is no possibility of increasing the likelihood with only one coefficient change. Thus, each step has a cost $O(Ld)$, and although there is no guarantee of less than L^d steps, in practice, we do approximately d, leading to an empirical $O(Ld^2)$ running time.

In the applications we considered, we believe the restriction for Q^* is an advantage of our approach. The evident drawback is that it degrades the optimal attainable likelihood. However, our applications are misspecified, and we believe our algorithm reduces overfitting while maintaining enough flexibility. In particular, we start from $Q^{*(0)} = 0$, and consequently, we obtain a sparse covariance matrix.

5.4 Experiment on Time Series With Delayed Observations

In most practical applications, the setting is not strictly online. The first challenge is the delay in the availability of the observations. In the context of electricity consumption, it is even more complex. The load is never perfectly known, but its estimation improves with time. An estimate is available within a short period, but the demand estimate is final only months afterward. In our applications in Part III, we didn't consider noisy data, but we assume that the estimate is final and exact with a delay k. As the state noise is centered and i.i.d., we can easily obtain the mean and covariance matrix of θ_t conditionally to the observations we know:

$$\hat{\theta}_{t|t-k} = \hat{\theta}_{t-k|t-k}, \qquad P_{t|t-k} = P_{t-k|t-k} + kQ\,. \tag{5.17}$$

A second operational constraint is that the model is not really updated at each time step. Indeed, models are updated once a week, for example. From a prediction perspective, this constraint means that the delay k is not constant anymore. The closer from the last update we are, the smaller k.

The literature is limited concerning estimating the variances in a state-space model with delayed observations. This is because if the state-space model is well-specified (that is, if Equations (5.1) and (5.2) generate the data), the introduction of a delay does not alter the data-generating process and the likelihood. In that case, the best forecaster is still the Kalman filter with the true state noise covariance matrix, using the estimate of Equation (5.17).

However, the data is misspecified in most practical applications, at least for those considered in this manuscript. A Kalman filter handles conveniently misspecified data due to its robustness. Still, for this sort of data, we should see the Kalman filter as a way to parametrize a gradient descent algorithm, and we claim that it is natural to apply a more empirical hyper-parameter selection.

Indeed, if we have some delay, we would prefer a careful model, not adapting too fast. Instead of maximizing Equation (5.5) we maximize an altered likelihood:

$$\sum_{t=1}^{n} \left(-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2 + x_t^\top P_{t|t-k} x_t) - \frac{1}{2}\frac{(y_t - \hat{\theta}_{t|t-k}^\top x_t)^2}{\sigma^2 + x_t^\top P_{t|t-k} x_t} \right).$$

The delay can also be a time-varying $k(t)$. We expect this variant of the *iterative grid search* to yield a smaller Q than the EM algorithm.

We confirm that intuition in practice on a toy data set. We use $d = 2$, a deterministic $\theta_t = \cos(\frac{2\pi t}{100}) \cdot (1,1)^\top$, then $x_t \sim \mathcal{N}(0, I)$ and $y_t - \theta_t^\top x_t \sim \mathcal{N}(0,1)$. The time period is $n = 10^3$ and the delay is $k = 50$, worst case scenario in terms of phase offset. We divide the data set in two, we learn the hyper-parameters on the first half and we compute the root-mean-square-error on the second. We obtain an error of 5.4 for the EM and 2.2 for the grid search. We display in Figure 5.2 the evolution of the first coordinate of θ_t as well as the two Kalman estimates.

5.5 Conclusion

In this chapter, we considered the choice of the hyper-parameters in a state-space model under the assumption of constant variances. We proposed two heuristics to maximize the likelihood. The first one is the expectation-maximization algorithm. Then we introduced the *iterative grid search* procedure. Three issues motivate this more empirical approach. The log-likelihood is not convex, therefore we claim there is no efficient algorithm to maximize it. In general, the

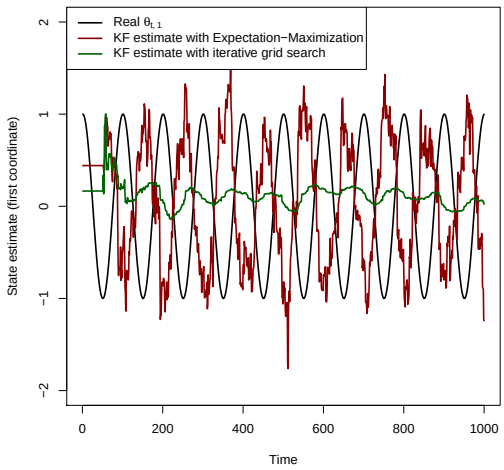


Figure 5.2 – Experiment on synthetic data with a constant delay $k = 50$.

data is not well-specified (generated by the state-space model with constant variances). Finally, real-world data are often available with some delay.

Both methods crucially rely on the constant assumption and have the same goal, maximum likelihood estimation. We present in Chapter 6 a totally different approach where the variances are estimated adaptively.

Chapter 6

Viking: Variational Bayesian Variance Tracking

We consider the problem of time series forecasting in an adaptive setting. We focus on the inference in state-space models under unknown and potentially time-varying noise variances. We introduce an augmented model in which the variances are represented as auxiliary Gaussian latent variables in a tracking mode. As variances are nonnegative, a transformation is chosen and applied to these latent variables. The inference relies on the online variational Bayesian methodology, which minimizes a Kullback-Leibler divergence at each time step. We observe that the minimum of the Kullback-Leibler divergence is an extension of the Kalman filter, taking into account the variance uncertainty. We design a novel algorithm named Viking, using these optimal recursive updates. We use second-order bounds for the auxiliary latent variables, whose optimum admit closed-form solutions. Experiments on synthetic data show that Viking behaves well and is robust to misspecification (violation of the state-space data generation assumption).

Contents

6.1 Introduction

Linear state-space models have been widely used to model a time series as a Gaussian random variable whose mean is a time-varying linear function of covariates. The linear parameter is a latent variable called state, and the hyper-parameters of the state-space model are the covariance matrices of the state and space noises. When these variances are known, optimal recursive estimation is achieved by the Kalman filter (Kalman and Bucy, 1961).

However, in most practical applications, the state and space noise variances are unknown. A vast amount of literature has emerged to choose them. The paradigm of time-invariant variances leads to maximum likelihood estimation on a historical data set (see for instance Brockwell and Davis, 2016; Durbin and Koopman, 2012, as well as Chapter 5). Another approach is to estimate these variances (fixed or not) in an online fashion, that is adaptive filtering (Mehra, 1972).

Recently, recursive variational Bayesian (VB) methods as introduced in (Beal, 2003; Šmídl and Quinn, 2006) have gathered attention in the Kalman filtering community. The objective is the online estimation of potentially time-variant parameters. The difference with the classical Bayesian approach is that an approximation is realized at each step in order to make the inference tractable: the distribution of the parameters is estimated by simple factorized distributions. The best factorized distribution is defined as the one minimizing its Kullback-Leibler divergence with the posterior.

A VB approach was first applied to estimate the observation noise covariance matrix in a Kalman filter (Sarkka and Nummenmaa, 2009), then extended in Agamennoni, Nieto, and Nebot, 2012 to be robust to non-Gaussian noise and in Särkkä and Hartikainen, 2013 to nonlinear state-space models. The covariance matrix is assumed diagonal and the prior used is a product of inverse Gamma distributions. To allow for a dynamical noise variance, the authors use a forgetting factor, multiplying the variances of the inverse Gamma posterior distributions by a constant. The method was extended with an inverse Wishart prior (Huang et al., 2017). At the same time, the authors apply the VB approach to correct the covariance matrix of the state after applying Kalman recursions with an inaccurate state noise covariance matrix. The inverse Wishart distribution appears as a nice conjugate prior to generalize the inverse Gamma distribution. More recently, another adaptive Kalman filter was proposed in Huang et al., 2020 to estimate the state and space noise covariance matrices simultaneously. The method applies Kalman filtering and smoothing on a slide window and could be described as an online Expectation-Maximization algorithm. In all these methods, the dynamics of the variances are introduced through a forgetting factor.

Up to our knowledge, to deal with unknown covariance matrices in state-space models, all existing methods apply at each step the standard Kalman filter with an estimate of the variances updated in an adaptive fashion. In other words, the Kalman filter is combined with a variance estimation algorithm. We claim that it is suboptimal and that the recursive update of the state estimates should leverage the variance uncertainty. This article treats the variances as auxiliary latent variables yielding an essential degree of freedom in an augmented latent representation.

We apply the VB approach, and we rely on second-order upper bounds to tackle the intractability of the VB step.

6.1.1 Overview

In Section 6.2 we present the state-space inference problem, we introduce the VB principle, and we motivate our augmented dynamical model. In Section 6.3 we study the VB minimization problem. The algorithm is detailed in Section 6.4, and we provide experimental results in Section 6.5.

6.1.2 Notations

Besides canonical notations we use the following.

— For any distribution P of probability density function (PDF) p, and any function ϕ, $\mathbb{E}_{x\sim P}[\phi(x)]$ is defined as $\int p(x)\phi(x)dx$.
— $\mathcal{N}(x \mid \mu, \Sigma)$ is the PDF at point x of the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$.
— For any matrix M, Δ_M is the vector composed of the diagonal coefficients of M. Reciprocally, for any vector v, D_v is the diagonal matrix whose diagonal is composed of the coefficients of v.
— If $\phi : \mathbb{R} \to \mathbb{R}$ and $x \in \mathbb{R}^d$, $\phi(x)$ is the d-dimensional vector obtained by applying ϕ to each coordinate of x.

6.2 Variance Tracking

We consider the problem of time series forecasting in the univariate setting for simplicity. At each time t we aim at forecasting $y_t \in \mathbb{R}$. To that end we have access to covariates $x_t \in \mathbb{R}^d$ as well as the past observations $x_1, y_1, \dots x_{t-1}, y_{t-1}$. We focus on a state-space representation where y_t is modelled as a linear function of x_t whose linear parameter evolves dynamically:

$$\theta_t = K\theta_{t-1} + \eta_t \,,$$
$$y_t = \theta_t^\top x_t + \varepsilon_t \,,$$

where $\eta_t \sim \mathcal{N}(0, Q_t)$ and $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)$ are the (independent) state and space noises, and the state follows the initial distribution $\theta_0 \sim \mathcal{N}(\hat{\theta}_0, P_0)$. When σ_t^2 and Q_t are known, the state vector θ_t given the past observations follows a Gaussian distribution whose mean and covariance can be estimated recursively by the standard Kalman filter (Kalman and Bucy, 1961). We focus on the setting where these variances are unknown and need to be estimated jointly with the state.

6.2.1 Bayesian Approach

We apply a Bayesian approach in order to estimate jointly the state θ_t and the variances σ_t^2, Q_t given the past observations. Remark, however, that the problem at hand remains the forecast of y_t; thus, the latent variable of interest is θ_t. Estimating σ_t^2 is necessary for a probabilistic forecast of y_t since it drives the noise variance. The covariance matrix Q_t is added to open flexibility for the estimation of the other variables in a dynamical way.

We introduce the filtration of the past observations $\mathcal{F}_t = \sigma(x_1, y_1, \dots, x_t, y_t)$. Then we define a prior on the latent variables $p(\theta_0, \sigma_0^2, Q_0 \mid \mathcal{F}_0)$, as well as a model on the dynamics of the three latent variables represented by a transition density $p(\theta_t, \sigma_t^2, Q_t \mid \theta_{t-1}, \sigma_{t-1}^2, Q_{t-1})$. At

each iteration t, the Bayesian approach consists of a prediction step using the dynamical model assumed and a filtering step using Bayes' rule:

$$\begin{aligned} &\text{Prediction:} && p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_{t-1})\,, \\ &\text{Filtering:} && p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_t)\,. \end{aligned}$$

In the case of known variances, the standard Kalman filter is applied. The prediction step yields $\hat{\theta}_{t|t-1}$ and $P_{t|t-1}$ such that $p(\theta_t \mid \mathcal{F}_{t-1}) = \mathcal{N}(\theta_t \mid \hat{\theta}_{t|t-1}, P_{t|t-1})$. Then the filtering step yields $\hat{\theta}_{t|t}$ and $P_{t|t}$ such that the posterior is $p(\theta_t \mid \mathcal{F}_t) = \mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})$.

We explain in the following paragraphs that for natural dynamical models, the posterior distribution is analytically intractable. Therefore we estimate it with simple distributions.

6.2.2 Variational Bayesian Approach

A standard approach, referred to as recursive variational Bayes (VB), is to approximate the posterior distribution recursively with a factorized distribution where each component is of a simple form (Šmídl and Quinn, 2006). We apply this framework, and we estimate our posterior distribution with a product of three distributions. We keep a Gaussian marginal for θ_t in order to coincide with the exact posterior in the degenerate setting where the variances are known. We introduce parametric distributions on σ_t^2 and Q_t of the form $P_{\Phi_{t|t}}$ and $P_{\Psi_{t|t}}$, where $\Phi_{t|t}$ and $\Psi_{t|t}$ are the parameters we want to estimate recursively. We denote by $p_{\Phi_{t|t}}$ and $p_{\Psi_{t|t}}$ their PDFs. We look for $\hat{\theta}_{t|t}, P_{t|t}, \Phi_{t|t}, \Psi_{t|t}$ such that the product $\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times P_{\Phi_{t|t}} \times P_{\Psi_{t|t}}$ is the best approximation of the posterior distribution denoted by $P_{\mathcal{F}_t}$. The approximation is quantified by the Kullback-Leibler (KL) divergence:

$$KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times P_{\Phi_{t|t}} \times P_{\Psi_{t|t}} \mid\mid P_{\mathcal{F}_t}\Big), \tag{6.1}$$

where $KL(P \mid\mid Q) = \int_x \log(p(x)/q(x))p(x)dx$ for any distributions P and Q of PDFs p and q. At each step, the VB approach yields a coupled optimization problem in three distributions.

The prediction step is determined by the dynamics we propose in the model. The state equation yields the following transition density:

$$p(\theta_t, \sigma_t^2, Q_t \mid \theta_{t-1}, \sigma_{t-1}^2, Q_{t-1}) = \mathcal{N}(\theta_t - K\theta_{t-1} \mid 0, Q_t)p(\sigma_t^2, Q_t \mid \sigma_{t-1}^2, Q_{t-1})\,.$$

Propagating the factorized approximation

$$p(\theta_{t-1}, \sigma_{t-1}^2, Q_{t-1} \mid \mathcal{F}_{t-1}) \approx \mathcal{N}(\theta_{t-1} \mid \hat{\theta}_{t-1|t-1}, P_{t-1|t-1})p_{\Phi_{t-1|t-1}}(\sigma_{t-1}^2)p_{\Psi_{t-1|t-1}}(Q_{t-1})\,,$$

the prediction step becomes:

$$\begin{aligned} p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_{t-1}) \approx &\iiint \mathcal{N}(\theta_t - K\theta_{t-1} \mid 0, Q_t)p(\sigma_t^2, Q_t \mid \sigma_{t-1}^2, Q_{t-1}) \\ &\qquad \mathcal{N}(\theta_{t-1} \mid \hat{\theta}_{t-1|t-1}, P_{t-1|t-1})p_{\Phi_{t-1|t-1}}(\sigma_{t-1}^2)p_{\Psi_{t-1|t-1}}(Q_{t-1})d\theta_{t-1}d\sigma_{t-1}^2 dQ_{t-1} \\ \approx &\mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^\top + Q_t) \\ &\qquad \iint p(\sigma_t^2, Q_t \mid \sigma_{t-1}^2, Q_{t-1})p_{\Phi_{t-1|t-1}}(\sigma_{t-1}^2)p_{\Psi_{t-1|t-1}}(Q_{t-1})d\sigma_{t-1}^2 dQ_{t-1}\,. \end{aligned}$$

The last double integral depends on the dynamical model imposed on σ_t^2 and Q_t. It is natural

to assume a dynamical model where the prediction step yields a factorized distribution:

$$p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_{t-1}) \approx \mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^\top + Q_t) p_{\Phi_{t|t-1}}(\sigma_t^2) p_{\Psi_{t|t-1}}(Q_t)\,.$$

Remark that this is in particular the case when the model is $\sigma_t^2 = \sigma_{t-1}^2$ and $Q_t = Q_{t-1}$.

The filtering step uses the prediction step as a prior and yields the following posterior PDF:

$$\begin{aligned} p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_t) = \frac{p(x_t, \mathcal{F}_{t-1})}{p(\mathcal{F}_t)} &\mathcal{N}(y_t \mid \theta_t^\top x_t, \sigma_t^2) \\ &\mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^\top + Q_t) p_{\Phi_{t|t-1}}(\sigma_t^2) p_{\Psi_{t|t-1}}(Q_t)\,. \end{aligned}$$

We remark that whatever the parametric distributions on σ_t^2 and Q_t, the joint posterior distribution of the state and variances cannot be factorized. We mean that when we assert that the posterior distribution is intractable. One could opt for numerical approaches. However, our objective is to obtain recursive algorithms. Therefore, it is natural to apply the VB approach in order to estimate the posterior distribution with a simple factorized one.

6.2.3 Absence of Conjugate Prior for Q_t

The term $\mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^\top + Q_t)$ in the posterior distribution makes a conjugate prior for Q_t impractical in a VB method. Indeed, the standard procedure to maximize the KL divergence (Tzikas, Likas, and Galatsanos, 2008) builds on the important following property:

Proposition 6.1. *Let $\mathcal{V} = \{\theta_t, \sigma_t^2, Q_t\}$ be the set of variables and $(p^\star_{\theta_t}, p^\star_{\sigma_t^2}, p^\star_{Q_t})$ the PDFs minimizing the KL divergence* (6.1)*. For any variable $x \in \mathcal{V}$, if $\mathbb{E}_{\mathcal{V}\setminus\{x\}}[\ln p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_t)]$ is of the form $\ln p_x(x)$ then there exists a constant c such that*

$$\ln p_x^\star(x) = \mathbb{E}_{\mathcal{V}\setminus\{x\}}[\ln p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_t)] + c\,.$$

Therefore the standard algorithm to solve the KL minimization problem (6.1) starts from some PDFs $(p_{\theta_t}, p_{\sigma_t^2}, p_{Q_t})$ and updates alternately one out of three, optimizing the KL with respect to this one distribution and fixing the other two. For each $x \in \mathcal{V}$, the update of its distribution fixing the other two consists of the computation of $\mathbb{E}_{\mathcal{V}\setminus\{x\}}[\ln p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_t)]$ and then the update

$$p_x(x) \propto \exp\left(\mathbb{E}_{\mathcal{V}\setminus\{x\}}[\ln p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_t)]\right).$$

This update works well for θ_t. Indeed, we obtain the logarithm of a Gaussian up to a constant:

$$\begin{aligned} \mathbb{E}_{\sigma_t^2, Q_t \sim P_{\Phi_{t|t}} \times P_{\Psi_{t|t}}}[\ln p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_t)] = &-\frac{1}{2}(y_t - \theta_t^\top x_t)^2 \mathbb{E}_{\sigma_t^2 \sim P_{\Phi_{t|t}}}[(\sigma_t^2)^{-1}] \\ &-\frac{1}{2}(\theta_t - K\hat{\theta}_{t-1|t-1})^\top \mathbb{E}_{Q_t \sim P_{\Psi_{t|t}}}[(KP_{t-1|t-1}K^\top + Q_t)^{-1}](\theta_t - K\hat{\theta}_{t-1|t-1}) + c\,, \end{aligned}$$

where c is a constant independent of θ_t.

We can also obtain a conjugate prior for σ_t^2. The inverse Gamma distribution, generalized by the inverse Wishart distribution, yields an exact closed-form solution of the KL minimization with respect to the distribution of σ_t^2. That is well explained by Sarkka and Nummenmaa, 2009.

However there is no conjugate prior for Q_t. Precisely, we have

$$\mathbb{E}_{\theta_t,\sigma_t^2\sim\mathcal{N}(\hat{\theta}_{t|t},P_{t|t})\times P_{\Phi_{t|t}}}[\ln p(\theta_t,\sigma_t^2,Q_t \mid \mathcal{F}_t)] = \ln p_{\Psi_{t|t-1}}(Q_t) - \frac{1}{2}\ln\det(KP_{t-1|t-1}K^\top + Q_t)$$
$$- \frac{1}{2}\operatorname{Tr}\left(((\hat{\theta}_{t|t} - K\hat{\theta}_{t|t-1})(\hat{\theta}_{t|t} - K\hat{\theta}_{t|t-1})^\top + P_{t|t})(KP_{t-1|t-1}K^\top + Q_t)^{-1}\right) + c\,,$$

where c is a constant independent of Q_t. In particular the usual inverse Wishart distribution is not a conjugate prior for Q_t.

Proposition 6.1 is not applicable. Therefore we directly optimize the KL divergence, see Section 6.3. We need approximations in the posterior estimation.

On the contrary, related work has focused on avoiding the $KP_{t-1|t-1}K^\top + Q_t$ matrix in order to apply the approach of Tzikas, Likas, and Galatsanos, 2008. In Huang et al., 2017 the authors focus on the correction of the matrix $P_{t|t-1}$, equivalent of $KP_{t-1|t-1}K^\top + Q_t$. That is the quantity obtained in the posterior distribution and the natural quantity to correct based on the observations. However, the authors do not estimate Q_t. Hence they cannot learn from previous observations what correction should be applied on the Kalman filter. More recently, the approach proposed by Huang et al., 2020 consists of applying a few iterations of Kalman smoothing with the previous estimates of the variances $\hat{\sigma}_{t-1}^2$ and $\hat{Q}_{t-1}$. Then the authors estimate the posterior distribution of σ_t^2, Q_t given $\mathcal{F}_t$ and the distribution of $(\theta_{t-L},\dots,\theta_t)$ obtained by Kalman smoothing. In that way they get rid of the crossed matrix $KP_{t-1|t-1}K^\top + Q_t$.

6.2.4 The Variance Tracking Model

We have presented the VB approach with general parametric distributions for σ_t^2 and Q_t. Let us detail how we define them in order to derive the algorithm Viking.

The recursive estimation of the posterior distribution does not suggest a natural distribution for the covariance matrix Q_t simplifying the KL divergence minimization. Therefore, we choose Gaussian representations. A significant advantage of a Gaussian latent variable is that dynamics is naturally introduced in the form of a random walk. In contrast, the dynamics is imposed on inverse Wishart distribution with a forgetting factor (Sarkka and Nummenmaa, 2009; Huang et al., 2020). However, as variances must be nonnegative, we transform these Gaussian variables. Specifically, we use $\sigma_t^2 = \exp(a_t)$ (log-normal distribution) and $Q_t = f(b_t)$, where a_t, b_t follow Gaussian distributions. We detail the choice of f in Section 6.3.2 where we define either scalar covariance matrices (proportional to I) or diagonal ones. Remark that b_t can be of any dimension, as long as $f(b_t)$ is a $d\times d$ positive semidefinite matrix. Thanks to this full Gaussian representation, the approach may be summarized as follows:

$$\begin{aligned}
&\theta_0 \sim \mathcal{N}(\hat{\theta}_0, P_0)\,, \quad a_0 \sim \mathcal{N}(\hat{a}_0, s_0)\,, \quad b_0 \sim \mathcal{N}(\hat{b}_0, \Sigma_0)\,,\\
&a_t - a_{t-1} \sim \mathcal{N}(0, \rho_a)\,, \quad b_t - b_{t-1} \sim \mathcal{N}(0, \rho_b I)\,,\\
&\theta_t - K\theta_{t-1} \sim \mathcal{N}(0, f(b_t))\,,\\
&y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \exp(a_t))\,,
\end{aligned}$$

where we introduce the (nonnegative) parameters ρ_a and ρ_b. These parameters govern the dynamics of the latent variables a_t and b_t, representing the variances σ_t^2 and Q_t.

Although we remarked in Section 6.2.3 that the inverse Gamma yields an analytical KL minimum with respect to the distribution of σ_t^2, we deliberately choose a log-normal distribution, for which KL minimization is inexact. Two reasons motivate that choice. First, we highlight the unity of the approach with a full Gaussian latent representation, with the appealing random

walk interpretation of their trajectories. Second, this choice demonstrates the robustness of Gaussian representations for the variances. Indeed, we show experimentally in Section 6.5.1 that, for known Q_t, our model with log-normal σ_t^2 is equivalent to the estimation of Sarkka and Nummenmaa, 2009 with an inverse Gamma. The introduced parameter ρ_a is the equivalent of the forgetting factor ρ of Sarkka and Nummenmaa, 2009 (ρ_a is close to 0 whereas ρ is close to 1).

In the preceding set of equations we implicitly assume that we have

$$p(\theta_t, a_t, b_t \mid \theta_{t-1}, a_{t-1}, b_{t-1}) = p(\theta_t \mid \theta_{t-1}, b_t) p(a_t \mid a_{t-1}) p(b_t \mid b_{t-1})\,.$$

Applying the VB approach, we look for $\hat{\theta}_{t|t}, P_{t|t}, \hat{a}_{t|t}, s_{t|t}, \hat{b}_{t|t}, \Sigma_{t|t}$ such that the product of Gaussian distributions $\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})$ is the best approximation of the posterior distribution.

Treating the approximation at time $t-1$ as a prior at time t, following the posterior derivation of Section 6.2.2 we obtain:

$$\begin{aligned} p(\theta_t, a_t, b_t \mid \mathcal{F}_t) = \frac{p(x_t, \mathcal{F}_{t-1})}{p(\mathcal{F}_t)} &\mathcal{N}(y_t \mid \theta_t^\top x_t, \exp(a_t))\mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^\top + f(b_t)) \\ &\mathcal{N}(a_t \mid \hat{a}_{t-1|t-1}, s_{t-1|t-1} + \rho_a)\mathcal{N}(b_t \mid \hat{b}_{t-1|t-1}, \Sigma_{t-1|t-1} + \rho_b I)\,. \end{aligned} \tag{6.2}$$

6.3 Kullback-Leibler Minimization

We first present a detailed expression of the KL divergence in the variance tracking model. The proof of the results of this section are deferred to Appendix C.

Lemma 6.1. *There exists a constant c independent of $\hat{\theta}_{t|t}, P_{t|t}, \hat{a}_{t|t}, s_{t|t}, \hat{b}_{t|t}, \Sigma_{t|t}$ such that*

$$\begin{aligned} &KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times \mathcal{N}(\hat{a}_{t|t}, s_{t|t}) \times \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; P_{\mathcal{F}_t}\Big) \\ &\quad = -\frac{1}{2}\log\det P_{t|t} - \frac{1}{2}\log(s_{t|t}) + \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)\exp(-\hat{a}_{t|t} + \frac{1}{2}s_{t|t}) + \frac{1}{2}\hat{a}_{t|t} \\ &\qquad - \frac{1}{2}\log\det\Sigma_{t|t} + \frac{1}{2}\mathbb{E}_{b_t \sim \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})}[\psi_t(b_t)] + \frac{1}{2(s_{t-1|t-1} + \rho_a)}(s_{t|t} + (\hat{a}_{t|t} - \hat{a}_{t-1|t-1})^2) \\ &\qquad + \frac{1}{2}\operatorname{Tr}\Big((\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^\top)(\Sigma_{t-1|t-1} + \rho_b I)^{-1}\Big) + c\,, \end{aligned}$$

where

$$\begin{aligned} \psi_t(b_t) &= \log\det(KP_{t-1|t-1}K^\top + f(b_t)) \\ &\quad + \operatorname{Tr}\Big((P_{t|t} + (\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})^\top)(KP_{t-1|t-1}K^\top + f(b_t))^{-1}\Big)\,. \end{aligned}$$

The rest of this section is devoted to minimizing the KL divergence expressed in Lemma 6.1. We present the optimization in the state distribution in Section 6.3.1. This yields insights on how to choose f, see Section 6.3.2. While the minimum of the KL divergence admits closed-form solutions with respect to $\hat{\theta}_{t|t}, P_{t|t}$, it does not with respect to the other parameters. We derive closed-form approximations to the VB recursive step in Sections 6.3.3 and 6.3.4. We use the first two moments of Gaussian distributions in second-order upper bounds. Minimizing the upper bounds does not necessarily lead to the minimization of the KL divergence, but it guarantees the decrease of the instantaneous KL divergence at each step.

6.3.1 State Estimation

We easily obtain a closed-form solution to minimize the KL divergence with respect to $\hat{\theta}_{t|t}, P_{t|t}$.

Theorem 6.1. *Given $\hat{a}_{t|t}, s_{t|t}, \hat{b}_{t|t}, \Sigma_{t|t}$, the values of $\hat{\theta}_{t|t}, P_{t|t}$ minimizing the KL divergence are given by*

$$A_t = \mathbb{E}_{b_t \sim \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})}[(KP_{t-1|t-1}K^\top + f(b_t))^{-1}]\,, \tag{6.3}$$

$$P_{t|t} = A_t^{-1} - \frac{A_t^{-1} x_t x_t^\top A_t^{-1}}{x_t^\top A_t^{-1} x_t + \exp(\hat{a}_{t|t} - \frac{1}{2} s_{t|t})}\,, \tag{6.4}$$

$$\hat{\theta}_{t|t} = K\hat{\theta}_{t-1|t-1} + \frac{P_{t|t} x_t}{e^{\hat{a}_{t|t} - s_{t|t}/2}}(y_t - x_t^\top K\hat{\theta}_{t-1|t-1})\,. \tag{6.5}$$

The updates defined above are the ones of the Kalman filter with known variances σ_t^2 and Q_t, where we have replaced σ_t^2 with $\exp(\hat{a}_{t|t} - \frac{1}{2} s_{t|t})$ which is $\mathbb{E}_{a_t \sim \mathcal{N}(\hat{a}_{t|t}, s_{t|t})}[\exp(a_t)^{-1}]^{-1}$ and $KP_{t-1|t-1}K^\top + Q_t$ with $\mathbb{E}_{b_t \sim \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})}[(KP_{t-1|t-1}K^\top + f(b_t))^{-1}]^{-1}$. If $s_{t|t} = 0, \Sigma_{t|t} = 0$ then we know the variances and we obtain the Kalman filter with $\sigma_t^2 = \exp(\hat{a}_{t|t})$ and $Q_t = f(\hat{b}_{t|t})$. Otherwise if $\Sigma_{t|t} \neq 0$, the result states that the update of the Kalman filter with unbiased estimated variances in place of the unknown variances is suboptimal in the sense of the Kullback-Leibler divergence. It implies also that we do not expect to obtain unbiased estimates of the variances.

It is important to remark that as long as $\rho_b > 0$ we do not have the convergence of $\Sigma_{t|t}$ to 0. Therefore we do not recover the standard Kalman filter asymptotically. On the contrary, existing adaptive Kalman filters use the standard Kalman recursive updates with estimates of the variances (Sarkka and Nummenmaa, 2009; Agamennoni, Nieto, and Nebot, 2012; Särkkä and Hartikainen, 2013; Huang et al., 2017; Huang et al., 2020). Therefore, in a well-specified setting where the underlying generating process is the state-space model with time-invariant variances, our method should be outperformed by adaptive Kalman filters with consistent variance estimates. We believe this drawback is a reasonable price to pay to get robustness to misspecification.

Furthermore note that (6.5) may be interpreted as a gradient step on the quadratic loss, where instead of a gradient step size we have the *preconditioning* matrix $P_{t|t}/\exp(\hat{a}_{t|t} - \frac{1}{2} s_{t|t})$. Therefore the algorithm derived in this article may be seen as a way to parameterize a second-order stochastic gradient algorithm.

6.3.2 Choice of f

The natural transformation for the latent variables a_t and b_t is the exponential, see Tyagi and Davis, 2008 for a filter on latent variables lying in a Riemannian manifold. We use the exponential to represent σ_t^2. However setting $f(b_t) = \exp(b_t)I$ for a unidimensional b_t contradicts a *careful* property that we define as follows using the gradient interpretation of Section 6.3.1. We claim that the algorithm should be more careful with uncertainty ($\Sigma_{t|t} \succ 0$) than without ($\Sigma_{t|t} = 0$). By more careful we mean smaller gradient steps, that is formally $A_t^{-1} \preccurlyeq KP_{t-1|t-1}K^\top + f(\hat{b}_{t|t})$. By Jensen's inequality, we have

$$A_t \succcurlyeq \left(KP_{t-1|t-1}K^\top + \mathbb{E}_{b_t \sim \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})}[f(b_t)]\right)^{-1}.$$

Therefore a sufficient (but not necessary) condition providing the careful property is f concave, again thanks to Jensen, and that is the contrary of the exponential. Unfortunately we cannot have both f concave and $f \succcurlyeq 0$ (unless f is constant). We propose to use a function which is zero on negative numbers and concave elsewhere:

$$\phi(b) = \begin{cases} 0 & if\ b < 0\,, \\ \log(1+b) & if\ b \geq 0\,. \end{cases}$$

Then we consider two settings for f: First a scalar setting where $f(b_t) = \phi(b_t)I$ for a unidimensional b_t. Second, a diagonal setting where $b_t \in \mathbb{R}^d$ and $f(b_t) = D_{\phi(b_t)}$ is a diagonal matrix whose diagonal coefficients are defined by the function ϕ applied to each coefficient of b_t.

6.3.3 Observation Noise Variance Estimation

We present recursive updates for the observation variance distribution. As the KL divergence does not admit analytical solutions with respect to $\hat{a}_{t|t}, s_{t|t}$, we optimize an upper bound of the KL. Instead of the true optimum, we find approximations where the guarantee is to decrease the instantaneous KL at each iteration.

Optimum in $s_{t|t}$

We are looking for $s_{t|t} \geq 0$ minimizing the KL divergence. As the conditional variance of a_t given $\mathcal{F}_{t-1}$ is $s_{t-1|t-1} + \rho_a$, we look for $s_{t|t}$ in the interval $[0, s_{t-1|t-1} + \rho_a]$. In this interval we simply use a linear upper bound for the exponential:

Proposition 6.2. *For any $s_{t|t} \in [0, s_{t-1|t-1} + \rho_a]$ we have*

$$\begin{aligned} &KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times \mathcal{N}(\hat{a}_{t|t}, s_{t|t}) \times \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; P_{\mathcal{F}_t}\Big) \\ &\quad \leq \frac{1}{4}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t) e^{-\hat{a}_{t|t}} s_{t|t} + \frac{1}{2}(s_{t-1|t-1} + \rho_a)^{-1} s_{t|t} - \frac{1}{2}\log(s_{t|t}) + c_s\,, \end{aligned}$$

where c_s is a constant independent of $s_{t|t}$. Furthermore, the upper bound is minimized by:

$$s_{t|t} = \Big((s_{t-1|t-1} + \rho_a)^{-1} + \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t) e^{-\hat{a}_{t|t}}\Big)^{-1}. \tag{6.6}$$

Optimum in $\hat{a}_{t|t}$

To upper bound the exponential with a polynomial form also in $\hat{a}_{t|t}$ we need to bound $\hat{a}_{t|t}$, and we consider the segment $[\hat{a}_{t-1|t-1} - M_a, \hat{a}_{t-1|t-1} + M_a]$ (we set arbitrarily $M_a = 3s_{t-1|t-1}$ to include more than 99% of the Gaussian distribution).

Proposition 6.3. *For any $\hat{a}_{t|t} \in [\hat{a}_{t-1|t-1} - M_a, \hat{a}_{t-1|t-1} + M_a]$ we have*

$$\begin{aligned} &KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times \mathcal{N}(\hat{a}_{t|t}, s_{t|t}) \times \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; P_{\mathcal{F}_t}\Big) \\ &\quad \leq \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t) e^{-\hat{a}_{t-1|t-1} + s_{t|t}/2}\Big(-(\hat{a}_{t|t} - \hat{a}_{t-1|t-1}) + \frac{e^{M_a}}{2}(\hat{a}_{t|t} - \hat{a}_{t-1|t-1})^2\Big) \\ &\qquad + \frac{1}{2}(s_{t-1|t-1} + \rho_a)^{-1}(\hat{a}_{t|t} - \hat{a}_{t-1|t-1})^2 + \frac{1}{2}\hat{a}_{t|t} + c_a\,, \end{aligned}$$

where c_a is a constant independent of $\hat{a}_{t|t}$. Furthermore the upper bound is minimized by:

$$\begin{aligned}\hat{a} &= \hat{a}_{t-1|t-1} + \frac{1}{2}\Big(\frac{1}{s_{t-1|t-1}+\rho_a} + \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)e^{-\hat{a}_{t-1|t-1}+s_{t|t}/2+M_a}\Big)^{-1} \\ &\qquad \Big(((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)e^{-\hat{a}_{t-1|t-1}+s_{t|t}/2} - 1\Big)\,, \\ \hat{a}_{t|t} &= \max(\min(\hat{a}, \hat{a}_{t-1|t-1} + M_a), \hat{a}_{t-1|t-1} - M_a)\,. \end{aligned} \tag{6.7}$$

We remark that $((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)e^{-\hat{a}_{t-1|t-1}+s_{t|t}/2} - 1$ is the gradient with respect to $\hat{a}$ of

$$\begin{aligned}&\mathbb{E}_{(\theta_t,a_t)\sim\mathcal{N}(\hat{\theta}_{t|t},P_{t|t})\times\mathcal{N}(\hat{a},s_{t|t})}[\log\mathcal{N}(y_t \mid \theta_t^\top x_t, \exp(a_t))] \\ &\quad = -\frac{1}{2}\hat{a} - \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)e^{-\hat{a}+s_{t|t}/2}\,,\end{aligned}$$

therefore (6.7) may be seen as a projected gradient step on an expected log-likelihood.

6.3.4 State Noise Covariance Matrix Estimation

The minimum of the Kullback-Leibler is also intractable in $\hat{b}_{t|t}, \Sigma_{t|t}$ due to the absence of analytical form for the expected value of ψ_t. In the following we focus on the specific settings that are introduced in Section 6.3.2, namely the scalar setting $f(b_t) = \phi(b_t)I$ and the diagonal setting $f(b_t) = D_{\phi(b_t)}$. For these two possible choices of f we have the following second-order upper bound for ψ_t:

Proposition 6.4. *In the scalar and diagonal settings defined in Section 6.3.2, for any t such that $f(\hat{b}_{t-1|t-1}) \succ 0$, the following holds for any b_t in a neighborhood of $\hat{b}_{t-1|t-1}$:*

$$\psi_t(b_t) \leq \psi_t(\hat{b}_{t-1|t-1}) + \frac{\partial\psi_t}{\partial b_t}\Big|_{\hat{b}_{t-1|t-1}}^\top (b_t - \hat{b}_{t-1|t-1}) + \frac{1}{2}(b_t - \hat{b}_{t-1|t-1})^\top H_t (b_t - \hat{b}_{t-1|t-1})\,,$$

where $B_t = P_{t|t} + (\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})^\top$, $C_t = KP_{t-1|t-1}K^\top + f(\hat{b}_{t-1|t-1})$, and then

$$\begin{aligned}\frac{\partial\psi_t}{\partial b_t}\Big|_{\hat{b}_{t-1|t-1}} &= \operatorname{Tr}(C_t^{-1}(I - B_t C_t^{-1}))\phi'(\hat{b}_{t-1|t-1})\,, \\ H_t &= -\operatorname{Tr}(C_t^{-1}B_tC_t^{-1})\phi''(\hat{b}_{t-1|t-1}) + 2\operatorname{Tr}(C_t^{-2}B_tC_t^{-1})\phi'(\hat{b}_{t-1|t-1})^2\,,\end{aligned}$$

in the scalar setting, and

$$\begin{aligned}\frac{\partial\psi_t}{\partial b_t}\Big|_{\hat{b}_{t-1|t-1}} &= \Delta_{C_t^{-1}(I-B_tC_t^{-1})} \odot \phi'(\hat{b}_{t-1|t-1})\,, \\ H_t &= -\Big(C_t^{-1}B_tC_t^{-1}D_{\phi''(\hat{b}_{t-1|t-1})}\Big) \odot I + 2C_t^{-1}B_tC_t^{-1} \odot C_t^{-1} \odot \phi'(\hat{b}_{t-1|t-1})\phi'(\hat{b}_{t-1|t-1})^\top\,,\end{aligned}$$

in the diagonal setting, with $\odot$ the Hadamard (pointwise) product.

The upper bound of the Kullback-Leibler divergence obtained thanks to the proposition above admits a closed-form minimum:

Proposition 6.5. *In the scalar and diagonal settings, for any t such that $f(\hat{b}_{t-1|t-1}) \succ 0$ and any $\hat{b}_{t|t}, \Sigma_{t|t}$,*

$$\begin{aligned} &KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times \mathcal{N}(\hat{a}_{t|t}, s_{t|t}) \times \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; P_{\mathcal{F}_t}\Big) \\ &\quad \leq -\frac{1}{2}\log\det\Sigma_{t|t} + \frac{1}{2}\frac{\partial\psi_t}{\partial b_t}\Big|_{\hat{b}_{t-1|t-1}}^{\top}(\hat{b}_{t|t} - \hat{b}_{t-1|t-1}) \\ &\qquad + \frac{1}{2}\operatorname{Tr}\Big((\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^{\top})\Big((\Sigma_{t-1|t-1} + \rho_b I)^{-1} + \frac{1}{2}H_t\Big)\Big) + c_b\,, \end{aligned}$$

where H_t is defined in Proposition 6.4 and c_b is a constant independent of $\hat{b}_{t|t}, \Sigma_{t|t}$. The minimum of the upper bound detailed above is obtained with:

$$\Sigma_{t|t} = \Big((\Sigma_{t-1|t-1} + \rho_b I)^{-1} + \frac{1}{2}H_t\Big)^{-1}, \tag{6.8}$$

$$\hat{b}_{t|t} = \hat{b}_{t-1|t-1} - \frac{1}{2}\Sigma_{t|t}\frac{\partial\psi_t}{\partial b_t}\Big|_{\hat{b}_{t-1|t-1}}. \tag{6.9}$$

Similarly as (6.7) we can interpret (6.9) as a gradient step on ψ_t and we can remark that $\psi_t(\hat{b})$ is the following expected log-likelihood:

$$\psi_t(\hat{b}) = \mathbb{E}_{\theta_t \sim \mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})}[\log\mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^{\top} + f(\hat{b}))]\,.$$

Thus, except the exact recursive steps on $\hat{\theta}_{t|t}, P_{t|t}$ which are extensions of the Kalman filter steps, our recursive steps resemble stochastic gradient variational Bayes as described in Knowles, 2015. This novel class of algorithms is very popular for tuning complex deep learning networks; see for instance Kingma and Welling, 2014; Tjandra et al., 2015. There, the expectation of the log-likelihood is approximated by Monte-Carlo simulation, and only the first order of the gradient is used.

6.4 Viking

We now introduce the algorithm following the recursive updates described in the previous section.

6.4.1 Definition of the Algorithm

Theorem 6.1 yields exact recursive updates for $\hat{\theta}_{t|t}, P_{t|t}$ but A_t^{-1} does not admit an explicit form. We propose to run Monte-Carlo estimation of A_t with very small samples ($n_{\text{mc}} = 10$ draws by default). As the KL optimization is a coupled problem we solve it in a classical iterative fashion, that is, we repeat N times the updates alternately ($N = 2$ is a good default value). We summarize the procedure in Algorithm 6. We name it Viking (**V**ariational Bayes**i**an Variance Trac**king**).

6.4.2 Complexity

We decompose the number of operations of Viking in Table 6.1. Although matrix multiplication and inversion have the same asymptotic complexity, inversion is more costly in practice.

Algorithm 6 : Viking at time step t

Time-invariant parameters: $\rho_a, \rho_b, n_{\text{mc}}, f$.
Default: $\rho_a = e^{-9}, \rho_b = e^{-6}, n_{\text{mc}} = 10, f(\cdot) = D_{\phi(\cdot)}$.
Inputs: $\hat{\theta}_{t-1|t-1}$, $P_{t-1|t-1}$, $\hat{a}_{t-1|t-1}$, $s_{t-1|t-1}$, $\hat{b}_{t-1|t-1}$, $\Sigma_{t-1|t-1}$, x_t, y_t.
Initialize:
Set $\hat{a}_{t|t}^{(0)} = \hat{a}_{t-1|t-1}$, $s_{t|t}^{(0)} = s_{t-1|t-1} + \rho_a$.
Set $\hat{b}_{t|t}^{(0)} = \hat{b}_{t-1|t-1}$, $\Sigma_{t|t}^{(0)} = \Sigma_{t-1|t-1} + \rho_b$.
Iterate: for $i = 1, \ldots, N$:
— 1. Set A_t using (6.3) with Monte-Carlo from n_{mc} samples of $\mathcal{N}(\hat{b}_{t|t}^{(i-1)}, \Sigma_{t|t}^{(i-1)})$. Compute A_t^{-1}.
2. Set $P_{t|t}^{(i)}, \hat{\theta}_{t|t}^{(i)}$ using (6.4) and (6.5), with A_t^{-1} from Step 1 and $\hat{a}_{t|t}^{(i-1)}, s_{t|t}^{(i-1)}$.
— **If we learn σ_t^2:**
3. Set $s_{t|t}^{(i)}$ using (6.6) with $\hat{\theta}_t^{(i)}, P_t^{(i)}, \hat{a}_{t|t}^{(i-1)}$.
4. Set $\hat{a}_{t|t}^{(i)}$ using (6.7) with $\hat{\theta}_t^{(i)}, P_t^{(i)}, s_{t|t}^{(i)}$.
— **If we learn Q_t:**
5. Set $\Sigma_{t|t}^{(i)}, \hat{b}_{t|t}^{(i)}$ using (6.8) and (6.9). Apply threshold $\hat{b}_{t|t}^{(i)} = \max(\hat{b}_{t|t}^{(i)}, 0)$.
Outputs: $\hat{\theta}_{t|t} = \hat{\theta}_{t|t}^{(N)}, P_{t|t} = P_{t|t}^{(N)}, \hat{a}_{t|t} = \hat{a}_{t|t}^{(N)}, s_{t|t} = s_{t|t}^{(N)}, \hat{b}_{t|t} = \hat{b}_{t|t}^{(N)}, \Sigma_{t|t} = \Sigma_{t|t}^{(N)}$.

Steps	Operations
1	$n_{\text{mc}}S + (n_{\text{mc}} + 1)I(d) + \mathcal{O}(M(d))$
2	$\mathcal{O}(d^2)$
3 and 4	$\mathcal{O}(d^2)$
5	$3I(d) + \mathcal{O}(M(d))$
Whole	$N\big(n_{\text{mc}}S + (n_{\text{mc}} + 4)I(d) + \mathcal{O}(M(d))\big)$

Table 6.1 – Complexity of Algorithm 6. S denotes the complexity of gaussian draw, $M(d)$ and $I(d)$ denote the complexity of matrix multiplication and inversion.

We suggest $N = 2$ and $n_{\text{mc}} = 10$ as default. Therefore, the complexity of Viking is essentially driven by the complexity of matrix inversion. Consequently, it is proportional to the one of methods relying on Kalman smoothing as in Huang et al., 2020.

6.5 Experiments

We run several experiments, arguing that our method behaves well for misspecified data. We begin with well-specified data generated under a state-space model with smoothly varying variances. Then we focus on misspecified data.

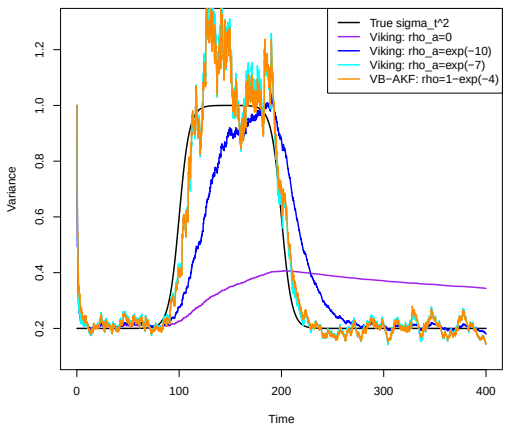


Figure 6.1 – Trajectory of the observation variance estimated by our algorithm and compared to the estimate provided by Sarkka and Nummenmaa, 2009. For both methods, we display the expected value of the estimated distributions.

6.5.1 Well-Specified Data with Unknown σ_t^2 and Known Q_t

We reproduce the experiment presented in Sarkka and Nummenmaa, 2009 on the stochastic resonator model:

$$\theta_{t+1} - \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\omega\Delta t) & \frac{\sin(\omega\Delta t)}{\omega} \\ 0 & -\omega\sin(\omega\Delta t) & \cos(\omega\Delta t) \end{pmatrix} \theta_t \sim \mathcal{N}(0, Q)\,,$$
$$y_t - (\theta_{t,1} + \theta_{t,2}) \sim \mathcal{N}(0, \sigma_t^2)\,,$$

where we set $\omega = 0.05$ and $\Delta t = 0.1$ and the known covariance matrix of the process noise is $Q = D_{(0.01,0,0.0001)}$. We display the variance trajectories for one simulation in Figure 6.1, and we observe that Viking almost coincide with VB-AKF (Sarkka and Nummenmaa, 2009). Also, running the experiment 100 times, both methods obtain the same mean squared error (MSE): 0.46981 for Viking and 0.46989 for VB-AKF (the 100 MSE have a standard deviation of approximately 0.01 for both). In this comparison, we take the best value of ρ_a for Viking as well as the best ρ for the VB-AKF in the list $e^{-i}, 1 \leq i \leq 10$. These very similar performances are a good *a posteriori* justification of the use of a log-normal distribution for σ_t^2. This distribution is close to the inverse Gamma, and we don't see any deterioration of performances.

6.5.2 Well-Specified Data with Unknown σ_t^2 and Q_t

We run a second simulation inspired by Huang et al., 2020 in a well-specified setting. We generate $x_t \in [0, 1]^5$ using two possible alternatives:

1. **Uniform i.i.d. design:** (x_t) is independent identically distributed. For each t, x_t is composed of 4 independent coefficients generated with uniform distributions on $[0, 1]$ and one deterministic 1 coefficient.

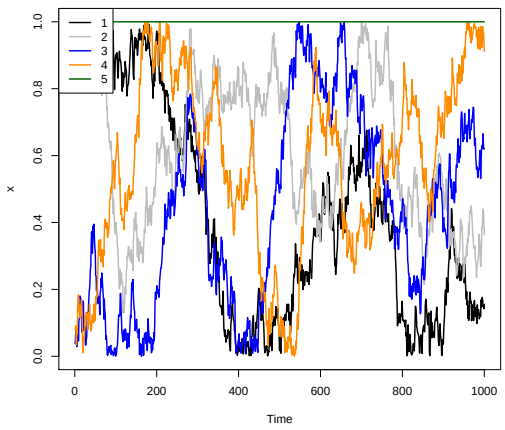


Figure 6.2 – Example of trajectory of the 5 components of the vector x_t considered in the setting *uniform non-iid*.

2. **Uniform non-i.i.d. design:** (x_t) has the same distribution but is not i.i.d., a sample is displayed in Figure 6.2. Precisely x_1 is generated as before. Then for $j \in \{1, 2, 3, 4\}$ and $t \geq 2$, we consider $z_{t,j} = x_{t-1,j} + \varepsilon_{t,j}$ where $\varepsilon_{t,j} \sim \mathcal{N}(0, 10^{-3})$ and we generate

$$x_{t,j} = \begin{cases} z_{t,j} \text{ if } 0 \leq z_{t,j} \leq 1\,, \\ \lceil z_{t,j} \rceil - z_{t,j} \text{ otherwise.} \end{cases}$$

Then we generate y_t by the following state-space model:

$$\begin{aligned} \theta_0 &\sim \mathcal{N}(0, I)\,, \\ \theta_t - \theta_{t-1} &\sim \mathcal{N}(0, Q_t)\,, \\ y_t - \theta_t^\top x_t &\sim \mathcal{N}(0, \sigma_t^2)\,, \end{aligned}$$

where

$$\begin{aligned} \sigma_t^2 &= 1 + 0.1 \cos \frac{4\pi t}{n}\,, \\ Q_t &= \Big(0.25 + 0.2 \cos \frac{4\pi t}{n}\Big) D_{(0,0,1,1,1)}\,. \end{aligned}$$

The simulation time is $n = 10^3$. In Figure 6.3 we compare Viking to the slide window variational adaptive Kalman filter (SWVAKF) introduced by Huang et al., 2020, which we tune in several ways. First, we increase the window length from 5 to 20, resulting in a significant improvement at the cost of more computations. Second, we tune the forgetting factor, and to play fair with Viking, we define different forgetting factors for the estimation of σ_t^2 and Q_t. We select the best *a posteriori*, and we do the same for Viking. Third, we enforce diagonal and scalar variants of the SWVAKF: the diagonal variant is defined by replacing by 0 each non-diagonal coefficient after each update. On top of that, in the scalar variant, we replace each diagonal coefficient with the averaged diagonal.

6.5.3 Misspecified Data with Unknown σ_t^2 and Q_t

To analyze the sensitivity to misspecification, we consider a state-space model with two states evolving independently with identical processes, and the observation is generated using one of them uniformly at random. That is summarized by the following set of equations:

$$\begin{aligned}
&\theta_0^{(i)} \sim \mathcal{N}(0, I)\,, && i \in \{0,1\}\,,\\
&\theta_t^{(i)} - 0.9\,\theta_t^{(i)} \sim \mathcal{N}(0, Q_t)\,, && i \in \{0,1\}\,,\\
&i_t \sim \mathcal{B}(1/2)\,,\\
&y_t - \theta_t^{(i_t)\top} x_t \sim \mathcal{N}(0, \sigma_t^2)\,,
\end{aligned}$$

where we assume all Gaussian noises to be independent of each other and of (i_t). We consider the same settings for x_t as well as the same variances σ_t^2, Q_t defined in Section 6.5.2.

The contraction (here by a coefficient 0.9) is necessary to have the convergence of the distribution of y_t as well as of the conditional distribution of y_t given the filtration $\mathcal{F}_{t-1}$. In the tracking mode (no contraction), the variance of the conditional distribution would diverge to ∞, and therefore the error of any forecasting strategy would also diverge to ∞.

We refer to Figure 6.3 for an evaluation on the mean squared error. We observe that Viking in the diagonal setting behaves poorly compared to the SWVAKF for well-specified data with i.i.d. design but better in the other three experiments. As mentioned in Section 6.3.1, we believe it is natural that a consistent adaptive Kalman filter should be closer to the true Kalman filter than our algorithm, which cannot be written using Kalman recursion. However, the careful property (see the design of f in Section 6.3.2) allows us to outperform existing methods for misspecified data. To a minor extent, this interpretation of the observation generation may be transposed to the design generation. Indeed, in our non-i.i.d. design, a shift in the data should be harder to attribute to one state coefficient, and therefore it should be harder to learn the variances. That is why the difference between the two Kalman filters with constant variances is more minor. Thus the latent model should not be trusted too much.

6.5.4 Impact of n_{mc}

The number of Monte-Carlo samples used at each step to compute A_t^{-1} is a crucial factor of the complexity of Viking. Therefore, it is necessary to evaluate its impact on performance in order to reach the best compromise between forecasting and computational efficiencies. We refer to Figure 6.4 for an evaluation of the error with different values of n_{mc}. Setting $n_{\mathrm{mc}} = 10$ as default seems reasonable.

6.6 Conclusion

We have introduced Viking, an adaptive time series forecasting algorithm relying on state-space models with unknown state and space variances. We derived an augmented latent model and applied variational Bayes for the inference. We extend the Kalman filter to an uncertain environment. For the additional latent variables, we use approximative steps close to the recursive steps of stochastic gradient variational Bayes. The prediction performances are better than state-of-the-art in misspecified settings at the same computational cost.

The choice of the function applied to the latent variable to obtain the state noise covariance matrix is a perspective of future research. We provide a specific choice leading to promising experimental results on simulations in both well-specified and misspecified settings. However, we

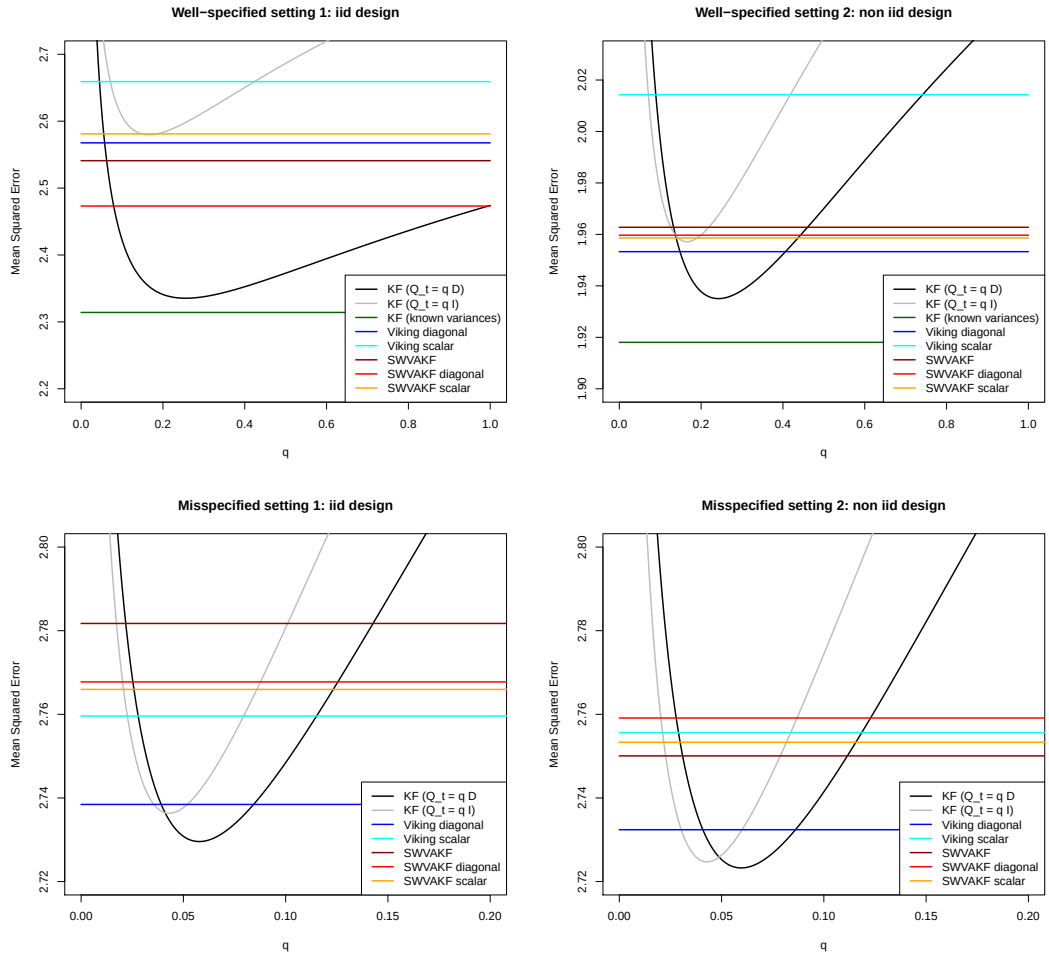


Figure 6.3 – Mean Squared Error in the four settings introduced in Sections 6.5.2 and 6.5.3: i.i.d. (left) or non-i.i.d. (right) design, well-specified (top) or misspecified (bottom). We compare Viking to the SWVAKF of Huang et al., 2020 in the scalar and diagonal settings. For Viking we set $n_{\text{mc}} = 10$. The oracles to which we compare are the Kalman filter with known variances when they exist (well-specified settings) and two Kalman filters with constant variances: the state noise covariance is either $Q = q \cdot D_{(0,0,1,1,1)}$ or $Q = q \cdot I$ and in both we set the space noise variance to $\sigma^2 = 1$. We evaluate through the mean squared error on the second half of the experiment in order to not depend too much on the initialization (even if we have same initial expected variances for Viking and SWVAKF).

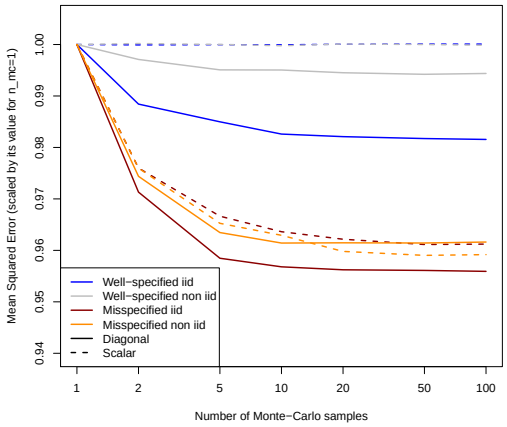


Figure 6.4 – Mean Squared Error of Viking as a function of n_{mc}. We scale by the mean squared error of the algorithm with $n_{\text{mc}} = 1$ in order to fit the different algorithms (diagonal and scalar settings) as well as the different experiments (i.i.d. or non-i.i.d. design, well-specified or misspecified) in the same graph.

wrote most of the article considering this function is a parameter of Viking because we believe other functions may be of interest.

Part III

Application to Electricity Load Forecasting

Chapter 7

Electricity Load Forecasting in France During Covid

The coronavirus disease 2019 (COVID-19) pandemic has urged many governments in the world to enforce a strict lockdown where all nonessential businesses are closed and citizens are ordered to stay at home. One of the consequences of this policy is a significant change in electricity consumption patterns. Since load forecasting models rely on calendar or meteorological information and are trained on historical data, they fail to capture the significant break caused by the lockdown and have exhibited poor performances since the beginning of the pandemic. In this paper we introduce two methods to adapt generalized additive models, alleviating the aforementioned issue. Using Kalman filters and fine-tuning allows to adapt quickly to new electricity consumption patterns without requiring exogenous information. The proposed methods are applied to forecast the electricity demand during the French lockdown period, where they demonstrate their ability to significantly reduce prediction errors compared to traditional models. Finally, expert aggregation is used to leverage the specificities of each predictions and enhance results even further.

This chapter is based on a joint work with David Obst and Yannig Goude published in IEEE Transactions on Power Systems.

Contents

7.1 Introduction

Accurate electricity load forecasting is of paramount importance for the balancing of the electricity grid, since they are the main inputs of the production planning at different horizons (Bunn and Farmer, 1985) and storage capacities are still limited regarding the consumption needs. Load forecasting is performed at different horizons of time, ranging from intra-day (10 minutes to 24 hours ahead) to daily, weekly, monthly or even a few years in advance for industrial needs covering production planning, demand response, grid management, electricity trading, risk management, optimization of production units maintenance and commercialization.

The field has been thoroughly studied the past decades, especially by the time series, statistics and machine learning communities. Time series approaches are very efficient for very-short term forecasts (typically less than 24 hours ahead). They rely on auto-regressive moving-average (ARIMA) models (Huang and Shih, 2003) or functional approaches (Antoniadis et al., 2016; Cho et al., 2013) exploiting daily and weekly patterns in the electricity load data. Machine learning models are usually stronger at incorporating exogenous information for short and mid-term predictions (more than 1 day ahead). They use calendar characteristics (such as the time of the year, day of the week...) as well as meteorological effects (temperature, wind speed) or tariff options as inputs and are then trained on a large set of historical data (usually 3 to 5 years). A good overview of load forecasting practices has been given by the Global Energy Forecasting Competitions (GEFCOM) (Hong, Pinson, and Fan, 2014). Popular algorithms include black box machine learning models such as gradient boosting machines (Lloyd, 2014) and neural networks (Park et al., 1991; Ryu, Noh, and Kim, 2017) or statistical models like Generalized Additive Models (GAM) (Pierrot and Goude, 2011; Fan and Hyndman, 2012; Goude, Nedellec, and Kong, 2013; Fasiolo et al., 2021). Black box models are attractive due to their good forecasting performances but generally suffer from their lack of interpretability. GAMs are very attractive to electric utilities as they combine the flexibility of fully nonparametric models, the simplicity of multiple regression model and are computationally efficient to scale with big data (Wood, Goude, and Shaw, 2015). The main French electricity provider, EDF (Électricité de France) uses GAM as their lead forecasting tool.

However, the coronavirus pandemic has significantly affected consumption patterns all over the world. As presented by Narajewski and Ziel, 2020; IEA, 2020, the closure of nonessential businesses as well as the stay-at-home directives have led to a significant drop of the power demand and changes in the daily consumption patterns. Figure 7.1 shows the French and Italian electricity load over time in 2020, whose decrease due to the lockdown (which happens before in Italy) is clearly seen. Daily profiles of the French consumption before and after the lockdown are represented in Figure 7.2. After lockdown for both countries the daily shapes of the load have converged towards the one of Saturdays.

Since models are trained on historical data and make the underlying assumption that future behavior will be similar to past one, they will fail to produce satisfactory predictions during the lockdown period. For instance in France GAM usually achieve around 1% MAPE (mean absolute percentage error) (Pierrot and Goude, 2011), but were around 5% during the first few weeks of the lockdown thus requiring manual intervention to correct the model forecasts. Not only do these poor forecasts have a high cost for electricity producers and system operators, but they represent a threat to the proper functioning of the electrical network as well, which could have even more consequences than usual during a pandemic.

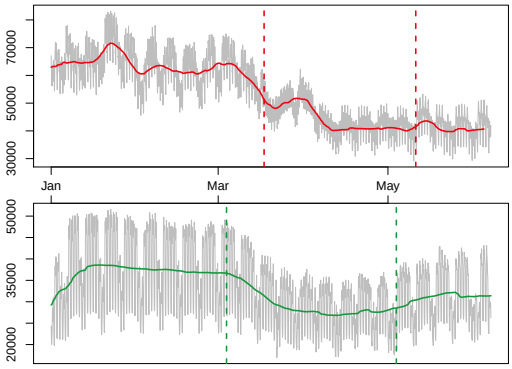


Figure 7.1 – French and Italian electricity load (in MW) at resp. half-hourly and hourly resolution in 2020. Dashed lines are the starting and ending date of the lockdown.

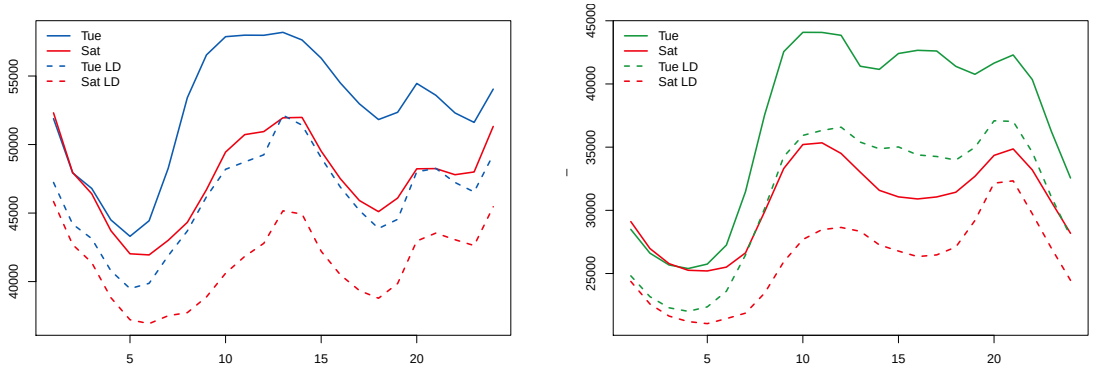


Figure 7.2 – French and Italian electricity Tuesday and Saturday load profiles before and during the lockdown (Dashed lines).

This is why finding novel approaches to better predict the load demand during these troubled times is of paramount importance. However to our knowledge, up to this date only a few papers have addressed this problem. Nagbe, 2020 is among the first to propose an efficient strategy to improve the predictions during the COVID-19 lockdown period in France. Using an adaptive functional state-space model and assimilating the period to non-workable days, the author was able to achieve significantly better performance compared to the French system operator. However, these models lack of interpretability, making other approaches preferred in the industrial context. Furthermore the aforementioned work requires to artificially set all days to weekends or holidays, which may be unviable in the long-term. Chen, Yang, and Zhang, 2020 combine the integration of mobility data with multi-task learning to improve the forecasting during the lockdown. They show that mobility is indeed a relevant feature that should be integrated in load demand models, and that joint training of a neural network for multiple geographical areas yields additional benefits and compensates for the lack of data. Nonetheless their forecasting errors remain high compared to pre-COVID standards, neural networks lack of interpretability as well and the introduction of exogenous features can be problematic in the future due to the sustainability of such data in operations.

We consider here the framework of GAM and propose two new adaptive versions of these models. The idea of adaptive models is to take advantage of data observed in an online fashion to update an initial model. This will make them able to adapt to the changes in consumption patterns spontaneously, without exogenous variable or intervention. In every adaptive forecasting method a trade-off has to be found between a good reactivity to a change (whether it is a smooth drift or a break) and a good behavior during stable periods. One of the most popular algorithm for that is the Kalman filter (Kalman and Bucy, 1961) already applied to electricity load forecasting by Harvey and Koopman, 1993 and Dordonnat, Koopman, and Ooms, 2009. We propose here to couple Kalman filters with GAM to obtain a forecasting procedure which performs well before the lockdown, exploiting the nice properties of GAM but also reacting quickly to the sudden change in the data at the beginning of the lockdown. The second approach we present leverages ideas from transfer learning to fine-tune a GAM on the lockdown period. Transfer learning (also referred as learning-to-learn or knowledge transfer) is a branch of machine learning that aims at reusing knowledge from one source task on another target one (Pan and Yang, 2010; Weiss, Khoshgoftaar, and Wang, 2016). It has shown great success, particularly when the source data is plentily available and the target one scarce. Recently it has even found applications for electricity load forecasting to transfer information from one set of customers to another one (Cai et al., 2020). In our case our source data will be the data before the lockdown and the target one the data during the lockdown in the country of interest (France in our study), or even a similar one where the lockdown came before (e.g. Italy here). The contributions of our work are the following:

1. Two mathematical approaches are proposed to efficiently adjust a historical model to consumer behavior change over time, even in the case where data is scarce. Furthermore they do not require the integration of additional features.
2. The two methodologies have been successfully applied on the difficult period of the COVID-19 lockdown in France, achieving forecast accuracy close to the one observed before the pandemic.
3. An empirical strategy is suggested to anticipate the impact of the lockdown on the load using another country's data, thus enabling satisfactory predictions from the very first day of stay-at-home order.

The rest of the paper is organized as following. In Section 7.2 we introduce the two model adaptation methods relying on Kalman filtering and fine-tuning. Section 7.3 presents the data

and the GAM model used for the French load and Section 7.4 summarizes the main results of our experiments. Finally Section 7.5 concludes our study and suggests further work.

7.2 Adaptation of Additive Models

We consider additive models whose assumption is that the response variable y_t is decomposed as

$$y_t = \beta_0 + \sum_{j=1}^{d} f_j(x_{t,j}) + \varepsilon_t \,,$$

where (ε_t) is an independent identically distributed (i.i.d.) random noise, $\boldsymbol{x}_t = (x_{t,1}, ..., x_{t,d})$ are the explanatory variables at time t, and each nonlinear effect f_j is decomposed on a spline basis $(B_{j,k})$ with coefficients $\boldsymbol{\beta}_j$:

$$f_j(x) = \sum_{k=1}^{m_j} \beta_{j,k} B_{j,k}(x) \,.$$

where m_j depends on the dimension of the spline basis. The f_j's are centered to ensure the identifiability of the model, and more details concerning the basis are given in Section 7.3.2. The coefficients $\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ then are estimated by penalized least-squares. The penalty term involves the second derivatives of the functions f_j, forcing the effects to be smooth (see Wood, 2017).

The random residuals ε_t are supposed to be Gaussian i.i.d. in the first place. Later in the numerical experiments we will introduce another variant of this model, where the residuals are supposed to be an ARIMA model optimised with classical time series methods. We focus here on structural adaptation of the GAM over time. We present two different levels of adaptation. Firstly, we consider the reduced problem of adapting a linear combination of the frozen effects $f_1, ..., f_d$. Secondly we try to adapt the whole model by fine-tuning.

7.2.1 Multiplicative Correction of the Effects

In order to reduce the dimension of the adaptation problem, a strategy is to freeze the nonlinear effects, and to correct these effects by a multiplicative factor. Precisely, we define $f(\boldsymbol{x}_t) = (1, \overline{f}_1(x_{t,1}), ..., \overline{f}_d(x_{t,d}))^\top$ where $\overline{f}_j$ is a normalized version of f_j obtained by subtracting the mean on the train set and dividing by the standard deviation. Then we adaptively estimate a vector $\boldsymbol{\theta}_t$ such that

$$\mathbb{E}[y_t \mid \boldsymbol{x}_t] = \boldsymbol{\theta}_t^\top f(\boldsymbol{x}_t) \,.$$

The estimator at time t will be denoted as $\hat{\boldsymbol{\theta}}_t$ in both Section 7.2.1 and Section 7.2.1. Thus we reduce the number of coefficients from $1 + \sum_{j=1}^{d} m_j$ to $1 + d$. This is a good trade-off to obtain a simple model which will react quickly to a break in the data generation process but also complex enough to fit well with the nonlinear properties of the load.

Exponential Least-Squares (exp-LS)

An empirical method consists in solving at each step a least-squares problem where we specify a weight decreasing exponentially with the time difference. Precisely we define

$$\hat{\boldsymbol{\theta}}_t \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{s=1}^{t-1} e^{-\mu(t-s)} \Big(y_s - \boldsymbol{\theta}^\top f(\boldsymbol{x}_s)\Big)^2 ,$$

and we predict $\hat{y}_t = \hat{\boldsymbol{\theta}}_t^\top f(\boldsymbol{x}_t)$. This formalisation leads to a single parameter, the exponential forgetting factor μ. The advantage of this type of adaptation lies in its simplicity. The forgetting factor μ is determined by minimizing the RMSE on a validation set composed of the last year of the train set for a GAM trained on the beginning of the train set, then we keep the same μ for the GAM trained on the whole train set. Previous work has been done on estimating this parameter online, but leads to computational issues and potential instability of the model (Ba et al., 2012).

Kalman Filter

We present also a state-space model approach. We assume the following equations:

$$\begin{aligned} y_t &= \boldsymbol{\theta}_t^\top f(\boldsymbol{x}_t) + \varepsilon_t \,, \\ \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \boldsymbol{\eta}_t \,, \end{aligned}$$

where (ε_t) and $(\boldsymbol{\eta}_t)$ are Gaussian white noises of respective variance / covariance σ^2 and Q. This is the setting of Kalman filtering (Kalman and Bucy, 1961), thus we use the recursive formulae of Kalman providing the expectation and covariance of the state $\boldsymbol{\theta}_t$ given the past observations, and these estimators yield the mean and variance of y_t given the past. This is described in Algorithm 7. Note that the exp-LS method has a very similar recursive form starting from t_0 such that $P_{t_0} = (\sum_{s=1}^{t_0-1} e^{-\mu(t_0-s)} f(\boldsymbol{x}_s) f(\boldsymbol{x}_s)^\top)^{-1}$ exists. Indeed, the same update rule stands for $\hat{\boldsymbol{\theta}}_t$ (with $\sigma = 1$) and the update on P_t is the following:

$$P_{t+1} = e^{\mu} \left(P_t - \frac{P_t f(\boldsymbol{x}_t) f(\boldsymbol{x}_t)^\top P_t}{f(\boldsymbol{x}_t)^\top P_t f(\boldsymbol{x}_t) + 1} \right) .$$

The simplicity stands in a single scalar parameter e^μ as multiplicative factor for the update of P_t, whereas Kalman Filter needs a matrix parameter Q added in the recursion.

There is a wide literature concerning the setting of the hyper-parameters $\hat{\boldsymbol{\theta}}_1, P_1, \sigma^2, Q$ on which the Kalman Filter crucially relies (Brockwell and Davis, 2016; Durbin and Koopman, 2012; Fahrmeir and Tutz, 2013). We refer to Chapter 5 for a detailed presentation. We observe that the iterates of $\hat{\boldsymbol{\theta}}_t$ depend only on $\hat{\boldsymbol{\theta}}_1, P_1^* = P_1/\sigma^2$ and $Q^* = Q/\sigma^2$, reducing the set of hyper-parameters as in Brockwell and Davis, 2016.

An interesting degenerate covariance matrix is the static setting $Q^* = 0$ (the state equation becomes $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$). Defining $\hat{\boldsymbol{\theta}}_1 = 0,\ P_1^* = I$, the estimate $\hat{\boldsymbol{\theta}}_t$ becomes a Ridge Forecaster:

$$\hat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \left(\sum_{s=1}^{t-1} (y_s - \boldsymbol{\theta}^\top f(\boldsymbol{x}_s))^2 + \|\boldsymbol{\theta}\|^2 \right) .$$

To obtain a dynamic setting we maximize the likelihood on the training set. The expectation-maximization algorithm is a renowned algorithm allowing to find a local optimum. However the

Algorithm 7 : Kalman Filter

Initialization: the prior $\boldsymbol{\theta}_1 \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_1, P_1)$ where $P_1 \in \mathbb{R}^{d\times d}$ is positive definite and $\hat{\boldsymbol{\theta}}_1 \in \mathbb{R}^d$.

Recursion: at each time step $t = 1, 2, \dots$

1. Prediction:

$$\mathbb{E}\left[y_t \mid (\boldsymbol{x}_s, y_s)_{s<t}, \boldsymbol{x}_t\right] = \hat{\boldsymbol{\theta}}_t^\top f(\boldsymbol{x}_t)\,,$$
$$Var\left[y_t \mid (\boldsymbol{x}_s, y_s)_{s<t}, \boldsymbol{x}_t\right] = \sigma^2 + f(\boldsymbol{x}_t)^\top P_t f(\boldsymbol{x}_t)\,.$$

2. Estimation:

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t + \frac{P_t f(\boldsymbol{x}_t)}{f(\boldsymbol{x}_t)^\top P_t f(\boldsymbol{x}_t) + \sigma^2}(y_t - \hat{\boldsymbol{\theta}}_t^\top f(\boldsymbol{x}_t))\,,$$
$$P_{t+1} = P_t - \frac{P_t f(\boldsymbol{x}_t) f(\boldsymbol{x}_t)^\top P_t}{f(\boldsymbol{x}_t)^\top P_t f(\boldsymbol{x}_t) + \sigma^2} + Q\,.$$

lack of global guarantee makes it inefficient in our case, and we applied instead some kind of grid search. Precisely we decided to set $P_1^* = I$ as in the static setting, and for a given Q^* the optimal $\hat{\boldsymbol{\theta}}_1$ for the likelihood has a closed-form solution. Q^* is of dimension 10×10 and we chose to restrict ourselves to diagonal matrices whose coefficients are in the set $\{2^j, -30 \leq j \leq 0\}$. This is still a set of around $8 \cdot 10^{14}$ elements, thus we used an iterative greedy procedure: we start from $Q^{*(0)} = 0$ and at each step, having $Q^{*(k)}$ in hand, we compute the likelihood of each matrix where only one coefficient differ from $Q^{*(k)}$, and we define $Q^{*(k+1)}$ as the one maximizing the likelihood among those tested. This algorithm yielded less than 10^4 evaluations of the likelihood.

In order to take the lockdown into account in the state-space representation, it is natural to consider a varying state noise covariance Q_t. Indeed, we expect the model to change much faster during and after the lockdown than before. It motivates a dynamic estimation of Q_t, however due to the amplitude of the crisis we modelled a break in the data at the lockdown beginning. We chose to change only the state noise covariance at the break time T, and for $t \neq T$ we use $Q_t^* = 0$ in the static setting or $Q_t^* = Q^*$ in the dynamic setting. We don't want to put any *a priori* on the break, therefore we defined $Q_T^* = P_1^* = I \gg Q^*$.

7.2.2 Correction of the Full Model

In the previous methods the nonlinear effects $f_j(\cdot)$ were frozen and adjusted with a multiplicative factor. However it may be insufficient on certain new types of behavior. Since learning a new model from scratch is inadvisable considering the few samples of target data available, we would like to start from the model trained on historical data and adapt it on the few instances available. This is a particular case of the framework of transfer learning, more specifically of model fine-tuning (FT). It consists in reusing a part of the parameters learned on the source set (typically neural network layers) and adjust them with a few gradient iterations on the target one for instance. Model fine-tuning has been successful in different fields such as computer vision (Shin et al., 2016) or even time series forecasting (Laptev, Yu, and Rajagopal, 2018).

In our case we will fine-tune the parameters of our GAM. Since it boils down to a penalized linear regression problem, it consists in fine-tuning a linear model. This framework was elaborated

by Obst et al., 2021. Starting from the coefficients $\hat{\boldsymbol{\beta}}_0$ learned on the historical source data, for each time step we perform K iterations of batch gradient descent with fixed step size α on the following objective function to yield an adjusted parameter vector $\hat{\boldsymbol{\beta}}_t$:

$$\mathcal{L}_t(\boldsymbol{\beta}) = \sum_{s=1}^{t-1} \Big(y_s - \sum_{j=1}^{d} \sum_{k=1}^{m_j} \beta_{j,k} B_{j,k}(x_{s,j}) \Big)^2$$

Let $B(\boldsymbol{x}_s)$ be the vector of the $B_{j,k}(x_{s,j})$ and $B(X_t)$ denote the matrix made by the concatenation (by row) of the $B(\boldsymbol{x}_s)$ for $s = 1, \ldots, t-1$. Again more details concerning the basis $(B_{j,k})$ is found in Section 7.3.2. As discussed by the aforementioned paper, the choice of the step size α is not crucial, as long as it is small enough. In practice a good step size is $\alpha = \alpha^*/5$ where $\alpha^* = 2/\Big(\lambda_{\max}(B(X_t)^\top B(X_t)) + \lambda_{\min}(B(X_t)^\top B(X_t))\Big)$ with $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ being respectively the maximum and minimum eigenvalue of M. Ergo the major hyper-parameter to tune is K the number of gradient iterations to perform. Theoretical methods are currently being investigated in the aforementioned paper and have been used to guide our choice here, but it was also observed empirically that K between 50 and 100 yields good results. Therefore a number of iterations in that range is always considered, and this choice usually coincides with the suggested theoretical guidelines.

7.3 Data and Model Presentation

In this section we detail the GAM model that has been used to forecast the French electricity consumption, as well as the data on which is has been applied.

7.3.1 Presentation of the Data

The French electricity consumption is freely available on the website of the system operator RTE (Réseau et Transport d'Électricité) [1]. Our dataset ranges from the 1st of January 2012 to the 7th of June 2020 with a 30 minutes temporal resolution.

As explanatory variables we obtained national averaged temperature on the website of the French weather forecaster Météo-France [2]. We took observed temperatures instead of forecasts in order to use only open data and make the results reproducible. As our goal is to compare different forecasting strategies on the same data this choice is relevant and allows a more precise comparison as we don't include in the score the uncertainty due to physical meteorological forecast.

We train the models on historical data from the beginning of 2012 to the end of August 2019. In this paper we are interested in predicting the load during and after the COVID-19 lockdown period in France. Since the consumer behavior changed abruptly during the first month and stabilized during the second one, we divide the crisis test data in two periods. The first one ranges from March 16th to April 15th and the second one from April 16th to June 7th. Note that although the lockdown officially begun Tuesday the 17th of March 2020 at midday in France, we consider March 16th as the first day of our lockdown period as the behavior had already changed. Finally, in order to assess the suitability of the offline methods and of the ones that do not model the break we consider the pre-lockdown period between September 1st 2019 and March 15th 2020.

1. `https://opendata.rte-france.com`
2. `https://donneespubliques.meteofrance.fr/`

7.3.2 The Additive Model

The time of day is crucial for load forecasting. It doesn't appear in the following definition of the additive model because we build one model for each instant of day, i.e. we treat the 48 half-hour time series independently:

$$\begin{aligned} y_t = & \sum_{i=1}^{7}\sum_{j=0}^{1} \alpha_{i,j} \mathbb{1}_{\mathrm{DayType}_t=i} \mathbb{1}_{\mathrm{DLS}_t=j} \\ & + \sum_{i=1}^{7} \beta_i \mathrm{Load1D}_t \mathbb{1}_{\mathrm{DayType}_t=i} + \gamma \mathrm{Load1W}_t \\ & + f_1(t) + f_2(\mathrm{ToY}_t) + f_3(t, \mathrm{Temp}_t) + f_4(\mathrm{Temp95}_t) \\ & + f_5(\mathrm{Temp99}_t) + f_6(\mathrm{TempMin99}_t, \mathrm{TempMax99}_t) + \varepsilon_t \,, \end{aligned} \tag{7.1}$$

where at each day t,

- y_t is the electricity load for the considered instant,
- $\mathrm{DayType}_t$ is a categorical variable indicating the type of the day of the week,
- DLS_t is a binary variable indicating whether t is in summer hour or winter hour,
- Load1D and Load1W are the load of the day before and the load of the week before,
- ToY_t is the time of year whose value grows linearly from 0 on the 1st of January 00h00 to 1 on the 31st of December 23h30,
- Temp_t is the national average temperature,
- $\mathrm{Temp95}_t$ and $\mathrm{Temp99}_t$ are exponentially smoothed temperatures of factor $\alpha = 0.95$ and 0.99. E.g. for $\alpha = 0.95$ at a given instant i, $\mathrm{Temp95}_i = \alpha \mathrm{Temp95}_{i-1} + (1-\alpha)\mathrm{Temp}_i$,
- $\mathrm{TempMin99}_t$ and $\mathrm{TempMax99}_t$ are the minimal and maximal value of $\mathrm{Temp99}_t$ at the current day.

The models are trained in `R` using the library `mgcv` (Wood, 2015). We use the default thin plate spline basis to represent the f_j's, except for the time of year effect f_2 for which we choose cyclic cubic splines (see Wood, 2017 for a complete description of spline basis). The dimensions of the bases are usually below 5, excluding f_2 which uses a basis of dimension 20.

As previously mentioned in Section 7.2, we suppose that ε_t is a Gaussian noise with 0 mean and constant variance. However this hypothesis is rarely true in practice and we observe an auto-correlation structure in the error. We thus propose to model it with an ARIMA model by selecting the best model with AIC criteria (Akaike, 1978) in the family of ARIMA(p,d,q) where $p, q \leqslant 100$ and $d \leqslant 1$ (we use the `R` function `auto.arima` of R. Hyndman). In that case the forecast are performed adding GAM forecasts and the short term correction of the ARIMA models exploiting recent observations.

7.3.3 Knowledge Transfer from Italy

Italy was the first country to be massively affected by the novel coronavirus in Europe. The Italian government decreed a total lockdown from the 9th of March 2020, hence 7 days before the French one. Since GAM models for both countries usually exhibit similar behavior (see Figure 7.3 for a comparison of residuals) and indices such as the Oxford COVID-19 Government Response Tracker (Hale et al., 2021) show that both countries took comparable measures during the lockdown, our idea is to use this one week head-start and to adjust our GAM model for France accordingly to the changes observed in Italy. We have at our disposal data from the

Italian system operator Terna [3] and meteorological data gathered through the `R` package `Riem` available from the 1st of January 2015 to the 28th of June 2020 with a 1 hour temporal resolution. For each instant, a model similar to (7.1) is constructed on the data on the range 2015-2019, with the main differences being that the effects $f_3(\cdot)$ and $f_6(\cdot)$ are removed, and that $f_2(\cdot)$ is replaced by a sum of 7 effects, one for each day of the week. Then the same procedure as described in Section 7.2.2 is applied. Let $\hat{\boldsymbol{\beta}}_0^{IT}$ be the coefficient learned on the Italian source data, and $\hat{\boldsymbol{\beta}}_t^{IT}$ be the coefficients obtained by performing the aforementioned fine-tuning on Italian data ranging from the 28th of February up to date t (typically t could correspond to the 15th of March, the day before the stay-at-home order begun in France). We thus obtain $\hat{\boldsymbol{\delta}}_t = \hat{\boldsymbol{\beta}}_t^{IT} - \hat{\boldsymbol{\beta}}_0^{IT}$ the adjustment of the model on the beginning of the lockdown period. We then use $\hat{\boldsymbol{\beta}}_t^{FR} = \hat{\boldsymbol{\beta}}_0^{FR} + \rho\,\hat{\boldsymbol{\delta}}_t$ to perform the predictions for France, where ρ is a scale parameter accounting for the difference of load levels between the two countries. We refer to this model as GAM-δ. Since the ToY effect is modelized differently for the Italian model (one function per day of the week), we will not adjust the corresponding coefficients in the French model. This is further justified by the fact that in general the ToY effect is very specific to a country, and it should be learned on a whole year at least. As for the choice of ρ, making the assumption that the consumption in France and Italy are proportional with a factor ρ allows us to use the simple estimate $\hat{\rho} = \sum_t y_t^{FR} / \sum_t y_t^{IT}$ summed over a year for instance. The advantage of GAM-δ is that it can be applied to reduce the prediction error starting at the very first day of lockdown. One can afterward combine this procedure with fine-tuning on the eventually available French data. The procedures for both regular fine-tuning and GAM-δ are summarized in Algorithm 8.

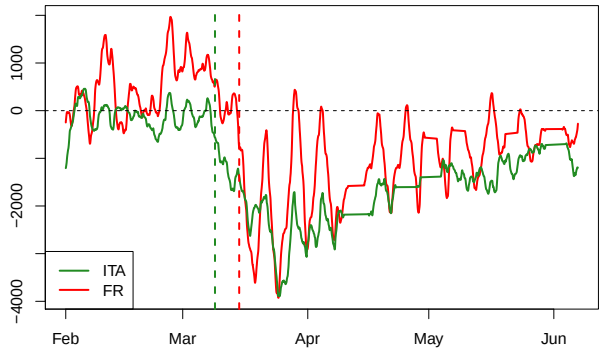


Figure 7.3 – Comparison of the smoothed residuals of the French and Italian GAMs in 2020. The dashed lines represent the start of the respective lockdowns.

7.4 Experiments

We present the application of our methods to the French dataset. While accuracy metrics are of paramount importance, we also focus on the interpretation of our results and on model behavior.

3. `https://www.terna.it`

Algorithm 8 : Transfer learning at time step t

Inputs: Step size α, number of iterations K, French and Italian historical source parameters $\hat{\boldsymbol{\beta}}_0^{FR}, \hat{\boldsymbol{\beta}}_0^{IT}$, scale parameter ρ.

If GAM fine-tuned:

1. Initialize $\hat{\boldsymbol{\beta}}_t^{FR} \leftarrow \hat{\boldsymbol{\beta}}_0^{FR}$.
2. Repeat K times:
   $$\hat{\boldsymbol{\beta}}_t^{FR} \leftarrow \hat{\boldsymbol{\beta}}_t^{FR} - \alpha \nabla \mathcal{L}_{t-1}^{FR}(\hat{\boldsymbol{\beta}}_t^{FR}).$$
3. Predict $\hat{y}_t = \hat{\boldsymbol{\beta}}_t^{FR\top} B(\boldsymbol{x}_t)$.

If GAM-δ:

1. Initialize $\hat{\boldsymbol{\beta}}_t^{IT} \leftarrow \hat{\boldsymbol{\beta}}_0^{IT}$.
2. Repeat K times:
   $$\hat{\boldsymbol{\beta}}_t^{IT} \leftarrow \hat{\boldsymbol{\beta}}_t^{IT} - \alpha \nabla \mathcal{L}_{t-1}^{IT}(\hat{\boldsymbol{\beta}}_t^{IT}).$$
3. Set $\hat{\boldsymbol{\delta}}_t = \hat{\boldsymbol{\beta}}_t^{IT} - \hat{\boldsymbol{\beta}}_0^{IT}$, $\hat{\boldsymbol{\beta}}_t^{FR} = \hat{\boldsymbol{\beta}}_0^{FR} + \rho\, \hat{\boldsymbol{\delta}}_t$.
4. Predict $\hat{y}_t = \hat{\boldsymbol{\beta}}_t^{FR\top} B(\boldsymbol{x}_t)$.

If GAM-δ fine-tuned:

1. Perform steps 1) to 3) of **GAM-δ**, obtaining $\hat{\boldsymbol{\beta}}_t^{FR} = \hat{\boldsymbol{\beta}}_0^{FR} + \rho\, \hat{\boldsymbol{\delta}}_t$.
2. Repeat K times:
   $$\hat{\boldsymbol{\beta}}_t^{FR} \leftarrow \hat{\boldsymbol{\beta}}_t^{FR} - \alpha \nabla \mathcal{L}_{t-1}^{FR}(\hat{\boldsymbol{\beta}}_t^{FR}).$$
3. Predict $\hat{y}_t = \hat{\boldsymbol{\beta}}_t^{FR\top} B(\boldsymbol{x}_t)$.

7.4.1 Model Dynamics

The moving average of the error of the different models are represented in Figure 7.4. At the beginning of the lockdown all the models will tend to overpredict the load. However most of our adaptive methods quickly accommodate to the lower demand and progressively reduce their bias, notably Kalman with Dynamic Break and GAM fine-tuned. On the contrary regular GAM does not succeed in reducing the error (even with the help of an ARIMA) as it keeps overpredicting the demand. GAM-δ on the other hand is very good during the first couple of days, efficiently taking advantage of the change in patterns observed in Italy. However it quickly drifts away over time because the Italian consumption recovers faster than the French one during the second month of lockdown (see Figure 7.1). However since the objective of GAM-δ is to provide an initial boost of performance during the first couple of weeks while the other models adjust, this is only a minor issue (see Section 7.4.2).

We test the Kalman filter in a static and a dynamic setting as described in Section 7.2.1. For both we assess the introduction of a break state noise covariance matrix at lockdown. The evolution of the state estimate $\hat{\boldsymbol{\theta}}_t$ is displayed in Figure 7.5 for different settings. In the static setting the Kalman filter optimizes a state which is assumed to be constant, hence explaining a slow evolution compared to the faster changes of the dynamic one. However both variants change faster during lockdown than they did before. As expected the introduction of a break covariance matrix at the beginning of the lockdown allows the model to adapt much faster.

The model dynamics can be analysed for the fine-tuning too. For GAM-δ the only coefficients of $\hat{\boldsymbol{\delta}}_t$ with a significant evolution after fine-tuning are the ones pertaining to the lagged load (γ for

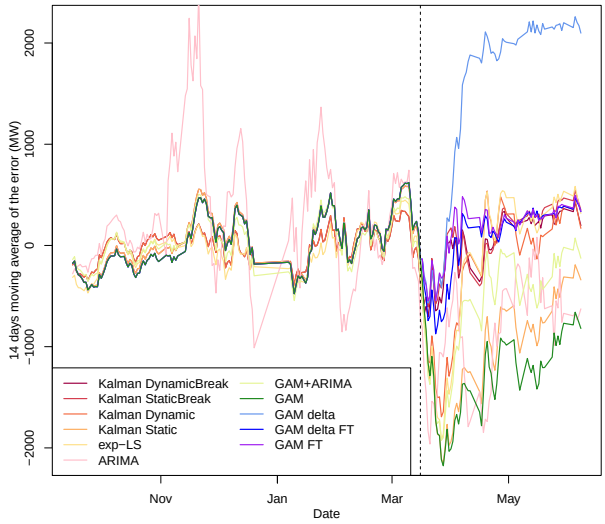
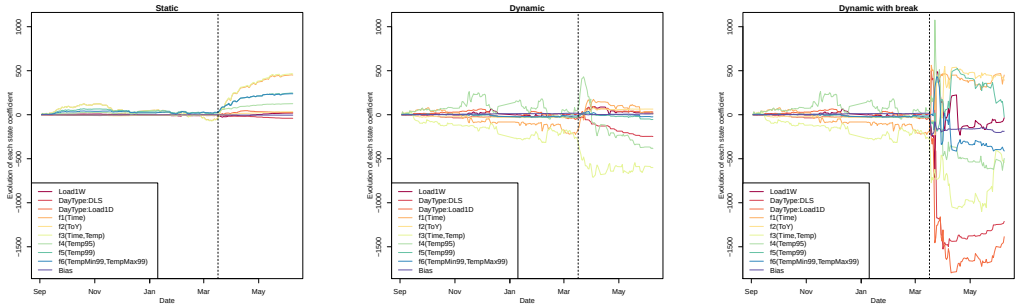


Figure 7.4 – Moving average of the error of the different models at 8-8:30 PM.



(a) Static (b) Dynamic (c) Dynamic with break

Figure 7.5 – Evolution of the state coefficients for different Kalman variants at 8-8:30 PM (subtracting the coefficients on September 1st 2019).

Load1W and $\beta_i, i = 1..7$ for Load1D) and have been represented in Figure 7.6. The other ones are zero and have been omitted for clarity. The coefficients of the working days drop, especially the Monday, whereas the ones of the weekend increase, notably Saturday. It can be interpreted as follows: the historical model learned a certain transition between the different days of the week. With the lockdown now all the days are similar and close to a Saturday, which has a lower demand than Monday and thus the associated coefficient plummets. The coefficient of Saturday soars because the demand on Fridays is now much lower than it used to be and that daily profiles are similar. Finally since during the first weeks the electricity demand progressively decreases (see Figure 7.1) the coefficient of γ drops as well.

7.4.2 Aggregation

We proposed 2 load forecasting models (ARIMA, GAM) and different variants to adapt them to the lockdown period (exp-LS, Kalman adaptation, transfer learning) leading to 11 candidates

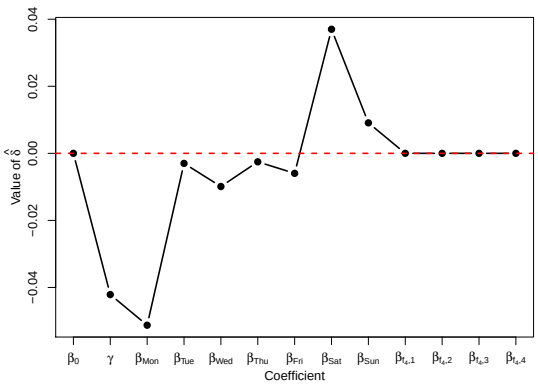


Figure 7.6 – Value of $\hat{\boldsymbol{\delta}}_t$ fine-tuned on the period 16/03-15/04 at 8-8:30 PM.

which we call experts in the following. A natural approach is then to aggregate them in a single forecast which will take benefit of the best one in function of time. This is the main idea behind online aggregation methods which have already demonstrated their benefits in the field of electricity load forecasting (Gaillard and Goude, 2015; Goehry et al., 2019). Since Figure 7.2 shows the convergence of the daily profiles towards the Saturday shape, this as well as Nagbe, 2020 motivates adding another expert named GAM Saturday, whose prediction is made by the regular GAM as if every day was a Saturday.

We recall briefly the main principles of the online aggregation approach and refer the interested reader to Cesa-Bianchi and Lugosi, 2006 for a complete presentation. A bounded sequence of observations (here half-hourly French electricity consumption) $y_1, \dots, y_n \in [0, B]$ is observed (B being an unknown constant). We have access to a set of N experts producing forecasts of the sequence at each instant t based on past values of y. After that, aggregation is computed step by step: $\hat{y}_t = \sum_{j=1}^{N} \hat{p}_{j,t} \hat{y}_t^j$ where the weights are updated according to past performances of each experts which are measured with a convex loss function. In accordance to the RMSE criterion used in our case study we consider the square loss $\ell_t(x) = (y_t - x)^2$. At time t expert e suffers loss $\ell_t(\hat{y}_t^e) = (y_t - \hat{y}_t^e)^2$ and the aggregation $\ell_t(\hat{y}_t) = (y_t - \hat{y}_t)^2$. We call Oracle an optimal forecast which is unknown in advance and usually hard to beat in terms of forecasting accuracy (Cesa-Bianchi and Lugosi, 2006). We denote it by $\hat{y}_t^*$. For example, it could be the best fixed convex aggregation or the best expert (best w.r.t the entire time interval performance, of course unknown a priori). The goal of aggregation algorithms is to minimise the total loss $\sum_{t=1}^{T}(y_t - \hat{y}_t)^2$ that can be expressed:

$$\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 \quad \triangleq \quad \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t^*)^2 + R_T\,,$$

where R_T is the so-called regret term, it is the error suffered by our algorithm relatively to the error of the oracle (Cesa-Bianchi and Lugosi, 2006). The aim is thus to propose algorithms that, regarding competitive oracles, achieve low regrets. In our study we use the ML-Poly algorithm of Gaillard, Stoltz, and Van Erven, 2014, implemented in the `R` package `opera` (Gaillard and Goude, 2016) and already used successfully for load and price forecasting (Gaillard and Goude, 2015; Gaillard, Goude, and Nedellec, 2016). It is described in Algorithm 9. An expert who has a high regret, which means that he suffers a higher loss than the aggregation, will receive less weight for the next round. The time-varying learning rate $\eta_{e,t}$ could be seen as a vector step size

Algorithm 9 : ML-Poly

Initialization: $\hat{\boldsymbol{p}}_1 = (1/N, \dots, 1/N)$ and $\boldsymbol{R}_0 = (0, \dots, 0)$
Recursion: at each time step $t = 1, 2, \dots$
- Pick the learning rates: $\eta_{e,t-1} = 1/\left(1 + \sum_{s=1}^{t-1} \big(\ell_s(\hat{y}_s) - \ell_s(y_s^e)\big)^2\right)$.
- Compute the weights $\hat{\boldsymbol{p}}_t$: $\hat{p}_{e,t} = \eta_{e,t-1} \left(R_{e,t-1}\right)_+ \ / \ \boldsymbol{\eta}_{t-1} \cdot \left(\boldsymbol{R}_{t-1}\right)_+$ where $\boldsymbol{R}_+$ is the non-negative parts of $\boldsymbol{R}$.
- Output prediction $\hat{y}_t = \sum_{e=1}^{N} \hat{p}_{e,t} \hat{y}_t^e$.
- For each expert e update the regret: $R_{e,t} = R_{e,t-1} + \ell_t(\hat{y}_t) - \ell_t(y_t^e)$, $\boldsymbol{R}_t = (R_{1,t}, \dots, R_{N,t})$.

parameter of gradient descent varying with time so that no parameter tuning is needed.

Finally a few experts are introduced in the aggregation only at lockdown. Indeed, before lockdown the transfer learning experts don't make sense (there is no target data), the Kalman experts modelling the break coincide with the other ones, and the expert GAM Saturday was only introduced for the lockdown period. These specialized experts are added to the aggregation at the lockdown period with a uniform weight (1/12), and the experts present before share the rest of the weight proportionally to their previous weight (Devaine et al., 2013).

The evolution of the weights of the experts over time is displayed in Figure 7.7. It gives insight on which predictions are the most useful in the aggregation at a given time. The lockdown acts as a break and causes a significant shift in the weights distribution. As such, GAM Saturday immediately takes a large weight: this is due to the aforementioned resemblance between the daily profiles during the lockdown with Saturdays. Moreover, this expert predicts a lower consumption than reality, compensating for the overestimation of the other experts at the beginning of the lockdown. GAM-δ also has high importance, as it has knowledge of what happened in Italy and thus suits the new patterns of load demand in France. For instance on the two first days of lockdown (16 and 17$^{\text{th}}$ of March) GAM-δ yields 1984 MW of RMSE, compared to 2674 and 3005 for Kalman Dynamic Break and regular GAM respectively. However their importance dwindle with time as the adaptive Kalman and fine-tuning methods have seen enough data and have become more competitive.

7.4.3 Numerical Results

As usual in electricity load forecasting, the performance metrics are the root mean squared error (in MW) and the mean absolute percentage error (in %):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \big(y_t - \hat{y}_t\big)^2}, \quad \text{MAPE} = \frac{100}{n}\sum_{t=1}^{n} \Big|\frac{y_t - \hat{y}_t}{y_t}\Big|,$$

where n is the number of instances in the test set.

We display the numerical performance of our methods in Table 7.1. We observe that any of our methods have lower RMSE or MAPE than GAM + ARIMA on both post-COVID test sets. As expected, the Kalman Dynamic with break yields the best results for the two error metrics during the first part of the lockdown period but the fine-tuned methods are very close to it. Similarly, the two break approaches are the best ones after the lockdown. The additional benefits brought by expert aggregation is emphasized by the two last rows. The algorithm manages to take advantage of the individual specificities of the different predictions, leading to

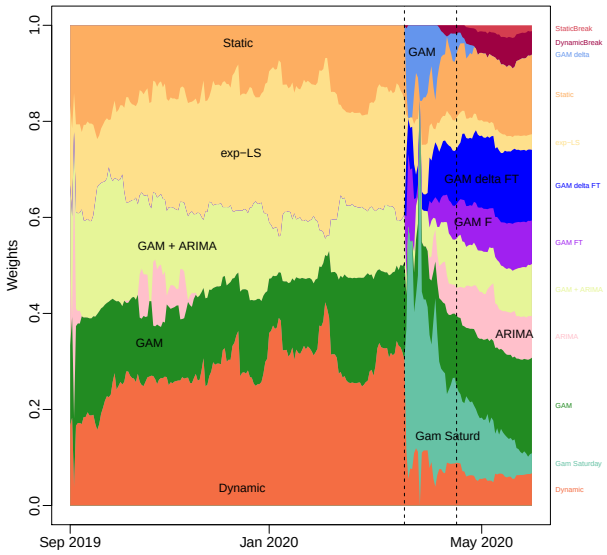


Figure 7.7 – Weights attributed to each expert by the aggregation method at 8-8:30 PM. Dashed lines split the test sets.

further error reduction. While individually poor, the inclusion of GAM Saturday in the mixture brings significant improvement for the first testing period (see the end of Section 7.4.2) because it compensates for the overestimation of the demand at the beginning of the lockdown.

The significativity of our results was assessed with two statistical tests: a Diebold-Mariano (DM) test (Diebold and Mariano, 2002) and a Wilcoxon test as proposed by Zhang, Ding, and Sun, 2020, both on the absolute error. For two methods A and B it allows to test the null hypothesis that method B outperforms or is equivalent to method A, against the alternative hypothesis that method A outperforms method B. In Table 7.2 we display the results of the tests for the most relevant forecasting models at the significance level 0.01. At each row i and column j we display the p-values of Wilcoxon test in blue and of the Diebold-Mariano test in purple, the alternative hypothesis is "method i outperforms j". We use the symbol ε when the p-value is below 0.01 and otherwise we give a 0.01 approximation. For clarity we consider only the best non-adaptive method and selected adaptive ones, and we order them according to the performance on the last test set. These tests confirm that on both post-COVID test sets, the improvement brought by our adaptive procedures on an ARIMA correction of the GAM is statistically significant, and so is the improvement of the aggregation compared to any of our method. Results coincide for the two tests, ergo consolidating our results even further.

7.5 Conclusion

In this paper, we proposed two novel approaches of adaptive generalized additive models to improve load forecast during the COVID-19 pandemic, one relying on Kalman filtering and the other on transfer learning with GAM fine-tuning. We showed that Kalman filtering approaches can be significantly improved by re-initializing the online update at the beginning of the lockdown period (Break approach). This helps the Kalman filter to adapt quickly to a change in the

Method	2019/09/01 - 2020/03/15	2020/03/16 - 2020/04/15	2020/04/16 - 2020/06/07
ARIMA	4.10 %, 3341 MW	5.44 %, 3248 MW	5.59 %, 3135 MW
GAM	1.39 %, 1085 MW	4.83 %, 2961 MW	3.12 %, 1753 MW
GAM + ARIMA	1.34 %, 1050 MW	4.28 %, 2654 MW	2.65 %, 1464 MW
exp-LS	1.26 %, 982 MW	3.94 %, 2521 MW	1.97 %, 1029 MW
Kalman Static	1.38 %, 1077 MW	4.81 %, 2923 MW	2.85 %, 1588 MW
Kalman Static Break	-	2.79 %, 1954 MW	1.59 %, 855 MW
Kalman Dynamic	**1.26 %, 979 MW**	3.66 %, 2351 MW	1.89 %, 1002 MW
Kalman Dynamic Break	-	**2.73 %, 1902 MW**	**1.62 %, 854 MW**
Fine-tuned	-	2.78 %, 1917 MW	1.80 %, 938 MW
GAM δ	-	4.11 %, 2364 MW	6.09 %, 2713 MW
GAM δ - Fine-tuned	-	2.81%, 1912 MW	1.72 %, 905 MW
GAM Saturday	8.33 %, 6425 MW	6.09 %, 3970 MW	8.40 %, 4616 MW
Aggregation without GAM Saturday	1.28 %, 1005 MW	3.01 %, 2014 MW	**1.44 %, 745 MW**
Aggregation with GAM Saturday	-	**2.54 %, 1636 MW**	**1.49 %, 766 MW**

Table 7.1 – Numerical performance in MAPE (%) and RMSE (MW).

2020/03/16-2020/04/15	1	2	3	4	5	6
1. Aggregation	**1** **1**	**0.44** ε	ε ε	ε ε	ε ε	ε ε
2. Kalman Dynamic Break	**0.56** **1**	**1** **1**	**0.12** **0.06**	ε ε	ε ε	ε ε
3. GAM δ Fine-Tuned	**1** **1**	**0.88** **0.94**	**1** **1**	ε ε	ε ε	ε ε
4. exp-LS	**1** **1**	**1** **1**	**1** **1**	**1** **1**	ε ε	**0.08** **0.28**
5. GAM + ARIMA	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**0.99** **1**
6. GAM δ	**1** **1**	**1** **1**	**1** **1**	**0.92** **0.72**	**0.01** ε	**1** **1**
2020/04/16-2020/06/07	1	2	3	4	5	6
1. Aggregation	**1** **1**	ε ε	ε ε	ε ε	ε ε	ε ε
2. Kalman Dynamic Break	**1** **1**	**1** **1**	ε ε	ε ε	ε ε	ε ε
3. Gam δ Fine-Tuned	**1** **1**	**1** **1**	**1** **1**	ε ε	ε ε	ε ε
4. exp-LS	**1** **1**	**1** **1**	**1** **1**	**1** **1**	ε ε	ε ε
5. GAM + ARIMA	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	ε ε
6. GAM δ	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**

Table 7.2 – Wilcoxon test and Diebold-Mariano test on the absolute error on the last two test sets. ε for p-value below 0.01.

data and update the forecasts taking advantage of recent observations. Transfer learning was successfully adapted to this problem in two ways: we fine-tuned a GAM learned before the COVID-19 crisis on the lockdown period, and we transferred information from Italian data to French data. We illustrated the benefits of the transfer from Italy at the beginning of the lockdown, as well as the efficiency of adaptive methods to significantly improve predictions, all without relying on the inclusion of new exogenous features. As all these new approaches have time-varying performances (the best forecasts vary with time), we proposed to use online expert aggregation to enhance results even further.

While in this paper we focused on adapting GAM, the proposed framework can be applied to other approaches. The use of neural networks, for instance, will soon be investigated. We also plan to include exogenous information such as mobility data proposed by Chen, Yang, and Zhang, 2020, macro-economic indicators, or data from social media such as Twitter. Regarding load data, exploiting regional data could be relevant as the propagation of the pandemic and its impact on consumption was different depending on the region in France and Italy. The inclusion of more countries could be helpful as well. For these next steps, transfer approaches will be of fundamental importance but also adaptive ones, as the effects of exogenous variables are likely to vary with time or even be added at some point.

Chapter 8

Competition Day-Ahead Electricity Load Forecasting: Post-Covid Paradigm

We present the winning strategy for the IEEE DataPort Competition on Day-Ahead Electricity Load Forecasting: Post-Covid Paradigm. This competition was organized to design new forecasting methods for unstable periods, such as the one starting in Spring 2020. First, we pre-process the data with a statistical correction of the meteorological variables. Second, we apply standard statistical and machine learning models. Third, we rely on state-space models to adapt the aforementioned forecasters. It achieves the right compromise between two extremes. Indeed, machine learning methods allow to learn complex dependence on explanatory variables on a historical data set but fail to forecast non-stationary data accurately. Conversely, purely time-series models such as autoregressive are adaptive in essence but fail to capture dependence on exogenous variables. Finally, we use aggregation of experts, and we leverage the diversity of the set of obtained forecasters to improve our final predictions. The evaluation period of the competition was the occasion of trial and error, and we put the focus on the final procedure.

This chapter is based on a joint work with Yannig Goude published in IEEE Open Access Journal of Power and Energy.

Contents

8.1 Introduction

Electricity demand forecasting is a crucial task for grid operators. Indeed the production must balance the consumption as storage capacities are still negligible compared to the load. Time series methods have been applied to address that problem, relying on calendar information and lags of the electricity consumption. Statistical and machine learning models have been designed to use exogenous information such as meteorological forecasts (the load usually depends on the temperature, for instance, due to electric heating and cooling).

The field has been thoroughly studied over the past decades, as shown in the bibliometric review of Yang et al., 2021. We will not propose here an exhaustive bibliographic study and refer to the recent surveys (Hong and Fan, 2016; Almalaq and Edwards, 2017; Nti et al., 2020). We focus on recent results in the different forecasting challenges related to this field. The Global Energy Forecasting Competitions (GEFCOM) (Hong, Pinson, and Fan, 2014; Hong et al., 2016; Hong, Xie, and Black, 2019) provide a large benchmark of popular and efficient load forecasting methods. Black box machine learning models such as gradient boosting machines (Lloyd, 2014) and neural networks (Ryu, Noh, and Kim, 2017; Dimoulkas, Mazidi, and Herre, 2019) rank among the first as well as statistical models like generalized additive models (GAM) (Nedellec, Cugliari, and Goude, 2014; Dordonnat, Pichavant, and Pierrot, 2016) or parametric regression models (Charlton and Singleton, 2014; Ziel, 2019). Ensemble methods or expert aggregation are also a common practice for competitors (Gaillard, Goude, and Nedellec, 2016; Smyl and Hua, 2019).

The consumption behavior changed abruptly during the coronavirus crisis, especially during lockdowns imposed by many governments. These changes in consumption mode have been challenging for electricity grid operators as historical forecasting procedures performed poorly. Therefore, designing new forecasting strategies to take that evolution into account is important to reduce the cost of forecasting errors and ensure the stability of the network in the future.

It is to be noted that purely time series methods like autoregressive didn't drift as they are very adaptive in essence. However, they fail to capture the dependence of the load on, for instance, meteorological variables. We claim that state-space models allow the best of both worlds. First, machine learning models trained on historical data are used to design new feature representations. Second, a state-space representation yields a methodology to adapt these complex forecasting models.

Our work extends a previous study on the French electricity load (Obst, Vilmarest, and Goude, 2021) where a state-space approach was presented to adapt generalized additive models in the context of online learning. The idea is to plug a Kalman filter on the estimated effects of a GAM to gain in online reactivity. The novelty of this article lies both in the forecasting method and in the application. First, besides generalized additive models, we extend our procedure to other widely used machine learning models, including neural networks. Second, after applying the standard Kalman filter (Kalman and Bucy, 1961), we apply another state-space approach named Viking (Vilmarest, Goude, and Wintenberger, 2021). Viking is a generalization of Kalman filter allowing to estimate jointly the state and the variances in a state-space model. Third, our procedure resulted in the winning strategy in a competition on post-covid day-ahead electricity demand forecasting (Farrokhabadi, 2020), motivating the efficiency of the proposed approach.

A diagram of our forecasting strategy is provided in Figure 8.1. The article follows its

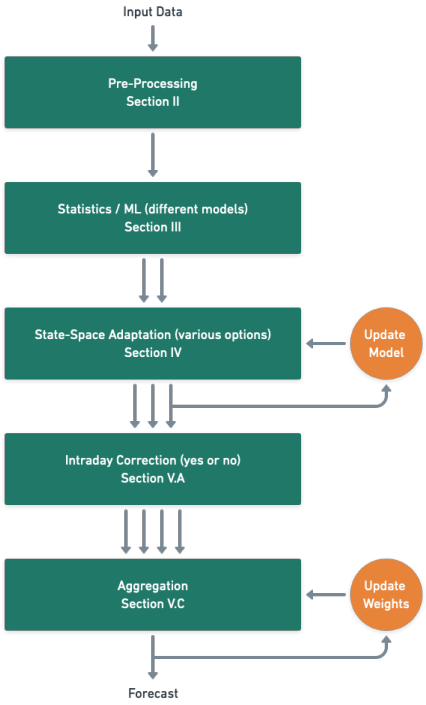


Figure 8.1 – Diagram of the forecasting strategy. The multiple arrows mean multiple outputs. Precisely, the pre-processing step yields one unique output. Then we have several classical forecasting methods, each of which has different adaptation flavors. The intraday correction doubles the number of forecasts, and all of them are combined in the aggregation step, yielding the final forecast.

structure. Section 8.2 is an introduction to the data set, and we detail our pre-processing with a focus on meteorological variables. In Section 8.3 we present standard forecasting models. The core of our strategy is Section 8.4 where we propose a generic state-space framework to adapt these methods. We discuss the numerical performances of the various models in Section 8.5, and we combine them through aggregation of experts to leverage each model's advantages.

8.2 Data Presentation and Pre-Processing

The objective of the competition was to predict the electricity load of an undisclosed location of average consumption 1.1 GW, that is of the order of one million people in western countries. The break in the electricity demand in March 2020 is clear in Figure 8.2. The objective of the competition was to design new strategies for day-ahead forecasting in order to be robust to this unstable period.

8.2.1 Time segmentation

The competition's setting was to forecast the hourly load 16 to 40 hours ahead in an online manner. Precisely, we had to predict the consumption of each hour of day d with data up to

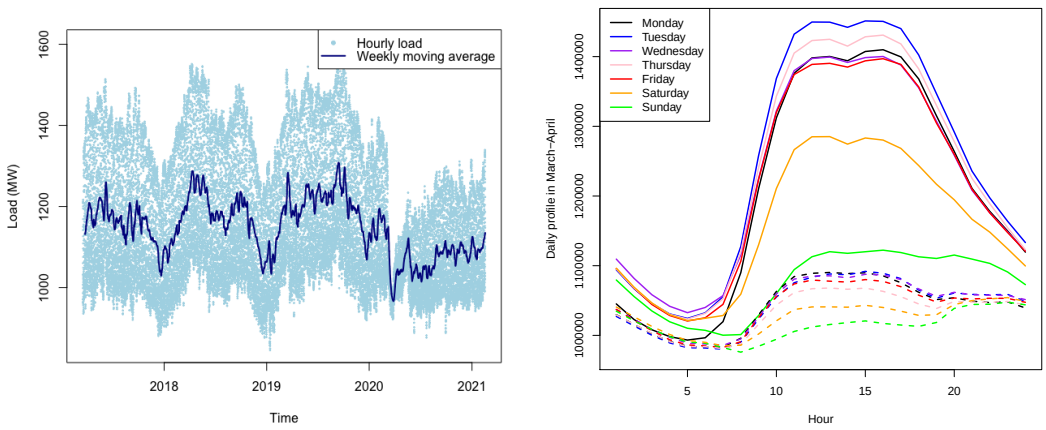


Figure 8.2 – On top: electricity load from March 18$^{\text{th}}$ 2017 to February 16$^{\text{th}}$ 2021. On the bottom: daily profiles of the electricity load in March-April 2019 (solid lines) compared to March-April 2020 (dashed lines).

8 AM day d-1. After our prediction was sent, a new batch of data up to 8 AM day d was released so that we had to predict day d+1 ...

The evaluation was based on the Mean absolute error on the period ranging from January 18$^{\text{th}}$ to February 16$^{\text{th}}$ 2021. To build a forecasting model, the historical load starting from March 18$^{\text{th}}$ 2017 was provided, as well as meteorological forecasts and realizations during the same period.

8.2.2 Meteorological Forecasts

Aside from calendar variables, it is usual that the most important exogenous factor explaining the electricity demand is meteorology. The dependence of the load on the temperature, for example, is due to electric heating and cooling. Moreover, the dependence of the electricity demand on meteorology is augmented by the development of decentralized renewables. Indeed, small renewable production is often used by its owner, yielding a net consumption that highly depends on wind or solar radiation. Therefore the error of a forecasting model for the electricity demand crucially depends on the performance of the meteorological forecasts.

The competition data include forecasts and realization of the temperature, the cloud cover, the pressure, the wind direction and speed. These forecasts are assumed to be known 48 hours in advance and invariant after. Thus they can be used to forecast the load at the 16 to 40 hours horizon.

However, from the statistical properties of the meteorological forecasting residuals (c.f. Figure 8.3), we conjecture that the forecasts come from physical models that need to be statistically corrected. Indeed, as the forecasts are available 48 hours in advance, if a statistical correction had been applied, then auto-correlations of the residuals over 48 hours would be negligible. We thus use correction models close to autoregressives on the residuals.

Formally, let (z_t) be any of the meteorological variable and $(\hat{z}_t)$ the forecast given in the data

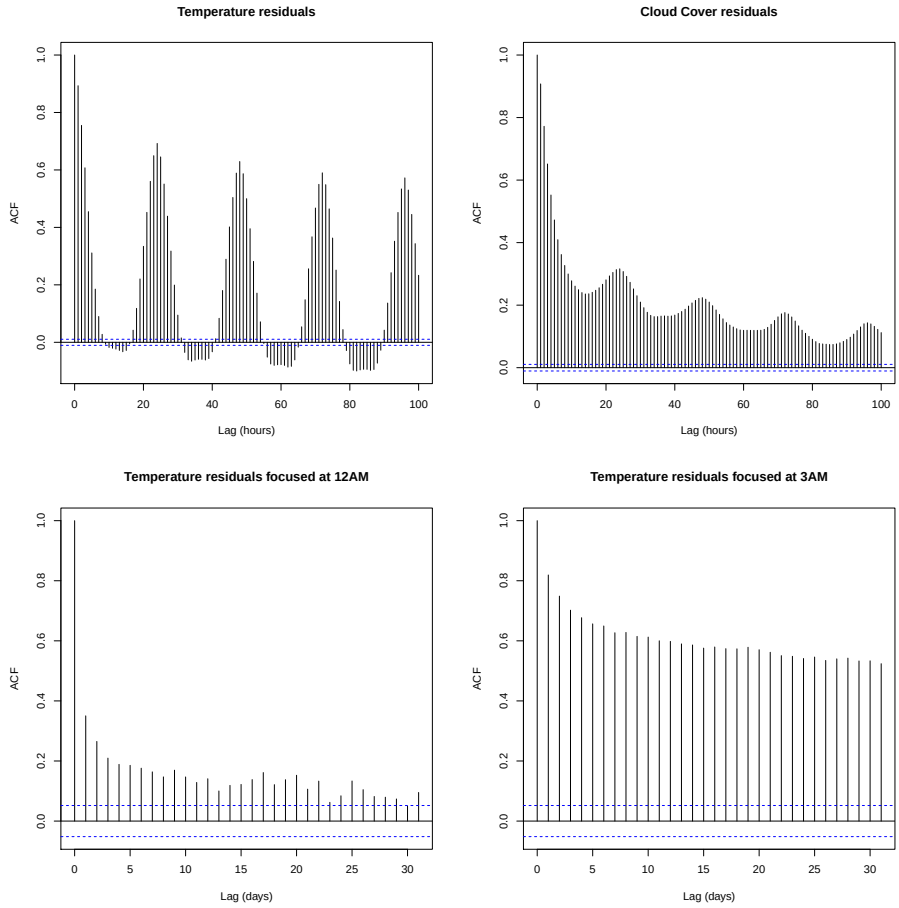


Figure 8.3 – On top: auto-correlation plots of the temperature (left) and cloud cover (right) forecasting residuals, with lags in hours. On the bottom: auto-correlation plots of the temperature residuals focused on a specific hour of the day (midnight on the left, 3 AM on the right), with lags in days.

	Initial	Last daily lag	Corrected
Temperature (°C)	3.00	2.11	1.69
Cloud cover (%)	17.28	18.74	14.99
Pressure (kPa)	0.506	1.30	0.423
Wind Speed (km/h)	4.53	3.49	2.53

Table 8.1 – Mean absolute error of different meteorological forecasts. The first column is the forecast given in the data set. The second one consists in using the variable of interest with a 24 or 48-hour delay. The last is our corrected forecast. We evaluate through the mean absolute error during 2020 while we train the corrections on the data before 2020.

set. Then we use the model

$$z_t = \alpha \hat{z}_t + \sum_{l \in \mathcal{L}_{p,P,h(t)}} \beta_l (z_{t-l} - \hat{z}_{t-l}) + \gamma z_{t-l_0(t)} + \delta + \varepsilon_t \,,$$

where $h(t) \in \{0, \dots, 23\}$ is the hour of the day of time t and

$$l_0(t) = \begin{cases} 24 & \text{if } h \leq 7 \,, \\ 48 & \text{if } h > 7 \,, \end{cases}$$

$$\mathcal{L}_{p,P,h} = \begin{cases} \{24, \dots, 24 * P, h + 17, \dots, h + 16 + p\} & \text{if } h \leq 7 \,, \\ \{48, \dots, 24 * (P+1), h + 17, \dots, h + 16 + p\} & \text{if } h > 7 \,. \end{cases}$$

In other words, we forecast the residual of the variable of interest with a linear model on
— the last P available daily lags of the residual,
— the last p available lags of the residual (up to 7 AM of the previous day),
— the forecast,
— the last daily lag of the variable of interest.

We optimize the coefficients separately for each hour of the day for the temperature, whereas we use the same coefficients at each hour of the day for the cloud cover, pressure, and wind speed (except the intercept term). We don't correct the wind direction. The parameters p and P are selected based on BIC. We display in Table 8.1 the error of the initial forecast, compared to simply using the last daily lag of the variable of interest, and our corrected forecast.

8.3 Time-Invariant Experts

We summarize the explanatory variables used in our forecasting models:
— calendar variables: the hour of the day, the day of the week, the time of year (Toy) growing linearly from 0 on January 1st to 1 on December 31st, and a variable growing linearly with time to account for a trend,
— meteorological forecasts after statistical correction: the temperature along with exponential smoothing variants of parameters 0.95 and 0.99 (respectively $Temps95$ and $Temps99$), the cloud cover, the pressure, the wind direction and speed,
— lags of the electricity load: the load a week ago $LoadW$ and the last load available $LoadD$ (a day ago for the forecast before 8 AM and two days ago after 8 AM, this constraint coming from the availability of the online data during the competition).

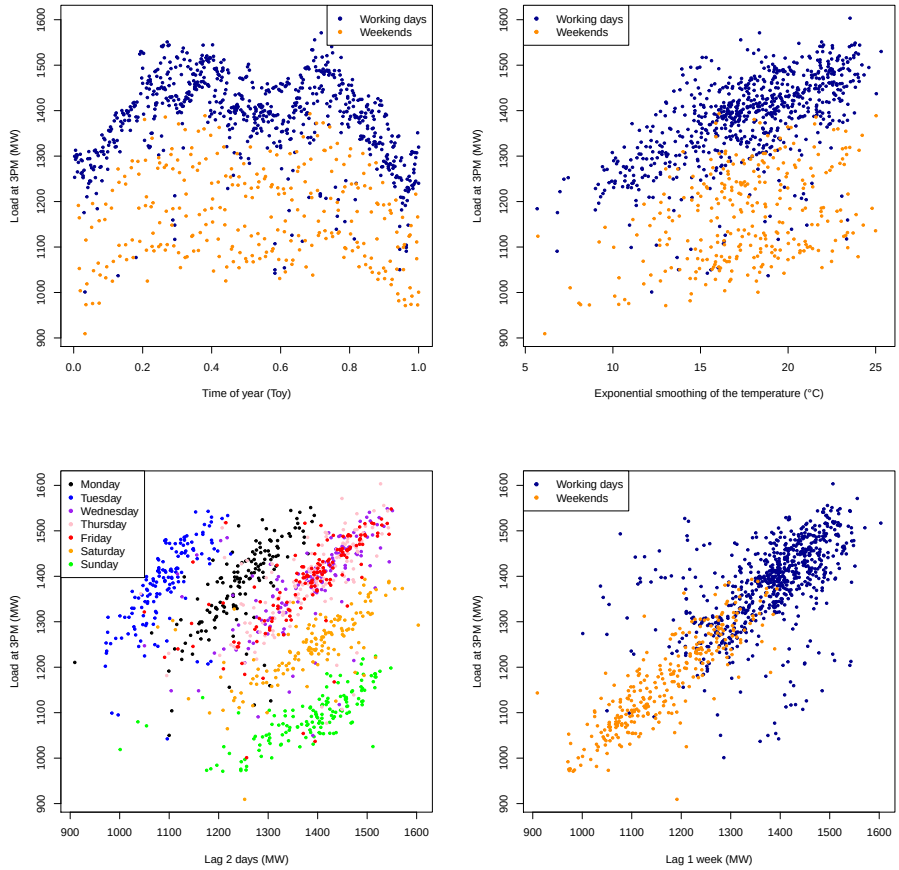


Figure 8.4 – Dependence of the load at 3PM on different covariates on the data up to January 1st 2020.

The dependence on the hour of the day and the day of the week is well observed in Figure 8.2. We display in Figure 8.4 the dependence of the load on a few of the aforementioned covariates.

We define independent models for the different hours of the day as is usual in electricity load forecasting. For each model, we use the same structure for the different hours, but we learn the model parameters independently for each time of day based on the training data of that particular time of day. In what follows, we denote by y_t the load at time t.

— **Autoregressive**. We consider a seasonal autoregressive model based on the daily and weekly lags of the load:

$$y_t = \sum_{l \in \mathcal{L}_{h(t)}} \alpha_l y_{t-l} + \sum_{1 \leq l \leq 6} \alpha_{7\times 24 l} y_{t-7\times 24 l} + \varepsilon_t \,, \tag{8.1}$$

$$\mathcal{L}_h = \begin{cases} \{24, 48, 72\} & \text{if } h \leq 7 \,, \\ \{48, 72, 96\} & \text{if } h > 7 \,. \end{cases}$$

— **Linear regression**. We use a linear model with the following variables: temperature,

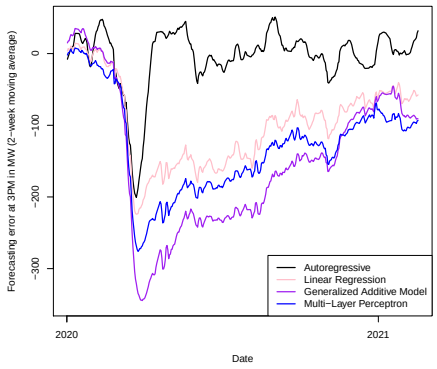


Figure 8.5 – Evolution of the forecasting error for the different models introduced in Section 8.3 trained on the data up to January 1st 2020.

cloud cover, pressure, wind direction and speed, day type (7 booleans), time of year, linear trend variable, and the two lags $LoadW$ and $LoadD$.

— **Generalized Additive Model (GAM)**. We propose a Gaussian generalized additive model (Wood, 2017):

$$y_t = \sum_{i=1}^{6} \beta_i \mathbb{1}_{DayType_t=i} + \gamma Temps95_t + f_1(Toy_t) + f_2(LoadD_t) + f_3(LoadW_t) + \alpha t + \beta_0 + \varepsilon_t \,,$$

where f_1 is obtained by penalized regression on cubic cyclic splines and f_2, f_3 on cubic regression splines.

— **Random Forest (RF)**. We build a random forest (Breiman, 2001) with the following covariates: linear trend variable, time of year, day type, the two lags, and the two exponential smoothing variables of the temperature. Quantile variants were also computed.

— **Random Forest (RF_GAM)**. We also correct the GAM using a random forest on the GAM residuals, with the same covariates as in RF to which we add the GAM effects $f_1(Toy_t)$, $f_2(LoadD_t)$, $f_3(LoadW_t)$ as well as lags (one week, one or two days) of the GAM residuals.

— **Multi-Layer Perceptron (MLP)**. Finally, we test a multi-layer perceptron of 2 hidden layers of 15 and 10 neurons using hyperbolic tangent activation. We take as input: the linear trend variable, time of year, day type, the exponential smoothing variable $Temps95$, and the two lags.

8.4 Adaptation using State-Space Models

Due to the lockdowns the consumer's behaviors changed abruptly and therefore the models presented in Section 8.3 perform poorly during Spring 2020 and afterward, see Figure 8.5. To

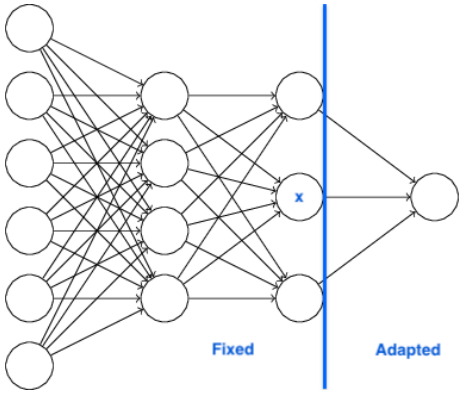


Figure 8.6 – Diagram of the definition of the features to adapt the MLP. The network has two hidden layers of 15 and 10 neurons, we freeze all the weights except the last ones.

adapt the models in time, we rely on linear Gaussian state-space models, summarized as

$$\theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t)\,,$$
$$y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma_t^2)\,,$$

where θ_t is the latent state, Q_t the process noise covariance matrix and σ_t^2 is the observation variance.

8.4.1 Definition of x_t

This state-space representation is natural for linear regression for which x_t is the vector containing the explanatory variables detailed in Section 8.3. Autoregressive models also fit directly in that framework, as they are, in fact, linear models on lags of the load, see Equation (8.1). We linearize the models to adapt GAM and MLP, and x_t is just another feature representation. We freeze the nonlinear effects in the GAM as in Section 7.2.1, and x_t contains the different effects, linear and nonlinear. We apply a similar approach for the MLP, for which we freeze the deepest layers and we learn the last one, that is, x_t is the final hidden state, see Figure 8.6.

The state-space approach is not applied to the random forest. For the latter, we compare with incremental offline random forests, consisting of re-training the random forest each day with all the data available at the time.

8.4.2 Kalman Filter

Bayesian estimation of the state θ_t in linear Gaussian state-space models is well understood under known variances σ_t^2, Q_t. The best estimator is obtained by the well-known Kalman filter (Kalman and Bucy, 1961). It yields an exact recursive estimation of the mean and covariance matrix of the state given the past observations, denoted by $\hat{\theta}_t$ and P_t. However, there is no consensus in the literature as to how to tune the hyper-parameters, see Chapter 5. The widely used expectation-maximization algorithm is an iterative algorithm that guarantees convergence to a local maximum of the likelihood. However, there is no global guarantee, and in our case, it performs poorly. We propose the following settings instead, building on Section 7.2.1:

— **Static**. We consider the degenerate setting where $Q_t = 0$ and $\hat{\theta}_1 = 0, P_1 = I, \sigma_t^2 = 1$.

— **Static break**. We consider a break at March 1st 2020 by setting $\hat{\theta}_1 = 0, P_1 = I, \sigma_t^2 = 1, Q_t = 0$ except $Q_T = I$ where T is March 1st 2020.
— **Dynamic**. We approximate the maximum-likelihood for constant variances $\sigma_t^2 = \sigma^2$ and $Q_t = Q$. We set $P_1 = \sigma^2 I$ and we observe that for a given Q/σ^2 we have closed-form solutions for $\hat{\theta}_1, \sigma^2$. Then we restrict ourselves to diagonal matrices Q/σ^2 whose nonzero coefficients are in $\{2^j, -30 \leq j \leq 0\}$ and we apply a greedy procedure: starting from $Q/\sigma^2 = 0$ we change at each step the coefficient improving the most the likelihood. That procedure is designed to optimize Q on the training data (up to January 1st 2020).
— **Dynamic break**. We use similar $\hat{\theta}_1, P_1, \sigma_t^2 = \sigma^2, Q_t = Q$ as in the dynamic setting except $Q_T = P_1 = \sigma^2 I$ where T is March 1st 2020.
— **Dynamic big**. We simply use $\sigma^2 = 1$ and a matrix Q proportional to I defined based on the 2020 data.

Also, it is important that we estimate a Gaussian posterior distribution, therefore we have a probabilistic forecast for the load. Precisely, our estimate is $\theta_t \sim \mathcal{N}(\hat{\theta}_t, P_t)$, thus we have $y_t \sim \mathcal{N}(\hat{\theta}_t^\top x_t, \sigma^2 + x_t^\top P_t x_t)$. The likelihood that is optimized to obtain the dynamic setting is built on that probabilistic forecast of y_t given the past observations. In the competition, we added quantiles of these Gaussian distributions as forecasters in the expert aggregation.

8.4.3 Dynamical Variances

The idea behind the break settings introduced in the previous paragraph is that we would like the model to adapt faster during an evolving period such as a lockdown than before. However, it consists in modeling a break in the data, a sudden change of state resulting from a noise of much bigger variance at a specific time specified *a priori*. A way to extend the approach would be to define a time-varying covariance matrix depending, for instance, on a lockdown stringency index such as defined by Hale et al., 2021. However, the competition policy forbade the use of external data, and the location was undisclosed.

In a more long-term perspective, let it be hoped that lockdowns won't drive the evolution of the electricity load. Therefore, it is more generic to learn the variances of the state-space model in an adaptive fashion. Consequently, we apply a novel approach for time-series forecasting introduced in Vilmarest, Goude, and Wintenberger, 2021 and named Variational Bayesian Variance Tracking, alias Viking. We briefly recall how the method works. This method was designed in parallel of the competition and was improved afterward. We present the last version only.

We treat the variances as latent variables and we augment the state-space model:

$$\begin{aligned}
&a_t - a_{t-1} \sim \mathcal{N}(0, \rho_a)\,, \quad b_t - b_{t-1} \sim \mathcal{N}(0, \rho_b)\,,\\
&\theta_t - \theta_{t-1} \sim \mathcal{N}(0, \exp(b_t) I)\,,\\
&y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \exp(a_t))\,.
\end{aligned}$$

Instead of estimating the state θ_t with variances fixed *a priori*, we estimate both the state and the variances represented by a_t, b_t. Although we have removed σ_t^2, Q_t as hyper-parameters, we now have to set priors on a_0, b_0 along with the parameters ρ_a, ρ_b controlling the smoothness of the dynamics on the variances.

We apply a Bayesian approach. At each step, we start from a prior $p(\theta_{t-1}, a_{t-1}, b_{t-1} \mid \mathcal{F}_{t-1})$ obtained at the last iteration, where we introduce the filtration $\mathcal{F}_t = \sigma(x_1, y_1, ..., x_{t-1}, y_{t-1})$. Then we obtain a prediction step thanks to the dynamical equations yielding $p(\theta_t, a_t, b_t \mid \mathcal{F}_{t-1})$. Finally Bayes' rule yields the posterior distribution $p(\theta_t, a_t, b_t \mid \mathcal{F}_t)$.

However the posterior distribution is analytically intractable, therefore the principle of Viking is to apply the classical variational Bayesian approach (Šmídl and Quinn, 2006). The posterior

distribution is recursively approximated with a factorized distribution. In our setting we look for the best product $\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})$ approximating $p(\theta_t, a_t, b_t \mid \mathcal{F}_t)$. The criterion minimized is the Kullback-Leibler (KL) divergence

$$KL(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \mid\mid p(\theta_t, a_t, b_t \mid \mathcal{F}_t)),$$

where $KL(p, q) = \int \log(dp/dq)dp$. At each step it yields a coupled optimization problem in the three Gaussian distributions. The classical iterative method (see for instance Tzikas, Likas, and Galatsanos, 2008) consists in computing alternately $\exp(\mathbb{E}[\log p(\theta_t, a_t, b_t \mid \mathcal{F}_t)])$ where the expected value is taken with respect to two of the three latent variables, and identifying the desired first two moments with respect to the other latent variable. However the expression $\exp(\mathbb{E}_{\theta_t, b_t}[\log p(\theta_t, a_t, b_t \mid \mathcal{F}_t)])$ doesn't match a Gaussian distribution in a_t, and similarly for b_t. We therefore use the first two moments of the Gaussian distribution to derive an upper-bound of the KL divergence for which we have an analytical solution. We refer to Vilmarest, Goude, and Wintenberger, 2021 for the detailed derivation of the algorithm. Chapter 6 presents a final version of Viking.

8.5 Experiments

We display the performance of the introduced methods that we call experts. Then we use aggregation of experts to leverage specificities of each forecaster. The end of the section is devoted to discussing our day-to-day strategy during the competition. Finally, we refer to the implementation for more details [1].

8.5.1 Intraday Correction

Although it performs better to use different models at different hours of the day, let it be noted that the correlation between different hours is important. To capture intraday information, we fit on the residuals of each model an autoregressive model incorporating lags of the 24 last available hours and optimized for each forecast horizon. This follows from the intuition that to predict the load at 8 AM, instead of using as the latest available data a delay of 48 hours, we can use a 25-hour delay.

We apply this correction to the models presented in Section 8.3 as well as to the ones resulting from the adaptation framework of Section 8.4. We display in Table 8.2 the gain for the statistical and machine learning methods of Section 8.3. To present the improvement brought by the intraday correction, we give the performance during a stable period (after the training of the model, but before the covid crisis). We observe that the only model for which the intraday correction doesn't improve the performance (RF_GAM) is the one including already a residual correction. The improvement during the evaluation period (2021) is much bigger (57% decrease of the MAE for the MLP, for instance). It is natural as the intraday correction is an autoregressive, that is, an adaptive model.

8.5.2 Adaptation of Individual Experts

Then we focus on adaptive models to show the improvements due to each setting, see Table 8.3. We have four different models (autoregressive, linear, GAM, and MLP). We try the various adaptation settings (no adaptation, Kalman filters, and Viking) for each one. Kalman

1. https://gitlab.com/JosephdeVilmarest/state-space-post-covid-forecasting

Adaptation	AR	Linear	GAM	RF	RF_GAM	MLP
No intraday	29.3	20.8	20.7	24.6	23.0	21.2
With intraday	27.0	19.9	19.3	24.4	23.7	20.6

Table 8.2 – Mean absolute error (in MW) of each method of Section 8.2 during normal test period. Models are trained up to Jan. 1st 2020 and tested during the next two months before the break of March.

Adaptation	AR	Linear	GAM	MLP
Offline	14.6	22.8	22.7	16.7
Static	20.5	15.7	17.0	22.9
Static break	27.9	14.4	28.4	35.4
Dynamic	14.4	14.9	15.3	13.0
Dynamic break	16.2	13.6	14.3	12.3
Dynamic big	14.3	11.2	12.4	12.4
Viking	14.4	11.5	12.7	12.5

Table 8.3 – Mean absolute error (in MW) of each method during the competition evaluation set (2021-01-18 to 2021-02-16). The performances are displayed for each model after intraday correction. As a comparison, re-training the random forest every day yields an online RF of MAE 15.0 MW, and an online GAM_RF of MAE 18.1 MW. The organizers propose a naive benchmark (relying on persistence) of MAE 15.5 MW.

filters with a constant covariance matrix proportional to the identity obtain the best results. That is not the case on the data previous to the competition, and it depends on the intrinsic evolution of the data.

We illustrate the different settings in Figure 8.7 where we display the evolution of the state coefficient for the GAM adaptation strategies.

Furthermore, in Figure 8.8 we present the evolution of the GAM model adapted by Viking for the 24 different hours of the day. This, as well as Figure 8.2, shows the necessity of different models for the different hours of the day. However, the resemblance of close hours motivates the intraday correction to benefit from the correlation between hours, see Section 8.5.1.

8.5.3 Aggregation

Online robust aggregation of experts (Cesa-Bianchi and Lugosi, 2006) is a powerful model agnostic approach for time series forecasting, already applied to load forecasting during the lockdown (see for instance Section 7.4.2). We use the ML-Poly algorithm proposed by Gaillard, Stoltz, and Van Erven, 2014 and implemented in the `R` package `opera` (Gaillard and Goude, 2016) to compute these online weights.

The aggregation weights are estimated independently for each hour of the day. We summarize different variants in Table 8.4. First, for each family of models we compute the aggregation of all the adaptation settings (7 for each). Then we aggregate all of them (28 models). An example of the weights obtained at 3 PM is displayed in Figure 8.9. The aggregation presented in this paper obtains a performance close to our strategy winning the competition (degradation of about 0.05 MW).

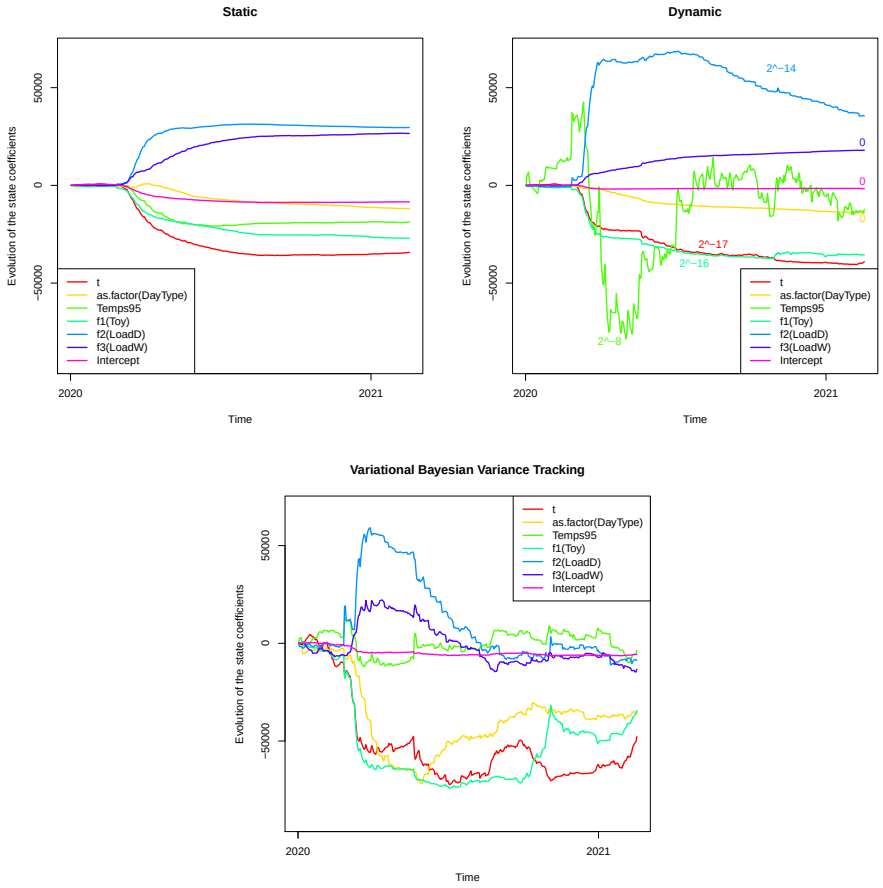


Figure 8.7 – Evolution of the state coefficients for various adaptations of the GAM, see Section 8.4. On the top left, we use the Kalman filter in the static setting (degenerate covariance matrix $Q_t = 0$). On the top right, the dynamic setting where the variances are constant, and we provide the ratio $Q/\sigma^2 = diag(2^{-17}, 0, 2^{-8}, 2^{-16}, 2^{-14}, 0, 0)$: we observe that the coefficient corresponding to the largest coefficient of Q (the effect of $Temps95$) evolves much faster. On the bottom, the Viking setting where we estimate the variances adaptively.

Adaptation	AR	Linear	GAM	MLP	All
Best expert	14.3	11.2	12.7	12.3	11.2
Aggregation	14.4	11.4	11.7	11.9	**10.9**

Table 8.4 – Mean absolute error of aggregation strategies (in MW) during the competition evaluation set (2021-01-18 to 2021-02-16).

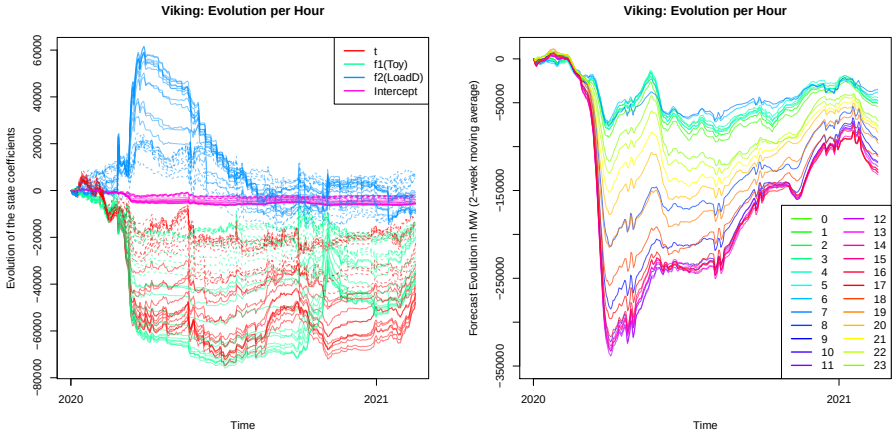


Figure 8.8 – Evolution of the Viking adaptation of the GAM for the 24 different hours of the day. On the left: evolution of 4 of the 7 coefficients as in Figure 8.7. We plot night hours (8 PM - 8 AM) in dashed lines and daylight hours (8 AM - 8 PM) in solid lines to show the groups. On the right: impact of the evolution on the forecast. Precisely, on the right graph, we display the difference between the forecast of Viking and the forecast that would have been made by Viking on January 1st 2020.

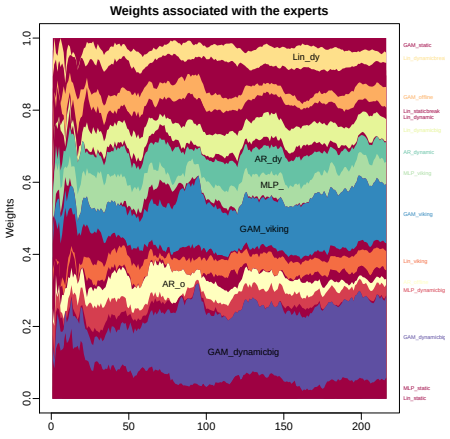


Figure 8.9 – Evolution of the aggregation weights at 3 PM from July 1st 2020 to February 16th 2021.

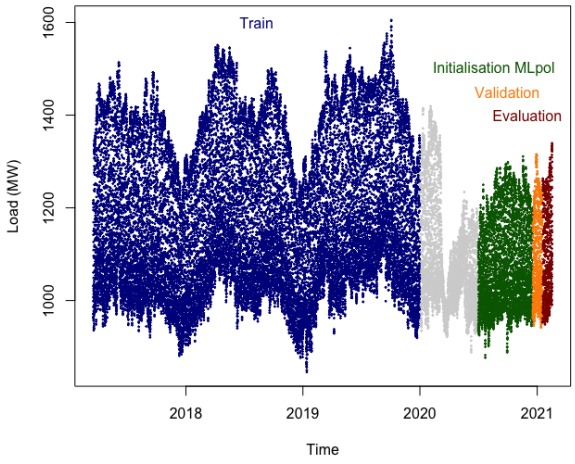


Figure 8.10 – Segmentation of the data set. Meteorological corrections as well as time-invariant forecasts were trained on the train period (up to January 1st 2020). Adaptive forecasting methods were evolving on the whole period. Then the aggregation weights were trained by MLpol from July 1st 2020, and the expert selection was determined with respect to the validation set.

8.5.4 Day-to-day Forecasts

During the competition, our predictions were not exactly the ones of the aggregation method presented in the previous subsection. There are mainly two reasons for that.

First, we considered a bigger set of forecasting methods (we had 72 experts). It seemed reasonable to prune the strategy for the sake of the paper's clarity, at the cost of a very small change of error. Still, it is interesting to present also the predictions used during the evaluation. We found a trade-off in the selection of experts. Indeed, too many experts in the aggregation yield poor performances. We applied a greedy procedure to select the experts we keep in the aggregation: we begin with an empty set, and at each step, we add the one improving the performance the most. That performance was evaluated with the MAE on the last month of the training data set. We provide in Figure 8.10 a graphical representation of how we defined different time periods. We refer to Figure 8.11 for the evolution of the validation MAE as the selection grows. We observe a sharp decrease of the error as experts are added with high diversity, and then a slow increase of the loss as the set of experts becomes too large.

Second, we were constantly experimenting the different strategies. We used a variational Bayesian method that was a prior version of the one of this paper. We also changed a lot the aggregation procedure.

We refer to Appendix D for a detailed presentation of our daily strategy. Official results of the challenge and additional significance analysis are described by Farrokhabadi et al., 2021. Overall, these day-by-day changes degraded the performance; if we had stayed on the first strategy with no change at all, our MAE would have been 10.51 MW instead of 10.84 MW. The critical issue in such unstable periods is to find the suitable validation period to select the prediction procedure. The month before the evaluation period seems *a posteriori* a good compromise. During the competition, we changed "manually" based on the performances in a shorter range, considering for instance an expert performing well on the last few weeks for a specific day type ... We should

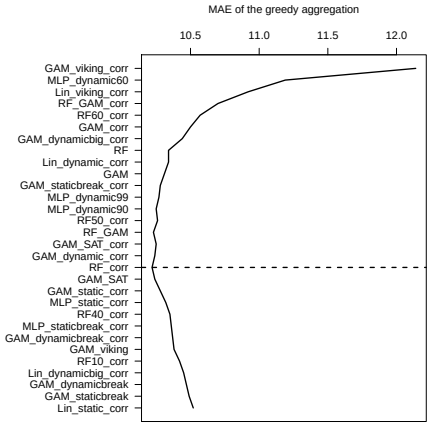


Figure 8.11 – Evolution of the validation MAE as the expert selection grows from 1 to 30 experts. The nomenclature is provided in Appendix D.

have trusted the aggregation's robustness.

8.6 Conclusion

In this paper, we presented our procedure to win a competition on electricity load forecasting during an unstable period. Our approach relies heavily on state-space models, and the competition was the first data set on which was applied a recent approach to adapt the variances of a state-space model. Some perspectives have been raised during the competition, such as interpretability of the global approach and a better understanding of the error propagation along the different adaptations (intraday correction, Kalman filtering, variance tracking, and aggregation).

Finally, similar state-space methods have been applied to obtain first place in another competition in which the objective was to forecast the electricity consumption of a building. We present this competition in Chapter 9.

Chapter 9

Competition on Building Energy Consumption Forecasting

We participated in a competition where the objective was to forecast the electricity consumption of a building. Our aim was to motivate state-space models to forecast the electricity load at a low level of aggregation. Indeed, the load at a small scale is expected to be unstable because a change of behavior of a single person may have a non-negligible impact on the variable of interest. Therefore, we claim that adaptive methods are necessary. We won the competition and we present in this chapter the strategy implemented.

Contents

9.1 Introduction

The objective of the competition was to forecast the electricity load of a building at a 15-minute granularity. To predict the 96 quarters of an hour of day d, we had access to the whole day $d-1$. It corresponds to predicting at midnight the whole following day, having access to the data with no delay. The evaluation period of the competition was February 10^{th} to 14^{th} 2020. The training period started January 1^{st} 2019 and ended at the beginning of the evaluation period.

We describe the electricity load in Figure 9.1. We observe an almost constant base consumption during nights and weekends, probably due to inactivity in the building. The main variation of the load occurs during weekdays between 8 AM and 8 PM. The metric used to evaluate the participants of the competition is non-standard (Pinto et al., 2021). We detail it in Section 9.5.

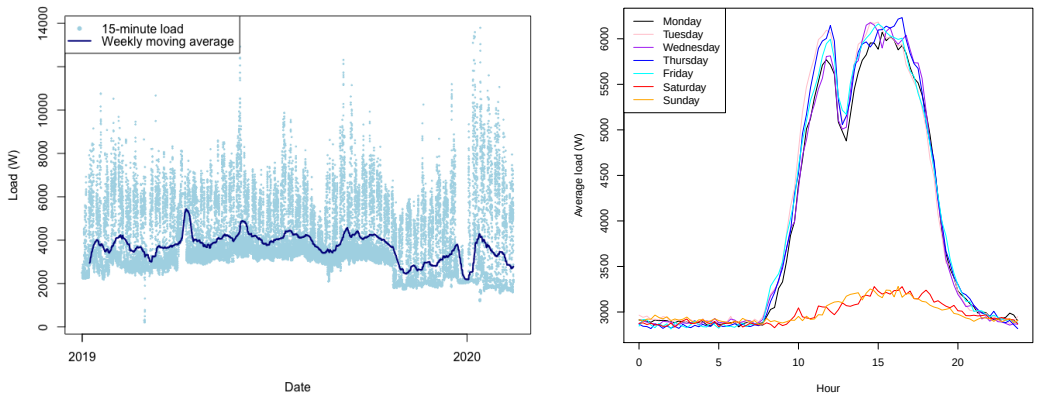


Figure 9.1 – Electricity load of the building considered in the competition. On the left: load per quarter of an hour in the whole data set, and 7-day moving average. On the right: daily profiles of the electricity load for each day of the week.

We have access to many explanatory variables. First, the data set contains the consumption and generation, as there are solar panels on the building. We don't see any possible use of the generation variable for the forecasting task at hand. However, as presented in Chapter 10, this data may be helpful if one wishes to forecast the net-load (reduced of the local generation). Second, meteorological data from a near weather station was provided (temperature, humidity, and solar radiation), as well as a second temperature variable that we deem likely to be measured closer to the building. Third, the building is decomposed into various zones, and we had access to the past load of each zone, as well as sensor information (temperature, humidity, lamp intensity). Our method doesn't take that disaggregated data into account. Finally, each day of the evaluation period, after we sent our prediction, we were given the data of the forecasted day.

Our strategy is decomposed into four steps, following the diagram of Figure 1.5. We first pre-process the data in Section 9.2, then we present two statistical models in Section 9.3. In Section 9.4, we adapt the statistical models, and finally we combine different forecasts in Section 9.5.

9.2 Pre-Processing

We aim to forecast at a 15-minute granularity while the data is provided with 5-minute intervals. We don't use that subdivision. Instead, we average each variable at 15-minute intervals, yielding 96 times of day. When the data is missing, we use linear interpolation per time of day.

Among the explanatory variables that were given we assess that the most useful ones are the humidity and radiation features from the *weather_ data* table, and the temperature variable from the *building_ sensor* table, see Figure 9.2. However, these covariates are realized meteorological variables, and we are not given forecasts for them. Thus we have two options. Either we use lags of the realized variables, or we forecast them. In some sense, we do both simultaneously: we forecast them with autoregressive models. Formally, each variable of interest z_t is modeled

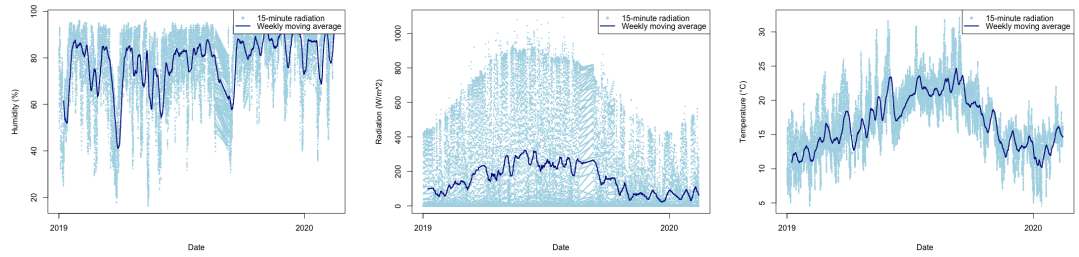


Figure 9.2 – Humidity (left), radiation (middle) and temperature (right) variables.

by the following equation:

$$z_t = \sum_{l \in \mathcal{L}} \alpha^l_{h(t)} z_{t-l} + \varepsilon_t \,,$$

where $h(t) \in \{0, 1, \ldots, 95\}$ is the time of day of t, ε_t is an i.i.d. Gaussian noise and $\mathcal{L}$ is a set of lags. Therefore, for each variable of interest and each time of day, we learn a linear model on lags of the variable of interest, and we use that linear model to make forecasts using the data provided at the prediction time. For humidity and temperature, we use the last 10 available values (end of the previous day) and the last 10 days at the same time of day. For radiation, we use the last 2 days at the same time of day.

Finally, our forecast for time t depends on the following covariates:

— $Hum_t, Rad_t, Temp_t, Temps99_t$: the forecasts for the humidity, radiation, and temperature variables, and an exponential smoothing of parameter 0.99 of the temperature forecast.

— $LoadDay_t, LoadWeek_t, LoadLast_t$: lags one day and one week of the electricity load, as well as the last load available (the day before at 11:45 PM).

— $DayType_t$: day of the week ranging from 1 to 7.

9.3 Statistical Models

We forecast the load using the variables defined in Section 9.2. Our models are defined per time of day, that is, we build 96 models independently of each other but using the same structure as follows.

We first define a linear regression where we use different effects of the lags and different intercepts for the different days of the week:

$$\begin{aligned} Load_t = {} & \alpha_1 Hum_t + \alpha_2 Rad_t + \alpha_3 Temp_t + \alpha_4 Temps99_t + \alpha_5 LoadLast_t \\ & + \sum_{i=1}^{7} \beta_i \mathbb{1}_{DayType_t = i} LoadDay_t + \sum_{i=1}^{7} \gamma_i \mathbb{1}_{DayType_t = i} LoadWeek_t + \sum_{i=1}^{7} \delta_i \mathbb{1}_{DayType_t = i} + \varepsilon_t \,. \end{aligned}$$

Then we design a generalized additive model (GAM). The difference from the linear model is limited to the meteorological forecasts. In the GAM, we define a nonlinear effect of the

temperature, and we remove the other meteorological forecasts:

$$Load_t = s(Temp_t) + \sum_{i=1}^{7} \beta_i \mathbb{1}_{DayType_t=i} LoadDay_t + \sum_{i=1}^{7} \gamma_i \mathbb{1}_{DayType_t=i} LoadWeek_t + \sum_{i=1}^{7} \delta_i \mathbb{1}_{DayType_t=i} + \varepsilon_t \,,$$

where the effect of the temperature is decomposed on a spline basis. The optimization is realized by penalized least-squares using the R package `mgcv` (Wood, 2015).

9.4 Adaptation

As mentioned previously, we claim that adaptive methods are crucial to forecast at a low level of aggregation. It is remarkable that the models detailed above are already very adaptive as they essentially depend on the lags. However, we reduce the error of the predictive models using a Kalman filter. We use the following state-space model:

$$\begin{aligned} &\text{State:} && \theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q) \,, \\ &\text{Space:} && Load_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma^2) \,, \end{aligned}$$

where the vector x_t is defined differently for the linear regression and for the GAM. It is naturally defined as the initial covariate vector for the linear regression. For the GAM, it is composed of the 5 effects (linear or nonlinear) of the GAM as in Section 7.2.1. We use the standard Kalman filter with the following hyper-parameters: $\hat{\theta}_1 = 0, P_1 = I, Q = 10^{-4} I, \sigma^2 = 1$. Indeed, estimating the state and space variances as in Part II didn't lead to improving the performances, and we believe that may be due to the lack of a sufficiently wide data set.

9.5 Final Forecasts and Performances

In this section, we assess the performances of the different methods, and we present the strategy we followed during the competition.

First, we define precisely the competition metric, which may be interpreted as a combination of mean absolute error (MAE) and root mean squared error (RMSE). For each day, let $(e_1, \dots e_{96})$ denote the difference between the realized loads and the forecasts. Then we define a slight modification of the MAE to penalize high errors and to penalize business hours. Precisely, we define $c_i = 4$ for the 12 times between 7 AM and 8 PM of highest errors, $c_i = 2$ for the other times between 7 AM and 8 PM, and $c_i = 1$ otherwise. It allows to define $m = \frac{1}{96} \sum_{i=1}^{96} c_i |e_i|$. Then the metric is defined as the weighted average $\frac{5}{6} m + \frac{1}{6} s$ where s is the standard deviation of $(e_1, \dots e_{96})$. Finally, on a period of more than one day, the metric is computed per day, then averaged.

We present in Table 9.1 the numerical performances on a validation set consisting of January 1$^{\text{st}}$ to February 9$^{\text{th}}$ 2020, while the models are trained on 2019. We compare the models introduced previously with naive persistent benchmarks:

— Last Load: we predict the load of the whole day as being constant equal to the one of the day before at 11:45 PM.

— Load 1 Day and Load 1 Week are the lags of the electricity load. We also compare with the average of these two lags.

	Val. MAE (W)	Val. RMSE (W)	Val. Metric (W)	Eval. Metric (W)
Last Load	1509	2739	3703	3615
Load 1 Day	1179	2074	2813	1567
Load 1 Week	968	1748	2252	1544
Average of lags	929	1563	2179	1151
Weekly Profile	1330	1635	2616	1706
Linear Offline	721	1292	1761	1067
Linear Kalman	693	1263	1681	968
GAM Offline	710	1246	1733	1058
GAM Kalman	686	1206	1674	913
Average Kalman	681	1222	1656	916
Final Forecast	648	1211	1625	883

Table 9.1 – Performances of different benchmarks and our models. We display the mean absolute error (MAE), the root mean squared error (RMSE), as they are more interpretable, and then the competition metric, all on the validation set. The last column is the result on the evaluation period. Average Kalman is the average of both Kalman filters.

— Weekly Profile: we forecast the load with the average load in 2019 at the specific day of the week and the specific time of day.

To boost the performances, it is natural to combine different forecasts to leverage specificities of each one. We first reduce the error slightly using the average of both Kalman filters.

Furthermore, it is interesting to study the daily profile of the error. We see that the best method is not the same for all the 15-minute intervals, see Figure 9.3. We display the mean absolute error as the competition metric is not defined per time of day. We observe that the last load available is the best forecast up to 8 AM, which seems reasonable as the load is almost constant during the night. Then, from 8:15 AM to 8 PM, the average of the Kalman filters performs better. Finally, from 8:15 PM, the average of lags is the best method. Our final forecast selects the best of these three methods depending on the time of day. Compared with the average of Kalman filters, this selection reduces the competition metric by about 2% on the validation period and 4% on the evaluation period, see Table 9.1.

9.6 Conclusion

We obtained first place in this competition using a straightforward application of our methodology. Therefore we believe there are promising perspectives in applying our work to low-level data.

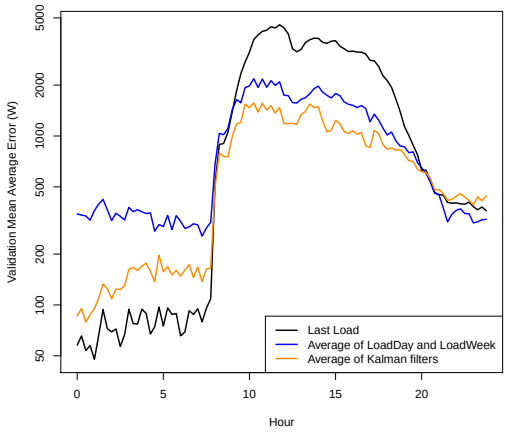


Figure 9.3 – Mean Absolute Error of different forecasts on the validation set depending on the time of day. Notice that the scale is logarithmic on the y-axis. Therefore, although there is an important relative gain before 8 AM using the last load, it has a moderate impact because the error is small for all models.

Chapter 10

Adaptive Probabilistic Forecasting of Electricity (Net-)Load

In this chapter, we present adaptive methods for probabilistic forecasting. We consider two data sets: the electricity net-load of Great Britain (load reduced of embedded solar and wind generation) at a half-hour time granularity, and the daily load of New York City.

Contents

10.1 Introduction

In Chapters 7, 8 and 9, we were interested in point forecasting. The objective was to make a prediction achieving the minimum of a loss function. In Chapter 7 we considered mostly the root-mean-square-error (RMSE) for which the best forecast is the expected value of the load. In the competition presented in Chapter 8, the objective was to minimize the mean absolute error

(MAE), and the best forecast is the median. The evaluation of the competition presented in Chapter 9 was more intricate and may be interpreted as a combination of MAE and RMSE.

However, the electricity system relies on more than point forecasting. Indeed, the expected value of the load is not sufficient for risk management, and it is better to have some information on the distribution of the load around its mean. An interesting indicator is its variance, which may also be estimated by existing approaches, including the Kalman filter. In this chapter, we are interested in forecasting the quantiles of the load, that is called probabilistic forecasting.

Probabilistic load forecasting has been widely studied. We refer to the review of Hong and Fan, 2016, as well as to the last two Global Energy Forecasting Competitions (Hong et al., 2016; Hong, Xie, and Black, 2019). Our aim is to design adaptive methods in the vein of the state-space approach presented in the previous chapters for point forecasting.

In Section 10.2 we present adaptive methods for probabilistic forecasting. Then we apply them to two data sets. In Section 10.3 we consider the regional net-load in Great Britain (Browell and Fasiolo, 2021). As the renewable generation increases, the available production means must meet the net-demand, that is, the load reduced of the wind and solar production (that are not controllable). Therefore it becomes fundamental to forecast the net-load. In Section 10.4 we apply the methods on the load in New York City (Ruan et al., 2020). In this latter application, we use daily data to determine whether our approach works well at this time granularity.

10.2 Theoretical Framework

This section presents the procedure we apply to obtain adaptive probabilistic forecasting.

10.2.1 Offline Model

Motivated by Browell and Fasiolo, 2021, we decompose our model into two steps. We use a generalized additive model (GAM) to forecast the mean, and then we use quantile regression on the GAM residuals to forecast the quantiles. This two-step procedure is essentially motivated by computational time. Indeed, we could also use quantile GAM (Fasiolo et al., 2021), but it is more time-consuming than simple quantile regressions. Formally, let $y_t \in \mathbb{R}$ be the variable of interest.

1. We model y_t as a Gaussian random variable whose mean is predicted using the covariates $x_{t,1}, \dots x_{t,d}$ and the following formula:

$$\hat{y}_t = \sum_{j=1}^{d} f_j(x_{t,j}) \,,$$

   where the effects $f_1, \dots f_d$ are either linear or nonlinear. In the latter case, the effects are decomposed on spline bases (Wood, 2017).

2. The Gaussian assumption $y_t - \hat{y}_t \sim \mathcal{N}(0, \sigma^2)$ is violated in practice, and therefore we fit a set of quantile regressions (Koenker and Bassett Jr, 1978) on the residual to predict the distribution of the variable. We use a different vector of covariates $z_t \in \mathbb{R}^{d_0}$, and for some q, we define a vector $\beta_q \in \mathbb{R}^{d_0}$ by the following minimization problem on a training set $\mathcal{T}$:

$$\beta_q \in \arg\min_{\beta \in \mathbb{R}^{d_0}} \sum_{t \in \mathcal{T}} \rho_q(y_t - \hat{y}_t, \beta^\top z_t) \,.$$

where $\rho_q(y, \hat{y}_q) = (\mathbb{1}_{y<\hat{y}_q} - q)\ (\hat{y}_q - y)$ is the pinball loss. This pinball loss is motivated by the following lemma (Koenker and Hallock, 2001):

Lemma 10.1. *Let Y be an integrable real-valued random variable. For any $0 < q < 1$ the q-quantile of Y denoted by Y_q satisfies $Y_q \in \arg\min \mathbb{E}[\rho_q(Y, Y_q)]$.*

Finally we predict the quantile for probability level q with $\hat{y}_t + \beta_q^\top z_t$.

10.2.2 Adaptation of the Mean Forecast

We adapt the GAM as described in Section 7.2.1. We freeze the nonlinear effects, and we rely on a linear Gaussian state-space model to adapt a linear model on the new covariates $f_1(x_{t,1}), \dots f_d(x_{t,d})$. We use the standard Kalman filter applied with the variances obtained from our iterative grid search algorithm introduced in Section 5.3.3.

10.2.3 Adaptation of the Quantile Forecast

The specificity of this chapter lies in the adaptation of probabilistic forecasts. We compare different methods, besides incremental offline models where we re-train the model frequently with an increasing data set.

We first remark that the Kalman filter does not only yield a mean forecast but already a probabilistic forecast. Indeed, the Kalman filter yields estimates of the expected value and covariance matrix of the state vector given the past observations. Moreover, the *a posteriori* distribution of the state is Gaussian as long as the state-space assumption is satisfied. The observation is thus predicted as a Gaussian random variable whose mean and variance are estimated by the Kalman filter. It yields a first adaptive probabilistic forecast.

However, this first adaptive forecaster would be the generalization of an offline Gaussian distribution on the offline GAM residuals. The quantile regressions were specifically introduced because the Gaussian assumption is violated in practice. We then remark that if the conditional distribution of the residual $y_t - \hat{y}_t$ given the covariates z_t is constant, then the offline quantile regression model should work well. The need for adaptive methods is motivated only by changes in the residual distribution. Therefore we test the combination of the state-space adaptation of the GAM and the offline quantile regressions. Indeed, a critical property of the Kalman filter is that the residuals are stationary, provided that the state-space model is well-specified. Therefore, it is natural that the dependence of the residuals of the state-space GAM on the quantile covariates should be more stable than for the offline GAM.

Finally, we define an online quantile regression by simply applying online gradient descent (OGD) on the pinball loss. Precisely, we defined the offline quantile regression for the q-quantile as $\beta_q \in \arg\min_{\beta\in\mathbb{R}^{d_0}} \sum_{t\in\mathcal{T}} \rho_q(y_t - \hat{y}_t, \beta^\top z_t)$. The OGD allows to estimate recursively a vector $\beta_{t,q}$. We start from $\beta_{1,q} \in \mathbb{R}^{d_0}$ and at each step we update it with a step on the direction opposite to the gradient of the loss:

$$\beta_{t+1,q} = \beta_{t,q} - \alpha \frac{\partial \rho_q(y_t - \hat{y}_t, \beta^\top z_t)}{\partial \beta}\Big|_{\beta_{t,q}}.$$

We use a constant gradient step size α, and we standardize the covariates z_t.

10.3 Net-Load Forecasting in Great Britain

We first study the data set created by Browell and Fasiolo, 2021. We thank them for updating the data for us. Indeed, they studied the data from 2014 to 2018, and they augmented the period

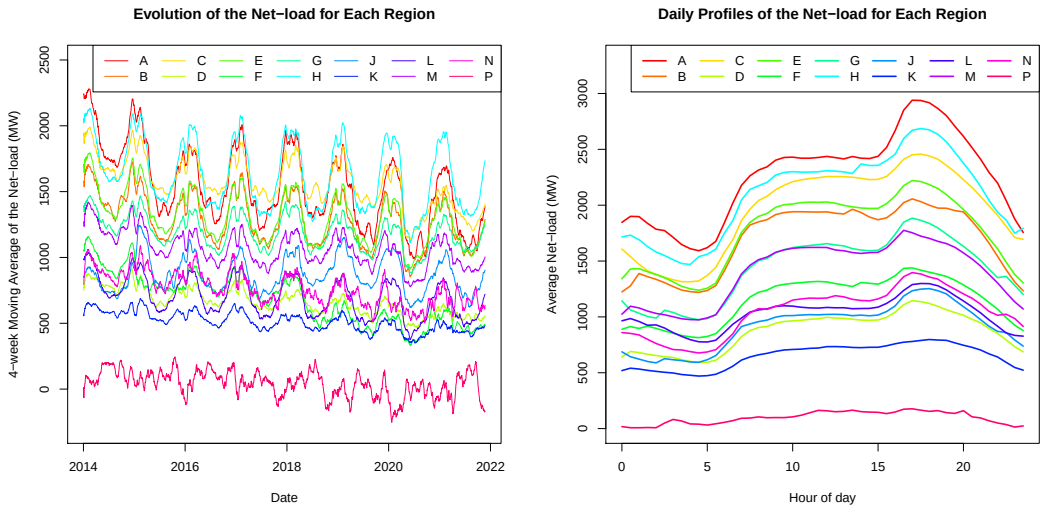


Figure 10.1 – On the left: evolution of the net-load of the 14 regions. On the right: daily profiles. We observe that in region P (North Scotland) the embedded generation often exceeds the consumption and the daily profile is close to 0. Also, this high generation means higher volatility, and the net-load does not have a clear yearly profile as in the other regions.

to range from 2014 to 2021, allowing us to integrate the unstable covid period. We refer to Browell and Fasiolo, 2021 and Browell, 2021 for more details on the data set. Due to Brexit, the day-ahead electricity price definition changed. Thus, we decided to remove the variable from the data set.

10.3.1 Data Presentation and Offline Model

We are interested in forecasting the electricity net-load, defined as the difference between electricity consumption and embedded generation (wind and solar production). We consider the data from Great Britain at a region-scale. Great Britain is divided into 14 regions called Grid Supply Point Groups. The time granularity is half an hour. We display in Figure 10.1 the evolution of the net-load in each region, as well as the daily profiles. We observe different behaviors. In particular, region P (North Scotland) often has a negative net-load, meaning the consumption is smaller than the embedded production.

For each region independently, the model designed by Browell and Fasiolo, 2021 to predict quantiles of the net-load is decomposed into three steps. First, a GAM is fitted to forecast the mean. Second, a set of quantile regressions is fitted on the GAM residuals (between 2.5% and 97.5%). Third, extreme quantiles are modeled by a generalized Pareto distribution. We only interest ourselves in non-extreme quantiles, and we consider the two steps introduced in Section 10.2.1.

1. The mean model we detail here is the only model of the present manuscript where we consider a unified model for all times of day. Indeed, most of the models presented are designed independently for each time of day (hour, half-hour), each following the same model structure. We use the formula of the *GAM Point* of Browell and Fasiolo, 2021, where we removed the electricity price. The normalized electricity net-load at time t is

modeled as a Gaussian random variable whose mean is the sum of
- linear functions of t and t^2 (polynomial trend),
- linear functions of $\sin(2\pi Toy_t)$, $\sin(4\pi Toy_t)$, $\cos(2\pi Toy_t)$ and $\cos(4\pi Toy_t)$ where Toy_t is the time of year, growing linearly from 0 on January 1^{st} to 1 on December 31^{st},
- a linear function of the one-day lag of a moving average of the normalized net-load (at any time of day d this lag is the average of the normalized net-load between day $d-15$ and day $d-2$ included),
- a linear effect of the categorical variable $WeekDay_t$, which is the day of the week except during school holidays (it is either Monday to Sunday, either Christmas, February half-term, Easter, Summer or Autumn holidays),
- a linear effect of the categorical variable $SchoolHol_t$, defining if the day is a Christmas holiday, another school holiday or a normal day,
- three nonlinear effects of the time of the day $Tod_t \in \{0, \dots, 47\}$: one global, one depending on the value of $WeekDay$ and one depending on the value of $SchoolHol$,
- two nonlinear effects: one of the temperature at maximum population density and one of a two-day moving average of that temperature,
- a nonlinear effect of the product between the solar radiation and the embedded solar capacity,
- a linear effect of the wind speed, as well as a nonlinear effect of the wind speed depending on the embedded wind capacity,
- a nonlinear effect of the precipitation,
- two tensor products: one of the temperature and the time of day, one of the precipitation and the time of day.

All nonlinear effects are decomposed on cubic regression splines.

2. For the quantile regressions, we also use the model of Browell and Fasiolo, 2021: the GAM residuals are modeled as a linear function of the GAM prediction, the squared prediction, the product of solar radiation and embedded solar capacity, the wind speed, the temperature, as well as of the categorical versions of the time of day and $WeekDay$. For these two latter variables, we have an additive constant defined for each value of the categorical variable.

   When we use quantile regression on the residuals of the GAM adapted by the Kalman filter, we replace the square of the prediction with the estimated variance of the observation. Indeed, the Kalman filter estimates the mean and covariance matrix of the state vector, from which we obtain the mean and variance of the observation.

10.3.2 Performances of Mean Forecast

Similarly as in the previous chapter we improve the mean forecast performance in almost all regions and all years using state-space adaptation, we display the root-mean-square error (RMSE) for each region and each year in Figure 10.2. We also compute the RMSE of the aggregated data: the root of the mean squared error for the standardized net-load of the 14 regions. We display it in Table 10.1 for different methods and different years. When we compare the dynamic Kalman filter to the incremental offline GAM, we have a much lower computational cost per day; we reduce the RMSE by approximately 3% in 2019, 11% in 2020 and 11% in 2021.

10.3.3 Performances of Quantile Forecast

There are many ways to evaluate quantile forecasts. A first qualitative evaluation is *reliability*, also known as *calibration*. A quantile forecast is reliable if the observed frequency coincides with

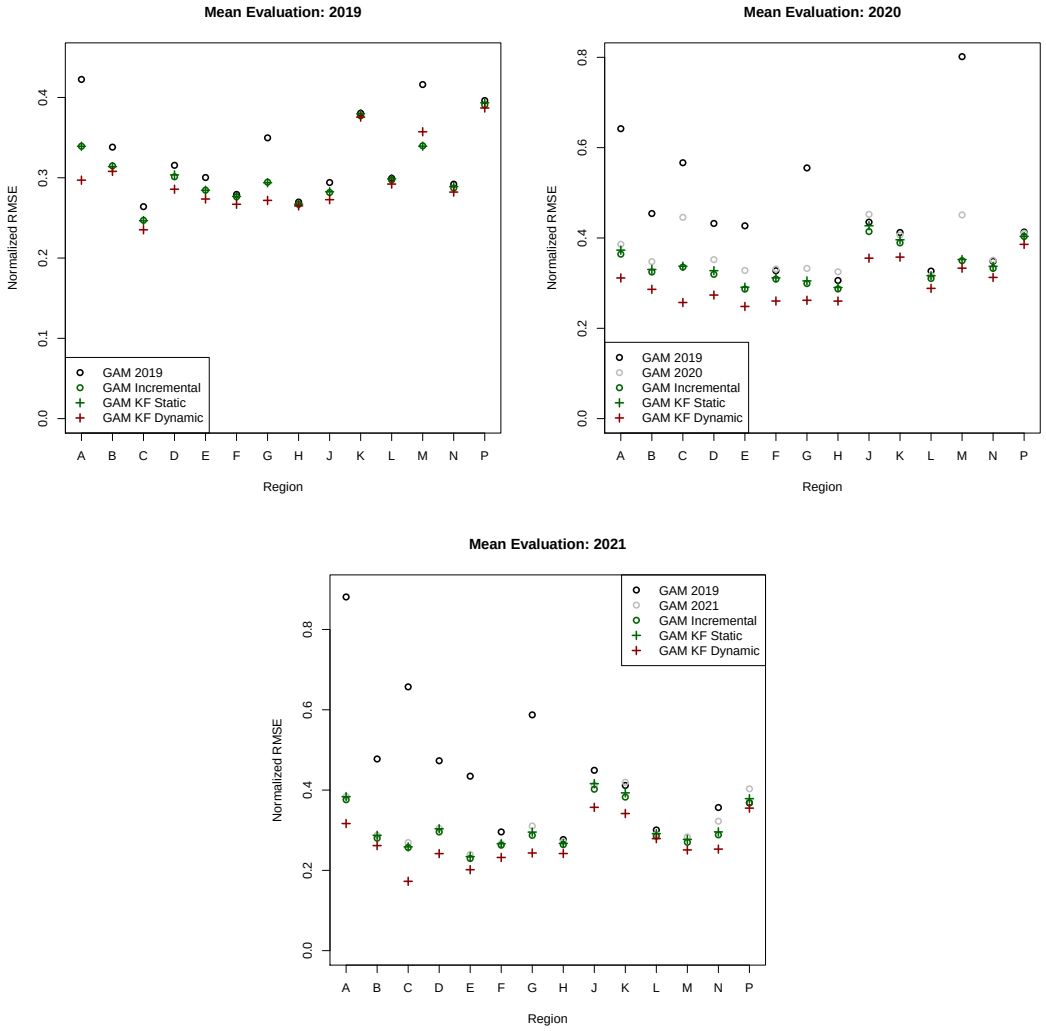


Figure 10.2 – Normalized RMSE for each region in Great Britain. We compare the offline GAM to the incremental offline GAM (trained each day on incremental training data), the Kalman filter static (degenerate covariance matrix of the state noise $Q = 0$), and the Kalman filter in the dynamic setting with *iterative grid search* method. We divide the test set in three (2019, 2020 and 2021). For the last two years we compare also with the GAM re-trained each year: GAM y is the GAM trained with the data up to year $y - 1$ included.

Method	2019	2020	2021
Offline GAM	0.334	0.479	0.560
Incremental offline GAM (yearly)	-	0.377	0.324
Incremental offline GAM (daily)	0.310	0.340	0.308
Kalman GAM	0.301	0.302	0.273

Table 10.1 – Aggregate RMSE on the standardized net-load of the 14 regions. Two incremental offline methods are presented, updated either at the end of each year, or each day.

	2019	2020	2021
Offline method	0.179	0.278	0.334
GAM Kalman (Gaussian Quantiles)	0.161	0.163	0.148
GAM Kalman + QR Offline	**0.155**	**0.161**	**0.144**
GAM Kalman + QR OGD (10^{-2})	**0.155**	**0.155**	**0.141**
GAM Offline + QR OGD (10^{-3})	0.167	0.213	0.202
GAM Offline + QR OGD (10^{-2})	0.156	0.169	0.146
GAM Offline + QR OGD (10^{-1})	0.191	0.189	0.202

Table 10.2 – Average of the Ranked Probability Score on the 14 regions for the different models and the three test years.

the quantile level. The forecast of a q-quantile is expected to be empirically bigger than the quantity of interest for a fraction q of the data set and smaller for a fraction $1-q$. We display reliability diagrams in Figure 10.3. We observe that the offline model is not reliable, probably because of the bias in the mean model. The two adaptation methods proposed yield more reliable forecasts. On the one hand, when a Kalman filter adapts the mean model, we obtain reliable quantile forecasts either using Gaussian Kalman quantiles or offline quantile regressions. On the other hand, it is also sufficient to adapt the quantile regression model with OGD on the residuals of the offline GAM.

Numerical evaluation of forecasts is obtained by the pinball loss. However, we have 14 regions, and many quantile levels: as in Browell and Fasiolo, 2021 we use 0.0005, 0.001, 0.0025, 0.005, 0.01, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, and symmetrically for the right tail.

A way to combine the pinball losses at different quantile levels is to use the continuous ranked probability score (CRPS) (Gneiting and Raftery, 2007), defined equivalently by the two following expressions:

$$CRPS(F,y) = \int_{-\infty}^{+\infty} (F(x) - \mathbb{1}_{y\leq x})^2 dx = 2\int_0^1 \rho_q(y, F^{-1}(q))dq\,,$$

where y is the observation and F the cumulative distribution function. Remark that we have $\rho_0(y, F^{-1}(0)) = \rho_1(y, F^{-1}(1)) = 0$, therefore the CRPS is not a good performance indicator for the tail estimation.

We use the discrete approximation of last integral. We have a set of forecasted quantiles $\hat{y}_{q_1}, \dots, \hat{y}_{q_l}$, and we define the RPS as the integral of the piecewise linear function interpolating $0, \rho_{q_1}(y, \hat{y}_{q_1}), \dots \rho_{q_l}(y, \hat{y}_{q_l}), 0$ at the points $0, q_1, \dots q_l, 1$. It yields

$$RPS\big((\hat{y}_{q_1}, \dots, \hat{y}_{q_l}), y\big) = \sum_{i=1}^{l} \rho_{q_i}(y, \hat{y}_{q_i})(q_{i+1} - q_{i-1})\,,$$

where we define $q_0 = 0, q_{l+1} = 1$. Then we naturally define the RPS on a test set for the forecasts $(\hat{y}_{t,q_1}, \dots \hat{y}_{t,q_l})$ of (y_t) as the average of $RPS((\hat{y}_{t,q_1}, \dots \hat{y}_{t,q_l}), y_t)$. We display this RPS in Figure 10.4, and we provide in Table 10.2 the RPS averaged on the 14 regions for the different models. We obtain an important gain in RPS by adapting the GAM using the Kalman filter and keeping an offline quantile regression, even for the stable period (the year 2019). Adapting the quantile regression with an OGD is outperformed by this adaptation of the mean model. However, the difference is tenuous for a well-chosen gradient step size on which the OGD result crucially depends. Combining both levels of adaptation boosts the performances a little.

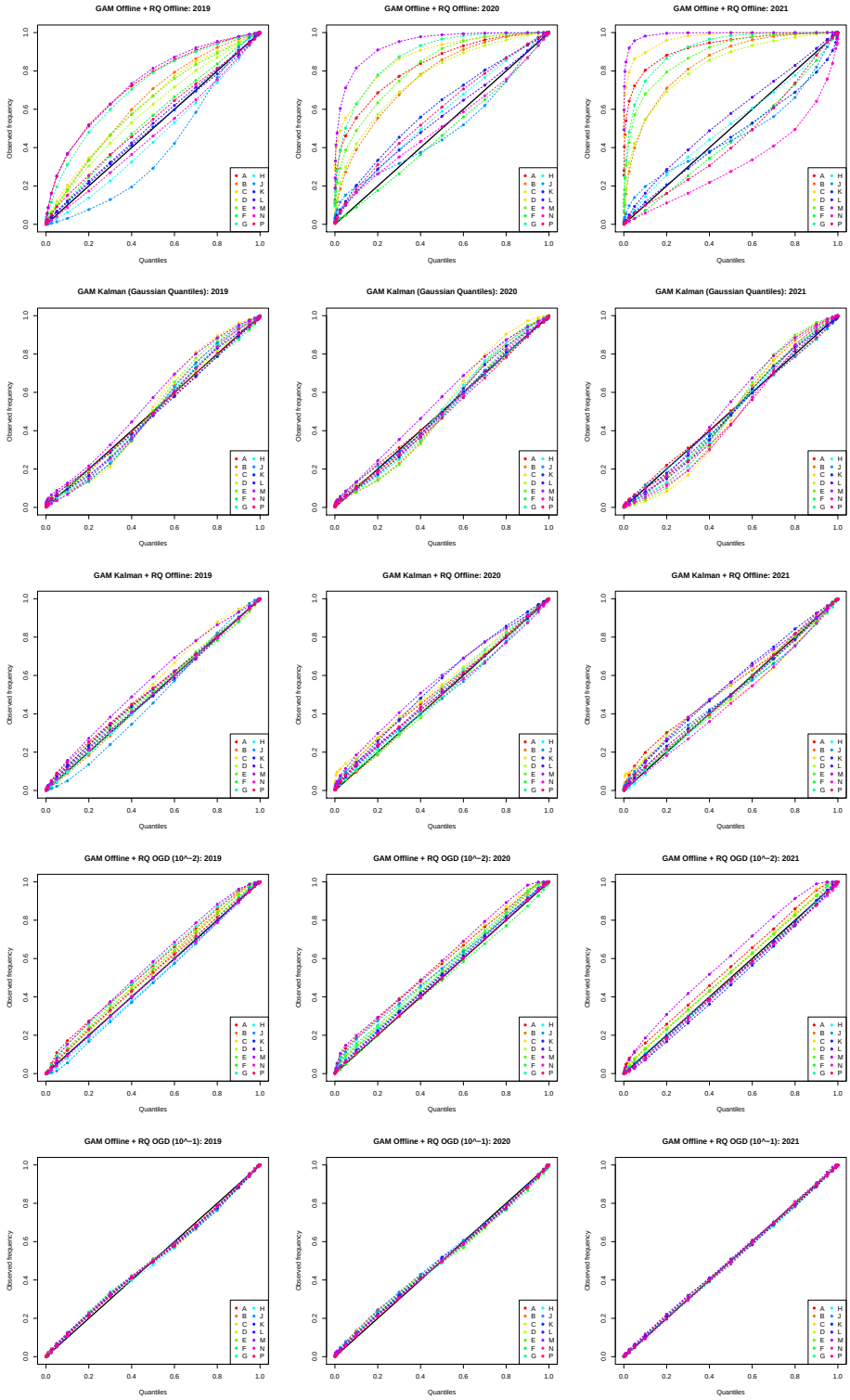


Figure 10.3 – Reliability diagrams for the different methods and the different years.

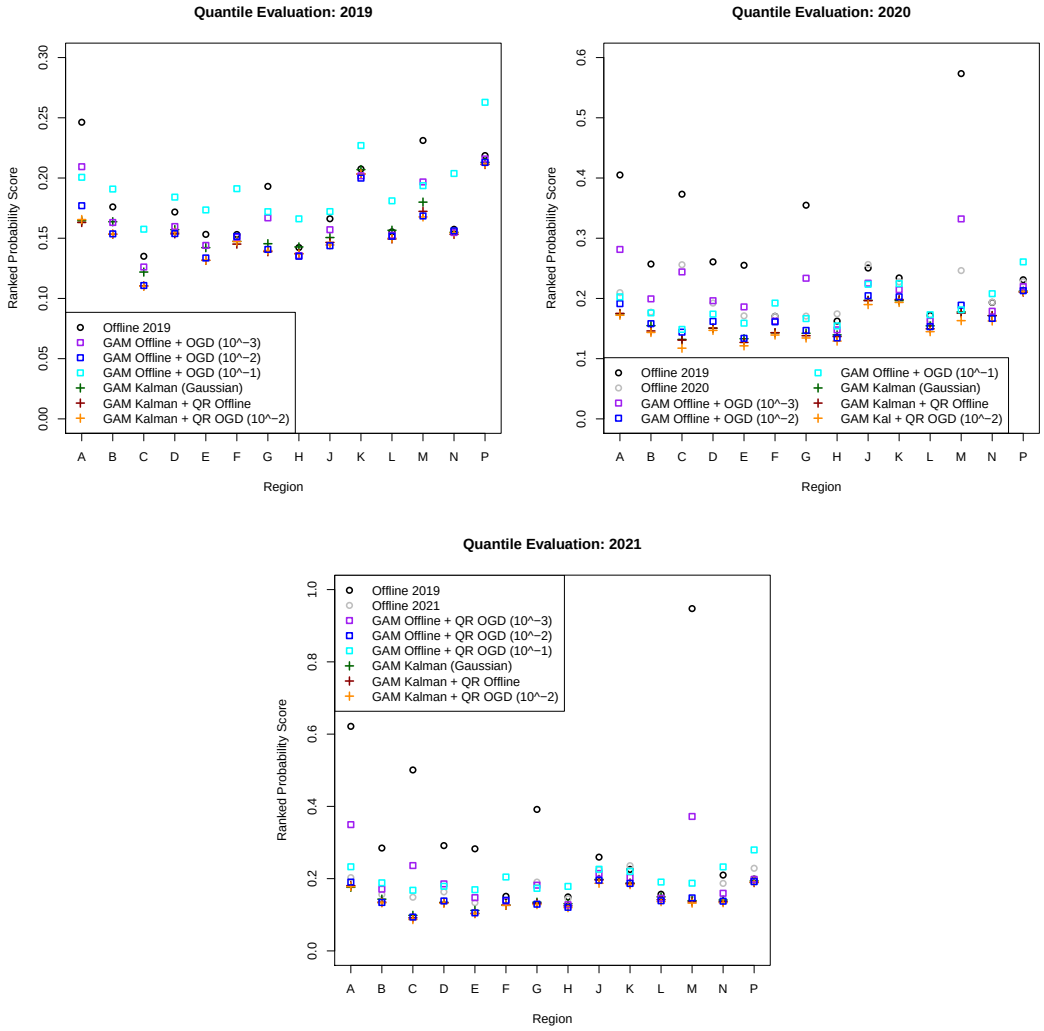


Figure 10.4 – Ranked Probability Score for each region in Great Britain.

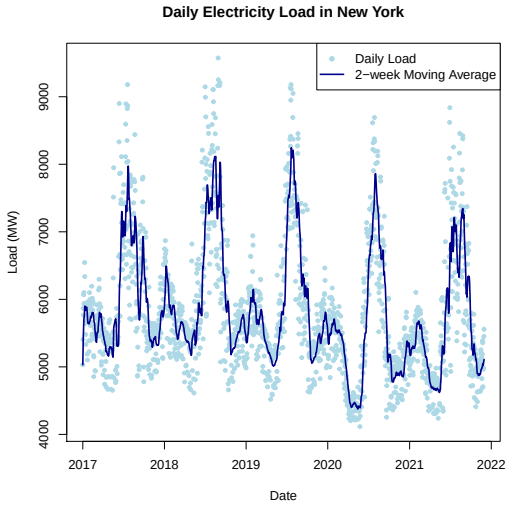


Figure 10.5 – Daily electricity load in New York.

10.3.4 Learning the Embedded Capacities

Another interesting advantage of model adaptation is that it reduces the need for good explanatory variables that can prove difficult to obtain. In particular, we show that our Kalman GAM can learn the embedded generation capacities, in the sense that removing these variables does not significantly change the predictions, contrary to the offline method.

Precisely, in the GAM and in the quantile regression model presented in Section 10.3.1, we remove the solar and wind generation capacities. We consider solar radiation only instead of its product with solar capacity. In the nonlinear GAM effect of the wind speed, we remove the dependence on the embedded wind capacity. We thus evaluate removing the capacities in the offline model, as well as in the Kalman adaptation of the GAM. We evaluate during 2019 to not include the coronavirus crisis.

For mean prediction, removing the capacities increases the RMSE of the offline method by more than 7%, while it reduces the RMSE by 0.2% for the Kalman adaptation of the GAM.

For probabilistic prediction, removing the capacities increases the RPS by more than 7% for the offline model and by less than 0.1% for the offline quantile regression on the residuals of GAM Kalman.

10.4 Load Forecasting in New York

We also test our framework on US data during the coronavirus crisis (Ruan et al., 2020).

10.4.1 Data Presentation and Offline Model

We focus on New York City and we consider daily data, represented in Figure 10.5. We display the dependence of the load on meteorological variables in Figure 10.6.

We follow the structure presented in Section 10.2. We forecast the mean with a GAM, then we fit quantile regressions on the residuals.

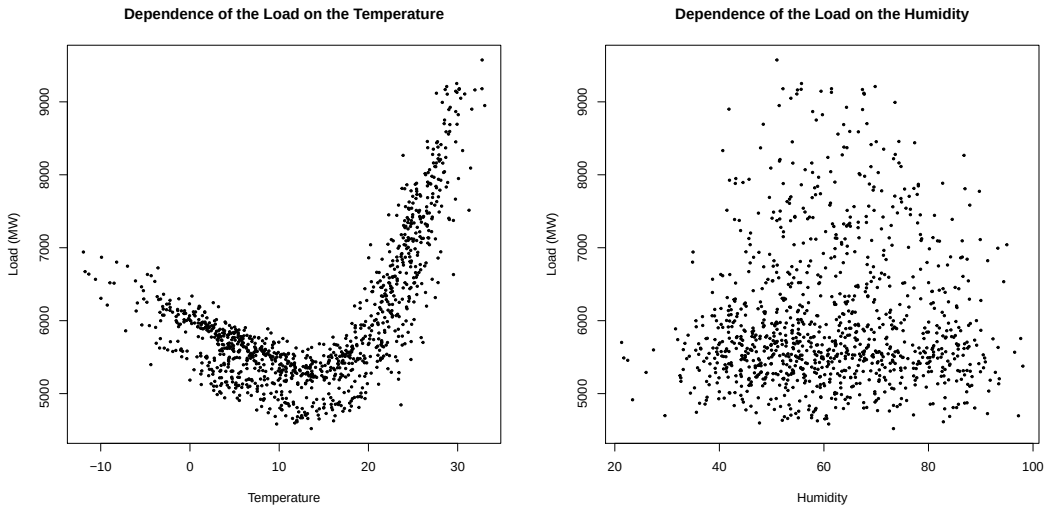


Figure 10.6 – Dependence of the load on the temperature and humidity covariates. Air-conditioning is more important than heating. It is the opposite in France.

— The generalized additive model we consider has the following form:

$$Load_t = \sum_{i=1}^{7} \alpha_i \mathbb{1}_{WeekDay_t=i} + \beta BH_t + \gamma WB_t + \delta LoadD_t + f_1(LoadW_t) + f_2(t) + f_3(Temp_t) + f_4(Hum_t) + f_5(Toy_t) + \varepsilon_t\,,$$

where ε_t is a Gaussian i.i.d. noise and for each day t,
— $WeekDay_t$ is the day of the week,
— BH_t (respectively WB_t) is a boolean denoting if the day is a bank holiday (respectively in the winter break),
— $LoadD_t$ and $LoadW_t$ are lags of the load with a one-day and one-week delays,
— t is the day (variable growing linearly with time),
— $Temp_t$ and Hum_t are the temperature and humidity,
— Toy_t is the time of year (variable growing linearly from 0 on January 1$^{\text{st}}$ to 1 on December 31$^{\text{st}}$).
The nonlinear effects f_1, f_2, f_3, f_4, f_5 are decomposed on a spline basis of thin plate splines for the first four, and of cubic cyclic splines for f_5, as the effet of the time of the year is cyclic.
— The quantile regression we fit on the residuals uses as covariates the (linear and nonlinear) effects of the GAM, as well as the squared prediction as in Section 10.3.

10.4.2 Performances of Mean Forecast

We display the evolution of the error in Figure 10.7. The error is evaluated through the RMSE. From January 1$^{\text{st}}$ 2020 to November 30$^{\text{th}}$ 2021, the offline GAM achieves 289 MW of RMSE. Applying the Kalman filter in the degenerate static setting ($Q_t = 0$) yields a decrease of 33% (195 MW RMSE). A dynamic Kalman filter, with a constant covariance matrix Q learned

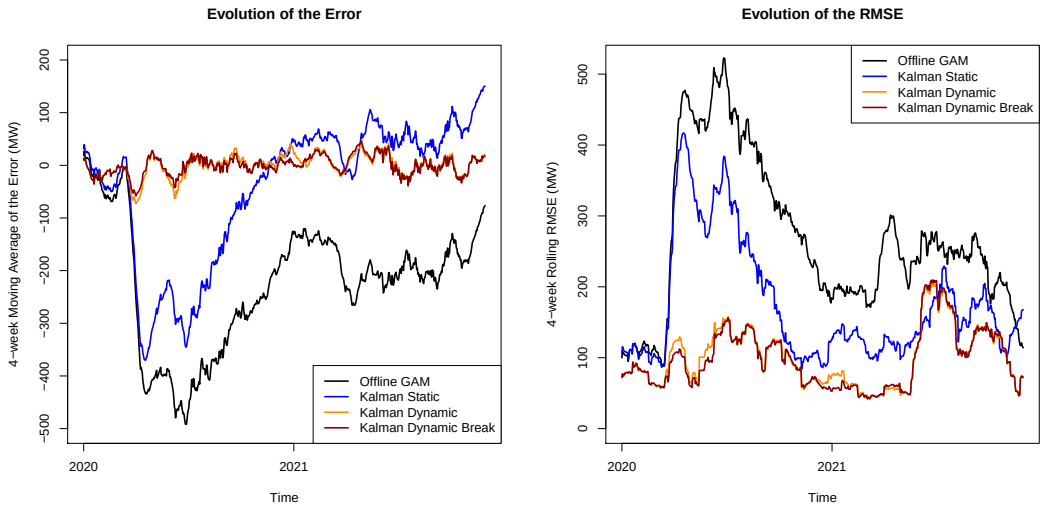


Figure 10.7 – Evolution of the error for the Kalman variants, compared to the base GAM. On the left: rolling version of the error. On the right: rolling version of the RMSE.

with the iterative grid search of Chapter 5 yields a decrease of 73% (108 MW RMSE). Introducing a break on March 16$^{\text{th}}$ as in Section 7.2.1 yields a small boost (106 MW RMSE). This boost essentially comes from a 15% decrease of the RMSE in the two months following the break introduced.

10.4.3 Performances of Quantile Forecast

We display in Figure 10.8 the reliability diagram for the different forecasts, as well as the pinball loss for different quantiles. We observe that contrary to the study on the GB data set, the offline quantile regression on the residuals of Kalman GAM does not yield well-calibrated forecasts. Gaussian quantiles of the Kalman filter yield better-calibrated forecasts, and the OGD method is also reliable.

The Gaussian quantiles of the Kalman filter are not far from achieving the best pinball loss. While the OGD adaptation of the quantile regression yields a smaller RPS, there is a significant sensibility to the gradient step size.

The RPS obtained is of 193 MW for the offline model, 56 MW for the Gaussian Kalman quantiles, 80 MW for the offline quantile regression on the residuals of Kalman GAM, 49 MW (resp. 47 MW) for the OGD adaptation of the quantile regression with best step size for the GAM (resp. Kalman GAM) residuals.

10.5 Conclusion and Future Work

In this chapter, we presented adaptive methods for probabilistic forecasting. The applications on Great Britain regional net-load and New York load show that state-space models might also be helpful for probabilistic forecasting. Indeed, the Kalman filter already provides a probabilistic forecast performing quite well, considering it relies on Gaussian quantiles. Quantile regressions on the residuals of the mean model improve the quantile forecasts. We can adapt these quantile

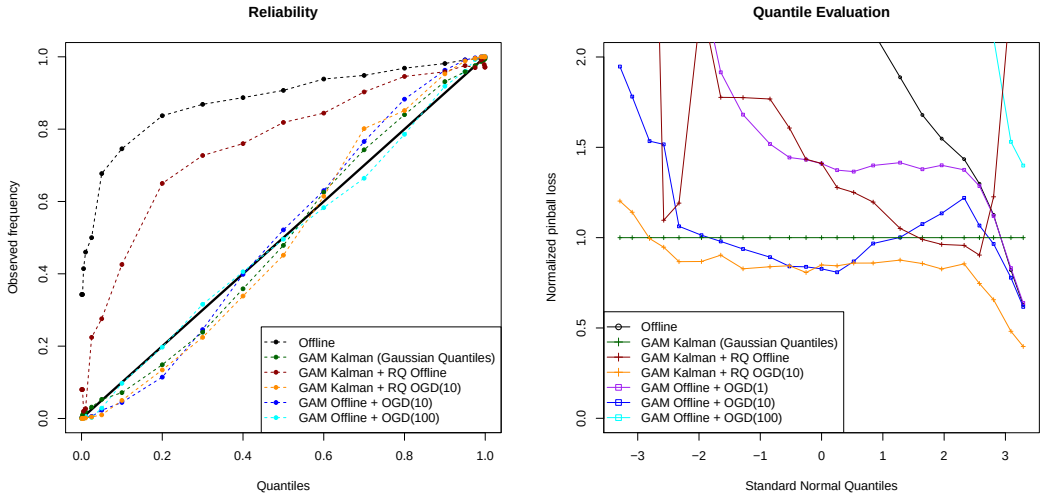


Figure 10.8 – On the left: reliability diagram for the different methods. On the right: pinball loss for the quantiles estimated. On the x-axis of the right graph we use the quantiles of the standard normal distribution (to be more readable in the extremes). It corresponds to the quantile levels 0.0005, 0.001, 0.0025, 0.005, 0.01, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5 and symmetrically.

regressions by online gradient descent, but it is very sensitive to the choice of the gradient step size.

We see two interesting leads for future research. The first one is expectile regression (Newey and Powell, 1987). Expectiles generalize the mean as quantiles generalize the median. While the mean is the minimum of the squared loss, expectiles minimize an asymmetric squared loss. We believe that adaptive expectile regression should be easier to obtain than adaptive quantile regression. Indeed, the squared loss is strongly convex while the absolute loss is not even strictly convex, and it should be possible to use second-order methods similar to the Kalman filter. The main issue concerns the expectile-to-quantile transform (Schnabel and Eilers, 2013).

The second is to study extreme quantiles. For the quantiles outside the $2.5\% - 97.5\%$ interval, Browell and Fasiolo, 2021 use the generalized Pareto distribution. Indeed, as extreme quantile regressions are not well-calibrated, the distribution they consider is piecewise linear using the quantile regressions between 2.5% and 97.5%, and the generalized Pareto distribution models the tails. Adaptive extreme estimation is difficult because of the inherent scarcity of extreme values. We believe that adapting the mean model would improve extreme forecasts in the same way it improves quantile regression forecasts.

Conclusion and Perspectives

In this thesis, we have presented several contributions to state-space estimation in the view of time series and the application of state-space models to the electricity market. Throughout the manuscript, we have highlighted the connection between Bayesian methods such as the Kalman filter and stochastic optimization algorithms. In Part I, we have provided an analysis of the extended Kalman filter in the degenerate static setting where the state is assumed constant and estimated recursively. We obtained a theoretical comparison between an extension of the Kalman filter and gradient descent algorithms of annealing step sizes. In Part II, we have detailed the critical issue state-space models encounter in most real-world applications, the estimation of the variances governing the dynamics of the underlying process. We extended the gradient descent parallel, interpreting the choice of the variances as the estimation of optimal gradient step size in stochastic optimization. In Part III, we have detailed the application of state-space models to electricity load forecasting. The objective we pursued was to demonstrate the generality of the approach to various countries, various scales and various objectives. We claim that the methods employed have not unleashed their full potential yet.

Estimating the state and space noise variances in a state-space model is challenging, especially in the adaptive fashion. It has occupied many researchers for decades, and there is still no consensus. The estimation algorithm we presented in Chapter 6, Viking, is a promising lead. However, we claim there is still much room for improvement. We applied the variational Bayes framework, and our approach crucially relies on the definition of the dynamical model we assume. Two leads seem particularly interesting to investigate. The first one is a better understanding of the representation of the state noise covariance matrix, leading to a better selection of the function applied to the Gaussian latent variable to recover the covariance matrix. A second opportunity for enhancement may come from a combined model on both variances. Indeed, it is intuitive that their evolution should not be independent. In our simulation, we observe that Viking tends to correlate their evolution anyway, so we concluded that it was robust to this hypothesis. Our heuristics for constant variances presented in Chapter 5 suggests the important parameter is the variance ratio Q/σ^2. In light of that analysis, it could be interesting to study a dynamical model with independent evolutions of σ_t^2 and Q_t/σ_t^2.

To apply state-space models to the problem of electricity load forecasting, we considered a hybrid approach. We first use statistical and machine learning methods. We linearize these models in order to apply linear Gaussian state-space models. We demonstrate that it achieves a nice compromise between the complex relations learned by machine learning models and the reactivity of state-space methods. This work considered simple offline models: mainly linear regression, generalized additive models, and multi-layer perceptron. It would be interesting to combine state-space models with deeper neural networks. Indeed, deep learning is known to capture complex relations, and deep neural networks may provide better features.

While we focused on mean forecasting in most of Part III, probabilistic forecasting has much interest in practice, and the approach has potential in this domain. Expectiles seem an interesting

lead to obtain second-order methods similar to the Kalman filter or Viking. Also, as the Kalman filter already provides a probabilistic forecast performing quite well, improvements on Viking may turn Viking into a method performing well in the task of probabilistic forecasting.

Finally, while this thesis spotlights the electricity load forecasting task, we believe the methods developed are pretty general. It would be interesting to apply them to other time series forecasting problems.

Appendix A

Supplementary Material for Chapter 3

We provide the proofs for all the claims of Chapter 3.

Contents

A.1 Proof of Proposition 3.1

Proof of Proposition 3.1. Provided that $\|\theta - \theta^\star\| \leq \varepsilon$ we have

$$\begin{aligned}
\|\nabla L(\theta)\|^2 &\geq \frac{1}{\|\theta - \theta^\star\|^2}(\nabla L(\theta)^\top(\theta - \theta^\star))^2 \\
&\geq \frac{1}{\|\theta - \theta^\star\|^2}\Big(L(\theta) - L(\theta^\star) + \frac{\mu_\varepsilon}{2}\|\theta - \theta^\star\|^2\Big)^2 \\
&\geq \frac{4}{\|\theta - \theta^\star\|^2}(L(\theta) - L(\theta^\star))\frac{\mu_\varepsilon}{2}\|\theta - \theta^\star\|^2 \\
&= 2\mu_\varepsilon(L(\theta) - L(\theta^\star)) \qquad \text{(A.1)}
\end{aligned}$$

The first inequality is Cauchy-Schwarz inequality, the second is Assumption 3.5 and the third line comes from $(a+b)^2 \geq 4ab$ for any $a, b \in \mathbb{R}$.

Moreover, using again Assumption 3.5:

$$\|\theta - \theta^\star\| = \varepsilon \implies L(\theta) - L(\theta^\star) \geq \frac{\mu_\varepsilon}{2}\|\theta - \theta^\star\|^2 ,$$

therefore Equation (A.1) yields

$$\|\theta - \theta^\star\| = \varepsilon \implies \|\nabla L(\theta)\|^2 \geq \mu_\varepsilon^2 \varepsilon^2 .$$

The convexity of L yields

$$\|\theta - \theta^\star\| > \varepsilon \implies \|\nabla L(\theta)\|^2 \geq \mu_\varepsilon^2 \varepsilon^2 \,.$$

To conclude, let $\eta > 0$, and $\theta \in \mathbb{R}^d$ such that $L(\theta) - L(\theta^\star) > \eta/2$. On the one hand, if $\|\theta - \theta^\star\| > \varepsilon$ we have $\|\nabla L(\theta)\|^2 \geq \mu_\varepsilon^2 \varepsilon^2$. On the other hand, if $\|\theta - \theta^\star\| \leq \varepsilon$, Equation (A.1) yields $\|\nabla L(\theta)\|^2 \geq \mu_\varepsilon \eta$. □

A.2 Proof of Corollary 3.1

Proof of Corollary 3.1. $\beta = 3/4$ yields

$$\frac{2(1-\beta)}{(1-(1/2)^{1-\beta})} \approx 3.14 \leq 4\,, \qquad \Big(\frac{1-(1/2)^{1-\beta}}{2(1-\beta)}\Big)^2 \approx 0.101 \geq 1/10$$

We therefore apply Theorem 3.3 with these bounds. For any η and

$$t \geq \max\Big(\Big(\frac{4}{\eta\mu}\Big(4g^2h + L(\theta_1) - L(\theta^\star)\Big)\Big)^4, \quad 2 + 2\Big(\frac{4g^2h}{\eta}\Big)^2\Big)\,,$$

it holds:

$$\begin{aligned}\mathbb{P}\left(L(\theta_t) - L(\theta^\star) > \eta\right) \leq (1+t/2)\exp\Big(-\eta^2\mu^2 t^{1/2}\frac{1}{320g^4}\Big) \\ + (1+t/2)\exp\Big(-\eta^2\frac{(t/2-2)^{1/2}}{256g^4}\Big)\,.\end{aligned}$$

If $t \geq 8$ then $t/2 - 2 \geq t/4$ and we obtain

$$\mathbb{P}\left(L(\theta_t) - L(\theta^\star) > \eta\right) \leq (t+2)\exp\Big(-\eta^2 t^{1/2}\frac{\min(\mu^2, 1/2)}{320g^4}\Big)\,.$$

Therefore, fixing $t \geq 8$ and $0 < \delta \leq 1$, we define $\eta_t = \sqrt{\frac{320g^4(\ln\delta^{-1} + \ln(t+2))}{t^{1/2}\mu^2}}$, and we obtain $\mathbb{P}\left(L(\theta_t) - L(\theta^\star) > \eta_t\right) \leq \delta$ provided that

$$t \geq \max\Big(\Big(\frac{4}{\eta_t\mu}\Big(4g^2h + L(\theta_1) - L(\theta^\star)\Big)\Big)^4, \quad 2 + 2\Big(\frac{4g^2h}{\eta_t}\Big)^2\Big)\,.$$

We write sufficient conditions such that last equation holds.

— We have

$$\begin{aligned}\Big(\frac{4}{\eta_t\mu}\Big(4g^2h + L(\theta_1) - L(\theta^\star)\Big)\Big)^4 &= \frac{t\mu^4 256\Big(4g^2h + L(\theta_1) - L(\theta^\star)\Big)^4}{320^2 g^8(\ln\delta^{-1} + \ln(t+2))^2\mu} \\ &\leq \frac{1}{2}t\mu^3\Big(h + \frac{L(\theta_1) - L(\theta^\star)}{4g^2}\Big)\,,\end{aligned}$$

because $\ln\delta^{-1} \geq 0$ and $\ln(t+2) \geq \ln 10 \geq 2$. Therefore if $\mu^3 \leq \Big(2h + \frac{L(\theta_1)-L(\theta^\star)}{2g^2}\Big)^{-1}$ we obtain $t \geq \Big(\frac{4}{\eta_t\mu}\Big(4g^2h + L(\theta_1) - L(\theta^\star)\Big)\Big)^4$.

— Furthermore, we write:

$$2 + 2\Big(\frac{4g^2h}{\eta_t}\Big)^2 = 2 + t^{1/2}\mu^2 \frac{32g^2h}{320g^4(\ln\delta^{-1} + \ln(t+2))} \\ \leq 2 + t\mu^2\frac{h}{20g^2}\,.$$

Therefore if $\mu^2 \leq \frac{10g^2}{h}$, as $t \geq 8$, we have $t \geq t/2 + 2$ and $t \geq 2 + 2\Big(\frac{4g^2h}{\eta_t}\Big)^2$.

To summarize, we have obtained for $t \geq 8$ and $0 < \delta \leq 1$, that with probability at least $1 - \delta$ it holds:

$$L(\theta_t) - L(\theta^\star) \leq \sqrt{\frac{320g^4(\ln\delta^{-1} + \ln(t+2))}{t^{1/2}}}\frac{1}{\mu}\,,$$

as long as $\mu^3 \leq \Big(2h + \frac{L(\theta_1) - L(\theta^\star)}{4g^2}\Big)^{-1}$ and $\mu^2 \leq \frac{10g^2}{h}$. If we define

$$\mu_0 = \min\Big(\mu, \Big(2h + \frac{L(\theta_1) - L(\theta^\star)}{4g^2}\Big)^{-1/3}, \Big(\frac{10g^2}{h}\Big)^{1/2}\Big) \\ = \max\Big(\mu^{-1}, \Big(2h + \frac{L(\theta_1) - L(\theta^\star)}{4g^2}\Big)^{1/3}, \Big(\frac{h}{10g^2}\Big)^{1/2}\Big)^{-1}\,,$$

then Assumption 3.4 is satisfied with $D_\eta^2 = \eta\mu_0$, and the result follows, remarking $\sqrt{320} \approx 17.9 \leq 18$ and $\sqrt{\ln\delta^{-1} + \ln(t+2)} \leq \sqrt{\ln\delta^{-1}} + \sqrt{\ln(t+2)}$. □

A.3 Proofs for the Averaged Stochastic Gradient Descent

Proof of Lemma 3.2. We apply the standard approach for averaged SGD. We start from the recursive update: for any $t \geq 1$ we have $\theta_{t+1} = \theta_t - \gamma_t\nabla\ell_t(\theta_t)$ and therefore

$$\|\theta_{t+1} - \theta\|^2 = \|\theta_t - \theta\|^2 - \frac{2}{t^\beta}\nabla\ell_t(\theta_t)^\top(\theta_t - \theta) + \frac{1}{t^{2\beta}}\|\nabla\ell_t(\theta_t)\|^2\,.$$

Defining $\Delta N_t = (\nabla L(\theta_t) - \nabla\ell_t(\theta_t))^\top(\theta_t - \theta)$, we obtain

$$\nabla L(\theta_t)^\top(\theta_t - \theta) = \Delta N_t + \nabla\ell_t(\theta_t)^\top(\theta_t - \theta) \\ = \Delta N_t + \frac{t^\beta}{2}(\|\theta_t - \theta\|^2 - \|\theta_{t+1} - \theta\|^2) + \frac{1}{2t^\beta}\|\nabla\ell_t(\theta_t)\|^2\,.$$

Summing from k to n yields

$$\sum_{t=k}^n \nabla L(\theta_t)^\top(\theta_t - \theta) = \sum_{t=k}^n \Delta N_t + \frac{1}{2}\sum_{t=k}^n \|\theta_t - \theta\|^2(t^\beta - (t-1)^\beta) - \frac{n^\beta}{2}\|\theta_{n+1} - \theta\|^2 \\ + \frac{1}{2}\sum_{t=k}^n \frac{\|\nabla\ell_t(\theta_t)\|^2}{t^\beta} \\ \leq \sum_{t=k}^n \Delta N_t + \frac{1}{2}\sum_{t=k}^n \|\theta_t - \theta\|^2 t^\beta(1 - (1-1/t)^\beta) + \frac{1}{2}\sum_{t=k}^n \frac{\|\nabla\ell_t(\theta_t)\|^2}{t^\beta}\,.$$

We then use Bernoulli's inequality $(1-1/t)^\beta \geq 1-\beta/t$. Plugging it into last inequality, we obtain

$$\sum_{t=k}^{n} \nabla L(\theta_t)^\top (\theta_t - \theta) \leq \sum_{t=k}^{n} \Delta N_t + \frac{1}{2}\sum_{t=k}^{n} \|\theta_t - \theta\|^2 \frac{\beta}{t^{1-\beta}} + \frac{1}{2}\sum_{t=k}^{n} \frac{\|\nabla \ell_t(\theta_t)\|^2}{t^\beta}\,.$$

□

Proof of Corollary 3.2. We use a union bound for any k:

$$\mathbb{P}\left(\bigcup_{t=k+1}^{\infty} (\|\theta_t - \theta^\star\| > \varepsilon)\right) \leq \sum_{t=k+1}^{\infty} \mathbb{P}(\|\theta_t - \theta^\star\| > \varepsilon)\,.$$

Assumption 3.5 yields $\|\theta - \theta^\star\| > \varepsilon \implies L(\theta_t) - L(\theta^\star) > \frac{\mu_\varepsilon \varepsilon^2}{2}$. Thus, for any t, it holds:

$$\mathbb{P}(\|\theta_t - \theta^\star\| > \varepsilon) \leq \mathbb{P}\Big(L(\theta_t) - L(\theta^\star) > \frac{\mu_\varepsilon \varepsilon^2}{2}\Big)\,.$$

We apply Theorem 3.3 with $D_{(\mu_\varepsilon \varepsilon^2)/2} = \sqrt{\min(\mu_\varepsilon^2 \varepsilon^2, \mu_\varepsilon^2 \varepsilon^2/2)} = \sqrt{\mu_\varepsilon^2 \varepsilon^2/2}$. As in the proof of Corollary 3.1, we observe that with $\beta = 3/4$ we have

$$\frac{2(1-\beta)}{(1-(1/2)^{1-\beta})} \approx 3.14 \leq 4\,, \qquad \Big(\frac{1-(1/2)^{1-\beta}}{2(1-\beta)}\Big)^2 \approx 0.101 \geq 1/10\,.$$

It yields, as long as

$$t \geq \max\Big(\Big(\frac{8}{\mu_\varepsilon^2 \varepsilon^2}\Big(4g^2 h + L(\theta_1) - L(\theta^\star)\Big)\Big)^4, \quad 2 + 2\Big(\frac{8g^2 h}{\mu_\varepsilon \varepsilon^2}\Big)^2\Big)\,,$$

it holds:

$$\mathbb{P}\left(L(\theta_t) - L(\theta^\star) > \frac{\mu_\varepsilon \varepsilon^2}{2}\right) \leq (1+t/2)\exp\Big(-\mu_\varepsilon^4 \varepsilon^4 t^{1/2} \frac{1}{1280 g^4}\Big) + (1+t/2)\exp\Big(-\mu_\varepsilon^2 \varepsilon^4 \frac{(t/2-2)^{1/2}}{1024 g^4}\Big)\,.$$

If $t \geq 8$ then $t/2 - 2 \geq t/4$ and as long as $\mu_\varepsilon \varepsilon^2 \leq 1/2$ we obtain

$$\mathbb{P}\left(L(\theta_t) - L(\theta^\star) > \frac{\mu_\varepsilon \varepsilon^2}{2}\right) \leq (t+2)\exp\Big(-\mu_\varepsilon^4 \varepsilon^4 t^{1/2} \frac{1}{1280 g^4}\Big)\,.$$

For any $a > 0$ and $t \geq (3/a)^2$ we have $\exp(-a t^{1/2}) \leq 1/t^{3/2}$. Therefore, for $t \geq \Big(\frac{10g}{\mu_\varepsilon \varepsilon}\Big)^8$ we obtain

$$\mathbb{P}\left(L(\theta_t) - L(\theta^\star) > \frac{\mu_\varepsilon \varepsilon^2}{2}\right) \leq \frac{t+2}{t^{3/2}}\exp\Big(-\mu_\varepsilon^4 \varepsilon^4 t^{1/2} \frac{1}{2560 g^4}\Big) \leq \frac{2}{t^{1/2}}\exp\Big(-\mu_\varepsilon^4 \varepsilon^4 t^{1/2} \frac{1}{2560 g^4}\Big)\,.$$

Moreover

$$\sum_{t=k+1}^{\infty} \mathbb{P}\left(L(\theta_t) - L(\theta^\star) > \frac{\mu_\varepsilon \varepsilon^2}{2}\right) \leq \int_k^\infty \frac{2}{u^{1/2}} \exp\Big(-\mu_\varepsilon^4 \varepsilon^4 u^{1/2} \frac{1}{2560 g^4}\Big) du$$
$$= \frac{5120 g^4}{\mu_\varepsilon^4 \varepsilon^4} \exp\Big(-\mu_\varepsilon^4 \varepsilon^4 k^{1/2} \frac{1}{2560 g^4}\Big).$$

For $k = \left(\frac{\ln \delta^{-1} + \ln\left(5120 g^4/(\mu_\varepsilon^4 \varepsilon^4)\right)}{\mu_\varepsilon^4 \varepsilon^4/(2560 g^4)}\right)^2$ we obtain $\mathbb{P}\left(\bigcup_{t=k+1}^{\infty} (\|\theta_t - \theta^\star\| > \varepsilon)\right) \leq \delta$. The result follow from $\ln 5120 \leq 9$. □

Proof of Theorem 3.4. Let $k \geq 1$ such that $\mathbb{P}(\cap_{t=k+1}^{\infty}(\|\theta_t - \theta^\star\| \leq \varepsilon)) \geq 1 - \delta$. Thanks to Assumption 3.5, it holds with probability $1 - \delta$ that

$$\sum_{t=k+1}^{n} (L(\theta_t) - L(\theta^\star)) \leq \sum_{t=k+1}^{n} \Big(\nabla L(\theta_t)^\top (\theta_t - \theta^\star) - \frac{\mu_\varepsilon}{2} \|\theta_t - \theta^\star\|^2\Big).$$

The rest of the proof consists in controlling the first-order term by $\|\theta_t - \theta^\star\|^2$ with constants smaller than $\mu_\varepsilon/2$. We apply Lemma 3.2:

$$\sum_{t=k+1}^{n} (L(\theta_t) - L(\theta^\star)) \leq \sum_{t=k+1}^{n} \Big(\Delta N_t + \frac{1}{2}\|\theta_t - \theta^\star\|^2 \Big(\frac{\beta}{t^{1-\beta}} - \mu_\varepsilon\Big) + \frac{\|\nabla \ell_t(\theta_t)\|^2}{2t^\beta}\Big), \tag{A.2}$$

We apply Lemma B.1 from Bercu and Touati, 2008 in order to control the martingale difference: for any $n \geq k$ and any $\lambda > 0$, it holds:

$$\mathbb{E}\Big[\exp\Big(\lambda \sum_{t=k+1}^{n} \Delta N_t - \frac{\lambda^2}{2} \sum_{t=k+1}^{n} (\Delta N_t^2 + \mathbb{E}[\Delta N_t^2 \mid \mathcal{F}_{t-1}])\Big)\Big] \leq 1, \tag{A.3}$$

where $(\mathcal{F}_t)$ is the natural filtration $(\sigma(\nabla \ell_1, \ldots, \nabla \ell_t))_t$.

We apply Markov inequality: for any $\lambda, \delta > 0$, we have

$$\mathbb{P}\Big(\sum_{t=k+1}^{n} \Delta N_t > \frac{\lambda}{2} \sum_{t=k+1}^{n} (\Delta N_t^2 + \mathbb{E}[\Delta N_t^2 \mid \mathcal{F}_{t-1}]) + \frac{\ln \delta^{-1}}{\lambda}\Big)$$
$$= \mathbb{P}\Big(\exp\Big(\lambda \sum_{t=k+1}^{n} \Delta N_t - \frac{\lambda^2}{2} \sum_{t=k+1}^{n} (\Delta N_t^2 + \mathbb{E}[\Delta N_t^2 \mid \mathcal{F}_{t-1}])\Big) > \delta^{-1}\Big)$$
$$\leq \frac{1}{\delta^{-1}} \mathbb{E}\Big[\exp\Big(\lambda \sum_{t=k+1}^{n} \Delta N_t - \frac{\lambda^2}{2} \sum_{t=k+1}^{n} (\Delta N_t^2 + \mathbb{E}[\Delta N_t^2 \mid \mathcal{F}_{t-1}])\Big)\Big] \leq \delta,$$

where the last inequality is provided by (A.3). The square of the martingale difference is controlled by the second-order term that we seek: for any t, it holds:

$$\Delta N_t^2 \leq 4g^2 \|\theta_t - \theta^\star\|^2, \qquad \mathbb{E}[\Delta N_t^2 \mid \mathcal{F}_{t-1}] \leq 4g^2 \|\theta_t - \theta^\star\|^2.$$

Therefore, for any $\lambda, \delta > 0$, it holds

$$\sum_{t=k+1}^{n} \Delta N_t \leq 4\lambda g^2 \sum_{t=k+1}^{n} \|\theta_t - \theta^\star\|^2 + \frac{\ln \delta^{-1}}{\lambda}\,,$$

with probability at least $1 - \delta$. This yields the desired control for the martingale difference in Equation (A.2). The last term is controlled thanks to Assumption 3.6.

We now combine our findings. For any $\lambda, \delta > 0$, it holds

$$\sum_{t=k+1}^{n} (L(\theta_t) - L(\theta^\star)) \leq \frac{\ln \delta^{-1}}{\lambda} + \sum_{t=k+1}^{n} \Big(4\lambda g^2 + \frac{\beta}{2t^{1-\beta}} + \frac{C_{\mathrm{Lip}}^2}{2t^\beta} - \frac{\mu_\varepsilon}{2}\Big)\|\theta_t - \theta^\star\|^2\,,$$

with probability at least $1 - 2\delta$. The final step is the choice of λ, k such that

$$4\lambda g^2 + \frac{\beta}{2k^{1-\beta}} + \frac{C_{\mathrm{Lip}}^2}{2k^\beta} - \frac{\mu_\varepsilon}{2} \leq 0\,.$$

We observe that this inequality is satisfied if the following holds:

$$4\lambda g^2 \leq \frac{\mu_\varepsilon}{4}, \qquad \frac{\beta}{2k^{1-\beta}} \leq \frac{\mu_\varepsilon}{8}, \qquad \frac{C_{\mathrm{Lip}}^2}{2k^\beta} \leq \frac{\mu_\varepsilon}{8}\,.$$

This motivates the definition of $\lambda = \frac{\mu_\varepsilon}{16 g^2}$. We observe that $\frac{\beta}{2k^{1-\beta}} \leq \frac{1}{2k^\beta}$ because $1/2 < \beta < 1$, and therefore $k \geq (4\max(1, C_{\mathrm{Lip}}^2)/\mu_\varepsilon)^{1/\beta}$ is sufficient.

We conclude with Equation (3.2). □

Appendix B

Supplementary Material for Chapter 4

The Appendix follows the structure of the article:

— Appendix B.1 presents the EKF for generalized linear models.
— Appendix B.2 contains the proofs of Section 4.3. Precisely, Lemma 4.1 is proved in Section B.2.1, the intermediate results of Sections 4.3.1 and 4.3.2 are proved in Sections B.2.2 and B.2.3, then Theorem 4.1 is proved in Section B.2.4 and Theorem 4.2 in Section B.2.5.
— Appendix B.3 contains the proofs of Section 4.4. We derive the global bound (Theorem 4.3) in Section B.3.1, then we obtain the concentration result on P_t in Section B.3.2, and finally we prove the convergence of the truncated algorithm in Section B.3.3.
— Appendix B.4 contains the proofs of Section 4.5. We prove Theorem 4.5 in Section B.4.1 and then in Section B.4.2 we prove the convergence of the algorithm, and we define an explicit value of $T(\varepsilon, \delta)$ satisfying Assumption 4.5.

Contents

B.1 Derivation of the Static EKF for Generalized Linear Models

As in Section 10.2 of Durbin and Koopman, 2012 we consider the following state-space model:

$$y_t = Z_t(\theta_t) + \varepsilon_t\,,$$
$$\theta_{t+1} = T_t(\theta_t) + \eta_t\,.$$

where ε_t and η_t are independent with mean zero and variances $h_t(\theta_t), Q_t(\theta_t)$. The state-space version of equation (4.2) is

$$p(y_t \mid X_t) = h(y_t)\exp\left(\frac{y_t\theta_t^\top X_t - b(\theta_t^\top X_t)}{a}\right).$$

The preceding equation matches the space equation form with $Z_t(\theta_t) = b'(\theta_t^\top X_t)$ and $h_t(\theta_t) = ab''(\theta_t^\top X_t)$. Thus we can write the EKF as follows (see Equation 10.4 of Durbin and Koopman, 2012): denoting by $\dot{T}_t$ the derivative of T_t,

$$\begin{aligned}
v_t &= y_t - b'(\hat{\theta}_t^\top X_t)\,, & F_t &= X_t^\top P_t X_t b''(\hat{\theta}_t^\top X_t)^2 + ab''(\hat{\theta}_t^\top X_t)\,,\\
\hat{\theta}_{t|t} &= \hat{\theta}_t + P_t X_t b''(\hat{\theta}_t^\top X_t)F_t^{-1}v_t\,, & P_{t|t} &= P_t - P_t X_t F_t^{-1} X_t^\top P_t b''(\hat{\theta}_t^\top X_t)^2\,,\\
\hat{\theta}_{t+1} &= T_t(\hat{\theta}_{t|t})\,, & P_{t+1} &= \dot{T}_t P_{t|t}\dot{T}_t^{\top} + Q_t(\hat{\theta}_{t|t})\,.
\end{aligned}$$

We focus on the static setting where the state equation becomes $\theta_{t+1} = \theta_t$, thus we have $\hat{\theta}_{t+1} = \hat{\theta}_{t|t}$ and $P_{t+1} = P_{t|t}$. We rewrite the update on P_t as follows:

$$P_{t+1} = P_t - \frac{P_t X_t X_t^\top P_t b''(\hat{\theta}_t^\top X_t)/a}{X_t^\top P_t X_t b''(\hat{\theta}_t^\top X_t)/a + 1}\,.$$

Moreover we have $P_{t+1}X_t = P_t X_t F_t^{-1} a b''(\hat{\theta}_t^\top X_t)$ thus we can rewrite the update on $\hat{\theta}_t$ as follows:

$$\hat{\theta}_{t+1} = \hat{\theta}_t + \frac{P_{t+1}X_t(y_t - b'(\hat{\theta}_t^\top X_t)}{a}\,.$$

This yields Algorithm 1.

B.2 Proofs of Section 4.3

B.2.1 Proof of Lemma 4.1

We prove the following Lemma inspired by the stopping time technique of Freedman, 1975 from which we derive Lemma 4.1. We give a general form useful in several proofs.

Lemma B.1. *Let $(\mathcal{F}_n)$ be a filtration, and we consider a sequence of events (A_n) that is adapted to $(\mathcal{F}_n)$. Let (V_n) be a sequence of random variables adapted to $(\mathcal{F}_n)$ satisfying $V_0 = 1$, $V_n \geq 0$ almost surely for any n, and*

$$\mathbb{E}[V_n \mid \mathcal{F}_{n-1}, A_{n-1}] \leq V_{n-1}, \qquad n \geq 1.$$

Then for any $\delta > 0$, it holds

$$\mathbb{P}\left(\left(\bigcup_{n=1}^{\infty} V_n > \delta^{-1}\right) \cup \left(\bigcup_{n=0}^{\infty} \overline{A_n}\right)\right) \leq \delta + \mathbb{P}\left(\bigcup_{n=0}^{\infty} \overline{A_n}\right) .$$

An important particular case is when (V_n) is a super-martingale adapted to the filtration $(\mathcal{F}_n)$ satisfying $V_0 = 1$ and $V_n \geq 0$ almost surely: then we have simultaneously $V_n \leq \delta^{-1}$ for $n \geq 1$ with probability larger than $1 - \delta$.

Proof. We define

$$E_k = \bigcup_{n=1}^{k} \left(V_n > \delta^{-1} \cup \overline{A_{n-1}}\right) .$$

As (E_k) is increasing, we have, for any $k \geq 1$,

$$\begin{aligned}\mathbb{P}(E_k) &= \sum_{n=1}^{k} \mathbb{P}\left(E_n \cap \overline{E_{n-1}}\right) \\ &= \sum_{n=1}^{k} \mathbb{P}\left(\overline{A_{n-1}} \cap \overline{E_{n-1}}\right) + \sum_{n=1}^{k} \mathbb{P}\left(V_n > \delta^{-1} \cap \overline{E_{n-1}} \cap A_{n-1}\right) .\end{aligned}$$

First, we have

$$\sum_{n=1}^{k} \mathbb{P}\left(\overline{A_{n-1}} \cap \overline{E_{n-1}}\right) \leq \mathbb{P}\left(\bigcup_{n=0}^{k-1} \overline{A_n}\right) .$$

Second, we apply Markov's inequality:

$$\begin{aligned}\sum_{n=1}^{k} \mathbb{P}\left(V_n > \delta^{-1} \cap \overline{E_{n-1}} \cap A_{n-1}\right) &\leq \sum_{n=1}^{k} \mathbb{E}\left[\frac{V_n}{\delta^{-1}} \mathbb{1}_{E_n \cap \overline{E_{n-1}} \cap A_{n-1}}\right] \\ &= \delta \sum_{n=1}^{k} \mathbb{E}\left[V_n (\mathbb{1}_{\overline{E_{n-1}} \cap A_{n-1}} - \mathbb{1}_{\overline{E_n}})\right] \\ &= \delta \sum_{n=1}^{k} \left(\mathbb{E}\left[V_n \mathbb{1}_{\overline{E_{n-1}} \cap A_{n-1}}\right] - \mathbb{E}\left[V_n \mathbb{1}_{\overline{E_n}}\right]\right) .\end{aligned}$$

The second line is obtained since $\overline{E_n} \subset \left(\overline{E_{n-1}} \cap A_{n-1}\right)$. According to the tower property and the super-martingale assumption,

$$\begin{aligned}\mathbb{E}\left[V_n \mathbb{1}_{\overline{E_{n-1}} \cap A_{n-1}}\right] &= \mathbb{E}\left[\mathbb{E}[V_n \mid \mathcal{F}_{n-1}, A_{n-1}] \mathbb{1}_{\overline{E_{n-1}} \cap A_{n-1}}\right] \\ &\leq \mathbb{E}\left[\mathbb{E}[V_n \mid \mathcal{F}_{n-1}, A_{n-1}] \mathbb{1}_{\overline{E_{n-1}}}\right] \\ &\leq \mathbb{E}\left[V_{n-1} \mathbb{1}_{\overline{E_{n-1}}}\right] .\end{aligned}$$

Therefore, a telescopic argument along with $V_0 = 1$ and $V_k \mathbb{1}_{\overline{E_k}} \geq 0$ yields

$$\sum_{n=1}^{k} \mathbb{P}\left(V_n > \delta^{-1} \cap \overline{E_{n-1}} \cap A_{n-1}\right) \leq \delta\,.$$

Finally, for any $k \geq 1$, we obtain

$$\mathbb{P}\left(E_k\right) \leq \mathbb{P}\left(\bigcup_{n=0}^{k-1} \overline{A_n}\right) + \delta$$

and the desired result follows by letting $k \to \infty$. □

Proof of Lemma 4.1. Let $\lambda > 0$. For any $n \geq 1$, we define

$$V_n = \exp\left(\sum_{t=k+1}^{k+n} \left(\lambda \Delta N_t - \frac{\lambda^2}{2}((\Delta N_t)^2 + \mathbb{E}[(\Delta N_t)^2 \mid \mathcal{F}_{t-1}])\right)\right).$$

Lemma B.1 of Bercu and Touati, 2008 states that (V_n) is a super-martingale adapted to the filtration $(\mathcal{F}_{k+n})$. Moreover $V_0 = 1$ and for any n, it holds $V_n \geq 0$ almost surely. Therefore we can apply Lemma B.1. □

B.2.2 Proof of Lemma 4.2

Proof of Lemma 4.2. We start from the update formula $\hat{\theta}_{t+1} = \hat{\theta}_t + P_{t+1}\frac{(y_t - b'(\hat{\theta}_t^\top X_t))X_t}{a}$ yielding

$$\begin{aligned}(\hat{\theta}_{t+1} - \theta^\star)^\top P_{t+1}^{-1}(\hat{\theta}_{t+1} - \theta^\star) = (\hat{\theta}_t - \theta^\star)^\top P_{t+1}^{-1}(\hat{\theta}_t - \theta^\star) + 2\frac{(y_t - b'(\hat{\theta}_t^\top X_t))X_t^\top}{a}(\hat{\theta}_t - \theta^\star) \\ + X_t^\top P_{t+1} X_t \left(\frac{y_t - b'(\hat{\theta}_t^\top X_t)}{a}\right)^2.\end{aligned}$$

With a summation argument, re-arranging terms, we obtain:

$$\begin{aligned}&\sum_{t=1}^{n}\left(\frac{(b'(\hat{\theta}_t^\top X_t) - y_t)X_t^\top}{a}(\hat{\theta}_t - \theta^\star) - \frac{1}{2}(\hat{\theta}_t - \theta^\star)^\top(P_{t+1}^{-1} - P_t^{-1})(\hat{\theta}_t - \theta^\star)\right) \\ &= \frac{1}{2}\sum_{t=1}^{n} X_t^\top P_{t+1} X_t \left(\frac{y_t - b'(\hat{\theta}_t^\top X_t)}{a}\right)^2 \\ &\quad + \frac{1}{2}\sum_{t=1}^{n}\left((\hat{\theta}_t - \theta^\star)^\top P_t^{-1}(\hat{\theta}_t - \theta^\star) - (\hat{\theta}_{t+1} - \theta^\star)^\top P_{t+1}^{-1}(\hat{\theta}_{t+1} - \theta^\star)\right).\end{aligned}$$

We bound the telescopic sum: as $P_{n+1}^{-1} \succcurlyeq 0$, we have

$$\begin{aligned}&\sum_{t=1}^{n}\left((\hat{\theta}_t - \theta^\star)^\top P_t^{-1}(\hat{\theta}_t - \theta^\star) - (\hat{\theta}_{t+1} - \theta^\star)^\top P_{t+1}^{-1}(\hat{\theta}_{t+1} - \theta^\star)\right) \\ &\qquad \leq (\hat{\theta}_1 - \theta^\star)^\top P_1^{-1}(\hat{\theta}_1 - \theta^\star) \leq \frac{\|\hat{\theta}_1 - \theta^\star\|^2}{\lambda_{\min}(P_1)}\,.\end{aligned}$$

The result follows from the identities

$$\frac{(b'(\hat{\theta}_t^\top X_t) - y_t)X_t}{a} = \ell'(y_t, \hat{\theta}_t^\top X_t)X_t\,, \qquad P_{t+1}^{-1} - P_t^{-1} = \ell''(y_t, \hat{\theta}_t^\top X_t)X_t X_t^\top\,.$$

□

B.2.3 Proofs of Section 4.3.2

Proof of Proposition 4.1. The first-order condition satisfied by $\theta^\star$ is

$$\mathbb{E}\left[-\frac{(y - b'(\theta^{\star\top} X))X}{a}\right] = 0\,,$$

yielding $\mathbb{E}\left[yX\right] = \mathbb{E}[b'(\theta^{\star\top} X)X]$. Therefore

$$\mathbb{E}\left[\frac{(b'(\theta^\top X) - y)X}{a}\right]^\top (\theta - \theta^\star) = \frac{1}{a}(\theta - \theta^\star)^\top \mathbb{E}\left[X(b'(\theta^\top X) - b'(\theta^{\star\top} X))\right]\,.$$

Considering the function $f : \lambda \to (\theta - \theta^\star)^\top \mathbb{E}\left[Xb'(\theta^\top X + \lambda(\theta - \theta^\star)^\top X))\right]$, we know there exists $\lambda \in [0,1]$ such that $f'(\lambda) = f(1) - f(0)$. This translates into

$$\frac{\partial L}{\partial \theta}\Big|_\theta^\top (\theta - \theta^\star) = \frac{1}{a}(\theta - \theta^\star)^\top \mathbb{E}\left[Xb''\left(\theta^\top X + \lambda(\theta^\star - \theta)^\top X\right)(\theta - \theta^\star)^\top X\right]\,.$$

Then we use Assumption 4.3:

$$\frac{b''\left(\theta^\top X + \lambda(\theta^\star - \theta)^\top X\right)}{b''\left(\theta^\top X\right)} = \frac{\ell''\left(y_t, \theta^\top X + \lambda(\theta^\star - \theta)^\top X\right)}{\ell''\left(y_t, \theta^\top X\right)} \geq \rho_{\|\theta - \theta^\star\|}\,,$$

yielding

$$\begin{aligned}\frac{\partial L}{\partial \theta}\Big|_\theta^\top (\theta - \theta^\star) &\geq \rho_{\|\theta-\theta^\star\|}(\theta - \theta^\star)^\top \mathbb{E}\left[\ell''(y, \theta^\top X)XX^\top\right](\theta - \theta^\star) \\ &= \rho_{\|\theta-\theta^\star\|}(\theta - \theta^\star)^\top \frac{\partial^2 L}{\partial \theta^2}\Big|_\theta (\theta - \theta^\star)\,.\end{aligned}$$

□

Proof of Proposition 4.2. We first recall that $L(\theta) - L(\theta^\star) \leq \frac{\partial L}{\partial \theta}\Big|_\theta^\top (\theta - \theta^\star)$, then Proposition 4.1 yields

$$\frac{\partial L}{\partial \theta}\Big|_\theta^\top (\theta - \theta^\star) - c(\theta - \theta^\star)^\top \frac{\partial^2 L}{\partial \theta^2}\Big|_\theta (\theta - \theta^\star) \geq (1 - \frac{c}{\rho_{\|\theta-\theta^\star\|}})\frac{\partial L}{\partial \theta}\Big|_\theta^\top (\theta - \theta^\star)\,,$$

and the result follows. □

Proof of Lemma 4.3. We first develop $(\Delta M_t)^2$:

$$\begin{aligned}(\Delta M_t)^2 &= \Big(\left(\mathbb{E}\left[\nabla_t \mid \mathcal{F}_{t-1}\right] - \nabla_t\right)^\top (\hat{\theta}_t - \theta^\star)\Big)^2 \\ &= (\hat{\theta}_t - \theta^\star)^\top \Big(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^\top + \nabla_t\nabla_t^\top \\ &\qquad\qquad - \nabla_t\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^\top - \mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]\nabla_t^\top\Big)(\hat{\theta}_t - \theta^\star) \\ &\leq 2(\hat{\theta}_t - \theta^\star)^\top \Big(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^\top + \nabla_t\nabla_t^\top\Big)(\hat{\theta}_t - \theta^\star) \\ &\leq 2(\hat{\theta}_t - \theta^\star)^\top \Big(\mathbb{E}[\nabla_t\nabla_t^\top \mid \mathcal{F}_{t-1}] + \nabla_t\nabla_t^\top\Big)(\hat{\theta}_t - \theta^\star)\,.\end{aligned}$$

The third line holds because if $U, V \in \mathbb{R}^d$, it holds $-UV^\top - VU^\top \preccurlyeq UU^\top + VV^\top$. The last one comes from $\mathbb{E}\Big[(\nabla_t - \mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}])(\nabla_t - \mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}])^\top \mid \mathcal{F}_{t-1}\Big] \succcurlyeq 0$.

Also, we have the relation

$$\mathbb{E}[(\Delta M_t)^2 \mid \mathcal{F}_{t-1}] \leq (\hat{\theta}_t - \theta^\star)^\top \mathbb{E}[\nabla_t\nabla_t^\top \mid \mathcal{F}_{t-1}](\hat{\theta}_t - \theta^\star)\,.$$

It yields

$$(\Delta M_t)^2 + \mathbb{E}[(\Delta M_t)^2 \mid \mathcal{F}_{t-1}] \leq (\hat{\theta}_t - \theta^\star)^\top \left(3\mathbb{E}[\nabla_t\nabla_t^\top \mid \mathcal{F}_{t-1}] + 2\nabla_t\nabla_t^\top\right)(\hat{\theta}_t - \theta^\star)\,,$$

and the result follows from Lemma 4.1. □

We derive the following Lemma in order to control the right-hand side of Lemma 4.2, in both settings.

Lemma B.2. *Assume the second point of Assumption 4.3 holds. For any $k, n \geq 1$, if $\|\hat{\theta}_t - \theta^\star\|^2 \leq \varepsilon$ for any $k < t \leq k+n$ then we have*

$$\sum_{t=k+1}^{k+n} \mathrm{Tr}\left(P_{t+1}(P_{t+1}^{-1} - P_t^{-1})\right) \leq d \ln\left(1 + n\frac{h_\varepsilon \lambda_{\max}(P_{k+1})D_X^2}{d}\right)\,.$$

Proof. We apply Lemma 11.11 of Cesa-Bianchi and Lugosi, 2006:

$$\begin{aligned}\sum_{t=k+1}^{k+n} \mathrm{Tr}\left(P_{t+1}(P_{t+1}^{-1} - P_t^{-1})\right) &= \sum_{t=k+1}^{k+n}\left(1 - \frac{\det(P_t^{-1})}{\det(P_{t+1}^{-1})}\right) \\ &\leq \sum_{t=k+1}^{k+n} \ln\left(\frac{\det(P_{t+1}^{-1})}{\det(P_t^{-1})}\right) \\ &= \ln\left(\frac{\det(P_{k+n+1}^{-1})}{\det(P_{k+1}^{-1})}\right) \\ &\leq \ln\det\left(I + \sum_{t=k+1}^{k+n} \ell''(y_t, \hat{\theta}_t^\top X_t)(P_{k+1}^{1/2}X_t)(P_{k+1}^{1/2}X_t)^\top\right) \\ &= \sum_{i=1}^{d} \ln(1+\lambda_i)\,,\end{aligned}$$

where $\lambda_1, ..., \lambda_d$ are the eigenvalues of $\sum_{t=k+1}^{k+n} \ell''(y_t, \hat{\theta}_t^\top X_t)(P_{k+1}^{1/2} X_t)(P_{k+1}^{1/2} X_t)^\top$. Therefore we have

$$\sum_{t=k+1}^{k+n} \mathrm{Tr}\left(P_{t+1}(P_{t+1}^{-1} - P_t^{-1})\right) \leq d \ln\left(1 + \frac{1}{d}\sum_{i=1}^{d} \lambda_i\right) \leq d \ln\left(1 + \frac{1}{d} n h_\varepsilon \lambda_{\max}(P_{k+1}) D_X^2\right).$$

□

B.2.4 Bounded Setting (Assumption 4.3)

Proof of Theorem 4.1. Let $\delta > 0$. On the one hand, we sum Lemma 4.2 and 4.3. We obtain, for any $\lambda > 0$,

$$\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} \Big(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^\top(\hat{\theta}_t - \theta^\star) - \frac{1}{2} Q_t - \lambda(\hat{\theta}_t - \theta^\star)^\top\Big(\nabla_t \nabla_t^\top + \frac{3}{2}\mathbb{E}\big[\nabla_t \nabla_t^\top \mid \mathcal{F}_{t-1}\big]\Big)(\hat{\theta}_t - \theta^\star)\Big) \leq \frac{1}{2} \sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} X_t^\top P_{t+1} X_t \ell'(y_t, \hat{\theta}_t^\top X_t)^2 + \frac{\|\hat{\theta}_1 - \theta^\star\|^2}{\lambda_{\min}(P_{T(\varepsilon,\delta)+1})} + \frac{\ln \delta^{-1}}{\lambda}, \qquad n \geq 1, \tag{B.1}$$

with probability at least $1 - \delta$, where we define $Q_t = (\hat{\theta}_t - \theta^\star)^\top \left(\ell''(y_t, \hat{\theta}_t^\top X_t) X_t X_t^\top\right)(\hat{\theta}_t - \theta^\star)$ for any t.

On the other hand, thanks to Assumption 4.3, we can apply Proposition 4.2 with $c = 0.75$ to obtain, for any $t \geq 1$,

$$\begin{aligned} &\|\hat{\theta}_t - \theta^\star\| \leq \varepsilon \\ &\quad\implies L(\hat{\theta}_t) - L(\theta^\star) \leq \frac{\rho_\varepsilon}{\rho_\varepsilon - 0.75}\Big(\frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^\top (\hat{\theta}_t - \theta^\star) - 0.75(\hat{\theta}_t - \theta^\star)^\top \frac{\partial^2 L}{\partial \theta^2}\Big|_{\hat{\theta}_t} (\hat{\theta}_t - \theta^\star)\Big), \\ &\quad\implies L(\hat{\theta}_t) - L(\theta^\star) \leq 5\Big(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^\top(\hat{\theta}_t - \theta^\star) - 0.75\mathbb{E}[Q_t \mid \mathcal{F}_{t-1}]\Big), \end{aligned} \tag{B.2}$$

because $\rho_\varepsilon > 0.95$.

In order to bridge the gap between Equations (B.1) and (B.2), we need to control the quadratic terms of Equation (B.1) with $\mathbb{E}[Q_t \mid \mathcal{F}_{t-1}]$. First, for any t, if $\|\hat{\theta}_t - \theta^\star\| \leq \varepsilon$, we have $Q_t \in [0, h_\varepsilon \varepsilon^2 D_X^2]$, and we apply Lemma A.3 of Cesa-Bianchi and Lugosi, 2006 to the random variable $\frac{1}{h_\varepsilon \varepsilon^2 D_X^2} Q_t \in [0, 1]$: for any $s > 0$,

$$\mathbb{E}\left[\exp\left(\frac{s}{h_\varepsilon \varepsilon^2 D_X^2} Q_t - \frac{e^s - 1}{h_\varepsilon \varepsilon^2 D_X^2}\mathbb{E}\left[Q_t \mid \mathcal{F}_{t-1}\right]\right) \mid \mathcal{F}_{t-1}, \|\hat{\theta}_t - \theta^\star\| \leq \varepsilon\right] \leq 1.$$

We fix $s = 0.1$ and we define

$$V_n = \exp\left(\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} \left(\frac{0.1}{h_\varepsilon \varepsilon^2 D_X^2} Q_t - (e^{0.1} - 1)\mathbb{E}\left[\frac{1}{h_\varepsilon \varepsilon^2 D_X^2} Q_t \mid \mathcal{F}_{t-1}\right]\right)\right).$$

The sequence (V_n) is adapted to $(\mathcal{F}_{T(\varepsilon,\delta)+n})$, almost surely we have $V_0 = 1$ and $V_n \geq 0$. Finally,

$$\mathbb{E}\left[V_n \mid \mathcal{F}_{T(\varepsilon,\delta)+n-1}, \|\hat{\theta}_{T(\varepsilon,\delta)+n} - \theta^\star\| \leq \varepsilon\right] \leq V_{n-1}\,,$$

and $(\|\hat{\theta}_{T(\varepsilon,\delta)+n} - \theta^\star\| \leq \varepsilon)$ belongs to $\mathcal{F}_{T(\varepsilon,\delta)+n-1}$. We apply Lemma B.1:

$$\mathbb{P}\left(\left(\bigcup_{n=1}^{\infty} V_n > \delta^{-1}\right) \cup \left(\bigcup_{n=1}^{\infty} (\|\hat{\theta}_{T(\varepsilon,\delta)+n} - \theta^\star\| > \varepsilon)\right)\right) \leq \delta + \mathbb{P}\left(\bigcup_{n=1}^{\infty} (\|\hat{\theta}_{T(\varepsilon,\delta)+n} - \theta^\star\| > \varepsilon)\right)\,.$$

We define $A_k^\varepsilon = \bigcap_{n=k+1}^{\infty} (\|\hat{\theta}_n - \theta^\star\| \leq \varepsilon)$ for any k. The last inequality is equivalent to

$$\mathbb{P}\Big(\bigcup_{n=1}^{\infty}\Big(\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} Q_t > 10(e^{0.1}-1)\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} \mathbb{E}\left[Q_t \mid \mathcal{F}_{t-1}\right] + 10h_\varepsilon\varepsilon^2 D_X^2 \ln\delta^{-1}\Big) \cap A^\varepsilon_{T(\varepsilon,\delta)}\Big) \leq \delta\,. \tag{B.3}$$

We then bound the two quadratic terms coming from Lemma 4.3: using Assumption 4.3 we have the implications

$$\begin{aligned}
&\|\hat{\theta}_t - \theta^\star\| \leq \varepsilon \implies (\hat{\theta}_t - \theta^\star)^\top \nabla_t \nabla_t^\top (\hat{\theta}_t - \theta^\star) \leq \kappa_\varepsilon Q_t\,,\\
&\|\hat{\theta}_t - \theta^\star\| \leq \varepsilon \implies (\hat{\theta}_t - \theta^\star)^\top \mathbb{E}\big[\nabla_t \nabla_t^\top \mid \mathcal{F}_{t-1}\big](\hat{\theta}_t - \theta^\star) \leq \kappa_\varepsilon \mathbb{E}\left[Q_t \mid \mathcal{F}_{t-1}\right]\,.
\end{aligned}$$

Therefore, we get from (B.3)

$$\begin{aligned}
\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty}\Bigg(\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n}\Big(&\frac{1}{2}Q_t + \lambda(\hat{\theta}_t - \theta^\star)^\top \nabla_t \nabla_t^\top (\hat{\theta}_t - \theta^\star)\\
&+ \frac{3}{2}\lambda(\hat{\theta}_t - \theta^\star)^\top \mathbb{E}\big[\nabla_t \nabla_t^\top \mid \mathcal{F}_{t-1}\big](\hat{\theta}_t - \theta^\star)\Big) >\\
&\Big(10(e^{0.1}-1)(\frac{1}{2} + \lambda\kappa_\varepsilon) + \frac{3}{2}\lambda\kappa_\varepsilon\Big)\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} \mathbb{E}\left[Q_t \mid \mathcal{F}_{t-1}\right]\\
&+ 10(\frac{1}{2} + \lambda\kappa_\varepsilon)h_\varepsilon\varepsilon^2 D_X^2 \ln\delta^{-1}\Bigg) \cap A^\varepsilon_{T(\varepsilon,\delta)}\Bigg) \leq \delta\,.
\end{aligned}$$

We set $\lambda = \frac{0.75 - 5(e^{0.1}-1)}{(10(e^{0.1}-1)+\frac{3}{2})\kappa_\varepsilon}$, so that

$$\begin{aligned}
&10(e^{0.1}-1)(\frac{1}{2} + \lambda\kappa_\varepsilon) + \frac{3}{2}\lambda\kappa_\varepsilon = 0.75\,,\\
&\frac{1}{2} + \lambda\kappa_\varepsilon = \frac{1}{2} + \frac{0.75 - 5(e^{0.1}-1)}{10(e^{0.1}-1)+\frac{3}{2}} \approx 0.59 \leq 0.6\,,
\end{aligned}$$

and consequently

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty}\Bigg(\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n}\Big(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^\top(\hat{\theta}_t-\theta^\star)-0.75\mathbb{E}[Q_t \mid \mathcal{F}_{t-1}]\Big) > 6h_\varepsilon\varepsilon^2 D_X^2\ln\delta^{-1}$$
$$+\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n}\Bigg(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^\top(\hat{\theta}_t-\theta^\star)-\frac{1}{2}Q_t$$
$$-\lambda(\hat{\theta}_t-\theta^\star)^\top\Big(\nabla_t\nabla_t^\top+\frac{3}{2}\mathbb{E}\big[\nabla_t\nabla_t^\top \mid \mathcal{F}_{t-1}\big]\Big)(\hat{\theta}_t-\theta^\star)\Bigg)\Bigg)\cap A^\varepsilon_{T(\varepsilon,\delta)}\Bigg)\le\delta\,.$$

We plug Equation (B.2) in the last inequality:

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty}\Bigg(\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n}(L(\hat{\theta}_t)-L(\theta^\star)) > 30h_\varepsilon\varepsilon^2D_X^2\ln\delta^{-1}$$
$$+5\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n}\Bigg(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^\top(\hat{\theta}_t-\theta^\star)-\frac{1}{2}Q_t$$
$$-\lambda(\hat{\theta}_t-\theta^\star)^\top\left(\nabla_t\nabla_t^\top+\frac{3}{2}\mathbb{E}\big[\nabla_t\nabla_t^\top \mid \mathcal{F}_{t-1}\big]\right)(\hat{\theta}_t-\theta^\star)\Bigg)\Bigg)\cap A^\varepsilon_{T(\varepsilon,\delta)}\Bigg)\le\delta\,.$$

We then use Equation (B.1) with $\frac{1}{\lambda}=\frac{(10(e^{0.1}-1)+\frac{3}{2})\kappa_\varepsilon}{0.75-5(e^{0.1}-1)}\approx 11.4\kappa_\varepsilon\le 12\kappa_\varepsilon$. It yields

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty}\Bigg(\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n}(L(\hat{\theta}_t)-L(\theta^\star)) > \frac{5}{2}\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n}X_t^\top P_{t+1}X_t\ell'(y_t,\hat{\theta}_t^\top X_t)^2$$
$$+\frac{5\|\hat{\theta}_1-\theta^\star\|^2}{\lambda_{\min}(P_{T(\varepsilon,\delta)+1})}+30(2\kappa_\varepsilon+h_\varepsilon\varepsilon^2D_X^2)\ln\delta^{-1}\Bigg)\cap A^\varepsilon_{T(\varepsilon,\delta)}\Bigg)\le 2\delta\,.$$

Thanks to Assumption 4.3, we have

$$X_t^\top P_{t+1}X_t\ell'(y_t,\hat{\theta}_t^\top X_t)^2\le\kappa_\varepsilon\operatorname{Tr}\left(P_{t+1}(P_{t+1}^{-1}-P_t^{-1})\right),\qquad t>T(\varepsilon,\delta)\,,$$

therefore we apply Lemma B.2: for any n, it holds

$$\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n}X_t^\top P_{t+1}X_t\ell'(y_t,\hat{\theta}_t^\top X_t)^2\le d\kappa_\varepsilon\ln\left(1+n\frac{h_\varepsilon\lambda_{\max}(P_{T(\varepsilon,\delta)+1})D_X^2}{d}\right).$$

As $P_{T(\varepsilon,\delta)+1}\preccurlyeq P_1$, we obtain

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty}\Bigg(\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n}(L(\hat{\theta}_t)-L(\theta^\star)) > \frac{5}{2}d\kappa_\varepsilon\ln\left(1+n\frac{h_\varepsilon\lambda_{\max}(P_1)D_X^2}{d}\right)$$
$$+\frac{5\|\hat{\theta}_1-\theta^\star\|^2}{\lambda_{\min}(P_{T(\varepsilon,\delta)+1})}+30(2\kappa_\varepsilon+h_\varepsilon\varepsilon^2D_X^2)\ln\delta^{-1}\Bigg)\cap A^\varepsilon_{T(\varepsilon,\delta)}\Bigg)\le 2\delta\,.$$

To conclude, we use Assumption 4.5. □

B.2.5 Quadratic Setting (Assumption 4.4)

We recall two definitions introduced in the previous subsection:

$$A_k^\varepsilon = \bigcap_{n=k+1}^{\infty} (\|\hat{\theta}_n - \theta^\star\| \le \varepsilon), \qquad k \ge 1\,,$$
$$Q_t = (\hat{\theta}_t - \theta^\star)^\top X_t X_t^\top (\hat{\theta}_t - \theta^\star), \qquad t \ge 1\,.$$

The sub-gaussian hypothesis requires a different treatment of several steps in the proof. In the following proofs, we use a consequence of the first points of Assumption 4.4. We apply Lemma 1.4 of Rigollet and Hütter, 2015: for any $X \in \mathbb{R}^d$,

$$\mathbb{E}[(y - \mathbb{E}[y \mid X])^{2i} \mid X] \le 2i(2\sigma^2)^i \Gamma(i) = 2(2\sigma^2)^i i!, \qquad i \in \mathbb{N}^\star\,. \tag{B.4}$$

First, we control the quadratic terms in $\nabla_t = -(y_t - \hat{\theta}_t^\top X_t)X_t$ in the following lemma.

Lemma B.3. *1. For any $k \in \mathbb{N}$ and $\delta > 0$, we have*

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty} \Bigg(\sum_{t=k+1}^{k+n} (\hat{\theta}_t - \theta^\star)^\top \nabla_t \nabla_t^\top (\hat{\theta}_t - \theta^\star) \\ > 3\left(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2\right) \sum_{t=k+1}^{k+n} Q_t + 12\varepsilon^2 D_X^2 \sigma^2 \ln \delta^{-1} \Bigg) \cap A_k^\varepsilon \Bigg) \le \delta\,.$$

2. For any t, it holds almost surely

$$(\hat{\theta}_t - \theta^\star)^\top \mathbb{E}[\nabla_t \nabla_t^\top \mid \mathcal{F}_{t-1}](\hat{\theta}_t - \theta^\star) \le 3\left(\sigma^2 + D_{\text{app}}^2 + \|\hat{\theta}_t - \theta^\star\|^2 D_X^2\right) \mathbb{E}[Q_t \mid \mathcal{F}_{t-1}]\,.$$

Proof. 1. We recall that for any a, b, c, we have $(a+b+c)^2 \le 3(a^2+b^2+c^2)$. Thus

$$\begin{aligned}(\hat{\theta}_t - \theta^\star)^\top \nabla_t \nabla_t^\top (\hat{\theta}_t - \theta^\star) &= Q_t (y_t - \hat{\theta}_t^\top X_t)^2 \\ &\le 3Q_t\Big((y_t - \mathbb{E}[y_t \mid X_t])^2 + (\mathbb{E}[y_t \mid X_t] - \theta^{\star\top} X_t)^2 + ((\theta^\star - \hat{\theta}_t)^\top X_t)^2\Big) \\ &\le 3Q_t\left((y_t - \mathbb{E}[y_t \mid X_t])^2 + D_{\text{app}}^2 + \|\hat{\theta}_t - \theta^\star\|^2 D_X^2\right)\,. \end{aligned} \tag{B.5}$$

To obtain the last inequality, we use the second point of Assumption 4.4 to bound the middle term. Then we use Taylor series for the exponential, and we apply Equation (B.4).

For any t and any μ satisfying $0 < \mu \leq \frac{1}{4Q_t\sigma^2}$, we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\mu Q_t(y_t - \mathbb{E}[y_t \mid X_t])^2\right) \mid \mathcal{F}_{t-1}, X_t\right] &= 1 + \sum_{i\geq 1} \frac{\mu^i Q_t^i \mathbb{E}[(y_t - \mathbb{E}\left[y_t \mid X_t\right])^{2i} \mid X_t\right]}{i!} \\
&\leq 1 + 2\sum_{i\geq 1} \frac{\mu^i Q_t^i i! (2\sigma^2)^i}{i!} \\
&\leq 1 + 2\sum_{i\geq 1} \left(2\mu Q_t\sigma^2\right)^i \\
&\leq 1 + 8\mu Q_t\sigma^2, \qquad 2\mu Q_t\sigma^2 \leq \frac{1}{2} \\
&\leq \exp\left(8\mu Q_t\sigma^2\right) .
\end{aligned}$$

Therefore, for any t,

$$\mathbb{E}\left[\exp\left(\frac{1}{4\varepsilon^2 D_X^2\sigma^2} Q_t\left((y_t - \mathbb{E}[y_t \mid X_t])^2 - 8\sigma^2\right)\right) \mid \mathcal{F}_{t-1}, X_t, \|\hat{\theta}_t - \theta^\star\| \leq \varepsilon\right] \leq 1 .$$

We define the random variable

$$V_n = \exp\left(\frac{1}{4\varepsilon^2 D_X^2\sigma^2} \sum_{t=k+1}^{k+n} Q_t\left((y_t - \mathbb{E}[y_t \mid X_t])^2 - 8\sigma^2\right)\right), \qquad n \in \mathbb{N} .$$

$(V_n)_n$ is adapted to the filtration $(\sigma(X_1, y_1, ..., X_{k+n}, y_{k+n}, X_{k+n+1})_n$, moreover $V_0 = 1$ and $V_n \geq 0$ almost surely, and

$$\mathbb{E}\left[V_n \mid X_1, y_1, ..., X_{k+n-1}, y_{k+n-1}, X_{k+n}, \|\hat{\theta}_{k+n} - \theta^\star\| \leq \varepsilon\right] \leq V_{n-1} .$$

Therefore we apply Lemma B.1: for any $\delta > 0$,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} (V_n > \delta^{-1}) \cap A_k^\varepsilon\right) \leq \delta ,$$

which is equivalent to

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty}\left(\sum_{t=k+1}^{k+n} Q_t(y_t - \mathbb{E}[y_t \mid X_t])^2 > 8\sigma^2 \sum_{t=k+1}^{k+n} Q_t + 4\varepsilon^2 D_X^2\sigma^2 \ln\delta^{-1}\right) \cap A_k^\varepsilon\right) \leq \delta .$$

Substituting in Equation (B.5), we obtain the desired result.

2. We apply the same decomposition as for Equation (B.5): for any t,

$$\begin{aligned}
&(\hat{\theta}_t - \theta^\star)^\top \mathbb{E}[\nabla_t \nabla_t^\top \mid \mathcal{F}_{t-1}](\hat{\theta}_t - \theta^\star) \\
&\leq 3(\hat{\theta}_t - \theta^\star)^\top \mathbb{E}\left[X_t X_t^\top \left((y_t - \mathbb{E}[y_t \mid X_t])^2 + D_{\text{app}}^2 + \|\theta^\star - \hat{\theta}_t\|^2 D_X^2\right) \mid \mathcal{F}_{t-1}\right](\hat{\theta}_t - \theta^\star) .
\end{aligned}$$

Assumption 4.4 implies that for any X_t, $\mathbb{E}[(y_t - \mathbb{E}[y_t \mid X_t])^2 \mid X_t] \leq \sigma^2$. Thus, the tower

property yields

$$(\hat{\theta}_t - \theta^\star)^\top \mathbb{E}[\nabla_t \nabla_t^\top \mid \mathcal{F}_{t-1}](\hat{\theta}_t - \theta^\star) \\ \leq 3\left(\sigma^2 + D_{\text{app}}^2 + \|\hat{\theta}_t - \theta^\star\|^2 D_X^2\right)(\hat{\theta}_t - \theta^\star)^\top \mathbb{E}[X_t X_t^\top \mid \mathcal{F}_{t-1}](\hat{\theta}_t - \theta^\star)\,.$$

□

Second, we bound the right-hand side of Lemma 4.2, that is the objective of the following lemma.

Lemma B.4. *Let $k \in \mathbb{N}$. For any $\delta > 0$, we have*

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty}\Bigg(\sum_{t=k+1}^{k+n} X_t^\top P_{t+1} X_t (y_t - \hat{\theta}_t^\top X_t)^2 > 12\lambda_{\max}(P_1) D_X^2 \sigma^2 \ln \delta^{-1} \\ + 3\left(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2\right) d \ln\left(1 + n\frac{\lambda_{\max}(P_{k+1}) D_X^2}{d}\right)\Bigg) \cap A_k^\varepsilon\Bigg) \leq \delta\,.$$

Proof. We apply a similar analysis as in the proof of Lemma B.3 in order to use the sub-gaussian assumption, and then we apply the telescopic argument as in the bounded setting. We decompose $y_t - \hat{\theta}_t^\top X_t$:

$$\begin{aligned} &X_t^\top P_{t+1} X_t (y_t - \hat{\theta}_t^\top X_t)^2 \\ &\quad \leq 3 X_t^\top P_{t+1} X_t \Big((y_t - \mathbb{E}[y_t \mid X_t])^2 + (\mathbb{E}[y_t \mid X_t] - b'(\theta^{\star\top} X_t))^2 + ((\theta^\star - \hat{\theta}_t)^\top X_t)^2\Big) \\ &\quad \leq 3 X_t^\top P_{t+1} X_t \left((y_t - \mathbb{E}[y_t \mid X_t])^2 + D_{\text{app}}^2 + \|\hat{\theta}_t - \theta^\star\|^2 D_X^2\right)\,. \end{aligned} \tag{B.6}$$

To control $(y_t - \mathbb{E}[y_t \mid X_t])^2 X_t^\top P_{t+1} X_t$, we use its positivity along with Equation (B.4). Precisely, for any t and any $\mu > 0$ satisfying $0 < \mu \leq \frac{1}{4 X_t^\top P_{t+1} X_t \sigma^2}$, we have

$$\begin{aligned} &\mathbb{E}\left[\exp\left(\mu(y_t - \mathbb{E}[y_t \mid X_t])^2 X_t^\top P_{t+1} X_t\right) \mid \mathcal{F}_{t-1}, X_t\right] \\ &\quad = 1 + \sum_{i \geq 1} \frac{\mu^i (X_t^\top P_{t+1} X_t)^i \mathbb{E}\left[(y_t - \mathbb{E}[y_t \mid X_t])^{2i} \mid X_t\right]}{i!} \\ &\quad \leq 1 + 2\sum_{i \geq 1} \frac{\mu^i (X_t^\top P_{t+1} X_t)^i i! (2\sigma^2)^i}{i!} \\ &\quad = 1 + 2\sum_{i \geq 1} \left(2\mu X_t^\top P_{t+1} X_t \sigma^2\right)^i \\ &\quad \leq 1 + 8\mu X_t^\top P_{t+1} X_t \sigma^2, \qquad 0 < 2\mu X_t^\top P_{t+1} X_t \sigma^2 \leq \frac{1}{2} \\ &\quad \leq \exp\left(8\mu X_t^\top P_{t+1} X_t \sigma^2\right)\,. \end{aligned}$$

We apply the previous bound with a uniform $\mu = \frac{1}{4\lambda_{\max}(P_1) D_X^2 \sigma^2}$. As $\lambda_{\max}(P_{t+1}) \leq \lambda_{\max}(P_1)$ for

any t, we get $\mu \leq \frac{1}{4X_t^\top P_{t+1} X_t \sigma^2}$. Thus, we define

$$V_n = \exp\left(\frac{1}{4\lambda_{\max}(P_1) D_X^2 \sigma^2} \sum_{t=k+1}^{k+n} \left((y_t - \mathbb{E}[y_t \mid X_t])^2 - 8\sigma^2\right) X_t^\top P_{t+1} X_t\right), \qquad n \in \mathbb{N}\,.$$

(V_n) is a super-martingale adapted to the filtration $(\sigma(X_1, y_1, ..., X_{k+n-1}, y_{k+n-1}, X_{k+n}))_n$ satisfying almost surely $V_0 = 1, V_n \geq 0$, thus we apply Lemma B.1:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} (V_n > \delta^{-1})\right) \leq \delta\,,$$

or equivalently

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty}\Bigg(\sum_{t=k+1}^{k+n} X_t^\top P_{t+1} X_t (y_t - \mathbb{E}[y_t \mid X_t])^2 \\ > 8\sigma^2 \sum_{t=k+1}^{k+n} X_t^\top P_{t+1} X_t + 4\lambda_{\max}(P_1) D_X^2 \sigma^2 \ln \delta^{-1}\Bigg)\Bigg) \leq \delta\,.$$

Combining it with Equation (B.6), we get

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty}\Bigg(\sum_{t=k+1}^{k+n} X_t^\top P_{t+1} X_t (y_t - \hat{\theta}_t^\top X_t)^2 > 3\left(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2\right) \sum_{t=k+1}^{k+n} X_t^\top P_{t+1} X_t \\ + 12\lambda_{\max}(P_1) D_X^2 \sigma^2 \ln \delta^{-1}\Bigg) \cap A_k^\varepsilon\Bigg) \leq \delta\,.$$

Then we apply Lemma B.2: the second point of Assumption 4.3 holds with $h_\varepsilon = 1$, thus

$$\sum_{t=k+1}^{k+n} \operatorname{Tr}\left(P_{t+1}(P_{t+1}^{-1} - P_t^{-1})\right) \leq d \ln\left(1 + n\frac{\lambda_{\max}(P_{k+1}) D_X^2}{d}\right), \qquad n \geq 1.$$

We conclude with $X_t^\top P_{t+1} X_t = \operatorname{Tr}(P_{t+1}(P_{t+1}^{-1} - P_t^{-1}))$. □

We sum up our findings and we prove the result for the quadratic loss. The structure of the proof is the same as the one of Theorem 4.1.

Proof of Theorem 4.2. On the one hand, we sum Lemma 4.2 and Lemma 4.3: for any $\lambda, \delta > 0$

$$\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} \left(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^\top (\hat{\theta}_t - \theta^\star) - \frac{1}{2} Q_t - \lambda(\hat{\theta}_t - \theta^\star)^\top \left(\nabla_t \nabla_t^\top + \frac{3}{2}\mathbb{E}\left[\nabla_t \nabla_t^\top \mid \mathcal{F}_{t-1}\right]\right)(\hat{\theta}_t - \theta^\star)\right) \\ \leq \frac{1}{2} \sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} X_t^\top P_{t+1} X_t (y_t - \hat{\theta}_t^\top X_t)^2 + \frac{\|\hat{\theta}_{T(\varepsilon,\delta)+1} - \theta^\star\|^2}{\lambda_{\min}(P_{T(\varepsilon,\delta)+1})} + \frac{\ln \delta^{-1}}{\lambda}, \qquad n \geq 1\,, \quad \text{(B.7)}$$

with probability at least $1-\delta$. On the other hand, we have

$$\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} (L(\hat{\theta}_t) - L(\theta^\star)) \leq \frac{1}{1-0.8} \sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} \Big(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^\top (\hat{\theta}_t - \theta^\star) - 0.8\mathbb{E}[Q_t \mid \mathcal{F}_{t-1}] \Big) \,. \tag{B.8}$$

We aim to relate Equations (B.7) and (B.8) as in the proof of Theorem 4.1. To that end, we apply Lemma B.3:

$$\begin{aligned}
\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty} \Bigg(& \sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} \left(\frac{1}{2} Q_t + \lambda(\hat{\theta}_t - \theta^\star)^\top \left(\nabla_t \nabla_t^\top + \frac{3}{2} \mathbb{E}\big[\nabla_t \nabla_t^\top \mid \mathcal{F}_{t-1} \big] \right) (\hat{\theta}_t - \theta^\star) \right) \\
& > \Big(\frac{1}{2} + 3\lambda(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2) \Big) \sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} Q_t \\
& \quad + \frac{9}{2} \lambda \Big(\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2 \Big) \sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} \mathbb{E}\left[Q_t \mid \mathcal{F}_{t-1} \right] + 12\lambda\varepsilon^2 D_X^2 \sigma^2 \ln \delta^{-1} \Bigg) \\
& \cap A^{\varepsilon}_{T(\varepsilon,\delta)} \Bigg) \leq \delta \,.
\end{aligned}$$

As in the proof of Theorem 4.1 we apply Lemma A.3 of (Cesa-Bianchi and Lugosi, 2006) and Lemma B.1: for any $\delta > 0$,

$$\begin{aligned}
\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty} \Bigg(& \sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} Q_t > 10(e^{0.1} - 1) \sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} \mathbb{E}[Q_t \mid \mathcal{F}_{t-1}] + 10\varepsilon^2 D_X^2 \ln \delta^{-1} \Bigg) \\
& \cap A^{\varepsilon}_{T(\varepsilon,\delta)} \Bigg) \leq \delta \,.
\end{aligned}$$

We combine the last two inequalities:

$$\begin{aligned}
\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty} \Bigg(& \sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} \left(\frac{1}{2} Q_t + \lambda(\hat{\theta}_t - \theta^\star)^\top \left(\nabla_t \nabla_t^\top + \frac{3}{2} \mathbb{E}\big[\nabla_t \nabla_t^\top \mid \mathcal{F}_{t-1} \big] \right) (\hat{\theta}_t - \theta^\star) \right) \\
& > \left(10(e^{0.1} - 1) \Big(\frac{1}{2} + 3\lambda(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2) \Big) + \frac{9}{2} \lambda(\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2) \right) \\
& \qquad \sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} \mathbb{E}\left[Q_t \mid \mathcal{F}_{t-1} \right] \\
& \quad + \left(10\varepsilon^2 D_X^2 \Big(\frac{1}{2} + 3\lambda(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2) \Big) + 12\lambda\varepsilon^2 D_X^2 \sigma^2 \right) \ln \delta^{-1} \Bigg) \\
& \cap A^{\varepsilon}_{T(\varepsilon,\delta)} \Bigg) \leq 2\delta \,.
\end{aligned} \tag{B.9}$$

We set

$$\lambda = \big(0.8 - 5(e^{0.1}-1)\big)\left(30(e^{0.1}-1)(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2) + \frac{9}{2}(\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2)\right)^{-1}$$

in order to obtain

$$\begin{aligned}
&10(e^{0.1}-1)\Big(\frac{1}{2} + 3\lambda(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2)\Big) + \frac{9}{2}\lambda(\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2) = 0.8\,,\\
&\frac{1}{109\sigma^2 + 28D_{\text{app}}^2 + 28\varepsilon^2 D_X^2} < \lambda < \frac{1}{108\sigma^2 + 27D_{\text{app}}^2 + 27\varepsilon^2 D_X^2}\,,\\
&10\varepsilon^2 D_X^2\Big(\frac{1}{2} + 3\lambda(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2)\Big) + 12\lambda D_X^2\varepsilon^2\sigma^2 \le 8\varepsilon^2 D_X^2\\
&\frac{1}{\lambda} \le 28(4\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2)\,.
\end{aligned}$$

Combining Equations (B.7), (B.8) and (B.9), we obtain

$$\begin{aligned}
\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty}\Bigg(0.2\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n}&(L(\hat\theta_t) - L(\theta^\star))\\
&> \frac{1}{2}\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} X_t^\top P_{t+1} X_t (y_t - \hat\theta_t^\top X_t)^2 + \frac{\varepsilon^2}{\lambda_{\min}(P_{T(\varepsilon,\delta)+1})}\\
&\quad + 28(4\sigma^2 + D_{\text{approx}}^2 + \varepsilon^2 D_X^2)\ln\delta^{-1} + 8\varepsilon^2 D_X^2 \ln\delta^{-1}\Bigg) \cap A^\varepsilon_{T(\varepsilon,\delta)}\Bigg) \le 3\delta\,.
\end{aligned}$$

Finally, we apply Lemma B.4 with $P_{T(\varepsilon,\delta)+1} \preccurlyeq P_1$ and we use Assumption 4.5: it holds simultaneously

$$\begin{aligned}
\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} L(\hat\theta_t) - L(\theta^\star) \le 5\Bigg(&\frac{3}{2}\left(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2\right) d \ln\left(1 + n\frac{\lambda_{\max}(P_1)D_X^2}{d}\right)\\
&+ \lambda_{\max}\left(P_{T(\varepsilon,\delta)+1}^{-1}\right)\varepsilon^2 + 28(4\sigma^2 + D_{\text{approx}}^2 + \varepsilon^2 D_X^2)\ln\delta^{-1}\\
&+ 8\varepsilon^2 D_X^2 \ln\delta^{-1} + 6\lambda_{\max}(P_1)D_X^2\sigma^2\ln\delta^{-1}\Bigg), \qquad n \ge 1\,,
\end{aligned}$$

with probability at least $1 - 5\delta$. To conclude, we write

$$\begin{aligned}
&28(4\sigma^2 + D_{\text{approx}}^2 + \varepsilon^2 D_X^2) + 8\varepsilon^2 D_X^2 + 6\lambda_{\max}(P_1)D_X^2\sigma^2\\
&\quad \le 28\left(\sigma^2(4 + \frac{\lambda_{\max}(P_1)D_X^2}{4}) + D_{\text{app}}^2 + 2\varepsilon^2 D_X^2\right).
\end{aligned}$$

□

B.3 Proofs of Section 4.4

B.3.1 Proof of Theorem 4.3

Proof of Theorem 4.3. We check Assumption 4.3 with $\kappa_\varepsilon = e^{D_X(\|\theta^\star\|+\varepsilon)}, h_\varepsilon = \frac{1}{4}$ and $\rho_\varepsilon = e^{-\varepsilon D_X} > 0.95$. We can thus apply Theorem 4.1 with

$$\lambda_{\max}(P^{-1}_{T(\varepsilon,\delta)+1}) \le \lambda_{\max}(P_1^{-1}) + \frac{1}{4}\sum_{t=1}^{T(\varepsilon,\delta)} \|X_t\|^2\,,$$

$$\frac{5\kappa_\varepsilon}{2} < 3e^{D_X\|\theta^\star\|}, \qquad 30\Big(2\kappa_\varepsilon + \frac{\varepsilon^2 D_X^2}{4}\Big) < 64 e^{D_X\|\theta^\star\|}, \qquad 5\varepsilon^2 D_X^2 \le 1/75\,.$$

We then control the first terms. To that end, we use a rough bound at any time $t \ge 1$:

$$\begin{aligned} L(\hat\theta_t) - L(\theta^\star) &\le \mathbb{E}\left[\frac{yX}{1+e^{y\hat\theta_t^\top X}} \mid \hat\theta_t\right]^\top (\hat\theta_t - \theta^\star) \\ &\le D_X\|\hat\theta_t - \theta^\star\| \\ &\le D_X(\|\hat\theta_1 - \theta^\star\| + (t-1)\lambda_{\max}(P_1)D_X)\,, \end{aligned}$$

because for any $s \ge 1$, we have $P_s \preccurlyeq P_1$ and therefore $\|\hat\theta_{s+1} - \hat\theta_s\| \le \lambda_{\max}(P_1)D_X$. Summing from 1 to $T(\frac{1}{20D_X}, \delta)$ yields the result. □

B.3.2 Concentration of P_t

We prove a concentration result based on Tropp, 2012, which will be used on the inverse of P_t.

Lemma B.5. *If Assumption 4.1 is satisfied, then for any $0 \le \beta < 1$ and $t \ge 4^{1/(1-\beta)}$, it holds*

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{s=1}^{t-1}\frac{X_sX_s^\top}{s^\beta}\right) < \frac{\Lambda_{\min}t^{1-\beta}}{4(1-\beta)}\right) \le d\exp\left(-t^{1-\beta}\frac{\Lambda_{\min}^2}{10D_X^4}\right).$$

Proof. We wish to center the matrices $X_sX_s^\top$ by subtracting their (common) expected value. We use that if A and B are symmetric, $\lambda_{\min}(A-B) \le \lambda_{\min}(A) - \lambda_{\min}(B)$. Indeed, denoting by v any eigenvector of A associated with its smallest eigenvalue,

$$\begin{aligned} \lambda_{\min}(A-B) &= \min_x \frac{x^\top(A-B)x}{\|x\|^2} \\ &\le \frac{v^\top(A-B)v}{\|v\|^2} \\ &= \lambda_{\min}(A) - \frac{v^\top Bv}{\|v\|^2} \\ &\le \lambda_{\min}(A) - \min_x \frac{x^\top Bx}{\|x\|^2} \\ &= \lambda_{\min}(A) - \lambda_{\min}(B)\,. \end{aligned}$$

We obtain:

$$\begin{aligned}\lambda_{\min}\left(\sum_{s=1}^{t-1}\frac{X_sX_s^\top}{s^\beta}-\sum_{s=1}^{t-1}\mathbb{E}\left[\frac{X_sX_s^\top}{s^\beta}\right]\right) &\le \lambda_{\min}\left(\sum_{s=1}^{t-1}\frac{X_sX_s^\top}{s^\beta}\right)-\lambda_{\min}\left(\sum_{s=1}^{t-1}\mathbb{E}\left[\frac{X_sX_s^\top}{s^\beta}\right]\right)\\ &= \lambda_{\min}\left(\sum_{s=1}^{t-1}\frac{X_sX_s^\top}{s^\beta}\right)-\Lambda_{\min}\sum_{s=1}^{t-1}\frac{1}{s^\beta}\\ &\le \lambda_{\min}\left(\sum_{s=1}^{t-1}\frac{X_sX_s^\top}{s^\beta}\right)-\Lambda_{\min}\frac{t^{1-\beta}-1}{1-\beta}\,.\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}&\mathbb{P}\left(\lambda_{\min}\left(\sum_{s=1}^{t-1}\frac{X_sX_s^\top}{s^\beta}\right)<\frac{\Lambda_{\min}(t^{1-\beta}-2)}{2(1-\beta)}\right)\\ &\quad\le \mathbb{P}\left(\lambda_{\min}\left(\sum_{s=1}^{t-1}\left(\frac{X_sX_s^\top}{s^\beta}-\mathbb{E}\left[\frac{X_sX_s^\top}{s^\beta}\right]\right)\right)<\frac{\Lambda_{\min}(t^{1-\beta}-2)}{2(1-\beta)}-\Lambda_{\min}\frac{t^{1-\beta}-1}{1-\beta}\right)\\ &\quad= \mathbb{P}\left(\lambda_{\max}\left(\sum_{s=1}^{t-1}\left(\mathbb{E}\left[\frac{X_sX_s^\top}{s^\beta}\right]-\frac{X_sX_s^\top}{s^\beta}\right)\right)>\frac{\Lambda_{\min}t^{1-\beta}}{2(1-\beta)}\right)\,.\end{aligned}$$

We check the assumptions of Theorem 1.4 of Tropp, 2012:

— Obviously $\mathbb{E}\left[\frac{X_sX_s^\top}{s^\beta}\right]-\frac{X_sX_s^\top}{s^\beta}$ is centered,
— $\lambda_{\max}\left(\mathbb{E}\left[\frac{X_sX_s^\top}{s^\beta}\right]-\frac{X_sX_s^\top}{s^\beta}\right)\le\lambda_{\max}\left(\mathbb{E}\left[\frac{X_sX_s^\top}{s^\beta}\right]\right)\le D_X^2$ almost surely.

As $0\preccurlyeq\mathbb{E}\left[\left(\mathbb{E}\left[\frac{X_sX_s^\top}{s^\beta}\right]-\frac{X_sX_s^\top}{s^\beta}\right)^2\right]\preccurlyeq\mathbb{E}\left[\left(\frac{X_sX_s^\top}{s^\beta}\right)^2\right]\preccurlyeq\frac{D_X^4}{s^{2\beta}}I\preccurlyeq\frac{D_X^4}{s^\beta}I$, we get

$$0\preccurlyeq\sum_{s=1}^{t-1}\mathbb{E}\left[\left(\mathbb{E}\left[\frac{X_sX_s^\top}{s^\beta}\right]-\frac{X_sX_s^\top}{s^\beta}\right)^2\right]\preccurlyeq\left(\sum_{s=1}^{t-1}\frac{D_X^4}{s^\beta}\right)I\preccurlyeq\left(D_X^4\frac{t^{1-\beta}}{1-\beta}\right)I\,.$$

Therefore we can apply Theorem 1.4 of Tropp, 2012:

$$\begin{aligned}&\mathbb{P}\left(\lambda_{\max}\left(\sum_{s=1}^{t-1}\left(\mathbb{E}\left[\frac{X_sX_s^\top}{s^\beta}\right]-\frac{X_sX_s^\top}{s^\beta}\right)\right)>\frac{\Lambda_{\min}t^{1-\beta}}{2(1-\beta)}\right)\\ &\quad\le d\exp\left(-\frac{\Lambda_{\min}^2t^{2(1-\beta)}/(8(1-\beta)^2)}{D_X^4t^{1-\beta}/(1-\beta)+D_X^2\Lambda_{\min}t^{1-\beta}/(6(1-\beta))}\right)\\ &\quad= d\exp\left(-t^{1-\beta}\frac{\Lambda_{\min}^2}{8D_X^4}\frac{1/(1-\beta)^2}{1/(1-\beta)+\Lambda_{\min}/(6D_X^2(1-\beta))}\right)\\ &\quad= d\exp\left(-t^{1-\beta}\frac{\Lambda_{\min}^2}{8D_X^4}\left(1-\beta+\frac{\Lambda_{\min}(1-\beta)}{6D_X^2}\right)^{-1}\right)\,.\end{aligned}$$

Using $\Lambda_{\min}/D_X^2\le 1$ and $\beta\ge 0$, we obtain $8(1-\beta+\frac{\Lambda_{\min}(1-\beta)}{6D_X^2})\le 8(1+1/6)=28/3\le 10$,

therefore

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{s=1}^{t-1}\frac{X_sX_s^\top}{s^\beta}\right) < \frac{\Lambda_{\min}(t^{1-\beta}-2)}{2(1-\beta)}\right) \le d\exp\left(-t^{1-\beta}\frac{\Lambda_{\min}^2}{10D_X^4}\right).$$

The result follows from $\frac{1}{2}t^{1-\beta}-2>0$ for $t \ge 4^{1/(1-\beta)}$. □

We can now do a union bound to obtain Proposition 4.4.

Proof of Proposition 4.4. We first move our problem to the setting of Lemma B.5:

$$\begin{aligned}\lambda_{\max}(P_t) &= \lambda_{\min}\left(P_1^{-1}+\sum_{s=1}^{t-1}X_sX_s^\top\alpha_s\right)^{-1}\\ &\le \lambda_{\min}\left(P_1^{-1}+\sum_{s=1}^{t-1}\frac{X_sX_s^\top}{s^\beta}\right)^{-1},\end{aligned}$$

because $\alpha_s \ge 1/s^\beta$. Therefore, for $t \ge 8 \ge 4^{1/(1-\beta)}$,

$$\begin{aligned}\mathbb{P}\left(\lambda_{\max}(P_t) > \frac{4}{\Lambda_{\min}t^{1-\beta}}\right) &\le \mathbb{P}\left(\lambda_{\min}\left(P_1^{-1}+\sum_{s=1}^{t-1}\frac{X_sX_s^\top}{s^\beta}\right)^{-1} > \frac{4}{\Lambda_{\min}t^{1-\beta}}\right)\\ &= \mathbb{P}\left(\lambda_{\min}\left(P_1^{-1}+\sum_{s=1}^{t-1}\frac{X_sX_s^\top}{s^\beta}\right) < \frac{\Lambda_{\min}t^{1-\beta}}{4}\right)\\ &\le \mathbb{P}\left(\lambda_{\min}\left(\sum_{s=1}^{t-1}\frac{X_sX_s^\top}{s^\beta}\right) < \frac{\Lambda_{\min}t^{1-\beta}}{4}\right)\\ &\le d\exp\left(-t^{1-\beta}\frac{\Lambda_{\min}^2}{10D_X^4}\right),\end{aligned}$$

where we applied Lemma B.5 to obtain the last line. We take a union bound to obtain, for any $k \ge 7$,

$$\begin{aligned}\mathbb{P}\left(\exists t>k, \lambda_{\max}(P_t) > \frac{4}{\Lambda_{\min}t^{1-\beta}}\right) &\le \sum_{t>k} d\exp\left(-t^{1-\beta}\frac{\Lambda_{\min}^2}{10D_X^4}\right)\\ &\le d\sum_{t>k}\exp\left(-\lfloor t^{1-\beta}\rfloor\frac{\Lambda_{\min}^2}{10D_X^4}\right)\\ &= d\sum_{m\ge1}\exp\left(-m\frac{\Lambda_{\min}^2}{10D_X^4}\right)\sum_{t>k}\mathbb{1}_{\lfloor t^{1-\beta}\rfloor=m}\end{aligned}$$

We bound $\sum_{t>k}\mathbb{1}_{\lfloor t\rfloor=m}$: for any m

$$\lfloor t^{1-\beta}\rfloor = m \implies m^{1/(1-\beta)} \le t < (m+1)^{1/(1-\beta)},$$

then using $e^x \leq 1 + 2x$ for any $0 \leq x \leq 1$, we have

$$\begin{aligned}(m+1)^{1/(1-\beta)} &= m^{1/(1-\beta)}(1+1/m)^{1/(1-\beta)} \\ &= m^{1/(1-\beta)} \exp(\ln(1+1/m)/(1-\beta)) \\ &\leq m^{1/(1-\beta)} \exp(1/(m(1-\beta))) \\ &\leq m^{1/(1-\beta)}(1+2/(m(1-\beta)))\,,\end{aligned}$$

as long as $m \geq 2 \geq 1/(1-\beta)$. Therefore

$$(m+1)^{1/(1-\beta)} - m^{1/(1-\beta)} + 1 \leq 2m^{1/(1-\beta)-1}/(1-\beta) + 1 \leq 4m + 1 \leq 4(m+1)\,,$$

and that is true for $m = 1$ too. Hence

$$\begin{aligned}\mathbb{P}\left(\exists t > k, \lambda_{\max}(P_t) > \frac{4}{\Lambda_{\min} t^{1-\beta}}\right) &\leq 4d \sum_{m \geq \lfloor k^{1-\beta} \rfloor} (m+1) \exp\left(-m \frac{\Lambda_{\min}^2}{10 D_X^4}\right) \\ &= 4d \frac{\exp\left(-\frac{\Lambda_{\min}^2}{10 D_X^4}\right)^{\lfloor k^{1-\beta} \rfloor}}{1 - \exp\left(-\frac{\Lambda_{\min}^2}{10 D_X^4}\right)} \left(\lfloor k^{1-\beta} \rfloor + 1 + \frac{\exp\left(-\frac{\Lambda_{\min}^2}{10 D_X^4}\right)}{1 - \exp\left(-\frac{\Lambda_{\min}^2}{10 D_X^4}\right)}\right) \\ &\leq 4d \frac{\exp\left(\frac{\Lambda_{\min}^2}{10 D_X^4}\right)}{1 - \exp\left(-\frac{\Lambda_{\min}^2}{10 D_X^4}\right)} \left(k^{1-\beta} + \frac{1}{1 - \exp\left(-\frac{\Lambda_{\min}^2}{10 D_X^4}\right)}\right) \exp\left(-\frac{\Lambda_{\min}^2}{10 D_X^4}\right)^{k^{1-\beta}}\,,\end{aligned}$$

where the second line is obtained deriving both sides of $\sum_{m \geq \lfloor k^{1-\beta} \rfloor} r^{m+1} = \frac{r^{\lfloor k^{1-\beta} \rfloor + 1}}{1-r}$ with respect to r. Also, as $1 - e^{-x} \geq x e^{-x}$ for any $x \in \mathbb{R}$, we get

$$\begin{aligned}&\mathbb{P}\left(\exists t > k, \lambda_{\max}(P_t) > \frac{4}{\Lambda_{\min} t^{1-\beta}}\right) \\ &\quad \leq 4d \frac{10 D_X^4}{\Lambda_{\min}^2} \exp\left(2 \frac{\Lambda_{\min}^2}{10 D_X^4}\right) (k^{1-\beta} + \frac{10 D_X^4}{\Lambda_{\min}^2} \exp\left(\frac{\Lambda_{\min}^2}{10 D_X^4}\right)) \exp\left(-\frac{\Lambda_{\min}^2}{10 D_X^4}\right)^{k^{1-\beta}}\,.\end{aligned}$$

Furthermore, as $x e^{-x} \leq e^{-1}$ for any $x \geq 0$, we get for any $k \geq 7$:

$$\begin{aligned}&\left(k^{1-\beta} + \frac{10 D_X^4}{\Lambda_{\min}^2} \exp\left(\frac{\Lambda_{\min}^2}{10 D_X^4}\right)\right) \exp\left(-k^{1-\beta} \frac{\Lambda_{\min}^2}{20 D_X^4}\right) \\ &\quad \leq \frac{20 D_X^4 e^{-1}}{\Lambda_{\min}^2} \exp\left(\frac{10 D_X^4}{\Lambda_{\min}^2} \exp\left(\frac{\Lambda_{\min}^2}{10 D_X^4}\right) \frac{\Lambda_{\min}^2}{20 D_X^4}\right) \\ &\quad = \frac{20 D_X^4 e^{-1}}{\Lambda_{\min}^2} \exp\left(\frac{1}{2} \exp\left(\frac{\Lambda_{\min}^2}{10 D_X^4}\right)\right)\,.\end{aligned}$$

Combining the last two inequalities, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\exists t > k, \lambda_{\max}(P_t) > \frac{4}{\Lambda_{\min} t^{1-\beta}}\right) \\
&\quad \leq d\frac{800 D_X^8 e^{-1}}{\Lambda_{\min}^4} \exp\left(2\frac{\Lambda_{\min}^2}{10 D_X^4} + \frac{1}{2}\exp\left(\frac{\Lambda_{\min}^2}{10 D_X^4}\right)\right) \exp\left(-k^{1-\beta}\frac{\Lambda_{\min}^2}{20 D_X^4}\right) \\
&\quad \leq d\frac{625 D_X^8}{\Lambda_{\min}^4} \exp\left(-k^{1-\beta}\frac{\Lambda_{\min}^2}{20 D_X^4}\right),
\end{aligned}$$

and the result follows. The last line comes from $\Lambda_{\min} \leq D_X^2$ and consequently

$$800e^{-1}\exp\left(2\frac{\Lambda_{\min}^2}{10 D_X^4} + \frac{1}{2}\exp\left(\frac{\Lambda_{\min}^2}{10 D_X^4}\right)\right) \leq 800 e^{-1+0.2+0.5e^{0.1}} \approx 624.7 \leq 625\,.$$

The condition $k \geq 7$ is not necessary because

$$\left(\frac{20 D_X^4}{\Lambda_{\min}^2}\ln\left(\frac{625 d D_X^8}{\Lambda_{\min}^4 \delta}\right)\right)^{1/(1-\beta)} \geq 20\ln(625\delta^{-1})\,,$$

and either $\delta \geq 1$ and the result is trivial, either $\delta < 1$ and $20\ln(625\delta^{-1}) \geq 128$. □

B.3.3 Convergence of the Truncated Algorithm

In order to prove Theorem 4.4, we state and prove an intermediate lemma.

Lemma B.6. *Let $\theta \in \mathbb{R}^d$.*

1. *For any $\eta > 0$, we have*

$$L(\theta) - L(\theta^\star) > \eta \implies \left\| \frac{\partial L}{\partial \theta}\Big|_\theta \right\| \geq D_\eta$$

   where $D_\eta = \frac{\Lambda_{\min}\sqrt{\eta}}{\sqrt{2}D_X(1+e^{D_X(\|\theta^\star\|+\sqrt{8\eta/D_X^2})})}$.

2. *For any $\varepsilon > 0$, we have*

$$\|\theta - \theta^\star\| > \varepsilon \implies L(\theta) - L(\theta^\star) > \frac{\Lambda_{\min}}{4(1+e^{D_X(\|\theta^\star\|+\varepsilon)})}\varepsilon^2\,.$$

Proof. Both points derive from a second-order identity, turned in an upper-bound in the one case and in a lower-bound in the other. Using $\frac{\partial L}{\partial \theta}(\theta^\star) = 0$, there exists $0 \leq \lambda \leq 1$ such that

$$L(\theta) = L(\theta^\star) + \frac{1}{2}(\theta - \theta^\star)^\top \mathbb{E}\left[\frac{1}{(1+e^{(\lambda\theta+(1-\lambda)\theta^\star)^\top X})(1+e^{-(\lambda\theta+(1-\lambda)\theta^\star)^\top X})} XX^\top\right](\theta - \theta^\star)\,.$$

1. We first have

$$L(\theta) - L(\theta^\star) \leq \frac{D_X^2}{8}\|\theta - \theta^\star\|^2\,.$$

   Assume $L(\theta) - L(\theta^\star) > \eta$. Then $\|\theta - \theta^\star\| \geq \sqrt{8\eta/D_X^2}$. Also, using the Taylor expansion

of $\theta^\star$ around some $\theta_0 \in \mathbb{R}^d$, we get

$$L(\theta^\star) \geq L(\theta_0) + \frac{\partial L}{\partial \theta}\Big|_{\theta_0}^\top (\theta^\star - \theta_0) + \frac{1}{4(1+e^{D_X(\|\theta^\star\|+\|\theta_0-\theta^\star\|)})}(\theta_0-\theta^\star)^\top \mathbb{E}\left[XX^\top\right](\theta_0-\theta^\star)\,,$$

and that yields

$$\frac{\partial L}{\partial \theta}\Big|_{\theta_0}^\top (\theta_0 - \theta^\star) \geq L(\theta_0) - L(\theta^\star) + \frac{\Lambda_{\min}}{4(1+e^{D_X(\|\theta^\star\|+\|\theta_0-\theta^\star\|)})}\|\theta_0-\theta^\star\|^2\,.$$

Therefore, as $L(\theta_0) - L(\theta^\star) \geq 0$,

$$\left\|\frac{\partial L}{\partial \theta}\Big|_{\theta_0}\right\| \geq \frac{\Lambda_{\min}}{4(1+e^{D_X(\|\theta^\star\|+\|\theta_0-\theta^\star\|)})}\|\theta_0-\theta^\star\|\,.$$

Finally, as L is convex of minimum $\theta^\star$,

$$\begin{aligned}\left\|\frac{\partial L}{\partial \theta}\Big|_{\theta}\right\| &\geq \min_{\|\theta_0-\theta^\star\|=\sqrt{8\eta/D_X^2}} \left\|\frac{\partial L}{\partial \theta}\Big|_{\theta_0}\right\| \\ &\geq \frac{\Lambda_{\min}}{4(1+e^{D_X(\|\theta^\star\|+\sqrt{8\eta/D_X^2})})}\sqrt{8\eta/D_X^2} \\ &\geq \frac{\Lambda_{\min}}{\sqrt{2}D_X(1+e^{D_X(\|\theta^\star\|+\sqrt{8\eta/D_X^2})})}\sqrt{\eta}\,.\end{aligned}$$

2. On the other hand we have

$$L(\theta) \geq L(\theta^\star) + \frac{\Lambda_{\min}}{4(1+e^{D_X(\|\theta^\star\|+\|\theta-\theta^\star\|)})}\|\theta-\theta^\star\|^2\,.$$

Thus, as L is convex of minimum $\theta^\star$, if $\|\theta-\theta^\star\| > \varepsilon$ it holds

$$L(\theta) - L(\theta^\star) > \min_{\|\theta_0-\theta^\star\|=\varepsilon} L(\theta_0) - L(\theta^\star) \geq \frac{\Lambda_{\min}}{4(1+e^{D_X(\|\theta^\star\|+\varepsilon)})}\varepsilon^2\,.$$

□

Proof of Theorem 4.4. We prove the convergence of $(L(\hat{\theta}_t))_t$ to $L(\theta^\star)$ and then the convergence of $(\hat{\theta}_t)_t$ to $\theta^\star$ follows. The convergence of $(L(\hat{\theta}_t))_t$ comes from the first point of Lemma B.6. The link between the two convergences is stated in the second point.

To study the evolution of $L(\hat{\theta}_t)$ we first apply a second-order Taylor expansion: for any $t \geq 1$ there exists $0 \leq \alpha_t \leq 1$ such that

$$L(\hat{\theta}_{t+1}) = L(\hat{\theta}_t) + \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^\top (\hat{\theta}_{t+1}-\hat{\theta}_t) + \frac{1}{2}(\hat{\theta}_{t+1}-\hat{\theta}_t)^\top \frac{\partial^2 L}{\partial \theta^2}\Big|_{\hat{\theta}_t+\alpha_t(\hat{\theta}_{t+1}-\hat{\theta}_t)}(\hat{\theta}_{t+1}-\hat{\theta}_t)\,. \tag{B.10}$$

We have $\frac{\partial^2 L}{\partial \theta^2} \preccurlyeq \frac{1}{4}\mathbb{E}[XX^\top]$, therefore, using the update formula on $\hat{\theta}$, the second-order term

is bounded with

$$(\hat{\theta}_{t+1} - \hat{\theta}_t)^\top \frac{\partial^2 L}{\partial \theta^2}\Big|_{\hat{\theta}_t + \alpha_t(\hat{\theta}_{t+1} - \hat{\theta}_t)} (\hat{\theta}_{t+1} - \hat{\theta}_t) \leq \frac{1}{(1 + e^{y_t \hat{\theta}_t^\top X_t})^2} X_t^\top P_{t+1}^\top \frac{\mathbb{E}[XX^\top]}{4} P_{t+1} X_t$$
$$\leq \frac{1}{4} D_X^4 \lambda_{\max}(P_{t+1})^2 \leq \frac{1}{4} D_X^4 \lambda_{\max}(P_t)^2 \,.$$

The first-order term is controlled using the definition of the algorithm:

$$\hat{\theta}_{t+1} - \hat{\theta}_t = \left(P_t - \frac{P_t X_t X_t^\top P_t}{1 + X_t^\top P_t X_t \alpha_t} \alpha_t \right) \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t^\top X_t}} \,,$$

and as $\alpha_t \leq 1$,

$$\left\| -\alpha_t \frac{P_t X_t X_t^\top P_t}{1 + X_t^\top P_t X_t \alpha_t} \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t^\top X_t}} \right\| \leq D_X^3 \lambda_{\max}(P_t)^2 \,.$$

Also, $\left\| \frac{\partial L}{\partial \theta} \right\| \leq D_X$. Substituting our findings in Equation (B.10), we obtain

$$L(\hat{\theta}_{t+1}) \leq L(\hat{\theta}_t) + \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^\top P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t^\top X_t}} + 2 D_X^4 \lambda_{\max}(P_t)^2 \,. \tag{B.11}$$

We define

$$M_t = \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^\top P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t^\top X_t}} - \mathbb{E}\left[\frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^\top P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t^\top X_t}} \mid X_1, y_1, ..., X_{t-1}, y_{t-1} \right]$$
$$= \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^\top P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t^\top X_t}} + \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^\top P_t \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t} \,.$$

Hence we have

$$\frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^\top P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t^\top X_t}} \leq M_t - \lambda_{\min}(P_t) \left\| \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t} \right\|^2 \leq M_t - \frac{1}{t D_X^2} \left\| \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t} \right\|^2 \,,$$

because $P_s \preccurlyeq \frac{I}{s D_X^2}$. Combining it with Equation (B.11) and summing consecutive terms, we obtain, for any $k < t$,

$$L(\hat{\theta}_t) - L(\hat{\theta}_k) \leq \sum_{s=k}^{t-1} \left(M_s - \frac{1}{s D_X^2} \left\| \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_s} \right\|^2 + 2 D_X^4 \lambda_{\max}(P_s)^2 \right) \,. \tag{B.12}$$

We recall that there exists C_δ such that $\mathbb{P}(A_{C_\delta}) \geq 1 - \delta$ where

$$A_{C_\delta} := \bigcap_{t=1}^{\infty} \left(\lambda_{\max}(P_t) \leq \frac{C_\delta}{t^{1-\beta}} \right) .$$

On the previous inequality, we see that the right-hand side is the sum of a martingale and a term which is negative for s large enough, under the event A_{C_δ}.

We are then interested in $\mathbb{P}((L(\hat{\theta}_t) - L(\theta^\star) > \eta) \mid A_{C_\delta})$ for some $\eta > 0$. For $0 \leq k \leq t$, we

define $B_{k,t}$ the event $(\forall k < s < t, L(\hat{\theta}_s) - L(\theta^\star) > \eta/2)$. Then we use the law of total probability:

$$\begin{aligned}
&\mathbb{P}(L(\hat{\theta}_t) - L(\theta^\star) > \eta \mid A_{C_\delta}) \\
&\quad \leq \mathbb{P}\left((L(\hat{\theta}_t) - L(\theta^\star) > \eta) \cap B_{0,t} \mid A_{C_\delta}\right) \\
&\qquad + \sum_{k=1}^{t-1} \mathbb{P}\left((L(\hat{\theta}_t) - L(\theta^\star) > \eta) \cap \big(L(\hat{\theta}_k) - L(\theta^\star) \leq \frac{\eta}{2}\big) \cap B_{k,t} \mid A_{C_\delta}\right) \qquad \text{(B.13)} \\
&\quad \leq \mathbb{P}\left((L(\hat{\theta}_t) - L(\theta^\star) > \eta) \cap B_{0,t} \mid A_{C_\delta}\right) \\
&\qquad + \sum_{k=1}^{t-1} \mathbb{P}\left(\big(L(\hat{\theta}_t) - L(\hat{\theta}_k) > \frac{\eta}{2}\big) \cap B_{k,t} \mid A_{C_\delta}\right) .
\end{aligned}$$

Lemma B.6 yields

$$L(\hat{\theta}_s) - L(\theta^\star) > \frac{\eta}{2} \implies \left\| \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_s} \right\| \geq D_\eta \, .$$

We combine the last equation, along with Equation (B.12) and the definition of A_{C_δ} to get, for any $1 \leq k < t$,

$$\begin{aligned}
\mathbb{P}\left((L(\hat{\theta}_t) - L(\hat{\theta}_k) > \eta/2) \cap B_{k,t} \mid A_{C_\delta}\right) &\leq \mathbb{P}\left(\Big(\sum_{s=k}^{t-1} M_s > f(k,t)\Big) \cap B_{k,t} \mid A_{C_\delta}\right) \\
&\leq \mathbb{P}\left(\sum_{s=k}^{t-1} M_s > f(k,t) \mid A_{C_\delta}\right) ,
\end{aligned}$$

where $f(k,t) = \frac{\eta}{2} + \frac{D_\eta^2}{D_X^2} \sum_{s=k}^{t-1} \frac{1}{s} - 2D_X^4 C_\delta^2 \sum_{s=k}^{t-1} \frac{1}{s^{2(1-\beta)}}$ for any $1 \leq k < t$.

Similarly, we get

$$\mathbb{P}\left((L(\hat{\theta}_t) - L(\theta^\star) > \eta) \cap B_{0,t} \mid A_C\right) \leq \mathbb{P}\left(\sum_{s=1}^{t-1} M_s > f_0(t) \mid A_C\right) ,$$

with $f_0(t) = \eta - (L(\hat{\theta}_1) - L(\theta^\star)) + \frac{D_\eta^2}{D_X^2} \sum_{s=1}^{t-1} \frac{1}{s} - 2D_X^4 C_\delta^2 \sum_{s=1}^{t-1} \frac{1}{s^{2(1-\beta)}}$ for any $t \geq 1$.

We have $\mathbb{E}[M_s \mid X_1, y_1, ..., X_{s-1}, y_{s-1}] = 0$, and almost surely $|M_s| \leq 2D_X^2 \lambda_{\max}(P_s)$. We can therefore apply Azuma-Hoeffding inequality: for t, k such that $f(k,t) > 0$,

$$\mathbb{P}\left(\sum_{s=k}^{t-1} M_s > f(k,t) \mid A_{C_\delta}\right) \leq \exp\left(-f(k,t)^2 \frac{(1-2\beta)\max\big(1/2, (k-1)^{1-2\beta}\big)}{8D_X^4 C_\delta^2}\right) ,$$

because $\sum_{s=k}^{+\infty} \frac{1}{s^{2(1-\beta)}} \leq \frac{1}{(1-2\beta)\max(1/2,(k-1)^{1-2\beta})}$. Similarly, for t such that $f_0(t) > 0$,

$$\mathbb{P}\left(\sum_{s=1}^{t-1} M_s > f_0(t) \mid A_{C_\delta}\right) \leq \exp\left(-f_0(t)^2 \frac{1-2\beta}{16 D_X^4 C_\delta^2}\right) .$$

We need to control $f(k,t), f_0(t)$. We see that for t large enough, when k is small compared to t, $f(k,t)$ is driven by $\frac{D_\eta^2}{D_X^2}\ln(t)$ and when $k \approx t$, $f(k,t)$ is driven by $\eta/2$. The following Lemma formally states these approximations as lower-bounds. We prove it right after the end of this proof.

Lemma B.7. *For* $t \geq \max\left(e^{\frac{16D_X^6 C_\delta^2}{D_\eta^2(1-2\beta)}}, \left(1+\left(\frac{8D_X^4 C_\delta^2}{\eta(1-2\beta)}\right)^{\frac{1}{1-2\beta}}\right)^2\right)$, *it holds*

$$\begin{aligned} f(k,t) &\geq \frac{D_\eta^2}{4D_X^2}\ln(t), && 1 \leq k < \sqrt{t}, \\ f(k,t) &\geq \frac{\eta}{4}, && \sqrt{t} \leq k < t\,. \end{aligned}$$

Similarly, for $t \geq e^{\frac{2D_X^2}{D_\eta^2}\left(L(\hat{\theta}_1)-L(\theta^\star)+\frac{4D_X^4 C_\delta^2}{1-2\beta}\right)}$, *we have*

$$f_0(t) \geq \frac{D_\eta^2}{2D_X^2}\ln(t)\,.$$

Then, defining $C_1 = \frac{D_\eta^4(1-2\beta)}{256 D_X^8 C_\delta^2}$ and $C_2 = \frac{\eta^2(1-2\beta)}{128 D_X^4 C_\delta^2}$, we finally get for t large enough:

$$\begin{aligned} &\mathbb{P}\left((L(\hat{\theta}_t) - L(\theta^\star) > \eta) \cap B_{0,t} \mid A_{C_\delta}\right) \leq \exp\left(-4C_1\ln(t)^2\right), \\ &\mathbb{P}\left((L(\hat{\theta}_t) - L(\theta^\star) > \eta) \cap (L(\hat{\theta}_k) - L(\theta^\star) \leq \frac{\eta}{2}) \cap B_{k,t} \mid A_{C_\delta}\right) \\ &\qquad \leq \begin{cases} \exp\left(-C_1\ln(t)^2\right) & \text{if } 1 \leq k < \sqrt{t}\,, \\ \exp\left(-C_2(k-1)^{1-2\beta}\right) & \text{if } \sqrt{t} \leq k < t\,. \end{cases} \end{aligned}$$

Substituting in Equation (B.13) yields:

$$\begin{aligned} &\mathbb{P}(L(\hat{\theta}_t) - L(\theta^\star) > \eta \mid A_C) \\ &\quad \leq \exp\left(-4C_1\ln(t)^2\right) + \sum_{k=1}^{\lceil\sqrt{t}\rceil-1} \exp\left(-C_1\ln(t)^2\right) + \sum_{k=\lceil\sqrt{t}\rceil}^{t-1} \exp\left(-C_2(k-1)^{1-2\beta}\right) \\ &\quad \leq (\sqrt{t}+1)\exp\left(-C_1\ln(t)^2\right) + t\exp\left(-C_2(\sqrt{t}-1)^{1-2\beta}\right). \end{aligned}$$

Finally, Point 2 of Lemma B.6 allows to obtain the result: defining $\eta = \frac{\Lambda_{\min}\varepsilon^2}{4(1+e^{D_X(\|\theta^\star\|+\varepsilon)})}$, we obtain

$$\begin{aligned} \mathbb{P}(\|\hat{\theta}_t - \theta^\star\| > \varepsilon \mid A_{C_\delta}) &\leq \mathbb{P}(L(\hat{\theta}_t) - L(\theta^\star) > \eta \mid A_{C_\delta}) \\ &\leq (\sqrt{t}+1)\exp\left(-C_1\ln(t)^2\right) + t\exp\left(-C_2(\sqrt{t}-1)^{1-2\beta}\right). \end{aligned}$$

In order to obtain the constants involved in the Theorem, we write

$$D_\eta = \frac{\Lambda_{\min}\sqrt{\frac{\Lambda_{\min}\varepsilon^2}{4(1+e^{D_X(\|\theta^\star\|+\varepsilon)})}}}{2D_X(1+\exp\left(D_X(\|\theta^\star\|+\sqrt{\frac{\Lambda_{\min}\varepsilon^2}{D_X^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})}}\right))} \geq \left(\frac{\Lambda_{\min}}{1+e^{D_X(\|\theta^\star\|+\varepsilon)}}\right)^{3/2}\frac{\varepsilon}{4D_X},$$

$$C_1 \geq \frac{\Lambda_{\min}^6(1-2\beta)\varepsilon^4}{2^{16}D_X^{12}C_\delta^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^6},$$

$$C_2 \geq \frac{\Lambda_{\min}^2(1-2\beta)\varepsilon^4}{2^{11}D_X^4C_\delta^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^2},$$

and the conditions of Lemma B.7 become

$$t \geq \exp\left(\frac{2^8D_X^8C_\delta^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^3}{\Lambda_{\min}^3(1-2\beta)\varepsilon^2}\right),$$

$$t \geq \left(1+\left(\frac{32D_X^4C_\delta^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})}{(1-2\beta)\Lambda_{\min}\varepsilon^2}\right)^{\frac{1}{1-2\beta}}\right)^2,$$

$$t \geq \exp\left(\frac{32D_X^4(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^3}{\Lambda_{\min}^3\varepsilon^2}\left(L(\hat{\theta}_1)-L(\theta^\star)+\frac{4D_X^4C_\delta^2}{1-2\beta}\right)\right).$$

We would like to obtain a single condition on t, thus we write

$$\begin{aligned}
&\left(1+\left(\frac{32D_X^4C_\delta^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})}{(1-2\beta)\Lambda_{\min}\varepsilon^2}\right)^{\frac{1}{1-2\beta}}\right)^2 \\
&\quad= \exp\left(2\ln\left(1+\left(\frac{32D_X^4C_\delta^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})}{(1-2\beta)\Lambda_{\min}\varepsilon^2}\right)^{\frac{1}{1-2\beta}}\right)\right) \\
&\quad\leq \exp\left(\frac{2}{1-2\beta}\ln\left(1+\frac{32D_X^4C_\delta^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})}{(1-2\beta)\Lambda_{\min}\varepsilon^2}\right)\right) \\
&\quad\leq \exp\left(\frac{2}{1-2\beta}\sqrt{\frac{32D_X^4C_\delta^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})}{(1-2\beta)\Lambda_{\min}\varepsilon^2}}\right) \\
&\quad\leq \exp\left(\frac{2^8D_X^8C_\delta^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^3}{\Lambda_{\min}^3(1-2\beta)^{3/2}\varepsilon^2}\right),
\end{aligned}$$

The third line is obtained with the inequality $\ln(1+x) \leq \sqrt{x}$ for any $x > 0$. Obviously, as $0 < 1-2\beta < 1$, the first threshold on t is bounded by:

$$\exp\left(\frac{2^8D_X^8C_\delta^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^3}{\Lambda_{\min}^3(1-2\beta)\varepsilon^2}\right) \leq \exp\left(\frac{2^8D_X^8C_\delta^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^3}{\Lambda_{\min}^3(1-2\beta)^{3/2}\varepsilon^2}\right).$$

To handle the third one, we use $D_X^2C_\delta \geq \frac{4D_X^2}{\Lambda_{\min}} \geq 4$ and as $\hat{\theta}_1 = 0$ we obtain $L(\hat{\theta}_1) - L(\theta^\star) \leq$

$\ln 2 \leq \frac{4D_X^4 C_\delta^2}{1-2\beta}$, hence

$$\exp\left(\frac{32D_X^4(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^3}{\Lambda_{\min}^3\varepsilon^2}\left(L(\hat{\theta}_1)-L(\theta^\star)+\frac{4D_X^4C_\delta^2}{1-2\beta}\right)\right) \leq \exp\left(\frac{2^8D_X^8C_\delta^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^3}{\Lambda_{\min}^3(1-2\beta)^{3/2}\varepsilon^2}\right).$$

□

Proof of Lemma B.7. We recall that for any $k \geq 1$,

$$\sum_{s=k}^{t-1}\frac{1}{s} \geq \ln t - \ln k\,, \qquad \sum_{s=k}^{t-1}\frac{1}{s^{2(1-\beta)}} \leq \frac{1}{1-2\beta}\frac{1}{\max(1/2,(k-1)^{1-2\beta})}\,.$$

Therefore:

$$f(k,t) \geq \frac{\eta}{2} + \frac{D_\eta^2}{D_X^2}(\ln t - \ln k) - \frac{2D_X^4C_\delta^2}{1-2\beta}\frac{1}{\max(1/2,(k-1)^{1-2\beta})}\,,$$
$$f_0(t) \geq \eta - (L(\hat{\theta}_1)-L(\theta^\star)+\frac{D_\eta^2}{D_X^2}\ln t - \frac{4D_X^4C_\delta^2}{1-2\beta}\,.$$

— For any $1 \leq k < \sqrt{t}$, it holds $\ln k \leq \frac{1}{2}\ln t$, and we have

$$f(k,t) \geq \frac{D_\eta^2}{2D_X^2}\ln(t) - \frac{4D_X^4C_\delta^2}{1-2\beta}\,.$$

Taking $t \geq e^{\frac{16D_X^6C_\delta^2}{D_\eta^2(1-2\beta)}}$ yields $f(k,t) \geq \frac{D_\eta^2}{4D_X^2}\ln(t)$.

— For $t \geq 2$ and any $k \geq \sqrt{t}$, we have

$$f(k,t) \geq \frac{\eta}{2} - \frac{2D_X^4C_\delta^2}{(1-2\beta)(k-1)^{1-2\beta}} \geq \frac{\eta}{2} - \frac{2D_X^4C_\delta^2}{(1-2\beta)(\sqrt{t}-1)^{1-2\beta}}\,.$$

Then if $t \geq \left(1+\left(\frac{8D_X^4C_\delta^2}{\eta(1-2\beta)}\right)^{\frac{1}{1-2\beta}}\right)^2$, we get $f(k,t) \geq \frac{\eta}{4}$.

— Last point comes from $f_0(t) \geq \frac{D_\eta^2}{D_X^2}\ln t - (L(\hat{\theta}_1)-L(\theta^\star) - \frac{4D_X^4C_\delta^2}{1-2\beta}$.

□

Proof of Corollary 4.2. We apply Theorem 4.4: for any $t \geq \exp\left(\frac{2^8D_X^8C_{\delta/2}^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^3}{\Lambda_{\min}^3(1-2\beta)^{3/2}\varepsilon^2}\right)$,

$$\mathbb{P}(\|\hat{\theta}_t-\theta^\star\| > \varepsilon \mid A_{C_{\delta/2}}) \leq (\sqrt{t}+1)\exp\left(-C_1\ln(t)^2\right) + t\exp\left(-C_2(\sqrt{t}-1)^{1-2\beta}\right)\,,$$

where

$$C_1 = \frac{\Lambda_{\min}^6(1-2\beta)\varepsilon^4}{2^{16}D_X^{12}C_{\delta/2}^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^6}, \qquad C_2 = \frac{\Lambda_{\min}^2(1-2\beta)\varepsilon^4}{2^{11}D_X^4C_{\delta/2}^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^2}\,.$$

We use a union bound: for any $T \geq \exp\left(\frac{2^8 D_X^8 C_{\delta/2}^2 (1+e^{D_X(\|\theta^\star\|+\varepsilon)})^3}{\Lambda_{\min}^3 (1-2\beta)^{3/2}\varepsilon^2}\right)$,

$$\mathbb{P}\left(\bigcup_{t=T+1}^{\infty} (\|\hat{\theta}_t - \theta^\star\| > \varepsilon) \mid A_{C_{\delta/2}}\right) \leq \sum_{t>T} (\sqrt{t}+1)\exp\left(-C_1 \ln(t)^2\right) + \sum_{t>T} t \exp\left(-C_2(\sqrt{t}-1)^{1-2\beta}\right).$$

— If $T \geq e^{\frac{3}{2C_1}}$, we have

$$\sum_{t>T} (\sqrt{t}+1)\exp\left(-C_1 \ln(t)^2\right) \leq \sum_{t>T} (\sqrt{t}+1)\frac{1}{t^{5/2}} \leq 2/T\,.$$

— For $t \geq 4$, $1 - 1/\sqrt{t} \geq 1/2$, then for $t \geq \left(\frac{12}{C_2(1-2\beta)}\right)^{4/(1-2\beta)}$,

$$\begin{aligned} t^3 \exp\left(-C_2(\sqrt{t}-1)^{1-2\beta}\right) &\leq \exp\left(3\ln(t) - \frac{C_2}{2} t^{(1-2\beta)/2}\right) \\ &\leq \exp\left(\frac{12}{1-2\beta}\ln\left(\frac{12}{C_2(1-2\beta)}\right) - \frac{6}{1-2\beta}\left(\frac{12}{C_2(1-2\beta)}\right)\right) \\ &\leq 1\,, \end{aligned}$$

because for any $x > 0$, we have $\ln x \leq x/2$.
Thus for $T \geq \left(\frac{12}{C_2(1-2\beta)}\right)^{4/(1-2\beta)}$

$$\sum_{t>T} t \exp\left(-C_2(\sqrt{t}-1)^{1-2\beta}\right) \leq 1/T\,.$$

Finally, for T satisfying the previous conditions as well as $T \geq 6\delta^{-1}$, we obtain

$$\mathbb{P}\left(\bigcup_{t=T+1}^{\infty} (\|\hat{\theta}_t - \theta^\star\| > \varepsilon) \mid A_{C_{\delta/2}}\right) \leq 3/T \leq \delta/2\,.$$

We now compare the constants involved. As long as $\varepsilon D_X \leq 1$, we have

$$\exp\left(\frac{2^8 D_X^8 C_{\delta/2}^2 (1+e^{D_X(\|\theta^\star\|+\varepsilon)})^3}{\Lambda_{\min}^3 (1-2\beta)^{3/2}\varepsilon^2}\right) \leq \exp\left(\frac{3\cdot 2^{15} D_X^{12} C_{\delta/2}^2 (1+e^{D_X(\|\theta^\star\|+\varepsilon)})^6}{\Lambda_{\min}^6 (1-2\beta)^{3/2}\varepsilon^4}\right).$$

Furthermore, as $1-2\beta \leq 1$, we have

$$\begin{aligned} \exp\left(\frac{3}{2C_1}\right) &= \exp\left(\frac{3\cdot 2^{15} D_X^{12} C_{\delta/2}^2 (1+e^{D_X(\|\theta^\star\|+\varepsilon)})^6}{\Lambda_{\min}^6 (1-2\beta)\varepsilon^4}\right) \\ &\leq \exp\left(\frac{3\cdot 2^{15} D_X^{12} C_{\delta/2}^2 (1+e^{D_X(\|\theta^\star\|+\varepsilon)})^6}{\Lambda_{\min}^6 (1-2\beta)^{3/2}\varepsilon^4}\right). \end{aligned}$$

Finally,

$$\begin{aligned}
\left(\frac{12}{C_2(1-2\beta)}\right)^{4/(1-2\beta)} &= \exp\left(\frac{4}{1-2\beta}\ln\frac{12}{C_2(1-2\beta)}\right) \\
&= \exp\left(\frac{4}{1-2\beta}\ln\frac{12\cdot 2^{11}D_X^4 C_{\delta/2}^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^2}{\Lambda_{\min}^2(1-2\beta)^2\varepsilon^4}\right) \\
&= \exp\left(\frac{8}{1-2\beta}\ln\frac{12\cdot 2^{11}D_X^4 C_{\delta/2}^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^2}{\Lambda_{\min}^2(1-2\beta)\varepsilon^4}\right) \\
&\le \exp\left(\frac{8}{1-2\beta}\sqrt{\frac{3\cdot 2^{13}D_X^4 C_{\delta/2}^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^2}{\Lambda_{\min}^2(1-2\beta)\varepsilon^4}}\right) \\
&= \exp\left(\frac{\sqrt{6}2^9 D_X^2 C_{\delta/2}(1+e^{D_X(\|\theta^\star\|+\varepsilon)})}{\Lambda_{\min}(1-2\beta)^{3/2}\varepsilon^2}\right) \\
&\le \exp\left(\frac{3\cdot 2^{15}D_X^{12} C_{\delta/2}^2(1+e^{D_X(\|\theta^\star\|+\varepsilon)})^6}{\Lambda_{\min}^6(1-2\beta)^{3/2}\varepsilon^4}\right).
\end{aligned}$$

□

B.4 Proofs of Section 4.5

Proof of Proposition 4.5. The first order condition of the optimum yields

$$\arg\min_{\theta\in\mathbb{R}^d}\sum_{s=1}^{t-1}(y_s-\theta^\top X_s)^2+\frac{1}{2}(\theta-\hat\theta_1)^\top P_1^{-1}(\theta-\hat\theta_1)=\hat\theta_1+P_t\sum_{s=1}^{t-1}(y_s-\hat\theta_1^\top X_s)X_s\,.$$

Therefore we prove recursively that $\hat\theta_t-\hat\theta_1=P_t\sum_{s=1}^{t-1}(y_s-\hat\theta_1^\top X_s)X_s$. It is clearly true at $t=1$. Assuming it is true for some $t\ge 1$, we use the update formula

$$\begin{aligned}
\hat\theta_{t+1}-\hat\theta_1 &= (I-P_{t+1}X_tX_t^\top)(\hat\theta_t-\hat\theta_1)+P_{t+1}y_tX_t-P_{t+1}X_tX_t^\top\hat\theta_1 \\
&= (I-P_{t+1}X_tX_t^\top)P_t\sum_{s=1}^{t-1}(y_s-\hat\theta_1^\top X_s)X_s+P_{t+1}(y_t-\hat\theta_1^\top X_t)X_t\,.
\end{aligned}$$

We conclude with the following identity:

$$(I-P_{t+1}X_tX_t^\top)P_t=P_t-P_tX_tX_t^\top P_t+\frac{P_tX_tX_t^\top P_tX_tX_t^\top P_t}{X_t^\top P_tX_t+1}=P_t-\frac{P_tX_tX_t^\top P_t}{X_t^\top P_tX_t+1}=P_{t+1}\,.$$

□

B.4.1 Proof of Theorem 4.5

We first prove a result controlling the first estimates of the algorithm.

Lemma B.8. *Provided that Assumptions 4.1, 4.2 and 4.4 are satisfied, starting from any $\hat{\theta}_1 \in \mathbb{R}^d$ and $P_1 \succ 0$, for any $\delta > 0$, it holds simultaneously*

$$\|\hat{\theta}_t - \theta^\star\| \leq \|\hat{\theta}_1 - \theta^\star\| + \lambda_{\max}(P_1) D_X \left((3\sigma + D_{\text{approx}})(t-1) + 3\sigma \ln \delta^{-1} \right), \qquad t \geq 1,$$

with probability at least $1 - \delta$.

Proof. From Proposition 4.5, we obtain, for any $t \geq 1$, $\hat{\theta}_t - \hat{\theta}_1 = P_t \sum_{s=1}^{t-1} (y_s - \hat{\theta}_1^\top X_s) X_s$. Consequently,

$$\begin{aligned}
\hat{\theta}_t - \theta^\star &= P_t \sum_{s=1}^{t-1} (y_s - \hat{\theta}_1^\top X_s) X_s - P_t \left(P_1^{-1} + \sum_{s=1}^{t-1} X_s X_s^\top \right) (\theta^\star - \hat{\theta}_1) \\
&= P_t \sum_{s=1}^{t-1} (y_s - \theta^{\star\top} X_s) X_s + P_t P_1^{-1} (\hat{\theta}_1 - \theta^\star) \,,
\end{aligned}$$

and using $P_t P_1^{-1} \preccurlyeq I$, we obtain

$$\begin{aligned}
\|\hat{\theta}_t - \theta^\star\| &\leq \|\hat{\theta}_1 - \theta^\star\| + \lambda_{\max}(P_t) D_X \sum_{s=1}^{t-1} |y_s - \theta^{\star\top} X_s| \\
&\leq \|\hat{\theta}_1 - \theta^\star\| + \lambda_{\max}(P_1) D_X \sum_{s=1}^{t-1} \left(|y_s - \mathbb{E}[y_s \mid X_s]| + D_{\text{app}} \right) . \qquad \text{(B.14)}
\end{aligned}$$

We apply Lemma 1.4 of Rigollet and Hütter, 2015 in the second line of the following: for any μ such that $0 < \mu < \frac{1}{2\sqrt{2}\sigma}$,

$$\begin{aligned}
\mathbb{E}\left[\exp(\mu |y_t - \mathbb{E}[y_t \mid X_t]|)\right] &= 1 + \sum_{i \geq 1} \frac{\mu^i \mathbb{E}[|y_t - \mathbb{E}[y_t \mid X_t]|^i]}{i!} \\
&\leq 1 + \sum_{k \geq 1} \frac{\mu^i (2\sigma^2)^{i/2} i \Gamma(i/2)}{i!} \\
&\leq 1 + \sum_{i \geq 1} \left(\sqrt{2}\mu\sigma \right)^i, \qquad \text{because } \Gamma(i/2) \leq \Gamma(i) = (i-1)! \\
&\leq 1 + 2\sqrt{2}\mu\sigma, \qquad \text{because } 0 < \sqrt{2}\mu\sigma \leq \frac{1}{2} \\
&\leq \exp\left(2\sqrt{2}\mu\sigma \right) .
\end{aligned}$$

Therefore $\left(\exp\left(\frac{1}{2\sqrt{2}\sigma} \sum_{s=1}^{t} (|y_s - \mathbb{E}[y_s \mid X_s]| - 2\sqrt{2}\sigma) \right) \right)_t$ is a super-martingale to which we can apply Lemma B.1. We obtain, for any $\delta > 0$,

$$\sum_{s=1}^{t-1} |y_t - \mathbb{E}[y_t \mid X_t]| \leq 2\sqrt{2}(t-1)\sigma + 2\sqrt{2}\sigma \ln \delta^{-1}, \qquad t \geq 1,$$

with probability at least $1 - \delta$. The result follows from Equation (B.14) and $2\sqrt{2} \leq 3$. □

Proof of Theorem 4.5. We first apply Theorem 4.2: with probability at least $1 - 5\delta$, it holds

simultaneously for all $n \geq T(\varepsilon, \delta)$

$$\sum_{t=T(\varepsilon,\delta)+1}^{n} L(\hat{\theta}_t) - L(\theta^\star) \leq \frac{15}{2} d \left(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2\right) \ln\left(1 + (n - T(\varepsilon,\delta)) \frac{\lambda_{\max}(P_1) D_X^2}{d}\right) + 5\lambda_{\max}\left(P_{T(\varepsilon,\delta)+1}^{-1}\right) \varepsilon^2 + 115\left(\sigma^2(4 + \frac{\lambda_{\max}(P_1) D_X^2}{4}) + D_{\text{app}}^2 + 2\varepsilon^2 D_X^2\right) \ln \delta^{-1} .$$

Moreover, $\lambda_{\max}\left(P_{T(\varepsilon,\delta)+1}^{-1}\right) \leq \lambda_{\max}(P_1^{-1}) + T(\varepsilon,\delta) D_X^2$.

Then we derive a bound on the first $T(\varepsilon, \delta)$ terms. For any $t \geq 1$, we have $L(\hat{\theta}_t) - L(\theta^\star) \leq D_X^2 \|\hat{\theta}_t - \theta^\star\|^2$, thus, using $(a+b)^2 \leq 2(a^2+b^2)$ and applying Lemma B.8 we obtain the simultaneous property

$$L(\hat{\theta}_t) - L(\theta^\star) \leq 2D_X^2(\|\hat{\theta}_1 - \theta^\star\| + 3\lambda_{\max}(P_1) D_X \sigma \ln \delta^{-1})^2 + 2\lambda_{\max}(P_1)^2 D_X^4 (3\sigma + D_{\text{app}})^2 (t-1)^2, \qquad t \geq 1,$$

with probability at least $1 - \delta$. A summation argument yields, for any $\delta > 0$,

$$\sum_{t=1}^{T(\varepsilon,\delta)} L(\hat{\theta}_t) - L(\theta^\star) \leq 2D_X^2(\|\hat{\theta}_1 - \theta^\star\| + 3\lambda_{\max}(P_1) D_X \sigma \ln \delta^{-1})^2 T(\varepsilon,\delta) + \lambda_{\max}(P_1)^2 D_X^4 (3\sigma + D_{\text{app}})^2 \frac{(T(\varepsilon,\delta)-1) T(\varepsilon,\delta)(2T(\varepsilon,\delta)-1)}{3} ,$$

with probability at least $1 - \delta$. □

B.4.2 Definition of $T(\varepsilon, \delta)$

We now focus on the definition of $T(\varepsilon, \delta)$. We first transcript the result of Hsu, Kakade, and Zhang, 2012 to our notations in the following lemma.

Lemma B.9. *Provided that Assumptions 4.1, 4.2 and 4.4 are satisfied, starting from any $\hat{\theta}_1 \in \mathbb{R}^d$ and $P_1 = p_1 I, p_1 > 0$, we have, for any $0 < \delta < e^{-2.6}$ and $t \geq 6 \frac{D_X^2}{\Lambda_{\min}}(\ln d + \ln \delta^{-1})$,*

$$\|\hat{\theta}_{t+1} - \theta^\star\|_\Sigma^2 \leq \frac{3}{t}\left(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{2p_1} + \frac{D_X^2}{\Lambda_{\min}} D_{\text{app}}^2 \frac{4(1+\sqrt{8\ln\delta^{-1}})}{0.07^2} + \frac{3\sigma^2(d/0.035 + \ln\delta^{-1})}{0.07}\right) + \frac{12}{0.07^2 t^2}\left(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{p_1}\frac{D_X^2}{\Lambda_{\min}}(1+\sqrt{8\ln\delta^{-1}}) + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\text{app}} + D_X\|\theta^\star\|) + \frac{\|\hat{\theta}_1 - \theta^\star\|}{\sqrt{2p_1}}\right)^2 (\ln\delta^{-1})^2\right),$$

with probability at least $1 - 4\delta$.

Proof. We first observe that

$$\arg\min_{w\in\mathbb{R}^d} \frac{1}{t}\sum_{s=1}^{t}(y_s - w^\top X_s)^2 + \lambda\|w - \hat{\beta}_1\|^2 = \arg\min_{w\in\mathbb{R}^d} \frac{1}{t}\sum_{s=1}^{t}(y_s - \hat{\beta}_1^\top X_s - w^\top X_s)^2 + \lambda\|w\|^2 ,$$

therefore we apply the ridge analysis of Hsu, Kakade, and Zhang, 2012 to $(X_s, y_s - \hat{\beta}_1^\top X_s)$. We note that $(y_s - \hat{\beta}_1^\top X_s)$ has the same variance proxy and the same approximation error, therefore it only amounts to translate the optimal w, that is denoted by β.

For any $\lambda > 0$, we observe that

$$d_{2,\lambda} \le d_{1,\lambda} \le d\,, \qquad \rho_\lambda \le \frac{D_X}{\sqrt{d_{1,\lambda}\Lambda_{\min}}}\,, \qquad b_\lambda \le \rho_\lambda(D_{\rm app} + D_X\|\beta - \hat{\beta}_1\|)\,.$$

Therefore we can apply Theorem 16 of Hsu, Kakade, and Zhang, 2012: for $0 < \delta < e^{-2.6}$ and $t \ge 6\frac{D_X}{\sqrt{\Lambda_{\min}}}(\ln(d) + \ln\delta^{-1})$, it holds that $\|\hat{\beta}_{t+1,\lambda} - \beta\|_\Sigma^2 = 3(\|\beta_\lambda - \beta\|_\Sigma^2 + \varepsilon_{\rm bs} + \varepsilon_{\rm vr})$ with probability $1 - 4\delta$, with

$$\begin{aligned}
\varepsilon_{\rm bs} &\le \frac{4}{0.07^2}\Bigg(\frac{\frac{D_X^2}{\Lambda_{\min}}\mathbb{E}[(\mathbb{E}[y \mid X] - \beta^\top X)^2] + (1 + \frac{D_X^2}{\Lambda_{\min}})\|\beta_\lambda - \beta\|_\Sigma^2}{t}(1 + \sqrt{8\ln\delta^{-1}}) \\
&\qquad + \frac{(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\rm app} + D_X\|\beta - \hat{\beta}_1\|) + \|\beta_\lambda - \beta\|_\Sigma)^2}{t^2}(\ln\delta^{-1})^2\Bigg)\,, \\
\delta_f &\le \frac{1}{\sqrt{t}}\frac{D_X}{\sqrt{\Lambda_{\min}}}(1 + \sqrt{8\ln\delta^{-1}}) + \frac{1}{t}\frac{4\sqrt{\frac{D_X^4}{\Lambda_{\min}^2 d} + 1}}{3}\ln\delta^{-1}\,, \\
\varepsilon_{\rm vr} &\le \frac{\sigma^2 d(1+\delta_f)}{0.07^2 t} + \frac{2\sigma^2\sqrt{d(1+\delta_f)\ln\delta^{-1}}}{0.07^{3/2}t} + \frac{2\sigma^2\ln\delta^{-1}}{0.07t}\,.
\end{aligned}$$

Moreover $\mathbb{E}[(\mathbb{E}[y \mid X] - \beta^\top X)^2] \le D_{\rm app}^2$ and $\Lambda_{\min} \le D_X^2$, hence, using $\|\beta_\lambda - \beta\|_\Sigma \le \lambda\|\beta - \hat{\beta}_1\|$ we transfer the result in our KF notations, that is, $\hat{\theta}_t = \hat{\beta}_{t,p_1^{-1}/2(t-1)}, \hat{\beta}_1 = \hat{\theta}_1, \beta = \theta^\star$. We obtain, for any $0 < \delta < e^{-2.6}$ and $t \ge 6\frac{D_X}{\sqrt{\Lambda_{\min}}}(\ln(d) + \ln\delta^{-1})$,

$$\begin{aligned}
\varepsilon_{\rm bs} &\le \frac{4}{0.07^2}\Bigg(\frac{\frac{D_X^2}{\Lambda_{\min}}D_{\rm app}^2 + \frac{D_X^2}{\Lambda_{\min}}\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{p_1 t}}{t}(1 + \sqrt{8\ln\delta^{-1}}) \\
&\qquad + \frac{(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\rm app} + D_X\|\theta^\star\|) + \frac{\|\hat{\theta}_1 - \theta^\star\|}{\sqrt{2p_1 t}})^2}{t^2}(\ln\delta^{-1})^2\Bigg)\,, \\
\delta_f &\le \frac{1}{\sqrt{t}}\frac{D_X}{\sqrt{\Lambda_{\min}}}(1 + \sqrt{8\ln\delta^{-1}}) + \frac{1}{t}\frac{4\sqrt{\frac{D_X^4}{\Lambda_{\min}^2 d} + 1}}{3}\ln\delta^{-1}\,, \\
\varepsilon_{\rm vr} &\le \frac{\sigma^2 d(1+\delta_f)}{0.07^2 t} + \frac{2\sigma^2\sqrt{d(1+\delta_f)\ln\delta^{-1}}}{0.07^{3/2}t} + \frac{2\sigma^2\ln\delta^{-1}}{0.07t}\,, \\
\|\hat{\theta}_{t+1} - \theta^\star\|_\Sigma^2 &\le 3\left(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{2p_1 t} + \varepsilon_{\rm bs} + \varepsilon_{\rm vr}\right)\,,
\end{aligned}$$

with probability at least $1 - 4\delta$. For $t \ge \frac{D_X^2}{\Lambda_{\min}}\ln\delta^{-1}$, as $\ln\delta^{-1} \ge 1$, we get

$$\delta_f \le \frac{1}{\sqrt{6\ln\delta^{-1}}}(1 + \sqrt{8\ln\delta^{-1}}) + \frac{1}{6}\frac{4}{3}\sqrt{\frac{1}{d} + 1} \le \frac{1 + \sqrt{8}}{\sqrt{6}} + \frac{2\sqrt{2}}{9} \approx 1.9 \le 2\,.$$

Thus, as $\sqrt{ab} \leq \frac{a+b}{2}$ for any $a, b > 0$, we have

$$\begin{aligned}\varepsilon_{\text{vr}} &\leq \frac{\sigma^2}{0.07t}\left(\frac{3d}{0.07} + 2\sqrt{\frac{3d\ln\delta^{-1}}{0.07}} + 2\ln\delta^{-1}\right) \\ &\leq \frac{\sigma^2}{0.07t}\left(\frac{6d}{0.07} + 3\ln\delta^{-1}\right) \\ &\leq \frac{3\sigma^2(d/0.035 + \ln\delta^{-1})}{0.07t}\,.\end{aligned}$$

It yields the result. □

Lemma B.9 allows the definition of an explicit value for $T(\varepsilon, \delta)$, as displayed in the following Corollary.

Corollary B.1. *Assumption 4.5 is satisfied for* $T(\varepsilon,\delta) = \max(T_1(\delta), T_2(\varepsilon,\delta), T_3(\varepsilon,\delta))$ *where we define*

$$\begin{aligned}T_1(\delta) &= \max\left(12\frac{D_X^2}{\Lambda_{\min}}(\ln d + \ln\delta^{-1}), \frac{48D_X^2}{\Lambda_{\min}}\ln\frac{24D_X^2}{\Lambda_{\min}}\right), \\ T_2(\varepsilon,\delta) &= \frac{24\varepsilon^{-1}}{\Lambda_{\min}}\left(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{2p_1} + \frac{D_X^2}{\Lambda_{\min}}D_{\text{app}}^2\frac{4(1+\sqrt{8\ln\delta^{-1}})}{0.07^2} + \frac{3\sigma^2(d/0.035+\ln\delta^{-1})}{0.07}\right) \\ &\quad \ln\frac{12\varepsilon^{-1}}{\Lambda_{\min}}\left(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{2p_1} + \frac{D_X^2}{\Lambda_{\min}}D_{\text{app}}^2\frac{4(1+\sqrt{8\ln\delta^{-1}})}{0.07^2} + \frac{3\sigma^2(d/0.035+\ln\delta^{-1})}{0.07}\right), \\ T_3(\varepsilon,\delta) &= \sqrt{\frac{96\varepsilon^{-1}}{0.07^2\Lambda_{\min}}}\Bigg(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{p_1}\frac{D_X^2}{\Lambda_{\min}}(1+\sqrt{8\ln\delta^{-1}}) \\ &\qquad + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\text{app}} + D_X\|\theta^\star\|) + \frac{\|\hat{\theta}_1 - \theta^\star\|}{\sqrt{2p_1}}\right)^2(\ln\delta^{-1})^2\Bigg)^{1/2} \\ &\quad \ln\frac{96\varepsilon^{-1}}{0.07^2\Lambda_{\min}}\Bigg(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{2p_1}(1+\frac{D_X^2}{\Lambda_{\min}})(1+\sqrt{8\ln\delta^{-1}}) \\ &\qquad + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\text{app}} + D_X\|\theta^\star\|) + \frac{\|\hat{\theta}_1 - \theta^\star\|}{\sqrt{2p_1}}\right)^2(\ln\delta^{-1})^2\Bigg).\end{aligned}$$

We recall that for any $\eta \leq 1$, we have $\frac{\ln t}{t} \leq \eta$ for $t \geq 2\eta^{-1}\ln(\eta^{-1})$, and we use it in the following proof.

Proof of Corollary B.1. We define $\delta_t = \delta/t^2$ for any $t \geq 1$. In order to apply Lemma B.9 with a union bound, we need $t \geq 6\frac{D_X^2}{\Lambda_{\min}}(\ln d + \ln\delta_t^{-1})$. If $t \geq 12\frac{D_X^2}{\Lambda_{\min}}(\ln d + \ln\delta^{-1})$ and $t \geq \frac{48D_X^2}{\Lambda_{\min}}\ln\frac{24D_X^2}{\Lambda_{\min}}$,

we obtain

$$\begin{aligned} t &\geq \frac{t}{2} + \frac{\sqrt{t}}{2}\sqrt{t} \\ &\geq 6\frac{D_X^2}{\Lambda_{\min}}(\ln d + \ln \delta^{-1}) + \frac{12D_X^2}{\Lambda_{\min}} \ln t, \qquad \text{as } \ln t \leq \sqrt{t} \\ &= 6\frac{D_X^2}{\Lambda_{\min}}(\ln d + \ln \delta_t^{-1})\,. \end{aligned}$$

Therefore, we define $T_1(\delta) = \max\left(12\frac{D_X^2}{\Lambda_{\min}}(\ln d + \ln \delta^{-1}), \frac{48D_X^2}{\Lambda_{\min}} \ln \frac{24D_X^2}{\Lambda_{\min}}\right)$, and we apply Lemma B.9. We get the simultaneous property

$$\begin{aligned} \|\hat{\theta}_{t+1} - \theta^\star\|_\Sigma^2 \leq & \frac{3}{t}\left(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{2p_1} + \frac{D_X^2}{\Lambda_{\min}} D_{\text{app}}^2 \frac{4(1+\sqrt{8\ln \delta_t^{-1}})}{0.07^2} + \frac{3\sigma^2(d/0.035 + \ln \delta_t^{-1})}{0.07}\right) \\ & + \frac{12}{0.07^2 t^2}\left(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{p_1}\frac{D_X^2}{\Lambda_{\min}}(1+\sqrt{8\ln \delta_t^{-1}})\right. \\ & \qquad \left. + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\text{app}} + D_X\|\theta^\star\|) + \frac{\|\hat{\theta}_1 - \theta^\star\|}{\sqrt{2p_1}}\right)^2 (\ln \delta_t^{-1})^2\right) \end{aligned}$$

for all $t \geq T_1(\delta)$, with probability at least $1 - 4\delta \sum_{t \geq T_1(\delta)} t^{-2} \geq 1 - \delta$ because $T_1(\delta) > 4$.

Thus, as $\ln t \geq 1$ for $t \geq T_1(\delta)$ and $\|\hat{\theta}_{t+1} - \theta^\star\|_\Sigma^2 \geq \Lambda_{\min}\|\hat{\theta}_{t+1} - \theta^\star\|^2$, we obtain

$$\begin{aligned} \|\hat{\theta}_{t+1} - \theta^\star\| \leq & \frac{6\ln t}{\Lambda_{\min} t}\left(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{2p_1} + \frac{D_X^2}{\Lambda_{\min}} D_{\text{app}}^2 \frac{4(1+\sqrt{8\ln \delta^{-1}})}{0.07^2} + \frac{3\sigma^2(d/0.035 + \ln \delta^{-1})}{0.07}\right) \\ & + \frac{48(\ln t)^2}{0.07^2 \Lambda_{\min} t^2}\left(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{p_1}\frac{D_X^2}{\Lambda_{\min}}(1+\sqrt{8\ln \delta^{-1}})\right. \\ & \qquad \left. + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\text{app}} + D_X\|\theta^\star\|) + \frac{\|\hat{\theta}_1 - \theta^\star\|}{\sqrt{2p_1}}\right)^2 (\ln \delta^{-1})^2\right) \end{aligned}$$

for all $t \geq T_1(\delta$, with probability at least $1 - \delta$. Finally, both terms of the last inequality are bounded by $\varepsilon/2$. □

From Corollary B.1, we obtain the asymptotic rate by comparing $T_2(\delta)$ and $T_3(\delta)$. We write $T_2(\delta) = 2A_2(\delta)\ln A_2(\delta), T_3(\delta) = 2A_3(\delta)\ln A_3(\delta)$ with

$$\begin{aligned} A_2(\delta) &\lesssim \frac{\varepsilon^{-1}}{\Lambda_{\min}}\left(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{p_1} + \frac{D_X^2}{\Lambda_{\min}} D_{\text{app}}^2 \sqrt{\ln \delta^{-1}} + \sigma^2(d + \ln \delta^{-1})\right) \\ A_3(\delta) &\lesssim \sqrt{\frac{\varepsilon^{-1}}{\Lambda_{\min}}\left(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{p_1}\frac{D_X^2}{\Lambda_{\min}}\sqrt{\ln \delta^{-1}} + \left(\frac{D_X(D_{\text{app}} + D_X\|\theta^\star\|)}{\sqrt{\Lambda_{\min}}} + \frac{\|\hat{\theta}_1 - \theta^\star\|}{\sqrt{p_1}}\right)^2 (\ln \delta^{-1})^2\right)}\,. \end{aligned}$$

where the symbol $\lesssim$ means less than up to universal constants. As $\sqrt{a+b} \lesssim \sqrt{a} + \sqrt{b}$ and

$\sqrt{ab} \lesssim a + b$, we obtain

$$\begin{aligned} A_3(\delta) &\lesssim \sqrt{\frac{\varepsilon^{-1}}{\Lambda_{\min}}} \Bigg(\sqrt{\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{p_1} \frac{D_X^2}{\Lambda_{\min}} \sqrt{\ln \delta^{-1}}} \\ &\qquad + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}} (D_{\text{app}} + D_X \|\theta^\star\|) + \frac{\|\hat{\theta}_1 - \theta^\star\|}{\sqrt{p_1}} \right) \ln \delta^{-1} \Bigg) \\ &\lesssim \sqrt{\frac{\varepsilon^{-1}}{\Lambda_{\min}}} \Bigg(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{p_1} + \frac{D_X^2}{\Lambda_{\min}} \sqrt{\ln \delta^{-1}} \\ &\qquad + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}} (D_{\text{app}} + D_X \|\theta^\star\|) + \frac{\|\hat{\theta}_1 - \theta^\star\|}{\sqrt{p_1}} \right) \ln \delta^{-1} \Bigg). \end{aligned}$$

Thus, as long as $\frac{\varepsilon^{-1}}{\Lambda_{\min}} \leq 1$, we get

$$\begin{aligned} A_2(\delta), A_3(\delta) &\lesssim \frac{\varepsilon^{-1}}{\Lambda_{\min}} \Bigg(\frac{\|\hat{\theta}_1 - \theta^\star\|^2}{p_1} + \frac{D_X^2}{\Lambda_{\min}} (1 + D_{\text{app}}^2) \sqrt{\ln \delta^{-1}} + \sigma^2 d \\ &\qquad + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}} (D_{\text{app}} + D_X \|\theta^\star\|) + \frac{\|\hat{\theta}_1 - \theta^\star\|}{\sqrt{p_1}} + \sigma^2 \right) \ln \delta^{-1} \Bigg). \end{aligned}$$

Appendix C

Supplementary Material for Chapter 6

We provide the proofs for all the claims of Chapter 6.

Contents

C.1 Kullback-Leibler Derivation

Proof of Lemma 6.1. We start from the expression of (6.1) that we decompose as follows:

$$
\begin{aligned}
KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times \mathcal{N}(\hat{a}_{t|t}, s_{t|t}) \times \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; P_{\mathcal{F}_t}\Big) &= \mathbb{E}_{\theta_t \sim \mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})}[\log \mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})] \\
&+ \mathbb{E}_{a_t \sim \mathcal{N}(\hat{a}_{t|t}, s_{t|t})}[\log \mathcal{N}(a_t \mid \hat{a}_{t|t}, s_{t|t})] + \mathbb{E}_{b_t \sim \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})}[\log \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})] \\
&- \mathbb{E}_{(\theta_t, a_t, b_t) \sim \mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})}[\log p(\theta_t, a_t, b_t \mid \mathcal{F}_t)] \,.
\end{aligned}
$$

The last term can be split using the factorized form of (6.2). We observe that on the one hand,

$$
\begin{aligned}
&\mathbb{E}_{(\theta_t, a_t, b_t) \sim \mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})}[\log \mathcal{N}(y_t \mid \theta_t^\top x_t, \exp(a_t))] \\
&\quad = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\hat{a}_{t|t} - \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)\exp(-\hat{a}_{t|t} + \frac{1}{2}s_{t|t}) \,,
\end{aligned}
$$

and on the other hand,

$$
\begin{aligned}
&\mathbb{E}_{(\theta_t, a_t, b_t) \sim \mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})}[\log \mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^\top + f(b_t))] \\
&\quad = -\frac{d\log(2\pi)}{2} - \frac{1}{2}\mathbb{E}_{b_t \sim \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})}[\psi_t(b_t)] \,,
\end{aligned}
$$

where ψ_t is defined in the lemma. Combining the last equations with the value of the entropy of gaussian random variables yields the result. □

C.2 State Estimation

Proof of Theorem 6.1. Thanks to Lemma 6.1 we have

$$\begin{aligned}
&KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times \mathcal{N}(\hat{a}_{t|t}, s_{t|t}) \times \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; P_{\mathcal{F}_t}\Big) \\
&\quad = \frac{1}{2}\operatorname{Tr}\Big((P_{t|t} + (\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})^\top)A_t\Big) \\
&\qquad + \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)\exp(-\hat{a}_{t|t} + \frac{1}{2}s_{t|t}) - \frac{1}{2}\log\det P_{t|t} + c_\theta\,,
\end{aligned}$$

where c_θ is a constant independent of $\hat{\theta}_{t|t}, P_{t|t}$, and A_t is defined in the theorem. To conclude we write the first order conditions:

$$\begin{aligned}
&-\frac{1}{2}P_{t|t}^{-1} + \frac{1}{2}\Big(A_t + \frac{x_t x_t^\top}{\exp(\hat{a}_{t|t} - \frac{1}{2}s_{t|t})}\Big) = 0\,, \\
&-\frac{(y_t - \hat{\theta}_{t|t}^\top x_t)x_t}{\exp(\hat{a}_{t|t} - \frac{1}{2}s_{t|t})} + A_t(\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1}) = 0\,.
\end{aligned}$$

□

C.3 Observation Noise Variance Estimation

Proof of Proposition 6.2. Thanks to Lemma 6.1, we have

$$\begin{aligned}
&KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times \mathcal{N}(\hat{a}_{t|t}, s_{t|t}) \times \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; P_{\mathcal{F}_t}\Big) \\
&\quad = \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)e^{-\hat{a}_{t|t} + s_{t|t}/2} + \frac{1}{2}(s_{t-1|t-1} + \rho_a)^{-1}s_{t|t} - \frac{1}{2}\log(s_{t|t}) + c_s\,,
\end{aligned}$$

where c_s is a constant independent of $s_{t|t}$. Moreover, if $0 \le s_{t|t} \le s_{t-1|t-1} + \rho_a$ then

$$e^{s_{t|t}/2} \le e^{(s_{t-1|t-1} + \rho_a)/2} + \frac{1}{2}(s_{t|t} - (s_{t-1|t-1} + \rho_a))\,.$$

The last two equations yield the upper bound of the proposition. To obtain (6.6) we write the first order condition of optimality:

$$\frac{1}{4}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)e^{-\hat{a}_{t|t}} - \frac{1}{2}s_{t|t}^{-1} + \frac{1}{2}(s_{t-1|t-1} + \rho_a)^{-1} = 0\,.$$

□

Proof of Proposition 6.3. Thanks to Lemma 6.1 we have

$$\begin{aligned}
&KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times \mathcal{N}(\hat{a}_{t|t}, s_{t|t}) \times \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; P_{\mathcal{F}_t}\Big) \\
&\quad \le \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)e^{-\hat{a}_{t|t} + s_{t|t}/2} + \frac{1}{2}(s_{t-1|t-1} + \rho_a)^{-1}(\hat{a}_{t|t} - \hat{a}_{t-1|t-1})^2 + \frac{1}{2}\hat{a}_{t|t} + c_a\,,
\end{aligned}$$

with c_a a constant independent of $\hat{a}_{t|t}$. Moreover, if $\hat{a}_{t|t} \in [\hat{a}_{t-1|t-1} - M_a, \hat{a}_{t-1|t-1} + M_a]$ we have the following upper bound:

$$e^{-\hat{a}_{t|t}} \leq e^{-\hat{a}_{t-1|t-1}} \Big(1 - (\hat{a}_{t|t} - \hat{a}_{t-1|t-1}) + \frac{e^{M_a}}{2}(\hat{a}_{t|t} - \hat{a}_{t-1|t-1})^2\Big) .$$

The last two equations yield the upper bound of the proposition. To obtain (6.7) we write the first-order condition:

$$\frac{1}{s_{t-1|t-1} + \rho_a}(\hat{a}_{t|t} - \hat{a}_{t-1|t-1}) + \frac{1}{2} + \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t) e^{-\hat{a}_{t-1|t-1} + s_{t|t}/2} \Big(-1 + e^{M_a}(\hat{a}_{t|t} - \hat{a}_{t-1|t-1})\Big) = 0 ,$$

□

C.4 State Noise Covariance Matrix Estimation

To prove Propositions 6.4 and 6.5 we first compute the first and second derivatives of ψ_t for the scalar and diagonal settings:

Lemma C.1. *Let* $C_t = KP_{t-1|t-1}K^\top + f(b_t)$ *and* $B_t = P_{t|t} + (\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})^\top$.

— *If* $f(\cdot) = \phi(\cdot)I$ *then for any* b_t, *we have*

$$\psi_t'(b_t) = \operatorname{Tr}(C_t^{-1}(I - B_t C_t^{-1}))\phi'(b_t) ,$$
$$\psi_t''(b_t) = \operatorname{Tr}(C_t^{-1}(I - B_t C_t^{-1}))\phi''(b_t) + 2\operatorname{Tr}(C_t^{-2}(B_t C_t^{-1} - I/2))\phi'(b_t)^2 .$$

— *If* $f(\cdot) = D_{\phi(\cdot)}$ *then for any* b_t, *we have*

$$\frac{\partial \psi_t}{\partial b_t} = \Delta_{C_t^{-1}(I - B_t C_t^{-1})} \odot \phi'(b_t) ,$$
$$\frac{\partial^2 \psi_t}{\partial b_t^2} = C_t^{-1}(I - B_t C_t^{-1}) D_{\phi''(b_t)} \odot I + 2C_t^{-1}(B_t C_t^{-1} - I/2) \odot C_t^{-1} \odot \phi'(b_t)\phi'(b_t)^\top ,$$

where $\odot$ *is the Hadamard (pointwise) product.*

Proof. — In the scalar setting we recall that

$$\psi_t(b) = \log\det(KP_{t-1|t-1}K^\top + \phi(b)I) + \operatorname{Tr}(B_t(KP_{t-1|t-1}K^\top + \phi(b)I)^{-1}) .$$

We denote by log and exp the univariate logarithm and exponential and by Log the matrix logarithm. Note that if $A \succ 0$, it holds $\det A = \exp\operatorname{Tr}(\operatorname{Log} A)$. We define $C_t = KP_{t-1|t-1}K^\top + \phi(b_t)I$ and we obtain:

$$\begin{aligned}
&\log\det(KP_{t-1|t-1}K^\top + \phi(b)I) - \operatorname{Tr}\operatorname{Log}(C_t) \\
&\quad = \operatorname{Tr}\operatorname{Log}(KP_{t-1|t-1}K^\top + \phi(b)I) - \operatorname{Tr}\operatorname{Log}(C_t) \\
&\quad = \operatorname{Tr}\operatorname{Log}\Big(I + (\phi(b) - \phi(b_t))C_t^{-1}\Big) \\
&\quad = \operatorname{Tr}\Big(\Big(\phi'(b_t)(b - b_t) + \frac{1}{2}\phi''(b_t)(b - b_t)^2\Big)C_t^{-1} - \frac{1}{2}\big(\phi'(b_t)(b - b_t)C_t^{-1}\big)^2 + o((b - b_t)^2)\Big) .
\end{aligned}$$

The last line follows from the series expansion of the Logarithm. We apply another series expansion for the second term of ψ_t: we have

$$\begin{aligned}
&\mathrm{Tr}(B_t(KP_{t-1|t-1}K^\top + \phi(b)I)^{-1}) \\
&\quad = \mathrm{Tr}\left(B_t C_t^{-1}\Big(I + (\phi(b) - \phi(b_t))C_t^{-1}\Big)^{-1}\right) \\
&\quad = \mathrm{Tr}\left(B_t C_t^{-1}\Big(I - \Big(\phi'(b_t)(b-b_t) + \frac{1}{2}\phi''(b_t)(b-b_t)^2\Big)C_t^{-1} \right. \\
&\qquad\qquad \left. + \big(\phi'(b_t)(b-b_t)C_t^{-1}\big)^2 + o((b-b_t)^2)\Big)\right).
\end{aligned}$$

Summing the last two equations, and using the identity $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$, we can identify the first and second derivatives of ψ_t.

— We develop a similar argument in the diagonal setting:

$$\psi_t(b) = \log\det(KP_{t-1|t-1}K^\top + D_{\phi(b)}) + \mathrm{Tr}(B_t(KP_{t-1|t-1}K^\top + D_{\phi(b)})^{-1}),$$

then we apply the series expansion of the Logarithm:

$$\begin{aligned}
&\log\det(KP_{t-1|t-1}K^\top + D_{\phi(b)}) - \mathrm{Tr}\,\mathrm{Log}(C_t) \\
&\quad = \mathrm{Tr}\,\mathrm{Log}\left(I + D_{\phi(b)-\phi(b_t)}C_t^{-1}\right) \\
&\quad = \mathrm{Tr}\left(D_{\phi'(b_t)(b-b_t)+\frac{1}{2}\phi''(b_t)(b-b_t)^2}C_t^{-1} - \frac{1}{2}(D_{\phi'(b_t)(b-b_t)}C_t^{-1})^2 + o(\|b-b_t\|^2)\right),
\end{aligned}$$

where $C_t = KP_{t-1|t-1}K^\top + D_{\phi(b_t)}$ and $\phi'(b_t), \phi''(b_t)$ denote the coefficient-wise application of the first and second derivatives of ϕ to the vector b_t. We apply another series expansion for the second term of ψ_t:

$$\begin{aligned}
&\mathrm{Tr}\left(B_t(KP_{t-1|t-1}K^\top + D_{\phi(b)})^{-1}\right) \\
&\quad = \mathrm{Tr}\left(B_t C_t^{-1}\Big(I + D_{\phi(b)-\phi(b_t)}C_t^{-1}\Big)^{-1}\right) \\
&\quad = \mathrm{Tr}\left(B_t C_t^{-1}\Big(I - D_{\phi'(b_t)(b-b_t)+\frac{1}{2}\phi''(b_t)(b-b_t)^2}C_t^{-1} \right. \\
&\qquad\qquad \left. + (D_{\phi'(b_t)(b-b_t)}C_t^{-1})^2 + o(\|b-b_t\|^2)\Big)\right).
\end{aligned}$$

Summing the last two equations we obtain

$$\begin{aligned}
\psi_t(b) = \mathrm{Tr}\,\mathrm{Log}(C_t) + \mathrm{Tr}(B_t C_t^{-1}) + \mathrm{Tr}\left(C_t^{-1}(I - B_t C_t^{-1})D_{\phi'(b_t)(b-b_t)+\frac{1}{2}\phi''(b_t)(b-b_t)^2}\right) \\
+ \mathrm{Tr}\left(C_t^{-1}(B_t C_t^{-1} - I/2)D_{\phi'(b_t)(b-b_t)}C_t^{-1}D_{\phi'(b_t)(b-b_t)}\right) + o(\|b-b_t\|^2).
\end{aligned}$$

Then we use the identity $Tr(AD_vBD_v) = v^\top(A \odot B^\top)v$. We have

$$\begin{aligned}\psi_t(b) &= \operatorname{Tr}\operatorname{Log}(C_t) + \operatorname{Tr}(B_tC_t^{-1}) \\ &+ \frac{1}{2}(b-b_t)^\top\Big(C_t^{-1}(I - B_tC_t^{-1})D_{\phi''(b_t)} \odot I \\ &\qquad + 2C_t^{-1}(B_tC_t^{-1} - I/2) \odot C_t^{-1} \odot \phi'(b_t)\phi'(b_t)^\top\Big)(b-b_t) \\ &+ (\Delta_{C_t^{-1}(I-B_tC_t^{-1})} \odot \phi'(b_t))^\top(b-b_t) + o(\|b-b_t\|^2)\,.\end{aligned}$$

Thus we can identify the first and second derivatives of ψ_t.

□

Proof of Proposition 6.4. As long as $f(\hat{b}_{t-1|t-1}) \succ 0$ we know that f is twice differentiable in $\hat{b}_{t-1|t-1}$ and the local upper-bound property of Proposition 6.4 holds if $\frac{\partial^2\psi_t}{\partial b_t^2}|_{\hat{b}_{t-1|t-1}} \prec H_t$. We bound the expressions obtained in Lemma C.1.

— In the scalar setting,

$$\psi_t''(\hat{b}_{t-1|t-1}) = \operatorname{Tr}(C_t^{-1}(I - B_tC_t^{-1}))\phi''(\hat{b}_{t-1|t-1}) + 2\operatorname{Tr}(C_t^{-2}(B_tC_t^{-1} - I/2))\phi'(\hat{b}_{t-1|t-1})^2\,.$$

Furthermore, $C_t \succ 0$ thus $C_t^{-1} \succ 0$, $\operatorname{Tr}(C_t^{-1}) > 0$, and $\operatorname{Tr}(C_t^{-2}) > 0$. $\phi''(\hat{b}_{t-1|t-1}) = -1/(1+\hat{b}_{t-1|t-1})^2 < 0$ and $\phi'(\hat{b}_{t-1|t-1})^2 > 0$, therefore we obtain

$$\psi_t''(\hat{b}_{t-1|t-1}) < -\operatorname{Tr}(C_t^{-1}B_tC_t^{-1})\phi''(\hat{b}_{t-1|t-1}) + 2\operatorname{Tr}(C_t^{-2}B_tC_t^{-1})\phi'(\hat{b}_{t-1|t-1})^2\,.$$

— In the diagonal setting,

$$\begin{aligned}\frac{\partial^2\psi_t}{\partial b_t^2}\Big|_{\hat{b}_{t-1|t-1}} &= C_t^{-1}(I - B_tC_t^{-1})D_{\phi''(\hat{b}_{t-1|t-1})} \odot I \\ &\qquad + 2C_t^{-1}(B_tC_t^{-1} - I/2) \odot C_t^{-1} \odot \phi'(\hat{b}_{t-1|t-1})\phi'(\hat{b}_{t-1|t-1})^\top\,.\end{aligned}$$

Similarly we have $C_t^{-1} \succ 0, D_{\phi''(\hat{b}_{t-1|t-1})} \prec 0$ and as diagonal coefficients of C_t^{-1} are positive, it yields $(C_t^{-1}D_{\phi''(\hat{b}_{t-1|t-1})}) \odot I \prec 0$.
Moreover $\phi'(\hat{b}_{t-1|t-1})\phi'(\hat{b}_{t-1|t-1})^\top \succ 0$, and we can apply Schur product theorem: $C_t^{-1} \odot C_t^{-1} \odot \phi'(\hat{b}_{t-1|t-1})\phi'(\hat{b}_{t-1|t-1})^\top \succ 0$. Eventually:

$$\begin{aligned}\frac{\partial^2\psi_t}{\partial b_t^2}\Big|_{\hat{b}_{t-1|t-1}} &\prec -C_t^{-1}B_tC_t^{-1}D_{\phi''(\hat{b}_{t-1|t-1})} \odot I \\ &\qquad + 2C_t^{-1}B_tC_t^{-1} \odot C_t^{-1} \odot \phi'(\hat{b}_{t-1|t-1})\phi'(\hat{b}_{t-1|t-1})^\top\,.\end{aligned}$$

□

Proof of Proposition 6.5. Thanks to Lemma 6.1 we have:

$$\begin{aligned}KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; P_{\mathcal{F}_t}\Big) &= -\frac{1}{2}\log\det\Sigma_{t|t} + \frac{1}{2}\mathbb{E}_{b_t\sim\mathcal{N}(\hat{b}_{t|t},\Sigma_{t|t})}[\psi_t(b_t)] \\ &+ \frac{1}{2}\operatorname{Tr}\Big((\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^\top)(\Sigma_{t-1|t-1} + \rho_bI)^{-1}\Big) + c_b\,,\end{aligned}$$

where c_b is a constant independent of $\hat{b}_{t|t}, \Sigma_{t|t}$. Combining the last equation and Proposition 6.4, then using the first two moments of the gaussian distribution we obtain:

$$\begin{aligned}
&KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; P_{\mathcal{F}_t}\Big) \\
&\quad \leq -\frac{1}{2}\log\det\Sigma_{t|t} + \frac{1}{2}\psi_t(\hat{b}_{t-1|t-1}) + \frac{1}{2}\frac{\partial\psi_t}{\partial b_t}\Big|^{\top}_{\hat{b}_{t-1|t-1}}(\hat{b}_{t|t} - \hat{b}_{t-1|t-1}) \\
&\qquad + \frac{1}{4}\operatorname{Tr}\Big(H_t(\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^{\top})\Big) \\
&\qquad + \frac{1}{2}\operatorname{Tr}\Big((\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^{\top})(\Sigma_{t-1|t-1} + \rho_b I)^{-1}\Big) + c_b\,.
\end{aligned}$$

This yields the upper bound of Proposition 6.5. The recursive updates follow from the first order conditions:

$$\begin{aligned}
&-\frac{1}{2}\Sigma_{t|t}^{-1} + \frac{1}{2}\Big((\Sigma_{t-1|t-1} + \rho_b I)^{-1} + \frac{1}{2}H_t\Big) = 0\,, \\
&\Big((\Sigma_{t-1|t-1} + \rho_b I)^{-1} + \frac{1}{2}H_t\Big)(\hat{b}_{t|t} - \hat{b}_{t-1|t-1}) + \frac{1}{2}\frac{\partial\psi_t}{\partial b_t}\Big|_{\hat{b}_{t-1|t-1}} = 0\,.
\end{aligned}$$

□

Appendix D

Supplementary Material for Chapter 8

Contents

D.1 Nomenclature

The experts `AR`, `Lin`, `GAM`, `RF`, `RF_GAM`, `MLP` are the ones presented in that same order in Section 8.3.

Names of the form `model_setting` refer to the expert obtained by state-space adaptation of the model `model` with the setting `setting`. For instance, `Lin_dynamic` refers to a linear model adapted with the Kalman filter in the dynamic setting, *c.f.* Section 8.4.2.

We consider quantile variants of `RF`, denoted by `RFq` where `q` is the quantile order in percent (*e.g.* `RF40` is the quantile random forest of quantile value 0.4). We also consider a quantile variant of the dynamic MLP denoted by similar names (`MLP_dynamic60` is the quantile 0.6 of the MLP in the dynamic setting).

Furthermore, we introduce an expert named `GAM_SAT` forecasting each day with the GAM as if it were a Saturday motivated by Chapter 7.

Finally, each expert `x` yields another expert `x_corr` after intraday correction.

D.2 Day-to-day Evolution of the Forecasting Strategy

As explained in Section 8.5.4, our strategy evolved in time and we recall here every change.

— **From January 18$^{\text{th}}$ to January 24$^{\text{th}}$**: we used the following set of experts obtained by the greedy selection described in Section 8.5.4: `RF`, `RF_corr`, `RF50_corr`, `RF60_corr`, `Lin_dynamic_corr`, `Lin_viking_corr`, `GAM`, `GAM_corr`, `GAM_staticbreak_corr`, `GAM_dynamic_corr`, `GAM_viking_corr`, `GAM_dynamicbig_corr`, `RF_GAM`, `RF_GAM_corr`, `GAM_SAT_corr`, `MLP_dynamic60`, `MLP_dynamic90`, `MLP_dynamic99`. We aggregated with ML-poly with an aggregation estimated independently for each hour, with the absolute loss. We found afterward a bug in

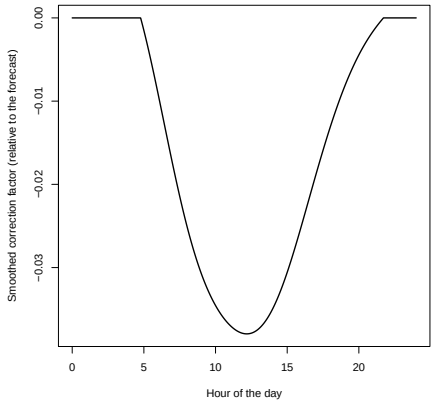


Figure D.1 – Special day correction applied on February 5^{th}. It is a multiplicative correction, *e.g.* at midday we reduce our forecast by about 3.8%.

`RF50_corr, RF60_corr`: the quantile RF were set to 0 on the test set so that these two experts were simple intraday autoregressive trained in an unwanted manner.

— **From January 25$^{\text{th}}$ to January 31$^{\text{st}}$**: we removed the experts `RF50_corr, RF60_corr` and we replaced them with `AR_corr`.

— **February 1$^{\text{st}}$ and 2$^{\text{nd}}$**: we used the uniform average between three forecasts. First, the previous aggregation. Second, another aggregation procedure called RF-stacking, consisting in a quantile random forest minimizing the MAE and taking as input the 72 experts as well as the day type and hour of the day. Third, a benchmark close to the one given by the competition organizers: we predict each time with the last available load of the same hour and the same day group (week days, saturdays and sundays).

— **From February 3$^{\text{rd}}$ to 7$^{\text{th}}$**: we removed the benchmark which damaged the performances. For **Feb. 5$^{\text{th}}$** we corrected the ML-poly prediction using a special day correction, once we observe that Feb. 5^{th} had a special behavior in the last three years. Precisely, we observed that the relative error of the model is significantly negative on the last three years, a behavior that may come from a bank holiday for instance. Therefore we fit a smoothed function of the time of day on the relative error and we applied it to our forecast. We truncated so that there is no correction during night. See the shape of the correction in Figure D.1.

— **February 8$^{\text{th}}$**: we used the single expert `Lin_dynamicbig_corr` as we observed that it was by far our best expert on the last week, and it seemed to perform especially well on Mondays.

— **February 9$^{\text{th}}$**: we came back to the average between ML-poly and the RF-stacking but we added to the aggregation ML-poly the expert `Lin_dynamicbig_corr`, and we replaced the expert `AR_corr` with another expert `AR_intra` incorporating directly the intraday correction in the autoregressive, instead of correcting an autoregressive based only on daily lags.

— **February 10$^{\text{th}}$ and 11$^{\text{th}}$**: we removed the RF-stacking which degraded our performances since its introduction and we kept only the aggregation ML-poly.

— **February 12$^{\text{th}}$ and 13$^{\text{th}}$**: we corrected *a posteriori* the electricity load for February 5^{th} with the special day correction. It was important to do it on that day as the weekly lags

is important in the models.
— **February 14$^{\mathbf{th}}$**: we used once again the average between the ML-poly aggregation and the RF-stacking, as we observed that the RF-stacking is especially good on Sunday.
— **February 15$^{\mathbf{th}}$ and 16$^{\mathbf{th}}$**: we used only the ML-poly aggregation.

Bibliography

Agamennoni, Gabriel, Juan I Nieto, and Eduardo M Nebot (2012). "Approximate inference in state-space models with heavy-tailed noise". In: *IEEE Transactions on Signal Processing* 60.10, pp. 5024–5037.

Akaike, Hirotugu (1978). "Time series analysis and control through parametric models". In: *Applied Time Series Analysis I*. Elsevier, pp. 1–23.

Almalaq, Abdulaziz and George Edwards (2017). "A Review of Deep Learning Methods Applied on Load Forecasting". In: *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 511–516. DOI: 10.1109/ICMLA.2017.0-110.

Amari, Shun-Ichi (1998). "Natural gradient works efficiently in learning". In: *Neural Computation* 10.2, pp. 251–276.

Antoniadis, Anestis et al. (2016). "A prediction interval for a function-valued forecast model: Application to load forecasting". In: *International Journal of Forecasting* 32.3, pp. 939–947.

Ba, Amadou et al. (2012). "Adaptive learning of smoothing functions: Application to electricity load forecasting". In: *Advances in neural information processing systems*, pp. 2510–2518.

Bach, Francis (2014). "Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression". In: *Journal of Machine Learning Research* 15.1, pp. 595–627.

Bach, Francis and Eric Moulines (2013). "Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n)". In: *Advances in Neural Information Processing Systems*, pp. 773–781.

Beal, Matthew James (2003). *Variational algorithms for approximate Bayesian inference*. University of London, University College London (United Kingdom).

Bercu, Bernard, Antoine Godichon, and Bruno Portier (2020). "An efficient stochastic Newton algorithm for parameter estimation in logistic regressions". In: *SIAM Journal on Control and Optimization* 58.1, pp. 348–367.

Bercu, Bernard and Abderrahmen Touati (2008). "Exponential inequalities for self-normalized martingales with applications". In: *The Annals of Applied Probability* 18.5, pp. 1848–1869.

Blackard, Jock A. and Denis J. Dean (1999). "Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables". In: *Computers and Electronics in Agriculture* 24.3, pp. 131–151.

Bougerol, Philippe (1992). "Some results on the filtering riccati equation with random parameters". In: *Applied Stochastic Analysis*. Springer, pp. 30–37.

Breiman, L. (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32.

Brockwell, Peter J. and Richard A. Davis (2016). *Introduction to time series and forecasting*. Springer.

Browell, Jethro (2021). *Supplementary Material for "Probabilistic Forecasting of Regional Net-load with Conditional Extremes and Gridded NWP"*. DOI: 10.5281/zenodo.5031704. URL: https://doi.org/10.5281/zenodo.5031704.

Browell, Jethro and Matteo Fasiolo (2021). "Probabilistic Forecasting of Regional Net-load with Conditional Extremes and Gridded NWP". In: *IEEE Transactions on Smart Grid* 12.6, pp. 5011–5019.

Bunn, D and E Dillon Farmer (1985). *Comparative models for electrical load forecasting*. John Wiley and Sons Inc., New York, NY.

Cai, Long et al. (2020). "Forecasting customers' response to incentives during peak periods: A transfer learning approach". In: *International Transactions on Electrical Energy Systems* 30.7, e12251.

Cesa-Bianchi, Nicolo and Gábor Lugosi (2006). *Prediction, Learning, and Games*. New York, NY, USA: Cambridge University Press. ISBN: 0521841089.

Charlton, Nathaniel and Colin Singleton (2014). "A refined parametric model for short term load forecasting". In: *International Journal of Forecasting* 30.2, pp. 364–368. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2013.07.003.

Chen, Yize, Weiwei Yang, and Baosen Zhang (2020). "Using Mobility for Electrical Load Forecasting During the COVID-19 Pandemic". In: *arXiv preprint arXiv:2006.08826*.

Cho, Haeran et al. (2013). "Modeling and forecasting daily electricity load curves: a hybrid approach". In: *Journal of the American Statistical Association* 108.501, pp. 7–21.

Devaine, Marie et al. (2013). "Forecasting electricity consumption by aggregating specialized experts". In: *Machine Learning* 90.2, pp. 231–260.

Diderrich, George T. (1985). "The Kalman Filter from the Perspective of Goldberger–Theil Estimators". In: *The American Statistician* 39.3, pp. 193–198.

Diebold, Francis X and Robert S Mariano (2002). "Comparing predictive accuracy". In: *Journal of Business & economic statistics* 20.1, pp. 134–144.

Dimoulkas, Ilias, Peyman Mazidi, and Lars Herre (2019). "Neural networks for GEFCom2017 probabilistic load forecasting". In: *International Journal of Forecasting* 35.4, pp. 1409–1423.

Dordonnat, Virginie (2009). "State-space modelling for high frequency data: Three applications to French national electricity load". PhD thesis. Université d'Amsterdam.

Dordonnat, Virginie, Siem Jan Koopman, and Marius Ooms (2009). "Dynamic factors in state-space models for hourly electricity load signal decomposition and forecasting". In: *2009 IEEE Power & Energy Society General Meeting*. IEEE, pp. 1–8.

Dordonnat, Virginie, Audrey Pichavant, and Amandine Pierrot (2016). "GEFCom2014 probabilistic electric load forecasting using time series and semi-parametric regression models". In: *International journal of forecasting* 32.3, pp. 1005–1011.

Duchi, John, Elad Hazan, and Yoram Singer (2011). "Adaptive subgradient methods for online learning and stochastic optimization." In: *Journal of machine learning research* 12.7.

Durbin, James and Siem Jan Koopman (2012). *Time series analysis by state space methods*. Oxford university press.

Fahrmeir, Ludwig (1992). "Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models". In: *Journal of the American Statistical Association* 87.418, pp. 501–509.

Fahrmeir, Ludwig and Gerhard Tutz (2013). *Multivariate statistical modelling based on generalized linear models*. New-York: Springer Science & Business Media.

Fan, Shu and Rob J Hyndman (2012). "Forecasting electricity demand in Australian national electricity market". In: *2012 IEEE Power and Energy Society General Meeting*. IEEE, pp. 1–4.

Farrokhabadi, Mostafa (2020). *Day-Ahead Electricity Demand Forecasting: Post-COVID Paradigm*. DOI: 10.21227/67vy-bs34. URL: https://dx.doi.org/10.21227/67vy-bs34.

Farrokhabadi, Mostafa et al. (2021). *Day-Ahead Electricity Demand Forecasting Competition: Post-COVID Paradigm*. Tech. rep.

Fasiolo, Matteo et al. (2021). "Fast calibrated additive quantile regression". In: *Journal of the American Statistical Association* 116.535, pp. 1402–1412.

Freedman, David A. (1975). "On tail probabilities for martingales". In: *the Annals of Probability*, pp. 100–118.

Gaillard, Pierre (2015). "Contributions à l'agrégation séquentielle robuste d'experts: Travaux sur l'erreur d'approximation et la prévision en loi. Applications à la prévision pour les marchés de l'énergie." PhD thesis. Université Paris Sud.

Gaillard, Pierre and Yannig Goude (2015). "Forecasting electricity consumption by aggregating experts; how to design a good set of experts". In: *Modeling and stochastic learning for forecasting in high dimensions*. Springer, pp. 95–115.

— (2016). *opera: Online Prediction by Expert Aggregation*.

Gaillard, Pierre, Yannig Goude, and Raphaël Nedellec (2016). "Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting". In: *International Journal of Forecasting* 32.3, pp. 1038 –1050. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2015.12.001.

Gaillard, Pierre, Gilles Stoltz, and Tim Van Erven (2014). "A second-order bound with excess losses". In: *Conference on Learning Theory*, pp. 176–196.

Gneiting, Tilmann and Adrian E Raftery (2007). "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American statistical Association* 102.477, pp. 359–378.

Goehry, Benjamin et al. (2019). "Aggregation of Multi-Scale Experts for Bottom-Up Load Forecasting". In: *IEEE Transactions on Smart Grid* 11.3, pp. 1895–1904.

Goude, Y., R. Nedellec, and N. Kong (2013). "Local Short and Middle term Electricity Load Forecasting with semi-parametric additive models". In: *IEEE transactions on smart grid* 5.1, pp. 440 –446.

Hale, Thomas et al. (2021). "A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)". In: *Nature Human Behaviour* 5.4, pp. 529–538.

Harvey, Andrew and Siem Jan Koopman (1993). "Forecasting hourly electricity demand using time-varying splines". In: *Journal of the American Statistical Association* 88.424, pp. 1228–1236.

Hazan, Elad, Amit Agarwal, and Satyen Kale (2007). "Logarithmic regret algorithms for online convex optimization". In: *Machine Learning* 69.2-3, pp. 169–192.

Hong, Tao and Shu Fan (2016). "Probabilistic electric load forecasting: A tutorial review". In: *International Journal of Forecasting* 32.3, pp. 914–938.

Hong, Tao, Pierre Pinson, and Shu Fan (2014). "Global Energy Forecasting Competition 2012". In: *International Journal of Forecasting* 30.2, pp. 357 –363. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2013.07.001.

Hong, Tao, Jingrui Xie, and Jonathan Black (2019). "Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting". In: *International Journal of Forecasting* 35.4, pp. 1389–1399.

Hong, Tao et al. (2016). "Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond". In: *International Journal of Forecasting* 32.3, pp. 896–913. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2016.02.001.

Hsu, Daniel, Sham M. Kakade, and Tong Zhang (2012). "Random design analysis of ridge regression". In: *Conference on Learning Theory*, pp. 9–1.

Huang, Shyh-Jier and Kuang-Rong Shih (2003). "Short-term load forecasting via ARMA model identification including non-Gaussian process considerations". In: *IEEE Transactions on power systems* 18.2, pp. 673–679.

Huang, Yulong et al. (2017). "A novel adaptive Kalman filter with inaccurate process and measurement noise covariance matrices". In: *IEEE Transactions on Automatic Control* 63.2, pp. 594–601.

Huang, Yulong et al. (2020). "A Slide Window Variational Adaptive Kalman Filter". In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 67.12, pp. 3552–3556.

IEA (2020). *Year-on-year change in weekly electricity demand, weather corrected, in selected countries*. `https://www.iea.org/data-and-statistics/charts/year-on-year-change-in-weekly-electricity-demand-weather-corrected-in-selected-countries-january-december-2020`.

Jazwinski, AH (1970). *Stochastic processes and filtering theory*. New York: Academic Press.

Julier, Simon J and Jeffrey K Uhlmann (1997). "New extension of the Kalman filter to nonlinear systems". In: *Signal processing, sensor fusion, and target recognition VI*. Vol. 3068. International Society for Optics and Photonics, pp. 182–193.

Kakade, Sham M. and Andrew Y. Ng (2005). "Online bounds for Bayesian algorithms". In: *Advances in Neural Information Processing Systems*, pp. 641–648.

Kalman, Rudolph E. and Richard S. Bucy (1961). "New results in linear filtering and prediction theory". In: *Journal of Basic Engineering* 83.1, pp. 95–108.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kingma, Diederik P and Max Welling (2014). "Stochastic gradient VB and the variational auto-encoder". In: *Second International Conference on Learning Representations, ICLR*. Vol. 19, p. 121.

Knowles, David A (2015). "Stochastic gradient variational Bayes for gamma approximating distributions". In: *arXiv preprint arXiv:1509.01631*.

Koenker, Roger and Gilbert Bassett Jr (1978). "Regression quantiles". In: *Econometrica: journal of the Econometric Society*, pp. 33–50.

Koenker, Roger and Kevin F Hallock (2001). "Quantile regression". In: *Journal of economic perspectives* 15.4, pp. 143–156.

Kohavi, Ron (1996). "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid." In: *The International Conference on Knowledge Discovery and Data Mining*. Vol. 96, pp. 202–207.

Koren, Tomer (2013). "Open problem: Fast stochastic exp-concave optimization". In: *Conference on Learning Theory*, pp. 1073–1075.

Laptev, Nikolay, Jiafan Yu, and Ram Rajagopal (2018). "Reconstruction and regression loss for time-series transfer learning". In: *Proceedings of the Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) and the 4th Workshop on the Mining and LEarning from Time Series (MiLeTS), London, UK*. Vol. 20.

Lloyd, James Robert (2014). "GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes". In: *International Journal of Forecasting* 30.2, pp. 369–374.

Mahdavi, Mehrdad, Lijun Zhang, and Rong Jin (2015). "Lower and upper bounds on the generalization of stochastic exponentially concave optimization". In: *Conference on Learning Theory*, pp. 1305–1320.

McCullagh, Peter and John A. Nelder (1989). *Generalized Linear Models*. 2nd ed. London Chapman and Hall.

Mehra, Raman (1972). "Approaches to adaptive filtering". In: *IEEE Transactions on Automatic Control* 17.5, pp. 693–698.

Murata, Noboru and Shun-ichi Amari (1999). "Statistical analysis of learning dynamics". In: *Signal Processing* 74.1, pp. 3–28.

Nagbe, Komi (2020). "France short-term load demand forecasting using a functional state space adaptative model: case of covid-19 lockdown period". In.

Narajewski, Michał and Florian Ziel (2020). "Changes in electricity demand pattern in Europe due to COVID-19 shutdowns". In: *arXiv preprint arXiv:2004.14864*.

Nedellec, Raphael, Jairo Cugliari, and Yannig Goude (2014). "GEFCom2012: Electric load forecasting and backcasting with semi-parametric models". In: *International Journal of forecasting* 30.2, pp. 375–381.

Newey, Whitney K and James L Powell (1987). "Asymmetric least squares estimation and testing". In: *Econometrica: Journal of the Econometric Society*, pp. 819–847.

Nti, Isaac Kofi et al. (2020). "Electricity load forecasting: a systematic review". In: *Journal of Electrical Systems and Information Technology* 7.1, pp. 1–19.

Obst, David, Joseph de Vilmarest, and Yannig Goude (2021). "Adaptive Methods for Short-Term Electricity Load Forecasting During COVID-19 Lockdown in France". In: *IEEE Transactions on Power Systems* 36.5, pp. 4754–4763.

Obst, David et al. (2021). "Transfer Learning for Linear Regression: a Statistical Test of Gain". In: *arXiv preprint arXiv:2102.09504*.

Ollivier, Yann (2018). "Online natural gradient as a Kalman filter". In: *Electronic Journal of Statistics* 12.2, pp. 2930–2961.

Ostrovskii, Dmitrii M. and Francis Bach (2021). "Finite-sample analysis of M-estimators using self-concordance". In: *Electronic Journal of Statistics* 15.1, pp. 326–391.

Pan, S. J. and Q. Yang (2010). "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10, pp. 1345–1359. DOI: 10.1109/TKDE.2009.191.

Park, Dong C et al. (1991). "Electric load forecasting using an artificial neural network". In: *IEEE transactions on Power Systems* 6.2, pp. 442–449.

Pierrot, Amandine and Yannig Goude (2011). "Short-term electricity load forecasting with generalized additive models". In: *Proceedings of ISAP power*.

Pinto, Tiago et al. (2021). *Competition on building energy consumption forecasting*. http://www.gecad.isep.ipp.pt/smartgridcompetitions/.

Polyak, Boris T. and Anatoli B. Juditsky (1992). "Acceleration of stochastic approximation by averaging". In: *SIAM Journal on Control and Optimization* 30.4, pp. 838–855.

Rigollet, Phillippe and Jan-Christian Hütter (2015). "High dimensional statistics". In: *Lecture notes for course 18S997*.

Robbins, Herbert and Sutton Monro (1951). "A stochastic approximation method". In: *The Annals of Mathematical Statistics*, pp. 400–407.

Robbins, Herbert and David Siegmund (1971). "A convergence theorem for non negative almost supermartingales and some applications". In: *Optimizing methods in statistics*. Elsevier, pp. 233–257.

Ruan, Guangchun et al. (2020). "A cross-domain approach to analyzing the short-run impact of COVID-19 on the US electricity sector". In: *Joule* 4.11, pp. 2322–2337.

Ruppert, David (1988). *Efficient estimations from a slowly convergent Robbins-Monro process*. Tech. rep. Cornell University Operations Research and Industrial Engineering.

Ryu, Seunghyoung, Jaekoo Noh, and Hongseok Kim (2017). "Deep neural network based demand side short term load forecasting". In: *Energies* 10.1, p. 3.

Särkkä, Simo and Jouni Hartikainen (2013). "Non-linear noise adaptive Kalman filtering via variational Bayes". In: *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, pp. 1–6.

Sarkka, Simo and Aapo Nummenmaa (2009). "Recursive noise adaptive Kalman filtering by variational Bayesian approximations". In: *IEEE Transactions on Automatic Control* 54.3, pp. 596–600.

Schnabel, Sabine K and Paul HC Eilers (2013). "A location-scale model for non-crossing expectile curves". In: *Stat* 2.1, pp. 171–183.

Shin, Hoo-Chang et al. (2016). "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning". In: *IEEE transactions on medical imaging* 35.5, pp. 1285–1298.

Šmídl, Václav and Anthony Quinn (2006). *The variational Bayes method in signal processing*. Springer Science & Business Media.

Smyl, Slawek and N Grace Hua (2019). "Machine learning methods for GEFCom2017 probabilistic load forecasting". In: *International Journal of Forecasting* 35.4, pp. 1424–1431.

Tjandra, Andros et al. (2015). "Stochastic gradient variational bayes for deep learning-based ASR". In: *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, pp. 175–180.

Tropp, Joel A. (2012). "User-Friendly Tail Bounds for Sums of Random Matrices". In: *Foundations of Computational Mathematics* 12.4, pp. 389–434.

Tyagi, Ambrish and James W Davis (2008). "A recursive filter for linear systems on Riemannian manifolds". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1–8.

Tzikas, Dimitris G, Aristidis C Likas, and Nikolaos P Galatsanos (2008). "The variational approximation for Bayesian inference". In: *IEEE Signal Processing Magazine* 25.6, pp. 131–146.

Vilmarest, Joseph de, Yannig Goude, and Olivier Wintenberger (2021). *VIKING: Variational Bayesian Variance Tracking Winning a Post-Covid Day-Ahead Electricity Load Forecasting Competition*. `https://roseyu.com/time-series-workshop/submissions/2021/TSW-ICML2021_paper_15.pdf`.

Watson, Mark W and Robert F Engle (1983). "Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models". In: *Journal of Econometrics* 23.3, pp. 385–400.

Weiss, Karl, Taghi M Khoshgoftaar, and DingDing Wang (2016). "A survey of transfer learning". In: *Journal of Big data* 3.1, p. 9.

Wood, Simon (2015). "Package 'mgcv'". In: *R package version* 1, p. 29.

Wood, Simon N (2017). *Generalized additive models: an introduction with R*. CRC press.

Wood, Simon N, Yannig Goude, and Simon Shaw (2015). "Generalized additive models for large data sets". In: *Journal of the Royal Statistical Society: Series C: Applied Statistics* 64, pp. 139–155.

Yang, Dongchuan et al. (2021). "Knowledge Mapping in Electricity Demand Forecasting: A Scientometric Insight". In: *Frontiers in Energy Research* 9, p. 633. ISSN: 2296-598X.

Zhang, Zichen, Shifei Ding, and Yuting Sun (2020). "A support vector regression model hybridized with chaotic krill herd algorithm and empirical mode decomposition for regression task". In: *Neurocomputing* 410, pp. 185 –201. ISSN: 0925-2312. DOI: `https://doi.org/10.1016/j.neucom.2020.05.075`.

Ziel, Florian (2019). "Quantile regression for the qualifying match of GEFCom2017 probabilistic load forecasting". In: *International Journal of Forecasting* 35.4, pp. 1400–1408. ISSN: 0169-2070. DOI: `https://doi.org/10.1016/j.ijforecast.2018.07.004`.

Zinkevich, Martin (2003). "Online convex programming and generalized infinitesimal gradient ascent". In: *International Conference on Machine Learning*, pp. 928–936.

Modèles espace-état pour la prévision de séries temporelles. Application aux marchés électriques.

Résumé

L'électricité étant difficile à stocker, prévoir la demande est un enjeu majeur pour maintenir l'équilibre entre la production et la consommation. L'évolution des usages de l'électricité, le déploiement des énergies renouvelables, et plus récemment la crise du coronavirus, motivent l'étude de modèles qui évoluent au cours du temps, pour tenir compte des changements de comportements. L'objectif de ce travail est de proposer des méthodes adaptatives de prévision, et nous nous sommes intéressés tout spécialement au cadre des modèles espace-état. Dans ce paradigme, on représente l'environnement (ou le contexte) par un état caché. À chaque instant, la demande dépend de cet état que nous cherchons donc à estimer grâce aux observations dont nous disposons, et selon les hypothèses que l'on effectue sur la dynamique du système. L'estimation de l'état nous permet ensuite de prévoir la demande.
Un premier objectif de la thèse est de contribuer au lien entre l'optimisation et l'estimation dans les modèles espace-état. Nous interprétons en effet les méthodes que nous utilisons comme diverses façons de paramétrer un algorithme de descente de gradient de second ordre, et nous avons détaillé ce lien dans un cas particulier. Une seconde contribution de la thèse est de proposer différentes méthodes d'estimation dans les modèles espace-état. Le principal enjeu nous semble être de définir la dynamique avec lequel évolue l'état, et nous proposons deux méthodes dans ce but. Le troisième apport de ce manuscrit est d'appliquer ces méthodes espace-état à la prévision de consommation d'électricité. Nos prévisions s'appuient sur des modèles de prévision existants, par exemple le modèle additif généralisé, que nous cherchons à adapter. Ainsi, nous tirons parti de certaines dépendances complexes capturées par les modèles existants, par exemple la sensibilité de la consommation d'électricité à la température, tout en profitant de la faculté d'adaptation des modèles espace-état.

Mots clés : modèles espace-état, prévision de consommation électrique, séries temporelles

---

Abstract

Electricity storage capacities are still negligible compared to the demand. Therefore, it is fundamental to maintain the equilibrium between consumption and production, and to that end, we need load forecasting. Numerous patterns motivate the study of time-varying models, including: changes in people's habits, increasing renewable capacities, more recently the coronavirus crisis. This thesis aims to propose adaptive methods for time series forecasting. We focus on state-space models, where the environment (or context) is represented by a hidden state on which the demand depends. Thus, we try to estimate that state based on the observations at our disposal. Based on our estimate, we forecast the load.
The first objective of the thesis is to enrich the link between optimization and state-space estimation. Indeed, we see our methods as second-order stochastic gradient descent algorithms, and we treat a particular case to detail that link. The second contribution concerns variance estimation in state-space models. Indeed, the variances are the parameters on which the models' dynamics crucially relies. The third part of the manuscript is the application of these methods to electricity load forecasting. Our methods build on existing forecasting methods like generalized additive models. The procedure allows to leverage advantages of both. On the one hand, statistical models learn complex relations to explanatory variables like temperature. On the other hand, state-space methods yield model adaptation.

Keywords: electricity load forecasting, state-space models, time series

---

Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France